UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
DOUTORADO EM CIÊNCIA DA INFORMAÇÃO

LEVI ALÃ NEVES DOS SANTOS

MÍNIMOS QUADRADOS ORDINÁRIOS (MQO) NA PRODUÇÃO CIENTÍFICA BRASILEIRA: A INTERDISCIPLINARIDADE ENTRE A ECONOMETRIA E AS METRIAS DA INFORMAÇÃO (BIBLIOMETRIA, INFORMETRIA E CIENTOMETRIA)

Salvador

2017

LEVI ALÃ NEVES DOS SANTOS

MÍNIMOS QUADRADOS ORDINÁRIOS (MQO) NA PRODUÇÃO CIENTÍFICA BRASILEIRA: A INTERDISCIPLINARIDADE ENTRE A ECONOMETRIA E AS METRIAS DA INFORMAÇÃO (BIBLIOMETRIA, INFORMETRIA E CIENTOMETRIA)

Tese apresentada ao Curso de Doutorado em Ciência da Informação do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Bahia (PPGCI/UFBA) como requisito para obtenção de grau de Doutor em Ciência da Informação.

Orientadora: Profa. Dra. Lidia Brandão Toutain

Salvador

2017

S237 Santos, Levi Alã Neves dos
Mínimos quadrados ordinários (MQO) na produção científica brasileira: a interdisciplinaridade entre a econometria e as metrias da informação (bibliometria, informetria e cientometria) / Levi Alã Neves dos Santos. – Salvador, 2017.
187 f.; il.

Orientadora: Profa. Dra. Lidia Brandão Toutain
Tese (doutorado) - Universidade Federal da Bahia. Instituto de Ciência da Informação, 2017.

1. Mínimos Quadrados Ordinários (MQO). 2. Produção científica. 3. Bibliometria. 4. Informetria. 5. Cienciometria. 6. Econometria. I. Universidade Federal da Bahia. II. Toutain, Lidia Brandão. III. Título.

LEVI ALÃ NEVES DOS SANTOS

MÍNIMOS QUADRADOS ORDINÁRIOS (MQO) NA PRODUÇÃO CIENTÍFICA BRASILEIRA: A INTERDISCIPLINARIDADE ENTRE A ECONOMETRIA E AS METRIAS DA INFORMAÇÃO (BIBLIOMETRIA, INFORMETRIA E CIENTOMETRIA)

Tese apresentada ao Curso de Doutorado em Ciência da Informação do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal da Bahia – PPGCI/UFBA como requisito para obtenção de grau de Doutor em Ciência da Informação.

Aprovado em: 05 / 12 / 2017

Banca Examinadora

Prof.ª Dra. Lídia Maria Batista Brandão Toutain – Orientadora – UFBA

Prof.ª Dra. Olivia Maria Cordeiro de Oliveira – Membro Externo Titular – UFBA

Prof. Dr. Marcelo Albano Morel Simões Gonçalves – Membro Externo Titular – SENAI

Prof. Dr. Valter Senna - Membro Externo Titular – SENAI

Prof.ª Drª Ana Paula de Oliveira Villalobos – Membro Interno Titular – UFBA

Heitor, Arthur e Cecília, meus amados filhos, para mim
tudo só tem significado por vocês existirem.

AGRADECIMENTOS

Á Deus, que permite todas as coisas. Muito obrigado, Senhor.

Aos meus pais, Martins e Tânia, irmãos, Hales e Aline, e sobrinho Eduardo, por serem perseverantes em me mostrar o caminho seguro para a felicidade. Amo vocês.

Á minha esposa, Silvana Bastos Paula, por ser tão especial em minha vida e a luz em horas escuras. Te amo.

Ao meu amigo Samir Elias Kalil Lion, sempre junto e decisivo. Obrigado meu irmão.

À minha orientadora, professora doutora e pesquisadora, Lidia Brandão Toutain, por acreditar na importância e relevância deste estudo e incentivar seu desenvolvimento.

À minha equipe do Instituto de Geociências, chefe atencioso, colegas compreensíveis e parceiros. Obrigado, vocês são 10^{10}.

Aos professores do ICI, por conduzirem um aprendizado que vai além da técnica, impregnado de valores morais e éticos. Verdadeiros mestres.

A toda equipe do ICI, sempre muito dispostos a ajudar. Valeu por tudo.

RESUMO

Analisa a produção científica brasileira (artigos nacionais, artigos internacionais, anais de eventos e livros) através dos Mínimos Quadrados Ordinários (MQO). Para tanto, discorre sobre o percurso histórico e de aplicação das metrias que a Ciência da Informação (CI) vem construindo, desde a mais primordial de todas, a bibliometria, oriunda da biblioteconomia, passando pelas visões modernas como a cienciometria até a informetria. Explica como a econometria constrói o seu modelo de análise, que é utilizado para pesquisas na economia e, ao mesmo tempo, reflete como esse método pode ser trazido para as metrias da informação. Explica e expõe o método de estimação por MQO para a análise de regressão, que é a proposta desta tese. Pesquisa aplicada descritiva com abordagem quantitativa com procedimentos baseados no tipo de pesquisa estudo de caso do levantamento de dados a partir do Portal do Plano Tabular do CNPq do ano de 2010. Os critérios para delineamento da pesquisa foram aprofundados, na revisão de literatura, em referências tanto da área da CI quanto da bibliometria, estatística e econometria. Este estudo, metodologicamente, conta com a abordagem conceitual da bibliometria e da CI em busca de teorias aplicáveis aos estudos em MQO e a aplicação empírica do MQO se aproxima da concepção econométrica. A tese conclui que a utilização de técnicas de análises das funções de regressão construída por meio de MQO possibilita a criação de um modelo de previsão da produção científica brasileira. Esse modelo é construído a partir da correlação e determinação detectada entre o número de doutores e a produção científica destes em cada estado do Brasil. Com a aplicação de estratégias econométricas (índice de correlação, índice de determinação, forma funcional de curva de regressão e cálculo dos parâmetros da função por MQO), foi possível construir um modelo de previsão.

Palavras-chave: Ciência da Informação. Mínimos Quadrados Ordinários (MQO). Informetria. Bibliometria. Lotka, Bradford, Zipf, Goffman. Cienciometria. Cientometria. Econometria.

ABSTRACT

It analyzes Brazilian scientific production (national articles, international articles, annals of events and books) through the Ordinary Least Squares (OLS). Then, it discusses the historical path and application metrics that the Information Science (CI) has been building, from the most primordial of all, bibliometrics, coming librarianship, and passing through modern visions such as scienometry to informetria. Explains how econometrics constructs its analysis model, used for research in the economy and, at the same time, reflects how this method can brought to the metrics of information. Explains the method of estimation by OLS for regression analysis and exposes the estimation method by OLS, applied to the regression analysis proposed in this thesis. Descriptive applied research with quantitative approach with procedures based on the of research case study of data collection from the Portal of the Tabular Plan of the CNPq of the year 2010. The criteria for the delineation of the research deepened in the literature review in references both in the IC area and in bibliometrics, statistics and econometrics. This study, methodologically, relies on the conceptual approach of bibliometry and CI in search of theories applicable to OLS studies and the empirical application of OLS approaches econometric conception. The thesis concludes that use of analysis techniques of the functions of regression constructed by of OLS makes possible the creation of a prediction model of Brazilian scientific production. This model constructed from the correlation and detected determination the number of doctors and the scientific production of these in eachstate of Brazil. With the application of econometric strategies (correlation index, determination index, functional form of regression curve and calculation of the parameters of the function by OLS), it was possible to construct a prediction model.

Keywords: Information Science-Ordinary Least Squares (OLS). Informetrics. Bibliometrics- Lotka, Bradford, Zipf, Goffman. Scientometrics. Econometrics.

.

LISTA DE ABREVIATURAS

CI – Ciência da Informação

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico

FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo

FRA – Função de Regressão Amostral

FRP – Função de Regressão Populacional

IBGE – Instituto Brasileiro de Geografia e Estatística

MCT – Ministério de Ciência e Tecnologia

MQO – Mínimos Quadrados Ordinários

PCT – Pesquisa Científica e Tecnológica

SQE – Soma dos Quadrados Explicados pela Regressão

SQR – Soma dos Quadrados dos Resíduos

STQ – Soma Total de Quadrados

TIC(s) – Tecnologia(s) da Informação e Comunicação

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE GRÁFICOS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O método dos Quadrados Mínimos Ordinários (MQO) é uma técnica de otimização estatística utilizada para encontrar o melhor ajuste para um conjunto de dados. Essa técnica consiste em minimizar a soma dos quadrados das diferenças (resíduos) entre o valor estimado e os dados observados (reais). É amplamente utilizada por diversas áreas por permitir, através da minimização da Soma dos Quadrados dos Resíduos da regressão, maximizar o grau de ajuste do modelo aos dados observados.

O MQO é um método amplamente utilizado em diversas áreas (engenharia, medicina, biologia etc.), aqui aplicado segundo a experiência da economia e sua ciência auxiliar a econometria. Propõe-se ser uma ferramenta adicional para as tradicionais formas de medir os fenômenos informacionais (bibliometria, econometria, informetria etc.) que, além de demonstrar a relação entre valores, permite determinar modelos de previsão. Os modelos criados por meio de MQO trazem a força da relação entre variáveis e permitem simular resultados a partir de variáveis, determinando a precisão desses valores.

A visão econométrica das aplicações do MQO foi escolhida por trazer modelos de análises como métodos de interpretação de teorias econômicas. A econometria vai além da utilização da estatística como mera ferramenta, pois os modelos criados fazem parte do corpo de conhecimento e dos estudos econômicos. Apresenta uma relação entre dados, métodos e teoria, o que aparentemente não acontece em outras áreas que aplicavam o MQO como ferramenta estatística, ou seja, sem uma relação teórica com a área. Empiricamente, os métodos de análise de dados, que utilizam conceitos e teorias estatísticas, estão mais consolidados na disciplina Econometria (aplicabilidade em pesquisas de análises quantitativas), do que nas áreas que medem a informação. Isso porque a Econometria enseja metodologias bastante consistentes para os estudos de dados econômicos (produção, consumo, distribuição, demanda, oferta, juros, renda nacional, balança internacional etc.).

A economia antes das ferramentas métricas já possui explicações (teorias) para a produção, consumo, distribuição, demanda, oferta, juros, renda nacional e balança internacional. Com a utilização das ferramentas matemáticas e estatísticas, as teorias econômicas encontraram explicações quantitativas para os fenômenos econômicos, inclusive com a possibilidade de contar com modelos de previsão. A econometria é o encontro dessas áreas (teoria econômica e ferramentas matemáticas e estatísticas). Isto é, é uma interdisciplina

já consolidada que dá objetividade a fenômenos sociais, coletivos e comportamentais que caracterizam a teoria econômica.

No contexto de competitividade, baseada na contínua evolução de tecnologias e na crescente modernização das práticas administrativas, existe a necessidade da utilização de métodos capazes de determinar o que foi produzido e, em um ciclo contínuo, entender, adaptar e modificar tais práticas de modo a ajustar os referidos avanços às condições concretas de cada cenário real observado.

Dentre as preocupações crescentes no mundo moderno, no que tange ao desenvolvimento científico e tecnológico, existe a preocupação de medir a produção científica nos países através de seus produtores, produtos e produtividade efetiva. No Brasil, o Ministério de Ciência e Tecnologia (MCT) vem adotando, há algumas décadas, várias medidas para implementar tais mecanismos que gerenciem o controle do processo produtivo retratando o cenário nacional. Essa preocupação se ocupa não só com o desenvolvimento nacional, mas também em tornar o país mais competitivo no mercado mundial com base na exploração e compreensão dos dados de sua produção.

É nesse sentido que o presente estudo tem como objetivo geral analisar os processos métricos da produção científica brasileira (artigos nacionais, artigos internacionais, anais de eventos e livros) através do MQO frente aos modelos tradicionais de metrias da informação, que poderá auxiliar na construção de instrumentos que permitam maior suporte à avaliação e às decisões tanto do MCT quanto das agências de fomento e, assim, racionalizar e flexibilizar tanto a aplicação de recursos públicos quanto a definição de políticas nos estados e na Federação. Temos como objetivos específicos: a) discorrer sobre o percurso histórico e de aplicação das metrias, que a Ciência da Informação (CI) vem construindo, desde a mais primordial de todas, a bibliometria oriunda da biblioteconomia, passando pelas visões modernas como a cienciometria até a informetria; b) explicar como a econometria constrói o seu modelo de análise, que é utilizado para pesquisas na economia e, ao mesmo tempo, refletir como esse método pode ser trazido para as metrias da informação; c) explicar o método de estimação por MQO para a análise de regressão; d) expor o método de estimação por MQO, aplicado à análise de regressão a que se propõe esta tese; e) apresentar o modelo de previsão da produção científica nacional.

A partir desse percurso, respondemos à questão de pesquisa: como analisar a produção científica brasileira (artigos nacionais, artigos internacionais, anais de eventos e livros) utilizando os MQO? Os critérios para delineamento da pesquisa foram aprofundados na revisão de literatura, em referências tanto da área da CI quanto da bibliometria, informetria,

cientometria, estatística e econometria. Este estudo está metodologicamente dividido em duas partes: a abordagem conceitual da bibliometria e da CI e a aplicação empírica das técnicas da econometria e estatística.

O modelo foi construído com a técnica de MQO a partir da correlação entre o número de doutores e a produção científica. Com a aplicação de estratégias econométricas foi possível construir uma análise da produção científica brasileira, contendo inclusive um modelo de previsão.

Quanto ao nível, é uma pesquisa aplicada descritiva com abordagem estatística, por haver o uso de métodos quantificáveis. O procedimento é baseado em um estudo de caso, por se tratar de um levantamento de dados a partir do Portal do Plano Tabular do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Para viabilizar a pesquisa, foram reunidos dados da produção científica brasileira do período 2007-2010. Também se observou a existência de correlação entre a produção científica e o número de doutores autores.

Formulários na forma de tabelas foram construídos como instrumento de coleta de dados para que nelas constassem organizados os dados gerados. Isso pode ser feito em tabelas dinâmicas configuradas pelo usuário na consulta ao perfil de pesquisa no Brasil do Portal do Plano Tabular do CNPq.

Nesta tese, o medir foi explorado a partir de conceitos da bibliometria de Otlet, aliados às técnicas estatísticas para análise de dados. Esta pesquisa aponta para a valorização da possibilidade de recorrer aos conhecimentos da econometria (matemática e estatística através de MQO) para ampliar a área da compreensão de dados (medidas da informação) sobre a produção científica brasileira (fenômeno informacional). Conceitos da informetria, bibliometria e cientometria foram trazidos para reforçar a interdisciplinaridade entre as metrias da informação e a econometria.

O interesse do estudo surgiu através do entendimento que o número de publicações (artigos nacionais, artigos internacionais, anais de eventos e livros) apresentado por cada estado da Federação pode ser explicado pelo número de doutores de cada um desses estados. Para realizar a pesquisa, foram utilizadas técnicas estatísticas com base em métodos econométricos.

A tese está estruturada em capítulos delineados de acordo com os elementos considerados necessários à formulação e execução da atividade de estudo acadêmico. As seções dos capítulos encontram-se articulados do ponto de vista conceitual e da argumentação proposta pelo autor para alcançar o objetivo da tese.

Nesta introdução, destacamos resumidamente o campo conceitual, apresentamos a questão pesquisa, os objetivos gerais e específicos e a metodologia do procedimento de

pesquisa. A tese foi construída tomando por base duas áreas de conhecimentos interdisciplinares. O arcabouço teórico e a estrutura metodológica são emprestadas da CI e é também de onde surge a questão da pesquisa. As ferramentas e técnicas são oriundas da econometria.

A pesquisa utilizou a técnica dos MQO como forma de estudar as metrias da informação e propor o seu uso na análise da produção científica brasileira. A criação do MQO é atribuída ao matemático alemão Carl Friedrich Gauss e tem propriedades estatísticas que o torna um dos métodos de análise de regressão mais poderosos e difundidos em análise de dados estatísticos, justificando a sua proposição no âmbito desta tese. Para atingir o objetivo geral e os objetivos específicos desta pesquisa, a presente o referencial teórico estruturada em quatro capítulos.

O referencial possui capítulos que se constituem da apresentação dos fundamentos teóricos identificados durante a revisão de literatura, que subsidiaram a análise dos dados coletados acerca do objeto estudado, a partir das aplicações econométricas do MQO. Também apresenta o percurso histórico do termo "bibliometria", dos antecedentes até a sua consolidação, suas questões filosóficas e epistemológicas a partir das impressões apontadas por Edson Nery da Fonseca (1986), em sua obra *Bibliometria: teoria e prática*, a partir dos trabalhos "Traitê de documentation", de Otlet (1934), e as ideias de Robert Estivals (1965), com a obra *Création, consommation et production intellectuelles*, que apresentam a relação histórica como disciplina auxiliar, inicialmente da bibliografia, e suas aproximações com as ciências naturais (quantitativamente) e com as ciências sociais (qualitativamente), bem como todo panorama histórico das metrias e o escopo dos estudos.

O segundo capítulo traz como objetivo discorrer sobre o percurso histórico e de aplicação das metrias, que a CI vem construindo desde a mais primordial de todas, a bibliometria oriunda da biblioteconomia, passando pelas visões modernas, como a cienciometria, até a informetria, que é uma das metrias que mais se aproxima desta tese.

O terceiro explica como a econometria constrói o seu modelo de análise, que é utilizado para pesquisas na economia e, ao mesmo tempo, refletir como esse método pode ser trazido para as metrias da informação. Essa reflexão é necessária porque a econometria aplica eficientemente a estatística matemática a dados econômicos e, ao nosso ver, isso ainda não acontece com as metrias da informação, o que quer dizer que a informetria, a bibliometria, a cientometria, a webometria, entre outras, ainda não conseguem aplicar eficientemente a estatística matemática aos fenômenos informacionais.

O quarto capítulo explica o método de estimação por MQO para a análise de regressão a que se propõe esta tese, com a explicação do método através de percurso técnico. Isso parte

do entendimento de que a área das metrias da informação (bibliometria, informetria, cienciometria) se apresenta, de maneira geral, como um conjunto de técnicas estatísticas e matemáticas para mensurar as atividades informacionais, agregadas ao longo do tempo, da qual a mais antiga é a bibliometria, o que permite uma abertura para a agregação de novas técnicas e/ou abordagens (MQO) que possam contribuir com a ampliação da eficiência da mensuração do conhecimento e de sua produção.

A última parte da construção teórica, e quinto capítulo, apresenta um estudo pré-existente de Santos[1] (2011), sobre métodos de análise de dados da produção científica brasileira, utilizando informação da tabela de produção bibliográfica segundo unidades da Federação para pesquisadores doutores, 2005-2008, Censo 2008, cujos gráficos são retomados para a análise destes com a utilização do método proposto por esta tese, o MQO.

O sexto capítulo descreve a construção do quadro metodológico. A análise da produção científica brasileira através da aplicação do MQO foi realizada no sétimo capítulo e as considerações finais se encontram no oitavo capítulo, com recomendação para utilização do modelo pela CI como uma nova proposta de abordagem para a análise dos fenômenos informacionais.

A contribuição desta pesquisa segue a direção de entender os números provenientes do quantitativo do registro do conhecimento científico produzido no Brasil. A história do conhecimento passou por diversas etapas como o saber, o registro desse saber e o medir a quantidade de registros do saber humano. Isso na verdade é um ciclo, em que cada etapa passa, ocasionalmente, por processos de evolução.

[1] Dissertação defendida no Programa de Pós-graduação em Ciência da Informação da UFBA, elaborada por mim, sob a orientação da professora Dra. Ana Paula de Oliveira Villalobos.

2 METRIAS DA INFORMAÇÃO NA ÁREA DA CIÊNCIA DA INFORMAÇÃO

A Ciência da Informação (CI), que introduz o novo cenário da informação, tem em sua história fatos complexos e multidimensionais envolvendo biblioteconomia, arquivologia e museologia, e tem, segundo Borko (1968), tanto um componente de ciência pura, que investiga o assunto sem considerar a sua aplicação, quanto um componente de ciência aplicada, que desenvolve serviços e produtos. Mas também existem alguns fatos que trouxeram novas reflaxões teóricas advindas de diversas outras áreas que contribuiram e continuam a contribuir para a edificação e avanço da CI.

> [...] um primeiro marco daquilo que viria a ser a Ciência da Informação encontra-se na área da Documentação, criada por Otlet e La Fontaine no início do século XX. Voltados inicialmente para a questão da Bibliografia, estes dois pesquisadores empreenderam uma série de esforços para garantir uma rede de atuação internacional em prol da inventariação de toda a produção intelectual humana. [...] Em 1934, Otlet, preocupado com a sustentação teórica de sua proposta, escreve seu *Traité de Documentation*, no qual desenvolveu o conceito de 'documento', alargando o campo de intervenção para além dos livros e registros impressos (ARAÚJO, 2013, p. 281-282).

Com isso, tem-se a impressão de que existem mais reflexões paradigmáticas em um campo do conhecimento do que em outros. Porém, segundo Wersig (1993), parece que algumas discussões paradigmáticas não demonstram serem causadas por mudanças paradigmáticas dramáticas nem por sérias competições entre paradigmas alternativos. Isso quer dizer que, baseado em Araújo (2013), apesar de integrar as atividades de arquivos, biblioteca, museus e outras instituições que trabalham com a informação registrada, a "documentação" de Otlet e La Fontaine acabou por se desenvolver menos como uma nova ciência e mais como uma atividade profissional que atuava no campo da informação científica e tecnológica por meio dos centros de documentação.

Já se percebe uma tendência interdisciplinar para a CI devido às relações que constrói com várias outras disciplinas (SARACEVIC, 1995). As novas abordagens que se formam desses estudos interdisciplinares não abandonam, necessariamente, as propostas de Otlet e La Fontaine e se integram ao posterior pensamento de Vannevar Bush, anunciado em seu artigo "As we may think" ("Como nós pensamos"), de 1945, de criar um planetário universal que recupere qualquer conhecimento produzido pelo homem em qualquer tempo e lugar.

> Um segundo marco da história da Ciência da Informação é o desenvolvimento da área de Recuperação da Informação. Sua origem remonta às décadas de 1930 e 1940, quando começaram a ser utilizados os microfilmes como alternativa de guarda e disponibilização de acervos documentais. Tal fato despertou alguns teóricos a refletir sobre a distinção entre os suportes físicos

> do conhecimento e seu conteúdo, na medida em que se permitia a transposição do conteúdo para outros suportes. Tal percepção se aprofundou com o progressivo desenvolvimento das tecnologias computacionais, e teve uma inspiração teórica no artigo *As we may think*, de Vannevar Bush, publicado em 1945, onde é apresentado a ideia de uma possibilidade de recuperação automatizada da informação (ARAÚJO, 2013, p. 282).

Wersing (1993) afirma que existem tradições dentro da CI que não se ajustam à área da biblioteca ou da recuperação da informação, por exemplo, os estudos relativos à citação, estudos do fluxo da informação, estudos das consequências sociais das tecnologias da informação e estudos sobre a produção do conhecimento. Mesmo assim, segundo Araújo (2013), a área da recuperação da informação surgiu em 1951, quando Calvin Mooers criou essa expressão, com as pesquisas voltadas para os sistemas de recuperação da informação impulsionados pelo Second International Congress of Information System Sciences, evento ocorrido em 1962, no Georgia Institute of Technology, da Virgínia no Estados Unidos.

> Um terceiro fato importante foi a atuação de alguns cientistas, entre os anos de 1920 e 1940, primeiro na Inglaterra, depois nos Estados Unidos e na União Soviética, que passaram não mais a se dedicar aos assuntos específicos de suas ciências (a química, a física, entre outras), mas ao trabalho de coleta, seleção, produção de resumos e disseminação da produção científica para os demais cientistas de seus respectivos campos (ARAÚJO, 2013, p. 283).

Talvez isso sinalizasse uma tendência para pesquisas que tivessem a liberdade de passear por onde fosse conveniente na busca por produzir conhecimento, algo bastante conexo com um olhar mais voltado para o fluxo da informação do que para a posse dos acervos. Mesmo que a "documentação" fosse tida como sendo um modo não muito eficiente de lidar com algo que era sentido como um novo problema (a informação), ela auxiliou a CI a percorrer o seu caminho ao longo do século XX, começando justamente por estes cientistas que passaram a ser conhecidos como "cientistas da informação".

Tudo o que Araújo (2013) discorre sobre a CI evidencia que, através de muitos séculos, o papel do conhecimento para indivíduos, organizações e sociedades mudou de várias formas e, prevista por Wersig (1993), essas mudanças se tornaram mais aparentes no começo do século XX e, aproximadamente desde a década de 1960, estão se tornando parte de uma ampla mudança característica da pós-modernidade. Isso fica claro quando Borko, em 1968, sacramenta a CI como uma ciência interdisciplinar derivada e relacionada com áreas como a matemática, lógica, linguística, psicologia, informática, pesquisa de operações, artes gráficas, comunicação, biblioteconomia, administração e outras áreas afins e também quando Saracevic (1995), um pouco depois, diz ser a CI uma participante ativa na evolução da chamada sociedade da informação.

Isso é confirmado por Saracevic (1995) quando diz que a CI é um campo dedicado à investigação científica e à prática profissional de resolver os problemas da comunicação efetiva dos registros do conhecimento e do conhecimento entre os seres humanos nos âmbitos sociais, institucionais e/ou individuais e das necessidades de informação, sendo inexoravelmente ligada à tecnologia da informação. Principalmente quando o autor salienta que as investigações em CI devem levar em consideração os seres humanos nos âmbitos sociais, institucionais e/ou individuais, ele insere uma dimensão social para as pesquisas da área.

> É possível dizer, assim, que a história da Ciência da Informação ao longo das décadas seguintes à sua estruturação foi o de, progressivamente, tentar incorporar à sua agenda de pesquisa as dimensões semântica e pragmática inicialmente 'expulsas' do campo com a adoção integral da Teoria Matemática para a definição do conceito de informação (ARAÚJO, 2013, p. 285).

A informação no contexto da produção do conhecimento gera mais informação em um ciclo interminável, para alguns até incontrolável. Como medir uma produção informacional cada vez mais crescente? Não importa em que área esse conhecimento é produzido e nem onde as informações oriundas desse conhecimento vão chegar, o fato é que são crescentes e necessitam serem organizadas para serem recuperadas. Para dar conta disso, uma das fronteiras da CI está nas metrias da informação preocupadas com métodos estatísticos e matemáticos, em dar racionalidade para a universalidade e velocidade do conhecimento e isso se inicia com a teoria da informação.

> O último ingrediente para a [...] construção de uma Ciência da Informação veio do livro *Teoria Matemática da Comunicação*, publicado em 1949 por Shannon e Weaver, dois engenheiros de telecomunicações dos laboratórios da Bell System, dos Estados Unidos, diretamente envolvidos com os esforços de inteligência de guerra na época da Guerra Fria. Essa teoria é normalmente conhecida como 'teoria da informação' e tal denominação não se deu sem motivos: trata-se da teoria que, pela primeira vez, enunciou um conceito científico de 'informação'. Os autores estão preocupados com a eficácia do processo de comunicação e, para tanto, elegem como conceito central de seu trabalho a noção de informação (ARAÚJO, 2013, p. 283-284).

Teoria e prática devem estar ligadas, visto que cada uma se alimenta da trabalho da outra, a ponto de Borko (1968) afirmar que na CI há espaço para ambas porque são necessárias. A teoria da informação tem importância histórica para a CI, mas ela se encaixa junto a outros conceitos formulados nos anos de 1960 que descrevem a CI como meramente voltada para os estudos das "propriedades objetivas" da informação. Araújo (2013) acrescenta que aí se inclui a definição publicada por Borko em 1968, em seu importante artigo para a CI denominado "Information Science: what is it".

> Shannon e Weaver (1975) apontaram que as questões relativas à comunicação envolvem três níveis de problemas. O primeiro se refere aos problemas técnicos, relativos ao transporte físico da materialidade que compõe a informação (como, por exemplo, o volume do som numa conversa ou a qualidade da impressão em um papel). O segundo nível se refere aos problemas semânticos, isto é, se relaciona com a atribuição de significado. Enquanto que o primeiro nível envolve uma operação mecânica (reconhecer as letras num papel, captar os sons de uma fala), o segundo se relaciona a uma operação mental especifica, a de depreender, de uma determinada materialidade (sonora, visual, etc), um sentido, que pode se dar de maneira conotativa ou denotativa, literal ou irônica, metafórica, etc. O terceiro nível é o pragmático, relaciona-se com a eficácia. Quem emite informações a outrem deseja, de alguma forma, provocar um comportamento, causar alguma reação (convencer alguém a comprar um produto, eleger um candidato, pedir um favor, etc) (ARAÚJO, 2013, p. 284).

Então, segundo alguns autores, as metrias oriundas da biblioteconomia eram limitadas porque essa área tem por objeto o suporte da informação, já a CI, com seus diálogos interdisciplinares, proporciona ambientes de desenvolvimento de metrias mais úteis. Talvez uma das diferenças entre as antigas e novas metrias da informação esteja, por exemplo, na própria diferença entre a biblioteconomia e a CI. A primeira se consolidou criando todo um cenário que privilegiava as bibliotecas com a visão custodial, mais voltada para a posse das coleções do que para o fluxo da informação, envolvendo medições de operações mecânicas (contar as coleções mais usadas, títulos mais consultados, número de leitores etc.). A segunda se apropria da informação como seu objeto em uma visão pós-custodial mais voltada para o fluxo e difusão da informação, envolvendo estudos e métricas de operações mentais e materialidades da informação.

> Em termos genéricos, estas são algumas possibilidades de aplicação das técnicas bibliométricas, cienciométricas e informétricas: identificar as tendências e o crescimento do conhecimento em uma área; identificar as revistas do núcleo de uma disciplina; mensurar a cobertura das revistas secundárias; identificar os usuários de uma disciplina; prever as tendências de publicação; estudar a dispersão e a obsolescência da literatura científica; prever a produtividade de autores individuais, organizações e países; medir o grau e padrões de colaboração entre autores; analisar os processos de citação e co-citação; determinar o desempenho dos sistemas de recuperação da informação; avaliar os aspectos estatísticos da linguagem, das palavras e das frases; avaliar a circulação e uso de documentos em um centro de documentação; medir o crescimento de determinadas áreas e o surgimento de novos temas (VANTI, 2002, p. 155).

De posse dos fundamentos e ambientes propulsores das metrias da informação se fez necessário aprofundar os aspectos históricos e as possibilidades de aplicação das técnicas bibliométricas, cienciométricas e informétricas. As próximas subseções foram elaboradas para percorrer os principais conceitos envolvendo as metrias da informação para, assim, identificar

aproximações e distâncias, no momento da análise dos dados, de aplicações das técnicas econométricas e do MQO.

O objetivo desta seção é discorrer sobre o percurso histórico e sobre a aplicação das metrias que a CI vem construindo, desde a mais primordial de todas, a bibliometria oriunda da biblioteconomia, passando pelas visões modernas como a cienciometria, até a informetria, que é, dentre as metrias, a que mais se aproxima desta tese.

2.1 BIBLIOMETRIA: LEIS ZIPF, LOTKA E BRADFORD

A bibliometria se constitui em um conjunto coordenado de medidas relativas ao livro e ao documento. Essa é a ideia com a qual Paul Otlet, em 1934, inicia sua explicação do que é a bibliometria em uma de suas principais obras, "Traitê de documentation", em um capítulo próprio intitulado "Le livre et Ia mesure Bibliométrie".

> [Modernamente] a Bibliometria é uma ferramenta estatística que permite mapear e gerar diferentes indicadores de tratamento e gestão da informação e do conhecimento, especialmente em sistemas de informação e de comunicação científicos e tecnológicos, e de produtividade, necessários ao planejamento, avaliação e gestão da ciência e da tecnologia, de uma determinada comunidade científica ou país. [...] é também um instrumento quantitativo, que permite minimizar a subjetividade inerente à indexação e recuperação das informações, produzindo conhecimento, em determinada área de assunto. Em última análise ela contribui para tomadas de decisão na gestão da informação e do conhecimento, uma vez que auxilia na organização e sistematização de informações científicas e tecnológicas (GUEDES; BORSCHIVER, 2005, p. 15).

Otlet pode ser considerado o pai do termo "bibliometria", como também o primeiro a lançar as bases para o surgimento dessa área como um campo de conhecimento interdisciplinar, que vai muito além de "medir" o livro. Apesar do radical "biblio", o termo já surge com nítida preocupação com a produção do conhecimento que vai além do suporte. Isso quer dizer que desde a sua origem, a bibliometria já se inclinava para as medições das ciências, o que coincide com o dito por Guedes e Borschiver (2005).

A bibliometria, ao longo do tempo, construiu fundamentos e leis próprias a partir de estruturas das ciências aplicadas, em especial da matemática e da estatística. Para entender melhor os estudos de aplicação matemática e estatística para medir a produção do conhecimento, se faz necessário apresentar aqui um panorama histórico do desenvolvimento dessa disciplina.

Em 1896 surge a primeira lei relacionada com a distribuição de renda, a regra que 80% dos efeitos (consequências) são provenientes de 20% das causas (80/20), apresentada pelo economista italiano Vilfredo Pareto (BUNKLEY, 2008). Em 1917, a publicação de *The History of Comparative Anatomy, Parte I: A Statistical Analysis of the Literature* marca um estudo bibliométrico pioneiro dos ingleses F. J. Cole e Nellie Eales, que analisaram estatisticamente uma bibliografia de anatomia comparada (FONSECA, 1986).

> O termo statistical bibliography – hoje Bibliometria – foi usado pela primeira vez em 1922 por E. Wyndham Hulme, antecedendo à data a qual se atribui a formação da área de Ciência da Informação, com a conotação de esclarecimento dos processos científicos e tecnológicos, por meio da contagem de documentos (GUEDES; BORSCHIVER, 2005, p. 2).

Ainda em 1922, Dresden apresentou uma análise dos trabalhos apresentados nas reuniões da Seção de Chicago (matemática), fazendo uma relação entre números de autores e trabalhos. O objetivo principal do artigo foi apontar qual a contribuição de cada autor para uma das cinco áreas de classificação enciclopédica e a relação delas com o valor total de artigos apresentados (DRESDEN, 1922).

Em 1926, Lotka disse que seria interessante determinar, se possível, a parte de cada contribuição dos pesquisadores de diferentes áreas para o progresso da ciência. E essa ideia, presente em seu artigo, seria posteriormente conhecida como a Lei de Lotka. Em 1927, surge a aplicação da análise de citações, que tem a função de promover uma relação entre dois documentos. Tal técnica, baseada na contagem de referências, é utilizada nesse ano pela primeira vez por P. Gross e E. Gross, depois por Allan, em 1929, e por Gross e Woodford, em 1931 (BORGES, 2008).

A década de 1930 foi marcada pelos estudos relacionados ao princípio da frequência relativa, pelo surgimento do termo "bibliométria" através do *Traité de Documentation* de Paul Otlet e pela publicação de *The psychobiology of language*, com a primeira formulação clara da "Lei de Zipf" pelo próprio Zipf.

Os anos de 1948 e 1949 sofrem o impacto das publicações das características da literatura de química e física com o uso de *key journals* por Herman Fussler, da "documentação" de Bradford, da Teoria Matemática da Comunicação, de Claude Shannon, da proposta do termo "librametrics" por Ranganathan e das ideias de Zipf sobre o comportamento humano e o princípio do menor esforço (ROUSSEAU R., 2014).

Nas décadas de 1950 e 1960, Borges (2008) e Rousseau R. (2014) destacam o surgimento de diversos outros estudos como a *Estrutura formelle des textos et comunicação*, de Mandelbrot, o "fator de impacto", de Garfield e Sher, o "efeito Matthew", de Merton, bem

como estudos aplicados, como o "índices de citação para a ciência", de Garfield e a "naukometria" (cientometria), de Nalimov. Em 1969, Alan Pritchard publica um artigo intitulado "Statistical Bibliography or Bibliometrics?" e Tague-Sutcliffe (1992) se refere a esse artigo como o primeiro a usar o termo, quando definiu a disciplina bibliometria como o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada, que desenvolve padrões e modelos matemáticos para medir esses processos, usando seus resultados para elaborar previsões e apoiar tomadas de decisão.

> O presente artigo nunca encontrou o termo *statistical bibliography*, e, a julgar por discussões com muitos outros pesquisadores da área, esta visão é bastante geral. [...] Portanto, sugere-se que um nome melhor para este assunto (como definido anteriormente) é *BIBLIOMETRICS*, ou seja, a aplicação de métodos matemáticos e estatísticos para livros e outros meios de comunicação. Uma pesquisa intensiva da literatura não conseguiu revelar qualquer uso anterior deste termo e uma abordagem para o OED novamente não conseguiu descobrir que o termo foi usado antes (PRITCHARD, 1969)[2].

Fonseca (1986), porém, já tratava dessa polêmica quando explicou que, embora tenha sido positivo para popularizar a disciplina, Otlet já tinha cunhado o termo "bibliometria" quase quatro décadas antes de Pritchard ao assumir a paternidade da terminologia, e que isso era confirmado por muitos outros autores.

Mesmo assim, o texto de Pritchard serviu para, a partir de 1969, consolidar a área de estudos que usa métodos matemáticos e estatísticos para investigar e quantificar os processos de comunicação escrita. Pritchard ajudou a notabilizar a literatura como o ingrediente-chave no processo de comunicação do conhecimento, além de mostrar que o atributo de uma unidade de literatura na forma publicada (artigos de periódicos e livros) pode ser estudado em termos estatísticos.

A partir de 1998, Macias-Chapula (1998) afirma que, apesar de a área da bibliometria ser mais antiga que a informetria, os cientistas da informação da Europa e dos Estados Unidos a têm incluído como subárea da informetria, principalmente porque algumas áreas de atuação são bastante comuns à bibliometria e à informetria e, como já definido por Tague-Sutcliffe (1992), possuem:

> a) características da relação autor-produtividade medidas por meio do número de artigos ou outros meios; grau de colaboração;

---

[2] The present writer has never found the terms *statistical bibliography* at all satisfactory, and, to judge from discussions with many other workers in the field, this feeling is fairly general. […] Therefore it is suggested that a better name for this subject (as previously defined) is BIBLIOMETRICS, i.e. the application of mathematics and statistical methods to books and other media of communication. An intensive search of the literature has failed to reveal any previous use of this term and an approach to the OED again failed to find that the term has been used before (PRITCHARD, 1969).

> b) características das publicações, sobretudo a distribuição em revistas de artigos relativos a uma disciplina;
> c) análise de citação: distribuição entre autores, artigos, instituições, revistas, países; uso em avaliação; mapa de disciplinas baseado na cocitação;
> d) uso da informação registrada: circulação em bibliotecas e uso de livros e revistas da própria instituição; uso de bases de dados;
> e) obsolescência da literatura, avaliada pelo uso e pela citação;
> f) crescimento de literaturas especializadas, bases de dados, bibliotecas; crescimento simultâneo de novos conceitos;
> g) tipos e características dos níveis de desempenho da recuperação.
> h) aspectos estatísticos da linguagem e frequência de citação de frases, tanto em textos (linguagem natural), como em índices impressos e em formato eletrônico;
> i) definição e medida da informação.

Dentre todas as áreas especificadas, as letras "a", "b", "c", "d" e "e" são as que mais aproximam a bibliometria das ciências exatas pelo fato de que seus princípios convergem para quantificar a produção de tipos documentais (livro, artigos, periódicos, teses, dissertações etc.) de forma empírica, o que no início era centrado na medição do livro como principal suporte da informação. Em geral, é possível afirmar que os seus métodos empíricos foram construídos através de tentativas e erros, baseados nas observações e experiências acumuladas ao longo do tempo. É dessa dinâmica que surge as bases das pesquisas bibliométricas que, em seus estudos de caso, relacionam fortemente seus resultados à aplicação de técnicas estatísticas.

> Verifica-se [por exemplo] na Lei de Bradford, que permite estimar o grau de relevância de periódicos em dada área do conhecimento, que os periódicos que produzem o maior número de artigos sobre dado assunto formam um núcleo de periódicos, supostamente de maior qualidade ou relevância para aquela área. A Lei de Lotka considera que alguns pesquisadores, supostamente de maior prestígio em uma determinada área do conhecimento, produzem muito e muitos pesquisadores, supostamente de menor prestígio, produzem pouco. Nas Leis de Zipf, que permitem estimar as frequências de ocorrência das palavras de um determinado texto científico e tecnológico e a região de concentração de termos de indexação, ou palavras-chave, que um pequeno grupo de palavras ocorre muitas vezes e um grande número de palavras é de pequena frequência de ocorrência (GUEDES; BORSCHIVER, 2005, p. 4).

A história da bibliometria é fascinante e vasta, por isso é que inúmeros eventos não constam neste breve levantamento. Mas é importante apresentar, aqui, os principais destaques que representam tecnicamente essa área do conhecimento: Leis de Zipf, Lei de Lotka e Lei de Bradford.

Em 1913, Auerbach encontrou uma relação hiperbólica entre a classificação e o tamanho das cidades alemãs, o que hoje se chama "Lei de Zipf". Em 1916, o uso do termo "hyperbolicnature" ("natureza hiperbólica" ou "Lei de Zipf") foi apresentado por Jean-Baptiste

Estoup, que foi secretário-geral do Institut Sténographique de France e pioneiro na investigação da regularidade que mais tarde ficou conhecida como a Lei de Zipf.

> Zipf observou que, num texto suficientemente longo, existia uma relação entre a frequência que uma dada palavra ocorria e sua posição (localização em uma lista, confeccionada segundo a ordem decrescente de frequência de ocorrência das palavras, que compõem um determinado texto) na lista de palavras ordenadas segundo sua frequência de ocorrência. Essa lista era confeccionada, levando-se em conta a frequência decrescente de ocorrências. À posição nesta lista dá-se o nome de ordem de série (rank). Assim, a palavra de maior frequência de ocorrência tem ordem de série 1, a de segunda maior frequência de ocorrência, ordem de serie 2 e, assim, sucessivamente (GUEDES; BORSCHIVER, 2005, p. 6).

Na verdade, segundo os autores, deve-se falar em Leis de Zipf porque a lei mais conhecida se relaciona à frequência de ocorrência de palavras em um dado texto, porém esta foi enriquecida pelo Ponto de Transição (T) de Goffman, que se relaciona diretamente com a representação da informação, isto é, a indexação temática automática. O cálculo de tal ponto leva em consideração o que foi proposto por Hans Peter Luhn, em 1957: que uma análise da indexação poderia se basear em uma amostra representativa de documentos sobre determinado assunto.

> Zipf observou, também, que o produto da ordem de série (r) de uma palavra, pela sua frequência de ocorrência (f) era aproximadamente constante (c). Enunciou assim que **r · f = c**, o que ficou conhecido como Primeira Lei de Zipf. Fairthone (1969) ressaltou que Zipf considera essa relação observada como uma consequência do Princípio Geral do Menor Esforço. Pao (1978) reconhece que esta lei é elegante em sua simplicidade. Entretanto, ela se aplica somente a palavras de alta frequência de ocorrência, em um texto. Para palavras de baixa frequência de ocorrência, Zipf propôs uma segunda lei, revisada e modificada por Booth (1967) (GUEDES; BORSCHIVER, 2005, p. 6).

Enquanto a primeira Lei de Zipf postula que $c = r \cdot f$, sendo (r) o produto da ordem de série de uma palavra e (f) a sua frequência de ocorrência, a segunda Lei de Zipf diz que em um determinado texto várias palavras de baixa frequência de ocorrência (alta ordem de série) têm a mesma frequência, e Booth, em 1967, a representou matematicamente da seguinte forma (I_1 é o número de palavras que têm frequência 1, I_n é o número de palavras que têm frequência n, e 2 é a constante válida para a língua inglesa):

Equação 1 – Segunda Lei de Zipt

$$\frac{I_1}{I_n} = \frac{n(n+1)}{2}$$

Fonte: Guedes e Borschiver (2005)

Segundo Guedes e Borschiver (2005), parte da literatura tem se referido a essa fórmula como "Segunda Lei de Zipf-Booth". Os autores afirmam que Guedes e Valois, em 1988, perceberam que o segundo membro da fórmula, n(n+1)/2, corresponde à soma dos termos de uma progressão aritmética utilizada por Carl Friedrich Gauss (1777-1855), quando calculou a soma aritmética dos termos de 1 a 100.

Um exemplo pode ser dado: suponha que se queira somar 1 + 2 + 3 + 4 + 5. Sabemos que para esse n = 5, o resultado é 15, assim, é fácil calcular a soma dos cinco primeiros termos de uma Progressão Aritmética (PA) de razão um. Agora imagine calcular os 100 primeiros termos de uma PA de razão 1? Com certeza seria muito trabalhoso e difícil. Porém, com a fórmula n(n+1)/2, fica fácil e rápido. Para a soma de 1 a 5, fica da seguinte forma: 5(5+1)/2=5x6/2=30/2=15. E para a soma de 1 a 100, feita por Gauss, fica da seguinte forma: 100(100+1)/2=100x101/2=10100/2=5050.

Com isso, pode-se afirmar que a primeira Lei de Zipf se comporta como uma função linear e que o segundo membro da fórmula da segunda lei, n(n+1)/2, corresponde à soma dos termos de uma PA de razão 1. Outras observações que podem ser acrescidas são: a) Goffman observou que a Primeira Lei de Zipf tinha validade apenas para a região de palavras de alta frequência de ocorrência; b) na maioria das vezes, afirma Guedes e Borschiver, as palavras de alta frequência têm a propriedade de ocupar *ranking* (r) único na lista de distribuição de palavras, isto é, dentre as palavras de alta frequência de ocorrência, dificilmente existem duas palavras com a mesma frequência de ocorrência; c) a denominada "Lei de Zipf-Booth" ou "Segunda Lei de Zipf" descreve o comportamento das palavras de baixa frequência de ocorrência e pode ainda ocorrer que existam muitas palavras com a mesma frequência (GUEDES; BORSCHIVER, 2005).

As observações "b" e "c", para os autores, definem duas extremidades da lista de distribuição de palavras de um texto, formando uma região crítica na qual há a transição do comportamento das palavras de alta frequência para as de baixa frequência, a ponto de Goffman admitir a hipótese de que nessa região de transição estariam as palavras de maior conteúdo semântico do texto analisado.

Continuado os estudos bibliométricos, é formulada, em 1926, a Lei de Lotka, surgindo de um estudo sobre a produtividade de cientistas, a partir da contagem de autores presentes no *Chemical Abstracts*, entre 1909 e 1916 (ARAÚJO, 2006). Essa lei explica as diferenças de produtividade de cientistas em um determinado campo, demonstrando que a relação entre autor e itens é inversamente proporcional ao quadrado do número de artigos publicados. Isso quer dizer, segundo Araújo (2006), que Lotka descobriu que uma larga proporção da literatura científica é produzida por um pequeno número de autores, e um grande número de pequenos produtores se iguala, em produção, ao reduzido número de grandes produtores.

> A Lei de Lotka, relacionada à produtividade de autores e fundamentada na premissa básica de que 'alguns pesquisadores publicam muito e muitos publicam pouco' (VOOS, 1974), enuncia que 'a relação entre o número de autores' e o número de artigos publicados por esses, em qualquer área científica, segue a Lei do Inverso do Quadrado $1/n^2$. Isto é, em um dado período de tempo, analisando um número n de artigos, o número de cientistas que escrevem dois artigos seria igual a ¼ do número de cientistas que escreveram um. O número de cientistas que escreveram três artigos seria igual a 1/9 do número de cientistas que escreveram um, e assim sucessivamente (GUEDES; BORSCHIVER, 2005, p. 4).

Araújo (2006) acrescenta que a Lei dos Quadrados Inversos pode assumir a seguinte forma: $yx = 6/p^2x^a$, onde yx é a frequência de autores publicando número "x" de trabalhos e "a" é um valor constante para cada campo científico como, por exemplo, a = 2, para físicos, ou a = 1,89, para químicos. Mas essa relação, inversamente proporcional entre o número de autores e o número de artigos publicados por estes em qualquer área científica, pode também apresentar uma curva cujo comportamento seja o Inverso do Cubo ($1/n^3$).

Guedes e Borschiver (2005) já afirmavam que Price, em 1965, observou que, para as ciências em geral, o número de autores decresce mais rapidamente que o Inverso do Quadrado, o que estaria mais próximo da Lei do Inverso do Cubo ($1/n^3$) do que da Lei do Inverso do Quadrado ($1/n^2$). Os autores explicam que essa observação de Price foi apresentada em um Seminário de Estudos em Ciência da Informação, em 1963, denominado "Little Science, Big Science", onde, pela primeira vez, Price apresentou dados estatísticos sobre o fenômeno do crescimento exponencial da literatura.

> Desde 1926, época em que Lotka estabeleceu esta lei, muitos estudos têm sido conduzidos para investigar a produtividade dos autores em distintas disciplinas. Até dezembro de 2000, mais de 200 trabalhos, entre artigos, monografias, capítulos de livros, comunicações a congressos e literatura gris (cinzenta) tinham sido produzidos tentando criticar, replicar e/ou reformular esta lei bibliométrica (URBIZAGÁSTEGUI ALVARADO, 2002, p. 14).

Em 2010, Urbizagástegui Alvarado observou, em seu estudo sobre a produção de artigos sobre a Lei de Lotka, que pesquisadores familiarizados com aplicações matemáticas tendem a exercer domínio sobre o campo de estudos métricos, o que poderia ensejar um comportamento mais inversamente cúbico do que inversamente quadrado da curva.

> Na gestão da informação, do conhecimento e planejamento científico e tecnológico, sua aplicabilidade [Lei de Lotka] se verifica na avaliação da produtividade de pesquisadores, na identificação dos centros de pesquisa mais desenvolvidos, em dada área de assunto, e no reconhecimento da "solidez" de uma área científica. Ou seja, quanto mais solidificada estiver uma ciência, maior probabilidade de seus autores produzirem múltiplos artigos, em dado período de tempo (GUEDES; BORSCHIVER, 2005, p. 5).

Um exemplo pode ser dado através daqueles autores com formação matemática ou estatística, ou através daqueles que defenderam uma tese sobre bibliometria ou sobre a Lei de Lotka. Estes terão maiores possibilidades de publicar mais artigos sobre o assunto "Lei de Lotka". Mesmo assim, nos campos do conhecimento, a medida, ou medição, é uma forma objetiva, inteligível e palpável de construir o núcleo de conhecimento como, por exemplo, quando periódicos científicos apresentam as produtividades de artigos sobre determinado assunto, ou ainda quando, através destes, os assuntos permitem identificar grupos ou zonas de produção de pesquisas.

Por fim, surge a Lei de Bradford ou Lei da Dispersão dos periódicos, criada em 1934 por Samuel C. Bradford, que mensura a incidência de um determinado termo ou assunto de uma área especifica a verificar, ou seja, o grau de atração de periódicos sobre determinada temática (ARAÚJO, 2006).

> A Lei de Bradford, relacionada à dispersão da literatura periódica científica, enuncia que 'se periódicos científicos forem ordenados em ordem decrescente de produtividade de artigos sobre determinado assunto, poderão ser divididos em um núcleo de periódicos mais particularmente dedicados ao assunto e em vários grupos ou zonas, contendo o mesmo número de artigos que o núcleo. O número de periódicos (n), no núcleo e zonas subsequentes, variará na proporção 1:n:n2 [...]' (BROOKES, 1969) (GUEDES; BORSCHIVER, 2005, p. 4).

Nesse sentido, a Lei de Bradford, dentro da bibliometria, constitui um conjunto coordenado de medidas relativas a periódicos, artigos, assuntos, núcleos e grupos de produção do saber. Se for analisado sob um ponto de vista interdisciplinar, Bradford já trabalhava na perspectiva da CI, que tem a interdisciplinaridade como sua natureza e se coaduna, segundo Saracevic (1995), com as relações com várias disciplinas que estão mudando a evolução interdisciplinar da CI. Com esse entendimento, a CI nasce com a concepção de contribuir para que as pesquisas bibliométricas tenham a liberdade de passear por onde for conveniente na

busca de produzir conhecimento. A Lei de Bradford e a CI, surgida no início do século XX, originam-se com o advento da "documentação" como um modo prático de lidar com algo que era sentido como um problema: a transposição da medição do suporte (periódicos, artigos e assuntos) para a medição da informação (núcleos de conhecimento, zonas e grupos de pesquisas).

> A Lei de Bradford sugere que na medida em que os primeiros artigos sobre um novo assunto são escritos, eles são submetidos a uma pequena seleção, por periódicos apropriados, e se aceitos, esses periódicos atraem mais e mais artigos, no decorrer do desenvolvimento da área de assunto. Ao mesmo tempo, outros periódicos publicam seus primeiros artigos sobre o assunto. Se o assunto continua a se desenvolver, emerge eventualmente um núcleo de periódicos, que corresponde aos periódicos mais produtivos em termos de artigos, sobre o tal assunto. Brookes (apud BROOKES, 1969), refere-se a esse fenômeno como o 'mecanismo do sucesso gerando o sucesso' (GUEDES; BORSCHIVER, 2005, p. 4).

Alguns questionamentos podem ser levantados quando se leva em conta as ideias de Bradford. Se essa lei está ajustada para a estrutura de bibliotecas ou para a recuperação da informação em periódicos, existem enfoques, segundo Wersig (1993), trazidos pela CI que não se ajustam a essas estruturas e que necessitam de novos usos para a Lei de Bradford, como os estudos relativos à citação, estudos de fluxo da informação, estudos das consequências sociais das tecnologias da informação e estudos sobre a produção do conhecimento.

> A Lei de Bradford é um instrumento útil para o desenvolvimento de políticas de aquisição e de descarte de periódicos, em nível de gestão de sistemas de recuperação da informação, gestão da informação e do conhecimento científico e tecnológico. É possível estimar a magnitude de determinada área bibliográfica e o custo de toda e qualquer fração específica da bibliografia, no todo (GUEDES; BORSCHIVER, 2005, p. 4).

Esses novos usos para a Lei de Bradford, expostos por Guedes e Borschiver (2005), coadunam com Borko (1968), quando este afirmava que dentro da CI há espaço tanto para o teórico quanto para o prático, porque teoria e prática estão inexoravelmente ligadas e se nutrem mutuamente. O fato de o conhecimento ser universal e permear todas as outras ciências faz com que as medições de periódicos e assuntos (Lei de Bradford em sentido estrito) tenham as suas fronteiras alargadas para a atuação no âmbito de uma ciência da informação (políticas de aquisição e de descarte de periódicos, gestão de sistemas de recuperação da informação, gestão da informação e do conhecimento científico e tecnológico, custos da bibliografia), que é um uso da lei em sentido ampliado.

> [Nesse sentido] Construtos teóricos, medidas de informação tais como precisão, revocação e relevância têm sido formuladas e reformuladas, mas faltam estudos empíricos, ou melhor, são inexpressivos em número, o que

> talvez explique uma certa estagnação da bibliometria durante algum tempo e seu retorno nos dias atuais, pelas facilidades tecnológicas para sua aplicação. [...] Estudos empíricos permitiriam testar hipóteses e métodos, não para buscar verdades e certezas matemáticas [...], mas identificar tendências e regularidades (PINHEIRO, 2005).

Baseado nesses autores pode-se afirmar que o foco da Lei de Bradford migra do suporte (livro, periódico, artigo) para o conteúdo (conhecimento), muito influenciado pela introdução de um novo conceito de informação, um assunto complexo e multidimensional, embrionário de uma nova ciência, a CI, justamente porque, como afirma Borko (1968), tem tanto um componente de ciência pura que investiga o assunto sem considerar a sua aplicação (Lei de Bradford enquanto estudos empíricos que buscam verdades e certezas matemáticas), quanto um componente de ciência aplicada que desenvolve serviços e produtos (Lei de Bradford enquanto estudos que identifiquem tendências e regularidades na precisão, revocação e relevância das medidas de informação para gerar conhecimento).

2.2 CIENTOMETRIA: FRENTE DE PESQUISA, FATOR DE IMPACTO, ACOPLAMENTO BIBLIOGRÁFICO, OBSOLESCÊNCIA DA LITERATURA, LEI DO ELITISMO, TEORIA EPIDÊMICA DE GOFFMAN E LEI DOS 80/20

A década de 1970 foi muito produtiva. Várias pesquisas e acontecimentos podem ser destacados, tais como: Price (1965), com o "Sucesso alimenta sucesso", o *Journal Citation Reports*, a fundação da revista *Scientometrics* e o uso do termo "informetrics", por Nacke.

Nas décadas seguintes, Borges (2008) lista a contribuição teórica de Callon e Michel, em 1980, juntamente com Bruno Latour e John Law, da atual e inovadora Teoria Ator-Rede (ou Actor-Network Theory, ATN). Essa teoria consiste na progressiva constituição de uma rede nas quais atores humanos e não humanos assumem identidades de acordo com a sua estratégia de interação. Há também a Big Science de Martin e Irvine. E Hamade (fator político nos hábitos da informação) que comparou a comunicação universitária e o número de publicações dos cientistas sociais que viviam ou foram educados em países do leste europeu.

Em 1992, prossegue Borges, Nuria Amat traz a questão dos "colégios invisíveis", afirmando que uma das causas do desenvolvimento destes é a ínfima qualidade dos documentos que os cientistas recebem das bases de dados comerciais ou devido à escassa eficácia demonstrada pelos centros de informação para fornecer os documentos que os cientistas solicitam e que, por isso, preferiam comunicarem-se entre eles.

Em 1994, Le Coadic afirmava que o objeto da cientometria é medir as atividades de Pesquisa Científica e Tecnológica (PCT), mediante insumos e produtos (investimentos, recursos humanos, equipamentos, publicações). Nesse sentido, o autor considera ciência e tecnologia, em sentido amplo, como sendo as ciências da matéria, do homem, da vida e da sociedade, com a cientometria sendo a sua medida.

Tague-Sutcliffe (1992), citado por Macias-Chapula (1998), assim definiu essa disciplina: cienciometria é o estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. A cienciometria é um segmento da sociologia da ciência, sendo aplicada no desenvolvimento de políticas científicas. Envolve estudos quantitativos das atividades científicas, incluindo a publicação e, portanto, sobrepondo-se à bibliometria.

Para Spinak (1998), a cienciometria examina o desenvolvimento e as políticas científicas e essa análise quantitativa considera a ciência como uma disciplina ou atividade econômica. Por isso, a cientometria pode fazer comparações entre políticas de investigação entre países no que tange aos aspectos econômicos e sociais das ciências. Pode também a cienciometria ser aplicada a uma grande variedade de campos como, por exemplo, história da ciência, ciências sociais, documentação e biblioteconomia, política científica e indústria da informação.

McGrath (1989), citado por Macias-Chapula (1998), na International Conference on Bibliometrics, Scientometrics and Informetrics, ocorrida no ano 1989, na cidade Ontário, no Canadá, apresentou sua tipologia para a definição e classificação da cientometria, identificando seu objeto de estudo, variáveis, métodos e objetivos, que seriam: a) objeto de estudo: disciplinas, assuntos, áreas e campos; b) variáveis: fatores que diferenciam as subdisciplinas, revistas, autores, documentos, e como os cientistas se comunicam; c) métodos: análise de conjunto e de correspondência; d) objetivos: identificar domínios de interesse, local onde os assuntos estão concentrados, como e quando os cientistas se comunicam.

> [Isso porque a] Ciência é um processo social. As ações e o comportamento de cientistas dependem do contexto. Para compreender isso, precisamos conhecer os cenários e as personalidades, estejamos olhando quer para o modo como os cientistas fazem ciência, quer para as formas pelas quais os resultados de seus esforços profissionais são comunicados. Na verdade, a ciência necessita ser considerada como um amplo sistema social, no qual uma de suas funções é disseminar conhecimentos. Sua segunda função é assegurar a preservação de padrões, e a terceira é atribuir créditos e reconhecimento para aqueles cujos trabalhos têm contribuído para o desenvolvimento das ideias em diferentes campos (MACIAS-CHAPULA, 1998, p. 136).

Porém, existem outros enfoques que não se ajustam às unidades de medida dessas leis (relevância de periódicos, produção de pesquisadores e frequências de ocorrência das palavras),

como os estudos relativos à citação, estudos de fluxo da informação, estudos das consequências sociais das tecnologias da informação e, principalmente, os estudos sobre a produção do conhecimento. Guedes e Borschiver (2005, p. 11) observam que "os estudos baseados na Análise de Citações partem da hipótese de que citação é um indicador válido de influência de um determinado trabalho sobre outro(s), evidenciando conexões intelectuais".

Araújo (2006) destaca a análise de citações (conjunto de uma ou mais referências bibliográficas que, incluídas em uma publicação, evidenciam elos entre indivíduos, instituições e áreas de pesquisa) como a área mais importante da bibliometria porque evidencia o relacionamento entre publicações. Foresti (1989) já acrescentava que a análise de citação investiga as relações entre os documentos citantes e os documentos citados, considerados no todo ou em suas diversas partes: autor, título, origem geográfica, ano, idioma de publicação etc.

Para Macias-Chapula (1998), um dos motivos de ocorrer a citação é porque é o meio mais comum de atribuir créditos e reconhecimento na ciência. Foresti (1989) determinava também outros motivos: a) contribuir para o desenvolvimento da ciência; b) prover o necessário reconhecimento de um cientista por seus colegas; c) estabelecer os direitos de propriedade e prioridade da contribuição científica de um autor; d) constituir importantes fontes de informação; e) ajudar a julgar os hábitos de uso da informação; e f) mostrar a literatura que é indispensável para o trabalho dos cientistas.

Logo, a análise de citações permite estudos de um conjunto de eventos relacionados à produção científica e ao conhecimento daí decorrente. Com isso, destacamos as seguintes metrias da produção científica (cienciometria): frente de pesquisa, fator de imediatismo ou de impacto, acoplamento bibliográfico, obsolescência da literatura, Lei do Elitismo e Teoria Epidêmica de Goffman. Nesse sentido, Araújo (2006, p. 18) acrescenta que ainda são possíveis as seguintes medições: autores mais citados, autores mais produtivos, procedência geográfica e/ou institucional dos autores mais influentes em um determinado campo de pesquisa, tipo de documento mais utilizado, idade média da literatura utilizada, procedência geográfica e/ou institucional da bibliografia utilizada, periódicos mais citados e "core" de periódicos que compõem um campo.

A análise de citações permite identificar a frente de pesquisa de uma determinada área científica por meio da contagem de citações de autores que se citam na literatura recente.

> [A] frente de pesquisa [...] correlaciona os índices absolutos de citação obtidos por cada autor com a data dos trabalhos publicados por cada autor. Assim, para a contagem da frente de pesquisa, só são contabilizados os trabalhos mais recentes desse autor. O período de tempo considerado relevante para se determinar se o trabalho é recente ou não varia conforme o objetivo que se quer atingir. Por exemplo, pode-se definir que se quer identificar a frente de

> pesquisa para cinco anos. Neste caso, contabilizasse apenas as citações recebidas referentes a trabalhos publicados nos últimos cinco anos (ARAÚJO, 2006, p. 20).

Os estudos sobre frente de pesquisa revelam um estreito padrão de relações múltiplas na literatura sobre o assunto. Permite revelar também os colégios invisíveis na medida que identifica em um pequeno grupo de artigos entrelaçados o trabalho de pesquisadores que colaboram entre si e que constituem os "colégios invisíveis". Price (1965), nesse sentido, observou que um grupo pequeno de autores e de publicações exerce maior influência em uma dada área de assunto, se constituindo em líderes nessas áreas, nas quais seus trabalhos são mais citados. Uma constatação matemática é que, em alguns casos, as citações são distribuídas regularmente e uniformemente por toda a literatura passada, com frequência decrescente (GUEDES; BORSCHIVER, 2005).

Araújo (2006) acrescenta que entram para frente de pesquisa os autores que tiverem pelo menos cinco citações, pois a condição para um autor fazer parte da frente de pesquisa é receber pelo menos uma citação por ano e que, atualmente, no Brasil, são realizados estudos referentes à frente de pesquisa. Um exemplo são os trabalhos de Mostafa (2002), que consistem em estudos de citação com o objetivo de se identificar a frente de pesquisa e, com isso, traçar as tendências epistemológicas de um determinado campo de estudo (MOSTAFA, 2002; MOSTAFA; MÁXIMO, 2003).

Outra forma de medição da informação científica é o fator de imediatismo ou de impacto. Para Guedes e Borschiver (2005), a análise de citações é utilizada para estimar o fator de imediatismo de um artigo publicado através do estudo da concentração de citações a esse artigo em documentos publicados. Para Araújo (2006), é um conceito extremamente relevante na análise de citações.

> Formulado por Garfield para constar das análises realizadas pelo ISI (*Institute of Scientific Information*), esse conceito consiste em '[...] dividir o número total de citações obtidas por um periódico em um ano qualquer pelo número de artigos publicados naquele ano' (RODRIGUES, 1981, p. 10). A bibliometria, ao longo de sua evolução, contudo, acabou se apropriando do conceito para a análise de autores, correlacionando índices absolutos de citação à quantidade de trabalhos citados. Assim, o fator de impacto é a divisão do número de citações recebidas por um autor dividido pelo número de trabalhos de receberam pelo menos uma citação. Com esse índice, se quer identificar autores que, apesar de terem tido pouca produção, produziram um material muito significativo, isto é, que receberam muitas citações, em oposição a autores que podem ter tido muitas citações porque publicaram muitos trabalhos, mas cada um desses trabalhos isoladamente com pouca relevância no campo científico. O uso do fator de impacto para a avaliação da produção científica continua a ser atual, algumas vezes relacionado a outros como obsolescência e a idade das referências (MEADOWS, 1999, p. 85-99; STREHL, 2005) (ARAÚJO, 2006, p. 19).

Araújo (2006) lembra que para Price (1965), o período de cálculo do fator de impacto é maior que um ano e o período de contagem é de 15 anos. Porém, Jones (2003) expressa sua preocupação com o uso e abuso do fator de impacto na avaliação da importância e do prestígio de periódicos científicos e de cientistas, e defende que o fator de impacto de um determinado periódico deve ser calculado em um período de dois anos, ou seja, o cálculo deve ser a divisão do número de citações correntes a um dado artigo, publicado nos últimos dois anos, pelo total de artigos publicados, no mesmo período de tempo. O cálculo do fator se baseia na hipótese de que artigos de periódicos citados mais frequentemente em determinada área científica são mais relevantes que artigos menos citados, o que pode levar a uma maior compreensão do que a citação acarreta para a indústria da informação.

> [...] a comercialização de citações significa que aqueles que produzem as citações (autores), aqueles que as utilizam (outros cientistas), aqueles que as processam e comercializam (indústria da informação) e aqueles que as distribuem (bibliotecários e cientistas da informação) precisam ter uma compreensão segura do que a citação acarreta e significa (MACIAS-CHAPULA, 1998, p. 136).

O autor afirma que as práticas de citação são também influenciadas por fatores externos à ciência. Ou seja, muitos autores citam por um reflexo da moda sem se preocuparem com as implicações dessa prática, não tendo claro nas suas mentes porque citam de uma determinada maneira ou como as citações estão relacionadas com uma possível ideologia que esteja dominante em uma ciência.

Uma terceira forma de medição é através do acoplamento bibliográfico e cocitação. Em 1963, Kessler desenvolveu as teorias do acoplamento bibliográfico, que dizem respeito à força de associação entre dois ou mais documentos (FORESTI, 1989). Segundo Araújo (2006), o acoplamento bibliográfico é uma abordagem que vai além da contagem de palavras e frequência das publicações e de citações. Isso revela que Kessler se preocupou com a relação entre os artigos que geram e recebem citações, numa ordenação multidimensional.

> O Acoplamento Bibliográfico (retrospectivo) consiste na união de artigos, citando o(s) mesmo(s) documento(s), e a Co-citação (prospectivo), na união de artigos, citados pelos mesmos documentos. Na opinião de Marshakova (1981), o Acoplamento Bibliográfico mede o grau de ligação entre dois ou mais artigo, segundo o número de documentos idênticos citados por esses artigos, e Co-citação mede o grau de ligação de dois ou mais artigos, pelo número de documentos onde esses artigos são citados, simultaneamente (GUEDES; BORSCHIVER, 2005, p. 12).

O acoplamento bibliográfico traz questões como o grau em que as práticas da citação refletem os elementos normativos e de valor dos sistemas científicos ou, até mesmo, o lado social desses sistemas no sentido de serem baseados na compreensão dos objetivos do trabalho

de um cientista, na análise do conhecimento científico adquirido ou na abrangência do caráter de propriedade social da ciência enquanto produto de um processo histórico (MACIAS-CHAPULA, 1998). Tudo isso enfatiza a importância da objetividade nos resultados da atividade científica, algo que só é possível através de medições como, por exemplo, a obsolescência da literatura e vida-média.

A obsolescência da literatura consiste na análise do declínio do uso da literatura, no decorrer do tempo, e a vida-média é estimada a partir da razão de obsolescência e da razão de crescimento de um determinado corpo de literatura. Em termos de uso da literatura, a vida-média tem sido interpretada mediante o estudo do número de citações feitas a um determinado item. Segundo Pao (1989), na área de química, metade das referências citadas na literatura relaciona-se aos artigos com menos de oito anos de publicação, enquanto que, na matemática, a vida-média de uma unidade da literatura é estimada em cerca de 20 anos (GUEDES; BORSCHIVER, 2005).

Há ainda dois tipos de envelhecimento da literatura científica: clássico ou efêmero (vida média longa ou curta). Nos estudos realizados dentro dessa perspectiva, percebeu-se que existem disciplinas com forte componente de literatura clássica (matemática, geologia, botânica), enquanto outras são quase exclusivamente compostas por literatura efêmera (física, engenharia), sendo algumas de caráter intermediário (fisiologia, química). Stinton e Lancaster (1987) desenvolveram uma importante linha de estudos nessa área a partir dos conceitos de sincronia e diacronia (ARAÚJO, 2006, p. 20).

A contagem de citações se baseia, segundo Macias-Chapula (1998), na premissa de que uma citação é a expressão de uma relação entre dois documentos porque aquele que cita e aquele que é citado estabelecem uma ligação intelectual, de ideias, de pesquisa, entre pessoas que se esforçam para produzir conhecimento. Pode haver, então, indagações sobre por que o autor citou estes autores ao invés de aqueles, dando indícios de um elitismo científico. Para medi-lo se tem uma quarta forma de metria: a Lei do Elitismo.

> A Lei do Elitismo enuncia que toda população de tamanho N tem uma elite efetiva de tamanho $\sqrt{N}$ (PRICE, 1965). Price (1963), Crawford (1971), Crane (1972), Cole e Cole (1972) e Griffith e Mullins (1972) encontraram, em seus estudos, evidências de elites e elitismo na ciência (PAO, 1989) (GUEDES; BORSCHIVER, 2005, p. 12).

Os motivos que levam os autores a citar são os mais variados. Acredita-se que por trás de cada medida da informação aqui discorrida existe um desses motivos, mas o principal é que as menções em textos de informações extraídas de outras fontes (cocitações) revelam ligações intelectuais, de ideias e de pesquisas entre pessoas que produzem conhecimento (autores). Um

exemplo da existência disso é a proposição de uma Teoria Epidêmica da Informação Científica por parte de Goffman, em 1964.

A Teoria Epidêmica de Goffman determina que se as ideias difundidas, dentro de uma determinada população de cientistas, possuem propriedades epidemiológicas, essas podem ser investigadas como em um processo epidêmico (GUEDES; BORSCHIVER, 2005). Para Araújo (2006), é uma variação de enfoques bibliométricos que foi desenvolvida por Goffman e Newill, em 1967. Esses autores realizaram seu estudo por comparação do ciclo da esquistossomose e da informação, fazendo uma analogia entre os dois sistemas, em que quando o processo epidêmico na área de saúde identifica a infecção (por exemplo, a pessoa com uma doença), Goffman e Newill identificam o autor com uma ideia. A pessoa que pode contrair a doença é vista como o leitor que recebe a ideia e o material infectante que, no caso da saúde, são os germes da doença, mas na Teoria Epidêmica da Informação consiste nas próprias ideias contidas na literatura.

> Segundo esse modelo, as ideias científicas são materiais infecciosos, no curso de uma epidemia intelectual; transmitidas, por exemplo, por comunicações diretas, entre um conferencista e o público, ou através de conversações. Essas ideias podem também ser expostas por um autor, em artigos de periódicos, para um determinado público. A análise matemática de Goffman foi capaz de prognosticar as condições de controle da epidemia, a razão de crescimento e de declínio, de uma dada área do conhecimento, e permitiu definir as condições sob as quais a epidemia declinaria e se tornaria estável (PAO, 1989). Segundo Goffman (1966), sua teoria possibilita estimar os níveis de importância de linhas de pesquisa, em uma determinada área de assunto, e prognosticar o comportamento dessas linhas de pesquisas (GUEDES; BORSCHIVER, 2005, p. 13).

Além de estimar os níveis de importância de linhas de pesquisa em uma determinada área de assunto, ou prognosticar o comportamento dessas linhas de pesquisas, ou mesmo calcular indicadores válidos de influência de um determinado trabalho sobre outro, quais os motivos que levam os autores a citar? A resposta é dada por Weinstock que, em 1971, segundo Macias-Chapula (1998), identificou 15 motivos da citação. São eles:

> 1. prestar homenagem aos pioneiros; 2. dar crédito para trabalhos relacionados; 3. identificar metodologia, equipamento etc.; 4. oferecer leitura básica; 5. retificar o próprio trabalho; 6. retificar o trabalho de outros; 7. analisar trabalhos anteriores; 8. sustentar declarações; 9. informar aos pesquisadores de trabalhos futuros; 10. dar destaque a trabalhos pouco disseminados, inadequadamente indexados ou desconhecidos (não citados); 11. validar dados e categorias de constantes físicas e de fatos etc.; 12. identificar publicações originais nas quais uma ideia ou um conceito são discutidos; 13. identificar publicações originais que descrevam conceitos ou termos epônimos, por exemplo, Mal de Hodgkin; 14. contestar trabalhos ou ideias de outros; 15. debater a primazia das declarações de outros (MACIAS-CHAPULA, 1998, p. 136).

Também deve-se acrescentar a estes o fato de que as citações dão acessibilidade ao material citado, pois, segundo Guedes e Borschiver (2005), a citação indica uso do citado pelo citante. Permitem a possibilidade de contar artigos, periódicos, autores, departamentos acadêmicos, universidades, institutos de pesquisa, entre outros, além do que um artigo de periódico muito citado representa a aceitação da comunidade que o cita. Porém, nem todos os motivos para citar estão relacionados com as convenções reconhecidas pela publicação acadêmica.

> Os fatores sociais e psicológicos, por exemplo, têm aí uma função significativa, juntamente com as lembranças e esquecimentos subconscientes. Há também fatores extrínsecos; por exemplo, os leitores ou a percepção do autor em relação às necessidades e expectativas dos leitores. O perfil e o status da revista na qual o artigo será publicado; a abrangência, formato e extensão do artigo; o conhecimento do autor sobre a área e a sua habilidade/disposição em usar as fontes e os serviços apropriados de informação são outros fatores envolvidos (MACIAS-CHAPULA, 1998, p. 136).

Smith, citado por Guedes e Borschiver (2005), destaca que a citação como meio de quantificação da informação científica, apesar da premissa de que indica uma relação de assuntos, pode conter relações que não sejam especificadas. Isso é confirmado por May (1967), citado por Macias-Chapula (1998), quando desafiou a visão ortodoxa de que as citações oferecem um quadro preciso das relações intelectuais entre as publicações e concluiu que os autores selecionam citações com vistas mais a seus objetivos científicos, políticos e pessoais, do que para questões de cunho intelectual. Macias-Chapula (1998) destaca outros autores que relativizam a precisão da análise de citações, como Lipetz (1965), Chubin e Moitra (1975), Moravcsik e Murugesan (1975), Oppenheim e Renn (1978), Hodges (1978), Finney (1979), Frost (1979), Duncan et al. (1981), Bonzi (1982) e O'Connor (1982). Ora, se a análise de citações/cocitações não é uma unidade de medida da informação científica em 100% dos casos, talvez ela possa apresentar uma acurácia bem próxima disso (relação 80/20), onde 80% da demanda de informação é satisfeita com 20% do conjunto de fontes de informação. Então, estaríamos diante de uma "Lei dos 80/20", que se apresenta como a quinta forma de medir a informação científica, sendo que essa lei é uma aplicação da Lei de Potência de Pareto, de 1896, já apresentada aqui.

> [...] embora não baseada na análise de citações, vale destacar a Lei dos 80/20, que consiste em um fenômeno, inicialmente observado no comércio e na indústria, segundo o qual em sistemas de informação 80% da demanda de informação se satisfaz com 20% do conjunto de fontes de informação. (TRUESWELL, 1969). Em sistemas de informação, esta lei pode ser usada nas tomadas de decisão relacionadas à composição e redução de acervos (GUEDES; BORSCHIVER, 2005, p. 13).

Nesse sentido, a análise de citações deve levar em consideração, além dos aspectos quantitativos, os hábitos, atitudes, experiências e expectativas de grupos onde se inserem os autores que são os produtores do conhecimento. Macias-Chapula (1998) destaca quatro grupos de autores: os controladores de qualidade, os educadores, os consumidores e os produtores.

> É necessário pensar a citação como um processo. Os resultados desse processo são as listas de citações que acompanham os trabalhos acadêmicos. O tipo e a composição dessas listas refletem a personalidade do autor e seu meio profissional. Não existe uma teoria única da citação capaz de explicar por que os autores citam de uma determinada maneira. A maioria das análises de citação tem apresentado mais características internas do que externas, concentrando-se em dados quantitativos e em distribuições de frequência, em vez de tratar os contextos nos quais os autores utilizam as citações. A análise de citação exige este último tipo de estudo (MACIAS-CHAPULA, 1998, p. 136).

Finalmente, em uma perspectiva quantitativa ou qualitativa, a análise de citações para a medição da informação científica tem as seguintes aplicações: na gestão de coleções em bibliotecas; no mapeamento do desempenho dos autores nas diversas áreas das ciências; na administração de financiamentos de pesquisa, auxílio, bolsas, orçamento de sistemas de informação e bibliotecas; como ferramenta para a recuperação da informação e avaliação de periódicos, produtividade de autores, medida de qualidade de uma dada informação, medida do fluxo de informação em uma unidade, sociologia da ciência, indicador de estruturas e tendências científicas (GUEDES; BORSCHIVER, 2005).

2.3 A INFORMETRIA

Os estudos métricos na CI têm o objetivo de fornecer um conjunto de informações que conduzam melhorias para as instituições, governos e sociedades necessitando acumular, organizar e transmitir conhecimentos, ensejando o grande campo da informetria. Dito de outra forma, o aporte da CI fez os aspectos relacionados à tradicional bibliometria crescerem em direção às metrias da informação, fazendo surgir na agenda de pesquisas disciplinas ou campos que vão além da bibliometria e se inter-relacionam com esta, como bem identificou Araújo (2006) em sua pesquisa, e que permeará esta subseção juntamente com as ideias de outros autores.

> Um deles é a informetria, termo utilizado pela primeira vez na Alemanha, por Nacke, em 1979. Outro é a cientometria, também conhecida no Brasil como cienciometria, termo popularizado pelo periódico húngaro de mesmo nome, fundado em 1977 por Braun. Vários autores se preocuparam com a

> caracterização de cada um deles (TAGUE-SUTCLIFFE, 1992, tradução nossa; LE COADIC, 1996, p. 52). Uma caracterização bem consistente foi proposta por McGrath em 1989, para quem a bibliometria, a cienciometria e a informetria são subdisciplinas que se assemelham por serem métodos quantitativos, mas se diferenciam quanto ao objeto de estudo, as variáveis, os métodos específicos e os objetivos. Apenas para se ter uma ideia dessas diferenças, o autor argumenta que o objeto de estudo da primeira são livros, documentos, revistas, artigos, autores e usuários; da segunda, disciplinas, assuntos, áreas e campos; e da terceira, palavras, documentos e bases de dados (MACIAS-CHAPULA, 1998, p. 135). Mais recentemente, um outro subcampo surgiu, a webometria. Um trabalho que se utiliza da mesma estrutura de McGrath propõe que o objeto de estudo da webometria são os sítios na world wide web (VANTI, 2002, p. 160) (ARAÚJO, 2006, p. 22).

Além disso, autores como Pritchard (1969), Tague-Sutcliffe (1992), Le coadic (1996), Macias-Chapula (1998), Spinak (1998), Vanti (2002) e outros questionam a visão exclusivamente quantitativa nos estudos das metrias da informação, colocando a proposta de que a tais estudos deveriam ser somados outros com um olhar social. A consequência, apresentada aqui, é a junção das ciências sociais com a CI no tocante às metrias da informação. Um exemplo disso é a própria proposta desta pesquisa de juntar os estudos da produção científica brasileira com o método de MQO.

> A análise dos dados informétricos e cienciométricos oferece informações sobre a orientação e a dinâmica científica de um país, bem como sobre sua participação na ciência e na tecnologia mundial. Análises cooperativas tornam possível identificar redes científicas e revelar os elos entre países, instituições e pesquisadores, assim como permitem conhecer o impacto dos principais programas e organizações. A cienciometria também traz à luz a estrutura das disciplinas científicas e as conexões entre elas (MACIAS-CHAPULA, 1998, p. 137).

Macias-chalupa (1998), assim como Tague-Sutckiffe (1992), consideram que a informetria compreende um campo mais amplo que engloba a bibliometria, a cienciometria e a webometria. Porém, outros insistem que tais metrias da informação devem ser tratadas como sinônimas (GLÄZEL E SCHOEPFLIN, 1994; VANTI, 2002).

> Cabe, portanto, considerar ‘informetria’ como o termo ‘guarda-chuva’ que consegue abarcar os outros três conceitos dentro dele. [...] devemos consolidar a ideia de que todas estas especialidades encontram sua ligação a partir da sua origem comum: todas elas indubitavelmente constituem-se em subcampos da ciência da informação. Considerando-as, desta maneira, poderemos compreender mais satisfatoriamente a aplicabilidade de cada uma e a utilidade que oferecem para as mais diversas áreas do conhecimento (VANTI, 2002, p. 161).

Isso já era identificado por Le Coadic (1994) quando afirmava que o objeto da informetria é medir as atividades da informação científica e técnica porque a mensuração de elementos de informação permite elaborar indicadores quantitativos (medidas) e qualitativos

das atividades de construção, comunicação e uso das informações científicas e técnicas que, para tanto, aplicam-se métodos matemáticos e estatísticos. Talvez em 1994 e 1998, Le Coadic e Macias-Chapula tivessem uma visão das metrias da informação mais voltadas para a área científica e técnica, mas isso permite uma primeira aproximação para um conceito de informetria.

Le Coadic (2005) explica para aqueles que estranham a *mathématisation* ("matematização") que a aplicação da matemática à análise dos fenômenos sociais e humanos data de muito tempo e aponta os exemplos de Georges Buffon com sua aritmética moral, e de Marie-Jean Condorcet com o uso de análise quantitativas para a contagem dos votos. O autor cita também a economia do final do século XIX, a demografia, a psicologia da década de 1910 e a sociologia da década de 1950, como áreas que utilizaram a investigação matemática. Ainda segundo o autor, essa aplicação não aconteceu facilmente, por exemplo, para publicar e legitimar trabalhos quantitativos na área da biologia no início século XX, foi chamado para criar a revista *Biometrika* o estatístico Karl Pearson.

Isso demonstra que, apesar da diversidade de aplicações e do caráter interdisciplinar da matemática e da estatística, existe a necessidade de um esforço coletivo para o reconhecimento de estudos métricos nos campos do conhecimento humano, principalmente no sentido de que as análises das metrias da informação transcendam para estudos em qualquer campo de produção, organização e disseminação de informações.

Quanto à medida das atividades de informação, Le Coadic (2005) reforça que não há ciência ou técnica sem medida e, principalmente, sem medidas exatas, mas reconhece que não foi fácil introduzir a mensuração em centros de informação (bibliotecas, centros de documentação etc.) quando ainda não havia uma ciência própria e que só foi possível pensar o desenvolvimento da informetria a partir de uma ciência que tornasse as metrias um objeto de pesquisa e passasse a usar técnicas eletrônicas e fotônicas de informação, além de digitalizar os fluxos de informação disso resultantes.

As ideias de Le Coadic, principalmente a de que se existe o número (quantitativo) nos sistemas de informação é possível medir, não deixam dúvidas de que o campo das metrias da informação é amplo para pesquisas em diversos cenários e se complementa a autores que estudam a interdisciplinaridade. Mas a informetria tem seu surgimento "posterior aos dos outros dois termos [bibliometria e cienciomentria], pode-se dizer que esta [informetria] tem um escopo tanto mais distinto e abrangente do que a cienciomentria e bibliometria" (VANTI, 2002, p. 154).

Tague-Sutckiffe (1992), Macias-Chapula (1998) e Vanti (2002) definem a informetria como um estudo dos aspectos quantitativos da informação em qualquer formato, não apenas

registros catalográficos ou bibliografias, referente a qualquer grupo social, e não apenas aos cientistas. Assim, "A informetria pode incorporar, utilizar e ampliar os muitos estudos de avaliação da informação que estão fora dos limites tanto da bibliometria como da cienciometria" (MACIAS-CHAPULA, 1998, p. 135).

Tague-Sutckiffe (1992) expande o conceito, que continuou com as ideias de Macias-Chapula (1998), Vanti (2002), quando afirma que a diferença é que a informetria pode incorporar, utilizar e ampliar os muitos estudos de avaliação da informação que estão fora dos limites tanto da bibliometria como da cienciometria.

Segundo a literatura da área, todas essas áreas aproximam as metrias da informação das ciências exatas pelo fato de que seus princípios convergem para quantificar a produção, armazenamento e disseminação de informações, de forma empírica, para produção do conhecimento, que no início era centrado na medição do livro ou do suporte da informação. Pode-se afirmar que os métodos empíricos da informetria foram construídos através de tentativas e erros baseados nas observações e experiências acumuladas desde a bibliometria, e é dessa dinâmica que surgem as bases das pesquisas métricas em informação que, em seus estudos de caso, relacionam fortemente seus resultados à aplicação de técnicas estatísticas.

> O escopo da informetria é prático e teórico (Glanzel e Schoepflin), sendo que sua prioridade, em primeiro lugar, tem sido o desenvolvimento de modelos matemáticos e, em segundo lugar, a determinação de medidas para o fenômeno estudado. Os modelos oferecem uma base prática para a tomada de decisões, e seu valor está na sua capacidade de sintetizar, em poucos parâmetros, as características de muitos grupos de dados: formato completo, concentração, difusão e mudança através do tempo (MACIAS-CHAPULA, 1998, p. 135).

Segundo Macias-Chapula (1998, p. 137), os seguintes indicadores infométricos são atualmente bastante conhecidos pela sua importância no cenário nacional e internacional:

> a) Número de trabalhos – Reflete os produtos da ciência, medidos pela contagem dos trabalhos e pelo tipo de documentos (livros, artigos, publicações científicas, relatórios etc.). A dinâmica da pesquisa em um determinado país pode ser monitorada e sua tendência traçada ao longo do tempo.
> b) Número de patentes – Reflete as tendências das mudanças técnicas ao longo do tempo e avalia os resultados dos recursos investidos em atividades de P&D. Esses indicadores determinam o grau aproximado da inovação tecnológica de um país.
> c) Coautoria – Reflete o grau de colaboração na ciência em nível nacional e internacional. O crescimento ou o declínio da pesquisa cooperativa podem ser medidos.
> d) Mapas dos campos científicos e dos países – Auxiliam a localizar as posições relativas de diferentes países na cooperação científica global.
> e) Número de citações – Reflete o impacto dos artigos ou assuntos citados.
> f) Número de citações de patentes – Mede o impacto da tecnologia.

> O termo informetria designa, conforme Hjotgaard Christensen & Ingwersen, uma extensão recente das análises bibliométricas tradicionais ao abarcar o estudo das modalidades de produção da informação e de comunicação em comunidades não acadêmicas. Para estes autores, as análises de difusão de determinados assuntos nos bancos de dados *full-text* dos jornais seriam uma das novas possibilidades que surgem neste campo. A informetria se distinguiria claramente da cienciometria e da bibliometria no que diz respeito ao universo de objetos e sujeitos que estuda, não se limitando apenas à informação registrada, dado que pode analisar também os processos de comunicação informal, inclusive falada, e dedicar-se a pesquisar os usos e necessidades de informação dos grupos sociais desfavorecidos, e não só das elites intelectuais (VANTI, 2002, p. 155).

Na visão de Vanti (2002), a informetria tem um potencial de ciência guarda-chuva, devido à possibilidade de englobar metrias no âmbito de atuação de vários atores, desde os clássicos (bibliotecas, arquivistas, museus, centros de documentação etc.) até os modernos (bases de dados, bases de pesquisas, sistemas de recuperação da informação, avaliação de usos e usuários etc.). Para que os profissionais das mais diversas áreas possam resolver seus problemas informacionais, é importante perceber a versatilidade das metrias e a experiência da área da CI. Todas as áreas do conhecimento necessitam agregar suas informações oriundas de outros conhecimentos e de sua própria área e a informetria pode auxiliar nesse processo. Por isso, os métodos informétricos se valem de fundamentos de muitas outras bases e áreas do conhecimento, mas o objeto é sempre o mesmo: medir a informação.

Ainda para dar conta de medir a informação, Le Coadic (1994, p. 54-55) lista três métodos infométricos:

> Os monodimensionais: a) apoiam-se em classificações, nomenclaturas preestabelecidas e se baseiam na contagem do número de publicações que apresentem características similares (artigos ou patentes), como pertencer a um mesmo número de classificação; b) contam as citações em um período de tempo para medir a produtividade de um autor, país ou instituição; c) medem o impacto de uma revista, assunto ou autor;
>
> Os métodos bidimensionais ou relacionais: a) permitem a detecção de uma relação entre elementos de informação; b) identificam a estrutura de um campo de atividade científica ou técnica que é representada em forma gráfica e de mapas;
>
> E, por último, os métodos multidimensionais que utilizam métodos estatísticos e a análise fatorial.

Por fim, é possível afirmar que a informetria é uma disciplina social porque as metrias da informação apresentam dimensões objetivas (bibliometria, webometria, cienciometria) e dimensões culturais, quando, por exemplo, essas medições auxiliam nas políticas para o desenvolvimento das ciências. Isso faz com que as metrias da informação possam, através de seus elementos mensuráveis, ajudar na consolidação da produção do conhecimento. Nesse

ponto, devemos ressaltar que é forte a semelhança da informetria e a econometria porque ambas visam construir modelos que expliquem os comportamentos humanas dentro de determinados fenômenos. Tais aproximações e distanciamentos entre os métodos métricos da CI e da Econometria foram abordados no próximo capítulo.

3 ECONOMETRIA E AS METRIAS DA INFORMAÇÃO

Com alterações nos tradicionais setores da economia, a informação, ao lado da agropecuária, indústria e dos serviços, vem se constituindo em um quarto setor produtivo. A sociedade passa, assim, a dar um maior valor à informação. Ao longo dos anos, o ser humano vem aprimorando o processo de preservação do conhecimento e do desenvolvimento através do acesso e uso da informação. O uso da (in)formação de uma população pode ser avaliado pelo seu nível de desenvolvimento tecnológico e de acesso (BURNHAM, 2005). O Índice de Desenvolvimento Econômico (IDH) aponta como variável determinante o nível de acesso à informação, viabilizado pelas Tecnologias da Informação e da Comunicação (TICs) (RUA, 1998) (BOLAÑO; MELO, 2000).

Ao longo da década de 1980, segundo Araújo (2006), houve uma queda no interesse pela bibliometria, tanto no Brasil como no exterior. No início dos anos 1990, com as possibilidades do uso do computador, voltou a haver um grande interesse na exploração das metodologias quantitativas. Na verdade, desde a primeira International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval, na Bélgica, em 1987, essa disposição já havia se manifestado. Macias-Chapula (1998) acrescenta que esses problemas e limitações demandaram um diálogo entre autores, editores e as empresas que gerenciam as bases de dados e, particularmente, os países em desenvolvimento têm urgência em metodologias quantitativas para identificar e tratar problemas relacionados às novas tecnologias de informação.

Atualmente, as organizações já assimilaram a ideia de que a informação precisa e em tempo hábil é crucial para o desenvolvimento de planos estratégicos e para a própria gestão do empreendimento. O grande problema é que, em muitos casos, como no exemplo usado aqui referente aos órgãos de fomento, a informação disponibilizada acaba sendo obtusa e indireta. Para piorar, aparece como incontestável já que o órgão que a disponibiliza e a divulga sofre de dois males que são a extrema confiança e a crença quase absoluta em seus dados.

Entre os aspectos relacionados ao crescimento da bibliometria na agenda de pesquisa está o surgimento de algumas subdisciplinas ou subcampos da bibliometria. Um deles é a "informetria", termo utilizado pela primeira vez na Alemanha, por Nacke, em 1979. Outro é a cientometria, também conhecida no Brasil como "cienciometria", termo popularizado pelo periódico húngaro de mesmo nome, fundado em 1977, por Braun. Vários autores se preocuparam com a caracterização de cada um deles (tais como Tague-Sutcliffe, em 1992, e Le Coadic, em 1996). Para McGrath (1989), lembrado por Macias-Chapula (1998), a bibliometria,

a cienciometria e a informetria são subdisciplinas que se assemelham por serem métodos quantitativos, mas se diferenciam quanto ao objeto de estudo, as variáveis, os métodos específicos e os objetivos: o objeto de estudo da primeira são livros, documentos, revistas, artigos, autores e usuários; da segunda, disciplinas, assuntos, áreas e campos; e da terceira, palavras, documentos e bases de dados. Mais recentemente surgiu a webometria que, segundo Vanti (2002), tem como objeto de estudo os sítios na World Wide Web (ARAÚJO, 2006).

O ressurgimento da bibliometria e de outras áreas, como a cienciometria, a informetria e a webometria revelam e expressam que uma grande massa de informação processada deve ser mensurada para potencializar a transformação em conhecimento. Portanto, segundo Moita Neto (2004), cada vez mais precisamos de ferramentas estatísticas que apresentem uma visão mais global do fenômeno do que aquela que é possível em uma abordagem univariada. A abordagem univariada é algo bastante comum e encarado com normalidade nas pesquisas bibliométricas, principalmente na cienciometria, na qual, a partir das suas bases de dados, calculam-se médias, medianas, índices, percentuais, desvios e moda para os estudos das políticas de publicação científica.

> Glanzel (1993) declarou que as políticas de publicação e as possibilidades de recuperação oferecidas pelos negociantes de bases de dados tendem a limitar a pesquisa bibliométrica. As bases de dados são insuficientes diante das expectativas dos especialistas em bibliometria. Certas mudanças na estrutura dos dados ou dos padrões fariam sentido somente se pudessem retornar há pelo menos 10 ou 15 anos. Contudo, mudanças desse nível parecem ser definitivamente irreais, considerando os custos para os usuários (MACIAS-CHAPULA, 1998, p. 138).

A abordagem univariada tão somente pode gerar dados que sejam interpretados como informação. Choo (2003) alerta que dados brutos tomados como informação processada certamente prejudicam a caminhada em direção ao conhecimento, pois, embora a tomada de decisão seja um processo complexo, não há dúvida de que é essencial para a vida da organização: toda ação da organização é provocada por uma decisão e toda decisão é um compromisso para uma ação.

São exemplos, segundo Araújo (2006), do atual desenvolvimento de métodos e técnicas bibliométricas: a) os grupos de pesquisa: BIRG (de Sydney, Austrália), CEST (de Berna, Suíça), CINDOC (de Madrid, Espanha), CIS (de Copenhagen, Dinamarca), CRRM (de Marselha, França), CWTS (de Leiden, Holanda), FhG-ISI (de Karlsruhe, Alemanha), Inforsk (de Umeä, Suécia), OST (de Montreal, Canadá), OST (de Paris, França), SPRU (de Sussex, Inglaterra) e ISI Research Service Group (de Filadélfia, EUA); b) a realização, a cada dois anos, da International Conference on Scientometrics and Informetrics; c) os periódicos *Bibliometric*

Notes, *Cybermetrics* e *Scientometrics*, especializados no assunto, além de outros 14 que publicam trabalhos relacionados ao assunto; d) os indicadores de ciência e tecnologia produzidos atualmente: *Science and Technology Indicators* (EUA), *Science & Technology Indicators* (Ásia), *S&T Indicators for de European Research Area* (Europa) e *Main Science and Technology Indicators* (dos países da OECD); e) a existência, na história da elaboração de indicadores, dos manuais Frascati, de Oslo e de Canberra como importantes momentos de busca de consensos internacionais sobre os índices de inovação científica e tecnológica.

Tudo isso confirma que a informação deve proporcionar um novo olhar dentro de um determinado contexto para a produção de conhecimentos, de modo que seu potencial de análise é muito mais complexo do que uma análise simplesmente linear ou univariada. Exemplos disso não estão somente circunscritos a dados incompletos ou fontes não confiáveis, mas também a análises enviesadas.

> Na maioria dos casos, os editores de revistas ou mesmo os autores são os responsáveis por omissões de dados ou por informações incorretas. Outros fatores que prejudicam a confiabilidade das bases de dados bibliográficos são a elaboração superficial e arbitrária dos registros bibliográficos (por exemplo, autores com diferentes afiliações), os critérios de seleção não documentados e os problemas de compatibilidade entre diferentes versões da mesma base de dados (por exemplo, on-line e CD-ROM) (MACIAS-CHAPULA, 1998, p. 138).

Nas organizações, como as agências de fomento, a informação é um componente intrínseco de quase tudo que se faz. Sem uma clara compreensão dos processos organizacionais e humanos pelos quais a informação se transforma em percepção, conhecimento e ação, as organizações não são capazes de perceber a importância de suas fontes e tecnologias de informação (CHOO, 2003).

As informações, como na análise de citações, afirma Macias-Chapula (1998), só podem ser compreendidas mediante o exame das condições sociais em que os cientistas citam. O significado da citação revela uma realidade social, na qual a atitude de citar é um fenômeno que deve ser entendido em sua completude, uma vez que os autores não citam da mesma forma. Assim, as técnicas de indexação de citações não são suficientes para explicar as relações entre os documentos, pois é necessário que a indexação seja complementada por estudos quantitativos da literatura científica. Dessa forma, "A evolução dos estudos em produção científica, assim, assistiu à conversão da bibliometria, de um campo de pesquisa, em técnica – uma técnica útil, que deve ser adotada em conjunto com métodos qualitativos fornecidos pelas ciências sociais" (ARAÚJO, 2006, p. 23).

Castells (1999) aponta para o pensamento de que, independente da tecnologia, existe um ponto central que é o uso da informação para gerar conhecimento, não de maneira linear, mas de forma cíclica, onde a realimentação criaria uma espiral de inovação e usos. Nesse sentido, vimos nesta seção as leis clássicas da bibliometria (Leis de Zipf, Lotka e Bradford) e visões mais modernas como as da cientometria: frente de pesquisa, fator de imediatismo ou de impacto, acoplamento bibliográfico, obsolescência da literatura, Lei do Elitismo e Teoria Epidêmica de Goffman. Assim, atualmente, cremos que as técnicas bibliométricas devem se aliar a outros referenciais e métodos que levem em consideração o contexto sócio-histórico em que a atividade científica é produzida. Aliás, algo bastante conexo com a proposta desta tese. São exemplos disso, segundo Araújo (2006), trabalhos que estudam: a) a historicidade da produção científica a partir de conceitos da arqueologia do saber de Foucault (ALVARENGA, 1996); b) a região geográfica como fator interveniente na produção científica (TARGINO, 1998); c) a identidade dos pesquisadores em aspectos relacionados à carreira, motivações, produtividade, qualidade da produção, colaboração (MEADOWS, 1999; LACEY, 1998; WITTER, 1997); d) os conceitos de centro e periferia como base teórica para a compreensão da comunicação científica (MUELLER; OLIVEIRA, 2003).

> [...] o Centro de Estudos Informétricos de Copenhague busca uma nova abordagem para a área, qual seja, a combinação de teorias e metodologias avançadas de recuperação da informação com o estudo científico dos fluxos de informação. O Centro objetiva aplicar métodos bibliométricos não somente em estudos cienciométricos e em avaliações de pesquisa científica e tecnológica, mas também na análise de suas relações sociais, econômicas etc., ampliando as análises bibliométricas tradicionais [...] (ARAÚJO, 2006, p. 25).

No atual contexto de competitividade, baseada na contínua evolução de tecnologias e na crescente modernização das práticas científicas, verifica-se a necessidade de avanços para as metrias da informação, com a utilização de métodos capazes de captar o que existe para entender; e adaptar, modificar e ajustar os referidos avanços às condições concretas de cada cenário observado. Esta pesquisa se coaduna com isso quando insere o MQO no rol de técnicas e métodos que podem ampliar a análise da produção científica brasileira.

Mesmo com todos os avanços, dificuldades existem. A análise de citação é uma ferramenta importante que permite várias verificações infométricas, bibliométricas e cienciométricas. Mas, segundo Macias-Chapula (1998), MacRoberts e MacRoberts relacionaram os seguintes problemas da análise de citação, enquanto fenômeno e dados: a) influências formais não citadas; b) citação tendenciosa ou preconcebida; c) influências informais não citadas; d) autocitação; e) diferentes tipos de citação; f) variações nas médias de citação relacionadas ao tipo de publicação, nacionalidade, período, extensão e especialidade; g)

limitações técnicas de índices de citação e bibliografias: autoria múltipla, sinônimos, homônimos, erros de edição, cobertura da literatura.

Em relação às metrias da informação, preocupações crescentes também incidem sobre desenvolvimento científico e tecnológico referente a como medir a produção científica. No Brasil, o MCT vem adotando várias medidas para implementar tais mecanismos que gerenciem o controle do processo produtivo retratando o cenário nacional. Essa preocupação não só com o desenvolvimento nacional, mas também em tornar o país mais competitivo no mercado mundial é, sem dúvida, baseada na compreensão dos dados dessa produção.

> Diversas frentes de estudo são levadas a termo na atualidade com essa proposta. Há, por exemplo, estudos de usuários feitos com o auxílio de técnicas bibliométricas. É o caso do estudo de Oliveira (2004), que analisa a possibilidade de aquisição de itens para uma biblioteca universitária a partir de indicativos de necessidades de usuários obtidos com o estudo bibliométrico das referências bibliográficas de teses e dissertações. Ou, então, de estudos de uso de coleções também a partir da aplicação da bibliometria às referências de teses e dissertações defendidas em uma instituição específica (COITO et al, 2002; PAULA et al., 2002). E, ainda, estudos que evidenciam os processos de comunicação científica e formação de colégios invisíveis a partir da análise de citação (MELLO, 1996; NORONHA, 1998) (ARAÚJO, 2006, p. 25).

Através do uso de novas TICs é possível promover políticas públicas para o acesso e disseminação da informação. O ideal não é só participar das decisões no que tange a esse universo, mas também contribuir para a construção de projetos políticos que, para tanto, devem contar com indicadores que possam orientar a alocação de recursos para as políticas públicas voltadas para a área da informação. No caso da informação científica, Macias-Chapula, (1998) indica que os principais problemas referentes aos cálculos estatísticos de publicações são: a) as publicações oferecem contribuições diferentes ao conhecimento científico; b) existem variações nas médias de publicação em especialidade e contexto institucional.

Os números apontam um aumento significativo de publicações nacionais nos últimos anos, com destaque para o crescimento dessa participação em bases internacionais importantes como a *Web of Science* e segundo Araújo (2006),

> O uso de dados bibliométricos como indicadores da produção científica passou a ser cada vez mais freqüente, diante do conjunto de ações que vêm sendo desenvolvidas no sentido de dispor desses indicadores para o planejamento nacional das atividades de pesquisa científica (MUGNANI; JANNUZZI; QUONIAM, 2004, p. 123; GUEDES; BORSCHIVER, 2005; KRZYZANOWSKI; FERREIRA, 1998). Ou então para a análise do desenvolvimento da pesquisa científica e tecnológica dentro de uma instituição específica (PENTEADO FILHO et al., 2002; ROUSSEAU, 1998), da análise dos periódicos de uma área específica (CAMPOS, 2003; ELKIS, 1999) ou da produtividade de pesquisadores (POBLACIÓN; NORONHA, 2002). Há ainda estudos bibliométricos para determinar o "léxico básico" de um campo, com possibilidades de aplicação inclusive para a construção de

> tesauros e linguagens documentárias desse campo (ROBREDO; CUNHA, 1998). (ARAÚJO, 2006, p.25)

Durante muito tempo, o MCT estuda as melhores formas de traduzir esse crescimento e também as melhores formas de investir em ciência e tecnologia. Desenvolver políticas voltadas para o setor é um dos principais objetivos dos estudos, subsidiando com dados os órgãos de fomento que, por sua vez, utilizam-se destes na escolha da melhor forma de patrocinar o desenvolvimento científico e tecnológico do país.

Segundo Araújo e Freire (1999), torna-se evidente, por esse contexto, que na Era do Conhecimento a ciência conservará e, mais ainda, reforçará sua importância na disputa das capacidades produtivas dos países desenvolvidos.

> Percebe-se, dessa forma, que a bibliometria vem se consolidando como método de estudo dentro de uma preocupação com leituras mais ricas da realidade, mais atentas às reivindicações contemporâneas do pensamento complexo (MORIN, 1987; 2001) bem como às críticas ao projeto da mathesis universalis (a consideração exclusiva daquilo que é mensurável como cientificamente relevante) do projeto positivista de ciência (SANTOS, 1986). A imensa popularidade que a bibliometria passou a ter após as possibilidades digitais foi acompanhada, portanto, de uma série de avanços relativos ao aperfeiçoamento das leis bibliométricas mas, sobretudo, pela busca de fundamentação teórica e conceitos oriundos dos contextos concretos em que os fenômenos informacionais ocorrem (BORGMAN; FURNER, 2002, tradução nossa) (ARAÚJO, 2006, p. 26).

A preocupação em entender as relações entre as atividades científicas e suas relações com o desenvolvimento econômico e social tem elevado a importância da análise dos indicadores de produção científica. Apresentados como principal ferramenta para compreender a dinâmica dos processos científicos e fornecer elementos para direcionar políticas em ciência e tecnologia, os indicadores de produção científica são cada vez mais necessários ao planejamento e avaliação de resultados.

Segundo a FAPESP (2005, p. 5), os indicadores de produção científica, somados à família de indicadores de insumos para a Ciência e Tecnologia (C&T), têm contribuído de forma definitiva para a análise do desempenho e melhoria da eficiência dos sistemas nacionais de ciência, tecnologia e inovação. Os indicadores podem ser compreendidos como dados estatísticos usados para medir algo intangível, que ilustram aspectos de uma realidade multifacetada. A construção e o uso de indicadores de produção científica são objeto de estudo de várias áreas do conhecimento, sendo usados tanto para o planejamento e a execução de políticas para o setor, como também para que a comunidade científica conheça melhor o sistema no qual está inserida.

Mesmo que a bibliometria venha se consolidando como método de estudo dentro de uma preocupação com leituras mais ricas da realidade, atentas ao pensamento complexo, bem como àquilo que é mensurável como cientificamente relevante, ainda há muito o que se avançar. Mesmo diante das possibilidades que as metrias da informação apresentam com a era digital, ainda há que se avançar no aperfeiçoamento das teorias da informação. Os fenômenos informacionais atuais ocorrem em contextos concretos que exigem medições eficientes.

> As preocupações atuais dos pesquisadores em bibliometria caminham para direções distintas. Há um grupo particularmente preocupado com o aperfeiçoamento das fórmulas que expressam as leis. É o caso, por exemplo, dos estudos de Burrell (1992) que analisam a possibilidade de uso da curva de Leimkuhler, do índice de Gini e da distribuição de Pareto na análise das distribuições dos valores bibliométricos. Ou, então, de Urbizagástegui Alvarado (2004), que postula o uso da distribuição de Poisson, formulada ainda no século XIX para estimar a probabilidade de ocorrência de um evento durante um período determinado, para a análise da produtividade de autores (ARAÚJO, 2006, p. 23).

As áreas da bibliometria, cienciometria, informetria e webometria não podem se cingir tão somente de abordagens univariadas no seu campo de atuação, achando isso algo bastante comum e encarado com normalidade nas suas pesquisas. As áreas devem ir além do cálculo de médias, medianas, índices, percentuais, desvios e moda para os estudos informacionais, no geral, e para os estudos das políticas e de publicação científica, no específico desta tese, quando propõe o uso de MQO.

Isso é confirmado quando as referências de MQO na área da CI só retornam com dois artigos, onde, ao que parece, Rubén Urbizagástegui Alvarado é um dos poucos, senão o único, a tratar do assunto. Esses artigos são: “A Lei de Lotka na bibliometria brasileira” (2002); “A produtividade dos autores na literatura de enfermagem: um modelo de aplicação da Lei de Lotka” (2006).

Por sua vez, a estatística aplicada permite, de forma sistemática, organizar, descrever, analisar e interpretar dados obtidos de estudos ou experimentos, realizados em qualquer área do conhecimento. Porém, a questão é a sua aplicação direta a outros campos, entendendo que em alguns casos essa aplicação será simples, com obtenção direta de valores e/ou novos dados, e em outros haverá a necessidade de se encontrar uma relação interdisciplinar que justifique e/ou auxilie na interpretação desses resultados. A econometria é uma disciplina que surge justamente a partir de outras pré-existentes (economia e estatística matemática), para consolidar resultados de observações junto às teorias econômicas.

Gujarati e Porter (2011, p. 39) definem a econometria como a aplicação da estatística matemática à dados econômicos, ou seja, é uma ciência em que a teoria econômica, as

ferramentas da matemática e da inferência estatística são aplicadas juntas à análise dos fenômenos econômicos. A pesquisa econométrica visa, portanto, prover medições concretas para a teoria econômica, fazendo uma ponte entre a teoria econômica e as técnicas da inferência estatística. O próprio autor esclarece que a econometria teórica trata do desenvolvimento de métodos adequados para medir as relações econômicas especificadas nos modelos econométricos.

Logo, a econometria é dependente da estatística e da matemática. Essa visão já está presente em conceitos clássicos da área, como o de Surrey (1974, p. 9), que define a econometria como "a arte de confrontar o raciocínio a priori do economista teórico com a evidência estatística disponível". E também por Klein (1978, p. 11), quando afirma que "o principal objetivo da econometria é dar conteúdo empírico ao raciocínio econômico apriorístico. Este raciocínio apriorístico é composto, principalmente, pelo que chamamos de teoria econômica". Sendo possível também dizer, voltando às palavras de Surrey (1974, p. 11), que "um segundo modo de ver pelo menos uma parte da econometria é o de reconhecer nela um meio de se confrontar com a quase impossível tarefa de conduzir uma experiência controlada através dos conhecimentos da economia".

Aproximando tais princípios aos métodos de medição da informação, dentro dos parâmetros da CI, temos a definição de Guedes e Borschiver (2005, p. 15), que apresentam a bibliometria como "uma ferramenta estatística que permite mapear e gerar diferentes indicadores de tratamento e gestão da informação e do conhecimento […], necessários ao planejamento, avaliação e gestão da ciência e da tecnologia [...] é também um instrumento quantitativo, que contribui para tomadas de decisão na gestão da informação e do conhecimento". De certa forma, alguns dos métodos métricos da informação, como as leis bibliométricas de Lotka, Bradford e talvez Zipf, possuem formas da curva que permitam justificar um modelo matemático e estatístico, porém, em outros métodos, como a Teoria Epidêmica de Goffman, não é possível afirmar. Os demais métodos, no entendimento desta tese, valem-se tão somente de abordagens univariadas, com cálculos que se utilizam de médias, medianas, índices, percentuais, desvios e moda para os estudos informacionais.

Para entender o escopo da disciplina econometria, Klein (1978, p. 11) segue explicando que essa "é um ramo da economia no qual se estuda a medida das relações discutidas na análise econômica apriorística. […] [o que também] pode levar à descoberta de novas relações ou teorias até então insuspeitadas pelas considerações puramente a priori". A CI segue um caminho semelhante e, segundo Vanti (2002, p. 155), em termos genéricos, a aplicação das técnicas bibliométricas, cienciométricas e informétricas objetivam: identificar as tendências e o

crescimento do conhecimento em uma área; identificar as revistas do núcleo de uma disciplina; mensurar a cobertura das revistas secundárias; identificar os usuários de uma disciplina; prever as tendências de publicação; estudar a dispersão e a obsolescência da literatura científica; prever a produtividade de autores individuais, organizações e países; medir o grau e padrões de colaboração entre autores; analisar os processos de citação e cocitação; determinar o desempenho dos sistemas de recuperação da informação; avaliar os aspectos estatísticos da linguagem, das palavras e das frases; avaliar a circulação e uso de documentos em um centro de documentação; medir o crescimento de determinadas áreas e o surgimento de novos temas. A diferença está no fato de que os dados estatísticos para a econometria acompanham o comportamento econômico para tomada de decisão e tem como suporte um lastro teórico.

> O sangue vital da econometria são as séries estatísticas que dão as medidas das variáveis. […]. As estatísticas sobre preço, produção, emprego, compras, exportações, condições climáticas e muitas outras variáveis importantes para o comportamento econômico ou para tomada de decisão […] A grande massa de estudos econométricos baseia-se nas estatísticas oficiais publicadas por governos, agências internacionais ou mercados seguros. […] A abordagem econométrica está intimamente ligada à teoria matemática da estatística, uma teoria que nos diz como fazer inferência sobre população com base em amostras. […] Os métodos estatísticos usados em econometria seriam como os desenvolvidos em muitos ramos de estudo, apenas com traços diferentes. Os dados estatísticos da econometria são retirados de amostras de observações não experimentais. […] isto é, usamos dados que se originam do resultado real do processo econômico (KLEIN, 1978, p. 13-15).

Os métodos estatísticos utilizados para estudos sobre o crescimento da literatura publicada, segundo Urbizagástegui Alvarado (2010), inauguraram, no final do século XIX, as pesquisas sobre aplicação de métodos métricos na área da informação. A descrição quantitativa dos fenômenos informacionais, que são abordagens para medidas da informação, surgiu conjuntamente com os primeiros estudos em CI e as aplicações estatísticas e matemáticas datam da década de 1920 (RUSSELL; ROUSSEAU, R., 2002 e LE COADIC, 2005).

As áreas métricas da informação (bibliometria, cienciometria, informetria, webometria) se apresentam de maneira geral como um conjunto de técnicas estatísticas e matemáticas para mensurar as atividades informacionais, agregadas ao longo do tempo, da qual a mais antiga é a bibliometria. Isso permite uma abertura para a agregação de novas técnicas e/ou abordagens que possam contribuir com a ampliação da eficiência da mensuração do conhecimento e de sua produção. As reflexões acerca das estratégias econométricas permitem construir um modelo de previsão da produção científica.

As ideias aqui discutidas, para análise e tratamentos de dados, perpassam por um entendimento de uma análise que envolva um número significativo de variáveis que sustentem

uma teoria e também a existência de uma variável dependente, que será estimada, e outra(s) independente(s), que contribui(em) para a previsão (correlação). Lembrando as palavras de Quetelet (1835), o método estatístico possibilita obter, "de conjuntos complexos, representações simples e constatar se essas verificações simplificadas têm relações entre si". Para isso, um dos métodos mais usados, pela econometria, é o dos MQO ou, simplesmente, Mínimos Quadrados, sendo bastante apropriado aos estudos de inferência dos aspectos sociais, comportamentais e econômicos.

Para a econometria, especificar o modelo significa seguir os seguintes passos: exposição da teoria ou hipótese, especificação do modelo matemático e estatístico, definição da origem dos dados, estimação dos parâmetros do modelo econométrico, teste de hipóteses, definição do modelo de projeção ou previsão. Ou, de maneira mais ampla,

> [...] como fazem os econometristas para analisar um problema econômico? Qual metodologia utilizam? Embora existam várias escolas de pensamento sobre metodologia econométrica, aqui apresentamos a **tradicional** ou **clássica**, que ainda domina a pesquisa na economia e em outras ciências sociais e comportamentais. Em termos gerais, a metodologia econométrica tradicional segue os seguintes passos: 1. Exposição da teoria ou hipótese. 2. Especificação do modelo matemático da teoria. 3. Especificação do modelo estatístico ou econométrico. 4. Obtenção dos dados. 5. Estimação dos parâmetros do modelo econométrico. 6. Teste de hipóteses. 7. Projeção ou previsão. 8. Uso do modelo para fins de controle ou de política (GUJARATI; PORTER, 2011, p. 27).

Nesse ponto, os passos da especificação do modelo conduzirão a análise sobre aproximações e/ou distanciamentos entre as abordagens da aplicação dos MQO da econometria e a medição dos modelos métricos da CI. Como resultado, será apresentado um modelo estatístico para contribuir com a previsão da produção científica brasileira.

O objetivo deste capítulo é explicar como a econometria constrói o seu modelo de análise, que é utilizado para pesquisas na economia e, ao mesmo tempo, refletir como esse método pode ser trazido para as metrias da informação. Essa reflexão é necessária porque a econometria aplica eficientemente a estatística matemática a dados econômicos e, ao nosso ver, isso ainda não acontece com as metrias da informação, ou seja, a informetria, a bibliometria, a cientometria, a webometria, entre outros, ainda não conseguem aplicar eficientemente a estatística matemática aos fenômenos informacionais.

3.1 EXPOSIÇÃO DA TEORIA OU HIPÓTESE E ESPECIFICAÇÃO DO MODELO MATEMÁTICO E ESTATÍSTICO

O primeiro passo implica em expor antecipadamente a teoria ou hipótese com a qual se trabalhará a modelagem matemática, dito de outra forma, significa demonstrar previamente as relações de causa e efeito consagradas pela teoria. Por exemplo, se há uma relação direta e crescente entre anos de estudos e salários, há uma teoria ou hipótese subjacente de que quanto mais uma pessoa estuda mais ela ganhará. Ou, se o consumo cresce conforme a renda aumenta, há a Teoria do Consumo Keynesiana, que expõe antecipadamente a hipótese de uma relação direta e crescente entre consumo e renda.

Gujarati e Porter (2011) trazem, como ilustração do primeiro passo da metodologia econométrica (exposição da teoria ou hipótese), a conhecida Teoria do Consumo Keynesiana da seguinte forma:

> Keynes afirmou: A lei psicológica fundamental [...] é que os homens [as mulheres] estão dispostos, como regra e em média, a aumentar seu consumo conforme sua renda aumenta, mas não na mesma proporção que o aumento na renda. Em resumo, Keynes postulava que a **propensão marginal a consumir (PMC)**, a taxa de variação do consumo por variação de uma unidade (digamos, um dólar) de renda, é maior que zero, mas menor que 1 (GUJARATI; PORTER, 2011, p. 27).

Ora, se Keynes postulava que o consumo é função da renda é porque existe a possibilidade de expor antecipadamente a teoria ou hipótese com a qual se trabalhará a modelagem matemática. Ou seja, se na Propensão Marginal a Consumir (PMC) a taxa de variação do consumo por variação de uma unidade de renda está entre zero e um, significa que a teoria econômica consegue demonstrar previamente as relações de causa e efeito.

Em CI, o princípio de Pareto, também conhecido como a regra 80/20, apropriada pela cienciometria, estabelece que para muitos eventos aproximadamente 80% dos efeitos (consequências) são provenientes de 20% das causas. A Lei de Lotka, com a produtividade de autores em periódicos e a também Lei de Bradford, com estudos sobre produtividade de periódicos científicos, são boas opções para estudos que construam teorias ou hipóteses para a produção científica. Mas Urbizagástegui Alvarado (2002) apontou a fragilidade da aplicabilidade da Lei de Lotka e, em seu estudo, fez críticas ao trabalho de Targino e Caldeira (1988), que analisaram a produção científica dos docentes da Universidade Federal de Piauí de 1984 a 1985, explicando que nada concluíram em relação ao ajuste desta literatura produzida à luz da Lei de Lotka, pois apesar de incluírem tabelas específicas da distribuição da

produtividade, fazerem uma citação em cotações e citar Alfred J. Lotka na bibliografia, somente parecem sugerir um ajuste dessa produção à Lei de Lotka.

Urbizagástegui Alvarado (2002) também apontou incongruências no trabalho de Amaral (1996), que analisou a literatura brasileira de *marketing*. Nele a autora afirma que ficou comprovado o princípio da Lei de Lotka, mas a literatura revisada reporta resultados contraditórios que, segundo Urbizagástegui Alvarado, são: a não conformidade dos dados à Lei de Lotka, desvios consideravelmente altos, dados que se ajustam mais claramente à Lei de Price do que à Lei de Lotka e a necessidade de um adequado ajuste à Lei de Lotka.

Por fim, Urbizagástegui Alvarado concluiu, para o trabalho de Amaral (1996), que quando se usa o método dos Mínimos Quadrados e o Teste Kolmogorov-Smirnov para avaliar o ajuste da distribuição teórica à distribuição observada, somente alguns dos dados abordados no estudo se ajustam ao modelo do poder inverso generalizado de Lotka. E, ainda, que a distribuição do poder inverso generalizado de Lotka pelo método de Mínimos Quadrados mostrou um pobre ajuste ao modelo de Lotka.

É necessário lembrar que a relevância aqui é construir um modelo de regressão cuja análise possibilite estudar uma variável (variável dependente) em relação a uma ou mais variáveis (variáveis explanatórias) para confirmar as hipóteses expostas antecipadamente por uma teoria. Com isso, no entendimento deste estudo, dentro das metrias abordadas na revisão de literatura, algo que se possa denominar de "relações prévias de causa e efeito", consagradas pela teoria, estariam apenas nas leis bibliométricas (Zipf, Lotka e Bradford), de modo que inicialmente, em um estudo econométrico, a inclusão da fase de "exposição da hipótese" na proposta do modelo desta tese deve ser cautelosa por entendermos que relações prévias de causa e efeito consagrados pela teoria da informação são insipientes, ou seja, observado os ajustes quando os dados ou as pesquisas se comportarem como uma das leis da bibliometria, poder-se-á usar as suas hipóteses. Mas aqui esse passo não será incluído.

Com relação à especificação do modelo matemático e estatístico, aqui estão dois passos da metodologia econométrica que foram fundidos nesse item: especificação do modelo matemático da teoria e especificação do modelo estatístico ou econométrico. Tais passos implicam em especificar a forma exata da relação funcional (forma da curva) entre as duas variáveis e atribuir uma variável de erro.

Essa especificação exata se chama "Especificação do modelo matemático". O modelo matemático apresenta limites para explicar os fenômenos econômicos e informacionais porque supõe que existe uma relação exata ou determinística entre as variáveis e é sabido que variáveis sociais, econômicas e informacionais são dinâmicas e inexatas. Para resolver esse problema,

deve-se também fazer a "Especificação do modelo estatístico". Dito de outra forma, para levar em conta as relações inexatas entre as variáveis, deve-se modificar a função determinística da equação (forma da curva), acrescentando-se um *u*, que é conhecido como "distúrbio" ou "termo de erro" e representa na função uma variável aleatória (estocástica), que tem propriedades probabilísticas conhecidas.

Nesse ponto, os métodos de medir a informação em CI se restringem em apresentar o modelo matemático, ou em alguns casos nem isso, ignorando erros estocásticos naturais no processo de levantamento de dados. Voltando aos estudos de Urbizagástegui Alvarado, em 2009, seu artigo, ainda criticando trabalhos na área, apresenta um modelo matemático para generalização da Lei de Lotka, usando o modelo do poder inverso generalizado pelo método dos Mínimos Quadrados. E, coadunando-se com a abordagem econométrica, determina alguns passos para o "ajuste" do modelo e para a adoção de um modelo matemático em CI:

$$y_x = C\left(\frac{1}{x^n}\right)$$

Fonte: Urbizagástegui Alvarado (2009)

Na econometria, para a especificação do modelo matemático, Gujarati e Porter (2011), usando como exemplo a Teoria do Consumo Keynesiana, afirmam que, embora Keynes postulasse uma relação positiva entre consumo e renda, ele não especificou a forma exata da relação funcional entre as duas variáveis. Sob o ponto de vista matemático, um economista poderia sugerir a seguinte forma para a função de consumo keynesiana: $Y = \hat{\beta}_1 + \hat{\beta}_2 X$, com $0 < \hat{\beta}_2 < 1$, em que Y = despesas de consumo e X = renda, e $\hat{\beta}_1$ *e* $\hat{\beta}_2$, conhecidos como os "parâmetros do modelo", são, respectivamente, o "intercepto" e o "coeficiente angular".

> O coeficiente angular, $\hat{\beta}_2$, mede a PMC [propensão marginal a consumir] [...]. Essa equação, que especifica que o consumo se relaciona linearmente à renda, é um exemplo de modelo matemático da relação entre consumo e renda e é conhecida como **função consumo** em economia. O modelo é apenas um conjunto de equações matemáticas. Se o modelo tem apenas uma equação, como no apresentado, denomina-se **modelo uniequacional**, enquanto se tiver mais de uma equação será denominado **modelo de múltiplas equações [...].** Na Equação, a variável que aparece do lado esquerdo do sinal de igualdade é chamada de *variável dependente* e a(s) variável(eis) do lado direito é(são) chamada(s) de variável(eis) **independente(s)** ou **explanatória(s)**. Assim, na função consumo keynesiana, o consumo (despesa) é a variável dependente e a renda é a variável explanatória (GUJARATI; PORTER, 2011, p. 27).

Como o modelo matemático não leva em conta as relações inexatas entre as variáveis, deve-se modificar a função determinística da equação (forma da curva) acrescentando-se um *u*. As funções oriundas de MQO são funções de regressão que consideram os resíduos, que é a parte que o modelo não consegue explicar. Então, faz parte das técnicas utilizadas para investigar a adequação de um modelo de regressão a análise dos resíduos.

Ou seja, se o modelo estabelecido for bem ajustado, os resíduos encontrados devem refletir as propriedades dos parâmetros do modelo no sentido de reduzir ao máximo o erro. E isso é confirmado por Gujarati e Porter (2011), quando afirmam que o modelo puramente matemático da função consumo apresentado na equação (que apresenta uma forma linear e sem resíduos) é de interesse limitado para o econometrista, pois supõe que existe uma relação exata ou determinística entre o consumo e a renda (no caso da função consumo keynesiana), mas as relações entre variáveis econômicas são, em geral, inexatas.

> Portanto, se coletarmos dados sobre despesas de consumo e renda disponível (a renda depois de descontados os impostos) de uma amostragem de, digamos, 500 famílias americanas e traçarmos um gráfico em que o eixo vertical representa as despesas de consumo e o eixo horizontal, a renda disponível, não devemos esperar que as 500 observações se situem exatamente na reta dada pela Equação [...]. Isso porque, além da renda, outras variáveis afetam as despesas de consumo. O tamanho da família, a idade de seus integrantes, a religião etc., por exemplo, provavelmente exercem certa influência sobre o consumo. Para levar em conta as relações inexatas entre as variáveis econômicas, o econometrista deve modificar a função consumo determinística da Equação [...] do seguinte modo: $Y = \hat{\beta}_1 + \hat{\beta}_2 X + u$, em que *u*, conhecido como **distúrbio**, ou **termo de erro**, é uma **variável aleatória (estocástica)** que tem propriedades probabilísticas conhecidas. O termo de erro *u* pode representar bem todos esses fatores que afetam o consumo, mas que não são levados em conta explicitamente. A Equação [...] é um exemplo de modelo econométrico [estatístico]. Mais tecnicamente, é um exemplo de **modelo de regressão linear**, [...]. A função consumo econométrica baseia-se na hipótese de que a variável dependente *Y* (o consumo) se relaciona linearmente com a variável explanatória *X* (a renda), mas que a relação entre ambas não é exata: está sujeita a variações individuais (GUJARATI; PORTER, 2011, p. 27).

Gujarati e Porter (2011) destacam que na análise de regressão deve-se entender a dependência estatística entre as variáveis, evitando as relações não funcionais ou determinísticas. Ou seja, a relação entre as variáveis deve ser aleatória ou estocástica e as variáveis devem ter distribuições probabilísticas. A análise de regressão trata da dependência de uma variável em relação as outras, e isso implica em estabelecer uma conexão causal entre duas ou mais variáveis, significa, portanto, construir uma função que estime e preveja relações entre duas ou mais variáveis.

Algumas das metrias da informação abordadas na revisão de literatura possuem formas da curva que permitem alegar que estas possuem um modelo matemático e estatístico, como as

Leis Bibliométricas de Lotka e Bradford, e um pouco menos quanto à Lei Bibliométrica de Zipf, já em relação à Teoria Epidêmica de Goffman, nem isso podemos afirmar. As demais metrias, ao nosso olhar, se valem tão somente de abordagens univariadas com cálculos que se utilizam de médias, medianas, índices, percentuais, desvios e moda para os estudos informacionais.

No presente estudo, o modelo visa estudar o cenário de produção científica brasileira a partir dos valores conhecidos, sendo mantida a fase da "especificação do modelo matemático e estatístico" para a proposta do modelo.

3.2 DEFINIÇÃO DA ORIGEM DOS DADOS (OBTENÇÃO DOS DADOS) E ESTIMAÇÃO DOS PARÂMETROS DO MODELO ECONOMÉTRICO

Este passo implica em obter, coletar os dados do fenômeno social, econômico e informacional ao qual se trabalhará a modelagem matemática e estatística, além disso, é também necessário definir quais destes dados comporão a variável dependente (Y) e quais comporão a variável independente (X).

Esse é um passo importante para aplicação do método porque grande parte dos questionamentos que devem surgir sobre a aplicabilidade desse estudo é o fato de que dados bibliométricos, em sua essência, são discretos e os principais modelos estatísticos tratam valores contínuos. Abordagens estatísticas conseguem desfazer essas barreiras que, para esta pesquisa, contribuem, dentre outros, para que grande parte dos estudos em CI tenha um perfil mais pragmático do que conceitual.

Gujarati e Porter explicam que para estimarmos o modelo econométrico da equação $Y = \hat{\beta}_1 + \hat{\beta}_2 X + u$, isto é, para obtermos os valores numéricos de $\hat{\beta}_1$ e $\hat{\beta}_2$, precisamos de dados. E para demonstrar isso, os autores trazem uma Tabela (p. 30) como exemplo da função de consumo keynesiana. Tal Tabela contém os dados da economia dos Estados Unidos no período 1960-2005.

> Na tabela, a variável *Y* corresponde às despesas de consumo pessoal (DCP) *agregada* (isto é, para a economia como um todo) e a variável *X* ao produto interno bruto (PIB), um indicador de renda agregada, ambas medidas em termos de bilhões de dólares de 2000. Portanto, os dados são apresentados em termos 'reais', isto é, foram medidos a preços constantes (de 2000) (GUJARATI; PORTER, 2011, p. 29).

No caso desta tese, os dados foram coletados a partir do Portal do Plano Tabular do CNPq (http://dgp.cnpq.br/planotabular/), que tem a finalidade de divulgar o perfil da pesquisa no Brasil, em termos quantitativos, organizando esses dados em tabelas configuradas

dinamicamente pelo usuário. Esse sistema permite definir a origem dos dados ao possibilitar o cruzamento de indicadores, capazes de gerar um número muito grande de diferentes tabelas, que podem ser salvas em planilhas ou em arquivos de texto para futuras consultas e que permitem aos órgãos de fomento e planejamento estabelecer metas e prioridades, de modo que pesquisadores em quaisquer das áreas do conhecimento possam aplicar os dados em suas análises, incluindo esta pesquisa.

Quanto aos dados que compõem as variáveis, nesta pesquisa, a variável principal, produção científica brasileira (Yi), foi dividida em suas dimensões representadas neste estudo pela variável independente e dependente. Entendemos aqui, dentro desse aspecto, que o número de doutores cadastrados no Portal Plano Tabular do CNPq (X_i) representa os principais produtores da produção científica nacional (Censo 2010). Para variável dependente foi selecionada a produção bibliográfica de artigos publicados em periódicos nacionais, artigos internacionais, anais de eventos e livros. Estes são os Y_n que, juntos, determinam toda a produção científica brasileira, de todos os estados, em todas as áreas do conhecimento no Plano Tabular do CNPq (Censo 2010).

Em CI, a maior parte dos artigos que tratam sobre métodos métricos da informação buscam entender ou explicar conceitos (viés teórico). Em muitos casos, quando tratam de alguma aplicação prática (normalmente limitadas às leis bibliométicas e ao fator de impacto), apresentam resultados sem estabelecer controles e limites estatísticos. Ao demonstrar resultados são necessários critérios e testes para a possibilidade de estabelecer o método como modelo, como bem já afirmou Urbizagástegui Alvarado (2002 e 2008) em seus estudos para estabelecer a Lei de Lotka.

Bufrem e Prates (2006) construíram um quadro e analisaram 52 artigos de 1980 a 2001, para observar a aplicação práticas das teorias e, segundo as autoras, é certo que definições teóricas auxiliam a compreensão das configurações dos métodos, porém, para melhor entendimento, faz-se necessário tentar associar tais métodos à utilização de aplicações concretas dos termos a eles correspondentes. As autoras construíram esse quadro a partir das relações entre os termos, produzido por Willian McGrath e resumido por Macías-Chapula (1998). Analisando os 52 artigos, pelo menos 16 artigos tratam somente de termos, definições e históricos sobre metrias. Pode se dizer que seis trabalhos aplicam elementos de bibliometria em suas análises como um método e os demais esboçam preocupação em determinar padrões de indicadores, com a recuperação da informação, a importância dos estudos na área e também suas contribuições em pesquisas e resultados de pesquisas.

Aparentemente, após o estabelecimento da bibliometria, os estudos sobre métodos métricos em informação passam a ser tratados muito mais como filosofias do que aplicações empíricas de dados estatísticos. Conceitos filosóficos e sociológicos são trazidos para explicar o objeto informação, algumas vezes em um contexto matemático, mas afastado de alicerces estatísticos. Hayashi (2013), partilhando as palavras de Callon, Courtial e Penan (1993), explica que:

> [...] para alguns, entretanto, a Cientometria permaneceu associada à ciência da ciência e a seu positivismo: as estatísticas e as ferramentas matemáticas representam um papel essencial neste contexto. Para outros, ela se funda em análises mais qualitativas como as que foram desenvolvidas pelas correntes mais recentes da antropologia ou da história social das ciências: as estatísticas não constituem um fim em si mesmo, mas são mobilizadas para analisar a dimensão coletiva da atividade de pesquisa e o processo dinâmico de construção de conhecimentos. […] entendem que essencialmente os cientometristas partilham de três convicções inabaláveis que assegura à disciplina sua coerência: A primeira é que o estudo das ciências e das técnicas passa necessariamente pela análise sistemática das produções 'literárias' dos pesquisadores e dos engenheiros: decerto a Cientometria não se limita exclusivamente a este objeto, mas ela lhe concede um lugar essencial. A segunda é que os estudos quantitativos, desde que eles não constituam um fim em si, enriquecem a compreensão e a descrição da dinâmica das tecnociências. A terceira é a prioridade absoluta e quase que obsessiva admitida na concepção de ferramentas robustas e confiáveis (CALLON; COURTIAL; PENAN, 1993 apud HAYASHI, 2013).

Também autores reconhecidos pela busca de modelos teóricos justificáveis estatisticamente tendem a aproximar seus estudos a um viés mais filosófico. Em 2010, Urbizagástegui Alvarado apresentou um modelo teórico probabilístico da formação da elite de produtores e da frente de pesquisa sobre a Lei de Lotka, a partir de seus estudos anteriores, sob a ótica de Bourdieu e sua representação dos indivíduos através do *habitus* e a noção de campo:

> Pesquisas anteriores propiciaram a identificação dos autores produtores de literatura sobre a Lei de Lotka (Urbizagastegui, 2008) assim como a elite desses produtores (Urbizagastegui, 2009a) e aqueles que integram a chamada frente de pesquisa (Urbizagastegui, 2009b) neste subcampo da bibliometria. Com esses grupos isolados e conhecidos, é possível analisar as características comuns e divergentes de cada um desses integrantes colocados no topo da produção hierarquizada do campo da Bibliometria, subárea da lei de Lotka. Esses grupos serão analisados em relação às características que determinam suas posições hegemônicas ou hegemonizadas dentro do campo, bem como em relação às características de possessão ou des-possessão de um capital cultural e habitus como princípio gerador da prática de produção da literatura sobre a lei de Lotka. Os integrantes de ambos os grupos são conformados pelos mesmos autores, com pequenas variações (URBIZAGÁSTEGUI ALVARADO, 2010).

Dentro desse universo conceitual, Macías-Chapula (1998) apresentou o objetivo de cada área (bibliometria, cienciometria ou cientometria e informetria) a partir dos conceitos de Tague-Sutcliffe.

> A bibliometria é o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. Usada pela primeira vez por Pritchard, em 1969, a bibliometria desenvolve padrões e modelos matemáticos para medir esses processos, usando seus resultados para elaborar previsões e apoiar tomadas de decisão.
>
> Já a cienciometria é o estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. A cienciometria é um segmento da sociologia da ciência, sendo aplicada no desenvolvimento de políticas científicas. Envolve estudos quantitativos das atividades científicas, incluindo a publicação e, portanto, sobrepondo-se à bibliometria.
>
> E a informetria é o estudo dos aspectos quantitativos da informação em qualquer formato, e não apenas registros catalográficos ou bibliografias, referente a qualquer grupo social e não apenas aos cientistas. A informetria pode incorporar, utilizar e ampliar os muitos estudos de avaliação da informação que estão fora dos limites tanto da bibliometria como da cienciometria.

Macías-Chapula (1998) resumiu a tipologia de McGrath (1989) através de um quadro com a definição e classificação dessas três disciplinas, identificando seus objetos de estudo, suas variáveis, seus métodos e objetivos, conforme o quadro a seguir. A partir daí, surgem três perspectivas sob o ponto de vista estatístico: a primeira a respeito da definição das variáveis, a segunda revela os métodos para identificar as estatísticas a serem usadas e a última revela os objetivos a serem alcançados.

Quadro 1 – Classificação das três disciplinas Macías-Chapula (1998) baseado na tipologia de McGrath (1989)

Tipologia	Bibliometria	Cienciometria	Informetria
Objetos de estudo	Livros, documentos, revistas, artigos, autores, usuários	Disciplinas, assuntos, áreas, campos	Palavras, documentos, bases de dados
Variáveis	Número de empréstimos (circulação) e de citações, frequência de citação de palavras, extensão de frases etc.	Fatores que diferenciam as disciplinas e subdisciplinas. Revistas, autores, documentos. Como os cientistas se comunicam.	Difere da cienciometria no propósito das variáveis; por exemplo, medir a recuperação, a relevância, a revocação etc.
Métodos	Ranking, freqüência, distribuição.	Análise de conjunto e de correspondência.	Modelo vetor-espaço, modelos booleanos de recuperação, modelos probabilísticos; linguagem de processamento, abordagens baseadas no conhecimento, tesauros.
Objetivos	Alocar recursos: tempo, dinheiro etc.	Identificar domínios de interesse. Onde os assuntos estão concentrados. Compreender como e quanto os cientistas se comunicam.	Melhorar a eficiência da recuperação.

Fonte: Macías-Chapula (1998)

A estatística define as variáveis em dois tipos: qualitativas, quando seus valores são categorias organizadas de forma ordinal, nominal ou intervalar, e quantitativas, quando são numericamente mensuráveis e subdivididas em contínuas e discretas. Para analisar o Quadro 1, e oportunamente justificar o uso de análise de regressão desta pesquisa, é necessário distinguir entre variáveis contínuas e discretas. Uma variável contínua pode assumir qualquer valor matemático, inclusive o fracionário, dentro da classe da distribuição. Uma variável discreta, entretanto, pode apenas assumir valores que diferem entre si por certas quantidades fixas. Mas Reichmann (1975, p. 144) explica que "há assim uma diferença básica entre as variáveis contínuas e discretas, mas por muitas razões essa diferença pode ser ignorada". O autor justifica que:

> Quando os valores de uma variável contínua são agrupados em classes de frequência, sua distribuição se torna uma distribuição de frequência. As frequências podem apenas ser medidas em termos de números inteiros, pois impossível, por exemplo, para o valor x ocorrer, digamos, um quarto de vezes. Todas as distribuições de frequência são então discretas, apesar do caráter da variável subjacente. Igualmente, as limitações da capacidade de medida, a qual já se fez referência acima, inevitavelmente resultam discretas nas medidas reais registradas, mesmo quando a variável medida é efetivamente contínua (REICHMANN, 1975, p. 144).

Reichmann (1975) segue sua explicação afirmando que:

> Por outro lado, as variáveis discretas podem muitas vezes ser tratadas como se fossem de fato contínuas. Uma das diferenças elementares entre os tipos de variável reside no fato de que os valores de uma variável contínua teoricamente se fundem imperceptivelmente e podem então ser representados graficamente como uma curva regular. Uma variável discreta, porém, muda de valor aos saltos […], mas não se conformam às exigências do modelo matemático da mesma maneira que as curvas regulares, quando se deseja relacionar entre si esses fatos ou a outras variáveis. Por causa disso, os estatísticos apelam para o uso do conceito de aproximação geométrica. A área de um quadrado inscrito em um círculo não se aproximará da área do círculo, mas a área de um polígono de cinco faces se aproximara mais da área do círculo. Quando os polígonos de maior número de faces são inscritos dentro do círculo, veremos que suas áreas se aproximam progressivamente cada vez mais da área do círculo. O processo pode teoricamente ser continuado indefinidamente. Nenhum polígono coincidirá exatamente com a circunferência de um círculo, pois, por definição, um polígono tem lados retos e, portanto, não pode ser circular. No entanto, pode-se conceber um polígono tendo mil faces. Provavelmente não seria possível traçar tal polígono, mas se fosse traçado seria praticamente indistinguível do círculo; tão indistinguível de fato que por motivos práticos seria tratado como se fosse um círculo. Muito desse mesmo princípio e usado para justificar o tratamento de variáveis discretas como se fossem contínuas (REICHMANN, 1975, p. 145).

Segundo as variáveis apresentada por Macías-Chapula (1998), a bibliometria é quem apresenta um maior número de variáveis claramente quantitativas discretas. A cientometria apresenta uma mesclar, mas predominantemente qualitativa nominais. E a informetria descreve suas variáveis mais próximas dos conceitos matemáticos (computação) do que estatísticos, uma vez que, em sua maioria, são utilizadas para estudar as técnicas, metodologias e instrumentos que automatizam processos, e também as técnicas de modelagem de dados e de protocolos de comunicação, formalizando a matemática de algoritmos como forma de representar problemas de decisão a partir de métodos de resolução de problemas baseados em repetições previamente observadas.

Quanto aos métodos de análise de dados e seus objetivos, os bibliométricos e cientométricos são classicamente estatísticos, com destaque para aplicação das análises de conjunto e correspondência por serem técnicas estatísticas multivariadas de caráter exploratório e descritivo, utilizada para a análise de dados categóricos (variáveis qualitativas) e seus resultados oferecem interpretação similar àqueles obtidos pela análise de fatores, utilizada preferencialmente para variáveis contínuas. De acordo com Mingoti (2005), tais métodos são utilizados com o propósito de simplificar ou facilitar a interpretação do fenômeno que está sendo estudado por meio da construção de índices ou variáveis alternativas que sintetizem a informação original dos dados. Os objetivos bibliométricos e cientométricos podem ser

atingidos por estudos estatísticos, sendo que os da bibliometria foram definidos como econômicos e os da cienciometria foram eminentemente como sociais.

Já a informetria apresenta métodos computacionais voltados para a recuperação da informação. Tem aplicação baseada na indexação, armazenamento e busca da informação, utilizando modelos probabilísticos, redes neurais e lógica nebulosa, mas os resultados visam a busca da informação e não análises estatísticas.

A conclusão é que todos os métodos métricos da informação abordados na revisão de literatura possuem dados dos fenômenos informacionais, mas nem todos podem ser trabalhados através de modelagens matemáticas e estatísticas no sentido da análise da regressão, e alguns apenas em uma perspectiva univariada (médias, desvios, moda, mediana, índices, percentuais etc.). Como esta pesquisa visa estudar o cenário de produção científica a partir dos valores conhecidos, é fundamental manter a fase da "obtenção dos dados" para a proposta do modelo.

A estimação dos parâmetros do modelo econométrico. Esse passo tem a tarefa de estimar os parâmetros da função do modelo a partir dos dados coletados. A estimativa numérica dos parâmetros ($\hat{\beta}_1$ *e* $\hat{\beta}_2$) fornece um conteúdo empírico ao modelo matemático e estatístico (Função $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$). Os valores dessas estimativas serão obtidos a partir de uma amostra de n pares de valores (X_i, Y_i), i=1,...,n, que correspondem a n pontos em um gráfico.

O mecanismo para estimar os parâmetros está na própria econometria teórica que trata do desenvolvimento de métodos adequados para medir as relações econômicas especificadas nos modelos econométricos. Logo, a econometria se vale da estatística matemática e para isso, um dos métodos mais usados é o dos MQO ou, simplesmente, Mínimos Quadrados.

> Portanto, a função consumo [Keynesiana] estimada é: $\hat{Y}_i$ = -299,5913 + 0,7218*Xt*. O acento circunflexo em cima do *Y* indica que se trata de uma estimativa. [...] Como indica a Figura I.3 [p.31], a linha de regressão ajusta-se bem aos dados, no sentido de que os pontos no gráfico que representam os dados ficam muito próximos da linha de regressão. A figura nos mostra que, para o período 1960-2005, o coeficiente angular (a **PMC**) era de quase 0,72, indicando que, no período amostrado, um aumento de um dólar na renda real levava, *em média*, a um aumento de cerca de 72 centavos nas despesas reais de consumo. Dizemos *em média* porque a relação entre consumo e renda é inexata [...], nem todos os pontos dos dados estão exatamente sobre a linha de regressão. Em termos simples, podemos dizer que, de acordo com nossos dados, as despesas médias de consumo aumentam cerca de 70 centavos a cada aumento real de um dólar na renda real (GUJARATI; PORTER, 2011, p. 29).

O método dos Mínimos Quadrados escolhe $\hat{\beta}_1$ e $\hat{\beta}_2$ de tal forma que, para qualquer amostra ou conjunto de dados, o $\Sigma\hat{u}_i^2$ será o menor possível. Para uma dada amostra, o método dos Mínimos Quadrados sempre oferece estimativas únicas de $\hat{\beta}_1$ e $\hat{\beta}_2$ que proporcionam o menor valor possível de $\Sigma\hat{u}_i^2$. Esses valores de $\hat{\beta}_1$ e $\hat{\beta}_2$ são conhecidos como estimadores

(parâmetros) de Mínimos Quadrados, por serem derivados do princípio dos Mínimos Quadrados, assim como suas propriedades numéricas dos estimadores obtidos por meio do método de MQO. Para encontrar os valores de $\hat{\beta}_1$ e $\hat{\beta}_2$ devemos seguir o raciocínio em torno do $\Sigma\hat{u}_i^2$, bem como utilizar cálculos diferenciais e derivadas parciais para determinar as equações dos estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$.

O objetivo é estimar os parâmetros $\hat{\beta}_1$ e $\hat{\beta}_2$ de modo que os desvios ($\hat{u}_i$) entre os valores observados e estimados sejam mínimos. Isso equivale a minimizar o comprimento do vetor de erros. Na área da CI, Urbizagástegui Alvarado é um dos únicos, senão o único autor a fazê-lo. Urbizagástegui Alvarado (2008) apresenta essa preocupação para o ajuste através do MQO entre valores observados e estimados na aplicação da Lei de Lotka. Em suas considerações sobre dificuldades de ajuste, aponta outros autores, como Nicholls (1989), cujos resultados dos estudos não são comparáveis devido às diferenças substanciais na forma da medição, estimação dos parâmetros, formas dos testes e ainda devido às interpretações do modelo. Bem como autores como Oppenheim (1986), para quem Urbizagástegui Alvarado afirma que deve se enfatizar que a Lei de Lotka tem sido testada em muitas coleções de dados, porém o ajuste nem sempre tem sido bom.

De certa forma, aos métodos métricos da informação abordados na revisão de literatura podem ser atribuídos dados de fenômenos informacionais que possam ser trabalhados através de modelagens matemáticas e estatísticas, mas nem todos terão uma análise da regressão por motivos já explicados anteriormente. Esta tese busca a análise da regressão dos dados do cenário de produção científica, por isso foi mantida a fase da "estimação dos parâmetros" para a proposta do modelo.

3.3 TESTE DE HIPÓTESES E DEFINIÇÃO DO MODELO DE PROJEÇÃO OU PREVISÃO

Este passo é, *grosso modo*, um prolongamento da estimação dos parâmetros, que, como parte do tratamento dos dados, serve para demonstrar o grau de confiança dos parâmetros do modelo. Às vezes a existência da normalidade dos parâmetros e dos resíduos não é suficiente para que os resultados do ajuste do modelo de regressão sejam tratados como confiáveis, o que significa que a curva linear não é um bom estimador e isso implica em encontrar outra forma da curva. O teste de níveis de confiança dos parâmetros pode ocorrer com o teste T de Student ou o teste F de Fisher-Snedecor, que são usados para verificar a adequação dos parâmetros estimados aos parâmetros populacionais.

> Considerando que o modelo ajustado seja uma aproximação razoavelmente boa da realidade, é preciso desenvolver critérios adequados para verificar se as estimativas obtidas [...] estão de acordo com as expectativas da teoria que está sendo testada. Segundo economistas 'positivos' como Milton Friedman, uma teoria ou hipótese que não for verificável com evidências empíricas pode não ser admissível como parte de uma pesquisa científica. [...] Keynes esperava que a PMC [propensão marginal a consumir] fosse positiva, mas menor que 1. Em nosso exemplo, a PMC é de cerca de 0,72. Entretanto, antes de aceitarmos esse valor como uma confirmação da teoria do consumo keynesiana, precisamos nos perguntar se essa estimativa está suficientemente abaixo da unidade para nos convencer de que não é um resultado devido ao acaso ou uma peculiaridade dos dados que utilizamos. Em outras palavras, 0,72 é *estatisticamente menor que 1*? Se for, será um respaldo para a teoria de Keynes (GUJARATI; PORTER, 2011, p. 31).

Para verificarmos se a estimativa está suficientemente próxima da unidade ($0 < \mathbf{r^2} < 1$) e que não é um resultado devido ao acaso ou uma peculiaridade dos dados, pode-se dispor da análise da regressão e da correlação. Deve ser, então, esclarecida a diferença entre ambos. A análise de correlação é conceitualmente diferente da análise de regressão ao medir a força ou o grau de associação linear entre duas variáveis. O coeficiente de correlação mede a força dessa associação (linear). A análise de regressão busca estimar ou prever o valor médio de uma variável com base nos valores fixos de outras variáveis.

Desses conceitos é estabelecida a análise da correlação que, inclusive, é complementada com a análise da regressão, embora apresentem algumas diferenças. Na análise de regressão, existe uma assimetria na maneira como as variáveis dependentes e explanatória são tratadas. Na análise de correlação são tratadas quaisquer duas variáveis sem levar em conta a distinção entre as variáveis dependente e explanatória. A Teoria da Correlação se baseia na premissa da aleatoriedade das variáveis e a Teoria da Regressão se baseia na premissa de que a variável dependente é estocástica, mas as variáveis explanatórias são fixas ou não estocásticas (GUJARATI; PORTER, 2011). Dentro da análise da regressão se tem ainda os estudos ou análise da regressão bivariada ou, ainda, com duas variáveis. Ela analisa como a variável dependente (regressando) se relaciona com mais de uma variável explanatória (regressores). Sua aplicação é extremamente útil, por apresentar mais de uma ideia ou hipótese fundamentais para a análise de regressão.

No caso da economia, as estimativas quantitativas da Teoria do Consumo Keynesiana proporcionam informações valiosas para a formulação da política econômica. Conhecendo o modelo (função, parâmetros, forma da curva etc.), podemos prever o curso futuro da renda, das despesas de consumo e do emprego após uma alteração da política fiscal do governo. No caso do presente estudo, a análise da regressão dos dados do cenário de produção científica nacional (função produção científica nacional, parâmetros, forma linear ou não linear, análise do r e do

r^2) proporciona explicações para as relações de número de doutores e a variação no número de publicações. Por exemplo, se o coeficiente de regressão (r^2) for o mais próximo de 1, a linha de regressão ajusta-se muito bem aos dados; se o coeficiente de correlação (r) for próximo da unidade (-1 ou 1), isso explica que publicações de artigos e quantidade de doutores têm uma correlação positiva e alta, ou negativa e alta. Outro exemplo é que a análise conjunta do r^2 e r pode revelar um modelo eficiente em prever, ou não, que investimentos na formação de novos doutores aumentam a produção de publicações.

Aqui, dentro das métricas abordadas na revisão de literatura, algo que se possa denominar de verificação da adequação dos parâmetros estimados aos parâmetros populacionais estariam apenas nas leis da bibliometria (Zipf, Lotka e Bradford). Desse modo, deve ser mantida a fase de "teste de hipótese" na proposta de modelo desta pesquisa, mas desmembrado em dois momentos: "análise da correlação", para medir a força ou o grau de associação linear entre duas variáveis com o respectivo mapa de dispersão e decidir se vale a pena regredir o modelo, e a "análise da regressão", para verificar a qualidade do ajustamento através da relação funcional linear ou não linear (forma da curva linear ou não linear), isto é, a melhor adequação possível dos parâmetros estimados aos parâmetros populacionais.

A definição do modelo de projeção ou previsão é o último passo da modelagem econométrica, que é alcançado seguindo-se todos os passos anteriores, ou seja, no caso dos eventos econômicos, o modelo deve ser bem ajustado, as hipóteses não devem ser refutadas e a teoria econômica deve ser considerada dentro do modelo. Nesse sentido, esse passo permite utilizar o modelo para prever o(s) valor(es) futuro(s) da variável previsão Y (variável dependente), com base no(s) valor(es) futuro(s) conhecidos ou esperados da variável previsora X (variável explanatória).

Retome-se para fins de ilustração a Teoria do Consumo Keynesiana, exposta por Gujarati e Porter (2011, p. 32), como exemplo de previsão. Dizem os autores, supondo que desejemos prever as despesas médias de consumo para 2006, com o valor do PIB desse ano de $ 11.319,40 bilhões de dólares. Colocando o valor do PIB no lado direito da equação $\hat{Y}_i$ = - 299,5913 + 0,7218*Xt*, obtemos:

$$\hat{Y}2006 = - 299{,}5913 + 0{,}7218\ (11319{,}4) = 7.870{,}7516$$ ou cerca de $ 7.870 bilhões

Assim, dado o valor do PIB americano, as despesas de consumo médias previstas são de cerca de $ 7.870 bilhões de dólares para o ano de 2006. O valor real das despesas com consumo das famílias americanas para 2006 foi, na verdade, de $ 8.044 bilhões. Portanto,

segundo Gujarati e Porter (2011), esse modelo estimado subestimou as despesas de consumo reais em cerca de 174 bilhões de dólares: foi um erro de previsão de cerca $ 174 bilhões, que representa cerca de 1,5% do valor do PIB dos EUA registrado em 2006.

O importante nesse passo (definição do modelo de projeção ou previsão) é observar o quão importante é a existência de algo que nos permita prever eventos, principalmente para objetivos de política. Outra coisa importante, segundo os autores, é que esses erros de previsão são inevitáveis dada a natureza estatística da análise.

Nesse sentido, na área da CI, Urbizagástegui Alvarado (2010) afirma que os métodos estatísticos utilizados para estudos sobre o crescimento da literatura publicada inauguraram, no final do século XIX, as pesquisas sobre aplicação de métodos métricos na área da informação. E, segundo Russell e Rousseau (2002) e Le Coadic (2005), a descrição quantitativa dos fenômenos informacionais, que são abordagens para medidas da informação, surgiu conjuntamente com os primeiros estudos em CI e as aplicações estatísticas e matemáticas datam da década de 1920.

Tudo isso fez surgir, desde então, as áreas métricas da informação (informetria, cienciometria, webmetria), que se apresentam como um conjunto de técnicas estatísticas e matemáticas para mensurar as atividades informacionais, agregadas ao longo do tempo, da qual a mais antiga é a bibliometria.

Nesta tese, o exemplo da econometria permite uma abertura para a agregação de novas técnicas e/ou abordagens que possam contribuir com a ampliação da eficiência da mensuração das atividades informacionais, tais como o conhecimento e sua produção. Baseado nisso, nesta seção estão expostas as reflexões acerca das estratégias econométricas que podem permitir a construção de um modelo de previsão da produção científica nacional: um modelo que correlacione o número de doutores e a produção científica em cada estado da Federação a partir dos dados do Portal do Plano Tabular do CNPq (2007-2010).

Aqui, o cálculo da função por MQO abordou mais profundamente dois conceitos. O primeiro é o significado do termo "erro estocástico" ($\mathcal{E}$ ou u_i). Gujarati e Porter (2011, p. 65) explicam que um modelo de regressão não é, de forma alguma, uma descrição absolutamente fiel da realidade e, sendo assim, é certo que haverá diferenças entre o valor real do regressando (X_i) e seus valores estimados no modelo escolhido ($\hat{Y}_i$).

O segundo conceito é o da Função de Regressão Amostral (FRA). Esta aborda problemas nas amostras, onde valores de Y correspondentes a alguns valores X-fixados são necessários para estimar a função de regressão. Na próxima seção, ao adentrarmos ao cálculo da função por MQO, abordar-se-á mais aprofundadamente tais conceitos.

Gujarati e Porter (2011, p. 66) explicam que para formular o conceito de Função de Regressão Amostral (FRA), que representa a linha de regressão da amostra inicialmente determinada, a equação correspondente à linear para a amostra é escrita como $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$, onde: Y_i = estimador de $E(Y \mid X_i)$, $\hat{\beta}_1$ = estimador de $ß_1$, $\hat{\beta}_2$ = estimador de $ß_2$ e $\hat{u}_i$ representa o termo residual (na amostra). O $\hat{Y}_i$ é o valor estimado, média condicional, de Y_i, e a FRA primeiro deve expressar os $\hat{u}_i$ (os resíduos) estimados, que são as diferenças entre os valores observados e os valores estimados de Y, conforme $\hat{u}_i = Y_i - \hat{Y}_i$, que implica $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$.

Os dados de *n* pares de observações de Y e X, dentro de um diagrama de dispersão e para determinar a FRA, devem ficar o mais próximo possível do Y-observado. O critério de escolha da FRA deve ocorrer de tal forma que a soma dos resíduos ($\Sigma\hat{u}_i = \Sigma (Y_i - \hat{Y}_i)$) seja a menor possível. A FRA deve ser fixada de tal maneira que seja o mais ajustado possível, quando os $\hat{u}_i^2$ (resíduos elevados ao quadrado) dão mais peso aos resíduos que estão mais afastados da FRA do que os resíduos que estão mais próximos. A diferença está no fato de que quanto maior $\hat{u}_i$, em valores absolutos, maior $\Sigma\hat{u}_i^2$ e maior relevância no somatório, ou seja, se $\sum(\hat{u}_i)^2 = \sum(Y_i - \hat{Y}_i)^2$, então $(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 = \sum(\hat{u}_i)^2$, logo $f(\hat{\beta}_1, \hat{\beta}_2) = \sum\hat{u}_i^2 = > (\hat{u}_i^2 = 0, Y_i = \hat{Y}_i / \sum\hat{Y}_i = f \sum\hat{u}1^2, \hat{u}2^2, \hat{u}3^2 ... \hat{u}n^2)$.

Para os métodos métricos da informação abordadas na revisão de literatura, e já citados, somente aqueles que permitem ser trabalhados através de modelagens matemáticas e estatísticas possuem potencial para modelos de projeção e previsão. Como esta pesquisa visou estudar o cenário de produção científica com base em um modelo de projeção, a fase da "definição do modelo de previsão" é imperativa para a proposta do modelo. Isso, para esta tese, proporciona explicações para as relações de número de doutores e para a variação no número de publicações. Por exemplo, revelar um modelo para prever que investimentos na formação de novos doutores aumentam a produção de publicações.

Outra possibilidade de análise que esse passo (definição do modelo de projeção ou previsão) permite, neste estudo, é a análise da produção científica nacional a partir dos erros estocásticos ($\hat{u}_i^2$), ou seja, $\hat{Y}_i = f(\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2, \ldots, \hat{u}_n^2)$. Os erros ($\hat{u}_n^2$) indicam o quanto um valor Y se aproxima ou se afasta da estimativa esperada $\hat{Y}_i$. Com essa análise é possível, por exemplo, estabelecer um *ranking* dos estados brasileiros que apresentam maior eficiência na produção de publicações. Dito de outra forma, os estados que apresentam $\hat{u}_i^2$ menores se aproximam mais de uma combinação ótima entre número de doutores e publicações.

3.4 USO DO MODELO PARA FINS DE CONTROLE E POLÍTICA: CONSIDERAÇÕES

A informetria, a bibliometria, a cientometria, a webometria, entre outros, se apresenta de maneira geral como um conjunto de técnicas estatísticas e matemáticas para mensurar as atividades informacionais. Assim, como uma espécie de campo para as metrias da informação, permite a agregação de novas técnicas e/ou abordagens que possam contribuir com a ampliação da eficiência da mensuração do conhecimento e sua produção.

Com esse intuito, pode-se utilizar na área da CI o modelo calculado por meio de MQO para fins de controle ou política, e também como forma preditiva de estudos dos fluxos informacionais, bem como estudar modelos de previsão da produção científica nacional para fins de controle e políticas dos recursos públicos destinados à pesquisa, que é o que está conexo com esta tese.

Empiricamente, os métodos estatísticos de análise de dados a partir de conceitos e teorias da disciplina econometria estão mais consolidados no contexto de bagagem metodológica, pois a econometria enseja metodologias bastante consistentes para os estudos de dados econômicos (produção, consumo, distribuição, demanda, oferta, juros, renda nacional, balança internacional etc.) e, com isso, pode contribuir sobremaneira para as metrias da informação dentro deste enfoque.

O método dos MQO é uma técnica de otimização estatística utilizada para encontrar o melhor ajuste para um conjunto de dados. Essa técnica consiste em minimizar a soma dos quadrados das diferenças (resíduos) entre o valor estimado e os dados observados (reais). É amplamente utilizada por diversas áreas por permitir, através da minimização da Soma dos Quadrados dos Resíduos da regressão, maximizar o grau de ajuste do modelo aos dados observados.

O MQO é um método amplamente utilizado em diversas áreas (engenharia, medicina, biologia etc.), aqui aplicado segundo a experiência da economia e sua ciência auxiliar, a econometria. É um método alternativo aos tradicionais (bibliometria, econometria, informetria etc.) que, além de demonstrar a relação entre valores, permite determinar modelos de previsão. Os modelos criados trazem a força da relação entre variáveis, permitem simular resultados a partir de variáveis e determinam a precisão desses valores.

A visão econométrica das aplicações do MQO foi escolhida por trazer modelos de análises como métodos de interpretação de teorias econômicas. A econometria vai além da utilização da estatística como mera ferramenta. Os modelos criados fazem parte do corpo de conhecimento e estudos econômicos. Apresenta uma relação entre dados, métodos e teoria, o

que aparentemente não acontece em outras áreas que aplicavam o MQO como ferramenta estatística sem uma relação teórica com a área. Empiricamente, os métodos de análise de dados, que utilizam conceitos e teorias estatísticas, estão mais consolidados na disciplina de econometria (aplicabilidade em pesquisas de análises quantitativas) do que às da área que medem a informação.

4 MÍNIMOS QUADRADOS ORDINÁRIOS (MQO)

A informetria se apresenta de maneira geral como um conjunto de técnicas estatísticas e matemáticas para mensurar as atividades informacionais, agregadas ao longo do tempo, da qual a mais antiga é a bibliometria. Assim, como uma espécie de campo para as metrias da informação, permite a agregação de novas técnicas e/ou abordagens que possam contribuir com a ampliação da eficiência da mensuração do conhecimento e sua produção.

Com esse intuito, pode-se utilizar, na área da CI, a regressão (linear ou não linear), que também permite criar uma forma preditiva de estudos dos fluxos informacionais, bem como estudar modelos de previsão da produção científica nacional, que é o que está conexo com esta tese. Empiricamente, os métodos estatísticos de análise de dados a partir de conceitos e teorias da disciplina econometria estão mais consolidados no contexto de bagagem metodológica, pois a econometria enseja metodologias bastante consistentes para os estudos de dados econômicos (produção, consumo, distribuição, demanda, oferta, juros, renda nacional, balança internacional etc.) e, com isso, busca auxiliar as metrias da informação dentro deste enfoque.

A pesquisa econométrica visa, portanto, prover medições concretas para a teoria econômica, usando a teoria e técnica da inferência estatística como ponte. O próprio autor esclarece que a econometria teórica trata do desenvolvimento de métodos adequados para medir as relações econômicas especificadas nos modelos econométricos. Logo, a econometria é dependente da estatística matemática. Para isso, um dos métodos mais usados é o dos MQO.

O modelo de regressão através de MQO é tradicional na área da econometria por ser bastante apropriado aos estudos de inferência dos aspectos sociais, comportamentais e econômicos. No caminho de especificar o modelo, são necessários os seguintes passos: exposição da hipótese, especificação do modelo matemático e estatístico, definição da origem dos dados, estimação dos parâmetros do modelo econométrico, teste de hipóteses, definição do modelo de projeção ou previsão.

Com isso, é construído o modelo de regressão cuja análise vai estudar a dependência de uma variável (variável dependente) em relação a uma ou mais variáveis (variáveis explanatórias) para estimar, através de regressão e correlação, valores para o cenário de produção científica nacional a partir dos valores conhecidos. O termo “regressão”, segundo Gujarati e Porter (2011, p. 39), que auxiliará nosso percurso, foi criado por Francis Galton em seus estudos sobre uma tendência de que pais altos tivessem filhos altos e pais baixos tivessem filhos baixos. A estatura média das crianças nascidas de pais com uma dada altura tendia a se

mover ou “regredir” para a altura média da população como um todo. A Lei da Regressão Universal de Galton foi confirmada posteriormente por Karl Pearson.

Um ponto importante destacado por Gujarati e Porter (2011, p. 39) na análise de regressão é entender a dependência estatística entre as variáveis, evitando as relações não funcionais ou determinísticas. Ou seja, a relação entre as variáveis deve ser aleatória ou estocástica e as variáveis devem ter distribuições probabilísticas. A análise de regressão trata da dependência de uma variável em relação as outras, e isso implica em estabelecer uma conexão causal entre duas ou mais variáveis. Significa, portanto, construir uma função que estime e preveja relações entre duas ou mais variáveis.

Também deve ser esclarecida a diferença entre regressão e correlação. A análise de correlação é conceitualmente diferente da análise de regressão ao medir a força ou o grau de associação linear entre duas variáveis. Ainda como parte do tratamento dos dados existem os testes de hipóteses, que servem para adequar o problema às suposições do modelo. Às vezes, verificar a existência da normalidade dos parâmetros e dos resíduos não é suficiente para que os resultados do ajuste do modelo de regressão linear sejam tratados como confiável, o que implica em encontrar a melhor forma da curva. Outra questão são os níveis de confiança dos parâmetros. Nesse caso, o teste T de Student ou o teste F de Fischer-Snedcor podem ser usados para verificar a adequação dos parâmetros estimados aos parâmetros populacionais.

O objetivo desta seção é explicar o método de estimação por MQO para a análise de regressão a que se propõe esta tese. Seguem, nas subseções, os elementos considerados nesta pesquisa como importantes na construção do entendimento e aplicação dos conceitos e equações necessárias aos fundamentos do MQO. As explicações que se seguem foram obtidas a partir de observações e orientações didáticas dos trabalhos de Rocha (1975), Fonseca, Martins e Toledo (1976), Ravichandra Rao (1986), Sanz Casado (1994), Martins (2001) e Gujarati e Porter (2011), com informações e fórmulas disponíveis. Os trabalhos analisados tinham por objetivo orientar não estatísticos a utilizar ferramentas estatísticas para desenvolver pesquisas e observações dadas em pesquisas em qualquer área do conhecimento.

4.1 MEDIDAS DE TENDÊNCIA CENTRAL (MÉDIA E MEDIANA) E DE DISPERSÃO (VARIÂNCIA E DESVIO PADRÃO)

A apresentação de informações numéricas por meio de gráficos e tabelas é um método relativamente comum no que se refere à visualização de dados e serve para antecipar algumas

interpretações e análises. Esse aprofundamento interpretativo dos dados se inicia com uso de duas técnicas estatísticas, as observações sobre as medidas de tendência central, e consequentemente, as análises das medidas de dispersão. Na econometria, os princípios do uso do valor médio e do desvio-padrão para descrever o comportamento na regressão partem do interesse em saber se existe alguma força causal que a afeta. Caso exista, poderia ser determinado um melhor prognóstico desse valor médio com a análise de regressão, lembrando que os modelos econométricos são quase sempre desenvolvidos para testar uma ou mais teorias econômicas (GUJARATI, PORTER, 2011).

As medidas de tendência central, também conhecidas como "medidas de posição", consistem em encontrar um único número que represente o valor típico de um conjunto de dados e, com esse levantamento, é possível encontrar valores que representam e caracterizam um determinado conjunto de dados. Já as medidas de dispersão servem para verificar como os valores de um conjunto de dados se comportam em relação à dispersão e o quão distante essas podem estar. Ravichandra Rao (1986) já apontava que as medidas de tendência central complementadas com medidas de dispersão podem contribuir melhor para análises e estudos de conjuntos de dados, e isso possibilita melhores observações interpretativas acerca das metrias da informação.

As ferramentas mais utilizadas para a análise das dispersões nessa tarefa são denominadas de "variância" e "desvio padrão". Sanz Casado (1994) explica que as medidas de dispersão ajudam a conhecer a viabilidade que apresentam um grupo de observações e decidir se essas observações são muito parecidas ou distintas entre si. Lembra também que as medidas de dispersão completam as informações obtidas com as medidas de tendência central e que os desvios expressam as diferenças entre o valor médio e cada um dos valores no conjunto de dados.

Aqui analisamos o exemplo da produção brasileira de artigos, em específico os periódicos nacionais e estrangeiros, a partir das medidas de tendência central (média aritmética, média geométrica e mediana) e medidas de dispersão (variância, desvio médio, desvio padrão e coeficiente de variação). A auxiliar na explanação e exemplificação do assunto foi trazida na Tabela 1, com dados parciais do Plano Tabular, destacada para a presente análise, contendo dados da produção científica brasileira de artigos completos publicados em periódicos especializados. As colunas de "pesquisadores doutores" e "artigos completos publicados em periódicos especializados" foram construídas utilizando informações da tabela de produção bibliográfica elaborada segundo unidades da Federação para pesquisadores doutores, 2007-

2010, Censo 2010, cujos dados foram coletados do Plano Tabular do CNPq, que inclusive são também utilizados como base para o relatório do MCT.

Tabela 1 – Produção bibliográfica segundo UF para pesquisadores doutores, 2007-2010, Censo 2010

UF	pesquisadores doutores	Artigos completos publicados em periódicos especializados		Índices para artigos completos publicados em periódicos especializados	
		Circulação nacional	Circulação internacional	Circulação nacional	Circulação internacional
Acre	162	387	206	2,389	1,272
Alagoas	760	2.317	1.683	3,049	2,214
Amapá	65	182	193	2,800	2,969
Amazonas	1.112	3.017	3.723	2,713	3,348
Bahia	3.622	12.121	9.674	3,346	2,671
Ceará	1.975	8.980	7.556	4,547	3,826
Distrito Federal	2.686	10.814	8.683	4,026	3,233
Espírito Santo	979	3.880	2.486	3,963	2,539
Goiás	1.775	7.454	5.419	4,199	3,053
Maranhão	593	2.506	2.058	4,226	3,470
Mato Grosso	1.075	4.595	2.157	4,274	2,007
Mato Grosso do Sul	1.497	6.686	3.894	4,466	2,601
Minas Gerais	9.228	41.159	34.692	4,460	3,759
Pará	1.462	4.701	4.485	3,215	3,068
Paraíba	2.055	9.347	5.206	4,548	2,533
Paraná	6.508	28.586	21.122	4,392	3,246
Pernambuco	3.215	12.731	9.673	3,960	3,009
Piauí	626	2.846	1.672	4,546	2,671
Rio de Janeiro	10.997	36.693	42.933	3,337	3,904
Rio Grande do Norte	1.527	5.775	3.930	3,782	2,574
Rio Grande do Sul	7.841	36.627	30.596	4,671	3,902
Rondônia	221	705	723	3,190	3,271
Roraima	169	722	322	4,272	1,905
Santa Catarina	3.580	14.472	11.203	4,042	3,129
São Paulo	22.922	90.240	108.990	3,937	4,755
Sergipe	824	2.953	2.088	3,584	2,534
Tocantins	358	1.675	809	4,679	2,260

Fonte: Elaboração do autor

Na descrição ou na interpretação dos fenômenos estatísticos recorre-se, com frequência, à comparação de dois ou mais dados. Rocha (1975, p. 40) explica que a apreciação de um dado em relação a outra grandeza pode ser estabelecida de duas formas: por diferença ou por quociente. Para a aplicação das medidas de posição e de dispersão, foi utilizado quociente.

Rocha (1975, p. 48) também lembra que "habitualmente adotam-se definições especiais para algumas formas de comparação por quociente", índice para duas grandezas onde uma está incluída na outra (índice cefálico, densidade demográfica, quociente intelectual), e coeficiente ou taxa para quociente entre o número de ocorrências e a soma dos números de ocorrências com o de não ocorrência (taxa de natalidade, coeficiente de aprovação). Nessa seção, foi utilizado o índice *per capita* de produção científica.

A coluna "índice para artigos completos publicados em periódicos especializados" são índices que representam a produção científica a partir dos dados das colunas pesquisadores doutores e artigos completos publicados em periódicos especializados. Como explicado, matematicamente se utiliza a razão entre duas grandezas para fazer comparações entre variáveis. O índice da Figura 1 foi estabelecido para representar a razão entre a produção de artigos (periódicos nacionais e estrangeiros) e o total de doutores dentro do cenário do nacional por estados.

Figura 1 – Índice *per capita*

$$\text{Índice de produção} = \frac{\text{Produção de Artigos}}{\text{Número de doutores}}$$

Fonte: Santos (2011)

A análise dos índices construídos na Tabela 1 gerou a Tabela 2, com os resultados para as medidas de tendência central e medidas de dispersão.

Tabela 2 – Medidas de tendência central e medidas de dispersão para a Tabela 1

Artigos completos publicados em periódicos especializados	Máximos	Mínimos	Média	Média Geométrica	Mediana	Desvio Médio	Variância	Desvio Padrão
Circulação nacional	4,679	2,389	3,875	3,815	4,026	0,544	0,430	0,656
Circulação internacional	4,755	1,272	2,953	2,858	3,009	0,567	0,539	0,734

Fonte: Elaboração do autor

A produção nacional apresentou uma produção máxima de artigos, tanto em periódicos nacionais quanto em estrangeiros, de aproximadamente cinco por estado, com Tocantins (4,679) liderando a produção de artigos nacionais e São Paulo (4,755) os internacionais. Já a produção mínima (Acre) apresenta aproximadamente dois artigos em periódicos nacionais e

apenas um para estrangeiros. Os dados da média, ou média aritmética simples, ajudam a entender os valores centrais entre valores extremos. Para as 27 observações, os artigos nacionais apresentam uma média de aproximadamente quatro artigos nacionais por estado e três artigos internacionais por estado.

Levando em consideração as possíveis múltiplas propriedades numéricas da razão entre número de publicações e o número de doutores, foi aplicada outra medida de tendência central, a média geométrica. Os valores geometricamente médios dos artigos (3,815 para artigos nacionais e 2,858 para internacionais) são bastante próximos aos das médias aritméticas (3,875 para artigos nacionais e 2,953 para internacionais). O resultado aponta para uma diferença pouco significativa entre os valores, na casa dos decimais.

Enquanto a média é o resultado do somatório de um conjunto de dados dividido pelo total de elementos das observações (x_i), aplicando a mediana obtém-se o valor central de um conjunto de dados ordenado. A mediana foi utilizada para determinar o valor intermediário que separa a metade superior da metade inferior do conjunto de dados. Isso porque os valores observados na parte superior da mediana têm a mesma frequência que os valores inferiores e isso serve para excluir *outliers*, que são valores discrepantes que alterariam o entendimento no caso de aplicar média.

As medianas da Tabela 2 apontam para o Distrito Federal como a mediana com 4,026 artigos para publicações nacionais e Pernambuco como a mediana dos artigos internacionais, com 3,009 publicados. Assim, os 13 estados com valores abaixo da mediana para artigos nacionais são: Acre, Amazonas, Amapá, Alagoas, Rondônia, Pará, Rio de Janeiro, Bahia, Sergipe, Rio Grande do Norte, São Paulo, Pernambuco e Espírito Santo; e os estados abaixo da mediana para artigos internacionais são: Acre, Roraima, Mato Grosso, Alagoas, Tocantins, Paraíba, Sergipe, Espírito Santo, Rio Grande do Norte, Mato Grosso do Sul, Bahia, Piauí e Amapá. Os 13 estados acima da mediana para artigos nacionais são: Santa Catarina, Goiás, Maranhão, Roraima, Mato Grosso, Paraná, Minas Gerais, Mato Grosso do Sul, Piauí, Ceará, Paraíba, Rio Grande do Sul e Tocantins; e para os artigos internacionais acima da mediana são: Goiás, Pará, Santa Catarina, Distrito Federal, Paraná, Rondônia, Amazonas, Maranhão, Minas Gerais, Ceará, Rio Grande do Sul, Rio de Janeiro e São Paulo.

Quanto à aplicação das técnicas para obter as medidas de dispersão, o objetivo foi verificar como esses dados, os estados, se comportavam em relação à dispersão e o quão distante esses estados estavam da média. As ferramentas utilizadas nessa tarefa foram o desvio médio, a variância e o desvio padrão. Sanz Casado (1994, p. 140) explica que as medidas de dispersão ajudam a conhecer a viabilidade que apresentam um grupo de observações e decidir se essas

observações são muito parecidas ou distintas entre si, além de que as medidas de dispersão completam as informações obtidas com as medidas de tendência central.

Equação 2 – Desvio médio ou absoluto

$$\frac{\sum|x_i - \bar{x}|}{n}$$

Fonte: Sanz Casado (1994)

A primeira medida de dispersão é o desvio médio, ou absoluto, calculado pela Equação 2, que expressa a média dos desvios em termos absolutos, apontando um valor de 0,544 para uma média aritmética de 3,875 para publicações nacionais e um desvio médio de 0,567 para uma média de 2,953 para as publicações no estrangeiro.

A variância, segunda medida de dispersão, foi utilizada para verificar como cada medida estava distante do valor central (médio). Aqui os valores de variância apresentados foram de 0,430 para publicações nacionais e 0,539 para estrangeiras. A variância é calculada elevando ao quadrado a diferença entre a média aritmética e o valor de cada observação (x_i), ou seja, o desvio em torno da média. Mas ao elevar todas as medidas ao quadrado, é obtida uma visão distorcida para comparação do quanto as medidas se afastam da sua média aritmética. Para corrigir isso, se utiliza a raiz quadrada da variância e, assim, se obtém outra importante medida de dispersão, o desvio padrão.

Equação 3 – Variância da amostra

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Fonte: Sanz Casado (1994)

A variância foi aqui aplicada para entender a distribuição, uma vez que, quanto a variância menor indica que mais próximos os valores estão da média, e quanto maior é esse valor, mais os valores estão distantes da média.

Equação 4 – Desvio padrão da amostra

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Fonte: Sanz Casado (1994)

O desvio padrão, já citado, é a raiz quadrada da variância e identifica o "erro" em um conjunto de dados. Ele foi utilizado para determinar melhor o grau de dispersão (variação dos valores) observado em torno da média da distribuição. Aqui os desvios padrão são de 0,656 para artigos nacionais e de 0,734 para artigos publicados fora do Brasil.

A análise do desvio médio aponta que a variabilidade de todos os estados para publicações nacionais, mesmo sendo menor do que a variabilidade das publicações no estrangeiro, é relativamente próxima, ou seja, tem desvios similares.

Para ajudar a entender um pouco mais esses dados, outro cálculo útil é o coeficiente de variação, que é uma medida utilizada para saber o tamanho do desvio padrão em relação à média, expressa em porcentagem. Por não possuir unidades, o coeficiente de variação pode ser usado no lugar do desvio padrão para comparar a dispersão dos dados em conjuntos com unidades ou médias diferentes.

Equação 5 – Coeficiente de variação

$$\text{coeficiente de variação} = \frac{s}{\bar{x}}$$

$$\text{ou, coeficiente de variação em \%} = \left(\frac{s}{\bar{x}}\right) \cdot 100$$

Fonte: Sanz Casado (1994)

Para o cálculo do coeficiente de variação, é necessária a determinação prévia dos valores da média, do desvio médio, da variância e do desvio padrão, para então conhecer o quanto, percentualmente, a produção de artigos brasileiros é eficiente.

Isso é determinado pela diminuição do coeficiente de variação, o que representa um aumento da produção de artigos (nacionais e internacionais) em todos os estados, levando a um aumento geral dessa produção. Os coeficientes de variação encontrados são 17% para artigos nacionais e 25% para artigos publicados fora do Brasil. Isso explica que, para publicações nacionais, os estados são mais homogêneos.

Analisando individualmente os dados de números de doutores e os de números de publicações, fica claro que essas são variáveis aleatórias discretas, mas relacionando as duas na forma do índice *per capita*, e criando uma nova variável bidimensional, elas agora apresentam um comportamento aleatório contínuo.

Em estatística, essas diferenças entre os perfis das variáveis levam a análises e técnicas distintas. E, nesse sentido, foram mantidos os valores decimais representando a produção em cada estado.

Mas qual a escolha da medida de tendência central mais adequada? Com a ajuda de Rocha (1975, p. 149), e extrapolando para dados de produção científica, a escolha da medida depende dos objetivos fixados e, também, da definição e das propriedades da síntese.

Assim, a cada um dos intuitos seguintes corresponde um tipo de medida: a) verificar a suficiência (ou insuficiência) de valores totais, em comparação com o mínimo indispensável determinado pelas experiências. Nesse caso, basta somar a produção de cada estado e dividir pelo número de observações (27), o que pode ser satisfeito pela média aritmética; b) identificar o padrão de produtividade dominante. Nesse caso, basta determinar quais os valores de produção se encontra com maior frequência, utilizando moda (outra medida de tendência central) ou medidas de dispersão; c) pesquisar o nível de produção do estado alcançado ou exercido em cada fração dos produtores (por exemplo, a metade deles). Pode ser usado o valor mediano; d) verificar qual produtividade caberia a cada estado se o total fosse distribuído em partes iguais, em vez partes desiguais, o que também pode ser satisfeito pela média aritmética.

> As propriedades das sínteses devem ser consideradas, pois podem, ou não, estar de acordo com o objetivo da nossa indagação; além disso uma propriedade terá maior ou menor grau de adequação com a realidade, no sentido de que certas características do fato observado tenham comportamento compatível com a referida propriedade. […] Assim, pode-se concluir que, não há um motivo geral de preferência para uma ou outra medida de tendência central, já que conforme o fim da aplicação, uma delas poderá ser preferível (ROCHA, 1975, p. 149).

Ao relacionar duas ou mais variáveis, será apresentado outro olhar para os mesmos dados. Essa nova variável bidimensional é relativamente mais coerente para determinar medidas de tendência central e de dispersão.

O impacto do número de doutores deve ser considerado numa análise de produtividade para cada estado. Mas, finalmente, as medidas de tendências centrais e dos desvios são bons ou maus indicadores?

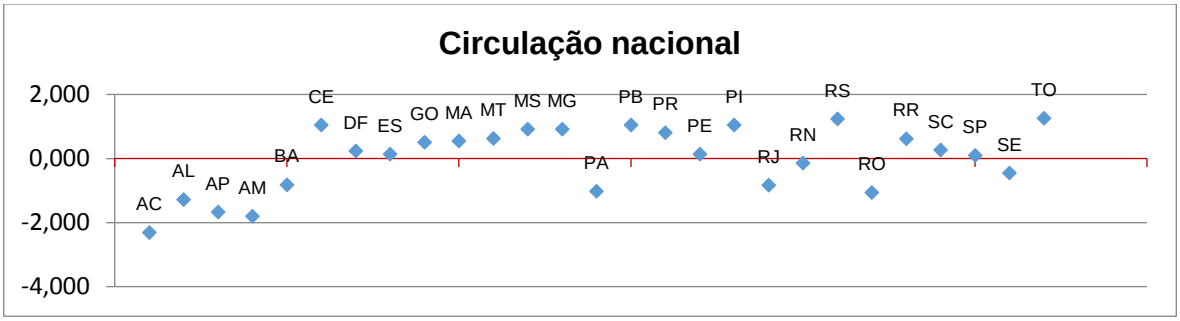
Bem, a aplicação desses dados deve constar dentro de arcabouço maior, para subsidiar a tomada de decisão dos interessados em aprimorar a produção científica nacional, pelo menos

em análises preliminares. Outro ponto é reforçar a necessidade de apropriação por parte das pesquisas em CI do uso de ferramentas estatísticas dentro de princípios metodológicos interdisciplinares, não só como ferramentas pontuais.

Outra forma de aplicar essas medidas para fazer análises é utilizar a padronização z escore, que informa a distância que, em desvios padrões, o valor está da média. Esse é um dado muito semelhante à distância da reta de regressão, erro (u_i), que será apresentada na próxima seção. A partir dessa informação, é possível visualizar o panorama do comportamento da variável.

A Figura 2 apresenta para publicações de artigos em periódicos nacionais as distâncias (quantidade de desvios padrões), de cada estado, em relação à média, valor zero no eixo y (linha vermelha). Quanto mais próximos da linha (zero), mais próximos da média. Valores abaixo da linha são abaixo da média e acima significa que também se está acima da média, com destaques para o Acre, o mais abaixo, e para o Tocantins, o mais acima.

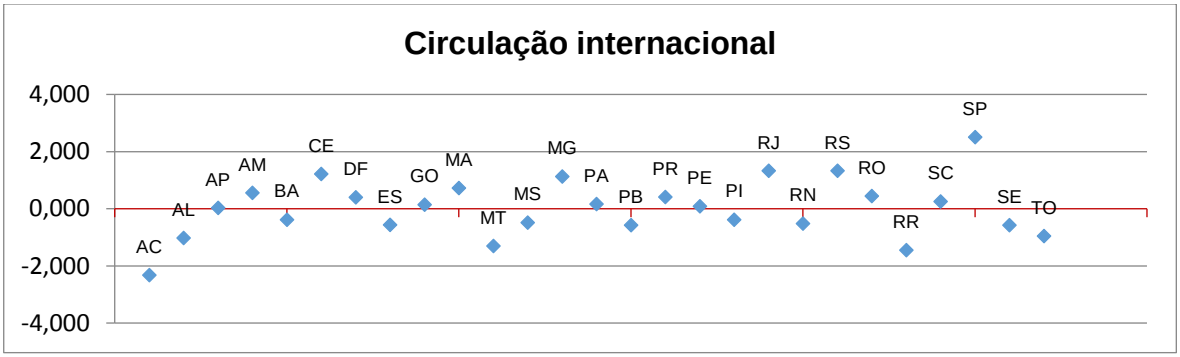
Figura 2 – Distâncias em desvio padrão dos estados em relação à média para artigos em publicações nacionais



Fonte: Elaboração do autor

A Figura 3 se apresenta para publicações de artigos em periódicos estrangeiros. Destaques para o Acre, o mais abaixo, e São Paulo, o mais acima.

Figura 3 – Distâncias em desvio padrão dos estados em relação à média para artigos em publicações estrangeiras



Fonte: Elaboração do autor

4.2 CORRELAÇÃO E REGRESSÃO: O R^2, UM EXEMPLO DA ECONOMETRIA

O termo correlação significa relação em dois sentidos (co + relação) e é usado em estatística para designar a força que mantém unidos dois conjuntos de valores (MARTINS, 2001). O estudo da correlação objetiva a verificação da existência e do grau de relação entre as variáveis. Tal estudo da correlação, quando encontrado certo grau de relacionamento verificado pela relação entre duas variáveis, viabiliza a construção de uma função de regressão, que é um modelo matemático que permite descrever e analisar a relação estimada partindo-se de observações reais através de uma função.

Na função de regressão, dentro do plano cartesiano, as variáveis são divididas em variável dependente, eixo y, e variável independente, eixo x. A correlação fornece uma medida perceptível da força, ou grau, de relacionamento entre essas duas variáveis, e isso indica se vale a pena construir a função de regressão e despender esforços para a análise e descrição, em termos matemáticos, do fenômeno observado.

O coeficiente de correlação (ρx, y) entre duas variáveis aleatórias X e Y com valores esperados μ_X e μ_Y e desvios padrão σ_X e σ_Y é definida como:

Equação 6 – Coeficiente de correlação

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y}$$

Fonte: Fonseca, Martins e Toledo (1976)

"E" é o valor esperado e "cov" significa "covariância". Como $\mu_X = E(X)$, $\sigma_X^2 = E(X^2) - E^2(X)$ e $\mu_y = E(Y)$, $\sigma y^2 = E(Y^2) - E^2(Y)$, a equação pode ser reescrita como:

Equação 7 – Coeficiente de correlação reescrito

$$\rho_{X,Y} = \frac{E(X,Y) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

Fonte: Fonseca, Martins e Toledo (1976)

O numerador de "$\rho_{X,Y}$" é dado pela covariância entre as duas variáveis aleatórias X e Y, ou seja, covariância ou variância conjunta é uma medida de como duas variáveis variam conjuntamente. A covariância de X e Y é dada pela equação:

Equação 8 – Covariância de X e Y

$$\mathrm{cov}(X,Y) = \sigma_{xy} = \sum_{i=1}^{n} [(x_i - \mu_i^x)(y_i - \mu_i^y)\rho(x_i, y_i)]$$

Fonte: Fonseca, Martins e Toledo (1976)

Onde "$p(x_i,y_i)$" é a frequência relativa (ou probabilidade de ocorrer o par (x_i,y_i)). Essa equação pode ser também escrita da seguinte forma:

Equação 9 – Covariância de X e Y reescrita

$$\mathrm{cov}(X,Y) = \sigma_{xy} = \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right]$$

Fonte: Fonseca, Martins e Toledo (1976)

Equação 10 – Coeficiente de correlação linear (r)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2)\,(\sum(y_i - \bar{y})^2)}}$$

Fonte: Fonseca, Martins e Toledo (1976)

Por exemplo, nas observações da relação entre a variável "número de doutores" com as demais variáveis indicativas de produção científica (artigos publicados em periódicos nacionais, artigos publicados em periódicos internacionais, trabalhos completos publicados em anais de eventos, livros publicados, capítulos de livros publicados e outras publicações bibliográficas) propostas nesta tese, pode ser utilizado o coeficiente de correlação linear de Pearson (ρ de Pearson) para medir o grau da correlação e mostrar a direção (positiva ou negativa). O coeficiente de correlação linear de Pearson "ρ" é obtido através da seguinte equação:

Equação 11 – Coeficiente de correlação linear de Pearson (ρ)

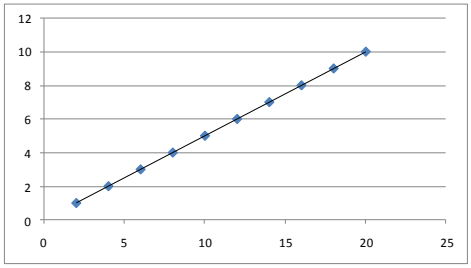
$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{cov(X,Y)}{\sqrt{var(X) \cdot var(Y)}}$$

Fonte: Fonseca, Martins e Toledo (1976)

Uma das aplicações possíveis na área da informetria é a análise para comprovar a existência de relação entre o número de doutores e as produções científicas, isto é, saber se as variações quantitativas presentes em uma das variáveis são acompanhadas por alterações nas outras. Por exemplo, variações do número de doutores em relação às variações na produção de artigos em revistas nacionais ou em relação às variações na produção de artigos em revistas internacionais ou, ainda, produção de livros etc.

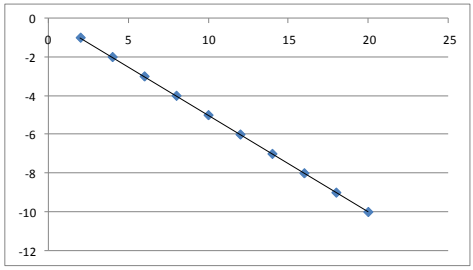
Os diagramas de dispersão abaixo demonstram que a correlação será tanto mais forte quanto mais próximo estiver o coeficiente de -1 ou +1, e será tanto mais fraca quanto mais próximo o coeficiente estiver de zero. O coeficiente linear de Pearson é o mais utilizado para se estudar a correlação entre duas variáveis, normalmente é representado por "ρ" e assume valores entre -1 e 1. Quando o ρ = 1, significa uma correlação perfeita positiva entre as duas variáveis. Quando o ρ = -1, significa uma correlação negativa perfeita entre as duas variáveis. Isto é, se uma variável aumenta, a outra diminui e vice-versa. Caso o ρ = 0, significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma dependência não linear. Assim, o resultado ρ = 0 deve ser investigado por outros meios.

Figura 4 – Correlação positiva r > 0



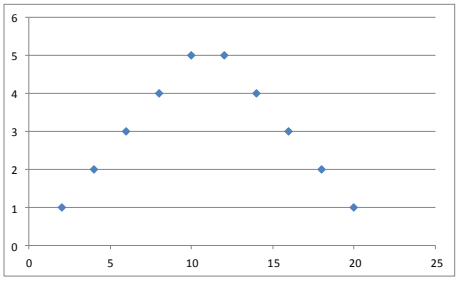
Fonte: Fonseca, Martins e Toledo (1976)

Figura 5 – Correlação negativa r < 0



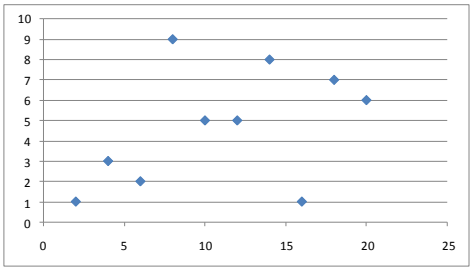
Fonte: Fonseca, Martins e Toledo (1976)

Figura 6 – Correlação não linear



Fonte: Fonseca, Martins e Toledo (1976)

Figura 7 – Sem correlação r = 0



Fonte: Fonseca, Martins e Toledo (1976)

Gujarati e Porter (2011, p. 78) explicam que a correlação pode ser calculada tanto por

Equação 12 – Correlação

$$r = \pm\sqrt{r^2}$$

Fonte: Gujarati e Porter (2011, p. 98, equação 3.5.12)

quanto com base em sua definição, que é conhecido como "coeficiente de correlação amostral".

Equação 13 – Correlação amostral

$$r = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n\sum X_i^2 - (\sum X_i)^2]\,[n\sum Y_i^2 - (\sum Y_i)^2]}}$$

Fonte: Gujarati e Porter (2011, p. 98, equação 3.5.13)

Gujarati e Porter (2011, p. 98) ainda apresentam algumas das propriedades de r:

> 1. Pode ser positivo ou negativo, o que dependerá do sinal do termo no numerador da Equação, que mede a covariação amostral das duas variáveis;
> 2. Se situa entre os limites de +1 e -1, isto é, $+1 \leq r \leq -1$.
> 3. Sua natureza é simétrica, isto é, o coeficiente de correlação entre X e Y (r_{XY}) é o mesmo que aquele entre Y e X (r_{XY}).
> 4. É independente da origem e da escala, isto é, se definindo $X_i^* = a\,X_i + C$ e $Y_i^* = bY_i + d$, onde a > 0, b > 0 e c e d são constantes, então o r entre X* e Y* é o mesmo que aquele entre as variáveis originais X e Y.
> 5. Se X e Y são estatisticamente independentes, o coeficiente de correlação entre elas é zero, mas se r = 0, isso não significa que as variáveis sejam independentes. Em outras palavras, correlação zero não implica necessariamente independência.
> 6. É uma medida de associação linear ou de dependência linear; não é significativa para descrever relações não lineares. Assim $Y = X^2$ é uma relação exata, embora r seja zero.
> 7. Mesmo sendo uma medida de associação linear entre duas variáveis, ela não implica necessariamente qualquer relação de causa e efeito.

Em se tratando do coeficiente de determinação r^2, este é, segundo a econometria, uma medida da "qualidade do ajustamento". Ou seja, se o r permite verificar se existe uma correlação linear, o r^2 permite complementar a análise apontando o quanto os dados da população são explicados pelos dados da regressão.

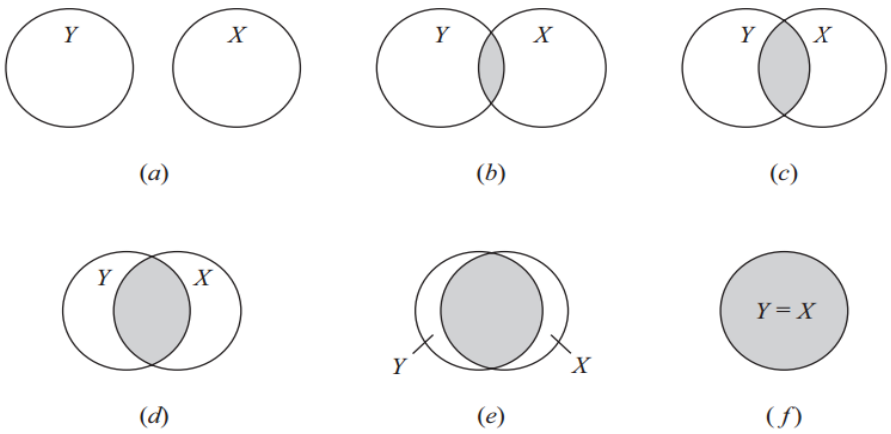
Gujarati e Porter (2011, p. 78) explicam que a tarefa de estimar a Função de Regressão Populacional (FRP) com base na Função de Regressão Amostral (FRA) da maneira mais precisa possível se faz utilizando o método dos MQO, por ser o mais utilizado para a análise de regressão, principalmente porque é intuitivamente convincente e matematicamente muito mais simples que outros métodos que costumam proporcionar resultados similares. Entendendo os princípios dos MQO, é possível perceber que algumas de suas propriedades estatísticas o tornaram um dos métodos de análise de regressão mais poderosos e difundidos em análises.

Ainda no entendimento de Gujarati e Porter (2011, p. 95), é necessário considerar a qualidade do ajustamento da linha de regressão em relação ao conjunto de dados, ou seja, verificar quão "bem" a linha de regressão é adequada aos dados. O interessante seria se todas

as observações estivessem sobre a linha de regressão, o que determinaria um ajustamento "perfeito", mas isso praticamente não acontece e normalmente alguns desses pontos apareceram, no eixo y, acima (positivos) ou abaixo (negativos) do ponto ideal.

O esperado é que esses resíduos em torno da linha de regressão sejam os menores possíveis. Para entender esse ajustamento para duas variáveis é utilizado o coeficiente de determinação r^2, uma importante medida para explicar a linha de regressão e o seu ajuste aos dados. Utilizando o diagrama de Venn, ou Ballentine, figura 12, Gujarati e Porter (2011, p. 95) explicam, de forma muito didática, como uma proporção cada vez maior da variação de Y é explicada por X.

Figura 8 – Diagrama de Venn para coeficiente de determinação r^2



Fonte: Gujarati e Porter (2011, p. 95, Figura 3.8)

A partir dos diagramas, que neste estudo sugerem o modelo de uma regressão de MQO, é obtida a sobreposição dos círculos, o que indica a extensão em que a variação de Y é explicada pela variação de X, ou seja, o r^2 é a medida numérica dessa sobreposição. A relação entre Y e X pode variar de quando não há sobreposição e X não explica a variação Y, a quando existe sobreposição total e X explica 100% a variação Y. Outra forma de representar o r^2 é no formato de desvio.

Equação 14 – Representação do r² no formato de desvio

$$y_i = \hat{y}_i + \hat{u}_i$$

Fonte: Gujarati e Porter (2011, p. 95)

Elevando ao quadrado os dois lados da equação e somando na amostra, é obtido:

Equação 15 – Representação do r² no formato de desvio a partir dos somatórios

$$\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma \hat{y}_i^2 + \Sigma \hat{u}_i^2 + 2\Sigma \hat{y}_i \hat{u}_i$$

$$= \Sigma \hat{y}_i^2 + \Sigma \hat{y}_i^2$$

$$= \hat{\beta}_2^2 \Sigma x_i^2 + \Sigma \hat{u}_i^2$$

Fonte: Gujarati e Porter (2011, p.95, equação 3.5.2)

Portanto, a equação também pode ser definida como:

STQ = SQE + SQR

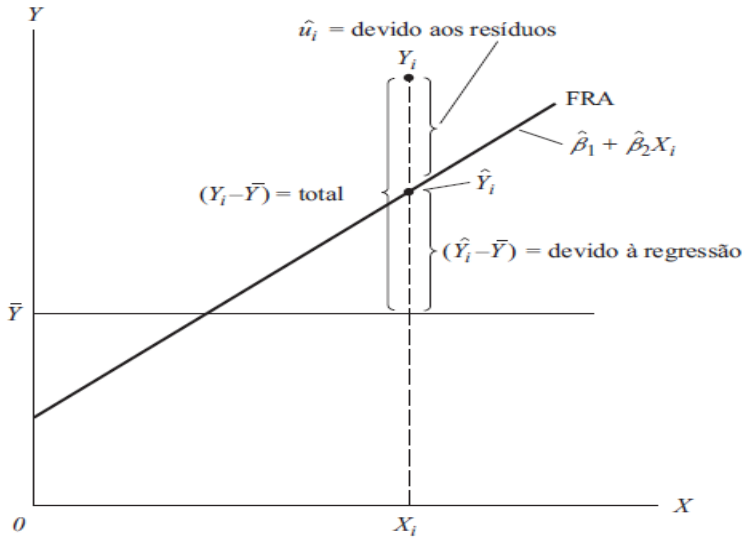
Fonte: Gujarati e Porter (2011)

Ou seja, a Soma Total de Quadrados (STQ) é $\sum y_i^2 = \sum (Y_i - \bar{Y})^2$, a Soma dos Quadrados Explicados pela Regressão (SQE) é $\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \hat{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2$ e a Soma dos Quadrados dos Resíduos (SQR), uma variação residual dos valores de Y em relação à linha de regressão, é $\sum \hat{u}_i^2$.

Isso mostra que a variação total dos valores observados de Y em torno de sua média pode ser dividida em duas partes, uma atribuível à linha de regressão e a outra a forças aleatórias, porque nem todas as observações efetivas de Y situam-se sobre a linha ajustada. Dividindo os dois lados da equação por STQ, é obtido:

Gráfico 1 – Separação da variação de Y em dois componentes

$$1 = \frac{SQE}{STQ} + \frac{SQR}{STQ} = \frac{\sum(\hat{Y}_i - \hat{Y})^2}{\sum(Y_i - \hat{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \hat{Y})^2}$$



Fonte: Gujarati e Porter (2011, p. 96, figura 3.9)

Agora definindo r² como:

Equação 16 – Definição do r²

$$r^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{SQE}{STQ}$$

Fonte: Gujarati e Porter (2011, p. 97, equação 3.5.5)

ou como:

Equação 17 – Coeficiente de determinação amostral

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{SQR}{STQ}$$

Fonte: Gujarati e Porter (2011, p. 97, equação 3.5.5a)

Gujarati e Porter (2011, p. 97) explicam esse r² como o coeficiente de determinação (amostral), sendo o indicador mais usado para medir a qualidade do ajustamento de uma linha

de regressão. Ou seja, o r^2 mede o percentual da variação total de Y explicada pelo modelo de regressão.

O autor lembra também que duas propriedades de r^2 devem ser destacadas. A primeira é um valor não negativo e a segunda seus limites são $0 \leq r^2 \leq 1$. Um r^2 igual a 1 significa um ajustamento perfeito, isto é, $\hat{Y}_i = Y_i$ para cada i. A segunda, um r^2 igual a zero significa que não há qualquer relação entre regressando e regressor ($\hat{\beta}_2 = 0$). Nesse caso, $\hat{Y}_i = \hat{\beta}_1 = \overline{Y}$, onde a melhor previsão para qualquer valor de Y é seu valor médio. Nessa situação, a linha de regressão será horizontal ao eixo dos X (GUJARATI; PORTER, 2011, p. 97).

Gujarati e Porter (2011, p. 97) também apresentam outra forma de calcular o r^2 rapidamente, que é a aplicação direta da definição da equação, com a seguinte fórmula:

Equação 18 – Dedução do r^2

$$r^2 = \frac{SQE}{STQ} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} = \hat{\beta}_2^2 \left(\frac{\sum x_i^2}{\sum y_i^2}\right)$$

Fonte: Gujarati e Porter (2011, p. 97, equação 3.5.6)

Então, Gujarati e Porter (2011, p. 97) indicam a divisão do numerador e do denominador pela amostra de tamanho n (ou por n-1, se o tamanho da amostra for muito pequeno), que é obtido:

Equação 19 – Dedução do r^2 a partir das variâncias amostrais

$$r^2 = \hat{\beta}_2^2 \left(\frac{S_x^2}{S_y^2}\right)$$

Fonte: Gujarati e Porter (2011, p. 97, equação 3.5.7)

em que S_y^2 *e* S_x^2 são as variâncias amostrais de Y e X, respectivamente.

Como $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$, também pode ser expressa como:

Equação 20 – Dedução do r² a partir das variâncias amostrais simplificado

$$r^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$$

Fonte: Gujarati e Porter (2011, p. 97, equação 3.5.8)

Equação 21 – Entendendo a Soma Total de Quadrados

$$STQ = SQE + SQR$$
$$\sum y_i^2 = r^2 \sum y_i^2 + (1 - r^2) \sum y_i^2$$

Fonte: Gujarati e Porter (2011, p. 98)

Na avaliação do peso das medidas, Gujarati e Porter (2011, p. 98) explicam que o r² é uma medida mais significativa que o r, já que essa "indica a proporção da variação da variável dependente explicada pela(s) variável(is) explanatória(s) e, portanto, proporciona uma medida geral da extensão em que a variação de uma variável determina a variação de outra. Já o r não tem esse valor".

Outra forma de calcular o r² é utilizar o coeficiente de correlação entre o Y_i observado e o Y_i estimado elevado ao quadrado, especificamente, o $\hat{Y}_i$. Usando a equação, como indicam Gujarati e Porter (2011, p. 99):

Equação 22 – r² através do quadrado do coeficiente de correlação

$$r^2 = \frac{[\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})]^2}{\sum(Y_i - \bar{Y})^2 \sum(\hat{Y}_i - \bar{Y})^2}$$

Fonte: Gujarati e Porter (2011, p. 99, equação 3.5.14)

Isto é:

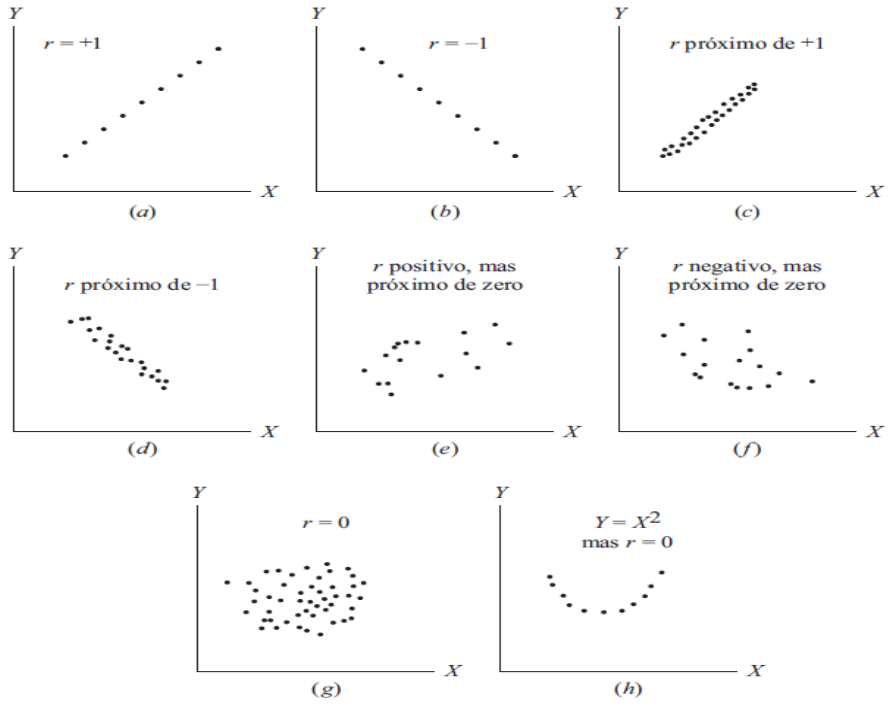
Equação 23 – r² através do quadrado do coeficiente de correlação simplificado

$$r^2 = \frac{(\sum y_i \hat{y}_i)^2}{(\sum y_i^2)(\sum \hat{y}_i^2)}$$

Fonte: Gujarati e Porter (2011, p. 99, equação 3.5.14)

Entendendo que Y_i = Y observado, $\hat{Y}_i$. = Y estimado, $\overline{Y} = \overline{\hat{Y}}$ = média de Y, informa quanto os valores estimados de Y estão próximos de seus valores observados. Gujarati e Porter (2011, p. 99) continuam com essa explanação graficamente com os padrões de correlação.

Figura 9 – Padrões de correlação



Fonte: Gujarati e Porter (2011, p. 99, figura 3.10)

4.3 CÁLCULO DA FUNÇÃO DE REGRESSÃO POR MEIO DE MQO

A pesquisa científica procura entender fenômenos e tratar com objetividade origens e repercussões nas mais diversas áreas. A confiabilidade na análise de dados experimentais pode ser determinante para creditar a explicação do fenômeno. Muitas vezes, se faz necessário um

ajuste de dados experimentais na busca da melhor forma de descrever esse conjunto de dados, ou seja, um conjunto finito de dados experimentais pode representar uma população e não só é possível conhecer como também estimar ocorrências futuras. Assim é o MQO, idealizado por Gauss, em 1795, com o objetivo de ser um estimador que maximiza o grau de ajuste do modelo matemático aos dados observados ao minimizar a Soma dos Quadrados dos Resíduos da regressão.

Inicialmente, mesmo antes de apresentar os elementos necessários para a construção da função de regressão por meio de MQO, é necessário abordar alguns conceitos básicos para uma melhor compreensão da análise por esse método.

O primeiro é o significado do termo "erro estocástico". Gujarati e Porter (2011, p. 65) explicam que um modelo de regressão não é, de forma alguma, uma descrição absolutamente fiel da realidade e, sendo assim, é certo que haverá diferenças entre o valor real do regressando e seus valores estimados no modelo escolhido. Considerando essa diferença, se estima o termo de erro estocástico u_i ou resíduo. Esse resíduo representa todas as variáveis omitidas. Para Gujarati e Porter (2011, p. 65), existem certas razões para não formular um modelo de regressão com o máximo de variáveis possíveis e, como contrapartida amostral, utilizar o termo de erro u_i no modelo.

Primeiramente por existir um caráter vago da teoria, quando se existe alguma, que explique o comportamento que Y pode apresentar, sendo, muitas vezes, incompleta. Ao saber com certeza que uma variável X influencia os valores de uma variável Y, é possível desconhecer, ou não ter certeza, quais são as outras variáveis que afetam Y. Portanto, o erro u_i pode ser usado como um substituto para todas as variáveis excluídas ou omitidas do modelo.

Ainda, em outro caso, mesmo conhecendo quais são algumas das variáveis excluídas e, portanto, considerando uma regressão múltipla em vez de uma simples, talvez não se tenha informações quantitativas a respeito dessas variáveis. Gujarati e Porter (2011, p. 65) entendem que a "indisponibilidade de dados é muito comum na análise empírica que os dados que gostaríamos idealmente de incluir não estejam disponíveis". A indisponibilidade pode obrigar a omissão de variável(eis) no modelo, mesmo que sua grande relevância teórica contribua para explicar Y.

Outro motivo é a relação entre variáveis essenciais *versus* variáveis periféricas/secundárias. Gujarati e Porter (2011, p. 65) apontam para o fato de que mesmo representando algum efeito para resultados de Y, "é bem possível que a influência conjunta de todas ou de algumas dessas variáveis seja tão pequena, e seja na melhor das hipóteses, não sistemática ou aleatória que, em termos práticos e para consideração de custos, não compense

incluí-las explicitamente no modelo". Para essas variáveis, é entendido que seu efeito combinado possa ser tratado como um termo de erro u_i.

Também existe o "caráter intrinsecamente aleatório do comportamento humano". Gujarati e Porter (2011, p. 65) lembram que "mesmo se conseguirmos incluir todas as variáveis relevantes no modelo, sempre haverá uma aleatoriedade 'intrínseca' nos Y individuais que não pode ser explicada por mais que nos esforcemos para tanto". Para isso, os termos de erro u_i podem refletir bem a aleatoriedade intrínseca.

Outro ponto levantado por Gujarati e Porter (2011, p. 66) é que, embora o modelo clássico de regressão proponha que as variáveis Y e X sejam medidas com exatidão, na prática, os dados podem estar infestados de erros de medição, ou seja, variáveis *proxy* pouco adequadas. O autor cita o exemplo da conhecida Teoria da Função Consumo, de Milton Friedman, que:

> [...] considera o *consumo permanente* (Y_P) como uma função da *renda permanente* (X_P). Mas, como os dados relativos a essas variáveis não são diretamente observáveis, na prática, utilizamos variáveis *proxy*, como consumo corrente (Y) e renda corrente (X), que são observáveis. Como os Y e X observados podem não ser iguais aos Y_P e X_P, há um problema de erro de medição. Nesse caso, o termo de erro u também pode representar erro de medição. […], se existirem tais erros de medição, eles podem ter sérias implicações na estimativa dos coeficientes da regressão, os ß.

Gujarati e Porter (2011, p. 66) também lembram do princípio da parcimônia, ou seja, o ideal seria formular o modelo de regressão mais simples possível. Isso quer diz que se for possível "explicar parte 'substancial' do comportamento de Y com duas ou três variáveis explanatórias e se nossa teoria não for suficientemente forte para sugerir quais outras variáveis podem ser incluídas, por que adicionar mais variáveis?". Nesse caso, é melhor deixar que u_i represente todas as outras variáveis, lembrando que não se deve excluir variáveis importantes e relevantes para apenas manter o modelo de regressão simples.

Por fim, a possibilidade da forma funcional errada, com Gujarati e Porter (2011, p. 66) reforçando a ideia de que "mesmo se as variáveis explanatórias de um fenômeno forem teoricamente corretas e mesmo se encontrarmos dados para essas variáveis, muitas vezes desconheceremos a forma funcional da relação entre o regressando e os regressores". Explica ainda que em modelos de duas variáveis, a forma funcional da relação pode muitas vezes ser inferida a partir do gráfico de dispersão, mas que em um modelo de regressão múltipla, não é possível determinar a relação funcional adequada, pois graficamente os diagramas de dispersão são apresentados com múltiplas dimensões.

Gujarati e Porter (2011, p. 66) finalizam sua explanação apontando que "por todas essas razões, o termo de erro estocástico u_i assume um papel fundamental na análise de regressão".

De uma maneira geral, o estimador de termo de erro estocástico u_i é o elemento aleatório ou não sistemático de todas as variáveis omitidas ou negligenciadas que podem afetar Y e que não foram incluídas nos modelos de regressão.

O próximo conceito abordado, antes da aplicação direta do MQO, é o da Função de Regressão Amostral (FRA), que aborda problemas nas amostras, onde valores de Y correspondentes a alguns valores X fixados são necessários para estimar a função de regressão. Para ilustrar, imagine que a população seja desconhecida e que a única informação disponível seja uma amostra selecionada aleatoriamente de valores de Y para os X fixados. Ou seja, ao contrário de uma população, se tem apenas um valor de Y para cada X escolhido aleatoriamente, dentre os Y correspondentes aos X_i dados de uma população.

Gujarati e Porter (2011, p. 66) explicam que não seria possível prever os valores de Y para a população como um todo correspondentes aos X escolhidos devido às variações amostrais, uma vez que, se selecionar outra amostra aleatória da mesma população, possivelmente será obtido um diagrama de dispersão, duas linhas de regressão amostral: FRA1 baseia-se na primeira amostra e FRA2 na segunda. Mas "Qual das duas linhas de regressão representa a linha de regressão populacional 'real'?". Nesse caso, tais linhas de regressão são conhecidas como linhas de regressão amostral e, tecnicamente, "representam a linha de regressão populacional, mas devido às variações amostrais, elas são, no máximo, aproximações da verdadeira regressão populacional. Em geral, obtemos N diferentes FRAs para N amostras diferentes, e estas FRAs provavelmente não serão as mesmas".

Gujarati e Porter (2011, p. 66) seguem esclarecendo a fórmula para o conceito de Função de Regressão Amostral (FRA), que representa a linha de regressão da amostra inicialmente determinada à equação correspondente à linear para a amostra escrita como:

Equação 24 – Função de Regressão Amostral (FRA)

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 \ X_i$$

Fonte: Gujarati e Porter (2011, p. 67, equação 2.6.1)

Onde:

$\hat{Y}_i$ = estimador de $E(Y \mid X_i)$

$\hat{\beta}_1$ = estimador de $ß_1$

$\hat{\beta}_2$ = estimador de $ß_2$

> Observe que um **estimador**, também conhecido como **estatística** (amostral), é apenas uma regra ou fórmula ou método que nos diz como estimar o parâmetro da população com base nas informações oferecidas pela amostra que temos à mão. Um valor numérico em particular obtido pela aplicação do estimador é conhecido como **estimativa**. Pode ser visto como aleatório, mas uma estimativa não é aleatória (GUJARATI; PORTER, 2011, p. 67).

Nessa mesma equação é possível aplicar $\hat{u}_i$, que representa o termo residual (na amostra).

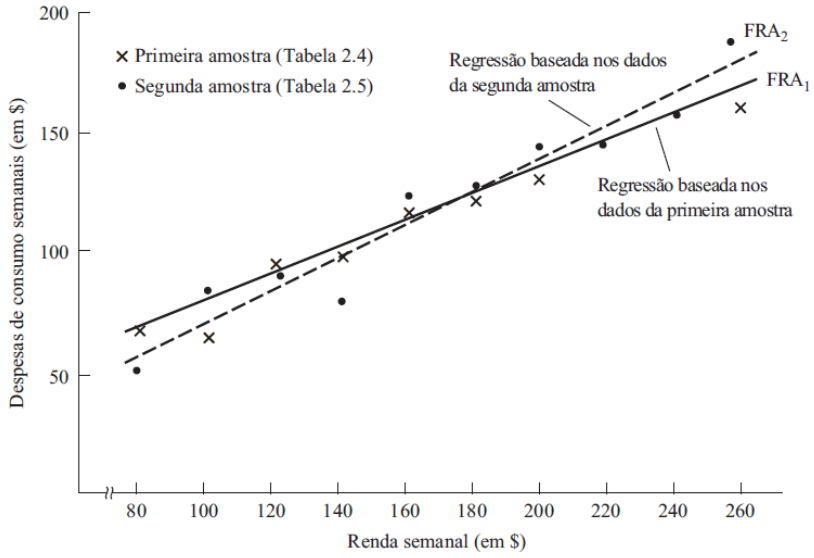
Equação 25 – Função de Regressão Amostral (FRA) com resíduo

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

Fonte: Gujarati e Porter (2011, p. 67, equação 2.6.2)

Para entender melhor, Gujarati e Porter (2011, p. 66) explicam que, durante a aplicação dessa técnica, o objetivo primordial na análise de regressão é estimar a Função de Regressão Populacional (FRP) com base na FRA, porque geralmente a análise está baseada em uma única amostra de alguma população. O autor diz que para $X = X_i$, se tem uma observação (amostral) $Y = Y_i$.

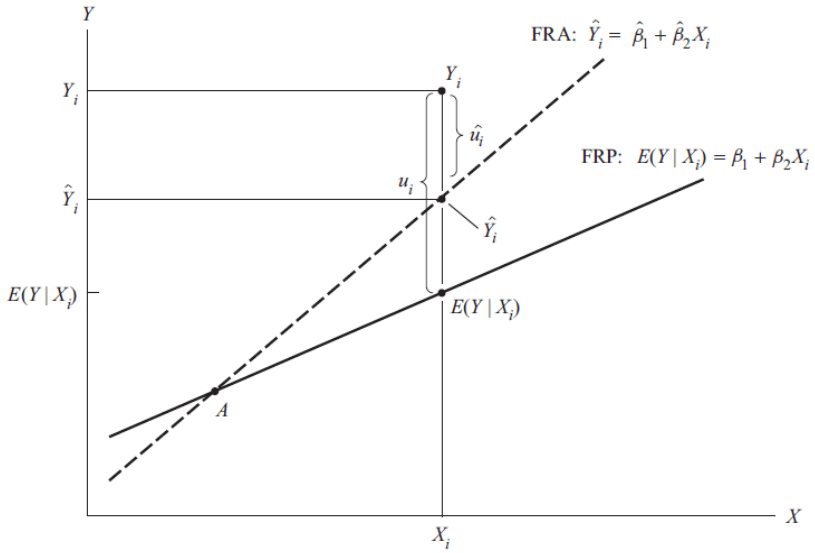
Gráfico 2 – Linhas de regressão baseadas em duas amostras diferentes



Fonte: Gujarati e Porter (2011, p. 69, figura 2.4)

Mas em casos de variações amostrais, as estimativas da FRP com base na Função de Regressão Amostral (FRA) será, na melhor das hipóteses, apenas uma aproximação.

Gráfico 3 – Funções de regressão para amostra e população



Fonte: Gujarati e Porter (2011, p. 69, figura 2.5)

Gujarati e Porter (2011, p. 66) explicam o Gráfico 3 informando que $\hat{Y}_i$ superestima a verdadeira $E(Y \mid X_i)$ para o X_i nela demonstrado. Assim como, para cada X_i à esquerda do ponto A, o valor da FRA subestimará os da verdadeira FRP devido às variações amostrais. E conclui trazendo as questões:

> A pergunta crítica agora é: sabendo que a FRA não é mais do que uma aproximação da FRP, podemos formular uma regra ou um método que torne essa aproximação a mais próxima possível? Em outras palavras, como devemos formular a FRA para que $\hat{\beta}_1$ fique o mais próximo possível do verdadeiro $ß_1$ e $\hat{\beta}_2$ do verdadeiro $ß_2$ mesmo que nunca venhamos a saber quais são os verdadeiros $ß_1$ e $ß_2$? [...] Aqui destacamos que é possível desenvolver procedimentos que nos digam como formular a FRA a fim de espelhar FRP o mais fielmente possível. É fascinante considerar que isso pode ser feito mesmo que nunca determinemos a FRP real.

Nesse ponto, de porte desses conhecimentos introdutórios, finalmente é possível explorar a aplicação do método dos Mínimos Quadrados Ordinários. A aplicação do MQO consiste em determinar a reta que melhor descreve um conjunto de dados dispersos no plano cartesiano. Gujarati e Porter (2011, p. 78), de uma forma muito didática, explicam todos os princípios para entender e aplicar o MQO. Partindo inicialmente da Função de Regressão Populacional (FRP), já explicada, para duas variáveis e entendido que:

Equação 26 – Função de Regressão Populacional (FRP)

$$Y_i = ß_1 + ß_2 X_i + u_i$$

Fonte: Gujarati e Porter (2011, p. 78, equação 2.4.2)

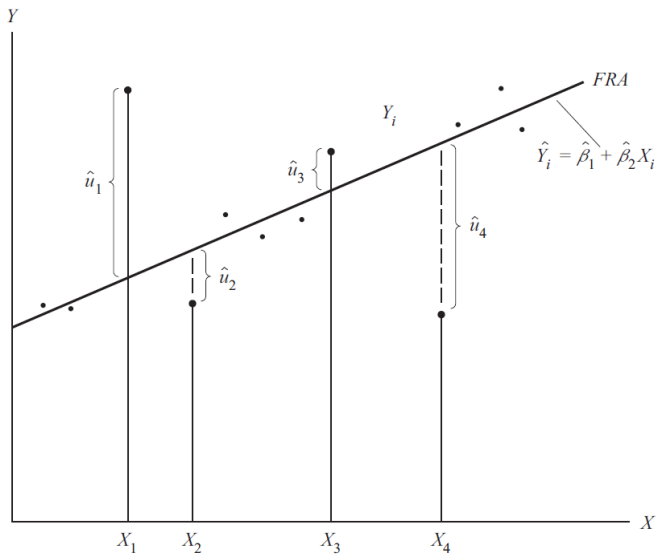
Não pode ser observada diretamente e será estimada por meio da Função de Regressão Amostral (FRA).

Equação 27 – Da Função de Regressão Amostral (FRA)

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$
$$= \hat{Y}_i + \hat{u}_i$$

Fonte: Gujarati e Porter (2011, p. 78, equação 2.6.2 e 2.6.3)

Gráfico 4 – Critério dos Mínimos Quadrados



Fonte: Gujarati e Porter (2011, p. 79, figura 3.1)

Ainda recapitulando, $\hat{Y}_i$ é o valor estimado, média condicional, de Y_i. Já a FRA primeiro deve expressar os $\hat{u}_i$ (os resíduos) estimados, que são simplesmente as diferenças entre os valores observados e estima dos de Y.

Equação 28 – Expressão dos resíduos

$$\hat{u}_i = Y_i - \hat{Y}_i$$
$$= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

Fonte: Gujarati e Porter (2011, p. 78, equação 3.1.1)

Analisando os dados de n pares de observações de Y e X, Gujarati e Porter (2011, p. 79) explicam, através do diagrama de dispersão hipotético, como determinar a FRA de maneira que fique o mais próximo possível do Y observado, inicialmente aplicando o critério de escolher a FRA de tal forma que a soma dos resíduos $\Sigma\hat{u}_i = \Sigma (Y_i - \hat{Y}_i)$ seja a menor possível. Mas o autor conclui que, mesmo convincente de que esse não é um critério adequado, pois minimizar o $\Sigma\hat{u}_i$, os resíduos $\hat{u}_1$ e $\hat{u}_4$ têm o mesmo peso na soma ($\hat{u}_1 + \hat{u}_2 + \hat{u}_3 + \hat{u}_4$), embora os dois primeiros estejam muito mais próximos da FRA que os dois últimos. "Em outras palavras, todos os resíduos recebem a mesma importância independentemente de quão próximos ou distantes estejam das observações individuais em relação à FRA".

Nesse instante, Gujarati e Porter (2011, p. 79) trazem a importância da aplicação dos Mínimos Quadrados, segundo o qual a FRA pode ser fixada de tal maneira que seja o menor possível, onde os $\hat{u}_i^2$ são os resíduos elevados ao quadrado dando mais peso aos resíduos como $\hat{u}_1$ e $\hat{u}_4$, do que aos resíduos $\hat{u}_2$ e $\hat{u}_3$. A diferença está no fato de que quanto maior $\hat{u}_i$, em valores absolutos, maior $\Sigma\hat{u}_i^2$ e maior relevância no somatório.

Equação 29 – Critério dos Mínimos Quadrados somatório

$$\sum(\hat{u}_i)^2 = \sum(Y_i - \hat{Y}_i)^2$$
$$\sum(\hat{u}_i)^2 = (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$
$$\sum\hat{u}_i^2 = f(\hat{\beta}_1, \hat{\beta}_2)$$

Fonte: Gujarati e Porter (2011, p. 79, equações 3.1.2 e 3.1.3)

A soma do quadrado dos resíduos é uma função dos estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$. "Para qualquer conjunto de dados, a escolha de valores diferentes para $\hat{\beta}_1$ e $\hat{\beta}_2$ resultará em û diferentes e, portanto, em valores diferentes de $\Sigma\hat{u}_i^2$" (GUJARATI; PORTER, 2011, p. 79). Para entender melhor essa afirmação, Gujarati e Porter (2011, p. 80) apresentam uma tabela de exemplos com os valores hipotéticos de Y e X e sugere dois experimentos.

Tabela 3 – Determinação experimental da FRA

Y_i (1)	X_t (2)	$\hat{Y}_{1i}$ (3)	$\hat{u}_{1i}$ (4)	$\hat{u}_{1i}^2$ (5)	$\hat{Y}_{2i}$ (6)	$\hat{u}_{2i}$ (7)	$\hat{u}_{2i}^2$ (8)
4	1	2,929	1,071	1,147	4	0	0
5	4	7,000	−2,000	4,000	7	2	4
7	5	8,357	−1,357	1,841	8	1	1
12	6	9,714	2,286	5,226	9	3	9
Soma: 28	16		0,0	12,214		0	14

Notas: $\hat{Y}_{1i} = 1{,}572 + 1{,}357X_i$ (isto é, $\hat{\beta}_1 = 1{,}572$ e $\hat{\beta}_2 = 1{,}357$)
$\hat{Y}_{2i} = 3.0 + 1{,}0X_i$ (isto é, $\hat{\beta}_1 = 3$ e $\hat{\beta}_2 = 1{,}0$)
$\hat{u}_{1i} = (Y_i - \hat{Y}_{1i})$
$\hat{u}_{2i} = (Y_i - \hat{Y}_{2i})$

Fonte: Gujarati e Porter (2011, p. 80, tabela 3.1)

Para entender o experimento, basta saber que a obtenção dos estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$, que serão explicados posteriormente, e que significam neste momento que $\hat{\beta}_2$ é coeficiente angular e determina a inclinação da reta, também determina se a função é crescente ou decrescente, $\hat{\beta}_2$ para valor positivo ou negativo. O valor $\hat{\beta}_1$ é o coeficiente linear da reta e é o ponto onde a reta corta o eixo y.

Agora, no primeiro, considerando os valores que foram obtidos aplicando-se o método dos Mínimos Quadrados, temos $\hat{\beta}_1$ = 1,572 e $\hat{\beta}_2$ = 1,357.1.

> Usando esses valores de $\hat{\beta}$ e os valores de X fornecidos da coluna (2) da Tabela 3.1, podemos calcular facilmente os Y_i estimados dados na coluna (3) como $\hat{Y}_i$ (em que o subscrito l indica o primeiro experimento). Agora, vamos conduzir outro experimento, desta vez utilizando os valores $\hat{\beta}_1$ = 3 e $\hat{\beta}_2$ = 1. Os valores estimados de Y_i neste experimento aparecem como $\hat{Y}_{2i}$ na coluna (6) da Tabela 3.1. Como os valores de $\hat{\beta}$ nos dois experimentos são diferentes, obtemos valores diferentes para os resíduos estimados, como se vê na tabela; os $\hat{u}_{1i}$ são os resíduos do primeiro experimento e os $\hat{u}_{2i}$ resíduos do segundo. Os quadrados desses resíduos estão nas colunas (5) e (8). Obviamente, como poderíamos esperar da Equação (3.1.3), a soma dos quadrados desses resíduos são diferentes, já que têm como base conjuntos diferentes de valores de $\hat{\beta}$.

Com o conjunto de valores, Gujarati e Porter (2011, p. 80) apontam para os dados com o menor somatório $\Sigma\hat{u}_i^2$, a partir dos valores de $\hat{\beta}$, de um dos experimentos como o "ideal". Ou seja, devem ser escolhidos os valores de $\hat{\beta}$ do primeiro experimento, que fornece um $\Sigma\hat{u}_i^2$ menor (= 12,214) do que os obtidos com os valores de $\hat{\beta}$ no segundo experimento (= 14), pois o método dos Mínimos Quadrados escolhe $\hat{\beta}_1$ e $\hat{\beta}_2$ de tal forma que, para qualquer amostra, ou conjunto de dados, o $\Sigma\hat{u}_i^2$ será menor possível. Para uma dada amostra, o método dos Mínimos Quadrados sempre oferece estimativas únicas de $\hat{\beta}_1$ e $\hat{\beta}_2$, que proporcionam o menor valor possível de $\Sigma\hat{u}_i^2$.

Esses valores de $\hat{\beta}_1$ e $\hat{\beta}_2$ são conhecidos como estimadores de Mínimos Quadrados, por serem derivados do princípio dos Mínimos Quadrados, assim como suas propriedades

numéricas dos estimadores obtidos por meio do método dos MQO. Para encontrar os valores de $\hat{\beta}_1$ e $\hat{\beta}_2$, basta seguir o raciocínio em torno do $\Sigma\hat{u}_i^2$. Gujarati e Porter (2011, p. 80) explicam que o processo é um exercício direto de cálculo diferencial e o processo de diferenciação resulta no seguinte sistema de equações, conhecidas como equações normais, para encontrar os estimadores em que n é o tamanho da amostra:

Equação 30 – Equações normais do MQO

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$
$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

Fonte: Gujarati e Porter (2011, p. 80 – equações 3.14 e 3.15 e 2.6.3)

Mas como encontrar as equações para obter os estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$? Para responder essa pergunta será necessário entender um pouco de cálculo diferencial, o que foge do escopo explicativo desta pesquisa, e aplicar derivadas parciais a partir da equação conhecida:

$$\sum(\hat{u}_i)^2 = \sum(Y_i - \hat{Y}_i)^2$$
$$\sum(\hat{u}_i)^2 = (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

Fonte: Gujarati e Porter (2011, p. 80)

Primeiro com a derivada parcial em relação a $\hat{\beta}_1$.

$$\frac{\partial(\sum \hat{\mu}_i^2)}{\partial \hat{\beta}_1} = \frac{\partial(\sum(Y_i - \hat{Y}_i)^2}{\partial \hat{\beta}_1} = \frac{\partial[(\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2]}{\partial \hat{\beta}_1}$$

$$= 2[\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)](-1)$$

$$= -2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)$$

Em seguida, a derivada parcial em relação a $\hat{\beta}_2$.

$$\frac{\partial(\sum \hat{\mu}_i^2)}{\partial \hat{\beta}_2} = \frac{\partial(\sum(Y_i - \hat{Y}_i)^2}{\partial \hat{\beta}_2} = \frac{\partial[(\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2]}{\partial \hat{\beta}_2}$$

$$= 2[\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)](-X_i)$$
$$= -2\sum(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)\, X_i$$

O próximo passo, com as equações resultantes das derivações que devem ser igualadas a zero para achar o ponto de mínimo.

$$-2\sum\left(Y_{\mathrm{i}} - \hat{\beta}_1 - \hat{\beta}_2 X_{\mathrm{i}}\right) = 0$$
$$\sum\left(Y_{\mathrm{i}} - \hat{\beta}_1 - \hat{\beta}_2 X_{\mathrm{i}}\right) = 0$$
$$\sum Y_{\mathrm{i}} - \sum \hat{\beta}_1 - \sum \hat{\beta}_2 X_{\mathrm{i}} = 0$$
$$\sum Y_{\mathrm{i}} - n\hat{\beta}_1 - \hat{\beta}_2 \sum X_{\mathrm{i}} = 0$$

O resultado da primeira é justamente o indicado por Gujarati e Porter (2011, p. 80):

$$\sum Y_{\mathrm{i}} - n\hat{\beta}_1 - \hat{\beta}_2 \sum X_{\mathrm{i}} = 0$$

E, igualando a segunda a zero, é obtida a outra equação do sistema.

$$-2\sum\left(Y_{\mathrm{i}} - \hat{\beta}_1 - \hat{\beta}_2 X_{\mathrm{i}}\right) X_{\mathrm{i}} = 0$$
$$\sum(Y_{\mathrm{i}} - \hat{\beta}_1 - \hat{\beta}_2 X_{\mathrm{i}})\, X_{\mathrm{i}} = 0$$
$$\sum(Y_{\mathrm{i}} X_{\mathrm{i}} - \hat{\beta}_1 X_{\mathrm{i}} - \hat{\beta}_2 {X_{\mathrm{i}}}^2) = 0$$
$$\sum Y_{\mathrm{i}} X_{\mathrm{i}} - \sum \hat{\beta}_1 X_{\mathrm{i}} - \sum \hat{\beta}_2 {X_{\mathrm{i}}}^2 = 0$$
$$\sum Y_{\mathrm{i}} X_{\mathrm{i}} - \hat{\beta}_1 \sum X_{\mathrm{i}} - \hat{\beta}_2 \sum {X_{\mathrm{i}}}^2 = 0$$
$$\sum Y_{\mathrm{i}} X_{\mathrm{i}} = \hat{\beta}_1 \sum X_{\mathrm{i}} + \hat{\beta}_2 \sum {X_{\mathrm{i}}}^2$$

Entendido como foi determinado o sistema de equações apresentas por Gujarati e Porter (2011, p. 80), o próximo passo é resolver as equações normais simultaneamente. O resultado é conseguido com manipulações algébricas simples. A partir da primeira é definido $\hat{\beta}_1$.

Equação 31 – Determinação de $\hat{\beta}_1$

$$\sum Y_{\mathrm{i}} - n\hat{\beta}_1 - \hat{\beta}_2 \sum X_{\mathrm{i}} = 0$$
$$n\hat{\beta}_1 = \sum Y_{\mathrm{i}} - \hat{\beta}_2 \sum X_{\mathrm{i}}$$
$$\hat{\beta}_1 = \frac{\sum Y_{\mathrm{i}}}{n} - \hat{\beta}_2 \frac{\sum X_{\mathrm{i}}}{n}$$
$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

Fonte: Gujarati e Porter (2011, p. 80)

Em seguida, com a outra, é definido $\hat{\beta}_2$.

Equação 32 – Determinação de $\hat{\beta}_2$

$$\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

$$\sum X_i Y_i = \left[\frac{\sum Y_i}{n} - \hat{\beta}_2 \frac{\sum X_i}{n} \right] \sum X_i + \hat{\beta}_2 \sum X_i^2$$

$$\sum X_i Y_i = \frac{\sum X_i \sum Y_i}{n} - \hat{\beta}_2 \frac{\left(\sum X_i\right)^2}{n} + \hat{\beta}_2 \sum X_i^2$$

$$\sum X_i Y_i = \frac{\sum X_i \sum Y_i}{n} + \hat{\beta}_2 \left[\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} \right]$$

$$\hat{\beta}_2 \left[\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} \right] = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}$$

$$\hat{\beta}_2 \left[\frac{n\sum X_i^2 - \left(\sum X_i\right)^2}{n} \right] = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n}$$

$$\hat{\beta}_2 \left\{ (1/n) \left[n\sum X_i^2 - \left(\sum X_i\right)^2 \right] \right\} = (1/n) \left[n\sum X_i Y_i - \sum X_i \sum Y_i \right]$$

$$\hat{\beta}_2 = \frac{(1/n) \left[n\sum X_i Y_i - \sum X_i \sum Y_i \right]}{(1/n) \left[n\sum X_i^2 - \left(\sum X_i\right)^2 \right]}$$

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

$$\hat{\beta}_2 = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2}$$

ou

$$\hat{\beta}_2 = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2}$$

$$\hat{\beta}_2 = \frac{\sum (x_i)(y_i)}{\sum (x_i)^2}$$

Fonte: Gujarati e Porter (2011, p. 80)

Para finalizar, Gujarati e Porter (2011, p. 100) apresentam um exemplo numérico muito oportuno para sanar possíveis dúvidas. Finalmente, é possível perceber que para encontrar os estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$ basta preencher uma tabela com valores X_i, Y_i, $\bar{x}, \bar{y}$, e u_i,, ou seja, utilizar os

valores de amostra, valores médios e erro estocástico, e aplicar às equações do MQO como seguem:

Tabela 4 – Tabela exemplo escolaridade *versus* salário

A Tabela 2.6 apresenta dados relativos ao nível de escolaridade (medido pelo número de anos de frequência escolar), o salário-hora médio das pessoas em cada nível de escolaridade e o número de pessoas em cada um desses níveis. Ernst Berndt obteve originalmente os dados apresentados na tabela com base em um levantamento da população conduzido em maio de 1985.[14]

Anos de estudo	Salário médio ($/hora)	Número de pessoas
6	4,4567	3
7	5,7700	5
8	5,9787	15
9	7,3317	12
10	7,3182	17
11	6,5844	27
12	7,8182	218
13	7,8351	37
14	11,0223	56
15	10,6738	13
16	10,8361	70
17	13,6150	24
18	13,5310	31
	Total	528

Fonte: Gujarati e Porter (2011, p. 69, tabela 2.6)

Tabela 5 – Dados brutos com base na Tabela 4

Obs	Y	X	x	y	x_i^2	$y_i x_i$	$\boxed{X_i^2}$	$\boxed{Y_i^2}$	$\boxed{\hat{Y}_i}$	$\boxed{\hat{u}_i = Y_i - \hat{Y}}$	$\boxed{\hat{u}_i^2}$
1	4,4567	6	6	4,218	36	25,308	36	19,86217	4,165294	0,291406	0,084917
2	5,77	7	5	2,9047	25	14,5235	49	33,2929	4,916863	0,853137	0,727843
3	5,9787	8	4	2,696	16	10,784	64	35,74485	5,668432	0,310268	0,096266
4	7,3317	9	3	1,343	9	4,029	81	53,75382	6,420001	0,911699	0,831195
5	7,3182	10	2	1,3565	4	2,713	100	53,55605	7,17157	0,14663	0,0215
6	6,5844	11	1	2,0903	1	2,0903	121	43,35432	7,923139	−1,33874	1,792222
7	7,8182	12	0	0,8565	0	0	144	61,12425	8,674708	−0,85651	0,733606
8	7,8351	13	1	0,8396	1	0,8396	169	61,38879	9,426277	−1,59118	2,531844
9	11,0223	14	2	2,3476	4	4,6952	196	121,4911	10,17785	0,844454	0,713103
10	10,6738	15	3	1,9991	9	5,9973	225	113,93	10,92941	−0,25562	0,065339
11	10,8361	16	4	2.1614	16	8,6456	256	117,4211	11,68098	−0,84488	0,713829
12	13,615	17	5	4,9403	25	24,7015	289	185,3682	12,43255	1,182447	1,398181
13	13,531	18	6	4,8563	36	29,1378	324	183,088	13,18412	0,346878	0,120324
Soma	112,7712	156	0	0	182	131,7856	2054	1083,376	112,7712	≈0	9,83017

Fonte: Gujarati e Porter (2011, p. 100, tabela 3.2)

$$x_i = X_i - \bar{X};\ y_i = Y_i = \bar{Y}$$

$$\hat{\beta}_2 = \frac{\Sigma y_i x_i}{\Sigma x_i^2} = \frac{131{,}7856}{182{,}0} = 0{,}7240967$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 8{,}674708 - 0{,}7240967x12 = -0{,}01445$$

$$\hat{\sigma}^2 = \frac{\Sigma \hat{u}_i^2}{n-2} = \frac{9{,}83017}{11} = 0{,}893652;\ \hat{\sigma} = 0{,}945332$$

$$\text{var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\Sigma x_i^2} = \frac{0{,}893652}{182{,}0} = 0{,}004910;\ \text{ep}(\hat{\beta}_2) = \sqrt{0{,}00490} = 0{,}070072$$

$$r^2 = 1 - \frac{\Sigma \hat{u}_i^2}{\Sigma (Y_i - \bar{Y})^2} = 1 - \frac{9{,}83017}{105{,}1188} = 0{,}9065$$

$$r = \sqrt{r^2} = 0{,}9521$$

$$\text{var}(\hat{\beta}_1) = \frac{\Sigma x_i^2}{n \Sigma x_i^2} = \frac{2054}{13(182)} = 0{,}868132;$$

$$\text{ep}(\hat{\beta}_1) = \sqrt{0{,}868132} = 0{,}9317359$$

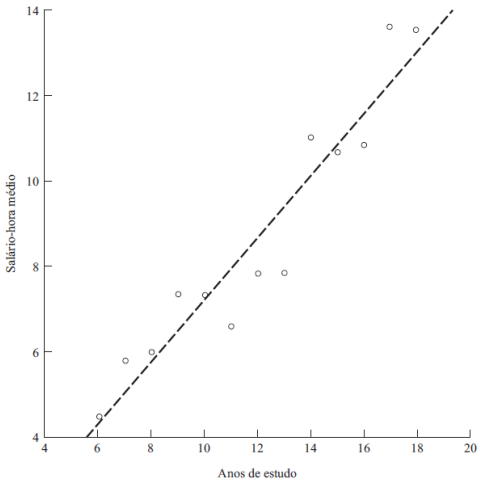
Fonte: Gujarati e Porter (2011, p. 100)

Com base nos dados da tabela, é obtida a linha de regressão estimada a partir da equação linear da função de regressão estimada para os dados salário *versus* escolaridade (GUJARATI; PORTER, 2011, p. 101, equação 3.6.1)

$$\mathbf{\hat{Y}_i = -0{,}0144 + 0{,}7240X_i}$$

Correspondente à função linear $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$.

Gráfico 5 – Linha de regressão estimada para os dados salário *versus* escolaridade



Fonte: Gujarati e Porter (2011, p. 101, figura 3.11)

Gujarati e Porter (2011, p. 101) lembram que cada ponto da linha de regressão representa uma estimativa do valor médio de Y correspondente ao valor de X escolhido; $\hat{Y}_i$ é uma estimativa de $E(Y \mid X_i)$.

No exemplo, o valor de $\hat{\beta}_2$ = 0,7240 explica que, dentro da faixa pesquisada de seis a 18 anos de estudo, um ano já determina o aumento estimado no salário/hora médio que é de cerca 0,72. Ou seja, para cada ano adicional de escolaridade (valor de X), em média, o salário/hora aumenta em 72 centavos de dólar. Essa é a capacidade preditiva do modelo.

Já o valor $\hat{\beta}_1$ = -0,0144 explica o nível médio do salário semanal quando a escolaridade for zero. Na verdade, pode ser interpretado que o valor médio do salário será baixo, mas literalmente o intercepto nesse caso não faz sentido, principalmente por ser um valor negativo. O resultado foi influenciado, provavelmente, pelo intervalo amostral que começar com seis anos de estudo e não inclui um nível de escolaridade zero.

Nesse exemplo, Gujarati e Porter (2011, p. 101) apresentam também o valor de r^2, em torno de 0,90, que sugere que a escolaridade explica cerca de 90% da variação no salário com uma linha de regressão que se ajusta muito bem aos dados, além do coeficiente de correlação r = 0,9521 demonstrado um ajuste de correlação positiva e alta entre salário e anos de estudo.

4.4 FORMAS DA FUNÇÃO DE REGRESSÃO

A regressão descreve, através de um modelo matemático, a relação existente entre duas variáveis a partir de n observações, que geralmente se expressa por meio de uma função. Vimos

que, em um modelo puro das possíveis relações de X e Y, a variável de erro não existe, por isso usam-se modelos estatísticos onde uma variável de erro (u) é considerada. Isso significa que haverá uma função que estabelece uma relação explicável de X sobre Y e uma parte não explicada pelo modelo (que é o erro). Isso é utilizado pela pesquisa científica para entender fenômenos e tratar com objetividade as origens e repercussões desses fenômenos nas mais diversas áreas do conhecimento.

Para que isso ocorra com confiabilidade, se faz necessário um ajuste de dados experimentais ou sociais, visando uma forma objetiva de descrever esse conjunto de dados, não só para explicar fenômenos atuais, como também para estimar ocorrências futuras. Assim, se usa o MQO idealizado por Gauss, em 1795, com o objetivo de ser um estimador que maximiza o grau de ajuste do modelo matemático aos dados observados ao minimizar a Soma dos Quadrados dos Resíduos da regressão.

Um ajustamento contém, portanto, duas características: a primeira é a determinação dos parâmetros da reta que melhor representa a relação verdadeira entre as variáveis X e Y; a segunda é que o processo de ajustamento que deve escolher a função (forma da curva ou forma funcional), através da qual os valores de X explicarão os de Y. Dito de outra forma, dado os valores observados de X e Y, busca-se construir um modelo de regressão linear de Y sobre X para obter uma reta a partir dos valores observados. Como vimos antes, uma das formas de facilitar este trabalho é construir o diagrama de dispersão, anotando em um gráfico os valores que elegemos para X e para Y. Em geral todos os processos de ajustamento consistem em supor uma relação funcional linear, ou seja, inicia-se a análise da regressão por uma regressão linear. O MQO, explicado anteriormente, cumpre o papel de encontrar essa reta de Mínimos Quadrados e de calcular os parâmetros de tal forma que seja mínima a soma dos quadrados das diferenças entre os valores observados de Y e os obtidos a partir da reta ajustada para os mesmos valores de X.

Se a forma linear for a melhor forma para o modelo, então o coeficiente de regressão e o coeficiente de correlação serão altos (próximos da unidade). Se isso não ocorrer é porque a relação funcional ou forma da curva pode ser não linear. Caberá então a tarefa de buscar outra forma de curva que expresse uma maior qualidade do ajustamento, ou seja, um maior r^2 (coeficiente de regressão) do que aquele expresso pelo modelo linear.

Tome-se como exemplo uma função potência. A forma da função potência é $Y = aX^b$. A linearização ($y = A + bx$) ocorre com a utilização de logaritmo, assim: $\log Y = \log a + b \log X$. Para a curva de função potência, temos: $y = \log Y$, $A = \log a$, $x = \log X$. Para calcular os parâmetros na forma linear, temos: $a = (\Sigma y/n)-(b\Sigma y/n)$ e $b = (\Sigma xy - (\Sigma x.\Sigma y/n))/(\Sigma x^2 – (\Sigma x^2/n))$.

Para calcular os parâmetros na forma de função potência, temos: a = (ΣlogY/n)-(bΣlogY/n) e b = (ΣlogX.logY - (ΣlogX.ΣlogY/n))/(ΣlogX^2 - (ΣlogX^2/n)).

Logo, podemos concluir que existem formas funcionais entre as variáveis X e Y que não são lineares, significando que Y é uma função não linear de X. Se isso ocorre, cada forma de curva terá sua maneira de calcular os parâmetros, isto é, cada função tem o seu MQO. Isso ficou claro com a função potência e o mesmo princípio ocorre com as demais formas da função, que podem ser facilmente encontradas na internet, onde estão disponíveis inclusive *sites*, programas e pacotes estatísticos que calculam a forma da curva e os parâmetros do modelo. Dentre as dezenas de funções que expressam formas de curvas, destacamos: função exponencial, função hipérbole, função parábola do segundo grau, função parábola do terceiro grau.

4.5 CONSTRUINDO UM MODELO: CONSIDERAÇÕES

A metria dos fenômenos econômicos (econometria) conta com oito passos para construir a metodologia econométrica, a saber: 1) exposição da teoria ou hipótese; 2) especificação do modelo matemático da teoria; 3) especificação do modelo estatístico ou econométrico; 4) obtenção dos dados; 5) estimação dos parâmetros do modelo econométrico; 6) teste de hipóteses; 7) projeção ou previsão; e 8) uso do modelo para fins de controle ou de política.

Empiricamente, os métodos estatísticos de análise de dados a partir de conceitos e teorias da economia, enquanto disciplina social, estão mais consolidados no contexto de bagagem metodológica do que as metrias da informação, pois a econometria enseja metodologias bastante consistentes para os estudos de dados econômicos (produção, consumo, distribuição, demanda, oferta, juros, renda nacional, balança internacional etc.).

Na presente tese, nosso modelo visa estudar o cenário de produção científica nacional a partir dos valores conhecidos. Com isso, ao nosso ver, dentro das metrias abordadas na revisão de literatura, algo que se possa denominar de relações prévias de causa e efeito consagrados pela teoria estariam apenas nas leis da bibliometria (Zipf, Lotka e Bradford). Desse modo, excluiremos a fase da "exposição da hipótese" da proposta de modelo desta tese, por entendermos que as metrias da informação são insipientes no sentido de teorias informacionais conseguirem demonstrar previamente relações de causa e efeito. Obviamente que quando os dados ou a pesquisa se comportarem como uma das leis da bibliometria, poder-se-á usar as suas hipóteses.

Algumas das metrias da informação abordadas na revisão de literatura possuem formas da curva que permitem alegar que estas possuem um modelo matemático e estatístico, como as Leis Bibliométricas de Lotka e Bradford, e um pouco menos quanto a Lei Bibliométrica de Zipf. Já em relação à Teoria Epidêmica de Goffman, nem isso podemos afirmar. As demais metrias, ao nosso olhar, se valem tão somente de abordagens univariadas com cálculos que se utilizam de médias, medianas, índices, percentuais, desvios e moda para os estudos informacionais. Como na presente tese nosso modelo visa estudar o cenário de produção científica nacional a partir dos valores conhecidos, manteremos a fase de "especificação do modelo matemático e estatístico" para a proposta do modelo.

Todas as metrias da informação abordadas na revisão de literatura possuem dados dos fenômenos informacionais, porém nem todos podem ser trabalhados através de modelagens matemáticas e estatísticas no sentido da análise da regressão, mas apenas em uma perspectiva univariada (médias, desvios, moda, mediana, índices, percentuais etc.). Como nesta tese visamos estudar o cenário de produção científica nacional a partir dos valores conhecidos, manteremos a fase da "obtenção dos dados" para a proposta do modelo.

Na revisão de literatura, algo que se possa denominar de verificação da adequação dos parâmetros estimados aos parâmetros populacionais estariam apenas nas leis da bibliometria (Zipf, Lotka e Bradford). Assim, manteremos a fase de "teste de hipótese" na proposta de modelo desta tese, mas a desmembramos em dois momentos: "análise da correlação", para medir a força ou o grau de associação linear entre duas variáveis com o respectivo mapa de dispersão e para decidir se vale a pena regredir o modelo; e "análise da regressão", para verificar a qualidade do ajustamento através da relação funcional linear ou não linear (forma da curva linear ou não linear), isto é, a melhor adequação possível dos parâmetros estimados aos parâmetros populacionais.

A todas as metrias da informação abordadas na revisão de literatura podem ser atribuídos dados de fenômenos informacionais que possam ser trabalhados através de modelagens matemáticas e estatísticas, mas nem todos terão uma análise da regressão. A pesquisa desta tese permite a análise da regressão dos dados do cenário de produção científica nacional, por isso manteremos a fase de "estimação dos parâmetros" para a proposta do modelo.

Todas as metrias da informação abordadas na revisão de literatura possuem dados dos fenômenos informacionais, mas só aqueles que puderem ser trabalhados através de modelagens matemáticas e estatísticas é que possuirão modelos de projeção e previsão.

No caso da economia, as estimativas quantitativas da teoria proporcionam informações valiosas para a formulação da política econômica. Conhecendo o modelo (função, parâmetros,

forma da curva etc.), podemos prever o curso futuro de fenômenos econômicos como renda, despesas de consumo e emprego após alterações de políticas governamentais.

No caso da presente tese, a análise da regressão dos dados do cenário de produção científica nacional (função produção científica nacional, parâmetros, forma linear ou não linear, análise do r e do r^2) proporciona explicações para as relações entre número de doutores e número de publicações. Por exemplo, o modelo de previsão permite, para a presente tese, análises da produção científica nacional a partir dos erros estocásticos ($\hat{u}_i^2$), ou seja, $\hat{Y}_i = f$ ($\hat{u}_1^2$, $\hat{u}_2^2$, $\hat{u}_3^2$..... $\hat{u}_n^2$). Os erros ($\hat{u}n^2$) indicam o quanto um valor Y se aproxima ou se afasta da estimativa esperada $\hat{Y}_i$. Com essa análise, é possível, por exemplo, estabelecer um *ranking* dos estados brasileiros que apresentam maior eficiência na produção de publicações.

Dito de outra forma, os estados que apresentam $\hat{u}_i^2$ menores se aproximam mais de uma combinação ótima entre número de doutores e publicações. Como nesta tese visamos estudar o cenário de produção científica nacional a partir dos valores conhecidos, manteremos a fase da "definição do modelo de previsão" para a proposta do modelo.

Com base em tudo que foi exposto, o nosso modelo infométrico e cienciométrico conta com os seguintes passos: 1) obtenção dos dados, para definir quais irão compor a variável dependente (Y) e quais irão compor a variável independente (X); 2) análise da correlação e do mapa de dispersão, para medir a força ou o grau de associação linear entre duas variáveis e decidir se vale a pena regredir o modelo; 3) especificação do modelo matemático e estatístico, para especificar a forma exata da relação funcional (forma da curva) entre as duas variáveis e atribuir uma variável de erro; 4) estimação dos parâmetros, para estimar os parâmetros ($ß_1$ e $\hat{\beta}_2$) da função do modelo a partir dos dados coletados; 5) análise da regressão, para verificar a qualidade do ajustamento através da relação funcional linear ou não linear (forma da curva linear ou não linear), isto é, a melhor adequação possível dos parâmetros estimados aos parâmetros populacionais; 6) definição do modelo de previsão, para se chegar aos erros estocásticos ($\hat{u}_n^2$) e construir o modelo de previsão da produção científica nacional a partir da análise dos erros estocásticos ($\hat{u}_i^2$), seguindo a seguinte função geral: $\hat{Y}_i = f$ ($\hat{u}_1^2$, $\hat{u}_2^2$, $\hat{u}_3^2$..... $\hat{u}_n^2$).

5 UM ESTUDO SOBRE METRIA DA PRODUÇÃO CIENTÍFICA BRASILEIRA

Em sua dissertação de mestrado, Santos (2011) apresentou métodos de análise de dados da produção científica brasileira utilizando informação da produção bibliográfica, elaborada segundo unidades da Federação para pesquisadores doutores, 2005-2008, Censo 2008. Essas informações foram originadas do Plano Tabular do CNPq e do relatório do MCT.

Todas as observações alcançadas por Santos (2011), preocupado com a responsabilidade social e o enriquecimento para a área da CI, partem da premissa de que a apresentação dessas informações de produção, nos veículos de divulgação e a sua posterior análise, não retratam a realidade e singularidade de estados e regiões do país. O formato de divulgação utilizado na apresentação desses dados, através de tabelas e gráficos, se faz útil, mas simplório demais para uma análise apurada da real produtividade nacional.

Macias-Chapula (1998) insere a produção científica num grande sistema social que é a ciência, com as funções de disseminar conhecimentos, assegurar a preservação de padrões e atribuir crédito e reconhecimento àqueles cujos trabalhos têm contribuído para o desenvolvimento dos diferentes campos das ciências. Disso resulta que a divulgação da informação científica é muito importante para ser tratada de forma parcial.

O Portal do Plano Tabular (http://dgp.cnpq.br/planotabular/) tem a finalidade de divulgar o perfil da pesquisa no Brasil, em termos quantitativos, organizando esses dados em tabelas configuradas dinamicamente pelo usuário. Esse sistema possibilita o cruzamento de indicadores, capazes de gerar um número muito grande de diferentes tabelas, que podem ser salvas em planilhas ou em arquivos de texto para futuras consultas. Isso facilita o acesso aos dados de produção e, com isso, o processo de pesquisa na área de metria da informação.

Conforme informação do Portal, e como já abordado aqui, o Plano Tabular se apoia nos conjuntos básicos de dados (unidades de análise), que possibilitam que essa extração seja inserida naquilo que Souza e Duarte (2011) chamam de “ciência cíclica”, que é aquela que está sempre em desenvolvimento e que leva a novas descobertas. Nesse aspecto, as autoras afirmam que a ciência se fundamenta em pesquisas científicas capazes de formar instrumentos responsáveis pela validação de tudo o que ela desenvolve e capazes de comunicar os resultados adquiridos à sociedade.

Vale lembrar que da conjunção das informações do Plano Tabular, originadas do cadastramento dos grupos de pesquisa, da base de currículos Lattes e do Sistema Coleta/Capes, onde Santos (2011), há um destaque para a fala do Portal, que explica que o fato de o inventário da produção científica, tecnológica e artística dos grupos ser construído a partir das informações

existentes nos currículos Lattes dos pesquisadores e estudantes tem como consequência quase sempre as duplas ou múltiplas contagens no número de produções. Em outras palavras, a produção CT&A dos grupos é sempre apresentada por uma *proxy*, que é a soma das produções individuais de seus componentes. Pesquisadores que participam de mais de um grupo de pesquisa terão a totalidade de sua produção remetida a cada um dos grupos de que participa.

> Nessa perspectiva, [o da existência de um portal de onde se possa extrair dados] a ciência exerce um compromisso com a sociedade, no sentido de expandir os horizontes. Por isso é fundamental publicar os resultados das pesquisas, visto que a Ciência se faz através de estudos e de testes exaustivos, que irão validar todo o trabalho de pesquisa gerada por um pesquisador ou grupo de pesquisa (SOUZA; DUARTE, 2011, p. 155).

O estudo de Santos (2011) teve como base a análise dos seguintes indicadores de produção científica disponibilizados pelo MCT: artigos completos publicados em periódicos especializados de circulação nacional, artigos completos publicados em periódicos especializados de circulação internacional, trabalhos completos publicados em anais de eventos, livros publicados, capítulos de livro publicados, outras publicações bibliográficas, resumos de trabalhos publicados em periódicos especializados, resumos de trabalhos publicados em anais de eventos.

Em seu estudo, Santos (2011) aplicou duas abordagens: técnicas univariadas (índices *per capita* e mapas temáticos) para comparar a produção dos estados e técnicas de análise multivariadas (componentes principais, faces de Chernoff e lógica difusa) para obter novas comparações. Entre os resultados obtidos, destaca-se a semelhança dos esforços produtivos nacionais, com uma larga proporção da literatura científica sendo produzida por um pequeno número de autores, enquanto um grande número de pequenos produtores se iguala, em produção, ao reduzido número de grandes produtores.

O objetivo desta seção é apresentar um estudo pré-existente sobre métodos de análise de dados da produção científica brasileira utilizando informação da tabela de produção bibliográfica, segundo unidades da Federação para pesquisadores doutores, 2005-2008, Censo 2008, cujas informações são originadas do Plano Tabular do CNPq. Vale ressaltar que os gráficos serão retomados para a análise com a utilização do método proposto por esta tese, o MQO.

5.1 VERIFICANDO A CORRELAÇÃO ENTRE AS VARIÁVEIS

Para entender a relação entre as variáveis, número de doutores e dados de produção científica apresentadas na Tabela 7, Santos (2011, p. 41) verificou o padrão de dispersão e aplicou a análise da correlação entre a variável "doutor" e as variáveis de produção (pares), observando o gráfico da reta de regressão e o coeficiente de correlação.

Tabela 6 – Produção bibliográfica segundo unidades da Federação de pesquisadores doutores, 2005-2008, Censo 2008

UF	Total de autores	Artigos completos publicados em periódicos especializados		Trabalhos completos publicados em anais de eventos	Livros ou capítulos de livro publicados		Outras publicações bibliográficas	Resumos de trabalhos publicados em	
		Nacional	Internacional		Livros	Capítulos de livros		Periódicos especializados	Anais de eventos
Acre	125	320	174	234	23	179	698	3	808
Alagoas	567	1344	1149	2817	160	835	1820	87	4137
Amapá	54	98	154	94	22	33	256	6	458
Amazonas	966	2418	2848	2836	326	1578	4410	136	8124
Bahia	3040	9935	7672	11215	896	5225	14234	501	22142
Ceará	1504	5960	5324	7711	434	2910	7621	290	14814
Distrito Federal	2009	7761	6202	9507	942	4885	12926	305	16098
Espírito Santo	753	2422	1632	4248	209	1545	3584	65	5810
Goiás	1367	5359	3721	6111	437	2619	8457	266	10865
Maranhão	410	1502	1491	1316	115	527	1582	157	3992
Mato Grosso	766	2623	1265	2646	276	1056	3456	73	6154
Mato Grosso do Sul	1088	4559	2425	5280	385	2124	8206	114	8978
Minas Gerais	6992	31778	23117	35499	2433	12552	37260	1770	66079
Pará	1023	2925	2470	4353	329	1995	4682	74	7737
Paraíba	1565	6624	3789	12449	430	2650	7078	260	11308
Paraná	5324	22782	15431	29746	1728	8341	29212	1224	47301
Pernambuco	2588	9636	6779	13963	730	4868	13633	520	19524
Piauí	406	1501	1008	1293	111	675	2902	108	3257
Rio de Janeiro	9200	29640	34190	42092	3400	18503	37513	1565	64942
Rio Grande do Norte	1035	3596	2548	6251	302	1544	4469	205	9450
Rio Grande do Sul	6491	29292	22705	37713	2502	15622	35713	1443	63013
Rondônia	116	319	209	400	35	233	650	4	595
Roraima	154	686	313	329	39	222	1160	5	1287
Santa Catarina	2933	10615	7642	19607	1159	4978	12657	308	20764
São Paulo	21014	80603	91929	96572	6830	45787	107170	8090	218998
Sergipe	515	1786	1332	3107	219	792	2318	26	5577
Tocantins	266	1135	608	800	91	278	1785	35	1709

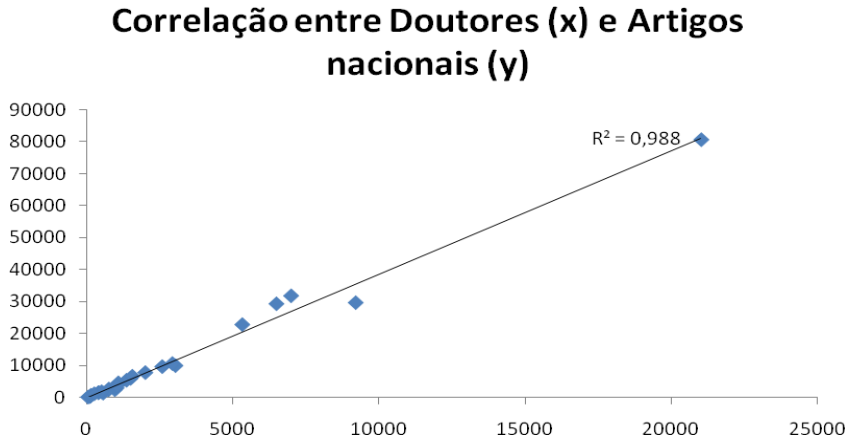
Fonte: Santos (2011, p. 41, Tabela 6.3.1)

O objetivo de Santos (2011) com a análise de dispersão foi verificar a curva que mais se aproximava dos dados e, ao analisar os pares de variáveis, conhecer se entre estas existia alguma possível dependência, isto é, se valores altos/baixos de uma das variáveis estão correlacionados com os valores altos/baixos da outra. Uma correlação alta (próxima de 1)

indicaria forte dependência linear entre as variáveis. Nesse caso, os pontos no diagrama de dispersão espalham-se em torno de uma reta. O autor afirma que nos diagramas de dispersão produzidos, a variável números de doutores foi representada pelo eixo x (horizontal) e as demais se revezaram no eixo y (vertical).

Na primeira análise, Santos (2011) utilizou os dados sobre artigos científicos publicados em revistas nacionais e obteve uma reta de regressão com os pontos distribuídos praticamente ao longo da reta, apresentando forte associação entre as variáveis. O diagrama de dispersão gerado a partir dos dados apresentou a existência de relacionamento entre as variáveis, com altos valores de uma das variáveis associados a altos valores da outra variável, com correlação positiva. O coeficiente de correlação, que marcou 0,99, também indica existir uma forte correlação entre a variável artigos nacionais e o número de doutores. O r^2 indica que 98,80% da variação do número de artigos nacionais é explicado pelo número de doutores em cada estado.

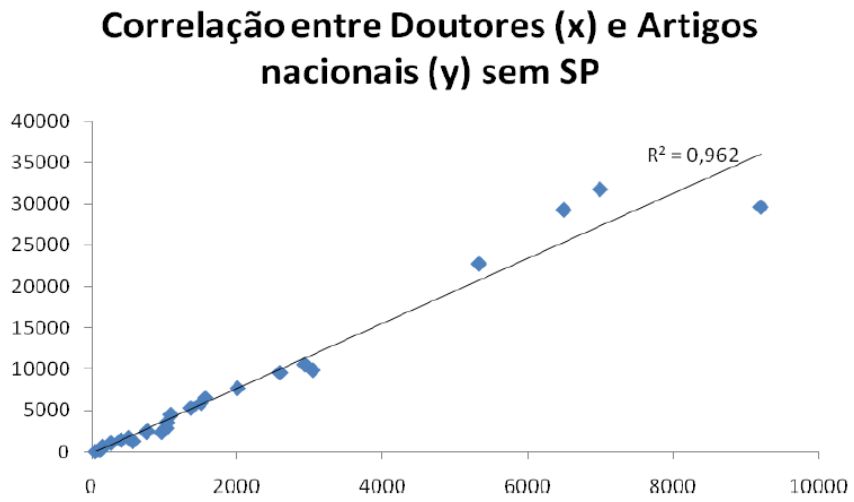
Gráfico 6 – Diagrama de dispersão artigos publicados em periódicos nacionais *versus* doutores



Fonte: Santos (2011, p.42, gráfico 6.3.1)

Segundo Santos (2011), o ponto que aparece no alto do gráfico em destaque representa o estado de São Paulo que, em números absolutos, é o maior produtor e também o maior em concentração de doutores do país. Para dissipar qualquer dúvida sobre um possível erro ou desvio causado pela distância do ponto gerado pelos dados do estado de São Paulo, a análise foi refeita omitindo esses dados para verificar se, de alguma forma, São Paulo estava influenciando na análise da correlação, uma vez que este aparece isolado.

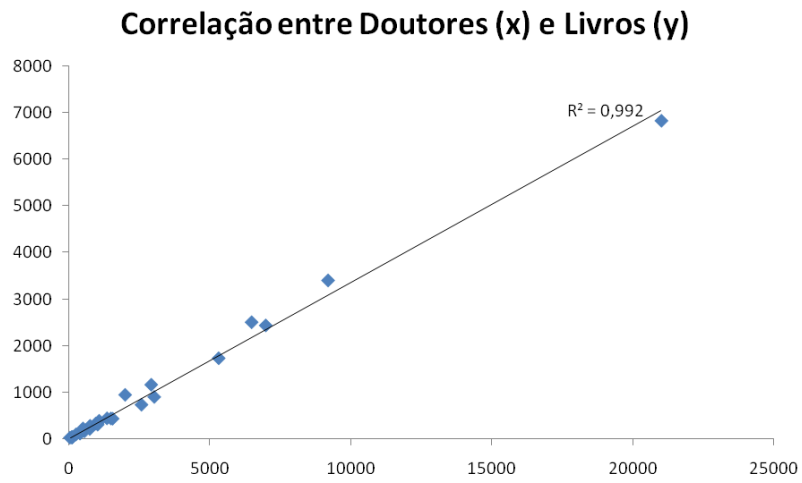
Gráfico 7 – Diagrama de dispersão artigos publicados em periódicos nacionais *versus* doutores sem São Paulo



Fonte: Santos (2011, p. 43, gráfico 6.3.2)

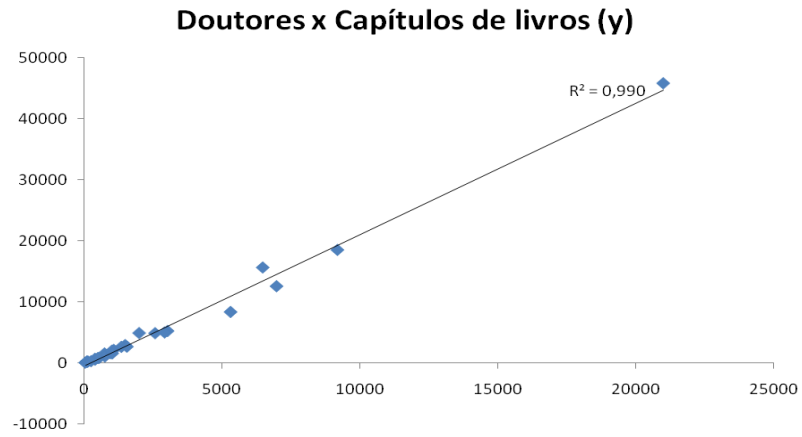
Ainda segundo Santos (2011), a segunda análise constatou que realmente existe uma correlação entre o número de doutores e a produção de artigos nacionais que explica 96,20% da produção de artigos, de modo que São Paulo não estava influenciando o resultado. Para uma maior segurança, todas as demais variáveis de produção sofreram a mesma observação sem os dados do estado de São Paulo e os resultados obtidos foram positivos para uma correlação entre o número de doutores e os demais dados da produção científica. O autor ainda destacou a publicação de livros e capítulos, que obteve coeficiente de correlação de 0,99 e r^2 de 99,20%. Ainda, capítulos de livros teve correlação acima de 0,99 e r^2 de 99% em relação à variável "número de doutores", conforme os seguintes gráficos.

Gráfico 8 – Diagrama de dispersão livros publicados *versus* doutores



Fonte: Santos (2011, p. 45, gráfico 6.3.5)

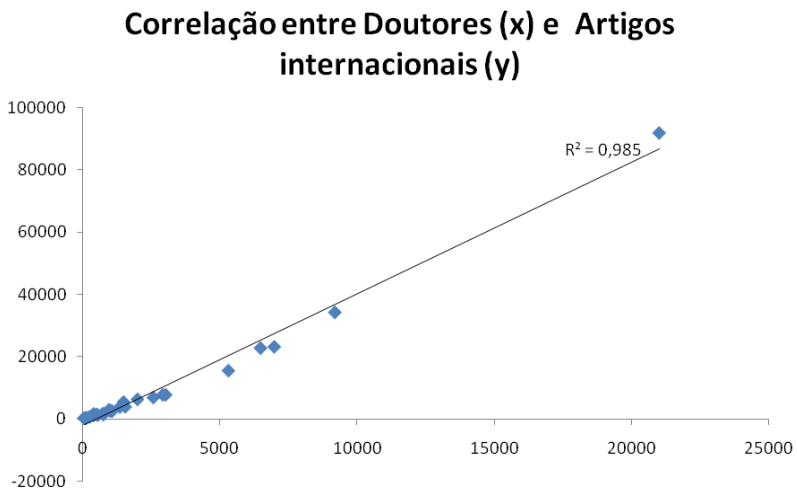
Gráfico 9 – Diagrama de dispersão capítulos de livro publicados *versus* doutores



Fonte: Santos (2011, p. 45, gráfico 6.3.6)

Para a correlação entre o número de doutores e a produção de artigos em periódicos internacionais, Santos (2011) afirma que apresentou um r^2 de aproximadamente 0,99, isso explica que 99% da variação da produção em periódicos internacionais se deve ao número de doutores envolvidos nesse processo e que existe dependência linear entre estas.

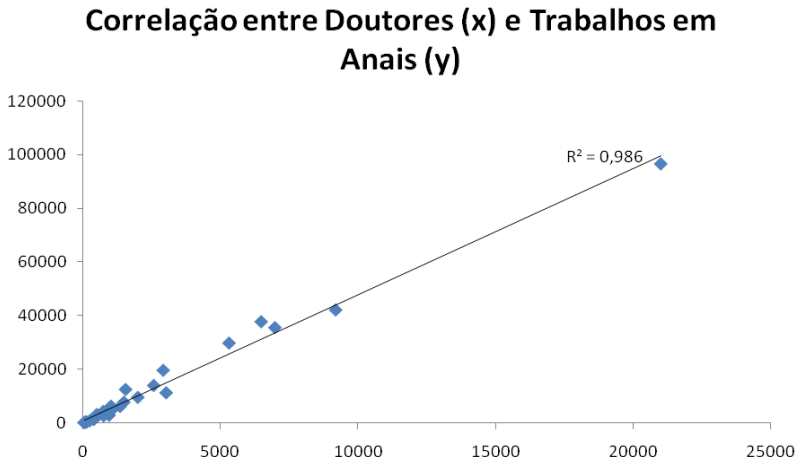
Gráfico 10 – Diagrama de dispersão artigos publicados em periódicos internacionais *versus* doutores



Fonte: Santos (2011, p. 44, gráfico 6.3.3)

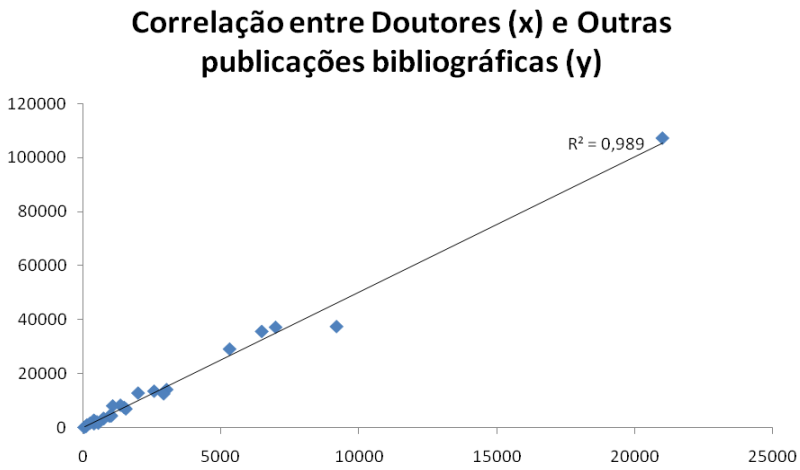
Os dados da produção de trabalhos completos em anais de eventos (gráfico 11) apresentaram correlação de 0,99 e r^2 de 98,60%, que indicam, segundo o autor, que os dados da produção científica têm forte correlação com o número de doutores em cada estado produtor. Os gráficos seguintes ajudam a identificar nos demais indicadores de produção científica a forte correlação positiva já apontada nas primeiras análises de correlação.

Gráfico 11 – Diagrama de dispersão trabalhos completos publicados em anais de eventos *versus* doutores



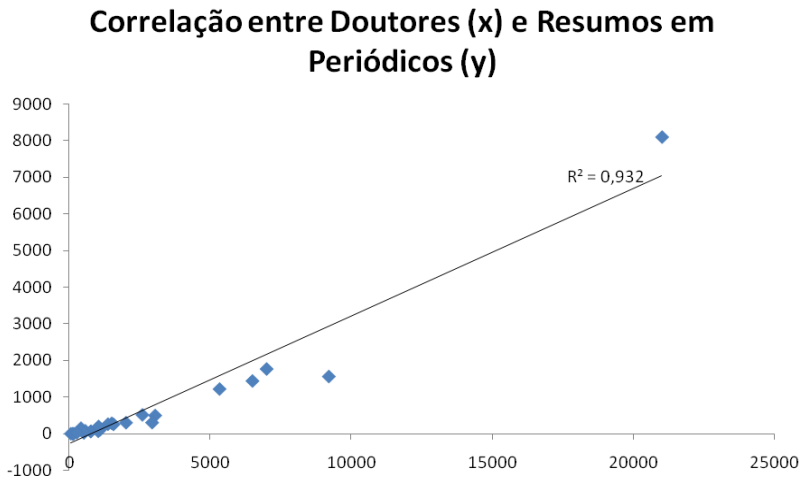
Fonte: Santos (2011, p. 44, gráfico 6.3.4)

Gráfico 12 – Diagrama de dispersão outras publicações bibliográficas *versus* doutores



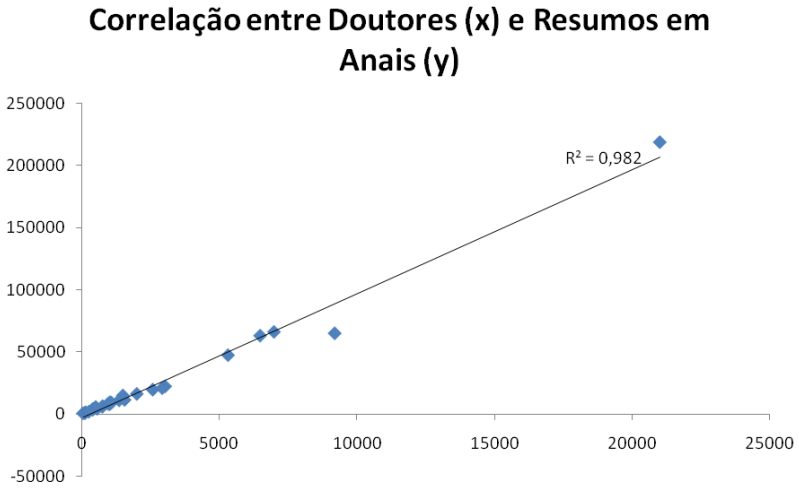
Fonte: Santos (2011, p. 45, gráfico 6.3.7)

Gráfico 13 – Diagrama de dispersão resumos de trabalhos publicados em periódicos especializados *versus* doutores



Fonte: Santos (2011, p. 45, gráfico 6.3.8)

Gráfico 14 – Diagrama de dispersão resumos de trabalhos publicados em anais de eventos *versus* doutores



Fonte: Santos (2011, p. 46, gráfico 6.3.9)

Santos (2011) afirmou que a análise preliminar da análise de correlação e determinação foram as garantias para a aplicação de técnicas de análise multivariada, que auxiliaram no processo de conhecer melhor a relação entre produção e produtores. Entender também essa relação proporcionou alicerces para a construção dos índices *per capita* que relacionou esses indicadores. Com o índice, foi possível determinar um novo olhar para os dados quantitativos de produção, agora proporcional ao número de doutores em cada estado, o que facilitou o entendimento dessa produção a partir de seus atores responsáveis por pesquisas fomentadas em todo o país.

Santos (2011) salienta, ainda, que seu estudo deve contribuir tanto para a CI quanto para a biblioteconomia, por demonstrar a importância de se utilizar métodos e procedimentos adequados à natureza do estudo que se quer desenvolver, além de demonstrar o quanto é possível e produtivo se ampliar as possibilidades de análise de dados sobre os produtos e serviços desenvolvidos nas bibliotecas e outros serviços de informação. Por fim, deve alargar os horizontes da CI para pesquisas semelhantes.

6 DELINEAMENTO METODOLÓGICO DA PESQUISA

O tema "produção científica e modelos de análises" foi delimitado, indo ao encontro das medidas do MCT, para implementar mecanismos que gerenciem o controle do processo de produção científica ao retratar o cenário nacional e estudar os modelos estatísticos que permitam a previsibilidade da produção bibliográfica.

Este estudo é justificado pelo interesse em auxiliar a construção de instrumentos que permitam um maior suporte à avaliação e às decisões, principalmente das agências de fomento e, assim, racionalizar e flexibilizar tanto a aplicação de recursos públicos quanto a definição de políticas nos estados e na Federação. Para isso, o objetivo geral foi analisar os processos métricos da produção científica brasileira (artigos nacionais, artigos internacionais, anais de eventos e livros) através do MQO frente aos modelos tradicionais de metrias da informação, com os objetivos específicos de: a) discorrer sobre o percurso histórico e de aplicação das metrias que a CI vem construindo, desde a mais primordial de todas, a bibliometria, oriunda da biblioteconomia, passando pelas visões modernas como a cienciometria até a informetria; b) explicar como a econometria constrói o seu modelo de análise, que é utilizado para pesquisas na economia e, ao mesmo tempo, refletir como esse método pode ser trazido para as metrias da informação; c) explicar o método de estimação por MQO para a análise de regressão; e d) expor o método de estimação por MQO, aplicado à análise de regressão a que se propõe esta tese.

Para atingir os objetivos foram utilizados os dados da produção científica brasileira do Censo 2010 do Portal do Plano Tabular do CNPq. Infelizmente, o portal foi descontinuado e não há mais, para anos posteriores, a disponibilização desses dados para coleta de forma direta no portal, sendo necessária a solicitação via e-mail ao Diretório dos Grupos de Pesquisa do CNPq. Com isso, finalizou-se a dinâmica de o próprio pesquisador gerar suas tabelas. Atualmente, o Diretório dos Grupos de Pesquisa disponibiliza apenas algumas tabelas (com dados até 2016) que não atendem à necessidade de dados dessa pesquisa (produção científica brasileira por unidade da Federação) e isso provavelmente dificultará sua replicação com dados mais atuais em outras pesquisas.

A principal variável nesta pesquisa é a produção científica brasileira que foi dividida em dependentes Y_n (artigos nacionais, artigos internacionais, anais de eventos e livros) e variável independente X (doutores). Para tanto, foi aplicado o método MQO, bem como a regressão e correlação, para gerar os modelos de previsão, uma para cada tipo de produção científica, sendo: função "artigos publicados em periódicos nacionais (Y) e doutores autores (X)"; função "artigos

publicados em periódicos internacionais (Y) e doutores autores (X)", função "anais de eventos (Y) e doutores autores (X)"; e função "livros (Y) e doutores autores (X)".

6.1 MÉTODO DE ABORDAGEM, MÉTODO DE PROCEDIMENTO E CLASSIFICAÇÃO DA PESQUISA

A abordagem lógica desta tese empregou o processo de investigação científica dos fatos com a construção de modelos de previsão da produção científica brasileira, dentro do necessário grau de abstração, o que possibilitou decidir acerca do alcance da investigação, das regras de explicação dos fatos e da validade de suas generalizações.

Este estudo abordou os dados a partir de técnicas, dentro da teoria das probabilidades, que permitem construir proposições de caráter probabilístico acerca da população de pesquisadores doutores, partindo da observação de alguns dos seus elementos, aqui os dados do Censo da Produção Científica (2007-2010), de acordo com conceitos da estatística indutiva ou inferencial. Na lógica indutiva, a generalização deriva de observações de casos da realidade concreta. Nessa perspectiva, a pesquisa utilizou os métodos de inferência estatística, propostas e cálculos, para inferir sobre os parâmetros dessa população (doutores) e sua produção científica, isto é, permitiu, com determinado grau de probabilidade, generalizar a essa população certas conclusões, por comparação com os resultados amostrais (Censo 2010, dados 2007-2010), entendendo que a:

> Indução é um processo mental por intermédio do qual, partindo de dados particulares, suficientemente constatados, infere-se uma verdade geral ou universal, não contida nas partes examinadas. Portanto, o objetivo dos argumentos indutivos é levar a conclusões cujo conteúdo é muito mais amplo do que o das premissas nas quais se basearam (LAKATOS; MARCONI, 2003, p. 86).

Resumindo, o objetivo da aplicação nesta pesquisa da inferência estatística foi determinar, a partir dos dados das observações, qual era a distribuição populacional, ou seja, qual é a distribuição da variável aleatória que caracteriza a população e, com isso, introduzir ordem às conclusões da evidência fornecida pelos dados, deduzindo-as como consequência, conclusão ou probabilidade. A estatística inferencial aborda dois tipos de problemas fundamentais (estimação de parâmetros de uma população e o teste de hipóteses) nesse estudo por ser aplicado ao MQO na construção do modelo preditivo ajustado ao problema de estimação de parâmetros.

Assim como apresentado ao longo da revisão de literatura, ao explicar o método dos MQO, a tese foi orientada pelo método estatístico, estabelecido por Adolphe Quetelet,

astrônomo, matemático, demógrafo, estatístico e sociólogo do século XIX. A utilização desse método significa a redução de aspectos quantitativos e de manipulação estatística dos fenômenos sociológicos, políticos, econômicos etc. Segundo Eknoyan (2008) e Pich (2013, p. 855), com base em probabilidades e estatística, Quetelet acreditava ser possível estender seu uso para todo tipo de fenômeno humano no que chamava de "física social" em sua publicação, de 1835, "Sur l'homme et le développement de ses facultés" ou "Essai de physique sociale". Introduziu a ideia do *average man* (homem médio). Pelo empenho em entender o funcionamento da sociedade, ficou conhecido como um dos estatísticos-sociais mais importantes de sua época. Quetelet (1835, p. 162), sobre estatística de crimes, reforça a apropriação do uso estatístico, quando explica que a interpretação dos dados (incluindo correlações) são possíveis mesmo que seus valores (quantitativos) sejam desconhecidos.

Desde então, o método estatístico assegura seu papel em diversas áreas do conhecimento e sua aplicação permite comprovar as relações dos fenômenos entre si, além de obter generalizações sobre sua natureza, ocorrência ou significado. O papel do método estatístico é, antes de tudo, fornecer uma descrição quantitativa da sociedade, considerada como um todo organizado, lembrando que, nesta tese, o método seguiu os contornos da econometria e suas aplicações nas ciências sociais e econômicas dentro da aplicação do MQO.

Sobre estatística inferencial, é possível trazer também para este estudo, agora com um viés mais aprofundado em raízes filosóficas de bases empíricas (claramente utilizadas por Quetelet), positivistas e até mesmo positivistas lógicas, o método indutivo, uma vez que os modelos de previsão com base no MQO objetivaram uma generalização para análises da produção científica brasileira. Embora possa haver algum entendimento sobre aspectos dedutivos nesta pesquisa, uma vez que argumentos matemáticos estejam mais associados ao modelo dedutivo por estes explicarem as observações, Marcone e Lakatos (2003, p. 93) afirmam que "a relação entre a evidência observacional e a generalização científica é de tipo indutivo. [...] Por sua vez, os argumentos matemáticos são dedutivos". E afirmam ainda, sobre a distinção entre os modelos lógicos, que

> [...] o dedutivo tem o propósito de explicar o conteúdo das premissas; o indutivo tem o desígnio de ampliar o alcance dos conhecimentos. Analisando isso sob outro enfoque, diríamos que os argumentos dedutivos ou estão corretos ou incorretos, ou as premissas sustentam de modo completo a conclusão ou, quando a forma é logicamente incorreta, não a sustentam de forma alguma; portanto, não há graduações intermediárias. Contrariamente, os argumentos indutivos admitem diferentes graus de força, dependendo da capacidade das premissas de sustentarem a conclusão. Resumindo, os argumentos indutivos aumentam o conteúdo das premissas, com sacrifício da precisão, ao passo que os argumentos dedutivos sacrificam a ampliação do conteúdo para atingir a 'certeza' (MARCONE; LAKATOS, 2003, p. 93).

A aplicação do método indutivo é reforçada quando se tenta responder às duas proposições feitas por Marcone e Lakatos (2003, p. 88): a) "temos expectativas e acreditamos que exista certa regularidade nas coisas, e por este motivo, o futuro será como o passado"; b) "São, principalmente, as observações feitas no passado. Em análises anteriores esse comportamento não se modificou, o que gerou a expectativa de certa regularidade no mundo, no que se refere a esses fatos e fenômenos. Por este motivo, analisando-se vários casos singulares do mesmo gênero, estende-se a todos (do mesmo gênero) as conclusões baseadas nas observações dos primeiros, através da 'constância das leis da natureza' ou do 'princípio do determinismo'". Isso equivale, no âmbito desta tese, ao seguinte questionamento: o comportamento da produção científica brasileira e a relação entre produtos e produtores devem permanecer a mesma?

Mas o emprego do método indutivo exige critérios e, segundo Marcone e Lakatos (2003, p. 87),

> Devemos considerar três elementos fundamentais para toda indução, isto é, a indução realiza-se em três etapas (fases): a) observação dos fenômenos - nessa etapa observamos os fatos ou fenômenos e os analisamos, com a finalidade de descobrir as causas de sua manifestação; b) descoberta da relação entre eles – na segunda etapa procuramos por intermédio da comparação, aproximar os fatos ou fenômenos, com a finalidade de descobrir a relação constante existente entre eles; c) generalização da relação – nessa última etapa generalizamos a relação encontrada na precedente, entre os fenômenos e fatos semelhantes, muitos dos quais ainda não observamos (e muitos inclusive inobserváveis).

Outro ponto importante a destacar é que esta tese aplica as formas de indução incompleta ou científica que, segundo os autores (p. 87), foi "criada por Galileu e aperfeiçoada por Francis Bacon. Não deriva de seus elementos inferiores, enumerados ou provados pela experiência, mas permite induzir, de alguns casos adequadamente observados (sob circunstâncias diferentes, sob vários pontos etc.)". Isso seguindo as regras de que os casos particulares devem ser provados e experimentados na quantidade suficientemente necessária, e que é possível afirmar, com certeza, que a própria natureza da coisa (fato ou fenômeno) foi o que provocou a sua propriedade (ou ação), além de entender que quanto maior a amostra, maior a força indutiva do argumento, e quanto mais representativa a amostra, maior a força indutiva do argumento.

O entendimento da aplicação dos dois métodos, indutivo e estatístico, fica explicitado pelas falas de Marcone e Lakatos (2003, p. 56) que, metodologicamente, categorizam os "métodos de abordagem e de procedimento utilizados": o método seria o indutivo e o procedimento seria o estatístico. Gil (2008, p. 9-15) também caracteriza os "métodos que

proporcionam as bases lógicas da investigação", aqui o indutivo, e os "métodos que indicam os meios técnicos da investigação", no caso desta tese, o método estatístico. Sobre a indicação dos meios técnicos da investigação, Gil (2008, p. 15) explica que:

> Estes métodos têm por objetivo proporcionar ao investigador os meios técnicos para garantir a objetividade e a precisão no estudo dos fatos sociais. Mais especificamente, visam fornecer a orientação necessária à realização da pesquisa social, sobretudo no referente à obtenção, processamento e validação dos dados pertinentes à problemática que está sendo investigada.

Autores como Prodanov e Freitas (2013, p. 26-36), que também citam Gil (2008) e Marcone e Lakatos (2007), dividem os procedimentos metodológicos em "métodos de abordagem – bases lógicas da investigação", nos quais incluem o método indutivo e "métodos de procedimentos – meios técnicos da investigação", nos quais apresentam o método estatístico, o que esclarece esses dois momentos da pesquisa. O primeiro determina o caminho, a forma, o modo de pensamento, ou seja, é a forma de abordagem em nível de abstração dos fenômenos, ou ainda, o conjunto de processos ou operações mentais empregados na pesquisa (PRODANOV; FREITAS, 2013, p. 26). E o segundo, que é menos abstrato, também chamado de "específico" ou "discreto", são etapas seguidas pelo pesquisador no processo da investigação do fenômeno (PRODANOV; FREITAS, 2013, p. 36).

Findada a apresentação do percurso do método desta tese, é necessário fazer a classificação da pesquisa. Nesse processo, Prodanov e Freitas (2013, p. 50) elaboraram um quadro que permite compreender as etapas clássicas de classificação a partir do ponto de vista da sua natureza, de seu nível, dos procedimentos técnicos e da forma de abordagem do problema.

A tese, sob o ponto de vista da sua natureza, é uma pesquisa aplicada, pois os conhecimentos gerados aqui podem e devem ser aplicados aos estudos relativos ao fomento em pesquisa científica que visem: prever a produção no cenário nacional; às práticas e técnicas bibliométricas; aos usos de métodos estatísticos e matemáticos em CI.

A pesquisa, sob o ponto de vista de seu nível, é descritiva, pois o fenômeno da produção científica foi registrado e descrito a partir das características da população e do estabelecimento de relações entre suas variáveis, procurando classificar, explicar e interpretar as relações que ocorrem. Apesar de existir uma aproximação com as pesquisas exploratórias, por proporcionar uma nova visão da questão da produção científica (produtos e produtores) no Brasil, e também por ultrapassar a identificação das relações entre as variáveis, procura estabelecer a natureza dessas relações.

Quanto aos procedimentos técnicos, ou seja, a maneira pela qual foram obtidos os dados necessários para a elaboração da pesquisa, é um estudo de caso, pois consistiu em analisar informações sobre a produção científica brasileira a partir dos dados do Censo 2010 do CNPq, a fim de estudar aspectos da relação entre as variáveis e propor um modelo de previsão com base em MQO e levando em consideração que seus resultados não devem extrapolar para outras relações e realidades.

> [A] dificuldade de generalização: a análise de um único ou mesmo de múltiplos casos fornece uma base muito frágil para a generalização científica. Todavia, os propósitos do estudo de caso não são os de proporcionar o conhecimento preciso das características de uma população a partir de procedimentos estatísticos, mas, sim, o de expandir ou generalizar proposições teóricas. O maior risco do estudo de caso único é que a explicação científica mostre-se frágil, devido a possíveis incidências de fenômenos encontrados apenas no universo pesquisado, o que pode comprometer a confiabilidade dos achados da pesquisa. Em qualquer das alternativas, o pesquisador deverá compor um cenário que corresponda à teoria que fundamenta a pesquisa e que se revele no estudo do caso, ou seja, construir uma análise que explique e preveja o fenômeno investigado (PRODANOV; FREITAS, 2013, p. 62).

Ainda sobre os procedimentos técnicos, esta tese poderia ser caracterizada também como uma pesquisa *ex-post facto*, mas se enquadra mais como estudo de caso. Segundo Prodanov e Freitas (2013, p. 65), "A pesquisa *ex-post facto* analisa situações que se desenvolveram naturalmente após algum acontecimento" e, devido à sua característica, "permite a investigação de determinantes econômicos e sociais do comportamento da sociedade em geral. Estudamos um fenômeno já ocorrido, tentamos explicá-lo e entendê-lo". Gil (2008, p. 54) explica que "na pesquisa *ex-post facto* a manipulação da variável independente é impossível. Elas chegam ao pesquisador já tendo exercido os seus efeitos. Também não é possível designar aleatoriamente sujeitos e tratamentos a grupos experimentais".

Sobre esse procedimento técnico, Gil (2008, p. 54) ressalta ainda que:

> Apesar de serem óbvias as limitações da pesquisa *ex-post facto*, isto não significa que devam ser descartadas como não científicas. Muitos problemas nas ciências sociais são problemas *ex-post facto* e requerem, portanto, pesquisas *ex-post facto* simplesmente porque as variáveis independentes não são manipuláveis. O que se faz necessário nesses estudos é considerar outras variáveis possivelmente relevantes e controlá-las estatisticamente, sobretudo por meio da análise multivariada. Dessa forma, a provável influência dessas variáveis poderia ser analisada e neutralizada na análise dos resultados da pesquisa.

Como a proposta aqui foi analisar dados do Censo 2010 do CNPq, reconhece-se a impossibilidade do controle sobre a variável "número de doutores" e suas repercussões. Como explicam Prodanov e Freitas (2013, p. 65), para esse procedimento técnico, o objetivo é identificar quais os possíveis relacionamentos entre as variáveis, ou seja, uma análise

correlacional após o fato ter sido consumado. Fato desejado para compreender o comportamento da variável independente.

Sobre a forma de abordagem do problema, a pesquisa é quantitativa por considerar que tudo pode ser quantificável e, a partir desse critério, estabelecer um modelo de previsão da produção científica brasileira. Esta pesquisa foi proposta com base no MQO e o uso de recursos e de técnicas estatísticas (percentagem, média, moda, mediana, desvio-padrão, coeficiente de correlação, análise de regressão etc.) foi fundamental para sua realização.

> Essa forma de abordagem [quantitativa] é empregada em vários tipos de pesquisas, inclusive nas descritivas, principalmente quando buscam a relação causa-efeito entre os fenômenos e também pela facilidade de poder descrever a complexidade de determinada hipótese ou de um problema, analisar a interação de certas variáveis, compreender e classificar processos dinâmicos experimentados por grupos sociais, apresentar contribuições no processo de mudança, criação ou formação de opiniões de determinado grupo e permitir, em maior grau de profundidade, a interpretação das particularidades dos comportamentos ou das atitudes dos indivíduos (PRODANOV; FREITAS, 2013, p. 70).

Com base em Prodanov e Freitas (2013, p. 71), pode-se determinar que, para pesquisas quantitativas, o foco deve ser: a quantidade (quantos, quanto); se suas raízes filosóficas são o positivismo, empirismo, lógico; se suas frases associadas serão experimentais, empíricas, estatísticas; se as metas de investigação são baseadas em predição, controle, descrição, confirmação, teste de hipótese; se o ambiente investigativo é artificial, não natural; o quanto a amostra deve ser grande, ampla, representativa; se a coleta de dados se faz a partir de instrumentos manipulados (escala, teste, questionário etc.); e, por fim, se o modo de análise empregado é o dedutivo (pelo método estatístico).

6.2 UNIVERSO E AMOSTRA

A divulgação de dados sobre a produção científica nasce da necessidade de saber e entender a situação atualizada por parte dos produtores e quais áreas exigem maior atenção para investimentos ou exploração por parte dos órgãos de fomento. Com essa finalidade foi criado o Diretório dos Grupos de Pesquisa do Portal Plano Tabular do CNPq. Para explicar sua importância e a confiabilidade dos dados utilizados nesta pesquisa, foi necessário conhecer sua construção e finalidades. Este estudo se valeu da forma como os dados de produção científica são disponibilizados no Portal Plano Tabular do CNPq, entendendo que existe uma

responsabilidade daqueles que divulgam resultados de dados sobre pesquisa no Brasil. As informações a seguir foram coletadas do Portal Diretório dos Grupos de Pesquisa[3].

Primeiramente, cabe o entendimento sobre o método estatístico e sua importância para a manipulação dos dados disponibilizados Portal Plano Tabular do CNPq que foram utilizados nesta tese. Segundo Marcone e Lakatos (2003, p. 108), o método estatístico permite "a redução de fenômenos sociológicos, políticos, econômicos etc, a termos quantitativos e a manipulação estatística, que permite comprovar as relações dos fenômenos entre si, e obter generalizações sobre sua natureza, ocorrência ou significado".

Gil (2008, p. 17) ainda afirma que "a utilização de testes estatísticos, torna-se possível determinar, em termos numéricos, a probabilidade de acerto de determinada conclusão, bem como a margem de erro de um valor obtido". E continua reconhecendo o razoável grau de precisão que o método proporciona e que faz ele ser bastante aceito pelos pesquisadores com preocupações de ordem quantitativa.

Em geral, para além do método estatístico, para Prodanov e Freitas (2013, p. 52), a própria pesquisa descritiva "visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis". A aplicação do método acontece pela identificação das variáveis e da definição de hipóteses sobre suas relações.

Sobre as variáveis, Marcone e Lakatos (2003, p. 137) trazem o conceito de Lipset e Bendix (1974), de que essas são operacionais, explicando que:

> [...] uma variável pode ser considerada como uma classificação ou medida; uma quantidade qua varia; um conceito operacional, que contém ou apresenta valores; aspecto, propriedade ou fator, discernível em um objeto de estudo e passível de mensuração. Os valores que são adicionados ao conceito operacional, para transformá-lo em variável, podem ser quantidades, qualidades, características, magnitudes, traços etc., que se alteram em cada caso particular e são totalmente abrangentes e mutuamente exclusivos. Por sua vez, o conceito operacional pode ser um objeto, processo, agente, fenômeno, problema etc. (MARCONE; LAKATOS, 2003, p. 137).

O conceito de variável de Gil (2008, p. 79) também perpassa a ideia do processo lógico de operacionalização que exige a "definição teórica da variável e a enumeração de suas dimensões" e, em seguida, outra definição (empírica) "chamada de operacional, fará referência a seus indicadores, ou seja, aos elementos que indicam seu valor de forma prática. A partir daí, torna-se possível a medição desses indicadores, o que possibilitará conhecer o valor da variável".

---

[3] Disponível em: <http://dgp.cnpq.br/censo2004/inf_gerais/index_que_eh.htm>.

Gil alerta também para o fato de que dados tratados isoladamente como variáveis podem ser, na verdade, só aspecto de uma variável mais complexa que nesta tese é apresentada pela produção científica dos estados da Federação, representando a variável que deve ser entendida e o número de doutores e sua produção como os indicadores relevantes para compreender essa variável.

Nesta pesquisa, a variável principal foi dividida em várias dimensões representadas nesta tese pelas variáveis independentes e dependentes, para o tratamento de dados quantitativo, oriundas também da própria dinâmica da coleta dados, processos de contagem e mensuração do Censo do CNPq. Ou seja, esta tese utiliza dados de uma contagem, inclusive, o censo. Tal termo vem do latim *censu*, que é uma das mais antigas formas de contagem utilizado pelos governantes para entender uma determinada realidade. Aparece inclusive na Bíblia, como um dos motivos da viagem de José e Maria até Belém. Amplamente conhecido como contagem de população ou demográfico, censos há muito tempo tiveram seu objetivo ampliado e hoje é um termo que pode ser utilizado para levantamentos que determinem um retrato, ou melhor, um recorte de uma determinada situação em um determinado momento. Com isso, observa-se que o pensamento dedutivo, positivista, há muito faz parte do levantamento de informações para a tomada de decisões.

No Brasil, talvez uma instituição que represente os princípios do modelo indutivo e dedutivo de pesquisa seja o Instituto Brasileiro de Geografia e Estatística (IBGE). O censo realizado a cada 10 anos pelo IBGE é o mais conhecido e, assim, essa grande contagem que mobiliza toda a população do país serve para apresentar um retrato social. Os dicionários Aurélio e Michaelis trazem uma ampliação do conceito e afirmam que pode também significar "rendimento coletável dos cidadãos que serve de base ao exercício de certos direitos políticos" ou "pensão ou renda anual, pela posse de uma terra ou em virtude de um contrato".

Possivelmente, a fim de deixar claro seu trabalho, o IBGE[4] utiliza o termo composto "Censo Demográfico" e explica que "se constituiu no grande retrato em extensão e profundidade da população brasileira e das suas características socioeconômicas e, ao mesmo tempo, na base sobre a qual deverá se assentar todo o planejamento público e privado da próxima década". Tudo isso revela que uma análise somente numérica é apriorística e provida de pouco sentido, pois através dos números, necessita-se chegar a informações políticas, socioeconômicas, de relações contratuais, de planejamento público e privado, portanto, os dados numéricos necessitam de algum nível de apropriação indutivista.

[4] Disponível em: <http://censo2010.ibge.gov.br/sobre-censo.html>.

Logo, é possível afirmar que as instituições públicas de pesquisa e coleta de informações trabalham com pesquisas que mesclam métodos dedutivos e indutivos. Nesse sentido, outro órgão que se destaca para esta pesquisa é o CNPq que, através do Diretório dos Grupos de Pesquisa no Brasil (DGP), mantém um inventário dos grupos de pesquisa científica e tecnológica em atividade no país. E, segundo o próprio DGP (http://lattes.cnpq.br/web/dgp/o-que-e/), é constituído a partir da coleta de informações referentes aos recursos humanos como pesquisadores, estudantes e técnicos, às linhas de pesquisa, às especialidades do conhecimento, aos setores de aplicação envolvidos, à produção científica, tecnológica e artística e às parcerias estabelecidas entre os grupos e as instituições, sobretudo com as empresas do setor produtivo. As informações disponíveis pelo DGP/CNPq permitem descrever os limites e o perfil geral da atividade científico-tecnológica no Brasil e, com os dados de sua base corrente, são realizados censos bianuais (até o momento da pesquisa foi 2010), que são verdadeiras fotografias da pesquisa no Brasil, algo que demonstra a necessidade de análises com aspectos dedutivos e indutivos.

É entendimento que manter um levantamento periódico de dados científicos e manter um sistema atualizado (DGP/CNPq) é um esforço institucional que requer a participação de muitos atores. Apenas para ter uma ideia da necessidade de uma proposta que seja construída coletivamente, para manter os levantamentos e atualizações DGP/Portal Plano Tabular do CNPq, são 551 instituições de pesquisa com 27.523 grupos cadastrados participantes no último censo. Esses grupos fazem parte de universidades, instituições isoladas de ensino superior com cursos de pós-graduação *stricto sensu*, institutos de pesquisa científica e institutos tecnológicos.

A disponibilização dos dados do Portal Plano Tabular do CNPq que permite aos órgãos de fomento e planejamento estabelecer metas e prioridades e que pesquisadores, em quaisquer das áreas do conhecimento, possam aplicar os dados em suas análises, o que inclui esta pesquisa, o que coaduna com o desejo de Quetelet (1834) de ser possível estender o uso de dados estatístico para todo tipo de fenômeno humano e do próprio papel do método estatístico de fornecer um panorama quantitativo da sociedade. Outro aspecto do método é o de oferecer diversas técnicas que permitem ampliar o olhar sobre esses dados brutos disponibilizados pelo Plano Tabular do CNPq e demonstrar a diversidade de suas facetas.

Essa foi, na verdade, a questão de pesquisa que indagou: como analisar a produção científica brasileira (artigos nacionais, artigos internacionais, anais de eventos e livros) utilizando os MQO? E que foi alcançada pelo objetivo geral, quando seu verbo de ação indicou: analisar a produção científica brasileira (artigos nacionais, artigos internacionais, anais de eventos e livros) através dos MQO. Para isso, entendendo, como explica Marcone e Lakatos

(2003, p. 137), que o "universo" é constituído de três níveis (onde ocorrem as observações de fatos, fenômenos, comportamentos e atividades reais, onde encontramos as hipóteses e finalmente, onde surgem as teorias, hipóteses válidas e sustentáveis), foi definido o universo desta tese como a totalidade da produção científica brasileira de doutores por estados, disponível no Plano Tabular do CNPq, e como subconjunto desse universo, por meio do qual foi possível estimar seu comportamento, foi estabelecido que a amostra é o Censo 2010 (2007-2010).

Primeiramente, se faz necessário explicar que, para atender ao modelo econométrico e estatístico, foi estabelecido a variável independente "x" para representar a dimensão "doutores na dinâmica da produção científica". Lembrando que, em estatística, estas são as variáveis determinantes para que ocorra um determinado resultado, ou seja, são os estímulos que condicionam uma resposta. Nesta tese, dentro desse aspecto, o número de doutores cadastrados no Portal Plano Tabular do CNPq representam os principais produtores da produção científica brasileira (Censo 2010). Quanto às variáveis dependentes, foram selecionadas todas aquelas referentes à produção bibliográfica dos autores doutores, onde os Y_n (artigos nacionais, artigos internacionais, anais de eventos e livros) determinam toda a produção bibliográfica da ciência nacional em todas as áreas do conhecimento por estados no Plano Tabular do CNPq (Censo 2010).

De modo geral, a modelagem da amostra, cujos dados estão no item "Coleta e tratamento dos dados", ficou da seguinte forma: X = número de autores doutores por estado para o ano de 2010 e $Y_{n=}$ número de produções bibliográficas para o ano de 2010, gerando as seguintes funções "artigos publicados em periódicos nacionais (Y) e doutores autores (X)", "artigos publicados em periódicos internacionais (Y) e doutores autores (X)", "trabalhos completos publicados em anais de eventos (Y) e doutores autores (X)" e "livros publicados (Y) e doutores autores (X)".

$$\hat{Y}_{Artigos\ nacionais} = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2).$$

$$\hat{Y}_{Artigos\ internacionais} = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2).$$

$$\hat{Y}_{Anais\ \text{de eventos}} = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2).$$

$$\hat{Y}_{Livros} = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2).$$

A construção dessas quatro funções é exposta no capítulo de apresentação e discussão dos resultados e no Apêndice.

6.3 COLETA E TRATAMENTO DOS DADOS

Cabe ressaltar o esforço de coleta do Diretório dos Grupos de Pesquisa no Brasil para apresentar um inventário dos grupos de pesquisa em atividade no país. Através de bases de dados, são disponibilizadas informações sobre os recursos humanos constituintes dos grupos, as linhas de pesquisa em andamento, as especialidades do conhecimento, os setores de atividade envolvidos, a produção científica, tecnológica e artística dos pesquisadores e estudantes que integram os grupos e os padrões de interação com o setor produtivo. Os dados representam grupos localizados em universidades, instituições isoladas de ensino superior, institutos de pesquisa científica, institutos tecnológicos, laboratórios de pesquisa e desenvolvimento de empresas estatais ou ex-estatais, além de algumas organizações não governamentais com atuação em pesquisa.

A disponibilidade dos dados do Diretório foi iniciada no CNPq em 1992 e mantém uma frequência quase bianual. A partir de 2002, o sistema passou à atualização contínua da base de dados, denominada “Base Corrente”, porém mantendo a frequência bianual para a divulgação de resultados. Os censos, como são denominados esses resultados, apresentam informações quantitativas sobre os grupos em suas diversas dimensões e oferecem recursos de buscas textuais sobre as bases de dados.

Através da tecnologia de Data Warehouse, o CNPq, com seu do Diretório dos Grupos de Pesquisa, organiza as diversas bases em um mesmo modelo de dados para a extração de informações padronizadas e atualizadas. O Data Warehouse do Diretório está apoiado em sete conjuntos básicos de informação, compostos pelos dados referentes: aos grupos de pesquisa, aos pesquisadores, aos estudantes, ao pessoal técnico, às linhas de pesquisa e à produção científica, tecnológica e artística e empresas. E, para recuperar informação textual no Diretório dos Grupos de Pesquisa no Brasil, é utilizada a API Lucene baseada em tecnologia Java e de código fonte aberto (OpenSource). A sigla API quer dizer “Application Programming Interface” e significa “Interface de Programação de Aplicativos”, que é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de *software* ou plataforma baseado na *web*, disponível em: http://jakarta.apache.org/lucene.

O Diretório dos Grupos de Pesquisa foi criado para três finalidades principais: ter precisão e rapidez, ser capaz de responder quem é quem, onde se encontra, o que está fazendo e o que produziu recentemente. No âmbito do planejamento e gestão das atividades de Ciência e Tecnologia (C&T), atende às necessidades de instituições, de sociedades científicas e/ou de várias instâncias de organização político-administrativa do país como uma fonte inesgotável de

informação, assim como uma importante fonte na preservação da memória da atividade científico-tecnológica no Brasil.

O portal do Diretório dos Grupos de Pesquisa é composto por oito módulos independentes, dos quais o denominado "Plano Tabular" foi o foco da coleta de dados desta pesquisa. A vantagem que o sistema apresenta para esse levantamento tem como base as informações originadas do cadastramento dos grupos de pesquisa e da base de currículos Lattes e do Sistema Coleta Capes. O sistema ainda apresenta a propriedade de correção das inconsistências próprias de uma captura de dados.

No modelo do Plano Tabular, utilizamos três subconjuntos para organizar as principais variáveis da base de dados divididas em: unidades de análise, que indicam o que o usuário deseja saber; variáveis de filtro, que apresentam como o usuário deseja distribuir a informação; e variáveis de conteúdo, que demonstram qual o nível de detalhamento que o usuário deseja visualizar.

A estrutura do portal foi definida em sete unidades de análise (grupos, linhas de pesquisa, pesquisadores, estudantes, pessoal técnico, setor produtivo e produção científica, tecnológica e artística). E, em cada unidade, estão disponíveis variáveis de filtro correspondentes que, combinadas entre si, estabelecem o conteúdo da tabela. Em alguns casos, esse conteúdo pode, ainda, ser configurado pelo usuário, pela indicação de variáveis de conteúdo. Explica ainda que essa é uma ferramenta versátil e muito útil para seleção e montagem de dados e amostras dos cenários de pesquisa no Brasil.

Para esta tese, bastou selecionar as opções disponíveis no próprio portal e gerar a tabela. O procedimento no portal foi: acessar a aba unidade de análise "Produção Científica, Tecnológica e Artística (C, T&A)", escolher o filtro "Geográfica por Unidade da Federação" e o Censo 2010. Em seguida, escolheu-se a opção "Produções Bibliográficas" (que serão as variáveis dependentes no cálculo da regressão por MQO), que são: artigos publicados em periódicos nacionais, artigos publicados em periódicos internacionais, trabalhos completos publicados em anais de eventos, livros publicados, capítulos de livros publicados e outras publicações bibliográficas. E, por fim, selecionou-se a opção "autores pesquisadores doutores". Em resumo, a coleta de dados ocorreu conforme descrito e o portal gerou uma tabela que equivale à apresentada (8), utilizada no tratamento de dados. O tratamento foi completado com o cálculo das funções de regressão pelo método de MQO utilizando-se a planilha do *software* Excel.

Definimos que a variável independente (X) seria o número de autores doutores por estado para o ano de 2010 e que as variáveis dependentes (Y) seriam as produções bibliográficas

destes para o ano de 2010 (artigos publicados em periódicos nacionais, artigos publicados em periódicos internacionais, trabalhos completos publicados em anais de eventos, livros publicados). Foram construídas ao total quatro funções, que são apresentadas e analisadas na seção seguinte. A tabela abaixo traz os dados estruturados que permitiram o cálculo das funções.

Tabela 7 – Produção bibliográfica segundo UF para pesquisadores doutores, 2007-2010, Censo 2010

UF	Total de autores doutores	Artigos completos em periódicos de circulação nacional	Artigos completos em periódicos de circulação internacional	Trabalhos completos publicados em anais de eventos	Livros
Acre	162	387	206	305	58
Alagoas	760	2.317	1.683	3.568	203
Amapá	65	182	193	221	14
Amazonas	1.112	3.017	3.723	2.981	406
Bahia	3.622	12.121	9.674	13.189	1.068
Ceará	1.975	8.980	7.556	10.259	609
Distrito Federal	2.686	10.814	8.683	11.029	1.141
Espírito Santo	979	3.880	2.486	5.013	309
Goiás	1.775	7.454	5.419	7.786	624
Maranhão	593	2.506	2.058	1.722	171
Mato Grosso	1.075	4.595	2.157	3.957	344
Mato Grosso do Sul	1.497	6.686	3.894	7.726	524
Minas Gerais	9.228	41.159	34.692	44.797	2.858
Pará	1.462	4.701	4.485	6.112	501
Paraíba	2.055	9.347	5.206	14.840	607
Paraná	6.508	28.586	21.122	34.752	2.118
Pernambuco	3.215	12.731	9.673	16.517	940
Piauí	626	2.846	1.672	1.789	146
Rio de Janeiro	10.997	36.693	42.933	44.523	4.076
Rio Grande do Norte	1.527	5.775	3.930	9.042	465
Rio Grande do Sul	7.841	36.627	30.596	42.966	2.846
Rondônia	221	705	723	769	75
Roraima	169	722	322	373	34
Santa Catarina	3.580	14.472	11.203	21.979	1.510
São Paulo	22.922	90.240	108.990	100.455	7.759
Sergipe	824	2.953	2.088	4.689	327
Tocantins	358	1.675	809	1.002	125
TOTAIS	**87.834**	**352.171**	**326.176**	**412.361**	**29.858**

Fonte: Elaboração do autor

Para clarear o entendimento dos processos de tratamentos dos dados, apresentamos um passo a passo da nossa metodologia, que é também o modelo desta tese, resumida no seguinte fluxograma:

Figura 10 - Processo do tratamento dos dados

Fonte: Elaboração do autor

A inserção dos dados foi definida segundo critérios já abordados aqui e definiu a variável dependente (Y), que são os itens de produção científica e que será relacionada com a variável independente (X), que é o número de doutores formando um total de quatro funções.

A obtenção dos dados, bem como a análise da correlação e do mapa de dispersão, foi executada de acordo com a experiência do pesquisador e comparações com exemplos obtidos na literatura.

A especificação do modelo matemático e estatístico também contou com a experiência do pesquisador e exemplos da literatura para especificar a forma da relação funcional (forma da curva) entre as duas variáveis e atribuir uma variável de erro.

A estimação dos parâmetros seguiu o modelo de MQO de Gujarati e Porter (2011). Com isso, a partir da econometria, se aplicou as fórmulas para estimar os parâmetros $\hat{\beta}_1$ e $\hat{\beta}_2$ necessários para determinar a função do modelo.

A análise da regressão verificou a qualidade do ajustamento (forma da curva) e foi determinante para identificar a curva mais adequada ao modelo proposto.

E, por fim, a definição do modelo de previsão, que também seguiu os ensinamentos de Gujarati e Porter (2011), permitindo a construção do modelo da tese, atendendo ao objetivo principal da pesquisa. Dito de outra forma, analisar a partir dos erros estocásticos ($\hat{u}_n^2$) os estados com o melhor ajuste, ou seja, os estados onde existem a melhor relação entre produção científica e número de doutores, atendendo à função geral desta tese: $\hat{Y}_i = f$ ($\hat{u}_1^2$, $\hat{u}_2^2$, $\hat{u}_3^2$..... $\hat{u}_n^2$)**.**

7 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

O objetivo deste capítulo é expor o método de estimação por MQO, aplicado a análise de regressão a que se propõe esta tese. Como recorte, escolhemos a função "variáveis artigos publicados em periódicos nacionais (Y) e doutores autores (X)". Ressaltamos que os cálculos para as demais funções (artigos nacionais, artigos internacionais, anais de eventos e livros) encontram-se no Apêndice.

7.1 OBTENÇÃO DOS DADOS

Os dados foram obtidos a partir do Portal do Plano Tabular do CNPq (http://dgp.cnpq.br/planotabular/), que tem a finalidade de divulgar o perfil da pesquisa no Brasil, em termos quantitativos, organizando esses dados em tabelas configuradas dinamicamente pelo usuário.

Esse sistema possibilita o cruzamento de indicadores, capazes de gerar um número muito grande de diferentes tabelas, que podem ser salvas em planilhas ou em arquivos de texto para futuras consultas, o que permite aos órgãos de fomento e planejamento estabelecer metas e prioridades e que pesquisadores em quaisquer das áreas do conhecimento possam aplicar os dados em suas análises, o que inclui esta pesquisa.

Infelizmente, no decorrer da pesquisa, o Portal do Plano Tabular do CNPq foi descontinuado e não há mais disponibilização desses dados de forma direta no portal, sendo necessária a solicitação via e-mail ao Diretório dos Grupos de Pesquisa do CNPq, para obtenção dos dados de produção científica por unidade da Federação de 2014, finalizando a dinâmica de o próprio pesquisador gerar suas tabelas. Atualmente, o Diretório dos Grupos de Pesquisa disponibiliza apenas algumas tabelas (com dados até 2016) que não atendem à necessidade de dados dessa pesquisa (produção científica por unidade da Federação) e provavelmente atrapalhando sua replicação em outras pesquisas.

A Tabela 7, já apresentada, apresenta os números dos quatro tipos de produção científica nacional disponibilizados pelo CNPq e escolhidos para análise. Como já explicado, nesta pesquisa a variável principal, "produção científica nacional", foi dividida em suas dimensões representadas neste estudo pela variável independente e dependente.

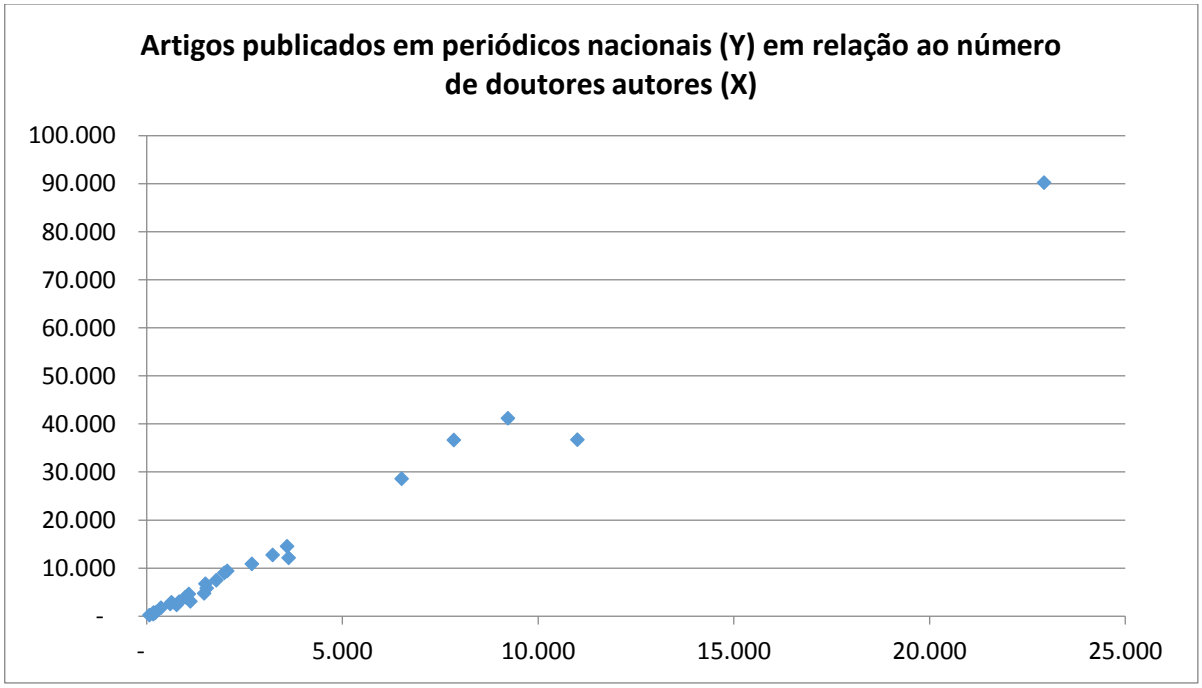
Limitamos a análise dos dados para artigos publicados em periódicos nacionais. Dentro desse aspecto, o número de doutores é a variável independente (x), cadastrados no Portal Plano Tabular do CNPq, que representa os principais produtores da produção científica nacional

(Censo 2010). Quanto à variável dependente, foi selecionada a produção bibliográfica de artigos publicados em periódicos nacionais, onde os Y_n determinam toda a produção nacional desse tipo em todas as áreas do conhecimento por estados no Plano Tabular do CNPq (Censo 2010).

7.2 ANÁLISE DA CORRELAÇÃO E DO MAPA DE DISPERSÃO

Seguindo o roteiro metodológico, o segundo passo é verificar a dispersão dos dados. Para isso foi construído o gráfico seguinte, com a dispersão gerada para artigos publicados em periódicos nacionais em relação ao número de doutores autores.

Gráfico 15 – Artigos publicados em periódicos nacionais em relação ao número de doutores autores



Fonte: Elaboração do autor

Isso serve para verificar a dispersão dos dados através do mapa de dispersão e calcular a coeficiente de correlação para medir a força ou o grau de associação linear entre duas variáveis e decidir se vale a pena regredir o modelo.

A dispersão aponta claramente para a formação de uma reta, bastante similar aos dados da pesquisa de Santos (2011, p. 42) para as mesmas variáveis. Aparentemente, a observação dos dados determina uma forma positiva da curva e a continuação do estudo e aplicação do MQO. Para confirmar, vamos ao cálculo do coeficiente de correlação através da seguinte fórmula:

$$r = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}$$

A tabela abaixo apresenta os dados para auxiliar com os elementos necessários ao cálculo do coeficiente de correlação.

Tabela 8 – Artigo nacional *versus* doutores

UF	X	Y	x	y	x_i^2	$y_i x_i$
Acre	162	387	-3091	-12656	9554968	39122247
Alagoas	760	2317	-2493	-10726	6215603	26742033
Amapá	65	182	-3188	-12861	10164052	41003478
Amazonas	1112	3017	-2141	-10026	4584357	21467573
Bahia	3622	12121	369	-922	136079	-340252
Ceará	1975	8980	-1278	-4063	1633568	5193439
Distrito Federal	2686	10814	-567	-2229	321615	1264301
Espírito Santo	979	3880	-2274	-9163	5171581	20838522
Goiás	1775	7454	-1478	-5589	2184812	8261710
Maranhão	593	2506	-2660	-10537	7076191	28030576
Mato Grosso	1075	4595	-2178	-8448	4744168	18401489
Mato Grosso do Sul	1497	6686	-1756	-6357	3083926	11164249
Minas Gerais	9228	41159	5975	28116	35699297	167987763
Pará	1462	4701	-1791	-8342	3208079	14942112
Paraíba	2055	9347	-1198	-3696	1435470	4428662
Paraná	6508	28586	3255	15543	10594302	50589532
Pernambuco	3215	12731	-38	-312	1452	11905
Piauí	626	2846	-2627	-10197	6901713	26789625
Rio de Janeiro	10997	36693	7744	23650	59967815	183140104
Rio Grande do Norte	1527	5775	-1726	-7268	2979460	12546015
Rio Grande do Sul	7841	36627	4588	23584	21048724	108199072
Rondônia	221	705	-3032	-12338	9193698	37411310
Roraima	169	722	-3084	-12321	9511741	38000475
Santa Catarina	3580	14472	327	1429	106856	467003
São Paulo	22922	90240	19669	77197	386865190	1518371931
Sergipe	824	2953	-2429	-10090	5900581	24510631
Tocantins	358	1675	-2895	-11368	8381668	32912695
Total	**87834**	**352171**	**0**	**0**	**616666969**	**2441458202**

Fonte: Elaboração do autor

Fazendo os cálculos, o coeficiente de correlação para artigos publicados em periódicos nacionais em relação ao número de doutores autores é de 0,994, r = 0,994.

Verificamos que a dispersão dos dados através do mapa de dispersão determina uma forma positiva da curva e o cálculo do coeficiente de correlação indica um alto grau de associação linear entre as variáveis “artigos publicados em periódicos nacionais (Y)” e “doutores autores (X)”. Por isso, vale a pena regredir o modelo.

7.3 ESPECIFICAÇÃO DO MODELO MATEMÁTICO E ESTATÍSTICO

A especificação do modelo matemático e estatístico implica em estabelecer a forma exata da relação funcional (forma da curva) entre as duas variáveis e atribuir uma variável de erro. É dizer qual a forma da curva que estabelece uma relação entre as variáveis “artigos publicados em periódicos nacionais (Y)” e “doutores autores (X)”.

Como a forma da curva pode conter relações inexatas entre as variáveis, deve-se acrescentar um u, que é conhecido como “distúrbio” ou “termo de erro” e representa na função uma variável aleatória (estocástica). Como obtivemos um coeficiente de correlação que indica um alto grau de associação linear entre as variáveis artigos publicados em periódicos nacionais (Y) e doutores autores (X), vamos supor previamente uma forma linear para a curva “variáveis artigos publicados em periódicos nacionais (Y) e doutores autores (X)”, da seguinte forma:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 \ \mathrm{X}_i + \hat{u}_i$$

7.4 ESTIMAÇÃO DOS PARÂMETROS

A estimação dos parâmetros serve para calcular o $\hat{\beta}_1$ e o $\hat{\beta}_2$ (parâmetros) da função do modelo a partir dos dados coletados (que no nosso caso supomos preliminarmente que seja linear). Uma vez estabelecida a equação (especificação do modelo matemático e estatístico), para encontrar os parâmetros é preciso construir uma tabela com os valores X (número de doutores autores) e Y (número de artigos publicados em periódicos nacionais), para calcular os estimadores $\hat{\beta}_1$ e $\hat{\beta}$, como segue:

Tabela 9 – Dados da produção científica para análise do MQO

UF	X	Y	x	y	x_i^2	$y_i x_i$	X_i^2	Y_i^2
Acre	162	387	-3091	-12656	9554968	39122247	26.244	149.769
Alagoas	760	2317	-2493	-10726	6215603	26742033	577.600	5.368.489
Amapá	65	182	-3188	-12861	10164052	41003478	4.225	33.124
Amazonas	1112	3017	-2141	-10026	4584357	21467573	1.236.544	9.102.289
Bahia	3622	12121	369	-922	136079	-340252	13.118.884	146.918.641
Ceará	1975	8980	-1278	-4063	1633568	5193439	3.900.625	80.640.400
Distrito Federal	2686	10814	-567	-2229	321615	1264301	7.214.596	116.942.596
Espírito Santo	979	3880	-2274	-9163	5171581	20838522	958.441	15.054.400
Goiás	1775	7454	-1478	-5589	2184812	8261710	3.150.625	55.562.116
Maranhão	593	2506	-2660	-10537	7076191	28030576	351.649	6.280.036
Mato Grosso	1075	4595	-2178	-8448	4744168	18401489	1.155.625	21.114.025
Mato Grosso do Sul	1497	6686	-1756	-6357	3083926	11164249	2.241.009	44.702.596
Minas Gerais	9228	41159	5975	28116	35699297	167987763	85.155.984	1.694.063.281
Pará	1462	4701	-1791	-8342	3208079	14942112	2.137.444	22.099.401
Paraíba	2055	9347	-1198	-3696	1435470	4428662	4.223.025	87.366.409
Paraná	6508	28586	3255	15543	10594302	50589532	42.354.064	817.159.396
Pernambuco	3215	12731	-38	-312	1452	11905	10.336.225	162.078.361
Piauí	626	2846	-2627	-10197	6901713	26789625	391.876	8.099.716
Rio de Janeiro	10997	36693	7744	23650	59967815	183140104	120.934.009	1.346.376.249
Rio Grande do Norte	1527	5775	-1726	-7268	2979460	12546015	2.331.729	33.350.625
Rio Grande do Sul	7841	36627	4588	23584	21048724	108199072	61.481.281	1.341.537.129
Rondônia	221	705	-3032	-12338	9193698	37411310	48.841	497.025
Roraima	169	722	-3084	-12321	9511741	38000475	28.561	521.284
Santa Catarina	3580	14472	327	1429	106856	467003	12.816.400	209.438.784
São Paulo	22922	90240	19669	77197	386865190	1518371931	525.418.084	8.143.257.600
Sergipe	824	2953	-2429	-10090	5900581	24510631	678.976	8.720.209
Tocantins	358	1675	-2895	-11368	8381668	32912695	128.164	2.805.625
Total	**87834**	**352171**	**0**	**0**	**616666969**	**2441458202**	**902400730**	**14.379.239.575**

Fonte: Elaboração do autor

De posse dos dados de X e Y, basta substituir e resolver as seguintes equações, tal como foi discorrido da revisão de literatura sobre MQO.

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{x} ;$$

$$\hat{\beta}_2 = \frac{\sum(x_i)(y_i)}{\sum(y_i x_i)^2}$$

Substituindo os valores nessas equações, procedemos da forma seguinte para calcular o $\hat{\beta}_2$ e o $\hat{\beta}_1$.

$x = (X-\bar{X})$ e $y = (Y-\bar{Y})$ → desvios

$\hat{\beta}_2 = \Sigma y_i x_i / \Sigma x_i^2$ → 2441458201,889/616666968,667→ $\hat{\beta}_2 = 3,96$

$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ → 13043,37037- (3,95911947 · 3253,111111)→ $\hat{\beta}_1 = 163,92$

Supomos que a forma da função do nosso modelo é $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + u_i$ A partir daí se pode construir a nossa função amostral, que explica as relações entre artigos publicados em periódicos nacionais e doutores, bastando substituir nela os parâmetros que calculados. Então, substituindo-se o $\hat{\beta}_1$ por 163,91 e o $\hat{\beta}_2$ por 3,95, a função fica assim:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \rightarrow \hat{Y}_i = 163,92 + 3,96X_i$$

As informações até aqui discorridas podem ser resumidas no seguinte quadro, onde se encontram as fórmulas do $\hat{\beta}_2$ e do $\hat{\beta}_1$, seus valores calculados e o modelo de regressão para as variáveis "artigos publicados em periódicos nacionais (Y)" e "doutores autores (X)", ou seja, a função "variáveis artigos publicados em periódicos nacionais (Y) e doutores autores (X)" .

Quadro 2 – Estimadores para a correlação artigo nacional *versus* doutores

Artigo nacional *versus* doutores		
Parâmetros	Valores	Função estimada
$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{x}$	$\hat{\beta}_1 = 163,92$	$Y_i = 163,92 + 3,96X_i$
$\hat{\beta}_2 = \frac{\sum(x_i)(y_i)}{\sum(y_i x_i)^2}$	$\hat{\beta}_2 = 3,96$	

Fonte: Elaboração do autor

A função amostral $Y_i = 163,92 + 3,96X_i$ expressa a esperança matemática $E(Y_i|X_i)$, ou seja, a esperança que a função amostral se comporte como a função populacional, mas para que isso ocorra é necessário considerar o erro $\hat{u}_i$ ($\hat{u}_i = Yi + \hat{Y}_i$).

O valor de $\hat{\beta}_2 = 3,96$ mede a inclinação da linha, mostrando que, dentro da faixa amostral de X, quando X aumenta em um doutor, o aumento estimado na produção média de artigos é de cerca de 3,96, ou quase 4. A interpretação do $\hat{\beta}_2$ é: para cada doutor a mais, em média, haverá um aumento de aproximadamente quatro artigos publicados em periódicos nacionais, por estado, para o triênio posterior ao período 2007-2010.

O valor de $\hat{\beta}_1$ = 163,92, que é o intercepto da linha, indica o nível médio da produção de artigos quando o número de doutores é zero. Essa interpretação literal do intercepto poderia não fazer sentido, já que como é possível que o número de doutores seja zero? O intercepto, em uma visão quantitativa, não apresenta um significado real viável e, como sugerido por Gujarati e Porter (2011), deve ser acrescido a este uma explicação qualitativa, por isso que vale salientar a importância da interdisciplinaridade entre conceitos da CI e da econometria. Baseado nisso, a interpretação correta do $\hat{\beta}_1$ é: se não houver a formação de nenhum novo doutor (o que equivale a zero doutores), haverá uma produção média de 164 artigos, em periódicos nacionais, por estado, para o triênio posterior ao período 2007-2010.

7.5 ANÁLISE DA REGRESSÃO

Analisar a regressão significa, para esta tese, verificar a qualidade do ajustamento através da relação funcional linear ou não linear (forma da curva linear ou não linear), isto é, a melhor adequação possível dos parâmetros estimados aos parâmetros populacionais.

Isso é feito através do teste de níveis de confiança dos parâmetros, que podem ser feitos por meio do teste T de Student ou do teste F de Fisher-Snedecor, que são usados para verificar a adequação dos parâmetros estimados aos parâmetros populacionais.

Analisar a função "variáveis artigos publicados em periódicos nacionais (Y) e doutores autores (X)" é um prolongamento da estimação dos parâmetros que, como parte do tratamento dos dados, serve para demonstrar o grau de confiança dos parâmetros do modelo.

Às vezes, a existência da normalidade dos parâmetros e dos resíduos não é suficiente para que os resultados do ajuste do modelo de regressão sejam tratados como confiáveis, o que significa que a curva linear não é um bom estimador e isso implica em encontrar outra forma da curva.

Para verificarmos isso, adotamos nesta tese o coeficiente de regressão (r^2). Se o valor desse coeficiente estiver suficientemente próximo da unidade ($0 < r^2 < 1$), podemos nos convencer que não é um resultado devido ao acaso ou uma peculiaridade dos dados. Em acréscimo, analisamos também o coeficiente de correlação.

Esclareça-se a contribuição de ambos. A análise de correlação mede a força ou o grau de associação linear entre duas variáveis e é feita antes desse passo, quando fazemos a "análise da correlação e do mapa de dispersão". O coeficiente de correlação mede a força dessa associação (que, no nosso caso, supomos linear). Se os dois valores convergirem para valores próximos da unidade e positivos, significa que a forma linear da curva é suficiente para

expressar um modelo de previsão, ou seja, a função de regressão da amostra se aproxima da função de regressão da população.

Para calcular o coeficiente de regressão, utilizamos a seguinte equação:

$$r^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$$

Os valores podem ser extraídos da tabela que utilizamos para calcular os parâmetros. Fazendo os cálculos, o coeficiente de regressão para artigos publicados em periódicos nacionais em relação ao número de doutores autores é de 0,99, $r^2 = 0{,}99$.

O valor de r^2, em torno de 0,99, explica que o número de doutores representa 99% da variação no número de publicações de artigos nacionais. Como r^2 pode ser no máximo igual a 1, a linha de regressão ajusta-se muito bem aos dados. O coeficiente de correlação r = 0,99 explica que publicações de artigos e quantidade de doutores têm uma correlação positiva e alta. A interpretação do r^2 e r é: investimentos na formação de novos doutores aumentará a produção de artigos em periódicos nacionais.

Verificamos que o coeficiente de regressão indica um elevado poder de explicação do modelo. O cálculo do coeficiente de correlação indica um alto grau de associação linear entre as variáveis "artigos publicados em periódicos nacionais (Y)" e "doutores autores (X)". Por isso, a forma linear da curva é suficiente para expressar um modelo de previsão porque a função (na verdade, as funções, porque podemos construir uma para cada estado da Federação) de regressão da amostra se aproxima da função de regressão da população.

7.6 DEFINIÇÃO DO MODELO DE PREVISÃO

Seguindo-se todos os passos anteriores, se chega aos quadrados dos erros estocásticos ($\hat{u}_n^2$) e, a partir daí, podemos construir o modelo de previsão da produção científica nacional desta tese, que se vale justamente da análise dos erros estocásticos ao quadrado ($\hat{u}_i^2$) para que se possa seguir a função geral desta tese, que é:

$$\hat{Y}_i = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$$

Ou, mais especificamente, como recorte, a função "variáveis artigos publicados em periódicos nacionais (Y) e doutores autores (X)". As demais funções estão em Apêndice.

$$\hat{Y}_{Artigos\ Nacionais} = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$$

Portanto, definir o modelo de projeção ou previsão neste estudo consiste na análise da produção científica brasileira a partir dos erros estocásticos ($\hat{u}_i^2$), ou seja, $\hat{Y}_i = f$ ($\hat{u}_1^2$, $\hat{u}_2^2$, $\hat{u}_3^2$..... $\hat{u}_n^2$). Os erros ($\hat{u}n^2$) indicam o quanto um valor Y se aproxima ou se afasta da estimativa esperada $\hat{Y}_i$, conforme a tabela seguinte.

Tabela 10 – Produção de artigos nacionais esperados ($\hat{Y}$) e erros ($\hat{u}_i$ e. $\hat{u}_i^2$)

UF	$\hat{Y}_i$	$\hat{u}_i = Y_i - \hat{Y}_i$	$\hat{u}_i^2$
Acre	805,292	- 418,292	174.968
Alagoas	3.172,846	- 855,846	732.472
Amapá	421,258	- 239,258	57.244
Amazonas	4.566,456	-1.549,456	2.400.813
Bahia	14.503,846	-2.382,846	5.677.953
Ceará	7.983,176	996,824	993.659
Distrito Federal	10.798,110	15,890	253
Espírito Santo	4.039,893	- 159,893	25.566
Goiás	7.191,352	262,648	68.984
Maranhão	2.511,673	- 5,673	32
Mato Grosso	4.419,968	175,032	30.636
Mato Grosso do Sul	6.090,717	595,283	354.362
Minas Gerais	36.698,669	4.460,331	19.894.550
Pará	5.952,147	-1.251,147	1.565.370
Paraíba	8.299,905	1.047,095	1.096.407
Paraná	25.929,864	2.656,136	7.055.057
Pernambuco	12.892,484	- 161,484	26.077
Piauí	2.642,324	203,676	41.484
Rio de Janeiro	43.702,352	-7.009,352	49.131.010
Rio Grande do Norte	6.209,490	- 434,490	188.782
Rio Grande do Sul	31.207,371	5.419,629	29.372.383
Rondônia	1.038,880	- 333,880	111.476
Roraima	833,006	- 111,006	12.322
Santa Catarina	14.337,563	134,437	18.073
São Paulo	90.914,851	- 674,851	455.424
Sergipe	3.426,229	- 473,229	223.946
Tocantins	1.581,280	93,720	8.784
Total	**352171**	**0**	**119.718.087**

Fonte: Elaboração do autor

Com essa análise, é possível estabelecer um *ranking* dos estados brasileiros que apresentam maior eficiência na produção de publicações, isto é, os estados que apresentam $\hat{u}_i^2$ menores se aproximam mais de uma combinação ótima entre número de doutores e publicações. Nesse caso, estamos usando o exemplo de artigos nacionais, mas o mesmo princípio vale para

os demais tipos de publicação (artigos internacionais, anais de eventos e livros) que estão no Apêndice.

O modelo possibilita análises a partir dos erros estocásticos ao quadrado ($\hat{u}_i^2$), onde os $\hat{u}_i^2$ com menor valor indicam a posição do estado dentro do *ranking*. Isso permite medir os mais eficientes em termos de produção científica, expondo do mais eficiente ao menos eficiente, conforme segue:

Tabela 11 – Classificação da produção de artigos nacionais por estado a partir dos erros estocásticos ($\hat{u}_i^2$)

UF	$\hat{u}_i^2$
Maranhão	32,179
Distrito Federal	252,501
Tocantins	8.783,513
Roraima	12.322,337
Santa Catarina	18.073,432
Espírito Santo	25.565,706
Pernambuco	26.077,059
Mato Grosso	30.636,109
Piauí	41.484,067
Amapá	57.244,199
Goiás	68.984,029
Rondônia	111.476,012
Acre	174.968,354
Rio Grande do Norte	188.781,789
Sergipe	223.945,948
Mato Grosso do Sul	354.362,231
São Paulo	455.424,298
Alagoas	732.471,743
Ceará	993.658,513
Paraíba	1.096.407,219
Pará	1.565.370,062
Amazonas	2.400.812,916
Bahia	5.677.952,927
Paraná	7.055.056,636
Minas Gerais	19.894.549,963
Rio Grande do Sul	29.372.382,892
Rio de Janeiro	49.131.010,421

Fonte: Elaboração do autor

Observe-se que os erros ao quadrado indicam o quanto a produção de publicações científicas se aproxima ou se afasta da estimativa esperada. Isso fica claro quando se visualiza

os dados da tabela anterior, perceba-se que é possível estabelecer um *ranking* dos estados que apresentam números mais eficientes de produtividade, que são os estados que apresentam $\hat{u}_i^2$ menores. Estes se aproximam mais de uma combinação ótima entre número de doutores e publicações de artigos em periódicos nacionais. Ressaltamos que os cálculos para as demais funções (artigos internacionais, anais de eventos e livros) encontram-se no Apêndice.

Então aparecem encabeçando a lista Maranhão, Distrito Federal, Tocantins, Roraima, Santa Catarina, Espírito Santo, Pernambuco, Mato Grosso, Piauí e Amapá, como os 10 estados onde a relação da produtividade (doutores *versus* artigos publicados) mais se aproxima da reta de regressão. Na outra ponta, São Paulo, Alagoas, Ceará, Paraíba, Pará, Amazonas, Bahia, Paraná, Minas Gerais, Rio Grande do Sul e Rio de Janeiro são os 10 que mais se afastam da reta. Esses dados explicam o fator de previsibilidade da produção de artigos nacionais, uma vez que, quanto menor o erro, mais próximo estará o valor observado do valor esperado.

Repare-se, olhando o Apêndice, que o estado do Maranhão apresenta a melhor combinação para artigos nacionais, já para artigos internacionais ele fica em 15º lugar, para anais de eventos fica em 16º lugar e para livros sua posição no cenário nacional fica em 11º lugar.

7.7 DISCUSSÃO DOS RESULTADOS

De posse das funções de regressão estimadas pelo procedimento de Mínimos Quadrados, a atenção agora se volta para a interpretação do modelo de regressão sobre o que os dados permitem explicar.

Geometricamente, na linha de regressão estimada, cada ponto da linha de regressão representa uma estimativa do valor médio de Y correspondente ao valor de X escolhido e $\hat{Y}_i$ é uma estimativa de $E(Y_i \mid X_i)$. É necessário entender os resultados dos coeficientes de correlação (r), coeficientes de regressão (r^2), coeficientes lineares ($ß_1$) e coeficientes angulares ($ß_2$) obtidos com os dados.

Observando o coeficiente de correlação (r), foi possível perceber que as associações lineares entre as duas variáveis (Xi; Yi), em todas as associações demonstradas pelas funções, indicam graus fortes e positivos de correlação. Para artigos nacionais, a associação ao número de doutores obteve um r de 0,99, para artigo internacional um r de 0,99, já anais de eventos um r de 0,99 e, por fim, para produção de livros um r de 0,99.

Ou seja, valores altos e muito próximos a 1, indicando uma forte associação linear entre os Yi e os doutores autores (Xi), o que justifica uma análise de previsão linear. Lembramos que

os valores para a função artigos nacionais estão neste capítulo, enquanto os demais valores encontram-se no Apêndice.

Analisando o coeficiente de determinação (r^2), foram encontrados valores que nos permitem afirmar percentualmente o quanto a variável "doutor" explica as variáveis "artigos nacionais, artigos internacionais, anais de eventos e livros" da produção científica.

Em relação aos artigos nacionais (Yi), o número de doutores (Xi) explica 98,80%. Para artigos internacionais, os doutores explicam 98,40%. Em anais de eventos, o total de doutores explica 98,30% e na relação da publicação de livros os doutores explicam 99,40%. Também é possível afirmar que a variável "número de doutores" é adequada e determinante, pois cabe aos erros, de uma maneira geral, explicar menos de 2% da variável dependente (produção científica). Ou seja, a despeito da diversidade dos dados, os r^2 estimados foram bastante altos.

No nosso modelo o valor de r^2, em torno de 0,99, explica que o número de doutores representa cerca de 99% da variação no número de publicações. Como r^2 pode ser no máximo igual a 1, a linha de regressão ajusta-se muito bem aos dados. O coeficiente de correlação r = 0,99 explica que publicações de artigos e quantidade de doutores têm uma correlação positiva e alta. A interpretação do r^2 e r é: investimentos na formação de novos doutores aumentará a produção de artigos em periódicos nacionais.

Os estimadores ($ß_1$ e $ß_2$), em cada função, permitem entender o grau de participação do número de doutores na produção científica nacional (artigos nacionais, artigos internacionais, anais de eventos e livros).

O coeficiente linear $ß_1$, que é o intercepto da reta de regressão, indica o nível médio da produção de artigos quando o número de doutores é zero. Essa interpretação literal do intercepto poderia não fazer sentido, mas o intercepto, em uma visão quantitativa, não apresenta um significado real viável. Deve ser acrescido a ele uma explicação qualitativa, por isso devemos salientar a importância da interdisciplinaridade entre a CI e a econometria. Assim, a interpretação do $\hat{\beta}_1$ é: se não houver formação de nenhum novo doutor (o que equivale a zero doutores), ou melhor, se o número de doutores estagnar, a produção se dará somente ao nível de $\hat{\beta}_1$.

Nessa perspectiva, para Hair Jr (2009, p. 158), a explicação é que o intercepto tem valor explanatório apenas dentro do domínio de valores para as variáveis independentes, ainda que sua interpretação se baseie nas características da variável independente. O mesmo autor alerta que o intercepto tem valor interpretativo somente quando zero é um valor conceitualmente válido para a variável independente, ou seja, quando a variável independente pode ter um valor nulo e ainda manter sua relevância prática, caso contrário, o intercepto auxilia no melhoramento

do processo de previsão, mas sem valor explanatório como no caso desta pesquisa. Hair Jr (2009) afirma ainda que para algumas situações especiais, nas quais sabe-se que a relação específica pode passar pela origem, o intercepto pode ser suprimido (conhecido como regressão pela origem) e que, nesses casos, a interpretação dos resíduos e dos coeficientes de regressão muda um pouco.

No nosso modelo, o valor de $\hat{\beta}_1$ = 163,91 indica o nível médio da produção de artigos quando o número de doutores é zero. Essa interpretação literal do intercepto poderia não fazer sentido, já que como é possível que o número de doutores seja zero? O intercepto, em uma visão quantitativa, não apresenta um significado real viável e, como sugerido por Gujarati e Porter (2011), deve ser acrescido a ele uma explicação qualitativa. Baseado nisso, a interpretação correta do $\hat{\beta}_1$ para nosso modelo é: se não houver a formação de nenhum novo doutor (o que equivale a zero doutores), haverá uma produção média de 164 artigos, em periódicos nacionais, por estado para o triênio posterior ao período 2007- 2010.

O parâmetro $ß_2$ é o coeficiente angular da relação linear entre a variável independente e a dependente. Determina, nesta pesquisa, a participação efetiva de cada doutor na produção, mede a inclinação da reta de regressão, mostrando que, dentro da faixa amostral de X, quando aumenta em um doutor, esse valor é multiplicado e propagado no aumento estimado da produção. Gujarati e Porter (2011, p. 175) explicam que "como o coeficiente angular $ß_2$ é apenas a taxa de variação, ele é medido nas unidades da razão das Unidades da variável dependente sobre Unidades da variável explanatória".

No nosso modelo, o valor de $\hat{ß}_2$ = 3,96 mede a inclinação da linha, mostrando que, dentro da faixa amostral de X, quando X aumenta em um doutor, o aumento estimado na produção média de artigos é de cerca de 3,96, ou quase 4. A interpretação do $\hat{ß}_2$ é: para cada doutor a mais, em média, haverá um aumento de aproximadamente quatro artigos publicados em periódicos nacionais, por estado, para o triênio posterior ao período 2007-2010.

Então, da relação produção de artigos nacionais pelo número de doutores, a partir da função estimada $Y_i = 163{,}92 + 3{,}96X_i$, foi possível afirmar que no investimento na formação de cada doutor, o retorno em termos de produtividade seria de aproximadamente quatro artigos no acréscimo de cada doutor. Os valores positivos de $ß_2$ significam que a reta crescerá em produção científica na medida em que número de doutores cresce, ou seja, a relação entre doutores e produção científica é de proporcionalidade direta, determinando claramente que o aumento do investimento em formação de doutores implica um aumento da produção científica.

A função de nosso modelo $Y_i = 163,92 + 3,96X_i$ expressa a esperança matemática ($E(Y_i|X_i)$) de que a função amostral se comporte como a função populacional, mas para isso temos de considerar o erro $\hat{u}_i$ ($\hat{u}_i=Y_i+ \hat{Y}_i$).

Para análise através do nosso modelo estimado, toma-se como exemplo o estado do Maranhão. De acordo com a função, é possível calcular a quantidade estimada de artigos publicados em periódicos nacionais para o estado do Maranhão ($\hat{Y}_{Maranhão}$) a partir do número observado de doutores autores ($X_{Maranhão}$), ou seja: $\hat{Y}_{Maranhão}=163,92+3,96(593) = 2.512$. Logo, pode-se calcular o erro para o estado do Maranhão, ou seja: $\hat{u}_{Maranhão} = 2.506-2.512 = -6$. Isso pode ser verificado na linha 10 da tabela seguinte e o mesmo cálculo pode ser feito para os demais estados da Federação.

Tabela 12 – O estado do Maranhão na produção de artigos esperados ($\hat{Y}$) e erros ($\hat{u}_i$ e. $\hat{u}_i^2$)

UF	$\hat{Y}_i$	$\hat{u}_i = Y_i - \hat{Y}_i$	$\hat{u}_i^2$
Maranhão	2.511,673	- 5,673	32
Distrito Federal	10.798,110	15,890	253
Tocantins	1.581,280	93,720	8.784
Roraima	833,006	- 111,006	12.322
Santa Catarina	14.337,563	134,437	18.073
Espírito Santo	4.039,893	- 159,893	25.566
Pernambuco	12.892,484	- 161,484	26.077
Mato Grosso	4.419,968	175,032	30.636
Piauí	2.642,324	203,676	41.484
Amapá	421,258	- 239,258	57.244
Goiás	7.191,352	262,648	68.984
Rondônia	1.038,880	- 333,880	111.476
Acre	805,292	- 418,292	174.968
Rio Grande do Norte	6.209,490	- 434,490	188.782
Sergipe	3.426,229	- 473,229	223.946
Mato Grosso do Sul	6.090,717	595,283	354.362
São Paulo	90.914,851	- 674,851	455.424
Alagoas	3.172,846	- 855,846	732.472
Ceará	7.983,176	996,824	993.659
Paraíba	8.299,905	1.047,095	1.096.407
Pará	5.952,147	-1.251,147	1.565.370
Amazonas	4.566,456	-1.549,456	2.400.813
Bahia	14.503,846	-2.382,846	5.677.953
Paraná	25.929,864	2.656,136	7.055.057
Minas Gerais	36.698,669	4.460,331	19.894.550
Rio Grande do Sul	31.207,371	5.419,629	29.372.383
Rio de Janeiro	43.702,352	-7.009,352	49.131.010
Total	**352171**	**0**	**119.718.087**

Fonte: Elaboração do autor

A partir daí, é possível verificar que muitas análises podem ser feitas, mas o modelo proposto por esta tese destaca uma classificação dos estados a partir dos erros estocásticos ($\hat{u}_i^2$). Ou seja, chega-se aos erros estocásticos ($\hat{u}_n^2$) para construir o modelo de previsão da produção científica nacional a partir da análise dos erros estocásticos ($\hat{u}_i^2$), seguindo a seguinte função geral: $\hat{Y}_i = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$.

Tabela 13 – Classificação dos estados produtores de artigos nacionais a partir dos erros estocásticos ($\hat{u}_i^2$), seguinte função geral: $\hat{Y}_i = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$

UF	$\hat{u}_i^2$
Maranhão	32,179
Distrito Federal	252,501
Tocantins	8.783,513
Roraima	12.322,337
Santa Catarina	18.073,432
Espírito Santo	25.565,706
Pernambuco	26.077,059
Mato Grosso	30.636,109
Piauí	41.484,067
Amapá	57.244,199
Goiás	68.984,029
Rondônia	111.476,012
Acre	174.968,354
Rio Grande do Norte	188.781,789
Sergipe	223.945,948
Mato Grosso do Sul	354.362,231
São Paulo	455.424,298
Alagoas	732.471,743
Ceará	993.658,513
Paraíba	1.096.407,219
Pará	1.565.370,062
Amazonas	2.400.812,916
Bahia	5.677.952,927
Paraná	7.055.056,636
Minas Gerais	19.894.549,963
Rio Grande do Sul	29.372.382,892
Rio de Janeiro	49.131.010,421

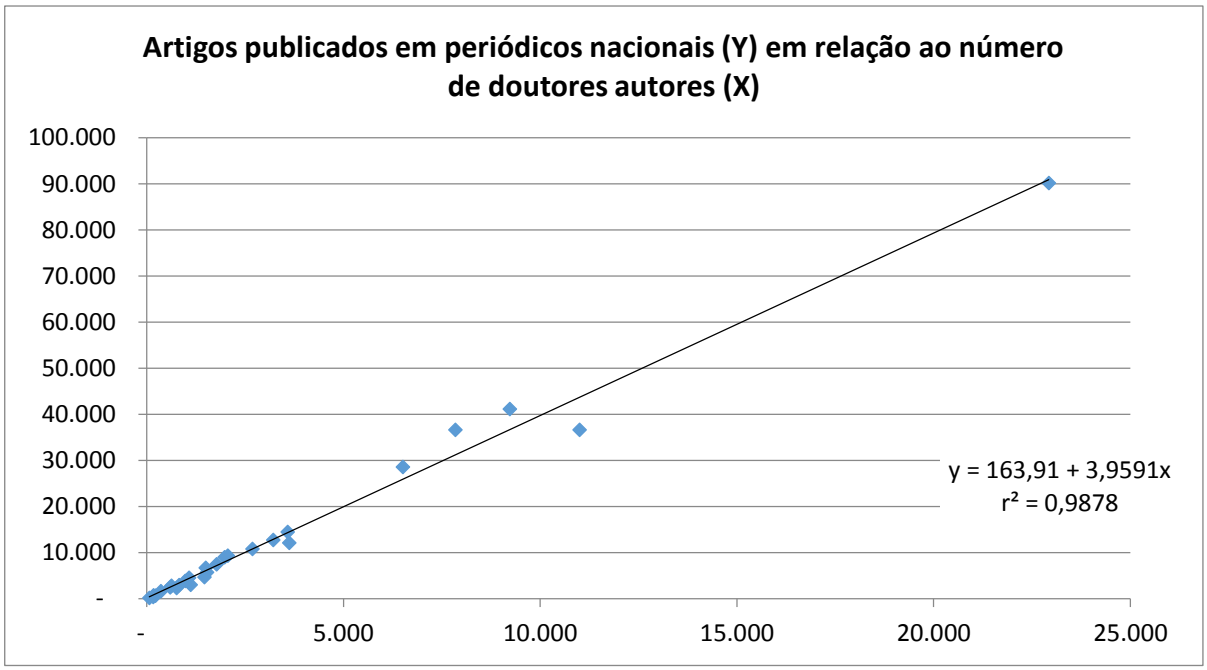
Fonte: Elaboração do autor

Observe-se que os erros indicam o quanto um valor se aproxima ou se afasta da estimativa esperada. Com esses dados é possível, por exemplo, estabelecer um *ranking* dos estados que apresentam números mais eficientes de produtividade, ou seja, os estados que apresentam $\hat{u}_i^2$ menores se aproximam mais de uma combinação ótima entre número de doutores e publicações de artigos em periódicos nacionais. E a mesma explicação cabe para as demais funções (artigos nacionais, artigos internacionais, anais de eventos e livros), conforme pode ser visto no Apêndice. Lembramos que os valores para artigos nacionais estão neste capítulo, os demais valores encontram-se no Apêndice.

O gráfico abaixo apresenta geometricamente a linha de regressão estimada. Como é sabido, cada ponto da linha de regressão representa uma estimativa do valor médio de Y correspondente ao valor de X escolhido; $\hat{Y}_i$ é uma estimativa de $E(Y_i|X_i)$. Esse gráfico representa as aproximações entre bibliometria e informetria, segundo Macias-Chapula (1998) e as áreas de atuação definidas por Tague-Sutcliffe (1992).

Com isso, é possível afirmar que os resultados têm uma grande conexão com as características da relação autor-produtividade, medida por meio do número de artigos ou outros meios (grau de colaboração), por ser possível determinar qual o impacto de cada estado em relação à produção científica. Como também é possível a definição e medida da informação definida por Tague-Sutcliffe (1992).

Gráfico 16 – Dispersão $E(Y_i|X_i)$ e reta de regressão



Fonte: Elaboração do autor

Ao analisar os dados das tabelas e o próprio gráfico é possível observar o medir da informação e os métodos infométricos indicados por Le Coadic (1994), passando pelos estágios anteriores, monodimensionais, que se apoiam em classificações, nomenclaturas preestabelecidas que se baseiam na contagem do número de publicações, e os métodos bidimensionais ou relacionais que permitem a detecção de uma relação entre elementos de informação e os métodos multidimensionais com a utilização de métodos estatísticos.

Então, a partir da classificação pelos erros estocásticos ($\hat{u}_i^2$), seguindo a função geral: $\hat{Y}_i = f(\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$, aparecem encabeçando a lista Maranhão, Distrito Federal, Tocantins, Roraima, Santa Catarina, Espírito Santo, Pernambuco, Mato Grosso, Piauí e Amapá, como os 10 estados onde a relação da produtividade (doutores *versus* artigos publicados) mais se aproxima da reta de regressão. Na outra ponta, São Paulo, Alagoas, Ceará, Paraíba, Pará, Amazonas, Bahia, Paraná, Minas Gerais, Rio Grande do Sul e Rio de Janeiro são os 10 que mais se afastam da reta. Esses dados explicam o fator de previsibilidade da produção de artigos nacionais, uma vez que, quanto menor o erro, mais próximo estará o valor observado do valor esperado. O que colabora com a fala, vista anteriormente, de Le Coadic (1994), ao afirmar que o objeto da informetria é medir produção científica e técnica, visto que a mensuração de elementos de informação permite elaborar indicadores e, para tanto, aplicam-se métodos matemáticos e estatísticos.

Outro ponto a ser explorado com a análise dos erros ($\hat{u}_i$) é o fato de que casos isolados podem ser estudados para entender suas variações. Por exemplo, valores com $\hat{u}_i$ acima da reta indicam variações positivas em Y, ou seja, produziram mais do que era esperado e, no caso de $\hat{u}_i$ negativos, abaixo da reta, há a indicação de que foi produzido abaixo do esperado. Como nas palavras de Macias-Chapula (1998, p. 137), ao afirmar que o indicador infométrico "número de trabalhos" "reflete os produtos da ciência, medidos pela contagem dos trabalhos e pelo tipo de documentos (livros, artigos, publicações científicas, relatórios etc.)", além disso, explica "a dinâmica da pesquisa em um determinado país pode ser monitorada e sua tendência traçada ao longo do tempo".

Como visto, Tague-Sutcliffe (1992) e Macias-Chapula (1998) entendem a cienciometria como o estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. Complementado por Spinak (1998), a cienciometria examina o desenvolvimento e as políticas científicas e esta análise quantitativa considera a ciência como uma disciplina ou atividade econômica. Ou seja, a cienciometria trata de aspectos (medidas) não cobertos por esta tese. Então, mesmo que para a cienciometria, segundo McGrath (1989), o objeto de estudo "corresponda apenas a disciplinas, assuntos, áreas e campos", esta tese pode colaborar para a área no que diz respeito aos seus objetivos que, ainda segundo McGrath (1989), buscam identificar domínios de interesse, local onde os assuntos estão concentrados, como e quando os cientistas se comunicam por proporcionar previsibilidade aos dados.

Da relação produção de artigos internacionais pelo número de doutores, de acordo com a função estimada $Y_i = -2715{,}198 + 4{,}548X_i$, foi possível afirmar que no investimento na formação de cada doutor, o retorno em termos de produtividade seria de aproximadamente

quatro artigos no acréscimo de cada mais um doutor, muito parecida com a análise anterior. Os valores também positivos de $ß_2$ indica que a relação entre doutores e produção científica é de proporcionalidade direta, e mais uma vez expressando que o aumento no investimento em formação de doutores implica aumento da produção científica. Vale reforçar que na função $Y_i = -2715{,}198 + 4{,}548X_i$, Hair Jr (2009, p. 158), assim como Gujarati e Porter (2011, p. 104), explicam que o valor negativo do intercepto ($ß_1$), nesse caso, não tem significado prático.

A próxima análise (em Apêndice) é a da função $Y_i = 727{,}757 + 4{,}47X_i$, que determina a relação entre publicações em anais de eventos pelo número de doutores. Da sua análise foi possível afirmar que a interpretação do coeficiente angular ($ß_2$) foi similar à anterior, onde se o número de doutores aumentar em média de um doutor, a produção de artigos internacionais também aumentará cerca de quatro artigos por estado.

A última função $Y_i = -7{,}1365 + 0{,}342X_i$, (também em Apêndice) determina a relação entre a publicação de livros e o número de doutores. A sua análise apresentou um coeficiente angular ($ß_2$), que representa bem como a divulgação científica acontece por meio da publicação de livros, onde será necessária a formação de pelo menos três doutores para aumentar em média a produção de um livro.

8 CONSIDERAÇÕES FINAIS

O objetivo geral desta tese foi analisar os processos métricos da produção científica brasileira (artigos nacionais, artigos internacionais, anais de eventos e livros) através do MQO frente aos modelos tradicionais de metrias da informação. Para alcançar o objetivo, idealizamos um modelo de análise que conta com os seguintes passos: 1) obtenção dos dados, para definir quais irão compor a variável dependente (Y) e quais irão compor a variável independente (X); 2) análise da correlação e do mapa de dispersão, para medir a força ou o grau de associação linear entre duas variáveis e decidir se vale a pena regredir o modelo; 3) especificação do modelo matemático e estatístico, para especificar a forma exata da relação funcional (forma da curva) entre as duas variáveis e atribuir uma variável de erro; 4) estimação dos parâmetros, para estimar os parâmetros ($\hat{\beta}_1 + \hat{\beta}_2$) da função do modelo a partir dos dados coletados; 5) análise da regressão, para verificar a qualidade do ajustamento através da relação funcional linear ou não linear (forma da curva linear ou não linear), isto é, a melhor adequação possível dos parâmetros estimados aos parâmetros populacionais; 6) definição do modelo de previsão, para se chegar aos erros estocásticos ($\hat{u}n^2$) e construir o modelo de previsão da produção científica nacional a partir da análise dos erros estocásticos ($\hat{u}_i^2$), seguindo a seguinte função geral: $\hat{Y}_i = f$ ($\hat{u}_1^2$, $\hat{u}_2^2$, $\hat{u}_3^2$..... $\hat{u}_n^2$).

A partir desse percurso, respondemos à questão de pesquisa: como analisar a produção científica brasileira (artigos nacionais, artigos internacionais, anais de eventos e livros) utilizando os MQO?

Esta tese estabelece que a utilização de técnicas de análise estatística possibilita a criação de um modelo de previsão da produção científica brasileira que poderá auxiliar na construção de instrumentos que permitam maior suporte à avaliação e às decisões, tanto do MCT quanto das agências de fomento e, assim, racionalizar e flexibilizar tanto a aplicação de recursos públicos quanto a definição de políticas nos estados e na Federação. As respostas para a pergunta de pesquisa são as considerações finais desta tese.

Então, analisando o coeficiente de correlação (r), foi possível perceber que as associações lineares entre as duas variáveis (X_i; Y_i), em todas as associações demonstradas pelas funções, indicam graus fortes e positivos de correlação. Para artigos nacionais, a associação ao número de doutores obteve um r de 0,99, para artigo internacional um r de 0,99, já anais de eventos um r de 0,99 e, por fim, para produção de livros um r de 0,99. Esses valores são altos e muito próximos a 1, indicando uma forte associação linear entre os Yi (publicações científicas brasileiras) e os doutores autores (Xi).

O valor de r^2, em torno de 0,99, explica que o número de doutores representa cerca de 99% da variação no número de publicações de artigos nacionais. Como r^2 pode ser no máximo igual a 1, a linha de regressão ajusta-se muito bem aos dados. O coeficiente de correlação r = 0,99 explica que publicações de artigos nacionais e quantidade de doutores têm uma correlação positiva e alta. A interpretação do r^2 e r é: investimentos na formação de novos doutores, aumentará a produção de artigos em periódicos nacionais.

A análise do $\hat{\beta}_1$ é: se não houver a formação de nenhum novo doutor (o que equivale a zero doutores), haverá uma produção média de 164 artigos, em periódicos nacionais, por estado, para o triênio posterior ao período 2007-2010.

Analisando o $\hat{\beta}_2$ = 3,96, que mede a inclinação da linha, percebe-se que, dentro da faixa amostral de X, quando X aumenta em um doutor, o aumento estimado na produção média de artigos é de cerca de 3,96, ou quase 4. A interpretação do $\hat{\beta}_2$ é: para cada doutor a mais, em média, haverá um aumento de aproximadamente quatro artigos publicados em periódicos nacionais, por estado, para o triênio posterior ao período 2007-2010.

O modelo proposto por esta tese destaca uma classificação dos estados a partir dos erros estocásticos ($\hat{u}_i^2$). Ou seja, chega-se aos erros estocásticos ($\hat{u}_n^2$) para construir o modelo de previsão da produção científica nacional a partir da análise dos erros estocásticos ($\hat{u}_i^2$), seguindo a seguinte função geral: $\hat{Y}_i = f$ ($\hat{u}_1^2$, $\hat{u}_2^2$, $\hat{u}_3^2$..... $\hat{u}_n^2$). Os erros indicam o quanto um valor se aproxima ou se afasta da estimativa esperada. Com isso, é possível estabelecer um *ranking* dos estados que apresentam números mais eficientes de produtividade, ou seja, os estados que apresentam $\hat{u}_i^2$ menores se aproximam mais de uma combinação ótima entre número de doutores e publicações de artigos em periódicos nacionais, artigos internacionais, anais de eventos e livros.

Da relação produção de artigos internacionais pelo número de doutores, de acordo com a função estimada $Y_i = -2715{,}20 + 4{,}55X_i$, foi possível afirmar que no investimento na formação de cada doutor, o retorno em termos de produtividade seria de aproximadamente quatro artigos no acréscimo de cada mais um doutor. A análise indica que a relação entre doutores e produção científica de artigos internacionais é de proporcionalidade direta, e mais uma vez expressa que o aumento no investimento em formação de doutores implica aumento da produção científica.

A análise da função $Y_i = 727{,}76 + 4{,}47X_i$, que determina a relação entre publicações em anais de eventos pelo número de doutores, permite afirmar que se o número de doutores aumentar em média um doutor, a produção de artigos internacionais também aumentará cerca de quatro artigos por estado para o triênio posterior ao período 2007-2010.

A função $Y_i = -7,14 + 0,34X_i$ determina a relação entre a publicação de livros e o número de doutores. A sua análise, que representa a divulgação científica por meio da publicação de livros, indica que é necessária a formação de pelo menos três doutores para aumentar em média a produção de um livro por estado para o triênio posterior ao período 2007-2010.

Dentre as preocupações crescentes no mundo moderno existe a de medir a produção científica nos países através de seus produtores, produtos e produtividade efetiva. No Brasil, o MCT vem adotando várias medidas para implementar tais mecanismos que gerenciem o controle desse processo produtivo. Isso visa ao desenvolvimento nacional e também a tornar o país mais competitivo no mercado mundial, com base na exploração e compreensão dos dados de sua produção.

O sucesso de qualquer análise depende da disponibilidade e relevância dos dados, algo que a metodologia econométrica preconiza como essencial dedicar algum tempo examinando a natureza, as fontes e as limitações dos dados que podem aparecer na análise empírica. Da análise da produção científica brasileira com o uso de MQO foi possível observar que, para cada doutor a mais, haverá um incremento na publicação de artigos nacionais, internacionais, anais de eventos e livros, por estado, para o triênio posterior ao período 2007-2010. Portanto, sugere-se o fomento à formação de novos doutores, bem como investimentos para o aumento da produção científica brasileira.

Também o modelo, a partir dos erros estocásticos ($\hat{u}_i^2$), deve estabelecer um *ranking* dos estados que apresentam números mais eficientes de produtividade, mesmo que os números absolutos indiquem o contrário. Sugere-se que isso seja usado pelos órgãos de fomento como um fator de previsibilidade da produção científica, de modo a orientar os investimentos públicos nessa área.

Por fim, este estudo, mesmo contribuindo com a CI e biblioteconomia, limita-se a demonstrar a utilização de métodos e procedimentos que ficaram circunscritos à natureza do presente estudo. É necessário, pois, ampliar as possibilidades de análise de dados sobre os produtos e serviços desenvolvidos nas bibliotecas e outros serviços de informação, bem como alargar os horizontes da CI para pesquisas semelhantes.

9 REFERÊNCIAS

ALCAÍN, M. D.; SAN MILLÁN, M. J. Uso y tendencias de lãs técnicas bibliomericas bibliométricas em enciencias sociales y humanas a nivel internacional. Revista Española de Documentación Cientifica, v. 16, n. 1, 1993.

ALVARENGA, Lídia. A institucionalização da pesquisa educacional no Brasil. 1996. Tese (Doutorado) – Faculdade de Educação da Universidade Federal de Minas Gerais, Belo Horizonte, 1996.

ARAÚJO, Carlos Alberto Avila. Arquivos, Bibliotecas e Museus: apontamentos para um possível modelo curricular de convergência. In: DUARTE, Zeny. *Arquivos, bibliotecas e museus: realidades de Portugal e Brasil*. Salvador: EDUFBA, 2013, p. 259-300.

_____. Bibliometria: evolução histórica e questões atuais. *Em Questão: Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS*, v. 12, n. 1, p. 11-32, 2006. Disponível em: <http://www.brapci.ufpr.br/brapci/v/6356>. Acesso em: 14 jul. 2016.

BAILON-MORENO, R.; JURADO ALAMEDA, E. Analysis of the field of physical chemistry of surfactants with the Unified Scienctometric Model: fit of relational and activity indicators. Scientometrics, v. 63, n. 2, p. 259-276, 2005.

BANDEIRA, M. (Org). *Texto 1B: tipos de pesquisa*. Disciplina: Modelos de Investigação e Produção em Psicologia do Laboratório de Psicologia Experimental, Departamento de Psicologia – FUNREI. Disponível em: <http://www.ufsj.edu.br/portal-repositorio/File/lapsam/texto%201b%20-%20TIPOS%20DE%20PESQUISA.pdf>. Acesso em: 21 out. 2015.

Bolaño, César Ricardo Siqueira; Melo, Ricardo Oliveira Lacerda de. Tecnologias da informação e da comunicação e desenvolvimento regional. *Revista de Economia Política das Tecnologias da Informação e Comunicação*, v. 2, n.2, p. 63-82, jul/ago 2000.

BORGES, Maria Manuel. Timeline de Bibliometria. 2008. Disponível em: https://mqgi.wordpress.com/2008/10/28/contributos-para-a-bibliometria/. Acesso em: 27 abr. 2015.

BORGES, Paulo César Rodrigues. Métodos quantitativos de apoio à bibliometria: a pesquisa operacional pode ser uma alternativa?. *Ci. Inf.*, Brasília, v. 31, n. 3, p. 5-17, Sept. 2002.

BORKO, H. Information science: what is it? *American Documentation*, Jan. 1968.

BRAMBILLA, Sônia Domingues Santos; STUMPF, Ida Regina Chitto. Produção Científica da UFRGS representada na WOS (2000-2009). *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 17, n. 3, 2012.

BUFREM, Leilah Santiago; PRATES, Yara. O saber científico registrado e as práticas de mensuração da informação. *Ci. Inf*, [S.l.], v. 34, n. 2, mar. 2006.

BUNKLEY, Nick. Joseph Juran, 103, Pioneer in Quality Control, Dies. New York Times, n. 3, mar. 2008. Disponível em: <http://www.nytimes.com/2008/03/03/business/03juran.html>. Acesso em: 27 abr. 2015.

BURNHAM, Terezinha Fróes. Da sociedade da informação à sociedade da aprendizagem: cidadania e participação sociopolítica na (in)formação do trabalhador. Encontro nacional de Ensino e pesquisa em Informação. Disponível em:< www.cinform.ufba.br/.../TeresinhaFroesBurnhamSociedadedaAprendizagem. pdf>. Acesso em: 26 Nov. 2014.

CALLON, M.; COURTIAL, J.-P.; PENAN, H. La scientométrie. Paris: PUF, 1993.

CARDOSO, Ana Maria Pereira. Pós-Modernidade e informação: conceitos complementares? *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 1, n. 1, p. 63-79, jan./jul. 1996.

CASTELLS, Manuel. *A Era da Informação*: economia, sociedade e cultura, vol. 1, São Paulo: Paz e terra, 1999.

CHOO, Chun Wei. A organização do conhecimento: uma visão holística de como as organizações usam a informação. In: _____. *A organização do conhecimento*: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. São Paulo: Senac, 2000. Cap. 1, p.27-61.

COILE, Russell C. Lotka and information science, *Journal of the American Society for Information Science*, v. 26, n. 2, p. 133, 1975.

COILE, Russell C. Lotka's frequency distribution of scientific productivity. *Journal of the American Society for Information Science*, v. 28, n. 6, p. 366-370, 1977.

DRESDEN, A. A report on the scientific work of the Chicago section, 1897-1922. *Bulletin of the American Mathematical Society*, v. 28, p. 303-307, July 1922. Disponível em: <http://projecteuclid.org/download/pdf_1/euclid.bams/1183485109>. Acesso em: 27 de abril de 2015.

EKNOYAN, Garabed. Adolphe Quetelet (1796–1874) - the average man and indices of obesity. *Nephrol. Dial. Transplant*. v. 23, n. 1, (2008) p. 47-51. Disponível em: <http://ndt.oxfordjournals.org/content/23/1/47.full>. Acesso em: 12 jan. 2015.

FAPESP. *Indicadores de ciência, tecnologia e inovação em São Paulo 2004*. São Paulo: FAPESP, 2005.

FONSECA, Jairo Simon da; MARTINS, Gilberto de Andrade; TOLEDO, Geraldo Luciano. *Estatística aplicada*. São Paulo, SP: Atlas, 1976. 273 p.

FONSECA. Edson Nery. *Bibliometria: teoria e prática*. São Paulo: Cultrix, 1986.

FORESTI, Nóris. *Estudo da contribuição das revistas brasileiras de biblioteconomia e ciência da informação enquanto fonte de referência para a pesquisa*. Brasília: Depto. de Biblioteconomia da UnB, 1989 (dissertação de mestrado).

GIL, Antonio Carlos. *Métodos e técnicas de pesquisa social*. 6. ed. São Paulo: Atlas, 2008.

GLÄNZEL, W.; SCHOEPFLIN, U. Little scientometrics, big scientometrics... and beyond? *Scientometrics*, v. 30, n. 2-3, p. 375-384, 1994.

GRÁCIO, Maria Cláudia Cabrini; OLIVEIRA, Ely Francina Tannuri de. A INSERÇÃO E O IMPACTO INTERNACIONAL DA PESQUISA BRASILEIRA EM ESTUDOS

MÉTRICOS: uma análise na base Scopus. *Tendências da Pesquisa Brasileira em Ciência da Informação*, João Pessoa, v. 5, n. 1, 2013.

GUEDES, Vânia L. S.; BORSCHIVER, Suzana. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. In: ENCONTRO NACIONAL DE CIÊNCIAS DA INFORMAÇÃO, 6, 2005, Salvador. *Anais eletrônicos*... Salvador, 2005. Disponível em: <http://www.cinform-anteriores.ufba.br/vi_anais/docs/VaniaLSGuedes.pdf >. Acesso em: 9 maio 2016.

GUJARATI, Damodar N.; PORTER, Dawn C. *Econometria básica*. 5. ed. Porto Alegre: AMGH, 2011.

HAIR JR, Joseph F [et al.] *Análise multivariada de dados*. 6. ed. Porto Alegre: Bookman, 2009.

HAYASHI, Maria Cristina Piumbato Innocentini. Afinidades eletivas entre a cientometria e os estudos sociais da ciência. *Filosofia e Educação* (Online), ISSN 1984-9605 – Volume 5, Número 2, Outubro de 2013.

HULME, E. W. *Statistical bibliography in relation to the growth of modern civilization*. London, 1923.

JONES, A W. Impact factors of forensic science and toxicology journals: what do the numbers really mean? *Forensic Science International* , V 133, n.1-2 ,p. 1-8,2003

JORGE, Ricardo Arencibia; ANEGÓN, Félix de Moya. La evaluación de lainvestigación científica: una aproximación teórica desde lacienciometría. *ACIMED*, v.17 n.4 Ciudad de La Habana abr. 2008.

KLEIN, Lawrence R. *Introdução a econometria*. São Paulo: Atlas, 1978.

KMENTA, J. *Elementos de Econometria*. São Paulo, Atlas, 1988.

LACEY, Hugh. Valores e atividade científica. São Paulo: Discurso Editorial, 1998.

LE COADIC, Yves F. Mathématique et statistique em science de linformation et enscience de la communication: infométriemathématique et infométrie statistique des revues scientifiques. *Ci. Inf.*, Brasília, v. 34, n. 3, p. 15. 2005.

LE COADIC, Yves-François. *A ciência da informação*. Tradução de Maria Yêda F.S. de Filgueiras Gomes. Brasilia, DF: Brinquet de Lemos, 1996. Original publicado em 1994.

LOTKA, Alfred J. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, v. 16, n. 12, p. 317-323, June 1926.

MACHADO, Raymundo das Neves; LETÃ, Jacqueline. Consumo da informação científica na ciência brasileira: estudo exploratório na temática ceratocone e extração de catarata.. *Em Questão: Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS*, Porto Alegre, v. 18, n. 1, p. 129-144, 2012.

MACÍAS-CHAPULA, C. A. O papel da informetria e da cientometria e sua perspectiva nacional e internacional. *Ci. Inf.*, Brasília, v. 27, n. 2, p. 134-140, maio/ago. 1998.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. *Fundamentos de metodologia científica*. 5. ed. São Paulo: Atlas, 2003.

MARTINS, G. de Andrade. *Estatística Geral e Aplicada*. São Paulo: Atlas, 2001.

MCGRATH, W. What bibliometricians, scientometricians and informetricians study; a typology for definition and classification; topics for discussion. In: INTERNATIONAL CONFERENCE ON BIBLIOMETRICS, SCIENTOMETRICS AND INFORMETRICS, 1989, Ontario. *Second Conference*... Ontario: The University of Western Ontario, 1989.

MEADOWS, A.J. *A comunicação científica*. Brasília: Briquet de Lemos, 1999.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada*: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005.

MOSTAFA, S. P.; MÁXIMO, L. F. A produção científica da Anped e da Intercom no GT da Educação e Comunicação. *Ci. Inf.*, v. 32, n. 1, 2003.

MOSTAFA, Solange. Citações epistemológicas na educomunicação. *Comunicação & Educação*, São Paulo, v. 8, n.24, p. 15-28, 2002.

MOSTAFA, Solange; MÁXIMO, Luís. A produção científica da ANPED e da Intercom no GT de Educação e Comunicação. *Ci. Inf.*, Brasília, v. 32, n. 1, p. 96-101, jan./abr. 2003.

MUELLER, Suzana; OLIVEIRA, Hamilton. Autonomia e dependência na produção da ciência: uma busca conceitual para estudar as relações na comunicação científica. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 8, n.1, p. 58- 65, jan./jun. 2003.

OLIVEIRA, Ely Francina Tannuri de; GRÁCIO, Maria Cláudia Cabrini. Indicadores bibliométricos em ciência da informação: análise dos pesquisadores mais produtivos no tema estudos métricos na base Scopus. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 16, n. 4, p. 16-28, 2011.

OLIVEIRA, Ely Francina Tannuri de; GRÁCIO, Maria Cláudia Cabrini. Visibilidade dos pesquisadores no periódico Scientometrics a partir da perspectiva brasileira: um estudo de cocitação. *Em Questão: Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS*, Porto Alegre, v. 18, n. 1, p. 99-113, 2012.

PAO, Miranda Lee. Automatic text analysis based on transition phenomena of word occurrences, Journal of the American Society for Information Science, *New York*, v. 29, n.3, p. 121-124, may 1978.

PETRUSZEWYCZ, M. L'histoire de laloi d'Estoup-Zipf : documents. Mathématiques et SciencesHumaines, v.44, 1973, p. 41-56. Disponível em: <http://www.numdam.org/item?id=MSH_1973__44__41_0>. Acesso em: 27 abr. 2015.

PICH, Santiago. Adolphe Quetelet e a biopolítica como teologia secularizada. *Hist. cienc. saude-Manguinhos*, Rio de Janeiro, v. 20, n. 3, p. 849-864, Sept. 2013

PINHEIRO, Lêna Vânia. Processo evolutivo e tendências contemporâneas da ciência da informação. *Informação & Sociedade*: Estudos, João Pessoa, v. 15, n. 1, p. 13-48, jan./jun. 2005.

POMBO, Olga. Interdisciplinaridade e integração dos saberes. *Liinc em Revista*, v. 1, n. 1, p. 3 -15, mar. 2005.

PRAT, Anna Maria. Avaliação da produção científica como instrumento para o desenvolvimento da ciência e da tecnologia. *Ci. Inf.*, Brasília, v. 27, n. 2, mai./ago. 1998.

PRICE, Derek J. De Solla. Networks of scientific papers. *Science*, [s.l.], v. 149, n.3683, p. 56-64, July 1965.

PRITCHARD, Alan. Statistical Bibliography or Bibliometrics?. *Journal of Documentation*, v. 25, n. 4, Dec 1969, 348-349. Disponível em: <http://independent.academia.edu/AlanPritchard/Papers/602982/Statistical_bibliography_or_bibliometrics>. Acesso em: 27 de abril de 2015.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. *Metodologia do trabalho científico*: métodos e técnicas de pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo/RS: Feevale, 2013. Disponível em: http://www.faatensino.com.br/wp-content/uploads/2014/11/2.1-E-book-Metodologia-do-Trabalho-Cientifico-2.pdf. Acesso em: 10 abr. 2015.

QUETELET, Adolphe. *Sur l'homme et le développement de ses facultés*: ou, Essai de physique sociale. Paris : Bachelier, imprimeur-libraire, 1835.

REICHMANN, W. J. *Uso e Abuso das Estatísticas*. Rio de Janeiro: Artenova S. A., 1975.

ROCHA, Elisa Maria Pinto; FERREIRA, Marta Araújo Tavares. Indicadores de ciência, tecnologia e inovação: Mensuração dos sistemas de CTeI nos Estados brasileiros. *Ci. Inf.*, Brasília, v. 33, n. 3, p. 61-68, set./dez. 2004.

ROUSSEAU, Ronald. Indicadores bibliométricos e econométricos para a avaliação de instituições científicas. *Ci. Inf.*, Brasília, v. 27, n. 2, p. 149-158, maio/ago. 1998.

_____. Timeline of bibliometrics. 2014. Disponível em: <http://users.telenet.be/ronald.rousseau/html/timeline_of_bibliometrics.html>. Acesso em: 27 abr. 2015.

RUA, Maria das Graças. Análise de políticas públicas: conceitos básicos. In.: RUA, M. G; CARVALHO, M. I. V. (org.). *O estudo da política: tópicos selecionados*. Brasília: Paralelo 15, 1998.

RUSSELL Jane; ROUSSEAU, Ronald. Bibliometrics and institutional evaluation. In: ENCYCLOPEDIA of Life Support Systems (EOLSS). [S.l.]: RigasArvantis, 2002. Part 19.3: Science and Technology Policy.

SANTOS, Levi Alã Neves dos. *Contribuição da mineração de dados e da otimização heurística para a interpretação dos dados da produção científica brasileira***.** Salvador, BA, 2011. 114f. Dissertação (Mestrado) - Universidade Federal da Bahia, Instituto de Ciência da Informação, 2011.

SARACEVIC, Tefko. A natureza interdisciplinar da ciência da informação *Ci. Inf.*, Brasília, v. 24, n. 1, apr. 1995.

SHANNON, Claude; WEAVER, Warren. *Teoria matemática da comunicação*. São Paulo, SP: Difel, 1975.

SOUZA, I.G.C.O; DUARTE, E.N. Ruben. Dimensões de um modelo de gestão da informação no campo da Ciência da Informação: uma revelação da produção científica do Enancib. Liinc em Revista, v.7, n.1,março 2011, Rio de Janeiro, p. 152 – 169 -. Disponível em: <http://www.ibict.br/liinc>. Acesso em: 6 maio 2016.

SPINAK, Ernesto. Indicadores cienciometricos. *Ci. Inf*., Brasília, v. 27, n. 2, 1998.

SURREY, M. J. C. *Uma introdução à econometria*. Rio de Janeiro: Zahar, 1974.

TAGUE-SUTCLIFFE, J. An introduction to informetrics. *Information Processing & Management*, v. 28, n. 1, p. 1-3, 1992.

TARGINO, Maria das Graças. *Comunicação científica*: o artigo de periódico nas atividades de ensino e pesquisa do docente universitário brasileiro na pós-graduação. 1998. Tese (Doutorado em Ciência da Informação) – Curso de Pós-Graduação em Ciência da Informação, Universidade de Brasília, Brasília, 1998.

TRZESNIAK, Piotr. Indicadores quantitativos: reflexões que antecedem seu estabelecimento. *Ci. Inf.*, Brasília, v. 27, n. 2, 1998.

URBIZAGASTEGUI ALVARADO, Ruben. A bibliometria no Brasil. *Ci. Inf.*, Brasília, v. 13, n. 2, p. 91-105, jul. 1984.

_____. A Bibliometria: história, legitimação e estrutura. In: TOUTAIN, Lídia Maria Batista Brandão (Org.). *Para entender a ciência da informação*. Salvador : EDUFBA, 2007. p. 185-217.

_____. A cienciometria como um campo científico. *Informação & Sociedade*: Estudos, v. 20, n. 3, p. 41-62, Sep-Dic. 2010.

_____. A frente de pesquisa na literatura sobre a produtividade dos autores. *Encontros BIBLI*, Florianópolis, Santa Catarina, Brasil, v. 14, n. 28, p. 38-56, 2009b.

_____. A lei de Lotka na bibliometria brasileira. . *Ci. Inf.*, Brasília, v.31, n.2,. p.14-20, maio/ago. 2002.

_____. A produtividade dos autores sobre a Lei de Lotka. *Ci. Inf***.**, Rio de Janeiro, Brasil, v. 37, n. 2, p. 87-102, maio/ago. 2008.

_____. Elitismo na literatura sobre a produtividade dos autores. . *Ci. Inf*, Rio de Janeiro, Brasil, v. 32, n. 2, p. 69-79, maio-ago., 2009a.

VANTI, Nadia Aurora Peres. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. *Ci. Inf***.**, Brasília , v. 31, n. 2, p. 369-379, Aug. 2002 .

_____. A cientometria revisitada à luz da expansão da ciência, da tecnologia e da inovação. *Ponto de Acesso*, Salvador, v. 5, n. 3, p. 5-31, dez 2011.

VINKLER, P. An attempt of surveying and classifying bibliometric indicators for scientometric purposes. *Scientometrics*, v. 13, n. 5-6, p. 239-259, 1988.

WERSIG, G. Information science: the study of postmodern knowledge usage. *Information Processing and Management: an International Journal*, Tarrytown-Nova Iorque, v. 29, n. 2, p. 229-239, Mar./Apr. 1993.

WITTER, Geraldina Porto (Org). *Produção científica*. Campinas: Átomo, 1997.

WORMELL, I. Informetrics: exploring databases as analytical tools. *Database*, v. 21, n. 5, p. 25-30, out./nov. 1998.

APÊNDICE – ANÁLISE DAS DEMAIS FUNÇÕES

Com base em tudo o que foi discorrido ao longo de nossa tese e, em especial, o Capítulo 4, expomos neste apêndice o modelo bibliométrico aplicado para as demais produções científicas nacionais, que são: artigo internacional *versus* doutores, anais de eventos *versus* doutores e livros *versus* doutores. Lembramos que a análise da função "variáveis artigos publicados em periódicos nacionais (Y) e doutores autores (X)" já foi feita no Capítulo 4. Os passos para aplicação do MQO são: definir os dados e as variáveis dependentes (y) e a variável independente (x), formando pares ordenados na análise de dispersão, de modo a: medir a força ou o grau de associação linear entre duas variáveis para decisão da regressão, especificar a forma exata da relação funcional (forma da curva) entre as duas variáveis e atribuir uma variável de erro, estimar os parâmetros (beta 1 e beta 2) da função do modelo a partir dos dados coletados, verificar a qualidade do ajustamento, isto é, a melhor adequação dos parâmetros estimados aos parâmetros populacionais, levantar os erros estocásticos ($\hat{u}n^2$) e construir o modelo de previsão da produção científica nacional a partir da análise dos erros estocásticos ($\hat{u}_i^2$), seguindo a função geral $\hat{y}_i = f\ (\hat{u}_1^2, \hat{u}_2\ ^2, \hat{u}_3\ ^2$ e $\hat{u}_4\ ^2)$.

- PASSO 1: OBTENÇÃO DOS DADOS

A tabela abaixo apresenta os números dos quatro tipos de produção científica nacional disponibilizados pelo CNPq e escolhidos para análise.

Tabela 14 – Tipos de produção científica nacional disponibilizados pelo CNPq

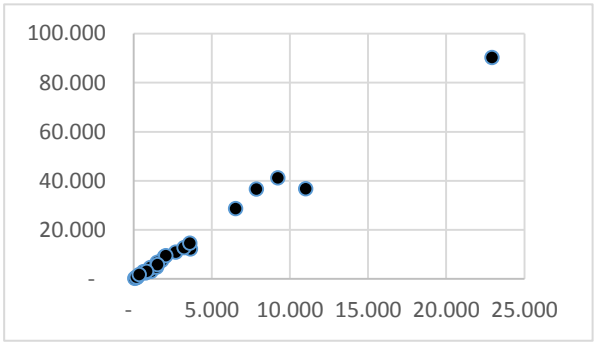
UF	Total de autores doutores	Artigos completos em periódicos de circulação nacional	Artigos completos em periódicos de circulação internacional	Trabalhos completos publicados em anais de eventos	Livros
Acre	162	387	206	305	58
Alagoas	760	2.317	1.683	3.568	203
Amapá	65	182	193	221	14
Amazonas	1.112	3.017	3.723	2.981	406
Bahia	3.622	12.121	9.674	13.189	1.068
Ceará	1.975	8.980	7.556	10.259	609
Distrito Federal	2.686	10.814	8.683	11.029	1.141
Espírito Santo	979	3.880	2.486	5.013	309
Goiás	1.775	7.454	5.419	7.786	624
Maranhão	593	2.506	2.058	1.722	171
Mato Grosso	1.075	4.595	2.157	3.957	344
Mato Grosso do Sul	1.497	6.686	3.894	7.726	524
Minas Gerais	9.228	41.159	34.692	44.797	2.858
Pará	1.462	4.701	4.485	6.112	501
Paraíba	2.055	9.347	5.206	14.840	607
Paraná	6.508	28.586	21.122	34.752	2.118
Pernambuco	3.215	12.731	9.673	16.517	940
Piauí	626	2.846	1.672	1.789	146
Rio de Janeiro	10.997	36.693	42.933	44.523	4.076
Rio Grande do Norte	1.527	5.775	3.930	9.042	465
Rio Grande do Sul	7.841	36.627	30.596	42.966	2.846
Rondônia	221	705	723	769	75
Roraima	169	722	322	373	34
Santa Catarina	3.580	14.472	11.203	21.979	1.510
São Paulo	22.922	90.240	108.990	100.455	7.759
Sergipe	824	2.953	2.088	4.689	327
Tocantins	358	1.675	809	1.002	125
TOTAIS	**87.834**	**352.171**	**326.176**	**412.361**	**29.858**

Fonte: Elaboração do autor

- PASSO 2: ANÁLISE DA CORRELAÇÃO E DO MAPA DE DISPERSÃO

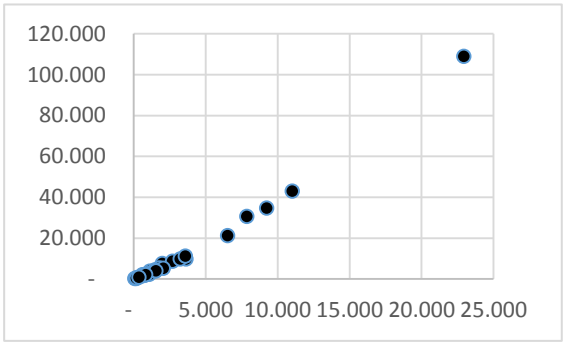
Os gráficos abaixo apresentam a dispersão dos dados da produção científica nacionais. É possível perceber o comportamento linear para os quatro tipos de produção e, além disso, um comportamento de semelhança entre os gráficos.

Gráfico 17 – Dispersão artigo nacional *versus* doutores



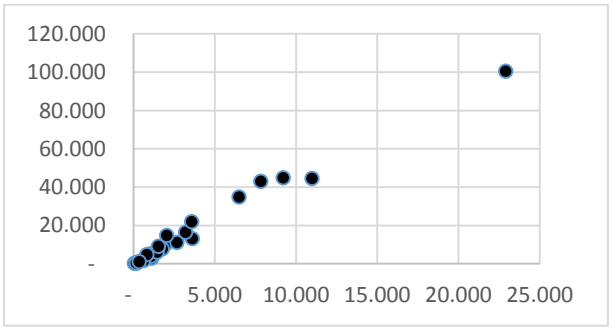
Fonte: Elaboração do autor

Gráfico 18 – Dispersão artigo internacional *versus* doutores



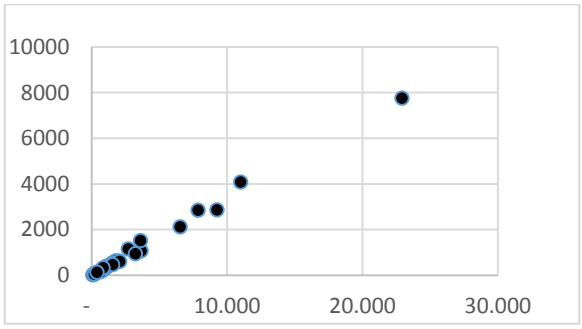
Fonte: Elaboração do autor

Gráfico 19 – Dispersão anais de eventos *versus* doutores



Fonte: Elaboração do autor

Gráfico 20 – Dispersão livros *versus* doutores



Fonte: Elaboração do autor

Verificamos que a dispersão dos dados através do mapa de dispersão determina uma forma positiva da curva e o cálculo do coeficiente de correlação indica um alto grau de associação linear entre as variáveis.

Quadro 3 – Coeficientes de correlação das curvas

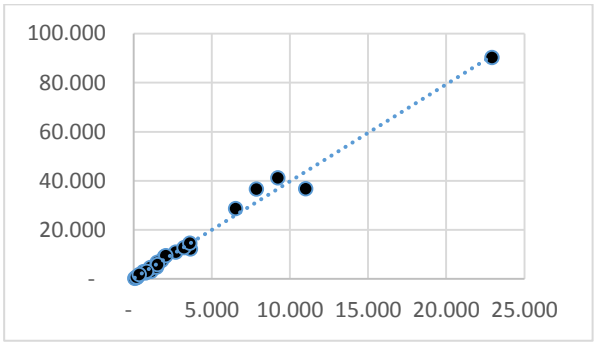
Correlações com o número de doutores	$r = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}$
	Linear
Artigo nacional	r =0,99
Artigo internacional	r =0,99
Anais de eventos	r =0,99
Livros	r =0,99

Fonte: Elaboração do autor

- PASSO 3: ESPECIFICAÇÃO DO MODELO MATEMÁTICO E ESTATÍSTICO

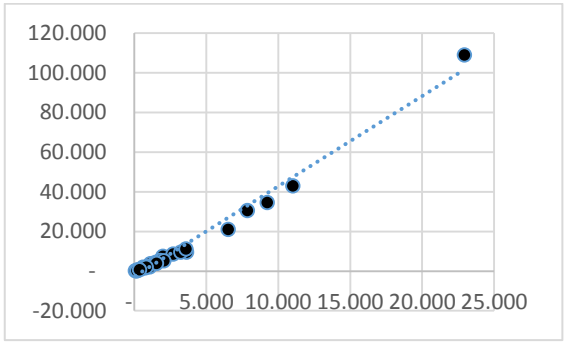
Os gráficos abaixo apresentam a dispersão dos dados com a linha de tendência para os quatro tipos de produção. Confirma a predisposição do modelo linear de curva.

Gráfico 21 – Curva de tendência artigo nacional *versus* doutores



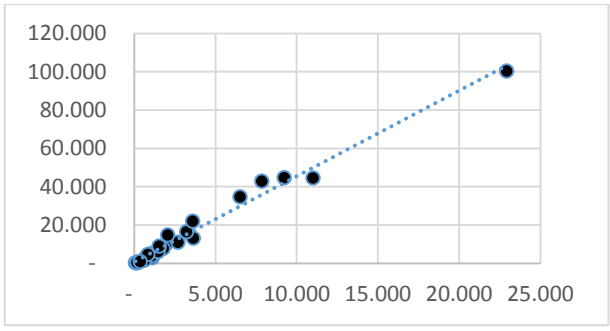
Fonte: Elaboração do autor

Gráfico 22 – Curva de tendência artigo internacional *versus* doutores



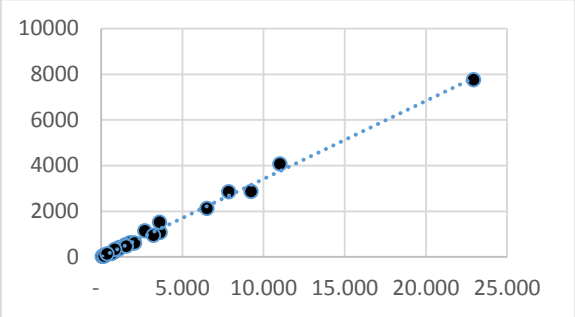
Fonte: Elaboração do autor

Gráfico 23 – Curva de tendência anais de eventos *versus* doutores



Fonte: Elaboração do autor

Gráfico 24 – Curva de tendência livros *versus* doutores



Fonte: Elaboração do autor

Outro fator de confirmação foram os resultados dos coeficientes de correlação, que apontam o modelo linear seguinte como o possível de ser escolhido para continuação da aplicação.

$$Y_i = ß_1 + ß_2 \ X_i + u_i$$

- PASSO 4: ESTIMAÇÃO DOS PARÂMETROS

A tabela e quadro seguintes apresentam os dados para estimação de parâmetros da função “variáveis artigos publicados em periódicos internacionais (Y) e doutores autores (X)”.

Tabela 15– Artigo internacional *versus* doutores

Obs.	UF	X	Y	x	y	x_i^2	y_ix_i
1	Acre	162	206	- 3.091,111	- 11.874,593	9.554.967,901	36.705.685,103
2	Alagoas	760	1.683	- 2.493,111	- 10.397,593	6.215.603,012	25.922.353,621
3	Amapá	65	193	- 3.188,111	- 11.887,593	10.164.052,457	37.898.966,029
4	Amazonas	1.112	3.723	- 2.141,111	- 8.357,593	4.584.356,790	17.894.534,362
5	Bahia	3.622	9.674	368,889	- 2.406,593	136.079,012	- 887.765,267
6	Ceará	1.975	7.556	- 1.278,111	- 4.524,593	1.633.568,012	5.782.932,066
7	Distrito Federal	2.686	8.683	- 567,111	- 3.397,593	321.615,012	1.926.812,510
8	Espírito Santo	979	2.486	- 2.274,111	- 9.594,593	5.171.581,346	21.819.169,621
9	Goiás	1.775	5.419	- 1.478,111	- 6.661,593	2.184.812,457	9.846.574,029
10	Maranhão	593	2.058	- 2.660,111	- 10.022,593	7.076.191,123	26.661.209,918
11	Mato Grosso	1.075	2.157	- 2.178,111	- 9.923,593	4.744.168,012	21.614.687,288
12	Mato Grosso do Sul	1.497	3.894	- 1.756,111	- 8.186,593	3.083.926,235	14.376.566,214
13	Minas Gerais	9.228	34.692	5.974,889	22.611,407	35.699.297,235	135.100.646,881
14	Pará	1.462	4.485	- 1.791,111	- 7.595,593	3.208.079,012	13.604.550,288
15	Paraíba	2.055	5.206	- 1.198,111	- 6.874,593	1.435.470,235	8.236.525,770
16	Paraná	6.508	21.122	3.254,889	9.041,407	10.594.301,679	29.428.776,510
17	Pernambuco	3.215	9.673	- 38,111	- 2.407,593	1.452,457	91.756,029
18	Piauí	626	1.672	- 2.627,111	- 10.408,593	6.901.712,790	27.344.529,251
19	Rio de Janeiro	10.997	42.933	7.743,889	30.852,407	59.967.815,123	238.917.614,918
20	Rio Grande do Norte	1.527	3.930	- 1.726,111	- 8.150,593	2.979.459,568	14.068.828,436
21	Rio Grande do Sul	7.841	30.596	4.587,889	18.515,407	21.048.724,457	84.946.631,918
22	Rondônia	221	723	- 3.032,111	- 11.357,593	9.193.697,790	34.437.482,695
23	Roraima	169	322	- 3.084,111	- 11.758,593	9.511.741,346	36.264.806,066
24	Santa Catarina	3.580	11.203	326,889	- 877,593	106.856,346	- 286.875,267
25	São Paulo	22.922	108.990	19.668,889	96.909,407	386.865.190,123	1.906.100.366,584
26	Sergipe	824	2.088	- 2.429,111	- 9.992,593	5.900.580,790	24.273.117,695
27	Tocantins	358	809	- 2.895,111	- 11.271,593	8.381.668,346	32.632.512,955
soma		**87834**	**326176**	**0**	**0**	**616666968,667**	**2804722996,222**

Fonte: Elaboração do autor

Quadro 4– Estimadores para a correlação artigo internacional *versus* doutores

Artigo internacional *versus* doutores		
Parâmetros	Valores	Função estimada
$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{x}$	$\hat{\beta}_1$ =-2715,20	$Y_i = -2715{,}20 + 4{,}55X_i$
$\hat{\beta}_2 = \frac{\sum(x_i)(y_i)}{\sum(y_i x_i)^2}$	$\hat{\beta}_2$ = 4,55	

Fonte: Elaboração do autor

A tabela e quadro seguintes apresentam os dados para estimação de parâmetros da função “variáveis publicações em anais de eventos (Y) e doutores autores (X)”.

Tabela 16 – Publicações em anais de eventos *versus* doutores

Obs.	UF	X	Y	x	y	x_i^2	$y_i x_i$
1	Acre	162	305	- 3.091,111	- 14.967,630	9.554.967,901	46.266.606,255
2	Alagoas	760	3.568	- 2.493,111	- 11.704,630	6.215.603,012	29.180.942,181
3	Amapá	65	221	- 3.188,111	- 15.051,630	10.164.052,457	47.986.267,663
4	Amazonas	1.112	2.981	- 2.141,111	- 12.291,630	4.584.356,790	26.317.744,774
5	Bahia	3.622	13.189	368,889	- 2.083,630	136.079,012	- 768.627,819
6	Ceará	1.975	10.259	- 1.278,111	- 5.013,630	1.633.568,012	6.407.975,737
7	Distrito Federal	2.686	11.029	- 567,111	- 4.243,630	321.615,012	2.406.609,514
8	Espírito Santo	979	5.013	- 2.274,111	- 10.259,630	5.171.581,346	23.331.537,737
9	Goiás	1.775	7.786	- 1.478,111	- 7.486,630	2.184.812,457	11.066.070,440
10	Maranhão	593	1.722	- 2.660,111	- 13.550,630	7.076.191,123	36.046.180,440
11	Mato Grosso	1.075	3.957	- 2.178,111	- 11.315,630	4.744.168,012	24.646.698,626
12	Mato Grosso do Sul	1.497	7.726	- 1.756,111	- 7.546,630	3.083.926,235	13.252.720,144
13	Minas Gerais	9.228	44.797	5.974,889	29.524,370	35.699.297,235	176.404.832,477
14	Pará	1.462	6.112	- 1.791,111	- 9.160,630	3.208.079,012	16.407.705,514
15	Paraíba	2.055	14.840	- 1.198,111	- 432,630	1.435.470,235	518.338,366
16	Paraná	6.508	34.752	3.254,889	19.479,370	10.594.301,679	63.403.186,181
17	Pernambuco	3.215	16.517	- 38,111	1.244,370	1.452,457	- 47.424,337
18	Piauí	626	1.789	- 2.627,111	- 13.483,630	6.901.712,790	35.422.993,218
19	Rio de Janeiro	10.997	44.523	7.743,889	29.250,370	59.967.815,123	226.511.618,107
20	Rio Grande do Norte	1.527	9.042	- 1.726,111	- 6.230,630	2.979.459,568	10.754.759,033
21	Rio Grande do Sul	7.841	42.966	4.587,889	27.693,370	21.048.724,457	127.054.106,218
22	Rondônia	221	769	- 3.032,111	- 14.503,630	9.193.697,790	43.976.616,551
23	Roraima	169	373	- 3.084,111	- 14.899,630	9.511.741,346	45.952.113,292
24	Santa Catarina	3.580	21.979	326,889	6.706,370	106.856,346	2.192.237,959
25	São Paulo	22.922	100.455	19.668,889	85.182,370	386.865.190,123	1.675.442.578,107
26	Sergipe	824	4.689	- 2.429,111	- 10.583,630	5.900.580,790	25.708.812,329
27	Tocantins	358	1.002	- 2.895,111	- 14.270,630	8.381.668,346	41.315.058,403
soma		**87834**	**412361**	**0**	**0**	**616666968,667**	**2757158257,111**

Fonte: Elaboração do autor

Quadro 5 – Estimadores para a correlação anais de eventos *versus* doutores

Anais de eventos *versus* doutores		
Parâmetros	Valores	Função estimada
$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{x}$	$\hat{\beta}_1$ =727,76	$Y_i = 727{,}76 + 4{,}47X_i$
$\hat{\beta}_2 = \frac{\sum(x_i)(y_i)}{\sum(y_i x_i)^2}$	$\hat{\beta}_2$ =4,47	

Fonte: Elaboração do autor

A tabela e quadro seguintes apresentam os dados para estimação de parâmetros da função “variáveis Livros (Y) e doutores autores (X)”.

Tabela 17 – Livros *versus* doutores

Obs.	UF	X	Y	x	y	x_i^2	$y_i x_i$
1	Acre	162	58	- 3.091,111	- 1.047,852	9.554.967,901	3.239.026,502
2	Alagoas	760	203	- 2.493,111	- 902,852	6.215.603,012	2.250.909,984
3	Amapá	65	14	- 3.188,111	- 1.091,852	10.164.052,457	3.480.945,021
4	Amazonas	1.112	406	- 2.141,111	- 699,852	4.584.356,790	1.498.460,576
5	Bahia	3.622	1068	368,889	- 37,852	136.079,012	- 13.963,128
6	Ceará	1.975	609	- 1.278,111	- 496,852	1.633.568,012	635.031,872
7	Distrito Federal	2.686	1141	- 567,111	35,148	321.615,012	- 19.932,905
8	Espirito Santo	979	309	- 2.274,111	- 796,852	5.171.581,346	1.812.129,650
9	Goiás	1.775	624	- 1.478,111	- 481,852	2.184.812,457	712.230,576
10	Maranhão	593	171	- 2.660,111	- 934,852	7.076.191,123	2.486.809,798
11	Mato Grosso	1.075	344	- 2.178,111	- 761,852	4.744.168,012	1.659.397,984
12	Mato Grosso do Sul	1.497	524	- 1.756,111	- 581,852	3.083.926,235	1.021.796,502
13	Minas Gerais	9.228	2858	5.974,889	1.752,148	35.699.297,235	10.468.890,502
14	Pará	1.462	501	- 1.791,111	- 604,852	3.208.079,012	1.083.356,872
15	Paraiba	2.055	607	- 1.198,111	- 498,852	1.435.470,235	597.679,947
16	Paraná	6.508	2118	3.254,889	1.012,148	10.594.301,679	3.294.429,761
17	Pernambuco	3.215	940	- 38,111	- 165,852	1.452,457	6.320,798
18	Piaui	626	146	- 2.627,111	- 959,852	6.901.712,790	2.521.637,465
19	Rio de Janeiro	10.997	4076	7.743,889	2.970,148	59.967.815,123	23.000.497,243
20	Rio Grande do Norte	1.527	465	- 1.726,111	- 640,852	2.979.459,568	1.106.181,502
21	Rio Grande do Sul	7.841	2846	4.587,889	1.740,148	21.048.724,457	7.983.606,354
22	Rondônia	221	75	- 3.032,111	- 1.030,852	9.193.697,790	3.125.657,354
23	Roraima	169	34	- 3.084,111	- 1.071,852	9.511.741,346	3.305.710,206
24	Santa Catarina	3.580	1510	326,889	404,148	106.856,346	132.111,539
25	São Paulo	22.922	7759	19.668,889	6.653,148	386.865.190,123	130.860.031,687
26	Sergipe	824	327	- 2.429,111	- 778,852	5.900.580,790	1.891.917,687
27	Tocantins	358	125	- 2.895,111	- 980,852	8.381.668,346	2.839.675,095
soma		**87834**	**29858**	**0**	**0**	**616666968,667**	**210980546,444**

Fonte: Elaboração do autor

Quadro 6 – Estimadores para a correlação livros *versus* doutores

Livros *versus* doutores		
Parâmetros	Valores	Função estimada
$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \cdot \bar{x}$	$\hat{\beta}_1$ =-7,14	$Y_i = -7{,}14 + 0{,}34X_i$
$\hat{\beta}_2 = \frac{\sum(x_i)(y_i)}{\sum(y_i x_i)^2}$	$\hat{\beta}_2$ = 0,34	

Fonte: Elaboração do autor

- PASSO 5: ANÁLISE DA REGRESSÃO

Verificamos no quadro abaixo que os coeficientes de regressão indicam um elevado poder de explicação do modelo. Se juntarmos isso ao cálculo do coeficiente de correlação indica um alto grau de associação linear entre as variáveis das publicações nacionais, perceberemos que a forma linear da curva é suficiente para expressar um modelo de previsão.

Quadro 7 – Coeficientes de regressão das publicações nacionais

Determinações	$r^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$
Artigo nacional *versus* doutores	*r² = 0,99*
Artigo internacional *versus* doutores	*r² = 0,98*
Anais de eventos *versus* doutores	*r² = 0,98*
Livros *versus* doutores	*r² = 0,99*

Fonte: Elaboração do autor

- PASSO 6: DEFINIÇÃO DO MODELO DE PREVISÃO

Seguindo os passos anteriores se chega aos quadrados dos erros estocásticos ($\hat{u}_n^2$) e, a partir daí, podemos construir o modelo de previsão da produção científica nacional desta tese, que se vale da análise dos erros estocásticos ao quadrado ($\hat{u}_i^2$) para que se possa seguir a função geral desta tese, que é:

$$\hat{Y}_i = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2).$$

Ou, mais especificamente, às funções "artigos publicados em periódicos internacionais (Y) e doutores autores (X)", "anais de eventos (Y) e doutores autores (X)" e "livros (Y) e doutores autores (X)".

.

$$\hat{Y}_{Artigos\ internacionais} = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2).$$

$$\hat{Y}_{Anais\ \text{de eventos}} = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2).$$

$$\hat{Y}_{Livros} = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2).$$

Portanto, definir o modelo de projeção ou previsão neste estudo consiste na análise da produção científica brasileira a partir dos erros estocásticos ($\hat{u}_i^2$), ou seja, $\hat{Y}_i = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$. Os erros ($\hat{u}_n^2$) indicam o quanto um valor Y se aproxima ou se afasta da estimativa esperada $\hat{Y}_i$, conforme demonstrado nas seguintes tabelas para as publicações que não foram abordadas no Capítulo 4 (artigos internacionais, anais de eventos e livros).

Tabela 18 – Dados para a função artigo internacional *versus* doutores com os erros estocásticos ($\hat{u}_i^2$)

X_i^2	Y_i^2	$\hat{Y}_i$	$\hat{u}_i = Y_i - \hat{Y}_i$	$\hat{u}_i^2$
26.244	42.436	- 1.978,390	2.184,390	4.771.561,367
577.600	2.832.489	741,432	941,568	886.551,140
4.225	37.249	- 2.419,566	2.612,566	6.825.498,591
1.236.544	13.860.729	2.342,397	1.380,603	1.906.064,708
13.118.884	93.586.276	13.758,372	- 4.084,372	16.682.094,746
3.900.625	57.093.136	6.267,491	1.288,509	1.660.254,973
7.214.596	75.394.489	9.501,259	- 818,259	669.548,462
958.441	6.180.196	1.737,487	748,513	560.272,092
3.150.625	29.365.561	5.357,852	61,148	3.739,110
351.649	4.235.364	- 18,117	2.076,117	4.310.263,391
1.155.625	4.652.649	2.174,114	- 17,114	292,878
2.241.009	15.163.236	4.093,453	- 199,453	39.781,463
85.155.984	1.203.534.864	39.255,566	- 4.563,566	20.826.131,555
2.137.444	20.115.225	3.934,266	550,734	303.307,933
4.223.025	27.102.436	6.631,347	- 1.425,347	2.031.613,958
42.354.064	446.138.884	26.884,469	- 5.762,469	33.206.051,371
10.336.225	93.566.929	11.907,256	- 2.234,256	4.991.898,724
391.876	2.795.584	131,973	1.540,027	2.371.682,776
120.934.009	1.843.242.489	47.301,327	- 4.368,327	19.082.276,892
2.331.729	15.444.900	4.229,899	- 299,899	89.939,305
61.481.281	936.115.216	32.947,216	- 2.351,216	5.528.217,196
48.841	522.729	- 1.710,047	2.433,047	5.919.716,496
28.561	103.684	- 1.946,553	2.268,553	5.146.332,748
12.816.400	125.507.209	13.567,348	- 2.364,348	5.590.140,189
525.418.084	11.878.820.100	101.538,578	7.451,422	55.523.682,751
678.976	4.359.744	1.032,516	1.055,484	1.114.046,104
128.164	654.481	- 1.086,944	1.895,944	3.594.602,634
902400730	**16900468284**	**326176**	**0**	**203635563,554**

Fonte: Elaboração do autor

Com essa análise é possível estabelecer um *ranking* dos estados brasileiros que apresentam maior eficiência na produção de publicações, isto é, os estados que apresentam $\hat{u}_i^2$ menores se aproximam mais de uma combinação ótima entre número de doutores e publicações de artigos internacionais.

O modelo possibilita análises a partir dos erros estocásticos ao quadrado ($\hat{u}_i^2$), onde os $\hat{u}_i^2$ com menor valor indicam a posição do estado dentro do *ranking*. Isso permite medir os mais eficientes em termos de produção científica de artigos internacionais, expondo do mais eficiente ao menos eficiente, conforme segue:

Tabela 19 – Classificação dos estados produtores de artigos internacionais a partir dos erros estocásticos ($\hat{u}_i^2$), seguinte função geral: $\hat{Y}_i = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$

UF	$\hat{u}_i^2$
Mato Grosso	292,878
Goiás	3.739,110
Mato Grosso do Sul	39.781,463
Rio Grande do Norte	89.939,305
Pará	303.307,933
Espírito Santo	560.272,092
Distrito Federal	669.548,462
Alagoas	886.551,140
Sergipe	1.114.046,104
Ceará	1.660.254,973
Amazonas	1.906.064,708
Paraíba	2.031.613,958
Piauí	2.371.682,776
Tocantins	3.594.602,634
Maranhão	4.310.263,391
Acre	4.771.561,367
Pernambuco	4.991.898,724
Roraima	5.146.332,748
Rio Grande do Sul	5.528.217,196
Santa Catarina	5.590.140,189
Rondônia	5.919.716,496
Amapá	6.825.498,591
Bahia	16.682.094,746
Rio de Janeiro	19.082.276,892
Minas Gerais	20.826.131,555
Paraná	33.206.051,371
São Paulo	55.523.682,751

Fonte: Elaboração do autor

Observe-se que os erros ao quadrado indicam o quanto a produção de publicações de artigos internacionais se aproxima ou se afasta da estimativa esperada. Perceba-se que é possível estabelecer um *ranking* dos estados que apresentam números mais eficientes de produtividade, que são os estados que apresentam $\hat{u}_i^2$ menores. Estes se aproximam mais de uma combinação ótima entre número de doutores e publicações de artigos em periódicos internacionais.

Então aparecem encabeçando a lista Mato Grosso, Goiás, Mato Grosso do Sul, Rio Grande do Norte, Pará, Espírito Santo, Distrito Federal, Alagoas, Sergipe e Ceará, como os 10 estados onde a relação da produtividade (doutores *versus* artigos internacionais publicados) mais se aproxima da reta de regressão. Na outra ponta, Roraima, Rio Grande do Sul, Santa Catarina, Rondônia, Amapá, Bahia, Rio de Janeiro, Mato Grosso, Paraná e São Paulo são os 10 estados que mais se afastam da reta. Esses dados explicam o fator de previsibilidade da produção de artigos internacionais, uma vez que quanto menor o erro mais próximo estará o valor observado

do valor esperado. Repare-se que o estado do Maranhão apresenta a melhor combinação para artigos nacionais, já para artigos internacionais ele fica em 15º lugar.

Tabela 20 7 – Dados para a função anais de eventos *versus* doutores com os erros estocásticos ($\hat{u}_i^2$)

X_i^2	Y_i^2	$\hat{Y}_i$	$\hat{u}_i = Y_i - \hat{Y}_i$	$\hat{u}_i^2$
26.244	93.025	1.452,070	- 1.147,070	1.315.769,914
577.600	12.730.624	4.125,767	- 557,767	311.104,211
4.225	48.841	1.018,377	- 797,377	635.809,783
1.236.544	8.886.361	5.699,582	- 2.718,582	7.390.688,825
13.118.884	173.949.721	16.921,956	- 3.732,956	13.934.959,926
3.900.625	105.247.081	9.558,111	700,889	491.244,760
7.214.596	121.638.841	12.737,039	- 1.708,039	2.917.396,697
958.441	25.130.169	5.104,930	- 91,930	8.451,209
3.150.625	60.621.796	8.663,898	- 877,898	770.705,598
351.649	2.965.284	3.379,099	- 1.657,099	2.745.977,984
1.155.625	15.657.849	5.534,153	- 1.577,153	2.487.410,704
2.241.009	59.691.076	7.420,942	305,058	93.060,226
85.155.984	2.006.771.209	41.986,748	2.810,252	7.897.517,769
2.137.444	37.356.544	7.264,455	- 1.152,455	1.328.152,467
4.223.025	220.225.600	9.915,797	4.924,203	24.247.778,436
42.354.064	1.207.701.504	29.825,450	4.926,550	24.270.892,464
10.336.225	272.811.289	15.102,232	1.414,768	2.001.567,461
391.876	3.200.521	3.526,644	- 1.737,644	3.019.408,135
120.934.009	1.982.297.529	49.896,062	- 5.373,062	28.869.797,129
2.331.729	81.757.764	7.555,074	1.486,926	2.210.948,288
61.481.281	1.846.077.156	35.785,380	7.180,620	51.561.300,259
48.841	591.361	1.715,863	- 946,863	896.549,528
28.561	139.129	1.483,368	- 1.110,368	1.232.916,208
12.816.400	483.076.441	16.734,171	5.244,829	27.508.229,322
525.418.084	10.091.207.025	103.213,515	- 2.758,515	7.609.406,839
678.976	21.986.721	4.411,915	277,085	76.775,908
128.164	1.004.004	2.328,399	- 1.326,399	1.759.334,130
902400730	**18842864465**	**412361**	**0**	**217593154,182**

Fonte: Elaboração do autor

A análise dos dados da tabela anterior permite estabelecer um *ranking* dos estados brasileiros que apresentam maior eficiência na produção de publicações em anais de eventos. Os estados que apresentam $\hat{u}_i^2$ menores se aproximam mais de uma combinação ótima entre número de doutores e publicações em anais de eventos. Nosso modelo possibilita análises a partir dos erros estocásticos ao quadrado ($\hat{u}_i^2$), onde os $\hat{u}_i^2$ com menor valor indicam a posição do estado dentro do *ranking*. Isso permite medir os mais eficientes em termos de produção científica de anais de eventos, expondo do mais eficiente ao menos eficiente, conforme segue:

Tabela 21 – Classificação dos estados produtores de anais de eventos a partir dos erros estocásticos ($\hat{u}_i^2$), seguinte função geral: $\hat{Y}_i = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$

UF	$\hat{u}_i^2$
Espírito Santo	8.451,209
Sergipe	76.775,908
Mato Grosso do Sul	93.060,226
Alagoas	311.104,211
Ceará	491.244,760
Amapá	635.809,783
Goiás	770.705,598
Rondônia	896.549,528
Roraima	1.232.916,208
Acre	1.315.769,914
Pará	1.328.152,467
Tocantins	1.759.334,130
Pernambuco	2.001.567,461
Rio Grande do Norte	2.210.948,288
Mato Grosso	2.487.410,704
Maranhão	2.745.977,984
Distrito Federal	2.917.396,697
Piauí	3.019.408,135
Amazonas	7.390.688,825
São Paulo	7.609.406,839
Minas Gerais	7.897.517,769
Bahia	13.934.959,926
Paraíba	24.247.778,436
Paraná	24.270.892,464
Santa Catarina	27.508.229,322
Rio de Janeiro	28.869.797,129
Rio Grande do Sul	51.561.300,259

Fonte: Elaboração do autor

Observe-se que os erros ao quadrado indicam o quanto a produção de publicações de anais de eventos se aproxima ou se afasta da estimativa esperada. Perceba-se que é possível estabelecer um *ranking* dos estados que apresentam números mais eficientes de produtividade, que são os estados que apresentam $\hat{u}_i^2$ menores. Estes se aproximam mais de uma combinação ótima entre número de doutores e publicações de anais de eventos.

Então aparecem encabeçando a lista Espírito Santo, Sergipe, Mato Grosso do Sul, Alagoas, Amapá, Goiás, Rondônia, Roraima, Acre e Pará, como os 10 estados onde a relação da produtividade (doutores *versus* anais de eventos publicados) mais se aproxima da reta de regressão. Na outra ponta, Piauí, Amazonas, São Paulo, Minas Gerais, Bahia, Paraíba, Paraná, Santa Catarina, Rio de Janeiro e Rio Grande do Sul são os 10 estados que mais se afastam da reta. Esses dados explicam o fator de previsibilidade da produção de anais de eventos, uma vez

que quanto menor o erro, mais próximo estará o valor observado do valor esperado. Repare-se que o estado do Maranhão apresenta a melhor combinação para artigos nacionais, já para anais de eventos ele fica em 16º lugar.

Tabela 22 – Dados para a função livros *versus* doutores com os erros estocásticos ($\hat{u}_i^2$)

X_i^2	Y_i^2	$\hat{Y}_i$	$\hat{u}_i = Y_i - \hat{Y}_i$	$\hat{u}_i^2$
26.244	3.364	48,289	9,711	94,311
577.600	41.209	252,883	- 49,883	2.488,277
4.225	196	15,102	- 1,102	1,214
1.236.544	164.836	373,313	32,687	1.068,470
13.118.884	1.140.624	1.232,060	- 164,060	26.915,675
3.900.625	370.881	668,571	- 59,571	3.548,719
7.214.596	1.301.881	911,826	229,174	52.520,780
958.441	95.481	327,809	- 18,809	353,786
3.150.625	389.376	600,145	23,855	569,059
351.649	29.241	195,747	- 24,747	612,406
1.155.625	118.336	360,654	- 16,654	277,346
2.241.009	274.576	505,033	18,967	359,756
85.155.984	8.168.164	3.150,043	- 292,043	85.289,269
2.137.444	251.001	493,058	7,942	63,072
4.223.025	368.449	695,942	- 88,942	7.910,601
42.354.064	4.485.924	2.219,448	- 101,448	10.291,787
10.336.225	883.600	1.092,813	- 152,813	23.351,776
391.876	21.316	207,037	- 61,037	3.725,534
120.934.009	16.613.776	3.755,272	320,728	102.866,431
2.331.729	216.225	515,297	- 50,297	2.529,756
61.481.281	8.099.716	2.675,508	170,492	29.067,408
48.841	5.625	68,474	6,526	42,585
28.561	1.156	50,684	- 16,684	278,340
12.816.400	2.280.100	1.217,690	292,310	85.444,847
525.418.084	60.202.081	7.835,178	- 76,178	5.803,031
678.976	106.929	274,779	52,221	2.727,035
128.164	15.625	115,346	9,654	93,196
902400730	**105649688**	**29858**	**0**	**448294,469**

Fonte: Elaboração do autor

A análise dos dados da tabela anterior permite estabelecer um *ranking* dos estados brasileiros que apresentam maior eficiência na produção de publicações de livros. Os estados que apresentam $\hat{u}_i^2$ menores se aproximam mais de uma combinação ótima entre número de doutores e publicações de livros. Nosso modelo possibilita análises a partir dos erros estocásticos ao quadrado ($\hat{u}_i^2$), onde os $\hat{u}_i^2$ com menor valor indicam a posição do estado dentro

do *ranking*. Isso permite medir os mais eficientes em termos de produção científica de livros, expondo do mais eficiente ao menos eficiente, conforme segue:

Tabela 23 – Classificação dos estados produtores de livros a partir dos erros estocásticos ($\hat{u}_i^2$), seguinte função geral: $\hat{Y}_i = f\ (\hat{u}_1^2, \hat{u}_2^2, \hat{u}_3^2 ..... \hat{u}_n^2)$

UF	$\hat{u}_i^2$
Amapá	1,214
Rondônia	42,585
Pará	63,072
Tocantins	93,196
Acre	94,311
Mato Grosso	277,346
Roraima	278,340
Espírito Santo	353,786
Mato Grosso do Sul	359,756
Goiás	569,059
Maranhão	612,406
Amazonas	1.068,470
Alagoas	2.488,277
Rio Grande do Norte	2.529,756
Sergipe	2.727,035
Ceará	3.548,719
Piauí	3.725,534
São Paulo	5.803,031
Paraíba	7.910,601
Paraná	10.291,787
Pernambuco	23.351,776
Bahia	26.915,675
Rio Grande do Sul	29.067,408
Distrito Federal	52.520,780
Minas Gerais	85.289,269
Santa Catarina	85.444,847
Rio de Janeiro	102.866,431

Fonte: Elaboração do autor

Observe-se que os erros ao quadrado indicam o quanto a produção de publicações de livros se aproxima ou se afasta da estimativa esperada. Perceba-se que é possível estabelecer um *ranking* dos estados que apresentam números mais eficientes de produtividade, que são os estados que apresentam $\hat{u}_i^2$ menores. Estes se aproximam mais de uma combinação ótima entre número de doutores e publicações de livros.

Então aparecem encabeçando a lista Amapá, Rondônia, Pará, Tocantins, Acre, Mato Grosso, Roraima, Espírito Santo, Mato Grosso do Sul e Goiás, como os 10 estados onde a relação da produtividade (livros publicados) mais se aproxima da reta de regressão. Na outra ponta, São Paulo, Paraíba, Paraná, Pernambuco, Rio Grande do Sul, Distrito Federal, Minas Gerais, Santa Catarina e Rio de Janeiro são os 10 estados que mais se afastam da reta. Esses dados explicam o fator de previsibilidade da produção de livros, uma vez que quanto menor o erro, mais próximo estará o valor observado do valor esperado. Repare-se que o estado do Maranhão apresenta a melhor combinação para artigos nacionais, já para livros ele fica em 11º lugar.