3. Modelos causais

Serão abordados nesse capítulo os modelos causais, utilizados como ferramentas para o desenvolvimento e análise desse trabalho.

3.1. Econometria

A base da Econometria é a aplicação de métodos matemáticos e estatísticos à análise de conjunto de dados históricos, visando dar suporte às teorias econômicas. Em realidade, define-se a Econometria como um método de análise econômica que agrega a Matemática, Teoria Econômica e Estatística. Evidentemente, não se constitui numa ciência adicional a congregação dessas 3 ciências, mas sim uma forma científica de traduzir um modelo teórico para uma formulação que possa ser passível de um teste empírico.

Ela é o ramo da Economia que lida com a mensuração de realizações econômicas, ou seja, relações entre variáveis de natureza econômicas. Segundo Michaelis (2004), Econometria é definida como o estudo dos fenômenos econômicos mediante a aplicação de métodos matemáticos e técnicas estatísticas, para verificar até que ponto as teorias dos ciclos econômicos encontram apoio na realidade concreta.

De outra fonte, a Econometria é, na verdade, uma combinação de teoria ou outro raciocínio a *priori* com matemática e estatística, com o objetivo de dar conteúdo empírico às formulações teóricas da Economia, Matos (2000).

Encontram-se na literatura duas abordagens para a modelagem econométrica, a abordagem tradicional, cujo método consiste em tomar um modelo econométrico simples, capaz de admitir a introdução de maiores generalizações, desenvolvida principalmente por Koopmans (1957), que procura, como mencionado no parágrafo acima *testar* um modelo previamente fornecido pela teoria e o paradigma alternativo apresentado por Hendry (1987), que se constitui exatamente na atitude inversa à do primeiro método: partindo-se dos dados existentes, tenta-se montar um modelo o mais adequado possível à ”história” contada pelos dados.

3.2. Processos Estocásticos

Um processo estocástico é uma variável que se desenvolve no tempo de uma maneira que é pelo menos parcialmente aleatória e imprevisível. De uma maneira mais formal, um processo estocástico é definido por uma lei de probabilidade para a evolução de uma variável x durante um tempo t.

3.2.1. Processos Estacionários

Um tipo de processo estocástico que recebe grande atenção pelos analistas é o processo estocástico estacionário. Um processo é assim chamado se sua média e variância são constantes ao longo do tempo, e o valor da covariância entre dois períodos de tempo dependem somente do tempo entre esses dois períodos e não do instante em que a covariância é computada. A estacionariedade assim definida é a de 2ª ordem.

Ilustrando:

$$Média : \mu = E(Y_t)$$

$$Variância : \text{var}(Y_t) = \sigma^2 = E(Y_t - \mu)^2$$

$$Co\,\text{var}\,iância : \gamma_k = E[(Y_t - \mu)(Y_{t+k} - \mu)]$$

Porque os processos estacionários são tão importantes? Porque se uma série é não-estacionária, pode-se estudá-la apenas ao longo do período sob consideração. Cada conjunto de dados deverá ser tratado em particular, e não é possível generalizar o comportamento da curva para outros períodos. Logo, para efeitos de previsão de futuro as séries não-estacionárias serão de pouca valia.

Como pode-se avaliar se uma série é estacionária? Antes de prosseguir, eis um tipo especial de série estocástica estacionária, chamada de puramente randômica ou de ruído branco. Nesse caso específico, a média é zero, a variância é constante e ela é serialmente não correlacionada.

3.2.2. Processos Não-Estacionários

Embora o interesse principal desse estudo sejam os processos estocásticos, frequentemente processos não-estocásticos representam a realidade. Exemplos clássicos desses processos são o de variáveis financeiras, como a bolsa de valores ou taxas de câmbio. Eles se dividem em dois grandes tipos: com ou sem tendência.

3.2.3. Testes de Estacionariedade

Uma pergunta que cabe ser colocada agora é: como saber se uma série é ou não estacionária? Ao se descobrir que uma séria é não-estacionária, há alguma maneira de torná-la estacionária? Apresentam-se as duas questões a seguir.

a) Análise Gráfica

A primeira análise que se pode fazer (na verdade deve-se fazer), antes dos testes formais, é plotar o gráfico da série ao longo do tempo e analisá-la. É particularmente fácil de verificar quando há uma tendência na média, tanto de alta quando de baixa. Obviamente, a análise aqui é muito subjetiva, mas muitas vezes já dá uma boa idéia do que pode-se encontrar nas análises seguintes.

b) Função de Autocorrelação e Correlograma

Um teste simples de estacionariedade é o teste da autocorrelação.

Define-se:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\text{cov}\,ariância_k}{\text{var}\,iância}$$

Notar que se k=0, ρ_0=1. Como ambas são medidas na mesma unidade, ρ_k é adimensional, e é compreendido entre -1 e +1, como qualquer outra correlação. Plotando-se ρ_k contra k, o gráfico é conhecido como correlograma da população.

Determinando a covariância amostral no período k, $\hat{\gamma}_k$, bem como a variância amostral, $\hat{\gamma}_0$:

$$\hat{\gamma}_k = \frac{\sum(\gamma_t - \bar{\gamma})(\gamma_{t+k} - \bar{\gamma})}{n}$$

$$\hat{\gamma}_0 = \frac{\sum(\gamma_t - \bar{\gamma})^2}{n}$$

onde:

n: número de pontos da amostra (tamanho da amostra)

$\bar{Y}$: média dos pontos da amostra

$\hat{\gamma}$: estimador de $\hat{\gamma}_k$

Com isso, a função de correlação amostral em k é:

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$$

onde:

$\hat{\gamma}_k$: estimador de $\hat{\rho}_k$

Plotando-se $\hat{\rho}_k$ contra k, o gráfico é conhecido como correlograma da amostra.

Então, como o correlograma da amostra indica se uma série é ou não estacionária? Uma análise, também gráfica, deve ser feita na plotagem dos pontos em torno de zero, da autocorrelação. Um correlograma de uma série

estacionária faz com que os pontos orbitem em torno de zero, randomicamente para os dois lados (positivo e negativo).

Também deve-se analisar a significância dos coeficientes $\hat{\rho}_k$. Para avaliar se os coeficientes são significantes ou não, deve-se analisar seu desvio padrão. Segundo Bartlet (1946), se uma série de dados é puramente randômica, ou seja, apresenta ruído branco, os coeficientes de autocorrelação $\hat{\rho}_k$. podem ser expressos aproximadamente por:

$$\hat{\rho}_k \sim N\left(0, \frac{1}{n}\right)$$

Isso significa que, em amostras grandes, os coeficientes de autocorrelações são normalmente distribuídos com média zero e variância igual a um sobre o tamanho da amostra.

Nesse caso, por exemplo com um intervalo de confiança de 95% para qualquer população $\hat{\rho}_k$, tem-se:

$$\hat{\rho}_k \pm 1{,}96\left(\frac{1}{\sqrt{n}}\right)$$

Onde:

n: número de observações da amostra.

Em outras palavras:

$$\Pr ob\left[\hat{\rho} - 1{,}96\left(\frac{1}{\sqrt{n}}\right) \leq \rho_k \leq \hat{\rho}_k + 1{,}96\left(\frac{1}{\sqrt{n}}\right)\right] = 0{,}95$$

Se o intervalo incluir o valor zero (dado um $\hat{\rho}_k$), não se rejeita a hipótese de ρ_k ser zero. Entretanto, se o intervalo encontrado não incluir o valor zero, rejeita-se a hipótese do verdadeiro ρ_k ser zero.

c) Testes de raiz unitária

Os testes de raiz unitárias se tornaram bastante populares nos últimos tempos. O ponto inicial desse teste é a fórmula:

$$Y_t = \rho . Y_{t-1} + \mu_t \qquad -1 \leq \rho \leq 1$$

Onde:

μ_t: termo de erro de ruído branco.

Sabe-se que se ρ=1, ou seja, no caso de raiz unitária se configura num Passeio Aleatório sem tendência ("*Random Walk model without drift*"), que é um processo não estacionário. A idéia por trás do teste de raiz unitária é regredir Y_t contra Y_{t-1}, e descobrir se o ρ estimado é igual a 1. Se assim for, então Y_t é não estacionária. Essa basicamente é a idéia por trás dos testes de raiz unitária.

Nessa linha, há diversas maneiras de efetuar o teste, como o teste *Augmented Dickey-Fuller* (ADF), o *Phillips-Perron* (PP) , entre outros.

3.2.4. Transformando séries não-estacionárias

Agora que já se conhece os problemas relacionados com séries não estacionárias, e sabe-se identificar quando uma série é ou não estacionária, a questão seguinte é o que fazer quando a série é não estacionária?

Existem dois tipos básicos de não-estacionariedade: Diferença Estacionária (DSP – "*Difference Stationary")* e Tendência Estacionária (TSP – "*Trend Stationary*").

No primeiro caso, onde as séries tem uma raiz unitária, o primeiro diferencial dessa série é estacionário. Logo, a solução aqui é tomar o primeiro diferencial da série.

Já no segundo caso, onde há tendência que impede a série de ser estacionária, basta fazer a regressão no tempo e os resíduos dessa série serão estacionários.

3.2.5. O Vetor Autoregressivo (VAR)

Nas análises precedentes, os modelos consideram algumas variáveis como endógenas e outras como exógenas. Antes de analisar esses modelos, é necessário ter certeza de que as equações do sistema estão corretamente identificadas, e essa identificação é freqüentemente obtida assumindo que algumas variáveis estão presentes somente em algumas das equações. Essa decisão é extremamente subjetiva e foi severamente criticada por Sims (1980).

De acordo com Sims, se há alguma simultaneidade entre um conjunto de variáveis, elas devem ser tratadas de maneira igual, não deveria haver distinção entre variáveis endógenas e exógenas. Nesse espírito foi criado o modelo VAR

Para estimar o modelo VAR, basea-se na seguinte fórmula geral:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \dots \\ y_{k,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \dots \\ c_k \end{bmatrix} + \begin{bmatrix} a_{1,1}^1 \, a_{1,2}^1 \dots a_{1,k}^1 \\ a_{1,1}^1 \, a_{1,2}^1 \dots a_{1,k}^1 \\ \dots \quad \dots \quad \dots \quad \dots \\ a_{1,1}^1 \, a_{1,2}^1 \dots a_{1,k}^1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \dots \\ y_{k,t-1} \end{bmatrix} + \dots + \begin{bmatrix} a_{1,1}^p \, a_{1,2}^p \dots a_{1,k}^p \\ a_{1,1}^p \, a_{1,2}^p \dots a_{1,k}^p \\ \dots \quad \dots \quad \dots \quad \dots \\ a_{1,1}^p \, a_{1,2}^p \dots a_{1,k}^p \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \dots \\ y_{k,t-p} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \\ \dots \\ e_{k,t} \end{bmatrix}$$

Notar que haverá uma equação para cada variável do sistema. No teste VAR, supõe-se que todas as variáveis tenham influência nas outras, ou seja, y_1 é influenciada por y_2, $y_3 \dots y_k$, y_2 é influenciada por y_1, $y_3 \dots y_k$, , y_k é influenciada por y_1, $y_2 \dots y_{k-1}$. As variáveis *e* são os erros estocásticos, chamados também de impulsos e choques na linguagem VAR.

Há que se decidir o tamanho da série (*lag*). É uma questão empírica, pois ao se obter um número muito grande, são consumidos muitos graus de liberdade, além de aumentar a possibilidade de introduzir uma multicolinearidade. Ao se obter um número pequeno, pode-se induzir a um erro de especificação. Uma maneira que auxilia nessa decisão é o método *Akaike* e *Schwarz*, e escolher o modelo que minimiza esses critérios. Não há dúvida que haverá tentativa e erro no processo.

Uma vez definido o número de *lags*, propõe-se a utilização do método VAR para previsões. Os coeficientes que são obtidos servirão de base para a equação de previsão.

3.3. Modelo de Regressão Linear

O método mais tradicional e comum da Econometria é a análise de regressão, cuja evolução foi associada ao desenvolvimento de técnicas de estimação com base no clássico modelo de regressão. Pode-se atribuir o sucesso desse método por conta de sua relativa simplicidade e facilidade computacional.

O referido método consiste em encontrar uma equação que melhor represente a relação entre duas ou mais variáveis, com base em uma série de dados. Após determinada essa equação, pode-se utilizar a mesma para previsões a respeito dos valores de uma das variáveis, dadas as demais.

Por conta da aplicação considerada pioneira por Francis Galton (1888), consistindo na estimação de uma reta que representasse a relação entre a altura dos filhos e a dos respectivos pais, homenageou-se denominando o termo *regressão*. Pelo fato de ser uma reta ou uma função linear das diversas variáveis envolvidas, utiliza-se então o termo *regressão linear*. Os dois autores pretendiam provar a suposta validade de uma "lei da regressão universal", que regia que as características do ser humano eram transmitidas hereditariamente de pai para filho de forma amortecida.

Considerando-se que o método consistia na estimação de relações, passou a ser utilizado bastante na estimação de relações entre variáveis econômicas, principalmente em funções demanda. Uma das pioneiras estimações de demanda

realizada foi atribuída ao estatístico italiano Benini, que em 1907 estimou uma demanda de café como função dos preços do café e do açúcar.

O objetivo dos modelos de regressão é estabelecer relações estatísticas entre um fenômeno em estudo e as variáveis independentes envolvidas, chamadas forças direcionadoras, que exercem influência sobre ele. Sendo assim, o modelo suporta a inclusão de variáveis exógenas como as sócio-econômicas. Isto contribui para o entendimento do fenômeno em estudo, mas é insuficiente para explicá-lo, pois a identificação de um relacionamento estatístico entre duas variáveis por si só não estabelece um relacionamento causal entre elas. Por exemplo, pode-se identificar através de um modelo de regressão que o crescimento populacional tem relação com o crescimento do desmatamento de uma determinada região, entretanto, o modelo de regressão não explica os mecanismos que ligam estas variáveis.

Dada uma relação funcional definida com base na teoria econômica, assume-se que Y é uma variável dependente explicada por um conjunto de variáveis independentes X_1, X_2, ..., X_k e por um termo aleatório ε, que representa a soma de todos os demais fatores que afetam a variável dependente Y, além de X_1, X_2, ..., X_k, mas que não estão presentes no modelo. Pressupõe-se, também, que o modelo é inerentemente linear, significando que só se aplica a equações lineares, ou possíveis de linearização.

Matematicamente, o modelo estabelece um relacionamento linear entre as variáveis dependentes e independentes através da expressão, segundo Hills, Griffiths, Judge (2003):

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

onde:

Y: variável dependente, cujo comportamento é explicado pela variável *X* (variável independente ou explicativa);

X: variável independente ou explicativa*;*

$\beta_{0:}$ interseção de *Y*;

β_1: inclinação de Y em relação à variável X_1, mantendo constantes as variáveis $X_2, X_3, ..., X_k$;

β_2: inclinação de Y em relação à variável X_2, mantendo constantes as variáveis $X_1, X_3, ..., X_k$;

β_k: inclinação de Y em relação à variável X_k, mantendo constantes as variáveis $X_1, X_2, ..., X_{k-1}$;

ε : erro *aleatório* em Y, para a observação i.

Outra restrição desse modelo é que ele se aplica apenas a processos estacionários. Será abordada em seguida a metodologia para avaliar se a série é estacionária ou não, e alguns métodos que podem "corrigir" a série, na hipótese da série não ser estacionária.

Um bom exemplo de modelo de regressão foi o apresentado por Reis e Margulis (1991), modelando o desmatamento da região Amazônica em função da densidade espacial das atividades econômicas da região. Numa primeira abordagem, as áreas desmatadas são relacionadas com a densidade populacional, áreas cultivadas, distância de centros urbanos e proximidade de rodovias, entre outras variáveis consideradas. Em seguida, num segundo estágio o modelo relaciona o crescimento de determinadas atividades (colonização, cultivo, pecuária entre 1980 e 1985 com a densidade destas atividades em 1980, obtendo dessa maneira o padrão de crescimento espacial de cada atividade. Com isso, e partindo do pressuposto de que este padrão de crescimento irá se manter no futuro, o modelo faz projeções sobre a tendência de desmatamento para o período compreendido entre o ano de 1985 e 2000.

Esse tipo de análise busca avaliar se existe uma associação entre alguma variável (ou variáveis) independente e a variável primária (ou dependente), de maneira que seja possível prever o desfecho a partir da variável independente. Para exemplificar, pode-se determinar se existe associação positiva entre peso e estatura, ou seja, quanto maior a estatura de uma pessoa, maior seu peso. Desta forma, pode-se definir uma equação de regressão que permita estimar o peso de uma pessoa, a partir do conhecimento de sua estatura. Evidentemente, a

correlação não é perfeita, e nem sempre funciona, principalmente quando se usa uma equação de correlação construída para determinada população. Por exemplo, fazendo-se essa avaliação para as crianças norte-americanas o resultado será bastante distinto se aplicar-se a mesma metodologia para crianças na África. Entretanto, observando-se alguns cuidados e restrições, as equações de regressão encontradas podem ser utilizadas para fazer algumas estimativas de valor. A regressão pode levar em conta não apenas a relação entre uma única variável independente, mas também pode levar em conta diversas variáveis independentes ao mesmo tempo, quando será chamada de regressão múltipla.

De acordo com Matos (2000), a análise de regressão linear em Econometria compreende quatro etapas principais:

Especificação do modelo: toma por base relações definidas por meio da teoria econômica, tais como equações de demanda, funções de produção, funções custo e modelos macroeconômicos.

Estimação: feita utilizando-se métodos econométricos que permitem estimar, para dada amostra de informações, parâmetros das relações econômicas como elasticidades e propensões a consumir, a poupar etc.

Análise de resultados: consiste na análise de resultados por meio de verificação de sua adequação à teoria subjacente e testes estatísticos e econométricos.

Fase final: utilização dos resultados para previsão.

Os métodos mais utilizados para estimação dos parâmetros da regressão são os dos Mínimos Quadrados Ordinários e da Máxima Verossimilhança, sendo o mais difundido o dos Mínimos Quadrados Ordinários (MQO).

3.3.1. Método dos Mínimos Quadrados

A análise da regressão linear basicamente busca encontrar uma linha reta que melhor se ajusta aos dados informados. Esse melhor ajuste em teoria poderia ser definido por intermédio de uma variedade de formas. O mais simples deles envolve encontrar a linha reta para qual as diferenças dos verdadeiros valores (Y_i) e os valores que seriam previstos a partir da linha ajustada de regressão ($\hat{Y}$) sejam

as menores possíveis. Cabe salientar que pelo fato das diferenças poderem ser positivas para algumas observações e negativas para outras, a soma das diferenças dos quadrados será minimizada. Pindyck, Rubinfeld (2001)

$$\text{(soma das diferenças ao quadrado)} = \sum_{i=1}^{n} (Yi - \hat{Y}i)^2$$

onde

Y_i : verdadeiro valor observado de Y para o dado i

$\hat{Y}i$: valor previsto de Y para a observação i

O método dos mínimos quadrados representa uma técnica matemática que determina quais os valores de β_1, β_2, ..., β_k que minimizam a soma das diferenças ao quadrado. Quaisquer valores correspondentes a β_1, β_2, ..., β_k, que não sejam aqueles determinados através do método dos mínimos quadrados, resultam em uma soma mais elevada das diferenças ao quadrado entre o verdadeiro valor de Y e o previsto de Y. Pindyck, Rubinfeld (2001)

3.3.2. Propriedades amostrais dos estimadores de Mínimos Quadrados

Os estimadores de mínimos quadrados *a* e *b* ($Y = a + bX$) são variáveis aleatórias e têm distribuições de probabilidade que podem ser estudadas antes da coleta de quaisquer dados. Segundo Pindyck e Rubinfeld (2001), as características de suas funções densidade de probabilidade são de grande interesse. Se essas funções são conhecidas, podem ser usadas para fazer afirmações probabilísticas sobre *a* e *b*. As médias (valores esperados) e as variâncias das variáveis aleatórias caracterizam a localização e a dispersão de suas distribuições de probabilidade. As médias e as variâncias de *a* e *b* informam os intervalos de valores que *a* e *b* provavelmente tomarão. O conhecimento desses intervalos é importante porque o objetivo é obter estimativas próximas dos verdadeiros valores dos parâmetros. Como *a* e *b* são variáveis aleatórias, podem ter covariância. Essa característica, junto com as médias e variâncias, que também são *pré-dados* de *a* e *b*, são

chamados *propriedades amostrais***,** porque a aleatoriedade dos estimadores decorre da extração de amostras de uma população.

As propriedades amostrais de um estimador de mínimos quadrados dizem como as estimativas variam de uma amostra para outra, dando uma base para avaliar a confiabilidade das estimativas. O estimador de mínimos quadrados é não tendencioso e não há outro estimador linear não tendencioso que apresente variância menor – se as hipóteses do modelo são corretas. Pindyck, Rubinfeld (2001)

3.3.3. Derivação dos Mínimos Quadrados

Para Pindyck e Rubinfeld (2001), a finalidade de construir uma relação estatística, em geral, é prever ou explicar os efeitos sobre uma variável que resultam de mudanças em uma ou mais variáveis explicativas (também conhecidas como antecipatórias ou explanatórias). Pode-se escrever a equação $Y = a + bX$, onde Y, a variável do lado esquerdo, é chamada de *variável dependente* e X, a variável do lado direito, é denominada *variável explicativa* (independente). Como a intenção é explicar ou prever movimentos em Y, é natural escolher como objetivo a minimização da soma vertical dos desvios ao quadrado da reta ajustada.

Para obter a fórmula de mínimos quadrados a fim de calcular os valores de *a* e de *b*, é preciso usar alguns instrumentos matemáticos básicos.

O critério dos mínimos quadrados pode ser representado formalmente como se segue:

$$Minimizar \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

Onde $\hat{Yi} = a + bX_i$ representa a equação para uma linha reta com intercepto *a* e inclinação *b*. Nessa notação Y_i é o valor efetivo de Y para a observação *i* e corresponde ao valor de X para aquela observação, enquanto N é o número de observações. $\hat{Yi}$, denominado valor ajustado ou previsto de Y_i, é o valor de Y na linha reta associada com a observação X_i. Para cada observação em X existe o

desvio correspondente entre o valor ajustado e o valor efetivo de Y. A soma dos quadrados desse desvio é o que se quer minimizar, pois permitirá calcular uma medida de grau de precisão com que a linha reta se ajusta aos dados.

O problema é escolher valores de *a* e de *b* que minimizem o resultado da equação. Isso pode ser feito usando cálculo elementar de álgebra. As soluções de mínimos quadrados para a inclinação e o intercepto são:

$$b = \frac{N\sum X_i Y_i - \sum X_i \sum Y_i}{N\sum X_i^2 - \left(\sum X_i\right)^2}$$

$$a = \frac{\sum Y_i}{N} - b\frac{\sum X_i}{N} = \overline{Y} - b\overline{X}$$

Onde $\overline{Y}$ e $\overline{X}$ são as médias de amostra de Y e X, respectivamente.

Considere agora como as fórmulas das duas equações anteriores simplificam no caso especial em que X e Y têm ambas médias de amostras iguais a 0. Primeiro, reescrevendo a equação, observa-se que

$$a = \overline{Y} - b\overline{X} = 0$$

Assim, quando as médias de amostras de X e Y são 0, o intercepto da reta de regressão ajustada será 0. Para obter a estimativa da inclinação correspondente nesse caso especial, dividiu-se tanto o numerador quanto o denominador da equação por N^2:

$$b = \frac{\sum \frac{X_i Y_i}{N} - \left(\sum \frac{X_i}{N}\right)\left(\sum \frac{Y_i}{N}\right)}{\sum \frac{X_i^2}{N} - \left(\sum \frac{X_i}{N}\right)^2}$$

Substituindo $\overline{Y}$ e $\overline{X}$ obtém-se

$$b = \frac{\sum \frac{X_i Y_i}{N} - \overline{XY}}{\sum \frac{X_i^2}{N} - \overline{X}^2}$$

Mas a pressuposição é que $\overline{X} = \overline{Y} = 0$. Por conseguinte,

$$b = \frac{\sum \frac{X_i Y_i}{N}}{\sum \frac{X_i^2}{N}} = \frac{\sum X_i Y_i}{\sum X_i^2}$$

O fato da nova equação ser menos complicada que a anterior sugere que a questão será mais simples e o entendimento melhor ao se escrever as estimativas de mínimos quadrados em termos de variáveis expressas como desvios de suas respectivas médias de amostra, sejam ou não essas médias iguais a zero. Para fazê-lo, transformam-se os dados para o formato de desvio, expressando cada observação em X e Y como desvios das respectivas médias.

$$x_i = X_i - \overline{X} \qquad\qquad y_i = Y_i - \overline{Y}$$

Com essa definição, a estimativa da inclinação de mínimos quadrados pode ser obtida (no caso geral) pois as variáveis *x* e *y* têm média zero. Na verdade, centralizam-se os dados ao mover a origem do gráfico que relaciona X e Y para a média da amostra. Nesse caso, as variáveis em caixa baixa são versões "centradas" das variáveis em caixa alta.

A estimativa da inclinação de mínimos quadrados torna-se:

$$b = \frac{\sum X_i Y_i}{\sum X_i^2}$$

3.3.4. O modelo de Regressão Múltipla

Apresenta-se a seguir o modelo de regressão com duas ou mais variáveis explicativas, isto é, o modelo de regressão múltipla. Descrevem-se os pressupostos subjacentes ao modelo clássico de regressão múltipla, e também como as estimativas dos parâmetros podem ser obtidas por intermédio do método dos mínimos quadrados.

Há possibilidade de surgir problemas por conta da iteração entre as variáveis explicativas da equação de regressão. Aqui, dá-se ênfase especial às várias estatísticas da regressão que ajudam na interpretação do modelo, inclusive coeficientes padronizados, elasticidades e coeficientes de correlação parcial.

Dada uma relação funcional definida com base na teoria econômica, assume-se que Y é uma variável dependente explicada por um conjunto de variáveis independentes X_1, X_2, ..., X_k e por um termo aleatório ε, que representa a soma de todos os demais fatores que afetam a variável dependente Y, além de X_1, X_2, ..., X_k, mas que não estão presentes no modelo. Pressupõe-se, também, que o modelo é inerentemente linear, significando que só se aplica a equações lineares, ou passíveis de linearização, Pindyck, Rubinfeld (2001).

Alguns pressupostos principais do modelo de regressão linear:

1. A relação entre Y e X é linear
2. Os X são variáveis não-estocásticas. Além disso, *não existe nenhuma relação linear exata entre duas ou mais variáveis explanatórias*.
3. O erro tem valor esperado (ou esperança matemática) zero para todas as observações.
4. O termo de erro tem variância constante para todas as observações.
5. Erros correspondentes a observações diferentes são independentes e portanto não há correlação entre eles.
6. O termo de erro tem distribuição normal.

Para simplificar, será utilizado um caso especial de regressão múltipla: o modelo de três variáveis

$$Y = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

O procedimento de mínimos quadrados equivale a buscar estimativas de parâmetros que minimizem a soma de quadrados dos desvios (ou dos resíduos) SQR, definida como

$$SQR = \sum \hat{\varepsilon}_i^2 = \sum (Y_i - \hat{Y})^2 \qquad \text{onde} \quad \hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$$

Podem-se encontrar os valores de β_1, β_2 e β_3 que minimizam SQR. Supondo que há mais de três observações e que as equações subjacentes são independentes, a solução é

$$\hat{\beta}_1 = \hat{Y} - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i}$$

Nesse modelo de três variáveis, o coeficiente β_2 mede a mudança de Y associada a uma mudança de uma unidade de X_2 supondo que a variável X_3 é mantida constante. Do mesmo modo, o coeficiente β_3 mede a mudança em Y associada com a mudança unitária em X_3 com X_2 mantido constante. Em ambos os casos a suposição de que os valores das demais variáveis explicativas são constantes é crucial para a interpretação dos coeficientes.