



Escola de Engenharia de Lorena

ECONOMETRIA

PROF. DR. DIOGO FERRAZ

1

Introdução



Era da Informação

- Uma firma necessita de informação sobre a melhor maneira de fabricar seus produtos, em face da limitação de recursos e da tecnologia corrente;
- Um consumidor necessita de informação para decidir o que comprará, diante de um orçamento limitado;
- Um governo necessita de informação sobre a sua sociedade e a comunidade global para melhor formular suas leis e regulamentos, sujeitos às restrições constitucionais.

Para que os agentes trabalhem com eficiência, consideradas as restrições sobre as quais operam, deve haver um *fluxo contínuo de informação*. Mas de onde provém essa informação?

Uma das fontes é a *Teoria Econômica* (Microeconomia, Macroeconomia, Desenvolvimento Econômico).

A Teoria Econômica nos proporciona uma base para identificarmos *variáveis econômicas importantes*, bem como a compreensão da alocação de recursos e da determinação de preços em uma economia de mercado.

Prof. Dr. Diogo Ferraz

2

Introdução



Contextualização:

Para que essas informações sejam úteis na tomada de decisões, precisamos *saber* mais sobre as *relações* entre *variáveis econômicas*, para responder questões do tipo:

- Qual é o efeito de uma aumento do déficit no orçamento federal sobre o nível da taxa de juros e sobre a inflação?
- Qual é o efeito das recentes fusões e aquisições de firmas sobre os ganhos dos acionistas?
- Qual a relação entre a quantidade de dinheiro (M1, por exemplo), e o nível de atividade econômica?

Se o Banco Central aumentar a taxa de desconto, isso causará uma estagnação com a inflação?

Se o governo estadual diminuir o número de alunos por sala de aula nas escolas da rede pública, haverá aumento do desempenho escolar?

Qual é o impacto, no nível de investimento, de uma modificação do imposto sobre ganhos de capital?

Prof. Dr. Diogo Ferraz

3

Introdução



Contextualização:

De onde provém informações como essas?

- Teoria Econômica
- Dados econômicos observados (preços, renda, consumo, oferta, aspectos sociais)

Desafio: Combinar teoria econômica com a informação fornecida pelos dados econômicos.

A *econometria* é baseada no desenvolvimento de *métodos estatísticos* para estimar *relações econômicas*, *testar teorias*, avaliar e implementar *políticas* de governo e de *negócios*.

Econometria pode fazer a previsão de importantes variáveis econômicas (juros, inflação, PIB, etc.).

Prof. Dr. Diogo Ferraz

4

Introdução



Exemplo 1.1: Para tratamento dos dados:

Disciplina Estatística: usamos Medidas de Tendência Central e Medidas de Dispersão para analisar um conjunto de dados X e outro conjunto Y → *Estatística Descritiva*.

Os dados da Tabela abaixo se referem à informações sobre a demanda de um produto (Y_i) e sobre o preço de venda (X_i):

X_i	Y_i
40	320
45	305
52	290
58	280
65	275
70	270
85	250
90	245
100	230
120	210

Prof. Dr. Diogo Ferraz

5

Econometria



O que é Econometria?

Econometria consiste na aplicação de *métodos matemáticos e estatísticos a problemas de economia*.

O econométrista combina conhecimentos de três ramos científicos:

1. Economia
2. Matemática e
3. Estatística.

A análise de regressão é o método mais importante da econometria.

Quais são os objetivos da Econometria?

A produção de *afirmações econômicas quantitativas* que permitem explicar o comportamento de variáveis que já observamos ou *prever comportamentos* ainda não observados, ou ambos. Em outras palavras, podemos conhecer os efeitos que algumas variáveis exercem, ou parecem exercer, sobre outras.

Prof. Dr. Diogo Ferraz

6

Econometria



Tipo de Dados:

Econometria utiliza **dados** econômicos **não experimentais**.

Estes dados não são acumulados por meio de **experimentos controlados** de indivíduos, firmas ou segmentos da economia.

Eles também são chamados **dados observacionais** ou **dados retrospectivos**, pois o econométrista é um coletor passivo destes dados.

Passos na Análise Econômica Empírica:

A análise empírica usa dados para testar uma teoria ou estimar uma relação.

1º passo: formulação cuidadosa da questão de interesse (**problema de pesquisa**).

Prof. Dr. Diogo Ferraz

7

Econometria



2º passo: construção de um **modelo econômico formal**, que consiste em equações matemáticas que descrevem várias relações.

Genericamente, tais relações funcionais podem ser representadas por:

$$Y = f(X_1, X_2, X_3, \dots, X_k)$$

Onde,

Y = variável dependente;

X_h (h=1, 2, 3,..., k) = representam as variáveis explanatórias (ou independentes).

Prof. Dr. Diogo Ferraz

8

Econometria



Alguns exemplos...

Se estamos lidando, em particular, com o **consumo (c)** podemos estabelecer a sua relação com a **renda (i)**:

$$c = f(i)$$

O que nos diz que o nível de consumo é uma função, f(i), da renda.

Sobre um **bem específico**, um automóvel Gol da Volkswagen, podemos escrever (q^g):

$$q^g = f(p, p^s, p^c, i)$$

O que nos diz que a quantidade demandada de Gols, q^g, é função do preço p dos Gols, do preço dos carros substitutos p^s, do preço de bens complementares (combustíveis) p^c, e do nível de renda i.

Prof. Dr. Diogo Ferraz

9

Econometria



Outros exemplos de relações funcionais entre as variáveis:

- Crescimento da **população** de um país (Y) em função dos **anos** (X);
- Variação da **produção** (Y) obtida numa cultura conforme a **quantidade de nitrogênio** (X1), **fósforo** (X2) e **potássio** (X3), utilizada na adubação;
- Variação do **preço** de um **produto** (Y) no mercado em função da **quantidade oferecida** (X).

Ou ainda, por meio da **Teoria do Capital Humano**, a relação funcional entre salário e as variáveis explanatórias:

$$\text{salário} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exp} + \beta_3 \text{sexo}$$

Onde:

salário é a variável dependente (y)

β₀ é o intercepto do modelo;

β₁ é o nível de escolaridade (anos médios de estudo);

β₂ é a experiência profissional;

β₃ é o sexo da pessoa;

Prof. Dr. Diogo Ferraz

10



Escola de Engenharia de Lorena

MODELO MATEMÁTICO E MODELO ESTATÍSTICO

PROF. DR. DIOGO FERRAZ

11

Modelo Matemático e Modelo Estatístico



Será que podemos considerar que todas as variáveis que afetam o salário estão contidas na equação?

$$\text{salário} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exp} + \beta_3 \text{sexo}$$

Ou seja, existem **variáveis não observadas**. Ex. salário: região onde a pessoa vive; aptidão inata; qualidade da educação, etc.

Na maior parte das vezes, a equação não possui todas as variáveis explicativas do modelo.



Prof. Dr. Diogo Ferraz

12

Modelo Matemático e Modelo Estatístico



3º passo: o economista precisa **transformar** o modelo econômico (matemático) em um **modelo econométrico**.

Consideremos duas variáveis, X e Y , relacionadas por uma função matemática $Y = f(X)$. Dado um conjunto de valores X_i ($i = 1, 2, \dots, n$) e os correspondentes valores de $Y_i = f(X_i)$, se colocarmos os pontos (X_i, Y_i) em um gráfico verificaremos que eles pertencem à curva que representa o **modelo matemático que relaciona as duas variáveis**.

É comum, entretanto, que a **variável dependente** seja **afetada** por **outros fatores**, além dos considerados no modelo adotado. Admitimos que a variável dependente sofra a influência de $k + m$ variáveis, isto é:

$$Y = f(X_1, X_2, X_3, \dots, X_k, X_{k+1}, X_{k+m})$$

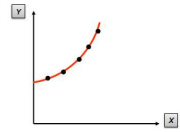


Figura 1.1 – Modelo Matemático $Y_i = f(X_i)$

Prof. Dr. Diego Ferraz

13

Modelo Matemático e Modelo Estatístico



Então, por vários motivos (**não disponibilidade** de valores, **impossibilidade de mensuração**, para simplificar a análise, etc.) **não consideramos a influência das variáveis**. Ao analisarmos Y como função das k primeiras variáveis **permanece**, então, **um resíduo ou erro**. Admitindo que esse erro seja aditivo, o modelo fica:

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i \quad \text{com } (i = 1, \dots, n)$$

Se apenas umas das variáveis independentes é considerada, temos:

$$Y_i = f(X_i) + u_i$$

Neste caso, o **conjunto de pares de valores** (X_i, Y_i) corresponde a um conjunto de pontos, **dispersos em torno da curva representativa da função**.

Dizemos que as duas variáveis estão relacionadas de acordo com um **modelo estatístico**.

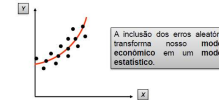


Figura 1.2 – Modelo Estatístico $Y_i = f(X_i) + u_i$

Prof. Dr. Diego Ferraz

14

Modelo Matemático e Modelo Estatístico



Outra justificativa para existência do erro (u_i) em um modelo estatístico é dada pelos **erros de mensuração da variável dependente**. Se os **verdadeiros valores** (V_i) da variável dependente são uma função matemática das variáveis explanatórias, isto é,

$$V_i = f(X_{1i}, X_{2i}, \dots, X_{ki})$$

e se os valores observados (Y_i) da variável dependente apresentam erros de mensuração (u_i), isto é,

$$Y_i = V_i + u_i$$

A relação entre Y_i e os X_{hi} ($h = 1, 2, \dots, k$) fica

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{ki}) + u_i$$

Prof. Dr. Diego Ferraz

15

Modelo Matemático e Modelo Estatístico



Em **casos reais** geralmente existem tanto **erros de mensuração como efeitos de outras variáveis**. Nestes casos, o **erro residual** do modelo será a **soma desses dois tipos de erro**.

Desde que existem erros de mensuração, é lógico admitir que os valores das variáveis explanatórias também são afetados.

Mas nesta etapa, **admitiremos apenas um erro residual devido à existência de fatores não incluídos no modelo e/ou erros de mensuração apenas na variável dependente**.

Veja que o erro aleatório (u) não só responde por diversos fatores que afetam as variáveis econômicas, como reflete a incerteza intrínseca na atividade econômica.

De fato, lidar com o termo de erro (ou termo de perturbação) é, talvez, o componente mais importante de qualquer análise econométrica.

Prof. Dr. Diego Ferraz

16

Modelo Matemático e Modelo Estatístico



Inferência Estatística:

Ao abordar o papel dos modelos econômicos e estatísticos, focalizamos as **relações entre variáveis econômicas observáveis**, como preços, consumo e investimento, e o desejo de **obter informações** sobre os correspondentes **parâmetros desconhecidos e não observáveis** como elasticidade e produtos marginais.

A **informação** sobre esses parâmetros econômicos desconhecidos ajuda-nos a entender as relações econômicas básicas e coloca-nos em posição de **discriminar entre hipóteses alternativas** sobre como o mundo funciona (teorias).

Ajuda-nos também a **formular implicações de política econômica** ou prever, e talvez controlar, os resultados de variáveis econômicas importantes.

A partir de uma amostra de dados coletados, cabe-nos delinear processos para utilizar tais informações para fins de estimação de parâmetros e inferência.

Prof. Dr. Diego Ferraz

17

Modelo Matemático e Modelo Estatístico



4º passo: Inferência Estatística.

Inferência significa a **utilização de informações contidas em uma amostra de dados para "inferir" algo sobre o mundo real**.

A teoria estatística permite-nos usar nossa informação amostral para **passarmos dos parâmetros teóricos** nos modelos econômicos e estatísticos **para seus correspondentes empíricos**, fazendo afirmações probabilísticas sobre os resultados possíveis.

Veja que trata-se de um conjunto completo de instrumentos: teoria econômica, modelo econômico, modelo estatístico, informação da amostra e teoria estatística.

Prof. Dr. Diego Ferraz

18



Escola de Engenharia de Lorena

TIPOS DE DADOS, CAUSALIDADE E CORRELAÇÃO

PROF. DR. DIOGO FERRAZ

19



Tipos de Dados

Corte Transversal (Cross-section):

Amostra de indivíduos: consumidores, empresas, cidades, estados, países ou uma variedade de outras unidades, tomada em um determinado ponto no tempo.

Às vezes, os dados de todas as unidades não correspondem precisamente ao mesmo período.

Não é possível considerar que os dados foram obtidos por amostragem aleatória da população.

Ex.: Altura de indivíduos selecionados aleatoriamente; salário e características de trabalhadores de um país.

Cross-section: i

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Conjunto de dados de corte transversal sobre salários e outras características individuais (educanos de estudo; experiancias de experiência; sexo=sexo do trabalhador, se homem (0) ou mulher (1)).

n. obs.	salário/h	educ	exper	sexo
1	3.10	11	2	1
2	3.24	12	22	1
3	3.00	9	2	0
4	6.00	12	16	1
5	5.30	12	7	0
6	5.45	13	8	0
...	...	...	...	...
525	11.56	14	10	0
526	3.50	6	5	1
527	15.20	14	30	0

Prof. Dr. Diogo Ferraz

20



Tipos de Dados

Séries temporais:

Consiste em observações sobre uma variável ou muitas variáveis ao longo do tempo (anuais, semanais, diários, etc.).

Ex.: Retornos diários do IBOVESPA, oferta de moeda, índice de preços, produto interno bruto, entre outros.

Neste tipo de dados, o tempo é uma variável importante, pois eventos passados podem influenciar eventos futuros. Desta forma, a ordenação cronológica das observações transmite informação importante.

Série temporal: t

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

Conjunto de dados de série temporal para salário mínimo (SM), taxa de cobertura média de trabalhadores que recebem o benefício (cobertura) e taxa de desemprego (desemp).

n. obs.	ano	SM	cobertura	desemp
1	1950	0.2	20.1	15.4
2	1951	0.21	20.7	16
3	1952	0.23	22.6	14.8
...	...	...	...	...
37	1986	3.35	58.1	18.9
38	1987	3.35	58.2	16.8
39	1988	3.37	58.4	16.7

Prof. Dr. Diogo Ferraz

21



Tipos de Dados

Dados em painel (dados longitudinais):

Consiste em uma série temporal para cada registro do corte transversal do conjunto de dados.

Eles têm como característica essencial que as mesmas unidades do corte transversal (indivíduos, cidades, empresas, etc.) são acompanhadas durante um determinado período.

Ex.: Trajetória de venda de uma mesma firma ao longo de um período de tempo. Número de homicídios em uma cidade ao longo dos anos.

Dado em painel: it

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

Conjunto de dados em painel sobre estatísticas de crime nas cidades para dois anos, sendo homic = número de homicídios, pop = população local em milhares, desemp = taxa de desemprego, policia é a quantidade de policiais na cidade para cada ano.

n. obs.	cidade	ano	homic	pop	desemp	policia
1	1	1986	5	350	8.70	441
2	1	1990	8	359	7.20	471
3	2	1986	2	64	5.40	75
4	2	1990	1	65	5.50	76
5	3	1986	6	20	6.00	30
6	3	1990	9	29	6.50	32
...	...	...	...	...	...	...
297	149	1986	20	300	10.00	89
298	149	1990	19	400	8.00	90
299	150	1986	26	500	16.00	115
300	150	1990	27	550	20.00	120

Prof. Dr. Diogo Ferraz

22



Causalidade e Correlação

O econometrista busca inferir se uma variável tem um efeito causal sobre outra variável.

Encontrar simplesmente uma associação entre duas ou mais variáveis pode ser sugestivo, mas, a não ser que se possa estabelecer uma causalidade, raramente é convincente.

Neste aspecto, a noção *ceteris paribus*, ou seja, quando outros fatores (relevantes) permanecem iguais, desempenha um papel fundamental na análise causal.

Exemplo 1: na análise de demanda do consumidor, estamos interessados em conhecer o efeito da variação do preço de um bem sobre a quantidade demandada, enquanto todos os outros fatores (renda, preços de outros bens e gostos individuais) permanecem constantes.

Se os demais fatores não permanecem constantes, não poderemos conhecer o efeito causal de uma variação do preço sobre a quantidade demandada.

Exemplo 2: Se uma pessoa é escolhida em uma população, e recebe um ano a mais de educação, em quanto terá o seu salário aumentado?

Essa é uma questão *ceteris paribus*, e implica que todos os outros fatores são mantidos fixos, enquanto a pessoa recebe um ano a mais de educação.

A questão fundamental é: foram mantidos fixos em número suficiente outros fatores, para que se possa inferir a causalidade?

Prof. Dr. Diogo Ferraz

23



Exercício

Refita...

Na equação de salários:

$$\text{salário} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exp} + \beta_3 \text{sexo}$$

Como transformar este modelo econômico em um modelo econométrico?

Quais variáveis estarão contidas no termo do erro (u_i)?

Imagine outras equações e defina as variáveis observáveis e não-observáveis do modelo.



Prof. Dr. Diogo Ferraz

24



Escola de Engenharia de Lorena

REGRESSÃO LINEAR SIMPLES

PROF. DR. DIOGO FERRAZ

25

Introdução

Objetivo:

Apresentar o método de Mínimos Quadrados Ordinários (MQO) para estimar parâmetros por meio da regressão linear simples.

Após este tópico, o aluno estará apto a:

- Apresentar conceitos sobre regressão linear simples;
- Apresentar o método de Mínimos Quadrados Ordinários (MQO);
- Estimar parâmetros;
- Analisar o modelo econométrico e os parâmetros;
- Verificar a qualidade do ajuste da regressão;
- Construir intervalos de confiança.

Prof. Dr. Diogo Ferraz

26

Introdução

Regressões lineares (*linhas retas*) entre *duas variáveis* quantitativas são fáceis de compreender e bastante comuns.

Vocês já viram que a *correlação* mede a *direção* e a *intensidade* dessa relação entre duas *variáveis*.

Uma reta de regressão também resume a *relação entre duas variáveis*, mas apenas em uma situação específica: *quando uma das variáveis ajuda a explicar ou a prever a outra*.

O termo regressão foi introduzido por Francis Galton (1822-1911). Ele estudou o fenômeno da *hereditariedade* e mostrou que, quando casais altos ou baixos têm filhos, as alturas destes tendem a regredir, ou reverter para uma altura média mais típica.

Os modelos de regressão são largamente utilizados em todas as áreas do conhecimento, tais como: computação, administração, engenharias, biologia, agronomia, economia, saúde, sociologia, etc.

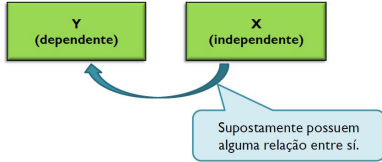
Prof. Dr. Diogo Ferraz

27

Introdução

Exemplos:

- Consumo e renda;
- Temperatura e pressão
- Vendas e gastos com MKT.



Prof. Dr. Diogo Ferraz

28

Introdução

APRENDER ESTATÍSTICA FÁCIL

EFEITO DA VAR. INDEPENDENTE SOBRE A VAR. DEPENDENTE

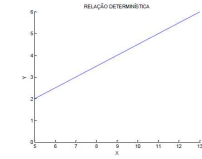
VARIÁVEL INDEPENDENTE	VARIÁVEL DEPENDENTE
Afeta a var. dependente (causa o efeito).	Afeta(a) pela var. independ (sofre o efeito).
Exemplo: - quantitativa, concentração do fertilizante; - qualitativa, tipo do fertilizante.	Exemplo: - quantitativa, tamanho da planta; - qualitativa, cor da flor.

Prof. Dr. Diogo Ferraz

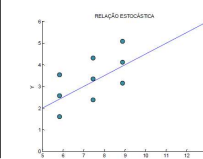
29

Introdução

Determinístico: Cada elemento do domínio (x) se relaciona com apenas um elemento da imagem (Y).

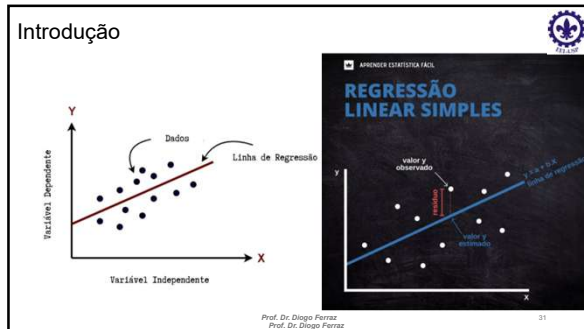


Estocástico: Para cada elemento do domínio (x) existe uma distribuição de probabilidade para os valores da imagem (Y).

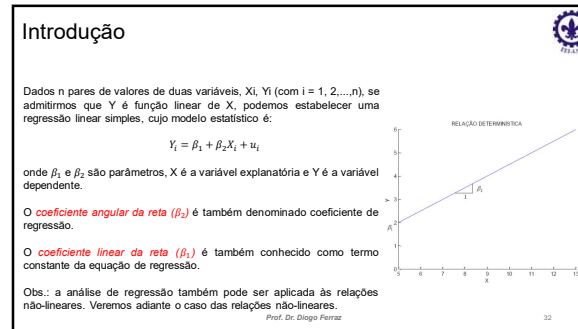


Prof. Dr. Diogo Ferraz

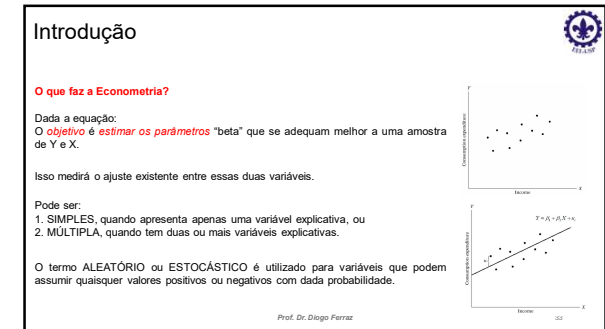
30



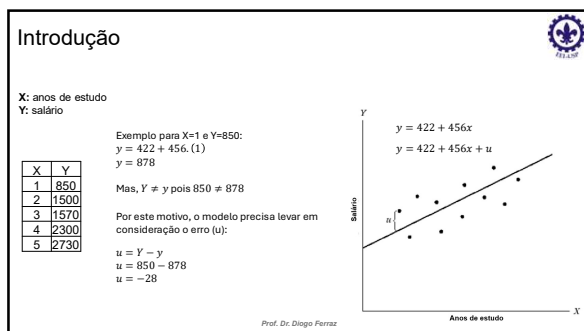
31



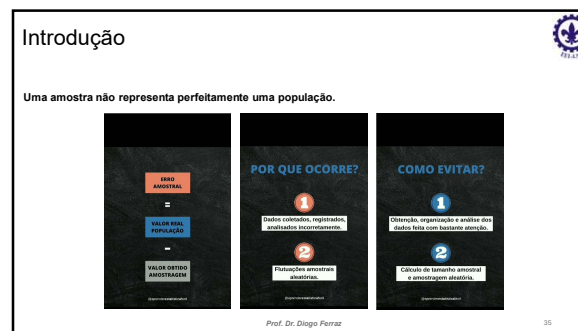
32



33



34



35



36

Modelo Populacional e Modelo Amostral



Do exemplo anterior, tínhamos um tamanho populacional igual a 5 (cinco observações).

Encontramos que:

$$\begin{aligned} \hat{Y} &= 422 + 456x_i \longrightarrow \hat{Y} = \beta_0 + \beta_1 x_i && \text{Função de regressão populacional} \\ y_i &= 422 + 456x_i + u_i \longrightarrow y_i = \beta_0 + \beta_1 x_i + u_i && \text{Modelo Populacional} \end{aligned}$$

Prof. Dr. Diogo Ferraz

37

37

Modelo Populacional e Modelo Amostral



Entretanto, podemos ter uma população igual a 500 indivíduos
Amostra = 5 observações

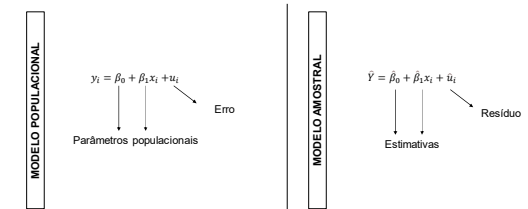
$$\begin{aligned} \hat{Y} &= 422 + 456x_i \longrightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i && \text{Função de regressão amostral} \\ \hat{Y} &= 422 + 456x_i + \hat{u}_i \longrightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i && \text{Modelo amostral} \end{aligned}$$

Prof. Dr. Diogo Ferraz

38

38

Modelo Populacional e Modelo Amostral



Prof. Dr. Diogo Ferraz

39

39

Modelo Populacional e Modelo Amostral



Pressuposições do Modelo:

- I) A relação entre X e Y é linear.
- II) Os valores de X são fixos, isto é, X não é uma variável aleatória.
- III) A média do erro é nula, isto é, $E(u_i) = 0$.
- IV) Para um dado valor de X, a variância do erro u é sempre σ^2 , denominada variância residual, isto é,

$$E(u^2) = \sigma^2 \quad \text{ou} \quad E[y_i - E(y_i | x_i)]^2 = \sigma^2$$

- Dizemos então, que o erro é **homocedástico** ou que temos homocedasticidade (do erro ou da variável dependente).
- V) O erro de uma observação é não correlacionado com o erro em outra observação, isto é, $E(u_i, u_j) = 0$ para $i \neq j$.
- VI) Os erros têm distribuição normal.

Combinando as pressuposições III, IV e VI, temos que $u_i \sim N(0, \sigma^2)$

Devemos, ainda, verificar se o número de observações disponíveis é maior do que o número de parâmetros da equação de regressão. Para ajustar uma regressão linear simples precisamos ter, no mínimo, 3 observações.

Prof. Dr. Diogo Ferraz

40

40



Escola de Engenharia de Lorena

ESTIMATIVA DOS PARÂMETROS

PROF. DR. DIOGO FERRAZ

41

Estimativa dos Parâmetros



O primeiro passo, na análise de regressão, é obter as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros da regressão. Os valores dessas estimativas serão obtidos a partir de uma amostra de n pares de valores X_i, Y_i (com $i = 1, 2, \dots, n$), que correspondem a n pontos num gráfico.

Obtemos, então

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Em que $\hat{\beta}_0, \hat{\beta}_1$ são, respectivamente, estimativas de $E(y_i) = \beta_0 + \beta_1 x_i$

Prof. Dr. Diogo Ferraz

42

42

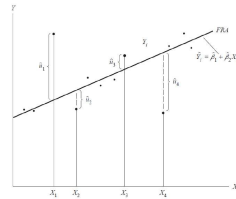
Estimativa dos Parâmetros

Para cada par de valores X_i, Y_i podemos estabelecer o desvio:

$$u_i = Y_i - \hat{Y}_i$$

$$u_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

O método dos Mínimos Quadrados Ordinários (MQO) consiste em adotar como estimativas dos parâmetros os valores que minimizam a soma dos quadrados dos desvios.



Prof. Dr. Diogo Ferraz

43

43

Estimativa dos Parâmetros

Temos o modelo econométrico: $y_i = \beta_0 + \beta_1 x_i + u_i$

Queremos minimizar a soma dos quadrados dos desvios $u_i = Y_i - \hat{Y}_i$

$$\text{Minimizar } u_i = Y_i - \hat{Y}_i$$

Primeiro, fazemos a diferença do valor real e do valor estimado

$$\text{Minimizar } \sum_{i=1}^n u_i = \sum_{i=1}^n (Y_i - \hat{Y}_i)$$

Em seguida, passamos o somatório, pois queremos minimizar todas as observações.

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

Finalmente, elevamos ao quadrado todos os termos.

Prof. Dr. Diogo Ferraz

44

44

Estimativa dos Parâmetros

A partir disso, podemos calcular a derivada parcial de cada parâmetro:

$$\frac{\partial \sum_{i=1}^n u_i^2}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial \sum_{i=1}^n u_i^2}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

Note que, apenas substituímos o $\hat{Y}_i$ pela equação

Estimativa para $\hat{\beta}_0$

$$\frac{\partial \sum_{i=1}^n u_i^2}{\partial \hat{\beta}_1} = 0$$

$$\frac{\partial \sum_{i=1}^n u_i^2}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i$$

Estimativa para $\hat{\beta}_1$

Prof. Dr. Diogo Ferraz

45

45

Estimativa dos Parâmetros

Simplificando, obtemos as equações denominadas Equações Normais de Mínimos Quadrados.

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n Y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i Y_i \end{cases}$$

A partir das estimativas de $\hat{\beta}_0, \hat{\beta}_1$ temos um sistema de equações que pode ser resolvido.

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n x_i}{n}$$

Primeiro podemos isolar o $\hat{\beta}_0$.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Assim, temos que $\hat{\beta}_0$ é a média de Y ($\bar{Y}$) menos $\hat{\beta}_1$ vezes a média de X ($\bar{x}$)

Prof. Dr. Diogo Ferraz

46

46

Estimativa dos Parâmetros

Desta forma, substituindo a equação anterior na segunda equação, temos que:

$$\hat{\beta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i Y_i - \hat{\beta}_0 \sum_{i=1}^n x_i$$

Retornando a segunda equação do sistema.

$$= \sum_{i=1}^n x_i Y_i - (\bar{Y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i$$

Substitua a definição anterior para $\hat{\beta}_0$.

$$= \sum_{i=1}^n x_i Y_i - \bar{Y} \sum_{i=1}^n x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i$$

Multiplique $\sum_{i=1}^n x_i$ por cada termo dentro dos parênteses.

$$= \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} + n\hat{\beta}_1 \bar{x}^2$$

Passar o por somatório.

$$\hat{\beta}_1 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}$$

Passar o $n\hat{\beta}_1 \bar{x}^2$ para o outro lado da equação e isole o $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Desta forma, temos o estimador para o $\hat{\beta}_1$.

Prof. Dr. Diogo Ferraz

47

47

Estimativa dos Parâmetros

Uma simplificação conveniente dos cálculos é obtida quando usamos a **variável centrada**.

Na representação gráfica, isso corresponde a tomar a média da variável X , como origem do eixo das abscissas. Nesse caso, o modelo estatístico fica:

$$Y_i = a + \beta x_i + u_i$$

Representando por "a" a estimativa de mínimos quadrados do parâmetro A , temos:

$$Y_i = a + \beta x_i + u_i$$

Prof. Dr. Diogo Ferraz

48

48

Estimativa dos Parâmetros



Em resumo, podemos representar as equações anteriores pelo modelo simplificado.

Centrando os valores na média, temos que $\sum z_i = 0$

Desta forma, temos que:

$$b = \frac{\sum_{i=1}^n xy}{\sum_{i=1}^n x^2} \quad a = \bar{Y} - b\bar{x}$$

Sendo que os valores de x , y e x^2 são valores centrados pela média previamente.

Calculamos primeiro o b e, posteriormente, encontramos o valor de a .

Prof. Dr. Diogo Ferraz

49

49

Estimativa dos Parâmetros



Uma indústria do setor metalmeccânico está analisando a relação entre o número de horas de funcionamento de uma máquina (X_i) e a quantidade de peças produzidas por turno (Y_i). Para isso, foi coletada uma amostra com 10 observações em diferentes dias da semana, conforme a tabela a seguir:

X_i	Y_i
0	3
1	2
1	3
2	5
3	4
3	4
4	7
5	6
5	7
6	9

Dicas resolução:

1º passo: Calcule o somatório de cada coluna (X_i e Y_i)

2º passo: Calcule a média de cada coluna (X_i e Y_i)

3º passo: Centre X_i e Y_i pela média (*subtraia cada valor (X_i ou Y_i) de sua média*);

4º passo: Com os valores centrados, calcule xy , x^2 e y^2

5º passo: Calcule o b .

6º passo: calcule o a .

7º passo: demonstre a equação estimada.

Prof. Dr. Diogo Ferraz

50

50



**EEL
USP**
Escola de Engenharia de Lorena

SOMA DE QUADRADOS

PROF. DR. DIOGO FERRAZ

51

Soma de Quadrados



A outra relação que nos interessa é a seguinte:

$$\sum y_i^2 = \sum e_i^2 + \sum \hat{y}_i^2$$

Soma de Quadrados Total = Soma de Quadrados Residual + Soma de Quadrados Da Regressão

Essa relação mostra que a variação dos valores de y em torno da sua média pode ser dividida em duas partes:

$\sum \hat{y}_i^2$ É a parte que é explicada pela regressão

$\sum e_i^2$ Esta é a parte não explicada pela regressão. Nem todos os pontos estão sobre a reta de regressão.

Prof. Dr. Diogo Ferraz

52

52

Soma de Quadrados



A outra relação que nos interessa é a seguinte:

Soma de Quadrados Total = Soma de Quadrados Residual + Soma de Quadrados Da Regressão

$$\begin{aligned} S.Q.Total &= \sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} \\ S.Q.Regr. &= b \sum xy \\ S.Q.Res. &= S.Q.Total - S.Q.Regr. \end{aligned}$$

Prof. Dr. Diogo Ferraz

53

53

Exemplo



Continuando com o Exemplo 1.0.

Use os dados da tabela abaixo para calcular a S.Q.Total, a S.Q.Resíduo e a S.Q.Regressão.

X_i	Y_i
0	3
1	2
1	3
2	5
3	4
3	4
4	7
5	6
5	7
6	9

Dicas resolução:

1º passo: Na tabela com valores centrados (Exemplo 1.0) insira uma coluna com o cálculo do y^2 ;

2º passo: Encontre a Soma de Quadrados Total;

3º passo: Com os valores da tabela (Exemplo 1.0), entre o S.Q. Regressão;

4º passo: Encontre o S.Q. Resíduo por meio da subtração entre S.Q. Total e S.Q. Regressão.

Prof. Dr. Diogo Ferraz

54

54



**EEL
USP**
Escola de Engenharia de Lorena

ANÁLISE DO MODELO

PROF. DR. DIOGO FERRAZ

55

Análise da Variância da Regressão

Na análise de variância da regressão, associamos às somas de quadrados total, de regressão e residual os graus de liberdade, respectivamente.

Os quadrados médios são obtidos dividindo as somas de quadrados pelos respectivos graus de liberdade.

$$Q.M. \text{ Regressão} = \frac{S.Q. \text{ Regressão}}{p - 1} \quad Q.M. \text{ Resíduo} = \frac{S.Q. \text{ Resíduo}}{n - p}$$

Se os erros têm distribuição normal, temos:

$$F = \frac{Q.M. \text{ Regr.}}{Q.M. \text{ Res.}}$$

Prof. Dr. Diogo Ferraz

56

Análise da Variância da Regressão

Outra forma de calcular é pela tabela abaixo.

Causas de variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios
Regressão	1	$b \sum xy$	$b \sum xy$
Resíduo	$n - 2$	$\sum y^2 - b \sum xy$	$(\sum y^2 - b \sum xy) / (n - 2)$
Total	$n - 1$	$\sum y^2$	

Este valor é o p-1 (na RLS sempre será igual a 1)

n é o número de observações da amostra

Passos:

1. Calcule os valores para os graus de liberdade;
2. Calcule a Soma de Quadrados da regressão e do resíduo;
3. Ache os quadrados médios dividindo a soma de quadrados pelo grau de liberdade correspondente.

Prof. Dr. Diogo Ferraz

57

Análise da Variância da Regressão

Utilizando o F calculado ($F_{calculado}$), podemos compará-lo com um F tabelado ($F_{tabelado}$), sendo que distribuição de F com 1 e $n-2$ graus de liberdade para a regressão linear simples.

A regra de rejeição é:

$$F_{calculado} > F_{tabelado}$$

Desta forma, podemos afirmar que o modelo é estatisticamente significativo ao nível de significância desejado.

Prof. Dr. Diogo Ferraz

58

Análise da Variância da Regressão

Utilizando a Tabela de Distribuição F (no exemplo ao lado, temos a tabela de 5% de significância).

Devemos encontrar:

$$F_{tabelado}(p - 1; n - p; \alpha\%)$$

Sendo que "p - 1" se refere à coluna da tabela (na regressão linear simples será sempre a primeira coluna);

E, "n - p" se refere ao número da linha na coluna selecionada.

Interpretação: Exemplo: 100 F(1; 100) com 5% de risco de rejeição (α = 0,05) = 161,254 (valor crítico de F)

Fonte: Christ (1966, p. 670) e Fincant Genes (1966, p. 406-407).

Prof. Dr. Diogo Ferraz

59

Exemplo

Continuando com o Exemplo 1.0.

Use os dados da tabela abaixo para calcular os quadrados médios da regressão.

X_i	Y_i
0	3
1	2
1	3
2	5
3	4
3	4
4	7
4	7
5	6
5	7
6	9

Dicas resolução:

1º passo: Utilize a soma de quadrados para encontrar os quadrados médios ou;

2º passo: utilize a tabela do slide anterior para encontrar os quadrados médios;

Prof. Dr. Diogo Ferraz

60

Exemplo



Tabela IV – Distribuição de F. Valor crítico F_{α} , tal que $P(F > F_{\alpha}) = 0,05$.
 Número de graus de liberdade do numerador

Gr. de liberdade do denominador	1	2	3	4	5	6	8	10	15	20	30	40	60	120	∞
1	161,4	191,6	215,7	230,2	240,0	247,6	253,7	258,9	263,9	268,9	273,8	278,6	283,3	287,9	292,4
2	18,51	18,00	17,59	17,28	16,98	16,70	16,44	16,19	15,95	15,72	15,50	15,29	15,09	14,90	14,72
3	10,13	9,78	9,49	9,25	9,02	8,80	8,58	8,36	8,14	7,93	7,72	7,52	7,32	7,13	6,95
4	6,59	6,35	6,12	5,90	5,68	5,46	5,24	5,02	4,80	4,59	4,38	4,17	3,96	3,76	3,57
5	5,09	4,86	4,63	4,41	4,20	3,98	3,76	3,54	3,32	3,11	2,90	2,69	2,48	2,28	2,09
6	4,21	4,00	3,77	3,55	3,34	3,12	2,90	2,68	2,46	2,25	2,04	1,83	1,62	1,42	1,23
7	3,59	3,39	3,16	2,94	2,72	2,50	2,28	2,06	1,84	1,63	1,42	1,21	1,00	0,80	0,61
8	3,10	2,91	2,68	2,46	2,24	2,02	1,80	1,58	1,36	1,15	0,94	0,73	0,52	0,31	0,12
9	2,75	2,57	2,34	2,12	1,90	1,68	1,46	1,24	1,02	0,81	0,60	0,39	0,18	0,07	0,00
10	2,48	2,31	2,08	1,86	1,64	1,42	1,20	0,98	0,76	0,55	0,34	0,13	0,02	0,00	0,00
11	2,28	2,12	1,89	1,67	1,45	1,23	1,01	0,79	0,57	0,36	0,15	0,04	0,00	0,00	0,00
12	2,12	1,96	1,73	1,51	1,29	1,07	0,85	0,63	0,41	0,20	0,09	0,00	0,00	0,00	0,00
13	2,00	1,84	1,61	1,39	1,17	0,95	0,73	0,51	0,29	0,08	0,00	0,00	0,00	0,00	0,00
14	1,90	1,74	1,51	1,29	1,07	0,85	0,63	0,41	0,20	0,08	0,00	0,00	0,00	0,00	0,00
15	1,82	1,66	1,43	1,21	0,99	0,77	0,55	0,33	0,11	0,00	0,00	0,00	0,00	0,00	0,00
16	1,75	1,59	1,36	1,14	0,92	0,70	0,48	0,26	0,04	0,00	0,00	0,00	0,00	0,00	0,00
17	1,69	1,53	1,30	1,08	0,86	0,64	0,42	0,20	0,00	0,00	0,00	0,00	0,00	0,00	0,00
18	1,64	1,48	1,25	1,03	0,81	0,59	0,37	0,15	0,00	0,00	0,00	0,00	0,00	0,00	0,00
19	1,60	1,44	1,21	0,99	0,77	0,55	0,33	0,11	0,00	0,00	0,00	0,00	0,00	0,00	0,00
20	1,56	1,40	1,17	0,95	0,73	0,51	0,29	0,07	0,00	0,00	0,00	0,00	0,00	0,00	0,00
25	1,45	1,29	1,06	0,84	0,62	0,40	0,18	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
30	1,37	1,21	0,98	0,76	0,54	0,32	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
40	1,28	1,12	0,89	0,67	0,45	0,23	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
60	1,21	1,05	0,82	0,60	0,38	0,16	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
120	1,15	0,99	0,76	0,54	0,32	0,10	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
∞	1,10	0,94	0,71	0,49	0,27	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Fonte: Christ (1966, p. 670) e FINECART Genex (1966, p. 406-407).
 Prof. Dr. Diogo Ferraz

61

Exemplo



Causas de variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios
Regressão	1	$b \sum xy$	$b \sum xy$
Resíduo	$n - 2$	$\sum y^2 - b \sum xy$	$(\sum y^2 - b \sum xy) / (n - 2)$
Total	$n - 1$	$\sum y^2$	

Prof. Dr. Diogo Ferraz

62


EEL
UST
 Escola de Engenharia de Lorena

COEFICIENTE DE DETERMINAÇÃO

PROF. DR. DIOGO FERRAZ

63

Coeficiente de Determinação (R^2)

Por meio da soma de quadrados, obtém-se o coeficiente de determinação, definido por:

$$r^2 = \frac{S.Q. \text{Regressão}}{S.Q. \text{Total}} = \frac{\sum y_i^2}{\sum y_i^2}$$

O coeficiente de determinação *indica a proporção da variação de y que é explicada pela regressão.*

Note que $0 \leq r^2 \leq 1$

Se estamos interessados em estimar os valores de Y a partir dos valores de X, a regressão será tanto mais útil quanto mais próximo de um estiver o valor de r^2 .

Prof. Dr. Diogo Ferraz

64

Coeficiente de Determinação (R^2)

Podemos verificar que:

$$r^2 = \frac{\sum y_i^2}{\sum y_i^2} = b^2 \frac{\sum x_i^2}{\sum x_i^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)} \quad \text{ou} \quad r^2 = \frac{S.Q. \text{Regressão}}{S.Q. \text{Total}}$$

Algumas fórmulas úteis para os exercícios:

$$\begin{aligned} S.Q. \text{Total} &= \sum y^2 = \sum y^2 - \frac{(\sum y)^2}{n} \\ S.Q. \text{Regr.} &= b \sum xy \\ S.Q. \text{Res.} &= S.Q. \text{Total} - S.Q. \text{Regr.} \end{aligned}$$

Prof. Dr. Diogo Ferraz

65

Coeficiente de Determinação (R^2)

R^2 é a razão entre a *variação explicada* e a *variação total*;

É interpretado como a fração da variação amostral em que y é explicado por x;

Quando interpretamos o R^2 , *usualmente multiplicamos por 100* para transformá-lo em *percentual*;

Se todos os pontos dos dados estiverem sobre a mesma reta, MQO fornece um ajuste perfeito aos dados ($R^2=1$ ou 100%);

$R^2=0$ indica um ajuste ruim da reta de MQO;

Entretanto, um R^2 *baixo* ainda pode apresentar *resultados interessantes* para a relação *ceteris paribus* (discutiremos isso em Regressão Linear Múltipla).

Prof. Dr. Diogo Ferraz

66

Exemplo

Continuando com o Exemplo 1.0 e posteriores...

Use os dados da tabela abaixo para calcular o R^2 .

X_i	Y_i
0	3
1	2
1	3
2	5
3	4
3	4
4	7
5	6
5	7
6	9

Dicas resolução (duas formas de resolução):

1º passo: utilize os valores da Soma de Quadrados para encontrar o r^2 .

Prof. Dr. Diogo Ferraz

67



Escola de Engenharia de Lorena

HIPÓTESES DO MODELO TEOREMA DE GAUSS-MARKOV

PROF. DR. DIOGO FERRAZ

68

Hipóteses do Modelo

Vamos analisar as propriedades de estimação do MQO.

Por meio do modelo populacional, estudaremos as propriedades estatísticas da estimação de MQO.

Primeiro, precisamos estabelecer a **inexistência de viés** do método MQO sob um conjunto simples de hipóteses da Regressão Linear Simples (RLS).

HIPÓTESE RLS 1. LINEAR NOS PARÂMETROS

No modelo populacional, a variável dependente y está relacionada à variável independente x e ao erro u como:

$$y = \beta_0 + \beta_1 x + u$$

Em que β_0 e β_1 são os parâmetros de intercepto e de inclinação populacionais, respectivamente.

Prof. Dr. Diogo Ferraz

69

69

Hipóteses do Modelo

Para especificar o modelo populacional, y , x e u são vistos como **variáveis aleatórias**.

HIPÓTESE RLS 2. AMOSTRAGEM ALEATÓRIA

Podemos usar uma amostra aleatória de tamanho n , $\{(x_i, y_i); i = 1, 2, \dots, n\}$ proveniente de um modelo populacional (Hipótese RLS 1).

Por isso, podemos escrever em termos de amostra aleatória:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad i = 1, 2, \dots, n$$

i pode ser: pessoa i , cidade i , país i , empresa i , etc.

Prof. Dr. Diogo Ferraz

70

70

Hipóteses do Modelo

Variação amostral em x_i como hipótese.

HIPÓTESE RLS 3. VARIAÇÃO AMOSTRAL NA VARIÁVEL EXPLICATIVA

Os resultados amostrais em x_i , ou seja, $\{x_i; i = 1, 2, \dots, n\}$ não são todos de mesmo valor.

Prof. Dr. Diogo Ferraz

71

71

Hipóteses do Modelo

A fim de obter estimadores não viesados, precisamos impor a hipótese de **média condicional zero**.

HIPÓTESE RLS 4. MÉDIA CONDICIONAL ZERO

O erro u tem um valor esperado de zero, dado qualquer valor de variável explicativa. Em outras palavras:

$$E(u|x) = 0$$

Para uma amostra aleatória, temos que: $E(u_i|x_i) = 0$

Prof. Dr. Diogo Ferraz

72

72

Hipóteses do Modelo



Em geral, a *inexistência de viés não é válida quando uma das quatro hipóteses é violada*.

É importante pensar na veracidade destas hipóteses antes de aplicar MQO em um caso particular.

Atenção deve ser dada para a Hipótese RLS 4. A *possibilidade de que x seja correlacionado com u* é quase sempre uma preocupação na análise de regressão linear simples.

Usar regressão linear simples quando u contém fatores que afetam y e que também estão correlacionados com x pode resultar em *correlação espúria*.

Correlação espúria significa achar uma relação entre y e x que se deve, na verdade, a outros fatores que afetam y e que também estão correlacionados com x.

Prof. Dr. Diogo Ferraz

73

73

Hipóteses do Modelo



É importante sabermos *quão distante*, em média, podemos esperar que $\hat{\beta}_1$ esteja distante *de* β_1 . Isso permite escolher o melhor estimado entre todos os estimadores não viesados.

A *medida de dispersão* da distribuição de $\hat{\beta}_1$ com a qual é mais fácil trabalhar é a *variância*, ou sua raiz quadrada, o *desvio padrão*.

A variância dos estimadores de MQO pode ser calculada sob as Hipóteses RLS 1 a RLS 4. Entretanto, vamos adicionar uma hipótese que é tradicional na análise de corte transversal.

HIPÓTESE RLS 5. HOMOSCEDASTICIDADE (variância constante)

O erro u tem a mesma variância, dado qualquer valor da variável explicativa. Em outras palavras:

$$\text{Var}(u|x) = \sigma^2$$

Prof. Dr. Diogo Ferraz

74

74

Hipóteses do Modelo



HIPÓTESE RLS 6. ERROS NORMALMENTE DISTRIBUÍDOS

O erro u tem distribuição normal. Em outras palavras:

$$u \sim N(0, \sigma^2)$$

Prof. Dr. Diogo Ferraz

75

75



Escola de Engenharia de Lorena

ANALISANDO OS PARÂMETROS DO MODELO

PROF. DR. DIOGO FERRAZ

76

Estimação da Variância do Erro



Para calcular a variância dos parâmetros, precisaremos *saber a variância do erro* σ^2 .

Entretanto, é muito raro sabermos esta medida, pois *desconhecemos as informações do erro*.

Uma *solução* é utilizarmos a *variância dos resíduos*, que chamaremos de s^2 .

Sendo assim, podemos definir o s^2 como:

$$s^2 = \frac{S.Q. \text{Resíduo}}{n - p}$$

Onde:

s^2 é a variância do resíduo da equação estimada;
n é o número de observações da amostra;
p é o número de parâmetros da regressão linear.

Prof. Dr. Diogo Ferraz

77

77

Estimação da Variância dos Parâmetros



Desta forma, podemos *calcular a variância dos parâmetros* do modelo, sendo:

$$V(b) = \frac{\sigma^2}{\sum x_i^2}$$

$$V(a) = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2} \right) \sigma^2$$

As respectivas estimativas são obtidas substituindo σ^2 por $s^2 = \text{Q.M.Res.}$, ou seja,

$$\hat{V}(b) = s^2(b) = \frac{s^2}{\sum x_i^2}$$

$$\hat{V}(a) = s^2(a) = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x_i^2} \right) s^2$$

As estimativas dos *desvios padrões* $s(a)$ e $s(b)$ são obtidas extraindo a *raiz quadrada* das respectivas estimativas de *variância*.

Prof. Dr. Diogo Ferraz

78

78

Exemplo

Utilizando os dados do Exemplo 1.0 (tabela abaixo), calcule a variância de a e b.

X_i	Y_i
0	3
1	2
1	3
2	5
3	4
3	4
4	7
5	6
5	7
6	9

Dicas resolução:

1º passo: calcule o s^2 ;

2º passo: utilize a fórmula da variância de b para o cálculo;

3º passo: utilize a fórmula da variância de a para o cálculo.

Prof. Dr. Diogo Ferraz

79

79

Exemplo

Utilizando os dados do Exemplo 1.0 (tabela abaixo), calcule a variância de a e b.

$$s^2 = \frac{S.Q. \text{ Resíduo}}{n - p}$$

$$V(b) = s^2(b) = \frac{s^2}{\sum x_i^2}$$

$$V(a) = s^2(a) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) s^2$$

Prof. Dr. Diogo Ferraz

80

80



TESTE DE HIPÓTESES

PROF. DR. DIOGO FERRAZ

81

Teste de Hipóteses

Um importante tópico para analisar a regressão é **testar a hipótese sobre um único parâmetro** da função de regressão.

Sendo assim, queremos testar hipóteses sobre um único β_j . Para isso, precisamos do seguinte resultado:

$$t(b) = \frac{b - \beta}{s(b)}$$

$$t(a) = \frac{a - \alpha}{s(a)}$$

$$t = \frac{(\text{estimativa} - \text{valor hipotético})}{\text{erro} - \text{padrão}}$$

tém distribuição de t com n-2 graus de liberdade.

Na maioria das aplicações, nosso principal interesse é **testar a hipótese nula**:

$$H_0: \beta_j = 0$$

Prof. Dr. Diogo Ferraz

82

82

Teste de Hipóteses

A **hipótese nula** $H_0: \beta_j = 0$ significa que o **parâmetro analisado não tem efeito sobre a variável dependente y**;

A estatística que usamos para testar a hipótese nula é a **estatística t**

$H_0: \beta_j = 0$ é útil porque fica fácil detectar $\beta_j \neq 0$

A questão que estamos analisando é quão distante o beta estimado está de zero? Um valor amostral de beta estimado muito distante de zero fornece evidência contra $H_0: \beta_j = 0$

A regra de **rejeição** exata depende da **hipótese alternativa** (H_1) e do **nível de significância** escolhido do teste.

Importante lembrar que estamos **testando hipóteses** sobre parâmetros **populacionais**. Nós não estamos testando hipóteses sobre estimativas de uma amostra particular.

Nunca fará sentido formular uma hipótese nula como $H_0: \hat{\beta}_1 = 0$ ou ainda $H_0: 0,237 = 0$

Nós estamos testando se o valor populacional desconhecido β_1 é zero.

Prof. Dr. Diogo Ferraz

83

83

Teste Unilateral

Precisamos decidir sobre a hipótese alternativa relevante.

HIPÓTESE ALTERNATIVA UNILATERAL:

$$H_1: \beta_j > 0$$

Neste teste, não nos preocupamos com alternativas de H_0 do tipo $H_1: \beta_j < 0$, por alguma razão, por exemplo, tomando como base a introspeção ou a teoria econômica.

Como devemos escolher sobre uma regra de rejeição?

Primeiro, devemos decidir sobre um **nível de significância** ou probabilidade de rejeitar H_0 quando ela é, de fato, verdadeira.

Um nível de significância usual é o de 5%, ou seja, estamos dispostos a rejeitar erroneamente H_0 , quando ela é verdadeira 5% das vezes.

Prof. Dr. Diogo Ferraz

84

84

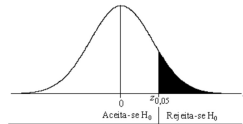
Teste Unilateral

A regra de rejeição é que H_0 é rejeitada em favor de H_1 no nível de significância de 5% se:

$$t_{\text{calculado}} > t_{\text{tabelado}}$$

De acordo com a escolha do valor crítico t_{tabelado} , a rejeição de H_0 , quando ela é verdadeira, ocorrerá em 5% de todas as amostras aleatórias.

Esta regra de rejeição é um exemplo de **teste monocaudal**. Para obtermos o t_{tabelado} precisamos do nível de significância e dos graus de liberdade.



Prof. Dr. Diogo Ferraz

85

85

Teste Unilateral

Por exemplo, para um teste no nível de significância de 5% e com 28 graus de liberdade, o valor crítico será:

1º passo: encontrar o nível de significância para o teste unilateral

2º passo: encontrar a linha com o grau de liberdade (gl) definido

3º passo: encontrar o valor correspondente à linha e à coluna definidos nos passos anteriores.

$$t_{\text{tabelado}} = 1,7011$$

Encontre as tabelas de distribuição no Moodle.

Prof. Dr. Diogo Ferraz

Distribuição t de Student

T	Teste Unilateral									
	15%	10%	5%	2.5%	1%	0.5%	0.1%	0.05%	0.01%	
1	1.753	1.777	1.812	1.895	1.943	1.983	2.059	2.145	2.282	
2	1.699	1.732	1.766	1.848	1.896	1.936	2.012	2.098	2.234	
3	1.645	1.678	1.712	1.794	1.842	1.882	1.958	2.044	2.180	
4	1.599	1.632	1.666	1.748	1.796	1.836	1.912	1.998	2.134	
5	1.559	1.592	1.626	1.708	1.756	1.796	1.872	1.958	2.094	
6	1.519	1.552	1.586	1.668	1.716	1.756	1.832	1.918	2.054	
7	1.479	1.512	1.546	1.628	1.676	1.716	1.792	1.878	2.014	
8	1.439	1.472	1.506	1.588	1.636	1.676	1.752	1.838	1.974	
9	1.399	1.432	1.466	1.548	1.596	1.636	1.712	1.798	1.934	
10	1.359	1.392	1.426	1.508	1.556	1.596	1.672	1.758	1.894	
11	1.319	1.352	1.386	1.468	1.516	1.556	1.632	1.718	1.854	
12	1.279	1.312	1.346	1.428	1.476	1.516	1.592	1.678	1.814	
13	1.239	1.272	1.306	1.388	1.436	1.476	1.552	1.638	1.774	
14	1.199	1.232	1.266	1.348	1.396	1.436	1.512	1.598	1.734	
15	1.159	1.192	1.226	1.308	1.356	1.396	1.472	1.558	1.694	
16	1.119	1.152	1.186	1.268	1.316	1.356	1.432	1.518	1.654	
17	1.079	1.112	1.146	1.228	1.276	1.316	1.392	1.478	1.614	
18	1.039	1.072	1.106	1.188	1.236	1.276	1.352	1.438	1.574	
19	0.999	1.032	1.066	1.148	1.196	1.236	1.312	1.398	1.534	
20	0.959	0.992	1.026	1.108	1.156	1.196	1.272	1.358	1.494	
21	0.919	0.952	0.986	1.068	1.116	1.156	1.232	1.318	1.454	
22	0.879	0.912	0.946	1.028	1.076	1.116	1.192	1.278	1.414	
23	0.839	0.872	0.906	0.988	1.036	1.076	1.152	1.238	1.374	
24	0.799	0.832	0.866	0.948	0.996	1.036	1.112	1.198	1.334	
25	0.759	0.792	0.826	0.908	0.956	0.996	1.072	1.158	1.294	
26	0.719	0.752	0.786	0.868	0.916	0.956	1.032	1.118	1.254	
27	0.679	0.712	0.746	0.828	0.876	0.916	0.992	1.078	1.214	
28	0.639	0.672	0.706	0.788	0.836	0.876	0.952	1.038	1.174	
29	0.599	0.632	0.666	0.748	0.796	0.836	0.912	0.998	1.134	
30	0.559	0.592	0.626	0.708	0.756	0.796	0.872	0.958	1.094	
40	0.449	0.482	0.516	0.598	0.646	0.686	0.762	0.848	0.984	
50	0.379	0.412	0.446	0.528	0.576	0.616	0.692	0.778	0.914	
60	0.319	0.352	0.386	0.468	0.516	0.556	0.632	0.718	0.854	
70	0.279	0.312	0.346	0.428	0.476	0.516	0.592	0.678	0.814	
80	0.239	0.272	0.306	0.388	0.436	0.476	0.552	0.638	0.774	
90	0.209	0.242	0.276	0.358	0.406	0.446	0.522	0.608	0.744	
100	0.189	0.222	0.256	0.338	0.386	0.426	0.502	0.588	0.724	
120	0.159	0.192	0.226	0.308	0.356	0.396	0.472	0.558	0.694	
140	0.139	0.172	0.206	0.288	0.336	0.376	0.452	0.538	0.674	
160	0.119	0.152	0.186	0.268	0.316	0.356	0.432	0.518	0.654	
180	0.109	0.142	0.176	0.258	0.306	0.346	0.422	0.508	0.644	
200	0.099	0.132	0.166	0.248	0.296	0.336	0.412	0.498	0.634	

Teste Unilateral

Quando o **nível de significância cai**, o **valor crítico** (t_{tabelado}) **aumenta**, de modo que, para rejeitar H_0 , exigimos um valor cada vez **maior do** $t_{\text{calculado}}$.

Sendo assim, se H_0 é rejeitada ao nível de significância de 5%, ela também será rejeitada ao nível de significância de 10%.

Por este motivo, não faz sentido rejeitar a hipótese nula ao nível de 5% e, em seguida, refazer o teste para testá-la ao nível de 10%.

A medida que os graus de liberdade da distribuição t ficam maiores, a **distribuição t se aproxima da distribuição normal padronizada**. Por este motivo, com **graus de liberdade acima de 120**, o valor crítico de 5% para a hipótese alternativa unilateral é $t_{\text{tabelado}} = 1,658$.

Para graus de liberdade acima de 120, podemos usar os valores críticos da distribuição normal padronizada $t_{\text{tabelado}} = 1,645$.

Prof. Dr. Diogo Ferraz

87

87

Teste Bilateral

É comum testar a hipótese nula $H_0: \beta_j = 0$ contra uma **hipótese alternativa bilateral**, ou seja:

$$H_1: \beta_j \neq 0$$

Esta hipótese alternativa diz que x_j tem um efeito *ceteris paribus* sobre y **sem especificar se o efeito é positivo ou negativo**.

Esta é uma hipótese relevante quando o sinal de β_j não é bem determinado pela teoria. Mesmo quando sabemos o sinal, o teste bilateral é a forma mais prudente.

Não é permitido utilizar as estimativas da regressão para ajudar a formular as hipóteses nula e alternativa, pois a inferência estatística clássica pressupõe que **formulamos as hipóteses nula e alternativa sobre a população antes de olhar os dados**.

Prof. Dr. Diogo Ferraz

88

88

Teste Bilateral

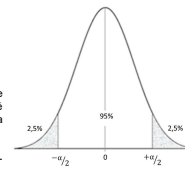
Quando a estatística é bilateral, estamos interessados no **valor absoluto** (|. |) da estatística t.

A regra de rejeição para $H_0: \beta_j = 0$ contra $H_1: \beta_j \neq 0$ será:

$$|t_{\text{calculado}}| > t_{\text{tabelado}}$$

Para encontrar o t_{tabelado} , devemos especificar novamente um nível de significância (por exemplo, 5%). Para um **teste bicaudal**, o t_{tabelado} é escolhido de tal forma que faça com que a área em cada cauda da distribuição t seja igual a 2,5%.

Em outras palavras, t_{tabelado} é o 97,5% percentil da distribuição t com n-1 graus de liberdade.



Prof. Dr. Diogo Ferraz

89

89

Teste de Hipóteses

Quando uma hipótese alternativa específica não é formulada, considera-se geralmente que ela é bilateral:

Nos livros de Econometria, o padrão é uma hipótese alternativa bilateral, e 5% será o nível de significância padrão.

Dizemos que:

Se H_0 é rejeitada em favor de $H_1: \beta_j \neq 0$ no nível de 5%, dizemos que x_j é estatisticamente significante, ou estatisticamente diferente de zero, no nível de 5%.

Se H_0 não é rejeitada contra $H_1: \beta_j \neq 0$ no nível de 5%, dizemos que x_j é estatisticamente não significante no nível de 5%.

90

90

Teste de Hipótese

Podemos testar se β_j é igual a alguma outra constante dada.

Exemplos comuns são: $\beta_j = 1$ e $\beta_j = -1$

A Hipótese nula pode ser expressa como $H_0: \beta_j = a_j$

Em que a_j é o nosso valor hipotético de β_j , então a estatística t apropriada será:

$$t = \frac{(\text{estimativa} - \text{valor hipotético})}{\text{erro} - \text{padrão}}$$

A definição de rejeição permanece a mesma dos casos anteriores.



Prof. Dr. Diogo Ferraz
Prof. Dr. Diogo Ferraz

91

91

Exemplo

Utilizando os dados do exemplo anterior, calcule o teste t de b, com o nível de significância de 5% em um teste bicaudal $H_1: \beta_j \neq 0$

X_i	Y_i
0	3
1	2
1	3
2	5
3	4
3	4
4	7
5	6
5	7
6	9

Dicas resolução:

1º passo: por meio da variância de b, calcule o desvio-padrão de b;

2º passo: aplique a fórmula do teste t;

3º passo: encontre o valor tabelado.

4º passo: compare o valor calculado com o valor tabelado e defina se é significativo ou não.



Prof. Dr. Diogo Ferraz

92

92



Escola de Engenharia de Lorena

INTERVALO DE CONFIANÇA

PROF. DR. DIOGO FERRAZ



Prof. Dr. Diogo Ferraz

93

93

Intervalo de Confiança

Podemos construir um Intervalo de Confiança (IC) para o parâmetro populacional β_j

Usando o fato de que $(\beta_j - \beta_j) / \text{er}(\beta_j)$ tem uma distribuição t com graus de liberdade, uma simples manipulação algébrica leva a um IC do β_j desconhecido.

Um intervalo de confiança de 95% é dado por:

$$b - t_{\alpha/2}(b) < \beta < b + t_{\alpha/2}(b)$$
$$a - t_{\alpha/2}(a) < \alpha < a + t_{\alpha/2}(a)$$

Sendo b ou a o parâmetro estimado;
 t_{α} é o t_{tabelado} ou seja, o valor de t encontrado na tabela de distribuição
 $s(a)$ ou $s(b)$ é o erro padrão de cada parâmetro.



Prof. Dr. Diogo Ferraz

94

94

Intervalo de Confiança

Quando os graus de liberdade são maiores do que 120, a distribuição t_{n-k-1} está suficientemente próxima da normal para usar o 97,5º percentil de uma distribuição normal padrão para construir um IC de 95%:

$$\beta_j \pm 1,96.s(\beta_j)$$

Quando os graus de liberdade são maiores do que 50, o valor do $t_{\text{calculado}}$ está próximo de 2, de modo que podemos usar como regra geral para intervalos de confiança de 95%:

$$\beta_j \pm 2,0.s(\beta_j)$$

Para graus de liberdade pequenos, os percentis exatos devem ser obtidos das tabelas t.



Prof. Dr. Diogo Ferraz

95

95

Exemplo

Utilizando os dados dos exemplos anteriores, calcule o intervalo de confiança de 95%.

X_i	Y_i
0	3
1	2
1	3
2	5
3	4
3	4
4	7
5	6
5	7
6	9

$$b - t_{\alpha/2}(b) < \beta < b + t_{\alpha/2}(b)$$



Prof. Dr. Diogo Ferraz

96

96



Escola de Engenharia de Lorena

PROBLEMA DE ESPECIFICAÇÃO

PROF. DR. DIOGO FERRAZ

97

Problema de Especificação

Quando aplicamos análise de regressão ao estudo da relação funcional entre duas variáveis, o problema de especificação consiste em **determinar a forma matemática na função que será ajustada**. Podemos escolher, por exemplo:

- $Y_i = \alpha + \beta X_i$
- $Y_i = \alpha \beta^{X_i}$
- $Y_i = \alpha X_i^\beta$
- $Y_i = \alpha + \frac{\beta}{X_i}$
- $Y_i = \alpha + \beta X_i + \gamma X_i^2$
- $Y_i = \alpha + \beta \rho^{X_i}$

A determinação da forma matemática da função pode ser feita de **duas formas diferentes**, muitas vezes complementares:

- Utilizando o **conhecimento** que temos **a priori** sobre o fenômeno.
- empregando o **conhecimento adquirido pela inspeção dos dados** numéricos disponíveis. É muito útil fazer um gráfico com os pontos (X_i, Y_i) e, eventualmente, gráficos com os pontos $(\ln X_i, Y_i)$, $(X_i, \ln Y_i)$ ou $(\ln X_i, \ln Y_i)$.

Frequentemente, **ajustamos mais de um modelo e escolhemos**, com base nos resultados estatísticos obtidos – coeficiente de determinação, quadrados médios residuais, etc. – o modelo que melhor se ajusta aos dados.

Prof. Dr. Diogo Ferraz

98

Problema de Especificação

Modelo I:

Admitindo um erro aditivo o modelo I fica:

$$Y_i = \alpha + \beta X_i + u_i$$

Esse é o modelo estatístico estudado até aqui.

Os modelos II, III e IV são exemplos de modelos não lineares que se transformam em funções lineares por anamorfose (substituição).

Prof. Dr. Diogo Ferraz

99

Problema de Especificação

Modelo II (Função Exponencial):

Admitindo um erro multiplicativo ε_i , obtemos o modelo estatístico:

$$Y_i = \alpha \beta^{X_i} \varepsilon_i$$

Aplicando logaritmos, obtemos

$$\log Y_i = \log \alpha + X_i \log \beta + \log \varepsilon_i$$

Fazendo:

$$\begin{aligned} \log Y_i &= Z_i \\ \log \alpha &= A \\ \log \beta &= B \\ \log \varepsilon_i &= u_i \end{aligned}$$

Propriedades dos Logaritmos

Log de um produto $\ln(X_1 \cdot X_2) = \ln X_1 + \ln X_2$
Log de uma potência $\ln X^a = a \ln X$

Este é o modelo estatístico de uma RLS de $Z_i = \log Y_i$ em relação a X_i . Se o erro $u_i = \log \varepsilon_i$ obedece às pressuposições dadas no item 2.1, podemos aplicar à amostra de pares de valores de X_i e Z_i os métodos de análise e regressão já estudados.

Prof. Dr. Diogo Ferraz

100

Problema de Especificação

Modelo III (Função Potência – Função Cobb-Douglas)

$$Y_i = \alpha X_i^\beta \varepsilon_i$$

Aplicando logaritmos, obtemos

$$\log Y_i = \log \alpha + \beta \log X_i + \log \varepsilon_i$$

Ou

$$Z_i = A + \beta V_i + u_i$$

Onde

$$\begin{aligned} Z_i &= \log Y_i \\ A &= \log \alpha \\ V_i &= \log X_i \\ u_i &= \log \varepsilon_i \end{aligned}$$

Essa função potência corresponde a um modelo de RLS nos logaritmos das duas variáveis.

Prof. Dr. Diogo Ferraz

101

Problema de Especificação

Modelo IV (Hipérbole)

Admitindo um erro aditivo o modelo IV fica:

$$Y_i = \alpha + \frac{\beta}{X_i} + u_i$$

Basta fazer a anamorfose

$$V_i = \frac{1}{X_i}$$

para obtermos o modelo de uma RLS:

$$Y_i = \alpha + \beta V_i + u_i$$

Modelos V e VI

O modelo V e VI são analisados como uma regressão múltipla com duas variáveis explanatórias, tema que será abordado no próximo tópico.

Prof. Dr. Diogo Ferraz

102

Exercício



Uma equipe de engenharia de produção está avaliando o impacto do número de horas de treinamento técnico (X) oferecidas a operadores de máquinas sobre o número médio de peças corretamente produzidas por turno (Y). A tabela a seguir apresenta os dados coletados em cinco unidades da fábrica:

É dada uma amostra de 5 pares de valores:

Admite-se que as variáveis X e Y estão relacionadas de acordo com o modelo $Y_i = \alpha + \beta X_i + u_i$ onde os u_i são variáveis aleatórias independentes com distribuição normal de média zero e variância σ^2 .

X	Y
1	3,0
2	7,5
3	7,0
4	11,5
5	11,0

- a) Determine as estimativas dos parâmetros da regressão linear.
- b) Calcule o coeficiente de determinação e faça análise de variância da regressão, considerando o nível de significância de 5%.
- c) Teste, ao nível de significância de 5%, a hipótese $H_0: \beta = -2$ contra a hipótese alternativa $H_1: \beta \neq -2$.
- d) Teste, ao nível de significância de 5%, a hipótese $H_0: \alpha = 13$ contra a hipótese alternativa $H_1: \alpha < 13$.
- e) Determine a estimativa de Y para X = 5.

Referências



Bibliografia Básica:

WOOLDRIDGE, J. Introdução à Econometria: Uma abordagem Moderna. 4 ed. Cengage Learning, 2010. (Capítulo 2)

HOFFMANN, Rodolfo; VIEIRA, Sônia. Análise de regressão: uma introdução à econometria. São Paulo, p. 32, 2016.