

Análise de Regressão Múltipla: Mínimos Quadrados Ordinários

**Ernesto F. L. Amaral
Magna M. Inácio**

26 de agosto de 2010

**Tópicos Especiais em Teoria e Análise Política:
Problema de Desenho e Análise Empírica (DCP 859B4)**

Fonte:

**Wooldridge, Jeffrey M. “Introdução à econometria: uma abordagem moderna”. São Paulo:
Cengage Learning, 2008.**

CAPÍTULO 2

INTRODUÇÃO

MODELO DE REGRESSÃO LINEAR SIMPLES

- Também chamado de modelo de regressão linear de duas variáveis ou modelo de regressão linear bivariada.

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

- Terminologia:

| y | x | Uso |
|----------------------|-----------------------|------------------------|
| Variável Dependente | Variável Independente | Econometria |
| Variável Explicada | Variável Explicativa | |
| Variável de Resposta | Variável de Controle | Ciências Experimentais |
| Variável Prevista | Variável Previsora | |
| Regressando | Regressor | |
| | Covariável | |

HIPÓTESE SOBRE A RELAÇÃO ENTRE x E u

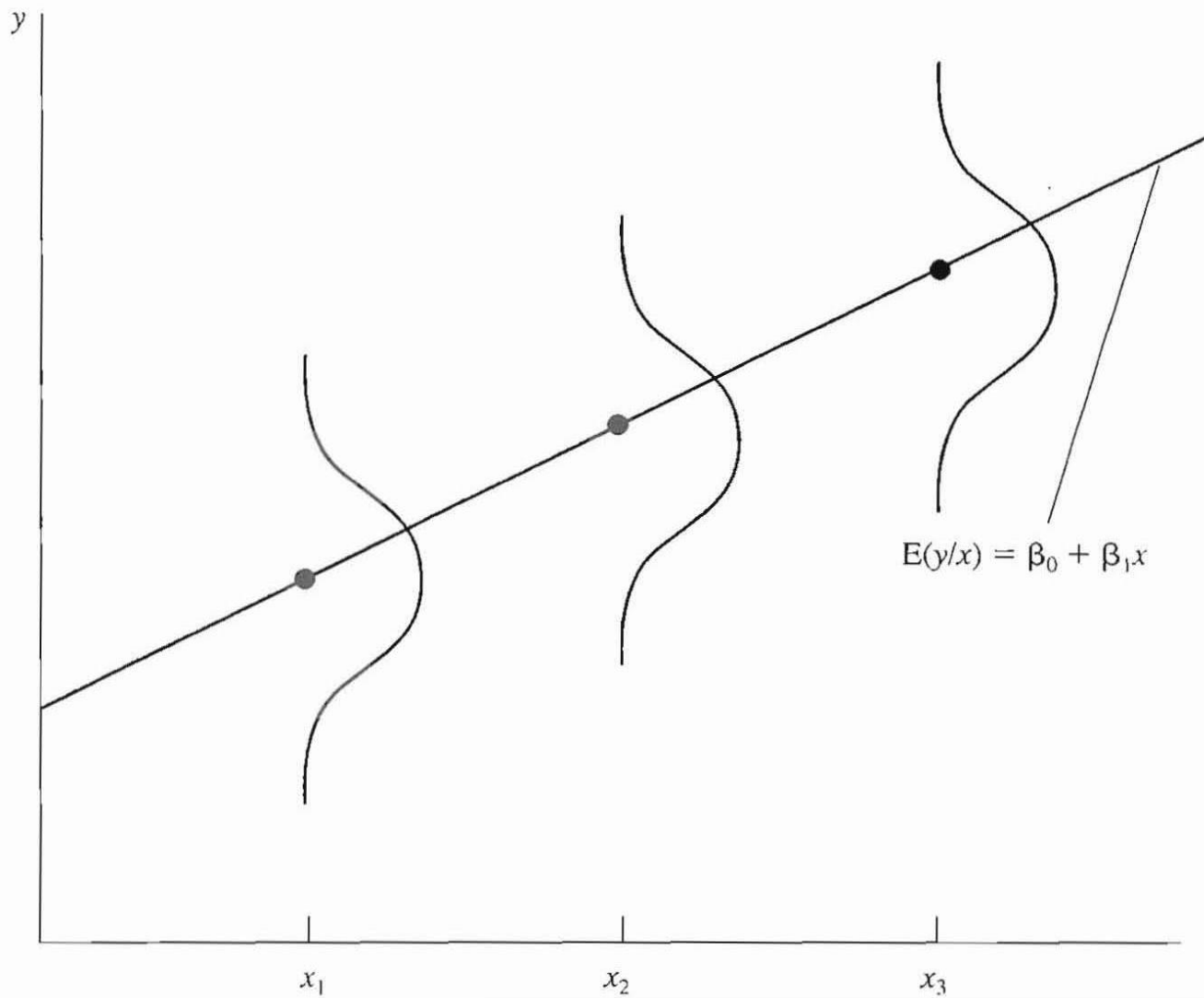
- Se u e x não estão correlacionados, então (como variáveis aleatórias) não são linearmente relacionados.
- No entanto, a correlação mede somente a dependência linear entre u e x .
- Na correlação, é possível que u seja não-correlacionado com x e seja correlacionado com funções de x , tal como x^2 .
- Melhor seria pensar na distribuição condicional de u , dado qualquer valor de x .
- Para um valor de x , podemos obter o valor esperado (ou médio) de u para um grupo da população.
- A hipótese é que o valor médio de u não depende de x :

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

- Ou seja, para qualquer valor de x , a média dos fatores não-observáveis é a mesma e, portanto, é igual ao valor médio de u na população (**hipótese de média condicional zero**).

Figura 2.1

$E(y|x)$ como função linear de x .



ESTIMATIVA DE MÍNIMOS QUADRADOS ORDINÁRIOS

- Para a estimação dos parâmetros β_0 e β_1 , é preciso considerar uma amostra da população:

$$\{(x_i, y_i): i=1, \dots, n\}$$

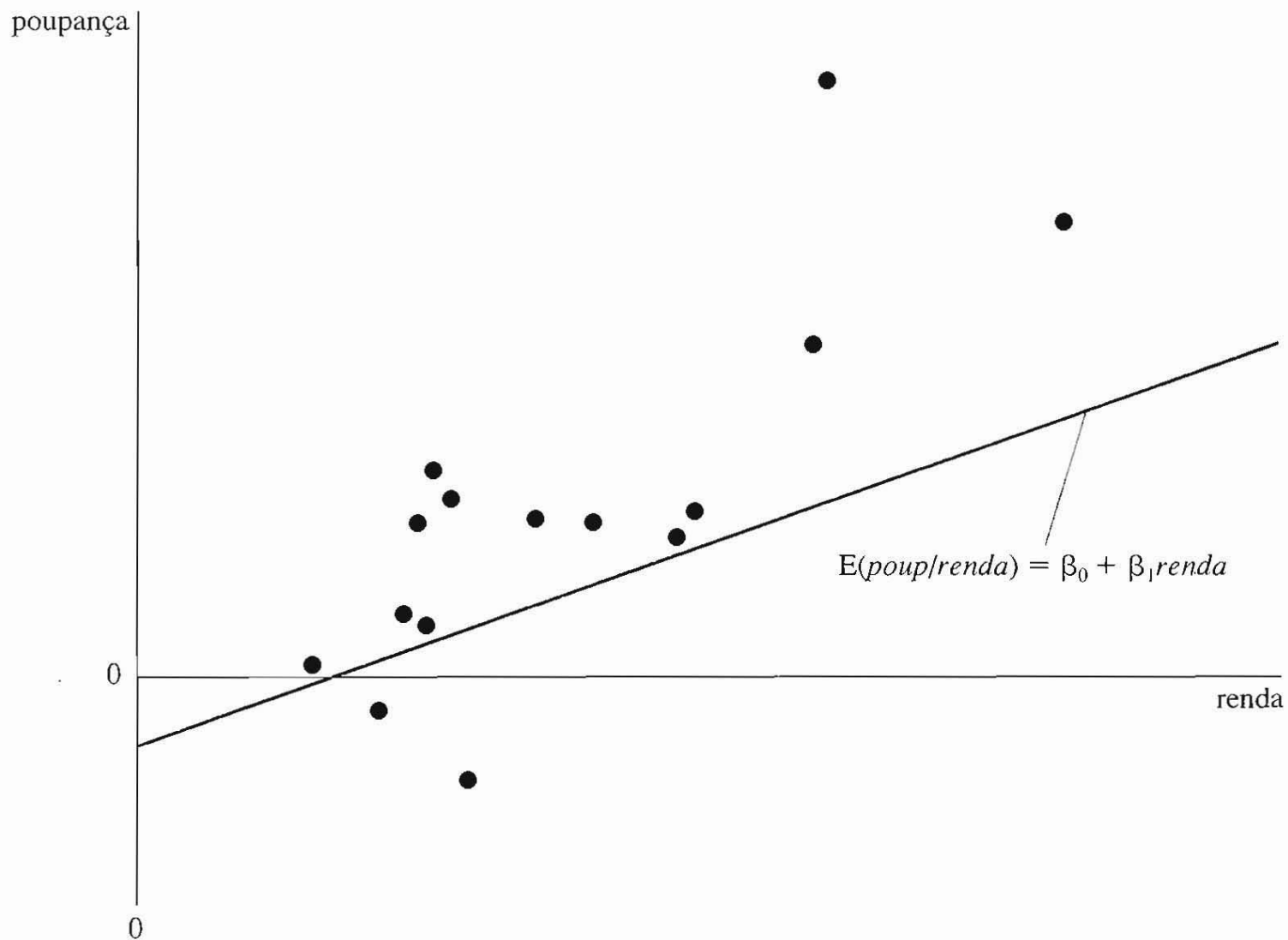
- A equação do modelo de regressão simples é escrito como:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- u_i é o termo erro para a observação i , já que contém todos os fatores, além de x_i , que afetam y_i .
- Um exemplo é a poupança anual para a família i (y_i), dependendo da renda anual desta família (x_i), em um determinado ano.


Figura 2.2

Gráfico da dispersão de poupança e renda de 15 famílias e a regressão populacional $E(\text{poup}|\text{renda}) = \beta_0 + \beta_1 \text{renda}$.



ESTIMATIVAS DE MQO DE $\hat{\beta}_0$ E $\hat{\beta}_1$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$


$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$



$$\hat{\beta}_1 = \frac{\text{Covariância amostral entre x e y}}{\text{Variância amostral de x}}$$

- Se x e y são positivamente correlacionados na amostra, $\hat{\beta}_1$ é positivo e vice-versa.

VALORES ESTIMADOS E RESÍDUOS

- Encontrados o intercepto e a inclinação, teremos um valor estimado para y para cada observação (x) na amostra:

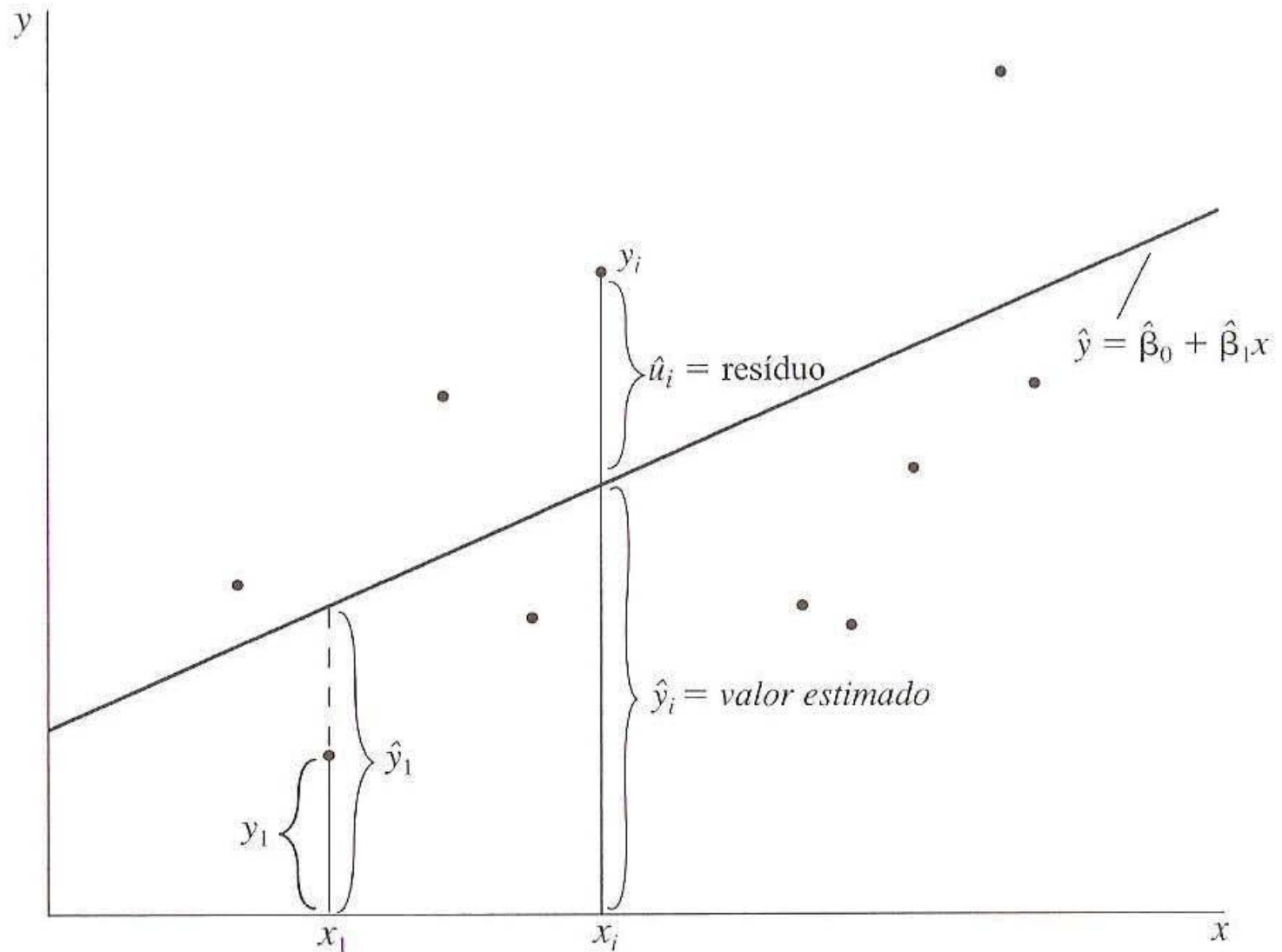
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- O resíduo é a diferença entre o valor verdadeiro de y_i e seu valor estimado:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Figura 2.4

Valores estimados e resíduos.



MINIMIZANDO A SOMA DOS RESÍDUOS QUADRADOS

- Suponha que escolhemos o intercepto e a inclinação estimados com o propósito de tornar a soma dos resíduos quadrados:

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- O nome “mínimos quadrados ordinários” é utilizado porque as estimativas do intercepto e da inclinação minimizam a soma dos resíduos quadrados.
- Não é utilizada a minimização dos valores absolutos dos resíduos, porque a teoria estatística para isto seria muito complicada

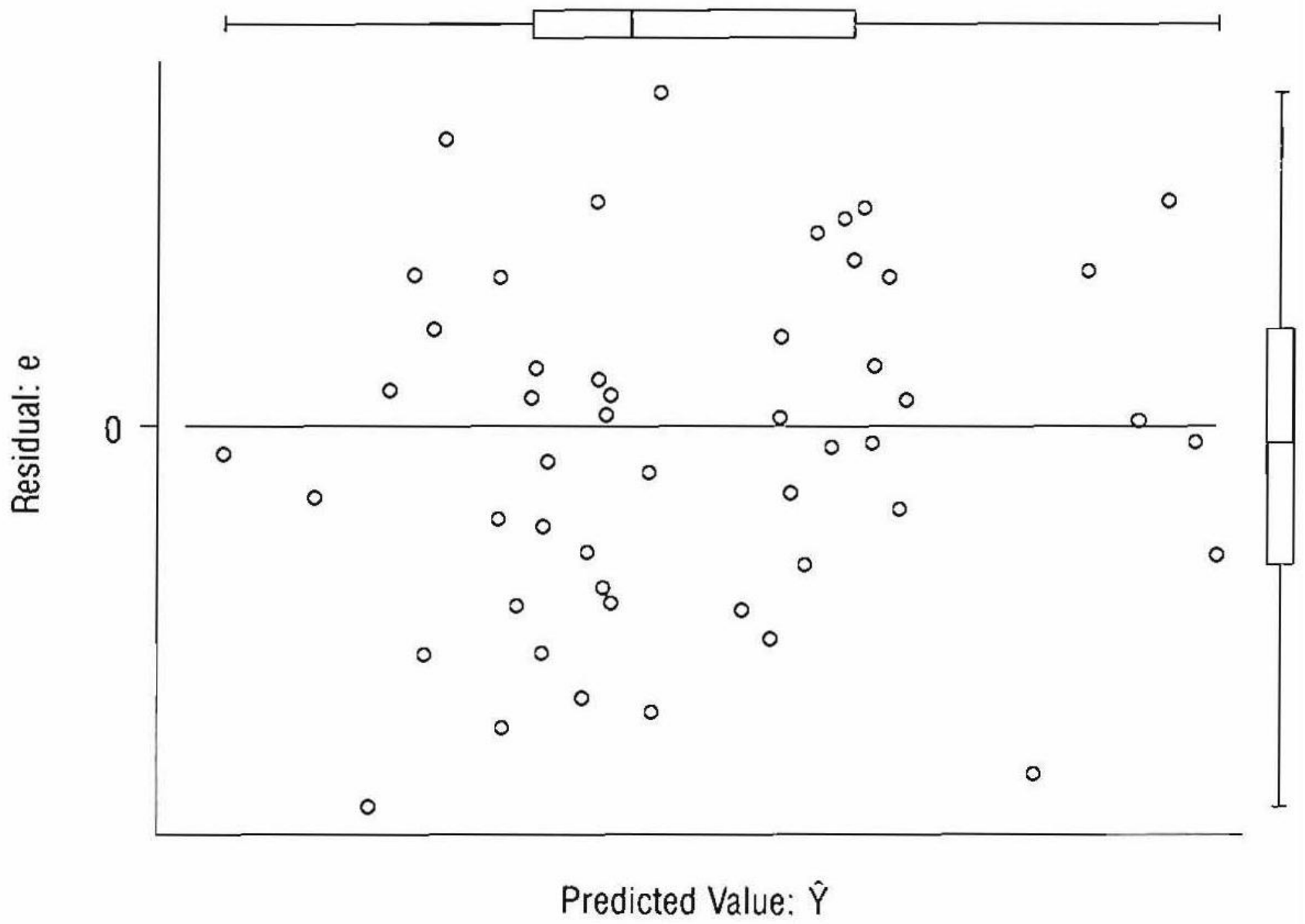
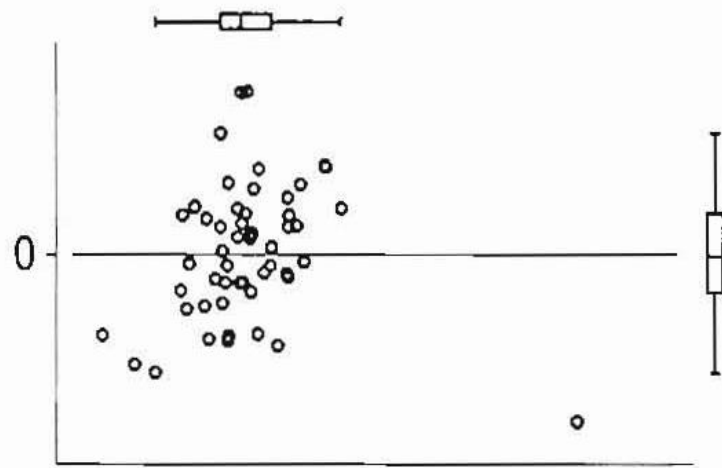
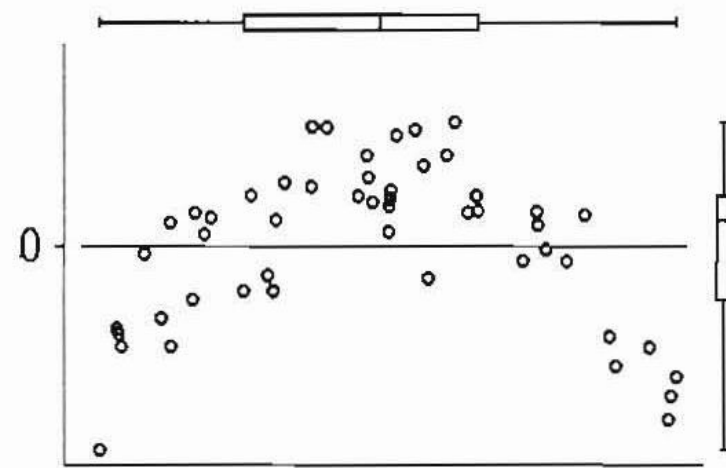


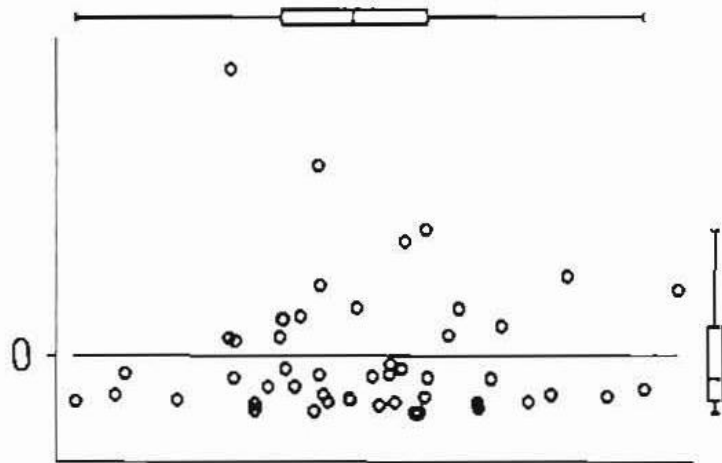
Figure 2.10 “All clear” e -versus- \hat{Y} plot (artificial data).



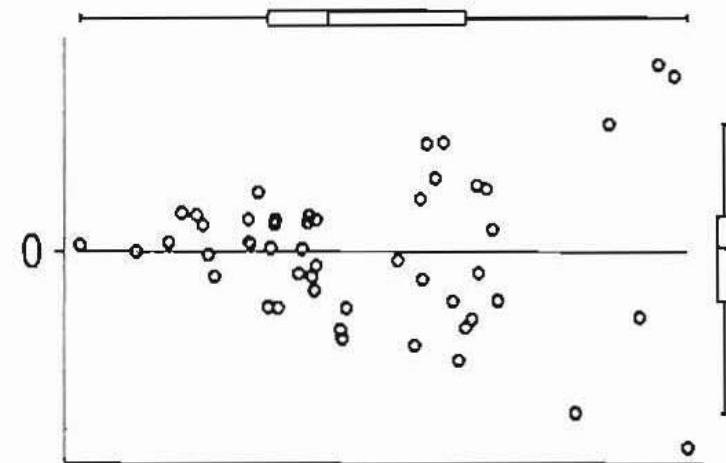
Influential Case



Curvilinear Relation



Nonnormal Residual Distribution



Heteroscedasticity

Figure 2.11 Examples of trouble seen in e -versus- \hat{Y} plots (artificial data).

SOMAS DOS QUADRADOS

- Soma dos quadrados total (SQT) é uma medida da variação amostral total em y_i (mede a dispersão dos y_i na amostra):

$$SQT = \sum_{i=1}^n (y_i - \bar{y})^2$$

- Soma dos quadrados explicada (SQE) mede a variação amostral em :

$$SQE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- Soma dos quadrados dos resíduos (SQR) mede a variação amostral em :

$$SQR = \sum_{i=1}^n \hat{u}_i^2$$

- Variação total em y é a soma da variação explicada e da variação não-explicada:

$$SQT = SQE + SQR$$

GRAU DE AJUSTE

- Visa mensurar o quanto bem a variável independente (x) explica a variável dependente (y).
- É um número que resume o quão bem a reta de regressão de MQO se ajusta aos dados.
- R^2 : razão entre a variação explicada (SQE) e a variação total (SQT).
- R^2 : fração da variação amostral em y que é explicada por x.

$$SQT = SQE + SQR$$

$$SQT/SQT = (SQE + SQR)/SQT$$

$$1 = SQE/SQT + SQR/SQT$$

$$SQE/SQT = 1 - SQR/SQT$$

- Usar o R^2 como principal padrão de medida de sucesso de uma análise econométrica pode levar a confusões.

NÃO-LINEARIDADE NA REGRESSÃO SIMPLES

- Formas funcionais populares usadas em economia podem ser incorporadas à análise de regressão.
- Até agora foram analisadas relações lineares entre as variáveis dependente e independente.
- No entanto, relações lineares não são suficientes para todas as aplicações econômicas e sociais.
- É fácil incorporar não-linearidade na análise de regressão simples.

EXEMPLO DE NÃO-LINEARIDADE

- Para cada ano adicional de educação, há um aumento fixo no salário. Esse é o aumento tanto para o primeiro ano de educação quanto para anos mais avançados:

$$\textit{salário} = \beta_0 + \beta_1 \textit{educ} + u$$

- Suponha que o aumento percentual no salário é o mesmo, dado um ano a mais de educação formal. Um modelo que gera um efeito percentual constante é dado por:

$$\log(\textit{salário}) = \beta_0 + \beta_1 \textit{educ} + u$$

- Se $\Delta u = 0$, então:

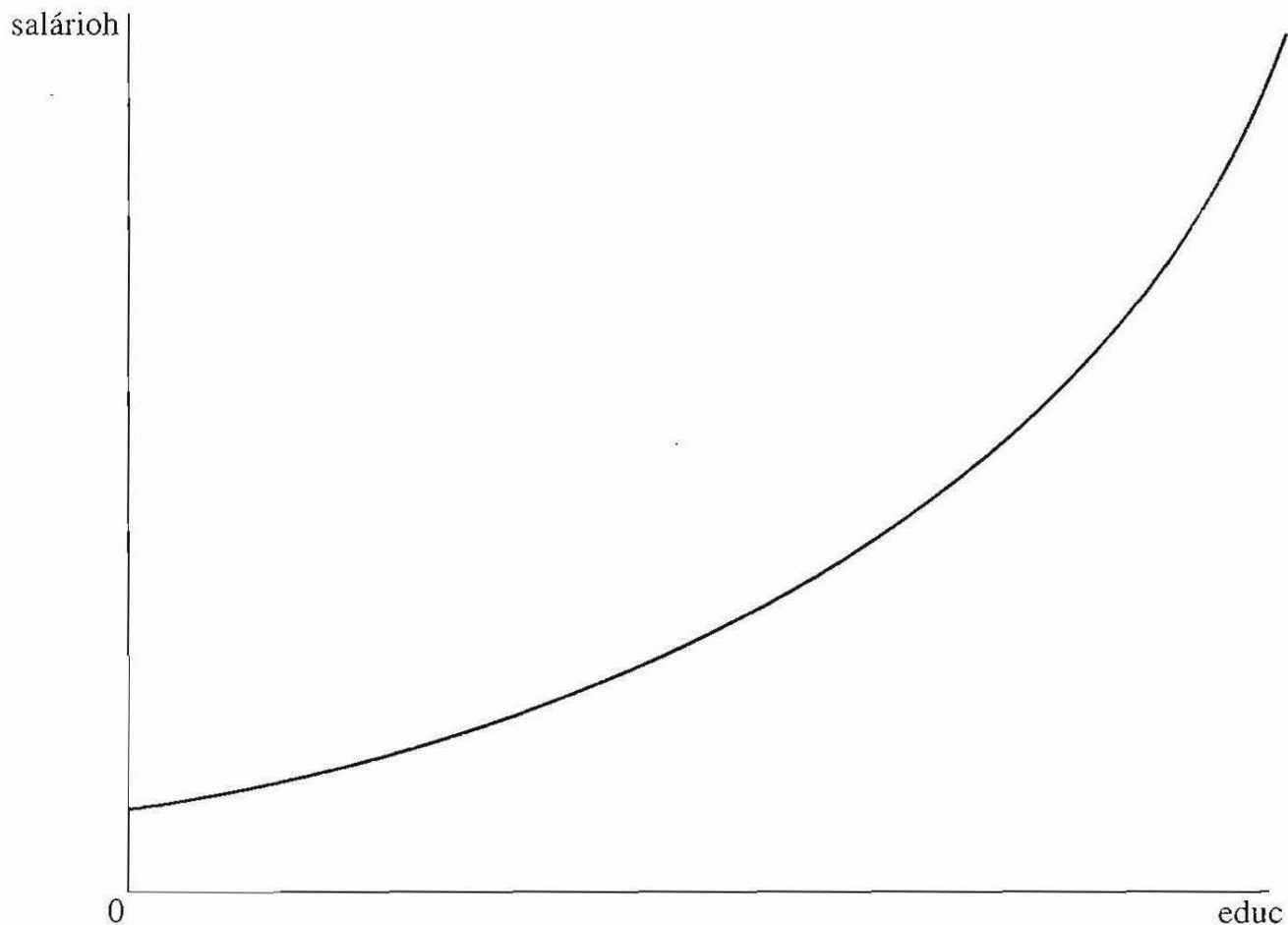
$$\% \Delta \textit{salário} = (100 * \beta_1) \Delta \textit{educ}$$

- Para cada ano adicional de educação, há um aumento de ?% sobre o salário.

- Como a variação percentual no salário é a mesma para cada ano adicional de educação, a variação no salário aumenta quando a educação formal aumenta.

Figura 2.6

$$\text{saláριο} = \exp(\beta_0 + \beta_1 \text{educ}), \text{ com } \beta_1 > 0.$$



INTERPRETAÇÃO DOS COEFICIENTES

- Aumento de uma unidade em x aumenta y em β_1 unidades:

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

- Aumento de 1% em x aumenta y em $(\beta_1/100)$ unidades:

$$y = \beta_0 + \beta_1 \log(x) + u$$

- Aumento de uma unidade em x aumenta y em $(100*\beta_1)\%$:

$$\log(y) = \beta_0 + \beta_1 x + u$$

- Aumento de 1% em x aumenta y em $\beta_1\%$:

$$\log(y) = \beta_0 + \beta_1 \log(x) + u$$

- Este último é o modelo de elasticidade constante.
- Elasticidade é a razão entre o percentual de mudança em uma variável e o percentual de mudança em outra variável.

FORMAS FUNCIONAIS ENVOLVENDO LOGARITMOS

| Modelo | Variável Dependente | Variável Independente | Interpretação de β_1 |
|-------------|---------------------|-----------------------|--|
| nível-nível | y | x | $\Delta y = \beta_1 \Delta x$ |
| nível-log | y | $\log(x)$ | $\Delta y = (\beta_1 / 100) \% \Delta x$ |
| log-nível | $\log(y)$ | x | $\% \Delta y = (100 \beta_1) \Delta x$ |
| log-log | $\log(y)$ | $\log(x)$ | $\% \Delta y = \beta_1 \% \Delta x$ |

SIGNIFICADO DE REGRESSÃO LINEAR

- O modelo de regressão linear permite relações não-lineares.
- Esse modelo é linear nos parâmetros: β_0 e β_1 .
- Não há restrições de como y e x se relacionam com as variáveis dependente e independente originais, já que podemos utilizar: logaritmo natural, quadrado, raiz quadrada...
- A interpretação dos coeficientes depende das definições de como x e y são construídos.
- “É muito mais importante tornar-se proficiente em interpretar coeficientes do que eficiente no cálculo de fórmulas.”
(Wooldridge, 2008: 45)

CAPÍTULO 3
ANÁLISE DE REGRESSÃO MÚLTIPLA:
ESTIMAÇÃO

MODELO DE REGRESSÃO MÚLTIPLA

- A desvantagem de usar análise de regressão simples é o fato de ser difícil que todos os outros fatores que afetam y não estejam correlacionados com x .
- Análise de regressão múltipla possibilita *ceteris paribus* (outros fatores constantes), pois permite controlar muitos outros fatores que afetam a variável dependente simultaneamente.
- Isso auxilia no teste de teorias econômicas e na avaliação de impactos de políticas públicas, quando possuímos dados não-experimentais.
- Ao utilizar mais fatores na explicação de y , uma maior variação de y será explicada pelo modelo.
- Este é o modelo mais utilizado nas ciências sociais.
- O método de MQO é usado para estimar os parâmetros do modelo de regressão múltipla.

MODELO COM DUAS VARIÁVEIS INDEPENDENTES

$$\text{salário}_h = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + u$$

- Salário é determinado por educação, experiência e outros fatores não-observáveis (Equação Minceriana).
- β_1 mede o efeito de educação sobre salário, mantendo todos os outros fatores fixos (*ceteris paribus*).
- β_2 mede o efeito de experiência sobre salário, mantendo todos os outros fatores fixos.
- Como experiência foi inserida na equação, podemos medir o efeito de educação sobre salário, mantendo experiência fixa.
- Na regressão simples, teríamos que assumir que experiência não é correlacionada com educação, o que é uma hipótese fraca.

MODELO GERAL DE DUAS VARIÁVEIS INDEPENDENTES

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- β_0 é o intercepto.
- β_1 mede a variação em y com relação a x_1 , mantendo os outros fatores constantes.
- β_2 mede a variação em y com relação a x_2 , mantendo os outros fatores constantes.

HIPÓTESE SOBRE u EM RELAÇÃO A x_1 E x_2

$$E(u/x_1, x_2) = 0$$

- Para qualquer valor de x_1 e x_2 na população, o fator não-observável médio é igual a zero.
- Isso implica que outros fatores que afetam y não estão, em média, relacionados com as variáveis explicativas.
- Os níveis médios dos fatores não-observáveis devem ser os mesmos nas combinações das variáveis independentes.
- A esperança igual a zero significa que a relação funcional entre as variáveis explicada e as explicativas está correta.

MODELO COM k VARIÁVEIS INDEPENDENTES

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

- Esse é o modelo de regressão linear múltipla geral ou, simplesmente, modelo de regressão múltipla.
- Há $k + 1$ parâmetros populacionais desconhecidos, já que temos k variáveis independentes e um intercepto.
- Os parâmetros β_1 a β_k são chamados de parâmetros de inclinação, mesmo que eles não tenham exatamente este significado.
- **A regressão é “linear” porque é linear nos β_j , mesmo que seja uma relação não-linear entre a variável dependente e as variáveis independentes:**

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_3 x_2^2 + u$$

INTERPRETAÇÃO DA EQUAÇÃO DE REGRESSÃO

- Novamente a reta de regressão de MQO:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k + u$$

- O intercepto é o valor previsto de y quando todas as variáveis independentes são iguais a zero.
- As estimativas dos demais parâmetros têm interpretações de efeito parcial (*ceteris paribus*).
- Da equação acima, temos:

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \dots + \hat{\beta}_k \Delta x_k$$

- O coeficiente de x_1 mede a variação em y devido a um aumento de uma unidade em x_1 , mantendo todas as outras variáveis independentes constantes:

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1, \text{ sendo: } \Delta x_2 = \dots = \Delta x_k = 0$$

SIGNIFICADO DE “MANTER OUTROS FATORES FIXOS”

- Regressão múltipla permite interpretação *ceteris paribus* mesmo que dados não sejam coletados de maneira *ceteris paribus*.
- Os dados são coletados por amostra aleatória que não estabelece restrições sobre os valores a serem obtidos das variáveis independentes.
- Ou seja, a regressão múltipla permite simular situação de outros fatores constantes, sem restringir a coleta de dados.
- Essa modelagem permite realizar em ambientes não-experimentais o que cientistas naturais realizam em experimentos de laboratório (mantendo outros fatores fixos).
- A avaliação de impacto de políticas pode ser realizada com regressão múltipla, mensurando relação entre variáveis independentes e dependente, com noção de *ceteris paribus*.

GRAU DE AJUSTE

- O R^2 nunca diminui quando outra variável independente é adicionada na regressão.
- Isso ocorre porque a soma dos resíduos quadrados nunca aumenta quando variáveis explicativas são acrescentadas ao modelo.
- Essa característica faz de R^2 um teste fraco para decidir pela inclusão de variáveis no modelo.
- O efeito parcial da variável independente (β_k) sobre y é o que deve definir se a variável deve ser inserida no modelo.
- R^2 é um grau de ajuste geral do modelo, assim como um teste para indicar o quanto um grupo de variáveis explica variações em y .

VALOR ESPERADOS DOS ESTIMADORES DE MQO

HIPÓTESE RLM.1 (LINEAR NOS PARÂMETROS)

- Modelo na população pode ser escrito como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

- $\beta_0, \beta_1, \dots, \beta_k$ são parâmetros desconhecidos (constantes) de interesse, e u é um erro aleatório não-observável ou um termo de perturbação aleatória.

HIPÓTESE RLM.2 (AMOSTRAGEM ALEATÓRIA)

- Temos uma amostra aleatória de n observações do modelo populacional acima.

HIPÓTESE RLM.3 (MÉDIA CONDICIONAL ZERO)

- O erro u tem um valor esperado igual a zero, dados quaisquer valores das variáveis independentes:

$$E(u|x_1, x_2, \dots, x_k) = 0$$

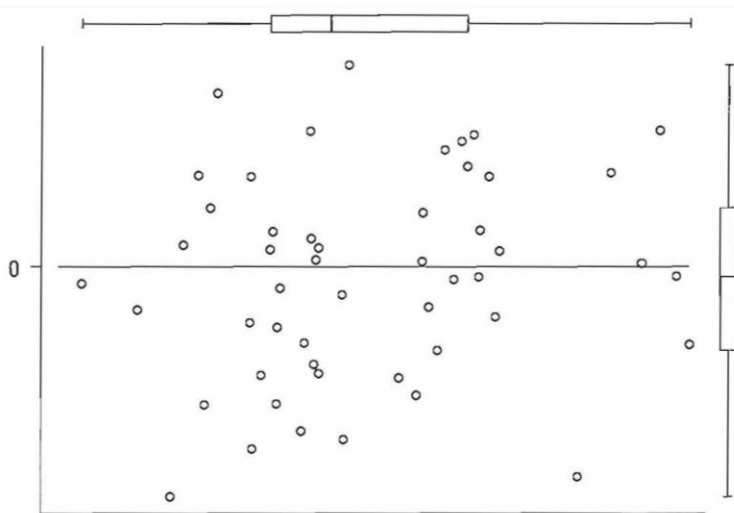
HIPÓTESE RLM.4 (COLINEARIDADE NÃO PERFEITA)

- Na amostra e na população, nenhuma das variáveis independentes é constante, e não há relações lineares exatas entre as variáveis independentes.
- As variáveis independentes devem ser correlacionadas entre si, mas não deve haver **colinearidade perfeita** (por exemplo, uma variável não pode ser múltiplo de outra).
- Altos graus de correlação entre variáveis independentes e tamanho pequeno da amostra aumentam variância de beta.
- Correlação alta (mas não perfeita) entre duas ou mais variáveis não é desejável (**multicolinearidade**).
- Por outro lado, se a correlação for nula, não é necessário regressão múltipla, mas sim regressão simples, já que o termo de erro englobaria todos fatores não-observáveis e não-relacionados com as variáveis independentes.

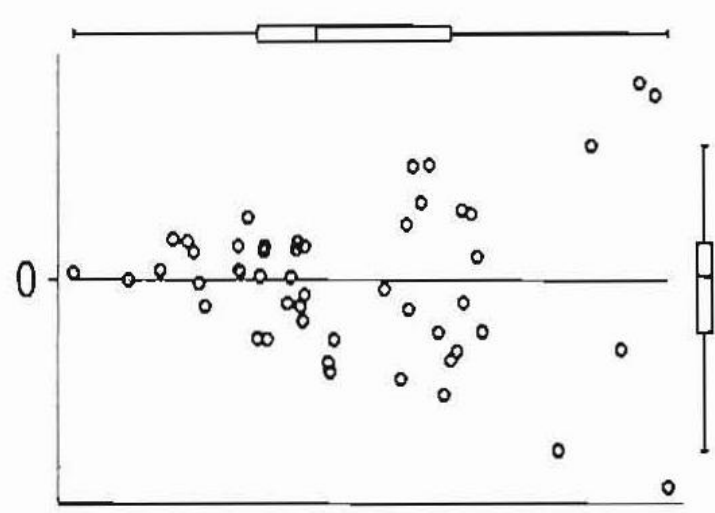
HIPÓTESE RLM.5 (HOMOSCEDASTICIDADE)

- A variância do termo erro (u), condicionada às variáveis explicativas, é a mesma para todas as combinações de resultados das variáveis explicativas.
- Se essa hipótese é violada, o modelo exibe heteroscedasticidade.

HOMOSCEDASTICIDADE



HETEROSCEDASTICIDADE



Fonte: Hamilton, 1992: 52-53.

HOMOSCEDASTICIDADE E HETEROSCEDASTICIDADE

- A hipótese de homoscedasticidade para a regressão múltipla significa que a variância do erro não observável (u), condicional nas variáveis explicativas, é constante.
- A homoscedasticidade não se mantém quando a variância dos fatores não-observáveis muda ao longo de diferentes segmentos da população.
- Por exemplo, a heteroscedasticidade está presente se a variância dos fatores não-observados (u) que afetam a renda (y) aumenta com a idade (x).
- A homoscedasticidade é necessária para estimar os testes de t e F , além dos intervalos de confiança.

INFERÊNCIA ROBUSTA

- É possível ajustar erros-padrão, estatísticas t e F de forma a torná-las válidas na presença de heteroscedasticidade de forma desconhecida.
- Isso significa que é possível descrever novas estatísticas que funcionam independentemente do tipo de heteroscedasticidade presente na população.
- Esses métodos são os procedimentos robustos em relação à heteroscedasticidade, já que são válidos mesmo que a variância dos erros não seja constante.
- É possível então estimar variâncias consistentes na presença de heteroscedasticidade.
- A aplicação de métodos robustos em relação à heteroscedasticidade é bastante fácil, pois muitos programas estatísticos e econométricos calculam essas estatísticas como uma opção.

TEOREMA DE GAUSS-MARKOV

- Sob as hipóteses RLM.1 a RLM.5, os parâmetros estimados do intercepto e de inclinação são os melhores estimadores lineares não-viesados dos parâmetros populacionais:

Best Linear Unbiased Estimators (BLUEs)

- Em outras palavras, os estimadores de mínimos quadrados ordinários (MQO) são os melhores estimadores lineares não-viesados.

CAPÍTULO 4
ANÁLISE DE REGRESSÃO MÚLTIPLA:
INFERÊNCIA

TRANSFORMAÇÃO É QUESTÃO EMPÍRICA

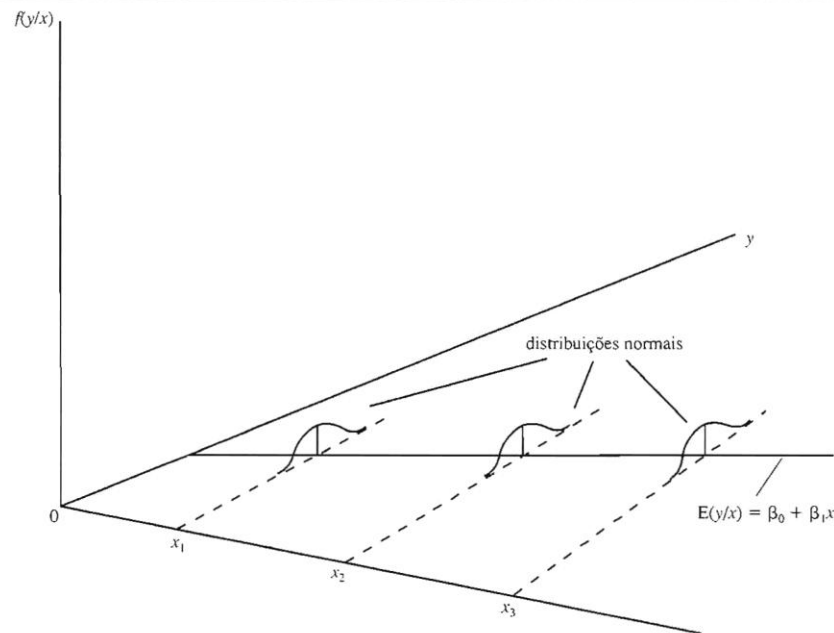
- Os objetivos de realizar transformações de variáveis independentes e dependente são:
 - Alcançar distribuição normal da variável dependente.
 - Estabelecer correta relação entre variável dependente e independentes.
- Fazer uma transformação de salário, especialmente tomando o log, produz uma distribuição que está mais próxima da normal.
- Sempre que y assume apenas alguns valores, não podemos ter uma distribuição próxima de uma distribuição normal.
- “Essa é uma questão empírica.” (Wooldridge, 2008: 112)

MODELO LINEAR CLÁSSICO

- As hipóteses BLUE, adicionadas à hipótese da normalidade (erro não-observado é normalmente distribuído na população), são conhecidas como hipóteses do modelo linear clássico (MLC).
- Distribuição normal homoscedástica com uma única variável explicativa:

Figura 4.1

A distribuição normal homoscedástica com uma única variável explicativa.



TESTES DE HIPÓTESE

- Podemos fazer testes de hipóteses sobre um único parâmetro da função de regressão populacional.
- Os β_j são características desconhecidas da população.
- Na maioria das aplicações, nosso principal interesse é testar a hipótese nula ($H_0: \beta_j = 0$).
- Como β_j mede o efeito parcial de x_j sobre o valor esperado de y , após controlar todas as outras variáveis independentes, a hipótese nula significa que, uma vez que x_1, x_2, \dots, x_k foram considerados, x_j não tem nenhum efeito sobre o valor esperado de y .
- O teste de hipótese na regressão múltipla é semelhante ao teste de hipótese para a média de uma população normal.
- É difícil obter os coeficientes, erros-padrão e valores críticos, mas os programas econométricos (nosso amigo Stata) calculam estas estimativas automaticamente.

TESTE t

- A estatística t é a razão entre o coeficiente estimado (β_j) e seu erro padrão: $ep(\beta_j)$.
- O erro padrão é sempre positivo, então a razão t sempre terá o mesmo sinal que o coeficiente estimado.
- Valor estimado de beta distante de zero é evidência contra a hipótese nula, mas devemos ponderar pelo erro amostral.
- Como o erro-padrão de β_j é uma estimativa do desvio-padrão de β_j , o teste t mede quantos desvios-padrão estimados β_j está afastado de zero.
- Isso é o mesmo que testar se a média de uma população é zero usando a estatística t padrão.
- A regra de rejeição depende da hipótese alternativa e do nível de significância escolhido do teste.
- Sempre testamos hipótese sobre parâmetros populacionais, e não sobre estimativas de uma amostra particular.

p*-VALORES DOS TESTES *t

- Dado o valor observado da estatística t , qual é o menor nível de significância ao qual a hipótese nula seria rejeitada?
- Não há nível de significância “correto”.
- O p -valor é a probabilidade da hipótese nula ser verdadeira:
 - p -valores pequenos são evidências contra hipótese nula.
 - p -valores grandes fornecem pouca evidência contra H_0 .
- Se α é o nível de significância do teste, então H_0 é rejeitada se $p\text{-valor} < \alpha$.
- H_0 não é rejeitada ao nível de $100*\alpha\%$.

TESTE: HIPÓTESES ALTERNATIVAS BILATERAIS

$$H_1: \beta_j \neq 0$$

- Essa hipótese é relevante quando o sinal de β_j não é bem determinado pela teoria.
- Usar as estimativas da regressão para nos ajudar a formular as hipóteses nula e alternativa não é permitido, porque a inferência estatística clássica pressupõe que formulamos as hipóteses nula e alternativa sobre a população antes de olhar os dados.
- Quando a alternativa é bilateral, estamos interessados no valor absoluto da estatística t . $|t| > c$.
- Para um nível de significância de 5% e em um teste bicaudal, c é escolhido de forma que a área em cada cauda da distribuição t seja igual a 2,5%.
- Se H_0 é rejeitada, x_j é estatisticamente significativa (ou estatisticamente diferente de zero) ao nível de 5%.

REGRA DE REJEIÇÃO DE H_0 (BILATERAL)

Figura 4.4

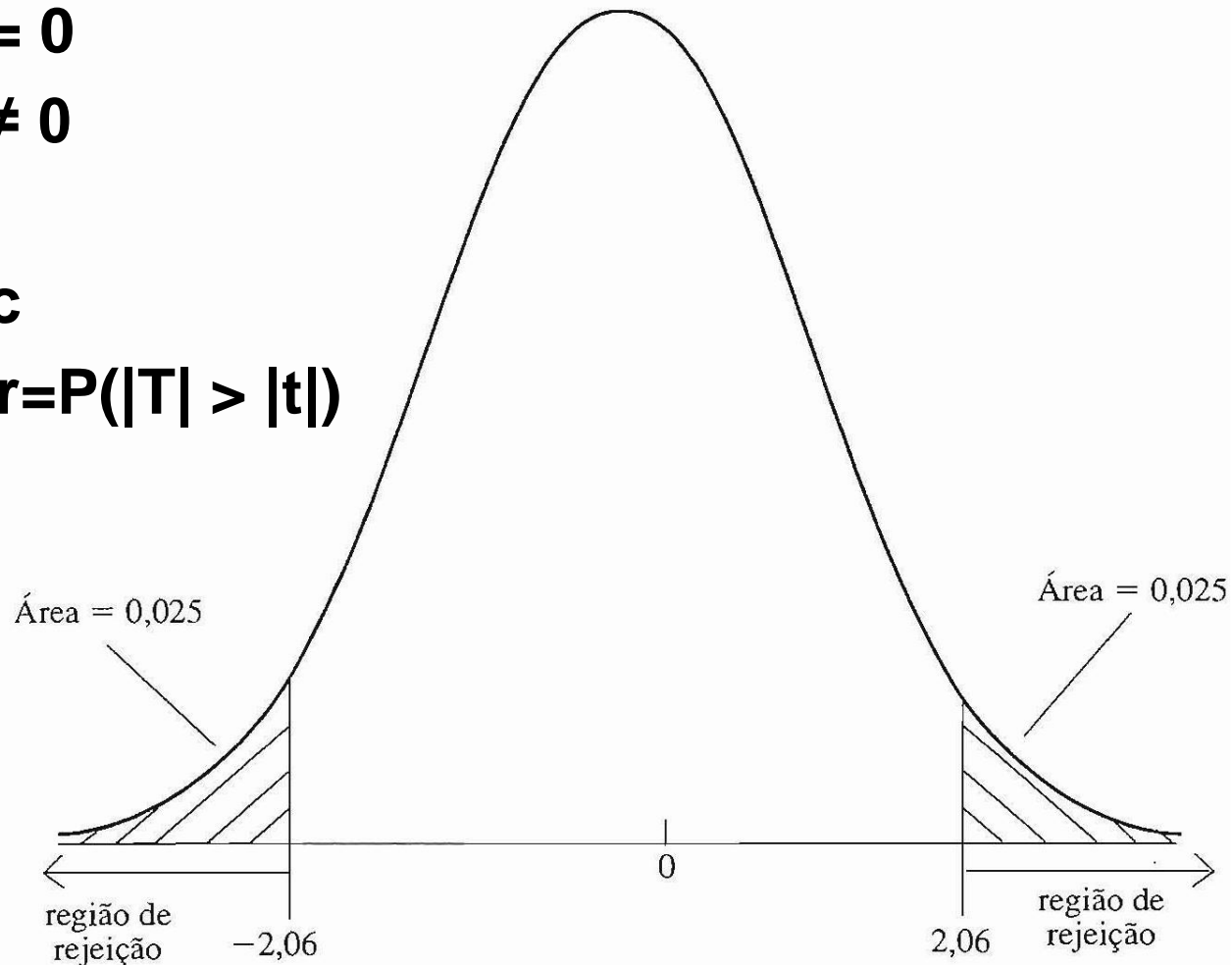
Regra de rejeição a 5% para a hipótese alternativa $H_1: \beta_j \neq 0$ com 25 gl.

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

$$|t_{\beta_j}| > c$$

$$\text{p-valor} = P(|T| > |t|)$$



SIGNIFICÂNCIA ECONÔMICA X ESTATÍSTICA

- É importante levar em consideração a magnitude das estimativas dos coeficientes, além do tamanho das estatísticas t .
- A **significância estatística** de uma variável x_j é determinada completamente pelo tamanho do teste t .
- A **significância econômica** (ou significância prática) da variável está relacionada ao tamanho e sinal do coeficiente beta estimado.
- Colocar muita ênfase sobre a significância estatística pode levar à conclusão falsa de que uma variável é importante para explicar y embora seu efeito estimado seja moderado.
- Com amostras grandes, os erros-padrão são pequenos, o que resulta em significância estatística.
- Erros-padrão grandes podem ocorrer por alta correlação entre variáveis independentes (multicolinearidade).

DISCUTINDO AS SIGNIFICÂNCIAS

- Verifique a **significância econômica**, lembrando que as unidades das variáveis independentes e dependente mudam a interpretação dos coeficientes beta.
- Verifique a **significância estatística**, a partir do teste t de cada variável.
- Se: (1) sinal esperado e (2) teste t grande, a variável é **significante economicamente e estatisticamente**.
- Se: (1) sinal esperado e (2) teste t pequeno, podemos aceitar p -valor maior, quando amostra é pequena (mas é arriscado, pois pode ser problema no desenho amostral).
- Se: (1) sinal não esperado e (2) teste t pequeno, variável **não significativa economicamente e estatisticamente**.
- Se: (1) sinal não esperado e (2) teste t grande, é problema sério em variáveis importantes (falta incluir variáveis ou há problema nos dados).

EXEMPLO DE REGRESSÃO MÚLTIPLA

Exemplo 3.5 (páginas 78 e 79):

narr86 = número de vezes que determinado homem foi preso em 1986.

pcnv = proporção de prisões anteriores a 1986 que levaram à condenação.

avgsen = duração média da sentença cumprida por condenação prévia.

ptime86 = meses passados na prisão em 1986.

qemp86 = número de trimestres que determinado ficou empregado em 1986.

```
reg narr86 pcnv avgsen ptime86 qemp86
```

| Source | SS | df | MS | | |
|----------|------------|------|------------|-----------------|--------|
| Model | 84.8242895 | 4 | 21.2060724 | Number of obs = | 2725 |
| Residual | 1925.52287 | 2720 | .707912819 | F(4, 2720) = | 29.96 |
| Total | 2010.34716 | 2724 | .738012906 | Prob > F = | 0.0000 |
| | | | | R-squared = | 0.0422 |
| | | | | Adj R-squared = | 0.0408 |
| | | | | Root MSE = | .84138 |

| narr86 | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|---------|-----------|-----------|--------|-------|----------------------|-----------|
| pcnv | -.1508319 | .0408583 | -3.692 | 0.000 | -.2309484 | -.0707154 |
| avgsen | .0074431 | .0047338 | 1.572 | 0.116 | -.0018392 | .0167254 |
| ptime86 | -.0373908 | .0087941 | -4.252 | 0.000 | -.0546345 | -.0201471 |
| qemp86 | -.103341 | .0103965 | -9.940 | 0.000 | -.1237268 | -.0829552 |
| _cons | .7067565 | .0331515 | 21.319 | 0.000 | .6417519 | .771761 |

TESTE *F*: TESTE DE RESTRIÇÕES DE EXCLUSÃO

- Testar se um grupo de variáveis não tem efeito sobre a variável dependente.
- A hipótese nula é que um conjunto de variáveis não tem efeito sobre y (β_3 , β_4 e β_5 , por exemplo), já que outro conjunto de variáveis foi controlado (β_1 e β_2 , por exemplo).
- Esse é um exemplo de restrições múltiplas.
- $H_0: \beta_3=0, \beta_4=0, \beta_5=0$.
- $H_1: H_0$ não é verdadeira.
- Quando pelo menos um dos betas for diferente de zero, rejeitamos a hipótese nula.

ESTATÍSTICA F (OU RAZÃO F)

- Precisamos saber o quanto SQR aumenta, quando retiramos as variáveis que estamos testando.
- Modelo restrito terá β_0 , β_1 e β_2 .
- Modelo irrestrito terá β_0 , β_1 , β_2 , β_3 , β_4 e β_5 .
- A estatística F é definida como:

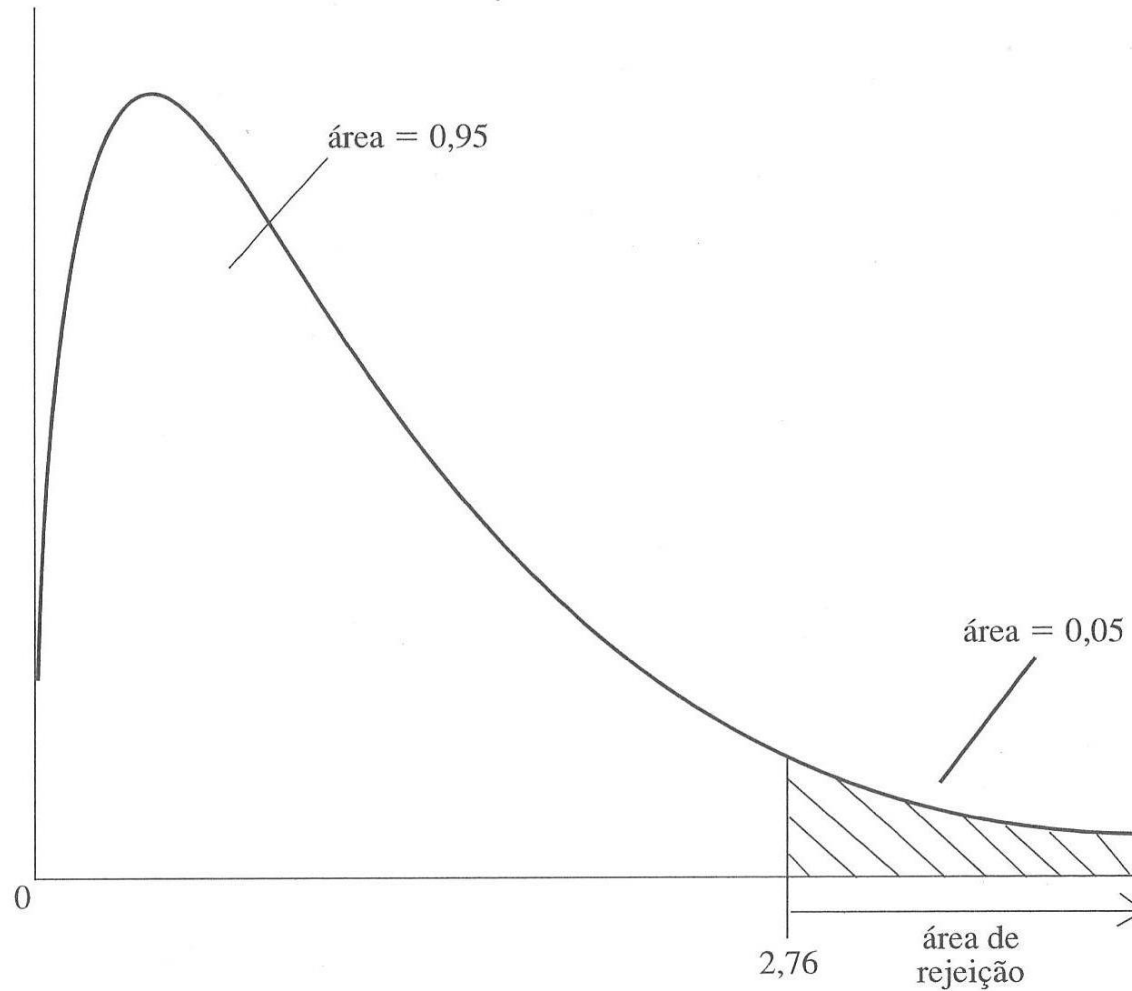
$$F \equiv \frac{(SQR_r - SQR_{ir})/q}{SQR_{ir}/(n - k - 1)}$$

- SQR_r é a soma dos resíduos quadrados do modelo restrito.
- SQR_{ir} é a soma dos resíduos quadrados do modelo irrestrito.
- q é o número de variáveis independentes retiradas (neste caso temos três: β_3 , β_4 e β_5), ou seja, $q = gl_r - gl_{ir}$.

CURVA DA DISTRIBUIÇÃO F

Figura 4.7

O valor crítico de 5% e a região de rejeição em uma distribuição $F_{3,60}$.



DESCRIÇÃO DOS RESULTADOS DA REGRESSÃO

- Informar os **coeficientes** estimados de MQO (betas).
- Interpretar **significância econômica** (prática) dos coeficientes das variáveis fundamentais, levando em consideração as unidades de medida.
- Interpretar **significância estatística**, ao incluir erros-padrão entre parênteses abaixo dos coeficientes (ou estatísticas t , ou p -valores, ou asteriscos).
 - Erro padrão é preferível, pois podemos: (1) testar hipótese nula quando parâmetro populacional não é zero; (2) calcular intervalos de confiança.
- Informar o **R-quadrado**: (1) grau de ajuste; (2) cálculo de F.
- **Número de observações** usado na estimação (n).
- Apresentar resultados em **equações** ou **tabelas** (indicar variável dependente, além de independentes na 1ª coluna).
- Mostrar **SQR** e **erro-padrão** (Root MRE), mas não é crucial.

CAPÍTULO 6
ANÁLISE DE REGRESSÃO MÚLTIPLA:
PROBLEMAS ADICIONAIS
(INTERAÇÕES)

MODELOS COM TERMOS DE INTERAÇÃO

- O efeito de uma variável independente, sobre a variável dependente, pode depender de outra variável explicativa:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

- O efeito parcial de x_2 sobre y é: $\Delta y / \Delta x_2 = \beta_2 + \beta_3 x_1$.
- β_2 é o efeito parcial de x_2 sobre y , quando $x_1=0$, o que pode não ser de interesse prático.
- Podemos então reparametrizar o modelo, tal como:

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u, \text{ sendo:}$$

μ_1 e μ_2 médias populacionais de x_1 e x_2 .

- δ_2 é o efeito parcial de x_2 sobre y , quando $x_1 = \mu_1$:

$$\delta_2 = \beta_2 + \beta_3 \mu_1$$

- É complicado interpretar modelos com termos de interação.

INTERAÇÕES ENTRE VARIÁVEIS BINÁRIAS

- Podemos realizar interações entre variáveis binárias:

$$\log(\text{salário}) = B_0 + B_1 \text{ hcasados} + B_2 \text{ mcasadas} + B_3 \text{ msolteiras}$$

- A equação acima permite testar diretamente diferenças entre qualquer grupo e homens solteiros.

- Podemos adicionar um termo de interação diretamente:

$$\log(\text{salário}) = B_0 + B_1 \text{ feminino} + B_2 \text{ casado} + B_3 \text{ fem}^* \text{casado}$$

- Os coeficientes de cada grupo serão:

- Homens solteiros: B_0

- Homens casados: $B_0 + B_2$

- Mulheres solteiras: $B_0 + B_1$

- Mulheres casadas: $B_0 + B_1 + B_2 + B_3$

- Nessa equação, B_3 permite testar diretamente se diferencial de sexo depende do estado civil e vice-versa.

INCLINAÇÕES DIFERENTES

- Existem casos de interação de variáveis binárias com variáveis explicativas que não são binárias para permitir diferença nas inclinações.
- Podemos testar se retorno da educação é o mesmo para homens e mulheres, considerando um diferencial de salários constante entre homens e mulheres:

$$\log(\text{salário}) = (\beta_0 + \delta_0 \text{feminino}) + (\beta_1 + \delta_1 \text{feminino}) * \text{educ} + u$$

- Homens: intercepto (β_0) e inclinação (β_1)
- Mulheres: intercepto ($\beta_0 + \delta_0$) e inclinação ($\beta_1 + \delta_1$)
- δ_0 : diferença nos interceptos entre mulheres e homens.
- δ_1 : diferença no retorno da educação entre sexos.

- No Stata:

$$\log(\text{salário}) = \beta_0 + \delta_0 \text{feminino} + \beta_1 \text{educ} + \delta_1 \text{fem} * \text{educ} + u$$

- Quando $\delta_0 + (\delta_1 * \text{educ}) = 0$, salário é igual entre sexos.

GRÁFICO A: intercepto e inclinação das mulheres é inferior.

GRÁFICO B: intercepto das mulheres é inferior, mas inclinação é superior.

Figura 7.2

Gráficos da equação (7.16). (a) $\delta_0 < 0, \delta_1 < 0$; (b) $\delta_0 < 0, \delta_1 > 0$.

