

# **AULAS 17 E 18**

# **Análise de Regressão Múltipla: Inferência**

**Ernesto F. L. Amaral**

**12 e 17 de maio de 2011**  
**Avaliação de Políticas Públicas (DCP 046)**

**Fonte:**

**Wooldridge, Jeffrey M. “Introdução à econometria: uma abordagem moderna”. São Paulo:  
Cengage Learning, 2008. Capítulo 4 (pp.110-157).**

# TRANSFORMAÇÃO É QUESTÃO EMPÍRICA

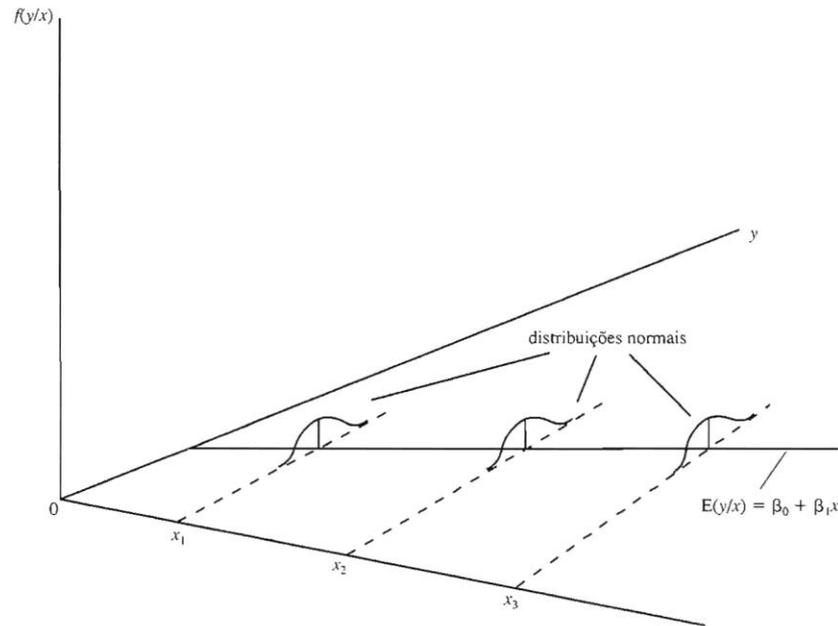
- Os objetivos de realizar transformações de variáveis independentes e dependente são:
  - Alcançar distribuição normal da variável dependente.
  - Estabelecer correta relação entre variável dependente e independentes.
- Fazer uma transformação de salário, especialmente tomando o log, produz uma distribuição 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  $\beta_j$  são características desconhecidas da população.
- Na maioria das aplicações, nosso principal interesse é testar a hipótese nula ( $H_0: \beta_j = 0$ ).
- Como  $\beta_j$  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 ( $\beta_j$ ) 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  $\beta_j$  é uma estimativa do desvio-padrão de  $\beta_j$ , **o teste  $t$  mede quantos desvios-padrão estimados  $\beta_j$  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  $\alpha$  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 UNILATERAIS

$$H_1: \beta_j > 0 \quad \text{OU} \quad H_1: \beta_j < 0$$

- Devemos decidir sobre um nível de significância (geralmente de 5%).
- Estamos dispostos a rejeitar erroneamente  $H_0$ , quando ela é verdadeira 5% das vezes.
- Um valor suficientemente grande de  $t$ , com um nível de significância de 5%, é o 95<sup>o</sup> percentil de uma distribuição  $t$  com  $n-k-1$  graus de liberdade (ponto c).
- **Regra de rejeição** é que  $H_0$  é rejeitada em favor de  $H_1$ , se  $t > c$  ( $H_1: \beta_j > 0$ ) ou  $t < -c$  ( $H_1: \beta_j < 0$ ), em um nível específico.
- Quando os graus de liberdade da distribuição  $t$  ficam maiores, a distribuição  $t$  aproxima-se da distribuição normal padronizada.
- Para graus de liberdade maiores que 120, pode-se usar os valores críticos da distribuição normal padronizada...

# GRAUS DE LIBERDADE (n-k-1) MAIORES QUE 120

## 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.

$$gl = n-k-1 = 2725-4-1 = 2720$$

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

Source	SS	df	MS
Model	84.8242895	4	21.2060724
Residual	1925.52287	2720	.707912819
Total	2010.34716	2724	.738012906

Number of obs	=	2725
F( 4, 2720)	=	29.96
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

# EXEMPLO DO “WORLD VALUES SURVEY”

## Variável dependente:

\*Índice tradicional/secular (tradrat5)

## Variável independente

- \* Homem (x001): indicador de sexo masculino.
- \* Escolaridade (x025r): (1) baixa; (2) média; (3) alta.
- \* Estado civil (x007): (1) casado; (2) separado; (3) solteiro.
- \* Religião é muito importante (a006): (0) não; (1) sim.
- \* Acredita no céu (f054): (0) não; (1) sim.
- \* Objetivo é de fazer pais orgulhosos (d054): (1) concorda muito; (2) concorda; (3) discorda; (4) discorda muito.
- \* Acredita no inferno (f053): (0) não; (1) sim.
- \* Tempo com pessoas da igreja (a060): (1) semanalmente; (2) 1 ou 2 vezes por semana; (3) algumas vezes por ano; (4) nunca.

# GRAUS DE LIBERDADE (n-k-1) MAIORES QUE 120

$$gl = n - k - 1 = 17.245 - 14 - 1 = 17230$$

```
. xi: reg tradrat5 homem i.educ i.estciv religiao ceu i.pais inferno i.igreja
i.educ          _Ieduc_1-3      (naturally coded; _Ieduc_1 omitted)
i.estciv        _Iestciv_1-3    (naturally coded; _Iestciv_1 omitted)
i.pais          _Ipais_1-4      (naturally coded; _Ipais_1 omitted)
i.igreja        _Iigreja_1-4    (naturally coded; _Iigreja_1 omitted)
```

Source	SS	df	MS
Model	2919.01365	14	208.500975
Residual	12914.5224	17230	.749536992
Total	15833.536	17244	.918205522

Number of obs = 17245

F( 14, 17230) = 278.17  
 Prob > F = 0.0000  
 R-squared = 0.1844  
 Adj R-squared = 0.1837  
 Root MSE = .86576

tradrat5	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
homem	-.0915848	.0134579	-6.81	0.000	-.1179637	-.0652059
_Ieduc_2	.1611334	.0160693	10.03	0.000	.1296359	.1926309
_Ieduc_3	.4183285	.0182525	22.92	0.000	.3825517	.4541053
_Iestciv_2	.0823282	.0244348	3.37	0.001	.0344336	.1302229
_Iestciv_3	.033135	.0150337	2.20	0.028	.0036675	.0626025
religiao	-.2900597	.0163472	-17.74	0.000	-.322102	-.2580175
ceu	.1481911	.0246461	6.01	0.000	.0998822	.1965
_Ipais_2	.1559776	.0144202	10.82	0.000	.1277126	.1842426
_Ipais_3	.4766756	.024395	19.54	0.000	.428859	.5244922
_Ipais_4	.807771	.0485836	16.63	0.000	.7125423	.9029998
inferno	-.4142113	.0221687	-18.68	0.000	-.4576642	-.3707584
_Iigreja_2	.034723	.0200903	1.73	0.084	-.0046561	.0741021
_Iigreja_3	-.012683	.0211482	-0.60	0.549	-.0541357	.0287697
_Iigreja_4	.1743971	.0189101	9.22	0.000	.1373314	.2114628
_cons	.255215	.0276124	9.24	0.000	.2010918	.3093382

# REGRA DE REJEIÇÃO DE $H_0$ (UNILATERAL)

Figura 4.2

Regra de rejeição a 5% para a hipótese alternativa  $H_1: \beta_j > 0$  com 28 gl.

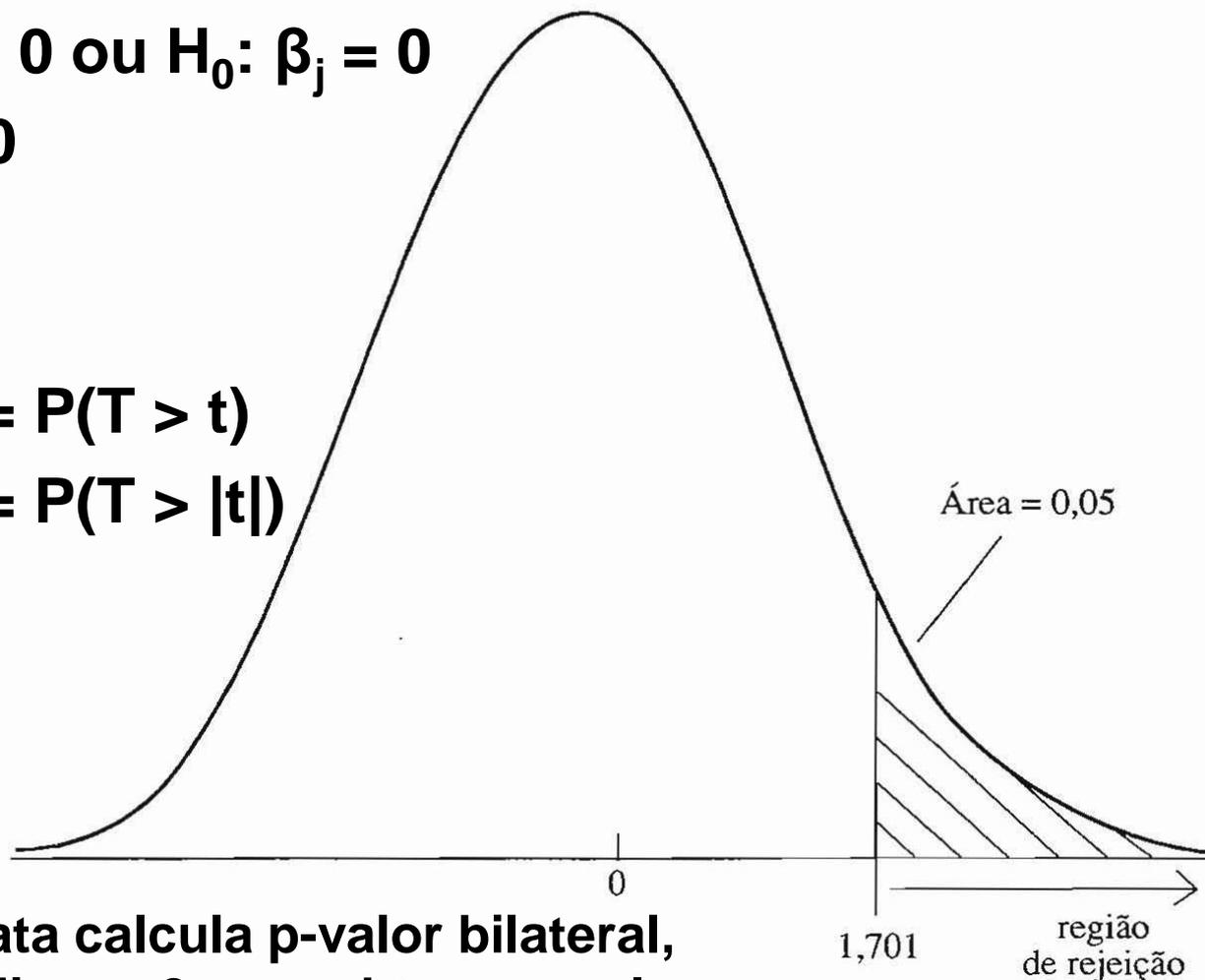
$$H_0: \beta_j \leq 0 \text{ ou } H_0: \beta_j = 0$$

$$H_1: \beta_j > 0$$

$$t_{\beta_j} > c$$

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

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



Como Stata calcula p-valor bilateral, é só dividir por 2 para obter o p-valor unilateral.

# REGRA DE REJEIÇÃO DE $H_0$ (UNILATERAL)

Figura 4.3

Regra de rejeição a 5% para a hipótese alternativa  $H_1: (\beta_j) < 0$ , com 18 gl.

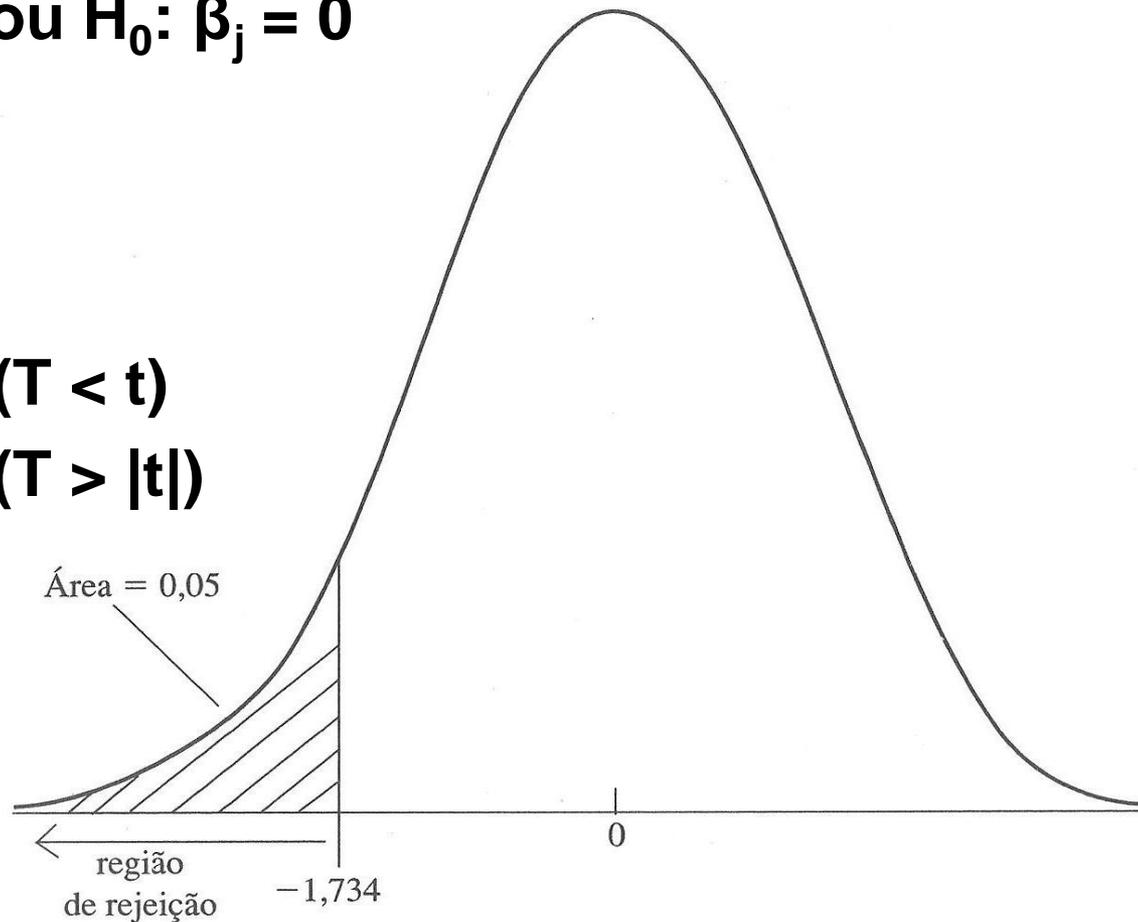
$$H_0: \beta_j \geq 0 \text{ ou } H_0: \beta_j = 0$$

$$H_1: \beta_j < 0$$

$$t_{\beta_j} < -c$$

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

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



Como Stata calcula p-valor bilateral, é só dividir por 2 para obter o p-valor unilateral.

# TESTE: HIPÓTESES ALTERNATIVAS BILATERAIS

$$H_1: \beta_j \neq 0$$

- Essa hipótese é relevante quando o sinal de  $\beta_j$  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 bi-caudal,  $c$  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

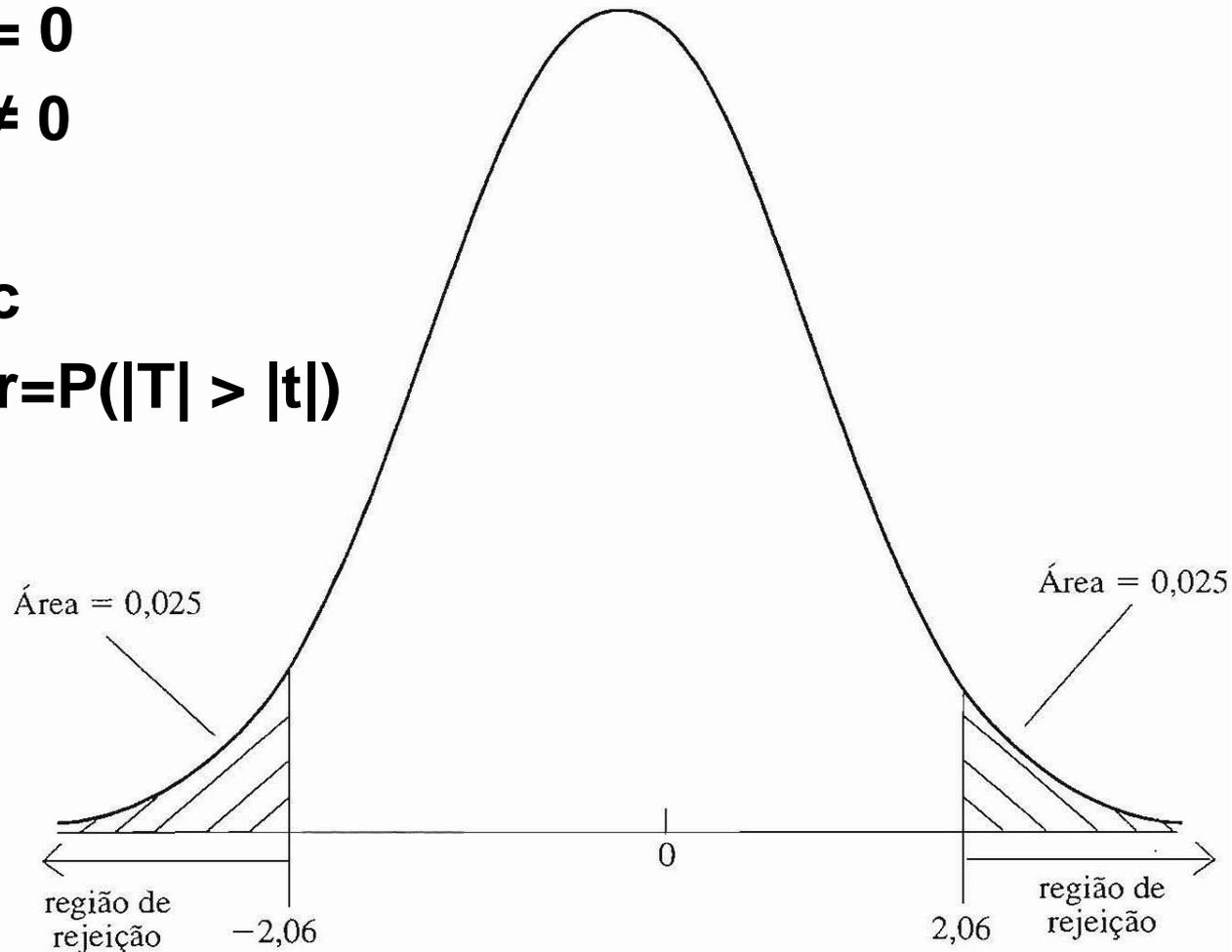
Regra de rejeição a 5% para a hipótese alternativa  $H_1: \beta_j \neq 0$  com 25 gl.

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

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

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



# EXEMPLO DE NÃO-REJEIÇÃO DE $H_0$ (BILATERAL)

Figura 4.6

Obtendo o  $p$ -valor contra uma alternativa bilateral, quando  $t = 1,85$  e  $gl = 40$ .

**p-valor**

$$= P(|T| > |t|)$$

$$= P(|T| > 1,85)$$

$$= 2P(T > 1,85)$$

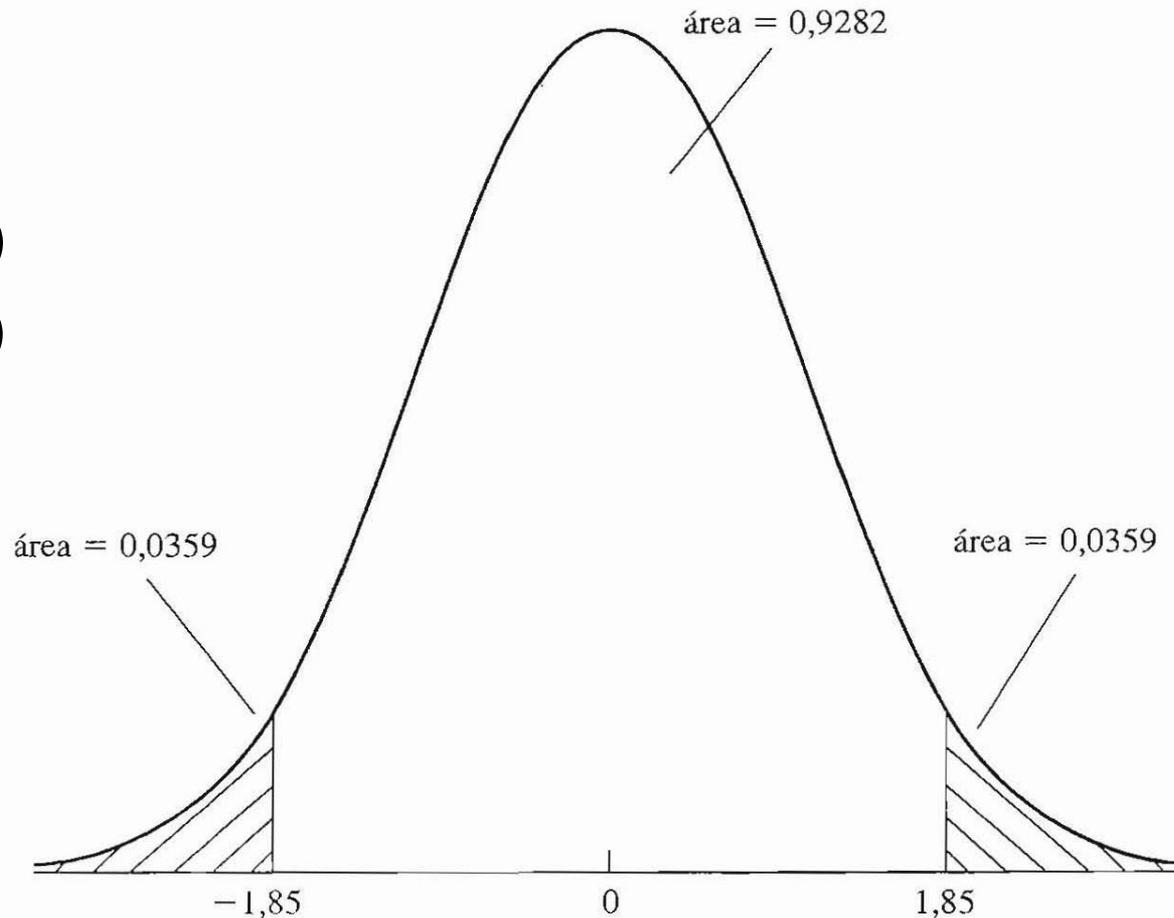
$$= 2(0,0359)$$

$$= 0,0718$$

**p-valor  $> \alpha$**

$$0,0718 > 0,05$$

**$H_0 : \beta_j = 0$  não é rejeitada**



## TESTES DE OUTRAS HIPÓTESES SOBRE $\beta_j$

- Poderíamos supor que uma variável dependente (log do número de crimes) necessariamente será relacionada positivamente com uma variável independente (log do número de estudantes matriculados na universidade).
- A hipótese alternativa testará se o aumento de 1% nas matrículas aumentará o crime em mais de 1%:

$$H_0: \beta_j = 1$$

$$H_1: \beta_j > 1$$

- $t = (\text{estimativa} - \text{valor hipotético}) / (\text{erro-padrão})$
- Neste exemplo,  $t = (\beta_j - 1) / \text{ep}(\beta_j)$
- Observe que adicionar 1 na hipótese nula, significa subtrair 1 no teste  $t$ .
- Rejeitamos  $H_0$  se  $t > c$ , em que  $c$  é o valor crítico unilateral.

# 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).

## INTERVALOS DE CONFIANÇA

- Os intervalos de confiança (IC), ou estimativas de intervalo, permitem avaliar uma extensão dos valores prováveis do parâmetro populacional, e não somente estimativa pontual:
  - Valor inferior:  $\beta_j - c \cdot ep(\beta_j)$
  - Valor superior:  $\beta_j + c \cdot ep(\beta_j)$
- A constante  $c$  é o 97,5º percentil de uma distribuição  $t_{n-k-1}$ .
- Quando  $n-k-1 > 120$ , podemos usar a distribuição normal para construir um IC de 95% ( $c=1,96$ ).
- Se amostras aleatórias fossem repetidas, então valor populacional estaria dentro do IC em 95% das amostras.
- Esperamos ter uma amostra que seja uma das 95% de todas amostras em que estimativa de intervalo contém beta.
- Se a hipótese nula for  $H_0: \beta_j = a_j$ ,  $H_0$  é rejeitada contra  $H_1: \beta_j \neq a_j$ , ao nível de significância de 5%, se  $a_j$  não está no IC.

## 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  $y$  ( $\beta_3$ ,  $\beta_4$  e  $\beta_5$ , por exemplo), já que outro conjunto de variáveis foi controlado ( $\beta_1$  e  $\beta_2$ , 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  $\beta_0$ ,  $\beta_1$  e  $\beta_2$ .
- Modelo irrestrito terá  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ,  $\beta_4$  e  $\beta_5$ .
- 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:  $\beta_3$ ,  $\beta_4$  e  $\beta_5$ ), ou seja,  $q = gl_r - gl_{ir}$ .

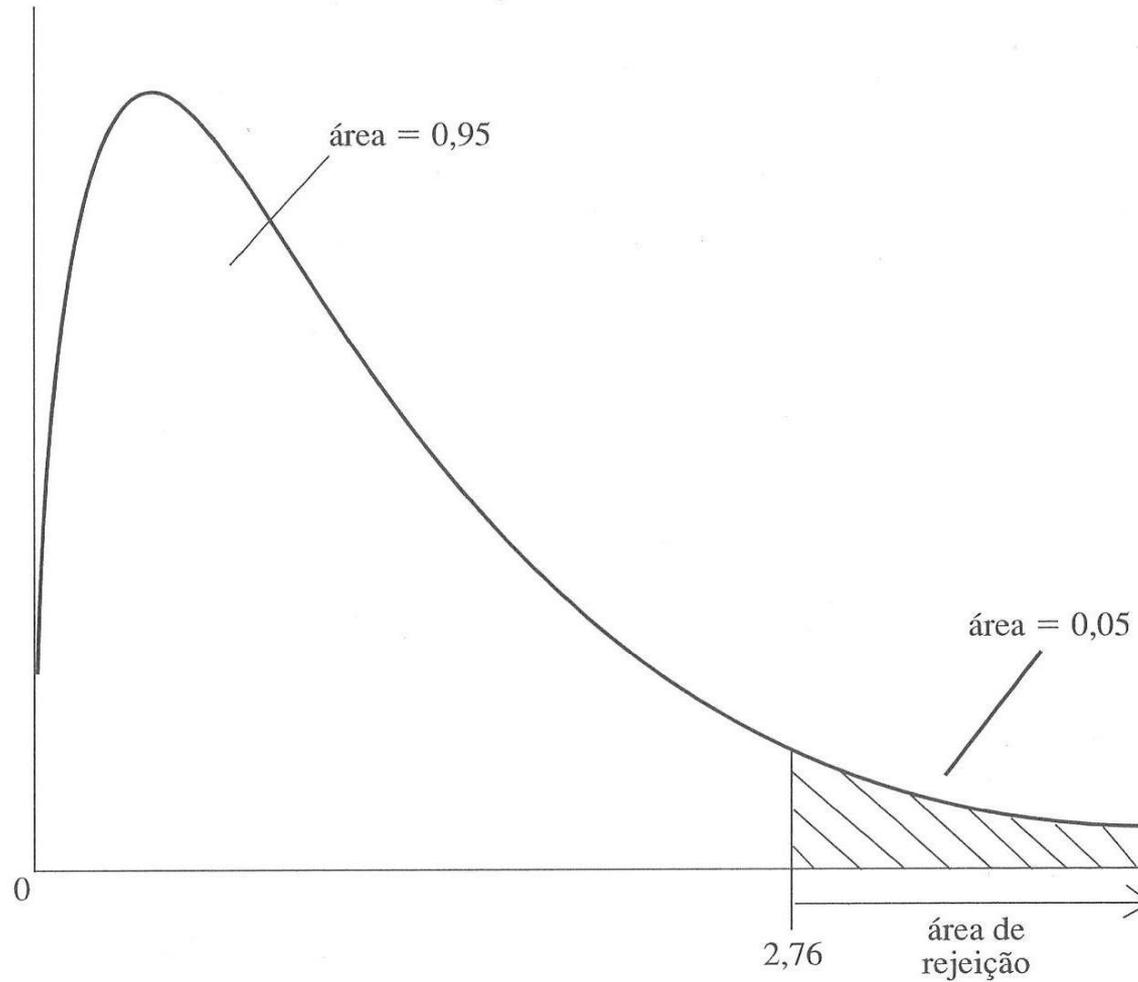
## REGRAS DE REJEIÇÃO DE $F$

- O valor crítico ( $c$ ) depende de:
  - Nível de significância (10%, 5% ou 1%, por exemplo).
  - Graus de liberdade do numerador ( $q = gl_r - gl_{ir}$ ).
  - Graus de liberdade do denominador ( $n - k - 1$ ).
  - Quando os  $gl$  do denominador chegam a 120, a distribuição  $F$  não é mais sensível a eles (usar  $gl = \infty$ ).
- Uma vez obtido  $c$ , rejeitamos  $H_0$ , em favor de  $H_1$ , ao nível de significância escolhido se:  $F > c$ .
- Se  $H_0$  ( $\beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ ) é rejeitada,  $\beta_3, \beta_4$  e  $\beta_5$  são **estatisticamente significantes conjuntamente**.
- Se  $H_0$  ( $\beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ ) não é rejeitada,  $\beta_3, \beta_4$  e  $\beta_5$  são **conjuntamente não significantes**.

# CURVA DA DISTRIBUIÇÃO F

Figura 4.7

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



## RELAÇÃO ENTRE ESTATÍSTICAS $F$ E $t$

- A estatística  $F$  para testar a exclusão de uma única variável é igual ao quadrado da estatística  $t$  correspondente.
- As duas abordagens levam ao mesmo resultado, desde que a hipótese alternativa seja bilateral.
- A estatística  $t$  é mais flexível para testar uma única hipótese, porque pode ser usada para testar alternativas unilaterais.
- As estatísticas  $t$  são mais fáceis de serem obtidas do que o teste  $F$ .

## FORMA R-QUADRADO DA ESTATÍSTICA $F$

- O teste  $F$  pode ser calculado usando os R-quadrados dos modelos resitrito e irrestrito.
- É mais fácil utilizar números entre zero e um ( $R^2$ ) do que números que podem ser muito grandes (SQR).
- Como  $SQR_r = SQT(1 - R_r^2)$ ,  $SQR_{ir} = SQT(1 - R_{ir}^2)$  e:

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

- ... os termos SQT são cancelados:

$$F \equiv \frac{(R_{ir}^2 - R_r^2)/q}{(1 - R_{ir}^2)/(n - k - 1)}$$

# CÁLCULO DOS $p$ -VALORES PARA TESTES $F$

$$p\text{-valor} = P(\mathcal{F} > F)$$

- O  $p$ -valor é a probabilidade de observarmos um valor de  $F$  pelo menos tão grande ( $\mathcal{F}$ ) quanto aquele valor real que encontramos ( $F$ ), dado que a hipótese nula é verdadeira.
- **Um  $p$ -valor pequeno é evidência para rejeitar  $H_0$** , porque a probabilidade de observarmos um valor de  $F$  tão grande quanto aquele para o qual a hipótese nula é verdadeira é muito baixa.
- **Um  $p$ -valor alto é evidência para NÃO rejeitar  $H_0$** , porque a probabilidade de observarmos um valor de  $F$  tão grande quanto aquele para o qual a hipótese nula é verdadeira é muito alta.

# TESTE $F$ PARA SIGNIFICÂNCIA GERAL DA REGRESSÃO

- No modelo com  $k$  variáveis independentes, podemos escrever a hipótese nula como:
  - $H_0: x_1, x_2, \dots, x_k$  não ajudam a explicar  $y$ .
  - $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ .
- Modelo restrito:  $y = \beta_0 + u$ .
- Modelo irrestrito:  $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$ .
- Número de variáveis independentes retiradas ( $q =$  graus de liberdade do numerador) é igual ao próprio número de variáveis independentes ( $k$ ):

$$F \equiv \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

- Mesmo com  $R^2$  pequeno, podemos ter teste  $F$  significativo para o conjunto, por isso não podemos olhar somente o  $R^2$ .

# 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.