AULAS 24 E 25
Análise de Regressão Múltipla: Inferência

Ernesto F. L. Amaral

06 e 08 de novembro de 2012
Metodologia de Pesquisa (DCP 854B)

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)

EXEMPLOS DE TRANSFORMAÇÕES

– **Variável dependente em alguns modelos:**

– Lineares (MQO): deve ter nível de mensuração de razão (contínua) e distribuição normal (logaritmo reduz concentração à esquerda de variáveis com valores positivos diferentes de zero).

– Logísticos e probit: variável dicotômica.

– Multinomiais: variável nominal com mais de duas categorias.

– Poisson: variável é contagem com concentração em zero.

– **Variável independente:**

– Se for contínua, deve ter distribuição normal (logaritmo).

– Se for contínua, também podemos calcular o quadrado (x^2) e incluir junto com variável independente original (x).

– Se for categórica, buscamos distribuição uniforme entre categorias, mas nem sempre é possível (categoria de referência geralmente possui vários casos).

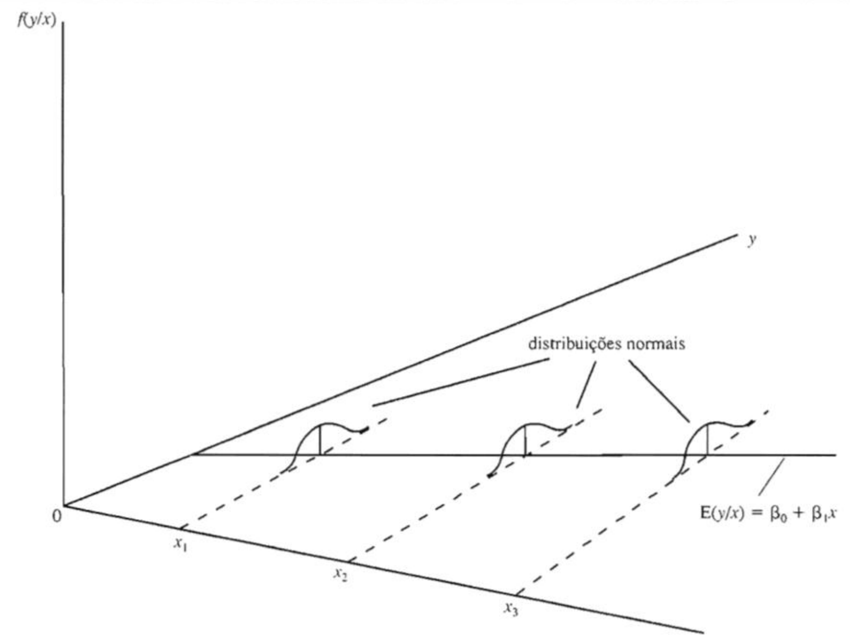
MODELO LINEAR CLÁSSICO

– 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.



Fonte: Wooldridge, 2008: 111.

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, ..., 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(β_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*-valor < α.

– H_0 não é rejeitada ao nível de 100*α%.

TESTE: HIPÓTESES ALTERNATIVAS UNILATERAIS

H_1: $\boldsymbol{\beta_j} > 0$ OU H_1: $\boldsymbol{\beta_j} < 0$

– 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 95º 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...

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 mês; (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*.

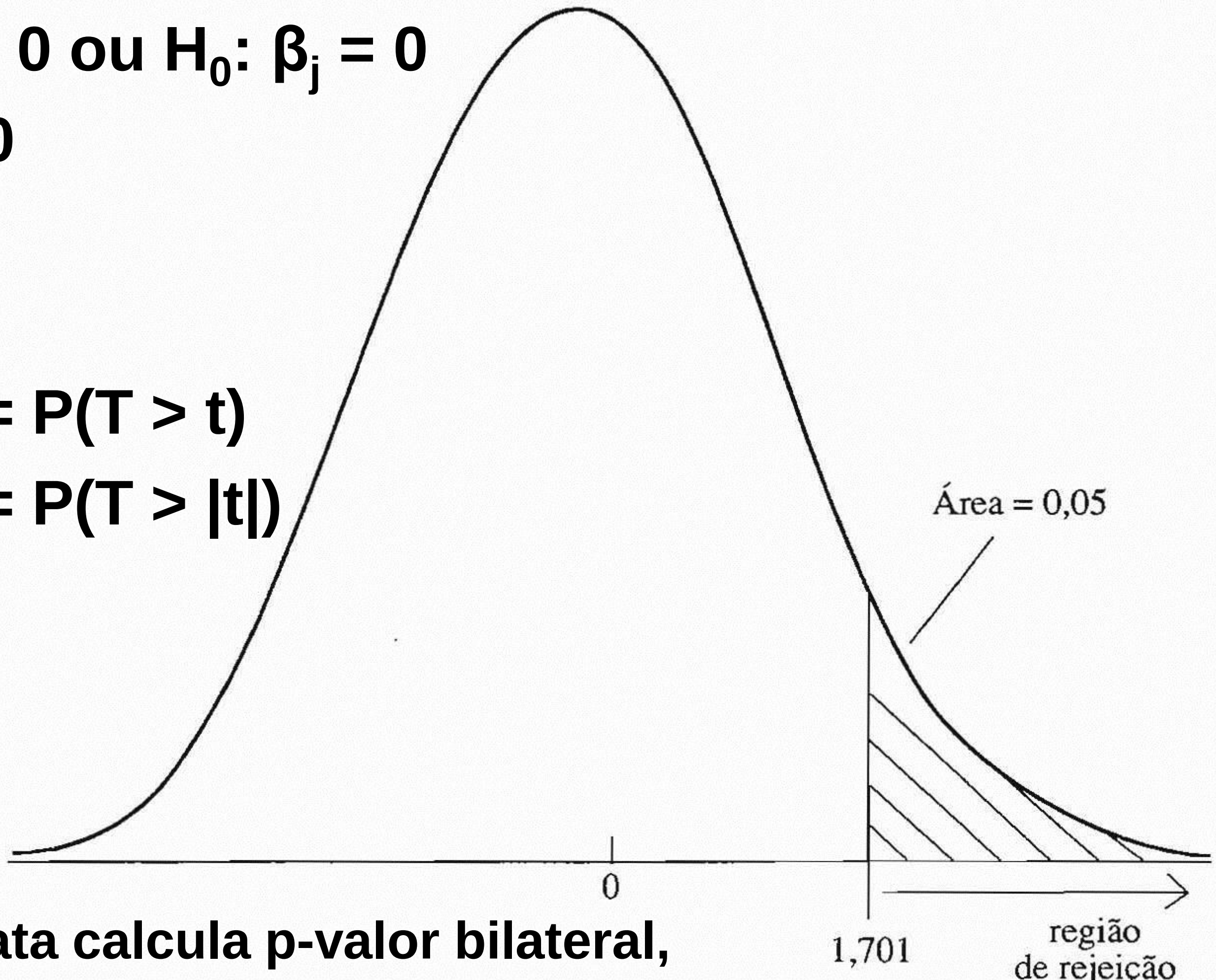
H_0: β_j <= 0 ou H_0: β_j = 0

H_1: β_j > 0

$t_{\beta j} > c$

p-valor = P(T > t)

p-valor = P(T > |t|)



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

Fonte: Wooldridge, 2008: 117.

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*.

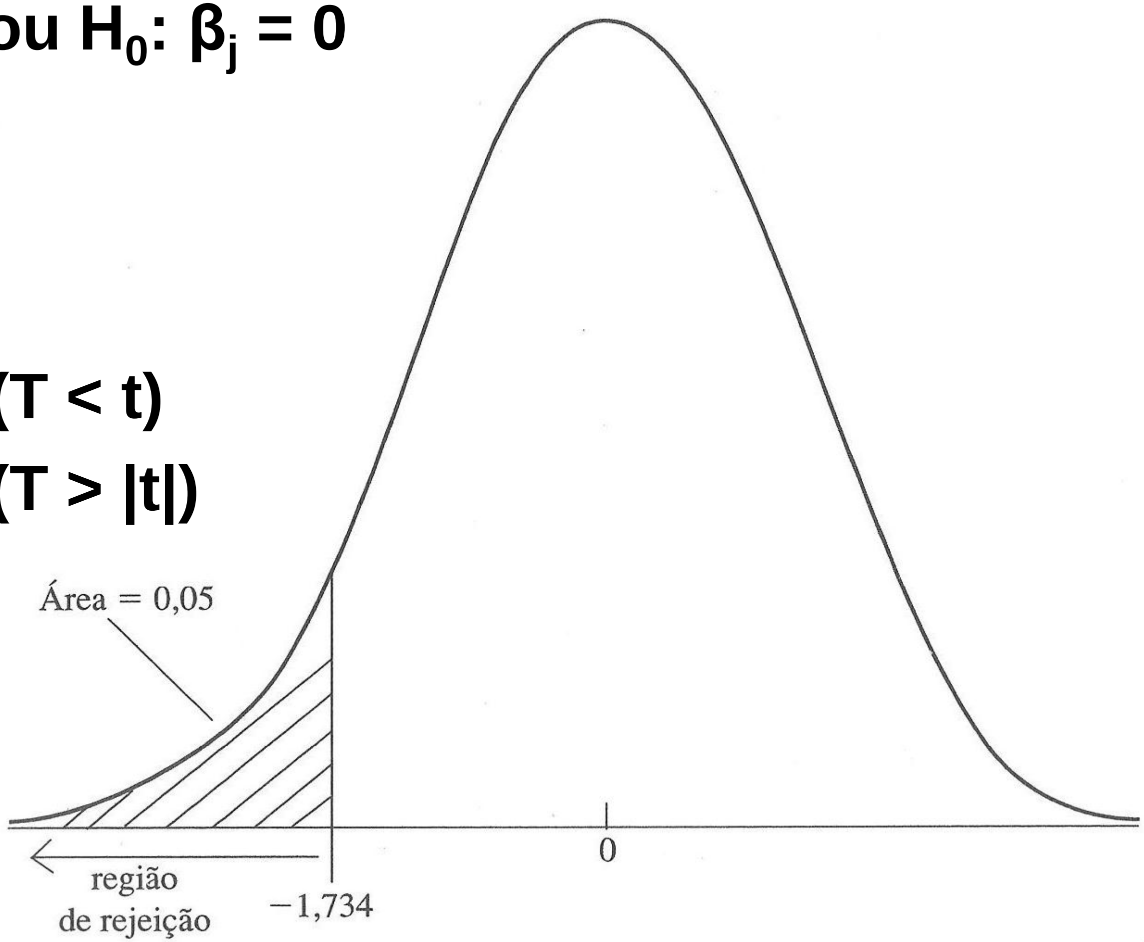
H_0: β_j >= 0 ou H_0: β_j = 0

H_1: β_j < 0

$t_{\beta j}$ < - c

p-valor = P(T < t)

p-valor = P(T > |t|)



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

Fonte: Wooldridge, 2008: 119.

TESTE: HIPÓTESES ALTERNATIVAS BILATERAIS

H_1: $\boldsymbol{\beta_j \neq 0}$

– 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 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 significante (ou estatisticamente diferente de zero) ao nível de 5%.

REGRA DE REJEIÇÃO DE H_0 (BILATERAL)

Figura 4.4

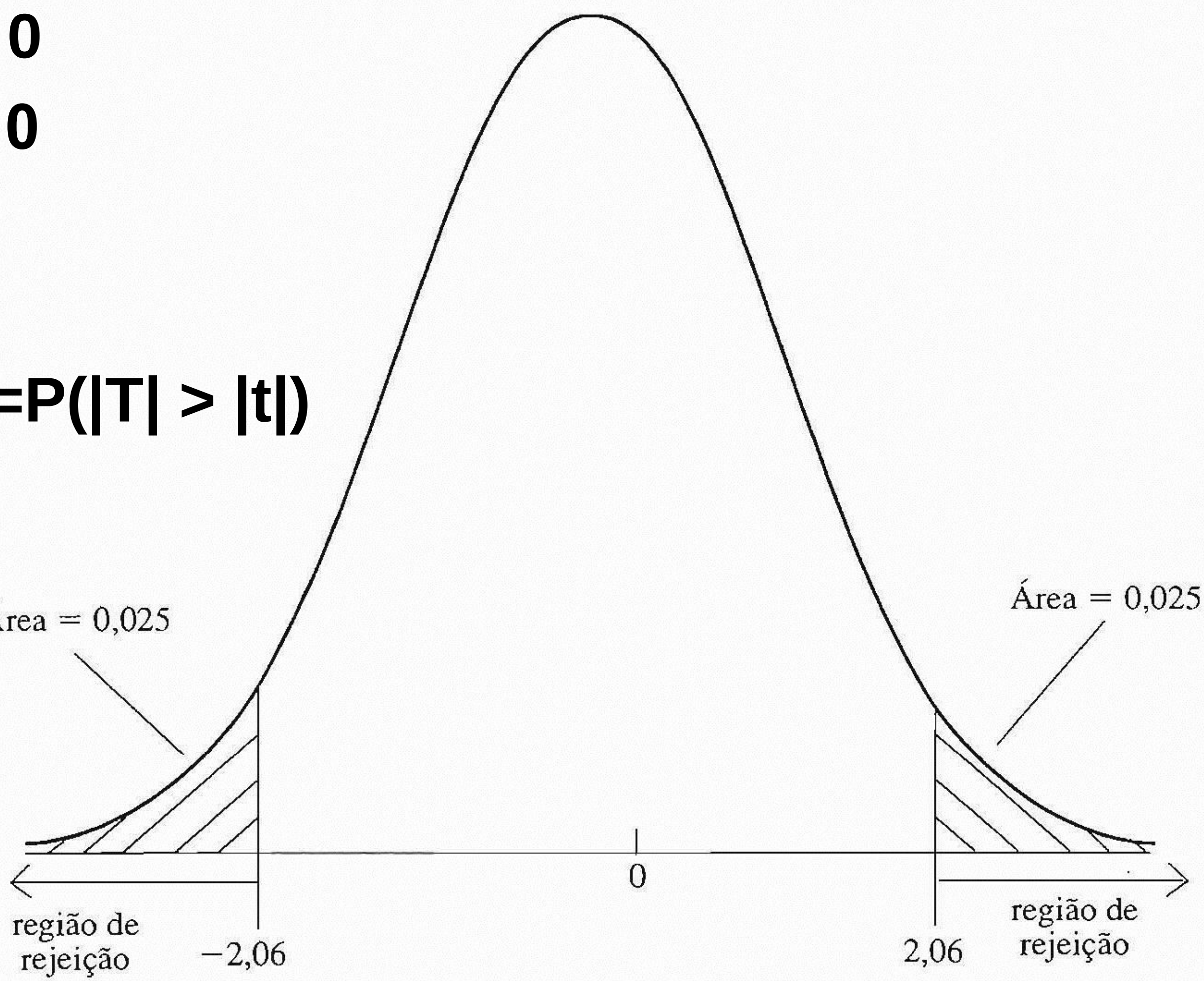
Regra de rejeição a 5% para a hipótese alternativa H_1: $\beta_j \neq 0$ com 25 *gl*.

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

$|t_{\beta j}| > c$

p-valor=P(|T| > |t|)



Fonte: Wooldridge, 2008: 122.

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

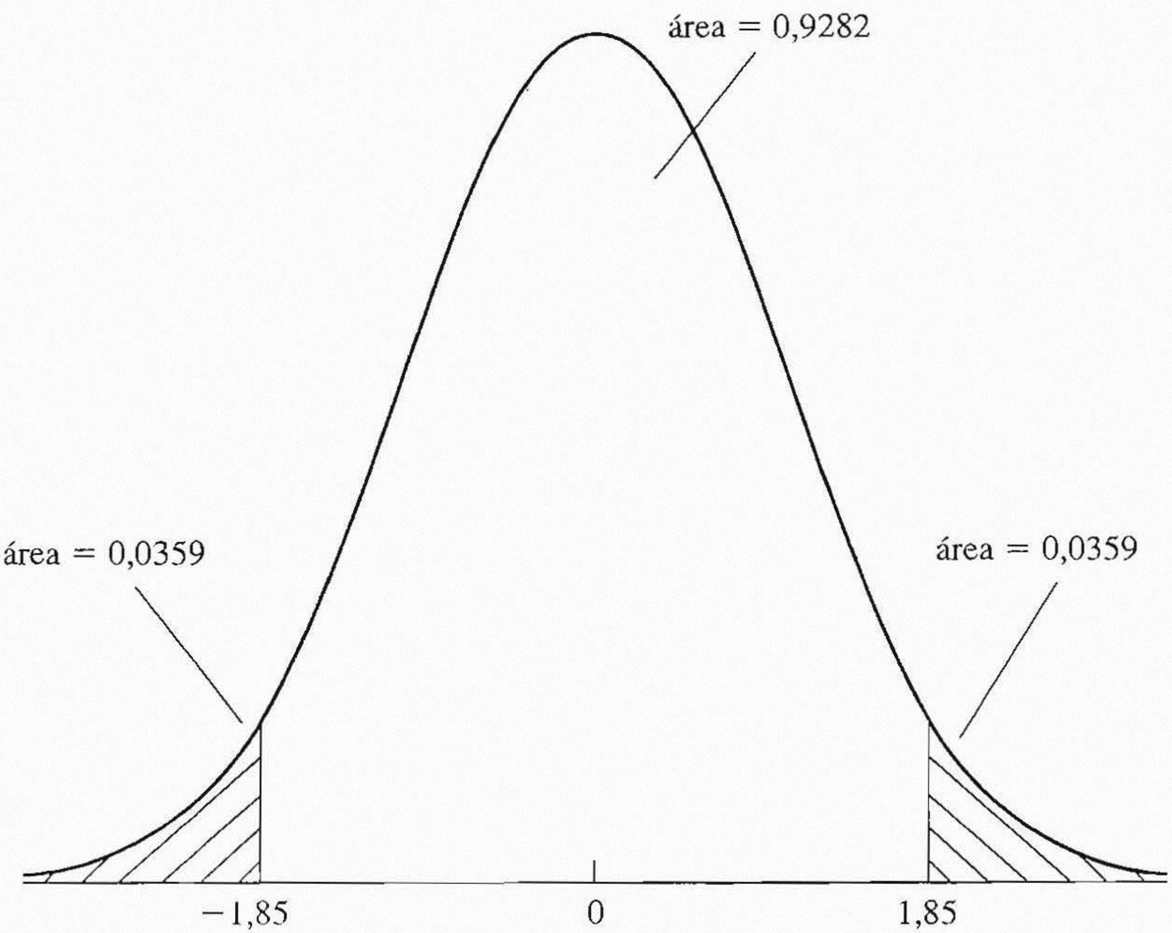
= P(|T| > |t|)

= P(|T| > 1,85)

= 2P(T > 1,85)

= 2(0,0359)

= 0,0718



p-valor > α

0,0718 > 0,05

H_0 : β_j=0 não é rejeitada

Fonte: Wooldridge, 2008: 127.

TESTES DE OUTRAS HIPÓTESES SOBRE β_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 = (estimativa - valor hipotético) / (erro-padrão)

– Neste exemplo, $t = (\beta_j - 1) / ep(\beta_j)$

– 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.

VARIÂNCIAS AMOSTRAIS DOS ESTIMADORES β_j (p.91-95)

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SQT_j(1 - R_j^2)} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}$$

– j = 1, 2, ..., k (cada variável independente do modelo).

– σ^2 é a variância do erro populacional.

– SQT_j é a variação amostral total em x_j.

– R_j^2 é o R-quadrado da regressão de cada x_j sobre todas as outras independentes.

– Vamos discutir cada termo desta fórmula...

VARIÂNCIA DO ERRO (σ^2)

– Uma variância do erro maior significa variâncias maiores dos estimadores de MQO.

– Um maior erro populacional torna mais difícil estimar o efeito parcial das variáveis independentes sobre y, o que é refletido em maiores variâncias dos estimadores de inclinação.

– Visto que σ^2 é uma característica da população, ela não tem nada a ver com o tamanho da amostra.

– Este é o componente desconhecido da fórmula da variância amostral do estimador de inclinação de MQO.

– Há somente uma maneira de reduzir a variância do erro, que é adicionar mais variáveis explicativas à equação (retirar alguns fatores do termo erro).

VARIÂNCIA AMOSTRAL TOTAL EM x_j (SQT_j)

– Mesmo não sendo possível escolher os valores amostrais das variáveis independentes, a variância amostral (SQT_j) aumenta ao aumentar o tamanho da amostra.

– Por outro lado, se temos uma amostra pequena, não haverá muita variação de x_{ij} em relação à sua média.

– Portanto, SQT_j aumenta sem limite quando o tamanho da amostra aumenta, o que diminui o erro padrão do beta estimado, aumentando a significância estatística (teste *t*).

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SQT_j(1 - R_j^2)}$$

RELAÇÕES LINEARES ENTRE AS INDEPENDENTES (R_j^2)

– Este R-quadrado é distinto do R-quadrado da regressão de y sobre x_1, x_2, ..., x_k.

– O R_j^2 é obtido de uma regressão em que cada x_j desempenha o papel de variável dependente, tomando as demais variáveis como independentes.

– Quando R_j^2 cresce, a variância amostral do estimador de inclinação torna-se maior.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SQT_j(1 - R_j^2)}$$

– Ou seja, um grau elevado de relação linear entre as variáveis independentes pode levar a variâncias grandes dos estimadores de inclinação de MQO.

DIFERENTES VALORES DE R_j^2

– R_j^2 é a proporção da variação total de x_j que pode ser explicada pelas outras variáveis independentes.

– Para dados σ^2 e SQT_j, a menor variância estimada de β_j é obtida quando x_j tem correlação amostral zero com as demais variáveis independentes ($R_j^2=0$).

– Se $R_j^2=1$, x_j é uma **combinação linear perfeita** de algumas variáveis independentes da regressão.

– Quando R_j^2 está próximo de um, há correlação alta (mas não perfeita) entre duas ou mais variáveis independentes, o que denota **multicolinearidade** no modelo.

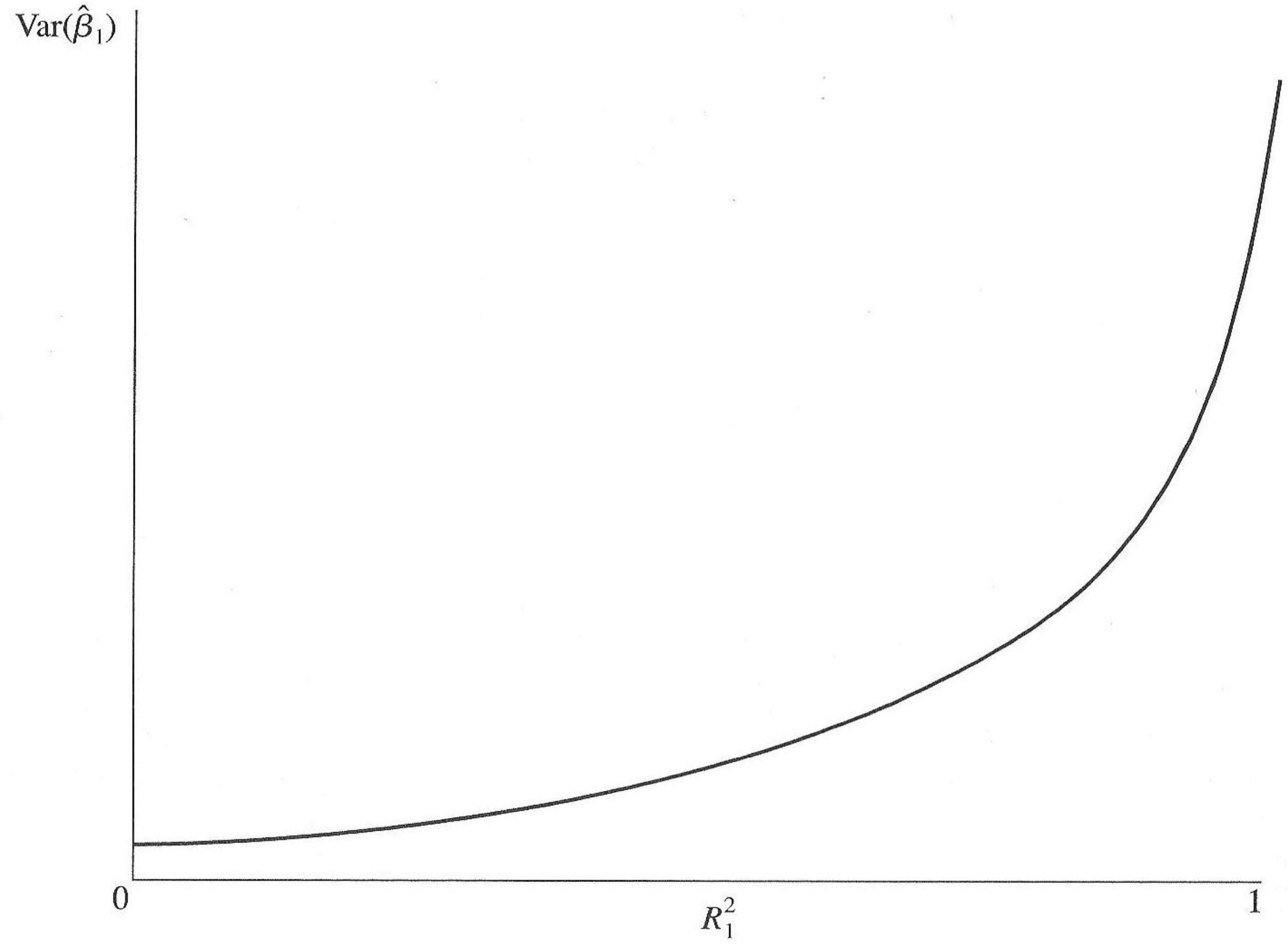
– Não há um valor definido de R_j^2 para concluir que a multicolinearidade é um problema em nosso modelo.

– O fato de R_j^2 ser alto, somente ocasionará uma variância grande do estimador de inclinação, em combinação com valores específicos de σ^2 e SQT_j.

VARIÂNCIA DO ESTIMADOR DE INCLINAÇÃO POR R_j^2

Figura 3.1

$\mathrm{Var}(\hat{\beta}_j)$ como uma função de R_j^2.



Fonte: Wooldridge, 2008: 94.

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 significante 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: β_j - c*ep(β_j)

  – Valor superior: β_j + c*ep(β_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:β_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 (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:

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

– SQRr é a soma dos resíduos quadrados do modelo restrito.

– SQRir 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, $q=gl_r-gl_{ir}$.

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=∞).

– Uma vez obtido *c*, rejeitamos H_0, em favor de H_1, ao nível de significância escolhido se: $F > c$.

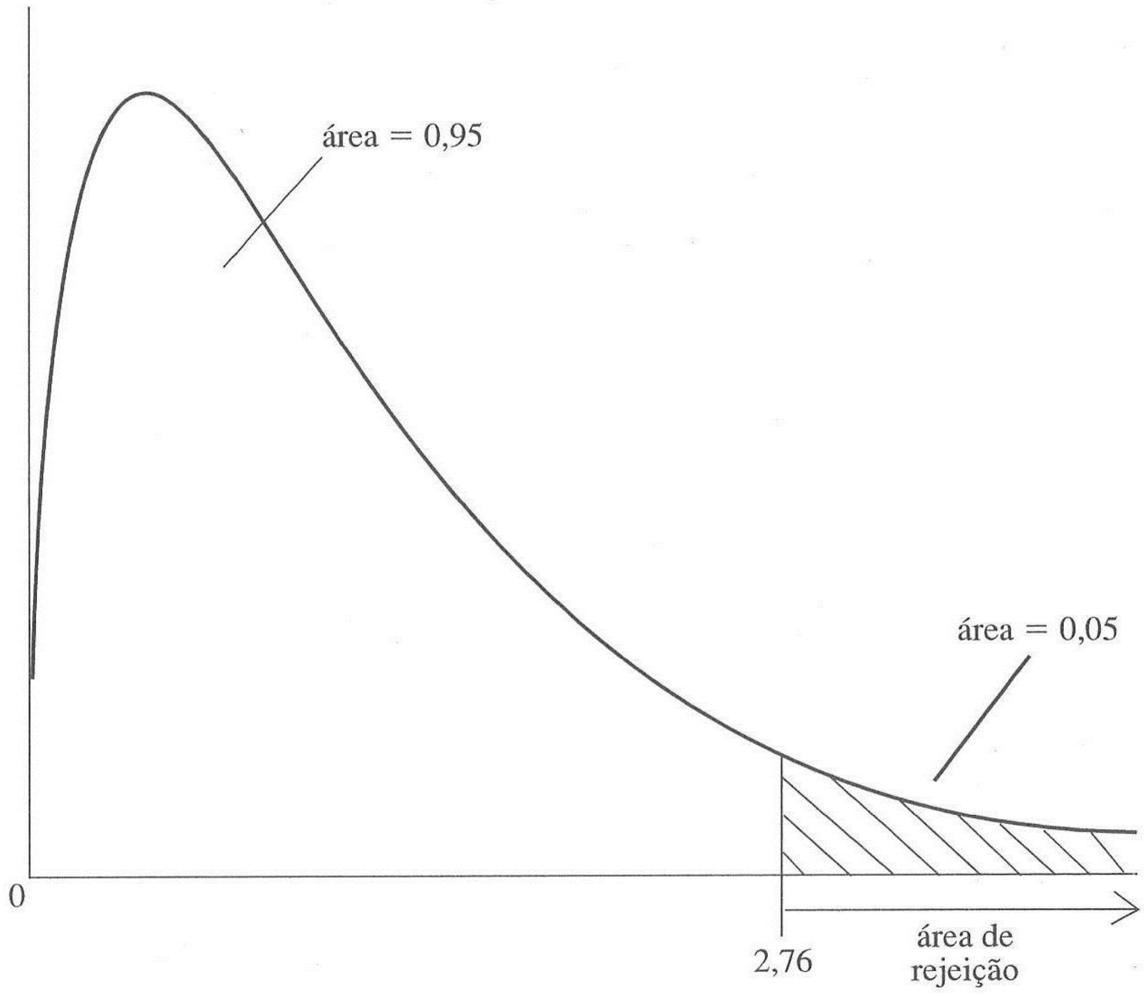
– Se H_0 ($\beta_3=0, \beta_4=0, \beta_5=0$) é rejeitada, β_3, β_4 e β_5 são **estatisticamente significantes conjuntamente**.

– Se H_0 ($\beta_3=0, \beta_4=0, \beta_5=0$) não é rejeitada, β_3, β_4 e β_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 $F_{3,60}$.



Fonte: Wooldridge, 2008: 142.

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:

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

CÁLCULO DOS *p*-VALORES PARA TESTES *F*

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

– O *p*-valor é a probabilidade de observarmos um valor de F pelo menos tão grande ($\mathscr{F}$) 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, ..., x_k não ajudam a explicar y.

  – H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$.

– Modelo restrito: $y = \beta_0 + u$.

– Modelo irrestrito: $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$.

– 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 significante 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 da 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.