AULAS 06, 07, 08 E 09
Análise de Regressão Múltipla: Estimação e Inferência

Ernesto F. L. Amaral

18, 23, 25 e 30 de março de 2010
Métodos Quantitativos de Avaliação de Políticas Públicas (DCP 030D)

Fonte:

Cohen, Ernesto, e Rolando Franco. 2000. "Avaliação de Projetos Sociais". São Paulo, SP: Editora Vozes. pp.118-136.

Wooldridge, Jeffrey M. "Introdução à econometria: uma abordagem moderna". São Paulo: Cengage Learning, 2008. Capítulos 3 e 4 (pp.64-157).

AUXILIANDO O EXERCÍCIO 1

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

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

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

```
Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
  salary |     209     1281.12    1372.345       223      14822
     roe |     209    17.18421    8.518509        .5       56.3
```

reg salary roe

n-k-1 = 209-1-1

=SS/df

```
      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  1,   207) =    2.77
       Model |  5166419.04     1  5166419.04           Prob > F      =  0.0978
    Residual |   386566563   207  1867471.32           R-squared     =  0.0132
-------------+------------------------------           Adj R-squared =  0.0084
       Total |   391732982   208  1883331.64           Root MSE      =  1366.6

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |   18.50119   11.12325     1.663   0.098    -3.428195    40.43057
       _cons |   963.1913   213.2403     4.517   0.000     542.7902    1383.592
------------------------------------------------------------------------------
```

R-squared = R^2 = SQE/SQT = 1 - SQR/SQT:

- É a proporção da variação em y explicada pelas variáveis independentes.
- É usado para calcular o teste F (significância conjunta das variáveis independentes).
- Ao incluir variável independente, R^2 aumenta (SQR diminui).

Adj R-squared = $\bar{R}^2$ = 1 - (1 - R^2)(n - 1)/(n - k - 1):

- Ao incluir variável independente, R^2 ajustado pode aumentar ou diminuir: SQR diminui e k aumenta.
- Pode ser usado para escolher modelo que não tenha regressores redundantes.
- Pode ter valor negativo, indicando ajuste ruim para número de "df" (p.190-193).

Root MSE = Raiz Quadrada do Erro Quadrado Médio = Raiz(MS Residual)

- Unidade é a mesma da variável dependente (comparar com quadro descritivo).

CAPÍTULO 7 - COHEN & FRANCO
MODELOS PARA A AVALIAÇÃO DE IMPACTOS

AVALIAÇÃO DE IMPACTO DE POLÍTICAS

– Os métodos de estimação de impacto dependem do desenho da avaliação, isto é, se há dados para grupos de tratamento (beneficiários) e controle (comparação).

GRUPO	ANTES	POLÍTICA	DEPOIS
Tratamento	T_0	X	T_1
Controle	C_0		C_1

– “Diferença em diferenças” ou “dupla diferença” (DD) estima:

1) Diferença dentro de cada grupo (tratamento e controle).

2) Diferença dessas duas médias.

$$DD = (T_1 – T_0) – (C_1 – C_0)$$

DESENHOS EXPERIMENTAIS

– Atribuição aleatória, dentre grupos de indivíduos, da oportunidade de participar em programas de intervenção, definindo grupos de tratamento e controle:
  – Realização de pesquisa para averiguar as regiões pobres.
  – Seleção aleatória de regiões incluídas na política e daquelas que serão o controle.
  – Única diferença entre grupos é o ingresso no programa.
– Avaliação sistemática e mensuração dos resultados em distintos momentos da implementação do programa.
– Se a seleção é aleatória, pode-se dispensar a avaliação anterior à política para ambos os grupos.

	X	**T_1**
		C_1

DESENHOS QUASE-EXPERIMENTAIS

– O controle é construído com base na propensão do indivíduo de ingressar no programa.

– Busca-se obter grupo de comparação que corresponda ao grupo de beneficiários:

  – Com base em certas características (sociais, econômicas...) estima-se a probabilidade de um indivíduo de participar do programa.

  – Com base nessa propensão (exercício de emparelhamento), constitui-se o grupo de controle.

– Estima-se os efeitos na comparação entre o grupo de tratamento e o grupo de controle, antes e depois do programa.

T_0	**X**	**T_1**
C_0		**C_1**

DESENHOS NÃO-EXPERIMENTAIS

– Ausência de grupos de controle torna mais difícil isolar causas que geram impactos na variável de interesse.

– Pode ser realizada análise reflexiva para estimar efeitos dos programas, com comparação dos resultados obtidos pelos beneficiários antes e depois do programa.

– Modelo antes-depois:

T_0	X	T_1

– Modelo somente depois com grupo de comparação:

	X	T_1	T_2
		C_1	C_2

– Modelo somente depois:

	X	T_1	T_2

DESENHO DA AVALIAÇÃO	MÉTODO DE ESTIMAÇÃO DE IMPACTO
EXPERIMENTAL	COMPARAÇÃO DE MÉDIAS
QUASE-EXPERIMENTAL	REGRESSÃO MÚLTIPLA & DIFERENÇA EM DIFERENÇAS
NÃO-EXPERIMENTAL	REGRESSÃO MÚLTIPLA

CAPÍTULO 3 - WOOLDRIDGE
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

$$salárioh = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

– 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.

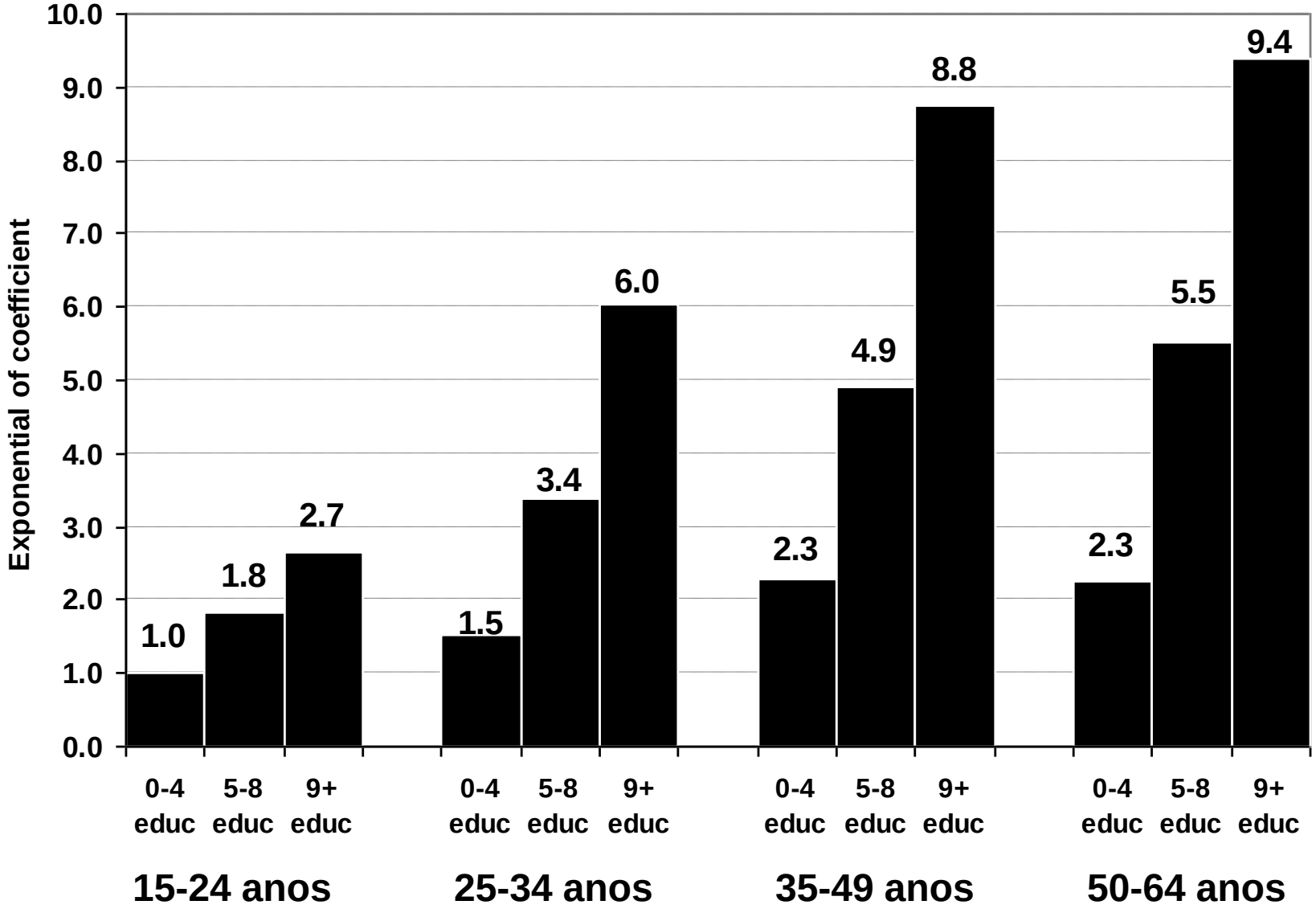
– Como experiência foi inserida na equação, podemos medir o efeito de educação sobre salário, mantendo experiência fixa.

– Na regressão simples, teríamos que assumir que experiência não é correlacionada com educação, o que é uma hipótese fraca.

APLICANDO A EQUAÇÃO DE MINCER (1974) NO BRASIL

Micro-região	Ano	Grupo Idade-Escol.	G11	G12	G13	...	G43	Log da Média da Renda
110006	1970	15-24 anos & 0-4 escol.	1	0	0	...	0	5.82
110006	1970	15-24 anos & 5-8 escol.	0	1	0	...	0	6.21
110006	1970	15-24 anos & 9+ escol.	0	0	1	...	0	6.75
...	...	...	...	...	...	...	...	...
110006	1970	50-64 anos & 9+ escol.	0	0	0	...	1	7.73
...	...	...	...	...	...	...	...	...

EFEITOS DE GRUPOS DE IDADE-ESCOLARIDADE NA RENDA DOS TRABALHADORES: BRASIL, 1970–2000



Fonte: Censos Demográficos Brasileiros 1970 a 2000 (IBGE).

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.

RELAÇÕES FUNCIONAIS ENTRE VARIÁVEIS

– A regressão múltipla é útil para generalizar relações funcionais entre variáveis.

– Por exemplo:

$$cons = \beta_0 + \beta_1 rend + \beta_2 rend^2 + u$$

– Variação no consumo decorrente de variação na renda é:

$$\frac{\Delta cons}{\Delta rend} \approx \beta_1 + 2\beta_2 rend$$

– O efeito marginal da renda sobre o consumo depende tanto de β_2 como de β_1 e do nível de renda.

– A definição das variáveis independentes é sempre importante na interpretação dos parâmetros.

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.

– No exemplo da renda ao quadrado, não é preciso incluir $rend^2$, já que ela é conhecida quando se conhece *rend*:

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

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:**

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

OBTENÇÃO DAS ESTIMATIVAS DE MQO

– Reta de regressão de MQO ou função de regressão amostral (FRA):

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

– O método de mínimos quadrados ordinários escolhe as estimativas que minimizam a soma dos resíduos quadrados.

– Dadas *n* observações de *y*, x_1, x_2, ... e x_k, as estimativas dos parâmetros são escolhidas para fazer com que a expressão abaixo tenha o menor valor possível:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik} \right)^2$$

INTERPRETAÇÃO DA EQUAÇÃO DE REGRESSÃO

– Novamente a reta de regressão de MQO:

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

– 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:

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

– 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:

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

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*.

COMPARAÇÃO DAS ESTIMATIVAS

– Relação entre parâmetros da regressão simples e múltipla.

– Tomemos como exemplo de regressão simples:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

– ... e de regressão múltipla:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

– Relação entre os β_1:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$

– δ_1: coeficiente de inclinação da regressão de x_{i2} sobre x_{i1}.

– Os parâmetros são iguais $(\tilde{\beta}_1 = \hat{\beta}_1)$, quando:

1) Efeito parcial de x_2 sobre y estimado é zero na amostra:

$$\hat{\beta}_2 = 0$$

2) x_1 e x_2 são não-correlacionados na amostra:

$$\tilde{\delta}_1 = 0$$

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*.

REGRESSÃO ATRAVÉS DA ORIGEM

– Em alguns modelos, pode-se avaliar que o ideal seria ter β_0 igual a zero:

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \cdots + \tilde{\beta}_k x_k$$

– R^2 pode ser negativo, o que significa que a média amostral de y "explica" mais da variação em y_i do que as variáveis independentes.

– Nesse caso, devemos incluir um intercepto ou procurar novas variáveis explicativas.

– Se β_0 for diferente de zero na população, a regressão através da origem gera estimadores dos parâmetros de inclinação (β_k) viesados.

– Se β_0 for igual a zero na população, a regressão com intercepto gera maiores variâncias dos estimadores de inclinação.

VALOR ESPERADOS DOS ESTIMADORES DE MQO

HIPÓTESE RLM.1 (LINEAR NOS PARÂMETROS)

– Modelo na população pode ser escrito como:

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

– β_0, β_1,..., β_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:

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

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**).

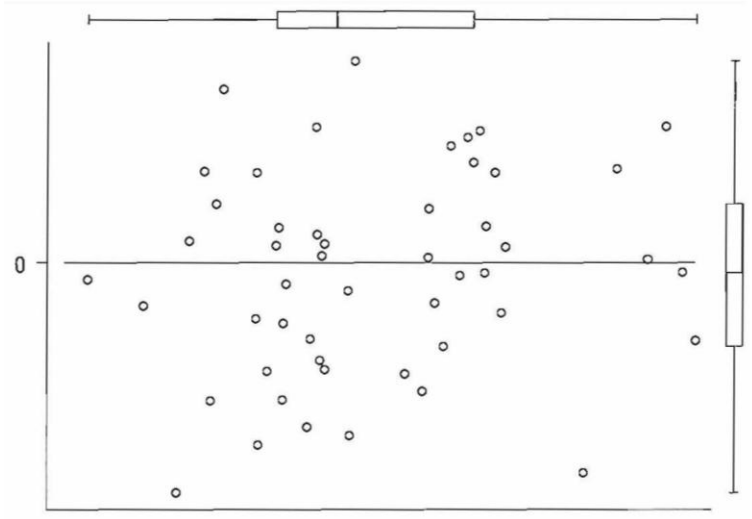
– Por outro lado, se a correlação for nula, não é necessário regressão múltipla, mas sim regressão simples, já que o termo de erro englobaria todos fatores não-observáveis e não-relacionados com as variáveis independentes.

HIPÓTESE RLM.5 (HOMOSCEDASTICIDADE)

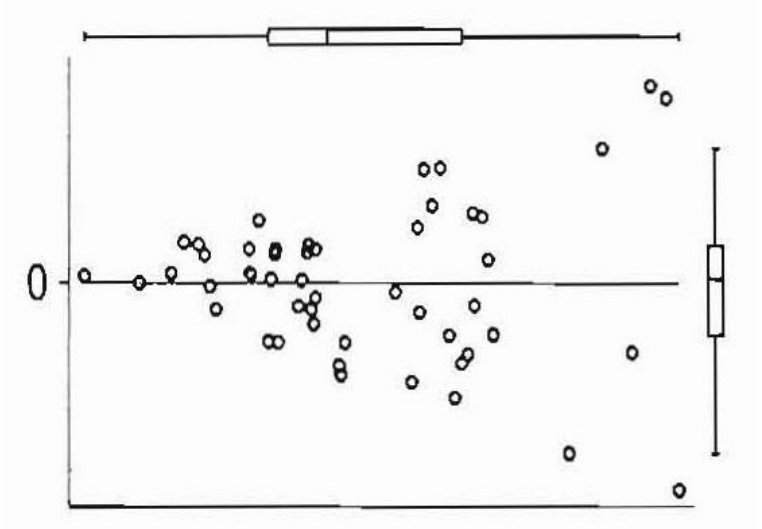
– A variância do termo erro (*u*), condicionada às variáveis explicativas, é a mesma para todas as combinações de resultados das variáveis explicativas.

– Se essa hipótese é violada, o modelo exibe heteroscedasticidade.

HOMOSCEDASTICIDADE



HETEROSCEDASTICIDADE



Fonte: Hamilton, 1992: 52-53.

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.

EXEMPLO DE TRANSFORMAÇÃO DA VARIÁVEL INDEPENDENTE

IMPACTO ECONÔMICO DA RELIGIÃO

– Unidade de análise: quatro grupos de idade (15-24; 25-34; 35-49; 50-64) e três grupos de escolaridade (0-4; 5-8; 9+) geram doze grupos de idade-escolaridade.

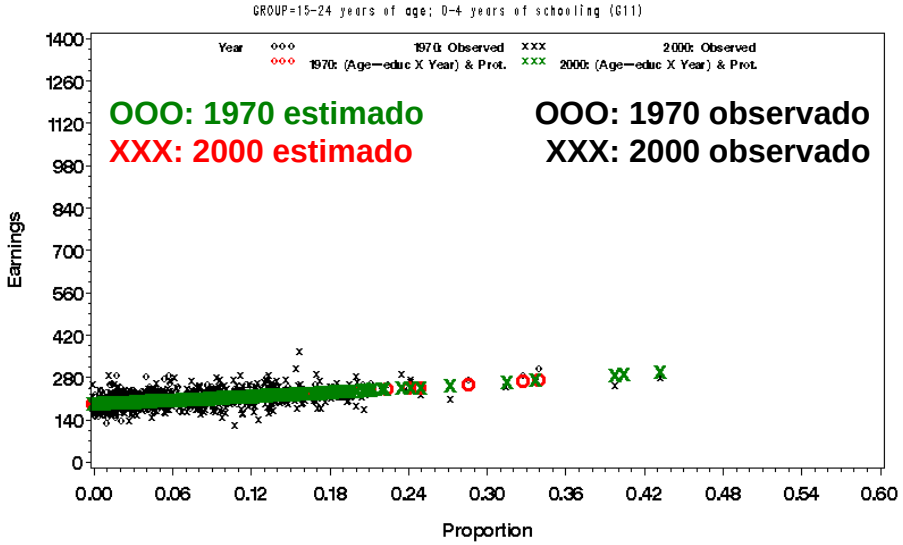
– Há informações para 502 microrregiões e quatro anos censitários (1970; 1980; 1991; 2000).

– Variável dependente: logaritmo da renda média do grupo de idade-escolaridade em cada microrregião e ano.

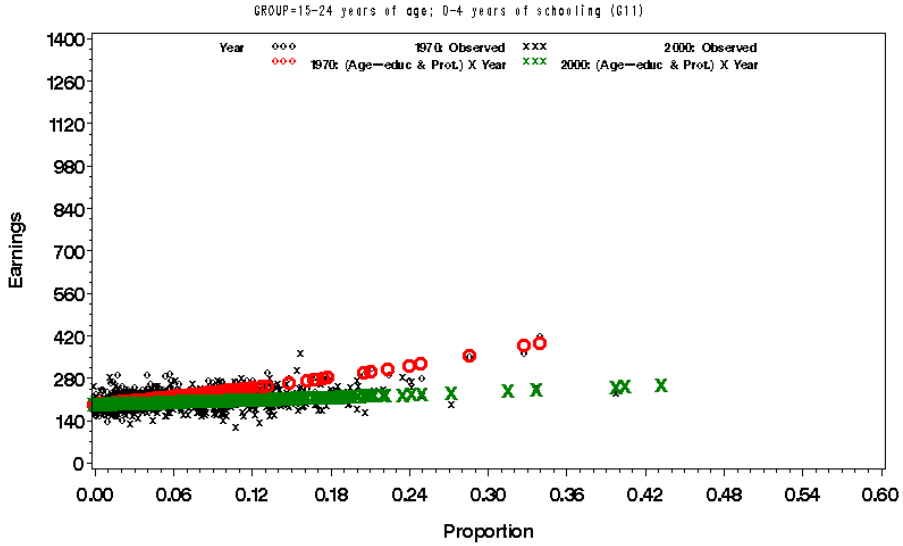
– Variáveis independentes: variáveis dicotômicas dos grupos de idade-escolaridade, proporção de protestantes em cada grupo de idade-escolaridade, efeitos fixos de microrregião e ano censitário.

IDADE 15-24 / ESCOLARIDADE 0-4

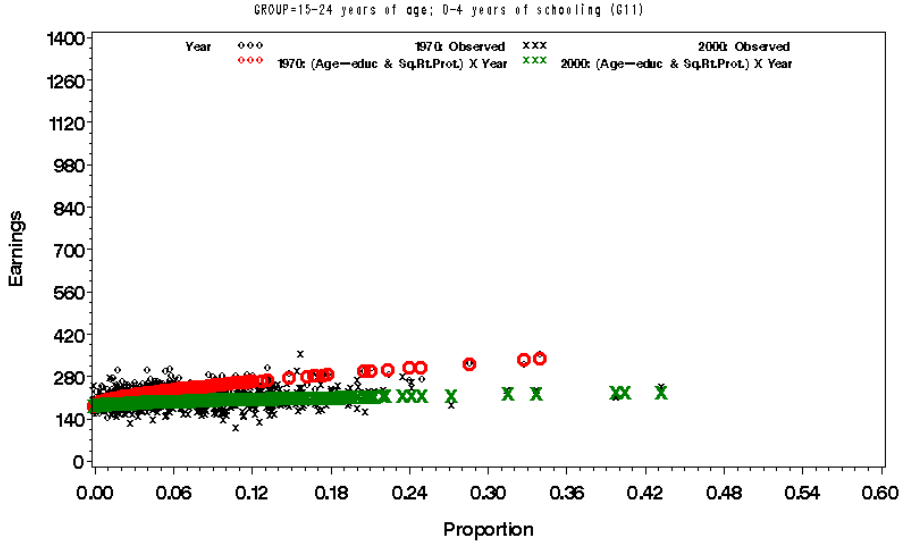
Prop. protestantes



Prop. protestantes * Ano

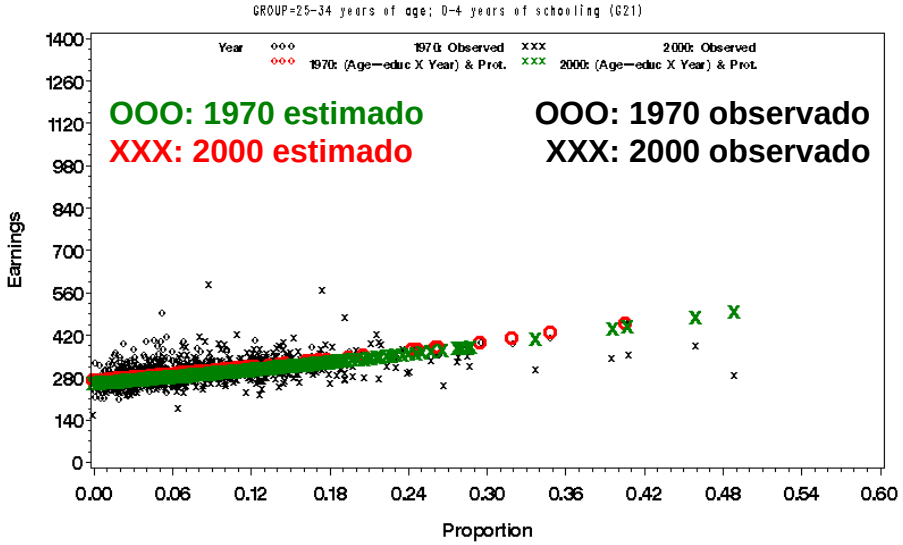


Raiz quadrada(Prop. protestantes) * Ano

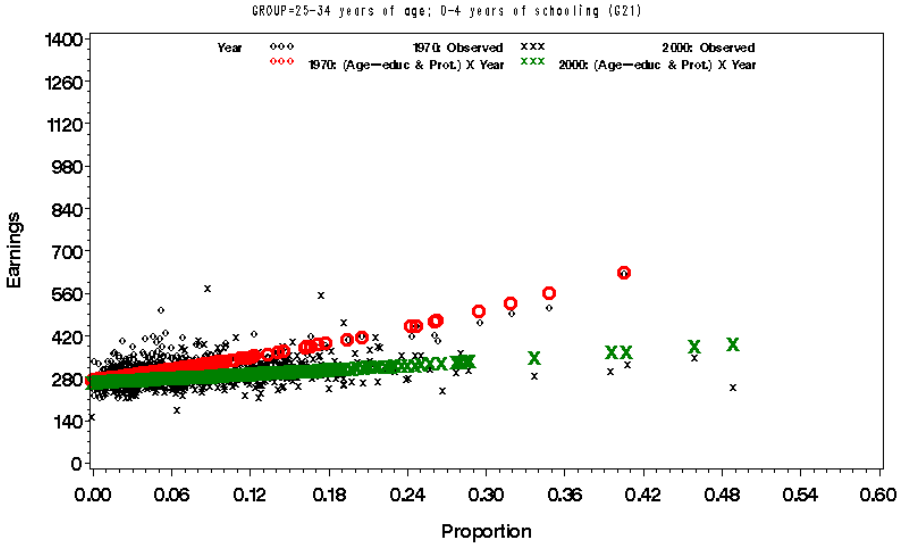


IDADE 25-34 / ESCOLARIDADE 0-4

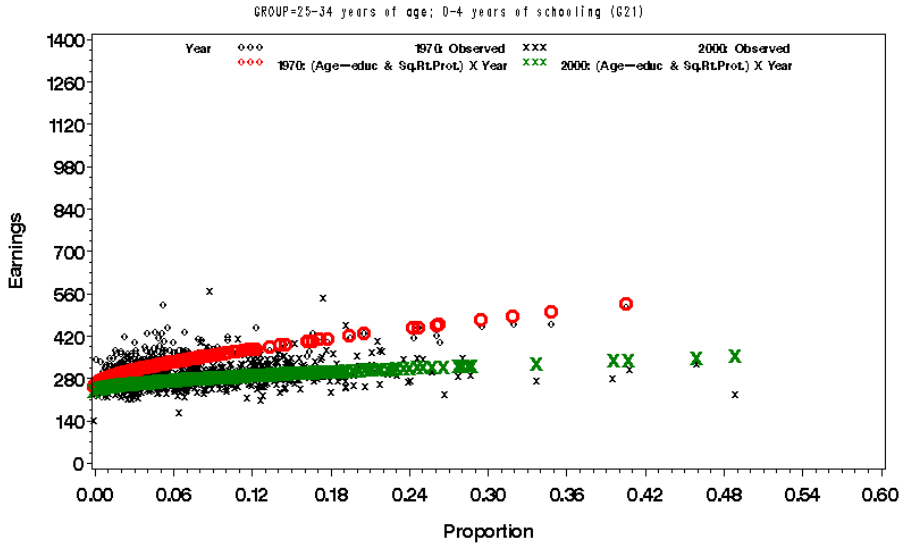
Prop. protestantes



Prop. protestantes * Ano

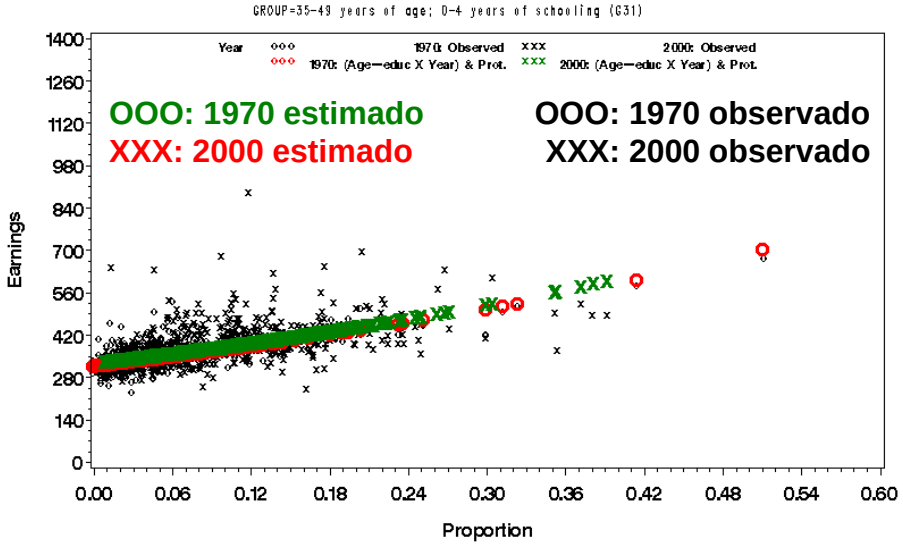


Raiz quadrada(Prop. protestantes) * Ano

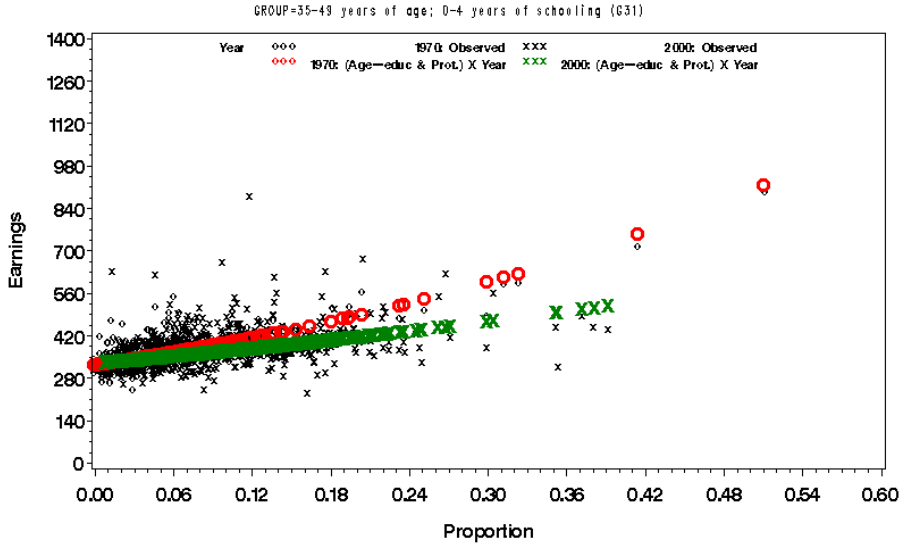


IDADE 35-49 / ESCOLARIDADE 0-4

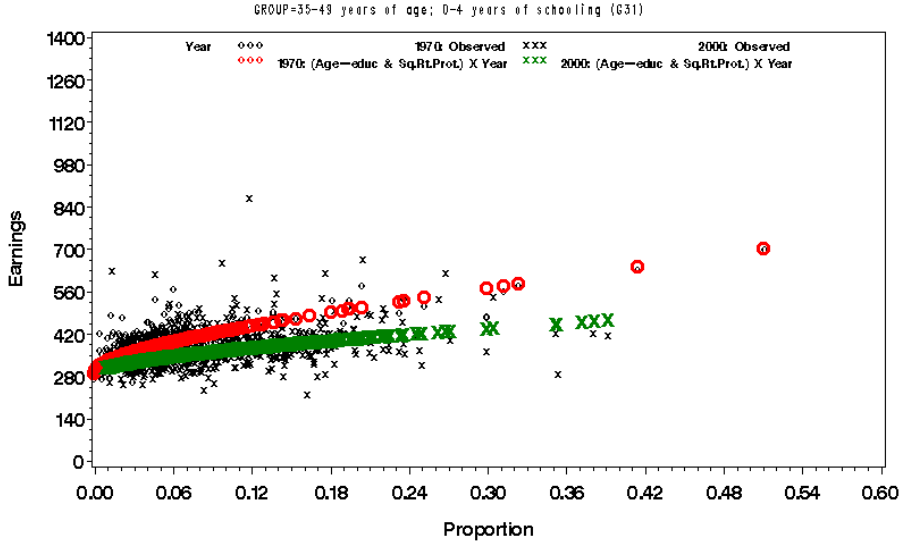
Prop. protestantes



Prop. protestantes * Ano

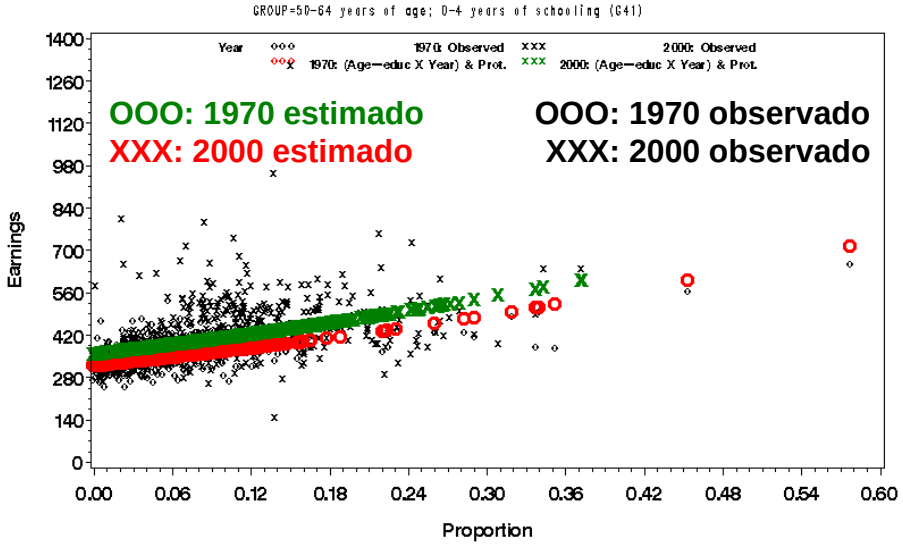


Raiz quadrada(Prop. protestantes) * Ano

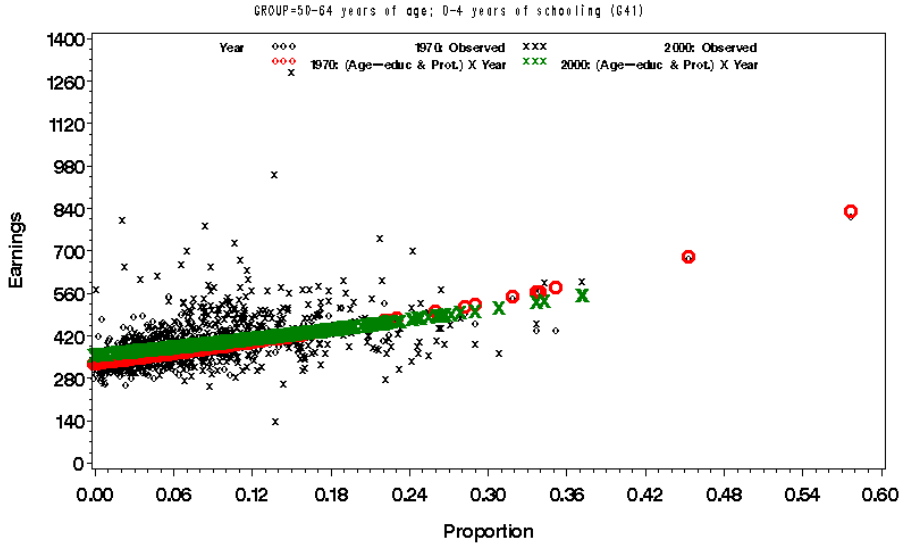


IDADE 50-64 / ESCOLARIDADE 0-4

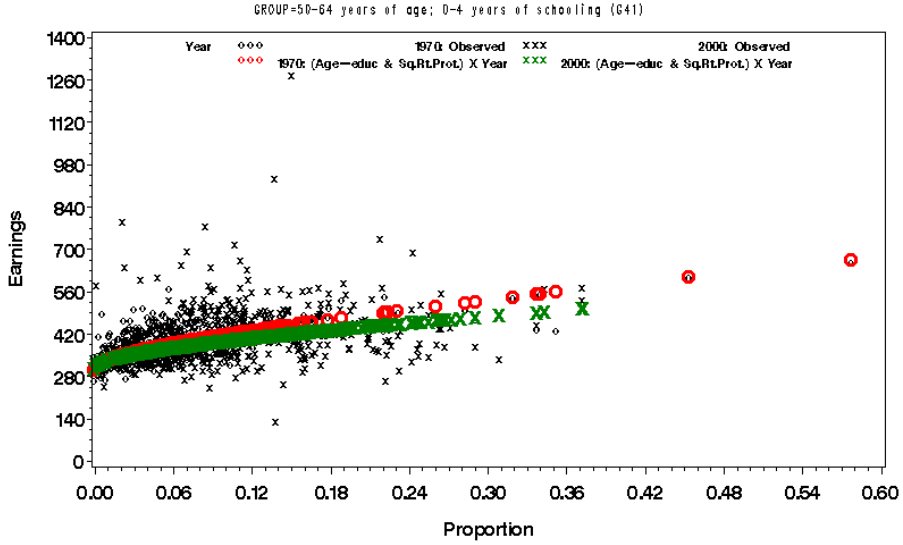
Prop. protestantes



Prop. protestantes * Ano



Raiz quadrada(Prop. protestantes) * Ano



CAPÍTULO 4 - WOOLDRIDGE
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)

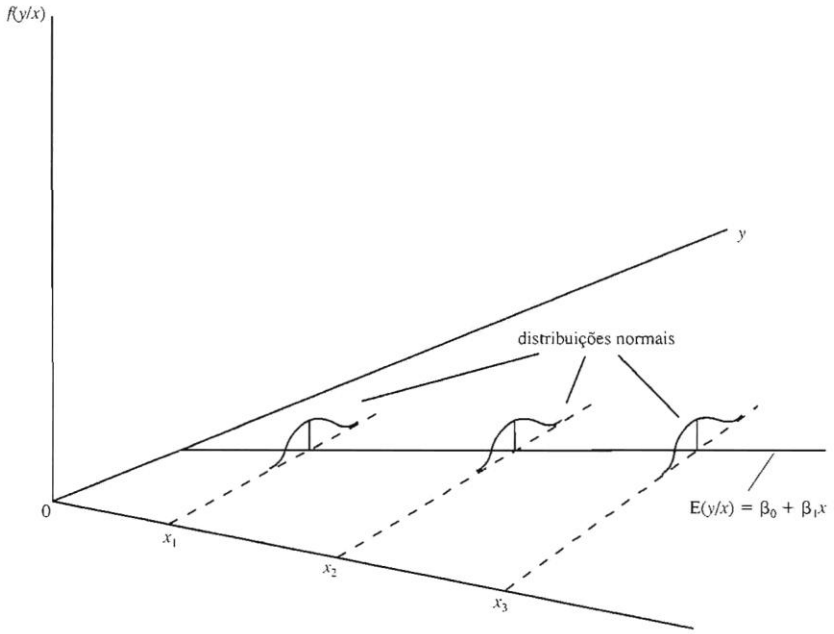
MODELO LINEAR CLÁSSICO

– 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.



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

$H_1: \beta_j > 0$ OU $H_1: \beta_j < 0$

– 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 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...

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.

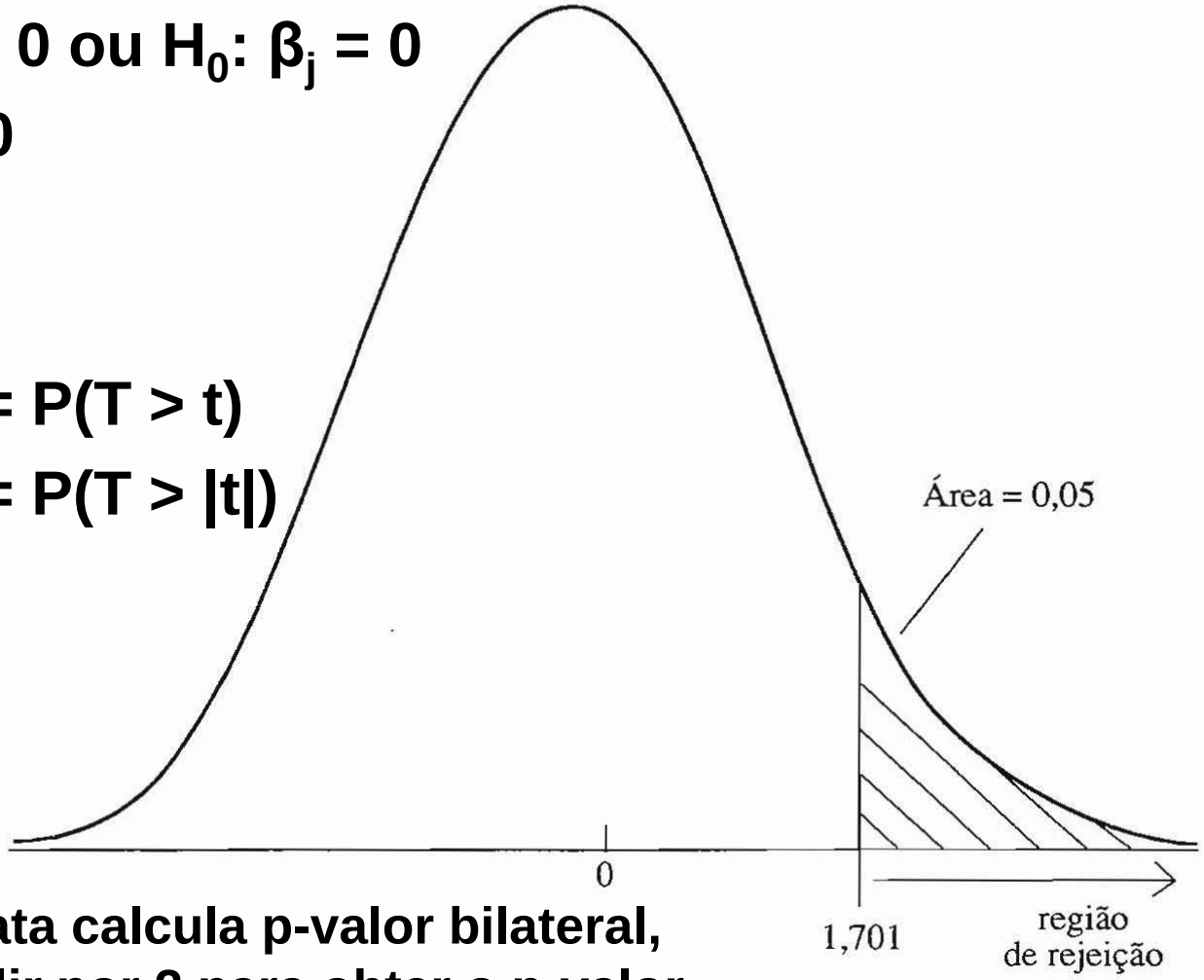
gl = n-k-1 = 2725-4-1 = 2720

```
reg narr86 pcnv avgsen ptime86 qemp86

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

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*.

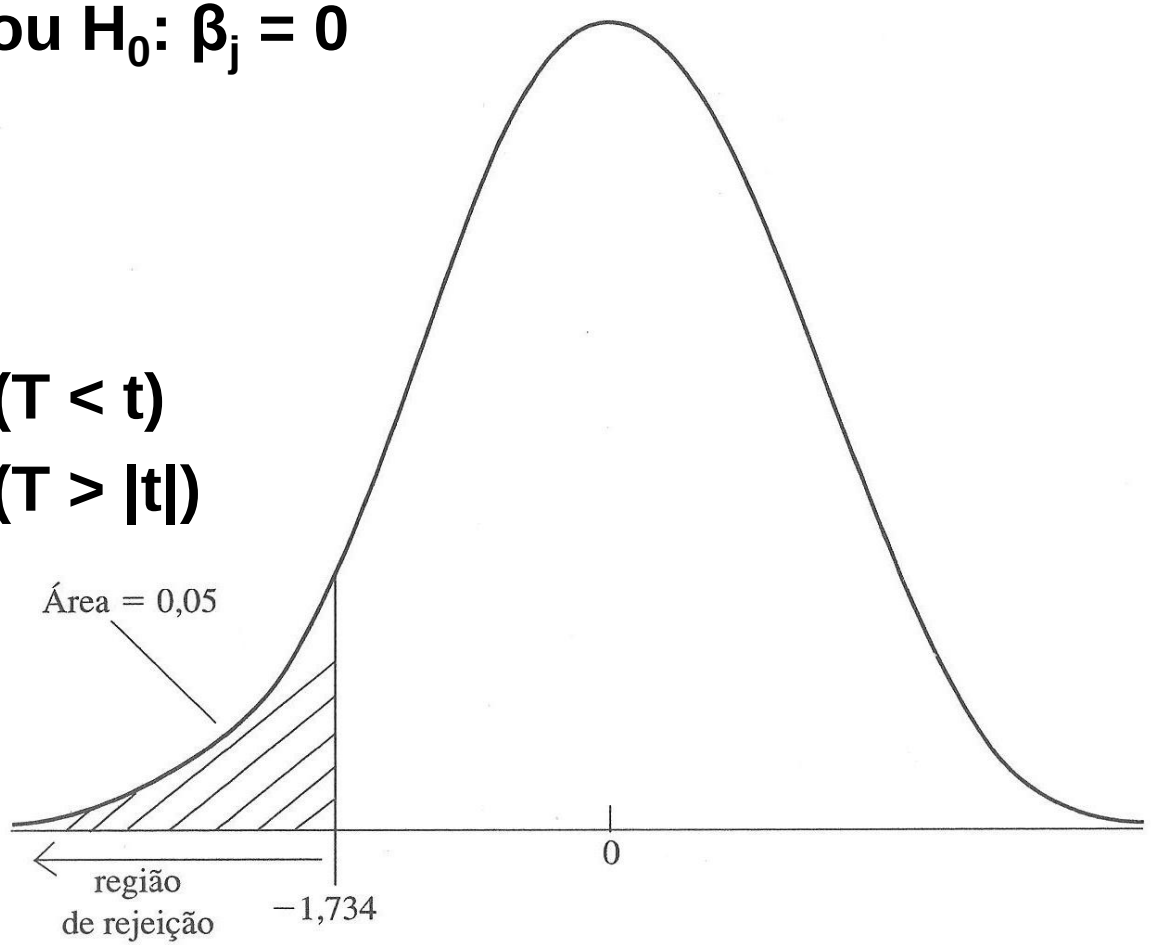
H_0: β_j >= 0 ou H_0: β_j = 0

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

$$H_1: \beta_j \neq 0$$

– 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

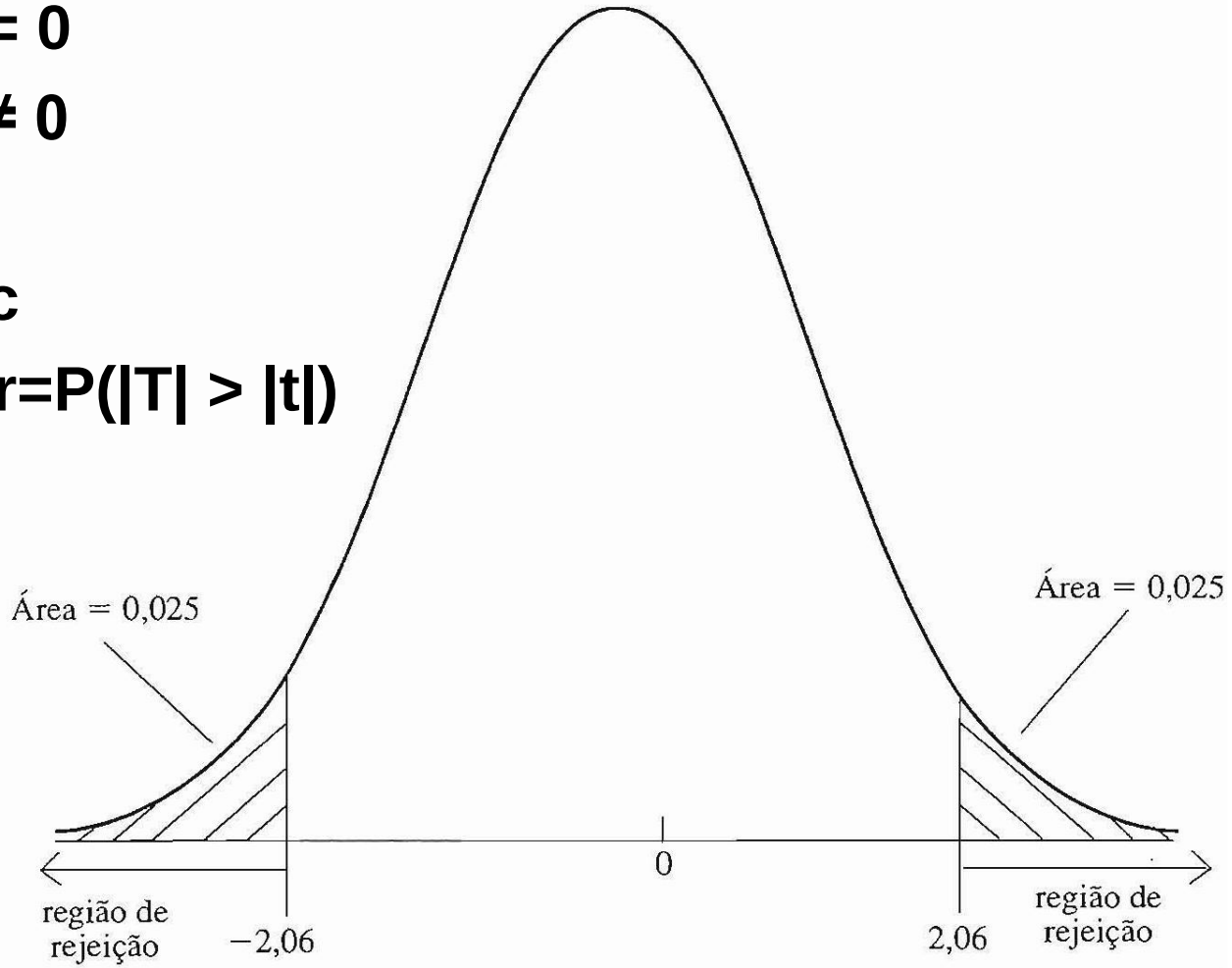
Regra de rejeição a 5% para a hipótese alternativa $H_1: \beta_j \neq 0$ com 25 *gl*.

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

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

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

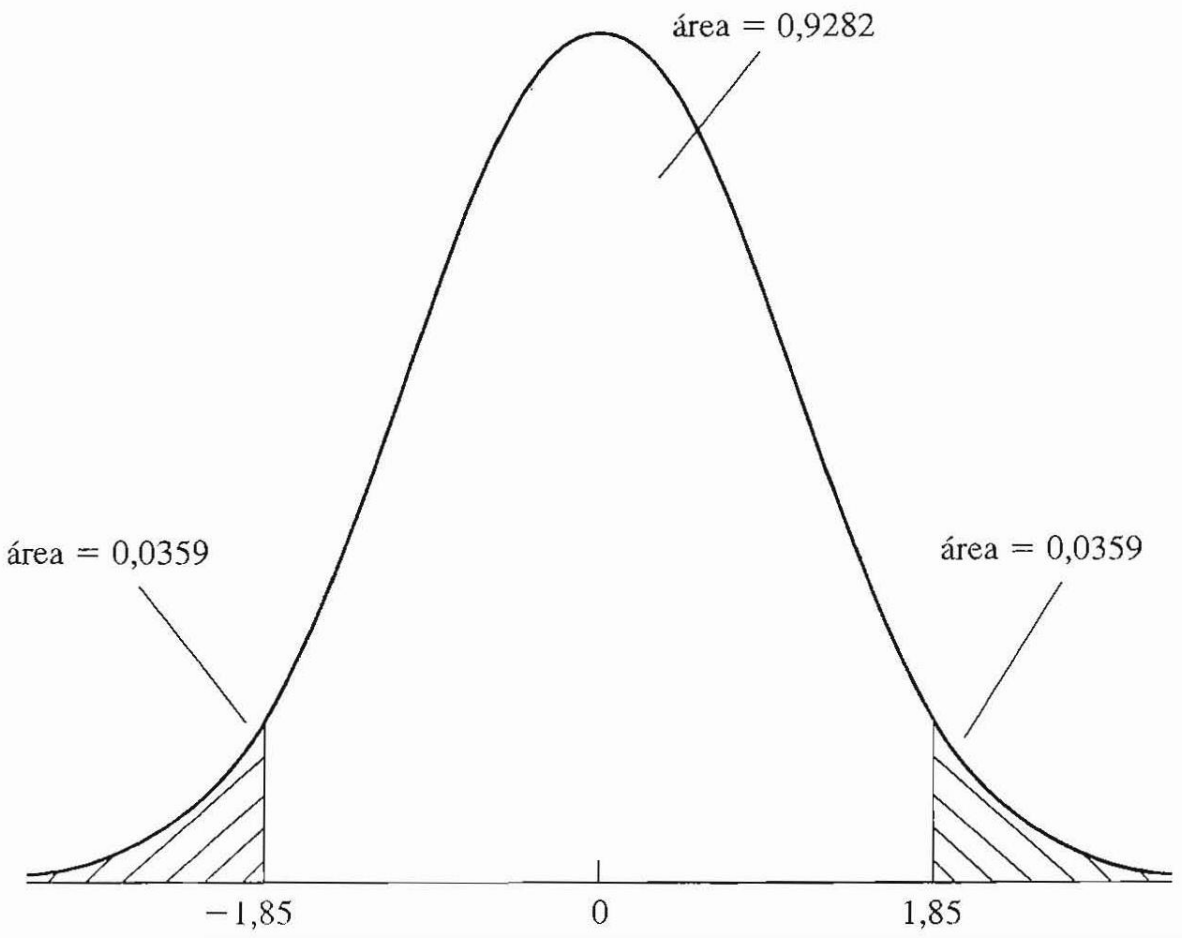


Fonte: Wooldridge, 2008: 122.

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.

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: $\beta_j - c*ep(\beta_j)$
  – Valor superior: $\beta_j + c*ep(\beta_j)$
– 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 (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: $F > c$.

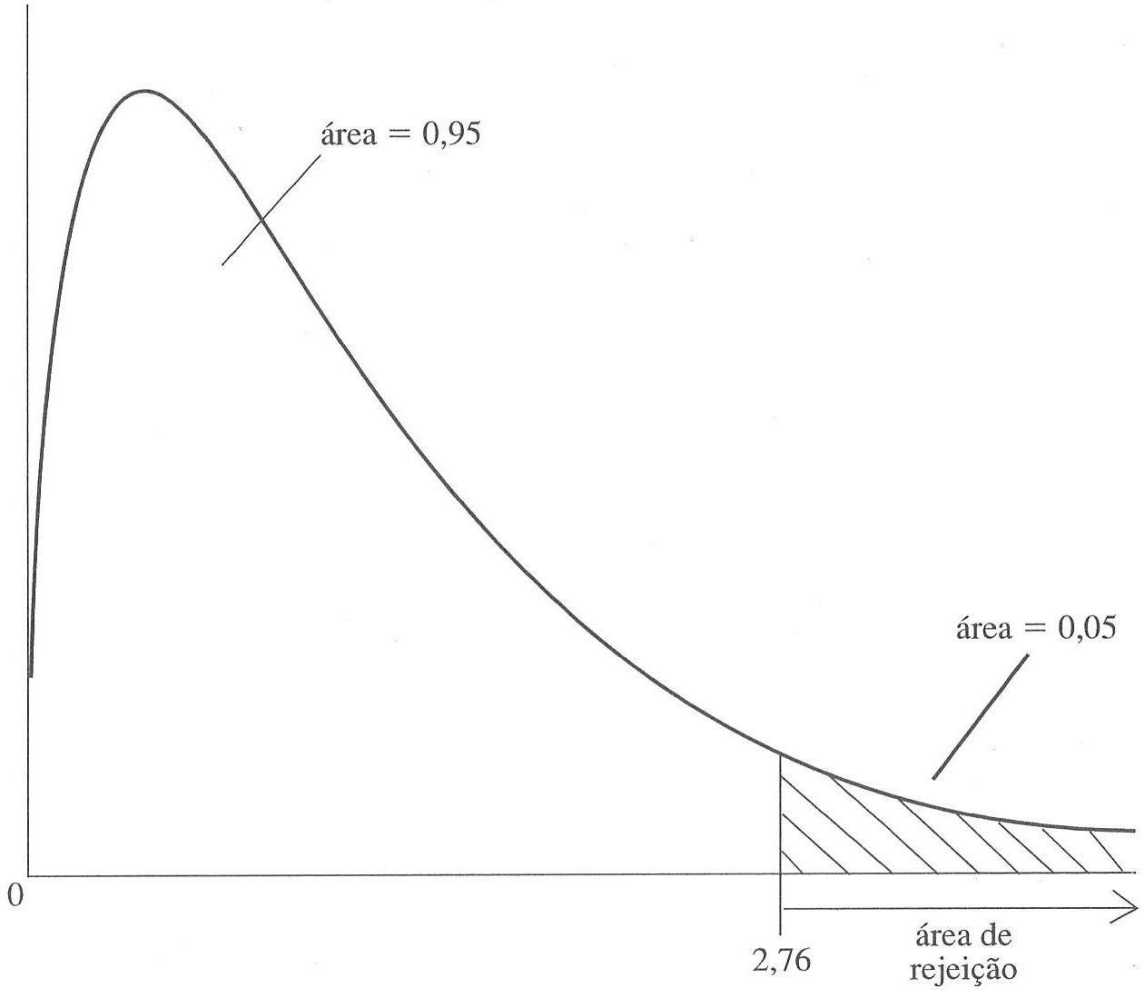
– Se H_0 ($\beta_3=0$, $\beta_4=0$, $\beta_5=0$) é rejeitada, β_3, β_4 e β_5 são **estatisticamente significantes conjuntamente**.

– 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.

PESO POPULACIONAL ≠ PESO AMOSTRAL

INDIVÍDUO	NÚMERO DE OBSERVAÇÕES	PESO POPULACIONAL	PESO AMOSTRAL
João	1	4	0,8
Maria	1	6	1,2
TOTAL	2	10	2

EXEMPLO:
Peso amostral do João =
Peso populacional do João * Peso amostral total / Peso populacional total

PESO POPULACIONAL NO STATA

- **FWEIGHT**:
  - Expande os resultados da amostra para o tamanho populacional.
  - Utilizado em tabelas para gerar frequências.
  - O uso desse peso é importante na amostra do Censo Demográfico e na Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatística (IBGE) para expandir a amostra para o tamanho da população do país, por exemplo.

```
tab x [fweight = peso]
```

PESO AMOSTRAL PARA PROGRAMADORES NO STATA

– **IWEIGHT**:
  – Não tem uma explicação estatística formal.
  – Esse peso é utilizado por programadores que precisam implementar técnicas analíticas próprias.

```
regress y x1 x2 [iweight = peso]
```

PESO AMOSTRAL ANALÍTICO NO STATA

- **AWEIGHT**:
  - Inversamente proporcional à variância da observação.
  - Número de observações na regressão é escalonado para permanecer o mesmo que o número no banco.
  - Utilizado para estimar uma regressão linear quando os dados são médias observadas, tais como:

group	x	y	n
1	3.5	26.0	2
2	5.0	20.0	3

  - Ao invés de:

group	x	y
1	3	22
1	4	30
2	8	25
2	2	19
2	5	16

UM POUCO MAIS SOBRE O AWEIGHT

– De uma forma geral, não é correto utilizar o **AWEIGHT** como um peso amostral, porque as fórmulas utilizadas por esse comando assumem que pesos maiores se referem a observações medidas de forma mais acurada.

– Uma observação em uma amostra não é medida de forma mais cuidadosa que nenhuma outra observação, já que todas fazem parte do mesmo plano amostral.

– Usar o **AWEIGHT** para especificar pesos amostrais fará com que o Stata estime valores incorretos de variância e de erros padrões para os coeficientes, assim como valores incorretos de "p" para os testes de hipótese.

```
regress y x1 x2 [aweight = peso]
```

PESO AMOSTRAL NAS REGRESSÕES DO STATA

- **PWEIGHT**:
  - Ideal para ser usado nas regressões do Stata.
  - Usa o peso amostral como o número de observações na população que cada observação representa.
  - São estimadas proporções, médias e parâmetros da regressão corretamente.
  - Há o uso de uma técnica de estimação robusta da variância que automaticamente ajusta para as características do plano amostral, de tal forma que variâncias, erros padrões e intervalos de confiança são calculados de forma mais precisa.
  - É o inverso da probabilidade da observação ser incluída no banco, devido ao desenho amostral.

```
regress y x1 x2 [pweight = peso]
```