AULAS 20 E 21
Análise de regressão múltipla: problemas adicionais

Ernesto F. L. Amaral

21 e 23 de maio de 2013
Avaliação de Políticas Públicas (DCP 046)

Fonte:
Wooldridge, Jeffrey M. "Introdução à econometria: uma abordagem moderna". São Paulo: Cengage Learning, 2008. pp.174-206 (capítulo 6).

EFEITOS DA DIMENSÃO DOS DADOS NAS ESTATÍSTICAS

– Mudanças das unidades de medida das variáveis não afeta o R^2.

– A intenção agora é de examinar o efeito do redimensionamento das variáveis dependente ou independente sobre:

  – Erros-padrão.

  – Estatísticas *t*.

  – Estatísticas *F*.

  – Intervalos de confiança.

– Escolhendo as unidades de medida, a aparência da equação estimada pode melhorar, sem alterar a essência do modelo.

– É geralmente realizada com valores monetários, especialmente quando os montantes são muito grandes.

EXEMPLO

– *pesônas*: peso dos recém-nascidos, em onças.

– *cigs*: número médio de cigarros que a mãe fumou por dia durante a gravidez.

– *rendfam*: renda anual familiar, em milhares de dólares.

– Equação 1:

$$\widehat{pesonas} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 rendfam$$

EFEITOS DA DIMENSÃO DOS DADOS

Variável Dependente	(1) pesonas	(2) pesonaslb = pesonas/16	(3) pesonas
Variáveis Independentes			
cigs	-0,4634 (0,0916)	-0,0289 (0,0057)	-----
maços = cigs/20	-----	-----	-9,268 (1,832)
rendfam	0,0927 (0,0292)	0,0058 (0,0018)	0,0927 (0,0292)
intercepto	116,974 (1,049)	7,3109 (0,0656)	116,974 (1,049)
Observações	1.388	1.388	1.388
R-quadrado	0,0298	0,0298	0,0298
SQR	557.485,51	2.177,6778	557.485,51
EPR	20,063	1,2539	20,063

MUDANÇA NA DEPENDENTE

– Não importa como a variável dependente seja medida, os efeitos da constante e coeficientes são transformados nas mesmas unidades.

– Equação 2:

$$\widehat{pesonas}/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16)cigs + (\hat{\beta}_2/16)rendfam$$

E A SIGNIFICÂNCIA ESTATÍSTICA?

– A alteração da variável dependente de onças para libras não tem efeito sobre o quanto são estatisticamente importantes as variáveis independentes.

– Os erros-padrão na coluna (2) são 16 vezes menores que os da coluna (1).

– As estatísticas *t* na coluna (2) são idênticas às da coluna (1).

– Os pontos extremos dos intervalos de confiança na coluna (2) são exatamente os pontos extremos na coluna (1) divididos por 16, já que ICs mudam pelos mesmos fatores dos erros-padrão.

– IC de 95% é beta estimado +/- 1,96 erro padrão estimado.

E O GRAU DE AJUSTE? E O SQR? E O EPR?

– Os R^2 das duas regressões são idênticos, como esperado.

– A soma dos resíduos quadrados (SQR) e o erro-padrão da regressão (EPR) possuem diferentes equações.

– Quando *pesonaslb* é a variável dependente, o resíduo da observação *i* na equação (2) é: $\hat{u}_i/16$

– O resíduo quadrado em (2) é: $(\hat{u}_i/16)^2 = \hat{u}_i^2/256$

– Por isso, **SQR(2) = SQR(1) / 256**.

– Como: $EPR = \hat{\sigma} = \sqrt{SQR/(n-k-1)} = \sqrt{SQR/1.385}$

– Por isso, **EPR(2) = EPR(1) / 16**.

REDUZIMOS O ERRO?

– O erro na equação com *pesonaslb* como a variável dependente tem um desvio-padrão 16 vezes menor do que o desvio-padrão do erro original.

– Isso não significa reduzir o erro por mudar a medida da variável dependente.

– O EPR menor simplesmente reflete uma diferença nas unidades de medida.

EXEMPLO 1: PNAD DE MINAS GERAIS DE 2007

– Rendimento em reais como variável dependente:

```
. *Modelo com reais
. reg renpri mulher idpia anest negra [aweight=v4729]
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	4.7726e+09	4	1.1932e+09
Residual	2.0236e+10	15615	1295928.74
Total	2.5009e+10	15619	1601162.78

Number of obs	15620
F(4, 15615)	920.70
Prob > F	0.0000
R-squared	0.1908
Adj R-squared	0.1906
Root MSE	1138.4

renpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mulher	-456.3604	18.81349	-24.26	0.000	-493.2371	-419.4838
idpia	26.5257	.7831009	33.87	0.000	24.99073	28.06067
anest	118.2471	2.375738	49.77	0.000	113.5904	122.9038
negra	-173.1548	18.80343	-9.21	0.000	-210.0117	-136.2979
_cons	-813.123	42.34102	-19.20	0.000	-896.1164	-730.1297

EXEMPLO 2: PNAD DE MINAS GERAIS DE 2007

– Rendimento em dólares (reais dividido por dois) como variável dependente:

```
. *Modelo com dólares
. reg rendol mulher idpia anest negra [aweight=v4729]
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	1.1932e+09	4	298289636
Residual	5.0590e+09	15615	323982.185
Total	6.2521e+09	15619	400290.695

Number of obs	15620
F(4, 15615)	920.70
Prob > F	0.0000
R-squared	0.1908
Adj R-squared	0.1906
Root MSE	569.19

rendol	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mulher	-228.1802	9.406745	-24.26	0.000	-246.6185	-209.7419
idpia	13.26285	.3915504	33.87	0.000	12.49536	14.03033
anest	59.12355	1.187869	49.77	0.000	56.79519	61.45191
negra	-86.57738	9.401716	-9.21	0.000	-105.0058	-68.14893
_cons	-406.5615	21.17051	-19.20	0.000	-448.0582	-365.0649

– Coeficientes, erros padrões, intervalo de confiança e erro padrão do resíduo são 2 vezes menores.

– R^2 é o mesmo.

MUDANÇA NA INDEPENDENTE

– *maços*: quantidade de maços de cigarros fumados por dia:

maços = cigs / 20

$$\widehat{pesonas} = \hat{\beta}_0 + (20\hat{\beta}_1)\left(\frac{cigs}{20}\right) + \hat{\beta}_2 rendfam$$
$$= \hat{\beta}_0 + (20\hat{\beta}_1)maços + \hat{\beta}_2 rendfam$$

– O intercepto e o coeficiente de inclinação de *rendfam* não se alteraram.

– O coeficiente de *maços* é 20 vezes o de *cigs*.

– O erro-padrão de *maços* é 20 vezes o de *cigs*, o que significa que a estatística *t* é a mesma.

– Se *maços* e *cigs* fossem inseridos conjuntamente, teríamos multicolinearidade perfeita.

EXEMPLO 3: PNAD DE MINAS GERAIS DE 2007

– Idade original como variável independente:

```
. *Modelo com idpia
. reg renpri mulher idpia anest negra [aweight=v4729]
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	4.7726e+09	4	1.1932e+09
Residual	2.0236e+10	15615	1295928.74
Total	2.5009e+10	15619	1601162.78

Number of obs	15620
F(4, 15615)	920.70
Prob > F	0.0000
R-squared	0.1908
Adj R-squared	0.1906
Root MSE	1138.4

renpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mulher	-456.3604	18.81349	-24.26	0.000	-493.2371	-419.4838
idpia	26.5257	.7831009	33.87	0.000	24.99073	28.06067
anest	118.2471	2.375738	49.77	0.000	113.5904	122.9038
negra	-173.1548	18.80343	-9.21	0.000	-210.0117	-136.2979
_cons	-813.123	42.34102	-19.20	0.000	-896.1164	-730.1297

EXEMPLO 4: PNAD DE MINAS GERAIS DE 2007

– Idade dividida por 5 como variável independente (impacto de 5 em 5 anos de idade):

```
. *Modelo com id5
. reg renpri mulher id5 anest negra [aweight=v4729]
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	4.7726e+09	4	1.1932e+09
Residual	2.0236e+10	15615	1295928.74
Total	2.5009e+10	15619	1601162.78

Number of obs	15620
F(4, 15615)	920.70
Prob > F	0.0000
R-squared	0.1908
Adj R-squared	0.1906
Root MSE	1138.4

renpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mulher	-456.3604	18.81349	-24.26	0.000	-493.2371	-419.4838
id5	132.6285	3.915504	33.87	0.000	124.9536	140.3033
anest	118.2471	2.375738	49.77	0.000	113.5904	122.9038
negra	-173.1548	18.80343	-9.21	0.000	-210.0117	-136.2979
_cons	-813.1231	42.34102	-19.20	0.000	-896.1164	-730.1297

– Coeficiente e erro padrão de idade são 5 vezes maiores.

– As demais estimativas são as mesmas.

COEFICIENTES BETA

– Algumas vezes, uma variável-chave é medida em uma dimensão de difícil interpretação.

– Primeiro exemplo: ao invés de perguntar o efeito sobre o salário, proveniente do aumento em dez pontos em um teste, talvez faça mais sentido perguntar sobre efeito proveniente do aumento de um desvio-padrão.

– Segundo exemplo: é o caso de variáveis criadas com análise fatorial, já que não sabemos exatamente o que a unidade de medida significa.

– Como o desvio-padrão da variável “fatorial” é geralmente próximo de uma unidade, verificamos o efeito na unidade da variável dependente (beta), após a alteração de um desvio-padrão na variável independente.

COEFICIENTES PADRONIZADOS

– Algumas vezes é útil obter resultados de regressão quando todas as variáveis tenham sido padronizadas.

– Uma variável é padronizada pela subtração de sua média e dividindo o resultado por seu desvio-padrão.

– Ou seja, computamos a transformação *z* de cada variável e depois fazemos a regressão usando esses valores *z*.

– Portanto, partimos de:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + \hat{u}$$

– Novo beta = beta original * (dp de x / dp de y)

– Intercepto (beta zero) não existe mais:

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \cdots + \hat{b}_k z_k + erro$$

– Para j = 1,...,k, os coeficientes são: $\hat{b}_j = \left(\hat{\sigma}_j / \hat{\sigma}_y\right)\hat{\beta}_j$

INTERPRETANDO COEFICIENTES PADRONIZADOS

– Os coeficientes padronizados são também chamados de coeficientes beta.

– Se x_1 aumentar em um desvio-padrão, então o y predito será alterado em b_1 desvios-padrão.

– Os efeitos não estão sendo medidos em termos das unidades originais de y ou de x_j, mas em unidades de desvios-padrão.

– A dimensão das variáveis independentes passa a ser irrelevante, colocando-as em igualdade.

– Quando cada x_j é padronizado, a comparação das magnitudes dos coeficientes (significância econômica) é mais convincente. Ou seja, a variável com maior coeficiente é a “mais importante”.

– O Stata apresenta os beta padronizados com opção “, beta”.

EXEMPLO 5: PNAD DE MINAS GERAIS DE 2007

– Coeficiente padronizado:

```
. *Modelo com betas padronizados
. reg renpri mulher idpia anest negra [aweight=v4729], beta
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	4.7726e+09	4	1.1932e+09
Residual	2.0236e+10	15615	1295928.74
Total	2.5009e+10	15619	1601162.78

Number of obs	15620
F(4, 15615)	920.70
Prob > F	0.0000
R-squared	0.1908
Adj R-squared	0.1906
Root MSE	1138.4

renpri	Coef.	Std. Err.	t	P>|t|	Beta
mulher	-456.3604	18.81349	-24.26	0.000	-.1772221
idpia	26.5257	.7831009	33.87	0.000	.2537558
anest	118.2471	2.375738	49.77	0.000	.386066
negra	-173.1548	18.80343	-9.21	0.000	-.0682722
_cons	-813.123	42.34102	-19.20	0.000	.

USO DE FORMAS FUNCIONAIS LOGARÍTMICAS

– O uso de logaritmos das variáveis dependentes ou independentes é o artifício mais comum em econometria para permitir relações não lineares entre a variável explicada e as variáveis explicativas.

$$\widehat{log}(y) = \hat{\beta}_0 + \hat{\beta}_1 log(x_1) + \hat{\beta}_2 x_2$$

– β_1 é a elasticidade de y, em relação a x_1:

  – Quando x_1 aumenta em 1%, y aumenta em β_1%, mantendo x_2 fixo.

– 100*β_2 é a semi-elasticidade de y, em relação a x_2:

  – Quando x_2 aumenta em 1, y aumenta em 100*[exp(β_2)-1]%, mantendo x_1 fixo.

  – No entanto, podemos utilizar (100*β_2)%, quando temos pequenas mudanças percentuais.

PECULIARIDADES DO USO DE LOGARITMOS

– Com log, ignoramos unidades de medida das variáveis, pois coeficientes de inclinação não variam pelas unidades.

– Quando y>0, os modelos que usam log(y) satisfazem MQO mais do que os modelos que usam o nível original de y.

– Log é útil para variáveis estritamente positivas com grandes valores e distribuição concentrada, tais como: renda, vendas de empresas, população, matrículas, empregados, votação.

– Log estreita amplitude dos valores, tornando estimativas menos sensíveis a observações extremas (*outliers*).

– Variáveis medidas em anos aparecem em forma original.

– Taxas geralmente aparecem em forma original.

– Log não é usado se variável tem valor zero ou negativo.

– Não é válido comparar R^2 entre modelos com y e log(y).

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:

$$salário = \beta_0 + \beta_1 educ + u$$

– 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:

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

– Se $\Delta u = 0$, então:

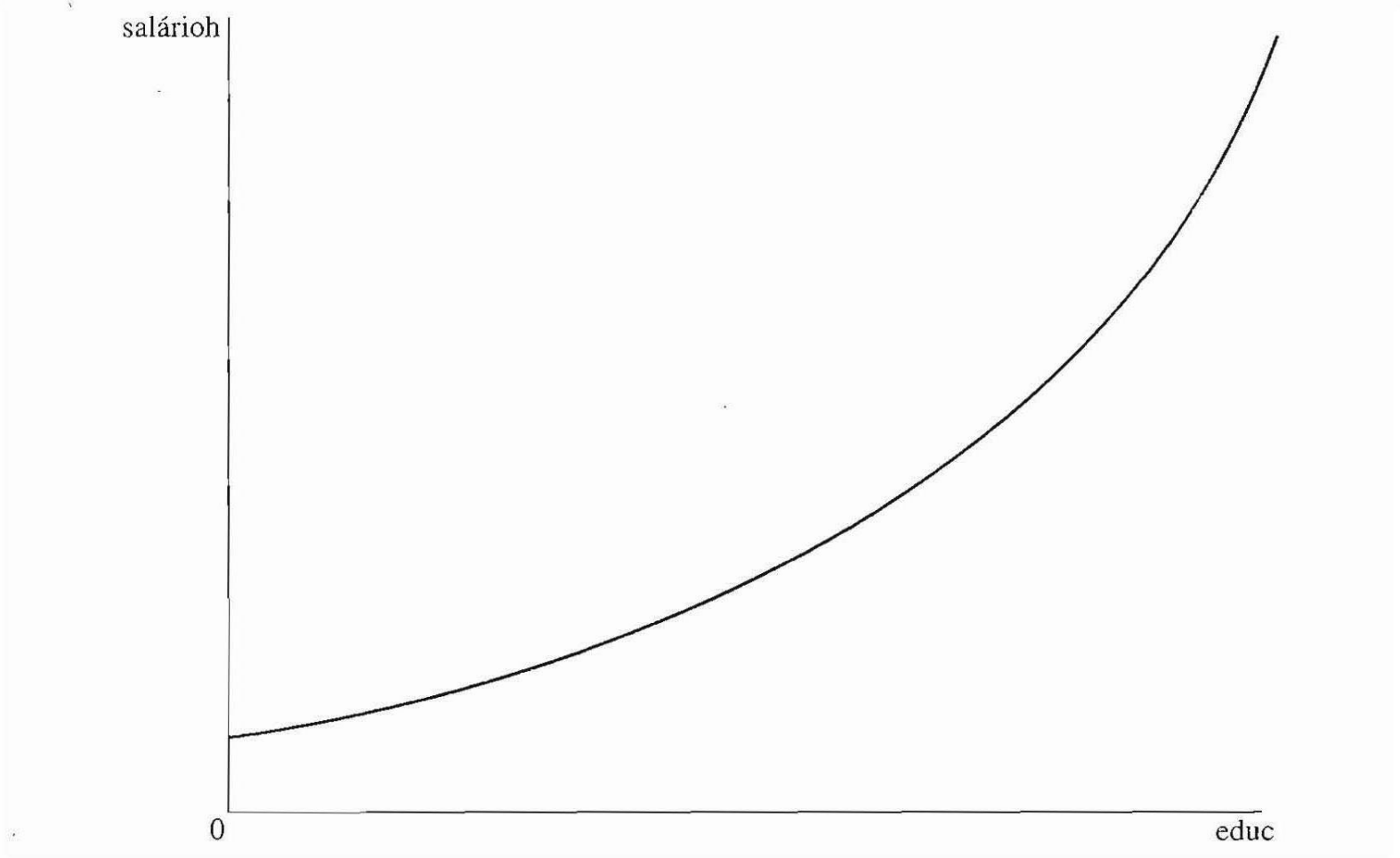
$$\%\Delta salário = (100 * \beta_1)\Delta educ$$

– 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

salárioh = $\exp(\beta_0 + \beta_1 educ)$, *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 (β_1/100) unidades:

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

– Aumento de uma unidade em x aumenta y em 100*[exp(β_1)-1]% ou, aproximadamente, em (100*β_1)%:

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

– Aumento de 1% em x aumenta y em β_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)	$\%\Delta y=\beta_1\%\Delta x$

EXEMPLO 6: PNAD DE MINAS GERAIS DE 2007

– Logaritmo do rendimento como variável dependente:

```
. reg lnrenpri mulher idpia anest negra [aweight=v4729]
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	4047.56788	4	1011.89197
Residual	7092.0144	15615	.454179597
Total	11139.5823	15619	.713207137

Number of obs	15620
F(4, 15615)	2227.96
Prob > F	0.0000
R-squared	0.3634
Adj R-squared	0.3632
Root MSE	.67393

lnrenpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mulher	-.5330765	.0111376	-47.86	0.000	-.5549075	-.5112455
idpia	.0215629	.0004636	46.51	0.000	.0206542	.0224716
anest	.1086479	.0014064	77.25	0.000	.1058911	.1114047
negra	-.1301593	.0111317	-11.69	0.000	-.1519786	-.1083399
_cons	4.921049	.025066	196.32	0.000	4.871917	4.970181

MODELOS COM FUNÇÕES QUADRÁTICAS

– Funções quadráticas são usadas para capturar efeitos marginais crescentes ou decrescentes.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

– Sempre existe um valor positivo de *x*, no qual o efeito de *x* sobre *y* é zero, chamado de ponto crítico: $x^* = |\beta_1/(2\beta_2)|$.

– Interpretações: (1) após ponto crítico, a relação se inverte; (2) após/antes ponto crítico, há poucos casos; (3) falta incluir variáveis; ou (4) falta transformar variáveis.

– Quando o coeficiente de *x* é positivo e o coeficiente de x^2 é negativo, a função quadrática tem um formato parabólico (∩):

  – Antes desse ponto, *x* tem um efeito positivo sobre *y*.

  – Após esse ponto, *x* tem um efeito negativo sobre *y*.

– Se β_1 é negativo e β_2 é positivo, função tem formato U.

EXEMPLO 7: PNAD DE MINAS GERAIS DE 2007

– Transformação quadrática da idade:

```
. reg lnrenpri mulher idpia idquad anest negra [aweight=v4729]
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	4332.2922	5	866.458439
Residual	6807.29008	15614	.43597349
Total	11139.5823	15619	.713207137

Number of obs	15620
F(5, 15614)	1987.41
Prob > F	0.0000
R-squared	0.3889
Adj R-squared	0.3887
Root MSE	.66028

lnrenpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mulher	-.5440937	.0109206	-49.82	0.000	-.5654994	-.522688
idpia	.088968	.0026764	33.24	0.000	.0837219	.0942141
idquad	-.0008933	.000035	-25.56	0.000	-.0009618	-.0008248
anest	.1067622	.0013799	77.37	0.000	.1040573	.109467
negra	-.1368042	.0109094	-12.54	0.000	-.1581878	-.1154205
_cons	3.805854	.0500742	76.00	0.000	3.707703	3.904005

– Se coeficiente de idade original é negativo e idade ao quadrado é positivo, função tem formato U.

– Se coeficiente de idade original é positivo e idade ao quadrado é negativo, função tem formato parabólico.

– Ponto crítico = $|\beta_1/(2\ \beta_2)|$ = |0.088968/(2*-0.0008933)| = 49,8

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:

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u$$, 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.

EXEMPLO 8: PNAD DE MINAS GERAIS DE 2007

– Termo de interação entre idade e escolaridade (“idest”):

```
. reg lnrenpri mulher idpia anest negra idest [aweight=v4729]
(sum of wgt is   8.4198e+06)

      Source |       SS       df       MS              Number of obs =   15620
-------------+------------------------------           F(  5, 15614) = 1846.84
       Model |  4139.7391     5   827.94782             Prob > F      =  0.0000
    Residual |  6999.84317 15614   .44830557            R-squared     =  0.3716
-------------+------------------------------           Adj R-squared =  0.3714
       Total |  11139.5823 15619  .713207137            Root MSE      =  .66956

------------------------------------------------------------------------------
    lnrenpri |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mulher |  -.5244468   .0110817   -47.33   0.000    -.5461683   -.5027253
       idpia |   .0090526   .0009866     9.18   0.000     .0071188    .0109865
       anest |   .0449341    .004658     9.65   0.000     .0358039    .0540643
       negra |  -.1319659   .0110602   -11.93   0.000    -.1536451   -.1102867
       idest |   .0016014   .0001117    14.34   0.000     .0013825    .0018203
       _cons |   5.437715   .0438012   124.15   0.000      5.35186    5.523571
------------------------------------------------------------------------------
```

– Efeito parcial de idade sobre renda em 2 anos de estudo:

“idpia” + (“idest” * escolaridade) = 0,01 + (0,002*2) = 1,2%

– Efeito parcial de escolaridade sobre renda em 15 anos:

“anest” + (“idest” * idade) = 0,05 + (0,002*15) = 6,9%

EXEMPLO 9: PNAD DE MINAS GERAIS DE 2007

– Interação de idade e escolaridade, centralizada na média:

```
. reg lnrenpri mulher idpia anest negra idestmed [aweight=v4729]
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	4139.7391	5	827.94782
Residual	6999.84317	15614	.44830557
Total	11139.5823	15619	.713207137

Number of obs	= 15620
F(5, 15614)	= 1846.84
Prob > F	= 0.0000
R-squared	= 0.3716
Adj R-squared	= 0.3714
Root MSE	= .66956

lnrenpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mulher	-.5244468	.0110817	-47.33	0.000	-.5461683	-.5027253
idpia	.0220666	.0004619	47.77	0.000	.0211611	.022972
anest	.1027652	.0014563	70.57	0.000	.0999107	.1056197
negra	-.1319659	.0110602	-11.93	0.000	-.1536451	-.1102867
idestmed	.0016014	.0001117	14.34	0.000	.0013825	.0018203
_cons	4.967743	.0251154	197.80	0.000	4.918514	5.016972

– Efeito parcial de anos de estudo sobre renda indica que o aumento de um ano de estudo aumenta a renda em 10,3%, considerando uma pessoa de idade média (36 anos).

“anest” + “idestmed” * (idade - idade média)

– Efeito parcial de idade sobre renda indica que o aumento de um ano de idade aumenta a renda em 2,2%, considerando uma pessoa de escolaridade média (8 anos de estudo).

GRAU DE AJUSTE E SELEÇÃO DE REGRESSORES

– Seleção de variáveis explicativas com base no tamanho do R^2 pode levar a modelos absurdos.

– Nada nas hipóteses do modelo linear clássico exige que o R^2 esteja acima de qualquer valor em particular.

– O R^2 é simplesmente uma estimativa do quanto da variação em *y* é explicado por x_1, x_2, ..., x_k na população.

– Modelos com R^2 pequenos significam que não incluímos fatores importantes, mas não necessariamente significam que fatores em *u* estão correlacionados com os *x's*.

– O tamanho de R^2 não tem influência sobre a média dos resíduos ser igual a zero.

– R^2 pequeno sugere que variância do erro é grande em relação à variância de *y*, mas isso pode ser compensado por amostra grande.

R^2 AJUSTADO

– Sendo σ_y^2 a variância populacional de *y* e σ_u^2 a variância populacional do erro, R^2 da população é a proporção da variação em *y* na população, explicada pelas independentes:

$$R^2 = 1 - \sigma_u^2 / \sigma_y^2$$

– R^2 usual = SQE/SQT = 1 - SQR/SQT = 1 - (SQR/*n*) / (SQT/*n*)

– Podemos substituir o SQR/*n* e SQT/*n*, por termos não-viesados de σ_u^2 e σ_y^2 , e chegamos ao R^2 ajustado:

$$\bar{R}^2 = 1 - [SQR/(n-k-1)] / [SQT/(n-1)] = 1 - \hat{\sigma}^2 / [SQT/(n-1)]$$

$$= 1 - (1 - R^2)(n - 1)/(n - k - 1)$$

– R^2 ajustado não corrige viés de R^2 na estimativa do R^2 da população, mas penaliza inclusão de independentes.

– R^2 ajustado negativo indica adaptação ruim do modelo, relativo ao número de graus de liberdade.

$\bar{R}^2$ NA ESCOLHA DE MODELOS NÃO-ANINHADOS

– O R^2 ajustado auxilia na escolha de modelo sem variáveis independentes redundantes (entre modelos não-aninhados).

– A estatística F (*test*) permite testar somente modelos aninhados.

– Por exemplo, podemos testar se modelo com informação de idade é melhor do que modelo com experiência no mercado de trabalho:

$$\text{renda} = \beta_0 + \beta_1 \text{escolaridade} + \beta_2 \text{idade} + u$$

$$\text{renda} = \beta_0 + \beta_1 \text{escolaridade} + \beta_2 \text{experiência} + u$$

– Neste caso, não queremos incluir as duas variáveis em conjunto, pois teoricamente medem a mesma dimensão.

– Estes são modelos não-aninhados, exigindo comparação do R^2 ajustado.

$\bar{R}^2$ E MODELOS COM DIFERENTES FORMAS FUNCIONAIS

– Comparação dos R^2 ajustados pode ser feita para optar entre modelos com formas funcionais diferentes das variáveis independentes:

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

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

– Não podemos usar nem o R^2 nem o R^2 ajustado para escolher entre modelos não-aninhados com diferentes formas funcionais da variável dependente.

– Os R^2 medem a proporção explicada do total da variação de qualquer variável dependente:

  – Portanto, diferentes funções da variável dependente terão diferentes montantes de variação a serem explicados.

Coeficientes estimados por modelos de mínimos quadrados ordinários para explicação do logaritmo do rendimento no trabalho principal (variável dependente), Minas Gerais, 2007.

Variáveis independentes	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 4 (beta padronizado)
Constante	4,5830***	3,6660***	3,7810***	3,8060***	
	(0,0590)	(0,0532)	(0,0539)	(0,0501)	
Idade	0,0858***	0,0831***	0,0832***	0,0890***	1,263
	(0,0033)	(0,0029)	(0,0029)	(0,0027)	
Idade ao quadrado	-0,0010***	-0,0008***	-0,0008***	-0,0009***	-0,973
	(4,31e-05)	(3,78e-05)	(3,76e-05)	(3,50e-05)	
Anos de escolaridade		0,0996***	0,0956***	0,1070***	0,516
		(0,0014)	(0,0015)	(0,0014)	
Cor/raça					
Branca			ref.	ref.	ref.
Negra (preta e parda)			-0,1360***	-0,1370***	-0,0801
			(0,0117)	(0,0109)	
Sexo					
Homem				ref.	ref.
Mulher				-0,5440***	-0,315
				(0,0109)	
R^2	0,0643	0,2860	0,2920	0,3890	0,3890
R^2 ajustado	0,0640	0,2850	0,2920	0,3890	0,3890
Observações	15.620	15.620	15.620	15.620	15.620

Obs.: Erros padrão em parênteses.

* Significativo ao nível de confiança de 90%; ** Significativo ao nível de confiança de 95%; *** Significativo ao nível de confiança de 99%.

Fonte: Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2007.

EXEMPLO 10: PNAD DE MINAS GERAIS DE 2007

– Modelo para gerar gráfico de bolhas:

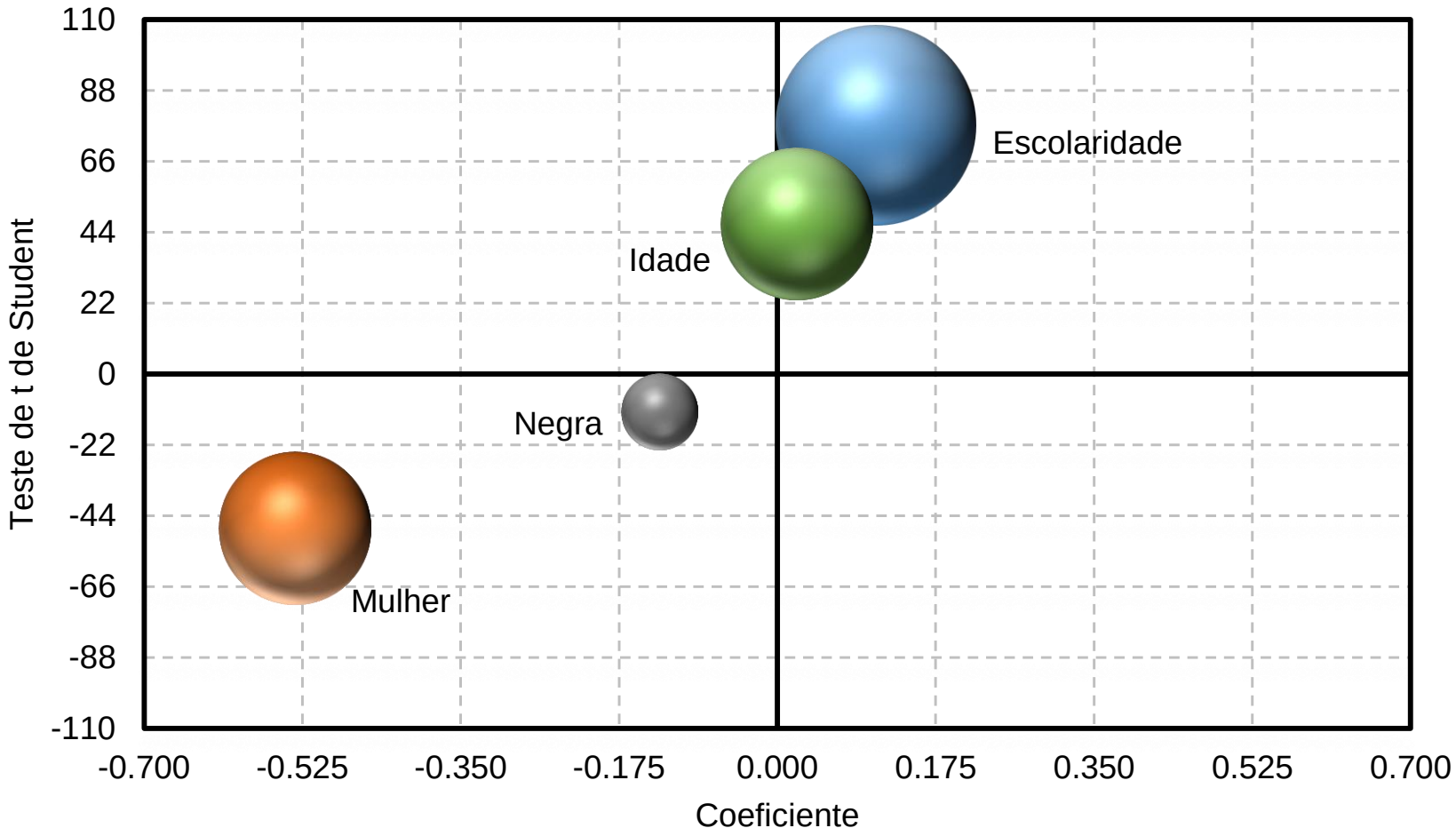
```
. reg lnrenpri mulher anest idpia negra [aweight=v4729], beta
(sum of wgt is   8.4198e+06)
```

Source	SS	df	MS
Model	4047.56788	4	1011.89197
Residual	7092.0144	15615	.454179597
Total	11139.5823	15619	.713207137

Number of obs	= 15620
F(4, 15615)	= 2227.96
Prob > F	= 0.0000
R-squared	= 0.3634
Adj R-squared	= 0.3632
Root MSE	= .67393

lnrenpri	Coef.	Std. Err.	t	P>|t|	Beta
mulher	-.5330765	.0111376	-47.86	0.000	-.3101769
anest	.1086479	.0014064	77.25	0.000	.5314987
idpia	.0215629	.0004636	46.51	0.000	.3090769
negra	-.1301593	.0111317	-11.69	0.000	-.0768943
_cons	4.921049	.025066	196.32	0.000	.

Resultado de modelo de mínimos quadrados ordinários para explicação do logaritmo de renda (variável dependente): coeficiente (eixo horizontal), teste de *t* de *Student* (eixo vertical) e beta padronizado (área da bolha) de variáveis independentes, Minas Gerais, 2007



Fonte: Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2007.

CONTROLE DE MUITOS FATORES NA REGRESSÃO

– Estamos preocupados com omissão de fatores importantes que possam estar correlacionados com as variáveis independentes.

– Se enfatizarmos o R^2, tenderemos a controlar fatores em um modelo que não deveriam ser controlados.

– Ao estudar o efeito da qualidade do ensino sobre a renda, talvez não faça sentido controlar os anos de escolaridade, pois subestimará o retorno da qualidade. Podemos estimar a equação com e sem anos de estudo.

– A questão de decidir se devemos ou não controlar certos fatores nem sempre é bem definida.

– Se nos concentrarmos na interpretação *ceteris paribus* da regressão, não incluiremos fatores no modelo, mesmo que estejam correlacionadas com a dependente.

ADIÇÃO DE FATORES: REDUZIR VARIÂNCIA DO ERRO

– A adição de uma nova variável independente pode aumentar o problema da multicolinearidade.

– Porém, ao adicionar uma variável, estamos reduzindo a variância do erro.

– Devemos incluir variáveis independentes que afetem *y* e que sejam não-correlacionadas com todas variáveis independentes, pois:

  – Não induzirá multicolinearidade.

  – Reduzirá variância do erro.

  – Diminuirá erros-padrão dos coeficientes beta, gerando estimativas mais precisas (estimador com menor variância do erro amostral).

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:

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

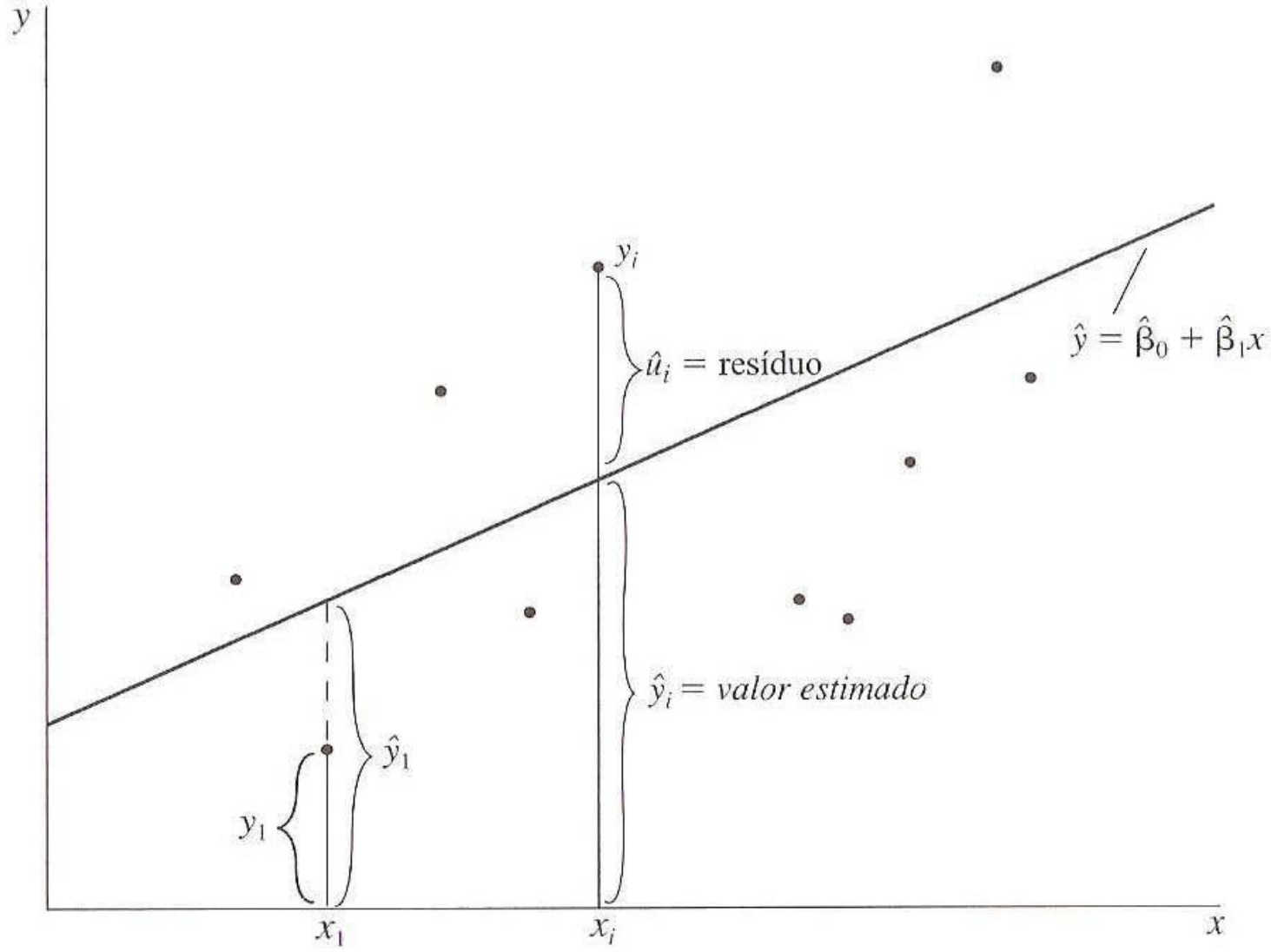
ANÁLISE DE RESÍDUOS

– É importante analisar os resíduos das observações individuais e examinar se valor efetivo da variável dependente está acima ou abaixo do valor previsto:

$$\hat{u}_i = y_i - \hat{y}_i$$

– Resíduo mais negativo indica valor observado mais baixo do que o previsto na regressão e vice-versa.

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:

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

– 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.

HOMOSCEDASTICIDADE DOS RESÍDUOS

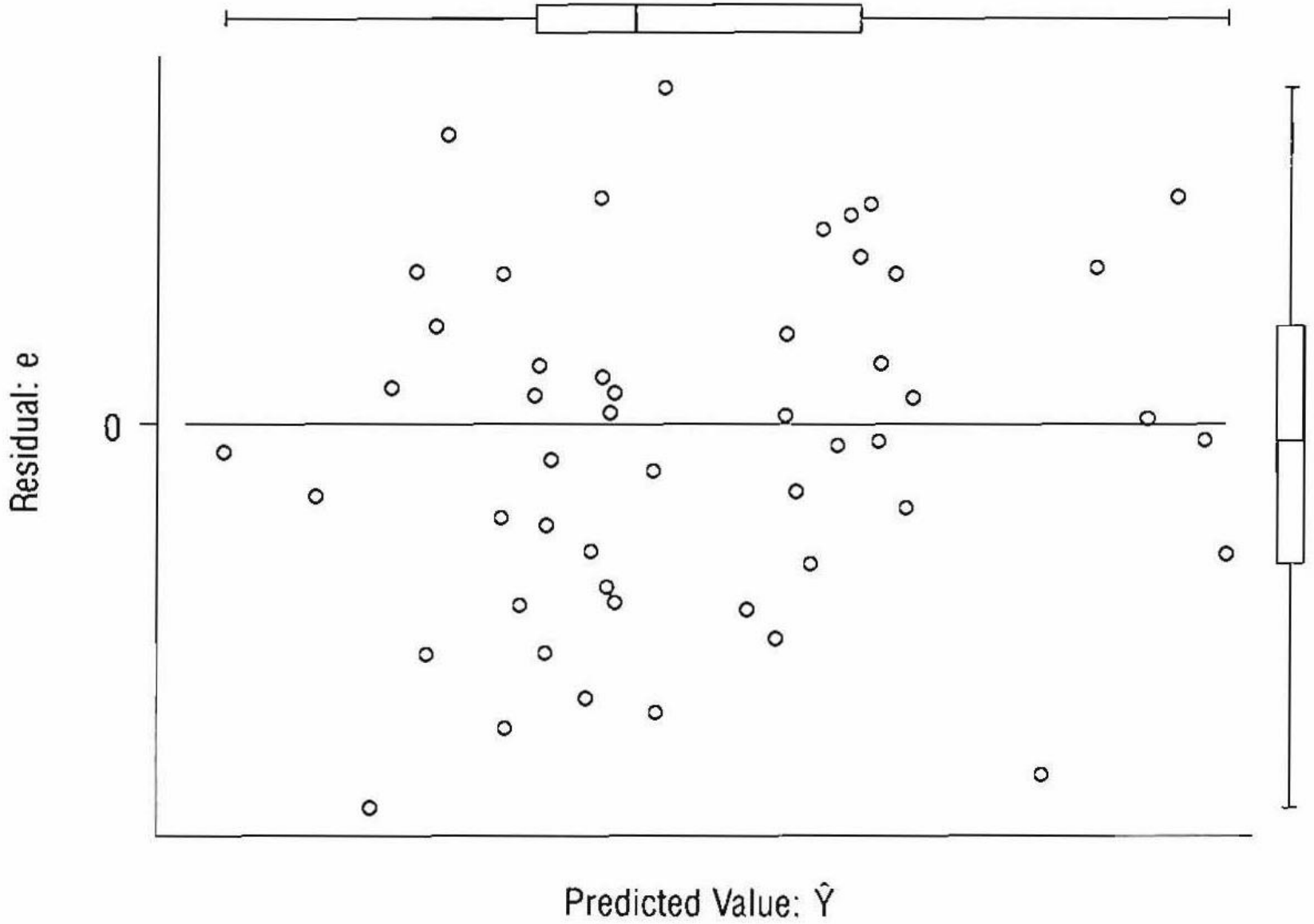


Figure 2.10 "All clear" *e*-versus-$\hat{Y}$ plot (artificial data).

Fonte: Hamilton, 1992: 52.

HETOROCEDASTICIDADE DOS RESÍDUOS

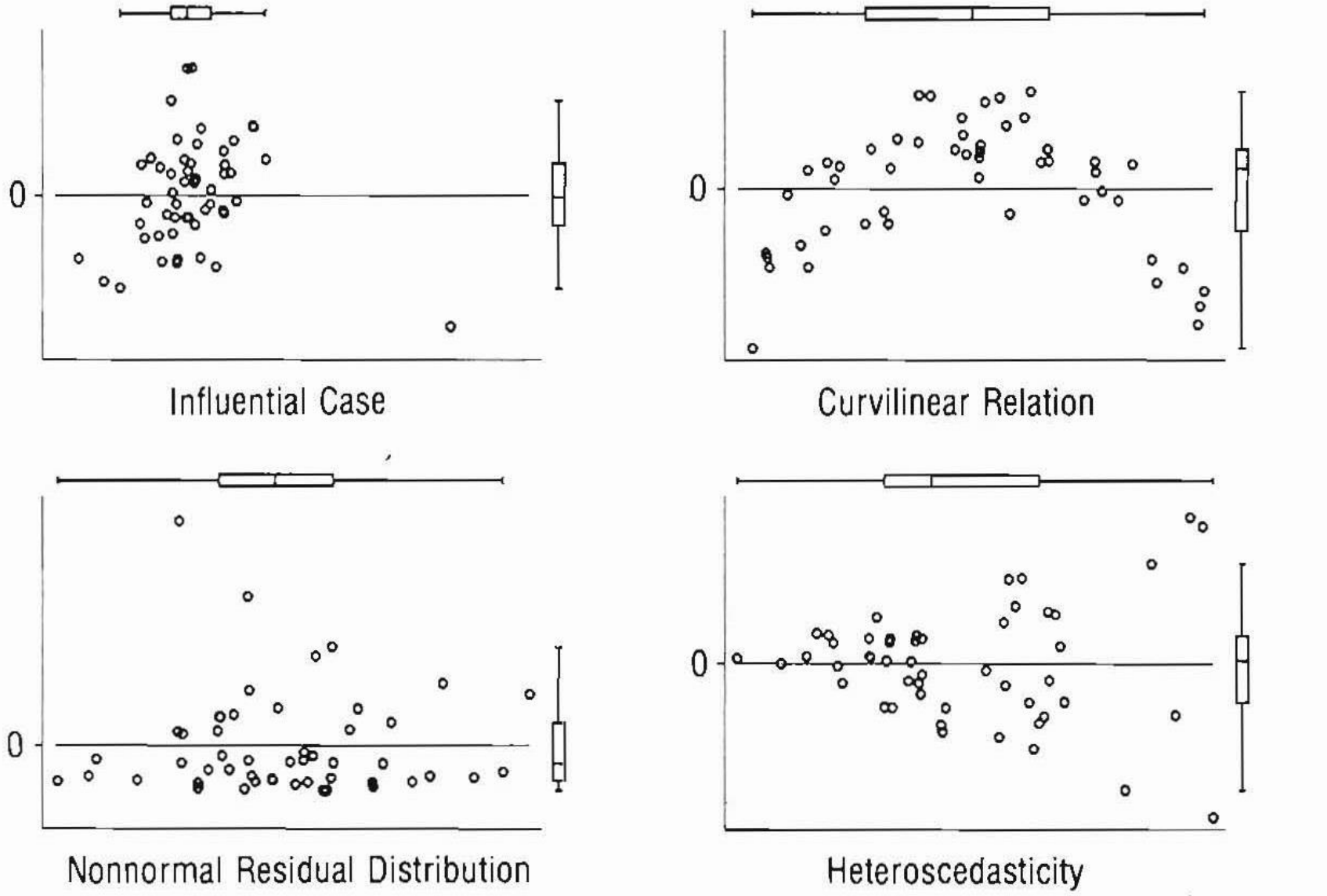


Figure 2.11 Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).

Fonte: Hamilton, 1992: 53.

LEMBRE-SE

– “É muito mais importante tornar-se proficiente em interpretar coeficientes do que eficiente no cálculo de fórmulas.” (Wooldridge, 2008: 45)

DIFERENTES PESOS

Indivíduo	Número de observações coletadas na amostra	Peso para expandir para o tamanho da população (N)	Peso para manter o tamanho da amostra (n)
João	1	4	0,8
Maria	1	6	1,2
Total	2	10	2

EXEMPLO:
Peso amostral do João =
Peso de frequência do João * (Peso amostral total / Peso de frequência total)

PESO DE FREQUÊNCIA 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.
  - Somente pode ser usado em tabelas de frequência quando o peso é uma variável discreta (não decimal).

```
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.
  – Pode ser utilizado em tabelas de frequência, mesmo que o peso seja decimal.

```
tab x [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]
```

OUTRAS OBSERVAÇÕES SOBRE PESOS NO STATA

PESOS EM TABELAS DE FREQUÊNCIA		
Tipo do peso	**Expandir para o tamanho da população (N)**	**Manter o tamanho da amostra (n)**
Discreto	**fweight**	**aweight**
Decimal	**iweight**	

PESOS EM MODELOS DE REGRESSÃO devem manter o tamanho da amostra (n)	
Erro padrão robusto	**R^2 ajustado, SQT, SQE, SQR**
pweight **reg y x, robust**	**aweight** **outreg2**

PLANO AMOSTRAL COMPLEXO

– Estatísticas descritivas e modelos de regressão devem levar em consideração a estrutura de planos amostrais complexos.

– PNAD tem amostra complexa (Silva, Pessoa, Lila, 2002):

  – Considerar variáveis de estrato de município autorrepresentativo e não autorrepresentativo (v4617) e de unidade primária de amostragem (v4618), do banco de domicílios.

  – Agregar variáveis acima ao banco de pessoas, o qual possui peso da pessoa (v4729).

  – Lidar com problema de alguns estratos terem somente uma unidade primária de amostragem. Pode-se especificar média deste estrato como sendo a média geral, ao invés da média do próprio estrato.

```
svyset [pweight=v4729], strata(v4617) psu(v4618) singleunit(centered)
```

  – Tabelas e regressões devem ser precedidas de “svy:”.