

AULAS 14 E 15

Modelo de regressão

simples

Ernesto F. L. Amaral

30 de abril e 02 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.20-63 (capítulo 2).

ESTRUTURA DO LIVRO

- **Parte 1:** trata de análise de regressão com dados de corte transversal (capítulos 2 ao 9).
- **Parte 2:** análise de regressão com dados de séries temporais (capítulos 10 ao 12).
- **Parte 3:** tópicos avançados (capítulos 13 ao 19).

DOCUMENTAÇÃO DO LIVRO

– UCLA Academic Technology Services:

<http://www.ats.ucla.edu>

– Introductory Econometrics: A Modern Approach
by Jeffrey M. Wooldridge:

<http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge.html>

MODELO DE REGRESSÃO SIMPLES

- O modelo de regressão linear simples explica uma variável (y) com base em modificações em outra variável (x).
- Ou seja, é usado para avaliar a relação entre duas variáveis.
- Esse tipo de regressão não é muito utilizada em ciências sociais aplicadas, devido à sua simplicidade.
- No entanto, serve como ponto de partida, já que sua álgebra e interpretações são fáceis de entender.
- O entendimento do modelo de regressão simples é importante para estudar a regressão múltipla.

PREMISSA E EXEMPLOS

- Premissa da análise econométrica:
 - y e x são duas variáveis que representam uma população.
 - Estamos interessados em explicar y em termos de x .
 - Ou seja, queremos estudar como y varia com variações em x .

- Exemplos:
 - y é o rendimento do trabalhador, e x são os anos de escolaridade.
 - y é a escala ideológica esquerda/direita, e x é o partido político do deputado.
 - y é o índice de tradicionalismo/secularismo, e x é o nível de escolaridade.

PERGUNTAS IMPORTANTES

- Como nunca há uma relação exata entre duas variáveis, como consideramos outros fatores que afetam y ?
- Qual é a relação funcional entre y e x ?
- Como podemos estar certos de que estamos capturando uma relação *ceteris paribus* (outros fatores constantes) entre y e x ?

MODELO DE REGRESSÃO LINEAR SIMPLES

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

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

- Terminologia:

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

VOLTANDO ÀS PERGUNTAS IMPORTANTES

- **Como nunca há uma relação exata entre duas variáveis, como consideramos outros fatores que afetam y ?**
 - Variável u é o termo erro ou perturbação da relação.
 - Na análise de regressão simples, todos fatores (além de x) que afetam y são tratados como não-observados.

OUTRA PERGUNTA

– Qual é a relação funcional entre y e x ?

- Se os outros fatores em u são mantidos fixos, de modo que a variação em u é zero ($\Delta u=0$), então x tem um efeito linear sobre y , tal como: $\Delta y=\beta_1\Delta x$; se $\Delta u=0$.
- A linearidade do modelo de regressão linear simples implica que uma variação de uma unidade em x tem o mesmo efeito sobre y , independentemente do valor inicial de x .
- Isso não é realista. Por exemplo, o próximo ano de escolaridade teria um efeito maior sobre os salários, em relação ao anterior. Esse problema será tratado adiante.

E O PROBLEMA DO *CETERIS PARIBUS*?

- Estamos capturando uma relação *ceteris paribus* (outros fatores constantes) entre y e x ?
 - A variação em y é β_1 multiplicado pela variação em x .
 - β_1 : **parâmetro de inclinação** da relação entre y e x , mantendo fixos os outros fatores em u .
 - β_0 : **parâmetro de intercepto** é raramente analisado.
 - β_1 mede o efeito de x sobre y , mantendo todos os outros fatores (em u) fixos.
 - No entanto, estamos ignorando todos os outros fatores.
 - Os estimadores de β_0 e β_1 serão confiáveis em uma amostra aleatória, se o termo não-observável (u) estiver relacionado à variável explicativa (x) de modo que o valor médio de u na população seja zero: $E(u)=0$.

HIPÓTESE SOBRE A RELAÇÃO ENTRE x E u

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

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

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

FUNÇÃO DE REGRESSÃO POPULACIONAL

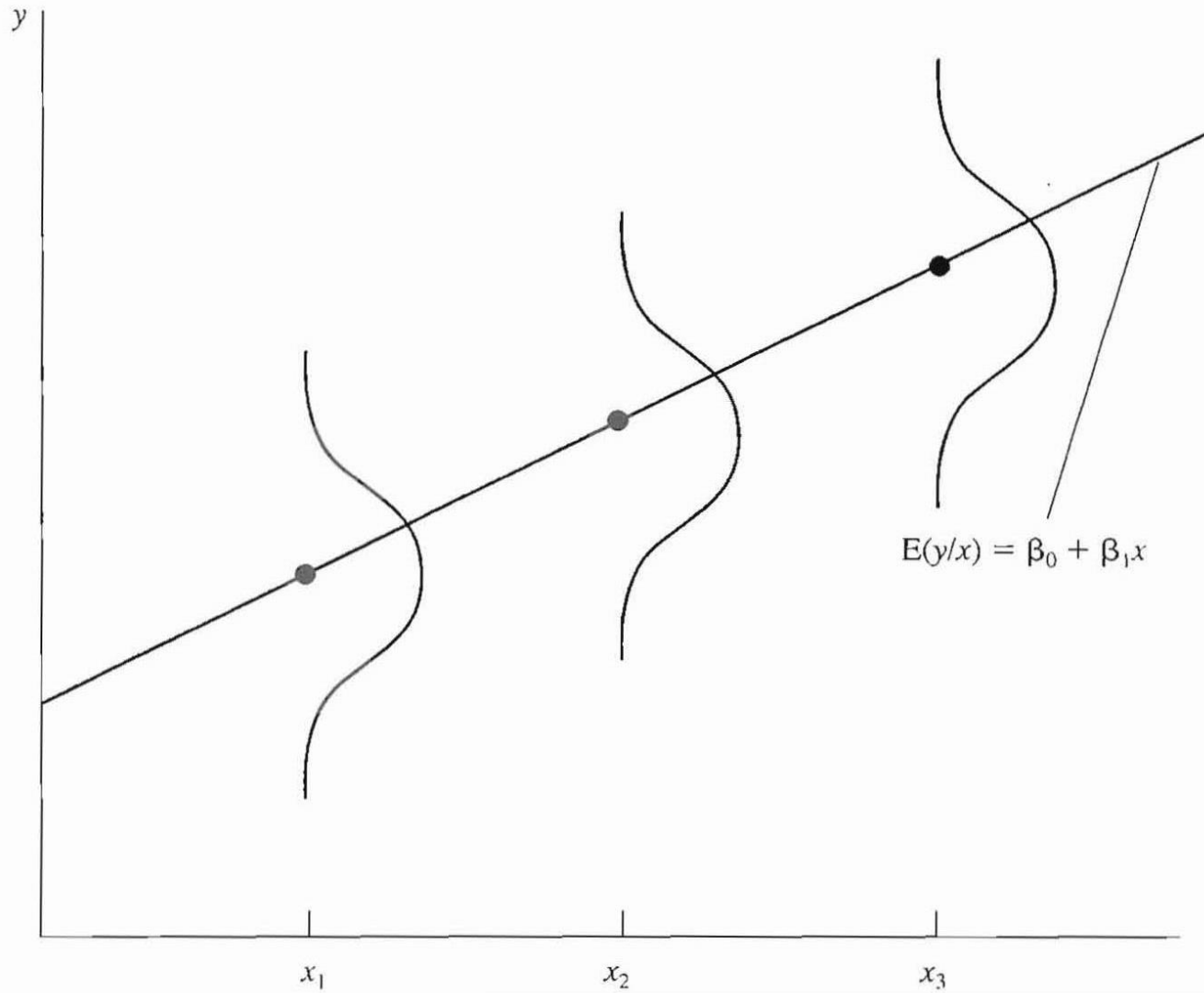
- Quando $E(u|x)=E(u)=0$ é verdadeiro, é útil dividir y em:
 - Parte sistemática (parte de y explicada por x): $\beta_0 + \beta_1 x$
 - Parte não-sistemática (parte de y não explicada por x): u
- Considerando o valor esperado de $y=\beta_0+\beta_1 x+u$ condicionado a x , e usando $E(u|x)=0$, temos a **função de regressão populacional** (FRP), que é uma função linear de x :

$$E(y|x) = \beta_0 + \beta_1 x$$

- **Linearidade**: o aumento de uma unidade em x faz com que o valor esperado de y varie segundo a magnitude de β_1 .
- Para qualquer valor de x , a distribuição de y está centrada ao redor de $E(y|x)$.

Figura 2.1

$E(y/x)$ como função linear de x .



ESTIMATIVA DE MÍNIMOS QUADRADOS ORDINÁRIOS

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

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

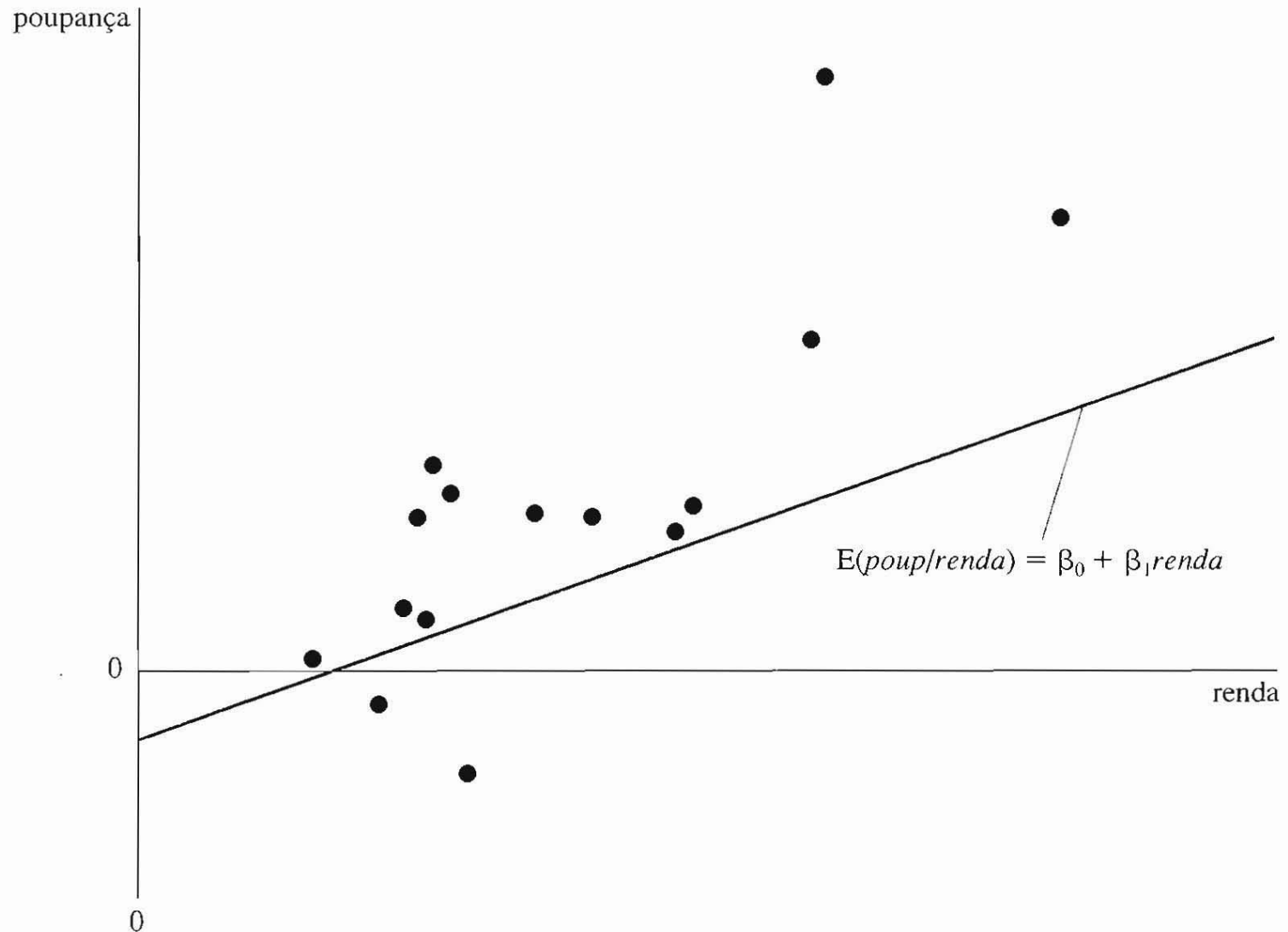
- A equação do modelo de regressão simples é escrito como:

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

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

Figura 2.2

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



ESTIMATIVA DE MÍNIMOS QUADRADOS ORDINÁRIOS

- Como obter estimativas do intercepto (β_0) e da inclinação (β_1) na regressão populacional da poupança sobre a renda?
- Na população, u tem média zero. O valor esperado de u é zero: $E(u)=0$
- Além disso, u é não-correlacionado com x . A covariância entre x e u é zero: $Cov(x,u)=E(xu)=0$
- $E(u)=0$ pode ser escrita como: $E(y-\beta_0-\beta_1x)=0$
- $Cov(x,u)=E(xu)=0$ pode ser escrita como: $E[x(y-\beta_0-\beta_1x)]=0$
- Como há dois parâmetros desconhecidos para estimar (β_0 e β_1), é possível utilizar uma amostra de dados para calcular as estimativas:

$$\hat{\beta}_0 \quad \text{e} \quad \hat{\beta}_1$$

EQUAÇÕES DA POPULAÇÃO E AMOSTRA

– Média de u na população:

$$E(y - \beta_0 - \beta_1 x) = 0$$

– Média de u na amostra:

$$\frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n} = 0$$

– Covariância entre x e u na população:

$$E[x(y - \beta_0 - \beta_1 x)] = 0$$

– Covariância entre x e u na amostra:

$$\frac{\sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n} = 0$$

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

$$\frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n} = 0$$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

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

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

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

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

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



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

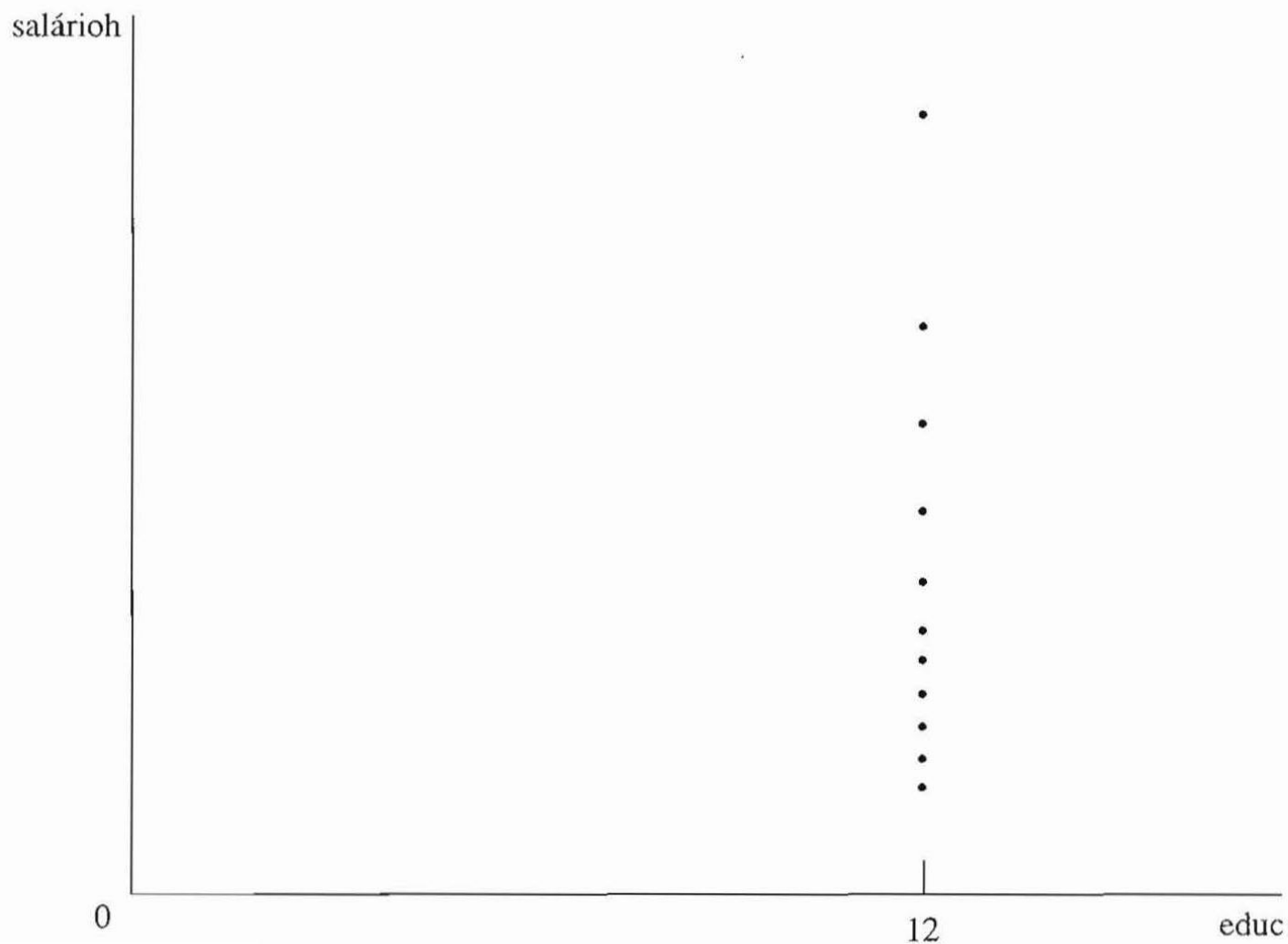
- Se x e y são positivamente correlacionados na amostra, $\hat{\beta}_1$ é positivo e vice-versa.

VARIÂNCIA DE x DEVE SER MAIOR QUE ZERO

- A hipótese necessária para calcular estimativas de mínimos quadrados ordinários (MQO) é que a variância amostral de x seja maior que zero.
- Ou seja, os valores de x_i na amostra não devem ser todos iguais a um mesmo valor.

Figura 2.3

Gráfico da dispersão de salários e educação, quando $educ_i = 12$ para todo i .



VALORES ESTIMADOS E RESÍDUOS

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

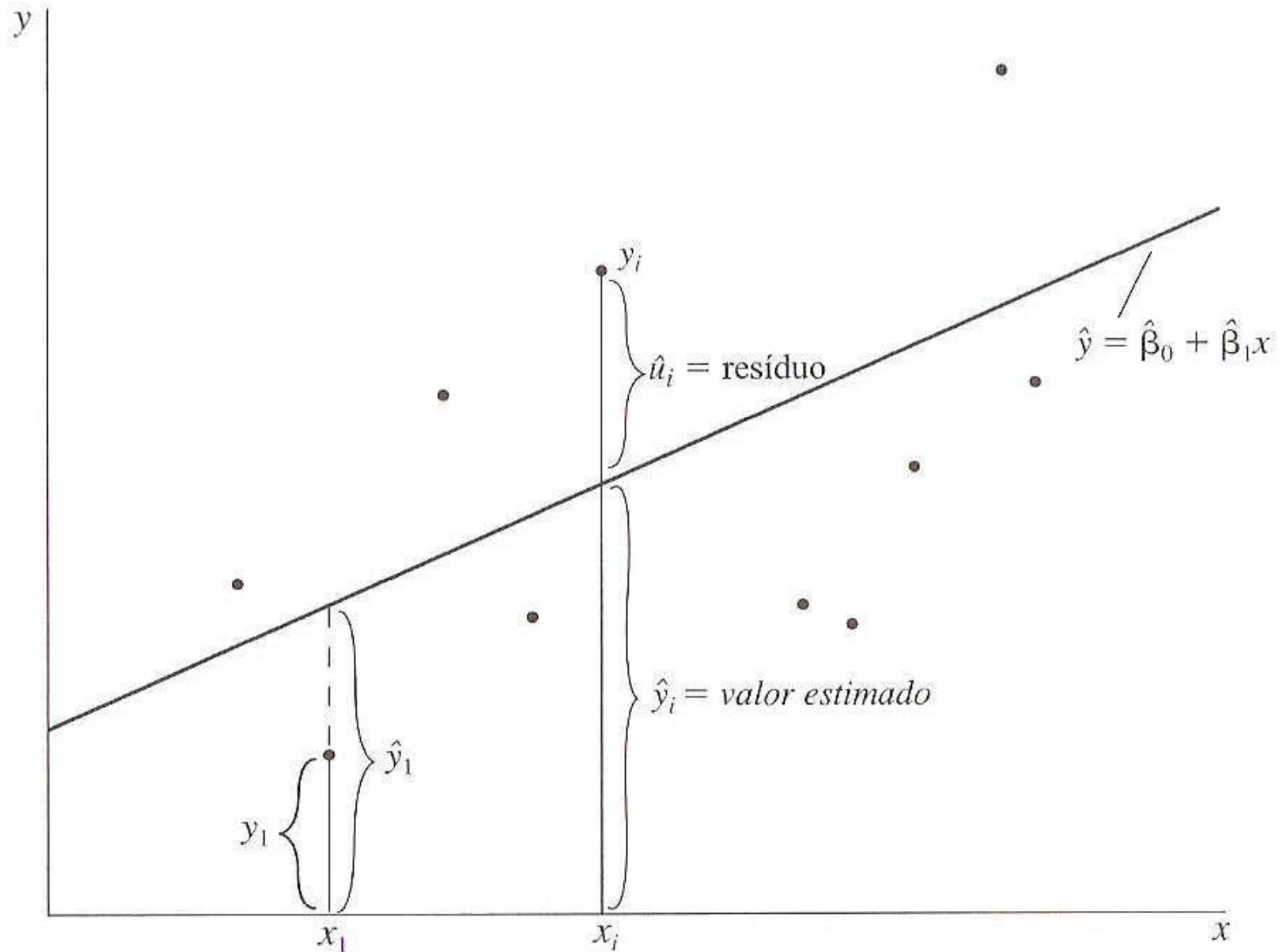
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

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

Figura 2.4

Valores estimados e resíduos.



MINIMIZANDO A SOMA DOS RESÍDUOS QUADRADOS

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

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

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

MINIMIZANDO A SOMA DOS RESÍDUOS QUADRADOS

- Reta de regressão de MQO ou função de regressão amostral (FRA) é a versão estimada da função de regressão populacional (FRP):

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

- O coeficiente de inclinação indica o quanto o valor estimado (previsto) de y varia quando x aumenta em uma unidade:

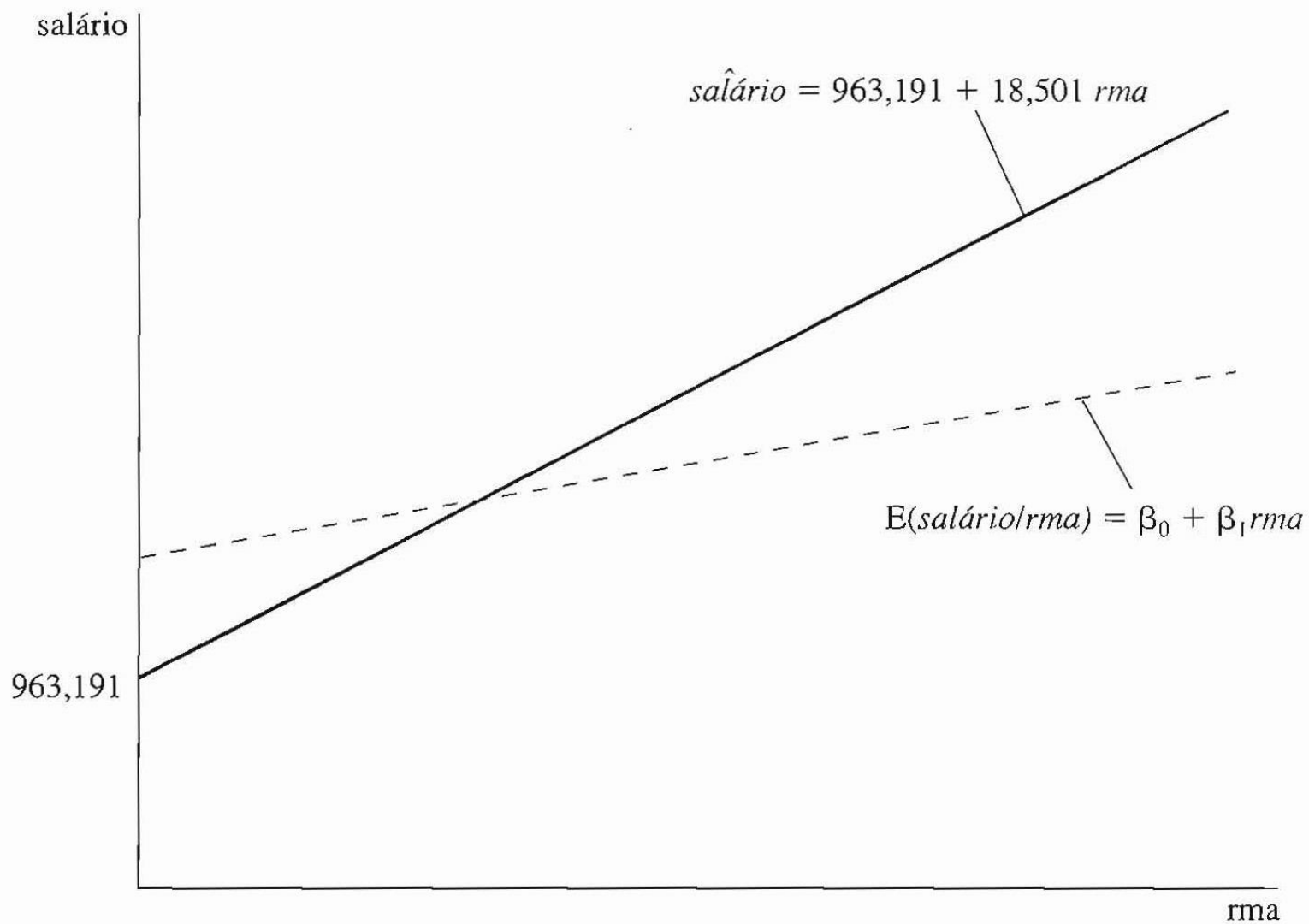
$$\hat{\beta}_1 = \Delta \hat{y} / \Delta x$$

- Da mesma forma, dada qualquer variação em x , podemos calcular a variação prevista em y :

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x$$

Figura 2.5

A reta de regressão de MQO $\hat{\text{salário}} = 963,191 + 18,501 rma$ e a função de regressão populacional (desconhecida).



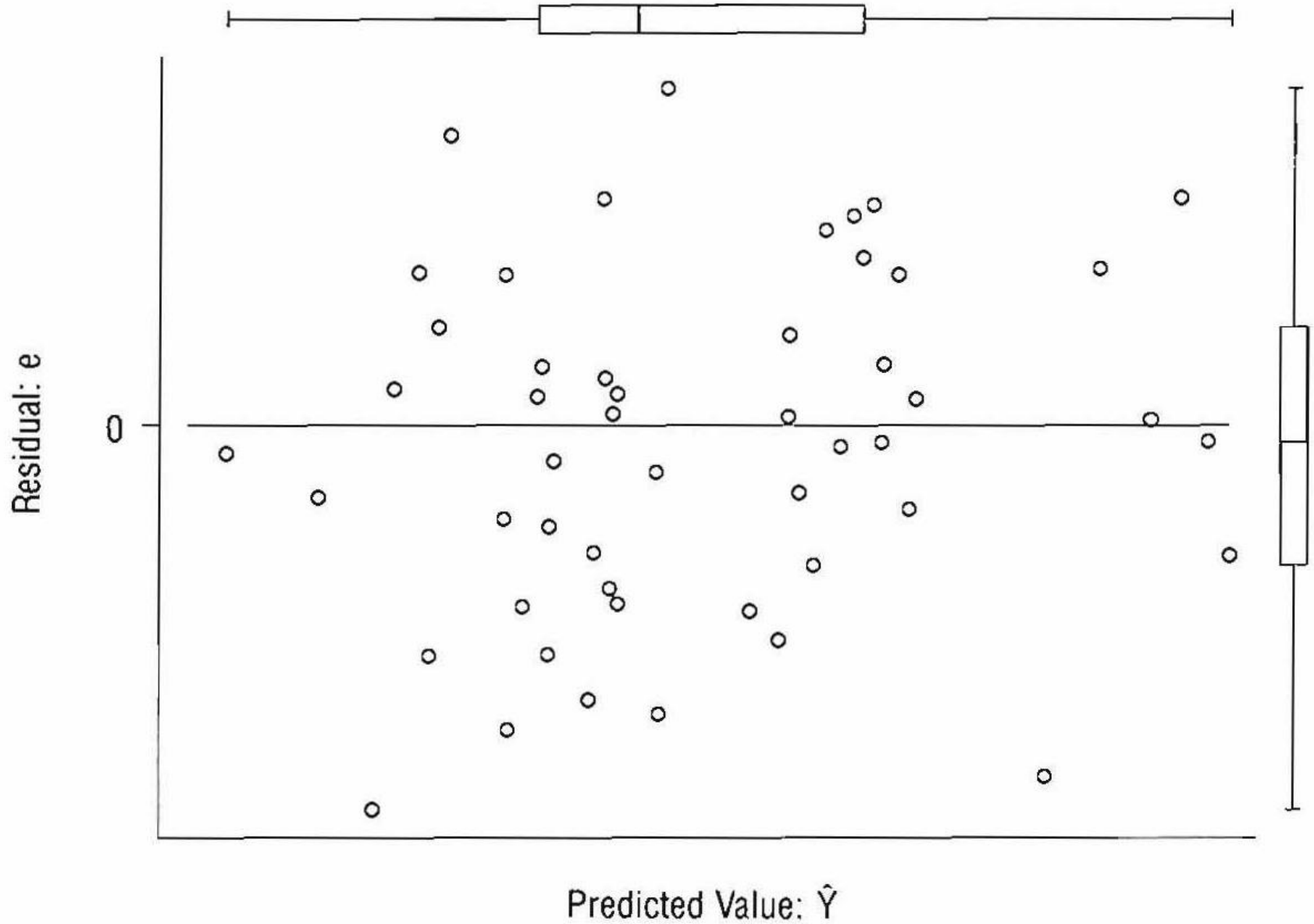
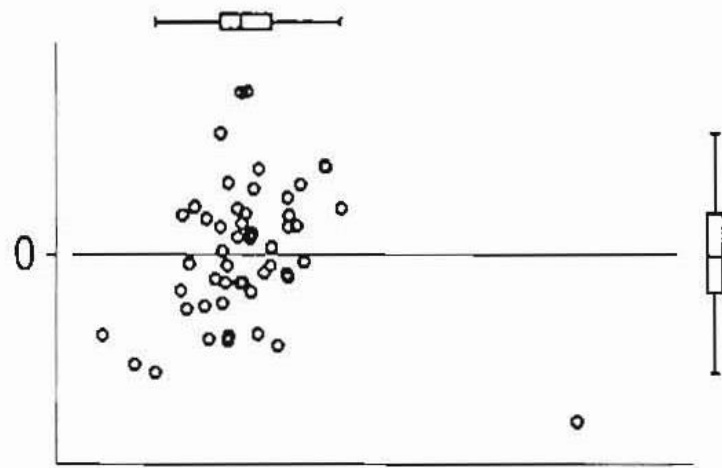
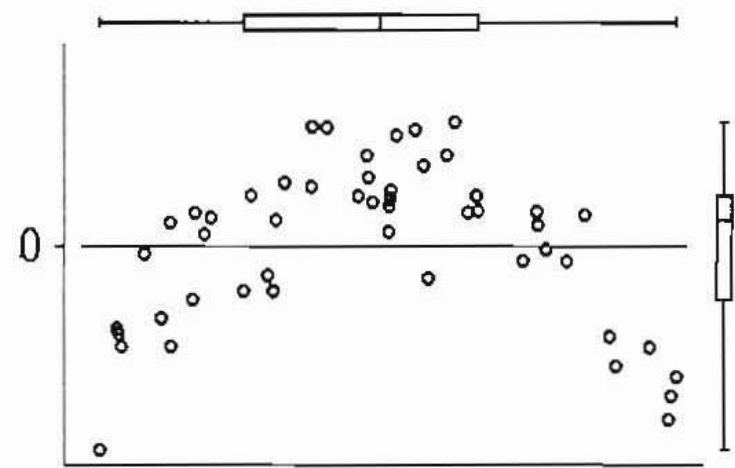


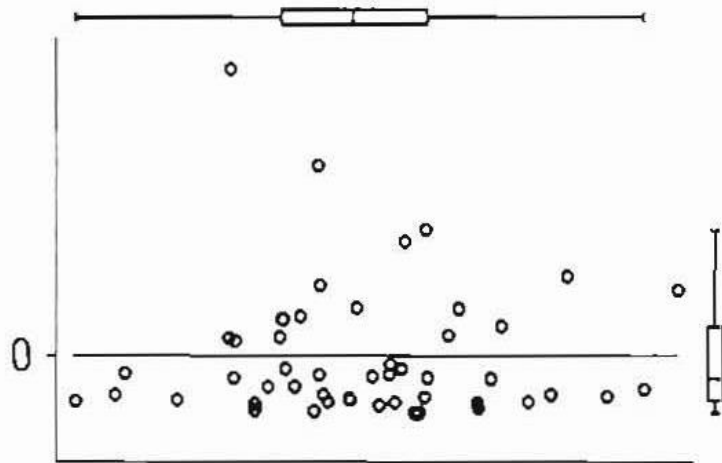
Figure 2.10 “All clear” e -versus- \hat{Y} plot (artificial data).



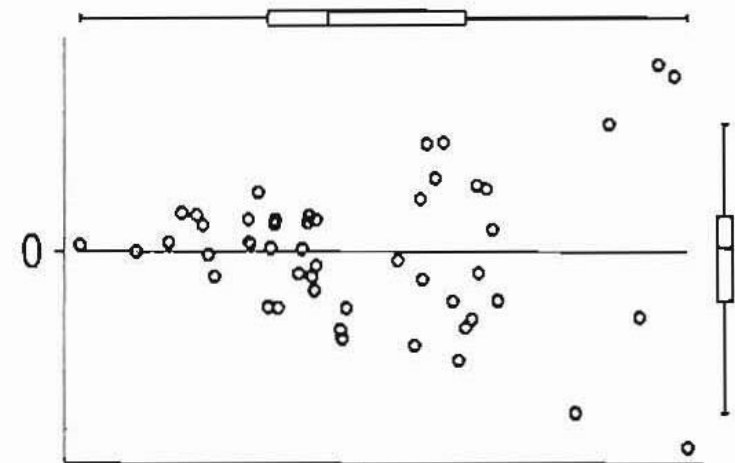
Influential Case



Curvilinear Relation



Nonnormal Residual Distribution



Heteroscedasticity

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

PROPRIEDADES ALGÉBRICAS DAS ESTATÍSTICAS

- A soma dos resíduos de MQO é zero, já que as estimativas de MQO de $\hat{\beta}_0$ e $\hat{\beta}_1$ são escolhidas para fazer com que a soma dos resíduos seja zero:

$$\sum_{i=1}^n \hat{u}_i = 0$$

- A covariância amostral entre os regressores e os resíduos de MQO é zero:

$$\frac{\sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}{n} = \sum_{i=1}^n x_i \hat{u}_i = 0$$

- Se inserirmos a média de x no lugar de x_i , o valor estimado é a média de y (este ponto está sempre sobre a reta):

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

SOMAS DOS QUADRADOS

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

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

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

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

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

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

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

$$SQT = SQE + SQR$$

GRAU DE AJUSTE

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

$$SQT = SQE + SQR$$

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

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

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

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

MUDANÇAS DAS UNIDADES DE MEDIDA

- Ao mudar unidades de medida das variáveis dependente e/ou independente, estimativas de MQO são afetadas.
- Se a **variável dependente** é multiplicada pela constante c (cada valor na amostra é multiplicado por c), então as estimativas de MQO de intercepto e de inclinação também são multiplicadas por c .
- Se a **variável independente** é dividida (ou multiplicada) por alguma constante diferente de zero (c) então o coeficiente de inclinação de MQO é multiplicado (ou dividido) por c , respectivamente.
- Mudar as unidades de medida da variável independente não afeta o intercepto.
- O grau de ajuste do modelo (R^2) não depende das unidades de medida das variáveis.

NÃO-LINEARIDADE NA REGRESSÃO SIMPLES

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

EXEMPLO DE NÃO-LINEARIDADE

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

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

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

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

- Se $\Delta u = 0$, então:

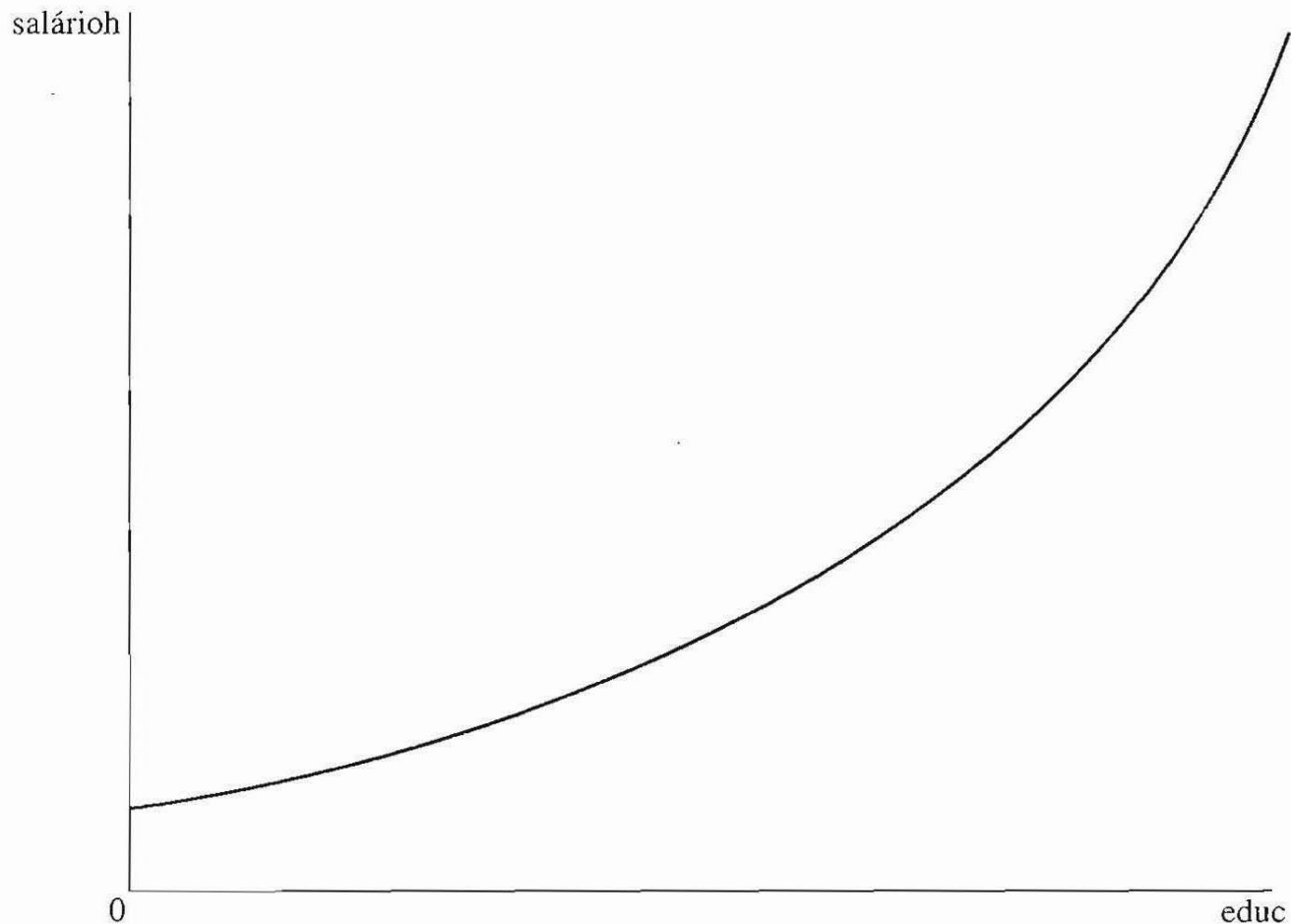
$$\% \Delta \textit{salário} = (100 * \beta_1) \Delta \textit{educ}$$

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

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

Figura 2.6

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



INTERPRETAÇÃO DOS COEFICIENTES

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

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

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

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

- Aumento de uma unidade em x aumenta y em $100 \cdot [\exp(\beta_1) - 1]\%$ ou, aproximadamente, em $(100 \cdot \beta_1)\%$:

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

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

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

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

FORMAS FUNCIONAIS ENVOLVENDO LOGARITMOS

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

SIGNIFICADO DE REGRESSÃO LINEAR

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

EXEMPLOS COM PNAD DE MINAS GERAIS DE 2007

- O banco de dados de pessoas possui informação de anos de escolaridade (anest), rendimento no trabalho principal (renpri) e logaritmo do rendimento no trabalho principal (lnrenpri).

	anest	renpri	lnrenpri
1	4	380	5,940171
2	4	530	6,272877
3	11	800	6,684612
4	6	350	5,857933
5	11	1600	7,377759
6	11	743	6,610696
7	11	500	6,214608
8	14	580	6,363028
9	4	380	5,940171
10	11	400	5,991465
11	11	8000	8,987197
12	8	459	6,12905
13	8	380	5,940171
14	0	120	4,787492
15	8	600	6,39693
16	8	550	6,309918
17	8	600	6,39693
18	10	400	5,991465
19	4	380	5,940171
20	4	380	5,940171

...

EXEMPLO 1: PNAD DE MINAS GERAIS DE 2007

– Escolaridade explicando rendimento:

```
. reg renpri anest [aweight=v4729]
(sum of wgt is 8.7563e+06)
```

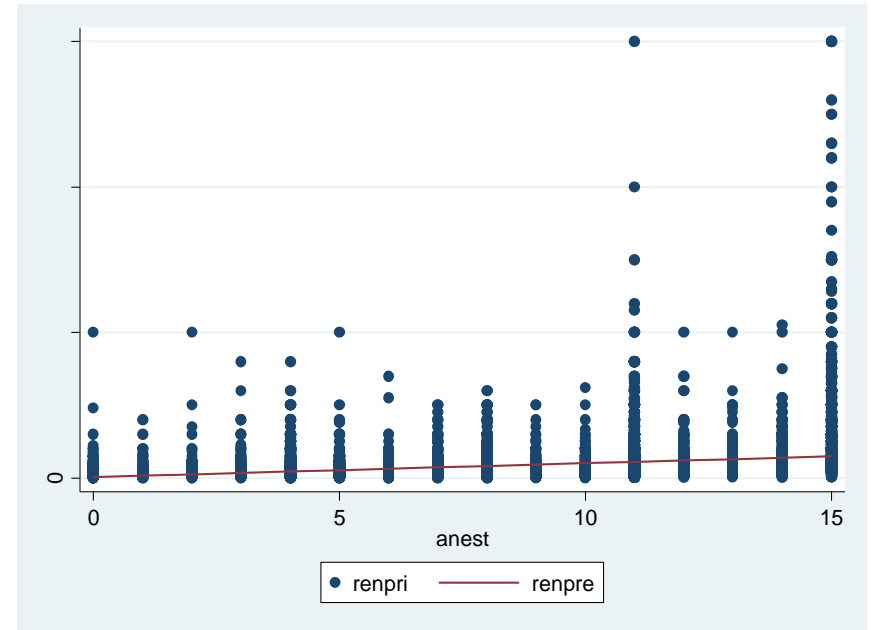
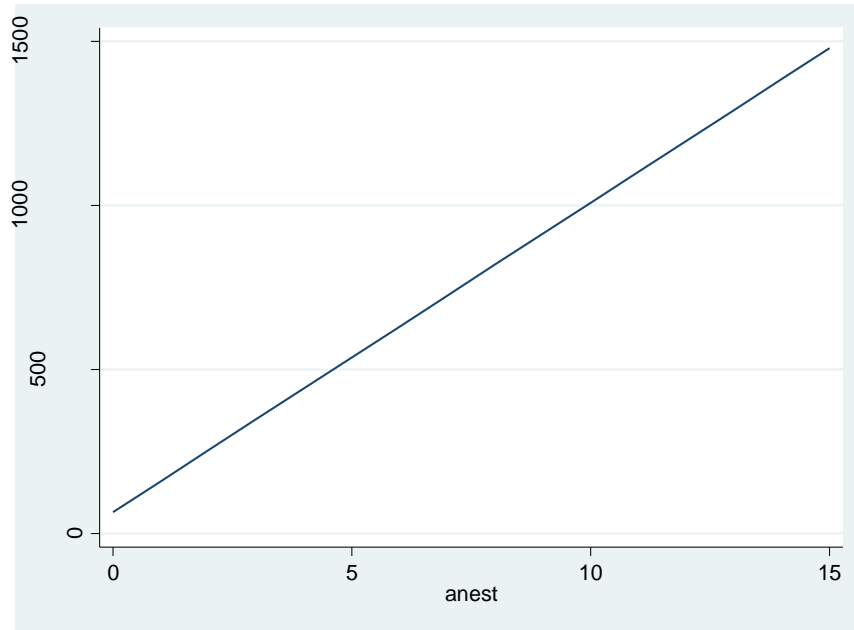
Source	SS	df	MS
Model	2.5086e+09	1	2.5086e+09
Residual	2.3809e+10	16230	1466951.75
Total	2.6317e+10	16231	1621416.61

```
Number of obs = 16232
F( 1, 16230) = 1710.07
Prob > F      = 0.0000
R-squared     = 0.0953
Adj R-squared = 0.0953
Root MSE     = 1211.2
```

renpri	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
anest	94.24418	2.279019	41.35	0.000	89.77705 98.71131
_cons	65.81278	20.36991	3.23	0.001	25.88551 105.7401

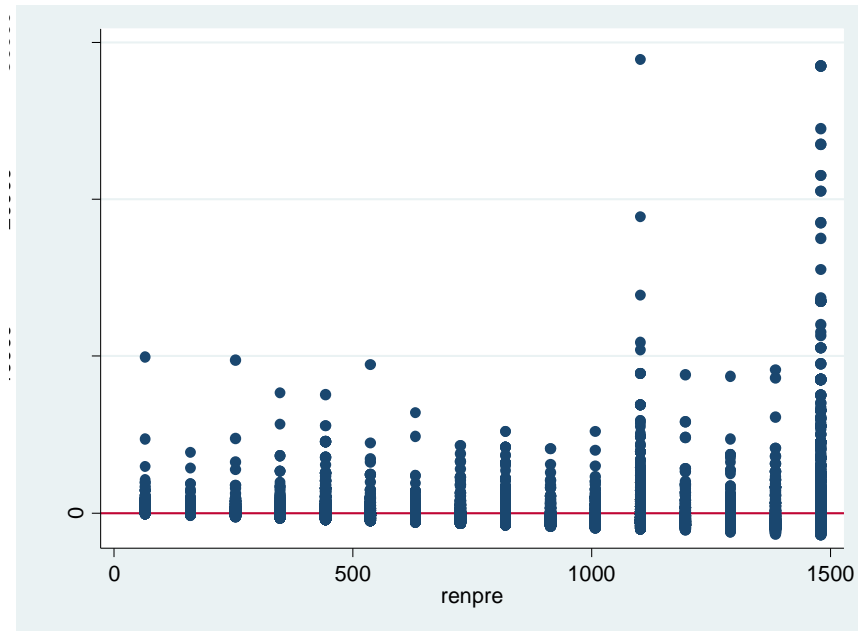
EXEMPLO 1: PNAD DE MINAS GERAIS DE 2007

– Renda predita por anos de escolaridade:



EXEMPLO 1: PNAD DE MINAS GERAIS DE 2007

– Resíduos por renda predita:



EXEMPLO 2: PNAD DE MINAS GERAIS DE 2007

– Escolaridade explicando logaritmo do rendimento:

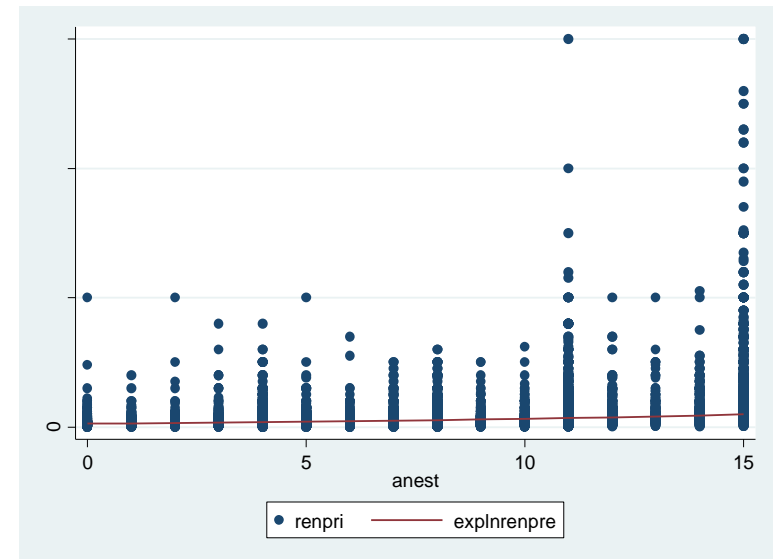
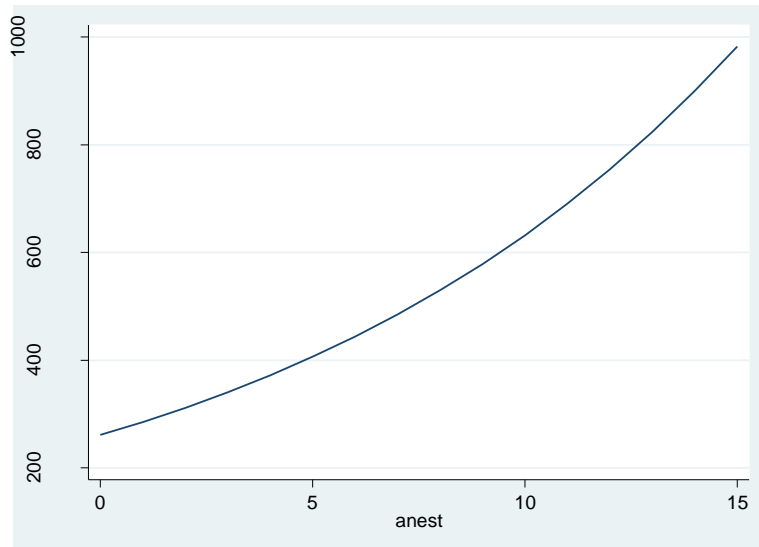
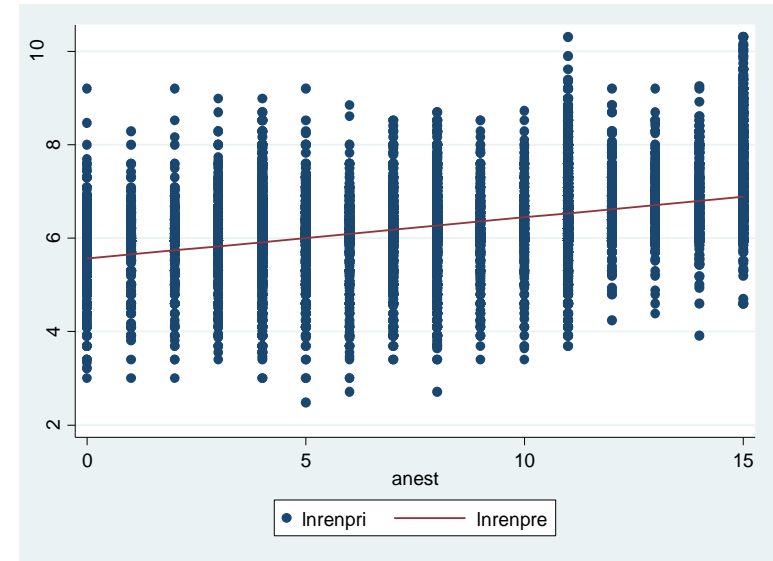
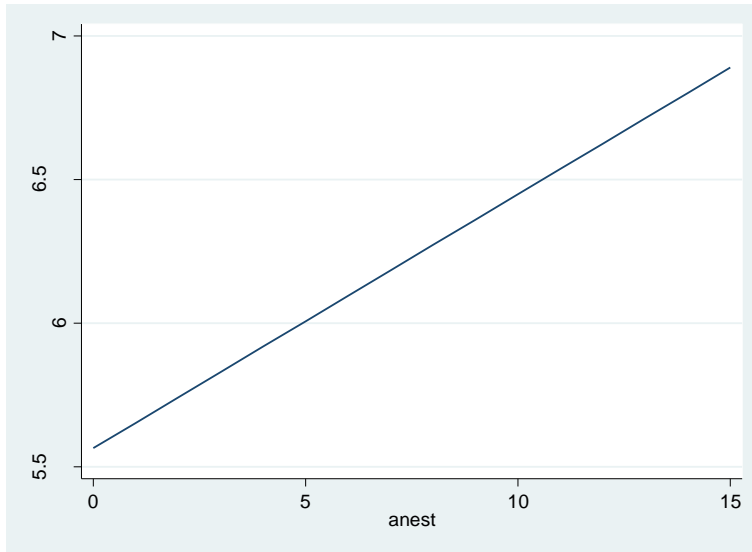
```
. reg lnrenpri anest [aweight=v4729]
(sum of wgt is 8.7563e+06)
```

Source	SS	df	MS			
Model	2204.86541	1	2204.86541	Number of obs =	16232	
Residual	10035.5653	16230	.618334278	F(1, 16230) =	3565.81	
Total	12240.4307	16231	.754139039	Prob > F =	0.0000	
				R-squared =	0.1801	
				Adj R-squared =	0.1801	
				Root MSE =	.78634	

lnrenpri	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
anest	.088355	.0014796	59.71	0.000	.0854548	.0912552
_cons	5.565065	.0132249	420.80	0.000	5.539142	5.590987

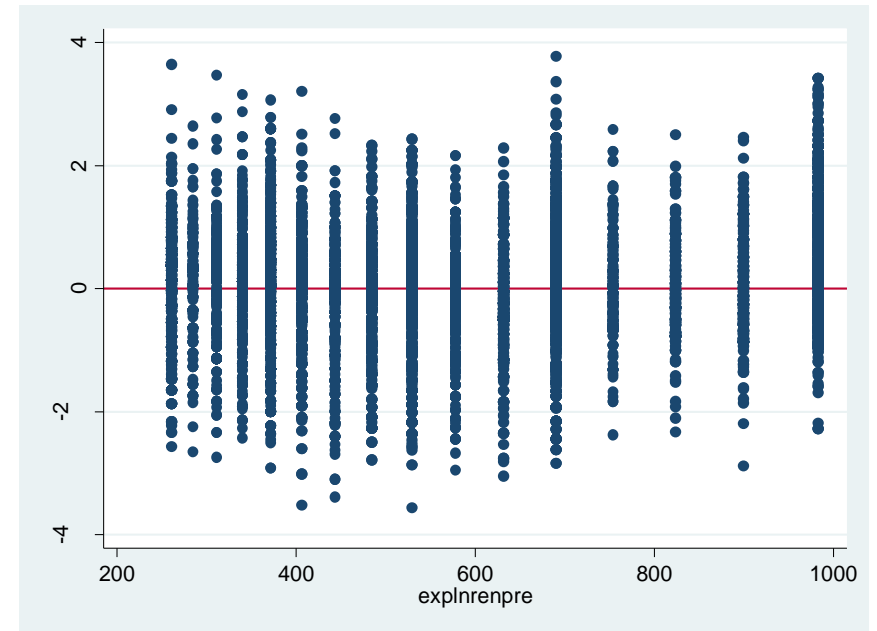
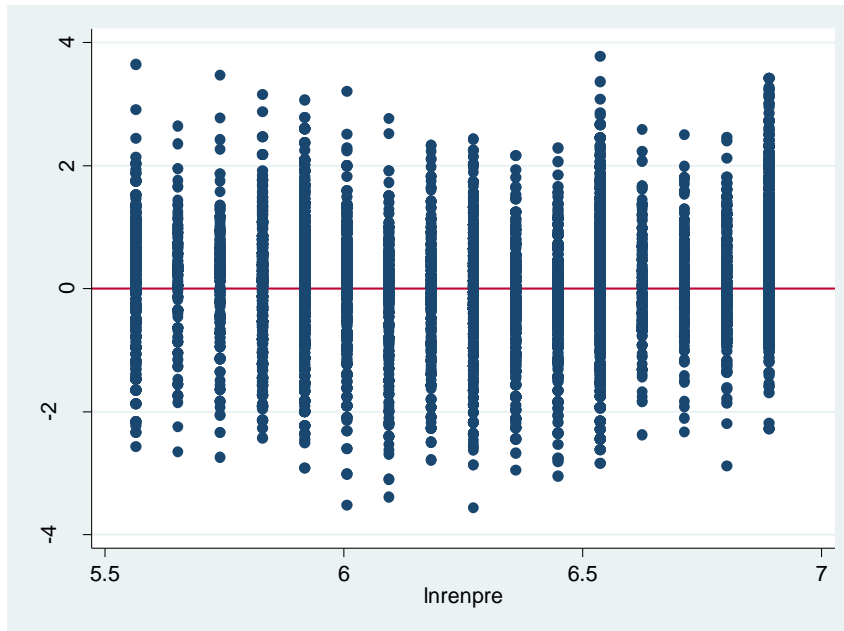
EXEMPLO 2: PNAD DE MINAS GERAIS DE 2007

– Renda predita por anos de escolaridade:



EXEMPLO 2: PNAD DE MINAS GERAIS DE 2007

– Resíduos por renda predita:



GRÁFICOS FORAM GERADOS COM ESTAS VARIÁVEIS

- Cálculo do valor predito: $y\text{-predito} = \beta_0 + \beta_1 x$
- Cálculo do resíduo: $u = y\text{-observado} - y\text{-predito}$
- Na 2ª regressão, calculamos ainda o exponencial do predito.

	anest	renpri	renpre	renres	lnrenpri	lnrenpre	lnrenres	exp lnrenpre
1	4	380	442.7895	-62.78951	5.940171	5.918485	.0216866	371.8478
2	4	530	442.7895	87.21049	6.272877	5.918485	.3543926	371.8478
3	11	800	1102.499	-302.4988	6.684612	6.53697	.1476422	690.1919
4	6	350	631.2779	-281.2779	5.857933	6.095195	-.2372616	443.7205
5	11	1600	1102.499	497.5012	7.377759	6.53697	.8407894	690.1919
6	11	743	1102.499	-359.4988	6.610696	6.53697	.0737262	690.1919
7	11	500	1102.499	-602.4988	6.214608	6.53697	-.3223614	690.1919
8	14	580	1385.231	-805.2313	6.363028	6.802034	-.4390065	899.6757
9	4	380	442.7895	-62.78951	5.940171	5.918485	.0216866	371.8478
10	11	400	1102.499	-702.4988	5.991465	6.53697	-.545505	690.1919
11	11	8000	1102.499	6897.501	8.987197	6.53697	2.450227	690.1919
12	8	459	819.7662	-360.7662	6.12905	6.271904	-.1428544	529.4848
13	8	380	819.7662	-439.7662	5.940171	6.271904	-.3317334	529.4848
14	0	120	65.81278	54.18722	4.787492	5.565065	-.7775729	261.1422
15	8	600	819.7662	-219.7662	6.39693	6.271904	.1250251	529.4848
16	8	550	819.7662	-269.7662	6.309918	6.271904	.0380138	529.4848
17	8	600	819.7662	-219.7662	6.39693	6.271904	.1250251	529.4848
18	10	400	1008.255	-608.2546	5.991465	6.448615	-.45715	631.8264
19	4	380	442.7895	-62.78951	5.940171	5.918485	.0216866	371.8478
20	4	380	442.7895	-62.78951	5.940171	5.918485	.0216866	371.8478