AULA 01
Introdução & Modelo de regressão simples

Ernesto F. L. Amaral

15 de julho de 2013
Análise de Regressão Linear (MQ 2013)
www.ernestoamaral.com/mq13reg.html

Fonte:

Wooldridge, Jeffrey M. "Introdução à econometria: uma abordagem moderna". São Paulo: Cengage Learning, 2008. Capítulo 1 (1-17) e Capítulo 2 (pp.19-63).

ESTRUTURA DO LIVRO

– **Introdução:** principais conceitos em econometria (capítulo 1).

– **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

DOCUMENTAÇÃO PARA EXERCÍCIO

– Vamos utilizar a Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2007 de Minas Gerais para as demonstrações em sala de aula e a PNAD de 2011 do Brasil para o exercício final do curso.

– Os bancos de dados, questionário, livro de códigos e demais arquivos estão disponíveis no site do Instituto Brasileiro de Geografia e Estatística (IBGE):

http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2011/microdados.shtm

CAPÍTULO 1 - WOOLDRIDGE
INTRODUÇÃO:
PRINCIPAIS CONCEITOS EM ECONOMETRIA

ECONOMETRIA

– A econometria evoluiu como uma disciplina separada da estatística matemática, porque enfoca problemas inerentes à coleta e à análise de dados econômicos não-experimentais.

– **Dados não-experimentais** não são acumulados por meio de experimentos controlados de indivíduos, firmas ou segmentos da economia.

– Dados não-experimentais são também chamados de **dados observacionais** para enfatizar o fato de que o pesquisador é um coletor passivo de dados.

– **Dados experimentais** são frequentemente coletados em ambientes de laboratório nas ciências naturais, mas são muito mais difíceis de serem obtidos nas ciências sociais.

– O método de análise da **regressão múltipla** é utilizado por econometristas e estatísticos matemáticos, mas o foco e interpretação pode diferir significantemente.

ANÁLISE ECONÔMICA EMPÍRICA

– Os métodos econométricos são usados para testar uma teoria econômica ou para analisar relações que apresentam importância para análises de políticas públicas.

– Uma análise empírica usa dados para testar uma teoria ou estimar uma relação.

– O primeiro passo em qualquer análise empírica é a formulação cuidadosa da questão de interesse, a qual pode ser a de testar efeitos de uma política governamental ou, até mesmo, de testar hipóteses e teorias.

– O modelo econômico formal consiste em equações matemáticas que descrevem relações para testar teorias.

MICROECONOMIA

– Os indivíduos fazem escolhas para maximizar seu bem-estar (**maximização da utilidade**), sujeitas às restrições de recursos.

– Isso oferece um arcabouço para criar modelos econômicos para fazer previsões entre variáveis.

– A maximização da utilidade leva a um conjunto de **equações de demanda**, no contexto das decisões de consumo.

– Em uma equação de demanda, a quantidade demandada de cada produto depende do seu próprio preço, do preço dos bens substitutos e complementares, da renda do consumidor e das características individuais que influem no gosto.

MODELO ECONÔMICO

– O modelo econômico é a formulação teórica de uma relação entre variáveis econômicas.

– A quantidade de tempo gasto na atividade criminosa é uma função de vários fatores (Gary Becker 1968):

$$y=f(x_1, x_2, x_3, x_4, x_5, x_6, x_7),$$

y = horas gastas em atividades criminosas.

x_1 = "salário" por hora ocupada em atividade criminosa.

x_2 = salário-hora em emprego legal.

x_3 = renda de outras atividades que não o crime ou um emprego legal.

x_4 = probabilidade de ser capturado.

x_5 = probabilidade de ser condenado se capturado.

x_6 = sentença esperada se condenado.

x_7 = idade.

MODELO ECONOMÉTRICO

– Após elaborar o modelo econômico, é especificado um modelo econométrico, que será aplicado a dados existentes.

– A forma da função *f(.)* deveria ser especificada antes de realizar uma análise econométrica.

– Se uma variável não pode ser obtida, é possível utilizar uma variável que se aproxima desta que se quer medir (***proxy***).

– Outros fatores são considerados no termo de erro *u* (ou termo de disturbância):

  – **Erro amostral** é a diferença entre o resultado amostral e o verdadeiro resultado da população (devidos ao acaso).

  – **Erro não-amostral** ocorre quando os dados amostrais são coletados, registrados ou analisados incorretamente.

– Modelo econométrico de Becker (1968):

$$crime = \beta_0 + \beta_1 salário + \beta_2 outrenda + \beta_3 freqpris + \beta_4 freqcond + \beta_5 sentmed + \beta_6 idade + u$$

MODELO ECONOMÉTRICO NA PRÁTICA

– Na maioria dos casos, a análise econométrica começa pela especificação de um modelo econométrico, sem consideração de detalhes da criação do modelo econômico.

– É comum começar com um modelo econométrico e usar o raciocínio econômico e conhecimentos científicos como guias para escolher as variáveis.

– Após a especificação do modelo econométrico, várias hipóteses podem ser formuladas em termos das direções e influências dos parâmetros desconhecidos (independentes) sobre a variável de interesse (dependente).

– Após os dados terem sido coletados, os métodos econométricos são usados para estimar os parâmetros do modelo econométrico e para testar as hipóteses de interesse.

DESENHOS BÁSICOS DE *SURVEY*: BANCOS DE DADOS

– Após especificar os objetivos e unidades de análise da pesquisa, é preciso escolher entre diversos desenhos diferentes:

  – *Surveys* interseccionais (*cross-sectional*).
  – *Surveys* longitudinais (tendências, coortes ou painel).
  – *Surveys* interseccionais servindo como longitudinais.

– Wooldridge (2008) classifica os dados econômicos em:

  – Dados de corte transversal = *surveys* interseccionais.
  – Cortes transversais agrupados = estudos de tendências.
  – Dados de séries de tempo = estudos de coortes.
  – Dados de painel ou longitudinais = estudos de painel.

DADOS DE CORTE TRANSVERSAL (Wooldridge) *SURVEYS* INTERSECCIONAIS (Babbie)

– Um conjunto de dados de corte transversal consiste em uma amostra de uma unidade de análise, tomada em um determinado ponto no tempo.

– Esses dados são muito utilizados em economia e em outras ciências sociais.

– Dados em um determinado ponto do tempo são importantes para testar hipóteses e avaliar políticas.

– Dados podem ter problemas de seleção amostral, no caso de determinados indivíduos não revelarem informações acuradas.

– Amostragem deve ser realizada de forma acurada para evitar que coleta se concentre em unidades com características semelhantes.

EXEMPLO DE DADOS DE CORTE TRANSVERSAL

– Conjunto de dados de corte transversal para o ano de 1976 de 526 trabalhadores (Wooldridge 2008):

Número da observação	Salário por hora	Anos de escolaridade	Anos de experiência no mercado de trabalho	Feminino	Estado civil (casado)
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
...	...	...	...	...	...
525	11,56	16	5	0	1
526	3,50	14	5	1	0

CORTES TRANSVERSAIS AGRUPADOS (Wooldridge) ESTUDOS DE TENDÊNCIAS (Babbie)

– Uma população pode ser amostrada e estudada em ocasiões diferentes.

– Um mesmo conjunto de variáveis é coletado em diferentes períodos do tempo, em **distintas** amostras aleatórias de uma mesma população (Censo Demográfico, Pesquisa Nacional por Amostra de Domicílios – PNAD).

– Agrupar cortes transversais de diferentes anos é eficaz para analisar os efeitos de uma política pública.

– O ideal é coletar dados de anos anteriores e posteriores a uma importante mudança de política governamental.

– Além de aumentar o tamanho da amostra, a análise de corte transversal agrupada é importante para estimar como uma relação fundamental mudou ao longo do tempo.

– Geralmente são utilizados dados secundários, coletados por outros pesquisadores ou instituições.

EXEMPLO DE CORTES TRANSVERSAIS AGRUPADOS

– Conjunto de dados sobre os preços da moradia em 1993 e 1995 nos Estados Unidos (Wooldridge 2008):

Número da observação	Ano	Preço comercializado	Imppro	Arquad	Quantidade de dormitórios	Quantidade de banheiros
1	1993	85.500	42	1.600	3	2,0
2	1993	67.300	36	1.440	3	2,5
3	1993	134.000	38	2.000	4	2,5
...	...	...	...	...		...
250	1993	243.600	41	2.600	4	3,0
251	1995	65.000	16	1.250	2	1,0
252	1995	182.400	20	2.200	4	2,0
253	1995	97.500	15	1.540	3	2,0
...	...	...	...	...	...	...
520	1995	57.200	16	1.100	2	1,5

DADOS DE SÉRIES DE TEMPO (Wooldridge)
ESTUDOS DE COORTES (Babbie)

– Um conjunto de dados de séries de tempo consiste em observações sobre variáveis ao longo do tempo.

– Como eventos passados podem influenciar eventos futuros, o tempo é uma dimensão importante em um conjunto de dados de séries de tempo.

– A análise desses dados pode ser dificultada, porque observações econômicas não são independentes ao longo do tempo (variáveis possuem padrões sazonais).

– Há uma série de frequências possíveis: diárias, semanais, mensais, trimestrais, anuais, decenais...

– Estes dados são também chamados de estudos de coorte, em que mesma população é analisada, mas amostras estudadas podem ser diferentes:

  – Pessoas com 10 anos em 2000, 20 anos em 2010, 30 anos em 2020, 40 anos em 2030...

EXEMPLO DE DADOS DE SÉRIES DE TEMPO

– Conjunto de dados de séries de tempo sobre efeitos do salário mínimo em Porto Rico (apud Wooldridge 2008):

Número da observação	Ano	Salário mínimo médio no ano	Taxa de trabalhadores cobertos pela lei de salário mínimo	Taxa de desemprego	Produto Nacional Bruto (PNB)
1	1950	0,20	20,1	15,4	878,7
2	1951	0,21	20,7	16,0	925,0
3	1952	0,23	22,6	14,8	1.015,9
...	...	...	...	...	...
37	1986	3,35	58,1	18,9	4.281,6
38	1987	3,35	58,2	16,8	4.496,7

DADOS DE PAINEL OU LONGITUDINAIS (Wooldridge) ESTUDOS DE PAINEL (Babbie)

– Um conjunto de dados de painel consiste em uma série de tempo para **cada** membro do corte transversal.

– Os dados de painel são distintos dos dados de corte transversal agrupados (tendências) e de séries de tempo (coortes), porque as **mesmas** unidades são acompanhadas ao longo de um determinado período.

– Dados de painel podem ser coletados para indivíduos, domicílios, instituições ou unidades geográficas.

– Esses dados são os mais sofisticados para fins explicativos, mas são mais difíceis e caros de se obter.

– Pode haver problema de grande número de não respostas nas últimas ondas de entrevistas.

– A análise dos dados pode se tornar complicada quando se tentar avaliar as mudanças dos indivíduos no tempo.

EXEMPLO DE DADOS DE PAINEL OU LONGITUDINAIS

– Conjunto de dados de painel sobre crime e estatísticas relacionadas em 1986 e 1990 em 150 cidades nos Estados Unidos (Wooldridge 2008):

Número da observação	Cidade	Ano	Homicídios	População	Desemprego	Polícia
1	1	1986	5	350.000	8,7	440
2	1	1990	8	359.200	7,2	471
3	2	1986	2	64.300	5,4	75
4	2	1990	1	65.100	5,5	75
...	...	...	...	...	...	...
297	149	1986	10	260.700	9,6	286
298	149	1990	6	245.000	9,8	334
299	150	1986	25	543.000	4,3	520
300	150	1990	32	546.200	5,2	493

CORTE TRANSVERSAL USADO COMO LONGITUDINAL

- Alguns mecanismos podem ser utilizados num *survey* interseccional (corte transversal) para aproximar o estudo de processo ou mudança (longitudinal).
- Podem ser realizadas perguntas referentes ao passado (renda no ano anterior, local de residência anterior):
  - Há problemas de erro de memória.
  - Os dados devem ser interpretados como amostra da população atual, e não de população passada.
- Por exemplo, é possível utilizar um único banco de dados de corte transversal para comparar pessoas de diferentes idades (jovens e idosos) e coortes (calouros e veteranos).

VARIAÇÕES DOS DESENHOS BÁSICOS

– Os desenhos básicos de pesquisa apresentados anteriormente podem ser modificados para se enquadrarem aos objetivos de um estudo:

  – **Amostras paralelas:** amostras separadas de populações diferentes, utilizando mesmo questionário (exemplo é a pesquisa sobre preconceito na UFMG).

  – **Estudos contextuais:** uso de dados sobre o ambiente ou meio da pessoa para descrever o contexto do indivíduo.

  – **Estudos sociométricos:** intenção é de observar as inter-relações entre membros da população estudada (redes de amizades, por exemplo).

ESCOLHENDO O DESENHO APROPRIADO

- **Dados de corte transversal** são mais apropriados se objetivo é descrição de tempo único.
- **Mudanças ao longo do tempo** são mais difíceis de realizar, porque dados de painel exigem tempo e recursos:
  - É possível utilizar dados de corte transversal e comparar pessoas que passaram por uma experiência no passado, com aqueles que não passaram.
- **Estudos de painel** são mais viáveis economicamente quando o fenômeno estudado tem duração curta (por exemplo, opinião de voto durante uma campanha eleitoral).
- **Estudos de tendências** podem ser realizados quando dados antigos são complementados com dados coletados pelo pesquisador.

CAUSALIDADE

– Na avaliação de políticas públicas, o objetivo do pesquisador é inferir que uma variável tem um **efeito causal** sobre outra variável.

– Encontrar uma associação entre duas ou mais variáveis pode ser sugestivo (correlação), mas somente será convincente se for possível estabelecer uma causalidade.

– A noção de ***ceteris paribus*** é importante, já que significa "outros fatores (relevantes) permanecendo iguais".

– Se outros fatores não forem mantidos fixos, não poderemos conhecer o efeito causal de uma variável sobre outra.

– Como a maioria dos dados coletados nas ciências sociais são não-experimentais (não são experimentos controlados como nas ciências naturais), descobrir relações causais é uma tarefa complexa.

CAPÍTULO 2 - WOOLDRIDGE
MODELO DE REGRESSÃO SIMPLES

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:

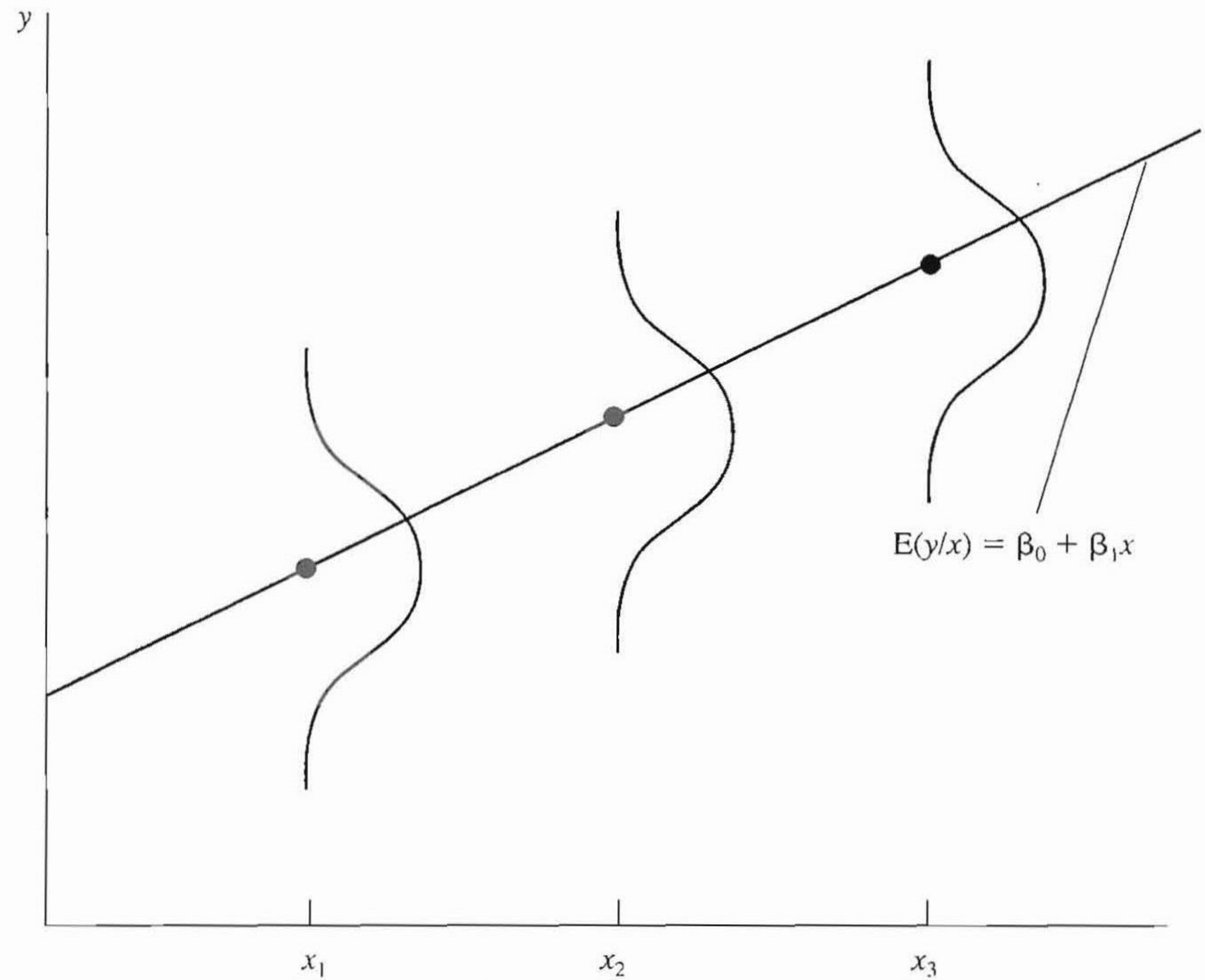
$$E(y|x) = \beta_0 + \beta_1 x$$

– **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:

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

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

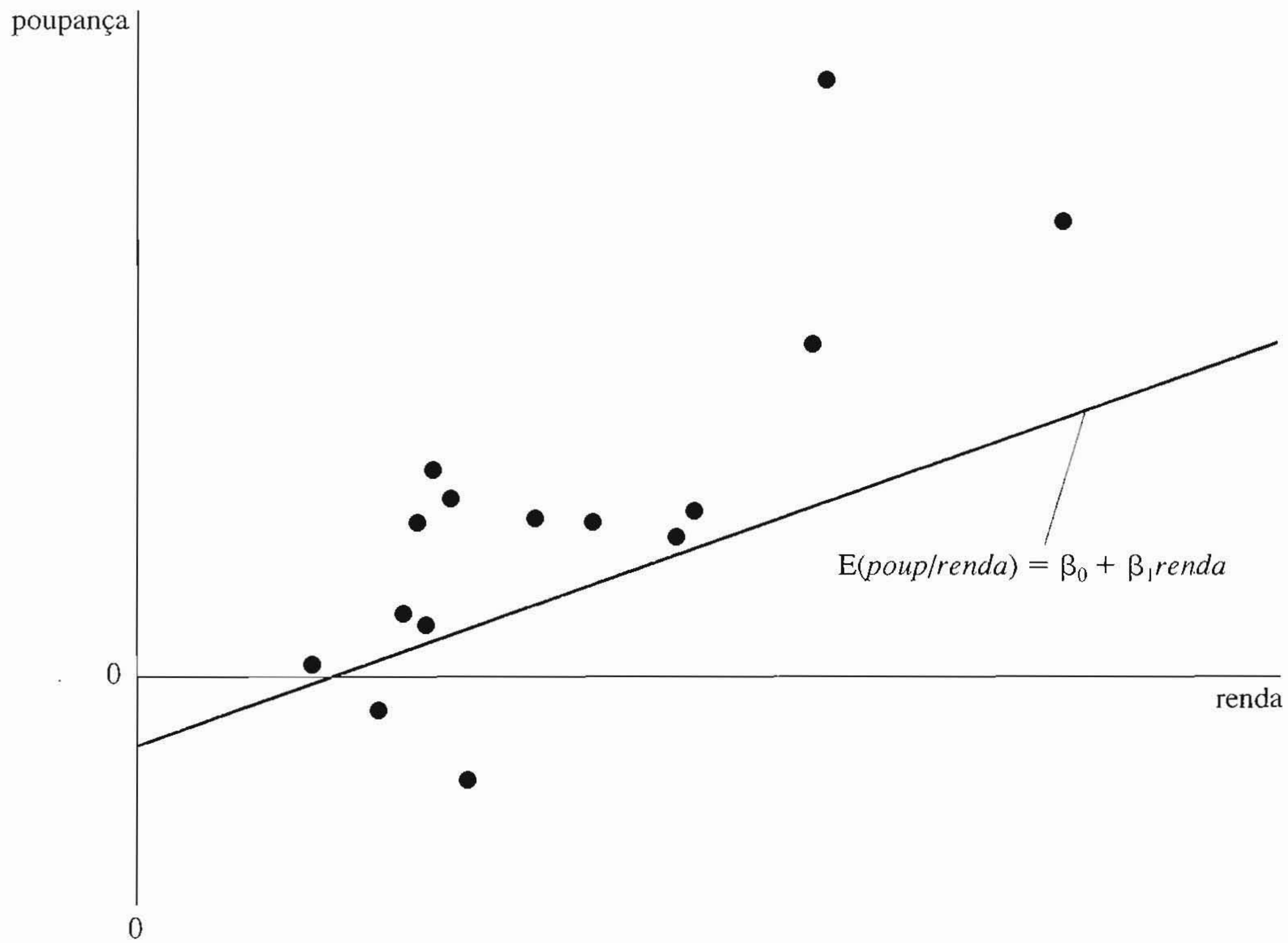
$$y_i = \beta_0 + \beta_1 x_i + u_i$$

– 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(*poup*|*renda*) = $\beta_0 + \beta_1 renda$.



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

– Cov(x,u)=E(xu)=0 pode ser escrita como: $E[x(y-\beta_0-\beta_1 x)]=0$

– 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 $\hat{\beta}_1$

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

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

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

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

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

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\Updownarrow$$

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

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

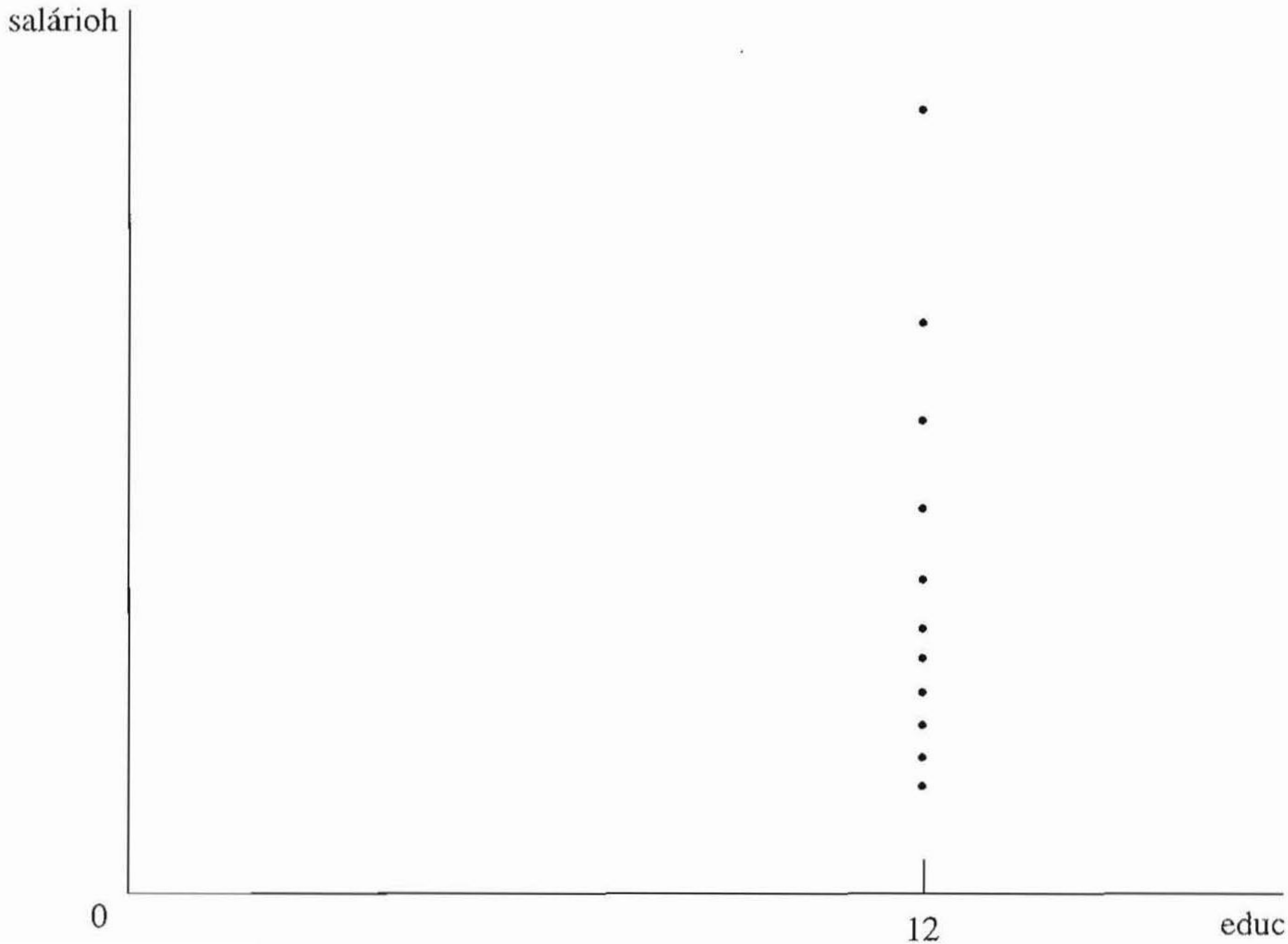
VARIÂNCIA DE x DEVE SER MAIOR QUE ZERO

– 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

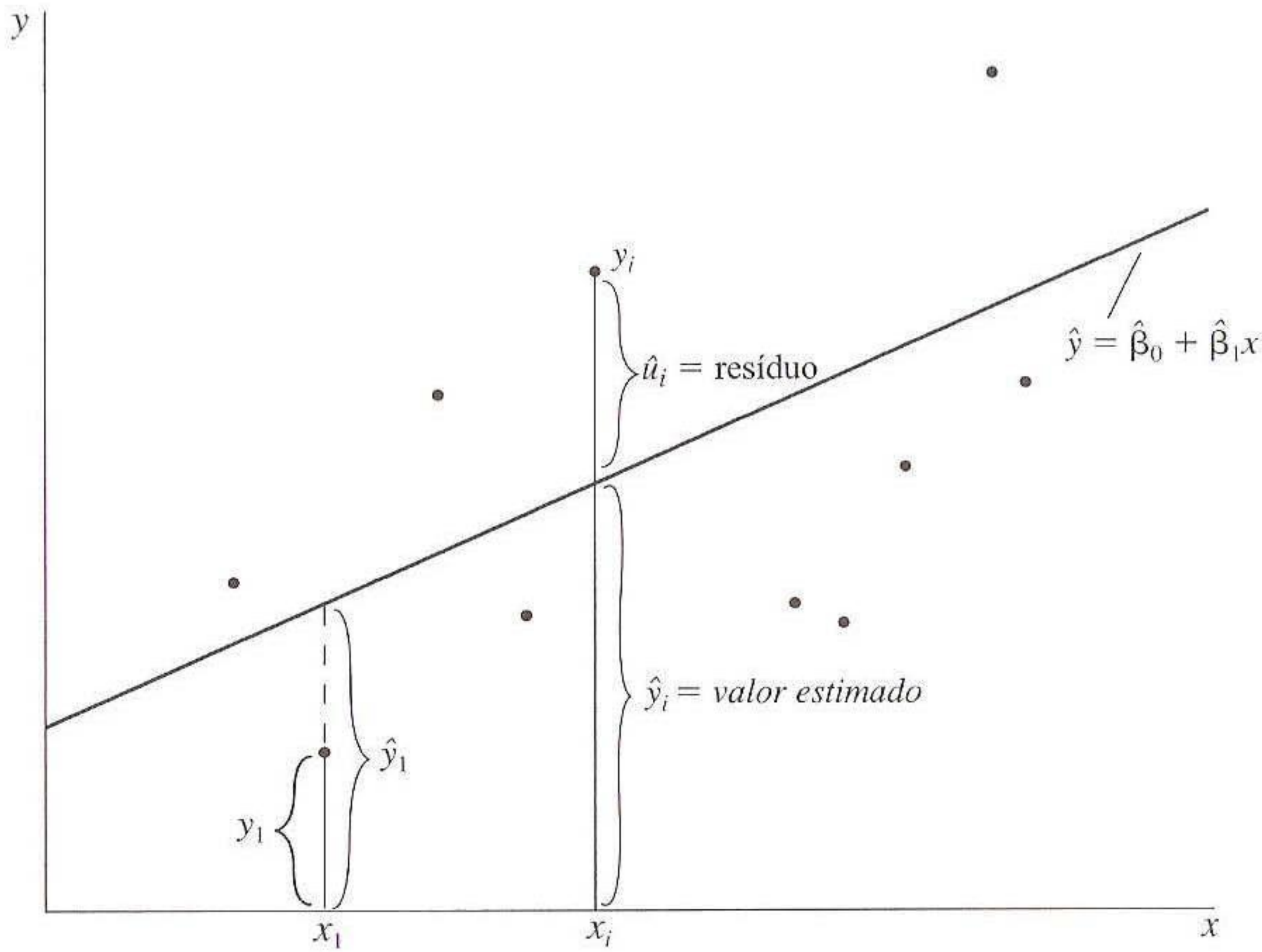
– Encontrados o intercepto e a inclinação, teremos um valor estimado para y para cada observação (x) na amostra:

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

– 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:

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

– 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.

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:

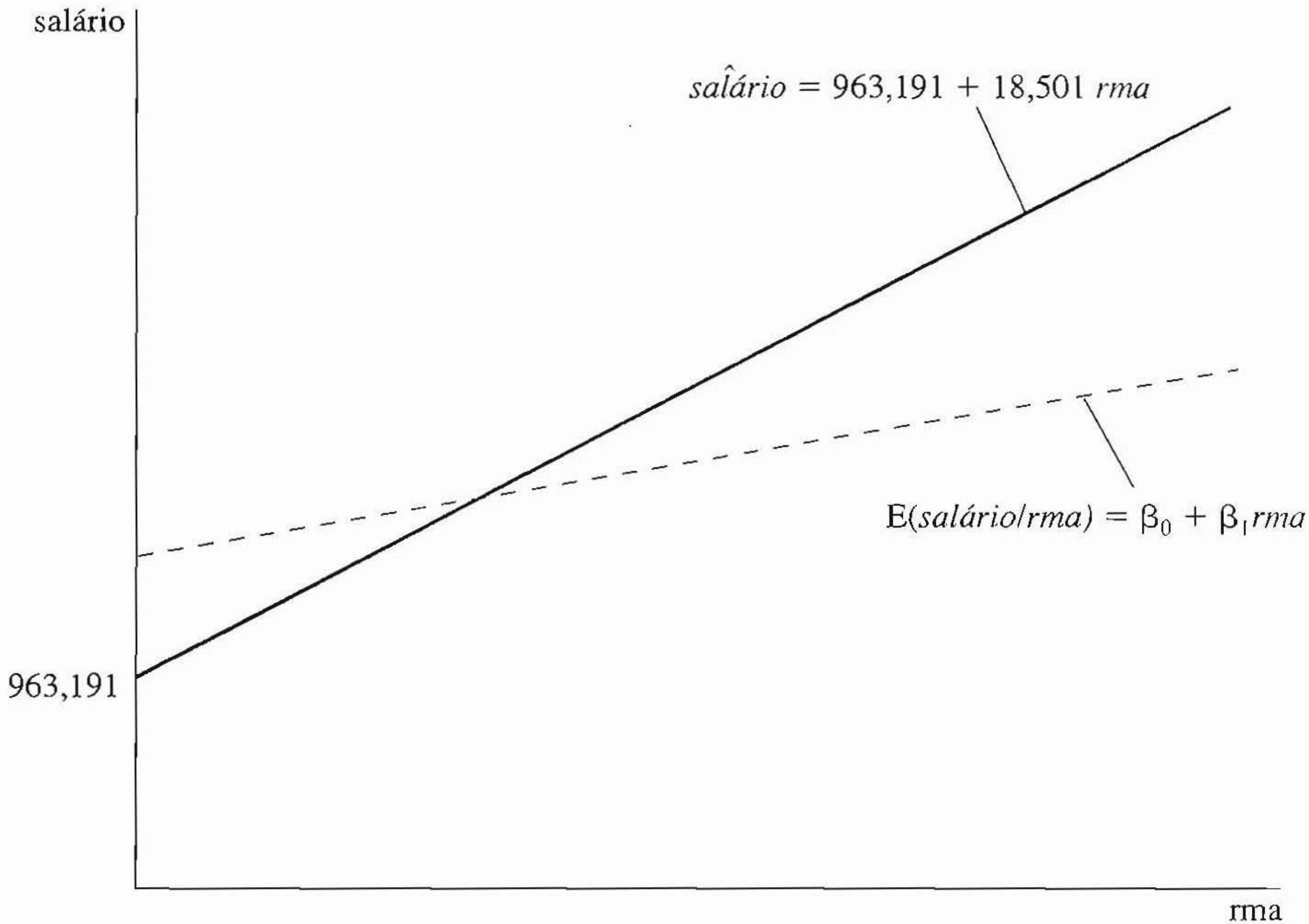
$$\hat{\beta}_1 = \Delta\hat{y}/\Delta x$$

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



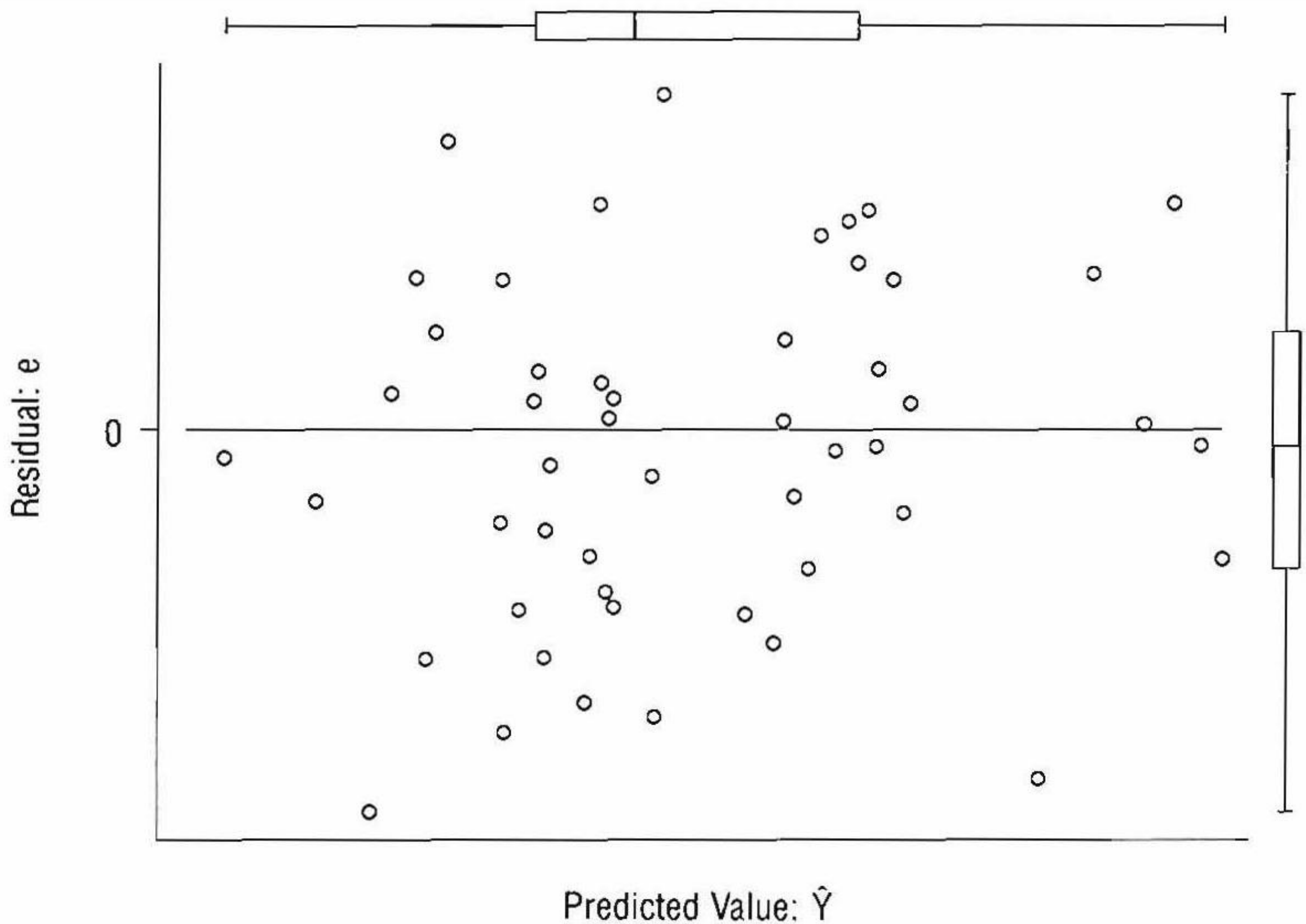


Figure 2.10 "All clear" e-versus-$\hat{Y}$ plot (artificial data).

Fonte: Hamilton, 1992: 52.

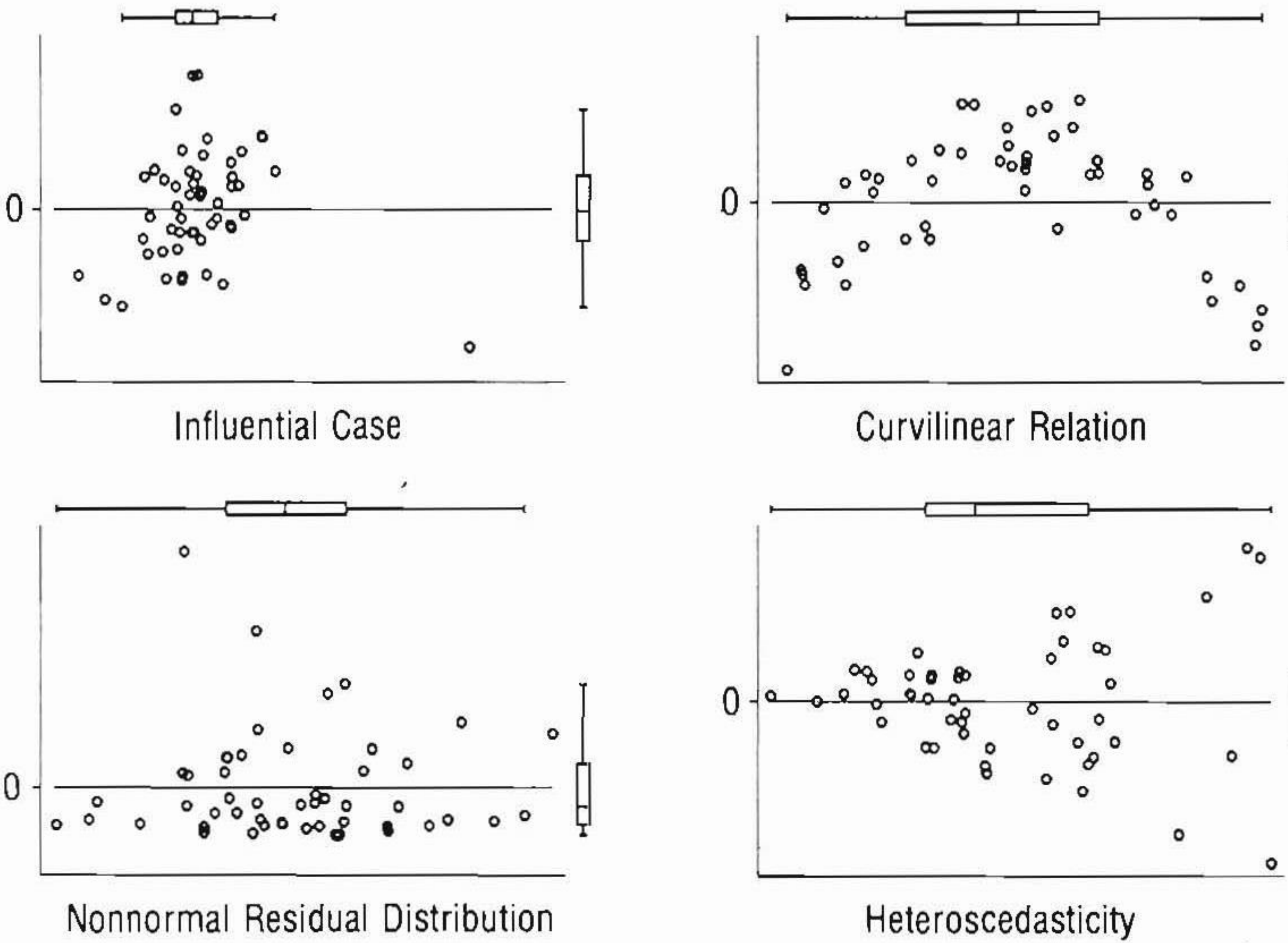


Figure 2.11 Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).

Fonte: Hamilton, 1992: 53.

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:

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

– 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:

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

– 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:

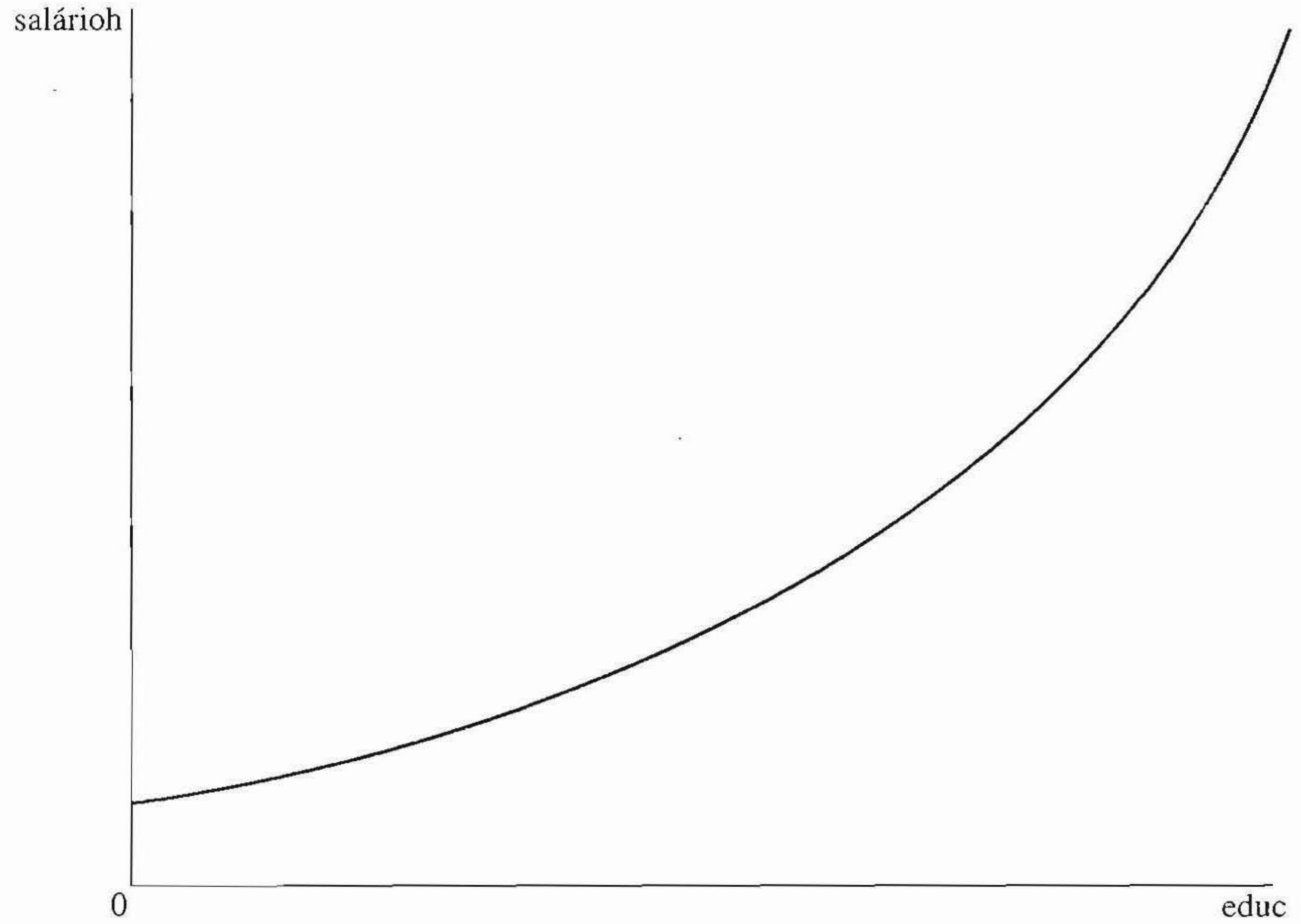
$$\%\Delta salário = (100 * \beta_1)\Delta educ$$

– 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*β_1)%. O cálculo da semi-elasticidade {[exp(β_1) – 1]*100} indica a diferença percentual exata:

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

– Aumento de 1% em x aumenta y em β_1% (modelo de elasticidade constante):

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

– 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)

UTILIZAÇÃO DE PESOS

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	aweight
reg y x, robust	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:”.

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), logaritmo do rendimento no trabalho principal (lnrenpri) e peso da pessoa (v4729):

	anest	renpri	lnrenpri	v4729
1	4	380	5,940171	613
2	4	530	6,272877	613
3	11	800	6,684612	613
4	6	350	5,857933	613
5	11	1600	7,377759	613
6	11	743	6,610696	613
7	11	500	6,214608	613
8	14	580	6,363028	613
9	4	380	5,940171	613
10	11	400	5,991465	613
11	11	8000	8,987197	612
12	8	459	6,12905	613
13	8	380	5,940171	613
14	0	120	4,787492	612
15	8	600	6,39693	612
16	8	550	6,309918	612
17	8	600	6,39693	612
18	10	400	5,991465	613
19	4	380	5,940171	613
20	4	380	5,940171	613

...

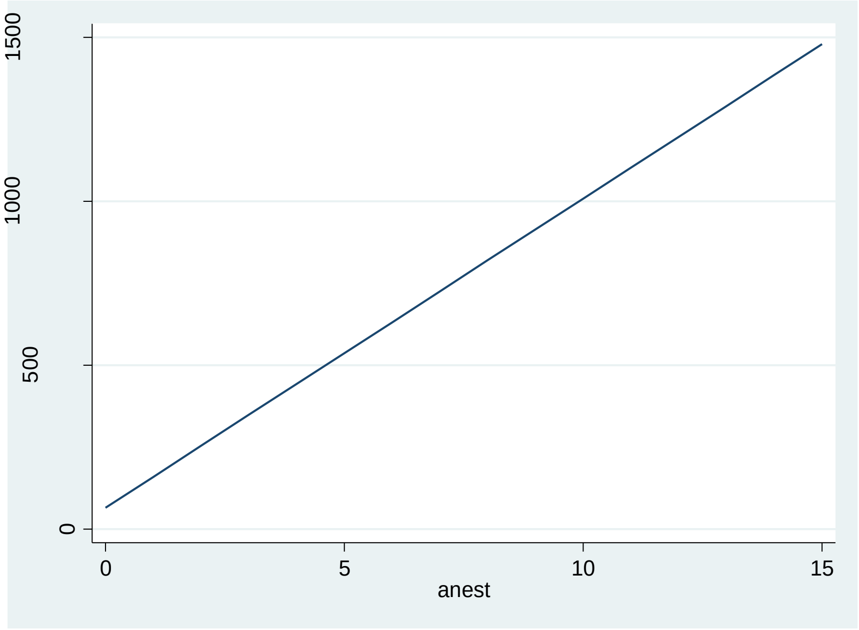
EXEMPLO 1: PNAD DE MINAS GERAIS DE 2007

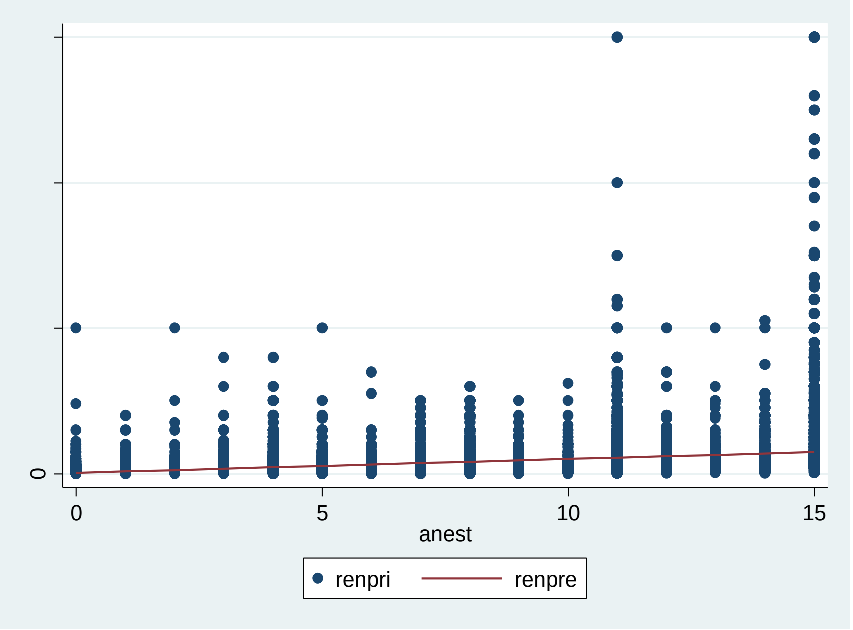
– Escolaridade explicando rendimento:

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

      Source |       SS       df       MS              Number of obs =   16232
-------------+------------------------------           F(  1, 16230) = 1710.07
       Model |  2.5086e+09     1  2.5086e+09           Prob > F      =  0.0000
    Residual |  2.3809e+10 16230  1466951.75           R-squared     =  0.0953
-------------+------------------------------           Adj R-squared =  0.0953
       Total |  2.6317e+10 16231  1621416.61           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

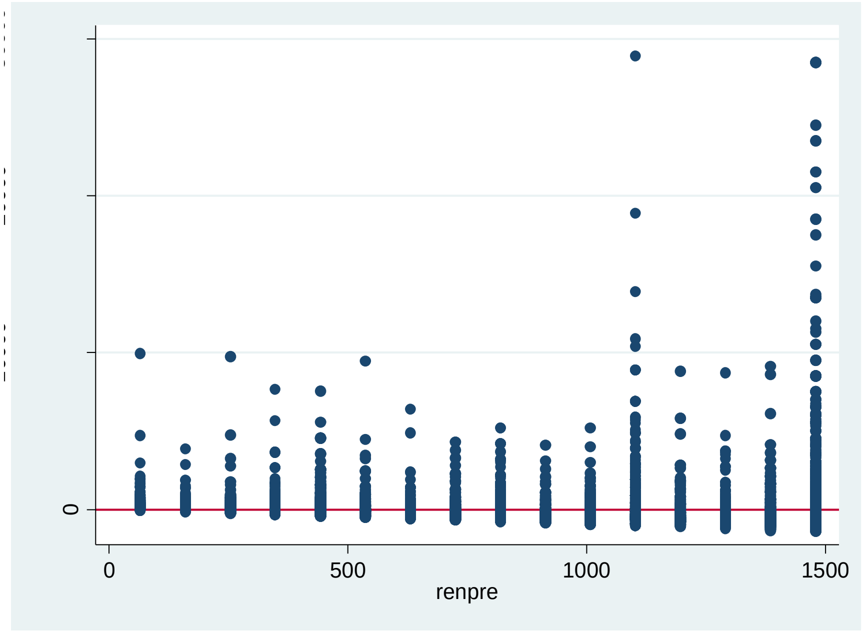
– Renda predita por anos de escolaridade:





EXEMPLO 1: PNAD DE MINAS GERAIS DE 2007

– Resíduos por renda predita:



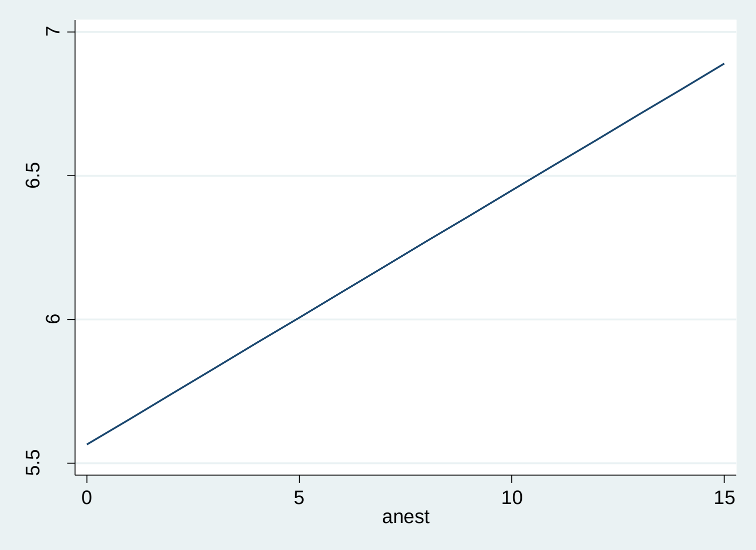
EXEMPLO 2: PNAD DE MINAS GERAIS DE 2007

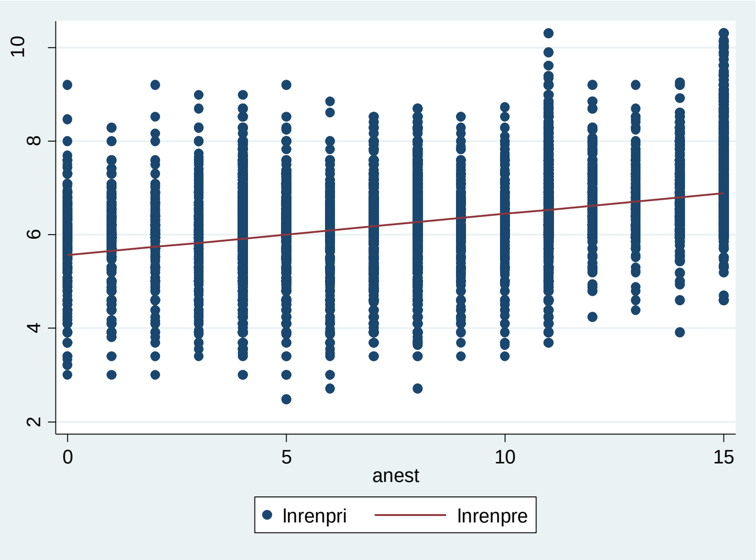
– Escolaridade explicando logaritmo do rendimento:

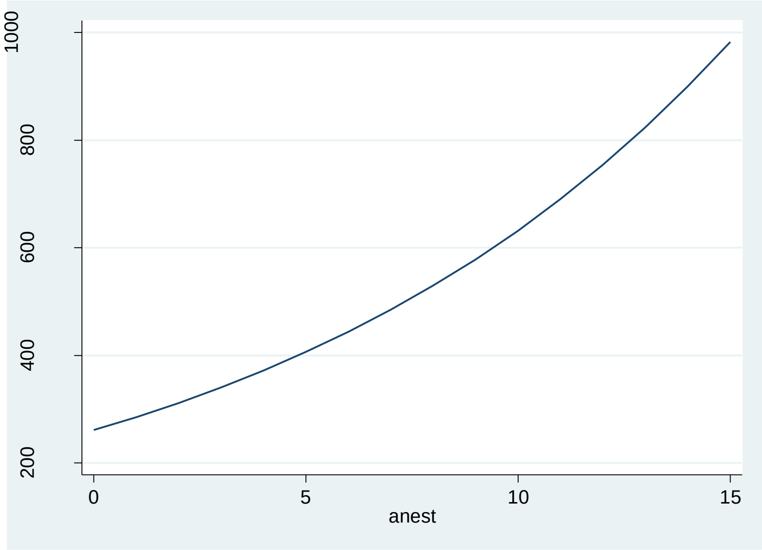
```
. reg lnrenpri anest [aweight=v4729]
(sum of wgt is   8.7563e+06)

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

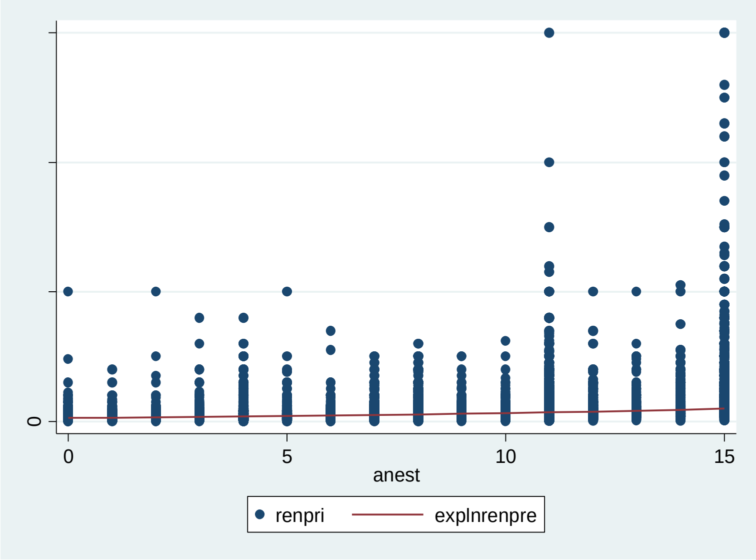
EXEMPLO 2: PNAD DE MINAS GERAIS DE 2007

– Renda predita por anos de escolaridade:



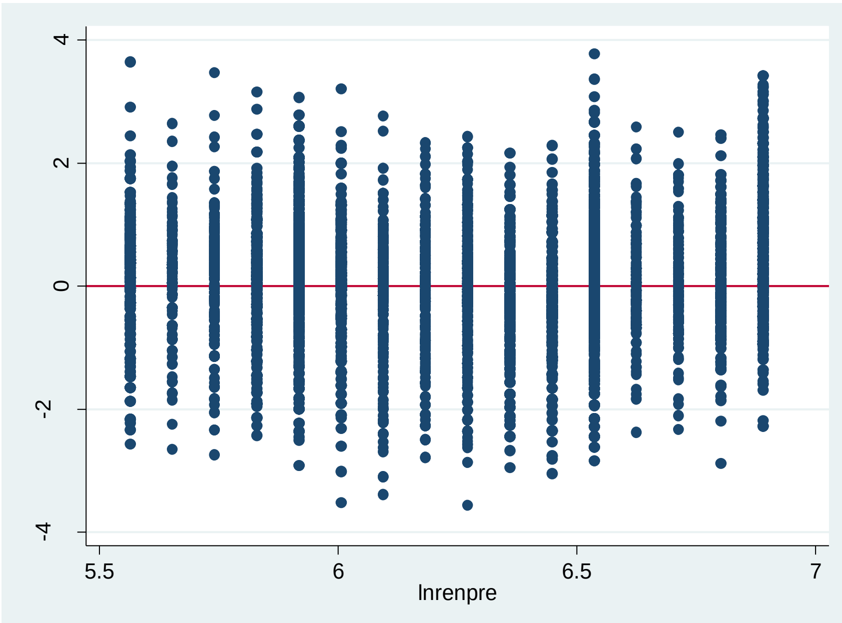


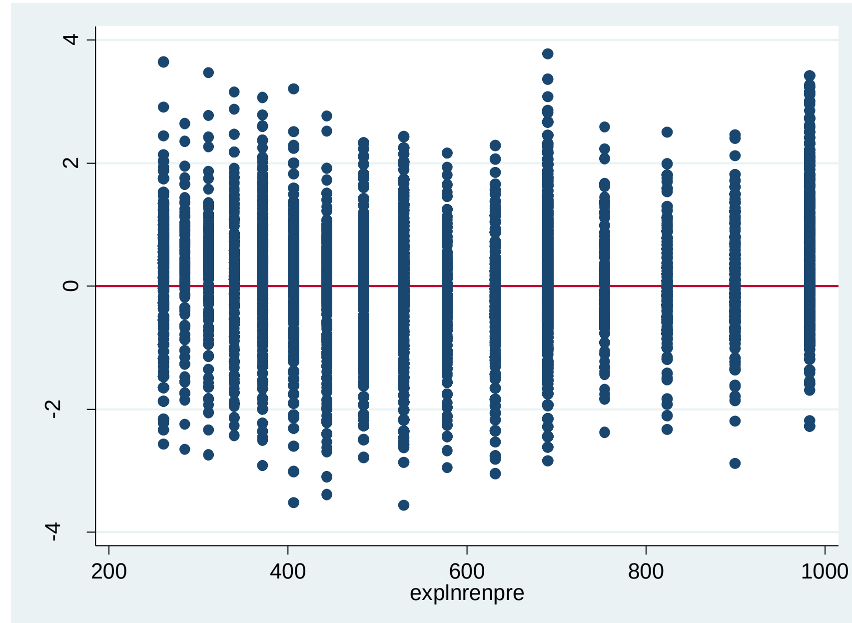




EXEMPLO 2: PNAD DE MINAS GERAIS DE 2007

– Resíduos por renda predita:





GRÁFICOS FORAM GERADOS COM ESTAS VARIÁVEIS

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

	anest	renpri	renpre	renres	lnrenpri	lnrenpre	lnrenres	explnrenpre	v4729
1	4	380	437,8324	-57,83244	5,940171	5,941878	-,0017066	380,649	613
2	4	530	437,8324	92,16756	6,272877	5,941878	,3309994	380,649	613
3	11	800	1094,848	-294,8484	6,684612	6,538908	,145704	691,531	613
4	6	350	625,5513	-275,5513	5,857933	6,112458	-,2545248	451,4469	613
5	11	1600	1094,848	505,1516	7,377759	6,538908	,8388512	691,531	613
6	11	743	1094,848	-351,8484	6,610696	6,538908	,071788	691,531	613
7	11	500	1094,848	-594,8484	6,214608	6,538908	-,3242996	691,531	613
8	14	580	1376,427	-796,4268	6,363028	6,794778	-,4317498	893,1708	613
9	4	380	437,8324	-57,83244	5,940171	5,941878	-,0017066	380,649	613
10	11	400	1094,848	-694,8484	5,991465	6,538908	-,5474432	691,531	613
11	11	8000	1094,848	6905,151	8,987197	6,538908	2,448289	691,531	612
12	8	459	813,2701	-354,2701	6,12905	6,283038	-,1539876	535,4126	613
13	8	380	813,2701	-433,2701	5,940171	6,283038	-,3428666	535,4126	613
14	0	120	62,39473	57,60527	4,787492	5,600718	-,813226	270,6206	612
15	8	600	813,2701	-213,2702	6,39693	6,283038	,1138919	535,4126	612
16	8	550	813,2701	-263,2701	6,309918	6,283038	,0268806	535,4126	612
17	8	600	813,2701	-213,2702	6,39693	6,283038	,1138919	535,4126	612
18	10	400	1000,989	-600,989	5,991465	6,453618	-,4621532	634,9956	613
19	4	380	437,8324	-57,83244	5,940171	5,941878	-,0017066	380,649	613
20	4	380	437,8324	-57,83244	5,940171	5,941878	-,0017066	380,649	613

...