

# **AULAS 17 E 18**

# **Análise de regressão múltipla: estimação**

**Ernesto F. L. Amaral**

**22 e 24 de outubro de 2013**  
**Avaliação de Políticas Públicas (DCP 046)**

**Fonte:**

**Cohen, Ernesto, e Rolando Franco. 2000. “Avaliação de Projetos Sociais”. São Paulo, SP: Editora Vozes. pp.118-136 (capítulo 7).**

**Wooldridge, Jeffrey M. “Introdução à econometria: uma abordagem moderna”. São Paulo: Cengage Learning, 2008. pp.64-109 (capítulo 3).**

# **CAPÍTULO 7 - COHEN & FRANCO**

## **MODELOS PARA A AVALIAÇÃO DE IMPACTOS**

# DESENHO DE PESQUISA DE AVALIAÇÃO DE IMPACTO

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

<b>GRUPO</b>	<b>ANTES</b>	<b>POLÍTICA</b>	<b>DEPOIS</b>
<b>Tratamento</b>	$T_0$	<b>X</b>	$T_1$
<b>Controle</b>	$C_0$		$C_1$

- “Diferença em diferenças” ou “dupla diferença” (DD) estima:
  - 1) Diferença dentro de cada grupo (tratamento e controle).
  - 2) Diferença dessas duas médias.

$$DD = (T_1 - T_0) - (C_1 - C_0)$$

## DESENHOS EXPERIMENTAIS

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

	<b>X</b>	<b>T<sub>1</sub></b>
		<b>C<sub>1</sub></b>

## DESENHOS QUASE-EXPERIMENTAIS

- O controle é construído com base na propensão do indivíduo de ingressar no programa.
- Busca-se obter grupo de comparação que corresponda ao grupo de beneficiários:
  - Com base em certas características (sociais, econômicas...) estima-se a probabilidade de um indivíduo de participar do programa.
  - Com base nessa propensão (exercício de emparelhamento), constitui-se o grupo de controle.
- Estima-se os efeitos na comparação entre o grupo de tratamento e o grupo de controle, antes e depois do programa.

$T_0$	$X$	$T_1$
$C_0$		$C_1$

## DESENHOS NÃO-EXPERIMENTAIS

- Ausência de grupos de controle torna mais difícil isolar causas que geram impactos na variável de interesse.
- Pode ser realizada análise reflexiva para estimar efeitos dos programas, com comparação dos resultados obtidos pelos beneficiários antes e depois do programa.
- Modelo antes-depois:

$T_0$	$X$	$T_1$

- Modelo somente depois com grupo de comparação:

	$X$	$T_1$	$T_2$
		$C_1$	$C_2$

- Modelo somente depois:

	$X$	$T_1$	$T_2$

<b>DESENHO DA AVALIAÇÃO</b>	<b>MÉTODO DE ESTIMAÇÃO DE IMPACTO</b>
<b>EXPERIMENTAL</b>	<b>COMPARAÇÃO DE MÉDIAS</b>
<b>QUASE-EXPERIMENTAL</b>	<b>REGRESSÃO MÚLTIPLA &amp; DIFERENÇA EM DIFERENÇAS</b>
<b>NÃO-EXPERIMENTAL</b>	<b>REGRESSÃO MÚLTIPLA</b>

**CAPÍTULO 3 - WOOLDRIDGE  
ANÁLISE DE REGRESSÃO MÚLTIPLA:  
ESTIMAÇÃO**



# MODELO DE REGRESSÃO MÚLTIPLA

- A desvantagem de usar análise de **regressão simples** é o fato de ser difícil que todos os outros fatores que afetam  $y$  não estejam correlacionados com  $x$ .
- Análise de **regressão múltipla** possibilita *ceteris paribus* (outros fatores constantes), pois permite controlar muitos outros fatores que afetam a variável dependente simultaneamente.
- Isso auxilia no teste de teorias e hipóteses, quando possuímos dados não-experimentais.
- Ao utilizar mais fatores na explicação de  $y$ , uma maior variação de  $y$  será explicada pelo modelo.
- Este é o modelo mais utilizado nas ciências sociais.
- O método de MQO é usado para estimar os parâmetros do modelo de regressão múltipla.

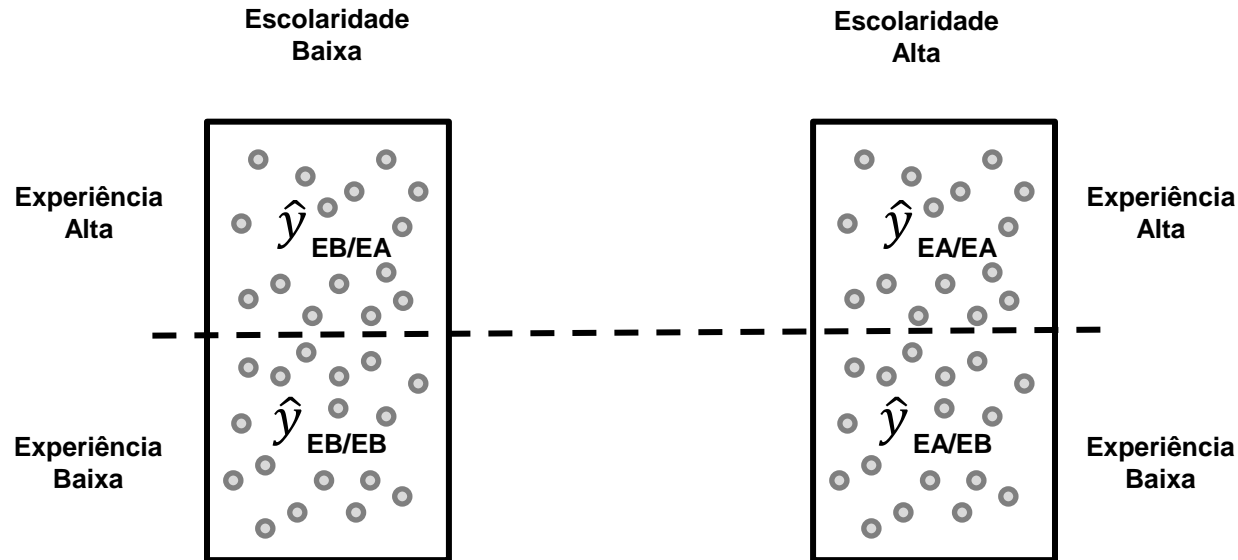
## MODELO COM DUAS VARIÁVEIS INDEPENDENTES

$$\textit{saláριο}h = \beta_0 + \beta_1 \textit{escol} + \beta_2 \textit{exper} + u$$

- Salário é determinado por escolaridade, experiência e outros fatores não-observáveis (Equação Minceriana).
- $\beta_1$  mede o efeito de escolaridade sobre salário, mantendo todos os outros fatores fixos (*ceteris paribus*).
- $\beta_2$  mede o efeito de experiência sobre salário, mantendo todos os outros fatores fixos.
- Como experiência foi inserida na equação, podemos medir o efeito de escolaridade sobre salário, mantendo experiência fixa.
- Na regressão simples, teríamos que assumir que experiência não é correlacionada com escolaridade, o que é uma hipótese fraca.

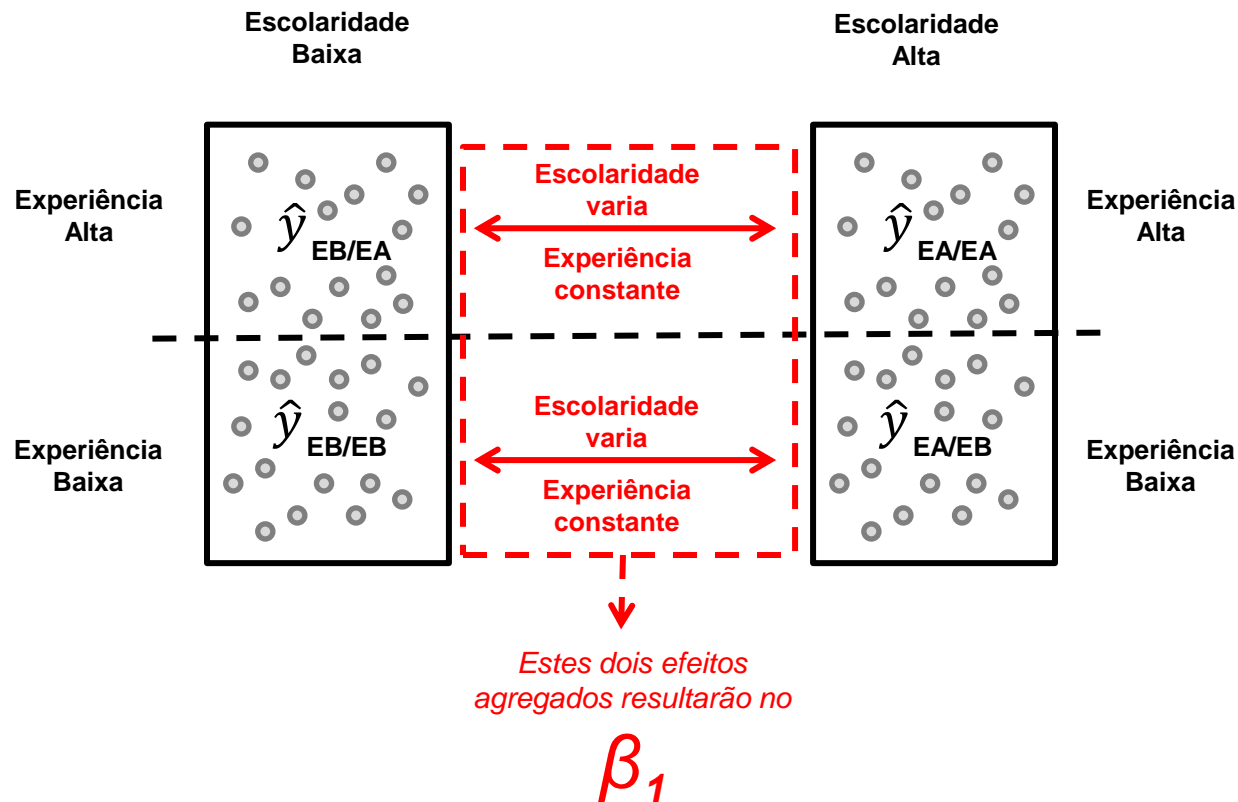
# ENTENDENDO ANÁLISE *CETERIS PARIBUS* NA REGRESSÃO MÚLTIPLA

$$salário_{ih} = \beta_0 + \beta_1 escol + \beta_2 exper + u$$



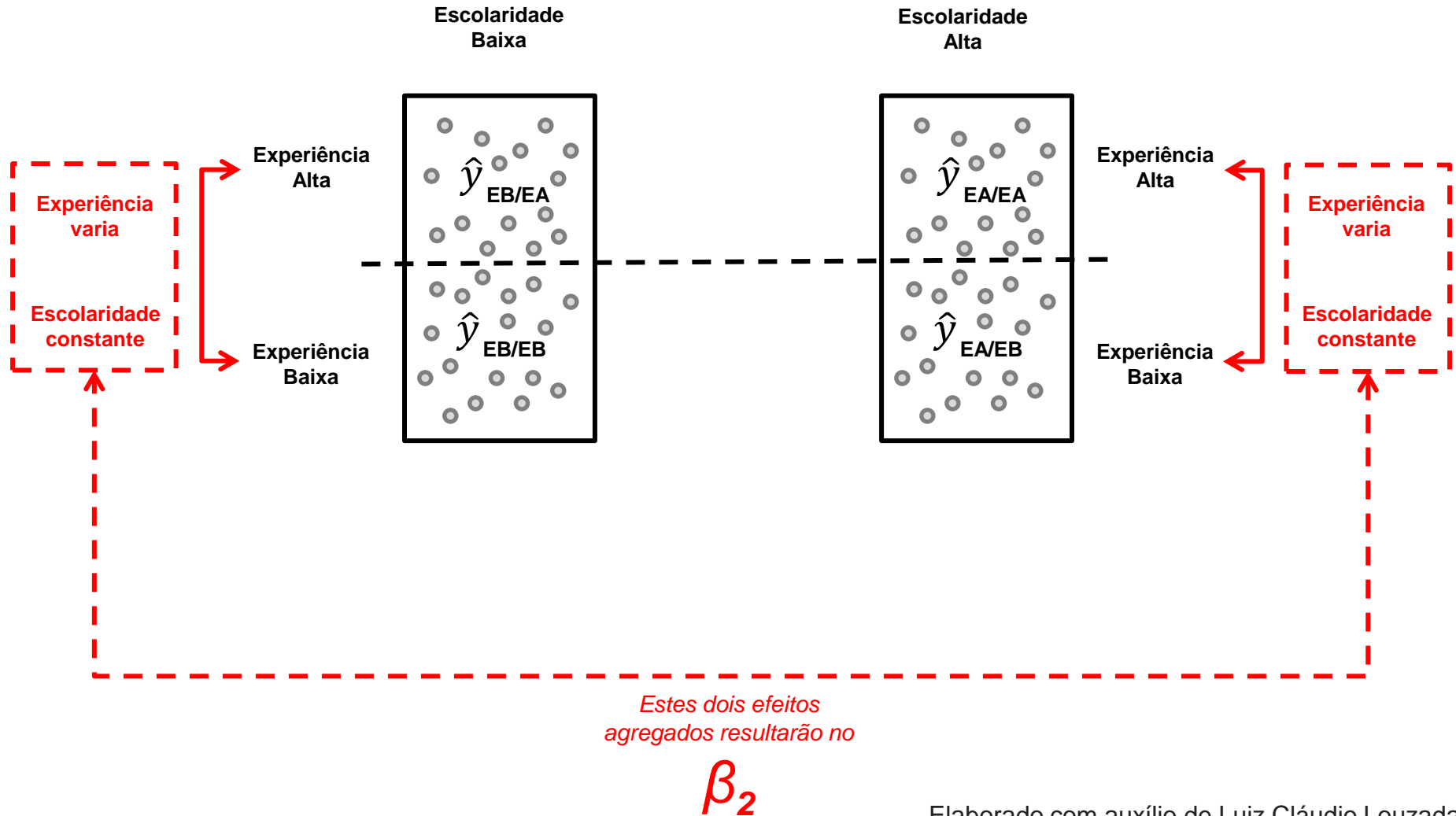
# ENTENDENDO ANÁLISE *CETERIS PARIBUS* NA REGRESSÃO MÚLTIPLA

$$\text{salário}_{ih} = \beta_0 + \beta_1 \text{escol} + \beta_2 \text{exper} + u$$



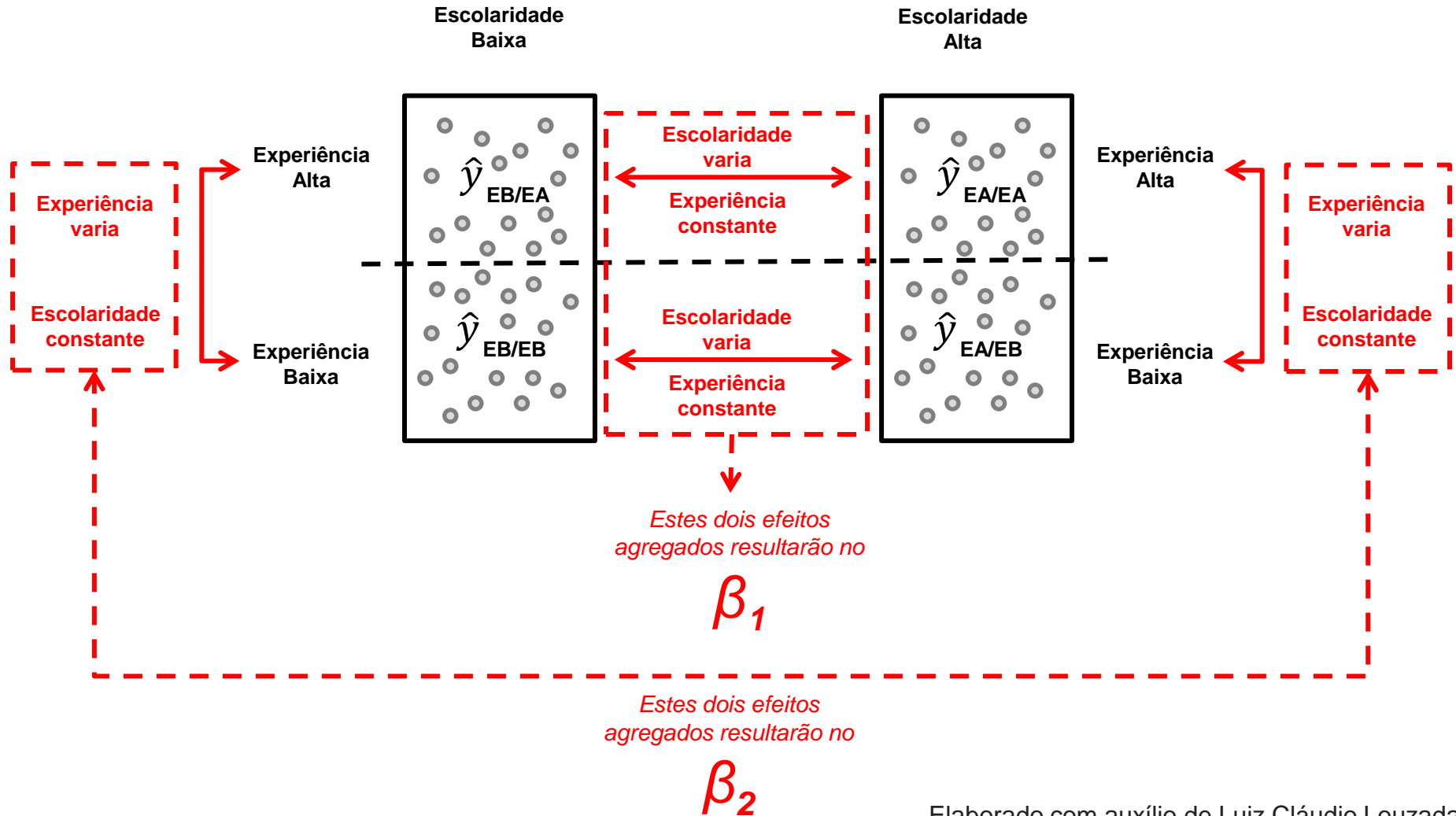
# ENTENDENDO ANÁLISE *CETERIS PARIBUS* NA REGRESSÃO MÚLTIPLA

$$\text{salário}_{ih} = \beta_0 + \beta_1 \text{escol} + \beta_2 \text{exper} + u$$



# ENTENDENDO ANÁLISE *CETERIS PARIBUS* NA REGRESSÃO MÚLTIPLA

$$\text{salário}_{ih} = \beta_0 + \beta_1 \text{escol} + \beta_2 \text{exper} + u$$



## EXEMPLOS COM PNAD DE MINAS GERAIS DE 2007

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

	anest	idpia	renpri	lnrenpri
1	4	42	380	5.940171
2	4	62	530	6.272877
3	11	33	800	6.684612
4	6	25	350	5.857933
5	11	33	1600	7.377759
6	11	45	743	6.610696
7	11	38	500	6.214608
8	14	36	580	6.363028
9	4	39	380	5.940171
10	11	25	400	5.991465
11	11	31	8000	8.987197
12	8	33	459	6.12905
13	8	33	380	5.940171
14	0	50	120	4.787492
15	8	46	600	6.39693
16	8	40	550	6.309918
17	8	34	600	6.39693
18	10	27	400	5.991465
19	4	39	380	5.940171
20	4	33	380	5.940171

# EXEMPLO 1: PNAD DE MINAS GERAIS DE 2007

– Escolaridade e idade explicando rendimento:

```
. reg renpri anest idpia [aweight=v4729]
(sum of wgt is 8.4521e+06)
```

Source	SS	df	MS	
Model	3.9215e+09	2	1.9608e+09	Number of obs = 15682
Residual	2.1184e+10	15679	1351081.24	F( 2, 15679) = 1451.25
Total	2.5105e+10	15681	1600989.85	Prob > F = 0.0000
				R-squared = 0.1562
				Adj R-squared = 0.1561
				Root MSE = 1162.4

renpri	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
anest	113.8647	2.322679	49.02	0.000	109.312	118.4175
idpia	26.9488	.7925765	34.00	0.000	25.39526	28.50234
_cons	-1071.595	39.02647	-27.46	0.000	-1148.091	-995.0986



## EXEMPLO 2: PNAD DE MINAS GERAIS DE 2007

– Escolaridade e idade explicando logaritmo do rendimento:

```
. reg lnrenpri anest idpia [aweight=v4729]
(sum of wgt is 8.4521e+06)
```

Source	SS	df	MS			
Model	2962.15421	2	1481.07711	Number of obs =	15682	
Residual	8237.31092	15679	.525372212	F( 2, 15679) =	2819.10	
Total	11199.4651	15681	.714206054	Prob > F =	0.0000	
				R-squared =	0.2645	
				Adj R-squared =	0.2644	
				Root MSE =	.72483	

lnrenpri	Coef.	Std. Err.	t	P> t	[95% Conf. Interva]	
anest	.1014244	.0014484	70.03	0.000	.0985854	.1042634
idpia	.0217022	.0004942	43.91	0.000	.0207335	.022671
_cons	4.687027	.0243362	192.60	0.000	4.639325	4.734729

# MODELO GERAL DE DUAS VARIÁVEIS INDEPENDENTES

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- $\beta_0$  é o intercepto.
- $\beta_1$  mede a variação em  $y$  com relação a  $x_1$ , mantendo os outros fatores constantes.
- $\beta_2$  mede a variação em  $y$  com relação a  $x_2$ , mantendo os outros fatores constantes.

# RELAÇÕES FUNCIONAIS ENTRE VARIÁVEIS

- A regressão múltipla é útil para generalizar relações funcionais entre variáveis.
- Por exemplo:

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

- Variação no consumo decorrente de variação na renda é:

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

- O efeito marginal da renda sobre o consumo depende tanto de  $\beta_2$  como de  $\beta_1$  e do nível de renda.
- A definição das variáveis independentes é sempre importante na interpretação dos parâmetros.

# EXEMPLO 3: PNAD DE MINAS GERAIS DE 2007

– Idade e idade ao quadrado explicando rendimento:

```
. reg renpri idpia idquad [aweight=v4729]
(sum of wgt is 8.4521e+06)
```

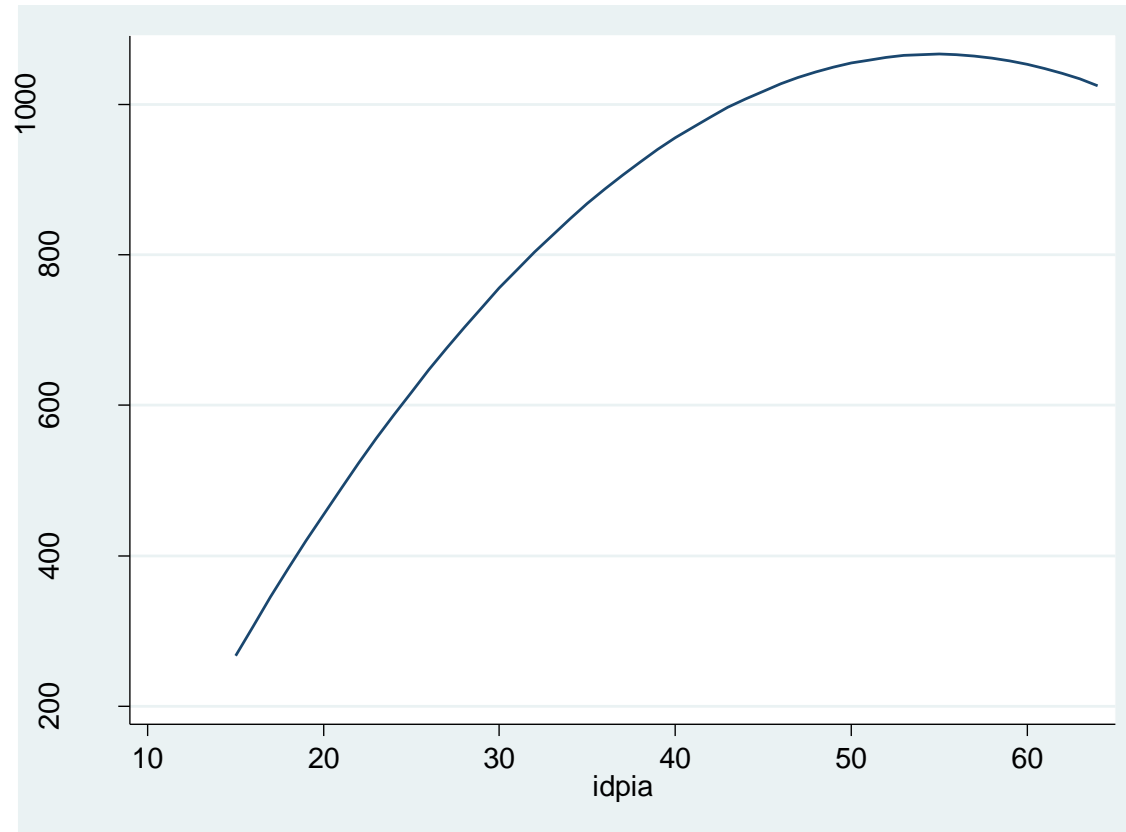
Source	SS	df	MS			
Model	765311134	2	382655567	Number of obs =	15682	
Residual	2.4340e+10	15679	1552382.85	F( 2, 15679) =	246.50	
Total	2.5105e+10	15681	1600989.85	Prob > F =	0.0000	
				R-squared =	0.0305	
				Adj R-squared =	0.0304	
				Root MSE =	1245.9	

renpri	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
idpia	55.13908	5.037051	10.95	0.000	45.26588	65.01228
idquad	-.5022615	.0656772	-7.65	0.000	-.6309963	-.3735267
_cons	-446.5655	89.82328	-4.97	0.000	-622.6295	-270.5015

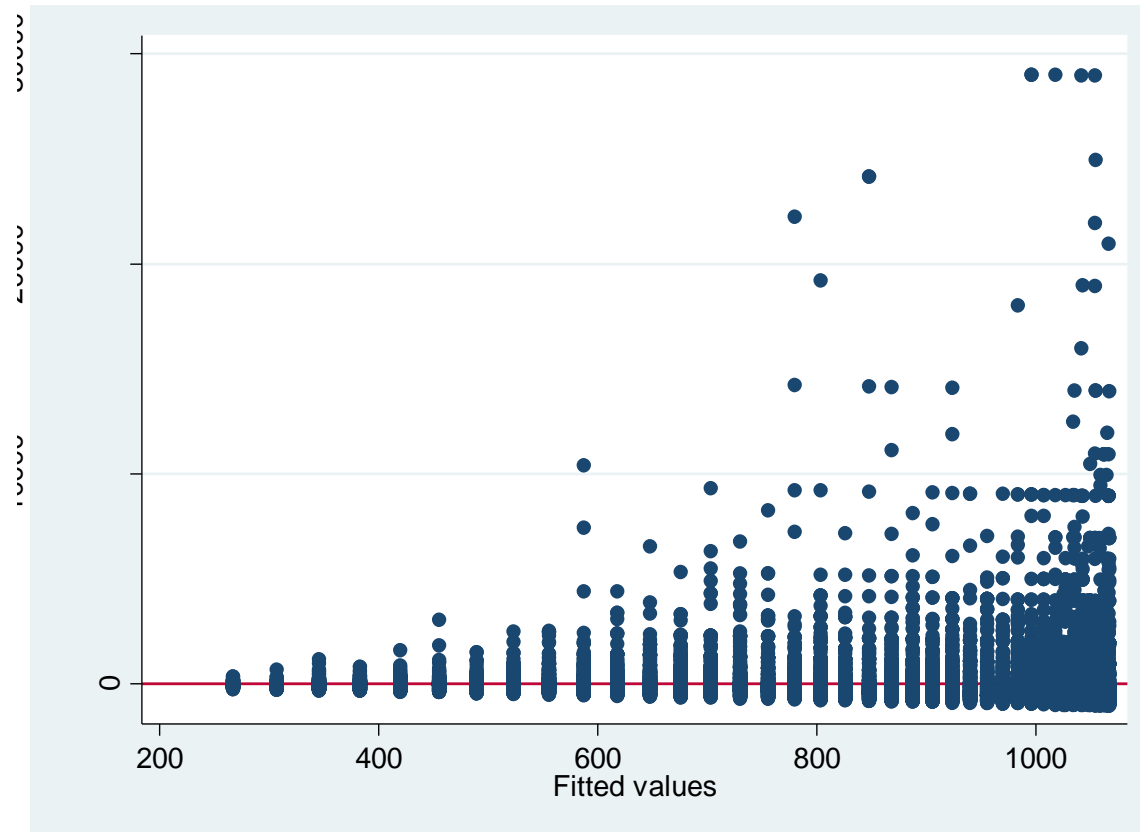
## EXEMPLO 3: PNAD DE MINAS GERAIS DE 2007

– Renda predita por idade:



# EXEMPLO 3: PNAD DE MINAS GERAIS DE 2007

– Resíduos por renda predita:



## EXEMPLO 4: PNAD DE MINAS GERAIS DE 2007

– Idade e idade ao quadrado explicando logaritmo do rendimento:

```
. reg lnrenpri idpia idquad [aweight=v4729]
(sum of wgt is 8.4521e+06)
```

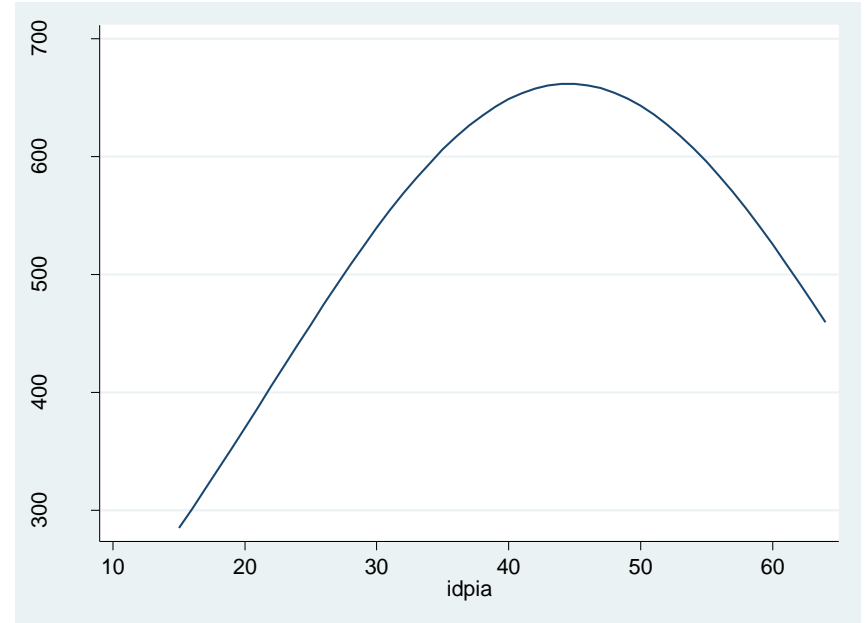
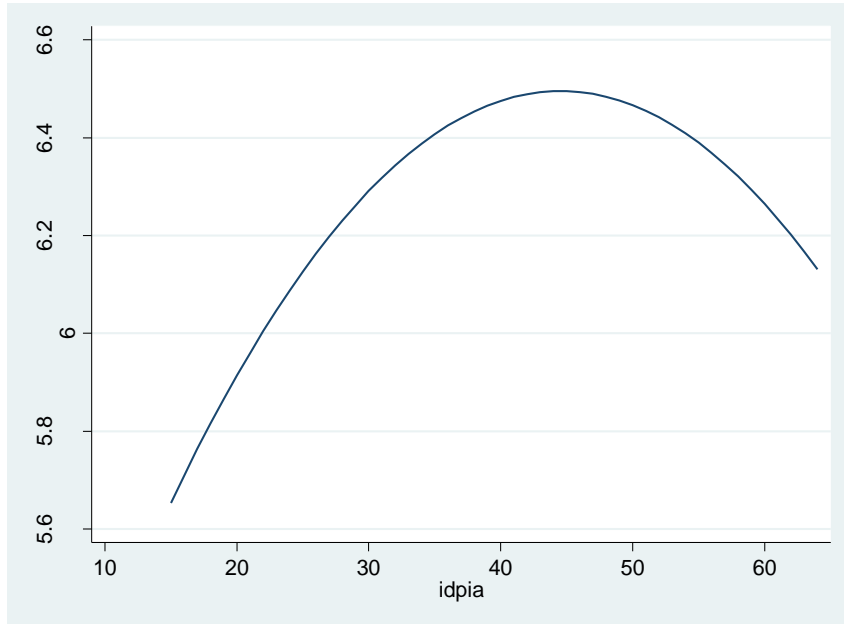
Source	SS	df	MS			
Model	720.724017	2	360.362009	Number of obs =	15682	
Residual	10478.7411	15679	.668329684	F( 2, 15679) =	539.20	
Total	11199.4651	15681	.714206054	Prob > F =	0.0000	
				R-squared =	0.0644	
				Adj R-squared =	0.0642	
				Root MSE =	.81751	

lnrenpri	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
idpia	.0859434	.003305	26.00	0.000	.0794652	.0924216
idquad	-.0009645	.0000431	-22.38	0.000	-.001049	-.0008801
_cons	4.580841	.0589366	77.72	0.000	4.465319	4.696364

# EXEMPLO 4: PNAD DE MINAS GERAIS DE 2007

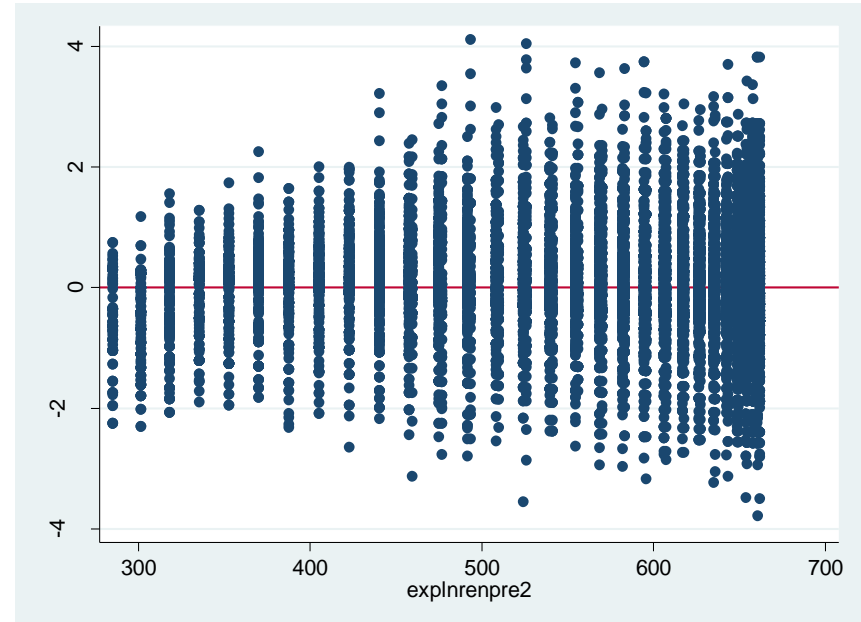
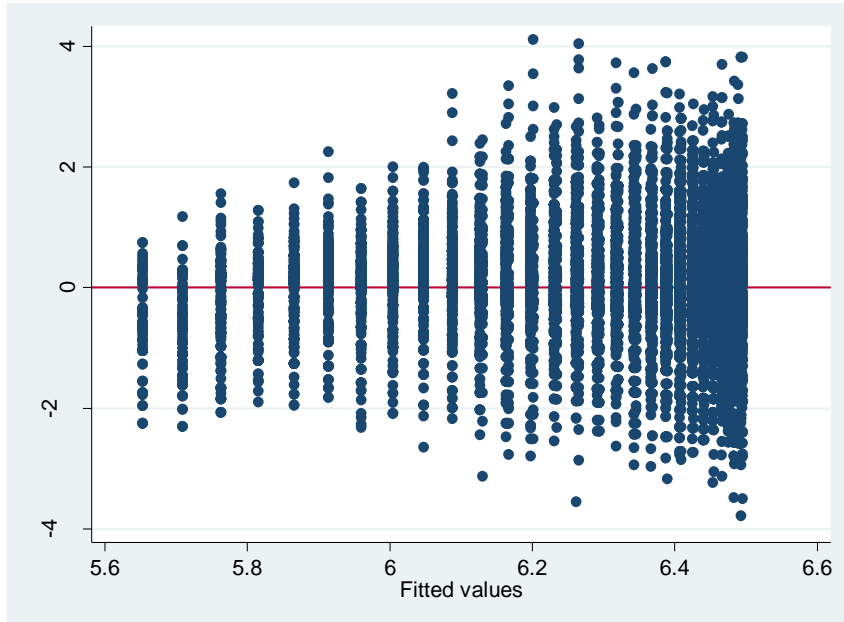
– Renda predita por idade:





# EXEMPLO 4: PNAD DE MINAS GERAIS DE 2007

– Resíduos por renda predita:



## HIPÓTESE SOBRE $u$ EM RELAÇÃO A $x_1$ E $x_2$

$$E(u/x_1, x_2)=0$$

- Para qualquer valor de  $x_1$  e  $x_2$  na população, o fator não-observável médio é igual a zero.
- Isso implica que outros fatores que afetam  $y$  não estão, em média, relacionados com as variáveis explicativas.
- Os níveis médios dos fatores não-observáveis devem ser os mesmos nas combinações das variáveis independentes.
- A esperança igual a zero significa que a relação funcional entre as variáveis explicada e as explicativas está correta.
- No exemplo da renda ao quadrado, não é preciso incluir  $rend^2$ , já que ela é conhecida quando se conhece  $rend$ :

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

## MODELO COM $k$ VARIÁVEIS INDEPENDENTES

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

- Esse é o modelo de regressão linear múltipla geral ou, simplesmente, modelo de regressão múltipla.
- Há  $k + 1$  parâmetros populacionais desconhecidos, já que temos  $k$  variáveis independentes e um intercepto.
- Os parâmetros  $\beta_1$  a  $\beta_k$  são chamados de parâmetros de inclinação, mesmo que eles não tenham exatamente este significado.
- **A regressão é “linear” porque é linear nos  $\beta_j$ , mesmo que seja uma relação não-linear entre a variável dependente e as variáveis independentes:**

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

## OBTENÇÃO DAS ESTIMATIVAS DE MQO

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

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

- O método de mínimos quadrados ordinários escolhe as estimativas que minimizam a soma dos resíduos quadrados.
- Dadas  $n$  observações de  $y$ ,  $x_1$ ,  $x_2$ , ... e  $x_k$ , as estimativas dos parâmetros são escolhidas para fazer com que a expressão abaixo tenha o menor valor possível:

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

# INTERPRETAÇÃO DA EQUAÇÃO DE REGRESSÃO

- Novamente a reta de regressão de MQO:

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

- O intercepto é o valor previsto de  $y$  quando todas as variáveis independentes são iguais a zero.
- As estimativas dos demais parâmetros têm interpretações de efeito parcial (*ceteris paribus*).
- Da equação acima, temos:

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

- O coeficiente de  $x_1$  mede a variação em  $y$  devido a um aumento de uma unidade em  $x_1$ , mantendo todas as outras variáveis independentes constantes:

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

## SIGNIFICADO DE “MANTER OUTROS FATORES FIXOS”

- Regressão múltipla permite interpretação *ceteris paribus* mesmo que dados não sejam coletados de maneira *ceteris paribus*.
- Os dados são coletados por amostra aleatória que não estabelece restrições sobre os valores a serem obtidos das variáveis independentes.
- Ou seja, a regressão múltipla permite simular situação de outros fatores constantes, sem restringir a coleta de dados.
- Essa modelagem permite realizar em ambientes não-experimentais o que cientistas naturais realizam em experimentos de laboratório (mantendo outros fatores fixos).
- A **avaliação de impacto de políticas** pode ser realizada com regressão múltipla, mensurando relação entre variáveis independentes e dependente, com noção de *ceteris paribus*.

## GRAU DE AJUSTE

- O  $R^2$  nunca diminui quando outra variável independente é adicionada na regressão.
- Isso ocorre porque a soma dos resíduos quadrados nunca aumenta quando variáveis explicativas são acrescentadas ao modelo.
- Essa característica faz de  $R^2$  um teste fraco para decidir pela inclusão de variáveis no modelo.
- O efeito parcial da variável independente ( $\beta_k$ ) sobre  $y$  é o que deve definir se a variável deve ser inserida no modelo.
- $R^2$  é um grau de ajuste geral do modelo, assim como um teste para indicar o quanto um grupo de variáveis explica variações em  $y$ .

## REGRESSÃO ATRAVÉS DA ORIGEM

- Em alguns modelos, pode-se avaliar que o ideal seria ter  $\beta_0$  igual a zero:

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

- $R^2$  pode ser negativo, o que significa que a média amostral de  $y$  “explica” mais da variação em  $y_i$  do que as variáveis independentes.
- Nesse caso, devemos incluir um intercepto ou procurar novas variáveis explicativas.
- Se  $\beta_0$  for diferente de zero na população, a regressão através da origem gera estimadores dos parâmetros de inclinação ( $\beta_k$ ) viesados.
- Se  $\beta_0$  for igual a zero na população, a regressão com intercepto gera maiores variâncias dos estimadores de inclinação.



# VALOR ESPERADOS DOS ESTIMADORES DE MQO

## HIPÓTESE RLM.1 (LINEAR NOS PARÂMETROS)

- Modelo na população pode ser escrito como:

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

- $\beta_0, \beta_1, \dots, \beta_k$  são parâmetros desconhecidos (constantes) de interesse, e  $u$  é um erro aleatório não-observável ou um termo de perturbação aleatória.

## HIPÓTESE RLM.2 (AMOSTRAGEM ALEATÓRIA)

- Temos uma amostra aleatória de  $n$  observações do modelo populacional acima.

## HIPÓTESE RLM.3 (MÉDIA CONDICIONAL ZERO)

- O erro  $u$  tem um valor esperado igual a zero, dados quaisquer valores das variáveis independentes:

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

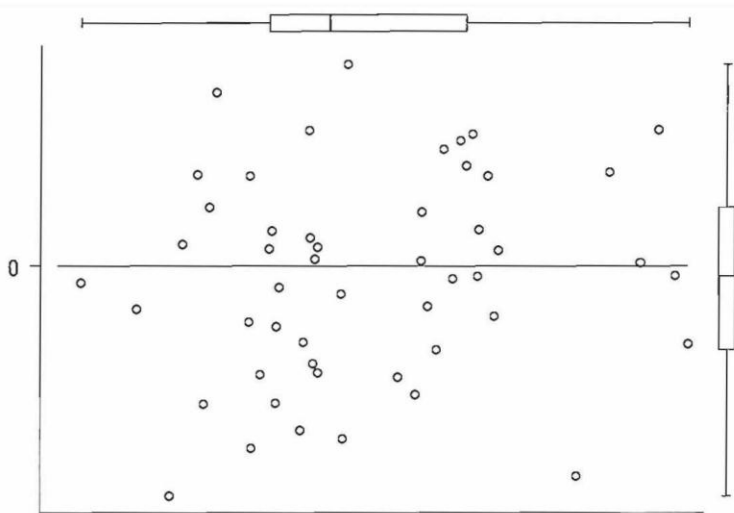
## HIPÓTESE RLM.4 (COLINEARIDADE NÃO PERFEITA)

- Na amostra e na população, nenhuma das variáveis independentes é constante, e não há relações lineares exatas entre as variáveis independentes.
- As variáveis independentes devem ser correlacionadas entre si, mas não deve haver **colinearidade perfeita** (por exemplo, uma variável não pode ser múltiplo de outra).
- Altos graus de correlação entre variáveis independentes e tamanho pequeno da amostra aumentam variância de beta.
- Correlação alta (mas não perfeita) entre duas ou mais variáveis não é desejável (**multicolinearidade**).
- Por outro lado, **se a correlação for nula**, não é necessário regressão múltipla, mas sim regressão simples, já que o termo de erro englobaria todos fatores não-observáveis e não-relacionados com as variáveis independentes.

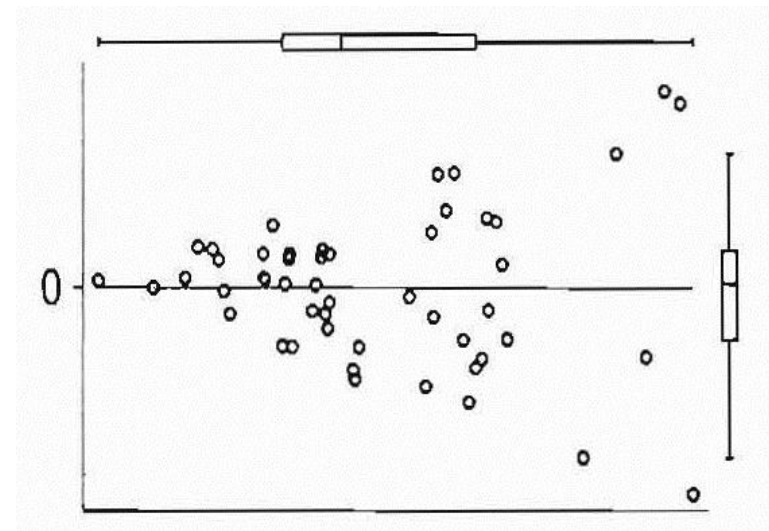
## HIPÓTESE RLM.5 (HOMOSCEDASTICIDADE)

- A variância do termo erro ( $u$ ), condicionada às variáveis explicativas, é a mesma para todas as combinações de resultados das variáveis explicativas.
- Se essa hipótese é violada, o modelo exibe heteroscedasticidade.

### HOMOSCEDASTICIDADE



### HETEROSCEDASTICIDADE



Fonte: Hamilton, 1992: 52-53.

## TEOREMA DE GAUSS-MARKOV

- Sob as hipóteses RLM.1 a RLM.5, os parâmetros estimados do intercepto e de inclinação são os melhores estimadores lineares não-viesados dos parâmetros populacionais:

*Best Linear Unbiased Estimators (BLUEs)*

- Em outras palavras, os estimadores de mínimos quadrados ordinários (MQO) são os melhores estimadores lineares não-viesados.