


Econometria

Parte 1

Prof. Adalto Acir Althaus Junior oe

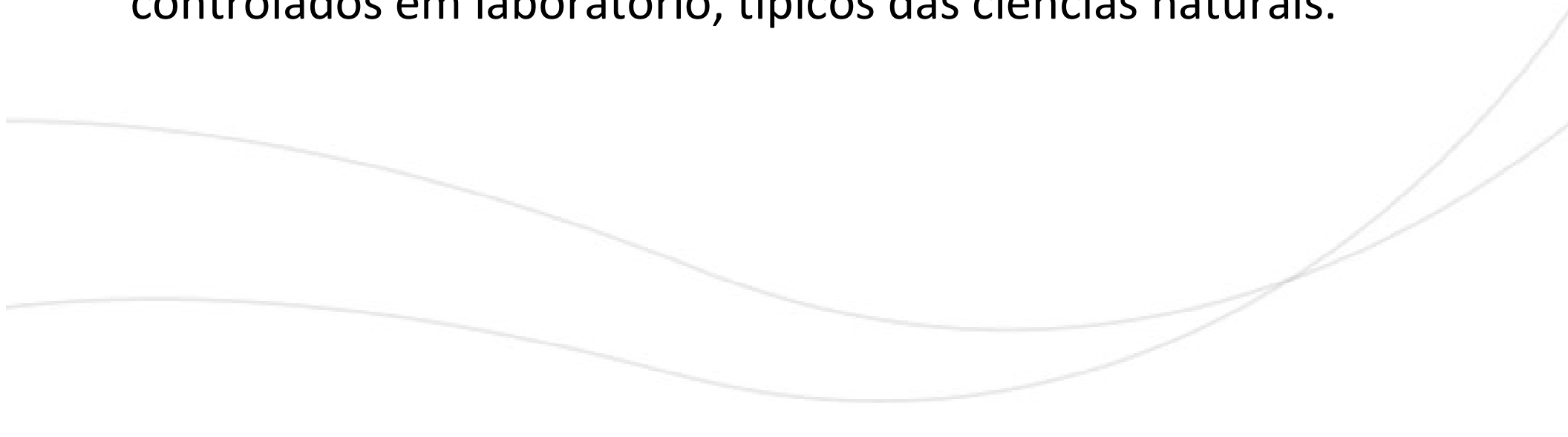
Sumário

- O que é?
 - Por que usar?
 - Modelos econométricos
 - Natureza dos dados
 - Causalidade
- 

O que é?

- Econometria é um conjunto de métodos estatísticos usados para estimação de relações econômicas
- Econometria teórica: trata do desenvolvimento de métodos adequados para medir as relações econômicas especificadas nos modelos econométricos. Sob esse aspecto, ela depende fortemente da estatística matemática
- Econometria aplicada: utiliza as ferramentas da econometria teórica para estudar diversos tópicos econômicos, como a função de produção, a função investimento, as funções de oferta e de demanda, a teoria do portfólio, etc.


Por que usar?

- Método mais adequado para lidar com dados não-experimentais.
 - Dados não-experimentais: coletados pela observação da realidade, típicos das ciências sociais.
 - Dados experimentais: coletados a partir de experimentos controlados em laboratório, típicos das ciências naturais.
- 


Modelos econométricos

- Primeiro passo para uma análise econométrica é a escolha do modelo econométrico a ser usado.
- Modelo econométrico: função matemática que representará a relação econômica a ser estudada.
 - $\text{Renda} = f(\text{Educação, Habilidade, ...})$
- Duas questões a se considerar:
 - ✓ Que relação será analisada? E quais os fatores associados a ela?
 - ✓ Que informações devem ser usadas?


Que relação será analisada?

- Se existir modelos teóricos que expliquem o fenômeno estudado, basta decidir qual deles usar.
 - Exemplo: para estudar o impacto de um imposto sobre um mercado há diversos modelos teóricos para se basear (competição perfeita, monopólio, oligopólio etc.); basta escolher o mais adequado ao problema.
- 

Que relação será analisada?

- Não havendo modelos econômicos que explicitem as relações que se pretende estudar, confia-se na intuição sobre o problema ou em conhecimentos de outras áreas.
 - Exemplo: para estudar os fatores associados aos níveis de saúde e educação das pessoas, os modelos econométricos se baseiam no que se sabe em outras áreas de conhecimento.
- 

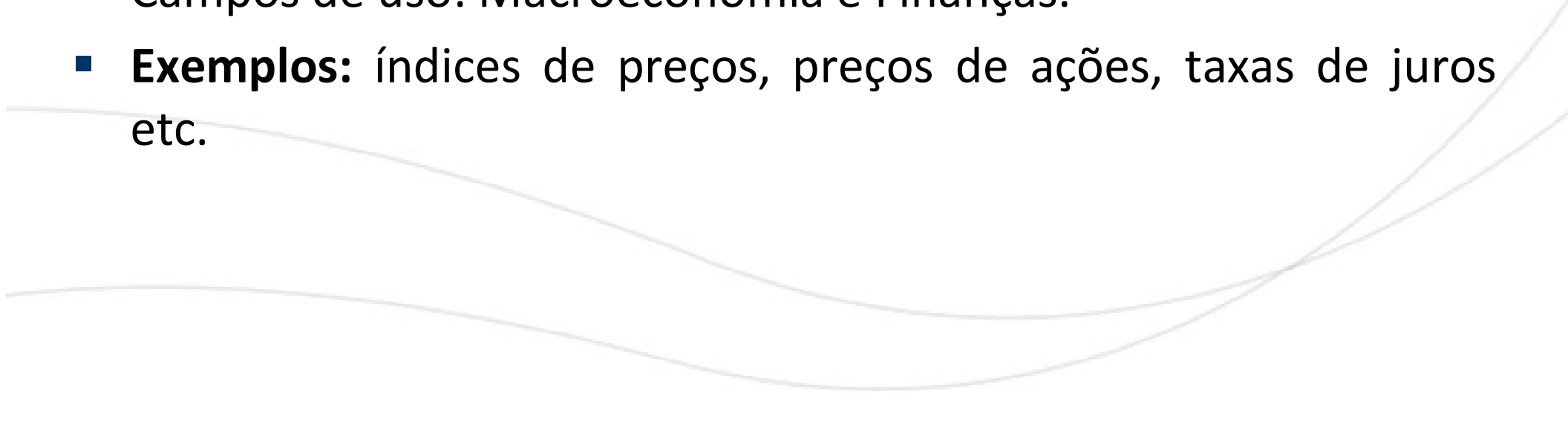
Que informações devem ser usadas?

- Depois de decidida a questão a ser respondida, deve-se decidir que tipo de dados são necessários.
 - Natureza dos dados econométricos:
 - ✓ cross-section
 - ✓ séries temporais
 - ✓ painel
- 

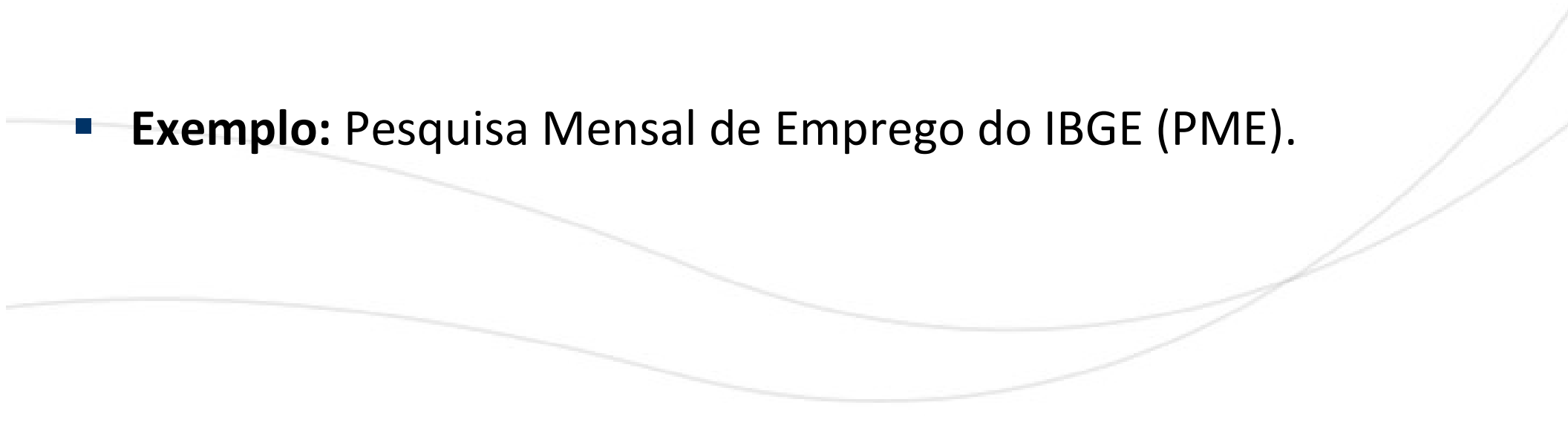
Dados em *cross-section*

- Dados de amostras de unidades (pessoas, firmas, países etc.) observadas em **um ponto no tempo**.
- A amostra deve ser **aleatória**, isto é, cada unidade é selecionada de forma independente.
- Campos de uso: Organização Industrial, Economia do Trabalho, Economia do Setor Público, etc.
- **Exemplo:** Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE – contém informações demográficas e socioeconômicas de domicílios em todo o país.

Séries de tempo

- Observações de uma variável **ao longo do tempo**, como PIB, inflação, câmbio, etc.
 - Informação depende do tempo, cada observação está relacionada com seu passado.
 - Não é possível obter amostras aleatórias.
 - Métodos econométricos diferentes dos usados em *cross-section*.
 - Campos de uso: Macroeconomia e Finanças.
 - **Exemplos:** índices de preços, preços de ações, taxas de juros etc.
- 

Dados em Painel

- Combinação de estrutura de série de tempo com as unidades de uma cross-section.
 - Unidades selecionadas de forma aleatória e então acompanhadas ao longo do tempo.
 - Ferramentas econométricas parecidas com as usadas em cross-section.
 - **Exemplo:** Pesquisa Mensal de Emprego do IBGE (PME).
- 

Tipos de Dados

Cross Sectional

	Ano 2000					
	Variável Y	Variável X1	Variável X2	Variável X3	...	Variável Xn
Empresa A						
Empresa B						
Empresa C						
Empresa D						
Empresa E						
Empresa F						
...
Empresa Z						

Time Series

	Variável Y
Ano 2000	
Ano 2001	
Ano 2002	
Ano 2003	
Ano 2004	
Ano 2005	
...	...
Ano 20nn	

Panel Data

	Ano	Variável Y	Variável X1	Variável X2	...	Variável Xn
Empresa A	2000					
Empresa A	2001					
Empresa A	...					
Empresa A	20nn					
Empresa B	2000					
Empresa B	2001					
Empresa B	...					
Empresa B	20nn					
Empresa C	2000					
Empresa C	2001					
Empresa C	...					
Empresa C	20nn					
...
Empresa Z	2000					
Empresa Z	2001					
Empresa Z	...					
Empresa Z	20nn					

Tipos de Dados

Cross Sectional

	Ano 2000					
	Variável Y	Variável X1	Variável X2	Variável X3	...	Variável Xn
Empresa A						
Empresa B						
Empresa C						
Empresa D						
Empresa E						
Empresa F						
...
Empresa Z						

Time Series

	Empresa A					
	Variável Y	Variável X1	Variável X2	Variável X3	...	Variável Xn
Ano 2000						
Ano 2001						
Ano 2002						
Ano 2003						
Ano 2004						
Ano 2005						
...
Ano 20nn						

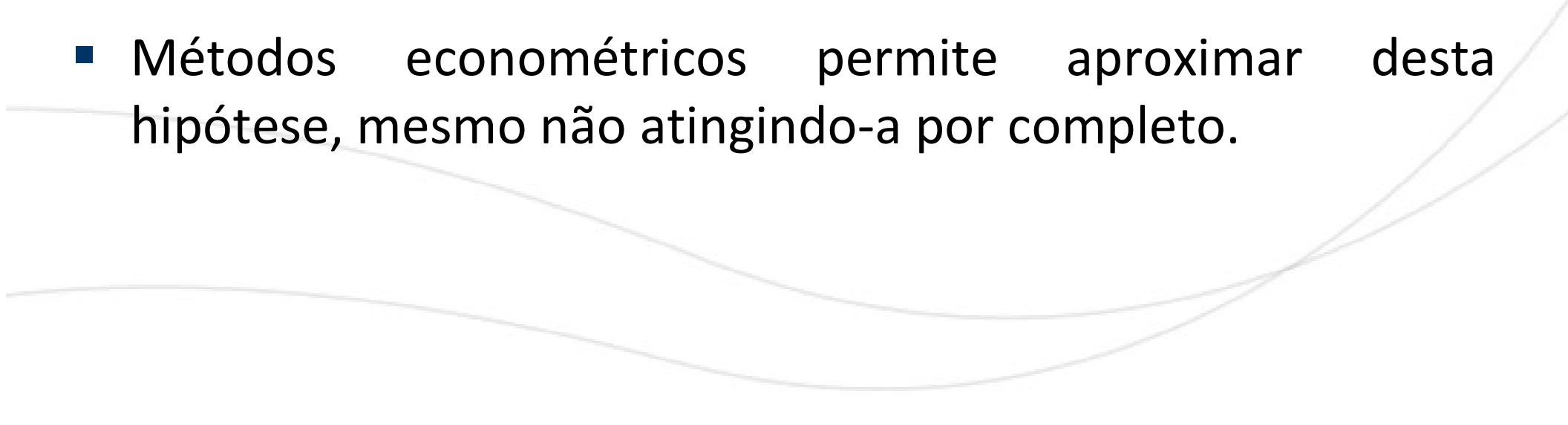
Panel Data

	Ano	Variável Y	Variável X1	Variável X2	...	Variável Xn
Empresa A	2000					
Empresa A	2001					
Empresa A	...					
Empresa A	20nn					
Empresa B	2000					
Empresa B	2001					
Empresa B	...					
Empresa B	20nn					
Empresa C	2000					
Empresa C	2001					
Empresa C	...					
Empresa C	20nn					
...
Empresa Z	2000					
Empresa Z	2001					
Empresa Z	...					
Empresa Z	20nn					

Time Series

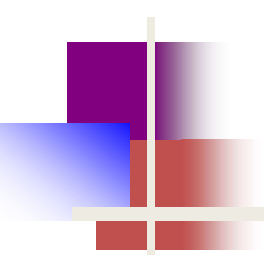
	Variável Y
Ano 2000	
Ano 2001	
Ano 2002	
Ano 2003	
Ano 2004	
Ano 2005	
...	...
Ano 20nn	

Causalidade

- Em geral, procura-se **relações de causalidade** entre os fenômenos econômicos
 - Tarefa difícil: há inúmeros fatores associados a um fenômeno econômico.
 - Hipótese ***ceteris paribus*** (“tudo mais constante”): permite estabelecer uma causalidade mais ‘pura’ entre os fenômenos econômicos.
 - Métodos econométricos permite aproximar desta hipótese, mesmo não atingindo-a por completo.
- 

Conceitos e definições gerais





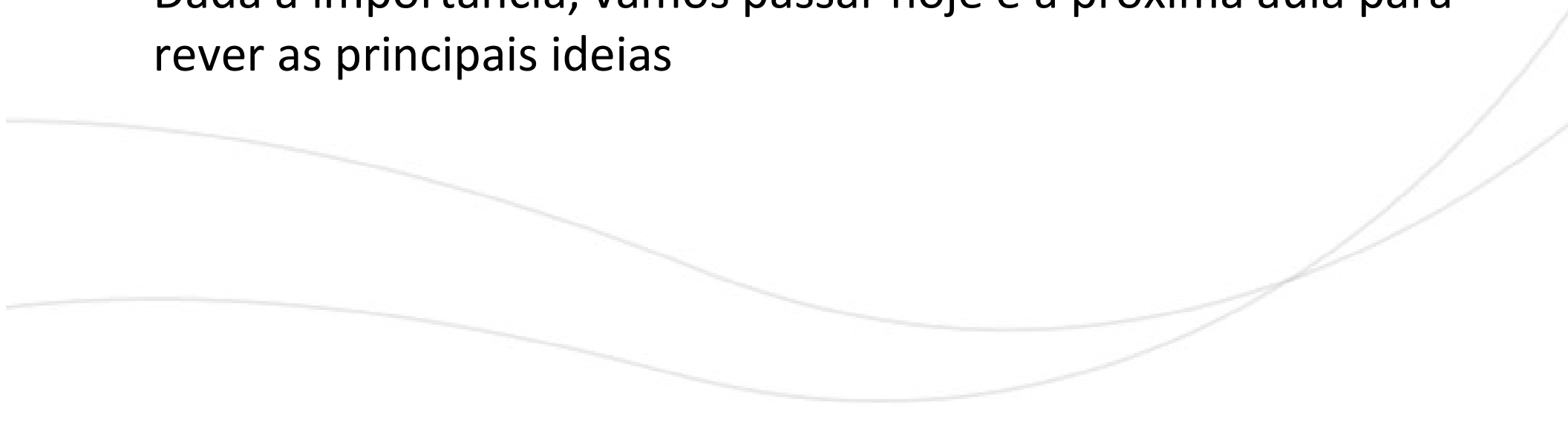
Econometria:

2 - Regressão Simples

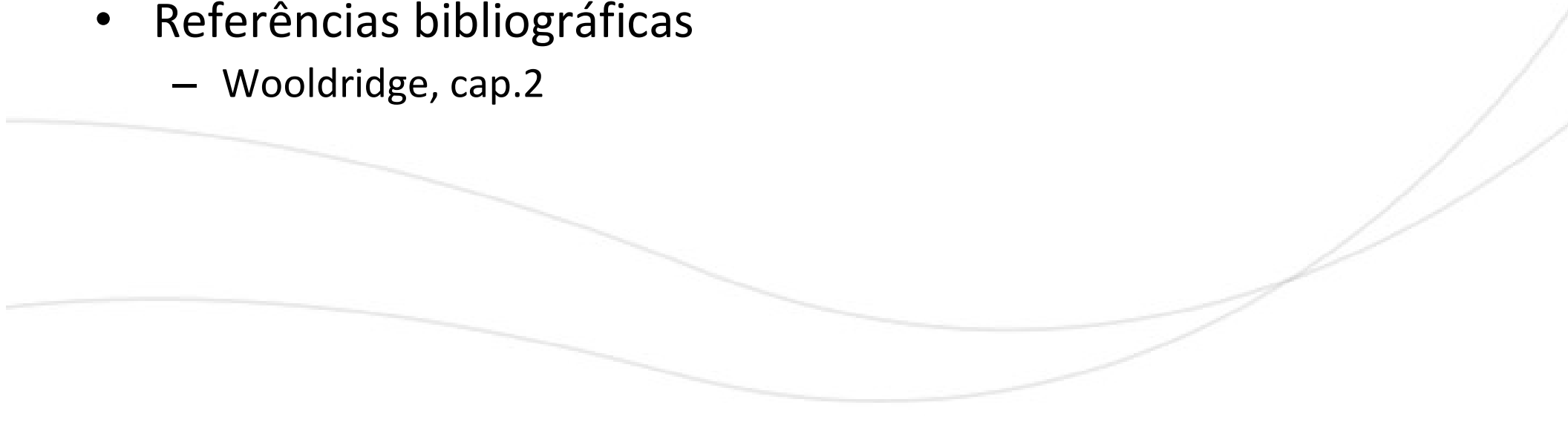
Parte deste material foi gentilmente cedido pelo Prof. Marco A.F.H.
Cavalcanti
cavalcanti@ipea.gov.br

Pontifícia Universidade Católica do Rio de Janeiro
PUC-Rio

Motivação

- Regressões lineares são indiscutivelmente a abordagem de modelagem mais popular em finanças e economia
 - Transparente e intuitivo
 - Técnica muito robusta; fácil de construir
 - Mesmo que não esteja interessado em causalidade, é útil para descrever os dados
 - Dada a importância, vamos passar hoje e a próxima aula para rever as principais ideias
- 

Sumário

- O modelo de regressão linear simples
 - Definição e terminologia
 - Estimação
 - Propriedades algébricas
 - Qualidade do ajuste
 - Unidades de medida
 - Forma funcional e não-linearidade
 - Propriedades estatísticas
 - Referências bibliográficas
 - Wooldridge, cap.2
- 

Introdução

- Origem histórica
 - Em um importante artigo, Francis Galton (1886), observou que a altura média de crianças, nascidas de pais com uma dada estatura, possuía uma tendência a se mover - *regress* - na direção da altura média da população como um todo.
 - Karl Pearson, em 1886, confirmou os resultados de Galton após coletar mais de 1000 registros dos membros de um grupo familiar.
- Interpretação moderna
 - Análise de regressão é o estudo da dependência de uma variável, *a variável dependente*, em uma ou mais variáveis, *variáveis explicativas ou independentes*.

Regressão Simples

Definição e Terminologia

- Sejam y e x duas variáveis representando alguma população.
 - O objetivo é explicar y em função de x , ou seja, como y varia de acordo com mudanças em x .
- Regredir y contra x
- 3 pontos importantes:
 - Dado que não há uma relação precisa entre y e x , como levar em conta outros fatores que afetam y ?
 - Qual a relação funcional entre y e x ?
 - Como capturar uma relação *ceteris paribus* entre y e x (se for o caso)?

Regressão Simples

Definição e Terminologia

- Solução:
 - Considere a seguinte equação relacionando y e x

$$y = \beta_0 + \beta_1 x + u$$
 - Esta equação linear é conhecida como modelo de regressão simples.
- Terminologia:
 - y : variável dependente, variável explicada, variável de resposta, variável prevista, regressando, saída, efeito.
 - x : variável independente, variável explicativa, variável de controle, preditor, regressor, entrada, causa.
 - u : erro, distúrbio ou ruído.

Regressão Simples

Definição e Terminologia

- A variável u representa:
 - todos os outros fatores além de x que afetam a variável y ;
 - erros de medição;
 - forma funcional inadequada e
 - inerente variabilidade nos agentes econômicos.
- Em análise de regressão consideramos que u é não-observável.
- Repare que se todos os outros fatores além de x são mantidos fixos, então

$$\Delta y = \beta_1 \Delta x$$

Regressão Simples

Definição e Terminologia

- Exemplo:
 - Safra de soja e quantidade de fertilizante

$$safra = \beta_0 + \beta_1 fert + u$$

- Se todos os outros fatores que afetam a safra permanecerem constantes, então:

$$\Delta safra = \beta_1 \Delta fert$$

Regressão Simples

Definição e Terminologia

- A linearidade da equação anterior implica que uma mudança de uma unidade em x , tem o mesmo efeito em y .
- O modelo de regressão linear realmente permite que conclusões *ceteris paribus* sejam obtidas?
 - Infelizmente NÃO!!!!
- Existe solução?
 - Sim, impondo restrições na variável u .

Regressão Simples

Definição e Terminologia

- Algumas hipóteses sobre a variável u :
 - Média nula $E(u) = 0$
 - Média condicional nula $E(u | x) = E(u) = 0$
- Questão:
 - Suponha que a nota final dos alunos em um exame depende da frequência dos alunos e de fatores não-observáveis, tais como habilidade. A equação $nota = \beta_0 + \beta_1 freq + u$ satisfaz a premissa de média condicional nula do erro?

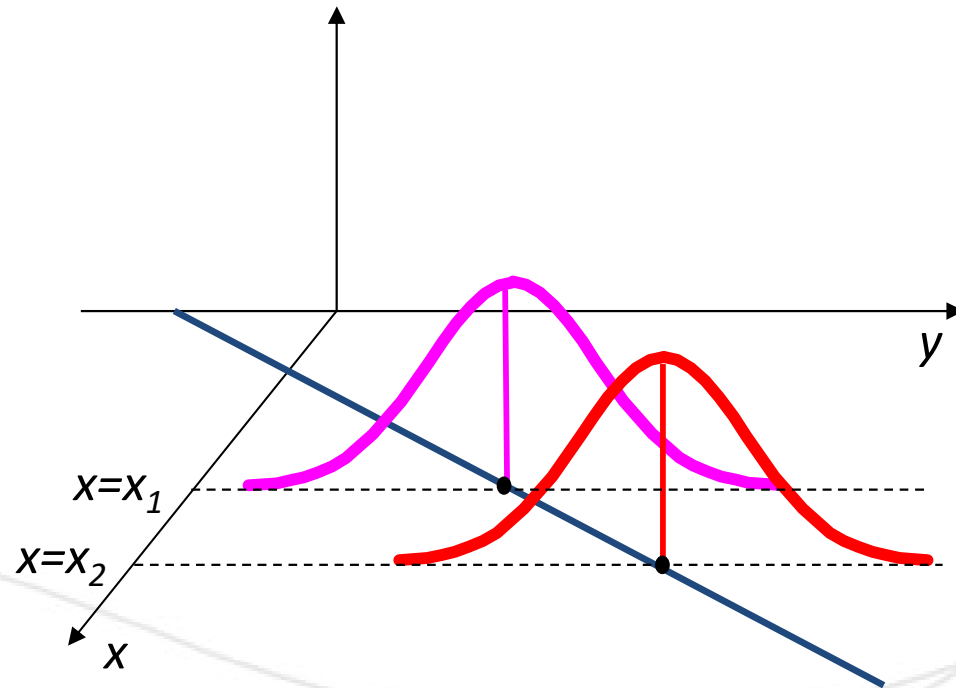
Função de Regressão Populacional

Definição

- Se $E(u|x)=0$, a função de regressão populacional (FRP) é definida por

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

- A FRP é uma função linear de x . Para qualquer valor de x , a distribuição de y está centrada em torno de $E(y|x)$.



Estimação dos Parâmetros

- Como estimar os parâmetros β_0 e β_1 na equação de regressão?
 - É necessário uma amostra da população!

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

uma amostra aleatória de tamanho n da população.

- Como esta amostra veio do modelo

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

pode-se escrever

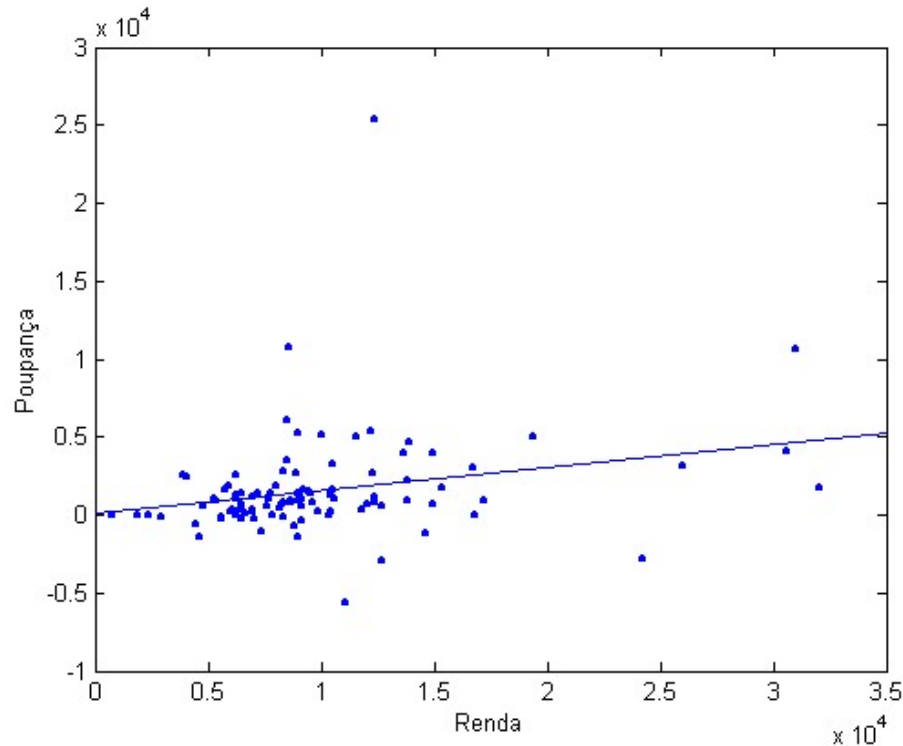
$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Estimação dos Parâmetros

Exemplo: Poupança e Renda

- Os dados

- Dados de poupança e da renda de 100 famílias no ano de 1970.



- A reta representa a equação

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

Estimação dos Parâmetros

- Como utilizar os dados para estimar os parâmetros?

- Deve-se lembrar que

$$E(u) = 0 \text{ e } E(xu) = 0$$

- Logo,

$$\begin{aligned} E(y - \beta_0 - \beta_1 x) &= 0 \\ E[x(y - \beta_0 - \beta_1 x)] &= 0 \end{aligned}$$

- Portanto, pode-se usar os dados amostrais para resolver o problema

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

Estimação dos Parâmetros

- Solução

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

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

onde

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{e} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- Atenção:

$$\sum_{i=1}^n (x_i - \bar{x})^2 > 0$$

Estimação dos Parâmetros

- Os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$

são chamados de estimadores de mínimos quadrados.

- Para justificar este nome, define-se o valor estimado para variável y dado que $x=x_i$ como

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

- O resíduo para a observação i é a diferença entre o valor real y_i e o seu valor estimado

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

- ATENÇÃO**: os resíduos não são os erros u_i definidos anteriormente!

Estimação dos Parâmetros

- Voltando ao problema...
- Suponha que os parâmetros do modelo de regressão linear simples são estimados de forma a tornar a soma dos quadrados dos resíduos

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

o menor possível.

- A solução é obtida ao derivar $S(\hat{\beta}_0, \hat{\beta}_1)$ em relação a $\hat{\beta}_0$ e $\hat{\beta}_1$

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Estimação dos Parâmetros

- Dividindo a primeira por $2n$ e com um pouco de álgebra têm-se

$$-\frac{2 \sum_1^n y_i}{2n} + \frac{2 \sum_1^n \hat{\beta}_0}{2n} + \frac{2 \sum_1^n \hat{\beta}_1 x_i}{2n} = \frac{0}{2n}$$

$$0 = -\bar{y} + \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

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

- Substituindo $\hat{\beta}_0$ na segunda equação

$$-2 \sum_1^n x_i (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) = 0$$

$$\sum_1^n x_i (y_i - \bar{y}) + \hat{\beta}_1 \sum_1^n x_i (\bar{x} - x_i) = 0$$

- A solução é

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

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

$$\hat{\beta}_1 = \frac{COV XY}{VAR X}$$

Função de Regressão Amostral

Definição

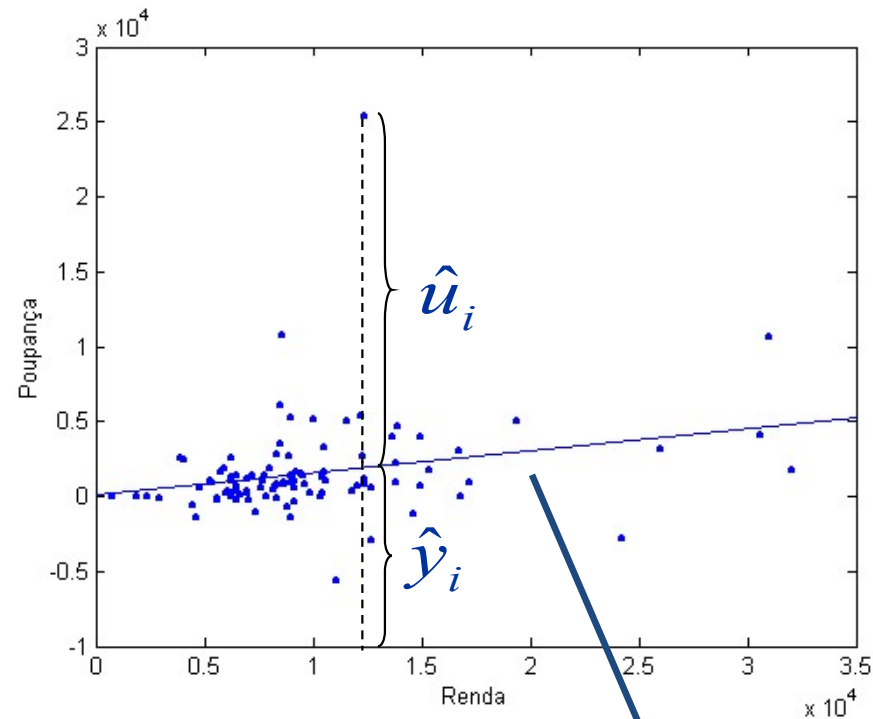
- Um vez estimados os parâmetros pode-se construir a função de regressão amostral (FRA)

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

- A FRA é a versão estimada da função de regressão populacional, FRP.
- Vale lembrar que a FRP é desconhecida.
- Como a FRA é obtida a partir de uma amostra, uma nova amostra irá gerar uma nova FRA.

Função de Regressão Amostral

Definição



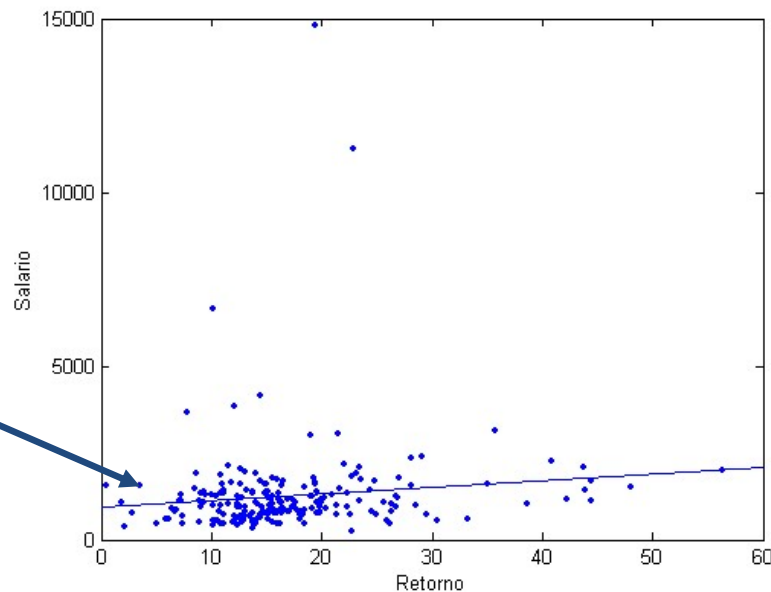
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Exemplos

Salário de executivos e retornos financeiros

- Sejam:
 - sal: salário anual de executivos em milhares de dólares
 - ret: retorno médio das ações da empresa
- Dados:
 - 209 registros para o ano de 1990 (fonte: Business Week, 06/05/1991)

$$\hat{sal} = 963.191 + 18.501ret$$

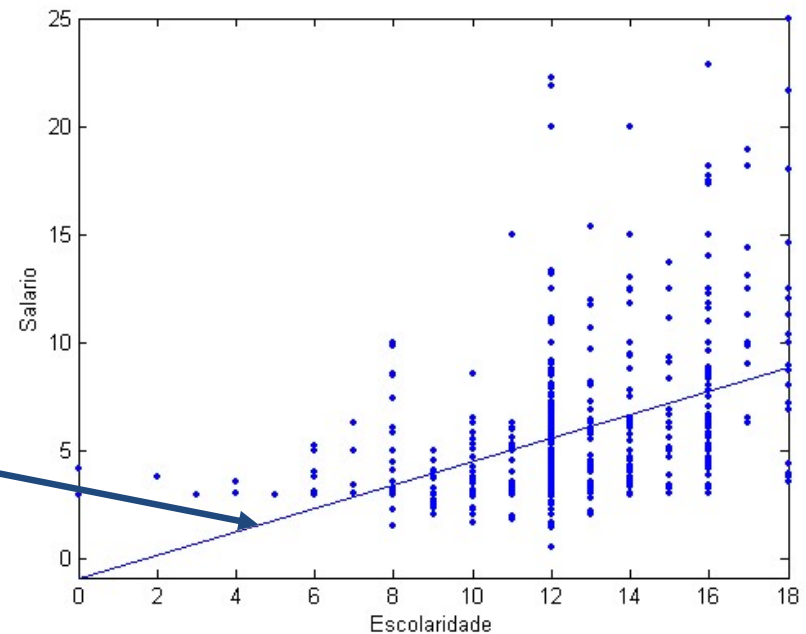


Exemplos

Salário e Educação

- Sejam:
 - sal: salário por hora
 - educ: anos de escolaridade
- Dados: Dados de 526 indivíduos no ano de 1976

$$\hat{s\acute{a}l} = -0.90 + 0.54educ$$



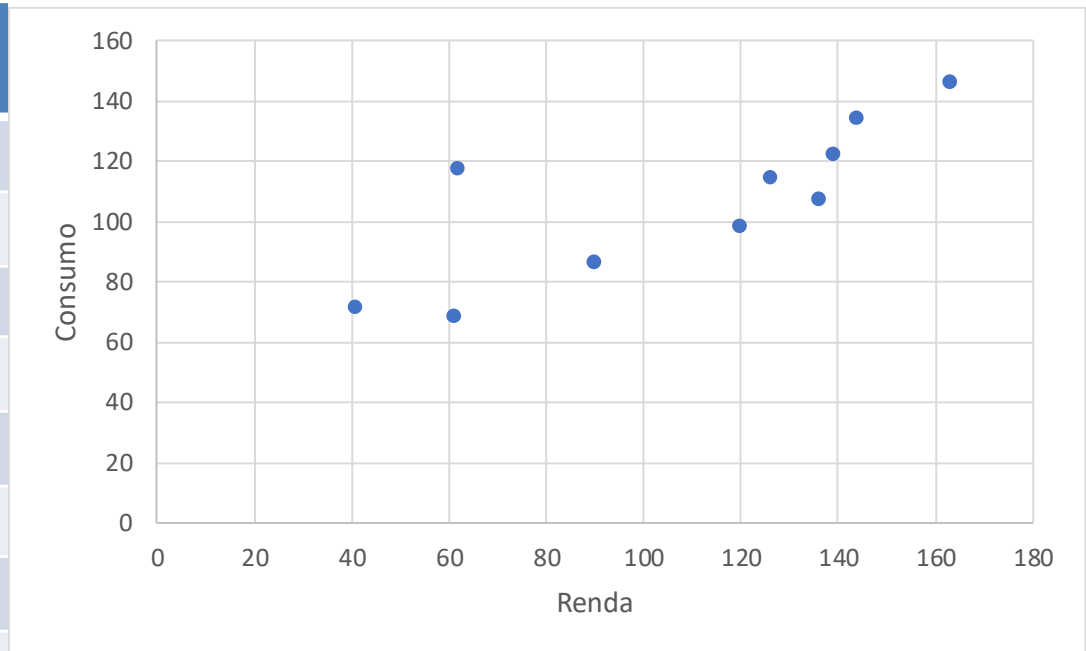
- Qual o problema com o resultado obtido?

Exercício

Salário e Consumo

- Estime a relação entre consume e renda com os dados abaixo:

Indivíduo "i"	Consumo	Renda
1	122	139
2	114	126
3	86	90
4	134	144
5	146	163
6	107	136
7	68	61
8	117	62
9	71	41
10	98	120

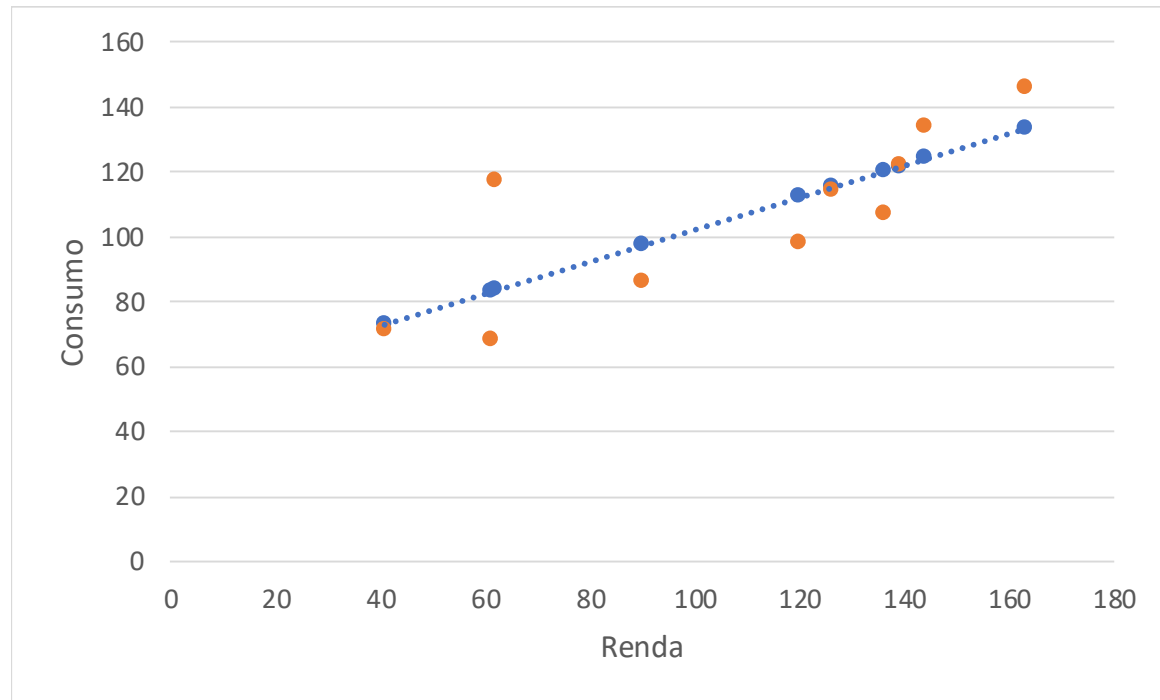


Exercício

Salário e Consumo

- Estime a relação entre consume e renda com os dados abaixo:

Indivíduo "i"	Consumo	Renda
1	122	139
2	114	126
3	86	90
4	134	144
5	146	163
6	107	136
7	68	61
8	117	62
9	71	41
10	98	120



$$\text{Consumo} = 52,69 + 0,4954 \times \text{Renda} + e$$

Mínimos Quadrados Ordinários

Propriedades Algébricas dos Estimadores

- A soma dos resíduos, e conseqüentemente a média, é ZERO.

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

- A covariância amostral entre os regressores e os resíduos é ZERO implicando que

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

- O ponto (\bar{x}, \bar{y})
esta sempre sobre a reta de mínimos quadrados.

Mínimos Quadrados Ordinários

Propriedades Algébricas dos Estimadores

- A variável dependente pode ser decomposta em dois termos: o valor estimado e o resíduo da regressão, isto é

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

- Pela decomposição acima nota-se que a média da variável dependente estimada é igual a média da própria variável dependente.
- A covariância amostral entre os resíduos e o valor estimado da variável dependente é ZERO, implicando que

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

Mínimos Quadrados Ordinários

Qualidade do Ajuste

- Define-se
 - Soma total dos quadrados (SST – *Total Sum of Squares*)

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

- Soma dos quadrados ajustados ou explicados (SSE – *Explained Sum of Squares*)

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

- Soma dos quadrados dos resíduos (SSR – *Residual Sum of Squares*)

$$SSR \equiv \sum_{i=1}^n \hat{u}_i^2$$

Mínimos Quadrados Ordinários

Qualidade do Ajuste

- Pela definição de SST , SSE e SSR , chega-se a seguinte relação

$$SST = SSE + SSR$$

- Qual a interpretação para SST , SSE e SSR ?
- Como medir a qualidade do ajuste a partir dos valores de SST , SSE e SSR ?
 - Coeficiente de determinação ou R^2

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

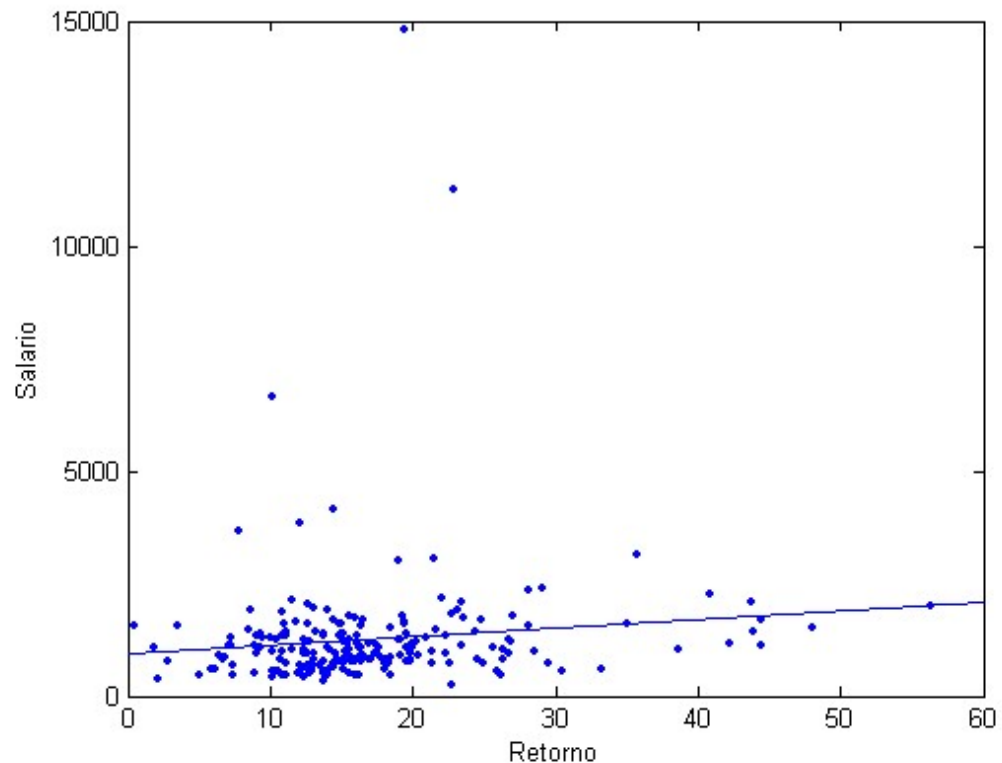
- Qual a interpretação para o coeficiente de determinação?

Exemplos

Salário de executivos e retornos financeiros

- Sejam:
 - sal: salário anual de executivos em milhares de dólares
 - ret: retorno médio das ações da empresa
- Dados:
 - 209 registros para o ano de 1990 (fonte: Business Week, 06/05/1991)

$$\hat{sal} = 963.191 + 18.501ret$$
$$n = 209, \quad R^2 = 0.0132$$



Unidade de Medida

- Qual o efeito da mudança da unidade de medida das variáveis nos resultados da estimação por mínimos quadrados?
 - Considere o exemplo anterior. Suponha que, em vez de milhares de dólares, o salário seja medido em dólares.
 - Neste caso o resultado da regressão é
$$\hat{s\hat{a}l} = 963191 + 18501ret$$
$$n = 209, \quad R^2 = 0.0132$$
- O que acontece com os parâmetros estimados quando a variável dependente é multiplicada por k e a variável explicativa é multiplicada por c ?

Unidade de Medida

- Ou seja, se nós deslocarmos y e x para cima por c e k respectivamente, o que ocorre?
- A inclinação estimada mudará?

Unidade de Medida

- Apenas o intercepto estimado será alterado
- Matematicamente, é fácil ver porque...

$$y = \alpha + \beta x + u$$

$$y + c = \alpha + c + \beta x + u$$

$$y + c = \alpha + c + \beta(x + k) - \beta k + u$$

$$y + c = (\alpha + c - \beta k) + \beta(x + k) + u$$



New intercept

Slope the same

Unidade de Medida

- Apenas o intercepto estimado será alterado
- Matematicamente, é fácil ver porque...

$$y = \alpha + \beta x + u$$

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

$$y + c = \alpha + c + \beta x + u$$

$$y + c = \alpha + c - \beta k + \beta(x + k) + u$$

$$y + c = (\alpha + c - \beta k) + \beta(x + k) + u$$

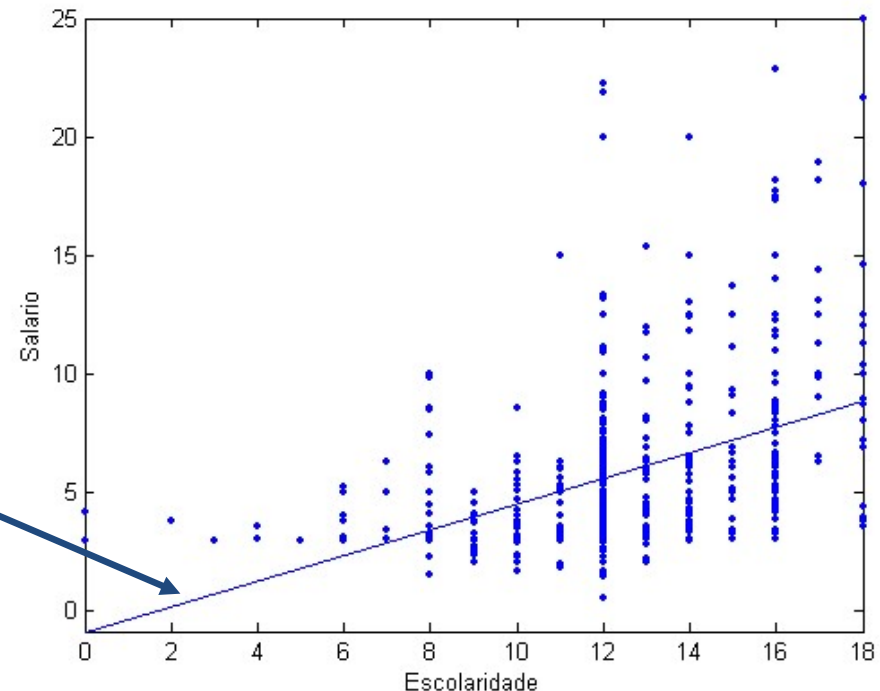
New intercept

Slope the same

Forma Funcional e Não-linearidade

- Não-linearidade nos parâmetros x não-linearidade nas variáveis.
- Neste curso será tratada apenas não-linearidade nas variáveis
- Parâmetros serão tratados como lineares
- Qual a vantagem de se trabalhar com modelos não-lineares nas variáveis?
 - Exemplo: salário x educação


$$\hat{s\acute{a}l} = -0.90 + 0.54educ$$



Forma Funcional e Não-linearidade

- Assumir que o CEF causal é linear nem sempre é tão realista
- Por exemplo. considere a seguinte regressão já comentada

$$\textit{Salário} = \alpha + \beta \textit{educ} + u$$

- Você acredita que uma relação linear entre o número de anos de educação e o nível de salários seria realista?
 - Como podemos ajustar/corrigir/melhorar isso?
- 

Forma Funcional e Não-linearidade

- Melhor suposição é que cada ano de educação leva a um aumento constante proporcional (ou seja, percentual) dos salários
- Aproximação desta intuição pode ser capturada por...

$$\ln(\text{Salário}) = \alpha + \beta \text{educ} + u$$

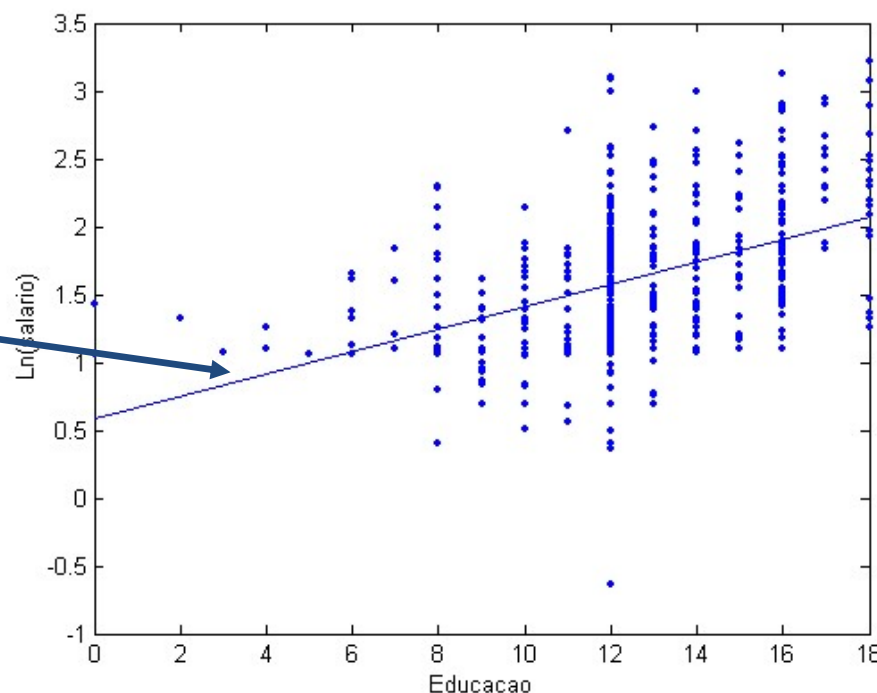
- Ou seja a especificação linear é muito flexível porque pode capturar relações lineares entre variáveis não-lineares

Forma Funcional e Não-linearidade

Exemplo: Salário e Educação

- Sejam:
 - $\ln sal$: logaritmo natural do salário horário
 - $educ$: anos de escolaridade
- Dados:
 - Dados de 526 indivíduos no ano de 1976

$$\widehat{\ln sal} = 0.584 + 0.083educ$$



Forma Funcional e Não-linearidade

- Quadro resumo:

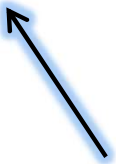
Modelo	Variável Dependente	Variável Independente	Interpretação de β_1
Nível-nível	y	x	$\Delta y = \beta_1 \Delta x$
Nível-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\% \Delta x$
Log-nível	$\log(y)$	x	$\% \Delta y = (100\beta_1) \Delta x$
Log-log	$\log(y)$	$\log(x)$	$\% \Delta y = \% \Delta x$

- Pergunta: como mudam os coeficientes estimados em cada um dos casos acima quando mudamos as unidades de medida de y e/ou x ?

Utilidade das regressões em Log

- As variáveis de log são úteis porque

$$100 * \Delta \ln(y) \approx \% \Delta y$$



Eu pessoalmente não gosto dessa notação para "alteração percentual", mas todo mundo usa isso.


- Nota: Quando eu (e outros) dizem “Log”, nós realmente nos referimos ao logaritmo natural, “Ln”. Por exemplo. se você usar a função "log" no Stata, isso significa que você quer dizer "ln"

Interpretando regressões log-nível

- Na estimativa da equação de $\ln(\text{salário})$, 100β irá nos dizer $\%\Delta\text{salário}$ que dever ocorrer para um ano adicional de educação. Para ver isso ...

$$\ln(\text{salario}) = \alpha + \beta \text{educ} + u$$

Cuidado com isso "=". É apenas "igual a" se este for o verdadeiro β , e a regressão for univariada. O melhor entendimento é: "está associado a".


$$\Delta \ln(\text{salario}) = \beta \Delta \text{educ}$$

$$100 \times \Delta \ln(\text{salario}) = (100\beta) \Delta \text{educ}$$

$$\%\Delta \text{salario} \approx (100\beta) \Delta \text{educ}$$

Interpretando regressões log-nível

- A mudança proporcional em y para uma dada mudança em x é assumida como constante
- A mudança em y não é assumida como constante ... ela aumenta à medida que x aumenta
- Especificamente, $\ln(y)$ é assumido como linear em x ; mas y não é uma função linear de x ...

$$\ln(y) = \alpha + \beta x + u$$

$$y = \exp(\alpha + \beta x + u)$$

Interpretando regressões log-nível

- Voltando a interpretação
- Suponha que você tenha estimado a equação salarial (onde os salários são \$ / hora) e tenha...

$$\ln(\text{salarío}) = 0.584 + 0.083\text{educ}$$

- O que um ano adicional de educação leva a você?
 - Resposta = aumento de 8,3% nos salários.
- Algum problema potencial com a especificação?
- Devemos interpretar a interceptação?

Interpretando regressões log-log

- Se estimarmos

$$\ln(y) = \alpha + \beta \ln(x) + u$$

- β é a elasticidade de y em relação a x !
- ou seja, β é a variação percentual em y para uma variação de, por exemplo, 1% em x .
- Uma variação percentual em x corresponde a $\beta \Delta x\%$ em y
- Nota: a regressão assume que a elasticidade é constante entre y e x independentemente do nível de x

Interpretando regressões log-log

- Suponha que você calculou que o modelo de salário do CEO usando registros obteve o seguinte:
- $\ln(\text{salário}) = 4,822 + 0,257 \ln(\text{vendas})$
- Qual é a interpretação de 0,257?

$$\ln(\text{salario}) = 4,822 + 0,257 \ln(\text{vendas})$$

Resposta = Para cada aumento de 1% nas vendas, o salário aumenta em 0,257%

Interpretando regressões nível-log

- Se estimarmos

$$y = \alpha + \beta \ln(x) + u$$

- $\beta/100$ é a mudança em y para 1% de mudança x

Interpretando regressões nível-log


- Suponha que você calculou que o modelo de salário do CEO usando registros obteve o seguinte quando o salário é expresso em milhares:

$$salario = 4.822 + 1.812,5 \ln(vendas)$$

- Qual a interpretação do valor 1.812,50?

Resposta = Para cada aumento de 1% nas vendas, o salário aumenta em \$18,125

Interpretando Regressões

- Voltando a questão:
 - Pergunta: como mudam os coeficientes estimados em cada um dos casos acima quando mudamos as unidades de medida de y e/ou x ?
- 

Interpretando regressões em log

- Voltando a questão:
- Pergunta: como mudam os coeficientes estimados quando usamos regressões em log quando mudamos as unidades de medida de y ?

Resposta = Apenas o intercepto se altera; a inclinação não é afetada porque mede a mudança proporcional em y no modelo Log-nível

$$\log(y) = \alpha + \beta x + u$$

$$\log(c) + \log(y) = \log(c) + \alpha + \beta x + u$$

$$\log(cy) = (\log(c) + \alpha) + \beta x + u$$

Interpretando regressões em log

- E quando mudamos as unidades de x ?

Resposta = É a mesma lógica

$$y = \alpha + \beta \log(x) + u$$

$$y + \beta \log(c) = \alpha + \beta \log(x) + \beta \log(c) + u$$

$$y = (\alpha - \beta \log(c)) + \beta \log(cx) + u$$

Interpretando regressões em log

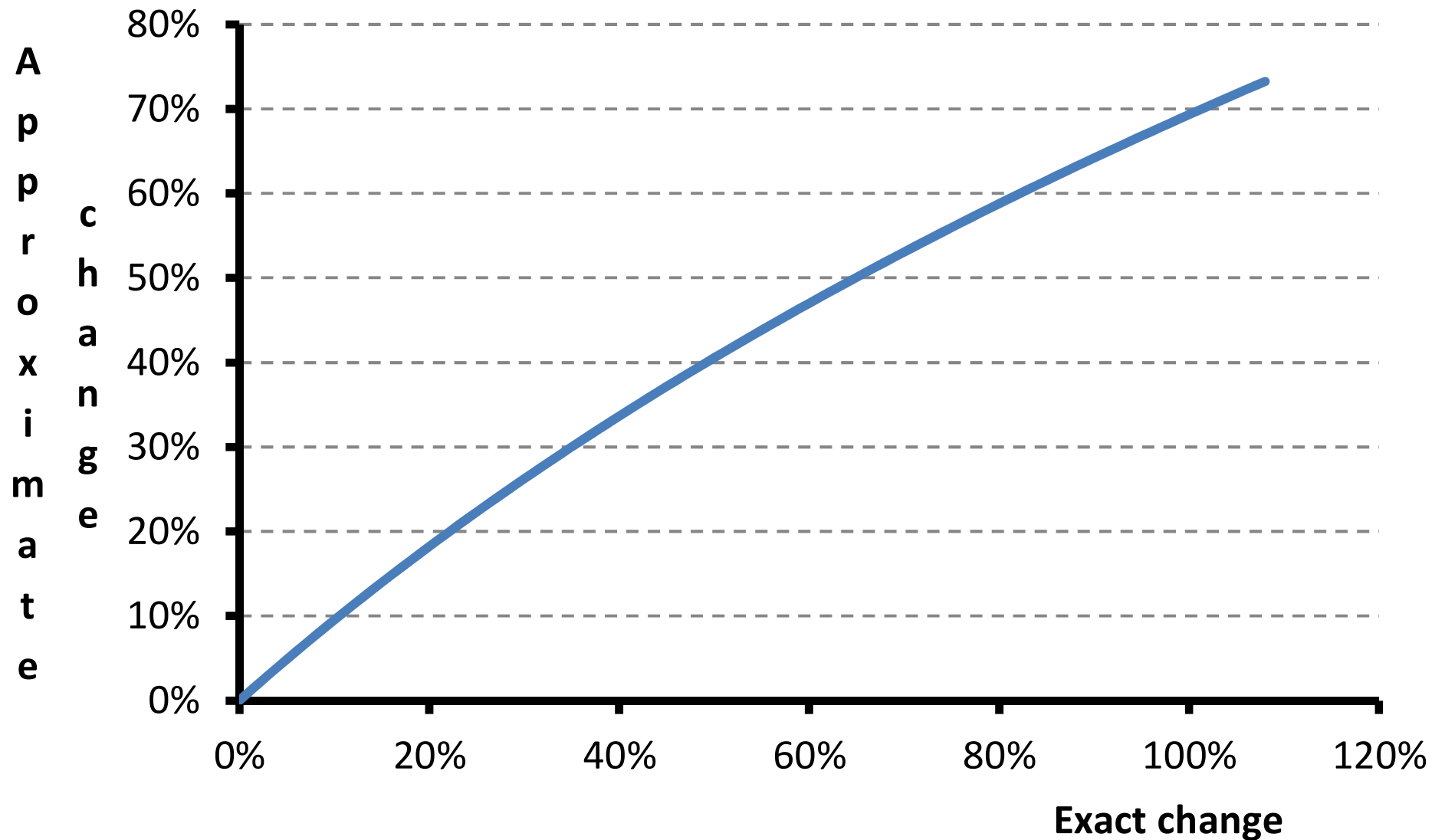
- Mensagem principal - Se você redimensionar uma variável, isso não afetará o coeficiente de inclinação porque você está apenas observando as alterações proporcionais
- Certa vez, um autor argumentou que permitir a entrada de capital no país causou uma variação de **-120%** nos preços das ações durante os períodos de crise.
- Você vê algum problema com isso?

Claro! Uma queda de 120% nos preços das ações não é possível. A verdadeira mudança percentual foi de 70%. Aqui é onde esse autor deu errado ...

Problemas da aproximação em log

- Erro de aproximação ocorre porque quando a verdadeira % Δy se torna maior, $100\Delta \ln(y) \approx \% \Delta y$ se torna uma aproximação cada vez pior
- Para ver isso, considere uma mudança de y para y' ...
 - **Ex. #1:** $\frac{y' - y}{y} = 5\%$, e $100\Delta \ln(y) = 4.9\%$
 - **Ex. #2:** $\frac{y' - y}{y} = 75\%$, mas $100\Delta \ln(y) = 56\%$

Problemas da aproximação em log



Problemas da aproximação em log

- O problema também ocorre para variações negativas
 - **Ex. #1:** $\frac{y' - y}{y} = -5\%$, e $100\Delta \ln(y) = -5.1\%$
 - **Ex. #2:** $\frac{y' - y}{y} = -75\%$, mas $100\Delta \ln(y) = -139\%$

Problemas da aproximação em log

- Portanto, se a mudança percentual implícita for grande, é melhor convertê-la em % real antes de interpretar a estimativa

$$\ln(y) = \alpha + \beta x + u$$

$$\ln(y') - \ln(y) = \beta(x' - x)$$

$$\ln(y'/y) = \beta(x' - x)$$

$$y'/y = \exp(\beta(x' - x))$$

$$[(y' - y)/y]\% = 100[\exp(\beta(x' - x)) - 1]$$

Coloque isso em uma nota de rodapé em seu artigo/trabalho!

Problemas da aproximação em log

- Agora podemos usar essa fórmula para ver qual é a variação real de% em y para $x' - x = 1$

$$[(y' - y)/y]\% = 100[\exp(\beta(x' - x)) - 1]$$

$$[(y' - y)/y]\% = 100[\exp(\beta) - 1]$$

- Se $\beta = 0.56$, a mudança percentual não é 56%, ela é:

$$100[\exp(0.56) - 1] = 75\%$$

Resumindo...

- Duas coisas para manter em mente sobre o uso de logs
 - O reescalonamento de uma variável não afeta os coeficientes de inclinação. Isso só afetará o intercepto
 - Log é apenas uma aproximação para % de alteração; pode ser uma aproximação muito ruim para grandes variações/mudanças
- Mas o uso de logs é útil
 - O uso de logs fornece coeficientes com interpretação atraente
 - Pode ignorar a unidade de medida das variáveis, já que elas são proporcionais às suas variações (Δs)
 - Logs de y ou x podem atenuar a influência de outliers

Rules of Thumb quando usar logs

- Útil para utilizar logs para variáveis com...
 - Quantidade monetária positiva
 - Valores integrais grandes (por exemplo, ativos totais)
- Não use logs para variáveis medidas em anos ou como proporções
- Se $y \in [0, \infty)$, pode utilizar $\ln(1 + y)$, mas tenha cuidado... a interpretação costumeira não será mais verdadeira...

O que há sobre usar $\ln(1 + y)$?

- Como $\ln(0)$ não existe, as pessoas usam $\ln(1 + y)$ para variáveis não negativas, ou seja, $y \in [0, \infty)$
- Seja cuidadoso ao interpretar as estimativas! A interpretação “normal” já não é mais verdadeira, especialmente se houver muitos zeros ou muitos valores pequenos em y [**Why?**]
- Ex. # 1: O que significa ir de $\ln(0)$ até $\ln(x > 0)$?
- Ex. # 2: $\ln(x' + 1) - \ln(x + 1)$ não é uma mudança percentual de x
- Nesse caso, pode ser melhor dimensionar y por outra variável, como tamanho da empresa, valor do PIB, ...

Mudança percentual...

- Qual é a mudança percentual no desemprego se passar de 10% para 9%?
 - Isso é 10% de queda
 - É uma queda de 1 ponto percentual
 - A alteração percentual é de $[(x_1 - x_0) / x_0] \times 100$
 - A mudança de ponto percentual é a mudança bruta em porcentagens

Por favor, tome cuidado para descrever e interpretar corretamente de seus resultados empíricos

Muitas pessoas (e eu quero dizer muitos mesmo) cometem erros sobre isso!

Modelos com formas quadráticas

- Considere $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$
- Efeito parcial de x é dado por...

$$\Delta y = (\beta_1 + 2\beta_2 x)\Delta x$$

- O que há de diferente nesse efeito parcial em tudo o que vimos até agora?

Resposta = Depende do valor de x . Então, precisaremos escolher um valor de x para avaliação (por exemplo, \bar{X})

Modelos com formas quadráticas

- Se $\hat{\beta}_1 > 0, \hat{\beta}_2 < 0$, então há uma relação parabólica
 - Ponto de inflexão = Máximo = $|\hat{\beta}_1 / 2\hat{\beta}_2|$
 - Saiba onde está esse ponto de inflexão! Não espere uma relação parabólica se estiver fora do alcance de x
 - Por exemplo: não espere que "para empresas com ativos totais maiores que 1 trilhão de reais, a relação se torne negativa.."
 - Valores estranhos ou improváveis podem implicar erros de especificação ou simplesmente significar que os termos quadráticos são irrelevantes e devem ser excluídos da regressão (nem sempre)

Mínimos Quadrados Ordinários

Propriedades Estatísticas dos Estimadores

- **VALIDADE DO MODELO**

- Algumas hipóteses importantes:

- (H1) No modelo populacional, a variável dependente y está relacionada à variável independente x e ao erro u da seguinte forma:

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

- (H2) Uma amostra aleatória de tamanho n

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

pode ser construída a partir do modelo populacional.

Mínimos Quadrados Ordinários

Propriedades Estatísticas dos Estimadores

– (H3) Média condicional nula

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

– (H4) Na amostra, as variáveis independentes x_i , $i = 1, \dots, n$, não são todas iguais. Ainda, nenhuma variável independente deve ser combinação linear perfeita de outra variável independente do modelo.

Mínimos Quadrados Ordinários

Propriedades Estatísticas dos Estimadores

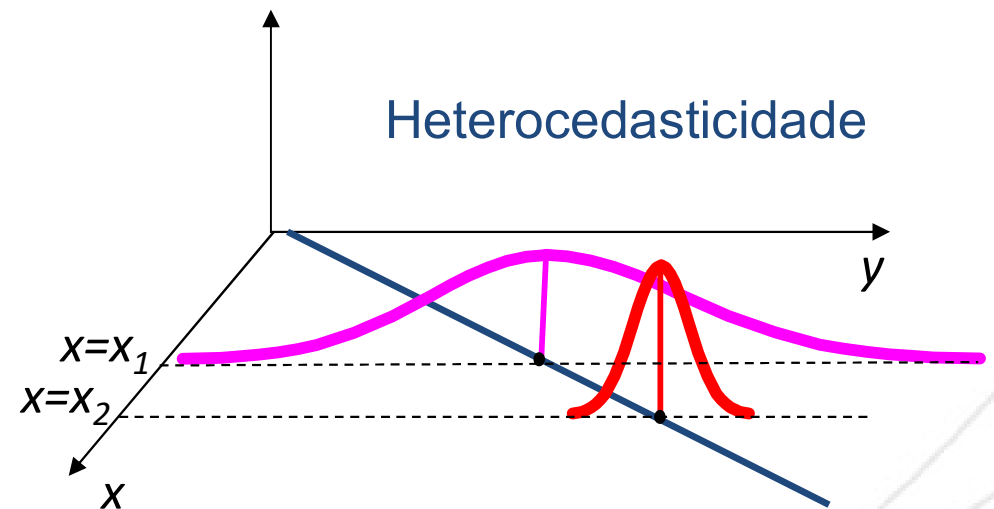
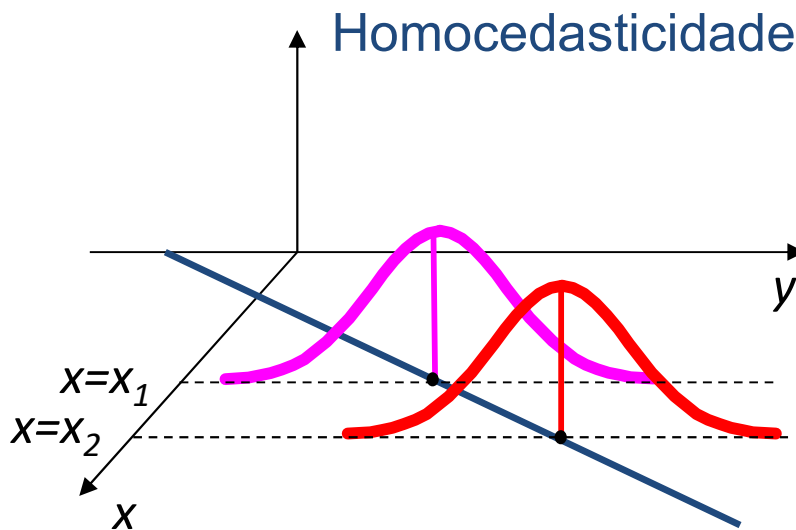
- Sob $H(4)$:
- Uma importante fonte de falha desta hipótese é a situação em que o número de observações na amostra é menor que o número de parâmetros a serem estimados no modelo ($n < k + 1$).
- Assim, uma importante implicação desta hipótese é que a amostra tenha ao menos uma observação para cada parâmetro que se pretende estimar, isto é, $n \geq k + 1$.

Mínimos Quadrados Ordinários

Propriedades Estatísticas dos Estimadores

- (H5) Homocedasticidade

$$\text{Var}(u | x) = \sigma^2$$



Mínimos Quadrados Ordinários

Propriedades Estatísticas dos Estimadores

- Teorema 1: sob as hipóteses (H1) - (H4) os estimadores de mínimos quadrados ordinários são não-tendenciosos, isto é

$$\begin{array}{l} E(\hat{\beta}_0) = \beta_0 \\ E(\hat{\beta}_1) = \beta_1 \end{array}$$

Mínimos Quadrados Ordinários

Propriedades Estatísticas dos Estimadores

- Teorema 2: sob as hipóteses (H1) - (H5)

$$\text{Var}(\hat{\beta}_0) = \frac{\frac{\sigma_u^2}{n} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

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

$$\text{Var}(\hat{\beta}_0) = \frac{\frac{\sigma^2}{n} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Por simplicidade de notação: $\sigma_u^2 = \sigma^2$

Mínimos Quadrados Ordinários

Propriedades Estatísticas dos Estimadores

- Como estimar σ^2 ?

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{(n-2)}$$


- Teorema 3: sob as hipótese (H1) - (H5)

$$E(\hat{\sigma}^2) = \sigma^2$$

Conceitos e definições gerais



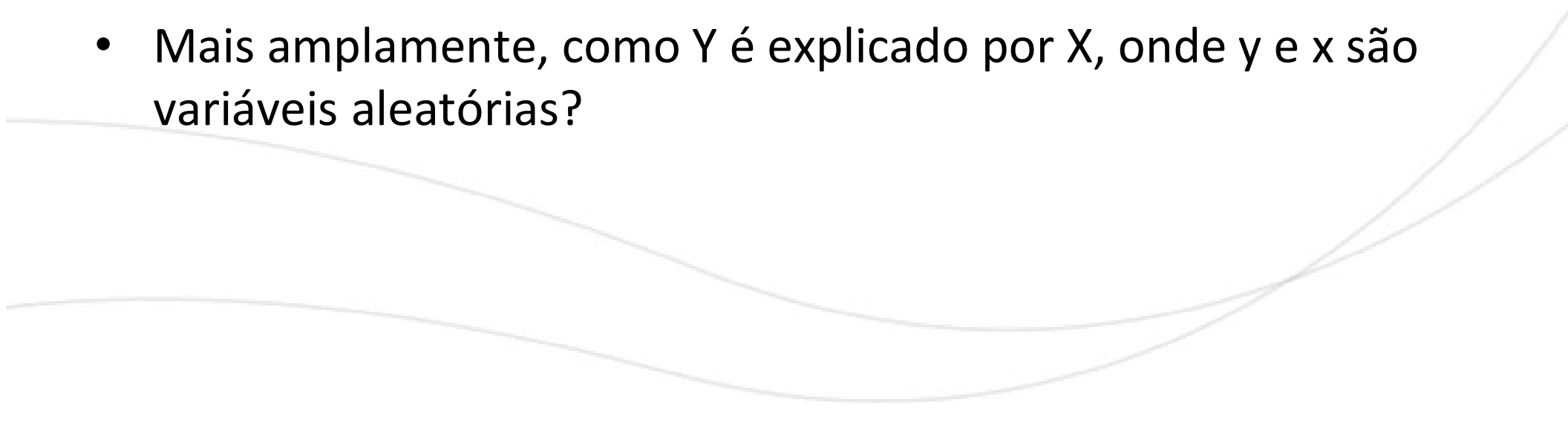
Regressão Linear

- Função de expectativa Condicional (CEF)
 - Modelo linear OLS
 - Estimação multivariada
 - Teste de hipóteses
 - Outras questões
- 

Motivação

- Regressões lineares são indiscutivelmente a abordagem de modelagem mais popular em finanças e economia
 - Transparente e intuitivo
 - Técnica muito robusta; fácil de construir
 - Mesmo que não esteja interessado em causalidade, é útil para descrever os dados

Motivação

- Como estudiosos, analistas e pesquisadores, estamos interessados em explicar e entender como o mundo funciona
 - Por exemplo: como as escolhas das empresas em relação à alavancagem são explicadas por suas oportunidades de investimento
 - Ou seja, se as oportunidades de investimento saltassem repentinamente por alguma razão aleatória, como poderíamos esperar que a alavancagem das empresas respondesse em média?
 - Mais amplamente, como Y é explicado por X , onde y e x são variáveis aleatórias?
- 

Variáveis Aleatórias

- É útil saber que qualquer variável aleatória y pode ser escrita como

$$y = E(y|x) + \varepsilon$$

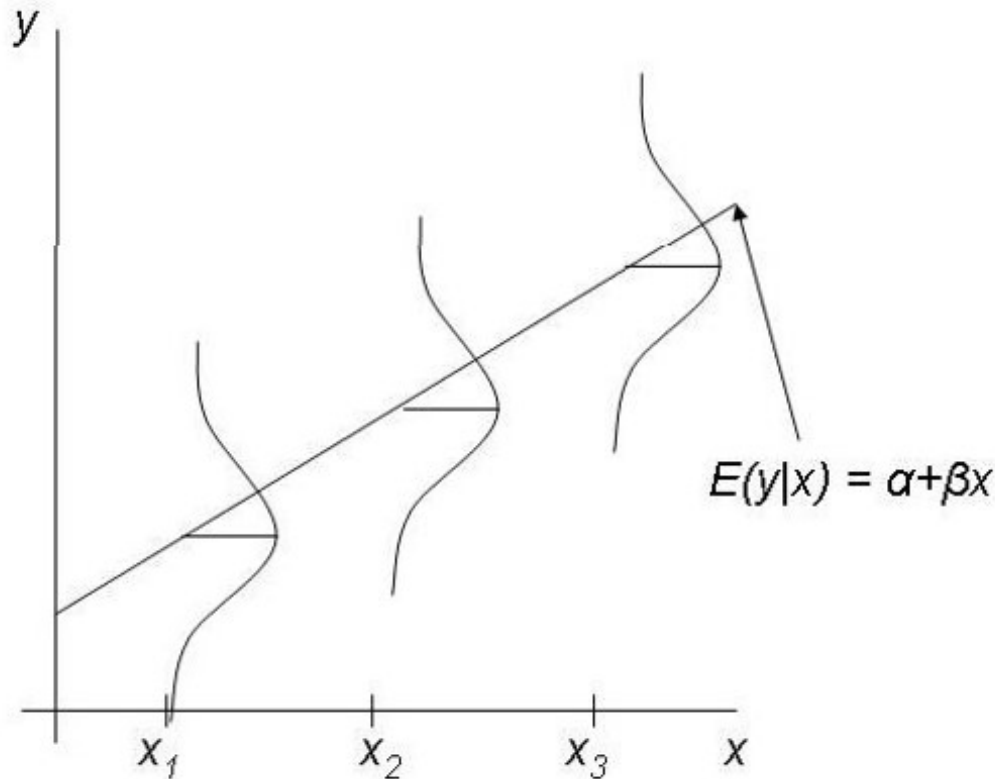
- onde (y, x, ε) são variáveis aleatórias e $E(\varepsilon | x) = 0$
- $E(y | x)$ é o valor esperado de y dado x
- Em outras palavras, y pode ser dividido em parte "explicado" por x , $E(y | x)$, e uma parte que é média independente de x , ε

Função de Expectativa Condicional

- $E(y|x)$ é o que chamamos de CEF (Conditional expectation function), e tem propriedades muito desejáveis
 - Maneira natural de pensar sobre a relação entre x e y
 - E , é melhor preditor de y dado x em um sentido de erro médio-quadrado mínimo
- Ou seja $E(y|x)$ minimiza $E[(y - m(x))^2]$, onde $m(x)$ pode ser qualquer função de x .


CEF visualmente...

- $E(y | x)$ é fixo, mas não observável

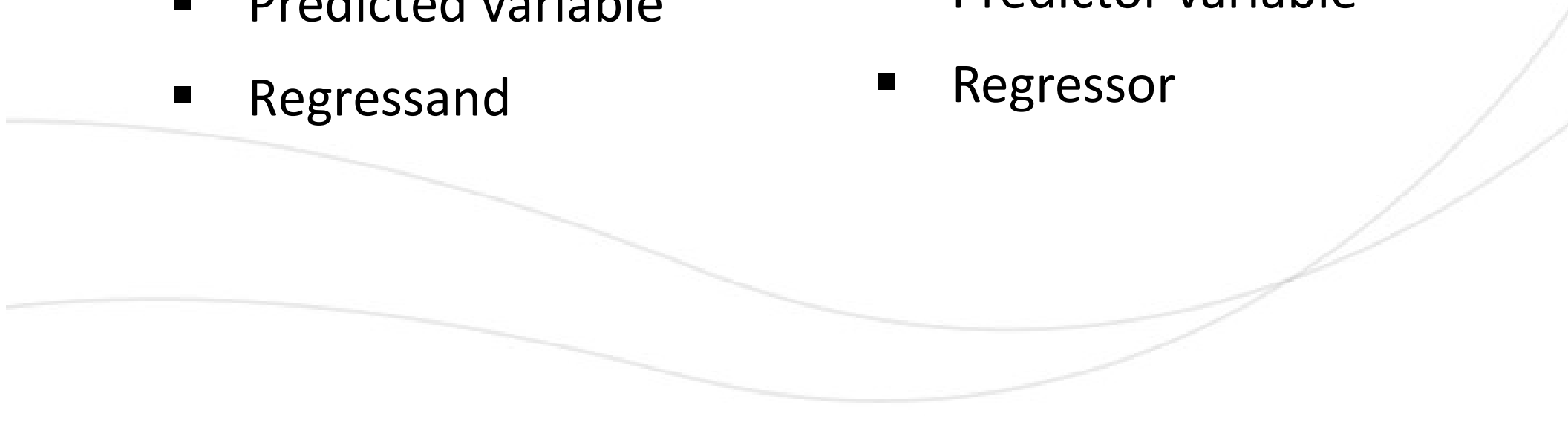


- Intuição: para qualquer valor de x , a distribuição de y é centrada em $E(y | x)$


Regressão Linear e a CEF

- Se feito corretamente, uma regressão linear pode nos ajudar a descobrir o que a CEF é
 - Considere o modelo de regressão linear, $y = \beta x + u$
 - y = variável dependente
 - x = variável independente
 - u = termo de erro (ou perturbação)
 - β = parâmetro de inclinação $E(y \mid x)$ é fixo, mas não observável
- 

Regressão Linear e a CEF

- Outros Termos para y ...
 - Outcome variable
 - Response variable
 - Explained variable
 - Predicted variable
 - Regressand
 - Outros Termos para x ...
 - Covariate
 - Control variable
 - Explanatory variable
 - Predictor variable
 - Regressor
- 

Detalhes sobre $y = \beta x + u$

- (y, x, u) são variáveis aleatórias
 - (y, x) são observáveis
 - (u, β) não são observáveis
 - u captura tudo o que determina y depois de contabilizar x
[Isso pode ser um monte de coisas!]
 - Queremos estimar β
- 

Ordinary Least Squares (OLS)

- Simplificando, o OLS encontra o β que minimiza o erro médio quadrático

$$\beta = \arg \min_b E[(y - bx)^2]$$

- Usando a condição de primeira ordem: $E[x(y - \beta x)] = 0$, temos $\beta = E[(xy)] / [E(x^2)]$
- Nota: por definição, o resíduo desta regressão, $y - \beta x$, não é correlacionado com x

Ordinary Least Squares (OLS)

- Simplificando, o OLS encontra o β que minimiza o erro médio quadrático

$$\beta = \arg \min_b E[(y - bx)^2]$$

- Usando a condição de primeira ordem: $E[x(y - \beta x)] = 0$, temos $\beta = E[(xy)] / [E(x^2)]$
- Nota: por definição, o resíduo desta regressão, $y - \beta x$, não é correlacionado com x

Ordinary Least Squares (OLS)

- Pode ser provado que...
 - βx é melhor* predição linear de y dado x
 - βx é melhor* aproximação linear de $E(y|x)$
 - * "Melhor" em termos de erro mínimo de média quadrática
- Isso é bastante útil. Ou seja mesmo se $E(y|x)$ é não-linear, a regressão nos dá a melhor aproximação linear do mesmo

Causalidade



Causalidade

- Precisamos ter cuidado aqui ...
 - Como x explica y , o que a regressão nos ajuda a entender, não é o mesmo que determinar o efeito causal de x em y
- Para isso, precisamos de mais suposições ...

Causalidade

- Suposição 1: $E(u) = 0$
 - Com o intercepto, isso é totalmente inócuo
 - Apenas mude a regressão para $y = \alpha + \beta x + u$, onde α é o termo de interceptação
 - Agora suponha que $E(u) = k \neq 0$
 - Poderíamos reescrever $u = k + w$, onde $E(w) = 0$
 - Então, o modelo se torna $y = (\alpha + k) + \beta x + w$
 - Intercepto é agora apenas $\alpha + k$ e erro, w , significa zero
- Ou seja Qualquer média diferente de zero é absorvida pelo intercepto

Causalidade

- Suposição 2: $E(u|x) = E(u) \rightarrow \text{CMI}$
 - Em outras palavras, a média de u (ou seja, porção inexplicável de y) não depende do valor de x
 - Isto é "**independência da média condicional**" (CMI)
 - Verdade se x e u forem independentes um do outro
 - Implica que u e x são não correlacionados
- Esta é a suposição fundamental que está sendo feita quando as pessoas fazem inferências causais

Causalidade

- Basicamente, a suposição diz que você tem o modelo CEF (Conditional expectation function) correto para o efeito causal de x em y
 - O CEF é causal se descreve diferenças nos resultados médios para uma mudança em x
 - isto é, aumentar em x os valores a e b é igual a

$$E(y|x = b) - E(y|x = a) \quad \textbf{[In words?]}$$

- É fácil ver que isso só é verdade se $E(u | x) = E(u)$
[Isso é feito no próximo slide...]

Causalidade


- Com o modelo $y = \alpha + \beta x + u$,
 - $E(y|x = a) = \alpha + \beta a + E(u|x = a)$
 - $E(y|x = b) = \alpha + \beta b + E(u|x = b)$
 - Assim, $E(y|x = b) - E(y|x = a) = \beta(b - a) + E(u|x = b) - E(u|x = a)$
- Isso apenas é igual ao que pensamos como o efeito "causal" de x mudar de a para b se **$E(u|x = b) = E(u|x = a)$** ... Isto é, a suposição de CMI se mantém

Causalidade

CMI versus Correlação

- CMI (o que implica x e u são não correlacionados) é necessário para não viés
[que é uma propriedade de amostra finita]
- Mas, nós só precisamos assumir uma correlação zero entre x e u para consistência
[que é uma propriedade de amostra grande]
- Mais sobre viés vs. consistência depois; mas normalmente nos preocupamos com a consistência, e é por isso que muitas vezes me refiro a correlações em vez de CMI

Causalidade – É plausível?

- É certo que existem muitas razões pelas quais essa suposição pode ser violada
 - Lembre-se, u captura todos os fatores que afetam y além de x ... E ele conterá muitos!
 - Vamos apenas fazer alguns exemplos ...
- 

Ex. # 1 - regressão de estrutura de capital

- Considere seguir a regressão em nível de empresa:

$$Leverage_i = \alpha + \beta Profitability_i + u_i$$

- CMI implica que média u é a mesma para cada rentabilidade
- É fácil encontrar algumas histórias porque isso não é verdade ...
 - # 1 - empresas não lucrativas tendem a ter maior risco de falência, o que, deveria significar uma menor alavancagem (endividamento)

Por outro lado....

 - # 2 - empresas não lucrativas acumularam menos dinheiro, o que, deveria significar que elas teriam mais alavancagem (endividamento)

Ex. # 2 - Investments

- Considere a regressão a seguir em nível de empresa:

$$Investment_i = \alpha + \beta Q_i + u_i$$

- CMI implica que média de u é a mesma para cada Q de Tobin
- É fácil encontrar algumas histórias que isso não é verdade ...
 - # 1 - Empresas com Q baixo podem estar em perigo e investir menos
 - # 2 - Empresas com Q alto podem ser menores, empresas mais jovens que têm mais dificuldade em levantar capital para financiar investimentos

Existe uma maneira de testar o CMI?

- Seja \hat{y} o valor previsto de y , ou seja,
 - $\hat{y} = \alpha + \beta x$, onde α e β são estimativas de OLS
 - E, deixe \hat{u} ser o residual, ou seja, $\hat{u} = y - \hat{y}$
- Podemos provar CMI se $E(\hat{u}) = 0$ e se \hat{u} não é correlacionado com x ?


Resposta: não! Por construção esses resíduos são médios zero e não correlacionados com x . Veja a derivação anterior de estimativas do OLS

Existe uma maneira de testar o CMI?


- O que as pessoas chamam de “estratégia de identificação” são aquelas que procuram por violações da CMI
 - Ou seja, se procura uma razão pela qual a perturbação do modelo está correlacionada com x
- Infelizmente, não é tão difícil assim ...
- **Tentar encontrar maneiras de garantir que a suposição do CMI possa ser realizada** é o cuidado principal que você deve ter em seus estudos e trabalhos

$$E(u|x) = E(u) \rightarrow \text{CMI}$$


Endogeneidade

- Muitos “avaliadores” criticam um modelo dizendo que ele tem um “problema de endogeneidade”, mas não dizem mais nada...
 - Mas o que significa dizer que existe um "problema de endogeneidade"?
- 

Endogeneidade

- Minha opinião: essas críticas vagas sobre “endogeneidade” suspeitam que algo está potencialmente errado, mas não sabem realmente porque ou como
 - Não seja assim, Quando criticar, seja específico sobre qual é o problema!
 - Violações ao CMI podem ser categorizadas em três pontos ... quais são?
- 

Três razões pelas quais o CMI é violado

- Viés de variável omitida
 - Viés de erro de medida
 - Viés de simultaneidade
-
- Vamos ver cada um deles com muito mais detalhes na sessão “Causalidade”
- 

O que "endógeno" significa para mim

- Um x "endógeno" é quando seu valor depende de y (ou seja, é determinado juntamente com y de tal forma que há viés de simultaneidade).

to me: Endogeneidade = “Tostines effect”

- Mas, alguns usam uma definição mais ampla que significa qualquer correlação entre x e u $\rightarrow E(u | x) = k \neq 0$
 - [por exemplo. Roberts & Whited (2011)]
- Por causa da confusão, evito usar “endogeneidade”; Eu recomendaria o mesmo para você
 - seja específico sobre a violação do CMI; apenas diga variável omitida, erro de medida ou viés de simultaneidade

Em suas apresentações ...

- Pense em "causalidade" ao apresentar seu trabalho ou paper
 - Ainda não formalizei as várias razões pelas quais as inferências "causais" não deviam ser feitas; mas eu gostaria que você tentasse pensar nisso
- 