Regressão Linear Simples – Aula 06a

Statistics for Business and Economics 7 edição, by Paul Newbold , William Carlson , Betty Thorne (cap. Simple Regression)

Bussab e Morettin (Cap 16 Regressão Linear Simples)

Statistics for Economics, Accounting and Business Studies, capítulo 7, Barrow (Correlation and Simple Regression)

Marislei Nishijima

Estatística II

Econometria

- Consiste no desenvolvimento de métodos estatísticos para estimar relações econômicas, testar teorias, avaliar e implementar políticas de governo e de negócios.
- Foca problema de coleta de dados não experimentais

Campos de estudo assemelhados

- Sociometria
- Cliometria

Passos para análise empírica

- Formulação cuidadosa da questão de interesse

Ex. Modelo de crime de Gary Becker (crimes têm recompensas econômicas, mas também custos)

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

y horas gastas em atividades criminosas

x1 “salário” por hora ocupada em atividade criminosa

x2 salário hora em emprego legal

x3 renda de outras atividades não criminosas e não emprego legal

x4 probabilidade de ser capturado

x5 probabilidade de ser condenado caso seja capturado

x6 sentença esperada se condenado, e

x7 idade

Tipos de dados

- **Dados de corte transversal**: indivíduos, empresas, cidades, estados e outras unidades tomados num único ponto do tempo.
- **Dados de série de tempo**: observações de uma variável (ou conjunto de variáveis) ao longo do tempo. Ex. Índice de Gini no Brasil, renda per capita do país.
- **Dados de cortes transversais agrupados**: exemplo suplemento de saúde da PNAD 1998 e 2003 com informações por indivíduos (340 mil indivíduos entrevistados).
- **Dados de Painel ou longitudinal**: vários indivíduos ou unidades (os mesmos) observados ao longo do tempo.

Causalidade e a noção de Ceteris Paribus na análise econométrica

- Causalidade: encontrar uma simples relação entre duas ou mais variáveis pode ser sugestivo, mas é preciso determinar qual variável afeta (causa) qual.
- Ceteris Paribus: realiza-se alteração em x e verifica-se seu impacto sobre y quando todas as demais variáveis explicativas permanecem constantes.
- Veja Exemplos de correlação espúria:
- https://www.tylervigen.com/spurious-correlations

Exemplos

- Mais policiais na rua reduz criminalidade?
- Medindo o retorno da educação: a maior escolaridade afeta positivamente o salário recebido?
- Qual o efeito dos novos fertilizantes sobre a agricultura?
- Qual o efeito do salário mínimo sobre o desemprego?

Medidas de Posição

Moda é o elemento de maior freqüência.

Mediana é o valor que divide um conjunto ordenado sequencialmente ao meio.

A **mediana**, ao contrário da média, **não é sensível a valores extremos**.

Estimadores

- Média

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Variância

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Medidas de Dispersão

Variância

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Desvio-Padrão

$$\mathrm{dp}(X) \equiv \sigma_X = \sqrt{\mathrm{var}(X)}$$

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\mathrm{var}\ (X) = \frac{1}{n}\sum_{i=1}^{n}(X_i^{\ 2} - 2X_i\overline{X} + \overline{X}^2)$$

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}X_i^{\ 2} - \frac{1}{n}\sum_{i=1}^{n}2X_i\overline{X} + \frac{1}{n}\sum_{i=1}^{n}\overline{X}^2$$

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}X_i^{\ 2} - 2\overline{X}\frac{1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}n\overline{X}^2$$

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}X_i^{\ 2} - 2\overline{X}^2 + \overline{X}^2$$

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}X_i^{\ 2} - \overline{X}^2$$

Ou, em outras palavras:

var(X) = média dos quadrados - quadrado da média

Relações entre variáveis

Covariância: mede a existência de associação linear entre 2 variáveis aleatórias

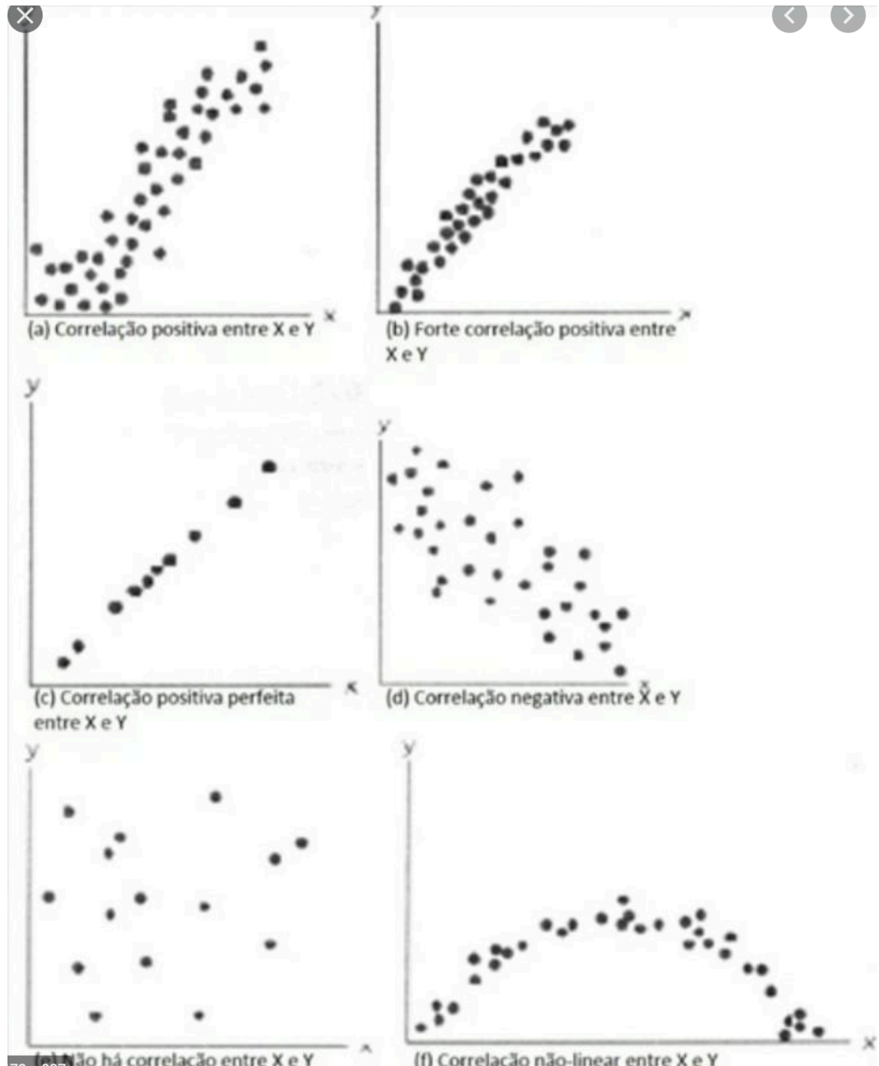
$$\text{covar}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Coeficiente de correlação: valores entre -1 e 1

Mede a relação (grau de associação) linear entre as variáveis

$$\text{corr}(X,Y) \equiv \rho_{XY} = \frac{\text{covar}(X,Y)}{\text{dp}(X).\text{dp}(Y)}$$

Exemplos de correlações



Sejam X e Y variáveis aleatórias (v.a.) Propriedades de variáveis aleatórias

covar(aX,bY) = a.b.covar(X,Y) , a e b constantes

I) var(aX) = a^2var(X)

II) dp(aX) = a.dp(X)

III) var(X+a) = var(X)

IV) dp(X+a) = dp(X)

V) covar(X,Y) = média dos produtos - produto da média

$$\mathrm{covar}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}$$

Sejam X e Y variáveis aleatórias (v.a.) Propriedades de variáveis aleatórias

VI) $covar(aX,bY) = a.b.covar(X,Y)$

VII) $var(X+Y) = var(X) + var(Y) + 2covar(X,Y)$

VIII) $var(X-Y) = var(X) + var(Y) - 2covar(X,Y)$

Se X e Y independes $covar(X, Y)=0$

Modelo de Regressão

- relação entre 2 variáveis X e Y
- Quanto de Y pode ser explicado por X?
- Nunca há relação exata entre X e Y, como considerar outros fatores que afetam X?

Equação Linear Simples

$$y = \beta_0 + \beta_1 X + \mu \qquad \mu \sim N(0, \sigma)$$

$\mu \rightarrow$ termo erro: representa outros fatores não observados, além de X, que afetam Y

Modelo de Regressão

Se outros fatores fixos $\Delta\mu = 0$ então $\Delta Y = \beta_1 \Delta X$

$\beta_1 \rightarrow$ Parâmetro de interesse (inclinação entre Y e X)

$\beta_0 \rightarrow$ Intercepto

Exemplo 1: Produção de Soja

Produção = $\beta_0 + \beta_1 . fertilizante + \mu$

Exemplo 2: Impacto da educação sobre rendimento

Salário= $\beta_0 + \beta_1 . educ + \mu$ (salário hora e número de anos de estudo)

$\beta_1 \rightarrow$ mede a variação no salário hora dado um ano a mis de educação, mantendo os demais fatores fixos.

$$y = \beta_0 + \beta_1 X + \mu \qquad \mu \sim N(0, \sigma)$$

Hipóteses do modelo de regressão

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

O valor médio de *u* não depende do valor de *X*.
Hipótese de média condicional zero

Exemplo:
Sob esta hipótese: E(aptidão | 8) = E(aptidão | 16)
Ou seja, aptidão não muda com anos de educação.

$$E(u/X) = 0 \Rightarrow E(y/X) = \beta_0 + \beta_1 X$$

A função de regressão populacional, *E(y/X)*, é uma função linear de *X*: o valor esperado de *y* varia de acordo com a magnitude de β_1.
Para qualquer valor de *X*, a distribuição de *y* está centrada em *E(y/X)*

Se E(u/X)=0 vale: $\beta_0 + \beta_1 X$ é a parte sistemática (explicada) de *y*;
u é a parte não sistemática (não explicada) de *y*.

$$y = \beta_0 + \beta_1 X + \mu \qquad \mu \sim N(0, \sigma)$$

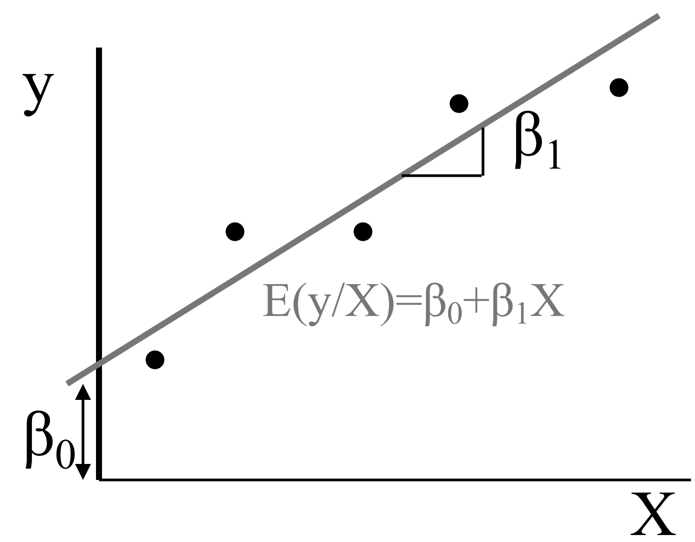
Derivação dos Mínimos Quadrados Ordinários

De $E(u/X) = E(u) = 0$, sabe-se que E(u)=0 $\Rightarrow E(y - \beta_0 - \beta_1 X) = 0$

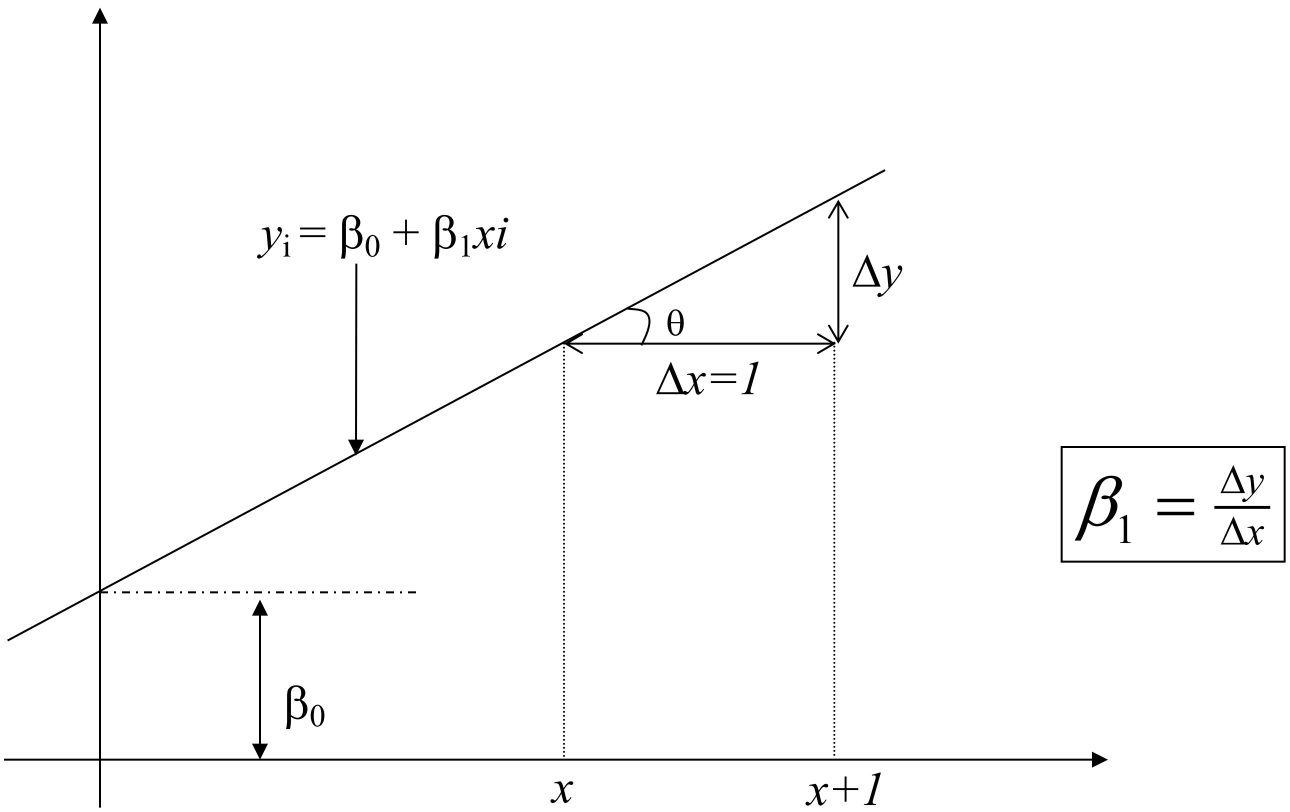
e Cov(X,u)=0 $\Rightarrow E[X(y - \beta_0 - \beta_1 X)] = 0$

$$\underbrace{Y = \beta_0 + \beta_1 X}_{\text{Componente determinística}} + \underbrace{\mu}_{\text{Componente aleatória}}$$

Y = variável dependente
X = variável independente
μ = erro aleatório
β_1 = declive da linha
β_0 = ponto que intercepta o eixo y



Significado dos parâmetros do modelo de regressão linear simples



β_0 (intercepto); quando a região experimental inclui *X=0,* β_0 é o valor da média da distribuição de Y em X=0, cc, não tem significado <u>prático</u> como um termo separado (isolado) no modelo; β_1 (inclinação) expressa a *taxa de mudança* em *Y,* isto é, é a mudança em *Y* quando ocorre a mudança de uma unidade em *X.* Ele indica a mudança na média da distribuição de probabilidade de *Y* por unidade de acréscimo em *X.*

Derivação dos Mínimos Quadrados Ordinários

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

Reescrevendo $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, onde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ e $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

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

.

Minimizando a soma dos resíduos quadrados:

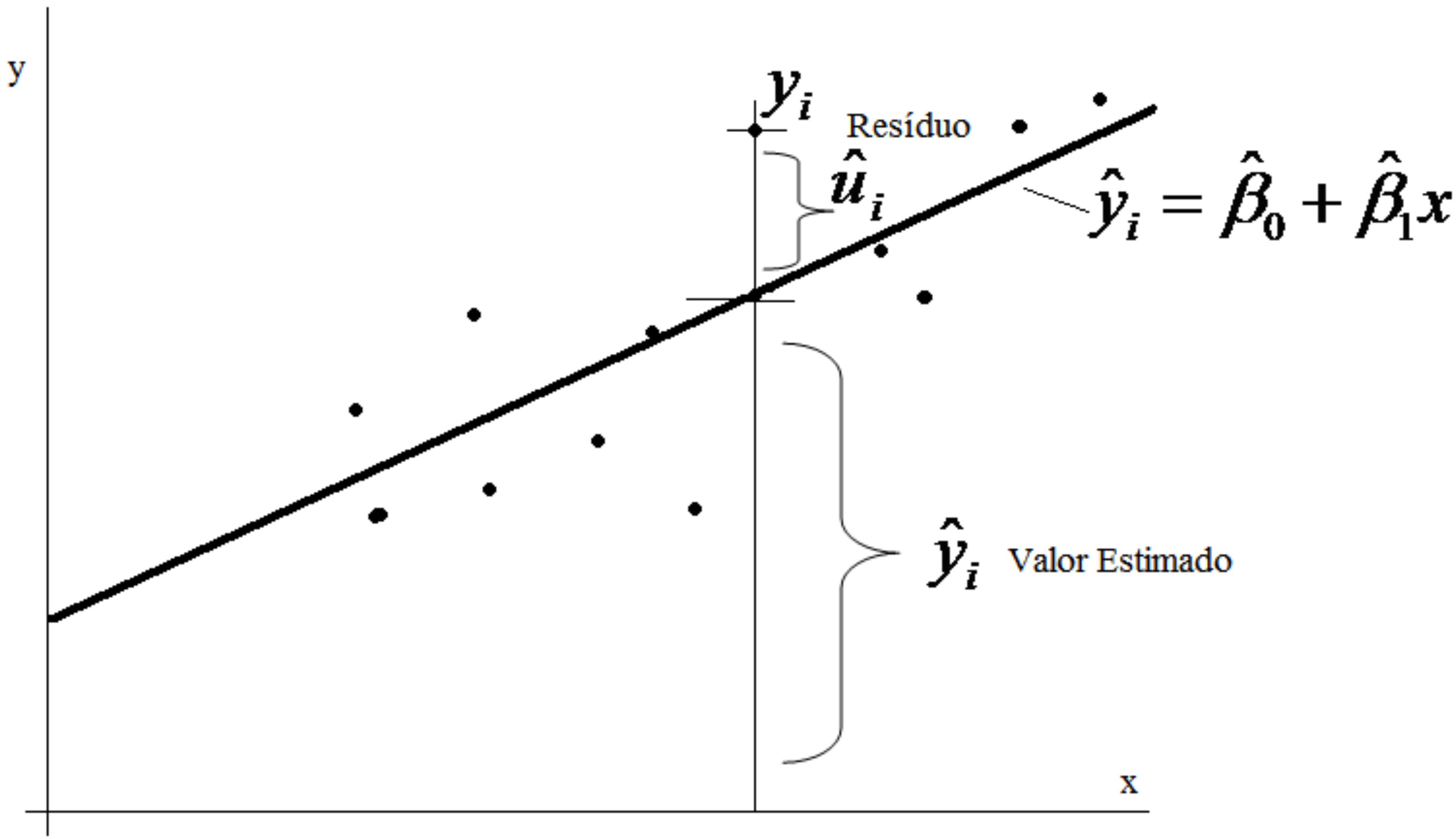
$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \hat{u}_i^2$$

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Resulta em $\boxed{\hat{\beta}_1 = \frac{\sum_{i=i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)}$

Derivação dos Mínimos Quadrados Ordinários

Importante: x deve ter variabilidade



Resultado

Estimamos o modelo populacional:

$$y = \beta_0 + \beta_1 X + \mu \qquad \mu \sim N(0, \sigma)$$

Usando uma amostra, por:

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

O resíduo μ que é não observado $\mu = y - \hat{y}$

pode ser estimado, então, por resíduo

como: $\hat{\mu} = y - \hat{y}$

Então, os MQO minimizam o erro ao quadrado

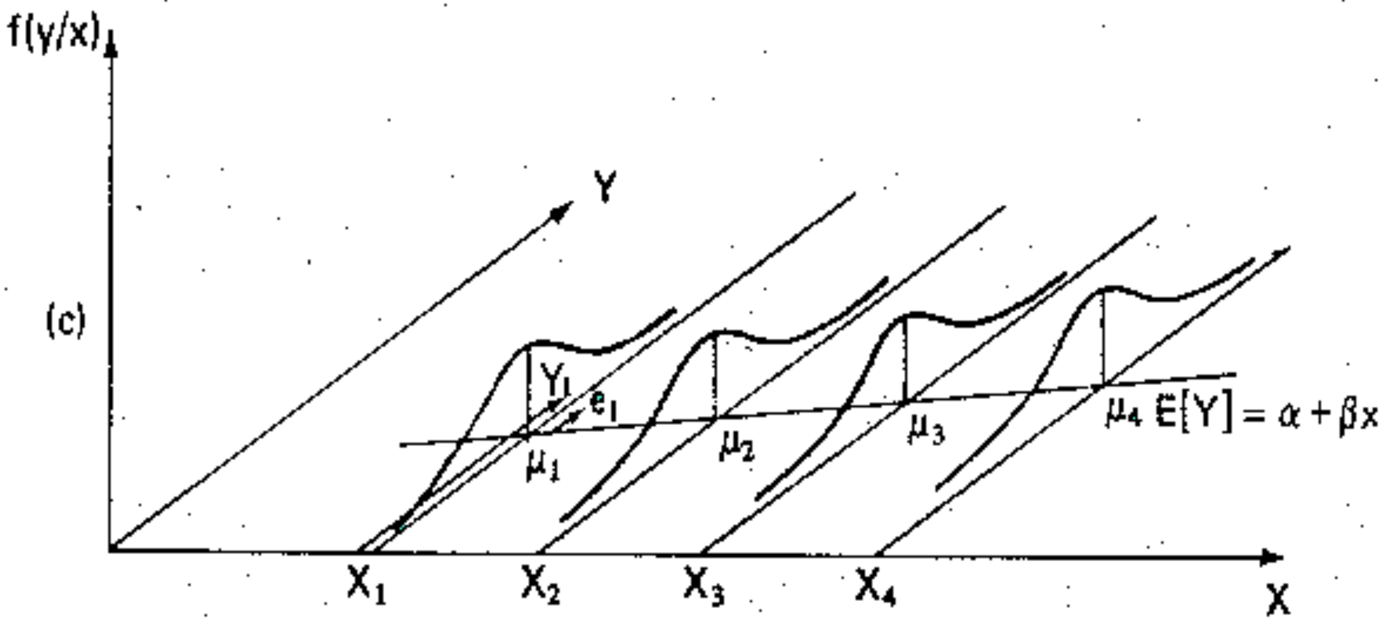
$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

A reta de Regressão de MQO

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (amostral, pode apresentar valores diferentes conforme a amostra)

É a função de regressão amostral (FRA) estimada da função de regressão populacional

$E(y/x) = \beta_0 + \beta_1 x$ (populacional, fixo e desconhecido)



(c) Médias alinhadas: distribuições normais e homocedásticas.

A figura mostra a distribuição de *Y* para vários valores de *X*. Mostra onde cai a observação Y_1. Mostra que o erro é a diferença entre Y_1 e $E(Y_1)$. Observe que as distribuições de probabilidade apresentam a mesma variabilidade.

Exemplo: Salários de Diretores Executivos e Retornos de Ações

Modelo postulado a ser estimado:

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

y → salário anual em mil dólares: salário

x → retorno médio da ação sobre o patrimônio: rma

n = 209 observações

Modelo de Regressão de MQO:

$$\widehat{salário} = 963{,}191 + 18{,}501.rma$$

$\hat{\beta}_0$ =963,191 $\hat{\beta}_1$ = 18,501

Se a amostra mudar, pode-se obter outra reta

Exemplo: Resultados Eleitorais e Gastos em Campanha

Modelo postulado a ser estimado:

$$votoA = \beta_0 + \beta_1 partA + u$$

y → percentagem de votos recebidos pelo candidato A: *votoA*

x → percentagem dos gastos totais de campanha que cabem ao candidato A: *partA*

n = 179 observações

Modelo de Regressão de MQO:

$$vo\hat{t}oA = 26{,}81 + 0{,}464.partA$$

$\hat{\beta}_0$ =26,81 $\hat{\beta}_1$ = 0,464

15 primeiras observações de Salários de Diretores Executivos e Retornos de Ações e valores estimados

n obs.	y=salário	x=rma	y chapéu	u chapéu
1	1095	14.1	1224.058	-129.058
2	1001	10.9	1164.854	-163.854
3	1122	23.5	1397.969	-275.969
4	578	5.9	1072.348	-494.348
5	1368	13.8	1218.508	149.492
6	1145	20	1333.215	-188.215
7	1078	16.4	1266.611	-188.611
8	1094	16.3	1264.761	-170.761
9	1237	10.5	1157.454	79.546
10	833	26.3	1449.773	-616.773
11	567	25.9	1442.372	-875.372
12	933	26.8	1459.023	-526.023
13	1339	14.8	1237.009	101.991
14	937	22.3	1375.768	-438.768
15	2011	56.3	2004.808	6.192

Propriedades Algébricas de MQO

(1) A soma, e a média amostral, dos resíduos de MQO é zero

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

(2) A covariância amostral entre os regressores e os resíduos de MQO é zero

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

Outra interpretação de MOQ

$y_i = \hat{y}_i + \hat{u}_i$, a média amostral dos valores estimados, $\hat{y}_i$, é a mesma média amostral de y_i, ou seja $\bar{y}_i$.

SQT=SQE+SQR

Soma dos quadrados total= soma dos quadrados Explicada+Soma dos quadrados dos resíduos

$$\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Grau de Ajuste: Mede o quanto a variável explicativa (x) mede a variável explicada (y)

R-quadrado da regressão : R^2 é a percentagem da variação amostral em y que é explicada por x

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

Então $0 \leq R^2 \leq 1$

Se $R^2 = 1$, ajuste perfeito dos dados

Se $R^2 \cong 0$, ajuste ruim da reta de MQO

Exemplo1: Salários de Diretores Executivos e Retornos de Ações

$$\hat{salário} = 963{,}191 + 18{,}501.rma$$

n=209 , R^2=0,0132

Vemos que a variação nos salários é pouco explicada pelo retorno das ações. O retorno das ações explica somente 1,3% das variações de salário desta amostra de 209 diretores executivos. Ou seja, 98,7% das variações salariais desses diretores são deixadas sem explicação.

Valores baixos são freqüentes nas ciências sociais

Exemplo2: resultados Eleitorais e Gastos de Campanha

$$\hat{votoA} = 26{,}81 + 0{,}464.partA$$

n=179 , R^2= 0,856

A participação nos gastos do candidato explica 85,6% dos votos eleitorais nesta pesquisa

Unidades de Medida e Forma Funcional

Efeitos de mudança de medida da variável

Salário anual medido em mil dólares

$$\hat{salário} = 963{,}191 + 18{,}501.rma$$

Salário anual medido em dólares

$$\hat{salário} = 963191{,}00 + 18501.rma$$

Retorno da ação medido (percentual) dividido por 100

$$\hat{salário} = 963{,}191 + 1850{,}1.rma$$

Unidades de Medida e Forma Funcional

Incorporação de Não-Linearidade na Regressão

Muitas vezes é preciso considerar modelos não lineares

Exemplo: efeito da escolaridade sobre salário

ganhos crescentes de salários a cada ano adicional de estudo

n = 526, R^2=0,186

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

$$sal\hat{a}rioh = -0,9 + 0,54 educ + u$$

Para cada ano adicional de escolaridade é previsto US$0,54 a mais no salário hora (o que vale tanto para o 1º ano de educação quanto para o 20º)

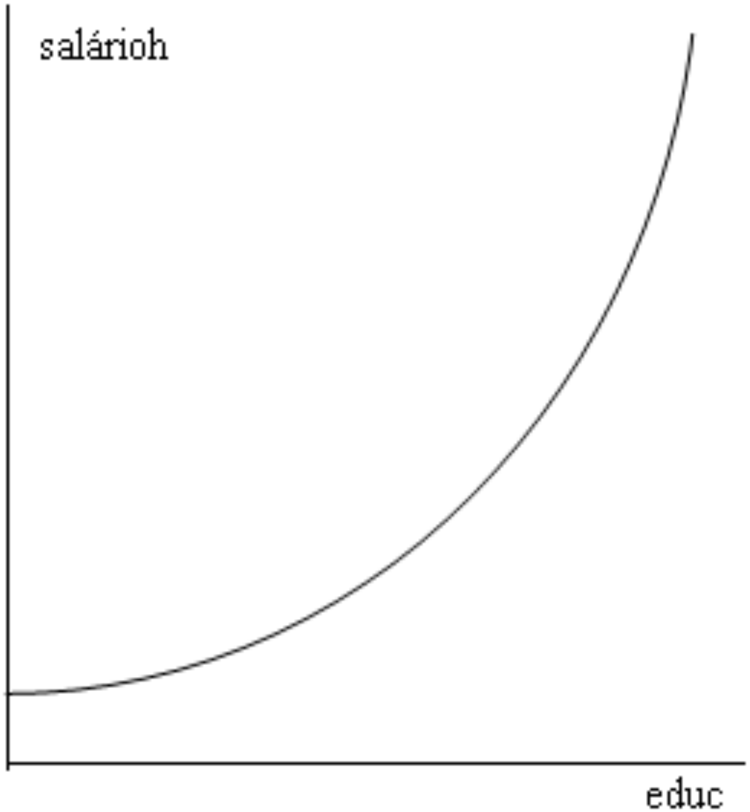
$$\log(salárioh) = \beta_0 + \beta_1 educ + u$$

$$\%\Delta salárioh \approx (100 \cdot \beta_1)\Delta educ$$

$$\log\hat{(}salárioh) = 0,584 + 0,083 \cdot educ$$

O aumento percentual depende do valor inicial

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



Note que

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

pode ser reescrito como:

$y=\exp(\beta 0+\beta 1x+u)$.

O método de estimação não muda. O que muda é a maneira como se usa a variável

Significado de “Linear”

O modelo de regressão é linear nos parâmetros β_0 e β_1. Pois é possível fazer transformações nas variáveis – logX, X^2,$X^{1/2}$, etc.- e usar o método de MQO.

Interpretação do coeficiente associado à variável explicativa X

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)	$\% \Delta y = \beta_1 . \% \Delta x$

Valores Esperados e Variâncias dos Estimadores de MQO

Hipótese 1 (linearidade nos parâmetros): No modelo populacional, a variável dependente y está relacionada à variável independente x e ao erro u como:

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

Hipótese 2 (amostragem aleatória)

Pode-se usar uma amostra aleatória de tamanho n, $\{(x_i, y_i): i=1,2,..,.n\}$, proveniente de um modelo populacional.

Hipótese 3 (média condicional zero)

$E(u/x)=0$ ou $E(y/x) = \beta_0 + \beta_1 x$

3 e 2 $\Rightarrow$ tecnicamente tratar o x_i fixo em amostras repetidas.

Hipótese 4 (variação amostral na variável independente)

Na amostra, as variáveis independentes $x_{i,}$ $i=1,2,..,.n$ não são todas iguais a uma mesma constante. Isso exige uma variação em x na população.

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

Vemos que $\hat{\beta}_1$ é uma variável aleatória.

Reescrevendo: $\boxed{\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{SQT_x} = \beta_1 + (1/SQT_x)\sum_{n=1}^{n} d_i \cdot u_i}$ em que $d_i = x_i - \bar{x}$

Condicionada aos valores de x_i, a aleatoriedade em $\hat{\beta}_1$ depende dos erros, u_i.O fato de que os erros sejam, em geral, diferentes de zero é que faz com que $\hat{\beta}_1$ seja diferente de β_1.

Teorema 1: As Hipóteses de 1 até 4 garantem a inexistência de viés em MQO :

$$E(\hat{\beta}_0) = \beta_0 \text{ } \mathbf{e} \text{ } E(\hat{\beta}_1) = \beta_1$$

Preocupações com a Hipótese 3, E(*u/x*)=0

Usar a regressão simples quando *u* contém fatores que afetam *y* e que também estão correlacionados com *x* pode resultar em correlação **espúria** (uma relação entre *x* e *y* que se deve a outros valores que afetam *y* e também são correlacionados com *x*)

Exemplo: efeito da merenda sobre o desempenho escolar

$$mate10 = \beta_0 + \beta_1 prgalm + u$$

y = % de alunos do primeiro ano do ensino médio aprovados num exame de matemática: *mate10*

x = % de estudantes aptos para participar do programa de merenda escolar: *prgalm*

u possui característica da escola e do aluno que afetam o desempenho escolar total.

$$\widehat{mate}10 = 32{,}14 + 0{,}319\, prgalm$$

n=408, R^2=0,171

Devemos aceitar que a maior participação na merenda piora o desempenho escolar? Muito provavelmente não.

Explicação: o termo erro, u, está correlacionado com prgalm (u contém fatores como taxa de pobreza das crianças que freqüentam a escola, qualidade e recursos da escola, que estão correlacionados com y (desempenho) e com x (elegibilidade ao programa de merenda)) [caracterizando um problema de variável relevante omitida].

Variância dos Estimadores de MQO

Além de saber que a distribuição amostral de $\hat{\beta}_1$ está centrada em torno de β_1 ($\hat{\beta}_1$ é não viesado), é preciso saber quão distante, em média, $\hat{\beta}_1$ está de β_1.

MQO oferecem o estimador com menor variância entre os estimadores lineares.

Hipótese 5 de Homoscedasticidade ou de variância constante: $Var(u/x)=\sigma^2$ ou $Var(y/x)= \sigma^2$

Implica que os estimadores de MQO são eficientes

Exemplo de heteroscedasticidade na equação de salários

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

Pessoas com maior nível de educação têm maiores possibilidades de emprego, o salário pode variar bastante. Pessoas com baixa escolaridade em geral trabalham pelo valor do salário mínimo, portanto apresentam menor variância de ganhos.

Variâncias amostrais em MQO sob as hipóteses de 1 a 5

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

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

Estimação da Variância do Erro

As fórmulas anteriores de variância dos estimadores são desconhecidas, a menos que σ^2 seja conhecido. Por isso é necessário estimá-las.

$y_i = \beta_0 + \beta_1 x_i + u_i$ modelo populacional descrito em termos de uma observação extraída aleatoriamente, em que u_i é o erro da observação.

$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \widehat{u}_i$ modelo estimado , $\hat{u}_i$ é o resíduo.

$$\hat{u}_i = y_i - \hat{y}_i = \beta_0 + \beta_1 x_i + u_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Resíduo é a "estimativa" do erro.

Como u_i é não observado e $\hat{u}_i$ é observado (calculado),

Usamos $\hat{u}_i = y_i - \hat{y}_i$ para calcular a variância do erros

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} \hat{u}_i^2$$ que é um estimador não viesado dos erros de MQO.

(perde-se 2 graus de liberdade por causa das restrições $\sum_{i=1}^{n} \hat{u}_i = 0$ e $\sum_{i=1}^{n} x_i \cdot \hat{u}_i = 0$)

Sob as hipóteses 1 até 5 $E(\hat{\sigma}^2) = \sigma_2$