Econometria

• O que é?

Origem (anos 30, XX): Medição dos fenómenos económicos com o objectivo de testar teorias sobre esses mesmos fenómenos com base na sua observação, recorrendo a procedimentos de inferência adequados

Hoje: é uma área científica mais virada para a aplicação de técnicas estatísticas à análise de dados económicos, financeiros , sociais, ..., com o objectivo de estimar relações entre uma determinada **variável dependente** e um conjunto de **variáveis explicativas**

Exemplos:

- $Consumo = f(rendimento\ disponível)$
- $Salário = f(escolaridade, experiência, idade, género)$

Econometria

• *Finalidade da Econometria*

- "**Testar**" validade de teorias
  Analisar se as teorias são confirmadas pela observação

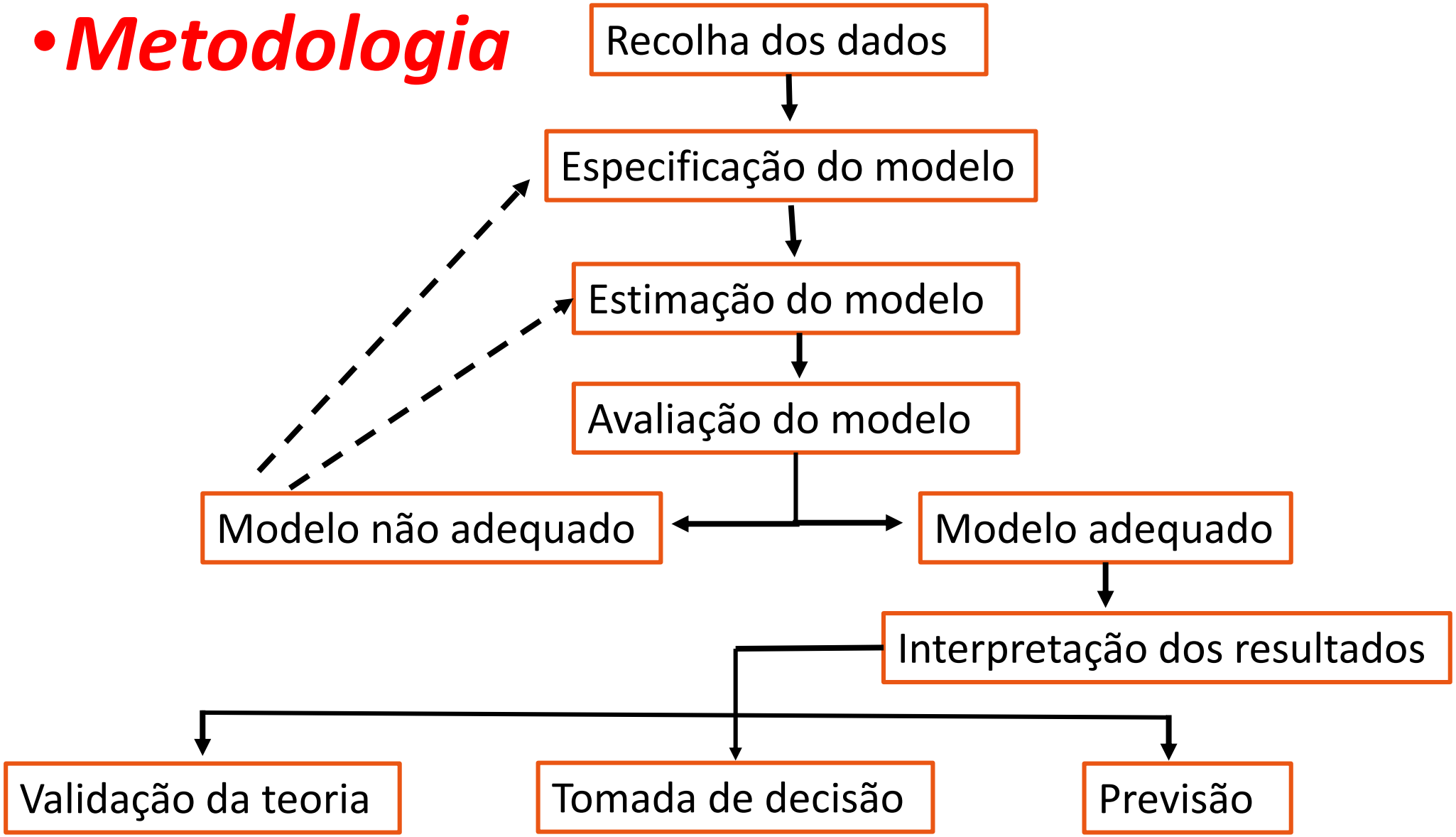
- "**Explicar**" determinada realidade, nomeadamente explicar o valor esperado da variável dependente como função dos valores assumidos pela(s) variável(is) explicativa(s);

- "**Prever**" o comportamento da variável dependente, conhecido(s) o(s) valor(es) assumido(s) pela(s) variável(is) explicativa(s).

- "**Fundamentar**" quantitativamente políticas
  Ex: saber se a implementação de medidas de política no sentido do aumento do nível escolaridade tem impacto na produtividade

Econometria

- ***Metodologia***



Econometria Tipos de dados:

Seccionais (cross section) (Estatística II)

. N entidades (indivíduos, empresas; famílias, ...)
. 1 observação por entidade
. Mesmo período temporal

→ Amostra corresponde a observações independentes

Temporais (time series) (Econometria)

. 1 entidade (empresas; país, ...)
. T observações por entidade
. Vários períodos no tempo

→ Amostra corresponde a observações ordenadas, dependentes

Dados Painel (panel data)

. N entidades
. T observações por entidade

Outros tipos de dados (pooled data)

Econometria Dados Seccionais:

Tempo está “fixo” ⟶ Todas as observações se referem ao mesmo período temporal

Pode considerar-se que a amostra corrresponde a **observações independentes** com **idêntica variância**

Mas ... Nem sempre a situação é tão “simples”. Podem existir vários problemas quer ao nível:

- do modelo (Variância não ser idêntica para todas as observações

- dos dados (Variáveis não observáveis ou observáveis com erro)

Econometria

Causalidade e significância estatística:

. O modelo procura identificar os determinantes do comportamento de uma variável de interesse (dependente): salário, preço de um imóvel

. Na avaliação do modelo:

- verifica-se se as variáveis explicativas são **estatísticamente significativas** para explicar o comportamento da variável de interesse, na procura de uma **possível relação de causalidade**

Causalidade e significância estatística – A causalidade vai para além da significância estatística pois envolve considerações teóricas

Econometria

Análise *ceteris paribus*:

. Num modelo envolvendo várias variáveis explicativas, procura-se, em geral, focarmo-nos sobre o efeito de uma delas sobre a variável de interesse (dependente).

$$Salário = f(escolaridade, experiência, idade, género)$$

. Analisa-se o efeito parcial/marginal de cada variável explicativa sobre a variável de interesse, sob a condição ***ceteris paribus***, isto é, assumindo que tudo o resto permanece constante.

Pode interessar: o sinal do efeito (positivo (relação directa)
\negativo (relação inversa)

Intensidade\magnitude da relação

Econometria Especificação do modelo:

Enquadramento do modelo

. Modelo com **dados seccionais**.

$$y = f(x_1, x_2, , \cdots, x_k)$$

. Os dados tem natureza não experimental (estão fora do controle de quem vai especificar o modelo) $\Rightarrow$ y e $x_1, x_2, , \cdots, x_k$ são consideradas v.a.(s).

. **Muito importante**: Para aliviar a notação, vai abandonar-se a convenção em torno das maiúsculas e minúsculas na distinção de v.a.(s) e suas realizações empíricas. As letras maiúsculas serão agora usadas para representar vectores e matrizes

Econometria Especificação do modelo:

Objectivo

. Modelo $\longrightarrow$ $y = f(x_1, x_2, , \cdots, x_k)$

$$Salário = f(escolaridade, experiência, idade, género)$$

. Vai modelar-se o **valor esperado** de y **condicionado** por $x_1, x_2, , \cdots, x_k$.

Relação é estatística e não matemática

. Um modelo é sempre uma representação simplificada da realidade que permite sublinhar os factores mais relevantes na explicação do comportamento da variável y

Existem sempre var.(s) explicativas não incluídas no modelo.

Modelo de Regressão linear

Manuel está a pensar ingressar no Ensino Superior como meio de melhorar as suas perspectivas de emprego e remuneração

Um amigo do Manuel, o João aconselha-o a desistir da ideia dando-lhe dois exemplos:

O amigo Frederico que tem doutoramento e está desempregado

O amigo David que não fez sequer o 12º e está rico

Questões subjacentes a esta discussão:

Será que mais escolaridade está associada a melhor salário?

Quanto aumenta o salário por cada ano adicional de escolaridade?

Modelo de Regressão Linear

A amostra recolhida produziu 100 pontos:

Inês 12 anos de escolaridade, 650 €

Pedro
17 anos de escolaridade, 1280 €

Joaquim

9 anos de escolaridade, 680 €

Teresa

11 anos de escolaridade, 700 €

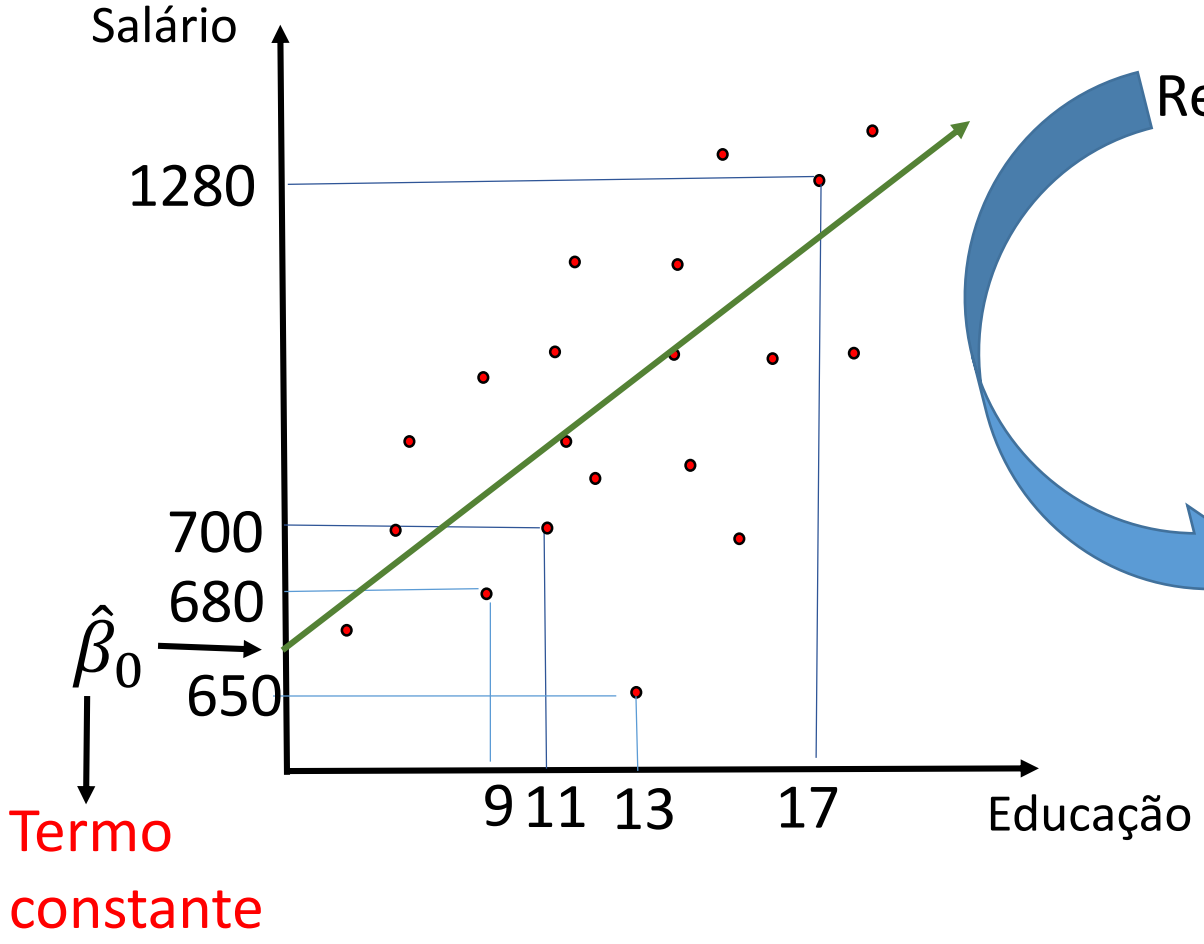
...

Modelo de Regressão Linear

Mas será possível saber mais sobre esta relação?

Sim, ajustando uma recta a esta nuvem de pontos

Qual o acréscimo de salário por cada ano de escolaridade adicional?



Recta $y = mx + b$ → Intersecção com eixo y

Inclinação da recta

Variável explicativa (v.a.)

Recta estimada

$E(salário|u) = \hat{\beta}_0 + \hat{\beta}_1$ Educação

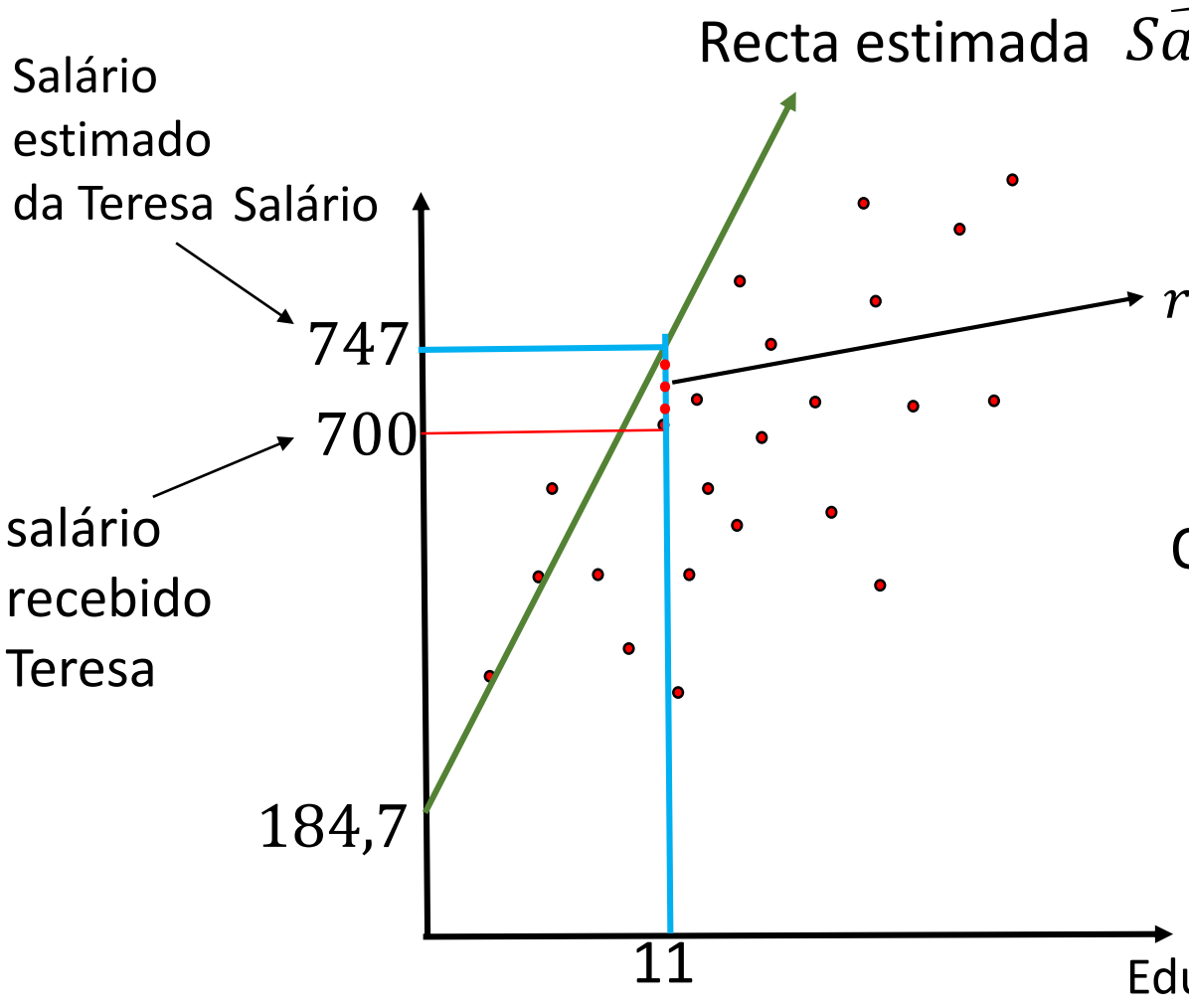
Variável explicada

Variável residual

Coeficientes Estimados

Modelo de Regressão Linear

Recta estimada $\widehat{Salário} = 184{,}7 + 51{,}12$ Educação



$$resíduo = \text{salário} - \widehat{salário}$$

$$= 700 - 747 = -47$$

Qual a explicação para o resíduo?

R: Modelo não se ajusta perfeitamente à realidade porque há várias outras variáveis que afectam o salário

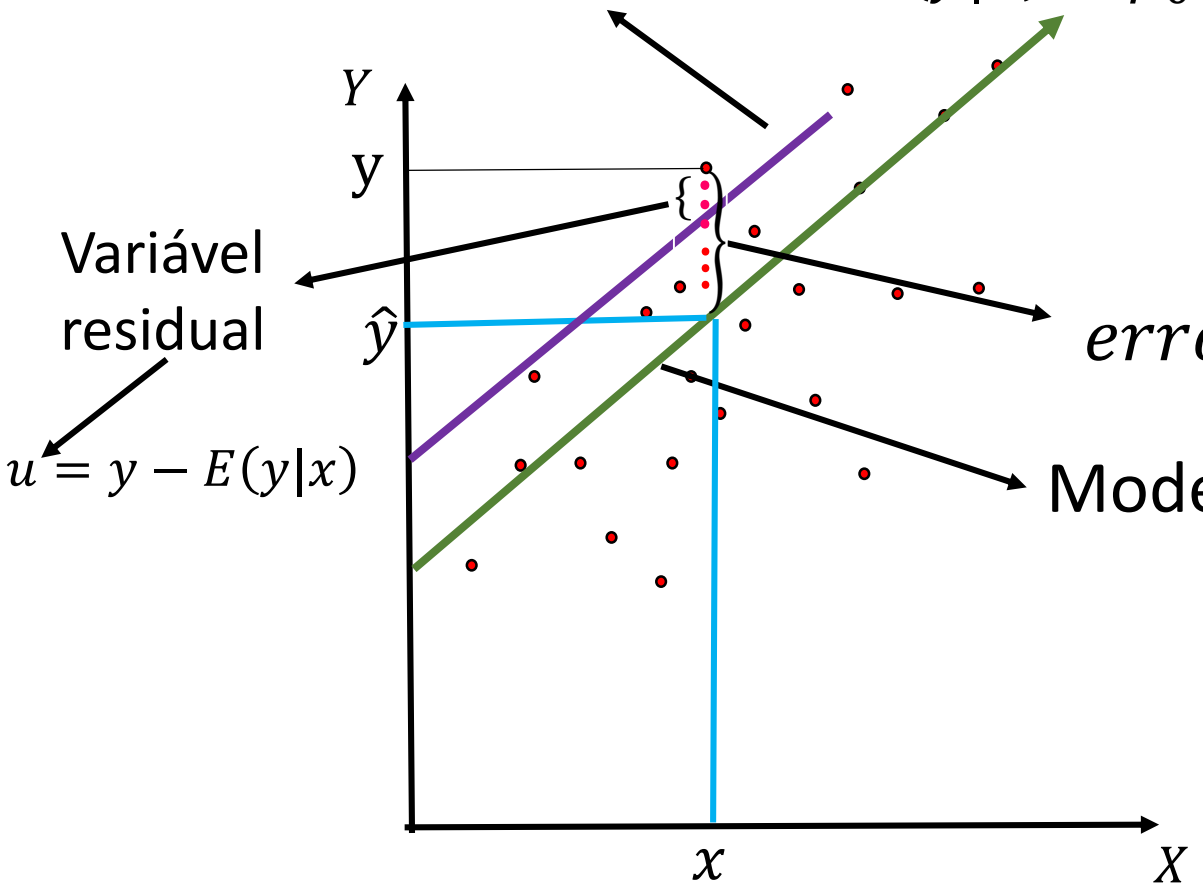
Modelo de Regressão Linear

Modelo da população $y_i = \beta_0 + \beta_1 x_i + u_i$ → Variável residual (não observável)

Representa o efeito de outras variáveis não consideradas no modelo

Exemplos: experiência, género, sector, …

Modelo Regressão **Linear** $E(y|x) = \beta_0 + \beta_1 x$



$erro \backslash \text{resíduo} = \hat{u} = \text{y} - \hat{y}$

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

$\widehat{E(y|x)}$

Modelo de Regressão Linear

Terminologia do Modelo de Regressão Linear:

y	x
Variável dependente	Variável independente
Variável explicada	Variável explicativa
Regressando	Regressor
β_0, β_1	
Coeficientes ou parâmetros (fixos e desconh.)	
u	
Variável residual (não observada)	

Modelo de Regressão Linear

Modelo da população: $y_i = \beta_0 + \beta_1 x_i + u_i \ (1)$

↓

Modelo Regressão Linear: $E(y|x) = \beta_0 + \beta_1 x \ \ (2)$

Relação entre var. dependente e explicativa é estatística e não matemática

Hipóteses do modelo:

$H_1: E(u) = 0 \qquad (3)$

$H_2: E(u|x) = E(u) \qquad (4)$

$H_3: Cov(x,u) = 0 \ \Rightarrow$ u e x não são correlacionadas

(3) O valor médio do efeito das variáveis não observadas (u) é nulo.

(4) O valor médio do efeito das variáveis não observadas (u) não depende do valor da var. explicativa (x) diz-se que a var. explicativa é exógena.

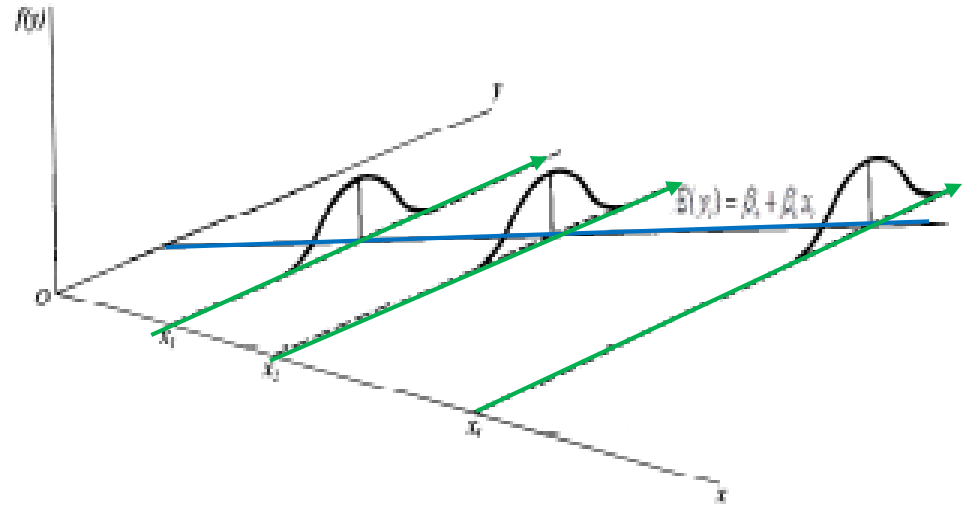
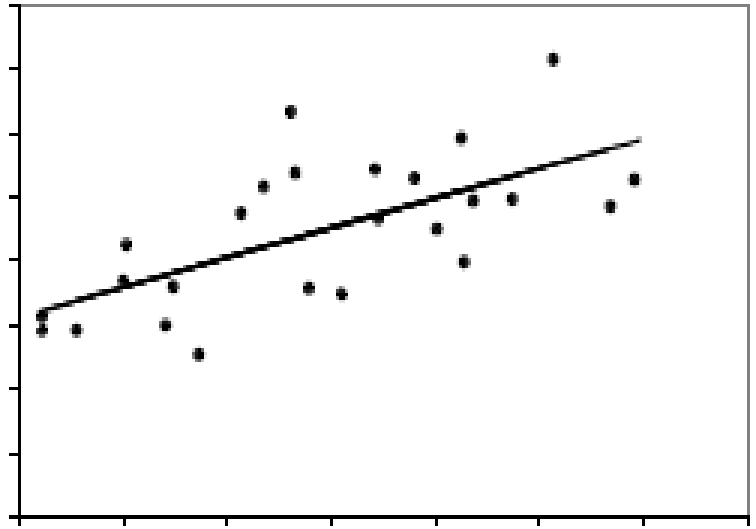
Modelo de Regressão Linear

Modelo da população ⟶ $\text{Salário}_i = \beta_0 + \beta_1 \text{Educação}_i + u_i$

Variável residual

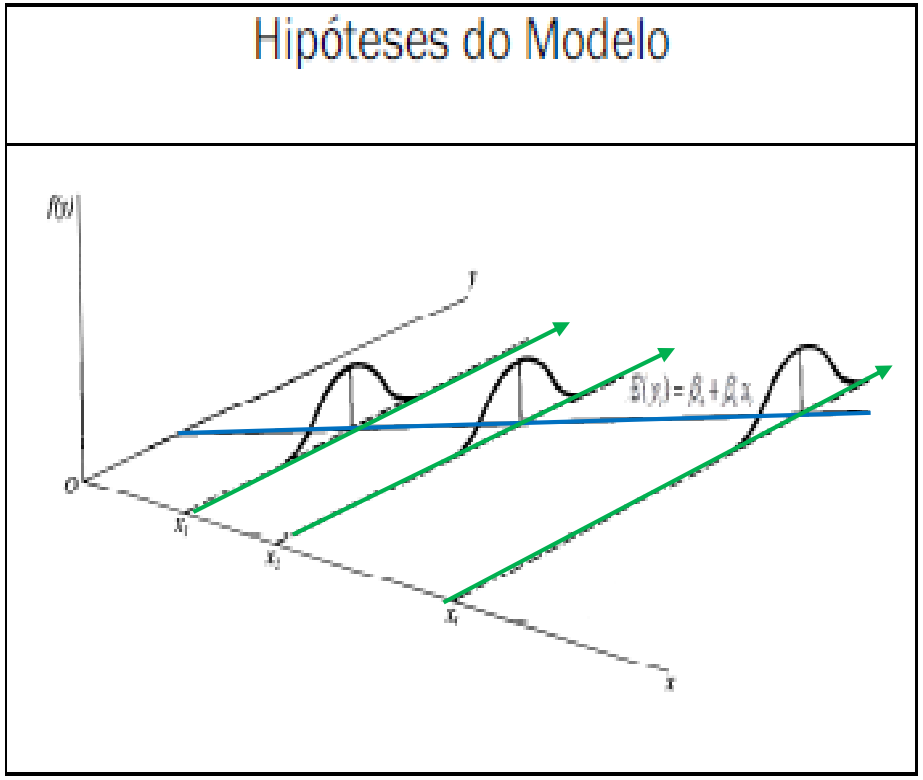
Flexibilidade

Modelo Regressão Linear ⟶ $E(salário|X) = \beta_0 + \beta_1 \text{Educação}$

Hipóteses do Modelo	Nuvem de pontos e recta ajustada
	

Muito importante: Esta equação diz-nos como é que o valor médio do salário varia com a educação. Não diz que o salário é igual a $\beta_0 + \beta_1 \text{Educação}$.

Modelo de Regressão Linear



$Var(u_i)$ é igual para todas as $i = 1, 2, \cdots, n$ observacões.

$$Var(u_i) = \sigma^2 \; i = 1, 2, \cdots, n$$

Homocedasticidade: variabilidade do salário à volta da média é constante .

Modelo de Regressão Linear

Apenas se vão estudar modelos que envolvem uma **relação linear** ou **linearizável** em relação aos parâmetros, porque:

- abrangem uma variedade significativa de situações

- são de tratamento mais fácil

Muito importante: Não confundir **linearidade relativa aos parâmetros** com **linearidade relativa às variáveis**

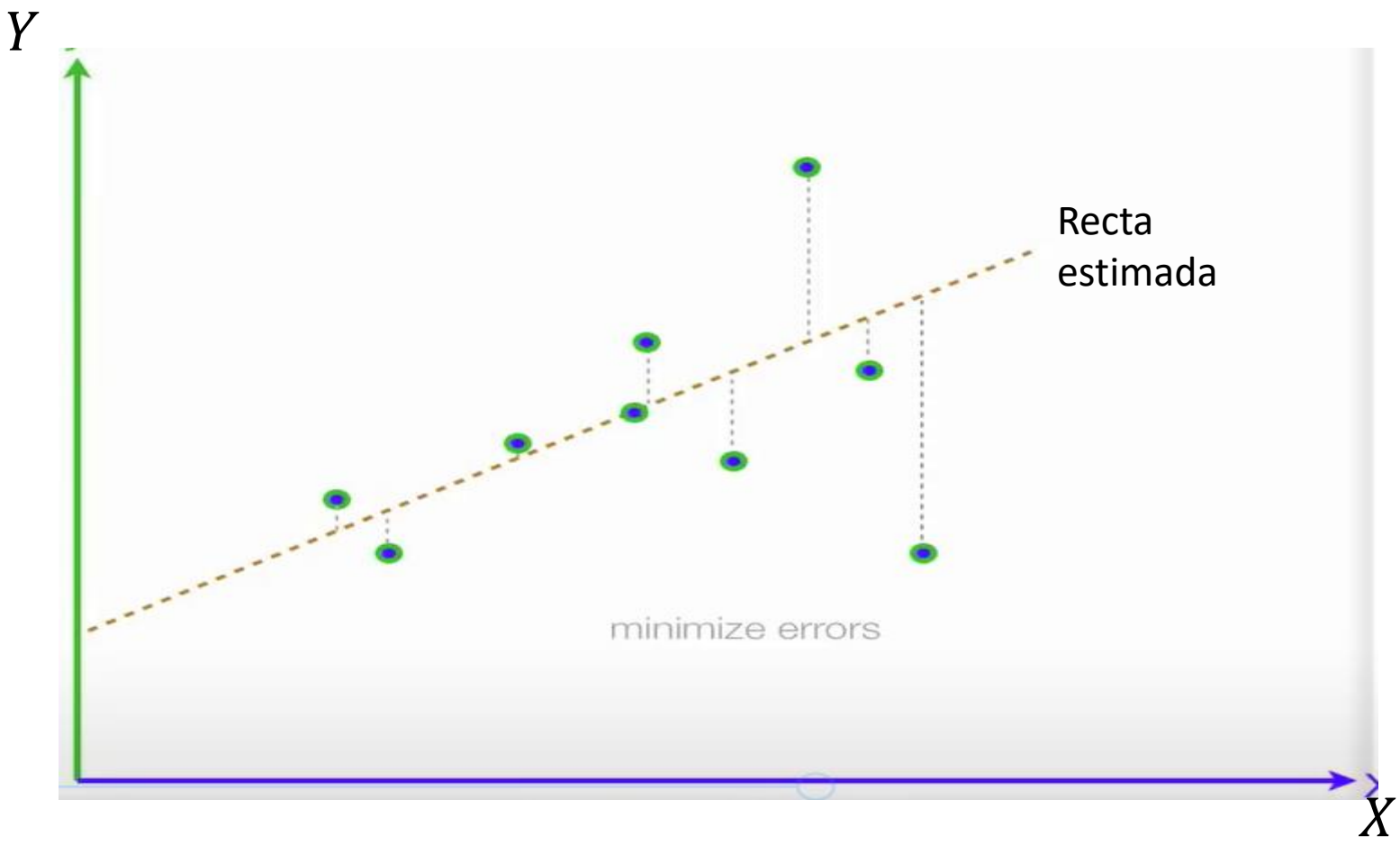
$Y = \beta_0 + \beta_1 X \longrightarrow$ É linear nos parâmetros

$Y = \beta_0 + \beta_1 X^2 \longrightarrow$ Não é linear mas é linearizável nos parâmetros

$Y = \beta_0 + {\beta_1}^2 X \longrightarrow$ Não é linear nos parâmetros

Modelo de Regressão linear

Estimação dos coeficientes de regressão pelo Método dos Mínimos Quadrados



IDEIA

Ajustamento da recta à nuvem de pontos será tanto melhor quanto menor for a distância dos pontos à recta

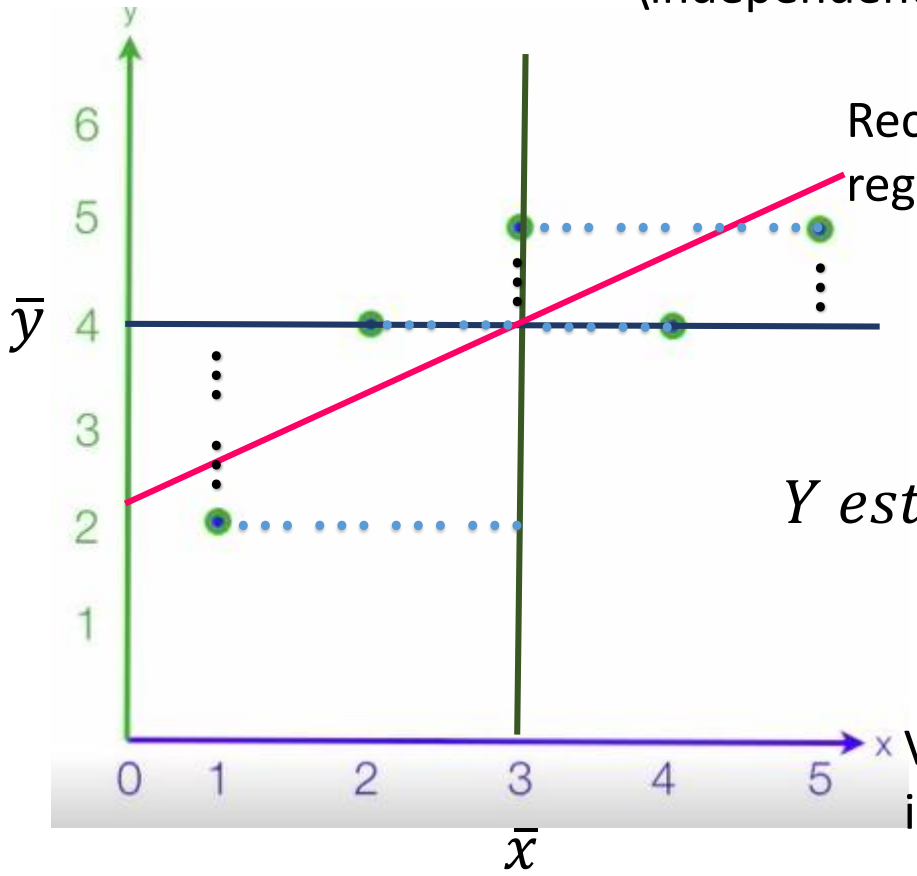
Minimizar a soma dos quadrados dos erros

Modelo de Regressão linear

Estimação dos coeficientes de regressão pelo Método dos Mínimos Quadrados

Variável dependente

Variável explicativa \independente



Variável dependente

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
				10	6

Coef. estimados

$Y\ estimado$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \longrightarrow$$ Recta de regressão estimada

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{6}{10}$$

$$E(B_1) = \beta_1\ (estimador\ não\ enviesado)$$

Qualquer recta de regressão linear passa pelo ponto $(\bar{x}, \bar{y}\ ) = (3,4)$

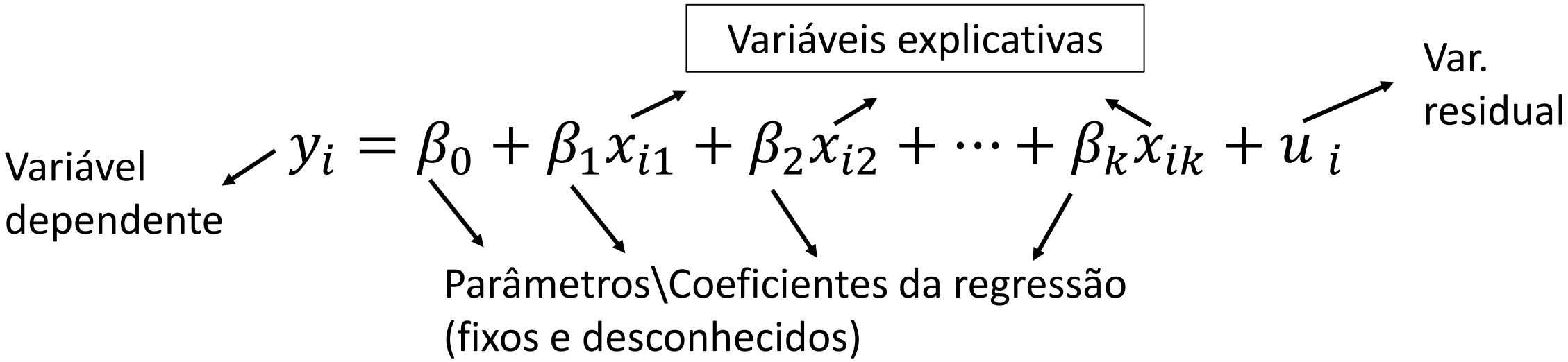
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \Leftrightarrow b_0 = 4 - 0.6 * 3 = 2.2$$

Modelo de Regressão Linear Múltipla

- Modelo vai ser estimado com base numa amostra com $\boldsymbol{N}$ observações

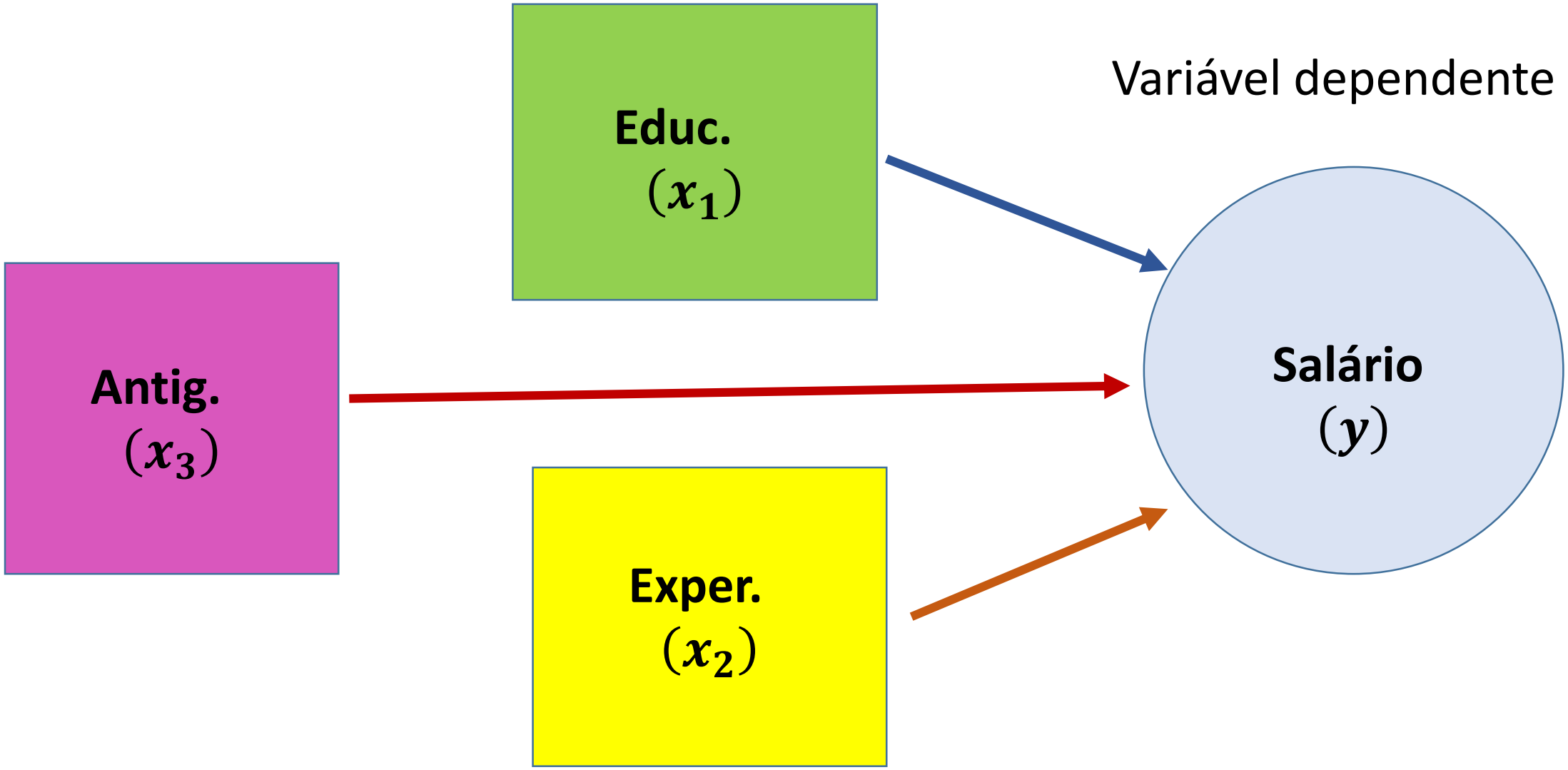
Cada observação é dada por: $(y_i, x_{i1}, x_{i2}, \cdots, x_{ik})$

Nota: $(n > k)$ **muito maior**. (n deve ser, no mínimo, 5 a 10 vezes maior que k)



Exemplo: $salário_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_k antig_i + u_i$

Modelo de Regressão Linear Múltipla



Modelo de Regressão Linear Múltipla (MRLM)

- Objectivo genérico: explicar $E(y|x)$

  Regressão linear: $E(y|x)$ é função linear de um conjunto de parâmetros

$$E(y|x) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

  Fazendo $\mathrm{u} = \mathrm{y} - E(y|x)$ pode escrever-se:

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

Determinam-se os valores $\widehat{\beta_0}, \widehat{\beta_1}, \cdots, \widehat{\beta_k}$ que minimizam $\sum_{i=1}^{n} \widehat{u_i}^2$

Método dos mínimos quadrados

Modelo de Regressão Linear Múltipla (MRLM)

- **OLS – Método dos mínimos quadrados**

$$\sum_{i=1}^{n} \widehat{u_i}^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

$$= \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \cdots - \widehat{\beta_k} x_{ik}\right)^2$$

As condições de 1ª ordem vêm ($\partial \sum_{i=1}^{n} \widehat{u_i}^2 / \partial \widehat{\beta_i} \quad i = 1,2,\cdots,k$):

$$\begin{cases} \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \cdots - \widehat{\beta_k} x_{ik}\right) = 0 \\ \sum_{i=1}^{n} x_{ij} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \cdots - \widehat{\beta_k} x_{ik}\right) = 0 \quad j = 1,2,\cdots,k \end{cases}$$

Modelo de Regressão Linear Múltipla (MRLM)

- **OLS – Método dos mínimos quadrados**

Em termos matriciais vem:

.Valores ajustados: $\hat{Y} = \boldsymbol{X}\hat{\beta}$ .Resíduos: $\hat{U} = Y - \hat{Y}$

Método dos mínimos quadrados: $\min \hat{U}^T \hat{U}$

As condições de 1ª ordem :

$$\boldsymbol{X}^T\left(Y - \boldsymbol{X}\hat{\beta}\right) = 0 \Leftrightarrow \boldsymbol{X}^T Y = \boldsymbol{X}^T \boldsymbol{X}\hat{\beta}$$

.Estimador $\boldsymbol{MQ}$: $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$

Modelo de Regressão Linear Múltipla (MRLM)

- **OLS – Método dos mínimos quadrados**

.Estimador $\boldsymbol{MQ}$: $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$

$$X^T X = \begin{bmatrix} N & \sum_{i=1}^{N} x_{i1} & \cdots & \sum_{i=1}^{N} x_{ik} \\ \sum_{i=1}^{N} x_{i1} & \sum_{i=1}^{N} x_{i1}^2 & \cdots & \sum_{i=1}^{N} x_{ik}x_{ik} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{N} x_{ik} & \sum_{i=1}^{N} x_{ik}x_{i1} & \cdots & \sum_{i=1}^{N} x_{ik}^2 \end{bmatrix} \qquad X^T Y = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{N} x_{ik}y_i \end{bmatrix}$$

Modelo de Regressão Linear Múltipla (MRLM)

Já que:

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{1k} & x_{2k} & x_{3k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} N & \sum_{i=1}^{N} x_{i1} & \sum_{i=1}^{N} x_{i2} & \cdots & \sum_{i=1}^{N} x_{ik} \\ \sum_{i=1}^{N} x_{i1} & \sum_{i=1}^{N} x_{i1}^2 & \sum_{i=1}^{N} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{N} x_{i1}x_{ik} \\ \sum_{i=1}^{N} x_{i2} & \sum_{i=1}^{N} x_{i2}x_{i1} & \sum_{i=1}^{N} x_{i2}^2 & \cdots & \sum_{i=1}^{N} x_{i2}x_{ik} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{N} x_{ik} & \sum_{i=1}^{N} x_{ik}x_{i1} & \sum_{i=1}^{N} x_{ik}x_{i3} & \cdots & \sum_{i=1}^{N} x_{ik}^2 \end{bmatrix}$$

Modelo de Regressão Linear Múltipla (MRLM)

- **OLS – Método dos mínimos quadrados**

e que:

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \cdots & x_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{1k} & x_{2k} & x_{3k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_{i1}\, y_i \\ \sum_{i=1}^{N} x_{i2}\, y_i \\ \\ \sum_{i=1}^{N} x_{ik}\, y_i \end{bmatrix}$$

Modelo de Regressão Linear Múltipla (MRLM)

- **Estimação OLS – exemplo:**

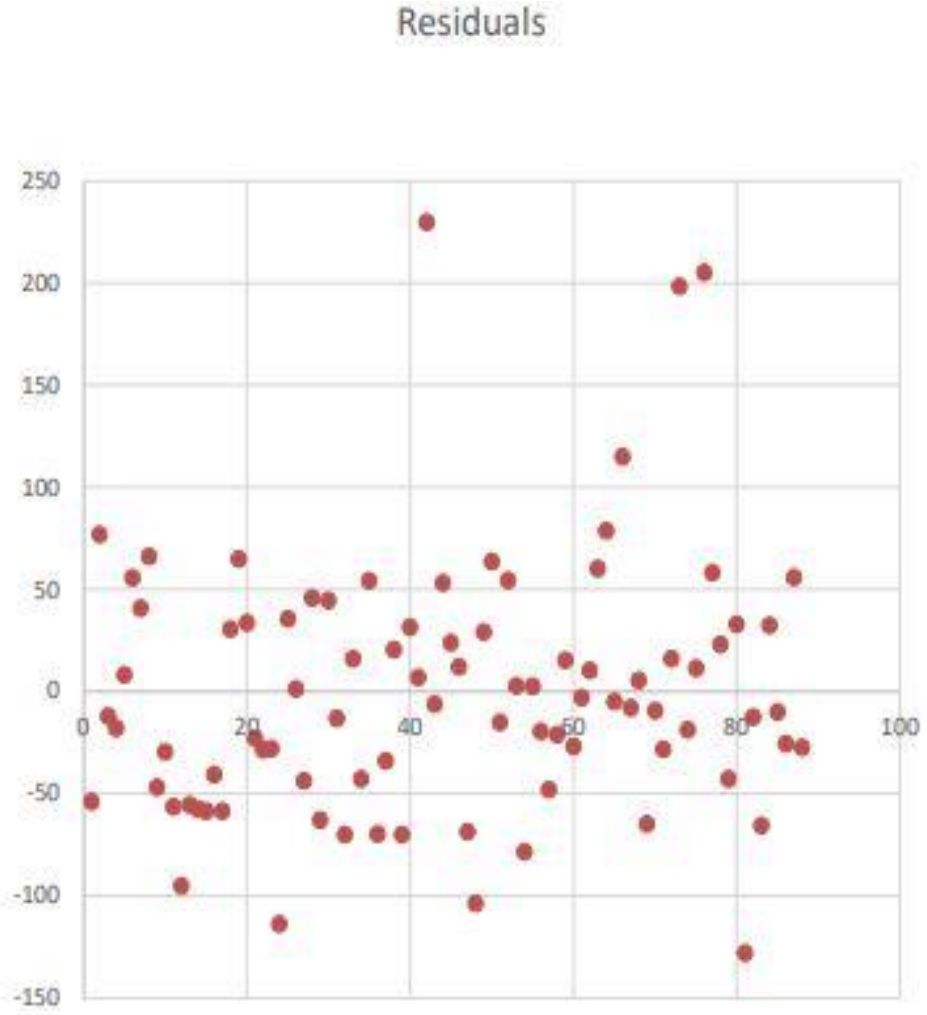
$$\widehat{preço} = -19.286 + 1.384 area + 15.121 quartos$$

Output EXCEL

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.794906945					
R Square	0.63187705					
Adjusted R Square	0.623215334					
Standard Error	63.04837628					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	579971.1994	289985.5997	72.95055817	3.58672E-19	
Residual	85	337883.3088	3975.097751			
Total	87	917854.5083				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-19.28550028	31.0475285	-0.621160563	0.536156232	-81.0163048	42.44530424
area(m2)	1.38360615	0.148943494	9.289470184	1.40049E-14	1.08746658	1.67974572
quartos	15.12133684	9.488597692	1.593632413	0.114730383	-3.744537442	33.98721112

Modelo de Regressão Linear Múltipla (MRLM)

Estimação OLS – exemplo:



A ordem da observação na amostra (i) está no eixo das abscissas enquanto os resíduos$(\hat{u}_i)$ estão no eixo das ordenadas.

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

Como para a 1ª obs, preço=300, area=226 e quartos=4 vem

$$\widehat{preço}_1 = -19.286 + 1.384 \times 226 + 15.121 \times 4 = 353.894$$

$\hat{u}_1$=300-353.894=-53.894

e, de forma semelhante,

$\hat{u}_2$ = 370 - 293.114 = 76.884

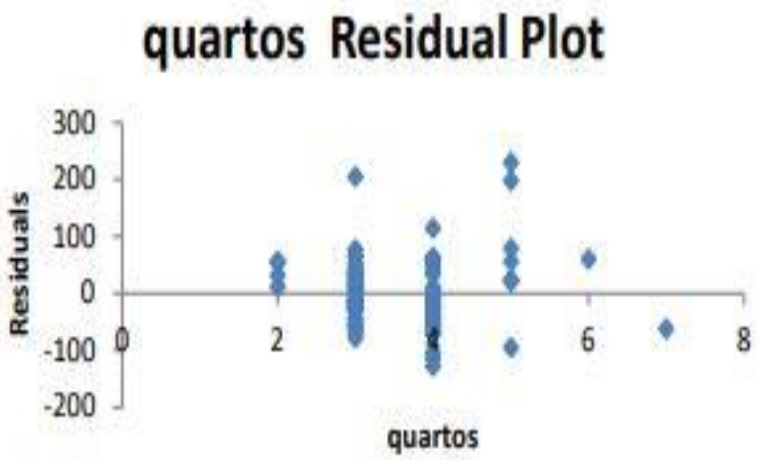
....

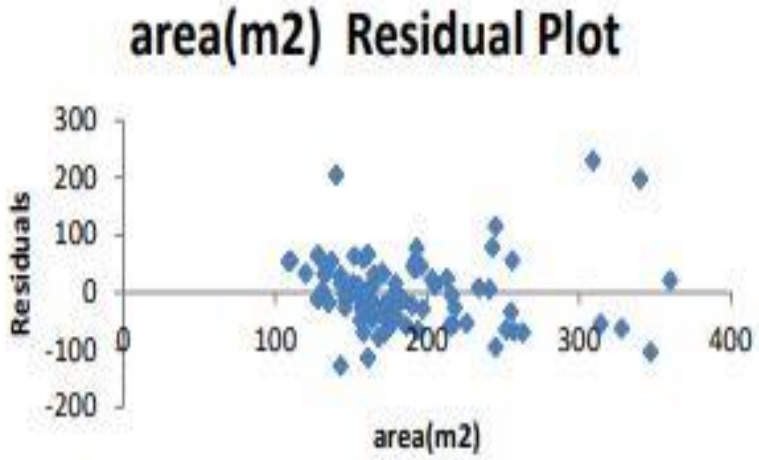
$\hat{u}_{88}$ =242 − 269.495 = −27.495

Modelo de Regressão Linear Múltipla (MRLM)

Estimação OLS – exemplo:

Como, num modelo com dados seccionais, a ordem das observações é arbitrária torna-se frequentemente interessante olhar para os resíduos em função de cada variável explicativa ou do valor da variável endógena





Modelo de Regressão Linear Múltipla

Propriedades dos resíduos OLS:

- $\sum_{i=1}^{n} \widehat{u_i} = 0 \quad (i = 1, 2, \cdots, n)$ (modelos com termo independente)
- $\sum_{i=1}^{n} \widehat{u_i} x_{ij} = 0 \quad (j = 1, 2, \cdots, k)$
- $\sum_{i=1}^{n} \widehat{u_i} \widehat{y_i} = 0$
- $\sum_{i=1}^{n} {y_i}^2 = \sum_{i=1}^{n} {\widehat{y_i}}^2 + \sum_{i=1}^{n} {\widehat{u_i}}^2$

Modelo de Regressão Linear Múltipla

Propriedades dos resíduos OLS:

Para demonstrar estas propriedades é necessário recordar as condições de 1ª ordem:

$$\begin{cases} \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \cdots - \widehat{\beta_k} x_{ik}\right) = 0 \\ \sum_{i=1}^{n} x_{ij} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i1} - \widehat{\beta_2} x_{i2} - \cdots - \widehat{\beta_k} x_{ik}\right) = 0 \quad j = 1,2,\cdots,k \end{cases}$$

- A propriedade 1 deriva da 1ª equação

Nota: Só é verdade se existir β_0

- A propriedade 2 deriva da 2ª equação

Modelo de Regressão Linear Múltipla

Propriedades dos resíduos OLS:

- Propriedade 3: $\sum_{i=1}^{n} \hat{u}_i \hat{y}_i = 0$

$$\sum_{i=1}^{N} \hat{u}_i \hat{y}_i = \sum_{i=1}^{N} \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ \ldots + \hat{\beta}_k x_{ik})$$

$$= \hat{\beta}_0 \sum_{i=1}^{N} \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^{N} \hat{u}_i x_{i1} + \cdots + \hat{\beta}_k \sum_{i=1}^{N} \hat{u}_i x_{ik}$$

$$= 0$$

aplicando

propriedade 1 ($\sum_{i=1}^{N} \hat{u}_i = 0$) e

propriedade 2 ($\sum_{i=1}^{N} \hat{u}_i x_{ij} = 0$ para $j = 1,2, \ldots, k$)

Modelo de Regressão Linear Múltipla

Propriedades dos resíduos OLS:

- Propriedade 4: $\sum_{i=1}^{n} y_i{}^2 = \sum_{i=1}^{n} \hat{y}_i{}^2 + \sum_{i=1}^{n} \hat{u}_i{}^2$

$$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} (\hat{y}_i + \hat{u}_i)^2 = \sum_{i=1}^{N} \hat{y}_i^2 + \sum_{i=1}^{N} \hat{u}_i^2 + 2 \sum_{i=1}^{N} \hat{y}_i \hat{u}_i$$
$$= \sum_{i=1}^{N} \hat{y}_i^2 + \sum_{i=1}^{N} \hat{u}_i^2$$

já que , pela propriedade 3, $\sum_{i=1}^{N} \hat{y}_i \hat{u}_i = 0$

Modelo de Regressão Linear

Interpretação dos parâmetros

Caso 1 Modelo lin-lin:

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

tem-se: $\Delta x_j = 1 \rightarrow \Delta E(y|\boldsymbol{x}) = \beta_j,\ ceteris\ paribus$

A dedução é imediata e pode ser ilustrada para $j = 1$

$$E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$$E(y|\boldsymbol{x}^*) = \beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + \cdots + \beta_k x_k = E(y|\boldsymbol{x}) + \beta_1$$

$$\Delta E(y|\boldsymbol{x}) = E(y|\boldsymbol{x}^*) - E(y|\boldsymbol{x}) = \beta_1,\ ceteris\ paribus$$

Modelo de Regressão Linear

Caso 1 Modelo lin-lin:

Exemplo: Considere o modelo

$$preço = \beta_0 + \beta_1 area + \beta_2 quartos + u$$

onde o **preço de uma casa**, em milhares de dólares, depende da **área** (m^2) e do **número de quartos** (dados disponíveis no AQUILA).

Estimou-se

$$\widehat{preço} = -19.286 + 1.384\, area + 15.121 quartos$$

Admitindo tudo o resto constante:

- por cada m^2 adicional (variação unitária de área), é de esperar que o preço esperado da casa aumente 1384 dólares;
- por cada quarto adicional (variação unitária de quartos) é de esperar que o preço esperado da casa aumente 15121 dólares.

Modelo de Regressão Linear

Interpretação dos parâmetros

Modelos com logaritmos – duas observações :

1. A função $\boldsymbol{ln(z)}$ apenas é definida para $z > 0$. Não pode ser aplicada a variáveis que assumam valores negativos ou nulos

2. A taxa de crescimento de $\boldsymbol{z}$ é dada por $\frac{\Delta z}{z} = \frac{(z+\Delta z)-z}{z}$.

Quando $\frac{\Delta z}{z}$ é pequeno, vem: $\frac{\Delta z}{z} \cong ln(z + \Delta z) - ln(z)$

Exemplo: $\boldsymbol{z = 100}$ *e* Δz=1, $\frac{\Delta z}{z} = \frac{101}{100} = 0.01$(taxa de crescimento de 1%)

$$ln(100 + 1) - ln(z) = 0.00995$$

Modelo de Regressão Linear

Interpretação dos parâmetros

Caso 2 Modelo log-lin:

regressando

A variável dependente é z mas $\boldsymbol{y = ln(z)}$. Suponha-se ainda que a variável explicativa é $\boldsymbol{x}$.

Modelo regressão transformado:

$$E[\boldsymbol{ln(z)}|x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Modelo estimado: $\widehat{\boldsymbol{ln(z)}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

Em termos directos, $\Delta x_j = 1 \rightarrow \Delta \widehat{\boldsymbol{ln(z)}} = \hat{\beta}_j$

Mas, o interesse reside em $\boldsymbol{z}$ e não em $\widehat{\boldsymbol{ln(z)}}$, pelo que aplicando o resultado anterior se tem: $\Delta x_j = 1 \longrightarrow \%\, \Delta\, \hat{\boldsymbol{y}} \cong \boldsymbol{100} * \hat{\beta}_j$, *ceteris paribus*

Modelo de Regressão Linear

Interpretação dos parâmetros

Caso 2 Modelo log-lin:

Exemplificando para $\Delta x_1 = 1 \Leftrightarrow \hat{y}^* = \hat{y} + \Delta\hat{y}$

Modelo inicial: $\widehat{ln(z)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

Modelo pós incremento: $\widehat{ln(z)}^* = \hat{\beta}_0 + \hat{\beta}_1 (x_1 + 1) + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

$$= \widehat{ln(z)} + \hat{\beta}_1$$

$$\widehat{ln(z)}^* - \widehat{ln(z)} = \hat{\beta}_1 \cong \frac{\Delta\hat{z}}{\hat{z}}$$

logo: $\%\Delta\hat{z} = 100 * \hat{\beta}_1\%$

→ Variação percentual de $\hat{y}$

Modelo de Regressão Linear

Interpretação dos parâmetros

Caso 2 Modelo log-lin:

Exemplo: Retomem-se os dados anteriores e considere-se agora o modelo alternativo que originou

$$\widehat{\ln preço} = 4.766 + 0.0041\ area + 0.0286\ quartos$$

Admitindo tudo o resto constante:

- Um aumento de 1 m^2 na área origina um acrescimo de aproximadamente 0.41% no preço esperado da casa
- por cada quarto adicional, o preço esperado das casas aumenta aproximadamente 2.86%

Modelo de Regressão Linear

Interpretação dos parâmetros

Caso 3 Modelo lin-log:

A variável explicada é y mas a variável explicativa $\boldsymbol{x}$ é logaritmizada.

Modelo regressão transformado:

$$E[\boldsymbol{y}|x] = \beta_0 + \beta_1 \ln(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k$$

Modelo estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_1) + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

Modelo pós incremento: $\widehat{\boldsymbol{y}^*} = \hat{\beta}_0 + \hat{\beta}_1 ln(x_1 + 1) + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

$$\widehat{\boldsymbol{y}^*} - \hat{\boldsymbol{y}} = \hat{\beta}_1 [ln(x_1 + 1) - \ln(x_1)] \cong \hat{\beta}_1 \frac{\Delta x_1}{x_1}$$

logo: $\%\Delta x_1 = 1 \rightarrow \%\Delta \hat{\boldsymbol{y}} = \frac{\hat{\beta}_1}{\mathbf{100}}$, *ceteris paribus*

Modelo de Regressão Linear

Interpretação dos parâmetros

Caso 3 Modelo lin-log: Exemplo

Exemplo: considere um novo modelo alternativo para o mesmo problema

$$\widehat{preço} = -1153.7 + 265.0 \ln(area) + 19.63\ quartos$$

Admitindo tudo o resto constante:

uma variação da área de 1% gera um aumento médio do preço das casas de aproximadamente 2.65 milhares de dólares

Notas:

. Como se pode ter $número\ de\ quartos = 0$, não se pode logaritmizar a variável quartos.

. Este modelo é menos utilizado que os outros

Modelo de Regressão Linear

Interpretação dos parâmetros

Caso 4 Modelo log-log:

A variável explicada é z mas $\boldsymbol{y = lnz}$. Suponha-se ainda que a variável explicativa é $\boldsymbol{x_J}$, e $\boldsymbol{w = ln(x_J)}$.

regressando

regressor

Modelo regressão transformado:

$$E[\boldsymbol{lnz}|x] = \beta_0 + \beta_1 \ln(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k$$

Modelo estimado: $\widehat{\boldsymbol{lnz}} = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_1) + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

Modelo pós incremento: $\widehat{\boldsymbol{lnz}^*} = \hat{\beta}_0 + \hat{\beta}_1 ln(x_1 + 1) + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

Logo $\widehat{\boldsymbol{y}^*} - \widehat{\boldsymbol{y}} = \hat{\beta}_1 [ln(x_1 + 1) - \ln(x_1)] \Rightarrow \%\Delta\widehat{\boldsymbol{y}} \cong \hat{\beta}_1 \%\Delta x_1$

Em geral, $\%\Delta x_j = 1 \rightarrow \%\Delta\widehat{\boldsymbol{y}} \cong \hat{\beta}_j \%$, *ceteris paribus*

Modelo de Regressão Linear

Caso 4 Modelo log-log:

Exemplo: considere-se agora o modelo

$$\widehat{ln(preço)} = 1.289 + 0.810\, ln(area) + 0.038\, quartos$$

Admitindo tudo o resto constante:

. uma variação da área de 1% gera um aumento no preço esperado das casas de aproximadamente 0.81%

. Por cada quarto adicional, o preço esperado das casas aumenta aproximadamente 3.8%

Modelo de Regressão Linear

Interpretação dos coeficientes do modelo

Quadro síntese: Table 2.3 Wooldrige

y^*	x^*	Interpretação de β_j	
y	x_j	$\Delta y = \beta_j \, \Delta x_j$	
y	$\ln x_j$	$\Delta y = (\beta_j / 100) \, \%\Delta x_j$	
$\ln y$	x_j	$\%\Delta y = (100 \, \beta_j) \, \Delta x_j$	Semi-elasticidade
$\ln y$	$\ln x_j$	$\%\Delta y = \beta_j \, \Delta x_j$	Elasticidade

Modelo de Regressão Linear Múltipla

Aprofundando a interpretação dos $\widehat{\beta_j}$:

Modelo: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

Modelo estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$ de resíduos $\hat{u}_i$

$\hat{\beta}_j$ é o elemento j de $\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Y$

Vamos agora analisar 2 pontos:

- Interpretação do "Partialling out" de $\widehat{\boldsymbol{\beta_j}}$
- O que acontece a $\widehat{\boldsymbol{\beta_j}}$ quando se acrescenta (ou se elimina) uma das (outras) variáveis explicativas

Modelo de Regressão Linear Múltipla

Efeito "Partialling out" :

$\widehat{\boldsymbol{\beta}_j}$ vai medir o efeito parcial da variável explicativa $\boldsymbol{x_j}$ depois de considerado o efeito das restantes variáveis explicativas o que é diferente do efeito marginal obtido numa regressão simples.

O teorema de Frisch-Waugh fundamenta este ponto, pois prova que o $\widehat{\boldsymbol{\beta}_j}$ que se obtém no $MRLM$ é idêntico ao que se obteria numa regressão linear simples de y em $\boldsymbol{r_j}$.

Resíduos da regressão de $\boldsymbol{x_j}$ nas restantes variáveis explicativas

O efeito "partialling out" consiste em retirar de $\boldsymbol{x_j}$ aquilo que é "comum" com as restantes variavéis explicativas

Modelo de Regressão Linear Múltipla

Efeito "Partialling out" :

Modelo: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$

Modelo auxiliar: $x_j = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \cdots + \hat{\alpha}_k x_k + \hat{r}_j$ (todas menos o x_j)

RL simples: $\hat{y} = \widehat{y_0} + \widehat{\beta_j}\widehat{r_j}$

Teorema de Frisch-Waugh:

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \hat{r}_{ij} y_i}{\sum_{i=1}^{n} \hat{r}_{ij}^{\,2}}, j = 1,2,\ldots,k$$

Resíduos da regressão de $\boldsymbol{x_{ij}}$ em $x_{i1}, x_{i2}, \cdots, x_{ik}$ (todas menos $\boldsymbol{x_{ij}}$)

Estimador MQ de $y_i \; em \; \widehat{r_{ij}}$

Nota: regressão simples com termo independente

Modelo de Regressão Linear Múltipla

Teorema de Frisch-Waugh - demonstração:

Iremos provar sem perda de generalidade, para $j = 1$ que $\widehat{\beta_1} = \frac{\sum_{I=1}^{n} \widehat{r_{i1}} y_i}{\sum_{i=1}^{n} \widehat{r_{i1}}^2}$

Considerem-se as 3 regressões:

1. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} + u_i$
2. $\hat{x}_1 = \widehat{\propto}_0 + \widehat{\propto}_2 x_{i2} + \cdots + \widehat{\propto}_k x_k \Leftrightarrow x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$
3. $\hat{y} = \hat{y}_0 + \hat{y}_1 \hat{r}_{i1}$

Partindo das condições 1ª ordem para $j = 1$, vem:

$$\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_k x_k) = \sum_{i=1}^{n} x_{i1}\,(y_i - \hat{y}_i) = \sum_{i=1}^{n} x_{i1}\,\hat{u}_i = 0$$

Substituindo x_{i1} por $\hat{x}_{i1} + \hat{r}_{i1}$, vem:

$$\sum_{i=1}^{n}(\hat{x}_{i1} + \hat{r}_{i1})\,\hat{u}_i = \sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i + \sum_{i=1}^{n} \hat{r}_{i1}\hat{u}_i = 0 \qquad \text{(A)}$$

Modelo de Regressão Linear Múltipla

Teorema de Frisch-Waugh - demonstração:

$$\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i = \sum_{i=1}^{n}(\hat{\alpha}_0 + \hat{\alpha}_2 x_{i2} + \cdots + \hat{\alpha}_k x_{ik})\hat{u}_i$$

$$= \hat{\alpha}_0 \sum \hat{u}_i + \hat{\alpha}_2 \sum x_{i2}\hat{u}_i + \cdots + \hat{\alpha}_k \sum x_{ik}\hat{u}_i = 0$$

(propriedades 1 e 2 dos resíduos OLS no mod 1)

$$\sum_{i=1}^{n} \hat{r}_{i1}\hat{u}_i = \sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{y}_i) = \sum \hat{r}_{i1} y_i - \sum \hat{r}_{i1}\hat{y}_i$$

$$= \sum \hat{r}_{i1} y_i - \sum \hat{r}_{i1}(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik})$$

$$= \sum \hat{r}_{i1} y_i - \hat{\beta}_0 \sum \hat{r}_{i1} - \hat{\beta}_1 \sum x_{i1}\hat{r}_{i1} - \hat{\beta}_2 \sum x_{i2}\hat{r}_{i1} - \cdots - \hat{\beta}_k \sum x_{ik}\hat{r}_{i1}$$

$$= \sum \hat{r}_{i1} y_i - \hat{\beta}_1 \sum x_{i1}\hat{r}_{i1}$$

porque: $\sum \hat{r}_{i1} = 0$ e $\sum x_{ij}\hat{r}_{i1} = 0$ para $j = 2, \ldots, k$

$$\hat{\beta}_1 = \frac{\sum \hat{r}_{i1} y_i}{\sum \hat{r}_{i1}^2}$$

Como: $\sum x_{i1}\hat{r}_{i1} = \sum(\hat{x}_{i1} + \hat{r}_{i1})\hat{r}_{i1} = \sum \overbrace{\hat{x}_{i1}\hat{r}_{i1}}^{0} + \sum \hat{r}_{i1}^2$ tem-se: $\sum \hat{r}_{i1} y_i - \hat{\beta}_1 \sum \hat{r}_{i1}^2 = 0$

Modelo de Regressão Linear Múltipla

Teorema de Frisch-Waugh - ilustração:

$$\widehat{preço} = -19.286 + 1.384 area + 15.121 quartos$$

Output EXCEL

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.794906945					
R Square	0.63187705					
Adjusted R Square	0.623215334					
Standard Error	63.04837628					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	579971.1994	289985.5997	72.95055817	3.58672E-19	
Residual	85	337883.3088	3975.097751			
Total	87	917854.5083				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-19.28550028	31.0475285	-0.621160563	0.536156232	-81.0163048	42.44530424
area(m2)	1.38360615	0.148943494	9.289470184	1.40049E-14	1.08746658	1.67974572
quartos	15.12133684	9.488597692	1.593632413	0.114730383	-3.744537442	33.98721112

Modelo de Regressão Linear Múltipla

Teorema de Frisch-Waugh - ilustração:

Obtenção $\hat{\beta}_1$ 1ª regressão auxilar: $area$ como função de $quartos$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.532120204					
R Square	0.283151911					
Adjusted R Square	0.274816468					
Standard Error	45.64604375					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	70777.80685	70777.80685	33.96963003	9.52941E-08	
Residual	86	179186.2727	2083.56131			
Total	87	249964.0795				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	66.14354244	21.31632029	3.102953114	0.002592652	23.76810017	108.5189847
quartos	33.89926199	5.816273698	5.828347109	9.52941E-08	22.33689256	45.46163143

$\widehat{area} = 66.1435 + 33.8993 quartos$

Os $\hat{r}_{i1}$ serão os resíduos desta regressão

Modelo de Regressão Linear Múltipla

Teorema de Frisch-Waugh - ilustração:

2ª regressão auxiliar $preço$ como função de $\hat{r}_1$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.611333125					
R Square	0.37372819					
Adjusted R Square	0.36644596					
Standard Error	81.75590471					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	343028.1042	343028.1042	51.32056696	2.5121E-10	
Residual	86	574826.4041	6684.027955			
Total	87	917854.5083				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	293.5460341	8.715208725	33.68204289	2.57951E-51	276.2207726	310.8712956
resid	1.38360615	0.193137569	7.163837446	2.5121E-10	0.99966137	1.76755093

$$\widehat{preço} = 293.546 + \mathbf{1.3836}\,\hat{r}_1$$

Modelo de Regressão Linear Múltipla

Efeito da adição de uma variável:

O que acontece quando se acrescenta uma nova variável explicativa ao modelo?

Exemplo: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \rightarrow \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$ → Inclinação da recta regressão de x_{i2} em x_{i1} $i = 1,2,\cdots,n$

Salvo em 2 casos especiais:

. O novo regressor é ortogonal em relação a qualquer dos outros

x_1 e x_2 não estão correlacionados $\Leftrightarrow \tilde{\delta}_1 = 0$

. O impacto do novo regressor na variável y é nulo $\Leftrightarrow \hat{\beta}_2 = 0$

Nota: Estas conclusões são **válidas** (após adaptação) no caso de **eliminação**

Modelo de Regressão Linear Múltipla

Efeito da adição de uma variável (Dem: Wooldrige):

Generalizando: $$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_r \, \tilde{\delta}_j \quad j, r = 1,2, \ldots, k \text{ e } r \neq j$$

Onde:

$\hat{\beta}_j$ coeficiente de x_j na regressão de y em $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$.
Inclui todas as variáveis explicativas

Var. adicionada

$\tilde{\beta}_j$ coeficiente de x_j na regressão de y em $\boldsymbol{x}$ com exceção de x_r.

$\tilde{\delta}_j$ coeficiente de x_j na regressão de x_r nas restantes variáveis explicativas, isto é, $\boldsymbol{x}$ com exceção de x_r.

$$\tilde{\beta}_j = \hat{\beta}_j \text{ sse } \hat{\beta}_r = 0 \text{ ou } \tilde{\delta}_j = 0$$

Modelo de Regressão Linear Múltipla

Efeito da adição de uma variável-ilustração:

Modelo: salário/hora em dolares; educação, tenure (nº anos na empresa) e exper (experiência profissional) em anos; female=1 se mulher e 0 se homem.

$$\widehat{wage.h} = -0.84503 + \mathbf{0.53799}educ + 0.16441\ tenure - 1.78839\ female$$

Estima-se uma 2ª regressão:

$$\widehat{wage.h} = -1.56794 + \mathbf{0.57150}educ + 0.14101\ tenure - 1.81085\ female + 0.02539\ exper$$

Calcula-se: $\widehat{\beta}_1 - \tilde{\beta}_1 = 0.57150 - 0.53799 = 0.03351 \Leftrightarrow \Delta \approx 6.3\%$

(alternativa) Estima-se uma 3ª regressão:

$$\widehat{exper} = 28.4654 - \mathbf{1.31952}educ + 0.92169\ tenure + 0.88433\ female$$

Calcula-se: $-\widehat{\beta}_4\tilde{\delta}_1 = -0.02539 * (-1.31952) = 0.03350$ (efeito fraco)

Modelo de Regressão Linear Múltipla

Avaliação da qualidade do ajustamento:

Modelo: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

Modelo estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \beta_k x_k$

1ª avaliação: comparar $\hat{y}_i$ com y_i e calcular o coef. correlação

$$r_{y\hat{y}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\left(\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)\right)^{1/2}} = \frac{s_{y\hat{y}}}{s_y\, s_{\hat{y}}}$$

Sempre positivo pq $\hat{y}$ melhor aproxima-ção a y

Então, usa-se: $r^2_{y\hat{y}}$ (varia entre 0 e 1. qto mais próximo de 1 melhpr o ajustamento)

Modelo de Regressão Linear Múltipla

Decomposição da variação total e R^2:

Avaliação mais habitual: calcular a proporção da variação da variável y que é "explicada" pelo modelo.

Ideia: decompor a variação total (SST) em variação explicada (SSE) e variação residual (SSR)

$$\boldsymbol{Variação\ total\ (SST)\ =\ Variação\ explicada\ (SSE)\ +\ Variação\ residual\ (SSR)}$$

E ver o peso da variação explicada na variação total → $\boldsymbol{R^2}$

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

Modelo de Regressão Linear Múltipla

Decomposição da variação total e R^2:

$\hat{y}_i$ com y_i

$$\textbf{Variação total } (SST) = \textbf{Variação explicada } (SSE) + \textbf{Variação residual } (SSR)$$

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

como $y_i = \hat{y}_i + \hat{u}_i \Rightarrow \sum y_i = \sum \hat{y}_i + \overbrace{\sum \hat{u}_i}^{0} = \sum \hat{y}_i \Rightarrow \bar{y} = \bar{\hat{y}}$

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

Como o modelo tem termo independente $\sum \hat{u}_i = 0 \Rightarrow \bar{\hat{u}} = 0$

Modelo de Regressão Linear Múltipla

Decomposição da variação total e R^2:

$\hat{y}_i$ com y_i

$$\textbf{Variação total (SST)} = \textbf{Variação explicada (SSE)} + \textbf{Variação residual (SSR)}$$

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

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

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

$$= SSR + SSE + 2\sum_{i=1}^{n}\hat{u}_i\hat{y}_i - 2\bar{y}\sum_{i=1}^{n}\hat{u}_i = SSE + SSR$$

Propriedades resíduos: $\sum_{i=1}^{n}\hat{u}_i = 0, \sum_{i=1}^{n}\hat{u}_i\hat{y}_i = 0$

Modelo de Regressão Linear Múltipla

Decomposição da variação total e R^2:

$\hat{y}_i$ com y_i

$$\textbf{Variação total } (SST) = \textbf{Variação explicada } (SSE) + \textbf{Variação residual } (SSR)$$

Tabela Anova:

	SS	DF	MS
Explicados	$\sum(\hat{y}_i - \bar{y})^2$	k	MSE
Residual	$\sum \hat{u}_i^{\,2}$	$n-k-1$	$MSR = \hat{\sigma}^2$
Total	$\sum(y_i - \bar{y})^2$	$n-1$	s_y^2

Modelo de Regressão Linear Múltipla

R^2 — Interpretação

Alguns comentários:

- Recorrendo às propriedades dos resíduos, mostra-se que no MRLM (com termo independente) $r^2_{y\hat{y}} = R^2$.

$$\bar{\hat{y}} = \frac{\sum \hat{y}_i}{n} = \frac{\sum (y_i - \hat{u}_i)}{n} = \frac{\sum y_i}{n} - \frac{\sum \hat{u}_i}{n} = \bar{y}$$ Propriedade 1 dos resíduos OLS

$$\begin{aligned}\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}) &= \sum (\hat{y}_i - \bar{y} + \hat{u}_i)(\hat{y}_i - \bar{y}) \\ &= \sum \left((\hat{y}_i - \bar{y})^2 + \hat{u}_i(\hat{y}_i - \bar{y})\right) \\ &= \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{u}_i \hat{y}_i - \bar{y} \sum \hat{u}_i \\ &= \sum (\hat{y}_i - \bar{y})^2 \qquad \text{Propriedades 1 e 3 dos resíduos OLS}\end{aligned}$$

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

$$= \frac{SSE}{SST} = R^2$$

Modelo de Regressão Linear Múltipla

R^2 — Interpretação

Alguns comentários:

- De um modo geral $SSE = R^2\ SST$ ou $\text{SSR} = (1 - R^2)\ SST$

$$SSE = R^2\ SST \Leftrightarrow SST - SSR = R^2\ SST \Leftrightarrow SST - R^2\ SST = SSR$$

- $R^2 = 1$ significa que $SSR = \sum_{i=1}^{n} \hat{u}_i^2 = 0$, isto é, que o ajustamento é perfeito $\hat{y}_i = y_i$.

- $R^2 = 0$ significa que $SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 0$, isto é, que $\hat{y}_i = \bar{y}$ ($\hat{y}_i$ é constante) e que o modelo nada adianta.

Modelo de Regressão Linear Múltipla

R^2 — Interpretação

Alguns comentários:

- A interpretação do valor do R^2 levanta ainda 2 problemas:
  - Consoante o problema que se esteja a modelar, o R^2 assume valores bem diferentes: um valor de 0.3 tanto pode corresponder a um bom valor como a um mau modelo (depende do que se esteja a modelar)
  - Quando se acrescentam variáveis explicativas ao modelo, mostra-se que o R^2 não pode decrescer! Assim, o R^2 tende a valorizar modelos com excesso de variáveis explicativas.

Modelo de Regressão Linear Múltipla

R^2 *ajustado* – Interpretação

- A solução mais comum para corrigir este último problema consiste em recorrer ao coeficiente de determinação corrigido (Adusted R square ou $\bar{R}^2$) que "penaliza" o modelo pela introdução de novas variáveis explicativas.

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

O inconveniente desta nova medida é não ter uma interpretação tão intuitiva quanto o R^2.

Também se mostra que

$$\bar{R}^2 = 1 - (1-R^2)\frac{(n-1)}{(n-k-1)}$$

Modelo de Regressão Linear Múltipla

R^2 – Comparação de modelos

	$R^2 = \frac{SSE}{SST}$	$r^2_{y\hat{y}} = cor(Y, \hat{Y})^2$	$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$
Os modelos têm todos termo independente	V		V
Os modelos têm todos a mesma variável dependente (y)	V		V
O número de variáveis explicativas (k) é o mesmo	V	V	

Modelo de Regressão Linear Múltipla

Caso particular – Regressão sem termo autónomo $(\beta_0 = 0)$:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$$

Recorrer à expressão geral para obter $\hat{\beta}_j$

Ter presente que este modelo **não verifica** $\sum_{i=1}^{N} \hat{u}_i = 0$, o que vai ter pesadas consequências nas propriedades dos estimadores.

Desde já se pode adiantar que não incluir termo independente quando se deveria é **muito mais grave** do que incluir quando não se deveria. Assim, só se utiliza este modelo em situações muito particlares.

Modelo de Regressão Linear Múltipla

Hipóteses do MRLM:

Para analisar a qualidade do estimador MQ (OLS) e poder definer procedimentos de inferência estatística torna-se necessário assumer um conjunto de hipóteses sobre o MRLM.

MRL 1 – Modelo linear nos parâmetros

O "verdadeiro" modelo na população é dado por:

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

Erro aleatório não observável

Observações:

1. Caso a relação entre y e $\boldsymbol{x}$ não for esta, o modelo estaria mal especificado

2. Deve ter-se presente a flexibilidade "adicional" dada pela possibilidade de "transformar as variáveis.

Modelo de Regressão Linear Múltipla

Hipóteses do MRLM:

MRL 2 – Amostra aleatória

As $\boldsymbol{n}$ observações $(y_i, x_{i1}, x_{i2}, \cdots, x_{ik}) \quad i = 1, 2, \cdots, n$ constituem uma amostra aleatória

Observações:

1. O modelo definido em **MRL 1** **aplica-se a todas as observações**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

2. As observações são consideradas independentes (dados seccionais).

Modelo de Regressão Linear Múltipla

Hipóteses do MRLM:

MRL 3 – Ausência de colinearidade perfeita

Nenhuma das variáveis explicativaspode ser combinação linear das restantes nem na amostra nem na população

Observações:

1. Garante que a matriz $\boldsymbol{X}^T\boldsymbol{X}$ tem inversa, i.é, existe $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

2. A existir colinearidade perfeita na amostra, ela é facilmente identificável no output

3. As variáveis explicativas podem estar correlacionadas sem pôr em causa esta hipótese desde que a correlação não atinja valores “significativos”. Os problemas de multicolinearidade serão tratados mais adiante.

Modelo de Regressão Linear Múltipla

Multicolinearidade:

Utilizando o teorema de Frisch-Waugh é possível aprofundar a interpretação da $var(\hat{\beta}_j|\boldsymbol{X})$ já que

$$var(\hat{\beta}_j|\boldsymbol{X}) = \frac{\sigma^2}{SST_j\left(1-R_j^2\right)}$$

Coeficiente de determinação da regressão de x_j nas outras var(s) explicativas

$$\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2$$

Ilustração com base no exemplo anterior (utilizando $\hat{\sigma}^2$):

- Do modelo principal tira-se $\hat{\sigma}_{\hat{\beta}_1} = 0.148943494$ e $\hat{\sigma}^2 = 3975.097751$
- Da regressão auxiliar tira-se $SST_1 = 249964.0795$ e $R_1^2 = 0.283151911$
- Como se pode verificar, $0.148943494^2 = \frac{3975.097751}{249964.0795\times 0.283151911}$

Modelo de Regressão Linear Múltipla

Multicolinearidade: $var(\hat{\beta}_j|\boldsymbol{X}) = \frac{\sigma^2}{SST_j\left(1-R_j^2\right)}$

Comentários:

. σ^2 é a variância da variável residual.Quanto menor for esta variância, melhor o modelo e menor a variabilidade dos estimadores (eficiência)

. SST_j é a variação total da variável x_j na amostra.

Se $n\uparrow \rightarrow \uparrow SST_j = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2 \rightarrow \downarrow var(\hat{\beta}_j|\boldsymbol{X})$

. Quanto mais dependente estiver x_j das restantes variáveis explicativas, maior será o valor de $R_j^2 \rightarrow maior\ var(\hat{\beta}_j|\boldsymbol{X})$

Modelo de Regressão Linear Múltipla

Multicolinearidade: $var(\hat{\beta}_j|\boldsymbol{X}) = \frac{\sigma^2}{SST_j\left(1-R_j^2\right)}$

Comentários:

. A situação ideal é ter-se $R_j^2 = 0 \rightarrow x_j$ é ortogonal a qualquer das outras variáveis explicativas.

. Quando $R_j^2 = 1$ →violação da MLR 3 e $var(\hat{\beta}_j|\boldsymbol{X}) = \infty$

. A multicolinearidade surge quando R_j^2 é grande, sem atingir o valor 1. Neste caso, existe o perigo de $var(\hat{\beta}_j|\boldsymbol{X})$ estar inflacionada, embora um valor grande de SST_j possa mitigar (pelo menos parcialmente) o problema.

O estimador contínua a ser **BLUE** embora de fraca qualidade.

Modelo de Regressão Linear Múltipla

Multicolinearidade:

Possíveis soluções:

. Recolher mais informação, i. é, uma amostra de maior dimensão.

. Introduzindo alguma restricção linear entre os β_j que permita reformular o modelo. Que restrição?. Não esquecer que tem de ser algo que tem de se verificar para a população.

. Eliminar alguma variável explicativa. Mas cuidado pois é muito mais grave eliminar erradamente uma variável que incluir erradamente uma variável adicional.

Modelo de Regressão Linear Múltipla

Hipóteses do MRLM:

MRL 4 – Exogeneidade $E(u|\boldsymbol{x}) = 0$

Observações:

1. Esta hipótese é **fundamental**. Não pode existir correlação entre as variáveis explicativas $\boldsymbol{x_j}$ e a variável residual u. Se tal acontecer a variável diz-se endógena e esta hipótese falha

2. A falha pode dever-se a má especificação do modelo: omissão de variáveis explicativas correlacionadas com as restantes, não transformação das variáveis quando necessária [a ser tratado mais adiante]

3. Enquanto **MRL 3** diz respeito às var.(s) explicativas, **MRL 4** tem a ver com as variáveis não incluídas no modelo, cujo efeito é representado por u

Modelo de Regressão Linear Múltipla

Hipóteses do MRLM:

MRL 5 – Homocedasticidade $var(u|\boldsymbol{x}) = \sigma^2$

Observações:

1. A variância da variável residual u é constante, i.é, não depende dos valores assumidos pelas variáveis explicativas

2. Esta hipótese nem sempre é realista. O seu levantamento será discutido mais adiante.

3. Tendo em conta que as observações da amostra são independentes (bastaria não correlacionadas), $var(U|\boldsymbol{X}) = \sigma^2 \, \mathrm{I}$, i. é,

$$var(u_i|\boldsymbol{X}_{i.}) = \sigma^2 \quad \text{e} \quad cov\left(u_i, u_j|\boldsymbol{X}\right) = 0, i, j = 1, 2, \cdots, n \qquad i \neq j$$

Modelo de Regressão Linear Múltipla

Hipóteses do MRLM:

MRL 5 – Homocedasticidade - ilustração

FIGURE 2.8

The simple regression model under homoskedasticity.

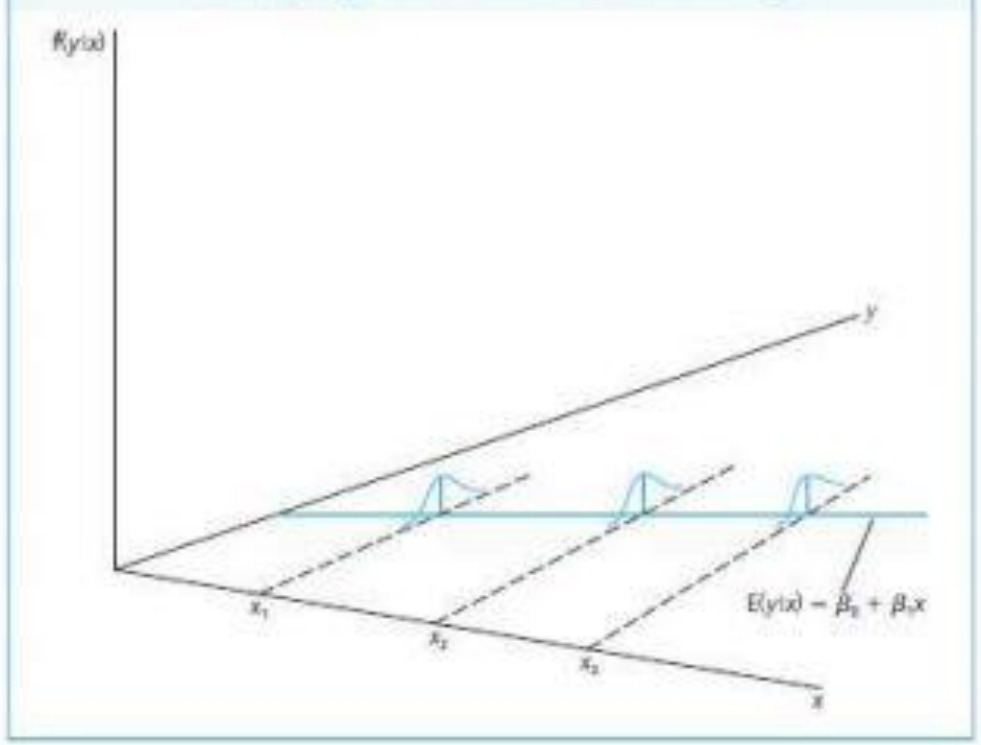
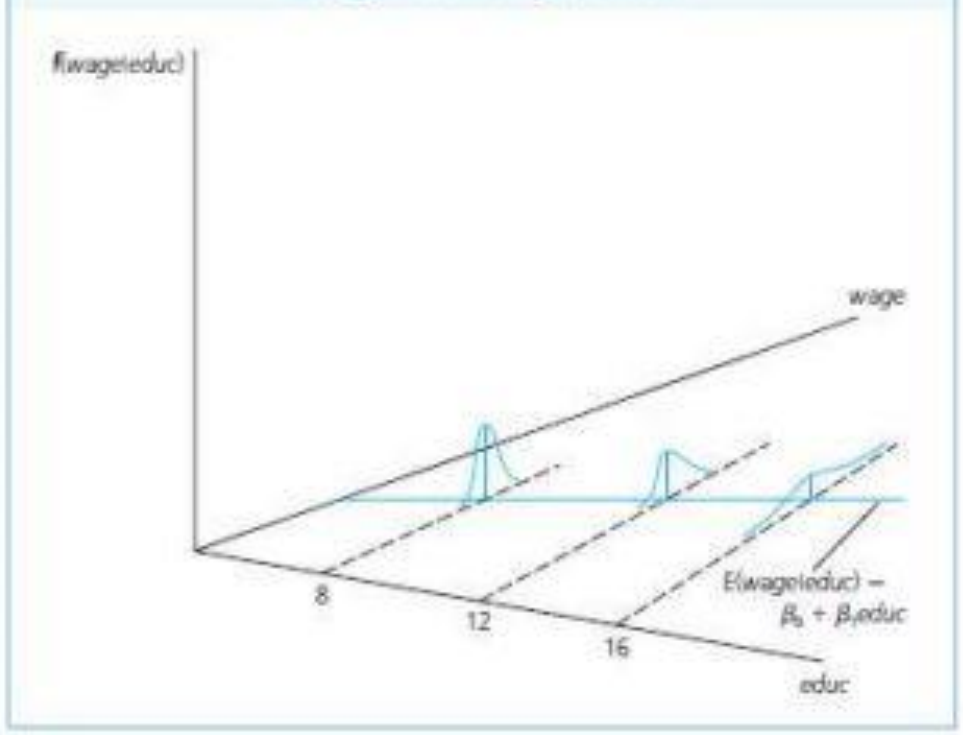


FIGURE 2.9

Var(wage|educ) increasing with educ.



Modelo de Regressão Linear Múltipla

Propriedades do estimador dos MQ $(\widehat{\boldsymbol{\beta}})$:

Não enviesamento: $E(\widehat{\boldsymbol{\beta}}_j) = \beta_j, \qquad j = 1, 2, \cdots, k$

Demonstração:

$$Y = X\beta + U \qquad \hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$$

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta + U = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T U$$

$$\hat{\beta} = \beta + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T U$$

$$\Rightarrow E(\hat{\beta}|\boldsymbol{X}) = E(\beta|\boldsymbol{X}) + E\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T U|\boldsymbol{X}\right) = \beta + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \underbrace{E(U|\boldsymbol{X})}_{0}$$

$$= \beta$$

$\hat{\beta}$ é um estimador centrado para β

Modelo de Regressão Linear Múltipla

Propriedades do estimador dos $MQ\ (\hat{\beta})$ - enviesamento:

Efeito da inclusão de variáveis irrelevantes (sobreespecificação):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_i$$

Modelo satisfaz as hipóteses MRL1 A MRL4, mas x_3 não não tem nenhum impacto sobre y após os efeitos de x_1 e x_2 terem sido controlados $\rightarrow \beta_3 = 0$.

Nota: x_3 pode ou não estar correlacionada com x_1 e x_2

Como não sabemos que $\beta_3 = 0$, o modelo estimado é:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

T. 3.1 diz-nos que o facto de $\beta_3 = 0$, não afecta o não enviesamento de $\hat{\beta}_3$, já que $E(\hat{\boldsymbol{\beta}}_j) = \beta_j$ para qualquer valor de β_j, incluindo $\beta_j = 0$.

Modelo de Regressão Linear Múltipla

Propriedades do estimador dos MQ $(\hat{\beta})$- enviesamento:

Efeito da omissão de variáveis relevantes (sub-especificação):

(verdadeiro modelo população) $\leftarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ (satisfaz MLR1 a MLR4)

Contudo, por ignorância ou indisponibilidade da dados, o modelo foi estimado sem a variável x_2.

O modelo estimado sem variável x_2: $y = \beta_0 + \beta_1 x_1 + v \ \ com\ v = \beta_2 x_2 + u$

O modelo estimado é: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$ Lembrando $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$

Hip(s) MRL1 a 4→ $\hat{\beta}_1$ e $\hat{\beta}_2$ são não enviesados pelo que se tem:

$$\Rightarrow E(\tilde{\beta}_1) = E(\hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1) = \mathrm{E}(\hat{\beta}_1) + \tilde{\delta}_1 \mathrm{E}(\hat{\beta}_2) = \beta_1 + \tilde{\delta}_1 \beta_2 \Leftrightarrow \overbrace{E(\tilde{\beta}_1) - \beta_1}^{env.\beta_1} = \tilde{\delta}_1 \beta_2$$

Modelo de Regressão Linear Múltipla

Propriedades do estimador dos $MQ\ (\widehat{\boldsymbol{\beta}})$ - enviesamento:

Efeito da omissão de variáveis relevantes (sub-especificação):

$$\Rightarrow E(\tilde{\beta}_1) = E(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1) = \mathrm{E}(\hat{\beta}_1) + \tilde{\delta}_1\mathrm{E}(\hat{\beta}_2) = \beta_1 + \tilde{\delta}_1\beta_2 \Leftrightarrow \overbrace{E(\tilde{\beta}_1) - \beta_1}^{env.\beta_1} = \tilde{\delta}_1\beta_2$$

β_1 será não enviesado sse $\beta_2 = 0$ ou $\tilde{\delta}_1 = 0$ mesmo com $\beta_2 \neq 0$ — desconhecido

Como $\tilde{\delta}_1$ é a $covar(x_1, x_2)$ no cálculo de $var(x_1)$, na amostra:

- $\tilde{\delta}_1 = 0$ sse x_1 e x_2 não estiverem correlacionadas na amostra
- $\tilde{\delta}_1 \neq 0 \rightarrow x_1$ e x_2 estiverem correlacionadas na amostra e $\tilde{\delta}_1$ terá o sinal da $corr(x_1, x_2)$ [Table3.2 pg.86 Wooldrige]

Dimensão do enviesamento é também importante. Um pequeno enviesamento (0.1%) não é motivo de preocupação mas um enviesamento de 3% ou superior é um problema muito sério

Modelo de Regressão Linear Múltipla

Propriedades do estimador dos $MQ\ (\widehat{\beta})$ - enviesamento:

Efeito da omissão de variáveis relevantes (sub-especificação) Exemplo:

$$log(sal) = \beta_0 + \beta_1 educ + \beta_2 capacidades + u$$

Não existem dados sobre capacidades. Por isso estimou-se:

$$\widehat{log(sal)} = 0.584 + 0.083\ educ \quad n = 526, \quad R^2 = 0.186$$

Capacidades e educação deverão estar positivamente correlacionadas pelo que:

$$log(sal) = \beta_0 + \beta_1 educ + v \quad com\ v = \beta_2 capacidades + u$$

→estimativas OLS, estarão em média, inflacionadas. Tal não significa que $0.083 > \beta_1$. Pode apenas dizer-se, numa perspectiva frequencista, que a média das estimativas para um grande nº amostras será maior que β_1.

Modelo de Regressão Linear Múltipla

Matriz das var/cov do estimador dos MQ (OLS) :

Recordando: $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$ e $Y = \boldsymbol{X\beta} + \boldsymbol{U}$ vem $\hat{\beta} = \beta + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T U$;

$\mathrm{var}(U|\boldsymbol{X}) = \sigma^2 I$

Demonstração:

$$var(\hat{\beta}|\boldsymbol{X}) = var\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T U|\boldsymbol{X}\right) \quad (\beta \text{ não é aleatório})$$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T var(U|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\,\mathrm{I})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$= \sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}}_{I} = \sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Como σ^2 é um parâmetro desconhecido, $var(\hat{\beta}|\boldsymbol{X})$ também o é.

Para estimar $var(\hat{\beta}|\boldsymbol{X})$, matriz simétrica de dimensão $(k+1) * (k+1)$ bastará estimar σ^2 pois a matriz $\boldsymbol{X}$ é observável

Modelo de Regressão Linear Múltipla

Matriz das var/cov do estimador dos $MQ\ (OLS)$:

A variância condicionada de um dado $\hat{\beta}_j, j = 1, 2, \cdots, k \rightarrow \sigma^2_{\hat{\beta}_j}$ é dada pelo elemento (j, j) na diagonal principal da matrix $\sigma^2\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}$

A covariância condicionada entre $\hat{\beta}_i$ e $\hat{\beta}_r$ é dada pelo elemento (i, r) ou (r, i)da matrix já que $cov(\hat{\beta}_i, \hat{\beta}_r) = cov(\hat{\beta}_r, \hat{\beta}_i)$

$$var(\hat{\beta}|\boldsymbol{X}) = \begin{bmatrix} var(\hat{\beta}_0|\boldsymbol{X}) & cov(\hat{\beta}_0, \hat{\beta}_1|\boldsymbol{X}) & \cdots \\ cov(\hat{\beta}_1, \hat{\beta}_0|\boldsymbol{X}) & var(\hat{\beta}_1|\boldsymbol{X}) & \cdots \\ \cdots & & \cdots \\ cov(\hat{\beta}_k, \hat{\beta}_0|\boldsymbol{X}) & cov(\hat{\beta}_k, \hat{\beta}_1|\boldsymbol{X}) & \end{bmatrix}$$

Modelo de Regressão Linear Múltipla

Estimação de σ^2:

Para estimar $\boldsymbol{\sigma^2}$ utiliza-se:

$$\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}} = \frac{\sum \hat{\boldsymbol{u}}_{\boldsymbol{i}}^{\boldsymbol{2}}}{\boldsymbol{n-k-1}} = \frac{\boldsymbol{U^T U}}{\boldsymbol{n-k-1}}$$

Estimador centrado para $\boldsymbol{\sigma^2}$

Observação: $\hat{\boldsymbol{\sigma}}$ - erro padrão da regressão

Porquê este estimador?

Sendo $\sigma^2 = var(u_i|\boldsymbol{x})$ a solução natural seria considerar a variância observada de u, S_u^2 ou $S_u'^2$ mas não sendo u observável, recorre-se à utilização de $\hat{u}_i$ introduzindo uma compensação no denominador para garantir que o estimador é centrado

A demonstração de $E\left(\hat{\boldsymbol{\sigma}}^{\boldsymbol{2}}\right) = \sigma^2$ está fora do âmbito do curso

Modelo de Regressão Linear Múltipla

Matriz estimada das var/cov de $\widehat{\boldsymbol{\beta}}$:

A matriz estimada das variâncias / covariâncias condicionais de $\widehat{\boldsymbol{\beta}}$ é:

$$\widehat{var}(\hat{\beta}|\boldsymbol{X}) = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \widehat{var}(\hat{\beta}_0|\boldsymbol{X}) & \widehat{cov}(\hat{\beta}_0, \hat{\beta}_1|\boldsymbol{X}) & \cdots \\ \widehat{cov}(\hat{\beta}_1, \hat{\beta}_0|\boldsymbol{X}) & \widehat{var}(\hat{\beta}_1|\boldsymbol{X}) & \cdots \\ \cdots & & \cdots \\ \widehat{cov}(\hat{\beta}_k, \hat{\beta}_0|\boldsymbol{X}) & \widehat{cov}(\hat{\beta}_k, \hat{\beta}_1|\boldsymbol{X}) & \end{bmatrix}$$

$$\hat{\sigma}_{\widehat{\boldsymbol{\beta}}_j} = \sqrt{\widehat{var}(\widehat{\boldsymbol{\beta}}_j)} \rightarrow \textbf{erro padrão do estimador } \widehat{\boldsymbol{\beta}}_j$$

Nota: No caso do $MRLS$, $\hat{\sigma}_{\widehat{\boldsymbol{\beta}}_1}{}^2 = \dfrac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Modelo de Regressão Linear Múltipla

Teorema de Gauss Markov- estimador BLUE:

Verificadas as hipóteses **MRL 1** a **MRL 5**, prova-se que, de entre os estimadores lineares e centrados, o estimador OLS é o de menor variância, sendo assim o mais eficiente.

Estimador OLS é BLUE (Best Linear Unbiased Estimator)

Best → menor variância

Linear → Cada $\hat{\beta}$ é uma combinação linear dos y_i

$$\hat{\beta} = \overbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T}^{A} Y = \underbrace{A}_{(k+1)xn} \quad Y \Rightarrow \hat{\beta}_j = \sum_{i=}^{n} a_{i,j}\, y_i$$

Elemento (i,j) de A

Unbiased → $E(\hat{\beta}|\boldsymbol{X}) = \beta$

Modelo de Regressão Linear Múltipla

Output EXCEL:

Modelo estimado:

$$\widehat{preço} = -19.286 + \mathbf{1.3836}\ area + 15.121\ quartos$$

Identificar:

$$\widehat{\beta} = \begin{bmatrix} -19.286 \\ 1.3836 \\ 15.121 \end{bmatrix}$$

$$\hat{\sigma}_{\widehat{\beta}_0} = 31.048$$
$$\hat{\sigma}_{\widehat{\beta}_1} = 0.1489$$
$$\hat{\sigma}_{\widehat{\beta}_2} = 9.4886$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.794906945					
R Square	0.63187705					
Adjusted R Square	0.623215334					
Standard Error	63.04837628					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	579971.1994	289985.5997	72.95055817	3.58672E-19	
Residual	85	337883.3088	3975.097751			
Total	87	917854.5083				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-19.28550028	31.0475285	-0.621160563	0.536156232	-81.0163048	42.44530424
area(m2)	1.38360615	0.148943494	9.289470184	1.40049E-14	1.08746658	1.67974572
quartos	15.12133684	9.488597692	1.593632413	0.114730383	-3.744537442	33.98721112

Modelo de Regressão Linear Múltipla

Output Eviews:

Modelo estimado:

$$\widehat{preço} = -19.286 + 1.3836\, area + 15.121\, quartos$$

Identificar: $R^2 = 0.6319$ $\bar{R}^2 = 0.6232$ $n = 88$ $SSR = \sum \hat{u}_i^2 = 337883.3$

$$\widehat{\beta} = \begin{bmatrix} -19.286 \\ 1.3836 \\ 15.121 \end{bmatrix}$$

Dependent Variable: PRECO
Method: Least Squares
Date: 03/24/20 Time: 17:00
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-19.28550	31.04753	-0.621161	0.5362
AREA	1.383606	0.148943	9.289470	0.0000
QUARTOS	15.12134	9.488598	1.593632	0.1147

R-squared	0.631877	Mean dependent var	293.5460
Adjusted R-squared	0.623215	S.D. dependent var	102.7134
S.E. of regression	63.04838	Akaike info criterion	11.15918
Sum squared resid	337883.3	Schwarz criterion	11.24363
Log likelihood	-488.0038	Hannan-Quinn criter.	11.19320
F-statistic	72.95056	Durbin-Watson stat	1.857617
Prob(F-statistic)	0.000000		

Modelo de Regressão Linear Múltipla

Output Eviews:

Coefficient Covariance Matrix

	C	AREA_M2_	QUARTOS
C	963.9490	-1.467339	-180.5499
AREA	-1.467339	0.022184	-0.752027
QUART	-180.5499	-0.752027	90.03349

Modelo de Regressão Linear Múltipla

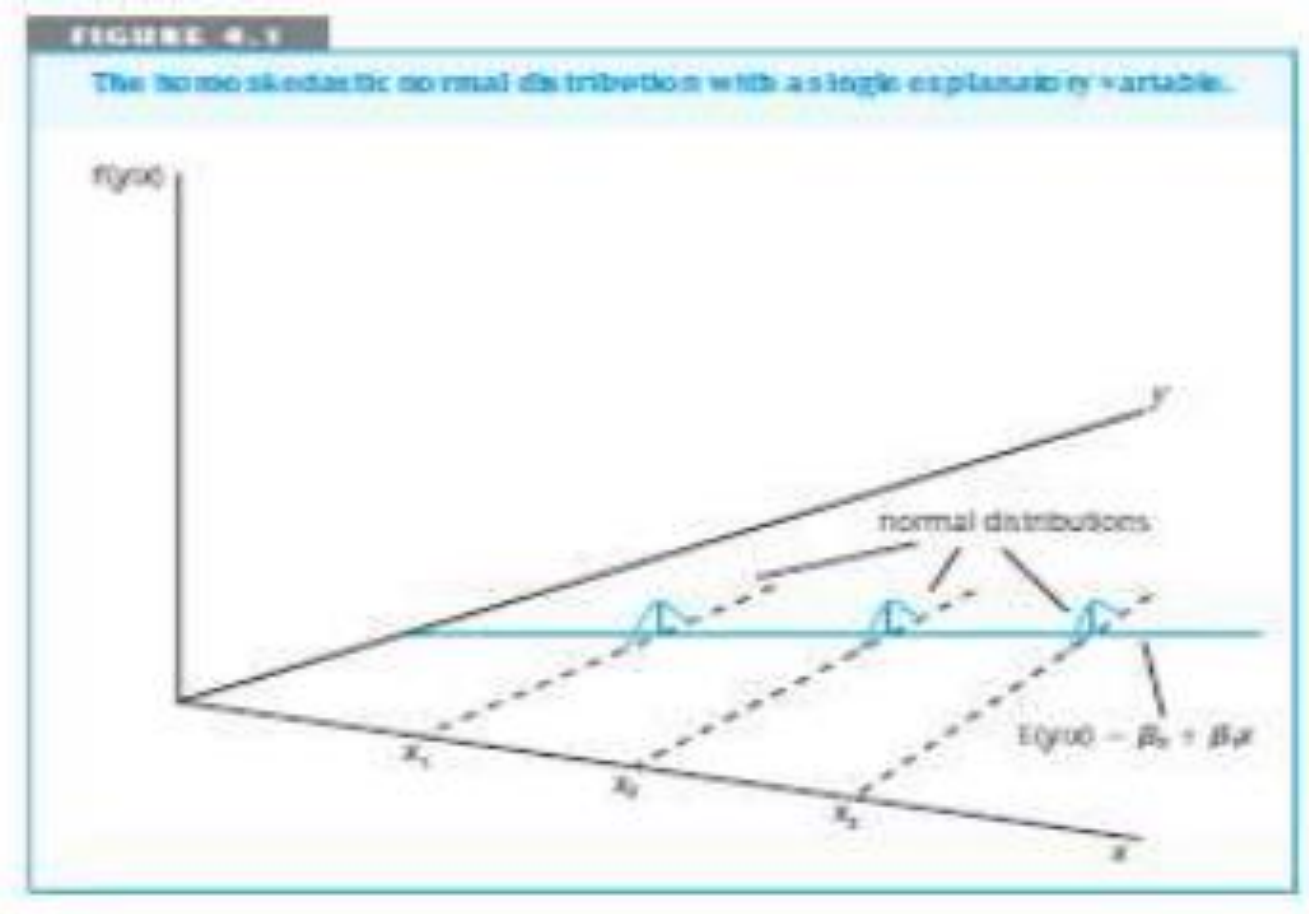
Inferência estatística sobre o modelo

$$
\begin{cases}
\beta_j \begin{cases}
\text{teste à significância estatística} & H_0: \beta_j = 0 \ \ vs \ H_1: \beta_j \neq 0 \\
\text{teste ao sinal} & \begin{cases} H_0: \beta_j \geq 0 \ \ vs \ H_1: \beta_j < 0 \\ H_0: \beta_j \leq 0 \ \ vs \ H_1: \beta_j > 0 \end{cases} \\
\text{teste a um valor particular} & \begin{cases} H_0: \beta_j = c \ \ vs \ H_1: \beta_j \neq c \\ H_0: \beta_j \geq c \ \ vs \ H_1: \beta_j < c \\ H_0: \beta_j \leq c \ \ vs \ H_1: \beta_j > c \end{cases} \\
\text{teste a uma combinação linear de coeficientes}
\end{cases} \\
\text{subconjunto regressores} \begin{cases}
\text{teste à significância conjunta de } q \text{ coef.(s)} \\
H_0: \beta_{q+1} = 0, \beta_{q+2} = 0, \cdots, \beta_k \ \ vs \ \ H_1: \exists \, \beta_j \neq 0 \ \ (j = q+1, \cdots, k) \\
\text{teste a várias combinações lineares de coeficientes}
\end{cases} \\
\text{teste à significância conjunta de todos os regressores (validade modelo)} \\
H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \ \ \ contra \ \ H_1: \exists \, \beta_j \neq 0 \ \ (j = 1, \cdots, k)
\end{cases}
$$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

$MRL\,6$ – Distribuição normal da variável residual $u_i|X \sim N(0, \sigma^2)$



Observações:

Hipótese forte porque implica normalidade dos resíduos, $E(u|\boldsymbol{x})$=0 (MLR4) e $var(u|\boldsymbol{x})$= var(u)=σ^2 (MLR5)

O estimador OLS, $\widehat{\boldsymbol{\beta}}$ passa a ser o mais eficiente entre os estimadores centrados.

Modelo de Regressão Linear Múltipla

$MRL\ 6$ – Distribuição normal da variável residual

Observações:

1. Esta hipótese é bastante mais forte que qualquer das outras já que, para além da distribuição normal implica MRL 4 e MRL 5.
   A independência entre u e $\boldsymbol{x} \rightarrow E(u|\boldsymbol{x}) = E(u) = 0$
   $$var(u|\boldsymbol{x}) = var(u) = \boldsymbol{\sigma^2}$$

2. O conjunto de hipóteses MRL 1 a MRL 6 – Hipóteses clássicas MRL no quadro dos modelos seccionais

3. No quadro das hipóteses MRL 1 a MRL 6 (Hipóteses clássicas MRL) o estimador $\widehat{\boldsymbol{\beta}}$ para além de **BLUE** passa a ser também o mais eficiente de entre os estimadores centrados.

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Distribuição por amostragem do estimador OLS $\hat{\boldsymbol{\beta}}$

Cada uma das $i = 1,2,\cdots,n$ observações:

$$y_i|\boldsymbol{x}\sim n(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + +\beta_k x_{ik};\ \sigma^2)$$

Em termos matriciais: $Y|\boldsymbol{X}\sim N(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}) \Rightarrow \hat{\beta}\sim N(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$

Já que, $\hat{\beta} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-\boldsymbol{1}}Y$ é uma combinação linear de v.a.(s) com distribuição normal

$\hat{\beta}_j\sim N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)$ Estandardizando, tem-se: $\dfrac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}}\sim N(0,1)$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Distribuição por amostragem do estimador OLS $\widehat{\boldsymbol{\beta}}$

Como $\sigma_{\widehat{\beta}_j}$ é desconhecido tem de utilizar-se o estimador de $\sigma_{\widehat{\beta}_j} \rightarrow$ $\hat{\sigma}_{\widehat{\beta}_j}$ pelo que se tem:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\widehat{\beta}_j}} \sim t(n-k-1)$$

Esta será a variável usada para fazer inferência sobre β_j:

- Variável fulcral em intervalos de confiança

- $\hat{\beta}_j$ - Estatística teste em ensaios de hipóteses

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Intervalo de confiança para β_j

Variável fulcral $\rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n-k-1)$

Intervalo de confiança a $(1-\alpha)100\%$ para β_j:

$$\left(\hat{\beta}_j - t_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j}; \hat{\beta}_j + t_{\alpha/2}\hat{\sigma}_{\hat{\beta}_j}\right)$$

Com $t_{\alpha/2}$: $P\left(T_{(n-k-1)} > t_{\alpha/2}\right) = \alpha/2$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Intervalo de confiança para β_j

Exemplo: $\widehat{preço} = -19.286 + 1.384\, area + 15.121\, quartos$

(31.046) (0.1489) (9.4886)

$IC_{\beta_1}^{95\%} = ?$

Variável fulcral $\rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n-k-$

$$\left(\hat{\beta}_j - t\alpha_{/2}\hat{\sigma}_{\hat{\beta}_j}; \hat{\beta}_j + t\alpha_{/2}\hat{\sigma}_{\hat{\beta}_j}\right)$$

Com $t\alpha_{/2}: P\left(T_{(n-k-1)} > t\alpha_{/2}\right) = \alpha/2$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Ensaio de hipóteses para β_j

$$\text{Estatística teste} \rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n-k-1)$$

Neyman-Pearson	Valor-p
Estabelecer as hipóteses em teste, H_0 e H_1	
Definir a estatística de teste (substituir β pelo valor (fronteira) em H_0)	
Definir α	Calcular $t_{j,obs}$
Construir W (região de rejeição)	Calcular o valor-p
Calcular $t_{j,obs}$	Decidir
Decidir	

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Significância de $x_j \Leftrightarrow$ *significância estatística* β_j

$H_0: \beta_j = 0 \quad contra\ H_1: \beta_j \neq 0$

Estatística teste $\rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n-k-1)$

Rejeitar H_0 significa que x_j é relevante para explicar o comportamento de y

Nota: este teste figura nos outputs de computador

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste ao sinal de β_j Estatística teste $\rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n-k-1)$

$H_0: \beta_j = 0 \quad contra\ H_1: \beta_j > 0$ ou $H_0: \beta_j \leq 0 \quad contra\ H_1: \beta_j > 0$

Testa um impacto positivo

$H_0: \beta_j = 0 \quad contra\ H_1: \beta_j < 0$ ou $H_0: \beta_j \geq 0 \quad contra\ H_1: \beta_j < 0$

Testa um impacto negativo

Atenção: a resposta é afirmativa quando se rejeita H_0

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste para um valor particular de β_j

$H_0: \beta_j = c \quad contra\ H_1: \beta_j > c$ ou $H_0: \beta_j \leq c \quad contra\ H_1: \beta_j > c$

$H_0: \beta_j = c \quad contra\ H_1: \beta_j < c$ ou $H_0: \beta_j \geq c \quad contra\ H_1: \beta_j < c$

$$\text{Estatística teste} \rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n-k-1)$$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Ensaio de hipóteses para β_j

Output EXCEL

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.794906945			
R Square	0.63187705			
Adjusted R Square	0.623215334			
Standard Error	63.04837628			
Observations	88			
ANOVA				
	df	SS	MS	F
Regression	2	579971.1994	289985.5997	72.95055817
Residual	85	337883.3088	3975.097751	
Total	87	917854.5083		
	Coefficients	Standard Error	t Stat	P-value
Intercept	-19.28550028	31.0475285	-0.621160563	0.536156232
area(m2)	1.38360615	0.148943494	9.289470184	1.40049E-14
quartos	15.12133684	9.488597692	1.593632413	0.114730383

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Relevância prática de um regressor

$H_0: \beta_j = 0 \quad contra\ H_1: \beta_j \neq 0$

Para além da significância estatística é necessário analisar a relevância prática do regressor, i. é, se $\boldsymbol{x_j}$ tem um impacto efectivo sobre y

Um valor de $|t_{j,obs}|$ elevado que leve à rejeição de H_0 pode derivar de:

- Um valor elevado de $|\beta_j|$ combinado com um valor razoável de $\hat{\sigma}_{\hat{\beta}_j}$
- Um valor muito pequeno de $\hat{\sigma}_{\hat{\beta}_j}$ sem que $|\beta_j|$ assuma um valor relevante em termos práticos

Modelo de Regressão Linear Múltipla

Relevância prática de um regressor: exemplo

Exemplo Wooldridge – Participation Rates in 401 (k) Plans

A ideia é explicar a taxa de participação *prate* (percentagem de trabalhadores elegíveis que aderem ao plano) no plano de pensões 401 (k) em função do peso da contribuição patronal *mrate* (montante que a empresa mete no fundo por cada dólar de contribuição do trabalhador), da idade do plano *age,* e do tamanho da empresa *totemp* (nº de trabalhadores). Dados na página da UC.

O modelo estimado foi

$$\widehat{prate} = 80.29 + 5.44\, mrate + 0.269\, age - 0.0001297\, totemp$$
$$(0.78) \quad (0.52) \qquad (0.045) \qquad (0.0000367)$$
$$n = 1534 \quad R^2 = 0.100 \quad \overline{prate} = 87.4 \quad \overline{totemp} = 3567.3$$

Como se pode verificar, rejeita-se claramente $H_0: \hat{\beta}_3 = 0$ $(t_{3,obs} = -3.53)$ logo $totemp$ é estatisticamente significante relevante $(p - value = 0.00042)$ mas de pouca relevância prática: Para um crescimento de 10000 trabalhadores na empresa a tx de participação diminuiria de 1.297pp!

Não olhar tanto para o valor do coeficiente mas sim este valor multiplicado pelos valores "razoáveis" da variável x!

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste ao sinal de β_j

Exemplo:

Pretende-se testar se a área tem um impacto positivo sobre o preço

$$H_0\colon \beta_1 \geq 0 \quad contra \quad H_1\colon \beta_1 < 0 \quad \left(10000\ \$ \text{ por } 10m^2 \Leftrightarrow 1000\ \$/m^2\right)$$

$$\text{Estatística teste} \rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t\left(\overbrace{88-2-1}^{85}\right)$$

$$t_{obs} = \frac{1.384 - 0}{0.1489} = 9.295 \Rightarrow P(T > 9.295) = 1$$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste para um valor particular de β_j

Exemplo:

Pretende-se testar se um acréscimo de $10m^2$ na área tem um impacto médio esperado superior a $10000\ US\$$.

$$H_0\colon \beta_1 \leq 1 \quad contra \quad H_1\colon \beta_1 > 1 \quad \left(10000\ \$ \text{ por } 10m^2 \Leftrightarrow 1000\ \$/m^2\right)$$

$$\text{Estatística teste} \rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t\left(\overbrace{88-2-1}^{85}\right)$$

$$t_{obs} = \frac{1.384 - 1}{0.1489} = 2.579 \Rightarrow P(T > 2.579) = 0.0058$$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Inferência estatística sobre a variância das variáveis residuais

Estatística Teste: $Q = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sigma^2} = \frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k-1)}$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de uma combinação linear dos coeficientes de regressão

$$H_0: \delta = \beta_0 c_0 + \beta_1 c_1 + \cdots + \beta_k$$

Observações:

1 – Os c_j são os valores pré-fixadosem função do que se pretende testar

2 – Exemplo: se no modelo $preço = \beta_0 + \beta_1 área + \beta_2 quartos + u$
se quisesse testar se o impacto de 10 m^2 de área equivale ao impacto de 1 quarto ter-se-ia: $\delta = 10\beta_1 - \beta_2$ e testar-se-ia : $\delta = 0$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de uma combinação linear dos coeficientes de regressão

A estimação de δ não levanta problema: $\delta = \hat{\beta}_0 c_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k$

Cálculo da variância de $\hat{\delta}$: Aplicam-se as propriedades da variância da soma sem esquecer as covariâncias entre os $\hat{\beta}_j$

Em termos matriciais:

$$var(\hat{\delta}) = \sigma^2_{\hat{\delta}} = var(c\hat{\beta}) = cvar(\hat{\beta})c^T = \sigma^2 c(X^T X)^{-1} c^T$$

$$\hat{\delta} \sim N(\delta, \sigma^2_{\hat{\delta}})$$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de uma combinação linear dos coeficientes de regressão

Como σ^2 é desconhecido→ $\sigma^2_{\hat{\delta}}$ também é desconhecido pelo que a distribuição por amostragem de $\hat{\delta}$ é dada por:

$$T = \frac{\hat{\delta} - \delta}{\hat{\sigma}_{\hat{\delta}}} \sim t(n-k-1)$$

Obtenção de $\hat{\sigma}_{\hat{\delta}}$: difícil porque muitos softwares não reportam a matriz estimada das var/cov de $\hat{\beta}$ e obriga a uma conta adicional

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de uma combinação linear dos coeficientes de regressão

Retorne-se ao exemplo: $\widehat{preço} = -19.286 + 1.3836\ area + 15.121\ quartos$
(31.046) (0.1489) (9.4886)

Testar se o impacto de 10 m^2 de área equivale ao impacto de 1 quarto.

$\delta = 10\beta_1 - \beta_2 \quad H_0: \delta = 0 \quad$ contra $\quad H_1: \delta \neq 0$

Estatística teste: $T = \frac{\hat{\delta}}{\hat{\sigma}_{\hat{\delta}}} \sim t(88 - 2 - 1)$

$\hat{\delta} = 10\hat{\beta}_1 - \hat{\beta}_2 = 10 * 1.836 - 15.121 = -1.285$

$\hat{\sigma}^2_{\hat{\delta}} = 10^2\sigma^2_{\widehat{\beta_1}} + \sigma^2_{\hat{\beta}_2} - 2 * 10 * \hat{\sigma}_{\hat{\beta}_1+\hat{\beta}_2}$ → Matriz estimada das var/cov

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de uma combinação linear dos coeficientes de regressão

Recordando a matriz apresentada anteriormente:

	(Intercept)	area	quartos
(Intercept)	963.949026	-1.46733923	-180.5499241
area	-1.467339	0.02218416	-0.7520268
quartos	-180.549924	-0.75202680	90.0334862

$$t_{\hat{\delta},obs} = \frac{\hat{\delta}}{\hat{\sigma}_{\hat{\delta}}} = \frac{-1.285}{\sqrt{107.29}} = 0.124$$

Valor-p= 0.9015

$$\hat{\sigma}^2_{\hat{\delta}} = 10^2\hat{\sigma}^2_{\hat{\beta}_1} + \hat{\sigma}^2_{\hat{\beta}_2} - 2 * 10 * \hat{\sigma}_{\hat{\beta}_1+\hat{\beta}_2}$$

Não se rejeita H_0

$$\sigma^2_{\hat{\delta}} = 10^2 0.02218 + 90.0335 - 2 * 10 * (-0.75203) = 107.29$$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de uma combinação linear dos coeficientes de regressão

Metodologia alternativa

Parte-se de $\delta = 10\beta_1 - \beta_2 \rightarrow \beta_2 = 10\beta_1 - \delta$

Substitui-se β_2 no modelo inicial por $10\beta_1 - \delta$

$$preço = \beta_0 + \beta_1 área + (10\beta_1 - \delta)quartos + u$$

$$preço = \beta_0 + \beta_1 área + 10\beta_1 quartos - \delta quartos + u$$

$$preço = \beta_0 + \beta_1 (área + 10quartos) - \delta quartos + u$$

$preço = \beta_0 + \beta_1 x^* + (-\delta)quartos + u$ e estima-se este novo modelo

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de uma combinação linear dos coeficientes de regressão

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.794906945					
R Square	0.63187705					
Adjusted R Square	0.623215334					
Standard Error	63.04837628					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	579971.1994	289985.5997	72.95055817	3.58672E-19	
Residual	85	337883.3088	3975.097751			
Total	87	917854.5083				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-19.28550028	31.0475285	-0.621160563	0.536156232	-81.0163048	42.44530424
x*	1.38360615	0.148943494	9.289470184	1.40049E-14	1.08746658	1.67974572
quartos	-1.285275338	10.35820635	-0.124082809	0.901542671	-21.8801646	19.30961392

Note-se que o valor $\hat{\beta}_1$ na linha de x^* não se alterou em relação ao modelo inicial

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Exemplo: Considere o modelo (ver o tópico relevância prática de um regressor)

$$prate = \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 totemp + u$$

a) Escreva a regressão auxiliar para fazer inferência sobre $\delta = 4\beta_1 - 10\beta_2$

b) Mesma questão para $\theta = \frac{\beta_2}{2} + 1000\beta_3$

c) Obter um IC a 95% para δ

d) Testar $H_0: \delta = 0$ contra $H_1: \delta \neq 0$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

$$prate = \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 totemp + u$$

a) Escreva a regressão auxiliar para fazer inferência sobre $\delta = 4\beta_1 - 10\beta_2$

Sol 1:

$\delta = 4\beta_1 - 10\beta_2$ logo $\beta_1 = 0.25\,\delta + 2.5\,\beta_2$

$prate = \beta_0 + (0.25\,\delta + 2.5\,\beta_2)mrate + \beta_2 age + \beta_3 totemp + u$

$prate = \beta_0 + \delta\,(0.25 \times mrate) + \beta_2(age + 2.5\,mrate) + \beta_3 totemp + u$

Sol 2:

$\delta = 4\beta_1 - 10\beta_2$ logo $\beta_2 = 0.4\,\beta_1 - 0.1\,\delta$

$prate = \beta_0 + \beta_1 mrate + (0.4\,\beta_1 - 0.1\,\delta\,)age + \beta_3 totemp + u$

$prate = \beta_0 + \beta_1\,(mrate + 0.4\,age) + \delta(-0.1 \times age) + \beta_3 totemp + u$

Desafio: Estime as 2 regressões auxiliares e verifique que as linhas referentes a δ são idênticas.

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

$$prate = \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 totemp + u$$

b) Mesma questão para $\theta = \frac{\beta_2}{2} + 1000\beta_3$

Sol 1:

$\theta = \frac{\beta_2}{2} + 1000\beta_3$ logo $\beta_2 = 2\,\theta - 2000\,\beta_3$

$prate = \beta_0 + \beta_1 mrate + (2\,\theta - 2000\,\beta_3) age + \beta_3 totemp + u$

$prate = \beta_0 + \beta_1 mrate + \theta\,(2 \times age) + \beta_3 (totemp - 2000 \times age) + u$

Sol 2:

$\theta = \frac{\beta_2}{2} + 1000\beta_3$ logo $\beta_3 = 0.001\,\theta - 0.0005\,\beta_2$

$prate = \beta_0 + \beta_1 mrate + \beta_2\, age + (0.001\,\theta - 0.0005\,\beta_2)\, totemp + u$

$prate = \beta_0 + \beta_1 mrate + \beta_2\,(age - 0.0005\, totemp) + \theta\,(0.001 \times totemp) + u$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

$$prate = \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 totemp + u$$

c) Obter um IC a 95% para δ - sol 1

$$prate = \beta_0 + \delta\,(0.25 \times mrate) + \beta_2(age + 2.5\,mrate) + \beta_3 totemp + u$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.315592361					
R Square	0.099598538					
Adjusted R Square	0.097833045					
Standard Error	15.87778005					
Observations	1534					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	42666.57348	14222.19116	56.41400708	1.37991E-34	
Residual	1530	385718.966	252.1038993			
Total	1533	428385.5394				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	80.29428539	0.777695208	103.2464706	0	78.76882403	81.81974675
0.25mrate	19.07165742	2.199785126	8.669781966	1.08405E-17	14.75674437	23.38657047
age+2.5mrate	0.269407284	0.04514857	5.967127681	2.99497E-09	0.180847655	0.357966913
totemp	-0.000129781	3.67184E-05	-3.534503198	0.000420703	-0.000201805	-5.77576E-05

d) $H_0: \delta = 0$ contra $H_1: \delta \neq 0$ $\quad t_{obs} = 8.67$ $\quad p - value \approx 0$ Rejeita-se H_0

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

$$prate = \beta_0 + \beta_1 mrate + \beta_2 age + \beta_3 totemp + u$$

c) Obter um IC a 95% para δ - sol 2

$$prate = \beta_0 + \beta_1 (mrate + 0.4\, age) + \delta(-0.1 \times age) + \beta_3 totemp + u$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.315592361					
R Square	0.099598538					
Adjusted R Square	0.097833045					
Standard Error	15.87778005					
Observations	1534					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	42666.57348	14222.19116	56.41400708	1.37991E-34	
Residual	1530	385718.966	252.1038993			
Total	1533	428385.5394				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	80.29428539	0.777695208	103.2464706	0	78.76882403	81.81974675
mrate+0.4*age	5.441432564	0.524408596	10.37632222	2.00817E-24	4.412796873	6.470068255
-0.1*age	19.07165742	2.199785126	8.669781966	1.08405E-17	14.75674437	23.38657047
totemp	-0.000129781	3.67184E-05	-3.534503198	0.000420703	-0.000201805	-5.77576E-05

d) $H_0: \delta = 0$ contra $H_1: \delta \neq 0$ $\quad t_{obs} = 8.67$ $\quad p - value \approx 0$ Rejeita-se H_0

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de q combinações lineares dos coeficientes

Embora se possa continuar a considerar as 2 vias de abordagem anteriores (“matricial” ou regressão auxiliar por reparametrização da regressão original) apenas se irá desenvolver a segunda por ser bastante mais eficiente.

Apresentaremos o problema do caso mais simples para o mais geral (complicado) e utilizaremos um novo exemplo com base numa regressão envolvendo mais parâmetros

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à significância conjunta de q coeficientes

Exemplo: Vamos modelar o logaritmo do salário dos jogadores profissionais de baseball (dados de 1993 no ficheiro MLB1_simplificado.xlsx) em função de 5 características:

- $years$ – nº de anos de carreira do jogador
- $gamesyr$ – nº médio de jogos que o jogador joga por ano
- $bavg$ – nº médio de batting/ano
- $hrunsyr$ – nº médio de *home runs* por ano
- $rbisyr$ – nº médio de "runs" por ano ("runs batted in per year")

Modelo

$$\ln salary = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à significância conjunta de q coeficientes

Modelo estimado

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.792340118					
R Square	0.627802862					
Adjusted R Square	0.622439791					
Standard Error	0.726577259					
Observations	353					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	5	308.9892282	61.79784565	117.0603271	2.93802E-72	
Residual	347	183.1863363	0.527914514			
Total	352	492.1755645				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.19242075	0.288822864	38.75185162	4.1871E-128	10.62435701	11.76048449
years	0.068862623	0.012114544	5.684293499	2.7876E-08	0.045035448	0.092689799
gamesyr	0.012552112	0.002646763	4.742438528	3.08865E-06	0.007346394	0.017757829
bavg	0.000978594	0.001103509	0.886802204	0.375799595	-0.001191814	0.003149002
hrunsyr	0.014429519	0.01605698	0.898644642	0.369465108	-0.017151734	0.046010772
rbisyr	0.01076574	0.007174962	1.50045958	0.134404707	-0.003346147	0.024877626

Olhando para o output vê-se que os 3 últimos coeficientes não são, individualmente considerados, estatisticamente significativos. Será que os podemos eliminar?

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à significância conjunta de q coeficientes

$$H_0: \beta_{q+1} = 0, \beta_{q+2} = 0, \cdots, \beta_k \quad contra \quad H_1: \exists\, \beta_j \neq 0 \quad (j = q+1, \cdots, k)$$

Estratégia de teste:

- Estimar o modelo incorporando a restrição definida em H_0, isto é, estimar o modelo sem considerar as variáveis $x_{p+1}, x_{p+2}, \ldots, x_k$.
- Comparar a qualidade do modelo com restrição com a qualidade do modelo inicial (sem a restrição):
  - Se a qualidade for semelhante, pode concluir-se que as *r* variáveis associadas com os coeficientes em teste não acrescentam nada de significativo ao modelo e poderão ser consideradas **conjuntamente não significantes** e, eventualmente, eliminadas.
  - Se a introdução da restrição originar uma quebra de qualidade significativa, a conclusão terá de ser a oposta. As variáveis $x_{p+1}, x_{p+2}, \ldots, x_k$ são conjuntamente significantes.

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à significância conjunta de q coeficientes

Para tal concebe-se um teste em 3 passos:

1 – estimar o modelo sem restricções , i.é, com todos os regressores e obter:

$$SSR_{nr} = \sum_{i=1}^{n} \hat{u}_i^2$$

2 – estimar o modelo com restricções, i.é, eliminando os regressores considerados nulos e obter: $SSR_r = \sum_{i=1}^{n} \hat{u}_i^2$

3 – Comparar os modelos utilizando a estatística teste:

$$F = \frac{(SSR_r - SSR_{nr})/\left(\overbrace{k-p}^{m}\right)}{SSR_{nr}/(n-k-1)} \sim F_{(m,n-k-1)}$$

A **região de rejeição** situa-se na **aba direita** da distribuição F.

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à significância conjunta de q coeficientes

$$H_0: \beta_{q+1} = 0, \beta_{q+2} = 0, \cdots, \beta_k \quad contra \quad H_1: \exists\, \beta_j \neq 0 \quad (j = q+1, \cdots, k)$$

Não rejeitar a hipótese nula **significa que os q regressores não tem qualquer impacto sobre a variável dependente pelo que se podem retirar do modelo**

Nota: Como neste caso a variável y é a mesma nos 2 modelos, pode mostrar-se que $\frac{(SSR_* - SSR)/q}{SSR/(n-k-1)} = \frac{(R^2 - R_*^2)/q}{(1-R^2)/(n-k-1)}$ e portanto, de forma equivalente, tem-se

$$F = \frac{(R^2 - R_*^2)/q}{(1-R^2)/(n-k-1)} \sim F(q, n-k-1)$$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à significância conjunta de q coeficientes

Exemplo: $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ *vs* H_1: H_0 falsa (pelo menos um dos $\beta \neq 0$)

Modelo com a restrição

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.772704072					
R Square	0.597071582					
Adjusted R Square	0.594769134					
Standard Error	0.752731258					
Observations	353					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	293.8640431	146.9320216	259.3203218	8.2202E-70	
Residual	350	198.3115214	0.566604347			
Total	352	492.1755645				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.22380399	0.108312013	103.6247386	9.966E-265	11.01077971	11.43682826
years	0.071317962	0.012505011	5.703150453	2.50422E-08	0.046723543	0.095912381
gamesyr	0.02017448	0.00134287	15.02340897	1.01913E-39	0.017533371	0.022815589

$$F_{obs} = \frac{(198.311-183.186)/3}{183.186/(353-1\quad)} = 9.55 \qquad p - value = P(F_{3;347} \geq 9.55) \approx 0 \quad (4.47E-06)$$

Ou $F_{obs} = \frac{(0.627803-0.597072)/3}{(1-0.627803)/(353-1-5)} = 9.55$ $\qquad F_{0.05} = 2.6306$

Rejeita-se H_0 logo as avariáveis são conjuntamente significantes -> Multicolinearidade

Modelo de Regressão Linear Múltipla

Teste à significância conjunta de todos os regressores

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad contra \quad H_1: \exists\, \beta_j \neq 0 \quad (j = 1, \cdots, k)$$

A não rejeição da hipótese H_0 leva a que o modelo deva ser posto de parte.

Não rejeitar a hipótese nula **corresponde a verificar que o modelo proposto não é adequado, na sua globalidade**, para descrever o comportamento do regressando.

Utiliza-se a estratégia de teste anterior mas o modelo com restricções não precisa de ser estimado já que não incluindo qualquer variável se tem: $R_r^2 = 0, SSR_r = SST$ porque $SSE_r = 0$, pelo que :

$$\text{Estatística teste: } F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{SSE/k}{SSR/(n-k-1)} \sim F_{(k,n-k-1)}$$

A **região de rejeição** situa-se na **aba direita** da distribuição F.

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Comentários:

- Se o teste individual de cada um dos coeficientes incluídos em H_0 não rejeita a nulidade e o teste conjunto a rejeita, **desconfiar de uma possível multicolinearidade**

- A situação inversa, não se rejeita a nulidade conjunta de todos os regressores, com o teste F, mas rejeita-se a nulidade para um particular coeficiente pelo teste ***t,*** também é possível, mas neste caso é geralmente preferível confiar no teste ***t***.

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de várias combinações lineares dos coeficientes – caso geral

A metodologia de teste que se apresentou pode ser estendida para qualquer sistema de restrições lineares.

Para tal:

- Definir o sistema de restrições que se quer testar
- Resolver o sistema em ordem a um sub-conjunto de β_j
- Definir a regressão auxiliar e estimá-la
- Calcular F_{obs} e realizar o teste (utilizando o *valor-p* ou valor crítico)

Nota importante: Caso a variável dependente da regressão auxiliar seja diferente da variável dependente da regressão original não se pode utilizar o resultado baseado no R^2.

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de várias combinações lineares dos coeficientes – caso geral

Exemplo: Considere-se o exemplo anterior e admita-se que se quer testar

$$H_0: \begin{cases} \beta_1 = 0.1 \\ \beta_4 - \beta_5 = 0 \end{cases} \text{ vs } H_1: \ H_0 \text{ falsa}$$

- Resolver o sistema

$$\begin{cases} \beta_1 = 0.1 \\ \beta_4 - \beta_5 = 0 \end{cases} \Leftrightarrow \begin{cases} \beta_1 = 0.1 \\ \beta_4 = \beta_5 \end{cases}$$

- Regressão auxiliar

$$\ln salary = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$
$$= \beta_0 + 0.1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_5 hrunsyr + \beta_5 rbisyr + u$$
$$\ln salary - 0.1\ years = \beta_0 + \beta_2 gamesyr + \beta_3 bavg + \beta_5 (hrunsyr + rbisyr) + u$$

(ver output no slide seguinte)

- $F_{obs} = \frac{(186.674 - 1 \quad .186)/2}{183.186/(353-5 \quad)} = 3.3037 \qquad F_{0.05}(2; 347) = 3.022$

$p - value = P(F \geq 3.3037) = 0.0379$ Rejeita H_0 para $\alpha = 0.05$

Modelo de Regressão Linear Múltipla

Inferência estatística - O modelo de regressão linear

Teste à existência de várias combinações lineares dos coeficientes – caso geral

Exemplo: output da regressão auxiliar

$$\ln salary - 0.1\ years = \beta_0 + \beta_2 gamesyr + \beta_3 bavg + \beta_5 (hrunsyr + rbisyr) + u$$

SUMMARY OUTPUT		ln(salary)-0.1*years				
Regression Statistics						
Multiple R	0.691355747					
R Square	0.477972769					
Adjusted R Square	0.473485429					
Standard Error	0.731357498					
Observations	353					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	170.9207773	56.97359242	106.5158331	5.65962E-49	
Residual	349	186.6744424	0.534883789			
Total	352	357.5952197				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.19091343	0.272181054	41.11569581	7.6181E-136	10.65559194	11.72623493
gamesyr	0.010532168	0.001858351	5.667481593	3.03747E-08	0.006877193	0.014187143
bavg	0.000878062	0.001074004	0.817559631	0.414166471	-0.001234272	0.002990396
hrunsyr+rbisyr	0.011681257	0.002298988	5.081043186	6.12886E-07	0.007159643	0.016202871

Modelo de Regressão Linear Múltipla

Propriedades assimptóticas do MRLM

Para fazer inferência sobre o MRLM, foi necessário introduzir a hipótese MRL 6 referente à distribuição normal.

Na 1ª parte da UC, em Estatística, quando se passou de populações normais para "grandes amostras" recorreu-se ao Teorema do Limite Central.

Agora, no quadro do MRLM vamos fazer algo semelhante e levantar a hipótese MRL 6 no quadro das grandes amostras.

Aproveitaremos o tratamento de grandes amostras para apresentar algumas propriedades assintóticas dos estimadores OLS, nomeadamente a consistência.

Em termos da UC de Estatística 2 apenas cobriremos uma versão "light" do capítulo 5.

Modelo de Regressão Linear Múltipla

Propriedades assimptóticas do MRLM

Consistência do estimador OLS

Assumindo as hipóteses MRL 1 a MRL 4, mostra-se que $\hat{\beta}$, estimador OLS de β **é consistente**, isto é, tende em probabilidade para o verdadeiro valor β:

$$\forall \varepsilon > 0, \lim_{n \to \infty} Pr\left(\left|\hat{\beta}_j - \beta_j\right| < \varepsilon\right) = 1 \quad \text{para } j = 0,1,\ldots,k$$

o que se resume escrevendo

$$plim(\hat{\beta}) = \beta$$

Demonstração no Apêndice E do Wooldridge.

Tecnicamente, tem de se garantir que $var(x_j) < \infty$.

Modelo de Regressão Linear Múltipla

Propriedades assimptóticas do MRLM

Consistência do estimador OLS

A existência de correlação entre o termo de erro, u, e qualquer das variáveis x_j incluídas no modelo origina a inconsistência de todos os $\hat{\beta}_j$ para além do enviesamento do estimador.

Este ponto é de grande importância já que mostra que o enviesamento não irá desaparecer com o aumento da amostra.

Não se verificando a hipótese MRL 4 em relação a uma qualquer variável $\boldsymbol{x_j}$ diz-se que a variável é endógena. Que existe endogeneidade no modelo (por oposição a exogeneidade quando a hipótese se verifica)

Modelo de Regressão Linear Múltipla

Propriedades assimptóticas do MRLM

Consistência do estimador OLS Omissão de var.(s) relevantes

A omissão de uma variável relevante no modelo – violação de MRL 1 – (o seu efeito passará a estar incluído em u), se correlacionada com alguma das variáveis incluídas, originará assim a inconsistência do estimador OLS o que é bastante mais grave do que a inclusão de variáveis irrelevantes (que só originam a perda de eficiência do estimador). Esta é uma razão para manter variáveis de controle que não se mostram estatisticamente significantes.

Modelo de Regressão Linear Múltipla

Variáveis explicativas omitidas e Variáveis explicativas irrelevantes

Quando da escolha dos regressores a incluir no modelo deve ter-se em conta que:

. Regressores em excesso deram redução de eficiência

. Omissão de regressores (tendo em conta que o seu impacto estará incluido na var. residual) gera:

- Inconsistência se $E(U|X) \neq 0 \rightarrow$ endogeneidade
- Consistência se $E(U|X) = 0 \rightarrow$ exogeneidade

Modelo de Regressão Linear Múltipla

Teste assintótico para um β_j

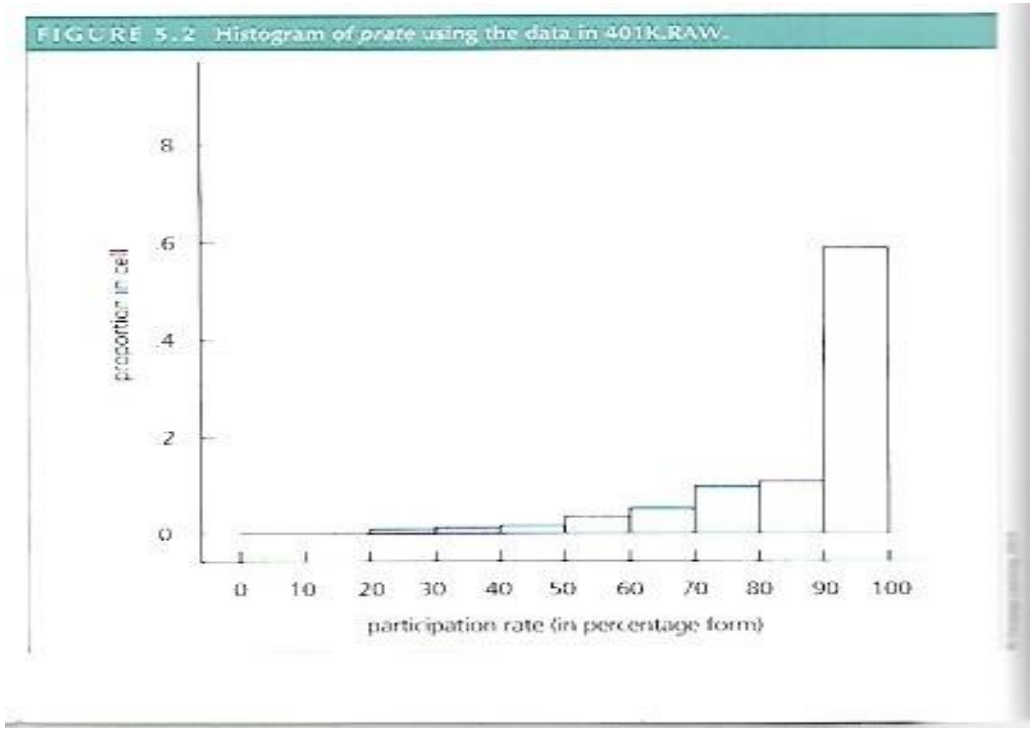
A normalidade dos estimadores OLS deriva da normalidade da distribuição dos resíduos na população. Se os resíduos seguem uma distribuição não normal → $\hat{\beta}_j$ não tem distribuição normal pelo que a estatística t não tem distribuição t-Student e a estatística F não tem distribuição F-Snédecor.

Exemplo 4.6:

Percentagens de participação em planos de pensões - *prate*

A distribuição de frequências de *Prate* está longe de ser normal.

O que fazer?



Modelo de Regressão Linear Múltipla

Teste assintótico para um β_j - TLC (grandes amostras)

Estatística teste: $\rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \dot{\sim} N(0.1)$

Também se pode utilizar:

Estatística teste: $\rightarrow T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \dot{\sim} t(n-k-1)$

Observação: Esta abordagem também se aplica, sem dificuldade quando se quer fazer inferência sobre uma combinação linear de β_j.

Estatística teste: $\rightarrow T_j = \frac{\hat{\delta} - \delta}{\hat{\sigma}_{\hat{\delta}}} \dot{\sim} N(0.1)$ ou $T_j = \frac{\hat{\delta} - \delta}{\hat{\sigma}_{\hat{\delta}}} \dot{\sim} t(n-k-1)$

Modelo de Regressão Linear Múltipla

Testes assintóticos para um conjunto de coeficientes

- Pode utilizar-se o teste F, apesar da distribuição ser apenas aproximada (neste caso a aproximação é mais lenta do que com a *t-Student*)

. Pode aplicar-se o teste **LM** (**L**agrange **M**ultiplier) que apenas envolve o modelo restrito supondo q restrições.

- Estimar a regressão assumindo H_0 verdadeira e guardar os resíduos desta regressão que se designarão por $\tilde{u}$ (para os diferenciar dos resíduos do modelo sem restrição)
- Estimar a regressão de $\tilde{u}$ em **todas** as variáveis x_j e obter o coeficiente de determinação desta regressão, R_u^2.
- A estatística de teste será $LM = nR_u^2 \overset{a}{\rightarrow} X_q^2$
- **Região de rejeição:** aba direita da distribuição *Qui-quadrado*

Modelo de Regressão Linear Múltipla

Teste Lagrange Multiplier - exemplo

Retoma-se o exemplo dos salários dos jogadores de baseball que se usou pra o "teste à significância de q coef.(s)" e vai testar-se:

$H_0: \beta_3 = \beta_4 = \beta_5 = 0 \; vs \;\; H_1: \exists \beta_j \neq 0 \; j = 3,4,5$

1. Estimação do modelo restrito:

2. Estima-se a regressão dos resíduos deste modelo em todas as variáveis x_j

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.772704072					
R Square	0.597071582					
Adjusted R Square	0.594769134					
Standard Error	0.752731258					
Observations	353					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	293.8640431	146.9320216	259.3203218	8.2202E-70	
Residual	350	198.3115214	0.566604347			
Total	352	492.1755645				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.22380399	0.108312013	103.6247386	9.966E-265	11.01077971	11.43682826
years	0.071317962	0.012505011	5.703150453	2.50422E-08	0.046723543	0.095912381
gamesyr	0.02017448	0.00134287	15.02340897	1.01913E-39	0.017533371	0.022815589

Modelo de Regressão Linear Múltipla

Teste Lagrange Multiplier – exemplo (continuação)

2. Estima-se a regressão dos resíduos deste modelo em todas as variáveis x_j

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.276169921					
R Square	0.076269825					
Adjusted R Square	0.062959592					
Standard Error	0.726577259					
Observations	353					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	5	15.12518511	3.025037023	5.730164532	4.26556E-05	
Residual	347	183.1863363	0.527914514			
Total	352	198.3115214				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.031383232	0.288822864	-0.1086591	0.913535688	-0.599446973	0.536680508
years	-0.002455338	0.012114544	-0.202676916	0.839506351	-0.026282514	0.021371837
gamesyr	-0.007622368	0.002646763	-2.879882894	0.004225253	-0.012828086	-0.002416651
bavg	0.000978594	0.001103509	0.886802204	0.375799595	-0.001191814	0.003149002
hrunsyr	0.014429519	0.01605698	0.898644642	0.369465108	-0.017151734	0.046010772
rbisyr	0.01076574	0.007174962	1.50045958	0.134404707	-0.003346147	0.024877626

$$LM_{obs} = 353 * 0.0762 \ldots = 26.92 \qquad q_{0.05} = 7.814\ (3g.l.)$$

$$Valor - p = P(LM \geq 26.92) = 6.11E^{-06} \Rightarrow rejeita - se\ H_0$$

Modelo de Regressão Linear Múltipla

Teste Lagrange Multiplier para os k declives

Neste caso o teste é quase imediato (alguns softwares reportam mesmo o seu resultado)

$$LM = nR^2 \sim X_k^2$$

Em que R^2 é o coeficiente de determinação do modelo. Como é evidente, a região de rejeição mantém-se na cauda direita da distribuição *Qui-quadrado*

Modelo de Regressão Linear Múltipla

Propriedades do estimador OLS - Síntese

Hipóteses:

1. Modelo linear nos parâmetros: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
2. Amostra aleatória
3. Ausência de colinearidade perfeita
4. Exogeneidade: $E(u|\boldsymbol{x}) = 0$
5. Homoscedasticidade: $var(u|\boldsymbol{x}) = \sigma^2$
6. Normalidade do erro: $u \sim Normal(0, \sigma^2)$

Pequenas amostras	Propriedades assimptóticas
1-4: estimadores centrados	1-4: estimadores consistentes
1-5: estimadores centrados e eficientes	1-5: estimadores consistentes, eficientes e com dist aprox normal
1-6: estimadores centrados, eficientes e normalmente distribuídos	

Modelo de Regressão Linear Múltipla

Previsão

Modelo $\rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

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

Como prever, usando o modelo estimado?

Duas questões:

1. O que prever? O valor de y ou $E(y|x)$?

2. Como prever? Previsão pontual ou por intervalos

Em qualquer dos casos assume-se que os valores das variáveis explicativas são conhecidos $x_1 = c_1, \cdots, x_k = c_k$

Modelo de Regressão Linear Múltipla

Previsão em média - $E(y|x = c)$

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Previsão pontual: $\widehat{E(y|x)} = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$

Nota: $E(y|\boldsymbol{x} = \boldsymbol{c})$=$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$ pode ser considerada uma combinação linear de β_j.

Previsão por intervalos:

Prever $\rightarrow \theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$

Previsor $\rightarrow \hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$

É necessário obter $\widehat{\sigma_{\hat{\theta}_0}} = ?$

Modelo de Regressão Linear Múltipla

Previsão em média - $E(y|x = c)$

Previsão por intervalos – obtenção $\widehat{\sigma_{\hat{\theta}_0}}$: Dois métodos:

1. Recorrer à matriz estimada das variâncias- covariâncias dos $\hat{\beta}_j$

2. Utilizar a regressão auxiliar:

   - Fazer $\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - \cdots - \beta_k c_k$

   - Construir a regressão auxiliar substituindo β_0 pela expressão anterior $\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - \cdots - \beta_k c_k$

$$y = \theta_0 + \beta_1(x_1 - c_1) + \cdots + \beta_k(x_k - c_k) + u$$

   - Ao estimar-se a regressão obtém-se directamente $\hat{\theta}_0$ e $\widehat{\sigma_{\hat{\theta}_0}}$

Modelo de Regressão Linear Múltipla

Previsão em média - $E(y|x = c)$ - exemplo:

Voltemos ao exemplo do preço de um imóvel como função da área, quadrado da área e nº quartos.

Pretende-se obter uma previsão por intervalos com um grau de confiança de 95% para o valor esperado de um imóvel com 4 quartos e 220 m^2.

Modelo estimado:

$$\widehat{preço} = \underset{(85.626)}{149.92} - \underset{(0.7853)}{0.2477}\, area + \underset{(0.0017)}{0.0037}\, area^2 + \underset{(9.3055)}{14.487}\, quartos$$

Modelo auxiliar estimado

$$\widehat{preço} = \underset{(10.726)}{330.206} - \underset{(0.7853)}{0.4277}\,(area - 220) + \underset{(0.0017)}{0.0037}\,(area^2 - 220^2) + \underset{(9.3055)}{14.487}\,(quartos - 4)$$

$$I.Prev^{95\%}_{\hat{\boldsymbol{\theta}}_0} = \left(\hat{\boldsymbol{\theta}}_0 - t\alpha_{/2}\widehat{\boldsymbol{\sigma}_{\hat{\boldsymbol{\theta}}_0}}; \hat{\boldsymbol{\theta}}_0 + t\alpha_{/2}\widehat{\boldsymbol{\sigma}_{\hat{\boldsymbol{\theta}}_0}}\right)$$

Modelo de Regressão Linear Múltipla

Previsão em média - $E(y|x = c)$ - exemplo (continuação):

SUMMARY OUTPUT			preço	area	area^2	quartos	area-220	area^2-220^2	quartos-4
			300	226	51076	4	6	2676	0
Regression Statistics									
Multiple R	0.80652046								
R Square	0.650475252								
Adjusted R Square	0.637992226								
Standard Error	61.79968268								
Observations	88								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	3	597041.6428	199013.8809	52.1087768	4.01148E-19				
Residual	84	320812.8654	3819.200779						
Total	87	917854.5083							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%			
Intercept	330.2059972	10.72571732	30.78637888	1.60398E-47	308.8767321	351.5352624			
area-220	-0.247693118	0.785299393	-0.315412339	0.753230872	-1.809347083	1.313960846			
area^2-220^2	0.003653521	0.001728126	2.114151745	0.037467195	0.000216953	0.00709009			
quartos-4	14.48683014	9.305514014	1.556800636	0.123277661	-4.018204989	32.99186526			

Modelo de Regressão Linear Múltipla

Previsão pontual para um caso particular $y|x = c$

Modelo $\rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

Problema: prever $\rightarrow y^0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k + u^0$

Previsor pontual $\rightarrow \hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k \quad E(u^0) = 0$

Erro de Previsão $\rightarrow \hat{e}^0 = y^0 - \hat{y}^0 = \theta^0 + u^0 - \hat{\theta}^0 = \theta^0 - \hat{\theta}^0 + u^0$

. $\mathrm{E}(\hat{e}^0) = E\left(\theta^0 - \hat{\theta}^0 + u^0\right) = E\left(\theta^0 - \hat{\theta}^0\right) + E(u^0) = 0$

. $\mathrm{var}(\hat{e}^0) = var\left(\theta^0 - \hat{\theta}^0 + u^0\right) = var\left(\hat{\theta}^0\right) + var(u^0)$

$= var\left(\hat{\theta}^0\right) + \sigma^2$

pois $\hat{\theta}^0$ é uma combinação linear $\hat{\beta}_j$ e u^0 é não correlacionado com os u_i da amostra

Modelo de Regressão Linear Múltipla

Previsão por intervalos para um caso particular $y|x = c$

Modelo $\rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

Problema: prever $\rightarrow y^0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k + u^0$

Previsor pontual $\rightarrow \hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$

Erro de Previsão $\rightarrow \hat{e}^0 = y^0 - \hat{y}^0$, $\mathrm{E}(\hat{e}^0) = 0$, $\mathrm{var}(\hat{e}^0) = var(\hat{\theta}^0) + \sigma^2$

. Hip. MRL 6, $\hat{e}^0 \sim N\left(0, var(\hat{\theta}^0) + \sigma^2\right)$, então tem-se:

$$\frac{y^0 - \hat{y}^0}{\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \hat{\sigma}^2}} \sim t_{(n-k-1)}$$

$$I.Prev^{95\%}_{\hat{\theta}_0} = \left(\hat{y}^0 - t_{\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \hat{\sigma}^2}; \hat{y}^0 + t_{\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \hat{\sigma}^2}\right)$$

Modelo de Regressão Linear Múltipla

Previsão pontual para um caso particular $y|x = c$ - exemplo

Modelo auxiliar estimado

$$\widehat{preço} = \underset{(10.726)}{330.206} - \underset{(0.7853)}{0.4277}\,(area - 220) + \underset{(0.0017)}{0.0037}\,(area^2 - 220^2) + \underset{(9.3055)}{14.487}\,(quartos - 4)$$

$\hat{\sigma}^2 = 3819.20$, $\hat{y}^0 = 330.206$, $\hat{\sigma}^2_{\hat{\boldsymbol{\theta}}_0} = 10.725717^2$, $\sqrt{\hat{\sigma}^2_{\hat{\boldsymbol{\theta}}_0} + \hat{\sigma}^2} = 62.7235$

Como é habitual, $\hat{\sigma}^2$ é muito maior que $\hat{\sigma}^2_{\hat{\boldsymbol{\theta}}_0}$ o que tende a gerar intervalos com amplitude excessiva → fraca utilidade

$$I.Prev^{95\%}_{\hat{\boldsymbol{\theta}}_0} = (330.206 - 1.96 * 62.7235;\ 330.206 + 1.96 * 62.7235)$$

$$I.Prev^{95\%}_{\boldsymbol{y_0}} = (205.473; 454.939)$$

Modelo de Regressão Linear Múltipla

Previsão pontual para um caso particular $y|x = c$ - exemplo

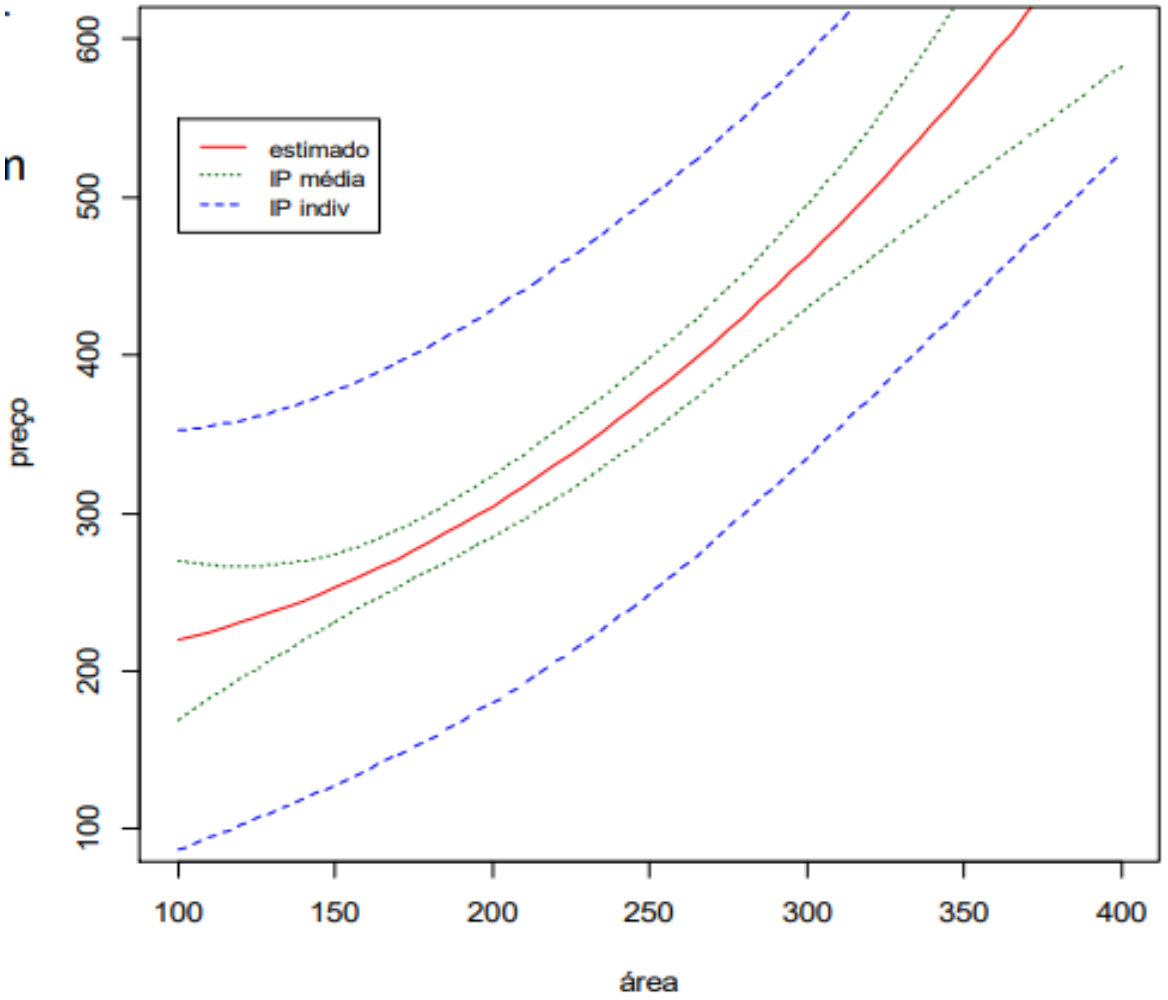
Como referido anteriormente, a amplitude do intervalo de previsão vai depender de 3 factores:

1. A previsão para um caso particular ou previsão em média

2. O nível de confiança

3. Os valores das variáveis explicativas. Quanto mais próximos das médias amostrais, menor a amplitude.

Gráfico ilustra a situação fazendo variar área com quartos=4



Modelo de Regressão Linear Múltipla

Previsão: ln(y) é a variável dependente

As previsões obtidas até agora dizem respeito a y. Quando y resulta de uma transformação as previsões em termos da variável original (aquelas que nos vão interessar) nem sempre são diretas.

Das várias transformações possíveis, apenas se vai ver o caso em que a variável de interesse foi logaritmizada, já que é a situação mais frequente.

A primeira ideia que surge é aplicar a transformação inversa (isto é a exponencial), ideia que funciona perfeitamente para os intervalos de previsão (pontual e em média) mas não para a previsão pontual.

Assim, se o intervalo de previsão para $\ln y$ for dado por $(a; b)$, o intervalo de previsão para y será dado por $\left(e^{a};\ e^{b}\right)$.

Modelo de Regressão Linear Múltipla

Previsão: ln(y) é a variável dependente

Previsão pontual: situação mais delicada pois $E(lny|\boldsymbol{x}) \neq lnE(y|\boldsymbol{x})$.
Prova-se que $E(lny|\boldsymbol{x}) < lnE(y|\boldsymbol{x})$.

Aplicando a transformação inversa tem-se: $e^{E(lny|\boldsymbol{x})} < e^{lnE(y|\boldsymbol{x})}$,
isto é, a uma sub-previsão

Como corrigir este enviesamento?

Se a hipótese de normalidade dos resíduos não levantar problemas, a solução óptima consiste em utilizar:

$$\hat{y} = \widehat{E(y|x)} = e^{\left(\frac{\hat{\sigma}^2}{2}\right)} * e^{(\widehat{lny})}$$

Previsor enviesado mas consistente

Modelo de Regressão Linear Múltipla

Previsão: ln(y) é a variável dependente

Caso se procure uma solução que não dependa tanto da normalidade dos resíduos (grandes amostras) a solução passa por estimar a constante de proporcionalidade entre $\hat{y}$ e $e^{(\widehat{lny})}$ em vez de $e^{\left(\frac{\hat{\sigma}^2}{2}\right)}$

Assim:

1. Obter $m_i = e^{(\widehat{lny})}$ para os n valores da amostra
2. Estimar o parâmetro α_0 da regressão simples sem termo independente $y_i = \alpha_0 m_i$ obtendo-se $\hat{\alpha}_0$
3. As previsões pontuais para y ou $E(y|x)$ serão dadas por $\hat{y} = \hat{\alpha}_0 e^{(\widehat{lny})}$

Previsor enviesado mas consistente

Modelo de Regressão Linear Múltipla

Previsão: ln(y) é a variável dependente - exemplo

Retomemos o exemplo anterior, considerando agora que se tinha definido como variável dependente o logaritmo do preço e não o preço. Com base neste **novo modelo** vai-se obter uma previsão pontual e um intervalo de precisão a 95% para o valor esperado de um imóvel (preço de mercado, preço esperado) com 4 quartos e 220 m2 de área.

Modelo estimado: $\widehat{\ln preço} = 1.2893 + \underset{(0.0988)}{0.8101} \ln area + \underset{(0.0303)}{0.0376}\, quartos$

Modelo auxiliar estimado (ver slide seguinte):

$$\widehat{\ln preço} = \underset{(0.0275)}{5.8090} + \underset{(0.0988)}{0.8101} (\ln area - \ln 220) + \underset{(0.0303)}{0.0376}\, (quartos - 4)$$

Fazendo a conta (ou output) vem o IP para o VE de $\ln preço$: (5.7544; 5.8637)

Aplicando a exponencial a cada extremidade obtém-se (315.588; 352.016).

Modelo de Regressão Linear Múltipla

Previsão: ln(y) é a variável dependente - exemplo

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.749479308					
R Square	0.561719234					
Adjusted R Square	0.551406745					
Standard Error	0.203324195					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	4.503642728	2.251821364	54.46980398	5.94946E-16	
Residual	85	3.513961901	0.041340728			
Total	87	8.017604629				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	5.809057324	0.027471396	211.4583978	1.8092E-117	5.754436829	5.863677818
ln(area)-ln(220)	0.810063655	0.098761107	8.202253716	2.2161E-12	0.613700116	1.006427194
quartos-4	0.037646372	0.030344583	1.240629064	0.21815627	-0.022686789	0.097979533

Modelo de Regressão Linear Múltipla

Previsão: ln(y) é a variável dependente - exemplo

Para obter uma previsão pontual para $preço$

Solução 1 – (baseada na normal)

$$\widehat{preço}^0 = \exp\left(\frac{0.20332^2}{2}\right)\exp(5.80906) = 340.266$$

Solução 2 – (mais robusta) – ver slide seguinte

- Obter $m_i = \exp(\widehat{\ln y_i})$
- $\hat{y} = \hat{\alpha}_o \exp(\widehat{lny}) = 1.0286 \exp(\widehat{lny})$
- $\widehat{preço}^0 = 1.0286 \ \exp(5.80906) = 342.842$

Modelo de Regressão Linear Múltipla

Previsão: ln(y) é a variável dependente - exemplo

preço	ln(preço)	Predicted	exp(lny)
300	5.703782475	5.830854073	340.6494945

Regressão auxiliar:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.979692285					
R Square	0.959796974					
Adjusted R Square	0.948302721					
Standard Error	62.6755834					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	8158994.73	8158994.73	2077.016192	5.22273E-62	
Residual	87	341755.9016	3928.228754			
Total	88	8500750.632				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
exp(lny)	**1.028612766**	0.022570021	45.57429311	1.67839E-62	0.983752405	1.073473128

Modelo de Regressão Linear Múltipla

Complementos sobre a forma funcional

Este tópico cobre o essencial das secções 6.1, 6.2 e 6.3 e também a secção 9.1, pontos que se podem mostrar interessantes em termos de modelação de determinado fenómeno. Destacam-se 5 pontos:

- Efeitos da alteração de escala (numa ou mais variáveis) no MRLM
- Uma segunda leitura das transformações logaritmicas
- Introdução de termos quadráticos no MRLM
- Introdução de termos de interacção no MRLM
- Teste RESET

Modelo de Regressão Linear Múltipla

- **Efeitos da alteração de escala (numa ou mais variáveis) no MRLM**

Modelo: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

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

Questão? O que acontece aos coeficientes do modelo quando se alteram as unidades em que são medidas uma ou mais variáveis?

Exemplo: Regresse-se ao exemplo referente ao preço de um imóvel onde se tinha

$$\widehat{preço} = -19.286 + 1.3836\ area + 15.121\ quartos$$

Qual o efeito nos resultados do modelo (coeficientes e demais estatísticas de interesse) de se medir a área em pés-quadrados (em vez de m2) ou o preço em dólares (em vez de milhares de dólares)?

Modelo de Regressão Linear Múltipla

- **Efeitos da alteração de escala (numa ou mais variáveis) no MRLM**

Alteração em $y \rightarrow y^* = cy$

Modelo inicial: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

Novo modelo: $y^* = cy = c(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u)$

$$= c\beta_0 + c\beta_1 x_1 + c\beta_2 x_2 + \cdots + c\beta_k x_k + cu$$

$$= \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 + \cdots + \beta_k^* x_k + u^*$$

Repare-se que $E(cu) = cE(u) = 0$ (MRL 3 mantém-se)

$\text{var}(cu) = c^2 var(u) = c^2\sigma^2$ (essência MRL 4 mantém-se)

Ir-se-á obter $\beta_j^* = c\hat{\beta}_j$ e $\hat{u}_i^* = c\hat{u}_i \rightarrow \hat{\sigma}_{\hat{\beta}_j}^* = c\hat{\sigma}_{\hat{\beta}_j}, SST^* = c^2\text{SST}, SSR^* = c^2\text{SSR}$,

mantendo-se inalterados o coef. Correlação quer os coeficientes de determinação.

Modelo de Regressão Linear Múltipla

- **Efeitos da alteração de escala (numa ou mais variáveis) no MRLM**

Exemplo: Regressem ao modelo explicação preços dos imóveis, multipliquem-nos por 1000 e voltem a estimar o modelo.

Suponham agora que $y^* = y - c$. O que aconteceria?

Regressão original (preço em 10^3 dólares, área em m^2)

$$\widehat{preço} = -19.2855 + \mathbf{1.3836}\ area + 15.1213\ quartos \qquad R^2 = 0.6319$$

ANOVA		
	df	SS
Regression	2	579971.1994
Residual	85	337883.3088
Total	87	917854.5083

	Coefficients
Intercept	-19.28550028
area(m2)	1.38360615
quartos	15.12133684

Modelo de Regressão Linear Múltipla

- **Efeitos da alteração de escala (numa ou mais variáveis) no MRLM**

Alteração 1 (preço em dólares, área em m^2)

$$\widehat{preço} = -19285.5 + 1383.6\ area + 15121.3\ quartos \quad R^2 = 0.6319$$

ANOVA

	df	*SS*
Regression	2	5.79971E+11
Residual	85	3.37883E+11
Total	87	9.17855E+11

	Coefficients
Intercept	-19285.50028
area	1383.60615
quartos	15121.33684

Modelo de Regressão Linear Múltipla

- **Efeitos da alteração de escala (numa ou mais variáveis) no MRLM**

Alteração 2 (preço em 10^3 dólares decontando 100 (mil dólares), área em m^2)

$$\widehat{preço} = -119.2855 + 1.3836\ area + 15.1213\ quartos \qquad R^2 = 0.6319$$

ANOVA

	df	*SS*
Regression	2	579971.1994
Residual	85	337883.3088
Total	87	917854.5083

	Coefficients	*S*
Intercept	-119.2855003	
area	1.38360615	
quartos	15.12133684	

Modelo de Regressão Linear Múltipla

- **Efeitos da alteração de escala (numa ou mais variáveis) no MRLM**

Alteração em $x_j \rightarrow x_j{}^* = cx_j$ e façamos sem perda de generalidade ***j***=1

Modelo inicial: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

Novo modelo: $y = \beta_0 + \beta_1 \frac{x_1{}^*}{c} + \beta_2 x_2 + \cdots + \beta_k x_k + u$

$$= \beta_0 + \frac{\beta_1}{c} x_1{}^* + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

$$= \beta_0 + \beta_1^* x_1{}^* + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Ter-se-á apenas 2 alterações: $\beta_1^* = \frac{\beta_1}{c}$; $\hat{\sigma}^*_{\hat{\beta}_1} = \frac{\hat{\sigma}_{\hat{\beta}_1}}{c}$

rácio t não se altera→ não se altera a significância estatística do regressor

Modelo de Regressão Linear Múltipla

- **Efeitos da alteração de escala (numa ou mais variáveis) no MRLM**

Exemplo: Sabendo que $m^2 = 10.764 pés^2$ multiplique x_1 por este valor e comprove que apenas $\hat{\beta}_1$ e $\hat{\sigma}_{\hat{\beta}_1}$ se alteram

SUMMARY OUTPUT

Preço em milhares de dolares
Área em pés quadrados

Regression Statistics	
Multiple R	0.794906945
R Square	0.63187705
Adjusted R Square	0.623215334
Standard Error	63.04837628
Observations	88

ANOVA

	df	*SS*	*MS*	*F*
Regression	2	579971.1994	289985.5997	72.95055817
Residual	85	337883.3088	3975.097751	
Total	87	917854.5083		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-19.28550028	31.0475285	-0.621160563	0.536156232
area(pé^2)	0.128540148	0.013837188	9.289470184	1.40049E-14
quartos	15.12133684	9.488597692	1.593632413	0.114730383

Modelo de Regressão Linear Múltipla

Complementos sobre o uso de logaritmos

No capítulo 3 interpretaram-se os parâmetros do modelo (embora de forma aproximada) no quadro dos modelos log-log e log-lin. Para recordar a situação, considere-se o modelo

$$\widehat{\ln y} = \hat{\beta}_0 + \hat{\beta}_1 \ln x_1 + \hat{\beta}_2 x_2$$

$\hat{\beta}_1$ – uma variação de 1% em x_1 origina, tudo o resto constante, uma variação percentual aproximada de $\hat{y}$ dada por $\hat{\beta}_1$

$$\%\Delta\, x_1 = 1 \rightarrow \%\Delta\hat{y} \cong \beta_1\%$$

$\hat{\beta}_2$ – uma variação unitária de x_2 origina, tudo o resto constante, uma variação percentual aproximada de $\hat{y}$ dada por $100\ \hat{\beta}_2$

$$\Delta x_2 = 1 \rightarrow \%\Delta\hat{y} \cong\ 100\ \hat{\beta}_2\%$$

Em qualquer dos casos a aproximação é valida para pequenas variações de y, nomeadamente para valores pequenos de $\hat{\beta}_j$.

Quando Δx_2 são grandes os $\%\Delta\hat{y}$ exactos $\neq s$ dos aproximados

Modelo de Regressão Linear Múltipla

Complementos sobre o uso de logaritmos

Caso da semi-elasticidade constante $\widehat{\boldsymbol{\beta}}_2$

Modelo inicial: $\widehat{lny} = \hat{\beta}_0 + \hat{\beta}_1 lnx_1 + \hat{\beta}_2 x_2$

Modelo pós-incremento: $\widehat{lny^*} = \hat{\beta}_0 + \hat{\beta}_1 lnx_1 + \hat{\beta}_2 (x_2 + \Delta x_2)$

como $\widehat{lny^*} - \widehat{lny} = \widehat{ln\left(\frac{y^*}{y}\right)} = \widehat{ln\left(\frac{y+\Delta y}{y}\right)} = ln\left(1 + \frac{\Delta\hat{y}}{\hat{y}}\right)$

então $\widehat{lny^*} - \widehat{lny} = \hat{\beta}_2 \Delta x_2 \Leftrightarrow \widehat{ln\left(1 + \frac{\Delta\hat{y}}{\hat{y}}\right)} = \hat{\beta}_2 \Delta x_2$

$1 + \frac{\Delta\hat{y}}{\hat{y}} = e^{\widehat{\beta}_2 \Delta x_2} \Leftrightarrow \frac{\Delta\hat{y}}{\hat{y}} = e^{\widehat{\beta}_2 \Delta x_2} - 1,$

isto é, $\%\Delta\hat{y} = 100\left(e^{\widehat{\beta}_2 \Delta x_2} - 1\right)$ pelo que $\Delta x_2 = 1 \rightarrow \%\Delta\hat{y} = 100\left(e^{\widehat{\beta}_2 \Delta x_2} - 1\right)$

Modelo de Regressão Linear Múltipla

Complementos sobre o uso de logaritmos

Caso da semi-elasticidade constante $\widehat{\beta}_2$- exemplo

Modelo inicial: $\widehat{ln(preço)} = 1.289 - 0.810ln(área) + 0.038quartos$

Pelo que, $\Delta x_2 = 1 \rightarrow \%\Delta\hat{y} = 100 * 0.038 = 3.8\%$ (variação aproximada)

A variação exacta será: $\Delta x_2 = 1 \rightarrow \%\Delta\hat{y} = 100\left(e^{\widehat{\beta}_2 \Delta x_2} - 1\right) = 3.873\%$

Pelo que, $\Delta x_2 = 5 \rightarrow \%\Delta\hat{y} = 100 * 5 * 0.038 = 19\%$ (variação aproximada)

A variação exacta será: $\Delta x_2 = 5 \rightarrow \%\Delta\hat{y} = 100\left(e^{\widehat{\beta}_2 \Delta x_2} - 1\right) = 20.925\%$

A diferença entre o valor aproximado e o valor exacto poderá não ser muito significativa se $\hat{\beta}_2$ e e Δx_2 tiverem valores pequenos mas será significativa no caso contrário

Modelo de Regressão Linear Múltipla

Complementos sobre o uso de logaritmos

Caso da elasticidade constante $\widehat{\beta}_1$

Modelo inicial: $\widehat{lny} = \hat{\beta}_0 + \hat{\beta}_1 lnx_1 + \hat{\beta}_2 x_2$

Modelo pós-incremento: $\widehat{lny^*} = \hat{\beta}_0 + \hat{\beta}_1 ln(x_1 + \Delta x_1) + \hat{\beta}_2 x_2$

como $\widehat{lny^*} - \widehat{lny} = ln\left(1 + \frac{\Delta\hat{y}}{\hat{y}}\right) = \hat{\beta}_1 ln\left(\frac{x_1 + \Delta x_1}{x_1}\right) = \hat{\beta}_1 ln\left(1 + \frac{\Delta x_1}{x_1}\right) = ln\left(1 + \frac{\Delta x_1}{x_1}\right)^{\widehat{\beta}_1}$

Logo, $1 + \frac{\Delta\hat{y}}{\hat{y}} = \left(1 + \frac{\Delta x_1}{x_1}\right)^{\widehat{\beta}_1} \Leftrightarrow \frac{\Delta\hat{y}}{\hat{y}} = \left(1 + \frac{\Delta x_1}{x_1}\right)^{\widehat{\beta}_1} - 1$

isto é, $\%\Delta\hat{y} = 100\left[\left(1 + \frac{\Delta x_1}{x_1}\right)^{\widehat{\beta}_1} - 1\right] \Rightarrow \%\Delta x_1 = 1 \rightarrow \%\Delta\hat{y} = 100\left[(1 + 0.01)^{\widehat{\beta}_1} - 1\right]$

Modelo de Regressão Linear Múltipla

- **Complementos sobre o uso de logaritmos**

Caso da elasticidade constante $\hat{\beta}_1$- exemplo

Modelo inicial: $\widehat{ln(preço)} = 1.289 - 0.810ln(área) + 0.038quartos$

Pelo que, $\%\Delta x_1 = 1 \rightarrow \%\Delta\hat{y} = 0.810\%$ (variação aproximada)

A variação exacta será: $\%\Delta x_1 = 1 \rightarrow \%\Delta\hat{y} = 100[(1 + 0.01)^{0.810} - 1]$
$= 0.809\%$

Pelo que, $\%\Delta x_1 = 5 \rightarrow \%\Delta\hat{y} = 0.810 * 5 = 4.05\%$ (variação aproximada)

A variação exacta será: $\%\Delta x_1 = 1 \rightarrow \%\Delta\hat{y} = 100[(1 + 0.05)^{0.810} - 1]$
$= 4.03\%$

Modelo de Regressão Linear Múltipla

- **Complementos sobre o uso de logaritmos**

Qualidade da aproximação: Depende de $\hat{\beta}_j$ e de Δx ou $\%\Delta x$

½ Elasticidade Constante Δx=1				Elasticidade Constante $\%\Delta x$=1		
$\hat{\beta}$	$\%\Delta y$ apox $100 \times \hat{\beta}$	$\%\Delta y$		$\hat{\beta}$	$\%\Delta y$ apox $\hat{\beta}$	$\%\Delta y$
-0.05	-5	-4.877		-0.05	-0.05	-0.050
0.05	5	5.127		0.05	0.05	0.050
0.1	10	10.517		0.1	0.1	0.100
0.3	30	34.986		0.3	0.3	0.299
0.5	50	64.872		0.5	0.5	0.499
0.7	70	101.375		0.7	0.7	0.699
1	100	171.828		1	1	1.000

A aproximação é bem mais robusta no modelo log-log

Modelo de Regressão Linear Múltipla

• Logaritmizar ou não as variáveis

Algumas regras práticas:

- *Logaritmizar* variáveis expressas em unidades monetárias como salários, preços, vendas, ….
- *Logaritmizar* contagens (valores inteiros elevados) como populações, nº de empregados (grandes empresas) , nº de estudantes numa universidade, ….
- *Não logaritmizar* variáveis medidas em unidades de tempo (anos, semanas,…) como educação, experiência profissional, idade ,….
- Situação menos clara para variáveis medidas em % ou para proporções como taxa de desemprego, % de sucesso num exame, …. (para estas variáveis é importante sublinhar a diferença entre variação percentual e pontos percentuais).

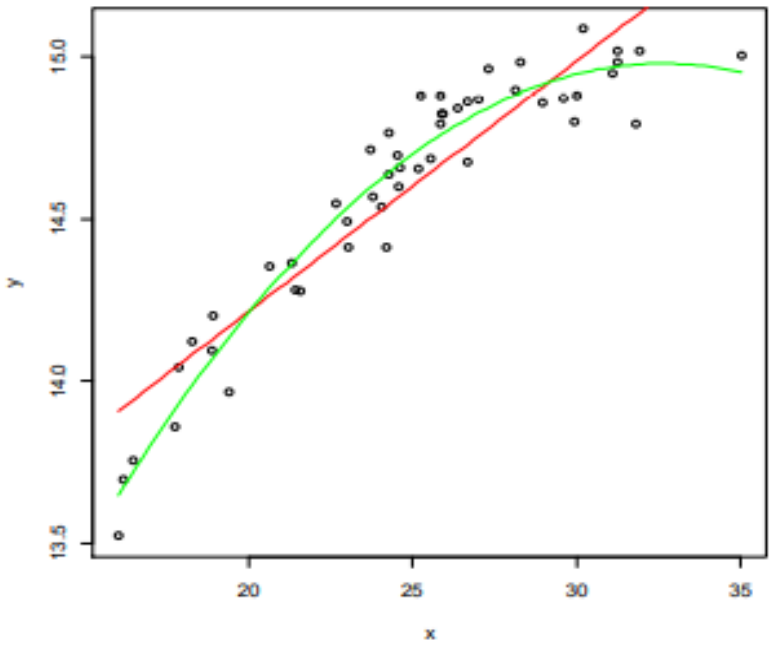
• Logaritmizar y pode reduzir uma possível assimetria ou o peso de alguns *outliers*

Modelo de Regressão Linear Múltipla

- **Termos quadráticos:**

Ideia → "suavizar" o impacto marginal da variável x_j na variável y pela introdução de um termo quadrático

Exemplo: Tendo-se observado uma amostra de 50 observações do par (y_i, x_i) , obteve-se o diagrama de dispersão que se segue



Em vermelho traçou-se o MRL ajustado: $\hat{y} = 12.688 + 0.0733\, x$ enquanto a verde se traçou o modelo com termo quadrático em x, $\hat{y} = 9.833 + 0.3157\, x - 0.0048\, x^2$

A ideia por trás do termo quadrático é flexibilizar o modelo e assim melhorar o ajustamento.

Modelo de Regressão Linear Múltipla

- **Termos quadráticos:**

MRLM: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + +\beta_3 x_3 + \cdots + \beta_k x_k + u$

Modelo estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_k x_k$

Impacto de x_1 em $\hat{y}$ será dado por: $\frac{\partial \hat{y}}{\partial x_1} = \hat{\beta}_1 + 2\hat{\beta}_2 x_1$

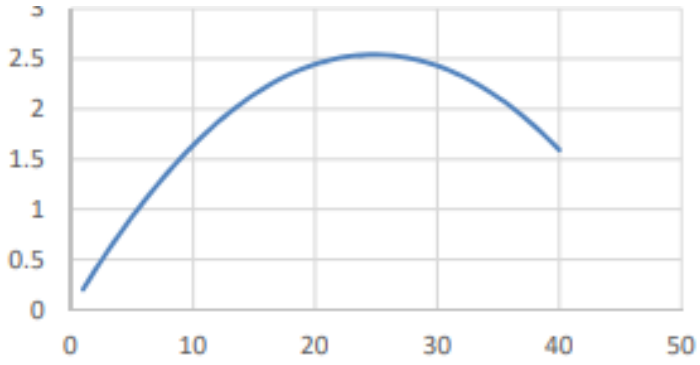
Isto é, deixa de ser constante e pode mesmo trocar de sinal

	$\hat{\beta}_2 < 0$	$\hat{\beta}_2 > 0$
$\frac{\partial \hat{y}}{\partial x_1} > 0$	$x_1 < \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$	$x_1 > \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$
$\frac{\partial \hat{y}}{\partial x_1} = 0$	$x_1 = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$	$x_1 = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$
$\frac{\partial \hat{y}}{\partial x_1} < 0$	$x_1 > \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$	$x_1 < \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$

Modelo de Regressão Linear Múltipla

- **Termos quadráticos – exemplo:**

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.631388521					
R Square	0.398651465					
Adjusted R Square	0.392869268					
Standard Error	2.877600715					
Observations	526					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	5	2854.509655	570.901931	68.94463019	3.01595E-55	
Residual	520	4305.904656	8.280585876			
Total	525	7160.41431				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-2.120092201	0.712046238	-2.977464224	0.003041951	-3.518933027	-0.721251375
educ.years	0.53009073	0.048588114	10.90988476	4.27791E-25	0.434637605	0.625543854
tenure.years	0.133669401	0.020632481	6.478590626	2.14967E-10	0.093136138	0.174202663
female	-1.790225875	0.257691666	-6.94716249	1.11859E-11	-2.296470559	-1.283981192
exper.years	0.204841793	0.034457597	5.944749778	5.0807E-09	0.137148586	0.272535001
exper^2	-0.004126562	0.000748917	-5.510039898	5.65241E-08	-0.005597837	-0.002655287



Modelo de Regressão Linear Múltipla

• Termos quadráticos – exemplo:

Usando dados do ficheiro WAGE1, estimou-se:

$$\widehat{wage} = -2.120 + 0.530\ educ + 0.134\ tenure - 1.790\ female + 0.205\ exper - 0.004\ exper^2$$
$$\qquad\qquad\quad (0.049) \qquad (0.021) \qquad\quad (0.258) \qquad\quad (0.034) \qquad\quad (0.001)$$

Do quadro anterior retira-se que o impacto de $exper$ no salário será positivo enquanto $exper < \frac{0.205}{2*0.004} = 24.82$.

A experiência profissional cresce até aos 24.82 anos experiência e decresce a partir daí.

Vale a pena considerar aqui o possível efeito da variável idade (omitida no modelo). A sua omissão conjugada com a correlação com $exper$ pode levantar problemas de endogeneidade.

Modelo de Regressão Linear Múltipla

Interacção entre variáveis explicativas

A introdução de um termo de interação é feita quando o efeito parcial de uma variável explicativa depende de outra variável explicativa. Supondo, sem perda de generalidade que se assume que o efeito parcial de x_1 depende de x_2,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_4 + \cdots + \hat{\beta}_k x_k$$

isto é constroi-se uma variável x_3 em que cada observação é o produto das observações de x_1 e de x_2 ou seja $x_{i3} = x_{i1} \times x_{i2}$, $i = 1,2,\ldots,n$. A estimação e a inferência são feitas nos termos habituais.

Impacto de x_1 em $\hat{y}$ será dado por: $\frac{\partial \hat{y}}{\partial x_1} = \hat{\beta}_1 + \hat{\beta}_3 x_2$

O impacto parcial de x_1 em $\hat{y}$ depende de x_2 (e inversamente)

Modelo de Regressão Linear Múltipla

Interacção entre variáveis explicativas – exemplo:

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.642556997					
R Square	0.412879494					
Adjusted R Square	0.406091974					
Standard Error	2.846092569					
Observations	526					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	6	2956.388241	492.7313735	60.82920957	5.66707E-57	
Residual	519	4204.02607	8.100242909			
Total	525	7160.41431				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.451983047	0.84688127	-0.533702967	0.59377555	-2.115719691	1.211753598
educ.years	0.406363274	0.059384744	6.84289003	2.19453E-11	0.289699252	0.523027296
tenure.years	-0.104344931	0.07014746	-1.48750832	0.137487889	-0.242152796	0.033462935
female	-1.832956317	0.255154726	-7.183705154	2.36962E-12	-2.334219342	-1.331693292
exper.years	0.186334006	0.034477558	5.404501249	9.92236E-08	0.11860128	0.254066732
exper^2	-0.003727954	0.000749196	-4.975940865	8.84564E-07	-0.005199783	-0.002256125
educ*tenure	0.019806316	0.005584847	3.546438498	0.000425829	0.008834631	0.030778001

Modelo de Regressão Linear Múltipla

Interacção entre variáveis explicativas

$$\widehat{wage} = -0.452 + 0.406\, educ - 0.104\, tenure - 1.833\, female + 0.186\, exper$$
$$-0.004\, exper^2 + 0.020\, (educ \times tenure)$$

Repare-se que:

- Na alteração dos coeficientes, nomeadamente de *educ* e *tenure*. Este último mudou até de sentido e deixou de ser estatisticamente significante.
- A interação é estatisticamente significante, tal como *educ* mas *tenure* deixou de o ser. Reforça a ideia que o impacto de *tenure* é feito em termos de *educ,* não existindo um impacto autónomo significativo (isto é independente de *educ*) . Já *educ* tem um impacto autónomo muito forte, com seria aliás de esperar.
- Neste como em todas as regressões baseadas neste conjunto de dados a discriminação de género é muito marcada!

Modelo de Regressão Linear Múltipla

Teste à forma funcional – RESET

Modelo: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

Objetivo: O teste RESET procura detetar uma má especificação da forma funcional, nomeadamente a omissão de uma variável relevante que esteja correlacionada com as variáveis explicativas incluídas no modelo ou uma não transformação da variável dependente.

Intuição: Se o modelo estiver bem especificado então nenhuma função das variáveis explicativas acrescenta algo ao modelo. Assim vamos compara o modelo estimado com um modelo auxiliar onde se acrescente às variáveis explicativas $\hat{y}^2$, $\hat{y}^3$.

Modelo de Regressão Linear Múltipla

Teste à forma funcional – RESET

Procedimento:

1. Estimar o modelo: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$ e guardar os $\hat{y}_i$

2. Definir o modelo auxiliar:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \varepsilon$$

3. Testar: $H_0: \gamma_1 = \gamma_2 = 0$ contra $H_1: \gamma_1 \neq 0 \; ou \; \gamma_2 \neq 0$

   Utilizar o teste F para a nulidade de um subconjunto de parâmetros (regressores). O teste F pode ser feito com base na variação residual (SSR) ou no R^2 uma vez que a variável dependente é a mesma nas duas regressões. Também se pode utilizer o teste LM.

Nota: A rejeição de H_0 requer que se procure uma forma functional alternativa

Modelo de Regressão Linear Múltipla

Teste à forma funcional – RESET

Comentários:

- Porquê as potências 2 e 3 para $\hat{y}$ na regressão auxiliar? Simplicidade e as não linearidades são geralmente bem apanhadas;
- Como se dá o benefício da dúvida a H_0 **a não rejeição de H_0 não garante que o nosso modelo é o mais adequado**. Apenas nos diz que não é rejeitado. Para quase todos os fenómenos é habitual encontrar vários modelos que passam o teste RESET!
- Ao invés, **se rejeitarmos H_0 o modelo tem de ser reformulado**

Modelo de Regressão Linear Múltipla

Teste à forma funcional – RESET

Exemplo: testar a forma funcional de

$$preço = \beta_0 + \beta_1 area + \beta_2 quartos + u$$

```
 regress preco quartos area
```

```
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   72.95
       Model |  579971.198     2  289985.599            Prob > F      =  0.0000
    Residual |  337883.308    85  3975.09774            R-squared     =  0.6319
-------------+------------------------------           Adj R-squared =  0.6232
       Total |  917854.506    87  10550.0518            Root MSE      =  63.048

------------------------------------------------------------------------------
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |   15.12134   9.488598     1.59   0.115    -3.744538    33.98721
        area |   1.383606   .1489435     9.29   0.000     1.087467    1.679746
       _cons |   -19.2855   31.04753    -0.62   0.536     -81.0163     42.4453
------------------------------------------------------------------------------
```

```
. predict precohat
(option xb assumed; fitted values)

. generate precohat2=precohat^2

. generate precohat3=precohat^3
```

Modelo de Regressão Linear Múltipla

Teste à forma funcional – RESET

Exemplo (continuação):

```
. regress preco quartos area precohat2 precohat3

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  4,    83) =   41.17
       Model |  610249.039     4  152562.26             Prob > F      =  0.0000
    Residual |  307605.467    83  3706.08996            R-squared     =  0.6649
-------------+------------------------------           Adj R-squared =  0.6487
       Total |  917854.506    87  10550.0518            Root MSE      =  60.878

------------------------------------------------------------------------------
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |  -58.37904   38.71904    -1.51   0.135    -135.3897    18.63158
        area |  -5.680895   3.613211    -1.57   0.120    -12.86743    1.505637
   precohat2 |   .0133394   .0076821     1.74   0.086    -.0019399    .0286187
   precohat3 |  -.0000109   7.20e-06    -1.52   0.133    -.0000252    3.40e-06
       _cons |   675.0476   328.2222     2.06   0.043     22.22683    1327.868
------------------------------------------------------------------------------
```

$H_0: \gamma_1 = \gamma_2 = 0 \rightarrow$ FF correcta

$$F_{obs} = \frac{(0.6649 - 0.631\ *)/2}{(1 - 0.6649)/(88 - 5)} = 4.08$$

A 5% de significância, $F(2,83) \simeq 3.15$

Rejeita-se H_0: a FF não é válida

Modelo de Regressão Linear Múltipla

Teste à forma funcional – RESET

Exemplo (continuação): procura de outra FF – adiciona-se o quadrado de area

```
. generate area2=area^2
. regress preco quartos area area2
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,     84) =   52.11
       Model |  597041.642     3  199013.881            Prob > F      =  0.0000
    Residual |  320812.864    84  3819.20076            R-squared     =  0.6505
-------------+------------------------------           Adj R-squared =  0.6380
       Total |  917854.506    87  10550.0518            Root MSE      =    61.8
------------------------------------------------------------------------------
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |    14.48683   9.305514     1.56   0.123    -4.018205    32.99187
        area |   -.2476931   .7852994    -0.32   0.753    -1.809347    1.313961
       area2 |    .0036535   .0017281     2.11   0.037      .000217    .0070901
       _cons |    149.9207   85.62566     1.75   0.084    -20.35527    320.1968
------------------------------------------------------------------------------

. predict precohatA
(option xb assumed; fitted values)
. generate precohatA2=precohatA^2
. generate precohatA3=precohatA^3
```

Modelo de Regressão Linear Múltipla

Teste à forma funcional – RESET

Exemplo (continuação):

```
. regress preco quartos area area2 precohatA2 precohatA3
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  5,    82) =   33.15
       Model |   614077.42     5  122815.484            Prob > F      =  0.0000
    Residual |  303777.086    82  3704.59861            R-squared     =  0.6690
-------------+------------------------------           Adj R-squared =  0.6489
       Total |  917854.506    87  10550.0518            Root MSE      =  60.865
------------------------------------------------------------------------------
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |  -138.0947   77.28978    -1.79   0.078    -291.8487    15.65929
        area |   6.140899   4.174249     1.47   0.145    -2.163011    14.44481
       area2 |  -.0453309   .0264509    -1.71   0.090    -.0979502    .0072884
  precohatA2 |   .0277233   .0134951     2.05   0.043     .0008772    .0545695
  precohatA3 |  -.0000214   .0000101    -2.12   0.037    -.0000415   -1.33e-06
       _cons |  -529.7486   404.1627    -1.31   0.194    -1333.757    274.2597
------------------------------------------------------------------------------
```

$H_0: \gamma_1 = \gamma_2 = 0 \rightarrow$ FF correcta

$$F_{obs} = \frac{(0.6690 - 0.6505)/2}{(1 - 0.6690)/(88 - 6)} = 2.30$$

A 5% de significância, $F(2,82) \simeq 3.15$

Não se rejeita H_0: a FF que inclui o quadrado de área **não é rejeitada (passa o RESET)**

Modelo de Regressão Linear Múltipla

Teste à forma funcional – RESET

Recorda-se que para um mesmo problema existem geralmente vários modelos que passam o teste RESET.

Voltando ao exemplo e considerando agora o dataset completo (ver ficheiro "hprice1 com nomes.xlsx") onde apenas se transformou a unidade de medida das áreas de pés-quadrados para m2, pode verificar-se que o modelo

$$\widehat{\ln preço} = \underset{}{0.766} + \underset{(0.0383)}{0.168 \ln arealote} + \underset{(0.0929)}{0.700 \ln areacasa} + \underset{(0.0275)}{0.037\ quartos}$$

Também passa o teste RESET

$$F_{obs} = \frac{(0.66399 - 0.64300)/2}{(1 - 0.66399)/(88 - 6)} = 2.565$$

A 5% de significância, $F(2{,}82) \simeq 3.15$ ou $p - value = 0.083$

Em suma: O teste RESET serve essencialmente para rejeitar formalizações incorretas e não tanto para escolher **a** formalização adequada

Modelo de Regressão Linear Múltipla

Heterocedasticidade O que é?

Violação da hipótese MRL 5 – Homocedasticidade – $var(u|\boldsymbol{x}) = \sigma^2$, isto é **a variância da variável residual u deixa de ser constante.**

A variância pode:

. modelar-se como função (total ou parcialmente conhecida) das variáveis explicativas ($var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$)

. depender de variáveis que não estão no modelo, i.é, as variáveis que explicam a variância podem não ser as mesmasque explicam o valor esperado

. Pode ter padrão desconhecido

Modelo de Regressão Linear Múltipla

Heterocedasticidade O que é?

Violação da hipótese MRL 5 – Homocedasticidade – $var(u|\boldsymbol{x}) = \sigma^2$, isto é **a variância da variável residual u deixa de ser constante.**

FIGURE 2.8

The simple regression model under homoskedasticity.

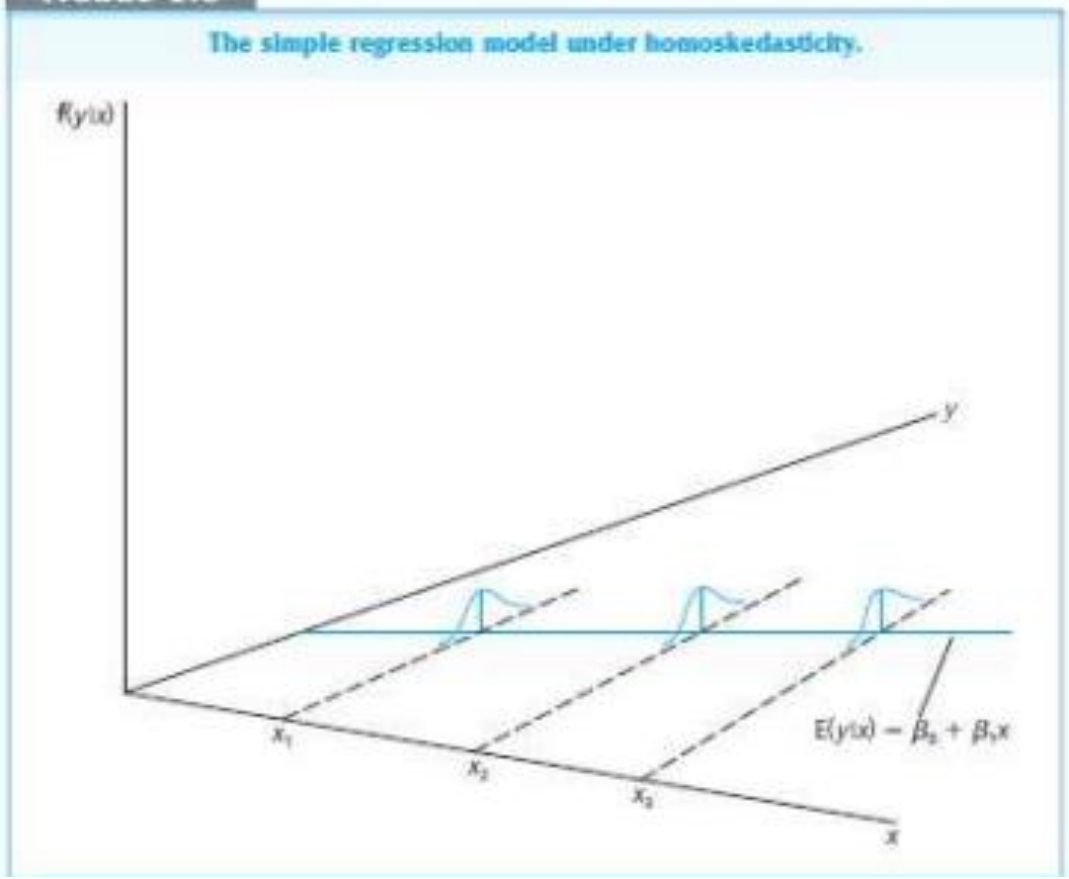
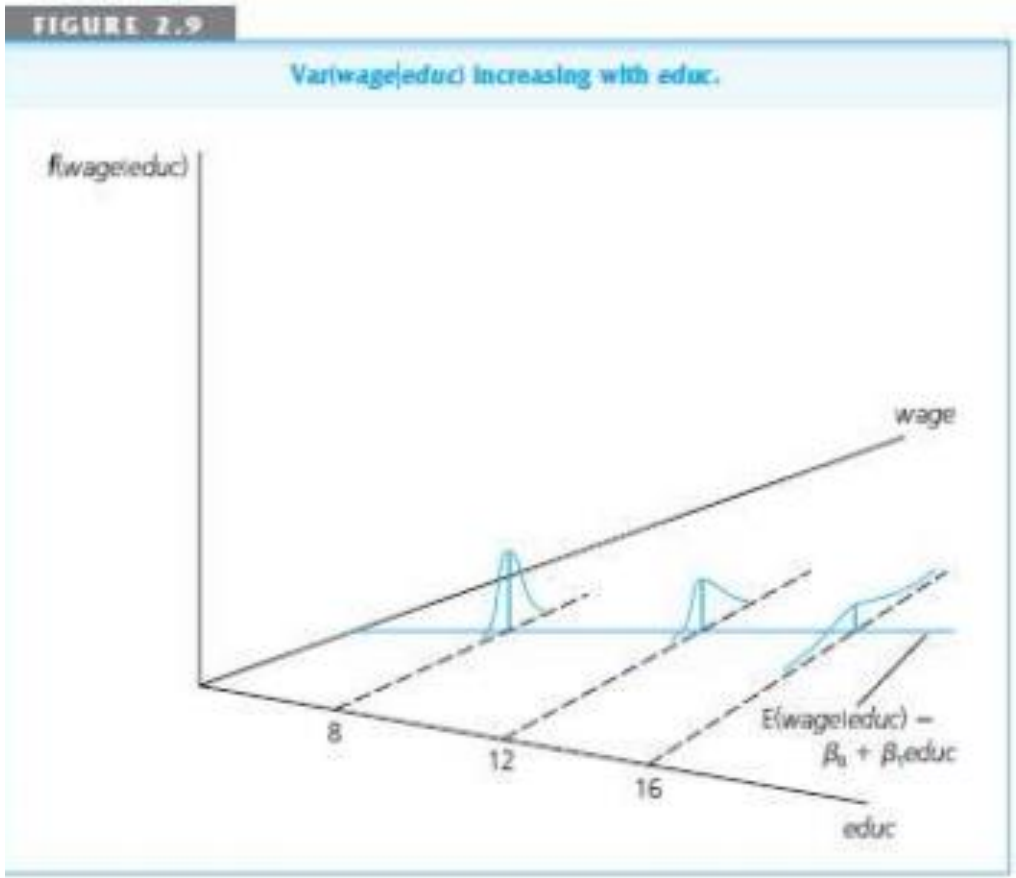


FIGURE 2.9

Var(wage|*educ*) increasing with *educ*.



Modelo de Regressão Linear Múltipla

Heterocedasticidade Porque acontece?

- Má especificação do modelo, nomeadamente:
  - ausência de algumas variáveis relevantes
  - Não transformação de variáveis (logaritmização)

- Natureza do problema

  É natural que um modelo onde a poupança de um agregado familiar é explicada pelo rendimento disponível do agregado sofra de heterocedasticidade por não ter em conta uma diferente dimensão do agregado familiar ou o número de elementos com menos de 22 anos e que a heterocedasticidade seja de alguma forma proporcional ao rendimento disponível. Maior rendimento disponível estará associado a maior variabilidade na poupança.

Modelo de Regressão Linear Múltipla

Heterocedasticidade Que consequências?

Ao cair a hipótese MRL 5, o estimador MMQ(OLS) (β)

. continua a ser centrado e consistente (MRL 1 a MRL 4)

. deixa de ser o mais eficiente

. A estimação da variância do estimador, tal como vimos, deixa de ser válida

. As estatísticas t e F deixam de ter distribuição de t-Student e f-Snédecor, respectivamente

. Mantém-se válidas as estatísticas R^2 e $\bar{R}^2$

Nota: assume-se que se mantém as restantes hipóteses do MRL, nomeadamente, MRL 6 e MRL 2→ ausência correlação entre os u_i → matriz das variâncias/covariâncias dos u_i é diagonal

Modelo de Regressão Linear Múltipla

Heterocedasticidade Que soluções?

A solução está dependente do conhecimento que se tenha (ou não) da razão pela qual ela existe.

skedastic function

1. Se se pode assumir que $var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$, com $h(\boldsymbol{x})$ conhecida, substitui-se MRL 5 por uma hipótese mais fraca MRL5'

Os resultados irão depender da validade desta nova hipótese.

Homocedasticidade corresponde a assumir $h(\boldsymbol{x}) = \mathbf{1}$

Assim, tem-se 2 alternativas:

- 1. Estimação robusta – assumir $var(u_i|\boldsymbol{x}) = \sigma_i^2$
- 2. Estimação por GLS (*skedastic function conhecida)*
  assumir $var(u_i|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação robusta

$\hat{\beta} = (X^T X)^{-1} X^T Y$ continuam a ser centrados e consistentes

É necessário apenas corrigir a respectiva variância de modo a que o teste *t* seja válido

Como estimar a $var(\hat{\beta})$?

Sabe-se que:

- $var(u_i|\boldsymbol{X}) = \sigma_i^2$
- $cov(u_i, u_r|\boldsymbol{X}) = 0$, para $i \neq r$ (a hipótese MRL 2 continua válida)
- $var(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T var(U|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação robusta

Enquanto com MLR 5 se utilizava $var(U|\boldsymbol{X}) = \sigma^2\mathbf{I}$, tem-se agora de assumir $var(U|\boldsymbol{X}) = \boldsymbol{\Sigma}$, sendo $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$.

$$var(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Como estimar Σ de forma robusta ?

Estimador de White:

Intuição: $var(u_i|\boldsymbol{X}) = E(u_i^2|\boldsymbol{X}) - \left(E(u_i|\boldsymbol{X})\right)^2 = E(u_i^2|\boldsymbol{X})$ MLR4

$\hat{\sigma}_i^2 = \hat{u}_i^2$ e portanto $\hat{\boldsymbol{\Sigma}} = diag(\hat{\boldsymbol{\sigma}}_1^2, \hat{\boldsymbol{\sigma}}_2^2, \ldots, \hat{\boldsymbol{\sigma}}_n^2)$

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação robusta

A matriz $\left(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{X}\right)$ vem assim

$$\left(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{X}\right) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ \dots & \dots & & \dots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix} \times \begin{bmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \dots & \dots & & \dots \\ 0 & 0 & \dots & \hat{\sigma}_n^2 \end{bmatrix} \times \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_2^2 & \dots & \hat{\sigma}_n^2 \\ \hat{\sigma}_1^2 x_{11} & \hat{\sigma}_2^2 x_{21} & \dots & \hat{\sigma}_n^2 x_{n1} \\ \dots & \dots & & \dots \\ \hat{\sigma}_1^2 x_{1k} & \hat{\sigma}_2^2 x_{2k} & \dots & \hat{\sigma}_n^2 x_{nk} \end{bmatrix} \times \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} \sum \hat{\sigma}_i^2 & \sum \hat{\sigma}_i^2 x_{i1} & \dots & \sum \hat{\sigma}_i^2 x_{ik} \\ \sum \hat{\sigma}_i^2 x_{i1} & \sum \hat{\sigma}_i^2 x_{i1}^2 & \dots & \sum \hat{\sigma}_i^2 x_{i1} x_{ik} \\ & \dots & \dots & \dots \\ \sum \hat{\sigma}_i^2 x_{ik} & \sum \hat{\sigma}_i^2 x_{ik} x_{i1} & \dots & \sum \hat{\sigma}_i^2 x_{ik}^2 \end{bmatrix}$$

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação robusta

O estimador para a matriz das variâncias covariâncias de $\hat{\beta}$, será:

$$\widehat{var}(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Que se demonstra ser um **estimador consistente**

[Wooldrige] Uma forma alternativa, mais fácil, de obter os elementos da diagonal principal desta matriz consiste em calcular:

$$\widehat{var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \tilde{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

$\tilde{r}_{i,j}^2$ - resíduos da regressão de x_j nas restantes variáveis explicativas

$$\sum_{i=1}^{n} \tilde{r}_{i,j}^2$$

Modelo de Regressão Linear Múltipla

Heteroscedataticidade: estimação robusta de $\widehat{var}(\hat{\beta})$

Exemplo: Considere-se o exemplo habitual dos imóveis (na sua versão mais simples) em que o modelo estimado era

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

e calcule-se os erros-padrão de forma robusta.

Solução 1:

Com base nesta regressão obtiveram-se os $\hat{u}_i$ e contruiu-se $\hat{\Sigma} = diag(\hat{u}_i^2)$.

Recorrendo ao software R obteve-se:

$$\widehat{var}(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 1666.8 & -5.9835 & 179.80 \\ -5.9835 & 0.04307 & -0.4474 \\ -179.80 & -0.4474 & 77.6484 \end{bmatrix}$$

Isto. É, os erros padrão robustos para $\hat{\beta}_1$ e $\hat{\beta}_2$ são respectivamente: 0.2075 e 8.8118

Modelo de Regressão Linear Múltipla

Heteroscedaticidade: estimação robusta de $var(\hat{\beta})$

Solução 2:

Com base na regressão original obtiveram-se os $\hat{u}_i$

Como se tem 2 parâmetros são necessárias 2 regressões auxiliares:

- $\widehat{area} = -19.286 + 15.121\ quartos \rightarrow$ obter $\tilde{r}_{i1}$ e calcular

$$\widehat{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} \tilde{r}_{i1}^2\ \hat{u}_i^2}{SSR_1^2} = 0.04307$$

Nota: $SSR_1^2 = \left(\sum_{i=1}^{n} \tilde{r}_{i1}^2\right)^2$

- $\widehat{quartos} = -19.286 + 1.384\ area \rightarrow$ obter $\tilde{r}_{i2}$ e calcular

$$\widehat{var}(\hat{\beta}_2) = \frac{\sum_{i=1}^{n} \tilde{r}_{i2}^2\ \hat{u}_i^2}{SSR_2^2} = 77.6484$$

Recorre-se normalmente a software para calcular os erros padrão robustos à heteroscedaticidade.

Modelo de Regressão Linear Múltipla

Heteroscedaticidade: estimação robusta de $var(\hat{\beta})$

Estes estimadores são conhecidos como White, Huber ou Eicker (ou combinações de 2 ou 3 destes nomes)

Aparecem geralmente na forma "corrigida" que consiste em multiplicar as variâncias referidas anteriormente por $\frac{n}{n-k-1}$ para melhorar a compatibilidade com o caso homocedástico.

Como é evidente o fator $\frac{n}{n-k-1}$ não altera a validade assintótica do estimador, (já que $\frac{n}{n-k-1} \rightarrow 1$ quando $n \rightarrow \infty$)

Corrigidos os erros-padrão a inferência é feita nos termos habituais.

Modelo de Regressão Linear Múltipla

Heteroscedaticidade: estimação robusta de $var(\hat{\beta})$

Output EVIEWS

Dependent Variable: PRECO
Method: Least Squares
Date: 04/27/20 Time: 17:33
Sample: 1 88
Included observations: 88
Huber-White-Hinkley (HC1) heteroskedasticity consistent standard errors and covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-19.28550	41.54017	-0.464261	0.6436
AREA_M2_	1.383606	0.211163	6.552317	0.0000
QUARTOS	15.12134	8.965990	1.686522	0.0954

R-squared	0.631877	Mean dependent var	293.5460
Adjusted R-squared	0.623215	S.D. dependent var	102.7134
S.E. of regression	63.04838	Akaike info criterion	11.15918
Sum squared resid	337883.3	Schwarz criterion	11.24363

Observação: Como é evidente os erros-padrão robustos são iguais àqueles que se obtiveram no slide 9 depois de multiplicados pelo fator (88/85).

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação robusta

Para o teste F a situação é mais complicada e requer software adequado (embora com heterocedasticidade moderada se possa continuar a utilizar o teste). Pode no entanto utilizar-se o teste LM (multiplicadores de Lagrange) adaptado para testar a nulidade de vários coeficientes. Neste caso convém que a amostra tenha dimensão adequada dado que o teste é assintótico.

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação WLS e GLS

Quando se substitui MRL 5 por $var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$ existem duas alternativas:

1. $h(\boldsymbol{x})$ é uma função conhecida **sem** parâmetros desconhecidos
2. $h(\boldsymbol{x})$ é uma função conhecida **com** parâmetros desconhecidos

O caso 1 onde se assume que a heterocedasticidade é conhecida a menos de uma constante (estimação WLS) tem naturalmente uma solução mais simples do que o caso 2 (estimação GLS).

Os slides apenas abordam o caso 1, sendo que o caso 2 pode ser visto no Wooldridge (sub-secção *Feasible GLS* no quadro da secção *Weighted Least Squares Estimation*)

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação WLS

$var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$ sendo $h(\boldsymbol{x})$ uma função conhecida **sem** parâmetros desconhecidos

O método **WLS** produz estimadores centrados, consistentes e eficientes (independentemente da dimensão da amostra) cuja validade está obviamente ligada à hipótese que se assumiu.

A ideia base é partir do modelo de interesse

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \text{ com } var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$$

e obter um modelo transformado que verifique MRL 5.

Ao dividir os 2 lados da igualdade por $\sqrt{h(\boldsymbol{x})}$ teremos no termo de erro $u^* = \frac{u}{\sqrt{h(\boldsymbol{x})}}$ e portanto $var(u^*|\boldsymbol{x}) = var\left(\frac{u}{\sqrt{h(\boldsymbol{x})}}|\boldsymbol{x}\right) = \frac{var(u|\boldsymbol{x})}{h(\boldsymbol{x})} = \sigma^2$.

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação WLS

Modelo de interesse (escrito em termos da amostra)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i \text{ com } var(u_i|\boldsymbol{x}_i) = \sigma^2 h_i \text{ e } h_i = \mathrm{h}(\boldsymbol{x}_i)$$

Modelo transformado:

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$$

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + +\beta_2 x_{i2}^* + \cdots + \beta_1 x_{ik}^* + u_i^*$$

$$\text{com } y_i^* = \frac{y_i}{\sqrt{h_i}}, x_{i0}^* = \frac{1}{\sqrt{h_i}}, \; x_{ij}^* = \frac{x_{ij}}{\sqrt{h_i}} \; (j = 1,2, \ldots, k), u_i^* = \frac{u_i}{\sqrt{h_i}}$$

Notas: - O modelo transformado não tem termo constante

- Interpretação dos $\beta(s)$ é feita em função do modelo original

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação WLS

Modelo de interesse: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$

Modelo transformado: $y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + +\beta_2 x_{i2}^* + \cdots + \beta_1 x_{ik}^* + u_i^*$

Observações:

- A inferência estatística (R^2, $\bar{R}^2$, testes t, teste F, ...) é feita com base no modelo transformado.
- **A interpretação dos β's é feita em função do modelo original.**
- O estimador GLS (*Generalized Least Squares*) para esta situação é conhecido como WLS (*Weighted Least Squares*) uma vez que minimiza a soma dos quadrados dos resíduos ponderada em que cada termo é ponderado por $\frac{1}{h_i}$,

$$\sum \hat{u}_i^{*2} = \sum (y_i^* - \hat{y}_i^*)^2 = \sum \left(\frac{y_i}{\sqrt{h_i}} - \frac{\hat{y}_i}{\sqrt{h_i}} \right)^2 = \sum \frac{(y_i - \hat{y}_i)^2}{h_i}$$

Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação WLS

Retome-se o exemplo habitual do preço de um imóvel como função da área e do número de quartos.

A estimação do modelo feita anteriormente originava

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$
$$(0.149) \qquad (9.489)$$
$$[0.211] \qquad [8.966]$$

Modelo transformado assumindo $var(u|\boldsymbol{x}) = \sigma^2 area$

$$\frac{preço_i}{\sqrt{area_i}} = \beta_0 \frac{1}{\sqrt{area_i}} + \beta_1 \frac{area_i}{\sqrt{area_i}} + \beta_2 \frac{quartos_i}{\sqrt{area_i}} + u_i^*$$

Estimação OLS (ver slides seguintes)

$$\left(\frac{\widehat{preço_i}}{\sqrt{area_i}}\right) = 7.8960 \frac{1}{\sqrt{area_i}} + 1.3080 \frac{area_i}{\sqrt{area_i}} + 11.4685 \frac{quartos_i}{\sqrt{area_i}}$$
$$(0.1555) \qquad (8.9630)$$

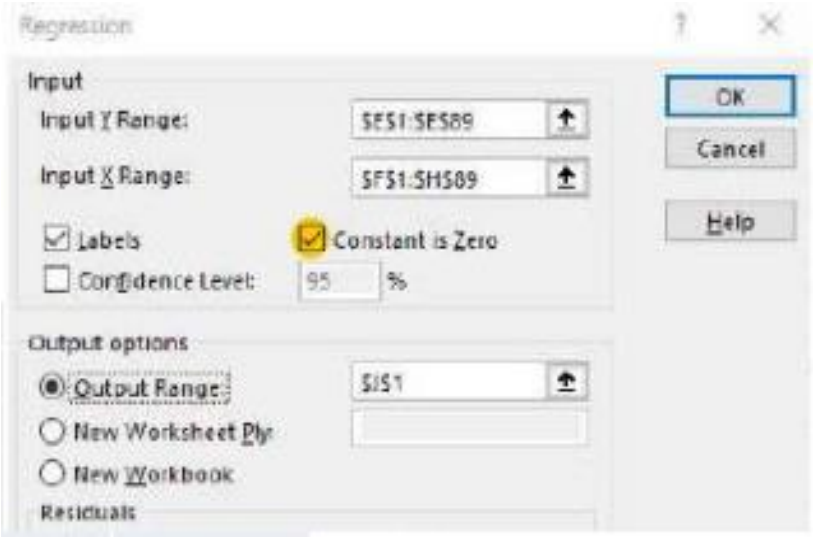
Modelo de Regressão Linear Múltipla

Heterocedasticidade: estimação WLS EXCEL

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.979954982
R Square	0.960311767
Adjusted R Square	0.947613221
Standard Error	4.442557176
Observations	88

preço	area	quartos	preço*	x*_0	area*	quartos*
300	226	4	19.956	0.067	15.033	0.266
370	193	3	26.633	0.072	13.892	0.216

ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	40591.53428	13530.51143	685.564247	6.25012E-59	
Residual	85	1677.586712	19.73631426			
Total	88	42269.12099				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
x*_0	7.895976487	30.70298548	0.257172922	0.79766681	-53.1497842	68.94173718
area*	1.307992309	0.155539391	8.409395838	8.46173E-13	0.998738329	1.617246289
quartos*	11.46850858	8.963038244	1.279533599	0.204190848	-6.352412713	29.28942988



Modelo de Regressão Linear Múltipla

Heterocedasticidade: testes para detecção

Três testes, todos baseados em regressões auxiliares com variável dependente $\hat{u}^2$, onde a significância global dos parâmetros é testada através de um teste F / LM.

Modelo base:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

H_0: Homocedasticidade $\rightarrow var(u|\boldsymbol{x}) = \sigma^2$ (usar OLS)

H_1: Heterocedasticidade (usar GLS ou OLS com erros-padrão robustos)

Tal como os testes estão formulados, só se rejeita a homocedasticidade quando os dados aponta claramente para a heterocedasticidade.

Modelo de Regressão Linear Múltipla

Heterocedasticidade: testes para detecção

1. **Procedimento comum**: estimar o modelo base por OLS e obter $\hat{u}^2$
2. Estimar a regressão auxiliar (depende do teste escolhido)
   - Breusch-Pagan: $\hat{u}^2 = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_k x_k + e$

Assume que heterocedasticidade é função das variáveis explicativas.

Nota: pode restringir-se a um subconjunto destas se se pensar que a heterocedasticidade está relacionada apenas com este subconjunto.

Modelo de Regressão Linear Múltipla

Heterocedasticidade: testes para detecção

2. Estimar a regressão auxiliar (depende do teste escolhido)

- White : Incluir todas as variáveis mais os seus quadrados e produtos cruzados – Exemplo para $k = 2$

$$\hat{u}^2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1^2 + \gamma_4 x_2^2 + \gamma_5 x_1 x_2 + e$$

- White simplificado: $\hat{u}^2 = \gamma_0 + \gamma_1 \hat{y} + \gamma_2 \hat{y}^2 + e$

em qualquer das alternativas, obter $R^2_{\hat{u}^2}$

Em termos de filosofia, o teste BP procura um padrão para a heterocedasticidade enquanto o teste de White tem uma filosofia mais "robusta"

Modelo de Regressão Linear Múltipla

Heterocedasticidade: testes para detecção

3. Estatística de teste e distribuição:

$$F = \frac{R^2_{\hat{u}^2}/m}{(1 - R^2_{\hat{u}^2})/(n - m - 1)} \sim F(m, n - m - 1)$$

ou

$$LM = nR^2_{\hat{u}^2} \sim \chi^2_m$$

onde m corresponde ao número de declives na regressão auxiliar

Breusch-Pagan$\rightarrow m = k$

White simplificado $\rightarrow m = 2$

White simplificado $\rightarrow m = k + k + \frac{k(k+1))}{2} = \frac{k(k+3))}{2}$

Modelo de Regressão Linear Múltipla

Heterocedasticidade: testes para detecção – exemplo 1

Exemplo: Considere-se o modelo habitual

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

BP Output Eviews

Heteroskedasticity Test: Breusch-Pagan-Godfrey			
Null hypothesis: Homoskedasticity			
F-statistic	5.873654	Prob. F(2,85)	0.0041
Obs*R-squared	10.68519	Prob. Chi-Square(2)	0.0048
Scaled explained SS	23.45408	Prob. Chi-Square(2)	0.0000

Rejeita-se homocedasticidade

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/26/20 Time: 14:40
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8300.608	3911.614	-2.122042	0.0367
AREA_M2_	42.12340	18.76508	2.244776	0.0274
QUARTOS	1193.551	1195.449	0.998413	0.3209

R-squared	0.121423
Adjusted R-squared	0.100750
S.E. of regression	7943.334
Sum squared resid	5.36E+09
Log likelihood	-913.5882
F-statistic	5.873654

Modelo de Regressão Linear Múltipla

Heterocedasticidade: testes para detecção – exemplo 1

Exemplo: Considere-se o modelo habitual

White Output Eviews

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

Heteroskedasticity Test: White
Null hypothesis: Homoskedasticity

F-statistic	4.007899	Prob. F(5,82)	0.0027
Obs*R-squared	17.28228	Prob. Chi-Square(5)	0.0040
Scaled explained SS	37.93476	Prob. Chi-Square(5)	0.0000

Rejeita-se homocedasticidade

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/26/20 Time: 14:40
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10706.38	13167.06	0.813118	0.4185
AREA_M2_^2	0.465846	0.255698	1.821857	0.0721
AREA_M2_*QUARTOS	21.37083	19.60682	1.089970	0.2789
AREA_M2_	-251.8630	108.9744	-2.311212	0.0233
QUARTOS^2	-1281.901	840.9803	-1.524294	0.1313
QUARTOS	7025.944	5680.170	1.236925	0.2196

R-squared	0.196390
Adjusted R-squared	0.147389
S.E. of regression	7734.605
Sum squared resid	4.91E+09
Log likelihood	-909.6639
F-statistic	4.007899

Modelo de Regressão Linear Múltipla

Heterocedasticidade: testes para detecção – exemplo 1

Exemplo: Considere-se o modelo habitual **White simplificado**

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.160192	0.115627	1.385425	0.1695
YC	-0.000731	0.000697	-1.048718	0.2973
YC2	1.02E-06	9.84E-07	1.032361	0.3048

R-squared	0.012779
Adjusted R-squared	-0.010450
S.E. of regression	0.083458
Sum squared resid	0.592050
Log likelihood	95.19945
F-statistic	0.550126

$$RES2 = (preço - \widehat{preço})^2$$

$$YC = \widehat{preço}$$

$$YC2 = \widehat{preço}^2$$

$$F_{obs} = \frac{0.012779/2}{(1-0.012779)/85} = 0.550$$

$$p - value = 0.579$$

$$LM_{obs} = 88 * 0.012779$$

$$p - value = 0.570$$

Modelo de Regressão Linear Múltipla
Heterocedasticidade: testes para detecção – exemplo 2

Exemplo: testar heterocedasticidade no âmbito do modelo

$$\widehat{ln(preço)} = 1.289 + 0.8101 \ln(\text{area}) + 0.0376 \text{ quartos}$$

```
gen larea=log(area)
gen lpreço=log(preco)
regress  lpreço larea quartos
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   54.47
       Model |  4.50364223     2  2.25182112            Prob > F      =  0.0000
    Residual |  3.51396129    85  .041340721           R-squared     =  0.5617
-------------+------------------------------           Adj R-squared =  0.5514
       Total |  8.01760352    87  .092156362           Root MSE      =  .20332
------------------------------------------------------------------------------
      lpreço |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .8100637   .0987611     8.20   0.000     .6137002    1.006427
     quartos |   .0376464   .0303446     1.24   0.218    -.0226868    .0979795
       _cons |    1.28929   .4666125     2.76   0.007     .3615395    2.217041
------------------------------------------------------------------------------
predict uhat, resid
gen uhat2=uhat^2
predict yhat
gen yhat2=yhat^2
```

Modelo de Regressão Linear Múltipla
Heterocedasticidade: testes para detecção – exemplo 2 (cont.)

Heteroskedasticity Test: Breusch-Pagan-Godfrey			
Null hypothesis: Homoskedasticity			
F-statistic	1.928666	Prob. F(2,85)	0.1516
Obs*R-squared	3.820114	Prob. Chi-Square(2)	0.1481
Scaled explained SS	7.616429	Prob. Chi-Square(2)	0.0222

Teste BP - Eviews

R-squared	0.043410
Adjusted R-squared	0.020902
S.E. of regression	0.082153
Sum squared resid	0.573680
Log likelihood	96.58631
F-statistic	1.928666

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.274437	0.188535	1.455625	0.1492
LAREA	-0.060722	0.039905	-1.521670	0.1318
QUARTOS	0.022712	0.012261	1.852374	0.0674

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/26/20 Time: 14:50
Sample: 1 88
Included observations: 88

Não se rejeita H_0

Existe evidência de homocedasticidade

Modelo de Regressão Linear Múltipla
Heterocedasticidade: testes para detecção – exemplo 2 (cont.)

Teste White - Eviews

F-statistic	1.297465	Prob. F(5,82)	0.2731
Obs*R-squared	6.451600	Prob. Chi-Square(5)	0.2647
Scaled explained SS	12.86301	Prob. Chi-Square(5)	0.0247

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.673631	3.110976	1.502304	0.1369
LAREA^2	0.167924	0.123342	1.361455	0.1771
LAREA*QUARTOS	-0.007376	0.042017	-0.175542	0.8611
LAREA	-1.805730	1.235645	-1.461366	0.1477
QUARTOS^2	-0.008546	0.008595	-0.994323	0.3230
QUARTOS	0.127937	0.197192	0.648794	0.5183

F-statistic	1.297465

Não se rejeita H_0

Existe evidência de homocedasticidade

Modelo de Regressão Linear Múltipla
Heterocedasticidade: testes para detecção – exemplo 2 (cont.)

Teste White simplificado - Eviews

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.334571	4.153411	1.043617	0.2996
YC	-1.488248	1.453358	-1.024006	0.3087
YC2	0.128648	0.126999	1.012978	0.3139

R-squared	0.013964
Adjusted R-squared	-0.009237
S.E. of regression	0.083408
Sum squared resid	0.591339
Log likelihood	95.25231
F-statistic	0.601882

$$RES2 = \left(\ln preço - \ln \widehat{preço}\right)^2$$

$$YC = \ln \widehat{preço}$$

$$YC2 = \left(\ln \widehat{preço}\right)^2$$

$$Q_{obs} = 88 \times 0.013964 = 1.229$$

$$p - value = 0.541$$

Não se rejeita H_0

Existe evidência de homocedasticidade