

Qualidade do Ajuste e Binárias

Prof. Alexandre Gori Maia

Universidade Estadual de Campinas

Disciplina: Econometria Aplicada II

Universidade Nacional Agraria La Molina

Ementa

Teorema de Gauss-Markov

Formas Funcionais

Quebra de Pressupostos

Variáveis Binárias

Bibliografia

Maia, A. G. 2014. Econometria: conceitos e aplicações. Insittuto de Economia. Caps. 4, .

Teorema de Gauss-Markov

- Para que os estimadores de MQO sejam os Melhores Estimadores Lineares Não Viesados (MELNV, ou *BLUE*):

1) Relação Linear entre x e Y :

O ajuste só é válido para relações lineares.

**Viés e
Consistência**

2) Os valores de x são fixos em repetidas amostras, e não aleatórios:

Quem varia é o regressando. O regressor é fixo e dado, qualquer que seja a amostra. Fazemos a pressuposição que dados os valores de X , Y variará segundo uma distribuição de probabilidade com valor esperado dado por $E(Y/x_i)$.

3) Os erros possuem esperança condicional zero, ou seja, $E(e | x_i) = 0$:

É a mesma coisa afirmar que $E(Y/x_i) = x_i\beta$.

4) A variabilidade dos erros é constante, qualquer que seja x : **Eficiência**

Não há relação entre os erros e as variáveis independentes.

5) Os erros são não autocorrelacionados, ou seja, $E(e_{it}e_{js}) = 0$:

Não há relação entre valores ordenados dos erros segundo tempo ou espaço.

6) Os erros apresentam distribuição normal:

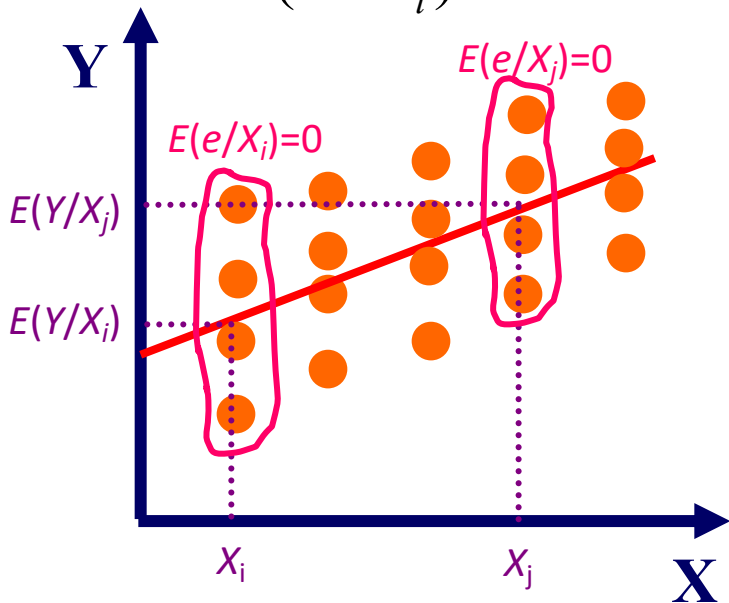
Inferência

Não é um pressuposto necessário para que os estimadores de MQO sejam MELNV, mas necessário para que estes tenham distribuição normal.

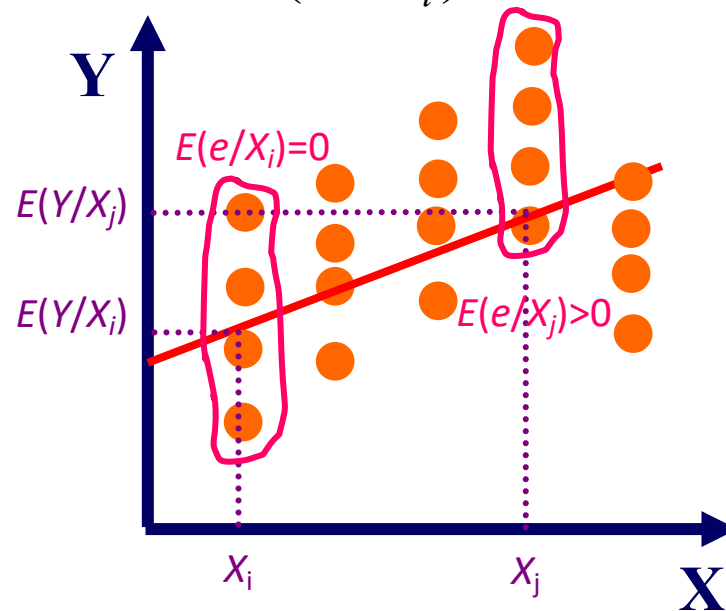
Esperança Condicional Zero

- Para que os estimadores de MQO sejam não tendenciosos, as esperanças condicionais dos erros devem ser iguais a zero:

$$E(e / X_i) = 0$$



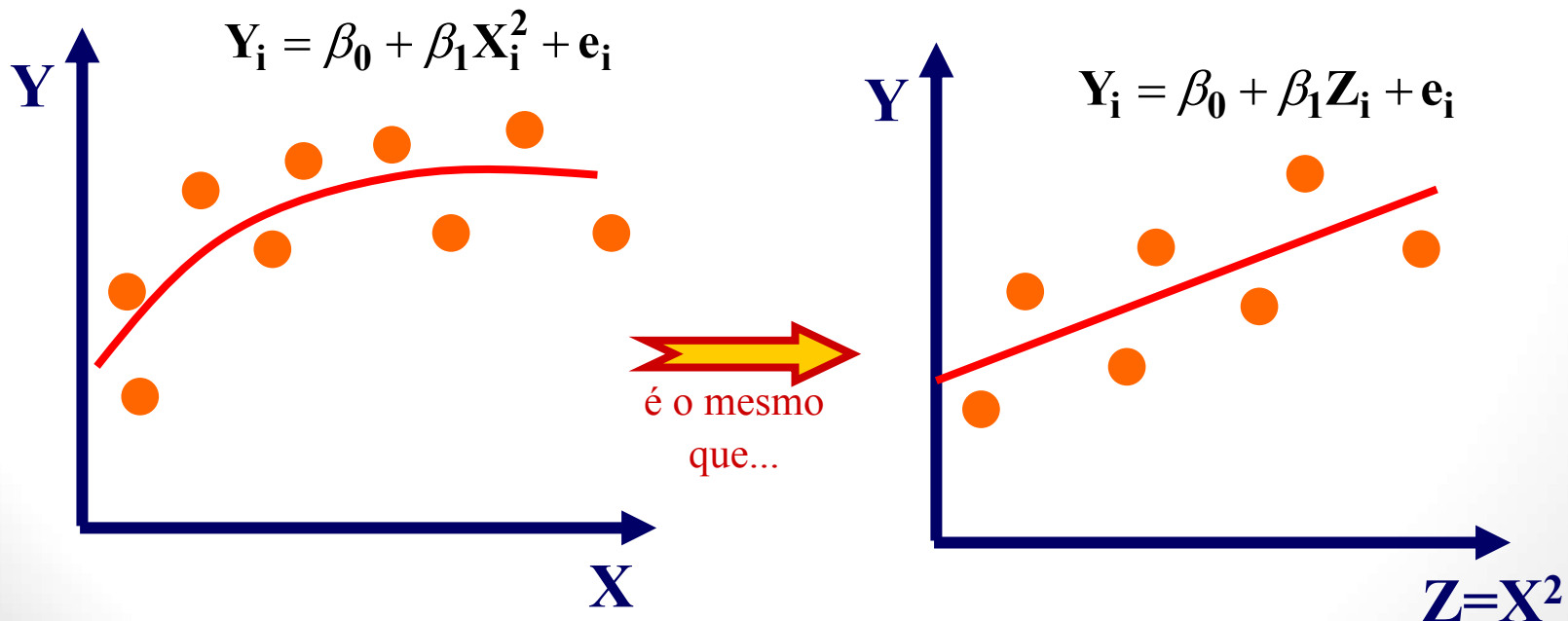
$$E(e / X_i) \neq 0$$



- É a mesma coisa que pressupor: $E(Y/X_i) = \alpha + \beta X_i$
- É um pressuposto mais forte que $Cov(e, X)=0$, pois este considera apenas relações lineares entre erros e regressores;

Relação Linear

- Relações não lineares nas variáveis podem ser transformadas em relações lineares por anamorfose, ou seja, por transformações de suas variáveis originais;
- A escolha da forma funcional (tipo de transformação das variáveis) dependerá da análise da distribuição dos valores e, principalmente, do conhecimento prévio das relações por parte do pesquisador;



Erro de Especificação

Seja o modelo de regressão inicialmente dado por:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + e_i$$

Haverá erro de especificação da forma funcional caso a correta relação entre as variáveis seja, por exemplo, dada pelas funções:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{2_i}^2 + e_i$$

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{1_i} \cdot X_{2_i} + e_i$$

$$Y_i = \alpha + \beta_1 \ln(X_{1_i}) + \beta_2 \ln(X_{2_i}) + e_i$$

Em qualquer das situações, a especificação inadequada levará a estimativas tendenciosas e inconsistentes dos coeficientes.

Teste RESET (Ramsey) de Especificação

Seja o modelo de regressão inicialmente dado por:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_{1_i} + \hat{\beta}_2 X_{2_i} + \hat{Y}_i + e_i$$

Uma maneira de avaliarmos a necessidade de, por exemplo, termos quadráticos e combinações das variáveis independentes, seria ajustar o modelo:

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{1_i} \cdot X_{2_i} + \beta_4 X_{1_i}^2 + \beta_5 X_{2_i}^2 + e_i$$

E testar a hipótese $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$

Para economizarmos graus de liberdade, podemos alternativamente ajustar o modelo:

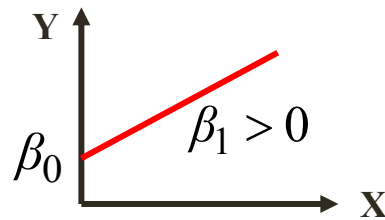
$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + e_i$$

E testarmos as hipóteses: $H_0 : \beta_3 = \beta_4 = 0$

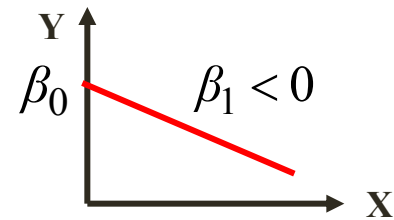
Formas Funcionais - Exemplos

1) Modelo linear:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

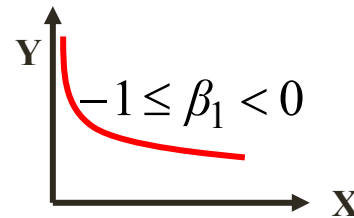


ou

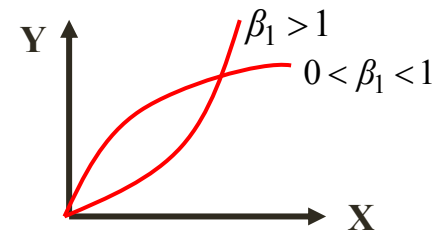


2) Modelo log-log:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + e_i$$

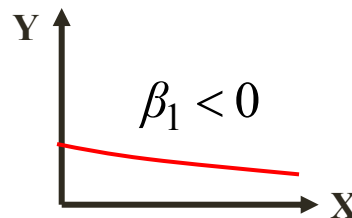


ou

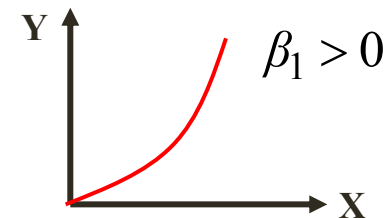


3) Modelo log-lin:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + e_i$$

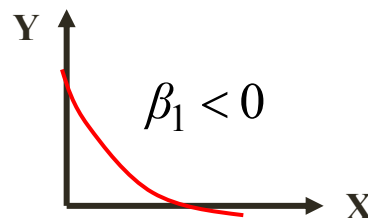


ou

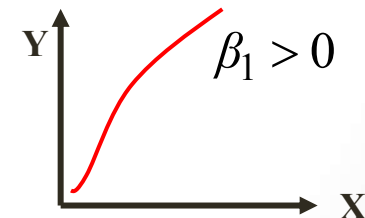


4) Modelo lin-log:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + e_i$$

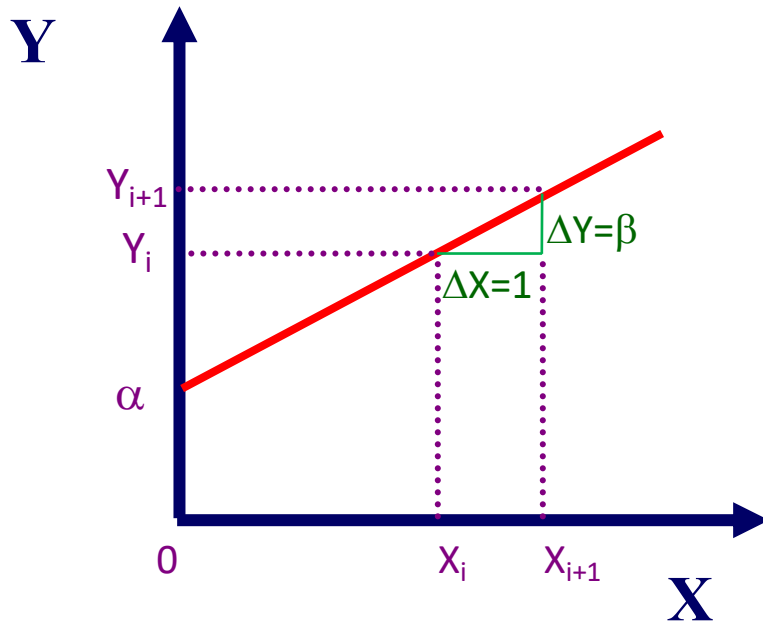


ou



Modelo Linear - Interpretação

- Um modelo linear pressupõe que Y apresente variações absolutas constantes para variações absolutas constantes em X ;
- A variação marginal no valor esperado de Y é a mesma para qualquer valor de X ;



$$E[Y / 0] = \alpha + \beta(0) = \alpha$$

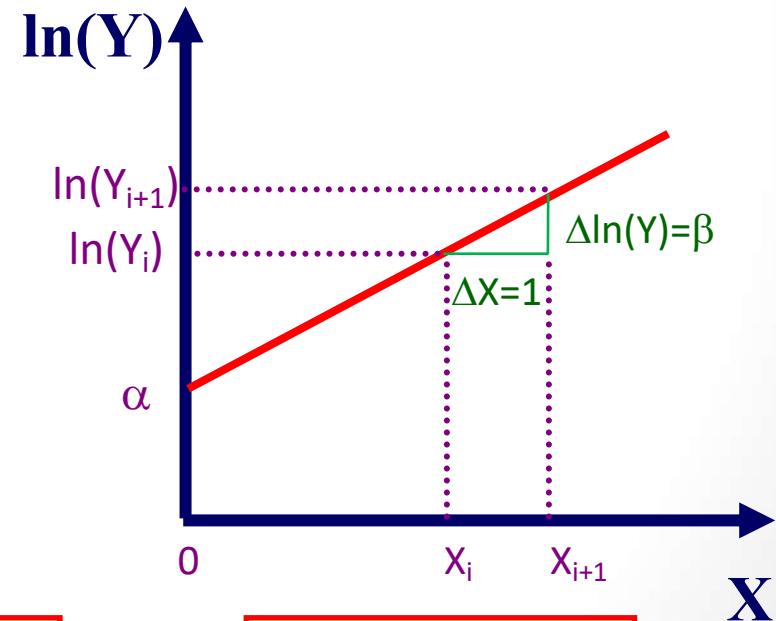
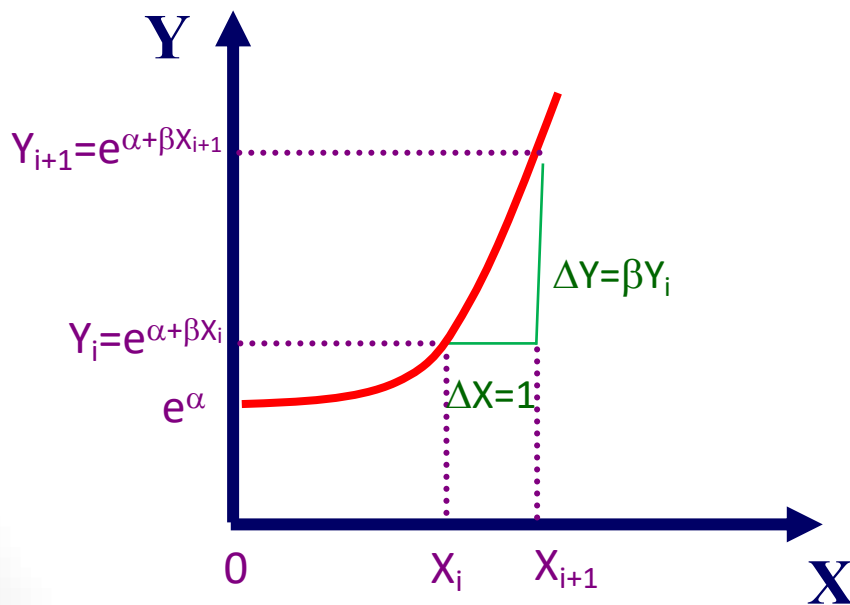
α é o valor esperado de Y para valores nulos de X

$$\frac{\Delta Y}{\Delta X} = \frac{dY}{dX} = \frac{d(\alpha + \beta X)}{dX} = \beta$$

β é a variação marginal absoluta no valor esperado de Y (ΔY) dada uma variação unitária em X ($\Delta X=1$).

Modelo Log-Lin - Interpretação

- Um modelo log-lin pressupõe que Y apresente variações relativas constantes em função de variações absolutas constantes de X ;
- A variação absoluta no valor esperado de Y será diferente para cada valor de X ;
- Por exemplo, uma variação de 0,01 no $\ln(Y)$ significa uma variação de 1% no valor de Y ;



Pelo cálculo diferencial, sabemos que pequenas variações no $\ln Y$ representam variações relativas em Y . Ou seja:

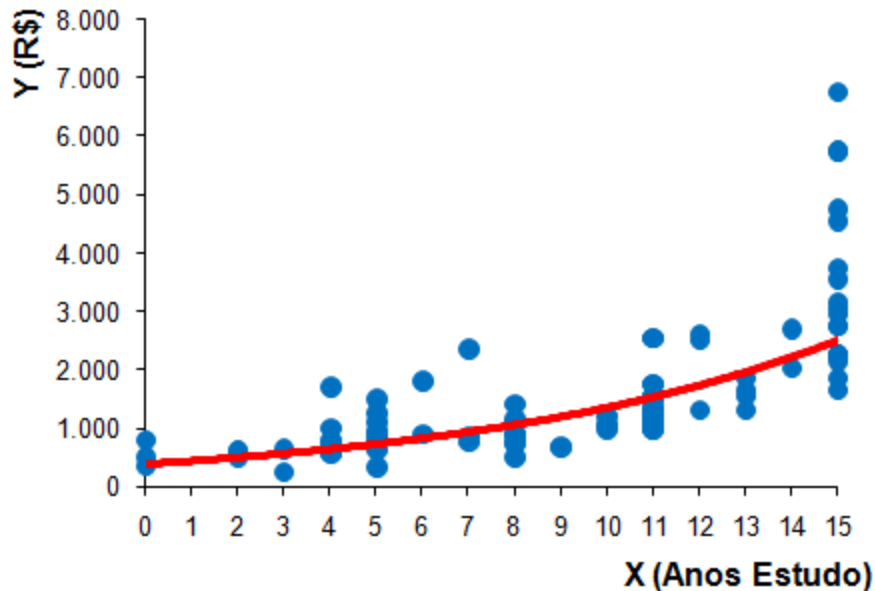
$$\Delta \ln(Y) = \frac{\Delta Y}{Y_i}$$

então...

$$\beta = \frac{\ln(Y_i)}{\Delta X} = \frac{\Delta Y / Y_i}{\Delta X}$$

Modelo Log-Lin - Exemplo

Qual o impacto da escolaridade sobre a renda do trabalho?



Pressupondo que a renda do trabalho (Y , em R\$ mensais) cresça exponencialmente com os anos de escolaridade (X):

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + e_i$$

De onde obtemos:

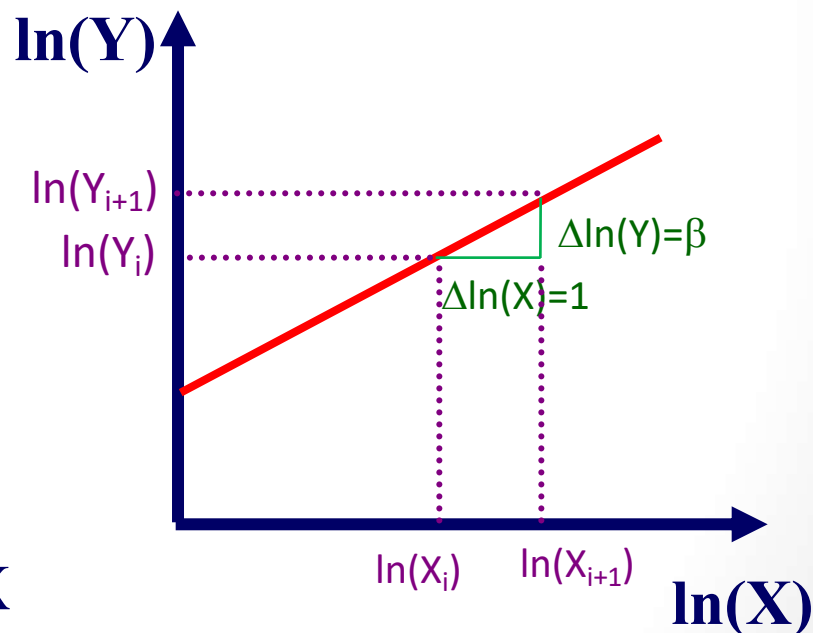
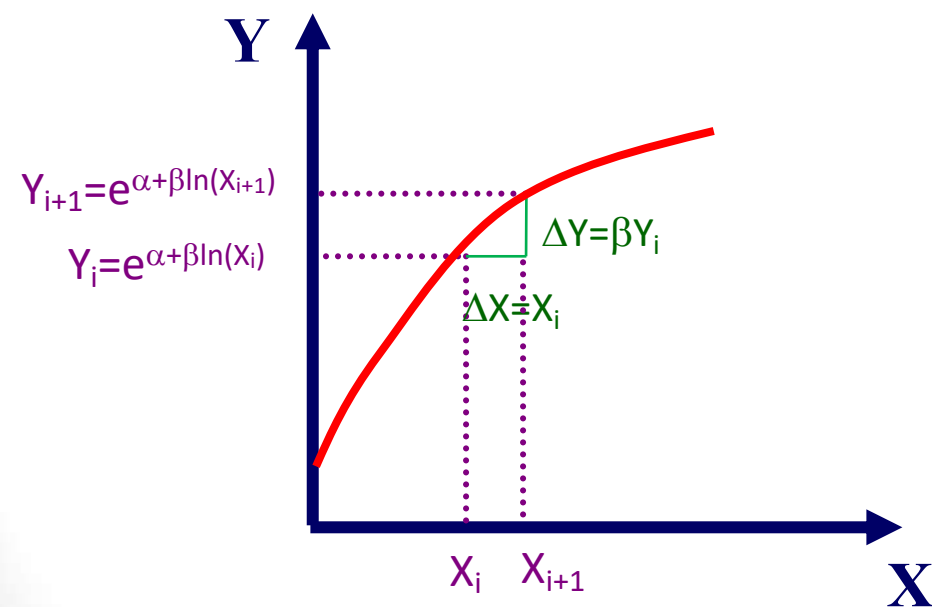
$$\ln(Y_i) = 6,006 + 0,121X_i + \hat{e}_i$$

Espera-se, para cada ano adicional de escolaridade ($\Delta X=1$), um incremento relativo de 12,1% no rendimento do trabalho ($\Delta Y=0,121 Y_i$).

O rendimento esperado para aqueles sem escolaridade ($X=0$) seria de aproximadamente 406 reais ($e^{6,006}$).

Modelo Log-Log - Interpretação

- Um modelo log-log pressupõe que Y apresente variações relativas constantes em função de variações relativas constantes de X ;
- O coeficiente angular será uma medida constante da elasticidade de Y em relação a X , ou seja, qual será a variação percentual em Y (β) para uma variação de 1% em X .



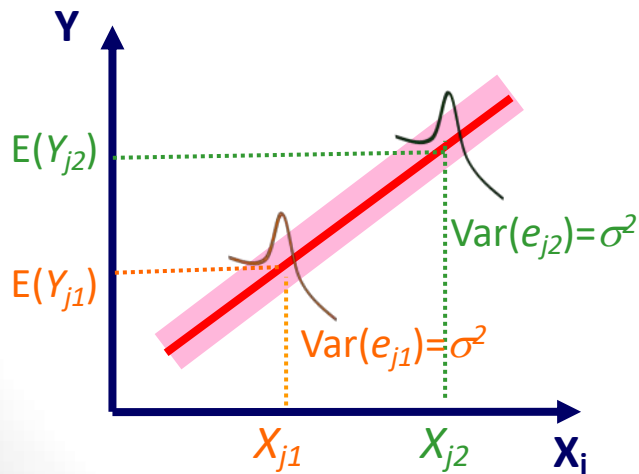
Se $\Delta \ln(X) = \frac{\Delta X}{X_i}$ e $\Delta \ln(Y) = \frac{\Delta Y}{Y_i}$ então

$$\beta = \frac{\Delta \ln(Y)}{\Delta \ln(X)} = \frac{\Delta Y / Y_i}{\Delta X / X_i}$$

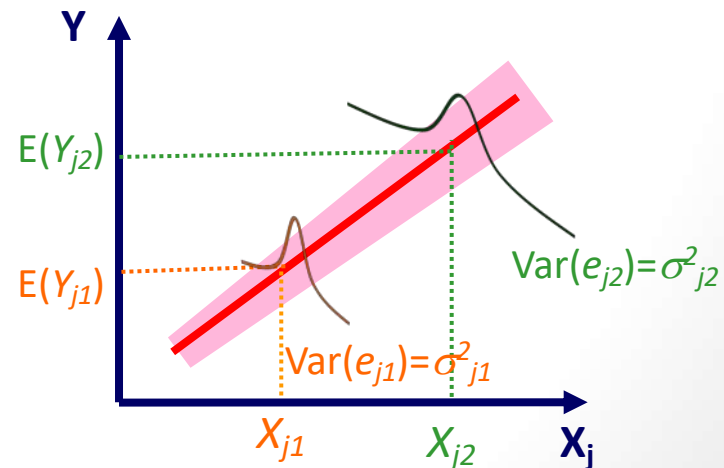
Heterocedasticidade – Definição

- Homocedasticidade: ocorre quando a variância dos erros e , condicionada aos valores das variáveis explanatórias, é constante: $Var(e / X_{1j}, X_{2j}, ..., X_{kj}) = \sigma^2$
- Heterocedasticidade: ocorre quando a dispersão em torno da regressão se altera em função dos valores de uma variável explanatória: $Var(e / X_{1j}, X_{2j}, ..., X_{kj}) = \sigma_j^2$

Homocedasticidade



Heterocedasticidade

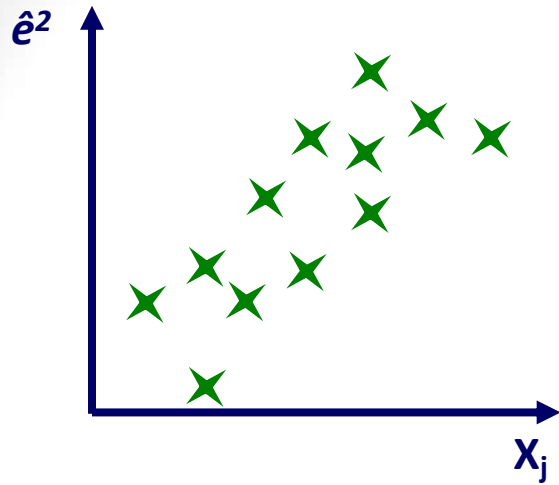


Heterocedasticidade – Causas

Os principais fatores que podem causar a heterocedasticidade:

- 1) Natureza das variáveis:** alguns relacionamentos apresentam tipicamente tendência à heterocedasticidade. Por exemplo, renda e poupança. Pessoas de baixa renda são limitadas pela renda ao poupar, e possuem pouca dispersão em relação aos valores médios de seu grupo econômico. Já entre os ricos o comportamento é mais disperso: há aqueles que poupam boa parcela de seus ricos rendimentos, até aqueles que nada poupam.
- 2) Valores extremos:** a ocorrência de um valor extremo na amostra pode inflacionar a variabilidade em um determinado ponto do ajuste.
- 3) Falhas na especificação do modelo:** a heterocedasticidade pode também ser devida à omissão de importantes variáveis no modelo.
- 4) Transformação dos dados:** a transformação das variáveis (por exemplo, proporção ao invés de valores absolutos) ou da forma funcional (modelo log-duplo ao invés de linear) pode eliminar a heterocedasticidade.

Teste de White



Seja o modelo de RLM com $k=2$:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

Par verificarmos se os resíduos quadráticos têm relação com os regressores, seus quadrados e seus produtos cruzados:

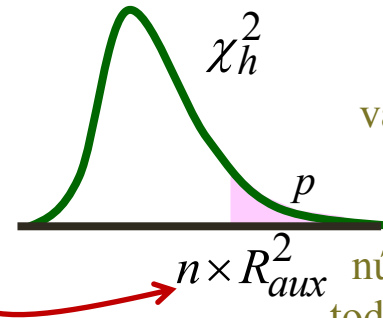
$$\hat{e}_i^2 = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{1i} X_{2i} + \delta_4 X_{1i}^2 + \delta_5 X_{2i}^2 + u_i$$

Para testar a hipótese nula da homocedasticia:

$$\begin{cases} H_0 : \delta_1 = \dots = \delta_5 = 0 \\ H_1 : \delta_j \neq 0 \end{cases}$$



$$LM = n \times R_{aux}^2$$



Onde h é o número de variáveis explanatórias, R_{aux}^2 o coeficiente de determinação e n o número de observações, todos referentes ao ajuste auxiliar

Passos para efetuar o teste de White

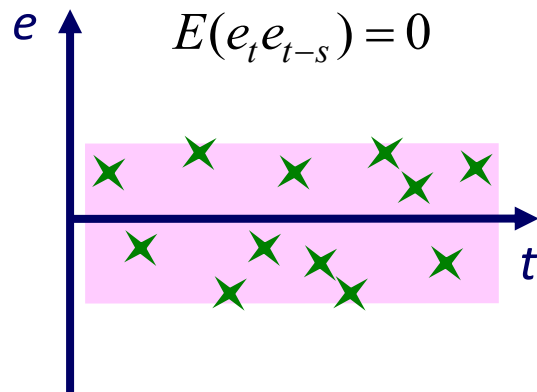
- 1- Estimar os resíduos do ajuste de MQO para o modelo original de RLM;
- 2- Ajustar um modelo auxiliar relacionando o quadrado dos resíduos às variáveis independentes do modelo original, seus quadrados e produtos cruzados;
- 3- Calcular a estatística LM pelo produto do número de observações e o R^2 do ajuste auxiliar;
- 4- Calcular o valor p associado à estatística em uma distribuição χ^2 com gl dado pelo número de variáveis explanatórias do ajuste auxiliar;

Autocorrelação – Definição

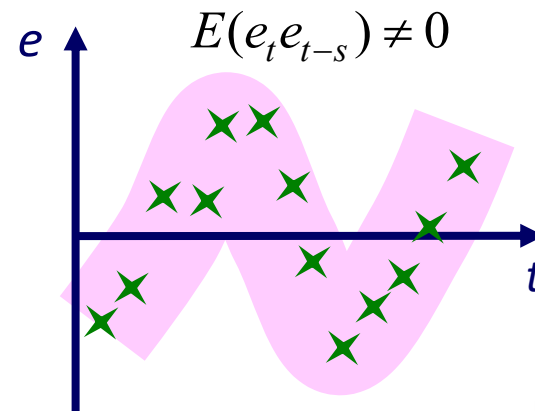
- Correlação de valores de uma mesma variável ordenados, por exemplo, no tempo (com dados de séries temporais) ou no espaço (com dados espaciais). No contexto da regressão, dizemos que há autocorrelação nos termos de erro quando

$$\text{Cov}(e_t, e_{t+s}) = E(e_t e_{t-s}) \neq 0$$

Não Autocorrelacionado



Autocorrelacionado



Correlação Serial – Causas

1) Inércia: é comum séries temporais econômicas apresentarem ciclos, ou seja, períodos de crescimento ou decaimento. Mudanças na tendência costumam ocorrer lentamente. Assim, em análise de regressão, observações sucessivas tendem a ser autocorrelacionadas, ou seja, valores acima da média costumam estar próximos aos períodos de crescimento, assim, como valores abaixo da média estariam próximos aos períodos de recessão.

2) Falhas de especificação: a autocorrelação pode ser devida à ausência de uma importante variável no modelo de regressão ou transformação das variáveis existentes. Os erros expressariam, assim, um padrão sistemático devido à ausência dessas informações. Por exemplo, omitir o preço de um substituto em uma função de demanda pode sujeitar os erros a um comportamento sistemático. Analogamente, ajustar uma função linear a uma relação quadrática também sujeitaria os resíduos a um padrão sistemático.

2) Defasagens: as decisões econômicas em um período t dependem, muitas vezes, de informações defasadas do período $t-1$. Desconsiderar esse tipo de relação sujeitaria os erros à correlação serial. Por exemplo, o consumo depende, além da renda presente, do consumo no período anterior;

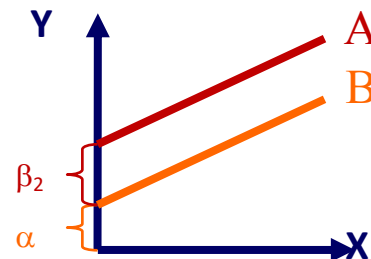
Variáveis Binárias – 2 Categorias

- Para representarmos duas categorias independentes nominais (A e B) em um modelo de regressão, precisamos de apenas uma variável binária D . A referência da análise será dada por $D=0$;

$$Y_i = \alpha + \beta_1 X_i + \beta_2 D_i + e_i$$

O coeficiente β_2 indicaria quanto Y seria, em média, maior (ou menor) para a categoria A ($D=1$) que a categoria de referência B ($D=0$), independente do valor de X .

Categoria	D_i
A	1
B	0



Para A: $Y_i = (\alpha + \beta_2) + \beta_1 X_i + e_i$

Para B: $Y_i = \alpha + \beta_1 X_i + e_i$

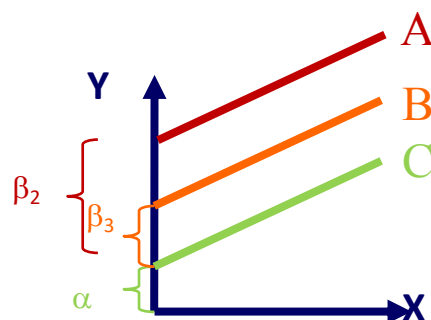
Variáveis Binárias – k Categorias

- Para representarmos k categorias nominais, precisamos de $k-1$ variáveis binárias D . A referência da análise será dada por uma das categorias.

$$Y_i = \alpha + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{2i} + e_i$$

O coeficiente β_2 indicaria quanto Y seria, em média, maior para a categoria A ($D_1=1$) que a categoria de referência C ($D_1=0$ e $D_2=0$), independente do valor de X . O coeficiente β_3 indicaria quanto Y seria, em média, maior para a categoria B ($D_2=1$) que a categoria de referência C.

Categoria	D_{1i}	D_{2i}
A	1	0
B	0	1
C	0	0



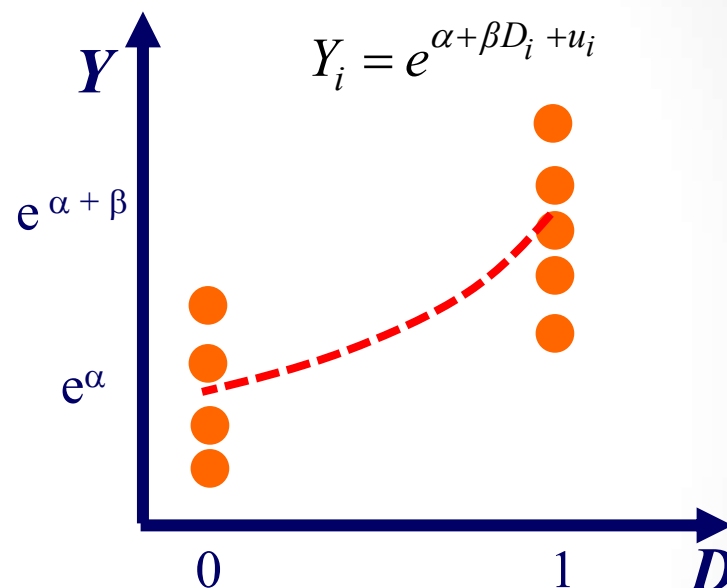
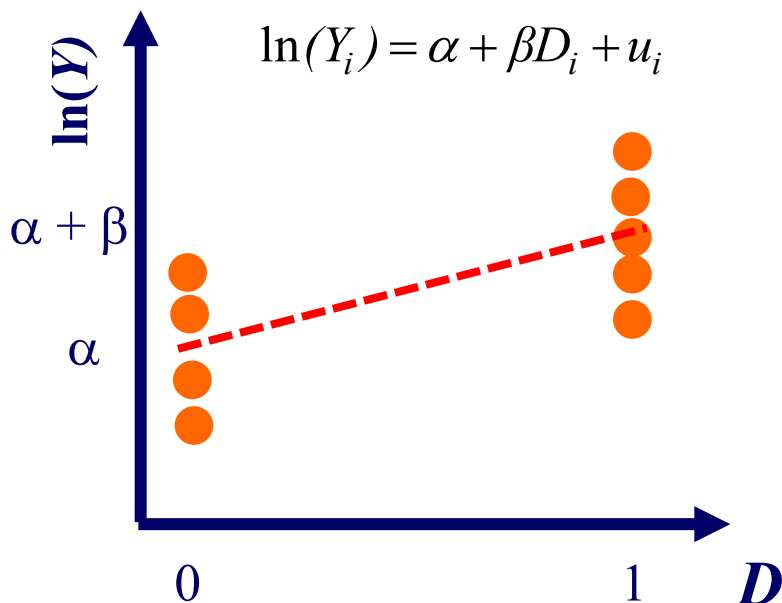
Para A: $Y_i = (\alpha + \beta_2) + \beta_1 X_i + e_i$

Para B: $Y_i = (\alpha + \beta_3) + \beta_1 X_i + e_i$

Para C: $Y_i = \alpha + \beta_1 X_i + e_i$

Binárias em Funções Logarítmicas

Seja a equação semi-logarítmica:



- Seja Y_1 o valor de Y para $D=1$ e Y_0 o valor para $D=0$;
- A interpretação do β associado à binária será dada por::

Para $D=0$:

$$\ln(Y_0) = \alpha$$

$$Y_0 = e^\alpha$$

Para $D=1$:

$$\ln(Y_1) = \alpha + \beta$$

$$Y_1 = e^{\alpha + \beta}$$

Então:

$$\frac{Y_1 - Y_0}{Y_0} = \frac{e^\alpha e^\beta - e^\alpha}{e^\alpha}$$



$$\frac{\Delta Y}{Y} = \frac{Y_1 - Y_0}{Y_0} = e^\beta - 1$$

Exercícios

- 1) O arquivo Data_TravelCosts.xls contém informações sobre o custos médios de deslocamentos de municípios a um parque nacional (MAIA, A. G., ROMEIRO, A. . Validade e confiabilidade do método de custo de viagem: um estudo aplicado ao Parque Nacional da Serra Geral. Revista de Economia Aplicada, v. 12, p. 103-123, 2008):
 - a) Analise a homocedasticidade dos erros;
 - b) Compare a qualidade do ajuste para diferentes especificações da forma funcional;
 - c) Inclua a binária RS, que assume 1 quando o município pertencer ao estado do Rio Grande do Sul (os dois primeiros dígitos da variável *code* devem ser iguais a 43);