Variáveis Binárias

HO 231 – Econometria

Profa. Rosangela Ballini

Instituto de Economia - UNICAMP

Ementa

- Variáveis Binárias para 2 ou *k* Categorias Nominais
- Coeficientes Angulares Interativos
- Binárias para categorias ordinais
- Teste de Mudança Estrutural

Bibliografia

WOLDRIDGE, J.M. (2006). Introductory Econometrics: a Modern Approach. Cap. 7.

Categoria de Variáveis

1) Escala Nominal: Valores representam categorias (*nomes*). Não se pode falar que um seja maior que o outro. Exemplo: sexo.

2) Escala Ordinal: Valores representam uma hierarquia de posições. Não se pode, entretanto, falar quão maior é um valor em relação a outro. Exemplo: classe de renda.

3) Escala Intervalar: Valores representam ordem e é possível mensurar intervalo entre eles. Não se pode, entretanto, dizer quantas vezes um é maior que outro. Exemplo: ano.

4) Escala de razão: Valores representam ordem, é possível mensurar intervalo entre eles e quantificar grandezas em uma escala de razão. Exemplo: renda.

Variáveis Binárias – Definição

Uma variável binária (variável *dummy*) pode representar dois estados possíveis::

$$X = \begin{cases} 0, \text{ ausência da característica de interesse (Fracasso)} \\ 1, \text{ presença da característica de interesse (Sucesso)} \end{cases}$$

Podemos, assim, estimar a influência de variáveis explicativas (independentes) nominais ou ordinais em modelo de regressão, da mesma maneira que fazemos com variáveis quantitativas de escala intervalar ou de razão.

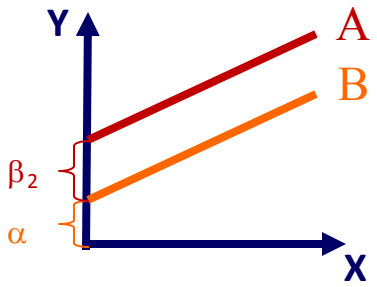
Variáveis Binárias – 2 Categorias

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

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

Categoria	D_i
A	1
B	0

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



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

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

Binária com 2 Categorias - Exemplo

Estime o modelo:

$$co2_i = \beta_0 + \beta_1 D_i + \beta_3 seto2_i + u_i$$

Em que D_i representa a variável binária *alta dada por:*

$$D_i = \begin{cases} 1, se\ PIB\ per\ capita\ maior\ ou\ igual\ a\ US\$\ 12\ mil \\ 0, caso\ contrário \end{cases}$$

Coefficients:	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-0.47168	0.68429	-0.689	0.492
alta	8.15935	0.72254	11.293	< 2e-16 ***
setor2	0.10560	0.02063	5.119	8.4e-07 ***

Mantendo-se constante a participação da indústria no PIB, os países de renda alta emitem, em média, 8,2 ton per capita de CO_2 a mais que os demais países.
Binária é uma variável nominal do pib, podendo ser redundante a incorporação desta última variável.
Especificação adequada, depende, entretanto, dos pressupostos da análise: **as emissões crescem linearmente com a renda ou podem simplesmente ser classificadas em dois grupos de renda?**

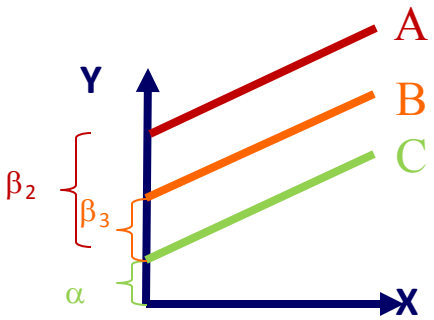
Variáveis Binárias – *k* Categorias

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

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

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

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.



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

Binária com 3 Categorias - Exemplo

- Podemos discriminar 3 categorias de renda (alta: 12 mil ou mais; média=4 a 12 mil; baixa=menos de 4 mil) com 2 binárias;

  Construa as binárias

  alta se 1 para PIB >= 12 mil; 0 c.c.

  media se 1 para PIB entre 4 e 12 mil; 0 c.c.

  O grupo de baixa renda (menos de 4 mil) será a referência de análise.

- Em seguida, ajuste o modelo incorporando as binárias como regressores, ou seja,

$$co2_i = \beta_0 + \beta_1 setor2_i + \beta_3 alta_i + \beta_4 media_i + u_i$$

Binária com 3 Categorias - Exemplo

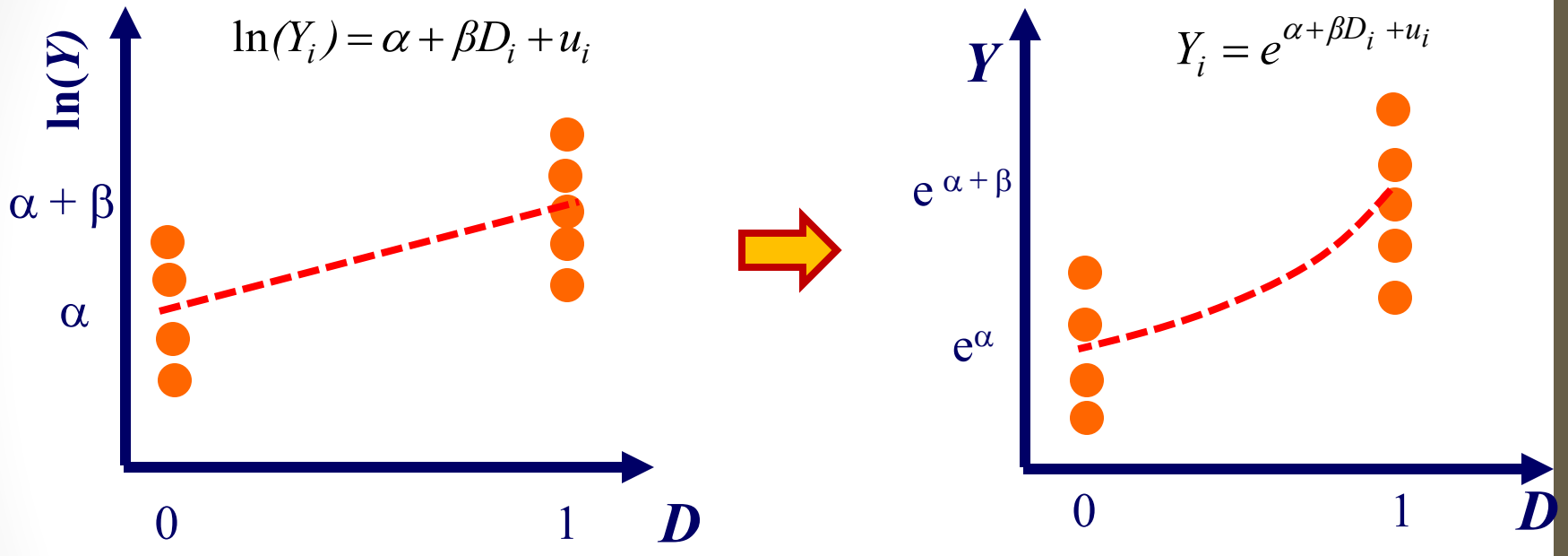
Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.99758	0.62561	-1.595	0.113
alta	9.05570	0.67052	13.505	< 2e-16 ***
media	3.97211	0.64931	6.117	6.67e-09 ***
setor2	0.09279	0.01880	4.937	1.93e-06 ***

Mantendo-se constante a participação da indústria, os países de renda alta emitem, em média, 9,1 ton per capita a mais de CO_2 que os países de renda baixa.

Por sua vez, os países de renda média emitem, em média, 4,0 ton per capita a mais que os países de renda baixa.

Binárias em Funções Logarítmicas

Seja a equação semi-logarítmica:



- Seja:

  Y_1 o valor de Y para D=1;

  Y_0 o valor de Y para D=0.

Binárias em Funções Logarítmicas

- A interpretação de β associada à binária é 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:

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

Log-Lin com Binárias - Exemplo

- Ajuste o modelo:

$$\log(co2_i) = \beta_0 + \beta_1 D_i + \beta_3 seto2_i + u_i$$

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.132564	0.251960	-4.495	1.30e-05 ***
alta	2.181787	0.266045	8.201	6.23e-14 ***
setor2	0.041874	0.007595	5.513	1.32e-07 ***

Independente da participação da indústria no PIB, as emissões per capita de CO_2 dos países de renda alta são, em média, quase 8 vezes superior

$$e^{2,182} - 1 = 7,862 = 786,2\%$$

à dos demais países.

Log-Lin com Binárias - Exemplo

- Raciocínio análogo é válido com *k* binárias para representar *k*+1 categorias nominais:

$$\log(co2_i) = \beta_0 + \beta_1 setor2_i + \beta_3 alta_i + \beta_4 media_i + u_i$$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.362877	0.219253	-6.216	4.02e-09	***
alta	2.574336	0.234993	10.955	< 2e-16	***
media	1.739567	0.227558	7.644	1.64e-12	***
setor2	0.036266	0.006587	5.505	1.39e-07	***

Mantendo-se constante a participação da indústria no PIB, as emissões per capita de CO_2 dos países de renda alta são, em média, mais de 12 vezes superior ($e^{2,574} - 1 = 12,123 = 1.212,3\%$) às dos países de renda baixa.

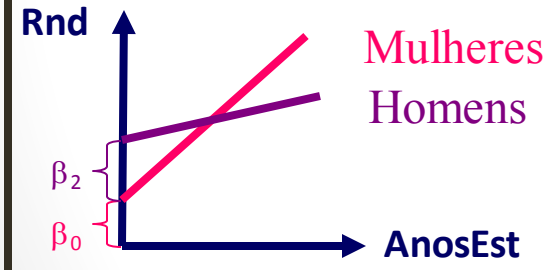
As dos países de renda média são, em média, 4,7 vezes superior às dos países de renda baíxa ($e^{1,740} - 1 = 4,695 = 469,5\%$).

Coeficiente Angular Interativo

$$Rnd = \beta_0 + \beta_1 Masc + \beta_2 AnosEst + e$$

Esse modelo pressupõe deslocamentos da função de rendimentos (β_1) mas retornos marginais da escolaridade (β_2) iguais para homens e mulheres.

$$Rnd = \beta_0 + \beta_1 Masc + \beta_2 AnosEst + \beta_3 Masc \cdot AnosEst + e$$



Esse modelo pressupõe desolocamentos da função de rendimentos (β_1) e retornos marginais da escolaridade diferentes para homens (β_2+β_3) e mulheres (β_2).

Efeito de Interação - Exemplo

Supondo que o efeito da indústria sobre as emissões seja diferente entre os grupos de países, estime o modelo:

$$\log(co2_i) = \beta_0 + \beta_1 setor2_i + \beta_3 alta_i + \beta_4 setor2_i * alta_i + u_i$$

A variável *setor2*alta* captará o efeito de interação entre a participação da indústria e o grupo de países de renda alta.

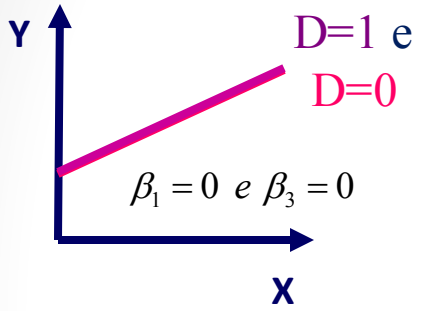
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.218660	0.266991	-4.564	9.74e-06	***
alta	2.818869	0.704934	3.999	9.59e-05	***
setor2	0.044759	0.008151	5.491	1.48e-07	***
I(alta * setor2)	-0.021930	0.022471	-0.976	0.331	

Países que não são de renda alta (*alta*=0), o impacto parcial de um acréscimo percentual na participação da indústria no PIB será, em média, de 4,5% nas emissões. Estimativas sugerem que, para os países de renda alta (*alta*=1), o impacto parcial seria, em média, de 2,3% (4,5% - 2,2%).
A diferença entre o impacto dos países de renda alta e dos demais países não se mostrou significativa.

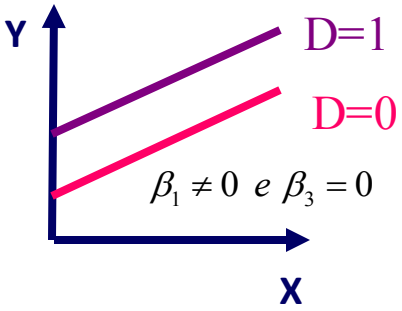
Mudança Estrutural - Conceito

Seja as seguintes hipóteses para a relação entre *Y*, *X* e a binária *D*:

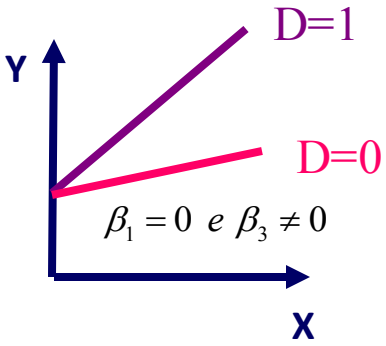
$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i \cdot X_i + e_i$$



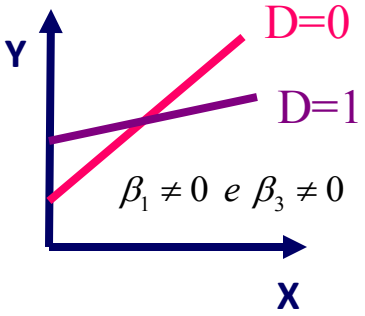
Regressões Coincidentes



Regressões Paralelas



Regressões Concorrentes



Regressões Dissimilares

Testar se há mudança estrutural significa testar se pelo menos um dos coeficientes β_1 ou β_3 é diferente de zero:

$$\begin{cases} H_0: \ \beta_1 = \beta_3 = 0 \\ H_1: \ \beta_1 \ e/ou \ \beta_3 \neq 0 \end{cases}$$

Esse teste corresponde à contribuição marginal das variáveis associadas aos coeficientes β_1 e β_3.

Teste de Mudança Estrutural

- $SQReg$ devido às variáveis X_1, D e $X_1 * D$ (Modelo Irrestrito):

$$SQReg(Y|X_1, D \text{ e } X_1 * D) \text{ ou } SQReg_I$$

✓ Variabilidade da variável dependente (Y) explicada pelo conjunto das variáveis independentes (X_1, D e $X_1 * D$)

Graus de liberdade: k=3 (igual ao número de regressores do modelo)

- $SQReg$ devido à variável X_1 (Modelo Restrito):

$$SQReg(Y|X_1) \text{ ou } SQReg_R$$

- ✓ Variabilidade da variável dependente explicada exclusivamente por X_1

Grau de liberdade: k=1 (igual ao número de regressor do modelo)

- Contribuição de $D, X_1 * D$:

$$Contribuição\ D, X_1 * D = SQReg_I - SQReg_R$$

✓ Variabilidade de Y explicada por D e $X_1 * D$, após considerada a variabilidade explicada por X_1

Grau de liberdade: q=2 (novos coeficientes de regressão incorporados ao modelo $(\beta_1$ e $\beta_3)$.

Teste de Mudança Estrutural - Definição

Seja o modelo *irrestrito* de RLM:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e$$

Para verificarmos se a contribuição de um grupo de q variáveis é significativa no modelo elaboramos um modelo com restrição aos parâmetros. Suponha que, por simplicidade, as q variáveis que desejamos testar são as últimas das k variáveis do modelo irrestrito (a ordem, obviamente, não faz importância).
Modelo *restrito* é dado por:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-q} X_{k-q} + e$$

Em outras palavras, estaríamos interesados em testar a hipótese nula de que os q coeficientes do modelo irrestrito são nulos:

$$H_0 : \beta_{k-q+1} = 0, \ \ldots, \ \beta_k = 0$$

Teste de Mudança Estrutural - Definição

Seja o modelo *irrestrito*:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u$$

Para verificarmos se a contribuição de um grupo de q variáveis é significativa no modelo, devemos inserir no modelo as restrições nos parâmetros.

Vamos supor que o modelo *restrito* seja dado por:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-q} X_{k-q} + u$$

Ou seja, estamos interessados em testar a seguinte hipótese nula:

$$H_0 \colon \beta_{k-q+1} = 0, \cdots, \beta_k = 0$$

Teste de Mudança Estrutural - Definição

O teste estatístico consiste em verificar se a contribuição marginal dessas q variáveis é significativa, ou seja,

$$F = \frac{(SQReg_I - SQReg_R)/q}{SQRes_I/(n-p)}$$

Ou,

$$F = \frac{(SQRes_R - SQRes_I)/q}{SQRes_I/(n-p)}$$

Rejeita-se H_0 se o valor de F calculado exceder o valor crítico para um dado nível de significância com *q e (n-p)* graus de liberdade(ou p-valor for suficientemente pequeno).

Mudança Estrutural - Exemplo

Considere o modelo irrestrito:

$$\log(co2_i) = \beta_0 + \beta_1 setor2_i + \beta_3 alta_i + \beta_4 setor2_i * alta_i + u_i$$

Testar se a contribuição marginal das variáveis *alta* e *setor2Xalta* é nula.

```
Model 1: restricted model
Model 2: log(co2) ~ alta + setor2 + interacao

   Res.Df    RSS      Df   Sum of Sq      F       Pr(>F)
1    167    396.36
2    165    280.46    2       115.9    34.093  4.045e-13 ***
```

Se afirmarmos que a variável *alta* ou sua interação com a variável *setor2* contribuam para explicar o logaritmo das emissões de carbono, estaremos sujeitos a um erro praticamente nulo.

Ou seja, há quebra estrutural na relação entre emissões de carbono e participação da indústria. A partir dos testes *t* para os coeficientes do modelo, podemos afirmar que há deslocamento do intercepto. Mas não podemos afirmar que haja mudança no efeito parcial da participação da indústria. Assim, sugere-se que as regressões sejam paralelas.

Exercícios

1) A partir do arquivo ***WAGE2.xlsx***:
   a) Ajuste um modelo linear para a renda (*wage*) como função dos anos de educação (*educ*), quociente de inteligência (*IQ*) e experiência profissional (*exper*). Considere ainda binárias para discriminar a cor (*black=1*) e estado civil (*married=1*);
   b) Ajuste um modelo log-linear a renda (*wage*) como função dos mesmos regressores;
   c) Crie duas binárias para discriminar 3 categorias de escolaridade: até 11 anos de estudo (referência); 12 a 15 anos (*educ2*), 16 anos ou mais (*educ3*). Incorpore essas variáveis no modelo de regressão log-linear;
   d) Incorpore a interação entre a cor e experiência profissional no modelo log-linear. Qual pressuposto teórico estaria sendo testado?
   e) Realize um teste de mudança estrutural para a relação entre renda e experiência profissional segundo cor do trabalhador;
   f) A partir de um teste de especificação, qual modelo é mais adequado?

Exercícios

2) A partir do arquivo **Dados_PIB.xlsx**, que contém informações do WDB sobre sobre gastos com P&D, PIB per capita, razão de dependência para jovens e gastos com educação dos países em 2010, pede-se:
   a) Considere no modelo log-lin uma binária para discriminar os países do G7;
   b) Considere no modelo log-lin duas binárias para discriminar os países do G7 e BRICS;
   c) Incorpore uma interação entre PIB e G7;
   d) Realize um teste de mudança estrutural;
   e) Aplique um teste de especificação para avaliar os modelos ajustados. Qual modelo mostra-se mais adequado?