Insper | Instituto de Ensino e Pesquisa

Modelo Linear Geral V

Aula 10

Heij et al., 2004 – Capítulo 5

Wooldridge, 2011 (4. ed) – Capítulo 7

ANÁLISE DE REGRESSÃO LINEAR MÚLTIPLA COM INFORMAÇÃO QUALITATIVA: O USO DA VARIÁVEL *DUMMY*

Variável *Dummy*

Uma forma de introduzir características qualitativas em modelos econométricos consiste na utilização de variáveis *dummy* (fictícia, postiça), frequentemente chamadas de variáveis binárias ou dicotômicas, uma vez que assumem apenas um de dois valores – em geral 0 ou 1 – para indicar a presença ou ausência de determinada característica.

Variável *Dummy*

Vale lembrar que a variável *dummy* representa estados ou níveis de fatores, ou seja representa algo que não possui valores numéricos ou, caso possua, estes valores não têm realmente um significado numérico.

Assim, uma variável *dummy*, D, pode ser descrita da seguinte maneira:

$$D = \begin{cases} 0, & \text{se a característica não estiver presente} \\ 1, & \text{se a característica estiver presente} \end{cases}$$

Voltando à Empresa TEMCO

A senhorita Rose Jolie, gerente do departamento de RH da empresa TEMCO, gostaria de estimar os parâmetros de um modelo de regressão linear que levasse em consideração as variáveis explicativas *educ* e *dept* na explicação da variável resposta *salário*. Auxilie a senhorita Jolie nesta proposição.

Voltando à Empresa TEMCO

Apenas para lembrar, a senhorita Jolie, coletou informações de uma amostra aleatória de 46 funcionários da empresa, sobre as seguintes variáveis:

id – número cadastral do funcionário;

salario – anual, em dólares;

anosemp – tempo (em anos) na empresa;

expprev – experiência anterior (em anos);

educ – anos de estudo após o segundo grau;

sexo – (feminino = 0, masculino = 1);

dept – departamento no qual o funcionário atua

(Compras = 1, Engenharia = 2, Propaganda = 3, Vendas = 4);

super – número de empregados sob responsabilidade do empregado.

Voltando à Empresa TEMCO

À primeira vista, como existem quatro departamentos na empresa *TEMCO*, Rose Jolie poderia optar por usar a variável *dept*, com os valores 1, 2, 3 e 4.

Dessa maneira,

$$salário = \beta_1 + \beta_2 educ + \beta_3 dept + \varepsilon$$

No entanto, ao fazer isto, Rose Jolie estaria introduzindo uma ideia de espaçamento, que ficará mais clara nos resultados descritos nos *slides* a seguir.

Voltando à Empresa TEMCO

Escrevendo a equação de regressão de interesse, para cada um dos departamentos, temos que:

$$E(salário|educ, dept = 1) = (\beta_1 + \beta_3) + \beta_2 educ$$

$$E(salário \mid educ, dept = 2) = (\beta_1 + 2\beta_3) + \beta_2 educ$$

$$E(salário \mid educ, dept = 3) = (\beta_1 + 3\beta_3) + \beta_2 educ$$

$$E(salário \mid educ, dept = 4) = (\beta_1 + 4\beta_3) + \beta_2 educ$$

Voltando à Empresa TEMCO

Dessa forma, admitiríamos, por exemplo, que

$$E(salário \mid educ, dept = 2) - E(salário \mid educ, dept = 1) =$$

$$= E(salário \mid educ, dept = 4) - E(salário \mid educ, dept = 3) =$$

$$= \beta_3$$

ou seja, que a diferença entre os salários esperados dos funcionários dos departamentos de Engenharia e Compras é a mesma que a dos funcionários dos departamentos de Propaganda e Engenharia, mantendo constante o tempo de escolaridade.

Voltando à Empresa TEMCO

Assim, se Rose Jolie utilizasse *dept* da forma como foi construída, então ela estaria impondo uma restrição ao modelo, que não sabemos se é real.

Ainda, se a ordem das categorias da variável departamento fosse alterada, estaríamos propondo um novo conjunto de restrições ao modelo, o que muito provavelmente nos levaria a resultados completamente diferentes do caso anterior.

Voltando à Empresa TEMCO

Portanto, o ideal seria utilizar um grupo de variáveis que representasse os estados de interesse, que no nosso caso não apresentam nenhuma ordenação natural, de tal sorte a nunca alterar o resultado final, qualquer que seja o critério de criação adotado para a construção destas variáveis.

Variável *Dummy*

A solução é, portanto, trabalharmos com algumas variáveis *dummy*.

No geral, se temos *p* estados, devemos trabalhar com *p – 1* variáveis *dummy*.

Variável *Dummy*

Para o nosso exemplo, poderíamos definir as variáveis *dummy* D_C, D_E e D_P da seguinte maneira, para representar os estados da variável departamento:

dept	D_C	D_E	D_P
Compras	1	0	0
Engenharia	0	1	0
Propaganda	0	0	1
Vendas	0	0	0

Variável *Dummy*

Assim, partindo do modelo de regressão linear

$$y_i = \beta_1 + \beta_2\, educ_i + \delta_1\, D_{Ci} + \delta_2\, D_{Ei} + \delta_3\, D_{Pi} + \varepsilon_I$$

temos que:

Compras: $y_i = (\beta_1 + \delta_1) + \beta_2 educ_i + \varepsilon_i$

Engenharia: $y_i = (\beta_1 + \delta_2) + \beta_2 educ_i + \varepsilon_i$

Propaganda: $y_i = (\beta_1 + \delta_3) + \beta_2 educ_i + \varepsilon_i$

Vendas: $y_i = \beta_1 + \beta_2\, educ_i + \varepsilon_i$

Variável *Dummy*

Do *slide* 14, o parâmetro δ_1, por exemplo, pode ser interpretado como a diferença esperada entre os salários dos profissionais das áreas de Compras e Vendas, que apresentam o mesmo tempo de escolaridade.

Ainda, vale lembrar que, estamos admitindo que o acréscimo médio no salário correspondente ao acréscimo em um ano de escolaridade é o mesmo para os quatro departamentos.

Variável *Dummy*

Variáveis binárias como D_C, D_E e D_P, que são incorporadas num modelo de regressão para dar conta de um deslocamento do intercepto como resultado de algum fator qualitativo, são chamadas de variáveis binárias de intercepto ou, simplesmente, variáveis *dummy* de intercepto.

Variável *Dummy*

Como criar variáveis *dummy* no Eviews?

Exemplo

(criação da variável D_C)

(i) Clicar em QUICK;

(ii) Depois em GENERATE SERIES;

(iii) Digitar DC=(dept=1).

O que aconteceu ao realizar o procedimento anterior?

Voltando à Empresa TEMCO

Estimação dos Parâmetros do Modelo de Interesse

Dependent Variable: SALARIO
Method: Least Squares
Date: 03/07/12 Time: 12:32
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*DC+C(4)*DE+C(5)*DP

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	19235.72	2662.597	7.224419	0.0000
C(2)	2952.957	341.8007	8.639412	0.0000
C(3)	5393.973	3070.062	1.756959	0.0864
C(4)	8065.517	2484.109	3.246845	0.0023
C(5)	6664.357	3181.833	2.094502	0.0424

R-squared	0.686058	Mean dependent var	39827.39
Adjusted R-squared	0.655429	S.D. dependent var	10999.24
S.E. of regression	6456.572	Akaike info criterion	20.48591
Sum squared resid	1.71E+09	Schwarz criterion	20.68467
Log likelihood	-466.1759	Hannan-Quinn criter.	20.56037
F-statistic	22.39933	Durbin-Watson stat	1.621506
Prob(F-statistic)	0.000000		

$$\hat{salario} = 19235{,}72 + 2952{,}96 \cdot educ + 5393{,}97 \cdot D_C + 8065{,}52 \cdot D_E + 6664{,}36 \cdot D_P$$

Voltando à Empresa TEMCO

$$\hat{y}_{vendas} = 19235{,}72 + 2952{,}96 \cdot educ$$

$$\hat{y}_{compras} = 24629{,}69 + 2952{,}96 \cdot educ$$

$$\hat{y}_{engenharia} = 27301{,}24 + 2952{,}96 \cdot educ$$

$$\hat{y}_{propaganda} = 25900{,}08 + 2952{,}96 \cdot educ$$

Interprete as estimativas dos parâmetros

Observação 1

Vale recordar que a escolha dos valores de D_C, D_E e D_V não é única. Entretanto, qualquer que seja a escolha, os resultados finais da estimação deverão ser sempre os mesmos.

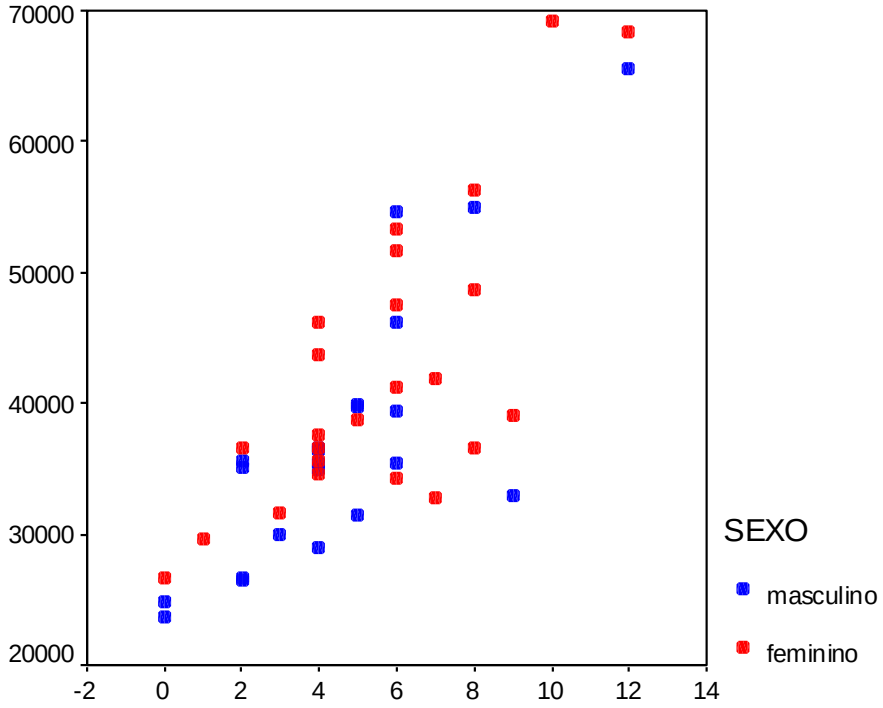
Observação 2

INTERPRETAÇÃO DOS COEFICIENTES LIGADOS ÀS VARIÁVEIS *DUMMY*

Correspondem à diferença em relação ao valor do intercepto e, portanto, à categoria que ele representa ("*benchmark*", ou categoria de referência)

Exercício

Num modelo de regressão linear que já que acomodou *educ* como variável explicativa para *salário*, seria interessante inserir a variável *sexo* em tal modelo?



Exercício (cont.)

Sexo	D_S
Masculino	1
Feminino	0

Modelo:

$$y_i = \beta_1 + \beta_2 \, educ_i + \beta_3 \, D_{Si} + \varepsilon_i$$

Feminino: $y_i = \beta_1 + \beta_2 educ_i + \varepsilon_i$

Masculino: $y_i = (\beta_1 + \beta_3) + \beta_2 educ_i + \varepsilon_i$

Exercício (cont.)

Estimação dos Parâmetros do Modelo de Interesse

Dependent Variable: SALARIO
Method: Least Squares
Date: 03/07/12 Time: 12:37
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*DS

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	26040.75	2529.704	10.29399	0.0000
C(2)	2933.164	374.0873	7.840853	0.0000
C(3)	-2238.262	2103.155	-1.064240	0.2932

R-squared	0.613885	Mean dependent var	39827.39
Adjusted R-squared	0.595927	S.D. dependent var	10999.24
S.E. of regression	6991.865	Akaike info criterion	20.60588
Sum squared resid	2.10E+09	Schwarz criterion	20.72513
Log likelihood	-470.9351	Hannan-Quinn criter.	20.65055
F-statistic	34.18294	Durbin-Watson stat	1.329805
Prob(F-statistic)	0.000000		

Exercício (cont.)

Forma usual

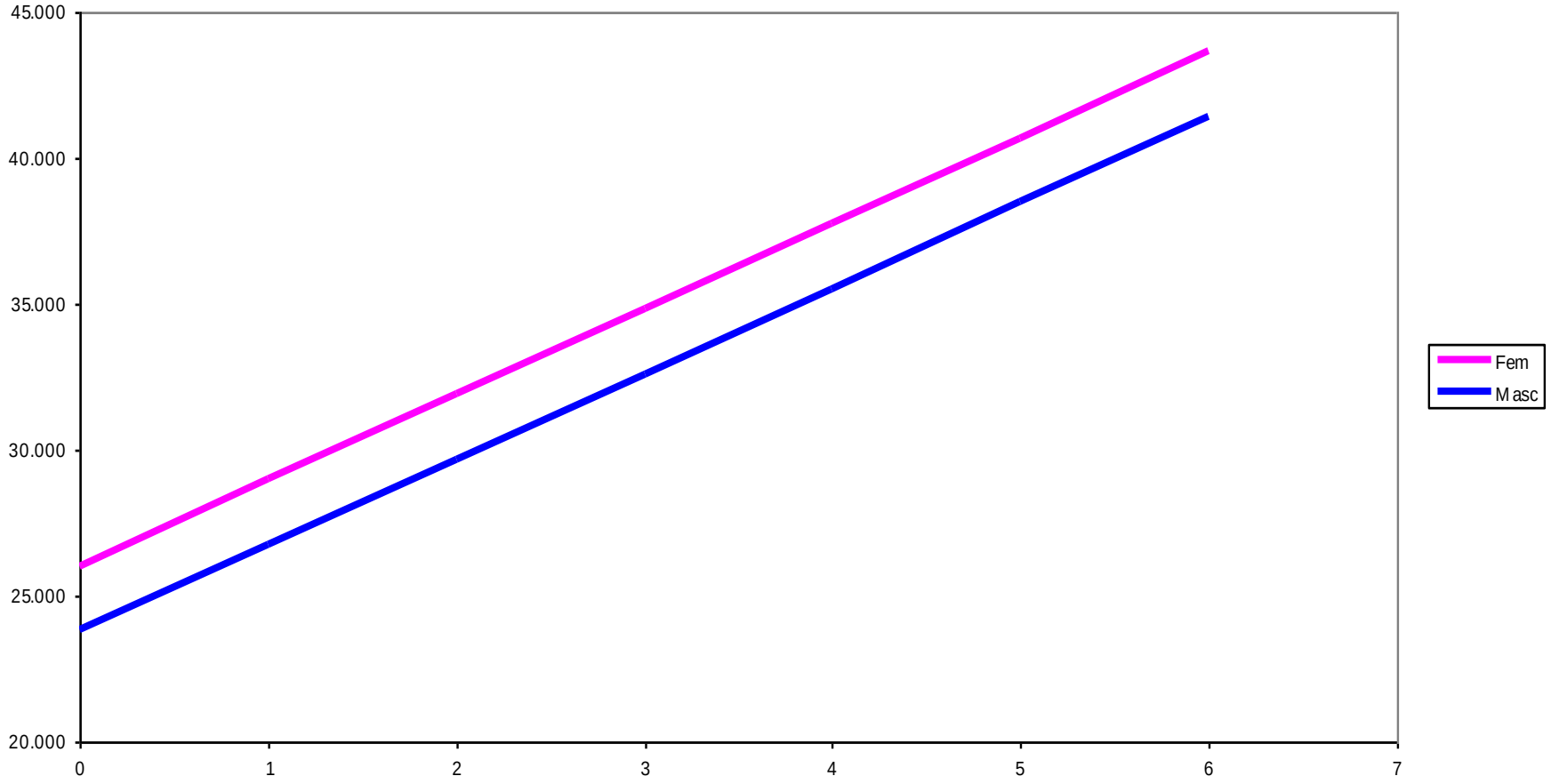
$$\hat{salário} = 26040{,}75 + 2933{,}16 \cdot educ - 2238{,}26 \cdot D_S$$

$$\hat{y}_{fem} = 26040{,}75 + 2933{,}16 \cdot educ$$

$$\hat{y}_{masc} = 23802{,}49 + 2933{,}16 \cdot educ$$

Interprete as estimativas dos parâmetros

Modelo estimado com *EDUC* e *SEXO*



Deste modo, estamos admitindo que a reta de regressão do salário em função da educação para homens é paralela à reta de regressão para as mulheres.

Variável *Dummy* de Inclinação

Variável *Dummy* de Inclinação

No exemplo anterior, utilizando variáveis *dummy* de intercepto, ajustamos quatro retas com a mesma inclinação e diferentes interceptos.

Veremos agora como podemos ajustar um modelo mais geral, no qual, por exemplo, também as inclinações podem ser distintas.

Variável *Dummy* de Inclinação

Sejam D_C, D_E e D_P as variáveis *dummy* do exemplo anteriormente citado.

Considere, ainda, o seguinte modelo

$$y = \beta_1 + \beta_2 educ +$$

$$+ D_C(\delta_0 + \delta_1 educ) + D_E(\delta_2 + \delta_3 educ) + D_P(\delta_4 + \delta_5 educ) + \varepsilon$$

Variável *Dummy* de Inclinação

Assim, para cada um dos departamentos, teríamos os seguintes modelos de regressão:

$$y_{vendas} = \beta_1 + \beta_2 educ + \varepsilon$$

$$y_{compras} = (\beta_1 + \delta_0) + (\beta_2 + \delta_1) educ + \varepsilon$$

$$y_{engenharia} = (\beta_1 + \delta_2) + (\beta_2 + \delta_3) educ + \varepsilon$$

$$y_{propaganda} = (\beta_1 + \delta_4) + (\beta_2 + \delta_5) educ + \varepsilon$$

Variável *Dummy* de Inclinação

Ou seja, o modelo de regressão linear

$$y = \beta_1 + \beta_2\, educ + D_C(\delta_0 + \delta_1 educ) +$$

$$+ D_E(\delta_2 + \delta_3 educ) + D_P(\delta_4 + \delta_5 educ) + \varepsilon$$

faz com que sejam ajustadas quatro retas com interceptos e inclinações diferentes.

Variável *Dummy* de Inclinação

Observe que o modelo anterior pode ser reescrito como

$$y = \beta_1 + \beta_2 educ + \delta_0 D_C + \delta_2 D_E + \delta_4 D_P +$$

$$+ \delta_1 educ D_C + \delta_3 educ D_E + \delta_5 educ D_P + \varepsilon$$

Donde, não é difícil observar que os parâmetros associados às variáveis *dummy* D_C, D_E e D_P, isoladamente, serão responsáveis pela alteração dos interceptos.

Ainda, os parâmetros associados aos produtos de D_C, D_E e D_P por *educ* serão responsáveis pela alteração dos coeficientes angulares.

Variável *Dummy* de Inclinação

Finalmente, as variáveis $educD_C$, $educD_E$ e $educD_P$ são chamadas de variáveis de interação, pois são responsáveis por capturar o efeito de interação entre a escolaridade e departamento sobre o salário. Traduzindo, o impacto na variação do salário esperado de indivíduos de setores diferentes, dada a variação de um ano na escolaridade desses indivíduos, podem ser diferentes.

Variável *Dummy* de Inclinação

Modelo Estimado

Dependent Variable: SALARIO
Method: Least Squares
Date: 03/07/12 Time: 12:42
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*DC+C(4)*DE+C(5)*DP+C(6)*EDUC
*DC+C(7)*EDUC*DE+C(8)*EDUC*DP

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	28013.06	3701.785	7.567446	0.0000
C(2)	1197.488	638.6186	1.875123	0.0685
C(3)	-8891.294	6797.504	-1.308023	0.1987
C(4)	-3898.900	4498.174	-0.866774	0.3915
C(5)	-1738.333	6464.406	-0.268908	0.7895
C(6)	3014.423	1369.996	2.200315	0.0339
C(7)	2347.757	758.7415	3.094277	0.0037
C(8)	1680.538	1153.614	1.456759	0.1534

R-squared	0.755191	Mean dependent var	39827.39
Adjusted R-squared	0.710095	S.D. dependent var	10999.24
S.E. of regression	5922.305	Akaike info criterion	20.36761
Sum squared resid	1.33E+09	Schwarz criterion	20.68563
Log likelihood	-460.4550	Hannan-Quinn criter.	20.48674
F-statistic	16.74617	Durbin-Watson stat	2.014196
Prob(F-statistic)	0.000000		

Variável *Dummy* de Inclinação

Resultado da estimação com *EDUC*, *DEPT* e *interações*

$$\hat{y}_{vendas} = 28013{,}06 + 1197{,}49 \cdot educ$$

$$\hat{y}_{compras} = 19121{,}77 + 4211{,}91 \cdot educ$$

$$\hat{y}_{engenharia} = 24114{,}16 + 3545{,}25 \cdot educ$$

$$\hat{y}_{propaganda} = 26274{,}73 + 2878{,}03 \cdot educ$$

Interprete as estimativas dos parâmetros

Observação

As quatro retas ajustadas simultaneamente, neste exemplo, são equivalentes às retas que obteríamos se ajustássemos separadamente um modelo para cada departamento.

No entanto, este procedimento tem a vantagem de facilitar a construção dos testes de hipóteses envolvendo simultaneamente parâmetros das quatro retas.

EXERCÍCIO PARA ENTREGA

Ajuste um modelo de regressão para a variável *salário* que contenha as variáveis explicativas *educ*, *anosemp*, *sexo* e *dept.* Inclua, ainda, neste modelo todas as interações de primeira ordem. Escreva o modelo estimado e interprete os resultados.