Insper | Instituto de Ensino e Pesquisa

Análise de Regressão Linear Múltipla VII

Aula 10

Heij et al., 2004 – Seções 3.2 e 3.4

Hipótese Linear Geral

Seja

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

um modelo de regressão linear múltipla, que pode ser escrito na forma linear geral, dada por

$$\underset{\sim}{\mathbf{y}} = \underset{\sim}{\mathbf{X}}\,\underset{\sim}{\boldsymbol{\beta}} + \underset{\sim}{\boldsymbol{\varepsilon}}$$

Hipótese Linear Geral

Ainda, baseando-se no modelo anterior, uma hipótese formulada como

$$H_0 : \underset{\sim}{\mathbf{R}}\,\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}} \Leftrightarrow \underset{\sim}{\mathbf{R}}\,\underset{\sim}{\boldsymbol{\beta}} - \underset{\sim}{\mathbf{r}} = \underset{\sim}{\mathbf{0}},$$

em que

- $\underset{\sim}{\mathbf{R}}$ é uma matriz de dimensão g x $(k+1)$ de constantes
- $\underset{\sim}{\mathbf{r}}$ é um vetor de constantes especificadas de dimensão g

é conhecida como hipótese linear geral.

Observação

A condução do teste de hipóteses associado a tal formulação é muito flexível e serve para testar quaisquer tipos de hipóteses lineares de interesse (restrições nos parâmetros).

Exemplo

O gerente de uma empresa terceirizada, responsável pelo recrutamento e seleção de novos funcionários para a empresa TEMCO, acredita que os salários dos funcionários da TEMCO sofrem um acréscimo médio de 700,00 dólares, por ano a mais na empresa, e que a experiência prévia na função não tem impacto no salário, uma vez que a TEMCO mantém uma política de contratar recém-formados e trabalhadores sem experiência, pois prefere fornecer um treinamento customizado aos recém-contratados, *ceteris paribus*.

Exemplo (cont.)

Para tanto, a análise inferencial deve ser feita a partir da estimação dos parâmetros de um modelo de regressão linear múltipla que apresenta *educ*, *anosemp* e *expprev* como regressores e *salario* como regressando. Adotando um nível de significância de 5%, a desconfiança do gerente procede ou não.

Exemplo (cont.)

Modelo proposto:

$$salario_i = \beta_0 + \beta_1 educ_i + \beta_2 anosemp_i + \beta_3 \exp prev_i + \varepsilon_i$$

em que

***salario* – anual, em dólares;**

***anosemp* – tempo (em anos) na empresa;**

***expprev* – experiência anterior (em anos);**

***educ* – anos de estudo após o segundo grau.**

Exemplo (cont.)

Modelo proposto:

$$salario_i = \beta_0 + \beta_1 educ_i + \beta_2 anosemp_i + \beta_3 \exp prev_i + \varepsilon_i$$

Hipóteses de Interesse:

$$H_0: \beta_2 = 700 \quad e \quad \beta_3 = 0$$
$$H_A: \beta_2 \neq 700 \quad e/ou \quad \beta_3 \neq 0$$

Exemplo (cont.)

Que é equivalente a escrever:

$$\mathrm{H}_0 : \begin{cases} \beta_2 = 700 \\ \beta_3 = 0 \end{cases} \qquad \mathrm{H}_\mathrm{A} : \begin{cases} \beta_2 \neq 700 \\ e/ou \\ \beta_3 \neq 0 \end{cases}$$

Ou, ainda

$$\mathrm{H}_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 700 \\ 0 \end{pmatrix} \qquad \mathrm{H}_\mathrm{A} : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 700 \\ 0 \end{pmatrix}$$

(hipótese linear geral)

Exemplo (cont.)

Vale observar que a última formulação é obtida a partir da representação geral, dada por

$$H_0 : \underset{\sim}{\mathbf{R}}\,\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}}$$

$$H_A : \underset{\sim}{\mathbf{R}}\,\underset{\sim}{\boldsymbol{\beta}} \neq \underset{\sim}{\mathbf{r}}$$

com

$$\underset{\sim}{\mathbf{R}} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \underset{\sim}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \text{e} \quad \underset{\sim}{\mathbf{r}} = \begin{pmatrix} 700 \\ 0 \end{pmatrix}$$

TESTE F-parcial

Teste F-parcial

Prova-se que a estatística

(1) $$F = \frac{\left(\tilde{\hat{\boldsymbol{\varepsilon}}}_{\mathbf{R}}' \tilde{\hat{\boldsymbol{\varepsilon}}}_{\mathbf{R}} - \tilde{\hat{\boldsymbol{\varepsilon}}}_{\mathbf{IR}}' \tilde{\hat{\boldsymbol{\varepsilon}}}_{\mathbf{IR}}\right) / g}{\tilde{\hat{\boldsymbol{\varepsilon}}}_{\mathbf{IR}}' \tilde{\hat{\boldsymbol{\varepsilon}}}_{\mathbf{IR}} / [n-(k+1)]}$$

ou

(2) $$F = \frac{\left(R_{IR}^2 - R_R^2\right)/g}{\left(1-R_{IR}^2\right)/[n-(k+1)]}$$

sob a hipótese nula e, ainda, admitindo a validade das suposições MLR.1 a MLR.6, segue uma distribuição

$$F_{[g;n-(k+1)]}$$

Teste F-parcial

em que

$\underset{\sim}{\hat{\boldsymbol{\varepsilon}}}_{\mathbf{R}}$ – vetor de resíduos associado à estimação dos parâmetros do modelo restrito (modelo definido sob H_0);

$\underset{\sim}{\hat{\boldsymbol{\varepsilon}}}_{\mathbf{IR}}$ – vetor de resíduos associado ao modelo irrestrito;

g – número de restrições a serem testadas, sob H_0;

R_R^2 – coeficiente de determinação associado à estimação dos parâmetros do modelo restrito (modelo definido sob H_0);

R_{IR}^2 – coeficiente de determinação associado à estimação dos parâmetros do modelo irrestrito.

Teste F-parcial – Exercício

Mostre que (1) e (2) são equivalentes.

Observação

O teste F-parcial pode ser utilizado como:

i. forma de verificar a contribuição de uma ou mais variáveis explicativas como se estas fossem as últimas variáveis que entraram no modelo;

ii. critério de seleção da melhor equação de regressão.

Voltando ao Exemplo

Modelo Irrestrito

$$salario_i = \beta_0 + \beta_1 educ_i + \beta_2 anosemp_i + \beta_3 \exp prev_i + \varepsilon_i$$

Hipóteses de Interesse

$$H_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 700 \\ 0 \end{pmatrix} \qquad H_A : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 700 \\ 0 \end{pmatrix}$$

Modelo Restrito

$$salario_i = \beta_0 + \beta_1 educ_i + 700 * anosemp_i + 0 * \exp prev + \varepsilon_i$$

Resolução

(Modelo Irrestrito)

Dependent Variable: SALARIO
Method: Least Squares
Date: 09/06/10 Time: 17:06
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*ANOSEMP+C(4)*EXPPREV

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	23480.46	2027.696	11.57987	0.0000
C(2)	1925.882	384.4395	5.009586	0.0000
C(3)	671.3254	143.2125	4.687618	0.0000
C(4)	-73.82734	232.7840	-0.317150	0.7527

R-squared	0.740548	Mean dependent var	39827.39
Adjusted R-squared	0.722016	S.D. dependent var	10999.24
S.E. of regression	5799.262	Akaike info criterion	20.25179
Sum squared resid	1.41E+09	Schwarz criterion	20.41080
Log likelihood	-461.7912	Hannan-Quinn criter.	20.31136
F-statistic	39.95994	Durbin-Watson stat	1.250596
Prob(F-statistic)	0.000000		

Resolução

(Modelo Restrito)

Dependent Variable: SALARIO
Method: Least Squares
Date: 09/06/10 Time: 17:16
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+700*ANOSEMP+0*EXPPREV

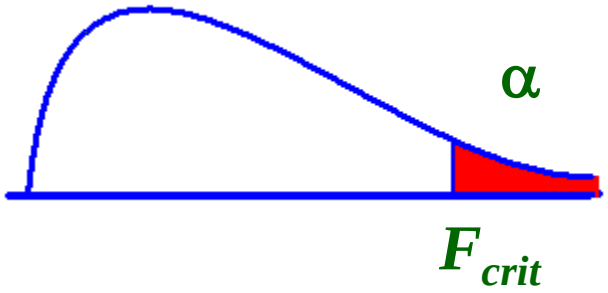
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	23119.67	1725.643	13.39772	0.0000
C(2)	1871.481	297.9524	6.281143	0.0000

R-squared	0.739696	Mean dependent var	39827.39
Adjusted R-squared	0.733780	S.D. dependent var	10999.24
S.E. of regression	5675.225	Akaike info criterion	20.16811
Sum squared resid	1.42E+09	Schwarz criterion	20.24762
Log likelihood	-461.8666	Hannan-Quinn criter.	20.19790
F-statistic	125.0332	Durbin-Watson stat	1.237254
Prob(F-statistic)	0.000000		

Resolução

$$H_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 700 \\ 0 \end{pmatrix} \qquad H_A : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 700 \\ 0 \end{pmatrix}$$

$$F = \frac{(0{,}740548 - 0{,}739696)/2}{(1 - 0{,}740548)/42} = 0{,}06896$$



Rejeito H_0 se $F_{obs} > F_{crit}$

$$F_{crit} = F^{(0,05)}_{[4-2;46-4]} = F^{(0,05)}_{[2;42]} \overset{No\ Eviews}{=} @qfdist(0.95,2,42) = 3{,}21994$$

No Eviews

Para realizar um teste de restrição nos parâmetros utilizando o *software Eviews*, basta estimar o modelo completo (sem restrições) e, posteriormente,

i. **clicar no ícone *view* (que fica no lado esquerdo da janela que mostra os resultados da estimação);**
ii. **em seguida clicar no menu de opções de *coefficient diagnostics*;**
iii. **selecionar, então, a opção *coefficient restrictions,* e digitar a hipótese nula de interesse.**

Resolução (direto no *Eviews*)

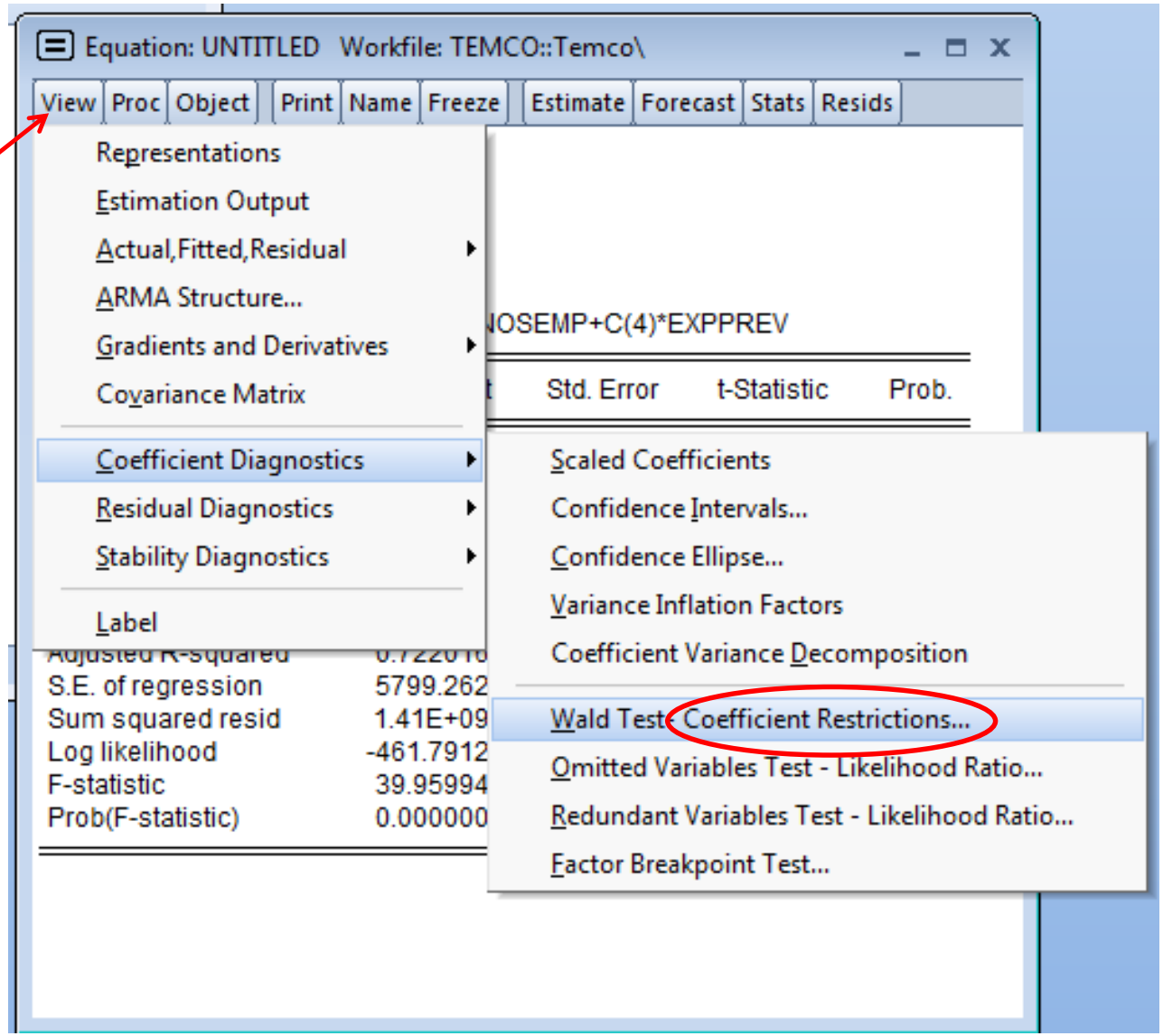
Dependent Variable: SALARIO
Method: Least Squares
Date: 09/06/10 Time: 17:06
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*ANOSEMP+C(4)*EXPPREV

(Modelo Irrestrito)

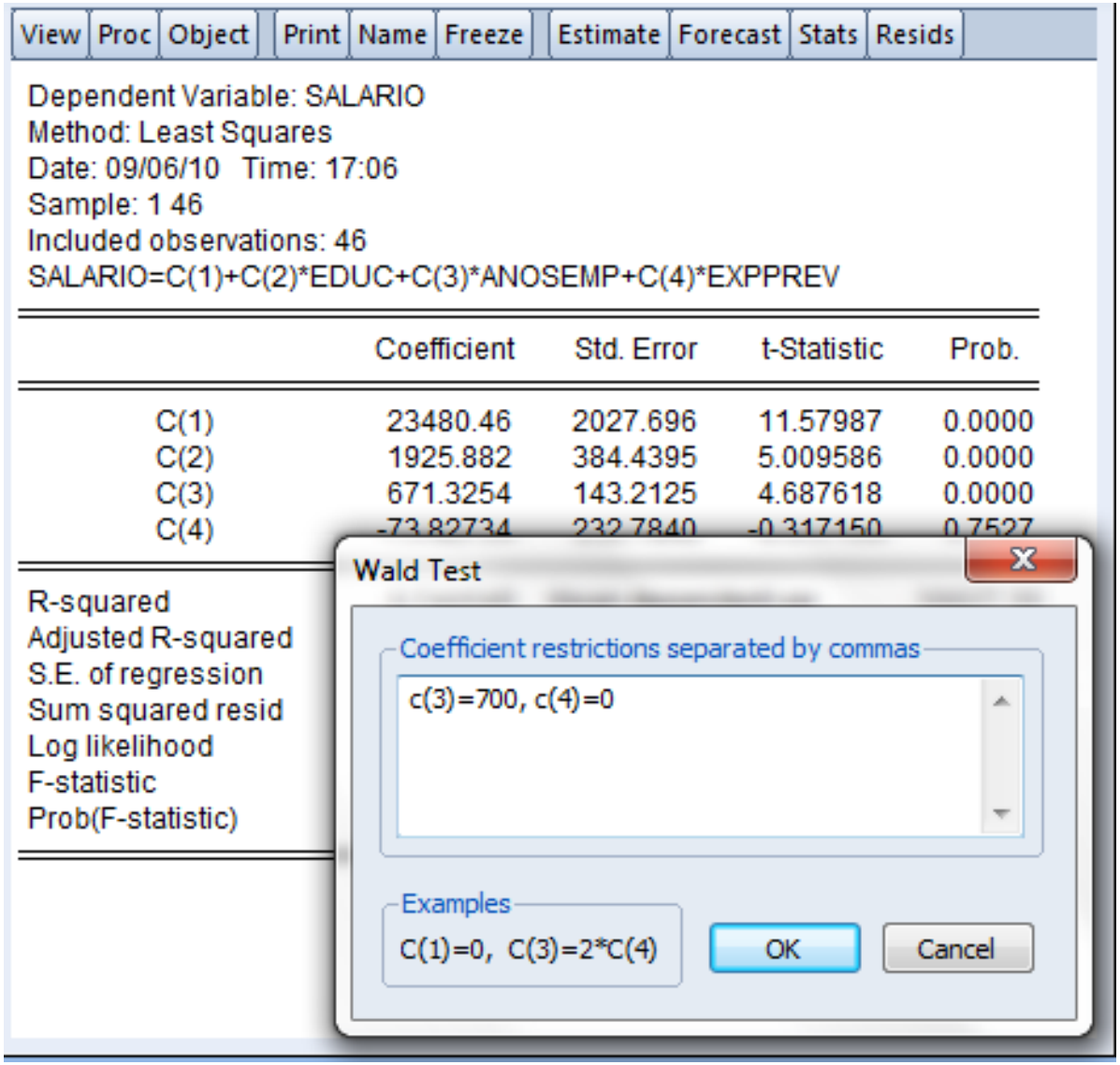
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	23480.46	2027.696	11.57987	0.0000
C(2)	1925.882	384.4395	5.009586	0.0000
C(3)	671.3254	143.2125	4.687618	0.0000
C(4)	-73.82734	232.7840	-0.317150	0.7527

R-squared	0.740548	Mean dependent var	39827.39
Adjusted R-squared	0.722016	S.D. dependent var	10999.24
S.E. of regression	5799.262	Akaike info criterion	20.25179
Sum squared resid	1.41E+09	Schwarz criterion	20.41080
Log likelihood	-461.7912	Hannan-Quinn criter.	20.31136
F-statistic	39.95994	Durbin-Watson stat	1.250596
Prob(F-statistic)	0.000000		

Resolução (direto no *Eviews*)



Resolução (direto no *Eviews*)



Resolução (direto no *Eviews*)

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	0.068973	(2, 42)	0.9335
Chi-square	0.137945	2	0.9334

Null Hypothesis: C(3)=700, C(4)=0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
-700 + C(3)	-28.67455	143.2125
C(4)	-73.82734	232.7840

Restrictions are linear in coefficients.

Exercício Resolvido

O sindicato, ao qual pertencem os funcionários da empresa TEMCO, afirma ao diretor que deve haver um acréscimo médio anual de U$ 2.700,00 quando aumenta-se conjuntamente 1 ano no tempo de empresa e 1 ano de estudo após o 2º grau, mantendo-se o tempo de experiência prévia fixo. Conclua se a empresa segue a norma com 95% de confiança.

Exercício Resolvido

Modelo proposto:

$$salario_i = \beta_0 + \beta_1 educ_i + \beta_2 anosemp_i + \beta_3 \exp prev_i + \varepsilon_i$$

Hipóteses de Interesse:

$$H_0: \beta_1 + \beta_2 = 2700$$
$$H_A: \beta_1 + \beta_2 \neq 2700$$

Exercício Resolvido

Ainda, as hipóteses escritas na forma Linear Geral (HLG) ficam dadas por

$$\mathrm{H}_0 : \underset{\sim}{\mathbf{R}}\,\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}} \Leftrightarrow \beta_1 + \beta_2 = 2700$$

$$\mathrm{H}_\mathrm{A} : \underset{\sim}{\mathbf{R}}\,\underset{\sim}{\boldsymbol{\beta}} = \underset{\sim}{\mathbf{r}} \Leftrightarrow \beta_1 + \beta_2 \neq 2700$$

em que

$$\underset{\sim}{\mathbf{R}} = \begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}, \quad \underset{\sim}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \text{e} \quad \underset{\sim}{\mathbf{r}} = \begin{pmatrix} 2700 \end{pmatrix}$$

Exercício Resolvido

Dependent Variable: SALARIO
Method: Least Squares
Date: 10/19/12 Time: 11:28
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*ANOSEMP+C(4)*EXPPREV

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	23480.46	2027.696	11.57987	0.0000
C(2)	1925.882	384.4395	5.009586	0.0000
C(3)	671.3254	143.2125	4.687618	0.0000
C(4)	-73.82734	232.7840	-0.317150	

R-squared	0.740548	Mean dependent var	
Adjusted R-squared	0.722016	S.D. dependent var	
S.E. of regression	5799.262	Akaike info criterion	
Sum squared resid	1.41E+09	Schwarz criterion	
Log likelihood	-461.7912	Hannan-Quinn criter.	
F-statistic	39.95994	Durbin-Watson stat	
Prob(F-statistic)	0.000000		

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
t-statistic	-0.322742	42	0.7485
F-statistic	0.104163	(1, 42)	0.7485
Chi-square	0.104163	1	0.7469

Null Hypothesis: C(2)+C(3)=2700
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
-2700 + C(2) + C(3)	-102.7922	318.4960

Restrictions are linear in coefficients.

Exercícios

Exercício 1

Entregar na próxima aula

A senhorita Rose Jolie, gerente do departamento de RH da empresa TEMCO, gostaria de estimar os parâmetros de um modelo de regressão linear múltipla que levasse em consideração os regressores educ, anosemp e dept na explicação do ln(salário). Ainda, fazendo uma revisão da literatura, a senhorita Rose Jolie notou que muitos autores dizem que o tempo de escolaridade, dependendo do departamento onde o funcionário trabalha, costuma apresentar um efeito diferenciado na variável resposta.

Exercício 1 (Cont.)

Entregar na próxima aula

a. **Estime o modelo de interesse da senhorita Rose Jolie e escreva os resultados na forma usual.**
b. **Interprete as estimativas dos parâmetros em termos do problema em questão.**
c. **Escreva a equação na forma usual para cada um dos departamentos da empresa.**

Exercício 1 (Cont.)

Entregar na próxima aula

d. **Pode-se dizer que o modelo é significante com 95% de confiança? Justifique sua resposta.**

e. **Verifique se há um efeito diferenciado de *educ* no ln(salário) dos funcionários dos diversos departamentos da empresa, com 95% de confiança.**

f. **Verifique se o departamento do funcionário influencia o ln(salário) com 95% de confiança.**

Exercício 2

Utilizando os dados do arquivo ***GPA2.wf1***, estimou-se a seguinte equação

$$\widehat{sat} = 1028{,}10 + 19{,}30\ hsize - 2{,}19\ hsize^2 - 45{,}09\ female - 169{,}81\ black + 62{,}31\ female \cdot black$$
$$(6{,}29) \quad (3{,}83) \qquad (0{,}53) \qquad (4{,}29) \qquad (12{,}71) \qquad (18{,}15)$$

$$n = 4137,\ R^2 = 0{,}0858$$

A variável *sat* é um escore combinado, *hsize* é o tamanho das salas de aulas dos estudantes da graduação, em centenas de estudantes, *female* é uma variável *dummy* de gênero (1 - feminino, 0 = caso contrário) e *black* é uma variável *dummy* de raça (1 – negro, 0 – caso contrário).

a) Há fortes evidências para que $hsize^2$ seja incluída no modelo? Justifique sua resposta.

b) Segundo o modelo estimado, qual o tamanho ótimo para as salas de aulas?

c) Fixando-se as demais variáveis, qual é a diferença estimada na variável resposta quando comparamos mulheres não negras com homens não negros? Esta diferença é estatisticamente significante?

d) Qual é a diferença estimada na variável resposta quando comparamos homens negros com homens não negros? Teste H_0: ausência de diferença entre os escores, contra H_A: existem diferenças.

e) Qual é a diferença estimada na variável resposta quando comparamos mulheres negras com mulheres não negras?

Exercício 3

A senhorita Jolie, gerente do departamento de RH da empresa TEMCO, agora desconfia que, dependendo do departamento onde o funcionário trabalha, cada ano a mais de escolaridade tenha um efeito diferenciado no valor esperado do salário. Assim sendo, proponha um modelo de regressão linear que seja adequado para testar tal desconfiança.

Exercício 3 (cont.)

Modelo proposto:

$$salario = \beta_0 + \beta_1 educ + \beta_2 D_C + \beta_3 D_E + \beta_4 D_P +$$
$$+ \beta_5 educ D_C + \beta_6 educ D_E + \beta_7 educ D_P + \varepsilon$$

em que

D_C – variável *dummy* que assume o valor 1 caso o funcionário seja do departamento de compras;

D_E – variável *dummy* que assume o valor 1 caso o funcionário seja do departamento de engenharia;

D_P – variável *dummy* que assume o valor 1 caso o funcionário seja do departamento de propaganda.

Exercício 3 (cont.)

Modelo proposto:

$$salario = \beta_0 + \beta_1 educ + \beta_2 D_C + \beta_3 D_E + \beta_4 D_P +$$
$$+ \beta_5 educ D_C + \beta_6 educ D_E + \beta_7 educ D_P + \varepsilon$$

Hipóteses de Interesse:

$$H_0: \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_A: ao\ menos\ um\ parâmetro\ diferente\ de\ zero$$

Exercício 4

Utilizando a base de dados *TEMCOPROD.wf1*, responda:

(a) Existe relação entre o ln(*salário*) e a *produtividade* dos funcionários da empresa TEMCO?

(b) Proponha e estime os parâmetros de um modelo de regressão linear simples para prever o ln(*salário*) com base na *produtividade* dos funcionários analisados. Escreva os resultados na forma usual e interprete as estimativas dos parâmetros e o coeficiente de determinação.

Exercício 4 (cont.)

Utilizando a base de dados *TEMCOPROD.wf1*, responda: (cont.)

(c) Com base nas informações coletadas de 46 funcionários da empresa TEMCO, proponha e estime os parâmetros de um modelo de regressão linear múltipla para prever o ln(*salário*) com base nas variáveis explicativas *educ* e *anosemp*. Escreva os resultados na forma usual e interprete as estimativas dos parâmetros e o coeficiente de determinação.

Exercício 4 (cont.)

Utilizando a base de dados *TEMCOPROD.wf1*, responda: (cont.)

(d) De resultados anteriores, foi possível observar que as variáveis *educ* e *anosemp* são conjuntamente relevantes para explicar o ln(*salário*). Pergunta-se, então: a variável *produtividade* traz alguma informação relevante para explicar o ln(*salário*), num modelo que já apresenta *educ* e *anosemp* como variáveis explicativas?

Exercício 5

Utilizando a base de dados *TEMCOPROD.wf1*, responda:

A introdução de *educ* e *anosemp*, num modelo que já possui *produtividade*, traz alguma informação relevante para explicar o ln(*salário*)?