Insper Instituto de Ensino e Pesquisa

Análise de Regressão Linear Múltipla IV

Aula 07

Gujarati e Porter, 2011 – Capítulos 7 e 8

Heij et al., 2004 – Capítulo 3

Exemplo

Tomando por base o modelo

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

a senhorita Jolie, gerente do departamento de RH da empresa TEMCO, desconfia que ao menos um dos regressores é relevante para explicar a variável resposta. Utilizando um nível de significância de 1%, conduza um teste de hipóteses adequado.

Exemplo

Modelo

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

Hipóteses de Interesse

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_A: \text{pelo menos um parâmetro difere de zero}$$

SST = SSR + SSE

Se H_0 for verdadeira, espera-se que SSE seja pequena e SSR grande.

TESTE F

(Análise de Variâncias – ANOVA)

Teste F

É possível demonstrar que, sob certas condições, as v.a. SSR, SSE e SST apresentam as seguintes características:

1. $\frac{SSR}{\sigma^2} \sim \chi^2_{[n-(k+1)]}$;

2. $\frac{SSE}{\sigma^2} \sim \chi^2_{(k)}$, $se \quad \beta_1 = \ldots = \beta_k = 0$;

3. SSR e SSE são independentes.

Teste F

Consequências:

(a) $E\left(\frac{SSR}{\sigma^2}\right) = n-(k+1) \Rightarrow E\left(\frac{SSR}{n-(k+1)}\right) = E(MSR) = \sigma^2$

Logo, MSR é um estimador não-viesado de σ^2

(b) $\mathrm{E}\left(\frac{\mathrm{SSE}}{\sigma^2}\right) = \mathrm{k},\ se\ \beta_1 = \ldots = \beta_k = 0 \Rightarrow \mathrm{E}\left(\frac{\mathrm{SSE}}{\mathrm{k}}\right) = \mathrm{E}(\mathrm{MSE}) = \sigma^2$

Se $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, então MSE = SSE/k é um estimador não-viesado de σ^2.

Teste F

Consequências: (cont.)

(c) Se $\beta_1 = \ldots = \beta_k = 0,$

$$E(SST) = E(SSR) + E(SSE) =$$
$$= [n - (k+1)]\sigma^2 + (k)\,\sigma^2 = (n\text{-}1)\,\sigma^2$$

Logo, SST/(n-1) é estimador não-viesado de σ^2

Teste F

Consequências: (cont.)

(d) Se $\beta_1 = \ldots = \beta_k = 0,$

$$F = \frac{\frac{SSE/\sigma^2}{k}}{\frac{SSR/\sigma^2}{[n-(k+1)]}} = \frac{\frac{SSE}{k\text{-}1}}{\frac{SSR}{[n-(k+1)]}} = \frac{MSE}{MSR} \sim F_{[k,n-(k+1)]}$$

Teste F

Consequências: (cont.)

Fonte de variação	SS	gl	MS	F
Regressão	SSE	k	MSE	MSE/MSR
Erro	SSR	n-(k+1)	MSR	
Total	**SST**	n-1		

$$MSE = \frac{SSE}{k}$$

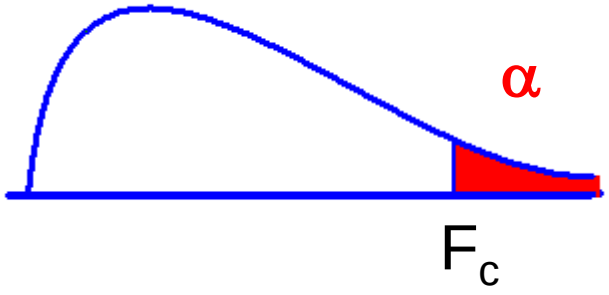
$$MSR = \frac{SSR}{n - (k+1)}$$

Teste F

Consequências: (cont.)

$$F = \frac{MSE}{MSR} = \frac{R^2 / (k)}{(1-R^2) / [n-(k+1)]} \overset{\text{sob } H_0}{\sim} F_{[k,\ n-(k+1)]}$$

Região crítica:



Exemplo

Tomando por base o modelo

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

a senhorita Jolie, gerente do departamento de RH da empresa TEMCO, desconfia que ao menos um dos regressores é relevante para explicar a variável resposta. Utilizando um nível de significância de 1%, conduza um teste de hipóteses adequado.

Resolução do Exemplo

Modelo

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

Hipóteses de Interesse

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_A: \text{pelo menos um parâmetro difere de zero}$$

Resolução do Exemplo

Dependent Variable: LOG(SALARIO)
Method: Least Squares
Date: 08/26/13 Time: 14:06
Sample: 1 46
Included observations: 46
LOG(SALARIO)=C(1)+C(2)*EDUC+C(3)*ANOSEMP+C(4)*EXPPREV

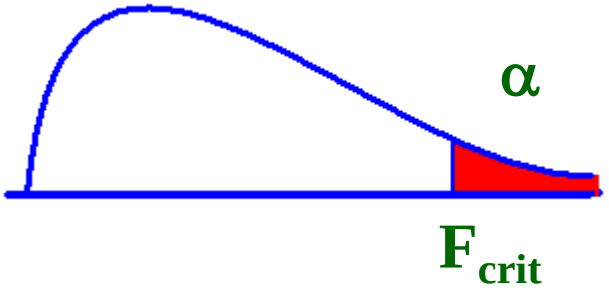
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	10.15234	0.046720	217.3026	0.0000
C(2)	0.045025	0.008858	5.083099	0.0000
C(3)	0.016009	0.003300	4.851557	0.0000
C(4)	0.002736	0.005364	0.510041	0.6127

R-squared	0.751687	Mean dependent var	10.55832
Adjusted R-squared	0.733950	S.D. dependent var	0.259053
S.E. of regression	0.133620	Akaike info criterion	-1.104695
Sum squared resid	0.749879	Schwarz criterion	-0.945683
Log likelihood	29.40798	Hannan-Quinn criter.	-1.045128
F-statistic	42.38040	Durbin-Watson stat	1.167563
Prob(F-statistic)	0.000000		

Resolução do Exemplo

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

H_A: *pelo menos um parâmetro difere de zero*



$$F_{crit} = F_{[3;42]}^{(0,01)} \overset{\text{No Eviews}}{=} @qfdist\,(0.99,3,42) = 4,285$$

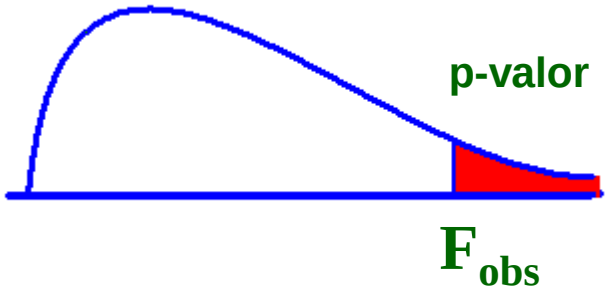
$$F_{obs} = 42,38$$

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

Resolução do Exemplo

$H_0: \beta_2 = \beta_3 = \beta_4 = 0$

H_A: *pelo menos um parâmetro difere de zero*



$$p-valor = P\left(F_{[3;42]} > F_{obs}\right) \overset{No\ Eviews}{=} 1-@cfdist\ (42.38,3,42) = 9{,}07 \cdot 10^{-13}$$

Rejeito H_0 se p - valor $< \alpha$

Voltando ao Exemplo

A senhorita Jolie sabe que, a 1% de significância, ao menos um dos regressores é relevante para explicar a variável resposta. Todavia, a senhorita Jolie desconfia que *expprev* seja irrelevante, dado que os funcionários da TEMCO passam por um processo de treinamento assim que são admitidos na empresa. Dessa forma, adotando um nível de significância de 1%, existem evidências favoráveis à desconfiança da gerente de RH?

Voltando ao Exemplo

Modelo

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

Hipóteses de Interesse

$$H_0: \beta_3 = 0$$

$$H_A: \beta_3 \neq 0$$

Teste t

Teste t

Já foi visto que

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{\sigma^2}{SQT_{x_j}(1-R^2_{x_j})}}} \sim N(0;1)$$

e como σ^2 é um parâmetro desconhecido, então deverá ser estimado. Dessa maneira, será necessário estudar a distribuição de probabilidades da nova v.a. resultante.

Teste t

Nos slides anteriores foi dito que SSR, SSE e SST são v.a. e, ainda, não é difícil provar que, sob certas condições:

$$\frac{SSR}{\sigma^2} \sim \chi^2_{n-(k+1)};$$

Assim,

$$E\left(\frac{SSR}{\sigma^2}\right) = n-(k+1) \Rightarrow E\left(\frac{SSR}{n-(k+1)}\right) = E(MSR) = \sigma^2$$

MSR é um estimador não-viesado de σ^2

Teste t

Assim, substituindo σ^2, pelo seu estimador, MSR, na expressão do *slide* 19, temos que

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{MSR}{SQT_{x_j}(1-R^2_{x_j})}}}$$

em que

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\frac{\hat{\sigma}^2}{(n-1)s^2_{X_j}(1-R^2_j)}} : erro\ padrão\ de\ \hat{\beta}_j$$

$$\sqrt{\hat{\sigma}^2} = \sqrt{MSR} = \hat{\sigma} : erro\ \ padrão\ da\ \ regressão$$

Teste t

Logo, para testarmos as hipóteses

H_0: $\beta_j = b$ (em particular b = 0)

H_A: $\beta_j \neq b$ (H_A: $\beta_j < b$ ou H_A: $\beta_j > b$),

utilizaremos o fato que, sob H_0,

$$\frac{\hat{\beta}_j - b}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-(k+1)}$$

e construiremos a região crítica de acordo com a hipótese alternativa adotada.

Voltando ao Exemplo

Modelo

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

Hipóteses de Interesse

$$H_0: \beta_3 = 0$$

$$H_A: \beta_3 \neq 0$$

Resolução do Exemplo

Dependent Variable: LOG(SALARIO)
Method: Least Squares
Date: 08/26/13 Time: 14:06
Sample: 1 46
Included observations: 46
LOG(SALARIO)=C(1)+C(2)*EDUC+C(3)*ANOSEMP+C(4)*EXPPREV

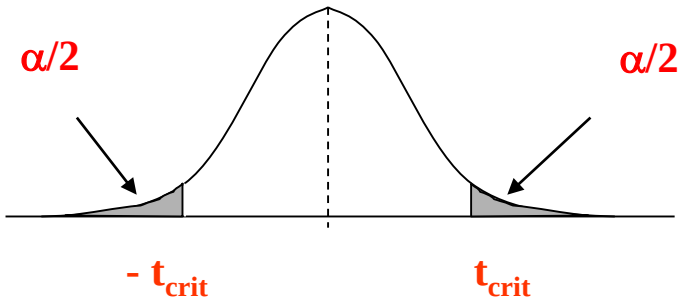
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	10.15234	0.046720	217.3026	0.0000
C(2)	0.045025	0.008858	5.083099	0.0000
C(3)	0.016009	0.003300	4.851557	0.0000
C(4)	0.002736	0.005364	0.510041	0.6127

R-squared	0.751687	Mean dependent var	10.55832
Adjusted R-squared	0.733950	S.D. dependent var	0.259053
S.E. of regression	0.133620	Akaike info criterion	-1.104695
Sum squared resid	0.749879	Schwarz criterion	-0.945683
Log likelihood	29.40798	Hannan-Quinn criter.	-1.045128
F-statistic	42.38040	Durbin-Watson stat	1.167563
Prob(F-statistic)	0.000000		

Resolução do Exemplo

$$H_0: \beta_3 = 0$$

$$H_A: \beta_3 \neq 0$$



$$t_{crit} = t^{(0,005)}_{[46-4]} = t^{(0,005)}_{[42]} \overset{No\ Eviews}{=} @\,qtdist(0.995,42) = 2,698$$

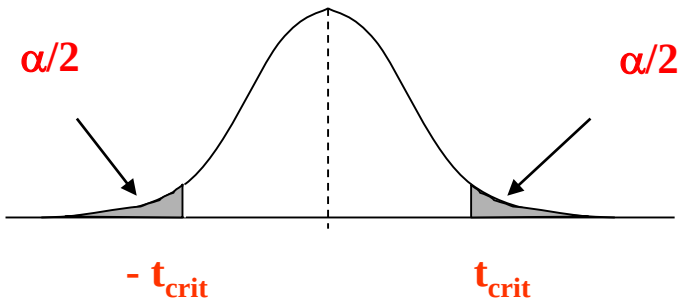
$$t_{obs} = \frac{0,002736 - 0}{0,005364} = 0,510041$$

Rejeito H_0 se $|t_{obs}| > t_{crit}$

Resolução do Exemplo

$$H_0: \beta_3 = 0$$

$$H_A: \beta_3 \neq 0$$



$$t_{obs} = \frac{0{,}002736 - 0}{0{,}005364} = 0{,}510041$$

$$p - valor = 2 \cdot P\left(t_{[42]} \geq 0{,}510041\right) \overset{No\ Eviews}{=} 2 \cdot \left[1 - @ctdist(0{,}510041)\right] = 0{,}6127$$

Rejeito H_0 se p - valor < α

Voltando ao Exemplo

Tomando por base o modelo

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

existem evidências sobre a relevância da variável *educ*, com 99% de confiança? Toda a sua análise deve ser baseada na construção de um intervalo de confiança.

Intervalo de Confiança

Intervalo de Confiança para β_j

Prova-se que

$$IC(\beta_j;\gamma) = \left(\hat{\beta}_j \pm t^{\alpha/2}_{n-(k+1)} \cdot \hat{\sigma}_{\hat{\beta}_j}\right)$$

em que

$$\hat{\sigma}_{\hat{\beta}_j} - erro\ padrão\ \ associado\ a\ \hat{\beta}_j$$

é um intervalo de confiança para o parâmetro β_j, com coeficiente de confiança de 1-α.

Voltando ao Exemplo

Modelo

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

Hipóteses de Interesse

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

Resolução do Exemplo

Dependent Variable: LOG(SALARIO)
Method: Least Squares
Date: 08/26/13 Time: 14:06
Sample: 1 46
Included observations: 46
LOG(SALARIO)=C(1)+C(2)*EDUC+C(3)*ANOSEMP+C(4)*EXPPREV

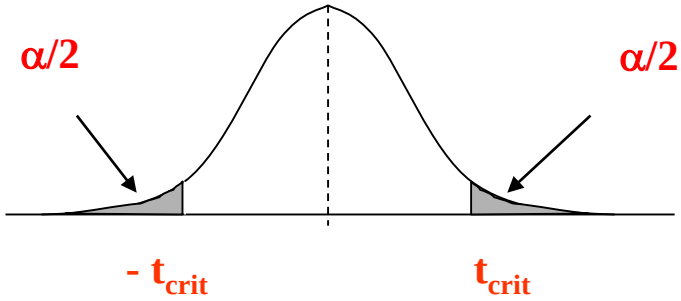
	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	10.15234	0.046720	217.3026	0.0000
C(2)	0.045025	0.008858	5.083099	0.0000
C(3)	0.016009	0.003300	4.851557	0.0000
C(4)	0.002736	0.005364	0.510041	0.6127

R-squared	0.751687	Mean dependent var	10.55832
Adjusted R-squared	0.733950	S.D. dependent var	0.259053
S.E. of regression	0.133620	Akaike info criterion	-1.104695
Sum squared resid	0.749879	Schwarz criterion	-0.945683
Log likelihood	29.40798	Hannan-Quinn criter.	-1.045128
F-statistic	42.38040	Durbin-Watson stat	1.167563
Prob(F-statistic)	0.000000		

Resolução do Exemplo

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$



$$t_{crit} = t^{(0,005)}_{[46-4]} = t^{(0,005)}_{[42]} \overset{No\ Eviews}{=} @\,qtdist(0.995,42) = 2,698$$

$$IC(\beta_j;\gamma) = \left(0,045025 \pm \overbrace{2,698 \cdot 0,008858}^{0,023899} \right) = (0,021126;\ 0,068924)$$

Como o IC não engloba o zero, então, com 99% de confiança, existem evidências contrárias à hipótese nula.

Exercício Completo

A administração de um hospital particular deseja estudar a relação entre a satisfação dos pacientes em relação ao atendimento do hospital (escore de 0 a 100) em função do índice de severidade da doença do paciente (escore de 0 a 100), do custo hospitalar pago pelo paciente (em milhares de reais) e do nível de ansiedade do paciente (escore de 0 a 5). Para o estudo foi coletada uma amostra aleatória de 23 pacientes atendidos no último mês e os resultados estão no arquivo *sat_pacientes.xls*.

Exercício Completo

a) Faça a análise descritiva dos dados obtidos em função do problema a ser analisado.
b) Proponha e estime os parâmetros de um modelo de regressão adequado. Escreva os resultados na forma usual. Interprete as estimativas dos parâmetros do modelo em termos do problema em questão.
c) Calcule uma medida de qualidade do ajuste do modelo e interprete-a em termos do problema em questão.
d) Qual(is) das variáveis explicativas utilizadas no modelo influenciam a satisfação dos pacientes? Justifique sua resposta com base numa técnica inferencial adequada e use um nível de significância de 10%.

Exercício Completo

e) Estime um novo modelo com base nos resultados obtidos no item (d).

f) O diretor do hospital afirmou que o aumento de uma unidade no índice de severidade da doença de um paciente provoca um decréscimo de pelo menos 1,5 no escore de satisfação esperado. Verifique se esta afirmação procede com 90% de confiança.

g) Com base num intervalo com 95% de confiança, verifique se há um decréscimo médio de 2 unidades no escore de satisfação com o aumento do custo hospitalar em mil reais.

Exercício Completo

h) **(DESAFIO)** O diretor ainda afirmou que o escore de satisfação decresce em média 2 unidades quando se aumenta o índice de severidade em 1 unidade e o custo hospitalar em mil reais, conjuntamente. Teste a afirmação do diretor com 10% de significância.

Dica: Leitura complementar!!!!

LEITURA COMPLEMENTAR

Teste de Hipóteses sobre uma única Combinação Linear de Parâmetros (teste t)

Exemplo

Seja a equação de regressão múltipla,

$$salário = \beta_0 + \beta_1 educ + \beta_2 produtividade + \varepsilon$$

Verifique, a partir da formulação e construção de um teste de hipóteses, se a variável *educ* apresenta um impacto superior ao da variável *produtividade* na variável resposta. Nesse exemplo, utilize o banco de dados *TEMCOPROD.wf1*.

Exemplo

Modelo

$$salário = \beta_0 + \beta_1 educ + \beta_2 produtividade + \varepsilon$$

Hipóteses

$$H_0 : \beta_1 = \beta_2 \Leftrightarrow \beta_1 - \beta_2 = 0$$

$$H_A : \beta_1 > \beta_2$$

Teste de hipóteses sobre uma única combinação linear de parâmetros

Sejam as hipóteses

$$H_0 : \beta_i = \beta_j \Leftrightarrow \beta_i - \beta_j = 0$$

$$H_A : \beta_i \neq \beta_j \ (\beta_i < \beta_j \quad \text{ou} \quad \beta_i > \beta_j)$$

Estatística do teste

$$t = \frac{(\hat{\beta}_i - \hat{\beta}_j) - 0}{se(\hat{\beta}_i - \hat{\beta}_j)}$$

$$se(\hat{\beta}_i - \hat{\beta}_j) = \sqrt{\left[se(\hat{\beta}_i)\right]^2 + \left[se(\hat{\beta}_j)\right]^2 - 2C\hat{o}v(\hat{\beta}_i, \hat{\beta}_j)}$$

Teste de hipóteses sobre uma única combinação linear de parâmetros

ALTERNATIVAS DE SOLUÇÃO

A) Calcular todos os componentes do erro padrão (o *software Eviews* gera a matriz de variâncias e covariâncias para os estimadores dos parâmetros do modelo de regressão. Lembra onde está?).

$$se(\hat{\beta}_i - \hat{\beta}_j) = \sqrt{\left[se(\hat{\beta}_i)\right]^2 + \left[se(\hat{\beta}_j)\right]^2 - 2C\hat{o}v(\hat{\beta}_i, \hat{\beta}_j)}$$

Teste de hipóteses sobre uma única combinação linear de parâmetros

ALTERNATIVAS DE SOLUÇÃO (cont.)

B) Trabalhar com um modelo transformado para obter o resultado diretamente

Seja

$$\theta = \beta_i - \beta_j,$$

por exemplo, β_j pode ser escrito como $\beta_j = \beta_i - \theta$ e, substituindo este resultado na equação de regressão múltipla, podemos testar H_0: $\theta = 0$, contra uma alternativa apropriada.

Teste de hipóteses sobre uma única combinação linear de parâmetros

ALTERNATIVAS DE SOLUÇÃO (cont.)

C) Utilizar o *menu* do *Eviews* para solucionar o problema (teste de restrições nos coeficientes).

D) Trabalhar com uma hipótese linear geral ($R\beta = r$) e usar, por exemplo, o MATLAB.

E) Estimar o modelo restrito e o irrestrito e, através dos coeficientes de determinação de ambos, conduzir o teste de hipóteses de interesse.

Voltando ao Exemplo (solução *B*)

Modelo

$$salário = \beta_0 + \beta_1 educ + \beta_2 produtividade + \varepsilon$$

Hipóteses

$$H_0 : \beta_1 = \beta_2 \Leftrightarrow \beta_1 - \beta_2 = 0$$

$$H_A : \beta_1 > \beta_2$$

Escrevendo $\theta = \beta_1 - \beta_2$, vem que

$$\mathbf{H_0} : \theta = 0$$

$$\mathbf{H_A} : \theta > 0$$

Voltando ao Exemplo (solução *B*)

Mas, $\theta = \beta_1 - \beta_2 \Leftrightarrow \beta_1 = \theta + \beta_2$, e substituindo este resultado no modelo proposto, vem que

$$salário = \beta_0 + \beta_1 educ + \beta_2 produtividade + \varepsilon$$

$$\Updownarrow$$

$$salário = \beta_0 + (\theta + \beta_2) educ + \beta_2 produtividade + \varepsilon$$

$$\Updownarrow$$

$$salário = \beta_0 + \theta\, educ + \beta_2 educ + \beta_2 produtividade + \varepsilon$$

$$\Updownarrow$$

$$salário = \beta_0 + \theta\, educ + \beta_2 (educ + produtividade) + \varepsilon$$

Voltando ao Exemplo (solução *B*)

Dependent Variable: SALARIO
Method: Least Squares
Date: 09/06/10 Time: 17:24
Sample: 1 46
Included observations: 46
SALARIO=C(1)+C(2)*EDUC+C(3)*(EDUC+PRODUTIVIDADE)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	12281.96	3944.548	3.113654	0.0033
C(2)	1255.717	591.9782	2.121222	0.0397
C(3)	470.0954	132.2545	3.554476	0.0009

R-squared	0.693710	Mean dependent var	39827.39
Adjusted R-squared	0.679464	S.D. dependent var	10999.24
S.E. of regression	6227.324	Akaike info criterion	20.37427
Sum squared resid	1.67E+09	Schwarz criterion	20.49353
Log likelihood	-465.6083	Hannan-Quinn criter.	20.41895
F-statistic	48.69484	Durbin-Watson stat	1.198179
Prob(F-statistic)	0.000000		

Voltando ao Exemplo (solução *B*)

Hipóteses

$$H_0 : \theta = 0$$

$$H_A : \theta > 0$$

Sob H_0

$$t_{obs} = \frac{1255{,}717 - 0}{591{,}978} = 2{,}121222$$

$$t_{crítico} = t_{n\text{-}k}^{\alpha} = t_{43}^{0{,}05} = @\,qtdist(0.95{,}43) = 1{,}681$$

$$p - valor = 1 - @\,ctdist(2.121222{,}43) = 0{,}01985$$