Estatística II – Licenciatura em Gestão

EN – 3 Junho 2013

Parte teórica

Nome: ____ **Nº**____________

1. **Perguntas de Verdadeiro/Falso** (**1.5 valores**) - Para cada afirmação, assinale se esta é Verdadeira (V) ou Falsa (F). Uma resposta certa vale 0.3 e uma resposta errada penaliza em idêntico valor.

	V	F
A potência de um teste é igual ao seu valor-*p*		X
Quando num teste de hipóteses com $\alpha = 0.03$ se obtém um valor-p de 0.041 rejeita-se H_0		X
Para que $\hat{\theta}$, estimador de θ, seja consistente é necessário garantir que $E(\hat{\theta}) = \theta$, $\forall \theta \in \Theta$		X
No MRL $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t$, $t = 1, 2, ..., 50$, a verificar as hipóteses habituais, o teste de $H_0 : \beta_3 = \beta_4 = 0$ pode ser feito recorrendo a um teste *t*		X
Quando, num modelo de regressão linear, se introduz uma restrição linear nos parâmetros, a soma dos quadrados dos resíduos não pode diminuir.	**X**	

2. **Perguntas de resposta múltipla** (**2.25 valores**) - Para cada pergunta escolha **a** alternativa correcta. Uma resposta certa vale 0.75 valores e uma resposta errada penaliza em 0.25 valores.

a. Num universo normal de média e variância desconhecidas pretende-se testar $H_0 : \mu = 0$ contra $H_1 : \mu < 0$. Definiu-se a seguinte região de rejeição $W = \{(x_1, x_2, \cdots, x_n) : \bar{x} < k\}$. A função potência do ensaio é dada por:

☐ $P(\bar{X} < k \mid \mu > 0)$

X $P(\bar{X} < k \mid \mu < 0)$

☐ $P(\bar{X} < k \mid \mu = 0)$

☐ $P(\bar{X} > k \mid \mu < 0)$

b. Assinale a alternativa correcta

☐ Todas as variáveis fulcrais são estimadores

☐ Alguns estimadores podem ser variáveis fulcrais

X Todos os estimadores são estatísticas

☐ Algumas variáveis fulcrais podem ser estimadores

c. Considere o MRL estimado pelos mínimos quadrados $\hat{y}_t = b_1 + b_2 x_t$, em que y representa o preço de uma propriedade agrícola em determinada região e x a sua área. Qual a estimativa da elasticidade preço-área ?

☐ b_2

X $b_2 x_t / \hat{y}_t$

☐ $b_2 x_t$

☐ $b_2 \hat{y}_t$

3. Perguntas de desenvolvimento (2.25 valores) **–** alínea a) 1 valor; alínea b) 1.25 valores..

a. Defina o conceito de variável fulcral e explique o seu interesse.

Definição 7.9 do livro

A variável fulcral serve para a construção de intervalos de confiança

b. Considere o MRL $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t$ a verificar as hipóteses habituais. Pretende-se estimar $\theta = \beta_2 + \beta_3$. Recordando-se que o vector b é estimador BLUE para β, mostre que $\hat{\theta} = b_2 + b_3$ é estimador centrado para θ. Reparametrize o modelo de forma a que θ seja um dos coeficientes da regressão.

$E(\hat{\theta}) = E(b_2 + b_3) = E(b_2) + E(b_3) = \beta_2 + \beta_3$ O estimador sendo BLUE, é centrado logo $E(b_j) = \beta_j$

$\theta = \beta_2 + \beta_3 \Leftrightarrow \beta_2 = \theta - \beta_3$

Logo

$$
\begin{aligned}
y_t &= \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t = \beta_1 + (\theta - \beta_3) x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t \\
&= \beta_1 + \theta x_{t2} - \beta_3 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + u_t = \beta_1 + \theta x_{t2} + \beta_3 \left(x_{t3} - x_{t2}\right) + \beta_4 x_{t4} + u_t
\end{aligned}
$$

Estatística II – Licenciatura em Gestão

EN - 3 Junho 2013

Parte Prática

Nome: ____ **Nº**__________

Espaço reservado para classificações

1. (15)	3. (15)	4a. (15)	4d. (15)	**T:**
2a. (15)		4b. (15)	4e. (15)	**P:** _____
2b. (15)		4c. (20)		

Em todos os testes de hipóteses que fizer, formule as hipóteses em teste, indique a estatística de teste e a sua distribuição. Para os intervalos de confiança proceda de forma semelhante para a variável fulcral.

Se necessitar de espaço dispõe de uma folha em branco no fim do enunciado, antes do anexo. Pode arrancar a folha de anexo se lhe der mais jeito

1) De uma população X, cuja distribuição depende dos parâmetros α e β, sabe-se que $E(X) = \alpha$ e $Var(X) = 2\beta^2$, onde $\alpha \in \mathbb{R}$ e $\beta > 0$. Retirada uma amostra casual de 100 observações, obteve-se: $\sum_{i=1}^{100} x_i = 16.63$ e $\sum_{i=1}^{100} x_i^2 = 70.72$. Com base nesta informação, apresente uma estimativa para os parâmetros desconhecidos, α e β.

Estimação pelo método dos momentos.

$$E(X) = \alpha$$
$$E(X^2) = Var(X) + [E(X)]^2 = 2\beta^2 + \alpha^2$$

Logo, igualando os momentos populacionais aos momentos amostrais, obtém-se:

$$\tilde{\alpha} = \bar{X}$$

$$2\beta^2 + \alpha^2 = \frac{\sum_{1=1}^{n} X_i^2}{n} \Rightarrow 2\tilde{\beta}^2 + \bar{X}^2 = \frac{\sum_{1=1}^{n} X_i^2}{n} \Rightarrow \cdots \Rightarrow \tilde{\beta} = \sqrt{\frac{\sum_{1=1}^{n} X_i^2}{2n} - \frac{\bar{X}^2}{2}} = \frac{S}{\sqrt{2}}$$

Pelo que, as estimativas são ($n = 100$):

$$\tilde{\alpha} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{16.63}{100} = 0.1663$$

$$\tilde{\beta} = \sqrt{\frac{\sum_{1=1}^{n} x_i^2}{2n} - \frac{\bar{x}^2}{2}} = \sqrt{\frac{70.72}{200} - \frac{0.1663^2}{2}} \approx 0.5829$$

2) Suponha que lhe é pedido para analisar os preços de duas acções, representados pelas variáveis aleatórias X e Y, e entre elas, escolher a mais "desejável" tendo em conta dois critérios: preço e respectiva volatilidade. Admita que X e Y têm distribuição normal. Ao analisar os referidos preços durante um determinado mês (21 dias úteis), obtiveram-se os seguintes resultados:

	Preço médio	Desvio-padrão corrigido
Empresa X	4.981	1.011
Empresa Y	4.744	0.722

a) Assumindo que as variâncias das 2 populações são iguais**,** construa um intervalo de confiança a 95% para a diferença dos preços médios das acções. Será que pode concluir que uma das acções é "preferível" à outra?

IC a 95% para $\mu_X - \mu_Y$

$$\mathbf{VF}: T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{m} + \frac{1}{n}} \sqrt{\frac{(m-1)S_X'^2 + (n-1)S_Y'^2}{m+n-2}}} \sim t(m+n-2) = t(40)$$

IC: $(\bar{x} - \bar{y} \mp t_{0.025} \times s^*)$, onde:

- $\bar{x} = 4.981, \bar{y} = 4.744$
- $t_{0.025} = 2.021$
- $s^* = \sqrt{\frac{1}{21} + \frac{1}{21}} \sqrt{\frac{20 \times 1.011^2 + 20 \times 0.722^2}{40}} \approx 0.2711$

IC a 95% para $\mu_X - \mu_Y$: $(4.981 - 4.744 \mp 2.021 \times 0.2711) = (-0.311, 0.785)$

Porque o valor zero pertence ao IC obtido, não se pode afirmar que uma acção é preferível à outra (note que se assumiu iguais volatilidades).

b) Teste a 5% a hipótese feita na alínea anterior no que diz respeito à igualdade de variâncias.

$H_0: \frac{\sigma_X^2}{\sigma_Y^2} = 1 \quad vs \quad H_1: \frac{\sigma_X^2}{\sigma_Y^2} \neq 1$ **ou** $H_0: \sigma_X^2 = \sigma_Y^2 \quad vs \quad H_1: \sigma_X^2 \neq \sigma_Y^2$

$$\mathbf{ET}: F = \frac{S_X'^2}{S_Y'^2} \sim F(m-1, n-1)$$

Onde $m = n = 21$. Logo, $F \sim F(20,20)$.

Região de rejeição: $W_F = \{f_{obs}: f_{obs} < 0.407 \ \vee \ f_{obs} > 2.46\}$

$f_{obs} = \frac{s_X'^2}{s_Y'^2} = \frac{1.011^2}{0.722^2} \approx 1.961 \ \notin \ W_F$ $\qquad$ $valor-p = 0.1406$

Logo, não se rejeita H_0 ao nível de 5%, isto é, não se rejeita que as variâncias sejam iguais.

3) A Ana e o João estão a divertir-se com o jogo "Maratona Louca", que consiste em ver qual o peão que corta primeiro a meta de chegada. Para tal, precisam de lançar um dado e andar o número de casas determinado pelos pontos do dado. No fim do jogo, a Ana ganhou! Porque o jogo pertencia à Ana e cada um deles tinha um dado, o João, algo desconfiado, pediu o dado da Ana para averiguar se este não estaria viciado para sair pontuação mais alta. Após 120 lançamentos do dado, e já saturado, o João obteve os seguintes resultados:

Pontos	1P	2P	3P	4P	5P	6P
Nº vezes	15	18	16	24	24	23

Com base num teste adequado ($\alpha = 0.05$), diga se o dado da Ana é equilibrado. De acordo com o resultado do teste, será que o João tem razão para contestar o jogo?

Teste de ajustamento

$$H_0: p_{0j} = \frac{1}{6} \ (j = 1, 2, \dots, 6) \quad vs \quad H_1: \exists j, p_{0j} \neq \frac{1}{6} \quad (j = 1, 2, \dots, 6)$$

$$\mathbf{ET}: Q = \sum_{j=1}^{m} \frac{\left(N_j - np_{0j}\right)^2}{np_{0j}} \overset{a}{\sim} \chi^2(m-1)$$

Frequências esperadas: $np_{0j} = 120 \times (1/6) = 20 \quad (j = 1, 2, \ldots, 6)$

Todas $fe_j \geq 5 \Rightarrow m = 6 \Rightarrow Q \overset{a}{\sim} \chi^2(5)$

Nº pontos	n_j	np_{0j}	$\frac{(n_j - np_{0j})^2}{np_{0j}}$
1P	15	20	1.25
2P	18	20	0.2
3P	16	20	0.8
4P	24	20	0.8
5P	24	20	0.8
6P	23	20	0.45
TOTAL	**120**	**120**	**4.3**

$q_{obs} = 4.3$ e $\quad valor - p = P(Q \geq q_{obs}|H_0) = P(Q \geq 4.3) \approx 0.507$

Ou região de rejeição ($\alpha = 0.05$): $W_Q = \{q_{obs}: q_{obs} > 11.070\}$

Com este valor-p, não se deve rejeitar H_0 aos níveis habituais, isto é, a evidência estatística é favorável a que o dado da Ana seja equilibrado. O João perdeu honestamente e, como tal, não tem razão para contestar o jogo.

4) O Jorge é um investigador ligado à educação. No seu próximo estudo, ele pretende averiguar se o peso das mochilas que os alunos carregam às costas é excessivo. Para tal, ele pede o seu auxílio para que responda a umas questões preliminares para começar a desenvolver o seu trabalho. Após alguma reflexão, chega-se ao seguinte modelo explicativo do peso das mochilas:

$$peso\_m_t = \beta_1 + \beta_2 peso\_c_t + \beta_3 rapaz_t + \beta_4 num_t + \beta_5 dist_t + u_t \quad (t = 1, 2, \ldots, 250)$$

Onde,

- $peso\_m$ — Peso da mochila, em kg;
- $peso\_c$ — Peso da criança, em kg;
- $rapaz$ — Variável binária (=1 se rapaz / =0 se rapariga);
- num — Número de disciplinas a assistir;
- $dist$ — Distância entre casa e escola, em km.

Recolhida uma amostra casual de 250 alunos, que iniciam o 5º ano de escolaridade, estimou-se o modelo acima recorrendo ao EXCEL, cujos resultados se encontram em anexo, no **Modelo 1**.

a) Interprete as estimativas dos coeficientes β_2 e β_3 e, com base no valor-p, teste a significância estatística de cada um deles ao nível de 10%. (Não se esqueça de apresentar a **estatística-teste**).

$b_2 = 0.07 \rightarrow$ *Ceteris paribus*, se o peso da criança for superior em 1kg, o peso da mochila aumenta, em média, 0.07 kg (70 gramas).

$b_3 = -0.2034 \rightarrow$ *Ceteris paribus*, a mochila dos meninos pesa, em média, menos 0.2034 kg (203.4 gramas) que a mochila das meninas.

$$H_0: \beta_j = 0 \quad vs \quad H_1: \beta_j \neq 0 \quad (j = 2, 3)$$

$$\textbf{ET}: t_j = \frac{b_j - \beta_j^0}{s_{b_j}} \sim t(n-k) = t(245) \quad (j = 2, 3)$$

Pelo output, conclui-se que os valores-p para os testes acima são 0.091 (j=2) e 0.3237 (j=3). Portanto, ao nível de 10%, β_2 é estatisticamente significativo, mas β_3 já não o é.

b) Construa um intervalo de confiança a 95% para o efeito de mais 1 disciplina a assistir sobre o peso da mochila, *ceteris paribus*.

IC a 95% para β_4

$$\mathbf{VF}: t_4 = \frac{b_4 - \beta_4}{s_{b_4}} \sim t(n-k) = t(245)$$

$$\left(b_4 \mp t_{0.025} \times s_{b_4}\right) = (0.0204 \mp 1.96 \times 0.0094) = (0.0020, 0.0388)$$

c) Para testar a hipótese $H_0: \beta_3 = -10\beta_4 \wedge \beta_5 = -0.5$, estimou-se uma regressão auxiliar da qual resultou: $R^2 = 0.0328$, $\bar{R}^2 = 0.02497$ e $VR = \sum_{t=1}^{250} \hat{u}_t^2 = 222.7167$. Escreva a forma da regressão auxiliar estimada e, em seguida, efectue o teste. (Não se esqueça de apresentar as **hipóteses** e a **estatística-teste**).

Regressão auxiliar

$$peso\_m_t = \beta_1 + \beta_2 peso\_c_t + \beta_3 gen_t + \beta_4 disc_t + \beta_5 dist_t + u_t \Rightarrow$$
$$\Rightarrow peso\_m_t = \beta_1 + \beta_2 peso\_c_t + (-10\beta_4) gen_t + \beta_4 disc_t + (-0.5) dist_t + u_t \Rightarrow$$
$$\Rightarrow peso\_m_t + 0.5 dist_t = \beta_1 + \beta_2 peso\_c_t + \beta_4 (disc_t - 10 gen_t) + u_t \Rightarrow$$
$$\Rightarrow y_t^* = \alpha_1 + \alpha_2 peso\_c_t + \alpha_3 x_t^* + v_t$$

Onde: $y_t^* = peso\_m_t + 0.5 dist_t$ e $x_t^* = disc_t - 10 gen_t$

Teste de Hipóteses

$$H_0: \beta_3 = -10\beta_4 \wedge \beta_5 = -0.5 \quad vs \quad H_1: \beta_3 \neq -10\beta_4 \vee \beta_5 \neq -0.5$$

$$\mathbf{ET}: F = \frac{(VR_0 - VR_1)/m}{VR_1/(n-k)} \sim F(m, n-k) = F(2, 245)$$

$$f_{obs} = \frac{(222.7167 - 222.2807)/2}{222.2807/(250-5)} \approx 0.2403$$

$$valor - p = P(F \geq f_{obs} | H_0) = P(F \geq 0.2403) \approx 0.7866$$

Sendo o valor-p elevado, não se rejeita H_0 aos níveis habituais. As restrições são validadas pelos dados.

Região de rejeição (5%): $W_F = \{f_{obs}: f_{obs} > 3\}$

d) Apresente um intervalo de previsão a 90% para o peso médio de uma mochila transportada por um menino de 32 quilos, que vai assistir a 5 disciplinas e que mora a uma distância da escola de 3 quilómetros. Para tal, recorra ao output do **Modelo 2**, em anexo. Escreva também a que correspondem as variáveis identificadas como Z1, Z2, Z3 e Z4 no referido output.

Intervalo de previsão em média

$$\mathbf{VF}: t_{\hat{\theta}} = \frac{\hat{\theta} - \theta}{s_{\hat{\theta}}} \sim t(n-k) = t(245)$$

IC a 90% para θ: $(\hat{\theta} \mp t_{0.05} \times s_{\hat{\theta}})$

Pelo Modelo 2, obtém-se: $\hat{\theta} = 3.4348$ e $s_{\hat{\theta}} = 0.7355$. Logo, o IC vem:

$$\left(\hat{\theta} \mp t_{0.05} \times s_{\hat{\theta}}\right) = (3.4348 \mp 1.645 \times 0.7355) = (2.225, 4.645)$$

$Z1 = peso\_c - 32$; $Z2 = gen - 1$; $Z3 = disc - 5$; $Z4 = dist - 3$

e) O **Modelo 3**, em anexo, fornece a estimação de uma equação auxiliar para efectuar um teste específico. Indique qual a utilidade desta equação e efectue o teste em questão, considerando uma dimensão 5%.

Trata-se da equação auxiliar do Teste de White simplificado, que tem como objectivo testar a hipótese básica do MRL de homocedasticidade condicionada sobre o termo de erro do Modelo 1. Logo, quer-se testar,

$$H_0: \alpha_2 = \alpha_3 = 0 \quad vs \quad H_1: \alpha_2 \neq 0 \vee \alpha_3 \neq 0$$

$$\mathbf{ET}: W_s = nR^2 \overset{a}{\sim} \chi^2(2)$$

Pelo Modelo 3, obtém-se $R^2 = 0.00114$. Logo,

$w_{s,obs} = 250 \times 0.00114 = 0.285$

$valor - p = P\left(W_s \geq w_{s,obs} \middle| H_0\right) = P(W_s \geq 0.285) \approx 0.8672$

Com este valor-p, não se deve rejeitar a hipótese de homocedasticidade aos níveis habituais.
Região de rejeição: $W_{5\%} = \{w_{s,obs}: w_{s,obs} > 5.991\}$

ANEXO

(Modelo 1)

$$peso\_m_t = \beta_1 + \beta_2 peso\_c_t + \beta_3 gen_t + \beta_4 num_t + \beta_5 dist_t + u_t$$

Regression Statistics	
Multiple R	0.2166
R Square	0.0469
Adjusted R Square	0.0314
Standard Error	0.9525
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	10.9434	2.7358	3.0155	0.0187
Residual	245	222.2807	0.9073		
Total	249	233.2241			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.4146	0.7363	3.2794	0.0012
peso_c	0.0700	0.0413	1.6970	0.0910
gen	-0.2034	0.2057	-0.9888	0.3237
num	0.0204	0.0094	2.1576	0.0319
dist	-0.3729	0.1836	-2.0308	0.0434

(Modelo 2)

$$peso\_m_t = \theta + \beta_2\, Z1_t + \beta_3\, Z2_t + \beta_4\, Z3_t + \beta_5\, Z4_t + u_t$$

Regression Statistics	
Multiple R	0.2166
R Square	0.0469
Adjusted R Square	0.0314
Standard Error	0.9525
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	10.9434	2.7358	3.0155	0.0187
Residual	245	222.2807	0.9073		
Total	249	233.2241			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3.4348	0.7355	4.6698	0.0000
Z1	0.0700	0.0413	1.6970	0.0910
Z2	-0.2034	0.2057	-0.9888	0.3237
Z3	0.0204	0.0094	2.1576	0.0319
Z4	-0.3729	0.1836	-2.0308	0.0434

(Modelo 3)

$$\hat{u}_t^2 = \alpha_1 + \alpha_2 \widehat{peso\_m}_t + \alpha_3 \widehat{peso\_m}_t^2 + v_t$$

Regression Statistics	
Multiple R	0.03377
R Square	0.00114
Adjusted R Square	-0.00695
Standard Error	1.22419
Observations	250

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.4226	0.2113	0.1410	0.8686
Residual	247	370.1666	1.4987		
Total	249	370.5892			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4.3173	11.3154	0.3815	0.7031
$\widehat{peso\_m}_t$	-2.5042	7.7356	-0.3237	0.7464
$\widehat{peso\_m}_t^2$	0.4529	1.3185	0.3435	0.7315

Notas: $\hat{u}_t$ – resíduos MQ do Modelo 1; $\widehat{peso\_m}_t$ – valores ajustados do Modelo 1