Instituto Superior de Economia e Gestão
UNIVERSIDADE TÉCNICA DE LISBOA

Estatística II – Licenciatura em Gestão

ER – 2 Julho 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
Num teste do χ^2 à bondade do ajustamento a região de rejeição pode ser bilateral		X
Quando num teste de hipóteses com $\alpha = 0.03$ se obtém um valor-p de 0.047 rejeita-se H_0		X
Seja X uma população com distribuição exponencial. A conjectura $P(X < 2) = 0.4$ pode ser testada recorrendo a um teste de hipóteses paramétricas.	X	
Na estimação do MRL $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t$, $t = 1,2,...,105$, a verificar as hipóteses habituais obteve-se $R^2 = 0.971$. Pode-se então garantir que se rejeita, para os níveis de significância habituais, H_0 no teste $H_0 : \beta_2 = 0$ contra $H_1 : \beta_2 \neq 0$.		X
Quando uma população não é normal a média da amostra é um estimador enviesado da média da população, se esta existir		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. Considere um teste de independência do χ^2 construído com base numa tabela de contingência. Assinale a alternativa correcta
   - ☐ O número de elementos observados em cada célula tem de ser ≥ 5
   - X A hipótese H_0 traduz a independência entre os 2 factores em análise
   - ☐ Os graus de liberdade da qui-quadrado dependem de n, dimensão da amostra.
   - ☐ A região de rejeição é bilateral

   b. Seja o modelo de regressão linear $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t$ com $t = 1,2,...,n$. Quando se refere que o modelo não sofre de autocorrelação está-se a dizer que, para $t, s = 1,2,...,n$,
   - ☐ $\text{cov}(x_{t2}, x_{s3} \mid X) = 0$
   - ☐ $\text{cov}(u_t, u_s \mid X) = \sigma^2$ para $t \neq s$ e $\text{cov}(u_t, u_s \mid X) = 0$ para $t = s$
   - X $\text{cov}(u_t, u_s \mid X) = 0$ para $t \neq s$
   - ☐ Todas as afirmações anteriores são falsas

   c. No modelo de regressão linear $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + u_t$ com $t = 1,2,...,n$ pretende-se testar $H_0 : \beta_2 + \beta_3 = 0 \wedge \beta_4 = \beta_3$ contra $H_1 : H_0$ falsa . Para tal recorre-se à regressão auxiliar dada por
   - X $y_t = \beta_1 + \beta_3(-x_{t2} + x_{t3} + x_{t4}) + \beta_5 x_{t5} + u_t$
   - ☐ $y_t = \beta_1 + \beta_3(x_{t2} + x_{t3} + x_{t4}) + \beta_5 x_{t5} + u_t$
   - ☐ $y_t = \beta_1 + \beta_3 x_{t3} + \beta_5 x_{t5} + u_t$
   - ☐ $y_t - x_{t2} = \beta_1 + \beta_3 x_{t3} + \beta_5 x_{t5} + u_t$

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

a. Defina o conceito de amostra emparelhada e explique o seu interesse.

Uma amostra emparelhada é uma amostra formada por pares de valores $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$. Procura-se que os elementos de cada par sejam tanto quanto possível semelhantes em todos os aspectos menos em relação ao factor cujo efeito se pretende avaliar. Os pares são independentes entre si e identicamente distribuídos.
O interesse das amostras emparelhadas reside em associar a rejeição de $H_0 : \mu_X = \mu_Y$ à influência de um factor específico, aquele que diferencia o comportamento de X_i do de Y_i

b. Considere o MRL, $y_t = \beta_1 + \beta_2 x_t + u_t$, $t = 1, 2, \cdots, n$, onde $u_t = v_t \sqrt{x_t}$, com $E(v_t) = 0$, $\mathrm{var}(v_t) = \sigma^2$ e a variável v independente da variável x. Mostre que a hipótese H2, $E(u_t \mid X) = 0$, (exogeneidade condicionada) se encontra verificada e que a hipótese H4, $\mathrm{var}(u_t \mid X) = \sigma^2$, (homocedasticidade condicionada) se encontra violada.

$$
\begin{aligned}
E(u_t \mid X) &= E\left(v_t \sqrt{x_t} \mid X\right) = \sqrt{x_t}\, E\left(v_t \mid X\right) \\
&= \sqrt{x_t}\, E\left(v_t\right) && \text{indep. entre } v_t \text{ e } X \\
&= 0 && \text{já que } E(v_t) = 0
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}(u_t \mid X) &= \mathrm{var}\left(v_t \sqrt{x_t} \mid X\right) = x_t\, \mathrm{var}\left(v_t \mid X\right) \\
&= x_t\, \mathrm{var}\left(v_t\right) && \text{indep. entre } v_t \text{ e } X \\
&= x_t\, \sigma^2 && \text{já que } \mathrm{var}(v_t) = \sigma^2
\end{aligned}
$$

e portanto $\mathrm{var}(u_t \mid X)$ não é constante

Estatística II – Licenciatura em Gestão

ER – 3 Julho 2013

Parte prática

Nome: ____ Nº____________

Espaço reservado para classificações				
1a. (15)	2a. (15)	4a. (15)	4d. (15)	T:
1b. (10)	2b.(15)	4b. (10)	4e. (15)	P: _____
	3.(15)	4c. (15)		

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. Seja X uma população com $E(X) = \frac{2}{1+\theta}$ e $Var(X) = \frac{2\theta(1-\theta)}{(1+\theta)^2}$, $\theta > 0$ da qual se recolheu uma amostra casual de dimensão $n > 5$.

a. Para estimar $E(X)$ foi proposto o estimador $T = \frac{1}{8}(2X_1 + 5X_2 + kX_5 + 4X_n)$, sendo k uma constante desconhecida. Determine k de forma a garantir que T é estimador centrado para a média da população. Obtenha também a variância do estimador T como função de k e de θ.

T será centrado para a média da população se verificar $E(T) = E(X)$

$$E(T) = E\left[\frac{1}{8}(2X_1 + 5X_2 + kX_5 + 4X_n)\right] = \frac{1}{8}[2E(X_1) + 5E(X_2) + kE(X_5) + 4E(X_n)] = \frac{11+k}{8}\ E(X)$$

Logo só será centrado se $\frac{11+k}{8} = 1 \Rightarrow k = -3$

$$\text{var}(T) = \text{var}\left(\frac{1}{8}\left(2X_1 + 5X_2 + kX_5 + 4X_n\right)\right) = \frac{1}{64}\left(4 + 25 + k^2 + 16\right)\text{var}(X)$$

$$= \frac{45+k^2}{64} \times \frac{2\theta(1-\theta)}{(1+\theta)^2}$$

b. Estime θ pelo método dos momentos.

$$E(X) = \frac{2}{1+\theta}$$

O estimador dos momentos será dado pela solução de $\bar{X} = \frac{2}{1+\theta}$ em ordem a θ, isto é $\tilde{\theta} = \frac{2}{\bar{X}} - 1$

2. Uma máquina de café está afinada para encher um copo com 4 cl de café. A equipa de manutenção suspeita que a máquina está a fornecer menos quantidade de café do que aquela que estava prevista. Como o arranjo da máquina envolve despesas de algum montante a equipa de manutenção decidiu recolher uma amostra casual de 16 cafés tendo obtido $\sum_{i=1}^{16} x_i = 60.0$, $s'^2 = 0.16$. Assuma que a quantidade de café por copo segue uma distribuição normal.

a. Efectue um teste ($\alpha = 0.05$) que permita saber se a máquina deve ser sujeita a manutenção e conclua.

X quantidade de café por copo.

$H_0 : \mu = 4$ contra $H_1 : \mu < 4$

Estatística de teste $T = \dfrac{\bar{X} - \mu_0}{S' / \sqrt{n}} = \dfrac{\bar{X} - 4}{S' / 4} \sim t_{(15)}$

Valor observado da estatística de teste: $t_{obs} = \dfrac{(60/16) - 4}{0.4/4} = -2.5$

Região de rejeição: $W = \{(x_1, x_2, \cdots, x_{16}) : t < -1.753\}$ logo como $t_{obs} \in W$, rejeita-se H_0 ou seja manda-se afinar a máquina.

Alternativamente, $valor - p = P(T < -2.5) = 0.0123 < \alpha$ e tira-se a mesma conclusão

b. Com base na amostra observada, construiu-se, pelo processo habitual, o seguinte intervalo de confiança para a variância do conteúdo do copo (0.096; 0.331). Qual o grau de confiança associado ao intervalo?

O processo habitual utiliza a variável fulcral $Q = \dfrac{(n-1)S'^2}{\sigma^2} \sim \chi^2_{(n-1)}$ e conduz ao intervalo de confiança $\left(\dfrac{15 \times 0.16}{q_2} ; \dfrac{15 \times 0.16}{q_1} \right)$ onde $q_1 : P\left[\chi^2_{(15)} < q_1\right] = \dfrac{\alpha}{2}$ e $q_2 : P\left[\chi^2_{(15)} > q_2\right] = \dfrac{\alpha}{2}$. Resolvendo vem $q_2 = 25$ e $q_1 = 7.25$ logo o grau de confiança é de cerca de 90%.

3. Para melhor identificar o seu mercado alvo a Alfa Romeo conduziu um estudo de mercado. Uma amostra aleatória com 300 observações foi recolhida e cada pessoa seleccionada foi sujeita a um teste de condução findo o qual se classificou a sua atitude ao volante (Defensiva, Agressiva ou Equilibrada). Também se inquiriu, para cada pessoa, o seu modelo de Alfa Romeo preferido de entre 2 alternativas.

	Defensiva	Agressiva	Equilibrada	Total
Mito	60	10	30	100
Giulietta	60	60	80	200
Total	120	70	110	300

Teste (significância de 5%) se existe independência entre a atitude ao volante e o modelo de Alfa Romeo preferido.

$H_o : p_{ij} = p_{io} \times p_{oj}$ $\;(i=1,2;\, j=1,2,3)$ (independência) $H_1 : H_o$ *falsa*

$Q = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(N_{ij} - fe_{ij})^2}{fe_{ij}} \overset{a}{\sim} \chi^2_{(2)}$ (N_{ij} - freq da classe *ij*; $fe_{ij} = \frac{N_{i\circ}N_{\circ j}}{n}$ freq esperada da classe ij)

$W_{5\%} = \{q : q > 5.991\}$ tabela $\chi^2_{(2)}$

Frequências esperadas

	Defensiva	Agressiva	Equilibrada
Mito	40.00	23.33	36.67
Giulietta	80.00	46.67	73.33

$$Q_{obs} = \frac{(60-40)^2}{40} + \frac{(10-23.33)^2}{23.33} + \cdots + \frac{(80-73.33)^2}{73.33} = 28.247$$

Logo como $Q_{obs} \in W_{5\%}$ rejeita-se H_0, isto é, não existe independência entre a atitude ao volante e o modelo de Alfa Romeo preferido.

Alternativamente: valor-p ≈ 0 e tira-se a mesma conclusão com redobrada convicção.

4. Para analisar se o valor médio das rendas numa cidade é influenciado pela existência de uma universidade considerou-se o seguinte modelo:

$$\log(rendas_t)=\beta_1 + \beta_2 \log(pop_t) + \beta_3 \log(rendmedio_t) + \beta_4\, univ_t + \beta_5\, propalug_t + u_t \qquad t=1,2,\cdots,n$$

onde *log(rendas)* é o logaritmo do valor médio das rendas (em euros) na cidade t, *log(pop)* é o logaritmo da população da cidade t, *log(rendmedio)* é o logaritmo do rendimento médio em euros, *univ* variável binária que assume o valor 1 se existe uma universidade na cidade e *propalug* é a proporção de habitações alugadas no total de habitações ocupadas. No **anexo** são disponibilizados os resultados de algumas estimações. Em todas elas a variável dependente é o logaritmo do valor médio das rendas.

a. Interprete as estimativas obtidas para os coeficientes β_2 e β_4. Recorrendo ao valor-p o que pode dizer quanto à significância individual do regressor associado com β_2 ($\alpha = 0.10$)?

$b_2 = -0.0279$ (elasticidade) Tudo o resto igual, quando a população aumenta 1% estima-se que o valor médio esperado das rendas diminua de 0.0279 % aproximadamente, isto é uma variação muito pequena em termos práticos.

$b_4 = 0.1492$ (semi-elasticidade) Tudo o resto igual, quando se trata de uma cidade com universidade o valor médio esperado das rendas é 14.92% mais elevado (aproximadamente) do que nas cidades sem universidade.

A significância individual dos regressores é testada fazendo:

$H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$ Estatística de teste $T = b_j / s_{\hat{b}_j} \sim t(123)$

O valor-p associado com o coeficiente associado com a variável *log(pop)*, 0.1718, claramente superior a $\alpha = 0.1$ leva-nos a aceitar H_0 e concluir pela não significância estatística do regressor log(pop).

b. Interprete o valor obtido para o coeficiente de determinação e teste a significância global da regressão.

$R^2 = 0.8426$ – O modelo explica cerca de 84,27% da variação total do logaritmo do valor médio das rendas.

$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ contra $H_1 : \exists \beta_j \neq 0,\ j = 2,3,4,5$

$$F = \frac{R^2 / 4}{(1-R^2)/123} \sim F_{(4,123)}$$

$f_{obs} = 164.6121$

$valor - p = P(F > 164.6359) \approx 0$ ou $w_F^{5\%} = \{f : f > 2.445\}$

logo rejeita-se H_0 e conclui-se pela significância global da regressão.

c. Teste a 5% se é razoável admitir que um acréscimo de 5% no rendimento médio origine um aumento de 4% no valor médio das rendas de determinada cidade.

$H_0 : \beta_3 = 0.8 \quad contra \quad H_0 : \beta_3 \neq 0.8$

Estatística de teste: $T = \frac{b_3 - 0.8}{s_{b_3}} \sim t(123)$

Valor observado da estatística de teste: $t_{obs} = \frac{0.7996-0.8}{0.043} = -0.0093$

$valor - p = 2P(T > |-0.0093| | H0) \approx 1$ ou $w_T^{5\%} = \{t : |t| > 1.979\}$

Como $t_{obs} \notin w_{0.05}$, não se rejeita H_0 para uma dimensão de teste de 5%, logo parece ser razoável a afirmação do enunciado.

d. Construa um intervalo de previsão a 90% para o valor médio das rendas numa cidade particular, a cidade A que é uma cidade universitária com 50500 habitantes. Considere ainda que, das habitações ocupadas, metade possui um contrato de aluguer e que o rendimento médio na cidade A é de 19 mil euros.

$t = \dfrac{d}{s_d} \sim t_{(123)}$ com $d = y_0 - \hat{y}_0$

Construção do intervalo de previsão pontual para a variável *rendas*

Intervalo de previsão para *log(rendas)*: $\hat{y}_0 \pm t_{0.05} \times s_d$

$\hat{y}_0 = 5.885$

$t_{0.05}(123) \approx 1.645$

$s_d = \sqrt{s^2 + s_{\hat{\theta}}^2} = \sqrt{0.1341^2 + 0.0269^2} \approx 0.138$

Intervalo de confiança para *log(rendas)*=$(5.66; 6.11)$

Intervalo de confiança para *rendas* = (287.188;450.404)

e. Após a estimação do modelo proposto decidiu-se averiguar se era razoável admitir que os erros no modelo inicialmente proposto eram homocedásticos. Apresente uma regressão auxiliar adequada para o efeito. Assuma que se estimou a regressão auxiliar que apresentou tendo-se obtido $R^2 = 0.424$. Efectue o teste considerando uma dimensão de 5%.

2 soluções possíveis: Teste de White ou teste de White simplificado.

<u>Teste de White</u>

Regressão auxiliar

$$\begin{aligned}\hat{u}_t^2 = {} & \alpha_1 + \alpha_2 \log(pop) + \alpha_3 \log(rendmedio) + \alpha_4 univ + \alpha_5 percalug + \alpha_6 \log(pop)^2 + \alpha_7 \log(rendmedio)^2 \\ & + \alpha_8 percalug^2 + \alpha_9 (\log(pop) \times \log(rendmedio)) + \alpha_{10} (\log(pop) \times univ) \\ & + \alpha_{11} (\log(pop) \times percalug) + \alpha_{12} (\log(rendmedio) \times univ) \\ & + \alpha_{13} (\log(rendmedio) \times percalug) + \alpha_{14} (univ \times percalug) + v_t\end{aligned}$$

$H_0: \alpha_2 = \cdots = \alpha_{14} = 0$ contra $H_1: \exists \alpha_j \neq 0 \;\; j = 2, \ldots, 14$

$W = nR^2 \sim \chi^2_{(p-1)}$ com $p - 1 = 13$ e $n = 128$

$W_{obs} = 128 * 0.424 = 54.272$

$valor - p = P(W > 54.272 | H0) \approx 0 \quad ou \quad w_w^{5\%} = \{w: w > 22.362\}$

Como $W_{obs} \in w_w^{5\%}$ (ou porque o valor-p é inferior ao níveis de significância habituais) rejeita-se a hipótese nula de homocedasticidade nos erros.

ou

<u>Teste de White simplificado</u>

Regressão auxiliar

$\hat{u}_t^2 = \alpha_1 + \alpha_2 \log(\widehat{\text{rendas}_\text{t}}) \quad + \alpha_3 \log(\widehat{\text{rendas}_\text{t}})^2 + v_t$

$H_0: \alpha_2 = \alpha_3 = 0$ contra $H_1: \exists \alpha_j \neq 0 \;\; j = 2,3$

$W = nR^2 \sim \chi^2_{(2)}$

$W_{obs} = 128 * 0.424 = 54.272$

$valor - p = P(W > 54.272 | h0) \approx 0 \quad ou \quad w_w^{5\%} = \{w: w > 5.991\}$

Como $W_{obs} \in w_w^{5\%}$ (ou porque o valor-p é inferior ao níveis de significância habituais) rejeita-se a hipótese nula de homocedasticidade nos erros.