ESTATÍSTICA II – Economia/Finanças

Duração: 2h – Época Normal – Maio 2022

Nome:_____ Nº:_____

Espaço reservado para classificações					
1.a)[15]	2.a)[15]	3.[20]	4.a)[10]	4.d)[15]	4.g)[15]
1.b)[15]	2.b)[15]		4.b)[15]	4.e)[15]	
1.c)[20]			4.c)[15]	4.f)[15]	

Atenção: todas as questões devem ser devidamente formalizadas e justificadas. Sempre que fizer um teste de hipóteses formule as hipóteses em teste e apresente a estatística de teste e sempre que fizer um intervalo de confiança apresente a variável fulcral e respectivas distribuições.

1. Assuma que X segue uma distribuição de Poisson de parâmetro desconhecido λ e que foi recolhida uma amostra casual $(X_1, \dots, X_n)$. Foram propostos 2 estimadores para λ: $T_1 = \bar{X}$ e $T_2 = S'^2$.

   a) **[15]** Mostre que T_1 é o estimador de máxima verosimilhança para λ e apresente o estimador de máxima verosimilhança para $P(X = 0)$ referindo a propriedade que está a utilizar.

$$X \sim Po(\lambda) \Leftrightarrow f(x|\lambda) = \frac{e^{-\lambda}\,\lambda^x}{x!} \quad (x = 0, 1, \dots;\ \lambda > 0)$$

$$L(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n}\left(\frac{e^{-\lambda}\,\lambda^{x_i}}{x_i!}\right) = \frac{e^{-n\lambda}\,\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)} = \frac{e^{-n\lambda}\,\lambda^{n\bar{x}}}{\prod_{i=1}^n (x_i!)}\ ,\ \ \lambda > 0$$

$$l(\lambda) = \ln[L(\lambda)] = \ln\left[\frac{e^{-n\lambda}\,\lambda^{n\bar{x}}}{\prod_{i=1}^n (x_i!)}\right] = -n\lambda + n\bar{x}\ln(\lambda) - \sum_{i=1}^{n}\ln(x_i!)\ ,\ \ \lambda > 0$$

$$l'(\lambda) = 0 \Leftrightarrow -n + \frac{n\bar{x}}{\lambda} = 0 \Leftrightarrow \frac{-n\lambda + n\bar{x}}{\lambda} = 0 \Leftrightarrow -n\lambda + n\bar{x} = 0 \Leftrightarrow n\lambda = n\bar{x} \Leftrightarrow \lambda = \bar{x}$$

$$l''(\lambda) = -\frac{n\bar{x}}{\lambda^2} < 0\,,\ \ \forall \lambda > 0$$

Logo, o estimador de máxima verosimilhança para λ é: $\hat{\lambda} = \bar{X} = T_1$.

$P(X = 0) = f(0|\lambda) = \frac{e^{-\lambda}\,\lambda^0}{0!} = e^{-\lambda}$ é uma função biunívoca de λ. Logo, pela propriedade da invariância dos estimadores de máxima verosimilhança, o estimador de máxima verosimilhança para $P(X = 0)$ é dado por:

$$\hat{P}(X = 0) = \widehat{(e^{-\lambda})} = e^{-\hat{\lambda}} = e^{-\bar{X}}$$

b) **[15]** Sabendo que $Var(T_2) = \frac{\lambda}{n} \cdot \frac{2n\lambda+n-1}{n-1}$ compare os dois estimadores em termos de eficiência. Qual deles lhe parece ser o mais adequado segundo este critério?

Note-se que ambos os estimadores, T_1 e T_2 são centrados para λ já que:

$$E(T_1) = E(\bar{X}) = E(X) = \lambda \quad \text{e} \quad E(T_2) = E(S'^2) = Var(X) = \lambda$$

Comparando as respectivas variâncias:

$Var(T_1) = Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\lambda}{n}$ e $Var(T_2) = \frac{\lambda}{n} \cdot \frac{2n\lambda+n-1}{n-1} = \frac{\lambda}{n}\left(\frac{2n\lambda}{n-1} + 1\right)$,

conclui-se que $Var(T_1) < Var(T_2)$, pois $\frac{2n\lambda}{n-1} + 1 > 1, \ \forall\lambda > 0$.

Logo, T_1 é mais eficiente que T_2 sendo o estimador preferível segundo este critério.

c) **[20]** Considere agora um outro estimador alternativo para λ, definido como: $T_3 = aT_1 + bT_2$, onde $a, b \in \mathbb{R}$. Qual a relação que deve existir entre as constantes a e b de maneira que T_3 seja um estimador centrado para λ? A partir da relação obtida e sem fazer contas adicionais, será que se consegue definir os valores que originam a menor variância para o estimador T_3? Justifique.

T_3 centrado para $\lambda \Leftrightarrow E(T_3) = \lambda \Leftrightarrow E(aT_1 + bT_2) = \lambda \Leftrightarrow aE(T_1) + bE(T_2) = \lambda \Leftrightarrow a\lambda + b\lambda = \lambda \Leftrightarrow$

$\Leftrightarrow a + b = 1$. Logo, para que T_3 seja centrado para λ deve verificar-se a relação $a + b = 1$.

Sabe-se que a existir estimador mais eficiente, este será estimador de máxima verosimilhança. Com efeito, o limite mínimo para a variância dado pelo Teorema de Fréchet-Cramer-Rao:

$$\frac{1}{n\Im(\lambda)} = \frac{1}{n\frac{1}{\lambda}} = \frac{\lambda}{n} = Var(T_1).$$

Logo, T_3 terá menor variância se coincidir com T_1, ou seja, se $a = 1$ e $b = 0$.

2. A empresa Metropolitano de Lisboa tem em curso um inquérito para avaliar a satisfação dos seus utentes. Considere que uma de entre várias perguntas era:

> "Avalie numa escala de 0 (muito insatisfeito) a 100 (muito satisfeito) o quão satisfeito está com o serviço prestado".

Assuma sempre que necessário que as respostas à pergunta assumem uma distribuição normal. Da amostra observada (180 respostas) obteve-se uma média de 57.6 e um desvio-padrão corrigido de 12.5.

a) **[15]** Defina um intervalo de confiança a 90% para o valor médio do grau de satisfação.

$X -$ grau de satisfação; $X \sim N(\mu, \sigma^2)$

VF: $T = \dfrac{\bar{X} - \mu}{S'/\sqrt{n}} \sim t(179)$

IC a 90% para μ: $\left(\bar{x} \pm t_{0.05} \times \dfrac{s'}{\sqrt{n}}\right) = \left(57.6 \pm 1.645 \times \dfrac{12.5}{\sqrt{180}}\right) = (56.07, 59.13)$

Com uma confiança de 90% estima-se que o valor médio do grau de satisfação se situa entre 56.07 e 59.13.

b) **[15]** O Conselho de Administração da empresa definiu fixar, como objetivo principal, que o valor esperado do grau de satisfação fosse de pelo menos 60. Utilizando um teste adequado e com base no valor-p conclua sobre o cumprimento ou não desse objetivo.

$H_0: \mu \geq 60 \quad vs \quad H_1: \mu < 60$

ET: $T = \dfrac{\bar{X} - 60}{S'/\sqrt{180}} \sim t(179)$

$T_{obs} = \dfrac{57.6 - 60}{12.5/\sqrt{180}} \approx -2.576$

Valor-p (usando a Tabela 7): $p_{obs} = P(T \leq T_{obs} | H_0) = P(T \leq -2.576) = P(T \geq 2.576) = 0.005$

Sendo o valor-p inferior aos níveis de significância habituais, existe considerável evidência estatística desfavorável a H_0, devendo esta ser rejeitada. Logo, a conclusão do teste remete para o não cumprimento do objetivo estipulado.

3. **[20]** Para testar H_0: $X \sim N(0,1)$, obtiveram-se os seguintes dados:

x	$]-\infty,-3[$	$[-3,-2[$	$[-2,-1[$	$[-1,0[$	$[0,1[$	$[1,2[$	$[2,3[$	$[3,+\infty[$
n_j	3	8	40	72	80	35	10	2
fe_j	?	5.375	33.975	85.325	85.325	33.975	5.375	?

Efetue o teste em questão e conclua, usando uma dimensão de 5%.

$H_0: X \sim N(0,1) \quad vs \quad H_1: X \nsim N(0,1)$

ET: $Q = \sum_{j=1}^{m} \frac{(n_j - fe_j)^2}{fe_j} \underset{\sim}{a} \chi^2(m-1)$

$fe_1 = nP(X < -3) = (3+8+40+72+80+35+10+2)\Phi(-3) = 250[1-\Phi(3)] =$

$= 250 \times (1 - 0.9987) = 250 \times 0.0013 = 0.325$

Pela simetria da distribuição normal, tem-se que $P(X > 3) = P(X < -3)$, pelo que: $fe_8 = 0.325$.

Verifica-se que $fe_1 < 5$ e $fe_8 < 5$, pelo que efetuando os agrupamentos necessários, obtém-se:

x	$]-\infty,-2[$	$[-2,-1[$	$[-1,0[$	$[0,1[$	$[1,2[$	$[2,+\infty[$
n_j	11	40	72	80	35	12
fe_j	5.7	33.975	85.325	85.325	33.975	5.7

Tem-se $m = 6$, pelo que $Q \underset{\sim}{a} \chi^2(5)$ e $W_{0.05} = \{Q_{obs}: Q_{obs} > 11.07\}$

$$Q_{obs} = \frac{(11-5.7)^2}{5.7} + \frac{(40-33.975)^2}{33.975} + \frac{(72-85.325)^2}{85.325} + \frac{(80-85.325)^2}{85.325} + \frac{(35-33.975)^2}{33.975} +$$

$$+ \frac{(12-5.7)^2}{5.7} \approx 15.4$$

Porque $Q_{obs} \in W_{0.05}$, rejeita-se H_0: $X \sim N(0,1)$ ao nível de 5%.

4. Num estudo que visa as principais determinantes das vendas em lojas de roupa em certo país da Zona Euro, usou-se o seguinte modelo:

$$\ln(vendas) = \beta_0 + \beta_1 tfull + \beta_2 tpart + \beta_3 hfunc + \beta_4 dim + \beta_5 idade + \beta_6 idade^2 + \beta_7 \ln(inv) + u$$

Onde:

- $vendas -$ vendas anuais da loja, em milhares de euros;
- $tfull -$ número de trabalhadores em *full-time* na loja;
- $tpart -$ número de trabalhadores em *part-time* na loja;
- $hfunc -$ número mensal de horas com a loja em funcionamento;
- $dim -$ dimensão da loja, em m^2;
- $idade$ – idade da loja, em anos;
- $inv-$ investimento realizado na loja, em milhares de euros.

Recolhida uma amostra aleatória de 350 lojas desse país, estimou-se o modelo apresentado anteriormente pelo método dos Mínimos Quadrados Ordinários (OLS), cujo resultado encontra-se na **Equação 1**, em **Anexo**.

a) **[10]** Analise a significância global da regressão ao nível de 1%.

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0 \quad vs \quad H_1: \exists \beta_j \neq 0 \ (j = 1,2,3,4,5,6,7)$

ET: $F = \dfrac{R^2}{1-R^2} \times \dfrac{n-k-1}{k} \sim F(7,342)$

$W_{0.01} = \{F_{obs}: F_{obs} > 2.64\}$

$F_{obs} = \dfrac{0.5665}{1-0.5665} \times \dfrac{342}{7} \approx 63.85 \in W_{0.01}$

Logo, rejeita-se H_0 ao nível de 1% e conclui-se que o modelo é globalmente significativo.

b) **[15]** Interprete as estimativas dos coeficientes associados aos regressores "$hfunc$" e "$\ln(inv)$", e analise a respetiva significância individual ao nível de 10%.

$\hat{\beta}_3 = 0.0074 \rightarrow$ tudo o resto constante, por cada hora adicional em funcionamento por mês, estima-se que as vendas anuais aumentam em média aproximadamente 0.74%.

$\hat{\beta}_7 = 0.0253 \rightarrow$ tudo o resto constante, se o investimento aumentar 1%, estima-se que as vendas anuais aumentam em média aproximadamente 0.0253%.

$H_0: \beta_3 = 0 \quad vs \quad H_1: \beta_3 \neq 0$ ET: $t_3 = \dfrac{\hat{\beta}_3}{se(\hat{\beta}_3)} \sim t(342)$ Pelo output, observa-se que o valor-p deste teste é $p_{obs} = 0 < \alpha = 0.1$, pelo que se rejeita H_0 ao nível de 10%, o que significa que as horas de funcionamento é um fator estatisticamente relevante.	$H_0: \beta_7 = 0 \quad vs \quad H_1: \beta_7 \neq 0$ ET: $t_7 = \dfrac{\hat{\beta}_7}{se(\hat{\beta}_7)} \sim t(342)$ Pelo output, observa-se que o valor-p deste teste é $p_{obs} = 0.3093 > \alpha = 0.1$, pelo que não se rejeita H_0 ao nível de 10%, o que significa que o investimento não é um fator estatisticamente relevante.

c) **[15]** Comente sobre a veracidade da afirmação: em média, tudo o resto constante, o efeito sobre as vendas de mais um trabalhador em *full-time* é superior ao de mais um trabalhador em *part-time*. Justifique a sua resposta com base num teste com uma dimensão de 5%.

$H_0: \beta_1 - \beta_2 \leq 0 \quad vs \quad H_1: \beta_1 - \beta_2 > 0$

ET: $t = \dfrac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t(342)$

$W_{0.05} = \{t_{obs}: t_{obs} > 1.645\}$

Para calcular o valor observado da estatística-teste, tem-se:

- $\hat{\beta}_1 - \hat{\beta}_2 = 0.1191 - 0.0839 = 0.0352$
- $se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\widehat{Var}(\hat{\beta}_1 - \hat{\beta}_2|X)} = \sqrt{\widehat{Var}(\hat{\beta}_1|X) + \widehat{Var}(\hat{\beta}_2|X) - 2\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2|X)} =$

$= \sqrt{0.002058 + 0.001128 - 2 \times 0.000379}$ (Usando a matriz $\widehat{Cov}(\hat{\beta}|X)$ dada)

Então,

$t_{obs} = \dfrac{0.0352}{\sqrt{0.002058 + 0.001128 - 2 \times 0.000379}} \approx 0.714 \notin W_{0.05}$

Logo, não se rejeita H_0 ao nível de 5%, o que sugere que a afirmação não é verdadeira.

d) **[15]** Uma maneira de avaliar o impacto de mais um trabalhador em geral (isto é, independentemente do tipo de vínculo laboral) seria construir uma nova variável, $ntrab = tfull + tpart$, e incluí-la no modelo. Comente sucintamente sobre o principal problema deste procedimento e apresente uma possível solução.

Este procedimento originará colinearidade perfeita entre a nova variável *ntrab* e as outras existentes, *tfull* e *tpart*, o que é uma violação de uma das hipóteses básicas do modelo de regressão linear (a matriz dos regressores X não tem característica máxima), tornando impossível obter matematicamente as estimativas OLS para os parâmetros do modelo (já que a matriz $X^T X$ não é invertível).
De acordo com o objetivo pretendido, a solução passará por introduzir a nova variável *ntrab* e remover as outras duas, *tfull* e *tpart*, do modelo.

e) **[15]** Teste a 5% a significância estatística da idade da loja. A partir de que idade se estima que o impacto sobre as vendas passa a ser negativo?

$H_0: \beta_5 = 0 \wedge \beta_6 = 0 \quad vs \quad H_1: \beta_5 \neq 0 \vee \beta_6 \neq 0$

ET: $F = \dfrac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} \times \dfrac{n-k-1}{q} \sim F(2, 342)$

$W_{0.05} = \{F_{obs}: F_{obs} > 3\}$

$F_{obs} = \dfrac{0.5665 - 0.5274}{1 - 0.5665} \times \dfrac{342}{2} \approx 15.42 \in W_{0.05}$

Logo, rejeita-se H_0 ao nível de 5% e conclui-se que a idade é um fator estatisticamente significativo.

A relação quadrática estimada entre $\ln(vendas)$ e a idade é côncava e o efeito *ceteris paribus* estimado da idade sobre as vendas é aproximadamente dado por:

$$\frac{\partial \ln(\widehat{vendas})}{\partial idade} = 0.1353 + 2 \times (-0.0015)\, idade = 0.1353 - 0.003\, idade$$

E o maximizante,

$$\frac{\partial \ln(\widehat{vendas})}{\partial idade} = 0 \Leftrightarrow 0.1353 - 0.003\, idade = 0 \Leftrightarrow idade = \frac{-0.1353}{-0.003} = 45.1\,,$$

o que significa que o impacto estimado da idade sobre as vendas passa a ser negativo aproximadamente a partir dos 45.1 anos.

f) **[15]** Apresente uma previsão para as vendas de uma loja em particular com 7 trabalhadores em *full-time* e 3 em *part-time*, que funciona 360 horas por mês, tem uma dimensão de 150 m², 20 anos de idade e com um investimento realizado de 50 mil euros. Para tal, assuma a normalidade do termo de erro.

Assumindo a normalidade do termo de erro, a previsão para as vendas de uma loja nestas condições é dada por:

$$\widehat{vendas}_0 = \exp\left(\frac{\hat{\sigma}^2}{2}\right) \exp\left[\ln(\widehat{vendas})_0\right]$$

Onde:

- $\hat{\sigma} = 0.4791$
- $\ln(\widehat{vendas})_0 = 0.5958 + 0.1191 \times 7 + 0.0839 \times 3 + 0.0074 \times 360 + 0.0015 \times 150 +$
  $+0.1353 \times 20 - 0.0015 \times 20^2 + 0.0253 \times \ln(50) \approx 6.775$

Logo, prevê-se que o volume anual de vendas desta empresa seja de:

$$\widehat{vendas}_0 = \exp\left(\frac{0.4791^2}{2}\right) \exp(6.775) \approx 982.174 \text{ milhares de euros}$$

g) **[15]** Qual o objetivo da **Equação 3** em Anexo? Efetue o teste correspondente a 5% e teça um pequeno comentário à sua conclusão.

O objetivo da Equação 3 em Anexo,

$$\ln(vendas) = \beta_0 + \beta_1 tfull + \cdots + \beta_7 \ln(inv) + \delta_1 \ln(\widehat{vendas})^2 + \delta_2 \ln(\widehat{vendas})^3 + v$$

é o de efetuar um teste RESET ao modelo inicial.

$H_0: \delta_1 = 0 \ \wedge \ \delta_2 = 0 \quad vs \quad H_1: \delta_1 \neq 0 \ \vee \ \delta_2 \neq 0$

ET: $F = \dfrac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} \times \dfrac{n - k - 3}{q} \sim F(2, 340)$

$W_{0.05} = \{F_{obs}: F_{obs} > 3\}$

$F_{obs} = \dfrac{0.6604 - 0.5665}{1 - 0.6604} \times \dfrac{340}{2} \approx 47.005 \in W_{0.05}$

Logo, rejeita-se H_0 ao nível de 5% e tem-se evidência estatística de má especificação da forma funcional do modelo proposto. Neste caso, toda a inferência estatística e interpretações realizadas são inválidas, devendo-se então reformular o modelo inicial.

ANEXO

EQUAÇÃO 1 – Regressando: ln($vendas$)

Regression Statistics	
Multiple R	0.7526
R Square	0.5665
Adjusted R Square	0.5576
Standard Error	0.4791
Observations	350

ANOVA

	df	*SS*	*MS*	*F*
Regression	7	102.5878	14.6554	
Residual	342	78.5142	0.2296	
Total	349	181.1021		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.5958	0.8126	0.7332	0.4640
tfull	0.1191	0.0454	2.6242	0.0091
tpart	0.0839	0.0336	2.4981	0.0130
hfunc	0.0074	0.0007	11.1162	0.0000
dim	0.0015	0.0003	4.6600	0.0000
idade	0.1353	0.0292	4.6335	0.0000
idade2	-0.0015	0.0004	-4.0391	0.0001
ln(inv)	0.0253	0.0248	1.0182	0.3093

A respectiva matriz de variâncias-covariâncias estimada, $\widehat{Cov}(\hat{\beta}|X)$, é dada por:

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$
$\hat{\beta}_0$	0.660326	-0.023668	-0.012746	-0.000261	2.81E-05	-0.015360	0.000198	-0.002629
$\hat{\beta}_1$	-0.023668	0.002058	0.000379	1.52E-05	-3.34E-06	-2.48E-05	2.37E-08	8.79E-05
$\hat{\beta}_2$	-0.012746	0.000379	0.001128	1.20E-05	-1.06E-06	-6.47E-05	6.69E-07	9.55E-05
$\hat{\beta}_3$	-0.000261	1.52E-05	1.20E-05	4.44E-07	-1.08E-07	-3.38E-08	-1.04E-08	-4.32E-07
$\hat{\beta}_4$	2.81E-05	-3.34E-06	-1.06E-06	-1.08E-07	1.06E-07	7.34E-07	-9.38E-09	-1.20E-06
$\hat{\beta}_5$	-0.015360	-2.48E-05	-6.47E-05	-3.38E-08	7.34E-07	0.000853	-1.10E-05	-4.56E-05
$\hat{\beta}_6$	0.000198	2.37E-08	6.69E-07	-1.04E-08	-9.38E-09	-1.10E-05	1.47E-07	5.95E-07
$\hat{\beta}_7$	-0.002629	8.79E-05	9.55E-05	-4.32E-07	-1.20E-06	-4.56E-05	5.95E-07	0.000616

EQUAÇÃO 2 – Regressando: ln($vendas$)

Regression Statistics	
Multiple R	0.7262
R Square	0.5274
Adjusted R Square	0.5206
Standard Error	0.4988
Observations	350

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3.0865	0.6445	4.7887	0.0000
tfull	0.1367	0.0470	2.9098	0.0039
tpart	0.1020	0.0348	2.9314	0.0036
hfunc	0.0079	0.0007	11.7753	0.0000
dim	0.0014	0.0003	4.1292	0.0000
ln(inv)	0.0323	0.0258	1.2508	0.2119

EQUAÇÃO 3 – Regressando: ln($vendas$)

Regression Statistics	
Multiple R	0.8126
R Square	0.6604
Adjusted R Square	0.6514
Standard Error	0.4253
Observations	350

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-39.8663	12.8740	-3.0966	0.0021
tfull	2.6509	0.9495	2.7919	0.0055
tpart	1.7925	0.6579	2.7248	0.0068
hfunc	0.1697	0.0579	2.9285	0.0036
dim	0.0348	0.0119	2.9265	0.0037
idade	3.0052	1.0459	2.8733	0.0043
idade2	-0.0345	0.0120	-2.8813	0.0042
ln(inv)	0.5616	0.1989	2.8236	0.0050
lvendas_hat2	-2.6701	1.1532	-2.3154	0.0212
lvendas_hat3	0.1023	0.0565	1.8117	0.0709

Nota: "lvendas_hat2" e "lvendas_hat3" são os valores ajustados da **EQUAÇÃO 1** ao quadrado e ao cubo, respetivamente.