ESTATÍSTICA II (Eco+Fin) - 2º Ano/1º sem.

Época de Recurso – 04.02.2020

1 hora (14 valores)

Nome:_____

Espaço reservado para classificações			
1 a) (10)	2 a) (10)	4 a) (10)	5 a) (10)
1 b) (10)	2 b) (15)	4 b) i) (10)	5 b) (10)
1 c) (10)	2 c) (10)	4 b) ii) (10)	
	3 (10)	4 c) (15)	
T:			

NOTA: em todos os testes utilize um nível de significância de 5%.

1. Seja uma variável aleatória com distribuição uniforme $X \sim U(0,\theta)$, com $\theta > 0$.

(a) Determine o estimador pelo método dos momentos para o parâmetro θ.

$\mu_1' = \bar{X} \Leftrightarrow E(X) = \bar{X} \Leftrightarrow \frac{\theta}{2} = \bar{X} \Leftrightarrow \tilde{\theta} = 2\bar{X}$

(b) Mostre que a média da amostra é um estimador enviesado para θ e encontre um estimador centrado para o parâmetro θ.

$E(\bar{X})=\mu = \frac{\theta}{2} \neq \theta$, então $\bar{X}$ é um estimador enviesado para θ

$E(2\bar{X})=\mu = 2\frac{\theta}{2} = \theta$, então $2\bar{X}$ é um estimador centrado para θ

(c) De uma população normal, extraíram-se várias amostras de dimensão $\boldsymbol{n}$. Comente a afirmação " A variância da média da amostra é sempre igual à variância da população". Corrija, demonstrando, se achar necessário.

A afirmação está errada. De facto

$$Var(\bar{X}) = \overbrace{Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)}^{porque\ X_i\ são\ independentes} = \underbrace{\frac{1}{n^2}\sum_{i=1}^{n} Var(X)}_{porque\ X_i\ são\ i.d.a\ X} = \frac{1}{n^2} n * Var(X) = \frac{\sigma^2}{n} < \sigma^2$$

para $n > 1$, pelo que se pode concluir que a variância da média da amostra subestima a variância da população.

2. O controlo de qualidade de certos produtos, numa unidade fabril, é feito através da inspeção por amostragem desses produtos. Considere que o número de defeitos em cada produto é uma variável aleatória com distribuição de Poisson com média desconhecida. Numa amostra

aleatória de 90 produtos registou-se um total de 9 defeitos. O intervalo de confiança a 99% para o número esperado de defeitos por produto para a amostra particular selecionada é $(0.0141, 0.1859)$

a. A unidade fabril garante que "O número médio de defeitos por produto é igual a 0.05". Comente a afirmação.

   Não se pode afirmar que "O número médio de defeitos por produto é igual a 0.05", mas como $0.05 \in IC_{\lambda}^{99\%}$ pode afirmar-se que existem 99% de probabilidade de número médio de defeitos por produto ser igual a 0.05, isto é, a garantia é verdadeira

b. Determine o nível de confiança que reduz para 0.1 a amplitude do intervalo de confiança para a mesma amostra? Comente o resultado encontrado com base nos conceitos de precisão e confiança na estimação de parâmetros.

   Amplitude do $IC_{\lambda}^{(1-\alpha)100\%} = 2z\alpha_{/2}\sqrt{\frac{\bar{x}}{n}} = 0.1 \Leftrightarrow z\alpha_{/2} = \frac{0.1}{2\sqrt{\frac{\bar{x}}{n}}} \Leftrightarrow z\alpha_{/2} = \frac{0.1}{2\sqrt{\frac{0.1}{90}}} \Leftrightarrow z\alpha_{/2} = 1.5$

   $P(-1.5 < Z < 1.5) = 2\boldsymbol{\Phi}(1.5) - \mathbf{1} = \mathbf{2} * \mathbf{0.9332} - \mathbf{1} = \mathbf{0.866}$

   Pelo que o nível de confiança teria de ser aproximadamente 87%. Este é o resultado esperado, visto que uma maior precisão do resultado consubstanciado numa menor amplitude do intervalo só é possível reduzindo a confiança.

c. Explique **como** e **porque** varia a amplitude de um intervalo de confiança com a dimensão da amostra, mantendo-se constante o nível de confiança. Qual o efeito na precisão da estimativa?

   A amplitude de um intervalo de confiança diminui, o que significa que o estimador se torna mais preciso, com o aumento da dimensão da amostra porque qto maior for o número de elementos na amostra maior é a informação sobre o comportamento da população.

3. Uma sondagem sobre as intenções de voto feita a 1000 eleitores nos EUA resultou na seguinte tabela de contingência:

		Sexo	
		Masculino	**Feminino**
Intenção de voto	**Republicano**	200	250
	Democrata	150	300
	Independente	50	50

Com base na informação que consta desta tabela e para um nível de significância de 5%, comente a independência entre o sexo do individuo e a sua inclinação partidária.

$H_0: p_{ij} = p_i \ p_j \quad \forall(i,j)$ contra $H_1: p_{ij} \neq p_i \ p_j \quad algum\ (i,j)\ (i = 1,2,3; j = 1,2)$

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(\boldsymbol{n_{ij}} - n.\widehat{p_{i.}}\ \widehat{p_{J.}}\right)^2}{n\,\widehat{p_{i.}}\ \widehat{p_{J.}}} \sim \boldsymbol{\chi}^2_{\left[\frac{(3-1)(2-1)}{2}\right]}$$

Frequências esperadas			
	Sexo		
Row variable	**Masculino**	**Feminino**	Total
Republicano	180	270	450
Democrata	180	270	450
Independente	40	60	100
Total	400	600	1000

$q_{observ} = 16.2037 \Rightarrow valor - p = P\left(\boldsymbol{\chi}^2_{(2)} > 16.2037\right) = 0.0003$

Ou $W = \{q: q > q_\alpha\}$ com $q_\alpha: P(Q > q_\alpha) = 0.05 \Rightarrow q_\alpha = 5.9915$

Então $q_{observ} \in W \Rightarrow$ rejeita-se a hipótese nula de independência, isto é, o sexo do indivíduo não é independente da sua inclinação partidária.

4. Por forma a estudar os determinantes das margens de lucro dos *franchisados* da empresa *XPTOStyle*, foi estimado o seguinte modelo (via Mínimos Quadrados Ordinários):

$$margem = \beta_0 + \beta_1 \log(vendas) + \beta_2\, ncomp + \beta_3\, nparc + \beta_4\, ntemp + u$$

O *output* do modelo estimado:

Dependent Variable: MARGEM
Method: Least Squares
Included observations: 400

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	23.77211	3.932060	6.045713	0.0000
LOG(VENDAS)	1.703269	0.446501	3.814703	0.0002
NCOMP	0.195730	0.262909	0.744476	0.4570
NPARC	1.276986	0.371987	3.432877	0.0007
NTEMP	-1.473138	0.639404	-2.303925	0.0217

R-squared	0.083776	Mean dependent var	38.77423
Adjusted R-squared	0.074498	S.D. dependent var	5.218184
S.E. of regression	5.020050	Akaike info criterion	6.077178
Sum squared resid	9954.354	Schwarz criterion	6.127071
Log likelihood	-1210.436	Hannan-Quinn criter.	6.096936
F-statistic	9.029367	Durbin-Watson stat	1.859872
Prob(F-statistic)	0.000001		

Em que:

- $margem$, corresponde à margem de lucro do *franchisado i*;
- $vendas$, corresponde às vendas do *franchisado i* (em centenas de euros);
- $ncomp$, corresponde ao número de trabalhadores a tempo integral, do *franchisado i*;
- $nparc$, corresponde ao número de trabalhadores a tempo parcial, do *franchisado i*;

- $ntemp$, corresponde ao número de trabalhadores temporários (sazonais), do *franchisado i*.

(a) Interprete as estimativas para β_1 e para β_2.

$\beta_1 -$ um aumento de 1% nas vendas do *franchisado*, em média, tudo o resto constante, implicará um aumento aproximado de 0.017 unidades da margem de lucro do *franchisado i*

$\beta_2 -$ um aumento de 1 de trabalhador a tempo integral adicional, do *franchisado*, em média, tudo o resto constante, implicará um aumento aproximado de 0.1957 unidades da margem de lucro do *franchisado i*

(b) Foi estimado um outro modelo:

Dependent Variable: MARGEM
Method: Least Squares
Included observations: 400

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	19.43422	4.748420	4.092776	0.0001
LOG(VENDAS)	1.740787	0.444215	3.918790	0.0001
NCOMP	-0.801716	0.930328	-0.861756	0.3893
NPARC	3.475134	0.841012	4.132088	0.0000
NTEMP	1.676619	2.537189	0.660818	0.5091
NCOMP^2	0.207267	0.158792	1.305272	0.1926
NPARC^2	-0.409514	0.142599	-2.871792	0.0043
NTEMP^2	-0.782505	0.597350	-1.309961	0.1910

R-squared	0.106448	Mean dependent var	38.77423
Adjusted R-squared	0.090492	S.D. dependent var	5.218184
S.E. of regression	4.976485	Akaike info criterion	6.067122
Sum squared resid	9708.039	Schwarz criterion	6.146951
Log likelihood	-1205.424	Hannan-Quinn criter.	6.098736
F-statistic	6.671221	Durbin-Watson stat	1.862591
Prob(F-statistic)	0.000000		

i. Obtenha e interprete o ponto de viragem (caso exista) do regressor $nparc$.

Ponto de viragem$= nparc = \frac{-\hat{\beta}_3}{2*\hat{\beta}_6} = \frac{-3.475134}{2*(-0.409514)} = 4.243$

Interpretação: o ponto de viragem do regressor $nparc$ indica que a partir de aproximadamente 4 trabalhadores a tempo parcial, o efeito parcial sobre margem de lucro de um dado *franchisado* passa a ser negativo.

ii. O que pode concluir acerca da significância conjunta dos regressores adicionados ao modelo?

$$H_0: \beta_5 = \beta_6 = \beta_7 = 0, \quad contra \;\; H_1: \exists\, \beta_j \neq 0 \;\; (j = 5,6,7)$$

estatística teste: $F = \frac{(R^2 - R_*^2)/(m)}{(1-R^2)/(n-k-1)} \sim F_{(m,\, n-k-1)}$

$$f_{0bs} = \frac{(0.1064 - 0.0838)/3}{(1 - 0.1064)/(400 - 7 - 1)} = 3.3047$$

$valor - p = P(F_{(3, 392)} > 3.3047) = 0.0203 < 0.05$, então rejeita-se H_0 e os regressores adicionados ao modelo são conjuntamente significativos.

(c) Considere o seguinte *output*:

Dependent Variable: RES^2
Method: Least Squares
Included observations: 400

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	197.7107	51.99378	3.802583	0.0002
LOG(VENDAS)	-16.49139	4.864026	-3.390481	0.0008
NCOMP	20.35428	10.18682	1.998100	0.0464
NPARC	-11.85814	9.208827	-1.287693	0.1986
NTEMP	-54.85223	27.78146	-1.974419	0.0490
NCOMP^2	-2.769698	1.738730	-1.592943	0.1120
NPARC^2	1.821772	1.561414	1.166745	0.2440
NTEMP^2	14.55501	6.540804	2.225263	0.0266

R-squared	0.053599	Mean dependent var	24.27010
Adjusted R-squared	0.036699	S.D. dependent var	55.51930
S.E. of regression	54.49103	Akaike info criterion	10.85375
Sum squared resid	1163955.	Schwarz criterion	10.93358
Log likelihood	-2162.749	Hannan-Quinn criter.	10.88536
F-statistic	3.171537	Durbin-Watson stat	2.018326
Prob(F-statistic)	0.002825		

Onde $RES\text{^}2$ corresponde ao quadrado dos resíduos do modelo estimado na alínea (b). Qual a informação que dele podemos retirar? Com base na conclusão anterior, discuta a validade dos resultados obtidos até aqui.

$$F = \frac{R^2_{\hat{u}^2}/(m)}{\left(1 - R^2_{\hat{u}^2}\right)/(n - m - 1)} \sim F_{(m,\, n-m-1)}$$

$$f_{0bs} = \frac{0.0536/7}{(1 - 0.0536)/(400 - 7 - 1)} = 3.1716$$

$valor - p = P(F_{(7, 392)} > 3.1716) = 0.0028 < 0.05$,

Ou $LM = nR^2_{\hat{u}^2} \sim \chi^2_{(m)} \Rightarrow LM = 400 * 0.0536 = 21.44$ e

$$P\left(\chi^2_{(7)} > 21.44\right) = 0.0032 < 0.05$$

então rejeita-se H_0 e pode afirmar-se que o modelo sofre de heterocedasticidade

5. Considere o modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

em que u_i tem média nula e variância constante. Sabe-se ainda que uma amostra $\{(x_{1i}, x_{2i}, y_i): i = 1,2, \dots ,1000\}$ foi retirada da população. Um aluno de Estatística II estimou a seguinte equação (via Mínimos Quadrados Ordinários):

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i}$$

Formalize e discuta o enviesamento e consistência (assim como a sua direção) do estimador dos Mínimos Quadrados Ordinários, nas seguintes situações:

(a) A variável x_2 está positivamente correlacionada com a variável x_1, mas não tem qualquer efeito sobre y.

Se a introdução da variável x_2 não tem qualquer efeito sobre y_i é porque $\beta_2 = 0$, o que significa que a variável x_2 não aparece no modelo verdadeiro, logo o estimador dos MQ para β_1 é não enviesado apesar de x_2 está positivamente correlacionada com a variável x_1.

(b) A variável x_2 está negativamente correlacionada com a variável x_1 e tem um efeito positivo sobre y.

Se a introdução da variável x_2 tem um efeito positivo sobre y_i, é porque $\beta_2 > 0$ o que significa que a variável x_2 está omissa no modelo estimado pelo que sendo x_2 negativamente correlacionada com a variável x_1, $\tilde{\delta}_1 < 0$ logo o estimador dos MQ para β_1 é enviesado e o sinal do enviesamento será negativo.

TABLE 3.2 Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias