Econometria - Lista 3 - Regressão Linear Múltipla

Professores: Hedibert Lopes, Priscila Ribeiro e Sérgio Martins
Monitores: Gustavo Amarante e João Marcos Nusdeo

QUESTÃO 1. Você trabalha na consultoria "Fazemos Qualquer Negócio" que acabou de ser contratada para avaliar como as vendas de um setor de empresas de exportação reagem ao aumento da renda per capita do país em que estão instaladas. Com base no que foi exposto, você tem interesse em estimar o seguinte modelo:

$$v_i = \hat{\beta}_1 r_i + \hat{\varepsilon}_i$$

em que:

1) v_i é o valor das exportações da empresa do i-ésimo país em 2013, em milhares de reais;
2) r_i é a renda per capita do i-ésimo país em 2013, em milhares de reais.

Aplicando o método de mínimos quadrados ordinários, você encontrou o seguinte resultado:

$$\underset{(1,2)}{v_i = 2r_i} + \hat{\varepsilon}_i \qquad n = 31$$

em que entre parênteses está o erro-padrão associado a $\hat{\beta}_1$. Os resultados são apresentados ao seu chefe durante uma reunião.

a) Ele te pergunta, "r_i tem impacto positivo sobre v_i?". O que você responde? Utilizando um nível de significância de 5%, fundamente a sua resposta com base numa análise inferencial adequada.

b) Seu chefe, que já está exaltado com alguns resultados apresentados durante a reunião, começa a reclamar e te pergunta "a estimativa $\hat{\beta}_1$ sofreria alguma alteração se v_i e r_i fossem medidos em reais?". O que você responderia? Mostre matematicamente o fundamento da sua resposta.

c) Seu chefe desconfia que o aumento de um mil reais na renda do resto do mundo provoca um aumento médio nas vendas de 3 mil reais. Conduza um teste de hipótese para verificar se tal conjectura é válida. Adote um nível de confiança de 95%.

QUESTÃO 2. Um pesquisador está interessado em verificar o diferencial do salário devido ao gênero utilizando o seguinte modelo ajustado:

$$\ln(salario_i) = \hat{\beta}_1 + \hat{\beta}_2 D_{Mi} + \hat{\beta}_3 educ_i + \hat{\beta}_4 D_{Mi} educ_i + \hat{\beta}_5 \exp er_i + \hat{\varepsilon}_i \qquad (1)$$

em que $salario_i$ é o salário, em milhares de reais, do *i-ésimo* indivíduo da amostra; D_{Mi} é uma variável *dummy* que assume o valor 1 se o *i-ésimo* indivíduo da amostra for do gênero feminino e 0 se for do gênero masculino; $educ_i$ é o número de anos estudados pelo *i-ésimo* indivíduo da amostra; $\exp er_i$ é o número de anos que o *i-ésimo* indivíduo da amostra trabalha (experiência); $\hat{\varepsilon}_i$ é o resíduo do *i-ésimo* indivíduo da amostra.

A Tabela 1, a seguir, apresenta o resultado da estimação dos parâmetros do modelo proposto em (1), usando o método dos mínimos quadrados ordinários (MQO).

Tabela 1 – Estimação, por MQO, de (1).

Variável Dependente: ln(salário)		
Variáveis Explicativas	Estimativa	Erro-padrão
Constante	0,462	0,127
D_M	-0,296	0,179
Educ	0,093	0,009
D_M*Educ*	-0,004	0,014
Exper	0,009	0,001
Observações		60
Soma de Quadrado de Resíduos		123,22
R^2		0,65

a) A tabela 1apresenta a estimativa de cada parâmetro, seu respectivo erro-padrão, o R^2 e a Soma de Quadrado dos Resíduos. Interprete a estimativa $\hat{\beta}_5$.

b) Qual é o retorno esperado no salário para um ano adicional de educação para um homem? E para uma mulher?

c) Dê uma interpretação prática para as hipóteses formuladas a seguir:

$$H_0 : \beta_4 = 0$$

$$H_A : \beta_4 \neq 0 .$$

Ainda, adotando um nível de 5% de significância, quais seriam as suas conclusões a respeito deste teste de hipóteses? Não se esqueça de apresentar a estatística do teste, a distribuição de probabilidades da estatística do teste e a região crítica do mesmo. Justifique a sua resposta. Resposta sem justificativa será ignorada.

QUESTÃO 3. (PI 2007/02) Considere o seguinte modelo, proposto com o objetivo de explicar os salários:

$$\ln(salário_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper^2{}_i + \beta_4 \ln(QI_i) + \varepsilon_i$$

em que:
$salário_i$ é o salário do indivíduo i (em reais);
$educ_i$ é a educação formal do indivíduo i (número de anos completos de estudo);
$exper_i$ é a experiência do indivíduo i (número de anos de trabalho);
$exper^2{}_i$ é a experiência ao quadrado
QI_i é o QI do indivíduo i (medido em uma escala de 0 a 100)
ε_i é o termo de erro associado ao indivíduo i

Responda:

a) Qual a interpretação do coeficiente β_1?
b) Qual a interpretação do coeficiente β_4?
c) Qual seria a expectativa dos sinais dos coeficientes β_2 e β_3? Justifique sua resposta.
d) Qual é a derivada primeira de $\ln(salário_i)$ com respeito à $exper_i$? Comente.
e) Qual é o significado do valor da primeira derivada, gerada no item anterior, ser igual a zero?
f) Qual é a derivada segunda de $\ln(salário_i)$ com respeito à $exper_i$? Comente.
g) Qual é o significado do valor da segunda derivada, gerada no item anterior, ser igual a zero?

QUESTÃO 4. (PI 2009/01) Uma companhia de seguros deseja estudar o custo da seguradora com danos materiais a terceiros em relação à idade e ao gênero de seus clientes. A companhia coletou os dados de uma amostra aleatória de 90 clientes entre 18 e 90 anos dos gêneros masculino e feminino a fim de estimar o seguinte modelo de regressão linear múltipla.

$$\hat{y}_i = 4{,}55 + 0{,}22x_2 + 40{,}2x_4 - 0{,}67x_2x_4 - 10{,}8x_3x_4 + 0{,}18x_2x_3x_4$$

Em que:
y_i : custo da seguradora com danos materiais a terceiros para seus clientes, em milhares de reais;
x_2 : idade do principal condutor do veículo, em anos;
x_3 : dummy que assume o valor 1 caso o principal condutor do veículo seja do gênero feminino e 0 caso contrário;
x_4 : dummy que assume o valor 1 caso o condutor do veículo tenha até 60 anos de idade e 0 caso contrário.

a) Qual o custo médio da seguradora com danos materiais a terceiros para cliente cujo principal condutor seja do sexo feminino com 25 anos de idade?

b) Esboce graficamente o modelo estimado pela seguradora para cada um dos segmentos avaliados. Interprete as inclinações de acordo com as suas expectativas.

c) O diretor da seguradora afirma que, para os clientes que têm até 60 anos, quando se aumenta um ano na idade do principal condutor do veículo, há um decréscimo maior no custo esperado dos clientes do gênero masculino em comparação aos clientes do gênero feminino. Teste a veracidade desta afirmação com 99% de confiança.

QUESTÃO 5. Um grupo de alunos de econometria do Insper escolheu como tema para o trabalho do semestre "Fatores que influenciam o diferencial de salários entre homens e mulheres no Brasil". Para conduzir tal estudo coletaram os dados da PNAD (Pesquisa Nacional por Amostragem de Domicílios), uma base de dados do IBGE com mais de 200 variáveis e mais de 300 mil observações. Um dos integrantes do grupo disse: "Já que temos tantas variáveis para explicar o diferencial de salários, quero usar no trabalho o melhor modelo, com o maior R^2 possível e por isso devemos escolher para o nosso modelo as variáveis que maximizem nosso coeficiente de determinação!".

a) Comente o que o aluno disse. Você acha que usar apenas o R^2 como medida para escolha do modelo é opção razoável?
b) Explique intuitivamente porque o R^2 nunca diminui quando adicionamos uma variável no modelo.

Tendo aprendido o que foi dito no item b), o aluno torna a dizer: "Ok! Já que não podemos colocar todas as variáveis no modelo, podemos usar o R^2 ajustado como medida de escolha do modelo, já que ele é penalizado pelo número de variáveis".

c) Mostre que maximizar o R^2 ajustado é equivalente a minimizar o erro-padrão da regressão.
d) Mostre que se utilizarmos o critério da maximização do R^2 ajustado, a variável x_j será escolhida para o modelo se, e somente se, a estatística-F do teste $H_0: \beta_j = 0$ for maior que 1. Para amostras grandes (como a PNAD), qual o nível de significância deste teste? Você diria que escolher variáveis que maximizem o R^2 ajustado é um bom critério de seleção de variáveis do modelo?

QUESTÃO 6. Mariana é uma empreendedora de sucesso, dona de uma rede de docerias famosa por seus pães de mel. Mariana está interessada em entender melhor sobre a capacidade de produção de suas fábricas e por isso pediu que as estagiárias da empresa levantassem os seguintes dados para uma amostra de 26 fábricas:

- K_i: Estoque de capital físico da i-ésima fábrica em 2013.
- L_i: Total de funcionários na i-ésima fábrica em 2013.
- Q_i: Quantidade de pães de mel produzidos pela i-ésima fábrica em 2013.

Mariana contratou a consultoria *Nusdeo e Amarante Consulting Corporation* para estimar uma função de produção para suas fábricas. Os consultores disseram que, com base na teoria microeconômica, uma forma funcional razoável para a função de produção seria a função Cobb-Douglas:

$$Q_i = e^{\beta_o} K_i^{\beta_1} L_i^{\beta_2} e^{\varepsilon_i}$$

em que ε_i é um termo estocástico independente e identicamente distribuído com média zero e variância σ^2.

a) Os consultores querem utilizar o método de mínimos quadrados ordinários para estimar os parâmetros da função de produção. Porque isso não é possível para a forma funcional sugerida? Como eles podem contornar o problema?

Aplicando a transformação logaritmo na função de produção Cobb-Douglas e estimando o modelo resultante, os consultores chegaram no resultado:

$$\widehat{\log(Q_i)} = \underset{(0,415)}{0,701} + \underset{(0,110)}{0,242}\log(K_i) + \underset{(0,091)}{0,756}\log(L_i)$$

$$SSR = 1,825544 \quad R^2 = 0,956888 \quad n = 26$$

b) Faça um teste individual para a significância de cada parâmetro adotando uma significância de 5%. Interprete os parâmetros significantes.

c) Como você faria para testar a hipótese de que as fábricas de Mariana possuem retornos constantes de escala por meio de um teste-F? Neste caso, existe diferença entre montar a estatística-F com base no R^2 e com base na SSR? Qual delas é a correta para este caso? Justifique. Caso julgue necessário, utilize as informações abaixo para lhe ajudar no teste.

$$\widehat{\log\left(\frac{Q_i}{K_i}\right)} = \underset{(0,1319)}{0,6864} + \underset{(0,0887)}{0,7558}\log\left(\frac{L_i}{K_i}\right)$$

$$SSR = 1,825652 \quad R^2 = 0,751397 \quad n = 26$$

QUESTÃO 7. Houve controvérsia na eleição americana do ano 2000 (Gore VS Bush) quando o Partido Democrata alegou problemas na contagem de votos em dois municípios do estado da flórida (Miami-Dade e Palm Beach), estado que era decisivo para o resultado da eleição. Para tentar explicar essa controvérsia um econometrista tinha em mente o modelo:

$$VotosGore_i = \beta_0 + \beta_1 Dproblematico_i + \varepsilon_i$$

Em que $VotosGore_i$ é a porcentagem de votos recebidos por Al Gore no i-ésimo município (em pontos percentuais) e $Dproblematico_i$ é uma variável dummy que assume valor 1 para os dois municípios em que houveram problemas na contagem dos votos e valor 0 para os demais. O econometrista tem uma amostra de 3141 municípios dos estados unidos e com ela estimou os parâmetros do modelo acima.

	Estimativa	Erro-Padrão
$\hat{\beta}_0$	50.3	16.4
$\hat{\beta}_1$	14.1	5.2
n	3141	
R^2	0.0012	

a) Porque o R^2 deste modelo estimado é tão baixo?
b) Os coeficientes estimados são estatisticamente significantes? Dê uma interpretação para eles.
c) O econometrista fez a seguinte análise:
"*O coeficiente associado à variável dummy é positivo e significante. Então é errado dizer que um possível problema na contagem dos votos tenha prejudicado Al Gore, o candidato do Partido Democrata, já que ele teve uma porcentagem de votos maior nos municípios em que acusa haver problema de contagem de votos*".
Você concorda com a análise do econometrista? Você acha que algum dos partidos americanos gostou desta análise?

Um segundo econometrista, acreditava que o modelo acima tinha variáveis omitidas que poderiam estar viesando os estimadores. Uma das variáveis que ele gostaria de colocar em seu modelo é a porcentagem da população do município que apoia o Partido Democrata. O econometrista não tinha essa variável em sua base, mas ele se lembrou de que Al Gore era vice-presidente da gestão de Bill e concluiu que a porcentagem de votos recebidos por Bill Clinton nas eleições de 1996 em cada município na última eleição ($VotosClinton1996_i$) seria uma boa medida da porcentagem de pessoas que apoiam o Partido Democrata. Então ele estimou o seguinte modelo:

$$VotosGore_i = \beta_0 + \beta_1 Dproblematico_i + \beta_2 VotosClinton1996_i + \varepsilon_i$$

	Estimativa	Erro-Padrão
$\hat{\beta}_0$	40.1	16.4
$\hat{\beta}_1$	-2.9	0.98
$\hat{\beta}_2$	0.93	0.27
n	3141	
R^2	0.87	

d) Interprete novamente as estimativas. Os resultados acima corroboram as alegações do Partido Democrata?