Econometria 1

Pedro Henrique Rocha Mendes*

Lista 2

(Texto para as questões 1 e 2) A equação abaixo explica o desempenho no ENEM 2013, "nota_ENEM", de um aluno que concluiu ensino médio no ano de 2013 em função do valor da mensalidade, "mensalidade", da escola em que o aluno cursou o ensino médio. o valor da mensalidade é uma proxy para a qualidade da escola. Apenas alunos que estudaram em escolas particulares são considerados.

$$Nota\_ENEM_i = \beta_0 + \beta_1 \; mensalidade_i + u_i$$

1) Selecione um fator (variável aleatória) que você acredita explicar o desempenho do ENEM, mas que, estando omitido da equação, é captado pelo termo de perturbação, u_i. Explique com detalhe porque este fator influencia a variável dependente.
2) Verifique se a propriedade de ausência de viés do estimador do coeficiente de uma regressão linear simples se mantém caso a hipótese de expectativa condicional zero, i.e., de que $E(u_i|X) = 0, i = 1, \ldots, N$, for substituída pela hipótese de que $E(u_i|X) = \alpha_0 + \alpha_1 X_i, i = 1, \ldots, N, \alpha_0 \neq 0, \alpha_1 \neq 0$. Em sua resposta, apresente os passos lógicos necessários para estabelecer o resultado de ausência de viés sob as hipóteses convencionais, apenas tomando o cuidado de utilizar a hipótese aqui requerida. Para fins de fixação do conteúdo que seguirá nas próximas aulas, acrescenta-se que a hipótese $E(u_i|X) = 0$ é também referida como "exogeneidade".
3) Um analista trainee do setor bancário estimou, como parte de uma exploração descritiva dos dados da PNAD anual 2015, a regressão simples reportada na tabela abaixo. Teve-se como objetivo investigar a relação entre nível educacional e remuneração horária. Essas duas variáveis estavam disponíveis nos dados na forma de características individuais de pessoas empregadas em 2015.

*RA: 11201811516

Parâmetro	Estimativa pontual	Unidade de medida
Intercepto	3,67	R$/hora
Coeficiente	0,69	R$/hora/ano de estudo

a. O que significa exatamente o valor numérico da estimativa pontual do intercepto? Informe a leitura correta de tal número, tendo em vista o objetivo da análise.

b. O que significa exatamente o valor numérico da estimativa pontual do coeficiente? Informe a leitura correta de tal resultado, tendo em vista o objetivo da análise.

c. O analista também estimou a especificação alternativa na tabela abaixo, em que as duas variáveis foram incorporadas à regressão simples em forma logarítmica. Qual é a interpretação correta da estimativa pontual do valor numérico coeficiente neste caso? Não deixe de considerar o objetivo da análise e tenha atenção à unidade de medida.

Parâmetro	Estimativa pontual	Unidade de medida
Intercepto	1,73	Log (R$/hora) ou porcentagem
Coeficiente	0,25	Log (R$/hora) / Log (anos de estudo) ou porcentagem/porcentagem

4) Para esta questão você utilizará a planilha "dados_lista_2.xlsx". Nela se encontram dados para duas variáveis, $Y \equiv$ medida padronizada para o superávit de altura-para-idade para crianças com menos do que cinco anos (`z_nutri`) e $X \equiv$ renda domiciliar (`renda_percapita`) – a variável `id_dom` é o código que identifica univocamente os domicílios. Utilizando o Excel ou o software gratuito R (este último é objeto dos vídeos lab.1/lab.2/lab.3), realize as tarefas a seguir:

a. Calcule a estimativa pontual para o coeficiente da regressão simples $Y = \beta_0 + \beta_1 X + u$. Para isso, aplique a fórmula do estimador em questão aos dados, sendo ela:

$$\hat{\beta}_1 = \frac{N^{-1} \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{N^{-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Uma versão alternativa da fórmula acima é a seguinte:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \tilde{y}_i \tilde{x}_i}{\sum_{i=1}^{N} (\tilde{x}_i)^2}$$

Em que $\tilde{y}_i = y_i - \bar{y}$ e $\tilde{x}_i = x_i - \bar{x}$.

```r
# install.packages("tidyverse")
library("tidyverse")

df <- readxl::read_xlsx(path = "listas/lista_2/dados_lista_2.xlsx") |>
  dplyr::rename(renda_pc = `renda_percapita (X)`,
                z_nutri = `z_nutri (Y)`)
beta1 <- df |>
  dplyr::mutate(d_xy = (renda_pc - mean(renda_pc))*(z_nutri - mean(z_nutri)),
                dxquad = (renda_pc - mean(renda_pc))^2) |>
  dplyr::summarise(beta1 = sum(d_xy)/sum(dxquad)) |>
  round(6) |>
  dplyr::pull(beta1)
```

$\beta_1 = 2.35 \times 10^{-4}$

b. Julgue, de maneira justificada, se a estimativa pontual obtida no item anterior faz sentido e interprete-a. Considere que Y é medida em desvios padrão e X em R$.[1]

c. Calcule a estimativa pontual para o intercepto da regressão simples $Y = \beta_0 + \beta_1 X + u$. Para isso, aplique a fórmula do estimador em questão aos dados, sendo ela: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

```r
beta0 <- df |>
  dplyr::summarise(beta0 = mean(z_nutri)-beta1*mean(renda_pc)) |>
  round(2) |>
  dplyr::pull(beta0)
```

$Y = \beta_0 + \beta_1 X + u =$ -0.55 + 2.35 $\times 10^{-4}$ X.

d. Julgue, de maneira justificada, se a estimativa pontual obtida no item anterior faz sentido e interprete-a.

5) Com base na planilha `dados_lista_2.xlsx` e considerando a mesma regressão da questão anterior:

a. Calcule a fórmula para o coeficiente de determinação, conforme segue:

$$R^2 = 1 - \frac{SQR}{SQT} = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{SQT} = 1 - \frac{\sum_{i=1}^{N} (\hat{u}_i)^2}{\sum_{i=1}^{N} (\tilde{y}_i)^2}$$

Em que $\hat{u}_i$ é o resíduo.

```r
r2 <- df |>
  summarize(r2 = 1-sum((z_nutri - (beta0+beta1*renda_pc))^2)/
              sum((z_nutri - mean(z_nutri))^2)) |>
```

[1]Nota: a variável dependente é uma medida de nutrição e não de desnutrição, pois se trata da seguinte diferença: altura observada - altura de referência para a idade.

```
round(2) |>
pull(r2)
```

$R^2 = 0.01$

b. Interprete o valor numérico do coeficiente de determinação.

6) Um aluno da disciplina de econometria I armazenou os dados utilizados para estimar uma regressão simples em uma pasta temporária no disco duro de seu computador.

a. O aluno procurou calcular de maneira semi-manual a soma dos quadrados dos resíduos. Para isso ele fez as nove operações abaixo e, antes de realizar a última operação, o Windows automaticamente reiniciou o computador para instalar uma atualização. Ao reiniciar o software estatístico, os dados não estavam mais disponíveis. Complete o cálculo para o aluno, informando (i) o resíduo da décima observação, (ii) o quadrado do resíduo referente à décima observação e (iii) a soma dos quadrados dos resíduos, utilizando uma das propriedades algébricas da regressão linear simples.

Observação	Nível educacional (X)	Resíduo (A)	Resíduo ao quadrado (B = A^2)
1	8	-2,027923995	4,112475731
2	11	-1,309013586	1,713516569
3	11	0,996611414	0,99323431
4	15	4,242337959	17,99743136
5	0	-0,403581087	0,162877694
6	15	-0,994722041	0,989471938
7	11	-2,896600586	8,390294956
8	0	-0,403581087	0,162877694
9	11	-1,862363586	3,468398127
10	5	Não disponível	Não disponível

b. Agora ajude o estudante a calcular a covariância amostral entre a variável dependente e o resíduo, dada por $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(\hat{u}_i - \bar{u})$. Utilize, para isso, uma segunda propriedade algébrica de regressão linear simples. Confirme o resultado obtido algebricamente fazendo o cálculo numérico com base na tabela acima.

7) Derive o estimador de mínimos quadrados ordinários do seguinte modelo sem intercepto $Y_i = \beta_1 X_i + u_i$ (1). Agora verifique que, se o modelo populacional possui um intercepto β_0 diferente de zero, assumindo forma $Y_i = \beta_0 + \beta_1 X_i + u_i$, então o estimador de β_1 da equação (1) é viesado.