Lista Prática 3

Exercício 1. *Neste exercício, usaremos as funções* `runif()` *e* `rnorm()` *para gerar números aleatórios com distribuições uniforme e normal, respectivamente.*

a) *Gere o vetor* `x` *com 10.000 números aleatórios a partir de uma distribuição uniforme no intervalo [0, 10]. Qual é a média e o desvio padrão de* x*?*

b) *Gere o vetor* `z` *com 10.000 números aleatórios usando* $z = 2x + \tilde{u}, \quad \tilde{u} \sim N(0, 4^2)$*. Qual é a correlação entre* x *e* z*?*

c) *Gere o vetor* $\tilde{\varepsilon}$ *(*`e_til`*) com 10.000 números aleatórios a partir de uma distribuição* $N(0, 6^2)$*. Qual é a correlação entre* $\tilde{\varepsilon}$ *e cada uma das demais variáveis* x *e* z*? Além disso, verifique a correlação entre* x *e a soma* $3z + \tilde{\varepsilon}$*.*

d) *Gere o vetor* `y`*, considerando o seguinte modelo real:*

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 z + \tilde{\varepsilon}, \tag{1.1}$$

em que $\tilde{\beta}_0 = 10$*,* $\tilde{\beta}_1 = 2$ *e* $\tilde{\beta}_2 = 3$*. Agora, estime por MQO o seguinte modelo empírico:*

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{1.2}$$

A estimação conseguiu recuperar $\hat{\beta}_0 \approx \tilde{\beta}_0$ *e* $\hat{\beta}_1 \approx \tilde{\beta}_1$*? Explique.*

e) *Obtenha os resíduos de MQO,* $\hat{\varepsilon}$*, e verifique se valem os seguintes momentos amostrais (sujeitas a algum erro de arredondamento):*

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_i = 0 \qquad e \qquad \frac{1}{N}\sum_{i=1}^{N} x_i \hat{\varepsilon}_i = 0$$

f) *O que os resultados do item (e) dizem sobre as condições de momento populacionais* $E(\varepsilon) = 0$ *e* $E(x\varepsilon) = 0$*?*

g) *Denote como Caso I o modelo real visto até agora, em que* $\tilde{\beta}_2 = 3$ *e* $z = 2x + \tilde{u}$*, em que* $\tilde{u} \sim N(0, 4^2)$*. Gere novamente observações* z *e* y*, e estime por MQO o modelo empírico (1.2) para cada um dos seguintes modelos reais:*

- *Caso II:* $\tilde{\beta}_2 = -3$ *e* $z = 2x + \tilde{u}, \quad \tilde{u} \sim N(0, 4^2)$
- *Caso III:* $\tilde{\beta}_2 = 3$ *e* $z = -2x + \tilde{u}, \quad \tilde{u} \sim N(0, 4^2)$
- *Caso IV:* $\tilde{\beta}_2 = -3$ *e* $z = -2x + \tilde{u}, \quad \tilde{u} \sim N(0, 4^2)$

Considerando os sinais do parâmetro da variável omitida z, $\tilde{\beta}_2$, e da sua covariância com x, $cov(x,z)$, em quais casos a estimativa do parâmetro de x é sobre-estimada ($\hat{\beta}_1 > \tilde{\beta}_1$)? E em quais é sub-estimada ($\hat{\beta}_1 < \tilde{\beta}_1$)?

Exercício 2. *Neste exercício, usaremos a base de dados de Papke (1995), que possui informações sobre a participação e contribuição em planos previdência privada de empresas nos EUA, chamada de 401k:*

```
data(k401k, package="wooldridge")
```

- `prate`*: é o percentual de trabalhadores contribuindo ativamente à previdência privada.*
- `mrate`*: é a taxa de "generosidade" da empresa, isto é, a razão de quanto a empresa contribui para a previdência privada de seu funcionário.*
- `totemp`*: é número total de funcionários.*

Queremos saber a relação entre a taxa de participação de funcionários (`prate`) e a taxa de generosidade da empresa (`mrate`).

a) Estime analiticamente (sem usar a função `lm()`) o modelo:

$$prate = \beta_0 + \beta_1 mrate + \beta_2 totemp + \varepsilon$$

b) Usando operações matriciais, adapte as funções objetivo da seção de Otimização para o caso multivariado[1]. Depois, usando `optimx::opm()`, obtenha as estimativas que otimizam essas funções objetivo por

(i) Minimização da soma do quadrado dos resíduos

(ii) Método Generalizado dos Momentos (GMM), cujos momentos amostrais são

$$\begin{bmatrix} \sum_{i=1}^{N} \hat{\varepsilon}_i \\ \sum_{i=1}^{N} mrate_i.\hat{\varepsilon}_i \\ \sum_{i=1}^{N} totemp_i.\hat{\varepsilon}_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

(iii) Máxima Verossimilhança (ML)

Exercício 3. *Neste exercício, usaremos a base de dados `wage1` do pacote `wooldridge` e pode ser carregada no R usando o comando:*

```
data(wage1, package="wooldridge")
```

Queremos saber a relação entre anos de estudo (`educ`) e o logaritmo da renda das pessoas (`lwage`), considerando seus sexos/gêneros.

[1] Lembre-se de transformar os objetos em vetores/matrizes usando `matrix()` ou `as.matrix()` antes de fazer as operações matriciais.

a) *Gere duas bases a partir de* `wage1`*: uma apenas com mulheres (*`wage_female`*) e outra apenas com homens (*`wage_male`*), e estime os seguintes modelos (sem interceptos):*

$$coninc = \beta_F.educ + \varepsilon \quad \text{(base com mulheres)}$$
$$coninc = \beta_M.educ + \varepsilon \quad \text{(base com homens)}$$

Quais são as estimativas para β_F e β_M?

b) *Usando a base completa (*`wage1`*), plote um gráfico de dispersão (scatterplot) entre anos de estudo × renda, colorindo os pontos de acordo com o sexo da pessoa. Também, adicione as retas das regressões feitas no item (a) com cores distintas.*

c) *Na base completa (*`wage1`*), regrida um único modelo em que, com as estimativas obtidas, possamos calcular $\hat{\beta}_F$ e $\hat{\beta}_M$ encontrados no item (a). A diferença entre $\hat{\beta}_F$ e $\hat{\beta}_M$ é estatisticamente significante?*

Exercício 4. *Analisando* ***apenas trabalhadores em tempo integral e parcial ("Working Fulltime" e "Working Parttime")****, verifique a possível discriminação na renda em relação aos povos hispânicos povos hispânicos (mexicanos, porto riquenhos, cubanos, etc.). Para isto use a base de dados General Social Survey (GSS)[2], que pode ser carregada no R usando o comando:*

```
load(url("https://fhnishida.netlify.app/project/rec5004/gss.Rdata"))
```

a) *Regrida a renda (*`coninc`*) em relação às dummies dos povos hispânicos e liste os que possuem diferença significativa de renda em relação aos não-hispânicos. Utilize como variáveis de controle: o sexo, a idade, a idade^2, a raça/cor de pele, os anos de estudo (*`educ`*) e o status de trabalho (*`wrkstat`*).*

b) *Note que a regressão anterior "jogou fora" quase 70% das observações por causa de valores ausentes (NA's) das variáveis utilizadas, sobretudo de* `hispanic`*. Caso essas informações estejam faltando aleatoriamente (missing at random), pode ser razoável considerar o resultado do item (a) como representativo de toda população trabalhadora nos EUA. Para verificar isso, siga os passos abaixo:*

   i. *Para facilitar, selecione apenas as colunas/variáveis usadas na regressão do item (a)*
   ii. *Crie uma variável dummy* `missing` *que é igual a 1 se houver pelo menos um* `NA` *na linha/observação, e igual a 0 caso contrário.*
   iii. *Crie variáveis dummies para as variáveis categóricas* `sex` *e* `race`*.*
   iv. *Usando regressões de diferença de médias, verifique a significância das diferenças entre as observações retiradas da regressão do item (a) (*`missing==1 / NA`*), e as que foram mantidas (*`missing==0 / nonNA`*)*

[2]Disponibilizado por Bryan Wheeler (2014)

v. Crie e analise a seguinte tabela com os resultados:

	`nonNA`	`NA`	diferença	p-valor
coninc	-	-	-	-
sex_female	-	-	-	-
age	-	-	-	-
race_white	-	-	-	-
race_black	-	-	-	-
educ	-	-	-	-

Exercício 5. *O seguinte modelo pode ser usado para estudar se gastos na campanha afetam os resultados eleitorais:*

$$voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + \varepsilon$$

em que:

- `voteA` *é o percentual de votos recebidos pelo Candidato A*
- `expendA` *e* `expendB` *são gastos de campanha pelos Candidatos A e B, respectivamente*
- `prtystrA` *é uma medida de força do partido do Candidato A*

a) Declare a hipótese nula: o impacto nos votos do Candidato A por um aumento em 1% nos seus gastos é anulado por um aumento em 1% de gastos de B.

b) Usando a base de dados `vote1` *do pacote* `wooldridge`*, estime o modelo acima.*

c) Faça o teste hipótese do item (a). Dica: Você pode obter as variâncias e covariâncias das estimativas usando `vcov()` *no objeto de regressão gerado por* `lm()`*.*

Exercício 6. *Use a base de dados sobre salários na liga americana de beisebol* `mlb1`*, do pacote* `wooldridge`*, para este exercício e considere o modelo:*

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \varepsilon$$

em que:

- `salary` *é o salário do jogador*
- `years` *é a quantidade de anos como jogador profissional*
- `gamesyr` *é a média de jogos por ano do jogador*
- `bavg` *é o percentual de rebatida*
- `hrunsyr` *é a média de home runs por ano*

a) Adicione `runsyr` *(corridas por ano),* `fldperc` *(percentual de defesa), e* `sbasesyr` *(bases roubadas por ano) ao modelo. Quais destes fatores são individualmente significantes?*

b) A partir do modelo modelo do item (a), teste a significância conjunta de `bavg`*,* `fldperc`*, e* `sbasesyr`*. Faça "na mão" os dois testes possíveis para este caso e analise-os.*