Lista Prática 3

Exercício 1 (`swirl`). *Usando o pacote* `swirl`*, instale o curso "Exploratory Data Analysis" usando a seguinte função* `swirl::install_course("Exploratory Data Analysis")`[1] *e faça as seguintes lições:*

a) 5: Base Plotting System

b) 9: GGPlot2 Part2[2]

Exercício 2. *Neste exercício, usaremos a base de dados do General Social Survey (GSS), que coleta dados demográficos nos EUA e pode ser carregada no R usando o comando*[3]*:*

```
load(url("http://bit.ly/dasi_gss_data"))
```

Queremos saber a relação entre anos de estudo (`educ`*) e a renda das pessoas (*`coninc`*) que* ***trabalham em tempo integral*** *para homens e mulheres.*

a) Gere duas bases a partir da GSS: uma apenas com mulheres (`gss_mulher`*) e outra apenas com homens (*`gss_homem`*), e estime os seguintes modelos:*

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

Note que o modelo não possui uma constante (insira um regressor `0` *para não incluí-la na regressão via* `lm()`*). Quais são as estimativas para β_M e β_H?*

b) Usando a base completa (`gss`*), 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 (`gss`*), regrida um único modelo em que é possível calcular os mesmos β_M e β_H encontrados no item (a). A diferença entre β_M e β_H é estatisticamente significante?*

Resposta:

a) $\beta_M = 3576.56$ e $\beta_H = 4229.01$

[1] Caso tenha problemas, instalar manualmente: `http://swirlstats.com/scn/eda.html`

[2] O 1º exercício irá pedir para fazer um gráfico usando `qplot()`, você pode dar `skip()` para avançar para a parte com gráficos usando a função `ggplot()`

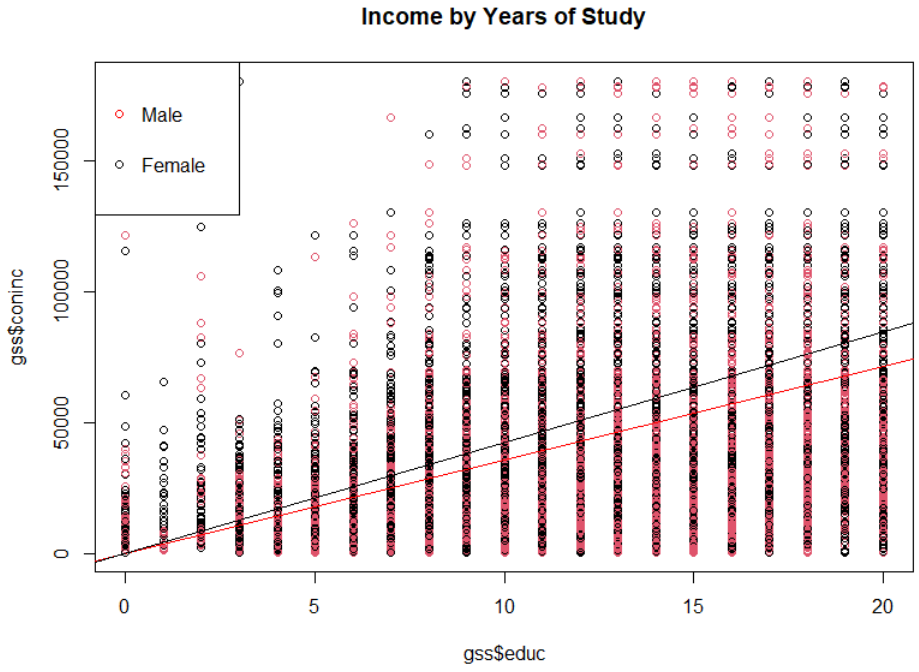
[3] Disponibilizado por Bryan Wheeler (2014)

```r
library(dplyr)

# Gerando 2 bases com trab. em tempo integral de mulheres e de homens
gss_homem = gss %>% filter(sex=="Male", wrkstat=="Working Fulltime")
gss_mulher = gss %>% filter(sex=="Female", wrkstat=="Working Fulltime")

# Estimando os modelos sem constante
fit_homem = lm(coninc ~ 0 + educ, gss_homem)
fit_mulher = lm(coninc ~ 0 + educ, gss_mulher)

# Mostrando os betas estimados
fit_homem$coef
fit_mulher$coef
```

```
4229.012
3576.562
```

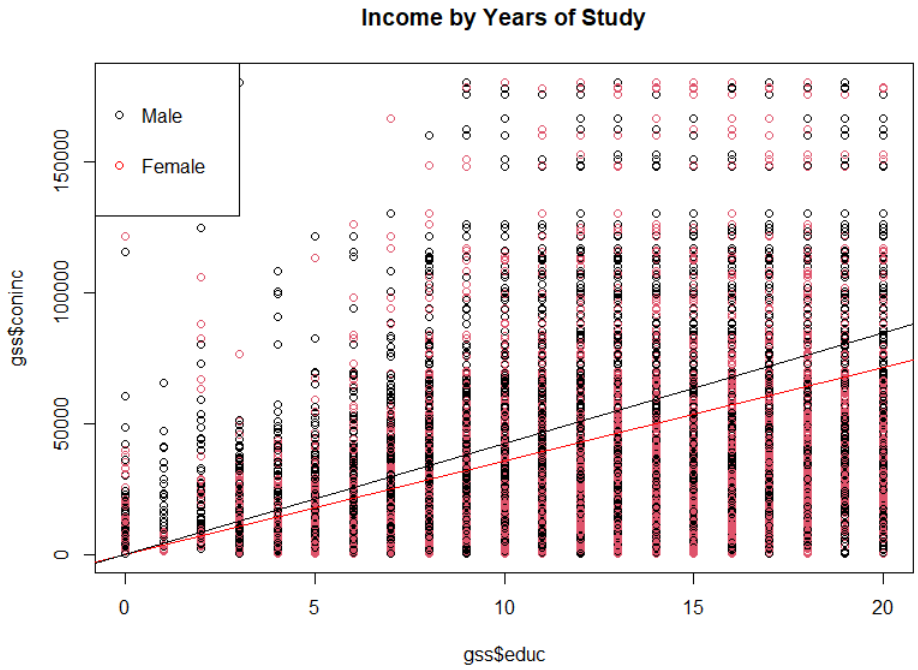
b)

```r
# Plotando anos de educação X renda
plot(gss$educ, gss$coninc, col=gss$sex, main="Income by Years of Study")
# Plotando as retas das regressões estimadas
abline(fit_mulher, col="red")
abline(fit_homem, col="black")
legend("topleft", pch=1, col=c("black", "red"), legend=c("Male", "Female"))
```



ou também por ggplot2:

```r
library(ggplot2)
g = ggplot(gss, aes(educ, coninc))
g + geom_point(aes(color=sex), size=.9, alpha=0.3) +
    geom_abline(intercept=c(0,0),
                slope=c(fit_homem$coef, fit_mulher$coef),
                color=c("darksalmon", "cyan3"))
```



c) Vamos estimar o modelo que, além da covariada de anos de estudo, vamos incluir a interação entre anos de estudo e a *dummy* `mulher`:

$$\text{coninc} = \beta_1.\text{educ} + \beta_2.\text{educ.mulher} + \varepsilon$$

```
# Na base completa, filtro trab tempo integral e criando Dummy mulher
gss = gss %>% filter(wrkstat=="Working Fulltime") %>%
    mutate(mulher = ifelse(sex=="Female", 1, 0))

# Estimando o modelo com interação entre anos de educ. e dummy mulher
lm(coninc ~ 0 + educ + educ:mulher, gss) %>% summary()
```

```
             Estimate Std. Error t value Pr(>|t|)
educ          4229.01      19.75  214.08   <2e-16 ***
educ:mulher   -652.45      29.84  -21.86   <2e-16 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1
```

Note que $\beta_H = \beta_1$ e $\beta_M = \beta_1 + \beta_2$. Além disso, a diferença entre as estimativas, $\beta_M - \beta_H = \beta_2$, é significativa a $0,1\%$.

□

Exercício 3. *Neste exercício, utilizaremos a base de dados de Card e Krueger (1994), a qual foi usada para analisar o efeito do aumento do salário mínimo estadual na quantidade de empregos. Uma versão modificada dela pode ser carregada no R usando:*

```
fastfood = read.csv("https://fhnishida.github.io/fearp/eco1/CardKrueger_1994_
    modificado.csv")
```

- *id: identificador da loja de fast food*
- *state: estado dos EUA (New Jersey ou Pennsylvania)*
- *emptot_feb92: total de empregados em fev/92*
- *emptot_nov92: total de empregados em nov/92*
- *chain: rede de fast food*
- *pct_fte: percentual de trabalhadores em tempo integral*
- *wage_st: salário por hora*
- *hrsopen: horas de funcionamento diário*

Idealmente (para avaliação), haveria um experimento em que estados seriam aleatoriamente selecionados para aumentarem o salário mínimo e, logo, o efeito da intervenção poderia ser estimado por meio da diferença das médias de variação de emprego entre os estados que elevaram seus salários mínimos (tratados) e os que não alteraram (não-tratados/controle). No entanto, em políticas públicas, a aleatorização pode ser custosa (financeira e politicamente) ou operacionalmente inviável. Neste caso, para identificar esse efeito, Card e Krueger (1994) assumiram o estado vizinho de Pennsylvania como o controle, por não ter elevado o seu salário mínimo e ser considerado, ao menos no mercado de trabalho de fast food, similar a New Jersey. Sob essa premissa, os estados manteriam o mesmo comportamento se ambos não alterassem seu nível de salário mínimo e, portanto, o efeito do tratamento pode ser calculado pela diferença entre as variações de empregados dos dois estados. Por exemplo, se a variação do nº de empregados em New Jersey foi consideravelmente maior em relação a de Pennsylvania, então haveria indício de um efeito positivo e significativo da mudança na política salarial no nível de emprego.

a) *Calcule a diferença de médias das variações de empregos, entre fev/92 e nov/92, das lojas tratadas e das lojas de controle. A diferença foi significativa?*

b) *Para avaliar o quão razoável é supor que ambos estados são similares no mercado de trabalho de fast food, compararemos características de New Jersey e de Pennsylvania por meio de diferenças de médias para as variáveis: (i-iv) dummies de cada rede de fast food, (v) nº de empregados em fev/92, (vi) proporção de empregados em tempo integral, (vii) salário por hora e (viii) tempo de funcionamento. Construa uma tabela com as médias de cada estado, a diferença entre eles e o p valor:*

	Pennsylvania	New Jersey	dif.	p valor
dmy_bk	-	-	-	-
dmy_kfc	-	-	-	-
dmy_roys	-	-	-	-
dmy_wendys	-	-	-	-
emptot_feb92	-	-	-	-
pct_fte	-	-	-	-
wage_st	-	-	-	-
hrsopen	-	-	-	-

A partir desta análise, é razoável supor que os mercados de trabalho das redes de fast food eram similares em New Jersey e Pennsylvania?

Resposta:

a)
```
library(dplyr)
fastfood = fastfood %>%
  mutate(treated = ifelse(state=="New Jersey", 1, 0),
           var_emp = emptot_nov92 - emptot_feb92)

lm(var_emp ~ treated, data = fastfood) %>% summary()
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.2833     0.7313  -3.122  0.00186 **
treated        2.7500     0.8152   3.373  0.00078 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

Houve um aumento significativo na variação do nº de empregados nas redes de fast food de New Jersey (tratado) em relação ao de Pennsylvania (controle).

b)
```
# Diferenças de médias - variáveis pré-tratamento
fastfood = fastfood %>%
  mutate(dmy_bk = ifelse(chain=="bk", 1, 0),
           dmy_kfc = ifelse(chain=="kfc", 1, 0),
           dmy_roys = ifelse(chain=="roys", 1, 0),
           dmy_wendys = ifelse(chain=="wendys", 1, 0))

lista_var = c("dmy_bk", "dmy_kfc", "dmy_roys", "dmy_wendys",
              "emptot_feb92", "pct_fte", "wage_st", "hrsopen")

resultados = matrix(0, length(lista_var), 4) %>%
as.data.frame()
colnames(resultados) = c("Pennsylvania", "New Jersey", "dif", "p valor")
rownames(resultados) = lista_var

for (var in lista_var) {
  reg = lm(eval(parse(text=var)) ~ treated, fastfood) %>% summary() %>% coef()
  resultados[var, 1] = reg["(Intercept)", "Estimate"]
  resultados[var, 2] = reg["(Intercept)", "Estimate"] + reg["treated", "
    Estimate"]
  resultados[var, 3] = reg["treated", "Estimate"]
  resultados[var, 4] = reg["treated", "Pr(>|t|)"]
}
resultados %>% round(4)
```

```
             Pennsylvania New Jersey     dif p valor
dmy_bk             0.4430     0.4109 -0.0322  0.4619
dmy_kfc            0.1519     0.2054  0.0535  0.1274
dmy_roys           0.2152     0.2477  0.0325  0.3910
dmy_wendys         0.1899     0.1360 -0.0539  0.0851
emptot_feb92      23.3312    20.4394 -2.8918  0.0009
pct_fte           32.7265    34.3406  1.6141  0.4582
wage_st            4.6240     4.8480  0.2240  0.0000
hrsopen           14.5892    14.4191 -0.1701  0.4917
```

*`eval(parse(text="nome_variavel"))` foi utilizado para transformar um texto em um nome de objeto. Assim, podemos "chamar" um objeto e, neste caso, usá-lo dentro da fórmula do modelo na função `lm()`.

Note que, para a maioria das variáveis analisadas, as diferenças não são estatisticamente significantes a 5%. Há diferença apenas para o salário por hora (4,8% superior em New Jersey) e na quantidade de empregados por loja (Pennsylvania possui 3 funcionários a mais em média), o que pode indicar alguma fragilidade na hipótese de que as tendências de comportamento dos dois estados seguiriam semelhantes na ausência de tratamento. □

Exercício 4. *Nos EUA, discute-se bastante a questão da discriminação dos povos hispânicos (mexicanos, porto riquenhos, cubanos, etc.). Analisando* ***apenas trabalhadores em tempo integral e parcial****, verificaremos a possível discriminação na renda usando a base de dados General Social Survey (GSS), a mesma utilizada no Exercício 2 e que pode ser carregada no R usando o comando:*

```
load(url("http://bit.ly/dasi_gss_data"))
```

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, os anos de estudo (`educ`) e o status de trabalho (`wrkstat`).*

b) *Você acha razoável considerar que o resultado do item (a) como representativo de toda população trabalhadora nos EUA? Justifique resumindo/analisando os dados.*
Dica: *Verifique se todas observações de `gss` foram utilizadas na regressão no item (a).*

Resposta:

a) Mexicanos a 0,1%, Filipinos a 5%, Argentinos e Porto Riquenhos a 10%:

```
library(dplyr)
load(url("http://bit.ly/dasi_gss_data"))

gss = gss %>%
  # filtrando apenas trabalhadores em tempo integral e parcial
  filter(wrkstat %in% c("Working Fulltime", "Working Parttime")) %>%

  # definindo Not Hispanic como referencia
  mutate(hispanic = relevel(hispanic, ref="Not Hispanic"))

# Regressão
fit = lm(coninc ~ hispanic + sex + age + I(age^2) + race + educ + wrkstat, gss)
summary(fit)
```

```
Call:
lm(formula = coninc ~ hispanic + sex + age + I(age^2) + race +
educ + wrkstat, data = gss)

Residuals:
Min       1Q   Median       3Q      Max
-102538  -25548   -8266   14571  163415

Coefficients:
```

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)                          -76942.435   4104.084 -18.748  < 2e-16 ***
hispanicMexican, Mexican American     -6158.952   1711.843  -3.598 0.000322 ***
hispanicPuerto Rican                  -5831.574   3195.995  -1.825 0.068082 .
hispanicCuban                          -727.544   5755.621  -0.126 0.899413
hispanicSalvadorian                   -6341.522   6006.062  -1.056 0.291060
hispanicGuatemalan                    -1289.261   7959.050  -0.162 0.871319
hispanicPanamanian                   -18409.278  12850.695  -1.433 0.152016
hispanicNicaraguan                   -13707.093  13653.226  -1.004 0.315428
hispanicCosta Rican                   15386.564  22241.821   0.692 0.489088
hispanicCentral American              -6364.602  11644.985  -0.547 0.584698
hispanicHonduran                      -3260.999   9975.247  -0.327 0.743743
hispanicDominican                    -11471.848   8454.734  -1.357 0.174857
hispanicWest Indian                    3739.035  38519.432   0.097 0.922674
hispanicPeruvian                      -6149.380  10701.734  -0.575 0.565564
hispanicEquadorian                   -10500.778  10714.220  -0.980 0.327070
hispanicColumbian                     -5315.638  10301.585  -0.516 0.605864
hispanicVenezuelan                     4602.093  13621.146   0.338 0.735473
hispanicArgentinian                   47851.664  27233.777   1.757 0.078935 .
hispanicSpanish                       -7198.788   5416.767  -1.329 0.183883
hispanicBasque                       -14484.157  38494.135  -0.376 0.706724
hispanicFilipino/A                    41203.692  19262.145   2.139 0.032451 *
hispanicLatin American               -12548.006  22230.069  -0.564 0.572453
hispanicSouth American                -5027.299  12859.208  -0.391 0.695843
hispanicLatino/A                     -14379.493  15748.125  -0.913 0.361215
hispanicHispanic                     -18581.959  22247.698  -0.835 0.403607
hispanicOther, Not Specified         -19508.177  14556.634  -1.340 0.180223
sexFemale                             -8196.156    766.173 -10.698  < 2e-16 ***
age                                    2946.157    176.359  16.705  < 2e-16 ***
I(age^2)                                -27.439      1.988 -13.804  < 2e-16 ***
raceBlack                            -14506.432   1119.148 -12.962  < 2e-16 ***
raceOther                              2511.628   1474.706   1.703 0.088572 .
educ                                   5041.284    135.727  37.143  < 2e-16 ***
wrkstatWorking Parttime               -6199.141   1047.625  -5.917 3.38e-09 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

Residual standard error: 38490 on 10422 degrees of freedom
(23594 observations deleted due to missingness) <<<<<<<<<<<<<<<<<<<<<<<<<<<<<
Multiple R-squared:  0.2148,  Adjusted R-squared:  0.2124
F-statistic: 89.09 on 32 and 10422 DF,  p-value: < 2.2e-16
```

b) Note que 23.594 das 34.049 observações (69,3%) foram excluídas da regressão por causa de missing values (NA's). Os resultados do item (a) podem ser representativas da população americana se a ocorrência de missing values se deu de maneira proporcional nos diferentes estratos da sociedade. Então vamos comparar as diferenças das médias de algumas variáveis para as duas amostras. Já temos a base "completa" `gss`, então precisamos gerar a base utilizada na regressão (sem missing values). Podemos obtê-la omitindo todos NA's das variáveis utilizadas na regressão:

```r
gss_reg = gss %>% filter(!is.na(coninc), !is.na(hispanic), !is.na(sex),
                         !is.na(age), !is.na(educ), !is.na(wrkstat))
nrow(gss_reg) #34049 - 23594 = 10455
```

Como pode ser visto abaixo, grande parte dos NA's ocorrem na variável `hispanic`:

```r
# Número de missing values em cada variável utilizada
```

```r
gss %>% select(coninc, hispanic, sex, age, race, educ, wrkstat) %>%
    is.na() %>% apply(2, sum)
```

```
  coninc  hispanic  sex   age   race  educ  wrkstat
    2600     22434    0   109      0    75        0
```

e, por isso, as proporções dos povos hispânicos alterou muito pouco entre as bases:

```r
# Comparando hispanic na base cheia e na usada em regressão
rbind(prop.table(table(gss$hispanic)), prop.table(table(gss_reg$hispanic))) %>%
    round(4) %>% t()
```

```
                                      [,1]   [,2]
Not Hispanic                        0.8830 0.8844
Mexican, Mexican American, Chicano/A 0.0710 0.0712
Puerto Rican                        0.0145 0.0145
Cuban                               0.0046 0.0043
Salvadorian                         0.0040 0.0040
Guatemalan                          0.0024 0.0023
Panamanian                          0.0009 0.0009
Nicaraguan                          0.0007 0.0008
Costa Rican                         0.0003 0.0003
Central American                    0.0012 0.0011
Honduran                            0.0016 0.0014
Dominican                           0.0022 0.0020
West Indian                         0.0001 0.0001
Peruvian                            0.0011 0.0012
Equadorian                          0.0016 0.0012
Columbian                           0.0016 0.0013
Venezuelan                          0.0007 0.0008
Argentinian                         0.0002 0.0002
Chilean                             0.0000 0.0000
Spanish                             0.0047 0.0049
Basque                              0.0001 0.0001
Filipino/A                          0.0003 0.0004
Latin American                      0.0003 0.0003
South American                      0.0009 0.0009
Latin                               0.0002 0.0000
Latino/A                            0.0007 0.0006
Hispanic                            0.0004 0.0003
Other, Not Specified                0.0007 0.0007
```

Em `gss`, criaremos a variável binária que indica se a observação (`idcase`) está presente em `gss_reg`, ou seja, se possui algum missing value nas variáveis utilizadas na regressão. Também, incluiremos dummies de mulher, raças e trabalho em tempo integral.

```r
# criando variáveis dummies
gss = gss %>% mutate(
    dmy_reg = ifelse(caseid %in% gss_reg$caseid, 1, 0),
    dmy_female = ifelse(sex=="Female", 1, 0),
    dmy_race_white = ifelse(race=="White", 1, 0),
    dmy_race_black = ifelse(race=="Black", 1, 0),
    dmy_race_other = ifelse(race=="Other", 1, 0),
    dmy_fulltime = ifelse(wrkstat=="Working Fulltime", 1, 0)
)
```

Agora, calcularemos as diferenças de média das variáveis de renda, sexo, idade, raça, anos de educação e status de trabalho:

```r
# diferenças de médias
lista_var = c("coninc", "dmy_female", "age", "dmy_race_white",
              "dmy_race_black", "educ", "dmy_fulltime")

resultados = matrix(0, length(lista_var), 4) %>%
as.data.frame()
colnames(resultados) = c("base completa", "base reg", "dif", "p valor")
rownames(resultados) = lista_var

for (var in lista_var) {
  reg = lm(eval(parse(text=var)) ~ dmy_reg, gss) %>% summary() %>% coef()
  resultados[var, 1] = reg["(Intercept)", "Estimate"]
  resultados[var, 2] = reg["(Intercept)", "Estimate"] + reg["dmy_reg", "
    Estimate"]
  resultados[var, 3] = reg["dmy_reg", "Estimate"]
  resultados[var, 4] = reg["dmy_reg", "Pr(>|t|)"]
}
resultados %>% round(4)
```

```
               base completa   base reg       dif p valor
coninc           48686.4779 56103.1536 7416.6757  0.0000
dmy_female           0.4752     0.4938    0.0186  0.0015
age                 39.8771    41.7068    1.8297  0.0000
dmy_race_white       0.8342     0.7646   -0.0696  0.0000
dmy_race_black       0.1299     0.1373    0.0073  0.0663
educ                13.2103    13.8810    0.6707  0.0000
dmy_fulltime         0.8279     0.8297    0.0018  0.6893
```

Note que apenas as proporções de negros e de trabalhadores em tempo integral são estatisticamente iguais a 5%. Já as diferenças de idade, de proporção de mulheres e de anos de educação, embora sejam significativas, não são tão diferentes em números absolutos.

□