Tutorial 1 aula 2

Mônica Rocabado

Resolucao individual do exercício.

library(tidyverse)  
pnad <- read\_csv("pnad2015.csv")

#### *PERGUNTA 1*

*O grupo começa estimando o impacto do sexo, uma variável binária (dummy) igual a 1 para mulheres (0 caso contrário), no salário (variável salariom) através do modelo de regressão linear simples. O grupo interpreta o coeficiente de sexo. Alguém do grupo pergunta qual seria o salário médio de homens de acordo com os resultados da regressão do grupo. A mesma pessoa pergunta se esse resultado seria suficiente para dizer que existe discriminação de gênero. O grupo responde as perguntas, justificando a resposta*

mqo <- lm(salariom ~ sexo, data = pnad, na.action=na.exclude)  
summary(mqo)

##   
## Call:  
## lm(formula = salariom ~ sexo, data = pnad, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1876 -1076 -591 124 198124   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1875.900 8.825 212.6 <2e-16 \*\*\*  
## sexo -485.204 13.516 -35.9 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2699 on 163044 degrees of freedom  
## (145553 observations deleted due to missingness)  
## Multiple R-squared: 0.007842, Adjusted R-squared: 0.007836   
## F-statistic: 1289 on 1 and 163044 DF, p-value: < 2.2e-16

**R:** O Salário médio dos homens, de acordo com o intercepto é 1875,9 reais, caso for 1, ou seja, mulheres, há uma queda de 485,2 reais de salário, sendo assim seu salário médio de 1390 reais. Todavia isso nao é o suficiente para determinar se o discriminacao de genero, dado que há outros fatores que podem levar a esse resultado, como tipo de trabalho que leva a essa remuneracao, anos de escolaridade, entre outros, ainda que seja um indicativo que isto pode ocorrer.

#### *PERGUNTA 2*

*A vice-diretora olha os resultados preliminares e diz que é importante colocar outras variáveis de controle no modelo acima, tais como escolaridade, idade,cor/raça e uma variável dummy igual a 1 se o indivíduo mora em região rural (0 caso contrário). O grupo roda a regressão incluindo as variáveis de controle e interpreta novamente o coeficiente de sexo, comparando com o coeficiente encontrado na questão anterior.*

mq1 <- lm(salariom ~ sexo+educ+idade+cor+rural, data = pnad, na.action=na.exclude)  
summary(mq1)

##   
## Call:  
## lm(formula = salariom ~ sexo + educ + idade + cor + rural, data = pnad,   
## na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5318 -942 -319 382 196800   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1130.7062 31.2512 -36.18 <2e-16 \*\*\*  
## sexo -802.3504 12.6039 -63.66 <2e-16 \*\*\*  
## educ 223.3453 1.5968 139.87 <2e-16 \*\*\*  
## idade 39.5017 0.4718 83.73 <2e-16 \*\*\*  
## cor -70.5098 2.1922 -32.16 <2e-16 \*\*\*  
## rural -333.6729 18.0959 -18.44 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2484 on 162569 degrees of freedom  
## (146024 observations deleted due to missingness)  
## Multiple R-squared: 0.162, Adjusted R-squared: 0.162   
## F-statistic: 6286 on 5 and 162569 DF, p-value: < 2.2e-16

Nota-se uma mudanca no valor do intercepto e também em sexo, quase dobrando o valor negativo para mulher, ou seja, as variáveis possuem alguma relacao entre si.

A lógica do modelo de regressao multipla é *“O modelo de regressão linear múltipla mantém os valores das outras variáveis independentes fixos mesmo se houver correlação”*

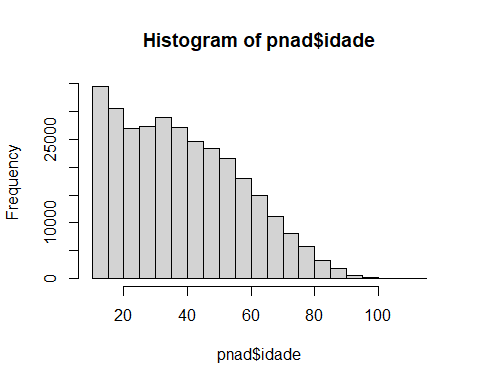
Logo no caso, mantendo as variáveis constantes:

* sexo: Ser mulher (1) diminui o salário em 802 reais
* educ: Um ano a mais de escolaridade aumenta o salário em 223 reais
* idade: Um ano a mais de vida aumenta a o salário em 39 reais
* cor: **?Dúvida:** como explicar se cor possui várias opcoes categoricas??
* rural: Morar em regiao rural diminui o salário em 333 reais

#### *PERGUNTA 3*

*O grupo mostra os resultados da regressão acima para a vice-diretora do centro de pesquisa. Ela pergunta se o grupo considerou um possível efeito não linear da idade no salário. O grupo cria uma variável de idade ao quadrado e estima um modelo com idade e idade ao quadrado na regressão incluindo as variáveis da questão anterior. O grupo interpreta o coeficiente de idade*

hist(pnad$idade)



pnad <- pnad %>%   
 mutate(idadequadrado = idade\*idade)

Agora fazendo a regressao:

mq2 <- lm(salariom ~ sexo+educ+idade+cor+rural+idadequadrado, data = pnad, na.action=na.exclude)  
summary(mq2)

##   
## Call:  
## lm(formula = salariom ~ sexo + educ + idade + cor + rural + idadequadrado,   
## data = pnad, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4466 -940 -312 379 196749   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1972.0934 50.9487 -38.71 <2e-16 \*\*\*  
## sexo -812.3797 12.5962 -64.49 <2e-16 \*\*\*  
## educ 219.1662 1.6072 136.37 <2e-16 \*\*\*  
## idade 87.8082 2.3596 37.21 <2e-16 \*\*\*  
## cor -71.8113 2.1902 -32.79 <2e-16 \*\*\*  
## rural -312.2448 18.1008 -17.25 <2e-16 \*\*\*  
## idadequadrado -0.5851 0.0280 -20.89 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2480 on 162568 degrees of freedom  
## (146024 observations deleted due to missingness)  
## Multiple R-squared: 0.1643, Adjusted R-squared: 0.1642   
## F-statistic: 5325 on 6 and 162568 DF, p-value: < 2.2e-16

**R:** As informacoes no geral se mantiveram, no entanto, se nota uma diferenca do coeficiente para variável idade e da variavel idadequadrado, em que o primeiro está positivo e o segundo negativo, logo no primeiro a cada ano de idade, um aumento de 87 reais, enquanto no segundo a cada ano adicional de idade, deve-se aplicar o efeito na formula

Efeito marginal da idade: 219.1662idade - 2(-0.5851)idade

Já que: “A interpretação da estimativa do estimador de 𝑥1 vai depender do seu valor”

(Comentário: modelo quadrático. Há valores ótimos existentes, o quadratico vai gerar um gráfico de curva com ponto ótimo, derivada igual a zero encontra o ponto ótimo. A interpretacao muda: o efeito é um aumento ou diminuicao de b1+b2+2b2x)

#### *PERGUNTA 4*

\_A vice-diretora do projeto sugere que seria interessante estimar usar a transformação logarítmica na variável dependente, e pede para estimar uma *regressão do logaritmo natural do salário mensal no sexo e no número de horas de afazeres domésticos não remunerado (variável horasdom)*, *além das outras variáveis da questão anterior. O grupo tenta rodar uma nova regressão usando o logaritmo natural do salário mensal como variável dependente e sexo e horas de afazeres domésticos não remunerado como variáveis explicativas, incluindo as variáveis de controle da questão anterior, mas obtém o erro.*

*O erro quer dizer que a transformação logarítmica produziu o valor -∞ para indivíduos sem salário mensal (𝑠𝑎𝑙á𝑟𝑖𝑜𝑚 = 0), pois o logaritmo natural de 0 não é determinado. O grupo conversa com uma aluna terminando sua dissertação e ela sugere rodar o código abaixo para transformar os valores -∞ em dados faltantes (NA), onde pnad é o nome do objeto que armazena a base de dados.* *O grupo corrige o problema e roda novamente a regressão linear do logaritmo natural do salário mensal no sexo, nas horas de trabalho doméstico não remunerado e nas outras variáveis de controle da questão anterior. Funcionou!!! Agora sim, O grupo interpreta os coeficientes das horas de afazeres doméstico não remunerado, escolaridade e sexo*

pnad <- pnad %>%   
 mutate(logsalariom = log(salariom))  
  
is.na(pnad) <- sapply(pnad, is.infinite) # transformando todos os - Inf # em missing data (NA)  
  
#modelo log linear, em que a variável dependente y fica com log  
lqo3 <- lm(logsalariom ~ sexo+educ+idade+cor+rural+idadequadrado+horasdom, data = pnad, na.action=na.exclude)  
summary(lqo3)

##   
## Call:  
## lm(formula = logsalariom ~ sexo + educ + idade + cor + rural +   
## idadequadrado + horasdom, data = pnad, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.5178 -0.3824 0.0157 0.4080 5.3312   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.002e+00 1.987e-02 251.71 <2e-16 \*\*\*  
## sexo -3.497e-01 4.862e-03 -71.92 <2e-16 \*\*\*  
## educ 1.004e-01 5.721e-04 175.58 <2e-16 \*\*\*  
## idade 7.266e-02 9.417e-04 77.15 <2e-16 \*\*\*  
## cor -3.542e-02 7.696e-04 -46.02 <2e-16 \*\*\*  
## rural -3.675e-01 7.325e-03 -50.17 <2e-16 \*\*\*  
## idadequadrado -7.063e-04 1.124e-05 -62.85 <2e-16 \*\*\*  
## horasdom -1.089e-02 2.101e-04 -51.82 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7004 on 105276 degrees of freedom  
## (203315 observations deleted due to missingness)  
## Multiple R-squared: 0.3804, Adjusted R-squared: 0.3804   
## F-statistic: 9234 on 7 and 105276 DF, p-value: < 2.2e-16

A interpretacao muda, pois nao tenho mais y, tenho log(y), assim um aumento de uma unidade de x nao leva a um aumento esperado de b1 em y. Aumentar 1 unidade X leva um aumento *relativo* de b1\*100% em y

* sexo: -0.34\*100 = Ser mulher e com as demais variáveis constantes leva a uma reducao de -34% do salário.
* educ: -0.10\*100 = Cada ano de escolaridade e com as demais variáveis constantes leva a uma reducao -10% do salário (???)
* idade: 0.07\*100 = 7%
* cor: -0.035\*100 = -3.5% (nao sei como interpretar)
* rural: -0.36\*100 = Ser do rural e com as demais variáveis constantes leva a uma reducao de -36% do salário
* idadequadrado: (0.07*100) - 2(-0.36*100) = 79%
* horasdom: -0.01\*100 = A cada hora de atividade domiciliar e com as demais variáveis constantes leva a uma reducao -1% do salário

#### *PERGUNTA 5*

*Um aluno ingressante do mestrado acadêmico pergunta sobre a(s) razão(ões) pela qual fazemos a transformação logarítmica em uma variável dependente, e o grupo explica o(s) principais motivo(s)*

Motivos para usar o log:

1. Quando a relacao entre x e y é exponencial e nao linear
2. Quando o resíduo nao possui distribuicao normal

Sendo assim, respectivamente: 1. lineariza as relacoes, reduz a complexidade das relacoes matemáticas. 2. O log de y, leva a um um padrao esperado de uma distribuicao normal, pois simetriza e concentra valores, entao tende tornar uma coisa assimétrica para simétrica e trazer valores distantes para mais perto.

#### *PEGUNTA 6*

*Por fim, a vice-diretora comenta que a transformação logarítmica nas horas de afazeres domésticos também poderia ser feita, pois poderia facilitar a interpretação. O grupo faz a transformação e estima o modelo da questão 4. O grupo interpreta os coeficientes do log(ℎ𝑜𝑟𝑎𝑠𝑑𝑜𝑚) e escolaridade*

pnad <- pnad %>%   
 mutate(loghorasdom = log(horasdom))  
  
lqo4 <- lm(logsalariom ~ sexo+educ+idade+cor+rural+idadequadrado+loghorasdom, data = pnad, na.action=na.exclude)  
summary(lqo4)

##   
## Call:  
## lm(formula = logsalariom ~ sexo + educ + idade + cor + rural +   
## idadequadrado + loghorasdom, data = pnad, na.action = na.exclude)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.5330 -0.3786 0.0185 0.4105 5.3632   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.163e+00 2.035e-02 253.65 <2e-16 \*\*\*  
## sexo -3.583e-01 4.936e-03 -72.60 <2e-16 \*\*\*  
## educ 1.009e-01 5.737e-04 175.89 <2e-16 \*\*\*  
## idade 7.242e-02 9.455e-04 76.60 <2e-16 \*\*\*  
## cor -3.543e-02 7.721e-04 -45.88 <2e-16 \*\*\*  
## rural -3.677e-01 7.349e-03 -50.04 <2e-16 \*\*\*  
## idadequadrado -7.047e-04 1.128e-05 -62.48 <2e-16 \*\*\*  
## loghorasdom -1.318e-01 2.948e-03 -44.71 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7027 on 105276 degrees of freedom  
## (203315 observations deleted due to missingness)  
## Multiple R-squared: 0.3764, Adjusted R-squared: 0.3764   
## F-statistic: 9080 on 7 and 105276 DF, p-value: < 2.2e-16

(Comentários: Em caso de modelo linear log, onde y se mantém e o log ocorre nas variáveis independentes Interpretacao aproximada/relativa. aumentar 1% induz um aumento ou diminuicao de b1/100 unidades em y)

* horasdom: -0.131\*100 = A cada hora de atividade domiciliar e com as demais variáveis constantes leva a uma reducao -13% do salário
* educ: 0.1\*100 = A cada ano de escolaridade e com as demais variáveis constantes leva a um aumento de 10% do salário

##### Duvidas gerais para aula

* Como analisar o resultado quando há todas as formas combinadas (log-linear, linear-log, quadrático)?
* Como avaliamos a relacao entre as variáveis independentes? Isso importa para análise?
* Qual o momento que retiramos uma variável ou quando percebemos que ela nao é relevante para o modelo?
* Como analisar o coeficiente quadrático? Realmente a A interpretação da estimativa do estimador de 𝑥1 vai depender do seu valor?
* Como interpretar o coeficiente de variáveis categoricas como cor?