Bruno, Camila Costa, Emerson ,Mônica, Thais

1. A vice-diretora do centro pede para estimar uma regressão do salário mensal (variável salariom) no número de filhos (variável filhos) e **interpretar os coeficientes do modelo.**

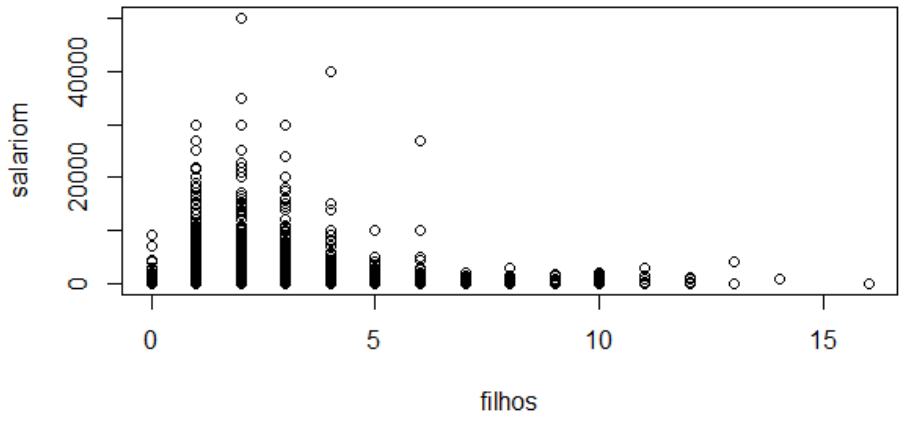
```
Call:
lm(formula = salariom ~ filhos, data = filhos_base, na.action = na.exclude)

Residuals:
   Min     1Q Median     3Q    Max
 -1541   -655   -339    108  48783

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1540.835     17.688   87.11   <2e-16 ***
filhos      -162.167      7.214  -22.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1635 on 32127 degrees of freedom
Multiple R-squared:  0.01548,   Adjusted R-squared:  0.01545
F-statistic: 505.3 on 1 and 32127 DF,  p-value: < 2.2e-16
```

Y= ax + b

A(intercept)

B (filhos)

Salariosm = 1540.835*x - 162



Comentários sobre a variável

Intercept – coeficiente angular (?) - is the expected mean value of Y when all X=0)

Filhos – quando y = 0

Desvio padrão (estimated std) - Rquadrado ajustado – R2 sempre aumenta quando você inclui uma variável. Então precisamos usar o R^2 ajustado é o mais indicado pois ele ‘ajusta’ esse ponto.

Pr(>|t|) - o coeficiente precisa ser baixo (significante). Geralmente precisa ser abaixo de 0,05

O valor do coeficiente angular da equação (-162.167, com desvio padrão de mais ou menos 7) aponta que, para cada filho adicional, o valor do salário mensal de mulheres entre 18 e 45 anos cai

162.167 unidades monetárias. Logo uma mulher sem filhos entre 18 e 45 anos, o salário médio estimado seria 1.540 unidades monetárias (o intercepto), com desvio padrão de mais menos 17.

Embora os coeficientes dos parâmetros sejam significantes, o valor observado no R quadrado (0.01548) é baixo, o que significa que a capacidade explicativa do modelo, especificado com apenas uma variável independente (número de filhos), é baixo.

2. Um aluno do primeiro ano do semestre do mestrado acadêmico que faz parte do grupo mostra o resultado do RStudio na figura abaixo. Ele não entende o que o R-quadrado (Multiple R-Squared) significa e pede que o grupo explique.

```
Residual standard error: 1635 on 32127 degrees of freedom
Multiple R-squared:  0.01548,   Adjusted R-squared:  0.01545
F-statistic: 505.3 on 1 and 32127 DF,  p-value: < 2.2e-16
```

O R-quadrado representa a proporção da variabilidade na variável resposta explicada pela variável preditora. Ou seja, apenas 1,5% da explicação da variação do salário pode ser explicado pelo número de filhos, utilizando uma regressão linear. Desta forma, pode-se concluir que existe uma baixa correlação.

3. A vice-diretora pede para acrescentar a variável de escolaridade no modelo como variável de controle. Em outras palavras, ela gostaria que o grupo estimasse a regressão de salário mensal em filhos e escolaridade (variável educ). Interprete o coeficiente de escolaridade. O que aconteceu com o coeficiente de filhos? Explique.

```
Call:
lm(formula = salariom ~ filhos + educ, data = filhos_base, na.action = na.exclude)

Residuals:
   Min     1Q Median     3Q    Max
 -2134   -686   -254    263  47896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -380.632     32.350  -11.77   <2e-16 ***
filhos        10.149      7.198    1.41    0.159
educ         164.263      2.373   69.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1527 on 31974 degrees of freedom
  (152 observations deleted due to missingness)
Multiple R-squared:  0.1437,    Adjusted R-squared:  0.1437
F-statistic:  2684 on 2 and 31974 DF,  p-value: < 2.2e-16
```

A Pr(>|t|) para filhos é igual a 0.159, o que indica que a variável filhos é não significante a 10%, logo não ajuda a explicar os efeitos sobre a renda mensal.

Entretanto, o parâmetro de número de anos de estudo (educ) se mostra significante, com valor de 164.263, o que significa que, para cada ano de estudo adicional, o adicional estimado de renda é de 164.263 unidades monetárias.

Adjusted R- squared: é maior do que o modelo anterior (o que em teoria é bom), porém valor ainda é baixo.

4. Por fim, a vice-diretora do centro pede para estimar a regressão do salário mensal no número de filhos, escolaridade e idade (variável idade), **e interpretar o coeficiente de idade.** Um aluno novato nota que o R-quadrado do modelo aumentou, mas não entende a razão. O grupo explica para ele. Por fim, a vice-diretora pede que o grupo crie um código no R que estime o salário mensal predito (estimado), $salariom\hat{}$, para uma mulher com 1 filho, 15 anos de escolaridade e 38 anos de idade.

```
Call:
lm(formula = salariom ~ filhos + educ + idade, data = filhos_base,
    na.action = na.exclude)

Residuals:
   Min     1Q Median     3Q    Max
 -2496   -645   -242    258  47582

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1452.955     51.576 -28.171  < 2e-16 ***
filhos        -42.940      7.397  -5.805 6.49e-09 ***
educ          161.383      2.350  68.668  < 2e-16 ***
idade          34.916      1.317  26.511  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1511 on 31973 degrees of freedom
  (152 observations deleted due to missingness)
Multiple R-squared:  0.1622,    Adjusted R-squared:  0.1621
F-statistic:  2063 on 3 and 31973 DF,  p-value: < 2.2e-16
```

O código utilizado:

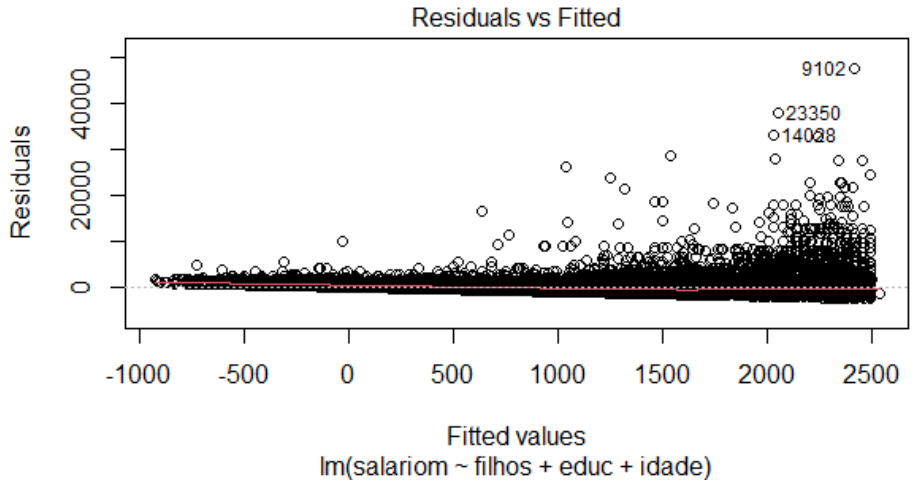
```
mqo3 <- lm(salariom ~ filhos+educ+idade, data = filhos_base, na.action=na.exclude)

summary(mqo3)

valpred<-data.frame(educ=15, filhos =1, idade=38)

predict(mqo3, valpred)
```

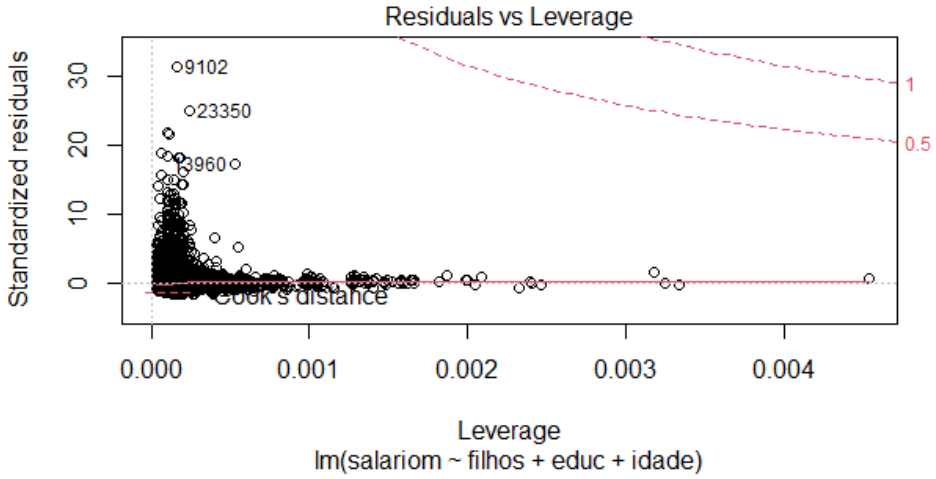
O Salário mensal é 2251.659 unidade monetárias.

O R –quadrado aumentou em relação aos modelos anteriores por conta da inclusão de nova variável.

5. Conversando com o grupo de pesquisa, surge a dúvida se o grupo deveria fazer a **transformação logarítmica no salário mensa**l. O grupo usa o **modelo da questão anterior para criar um gráfico** e responde à dúvida, justificando a resposta.

Devido a distribuição do salário ser distorcida, em que o valor mínimo é 0 e máximo é 50mil unidades monetárias, a transformação da variável em log nos permite normalizá-la.





Comentários

Summary ()

- Para uma variável só me da a estatistica descritiva (min, q1, med, q3, média, max
- Da para ver se segue uma distribuição normal (continua, quantitativa e discretas)
- Summary na regressão (summary (mqo)) faz de uma regressão (ou seja, posso aplicar para a variável ou para a regressão)

Salariom = b0 +b1*filhos + U

U - são todas os erros (ou seja, caracteristicas não observáveis)