Retorno sobre o investimento em um MBA

Objetivo

O objetivo desta análise é verificar se existe retorno financeiro em realizar um curso de MBA.

Desejamos saber se uma estudante com as características abaixo deve investir na realização do curso, determinando qual será o seu salário após a graduação.

A estudante possui as caracteristicas abaixo:

A) Mulher.
B) Língua mãe não é o inglês.
C) GMAT 680-700.
D) Aluna com desempenho no 1o quartil entre os alunos
E) 8 anos de experiência de trabalho.
F) 29 anos.

Dados da amostragem

Os seguintes dados, de 274 estudantes observados, foram coletados.

```
mba <- read.csv("data/mba.csv", stringsAsFactors = FALSE)
names(mba)
```

```
##  [1] "age"      "sex"      "gmat_tot" "gmat_qpc" "gmat_vpc" "gmat_tpc"
##  [7] "s_avg"    "f_avg"    "quarter"  "work_yrs" "frstlang" "salary"
## [13] "satis"
```

As variáveis de interesse presentes na amostragem são:

A) **sex**, o sexo do. Temos codificada esta informação com os valores 1 (*masculino*) ou 2 (*feminino*).
B) **frstlang**, first language, se a língua mãe é inglesa ou não. Temos codificada esta informação com os valores 1 (*língua mãe é inglês*) ou 2 (*língua mãe* ***não*** *é inglês*).
C) **gmat_tot**, o valor total que o estudante obteve na sua prova de GMAT.
D) **quarter**, em que quartil o estudante possui desempenho nos cursos já realizados entre os alunos.
E) **work_yrs**, a quantidade em anos que o estudante possui de experiência de trabalho.
F) **age**, a idade em anos do estudante.

Análise exploratória do salário e QA da amostra.

Desejamos saber qual o salário de um ex-estudante após se formar. Este atributo corresponde à variável **salary** em nossa base.

Uma análise exploratória sumarisa o salário está distribuido dentro da amostra:

Média salarial:

```r
mean(mba$salary)
```

```
## [1] 39025.69
```

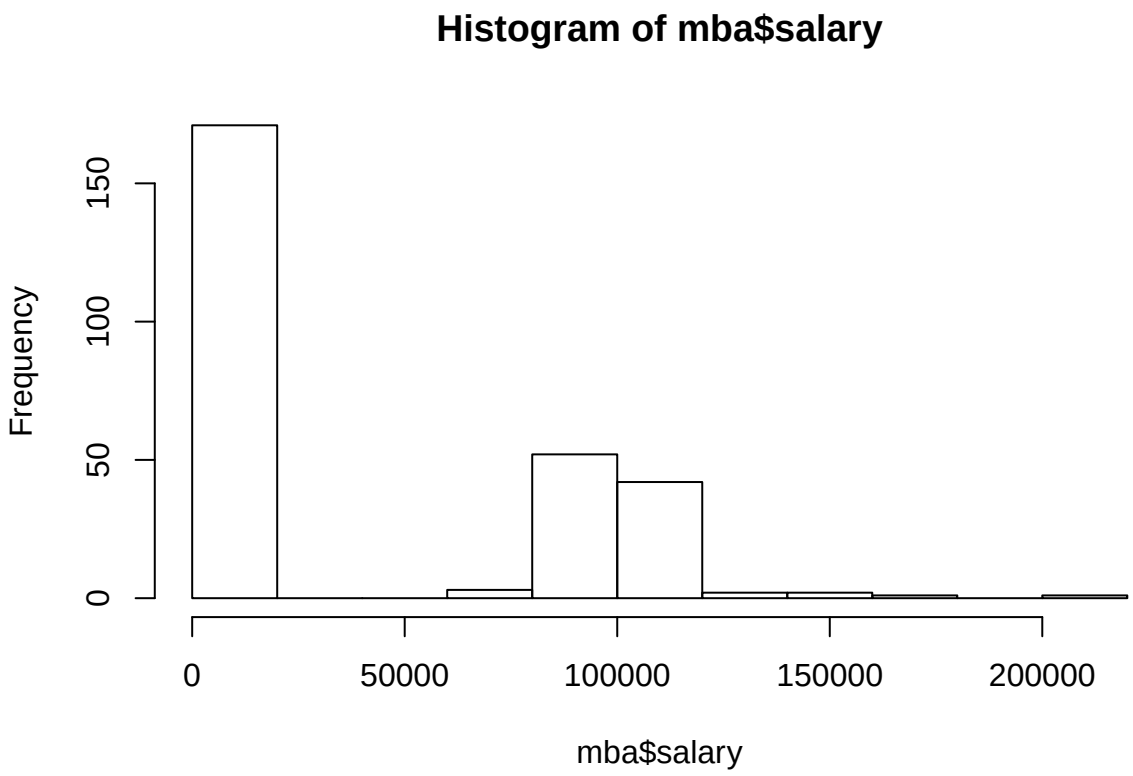
Mediana salarial:

```r
median(mba$salary)
```

```
## [1] 999
```

Histograma salário:

```r
hist(mba$salary)
```



Valores de salário inválidos.

A discrepância entre a média e a mediana leva ao questionamento da qualidade da variável salário na amostra. Verificando-se a amostra, pode-se notar uma grande quantidade de salários com valores 989 e 999. Estes valores foram utilizados na amostragem para denotar profissionais não alocados. Assim, estas observações, com valores 998 e 999, serão excluídas da análise.

```r
mba2 <- mba %>% filter(salary != 998 & salary != 999)
```

Os dados da amostragem após a filtragem possuem o perfil:

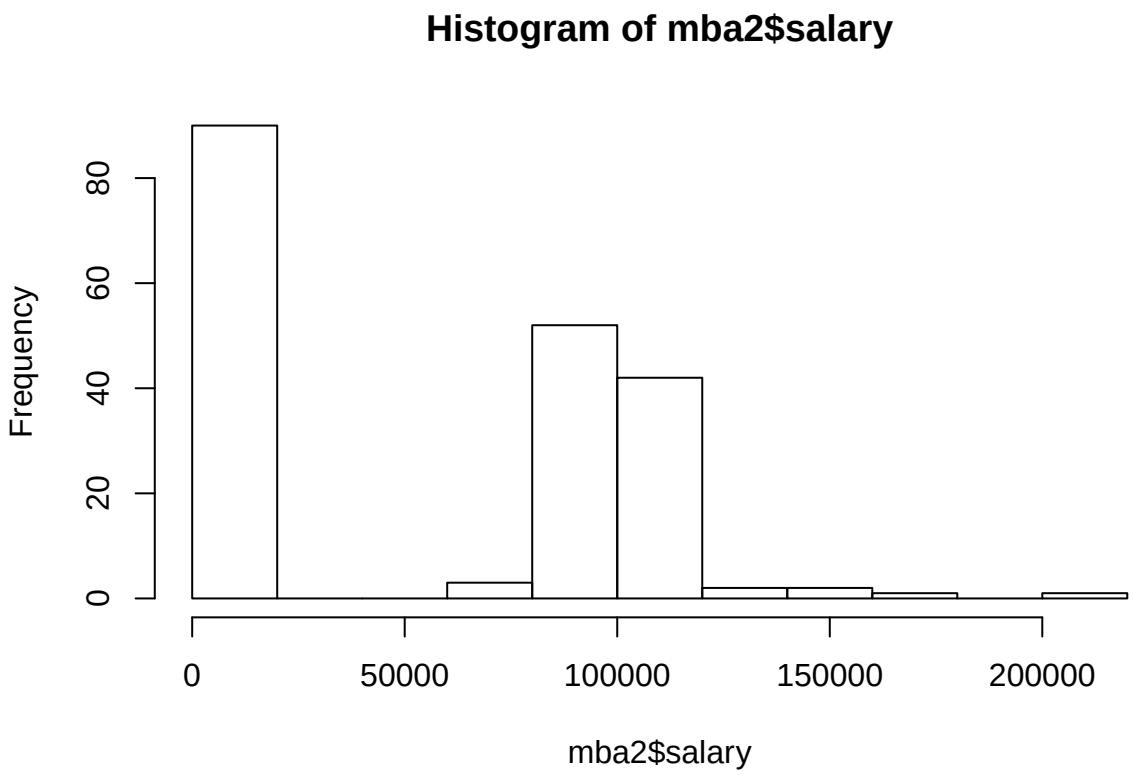
```r
mean(mba2$salary)
```

```
## [1] 54985.32
```

```r
median(mba2$salary)
```

```
## [1] 85000
```

```r
hist(mba2$salary)
```



Valores de salário com zero.

A discrepância entre a média e a mediana, apesar de haver diminuida, leva a uma nova análise do salário. Explorando a base de dados podemos constatar observações com valor igual a zero, totalizando 90 de um total de 193. Estes salários são referêntes a ex-alunos sem informação.

```r
length(mba2$salary)
```

```
## [1] 193
```

```r
sum(mba2$salary == 0)
```

```
## [1] 90
```

Optou-se por retirar-los da amostra da análise, visto que a amostragem ocorreu pouco tempo após a finalização do curso, e estes ex-alunos ainda podem não se encontrar empregados. Filtramos a amostragem retirando também estas amostras.

```r
mba3 <- mba2 %>% filter(salary != 0)
```

Os dados da amostra resultante possuem o perfil abaixo, com média, mediana e distribuição mais próximo ao esperado.

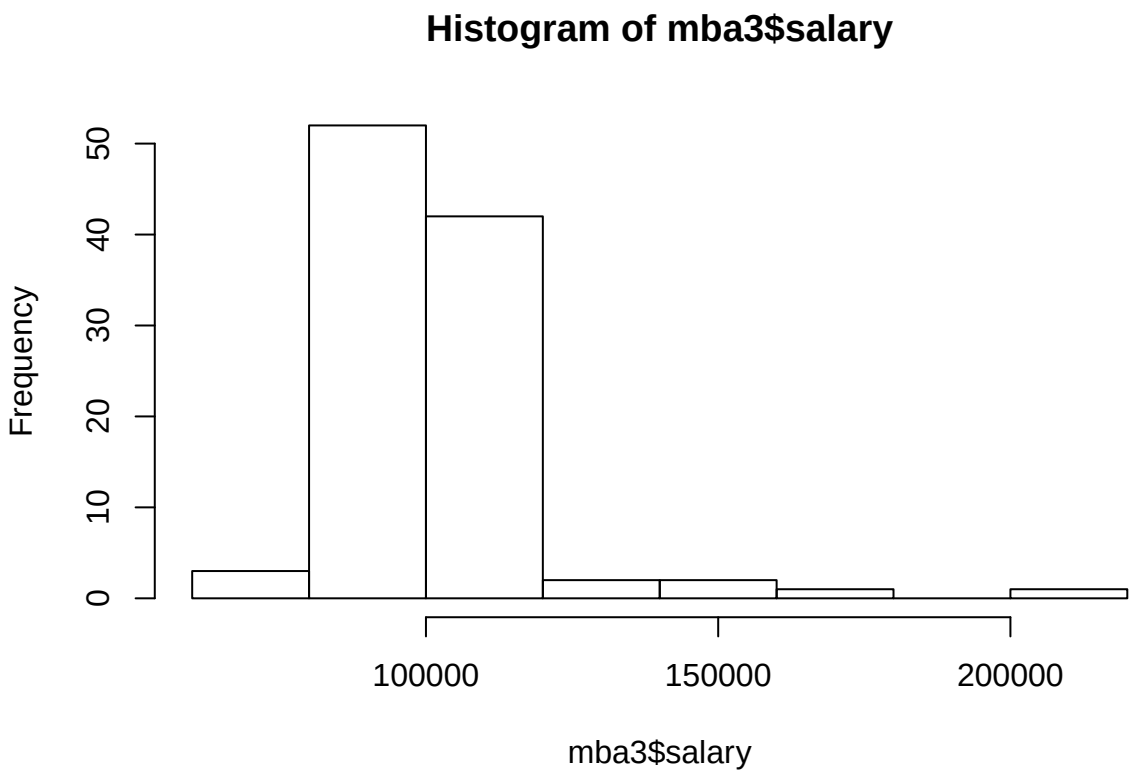
```r
mean(mba3$salary)
```

```
## [1] 103030.7
```

```r
median(mba3$salary)
```

```
## [1] 100000
```

```
hist(mba3$salary)
```



Análises de relevância

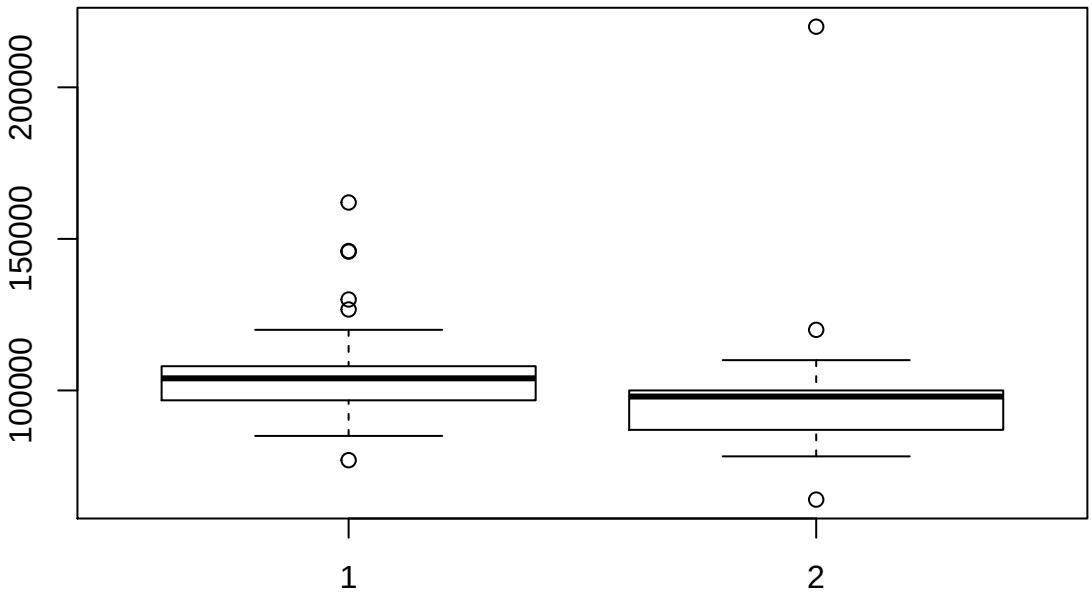
Neste ponto, desejamos identificar quais variáveis influenciam ou não o salário. Assim, iremos realizar uma análise de relevância para as variáveis que conhecemos da estudante alvo.

Gênero influencia no salário?

Desejamos saber se o gênero influencia no salário final. Iremos neste caso utilizar o teste "One way ANOVA". Temos portanto:

- Hipótese Nula, H_0: salário não é influenciada pelo gênero.
- Hipótese Alternativa, H_A: salário é influenciada pelo gênero.

```
boxplot(formula = mba3$salary ~ mba3$sex)
```




```
oneway.test(mba3, formula = salary ~ sex)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  salary and sex
## F = 1.8573, num df = 1.000, denom df = 38.115, p-value = 0.1809
```

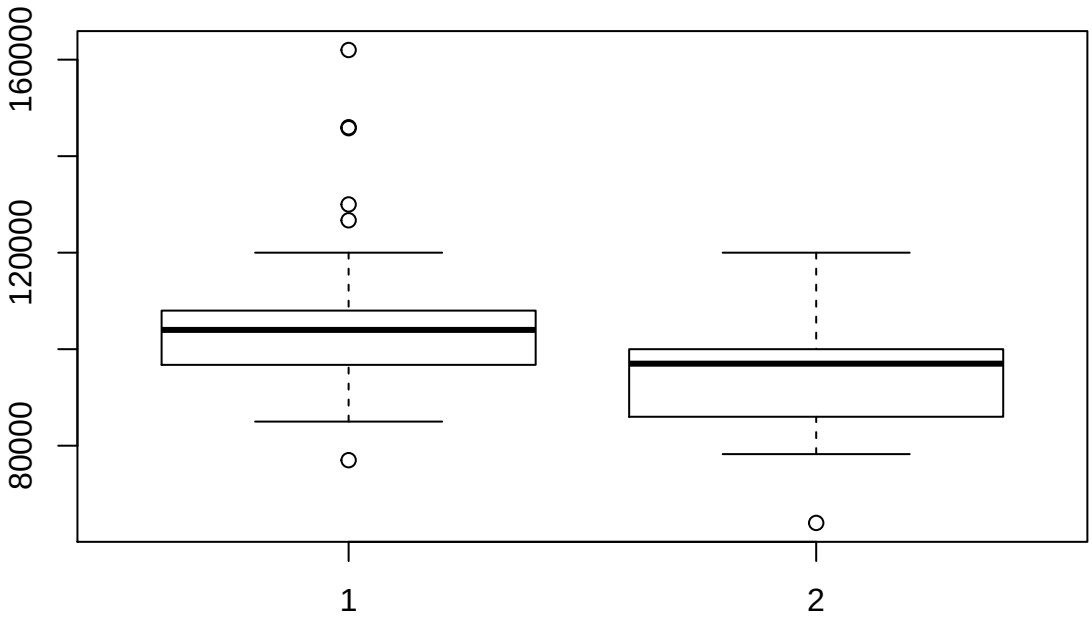
Verificamos que existe um salário que destoa fortemente das outras observações, Após pesquisa, descobriu-se que correspondente à uma executiva empregada em um grupo pertencente a sua familia, com salário muito acima do praticado no mercado, sendo assim estatisticamente distorcivo.

Retiramos portanto a observação dos dados.

```
mba4 <- filter(mba3, salary != max(mba3$salary))
```

Sem esta observação, podemos realizar o teste novamente e constatar a relevância do gênero para a análise

```
boxplot(formula = mba4$salary ~mba4$sex)
```




```
oneway.test(mba4, formula = salary ~ sex)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  salary and sex
## F = 17.729, num df = 1.000, denom df = 70.693, p-value = 7.384e-05
```

Com p-value perto de 0, concluímos que o salário é influenciado pelo gênero.

Língua materna influencia salário?

Desejamos saber se a língua materna influencia no salário final. Uma primeira análise das observações mostra um número estatisticamente baixo de estudantes que não possuem a língua inglesa como lingua mãe (apenas 6 do total de 102 observações)

```
# 1 - inglês | 2 - outros
unique(mba4$frstlang)
```

```
## [1] 1 2
```

```
sum(mba4$frstlang == 1)
```

```
## [1] 96
```

```
sum(mba4$frstlang == 2)
```

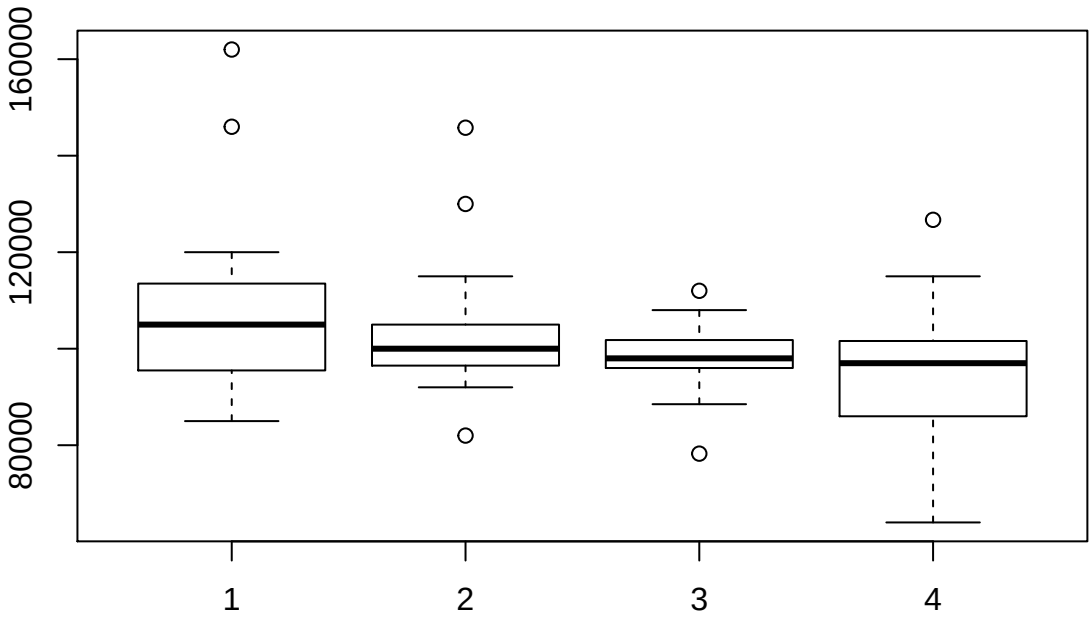
```
## [1] 6
```

Assim, optamos por não utilizar esta variável na análise.

Quartil influencia salário?

O desempenho passado do estudante pode ou não influenciar no salário. O alunos são divididos nos 4 quartis (de acordo com o seu desempenho academico) e seu salário analisado abaixo.

```r
boxplot(mba4$salary ~ mba4$quarter)
```



Vista a diferença entre os quartis, decidimos por continuar com esta variável, quartil, em nossa análise.

Modelos de regressão

Os modelos de regressão serão utilizados para validar ou invalidar diversas hipóteses. As principais perguntas são se uma variável impacta ou não no salário do profissional. Assim, via de regra, teremos como hipótese Nula que esta variável não impacta no salário, e como hipótese alternativa que dada variável impacta no salário. Abaixo analisamos as variáveis que temos para o caso em estudo.

Modelo 1 (Salário ~ Quartil)

Teste de hipótese: - H_0: O quartil de performance do aluno não influencia o seu salário. - H_A: O quartil de performance do aluno influência o salário.

```r
modelo1 <- lm(mba4, formula = salary ~ quarter)
summary(modelo1)
```

```
##
## Call:
## lm(formula = salary ~ quarter, data = mba4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31314  -6738  -1058   4986  55454
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   110290       2920  37.764  < 2e-16 ***
## quarter        -3744       1167  -3.209  0.00179 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13040 on 100 degrees of freedom
## Multiple R-squared:  0.09335,    Adjusted R-squared:  0.08428
## F-statistic:  10.3 on 1 and 100 DF,  p-value: 0.001792
```

Como o **p-value** para esta análise possui valor **abaixo de 0,05**, optamos por **manter** a variável quartil como preditor no modelo candidato.

Modelo 2 (Salário ~ Anos de Experiência)

Teste de hipótese: - H_0: Os anos de experiência não influenciam no salário. - H_A: Os anos de experiência influenciam no salário.

```
modelo2 <- lm(mba4, formula = salary ~ work_yrs)
summary(modelo2)
```

```
##
## Call:
## lm(formula = salary ~ work_yrs, data = mba4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35600  -7055  -1600   5194  63856
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  96687.8     2099.3  46.058   <2e-16 ***
## work_yrs      1456.1      463.3   3.143   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13060 on 100 degrees of freedom
## Multiple R-squared:  0.08989,    Adjusted R-squared:  0.08079
## F-statistic: 9.877 on 1 and 100 DF,  p-value: 0.002203
```

Como o **p-value** para esta análise possui valor abaixo de **0,05**, optamos por **manter** a variável **anos de experiência** como preditor no modelo candidato.

Modelo 3 (Salário ~ Idade)

- H_0: A idade não influencia no salário.
- H_A: A idade influencia no salário.

```
modelo3 <- lm(mba4, formula = salary ~ age)
summary(modelo3)
```

```
##
## Call:
## lm(formula = salary ~ age, data = mba4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33784  -6818  -1333   4118  62667
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60613.5    11404.6   5.315  6.5e-07 ***
## age           1548.8      425.3   3.642 0.000431 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12870 on 100 degrees of freedom
## Multiple R-squared:  0.1171, Adjusted R-squared:  0.1083
## F-statistic: 13.26 on 1 and 100 DF,  p-value: 0.0004314
```

Como o **p-value** para esta análise possui valor abaixo de **0,05**, optamos por **manter** a variável **anos de experiência** como preditor no modelo candidato.

Modelo 4 (Salário ~ Anos de experiência e Idade)

Dado que entendemos que a idade e a experiência do aluno podem estar potencialmente relacionados, analisamos a influência de ambas as variáveis em conjunto.

- H_0: A idade e os anos de experiência não influenciam no salário.
- H_A: A idade e os anos de experiência influenciam no salário.

```
modelo4 <- lm(mba4, formula = salary ~ work_yrs + age)
summary(modelo4)
```

```
##
## Call:
## lm(formula = salary ~ work_yrs + age, data = mba4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33842  -6734  -1325   4202  62797
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62469.9    19661.1   3.177  0.00198 **
## work_yrs       104.4      898.2   0.116  0.90773
## age           1465.1      837.1   1.750  0.08319 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12930 on 99 degrees of freedom
## Multiple R-squared:  0.1172, Adjusted R-squared:  0.09937
## F-statistic: 6.572 on 2 and 99 DF,  p-value: 0.00209
```

Vemos que, quando analisadas em conjunto, os anos de experiência possuem um valor p de 0,91. A colinearidade entre idade e anos de experiência deve ser excluída do modelo. Porém, como individualmente os anos de experiência são relevantes, e dado que a idade é um fator aparentemente refletido nos anos de experiência, optamos por **não utilizar a idade** no possível modelo.

Modelo 5 (Salário ~ Sexo + Anos de experiência + quartil de performance)

Buscamos enriquecer o modelo preditivo com 3 variáveis, sexo, anos de experiência e quartil de performance.

- H_0: A svariáveis em conjunto não influenciam no salário.
- H_A: As variáveis em conjunto influenciam no salário.

```r
modelo5 <- lm(mba4, formula = salary ~ sex + work_yrs + quarter)
summary(modelo5)
```

```
##
## Call:
## lm(formula = salary ~ sex + work_yrs + quarter, data = mba4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23336  -7639  -2103   5673  55214
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 119309.8     5123.3  23.288  < 2e-16 ***
## sex          -9973.4     2627.8  -3.795 0.000255 ***
## work_yrs       912.2      437.6   2.085 0.039715 *
## quarter      -3462.8     1090.7  -3.175 0.002004 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11890 on 98 degrees of freedom
## Multiple R-squared:  0.2609, Adjusted R-squared:  0.2382
## F-statistic: 11.53 on 3 and 98 DF,  p-value: 1.543e-06
```

Podemos verificar **p-value** abaixo de **0,05**, com aumento da qualidade preditiva. Assim manteremos como modelo candidato até o momento o modelo5.

Modelo 6 (Salário ~ Sexo + Anos de experiência + quartil de performance + GMAT Total)

Buscamos enriquecer o modelo preditivo, que além das 3 variáveis contidas no modelo 5, pode também adicionar o GMAT do estudante.

```r
modelo6 <- lm(mba4, formula = salary ~ sex + work_yrs + quarter + gmat_tot)
summary(modelo6)
```

```
##
## Call:
## lm(formula = salary ~ sex + work_yrs + quarter + gmat_tot, data = mba4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22318  -7187  -2013   6382  53744
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107057.65   15950.26   6.712  1.3e-09 ***
## sex          -9977.52    2632.42  -3.790 0.000261 ***
## work_yrs       932.41     439.09   2.124 0.036253 *
## quarter      -3388.74    1096.41  -3.091 0.002606 **
## gmat_tot        19.47      24.00   0.811 0.419184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11910 on 97 degrees of freedom
```

```
## Multiple R-squared:  0.2658, Adjusted R-squared:  0.2356
## F-statistic: 8.781 on 4 and 97 DF,  p-value: 4.301e-06
```

Verifica-se que a inclusão do **GMAT não é relevante**, possuindo um **p-value** de **0,42**.

Conclusão

Concluimos que dentre as variáveis estudadas, temos abaixo que:

Influenciam no resultado:

- Sexo
- Quartil
- Idade

Não influenciam no modelo

- GMAT

Colinearidade.

- Idade e Anos de Experiência, eliminado Idade.

Não conclusivos, devido a baixa variação na amostragem.

- Lingua materna

Valor experado

De posso do modelo final (**Modelo5**), o valor esperado pela estudante, ao final de seu MBA média esperado é de **103197 USD**. Em um intervalo de **confiança de 95%**, o salário da candidata é entre **96873 USD e 109521 USD**.

```
student <- data.frame(
  sex = 2,
  work_yrs = 8,
  quarter = 1
)
predict(modelo5, newdata = student, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 103197.8 96873.98 109521.7
```