Método de regressão linear aplicado ao dataset *Students Performance in Exams*

João Cunha (202103227)

Roberto Pinto (202104006)

1. Introdução

O sucesso escolar no ensino secundário é um tema com bastante relevância nos dias de hoje. Sendo um requisito praticamente obrigatória nos países da Europa Ocidental, este ensino situa-se entre o ensino universitário e a escolaridade básica. Sendo que, os sistemas educativos atuais são muito importantes para os jovens de todo o mundo, tendo um impacto crucial na vida dos estudantes, através da qual são feitas escolhas importantíssimas para o seu futuro pessoal e profissional.

Embora o estudo das variáveis associadas ao desempenho escolar tenha sido, historicamente, uma preocupação mundial, como é referido na publicação de Coleman (1966) (1), o papel central das variáveis socioeconômicas e a relevância das práticas e políticas escolares, deu início a uma pesquisa linha cuja relevância se estende por mais de cinco décadas, sendo ainda muito relevante nos dias de hoje. Apesar de, existirem muitas fontes diferentes de dados para a realização deste tipo de estudos com variáveis relacionadas ao desempenho dos alunos, as avaliações em larga escala demonstraram ser uma fonte valiosa, devido ao grande volume de variáveis e observações que oferecem aos investigadores. (2)

Para a avaliação destes fatores condicionam os resultados de grande maioria dos estudantes é possível recorrer a vários tipos de análises. Neste trabalho recorremos à utilização da Regressão Linear, avaliando 8 variáveis características de todo o tipo de alunos, mais concretamente: Género, Etnicidade, Educação Parental, Almoço, Curso de preparação para o teste, Notas de matemática, Pontuação de leitura e Pontuação de escrita. Esta mesma análise permitiu-nos retirar conclusões muito relevantes em relação à dependência e efeito destas variáveis no valor das “Notas de Matemática”, tomando-a como o *target* deste modelo estatístico.

2. Descrição do dataset *Students Performance in Exams*

Este conjunto de dados consiste numa amostra de 1000 alunos e pretende perceber o desempenho dos estudantes em várias disciplinas tendo em conta vários aspetos, como a preparação de cada aluno para um teste e a situação académica familiar.

As variáveis descritas neste dataset são:

- Género: Masculino ou feminino
- Etnicidade: 5 grupos (A, B, C, D, E)
- Educação Parental: 6 grupos (Algum Secundário, Ensino Secundário, Diploma Associado, Alguma Faculdade, Licenciatura, Mestrado)
- Almoço: 2 grupos (Grátis/Reduzido, Padrão)
- Curso de preparação para o teste: 2 grupos (Completo, Nenhum)
- Notas de matemática: de 0 a 100
- Pontuação de leitura: de 0 a 100
- Pontuação de escrita: de 0 a 100

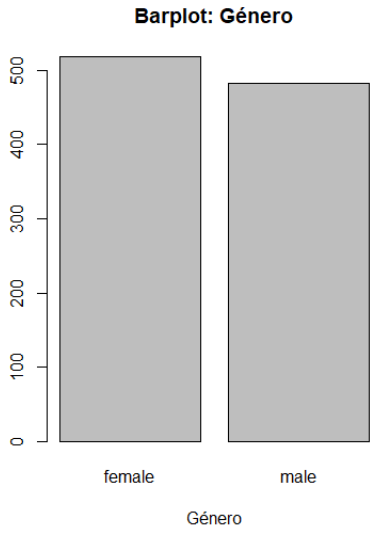
Assim sendo, este trabalho tem como finalidade perceber o impacto que todas as variáveis descritas têm nas notas de matemática e conseguir um modelo de regressão linear que demonstre a relação entre as várias variáveis.

	Gender	Race	Parent_Education	Lunch	Test_Prep	Math_Score	Reading_Score	Writing_Score
1	female	group B	bachelor's degree	standard	none	72	72	74
2	female	group C	some college	standard	completed	69	90	88
3	female	group B	master's degree	standard	none	90	95	93
4	male	group A	associate's degree	free/reduced	none	47	57	44
5	male	group C	some college	standard	none	76	78	75
6	female	group B	associate's degree	standard	none	71	83	78
7	female	group B	some college	standard	completed	88	95	92
8	male	group B	some college	free/reduced	none	40	43	39
9	male	group D	high school	free/reduced	completed	64	64	67
10	female	group B	high school	free/reduced	none	38	60	50
11	male	group C	associate's degree	standard	none	58	54	52
12	male	group D	associate's degree	standard	none	40	52	43
13	female	group B	high school	standard	none	65	81	73
14	male	group A	some college	standard	completed	78	72	70
15	female	group A	master's degree	standard	none	50	53	58
16	female	group C	some high school	standard	none	69	75	78
17	male	group C	high school	standard	none	88	89	86
18	female	group B	some high school	free/reduced	none	18	32	28
19	male	group C	master's degree	free/reduced	completed	46	42	46
20	female	group C	associate's degree	free/reduced	none	54	58	61
21	male	group D	high school	standard	none	66	69	63
22	female	group B	some college	free/reduced	completed	65	75	70
23	male	group D	some college	standard	none	44	54	53
24	female	group C	some high school	standard	none	69	73	73
25	male	group D	bachelor's degree	free/reduced	completed	74	71	80
26	male	group A	master's degree	free/reduced	none	73	74	72

Figura 1:Dataset Students Performance in Exams

3. Descrição das variáveis explicativas

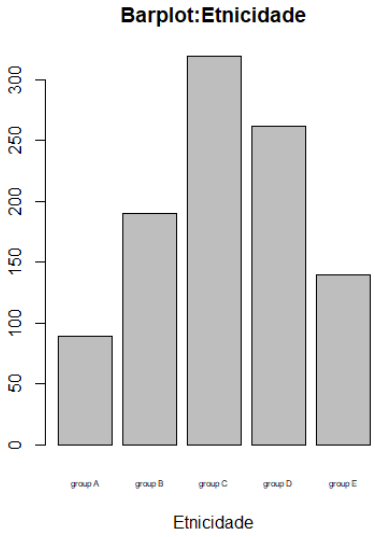
3.1. Género



Frequências absolutas e relativas	
Feminino	**Masculino**
518 (52%)	482 (48 %)

Tabela 1:Descrição do Género

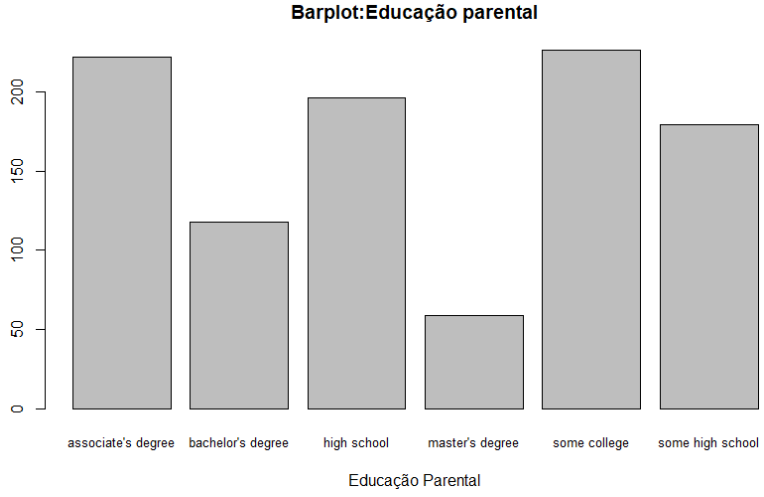
3.2. Etnicidade



Frequências absolutas e relativas				
A	**B**	**C**	**D**	**E**
89 (9%)	190 (19%)	319 (32%)	262 (26%)	140 (14%)

Tabela 2:Descrição da Etnicidade

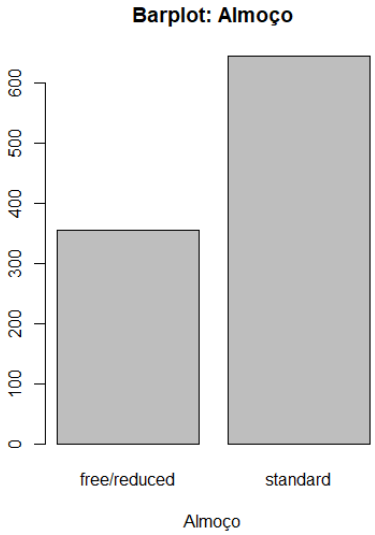
3.3. Educação Parental



Frequências absolutas e relativas					
Diploma Associado	**Licenciatura**	**Ensino Secundário**	**Mestrado**	**Alguma Faculdade**	**Algum Secundário**
222 (22%)	118 (12%)	196 (20%)	59 (6%)	226 (22%)	179 (18%)

Tabela 3:Descrição da Educação Parental

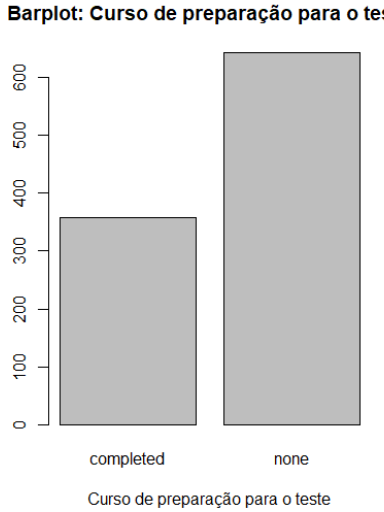
3.4. Almoço



Frequências absolutas e relativas	
Grátis/Reduzido	**Padrão**
355 (36%)	645 (64 %)

Tabela 4:Descrição do Almoço

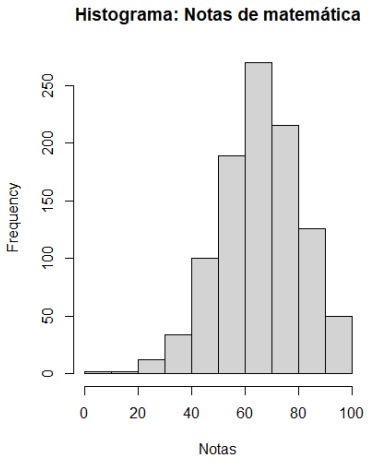
3.5. Curso de Preparação para o teste

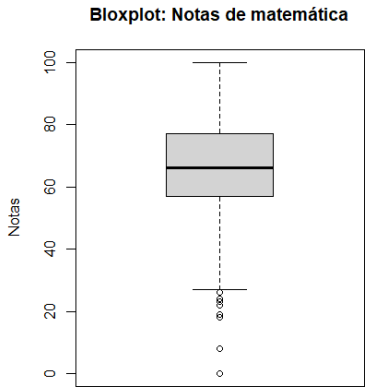


Frequências absolutas e relativas	
Completo	**Nenhum**
358 (36%)	642 (64%)

Tabela 5:Descrição do Curso de Preparação para o Teste

3.6. Notas de Matemática

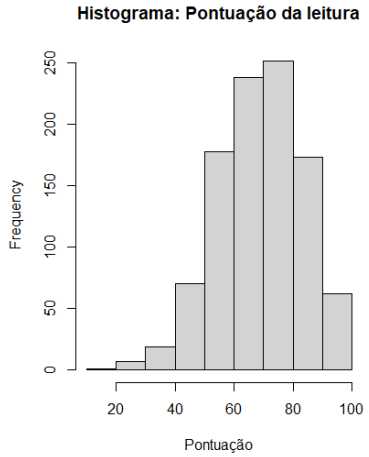


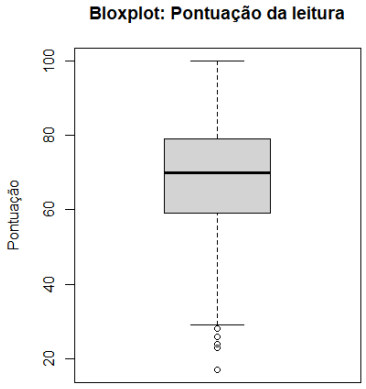


Mínimo: 0.00
Máximo: 100.00
Mediana: 66.00
Média: 66.089
Desvio padrão: 15.16308

Tabela 6:Descrição das Notas de Matemática

3.7. Pontuação de Leitura

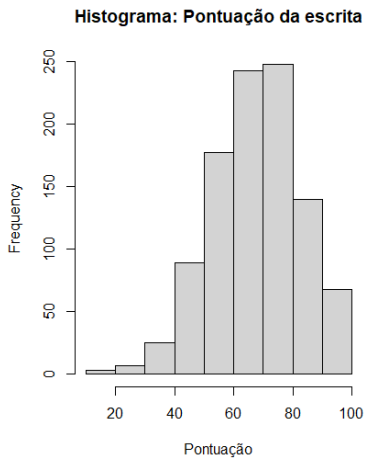


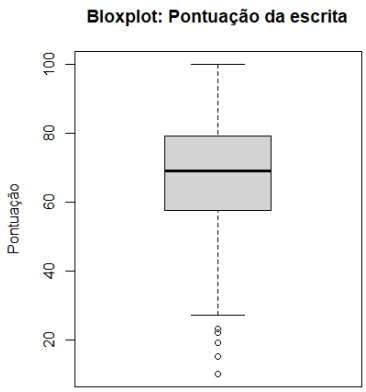


Mínimo: 0.00
Máximo: 100.00
Mediana: 70.00
Média: 69.169
Desvio padrão: 14.60019

Tabela 7:Descrição da Pontuação de Leitura

3.8. Pontuação de Escrita





Mínimo: 0.00
Máximo: 100.00
Mediana: 69.00
Média: 68.054
Desvio padrão: 15.19566

Tabela 8:Descrição da Pontuação de Escrita

4. Correlação entre variáveis

Para realizar a correlação entre todas as variáveis desta amostra, ou seja 8 variáveis (Género, Etnicidade, Educação Parental, Almoço, Curso de preparação para o teste, Notas de matemática, Pontuação de leitura, Pontuação de escrita), foi necessário transformar as 5 variáveis categóricas em variáveis numéricas, com o recurso à função *as.numeric().*

```
#Transformar variáveis categóricas em numéricas
#1:Rapariga, 2: Rapaz
data$gender1 <- as.numeric(factor(data$Gender))
# GroupA, GroupB, GroupC, GroupD =(1,2,3,4)
data$Race1<- as.numeric(factor(data$Race))
# parental level of education (1,2,3,4,5,6)
#('some high school','high school',"associate's degree",'some college',"bachelor's de
data$Parent_Education1 <-as.numeric(factor(data$Parent_Education,levels=c('some high s
#lunch
data$Lunch1 <- as.numeric(factor(data$Lunch, levels=c("standard",'free/reduced')))
#preparation course
data$Test_Prep1 <- as.numeric(factor(data$Test_Prep, levels =c('none','completed') ))
data1 <- data[,6:13]
```

Apenas após este passo foi possível realizar a correlação entre as variáveis.

```
cor(data1)
corrplot(cor(data1), method="color", addCoef.col = "grey", type="upper", diag=FALSE, tl.col="black")
```

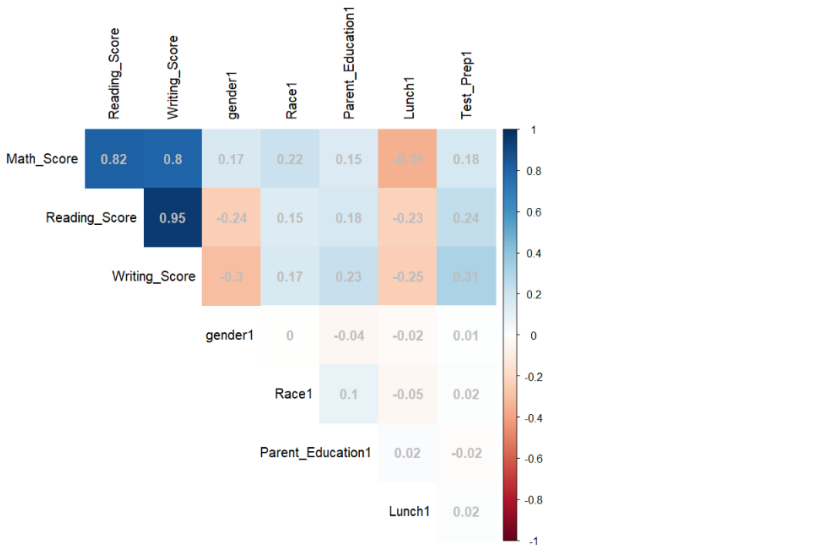


Figura 2: Matriz de Correlação entre as variáveis

Através da observação da matriz de correlação é possível perceber que as variáveis com maior correlação entre si são: Notas de matemática, Pontuação da leitura e Pontuação da escrita.

4.1. Gráficos de dispersão entre as variáveis com maior correlação

Um gráfico de dispersão mostra a relação entre duas variáveis quantitativas medidas para os mesmos indivíduos. Os valores de uma variável aparecem no eixo

horizontal e os valores da outra variável aparecem no eixo vertical. Cada indivíduo nos dados aparece como um ponto no gráfico. (3)

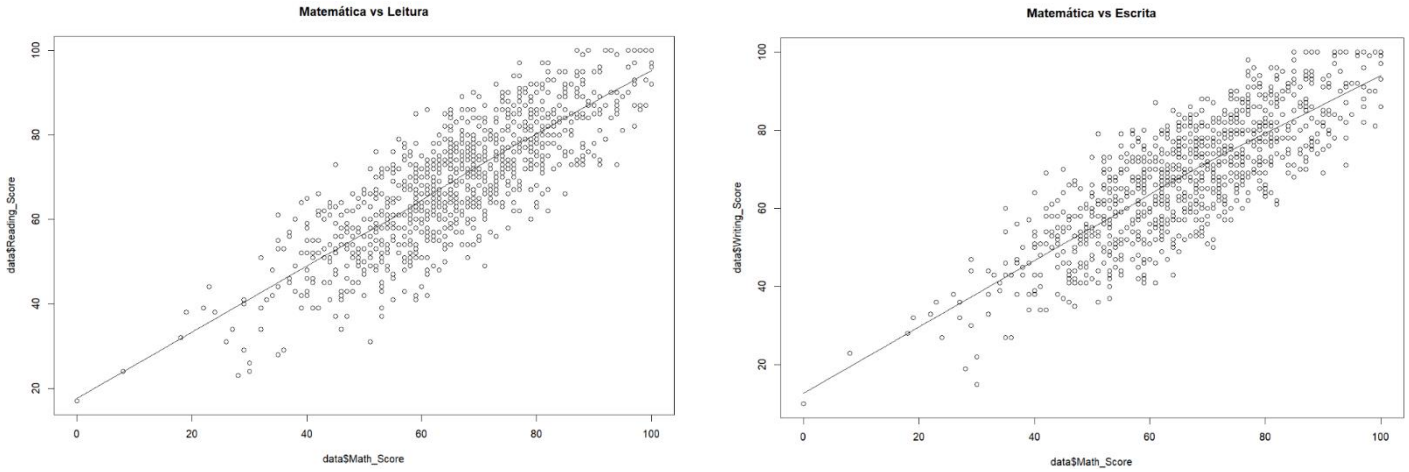


Figura 3:Gráficos de Dispersão entre as 3 variáveis

Em ambos os casos as duas variáveis têm uma associação positiva visto que tanto os valores acima da média como os valores abaixo da média tendem a ocorrer "juntos", seguem uma tendência semelhante.

Além disso demonstram uma relação linear, devido à forma do gráfico. Isso significa que os pontos no gráfico de dispersão são semelhantes a uma linha reta. Uma relação é linear se uma variável aumenta aproximadamente na mesma taxa que as outras variáveis mudam gradualmente.

4.2. Gráficos de Densidade

Realizamos também os gráficos de densidade para estas três variáveis, evidenciando ainda mais a sua relação e linearidade.

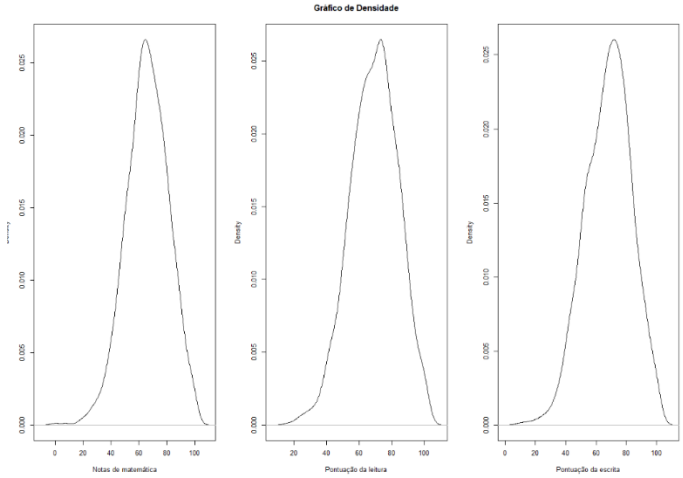


Figura 4:Gráficos de Densidade entre as 3 variáveis

É possível perceber, através destes gráficos, que a densidade máxima de alunos fica entre 50 e 90 pontos.

5. Discussão do modelo final selecionado

O conjunto de dados selecionado para este trabalho contém 8 variáveis, sendo que o y são as notas de matemática (*Math_Score*). Assim, para iniciar a regressão linear múltipla construiu-se um primeiro modelo onde constavam todas as variáveis. Esse modelo foi denominado de *model_a*.

```
#Linear Regression
model_a <- lm(Math_Score ~ Reading_Score + Writing_Score + as.factor(Gender) + as.factor(Race) +
              + as.factor(Lunch) + as.factor(Test_Prep) + as.factor(Parent_Education), data=data)
```

Através das funções *print()* e *summary()* conseguiu-se obter algumas informações sobre o modelo inicial (*model_a*).

```
Call:
lm(formula = Math_Score ~ Reading_Score + Writing_Score + as.factor(Gender) +
    as.factor(Race) + +as.factor(Lunch) + as.factor(Test_Prep) +
    as.factor(Parent_Education), data = data)

Coefficients:
                                (Intercept)                               Reading_Score
                                  -11.6045                                     0.2635
                              Writing_Score                  as.factor(Gender)male
                                     0.7016                                   13.2405
                    as.factor(Race)group B                    as.factor(Race)group C
                                     0.8354                                    0.1782
                    as.factor(Race)group D                    as.factor(Race)group E
                                     0.0984                                    5.0777
                 as.factor(Lunch)standard                 as.factor(Test_Prep)none
                                     3.2127                                    3.5023
as.factor(Parent_Education)bachelor's degree      as.factor(Parent_Education)high school
                                    -1.0469                                    0.5677
  as.factor(Parent_Education)master's degree      as.factor(Parent_Education)some college
                                    -1.8561                                    0.4003
 as.factor(Parent_Education)some high school
                                     0.5522
```

```
Call:
lm(formula = Math_Score ~ Reading_Score + Writing_Score + as.factor(Gender) +
    as.factor(Race) + +as.factor(Lunch) + as.factor(Test_Prep) +
    as.factor(Parent_Education), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4995  -3.6824   0.1218   3.3932  14.1178

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    -11.60449    1.24479  -9.322  < 2e-16 ***
Reading_Score                                    0.26351    0.04205   6.266 5.52e-10 ***
Writing_Score                                    0.70156    0.04352  16.120  < 2e-16 ***
as.factor(Gender)male                           13.24045    0.37193  35.599  < 2e-16 ***
as.factor(Race)group B                           0.83537    0.69230   1.207   0.2279
as.factor(Race)group C                           0.17823    0.64899   0.275   0.7837
as.factor(Race)group D                           0.09840    0.67014   0.147   0.8833
as.factor(Race)group E                           5.07770    0.73714   6.888 1.00e-11 ***
as.factor(Lunch)standard                         3.21271    0.37420   8.585  < 2e-16 ***
as.factor(Test_Prep)none                         3.50227    0.39658   8.831  < 2e-16 ***
as.factor(Parent_Education)bachelor's degree    -1.04690    0.61571  -1.700   0.0894 .
as.factor(Parent_Education)high school           0.56773    0.53518   1.061   0.2890
as.factor(Parent_Education)master's degree      -1.85607    0.79324  -2.340   0.0195 *
as.factor(Parent_Education)some college          0.40026    0.50814   0.788   0.4311
as.factor(Parent_Education)some high school      0.55216    0.54989   1.004   0.3156
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.362 on 985 degrees of freedom
Multiple R-squared:  0.8767,    Adjusted R-squared:  0.875
F-statistic: 500.3 on 14 and 985 DF,  p-value: < 2.2e-16
```

Através dos resultados apresentados anteriormente, consegue-se afirmar que o R quadrado (R^2) para este modelo é de 0,8767. Ainda, é possível verificar o nível de significância de cada variável, através do *p-value* dado pela linha *Signif. Codes*, em que se esse valor for inferior a 0.05, significa que essa variável não é significativa para o modelo e assim pode-se remover essa variável. É importante realçar que nas variáveis categóricas que tenham mais de que uma categoria, basta uma categoria satisfazer a condição *p-value* < 0.05, que essa variável passa a ser significativa para o modelo (4).

Observando então as figuras anteriores é possível verificar que a variável *Parent_Education*, em todas das categorias, tem um *p-value* superior a 0.05, podendo se então remover essa variável.

Assim sendo, procedeu-se á elaboração de um novo modelo, o model_b, onde se removeu a variável *Parent_Education* e que consiste em 6 variáveis (*Reading_Score, Writing_Score, Gender, Race, Lunch, Test_Prep*).

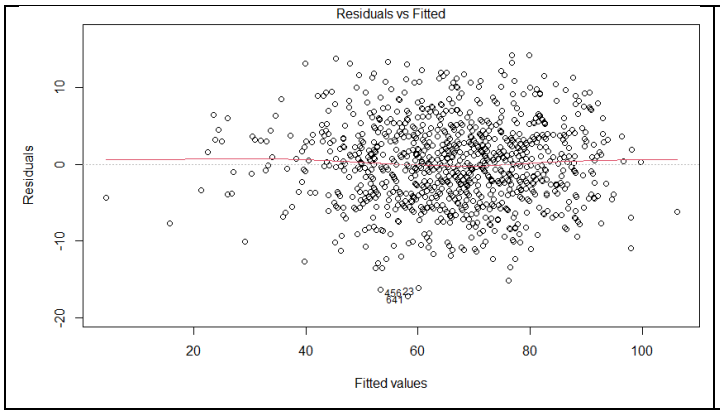
```
model_b <- lm(Math_Score ~ Writing_Score + Reading_Score + as.factor(Gender) + as.factor(Race) +
              +as.factor(Lunch) + as.factor(Test_Prep), data=data)
```

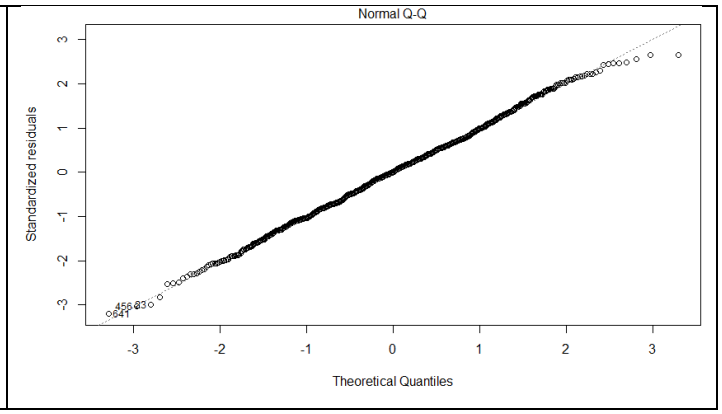
De seguida, e como foi feito para o modelo inicial (model_a), utilizou-se a função *print()* e *summary()* para obter informações relevantes sobre o *model_b*

```
> print(model_b)

Call:
lm(formula = Math_Score ~ Writing_Score + Reading_Score + as.factor(Gender) +
    as.factor(Race) + as.factor(Lunch) + as.factor(Test_Prep),
    data = data)

Coefficients:
                (Intercept)               Writing_Score               Reading_Score     as.factor(Gender)male     as.factor(Race)group B
                   -10.6876                      0.6669                      0.2862                   13.1465                   0.8766
  as.factor(Race)group C    as.factor(Race)group D    as.factor(Race)group E  as.factor(Lunch)standard  as.factor(Test_Prep)none
                    0.1335                    0.1143                    5.0305                    3.3733                    3.3372

> summary(model_b)

Call:
lm(formula = Math_Score ~ Writing_Score + Reading_Score + as.factor(Gender) +
    as.factor(Race) + as.factor(Lunch) + as.factor(Test_Prep),
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.1758  -3.7893   0.0466   3.5036  14.2145

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -10.68756    1.15393  -9.262  < 2e-16 ***
Writing_Score             0.66685    0.04227  15.775  < 2e-16 ***
Reading_Score             0.28620    0.04152   6.893 9.74e-12 ***
as.factor(Gender)male    13.14652    0.37221  35.320  < 2e-16 ***
as.factor(Race)group B    0.87655    0.69445   1.262    0.207
as.factor(Race)group C    0.13351    0.65007   0.205    0.837
as.factor(Race)group D    0.11435    0.67213   0.170    0.865
as.factor(Race)group E    5.03046    0.73774   6.819 1.60e-11 ***
as.factor(Lunch)standard  3.37328    0.37342   9.034  < 2e-16 ***
as.factor(Test_Prep)none  3.33720    0.39335   8.484  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.39 on 990 degrees of freedom
Multiple R-squared:  0.8748,    Adjusted R-squared:  0.8736
F-statistic: 768.5 on 9 and 990 DF,  p-value: < 2.2e-16
```

No output de ambas as funções é possível verificar que o valor de R quadrado desceu ligeiramente (0.8748). No entanto, olhando novamente para os níveis de significância das variáveis, isto é, o *p-value*, percebe-se que todas as variáveis têm um valor inferior a 0.05, podendo concluir, por isso, que este é o modelo final da regressão linear deste conjunto de dados.

Após escolher o modelo é necessário proceder á avaliação da sua qualidade. Isto é possível através da função *anova()* e *plot()*, obtendo-se, assim, uma tabela de análise da variância das variáveis e também 4 gráficos relativos ao *model_b*, que podem ser observados de seguida.





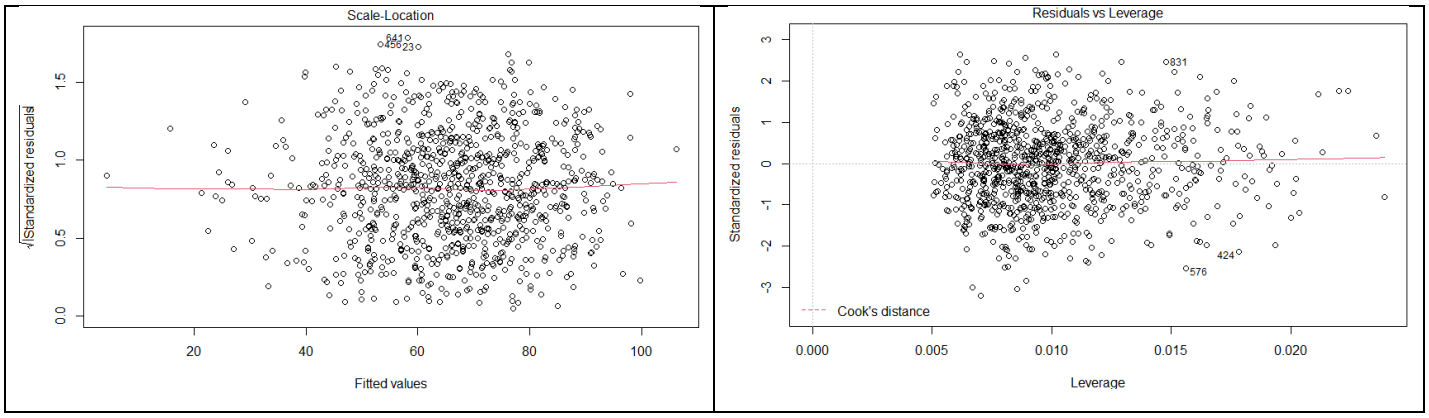


Figura 5:Gráficos do model_b

Observando os gráficos do *model_b* é possível verificar que a normalidade dos resíduos, visto que no gráfico normal Q-Q, estes coincidem todos com uma linha diagonal, que representa a distribuição normal teórica e os pontos a distribuição teórica dos resíduos. Através dos restantes gráficos consegue-se perceber a ausência de um padrão no comportamento dos resíduos em relação ao modelo. Assim, devido á normalidade dos resíduos e á ausência de um comportamento padronizado pode-se inferir que este modelo (model_b) é um modelo homocedástico (5).

6. Apresentação da equação matemática do modelo

G: *Gender*

G0: *female* (referência); G1: *male*

Esta variável será representada por 1 *Dummy*:

G1: 1 – male; 0 - otherwise

L: *Lunch*

L0: *Free/Reduced* (referência); L1: *standard*

Esta variável é representada por 1 *Dummy*:

L1: 1- *standard*; 0 - *otherwise*

TP: *Test_Prep*

TP0: *completed* (referência); TP1: *none*

Esta variável é representada por 1 *Dummy*:

TP1: 1 – *none*; 0 - *otherwise*

R: *Race*

R0: *group* A (referência); R1: *group* B; R2: *group* C; R3: *group* D; R4: *group* E

Esta variável é representada por 4 *Dummys*:

R1: 1 – *group* B; 0 - *otherwise*

R2: 1 – *group* C; 0 - *otherwise*

R3: 1 – *group* D; 0 - *otherwise*

R4: 1 – *group* E; 0 - *otherwise*

$$\text{Math\_Score} = \beta_0 + \beta_1 \text{Reading\_Score} + \beta_2 \text{Writing\_Score} + \beta_3 G1 + \beta_4 R1 + \beta_5 R2 + \beta_6 R3 + \beta_7 R4 + \beta_8 L1 + \beta_9 TP1 + \mu, \mu \sim N(0, \sigma^2)$$

7. Para uma variável continua X1 e uma variável categórica com mais de 2 categorias X2 que constem do modelo final

Para esta parte do trabalho, vai se considerar uma variável contínua, X1, que será a variável *Writing_Score* e uma variável categórica X2, que será a variável *Race*.

7.1. Interpretação do efeito bruto de *Writing_Score* e o efeito ajustado de *Race*.

Para se interpretar o efeito bruto da variável *Writing_Score* e o efeito ajustado de *Race*, tem-se que fazer um modelo de regressão linear simples apenas com essas varáveis (*model1* e *model2*) de modo a obter os efeitos brutos. De seguida, com o modelo selecionado (*model_b*) obtêm-se os efeitos ajustados.

```
model1 <- lm(Math_Score ~ Writing_Score, data=data)

model2 <- lm(Math_Score ~ as.factor(Race), data=data)
```

Através da função *print* foi possível obter os coeficientes dos 3 modelos e assim perceber o efeito bruto e o efeito ajustado de ambas as variáveis, demonstrado na tabela abaixo.

	Efeito Bruto	Efeito ajustado
Writing_Score	0.8009	0.6669
Racegroup B	1.823	0.8766
Racegroup C	2.835	0.1335
Racegroup D	5.733	0.1143
Racegroup E	12.192	5.0305

Interpretação – efeitos brutos:

Olhando primeiro para a variável contínua *Writing_Score*, é possível afirmar que as notas de matemática (*Math_Score*) aumentam 0.801 pontos.

No que respeita a variável categórica *Race*, caso um aluno seja de uma etnia do grupo B, existe um aumento das notas de matemática em 1,82 pontos. Esse aumento é verificado para todos os grupos étnicos, sendo que no grupo C as notas de matemática aumentam 2,84 pontos, no grupo D 5,73 e no grupo E existe um aumento mais

acentuado com 12,19 pontos na nota de matemática. Caso o aluno seja do grupo étnico A, a sua nota de matemática será 61,63 pontos.

Interpretação – efeitos ajustados:

No que concerne a variável *Writing_Score*, mantendo as restantes variáveis *Reading_Score*, *Gender*, *Test_Prep*, *Lunch* e *Race* constantes e após ajustamento dessas variáveis, as notas de matemática aumentam 0,67 pontos.

Na variável categórica *Race*, mantendo as restantes variáveis *Reading_Score*, *Gender*, *Test_Prep*, *Lunch* e *Writing_Score* constantes e após ajustamento dessas variáveis, as notas de matemática podem aumentar 0.88, 0.13, 0.11 ou 5.03, se o aluno pertencer ao grupo étnico B, C, D ou E, respetivamente.

7.2. Determinação gráfica das bandas de confiança e de predição em função dos valores de *Writing_Score*, fixando os restantes preditores contínuos nos seus valores medianos e os categorias nas respetivas modas.

Para proceder à determinação gráfica das bandas de confiança e de predição em função dos valores de Writing_Score, é necessário fixar os preditores contínuos na sua mediana e os preditores categóricos na sua moda criando um novo *dataframe* (*data3*), como é possível verificar na figura a seguir. No entanto, é importante realçar que se procedeu á elaboração de um novo modelo, igual ao modelo selecionado (model_b), de modo a não alterar nada desse modelo.

```
#Criação de modelo igual ao model_b, de modo a não alterar nada do modelo original
model_b2 <- lm(Math_Score ~ Writing_Score + Reading_Score + as.factor(Gender) + as.factor(Race)
               + as.factor(Lunch) + as.factor(Test_Prep), data=data)
summary(model_b2)

#fixação dos preditores continuos e categóricos e criação do dataframe
data3 <- data.frame(Writing_Score = c(69), Reading_Score = c(70), Gender = c('female'),
                    Race = c('group C'), Lunch = c('standard'), Test_Prep = c('none'))
```

Criou De seguida, elaborou-se a predição, através da função *predict()* e de seguida fez-se um gráfico com o resultado desta função, utilizando, posteriormente, a função *ggplot().*

```
#Predição
par(mfrow=c(1, 1))
predict(model_b2, data3, interval='prediction', level = 0.95)
predict(model_b2, data3, interval='confidence', level = 0.95)
#Gráfico da predição
pred <- predict(model_b2, data=data3, interval='prediction', level = 0.95)
library("ggplot2")
data4 <- cbind(data3,pred)
p <- ggplot(data4, aes(Writing_Score, Math_Score)) + geom_point() + stat_smooth(method = lm)
p + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+ geom_line(aes(y = upr), color = "red", linetype = "dashed")
```

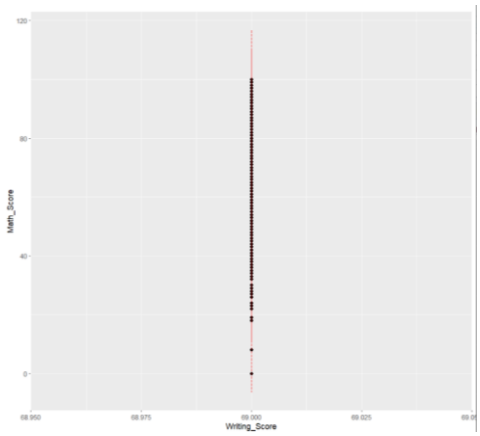


Figura 6:Gráfico da Previsão

7.3. Interpretação o efeito provocado na resposta por uma mudança da terceira categoria de X2 para a segunda, e indique um intervalo de confiança a 95% para esse efeito.

$$\text{Math\_Score (R1)} = \beta_0 + \beta_1 \text{Reading\_Score} + \beta_2 \text{Writing\_Score} + \beta_3 G1 + \beta_4 R1 + \beta_8 L1 + \beta_9 TP1$$

$$\text{Math\_Score (R2)} = \beta_0 + \beta_1 \text{Reading\_Score} + \beta_2 \text{Writing\_Score} + \beta_3 G1 + \beta_5 R2 + \beta_8 L1 + \beta_9 TP1 + \mu, \mu \sim N(0, \sigma^2)$$

$$\text{Math\_Score (R2 - R1)} = \beta_5 \text{R2} - \beta_4 \text{R1}$$

A mudança da terceira categoria de Race para segunda categoria tem um efeito de $\beta_5 - \beta_4$ na resposta. Para verificar a significância do efeito provocado, procedeu-se a um *T-test* para a diferença de médias.

```
> t.test(Math_Score[Race=='group B'], Math_Score[Race=='group C'], mu= 0, conf.level = 0.95)

        Welch Two Sample t-test

data:  Math_Score[Race == "group B"] and Math_Score[Race == "group C"]
t = -0.72407, df = 384.64, p-value = 0.4695
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.757483  1.734846
sample estimates:
mean of x mean of y
 63.45263  64.46395
```

Para um nível de significância de α=0.05, não se pode rejeitar a hipótese nula, visto que o *p-value* é superior a 0,05 (0.4695), podendo afirmar, assim, que a média das notas de matemática do grupo étnico B é igual ás médias das notas de matemática do grupo étnico C.

7.4. Interpretação do efeito provocado por um aumento em X1 correspondente a dois desvios padrão dos seus valores.

Para se conseguir verificar o efeito provocado por um aumento correspondente a dois desvios padrão de *Writing_Score*, tem-se que usar a função *lm.beta()*. Aplicando esta função ao *model_b* obtém-se o seguinte resultado.

```
Call:
lm(formula = Math_Score ~ Writing_Score + Reading_Score + as.factor(Gender) +
    as.factor(Race) + as.factor(Lunch) + as.factor(Test_Prep),
    data = data)

Standardized Coefficients::
                (Intercept)               Writing_Score               Reading_Score
                 0.000000000                 0.668283007                 0.275580211
   as.factor(Gender)male    as.factor(Race)group B    as.factor(Race)group C
                 0.433439989                 0.022689657                 0.004105948
  as.factor(Race)group D    as.factor(Race)group E as.factor(Lunch)standard
                 0.003317693                 0.115173020                 0.106506557
as.factor(Test_Prep)none
                 0.105565384
```

O coeficiente estandardizado de *Writing_Score* é 0.6683, significando que, sempre que a Pontuação da Escrita (*Writing_Score*) aumenta 1 desvio padrão, o modelo prevê um aumento das notas de matemática (*Math_Score*) em, aproximadamente, 0.67 desvios padrão, ajustando para. Sendo assim, mantendo as restantes variáveis, sempre que a Pontuação da Escrita (*Writing_Score*) aumenta 2 desvios padrão o modelo prevê um aumento das notas de matemática (*Math_Score*) em, aproximadamente, 1.34 desvios padrão.

7.5. Averiguação da existência de uma interação significativa entre *Writing_Score* e *Race* e interpretação dos efeitos estimados nessa interação.

Para averiguar a existência de uma interação significativa entre *Writing_Score* e *Race* foi criado um novo modelo, o *model3*.

```
model3 <- lm(Math_Score ~ Writing_Score + Reading_Score + as.factor(Gender) + as.factor(Race) +
               as.factor(Lunch) + as.factor(Test_Prep) + Writing_Score:as.factor(Race), data=data)
```

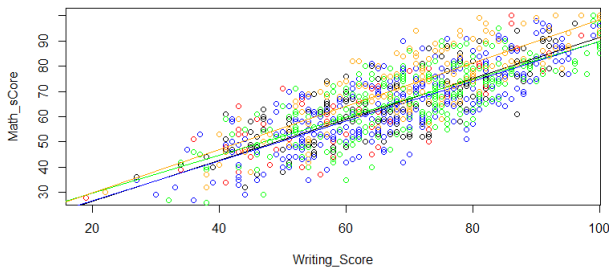
Após se criar o modelo, e como se tem feito para os outros modelos, tem se procedido á verificação do modelo através das funções *print()* e *summary()*.

```
Call:
lm(formula = Math_Score ~ Writing_Score + Reading_Score + as.factor(Gender) +
    as.factor(Race) + as.factor(Lunch) + as.factor(Test_Prep) +
    Writing_Score:as.factor(Race), data = data)

Coefficients:
                           (Intercept)                           Writing_Score
                              -6.61222                                 0.60080
                         Reading_Score                  as.factor(Gender)male
                               0.28737                                13.13151
                as.factor(Race)group B                as.factor(Race)group C
                              -4.22027                                -3.84330
                as.factor(Race)group D                as.factor(Race)group E
                              -2.81408                                -2.34278
             as.factor(Lunch)standard              as.factor(Test_Prep)none
                               3.40641                                 3.30314
Writing_Score:as.factor(Race)group B  Writing_Score:as.factor(Race)group C
                               0.08054                                 0.06351
Writing_Score:as.factor(Race)group D  Writing_Score:as.factor(Race)group E
                               0.04867                                 0.11109
```

```
> summary(model3)

Call:
lm(formula = Math_Score ~ Writing_Score + Reading_Score + as.factor(Gender) +
    as.factor(Race) + as.factor(Lunch) + as.factor(Test_Prep) +
    Writing_Score:as.factor(Race), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4814  -3.7122   0.1361   3.4134  14.2300

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -6.61222    2.50114  -2.644  0.00833 **
Writing_Score                           0.60080    0.05644  10.645  < 2e-16 ***
Reading_Score                           0.28737    0.04160   6.908 8.81e-12 ***
as.factor(Gender)male                  13.13151    0.37192  35.308  < 2e-16 ***
as.factor(Race)group B                 -4.22027    2.93384  -1.438  0.15061
as.factor(Race)group C                 -3.84330    2.77891  -1.383  0.16697
as.factor(Race)group D                 -2.81408    2.91534  -0.965  0.33465
as.factor(Race)group E                 -2.34278    3.25930  -0.719  0.47244
as.factor(Lunch)standard                3.40641    0.37377   9.114  < 2e-16 ***
as.factor(Test_Prep)none                3.30314    0.39319   8.401  < 2e-16 ***
Writing_Score:as.factor(Race)group B    0.08054    0.04482   1.797  0.07267 .
Writing_Score:as.factor(Race)group C    0.06351    0.04232   1.501  0.13376
Writing_Score:as.factor(Race)group D    0.04867    0.04382   1.111  0.26695
Writing_Score:as.factor(Race)group E    0.11109    0.04790   2.319  0.02060 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.384 on 986 degrees of freedom
Multiple R-squared:  0.8756,    Adjusted R-squared:  0.8739
F-statistic: 533.7 on 13 and 986 DF,  p-value: < 2.2e-16
```

Para estudar a possível interação entre as variáveis *Writing_Score* e *Race* foi feito um gráfico.

```
plot(Writing_Score[Race=='group A'], Math_Score[Race=='group A'], col='red',xlab='Writing_Score', ylab = 'Math_Score')
points(Writing_Score[Race=='group B'], Math_Score[Race=='group B'], col='black')
points(Writing_Score[Race=='group C'], Math_Score[Race=='group C'], col='blue')
points(Writing_Score[Race=='group D'], Math_Score[Race=='group D'], col='green')
points(Writing_Score[Race=='group E'], Math_Score[Race=='group E'], col='orange')

lm(Math_Score[Race=='group A'] ~ Writing_Score[Race=='group A'])
lm(Math_Score[Race=='group B'] ~ Writing_Score[Race=='group B'])
lm(Math_Score[Race=='group C'] ~ Writing_Score[Race=='group C'])
lm(Math_Score[Race=='group D'] ~ Writing_Score[Race=='group D'])
lm(Math_Score[Race=='group E'] ~ Writing_Score[Race=='group E'])

abline(14.5455, 0.7513, col='red')
abline(10.379, 0.809, col='black')
abline(10.9873, 0.7884,col='blue')
abline(14.8106, 0.7492,col='green')
abline(12.7809, 0.8548, col = 'orange')
```



O gráfico apresentado acima, sugere que poderá haver uma interação entre as variáveis *Writing_Score* e Race. No entanto, os termos de interação do modelo não são significativos.

Interpretação:

A Pontuação na Escrita (Writing_Score) faz aumentar 0.601 pontos nas notas de matemática, dependendo do grupo étnico a que o aluno pertence.

8. Referências Bibliográficas

1. COLEMAN JS |AN. O. EQUALITY OF EDUCATIONAL OPPORTUNITY. 1966.
2. Sing Chai C, Van Den Noortgate W, Leuven Kulak K, Bo Ning B, Martínez-Abad F, Gamazo A. Citation: Gamazo A and Martínez-Abad F (2020) An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques. Front Psychol. 2020;11:575167.
3. Marshall E. Scatterplots and correlation in SPSS SPSS. 2018.
4. Long J, Teetor P. R Cookbook [Internet]. 2019. Available from: https://rc2e.com/index.html
5. Helena M. Tutorial — Ajuste e Interpretação de Regressão Linear com R [Internet]. 2019. Available from: https://medium.com/data-hackers/tutorial-ajuste-e-interpretação-de-regressão-linear-com-r-5b23c4ddb72