Trabalho de Regressão Linear

Victor Hugo Constantino Lozano - 8910903

2022-11-05

Introdução

O objetivo deste projeto é primeiro analisar os fatores que influenciam os custos médicos explorando o dataset e todos os seus componentes para descobrir correlações entre os dados e, em segundo lugar, tentar construir um modelo adequado que possa prever com precisão os custos de seguro com base nos dados e otimizar seu desempenho.

Dataset O dataset utilizado neste trabalho foi encontrado na plataforma Kaggle e foi originalmente publicado junto ao livro *Machine Learning with R* por Brett Lantz como material de apoio ao conteúdo. O dataset tem 7 variáveis sendo elas:

- Idade do tipo numérico inteiro;
- Sexo do tipo caractere (mais tarde convertida para o tipo booleano);
- IMC do tipo numérico com casas decimais;
- Filhos do tipo numérico inteiro;
- Fumante do tipo caractere (mais tarde convertida para o tipo booleano);
- Região do tipo caractere (Não utilizada nas análises);
- Custos do tipo numerico inteiro.

Regressão Linear

A análise de regressão linear é usada para prever o valor de uma variável com base no valor de outra. A variável que deseja prever é chamada de variável dependente. A variável que é usada para prever o valor de outra variável é chamada de variável independente.

Essa forma de análise estima os coeficientes da equação linear, envolvendo uma ou mais variáveis independentes que melhor preveem o valor da variável dependente. A regressão linear se ajusta a uma linha reta ou superficial que minimiza as discrepâncias entre os valores de saída previstos e reais.

Modelos de regressão linear são relativamente simples e fornecem uma fórmula matemática fácil de interpretar que pode gerar previsões. A regressão linear pode ser aplicada a diversas áreas de estudo empresarial e acadêmico.

Métodos

```r
if(!require(pacman)) install.packages("pacman")
```

```
## Carregando pacotes exigidos: pacman
```

```
## Warning: package 'pacman' was built under R version 4.2.2
```

```r
library(pacman)
pacman::p_load(tidyverse, car, rstatix, lmtest, ggpubr, QuantPsyc, psych, scatterplot3d, outliers)
```

Lendo os dados do arquivo CSV

```r
insurance=read.csv("insurance.csv", header=TRUE, sep=",")
```

Checando o cabeçalho

```r
head(insurance)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Mostrando um resumo dos dados

```r
summary(insurance)
```

```
##       age            sex                 bmi           children     
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000  
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095  
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000  
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000  
##     smoker             region             charges     
##  Length:1338        Length:1338        Min.   : 1122  
##  Class :character   Class :character   1st Qu.: 4740  
##  Mode  :character   Mode  :character   Median : 9382  
##                                        Mean   :13270  
##                                        3rd Qu.:16640  
##                                        Max.   :63770
```

Procura se existe algum dado em branco

```r
any(is.na(insurance))
```

```
## [1] FALSE
```

Alguns dados do dataset estão em formato de strings (variáveis sexo e fumante), por isso dificultam a analise, então vamos transforma-los em dados booleanos (0 e 1) pois são informações binarias.

Transformaremos as variáveis nos seguintes equivalentes:

- Fumante, 0=não, 1=sim
- Sexo 0=Feminino, 1= Masculino

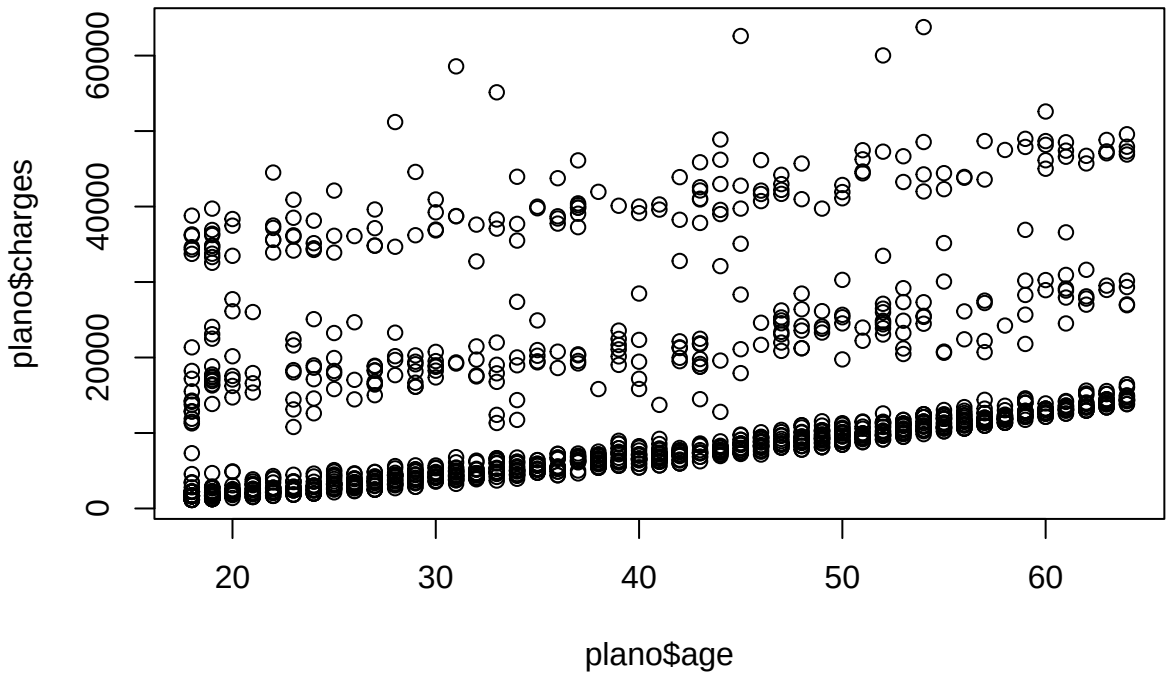
```
plano = insurance %>%
  mutate(smoker = if_else(smoker=="no", 0,1)) %>%
  mutate(sex = if_else(sex == "male", 1,0))

head(plano)
```

```
##   age sex    bmi children smoker    region   charges
## 1  19   0 27.900        0      1 southwest 16884.924
## 2  18   1 33.770        1      0 southeast  1725.552
## 3  28   1 33.000        3      0 southeast  4449.462
## 4  33   1 22.705        0      0 northwest 21984.471
## 5  32   1 28.880        0      0 northwest  3866.855
## 6  31   0 25.740        0      0 southeast  3756.622
```

```
summary(plano)
```

```
##       age             sex              bmi           children
##  Min.   :18.00   Min.   :0.0000   Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   1st Qu.:0.0000   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Median :1.0000   Median :30.40   Median :1.000
##  Mean   :39.21   Mean   :0.5052   Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00   Max.   :1.0000   Max.   :53.13   Max.   :5.000
##      smoker          region             charges
##  Min.   :0.0000   Length:1338        Min.   : 1122
##  1st Qu.:0.0000   Class :character   1st Qu.: 4740
##  Median :0.0000   Mode  :character   Median : 9382
##  Mean   :0.2048                      Mean   :13270
##  3rd Qu.:0.0000                      3rd Qu.:16640
##  Max.   :1.0000                      Max.   :63770
```

```
plot(plano$age, plano$charges)
```



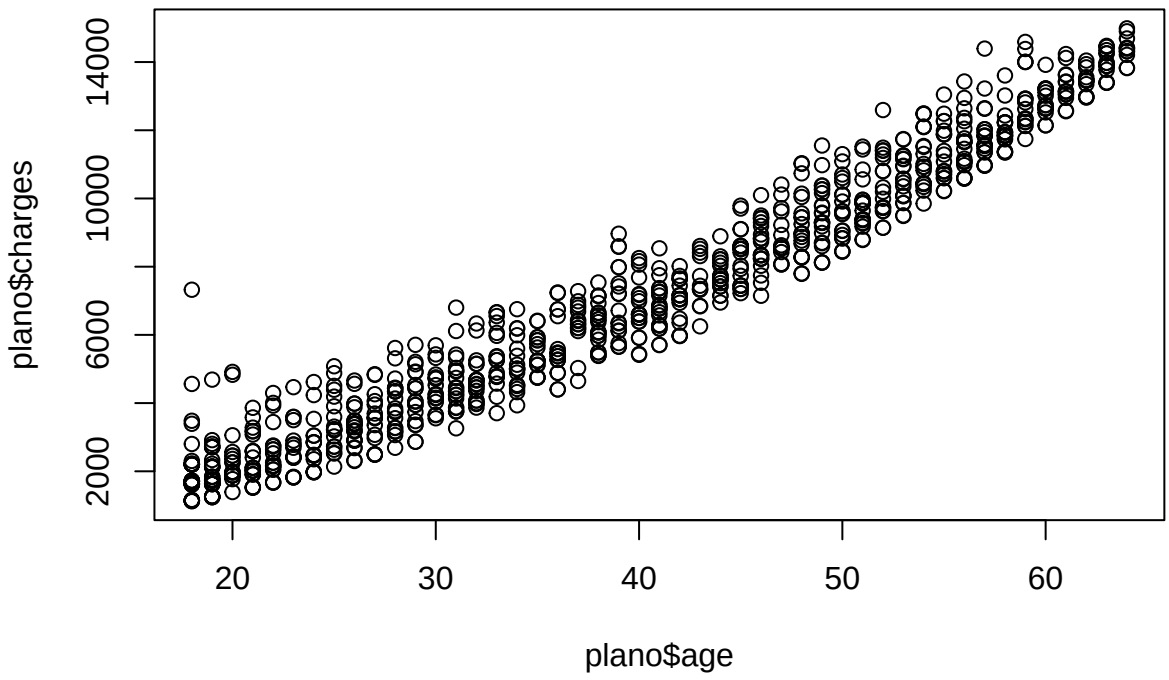
Podemos ver que existem muitos dados dispersos, os outliers, então vamos reduzir o dataset para remover os outliers.

```
plano=plano[plano$charges<=15000, ]

plano$charges = ifelse(plano$charges>9000 &plano$age<45, plano$charges==NA ,plano$charges)
```

Neste trabalho usaremos a viariável “charges” como Variável Dependente (VD).

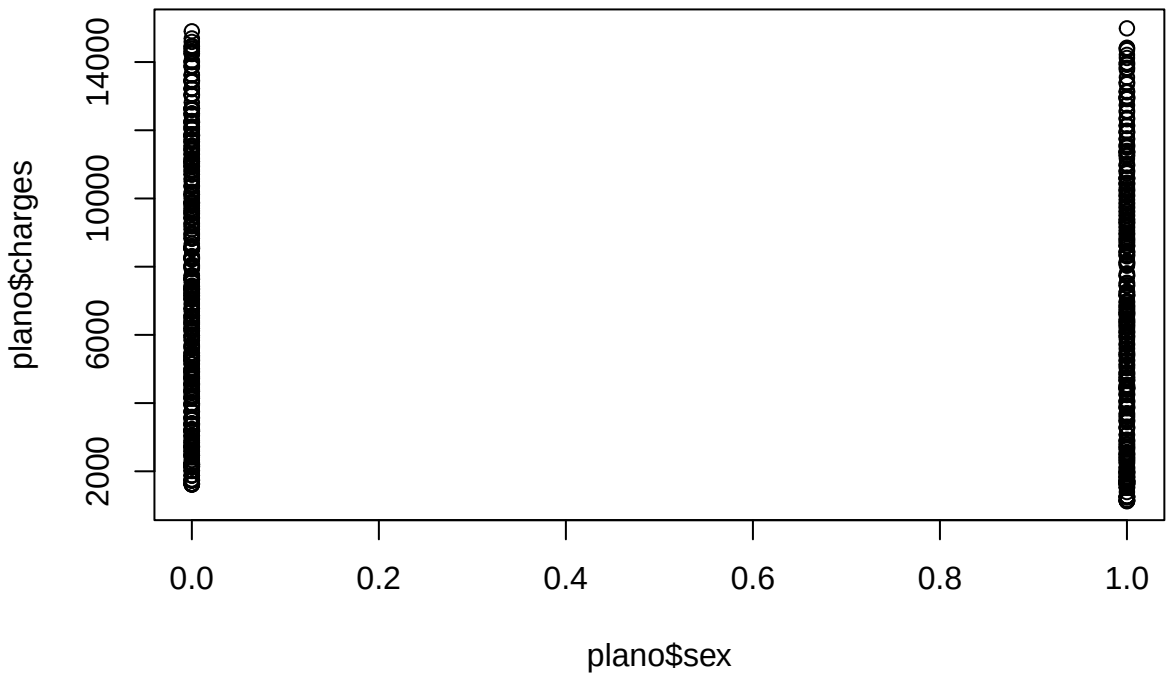
Sem fazer nenhum calculo estatístico podemos plotar a VD contra as outras, com a finalidade de observar se existe alguma relação entre elas.

```
plot(plano$age, plano$charges)
```



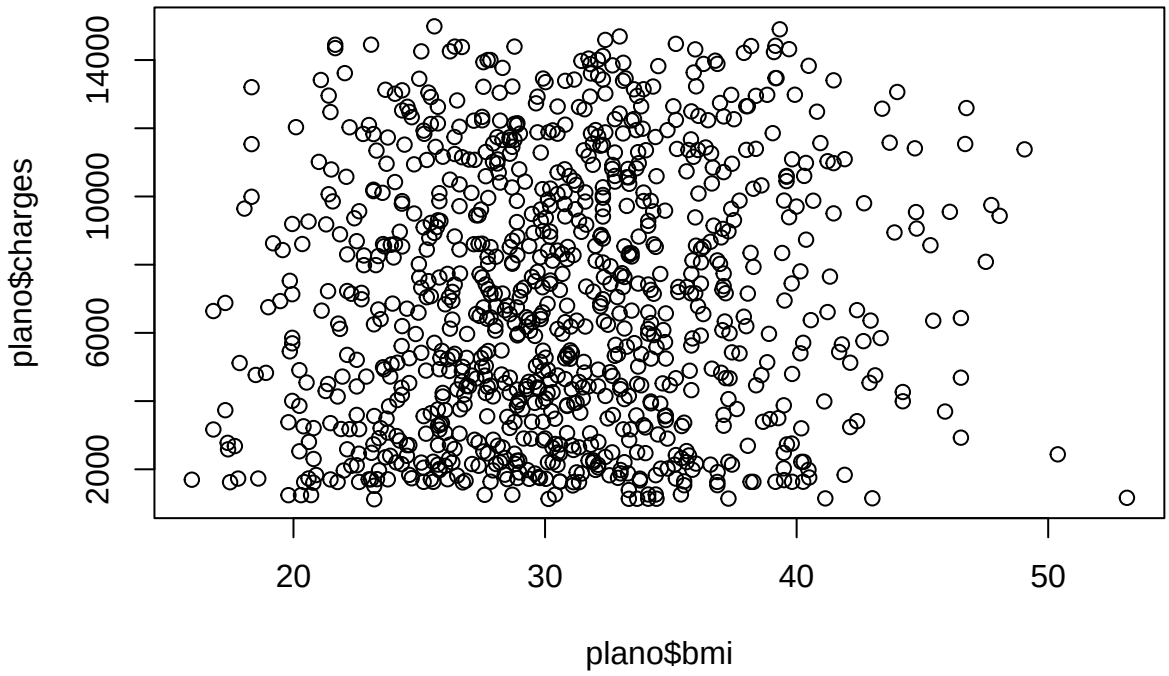
Existe uma relação entre as variáveis charges e age.

```
plot(plano$sex, plano$charges)
```



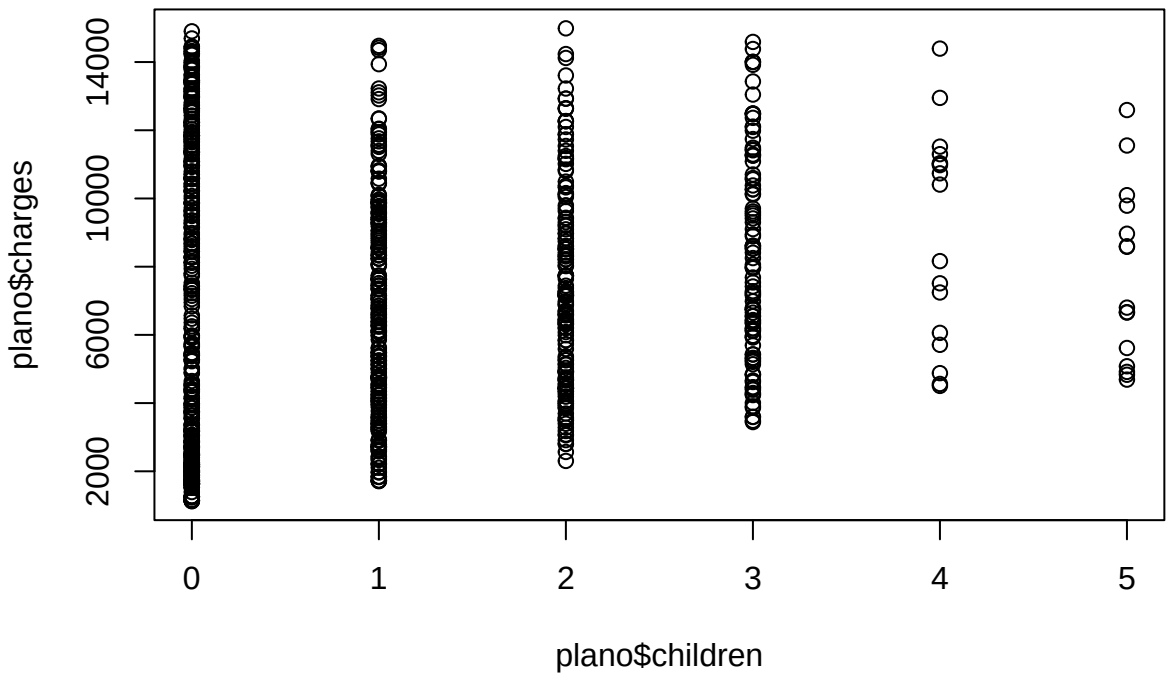
Não existe uma relação aparente.

```
plot(plano$bmi, plano$charges)
```



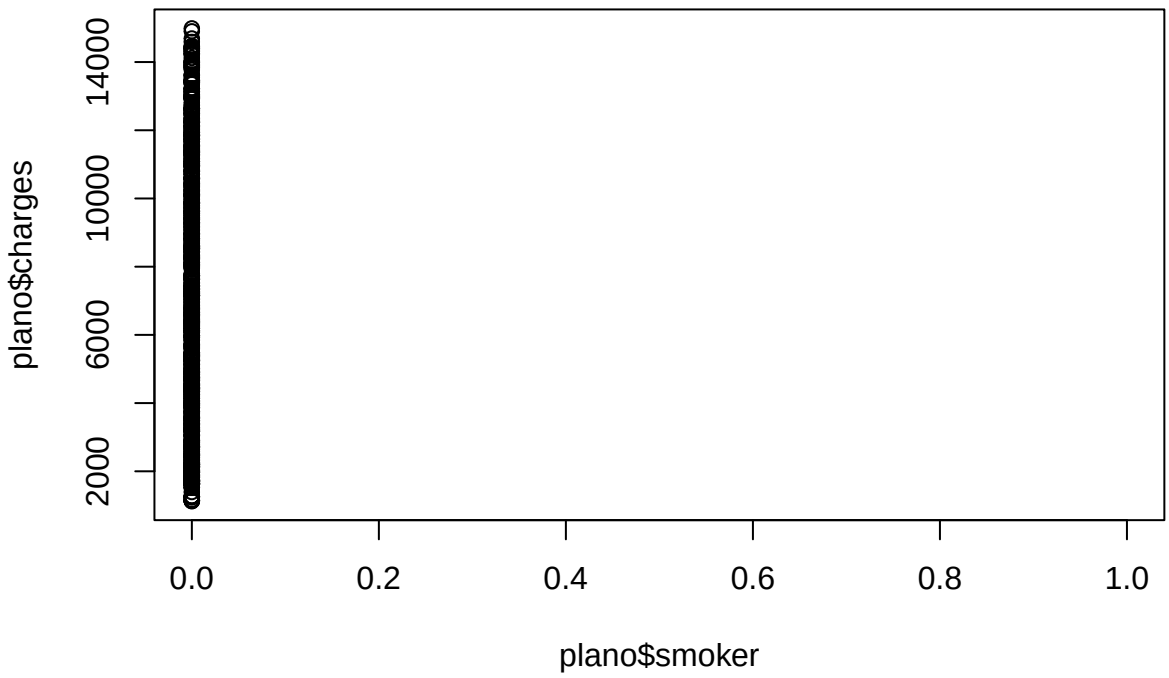
Parece existir uma relação entre as variáveis.

```
plot(plano$children, plano$charges)
```



Parece existir uma relação entre as variáveis.

```
plot(plano$smoker, plano$charges)
```



Depois da remoção dos outliers esta variável se tornou irrelevante para este projeto e, portanto, não vai mais ser levada em conta.

Após observar estes gráficos podemos começar a gerar modelos para se as relações aparentes são verdadeiras.

REGRESSÃO LINEAR SIMPLES

Modelos

Modelos de Regressão são utilizados para modelar o relacionamento entre duas ou mais variáveis explicativas e uma variável de resposta ajustando uma equação linear para dados observados. Todo valor da variável independente (x) é associada com um valor da variável dependente (y).

Após a criação dos modelos serão gerados quatro gráficos:

- No primeiro plot, **Residuals vs Fitted**, os resíduos são comparados com os valores ajustados. Se os resíduos estiverem distribuídos de forma homogênea e simétrica em torno da reta, indica que o modelo está adequado.
- No segundo plot, **Q-Q plot**, o modelo é adequado quanto mais próximos os pontos estiverem da diagonal, o que indicaria uma distribuição normal dos resíduos. É esperado que haja desvios, especialmente nos extremos. Os pontos que aparecem numerados indicam aqueles casos que merecem atenção pois são os que mais fogem das premissas.
- O terceiro plot, **Scale Location** indica se a variância é constante conforme o incremento da média. Para a regressão, se observa uma tendência de aumento da variância (representada pela raiz quadrada dos resíduos padronizados no eixo y) em relação aos valores ajustados pelo modelo (eixo x).

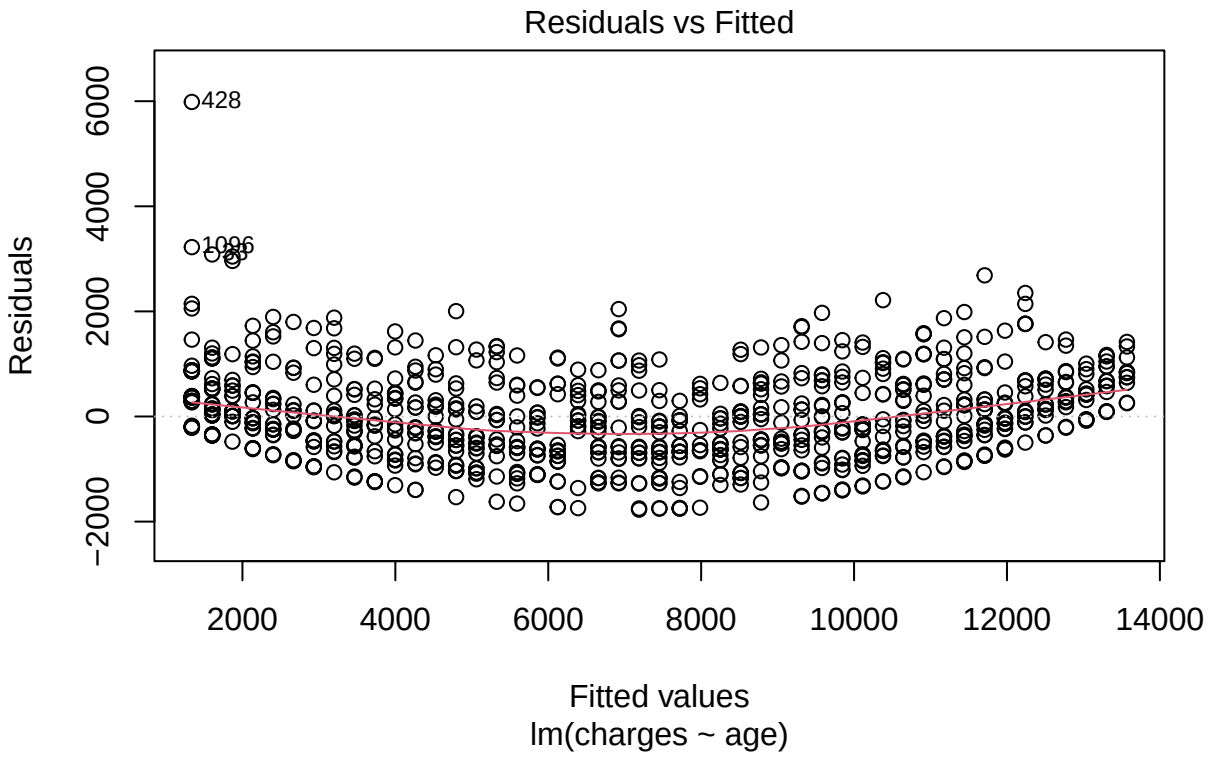
- O último plot, *Residuals vs Leverage* indica as observações, ou seja os valores de CW, que mais afetam o modelo. A linha vermelha deve passar próximo do valor 0 no eixo y, isto é coincidir com os a linha tracejada preta. Os valores acima e abaixo indicam o desvio padrão destes dados (oscilação entre -3 a +3 são típicos de uma distribuição normal). Além disso aparecem linhas pontilhadas que indicam a distância de Cook, que é uma medida de quanto a regressão mudaria caso um dos dados fosse retirado da análise. Distâncias menores que 0.5 são consideradas adequadas.

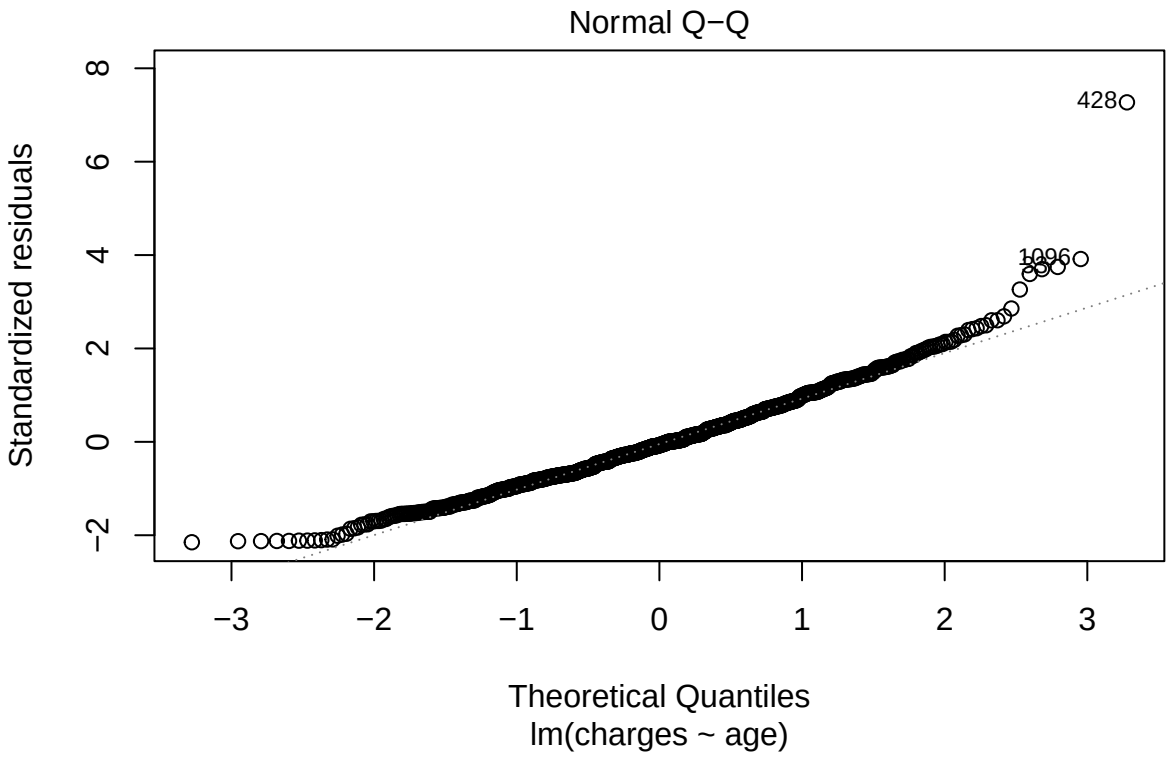
```
mod_charges = lm(charges~age,plano)

summary(mod_charges)
```

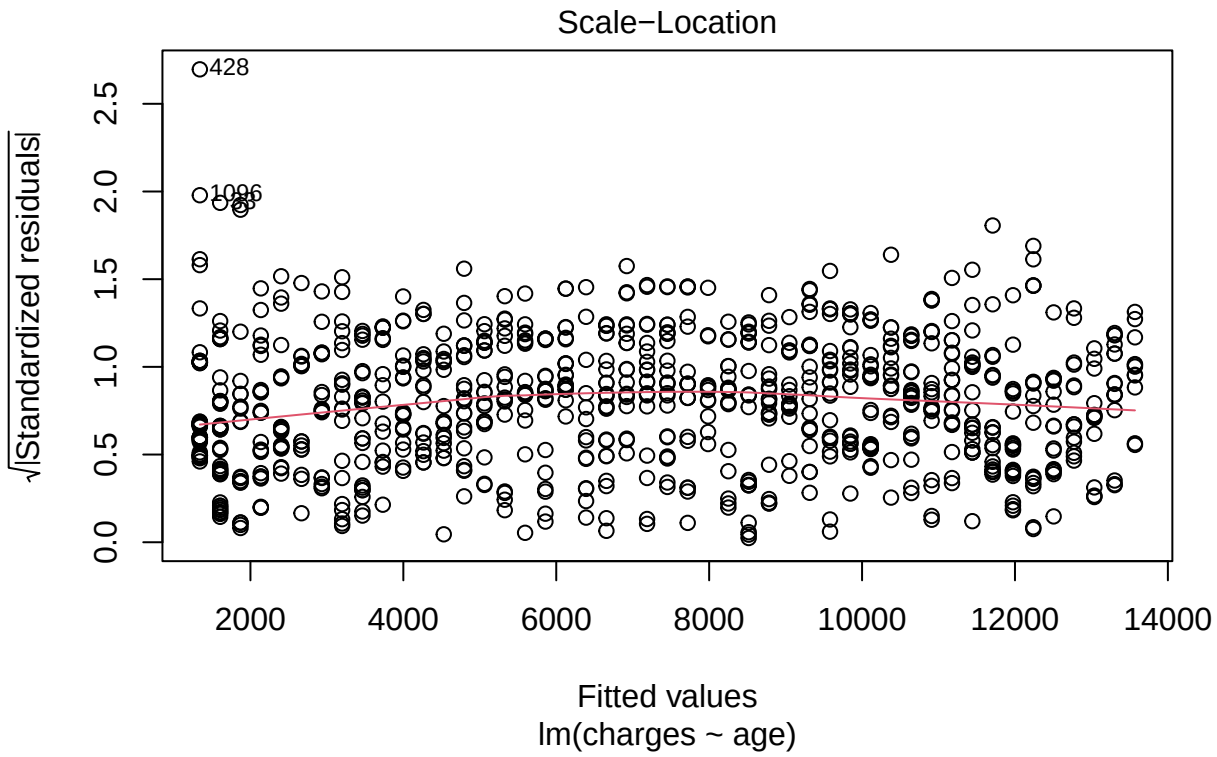
Charges vs Age

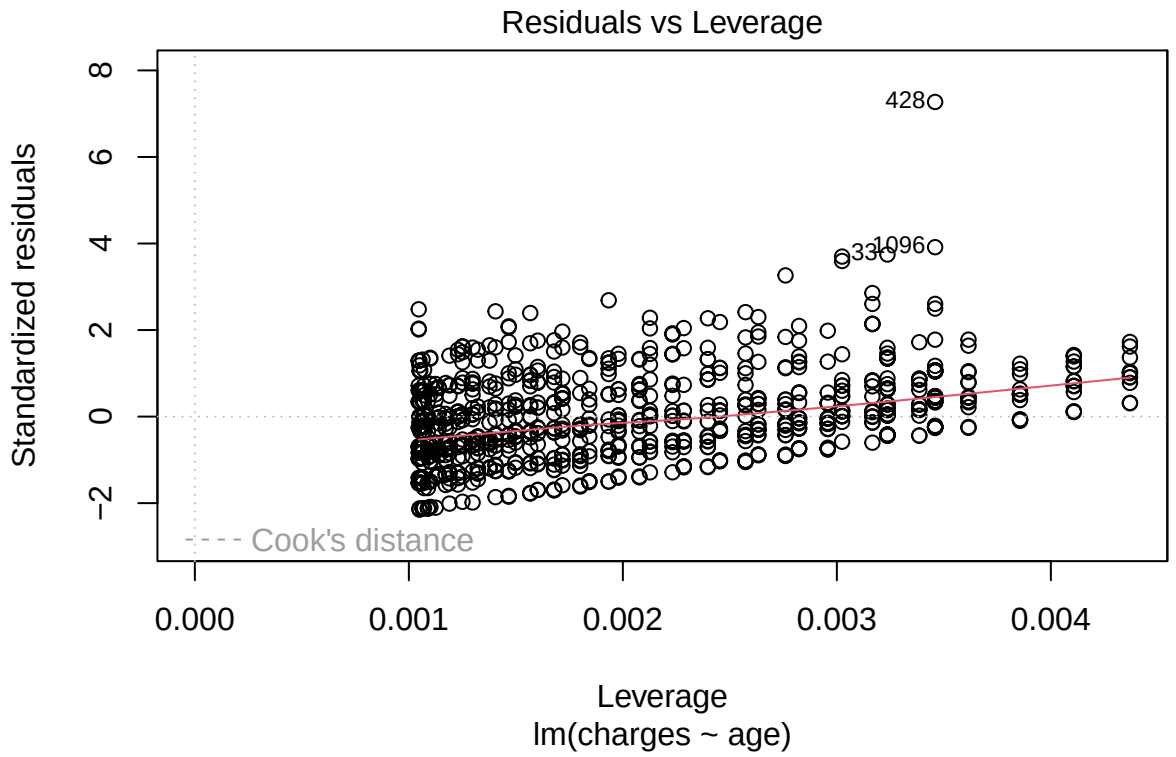
```
##
## Call:
## lm(formula = charges ~ age, data = plano)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1772.2  -580.8   -57.3   502.2  5985.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3447.998     79.569  -43.33   <2e-16 ***
## age           265.896      1.914  138.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 824.7 on 954 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.9529, Adjusted R-squared:  0.9528
## F-statistic: 1.93e+04 on 1 and 954 DF,  p-value: < 2.2e-16
```

```
plot(mod_charges)
```


Residuals vs Fitted
Residuals
Fitted values
lm(charges ~ age)
428
1086
−2000
0
2000
4000
6000
2000
4000
6000
8000
10000
12000
14000


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ age)
428
1096
−3
−2
−1
0
1
2
3
−2
0
2
4
6
8


Scale−Location
428
1096
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
2000
4000
6000
8000
10000
12000
14000
Fitted values
lm(charges ~ age)



Podemos observar resíduos que se destacam, estes serão removidos para a análise de regressão linear múltipla.

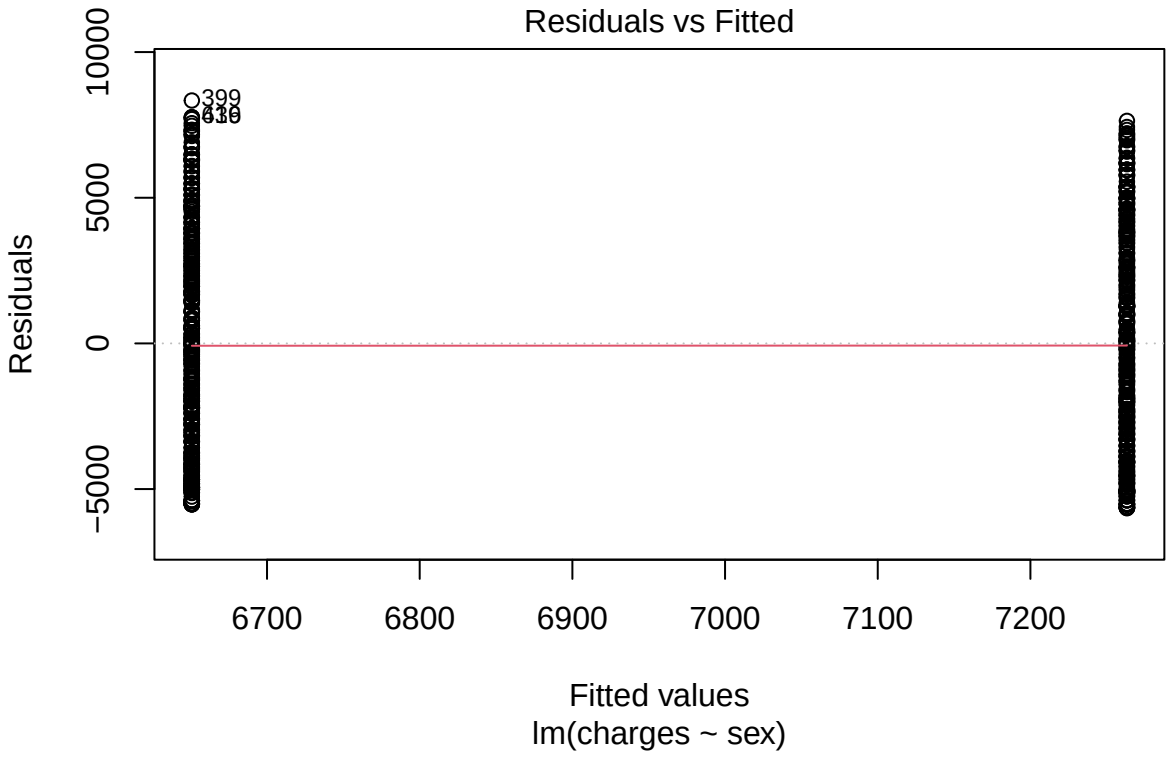
```
mod_sex = lm(charges~sex, plano)

summary(mod_sex)
```

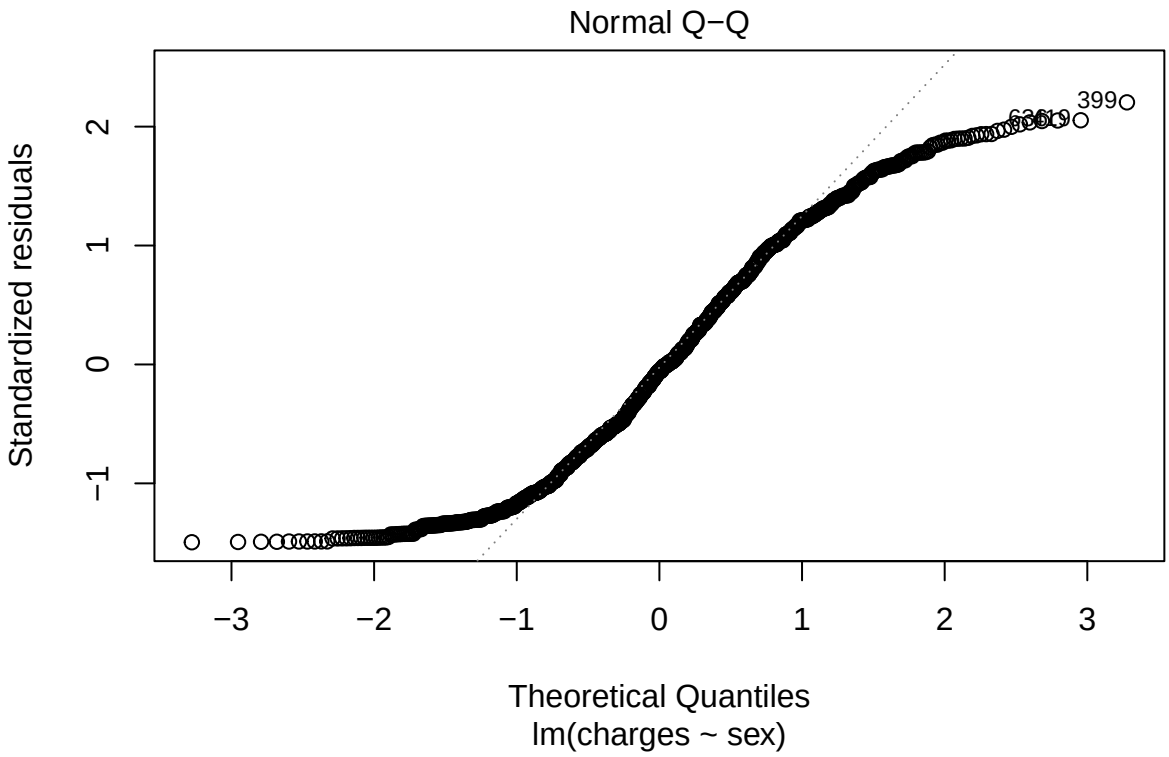
Charges vs Sex

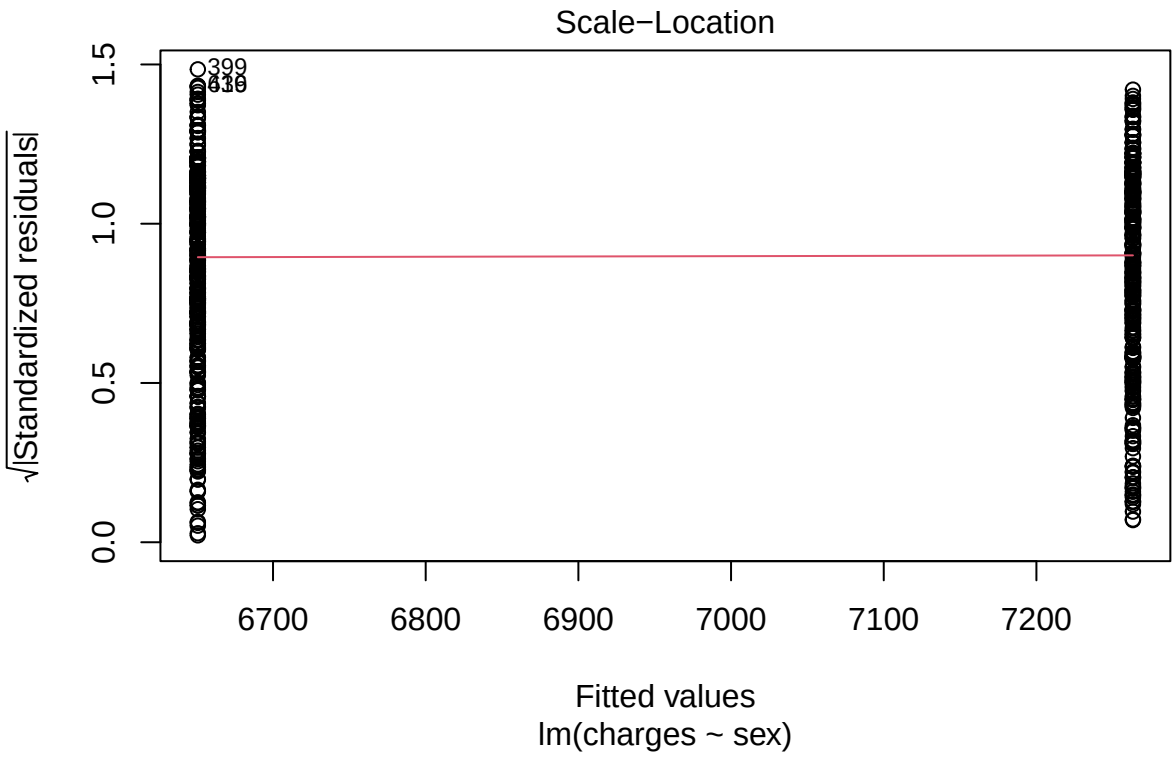
```
##
## Call:
## lm(formula = charges ~ sex, data = plano)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5655.7 -3345.3  -215.8  3153.5  8337.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7263.2      170.9  42.500   <2e-16 ***
## sex           -612.5      245.0  -2.499   0.0126 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

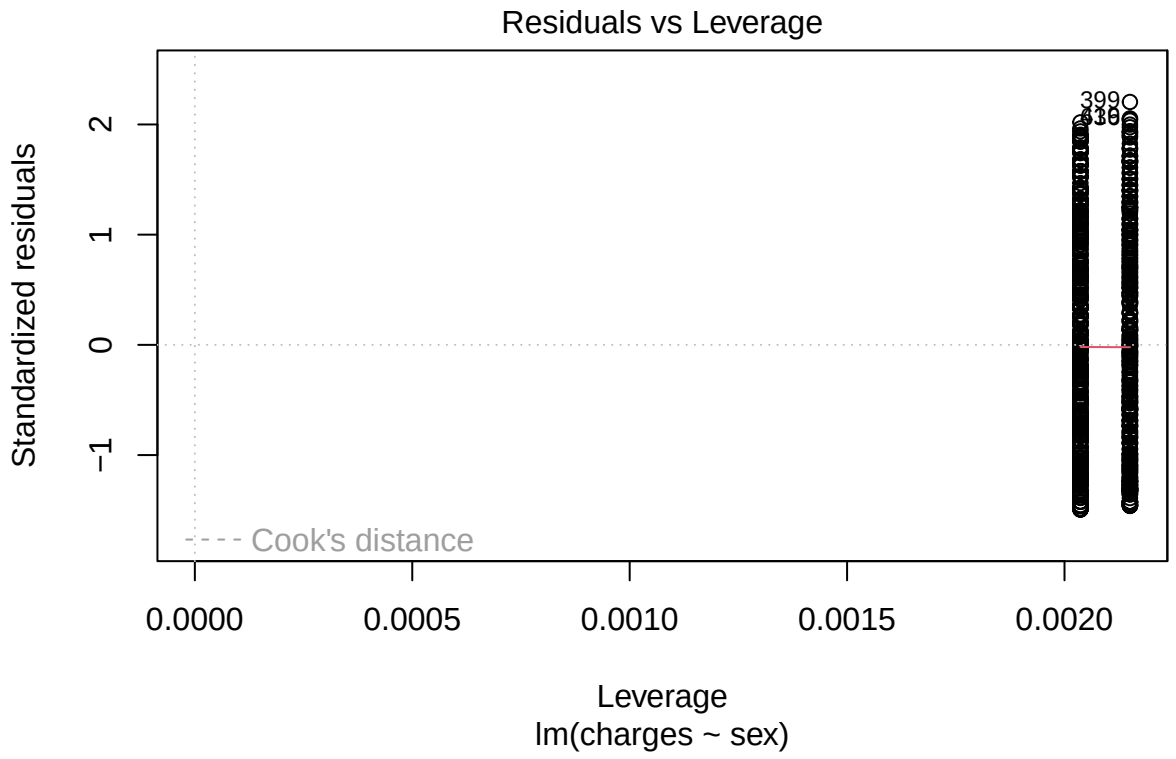
```
## Residual standard error: 3787 on 954 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.006506,   Adjusted R-squared:  0.005464
## F-statistic: 6.247 on 1 and 954 DF,  p-value: 0.01261
```

```
plot(mod_sex)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ sex)
399
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
399
6700
6800
6900
7000
7100
7200
Fitted values
lm(charges ~ sex)



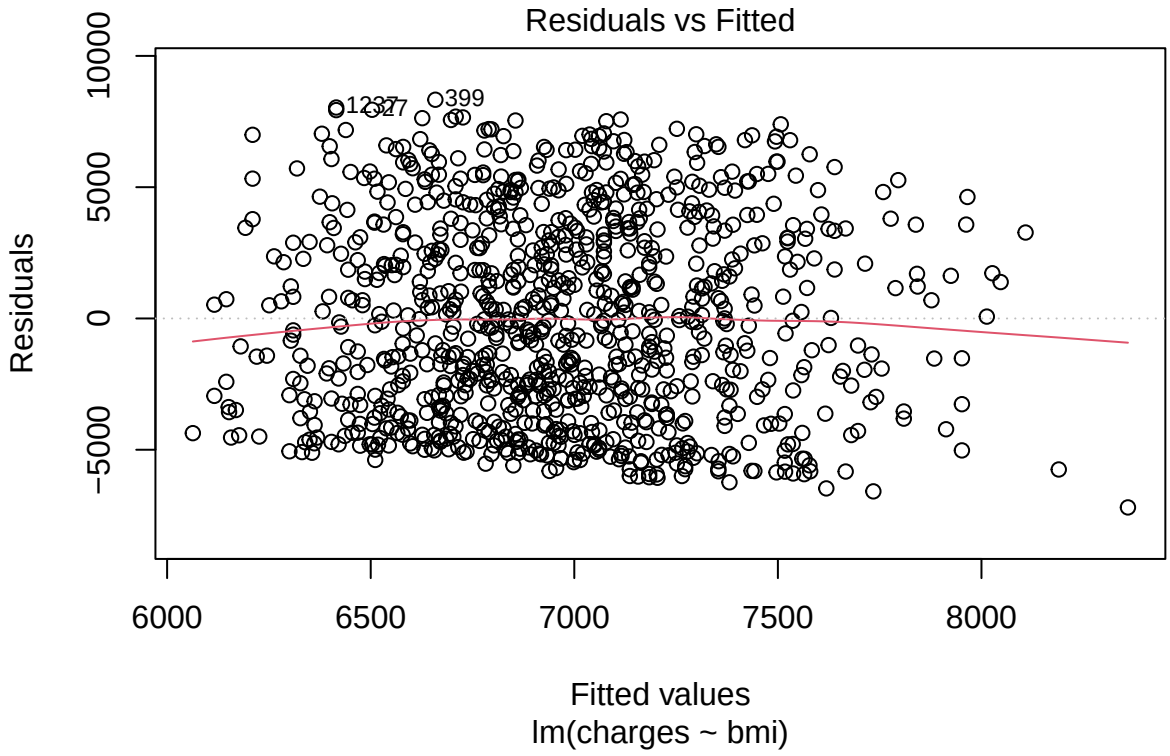

```
mod_bmi = lm(charges~bmi, plano)

summary(mod_bmi)
```

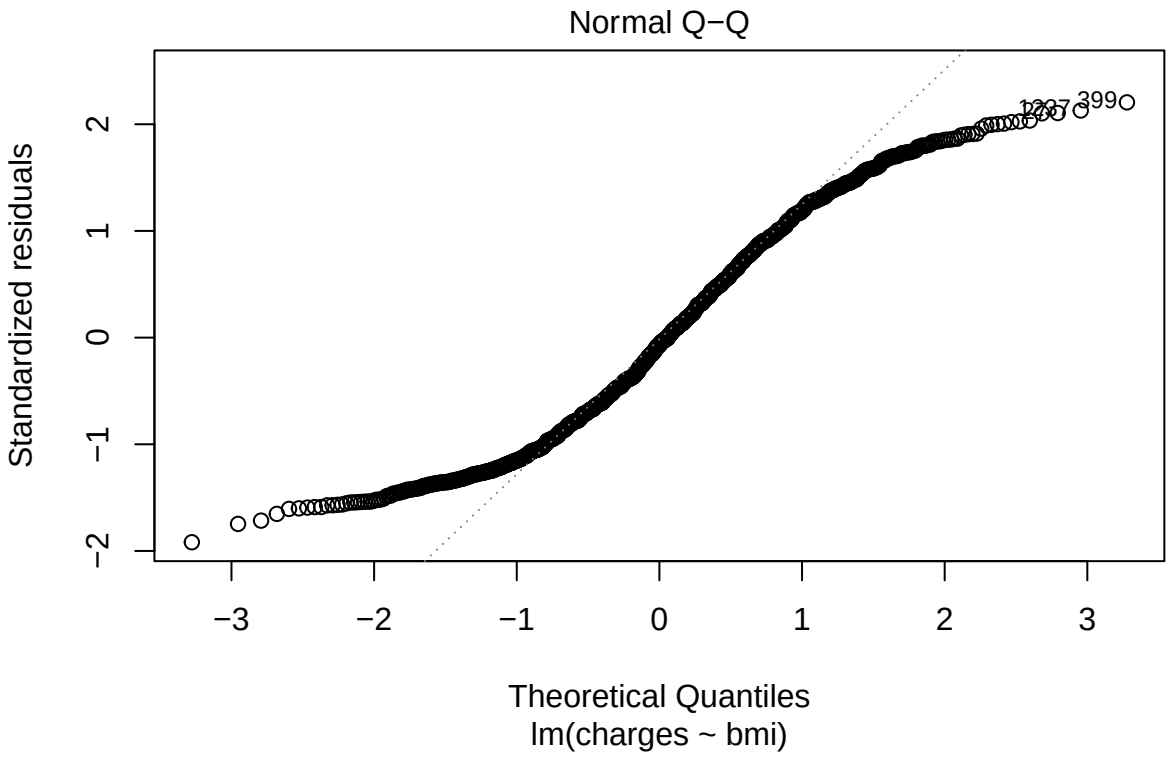
Charges vs IMC

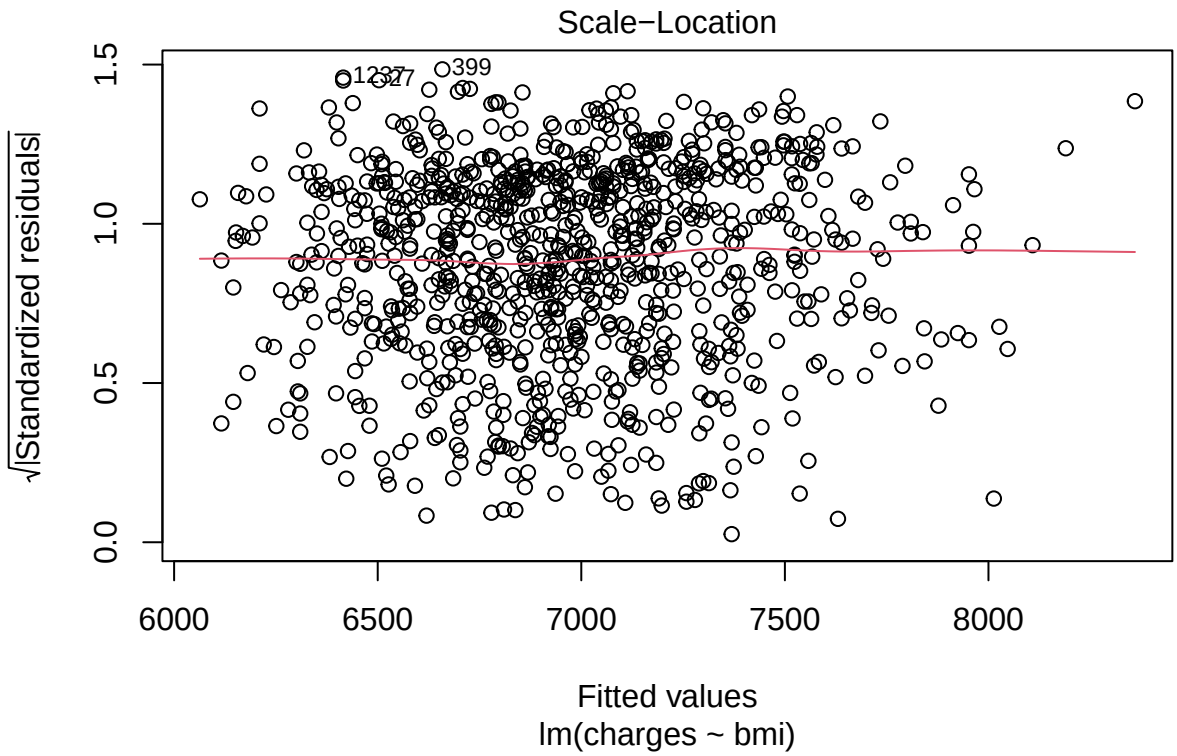
```
##
## Call:
## lm(formula = charges ~ bmi, data = plano)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7196.3 -3277.7  -266.6  3163.4  8329.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5077.04     623.97   8.137 1.26e-15 ***
## bmi            61.79      20.02   3.086  0.00209 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3780 on 954 degrees of freedom
##   (24 observations deleted due to missingness)
```

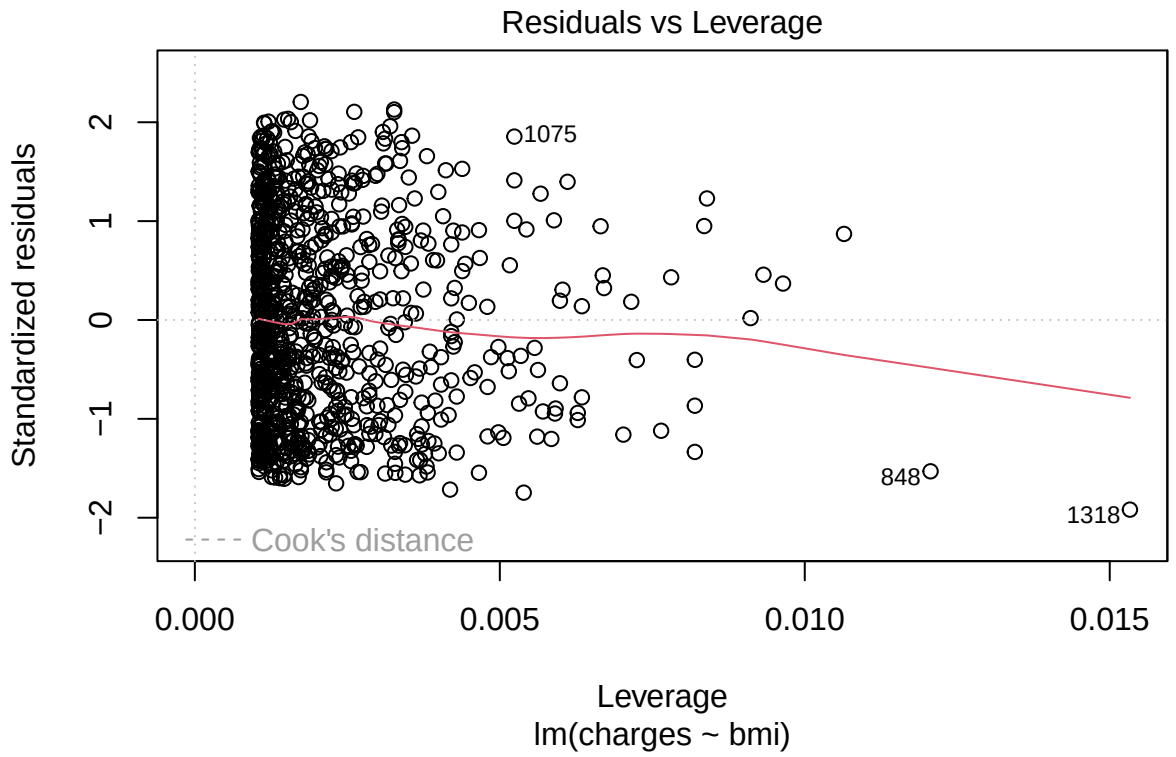
```
## Multiple R-squared:  0.009884,   Adjusted R-squared:  0.008847
## F-statistic: 9.524 on 1 and 954 DF,  p-value: 0.002087
```

```
plot(mod_bmi)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ bmi)
399


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1237
327
399
6000
6500
7000
7500
8000
Fitted values
lm(charges ~ bmi)



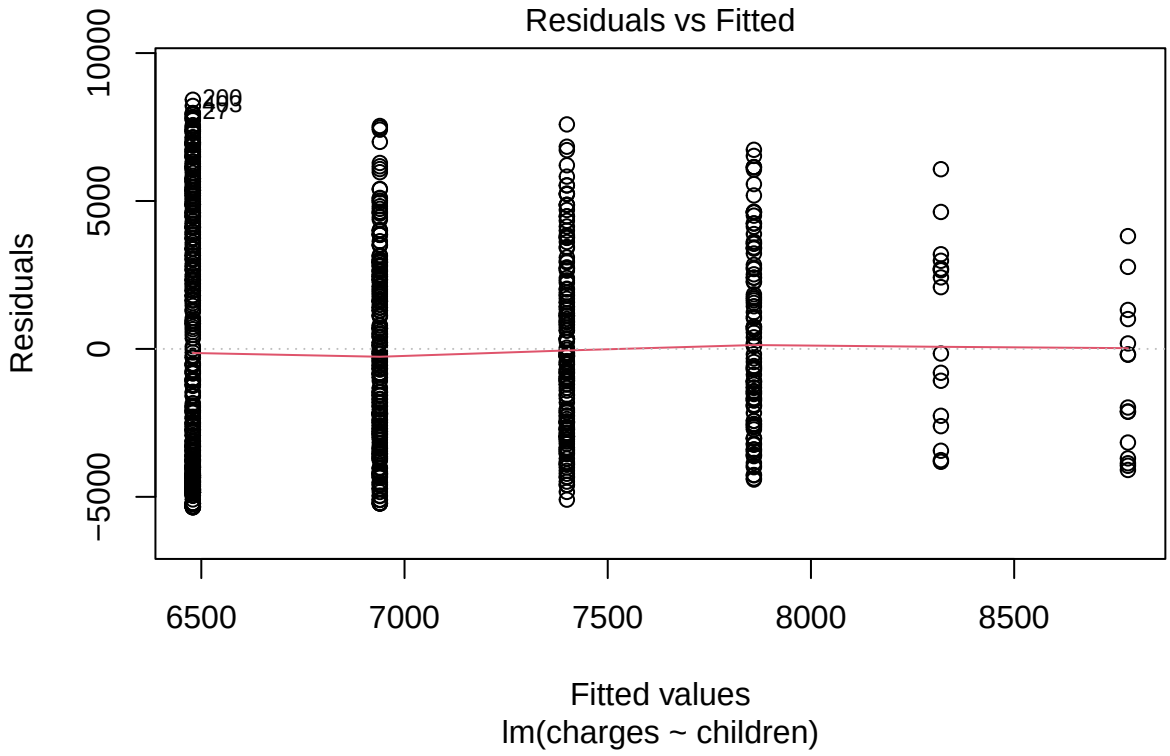

```
mod_children = lm(charges~children, plano)

summary(mod_children)
```

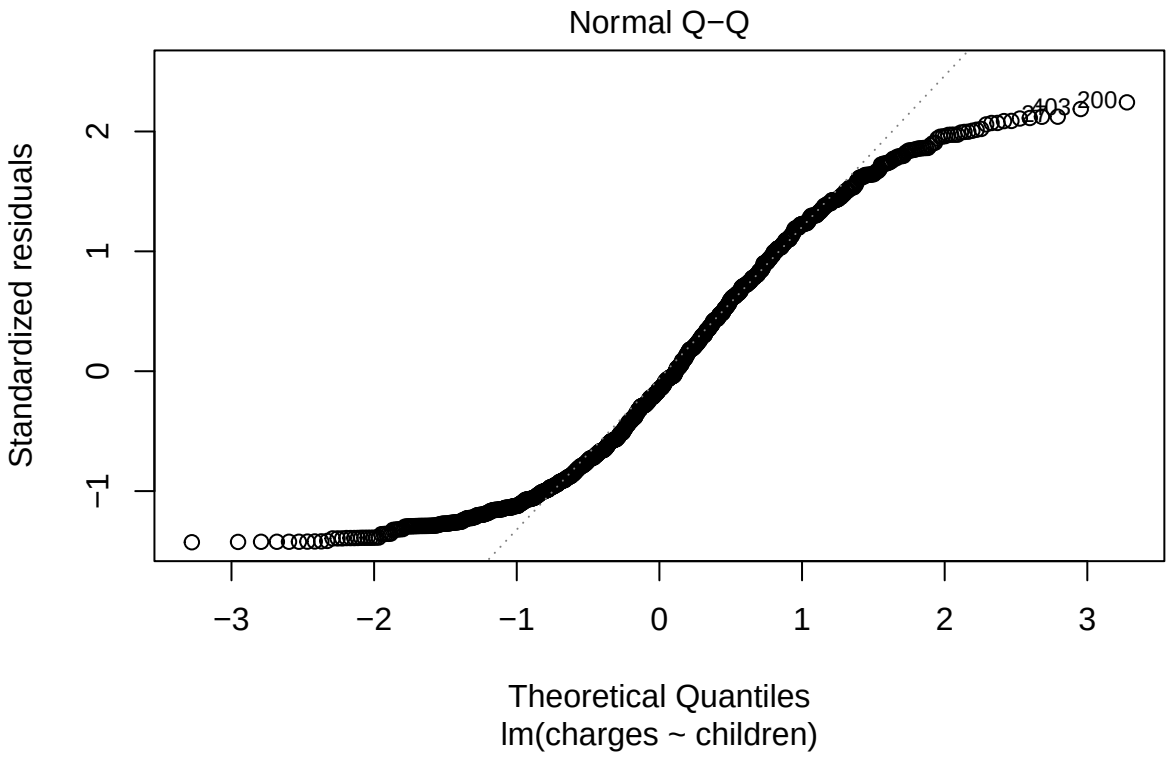
Charges vs Children

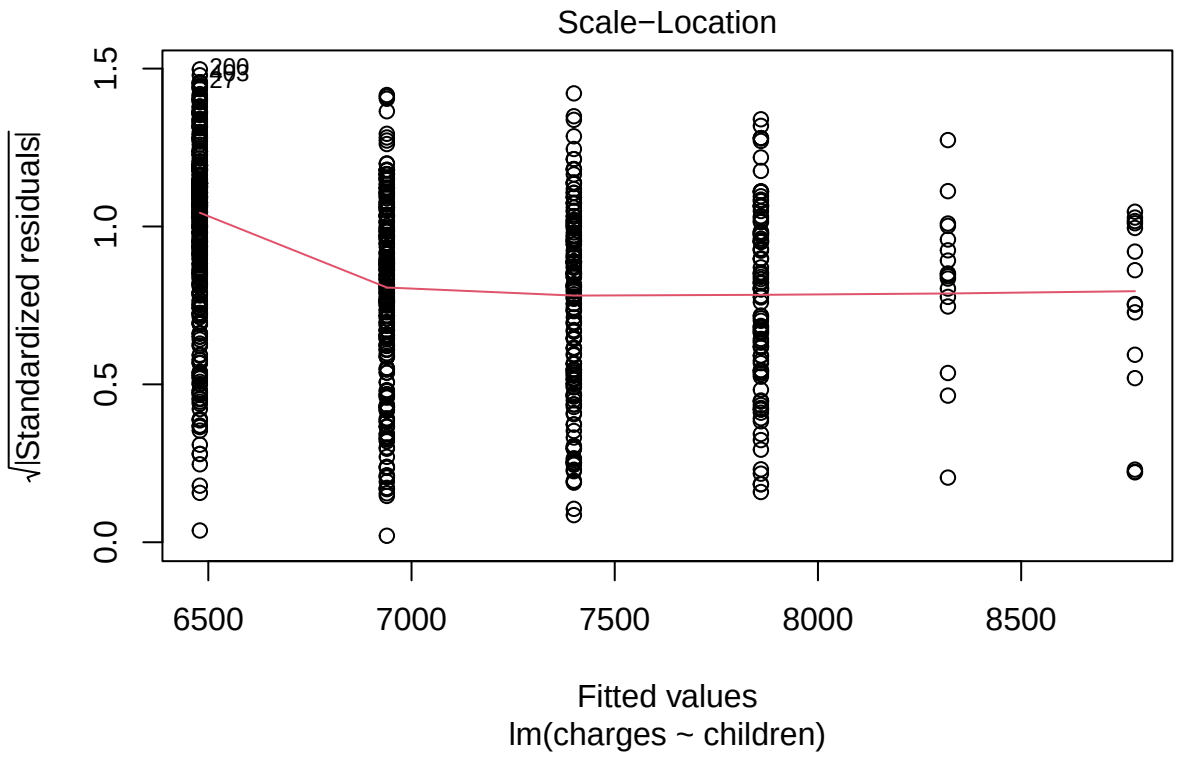
```
##
## Call:
## lm(formula = charges ~ children, data = plano)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5357.3 -3416.9  -562.4  2973.5  8422.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6479.2      161.6  40.083  < 2e-16 ***
## children       460.1      100.8   4.563 5.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3759 on 954 degrees of freedom
##   (24 observations deleted due to missingness)
```

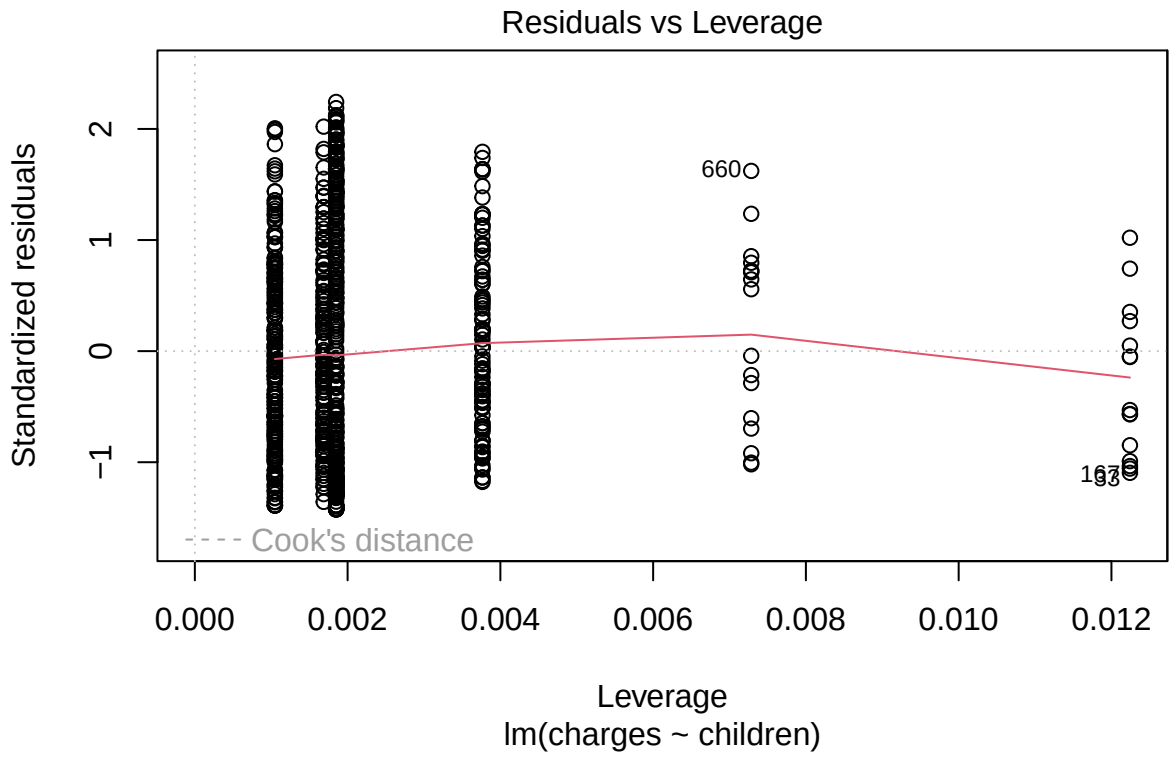
```
## Multiple R-squared:  0.02136,    Adjusted R-squared:  0.02033
## F-statistic: 20.82 on 1 and 954 DF,  p-value: 5.708e-06
```

```
plot(mod_children)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ children)
403
200
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
200
209
27
6500
7000
7500
8000
8500
Fitted values
lm(charges ~ children)



Testando normalidade dos resíduos

```
shapiro.test(mod_charges$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_charges$residuals
## W = 0.96882, p-value = 1.911e-13
```

```
shapiro.test(mod_sex$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_sex$residuals
## W = 0.94861, p-value < 2.2e-16
```

```
shapiro.test(mod_bmi$residuals)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  mod_bmi$residuals
## W = 0.95768, p-value = 5.406e-16
```

```
shapiro.test(mod_children$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_children$residuals
## W = 0.93879, p-value < 2.2e-16
```

Analise de outliers nos resíduos

```
summary(rstandard(mod_charges))
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.150121 -0.705134 -0.069623  0.000173  0.609351  7.270943
```

```
car::outlierTest(mod_charges)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 428 7.477263         1.7188e-13   1.6432e-10
```

```
summary(rstandard(mod_sex))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.49501 -0.88433 -0.05705  0.00000  0.83363  2.20407
```

```
car::outlierTest(mod_sex)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 399 2.208548           0.027443           NA
```

```
summary(rstandard(mod_bmi))
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -1.9183077 -0.8681366 -0.0705839 -0.0000392  0.8372362  2.2052498
```

```
car::outlierTest(mod_bmi)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 399 2.209733           0.027361           NA
```

```
summary(rstandard(mod_children))
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -1.4267013 -0.9099288 -0.1497494 -0.0000002  0.7923561  2.2429444
```

```
car::outlierTest(mod_children)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 200 2.247703           0.024823           NA
```

Analise de independência dos resíduos

```
durbinWatsonTest(mod_charges)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.06926558      1.861212   0.032
##  Alternative hypothesis: rho != 0
```

```
durbinWatsonTest(mod_sex)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02468099      1.946846   0.424
##  Alternative hypothesis: rho != 0
```

```
durbinWatsonTest(mod_bmi)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.03382719      1.928582   0.262
##  Alternative hypothesis: rho != 0
```

```
durbinWatsonTest(mod_children)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.03246344      1.931573   0.294
##  Alternative hypothesis: rho != 0
```

Teste de Homocedascidade

```
bptest(mod_charges)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod_charges
## BP = 1.1024, df = 1, p-value = 0.2937
```

```
bptest(mod_sex)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod_sex
## BP = 5.945e-05, df = 1, p-value = 0.9938
```

```
bptest(mod_bmi)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod_bmi
## BP = 2.2244, df = 1, p-value = 0.1358
```

```
bptest(mod_children)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod_children
## BP = 99.564, df = 1, p-value < 2.2e-16
```

Gráficos de dispersão

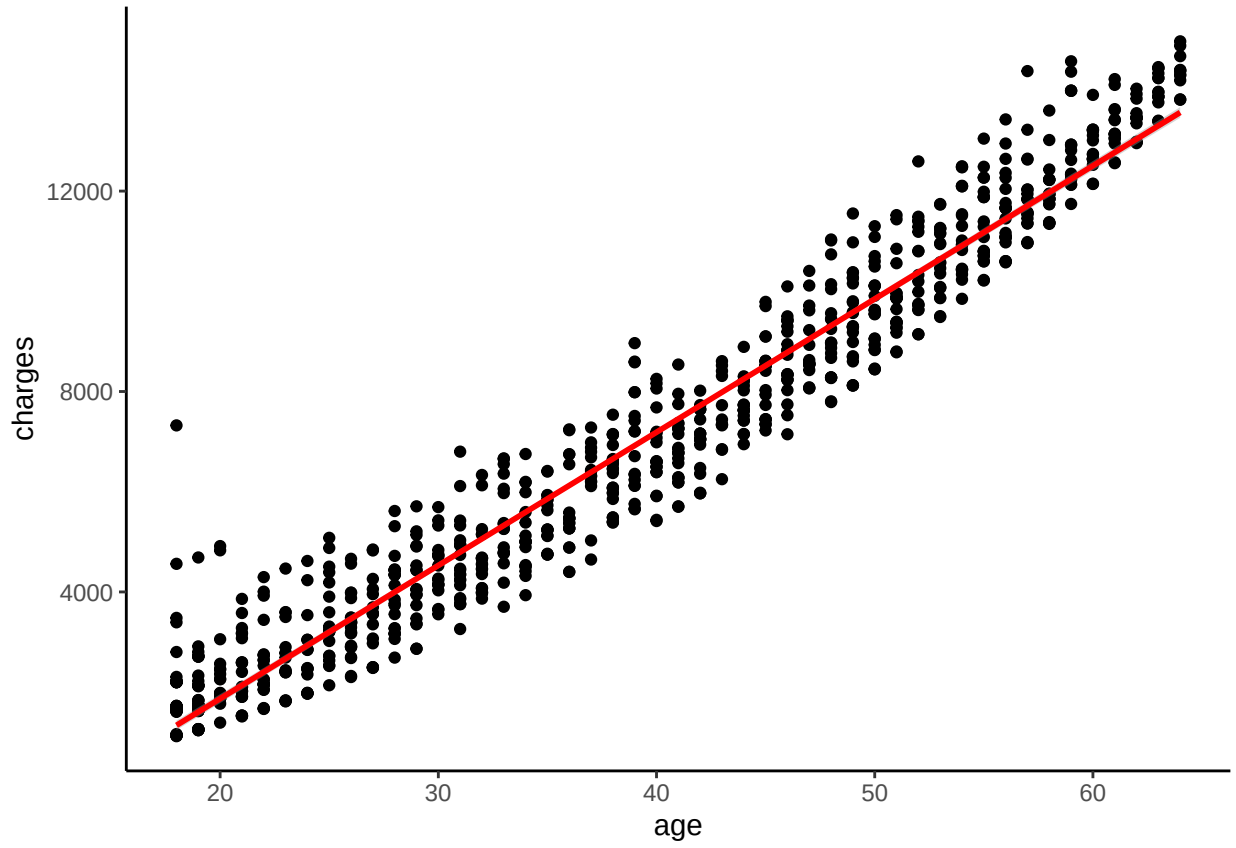
```
ggplot(data = plano, mapping = aes(x = age, y = charges)) +
  geom_point()+
  geom_smooth(method = "lm", col = "red")+
  theme_classic()
```

Charges vs Age

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 24 rows containing non-finite values (stat_smooth).
```

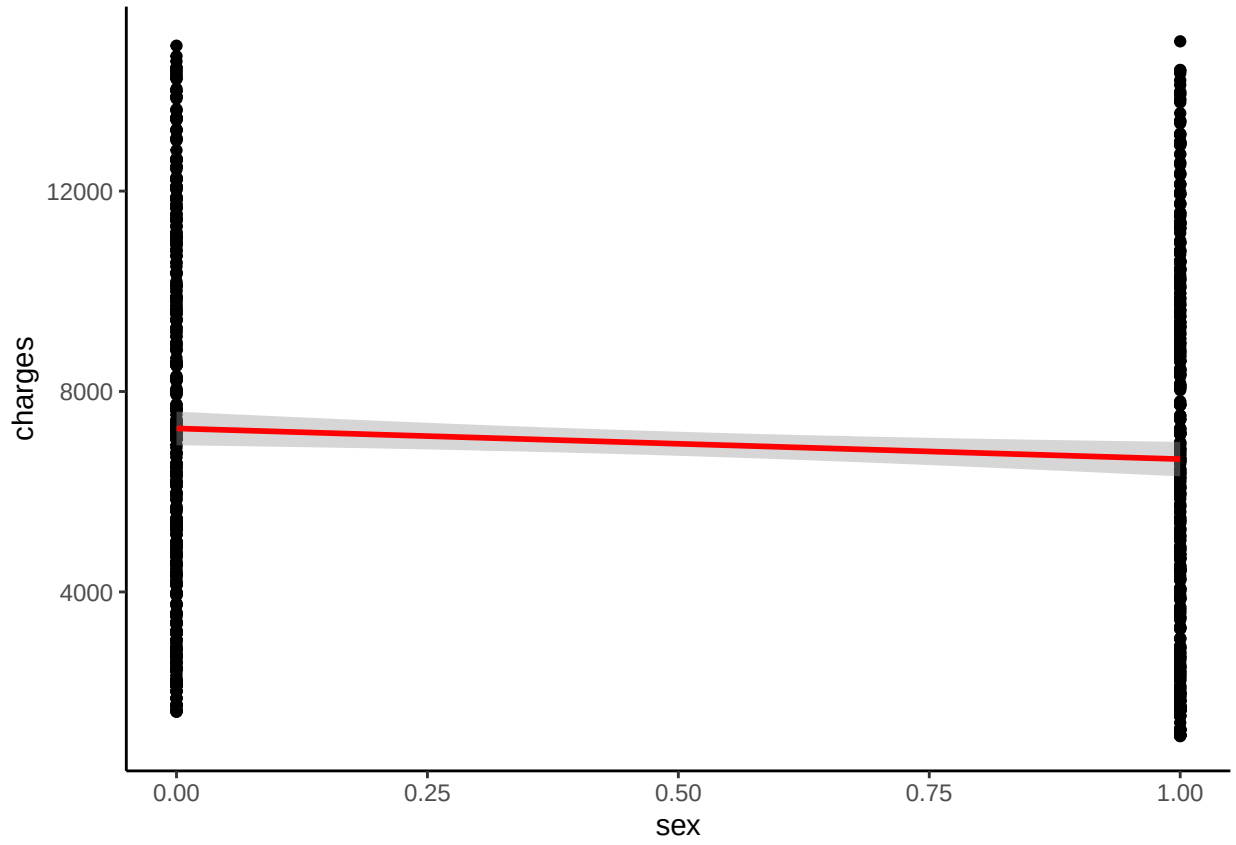
```
## Warning: Removed 24 rows containing missing values (geom_point).
```




```
ggplot(data = plano, mapping = aes(x = sex, y = charges)) +
  geom_point()+
  geom_smooth(method = "lm", col = "red")+
  theme_classic()
```

Charges vs Sex

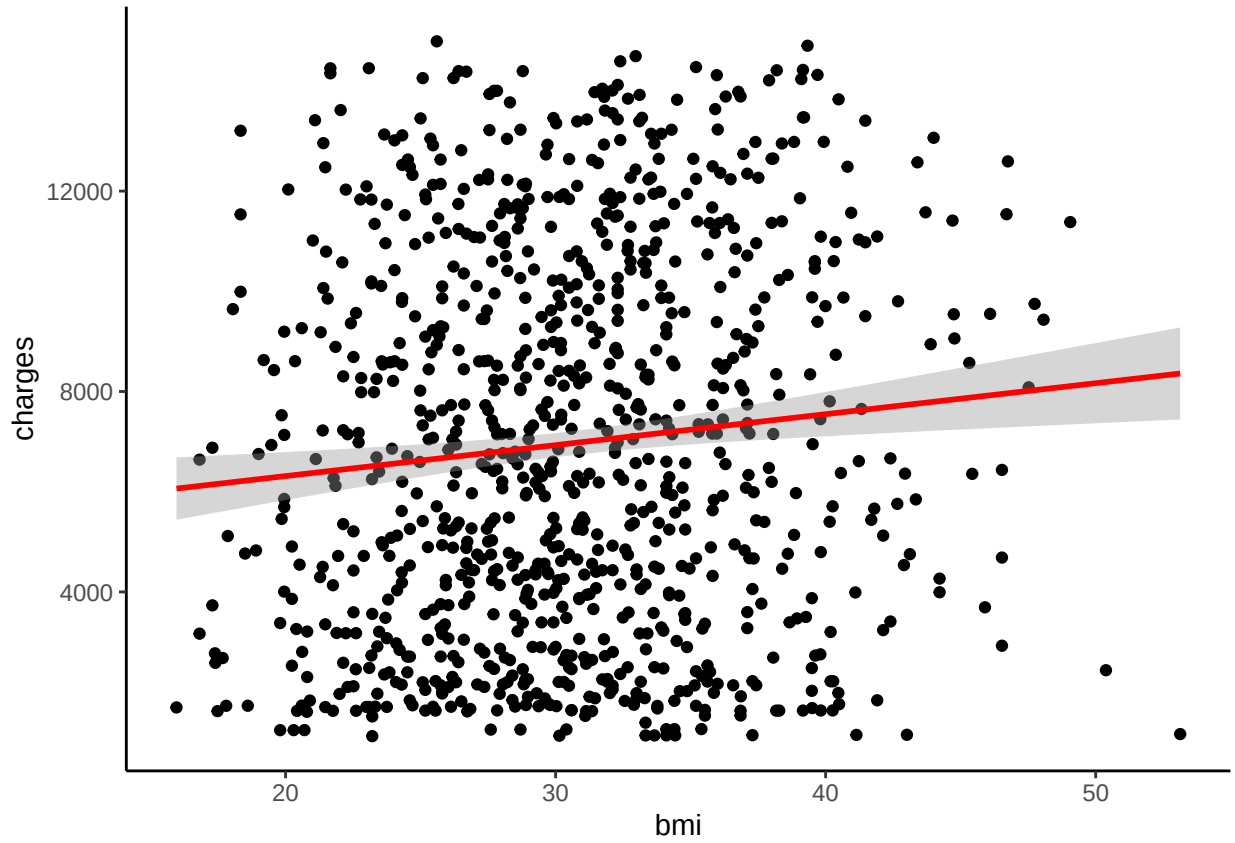
```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 24 rows containing non-finite values (stat_smooth).

## Warning: Removed 24 rows containing missing values (geom_point).
```




```
ggplot(data = plano, mapping = aes(x = bmi, y = charges)) +
  geom_point()+
  geom_smooth(method = "lm", col = "red")+
  theme_classic()
```

Charges vs IMC

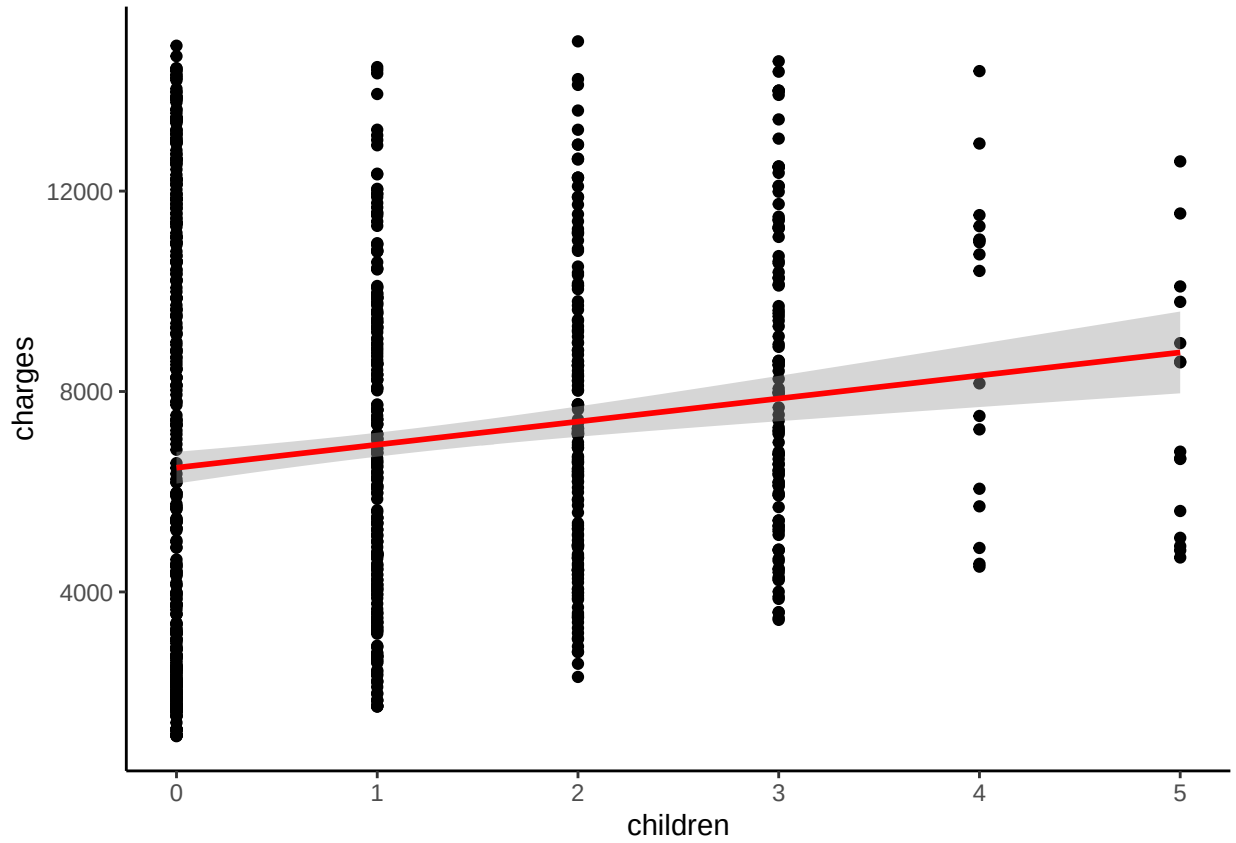
```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 24 rows containing non-finite values (stat_smooth).

## Warning: Removed 24 rows containing missing values (geom_point).
```




```
ggplot(data = plano, mapping = aes(x = children, y = charges)) +
  geom_point()+
  geom_smooth(method = "lm", col = "red")+
  theme_classic()
```

Charges vs Children

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 24 rows containing non-finite values (stat_smooth).

## Warning: Removed 24 rows containing missing values (geom_point).
```



REGRESSÃO LINEAR MULTIPLA

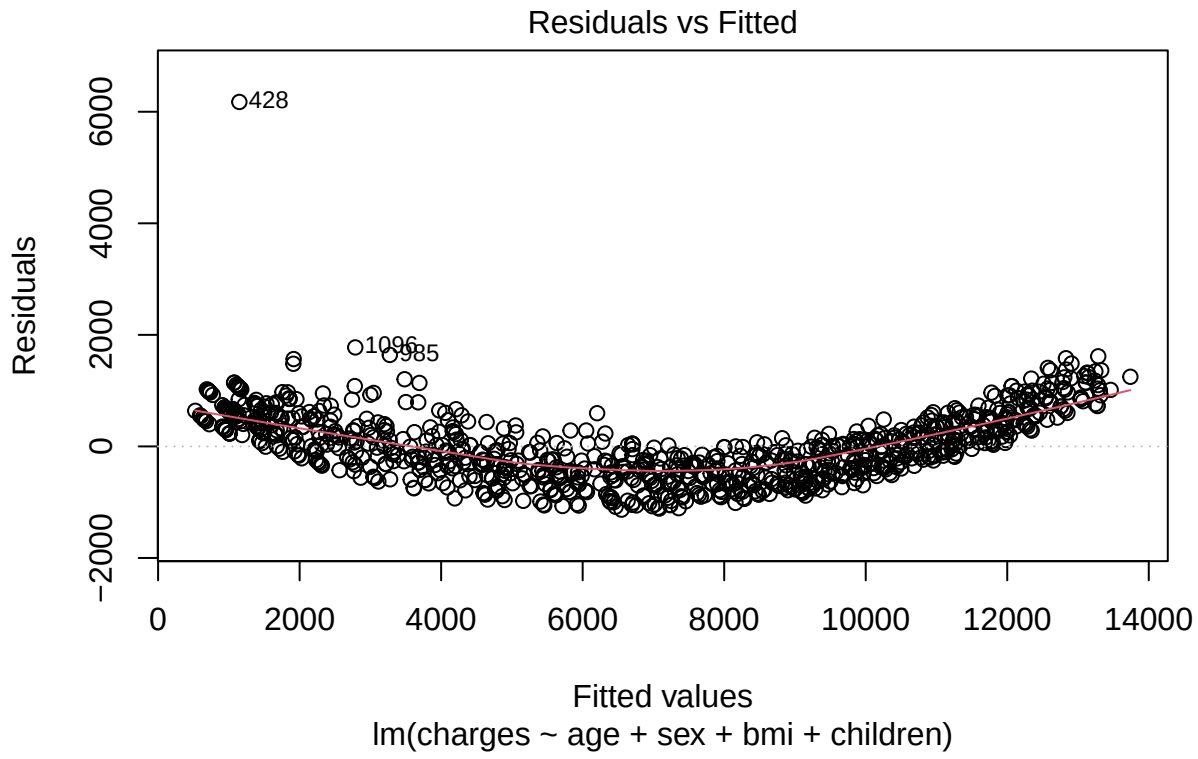
Objetivo

Verificar se idade, sexo, IMC e número de filhos influencia na quantia cobrada nos preços dos planos de saúde.

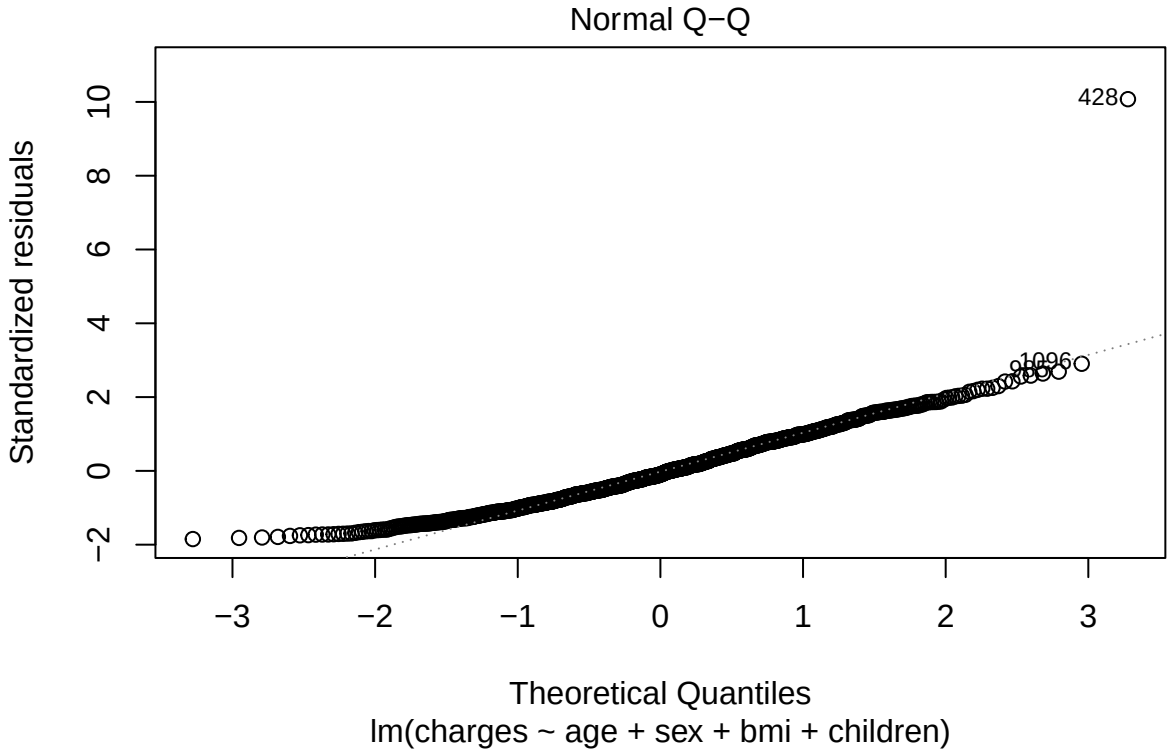
```
mod_mul = lm(charges~age+sex+bmi+children, plano)
```

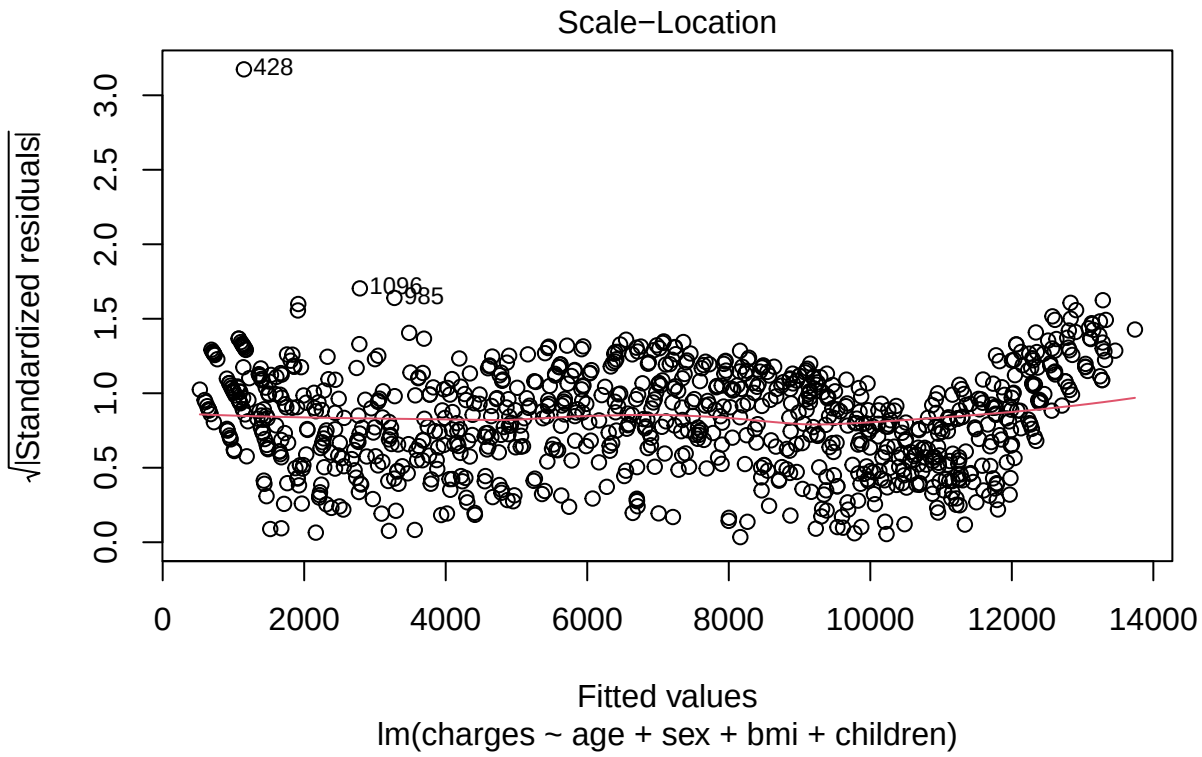
Construindo o modelo

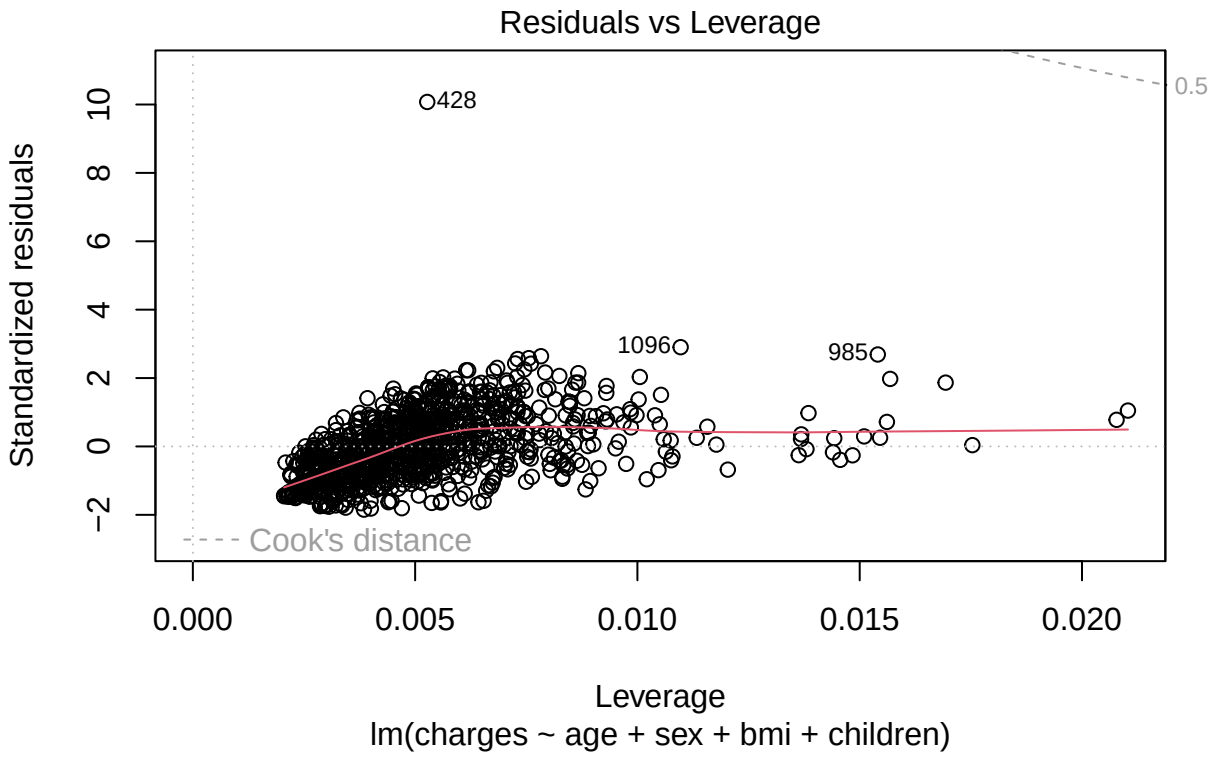
```
plot(mod_mul)
```



Analise gráfica




Scale−Location
428
1096
985
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
2000
4000
6000
8000
10000
12000
14000
Fitted values
lm(charges ~ age + sex + bmi + children)



Normalidade dos resíduos Como pudemos observar no Q-Q plot os resíduos parecem apresentar distribuição normal então podemos testar estes resíduos para verificar essa informação.

```
shapiro.test(mod_mul$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_mul$residuals
## W = 0.93659, p-value < 2.2e-16
```

O teste de normalidade de Shapiro-Wilk tem como H0 a distribuição normal dos dados e o H1 como uma distribuição diferente da normal. Neste caso o p-valor dos resíduos foi de 2.2e-16, este resultado rejeita a H0 pois é muito menor que a probabilidade de se encontrar um resultado significativo.

```
summary(rstandard(mod_mul))
```

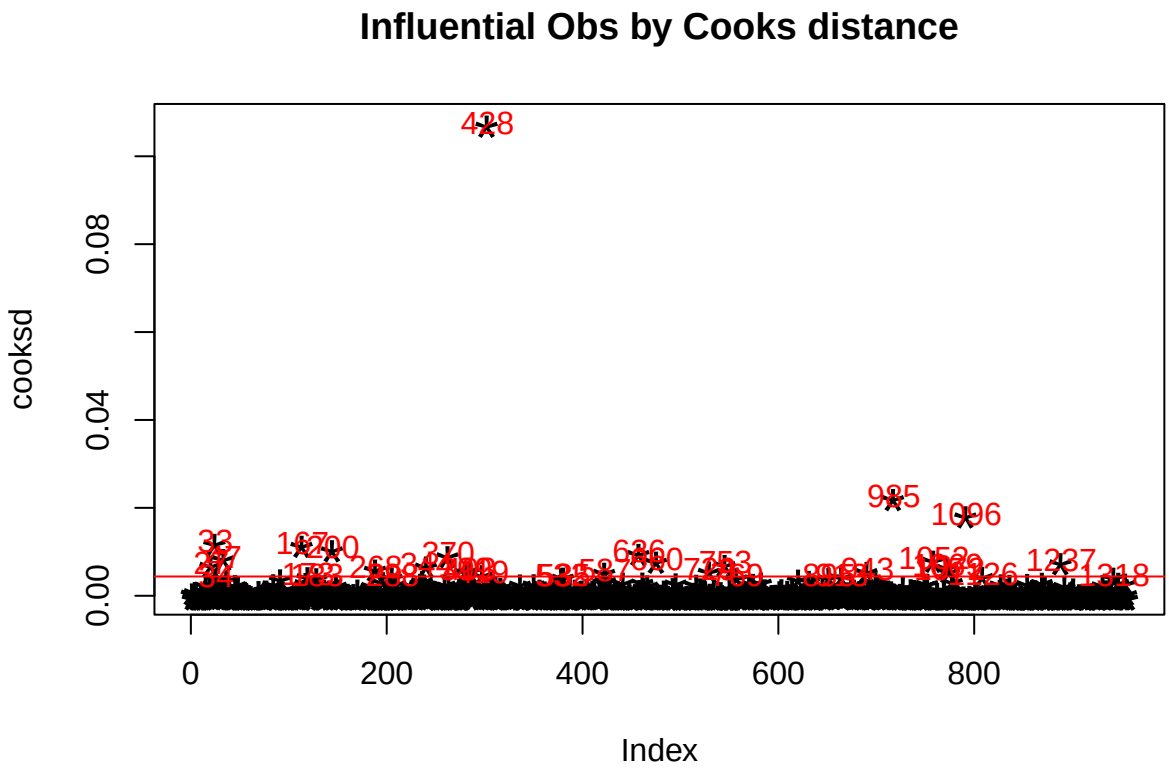
Outliers de resíduos

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.84773 -0.73233 -0.09500  0.00051  0.69050 10.07452
```

Utilizando a função summary em conjunto com a função rstandard podemos observar se existem outliers, neste caso podemos observar que sim, existem outliers pois o valor máximo dos resíduos padronizados é de 10,07452

Tratamento de Outliers Podemos fazer a identificação de outliers utilizando a distância de Cook, uma medida calculada em relação a um determinado modelo de regressão e, portanto, é afetada apenas pelas variáveis incluídas no modelo. Ele calcula a influência exercida por cada ponto de dados no resultado previsto.

```
cooksd = cooks.distance(mod_mul)
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")
abline(h = 4*mean(cooksd, na.rm=T), col="red")
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4*mean(cooksd, na.rm=T),names(cooksd),""), co
```



Utilizando a distância de Cook podemos exibir as linhas do dataframe que são consideradas influenciais (outliers) e utilizando a função outlierTest do pacote car podemos ver qual é a linha que tem a influência mais extrema e remove-la.

```
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T))])

head(plano[influential, ])
```

```
##     age sex    bmi children smoker    region   charges
## 38   26   1 20.800        0      0 southwest  2302.300
## 47   18   0 38.665        2      0 northeast  3393.356
```

```
## 48   28   0 34.770        0      0 northwest  3556.922
## 68   40   1 26.315        1      0 northwest  6389.378
## 227  28   1 38.060        0      0 southeast  2689.495
## 234  59   1 27.500        1      0 southwest 12333.828
```

```
car::outlierTest(mod_mul)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 428 10.65377         4.1014e-25   3.9209e-22
```

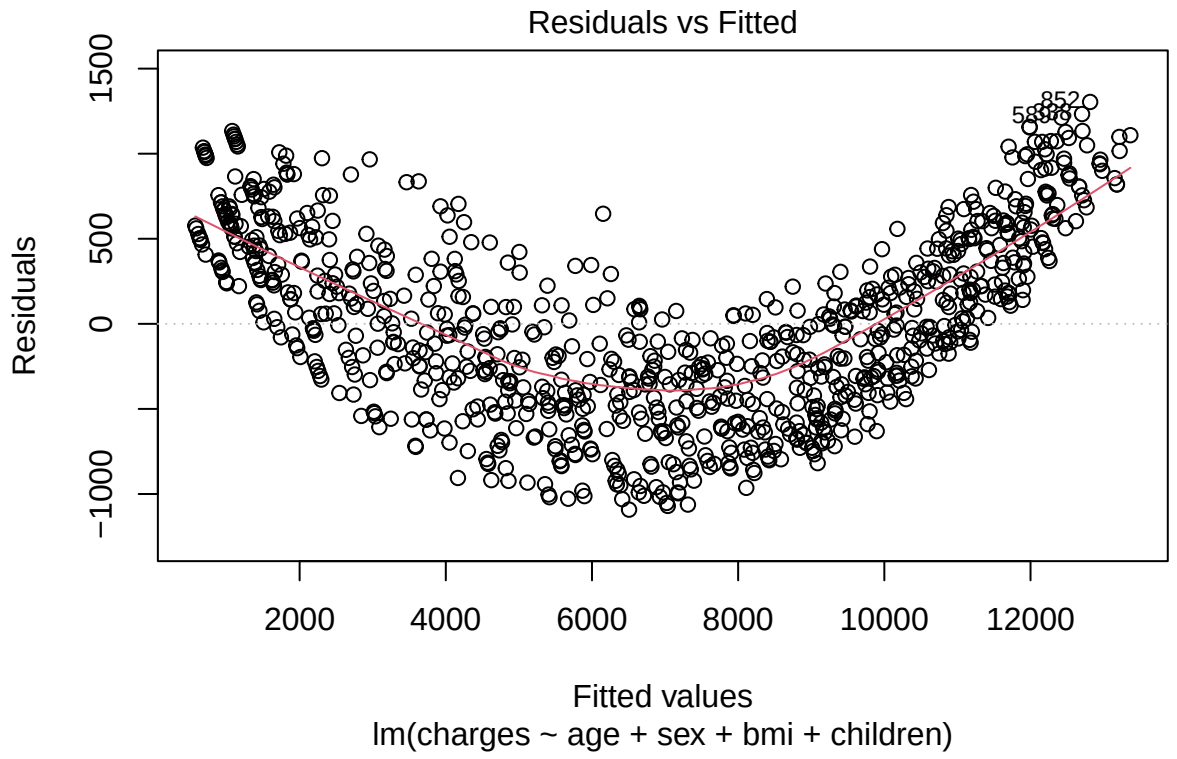
A partir desse cálculo de resíduos influenciais podemos retrabalhar o dataset para remover estes dados.

```
plano = insurance %>%
  mutate(smoking = if_else(smoker=="no", 0,1)) %>%
  mutate(sex = if_else(sex == "male", 1,0))

plano=plano[-influential,]

plano=plano[plano$charges<=15000, ]

plano$charges = ifelse(plano$charges>9000 &plano$age<45, plano$charges==NA, plano$charges)
```

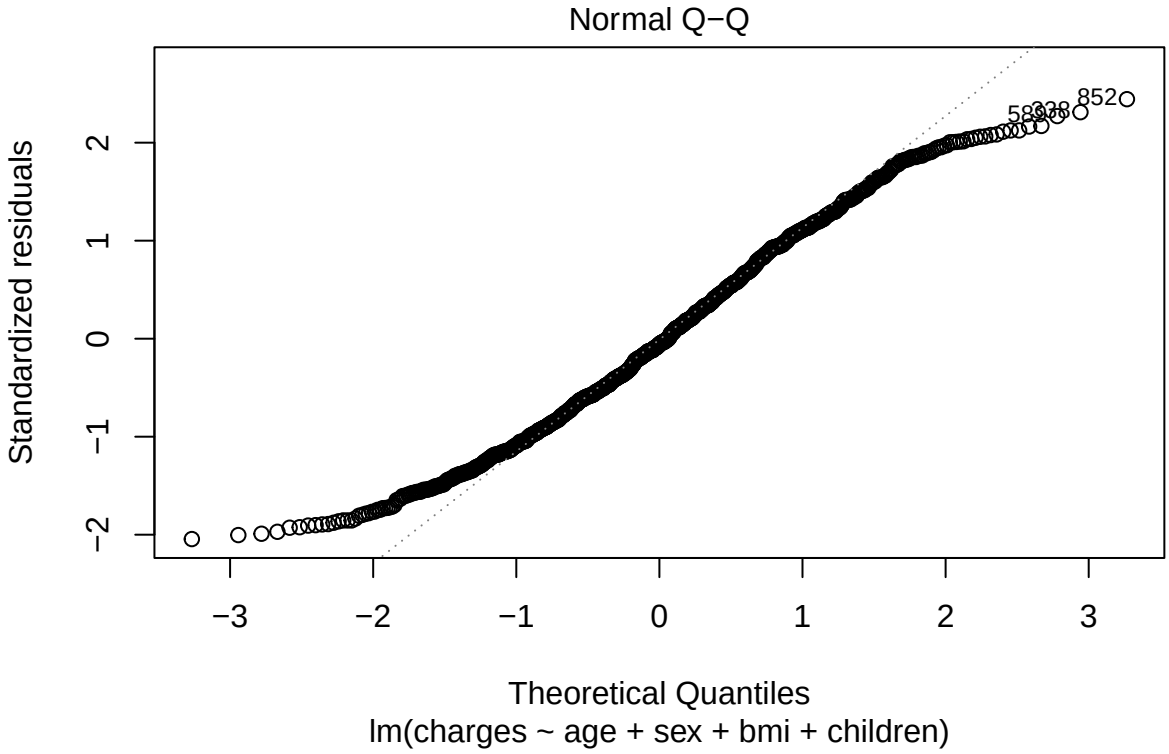
Depois de retrabalhar o dataset podemos reconstruir o modelo e fazer a análise a partir deste novo modelo

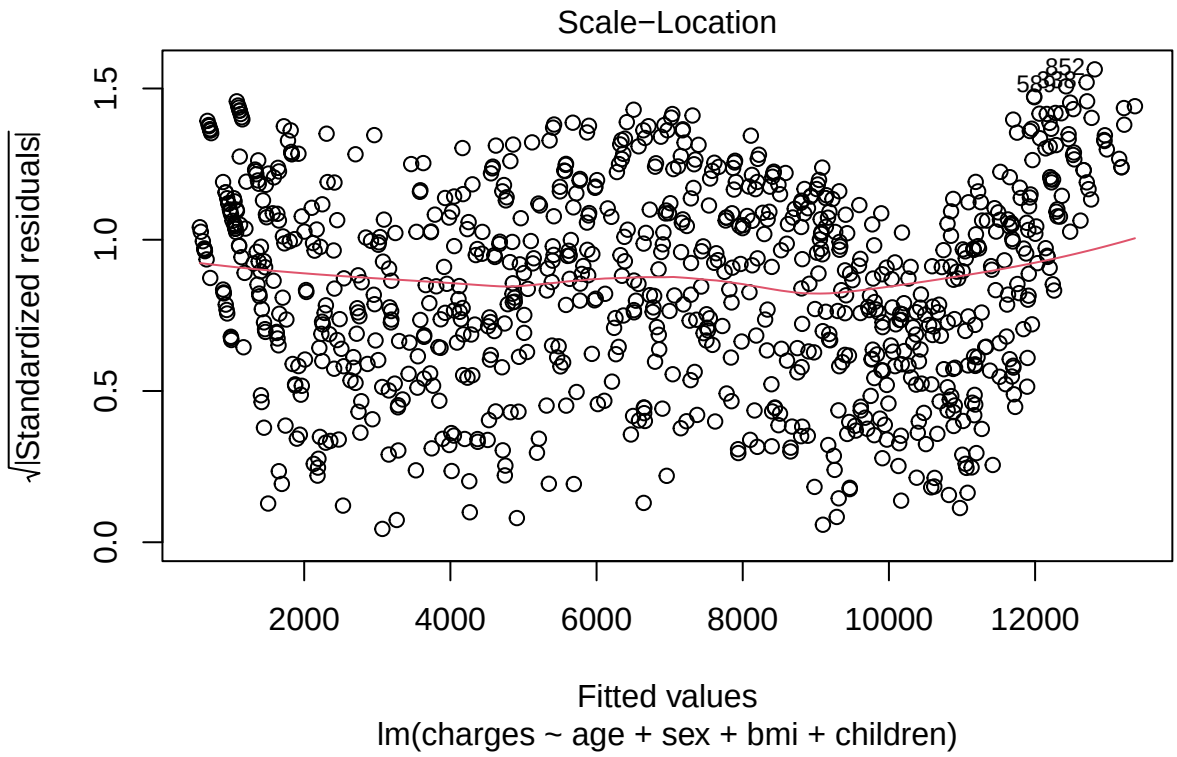
```
mod_mul_n = lm(charges~age+sex+bmi+children, plano)
```

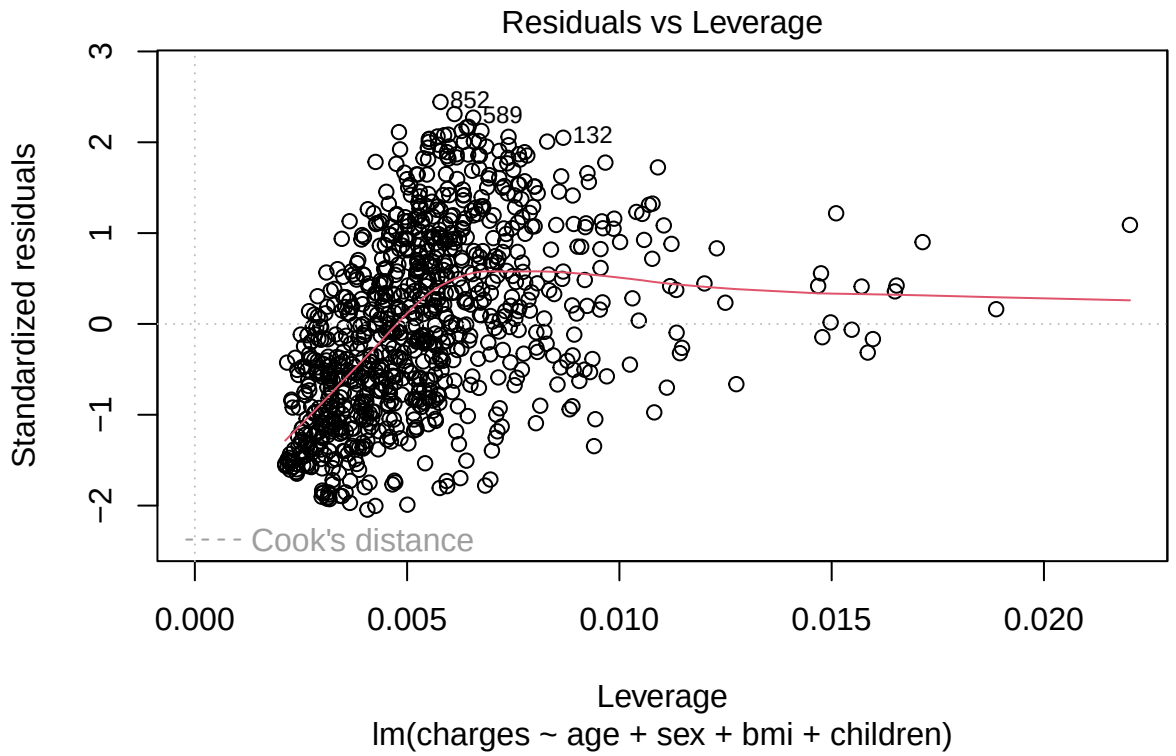
```
plot(mod_mul_n)
```



Análise gráfica




Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2000
4000
6000
8000
10000
12000
Fitted values
lm(charges ~ age + sex + bmi + children)



Podemos observar que depois da remoção de outliers conseguimos ver mais homogeneidade nos plots do modelo de regressão múltipla.

Teste de normalidade de resíduos

```
shapiro.test(mod_mul_n$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod_mul_n$residuals
## W = 0.98328, p-value = 9.096e-09
```

Podemos ver que apesar do Q-Q plot indicar uma distribuição perto da normal o teste Shapiro-Wilk confirma que a distribuição dos resíduos não é normal.

Outliers nos resíduos

```
summary(rstandard(mod_mul_n))
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -2.0444061 -0.7789558 -0.0608728  0.0004984  0.7602068  2.4437843
```

Testando os resíduos padronizados podemos ver que a mínima e a máxima dos resíduos são balanceadas e a média é bem próxima de zero, então podemos concluir que não existem outliers extremos.

Independência dos resíduos

```
durbinWatsonTest(mod_mul_n)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01055679      1.977118   0.746
##  Alternative hypothesis: rho != 0
```

Utilizando o teste de Durbin-Watson podemos testar se os resíduos são independentes através da estatística de Durbin-Watson e do p-valor. Neste caso ambos os produtos apontam para a independência dos resíduos.

Homocedasticidade

```
bptest(mod_mul_n)
```

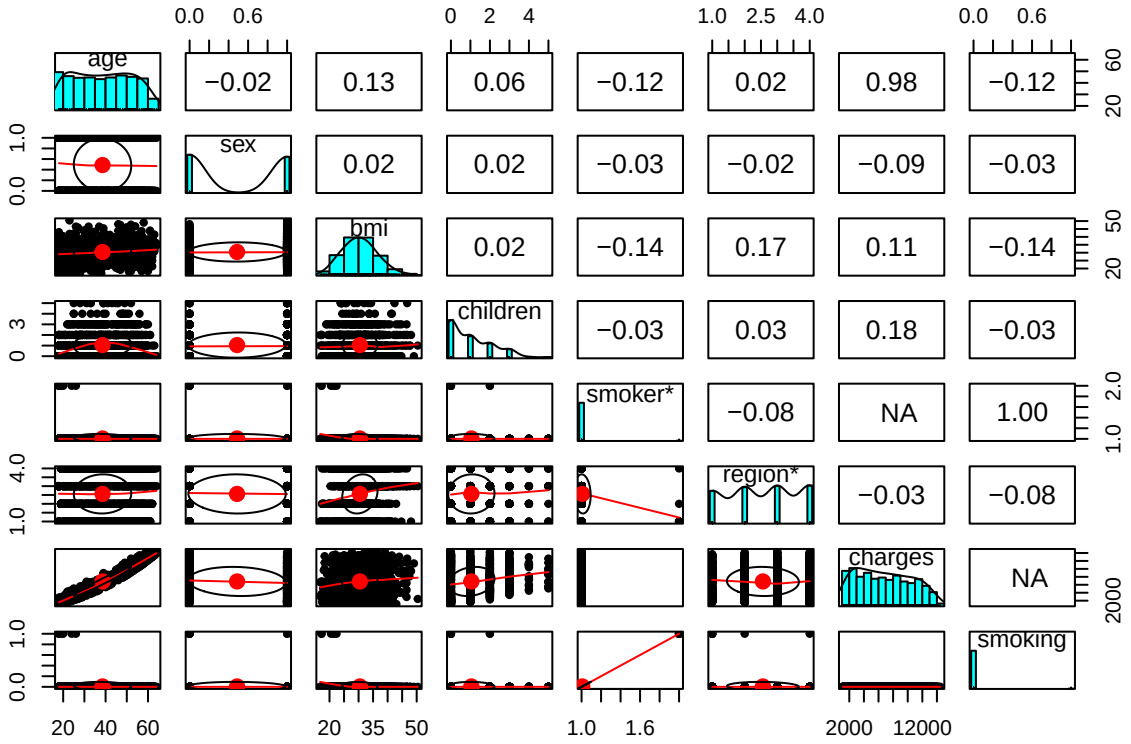
```
##
##  studentized Breusch-Pagan test
##
## data:  mod_mul_n
## BP = 55.33, df = 4, p-value = 2.771e-11
```

A homocedasticidade descreve uma situação em que o termo de erro (ou seja, o "ruído" ou perturbação aleatória na relação entre as variáveis independentes e a variável dependente) é o mesmo em todos os valores das variáveis independentes. Neste caso não há homocedasticidade, ou seja, o “ruido” da relação aleatória entre as variáves independêntes e a variável dependente é diferente entre todos os valores das variaveis independentes, demonstrado pelo p-valor do teste (p-value=2.771e-11).

Ausência de Multicolinearidade

```
pairs.panels(plano)
```

```
## Warning in cor(x, y, use = "pairwise", method = method): o desvio padrão é zero

## Warning in cor(x, y, use = "pairwise", method = method): o desvio padrão é zero

## Warning in cor(x, y, use = "pairwise", method = method): o desvio padrão é zero

## Warning in cor(x, y, use = "pairwise", method = method): o desvio padrão é zero
```




```
vif(mod_mul_n)
```

```
##      age      sex      bmi children
## 1.018883 1.001241 1.016084 1.003551
```

Uma maneira de medir a multicolinearidade é o fator de inflação da variância (VIF), que avalia o quanto a variância de um coeficiente de regressão estimado aumenta se as suas preditoras estiverem correlacionadas. Se nenhum fator estiver correlacionado, os VIFs serão todos 1. Neste caso podemos ver que todos os valores estão perto de 1, então podemos dizer que as variáveis independentes não são colineares.

Conclusão

```
summary(mod_mul_n)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children, data = plano)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1091.8  -416.2   -32.5   405.7  1304.0
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3415.909    101.344 -33.706   <2e-16 ***
## age           263.913      1.318 200.216   <2e-16 ***
## sex          -452.521     35.302 -12.819   <2e-16 ***
## bmi            -7.201      2.943  -2.447   0.0146 *
## children      408.984     14.980  27.303   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 535.1 on 916 degrees of freedom
##   (24 observations deleted due to missingness)
## Multiple R-squared:  0.9788, Adjusted R-squared:  0.9788
## F-statistic: 1.06e+04 on 4 and 916 DF,  p-value: < 2.2e-16
```

A regressão linear multipla mostrou que a idade e o número de de filhos tem efeito no preço dos planos de saúde. A cada ano que passa temos um aumento, em média, de U$263,913 (t=200,216; p<0,001) e a cada filho que a pessoa tem aumenta em média U$408,984 (t=27,303; p=<0,001).

Bibliografia e Referências

Medical Cost Personal Datasets, LANTZ, Brett - Machine Learning with R. Disponivel em : https://www.kaggle.com/datasets/mirichoi0218/insurance Último acesso em 22 de Novembro de 2022.

Sobre a Regressão Linear - Brasil | IBM - Disponivel em:https://www.ibm.com/br-pt/analytics/learn/linear-regression#:~:text=O%20que%20%C3%A9%20regress%C3%A3o%20linear,%C3%A9%20chamada%20de%20vari%C3%A1vel%20independente. Último acesso em 22 de Novembro de 2022.

Plots de diagnóstico de modelos lineares. Laboratório de Polychaeta, Departamento de Zoologia, Instituto de Biologia Universidade Federal do Rio de Janeiro. Disponivel em: https://www.labpoly.intranet.biologia.ufrj.br/diagnostico.ht
Último acesso em 22 de Novembro de 2022.

Regressão Linear Simples no R. Fernanda Peres. Youtube. Disponivel em: https://www.youtube.com/watch?v=E2bYIb81q4A&ab_channel=FernandaPeres Último acesso em 23 de Novembro de 2022.

Regressão Linear Múltipla no R. Fernanda Peres. Youtube. Disponivel em: https://www.youtube.com/watch?v=4YLOwyx_hxo&t=971s&ab_channel=FernandaPeres Último acesso em 23 de Novembro de 2022.