Ultimo Trabalho Regressao

Bruno Mesquita dos Santos

2023-11-15

# Biblioteca

if (!require('readr'))install.packages("readr");library(readr)

## Loading required package: readr

if (!require('faraway'))install.packages("faraway");library(faraway)

## Loading required package: faraway

if (!require('car'))install.packages("car");library(car)

## Loading required package: car

## Loading required package: carData

##   
## Attaching package: 'car'

## The following objects are masked from 'package:faraway':  
##   
## logit, vif

if (!require('olsrr'))install.packages("olsrr");library(olsrr)

## Loading required package: olsrr

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:faraway':  
##   
## hsb

## The following object is masked from 'package:datasets':  
##   
## rivers

if (!require('DescTools'))install.packages("DescTools");library(DescTools)

## Loading required package: DescTools

##   
## Attaching package: 'DescTools'

## The following object is masked from 'package:car':  
##   
## Recode

# Dados Consumo Cerveja

Consumo\_cerveja <- read\_csv("dados/Consumo\_cerveja.csv",   
 locale = locale(decimal\_mark = ","))

## Rows: 941 Columns: 7  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (5): Temperatura Media (C), Temperatura Minima (C), Temperatura Maxima ...  
## date (1): Data  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Funções

rm\_accent <- function(x) iconv(x, to = "ASCII//TRANSLIT")

# Limpeza

names(Consumo\_cerveja) <- tolower(  
 gsub(' ','\_',gsub('[)]','',gsub('[(]','',rm\_accent(names(Consumo\_cerveja)))))  
 )  
Consumo\_cerveja <- na.omit(Consumo\_cerveja)

# Separar Bases

data <- Consumo\_cerveja$data  
temperatura\_media\_c <- Consumo\_cerveja$temperatura\_media\_c  
temperatura\_minima\_c <- Consumo\_cerveja$temperatura\_minima\_c  
temperatura\_maxima\_c <- Consumo\_cerveja$temperatura\_maxima\_c  
precipitacao\_mm <- Consumo\_cerveja$precipitacao\_mm  
final\_de\_semana <- Consumo\_cerveja$final\_de\_semana  
consumo\_de\_cerveja\_litros <- Consumo\_cerveja$consumo\_de\_cerveja\_litros

# Análise de Regressão

## Questão A

Ajuste um modelo de regressão linear múltipla considerando todas as variáveis independentes. Verifique a multicolinearidade entre as variáveis independentes, e se há necessidade de excluir alguma delas por esse critério. Em caso afirmativo, ajuste novo modelo sem essa variável. Apresente todos os valores de Vif.

modelo\_completo = lm(consumo\_de\_cerveja\_litros ~ data+temperatura\_media\_c+  
 temperatura\_minima\_c+temperatura\_maxima\_c+  
 precipitacao\_mm+final\_de\_semana)  
  
summary(modelo\_completo)

##   
## Call:  
## lm(formula = consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c +   
## temperatura\_minima\_c + temperatura\_maxima\_c + precipitacao\_mm +   
## final\_de\_semana)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4823.8 -1794.4 -193.9 1804.6 5514.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -60845.145 19516.939 -3.118 0.001971 \*\*   
## data 4.027 1.167 3.451 0.000625 \*\*\*  
## temperatura\_media\_c -30.772 186.065 -0.165 0.868737   
## temperatura\_minima\_c 46.816 110.380 0.424 0.671725   
## temperatura\_maxima\_c 675.535 93.905 7.194 3.72e-12 \*\*\*  
## precipitacao\_mm -58.485 9.892 -5.913 7.86e-09 \*\*\*  
## final\_de\_semana 5198.875 267.020 19.470 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2298 on 358 degrees of freedom  
## Multiple R-squared: 0.7316, Adjusted R-squared: 0.7271   
## F-statistic: 162.6 on 6 and 358 DF, p-value: < 2.2e-16

summary(aov(modelo\_completo))

## Df Sum Sq Mean Sq F value Pr(>F)   
## data 1 1.335e+07 1.335e+07 2.528 0.113   
## temperatura\_media\_c 1 2.384e+09 2.384e+09 451.367 < 2e-16 \*\*\*  
## temperatura\_minima\_c 1 2.540e+08 2.540e+08 48.098 1.90e-11 \*\*\*  
## temperatura\_maxima\_c 1 3.247e+08 3.247e+08 61.481 5.18e-14 \*\*\*  
## precipitacao\_mm 1 1.752e+08 1.752e+08 33.169 1.82e-08 \*\*\*  
## final\_de\_semana 1 2.002e+09 2.002e+09 379.081 < 2e-16 \*\*\*  
## Residuals 358 1.891e+09 5.282e+06   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Anotacão:

Se o VIF for igual a 1 não há multicolinearidade entre os fatores, mas se o VIF for maior que 1, as preditoras podem estar moderadamente correlacionadas. A saída acima mostra que o VIF para os fatores de publicação e anos são cerca de 1.5, o que indica alguma correlação, mas não o suficiente para se preocupar demais com isso. Um VIF entre 5 e 10 indica alta correlação, o que pode ser problemático. E se o VIF for acima de 10, você pode assumir que os coeficientes de regressão estão mal estimados devido à multicolinearidade.

### VIF:

vif <- 1/(1-summary(modelo\_completo)[["r.squared"]])  
paste('vif =',vif)

## [1] "vif = 3.72548692094354"

Com relação ao nosso modelo temos um VIF para o modelo completo de 3.7254 o que quer dizer que ele é moderadamente correlacionadas.

vif(modelo\_completo)

## data temperatura\_media\_c temperatura\_minima\_c   
## 1.044990 24.129091 6.706766   
## temperatura\_maxima\_c precipitacao\_mm final\_de\_semana   
## 11.327898 1.039824 1.003914

Porém quando vamos para um VIF mais detalhado temos em temperatura\_media\_c, temperatura\_minima\_c, temperatura\_maxima\_c uma multicolinearidade alto, o que faz sentido devido serem dados de temperaturas tendo a mesmas representatividade, logo removeria dois ficando é com a temperatura\_media\_c.

### Removendo a multicolinearidade

modelo\_completo = lm(consumo\_de\_cerveja\_litros ~ data+temperatura\_media\_c+  
 precipitacao\_mm+final\_de\_semana)  
  
summary(modelo\_completo)

##   
## Call:  
## lm(formula = consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c +   
## precipitacao\_mm + final\_de\_semana)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6190.9 -1911.3 -251.5 1995.6 6504.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -69560.767 21094.827 -3.298 0.001073 \*\*   
## data 4.556 1.263 3.608 0.000353 \*\*\*  
## temperatura\_media\_c 854.735 41.965 20.368 < 2e-16 \*\*\*  
## precipitacao\_mm -74.587 10.681 -6.983 1.4e-11 \*\*\*  
## final\_de\_semana 5239.630 293.724 17.839 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2530 on 360 degrees of freedom  
## Multiple R-squared: 0.673, Adjusted R-squared: 0.6694   
## F-statistic: 185.2 on 4 and 360 DF, p-value: < 2.2e-16

summary(aov(modelo\_completo))

## Df Sum Sq Mean Sq F value Pr(>F)   
## data 1 1.335e+07 1.335e+07 2.087 0.149   
## temperatura\_media\_c 1 2.384e+09 2.384e+09 372.581 < 2e-16 \*\*\*  
## precipitacao\_mm 1 3.074e+08 3.074e+08 48.038 1.94e-11 \*\*\*  
## final\_de\_semana 1 2.036e+09 2.036e+09 318.216 < 2e-16 \*\*\*  
## Residuals 360 2.303e+09 6.399e+06   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Conclusão:

vif <- 1/(1-summary(modelo\_completo)[["r.squared"]])  
paste('vif =',vif)

## [1] "vif = 3.05811655234158"

Com relação ao nosso modelo temos um VIF para o modelo completo de 3.058 o que quer dizer que ele é moderadamente correlacionadas.

vif(modelo\_completo)

## data temperatura\_media\_c precipitacao\_mm final\_de\_semana   
## 1.010045 1.013175 1.000704 1.002719

Porém quando vamos para um VIF mais detalhado temos que não há multicolinearidade entre os fatores.

## Questão B

Escreva as hipóteses, decisão e conclusão do teste F para o modelo. Use o pvalor da saída do software para o teste. Faça a interpretação do coeficiente de determinação.

summary(modelo\_completo)

##   
## Call:  
## lm(formula = consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c +   
## precipitacao\_mm + final\_de\_semana)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6190.9 -1911.3 -251.5 1995.6 6504.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -69560.767 21094.827 -3.298 0.001073 \*\*   
## data 4.556 1.263 3.608 0.000353 \*\*\*  
## temperatura\_media\_c 854.735 41.965 20.368 < 2e-16 \*\*\*  
## precipitacao\_mm -74.587 10.681 -6.983 1.4e-11 \*\*\*  
## final\_de\_semana 5239.630 293.724 17.839 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2530 on 360 degrees of freedom  
## Multiple R-squared: 0.673, Adjusted R-squared: 0.6694   
## F-statistic: 185.2 on 4 and 360 DF, p-value: < 2.2e-16

### Hipótese:

### Conclusão

* Dado que nosso modelo completo possuí p-value: < 2.2e-16, podemos concluir que há qualquer nível de significância ele rejeita H0, logo pelo menos um beta difere de zero (2.2e-16 < 0.001). Temos uma F calc de 185.2 o que é bem expressivo e positivo.
* Sendo R-squared = 0.673 e Adjusted R-squared = 0.6694, o que relativamente aceitavel já que identifica a porcentagem a variabilidade no campo Y (consumo\_de\_cerveja\_litros) que é explicada pela variaveis regressoras.

## Questão C

Escreva as hipóteses, decisão e conclusão do teste t para todos os parâmetros do modelo. Decida quais variáveis não são importantes neste modelo e porque. Use 5% de significância, e considere a regra do pvalor para decisão.

### Hipótese 1:

modelo\_uma\_var = lm(consumo\_de\_cerveja\_litros ~ data)  
summary(modelo\_uma\_var)

##   
## Call:  
## lm(formula = consumo\_de\_cerveja\_litros ~ data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11094.7 -3396.6 -394.7 3309.0 12849.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -4766.744 36332.377 -0.131 0.896  
## data 1.815 2.186 0.830 0.407  
##   
## Residual standard error: 4401 on 363 degrees of freedom  
## Multiple R-squared: 0.001896, Adjusted R-squared: -0.0008538   
## F-statistic: 0.6895 on 1 and 363 DF, p-value: 0.4069

Concluimos a um nível de 5% de significância que aceita-se H0 dado que p-value = 0.4069 é maior que o nível significância de 0.05. Logo o coeficiente angular 𝛽1 é estatisticamente igual a 0, desta forma a variável data não serve para predizer os valores de y(consumo\_de\_cerveja\_litros).

### Hipótese 2:

modelo\_uma\_var = lm(consumo\_de\_cerveja\_litros ~ temperatura\_media\_c)  
summary(modelo\_uma\_var)

##   
## Call:  
## lm(formula = consumo\_de\_cerveja\_litros ~ temperatura\_media\_c)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9221.4 -2845.5 -315.3 2409.0 9392.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8528.91 1275.36 6.687 8.6e-11 \*\*\*  
## temperatura\_media\_c 794.88 59.42 13.377 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3605 on 363 degrees of freedom  
## Multiple R-squared: 0.3302, Adjusted R-squared: 0.3283   
## F-statistic: 178.9 on 1 and 363 DF, p-value: < 2.2e-16

Concluimos a um nível de 5% de significância que rejeita-se H0 dado que p-value = 2.2e-16 é menor que o nível significância de 0.05. Logo o coeficiente angular 𝛽2 é estatisticamente diferente de 0, desta forma a variável temperatura\_media\_c é útil para predizer os valores de y(consumo\_de\_cerveja\_litros).

### Hipótese 3:

modelo\_uma\_var = lm(consumo\_de\_cerveja\_litros ~ precipitacao\_mm)  
summary(modelo\_uma\_var)

##   
## Call:  
## lm(formula = consumo\_de\_cerveja\_litros ~ precipitacao\_mm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11415.1 -3252.5 -456.1 3227.2 12178.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 25758.12 245.27 105.021 < 2e-16 \*\*\*  
## precipitacao\_mm -68.65 18.24 -3.763 0.000195 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4322 on 363 degrees of freedom  
## Multiple R-squared: 0.03755, Adjusted R-squared: 0.0349   
## F-statistic: 14.16 on 1 and 363 DF, p-value: 0.0001954

Concluimos a um nível de 5% de significância que rejeita-se H0 dado que p-value = 0.0001954 é menor que o nível significância de 0.05. Logo o coeficiente angular 𝛽3 é estatisticamente diferente de 0, desta forma a variável precipitacao\_mm é útil para predizer os valores de y(consumo\_de\_cerveja\_litros).

### Hipótese 4:

modelo\_uma\_var = lm(consumo\_de\_cerveja\_litros ~ final\_de\_semana)  
summary(modelo\_uma\_var)

##   
## Call:  
## lm(formula = consumo\_de\_cerveja\_litros ~ final\_de\_semana)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9655.2 -2753.2 -61.2 2524.8 11862.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23998.2 235.2 102.04 <2e-16 \*\*\*  
## final\_de\_semana 4924.5 440.6 11.18 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3800 on 363 degrees of freedom  
## Multiple R-squared: 0.256, Adjusted R-squared: 0.254   
## F-statistic: 124.9 on 1 and 363 DF, p-value: < 2.2e-16

Concluimos a um nível de 5% de significância que rejeita-se H0 dado que p-value = 2.2e-16 é menor que o nível significância de 0.05. Logo o coeficiente angular 𝛽4 é estatisticamente diferente de 0, desta forma a variável final\_de\_semana é útil para predizer os valores de y(consumo\_de\_cerveja\_litros).

## Questão D

Utilize o método Backward de seleção de variáveis para encontrar o melhor conjunto de preditoras para essa variável y. Escreva a equação do modelo ajustado e a interpretação, para todas as variáveis que restaram no modelo. Considere 5% de significância. Apresente os valores dos testes em cada passo, com a interpretação.

Modelo completo -> consumo\_de\_cerveja\_litros ~ data+temperatura\_media\_c+precipitacao\_mm+final\_de\_semana

### Removendo data

modelo\_sem\_data = lm(consumo\_de\_cerveja\_litros ~ temperatura\_media\_c+precipitacao\_mm+final\_de\_semana)  
anova(modelo\_completo,modelo\_sem\_data)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c + precipitacao\_mm +   
## final\_de\_semana  
## Model 2: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + precipitacao\_mm +   
## final\_de\_semana  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 360 2303474722   
## 2 361 2386755919 -1 -83281197 13.016 0.0003526 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Removendo temperatura\_media\_c

modelo\_sem\_data = lm(consumo\_de\_cerveja\_litros ~   
 data++precipitacao\_mm+final\_de\_semana)  
anova(modelo\_completo,modelo\_sem\_data)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c + precipitacao\_mm +   
## final\_de\_semana  
## Model 2: consumo\_de\_cerveja\_litros ~ data + +precipitacao\_mm + final\_de\_semana  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 360 2303474722   
## 2 361 4957860285 -1 -2654385563 414.84 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Removendo precipitacao\_mm

modelo\_sem\_data = lm(consumo\_de\_cerveja\_litros ~   
 data+temperatura\_media\_c+final\_de\_semana)  
anova(modelo\_completo,modelo\_sem\_data)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c + precipitacao\_mm +   
## final\_de\_semana  
## Model 2: consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c + final\_de\_semana  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 360 2303474722   
## 2 361 2615517078 -1 -312042356 48.768 1.398e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Removendo final\_de\_semana

modelo\_sem\_data = lm(consumo\_de\_cerveja\_litros ~   
 data+temperatura\_media\_c+precipitacao\_mm)  
anova(modelo\_completo,modelo\_sem\_data)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c + precipitacao\_mm +   
## final\_de\_semana  
## Model 2: consumo\_de\_cerveja\_litros ~ data + temperatura\_media\_c + precipitacao\_mm  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 360 2303474722   
## 2 361 4339592657 -1 -2036117936 318.22 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

qf(0.95,1,361)

## [1] 3.867347

### Conclusão:

Dado que nosso menor F calc foi 13.016, devemos compara com F Tab de parametro (0.95,1,361). Logo temos que A 95% de confiança, escolhemos o Fcalc mínimo para comparar com Ftab, Fmin(13.016) > Ftab(3.867347), nesse casos rejeita-se Ho, e conclui-se que não se pode tirar a variável data. Então mantém o modelo com modelo completo.

## Questão E

Utilize o método Forward de seleção de variáveis para encontrar o melhor conjunto de preditoras para essa variável y. Escreva a equação do modelo ajustado e compare com o modelo obtido em (d).

Me basendo no F da questão C o melhor modelo reduzido para se começar é:

modelo\_reduzido = lm(consumo\_de\_cerveja\_litros ~ temperatura\_media\_c)

### Adicionar data

modelo\_mais\_data = lm(consumo\_de\_cerveja\_litros ~ temperatura\_media\_c+data)  
summary(aov(modelo\_mais\_data))

## Df Sum Sq Mean Sq F value Pr(>F)   
## temperatura\_media\_c 1 2.326e+09 2.326e+09 181.188 <2e-16 \*\*\*  
## data 1 7.143e+07 7.143e+07 5.564 0.0189 \*   
## Residuals 362 4.647e+09 1.284e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(modelo\_reduzido,modelo\_mais\_data)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c  
## Model 2: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + data  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 363 4718394688   
## 2 362 4646965029 1 71429659 5.5644 0.01886 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Adicionar precipitacao\_mm

modelo\_mais\_precipitacao\_mm = lm(consumo\_de\_cerveja\_litros ~   
 temperatura\_media\_c+precipitacao\_mm)  
anova(modelo\_reduzido,modelo\_mais\_precipitacao\_mm)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c  
## Model 2: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + precipitacao\_mm  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 363 4718394688   
## 2 362 4413993821 1 304400867 24.965 9.107e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Adicionar final\_de\_semana

modelo\_mais\_final\_de\_semana = lm(consumo\_de\_cerveja\_litros ~   
 temperatura\_media\_c+final\_de\_semana)  
anova(modelo\_reduzido,modelo\_mais\_final\_de\_semana)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c  
## Model 2: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + final\_de\_semana  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 363 4718394688   
## 2 362 2695619426 1 2022775262 271.64 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Conclusão:

O Fmax é igual 271.64 sendo final\_de\_semana o mais provável de ser adicionado.

### Validando a variavel

qf(0.95, 1 , 362)

## [1] 3.867275

#### Conclusão:

Fmax(271.64) > Ftab(3.867275) aceitamos o modelo completo como consumo\_de\_cerveja\_litros ~ temperatura\_media\_c+final\_de\_semana

### Novo Modelo Reduzido

modelo\_reduzido = lm(consumo\_de\_cerveja\_litros ~   
 temperatura\_media\_c+final\_de\_semana)

### Adicionar data

modelo\_mais\_data = lm(consumo\_de\_cerveja\_litros ~   
 temperatura\_media\_c+final\_de\_semana+data)  
summary(aov(modelo\_mais\_data))

## Df Sum Sq Mean Sq F value Pr(>F)   
## temperatura\_media\_c 1 2.326e+09 2.326e+09 321.03 < 2e-16 \*\*\*  
## final\_de\_semana 1 2.023e+09 2.023e+09 279.19 < 2e-16 \*\*\*  
## data 1 8.010e+07 8.010e+07 11.06 0.000975 \*\*\*  
## Residuals 361 2.616e+09 7.245e+06   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(modelo\_reduzido,modelo\_mais\_data)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + final\_de\_semana  
## Model 2: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + final\_de\_semana +   
## data  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 362 2695619426   
## 2 361 2615517078 1 80102349 11.056 0.0009747 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Adicionar precipitacao\_mm

modelo\_mais\_precipitacao\_mm = lm(consumo\_de\_cerveja\_litros ~ temperatura\_media\_c+final\_de\_semana+precipitacao\_mm)  
anova(modelo\_reduzido,modelo\_mais\_precipitacao\_mm)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + final\_de\_semana  
## Model 2: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + final\_de\_semana +   
## precipitacao\_mm  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 362 2695619426   
## 2 361 2386755919 1 308863507 46.716 3.508e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Conclusão:

O Fmax é igual 46.716 sendo precipitacao\_mm o mais provável de ser adicionado.

### Validando a variavel

qf(0.95, 1 , 361)

## [1] 3.867347

#### Conclusão:

Fmax(46.716) > Ftab(3.867347) aceitamos o modelo completo como consumo\_de\_cerveja\_litros ~ temperatura\_media\_c+final\_de\_semana+precipitacao\_mm

### Novo Modelo Reduzido

modelo\_reduzido = lm(consumo\_de\_cerveja\_litros ~ temperatura\_media\_c+final\_de\_semana+precipitacao\_mm)

### Adicionar data

modelo\_mais\_data = lm(consumo\_de\_cerveja\_litros ~ temperatura\_media\_c+final\_de\_semana+precipitacao\_mm+data)  
summary(aov(modelo\_mais\_data))

## Df Sum Sq Mean Sq F value Pr(>F)   
## temperatura\_media\_c 1 2.326e+09 2.326e+09 363.50 < 2e-16 \*\*\*  
## final\_de\_semana 1 2.023e+09 2.023e+09 316.13 < 2e-16 \*\*\*  
## precipitacao\_mm 1 3.089e+08 3.089e+08 48.27 1.75e-11 \*\*\*  
## data 1 8.328e+07 8.328e+07 13.02 0.000353 \*\*\*  
## Residuals 360 2.303e+09 6.399e+06   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(modelo\_reduzido,modelo\_mais\_data)

## Analysis of Variance Table  
##   
## Model 1: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + final\_de\_semana +   
## precipitacao\_mm  
## Model 2: consumo\_de\_cerveja\_litros ~ temperatura\_media\_c + final\_de\_semana +   
## precipitacao\_mm + data  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 361 2386755919   
## 2 360 2303474722 1 83281197 13.016 0.0003526 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Conclusão:

O Fmax é igual 13.016 sendo data provável de ser adicionado.

### Validando a variavel

qf(0.95, 1 , 360)

## [1] 3.867419

#### Conclusão:

Fmax(13.016) > Ftab(3.867419) aceitamos o modelo completo como consumo\_de\_cerveja\_litros ~ temperatura\_media\_c+final\_de\_semana+precipitacao\_mm+data

### Conclusão:

O Modelo da questão D e E chegaram no mesmo resultado.

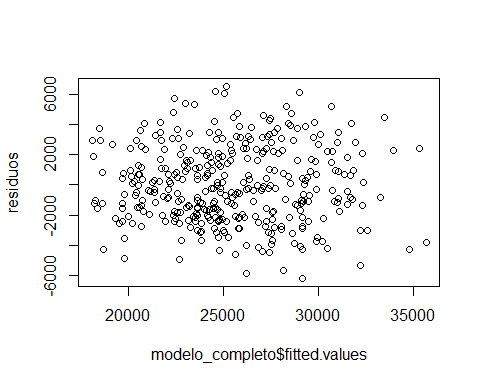
## Questão F

Escolha um dos modelos ajustados em (d) ou (e) e faça a análise completa dos resíduos do modelo, verificando todas as pressuposições do modelo. Apresente os gráficos dos resíduos padronizados contra: y estimado, variáveis independentes, ordem das observações. Apresente todas as conclusões. Complemente as conclusões com os testes de Shapiro Wilk, Durbin Watson. Discuta sobre a necessidade de transformação na variável resposta, ou de usar mínimos quadrados ponderados, justificando.

residuos <- residuals(modelo\_completo)

### Gráfico de Dispersão dos Resíduos vs. Valores Ajustados

plot(modelo\_completo$fitted.values, residuos)



#### Conclusões:

Os pontos distribuem-se de forma aleatória, indicando assim homocedasticidade e homogênea, ou seja, a variabilidade dos resíduos é constante em vários níveis de predição. Se houvesse padrões, como um aumento nos resíduos à medida que os valores ajustados aumentam, isso sugeriria a presença de heterocedasticidade. Tal cenário poderia indicar que o residuo do modelo não possui variabilidade costante.

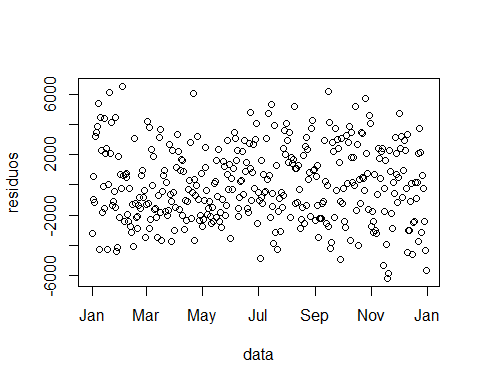
(Solução: fazer transformação em Y ou usar Mínimos Quadrados Ponderados.)

A homocedasticidade é provavelmente violada se…

1. Se os resíduos aumentam ou diminuem com os valores ajustados.
2. Se os pontos formam uma curva ao redor de zero e não estão dispostos aleatoriamente.
3. Poucos pontos no gráfico ficam muito distantes dos demais.

### Gráfico de Dispersão dos Resíduos vs. Data

plot(data, residuos)

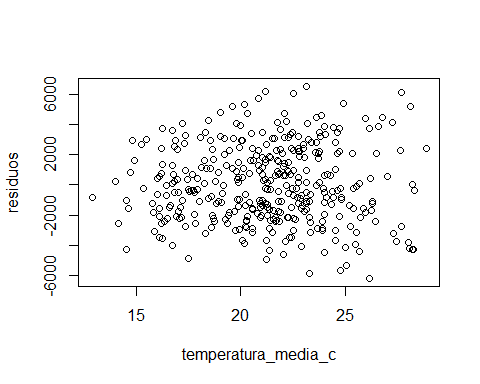


#### Conclusões:

1. Os resíduos se distribuem aleatoriamente em torno de zero.
2. Não se observa nenhum padrão.
3. Isso indica que: a variância é constante; e a relação entre as variáveis é linear.

### Gráfico de Dispersão dos Resíduos vs. Temperatura\_media\_c

plot(temperatura\_media\_c, residuos)



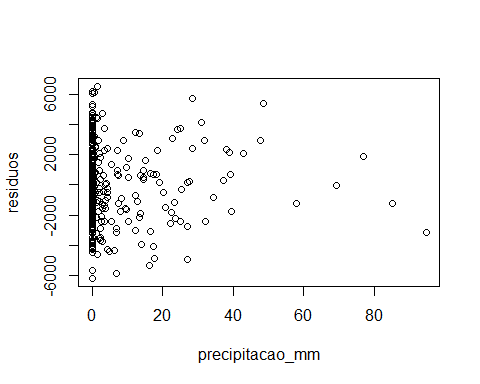
#### Conclusões:

Tem uma leve diferença entre os valores de 15 a 20 e depois tem uma aumento, pode repesentar uma variância não constante(Vereficar se é significativo com a professora).

1. Os resíduos se distribuem aleatoriamente em torno de zero.
2. Não se observa nenhum padrão.
3. Isso indica que: a variância é constante; e a relação entre as variáveis é linear.

### Gráfico de Dispersão dos Resíduos vs. precipitacao\_mm

plot(precipitacao\_mm , residuos)



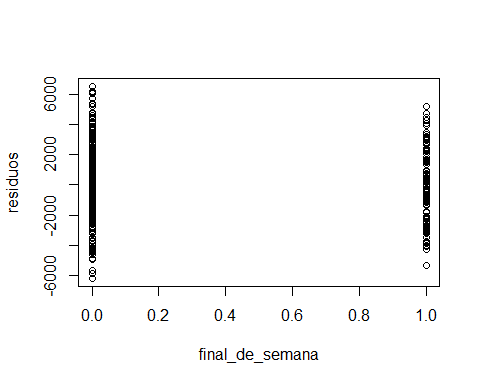
Conclusões:

Possui uma concentraçao no zero e tem uma leve diferença entre os valores de 0 a 20, em seguida tem uma diminuição, pode repesentar uma variância não constante (Vereficar se é significativo com a professora).

1. Os resíduos se distribuem aleatoriamente em torno de zero.
2. Não se observa nenhum padrão.
3. Isso indica que: a variância é constante; e a relação entre as variáveis é linear.

### Gráfico de Dispersão dos Resíduos vs. final\_de\_semana

plot(final\_de\_semana , residuos)

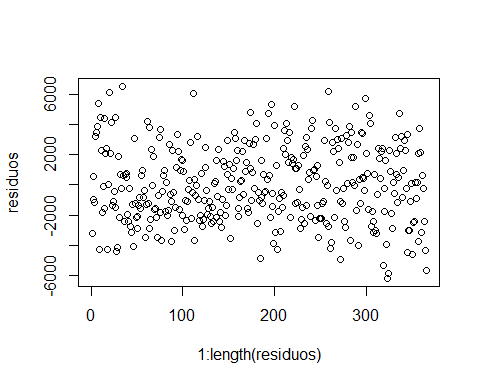


#### Conclusões:

1. Os resíduos se distribuem aleatoriamente em torno de zero.
2. Não se observa nenhum padrão.
3. Isso indica que: a variância é constante; e a relação entre as variáveis é linear.

# Gráfico de dispersão dos resíduos vs. ordem das observações

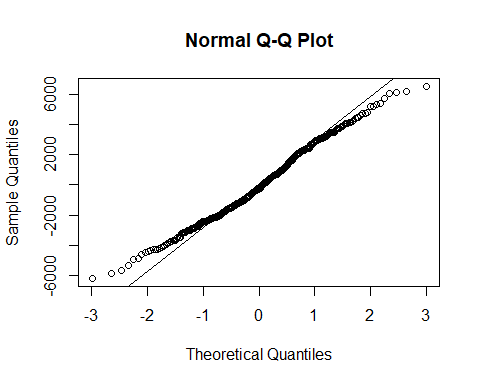
plot(1:length(residuos), residuos)



A dispersão aleatória dos resíduos ao longo da ordem das observações mostra que o modelo está bem ajustado aos dados. Não há indicações de que o modelo esteja cometendo erros de maneira sistemática ou previsível.

### Gráfico QQ-plots

qqnorm(residuos)  
qqline(residuos)



#### Conclusões:

Podemos ver que os resíduos tendem a se desviar um pouco da linha perto das caudas, o que pode indicar que eles não estão normalmente distribuídos.

### Teste de normalidade dos resíduos (Shapiro-Wilk)

shapiro.test(residuos)

##   
## Shapiro-Wilk normality test  
##   
## data: residuos  
## W = 0.98964, p-value = 0.01106

#### Hipótese:

##### Conclusões:

Como resultado, o teste retornará a estatística W, que terá um valor de significância associada, o valor-p. Para dizer que uma distribuição é normal, o valor p precisa ser maior do que 0,05. Logo rejeitar a hipótese nula concluimos que os erros não tem distribuição normal.

### Teste de autocorrelação dos resíduos (Durbin-Watson)

durbinWatsonTest(modelo\_completo)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.03100611 1.919447 0.384  
## Alternative hypothesis: rho != 0

#### Hipótese:

##### Conclusões:

Ele faz uma estatística que testa se eles são independentes. Aí, o p-valor que a gente compara é com 0.05, nesse caso, o p-valor não foi menor do que 0.05. Logo, a gente não tem evidência suficiente para rejeitar a hipótese nula. Se eu não consigo rejeitar é porque eu não tenho evidências para dizer que elas não são independentes. Então, eu aceito que são independentes e não há autocorrelação significativa.

### Conclusões Geral:

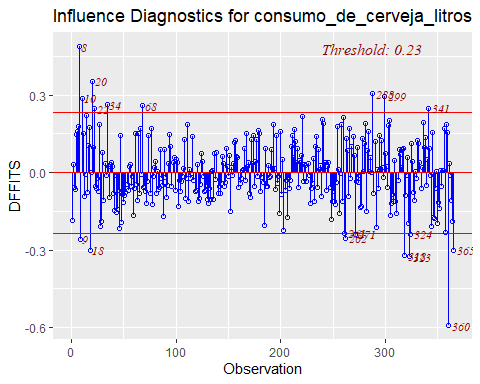
É possível que seja necessário transformar a variável de resposta para abordar a falta de normalidade nos resíduos, facilitando assim a aprovação no teste de Shapiro-Wilk. No entanto, em relação à aplicação de mínimos quadrados ponderados, a evidência de falta de homocedasticidade nos resíduos não é conclusiva, necessitando de mais testes.

## Questão G

Realize a análise de diagnóstico do modelo, apresentando os gráficos de todas as medidas estudadas (DFFITS, DFBETAS, Distância de Cook, leverage, resíduos estudentizados). Interprete cada uma usando o gráfico e CALCULE o ponto de corte visto no livro para comparar.

### DFFIT

ols\_plot\_dffits(modelo\_completo)



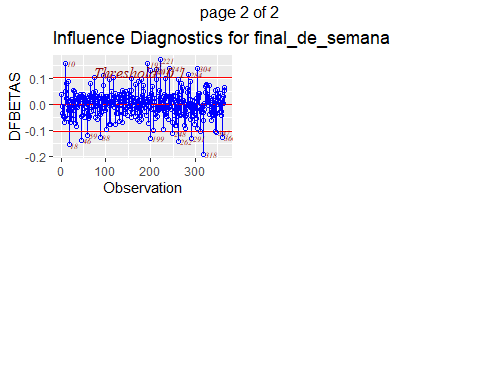
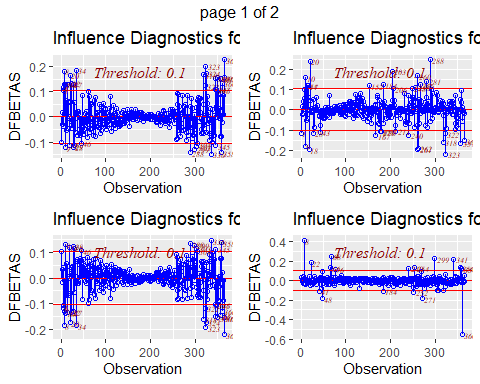
#### Conclusões:

DFFIT - diferença nos ajustes, é usado para identificar pontos de dados influentes. Ele quantifica o número de desvios padrão que o valor ajustado muda quando o i-ésimo ponto de dados é omitido.

Pontos ultrapassao a linha podem ser considerados pontos influentes.

### DFBETAS

ols\_plot\_dfbetas(modelo\_completo)

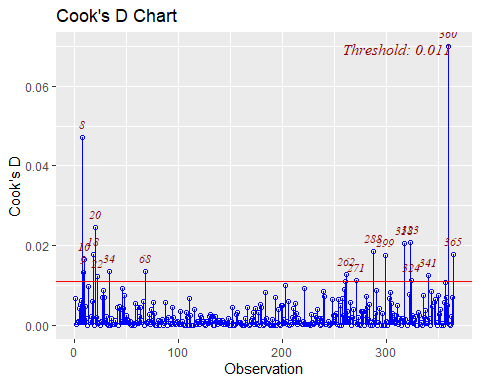


#### Conclusões:

O gráfico sugere que vários pontos têm uma influência considerável na estimativa da inclinação para x. Neste caso, a inspeção visual da relação linear entre x e y é possível. O painel mostra a influência de cada observação nas estimativas dos quatro coeficientes de regressão. As estatísticas são padronizadas para que todos os gráficos possam utilizar a mesma escala vertical. As linhas horizontais são desenhadas em ±2/sqrt(n). As observações são chamadas de influentes se tiverem uma estatística DFBETA que exceda esse valor.

### Cook’s distance

ols\_plot\_cooksd\_chart(modelo\_completo)



#### Conclusões:

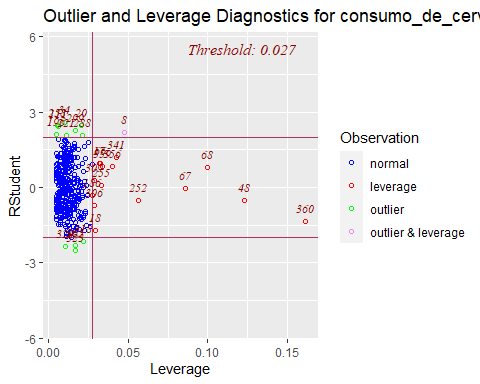
Gráfico de barras da distância de Cook para detectar observações que influenciam fortemente os valores ajustados do modelo. A distância de Cook foi introduzida pelo estatístico americano R Dennis Cook em 1977. É usada para identificar pontos de dados influentes. Depende tanto do resíduo quanto da alavancagem, ou seja, leva em consideração tanto o valor x quanto o valor y da observação. Passos para calcular a distância de Cook:

* exclua as observações uma de cada vez.
* reajuste o modelo de regressão no restante (n−1) observações
* examine quanto todos os valores ajustados mudam quando a i-ésima observação é excluída.

Um ponto de dados com distância Cook grande indica que o ponto de dados influencia fortemente os valores ajustados.

### Studentized Residuals vs Leverage Plot

ols\_plot\_resid\_lev(modelo\_completo)

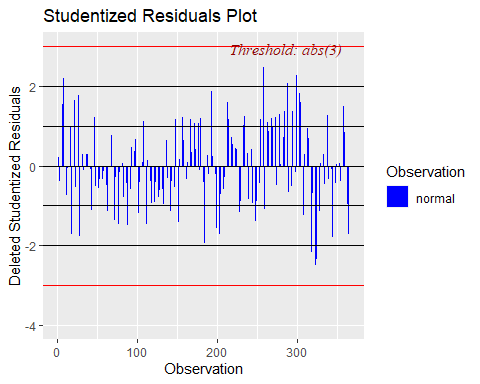


#### Conclusões:

Gráfico para detecção de outliers e/ou observações com alta influencia

### Resíduos Studentizados

ols\_plot\_resid\_stud(modelo\_completo)



#### Conclusões:

Gráfico para detecção de outliers. Resíduos excluídos estudantis (ou resíduos estudantis externamente) são os resíduos excluídos divididos por seu desvio padrão estimado. Os resíduos estudantis serão mais eficazes para detectar observações Y periféricas do que os resíduos padronizados. Se uma observação tiver um resíduo estudantil externamente maior que 3 (em valor absoluto), podemos chamá-la de outlier.

# Regressão Logística

## Questão A

Escreva a equação do modelo ajustado, capaz de descrever o relacionamento existente entre a ocorrência de sinistro e as variáveis em estudo.

sinistro <- read.delim("dados/dadosSinistro.txt")

modelo=glm(Sinistro~Idade+ECivil+Sexo,data = sinistro,  
 family=binomial(link="logit"))  
summary(modelo)

##   
## Call:  
## glm(formula = Sinistro ~ Idade + ECivil + Sexo, family = binomial(link = "logit"),   
## data = sinistro)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3375 -0.3070 -0.1441 0.2496 2.0702   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.78103 1.94470 1.944 0.05186 .   
## Idade -0.18957 0.07228 -2.623 0.00872 \*\*  
## ECivil -3.62511 1.65160 -2.195 0.02817 \*   
## Sexo 3.70268 1.67313 2.213 0.02690 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 42.541 on 35 degrees of freedom  
## Residual deviance: 20.990 on 32 degrees of freedom  
## AIC: 28.99  
##   
## Number of Fisher Scoring iterations: 6

## Questão B

Escreva as hipóteses, decisão e conclusão do teste F para o modelo. Use o pvalor da saída do software para o teste. Faça a interpretação do coeficiente de determinação.

#### Conclusões:

O modelo apresenta Pr(>|z|) é menos que o nível significância de 0.05. Logo os parâmetros são significativos. O modelo possui um Intercepto positivo o que siginifica que ele tem uma curva crescente.

PseudoR2(modelo,which="all")

## McFadden McFaddenAdj CoxSnell Nagelkerke AldrichNelson   
## 0.5065897 0.3185342 0.4504355 0.6497555 0.3744641   
## VeallZimmermann Efron McKelveyZavoina Tjur AIC   
## 0.6913541 0.5077785 0.7203881 0.5240548 28.9899924   
## BIC logLik logLik0 G2   
## 35.3240682 -10.4949962 -21.2703209 21.5506493

Quanto menores, melhor é o modelo.

O AIC estima a quantidade relativa de informação perdida por um determinado modelo: quanto menos informações um modelo perde, maior a qualidade desse modelo e menor a pontuação AIC.

## Questão C

Teste a significância do modelo por meio do Teste de razão de verossimilhança, apresentando as hipóteses e conclusão. Teste a significância dos parâmetros do modelo e apresente as conclusões, em termos das variáveis. Use 1% de significância.

paste0('TRV = ',modelo[["null.deviance"]]-modelo[["deviance"]],' e ',   
 qchisq(0.95,3), ' há 95% e 3p')

## [1] "TRV = 21.5506493356017 e 7.81472790325118 há 95% e 3p"

#### Conclusões:

Dado que o TRV é maior que qui quadrado há 95% e com 3 parametros o modelo como um todo é significativo.

Lembrando:

Considere dois modelos um “completo” (Mc ) e o outro “restrito” (Mr) quanto à quantidade de parâmetros.Então a estatística de teste é:

TRV = 2 [log(L(Mc )) − log(L(Mr))].

Sob a hipóotese de igualdade entre os dois modelos, TRV tem distribuição aproximada de Qui-Quadrado com graus de liberdade igual ao número de restrições (diferença entre as quantidades de parâmetros).

## Questão D

Interprete as razões de chance e os intervalos de confiança obtidas para todos os parâmetros;

### Intervalos de confiança

ICbeta=confint.default(modelo,level=0.95)  
ICbeta

## 2.5 % 97.5 %  
## (Intercept) -0.03051846 7.59258396  
## Idade -0.33123689 -0.04790661  
## ECivil -6.86218543 -0.38803500  
## Sexo 0.42339383 6.98195617

#### Conclusões:

Isso significa que você tem 95% de confiança de que a verdadeiro valor de beta/intercepto esteja entre os valores acima. O valor zero pertence ao intervalo, logo os parâmetros de beta sao estatisticamente igual a zero.

### odds ratio – OR (chance)

OR=exp(modelo$coefficients)  
OR

## (Intercept) Idade ECivil Sexo   
## 43.86131610 0.82731335 0.02664616 40.55564586

#### Conclusões:

A chance de haver sinistro(1) é de cerca de 0.83 vezes menos para cada ano de acrescentado na idade; A chance de haver sinistro(1) é de cerca de 0.027 vezes menos quando estado civil igual a Solteiro(1), ou seja, as chances de a haver sinistro é de 97.3% menor se for Solteiro; A chance de haver sinistro(1) é de cerca de 40.55 vezes maior se o sexo for Feminino(1).

### Intervalos de confiança para as razões de chance

ICOR=exp(ICbeta);  
round((cbind(OR, ICOR)),3)

## OR 2.5 % 97.5 %  
## (Intercept) 43.861 0.970 1983.432  
## Idade 0.827 0.718 0.953  
## ECivil 0.027 0.001 0.678  
## Sexo 40.556 1.527 1077.023

#### Conclusões:

Isso significa que você tem 95% de confiança de que a verdadeiro valor de odds ratio esteja entre os valores acima.

## Questão E

Estime a probabilidade de sinistro associada a um cliente do sexo masculino, casado com 25 anos de idade

p\_sexo = 0  
p\_ec = 0  
p\_idade = 25  
  
razaopiest = exp(modelo[["coefficients"]]['(Intercept)'][[1]] +   
 (modelo[["coefficients"]]['Idade'][[1]] \* p\_idade) +  
 (modelo[["coefficients"]]['ECivil'][[1]] \* p\_ec) +  
 (modelo[["coefficients"]]['Sexo'][[1]] \* p\_sexo)  
 )  
  
paste0('Estimação da probabilidade de sinistro = ',  
 (razaopiest /(1 + razaopiest) \* 100))

## [1] "Estimação da probabilidade de sinistro = 27.7226495853509"

## Questão F

Para o mesmo cliente citado no item anterior, qual a probabilidade de sinistro se ele for solteiro?

p\_sexo = 0  
p\_ec = 1  
p\_idade = 25  
  
razaopiest = exp(modelo[["coefficients"]]['(Intercept)'][[1]] +   
 (modelo[["coefficients"]]['Idade'][[1]] \* p\_idade) +  
 (modelo[["coefficients"]]['ECivil'][[1]] \* p\_ec) +  
 (modelo[["coefficients"]]['Sexo'][[1]] \* p\_sexo)  
 )  
  
paste0('Estimação da probabilidade de sinistro = ',  
 (razaopiest /(1 + razaopiest) \* 100))

## [1] "Estimação da probabilidade de sinistro = 1.0116983059195"

## Questão G

Compare os resultados obtidos nos dois itens anteriores e reflita sobre as estratégias que poderiam ser adotadas pela companhia para atrair novos clientes.

#### Conclusões:

Com base na estimação anterioir e para os parametros colocados a prova podemos dizer que para a companhia seria benefico focar em atrai solteiros dados que a chance de haver sinistro diminui muito.