Previsão tempo de nado

Pedro Sobreiro

Abstract

TODO.

1 Previsão do tempo de nado

1.1 Descrição variáveis

- N (ID)
- Sexo
- Idade
- Altura (m)
- Peso (kg)
- Envergadura (m)
- 50m L Tempo (s)
- 50m L FG (ciclos/min) 5-20m
- 50m L DC (m)
- 50m L IN (equação)
- 50m L Tempo Viragem (s) 5+10m
- MI Média Distância (m)
- MS Média Distância (m)

Geral Masculinos Femininos

1.2 Abrir ficheiros

```
library(janitor)
library(writexl)
library(xlsx)

print(caminho)
```

```
## [1] "C:/nuvem/OneDrive - Instituto Politécnico de Santarém/investigacao/1.emCurso/7
```

```
dados<-read.xlsx("../dados/dadosNadadores_parte2.xlsx",sheetName = 'Geral',to.data.fra
names(dados)
```

```
##  [1] "N"                                "Sexo"
##  [3] "Idade"                            "Altura..m."
##  [5] "Peso..kg."                        "Envergadura..m."
```

```
##  [7] "X50m.L.Tempo..s."                "X50m.L.FG..ciclos.min..5.20m"
##  [9] "X50m.L.DC..m."                   "X50m.L.IN..equaÃ.Ã.o."
## [11] "X50m.L.Tempo.Viragem..s..5.10m" "MI.MÃ.dia.DistÃ.ncia..m."
## [13] "MS.MÃ.dia.DistÃ.ncia..m."        "to.data.frame"
```

```
head(dados)
```

```
##   N Sexo Idade Altura..m. Peso..kg. Envergadura..m. X50m.L.Tempo..s.
## 1 1    2    14       1.69      54.0            1.69            31.04
## 2 2    2    14       1.60      53.4            1.66            31.50
## 3 3    2    14       1.58      55.0            1.67            30.70
## 4 4    2    14       1.56      48.5            1.65            32.42
## 5 5    2    14       1.64      53.0            1.68            31.66
## 6 6    2    14       1.63      56.4            1.59            30.34
##   X50m.L.FG..ciclos.min..5.20m X50m.L.DC..m. X50m.L.IN..equaÃ.Ã.o.
## 1                        55.05      1.708336              2.677642
## 2                        54.88      1.706495              2.663624
## 3                        41.15      2.252441              3.479569
## 4                        51.87      1.757960              2.671672
## 5                        46.00      1.992385              3.043357
## 6                        46.00      2.097022              3.371418
##   X50m.L.Tempo.Viragem..s..5.10m MI.MÃ.dia.DistÃ.ncia..m.
## 1                           9.21                1.550000
## 2                           9.14                1.646667
## 3                           9.07                1.530000
## 4                           9.75                1.516667
## 5                           9.28                1.590000
## 6                           8.97                1.480000
##   MS.MÃ.dia.DistÃ.ncia..m. to.data.frame
## 1                 3.240000          TRUE
## 2                 3.413333          TRUE
## 3                 3.650000          TRUE
## 4                 3.083333          TRUE
## 5                 3.383333          TRUE
## 6                 3.033333          TRUE
```

```
dados<-clean_names(dados)

names(dados)
```

```
##  [1] "n"                              "sexo"
##  [3] "idade"                          "altura_m"
##  [5] "peso_kg"                        "envergadura_m"
##  [7] "x50m_l_tempo_s"                 "x50m_l_fg_ciclos_min_5_20m"
##  [9] "x50m_l_dc_m"                    "x50m_l_in_equa_a_a_o"
## [11] "x50m_l_tempo_viragem_s_5_10m"   "mi_ma_dia_dist_a_ncia_m"
## [13] "ms_ma_dia_dist_a_ncia_m"        "to_data_frame"
```

```
dados<-dados[,-c(14)]
names(dados)
```

```
##  [1] "n"                                "sexo"
##  [3] "idade"                            "altura_m"
##  [5] "peso_kg"                          "envergadura_m"
##  [7] "x50m_l_tempo_s"                   "x50m_l_fg_ciclos_min_5_20m"
##  [9] "x50m_l_dc_m"                      "x50m_l_in_equa_a_a_o"
## [11] "x50m_l_tempo_viragem_s_5_10m" "mi_ma_dia_dist_a_ncia_m"
## [13] "ms_ma_dia_dist_a_ncia_m"
```

```
str(dados)
```

```
## 'data.frame':    2037 obs. of  13 variables:
##  $ n                           : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ sexo                        : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ idade                       : num  14 14 14 14 14 14 14 14 14 14 ...
##  $ altura_m                    : num  1.69 1.6 1.58 1.56 1.64 1.63 1.63 1.64 1.57 1
##  $ peso_kg                     : num  54 53.4 55 48.5 53 56.4 56.9 52.7 56.9 55.1
##  $ envergadura_m               : num  1.69 1.66 1.67 1.65 1.68 1.59 1.63 1.65 1.68
##  $ x50m_l_tempo_s              : num  31 31.5 30.7 32.4 31.7 ...
##  $ x50m_l_fg_ciclos_min_5_20m  : num  55 54.9 41.1 51.9 46 ...
##  $ x50m_l_dc_m                 : num  1.71 1.71 2.25 1.76 1.99 ...
##  $ x50m_l_in_equa_a_a_o        : num  2.68 2.66 3.48 2.67 3.04 ...
##  $ x50m_l_tempo_viragem_s_5_10m: num  9.21 9.14 9.07 9.75 9.28 8.97 9.38 9.11 9.03
##  $ mi_ma_dia_dist_a_ncia_m     : num  1.55 1.65 1.53 1.52 1.59 ...
##  $ ms_ma_dia_dist_a_ncia_m     : num  3.24 3.41 3.65 3.08 3.38 ...
```

```
dados<-remove_empty(dados,which = c("rows"))

# vamos retirar N
dados<-dados[,-c(1)]
```

1.3 Descritivas

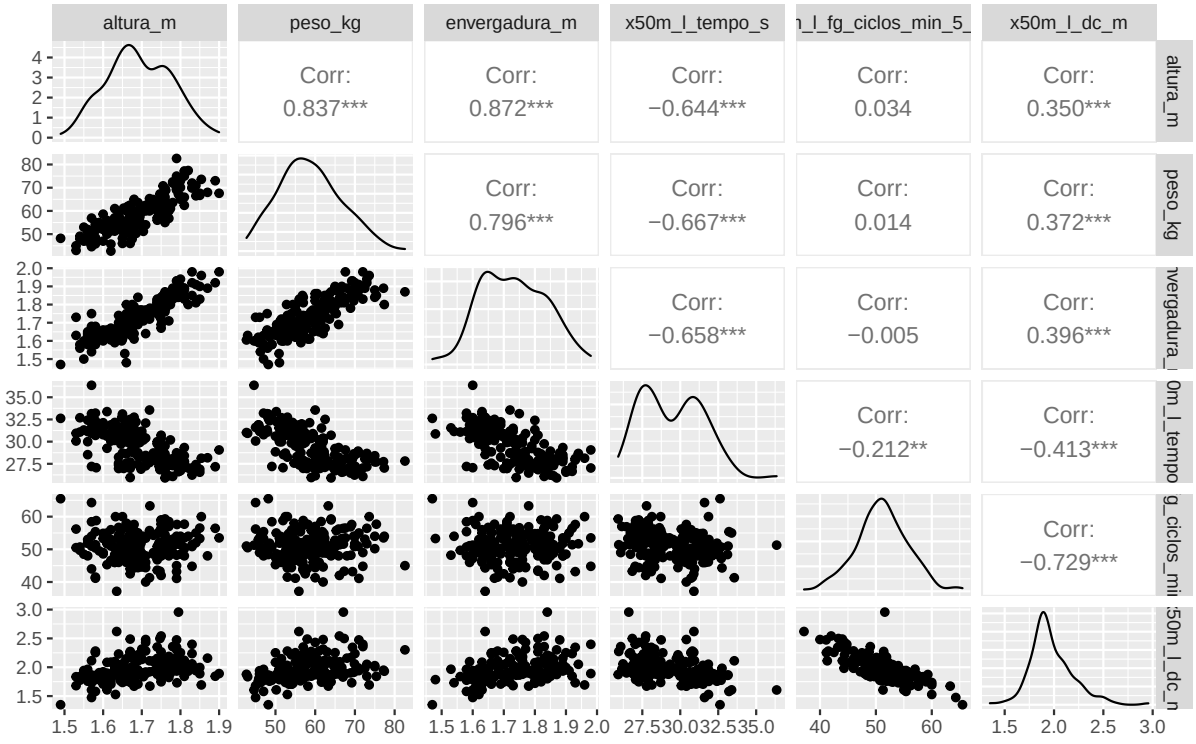
```
df<-dados
summary(df)
```

```
##       sexo          idade          altura_m        peso_kg       envergadura_m
##  Min.   :1.0   Min.   :12.00   Min.   :1.490   Min.   :42.70   Min.   :1.470
##  1st Qu.:1.0   1st Qu.:14.00   1st Qu.:1.639   1st Qu.:53.00   1st Qu.:1.650
##  Median :1.5   Median :14.00   Median :1.688   Median :57.95   Median :1.730
##  Mean   :1.5   Mean   :14.07   Mean   :1.693   Mean   :58.72   Mean   :1.732
##  3rd Qu.:2.0   3rd Qu.:15.00   3rd Qu.:1.755   3rd Qu.:63.70   3rd Qu.:1.810
##  Max.   :2.0   Max.   :16.00   Max.   :1.900   Max.   :82.60   Max.   :1.980
##
##  x50m_l_tempo_s  x50m_l_fg_ciclos_min_5_20m  x50m_l_dc_m
##  Min.   :25.91   Min.   :37.14              Min.   :1.350
##  1st Qu.:27.72   1st Qu.:48.40              1st Qu.:1.843
##  Median :29.29   Median :50.99              Median :1.922
##  Mean   :29.45   Mean   :51.12              Mean   :1.973
##  3rd Qu.:30.99   3rd Qu.:54.00              3rd Qu.:2.086
```

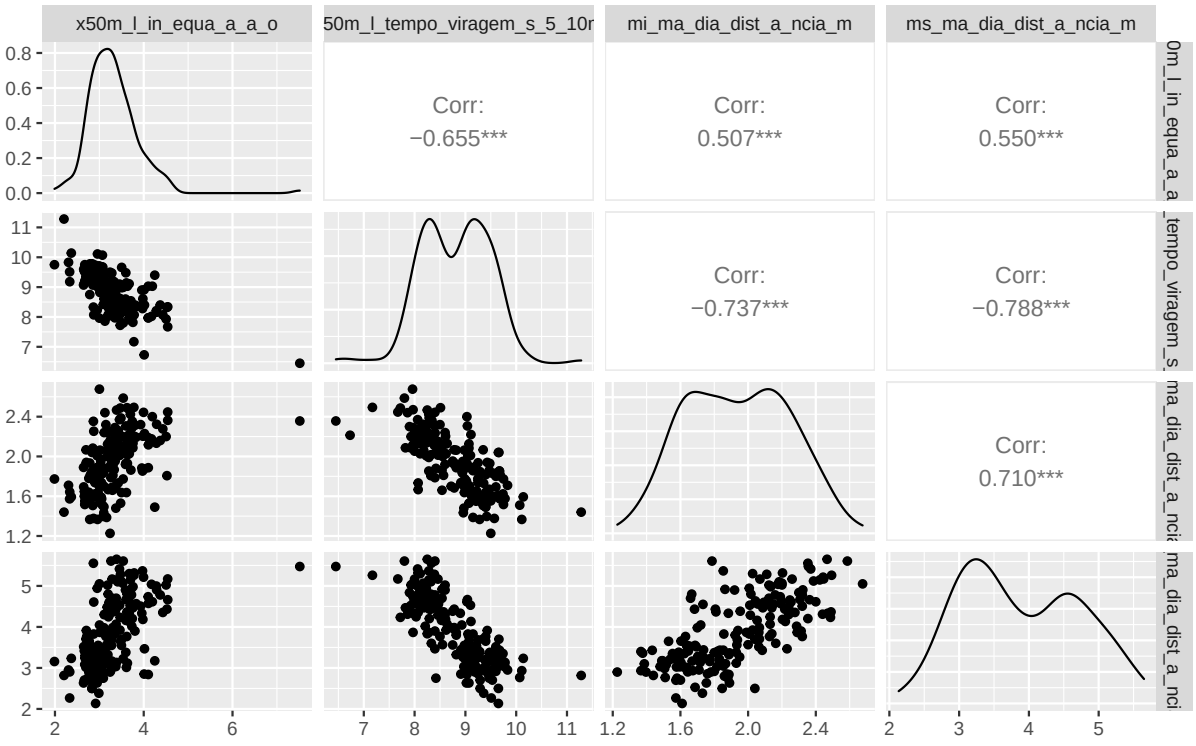
```
##  Max.   :36.35   Max.   :65.50                Max.   :2.956
##  NA's   :1       NA's   :1                    NA's   :1
##  x50m_l_in_equa_a_a_o x50m_l_tempo_viragem_s_5_10m mi_ma_dia_dist_a_ncia_m
##  Min.   :1.989        Min.   : 6.450               Min.   :1.227
##  1st Qu.:2.952        1st Qu.: 8.290               1st Qu.:1.684
##  Median :3.256        Median : 8.910               Median :1.930
##  Mean   :3.308        Mean   : 8.809               Mean   :1.937
##  3rd Qu.:3.600        3rd Qu.: 9.320               3rd Qu.:2.182
##  Max.   :7.516        Max.   :11.280               Max.   :2.677
##  NA's   :1            NA's   :1                    NA's   :2
##  ms_ma_dia_dist_a_ncia_m
##  Min.   :2.133
##  1st Qu.:3.182
##  Median :3.767
##  Mean   :3.890
##  3rd Qu.:4.587
##  Max.   :5.650
##  NA's   :1
```

```
library(corrplot)

# Visualize the data
library(GGally)
ggpairs(df[,c(3:8)])
```




```
ggpairs(df[,c(9:12)])
```



1.4 Verificação dos pressupostos

Peña, E. A., & Slate, E. H. (2006). Global Validation of Linear Model Assumptions. Journal of the American Statistical Association, 101(473), 341. https://doi.org/10.1198/016214505000000637

Pena, E. A., & Slate, E. H. (2019). gvlma: Global Validation of Linear Models Assumptions. https://CRAN.R-project.org/package=gvlma

- global stat:
  - Are the relationships between your X predictors and Y roughly linear?
  - Rejection of the null ($p < .05$) indicates a non-linear relationship between one or more of your X's and Y
- skewness:
  - Is your distribution skewed positively or negatively, necessitating a transformation to meet the assumption of normality?
  - Rejection of the null ($p < .05$) indicates that you should likely transform your data.
- kurtosis:
  - Is your distribution kurtotic (highly peaked or very shallowly peaked), necessitating a transformation to meet the assumption of normality?
  - Rejection of the null ($p < .05$) indicates that you should likely transform your data. measuring the distribution, outliers, influential data, etc
- link function:
  - Is your dependent variable truly continuous, or categorical?

  - Rejection of the null ($p < .05$) indicates that you should use an alternative form of the generalized linear model (e.g. logistic or binomial regression)

- heteroscedasticity:
  - Is the variance of your model residuals constant across the range of X (assumption of homoscedastiity)?
  - Rejection of the null ($p < .05$) indicates that your residuals are heteroscedastic, and thus non-constant across the range of X
  - Your model is better/worse at predicting for certain ranges of your X scales looking for equal variance in the residuals

```
names(df)
```

```
##  [1] "sexo"                          "idade"
##  [3] "altura_m"                      "peso_kg"
##  [5] "envergadura_m"                 "x50m_l_tempo_s"
##  [7] "x50m_l_fg_ciclos_min_5_20m"    "x50m_l_dc_m"
##  [9] "x50m_l_in_equa_a_a_o"          "x50m_l_tempo_viragem_s_5_10m"
## [11] "mi_ma_dia_dist_a_ncia_m"       "ms_ma_dia_dist_a_ncia_m"
```

```
library(gvlma)
myLModel <- lm(x50m_l_tempo_s ~ sexo+idade+altura_m+peso_kg+envergadura_m+x50m_l_fg_c
                 x50m_l_in_equa_a_a_o+x50m_l_tempo_viragem_s_5_10m+mi_ma_dia_dist_a_n
summary(myLModel)
```

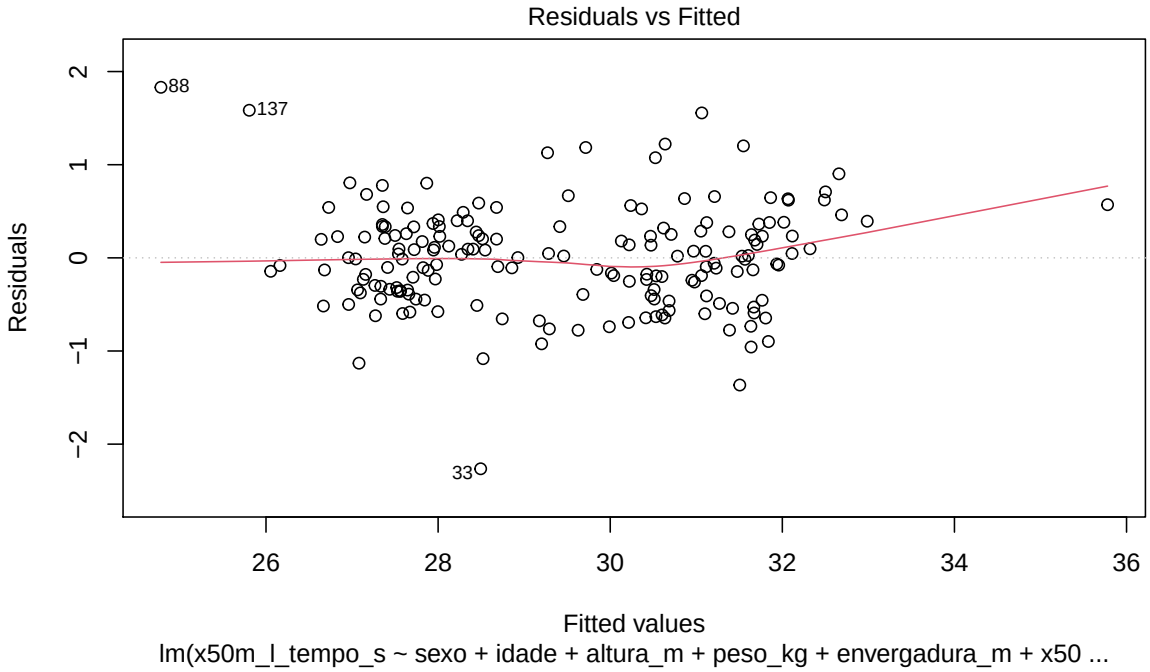
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ sexo + idade + altura_m + peso_kg +
##     envergadura_m + x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m +
##     x50m_l_in_equa_a_a_o + x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.26453 -0.35892  0.01059  0.33308  1.83107
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    52.854193   5.893646   8.968 5.73e-16 ***
## sexo                            0.534794   0.194033   2.756  0.00649 **
## idade                          -0.020960   0.086254  -0.243  0.80830
## altura_m                       -0.058279   1.202197  -0.048  0.96139
## peso_kg                        -0.008441   0.011881  -0.710  0.47841
## envergadura_m                  -0.449794   0.916633  -0.491  0.62428
## x50m_l_fg_ciclos_min_5_20m     -0.356209   0.050699  -7.026 5.06e-11 ***
## x50m_l_dc_m                   -14.706181   2.171662  -6.772 2.04e-10 ***
## x50m_l_in_equa_a_a_o            3.435431   0.574714   5.978 1.32e-08 ***
## x50m_l_tempo_viragem_s_5_10m    1.633439   0.150397  10.861  < 2e-16 ***
## mi_ma_dia_dist_a_ncia_m        -0.579702   0.237326  -2.443  0.01562 *
## ms_ma_dia_dist_a_ncia_m         0.012535   0.120556   0.104  0.91731
```

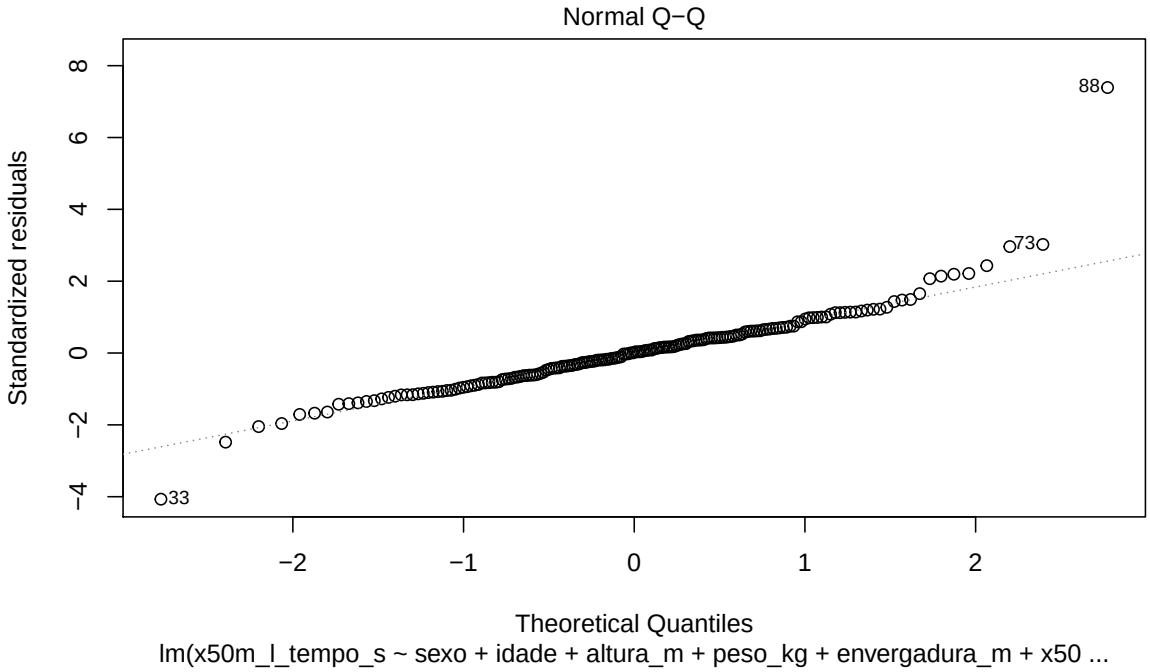
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5679 on 168 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.924,  Adjusted R-squared:  0.919
## F-statistic: 185.7 on 11 and 168 DF,  p-value: < 2.2e-16
```

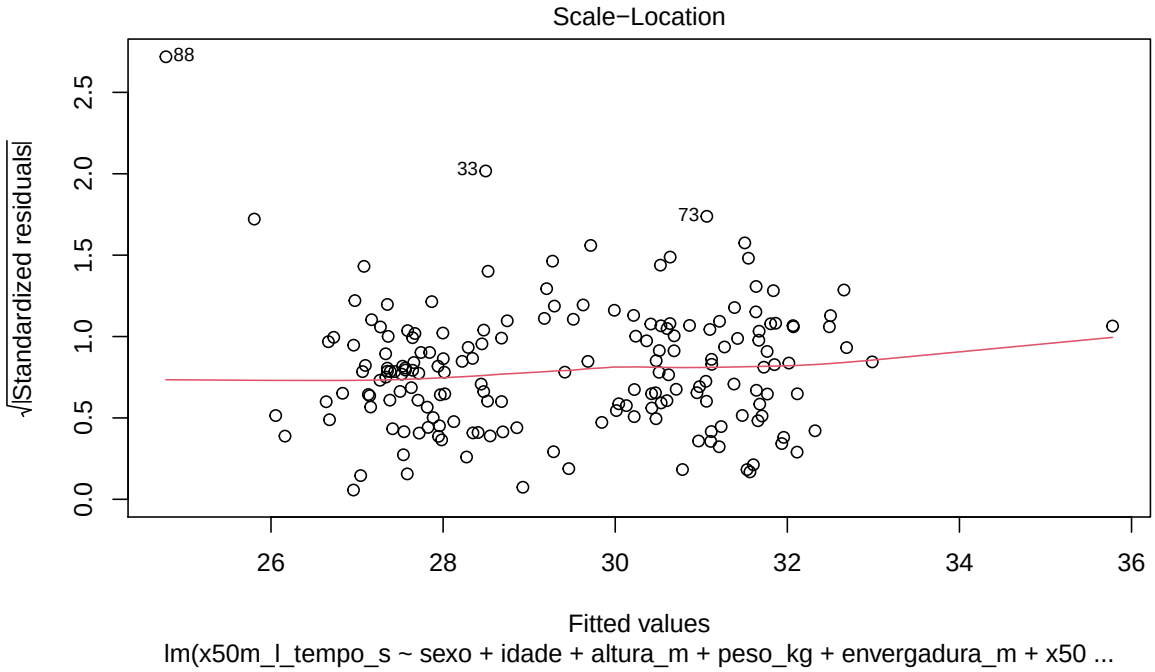
```
gvlma(myLModel,alphalevel = 0.05)
```

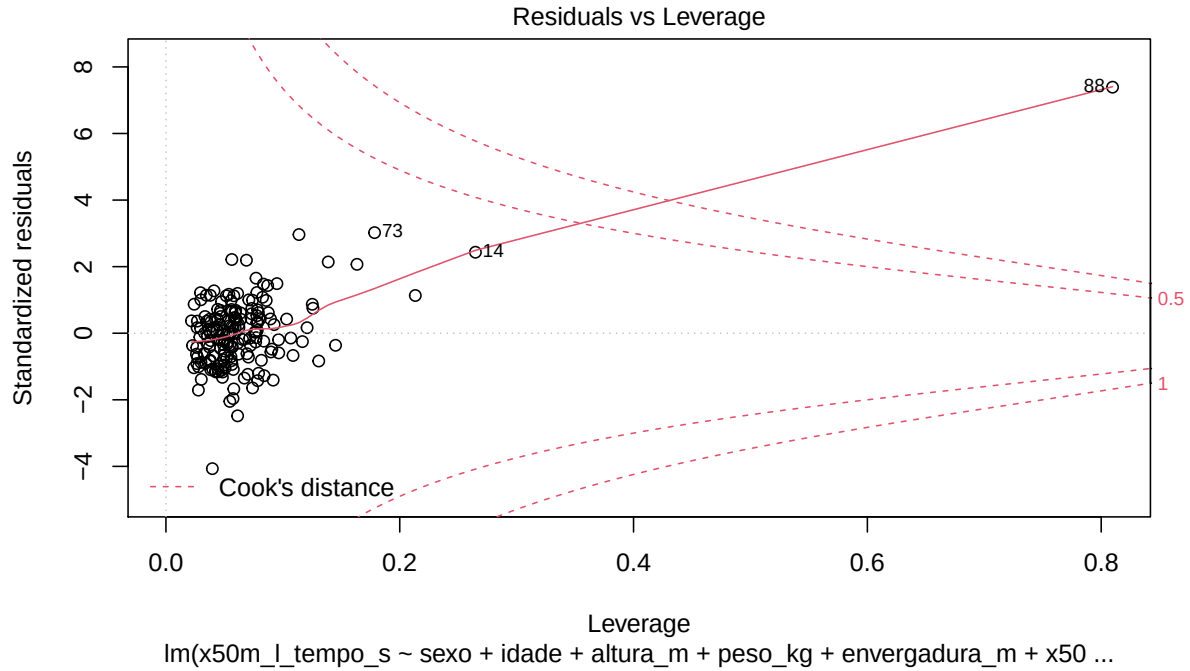
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ sexo + idade + altura_m + peso_kg +
##     envergadura_m + x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m +
##     x50m_l_in_equa_a_a_o + x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = df)
##
## Coefficients:
##                   (Intercept)                           sexo
##                     52.854193                       0.534794
##                         idade                       altura_m
##                     -0.020960                      -0.058279
##                       peso_kg                  envergadura_m
##                     -0.008441                      -0.449794
##    x50m_l_fg_ciclos_min_5_20m                    x50m_l_dc_m
##                     -0.356209                      -14.706181
##          x50m_l_in_equa_a_a_o  x50m_l_tempo_viragem_s_5_10m
##                      3.435431                       1.633439
##       mi_ma_dia_dist_a_ncia_m       ms_ma_dia_dist_a_ncia_m
##                     -0.579702                       0.012535
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = myLModel, alphalevel = 0.05)
##
##                       Value   p-value                   Decision
## Global Stat        39.2014 6.330e-08 Assumptions NOT satisfied!
## Skewness            0.1159 7.335e-01    Assumptions acceptable.
## Kurtosis           22.6217 1.972e-06 Assumptions NOT satisfied!
## Link Function      15.9697 6.437e-05 Assumptions NOT satisfied!
## Heteroscedasticity  0.4941 4.821e-01    Assumptions acceptable.
```

```
plot(myLModel)
```


Residuals vs Fitted
Residuals
Fitted values
lm(x50m_l_tempo_s ~ sexo + idade + altura_m + peso_kg + envergadura_m + x50 ...
88
137
33


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(x50m_l_tempo_s ~ sexo + idade + altura_m + peso_kg + envergadura_m + x50 ...
88
73
33






```
df<-df[complete.cases(df),]



myLModel <- lm(x50m_l_tempo_s ~ sexo+idade+altura_m+peso_kg+envergadura_m+x50m_l_fg_c
                x50m_l_in_equa_a_a_o+x50m_l_tempo_viragem_s_5_10m+mi_ma_dia_dist_a_n
myLModel
```

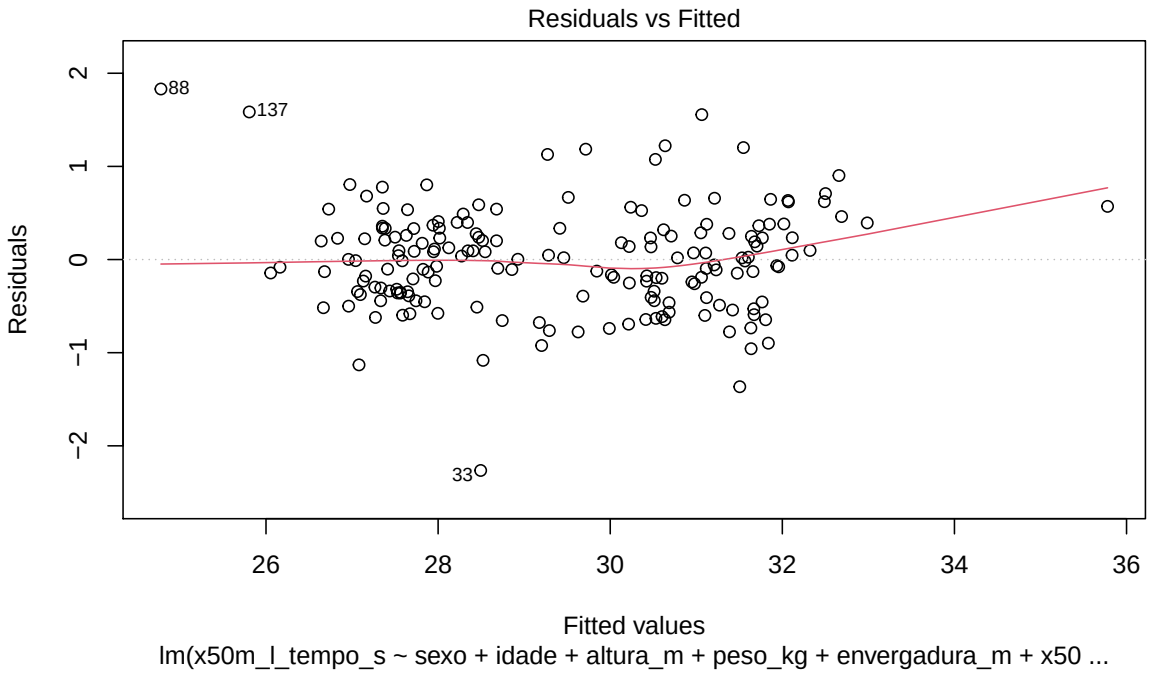
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ sexo + idade + altura_m + peso_kg +
##     envergadura_m + x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m +
##     x50m_l_in_equa_a_a_o + x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = df)
##
## Coefficients:
##                   (Intercept)                           sexo
##                     52.854193                       0.534794
##                         idade                       altura_m
##                     -0.020960                      -0.058279
##                       peso_kg                  envergadura_m
##                     -0.008441                      -0.449794
##   x50m_l_fg_ciclos_min_5_20m                    x50m_l_dc_m
##                     -0.356209                     -14.706181
##         x50m_l_in_equa_a_a_o  x50m_l_tempo_viragem_s_5_10m
##                      3.435431                       1.633439
##      mi_ma_dia_dist_a_ncia_m       ms_ma_dia_dist_a_ncia_m
##                     -0.579702                       0.012535
```

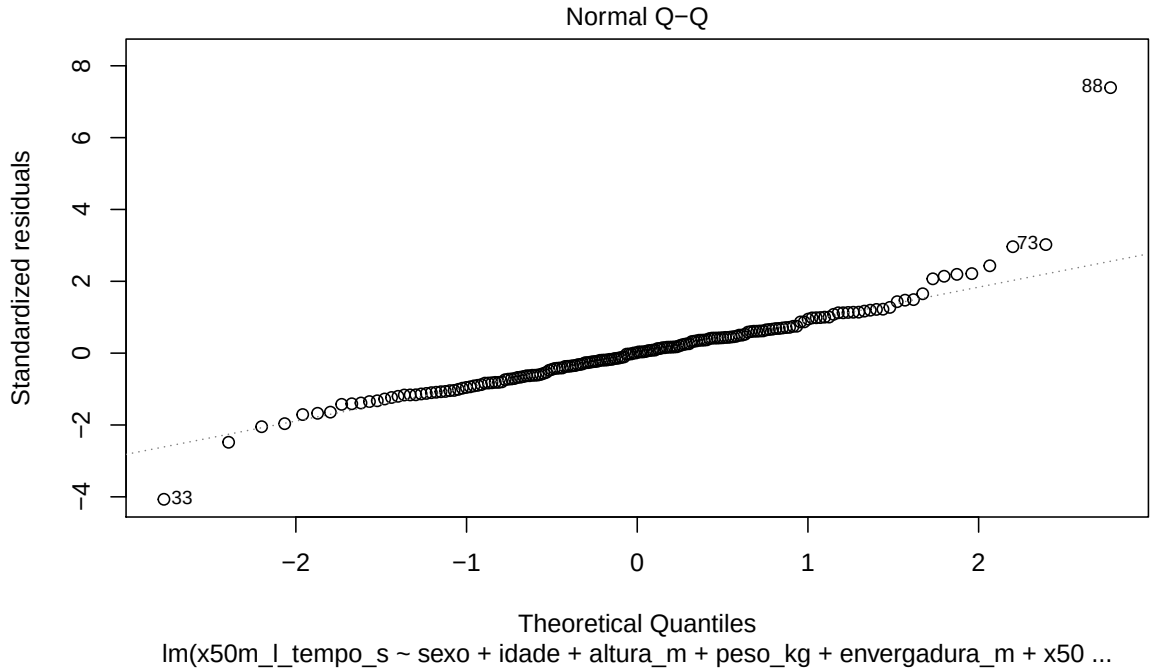
```
gvlma(myLModel,alphalevel = 0.05)
```

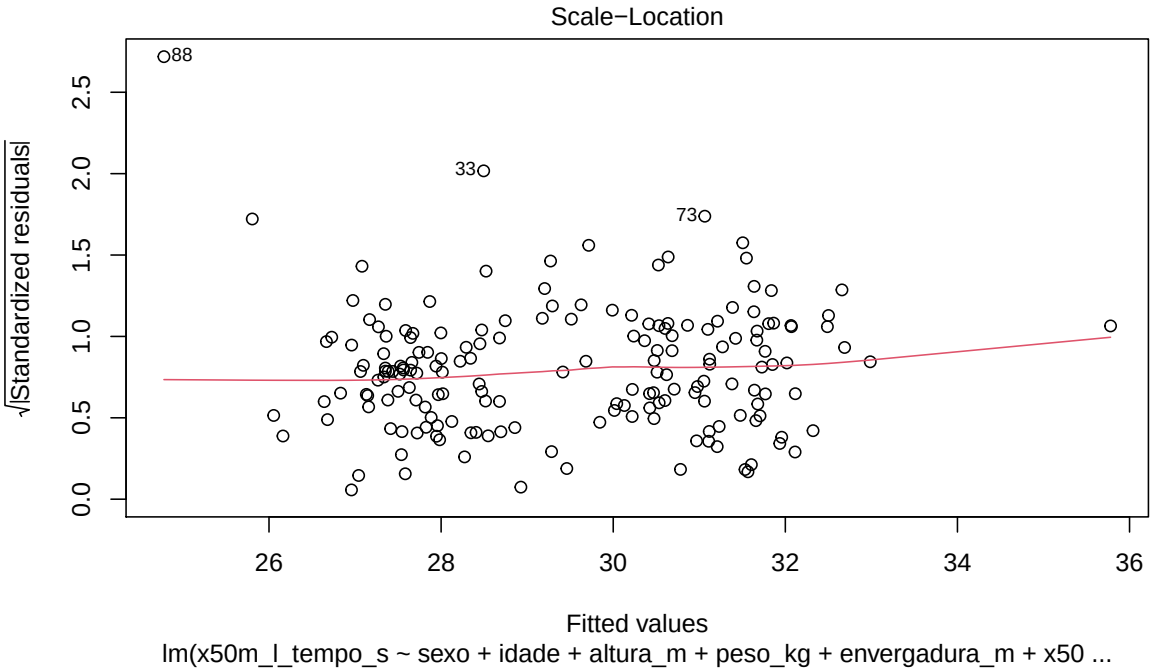
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ sexo + idade + altura_m + peso_kg +
##     envergadura_m + x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m +
##     x50m_l_in_equa_a_a_o + x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = df)
##
## Coefficients:
##                   (Intercept)                           sexo
##                     52.854193                       0.534794
##                         idade                       altura_m
##                     -0.020960                      -0.058279
##                       peso_kg                  envergadura_m
##                     -0.008441                      -0.449794
##   x50m_l_fg_ciclos_min_5_20m                    x50m_l_dc_m
##                     -0.356209                     -14.706181
##         x50m_l_in_equa_a_a_o  x50m_l_tempo_viragem_s_5_10m
##                      3.435431                       1.633439
##      mi_ma_dia_dist_a_ncia_m       ms_ma_dia_dist_a_ncia_m
##                     -0.579702                       0.012535
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
```

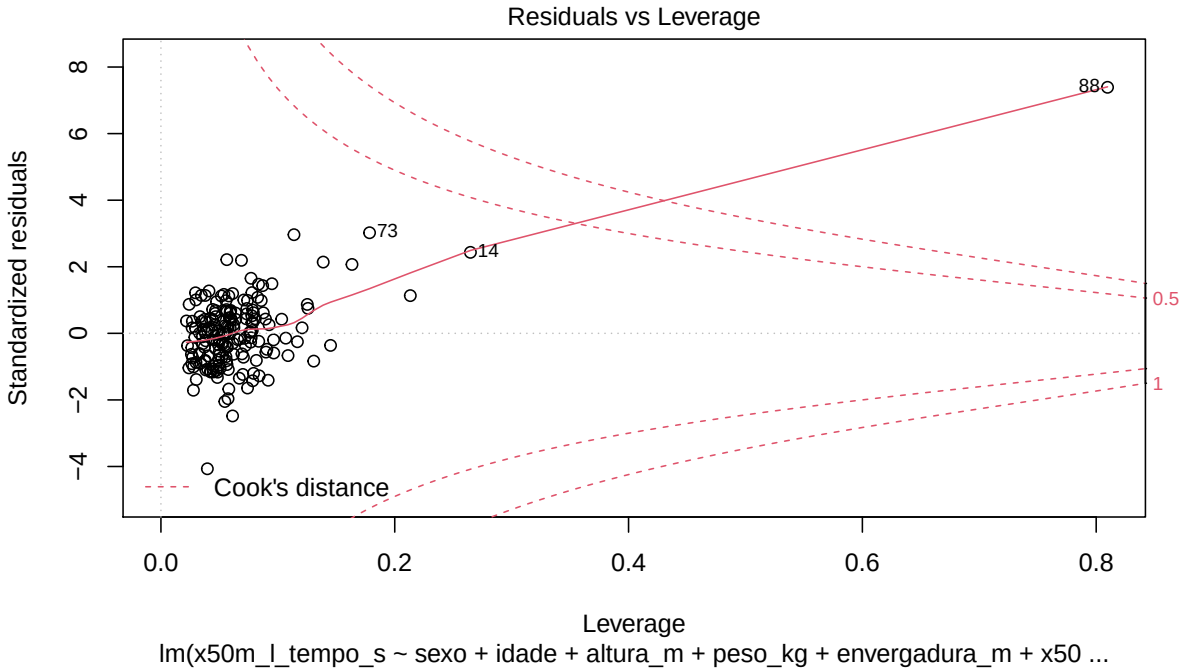
```
## Call:
##  gvlma(x = myLModel, alphalevel = 0.05)
##
##                      Value   p-value                   Decision
## Global Stat        39.2014 6.330e-08 Assumptions NOT satisfied!
## Skewness            0.1159 7.335e-01    Assumptions acceptable.
## Kurtosis           22.6217 1.972e-06 Assumptions NOT satisfied!
## Link Function      15.9697 6.437e-05 Assumptions NOT satisfied!
## Heteroscedasticity  0.4941 4.821e-01    Assumptions acceptable.
```

```
plot(myLModel)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(x50m_l_tempo_s ~ sexo + idade + altura_m + peso_kg + envergadura_m + x50 ...
88
73
33

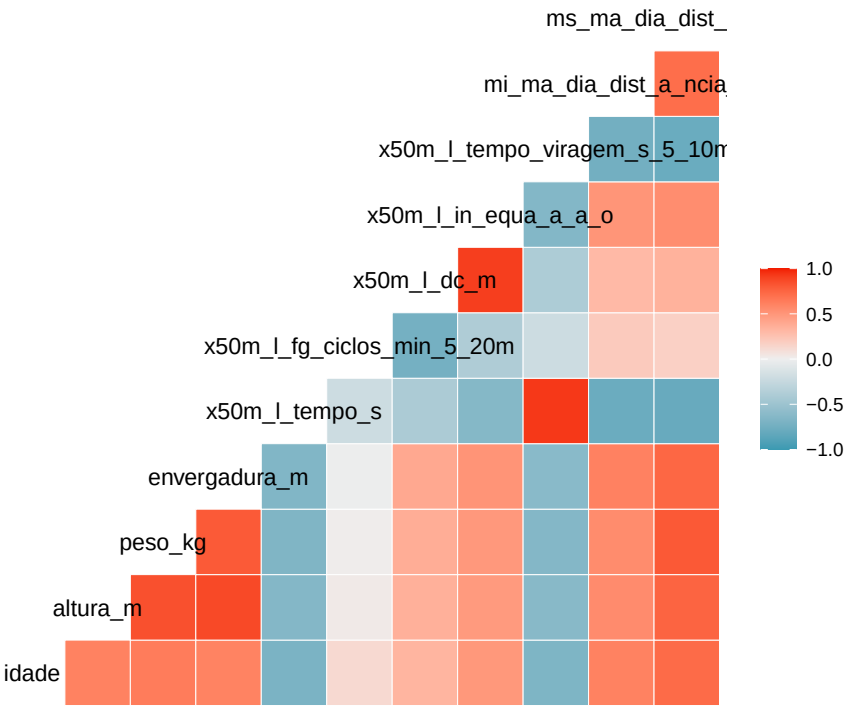

Scale−Location
√|Standardized residuals|
Fitted values
lm(x50m_l_tempo_s ~ sexo + idade + altura_m + peso_kg + envergadura_m + x50 ...
88
33
73



1.5 Failing assumptions

Existem missings - Vamos remover

```
ggcorr(df[,-c(1)])
```



Femininos

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
dfFem<-df %>% filter(sexo==2)
```

```r
myLModel <- lm(x50m_l_tempo_s ~ idade+altura_m+peso_kg+envergadura_m+x50m_l_fg_ciclos
                x50m_l_in_equa_a_a_o+x50m_l_tempo_viragem_s_5_10m+mi_ma_dia_dist_a_nc
myLModel
```

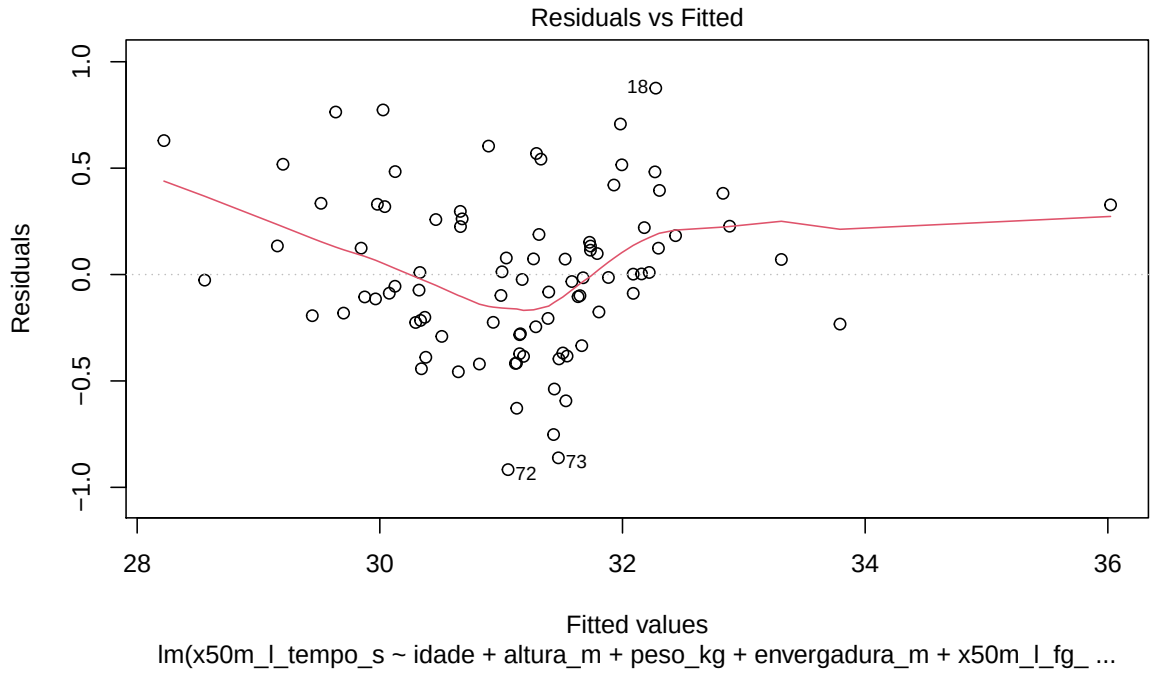
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m +
##     x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = dfFem)
##
## Coefficients:
##                   (Intercept)                         idade
##                       6.22888                       0.07669
##                    altura_m                        peso_kg
##                       3.74189                      -0.03224
##                 envergadura_m    x50m_l_fg_ciclos_min_5_20m
##                      -0.95717                       0.06244
##                  x50m_l_dc_m          x50m_l_in_equa_a_a_o
##                      11.26945                      -6.03879
## x50m_l_tempo_viragem_s_5_10m       mi_ma_dia_dist_a_ncia_m
##                       1.63615                      -0.63341
##      ms_ma_dia_dist_a_ncia_m
##                       0.12689
```

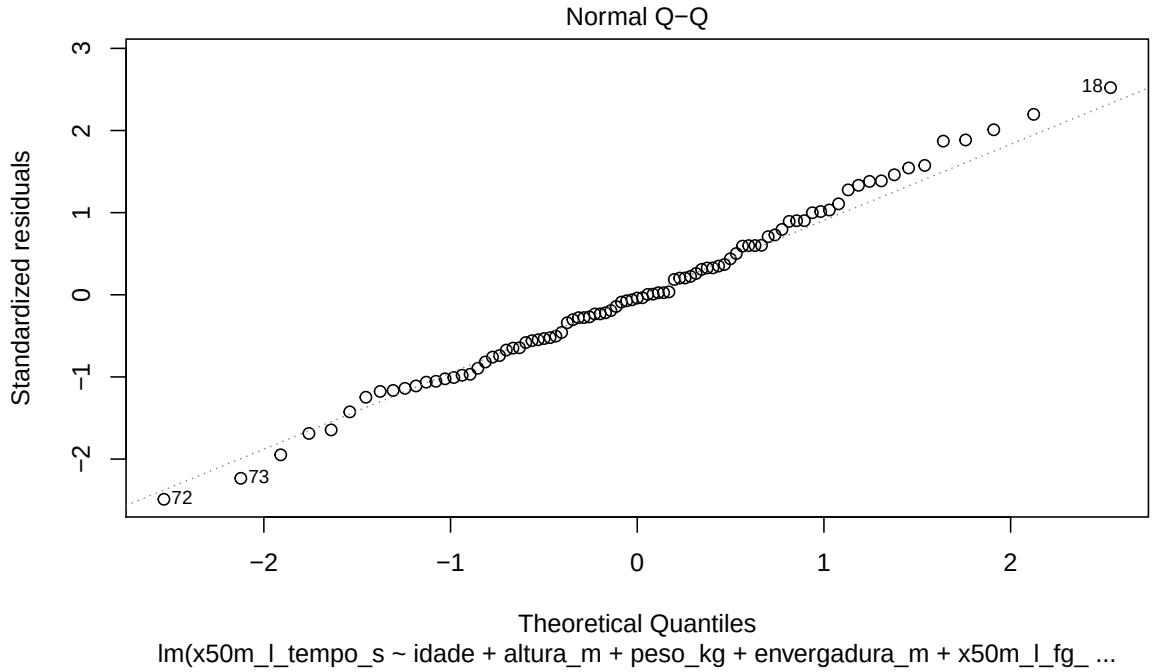
```r
gvlma(myLModel,alphalevel = 0.05)
```

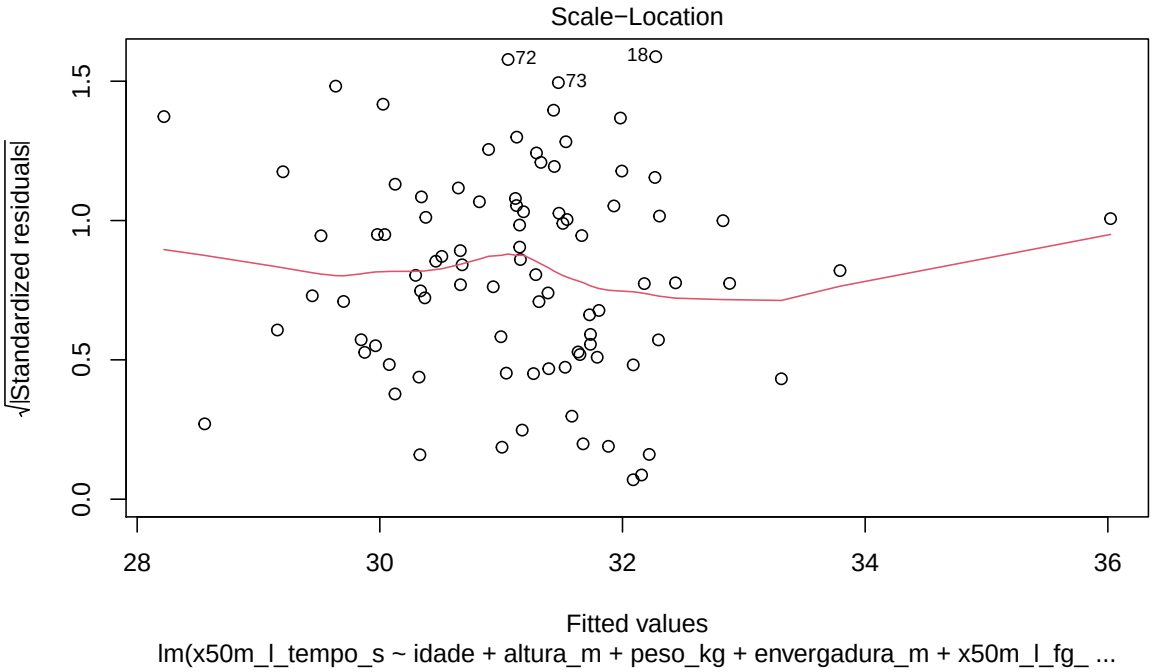
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m +
##     x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = dfFem)
##
## Coefficients:
##                   (Intercept)                         idade
```

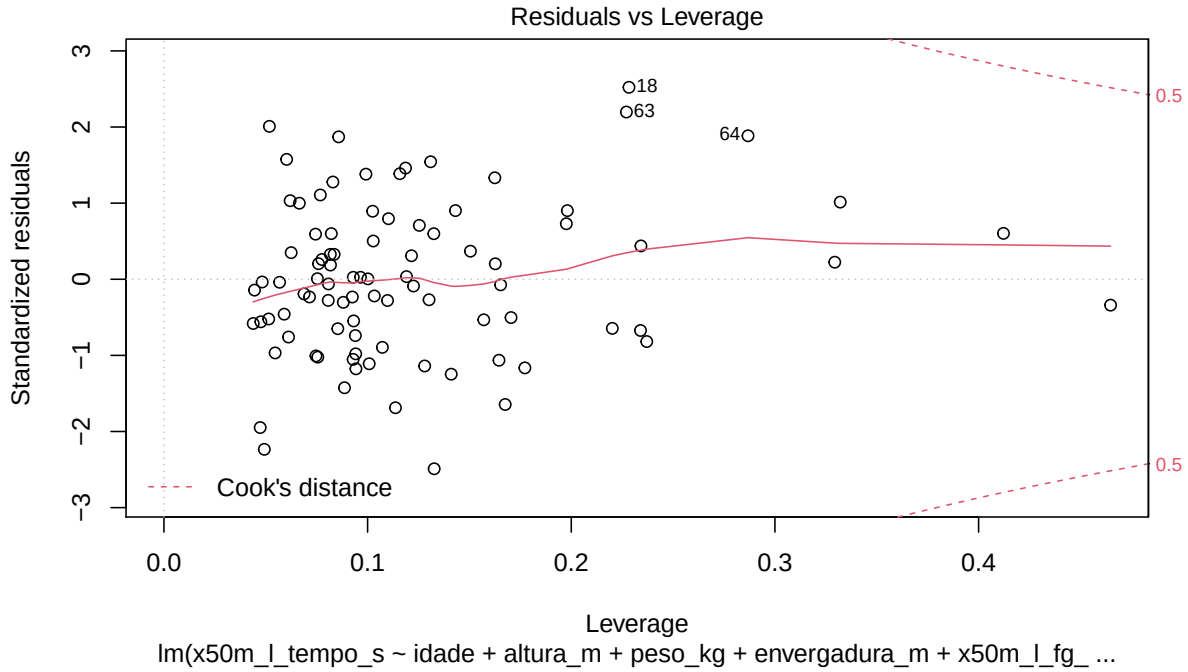
```
##                       6.22888                       0.07669
##                     altura_m                      peso_kg
##                       3.74189                      -0.03224
##                  envergadura_m   x50m_l_fg_ciclos_min_5_20m
##                       -0.95717                       0.06244
##                    x50m_l_dc_m         x50m_l_in_equa_a_a_o
##                       11.26945                      -6.03879
## x50m_l_tempo_viragem_s_5_10m       mi_ma_dia_dist_a_ncia_m
##                       1.63615                      -0.63341
##      ms_ma_dia_dist_a_ncia_m
##                       0.12689
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = myLModel, alphalevel = 0.05)
##
##                      Value p-value                   Decision
## Global Stat        7.88668 0.09582    Assumptions acceptable.
## Skewness           0.06985 0.79155    Assumptions acceptable.
## Kurtosis           0.11473 0.73482    Assumptions acceptable.
## Link Function      6.46615 0.01099 Assumptions NOT satisfied!
## Heteroscedasticity 1.23595 0.26625    Assumptions acceptable.
```

```
plot(myLModel)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m + x50m_l_fg_ ...
18
73
72


Scale−Location
√|Standardized residuals|
Fitted values
lm(x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m + x50m_l_fg_ ...
72
73
18

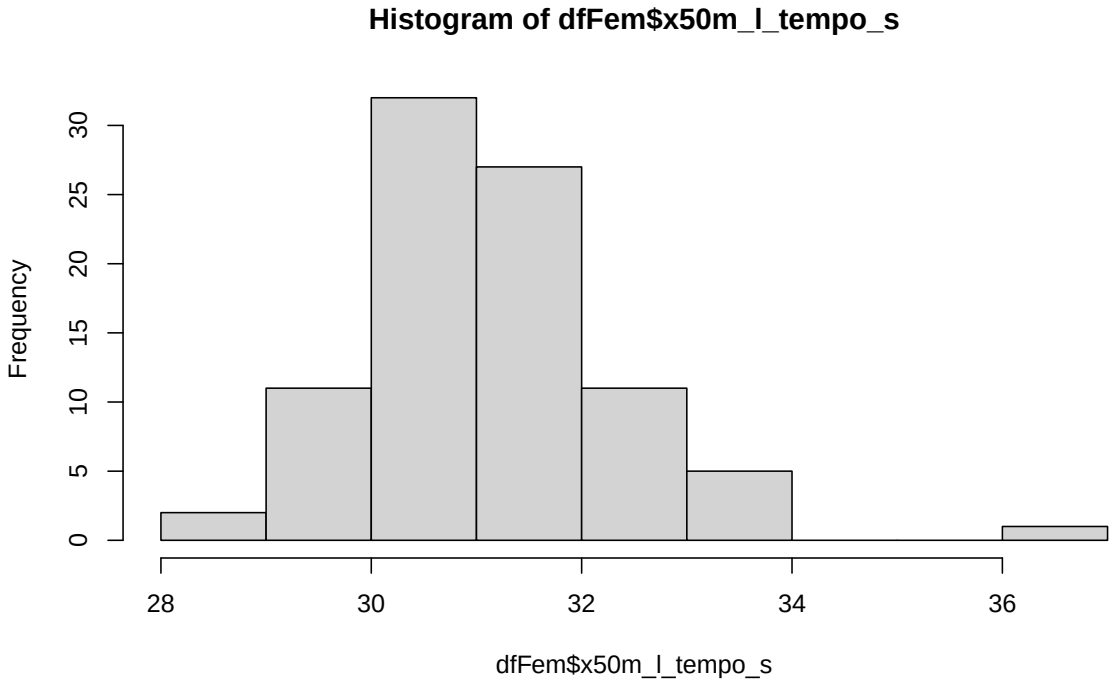


2 Obs

Apesar do teste link function ter um p<0.05 falhando o teste, a variável é contínua. Considerando que: - Is your dependent variable truly continuous, or categorical? - Rejection of the null (p < .05) indicates that you should use an alternative form of the generalized linear model (e.g. logistic or binomial regression)

2.1 Plot tempos

```
hist(dfFem$x50m_l_tempo_s)
```



2.2 Model summary Fem

```
summary(myLModel)
```

```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m +
##     x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = dfFem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91685 -0.23285 -0.01514  0.22714  0.87594
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    6.22888    6.50159   0.958 0.340995
## idade                          0.07669    0.08478   0.905 0.368430
## altura_m                       3.74189    1.11986   3.341 0.001282 **
## peso_kg                       -0.03224    0.01309  -2.464 0.015956 *
## envergadura_m                 -0.95717    0.85165  -1.124 0.264501
## x50m_l_fg_ciclos_min_5_20m     0.06244    0.05612   1.113 0.269267
## x50m_l_dc_m                   11.26945    3.04363   3.703 0.000396 ***
## x50m_l_in_equa_a_a_o          -6.03879    1.06665  -5.661  2.4e-07 ***
## x50m_l_tempo_viragem_s_5_10m   1.63615    0.14499  11.285  < 2e-16 ***
## mi_ma_dia_dist_a_ncia_m       -0.63341    0.25851  -2.450 0.016516 *
```

```
## ms_ma_dia_dist_a_ncia_m       0.12689    0.13086   0.970 0.335221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3954 on 78 degrees of freedom
## Multiple R-squared:  0.9013, Adjusted R-squared:  0.8887
## F-statistic: 71.27 on 10 and 78 DF,  p-value: < 2.2e-16
```

3 Masculinos

```
library(dplyr)
dfMas<-df %>% filter(sexo==1)

myLModel <- lm(x50m_l_tempo_s ~ idade+altura_m+peso_kg+envergadura_m+x50m_l_fg_ciclos_
                 x50m_l_in_equa_a_a_o+x50m_l_tempo_viragem_s_5_10m+mi_ma_dia_dist_a_nc
myLModel
```

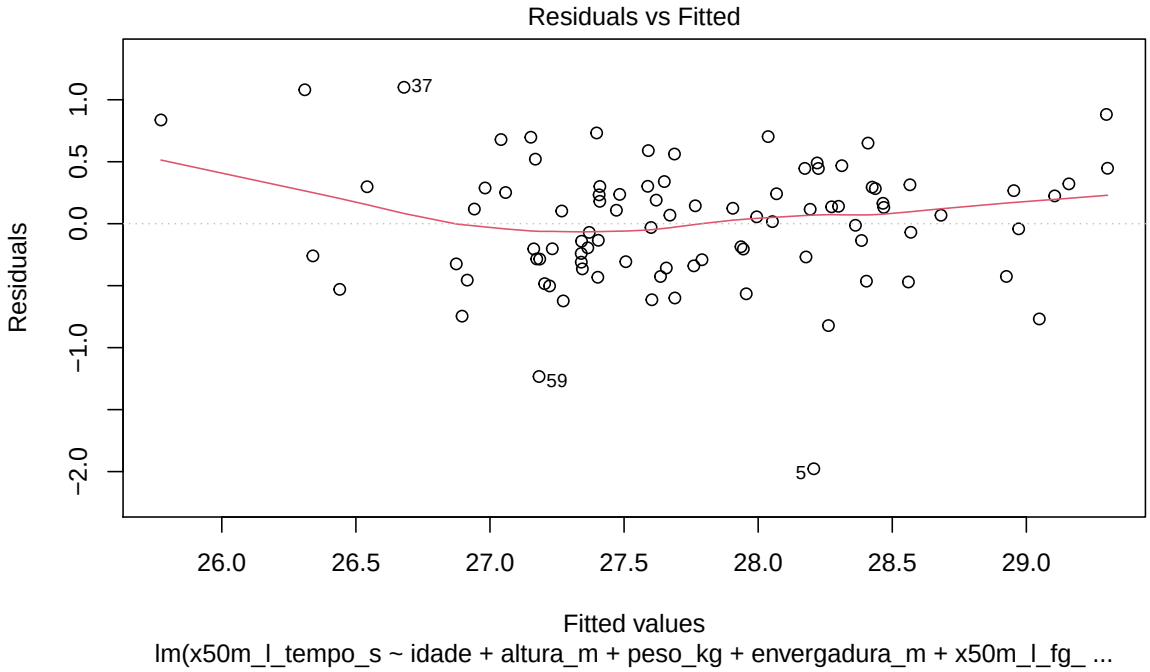
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m +
##     x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = dfMas)
##
## Coefficients:
##                   (Intercept)                         idade
##                     6.967e+01                    -4.878e-02
##                      altura_m                        peso_kg
##                    -2.888e+00                    -3.234e-04
##                 envergadura_m    x50m_l_fg_ciclos_min_5_20m
##                     2.053e+00                    -4.801e-01
##                   x50m_l_dc_m          x50m_l_in_equa_a_a_o
##                    -2.035e+01                     4.831e+00
## x50m_l_tempo_viragem_s_5_10m       mi_ma_dia_dist_a_ncia_m
##                     1.248e+00                    -2.013e-01
##      ms_ma_dia_dist_a_ncia_m
##                    -1.259e-01
```

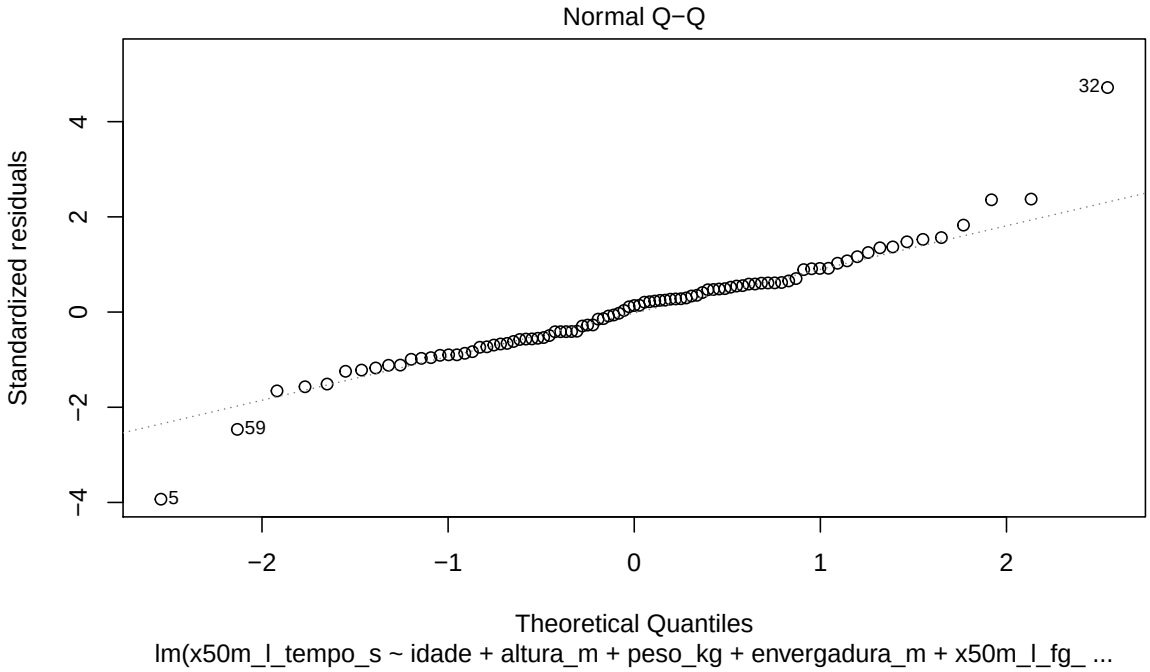
```
gvlma(myLModel,alphalevel = 0.05)
```

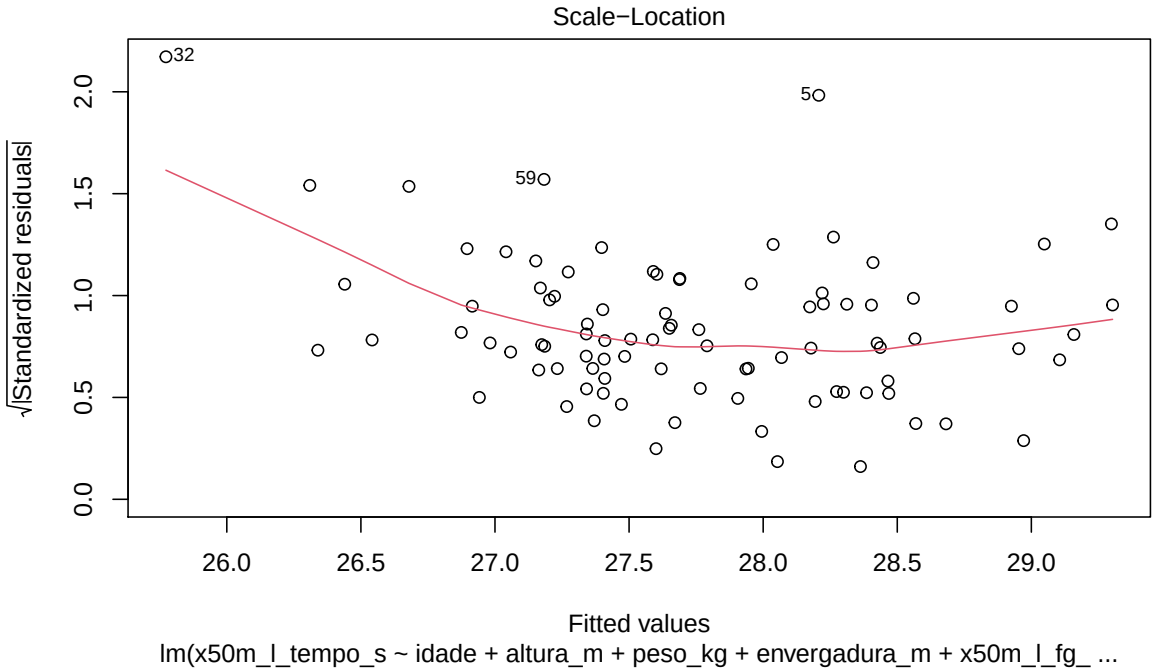
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m +
##     x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = dfMas)
##
## Coefficients:
```

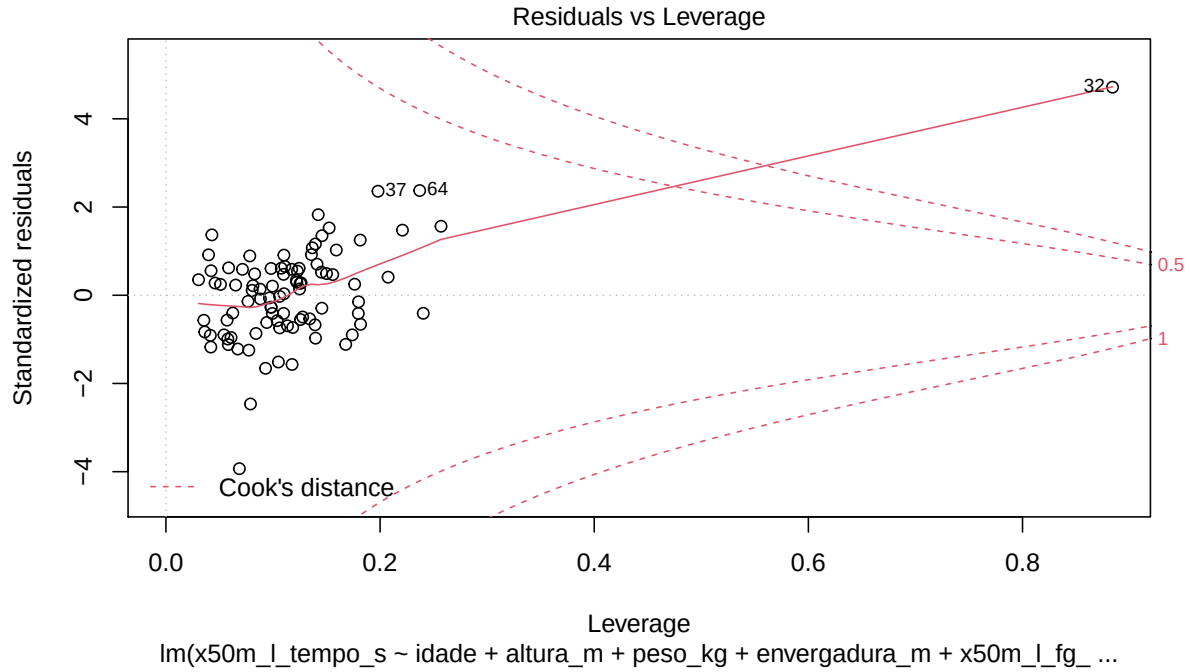
```
##                   (Intercept)                         idade
##                     6.967e+01                    -4.878e-02
##                      altura_m                        peso_kg
##                    -2.888e+00                    -3.234e-04
##                 envergadura_m    x50m_l_fg_ciclos_min_5_20m
##                     2.053e+00                    -4.801e-01
##                   x50m_l_dc_m          x50m_l_in_equa_a_a_o
##                    -2.035e+01                     4.831e+00
## x50m_l_tempo_viragem_s_5_10m       mi_ma_dia_dist_a_ncia_m
##                     1.248e+00                    -2.013e-01
##      ms_ma_dia_dist_a_ncia_m
##                    -1.259e-01
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = myLModel, alphalevel = 0.05)
##
##                      Value   p-value                   Decision
## Global Stat        28.495 9.897e-06 Assumptions NOT satisfied!
## Skewness            5.329 2.098e-02 Assumptions NOT satisfied!
## Kurtosis           13.921 1.906e-04 Assumptions NOT satisfied!
## Link Function       8.159 4.285e-03 Assumptions NOT satisfied!
## Heteroscedasticity  1.086 2.973e-01    Assumptions acceptable.
```

```
plot(myLModel)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−1.0
−2.0
37
59
5
26.0
26.5
27.0
27.5
28.0
28.5
29.0
Fitted values
lm(x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m + x50m_l_fg_ ...


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
32
59
5
−2
−1
0
1
2
Theoretical Quantiles
lm(x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m + x50m_l_fg_ ...





Vamos remover 5,59 e 32. Nos masculinos.

```
dfMas<-dfMas[-c(5,59,32),]

myLModel <- lm(x50m_l_tempo_s ~ idade+altura_m+peso_kg+envergadura_m+x50m_l_fg_ciclos_
                x50m_l_in_equa_a_a_o+x50m_l_tempo_viragem_s_5_10m+mi_ma_dia_dist_a_n
myLModel

##
```

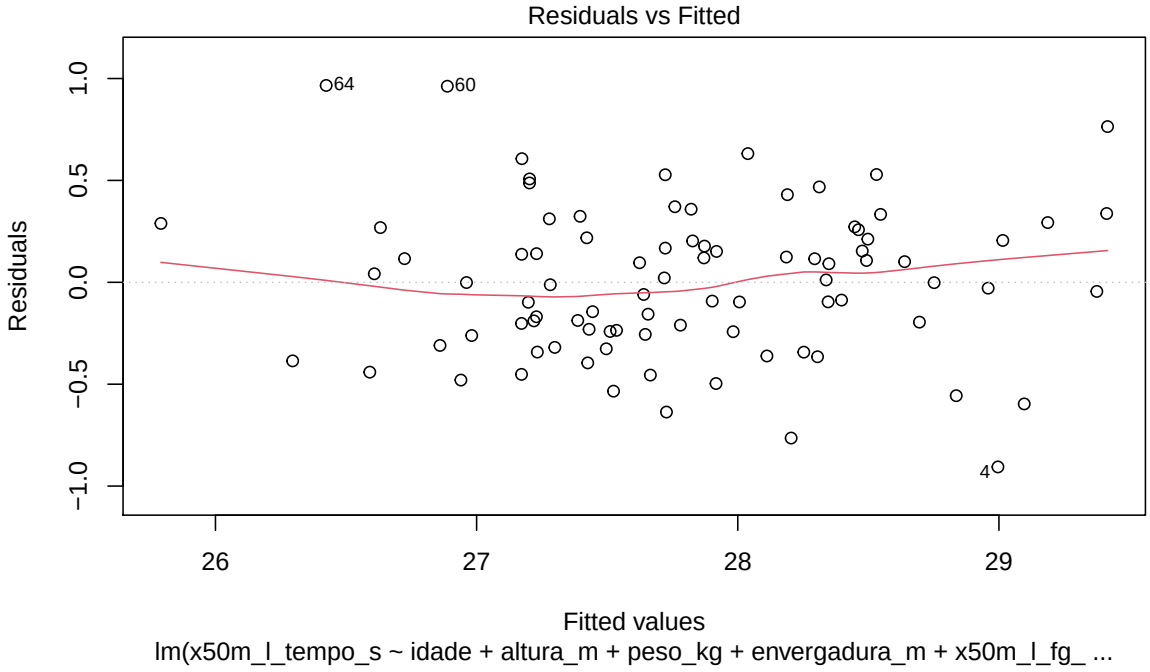
```
## Call:
## lm(formula = x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m +
##     x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = dfMas)
##
## Coefficients:
##                   (Intercept)                         idade
##                     45.197947                      0.053762
##                      altura_m                       peso_kg
##                     -2.282374                     -0.002136
##                 envergadura_m    x50m_l_fg_ciclos_min_5_20m
##                      1.168755                     -0.251845
##                   x50m_l_dc_m          x50m_l_in_equa_a_a_o
##                     -6.329625                      0.120339
## x50m_l_tempo_viragem_s_5_10m       mi_ma_dia_dist_a_ncia_m
##                      1.223299                     -0.224616
##      ms_ma_dia_dist_a_ncia_m
##                     -0.058869
```

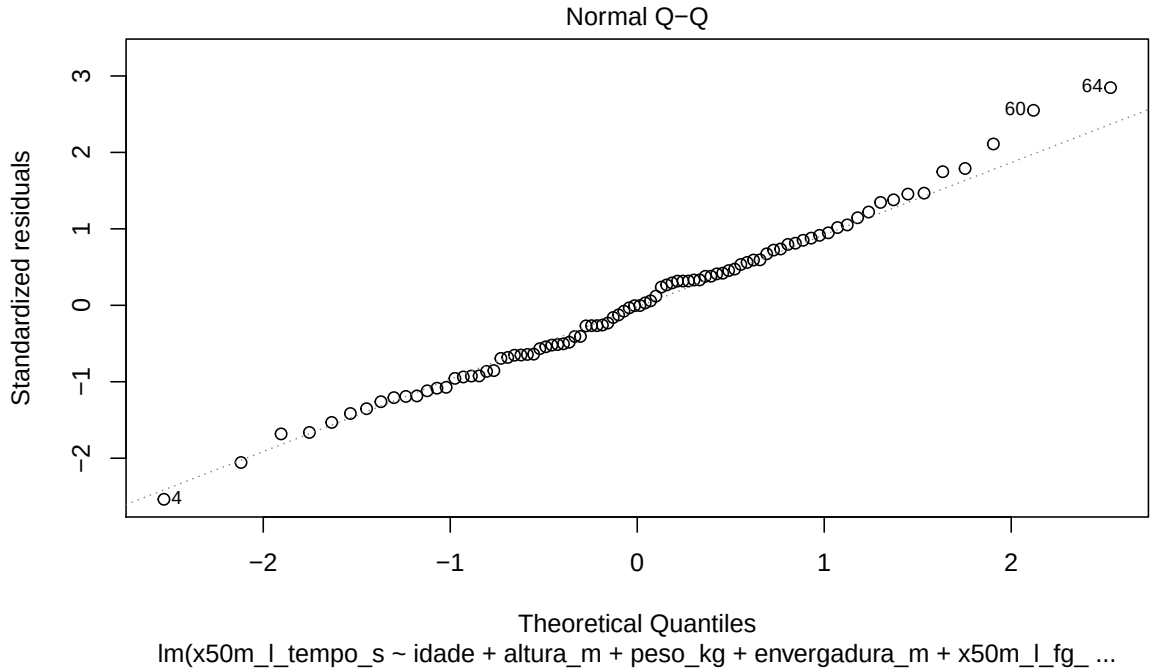
```
gvlma(myLModel,alphalevel = 0.05)
```

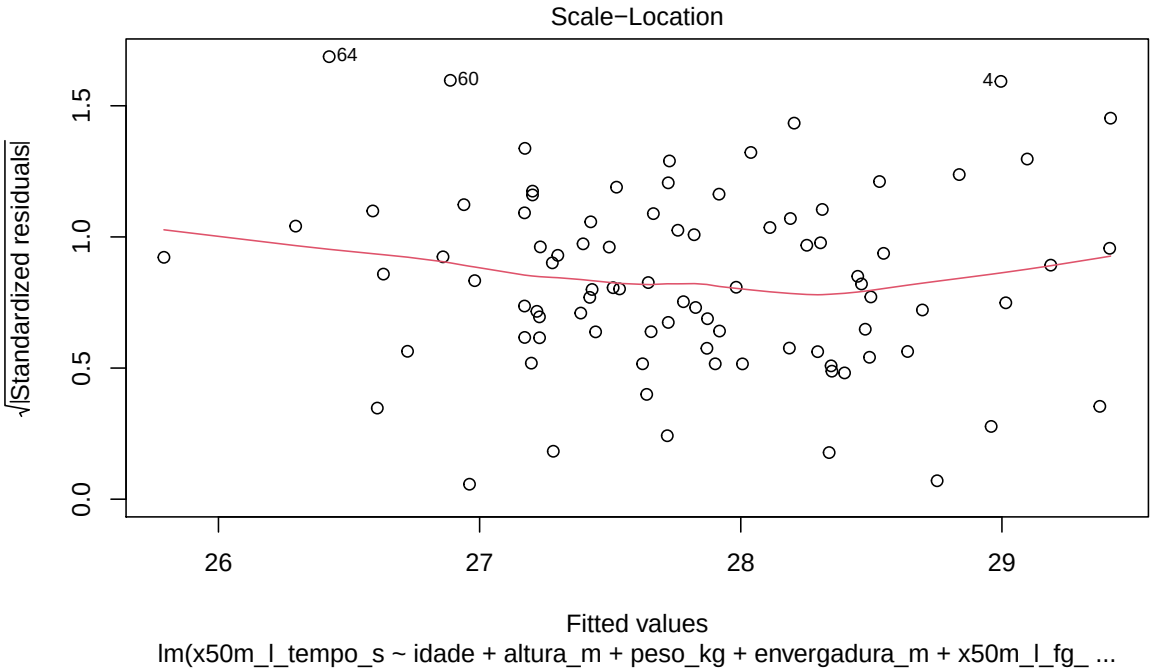
```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m +
##     x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = dfMas)
##
## Coefficients:
##                   (Intercept)                         idade
##                     45.197947                      0.053762
##                      altura_m                       peso_kg
##                     -2.282374                     -0.002136
##                 envergadura_m    x50m_l_fg_ciclos_min_5_20m
##                      1.168755                     -0.251845
##                   x50m_l_dc_m          x50m_l_in_equa_a_a_o
##                     -6.329625                      0.120339
## x50m_l_tempo_viragem_s_5_10m       mi_ma_dia_dist_a_ncia_m
##                      1.223299                     -0.224616
##      ms_ma_dia_dist_a_ncia_m
##                     -0.058869
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
```

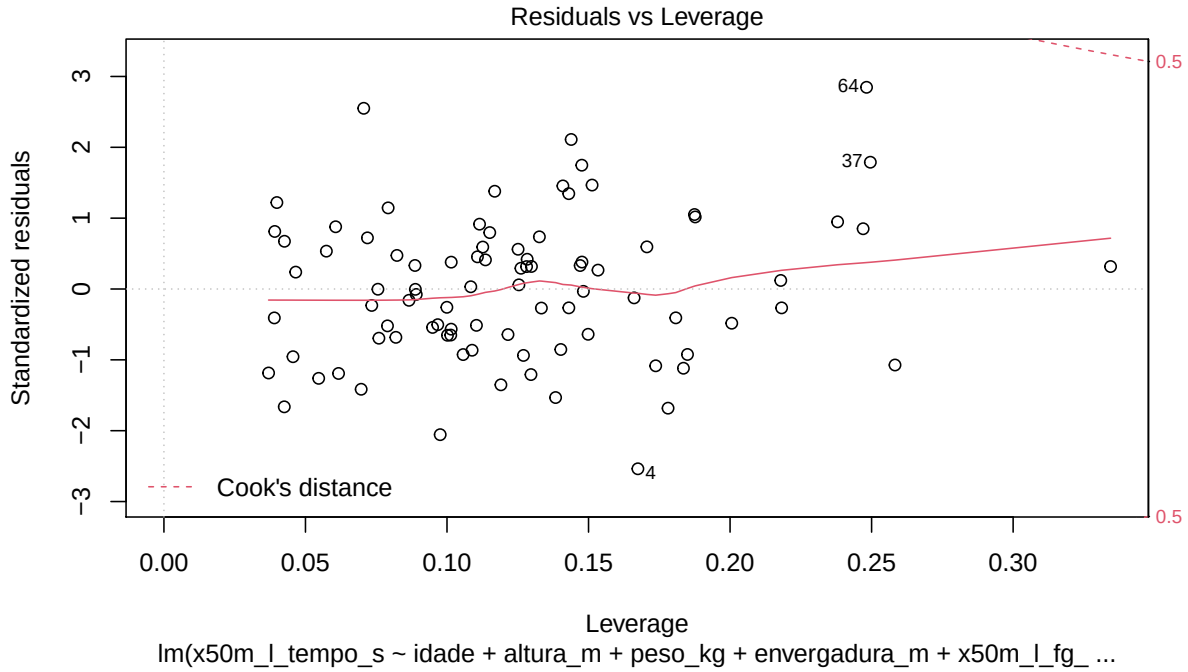
```
##  gvlma(x = myLModel, alphalevel = 0.05)
##
##                        Value p-value                Decision
## Global Stat        2.856e+00  0.5822 Assumptions acceptable.
## Skewness           5.328e-01  0.4654 Assumptions acceptable.
## Kurtosis           2.949e-06  0.9986 Assumptions acceptable.
## Link Function      2.295e+00  0.1298 Assumptions acceptable.
## Heteroscedasticity 2.803e-02  0.8670 Assumptions acceptable.
```

```
plot(myLModel)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m + x50m_l_fg_ ...
64
60
4


Scale−Location
√|Standardized residuals|
Fitted values
lm(x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m + x50m_l_fg_ ...
64
60
4



3.1 Model summary Masc

```
summary(myLModel)
```

```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ idade + altura_m + peso_kg + envergadura_m +
##     x50m_l_fg_ciclos_min_5_20m + x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m +
##     ms_ma_dia_dist_a_ncia_m, data = dfMas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90611 -0.24575 -0.00154  0.22816  0.96601
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   45.197947   8.420966   5.367 8.17e-07 ***
## idade                          0.053762   0.093637   0.574   0.5675
## altura_m                      -2.282374   1.374323  -1.661   0.1008
## peso_kg                       -0.002136   0.011016  -0.194   0.8467
## envergadura_m                  1.168755   1.067674   1.095   0.2771
## x50m_l_fg_ciclos_min_5_20m    -0.251845   0.079868  -3.153   0.0023 **
## x50m_l_dc_m                   -6.329625   3.927264  -1.612   0.1111
## x50m_l_in_equa_a_a_o           0.120339   1.163331   0.103   0.9179
## x50m_l_tempo_viragem_s_5_10m   1.223299   0.165025   7.413 1.36e-10 ***
## mi_ma_dia_dist_a_ncia_m       -0.224616   0.225603  -0.996   0.3226
```

```
## ms_ma_dia_dist_a_ncia_m      -0.058869   0.114466  -0.514   0.6085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3913 on 77 degrees of freedom
## Multiple R-squared:  0.8057, Adjusted R-squared:  0.7804
## F-statistic: 31.92 on 10 and 77 DF,  p-value: < 2.2e-16
```

3.2 Interpretações dos modelos

Intercept = expected x50m_l_tempo_s considering the average of all swimmers in the variables used

Slopes: - x50m_l_fg_ciclos_min_5_20m: para cada unidade x50m_l_fg_ciclos_min_5_20m o x50m_l_tempo_s reduz -0.251 - x50m_l_tempo_viragem_s_5_10m: para cada unidade freeSwimIndex50m o x50m_l_tempo_s reduz -0.55

Estes coeficientes não indicam a importância relativa de cada preditor para estimar a VD

- residuals:
  - difference between the actual observed response values and the response values that the model predicted
  - symmetrical distribution across these points on the mean value zero (0)
- coefficientes:
  - simple linear regression, the coefficients are two unknown constants that represent the intercept and slope terms in the linear model
  - find an intercept and a slope such that the resulting fitted line is as close as possible to the data points in our data set
- t value:
  - how many standard deviations our coefficient estimate is far away from 0
  - We want it to be far away from zero as this would indicate we could reject the null hypothesis
  - t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists. In general, t-values are also used to compute p-values.
- Pr(>t):
  - The Pr(>t) acronym found in the model output relates to the probability of observing any value equal or larger than t
  - A small p-value indicates that it is unlikely we will observe a relationship between the predictor and response variables due to chance.
  - Typically, a p-value of 5% or less is a good cut-off point
  - In our model example, the p-values are very close to zero. Note the 'signif. Codes' associated to each estimate.
  - Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis.

- residual std error:
  - measure of the quality of a linear regression fit
  - The Residual Standard Error is the average amount that the response (dist) will deviate from the true regression line
  - The Residual Standard Error was calculated with 164 degrees of freedom
  - degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters (restriction). In our case, we had 184 data points and 9 parameters
- r-squared:
  - R = coeficiente de correlação. Valores estimados vs valores observados (racio = VE/VO)
  - R Square = O quanto é que a variável dependente é explicada pelas variáveis utilizadas, mede a proporção da variação da variável dependente (t50mFree) que é explicada pelas variáveis independentes no modelo.
  - measure of how well the model is fitting the actual data
  - is a measure of the linear relationship between our predictor variable (speed) and our response / target variable (dist)
  - It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable)
  - In our example, the R2 we get is 0.6510794. Or roughly 65% of the variance found in the response variable can be explained by the predictor variable
  - A side note: In multiple regression settings, the R2 will always increase as more variables are included in the model.
- adjusted r-squared:
  - Adjusted R Square = medida a reportar para avaliação da qualidade do modelo, está corrigida para o número de variáveis independentes e n da amostra
  - is the preferred measure as it adjusts for the number of variables considered.
  - In multiple regression settings, the R2 will always increase as more variables are included in the model.
- f-statistics:
  - F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables
  - The further the F-statistic is from 1 the better it is
  - Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H0 : There is no relationship)

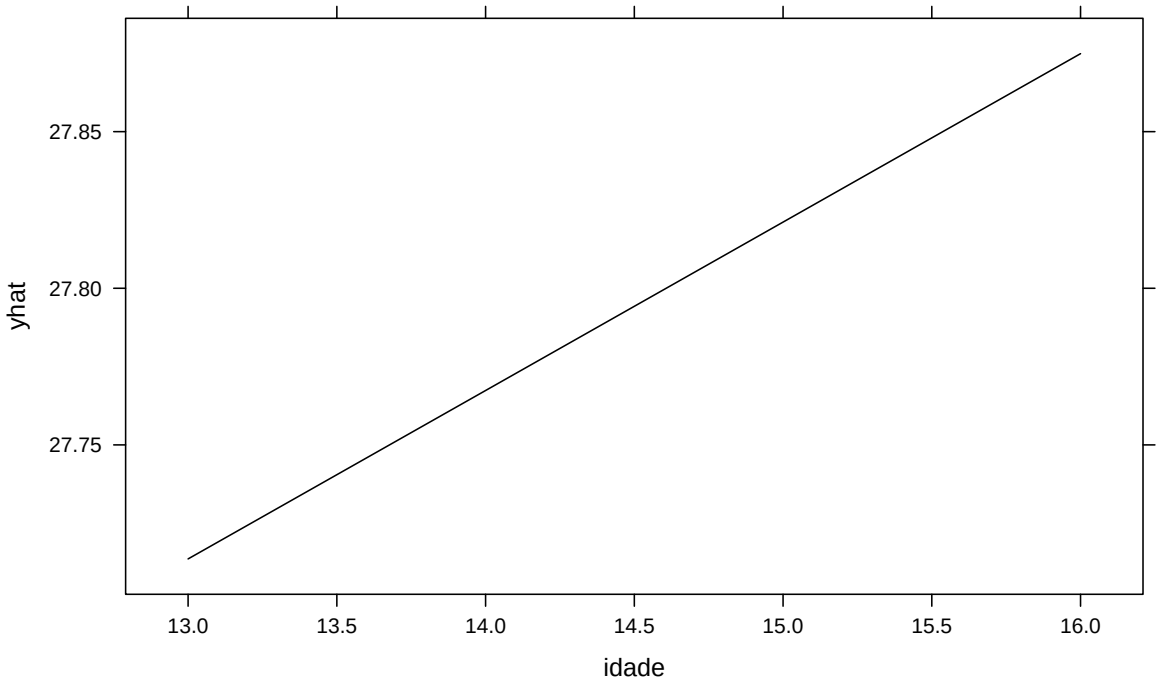
3.3 Partial Plots

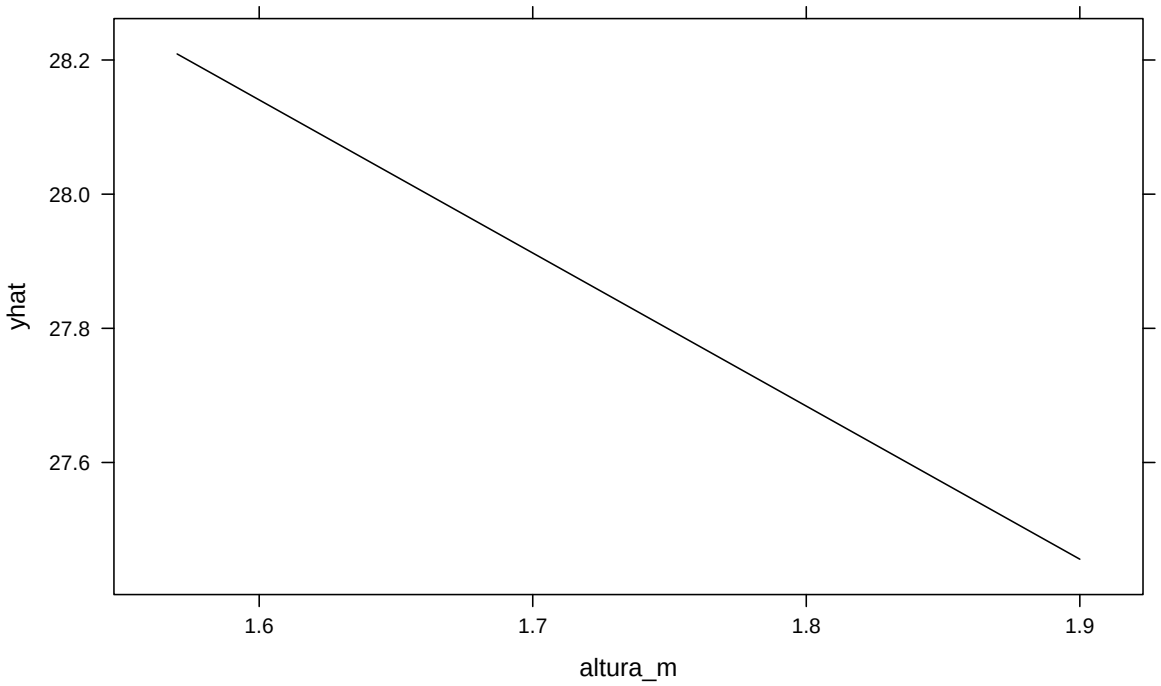
```
names(dfFem)
```

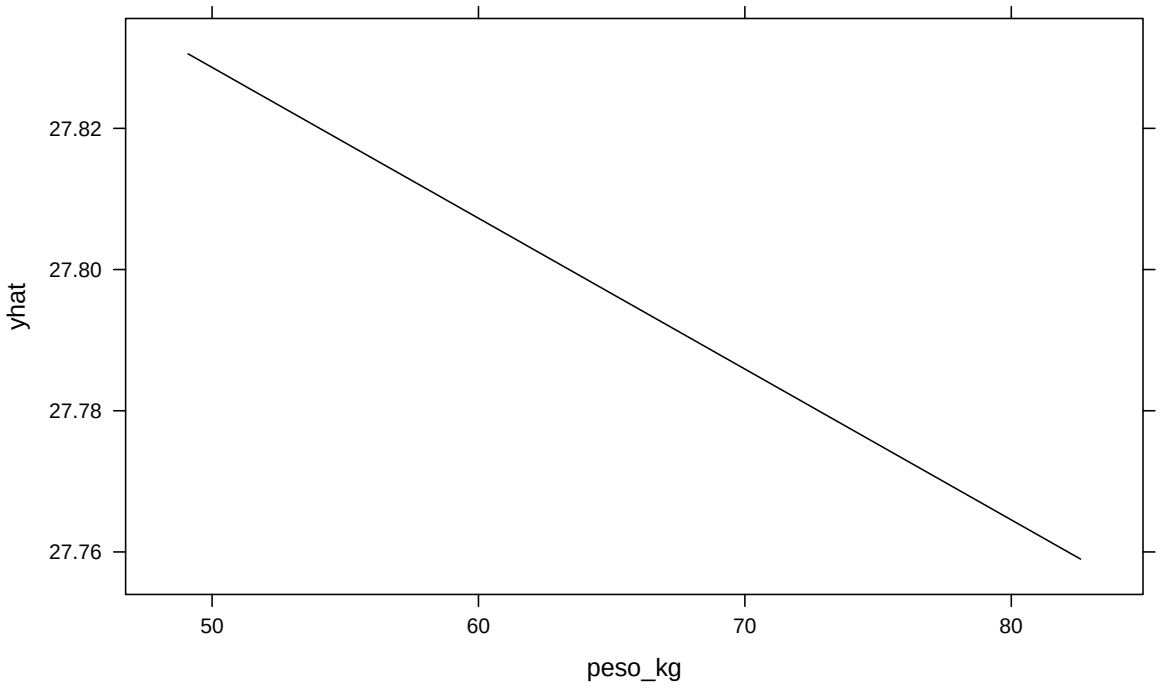
```
##  [1] "sexo"                          "idade"
##  [3] "altura_m"                      "peso_kg"
##  [5] "envergadura_m"                 "x50m_l_tempo_s"
```

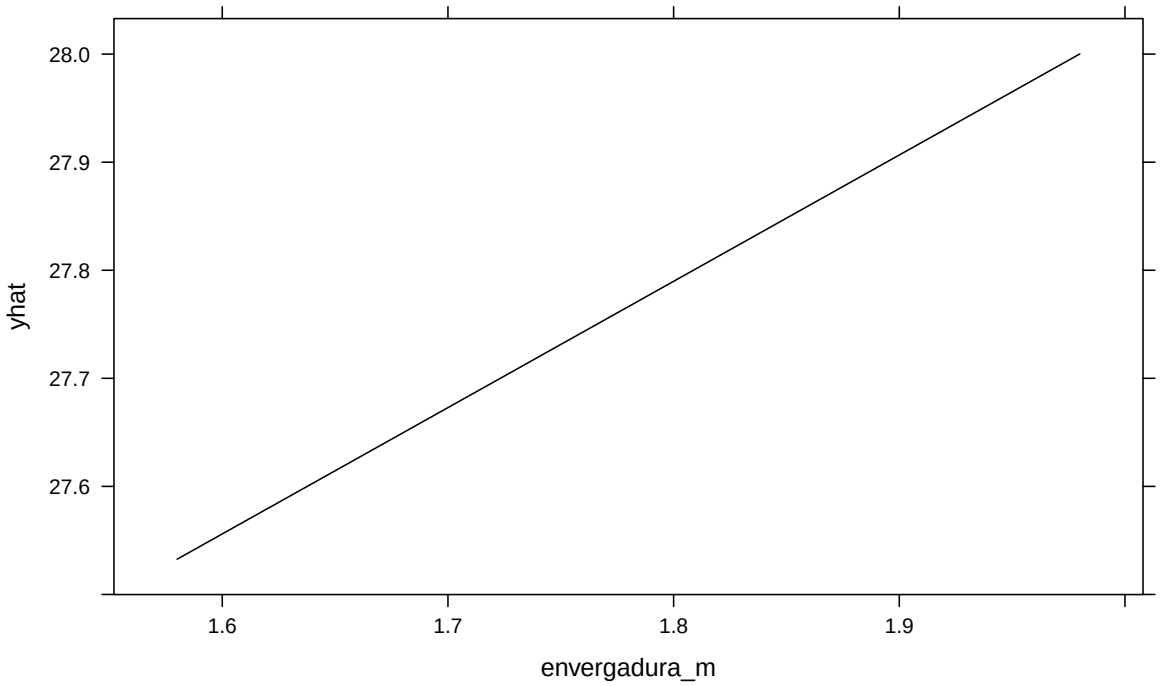
```
##  [7] "x50m_l_fg_ciclos_min_5_20m"   "x50m_l_dc_m"
##  [9] "x50m_l_in_equa_a_a_o"         "x50m_l_tempo_viragem_s_5_10m"
## [11] "mi_ma_dia_dist_a_ncia_m"      "ms_ma_dia_dist_a_ncia_m"
```

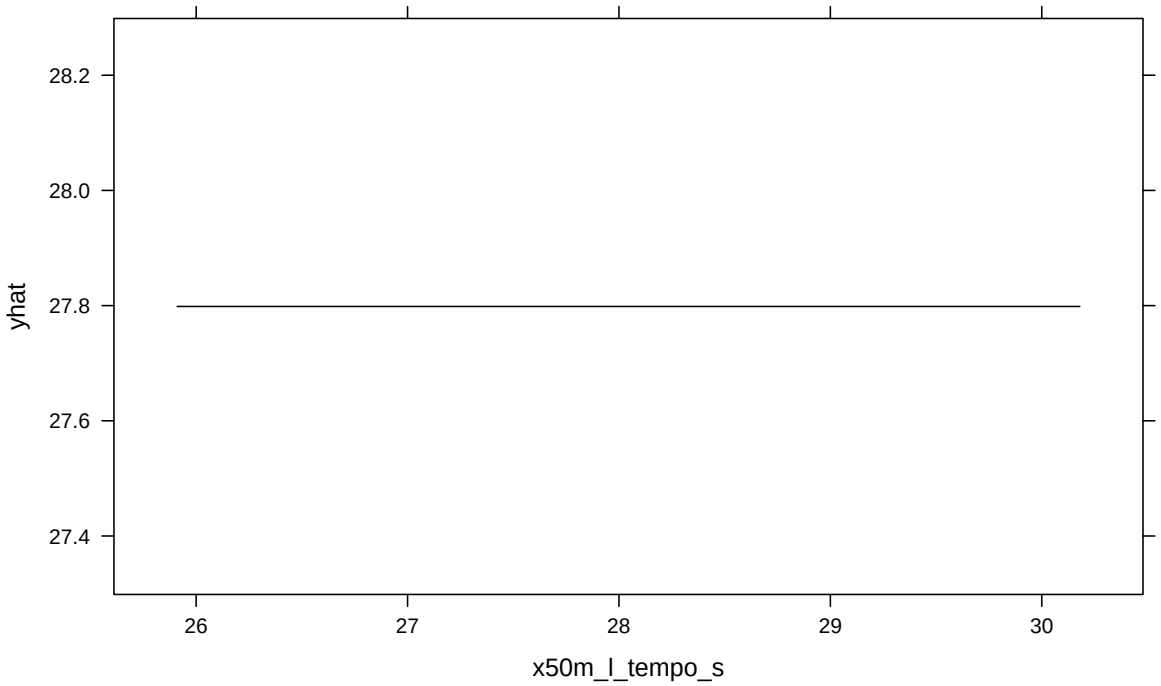
```
library(pdp)
vars=c("idade","altura_m","peso_kg","envergadura_m","x50m_l_tempo_s","x50m_l_fg_ciclos
       "x50m_l_in_equa_a_a_o","x50m_l_tempo_viragem_s_5_10m","mi_ma_dia_dist_a_ncia_m"
for (var in vars){
  print(partial(myLModel,pred.var = var,plot = TRUE))
}
```

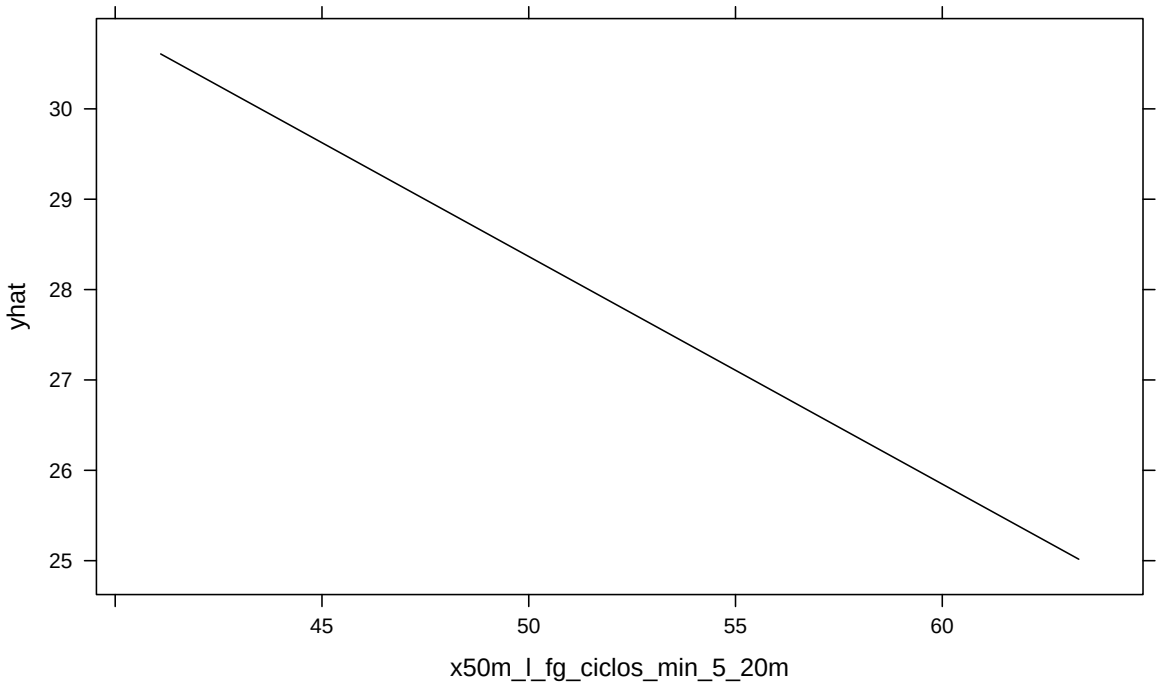


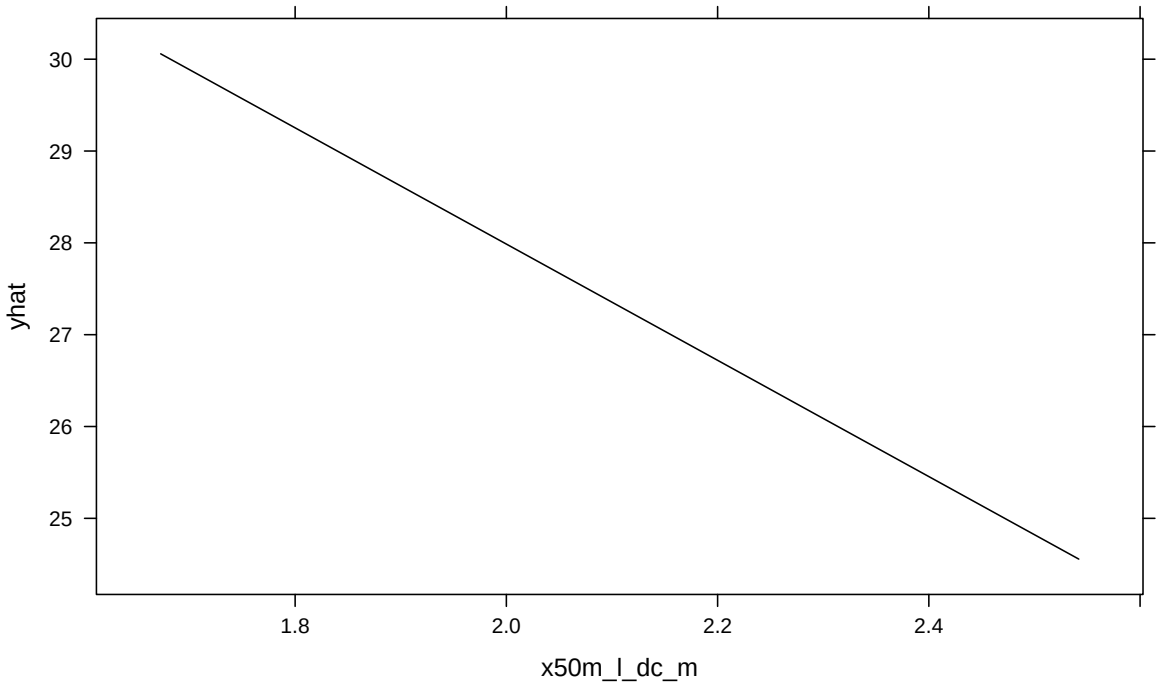

28.2
28.0
27.8
27.6
yhat
1.6
1.7
1.8
1.9
altura_m

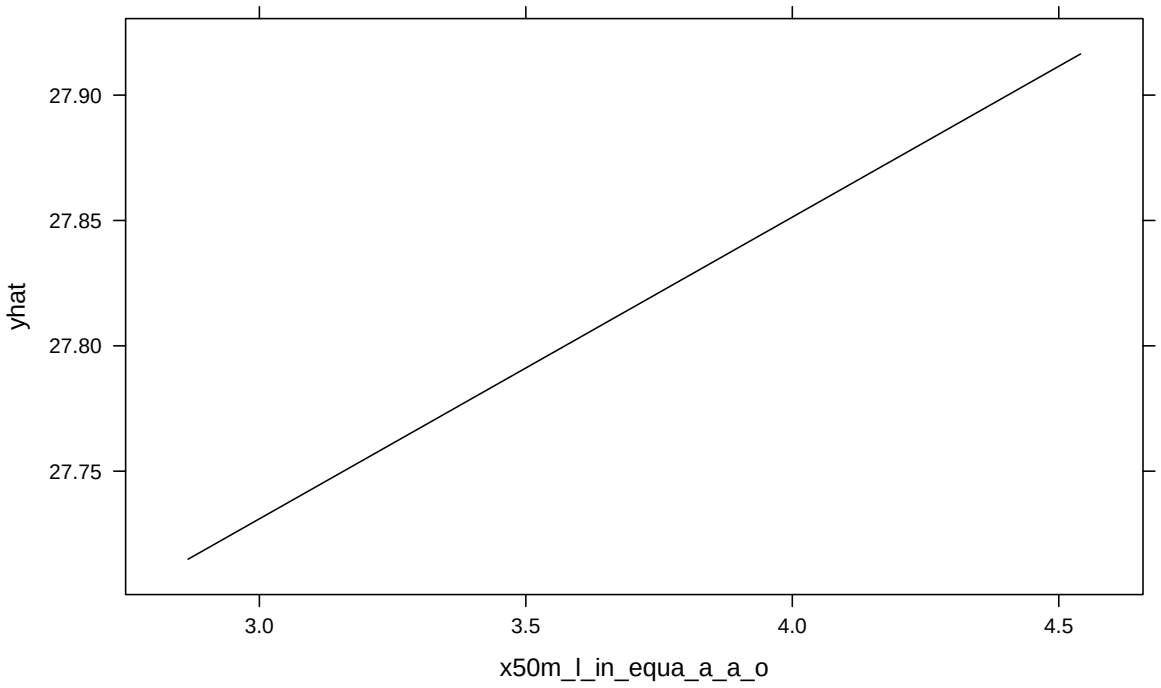

27.82
27.80
27.78
27.76
yhat
50
60
70
80
peso_kg

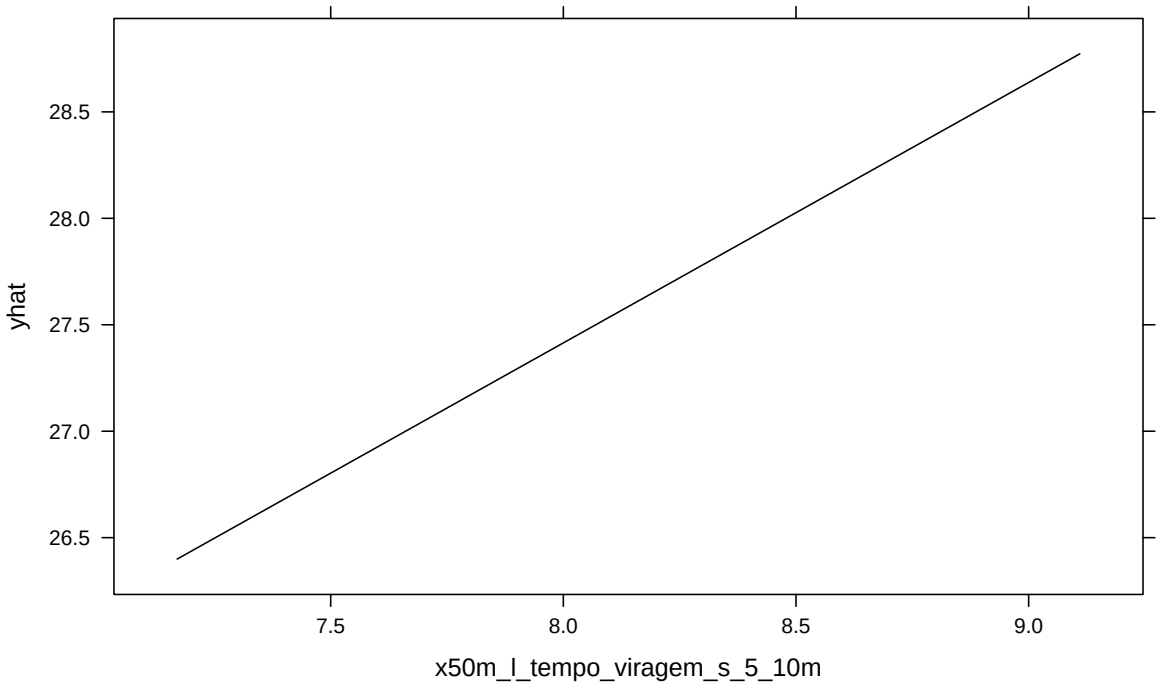

28.0
27.9
27.8
27.7
27.6
yhat
1.6
1.7
1.8
1.9
envergadura_m

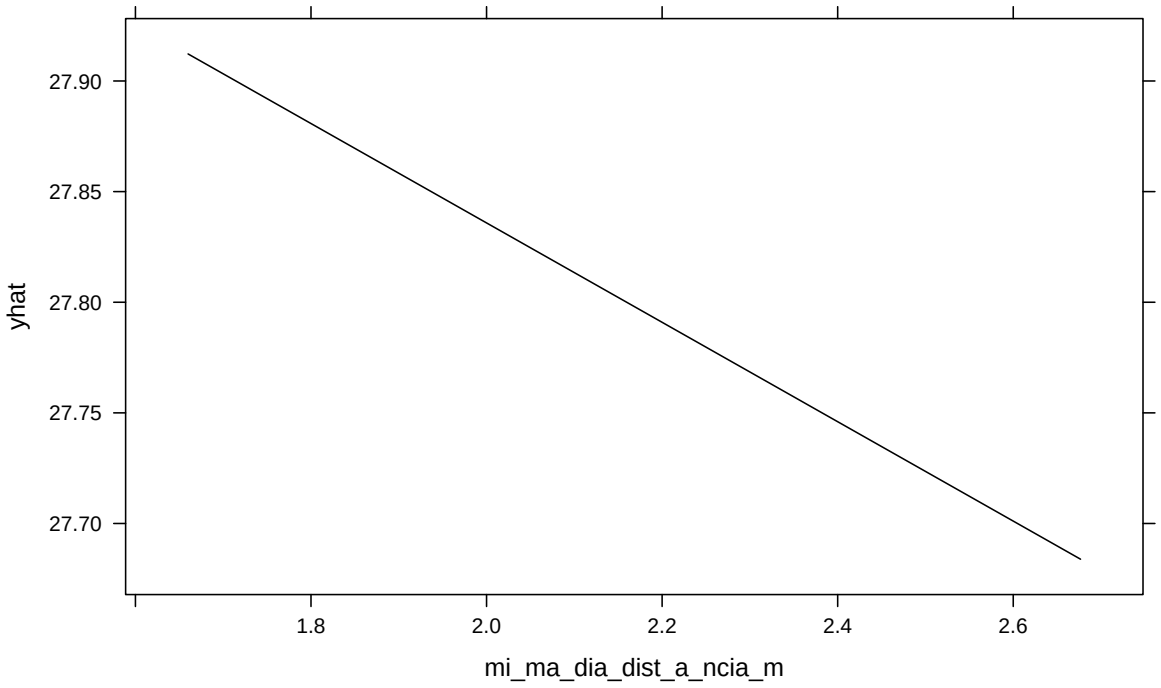

28.2
28.0
27.8
27.6
27.4
yhat
26
27
28
29
30
x50m_l_tempo_s

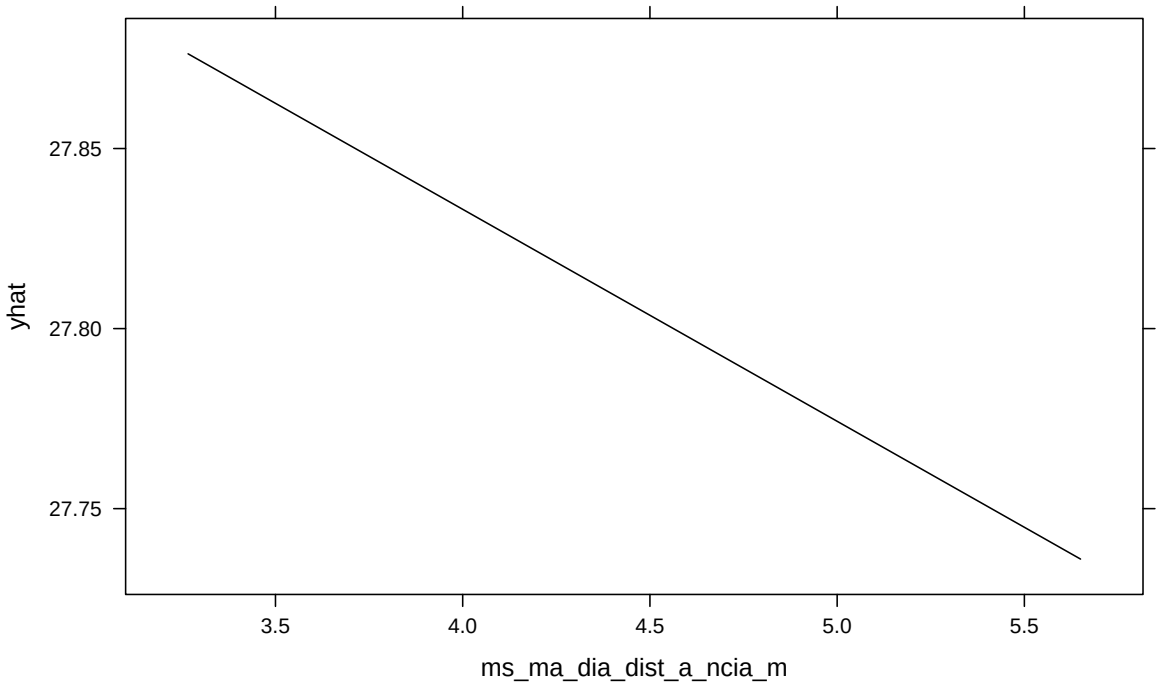

yhat
30
29
28
27
26
25
45
50
55
60
x50m_l_fg_ciclos_min_5_20m


yhat
30
29
28
27
26
25
1.8
2.0
2.2
2.4
x50m_l_dc_m


27.90
27.85
27.80
27.75
yhat
3.0
3.5
4.0
4.5
x50m_l_in_equa_a_a_o


28.5
28.0
27.5
27.0
26.5
yhat
7.5
8.0
8.5
9.0
x50m_l_tempo_viragem_s_5_10m





3.4 Best Model - Stepwise regression

Vamos que é o melhor modelo para prevermos o tempo nos 50 metros livres (t50mFree) testando todas as variáveis que temos disponíveis.

Vamos utilizar o modelo stepwise adicionando e removendo iterativamente variáveis preditoras (predictors) no modelo para identificar um subconjunto de variáveis que tem a melhor desempenho a prever o model, que é o modelo que tem um erro menor na previsão.

Existem três estratégias (James et al. 2014;P. Bruce and Bruce 2017):

- Forward selection: inicia sem preditores no modelo e iterativamente adiciona o que mais contribui para a previsão parando quando não existem melhorias estatisticamente significativas;
- Backward selection (or backward elimination): começa com todos os preditores no modelo (*full model*), iterativamente remove os que menos contribuem para a previsão. Para quando todos os preditores são significativos;
- Stepwise selection: combinação de forward e backward selections. Quando se começa sem variáveis preditoras e sequencialmente são adicionados os preditores que mais contribuem como a estratégia Forward selection. Depois de adicionar cada variável, são removidas as variáveis que não melhoram o modelo utilizando a aproximação backward selection;

Bruce, P., & Bruce, A. (2017). *Practical Statistics for Data Scientists: 50 Essential Concepts* (1st edition). O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R* (1st ed. 2013, Corr. 7th printing 2017 edition). Springer.

```
library(MASS)

# Fit do modelo com todas
dfTodas <- lm(x50m_l_tempo_s ~., data = dfFem)
# Stepwise regression model
stepModel <- stepAIC(dfTodas, direction = "both",  trace = FALSE,)
summary(stepModel)
```

```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ altura_m + peso_kg + x50m_l_dc_m +
##     x50m_l_in_equa_a_a_o + x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m,
##     data = dfFem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90212 -0.28190  0.01552  0.24554  0.78906
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    13.70548    1.97551   6.938 8.47e-10 ***
## altura_m                        2.92179    0.99315   2.942  0.00424 **
## peso_kg                        -0.02572    0.01188  -2.164  0.03335 *
## x50m_l_dc_m                     8.07194    0.92927   8.686 3.03e-13 ***
## x50m_l_in_equa_a_a_o           -5.04402    0.51157  -9.860 1.41e-15 ***
## x50m_l_tempo_viragem_s_5_10m    1.59517    0.14386  11.088  < 2e-16 ***
## mi_ma_dia_dist_a_ncia_m        -0.64419    0.25575  -2.519  0.01372 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3969 on 82 degrees of freedom
## Multiple R-squared:  0.8955, Adjusted R-squared:  0.8878
## F-statistic: 117.1 on 6 and 82 DF,  p-value: < 2.2e-16
```

3.5 Qual é o melhor modelo considerando todas as variáveis disponíveis para prever?

```
library(caret)
library(leaps)

models <- regsubsets(x50m_l_tempo_s~., data = dfFem, nvmax = 10, method = "seqrep")
```

```
## Reordering variables and trying again:
```

```
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(x50m_l_tempo_s ~ ., data = dfFem, nvmax = 10,
##     method = "seqrep")
## 11 Variables  (and intercept)
##                               Forced in Forced out
## idade                             FALSE      FALSE
## altura_m                          FALSE      FALSE
## peso_kg                           FALSE      FALSE
## envergadura_m                     FALSE      FALSE
## x50m_l_fg_ciclos_min_5_20m        FALSE      FALSE
## x50m_l_dc_m                       FALSE      FALSE
## x50m_l_in_equa_a_a_o              FALSE      FALSE
## x50m_l_tempo_viragem_s_5_10m      FALSE      FALSE
## mi_ma_dia_dist_a_ncia_m           FALSE      FALSE
## ms_ma_dia_dist_a_ncia_m           FALSE      FALSE
## sexo                              FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: 'sequential replacement'
##           sexo idade altura_m peso_kg envergadura_m x50m_l_fg_ciclos_min_5_20m
## 1  ( 1 )  " "  " "    " "      " "     " "           " "
## 2  ( 1 )  " "  " "    " "      " "     " "           " "
## 3  ( 1 )  " "  " "    " "      " "     " "           " "
## 4  ( 1 )  " "  " "    " "      " "     " "           " "
## 5  ( 1 )  " "  " "    "*"      " "     " "           " "
## 6  ( 1 )  " "  " "    "*"      "*"     " "           " "
## 7  ( 1 )  " "  " "    "*"      "*"     "*"           " "
## 8  ( 1 )  " "  " "    "*"      "*"     "*"           " "
## 9  ( 1 )  " "  " "    "*"      "*"     "*"           "*"
## 10  ( 1 ) " "  "*"    "*"      "*"     "*"           "*"
##           x50m_l_dc_m x50m_l_in_equa_a_a_o x50m_l_tempo_viragem_s_5_10m
## 1  ( 1 )  " "         " "                 "*"
## 2  ( 1 )  " "         "*"                 "*"
```

```
## 3  ( 1 )  "*"         "*"                  "*"
## 4  ( 1 )  "*"         "*"                  "*"
## 5  ( 1 )  "*"         "*"                  "*"
## 6  ( 1 )  "*"         "*"                  "*"
## 7  ( 1 )  "*"         "*"                  "*"
## 8  ( 1 )  "*"         "*"                  "*"
## 9  ( 1 )  "*"         "*"                  "*"
## 10  ( 1 ) "*"         "*"                  "*"
##           mi_ma_dia_dist_a_ncia_m ms_ma_dia_dist_a_ncia_m
## 1  ( 1 )  " "                     " "
## 2  ( 1 )  " "                     " "
## 3  ( 1 )  " "                     " "
## 4  ( 1 )  "*"                     " "
## 5  ( 1 )  "*"                     " "
## 6  ( 1 )  "*"                     " "
## 7  ( 1 )  "*"                     " "
## 8  ( 1 )  "*"                     "*"
## 9  ( 1 )  "*"                     "*"
## 10  ( 1 ) "*"                     "*"
```

```
# Set seed for reproducibility
set.seed(123)
# Set up repeated k-fold cross-validation
train.control <- trainControl(method = "cv", number = 10)
# Train the model
stepModel <- train(x50m_l_tempo_s ~., data = dfFem,method = "leapBackward",
                   tuneGrid = data.frame(nvmax = 1:5),
                   trControl = train.control
                   )
```

```
## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:
```

```
## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, : 1
## linear dependencies found

## Reordering variables and trying again:
```

```
stepModel$results
```

```
##   nvmax      RMSE  Rsquared       MAE    RMSESD RsquaredSD      MAESD
## 1     1 1.0366751 0.2369193 0.8121891 0.2657856  0.1918498 0.16306810
## 2     2 0.5933520 0.7287311 0.4741489 0.1141935  0.2006436 0.09312512
## 3     3 0.5376166 0.7570933 0.4318687 0.1177486  0.1905434 0.09892884
## 4     4 0.5487060 0.7465806 0.4423472 0.1214231  0.1922110 0.10427331
## 5     5 0.5532285 0.7477263 0.4495423 0.1270478  0.1943042 0.10282893
```

Quantas variáveis tem o melhor modelo? O que têm o RMSE e o MAE é o utilizado normalmente. R^2 indica a correlação entre as preditoras e a predicted (resultado), quanto mais alto melhor.

```
stepModel$bestTune
```

```
##   nvmax
## 3     3
```

3.6 Summary do melhor modelo

```
summary(stepModel$finalModel)
```

```
## Subset selection object
## 11 Variables  (and intercept)
##                              Forced in Forced out
```

```
## idade                              FALSE      FALSE
## altura_m                           FALSE      FALSE
## peso_kg                            FALSE      FALSE
## envergadura_m                      FALSE      FALSE
## x50m_l_fg_ciclos_min_5_20m         FALSE      FALSE
## x50m_l_dc_m                        FALSE      FALSE
## x50m_l_in_equa_a_a_o               FALSE      FALSE
## x50m_l_tempo_viragem_s_5_10m       FALSE      FALSE
## mi_ma_dia_dist_a_ncia_m            FALSE      FALSE
## ms_ma_dia_dist_a_ncia_m            FALSE      FALSE
## sexo                               FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: backward
##          sexo idade altura_m peso_kg envergadura_m x50m_l_fg_ciclos_min_5_20m
## 1  ( 1 ) " "   " "   " "      " "     " "           " "
## 2  ( 1 ) " "   " "   " "      " "     " "           " "
## 3  ( 1 ) " "   " "   " "      " "     " "           " "
## 4  ( 1 ) " "   " "   " "      " "     " "           " "
##          x50m_l_dc_m x50m_l_in_equa_a_a_o x50m_l_tempo_viragem_s_5_10m
## 1  ( 1 ) " "         " "                 "*"
## 2  ( 1 ) " "         "*"                 "*"
## 3  ( 1 ) "*"         "*"                 "*"
## 4  ( 1 ) "*"         "*"                 "*"
##          mi_ma_dia_dist_a_ncia_m ms_ma_dia_dist_a_ncia_m
## 1  ( 1 ) " "                     " "
## 2  ( 1 ) " "                     " "
## 3  ( 1 ) " "                     " "
## 4  ( 1 ) "*"                     " "
```

O melhor modelo contêm as variáveis x50m_l_dc_m, x50m_l_in_equa_a_a_o, x50m_l_tempo_viragem_s_5_10m e mi_ma_dia_dist_a_ncia_m

```
myLModel <- lm(x50m_l_tempo_s ~ x50m_l_dc_m+x50m_l_in_equa_a_a_o+x50m_l_tempo_viragem_
summary(myLModel)
```

```
##
## Call:
## lm(formula = x50m_l_tempo_s ~ x50m_l_dc_m + x50m_l_in_equa_a_a_o +
##     x50m_l_tempo_viragem_s_5_10m + mi_ma_dia_dist_a_ncia_m, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6998 -0.3912 -0.0554  0.3505  3.7954
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     12.5946     1.7502   7.196 1.75e-11 ***
## x50m_l_dc_m                     -0.1008     0.7044  -0.143    0.886
## x50m_l_in_equa_a_a_o            -0.1316     0.3244  -0.406    0.685
```

```
## x50m_l_tempo_viragem_s_5_10m   2.2840     0.1538  14.852  < 2e-16 ***
## mi_ma_dia_dist_a_ncia_m       -1.3683     0.2563  -5.339 2.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6936 on 175 degrees of freedom
## Multiple R-squared:  0.8819, Adjusted R-squared:  0.8792
## F-statistic: 326.7 on 4 and 175 DF,  p-value: < 2.2e-16
```