Previsão tempo de nado

Pedro Sobreiro

Abstract

TODO.

1 Previsão do tempo de nado

1.1 Descrição variáveis

- 50mFreeTime - Tempo total dos 50m Livres segundos
- 50mFreeTime5.20 - Tempo retirado entre os 5 e 20 metros, o qual foi utilizado para calcular a velocidade de nado, a frequência gestual, a distância de ciclo e o índice de nado segundos
- 50mFreeVelocity - Velocidade de nado calculada por 15/pelo variável anterior (tempo entre os 5 e 20m) m/s
- 50mFreeStrokeRate - Frequência gestual, retirada com o cronômetro entre os 5m e 20m ciclos/min
- 50mFreeStrokeLength - Distância de ciclo calculada por [(60xvelocidade)/frequência gestual] = [(60x@50mFreeVelocity)/@50mFreeStrokeRate] metros
- 50mFreeSwimIndex - Índice de nado calculada por [velocidadexdistância de ciclo] = [@50mFreeVelocityx@50mFreeStrokeLength]
- 50mFreeTurnTime5.10m - Tempo de viragem 5m antes da viragem + 10m depois da viragem segundos
- 50mFreeTurnIndex - Índice de viragem calculado pela divisão entre o tempo de de nado (entre os 5m e 20m) e o tempo da viragem = @50mFree-Time5.20/@50mFreeTurnTime
- SECO.MS1.Distância e as restantes são as tentativas que foram realizadas de força em seco. MS = membros superiores - lançamento da bola medicinal e MI = membros inferiores - salto horizontal, 1, 2 e 3 são as 3 tentativas
- LMAverageDistance - É a média da distância das 3 tentativas para os membros inferiores
- LMBestDistance - Foi a melhor distância encontrada do atleta nas 3 tentativas possiveis nos membros inferiores
- ULAverageDistance - É a média da distância das 3 tentativas para os membros superiores
- ULBestDistance - Foi a melhor distância encontrada do atleta nas 3 tentativas possíveis nos membros superiores

1.2 Abrir ficheiros

```
library(foreign)
library(writexl)

print(caminho)
```

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

```
dadosSPSS<-read.spss("../dados/BaseDadosEstudoMestrado.sav",to.data.frame = TRUE)
names(dadosSPSS)<-c("num","sportSeason","name", "esc", "nFPN", "club", "assoc","gender
                    "ageDec","height","weight","bodyMass","wingSpan","wingSpanHeight"
                    "adultHeight","adultHeightLevel", "t50mFree","t50mFree5.20m","free
                    "freeStrokeLength50m","freeSwimIndex50m","freeTurnTime5.10m_50m","
                    "SecoMI2Dist","SecoMI3Dist","LMAveDist","LMBestDist","SECOMS1Dist"
                    "ULAveDist","ULBestDist")
str(dadosSPSS)
```

```
## 'data.frame':    184 obs. of  38 variables:
##  $ num                  : num  29 30 31 32 33 34 35 36 37 38 ...
##  $ sportSeason          : Factor w/ 4 levels "2014-2015","2015-2016",..: 1 1 1 1 1
##  $ name                 : chr  "Afonso Meneses Sequeira     " "José Pedro Melo
##  $ esc                  : Factor w/ 2 levels "Juvenil A","Juvenil B": 1 1 1 1 1 1 1
##  $ nFPN                 : chr  "105711  " "117922  " "119635  " "107735  " ...
##  $ club                 : chr  "CNF            " "SCB            " "GDNVNF
##  $ assoc                : chr  "ANMAD     " "ANMIN     " "ANNP      " "ANNP      "
##  $ gender               : Factor w/ 2 levels "Male","Female": 1 1 1 1 1 1 1 1 1 1
##  $ yearOfBirth          : num  1999 1999 1999 1999 1999 ...
##  $ age                  : num  15 15 15 15 15 15 15 15 15 15 ...
##  $ ageDec               : num  15.8 15.7 15.8 16 15.8 ...
##  $ height               : num  1.72 1.73 1.81 1.79 1.83 1.84 1.66 1.89 1.75 1.82 .
##  $ weight               : num  65.3 64.9 75.4 69.4 69.9 66.5 60 73 57 77.4 ...
##  $ bodyMass             : num  22.1 21.7 23 21.7 20.9 ...
##  $ wingSpan             : num  1.75 1.76 1.85 1.93 1.85 1.81 1.75 1.92 1.84 1.8 ..
##  $ wingSpanHeight       : num  1.02 1.02 1.02 1.08 1.01 ...
##  $ fatherHeight         : num  1.73 1.79 1.85 1.79 1.85 1.78 1.73 1.87 1.75 1.84 .
##  $ motherHeight         : num  1.72 1.69 1.71 1.63 1.64 1.66 1.65 1.75 1.61 1.73 .
##  $ adultHeight          : num  1.75 1.8 1.85 1.8 1.85 1.85 1.73 1.93 1.78 1.85 ...
##  $ adultHeightLevel     : num  98.3 96.1 97.8 99.4 98.9 ...
##  $ t50mFree             : num  27.6 28.5 27.7 28.1 26.2 ...
##  $ t50mFree5.20m        : num  8.41 8.79 8.47 8.89 8.6 8.72 8.5 8.65 8.5 8.63 ...
##  $ freeVelocity50m      : num  1.78 1.71 1.77 1.69 1.74 ...
##  $ freeStrokeRate50m    : num  55.7 48.5 57.7 43.2 50.9 ...
##  $ freeStrokeLength50m  : num  1.92 2.11 1.84 2.35 2.06 ...
##  $ freeSwimIndex50m     : num  3.42 3.6 3.26 3.96 3.59 ...
##  $ freeTurnTime5.10m_50m: num  8.07 9.04 8.4 8.9 8.6 8.32 8.78 8.07 8.4 8 ...
##  $ freeTurnIndex50m     : num  1.042 0.972 1.008 0.999 1 ...
##  $ SecoMI1Dist          : num  1.75 2.3 1.78 1.8 1.97 2.1 2 1.7 2.2 1.99 ...
##  $ SecoMI2Dist          : num  1.75 2.32 1.8 1.95 1.9 2.25 2.12 1.65 2.22 2 ...
```

```
##  $ SecoMI3Dist          : num  1.7 2.3 1.78 1.9 1.89 2.2 2.08 1.65 2.3 2.17 ...
##  $ LMAveDist            : num  1.73 2.31 1.79 1.88 1.92 ...
##  $ LMBestDist           : num  1.75 2.32 1.8 1.95 1.97 2.25 2.12 1.7 2.3 2.17 ...
##  $ SECOMS1Dist          : num  4.35 4.2 5.4 4.2 5 5 4 4.6 4.5 4.6 ...
##  $ SECOMS2Dist          : num  4.3 4.8 5.82 4.5 4.7 5.1 3.9 4.8 4.4 4.5 ...
##  $ SECOMS3Dist          : num  5 5 5.6 4.3 5 5.2 4.05 5 4.55 4.6 ...
##  $ ULAveDist            : num  4.55 4.67 5.61 4.33 4.9 ...
##  $ ULBestDist           : num  5 5 5.82 4.5 5 5.2 4.05 5 4.55 4.6 ...
##  - attr(*, "variable.labels")= Named chr [1:38] "N" "Sport Season" "Name" "" ...
##   ..- attr(*, "names")= chr [1:38] "N" "SportSeason" "Name" "Escalão" ...
##  - attr(*, "codepage")= int 65001
```

```
write_xlsx(dadosSPSS,"../dados/dadosNadadores.xlsx")
```

1.3 Selecionar as variáveis

```
names(x = dadosSPSS)
```

```
##  [1] "num"                  "sportSeason"          "name"
##  [4] "esc"                  "nFPN"                 "club"
##  [7] "assoc"                "gender"               "yearOfBirth"
## [10] "age"                  "ageDec"               "height"
## [13] "weight"               "bodyMass"             "wingSpan"
## [16] "wingSpanHeight"       "fatherHeight"         "motherHeight"
## [19] "adultHeight"          "adultHeightLevel"     "t50mFree"
## [22] "t50mFree5.20m"        "freeVelocity50m"      "freeStrokeRate50m"
## [25] "freeStrokeLength50m"  "freeSwimIndex50m"     "freeTurnTime5.10m_50m"
## [28] "freeTurnIndex50m"     "SecoMI1Dist"          "SecoMI2Dist"
## [31] "SecoMI3Dist"          "LMAveDist"            "LMBestDist"
## [34] "SECOMS1Dist"          "SECOMS2Dist"          "SECOMS3Dist"
## [37] "ULAveDist"            "ULBestDist"
```

Vamos utilizar as variáveis: age; height; weight; wingSpan; 50mFreeTime; 50mFreeTime5.20; 50mFreeVelocity; 50mFreeStrokeRate; 50mFreeStrokeLength; 50mFreeSwimIndex; 50mFreeTurnTime5.10m; 50mFreeTurnIndex; LMAveDist; ULAveDist

```
library(dplyr)


df<- dadosSPSS %>%
  select(age,height,weight,wingSpan,t50mFree, t50mFree5.20m,freeVelocity50m,freeStroke
         freeStrokeLength50m,freeSwimIndex50m,freeTurnTime5.10m_50m, freeTurnIndex50m

str(df)
```

```
## 'data.frame':    184 obs. of  14 variables:
##  $ age                  : num  15 15 15 15 15 15 15 15 15 15 ...
##  $ height               : num  1.72 1.73 1.81 1.79 1.83 1.84 1.66 1.89 1.75 1.82 .
##  $ weight               : num  65.3 64.9 75.4 69.4 69.9 66.5 60 73 57 77.4 ...
##  $ wingSpan             : num  1.75 1.76 1.85 1.93 1.85 1.81 1.75 1.92 1.84 1.8 ..
```

```
##  $ t50mFree             : num  27.6 28.5 27.7 28.1 26.2 ...
##  $ t50mFree5.20m        : num  8.41 8.79 8.47 8.89 8.6 8.72 8.5 8.65 8.5 8.63 ...
##  $ freeVelocity50m      : num  1.78 1.71 1.77 1.69 1.74 ...
##  $ freeStrokeRate50m    : num  55.7 48.5 57.7 43.2 50.9 ...
##  $ freeStrokeLength50m  : num  1.92 2.11 1.84 2.35 2.06 ...
##  $ freeSwimIndex50m     : num  3.42 3.6 3.26 3.96 3.59 ...
##  $ freeTurnTime5.10m_50m: num  8.07 9.04 8.4 8.9 8.6 8.32 8.78 8.07 8.4 8 ...
##  $ freeTurnIndex50m     : num  1.042 0.972 1.008 0.999 1 ...
##  $ LMAveDist            : num  1.73 2.31 1.79 1.88 1.92 ...
##  $ ULAveDist            : num  4.55 4.67 5.61 4.33 4.9 ...
```

1.4 Descritivas

```
summary(df)
```

```
##       age            height          weight         wingSpan
##  Min.   :12.00   Min.   :1.490   Min.   :42.70   Min.   :1.470
##  1st Qu.:14.00   1st Qu.:1.639   1st Qu.:53.00   1st Qu.:1.650
##  Median :14.00   Median :1.688   Median :57.95   Median :1.730
##  Mean   :14.07   Mean   :1.693   Mean   :58.72   Mean   :1.732
##  3rd Qu.:15.00   3rd Qu.:1.755   3rd Qu.:63.70   3rd Qu.:1.810
##  Max.   :16.00   Max.   :1.900   Max.   :82.60   Max.   :1.980
##
##     t50mFree     t50mFree5.20m    freeVelocity50m freeStrokeRate50m
##  Min.   :25.91   Min.   : 5.900   Min.   :1.372   Min.   :37.14
##  1st Qu.:27.72   1st Qu.: 8.540   1st Qu.:1.572   1st Qu.:48.40
##  Median :29.29   Median : 8.960   Median :1.674   Median :50.99
##  Mean   :29.45   Mean   : 9.037   Mean   :1.669   Mean   :51.12
##  3rd Qu.:30.99   3rd Qu.: 9.540   3rd Qu.:1.756   3rd Qu.:54.00
##  Max.   :36.35   Max.   :10.930   Max.   :2.542   Max.   :65.50
##  NA's   :1       NA's   :1        NA's   :1       NA's   :1
##  freeStrokeLength50m freeSwimIndex50m freeTurnTime5.10m_50m freeTurnIndex50m
##  Min.   :1.350       Min.   :1.989    Min.   : 6.450        Min.   :0.9147
##  1st Qu.:1.843       1st Qu.:2.952    1st Qu.: 8.290        1st Qu.:1.0057
##  Median :1.922       Median :3.256    Median : 8.910        Median :1.0263
##  Mean   :1.973       Mean   :3.308    Mean   : 8.809        Mean   :1.0270
##  3rd Qu.:2.086       3rd Qu.:3.600    3rd Qu.: 9.320        3rd Qu.:1.0490
##  Max.   :2.956       Max.   :7.516    Max.   :11.280        Max.   :1.2392
##  NA's   :1           NA's   :1        NA's   :1             NA's   :1
##    LMAveDist       ULAveDist
##  Min.   :1.227   Min.   :2.133
##  1st Qu.:1.684   1st Qu.:3.182
##  Median :1.930   Median :3.767
##  Mean   :1.937   Mean   :3.890
##  3rd Qu.:2.182   3rd Qu.:4.587
##  Max.   :2.677   Max.   :5.650
##  NA's   :2       NA's   :1
```

1.5 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

```
library(gvlma)
myLModel <- lm(t50mFree ~ ULAveDist+LMAveDist+freeSwimIndex50m+wingSpan+age+freeTurnTi
summary(myLModel)
```

```
##
## Call:
## lm(formula = t50mFree ~ ULAveDist + LMAveDist + freeSwimIndex50m +
##     wingSpan + age + freeTurnTime5.10m_50m + height + weight,
##     data = df)
##
## Residuals:
```

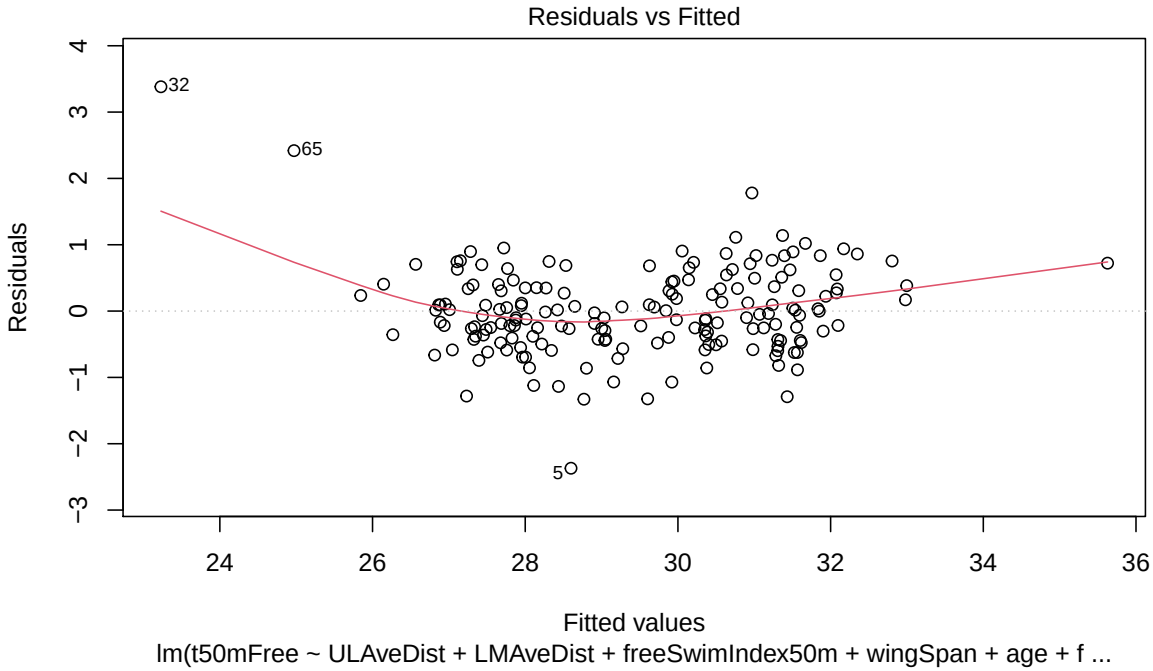
```
##     Min      1Q  Median      3Q     Max
## -2.3696 -0.4246 -0.0432  0.3717  3.3818
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           17.54767    2.70424   6.489 8.98e-10 ***
## ULAveDist             -0.12775    0.13428  -0.951 0.342759
## LMAveDist             -1.04865    0.27032  -3.879 0.000149 ***
## freeSwimIndex50m      -0.10985    0.11790  -0.932 0.352779
## wingSpan              -1.30271    1.06994  -1.218 0.225071
## age                   -0.12676    0.10165  -1.247 0.214087
## freeTurnTime5.10m_50m  2.05565    0.15428  13.324  < 2e-16 ***
## height                 0.69057    1.42375   0.485 0.628271
## weight                -0.00780    0.01388  -0.562 0.574904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6763 on 171 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8903, Adjusted R-squared:  0.8852
## F-statistic: 173.5 on 8 and 171 DF,  p-value: < 2.2e-16
```

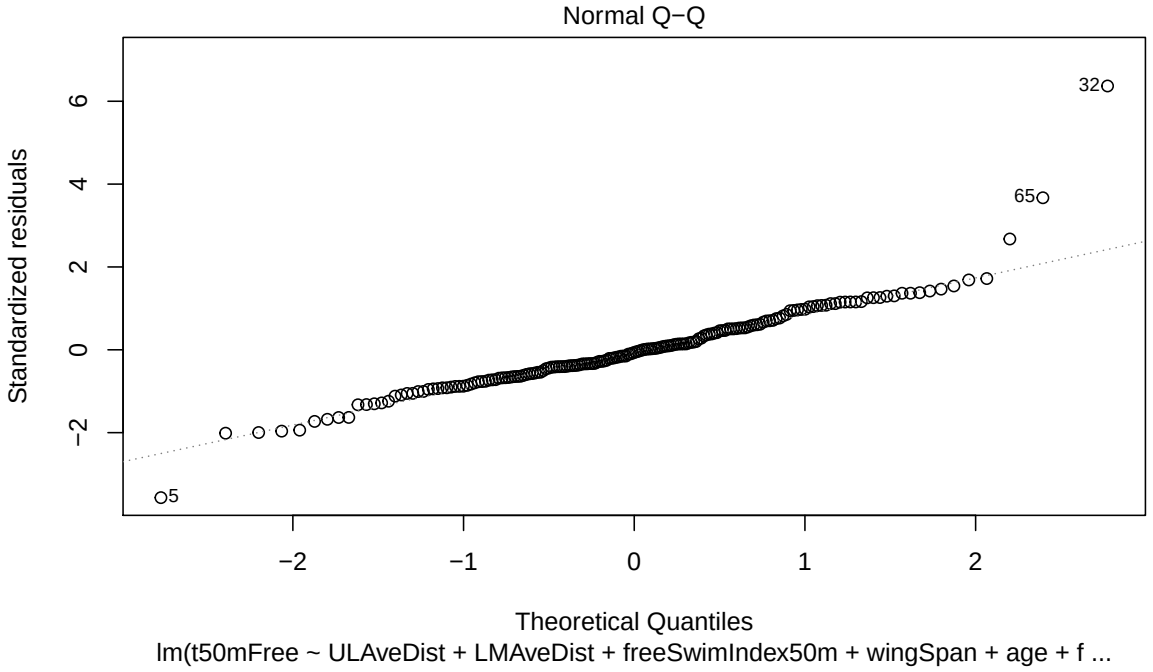
```
gvlma(myLModel,alphalevel = 0.05)
```

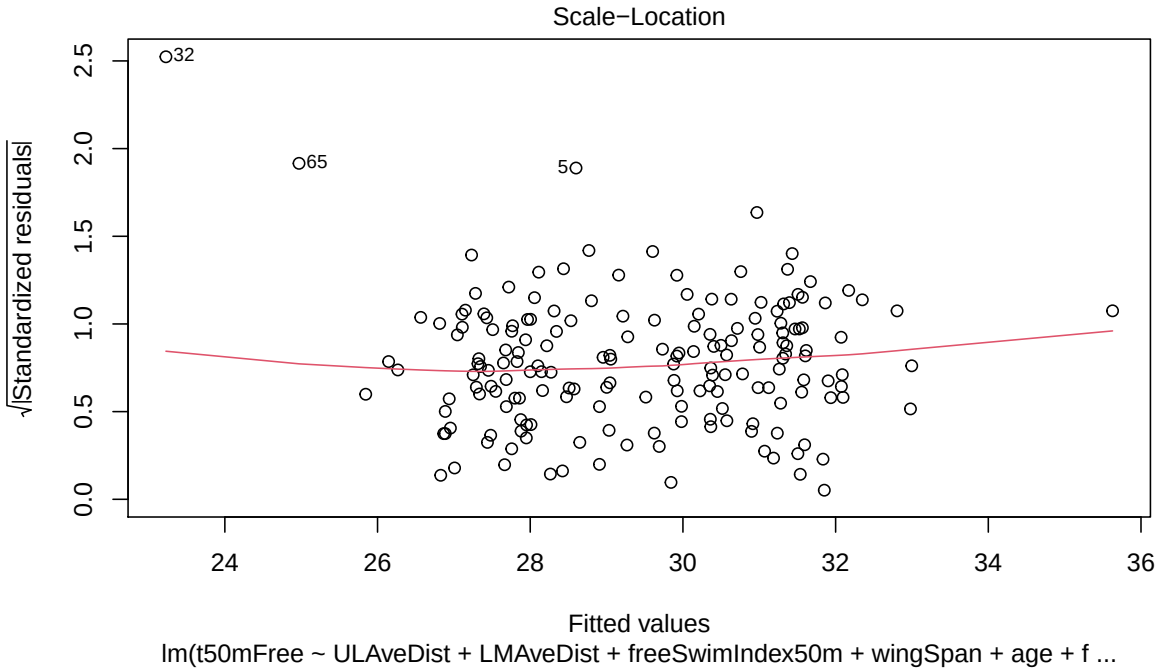
```
##
## Call:
## lm(formula = t50mFree ~ ULAveDist + LMAveDist + freeSwimIndex50m +
##     wingSpan + age + freeTurnTime5.10m_50m + height + weight,
##     data = df)
##
## Coefficients:
##           (Intercept)              ULAveDist              LMAveDist
##               17.5477                -0.1278                -1.0487
##      freeSwimIndex50m              wingSpan                    age
##               -0.1099                -1.3027                -0.1268
## freeTurnTime5.10m_50m                 height                 weight
##                2.0557                 0.6906                -0.0078
##
##
## 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        194.62 0.000e+00 Assumptions NOT satisfied!
## Skewness            17.67 2.623e-05 Assumptions NOT satisfied!
```

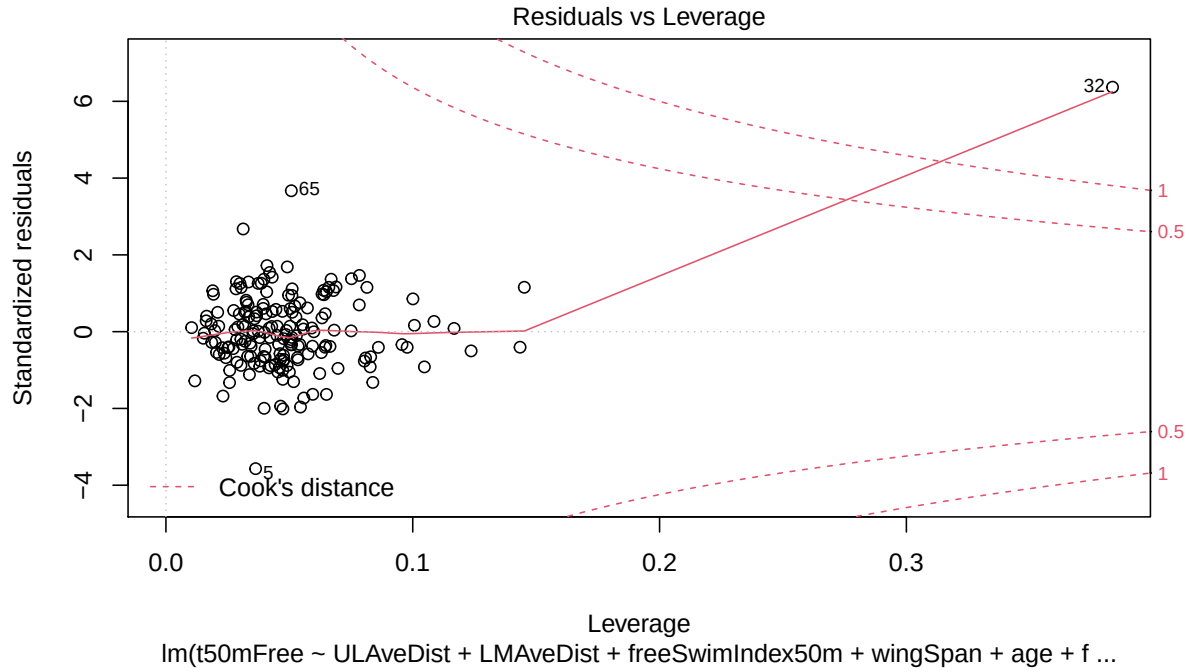
```
## Kurtosis            134.58 0.000e+00 Assumptions NOT satisfied!
## Link Function        38.00 7.067e-10 Assumptions NOT satisfied!
## Heteroscedasticity    4.36 3.678e-02 Assumptions NOT satisfied!
```

```
plot(myLModel)
```









Eliminar uns outliers subject 5,32,65

```
df<-df[-c(5,32,65), ]
```

```
myLModel <- lm(t50mFree ~ ULAveDist+LMAveDist+freeSwimIndex50m+wingSpan+age+freeTurnT
myLModel
```

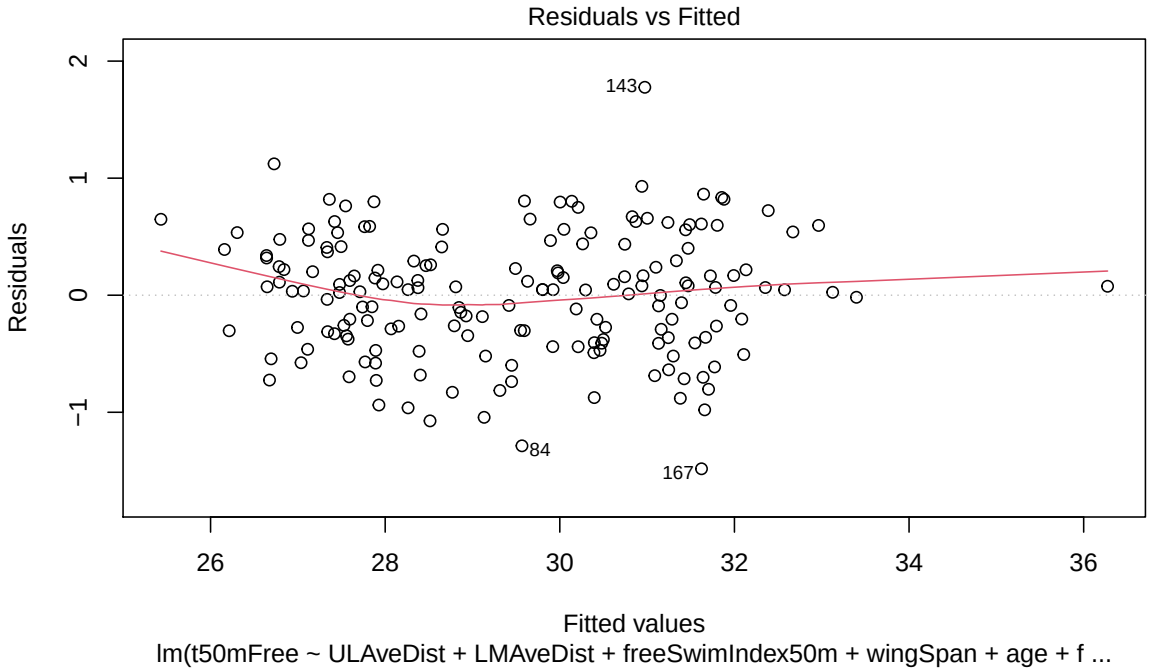
```
##
## Call:
```

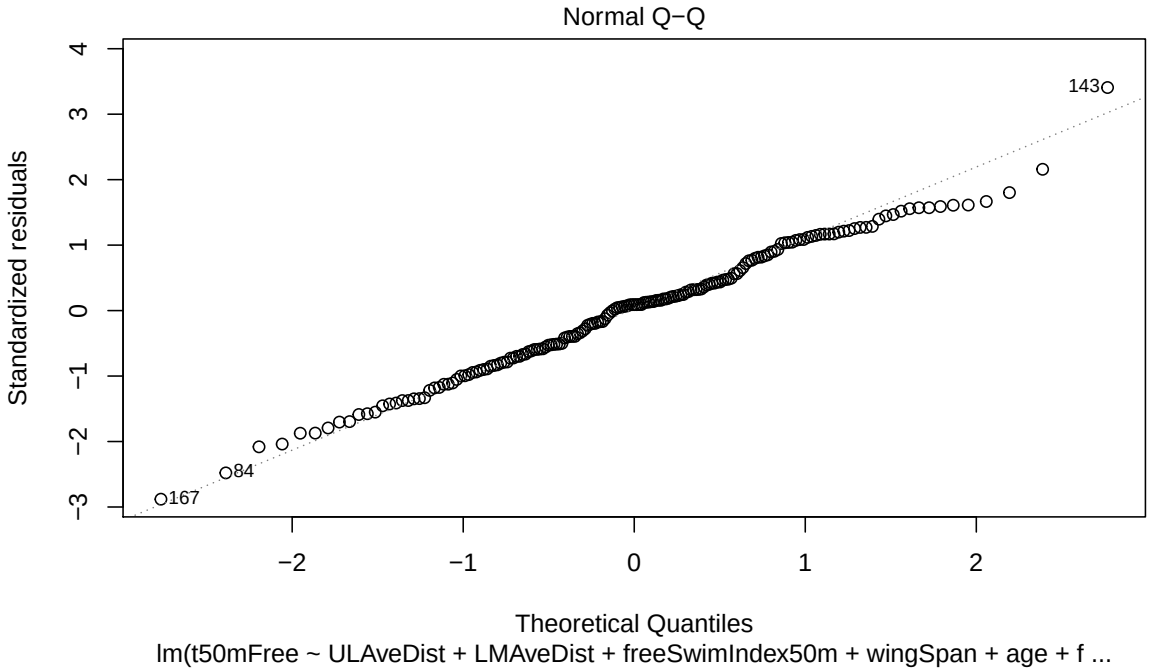
```
## lm(formula = t50mFree ~ ULAveDist + LMAveDist + freeSwimIndex50m +
##     wingSpan + age + freeTurnTime5.10m_50m + height + weight,
##     data = df)
##
## Coefficients:
##           (Intercept)              ULAveDist              LMAveDist
##             13.775888              -0.105987              -0.913388
##      freeSwimIndex50m               wingSpan                    age
##             -0.570190              -0.682804              -0.021497
## freeTurnTime5.10m_50m                 height                 weight
##              2.276746               0.729598              -0.001955
```

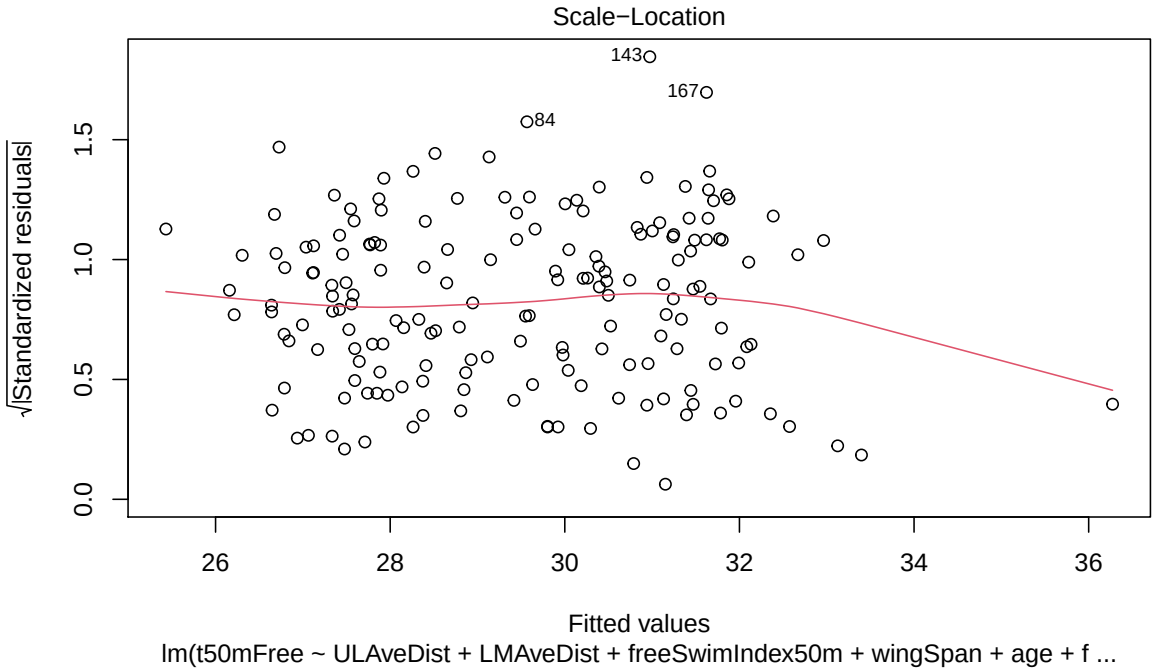
```
gvlma(myLModel,alphalevel = 0.05)
```

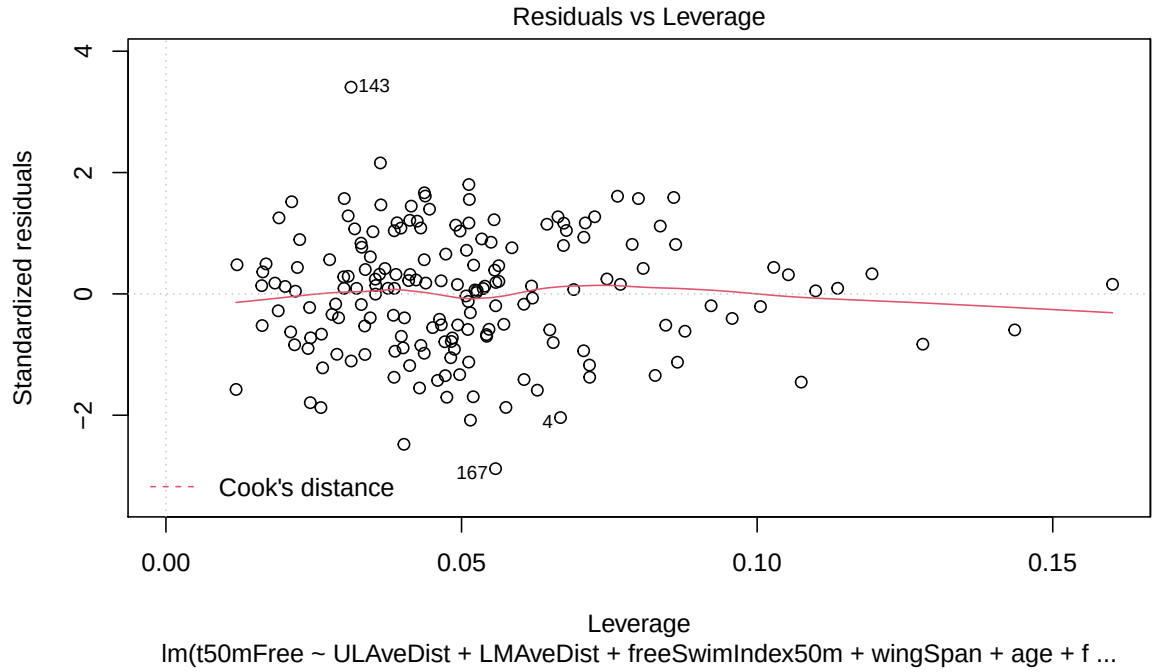
```
##
## Call:
## lm(formula = t50mFree ~ ULAveDist + LMAveDist + freeSwimIndex50m +
##     wingSpan + age + freeTurnTime5.10m_50m + height + weight,
##     data = df)
##
## Coefficients:
##           (Intercept)              ULAveDist              LMAveDist
##             13.775888              -0.105987              -0.913388
##      freeSwimIndex50m               wingSpan                    age
##             -0.570190              -0.682804              -0.021497
## freeTurnTime5.10m_50m                 height                 weight
##              2.276746               0.729598              -0.001955
##
##
## 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        5.20049  0.2673 Assumptions acceptable.
## Skewness           0.02290  0.8797 Assumptions acceptable.
## Kurtosis           0.06564  0.7978 Assumptions acceptable.
## Link Function      3.08320  0.0791 Assumptions acceptable.
## Heteroscedasticity 2.02875  0.1543 Assumptions acceptable.
```

```
plot(myLModel)
```


Residuals vs Fitted
Residuals
Fitted values
lm(t50mFree ~ ULAveDist + LMAveDist + freeSwimIndex50m + wingSpan + age + f ...
143
84
167


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(t50mFree ~ ULAveDist + LMAveDist + freeSwimIndex50m + wingSpan + age + f ...
143
84
167





1.6 Model summary

```
summary(myLModel)
```

```
##
## Call:
## lm(formula = t50mFree ~ ULAveDist + LMAveDist + freeSwimIndex50m +
##     wingSpan + age + freeTurnTime5.10m_50m + height + weight,
```

```
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48289 -0.35988  0.04731  0.39064  1.77631
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            13.775888   2.152323   6.400 1.49e-09 ***
## ULAveDist              -0.105987   0.105318  -1.006    0.316
## LMAveDist              -0.913388   0.213882  -4.271 3.26e-05 ***
## freeSwimIndex50m       -0.570190   0.110680  -5.152 7.17e-07 ***
## wingSpan               -0.682804   0.843234  -0.810    0.419
## age                    -0.021497   0.080274  -0.268    0.789
## freeTurnTime5.10m_50m   2.276746   0.123155  18.487  < 2e-16 ***
## height                  0.729598   1.119493   0.652    0.515
## weight                 -0.001955   0.010901  -0.179    0.858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5297 on 168 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.9317, Adjusted R-squared:  0.9284
## F-statistic: 286.5 on 8 and 168 DF,  p-value: < 2.2e-16
```

Intercept = expected t50mfree considering the average of all swimmers in the variables used

Slopes: - LMAveDist: para cada unidade a menos na LMAveDist o t50mFree reduz -0.908 - freeSwimIndex50m: para cada unidade a menos na freeSwimIndex50m o t50mFree reduz -0.55 - freeTurnTime5.10m_50m: para unidade a mais freeTurnTime5.10m_50m o t50mFree aumenta 2.29

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

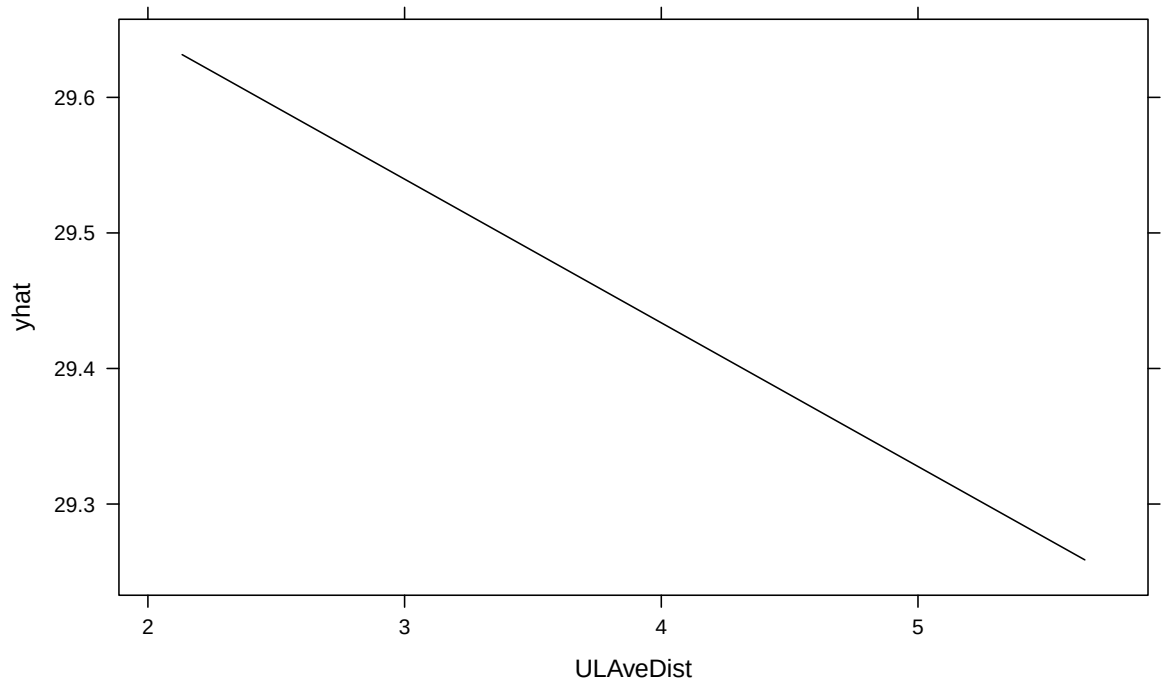
- r-squared:
  - R = coeficiente de correlação. Valores estimados vs valores observados (racio = VE/VO)
  - R Square = O quanto é que a variável dependente é explicada pelas variáveis utilizadas, mede a proporção da variação da variável dependente (t50mFree) que é explicada pelas variáveis independentes 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.

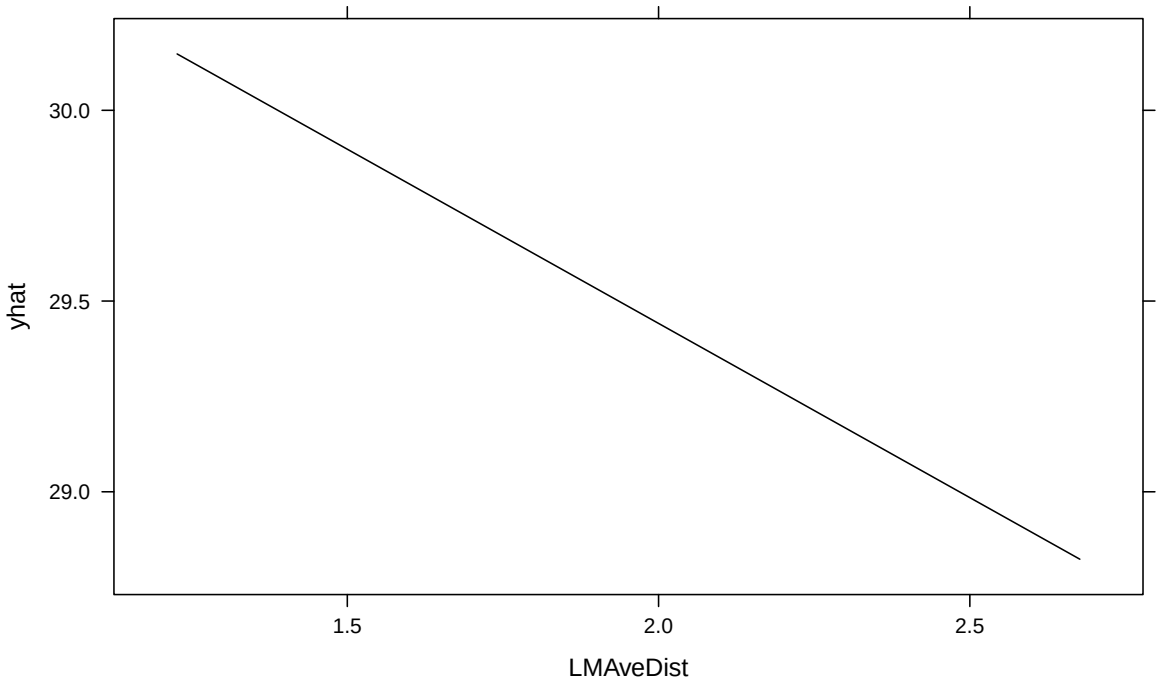
- adjusted r-squared:
  - Adjusted R Square = medida a reportar para avaliação da qualidade do modelo, está corrigida para o número de variáveis independentes e n da amostra
  - is the preferred measure as it adjusts for the number of variables considered.
  - In multiple regression settings, the R2 will always increase as more variables are included in the model.

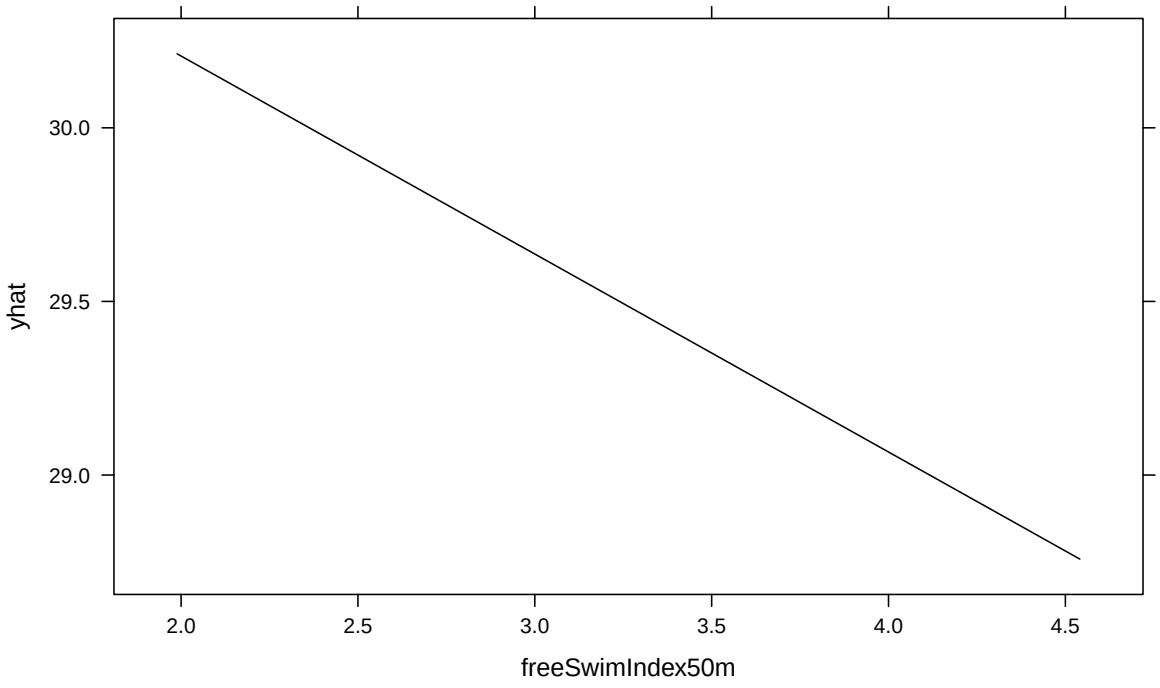
- f-statistics:
  - F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables
  - The further the F-statistic is from 1 the better it is
  - Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H0 : There is no relationship)

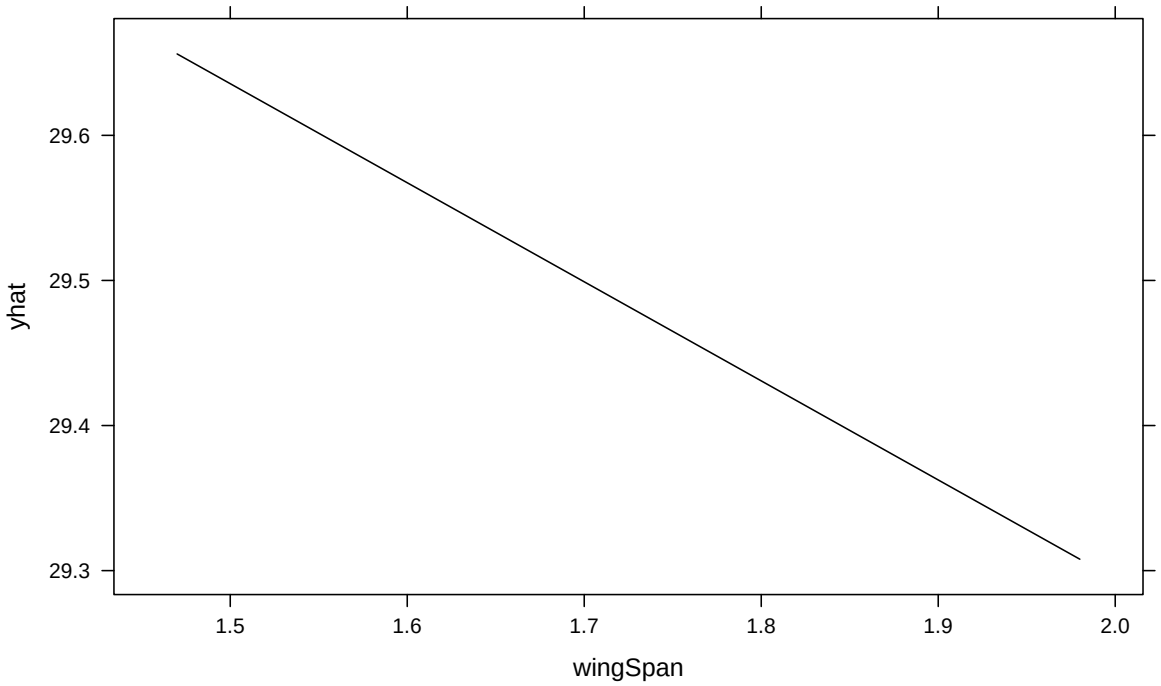
1.7 Partial Plots

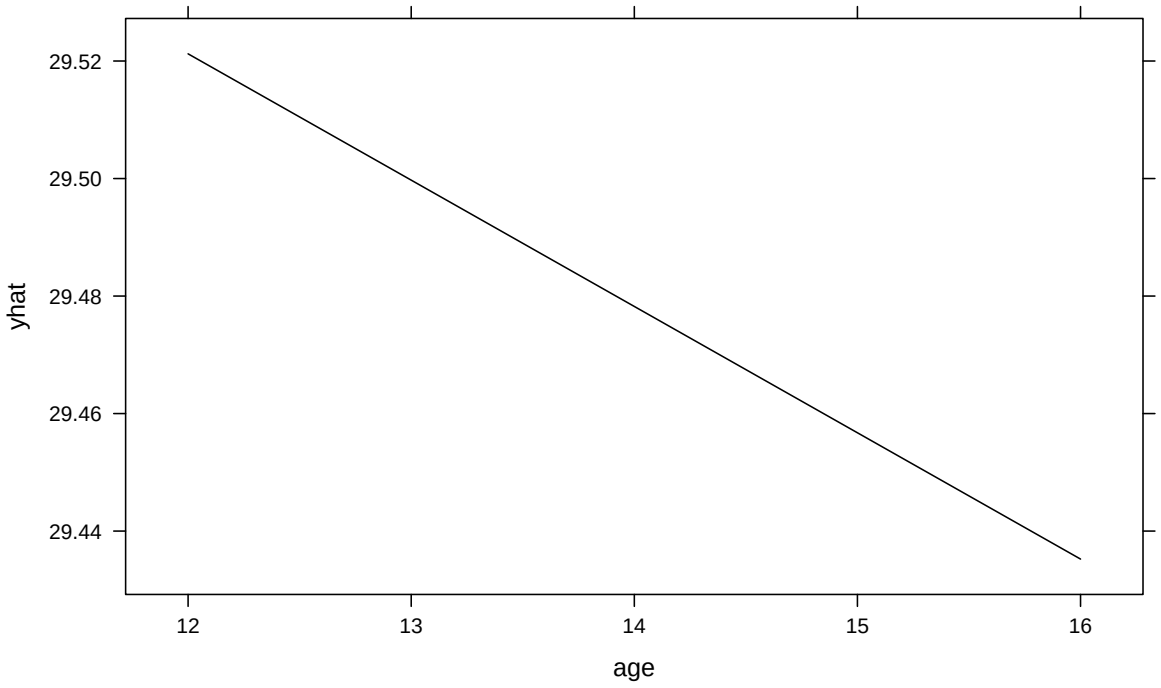
```
library(pdp)
vars=c("ULAveDist","LMAveDist","freeSwimIndex50m","wingSpan","age","freeTurnTime5.10m_
for (var in vars){
  print(partial(myLMModel,pred.var = var,plot = TRUE))
}
```

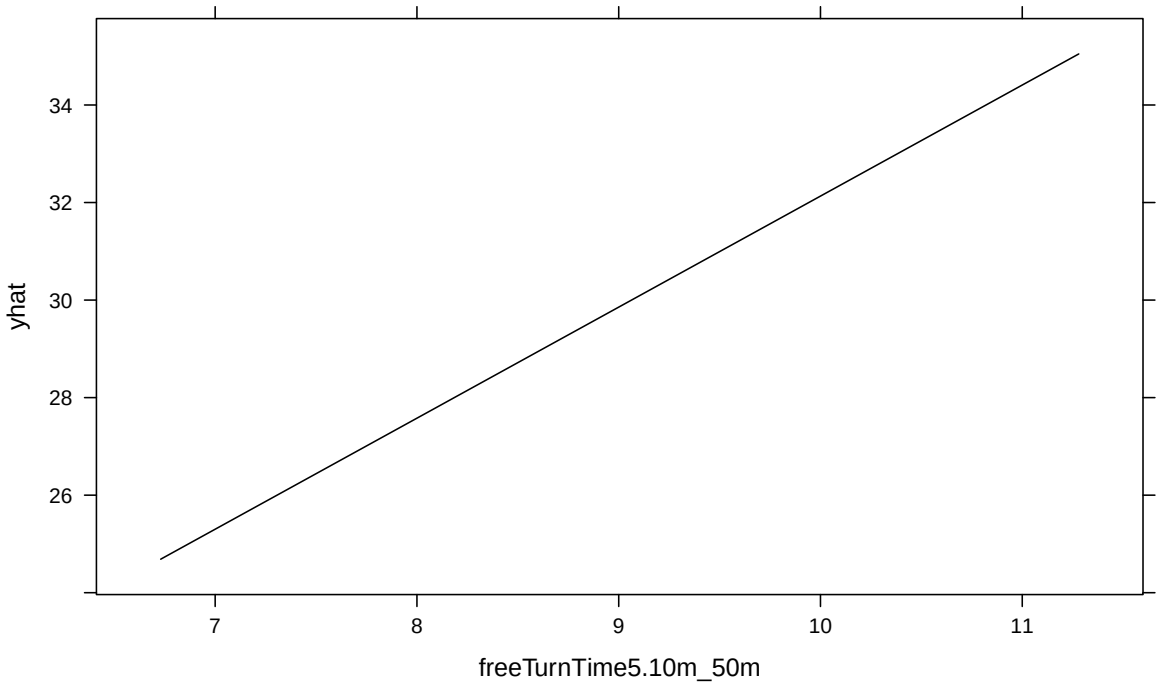


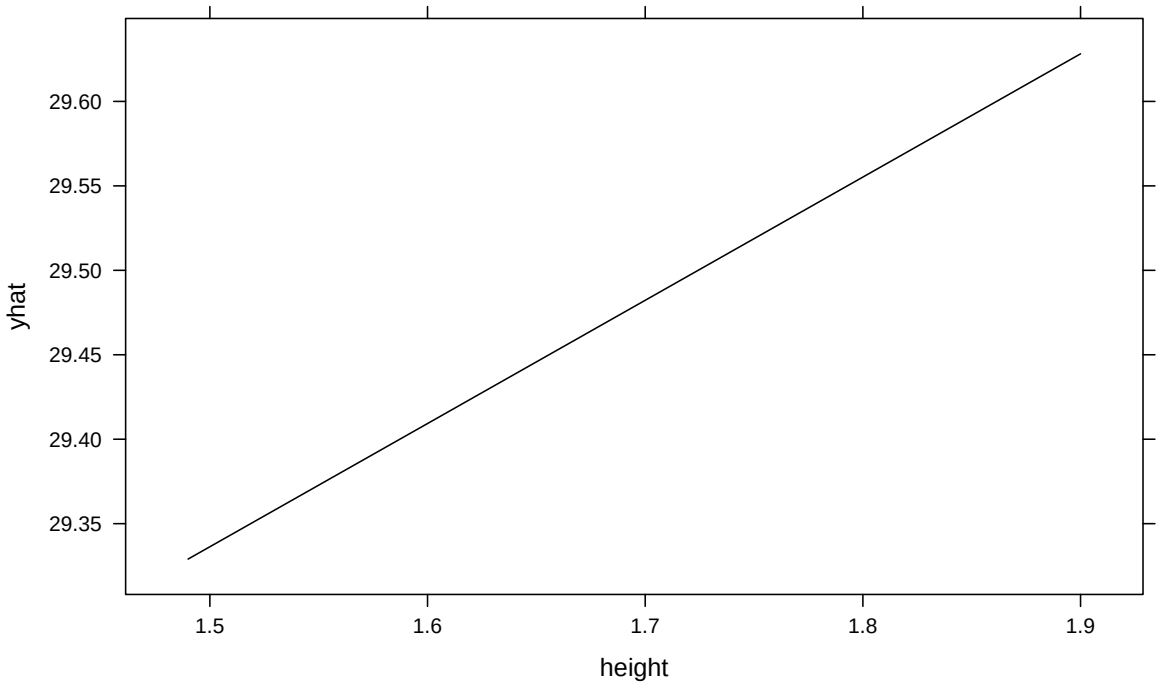

30.0
29.5
29.0
yhat
1.5
2.0
2.5
LMAveDist

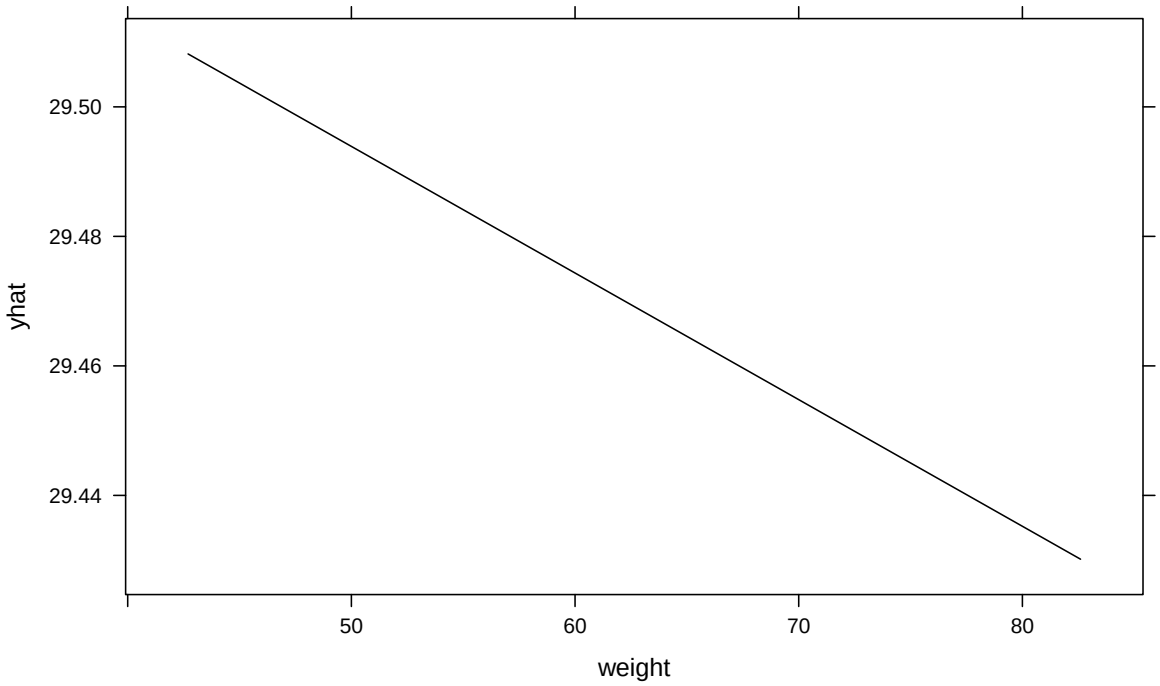

30.0
29.5
29.0
yhat
2.0
2.5
3.0
3.5
4.0
4.5
freeSwimIndex50m


yhat
29.6
29.5
29.4
29.3
1.5
1.6
1.7
1.8
1.9
2.0
wingSpan


yhat
29.52
29.50
29.48
29.46
29.44
12
13
14
15
16
age


yhat
34
32
30
28
26
7
8
9
10
11
freeTurnTime5.10m_50m


yhat
29.60
29.55
29.50
29.45
29.40
29.35
1.5
1.6
1.7
1.8
1.9
height




```
#partial(myLModel,pred.var = c("age"),plot = TRUE)
```

1.8 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)

# Vamos tirar NANs
df_semNANs<-na.omit(df)

# Fit do modelo com todas
dfTodas <- lm(t50mFree ~., data = df_semNANs)
# Stepwise regression model
stepModel <- stepAIC(dfTodas, direction = "both",  trace = FALSE,)
summary(stepModel)
```

```
##
## Call:
## lm(formula = t50mFree ~ height + weight + t50mFree5.20m + freeVelocity50m +
##     freeTurnIndex50m + LMAveDist, data = df_semNANs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12070 -0.22872 -0.00934  0.26407  1.06547
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -13.52233   11.39442  -1.187 0.236982
## height             1.00714    0.63280   1.592 0.113339
## weight            -0.01515    0.00677  -2.238 0.026512 *
## t50mFree5.20m      4.58127    0.61924   7.398 6.05e-12 ***
## freeVelocity50m    8.56516    3.38713   2.529 0.012358 *
## freeTurnIndex50m -12.29008    0.89418 -13.745  < 2e-16 ***
## LMAveDist         -0.48917    0.14600  -3.350 0.000994 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3664 on 170 degrees of freedom
## Multiple R-squared:  0.9669, Adjusted R-squared:  0.9658
## F-statistic: 828.4 on 6 and 170 DF,  p-value: < 2.2e-16
```

1.9 Qual é o melhor modelo considerando todas as variáveis disponíveis para prever t50mFree?

```
library(caret)
library(leaps)

models <- regsubsets(t50mFree~., data = df_semNANs, nvmax = 10, method = "seqrep")
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(t50mFree ~ ., data = df_semNANs, nvmax = 10,
```

```
##     method = "seqrep")
## 13 Variables  (and intercept)
##                       Forced in Forced out
## age                       FALSE      FALSE
## height                    FALSE      FALSE
## weight                    FALSE      FALSE
## wingSpan                  FALSE      FALSE
## t50mFree5.20m             FALSE      FALSE
## freeVelocity50m           FALSE      FALSE
## freeStrokeRate50m         FALSE      FALSE
## freeStrokeLength50m       FALSE      FALSE
## freeSwimIndex50m          FALSE      FALSE
## freeTurnTime5.10m_50m     FALSE      FALSE
## freeTurnIndex50m          FALSE      FALSE
## LMAveDist                 FALSE      FALSE
## ULAveDist                 FALSE      FALSE
## 1 subsets of each size up to 10
## Selection Algorithm: 'sequential replacement'
##           age height weight wingSpan t50mFree5.20m freeVelocity50m
## 1  ( 1 )  " " " "    " "    " "      "*"           " "
## 2  ( 1 )  " " " "    " "    " "      "*"           " "
## 3  ( 1 )  " " " "    " "    " "      "*"           " "
## 4  ( 1 )  " " " "    " "    " "      "*"           "*"
## 5  ( 1 )  " " " "    "*"    " "      "*"           "*"
## 6  ( 1 )  " " "*"    "*"    " "      "*"           "*"
## 7  ( 1 )  " " "*"    "*"    " "      "*"           "*"
## 8  ( 1 )  " " "*"    "*"    " "      "*"           "*"
## 9  ( 1 )  " " "*"    "*"    "*"      "*"           "*"
## 10  ( 1 ) " " "*"    "*"    "*"      "*"           "*"
##           freeStrokeRate50m freeStrokeLength50m freeSwimIndex50m
## 1  ( 1 )  " "               " "                 " "
## 2  ( 1 )  " "               " "                 " "
## 3  ( 1 )  " "               " "                 " "
## 4  ( 1 )  " "               " "                 " "
## 5  ( 1 )  " "               " "                 " "
## 6  ( 1 )  " "               " "                 " "
## 7  ( 1 )  " "               " "                 "*"
## 8  ( 1 )  " "               "*"                 "*"
## 9  ( 1 )  " "               "*"                 "*"
## 10  ( 1 ) "*"               "*"                 "*"
##           freeTurnTime5.10m_50m freeTurnIndex50m LMAveDist ULAveDist
## 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(t50mFree ~., data = df_semNANs,method = "leapBackward",
                   tuneGrid = data.frame(nvmax = 1:5),
                   trControl = train.control
                   )
stepModel$results
```

```
##   nvmax      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
## 1     1 0.5857796 0.9131446 0.4605571 0.11764088 0.03755425 0.09519304
## 2     2 0.3773555 0.9652974 0.3019877 0.11171716 0.01898411 0.08546053
## 3     3 0.3824550 0.9646816 0.3067812 0.09610711 0.01705362 0.07425697
## 4     4 0.3753466 0.9658798 0.3044413 0.09244116 0.01703067 0.07120074
## 5     5 0.3783242 0.9650260 0.3089391 0.08766592 0.01673086 0.07182098
```

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
## 4     4
```

1.10 Summary do melhor modelo

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

```
## Subset selection object
## 13 Variables  (and intercept)
##                       Forced in Forced out
## age                       FALSE      FALSE
## height                    FALSE      FALSE
## weight                    FALSE      FALSE
## wingSpan                  FALSE      FALSE
## t50mFree5.20m             FALSE      FALSE
## freeVelocity50m           FALSE      FALSE
## freeStrokeRate50m         FALSE      FALSE
## freeStrokeLength50m       FALSE      FALSE
## freeSwimIndex50m          FALSE      FALSE
## freeTurnTime5.10m_50m     FALSE      FALSE
## freeTurnIndex50m          FALSE      FALSE
## LMAveDist                 FALSE      FALSE
```

```
## ULAveDist                  FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: backward
##          age height weight wingSpan t50mFree5.20m freeVelocity50m
## 1  ( 1 ) " " " "    " "    " "      "*"           " "
## 2  ( 1 ) " " " "    " "    " "      "*"           " "
## 3  ( 1 ) " " " "    " "    " "      "*"           " "
## 4  ( 1 ) " " " "    " "    " "      "*"           "*"
##          freeStrokeRate50m freeStrokeLength50m freeSwimIndex50m
## 1  ( 1 ) " "               " "                 " "
## 2  ( 1 ) " "               " "                 " "
## 3  ( 1 ) " "               " "                 " "
## 4  ( 1 ) " "               " "                 " "
##          freeTurnTime5.10m_50m freeTurnIndex50m LMAveDist ULAveDist
## 1  ( 1 ) " "                   " "              " "       " "
## 2  ( 1 ) " "                   "*"              " "       " "
## 3  ( 1 ) " "                   "*"              "*"       " "
## 4  ( 1 ) " "                   "*"              "*"       " "
```

O melhor modelo contêm as variáveis t50mFree5.20m,freeStrokeLength50m, freeSwimIndex50m, freeTurnIndex50m e LMAveDist

```
myLModel <- lm(t50mFree ~ t50mFree5.20m+freeStrokeLength50m+freeSwimIndex50m+freeTurnI
summary(myLModel)
```

```
##
## Call:
## lm(formula = t50mFree ~ t50mFree5.20m + freeStrokeLength50m +
##     freeSwimIndex50m + freeTurnIndex50m + LMAveDist, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21787 -0.27536 -0.02307  0.25950  2.27510
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.3383     3.1237   1.389 0.166683
## t50mFree5.20m         3.9440     0.3140  12.560  < 2e-16 ***
## freeStrokeLength50m  -4.3070     1.4650  -2.940 0.003735 **
## freeSwimIndex50m      2.6923     0.8690   3.098 0.002275 **
## freeTurnIndex50m     -9.4790     0.8908 -10.641  < 2e-16 ***
## LMAveDist            -0.6249     0.1658  -3.770 0.000224 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.42 on 172 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.9563, Adjusted R-squared:  0.955
## F-statistic: 752.1 on 5 and 172 DF,  p-value: < 2.2e-16
```