Regression Model Selection

Josh Jarvey

06/21/2020

Knit a Word file from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word file.

Be advised, this homework will produce copious amounts of output.

## Exercise 1

Ninety members (ages 18.1 to 23.4 years) of three Division I women’s intercollegiate rowing teams (National Collegiate Athletic Association) within the Big Ten Conference volunteered to participate in a study to predict race time for female collegiate rowers from nineteen physical characteristics.

Data is in the file rowtime.rda in the DS705data package. The race times are in the variable named “racetime”.

### Part 1a

Load the data and use summary(rowtime) to see a numerical summary of the values of each.

1. What type of variable is the response variable racetime (categorical or quantitative)?
2. Does this indicate linear regression or logistic regression?
3. What types of variables are there in the pool of potential predictors? Categorical, quantitative, or a mixture of each?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

#load the data and review a summary of the variables.   
library(DS705data)  
data("rowtime")  
summary(rowtime)

## racetime tall weight armspan   
## Min. :406.0 Min. :155.6 Min. : 57.00 Min. :138.5   
## 1st Qu.:442.4 1st Qu.:169.3 1st Qu.: 69.90 1st Qu.:167.6   
## Median :460.0 Median :173.6 Median : 73.35 Median :172.7   
## Mean :458.5 Mean :172.9 Mean : 74.62 Mean :172.8   
## 3rd Qu.:475.3 3rd Qu.:176.3 3rd Qu.: 79.10 3rd Qu.:177.5   
## Max. :518.7 Max. :186.4 Max. :100.10 Max. :205.5   
## flexarm thighci calfcir tricep biceps   
## Min. :27.50 Min. :50.50 Min. :34.0 Min. :12.00 Min. : 4.00   
## 1st Qu.:30.50 1st Qu.:55.50 1st Qu.:37.0 1st Qu.:18.00 1st Qu.: 9.00   
## Median :31.90 Median :57.50 Median :38.0 Median :21.00 Median :12.00   
## Mean :31.60 Mean :57.89 Mean :38.4 Mean :21.09 Mean :12.46   
## 3rd Qu.:32.88 3rd Qu.:59.88 3rd Qu.:39.5 3rd Qu.:24.00 3rd Qu.:15.00   
## Max. :39.00 Max. :71.50 Max. :44.0 Max. :34.00 Max. :28.00   
## thigh estffm estfm bestsnr   
## Min. :13.00 Min. :44.86 Min. : 9.918 Min. :17.00   
## 1st Qu.:24.00 1st Qu.:54.52 1st Qu.:14.274 1st Qu.:34.00   
## Median :28.00 Median :56.97 Median :16.929 Median :40.00   
## Mean :28.73 Mean :57.83 Mean :16.786 Mean :38.72   
## 3rd Qu.:34.00 3rd Qu.:61.15 3rd Qu.:19.073 3rd Qu.:44.00   
## Max. :50.00 Max. :78.15 Max. :26.726 Max. :50.00   
## bestvj legpower endo meso   
## Min. : 8.00 Min. : 60.19 Min. :3.219 Min. :1.564   
## 1st Qu.:13.00 1st Qu.: 92.19 1st Qu.:4.768 1st Qu.:3.257   
## Median :15.00 Median : 99.84 Median :5.789 Median :4.033   
## Mean :14.82 Mean :100.33 Mean :5.696 Mean :4.042   
## 3rd Qu.:16.00 3rd Qu.:108.92 3rd Qu.:6.499 3rd Qu.:4.738   
## Max. :21.00 Max. :153.98 Max. :8.260 Max. :6.812   
## ecto expvarsity preexper   
## Min. :0.100 Min. :0.0000 Min. :0.0000   
## 1st Qu.:1.004 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :1.596 Median :1.0000 Median :0.0000   
## Mean :1.655 Mean :0.5667 Mean :0.3333   
## 3rd Qu.:2.218 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :3.887 Max. :1.0000 Max. :1.0000

The response variable racetime is a quantitative variable because it is numerical and is continuous. Because this is a quantitative type, we will use a linear regression. As for the predictor variables, there is a mix of quantitative (tall, weight, armspan ,etc) as well as categorical: expvarsity (indicating if they have varsity experience?) and preexper (indicating if they have previous experience?).

### Part 1b

Use the **regsubsets** function to find the “best” first-order model for predicting the response variable racetime with up to 8 of the 19 predictor variables in the data set. Produce the summary and the plot for the best single models with up to 8 predictors according to .

Which independent variables are in the best first-order model with 8 predictors when the is the criterion for selection?

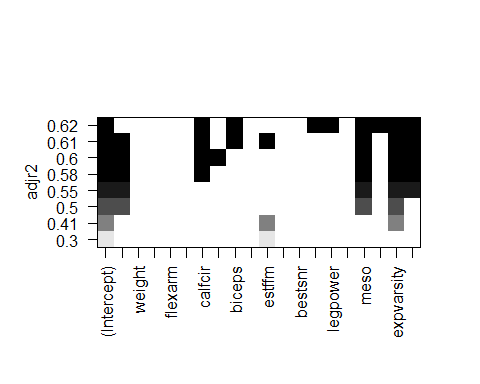
What is the for the best first-order model?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

#import the regsubsets() function from the leaps library  
library(leaps)  
 #using the "full" number of variables denoted by y~., step forward until you've determine the best model with 8 predictors.  
 #nvmax is the number of variables to step toward. Obviously there may be increases or decreases to adjusted r2, so it might be best to allow it to go all the way to the max number of variables within the dataset and then just look at which one gives the best adjusted r2 (or whatever metric were interested in as the grading criteria).  
allmodels=regsubsets(racetime~.,data = rowtime, nvmax = 8)  
 #because "all" 8 of the models are stored, run summary to see the output.   
summary(allmodels)

## Subset selection object  
## Call: regsubsets.formula(racetime ~ ., data = rowtime, nvmax = 8)  
## 19 Variables (and intercept)  
## Forced in Forced out  
## tall FALSE FALSE  
## weight FALSE FALSE  
## armspan FALSE FALSE  
## flexarm FALSE FALSE  
## thighci FALSE FALSE  
## calfcir FALSE FALSE  
## tricep FALSE FALSE  
## biceps FALSE FALSE  
## thigh FALSE FALSE  
## estffm FALSE FALSE  
## estfm FALSE FALSE  
## bestsnr FALSE FALSE  
## bestvj FALSE FALSE  
## legpower FALSE FALSE  
## endo FALSE FALSE  
## meso FALSE FALSE  
## ecto FALSE FALSE  
## expvarsity FALSE FALSE  
## preexper FALSE FALSE  
## 1 subsets of each size up to 8  
## Selection Algorithm: exhaustive  
## tall weight armspan flexarm thighci calfcir tricep biceps thigh estffm  
## 1 ( 1 ) " " " " " " " " " " " " " " " " " " "\*"   
## 2 ( 1 ) " " " " " " " " " " " " " " " " " " "\*"   
## 3 ( 1 ) "\*" " " " " " " " " " " " " " " " " " "   
## 4 ( 1 ) "\*" " " " " " " " " " " " " " " " " " "   
## 5 ( 1 ) "\*" " " " " " " " " "\*" " " " " " " " "   
## 6 ( 1 ) "\*" " " " " " " " " "\*" "\*" " " " " " "   
## 7 ( 1 ) "\*" " " " " " " " " "\*" " " "\*" " " "\*"   
## 8 ( 1 ) " " " " " " " " " " "\*" " " "\*" " " " "   
## estfm bestsnr bestvj legpower endo meso ecto expvarsity preexper  
## 1 ( 1 ) " " " " " " " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " " " " "\*" " "   
## 3 ( 1 ) " " " " " " " " " " "\*" " " "\*" " "   
## 4 ( 1 ) " " " " " " " " " " "\*" " " "\*" "\*"   
## 5 ( 1 ) " " " " " " " " " " "\*" " " "\*" "\*"   
## 6 ( 1 ) " " " " " " " " " " "\*" " " "\*" "\*"   
## 7 ( 1 ) " " " " " " " " " " "\*" " " "\*" "\*"   
## 8 ( 1 ) " " " " "\*" "\*" " " "\*" "\*" "\*" "\*"

#the summary is ok, but this plot is where you can really see what model gives the best adjusted r2 measure.   
plot(allmodels, scale = "adjr2")



Based on the output of the regsubsets() function, I will choose a model with 6 independent variables: tall, calfcir, tricep, meso, expvarsity, preexper - with an adjusted R2 score of 0.60. I choose this model because in comparing the models with 7 or 8 variables, the adjusted R2 score only increased by 0.10 for each additional variable. This didn’t seem like a fair trade off (add more complexity for minor gains in explainability), so i will settle with a slightly lesser adjusted R2 score for a simpler model.

### Part 1c

Use the **step** function with backward selection to find the “best” first-order model for predicting the response variable racetime. Recall that the formula structure y~. will produce the model using y as the response variable and all other variables in the data set as the predictors; in this set “racetime” is the response (not y) and all other variables are potential predictors.

Which independent variables are in this model? What is the AIC value for this model?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

#start with a "full" model denoted by y~.  
 #then do a backward step to remove variables from the equation until a best is reached.   
step(lm(racetime~., data=rowtime), direction = "backward")

## Start: AIC=512.17  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## tricep + biceps + thigh + estffm + estfm + bestsnr + bestvj +   
## legpower + endo + meso + ecto + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - bestsnr 1 1.1 17090 510.18  
## - thigh 1 1.3 17090 510.18  
## - endo 1 4.4 17094 510.20  
## - tricep 1 10.7 17100 510.23  
## - ecto 1 65.4 17155 510.52  
## - weight 1 76.6 17166 510.58  
## - estffm 1 77.2 17166 510.58  
## - estfm 1 83.8 17173 510.62  
## - flexarm 1 155.8 17245 510.99  
## - armspan 1 227.7 17317 511.37  
## - thighci 1 286.5 17376 511.67  
## <none> 17089 512.17  
## - legpower 1 452.9 17542 512.53  
## - bestvj 1 569.5 17659 513.13  
## - biceps 1 572.4 17662 513.14  
## - tall 1 807.5 17897 514.33  
## - calfcir 1 1453.5 18543 517.52  
## - preexper 1 1748.1 18837 518.94  
## - meso 1 2299.2 19388 521.54  
## - expvarsity 1 5176.8 22266 533.99  
##   
## Step: AIC=510.18  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## tricep + biceps + thigh + estffm + estfm + bestvj + legpower +   
## endo + meso + ecto + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - thigh 1 1.9 17092 508.19  
## - endo 1 3.7 17094 508.20  
## - tricep 1 9.6 17100 508.23  
## - ecto 1 67.8 17158 508.54  
## - weight 1 84.6 17175 508.62  
## - estffm 1 85.6 17176 508.63  
## - estfm 1 92.3 17183 508.67  
## - flexarm 1 159.8 17250 509.02  
## - armspan 1 234.9 17325 509.41  
## - thighci 1 285.5 17376 509.67  
## <none> 17090 510.18  
## - legpower 1 516.3 17607 510.86  
## - biceps 1 572.9 17663 511.15  
## - bestvj 1 662.2 17752 511.60  
## - tall 1 806.7 17897 512.33  
## - calfcir 1 1463.4 18554 515.57  
## - preexper 1 1766.9 18857 517.04  
## - meso 1 2303.2 19393 519.56  
## - expvarsity 1 5451.5 22542 533.10  
##   
## Step: AIC=508.19  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## tricep + biceps + estffm + estfm + bestvj + legpower + endo +   
## meso + ecto + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - endo 1 8.0 17100 506.23  
## - tricep 1 11.9 17104 506.25  
## - ecto 1 66.0 17158 506.54  
## - weight 1 87.8 17180 506.65  
## - estffm 1 88.6 17181 506.66  
## - estfm 1 96.5 17189 506.70  
## - flexarm 1 161.7 17254 507.04  
## - armspan 1 239.4 17332 507.44  
## - thighci 1 290.6 17383 507.71  
## <none> 17092 508.19  
## - legpower 1 514.7 17607 508.86  
## - biceps 1 572.7 17665 509.16  
## - bestvj 1 662.2 17754 509.61  
## - tall 1 834.4 17927 510.48  
## - calfcir 1 1482.3 18574 513.68  
## - preexper 1 1770.3 18863 515.06  
## - meso 1 2302.7 19395 517.57  
## - expvarsity 1 5464.4 22557 531.16  
##   
## Step: AIC=506.23  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## tricep + biceps + estffm + estfm + bestvj + legpower + meso +   
## ecto + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - tricep 1 6.4 17107 504.27  
## - ecto 1 61.9 17162 504.56  
## - weight 1 83.0 17183 504.67  
## - estffm 1 83.8 17184 504.67  
## - estfm 1 91.3 17191 504.71  
## - flexarm 1 181.0 17281 505.18  
## - armspan 1 235.7 17336 505.46  
## - thighci 1 282.8 17383 505.71  
## <none> 17100 506.23  
## - legpower 1 541.1 17641 507.04  
## - biceps 1 565.6 17666 507.16  
## - bestvj 1 695.8 17796 507.82  
## - tall 1 826.4 17927 508.48  
## - calfcir 1 1567.9 18668 512.13  
## - preexper 1 1763.4 18863 513.07  
## - meso 1 2295.3 19395 515.57  
## - expvarsity 1 5456.6 22557 529.16  
##   
## Step: AIC=504.27  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## biceps + estffm + estfm + bestvj + legpower + meso + ecto +   
## expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - ecto 1 72.5 17179 502.65  
## - weight 1 93.2 17200 502.76  
## - estffm 1 94.0 17200 502.76  
## - estfm 1 102.8 17209 502.81  
## - flexarm 1 220.2 17327 503.42  
## - armspan 1 242.2 17349 503.53  
## - thighci 1 279.9 17386 503.73  
## <none> 17107 504.27  
## - legpower 1 623.1 17730 505.49  
## - biceps 1 642.6 17749 505.59  
## - bestvj 1 782.5 17889 506.29  
## - tall 1 840.6 17947 506.58  
## - calfcir 1 1591.5 18698 510.27  
## - preexper 1 1758.8 18865 511.07  
## - meso 1 2290.4 19397 513.58  
## - expvarsity 1 5946.4 23053 529.12  
##   
## Step: AIC=502.65  
## racetime ~ tall + weight + armspan + flexarm + thighci + calfcir +   
## biceps + estffm + estfm + bestvj + legpower + meso + expvarsity +   
## preexper  
##   
## Df Sum of Sq RSS AIC  
## - weight 1 78.8 17258 501.06  
## - estffm 1 79.9 17259 501.06  
## - estfm 1 89.1 17268 501.11  
## - armspan 1 254.4 17433 501.97  
## - flexarm 1 265.2 17444 502.03  
## - thighci 1 304.3 17483 502.23  
## <none> 17179 502.65  
## - legpower 1 553.8 17733 503.50  
## - biceps 1 593.2 17772 503.70  
## - bestvj 1 715.5 17895 504.32  
## - calfcir 1 1644.1 18823 508.87  
## - preexper 1 1700.8 18880 509.14  
## - meso 1 2228.7 19408 511.62  
## - tall 1 3578.6 20758 517.68  
## - expvarsity 1 6028.6 23208 527.72  
##   
## Step: AIC=501.06  
## racetime ~ tall + armspan + flexarm + thighci + calfcir + biceps +   
## estffm + estfm + bestvj + legpower + meso + expvarsity +   
## preexper  
##   
## Df Sum of Sq RSS AIC  
## - estffm 1 9.8 17268 499.11  
## - armspan 1 238.2 17496 500.29  
## - flexarm 1 268.3 17526 500.45  
## - thighci 1 289.1 17547 500.55  
## <none> 17258 501.06  
## - legpower 1 499.5 17757 501.63  
## - estfm 1 548.7 17807 501.88  
## - biceps 1 586.3 17844 502.07  
## - bestvj 1 661.4 17919 502.44  
## - calfcir 1 1707.4 18965 507.55  
## - preexper 1 1889.5 19147 508.41  
## - meso 1 2159.5 19417 509.67  
## - tall 1 3502.6 20760 515.69  
## - expvarsity 1 5959.7 23217 525.76  
##   
## Step: AIC=499.11  
## racetime ~ tall + armspan + flexarm + thighci + calfcir + biceps +   
## estfm + bestvj + legpower + meso + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - armspan 1 248.4 17516 498.39  
## - thighci 1 286.5 17554 498.59  
## - flexarm 1 313.1 17581 498.73  
## <none> 17268 499.11  
## - biceps 1 599.0 17867 500.18  
## - estfm 1 831.1 18099 501.34  
## - legpower 1 1570.5 18838 504.94  
## - calfcir 1 1817.1 19085 506.11  
## - preexper 1 1889.0 19157 506.45  
## - bestvj 1 2062.9 19331 507.27  
## - meso 1 2150.4 19418 507.67  
## - tall 1 3592.7 20860 514.12  
## - expvarsity 1 6282.6 23550 525.04  
##   
## Step: AIC=498.39  
## racetime ~ tall + flexarm + thighci + calfcir + biceps + estfm +   
## bestvj + legpower + meso + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - flexarm 1 261.9 17778 497.73  
## <none> 17516 498.39  
## - thighci 1 400.0 17916 498.43  
## - biceps 1 721.4 18237 500.03  
## - estfm 1 723.8 18240 500.04  
## - legpower 1 1358.5 18874 503.12  
## - calfcir 1 1711.8 19228 504.79  
## - bestvj 1 1831.8 19348 505.35  
## - preexper 1 1966.2 19482 505.97  
## - meso 1 2104.6 19621 506.61  
## - tall 1 3345.4 20861 512.13  
## - expvarsity 1 6482.8 23999 524.73  
##   
## Step: AIC=497.73  
## racetime ~ tall + thighci + calfcir + biceps + estfm + bestvj +   
## legpower + meso + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## - thighci 1 297.6 18075 497.22  
## <none> 17778 497.73  
## - estfm 1 656.5 18434 498.99  
## - biceps 1 876.1 18654 500.06  
## - legpower 1 1115.6 18893 501.21  
## - calfcir 1 1450.1 19228 502.79  
## - bestvj 1 1616.2 19394 503.56  
## - meso 1 1894.0 19672 504.84  
## - preexper 1 2036.0 19814 505.49  
## - tall 1 3099.5 20877 510.19  
## - expvarsity 1 6567.6 24345 524.03  
##   
## Step: AIC=497.22  
## racetime ~ tall + calfcir + biceps + estfm + bestvj + legpower +   
## meso + expvarsity + preexper  
##   
## Df Sum of Sq RSS AIC  
## <none> 18075 497.22  
## - estfm 1 433.1 18508 497.36  
## - biceps 1 868.7 18944 499.45  
## - calfcir 1 1260.9 19336 501.29  
## - legpower 1 1421.3 19497 502.04  
## - meso 1 1885.0 19960 504.15  
## - bestvj 1 1887.1 19962 504.16  
## - preexper 1 2168.0 20243 505.42  
## - tall 1 2926.1 21001 508.73  
## - expvarsity 1 6381.6 24457 522.44

##   
## Call:  
## lm(formula = racetime ~ tall + calfcir + biceps + estfm + bestvj +   
## legpower + meso + expvarsity + preexper, data = rowtime)  
##   
## Coefficients:  
## (Intercept) tall calfcir biceps estfm bestvj   
## 797.881 -2.284 2.764 1.118 1.665 5.101   
## legpower meso expvarsity preexper   
## -1.140 -9.786 -18.370 -11.096

The best model selected using a backward step algorithm has an AIC of 497.22, and includes the variables: tall, calfcir, biceps, estffm, bestvj, legpower, meso, expvarsity, & preexper.

### Part 1d

Use the **step** function with forward selection to find the “best” model for predicting the response variable racetime. Recall that the formula structure y~1 will produce the model using y as the response variable and no variables in the data set as the predictors, only an intercept.

Which independent variables are in the model selected? What is the AIC value for this model?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

#to begin the "forward" step function, we set a "null" model which is the y~1  
nullmodel = lm(racetime~1, data = rowtime)  
 #we set a "full" model to be the "cap" for the forward step function y~.  
fullmodel = lm(racetime~., data = rowtime)  
 #starting with the null model, add variables one at a time and test their impact, within the bounds of the lower and upper limits.  
step(nullmodel,scope = list(lower=nullmodel, upper=fullmodel), direction = "forward")

## Start: AIC=575.99  
## racetime ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + estffm 1 16415.9 36552 544.60  
## + tall 1 13189.7 39778 552.21  
## + weight 1 12987.5 39980 552.67  
## + legpower 1 8478.5 44489 562.29  
## + expvarsity 1 7731.5 45236 563.79  
## + flexarm 1 7190.1 45777 564.86  
## + preexper 1 5346.5 47621 568.41  
## + thighci 1 4806.2 48161 569.43  
## + estfm 1 4288.2 48679 570.39  
## + armspan 1 3492.1 49476 571.85  
## + calfcir 1 2072.4 50895 574.39  
## <none> 52968 575.99  
## + meso 1 932.8 52035 576.39  
## + bestvj 1 203.9 52764 577.64  
## + ecto 1 110.7 52857 577.80  
## + thigh 1 102.8 52865 577.81  
## + biceps 1 76.4 52891 577.86  
## + bestsnr 1 49.6 52918 577.90  
## + tricep 1 42.6 52925 577.91  
## + endo 1 12.6 52955 577.97  
##   
## Step: AIC=544.6  
## racetime ~ estffm  
##   
## Df Sum of Sq RSS AIC  
## + expvarsity 1 5950.0 30602 530.61  
## + biceps 1 4411.1 32141 535.03  
## + preexper 1 3700.7 32851 536.99  
## + ecto 1 3276.0 33276 538.15  
## + tall 1 3072.6 33479 538.70  
## + endo 1 2677.3 33874 539.75  
## + calfcir 1 2326.6 34225 540.68  
## + tricep 1 2238.5 34313 540.91  
## + estfm 1 1274.1 35278 543.41  
## + weight 1 1270.1 35282 543.42  
## <none> 36552 544.60  
## + thighci 1 762.4 35789 544.70  
## + meso 1 705.0 35847 544.85  
## + flexarm 1 690.1 35862 544.89  
## + thigh 1 433.5 36118 545.53  
## + legpower 1 311.0 36241 545.83  
## + bestvj 1 53.8 36498 546.47  
## + armspan 1 53.4 36498 546.47  
## + bestsnr 1 3.1 36549 546.59  
##   
## Step: AIC=530.61  
## racetime ~ estffm + expvarsity  
##   
## Df Sum of Sq RSS AIC  
## + tall 1 3673.9 26928 521.10  
## + preexper 1 3345.3 27256 522.19  
## + calfcir 1 3235.8 27366 522.55  
## + ecto 1 3153.8 27448 522.82  
## + biceps 1 3060.2 27542 523.13  
## + legpower 1 1245.3 29356 528.87  
## + endo 1 1169.4 29432 529.10  
## + bestvj 1 1086.4 29515 529.36  
## + meso 1 992.0 29610 529.64  
## + tricep 1 681.0 29921 530.58  
## <none> 30602 530.61  
## + estfm 1 521.4 30080 531.06  
## + weight 1 516.1 30086 531.08  
## + flexarm 1 419.8 30182 531.37  
## + thighci 1 311.9 30290 531.69  
## + armspan 1 292.3 30309 531.75  
## + thigh 1 74.2 30528 532.39  
## + bestsnr 1 0.7 30601 532.61  
##   
## Step: AIC=521.1  
## racetime ~ estffm + expvarsity + tall  
##   
## Df Sum of Sq RSS AIC  
## + preexper 1 3687.9 23240 509.84  
## + bestvj 1 1256.1 25672 518.80  
## + meso 1 1247.8 25680 518.83  
## + calfcir 1 1212.2 25716 518.95  
## + legpower 1 1144.1 25784 519.19  
## + biceps 1 1141.2 25787 519.20  
## <none> 26928 521.10  
## + tricep 1 276.8 26651 522.17  
## + armspan 1 239.2 26689 522.30  
## + endo 1 198.0 26730 522.44  
## + estfm 1 123.0 26805 522.69  
## + weight 1 119.8 26808 522.70  
## + bestsnr 1 42.8 26885 522.96  
## + ecto 1 28.0 26900 523.01  
## + flexarm 1 19.4 26908 523.03  
## + thigh 1 8.9 26919 523.07  
## + thighci 1 7.7 26920 523.07  
##   
## Step: AIC=509.84  
## racetime ~ estffm + expvarsity + tall + preexper  
##   
## Df Sum of Sq RSS AIC  
## + biceps 1 1637.17 21603 505.27  
## + meso 1 827.19 22413 508.58  
## + calfcir 1 810.18 22430 508.65  
## + tricep 1 582.46 22657 509.56  
## <none> 23240 509.84  
## + legpower 1 479.60 22760 509.97  
## + bestvj 1 437.89 22802 510.13  
## + estfm 1 383.16 22857 510.35  
## + weight 1 382.67 22857 510.35  
## + endo 1 343.89 22896 510.50  
## + ecto 1 200.92 23039 511.06  
## + thigh 1 183.37 23057 511.13  
## + armspan 1 167.39 23073 511.19  
## + bestsnr 1 27.65 23212 511.74  
## + thighci 1 14.47 23225 511.79  
## + flexarm 1 0.53 23239 511.84  
##   
## Step: AIC=505.27  
## racetime ~ estffm + expvarsity + tall + preexper + biceps  
##   
## Df Sum of Sq RSS AIC  
## + meso 1 951.94 20651 503.21  
## + bestvj 1 523.92 21079 505.06  
## + calfcir 1 519.62 21083 505.08  
## <none> 21603 505.27  
## + legpower 1 263.49 21339 506.16  
## + armspan 1 153.44 21449 506.63  
## + thighci 1 153.08 21450 506.63  
## + flexarm 1 96.62 21506 506.87  
## + bestsnr 1 36.23 21567 507.12  
## + endo 1 32.94 21570 507.13  
## + weight 1 32.50 21570 507.13  
## + estfm 1 32.12 21571 507.13  
## + thigh 1 25.43 21577 507.16  
## + tricep 1 7.76 21595 507.24  
## + ecto 1 1.60 21601 507.26  
##   
## Step: AIC=503.21  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso  
##   
## Df Sum of Sq RSS AIC  
## + calfcir 1 1601.37 19049 497.95  
## + bestvj 1 746.45 19904 501.90  
## + legpower 1 553.04 20098 502.77  
## <none> 20651 503.21  
## + tricep 1 174.83 20476 504.45  
## + ecto 1 107.34 20544 504.74  
## + armspan 1 74.75 20576 504.89  
## + bestsnr 1 42.05 20609 505.03  
## + thighci 1 36.63 20614 505.05  
## + thigh 1 9.73 20641 505.17  
## + estfm 1 7.48 20643 505.18  
## + weight 1 6.80 20644 505.18  
## + flexarm 1 4.54 20646 505.19  
## + endo 1 3.15 20648 505.20  
##   
## Step: AIC=497.95  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir  
##   
## Df Sum of Sq RSS AIC  
## + bestvj 1 481.32 18568 497.65  
## <none> 19049 497.95  
## + legpower 1 286.15 18763 498.59  
## + flexarm 1 278.17 18771 498.62  
## + thighci 1 257.74 18792 498.72  
## + armspan 1 196.38 18853 499.02  
## + tricep 1 130.75 18919 499.33  
## + bestsnr 1 122.38 18927 499.37  
## + ecto 1 35.40 19014 499.78  
## + thigh 1 25.90 19024 499.83  
## + weight 1 10.02 19039 499.90  
## + estfm 1 9.57 19040 499.90  
## + endo 1 1.16 19048 499.94  
##   
## Step: AIC=497.65  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir + bestvj  
##   
## Df Sum of Sq RSS AIC  
## <none> 18568 497.65  
## + armspan 1 233.742 18334 498.51  
## + thighci 1 224.398 18344 498.55  
## + tricep 1 180.044 18388 498.77  
## + flexarm 1 155.779 18412 498.89  
## + legpower 1 145.988 18422 498.93  
## + bestsnr 1 38.290 18530 499.46  
## + ecto 1 28.280 18540 499.51  
## + endo 1 9.589 18559 499.60  
## + estfm 1 4.491 18564 499.62  
## + weight 1 4.321 18564 499.62  
## + thigh 1 0.008 18568 499.65

##   
## Call:  
## lm(formula = racetime ~ estffm + expvarsity + tall + preexper +   
## biceps + meso + calfcir + bestvj, data = rowtime)  
##   
## Coefficients:  
## (Intercept) estffm expvarsity tall preexper biceps   
## 867.706 -1.299 -17.579 -2.404 -10.983 1.155   
## meso calfcir bestvj   
## -10.389 2.810 1.002

The best model selected using a foward step algorithm has an AIC of 497.65, and includes the variables: estffm, expvarsity, tall, preexper, biceps, meso, calfcir, & bestvj.

### Part 1e

Use extractAIC to compute the AIC for the best first order-model with 8 predictors from the **regsubsets** function. How does it compare with the AIC for the two models produced by the backward and forward selection procedure?

So far, which first-order model is the “best” according to the AIC? (remember, smaller is better for AIC)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

#regsubsets w/8 variables  
extractAIC(lm(formula = racetime~calfcir+biceps+bestvj+legpower+meso+ecto+expvarsity+preexper, data=rowtime))

## [1] 9.0000 497.3196

#step backward (9 variables)  
extractAIC(lm(formula = racetime~tall+calfcir+biceps+estfm+bestvj+legpower+meso+expvarsity+preexper, data=rowtime))

## [1] 10.0000 497.2245

#step forward (8 variables)  
extractAIC(lm(formula = racetime~estffm+expvarsity+tall+preexper+biceps+meso+calfcir+bestvj, data=rowtime))

## [1] 9.0000 497.6452

The backward step() model containing 9 variables has the lowest AIC value of 497.2245.

### Part 1f

Find the VIF for each independent variable in the model produced by the forward step wise procedure. Is there a serious problem with collinearity? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1f -|-|-|-|-|-|-|-|-|-|-|-

#loading the car package for the vif() function  
library(car)

## Loading required package: carData

#checking the vif scores of the independent variables. greater than 10 suggests a collinearity problem.  
vif(lm(formula = racetime~tall+calfcir+biceps+estfm+bestvj+legpower+meso+expvarsity+preexper, data=rowtime))

## tall calfcir biceps estfm bestvj legpower meso   
## 4.984964 2.773274 2.628953 5.808960 8.373223 18.570555 5.220502   
## expvarsity preexper   
## 1.168600 1.135906

There does appear to be a collinearity problem with the “legpower” variable because the vif score of ~18.57 is greater than 10.

### Part 1g

What about the possibility of adding quadratic terms to the model? In this case, we have a fairly manageable number of quantitative predictors to check for quadratic relationship between the response variable racetime and any predictors.

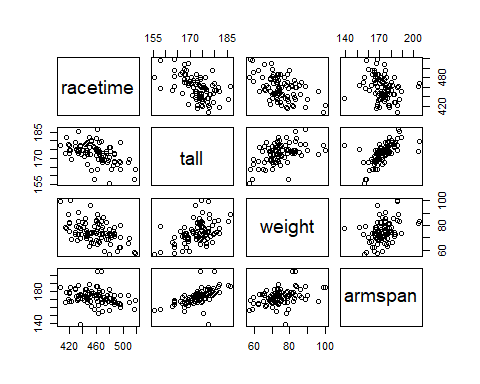
The R function pairs() can be used to look for quadratic relationships, but it will have to be restricted to about 4 predictors at a time so that the scatterplot matrices will be legible.

Since the response variable is in column 1 and the quantitative predictors are in columns 2 through 18, running the R code in the chunk shown below.

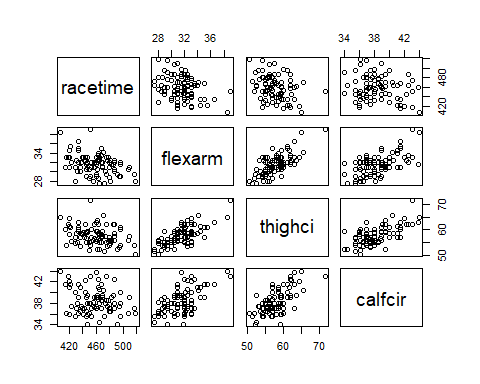
In each plot scatterplot matrix that is produced, look for any quadratic relationships between racetime and any of the predictor by examining the plots in the first row. Is there any obvious curvature in the trend for racetime with any of the predictors?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g -|-|-|-|-|-|-|-|-|-|-|-

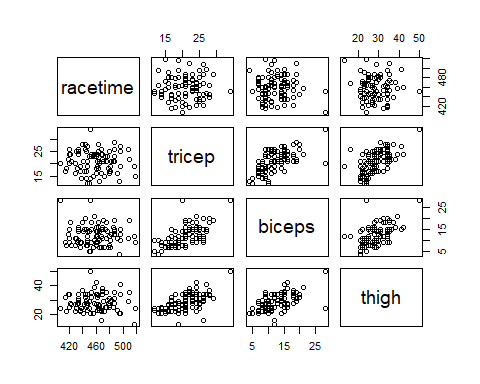
# The code in this chunk is provided for students  
library(DS705data)  
data("rowtime")  
pairs(rowtime[c(1,2,3,4)])



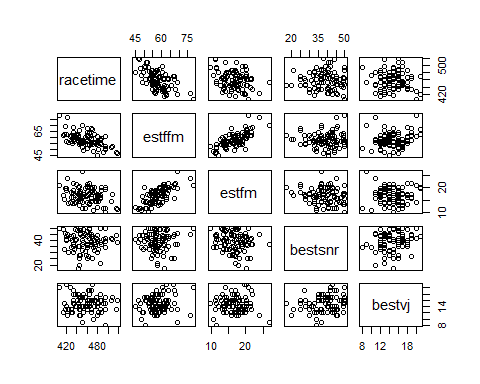
pairs(rowtime[c(1,5,6,7)])



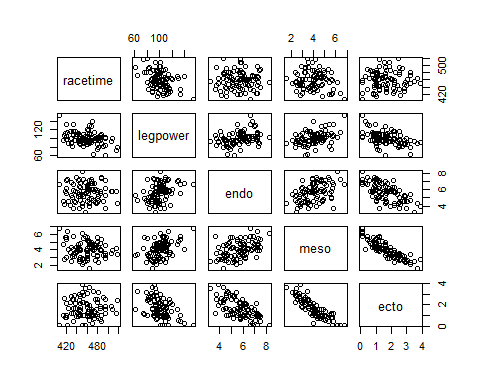
pairs(rowtime[c(1,8,9,10)])



pairs(rowtime[c(1,11,12,13,14)])



pairs(rowtime[c(1,15,16,17,18)])



After studying these scatterplots of the dependent vs. independent variables for a while, I don’t notice anything substantial that indicates a quadratic relationship.

### Part 1h

Something new will be covered in this part. All possible interactions can be examined using the step() function. This can be done using code like

step(initial\_model, scope = . ~ .^2, direction = ‘forward’)

where initial\_model is the output object from lm(). Using the output object for a first-order model would be a good initial model.

Higher order interactions can also be explored by replacing the 2 with the highest level of interaction desired, but we won’t go there in this assignment.

Fit the model from the forward step done previously

racetime~estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj

and use step() to look for the best model containing up to any two-way interaction terms. Report the model and the corresponding AIC for it.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1h -|-|-|-|-|-|-|-|-|-|-|-

#storing the results of the forward step() model into a variable  
model = lm(racetime~estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj, data = rowtime)  
 #using that model as a starting point, check the combination of all interaction terms to see if they add additional value.  
 #the first . means all variables. the \* means interaction, and the second . means all teams. So to summarize, check all terms interactions with all terms.   
step(model, scope = .~.\*., direction = "forward")

## Start: AIC=497.65  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir + bestvj  
##   
## Df Sum of Sq RSS AIC  
## + tall:calfcir 1 781.63 17787 495.77  
## + estffm:tall 1 675.93 17892 496.31  
## + estffm:bestvj 1 634.83 17933 496.51  
## + tall:meso 1 614.92 17953 496.61  
## <none> 18568 497.65  
## + biceps:meso 1 316.43 18252 498.10  
## + estffm:expvarsity 1 288.25 18280 498.24  
## + calfcir:bestvj 1 269.59 18299 498.33  
## + tall:biceps 1 250.94 18317 498.42  
## + meso:bestvj 1 194.76 18373 498.70  
## + estffm:biceps 1 156.30 18412 498.88  
## + expvarsity:tall 1 130.69 18437 499.01  
## + expvarsity:bestvj 1 126.94 18441 499.03  
## + tall:bestvj 1 125.77 18442 499.03  
## + meso:calfcir 1 124.59 18444 499.04  
## + preexper:biceps 1 109.68 18458 499.11  
## + biceps:bestvj 1 106.43 18462 499.13  
## + expvarsity:biceps 1 102.28 18466 499.15  
## + preexper:bestvj 1 53.98 18514 499.38  
## + expvarsity:meso 1 40.26 18528 499.45  
## + expvarsity:calfcir 1 34.38 18534 499.48  
## + tall:preexper 1 33.17 18535 499.48  
## + estffm:calfcir 1 32.73 18535 499.49  
## + expvarsity:preexper 1 28.70 18539 499.51  
## + estffm:preexper 1 25.63 18543 499.52  
## + biceps:calfcir 1 21.13 18547 499.54  
## + preexper:calfcir 1 3.07 18565 499.63  
## + estffm:meso 1 0.75 18567 499.64  
## + preexper:meso 1 0.31 18568 499.64  
##   
## Step: AIC=495.77  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir + bestvj + tall:calfcir  
##   
## Df Sum of Sq RSS AIC  
## + estffm:bestvj 1 1256.96 16530 491.18  
## + tall:bestvj 1 722.66 17064 494.04  
## + calfcir:bestvj 1 438.06 17348 495.53  
## <none> 17787 495.77  
## + estffm:biceps 1 386.87 17400 495.80  
## + estffm:expvarsity 1 304.16 17482 496.22  
## + biceps:meso 1 296.04 17490 496.26  
## + meso:bestvj 1 271.55 17515 496.39  
## + biceps:bestvj 1 234.31 17552 496.58  
## + expvarsity:bestvj 1 224.53 17562 496.63  
## + expvarsity:biceps 1 219.23 17567 496.66  
## + tall:meso 1 188.99 17598 496.81  
## + expvarsity:tall 1 165.31 17621 496.93  
## + estffm:tall 1 135.46 17651 497.09  
## + preexper:biceps 1 131.16 17655 497.11  
## + estffm:calfcir 1 114.93 17672 497.19  
## + expvarsity:meso 1 78.59 17708 497.38  
## + expvarsity:calfcir 1 56.78 17730 497.49  
## + estffm:preexper 1 54.12 17732 497.50  
## + meso:calfcir 1 53.99 17733 497.50  
## + estffm:meso 1 42.93 17744 497.56  
## + biceps:calfcir 1 25.32 17761 497.65  
## + preexper:bestvj 1 22.47 17764 497.66  
## + expvarsity:preexper 1 22.40 17764 497.66  
## + tall:biceps 1 6.84 17780 497.74  
## + preexper:calfcir 1 6.65 17780 497.74  
## + preexper:meso 1 2.65 17784 497.76  
## + tall:preexper 1 1.67 17785 497.77  
##   
## Step: AIC=491.18  
## racetime ~ estffm + expvarsity + tall + preexper + biceps + meso +   
## calfcir + bestvj + tall:calfcir + estffm:bestvj  
##   
## Df Sum of Sq RSS AIC  
## <none> 16530 491.18  
## + meso:bestvj 1 284.923 16245 491.61  
## + preexper:calfcir 1 283.368 16246 491.62  
## + preexper:meso 1 271.770 16258 491.69  
## + calfcir:bestvj 1 222.935 16307 491.96  
## + estffm:tall 1 220.346 16309 491.97  
## + biceps:bestvj 1 199.277 16330 492.09  
## + tall:bestvj 1 187.464 16342 492.15  
## + tall:meso 1 184.389 16345 492.17  
## + estffm:preexper 1 175.575 16354 492.22  
## + estffm:biceps 1 157.489 16372 492.32  
## + biceps:meso 1 142.848 16387 492.40  
## + expvarsity:tall 1 100.001 16430 492.63  
## + expvarsity:bestvj 1 79.425 16450 492.74  
## + estffm:expvarsity 1 66.885 16463 492.81  
## + tall:biceps 1 38.097 16491 492.97  
## + tall:preexper 1 31.958 16498 493.00  
## + expvarsity:calfcir 1 22.951 16507 493.05  
## + estffm:meso 1 15.721 16514 493.09  
## + estffm:calfcir 1 9.245 16520 493.13  
## + preexper:bestvj 1 5.020 16525 493.15  
## + biceps:calfcir 1 3.696 16526 493.16  
## + expvarsity:meso 1 2.004 16528 493.17  
## + meso:calfcir 1 1.923 16528 493.17  
## + expvarsity:biceps 1 0.669 16529 493.17  
## + expvarsity:preexper 1 0.391 16529 493.18  
## + preexper:biceps 1 0.046 16530 493.18

##   
## Call:  
## lm(formula = racetime ~ estffm + expvarsity + tall + preexper +   
## biceps + meso + calfcir + bestvj + tall:calfcir + estffm:bestvj,   
## data = rowtime)  
##   
## Coefficients:  
## (Intercept) estffm expvarsity tall preexper   
## 3560.2941 1.8653 -18.5377 -18.9470 -10.6704   
## biceps meso calfcir bestvj tall:calfcir   
## 1.3297 -11.7268 -72.4372 13.5107 0.4323   
## estffm:bestvj   
## -0.2074

Using the forward step() function to test the combination of all term interactions, we have identified a better model with AIC of 491.18. This model is:

racetime ~ estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj + tall:calfcir + estffm:bestvj

### Part 1i

Obtain the model summary for the model that resulted in Part 1h. Are there any predictors with coefficients that do not have coefficients that differ from 0 at the 5% level?

If so, drop those predictors from the model if they are not involved in any interactions and re-fit it without them. Compute both the and the AIC for that model.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1i -|-|-|-|-|-|-|-|-|-|-|-

#create a summary of the new model with the two additional interaction terms.   
model\_f = lm(racetime~estffm+expvarsity+tall+preexper+biceps+meso+calfcir+bestvj+tall:calfcir+estffm:bestvj, data=rowtime)  
summary(model\_f)

##   
## Call:  
## lm(formula = racetime ~ estffm + expvarsity + tall + preexper +   
## biceps + meso + calfcir + bestvj + tall:calfcir + estffm:bestvj,   
## data = rowtime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39.610 -9.961 -0.003 8.371 33.882   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3560.29413 1088.07630 3.272 0.001584 \*\*   
## estffm 1.86526 1.43120 1.303 0.196266   
## expvarsity -18.53770 3.37355 -5.495 4.62e-07 \*\*\*  
## tall -18.94699 6.38612 -2.967 0.003979 \*\*   
## preexper -10.67040 3.45390 -3.089 0.002769 \*\*   
## biceps 1.32970 0.44811 2.967 0.003974 \*\*   
## meso -11.72684 3.26783 -3.589 0.000575 \*\*\*  
## calfcir -72.43719 29.04882 -2.494 0.014733 \*   
## bestvj 13.51067 5.15213 2.622 0.010474 \*   
## tall:calfcir 0.43233 0.16691 2.590 0.011419 \*   
## estffm:bestvj -0.20740 0.08462 -2.451 0.016454 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.46 on 79 degrees of freedom  
## Multiple R-squared: 0.6879, Adjusted R-squared: 0.6484   
## F-statistic: 17.41 on 10 and 79 DF, p-value: 3.198e-16

Although the estffm variable is not significant on its own, it is considered significant when it is part of the interaction between estffm:bestvj, therefore i will not drop the individual term. This model results in staying the same, with an AIC of 491.18, and an Adjusted R2 of 0.6484.

### Part 1j

Let us refer to this final model as **Model F**. It should include the following terms:

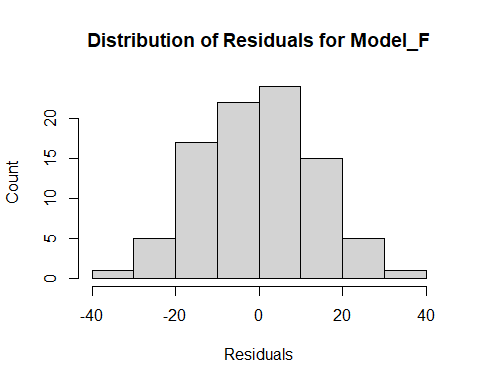
racetime ~ estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj + tall:calfcir + estffm:bestvj

If this will be our possible final model, it is necessary to evaluate the model assumptions.

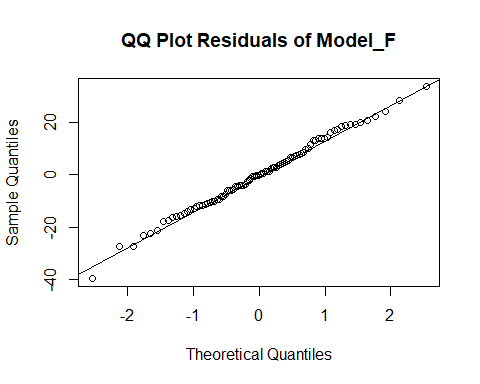
Are the residuals of **Model F** normal? Construct a histogram, normal probability plot, and boxplot of the residuals and perform a Shapiro-Wilk test for normality at a 5% level of significance. Justify your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1j -|-|-|-|-|-|-|-|-|-|-|-

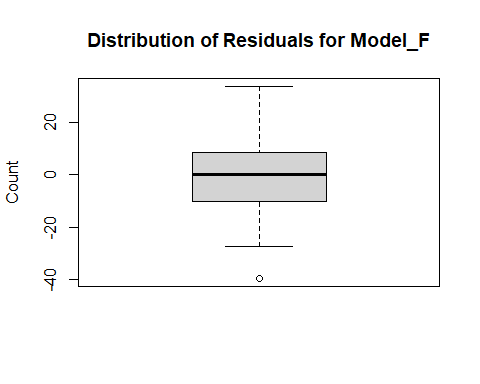
#create a histogram, qqplot, and boxplot of the residuals.   
hist(model\_f$residuals, xlab="Residuals", ylab = "Count", main = "Distribution of Residuals for Model\_F")



qqnorm(model\_f$residuals, main = "QQ Plot Residuals of Model\_F")  
qqline(model\_f$residuals)



boxplot(model\_f$residuals, ylab = "Count", main = "Distribution of Residuals for Model\_F")



#check normality of the residuals with the shapiro test.   
shapiro.test(model\_f$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model\_f$residuals  
## W = 0.99549, p-value = 0.9911

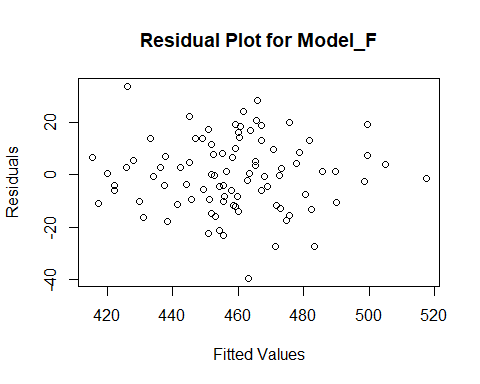
After reviewing the graphs above and performing the Shapiro test for normality, the residuals of the fitted model\_f appear to be normally distributed (shapiro’s pvalue = 0.9911)

### Part 1k

Construct a residual plot for **Model F**. Do you see any patterns indicating potential violations of model assumptions? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1k -|-|-|-|-|-|-|-|-|-|-|-

#create a residual plot of residuals vs. fitted values.   
plot(model\_f$fitted.values, model\_f$residuals, xlab="Fitted Values", ylab = "Residuals", main = "Residual Plot for Model\_F")



There doesn’t appear to be any visual patterns that may indicate any violations for the model’s assumptions. Perhaps an outlier or two, but more investigation would need to be done.

### Part 1l

Perform the Bruesch-Pagan test for equal variances of the residuals at a 5% level of significance. What does you conclude from this test?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1l -|-|-|-|-|-|-|-|-|-|-|-

#load in the lmtest library  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

#run the equal variance test.  
bptest(model\_f)

##   
## studentized Breusch-Pagan test  
##   
## data: model\_f  
## BP = 11.468, df = 10, p-value = 0.3222

At a 95% level of significance, there is not enough evidence to suggest that model\_f’s residuals are not equal (p-value = 0.3222).

### Part 1m

How do you feel about this last model being the “best”?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1m -|-|-|-|-|-|-|-|-|-|-|-

Given that all the assumptions check out, I feel pretty good about the model being the best as it stands today - the Adjusted R2 value on this final model is about 0.65 (thus this model should explain about 65% of the variability in NCAA D1 Women’s row time performance) and the terms are not overly complicated or extraneous.

## Exercise 2

In a study of small, constructed agricultural ponds in southeastern Minnesota, pond and the surrounding landscape features were used to assess their value as amphibian breeding sites. One measure of this was when the amphibian species richness was at least four.

The data frame is farmpond.rda in the DS705data package.

Species richness is the number of different species observed at each pond and the variable RICH is defined as:

RICH = 1 if species richness is at least 4; RICH = 0 otherwise.

Furthermore,

FISH = 1 if fish are present; FISH = 0 otherwise

and the remaining variables are quantitative.

### Part 2a

Suppose our goal is to build the “best” logistic regression model to predict species richness of at least 4 (i.e. RICH=1). Fit the first order logistic regression model using all of the available predictors in the file.

Also fit the null model (intercept only) and use step() with forward selection to search for the best first-order logistic regression model with these variables. Identify the resulting model.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

#load the farmpond dataset and build a logistic regression model for the RICH variable using all the other predictor variables.  
data("farmpond")  
 #building a full model (for the upper limit)  
full2model = glm(RICH~., data = farmpond, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

#building a null model (or starting point, for the lower bounds) so the "step forward" can add to it.  
null2model = glm(RICH~1, data = farmpond, family = "binomial")  
 #starting with the null model, add one variable at a time until you bottom out on the AIC score.   
 #dont forget make the lower and upper a LIST type.  
step(null2model, scope = list(lower = null2model, upper = full2model), direction = "forward")

## Start: AIC=52.45  
## RICH ~ 1  
##   
## Df Deviance AIC  
## + COND 1 40.162 44.162  
## + TOTNITR 1 40.446 44.446  
## + FISH 1 45.397 49.397  
## + TURB 1 45.860 49.860  
## + POND\_AREA 1 48.242 52.242  
## <none> 50.446 52.446  
## + DISOXY 1 48.712 52.712  
## + W\_DEPTH\_MEAN 1 49.372 53.372  
## + TEMP 1 50.429 54.429  
##   
## Step: AIC=44.16  
## RICH ~ COND

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## + TOTNITR 1 34.365 40.365  
## + TURB 1 36.667 42.667  
## <none> 40.162 44.162  
## + FISH 1 38.320 44.320  
## + POND\_AREA 1 38.869 44.869  
## + W\_DEPTH\_MEAN 1 39.908 45.908  
## + TEMP 1 40.107 46.107  
## + DISOXY 1 40.133 46.133

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=40.36  
## RICH ~ COND + TOTNITR

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## + FISH 1 25.587 33.587  
## + POND\_AREA 1 31.687 39.687  
## <none> 34.365 40.365  
## + DISOXY 1 33.016 41.016  
## + W\_DEPTH\_MEAN 1 33.660 41.660  
## + TURB 1 34.046 42.046  
## + TEMP 1 34.347 42.347

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=33.59  
## RICH ~ COND + TOTNITR + FISH

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## <none> 25.587 33.587  
## + DISOXY 1 24.131 34.131  
## + POND\_AREA 1 24.326 34.326  
## + W\_DEPTH\_MEAN 1 24.762 34.762  
## + TEMP 1 25.073 35.074  
## + TURB 1 25.578 35.578

##   
## Call: glm(formula = RICH ~ COND + TOTNITR + FISH, family = "binomial",   
## data = farmpond)  
##   
## Coefficients:  
## (Intercept) COND TOTNITR FISH   
## 4.4978820 -0.0002472 -4.2033973 -3.9906394   
##   
## Degrees of Freedom: 39 Total (i.e. Null); 36 Residual  
## Null Deviance: 50.45   
## Residual Deviance: 25.59 AIC: 33.59

The “best” model with an AIC score of 33.59 is:

RICH ~ COND + TOTNITR + FISH

### Part 2b

Construct a classification table (also known as a confusion matrix) for the model identified in the previous part. Use 0.5 as the cutoff probability.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## predRICH  
## Sp Rich<4 Sp Rich>=4 Sum  
## 0 9 4 13  
## 1 4 23 27  
## Sum 13 27 40

## [1] "Proportion correctly predicted = 0.8"

### Part 2c

Compute McFadden’s pseudo for the model identified in part 2a.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2c -|-|-|-|-|-|-|-|-|-|-|-

#load the function from the pscl package  
library(pscl)

## Classes and Methods for R developed in the  
## Political Science Computational Laboratory  
## Department of Political Science  
## Stanford University  
## Simon Jackman  
## hurdle and zeroinfl functions by Achim Zeileis

#using the pR2 function to get the "McFaddens pseudo R2" value. This is like an R2 value for linear regression, but for logistic.  
pR2(model2)[4]

## fitting null model for pseudo-r2

## McFadden   
## 0.4927857

### Part 2d

Conduct a Hosmer-Lemeshow test of goodness-of-fit for the model from part 2a. Use 5 groups and use a 5% level of significance.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2d -|-|-|-|-|-|-|-|-|-|-|-

#load the function from the ResourceSelection package  
library(ResourceSelection)

## ResourceSelection 0.3-5 2019-07-22

#conduct the Hoslem-Lemeshow GOF test for logistic regression. This basically tests if the model is a good fit to the data.  
 #null = is a good fit; alt = is not a good fit.  
hoslem.test(farmpond$RICH, fitted(model2), g=5)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: farmpond$RICH, fitted(model2)  
## X-squared = 0.90075, df = 3, p-value = 0.8252

At a 95% level of significance, there is not enough evidence to suggest that the model is not a good fit to the data (p-value = 0.8252).