Regression Model Selection

Matt Allen

07/17/2018

Knit a Word file from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word file via D2L Dropbox.

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.

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

1. What type of variable is the response variable racetime (categorical or quantitative)?

The response variable racetime is quantitive.

1. Does this indicate linear regression or logistic regression?

This indicates use of linear regression vs logistic regression, because the response variable can have a range of values versus a binary interpretation in logistic regression.

1. What types of variables are there in the pool of potential predictors? Categorical, quantitative, or a mixture of each?

There appears to be a mixture. The summary statement does not interpret expvarsity and preexper as categorical; however, from inspection of the data, they only takes on the values 0 and 1, and could be classified as categorical. All the other variables are quantitative.

### Part 1b

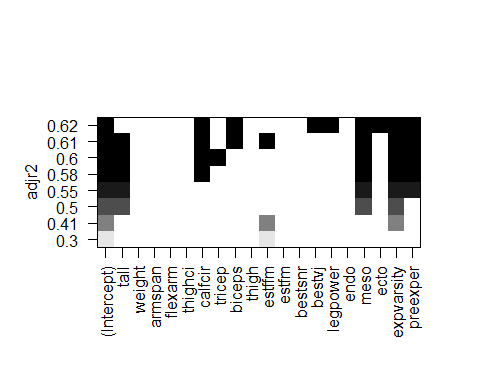
Use the **regsubsets** function to find the “best” first-order model for predicting the response variable racetime with up to 8 of the 19 predictor variables in the data set. Produce the summary and the plot for the best single models with up to 8 predictors according to .

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

summary(allmods)$adjr2

## [1] 0.3020824 0.4089748 0.4967972 0.5546830 0.5846293 0.5970301 0.6096559  
## [8] 0.6162122

plot(allmods, scale="adjr2")

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

The variables in the best first-order model with 8 predictors are calfcir, biceps, bestvj, legpower, meso, ecto, expvarsity, and preexper.

What is the for the best first-order model?

The for the best first-order model is 0.6162122.

### 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.

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

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

The independent variables in the model with the lowest AIC of 497.22 are tall, calfcir, biceps, estfm, bestvj, legpower, meso, expvarsity, and 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.

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

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

The independent variables in the model with the lowest AIC of 497.65 are estffm, expvarsity, tall, preexper, biceps, meso, calfcir and bestvj.

### Part 1e

Use extractAIC to compute the AIC for the 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 -|-|-|-|-|-|-|-|-|-|-|-

# Get AIC from first order-model with 8 predictors.  
regsubsets.best.aic <- lm(racetime~calfcir+biceps+bestvj+legpower+meso+ecto+expvarsity+preexper,data=rowtime)  
extractAIC(regsubsets.best.aic)

## [1] 9.0000 497.3196

The AIC for regsubsets function is in between the forward and backward selection procedure.

(ForwardStepAIC = 497.65) > (RegsubsetsAIC = 497.3196) > (BackwardStepAIC = 497.22)

The model with the lowest AIC was produced by the backward selection procedure.

racetime ~ tall + calfcir + biceps + estfm + bestvj + legpower + meso + expvarsity + preexper

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Find the VIF for each independent variable in the model produced by the forward step wise procedure  
forw.step.model <- lm(formula = racetime ~ estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj, data = rowtime)  
vif(forw.step.model)

## estffm expvarsity tall preexper biceps meso   
## 5.607151 1.178803 5.034022 1.134304 1.723637 5.152078   
## calfcir bestvj   
## 2.705303 1.267549

Since all the values for VIF are less that 10, there does not appear to be a serious problem with collinearity.

### 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.

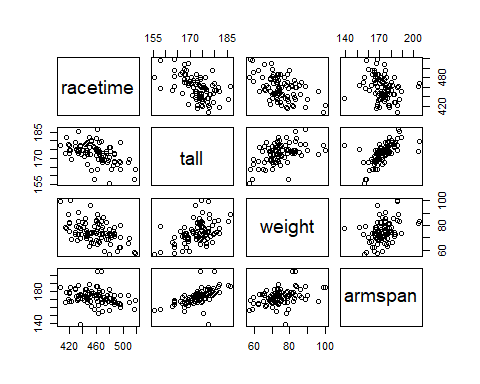
The R function pairs() can be used to look for quadratic relationships, but it will have to be restricted to about 4 predictors at a time so that the scatterplot matrices will be legible.

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.

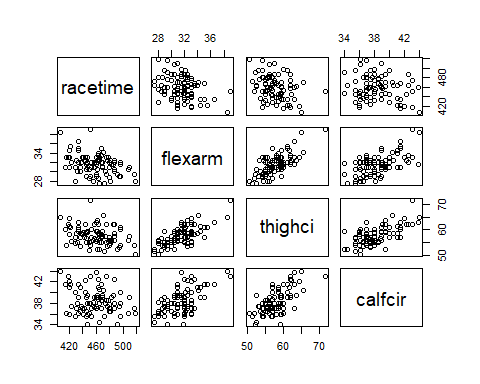
In each plot scatterplot matrix that is produced, look for any quadratic relationships between racetime and any of the predictor by examining the plots in the first row. Is there any obvious curvature in the trend for racetime with any of the predictors?

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

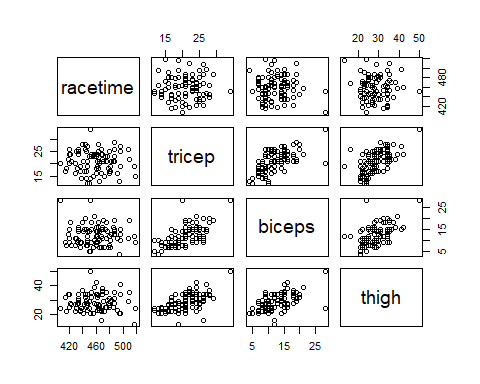
# The code in this chunk is provided for students  
library(DS705data)  
data("rowtime")  
pairs(rowtime[c(1,2,3,4)])



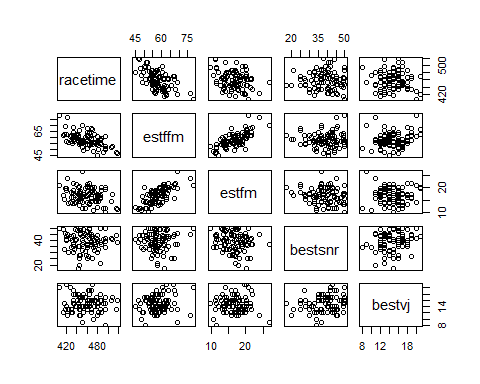
pairs(rowtime[c(1,5,6,7)])



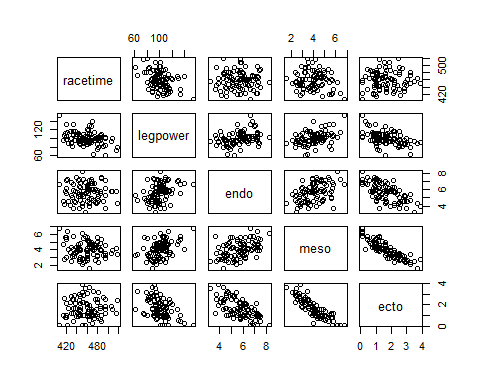
pairs(rowtime[c(1,8,9,10)])



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



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



I do not notice any obvious curvature in the relationship between racetime and its predictors.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

The model with the lowest AIC is given by

Step: AIC=491.18 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 -|-|-|-|-|-|-|-|-|-|-|-

# run summary on interaction model.   
model.f <- lm(formula = racetime ~ estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj + tall:calfcir + estffm:bestvj, data = rowtime)  
summary(model.f)$adj.r.squared

## [1] 0.6484292

extractAIC(model.f)

## [1] 11.0000 491.1784

The p-value for estffm is 0.196266. It is not significant at a 5% level, but its interaction with bestvj is significant at 5% level, so it can stay in the model. The of the model is 0.6484292. The AIC is 491.1784.

### Part 1j

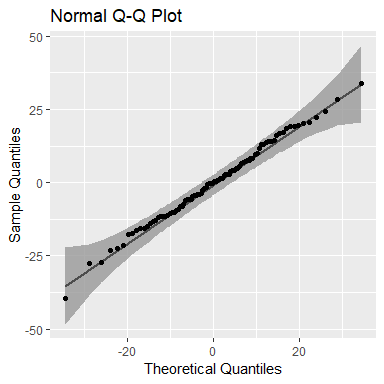
Let us refer to this final model as **Model F**. It should include the following terms:

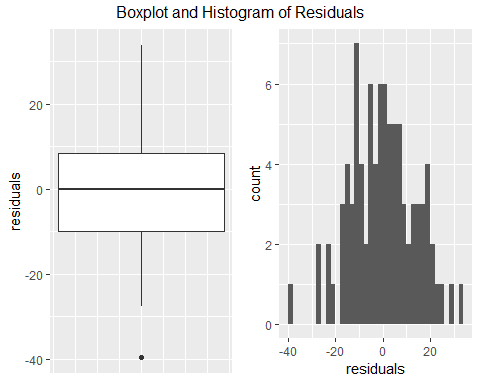
racetime ~ estffm + expvarsity + tall + preexper + biceps + meso + calfcir + bestvj + tall:calfcir + estffm:bestvj

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 -|-|-|-|-|-|-|-|-|-|-|-





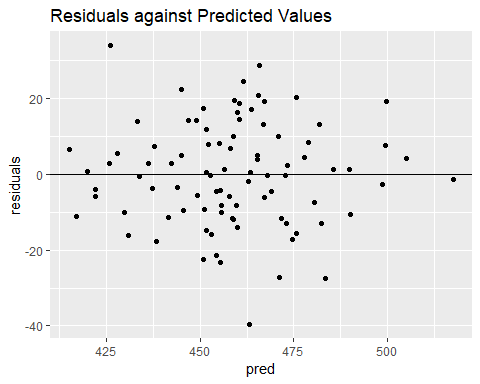
##   
## Shapiro-Wilk normality test  
##   
## data: resids  
## W = 0.99549, p-value = 0.9911

From the box plot of the residuals, the residuals appear appear to be centered at 0, but there is an outlier close to -40, which may indicate left skewness. The histogram as well looks normal, but may be left skewed. However based on the sharpiro-wilk test, we fail to reject that the residuals are normally distributed.

### Part 1k

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

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



The residuals appear to be scattered evenly. However, there appear to be a few outliers. It is possible that the errors do not have the same variance throughout, which would be a violation of a model assumption.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Perform the Bruesch-Pagan test  
require(lmtest)  
bptest(model.f)

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

The BP statistic is 11.468 and the p-value is 0.3222. At a 5% level of significance, we fail to reject the null hypothesis that the model has equal variances.

### Part 1m

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

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

It appears the model both satisfies all the assumptions and has the best AIC statistic. The model is the “best” in the sense that it does not use every variable, which makes it more interpretable, but has enough variables to make predictions meaningful.

## 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 -|-|-|-|-|-|-|-|-|-|-|-

The “best” model resulting from forward selection is given by

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 -|-|-|-|-|-|-|-|-|-|-|-

## 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 -|-|-|-|-|-|-|-|-|-|-|-

r2 <- pR2(forward.model) # use McFadden R-square, package = "pscl"  
 r2[4] # McFadden's R-square is in the 4th column of the output

## 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 -|-|-|-|-|-|-|-|-|-|-|-

hoslem.test(farmpond$RICH, fitted(forward.model), g=5)

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

When g=5, the p-value is .8252. At a 5% level of significance, there is insufficient evidence to say that this model does not fit.