Regression Model Selection

Your name here

mm/dd/yyyy

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.

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

# Insert your R code here.

Replace this text with your answer here.

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

# Insert your R code here.

Replace this text with your answer here.

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

# Insert your R code here.

Replace this text with your answer here.

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

# Insert your R code here.

Replace this text with your answer here.

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

# Insert your R code here.

Replace this text with your answer here.

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

# Insert your R code here.

Replace this text with your answer here.

### 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.

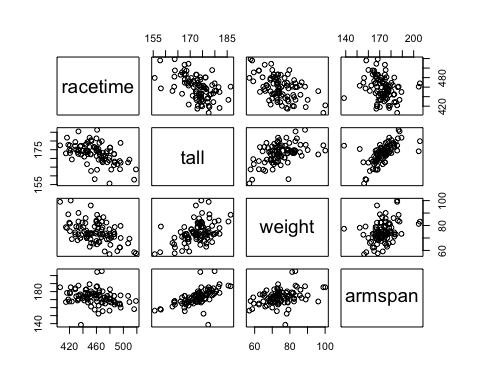
The R function pairs() can be used to look for quadratic relationships, but it will have to be restricted to about 4 predictors at a time so that the scatterplot matrices will be legible.

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.

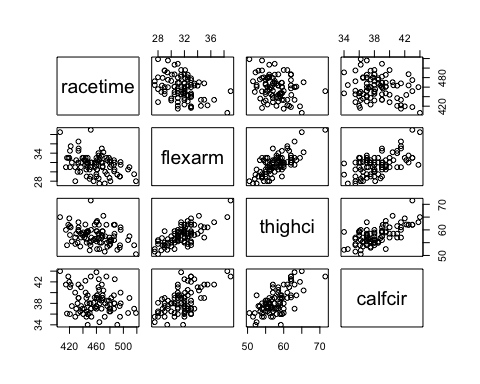
In each plot scatterplot matrix that is produced, look for any quadratic relationships between racetime and any of the predictor by examining the plots in the first row. Is there any obvious curvature in the trend for racetime with any of the predictors?

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

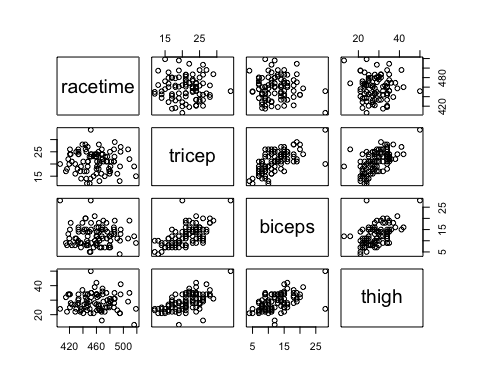
# The code in this chunk is provided for students  
library(DS705data)  
data("rowtime")  
pairs(rowtime[c(1,2,3,4)])



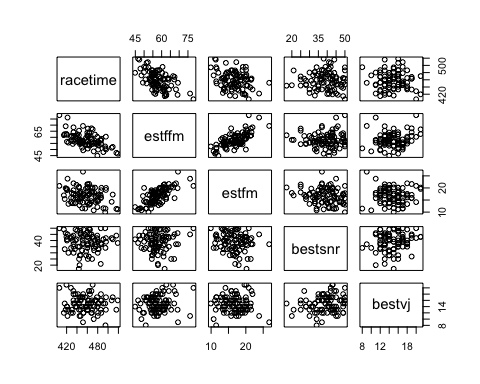
pairs(rowtime[c(1,5,6,7)])



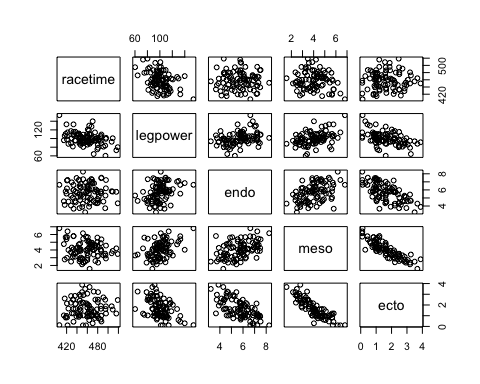
pairs(rowtime[c(1,8,9,10)])



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



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



Replace this text with your answer here.

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

# Insert your R code here.

Replace this text with your answer here.

### Part 1i

Obtain the model summary for the model that resulted in Part 1g. 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 -|-|-|-|-|-|-|-|-|-|-|-

# Insert your R code here.

Replace this text with your answer here.

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

# Insert your R code here.

Replace this text with your answer here.

### Part 1k

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

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

# Insert your R code here.

Replace this text with your answer here.

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

# Insert your R code here.

Replace this text with your answer here.

### Part 1m

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

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

Replace this text with your answer here.

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

# Insert your R code here.

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

# Insert your R code here.

### Part 2c

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

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

# Insert your R code here.

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

# Insert your R code here.

Replace this text with your answer here.