Lab 2: Multiple linear regression

Relationships of Body Dimensions

The data set we will utilize in this lab contains 21 body dimension measurements, age, weight, height, and gender of 507 individuals mostly in their twenties and thirties. All individuals exercise regularly and are considered physically active. The data was submitted to the Journal of Statistical Education by Grete Heinz and Louis J. Peterson, who took measurements at San Jose State University, the U.S. Naval Postgraduate School in Monterey, California, and dozens of California health and fitness clubs. The data set itself is available through the American Statiscial Association, and a description of the variables is provided here *http://www.amstat.org/publications/jse/datasets/body.txt*.

First, we will load the data and give the columns their respective names, found at the link above.

```r
require(RCurl)
URL <- getURL("http://www.amstat.org/publications/jse/datasets/body.dat.txt",
              ssl.verifypeer=FALSE)
body_dat <- read.table(text=URL)
names(body_dat) <- c("biac.diam", "pelvic.bredth", "bitro.diam", "chest.dep", "chest.diam",
                 "elbow.diam","wrist.diam", "knee.diam", "ankle.diam", "shoulder", "chest",
                 "waist", "navel", "hip", "thigh", "bicep", "forearm", "knee", "calf",
                 "ankle.min", "wrist.min", "age", "weight", "height", "gender" )
```

Data exploration

Let's start by just familiarizing ourselves with the dataset at hand.

> **Exercise 1** Refer to the link provided for the data description. What was this data originally collect for? What were the authors trying to investigate?

There are 21 different body dimension measurements provided as well as height and weight. We should first explore some of the data to see what the distribution of some of these measurements look like.

> **Exercise 2** First let's look at the distribution of weight in this sample. Create a histogram or boxplot to visualize the center and spread of **weight**. Describe.

> **Exercise 3** Choose 2 other variables (besides **weight**) you feel would be important measurements in predicting someone's weight. Look at their relationship using an appropriate visualization (scatterplot, side-by-side boxplots, or mosaic plot).

> **Exercise 4** The **gender** variable is stored as a numeric variable where 1 means female and 0 means male. Overwrite it as a factor variable using the **factor()** function. Remember, you can find help on a function using the **?** syntax, such as

```r
?factor
```

This is a product of statsTeachR that is released under a Creative Commons Attribution-ShareAlike 3.0 Unported. This lab was adapted for statsTeachR by Sara Nuñez, Nicholas Reich and Andrea Foulkes from an OpenIntro Statistics lab written by Andrew Bray and Mine Çetinkaya-Rundel.

Simple linear regression

It seems intuitive that most of these meaurements will be associated with a person's weight. Let's look at a few of these relationships using scatterplots (be sure to zoom in on the plots to see the relationships more clearly):

```
library(ggplot2)
qplot(hip, weight, data=body_dat)
qplot(chest, weight, data=body_dat)
qplot(wrist.diam, weight, data=body_dat)
qplot(thigh, weight, data=body_dat)
```

> **Exercise 5** Describe the 4 scatterplots. Do any of the plots seem to have more than one trend? If so, why do you think this is? Explore your hypothesis with another visualization of your choosing.

> **Exercise 6** Let's see if the apparent trend in the plot of **weight** and **chest** is something more than natural variation. Fit a linear model called **lm_chest** to predict average weight by average chest girth and add the line to your plot using **abline(**lm_chest**)**, or a **geom_smooth** in ggplot2. Write out the equation for the linear model and interpret the slope. Is chest girth a statistically significant predictor? Report the multiple R-squared value. Is this a positive or negative linear relationship?

> **Exercise 7** Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see Lab 7 for a reminder of how to make these).

Multiple linear regression

As you probably noticed in plots made above, there are more than 1 possible predictors for a persons weight in this data set. Fortunately, we are not restrained to using only one predictor. We can add another predictor to the model and check its significance and re-check the model diagnostics. This process is intuitively exhausting, as there are often hundreds of possible combinations of variables that are significant on their own as a single predictor of weight. However, this is does not necessarily mean that they including all of them will make the strongest model.

Let's start slow by adding one possible predictor to **lm_chest**. In order to see if chest is still a significant predictor of weight after we've accounted for the height of the person, we can add the **height** term into the model.

```
lm_chest_height <- lm(weight ~ chest + height, data = body_dat)
summary(lm_chest_height)
```

> **Exercise 8** P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

> **Exercise 9** Is **chest** still a significant predictor of **weight**? Has the addition of **height** to the model changed the parameter estimate for **chest**?

Let's try out a new model with a categorical variable. For starters, let's fit a model that uses **gender** and **thigh** to predict a person's weight.

```
lm_gender_thigh <- lm(weight ~ gender + thigh, data=body_dat)
summary(lm_gender_thigh)
```

Exercise 10 Are the two predictors in this model significant? Be sure to check the model diagnostics before considering the p-values.

Note that the estimate for **gender** is now called **gendermale**. You'll see this name change whenever you introduce a categorical variable. The reason is that R recodes **gender** from having the values of **female** and **male** to being an indicator variable called **gendermale** that takes a value of 0 for females and a value of 1 for males. (Such variables are often referred to as "dummy" variables.)

As a result, for females, $\hat{\beta}_{gender}$ is multiplied by zero, leaving the intercept and slope in a familiar simple regression format.

Exercise 11 What is the equation of the line corresponding to males? For two individuals who have the same thigh measurement, which gender tends to weigh more?

The interpretation of the coefficients in multiple regression is slightly different from that of simple regression. The estimate for **thigh** reflects how much more an individual is expected to weigh if they have a thigh measurement that is one unit higher *while holding all other variables constant*. In this case, that translates into considering only individuals of the same gender with **thigh** measurements that are one unit apart.

The search for the best model

We will start with a "full" model that predicts weight based on the girths of the shoulder, chest, waist, navel, hip, thigh, bicep, forearm, knee and calf.

Exercise 12 Which variable would you expect to have the highest p-value in this model? Why? *Hint:* Think about which variable would you expect to not have any or little association with weight.

Let's run the model...

```
lm_full <- lm(weight ~ shoulder + chest + waist + navel + hip + thigh + bicep +
    forearm + knee + calf, data = body_dat)
summary(lm_full)
```

Exercise 13 Check your suspicions from the previous exercise. Include the model output in your response. How much of the total variability in weight is explained by all of the predictors included?

Exercise 14 Interpret the coefficient associated with the hip variable, including reference to the actual units of observation of each variable.

Exercise 15 Drop the variable with the highest p-value and re-fit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficient estimates depend on the other variables that are included in the model.)

Exercise 16 Using "backward-selection" and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write out the linear model for predicting weight based on the final model you settle on.

Exercise 17 Verify that the conditions for this model are reasonable using diagnostic plots.

A cautionary note: Backwards and forwards selection is often used to find a "best" model fit, but should be used with hesitation. While it makes sense to some degree to eliminate or add variables based on their p-values, this does not consider every possible combination of variables. As we saw above, different combinations change the predictor parameters, and thus every possible combination should be considered. It is important to realize, however, that while many statisticians often think they can find the single most perfect model, *no such model exists!* The key here is to find a model that makes sense, has significant variables, follows the model diagnostics, and is interpretable when necessary.