Lab 4: Introduction to linear regression

Cigarettes and carbon monoxide emissions

An abundance of research has been done to assess the direct health impacts of cigarette smoke. Studies have also investigated the effects that different cigarette brands have on the environment based on their chemical make-ups. While each chemical in cigarettes are considered hazardous to the smoker's health by the United States Surgeon General, in this lab we will be interested in seeing if there is an association between the amount of chemicals and the amount of carbon monoxide emitted into the environment.

The data

The data set presented here is taken from the 3rd edition of *Statistics for Engineering and the Sciences* by Mendenhall and Sincich (1992) and is a subset of the data produced by the Federal Trade Commission. This data was found through the American Statistical Association website, and a fuller description of the data can be found at *http://www.amstat.org/publications/jse/datasets/cigarettes.txt*. Let's load the data and look at summary of the variables. Be sure to install the package *RCurl* in order to obtain the data from the internet.

```
library(RCurl)
URL <-getURL("http://www.amstat.org/publications/jse/datasets/cigarettes.dat.txt",
             ssl.verifypeer=FALSE)
cigs <-read.table(text=URL)
names(cigs)<-c("brand","tar","nicotene","weight","CO")

summary(cigs)
```

> **Exercise 1** What type of plot would you use to display the relationship between **CO** and one of the other numerical variables? Plot this relationship using the variable **tar** as the predictor. Does the relationship look linear? If you knew how much tar was in a given brand of cigarettes, would you be comfortable using a linear model to predict the carbon monoxide content of that brand?

If the relationship looks linear, we can quantify the strength of the relationship with the correlation coefficient.

```
cor(cigs$CO, cigs$tar)
```

Sum of squared residuals

Think back to the way that we described the distribution of a single variable. Recall that we discussed characteristics such as center, spread, and shape. It's also useful to be able to describe the relationship of two numerical variables, such as **CO** and **tar** above.

> **Exercise 2** Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

Just as we used the mean and standard deviation to summarize a single variable, we can summarize the relationship between these two variables by finding the line that best follows their association. Use the following interactive function to select the line that you think does the best job of going through the cloud of points.

Note: You need to access this function by sourcing a function on GitHub using the following commands:

```
u <- "https://raw.githubusercontent.com/nickreich/stat-modeling-2015/gh-pages/assets/labs/plot_ss.R"
script <- getURL(u, ssl.verifypeer = FALSE)
eval(parse(text = script))
plot_ss(x = cigs$tar, y = cigs$CO)
```

After running the last command above, you'll be prompted to click two points on the plot to define a line. Once you've done that, the line you specified will be shown in black and the residuals in red. Note that there are 25 residuals, one for each of the 25 observations. Recall that the residuals are the difference between the observed values and the values predicted by the line:

$$e_i = y_i - \hat{y}_i$$

The most common way to do linear regression is to select the line that minimizes the sum of squared residuals. The squared residuals are represented in this plot with blue dashed lines.

> **Exercise 3** Try running the above command again, this time with a line that is not a good fit. What happens to the squared residuals? Compare the sum of squares (given in the R output) of this poorly fit line to the first line you fit. Are you suprised at these results?

> **Exercise 4** Run this code several more times trying to minimize the sum of squares each time. What is the smallest sum of squares you can obtain? How does it compare to your neighbors? Compared to the first line you drew, what adjustments did you make to reduce the RSS?

The linear model

It is rather cumbersome to try to get the correct least squares line, i.e. the line that minimizes the sum of squared residuals, through trial and error. Instead we can use the **lm** function in R to fit the linear model (a.k.a. regression line).

```
m1 <- lm(CO ~ tar, data = cigs)
```

The first argument in the function **lm** is a formula that takes the form **y ~ x**. Here it can be read that we want to make a linear model of **CO** as a function of **tar**. The second argument specifies that R should look in the **cigs** data frame to find the **CO** and **tar** variables.

The output of **lm** is an object that contains all of the information we need about the linear model that was just fit. We can access this information using the summary function.

```
summary(m1)
```

Let's consider this output piece by piece. First, the formula used to describe the model is shown at the top. After the formula you find the five-number summary of the residuals. The "Coefficients" table shown next is key; its first column displays the linear model's y-intercept and the coefficient of **tar**. With this table, we can write down the least squares regression line for the linear model:

$$\hat{y} = 2.74328 + 0.80098 * tar$$

One last piece of information we will discuss from the summary output is the Multiple R-squared, or more simply, R^2. The R^2 value represents the proportion of variability in the response variable that is explained by the explanatory variable. For this model, 91.68% of the variability in carbon monoxide content is explained by the amount of tar in the cigarette.

Exercise 5 Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ from the model of **CO** as a function of **tar** by hand (i.e. using arithmetic/linear algebra and R as your calculator). Confirm that they match up with the values from the fitted model using **lm**.

Exercise 6 What does the slope tell us in the context of the relationship between the amount of carbon monoxide emitted into the environment and the amount of tar in the cigarette?

Exercise 7 Fit a new model **m2** that uses **weight** to predict **CO**. Using the estimates from the R output, write the equation of the regression line. How much of the variability in CO emission is explained by the weight of the cigarette? Which model, **m1** or **m2**, would you trust more to predict CO emission? Explain.

Prediction and prediction errors

Let's create a scatterplot of **CO** versus **tar** with the least squares line laid on top.

```
qplot(tar, CO, data=cigs)
ggplot(cigs, aes(tar, CO)) + geom_point() + geom_smooth(method="lm")
ggplot(cigs, aes(tar, CO)) + geom_point() + geom_smooth(method="lm", se=FALSE)
```

The fitted line can be used to predict y at any value of x. When predictions are made for values of x that are beyond the range of the observed data, it is referred to as *extrapolation* and is not usually recommended. However, predictions made within the range of the data are more reliable. They're also used to compute the residuals.

Exercise 8 If you saw the least squares regression line and not the actual data, how much CO (mg) would you predict to be emitted from a cigarette with 15 mg of tar? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

Model diagnostics

To assess whether the linear model is reliable, we need to check for (1) linearity, (2) nearly normal residuals, and (3) constant variability.

1. Linearity: You already checked if the relationship between CO content and amount of tar is linear using a scatterplot. We should also verify this condition with a plot of the residuals vs. **tar**.

```
qplot(tar, m1$residuals, data=cigs) + geom_hline(yintercept=0, linetype=3)
```

Exercise 9 Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between CO content and tar?

2. Nearly normal residuals: To check this condition, we can look at a histogram

```
qplot(m1$residuals)
```

or a normal probability plot of the residuals. Recall that any code following a # is intended to be a comment that helps understand the code but is ignored by R.

```
qqnorm(m1$residuals)
qqline(m1$residuals)  # adds diagonal line to the normal prob plot
```

Exercise 10 Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

3. Constant variability:

Exercise 11 Based on the plot in (1), does the constant variability condition appear to be met?

On Your Own

1. Produce a scatterplot of **CO** and **nicotene** and fit a linear model. At a glance, does there seem to be a linear relationship?
2. How does this relationship compare to the relationship between **CO** and **tar**? Use the R^2 values from the two model summaries to compare. Does **nicotene** seem to predict **CO** better than **tar**? How can you tell?
3. Which variable best predicts **CO** out of the three in this data set? Support your conclusion using the graphical and numerical methods we've discussed.
4. Check the model diagnostics for the regression model with the variable you decided was the best predictor for CO content.