Predicting Student Grades

SAMSI Undergraduate Workshop on Data Science

You may access the R code for this document at https://github.com/johnnardini/SAMSI_UG_Data_science.

We will use linear regression to predict student scores based on their gender, number of school absences, and daily study time.

1. To get started, we need to load in our data on students. To do so, open rstudio and enter

   ```
   setwd(filename)
   ```

   where filename is the folder on your computer where the file student.CSV is located. We can then read this file into memory by typing

   ```
   data = read.csv("student.csv",header=TRUE)
   ```

   "data" now represents a data frame with the students' information. To see what information we have for each student, enter

   ```
   colnames(data)
   ```

   for which we see

   ```
   "gender" "studytime" "absences" "grade"
   ```

   If for example, we want to view the students' genders, we can enter

   ```
   data$gender
   ```

2. Let's explore the data! Do you think that students who study longer tend to score better than students who study less? We can verify this by entering

   ```
   plot(data$studytime,data$grade,xlab="Hours Spent Studying",ylab="Grade")
   ```

   How do students who study 4 hours compare to students who study 1 hour?

3. Similar to the previous example, let's consider if student attendance impacts grade. We can investigate this visually by

   ```
   plot(data$absences,data$grade,xlab="Number of Absences",ylab="Grade")
   ```

4. Based on the two previous questions, we may suggest that a student's grade depends on their number of absences or hours spent studying. Mathematically, we may write this as a linear model given by:

$$y \approx \beta_0 + \beta_1 x$$

where y denotes a student's final grade, x denotes some other variable (such as hours spent studying or number of absences). The two terms β_0 and β_1 are coefficients that detail y's relation to x. Note that if $\beta_1 > 0$, then y increases when x increases. On the other hand, if $\beta_1 < 0$, then y decreases when x increases.

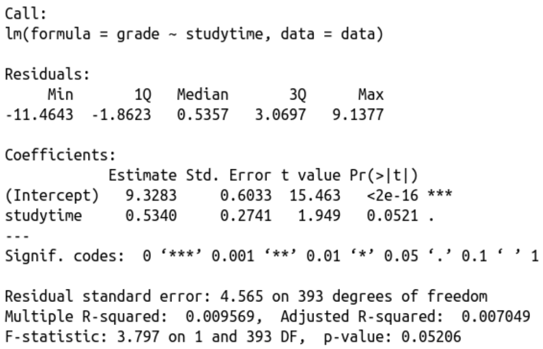
Statisticians are often interested in estimating the coefficients β_0 and β_1 from data. This is called **linear regression.** We can perform linear regression using the "lm" function in R. For example, if we want to let x denote the hours spent studying, then we can perform a linear model to our data by entering

```
model_hours <- lm(grade ~ studytime,data=data)
```

In this code, we are telling R that we think grades is a linear function of studytime (i.e., grade ~studytime), and we are letting R know that these variable are coming from the "data" dataframe. We can summarize the results of the linear model by entering:

```
summary(model_hours)
```

which prints



```
Call:
lm(formula = grade ~ studytime, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.4643  -1.8623   0.5357   3.0697   9.1377 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.3283     0.6033  15.463   <2e-16 ***
studytime     0.5340     0.2741   1.949   0.0521 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.565 on 393 degrees of freedom
Multiple R-squared:  0.009569,  Adjusted R-squared:  0.007049 
F-statistic: 3.797 on 1 and 393 DF,  p-value: 0.05206
```

Here, the (Intercept) Estimate value of 9.3283 corresponds to β_0 and the studytime Estimate value of 0.5340 corresponds to β_1. Based on these values of β_0 and β_1, how would you expect a student who studied for 1 hour to perform in comparison to a student who studied for 4 hours?

5. Perform the same analysis but let x denote the number of absences. Do you think grades increase or decrease with increasing absences?

```
model_absences <- lm(grade ~ absences,data=data)
summary(model_absences)
```

6. Furthermore, we can also assume that grades depend on both absences and hours spent studying as follows:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

   where y denotes a student's grade, x_1 denotes the number of absences, and x_2 denotes the number of hours spent studying. We can perform this linear model by

   ```
   model_absences_hours <- lm(grade ~ absences + studytime,data=data)
   summary(model_absences_hours)
   ```

7. An important application of linear regression is using a linear model to predict other outcomes. For example, if we fit a linear model to some students, then we may be able to predict how other students do. Let's see if we can predict how the females in our class do by first fitting the linear model to the males in the class. Begin by splitting the data in male and female groups

   ```
   males = subset(data,gender=='M')
   females = subset(data,gender=='F')
   ```

   We will now use the male data to find how grades depend on the number of hours studying

   ```
   model_hours_males <- lm(grade ~ studytime,data=males)
   summary(model_hours_males)
   ```

   We can now plot this against the female data as follows

   ```
   hours <- data.frame(studytime=seq(1,4,.1)) #useful for plotting
   predict_model = predict(model_hours_males,hours) #predict on hours

   plot(females$studytime,females$grade,xlab="Hours studying",ylab="Grade",main
       = "Predicting female grades")
   lines(hours$studytime,predict_model)
   ```