Lab 6: Linear Regression

In this lab, we will continue our examination of the FCAT data, and examine some possible relationships between the data using regression tools. Write up all these results and email us your pdf report and any suppport files. All files must be packaged in a tar, zip file or gzip file.

Assigned: Wed. March 5, 2014

Due: Fri. March 19, 2014

As a reminder, the FCAT dataset contains data from a reading experiment conducted in XXX in the state of Florida. The file contains 215 records of experiments performed on students.

```
ssrss03 = FCAT Reading Score
iiid___ = Study ID
iiage__  = Age
iige___ = Gender
gortcss = GORT Reading Comprehension
gortfss = GORT Fluency
orfwrcg = Oral Reading Fluency - AIMSWEB passages
Orfwrcf = Oral Reading Fluency - FCAT Passages
orfwrct = Oral Reading Fluency - Textbook Passages
tswessa = TOWRE Word Reading Efficiency task
tpdessa = TOWRE phonemic decoding task
tsum_ss = TOWRE Total Combined Score
rspatc  = Working Memory - Reading Span Task
lspatc = Working Memory - Listening Span Task
totalmq = Motivation to Read Questionnaire
wavoto2 = WASI (Weschler Abbreviated Scale of Intelligence) Vocabulary
wabdto2= WASI  Block Design Task
wasito2 = WASI Similarities
wamrto2 = WASI Matrix Reasoning
wafulIQ_  = WASI Full IQ
wapeIQ_ = WASI Performance IQ (combination of Block Design and Matrix Reasoning)
waveIQ_ = WASI Verbal IG (combination of Vocabulary and Similarities)
lc1sum = Listening comprehension task, passage 1
lc2sum = Listening comprehension task, passage 2
lc3sum = Listening comprehension task, passage 3
```

In the above list, to the left of the equal sign are the names of the variables that are found in the data file **FCAT_Mult_grade3.csv**. To the right of the equal sign are descriptions in the variable as I found them in the original SPSS file that was provided to me. We are interested in examining whether there is a gender difference in test results. We will concentrate on fluency (1), reading comprehension (2), and listening comprehension (3). To this end, we will model the data via linear regression, compute various correlations and examine the differences between male and female results as they relate to tasks (1) through (3). We will present our results in tabular and graphical format. One notes some interesting variables. Of course the variable of most interest is the overall FCAT reading score, which one seeks to maximize. The other variables are descriptive in nature, and serve to help better understand the origin (or influences) behind the scores.

Task 1 (10pts)

Create new variables **read_fluency** (the sum of the three columns related to "Oral Reading Fluency"),

list_compreh (the sum of the three listening comprehension variables). Also rename the variables **gortcss** and **gortfss** to **gort_read_compreh** and **gort_fluency**. Finally, replace **iige___** by **gender** and **ssrss03** by **score**.

Task 2 (5pts)

Remove the non-defined elements (NAs) in the scores (Hint: the functions is.na() or complete.cases() might be of use).

Task 3 (5 pts)

How many participants with no undefined data in the columns of interest are there in this study?

Task 4 (5 pts)

Compute the correlation between **score** and the following variables: **read_fluency**, **read_compreh**, and **list_compreh**. Make plots (scattergrams) of each pair of variables. There should be three plots, all on the same page.

Task 5 (10 pts)

Compute a linear regression model between **score** and the following variables: **read_fluency**, **read_compreh**, and **list_compreh**. For each regression line, create a plot with the scattergram, and the regression line superimposed. (Hint. The model line is **y = a + slope*x**. There is more than one way to plot it.) (Hint: one of these ways is to use **abline**)

Task 6 (10pts)

Create a single plot with two regression lines: one for **read_compreh** and one for **list_compreh**. Use the same approach as in task 5. Each regression line should have a different color. Compare these regression lines to one another. Is it possible to draw any conclusions? If so, what are they?

Task 7 (15pts)

Create your own function called **plot.my.data(df)** with a single argument, which is a dataframe. Execution of this function should generate the results of Task 6. Try it out and produce the results. The code and the results should appear in your report.

Task 8 (15pts)

Create the plots in Task 7 for women and men separately, in order to find out whether the two genders show any significant differences during the FCAT. The plots from Task 7 and those from this task should be combined on the same page.

Task 9 (10pts)

Compute the sum of the residual squares for your linear fit, both for the male case and for the female case using **read_compreh** and **list_compreh**. (This implies that you have four cases to consider.) What is the mean of the residual vector in both cases? What is the standard deviation of the residual vector in both cases? Only consider score versus fluency. (Hint: if **m = lm(...)** is the results of the linear model, the residuals are obtains with **m$residuals**.

Task 10 (15pts)

Set up a Hypothesis test where H0 is that the residual vectors for males and females have the same variance. The alternative Hypothesis H1 is that the variances are different. Note also, that we are comparing variances and not means, so using **t.test()** is not appopriate. Choose a confidence level of **p==0.1** (usually, $p=0.05$). Compare with the standard confidence level of 0.05. The residual vectors should be based on a linear regression of scores versus reading fluency, reading comprehension, and versus listening comprehension. The significance tests should be done with the functions **var.test()** and the function **fligner.test()**. Use the help facility in R to find out about these two functions and explain the differenes between them based on the documentation.