

Lab 5

This time we're going to examine some data from the example midterm in more detail, learn some more R, learn about transforming variables, and take a first step towards multiple regression. You can work alone or with a partner, but I recommend you sit next to someone so that you can discuss any results or problems.

Write-ups are due Friday 13th.

1. Here are some R commands to get you started with reading in the data and getting familiar with it. (You should always try to look at some helpful graphs when analysing a new data set).

```
# Read in the data: as a data frame, and have a look
# at it

scores_read.table("http://www.stat.washington.edu/stephens/S423/scores.txt",
                  header=T)

scores
names(scores)
dim(scores)
plot(scores$Midterm,scores$Final)

#note that you don't have to type out all of the name
#(only enough letters to identify it uniquely)

plot(scores$M,scores$F)

#note we can refer to specific lines of scores
scores[1,]
scores[3,]
scores[c(1,3),]

#What do these lines do?
didFinal_ scores$F>0
didFinal
scores_scores[didFinal,]
plot(scores$M,scores$F)
```

Now try fitting a linear regression to predict Final scores from Midterm scores. When your data are in a data frame like this you can do this as follows:

```
scores.lm _ lm(Final ~ Midterm, data = scores)
```

(but this time you do need to type the whole names).

Overlay the fitted line on your scatterplot of the points, and comment on the strength of the evidence in the data for dependence between midterm score and Final score. Do you think that this dependence is likely to be *causal*? Plot a scatterplot of the residuals against the x values, and plot a histogram of the residuals. What do you notice? Can you think of an explanation for this?

2. When the response variable ranges from 0 to 1, it is often helpful to *transform* it, to remove the problem we are having here. Eg let us call the response p (so $0 \leq p \leq 1$), and set $Y = \log(p/(1 - p))$. What is the range of possible values of Y ? Try fitting a linear regression to the transformed response variable, and show plots for the residuals against X . Does this model appear to predict final score better or worse than the previous model? (You may need to think a bit to answer this part. Explain how you get your answer.)
3. Try regressing the students *overall score* on i) their midterm score, and ii) their final score. You don't need to hand in any plots for this. Which seems to predict their overall score more precisely? (Explain how you obtained your answer to this, using appropriate excerpts from your results.) Can you think of any reasons this might be?
4. Now try fitting a (multiple) linear regression to predict overall scores from both Midterm scores *and* final scores simultaneously, as follow:

```
scores.lm _ lm(Overall ~ Midterm + Final, data = scores)
```

Write down the model that is being fitted here. Does this multiple regression appear to predict Overall score better than either of the simple regressions? (Explain how you get your answer to this.)