

## Lab 6

This time we're going to look at a data set that records various measurements relating to Poverty, Employment, Crime etc. in the 50 states plus DC. We'll also start to look at multiple linear regressions of the form

$$y_i = a + b_1x_{1i} + b_2x_{2i} + e_i$$

etc.

You need not hand in anything for 1 and 2 below.

1. Read the data file into R using

```
statesdata = read.table
('http://www.stat.washington.edu/stephens/S423/50states.txt',header=T)
```

2. 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
# the next line prints the data for the first 10 states
statesdata[1:10,]
names(statesdata)

#hopefully you can guess what kinds of thing most of these are
#measuring. Eg Urban measures the percentage of peopl living in cities;
#Drs measures the number of doctors per some fixed number of people;
#Poverty is the proportion of people living under some
#specified poverty line etc.

attach(statesdata)
plot(Crime,Prison)

#which do you think is the 'outlier?'
plot(Crime,Prison,type='n')
text(Crime,Prison,as.character(State))
```

3. Is the relationship between Crimes and Prison what you would expect? Try some other scatterplots, to see how different pairs of variables relate to one another. Note that you can do all at once using

```
pairs(statesdata)
```

(Include a copy of the plot produced by “pairs” in the work you hand in.)

4. Find 4 variables that appear to correlated with Poverty. Try fitting four linear regression, regressing Poverty with each in turn. Are the slopes all significantly different from 0? Look at residual plots to see if they appear to satisfy the usual assumptions OK. Are there any problems with outliers or high leverage points in any of the regressions? If so, can you correct the problem? Which of the four variables seems to be the best predictor of Poverty? Does this suggest a causal effect?
5. Now instead of regressing Poverty on the four variables separately, try regressing it on all four variables together, using

```
Poverty.lm ~ lm(Poverty ~ variable1 + variable2 + variable3 + variable4)
```

and look at the results using `summary(Poverty.lm)`. Try plotting the fitted values against the true values - do you see what you would expect? What about the usual residual plots? Are all the variables helpful in predicting Poverty? Are there any that were useful in the linear regression case, but are less useful in the multiple regression case? If so, why? Does the multiple regression predict better than any of the linear regressions? Are you sure? Give reasons for your answer.