Homework Assignment 1

Max H. Farrell – Chicago Booth
BUS41100 Applied Regression Analysis

Due at the beginning of class of week 2

1 Sample statistics and the regression coefficients

(a) If the sample variance for X is 1, the sample variance for Y is 2, and the sample correlation is 0.7, what is the slope of the least squares line?

(b) If the sample means for X and Y are 0 and 2 respectively, what is the intercept of this line?

2 Data Visualization Matters!

It is important to get a full and clear visualization of the data at hand. Summary statistics alone are not enough, and sometimes one way of looking at the data does not show you key features. We will now learn these lessons twice.

First, we will look at some box plots. The data set `boxplots.csv` on the course website contains data on two variables, `Y` and `X`.

(a) What type of data do we have in this case? That is, what type of variables are `Y` and `X`?

(b) Compute the *median* of `Y` for each value of `X`. What pattern do you see? (You may find the `by()` command useful in R, such as `by(Y, X, median)`.)

(c) Create a box plot of `Y` by `X`. What do you see? Remember the idea of the conditional distribution from class. In this case, what information does `X` have to offer about `Y`? Think about the center *and* the spread of the data.

(d) Compute the *mean* of `Y` for each value of `X`. What pattern do you see? Compare your findings to part **(b)**. Repeat for the *variance* of `Y` given `X`

(e) Create histograms of `Y` for each category of `X`. How does this inform what you have found above?

3 Data Visualization Still Matters!

Now, let us turn from box plots to scatter plots. The file `scatterplots.csv` on the course site contains 4 pairs of `x` and `y` variables

(a) **(i)** Create a scatter plot of `y1` against `x1`. What does this plot tell you about the function $\mathbb{E}[\texttt{y1} \mid \texttt{x1}]$? Remember intuitively what the function $\mathbb{E}[\texttt{y1}|\texttt{x1}]$ represents: the average of `y1` for a fixed value of `x1`. It's part of the *conditional distribution*, which contains all the information `x1` has about `y1`. So looking at this plot, if I tell you `x1`, what can you tell me about `y1`? Does knowing a particular value for `x1` tell you anything about what `y1` is likely to be?

(ii) Compute the correlation of `y1` and `x1`. Compute the least squares fit (the intercept and slope). Add the linear regression line of `y1` on `x1` to the scatter plot. Explain in words what the slope of the line is telling you in this case.

(b) **(i)** Create a scatter plot of `y2` against `x2`. For this pair, what does this plot tell you about the function $\mathbb{E}[\texttt{y2} \mid \texttt{x2}]$? Looking at this plot, if I tell you `x2`, what can you tell me about `y2`? Does knowing a particular value for `x2` tell you anything about what `y2` is likely to be?

(ii) Compute the least squares fit (the intercept and slope). Add the linear regression line of `y2` on `x2` to the scatter plot. Explain in words what the slope of the line is telling you in this case.

(iii) Compare this line, numerically and intuitively, to the one you found in part **(a)(ii)**.

(c) **(i)** Create a scatter plot of `y3` against `x3`. For this pair, what does this plot tell you about the function $\mathbb{E}[\texttt{y3} \mid \texttt{x3}]$? Looking at this plot, if I tell you `x3`, what can you tell me about `y3`? Does knowing a particular value for `x3` tell you anything about what `y3` is likely to be?

(ii) Compute the least squares fit (the intercept and slope). Add the linear regression line of `y3` on `x3` to the scatter plot. Explain in words what the slope of the line is telling you in this case.

(iii) Compare this line, numerically and intuitively, to the one you found in parts **(a)(ii)** and **(b)(ii)**.

(iv) How does this plot illustrate the difference between the conditional distribution of `y3` (given `x3`) and its marginal distribution?

(d) **(i)** Create a scatter plot of `y4` against `x4` and compute the least squares fit (the intercept and slope). .... I think you get the picture.[1]

4 Market Model Example

The CAPM (Capital Asset Pricing Model) relates asset returns to market returns through a simple linear regression model. Here we will model individual company returns as a function of the S&P500 index returns. This model assumes the rate of return R^s on a generic stock is linearly related to the rate of return (R^m) on the overall stock market as:

$$R_i^s = \alpha + \beta R_i^m + \epsilon_i$$

where the error term ϵ follows the assumptions of the SLR Model. The slope coefficient measures the sensitivity of the stock's rate of return to changes in the level of the overall market, and the intercept is market independent income. (The CAPM is discussed also in lecture 2.)

For this problem, use the file `mktmodel.csv` from the course website. The dataset contains 60 monthly returns (from 1992 to 1996) of the S&P500 and 30 individual US stocks (labelled by ticker).

(a) Use the code below to plot the return time series for the S&P and for each individual equity. Comment on what you see.

```
mkt <- read.csv("mktmodel.csv")
SP500 <- mkt$SP500
stocks <- mkt[,-1]
```

[1] Get it, the "picture"? I am very, very funny; tell your friends.

```
plot(SP500, col=0, ## Just get the plot up
     xlab = "Month", ylab = "Returns",
     main = "Monthly returns for 1992-1996",
     ylim=range(unlist(mkt)))
colors <- rainbow(30)  ## 30 different colors
## this is how you do 'loops' in R... this is useful!
for(i in 1:30){ lines(stocks[,i], col=colors[i], lty=2) }
lines(SP500, lwd=2)
```

(i) Calculate the market correlation for each stock. Based on this information alone, which CAPM fit would yield the highest R^2? Can you give a practical reasoning for this?

(ii) Estimate α and β for each stock and plot them against each other. Describe the results.

Coding hints:

- *Subset data with square brackets:* `mkt[3,4]` *gives the third row, fourth column,* `mkt[,1]` *gives the entire first column. You can also use names:* `stocks[,"GE"]` *gives the entire column named "GE".*
- *Coefficients from a regression can be extracted:*
  ```
  > GE.reg <- lm(stocks[,"GE"] ~ SP500)
  > GE.reg$coefficients
  ```
- *Run a bunch of regressions at once:* `mreg <- lm(as.matrix(stocks) ~ SP500)`

(b) *Pairs Trading* is a strategy which picks two stocks that generally move together and attempts to make money through arbitrage on differences within the pair. For example, if two stocks have the same market sensitivity (β), you could sell \$100 of the stock with low α (say α_{low}) and buy \$100 of the stock with high α (say α_{high}).

Suppose this is your trading strategy:

(i) Show that your average return is $\alpha_{\text{high}} - \alpha_{\text{low}}$. Do you lose money if the market goes down?

(ii) Based on the regressions you ran above, choose a pair of stocks for trading according to this strategy. Which would you buy and which would you sell?

(iii) Calculate what you would have made executing this strategy over the time span of our dataset. What is your average monthly return? How does this compare to the difference in alphas?

5 Teacher Salary Exploratory Analysis

The `teach.csv` data contains information on `salary` (in 1971 £ Sterling) for $n = 90$ teachers in the United Kingdom, along with the following characteristics of the teachers and the schools they work in: number of `months` of service (minus 12); `sex` (M/F); `marry` indicating (TRUE/FALSE) whether the female teachers were married or not[2]; type of `degree` offered to graduates ($\{0, 1, 2, 3\}$, with 3 being the "highest" type of degree); `type` of school (A/B); whether or not the teacher had special `training` (TRUE/FALSE); and `brk`, indicating whether or not the teacher had a break in service for two or more years (TRUE/FALSE).

You are going to explore how these variables affect teacher pay.

[2] Male teachers are coded as "single" (FALSE) whether they were married or not; apologies.

(a) Make a plot of `salary` versus the number of `months` in service using color, or otherwise, to indicate the sex of each teacher on the plot. Comment on what you see, and why the original article published using with this data may have been called "Sex differentials in teachers' pay" (Turnbull & Williams; JRSSA 1974).

(b) Now, ignore `months` and produce six sets of boxplots, one set for each other factor (`sex`, `marry`, `degree`, `type`, `train`, and `break`), showing the conditional distribution of `salary` for each level of each factor. Which seems to have the strongest effect on teacher salary?

(c) Using color, or otherwise, plot `salary` versus `months` in service [similar to **(a)**] with indications for the levels your chosen factor [from **(b)**] for each teacher. How does this new plot compare with the plot from **(a)**, and what do you conclude based on this new evidence?

Coding tip: careful with plotting and FALSE/TRUE variables. R *does not treat FALSE and TRUE like a factor variable. For factor variables, the alphabetically first level will map to '1', the second to '2', etc, but* R *maps FALSE to 0 and TRUE to 1. This makes sense, because if I take* `mean(teach$marry)` *I get the percent married, as I should. But if you use* `col=teach$marry` *in your plotting, you will get* **colors** *0 and 1, which are white and black. That means you'll plot white dots, which you don't want. A simple fix is to use* `col=(teach$marry+1)`*, giving 1 and 2.*

(d) Reconsider the questions in **(c)** through regression. That is, run two regressions: `salary` on your chosen factor and then again on `sex`.

(i) Explicitly write down the regression model you are fitting in each case.

(ii) How do you interpret the slope coefficient in the regression on `sex`?

(iii) Do you think that these factors make a meaningful difference in teacher's pay? What is your evidence?

(iv) Compare your results to the boxplots in **(b)**.

(v) (Looking ahead a little.) Run your first multiple linear regression (MLR) by regressing `salary` on `sex` and your chosen factor. What do you learn from the slope coefficients in this regression? How does this compare with what the two separate regressions indicated?

If you chose `marry` *in* **(b)***, the* R *code for the MLR would be*

```
> my.first.mlr <- lm(salary ~ sex + marry, data=teach)
> summary(my.first.mlr)
```

(e) Now, consider only the portion of the data corresponding to teachers whose school offers a `degree` of type "0":

```
> teach0 <- teach[teach$degree == 0,]
```

Investigate the effect of `months` of service on `salary` in this subset of the data. Calculate the correlaton between `months` and `salary` and use this to fit the regression line `salary` $= b_0 + b_1$`months`$+ e$. What does b_1 tell you about the influence of `months`? How would you predict the starting `salary` for teachers in schools which offer `degree` "0"?

(f) Consider the results from your regression in **(e)**. Plot the data (subset) and regression line. Plot the residuals both as a histogram and against `months`. Comment on any problems you see.