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 $y_1$ against $x_1$. What does this plot tell you about the function $E[y_1 \mid x_1]$? Remember intuitively what the function $E[y_1|x_1]$ represents: the average of $y_1$ for a fixed value of $x_1$. It’s part of the conditional distribution, which contains all the information $x_1$ has about $y_1$. So looking at this plot, if I tell you $x_1$, what can you tell me about $y_1$? Does knowing a particular value for $x_1$ tell you anything about what $y_1$ is likely to be?
(ii) Compute the correlation of $y_1$ and $x_1$. Compute the least squares fit (the intercept and slope). Add the linear regression line of $y_1$ on $x_1$ 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 $y_2$ against $x_2$. For this pair, what does this plot tell you about the function $E[y_2 | x_2]$? Looking at this plot, if I tell you $x_2$, what can you tell me about $y_2$? Does knowing a particular value for $x_2$ tell you anything about what $y_2$ is likely to be?

(ii) Compute the least squares fit (the intercept and slope). Add the linear regression line of $y_2$ on $x_2$ 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 $y_3$ against $x_3$. For this pair, what does this plot tell you about the function $E[y_3 | x_3]$? Looking at this plot, if I tell you $x_3$, what can you tell me about $y_3$? Does knowing a particular value for $x_3$ tell you anything about what $y_3$ is likely to be?

(ii) Compute the least squares fit (the intercept and slope). Add the linear regression line of $y_3$ on $x_3$ 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 $y_3$ (given $x_3$) and its marginal distribution?

(d) (i) Create a scatter plot of $y_4$ against $x_4$ and compute the least squares fit (the intercept and slope). . . . . I think you get the picture.

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:

$$R_s = \alpha + \beta R_m + \epsilon_i$$

where the error term $\epsilon$ 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.

```r
mkt <- read.csv("mktmodel.csv")
SP500 <- mkt$SP500
```

Get it, the “picture”? I am very, very funny; tell your friends.
```r
 stocks <- mkt[, -1]
 plot(SP500, col = 0, ## Just get the plot up
     xlab = "Month", ylab = "Returns",
     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 $\alpha$ and $\beta$ 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:
  ```r
  > 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 ($\beta$), you could sell $100 of the stock with low $\alpha$ (say $\alpha_{low}$) and buy $100 of the stock with high $\alpha$ (say $\alpha_{high}$).

Suppose this is your trading strategy:

(i) Show that your average return is $\alpha_{high} - \alpha_{low}$. 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; **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).

---

2Male teachers are coded as “single” (FALSE) whether they were married or not; apologies.
You are going to explore how these variables affect teacher pay.

(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?

(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)
      > 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 correlation between months and salary and use this to fit the regression line salary = \( b_0 + b_1 \text{months} + \epsilon \). 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.

6 Topic: Time Series Data

This question illustrates conceptual material, and thus it has extra exposition.

In class so far (in fact all the way until week 9) we have assumed that all the observations are independent of one another. In this question, you will get a first taste of dependent data, in the form of a time series. The overall message, which will be repeated ad nauseam in week 9, is that our same
regression tools will often (but not always!) apply just fine. Many students choose course projects
that involve time series data, so it is good to introduce some ideas now, instead of waiting all the way
until week 9.

The data set furniture.csv on the class website contains data from 1992–2001 on monthly furniture
sales (in millions of dollars). If you didn’t know any context this would look like an ordinary data
set (rows, columns, numbers, etc), but two things make it different. First, the order matters. These
rows are in time order, and have time labels. No data set that we’ve see so far has had an order. This
is what it means for the observations to be independent.

Second, there’s only one “real” variable: sales. The other columns are just the time. How can you
run a regression with just one variable? Or, put a different way, how are we going to predict sales?
This gives a big hint as to what we have to be careful about here: time dependence. We don’t care
(at least, not right now), how something like prices or advertising affect sales. What we care about,
and are going to try to capture, is how sales itself evolves over time.

(a) Create a variable called time that simply counts the months starting with the first, consecutively
to the end (i.e. January, 1992 = 1; February, 1992 = 1; . . . , December, 2001 = 120).

Plot the time series data for monthly furniture sales. This is just our usual plot with the outcome,
sales, on the y axis and time on the x axis, but now time has extra meaning. Comment on
what pattern(s) you see in the context of linear regression.

(b) Run a regression of sales against time, and add the regression line to the plot you made in
part (a). That is, fit the linear model

\[ \text{sales}_i = \beta_0 + \beta_1 \times \text{time}_i + \varepsilon_i. \]

Notice that this is just a plain simple linear regression, exactly like we had from the very
beginning in week 1!

(c) Let’s turn to prediction. Remember that the idea behind prediction is to try to form a good
guess for an outcome that you have not seen, Y_f, based on the data you have and a newly
observed X_f. (In our favorite example: Y_f is the price of a house that has not yet sold, X_f
would be the square footage, which can measure right now.) But here we don’t have an X
variable, just time. What we really want to do, then, is try to predict a future outcome.

Pretend you did not have the data for 2001. Re-run the regression above using only the data up
to, and including, December 2000. Use this regression to predict furniture sales for each month

We have ignored two major issues here. First, all the variance formulas we used in class relied on
independence (see the derivation handout), and so without independence, we can’t trust our inference.
That’s why we haven’t looked at any intervals or tests, and focused only on prediction. The second
thing we ignored was if there were any other time dependence patterns that we could capture. Is
there any seasonality in the data? A Christmas bump in sales? These questions we will try to answer
later in the course.