Statistics 101 Summer I 2011
Homework #5

Adam Kapelner, Instructor

Due 9AM, Wednesday, June 29, 2011

Instructions and Philosophy

There is not enough time to allow the new material to seep in before the final. Therefore in this assignment I'm going to require you to review lecture notes 17-20. In addition, there will be some material I cover on Tuesday that I cannot find in the book as a standalone section *i.e.* the section about $\hat{y} \sim \mathcal{N}(\cdot, \cdot)$. My exhortation to review the notes goes double for that material.

In Stine & Foster, read chapter 6 and 19. You can ignore the section on "properties of residuals" and the section on r^2 which are both concepts you will study in detail in Stat 102. I gave the last problem as practice and is *not required*. You should review the notes and the relevant parts of ch. 18 on 2-prop and 2-samp hypothesis testing before you do it. I would also recommend to read Wikipedia's treatment on regression to the mean which will be useful for the problem about Galton's data.

Once again, green means *easy*, yellow means *intermediate*, red means *difficult*, and purple means *extra credit*. This homework is worth 100 points and will be **graded on completeness**. Late homeworks are **not accepted**. This is for your benefit as you should use Wednesday to study for the final. As usual, 15 points are given as a bonus if the homework is typed using LaTeX (please delete all extraneous text).

Dependent Random Variables This section will cover the concepts of covariance and dependence.

Problem 1

We will prove some basic facts about covariances.

(a) Use the definition of covariance and variance to show that the covariance between any r.v. X and *itself* is just $\mathbb{V}\text{ar}[X]$.

(b) If X and Y are two independent r.v.'s, show that $\mathbb{C}\text{ov}[X, Y]$ is zero. This is in your notes.

(c) If X and Y are two independent r.v.'s, where $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y$, $\mathbb{V}\text{ar}[X] = \sigma_X^2$, $\mathbb{V}\text{ar}[Y] = \sigma_Y^2$, show that:

$$\mathbb{V}\text{ar}\left[XY\right] = \sigma_X^2\sigma_Y^2 + \mu_X^2\sigma_Y^2 + \mu_Y^2\sigma_X^2$$

Hint: this can be done in three lines if you the fact that you proved in homework #2 problem 2 part (g).

(d) X and Y are r.v.'s and a, b, c, d are numerical constants. Use the definition of covariance to show that the covariance between $aX + b$ and $cY + d$ is $(ac)\mathbb{C}\text{ov}\left[X, Y\right]$. If you are having trouble with this question, leave it until you finish the rest of the problem set.

Problem 2

Consider two spinners that people gamble on:

$$X_1, X_2 \overset{iid}{\sim} \begin{cases} \$0 & \text{w.p. } \frac{1}{2} \\ -\$2 & \text{w.p. } \frac{2}{5} \\ \$4 & \text{w.p. } \frac{1}{10} \end{cases}$$

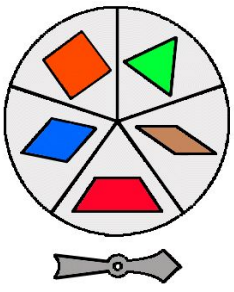
(a) Calculate the expectation of X_1 and X_2.

(b) Calculate the standard deviation of X_1 and X_2.

(c) Explain in English why S, the r.v. that is equal to the sum of the two spinners, and X_1 are dependent. Then, prove it mathematically.

(d) Calculate the covariance between S and X_1. Indicate units. I would recommend using the table method we did in class. If this is taking too long, do the rest of the problem set first and loop back to this one.

(e) Use the result from the previous question(s) to find the correlation between S and X_1.



Regression This section will cover the concepts of scatterplots, best fit lines, Simpson's paradox, and regression to the mean.

Problem 3

The goal of this is to learn how to build least squared lines from scratch, make predictions, calculate probabilities of responses, and to be able to leverage the power of JMP. Download `phila_housing.JMP` and load it up. This data set contains information related to house sales in Philadelphia and it has many variables. We are particularly interested in building a model that predicts house selling price (in dollars) using crime rate.

(a) What data types are the variables "selling price" and "crime rate"?

(b) Define "response variable" and "explanatory variable" and explain why they are sometimes called "dependent variable" and "independent variable" respectively.

(c) Is this an observational study or an experiment? Explain your answer.

(d) In the model we wish to create, what is the response variable (Y) and what is the explanatory variable (X)?

(e) Create a scatterplot using "Fit Y by X" for "selling price" and "crime rate". Analyze this scatterplot using the four criteria we went over in class (see bottom of p.108 for a refresher).

(f) Would you think it's appropriate to fit a line to this plot?

(g) Now, let JMP compute the univariate and bivariate statistics for you by clicking the red down arrow, then "density ellipse" and click the 0.95 option. Write down $\bar{x}, \bar{y}, s_x, s_y, r$.

(h) One of the rows was excluded from the data set (in JMP it appears with a cancel icon in the row header). Why is this row an outlier? To answer this question, un-exclude the row and create a new scatterplot, then explain what you see.

(i) Without using JMP, compute the best fit line's intercept (b_0) and slope (b_1) using the formulas from class and the univariate and bivariate statistics you wrote down previously.

(j) Use JMP to fit a least squares line by clicking the red down arrow and selecting the option "Fit Line". A line should appear as well as a bunch of information below the plot. Under the table header "Linear Fit", write down the equation. Is it the same as the one you came up with in the previous example? Print out this JMP analysis window and attach it to your homework.

(k) If the crime rate is 40, what is your best guess as to the sale price of the house?

(l) If the crime rate is 0, what is your best guess as to the sale price of the house?

(m) If the crime rate is 98.4, what is your best guess as to the sale price of the house?

(n) Is there anything wrong with making the predictions in the previous two questions?

(o) Under the table header "Summary of Fit", write down the RMSE.

(p) What is the probability of a house being worth more than \$200K if the crime rate is 30?

(q) What is the probability of a house being worth less than \$200K if the crime rate is 35?

(r) Consider the inverse problem where crime rate is viewed as a function of selling price. Use JMP to fit a least squares line for that data. Write down the equation. Now solve for selling price as a function of crime rate using that equation. Why is your equation different from the equation you got before?

(s) Solve the following equation for b_1:

$$\frac{\partial}{\partial b_1}\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] = 0$$

to prove that the slope estimate on the least squared line is: $b_1 = r\frac{s_y}{s_x}$. I showed you how to do this in class but I skipped over the messy details.

Problem 4 Now that you're a star in using JMP to create least squared regression lines, we will do another example. Download `doctor_salary.JMP` and load it up. This dataset contains information about doctors and their salary. We are particulary interested in whether or not there is an effect of the number of publications on their salary.

(a) Create a best fit line for salary in 2005 explained by publications per year using the JMP methods you learned in the last problem. Write down the prediction equation.

(b) Would you say there's an association between publications per year and salary in 2005? Does publishing less *cause* the doctor to make more money?

(c) Now we want to do the same analysis as before for each department the doctor is in. In JMP you would do this by doing a "Fit Y by X" where Y is salary in 2005, X is publications per year, and you would add Dept as the "By" variable. You should get 6 different scatterplots: one for each department. Fit a line to each of them. Do you see the negative association anymore? Is this an example of Simpson's paradox in action?

(d) Why was there a negative association in part (a)? Explain it by using what you learned in part (c).

Problem 5 The method of creating a least squared line was named after the discovery of regression to the mean by Sir Francis Galton in the late 1800's. We will analyze a portion of Galton's original parent-child height dataset using `R`. Download the file `galton.RData` from webcafe and load it into `R` by using "file...Load Workspace". None of the code in the question will work without this since the file contains the dataset. The two variables are "parent" and "child". The parent variable is an average of the two parents' heights in inches (the mother's height was multiplied by 1.08 to adjust for gender differences). The child variable is the height of a child. The number of rows is $n = 928$.

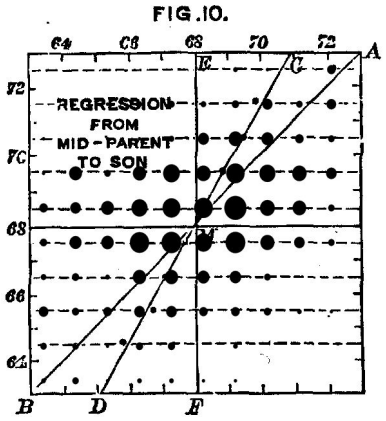


Figure 1: A table from Galton's original paper

(a) Explain in your own words what "regression toward the mean" means.

(b) Run the following `R` code:

```
plot(galton$child,
  galton$parent,
  xlim = c(60, 75),
  ylim = c(60, 75),
  ylab = "Child Height (in)",
  xlab = "Parent Height avg and adj (in)",
  main = "Child Height vs Parent Height");
#end of code placeholder
```

This will create a scatterplot. Use the criteria we discussed in class to describe the association in this plot.

(c) Explain what regression toward the mean would mean in this example. Do this before you do the next question.

(d) Run the following `R` code:

```
mod = lm(galton$child ~ galton$parent);
mod$coefficients;
#end of code placeholder
```

From the output of above, write down the least squares regression line with the values substituted for b_0 and b_1, the intercept and slope.

(e) Now run the following `R` code:

```
abline(mod, col = "red");
abline(a = 0, b = 1, col = "blue");
#end of code placeholder
```

This will draw in red the least squares line on the plot and in blue the 45^o line. This was one of the ways Galton originally studied regression to the mean. Do you see what he was surprised about? Explain.

More Hypothesis Testing This will be good practice for two-sample t-tests. This section is *not required to hand in* but is strongly recommended.

Problem 6

Amtrak maintains FRA Class 8 track on the NE corridor from Philadelphia to New York as well as other places. Class 8 track means it allows for top speeds of 125MPH on straightaways and 80MPH on turns. In order to run trains that quickly, the lines must maintain 10,000V AC which is converted to 6,500V DC to power the engines.

Sometimes voltages vary between stations. There are many reasons for such variation which are complicated. 12 samples were taken at New York Penn Station with a sample average of 9932V and a sample standard deviation of 1232V. 10 samples were taken at 30th Street Station in Philadelphia with an average of 9982V with a standard deviation of 705V.

(a) We are interested if there is a difference between the mean voltage at New York Penn station and at Philadelphia 30th Street Station. State hypotheses for the appropriate test.

(b) Are there any problems with running this test?

(c) Regardless of whether or not you thought there were problems, Run this test at $\alpha = 5\%$ disregarding the problems, if they exist.

(d) In order to do a 2-sample t-test, you must first check to see if the standard deviations should be pooled or left unpooled. Make sure you did this in the previous part. You should have used the unpooled protocol. However, the ratio of the sample variances was very near 3. Run the test again at the same α level using the pooled. Did your answer change much?