POL SCI 231b: Problem Set 3

Spring 2017

University of California, Berkeley

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

Due Monday, February 13 by 9:00 a.m.

Each group should turn in its problem set solution by email to Natalia Garbiras-Díaz (nataliagarbirasdiaz+231b@berkeley.edu) by Monday at 9 AM. Please work out the problems on your own, before you meet with your group to agree on solutions.

Questions:

1. Show that if $z_i = a + bx_i$ for all i, then $\bar{z} = a + b\bar{x}$. How does this relate to the fact that the regression line passes through the point of averages?

2. Show that adding a constant does not change the variance: if $z_i = x_i + d$ for all i, then $\text{var}(z) = \text{var}(x)$. (Hint: what is $z_i - \bar{z}$, expressed in terms of x_i, $\bar{x}$, and d?).

3. (a) Let $z_i = cx_i + d$ for all i. Show that $\text{Var}(z) = c^2\text{Var}(x)$. (Hints: use 1 and 2 to rewrite $z_i - \bar{z}$. Then, use the definition of variance, substitute for $z_i - \bar{z}$, and multiply out terms. You will also need to use the alternate definition of variance presented in class).

   (b) Now, recall that the equation for the regression line is $\hat{y}_i = a + bx_i$. So what is $\text{Var}(\hat{y}_i)$?

4. Let the variable z_i be x_i in standard units:

$$z_i = \frac{(x_i - \bar{x})}{\text{SD}_x}, \tag{1}$$

   where SD_x is the standard deviation of x. Show that

   (a) its average is 0;

(b) its variance and standard deviation are equal to 1.

5. Let Y_i be an outcome variable in an experiment, and let $D_i = 1$ if unit i is assigned to treatment and 0 otherwise. Suppose m out of N units are assigned to treatment. Thus, a fraction $\overline{D} = \frac{\sum_{i=1}^{N} D_i}{N} = \frac{m}{N}$ are assigned to treatment. The equation for the regression "line" is

$$\hat{Y}_i = a + bD_i. \tag{2}$$

   (a) Find the regression fit for a and b. (That is, use the algebra of variance and covariances to show what a linear regression returns as the intercept a and the slope b). Show your work.

   (b) Interpret b. What is another term for this quantity?

6. Consider the 2x2 matrix,

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

   (a) What is the determinant of this matrix?

   (b) Find the inverse of the matrix; or, if you cannot find the inverse, say why not.

   (c) How many linearly independent rows does this matrix have? How many linearly independent columns?

   (d) What is the rank of the matrix?

7. Let X be an $n \times 2$ matrix, where the first column is all 1's and the second column is $(x_1, x_2, ..., x_{n-1}, x_n)'$. Let Y be an $n \times 1$ column vector consisting of $(y_1, y_2, ..., y_{n-1}, y_n)'$.

   (a) What is the size of $X'X$? Of $X'Y$? What about $(X'X)^{-1}$ and $(X'X)^{-1}(X'Y)$? Can you multiply X and Y? Why or why not?

   (b) Find $X'X$. (That is, write out $X'X$, with typical elements given by n, $\sum_{i=1}^{n}(x_i)^2$, and so on).

   (c) Find $(X'X)^{-1}$.

   (d) Find $(X'Y)$.

   (e) Find $(X'X)^{-1}X'Y$.

   (f) Show that the $(2, 1)$ element of $(X'X)^{-1}X'Y = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$. That is, when there is a constant and one variable in an $n \times 2$ design matrix, the matrix representation reduces to the usual formula for the slope coefficient of the bivariate regression line.

   (g) Show that the $(1, 1)$ element of $(X'X)^{-1}X'Y = \bar{y} - b\bar{x}$, where $b = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$.

8. Hooke's law states that when a load (weight) is placed on a spring, the length is proportional to the weight. That is,

$$\text{length under load} = \text{length under no load} + \text{constant} \cdot \text{load}$$

Physicists test this prediction in the lab and obtain the results depicted in Table 1. Do not use *R*, other than as a calculator, to answer the following questions:

Load (kg)	Length (cm)
0	287.12
1	287.18
1	287.16
3	287.25
4	287.33
4	287.35
6	287.40
12	287.75

Table 1: A test of Hooke's Law.

(a) Find the regression equation for predicting length from load. (Show your work, including calculations of the relevant variances and covariances!).

(b) Use your result in (a) to write out a set of eight equations, where each equation is

$$Y_i = a + bX_i + e_i \tag{3}$$

for $i = 1, ..., 8$. Here, Y_i is the length of the spring (in cm); X_i is the load or weight (in kg), and e_i is the difference between the actual value of Y_i and the value on the regression line, given X_i. (In each of your equations, use actual numbers in place of Y_i, a, b, X_i and e_i). What units are the e_i measured in?

(c) Show that $\bar{e} = 0$; or, if it is not, say why not.

(d) Let X be the 8×2 matrix with typical element $[1 \;\; X_i]$, and e be the 8×1 vector of residuals. Show that $e'X = 0$, a $[1 \times 2]$ row vector; or if it is not, say why not.

(e) The two measured lengths for a load of 1 kg differ. Why might that be?

(f) Use the equation to predict length at the following loads: 2 kg, 3 kg, 5 kg, 105 kg.

(g) For a load of 3 kg, the answer to (f) is different from the number in the table. Under the load of 3 kg, would you use the number in the table, or the regression equation? Explain carefully. (You may want to refer to Chapter 12 of FPP).

(h) Do you think the regression equation provides a good basis for predicting the result of a hypothetical intervention, in which you hang a load on the spring? Why or why not?

9. Hibbs (1978) is interested in changes in industrial strike activity in advanced capitalist democracies in the twentieth century. His hypothesis is that changes in political economy after the Second World War—in particular, the rise of Left-Labor governments and the welfare state—shifts the locus of conflict away from private firms. Table 2 gives the approximate

numerical values for the change in strike activity and the change in Left-Labor cabinet representation, post war vs. inter-war means. (The values are approximate because they are eyeballed from a scatterplot in Hibbs 1978).

(a) Find the correlation between change in average strike volume and percentage change in Left-Labor cabinet representation, using the data in Table 3. What is the average and standard deviation of each variable? (Do this "by hand," not using R other than as a calculator).

(b) Find the regression equation for predicting change in average strike volume from percentage change in Left-Labor cabinet representation. (Do not use R; show your work, including calculations of the relevant variances and covariances!).

(c) Find the R^2 of the regression. Note that these data are based on differences of averages (the mean change in strike volume and the mean change in Left-Labor cabinet representation). What does this do to the R^2, for example relative to a situation in which we measure strikes and cabinet representation at yearly intervals?

(d) Now use R to create a plot showing Change in Left-Labor percentage cabinet representation on the horizontal axis and change in average strike volume, post-war mean minus inter-war mean, on the vertical axis. Label the axes, and label each point in the scatter plot with the name of the country. Superimpose the regression line you found in (b) on the plot. Create a caption describing the figure. Turn in your figure and your code.

(e) In general, do you think the regression equation provides a good basis for predicting the result of a hypothetical intervention to change Left-Labor cabinet representation? Why or why not? What are some similarities and differences between this setting and your discussion of Hooke's Law in the previous question?

Country	Change in Left-Labor cabinet representation (%)	Change in average strike volume
Norway	76	-1,980
Sweden	69	-1,700
U.K.	30	-1,000
Denmark	31	-650
Netherlands	19	-650
Belgium	10	-180
Italy	11	10
France	9	180
Finland	11	210
Canada	0	175
U.S.	1	190

Table 2: Change in average strike volume and average Socialist-Labour and Communist percentage of cabinet representation, inter-war to post-war period (Hibbs 1978)

10. Write an R function and code to perform the following simulation: for $d = b - 50$, where b is the value of the slope you found in part (b), calculate for each country the residual $f_i = Y_i - a - dX_i$. (For a, use the value that ensures the line $a + dX_i$ goes through the point of averages). Calculate the sum of squared residuals across all countries i and the R^2. Now increment d by 0.5 and repeat, again calculating the sum of squared residuals as well as the R^2. Continue incrementing d and repeating until you reach $d = b + 50$. Plot the sum of squared residuals and the R^2, both as a function of d (include two separate plots). What do these show?

11. Write a function that takes a vector for Y and a matrix with a number of independent variables and calculates multiple regression coefficients. Use the $family.rda$ data in the *problem_3_data.Rdata* file to show your function works (regress weight on height and bmi).