# POL SCI 231b: Problem Set 4

Spring 2017

University of California, Berkeley

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

Due Monday, February 20 by 11:59 pm.

Each group should turn in its problem set solution by email to Natalia Garbiras-Díaz (nataliagarbirasdiaz+231b@berkeley.edu) by Monday at 11:59 PM. Please work out the problems on your own, before you meet with your group to agree on solutions. Please remember to submit **compiled versions of your code**.

Before completing this problem set, you should work Exercise Sets A and B in Freedman (2009, Chapter 4). You should also read and think about the discussion questions.

1. Suppose that you have an $n \times 1$ column vector of observations on a variable $Y$ and an $n \times 1$ column vector of observations on a variable $X$. You convert each variable to standard units. Then you run a regression of standardized $Y$ on standardized $X$.

   (a) Show that the fitted slope coefficient on standardized $X$ is $r$, the coefficient of correlation between $Y$ and $X$.

   (b) What assumptions of the linear regression model did you need for part (a)?

2. Consider the linear regression model, $Y = X\beta + \epsilon$. The design matrix $X$ is $n \times p$; the first column of $X$ is all 1s. Make a list of the assumptions of the model.

3. Now, suppose we fit the regression model in the previous question to data, giving $Y_i = X_i\hat{\beta} + e_i$. Here, $\hat{\beta} = (X'X)^{-1}X'Y$, and $e_i = Y_i - X_i\hat{\beta}$, where $X_i$ is the $i$th row of $X$. Answer the following questions:

   (a) True or false: $\epsilon \perp X$

(b) True or false: $\epsilon \perp\!\!\!\perp X$

(c) True or false: $e \perp\!\!\!\perp X$

(d) True or false: $e \perp X$

(e) Does $e \perp X$ help validate $\epsilon \perp\!\!\!\perp X$?

(f) Does $\bar{e} = 0$ help validate $E(\epsilon_i) = 0$?

(g) Is $\sum_{i=1}^{n} \epsilon_i = 0$? Or is the sum typically around $\sigma\sqrt{n}$ in size?

4. (Interaction terms). Scholars sometimes counsel the use of interaction models for testing conditional hypotheses. This exercise gives an example. Assume the following OLS model:

$$Y_i = a + bX_i + cZ_i + dX_iZ_i + \epsilon_i \tag{1}$$

for all $i = 1, ..., n$. Here, $X_iZ_i$ is the product of $X_i$ and $Z_i$; this is usually called an "interaction term." The usual OLS assumptions apply, e.g., $\epsilon_i \perp\!\!\!\perp (X_i, Z_i, X_iZ_i)$ and the $\epsilon_i$ are i.i.d. with $E(\epsilon) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$ for all $i$.

(a) What is the marginal effect of intervening to change $X_i$ with $Z_i$ held fixed? That is, if

$$E(Y_i|X_i, Z_i) = a + bX_i + cZ_i + dX_iZ_i, \tag{2}$$

what is

$$\frac{\partial E(Y_i|X_i, Z_i)}{\partial X_i}? \tag{3}$$

And what is the marginal effect of intervening to change $Z_i$ with $X_i$ held fixed?

(Note: here, the "marginal effect" means, the change in the expected value of $Y_i$ due to a one-unit change in $X_i$ or $Z_i$).

(b) Now, suppose you run a regression of $Y_i$ on a constant, $X_i$, $Z_i$, and $X_iZ_i$, obtaining

$$\hat{Y}_i = \hat{a} + \hat{b}X_i + \hat{c}Z_i + \hat{d}X_iZ_i \tag{4}$$

for all $i$. Here, $\hat{Y}_i$ is the fitted ("predicted") value of $Y_i$, $\hat{a}$ is the fitted intercept, and $\hat{b}$, $\hat{c}$, and $\hat{d}$ are the fitted slope coefficients. Under the assumptions of the model in (1), $\hat{a}$ estimates $a$, $\hat{b}$ estimates $b$, and so on.

What is the estimated marginal effect of intervening to change $X_i$ with $Z_i$ held fixed? How about the estimated marginal effect of intervening to change $Z_i$ with $X_i$ held fixed?

(c) Let $\frac{\partial \hat{Y}_i}{\partial X_i}$ be the estimated marginal effect of intervening to change $X_i$ with $Z_i$ held fixed. Express the variance of $\frac{\partial \hat{Y}}{\partial X_i}$ in terms of variances and covariances of the variables and coefficient estimators in equation (4). (For convenience, treat $X_i$ and $Z_i$ as fixed, rather than as random variables).

(d) Suppose that after fitting equation (4), you find that

$$\hat{Y}_i = 1.2 + 2.3X_i + 0.5Z_i - 2.1X_iZ_i. \tag{5}$$

Moreover, the estimated variance-covariance matrix of $\hat{\beta} = (\hat{a}\ \hat{b}\ \hat{c}\ \hat{d})'$ is given by

$$\widehat{\text{cov}}(\hat{\beta}|X, Z, XZ) = \begin{pmatrix} 0.5 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0.9 & 0.5 & -0.3 \\ 0.4 & 0.5 & 0.4 & 0.3 \\ 0.7 & -0.3 & 0.3 & 0.7 \end{pmatrix} \tag{6}$$

Here, $n$ is large, so the sampling distribution of $\hat{\beta}$ is approximately normal.

i. Conduct a test of the null hypothesis that $b = 0$.
ii. Conduct a test of the null hypothesis that $d = 0$.
iii. Conduct a test of the null hypothesis that the marginal effect of intervening to change $X_i$, with $Z_i$ held fixed at $Z_i = 1$, is zero.
iv. Conduct a test of the null hypothesis that the marginal effect of intervening to change $X_i$, with $Z_i$ held fixed at $Z_i = 10$, is zero.
v. Comment on your results. Do they suggest that—given the model—the effect of $X_i$ is conditional on $Z_i$?

5. (Function). Extend the OLS function you wrote for problem set 3 to report the estimated standard errors of $\widehat{\beta}$ and the corresponding $t$ statistic and $p$-values from a two-sided test. Include a test comparing the results of your function to using the lm() command.

6. (Simulation). Sample 100 units from a normal distribution with mean 2 and variance 4. Call this vector x1, which will be fixed (will not vary across simulations). You will conduct a simulation with 10,000 replicates. For each replicate:

- Sample 100 realizations of i.i.d. $\epsilon$ from a normal distribution with mean 0 and sd = $\sqrt{2}$ and use that to construct Y as follows: $Y_i = 1.5 + 3x_{1i} + \epsilon_i$.
- Using your own regression function, fit a regression of $Y$ on $X_{100x2}$, where each row of $X$ is $[1 x_{1i}]$ for $i = 1, \ldots, 100$.
- Save $\widehat{\beta}$, $\widehat{\text{SE}}(\widehat{\beta})$, and the $t$-statistic.

Now,

(a) Plot the distribution of $\widehat{\beta}$ across the 10,000 replicates. Is $\widehat{\beta}$ unbiased? Justify your answer.

(b) Calculate the standard deviation of the distribution in (a). How does this s.d. compare to the average of the $\widehat{\text{SE}}(\widehat{\beta})$s across the 10,000 replicates? What does your answer indicate? And what is the theoretical standard error of $\widehat{\beta}$ in this simulation (i.e. based on $\sigma^2[X'X]^{-1}$)?

(c) Plot the distribution of the $t$-statistics. How close is the theoretical distribution of $t$ to normal? Superimpose a normal curve on the plot (with same mean as $t$).

7. **(From the 2015 midterm)** (The following is based on a real example, but some details are altered for purposes of the question). In November 1993, the state of Pennsylvania conducted elections for its state legislature. The result in the Senate election in the 2nd district (based in Philadelphia) was challenged in court. There, the Democratic candidate won 19,127 of the votes cast by voting machines on election day, while the Republican won 19,691 votes cast by voting machines, giving the Republican a lead of 564 votes. However, the Democrat won 1,396 absentee ballots, while the Republican won just 371 absentee ballots, more than offsetting the Republican lead based on the votes recorded by machines on election day. The Republican candidate sued, claiming that many of the absentee ballots were fraudulent. The judge in the case solicited analysis from an expert, who examined the relationship between absentee vote margins and machine vote margins in 21 previous Pennsylvania Senate elections in several districts in the Philadelphia area over the preceding decade. The analysis yielded a data set with the summary statistics depicted in Table 1. The disputed election is not included in the data set.

Table 1: Summary statistics

| Variable | Mean | Min | Max |
|---|---|---|---|
| Democratic Margin in Machine Balloting (percentage point lead among two-party votes) | 40.68 | -13.16 | 89.32 |
| Democratic Margin in Absentee Ballots (percentage point lead among two-party votes) | 29.06 | -34.67 | 72.97 |

The following characteristics of the data may be useful in answering the questions. Here, $X$ is a $21 \times 2$ matrix with 1's in the first column; the second column gives Democratic Margin in Machine Balloting in each election, as defined in Table 1. Also, $Y$ is a $21 \times 1$ column vector giving Democratic Margin in Absentee Ballots. We have:

$$X'X = \begin{pmatrix} 21 & 854.21 \\ 854.21 & 53530.38 \end{pmatrix}, \tag{7}$$

$$X'Y = \begin{pmatrix} 610.36 \\ 40597.39 \end{pmatrix}, \tag{8}$$

and

$$Y'Y = 35141.94 \tag{9}$$

(a) First, use $X'X$ and $X'Y$ to verify the means reported in Table 1. Explain your strategy.

(b) To conduct his analysis, the expert assumed the OLS regression model $Y_i = X_i\gamma + \epsilon_i$ with the usual assumptions. Here, $\gamma = [\alpha \ \beta]'$. What are the usual assumptions? Provide a substantive interpretation for each assumption in this context (i.e., say what they mean, given the details of the application).

(c) What are the ordinary least squares estimates of $\alpha$ and $\beta$? (Do this "by hand" using R or a calculator—do not use matrix manipulations in R such as `solve`. Show your work!). What assumptions of the model do you need to produce these estimates?

(d) Provide an interpretation of the estimated intercept $\hat{\alpha}$.

(e) What are the estimated standard errors of the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$? (Show your work!). What assumptions of the model do you need to justify these standard errors?

(f) Conduct a $t$-test to assess whether the estimate $\hat{\beta}$ is statistically significant. (Here, calculate the $t$ statistic "by hand" using your previous results, then refer the result to the appropriate $t$ distribution).

(g) Given the model and your estimates, what is the predicted Democratic margin in absentee ballots in the suspect election? What is the standard error of this prediction? Compare the predicted margin to the observed margin. Does it seem likely that something fishy is going on?

(h) In what ways is this problem like or unlike our example of Hooke's Law? Is this a prediction problem or a causal inference problem? Is a *response schedule* required to validate the analysis?

(i) Can you think of any reasons to question the validity of the expert's approach?

8. **(From the 2015 midterm)** In your view, which of these statements is closer to the truth?

   (a) Regression analysis can demonstrate causation;

   (b) Regression analysis assumes causation but can be used to estimate the size of a causal effect—if the assumptions of the regression models are correct.

   Pick one of these two statements and defend your choice in detail. Does your answer change, depending on whether we are analyzing experimental or observational data?