POL SCI: Problem Set 6

Spring 2017

University of California, Berkeley

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

Due Monday, March 20th, at 9:00 AM

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.

1. **Multicollinearity**. Let $Y_i = au_i + bv_i + \epsilon_i$ for $i = 1, ..., 100$. The ϵ_i are IID with mean 0 and variance 1. Here, u_i and v_i are fixed, not random. These two data variables have mean 0 and variance 1. The correlation between them is r. Let $M = [u\, v]$ denote the (partitioned) design matrix.

   (a) Show that the design matrix has rank 1 if $r = 1$ or $r = -1$.

   (b) Otherwise, let $(M'M)^{-1}M'Y = (\hat{a}\ \hat{b})'$ be the OLS estimator for a and b. Is the OLS estimator biased or unbiased?

   (c) Find the variance of $\hat{a}$; of $\hat{b}$; of $\hat{a} + \hat{b}$; and of $\hat{a} - \hat{b}$. What happens if $r = 0.99$?

   (d) What are the implications of multicollinearity for drawing inferences about a and b? What about their sum and their difference? What are the implications of exact collinearity? (Note: *exact collinearity* here means, $r = 1$ or $r = -1$; *multicollinearity* means $r \doteq 1$ or $r \doteq -1$).

   (e) True or false, and explain:

      i. Multicollinearity leads to bias in the OLS estimator.

      ii. Multicollinearity leads to bias in the estimated standard errors for the OLS estimates.

      iii. Multicollinearity leads to big standard errors for some estimates.

2. **fGLS vs. panel-corrected standard errors**. On p. 175, Freedman (2009) explains that White's method for estimating the SEs in OLS (what Beck and Katz (1995) call "panel-corrected standard errors" in the time-series cross-section context) "may have the same sort of problems as plug-in SEs, because estimated covariance matrices can be quite unstable." Explain why the estimated covariance matrix would be unstable in the settings discussed by Beck and Katz. How this would affect panel-correct standard errors? (Hint: look at p. 638 of Beck and Katz. Where does the covariance matrix of the errors appear in the formula for the covariance matrix of $\hat{\beta}$, and how is it estimated?).

3. **Measurement error**. In this exercise, you will build a plot like the one shown in lecture to depict the consequences of measurement error in dependent and independent variables, in the case of bivariate regression; use the code you wrote on the last problem set to build added variables plots, in the case of multivariate regressions with measurement error; and then run simulations to assess bias and mean squared error of estimators in the presence of measurement error, for the bivariate and multivariate case. (Note: you can write a function for the whole problem, but that is not required). Consider the model

$$Y^* = \beta X^* + \gamma W^* + \epsilon, \tag{1}$$

which conforms to the usual OLS assumptions. Here, Y^*, X^*, and W^* are true values of the respective variables. But each may be measured with error:

$$Y = Y^* + \delta, \tag{2}$$

$$X = X^* + \eta, \tag{3}$$

and

$$W = W^* + \nu, \tag{4}$$

where $E(\delta) = E(\eta) = 0 = E(\nu) = 0$ and the errors are independent of Y^*, X^*, W^*, ϵ, and each other, with variances σ_δ^2, σ_η^2, and σ_ν^2, respectively.

   (a) First, consider the case where $\gamma = 0$ and $\beta = 1$, so we are back to the bivariate model. Simulate 200 draws of X^*, ϵ, δ, and η, all distributed as $N(0, 1)$ random variables, all independent of each other, and use the draws to construct Y^*, Y, and X. Regress Y^* on X^*, Y on X^*, and Y^* on X. Add these three regression lines to a scatterplot of Y^* against X^*, and also add the line $Y^* = X^*$. What does the plot suggest?

   (b) Write a simulation to repeat (a) 20,000 times, and create a 3x2 table in R that has "no error", "error in Y", and "error in X" as the row labels and "bias" and "mse" as the column labels. (Here, MSE refers to mean-squared error; refer to the lecture for definitions). Fill in the table with the results of the simulation. What do you conclude about the effects of measurement error in the multivariate case, when independent variables are correlated?

   (c) Now suppose $\gamma = 2$ and $\beta = 1$. Simulate 200 draws of X^*, ϵ, δ, η, and ν, all distributed as $N(0, 1)$ random variables, all independent of each other. To create a correlation between the right-hand-side variables, generate $W^* = X^* + \psi$, where ψ

is $N(0,1)$. Now, use the draws to construct Y^*, Y, X, and W. Regress Y^* on X^* and W^*, Y on X^* and W^*, and Y^* on X and W. For each regression, construct added variable plots, using the code you wrote on your last problem set, where the (residuals of the) response variable are plotted against the residuals of one independent variable and then the other. (So there should be six plots – two for each of the three regressions). What do the plots suggest?

(d) Write a simulation to repeat (c) 20,000 times, and create a 6x2 table in R that has as the row labels "no error–est of beta", " "no error–est of beta", "error in Y–est of beta", "error in Y–est of gamma," and "error in X and W–est of beta," and "error in X and W–est of gamma" and as the column labels, "bias", "mse" and "average r." (Here, average r is the average correlation between X^* and W^* across the 20,000 simulations; it will be the same for each row of the table). What do you conclude about the effects of measurement error in the multivariate case, when the true values of the right-hand-side variables are correlated?

(e) Now modify the code you wrote for part (d) so that X^{star} and W^* are independent $N(0,1)$ random variables. What do you conclude about the effects of measurement error in the multivariate case, when the true values of the right-hand-side variables are generated as independent random variables?

4. **The Neyman model and regression**. Let $Y_i(1)$ denote the potential outcome if unit i is treated, and let $Y_i(0)$ denote the potential outcome for the same observation if it is not treated. The unit causal effect is the difference between $Y_i(1)$ and $Y_i(0)$. This causal effect may vary from one unit to the next. The random assignment of units to treatment ($X_i = 1$) and control ($X_i = 0$) (equivalently, the random sampling of units from the experimental study group into treatment and control groups) is the only random component in the modeling framework.

(a) Show that the potential outcomes model

$$Y_i = Y_i(0)(1 - X_i) + Y_i(1)X_i \tag{5}$$

may be expressed in the form of a regression model such that b represents the average causal effect of the treatment, and

$$Y_i = a + bX_i + u_i, \tag{6}$$

where the disturbance term $u_i \equiv Y_i(0) - \overline{Y(0)} + ((Y_i(1) - \overline{Y(1)}) - (Y_i(0) - \overline{Y(0)}))X_i$. Here, $\overline{Y(0)}$ is the average potential outcome under control for the study group, and $\overline{Y(1)}$ is the average potential outcome under treatment.

(b) Is it possible for the disturbance term u_i to be statistically independent of the independent variable? Explain your answer.

(c) Is it possible for the disturbance term u_i to be homoskedastic (i.e., have a constant variance denoted σ^2)? Explain your answer.

(d) In light of your answers to (b) and (c), is the OLS estimator of the model in equation (6) unbiased? Are the OLS standard errors (e.g. those calculated using $\hat{\sigma}^2[X'X]^{-1}$ and stored by R routines such as lm) correct? Explain your answers.

5. **Marginal effects plot.** Using the Miguel et al. data, plot the marginal effects of lagged rainfall as a function of land crop from the regression of growth (gdp_g) on a constant, lagged rainfall ($GPCP\_g\_l$), land crop ($land\_crop$) and a third term with the interaction of lagged rainfall and land crop. This is the plot shown in the lecture slides for week 6. You can write a function though this is not required. To build the plot, follow these steps:

   (a) Run the regression and store the coefficients for lagged rainfall and the interaction.

   (b) Find the variance for each of these coefficients and their covariance.

   (c) Produce a vector of simulated $land\_crop$ levels. For this, use seq() to get a sequence that goes from the minimum value of $land\_crop$ to its maximum by small increments.

   (d) Using (a), produce a vector with the marginal effects of lagged rainfall for each value of the vector created in (c).

   (e) Using (b), produce a vector with the variance of the marginal effects of lagged rainfall for each value of the vector created in (c). Use the variance to create two new vectors: one for the upper bound of the confidence interval and one for the lower bound.

   (f) Plot the marginal effects and the confidence interval lines. Use rug() to add a rug showing the distribution of the data.

   (g) Are the effects of rainfall on growth conditional on $land\_crop$? Comment. What needs to hold for you to answer in the affirmative?

6. A researcher is analyzing a data set with observations for two countries and twenty time periods (years). She fits the following model to this data set:

$$Y_{it} = \alpha_1 + \alpha_2 + \gamma Z_{it} + \epsilon_{it}. \tag{7}$$

   Here, Y_{it} is the value of the dependent variable for country $i = 1, 2$ in year $t = 1, \ldots, 20$, while Z_{it} is the key independent variable. The data set is stacked by country, so the first twenty observations are for country 1 and the second twenty are for country 2. The OLS fit for this model is $\hat{\beta} = (X'X)^{-1}X'Y$, where X is the design matrix and Y is the vector of observations on the dependent variable. Here, ϵ_{it} is a random error term, with $\epsilon \perp\!\!\!\perp X$, and α_1, α_2, and γ are parameters. The intercepts α_1 and α_2 for countries 1 and 2 are sometimes called "(country) fixed effects."

   (a) What is the size of Y and of X? What are the elements of X? (That is, describe each column of X). Under the model, what parameters does $\hat{\beta}$ estimate?

   (b) Explain in a few sentences a rationale for including α_1 and α_2 in a model like equation (7), when analyzing time-series cross-section data.

(c) Denote the vector of residuals after fitting equation (7) by $e = Y - X\hat{\beta}$, and let $\bar{e}_1 = \frac{1}{20}\sum_{t=1}^{20} e_{1t}$ be the average of the residuals for country 1. Is $\bar{e}_1 = 0$? Be specific in your answer (using algebra if appropriate).

(d) Suppose that the value of the error term for country i in year i is $\epsilon_{it} = \rho\epsilon_{i,t-1}$, where $\rho \in (0, 1)$. (This is called temporal "autocorrelation" of the errors; you can consider the initial value $\epsilon_{i,1}$ to be fixed for $i \in \{1, 2\}$, since we don't observe a realization of $\epsilon_{i,0}$). Is the OLS estimator $\hat{\beta}$ unbiased?

(e) (This continues the previous sub-question). What is the variance-covariance matrix of ϵ, given X? (Be specific: what are the elements of this matrix?) And what is the variance-covariance matrix of $\hat{\beta}$, given X?

(f) Propose two different ways of estimating the model (7), given temporal autocorrelation of the type discussed in parts (d) and (e). Discuss their advantages and disadvantages. Be as concrete as possible: e.g., in each case, give a formula for an estimator of the model parameters, and say how would you estimate the standard errors of those estimators.

(g) Suppose we included an overall intercept in the model, so that now $Y_{it} = \alpha + \alpha_1 + \alpha_2 + \gamma Z_{it} + \epsilon_{it}$. What is the new design matrix $\tilde{X}$? What are the consequences for the OLS fit $(\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$? Explain your answer.

7. Suppose that one seeks to estimate the parameter β in the time-series equation

$$Y_t = \alpha + \beta Y_{t-1} + u_t, \tag{8}$$

where the u_t are independent and identically distributed random error terms. Does OLS regression generate unbiased estimates of β? Explain your answer.

8. Write a simulation to show that the precision gain from using a difference in differences estimator is a function of the covariance of the potential outcomes and the pre-treatment covariate.