Gov 2002: Problem Set 9

Problem Set Instructions

This problem set is due on **Apr. 19th, 11:59 pm** Eastern time. Please upload a PDF of your solutions to Gradescope. We will accept hand-written solutions but we strongly advise you to typeset your answers in Rmarkdown. Please list the names of other students you worked with on this problem set.

Question 1 (30 points)

Often our data is collected with error, which we refer to as measurement error. For instance, for a dependent variable Y you're trying to measure in a survey, respondents may randomly mis-click, or they may systematically lie about having a socially undesirable trait. In this question, we will explore the impact of measurement error in regression analysis in the most favourable case where the measurement error is independent of the true values. Consider the linear projection:

$$L[Y \mid 1, X] = \beta_0 + \beta_1 X$$

with the projection error denoted as $e = Y - L[Y \mid 1, X]$, and $\mathbb{V}[X] = \sigma_X^2$. Unfortunately, we do not observe Y or X but instead noisy proxies for them $\{\tilde{Y}, \tilde{X}\}$, where

$$\tilde{Y} = Y + v, \quad \tilde{X} = X + w$$

Where v is one realization from $V \sim \mathcal{N}(0, \sigma_v^2)$ and w is one realization from $W \sim \mathcal{N}(0, \sigma_w^2)$, where W and V are independent of X and Y. This implies that $\text{Cov}(v, X) = \text{Cov}(v, e) = \text{Cov}(v, w) = \text{Cov}(w, X) = \text{Cov}(w, e) = \text{Cov}(w, v) = 0$. This is commonly referred to as classical measurement error.

(a) Consider the linear projection of these observable variables, $L[\tilde{Y} \mid 1, \tilde{X}] = \alpha_0 + \alpha_1 \tilde{X}$. Find α_1 in terms of $\{\beta_1, \sigma_w^2, \sigma_v^2, \sigma_X^2\}$. Hint: first derive an expression of the coefficients in terms of the $\tilde{X}$ and $\tilde{Y}$.

(b) From your expression, explain the effect of this type of measurement error in X on the sign and magnitude of the coefficient α_1 compared to β_1.

(c) From your expression, explain the effect of this type of measurement error in Y on the sign and magnitude of the coefficient α_1 compared to β_1.

Question 2 (30 points)

Public outrage about CEO pay finds its roots during the onset of the Great Recession. In this problem, we examine beliefs about CEO compensation. To begin, load `CEO.csv` from Ed. The data came from a survey of 632 Americans. They were asked questions on how much they thought CEOs do and should earn. (The variables are called `perceived` and `ideal`, respectively).

(a) To begin, produce a scatterplot with perceived CEO earnings among Americans on the x-axis and their ideal earnings on the y-axis. Estimate a least squares regression of ideal earnings on perceived earnings and report your results (coefficients and standard errors) in a neatly formatted table.

(b) A fellow researcher wants to know what units might be important for the fit of this regression. Identify the unit in the data with the largest leverage or hat value and describe where on the scatterplot that unit is. Calculate the change in the estimated slope from dropping that unit (you can either run another regression or use the closed-form expression given in lecture).

(c) We are going to walk through ways to check our regression assumptions. A common way of checking for non-linearity is to examine a plot in which the fitted values from a regression are plotted on the x-axis and the residuals from the regression are plotted on the y-axis (often called a *residuals versus fitted-values plot*). We can get the fitted values from a regression using the `fitted()` function and the residuals using the `residuals()` function. Intuitively, if linearity holds we would expect the residuals to be positive and negative in equal proportions at all points along the regression line. By plotting the residuals against the fitted values, we can see whether there are regions of the regression line where the residuals tend to be systematically positive or negative. Create a plot like this and interpret it. Based on this method, does there seem to be substantial non-linearity? If so, how would you correct it?

(d) Now we are going to introduce a second regressor. Estimate a linear model predicting ideal CEO salary using both `perceived` CEO salary and `age`. Report the coefficients and standard errors in a neatly formatted table.

(e) We can recreate the results from part (e) using bivariate regressions. To do so, follow these steps.

- Get residuals from a regression of `perceived` on `age`
- Plot `ideal` (on the y-axis) against the residuals from step 1 (on the x-axis).
- Calculate the regression line of `ideal` on the residuals from step 1 and add it to the plot.

What is the slope of the regression line that fits the residuals? How does this relate to the regression in part (e)?

Question 3 (40 points)

Now let's formalize the results we observed in 2(d) and 2(e). Consider the following multivariate regression:

$$Y = X\beta + Z\gamma + \epsilon$$

(a) Show that for any $\{X, Z\}$, we can decompose Z into $P_X Z + M_X Z$, where P_X and M_X are the projection matrix and annihilator matrix of X respectively (hint: use an add and subtract trick). Also show that P_X and M_X are orthogonal.

(b) Show that if $X \perp Z$, then the coefficients we get from regressing Y on X and Y on Z will be the same coefficients from the joint regression above. (Hint: there are different ways to show this, but one way is to show that the OLS minimization problem separates into two completely separate minimization problems. You could also derive an expression for the bivariate coefficient and then use the orthogonality condition.)

(c) Suppose $\hat{\beta}$ and $\hat{\gamma}$ are the OLS estimators for β and γ for the regression above. Find a $\hat{\beta}'$ such that:

$$\hat{Y} = X\hat{\beta}' + (M_X Z)\hat{\gamma}$$

Write $\hat{\beta}'$ in terms of $\hat{\beta}$ and $\hat{\gamma}$, and provide a substantive interpretation of $\hat{\beta}'$ in plain English (Hint: X and Z are not necessarily orthogonal anymore).

(d) Lastly, show that the following regression

$$M_X\hat{Y} = (M_X Z)\hat{\gamma}$$

will return the same OLS estimator $\hat{\gamma}$ as in the multivariate regression $Y = X\beta + Z\gamma + \epsilon$, explain this result in plain English.