Gov 2002: Problem Set 10

Problem Set Instructions

This problem set is due on **December 6, 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)

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, download `CEO.csv` from Ed and load it. 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. (**Hint:** Consider using `lm` and `stargazer`. The latter requires the `stargazer` package and you can install it by typing `install.packages("stargazer")`. (Of course, you are free to use other packages.) For running regression models and generating scatterplots, you may find the following slides helpful as well: **1. Regression**; **2. Scatterplot**.)

(b) 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?

(c) 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. Explain how the coefficients on `perceived` CEO salary changed after including the `age` variable.

Question 2 (30 points)

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**: Apply the definitions of the projection and annihilator matrices.) Also show that P_X and M_X are orthogonal. (**Hint**: Show that $P_X^T M_X = 0$.)

(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**: One way to work on this problem is to notice that $Y = X\beta + Z\gamma + \epsilon$, and thus $Cov(Y, X) = Cov(X\beta + Z\gamma + \epsilon, X)$.)

(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. Use results from part (a).).

(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.

Question 3 (40 points)

The standard output from OLS will give the standard errors for the estimated coefficients, but often we want to obtain measures of uncertainty for the predicted value of Y_i given some value of X_i (that is, the conditional expectation function). Using the example from lecture, we might be interested in the average wait times to vote for individuals making \$25,000, \$50,000, or \$100,000 in annual income, along with measures of uncertainty around those estimates. In this problem we will look at how to calculate interval estimates for these predicted values. Assume the following *true* population model for $Y_i | X_i$:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where the X_i are random variables and u_i are i.i.d. random variables with $E[u_i \mid X_i] = 0$ and $Var(u_i \mid X_i) = \sigma^2$. Suppose we observe a random sample of n paired observations $\{Y_i, X_i\}$. Assume the Gauss-Markov assumptions hold and that we have a large sample. Our goal is to estimate the predicted value at some value $X_i = x$:

$$\mu(x) = E[Y_i \mid X_i = x] = \beta_0 + \beta_1 x.$$

(a) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be OLS estimators of the regression of Y on X. Use what you know about the unbiasedness of OLS estimates to show that $\widehat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is an unbiased estimator of the population quantity $\mu(x) = E[Y_i \mid X_i = x]$.

(b) Find the conditional variance of $\hat{\beta}_0$, $Var(\hat{\beta}_0 \mid X_1, \ldots, X_n)$, using the following two facts. You answer should be in terms of σ^2 and functions of X_i.

$$Cov(\overline{Y}, \hat{\beta}_1 \mid X_1, \ldots, X_n) = 0 \quad \text{and} \quad Var(\hat{\beta}_1 \mid X_1, \ldots, X_n) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

(c) Find the covariance of the OLS estimates given our X values, $Cov(\hat{\beta}_0, \hat{\beta}_1 | X_1, \ldots, X_n)$, again in terms of σ^2 and functions of the X_i. (**Hint**: It's not zero.)

(d) Using what you found in (b) and (c), find the standard error of $\widehat{\mu}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

(e) Assume that we don't know σ^2 and instead construct our estimate of the standard error by plugging in for σ^2 our unbiased estimate s^2 using the residuals.

Give the formula for a large-sample 95% confidence interval estimator for $\mu(x) = E[Y \mid X = x]$ using what you found above and substituting s^2 for σ^2. How do we interpret this confidence interval?