Gov 2002: Problem Set 8

Problem Set Instructions

This problem set is due on April 12, 11:59 pm Eastern time. Please upload a PDF of your solutions to gradescope. We will accept hand-written solutions for problems 1-3 but we strongly advise you to typeset your answers in Rmarkdown. Problem 4 should be typeset. Please list the names of other students you worked with on this problem set.

Question 1

Let X and Y be random variables with finite variances, and let $W = Y - E(Y|X)$ be the CEF error. This is the population version of the sample residual: the difference between the true value of Y and the predicted value of Y via a conditional expectation function (CEF) involving X.

(a) Compute $E(W)$ and $E(W|X)$.

(b) Compute $Var(W)$, for the case that $W|X \sim \mathcal{N}(0, X^2)$ with $X \sim \mathcal{N}(0, 1)$.

(c) Now consider a third finite-variance random variable Z. Suppose the following CEF is true in the population:

$$E[Y|X,Z] = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 Z^2 + \beta XZ$$

Find the partial effects of X and Z on $E[Y|X,Z]$.

Question 2

In this problem we will explore how centering an independent (subtracting off the variable's mean) affects the interpretation of coefficients in linear projections.

(a) Suppose that $L[Y \mid 1, X] = \beta_0 + X\beta_1$. Let $Z = X - E[X]$. Find the coefficients of the linear projection $L[Y \mid 1, Z] = \alpha_0 + Z\alpha_1$ in terms of β_0 and β_1. Does centering X around its mean affect these parameters?

(b) Now suppose that $L[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$. Derive an expression for the partial effect of X_1, $\frac{\partial L[Y|X_1,X_2]}{\partial X_1}$, and the expectation of that partial effect (where the expectation is over the distribution of X_1 and X_2).

(c) A common trick with interactions is to center one of the variables for easier interpretation. Let $Z_2 = X_2 - \mu_2$, where $\mu_2 = E[X_2]$. Rewrite the linear projection $L[Y|X_1, X_2]$ as a function of Z_2 instead of X_2 and relate the new coefficient on X_1 to the linear projection in (b). That is, write

$$L[Y \mid X_1, X_2] = \alpha_0 + \alpha_1 X_1 + \alpha_2 Z_2 + \alpha_3 X_1 Z_2$$

and express the α coefficients in terms of $(\beta_0, \beta_1, \beta_2, \beta_3)$ and μ_2. How does the coefficient obtained in part (c) relate to the average of the partial effects in (b)? (Hint: you'll need to add and subtract certain values to obtain the new expression.)

(d) In a sentence or two, explain the substantive interpretation of α_1 and why using Z_2 instead of X_2 might be useful. (Hint: consider a case such where X_1 is assignment to some treatment and X_2 is birth year.) Does this transformation affect the interpretation of the interaction term?

Question 3

This question highlights the importance of the assumptions we make about the population regression function.

(a)

Suppose the following linear model is true in the population for some outcome variable Y:

$$Y = \mathbf{X}^T \boldsymbol{\beta} + u$$

Show that if this model is true and $E[u|\mathbf{X}] = 0$, then $E[Y|\mathbf{X}] = \mathbf{X}^T \boldsymbol{\beta}$.

(Note that this is the opposite of what we showed in lecture, where we saw that if we assume $E[Y|\mathbf{X}] = \mathbf{X}^T \boldsymbol{\beta}$, then the conditional mean zero assumption holds, $E[u|\mathbf{X}] = 0$.)

(b)

With regression we don't typically make many distributional assumptions about $\mathbf{X}$, except for a few crucial ones. In particular, we saw that for the linear projection to be well-defined we needed $\mathbf{Q_{XX}} = E[\mathbf{X}\mathbf{X}^T]$ to be positive definite and thus invertible.

Let $\mathbf{X} = (1, X)^T$ so we are in a bivariate regression setting. Show that if $Var(X) = 0$, then $\mathbf{Q_{XX}}$ is not positive definite. (Hint: look for linear dependencies in the columns of $\mathbf{Q_{XX}}$.)

Question 4: Regression Analysis of Subprime Loans

This problem will guide you through thinking about the conditional expectation function and how it relates to regression and how we can connect it back to hypothesis testing.

For this problem, we are going to use the subprime data. Recall that these are data collected by the U.S. government on all home lending transactions in Cape Coral and Fort Myers. They contain information on each loan applicant and give information on whether that applicant received a subprime loan (`high.rate`) as well as on the amount of the loan (`loan.amount`). They also contain basic demographic information such as race, gender, and income.

Assume the data represent the "truth" (i.e., an entire population). Also assume that the data in this population are distributed i.i.d. Take a sample of size 250, without replacement, from this population. Set your seed to `02138` before doing so. You will be working with this sample throughout this problem.

(a)

You care about the relationship between the variables `income` and `loan.amount` – seems like there should be a relationship between those two, right?

As per usual, you have a friend (you really need to get some new friends) who proposes that you use the following strategy to see if there is a relationship: Create a new income `income.bin` variable that takes on four values using the `cut()` function in `R`:

- a value for if income falls into the [0, 25] percentile range (which you can find via `quantile()`),
- a value if it falls into the (25, 50] range,
- a value for the (50, 75] range, and a value for the (75, 100] range.

Note that the lower bounds are NOT inclusive, except for the first range.

Run a regression of `loan.amount` on `income.bin`, and report the coefficients, standard errors, R^2 and sample size in a nicely formatted table (recall section).

(b)

Let's compare the approach in (a) to using a regression on the original continuous variable. What is an assumption we make in the approach in (a) that we don't make when we run a regression? We are looking for an assumption related to the fact that you have taken a continuous variable and stratified it into four categories.

(c)

Based on the results of the regression on binned income in (a), do you think the linearity assumption needed for a bivariate regression on the original continuous variable holds in this case? Why or why not?

(d)

In spite of your friend's opinion, you decide to run a regression of `loan.amount` on `income`. Run this regression within your sample, and report the coefficients, standard errors, R^2 and sample size in a nicely formatted table (recall section).

(e)

Interpret the estimated coefficients. Is the interpretation consistent with the results from (a)?