Gov 2002: Problem Set 9

Spring 2021

Problem Set Instructions

This problem set is due on April 22, 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: Measurement error

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:

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

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

$$\tilde{Y} = Y + v$$

$$\tilde{X} = X + w$$

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 $Cov(v, X) = Cov(v, e) = Cov(v, w) = Cov(w, X) = Cov(w, e) = Cov(w, v) = 0$. This is commonly referred to as **classical measurement error**.

(a) Consider the linear projection of these observable variables, $L[\tilde{Y}|1, \tilde{X}] = \alpha_0 + \alpha_1 \tilde{X}$. Find α_1 in terms of $\{\beta_1, \sigma_w^2, \sigma_v^2, \sigma_X^2\}$.

(b) What is the effect of the measurement error in X on the sign and magnitude of the coefficient α_1 compared to β_1?

(c) What is the effect of the measurement error in Y on the sign and magnitude of the coefficient α_1 compared to β_1?

Question 2: Omitted variables

Consider the linear projection

$$L[Y|1, X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

denoting the projection error again as e.

As in the previous problems, you only care about estimating the coefficient β_1 in this projection (the long regression). Unfortunately, your data consist only of a random sample on (Y, X_1). So the best you can do is estimate α_1 in the following projection (the short regression):

$$L[Y|1, X_1] = \alpha_0 + \alpha_1 X_1$$

(a)

Derive an expression for α_1 in terms of the model parameters $(\beta_0, \beta_1, \beta_2)$, $Var(X_1)$, and $Cov(X_1, X_2)$.

(b)

Derive an expression for the difference in X_1's coefficient between the short and long regressions, $\alpha_1 - \beta_1$.

(c)

Your results should imply that in order for α_1 to be the same as β_1, we need the omitted variable X_2 to either be (i) unrelated to the outcome ($\beta_2 = 0$) or (ii) unrelated to the explanatory variable of interest ($Cov(X_1, X_2) = 0$).

When a variable (X_2) is not available, it is common for applied researchers to make an educated guess about the sign of β_2 and $Cov(X_1, X_2)$ in order to know at least the sign of the bias $\alpha_1 - \beta_1$.

Suppose that for an individual:

- Y is voting for the conservative party in the upcoming election
- X_1 is years of education
- X_2 is distance from a city

Make a guess about the signs of β_2 and $Cov(X_1, X_2)$. Based on your guesses, will our short regression coefficient α_1 be biased upwards or downwards for the long regression coefficient β_1?

Question 3: Limitations of R-Squared

As mentioned in lecture, the R^2 of a linear regression mechanically increases as you add more covariates. You will formally show this in this exercise.

Consider two fitted least squares regressions (written here in matrix form):

$$\mathbf{Y} = \mathbf{X}_1\widehat{\alpha}_1 + \widehat{\mathbf{e}}_1$$

and

$$\mathbf{Y} = \mathbf{X}_1\widehat{\beta}_1 + \mathbf{X}_2\widehat{\beta}_2 + \widehat{\mathbf{e}}_2$$

Let R_1^2 and R_2^2 be the R-squared from the two regressions. Without making any additional assumptions, show that $R_2^2 \geq R_1^2$. Hint: consider the situation where $R_2^2 < R_1^2$ and think about the properties of OLS to arrive at a contradiction.

Is there a case when there is equality $R_1^2 = R_1^2$?

Question 4

Last week we used `lm()` to run a linear regression model in `R` on the `subprime` data. This week you are going to code your own function to generate OLS estimates and the corresponding R-squared value without using `lm()`.

(a) Write a generic function, that takes two arguments: a `formula` (such as `income ~ loan.amount`), and `data` (a data frame). You will need to take some pre-processing steps to construct the design matrix and outcome vector before you apply the OLS estimator. The following functions will be useful:

- `formula()`: Takes a input such as `income ~ loan.amount` and returns an object of class formula
- `model.matrix()`: Takes as inputs a formula and a data frame, and returns the corresponding design matrix (sometimes referred to as the model matrix), which will include a column of 1's.
- To get the outcome vector you will need two functions:
  - `model.frame()`: Takes as inputs a formula and a data frame
  - `model.response()`: Input is output from `model.frame()`

Your function should return a `list()` object, with two elements in the list. The first element should be the `coefficients`, and the second element should be `R.squared`. Do not use `lm()` in your function, we want you to code the estimates using `R`'s matrix operations.

(b) In this part of the question, you will test your function on the subprime data. Run the following regression model `income ~ loan.amount + black + woman` using your function from part (a) to `lm()` and compare the outputs.