Gov 2002: Problem Set 9

Problem Set Instructions

This problem set is due on **November 29, 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 in part (a), briefly explain (1-2 sentences) the effect of this type of measurement error in X on the sign and magnitude of the coefficient α_1 compared to β_1. Hint: what parameter controls the amount of measurement error in X?

(c) From your expression in part (a), briefly explain (1-2 sentences) the effect of this type of measurement error in Y on the sign and magnitude of the coefficient α_1 compared to β_1. Hint: what parameter controls the amount of measurement error in Y?

Question 2 (20 points)

Decide whether each of the following statements are true or false and explain your reasoning briefly.

(a) If $Y = X\beta + e$, $X \in \mathbb{R}$, and $E[e|X] = 0$, then $E[e] = 0$.

(b) If $Y = X\beta + e$, $X \in \mathbb{R}$, and $E[e|X] = 0$, then $E[X^3 e] = 0$.

(c) If $Y = X\beta + e$, $X \in \mathbb{R}$, and $E[Xe] = 0$, then $E[X^2 e] = 0$.

(d) If $Y = X\beta + e$, $X \in \mathbb{R}$, and $E[e|X] = 0$, then e and X are independent.

Question 3 (30 points)

In most linear regression models, the dependent variable Y is expressed as a function of independent variables $X_1, X_2, ..., X_k$ (or to use the vector notation, just X as a vector in $\mathbb{R}^k$). That is,

$$Y = X\beta + e$$

where β is a $k \times 1$ coefficient vector and e is the error.

(a) Explain briefly what it means for $g(X)$, a function of X, to be the best predictor of Y.

(b) Show that for $g(X)$ to be the best predictor, $g(X)$ must be equal to the conditional expectation function $E[Y|X]$. (**Hint:** You can assume that, for the CEF error $e = Y - E[Y|X]$, we have $E\Big|e\{E[Y|X] - g(X)\}\Big| < \infty$).

Now, for unifying definition, we also need to consider an *intercept-only* model, where there is no X and α is simply a constant:

$$Y = \alpha + e$$

(c) Find $\underset{\alpha}{\text{argmin}} E[(Y - \alpha)^2]$.

Question 4 (20 points)

Suppose you are interested in the association between income and race, gender, and education. The dependent variable Y is income in USD. The independent variables are age, gender, and education, where

- X_a = age (in years),
- $X_g = 1\{\text{gender} = \text{female}\}$, and
- $X_e = 1\{\text{college degree or higher}\}$.

You are especially interested in interactions among these variables, and run a linear regression model only with a three-way interaction term:

$$Y = \beta_0 + \beta_1 X_a X_g X_e$$

(a) Derive the marginal effects of X_a, X_g, and X_e.

Determine whether the following statements (b-d) are true or false and explain your reasoning briefly.

(b) Comparing (i) men whose age is 30 years and (ii) men whose age is 50 years, we find that (ii) men whose age is 50 years earn β_1 dollars more than (i) men whose age is 30 years.

(c) Comparing (i) women whose age is 30 years and who has no college degree and (ii) women whose age is 50 years and who has no college degree, we find that (ii) women whose age is 50 years and who has no college degree earn β_1 dollars more than (i) women whose age is 30 years and who has no college degree.

(d) Comparing (i) women whose age is 30 years and who has a college degree and (ii) women whose age is 50 years and who has a college degree, we find that (ii) women whose age is 50 years and who has a college degree earn $20 \times \beta_1$ dollars more than (i) women whose age is 30 years and who has a college degree.