Assignment 02

More Regression and Mathematics

This goal of this assignment is to review ideas from regression and mathematics that will be useful for the remainder of the course. Turn in a printed version of your responses to each of the questions on this assignment. Please adhere to the following guidelines for further formatting your assignment:

- All graphics should be set to an appropriate aspect ratio and sized so that they do not take up more room than necessary. They should also have an appropriate caption.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document.
- Syntax or computer output should not be included in your assignment unless it is specifically asked for.

This assignment is worth 10 points.

Use for Questions 1–3

Use the data in *canadian-prestige.csv* to answer the following questions.

1. The partial correlation between X_1 and Y, controlling for X_2 through X_k is defined as the simple correlation between the residuals $E_{Y|X_2,\ldots,X_k}$ and $E_{X_1|X_2,\ldots,X_k}$ (where $E_{Y|X_2,\ldots,X_k}$ are the residuals from regressing Y on the predictors X_2, X_3, …, and X_k). This partial correlation is denoted $r_{Y,1|2,\ldots,k}$. Use this idea to calculate the partial correlation between prestige and education, controlling for income and percentage of women.

2. An alternative procedure for calculating the partial regression coefficient estimate for β_1 (the partial regression coefficient for X_1 from the multiple regression of Y on $X_1, X_2, \ldots, X_k$) is to regress the residuals $E_{Y|X_2,\ldots,X_k}$ on $E_{X_1|X_2,\ldots,X_k}$. Use this idea to calculate the estimate of the partial regression coefficient for education based on the model that includes education, income, and percentage of women to predict variation in prestige.

3. In light of the procedures for computing the partial correlation (Question 1) and the partial regression coefficient (Question 2), explain why $r_{Y,1|2,\ldots,k} = 0$ only if $\beta_1 = 0$ (where β_1 is the partial regression coefficient for X_1 from the multiple regression of Y on $X_1, X_2, \ldots, X_k$).

Measurement Error

4. Derive Equations 6.12 (in Fox). To derive the first equation in 6.12 multiply Equation 6.11 by X_1. To derive the second equation, multiply Equation 6.11 by X_2. (*Hints:* Both X_1 and X_2 are uncorrelated with the regression error, ϵ. Likewise, X_2 is uncorrelated with the measurement error, δ.)

5. Show that the covariance of X_1 and δ is simply the measurement error variance, σ^2_δ, by multiplying $X_1 = \tau + \delta$ through by δ and taking expectations.

6. Show that the variance of $X_1 = \tau + \delta$ can be written as the sum of "true-score" variance, σ^2_τ and "measurement-error" variance, σ^2_δ. (*Hint:* Square both sides and take expectations.)

Use for Questions 7–9

Use the *duncan.csv* data to regress prestige on education and income (Model 1). Then, fit a series of multiple regression models that regresses prestige on income and a modified education variable. Create this modified education variable by adding random measurement errors to the education variable. Sample these errors from a normal distribution with mean of 0, repeating the exercise for each of the following measurement error variances: the following distributions: $\sigma^2_\delta = 100$ (Model 2); $\sigma^2_\delta = 625$ (Model 3); $\sigma^2_\delta = 2,500$ (Model 4); and $\sigma^2_\delta = 10,000$ (Model 5). In each case, re-regress prestige on income and the modified education variable.

7. Treat the initial multiple regression (Model 1) as corresponding to a model without measurement error ($\sigma^2_\delta = 0$). Create a plot of the estimates of the education coefficients as a function of the measurement error variances. Include this plot in your document.
8. Based on your plot, describe what happens to the education coefficient as measurement error increases.
9. Create and include a similar plot for the estimates of the income coefficient. Describe what happens to the income coefficient as measurement error in the education covariate increases. **(2pts)**