STA 2201 S 2011 Assignment 7

Closely related to the problem of measurement error in regression and perhaps even nastier is the problem of omitted variables in observational studies. In the following regression model, X_1 and Y are measured without error, but X_2, which has an impact on Y and is correlated with X_1, is not part of the data set. The true model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i,$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$. The independent variables are random, and for simplicity we'll make them normal. Let

$$\left[\begin{array}{c} X_{i,1} \\ X_{i,2} \end{array} \right] \sim N \left(\left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right], \left[\begin{array}{cc} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{array} \right] \right)$$

with ϵ_i independent of $X_{i,1}$ and $X_{i,2}$.

Since X_2 is not observed, it is swallowed up into the intercept and error term, as follows.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2 \mu_2) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} - \beta_2 \mu_2 + \epsilon_i) \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i'. \end{aligned}$$

The primes just denote a new β_0 and a new ϵ. And of course there could be more than one omitted variable. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

1. What is $Cov(X_{i,1}, \epsilon_i')$?
2. All we can observe are the pairs $(X_{i,1}, Y_i)$. Their distribution is bivariate normal. Calculate the mean and covariance matrix of $(X_{i,1}, Y_i)$ under the true model.
3. Are the parameters of the true model identifiable? Answer Yes or No and prove your answer. If the answer is No, all you have to do to prove it is produce two points in the parameter space that yield the same distribution of the observable data. A simple numerical example is fine.
4. Suppose we want to estimate β_1. Is the usual least squares estimator $\widehat{\beta}_1$ a consistent estimator of β_1 for all points in the parameter space under the true model? Answer Yes or no and show your work. Remember, X_2 is not available, so you are doing a regression with one independent variable.
5. Are there *any* points in the parameter space for which $\widehat{\beta}_1$ is a consistent estimator when the true model holds?
6. Do you think that there is the possibility of inflated Type I error rate here using the usual test F-test or t-test of $H_0 : \beta_1 = 0$? Briefly explain. No calculation is necessary here, just your opinion and some justification for it.
7. Do you think that there will be a problem in the *prediction* of Y from X_1 for a new set of data? Again I am only asking for your opinion. I don't actually know the answer to this one myself, though I have an opinion.