STA 431s17 Assignment Six[1]

This assignment is about problems caused by measurement error in regression, and some partial solutions. It covers lecture units 9 and 10. In the text, see pages 33-51 in Chapter 0. I think the treatment of identifiability is better in lecture than in the current version of the text.

1. This question explores the consequences of ignoring measurement error in the response variable. Independently for $i = 1, \ldots, n$, let

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
V_i &= Y_i + e_i,
\end{aligned}
$$

where $Var(X_i) = \phi$, $E(X_i) = \mu_x$, $Var(e_i) = \omega$, $Var(\epsilon_i) = \psi$, and X_i, e_i, ϵ_i are all independent. The explanatory variable X_i is observable, but the response variable Y_i is latent. Instead of Y_i, we can see V_i, which is Y_i plus a piece of random noise. Call this the *true model*.

   (a) Make a path diagram of the true model.

   (b) Strictly speaking, the distributions of X_i, e_i and ϵ_i are unknown parameters because they are unspecified. But suppose we are interested in identifying just the Greek-letter parameters. Does the true model pass the test of the Parameter Count Rule? Answer Yes or No and give the numbers.

   (c) Calculate the variance-covariance matrix of the observable variables as a function of the model parameters. Show your work.

   (d) Suppose that the analyst assumes that V_i is that same thing as Y_i, and fits the naive model $V_i = \beta_0 + \beta_1 X_i + \epsilon_i$, in which

$$
\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(V_i - \overline{V})}{\sum_{i=1}^n (X_i - \overline{X})^2}.
$$

   Assuming the *true* model (not the naive model), is $\widehat{\beta}_1$ a consistent estimator of β_1? Answer Yes or No and show your work.

   (e) Why does this prove that β_1 is identifiable?

2. This question explores the consequences of ignoring measurement error in the explanatory variable when there is only one explanatory variable. Independently for $i = 1, \ldots, n$, let

$$
\begin{aligned}
Y_i &= \beta X_i + \epsilon_i \\
W_i &= X_i + e_i
\end{aligned}
$$

where all random variables are normal with expected value zero, $Var(X_i) = \phi > 0$, $Var(\epsilon_i) = \psi > 0$, $Var(e_i) = \omega > 0$ and ϵ_i, e_i and X_i are all independent. The variables W_i and Y_i are observable, while X_i is latent. Error terms are never observable.

(a) What is the parameter vector $\boldsymbol{\theta}$ for this model?

(b) Denote the covariance matrix of the observable variables by $\boldsymbol{\Sigma} = [\sigma_{ij}]$. The unique σ_{ij} values are the moments, and there is a covariance structure equation for each one. Calculate the variance-covariance matrix $\boldsymbol{\Sigma}$ of the observable variables, expressed as a function of the model parameters. You now have the covariance structure equations.

(c) Does this model pass the test of the parameter count rule? Answer Yes or No and give the numbers.

(d) Are there any points in the parameter space where the parameter β is identifiable? Are there infinitely many, or just one point?

(e) The naive estimator of β is

$$\widehat{\beta}_n = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2}.$$

Is $\widehat{\beta}_n$ a consistent estimator of β? Why can you answer this question without doing any calculations?

(f) Go ahead and do the calculation. To what does $\widehat{\beta}_n$ converge?

(g) Are there any points in the parameter space for which $\widehat{\beta}_n$ converges to the right answer? Compare your answer to the set of points where β is identifiable.

(h) Suppose the reliability of W_i were known, or to be more realistic, suppose that a good estimate of the reliability were available; call it r^2_{wx}. How could you use r^2_{wx} to improve $\widehat{\beta}_n$? Give the formula for an improved estimator of β.

3. The improved version of $\widehat{\beta}_n$ in the last question is an example of *correction for attenuation* (weakening) caused by measurement error. Here is the version that applies to correlation. Independently for $i = 1, \ldots, n$, let

$$\begin{array}{lcl} D_{i,1} & = & F_{i,1} + e_{i,1} \\ D_{i,2} & = & F_{i,2} + e_{i,2} \end{array} \qquad cov\left(\begin{array}{c} F_{i,1} \\ F_{i,2} \end{array}\right) = \left(\begin{array}{cc} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{array}\right) \qquad cov\left(\begin{array}{c} e_{i,1} \\ e_{i,2} \end{array}\right) = \left(\begin{array}{cc} \omega_1 & 0 \\ 0 & \omega_2 \end{array}\right)$$

To make this concrete, it would be natural for psychologists to be interested in the correlation between intelligence and self-esteem, but what they want to know is the correlation between *true* intelligence and *true* self-esteem, not just the between score on an IQ test and score on a self-esteem questionnaire. So for subject i, let $F_{i,1}$ represent true intelligence and $F_{i,2}$ represent true self-esteem, while $D_{i,1}$ is the subject's score on an intelligence test and $D_{i,1}$ is score on a self-esteem questionnaire.

(a) Make a path diagram of this model.

(b) Show that $|Corr(D_{i,1}, D_{i,2})| \leq |Corr(F_{i,1}, F_{i,2})|$. That is, measurement error weakens (attenuates) the correlation.

(c) Suppose the reliability of $D_{i,1}$ is ρ_1^2 and the reliability of $D_{i,2}$ is ρ_2^2. How could you apply ρ_1^2 and ρ_2^2 to $Corr(D_{i,1}, D_{i,2})$, to obtain $Corr(F_{i,1}, F_{i,2})$?

(d) You obtain a sample correlation between IQ score and self-esteem score of $r = 0.25$, which is disappointingly low. From other data, the estimated reliability of the IQ test is $r_1^2 = 0.90$, and the estimated reliability of the self-esteem scale is $r_2^2 = 0.75$. Give an estimate of the correlation between true intelligence and true self-esteem. The answer is a number.

4. This is a simplified version of the situation where one is attempting to "control" for explanatory variables that are measured with error. People do this all the time, and it doesn't work. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ W_i &= X_{i,1} + e_i, \end{aligned}$$

where $V\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$, $V(\epsilon_i) = \psi$, $V(e_1) = \omega$, all the expected values are zero, and the error terms ϵ_i and e_i are independent of one another, and also independent of $X_{i,1}$ and $X_{i,2}$. The variable $X_{i,1}$ is latent, while the variables W_i, Y_i and $X_{i,2}$ are observable. What people usually do in situations like this is fit a model like $Y_i = \beta_1 W_i + \beta_2 X_{i,2} + \epsilon_i$, and test $H_0 : \beta_2 = 0$. That is, they ignore the measurement error in variables for which they are "controlling."

(a) Suppose $H_0 : \beta_2 = 0$ is true. Does the ordinary least squares estimator

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n W_i^2 \sum_{i=1}^n X_{i,2} Y_i - \sum_{i=1}^n W_i X_{i,2} \sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2 \sum_{i=1}^n X_{i,2}^2 - (\sum_{i=1}^n W_i X_{i,2})^2}$$

converge to the true value of $\beta_2 = 0$ as $n \to \infty$ everywhere in the parameter space? Answer Yes or No and show your work.

(b) Under what conditions (that is, for what values of other parameters) does $\widehat{\beta}_2 \stackrel{p}{\to} 0$ when $\beta_2 = 0$?

5. Finally we have a solution, though as usual there is a little twist. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta X_i + \epsilon_i \\ V_i &= Y_i + e_i \\ W_{i,1} &= X_i + e_{i,1} \\ W_{i,2} &= X_i + e_{i,2} \end{aligned}$$

where

- Y_i is a latent variable.
- V_i, $W_{i,1}$ and $W_{i,2}$ are all observable variables.
- X_i is a normally distributed *latent* variable with mean zero and variance $\phi > 0$.
- ϵ_i is normally distributed with mean zero and variance $\psi > 0$.
- e_i is normally distributed with mean zero and variance $\omega > 0$.

- $e_{i,1}$ is normally distributed with mean zero and variance $\omega_1 > 0$.
- $e_{i,2}$ is normally distributed with mean zero and variance $\omega_2 > 0$.
- X_i, ϵ_i, e_i, $e_{i,1}$ and $e_{i,2}$ are all independent of one another.

(a) Make a path diagram of this model.

(b) What is the parameter vector $\boldsymbol{\theta}$ for this model?

(c) Does the model pass the test of the Parameter Count Rule? Answer Yes or No and give the numbers.

(d) Calculate the variance-covariance matrix of the observable variables as a function of the model parameters. Show your work.

(e) Is the parameter vector identifiable at every point in the parameter space? Answer Yes or No and prove your answer.

(f) Some parameters are identifible, while others are not. Which ones are identifiable?

(g) If β (the paramter of main interest) is identifiable, propose a Method of Moments estimator for it and prove that your proposed estimator is consistent.

(h) Suppose the sample variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is

```
      W1    W2     V
W1 38.53 21.39 19.85
W2 21.39 35.50 19.00
V  19.85 19.00 28.81
```

Give a reasonable estimate of β. There is more than one right answer. The answer is a number. (Is this the Method of Moments estimate you proposed? It does not have to be.) **Circle your answer.**

(i) Describe how you could re-parameterize this model to make the parameters all identifiable, allowing you do maximum likelihood.