# STA 431s23 Assignment Six[1]

1. In a study of diet and health, suppose we want to know how much snack food each person eats, and we "measure" it by asking a question on a questionnaire. Surely there will be measurement error, and suppose it is of a simple additive nature. But we are pretty sure people under-report how much snack food they eat, so a model like $W = X + e$ with $E(e) = 0$ is hard to defend. Instead, let

$$W = \nu + X + e,$$

where $E(X) = \mu_x$, $E(e) = 0$, $Var(X) = \sigma^2_x$, $Var(e) = \sigma^2_e$, and $Cov(X, e) = 0$ The unknown constant $\nu$ could be called *measurement bias*. Calculate the reliability of $W$ for this model. Is it the same as the expression for reliability given in the text and lecture, or does $\nu \neq 0$ make a difference?

2. Continuing Question 1, suppose that two measurements of $W$ are available.

$$\begin{aligned} W_1 &= \nu_1 + X + e_1 \\ W_2 &= \nu_2 + X + e_2, \end{aligned}$$

where $E(X) = \mu_x$, $Var(X) = \sigma^2_x$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, and $X$, $e_1$ and $e_2$ are all independent. Calculate $Corr(W_1, W_2)$. Does this correlation still equal the reliability even when $\nu_1$ and $\nu_2$ are non-zero and potentially different from one another?

3. Let $X$ be a latent variable, $W = X + e_1$ be the usual measurement of $X$ with error, and $G = X + e_2$ be a measurement of $X$ that is deemed "gold standard," but of course it's not completely free of measurement error. It's better than $W$ in the sense that $0 < Var(e_2) < Var(e_1)$, but that's all you can really say. This is a realistic scenario, because nothing is perfect. Accordingly, let

$$\begin{aligned} W &= X + e_1 \\ G &= X + e_2, \end{aligned}$$

where $E(X) = \mu_x$, $Var(X) = \sigma^2_x$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = \sigma^2_1$, $Var(e_2) = \sigma^2_2$ and $X$, $e_1$ and $e_2$ are all independent of one another.

   (a) Make a path diagram of this model.

   (b) Prove that the squared correlation between $W$ and $G$ is strictly less than the reliability of $W$. Show your work.

   The idea here is that the squared *population* correlation[2] between an ordinary measurement and an imperfect gold standard measurement is strictly less than the actual

[1] This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/brunner/oldclass/431s23

[2] When we do Greek-letter calculations, we are figuring out what is happening in the population from which a data set might be a random sample.

reliability of the ordinary measurement. If we were to estimate such a squared correlation by the corresponding squared *sample* correlation, we would be estimating a quantity that is not the reliability. On the other hand, we would be estimating a lower bound for the reliability, and this could be reassuring if it were a high number.

4. Suppose we have two equivalent measurements with uncorrelated measurement error:

$$\begin{aligned} W_1 &= X + e_1 \\ W_2 &= X + e_2, \end{aligned}$$

where $E(X) = \mu_x$, $Var(X) = \sigma_x^2$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma_e^2$, and $X$, $e_1$ and $e_2$ are all independent. The "equivalent" part means the two measurements have the same amount of noise. What if we were to measure the true score $X$ by adding the two imperfect measurements together? Would the result be more reliable?

   (a) Let $S = W_1 + W_2$. Show that the reliability of $S$ is $\frac{\sigma_x^2}{\sigma_x^2 + \frac{1}{2}\sigma_e^2}$. Is this greater than $\frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}$?

   (b) Suppose you take $n$ independent measurements (in psychometric theory, these would be called equivalent test items). Show that the reliability of $S_n = \sum_{i=1}^n W_i$ is $\frac{\sigma_x^2}{\sigma_x^2 + \frac{1}{n}\sigma_e^2}$.

   (c) What is the reliability of $\overline{W}_n = \frac{1}{n}\sum_{i=1}^n W_i$? Show your work.

   (d) What happens to the reliability of $S_n$ and $\overline{W}_n$ as the number of measurements $n \to \infty$?

   Equivalent test items may be largely a fantasy, but this question shows how equivalent *tests* is a goal that can be closely approximated in practice. In the two equations displayed above, $W_1$ and $W_2$ might not be test items, but tests composed of multiple items. Each item might have a different error variance. But if the two *sums* or *averages* of the error variances are the same, the two tests are equivalent. This is nice, because it tells you that two tests do not need to be matched item for item in order to be equivalent.

5. Consider the two equivalent measurements at the start of Question 4. It is easy to imagine omitted variables that would affect both observed scores. For example, if $W_1$ and $W_2$ are two questionnaires about eating habits, some people will probably mis-remember or lie the same way on both questionnaires. Since $e_1$ and $e_2$ represent all other influences apart from the true quantity being measured, this means that $e_1$ and $e_2$ will have non-zero covariance. Furthermore, this covariance will be positive, since the omitted variables (there could be dozens of them) will tend to affect the two measurements in the same way. Accordingly, in the initial model of Question 4, let $Cov(e_1, e_2) = c > 0$.

   (a) Draw a path diagram of the model.

   (b) Show that $Corr(W_1, W_2)$ is strictly *greater* than the reliability.

   This means that in practice, omitted variables will result in over-estimates of reliability. There are almost always omitted variables.

6. 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 $X_i, e_i, \epsilon_i$ 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 $\epsilon_i$ 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 $\beta_1$? Answer Yes or No and show your work.

   (e) Why does this prove that $\beta_1$ is identifiable?

7. 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 $\epsilon_i$, $e_i$ 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 $\sigma_{ij}$ 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 $\beta$ is identifiable? Are there infinitely many, or just one point?

(e) The naive estimator of $\beta$ 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 $\beta$? To what does $\widehat{\beta}_n$ converge?

(f) 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 $\beta$ is identifiable.

(g) 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 $\beta$.

8. 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

$$\begin{array}{rcl} D_{i,1} & = & F_{i,1} + e_{i,1} \\ D_{i,2} & = & F_{i,2} + e_{i,2} \end{array} \qquad cov\begin{pmatrix} F_{i,1} \\ F_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix} \quad cov\begin{pmatrix} e_{i,1} \\ e_{i,2} \end{pmatrix} = \begin{pmatrix} \omega_1 & 0 \\ 0 & \omega_2 \end{pmatrix}$$

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 $\rho_1^2$ and the reliability of $D_{i,2}$ is $\rho_2^2$. How could you apply $\rho_1^2$ and $\rho_2^2$ 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. My answer is 0.304.

9. 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 $cov \begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$, $Var(\epsilon_i) = \psi$, $Var(e_1) = \omega$, all the expected values are zero, and the error terms $\epsilon_i$ 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." The usual fixed-$x$ estimator is

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

(a) $\widehat{\beta}_2$ converges in probability to a definite target. Give the target in terms of the model parameters. Remember that if $E(X) = 0$, then $E(X^2) = Var(X)$. This means you can use rules about variances to make some of the calculations easier.

(b) The target is a fairly complicated expression, but if it's correct, it should reduce to $\beta_2$ when $\omega = 0$ (no measurement error). Verify this.

(c) Now let $\omega > 0$ as before, and suppose that $H_0 : \beta_2 = 0$ is true. Does the $\widehat{\beta}_2$ converge to the true value of $\beta_2 = 0$ as $n \to \infty$ everywhere in the parameter space? Answer Yes or No.

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

10. 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 $\phi > 0$.
- $\epsilon_i$ 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 $\omega_2 > 0$.
- $X_i$, $\epsilon_i$, $e_i$, $e_{i,1}$ 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. Some of the variances and covariances you can just write down. For the others, 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 $\beta$ (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 $\beta$. 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.