Double Measurement Regression, Part Two[1]

STA 2101 Fall 2019

[1] See last slide for copyright information.

Overview

Double measurement

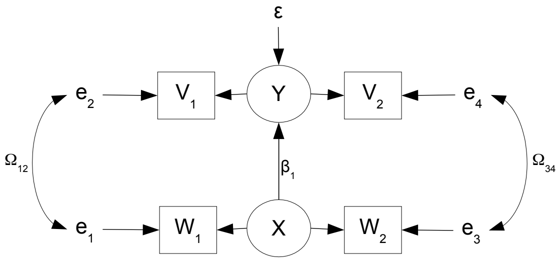
- We have studied an example where two independent measurements of a latent explanatory variable made all the model parameter identifiable.
- Extend the model.
- Double measurement can also help with correlated measurement error.

Correlated measurement error

- We are "measuring" exercise and snack food consumption by self-report.
- A simple additive model: What people report is the truth, plus a piece of noise that pushes the number up or down by a random amount that is different for each person.
- Is it reasonable to assume the error term for snack food is independent of the error term for exercise?
- This is another case of omitted variables.

- Acres planted by farmer's report and aerial photograph is a different story.
- Double measurement can help with correlated measurement error.

The general double measurement design



These are all matrices.

- The main idea is that $\mathbf{X}$ and $\mathbf{Y}$ are each measured twice, perhaps at different times using different methods.
- Measurement errors may be correlated within but not between sets of measurements.

Double Measurement Regression: A Two-Stage Model

Setting up a two-stage proof of identifiability

$$
\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\
\mathbf{D}_{i,1} &= \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\
\mathbf{D}_{i,2} &= \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2}
\end{aligned}
$$

Observable variables are $\mathbf{D}_{i,1}$ and $\mathbf{D}_{i,2}$: both are $(p+q) \times 1$.

$E(\mathbf{X}_i) = \boldsymbol{\mu}_x$, $cov(\mathbf{X}_i) = \boldsymbol{\Phi}_x$, $cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$, $cov(\mathbf{e}_{i,1}) = \boldsymbol{\Omega}_1$, $cov(\mathbf{e}_{i,2}) = \boldsymbol{\Omega}_2$. Also, $\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$, $\mathbf{e}_{i,1}$ and $\mathbf{e}_{i,2}$ are independent.

Measurement errors may be correlated
Look at the measurement model

$$
\begin{aligned}
\mathbf{F}_i &= \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} \\
\mathbf{D}_{i,1} &= \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\
\mathbf{D}_{i,2} &= \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2}
\end{aligned}
$$

$$
\begin{aligned}
cov(\mathbf{e}_{i,1}) &= \boldsymbol{\Omega}_1 = \left(\begin{array}{c|c} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \hline \boldsymbol{\Omega}_{12}^\top & \boldsymbol{\Omega}_{22} \end{array}\right) \\
cov(\mathbf{e}_{i,2}) &= \boldsymbol{\Omega}_2 = \left(\begin{array}{c|c} \boldsymbol{\Omega}_{33} & \boldsymbol{\Omega}_{34} \\ \hline \boldsymbol{\Omega}_{34}^\top & \boldsymbol{\Omega}_{44} \end{array}\right)
\end{aligned}
$$

Expected values of the observable variables

$\mathbf{D}_{i,1} = \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1}$ and $\mathbf{D}_{i,2} = \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2}$

$$
\begin{aligned}
E(\mathbf{D}_{i,1}) &= \left(\begin{array}{c} \boldsymbol{\mu}_{1,1} \\ \boldsymbol{\mu}_{1,2} \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\nu}_{1,1} + E(\mathbf{X}_i) \\ \boldsymbol{\nu}_{1,2} + E(\mathbf{Y}_i) \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\nu}_{1,1} + \boldsymbol{\mu}_x \\ \boldsymbol{\nu}_{1,2} + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{\mu}_x \end{array} \right) \\
E(\mathbf{D}_{i,2}) &= \left(\begin{array}{c} \boldsymbol{\mu}_{2,1} \\ \boldsymbol{\mu}_{2,2} \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\nu}_{2,1} + E(\mathbf{X}_i) \\ \boldsymbol{\nu}_{2,2} + E(\mathbf{Y}_i) \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\nu}_{2,1} + \boldsymbol{\mu}_x \\ \boldsymbol{\nu}_{2,2} + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{\mu}_x \end{array} \right)
\end{aligned}
$$

- $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\mu}_x$ parameters appear only in expected value, not covariance matrix.
- $\mathbf{X}_i$ is $p \times 1$ and $\mathbf{Y}_i$ is $q \times 1$.
- Even with $\boldsymbol{\beta}_1$ identified from the covariance matrix, have $2(p+q)$ equations in $3(p+q)$ unknown parameters.
- Identifying the expected values and intercepts is impossible.
- Re-parameterize, absorbing them into $\boldsymbol{\mu} = E \left(\begin{array}{c} \mathbf{D}_{i,1} \\ \mathbf{D}_{i,2} \end{array} \right)$.

Losing the intercepts and expected values by re-parameterization

- We cannot identify $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\mu}_x$ separately.
- Swallow them into $\boldsymbol{\mu}$.
- Estimate $\boldsymbol{\mu}$ with $\overline{\mathbf{D}}$.
- And it disappears from $L(\boldsymbol{\mu},\boldsymbol{\Sigma}) =$ $|\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{D}}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{D}}-\boldsymbol{\mu})\right\}$.
- And forget it. It's no great loss.
- Concentrate on the parameters that appear only in the covariance matrix of the observable data.
- Try to identify $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2)$ from $\boldsymbol{\Sigma} = cov\left(\begin{array}{c} \mathbf{D}_{i,1} \\ \mathbf{D}_{i,2} \end{array}\right)$.

Stage One: The latent variable model

$\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2)$

$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i$, where

- $cov(\mathbf{X}_i) = \boldsymbol{\Phi}_x$
- $cov(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$
- $\mathbf{X}_i$ and $\boldsymbol{\epsilon}_i$ are independent.

Vector of "factors" is $\mathbf{F}_i = \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right)$.

- Let $\boldsymbol{\Phi} = cov(\mathbf{F}_i)$.
- We know that $\boldsymbol{\Phi}_x$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\Psi}$ are functions of $\boldsymbol{\Phi}$.
- We've already shown it; this is a regression model.

That's Stage One. Parameters of the latent variable model are functions of $\boldsymbol{\Phi}$.

Stage Two: The measurement model

$$\begin{aligned}\mathbf{D}_{i,1} &= \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\ \mathbf{D}_{i,2} &= \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2}\end{aligned}$$

$cov(\mathbf{e}_{i,1}) = \boldsymbol{\Omega}_1$, $cov(\mathbf{e}_{i,2}) = \boldsymbol{\Omega}_2$. Also, $\mathbf{F}_i$, $\mathbf{e}_{i,1}$ and $\mathbf{e}_{i,2}$ are independent.

$$\boldsymbol{\Sigma} = cov\left(\begin{array}{c}\mathbf{D}_{i,1} \\ \mathbf{D}_{i,2}\end{array}\right) = \left(\begin{array}{cc}\boldsymbol{\Phi} + \boldsymbol{\Omega}_1 & \boldsymbol{\Phi} \\ \boldsymbol{\Phi} & \boldsymbol{\Phi} + \boldsymbol{\Omega}_2\end{array}\right)$$

$\boldsymbol{\Phi}$, $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ can easily be recovered from $\boldsymbol{\Sigma}$.

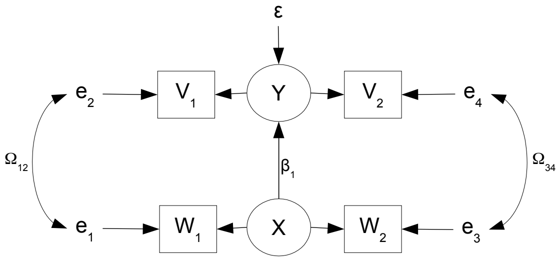
All the parameters in the covariance matrix are identifiable

$\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2)$

- $\boldsymbol{\Phi}_x$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\Psi}$ are functions of $\boldsymbol{\Phi} = cov(\mathbf{F}_i)$.
- $\boldsymbol{\Phi}$, $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are functions of $\boldsymbol{\Sigma} = cov\left(\begin{array}{c} \mathbf{D}_{i,1} \\ \mathbf{D}_{i,2} \end{array}\right)$.
- $\boldsymbol{\Sigma}$ is a function of the probability distribution of the observable data.
- So $\boldsymbol{\beta}_1, \boldsymbol{\Phi}_x, \boldsymbol{\Psi}, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2$ are all functions of the probability distribution of the observable data.
- They are identifiable.

Parameters of the double measurement regression model are identifiable

After re-parameterization



- Correlated measurement error within sets is allowed.
- This is a big plus, because omitted variables are a reality.
- Correlated measurement error between sets must be ruled out by careful data collection.
- No need to do the calculations ever again.

The BMI Health Study

- Body Mass Index: Weight in Kilograms divided by Height in Meters Squared.
- Under 18 means underweight, Over 25 means overweight, Over 30 means obese.
- High BMI is associated with poor health, like high blood pressure and high cholesterol.
- People with high BMI tend to be older and fatter.
- *But*, what if you have a high BMI but are in good physical shape (low percent body fat)?

The Question

- If you control for age and percent body fat, is BMI still associated with indicators for poor health?
- But percent body fat (and to a lesser extent, age) are measured with error. Standard ways of controlling for them with ordinary regression are highly suspect.
- Use the double measurement design.

True variables (all latent)

- X_1 = Age
- X_2 = BMI
- X_3 = Percent body fat
- Y_1 = Cholesterol
- Y_2 = Diastolic blood pressure

Measure twice with different personnel at different locations and by different methods

	Measurement Set One	Measurement Set Two
Age	Self report	Passport or birth certificate
BMI	Dr. Office measurements	Lab technician, no shoes, gown
% Body Fat	Tape and calipers, Dr. Office	Submerge in water tank
Cholesterol	Lab 1	Lab 2
Diastolic BP	Blood pressure cuff, Dr. office	Digital readout, mostly automatic

- Set two is of generally higher quality.
- Correlation of measurement errors is unlikely between sets.

Copyright Information