Double Measurement Regression[1]

STA431 Spring 2023

[1]See last slide for copyright information.

Overview

Seeking identifiability

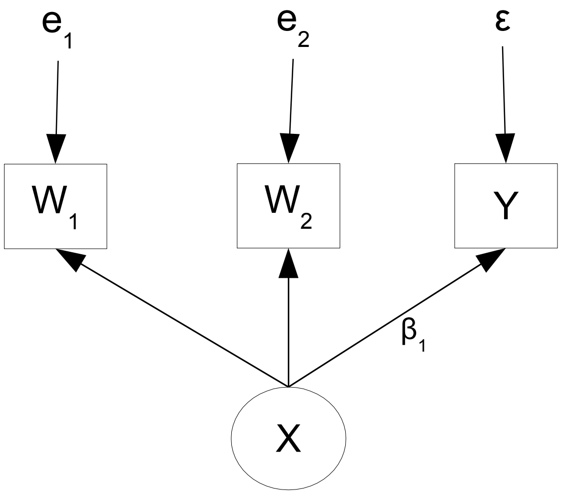
The parameters of this model are not identifiable.

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

- For example, X might be number of acres planted and Y might be crop yield.
- Plan the statistical analysis in advance.
- Take 2 independent measurements of the explanatory variable.
- Say, farmer's report and satellite photograph.

Double measurement

Of the explanatory variable



Model

Could have written this down based on the path diagram

Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} W_{i,1} &= \nu_1 + X_i + e_{i,1} \\ W_{i,2} &= \nu_2 + X_i + e_{i,2} \\ Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i, \end{aligned}$$

where

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

Parameter Count Rule

$$
\begin{array}{lcl}
W_{i,1} & = & \nu_1 + X_i + e_{i,1} \\
W_{i,2} & = & \nu_2 + X_i + e_{i,2} \\
Y_i & = & \beta_0 + \beta_1 X_i + \epsilon_i,
\end{array}
$$

- $\boldsymbol{\theta} = (\nu_1, \nu_2, \beta_0, \mu_x, \beta_1, \phi, \psi, \omega_1, \omega_2)$: 9 parameters.
- Three expected values, three variances and three covariances: 9 moments.
- Identifiability is possible, but not guaranteed.

Distribution of the sample data

$$
\begin{array}{rcl}
W_{i,1} & = & \nu_1 + X_i + e_{i,1} \\
W_{i,2} & = & \nu_2 + X_i + e_{i,2} \\
Y_i & = & \beta_0 + \beta_1 X_i + \epsilon_i,
\end{array}
$$

The model implies that the triples $\mathbf{d}_i = (W_{i,1}, W_{i,2}, Y_i)^\top$ are independent multivarate normal with

$$
E(\mathbf{d}_i) = E \left(\begin{array}{c} W_{i,1} \\ W_{i,1} \\ Y_i \end{array} \right) = \left(\begin{array}{c} \mu_1 \\ \mu_2 \\ \mu_3 \end{array} \right) = \left(\begin{array}{c} \mu_x + \nu_1 \\ \mu_x + \nu_2 \\ \beta_0 + \beta_1 \mu_x \end{array} \right),
$$

and variance covariance matrix $cov(\mathbf{d}_i) = \boldsymbol{\Sigma} =$

$$
\left(\begin{array}{ccc} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{array} \right) = \left(\begin{array}{ccc} \phi + \omega_1 & \phi & \beta_1 \phi \\ & \phi + \omega_2 & \beta_1 \phi \\ & & \beta_1^2 \phi + \psi \end{array} \right).
$$

Are the parameters in the covariance matrix identifiable?

Six equations in five unknowns

$$\left(\begin{array}{ccc} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{array} \right) = \left(\begin{array}{ccc} \phi+\omega_1 & \phi & \beta_1\phi \\ & \phi+\omega_2 & \beta_1\phi \\ & & \beta_1^2\phi+\psi \end{array} \right).$$

$$\begin{array}{rcl} \phi & = & \sigma_{12} \\ \omega_1 & = & \sigma_{11}-\sigma_{12} \\ \omega_2 & = & \sigma_{22}-\sigma_{12} \\ \beta_1 & = & \frac{\sigma_{13}}{\sigma_{12}} \\ \psi & = & \sigma_{33}-\beta_1^2\phi = \sigma_{33}-\frac{\sigma_{13}^2}{\sigma_{12}} \end{array}$$

Yes.

What about the expected values?

Model equations again:

$$
\begin{aligned}
W_{i,1} &= \nu_1 + X_i + e_{i,1} \\
W_{i,2} &= \nu_2 + X_i + e_{i,2} \\
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i,
\end{aligned}
$$

Expected values:

$$
\begin{aligned}
\mu_1 &= \nu_1 + \mu_x \\
\mu_2 &= \nu_2 + \mu_x \\
\mu_3 &= \beta_0 + \beta_1 \mu_x
\end{aligned}
$$

Four parameters appear only in the expected values: $\nu_1, \nu_2, \mu_x, \beta_0$.

- Three equations in four unknowns, even with β_1 identified from the covariance matrix.
- Parameter count rule applies.

Re-parameterize

$\mu_1 = \nu_1 + \mu_x \qquad \mu_2 = \nu_2 + \mu_x \qquad \mu_3 = \beta_0 + \beta_1 \mu_x$

- Absorb $\nu_1, \nu_2, \mu_x, \beta_0$ into $\boldsymbol{\mu}$.
- Parameter was $\boldsymbol{\theta} = (\nu_1, \nu_2, \beta_0, \mu_x,\ \beta_1, \phi, \psi, \omega_1, \omega_2)$
- Now it's $\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_3,\ \beta_1, \phi, \psi, \omega_1, \omega_2)$.
- Dimension of the parameter space is now one less.
- We haven't lost much, especially because the model was already re-parameterized.

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \epsilon \\
V &= \nu_0 + Y + e \\
&= \nu_0 + (\beta_0 + \beta_1 X + \epsilon) + e \\
&= (\nu_0 + \beta_0) + \beta_1 X + (\epsilon + e) \\
&= \beta_0^\prime + \beta X + \epsilon^\prime
\end{aligned}
$$

Re-parameterization

- Re-parameterization makes maximum likelihood possible.
- Otherwise the maximum is not unique and it's a mess.
- Estimate $\boldsymbol{\mu}$ with $\overline{\mathbf{d}}$ and it simply disappears from

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{d}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{d}} - \boldsymbol{\mu})\right\}$$

- This step is so common it becomes silent.
- Model equations are often written in centered form.

Back to the covariance structure equations

$$
\left(\begin{array}{ccc}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
 & \sigma_{22} & \sigma_{23} \\
 & & \sigma_{33}
\end{array}\right) = \left(\begin{array}{ccc}
\phi+\omega_1 & \phi & \beta_1\phi \\
 & \phi+\omega_2 & \beta_1\phi \\
 & & \beta_1^2\phi+\psi
\end{array}\right).
$$

- Notice that the model dictates $\sigma_{1,3} = \sigma_{2,3}$.
- There are two ways to solve for β_1:
  $\beta_1 = \frac{\sigma_{13}}{\sigma_{12}}$ and $\beta_1 = \frac{\sigma_{23}}{\sigma_{12}}$.
- Does this mean the solution for β_1 is not "unique?"

Testing goodness of fit.

$$\left(\begin{array}{ccc} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{array}\right) = \left(\begin{array}{ccc} \phi+\omega_1 & \phi & \beta_1\phi \\ & \phi+\omega_2 & \beta_1\phi \\ & & \beta_1^2\phi+\psi \end{array}\right).$$

- $\sigma_{1,3} = \sigma_{2,3}$ is a *model-induced constraint* upon $\mathbf{\Sigma}$.
- It's a testable null hypothesis.
- If rejected, the model is called into question.
- Likelihood ratio test comparing this model to a completely unrestricted multivariate normal model:

$$G^2 = -2\ln \frac{L\left(\overline{\mathbf{d}}, \mathbf{\Sigma}(\widehat{\boldsymbol{\theta}})\right)}{L(\overline{\mathbf{d}}, \widehat{\mathbf{\Sigma}})}$$

- Valuable even if the data are not normal.

The Reproduced Covariance Matrix

- $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ is called the *reproduced covariance matrix.*
- It is the covariance matrix of the observable data, written as a function of the model parameters and evaluated at the MLE.

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}) = \left(\begin{array}{ccc} \widehat{\phi} + \widehat{\omega}_1 & \widehat{\phi} & \widehat{\beta}_1 \widehat{\phi} \\ & \widehat{\phi} + \widehat{\omega}_2 & \widehat{\beta}_1 \widehat{\phi} \\ & & \widehat{\beta}_1^2 \widehat{\phi} + \widehat{\psi} \end{array} \right)$$

- The reproduced covariance matrix obeys all model-induced constraints, while $\widehat{\boldsymbol{\Sigma}}$ does not.
- But if the model is right they should be close.

$$G^2 = -2 \ln \frac{L\left(\overline{\mathbf{d}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\right)}{L(\overline{\mathbf{d}}, \widehat{\boldsymbol{\Sigma}})}$$

General pattern for testing goodness of fit

- Suppose there are k moment structure equations in p parameters, and all the parameters are identifiable.
- If $p < k$, call the parameter vector *over-identifiable*.
- Only need p equations to solve for $\boldsymbol{\theta}$.
- Substituting the solutions (in terms of σ_{ij}) back into the unused equations would yield $k-p$ equality constraints on $\boldsymbol{\Sigma}$.
- Test those constraints with $G^2 = -2\ln\frac{L\left(\overline{\mathbf{d}},\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\right)}{L(\overline{\mathbf{d}},\widehat{\boldsymbol{\Sigma}})}$.
- $df = k-p$
- Don't need to actually derive the constraints – just count them.

With the same number of equations and parameters

- If the parameter is identifiable, call it *just identifiable.*
- Parameters are 1-1 with those of an unrestricted multivariate normal.
- Call the model "saturated."
- There are no equality constraints on $\boldsymbol{\Sigma}$.
- No likelihood ratio test because $G^2 = -2\ln\frac{L\left(\overline{\mathbf{d}},\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\right)}{L(\overline{\mathbf{d}},\widehat{\boldsymbol{\Sigma}})} = 0$.
- This is what happens in regression with all observed variables.

Data analysis strategy

- Verify identifiability.
- If the model is over-identified, test goodness of fit.
- If it passes (non-significant), proceed.
- Now think of your model as the "full," or unrestricted model.
- Compared to some (even more) reduced model that is restricted by a null hypothesis like $\beta_1 = 0$.
- Fit the reduced model.
- Subtract goodness of fit (G^2 or "chi-square") statistics to test H_0.

Subtract Likelihood Ratio Fit Statistics

Badness of fit

G^2 tests the full model against the saturated model, and G_0^2 tests the restricted model against the saturated model.

$$\begin{aligned}
G_0^2 - G^2 &= -2\ln \frac{L\left(\overline{\mathbf{d}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_0)\right)}{L(\overline{\mathbf{d}}, \widehat{\boldsymbol{\Sigma}})} - -2\ln \frac{L\left(\overline{\mathbf{d}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\right)}{L(\overline{\mathbf{d}}, \widehat{\boldsymbol{\Sigma}})} \\
&= -2\left(\ln L\left(\overline{\mathbf{d}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_0)\right) - \ln L(\overline{\mathbf{d}}, \widehat{\boldsymbol{\Sigma}}) - \ln L\left(\overline{\mathbf{d}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\right) \right.\\
&\qquad \left. + \ln L(\overline{\mathbf{d}}, \widehat{\boldsymbol{\Sigma}})\right) \\
&= -2\ln \frac{L\left(\overline{\mathbf{d}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_0)\right)}{L\left(\overline{\mathbf{d}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\right)}
\end{aligned}$$

Further comments

- Models with non-identifiable parameters can imply testable equality constraints, but testing them is not automatic.
- Models can imply *inequality* constraints on $\boldsymbol{\Sigma}$, too.
- Using the solutions

$$\begin{aligned}
\phi &= \sigma_{12} \\
\omega_1 &= \sigma_{11} - \sigma_{12} \\
\omega_2 &= \sigma_{22} - \sigma_{12} \\
\beta_1 &= \frac{\sigma_{13}}{\sigma_{12}} \\
\psi &= \sigma_{33} - \beta_1^2\phi = \sigma_{33} - \frac{\sigma_{13}^2}{\sigma_{12}}
\end{aligned}$$

We get four inequality constraints.

Four inequality constraints on $\boldsymbol{\Sigma}$

$$\left(\begin{array}{ccc} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{array} \right) = \left(\begin{array}{ccc} \phi+\omega_1 & \phi & \beta_1\phi \\ & \phi+\omega_2 & \beta_1\phi \\ & & \beta_1^2\phi+\psi \end{array} \right).$$

$$\begin{array}{rcl} \phi & = & \sigma_{12} > 0 \\ \omega_1 & = & \sigma_{11}-\sigma_{12} > 0 \\ \omega_2 & = & \sigma_{22}-\sigma_{12} > 0 \\ \psi & = & \sigma_{33}-\dfrac{\sigma_{13}^2}{\sigma_{12}} > 0 \end{array}$$

Inequality constraints

- Inequality constraints arise because variances are positive.
- Or more generally, covariance matrices are positive definite.
- Could inequality constraints be violated in numerical maximum likelihood?
- Definitely.
- But only a little by sampling error if the model is correct.
- So maybe it's not so dumb to test hypotheses like $H_0 : \omega_1 = 0$.
- Since the model says $\omega_1 = \sigma_{11} - \sigma_{12}$ and it might not be true.

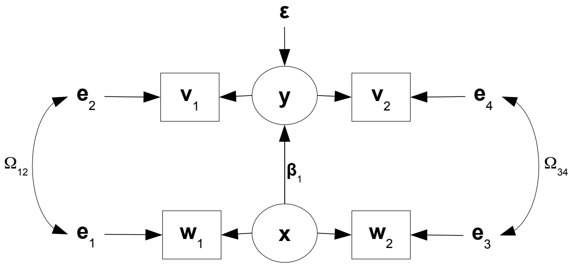
Computation with `lavaan`

`431s23Babydouble.pdf`

This link will probably be broken once the term is over. See the course website for another route to the output file:

`http://www.utstat.toronto.edu/brunner/oldclass/431s23`

The general double measurement design



These are all matrices.
Double measurement can help solve a big problem: Correlated measurement error.

- The main idea is that **x** and **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

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

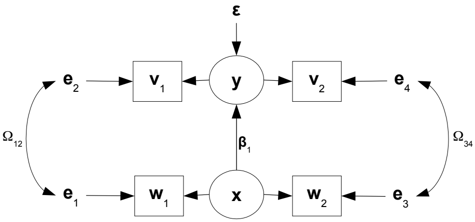
Observable variables are $\mathbf{d}_{i,1}$ and $\mathbf{d}_{i,2}$: both are $(p+q)\times 1$.

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

$$
\begin{aligned}
\mathbf{F}_i &= \left(\frac{\mathbf{x}_i}{\mathbf{y}_i}\right) \\
\mathbf{d}_{i,1} &= \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\
\mathbf{d}_{i,2} &= \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2}
\end{aligned}
$$



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

Expected values of the observable variables

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

$$\begin{aligned} E(\mathbf{d}_{i,1}) &= \left(\frac{\boldsymbol{\mu}_{1,1}}{\boldsymbol{\mu}_{1,2}}\right) = \left(\frac{\boldsymbol{\nu}_{1,1} + E(\mathbf{x}_i)}{\boldsymbol{\nu}_{1,2} + E(\mathbf{y}_i)}\right) = \left(\frac{\boldsymbol{\nu}_{1,1} + \boldsymbol{\mu}_x}{\boldsymbol{\nu}_{1,2} + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{\mu}_x}\right) \\ E(\mathbf{d}_{i,2}) &= \left(\frac{\boldsymbol{\mu}_{2,1}}{\boldsymbol{\mu}_{2,2}}\right) = \left(\frac{\boldsymbol{\nu}_{2,1} + E(\mathbf{x}_i)}{\boldsymbol{\nu}_{2,2} + E(\mathbf{y}_i)}\right) = \left(\frac{\boldsymbol{\nu}_{2,1} + \boldsymbol{\mu}_x}{\boldsymbol{\nu}_{2,2} + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{\mu}_x}\right) \end{aligned}$$

- $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $\boldsymbol{\beta}_0$ 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 $\boldsymbol{\mu} = E\left(\frac{\mathbf{d}_{i,1}}{\mathbf{d}_{i,2}}\right)$.

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 $\boldsymbol{\Sigma} = cov \left(\frac{\mathbf{d}_{i,1}}{\mathbf{d}_{i,2}} \right)$.

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 $\mathbf{F}_i = \left(\begin{array}{c} \mathbf{x}_i \\ \hline \mathbf{y}_i \end{array} \right)$.

- 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

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

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

$$
cov\left(\begin{array}{c} \mathbf{d}_{i,1} \\ \hline \mathbf{d}_{i,2} \end{array}\right) = \boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{12}^\top & \boldsymbol{\Sigma}_{22} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{\Phi} + \boldsymbol{\Omega}_1 & \boldsymbol{\Phi} \\ \hline \boldsymbol{\Phi} & \boldsymbol{\Phi} + \boldsymbol{\Omega}_2 \end{array}\right)
$$

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

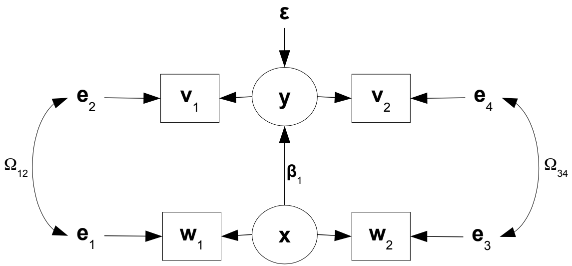
All the parameters in the covariance matrix are identifiable

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

- $\boldsymbol{\Phi}_x$, $\boldsymbol{\beta}_1$ 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?
- 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

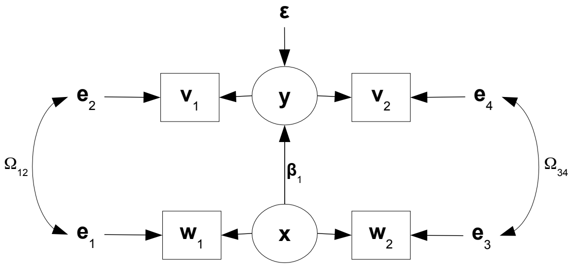
- Set two is of generally higher quality.
- Correlation of measurement errors is unlikely between sets.

Method of Moments

- What if the distributions are not normal?
- What's the parameter vector?

  $\theta = (\boldsymbol{\beta}, F_{\mathbf{x}}, F_{\boldsymbol{\epsilon}}, F_{\mathbf{e}})$.
  - $\boldsymbol{\Phi}$ is a function of $F_{\mathbf{x}}$.
  - $\boldsymbol{\Omega}$ is a function of $F_{\mathbf{e}}$.
  - $\boldsymbol{\Psi}$ is a function of $F_{\boldsymbol{\epsilon}}$.
- We are only interested in $\boldsymbol{\beta}$ anyway.
- Put hats on solution to covariance structure equations?

Path diagram again



Covariance structure equations

$$
\begin{aligned}
\boldsymbol{\Sigma} &= cov\begin{pmatrix} \mathbf{w}_{i,1} \\ \mathbf{v}_{i,1} \\ \mathbf{w}_{i,2} \\ \mathbf{v}_{i,2} \end{pmatrix} \\
&= \left(\begin{array}{c|c|c|c}
\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \hline
 & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \hline
 & & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} \\ \hline
 & & & \boldsymbol{\Sigma}_{44}
\end{array}\right) \\
&= \left(\begin{array}{c|c|c|c}
\boldsymbol{\Phi} + \boldsymbol{\Omega}_{11} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Omega}_{12} & \boldsymbol{\Phi} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top \\ \hline
 & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Psi} + \boldsymbol{\Omega}_{22} & \boldsymbol{\beta}\boldsymbol{\Phi} & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Psi} \\ \hline
 & & \boldsymbol{\Phi} + \boldsymbol{\Omega}_{33} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Omega}_{34} \\ \hline
 & & & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Psi} + \boldsymbol{\Omega}_{44}
\end{array}\right)
\end{aligned}
$$

The interesting equations

$$\begin{aligned}
cov(\mathbf{w}_{i,1}, \mathbf{w}_{i,2}) &= \boldsymbol{\Sigma}_{13} = \boldsymbol{\Phi} \\
cov(\mathbf{v}_{i,1}, \mathbf{w}_{i,2}) &= \boldsymbol{\Sigma}_{23} = \boldsymbol{\beta}\boldsymbol{\Phi} \\
cov(\mathbf{w}_{i,1}, \mathbf{v}_{i,2}) &= \boldsymbol{\Sigma}_{14} = \boldsymbol{\Phi}\boldsymbol{\beta}^\top
\end{aligned}$$

Solutions

$$\begin{aligned}
\boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{13} \\
\boldsymbol{\beta} &= \boldsymbol{\Sigma}_{23}\boldsymbol{\Phi}^{-1} = \boldsymbol{\Sigma}_{14}^\top\boldsymbol{\Phi}^{-1}
\end{aligned}$$

MOM estimates

Using the solutions

$$\begin{aligned}\boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{13} = cov(\mathbf{w}_{i,1}, \mathbf{w}_{i,2}) \\ \boldsymbol{\beta} &= \boldsymbol{\Sigma}_{23}\boldsymbol{\Phi}^{-1} = \boldsymbol{\Sigma}_{14}^{\top}\boldsymbol{\Phi}^{-1}\end{aligned}$$

$$\begin{aligned}\widehat{\boldsymbol{\Phi}}_M &= \frac{1}{2}(\widehat{\boldsymbol{\Sigma}}_{13} + \widehat{\boldsymbol{\Sigma}}_{13}^{\top}) \\ \widehat{\boldsymbol{\beta}}_M &= \frac{1}{2}\left(\widehat{\boldsymbol{\Sigma}}_{23} + \widehat{\boldsymbol{\Sigma}}_{14}^{\top}\right)\widehat{\boldsymbol{\Phi}}_M^{-1}\end{aligned}$$

The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_M$ is multivariate normal. Use Theorem A.1 from Appendix A, and the multivariate delta method, also given in Appendix A.

Asymptotic distribution of $\widehat{\boldsymbol{\beta}}_M = \frac{1}{2}\left(\widehat{\boldsymbol{\Sigma}}_{23} + \widehat{\boldsymbol{\Sigma}}_{14}^\top\right)\widehat{\boldsymbol{\Phi}}_M^{-1}$

Where $\widehat{\boldsymbol{\Phi}}_M = \frac{1}{2}(\widehat{\boldsymbol{\Sigma}}_{13} + \widehat{\boldsymbol{\Sigma}}_{13}^\top)$

- $\widehat{\boldsymbol{\beta}}_M$ is approximately multivariate normal with expected value $\boldsymbol{\beta}$
- And covariance matrix …

Bootstrap.

Copyright Information