Structural Equation Models

The General Case

An Extension of Multiple Regression

- More than one regression-like equation
- Includes latent variables
- Variables can be dependent in one equation and independent in another
- Modest changes in notation
- Vocabulary
- Path diagrams
- No intercepts, all expected values zero
- Serious modeling (compared to ordinary statistical models)
  - Causal models
  - Parameter identifiability

Variables can be dependent in one equation and independent in another

- Variables (IQ = Intelligence Quotient):
  - X_1 = Mother’s adult IQ
  - X_2 = Father’s adult IQ
  - Y_1 = Person’s adult IQ
  - Y_2 = Child’s IQ in Grade 8

$$\begin{aligned} Y_1 &= \alpha_1 + \gamma_1 X_1 + \gamma_2 X_2 + \epsilon_1 \\ Y_2 &= \alpha_2 + \beta Y_1 + \epsilon_2 \end{aligned}$$

- Of course all these variables are measured with error
- We will lose the intercepts very soon.

Modest changes in notation

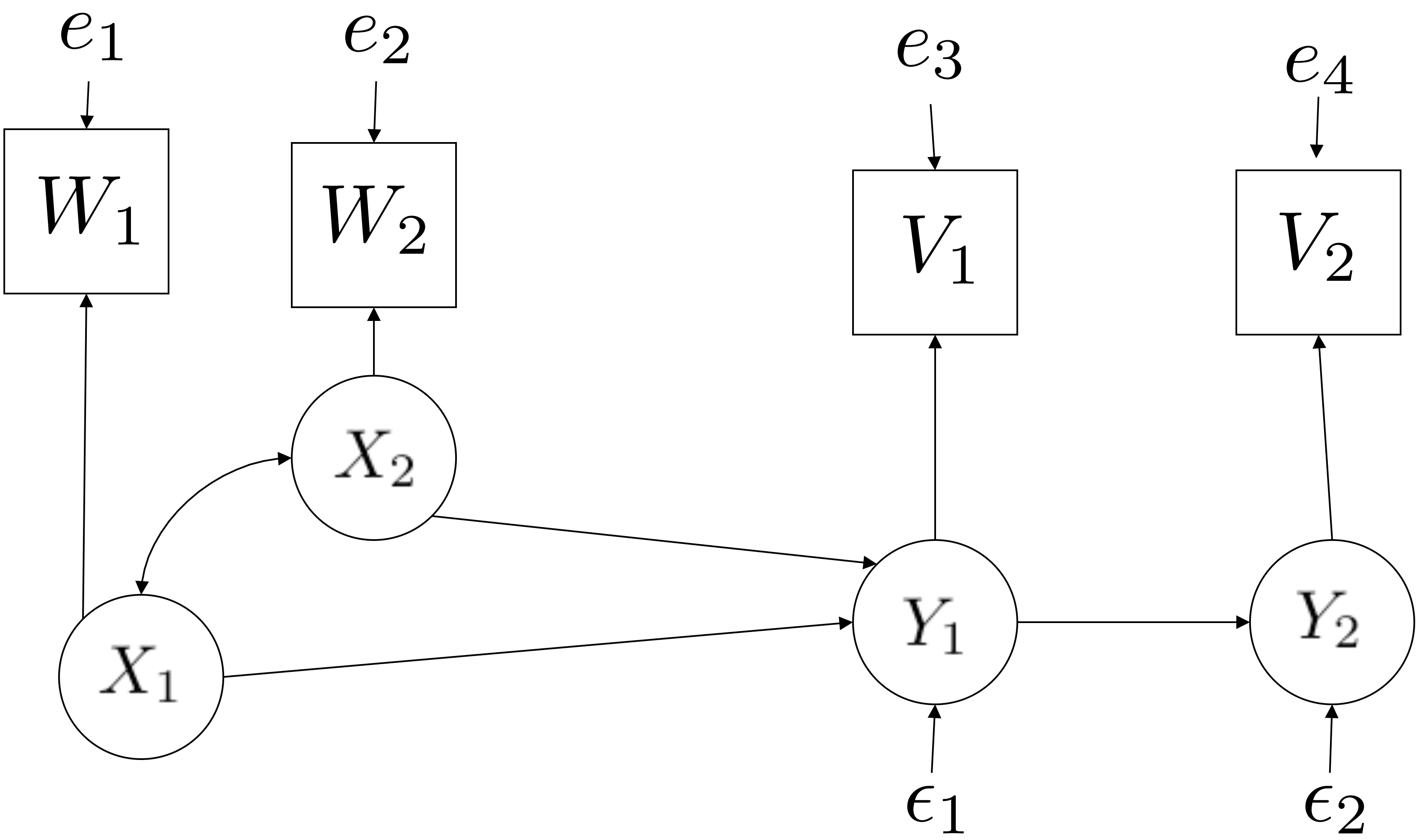
- Regression coefficients are now called gamma instead of beta
- Betas are used for links between Y variables
- Intercepts are alphas but they will soon disappear.
- Especially when model equations are written in scalar form, we feel free to drop the subscript *i*; implicitly, everything is independent and identically distributed for *i = 1, …, n*.

$$\begin{aligned} Y_1 &= \alpha_1 + \gamma_1 X_1 + \gamma_2 X_2 + \epsilon_1 \\ Y_2 &= \alpha_2 + \beta Y_1 + \epsilon_2 \end{aligned}$$

Strange Vocabulary

- Variables can be Latent or Manifest.
  - Manifest means observable
  - All error terms are latent
- Variables can be Exogenous or Endogenous
  - **Ex**ogenous variables appear only on the right side of the = sign.
    - Think “X” for independent variable.
    - All error terms are exogenous
  - **End**ogenous variables appear on the left of at least one = sign.
    - Think “end” of an arrow pointing from exogenous to endogenous
    - Betas link endogenous variables to other endogenous variables.

Path diagrams



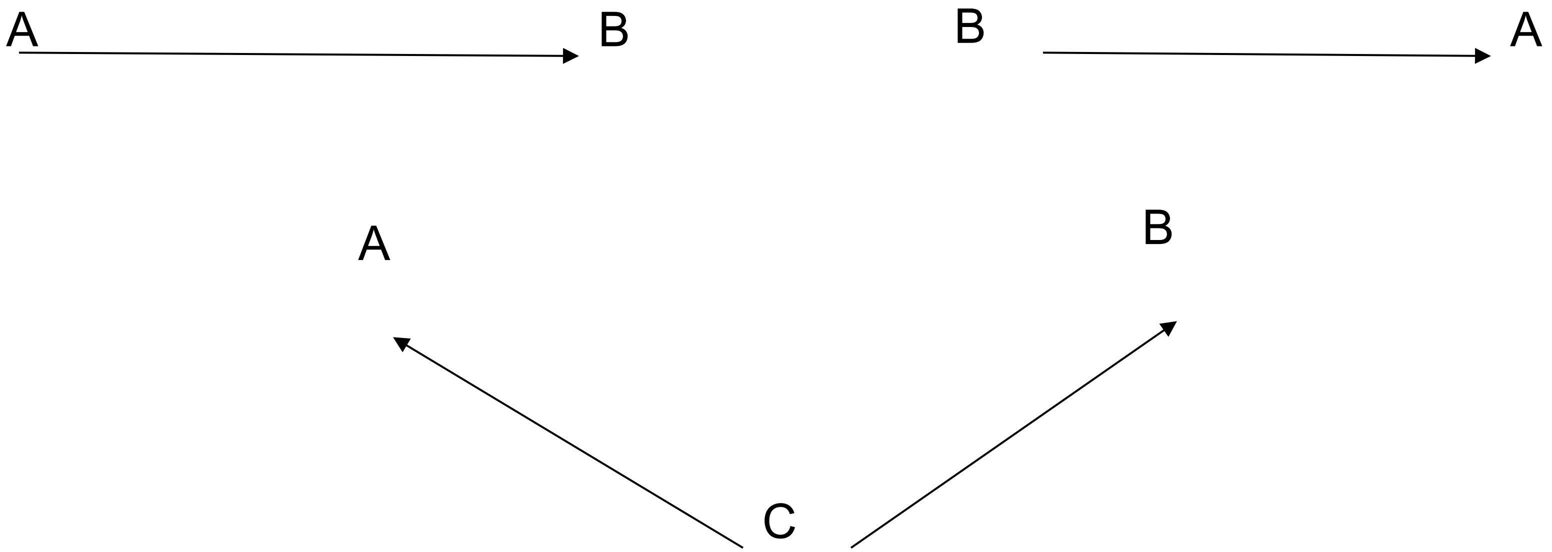
Path Diagram Rules

- Latent variables are enclosed by ovals.
- Observable (manifest) variables are enclosed by rectangles.
- Error terms are not enclosed
  - Sometimes the arrows from the error terms seem to come from nowhere. The symbol for the error term does not appear in the path diagram.
  - Sometimes there are no arrows for the error terms at all. It is just assumed that such an arrow points to each endogenous variable.
- Straight, single-headed arrows point from each variable on the right side of an equation to the endogenous variable on the left side.
  - Sometimes the coefficient is written on the arrow, but sometimes it is not.
- A curved, double-headed arrow between two variables (always exogenous variables) means they have a non-zero covariance.
  - Sometimes the symbol for the covariance is written on the curved arrow, but sometimes it is not.

Causal Modeling (cause and effect)

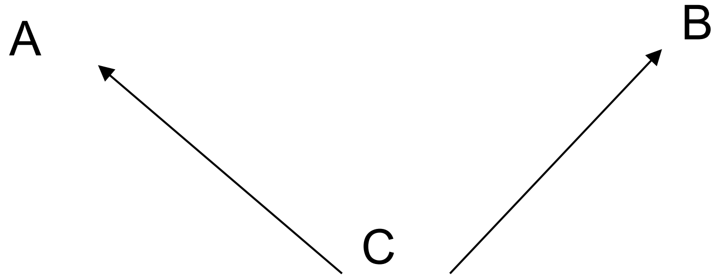
- The arrows deliberately imply that if A ➔ B, we are saying A *contributes* to B, or partly *causes* it.
- There may be other contributing variables. All the ones that are unknown are lumped together in the error term. It is a leap of faith to assume that these unknown variables are independent of the variables in the model.

But Correlation is not the same as causation!



Young smokers who buy contraband cigarettes tend to smoke more.

Confounding variable: A variable that contributes to both IV and DV, causing a misleading relationship between them.



Mozart Effect

- Babies who listen to classical music tend to do better in school later on.
- Does this mean parents should play classical music for their babies?
- Please comment. (What is one possible confounding variable?)

Experimental vs. Observational studies

- **Observational**: IV, DV just observed and recorded
- **Experimental**: Cases randomly assigned to values of IV
- Only a true experimental study can establish a causal connection between IV and DV

Structural equation models are mostly applied to observational data

- The correlation-causation issue is a logical problem, and no statistical technique can make it go away.
- So you (or the scientists you are helping) have to be able to defend the what-causes-what aspects of the model on other grounds.
- Parents’ IQ contributes to your IQ and your IQ contributes to your kid’s IQ. This is reasonable. It certainly does not go in the opposite direction.

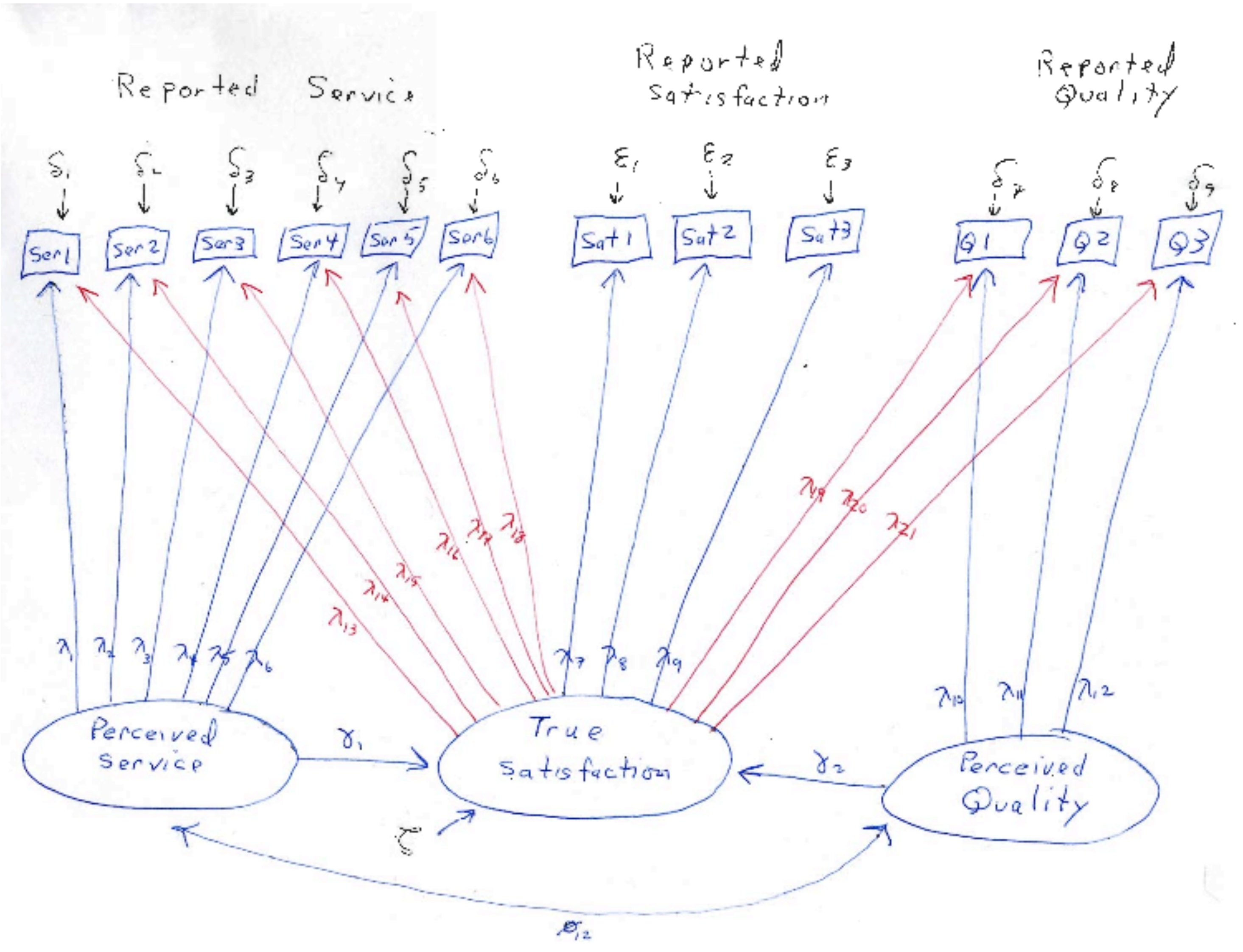
Models of Cause and Effect

- This is about the interpretation (and use) of structural equation models. Strictly speaking it is not a statistical issue and you don't have to think this way. However, …
- If you object to modeling cause and effect, structural equation modelers will challenge you.
- They will point out that regression models are structural equation models. Why do you put some variables on the left of the equals sign and not others?
  - You want to predict them.
  - It makes more sense that they are caused by the independent variables, compared to the other way around.
- If you want pure prediction, use standard tools.
- But if you want to discuss *why* a regression coefficient is positive or negative, you are assuming the independent variables in some way contribute to the dependent variable.

Serious Modeling

- Once you accept that model equations are statements about what contributes to what, you realize that structural equation models represent a rough *theory* of the data, with some parts (the parameter values) unknown.
- They are somewhere between ordinary statistical models, which are like one-size-fits-all clothing, and true scientific models, which are like tailor made clothing.
- So they are very flexible and potentially valuable. It is *good* to combine what the data can tell you with what you already know.
- But structural equation models can require a lot of input and careful thought to construct. In this course, we will get by mostly on common sense.
- In general, the parameters of the most reasonable model need not be identifiable. It depends upon the form of the data as well as on the model. Identifiability needs to be checked. Frequently, this can be done by inspection.

Example: Halo Effects in Real Estate


Reported Service
Reported Satisfaction
Reported Quality
δ_1 δ_2 δ_3 δ_4 δ_5 δ_6
ε_1 ε_2 ε_3
δ_7 δ_8 δ_9
Ser1 Ser2 Ser3 Ser4 Ser5 Ser6
Sat1 Sat2 Sat3
Q1 Q2 Q3
λ_{19} λ_{20} λ_{21}
λ_{16} λ_{17} λ_{18}
λ_{15}
λ_{14}
λ_{13}
λ_1 λ_2 λ_3 λ_4 λ_5 λ_6
λ_7 λ_8 λ_9
λ_{10} λ_{11} λ_{12}
Perceived Service
γ_1
True Satisfaction
γ_2
Perceived Quality
ζ
ϕ_{12}

Losing the intercepts and expected values

- Mostly, the intercepts and expected values are not identifiable anyway, as in multiple regression with measurement error.
- We have a chance to identify a *function* of the parameter vector – the parameters that appear in the covariance matrix **Σ** = V(**D**).
- Re-parameterize. The new parameter vector is the set of parameters in **Σ**, and also **μ** = E(**D**). Estimate **μ** with x-bar, forget it, and concentrate on inference for the parameters in **Σ**.
- To make calculation of the covariance matrix easier, write the model equations with zero expected values and no intercepts. The answer is also correct for non-zero intercepts and expected values, by the centering rule.

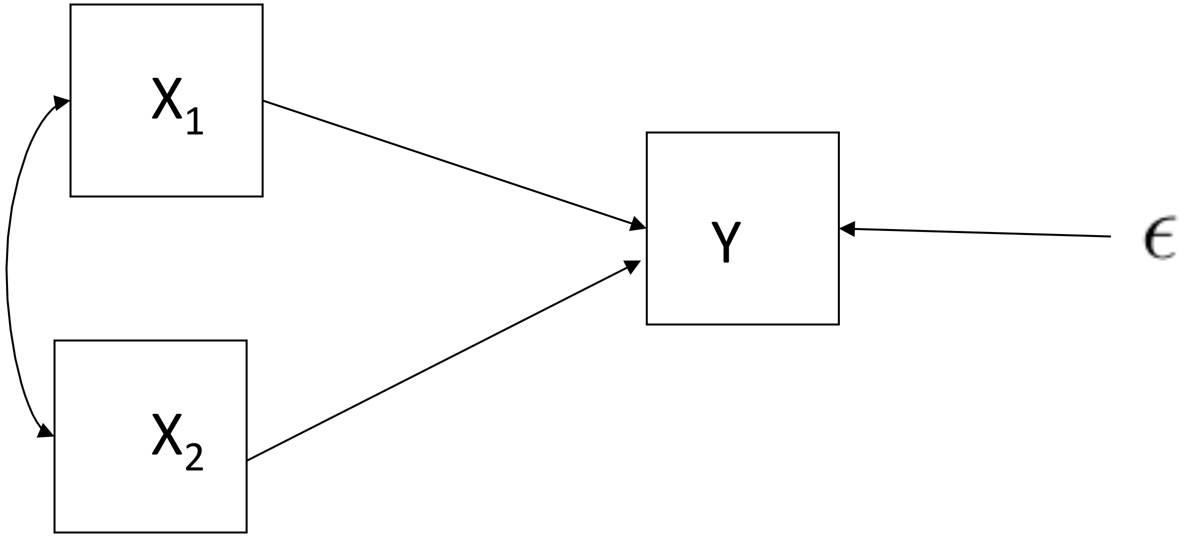
Centering the data

- This is another way to talk about the re-parameterization.
- Assume the data have been “centered” by subtracting off the sample mean D-bar from each data vector $\mathbf{D}_i$.
- Then assume a model with no intercepts and all expected values equal to zero.
- Data will be assumed multivariate normal with mean zero. All the (remaining) parameters are those in the covariance matrix.
- Because **D**-hat is close to $\boldsymbol{\mu}$ for large samples and all the inference is based on large-sample theory anyway, this does no harm.
- A good way to talk about it to clients, and not a bad way to think about it either. But really what we are doing is re-parameterizing.

Either way, from this point on the models have no means and no intercepts.

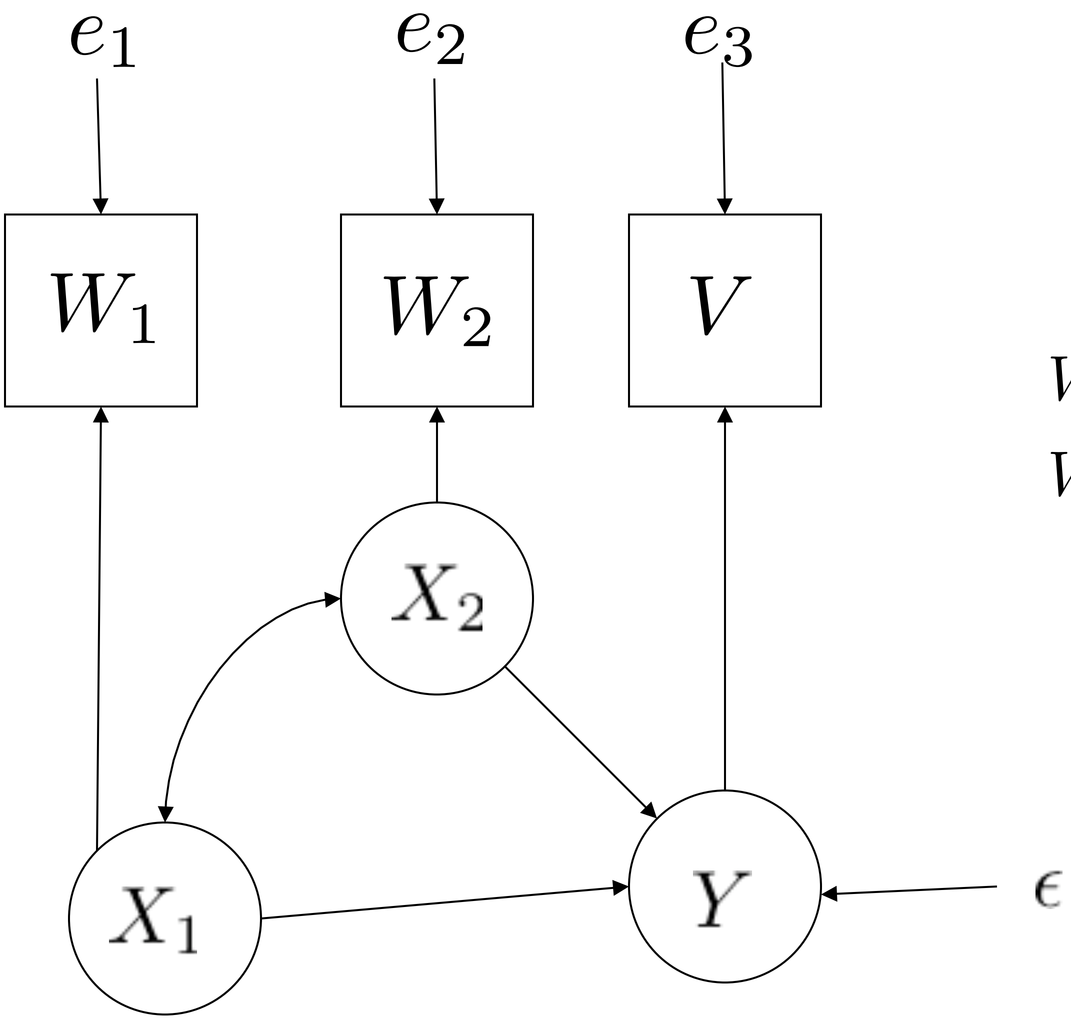
Now more examples

Multiple Regression



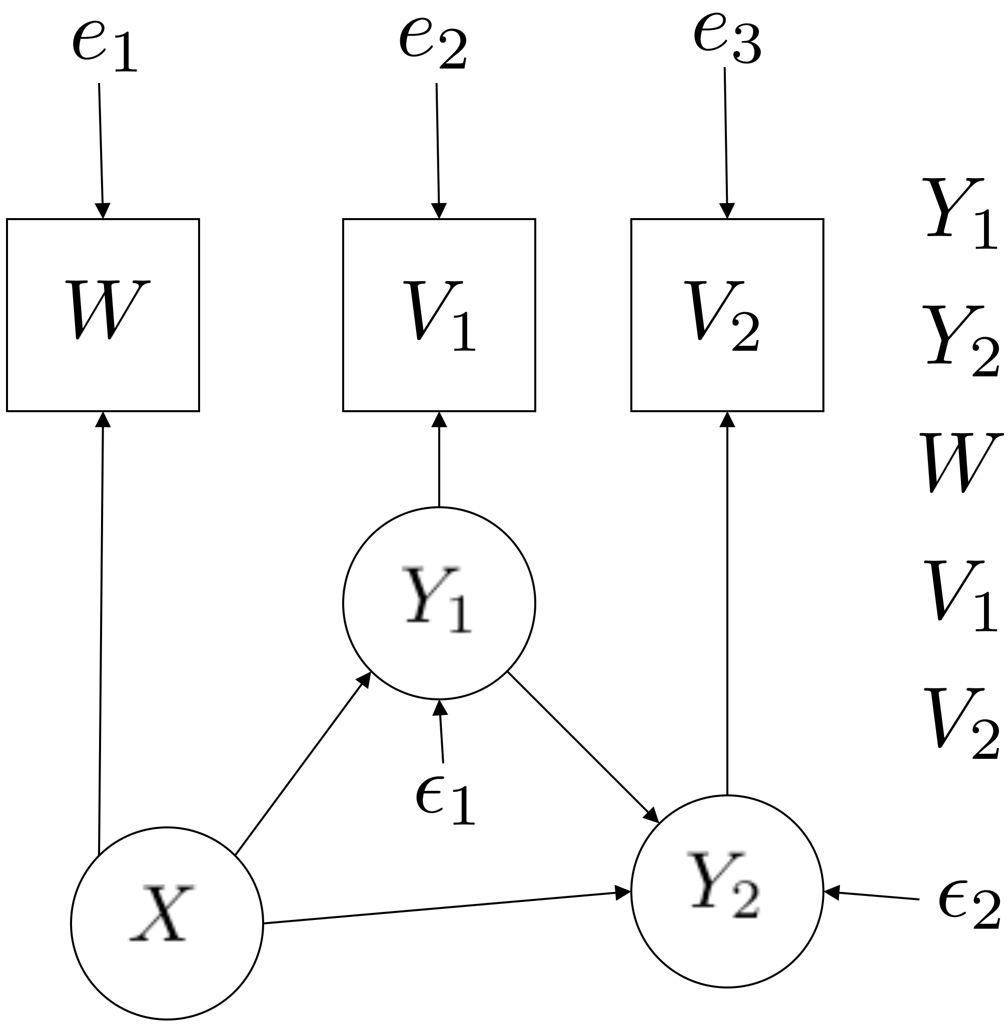
$$Y = \gamma_1 X_1 + \gamma_2 X_2 + \epsilon$$

Regression with measurement error



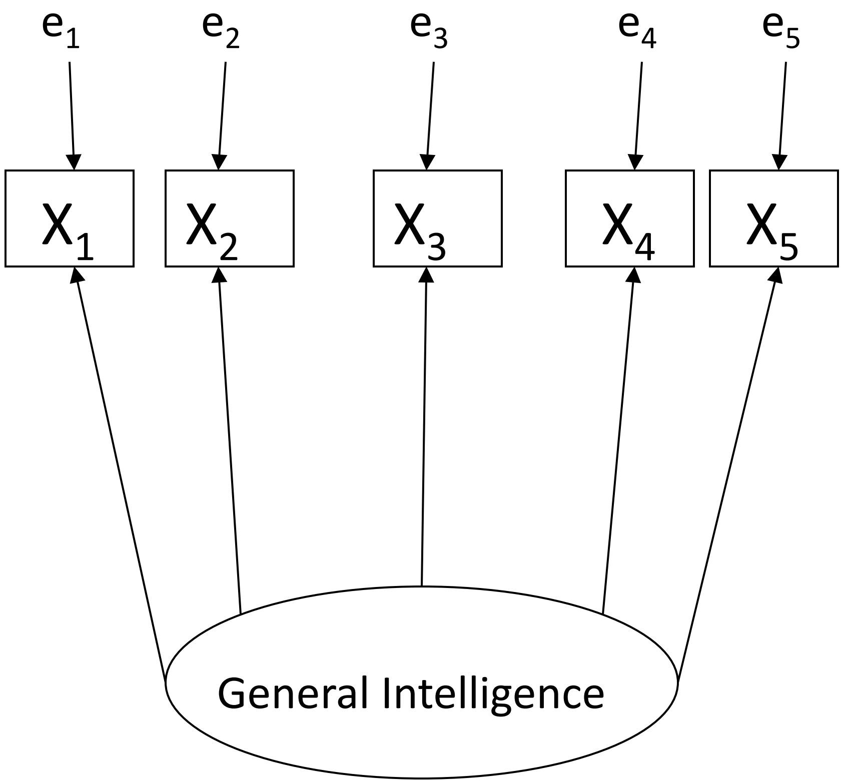
$$\begin{aligned}
Y &= \gamma_1 X_1 + \gamma_2 X_2 + \epsilon \\
W_1 &= X_1 + e_1 \\
W_2 &= X_2 + e_2 \\
V &= Y + e_3
\end{aligned}$$

A Path Model with Measurement Error



$$
\begin{aligned}
Y_1 &= \gamma_1 X + \epsilon_1 \\
Y_2 &= \beta Y_1 + \gamma_2 X + \epsilon_2 \\
W &= X + e_1 \\
V_1 &= Y_1 + e_2 \\
V_2 &= Y_2 + e_3
\end{aligned}
$$

A Factor Analysis Model



$$X_1 = \lambda_1 F + e_1$$

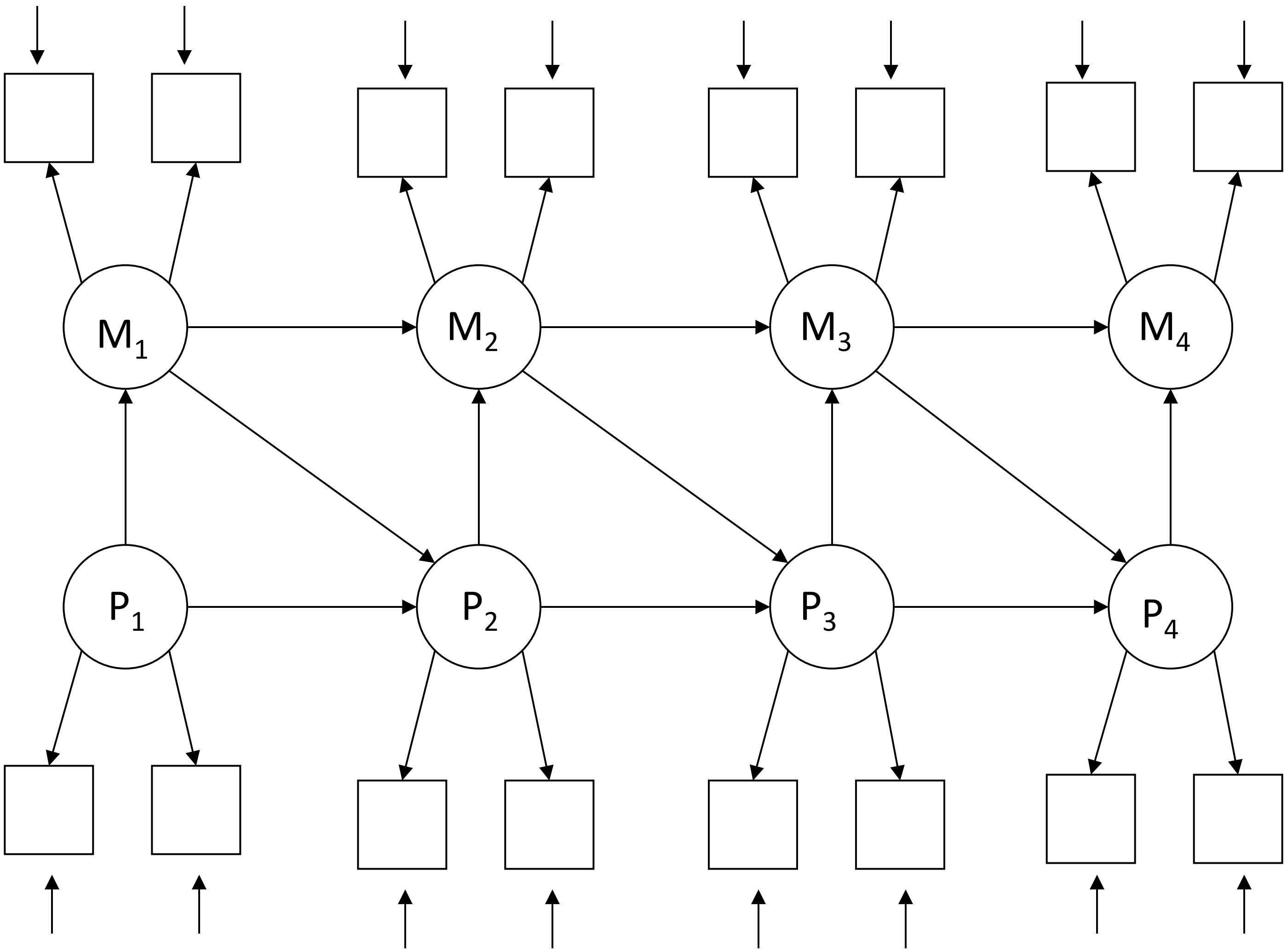
$$X_2 = \lambda_2 F + e_2$$

$$X_3 = \lambda_3 F + e_3$$

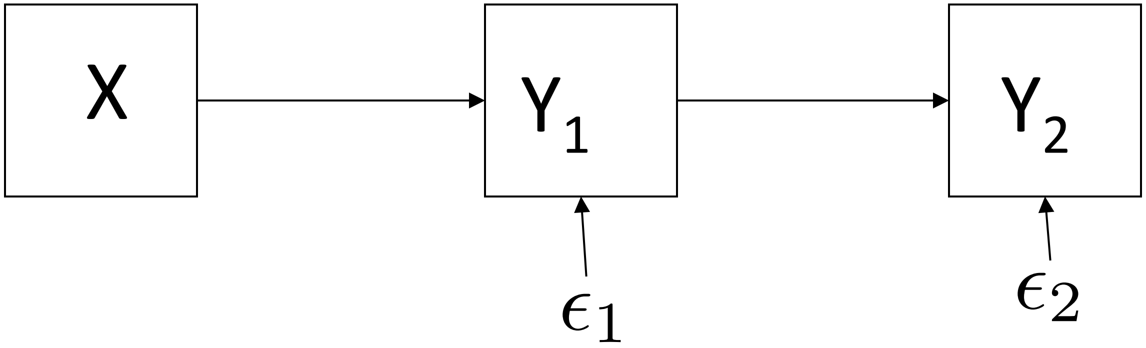
$$X_4 = \lambda_4 F + e_4$$

$$X_5 = \lambda_5 F + e_5$$

A Longitudinal Model



Estimation and Testing as Before



$$
\begin{aligned}
Y_1 &= \gamma X + \epsilon_1 \\
Y_2 &= \beta Y_1 + \epsilon_2
\end{aligned}
$$

All expected values equal zero.

$V(X) = \phi$, $V(\epsilon_1) = \psi_1$, $V(\epsilon_2) = \psi_2$,

$X, \epsilon_1, \epsilon_2$ are all independent.

Everything is normal.

Distribution of the data

$\begin{bmatrix} X_1 \\ Y_{1,1} \\ Y_{1,2} \end{bmatrix} \cdots \begin{bmatrix} X_n \\ Y_{n,1} \\ Y_{n,2} \end{bmatrix}$ are independent normal with mean zero

and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \phi & \gamma\phi & \beta\gamma\phi \\ \gamma\phi & \gamma^2\phi + \psi_1 & \beta(\gamma^2\phi + \psi_1) \\ \beta\gamma\phi & \beta(\gamma^2\phi + \psi_1) & \beta^2(\gamma^2\phi + \psi_1) + \psi_2 \end{bmatrix}$$

$$\boldsymbol{\theta} = (\gamma, \beta, \phi, \psi_1, \psi_2)$$

Maximum Likelihood

$$L(\mathbf{\Sigma}) = |\mathbf{\Sigma}|^{-n/2}(2\pi)^{-nk/2}\exp -\frac{n}{2}\left\{tr(\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1})\right\}$$

Minimize the “Objective Function”

$$\log|\mathbf{\Sigma}(\boldsymbol{\theta})| - \log|\widehat{\mathbf{\Sigma}}| + tr(\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}) - k$$

Tests

- Z tests for H_0: Parameter = 0 are produced by default
- “Chi-square” = (n-1) * Final value of objective function is the standard test for goodness of fit. Multiply by n instead of n-1 to get a true likelihood ratio test .
- Consider two nested models. One is more constrained (restricted) than the other. Then n * the difference in final objective functions is the large-sample likelihood ratio test, df = number of (linear) restrictions on the parameter.
- Other tests (for example Wald tests) are possible too.

A General Two-Stage Model

$$
\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\
\mathbf{D}_i &= \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i
\end{aligned}
$$

- $\mathbf{D}_i$ (the data) are observable. All other variables are latent.
- $\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$ is called the *Latent Variable Model*
- The latent vectors $\mathbf{X}_i$ and $\mathbf{Y}_i$ are collected into a "factor" $\mathbf{F}_i$. This is *not* a categorical independent variable, the usual meaning of factor in experimental design.
- $\mathbf{D}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i$ is called the *Measurement Model.*

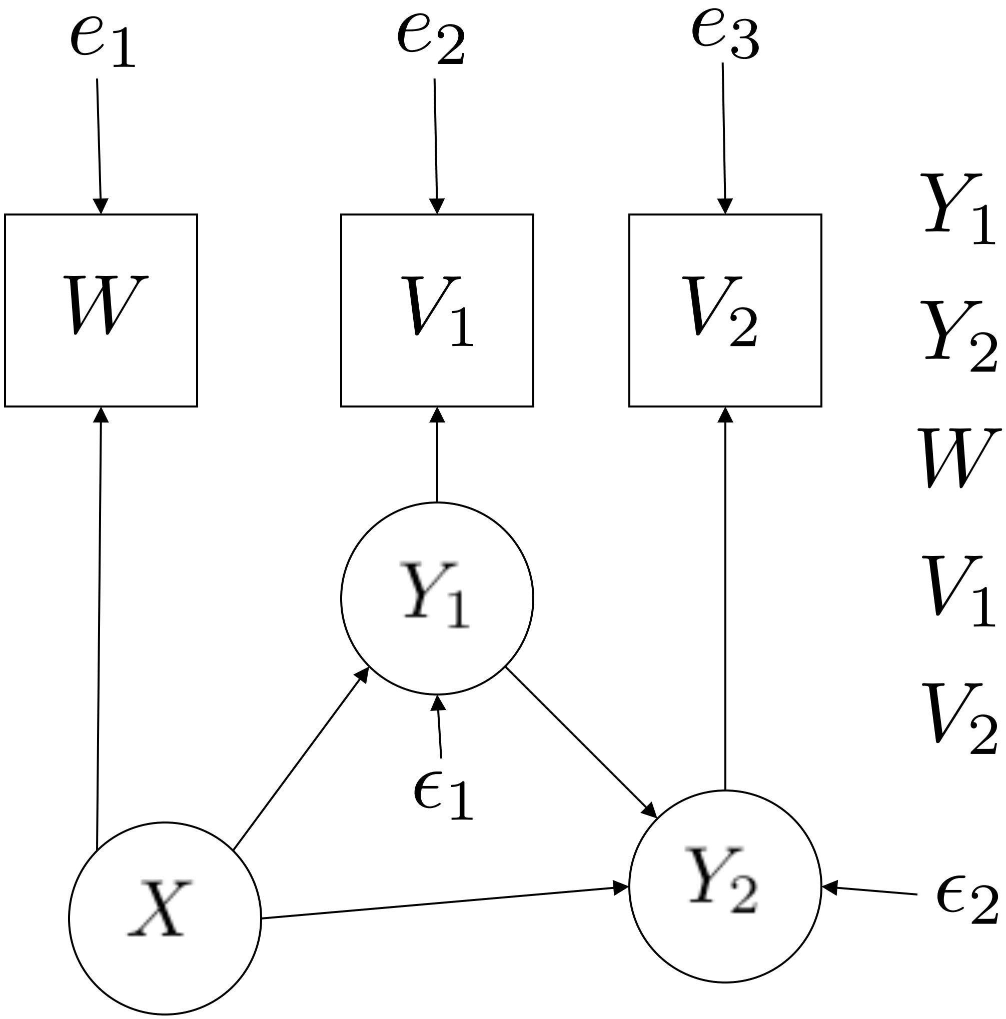
More Details

- $\mathbf{Y}_i$ is a $q \times 1$ random vector.
- $\boldsymbol{\beta}$ is a $q \times q$ matrix of constants with zeros on the main diagonal.
- $\mathbf{\Gamma}$ is a $q \times p$ matrix of constants.
- $\mathbf{X}_i$ is a $p \times 1$ random vector.
- $\boldsymbol{\epsilon}_i$ is a $q \times 1$ random vector.
- $\mathbf{F}_i$ (F for Factor) is just $\mathbf{X}_i$ stacked on top of $\mathbf{Y}_i$. It is a $(p+q) \times 1$ random vector.
- $\mathbf{D}_i$ is a $k \times 1$ random vector. Sometimes, $\mathbf{D}_i = \left(\begin{array}{c} \mathbf{W}_i \\ \mathbf{V}_i \end{array} \right)$
- $\mathbf{\Lambda}$ is a $k \times (p+q)$ matrix of constants.
- $\mathbf{D}_i$ is a $k \times 1$ random vector.
- $\mathbf{e}_i$ is a $k \times 1$ random vector.
- $\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ are independent.

$$
\begin{array}{rcl}
\mathbf{Y}_i & = & \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i & = & \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\
\mathbf{D}_i & = & \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i
\end{array}
$$

- $V(\mathbf{X}_i) = \boldsymbol{\Phi}_{11}$
- $V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$
- $V(\mathbf{F}_i) = V \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) = \left(\begin{array}{cc} V(\mathbf{X}_i) & C(\mathbf{X}_i, \mathbf{Y}_i) \\ C(\mathbf{Y}_i, \mathbf{X}_i) & V(\mathbf{Y}_i) \end{array} \right) = \boldsymbol{\Phi} = \left(\begin{array}{cc} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}_{12}' & \boldsymbol{\Phi}_{22} \end{array} \right)$
- $V(\mathbf{e}_i) = \boldsymbol{\Omega}$
- $V(\mathbf{D}_i) = \boldsymbol{\Sigma}$

Recall the example



$$\begin{aligned}
Y_1 &= \gamma_1 X + \epsilon_1 \\
Y_2 &= \beta Y_1 + \gamma_2 X + \epsilon_2 \\
W &= X + e_1 \\
V_1 &= Y_1 + e_2 \\
V_2 &= Y_2 + e_3
\end{aligned}$$

$$\begin{aligned}\mathbf{Y} &= \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\ \mathbf{D} &= \boldsymbol{\Lambda}\mathbf{F} + \mathbf{e}\end{aligned}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} + \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} X + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

$$\begin{pmatrix} W \\ V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y_1 \\ Y_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

- $V(\mathbf{X}) = \boldsymbol{\Phi}_{11} = \phi$
- $V(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix}$
- $V(\mathbf{e}) = \boldsymbol{\Omega} = \begin{pmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ 0 & 0 & \omega_3 \end{pmatrix}$

Observable variables in the latent variable model (fairly common)

- These present no problem
- Let $P(e_j=0) = 1$, so $Var(e_j) = 0$
- And $Cov(e_i,e_j)=0$ because if $P(e_j=0) = 1$

$$\begin{aligned} Cov(e_i, e_j) &= E(e_i e_j) - E(e_i)E(e_j) \\ &= E(e_i \cdot 0) - E(e_i) \cdot 0 \\ &= 0 - 0 = 0 \end{aligned}$$

- So in the covariance matrix **Ω**=V(**e**), just set $\omega_{ij} = \omega_{ji} = 0$, i=1,…,k

What should you be able to do?

- Given a path diagram, write the model equations and say which exogenous variables are correlated with each other.
- Given the model equations and information about which exogenous variables are correlated with each other, draw the path diagram.
- Given either piece of information, write the model in matrix form and say what all the matrices are.
- Calculate model covariance matrices
- Check identifiability

Recall the notation

$$\begin{aligned}\mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\ \mathbf{F}_i &= \left(\begin{array}{c}\mathbf{X}_i \\ \mathbf{Y}_i\end{array}\right) \\ \mathbf{D}_i &= \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i\end{aligned}$$

- $V(\mathbf{X}_i) = \boldsymbol{\Phi}_{11}$
- $V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$
- $V(\mathbf{F}_i) = V\left(\begin{array}{c}\mathbf{X}_i \\ \mathbf{Y}_i\end{array}\right) = \left(\begin{array}{cc}V(\mathbf{X}_i) & C(\mathbf{X}_i,\mathbf{Y}_i) \\ C(\mathbf{Y}_i,\mathbf{X}_i) & V(\mathbf{Y}_i)\end{array}\right) = \boldsymbol{\Phi} = \left(\begin{array}{cc}\boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}_{12}^{\prime} & \boldsymbol{\Phi}_{22}\end{array}\right)$
- $V(\mathbf{e}_i) = \boldsymbol{\Omega}$
- $V(\mathbf{D}_i) = \boldsymbol{\Sigma}$

For the latent variable model, calculate **Φ** = V(**F**)

$$
\begin{aligned}
& \mathbf{Y} = \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & \mathbf{Y} - \boldsymbol{\beta}\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & \mathbf{I}\mathbf{Y} - \boldsymbol{\beta}\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})^{-1}(\mathbf{I} - \boldsymbol{\beta})\mathbf{Y} = (\mathbf{I} - \boldsymbol{\beta})^{-1}(\boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\epsilon}) \\
\Rightarrow \quad & \mathbf{Y} = (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\mathbf{X} + (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\epsilon}
\end{aligned}
$$

So,

$$
\begin{aligned}
V(\mathbf{Y}) &= V\Big((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\mathbf{X} + (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\epsilon}\Big) \\
&= V\Big((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\mathbf{X}\Big) + V\Big((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\epsilon}\Big) \\
&= (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\, V(\mathbf{X}) \Big((\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\Big)' + (\mathbf{I} - \boldsymbol{\beta})^{-1} V(\boldsymbol{\epsilon})(\mathbf{I} - \boldsymbol{\beta})^{-1\prime} \\
&= (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\ \boldsymbol{\Phi}_{11}\boldsymbol{\Gamma}'(\mathbf{I} - \boldsymbol{\beta})^{-1\prime} + (\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \boldsymbol{\beta})^{-1\prime}
\end{aligned}
$$

For the measurement model, calculate **Σ** = V(**D**)

$$\begin{aligned}
\mathbf{D} &= \boldsymbol{\Lambda}\mathbf{F} + \mathbf{e} \\
\Rightarrow V(\mathbf{D}) &= V(\boldsymbol{\Lambda}\mathbf{F} + \mathbf{e}) \\
&= V(\boldsymbol{\Lambda}\mathbf{F}) + V(\mathbf{e}) \\
&= \boldsymbol{\Lambda}V(\mathbf{F})\boldsymbol{\Lambda}' + V(\mathbf{e}) \\
&= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Omega} \\
&= \boldsymbol{\Sigma}
\end{aligned}$$

Two-stage Proofs of Identifiability

- Show the parameters of the latent variable model ($\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}_{11}$, $\boldsymbol{\Psi}$) can be recovered from $\boldsymbol{\Phi}$ = V(**F**).
- Show the parameters of the measurement model ($\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Omega}$) can be recovered from $\boldsymbol{\Sigma}$= V(**D**).
- This means *all* the parameters can be recovered from $\boldsymbol{\Sigma}$.
- Break a big problem into two smaller ones.
- Develop *rules* for checking identifiability at each stage.