The General Structural Equation Model[1]

STA2101 F19

[1]See last slide for copyright information.

Features of Structural Equation Models

- Multiple equations.
- All the variables are random.
- An explanatory variable in one equation can be the response variable in another equation.
- Models are represented by path diagrams.
- Identifiability is always an issue.
- The statistical models are explicitly models of influence. They are often called *causal models*.

Correlation versus Causation

- The path diagrams deliberately imply influence. If $A \rightarrow B$, we are saying A *contributes* to B, or partly *causes* it.
- Data are usually observational. The correlation-causation issue does not go away.
- You may be able to argue on theoretical grounds that $A \rightarrow B$ is more believable than $B \rightarrow A$.
- A statistical model cannot "prove" causality, but if you have a causal model, you may be able to test whether it's compatible with the data.

Vocabulary

- **Exogenous variable**: In the regression-like equations of a structural equation model, the exogenous variables are ones that appear *only* on the right side of the equals sign, and never on the left side in any equation. If you think of Y being a function of X, this is one way to remember the meaning of **ex**ogenous.
- **Endogenous variable**: Endogenous variables are those that appear on the left side of at least one equals sign. Endogenous variables depend on the exogenous varables, and possibly other endogenous variables. Think of an arrow from an exogenous variable to an endogenous variable. The **end** of the arrow is pointing at the **end**ogenous variable.
- **Factor**: This term has a meaning that actually conflicts with its meaning in mainstream Statistics, particularly in experimental design. A *factor* is an underlying trait or characteristic that cannot be measured directly, like intelligence. It is a latent variable, period.

Notation

$$\begin{aligned} Y_{i,1} &= \alpha_1 + \gamma_1 X_{i,1} + \gamma_2 X_{i,2} + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta Y_{i,1} + \epsilon_{i,2} \end{aligned}$$

- Regression coefficients (links between exogenous variables and endogenous variables) are now called gamma instead of beta.
- Betas are used for links between endogenous variables.
- Intercepts are alphas but they will soon disappear.

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 $\boldsymbol{\Sigma}$ of an observable data vector. $\boldsymbol{\Sigma} = cov(\mathbf{D}_i)$.
- Denote the vector of parameters that appear in $\boldsymbol{\Sigma}$ by $\boldsymbol{\theta}$.
- Re-parameterize. The new parameter vector is $(\boldsymbol{\theta}, \boldsymbol{\mu})$, where $\boldsymbol{\mu} = E(\mathbf{D}_i)$.
- Estimate $\boldsymbol{\mu}$ with $\overline{\mathbf{D}}$, forget it, and concentrate on $\boldsymbol{\theta}$.
- From this point on the models *seemingly* have zero means, and no intercepts.

A General Two-Stage Model

Stage 1 is the latent variable model and Stage 2 is the measurement model.

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

- $\mathbf{D}_i$ (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 explanatory 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*.

$$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \qquad \mathbf{F}_i = \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} \qquad \mathbf{D}_i = \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i$$

- $\mathbf{Y}_i$ is a $q \times 1$ random vector.
- $\boldsymbol{\beta}$ is a $q \times q$ matrix of constants with zeros on the main diagonal.
- $\mathbf{X}_i$ is a $p \times 1$ random vector.
- $\boldsymbol{\Gamma}$ is a $q \times p$ matrix of constants.
- $\boldsymbol{\epsilon}_i$ 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, $\mathbf{D}_i = \begin{pmatrix} \mathbf{W}_i \\ \mathbf{V}_i \end{pmatrix}$.
- $\boldsymbol{\Lambda}$ is a $k \times (p+q)$ matrix of constants: "factor loadings."
- $\mathbf{e}_i$ is a $k \times 1$ random vector.
- $\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ are independent.

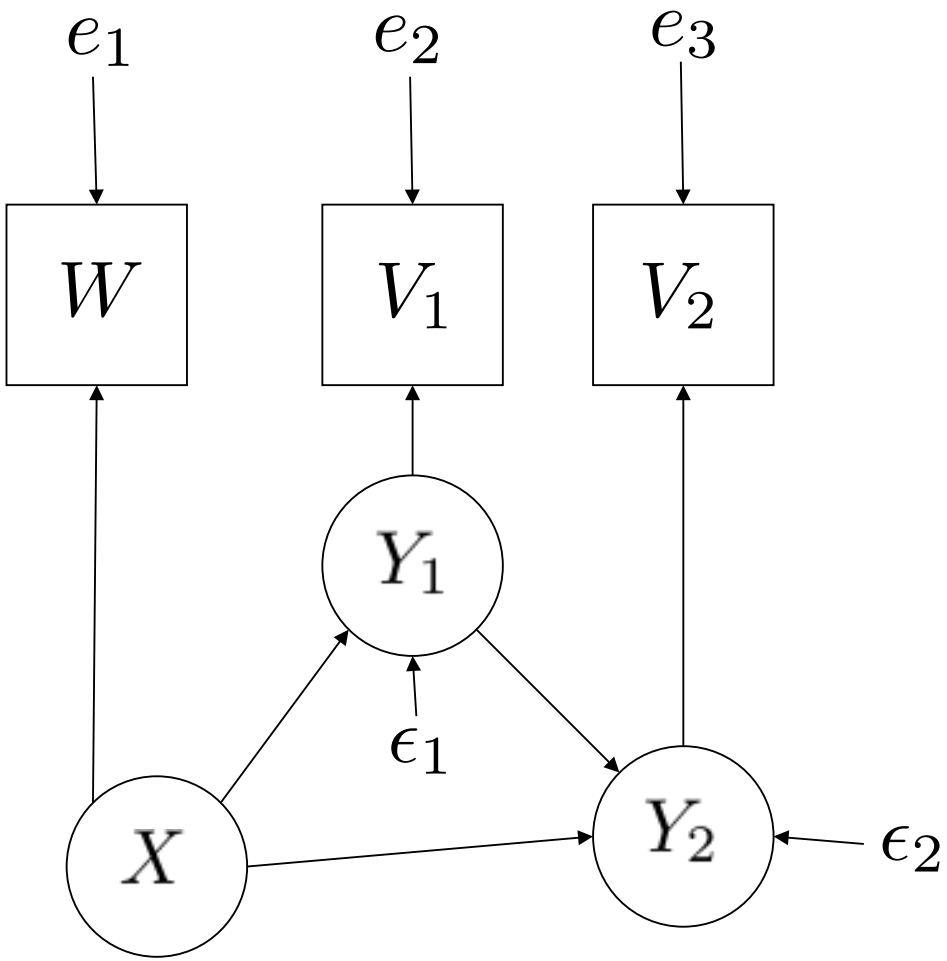
Covariance matrices

All assumed positive definite unless otherwise specified

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

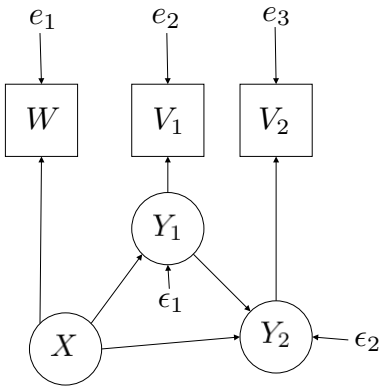
$$
\begin{aligned}
cov(\mathbf{X}_i) &= \boldsymbol{\Phi}_x \\
cov(\boldsymbol{\epsilon}_i) &= \boldsymbol{\Psi} \\
cov(\mathbf{F}_i) &= \boldsymbol{\Phi} = \begin{pmatrix} cov(\mathbf{X}_i) & cov(\mathbf{X}_i, \mathbf{Y}_i) \\ cov(\mathbf{Y}_i, \mathbf{X}_i) & cov(\mathbf{Y}_i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}_{12}^{\top} & \boldsymbol{\Phi}_{22} \end{pmatrix} \\
cov(\mathbf{e}_i) &= \boldsymbol{\Omega} \\
cov(\mathbf{D}_i) &= \boldsymbol{\Sigma}
\end{aligned}
$$

Example: A Path Model with Measurement Error



$$
\begin{aligned}
Y_{i,1} &= \gamma_1 X_i + \epsilon_{i,1} \\
Y_{i,2} &= \beta Y_{i,1} + \gamma_2 X_i + \epsilon_{i,2} \\
W_i &= X_i + e_{i,1} \\
V_{i,1} &= Y_{i,1} + e_{i,2} \\
V_{i,2} &= Y_{i,2} + e_{i,3}
\end{aligned}
$$

Matrix Form



$$
\begin{aligned}
Y_{i,1} &= \gamma_1 X_i + \epsilon_{i,1} \\
Y_{i,2} &= \beta Y_{i,1} + \gamma_2 X_i + \epsilon_{i,2} \\
W_i &= X_i + e_{i,1} \\
V_{i,1} &= Y_{i,1} + e_{i,2} \\
V_{i,2} &= Y_{i,2} + e_{i,3}
\end{aligned}
\qquad
\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} \\
\mathbf{D}_i &= \boldsymbol{\Lambda}\mathbf{F}_i + \mathbf{e}_i
\end{aligned}
$$

$$
\begin{array}{ccccccccc}
\mathbf{Y}_i & = & \boldsymbol{\beta} & & \mathbf{Y}_i & + & \boldsymbol{\Gamma} & \mathbf{X}_i & + & \boldsymbol{\epsilon}_i \\
\begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} & = & \begin{pmatrix} 0 & 0 \\ \beta & 0 \end{pmatrix} & & \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} & + & \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} & X_i & + & \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix}
\end{array}
$$

$$
\begin{array}{ccccccc}
\mathbf{D}_i & = & \boldsymbol{\Lambda} & & \mathbf{F}_i & + & \mathbf{e}_i \\
\begin{pmatrix} W_i \\ V_{i,1} \\ V_{i,2} \end{pmatrix} & = & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} & \begin{pmatrix} X_i \\ Y_{i,1} \\ Y_{i,2} \end{pmatrix} & + & \begin{pmatrix} e_{i,1} \\ e_{i,2} \\ e_{i,3} \end{pmatrix}
\end{array}
$$

Observable variables in the "latent" variable model

$\mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i$

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 $P(e_j = 0) = 1$,

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

- In $\boldsymbol{\Omega} = cov(\mathbf{e}_i)$, column j (and row j) are all zeros.
- $\boldsymbol{\Omega}$ singular, no problem.

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

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

$$
\begin{aligned}
cov(\mathbf{X}_i) &= \boldsymbol{\Phi}_x \\
cov(\boldsymbol{\epsilon}_i) &= \boldsymbol{\Psi} \\
cov(\mathbf{F}_i) &= \boldsymbol{\Phi} = \begin{pmatrix} cov(\mathbf{X}_i) & cov(\mathbf{X}_i, \mathbf{Y}_i) \\ cov(\mathbf{Y}_i, \mathbf{X}_i) & cov(\mathbf{Y}_i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}_{12}^{\top} & \boldsymbol{\Phi}_{22} \end{pmatrix} \\
cov(\mathbf{e}_i) &= \boldsymbol{\Omega} \\
cov(\mathbf{D}_i) &= \boldsymbol{\Sigma}
\end{aligned}
$$

Calculate a general expression for $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

For the latent variable model, calculate $\mathbf{\Phi} = cov(\mathbf{F}_i)$

Have $cov(\mathbf{X}_i) = \mathbf{\Phi}_x$, need $cov(\mathbf{Y}_i)$ and $cov(\mathbf{X}_i, \mathbf{Y}_i)$

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

So,

$$
\begin{aligned}
cov(\mathbf{Y}_i) &= (\mathbf{I} - \boldsymbol{\beta})^{-1} cov(\mathbf{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i)(\mathbf{I} - \boldsymbol{\beta})^{-1\top} \\
&= (\mathbf{I} - \boldsymbol{\beta})^{-1} \left(cov(\mathbf{\Gamma}\mathbf{X}_i) + cov(\boldsymbol{\epsilon}_i)\right) (\mathbf{I} - \boldsymbol{\beta}^\top)^{-1} \\
&= (\mathbf{I} - \boldsymbol{\beta})^{-1} \left(\mathbf{\Gamma}\mathbf{\Phi}_x\mathbf{\Gamma}^\top + \mathbf{\Psi}\right) (\mathbf{I} - \boldsymbol{\beta}^\top)^{-1}
\end{aligned}
$$

Does $(\mathbf{I} - \boldsymbol{\beta})^{-1}$ exist?

$$
\begin{aligned}
& (\mathbf{I} - \boldsymbol{\beta})\mathbf{Y}_i = \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\Rightarrow \quad & cov\left((\mathbf{I} - \boldsymbol{\beta})\mathbf{Y}_i\right) = cov\left(\boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i\right) \\
\Rightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})cov(\mathbf{Y}_i)(\mathbf{I} - \boldsymbol{\beta})^\top = \boldsymbol{\Gamma}cov(\mathbf{X}_i)\boldsymbol{\Gamma}^\top + cov(\boldsymbol{\epsilon}_i) \\
\Rightarrow \quad & (\mathbf{I} - \boldsymbol{\beta})cov(\mathbf{Y}_i)(\mathbf{I} - \boldsymbol{\beta})^\top = \boldsymbol{\Gamma}\boldsymbol{\Phi}_x\boldsymbol{\Gamma}^\top + \boldsymbol{\Psi}
\end{aligned}
$$

Now let the $q \times 1$ constant vector $\mathbf{a} \neq \mathbf{0}$, and

$$
\begin{aligned}
\mathbf{a}^\top(\mathbf{I} - \boldsymbol{\beta})cov(\mathbf{Y}_i)(\mathbf{I} - \boldsymbol{\beta})^\top\mathbf{a} &= \mathbf{a}^\top\boldsymbol{\Gamma}\boldsymbol{\Phi}_x\boldsymbol{\Gamma}^\top\mathbf{a} + \mathbf{a}^\top\boldsymbol{\Psi}\mathbf{a} \\
&> 0
\end{aligned}
$$

Because $cov\left(\boldsymbol{\Gamma}\mathbf{X}_i\right) = \boldsymbol{\Gamma}\boldsymbol{\Phi}_x\boldsymbol{\Gamma}^\top$ is non-negative definite and $\boldsymbol{\Psi} = cov(\boldsymbol{\epsilon}_i)$ is positive definite. Hence, $(\mathbf{I} - \boldsymbol{\beta})cov(\mathbf{Y}_i)(\mathbf{I} - \boldsymbol{\beta})^\top$ is positive definite.

Have $(\mathbf{I} - \boldsymbol{\beta})cov(\mathbf{Y}_i)(\mathbf{I} - \boldsymbol{\beta})^\top$ positive definite

- So the $q \times q$ matrix $(\mathbf{I} - \boldsymbol{\beta})cov(\mathbf{Y}_i)(\mathbf{I} - \boldsymbol{\beta})^\top$ is full rank: rank is q.
- $(\mathbf{I} - \boldsymbol{\beta})$, $cov(\mathbf{Y}_i)$ and $(\mathbf{I} - \boldsymbol{\beta})^\top$ are all $q \times q$.
- Rank of a product is minimum rank.
- Hence Rank$(\mathbf{I} - \boldsymbol{\beta}) = q$, and the inverse exists. ■

$(\mathbf{I} - \boldsymbol{\beta})^{-1}$ exists if the rest of the model is correct

- This forms a strange and unexpected hole in the parameter space.
- No need to "assume" it, as all the textbooks do.
- For example, if $\boldsymbol{\beta} = \mathbf{I}$, then

$$\begin{aligned} & \mathbf{Y}_i = \boldsymbol{\beta}\mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\ \Rightarrow \quad & \mathbf{Y}_i = \mathbf{Y}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\epsilon}_i \\ \Rightarrow \quad & \boldsymbol{\epsilon}_i = -\boldsymbol{\Gamma}\mathbf{X}_i \end{aligned}$$

  Impossible if $\mathbf{X}_i$ and $\boldsymbol{\epsilon}_i$ are independent.
- Summary: The rest of the model places subtle restrictions on $\boldsymbol{\beta}$.

But we were in the middle of a covariance calculation.

For the measurement model, calculate $\mathbf{\Sigma} = cov(\mathbf{D}_i)$

$$
\begin{aligned}
\mathbf{D}_i &= \mathbf{\Lambda}\mathbf{F}_i + \mathbf{e}_i \\
\Rightarrow cov(\mathbf{D}_i) &= cov(\mathbf{\Lambda}\mathbf{F}_i + \mathbf{e}_i) \\
&= cov(\mathbf{\Lambda}\mathbf{F}_i) + cov(\mathbf{e}_i) \\
&= \mathbf{\Lambda} cov(\mathbf{F}_i)\mathbf{\Lambda}^\top + cov(\mathbf{e}_i) \\
&= \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\top + \mathbf{\Omega} \\
&= \mathbf{\Sigma}
\end{aligned}
$$

Two-stage Proofs of Identifiability

Stage 1 is the latent variable model and Stage 2 is the measurement model.

- Show the parameters of the latent variable model $(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}_x, \boldsymbol{\Psi})$ can be recovered from $\boldsymbol{\Phi} = cov(\mathbf{F}_i)$.
- Show the parameters of the measurement model $(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Omega})$ can be recovered from $\boldsymbol{\Sigma} = cov(\mathbf{D}_i)$.
- 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.
- Just look at the path diagram.

Copyright Information