JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Introduction to Structural Equations

Statistics for Psychosocial Research II:

Structural Models

Qian-Li Xue PhD

Assistant Professor of Medicine, Biostatistics, Epidemiology

Johns Hopkins Medical Institutions

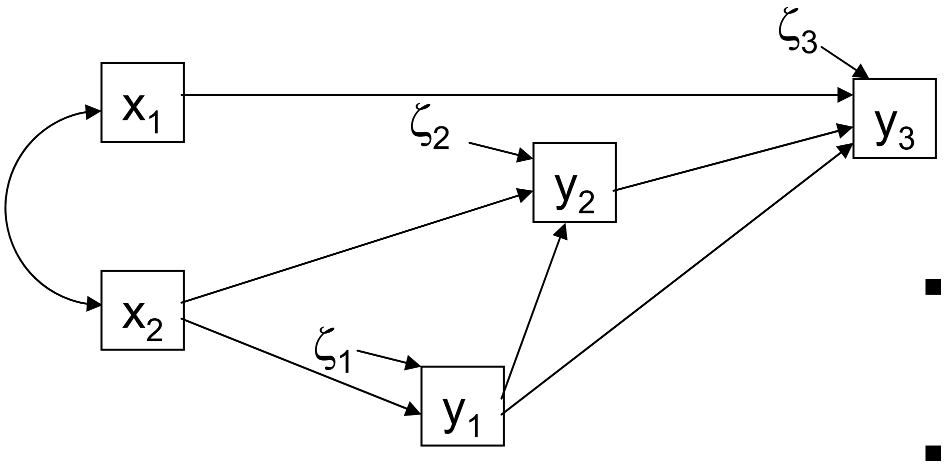
Course Overview

(1) Structural Regression/Path Analysis

(a) “effect mediation” versus “controlling for”

(b) causality

Course Overview: Structural Equation Models with Observed Variables



(McDonald and Clelland, 1984)

- Study Aim: union sentiment among southern nonunion textile workers
- y_1 – deference to managers
- y_2 – support for labor activism
- y_3 – sentiment toward unions
- x_1 – years in textile mill
- x_2 – age

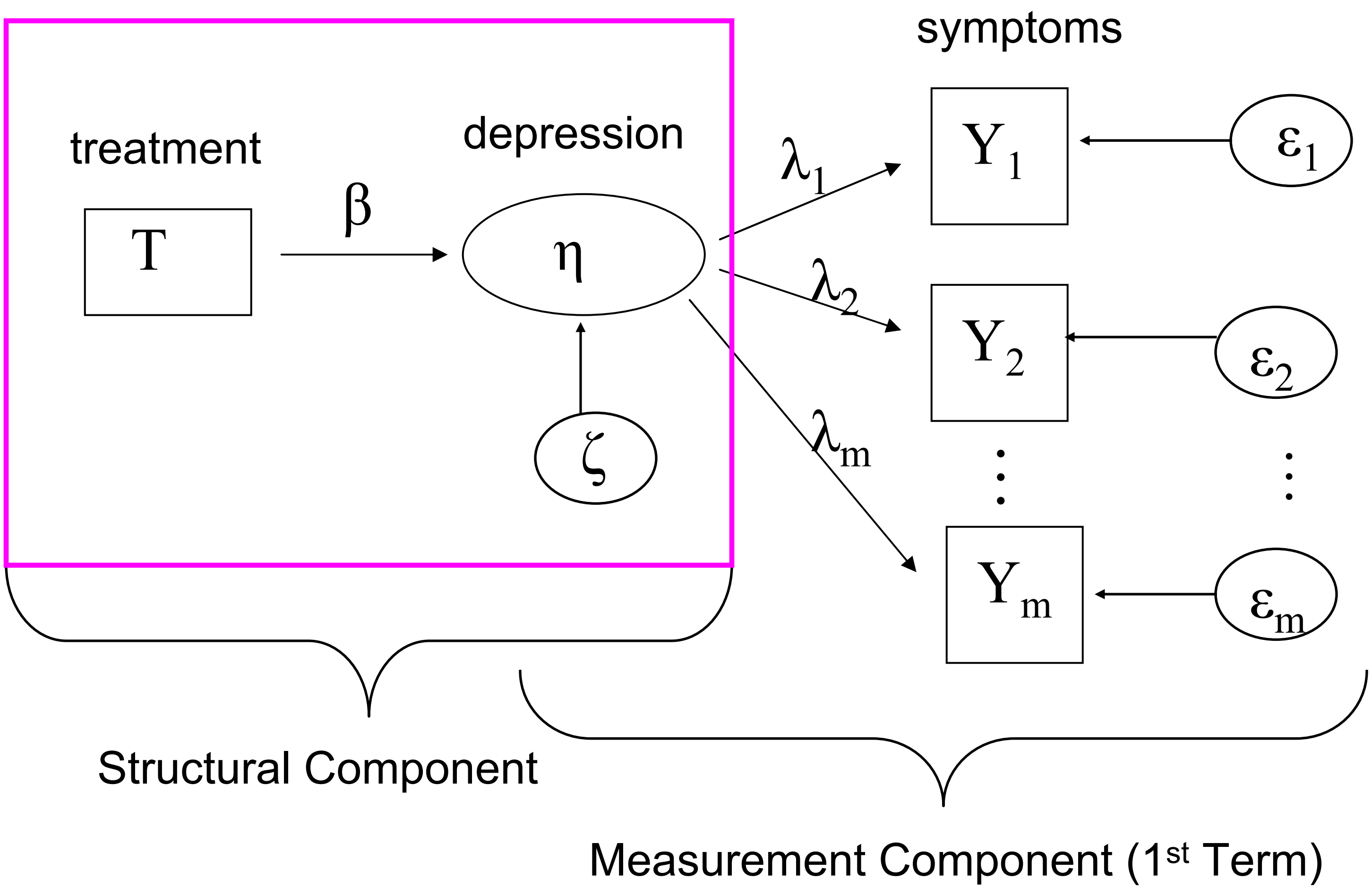
Course Overview: Structural Equation Models with Latent Variables

(2) Regression plus measurement structures from last term

- (a) if we ignore measurement, "item regression"
- (b) factor analysis: structural equations with latent variables
- (c) latent class analysis: latent class regression

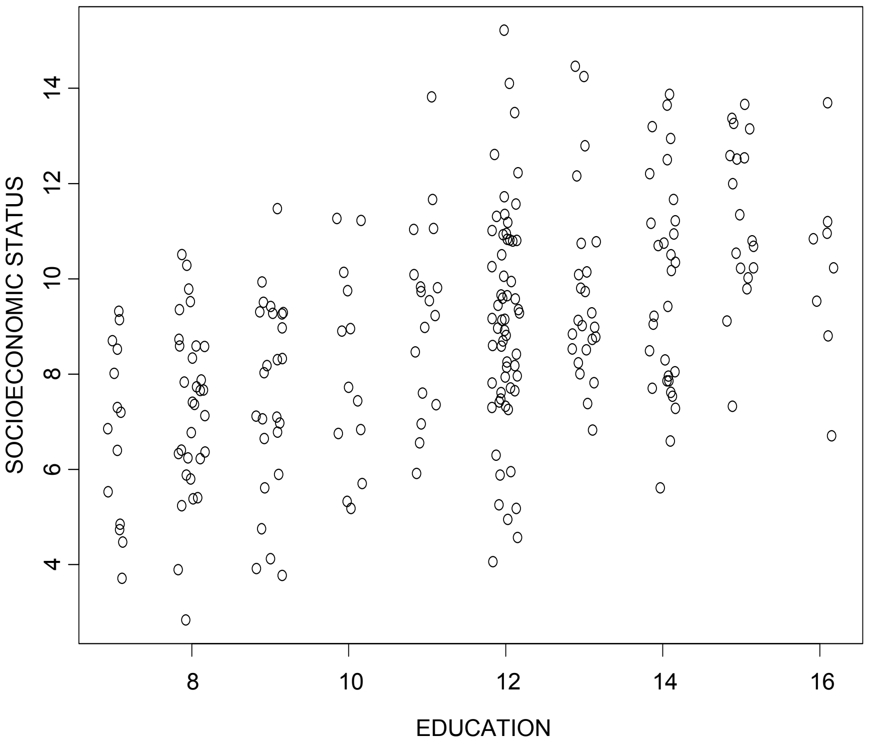
Course Overview:

Structural Equations = Structural + Measurement Components



General Idea

- How does outcome vary with predictors?
- Make inference on hypothesis about how predictors affect outcome
- Predict individual outcomes



Challenge

- How do we measure latent outcomes (and predictors)?
- There are multiple responses
- Approach 1:
  - $Y_1,\ldots,Y_n$ measure the same thing. Treat individually or summarize Y's.
- Approach 2:
  - Call ideal outcome η
  - If we <u>knew</u> η, then $\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots$
  - But we don't know it:
    - ❖ infer η from factor analysis or latent class analysis
    - ❖ regress η on X's

Three approaches to assessing association between covariates and multiple responses

(1) Summarize Then Analyze (STA)

(2) Analyze Then Summarize (ATS)

(3) Summarize AND Analyze: (SAA)

- Structural Equations
- 2 parts
  - measurement component
  - structural/regression component

Example: Depression Study
Summarize then Analyze (STA)

- Clinical trial of two anti-depressants
- Which anti-depressant is more effective for treating depression?
- Depression symptoms were based on the Hamilton Depression Rating Scale (HAM-D).

17 Symptoms

Depressed mood
Guilt feelings
suicide
Insomnia (x3)
Work and activities
Psychomotor retardation
agitation
anxiety
Somatic symptoms
…..

For each item, write the correct number on the line next to the item. (Only one response per item)

_____1. DEPRESSED MOOD (Sadness, hopeless, helpless, worthless)
0= Absent
1= These feeling states indicated only on questioning
2= These feeling states spontaneously reported
3= Communicates feeling states non-verbally—i.e., through facial expression, posture,   voice, and tendency to weep
4= Patient reports VIRTUALLY ONLY these feeling states in his spontaneous verbal and non-verbal communication

_____2. FEELINGS OF GUILT
0= Absent
1= Self reproach, feels he has let people down
2= Ideas of guilt or rumination over past errors or sinful deeds
3= Present illness is a punishment. Delusions of guilt
4= Hears accusatory or denunciatory voices and/or experiences threatening visual hallucinations

_____3. SUICIDE
0= Absent
1= Feels life is not worth living
2= Wishes he were dead or any thoughts of possible death to self
3= Suicidal ideas or gesture
4= Attempts at suicide (any serious attempt rates 4)

_____4. INSOMNIA EARLY
0= No difficulty falling asleep
1= Complains of occasional difficulty falling asleep—i.e., more than ½ hour
2= Complains of nightly difficulty falling asleep

_____5. INSOMNIA MIDDLE
0= No difficulty
1= patient complains of being restless and disturbed during the night
2= Waking during the night—any getting out of bed rates 2 (except for purposes of voiding)

_____6. INSOMNIA LATE
0= No difficulty
1= Waking in early hours of the morning but goes back to sleep
2= Unable to fall asleep again if he gets out of bed

Example:
Summarize then Analyze (STA)

- Summarize:
  - Add up the number of symptoms, or "score" the HAM-D.
  - Treat the score as "fixed" or "observed" outcome.
  - But, we know better! It is not measured perfectly.
  - What is the reliability of the HAM-D???
- Analyze: See how the outcome relates to predictor (i.e., treatment)

Summarize Then Analyze

Sum up HAM-D score pre and post and take difference:

Pre-treatment score: $Y_{i1} = Y_{i1,1} + Y_{i1,2} + \cdots + Y_{i1,21}$

Post-treatment score: $Y_{i2} = Y_{i2,1} + Y_{i2,2} + \cdots + Y_{i2,21}$

Difference: $D_i = Y_{i2} - Y_{i1}$

Evaluate association with Y_i and treatment

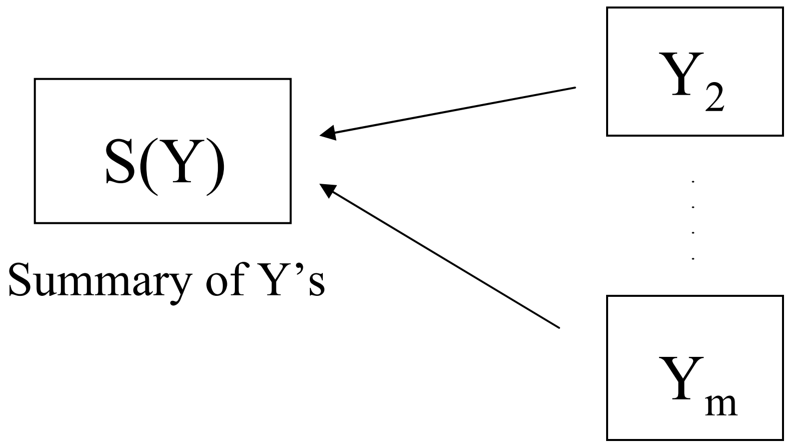
$$D_i = \beta_0 + \beta_1 trt_i$$

where trt_i = 1 of treatment A, and 0 if treatment B

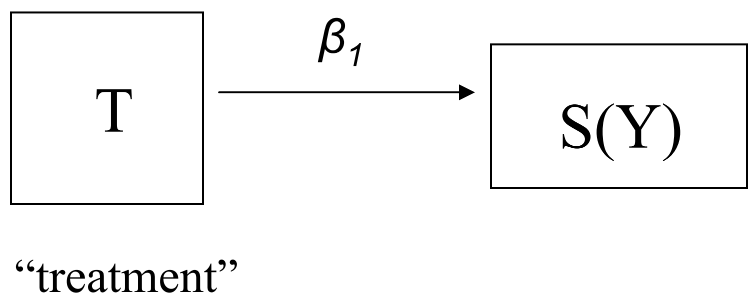
Make inference about treatment effect based on β_1

STA: Two models estimated separately

Model 1:



Model 2:



STA: so what is the problem???

- We are ignoring that S(Y) is measured with error.
- Note that that S(Y) has reliability less than 1.
- In our example: S(Y) represents an "imperfect measure" of depression with reliability of about 0.88 (depending on population).
- Aren't we then overestimating the variation in our outcome by using S(Y)?
- Recall: $Var(T_x) < Var(X)$, T_x is the true score of x
- What effect might that have on the standard error of β_1?

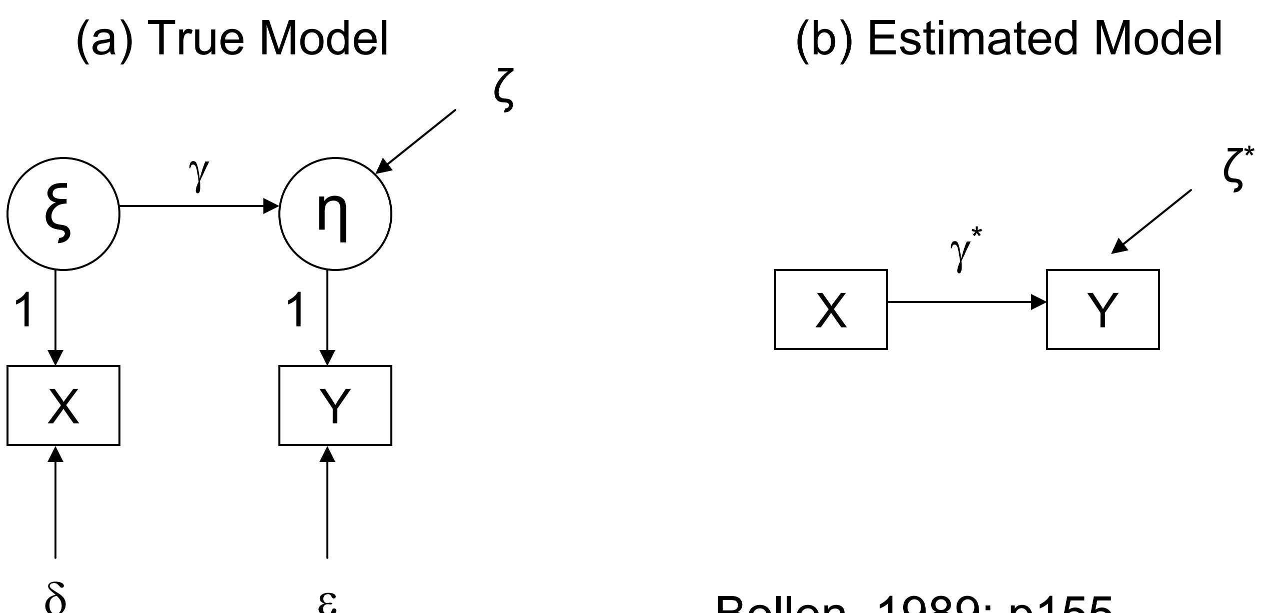
The Consequences of Measurement Error

True Model:

$$x = \xi + \delta$$

$$y = \eta + \varepsilon$$

$$\eta = \gamma\xi + \zeta$$



Bollen, 1989: p155

The Consequences of Measurement Error

True Model: $x = \xi + \delta$

$$y = \eta + \varepsilon$$

$$\eta = \gamma\xi + \zeta$$

$$\text{cov}(\xi, \eta) = \text{cov}(\xi, \gamma\xi + \zeta) = \gamma\phi$$

$$\text{cov}(x, y) = \text{cov}(\xi + \delta, \eta + \varepsilon) = \gamma\phi$$

$$\gamma^* = \frac{\text{cov}(x, y)}{\text{var}(x)} = \gamma\left[\frac{\phi}{\text{var}(x)}\right] = \gamma\rho_{xx}$$

$$\rho_{xx} = \text{reliability coefficient}$$

$$\text{Therefore, } |\gamma^*| < |\gamma|$$

Note: γ Is not affected by ρ_{yy}!

The Consequences of Measurement Error

True Model: $x = \xi + \delta$

$$y = \eta + \varepsilon$$

$$\eta = \gamma\xi + \zeta$$

$$\rho_{\xi\eta} = \frac{\rho_{xy}}{\sqrt{\rho_{xx}\rho_{yy}}}$$

correction for attenuation of correlation coefficient

Also, it can be shown that

$$\rho_{xy}^2 = \rho_{xx}\rho_{yy}\rho_{\xi\eta}^2$$

i.e. the squared correlation between the two observed measures is attenuated relative to the latent variables whenever the reliability of x or y is less than 1!

The Consequences of Measurement Error

In the case of multiple regression, the following is no longer true

$$|\gamma^*| < |\gamma|$$

However, the following still holds:

$$R^2 \geq R^{*2},$$

where R^2 and R^{*2} are the squared mulitple correlation coefficents for the regressions containing variables without and with measurement error, respectively.

Another Approach: Analyze Then Summarize (ATS)

Analyze: for each of the 21 items in the HAM-D, see if treatment is associated with improvement.

1. Define outcome per item:

$$D_{i,1} = Y_{i2,1} - Y_{i1,1}$$

$$\vdots$$

$$D_{i,21} = Y_{i2,21} - Y_{i1,21}$$

2. Estimate association per item with treatment:

$$D_{i,1} = \beta_{0,1} + \beta_{1,1} trt_i$$

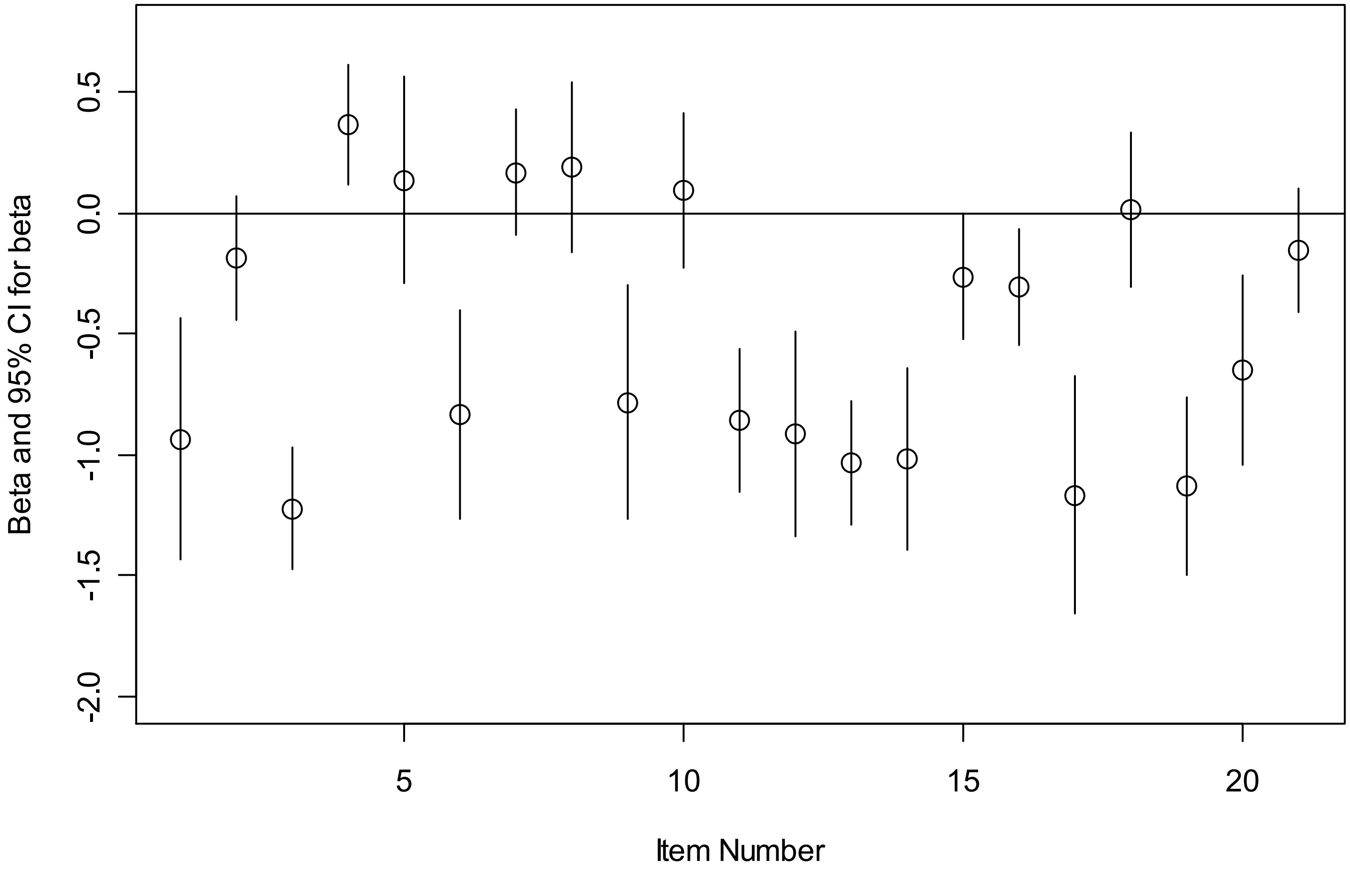
$$D_{i,2} = \beta_{0,2} + \beta_{1,2} trt_i$$

$$\vdots$$

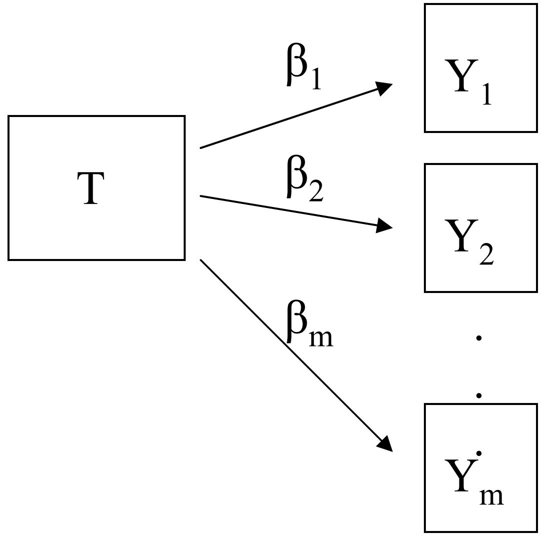
$$D_{i,21} = \beta_{0,21} + \beta_{1,21} trt_i$$

Another Approach: Analyze Then Summarize (ATS)

2. <u>Summarize</u>: Qualitatively or quantitatively evaluate the associations



Analyze then Summarize



Fit *m* regressions to individually describe association between T and each Y.

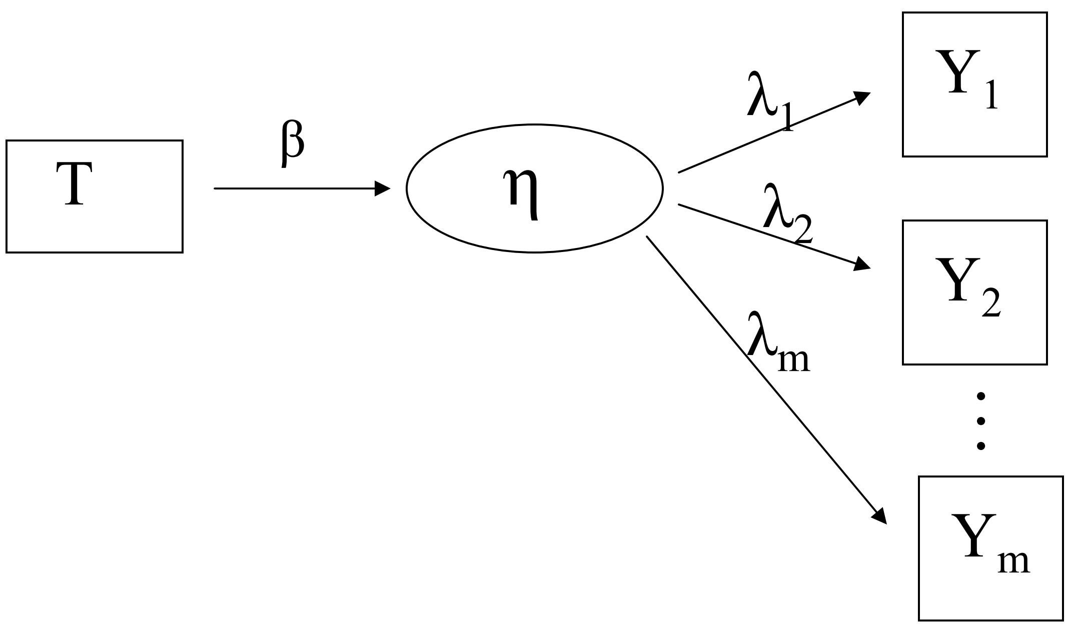
Then summarize associations.

So what is wrong with ATS?

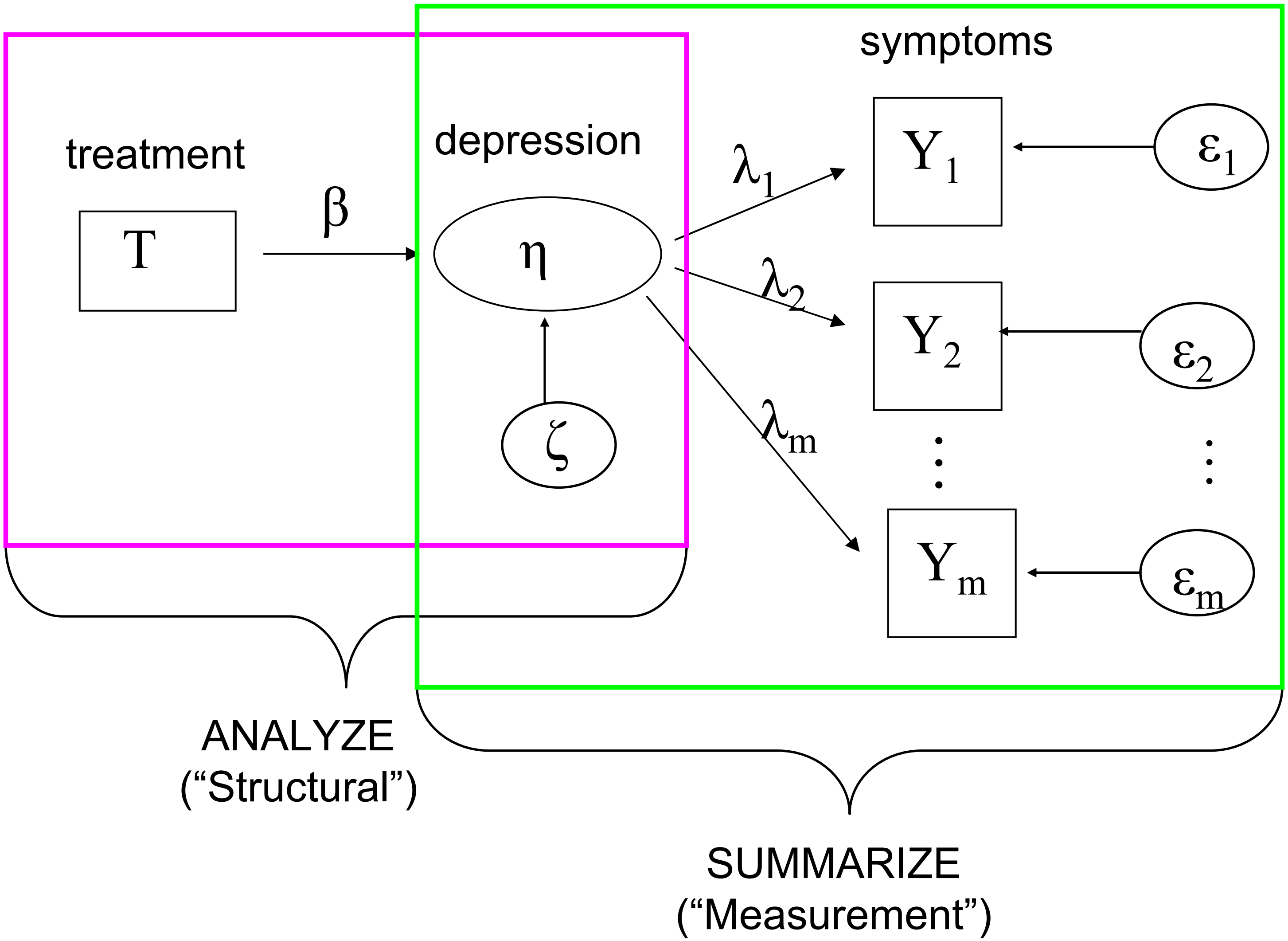
- How do we answer the question: “Which treatment works better?”
- We get individual answers.
- Often hard to summarize after the analysis has been done.
- (More about this in ‘Item Regression lecture’)

Summarize and Analyze Simultaneously (SAA)

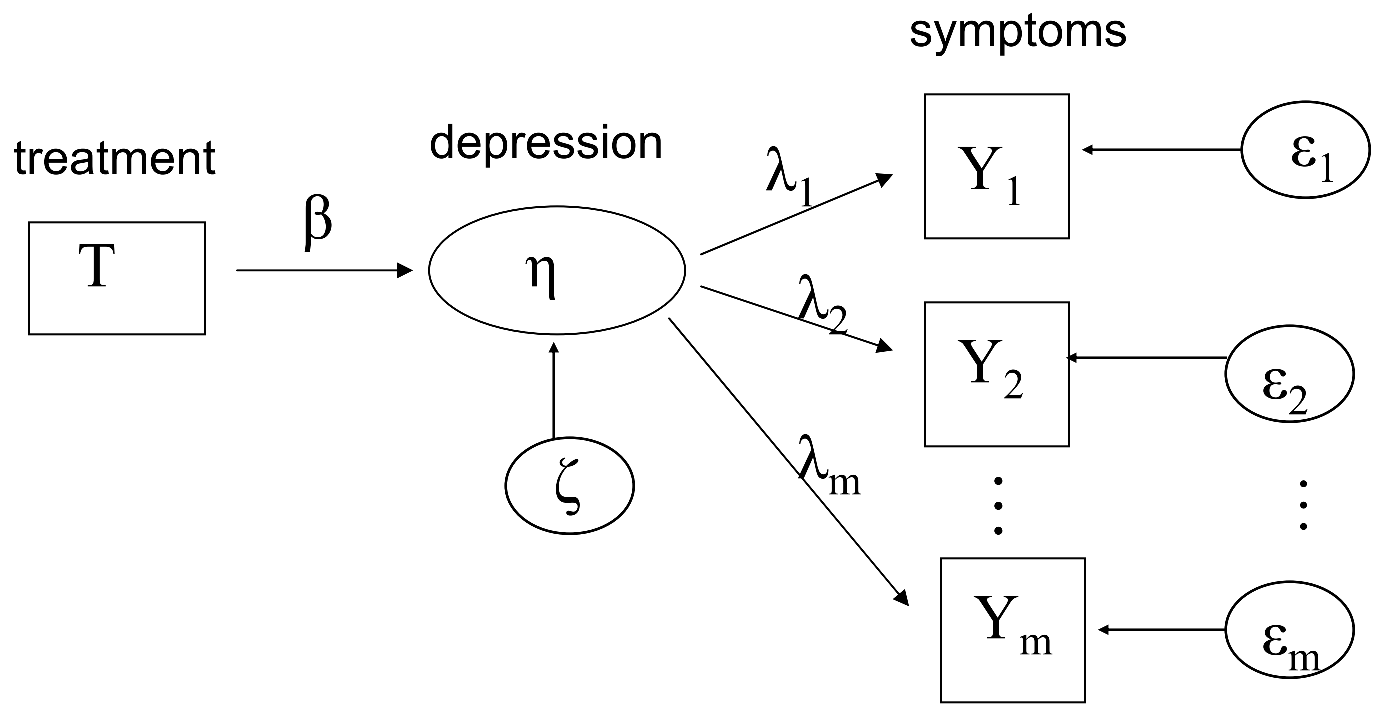
- Fit 'summarize' and 'analyze' components at the same time.
- One big model
- Accounts for measurement error via latent variable



Summarize and Analyze Simultaneously



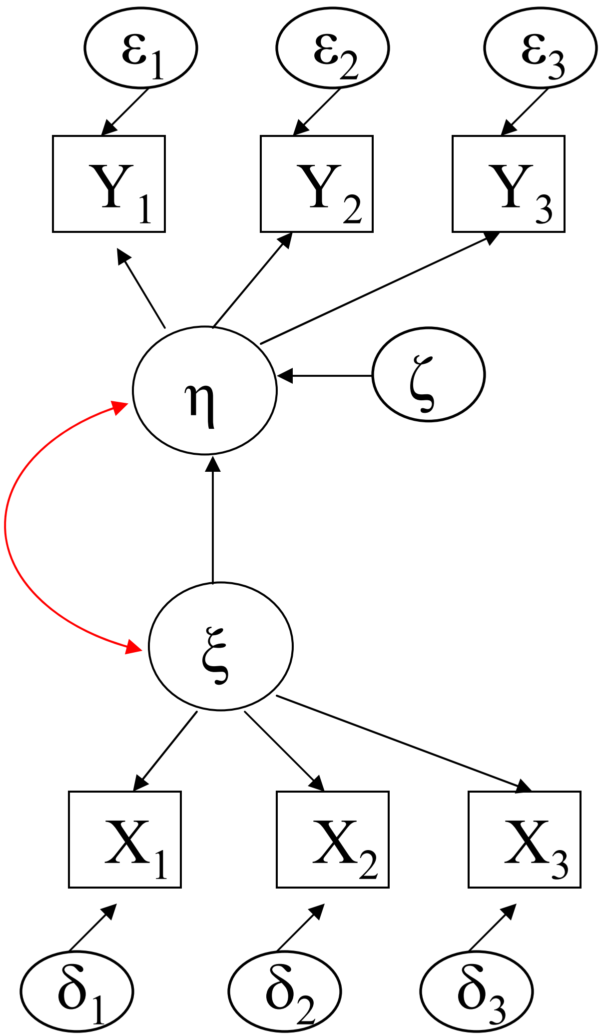
Summarize and Analyze Simultaneously



Example: $\eta = \beta T + \zeta$

$$Y_1 = \lambda_1 \eta + \varepsilon_1$$
$$Y_2 = \lambda_2 \eta + \varepsilon_2$$
$$\vdots$$
$$Y_m = \lambda_m \eta + \varepsilon_m$$

Path Notation

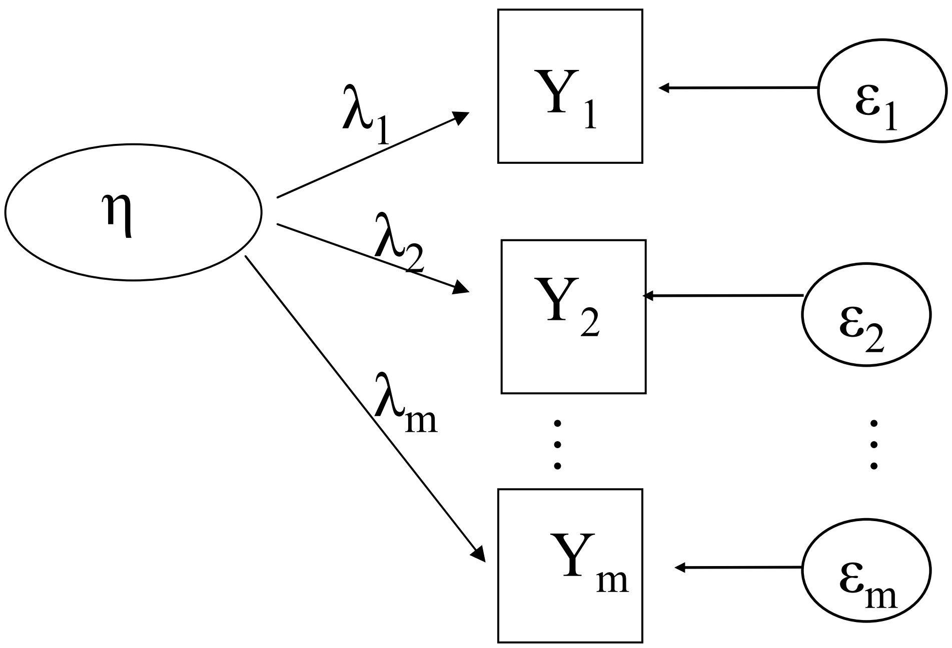


- Relationships
  - straight arrow (causal)
  - curved arrow (unspecified)
- Variables
  - circles vs. squares
  - exogenous (independent)
  - endogenous (dependent)
- Errors
  - one for every endogenous variable
  - unexplained component of predicted variables

Components of Structural Equation Model

(A) Measurement Piece

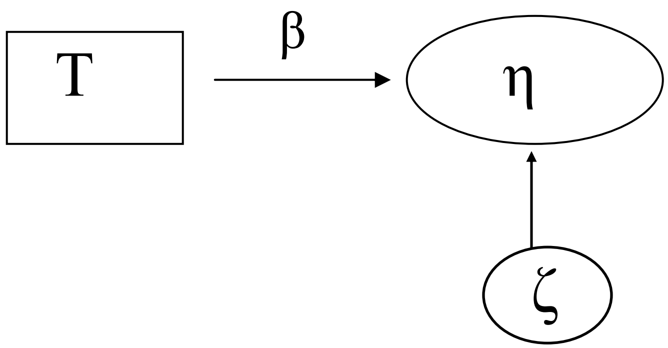
- how latent variable related to "surrogates"
- comprised of η's and Y's



Components of Structural Equation Model

(B) Structural Piece

- how latent variable is related to its predictors
- regression
- comprised of η's and T

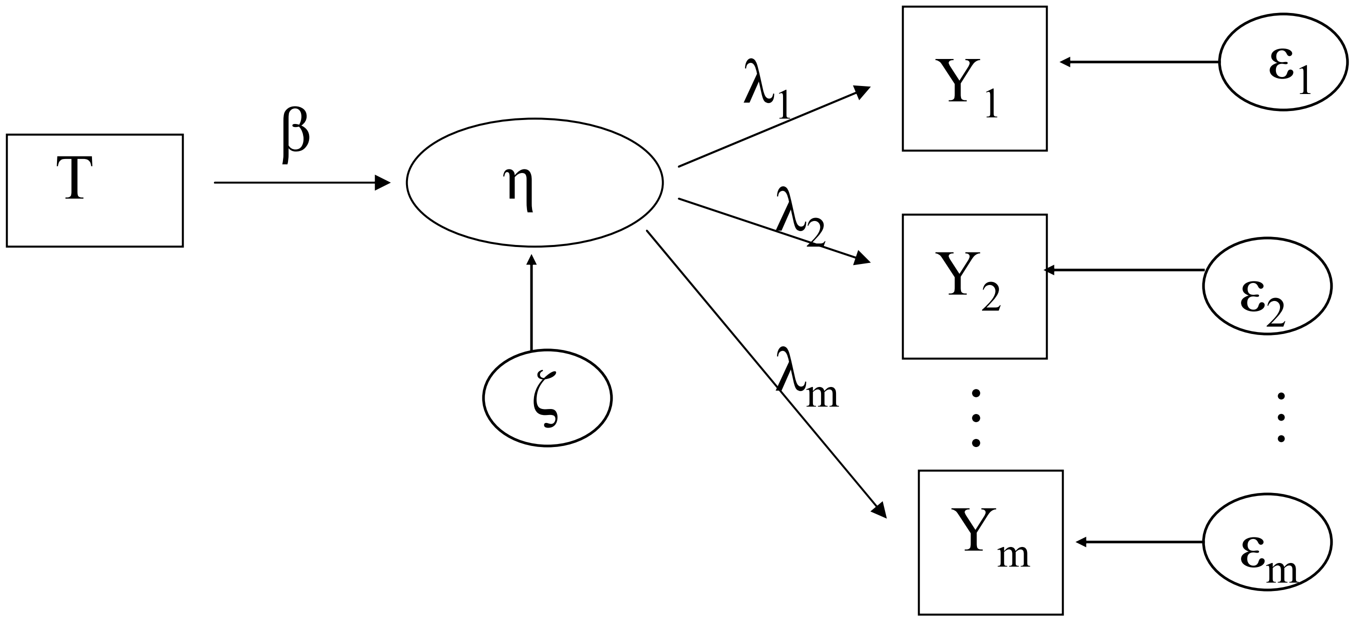


Components of Structural Equation Model

(C) Both components are fit in ONE step

Why better? Does not assume η (i.e., "summary" of Y's) known, which acknowledges measurement error.

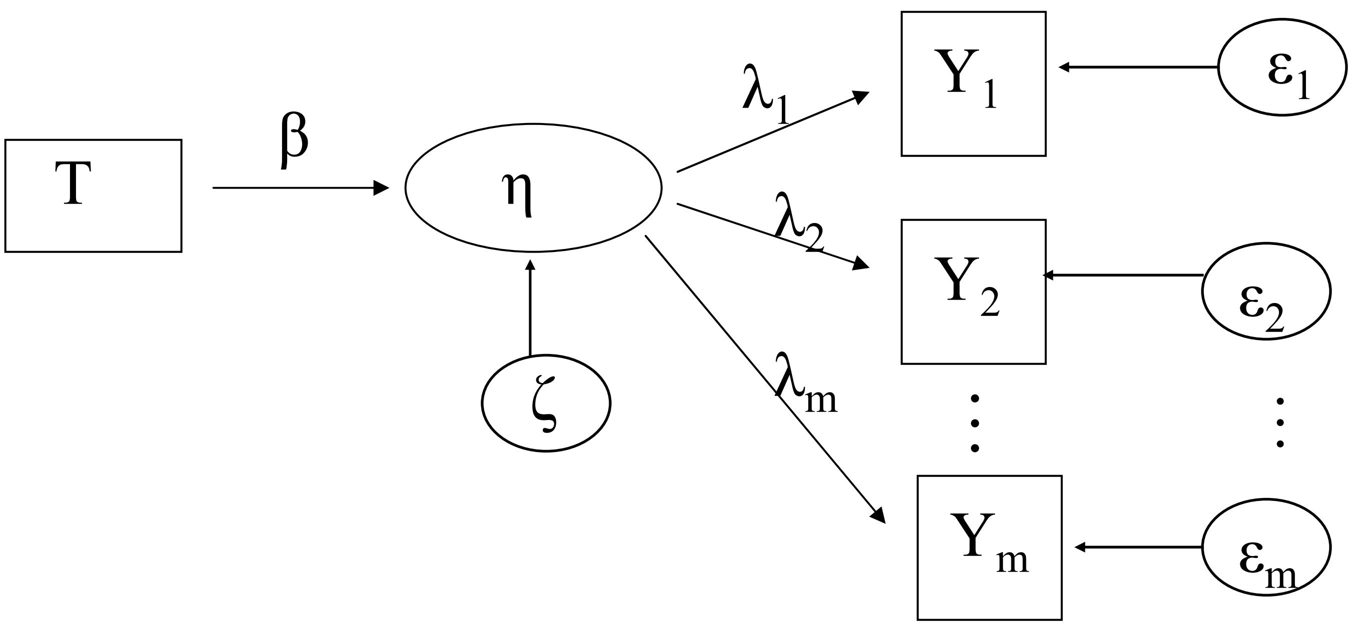
Why bad? If model is misspecified, then inference is misleading.



Statistical way of considering relationship between T and Y

$$P(Y=y \mid T)=\sum_{r=1}^{R} P(Y=y,\eta=r \mid T)$$

$$=\sum_{r=1}^{R} P(Y=y \mid \eta=r,T)P(\eta=r \mid T)$$



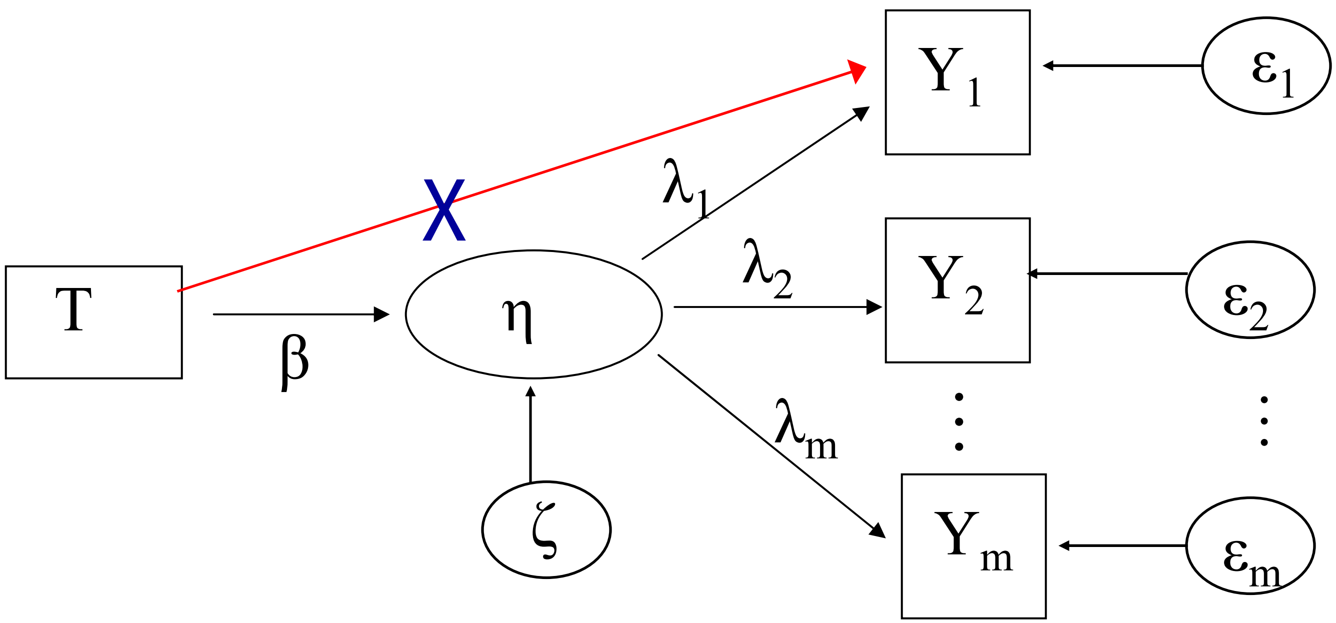
Assumption 1: Non-Differential Measurement

Equivalent interpretations:

- covariates do not predict observed responses after controlling for latent status
- no arrows between T and Y's
- Y and T independent given η

$$P(Y = y \mid \eta, T) = P(Y = y \mid \eta)$$

NOT OK UNDER NON-DIFFERENTIAL MEASUREMENT:



Assumption 2: Local/Conditional Independence

Equivalent Interpretations

- latent variable explains all association between observed variables
- no arrows among measurement errors
- observed variables are independent given η

$$P(Y_1 = y_1, Y_2 = y_2 \mid \eta) = P(Y_1 = y_1 \mid \eta) P(Y_2 = y_2 \mid \eta)$$

NOT OK UNDER CONDITIONAL INDEPENDENCE: