Consequences of measurement error

Psychology 588: Covariance structure and factor models

Scaling indeterminacy of latent variables

- Scale of a latent variable is arbitrary and "determined" by a convention for convenience
- Typically set to variance of one (factor analysis convention) or to be identical to an arbitrarily chosen indicator's scale

  By centering indicator variables, we set latent variables' means to zero
- Consider the following transformation:

$$x_j = \nu_j + \lambda_j \xi + \delta_j, \quad j = 1,...,J, \quad \xi^* = a + b\xi, \quad b \neq 0$$

$$= \left(\nu_j - \frac{a}{b}\lambda_j \right) + \left(\frac{\lambda_j}{b} \right) \xi^* + \delta_j$$

- If all J indicators are considered simultaneously, vector notation is more convenient:

$$\mathbf{x} = \mathbf{v} + \boldsymbol{\lambda}\xi + \boldsymbol{\delta}, \quad \xi^* = a + b\xi$$

$$= \left(\mathbf{v} - \frac{a}{b}\boldsymbol{\lambda} \right) + \left(\frac{1}{b}\boldsymbol{\lambda} \right) \xi^* + \boldsymbol{\delta}$$

meaning that the linear transformation of ξ can be exactly compensated in the accordingly transformed $\mathbf{v}^* = \mathbf{v} - \boldsymbol{\lambda} a/b$ and $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}/b$, leaving the errors $\boldsymbol{\delta}$ unchanged (i.e., same fit)

What's great about measurement errors in equation

- Regression weights and correlations are interpreted, implicitly assuming that the "operationally defined" variables involve no measurement error --- hardly realized for theoretical constructs (e.g., self esteem, IQ, etc.)
- Ignoring the measurement error will lead to inconsistent estimates
- We will see consequences of ignoring measurement errors

Univariate consequences

- Consider a mean-included equation for X (hours worked per week) to indicate ξ (achievement motivation):

$$X = \nu + \lambda\xi + \delta, \quad E(\xi) = \kappa, \quad E(\delta) = 0, \quad E(\xi\delta) = 0$$

$$E(X) = \mu_X = \nu + \lambda\kappa$$

$$\text{var}(X) = \lambda^2\phi + \text{var}(\delta)$$

Given only one indicator per latent variable, the intercept and loading (i.e., weight) are simply scaling constants for ξ

However, if the ξ scale is set comparable to the X scale (i.e., $\lambda = 1$), we see that $\text{var}(X)$ is an over-estimation of $\phi = \text{var}(\xi)$ if δ is not included in the equation

Bivariate relation and simple regression

- True data structure:

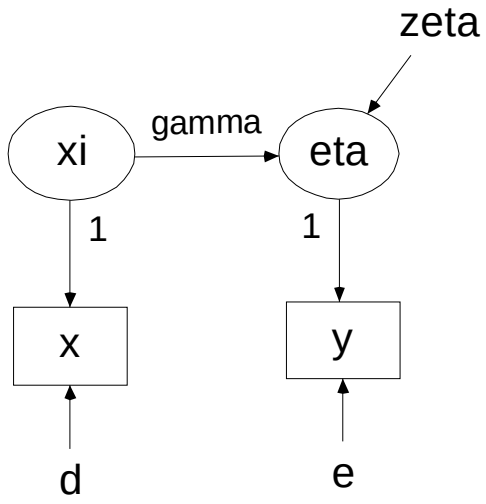
$$x = \lambda_1 \xi + \delta$$

$$y = \lambda_2 \eta + \varepsilon$$

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

η: job satisfaction

y: satisfaction scale



- $\mathrm{cov}(x, y)$ is unbiased estimate of $\mathrm{cov}(\xi, \eta)$ with $\lambda_1 = \lambda_2 = 1$, since no other variables (δ and ε) can explain $\mathrm{cov}(x, y)$

$$\mathrm{cov}\left(\xi, \eta\right) = \mathrm{cov}\left(\xi, \gamma\xi + \zeta\right) = \gamma\phi$$

$$\mathrm{cov}\left(x, y\right) = \mathrm{cov}\left(\xi + \delta, \gamma\xi + \zeta + \varepsilon\right) = \gamma\phi$$

- From the previous equations, $\gamma = \mathrm{cov}(\xi, \eta)/\phi$ and by analogy with $y = \gamma^* x + \zeta^*$ if measurement errors are ignored,

$$\gamma^* = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = \gamma \left(\frac{\phi}{\phi + \mathrm{var}(\delta)} \right) = \gamma \rho_{xx}$$

  The parenthesized ratio (reliability) becomes 1 only with no measurement error; otherwise, γ^* is an attenuated estimate of γ and $\hat{\gamma}^* = s_{xy}/s_{xx}$ is an inconsistent estimator of γ

- If $\lambda_1 \neq \lambda_2$, the bias of regression weight has an additional factor as $\gamma^* = (\lambda_2/\lambda_1)\gamma\rho_{xx}$ --- but such scaling is unusual when there is only one indicator per latent variable

- Correlations:

$$\rho_{\xi\eta}^2 = \frac{\mathrm{cov}\left(\xi,\eta\right)^2}{\phi\,\mathrm{var}\left(\eta\right)} = \frac{\gamma^2\phi}{\mathrm{var}\left(\eta\right)}$$

$$\rho_{xy}^2 = \frac{\mathrm{cov}\left(x,y\right)^2}{\mathrm{var}\left(x\right)\mathrm{var}\left(y\right)} = \frac{\gamma^2\phi^2}{\mathrm{var}\left(x\right)\mathrm{var}\left(y\right)}$$

$$= \frac{\phi}{\mathrm{var}\left(x\right)}\frac{\mathrm{var}\left(\eta\right)}{\mathrm{var}\left(y\right)}\frac{\gamma^2\phi}{\mathrm{var}\left(\eta\right)} = \rho_{xx}\rho_{yy}\rho_{\xi\eta}^2$$

which shows an attenuation of the “true” correlation due to measurement error, with the familiar correction formula:

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

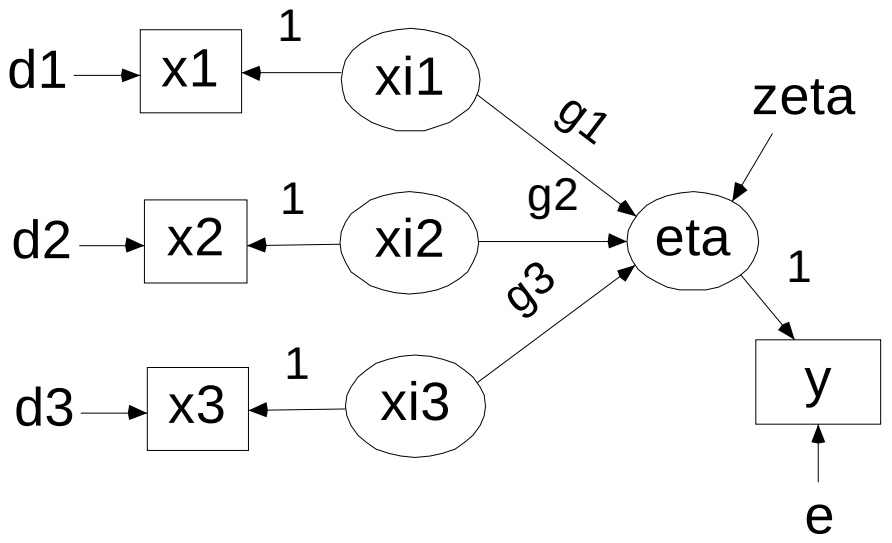
Consequences in multiple regression

- True data structure:

$$\eta = \boldsymbol{\gamma}'\boldsymbol{\xi} + \zeta$$

$$\mathbf{x} = \boldsymbol{\xi} + \boldsymbol{\delta}$$

$$y = \eta + \varepsilon$$



with $\boldsymbol{\Lambda}_x = \mathbf{I}$ and $\lambda_y = 1$

- Ignoring measurement errors: $y = \boldsymbol{\gamma}^{*\prime}\mathbf{x} + \zeta^*$
- $\boldsymbol{\sigma}_{\xi\eta} = \operatorname{cov}(\boldsymbol{\xi}, \eta) = \operatorname{cov}(\boldsymbol{\xi}, \boldsymbol{\xi}'\boldsymbol{\gamma} + \zeta) = \boldsymbol{\Phi}\boldsymbol{\gamma}$

$$\boldsymbol{\sigma}_{xy} = \operatorname{cov}(\mathbf{x}, y) = \operatorname{cov}(\boldsymbol{\xi} + \boldsymbol{\delta}, \boldsymbol{\xi}'\boldsymbol{\gamma} + \zeta + \varepsilon) = \boldsymbol{\Phi}\boldsymbol{\gamma}$$

- $\boldsymbol{\gamma} = \boldsymbol{\Phi}^{-1}\boldsymbol{\sigma}_{\xi\eta}$ and by analogy with $y = \boldsymbol{\gamma}^{*\prime}\mathbf{x} + \zeta^{*}$,

  $$\boldsymbol{\gamma}^{*} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\sigma}_{xy} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Phi}\boldsymbol{\gamma} = \left(\boldsymbol{\Phi} + \boldsymbol{\Theta}_{\delta}\right)^{-1}\boldsymbol{\Phi}\boldsymbol{\gamma}$$

  Without measurement error ($\boldsymbol{\Theta}_{\delta} = \mathbf{0}$), $\boldsymbol{\gamma}^{*} = \boldsymbol{\gamma}$; otherwise, $\boldsymbol{\gamma}^{*} \neq \boldsymbol{\gamma}$

- Alternatively written: $\boldsymbol{\gamma}^{*} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{x\xi}\boldsymbol{\gamma}$ since $\boldsymbol{\Sigma}_{x\xi} = \boldsymbol{\Phi}$ --- where $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{x\xi}$ is the OLS estimator of $\mathbf{B}$ in $\boldsymbol{\xi} = \mathbf{B}\mathbf{x} + \mathbf{e}$, i.e., regression weights for prediction of $\boldsymbol{\xi}$ by $\mathbf{x}$

  Again, without measurement error, $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{x\xi} = \mathbf{I}$

- Note: in Bollen (pp. 159-168), $\boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{\xi\eta}, \boldsymbol{\Sigma}_{xy}$ are meant to be $\boldsymbol{\gamma}, \boldsymbol{\sigma}_{\xi\eta}, \boldsymbol{\sigma}_{xy}$, respectively, for the multiple regression model

- As a very simplified case, suppose x_1 is the only fallible as:

$$x_1 = \xi_1 + \delta_1$$

$$x_i = \xi_i, \quad i = 2,...,q$$

with the true and estimated regression equations:

$$\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \cdots + \gamma_q\xi_q + \zeta$$

$$\eta = \gamma_1^*x_1 + \gamma_2^*\xi_2 + \cdots + \gamma_q^*\xi_q + \zeta^*$$

- In this special case, the regression weight matrix has a simple multiplicative form of bias (hint: use $\tilde{\mathbf{\Phi}} = \mathbf{\Phi} + \mathbf{\Theta}_\delta$):

$$\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{x\xi} = \left(\mathbf{\Phi} + \mathbf{\Theta}_\delta\right)^{-1}\mathbf{\Phi} = \mathbf{I} + \left[\mathbf{c}, \mathbf{0}_{q\times(q-1)}\right] = \left[\begin{array}{c|c} 1+c_1 & \mathbf{0}' \\ \hline \mathbf{c}_2 & \mathbf{I} \end{array}\right]$$

- Consequently, resulting bias factors are:

$$\gamma_1^* = b_{\xi_1 x_1 \cdot \xi_2 \cdots \xi_q} \gamma_1$$

$$\gamma_i^* = \gamma_i + b_{\xi_1 \xi_i \cdot x_1 \{\xi_{\sim i}\}} \gamma_1 , \quad i = 2, \ldots, q$$

  - Bias-factor for x_1 is less than 1 in absolute value (1 without measurement error), and so γ_1^* is biased toward 0 --- the bias factor $b_{\xi_1 x_1 \cdot \xi_2 \cdots \xi_q}$ indicates regression weight b_1 in $\xi_1 = b_0 + b_1 x_1 + b_2 \xi_2 + \ldots + b_q \xi_q$

  - Consequences for $x_i, \; i = 2, \ldots, q$ are additive, depending on relationships between ξ_1 and ξ_i holding all other IVs constant, and γ_1

- So far all reasoning is based on rather unrealistic assumptions:
  - Only single indicator per latent variable, and so its loading becomes simply scaling constant
  - Only one fallible IV
- Without such assumptions (e.g., all IVs fallible), consequences of measurement error become too complicated and hard to simplify algebraically --- no particular simple form of $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{x\xi}$
- One clear conclusion: all estimates are inconsistent --- systematically different from what they meant to be

- Consequence in standardization:

$$\text{standardized } \gamma_i^* = \gamma_i^* \sqrt{\frac{\phi_{ii} + \text{var}(\delta_i)}{\text{var}(\eta) + \text{var}(\varepsilon)}}$$

- Consequence in SMC is similar to the bivariate case:

$$\text{plim}(R^2) \geq \text{plim}(R^{*2})$$

- What should we do with essentially omnipresent measurement error?
  - Use SEM which allows for measurement errors in the model --- though we are limited in certain models regarding the model identification (e.g., Table 5.1, p. 164)

Correlated errors of measurement

- Consequence in regression weights further complicated:

$$\boldsymbol{\gamma}^{*} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{x\xi}\boldsymbol{\gamma} + \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\sigma}_{\delta\varepsilon}$$

For simple regression: $\gamma^{*} = \gamma\rho_{xx} + \dfrac{\mathrm{cov}(\varepsilon, \delta)}{\mathrm{var}(x)}$

Now, γ^{*} is not necessarily $<$ γ

- If correlated measurement errors are only within IVs (i.e., $\boldsymbol{\sigma}_{\delta\varepsilon} = \mathbf{0}$, $\boldsymbol{\Sigma}_{xx} = \boldsymbol{\Phi} + \boldsymbol{\Theta}_{\sigma}$ where $\boldsymbol{\Theta}_{\sigma}$ is not diagonal), $\boldsymbol{\gamma}^{*} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{x\xi}\boldsymbol{\gamma}$ still holds (but the bias factor will have a more complicated form, also involving off-diagonal entries of $\boldsymbol{\Theta}_{\sigma}$)

With multi-equations

- In path models with sequential causal paths, consequences of measurement errors very hard to simply generalize --- see the union sentiment (Fig. 5.2, p. 169) and SES (Fig. 5.4, p. 173) examples

- If reliabilities are known, the corresponding error variances can be constrained; if unknown, the error variances may be modeled as free parameters provided that they are identifiable

- To keep in mind: we need more than one indicator per latent variable for identifiability and statistical testing --- leading to measurement models with multiple indicators or CFA