Ignoring Measurement Error: Convergence[1]

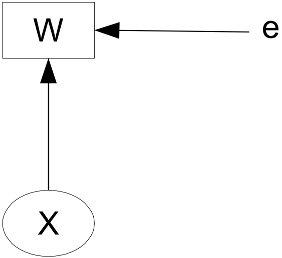
STA2053 Fall 2022

[1] See last slide for copyright information.

Overview

Additive measurement error

A very simple model



$$W = X + e$$

Where $E(X) = \mu_x$, $E(e) = 0$, $Var(X) = \sigma_x^2$, $Var(e) = \sigma_e^2$, and $Cov(X, e) = 0$.

Variance and Covariance

$W = X + e$

$$\begin{aligned} Var(W) &= Var(X) + Var(e) \\ &= \sigma_x^2 + \sigma_e^2 \end{aligned}$$

$$\begin{aligned} Cov(X, W) &= Cov\,(X, X + e) \\ &= Cov(X, X) + Cov(X, e) \\ &= \sigma_x^2 + 0 \\ &= \sigma_x^2 \end{aligned}$$

Definition of Reliability

Psychometric

Reliability is the squared correlation between the observed variable and the latent variable (true score).

Calculation of Reliability

Squared correlation between observed and true score

$$
\begin{aligned}
\rho^2 &= \left(\frac{Cov(X,W)}{SD(X)SD(W)}\right)^2 \\
&= \left(\frac{\sigma_x^2}{\sqrt{\sigma_x^2}\sqrt{\sigma_x^2+\sigma_e^2}}\right)^2 \\
&= \frac{\sigma_x^4}{\sigma_x^2(\sigma_x^2+\sigma_e^2)} \\
&= \frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2} \\
&= \frac{Var(X)}{Var(W)}.
\end{aligned}
$$

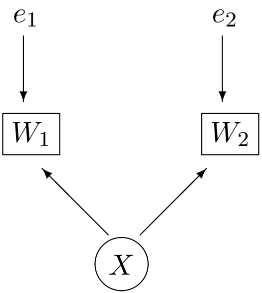
Reliability is the proportion of the variance in the observed variable that comes from the latent variable of interest, and not from random error.

How to estimate reliability from data

- Correlate usual measurement with “Gold Standard?”
- Not very realistic, except maybe for some bio-markers.
- One answer: Measure twice.

Measure twice

Called "equivalent measurements" because error variance is the same



$$
\begin{aligned}
W_1 &= X + e_1 \\
W_2 &= X + e_2,
\end{aligned}
$$

where $E(X) = \mu_x$, $Var(X) = \sigma^2_x$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, and X, e_1 and e_2 are all independent.

Reliability equals the correlation between two equivalent measurements

This is a population correlation

$$
\begin{aligned}
Corr(W_1, W_2) &= \frac{Cov(W_1, W_2)}{SD(W_1)SD(W_2)} \\
&= \frac{Cov(X + e_1, X + e_2)}{\sqrt{\sigma_x^2 + \sigma_e^2}\sqrt{\sigma_x^2 + \sigma_e^2}} \\
&= \frac{Cov(X, X) + Cov(X, e_2) + Cov(e_1, X) + Cov(e_1, e_2)}{\sigma_x^2 + \sigma_e^2} \\
&= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2},
\end{aligned}
$$

which is the reliability.

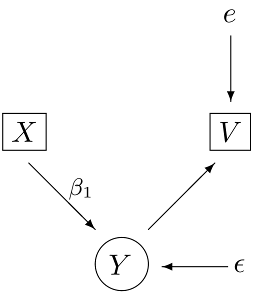
Estimate the reliability: Measure twice for a sample of size n

With a well-chosen time gap

Calculate $r = \frac{\sum_{i=1}^{n}(W_{i1}-\overline{W}_1)(W_{i2}-\overline{W}_2)}{\sqrt{\sum_{i=1}^{n}(W_{i1}-\overline{W}_1)^2}\sqrt{\sum_{i=1}^{n}(W_{i2}-\overline{W}_2)^2}}$.

- Test-retest reliability
- Alternate forms reliability
- Split-half reliability

Measurement error in the response variable only



True model:

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

Naive model: $V_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Is $\widehat{\beta}_1$ consistent?

Ignoring measurement error in Y

First calculate $Cov(X_i, V_i)$. Under the true model.

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

$$\begin{aligned} Cov(X_i, V_i) &= Cov(X_i, \beta_1 X_i + \epsilon_i) \\ &= \beta_1 \sigma_x^2 \end{aligned}$$

Target of $\widehat{\beta}_1$ as $n \to \infty$

Have $Cov(X_i, V_i) = \beta_1 \sigma_x^2$ and $Var(X_i) = \sigma_x^2$

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(V_i - \overline{V})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\widehat{\sigma}_{x,v}}{\widehat{\sigma}_x^2} \\
&\overset{a.s.}{\to} \frac{Cov(X_i, V_i)}{Var(X_i)} \\
&= \frac{\beta_1 \sigma_x^2}{\sigma_x^2} \\
&= \beta_1
\end{aligned}
$$

Consistent.

Why did it work?

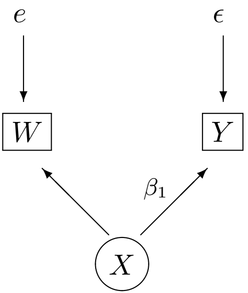
$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
V_i &= \nu + Y_i + e \\
&= \nu + (\beta_0 + \beta_1 X_i + \epsilon_i) + e_i \\
&= (\nu + \beta_0) + \beta_1 X_i + (\epsilon_i + e_i) \\
&= \beta_0' + \beta_1 X_i + \epsilon_i'
\end{aligned}
$$

- This is a re-parameterization.
- Most definitely *not* one-to-one.
- (ν, β_0) is absorbed into β_0'.
- (ϵ_i, e_i) is absorbed into ϵ_i'.
- Can't know everything, but all we care about is β_1 anyway.

Don't Worry

- If a response variable appears to have no measurement error, assume it does have measurement error but the problem has been re-parameterized.
- Measurement error in Y is part of ϵ.

Measurement error in a single explanatory variable



True model:

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

Naive model: $Y_i = \beta_0 + \beta_1 W_i + \epsilon_i$

Target of $\widehat{\beta}_1$ as $n \to \infty$

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ and $W_i = X_i + e_i$

Have $Cov(W_i, Y_i) = \beta_1 \sigma_x^2$ and $Var(W_i) = \sigma_x^2 + \sigma_e^2$

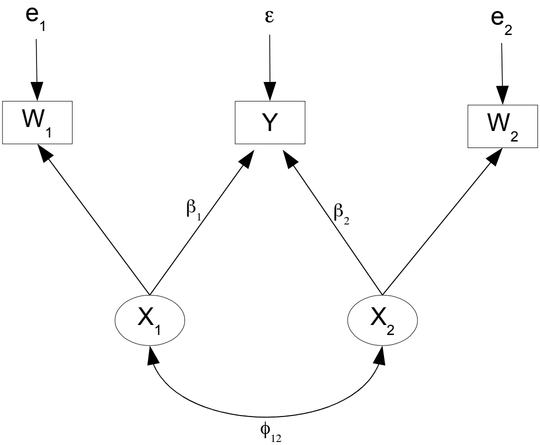
$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (W_i - \overline{W})(Y_i - \overline{Y})}{\sum_{i=1}^n (W_i - \overline{W})^2} \\
&= \frac{\widehat{\sigma}_{w,y}}{\widehat{\sigma}_w^2} \\
&\overset{a.s.}{\to} \frac{Cov(W,Y)}{Var(W)} = \frac{\beta_1 \sigma_x^2}{\sigma_x^2 + \sigma_e^2} \\
&= \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2} \right)
\end{aligned}
$$

$\widehat{\beta}_1 \stackrel{a.s.}{\rightarrow} \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2} \right)$

$W_i = X_i + e_i$

- $\widehat{\beta}_1$ converges to β times the reliability of W_i.
- It's inconsistent.
- Because the reliability is less than one, it's asymptotically biased toward zero.
- The worse the measurement of X_i, the more the asymptotic bias.
- Sometimes called "attenuation" (weakening).
- If a good estimate of reliability is available from another source, one can "correct for attenuation."
- When $H_0 : \beta_1 = 0$ is true, no problem.
- False sense of security?

Measurement error in two explanatory variables



Want to assess the relationship of X_2 to Y controlling for X_1 by testing $H_0 : \beta_2 = 0$.

Statement of the model

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

$$
\begin{array}{rcl}
Y_i & = & \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\
W_{i,1} & = & X_{i,1} + e_{i,1} \\
W_{i,2} & = & X_{i,2} + e_{i,2},
\end{array}
$$

where

$E(X_{i,1}) = \mu_1$, $E(X_{i,2}) = \mu_2$, $E(\epsilon_i) = E(e_{i,1}) = E(e_{i,2}) = 0$,

$Var(\epsilon_i) = \psi$, $Var(e_{i,1}) = \omega_1$, $Var(e_{i,2}) = \omega_2$,

The errors $\epsilon_i, e_{i,1}$ and $e_{i,2}$ are all independent,

$X_{i,1}$ and $X_{i,2}$ are independent of $\epsilon_i, e_{i,1}$ and $e_{i,2}$, and

$$cov \left(\begin{array}{c} X_{i,1} \\ X_{i,1} \end{array} \right) = \left(\begin{array}{cc} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{array} \right).$$

Note

- Reliability of W_1 is $\frac{\phi_{11}}{\phi_{11}+\omega_1}$.
- Reliability of W_2 is $\frac{\phi_{22}}{\phi_{22}+\omega_2}$.

True Model versus Naive Model

True model:

$$\begin{array}{rcl} Y_i & = & \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ W_{i,1} & = & X_{i,1} + e_{i,1} \\ W_{i,2} & = & X_{i,2} + e_{i,2}, \end{array}$$

Naive model: $Y_i = \beta_0 + \beta_1 W_{i,1} + \beta_2 W_{i,2} + \epsilon_i$

- Fit the naive model.
- See what happens to $\widehat{\beta}_2$ as $n \rightarrow \infty$ when the true model holds.
- Start by calculating $cov(\mathbf{d}_i)$.

Covariance matrix of the observable data

$$
\begin{aligned}
\boldsymbol{\Sigma} &= cov\left(\begin{array}{c} W_{i,1} \\ W_{i,2} \\ Y_i \end{array}\right) \\
&= \left(\begin{array}{rrr} \omega_1 + \phi_{11} & \phi_{12} & \beta_1\phi_{11} + \beta_2\phi_{12} \\ \phi_{12} & \omega_2 + \phi_{22} & \beta_1\phi_{12} + \beta_2\phi_{22} \\ \beta_1\phi_{11} + \beta_2\phi_{12} & \beta_1\phi_{12} + \beta_2\phi_{22} & \beta_1^2\phi_{11} + 2\,\beta_1\beta_2\phi_{12} + \beta_2^2\phi_{22} + \psi \end{array}\right)
\end{aligned}
$$

What happens to $\widehat{\beta}_2$ as $n \rightarrow \infty$?

Interested in $H_0 : \beta_2 = 0$

$$
\begin{aligned}
\widehat{\beta}_2 &= \frac{\widehat{\sigma}_{11}\widehat{\sigma}_{23} - \widehat{\sigma}_{12}\widehat{\sigma}_{13}}{\widehat{\sigma}_{11}\widehat{\sigma}_{22} - \widehat{\sigma}_{12}^2} \\
&\stackrel{a.s.}{\rightarrow} \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \\
&= \frac{\beta_1\omega_1\phi_{12} + \beta_2(\omega_1\phi_{22} + \phi_{11}\phi_{22} - \phi_{12}^2)}{(\phi_{1,1} + \omega_1)(\phi_{2,2} + \omega_2) - \phi_{12}^2} \\
&\neq \beta_2
\end{aligned}
$$

Inconsistent.

When $H_0 : \beta_2 = 0$ is true

$$\widehat{\beta}_2 \stackrel{a.s.}{\rightarrow} \frac{\beta_1 \omega_1 \phi_{12}}{(\phi_{1,1} + \omega_1)(\phi_{2,2} + \omega_2) - \phi_{12}^2}$$

So $\widehat{\beta}_2$ goes to the wrong target unless

- There is no relationship between X_1 and Y, or
- There is no measurement error in W_1, or
- There is no correlation between X_1 and X_2.

Also, t statistic goes to plus or minus ∞ and the p-value $\stackrel{a.s.}{\rightarrow} 0$. Remember, H_0 is true.

Copyright Information