Measurement Error in the Response Variable[1]

STA 2101 Fall 2019

[1] See last slide for copyright information.

Ignoring measurement error

- We have seen that ignoring measurement error in the explanatory variables can lead to disaster.
- What about measurement error in the response variable?

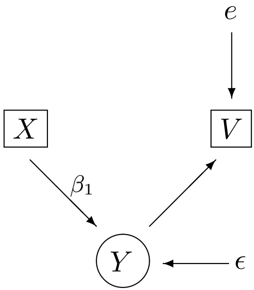
Example of Measurement Error in Y only

X could be drug dose, Y could be true anxiety, V could be reported anxiety

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

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

where $Var(X_i) = \sigma^2_x$, $Var(e_i) = \sigma^2_e$, $Var(\epsilon_i) = \sigma^2_\epsilon$, and X_i, e_i, ϵ_i are all independent.



Parameters of the true model are not identifiable from the means and covariance matrix

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

where $Var(X_i) = \sigma^2_x$, $Var(e_i) = \sigma^2_e$, and $Var(\epsilon_i) = \sigma^2_\epsilon$.

- Only the (X_i, V_i) pairs are observable.
- There are 5 moments.
- $\boldsymbol{\theta} = (\beta_0, \beta_1, \mu_x, \sigma^2_x, \sigma^2_\epsilon, \nu, \sigma^2_e)$: 7 parameters
- Fails the test of the parameter count rule.

Ignoring measurement error as usual

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:

$$V_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Fit the Naive Model, using V_i as the response variable

$V_i = \beta_0 + \beta_1 X_i + \epsilon_i$

First note that under the *true* model, $Cov(X_i, V_i) = \beta_1 \sigma_x^2$ and $Var(X_i) = \sigma_x^2$.

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

So $\widehat{\beta}_1$ is consistent, even though the model is mis-specified.

Why does the naive model work so well?

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

- This is a *re-parameterization*.
- Not a one-to-one re-parameterization – call it a "collapsing" re-parameterization.
- The pair (ν, β_0) is absorbed into β_0'.
- $Var(\epsilon_i + e_i) = \sigma_\epsilon^2 + \sigma_e^2$ is absorbed into a single unknown variance that will probably be called σ^2.
- ν and β_0 will never be knowable separately, and also σ_ϵ^2 and σ_e^2 will never be knowable separately.
- It's okay. All we care about is β_1 anyway.

This is very common

- In many models, it will appear that the response variable is being measured without error.
- Of course there really is measurement error in Y_i, but it has been absorbed into the error term.
- So any model without measurement error in the response variable should be viewed as a re-parameterized version of a more realistic model.
- The measurement error should be independent of X, or there is real trouble.

Copyright Information