Introduction to Regression with Measurement Error[1]

STA431 Spring 2017

[1]See last slide for copyright information.

Overview

Measurement Error

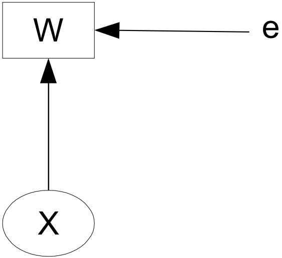
- Snack food consumption
- Exercise
- Income
- Cause of death (classification error)
- Even amount of drug that reaches animals blood stream in an experimental study.
- Is there anything that is *not* measured with error?

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

$W = X + e$

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

$$
\begin{aligned}
Cov(X, W) &= E(\overset{c}{X}\overset{c}{W}) \\
&= E(\overset{c}{X}\,(\overset{c}{X} + e)) \\
&= E(\overset{c}{X}^2) + E(\overset{c}{X})E(e) \\
&= \sigma_X^2
\end{aligned}
$$

Explained Variance

- Variance is an index of unit-to-unit variation in a measurement.
- Explaining unit-to-unit variation is an important goal of Science.
- How much of the variation in an observed variable comes from variation in the latent quantity of interest, and how much comes from random noise?

Definition of Reliability

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

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

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

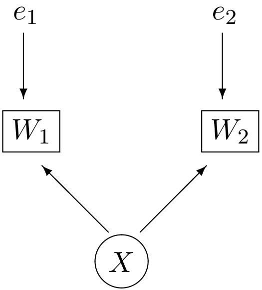
- Correlate usual measurement with “Gold Standard?”
- Not very realistic, except maybe for some bio-markers.
- One answer: Measure twice.

Measure twice

Called "equivalent measurements" because error variance is the same

$$W_1 = X + e_1$$
$$W_2 = X + e_2,$$

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

$$
\begin{aligned}
Corr(W_1, W_2) &= \frac{Cov(W_1, W_2)}{SD(W_1)SD(W_2)} \\
&= \frac{E(\overset{c}{W}_1 \overset{c}{W}_2)}{\sqrt{\sigma_x^2 + \sigma_e^2}\sqrt{\sigma_x^2 + \sigma_e^2}} \\
&= \frac{E(\overset{c}{X} + e_1)(\overset{c}{X} + e_2)}{\sigma_x^2 + \sigma_e^2} \\
&= \frac{E(\overset{c}{X}^2) + 0 + 0 + 0}{\sigma_x^2 + \sigma_e^2} \\
&= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2},
\end{aligned}
$$

which is the reliability.

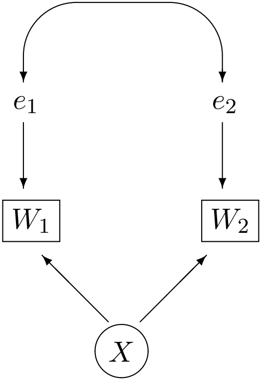
Estimate the reliability: Measure twice for a sample of size n

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

Omitted variables can cause correlated measurement error

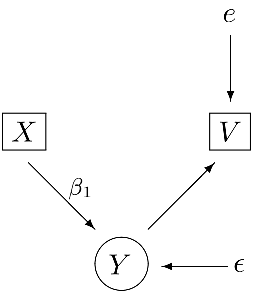


Leading to an over-estimate of reliability.

Measurement error in regression analysis

- Mostly we are interested in relationships between latent (true) variables.
- But all we have at best are the true variables measured with error.
- Models like $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i$ are mis-specified.
- The most common way of dealing with measurement error in regression is to ignore it.
- What effect does this have on estimation and inference?
- First consider ignoring measurement error just in the response variable.

Measurement error in the response variable



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

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

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

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

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

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

Why did it work?

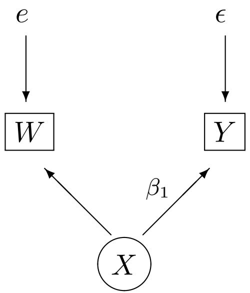
$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\
V_i &= \nu + Y_i + e \\
&= \nu + (\beta_0 + \beta_1 X_i + \epsilon_i) + e_i \\
&= (\nu + \beta_0) + \beta_1 X_i + (\epsilon_i + e_i) \\
&= \beta_0' + \beta_1 X_i + \epsilon_i'
\end{aligned}
$$

- 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 $n \rightarrow \infty$

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

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

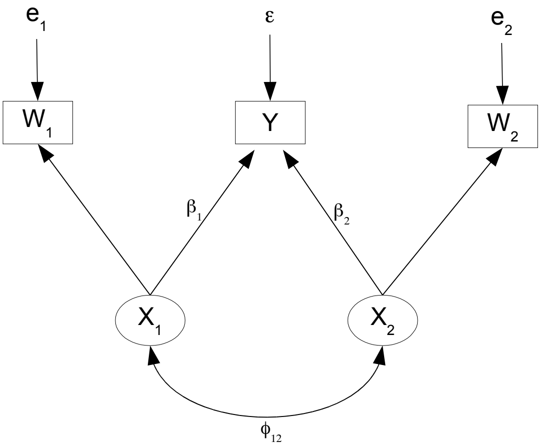
$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (W_i - \overline{W})(Y_i - \overline{Y})}{\sum_{i=1}^n (W_i - \overline{W})^2} \\
&= \frac{\widehat{\sigma}_{w,y}}{\widehat{\sigma}_w^2} \\
&\stackrel{p}{\rightarrow} \frac{Cov(W,Y)}{Var(W)} \\
&= \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}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

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

$W_i = X_i + e_i$

- $\widehat{\beta}_1$ 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 $i = 1, \ldots, n$

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

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:

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

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 $n \to \infty$ 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 $H_0 : \beta_2 = 0$

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

Inconsistent.

When $H_0 : \beta_2 = 0$ is true

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

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 p-value $\stackrel{p}{\rightarrow} 0$.
Remember, H_0 is true.

How bad is it for finite sample sizes?

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

A big simulation study (Brunner and Austin, 2009) with six factors

- Sample size: $n = 50, 100, 250, 500, 1000$
- $Corr(X_1, X_2)$: $\phi_{12} = 0.00, 0.25, 0.75, 0.80, 0.90$
- Proportion of variance in Y explained by X_1: 0.25, 0.50, 0.75
- Reliability of W_1: 0.50, 0.75, 0.80, 0.90, 0.95
- Reliability of W_2: 0.50, 0.75, 0.80, 0.90, 0.95
- Distribution of latent variables and error terms: Normal, Uniform, t, Pareto.

There were $5 \times 5 \times 3 \times 5 \times 5 \times 4 = 7{,}500$ treatment combinations.

Simulation study procedure

Within each of the $5 \times 5 \times 3 \times 5 \times 5 \times 4 = 7{,}500$ treatment combinations,

- 10,000 random data sets were generated
- For a total of 75 million data sets
- All generated according to the true model, with $\beta_2 = 0$.
- Fit naive model, test $H_0 : \beta_2 = 0$ at $\alpha = 0.05$.
- Proportion of times H_0 is rejected is a Monte Carlo estimate of the Type I Error Probability.
- It should be around 0.05.

Look at a small part of the results

- Both reliabilities = 0.90
- Everything is normally distributed
- $\beta_0 = 1$, $\beta_1 = 1$ and of course $\beta_2 = 0$.

Table 1 of Brunner and Austin (2009, p.39)

Canadian Journal of Statistics, Vol. 37, Pages 33-46, Used without permission

TABLE 1: Estimated Type I error rates when independent variables and measurement errors are all normal, and reliability of W_1 and W_2 both equal 0.90.

	Correlation between X_1 and X_2				
N	0.0	0.2	0.4	0.6	0.8
25% of variance in Y is explained by X_1					
50	0.0476†	0.0505†	0.0636	0.0715	0.0913
100	0.0504†	0.0521†	0.0834	0.0940	0.1294
250	0.0467†	0.0533†	0.1402	0.1624	0.2544
500	0.0468†	0.0595†	0.2300	0.2892	0.4649
1,000	0.0505†	0.0734	0.4094	0.5057	0.7431
50% of variance in Y is explained by X_1					
50	0.0460†	0.0520†	0.0963	0.1106	0.1633
100	0.0535†	0.0569†	0.1461	0.1857	0.2837
250	0.0483†	0.0625	0.3068	0.3731	0.5864
500	0.0515†	0.0780	0.5323	0.6488	0.8837
1,000	0.0481†	0.1185	0.8273	0.9088	0.9907
75% of variance in Y is explained by X_1					
50	0.0485†	0.0579†	0.1727	0.2089	0.3442
100	0.0541†	0.0679	0.3101	0.3785	0.6031
250	0.0479†	0.0856	0.6450	0.7523	0.9434
500	0.0445†	0.1323	0.9109	0.9635	0.9992
1,000	0.0522†	0.2179	0.9959	0.9998	1.00000

†Not significantly different from 0.05, Bonferroni corrected for 7,500 tests.

Marginal Mean Type I Error Probabilities

Base Distribution

normal	Pareto	t Distr	uniform
0.38692448	0.36903077	0.38312245	0.38752571

Explained Variance

0.25	0.50	0.75
0.27330660	0.38473364	0.48691232

Correlation between Latent Independent Variables

0.00	0.25	0.75	0.80	0.90
0.05004853	0.16604247	0.51544093	0.55050700	0.62621533

Sample Size n

50	100	250	500	1000
0.19081740	0.27437227	0.39457933	0.48335707	0.56512820

Reliability of W_1

0.50	0.75	0.80	0.90	0.95
0.60637233	0.46983147	0.42065313	0.26685820	0.14453913

Reliability of W_2

0.50	0.75	0.80	0.90	0.95
0.30807933	0.37506733	0.38752793	0.41254800	0.42503167

Summary

- Ignoring measurement error in the independent variables can seriously inflate Type I error probabilities.
- The poison combination is measurement error in the variable for which you are "controlling," and correlation between latent explanatory variables.
- If either is zero, there is no problem.
- Factors affecting severity of the problem are (next slide)

Factors affecting severity of the problem

Problem of inflated Type I error probability

- As the correlation between X_1 and X_2 increases, the problem gets worse.
- As the correlation between X_1 and Y increases, the problem gets worse.
- As the amount of measurement error in X_1 increases, the problem gets worse.
- As the amount of measurement error in X_2 increases, the problem gets less severe.
- As the sample size increases, the problem gets worse.
- Distribution of the variables does not matter much.

As the sample size increases, the problem gets worse

For a large enough sample size, no amount of measurement error in the explanatory variables is safe, assuming that the latent explanatory variables are correlated.

Other kinds of regression, other kinds of measurement error

- Logistic regression
- Proportional hazards regression in survival analysis
- Log-linear models: Test of conditional independence in the presence of classification error
- Median splits
- Even converting X_1 to ranks inflates Type I Error probability.

Moral of the story

Use models that allow for measurement error in the explanatory variables.

Copyright Information