Introduction to Regression with Measurement Error

STA302: Fall/Winter 2013

Measurement Error

- Snack food consumption
- Exercise
- Income
- Cause of death
- Even amount of drug that reaches animal's blood stream in an experimental study
- Is there anything that is *not* measured with error?

For categorical variables

Classification error is common

Simple additive model for measurement error: Continuous case

$$W = X + e$$

Where $E(X) = \mu$, $E(e) = 0$, $Var(X) = \sigma_X^2$, $Var(e) = \sigma_e^2$, and $Cov(X, e) = 0$. Because X and e are uncorrelated,

$$Var(W) = Var(X) + Var(e) = \sigma_X^2 + \sigma_e^2$$

How much of the variation in the observed variable comes from variation in the quantity of interest, and how much comes from random noise?

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

First, recall

$$\begin{aligned} Corr(X,Y) &= \frac{Cov(X,Y)}{SD(X)SD(Y)} \\ Var(X+a) &= Var(X) \\ Cov(X+a,Y+b) &= Cov(X,Y) \end{aligned}$$

Reliability

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

$$(Corr(X, W))^2 = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_e^2}$$

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

The consequences of ignoring measurement error in the explanatory (x) variables

Measurement error in the response variable is a less serious problem: Re-parameterize

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \epsilon_1 \\
V &= \nu + Y + \epsilon_2 \\
&= \nu + (\beta_0 + \beta_1 X + \epsilon_1) + \epsilon_2 \\
&= (\nu + \beta_0) + \beta_1 X + (\epsilon_1 + \epsilon_2) \\
&= \beta_0' + \beta_1 X + \epsilon'
\end{aligned}
$$

Can’t know everything, but all we care about is β_1 anyway.

Measurement error in the explanatory variables

- True model

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

- Naïve model

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

True Model (More detail)

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

where independently for $i = 1, \ldots, n$, $E(X_{i,1}) = \mu_1$, $E(X_{i,2}) = \mu_2$, $E(\epsilon_i) = E(e_{i,1}) = E(e_{i,2}) = 0$, $Var(\epsilon_i) = \sigma^2$, $Var(e_{i,1}) = \omega_1$, $Var(e_{i,2}) = \omega_2$, the errors $\epsilon_i, e_{i,1}$ and $e_{i,2}$ are all independent, $X_{i,1}$ is independent of $\epsilon_i, e_{i,1}$ and $e_{i,2}$, $X_{i,2}$ is independent of $\epsilon_i, e_{i,1}$ and $e_{i,2}$, and

$$
Var \left[\begin{array}{c} X_{i,1} \\ X_{i,1} \end{array} \right] = \left[\begin{array}{cc} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{array} \right]
$$

Reliabilities

- Reliability of W_1 is $$\frac{\phi_{11}}{\phi_{11} + \omega_1}$$

- Reliability of W_2 is $$\frac{\phi_{22}}{\phi_{22} + \omega_2}$$

Test X_2 controlling for (holding constant) X_1

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\frac{\partial}{\partial x_2} E(Y) = \beta_2$$

That's the usual conditional model

Unconditional: Test X_2 controlling for X_1

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon
\end{aligned}
$$

$$
\begin{aligned}
Cov(X_2, Y) &= \beta_1 Cov(X_1, X_2) + \beta_2 Var(X_2) \\
&= \beta_1 \phi_{12} + \beta_2 \phi_{22}
\end{aligned}
$$

Hold X_1 constant at fixed x_1

$$
Cov(X_2, Y | X_1 = x_1) = \beta_2 Var(X_2) = \beta_2 \phi_{22}
$$

Controlling Type I Error Probability

- Type I error is to reject H_0 when it is true, and there is actually no effect or no relationship
- Type I error is very bad. That's why Fisher called it an "error of the first kind."
- False knowledge is worse than ignorance.

Simulation study: Use pseudo-random number generation to create data sets

- Simulate data from the true model with $\beta_2=0$
- Fit naïve model
- Test H_0: $\beta_2=0$ at $\alpha = 0.05$ using naïve model
- Is H_0 rejected five percent of the time?

A **Big** Simulation Study (6 Factors)

- Sample size: n = 50, 100, 250, 500, 1000
- Corr(X_1,X_2): ϕ_{12} = 0.00, 0.25, 0.75, 0.80, 0.90
- 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

- 5x5x3x5x5x4 = 7,500 treatment combinations

Within each of the

- 5x5x3x5x5x4 = 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 naïve 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

Look at a small part of the results

- Both reliabilities = 0.90
- Everything is normally distributed
- β_0 = 1, β_1=1, β_2=0 (H_0 is true)

Weak Relationship between X_1 and Y: Var = 25%

N	Correlation Between X_1 and X_2 0.00	0.25	0.75	0.80	0.90
50	0.04760	0.05050	0.06360	0.07150	0.09130
100	0.05040	0.05210	0.08340	0.09400	0.12940
250	0.04670	0.05330	0.14020	0.16240	0.25440
500	0.04680	0.05950	0.23000	0.28920	0.46490
1000	0.05050	0.07340	0.40940	0.50570	0.74310

Moderate Relationship between X_1 and Y: Var = 50%

N	Correlation Between X_1 and X_2 0.00	0.25	0.75	0.80	0.90
50	0.04600	0.05200	0.09630	0.11060	0.16330
100	0.05350	0.05690	0.14610	0.18570	0.28370
250	0.04830	0.06250	0.30680	0.37310	0.58640
500	0.05150	0.07800	0.53230	0.64880	0.88370
1000	0.04810	0.11850	0.82730	0.90880	0.99070

Strong Relationship between X_1 and Y: Var = 75%

N	Correlation Between X_1 and X_2 0.00	0.25	0.75	0.80	0.90
50	0.04850	0.05790	0.17270	0.20890	0.34420
100	0.05410	0.06790	0.31010	0.37850	0.60310
250	0.04790	0.08560	0.64500	0.75230	0.94340
500	0.04450	0.13230	0.91090	0.96350	0.99920
1000	0.05220	0.21790	0.99590	0.99980	1.00000

Marginal Mean Type I Error Rates

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

Factors affecting severity of the problem

- 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 independent variables is safe, assuming that the latent independent variables are correlated.

The problem applies to other kinds of regression, and various 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 rate

If X_1 is randomly assigned

- Then it is independent of X_2: Zero correlation.
- So even if an experimentally manipulated variable is measured (implemented) with error, there will be no inflation of Type I error rate.
- If X_2 is randomly assigned and X_1 is a covariate observed with error (very common), then again there is no correlation between X_1 and X_2, and so no inflation of Type I error rate.
- Measurement error may decrease the precision of experimental studies, but in terms of Type I error it creates no problems.
- This is good news!

Need a statistical model that includes measurement error

Copyright Information