Large-sample target of least squares regression[1]

STA442/2101 Fall 2018

[1] See last slide for copyright information.

Background Reading: *Linear Models in Statistics* by Rencher and Schaalje

- Section 7.5 on the centered model
- Chapter 10 on random explanatory variables

Overview

1. The Centered Model
2. Estimation
3. Convergence
4. Model Mis-specification
5. Measurement Error

The centered model

Explanatory variable values are fixed, for now

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \epsilon_i \\
&= \beta_0 + \beta_1 \overline{x}_1 + \cdots + \beta_k \overline{x}_k \\
&\quad + \beta_1 (x_{i,1} - \overline{x}_1) + \cdots + \beta_k (x_{i,k} - \overline{x}_k) + \epsilon_i \\
&= \alpha_0 + \alpha_1 (x_{i,1} - \overline{x}_1) + \cdots + \alpha_k (x_{i,k} - \overline{x}_k) + \epsilon_i
\end{aligned}
$$

with

$$\alpha_0 = \beta_0 + \beta_1 \overline{x}_1 + \cdots + \beta_k \overline{x}_k.$$

$$\alpha_j = \beta_j \text{ for } j = 1, \ldots, k$$

This re-parameterization is one-to-one.

Invariance Principle

MLE of a function is that function of the MLE

- Since $\alpha_j = \beta_j$ for $j = 1, \ldots, k$, have $\widehat{\alpha}_j = \widehat{\beta}_j$ for $j = 1, \ldots, k$.
- Least-squares estimates are the same as MLEs under normality.
- So this conclusion applies to the least-squares estimates.
- When the explanatory variables are centered, the intercept of the least-squares plane changes, but the slopes remain the same.

Least-squares Estimation for the Centered Model

Working toward a useful formula

$$y_i = \alpha_0 + \beta_1(x_{i,1} - \overline{x}_1) + \cdots + \beta_k(x_{i,k} - \overline{x}_k) + \epsilon_i$$

Estimation:

- $\widehat{\alpha}_0 = \overline{y}$, regardless of the data.
- $\widehat{\beta}_j$ values are the same as for the uncentered model.
- To find the $\widehat{\beta}_j$ (once you have $\widehat{\alpha}_0 = \overline{y}$) minimize

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \overline{y} \ - \beta_1(x_{i,1} - \overline{x}_1) - \cdots - \beta_k(x_{i,k} - \overline{x}_k)^2$$

- This is the same as centering y as well as x, and fitting a regression through the origin.

Estimation of $\boldsymbol{\beta}$

- Center explanatory variables *and* the response variable by subtracting off sample means.
- Fit a regression through the origin.
- $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ as usual.
- But now the meaning of the notation is a little different because all the variables are centered.
- Again, this is the same as $\widehat{\boldsymbol{\beta}}$ for the uncentered model.

$\mathbf{X}^\top\mathbf{X}$ for the centered model

$k = 3$ example

$$\mathbf{X}^\top\mathbf{X} =$$

$$\begin{pmatrix} \sum_{i=1}^n (x_{i1}-\overline{x}_1)^2 & \sum_{i=1}^n (x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2) & \sum_{i=1}^n (x_{i1}-\overline{x}_1)(x_{i3}-\overline{x}_3) \\ \sum_{i=1}^n (x_{i2}-\overline{x}_2)(x_{i1}-\overline{x}_1) & \sum_{i=1}^n (x_{i2}-\overline{x}_2)^2 & \sum_{i=1}^n (x_{i2}-\overline{x}_2)(x_{i3}-\overline{x}_3) \\ \sum_{i=1}^n (x_{i3}-\overline{x}_3)(x_{i1}-\overline{x}_1) & \sum_{i=1}^n (x_{i3}-\overline{x}_3)(x_{i2}-\overline{x}_2) & \sum_{i=1}^n (x_{i3}-\overline{x}_3)^2 \end{pmatrix}$$

Multiply and divide by n, get

$$n \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)^2 & \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2) & \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(x_{i3}-\overline{x}_3) \\ \frac{1}{n}\sum_{i=1}^n (x_{i2}-\overline{x}_2)(x_{i1}-\overline{x}_1) & \frac{1}{n}\sum_{i=1}^n (x_{i2}-\overline{x}_2)^2 & \frac{1}{n}\sum_{i=1}^n (x_{i2}-\overline{x}_2)(x_{i3}-\overline{x}_3) \\ \frac{1}{n}\sum_{i=1}^n (x_{i3}-\overline{x}_3)(x_{i1}-\overline{x}_1) & \frac{1}{n}\sum_{i=1}^n (x_{i3}-\overline{x}_3)(x_{i2}-\overline{x}_2) & \frac{1}{n}\sum_{i=1}^n (x_{i3}-\overline{x}_3)^2 \end{pmatrix}$$

$$= n\widehat{\boldsymbol{\Sigma}}_x$$

$\mathbf{X}^\top\mathbf{y}$ for the centered model

Still for the $k = 3$ example

$$
\begin{aligned}
\mathbf{X}^\top\mathbf{y} &= \left(\begin{array}{c} \sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \\ \sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \\ \sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \end{array} \right) \\
&= n \left(\begin{array}{c} \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \\ \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \\ \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \end{array} \right) \\
&= n\widehat{\mathbf{\Sigma}}_{xy}
\end{aligned}
$$

$\mathbf{X}^\top\mathbf{X} = n\widehat{\boldsymbol{\Sigma}}_x$ and $\mathbf{X}^\top\mathbf{y} = n\widehat{\boldsymbol{\Sigma}}_{xy}$

For the centered model

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (n\widehat{\boldsymbol{\Sigma}}_x)^{-1} n\widehat{\boldsymbol{\Sigma}}_{xy} \\
&= \frac{1}{n}(\widehat{\boldsymbol{\Sigma}}_x)^{-1} n\widehat{\boldsymbol{\Sigma}}_{xy} \\
&= \widehat{\boldsymbol{\Sigma}}_x^{-1}\widehat{\boldsymbol{\Sigma}}_{xy}
\end{aligned}
$$

$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Sigma}}_x^{-1} \widehat{\boldsymbol{\Sigma}}_{xy}$

Where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)^\top$; the intercept is not included

The formula applies whether the data are centered or not, and whether the explanatory variables are fixed or random. Suppose they are random.

- $\widehat{\boldsymbol{\Sigma}}_x \stackrel{a.s.}{\rightarrow} \boldsymbol{\Sigma}_x$
- $\widehat{\boldsymbol{\Sigma}}_{xy} \stackrel{a.s.}{\rightarrow} \boldsymbol{\Sigma}_{xy}$
- Taking the inverse is a sequence of continuous operations.
- So by continuous mapping,

$$\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\Sigma}}_x^{-1} \widehat{\boldsymbol{\Sigma}}_{xy} \stackrel{a.s.}{\rightarrow} \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_{xy}$$

We have found $\widehat{\boldsymbol{\beta}}_n \stackrel{a.s.}{\rightarrow} \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}$

Is it the right target? There are two cases.

- The model is correct.
- The model is incorrect (mis-specified).

Correct model (*Uncentered*)

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

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i$$

where

- β_0 (the intercept) is an unknown scalar constant.
- $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters.
- $\mathbf{x}_i$ is a $k \times 1$ random vector with expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_x$.
- ϵ_i is a scalar random variable with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.
- $cov(\mathbf{x}_i, \epsilon_i) = \mathbf{0}$.

Calculate $cov(\mathbf{x}_i, y_i) = \boldsymbol{\Sigma}_{xy}$ for the *uncentered* model

$y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i$

$$
\begin{aligned}
cov(\mathbf{x}_i, y_i) &= E\left\{(\mathbf{x}_i - \boldsymbol{\mu})(y_i - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{\mu})^\top\right\} \\
&= E\left\{(\mathbf{x}_i - \boldsymbol{\mu})(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{\mu})^\top\right\} \\
&= E\left\{(\mathbf{x}_i - \boldsymbol{\mu})(\boldsymbol{\beta}^\top \mathbf{x}_i - \boldsymbol{\beta}^\top \boldsymbol{\mu} + \epsilon_i)^\top\right\} \\
&= E\left\{(\mathbf{x}_i - \boldsymbol{\mu})(\boldsymbol{\beta}^\top (\mathbf{x}_i - \boldsymbol{\mu}) + \epsilon_i)^\top\right\} \\
&= E\left\{(\mathbf{x}_i - \boldsymbol{\mu})\left((\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\beta} + \epsilon_i^\top\right)\right\} \\
&= E\left\{(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top\right\}\boldsymbol{\beta} + E\left\{(\mathbf{x}_i - \boldsymbol{\mu})(\epsilon_i - 0)^\top\right\} \\
&= \boldsymbol{\Sigma}_x \boldsymbol{\beta} + \mathbf{0}
\end{aligned}
$$

Convergence

Have $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_x \boldsymbol{\beta}$ for the uncentered model. So whether the variables are centered or not,

$$
\begin{array}{rcl}
\widehat{\boldsymbol{\beta}}_n & = & \widehat{\boldsymbol{\Sigma}}_x^{-1} \widehat{\boldsymbol{\Sigma}}_{xy} \\
 & \stackrel{a.s.}{\rightarrow} & \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_{xy} \\
 & = & \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_x \boldsymbol{\beta} \\
 & = & \boldsymbol{\beta}
\end{array}
$$

And $\widehat{\boldsymbol{\beta}}_n$ is strongly consistent for $\boldsymbol{\beta}$.

Model Mis-specification

What if the model is wrong (mis-specified)?

- Think of a particular way in which the regression model might be wrong.
- Call this the "true model."
- Still have $\widehat{\boldsymbol{\beta}}_n \stackrel{a.s.}{\rightarrow} \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}$.
- Calculate $\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}$ assuming the true model.
- This is the large-sample target of $\widehat{\boldsymbol{\beta}}$.
- Is it what you want?

Measurement Error

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

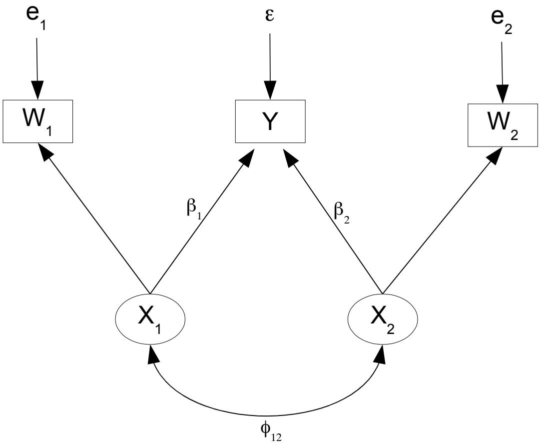
The problem with measurement error

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

- Trouble may arise if you take the regression model seriously as a model of how Y is produced from X.
- If your objective is pure prediction and *not interpretation*, there is no problem.
- In nature, there are relationships between true variables, and this is what we are interested in.
- Relationships between observable variables result from relationships between true variables, combined with the measurement error process.
- Measurement error does not just weaken the relationships.

Measurement error in two explanatory variables

An example



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

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

Reliability

As the term is used in psychometrics

$$\begin{aligned} W_{i,1} &= X_{i,1} + e_{i,1} \\ W_{i,2} &= X_{i,2} + e_{i,2}, \end{aligned}$$

where

$Var(X_{i,1}) = \phi_{11}$, $Var(X_{i,2}) = \phi_{22}$
$Var(e_{i,1}) = \omega_1$, $Var(e_{i,2}) = \omega_2$,

- Because X and e are independent, $Var(W) = Var(X) + Var(e) = \phi + \omega$.
- The proportion of the variance in W that comes from the "true" variable X (and not error) is $\frac{\phi}{\phi+\omega}$.
- Call it the "reliability."
- 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 $\widehat{\boldsymbol{\beta}}$ (especially $\widehat{\beta}_2$) as $n \to \infty$ when the true model holds.

$\widehat{\boldsymbol{\beta}}_n \stackrel{a.s.}{\to} \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{wy}$

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

Calculation of $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_{wy}$ by hand is not bad.

$$\boldsymbol{\Sigma}_w = \begin{pmatrix} \omega_1 + \phi_{11} & \phi_{12} \\ \phi_{12} & \omega_2 + \phi_{22} \end{pmatrix} \qquad \boldsymbol{\Sigma}_{wy} = \begin{pmatrix} \beta_1\phi_{11} + \beta_2\phi_{12} \\ \beta_1\phi_{12} + \beta_2\phi_{22} \end{pmatrix}$$

After some work

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

$$\widehat{\boldsymbol{\beta}}_n \stackrel{a.s.}{\rightarrow} \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\Sigma}_{wy} = \left(\begin{array}{c} \frac{\beta_2\omega_2\phi_{12}+\beta_1(\omega_2\phi_{11}+\phi_{11}\phi_{22}-\phi_{12}^2)}{(\phi_{1,1}+\omega_1)(\phi_{2,2}+\omega_2)-\phi_{12}^2} \\ \frac{\beta_1\omega_1\phi_{12}+\beta_2(\omega_1\phi_{22}+\phi_{11}\phi_{22}-\phi_{12}^2)}{(\phi_{1,1}+\omega_1)(\phi_{2,2}+\omega_2)-\phi_{12}^2} \end{array} \right) \neq \left(\begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right)$$

When $H_0 : \beta_2 = 0$ is true, this reduces to ...

The Target under $H_0 : \beta_2 = 0$

$$\begin{pmatrix} \widehat{\beta}_1 \\ \\ \widehat{\beta}_2 \end{pmatrix} \stackrel{a.s.}{\rightarrow} \begin{pmatrix} \beta_1 \left(\frac{\phi_{11}\phi_{22}-\phi_{12}^2}{(\phi_{1,1}+\omega_1)(\phi_{2,2}+\omega_2)-\phi_{12}^2} \right) \\ \\ \frac{\beta_1\omega_1\phi_{12}}{(\phi_{1,1}+\omega_1)(\phi_{2,2}+\omega_2)-\phi_{12}^2} \end{pmatrix}$$

Note $\phi_{11}\phi_{22} - \phi_{12}^2 = |\mathbf{\Sigma}_x|$, and $\omega_1 = Var(e_1)$, where $W_1 = X_1 + e_1$.

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 covariance between X_1 and X_2.

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

How bad is it for finite sample sizes?

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

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

Representative subset of the results

- All random variables are normally distributed.
- Both reliabilities equal 0.90.
- Separate slides for weak, moderate and strong relationship between X_1 and Y.

X_1 explains 25% of the variance in Y

Numbers in the table below are proportions of tests for which $H_0 : \beta_2 = 0$ was rejected in 10,000 simulated data sets.

	Correlation Between X1 and X2				
N	0.00	0.20	0.40	0.60	0.80
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

X_1 explains 50% of the variance in Y

Numbers in the table below are proportions of tests for which $H_0 : \beta_2 = 0$ was rejected in 10,000 simulated data sets.

	Correlation Between X1 and X2				
N	0.00	0.20	0.40	0.60	0.80
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

X_1 explains 75% of the variance in Y

Numbers in the table below are proportions of tests for which $H_0 : \beta_2 = 0$ was rejected in 10,000 simulated data sets.

	Correlation Between X1 and X2				
N	0.00	0.20	0.40	0.60	0.80
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

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.

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 probability.
- 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.
- Measurement error may decrease the precision of experimental studies, but in terms of Type I error it creates no problems.
- For observational studies, the news is not so good.

Observational studies

- Measurement error in the explanatory variables is almost universal.
- Standard statistical methods are almost guaranteed to yield inconsistent estimates.
- Conclusions may be incorrect – or they may not. With more than 2 explanatory variables, the impact of measurement error depends on the covariances between the x variables, in a complicated way.
- Instrumental variables can help.
- Statistical models that incorporate measurement error are available.
- But problems with identifiability prevent them from being applied to typical data sets.

Copyright Information