Introduction to Regression with Measurement Error

STA431: Spring 2013

Linear Regression

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

where $\epsilon_1, \ldots \epsilon_n$ are independent random variables with expected value zero and common variance σ^2, and $x_{i,1}, \ldots x_{i,p-1}$ are fixed constants.

Matrix Form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2 \mathbf{I}_n$

Are *X* values really constants?

Double Expectation

$$E\{Y\} = E\{E\{Y|X\}\}$$

E{Y} is a constant. E{Y|X} is a random variable, a function of X.

$$E\{E\{Y|X\}\} = \int E\{Y|X = x\}\, f(x)\, dx$$

Beta-hat is (conditionally) unbiased

$$E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x}\}=\boldsymbol{\beta}$$

Unbiased unconditionally, too

$$E\{\widehat{\boldsymbol{\beta}}\}=E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x}\}\}=E\{\boldsymbol{\beta}]=\boldsymbol{\beta}$$

Perhaps Clearer

$$
\begin{aligned}
E\{\widehat{\boldsymbol{\beta}}\} &= E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x}\}\} \\
&= \int \cdots \int E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x}\}\, f(\mathbf{x})\, d\mathbf{x} \\
&= \int \cdots \int \boldsymbol{\beta}\, f(\mathbf{x})\, d\mathbf{x} \\
&= \boldsymbol{\beta} \int \cdots \int f(\mathbf{x})\, d\mathbf{x} \\
&= \boldsymbol{\beta} \cdot 1 = \boldsymbol{\beta}.
\end{aligned}
$$

Conditional size α test, Critical region *A*

$$Pr\{F \in A | \mathbf{X} = \mathbf{x}\} = \alpha$$

$$\begin{aligned} Pr\{F \in A\} &= \int \cdots \int \alpha f(\mathbf{x})\, d\mathbf{x} \\ &= \alpha \int \cdots \int f(\mathbf{x})\, d\mathbf{x} \\ &= \alpha. \end{aligned}$$

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.

Correlate usual measurement with “Gold Standard?”

Not very realistic, except maybe for some bio-markers

Test-Retest

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

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

Equivalent measurements

Test-Retest Reliability

$$
\begin{aligned}
Corr(W_1, W_2) &= \frac{Cov(W_1, W_2)}{SD(W_1)SD(W_2)}, \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
Cov(W_1, W_2) &= Cov(\overset{c}{W}_1, \overset{c}{W}_2) \\
&= E(\overset{c}{W}_1 \overset{c}{W}_2) \\
&= E(\overset{c}{X} + e_1)(\overset{c}{X} + e_2) \\
&= E(\overset{c}{X}{}^2) + 0 + 0 + 0 \\
&= \sigma_X^2, \text{ so}
\end{aligned}
$$

$$
\begin{aligned}
Corr(W_1, W_2) &= \frac{\sigma_X^2}{\sqrt{\sigma_X^2 + \sigma_e^2}\sqrt{\sigma_X^2 + \sigma_e^2}} \\
&= \frac{\sigma_X^2}{\sigma_X^2 + \sigma_e^2}
\end{aligned}
$$

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

Calculate the sample correlation between

$$W_{1,1}, W_{2,1}, \ldots, W_{n,1}$$

$$W_{1,2}, W_{2,2}, \ldots, W_{n,2}$$

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

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 $\dfrac{\phi_{11}}{\phi_{11} + \omega_1}$

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

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 Rate

- 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 β_2=0
- Fit naïve model
- Test H_0: β_2=0 at α = 0.05 using naïve model
- Is H_0 rejected five percent of the time?

```
rmvn <- function(nn,mu,sigma)
# Returns an nn by kk matrix, rows are independent MVN(mu,sigma)
    {
    kk <- length(mu)
    dsig <- dim(sigma)
    if(dsig[1] != dsig[2]) stop("Sigma must be square.")
    if(dsig[1] != kk) stop("Sizes of sigma and mu are inconsistent.")
    ev <- eigen(sigma,symmetric=T)
    sqrl <- diag(sqrt(ev$values))
    PP <- ev$vectors
    ZZ <- rnorm(nn*kk) ; dim(ZZ) <- c(kk,nn)
    rmvn <- t(PP%*%sqrl%*%ZZ+mu)
    rmvn
    }# End of function rmvn
```

```
mereg <- function(beta0=1, beta1=1, beta2=0, sigmasq = 0.5,
          mu1=0, mu2=0, phi11=1, phi22=1, phi12 = 0.80,
          rel1=0.80, rel2=0.80, n=200)
####################################################################
# Model is    Y  = beta0 + beta1 X1 + beta2 X2 + epsilon
#             W1 = X1 + e1
#             W2 = W2 + e2
# Fit naive model
#             Y  = beta0 + beta1 W1 + beta2 W2 + epsilon
# Inputs are
#
#   beta0, beta1 beta2      True regression coefficients
#   sigmasq                 Var(epsilon)
#   mu1                     E(X1)
#   mu2                     E(X2)
#   phi11                   Var(X1)
#   phi22                   Var(X2)
#   phi12                   Cov(X1,X2) = Corr(X1,X1), because
#                           Var(X1) = Var(X2) = 1
#   rel1                    Reliability of W1
#   rel2                    Reliability of W2
#   n                       Sample size
# Note: This function uses rmvn, a multivariate normal random number
#       generator I wrote. The rmultnorm of the package MSBVAR does
#       the same thing but I am having trouble installing it.
####################################################################
```

```
{
# Calculate SD(e1) and SD(e2)
sd1 <- sqrt((phi11-rel1)/rel1)
sd2 <- sqrt((phi22-rel2)/rel2)
# Random number generation
epsilon <- rnorm(n,mean=0,sd=sqrt(sigmasq))
e1 <- rnorm(n,mean=0,sd=sd1)
e2 <- rnorm(n,mean=0,sd=sd2)
# X1 and X2 are bivariate normal. Need rmvn function.
Phi <- rbind(c(phi11,phi12),
             c(phi12,phi22))
X <- rmvn(n, mu=c(mu1,mu2), sigma=Phi) # nx2 matrix
X1 <- X[,1]; X2 <- X[,2]
# Now generate Y, W1 and W2

Y = beta0 + beta1*X1 + beta2*X2 + epsilon
W1 = X1 + e1
W2 = X2 + e2

# Fit the naive model
mereg <- summary(lm(Y~W1+W2))$coefficients
mereg # Returns table of beta-hats, SEs, t-statistics and p-values
} # End function mereg
```

```
> mereg()  # All the default values of inputs
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 0.9704708 0.05423489 17.893845 3.692801e-43
W1          0.6486972 0.06336434 10.237576 5.385982e-20
W2          0.2079601 0.06201811  3.353216 9.578634e-04
>
> mereg()[3,4] # Just the p-value for H0: beta2=0
[1] 0.0006340172
>
> # H0 rejected twice. Is the function okay?
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.03946133
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.2582209
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.08474088
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.5182614
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.2889913
```

```
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.1667587
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.4414364
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.2268087
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.8298779
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.3508289
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.05173589
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.243059
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.8818203
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.3430994
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.4860574
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.9644776
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.09245873
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.04757209
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.7947851
> mereg(rel1=1,rel2=1)[3,4] # No measurement error
[1] 0.8039931
```

Try it with measurement error

```
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 0.01080889
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 0.0007349183
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 0.01884786
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 0.003615565
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 0.003421935
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 3.895541e-07
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 3.328842e-07
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 0.0754436
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 0.0001274642
> mereg()[3,4] # Reliabilities both equal 0.80
[1] 6.900713e-05
```

9 out of 10

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 β_2=0

- Fit naïve model, test H_0: β_2=0 at α = 0.05
- Proportion of times H_0 is rejected is a Monte Carlo estimate of the Type I Error Rate

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 rates.
- 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!

What is going on theoretically?

First, need to look at some large-sample tools

Sample Space Ω, ω an element of Ω

- Observing whether a single individual is male or female:

$$\Omega = \{F, M\}$$

- Pair of individuals and observed their genders in order:

$$\Omega = \{(F, F), (F, M), (M, F), (M, M)\}$$

- Select *n* people and count the number of females:

$$\Omega = \{0, \ldots, n\}$$

- For limits problems, the points in Ω are infinite sequences

Random variables are functions from Ω into the set of real numbers

$$Pr\{X \in B\} = Pr(\{\omega \in \Omega : X(\omega) \in B\}$$

Random sample $X_1(\omega), \ldots, X_n(\omega)$

$$T = T(X_1, \ldots, X_n)$$

$$T = T_n(\omega)$$

$$\text{Let } n \to \infty$$

To see what happens for large samples

Modes of Convergence

- Almost Sure Convergence
- Convergence in Probability
- Convergence in Distribution

Almost Sure Convergence

We say that T_n converges *almost surely* to T, and write $T_n \stackrel{a.s.}{\to}$ if

$$Pr\{\omega : \lim_{n \to \infty} T_n(\omega) = T(\omega)\} = 1.$$

Acts like an ordinary limit, except possibly on a set of probability zero.

All the usual rules apply.

Strong Law of Large Numbers

$$\overline{X}_n \xrightarrow{a.s.} \mu$$

The only condition required for this to hold is the existence of the expected value.

Let X_1, …, X_n be independent and identically distributed random variables; let X be a general random variable from this same distribution, and $Y=g(X)$

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{a.s.} E(Y)$$

$$= E(g(X))$$

So for example

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \stackrel{a.s.}{\rightarrow} E(X^k)$$

$$\frac{1}{n}\sum_{i=1}^{n} U_i^2 V_i W_i^3 \stackrel{a.s.}{\rightarrow} E(U^2 V W^3)$$

That is, sample moments converge almost surely to population moments.

Convergence in Probability

We say that T_n converges *in probability* to T, and write $T_n \xrightarrow{P}$ if for all $\epsilon > 0$,

$$\lim_{n \longrightarrow \infty} P\{|T_n - T| < \epsilon\} = 1$$

Almost Sure Convergence => Convergence in Probability

Strong Law of Large Numbers => Weak Law of Large Numbers

Convergence in Distribution

Denote the cumulative distribution functions of $T_1, T_2, \ldots$ by $F_1(t), F_2(t), \ldots$ respectively, and denote the cumulative distribution function of T by $F(t)$.

We say that T_n converges *in distribution* to T, and write $T_n \xrightarrow{d} T$ if for every point t at which F is continuous,

$$\lim_{n \to \infty} F_n(t) = F(t)$$

Central Limit Theorem says

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim N(0,1)$$

Connections among the Modes of Convergence

- $T_n \xrightarrow{a.s.} T \Rightarrow T_n \xrightarrow{P} T \Rightarrow T_n \xrightarrow{d} T.$
- If a is a constant, $T_n \xrightarrow{d} a \Rightarrow T_n \xrightarrow{P} a.$

Consistency

$T_n = T_n(X_1, \ldots, X_n)$ is a statistic estimating a parameter θ

The statistic T_n is said to be *consistent* for θ if $T_n \stackrel{P}{\rightarrow} \theta$.

$$\lim_{n\rightarrow\infty} P\{|T_n - \theta| < \epsilon\} = 1$$

The statistic T_n is said to be *strongly consistent* for θ if $T_n \stackrel{a.s.}{\rightarrow} \theta$.

Strong consistency implies ordinary consistency.

Consistency is great but it's not enough

- It means that as the sample size becomes indefinitely large, you (probably) get as close as you like to the truth.
- It's the least we can ask. Estimators that are not consistent are completely unacceptable for most purposes.

$$T_n \stackrel{a.s.}{\rightarrow} \theta \Rightarrow U_n = T_n + \frac{100,000,000}{n} \stackrel{a.s.}{\rightarrow} \theta$$

Consistency of the Sample Variance

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2$$

By SLLN, $\overline{X}_n \stackrel{a.s.}{\to} \mu$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \stackrel{a.s.}{\to} E(X^2) = \sigma^2 + \mu^2$

Because the function $g(x, y) = x - y^2$ is continuous,

$$\widehat{\sigma}_n^2 = g(\frac{1}{n}\sum_{i=1}^{n} X_i^2, \overline{X}_n) \stackrel{a.s.}{\to} g(\sigma^2 + \mu^2, \mu) = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

Consistency of the Sample Covariance

$$\widehat{\sigma}_{1,2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}_n \overline{Y}_n$$

By SLLN, $\overline{X}_n \stackrel{a.s.}{\rightarrow} E(X)$, $\overline{Y}_n \stackrel{a.s.}{\rightarrow} E(Y)$, and $\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \stackrel{a.s.}{\rightarrow} E(XY)$

Because the function $g(x, y, z) = x - yz$ is continuous,

$$\begin{aligned}
\widehat{\sigma}_{1,2} &= g\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i, \overline{X}_n, \overline{Y}_n\right) \stackrel{a.s.}{\rightarrow} g\left(E(XY), E(X), E(Y)\right) \\
&= E(XY) - E(X)E(Y) = Cov(X, Y) \\
&= \sigma_{1,2}
\end{aligned}$$

Single Independent 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

$$
Y_i = \beta_0 + \beta_1 W_i + \epsilon_i
$$

where independently for $i = 1, \ldots, n$, $Var(X_i) = \sigma_X^2$, $Var(e_i) = \sigma_e^2$, and X_i, e_i, ϵ_i are all independent.

Least squares estimate of β_1 for the Naïve Model

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

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

- Goes to the true parameter times reliability of *W*.
- Asymptotically biased toward zero, because reliability is between zero and one.
- No asymptotic bias when β_1=0.
- No inflation of Type I error rate
- Loss of power when $\beta_1 \neq 0$
- Measurement error just makes relationship seem weaker than it is. Reassuring, but watch out!

Two Independent variables, β_2=0

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

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

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

Least squares estimate of β_2 for the Naïve Model when true $\beta_2 = 0$

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

$$= \left(\frac{\omega_1}{\phi_{1,1} + \omega_1} \right) \left(\frac{\beta_1 \phi_{1,2}}{\phi_{2,2} + \omega_2} \right)$$

Combined with estimated standard error going almost surely to zero,
Get *t* statistic for H_0: $\beta_2 = 0$ going to $\pm\infty$, and p-value going almost
Surely to zero, unless ....

Combined with estimated standard error going almost surely to zero, get *t* statistic for H_0: $\beta_2 = 0$ going to $\pm\infty$, and p-value going almost surely to zero, unless ....

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

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

And, anything that increases $Var(W_2)$ will decrease the bias.

Need a statistical model that includes measurement error

Copyright Information