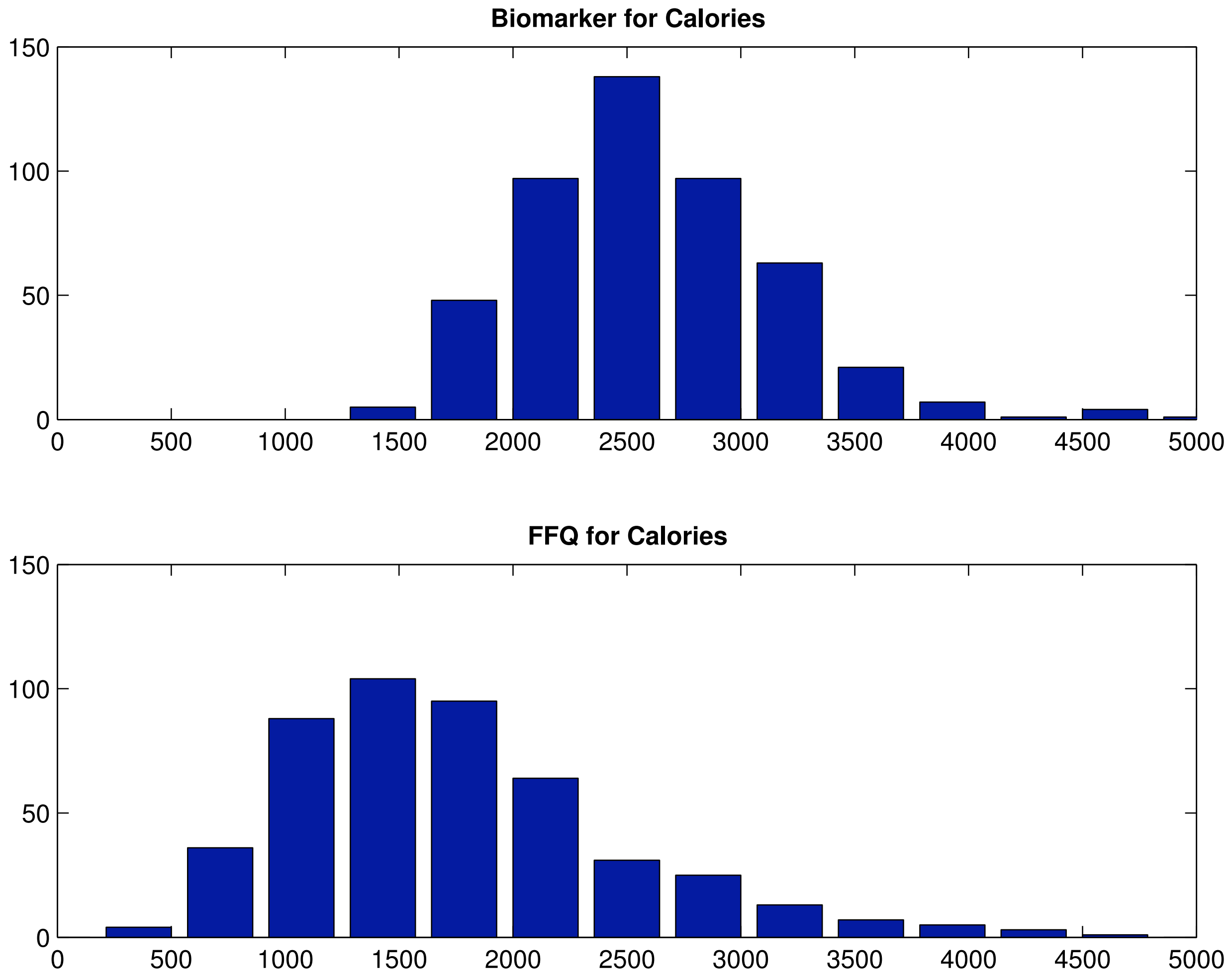


Figure 1.5 *OPEN Study data, histograms of energy (calories) using a biomarker (top panel) and a food frequency questionnaire (bottom panel). Note how individuals report far fewer calories than they actually consume.*

Measurement Error

- Exercise
- Income
- Snack food consumption
- 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?

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).

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

The consequences of ignoring measurement error in the independent variables

Measurement error in the dependent variable is a less serious problem

Y is a latent variable; X and V are observable

$$
\begin{aligned}
Y &= \beta_0 + \beta X + \epsilon_1 \\
V &= \nu_0 + \lambda Y + \epsilon_2 \\
&= \nu_0 + \lambda(\beta_0 + \beta X + \epsilon_1) + \epsilon_2 \\
&= (\nu_0 + \lambda\beta_0) + \lambda\beta X + (\lambda\epsilon_1 + \epsilon_2)
\end{aligned}
$$

Re-parameterize

Two Models

- 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

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

Test X_2 “controlling for” (holding constant) X_1

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

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

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

Need to control Type I error rate

```
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 $\beta_2=0$

- 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 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%

	Correlation Between X_1 and X_2				
N	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%

	Correlation Between X_1 and X_2				
N	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%

	Correlation Between X_1 and X_2				
N	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 W1
0.50         0.75         0.80         0.90         0.95
0.60637233   0.46983147   0.42065313   0.26685820   0.14453913


                      Reliability of W2
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, but there is a lot of bad news.

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 $\beta_2 = 0$

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

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 ±∞, 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

Perhaps the simplest case

$$\begin{aligned} Y &= \beta X + \epsilon \\ W &= X + e \end{aligned}$$

$X \sim N(0, \phi)$

$$\left[\begin{array}{c} W \\ Y \end{array}\right] \sim N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \boldsymbol{\Sigma} \right)$$

$\epsilon \sim N(0, \psi)$

$e \sim N(0, \omega)$

X, ϵ, e independent

All expected values zero

X is a latent variable; W and Y are observable

$$\begin{aligned} W &= X + e \\ Y &= \beta X + \epsilon \end{aligned}$$

$$\boldsymbol{\Sigma} = V \left[\begin{array}{c} W \\ Y \end{array} \right] = \left(\begin{array}{rr} \phi + \omega & \beta\phi \\ \beta\phi & \beta^2\phi + \psi \end{array} \right)$$

- Observable data are bivariate normal with mean zero and covariance matrix Sigma.
- With increasing sample size, all you can get is a better and better estimate of Sigma.

Cannot recover $\boldsymbol{\theta} = (\phi, \omega, \beta, \psi)$ from $(\sigma_{11}, \sigma_{12}, \sigma_{22})$.

Cannot recover $\boldsymbol{\theta} = (\phi, \omega, \beta, \psi)$ from $(\sigma_{11}, \sigma_{12}, \sigma_{22})$.

$$\boldsymbol{\Sigma} = \left(\begin{array}{cc} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{array} \right) = \left(\begin{array}{rr} \phi + \omega & \beta\phi \\ \beta\phi & \beta^2\phi + \psi \end{array} \right)$$

Let $\boldsymbol{\Sigma} = [\sigma_{ij}]$ be *any* 2×2 positive definite symmetric matrix, and let the other parameters be functions of ϕ as follows.

- $\omega = \sigma_{11} - \phi$
- $\beta = \frac{\sigma_{12}}{\phi}$
- $\psi = \sigma_{22} - \sigma_{12}\phi$

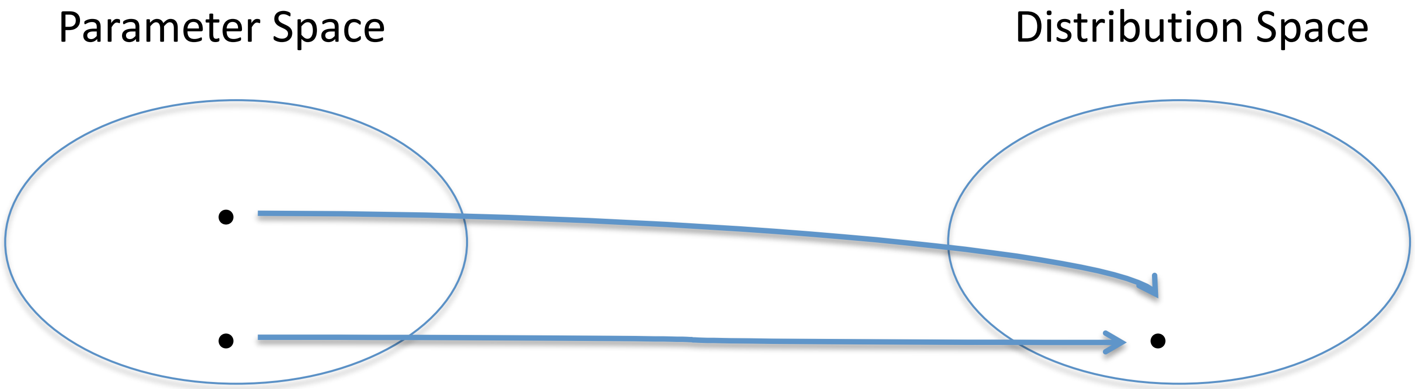
Every $\phi \in (0, \sigma_{11})$, yields the same $\boldsymbol{\Sigma}$.

For every possible (bivariate normal) distribution

- Infinitely many sets of different parameter values yield that distribution
- MLE is not unique
- Lots of trouble

Identifiability

Suppose a statistical model implies $\mathbf{D} \sim P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta$. If no two points in Θ yield the same probability distribution, then the parameter $\boldsymbol{\theta}$ is said to be *identifiable.* On the other hand, if there exist $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in Θ with $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$, the parameter $\boldsymbol{\theta}$ is *not identifiable.*



Consistent Estimation is Impossible

Suppose $\theta_1 \neq \theta_2$ with $P_{\theta_1} = P_{\theta_2}$

θ_1
θ_2

Need more information

- Bigger sample size will not help
- Sometimes, information from other studies may help. Recall simple naïve regression:

$$\widehat{\beta}_1 \overset{a.s.}{\rightarrow} \beta_1 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_e^2} \right)$$

- If you knew the reliability of W you could correct the estimator.
- Or, more variables can sometimes solve the problem.

Double measurement

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ W_{i,1} &= \nu_1 + X_i + e_{i,1} \\ W_{i,2} &= \nu_2 + X_i + e_{i,2}, \end{aligned}$$

$$E\left(\begin{array}{c} W_{i,1} \\ W_{i,1} \\ Y_i \end{array}\right) = \boldsymbol{\mu} = \left(\begin{array}{c} \mu_1 \\ \mu_2 \\ \mu_3 \end{array}\right) = \left(\begin{array}{c} \mu_x + \nu_1 \\ \mu_x + \nu_2 \\ \beta_0 + \beta_1 \mu_x \end{array}\right)$$

$$V\left(\begin{array}{c} W_{i,1} \\ W_{i,1} \\ Y_i \end{array}\right) = \boldsymbol{\Sigma} = [\sigma_{i,j}] = \left[\begin{array}{ccc} \phi + \omega_1 & \phi & \beta_1 \phi \\ & \phi + \omega_2 & \beta_1 \phi \\ & & \beta_1^2 \phi + \psi \end{array}\right]$$

Double Measurement Regression: A Two-Stage Model

$$\begin{array}{rcl} \mathbf{Y}_i & = & \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i \\ \mathbf{F}_i & = & \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\ \mathbf{D}_{i,1} & = & \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\ \mathbf{D}_{i,2} & = & \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2} \end{array}$$

Observable variables are $D_{i,1}$ and $D_{i,2}$: both p+q by 1

$$\begin{aligned}\mathbf{Y}_i &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i \\ \mathbf{F}_i &= \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\ \mathbf{D}_{i,1} &= \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\ \mathbf{D}_{i,2} &= \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2}\end{aligned}$$

$$E(\mathbf{X}_i) = \boldsymbol{\mu}_x,$$

$$\begin{aligned}E(\mathbf{D}_{i,1}) &= \left(\begin{array}{c} \boldsymbol{\mu}_{1,1} \\ \boldsymbol{\mu}_{1,2} \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\nu}_{1,1} + E(\mathbf{X}_i) \\ \boldsymbol{\nu}_{1,2} + E(\mathbf{Y}_i) \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\nu}_{1,1} + \boldsymbol{\mu}_x \\ \boldsymbol{\nu}_{1,2} + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{\mu}_x \end{array} \right) \\ E(\mathbf{D}_{i,2}) &= \left(\begin{array}{c} \boldsymbol{\mu}_{2,1} \\ \boldsymbol{\mu}_{2,2} \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\nu}_{2,1} + E(\mathbf{X}_i) \\ \boldsymbol{\nu}_{2,2} + E(\mathbf{Y}_i) \end{array} \right) = \left(\begin{array}{c} \boldsymbol{\nu}_{2,1} + \boldsymbol{\mu}_x \\ \boldsymbol{\nu}_{2,2} + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{\mu}_x \end{array} \right)\end{aligned}$$

Even with knowledge of β_1, identifying the expected values and intercepts is hopeless.

$$\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\
\mathbf{D}_{i,1} &= \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\
\mathbf{D}_{i,2} &= \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2}
\end{aligned}$$

$$V(\mathbf{X}_i) = \boldsymbol{\Phi}_{11},\ V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi},\ V(\mathbf{e}_{i,1}) = \boldsymbol{\Omega}_1,\ V(\mathbf{e}_{i,2}) = \boldsymbol{\Omega}_2,$$

$$\mathbf{X}_i,\ \boldsymbol{\epsilon}_i,\ \mathbf{e}_{i,1} \text{ and } \mathbf{e}_{i,2} \text{ independent.}$$

The main idea is that **D**$_1$ and **D**$_2$ are independent measurements of **F**, perhaps at different times using different methods. Measurement errors may be correlated within occasions (even For IV and DV), but not between occasions.

$$\begin{array}{rcl}\mathbf{Y}_i & = & \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i \\ \mathbf{F}_i & = & \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\ \mathbf{D}_{i,1} & = & \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\ \mathbf{D}_{i,2} & = & \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2} \end{array}$$

$$V(\mathbf{X}_i) = \boldsymbol{\Phi}_{11},\, V(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi},\, V(\mathbf{e}_{i,1}) = \boldsymbol{\Omega}_1,\, V(\mathbf{e}_{i,2}) = \boldsymbol{\Omega}_2,$$

$\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$, $\mathbf{e}_{i,1}$ and $\mathbf{e}_{i,2}$ independent.

Stage One

$$V(\mathbf{F}_i) = \boldsymbol{\Phi} = \left(\begin{array}{cc} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}_{12}^\prime & \boldsymbol{\Phi}_{22} \end{array} \right) = \left(\begin{array}{cc} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{11}\boldsymbol{\beta}_1^\prime \\ \boldsymbol{\beta}_1 \boldsymbol{\Phi}_{11} & \boldsymbol{\beta}_1 \boldsymbol{\Phi}_{11}\boldsymbol{\beta}_1^\prime + \boldsymbol{\Psi} \end{array} \right)$$

$\boldsymbol{\Phi}_{11}$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\Psi}$ can be recovered from $\boldsymbol{\Phi}$

The Measurement Model (Stage 2)

$$\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i \\
\mathbf{F}_i &= \left(\begin{array}{c} \mathbf{X}_i \\ \mathbf{Y}_i \end{array} \right) \\
\mathbf{D}_{i,1} &= \boldsymbol{\nu}_1 + \mathbf{F}_i + \mathbf{e}_{i,1} \\
\mathbf{D}_{i,2} &= \boldsymbol{\nu}_2 + \mathbf{F}_i + \mathbf{e}_{i,2}
\end{aligned}$$

$$\boldsymbol{\Sigma} = V \left(\begin{array}{c} \mathbf{D}_{i,1} \\ \mathbf{D}_{i,2} \end{array} \right) = \left(\begin{array}{cc} \boldsymbol{\Phi} + \boldsymbol{\Omega}_1 & \boldsymbol{\Phi} \\ \boldsymbol{\Phi} & \boldsymbol{\Phi} + \boldsymbol{\Omega}_2 \end{array} \right)$$

$\boldsymbol{\Phi}$, $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ can easily be recovered from $\boldsymbol{\Sigma}$

All the parameters in the covariance matrix are identifiable

- $\mathbf{\Phi}$, $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$ may be recovered from $\mathbf{\Sigma}$
- $\mathbf{\Phi}_{11}$, $\boldsymbol{\beta}_1$ and $\mathbf{\Psi}$ may be recovered from $\mathbf{\Phi}$

- Correlated measurement error within sets is allowed (a big plus), because it’s reality
- Correlated measurement error between sets must be ruled out by careful data collection
- No need to do the calculations ever again

The BMI Health Study

- Body Mass Index: Weight in Kilograms divided by Height in Meters Squared
- Under 18 means underweight, Over 25 means overweight, Over 30 means obese
- High BMI is associated with poor health, like high blood pressure and high cholesterol
- People with high BMI tend to be older and fatter
- **BUT**, what if you have a high BMI but are in good physical shape – low percent body fat?

The Question

- If you control for age and percent body fat, is BMI still associated with indicators for poor health?
- But percent body fat (and to a lesser extent, age) are measured with error. Standard ways of controlling for them with regression are highly suspect.
- Use the double measurement design.

True variables (all latent)

- X_1 = Age
- X_2 = BMI
- X_3 = Percent body fat
- Y_1 = Cholesterol
- Y_2 = Diastolic blood pressure

Measure twice with different personnel at different locations and by different methods

	Measurement Set One	Measurement Set Two
Age	Self report	Passport or Birth Certificate
BMI	Dr. Office Measurement	Lab technician, no shoes, gown
% Body Fat	Tape and calipers	Submerge in water tank
Cholesterol	Lab 1	Lab 2
Diastolic BP	Blood pressure cuff, Dr. Office	Digital readout, mostly automatic

Set two is of generally higher quality

Correlation of measurement errors is less likely between sets