The consequences of ignoring measurement error in the independent variables

Measurement error in the dependent variable is a less serious problem; we will deal with it later.

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

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

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 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

- 5x5x3x5x5x5 = 7,500 treatment combinations

Within each of the

- 5x5x3x5x5x5 = 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

Estimated Type I Error Rates:
Base Distribution Normal, both reliabilities = 0.90

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 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