

Lecture 12: Measurement Error

BIO144 Data Analysis in Biology

Owen Petchey & Stephanie Muff

University of Zurich

22 May, 2022

- ▶ Measurement error (ME) in explanatory variable (x) and in the response (y) of regression models.
- ▶ Effects of ME on regression parameters.
- ▶ When do I have to worry?
- ▶ Simple methods to correct for ME.

Course material covered today

The lecture material of today is partially based on the following literature:

- ▶ Chapter 6.1 in “Lineare regression” (BC reading)

Sources of measurement error (ME)

- ▶ **Measurement imprecision** in the field or in the lab (length, weight, blood pressure, etc.).
- ▶ Errors due to **incomplete** or **biased observations** (e.g., self-reported dietary aspects, health history).
- ▶ Biased observations due to **preferential sampling or repeated observations**.
- ▶ Rounding error, digit preference.
- ▶ **Misclassification error** (e.g., exposure or disease classification).
- ▶ ...

"Error" is often used synonymous to "uncertainty".

Another fundamental assumption (often neglected!)

- ▶ It is a **fundamental assumption** that explanatory variables are measured or estimated **without error**, for instance for
- ▶ the calculation of correlations.
- ▶ linear regression and ANOVA.
- ▶ Generalized linear and non-linear regressions (e.g. logistic and Poisson).
- ▶ Violation of this assumption may lead to **biased** parameter estimates, altered standard errors and p -values, incorrect variable importances, and to **misleading conclusions**.
- ▶ Even standard statistics textbooks do often not mention these problems.

Measurement error in the explanatory variables (x) violates an assumption of standard regression analyses!!

Classical measurement error

A very common error type:

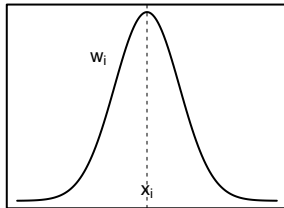
Let x_i be the *correct but unobserved* variable and w_i the observed variable with error u_i .
Then

$$w_i = x_i + u_i, \quad u_i \sim N(0, \sigma_u^2)$$

is the **classical ME model**.

(continued)

```
par(mar=c(0.1,0.1,0.1,0.1))
tx<-seq(-4,4,0.01)
par(mfrow=c(1,1))
plot(x = tx, dnorm(tx,0,1),type="l",xaxt="n",yaxt="n",xlab="",ylab="")
abline(v=0,lty=2,ldw=0.5)
text(0,0.02,labels=expression(x[i]),cex=0.6)
text(-1.5,0.3,labels=expression(w[i]),cex=0.6)
```



Examples: Imprecise measurements of a concentration, a mass, a length etc. → The observed value w_i varies around the true value x_i .

Illustration of the problem

Find regression parameters β_0 and β_x for the model with explanatory variable \mathbf{x} :

$$y_i = 1 \cdot x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

```
library(ggplot2)
set.seed(84522)
n <- 100
beta_0 <- 0
beta_1 <- 1
epsilon <- rnorm(n, 0, 0.2)
x <- rnorm(n, 0, 1)
u <- rnorm(n, 0, 1)
w <- x + u
# Classical
y <- beta_0 + beta_1*x + epsilon
m1 <- lm(y~x)
```


Plotting

```
cols <- c("X"="black", "W"="red")
ggplot(mapping=aes(x=x, y=y, col="X")) +
  geom_smooth(method="lm") + geom_point(size=0.9) +
  scale_colour_manual(name="Variable", values=cols) +
  xlab("X") +
  xlim(c(-3.5, 3.5)) +
  ylim(c(-2.7, 2.7)) +
  theme_bw()
```

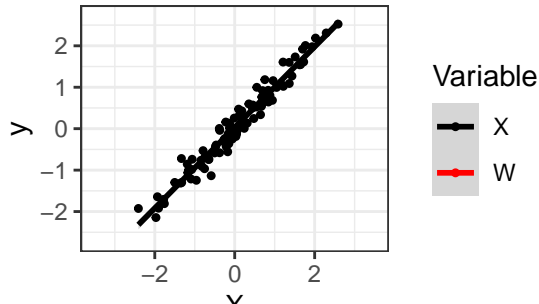


Illustration of the problem II

However, assume that only an erroneous proxy \mathbf{w} is observed with classical ME

$$w_i = x_i + u_i, \quad u_i \sim N(0, \sigma_u^2), \quad \sigma_u^2 = \sigma_x^2$$

```
w <- x + u ## error added to x in this step
```

```
## Classical
```

```
y <- beta_0 + beta_1*x + epsilon
```

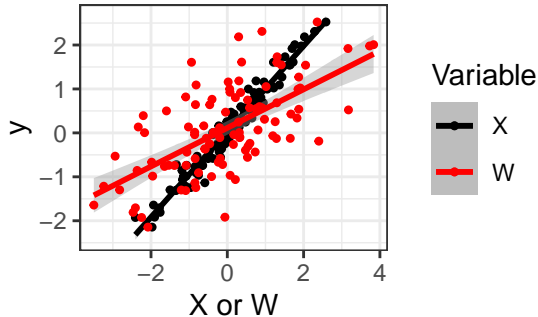
```
m1 <- lm(y~x)
```

```
## Now with added error
```

```
m2 <- lm(y~w)
```

Plotting

```
cols <- c("X"="black","W"="red")
ggplot(mapping=aes(x=x,y=y,color="X")) + geom_smooth(method="lm") +
  geom_point(size=0.9) +
  geom_smooth(mapping=aes(x=w,y=y,color="W"),method="lm") +
  geom_point(mapping=aes(x=w,y=y,color="W"),size=0.9) +
  scale_colour_manual(name="Variable",values=cols) + xlab("X or W") +
  theme_bw()
```



A tool you can have a play with...

► Illustration in a browser application

Classical measurement error in linear, logit and Poisson regression

Select regression model:

You can now check the effect of classical measurement error in a covariate of linear, logistic and Poisson regression. The linear predictor of the model is given as

$$\eta = \beta_0 + \beta_1 \cdot x + e$$

but covariate x is not directly observable. Instead, a substitute

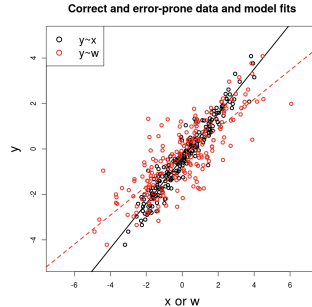
$$w = x + u$$

is observed, assuming that

$$u \sim N(0, \sigma_u^2).$$

To check what happens when the error increases, simply move the slider below to increase the error variance in the x covariate.

Select an error variance σ_u^2 (while $\sigma_e^2 = 1.$) :



The slope parameter of the error prone dataset is estimated as 0.64 (true slope: 1.0).
 The residual variance of the error prone model is estimated as 0.89 (true value: 0.25).

The “Triple Whammy of Measurement Error”

(Carroll et al. 2006)

- 1 **Bias**: The inclusion of erroneous variables in downstream analyses may lead to biased parameter estimates.
- 2 ME leads to a **loss of power** for detecting signals.
- 3 ME **masks important features** of the data, making graphical model inspection difficult.

How to correct for error?

- ▶ Generally, to correct for the error we need an **error model** and knowledge of the **error model parameters** **Example** If classical error $w_i = x_i + u_i$ with $u_i \sim N(0, \sigma_u^2)$ is present, knowledge of the **error variance** σ_u^2 is needed.

Strategy: Take repeated measurements to estimate the error variance!

- ▶ In **simple cases**, formulas for the bias exist.
- ▶ In most cases, such simple relations don't exist. Specific error modeling methods are then needed!

Attenuation in simple linear regression

Given the simple linear regression equation $y_i = \beta_0 + \beta_x x_i + \epsilon_i$ with $w_i = x_i + u_i$.
Assume that w_i instead of x_i is used in the regression:

$$y_i = \beta_0^* + \beta_x^* w_i + \epsilon_i$$

The **naive slope parameter** β_x^* is then underestimated with respect to the true slope β_x , with **attenuation factor** λ :

$$\beta_x^* = \underbrace{\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \right)}_{=\lambda} \beta_x$$

→ knowing σ_u^2 and σ_x^2 , the correct slope can be retrieved!

Example: $\sigma_x^2 = 5$, $\sigma_u^2 = 1$, → $\lambda = \frac{5}{6} = 0.83$

Error modeling

The **two most popular approaches**:

- ▶ **SIMEX**: SIMulation EXtrapolation, a heuristic and intuitive idea.
- ▶ **Bayesian methods**: Prior information about the error enters a model. Then use

$$\text{Likelihood} \times \text{prior} = \text{posterior}$$

to calculate the parameter distribution after error correction.

In any case, assessing the biasing effect of the error, as well as error modeling, can be done **only if the error structure (model) and the respective model parameters** (e.g., error variances) **are known!**

Therefore: Information about the error mechanism is essential, and potential errors must be identified already in the planning phase.

SIMEX: A very intuitive idea

Suggested by Cook & Stefanski (1994).

Idea:

- ▶ **Simulation phase:** The error in the data is progressively aggravated in order to determine how the quantity of interest is affected by the error.
- ▶ **Extrapolation phase:** The observed trend is then extrapolated back to a hypothetical error-free value.

Illustration of the SIMEX idea

Parameter of interest: β_x (e.g. a regression slope).

Problem: The respective explanatory variable x was estimated with error:

$$w = x + u, \quad u \sim N(0, \sigma_u^2)$$

```
set.seed(212356)
sigmax <- 1      # variance in x
sigmau <- 0.25   # measurement error

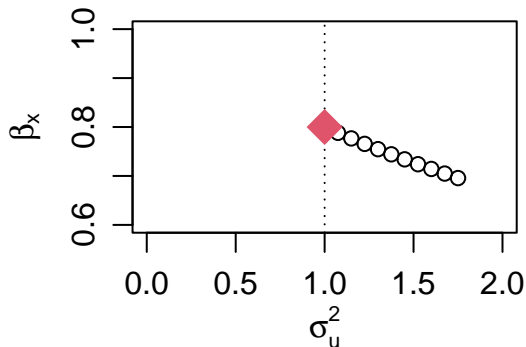
B <- 1 # True Beta
n <- 4 # number of measurements

# Additional observation error (simulated)
s <- seq(0, 3, 0.3)/n # note s=0 is the observed

# Observed Betas ~ sigmau
Bo <- B*sigmax/(sigmax + sigmau*(1+s))
```

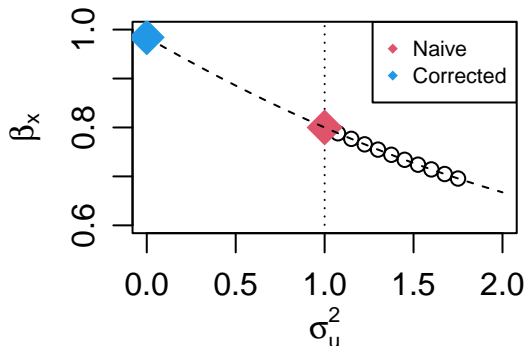
Plot

```
# plot the simulated data
par(mar=c(3,4,0.5,1), mgp=c(2.2,0.8,0))
plot((1+s), Bo, ylim=c(0.6,1), xlim=c(0,2), xlab=expression(sigma[u]^2),
     ylab=expression(beta[x]))
abline(v=1, lty=3)
points((1+s[1]),Bo[1],pch=18,cex=2.5,col=2) # highlight Bo
```



Find corrected Beta

```
print(p) # add plot again from above
model <- lm(Bo ~ poly(s,2)) # fit a quadratic line
newx <- seq(-1,1,0.1) # fit line to demonstrate fit
lines((1+newx), predict(model,newdata=data.frame(s=newx)), lty=2)
Be <- predict(model,newdata=data.frame(s=-1)) # B at s=-1 -> sigma_u=0
points(0,Be,pch=18,cex=2.5,col=4) # add to plot
legend("topright",legend=c("Naive","Corrected"),pch=18,col=c(2,4),cex=0.7)
```



Example of SIMEX use (part 1)

Let's consider a linear regression model

$$y_i = \beta_0 + \beta_x x_i + \beta_z z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

with

- ▶ $\mathbf{y} = (y_1, \dots, y_{100})^\top$: variable with % Bodyfat of 100 individuals.
- ▶ $\mathbf{x} = (x_1, \dots, x_{100})^\top$ the BMI of the individuals.

****Problem:*** The BMI was self-reported and thus suffers from measurement error! Not x_i are observed, but rather

$$w_i = x_i + u_i, \quad u_i \sim N(0, 4)$$

- ▶ $\mathbf{z} = (z_1, \dots, z_{100})^\top$ a binary explanatory variable that indicates if the i -th person was a male ($z_i = 1$) or female ($z_i = 0$).

→ Apply the SIMEX procedure!

Simulated example

```
set.seed(3243445)
x <- rnorm(100,24,4)
w <- x + rnorm(100,0,2)
z <- ifelse(x>25, rbinom(100,1,0.7), rbinom(100,1,0.3))
y <- -15 + 1.6*x - 2*z + rnorm(100,0,3)
data <- data.frame(cbind(w,z,y))
names(data) <- c("BMI", "sex", "bodyfat")
```

Check out the results

Use the error-prone BMI variable to fit a “naive” regression:

```
r.lm <- lm(bodyfat ~ BMI + sex,data,x=TRUE)  
summary(r.lm)$coef
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-8.003714	2.07060335	-3.865402	2.005407e-04
## BMI	1.271558	0.08821382	14.414504	7.478782e-26
## sex	-1.951735	0.73625960	-2.650879	9.376840e-03

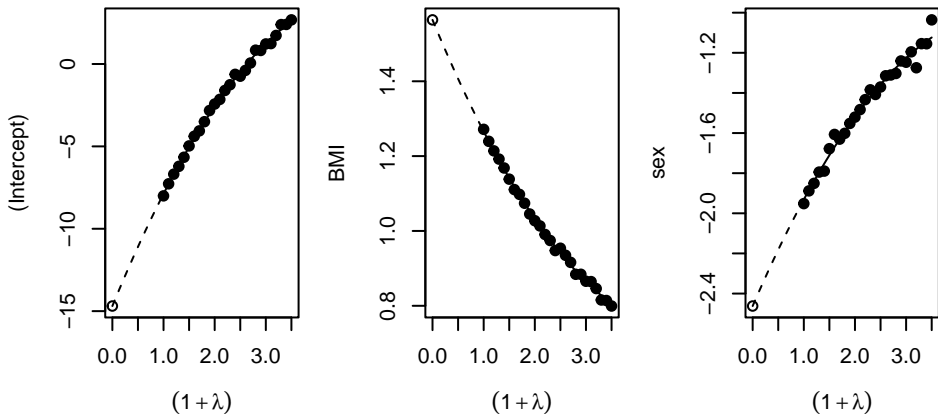
Now run simex procedure

Then run the simex procedure using the `simex()` function from the respective package:

```
library(simex)
r.simex <- simex(r.lm,
  SIMEXvariable = "BMI",
  measurement.error = sqrt(4),
  lambda = seq(0.1, 2.5, 0.1),
  B = 100,
  fitting.method = "quadratic")
summary(r.simex)$coef$asymptotic
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	-14.689940	2.6954519	-5.449899	3.825138e-07
##	BMI	1.564059	0.1159075	13.494022	5.467540e-24
##	sex	-2.462127	0.7906688	-3.113980	2.426632e-03

Graphical results with quadratic extrapolation function:



Note: The sex variable has *not* been mismeasured, nevertheless it is affected by the error in BMI! **Reason:** sex and BMI are correlated.

Practical advice

- ▶ Think about error problems **before** you start collecting your data!
- ▶ Ideally, take **repeated measurements**, maybe of a subset of data points.
- ▶ Figure out if error is a problem and what the bias in your parameters might be. You might need simulations to find out.
- ▶ If needed, model the error. **Seek help from a statistician!**

References

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). Measurement Error in Nonlinear Models: A Modern Perspective (2 ed.). Boca Raton: Chapman & Hall.

Cook, J. R. and L. A. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association 89, 1314–1328.