

# Lecture 12: Measurement Error

## BIO144 Data Analysis in Biology

Owen Petchey, Stephanie Muff & Erik Willems

University of Zurich

27 May, 2024

## A request from your TA's

“Help us help you... (even?) better next year!”



- ▶ Measurement error (ME) in one or more explanatory variable(s) ( $x$ )
- ▶ Effects of ME on model parameters
- ▶ When do you have to worry?
- ▶ An example of a method to correct for ME

# Course material covered today

The lecture material is partially based on:

- ▶ 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 **inaccurate observations** (e.g., self-reported dietary aspects, health history).
- ▶ Rounding error, digit preference.
- ▶ **Classification error** (e.g., exposure or disease classification).
- ▶ ...

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

## Yet another assumption. . .

It is an **implicit assumption** of most statistical tests that explanatory variables are measured or estimated **without error**. This is true for:

- ▶ correlation
- ▶ regression and ANOVA
- ▶ Generalized Linear Models (e.g. Poisson and binomial GLMs)

Violation of this assumption may lead to:

- ▶ biased parameter estimates, standard errors, and thus wrong  $p$ -values
- ▶ incorrect (relative) variable importance, and thus *even more* misguided conclusions

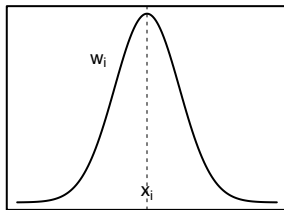
Standard statistics textbooks often do not mention this assumption at all!

# Classical measurement error

A very common type of error:

Let  $x_i$  be the **correct but unobserved** variable and  $w_i$  the observed variable with error  $u_i$ .  
Then the **classical ME model** is:

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

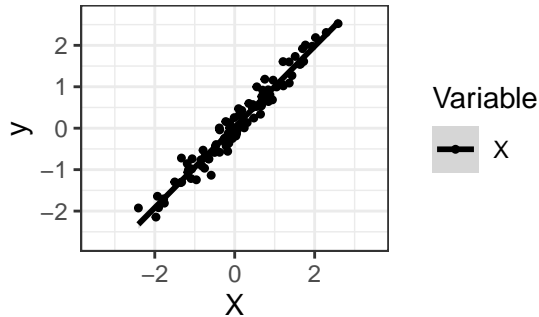


**Examples:** Inaccurate measurements of a concentration, a mass, a length etc. → the observed value  $w_i$  varies around the true value  $x_i$ .

# Illustration of the problem

Find regression parameters  $\beta_0$  and  $\beta_x$  for the model with explanatory variable  $\mathbf{x}$ :

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

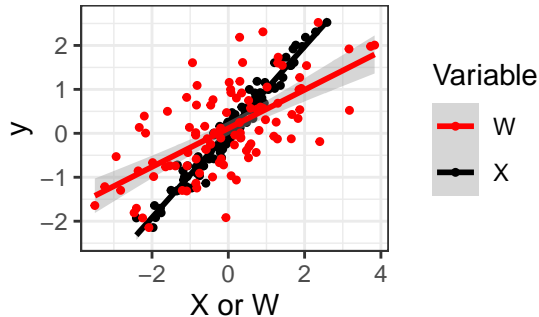




## Illustration of the problem

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

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



# A tool you can play around with...

► Illustration in a browser application

## Classical measurement error in linear, logit and Poisson regression

Select regression model:

Linear

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

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

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  $\sigma_u^2$  (while  $\sigma_\epsilon^2 = 1.$ ) :

0 1 2 3 4 5 6 7 8



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. **Biased** parameter estimates
2. **Loss of power** to detect signals
3. **Masks important features** of the data, making graphical model inspection difficult

## How to correct for ME?

- ▶ Generally, to correct for the error you 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**  $\sigma_u^2$  is required.

**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, and dedicated error modelling methods are needed.

## Attenuation in normal 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**  $\beta_x^*$  underestimates the true slope  $\beta_x$  by **attenuation factor**  $\lambda$ :

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

→ knowing  $\sigma_u^2$  and  $\sigma_x^2$ , the correct slope can be retrieved!

**Example:**  $\sigma_x^2 = 5$ ,  $\sigma_u^2 = 1$ ,  $\rightarrow \lambda = \frac{5}{5+1} = 0.83$

# Error modeling

## Two common approaches:

- ▶ **SIMEX**: SIMulation EXtrapolation, a heuristic and intuitive idea.
- ▶ **Bayesian methods**: Information about the error enters the model as a *prior*.

Both, however, require that the **error model**, and its respective parameters (e.g.,  $\sigma_u^2$ ) are known!

Thus, information about the error mechanism is essential, and potential sources of error must be identified at the planning stages of a study!

# SIMEX: An intuitive idea

Suggested by Cook & Stefanski (1994), SIMEX takes a two-step approach:

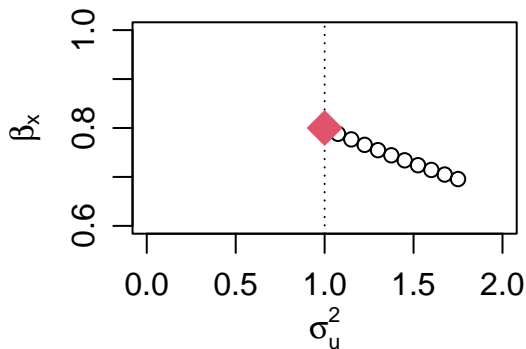
1. **Simulation phase:** The error in the data is progressively aggravated in order to determine how the model parameter of interest is affected.
2. **Extrapolation phase:** The simulated trend is then extrapolated back to a hypothetical error-free value of the model parameter.

# Illustration of the SIMEX idea

Parameter of interest:  $\beta_x$  (e.g. a regression slope).

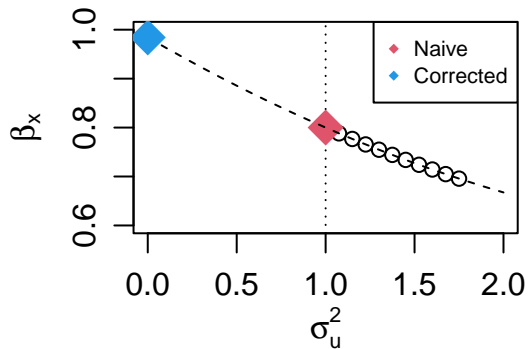
Problem: The respective explanatory variable  $x$  was estimated with error:

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





# Extrapolate to obtain an estimate of the corrected beta



## Example of SIMEX use (part 1)

Let's consider a linear regression model

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

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  $x_i$  was 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

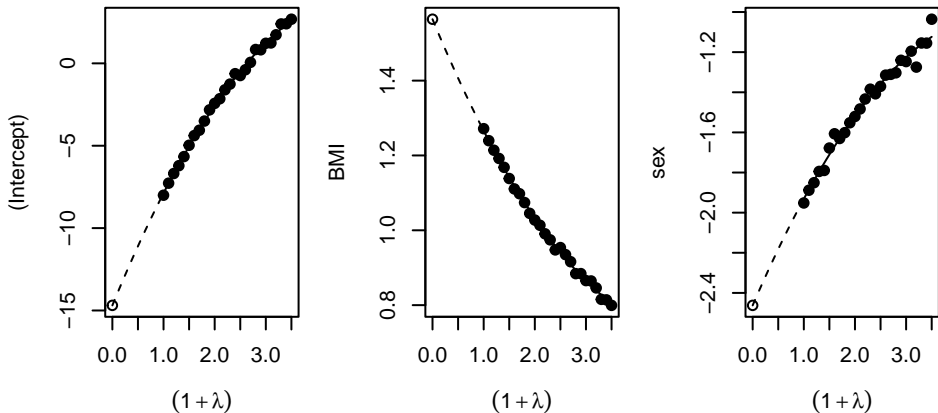
## Now run simex procedure

Then run the SIMEX procedure using the `simex()` function:

```
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 measurement error **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.