

# Kurs Bio144:

# Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Lecture 11: Measurement error in regression models

24./25. May 2018

# Overview

- ME in covariates ( $x$ ) 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 uncertainty / 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”.

# The fundamental assumptions of regression analyses

- **Linear regression including ANOVA:**

$$e_i \sim N(0, \sigma_e^2) .$$

- **Generalized linear model:** Implicit assumptions that can be checked by model diagnostic plots.
- Basically, a fundamental assumption is that the distributional assumptions are fulfilled.

## Another fundamental assumption that is 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 covariate importances, and to **misleading conclusions**.
- Even standard statistics textbooks do often not mention these problems.

→ Measurement error in the covariates ( $\mathbf{x}$ ) violates an assumption of standard regression analyses!!

# Classical measurement error

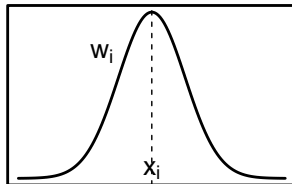
A very common error type:

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

$$w_i = x_i + u_i$$

$$u_i \sim \mathcal{N}(0, \sigma_u^2) ,$$

is the **classical ME model**.



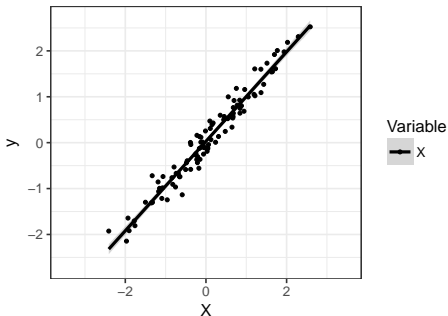
**Examples:** Imprecise measurements of a concentration, a mass, a length etc.

→ The observed value  $w_i$  varies around the true value  $x_i$ .

## Illustration of the problem

Find regression parameters  $\beta_0$  and  $\beta_x$  for the model with covariate  $x$ :

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

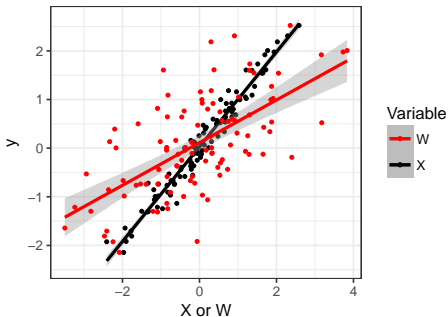




## Illustration of the problem II

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

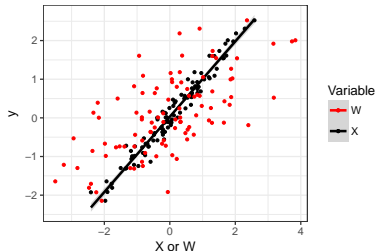
$$w_i = x_i + u_i \quad u_i \sim N(0, \sigma_u^2) \quad \text{with} \quad \sigma_u^2 = \sigma_x^2.$$



# 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.



# Simulations and apps

Illustration with shiny apps for two error types in linear, logistic and Poisson regression:

▶ Classical error

▶ Berkson error

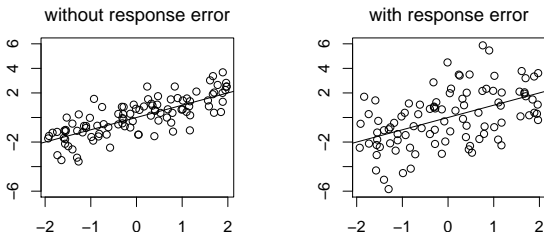
# Error in the outcome of regression models

Example: **Continuous** error in a linear regression outcome.

Note: In the case when the observed response

$$s_i = y_i + v_i \quad v_i \sim N(0, \sigma_v^2) ,$$

the error variance is simply absorbed in the residual variance  $\sigma_\epsilon^2$ .



→ Error in the response seems to be less of a problem. However, this is *not* true for other regression types (logistic, Poisson) or other error structures.

## 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**  $\sigma_u^2$  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  $y_i = \beta_0 + \beta_x x_i + e_i$  with  $w_i = x_i + u_i$ . Assume that  $w_i$  instead of  $x_i$  is used in the regression:

$$y_i = \beta_0^* + \beta_x^* w_i + e_i .$$

The **naive slope parameter**  $\beta_x^*$  is then underestimated with respect to the true slope  $\beta_x$ , with **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:**  $\sigma_x^2 = 5$ ,  $\sigma_u^2 = 1 \rightarrow \lambda = \frac{5}{6} = 0.83$ .

# 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 and 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.

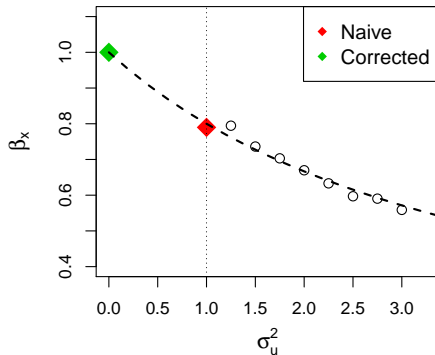


# Illustration of the SIMEX idea

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

Problem: The respective covariate  $x$  was estimated with error:

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



## Example of SIMEX use

Let's consider a linear regression model

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

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 covariate that indicates if the  $i$ -th person was a male ( $z_i = 1$ ) or female ( $z_i = 0$ ).

→ apply the SIMEX procedure!

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

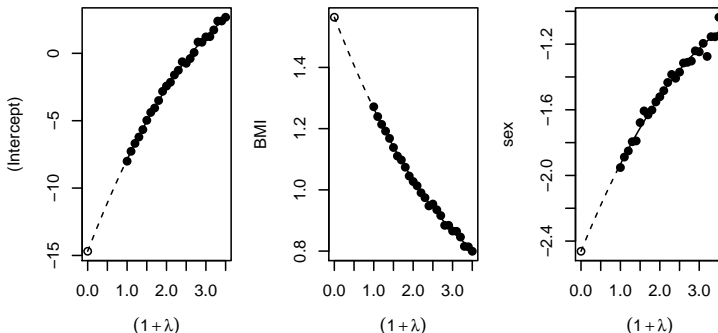
Then run the simex procedure using the `simex()` function from the respective package:

```
> 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:

```
> par(mfrow=c(1,3),mar=c(4,4,2,1))  
> plot(r.simex)
```



**Note:** The sex variable has *not* been mismeasured, but it is was also 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.