Measurement Error and Misclassification in statistical models: Basics and applications bcam Bilbao Part 2

Helmut Küchenhoff
Statistical Consulting Unit
Ludwig-Maximilians-Universität München

Bilbao
28-05-2019

Outline 2. Measurement error: Models and effect

- Models for the error
- Effect of measurement error
  - Response error
  - Linear model
  - Logistic model

Examples

- Munich bronchitis study:
  Average Occupational dust exposure
  Single measurement, expert ratings
- MONICA study:
  Long term Fat intake
  One week diary
- German radon study:
  Residential radon exposure
  Measurements in flats and estimation depending on the home
- Uranium miners study
  Radon exposure
  Job exposure matrix
- Erfurt study
  Individual exposure to a pollutant
  Data from two gauging stations
- Augsburg Study on the effect of PM_{10}

Models for measurement error

- Systematic vs random
- Classical vs Berkson
- Additive vs multiplicative
- Homoscedastic vs heteroscedastic
- Differential vs non differential

Classical additive random measurement error

X_i : True value
X_i^* : Measurement of X

$$
\begin{aligned}
X_i^* &= X_i + U_i \quad (U_i, X_i) \text{ indep.} \\
E(U_i) &= 0 \\
V(U_i) &= \sigma_U^2 \\
U_i &\sim N(0, \sigma_U^2)
\end{aligned}
$$

This model is suitable for

- Instrument m.e.
- One measurement is used for a mean

Accuracy, Validity and Reliability

- Accuracy: General term, describing how closely a measurement reproduces the attribute being measured
- Validity: How well the measurement captures the true attribute or how well it captures the concept which is targeted to be measured
- Reliability describes the differences between multiple measurements of an attribute

Statistical point of view:

Accuracy :	Mean square error
Validity :	Bias $E(U)$
Reliability:	Measurement error variance σ_U^2

Reliability measures

Two measurements

$$X_{ij}^* = X_i + U_{ij} \quad j = 1, 2$$

Assuming independence of the measurement errors U_{ij}

$$\begin{aligned}
Var(X_{ij}^*) &= Var(X_i) + Var(U_{ij}) \\
R &= \frac{Var(X_i)}{Var(X_{ij}^*)} \\
Cor(X_{i1}^*, X_{i2}^*) &= \frac{Cov(X_{i1}^*, X_{i2}^*)}{\sqrt{Var(X_{i1}^*) * Var(X_{i2}^*)}} = R \\
Cor(X_{i1}^*, X_i) &= \frac{Cov(X_{i1}^*, X_i)}{\sqrt{Var(X_{i1}^*) * Var(X_i)}} = \sqrt{R}
\end{aligned}$$

Intraclass Correlation and Reliability

Interpretation:

- R : Informative Part of measurement (Variance decomposition)
- R: Correlation between two independent measurements of the same unit
- R: Square of the correlation between true value and measurement

Estimation of reliability when two measurements per unit are available:

$$Corr(X^*_{i1}, X^*_{i2})$$

$$Var(X^*_{i1} - X^*_{i2}) = Var(U_{i1} - U_{i2}) = 2\sigma_u^2$$

General case

More than 2 measurements per unit, different measurement tools etc.
Use **variance component** model :

$$X_{ij}^* = X_i + U_{ij} (+\tau_j)$$

X_i : random true value

τ_j : random or fixed effect of the jth measurement tool

Then the variances and R can be estimated e.g. by ML or REML.

Problems

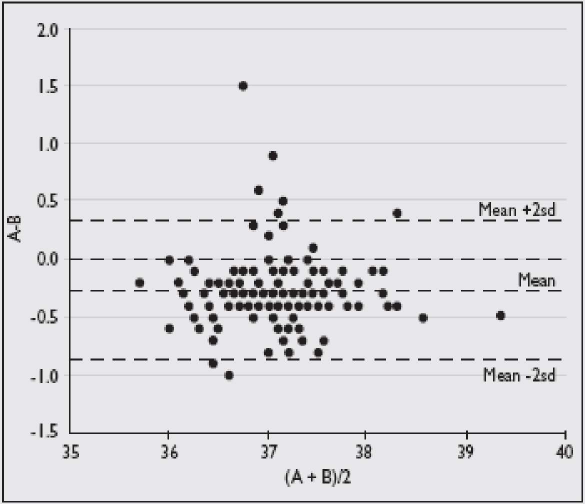
- Reliability dependent on $Var(X_i)$
- Intra Class Correlation invariant on change of the scale for one measurement
- Measurement error variance primary and intuitive characteristic for the simple measurement model
- Measurement error variance can be estimated from two independent (!!) measurements

Bland Altmann Plots

Main Idea: Explore relationship between measurement error and true value
Data: Two types of measurement
Plot difference between two measurements and the mean

Example

Fig. 6 Bland Altman plot.



Approaches for Assessment of agreement

Choudhary and Ng (Biometics 2006): Two measurement methods

- Find a model $D = X_{i1} - X_{i2} = f((X_{i1} + X_{i2})/2$
- Find a simultaneous $p\%$ probability range for the difference
- Use parametric or noparametric (Splines) regression models
- Bootstrap and approximations

Useful for assessment, but correction methods cannot be derived

Additive Berkson-error

$$
\begin{aligned}
X_i &= X_i^* + U_i \quad (U_i, X_i^*) \text{ indep.} \\
E(U_i) &= 0 \\
V(U_i) &= \sigma_U^2 \\
U_i &\sim N(0, \sigma_U^2)
\end{aligned}
$$

The model is suitable for

- Mean exposure of a region X^* instead of individual exposure X.
- Working place measurement
- Dose in a controlled experiment

Classical and Berkson

Note that in the Berkson case

$$\begin{aligned} E(X|X^*) &= X^* \\ Var(X) &= Var(X^*) + Var(U) \\ Var(X) &> Var(X^*) \end{aligned}$$

Note that in the Classical additive case

$$\begin{aligned} E(X^*|X) &= X \\ Var(X^*) &= Var(X) + Var(U) \\ Var(X^*) &> Var(X) \end{aligned}$$

Multiplicative measurement error

$$
\begin{aligned}
X_i^* &= X_i * U_i \quad (U_i, X_i) \text{ indep.} \\
\text{Classical} & \\
X_i &= X_i^* * U_i \quad (U_i, X_i^*) \text{ indep.} \\
\text{Berkson} & \\
E(U_i) &= 1 \\
U_i &\sim \mathit{Lognormal}
\end{aligned}
$$

- Additive on the logarithmic scale
- Used for exposure by chemicals or radiation

Measurement error in response

Simple linear regression

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \varepsilon \\
Y^* &= Y + U \;\; \text{additive measurement error} \\
\longrightarrow & \\
Y^* &= \beta_0 + \beta_1 X + \varepsilon + U
\end{aligned}
$$

New equation error: $\varepsilon + U$
Assumption : U and X independent, U and ϵ independent
$\longrightarrow$ Higher variance of ε
$\longrightarrow$ Inference still correct
Error in equation and measurement error are not discriminable.

Measurement error in covariates

We focus on covariate measurement error in regression models
Main model:

$$E(Y) = f(\beta, X, Z)$$

We are interested in Inference on β_1
Z is a further covariate measured without error
Error model:

$$X \longleftrightarrow X^*$$

Observed model:

$$E(Y) = f^*(X^*, Z, \beta^*)$$

Naive estimation:
Observed model = main model
but in most cases : $f^* \neq f,\ \beta^* \neq \beta$

Differential and non differential measurement error

Assumption of non differential measurement error:

$$[Y|X, X^*] = [Y|X]$$

For Y there is no further information in U or X^* when X is known. Then the error and the main model can be split.

$$[Y, X^*, X] = [Y|X][X^*|X][X]$$

From the substantive point of view:

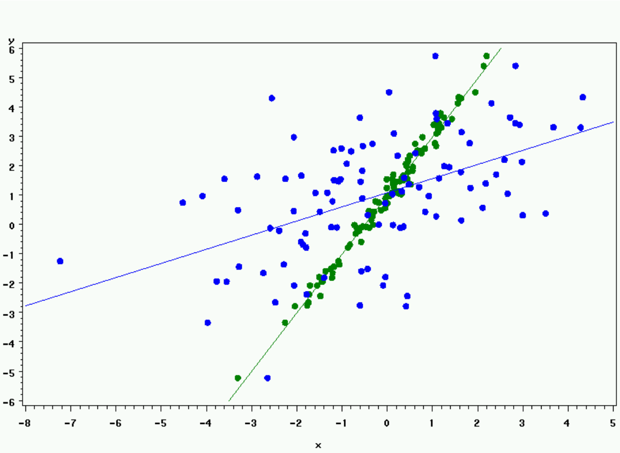
- Measurement process and Y are independent
- Blood pressure on a special day is irrelevant for CHD if long term average is known
- Mean exposure irrelevant if individual exposure is known
- **But** people with CHD can have a different view on their nutrition behavior

Simple linear regression

We assume a classical non differential additive normal measurement error

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \epsilon \\
X^* &= X + U, \;\; (U, X, \epsilon) \;\text{ indep.} \\
U &\sim N(0, \sigma_u^2) \\
\epsilon &\sim N(0, \sigma_\epsilon^2)
\end{aligned}
$$

Effect of additive measurement error on linear regression



The observed model in linear regression

$E(Y|X^*) = \beta_0 + \beta_1 E(X|X^*)$
Assuming $X \sim N(\mu_x, \sigma_x^2)$, the observed model is:

$$
\begin{aligned}
E(Y|X^*) &= \beta_0^* + \beta_1^* X^* \\
\beta_1^* &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \beta_1 \\
\beta_0^* &= \beta_0 + \left(1 - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right) \beta_1 \mu_x
\end{aligned}
$$

$$
Y - \beta_0^* - \beta_1^* X^* \sim N\left(0, \sigma_\epsilon^2 + \frac{\beta_1^2 \sigma_u^2 \sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right)
$$

- The observed model is still a linear regression !
- Attenuation of β_1 by the factor $\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$
  Reliability ratio"
- Loss of precision (higher error term)

Identification

$$
\begin{aligned}
(\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_\epsilon^2) &\longrightarrow [Y, X^*] \\
&\longrightarrow \mu_y, \mu_{x^*}, \sigma_y^2, \sigma_{x^*}^2, \sigma_{x^*y}
\end{aligned}
$$

$(\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_\epsilon^2)$ and $(\beta_0^*, \beta_1^*, \mu_x, \sigma_x^2 + \sigma_u^2, 0, \sigma_\epsilon)$ yield the identical distributions of (Y, X^*). $\Longrightarrow$ The model parameters are not identifiable

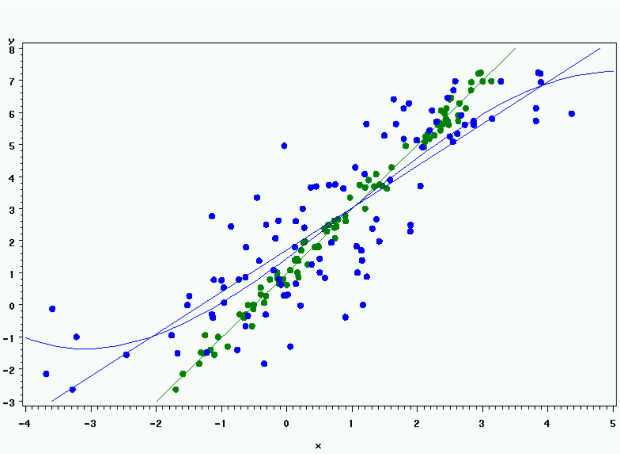
We need extra information, e.g

- σ_u is known or can be estimated
- σ_u/σ_ϵ is known (orthogonal regression)

The model with another distribution for X is identifiable by higher moments.

The observed model in linear regression (2)

Note that the observed model is dependent on the distribution of X. It is not a linear regression, if X is not normal. Ex: X is a mixture of Normals



Naive LS- estimation

For the slope :

$$
\begin{aligned}
\hat{\beta}_{1n} &= \frac{S_{yx^*}}{S^2_{x^*}} \\
\text{plim}(\hat{\beta}_{1n}) &= \frac{\sigma_{yx^*}}{\sigma^2_{x^*}} \\
&= \frac{\sigma_{yx}}{\sigma^2_x + \sigma^2_u} \\
&= \beta_1 * \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_u}
\end{aligned}
$$

For the intercept:

$$
\begin{aligned}
\hat{\beta}_{0n} &= \bar{Y} - \beta_{1n}\bar{X}^* \\
\text{plim}(\hat{\beta}_{0n}) &= \mu_y + \beta_1 * \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_u} * \mu_{x^*} \\
&= \beta_0 + \beta_1 * \left(1 - \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_u}\right) * \mu_x
\end{aligned}
$$

Naive LS- estimation (2)

For the residual variance:

$$
\begin{aligned}
MSE &= S_{Y-\hat{\beta}_{0n}-\hat{\beta}_{1n}X^*} \\
plim(MSE) &= \sigma_\epsilon^2 + \frac{\beta_1^2\sigma_u^2\sigma_x^2}{\sigma_x^2+\sigma_u^2}
\end{aligned}
$$

Multiple linear regression

The generalization form the simple model is straightforward:

$$\begin{aligned} Y &= \beta_0 + X'\beta_x + Z'\beta_z \\ X^* &= X + U \\ U &\sim N(0, \Sigma_u) \end{aligned}$$

Z is observed without error
If we use X^* instead of X then

$$\begin{pmatrix} \hat{\beta}_{x^*n} \\ \hat{\beta}_{zn} \end{pmatrix} \rightarrow \begin{pmatrix} \Sigma_x + \Sigma_u & \Sigma_{xz} \\ \Sigma_{xz} & \Sigma_z \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_x & \Sigma_{xz} \\ \Sigma_{xz} & \Sigma_z \end{pmatrix} \begin{pmatrix} \beta_{x^*} \\ \beta_z \end{pmatrix}$$

Multiple linear regression (2)

If Z and X are correlated then

- The attenuation factor is now
  $\frac{\sigma^2_{x|z}}{\sigma^2_{x|z}+\sigma^2_u}$
  $\sigma^2_{x|z}$ is residual variance from regressing X on Z
- $\hat{\beta}_{zn}$ is also biased
  $\hat{\beta}_{zn} \longrightarrow \beta_z + (1 - \frac{\sigma^2_{x|z}}{\sigma^2_{x|z}+\sigma^2_u})\beta_x\gamma_z$
  γ_z is regression coefficient when regressing X on Z

Correction for attenuation

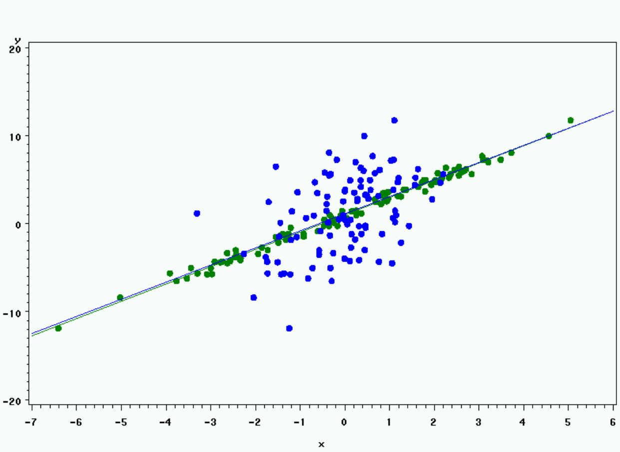
We have a first method: Solve the bias equation:

$$
\begin{aligned}
\hat{\beta}_1 &= \hat{\beta}_{1n} \frac{\sigma_x^2 + \sigma_u^2}{\sigma_x^2} \\
\hat{\beta}_1 &= \hat{\beta}_{1n} \frac{S_{x^*}^2}{S_{x^*}^2 - \sigma_u^2} \\
\hat{\beta}_0 &= \hat{\beta}_{0n} - \hat{\beta}_1 \left(\frac{S_{x^*}^2 - \sigma_u^2}{S_{x^*}^2} \right) \bar{X}^*
\end{aligned}
$$

Correction by reliability ratio.
$V(\hat{\beta}_1) > V(\beta_{1n})$ Bias Variance trade off

Berkson-Error in simple linear regression

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \epsilon \\ X &= X^* + U, \quad U, (X^*, Y) \text{ indep.}, \; E(U) = 0 \end{aligned}$$



Observed Model

$$
\begin{aligned}
E(Y|X^*) &= \beta_0 + \beta_1 X^* \\
V(Y|X^*) &= \sigma_\epsilon^2 + \beta_1^2 * \sigma_u^2
\end{aligned}
$$

- Regression model with identical β
- Measurement error ignorable
- Loss of precision

Binary Regression

Logistic with additive non differential measurement error

$$
\begin{aligned}
P(Y=1|X) &= G(\beta_0+\beta_1 X) \\
G(t) &= (1+\exp(-t))^{-1} \\
X^* &= X+U
\end{aligned}
$$

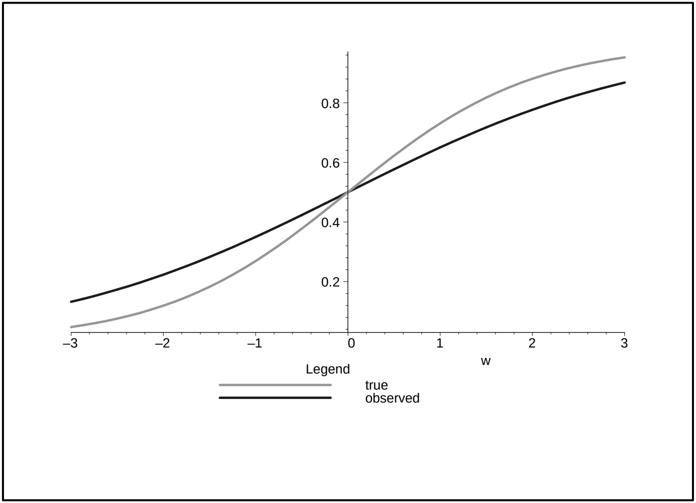
Observed model:

$$
\begin{aligned}
&P(Y=1|X^*) = \int P(Y=1|X,X^*) f_{X|X^*} dx \\
&= \int P(Y=1|X) f_{X|X^*} dx
\end{aligned}
$$

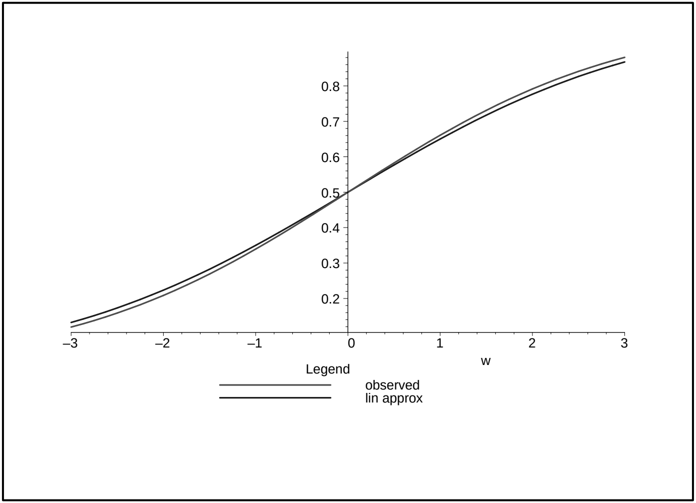
If we have additive measurement error and X and U are normal then $X|X^*$ is also normal

$$
P(Y=1|X^*) = \int G(\beta_0+\beta_1 X) f_{X|X^*} dx
$$

Simple Logistic



Linear Approximation



Probit Model

This integral is not easy to handle, but for the Probit model we can evaluate it:

$$
\begin{aligned}
P(Y=1|X^*) &= \Phi\left((\beta_0^* + \beta_1^* X^*)/\sqrt{1+\beta_1^2 \cdot v}\right) \\
\beta_1^* &= \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\beta_1 \\
\beta_0^* &= \beta_0 + \left(1 - \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\right)\beta_1\mu_x \\
v &= Var(X|X^*)
\end{aligned}
$$

This gives an exact correction for the Probit model

Probit approximation for logistic regression

$$G(t) = (1+\exp(-t))^{-1} \;\approx\; \Phi(t/h_*) \text{ mit } h_* \approx 1.70$$

$$\begin{aligned} E(Y|X^*) &= \int G(\beta_0 + \beta_1 X) f_{X|X^*} dx = \\ &= \int \Phi((\beta_0 + \beta_1 X) h_*^{-1}) f_{X|X^*} dx = \\ &= G\left((\beta_0^* + \beta_1^* X^*) / \sqrt{1 + \beta_1^2 v h_*^{-2}} \right) \end{aligned}$$

Effect of measurement error in logistic regression

- Similar to the linear Model
- Further attenuation by $\sqrt{1 + v\beta_1^2 h_*^{-2}}$