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

Helmut Küchenhoff
Statistical Consulting Unit
Ludwig-Maximilians-Universität München

Bilbao
27-05-2019

Schedule

Monday	1.	Introcduction and Misclassification: Basic Models
Tuesday	2.	Measurement error: Effect and Models
Wednesday	3.	Methods for Estimation in the presence of Measurement error
Thursday	4.	Simulation and Extrapolation (SIMEX) for Misclassification and measurement error: Concept and Examples
Friday	5.	Case studies: Uncertainty of diagnosis, Exposure assessment

Material

- Carroll R. J. , D. Ruppert , L. Stefanski and CRainiceanu, C : Measurement Error in Nonlinear Models. A Modern perspective. Chapman & Hall London 2006.
- Gustafson, P. : Measurement Error and Misclassification in Statistics and Epidemiology. Impacts and Bayesian Correction & CRC Press, Boca Raton 2004.

Outline

- Measurement
- Misclassification
- Model
- Effect
- Correction

Measurement

Museum of modern Art in Barcelona

Measurement

Measurement is the contact of reason with nature
(Henry Margenau)
Nearly all the grandest discoveries of science have been but the rewards of accurate measurement
(Lord Kelvin)
Measurement is the basis for producing data
Literature: David Hand: Measurement. Theory and practice . The world through quantification. (Arnold,2004)

Basics

Peter	$\longrightarrow$	1.84
Stefan	$\longrightarrow$	1.91
Laura	$\longrightarrow$	1.72

Measurement is a assignment of a number to a characteristic of an object. This measurement is to be compared with other objects.

Measurement: A structure preserving function (homomorphism)

Peter is smaller than $\Leftrightarrow 1.84 < 1.91$

Levels of scaling

This can be defined by the structure in the objects

only relation equal - non equal:	nominal
smaller :	ordinal scale
differences :	metric scale
differences and ratios :	Ratio scale

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
Reliability:	correlation or difference, agreement between two raters

Types of measurement

- Representational measurement
  Measurements **relate to existing attributes** of the objects
  Examples: Length, weight, blood parameter
- Pragmatic measurement
  An attribute is **defined by its measuring procedure**, no 'real' existence beyond that
  Examples: Pain score, intelligence

Sources of measurement error

- Induced by an instrument (laboratory value, blood pressure)
- Induced by medical doctors or patients
- Measurement error induced by definition, e.g. long term mean of daily fat intake"
- Surrogate -Variables e.g. mean of exposure in a region where the study participant lives instead of individual exposure

Misclassification: Examples

- Wrong diagnosis
  not diseased instead of *diseased*
- Wrong answer in a questionnaire *Voted for the greens*
  No drugs
  Do not smoke
- Technical problems , e. g. classification of genes
- Classification by machine learning tool (e.g. classification of images)
- Problem of definition, e .g. Caries
- Randomized response
- Anonymisation of data

General remarks

Before we start thinking of measurement error and misclassification, we should answers to the folowing questions:

- Is there a true value and how is it defined ?
- What is the aim of our study concerning the variable having measurement error ? (prediction or interpretation, outcome or predictor ...)

Notation

We have to distinguish between true (correctly measured, gold standard) variable
and its (possible incorrect) measurement
X ,W , Z - Notation (Carroll et al.)
X: correctly (unobservable) Variable
W: possibly incorrect measurement of X
Z: Further correctly measured variables

ξ - X- Notation (Schneeweiß , Fuller)
ξ : correctly (unobservable) Variable
X : possibly incorrect measurement of X

* - Notation (HK)
X, Z, Y : correctly (unobservable) Variable
X^*, Z^*, Y^*: Corresponding possibly incorrect measurements

One sample binary

Model for misclassification

Y : true binary variable, gold standard

Y^* : observed value of Y, surrogate

$$
\begin{aligned}
P(Y^* = 1 | Y = 1) &= \pi_{11} \quad \text{(Sensitivity)} \\
P(Y^* = 0 | Y = 0) &= \pi_{00} \quad \text{(Specificity)} \\
P(Y^* = 0 | Y = 1) &= 1 - \pi_{11} = \pi_{01} \\
P(Y^* = 1 | Y = 0) &= 1 - \pi_{00} = \pi_{10}
\end{aligned}
$$

$\rightarrow$ (mis-) classification matrix (diffusion matrix)

$$
\Pi = \begin{pmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{pmatrix}
$$

Effect of misclassification

Naive analysis: Simply ignore misclassification
We want to estimate $P(Y = 1)$
We use $\frac{1}{n}\sum_{i=1}^{n} Y_i^*$

$$\begin{aligned} P(Y^* = 1) &= \pi_{11}P(Y = 1) + \pi_{10}P(Y = 0) \\ P(Y^* = 1) - P(Y = 1) &= \pi_{10}P(Y = 0) - \pi_{01}P(Y = 1) \end{aligned}$$

$\rightarrow$ Examples:

No bias if $P(Y = 1) = \frac{1}{2}$ and $\pi_{00} = \pi_{11}$
Neg. bias if $P(Y = 1) = 0.9$ and $\pi_{00} = \pi_{11} = 0.9$
$\rightarrow$ Bias $= -0.1 \cdot 0.9 + 0.1 \cdot 0.1 = -0.08$
Pos. bias if $P(Y = 1) = 0.8$ and $\pi_{11} = 0.99, \pi_{00} = 0.9$
$\rightarrow$ Bias $= -0.01 \cdot 0.8 + 0.1 \cdot 0.9 = 0.01$

Effect of Misclassification

Everything can happen
dependent on π_{11}, π_{00} and $P(Y = 1)$.

However, if $\pi_{00} = \pi_{11}$ (in most times unrealistic) then

$$\begin{aligned} \text{Bias} &= \pi_{00}(1 - 2P(Y = 1)) \\ P(Y = 1) > 0.5 &\implies \text{bias} < 0 \\ P(Y = 1) < 0.5 &\implies \text{bias} > 0 \end{aligned}$$

Attenuation towards 0.5

Correction

Idea: Solve the bias equation

Note that X^* is still binomial and $\mathrm{P}(X^* = 1)$ can be consistently estimated from the observed data.

$$\begin{aligned} P(Y^* = 1) &= \pi_{11} P(Y = 1) + \pi_{10}(1 - P(Y = 1)) \\ \Rightarrow P(Y = 1) &= (P(Y^* = 1) - \pi_{10})/(\pi_{11} + \pi_{00} - 1) \end{aligned}$$

Assumptions

- π_{11} and π_{00} known
- $\pi_{11} + \pi_{00} > 1$

Variance factor $(\pi_{11} + \pi_{00} - 1)^{-2}$

Multinomial case

Y is multinomial with categories $1, \ldots, k$.
Y^* is observed
The error model is given by the classification Matrix

$$\Pi = \{\pi_{ij}\}$$

with $\pi_{ij} = P(Y^* = i | Y = j)$. Then we get for the probability vectors

$$\begin{aligned} P_Y &= (p_{y1}, \ldots, p_{yk})' \\ P_{Y^*} &= (p^*_{y1}, \ldots, p^*_{yk})' \\ P_{Y^*} &= \Pi * P_Y \end{aligned}$$

The matrix method

The correction method is given by

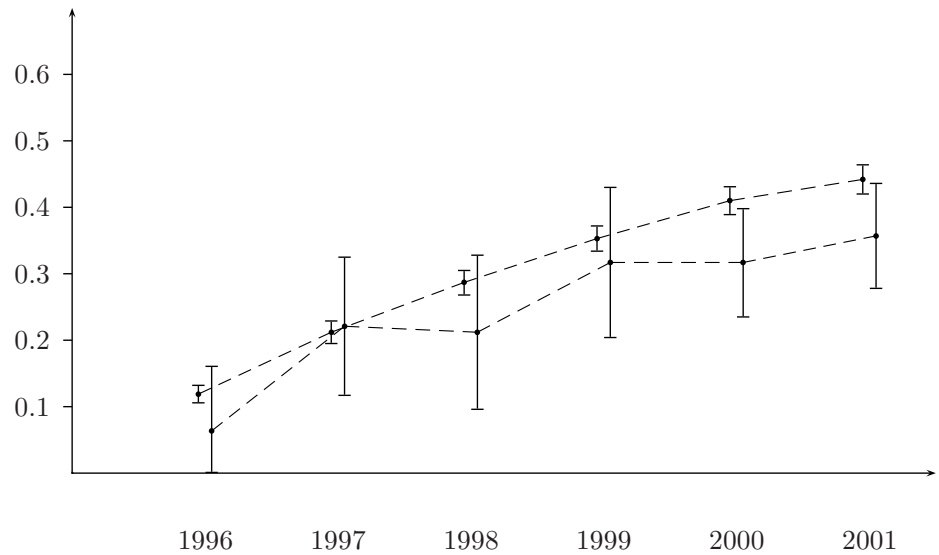
$$\hat{P}_Y := \Pi^{-1} * \hat{P}_{Y^*} \tag{1}$$

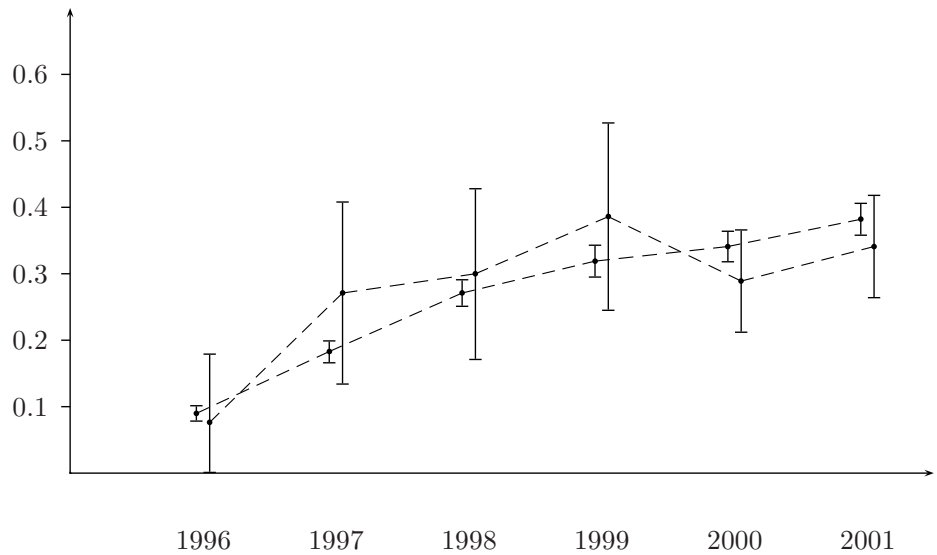
Properties

- Classification matrix has to be known or estimated
- Gives sometimes probabilities > 1 or < 0
- Variance calculation straight forward
- Use the delta method in the case of estimated Π, Greenland (1988)

Prevalence estimation from the Signal- Tandmobiel study

- Oral health study involving 4468 children in Flanders
- Y=1 if the tooth is decayed, missing due to caries or filled
- 16 examiners with high MC on Y
- Validation study also used for two regions
- Validation data from 3 validation studies
- Simple correction in two regions: East and West


0.6
0.5
0.4
0.3
0.2
0.1
1996
1997
1998
1999
2000
2001


0.6
0.5
0.4
0.3
0.2
0.1
1996
1997
1998
1999
2000
2001

Results

Estimated prevalence using data from the validation study

- Corrections
- Huge confidence limits
- MC Matrix possibly overestimated

Information about misclassification

There are three basic strategies:

- Assumption, external validation data
- Internal validation data
- Replication data

Assumption, external validation data

Examples

- Certain type of diagnosis
- Technical applications
- Results from other studies (be very careful!)
- Interpretation as sensitivity analysis

Note that ignoring misclassification assumes $\pi_{ij} = 0$!

Internal validation data

Examples

- Caries study: examiners were compared to a gold standard
- Controlling a part of a questionnaire by a doctor
- Ex post check of a diagnosis

Calibration Model

X : true binary variable, gold standard examiner
X^* : observed value of X, surrogate

$$P(X = 1|X^* = 1) \qquad \text{(positive predicted value)}$$
$$P(X = 0|X^* = 0) \qquad \text{(negative. Predicted value)}$$

can be calculated from MC-Matrix and marginal Distribution of X (i.e. from $P(X = 1)$)

Replication

If no gold standard is available measurements are replicated.

- Requirement: measurements have to be conditional independent on the true value
- Identifiability conditions for multinomial case

Two independent measurements

We observe X_{i1}^*, X_{i2}^* ,i.e a $2 * 2$ –table:

	$X_1^* = 0$	$X_1^* = 1$
$X_2^* = 0$	n_{00}	n_{10}
$X_2^* = 1$	n_{01}	n_{11}

Assuming independence and constant MC we get :

$$
\begin{aligned}
P(X_1^* = 0, X_2^* = 0) &= P(X = 0) * \pi_{00}^2 + P(X = 1) * \pi_{01}^2) \\
P(X_1^* = 1, X_2^* = 1) &= P(X = 0)\pi_{10}^2 + P(X = 1)\pi_{11}^2 \\
P(X_1^* = 0, X_2^* = 1) &= P(X = 0) * \pi_{00}\pi_{10} + P(X = 1)\pi_{01}\pi_{11} \\
P(X_1^* = 1, X_2^* = 0) &= P(X_1^* = 0, X_2^* = 1)
\end{aligned}
$$

Two independent equations, but three unknown parameters!
$\Rightarrow$ We cannot estimate the MC-Matrix and P(X=1)!!

Identified problems

Literature about diagnostic tests
Three independent Measurements: Three independent equations three unknowns. Explicit solution available
Further assumptions : Error in Haplotype reconstruction same MC matrix for each gene

Kappa Statistics

Basic idea : Evaluate agreement and adjust for agreement by chance
Measuring agreement:

$$\frac{n_{00} + n_{11}}{n}$$

	$X_1^* = 0$	$X_1^* = 1$
$X_2^* = 0$	10	2
$X_2^* = 1$	2	0

	$X_1^* = 0$	$X_1^* = 1$
$X_2^* = 0$	5	2
$X_2^* = 1$	2	5

Same proportion of agreement, but different situation !!

Definition of Kappa

$$
\begin{aligned}
P_o &= \frac{n_{00} + n_{11}}{n} \\
P_e &= \frac{n_{0.} n_{.0}}{n^2} + \frac{n_{1.} n_{.1}}{n^2} \\
\kappa &= (P_o - P_e)/(1 - P_e)
\end{aligned}
$$

	$X_1^* = 0$	$X_1^* = 1$
$X_2^* = 0$	10	2
$X_2^* = 1$	2	0

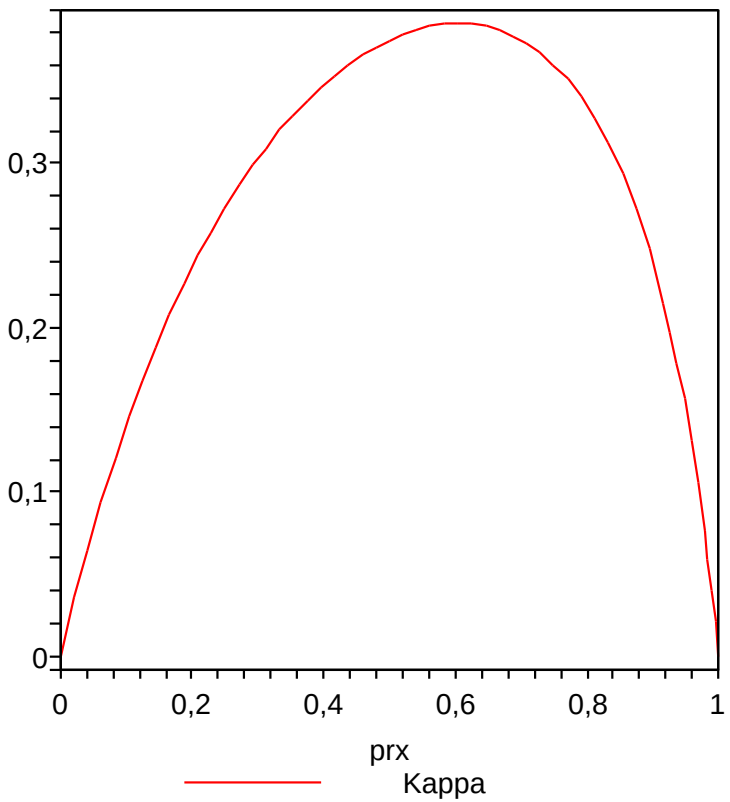
$\kappa < 0$

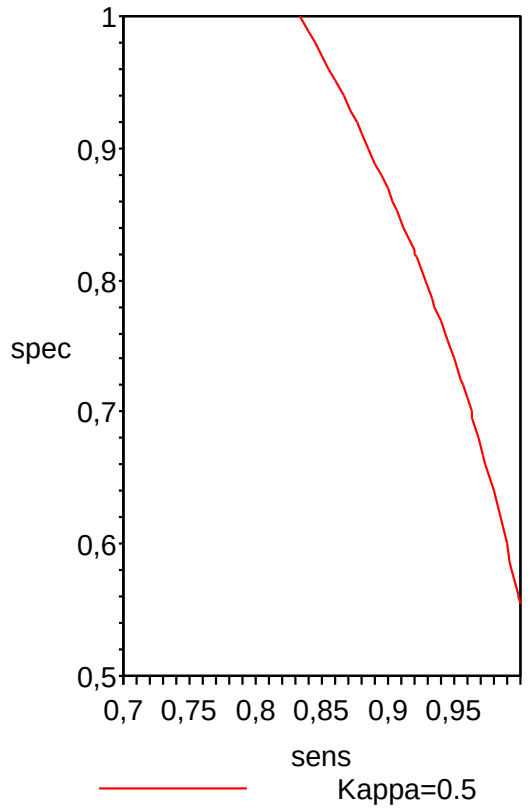
	$X_1^* = 0$	$X_1^* = 1$
$X_2^* = 0$	5	2
$X_2^* = 1$	2	5

$\kappa = 0.428$

Kappa and MC-Matrix

Kappa depends on the MC-Matrix and marginal distribution P(X=1)
Fixed MC-Matrix $\pi_{00} = 0.9, \pi_{11} = 0.7$ (I)
Sensitivity and specificity which result in $\kappa = 0.5$ for $P(X = 1) = 0.2$(r)





Bivariate analysis

Binary exposure: X
Disease status: Y
Measurement of disease: Y^*

Model for misclassification:

$$\begin{aligned}
\pi_{110} &= P(Y^* = 1 | Y = 1, X = 0) \\
\pi_{111} &= P(Y^* = 1 | Y = 1, X = 1) \\
\pi_{100} &= P(Y^* = 1 | Y = 0, X = 0) \\
\pi_{101} &= P(Y^* = 1 | Y = 0, X = 1)
\end{aligned}$$

Non differential misclassification if

$$\pi_{110} = \pi_{111} \text{ and } \pi_{100} = \pi_{101},$$

i.e. misclassification independent of exposure

Effect and correction

Use the results of one sample case:

$$P(Y^* = 1|X = 1) = \pi_{111}P(Y = 1|X = 1) + \pi_{101}P(Y = 0|X = 1)$$
$$P(Y^* = 1|X = 0) = \pi_{110}P(Y = 1|X = 0) + \pi_{100}P(Y = 0|X = 0)$$

If the misclassification error is non differential then:

$$P(Y^* = 1|X = 1) - P(Y^* = 1|X = 0) =$$
$$[P(Y = 1|X = 1) - P(Y = 1|X = 0)]\,(\pi_{11} + \pi_{00} - 1)$$

- Attenuation to 0
- Also for OR
- Correction by matrix method

Misclassification in exposure

We observe X^* instead of X
Model for misclassification:

$$\begin{aligned}
\pi_{110} &= P(X^* = 1|X = 1, Y = 0) \\
\pi_{111} &= P(X^* = 1|X = 1, Y = 1) \\
\pi_{100} &= P(X^* = 1|X = 0, Y = 0) \\
\pi_{101} &= P(X^* = 1|X = 0, Y = 1)
\end{aligned}$$

Non differential misclassification if
$\pi_{110} = \pi_{111}$ and $\pi_{100} = \pi_{101}$,
i.e. misclassification independent of disease
This is fulfilled in most cohort studies, but could be violated in case control studies

Example for non differential misclassification error

high fat	No	Yes	No	Yes	No	Yes
cases	450	250	360	340	410	290
controls	900	100	720	280	740	260
Odds ratio	5.0		2.4		2.0	
	Correct Classification		20%of No say Yes		20% of No s. Yes 20% of Yes s. No	

Attenuation to OR = 1 Note: Everything can happen in case of differential misclassification

Likelihood

We assume **non differential** misclassification error

$$P(Y=1, X^*=x^*) = \sum_x P(Y=1, X^*=x^*, X=x)$$

$$= \sum_x P(Y=1|X^*=x^*, X=x) * P(X^*=x^*, X=x)$$

$$= \sum_x P(Y=1|X=x) * P(X^*=x^*|X=x) * P(X=x)$$

We have three components of the likelihood:
Main model: $P(Y=1|X=x)$
Measurement model: $P(X^*=x^*|X=x)$
Exposure model: $P(X=x)$

Observed probabilities

$$P(Y=1|X^*=x^*) \quad = \quad \frac{P(Y=1, X^*=x^*|)}{P(X^*=x^*)}$$

$$\frac{P(Y=1|X^*=1)-P(Y=1|X^*=0)}{P(Y=1|X=1)-P(Y=1|X=0)} = \frac{(\pi_{11}+\pi_{00}-1)P(X=1)P(X=0)}{P(X^*=1)P(X^*=0)} < 1$$

see Gustafson (2004), p.35 ff Bias to 0 if $(\pi_{11}+\pi_{00}-1) > 0$

Misclassification in a confounder

X^* : Misclassified confounder
Z : Exposure
Y : Response
Even in the case of non differential measurement error with respect to Y and Z:

- Bias in both direction possible
- Residual confounding
- e.g. Savitz and Baron (1989)

Correction methods

- Matrix method: The two by two table can be seen as one multinomial variable
- Variance estimation see Greenland(1988)
- MLE for unrestricted sampling
- Alternatives by Tennebein (1972)

Misclassification in regression

General Regression Model

$$E(Y|X_1, \ldots, X_k) = h(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)$$ h: Link-function

Misclassification possibly on

- binary covariates: Observe X^* instead of X
- binary response : Observe Y^* instead of Y

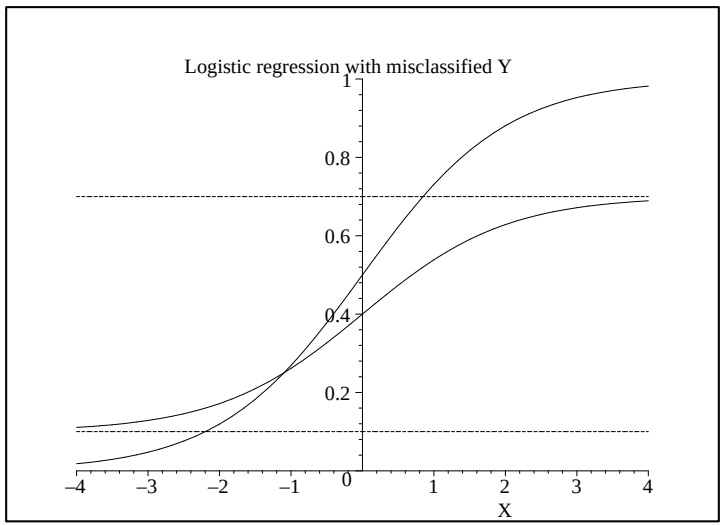
Handling misclassification in Y in binary regression

- Hausmann et al. (Journal of Econometrics, 1998)
- Neuhaus (Biometrika, 1999)

We observe Y^* instead of Y with misclassification matrix Π

$$\begin{aligned} P(Y^* = 1|X) &= \pi_{11} G(x'\beta) + (1 - \pi_{00})(1 - G(x'\beta)) = H(x'\beta) \\ H(t) &= \pi_{11} G(t) + (1 - \pi_{00})(1 - G(t)) \end{aligned}$$

Observed regression function



Logistic regression with misclassified Y

Misclassification in regressors

One binary regressor, normal Outcome:
$Y = \beta_0 + \beta_1 I_1, \beta_0 = \mu_0, \beta_1 = \mu 1 - \mu_0$
Naive analysis:
$E(Y|X^* = 0)) = P(X = 0|X^* = 0) * \mu_0 + P(X = 1|X^* = 0) * \mu_1$
$E(Y|X^* = 0) = P(X = 0|X^* = 1) * \mu_0 + P(X = 1|X^* = 1) * \mu_1$
These equations can be solved for μ_1 and μ_2, when MC Matrix and P(X=0) is known
Matrix Method

Likelihood

$$
\begin{aligned}
L(Y, X^*) &= \sum_x L(Y, X^*, X = x) \\
&= \sum_x L(Y|X^* = x^*, X = x) * P(X^* = x^*, X = x) \\
&= \sum_x L(Y|X = x) * P(X^* = x^*|X = x) * P(X = x)
\end{aligned}
$$

Likelihood for many regression models numerically easy to handle
Components of the misclassification model and its components can be added.

Effects of misclassification

- Biased and inconsistent estimates for parameters
- In most cases attenuation to 0
- In complex settings bias in any direction possible
- Effect dependent on the misclassification matrix
- Similar to effect of measurement error in continuous variables in regression

Hypothesis testing

Attenuation $\rightarrow$

- Naive tests (e. g. for no true effect in a $2x2$ table) have still correct significance level
- Power reduction
- Sample size calculation has to be corrected

Outlook

- Use of validation data
- Latent class analysis