

Scuola di Dottorato di Ricerca in Scienze Statistiche - XXII ciclo

Ciclo di seminari su

**MISCLASSIFICATION  
AND MEASUREMENT ERROR  
IN REGRESSION MODELS**

tenuti da

**Prof. Helmut Küchenhoff**  
*Ludwig-Maximilians-Universität München*

Calendario

<b>Martedì</b>	<b>2 ottobre 2007</b>	10.00 – 12.00	Aula Uggè
		14.30 – 16.30	Aula Cucconi
<b>Mercoledì</b>	<b>3 ottobre 2007</b>	10.00 – 12.00	Aula Cucconi
		14.30 – 16.30	Aula Cucconi

<http://www.stat.unipd.it/seminaridottorato>

*Prof.ssa Alessandra Salvan  
Direttore della Scuola*

# **MISCLASSIFICATION AND MEASUREMENT ERROR IN REGRESSION MODELS**

**Short course at**

**DIPARTIMENTO DI SCIENZE STATISTICHE**

**Universita' degli Studi di Padova**

Helmut Küchenhoff

Statistical Consulting Unit

Ludwig-Maximilians-Universität München

Padova

2./3.10.2007

## Schedule

Tuesday	10-12	1.	Misclassification: Basic Models
	14.30-16.30	2.	Measurement error: Effect and Models
Wednesday	10-12	3.	Methods for Estimation in the presence of Measurement error
	14.30-16.30	4.	Simulation and Extrapolation (SIMEX) for Misclassification and measurement error: Concept and Examples

# Material

- Carroll R. J. , D. Ruppert , L. Stefanski and CRainiceanu, C : Measurement Error in Nonlinear Models. A Modern perspective. Chapman & Hall London 2006.
- Kuha, J., C. Skinner and J. Palmgren. Misclassification error. In: Encyclopedia of Biostatistics, ed. by Armitage, P.and Colton, T., 2615-2621. Wiley, Chichester.

# Outline 1. Misclassification

- Examples
- One sample
  - Model
  - Effect
  - Correction
- Bivariate analysis
  - Error in disease status
  - Error in exposure status
  - Methods for correction

- Regression
  - Misclassification in binary response
  - Misclassification in regressors

## Outline 2. Measurement error: Models and effect

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

## Outline 3. Methods for correction

- Direct correction, method of moments
- Regression calibration
- Likelihood
- Quasi likelihood
- Bayes
- Corrected score

## Outline 4. SIMulation and EXtrapolation

- SIMEX for measurement error
- Misclassification SIMEX

# 1. Misclassification: Examples

- Wrong diagnosis  
„Not diseased“ instead of „diseased“
- Wrong answer in a questionnaire  
„No drugs“  
„Do not smoke“
- Technical problems , e. g. classification of genes
- Problem of definition, e .g. Caries
- Randomized response
- Anonymisation of data

## Notation

We have to distinguish between true (correctly measured ) 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

$\xi$  -  $X$ - Notation (Schneeweiß , Fuller)

$\xi$  : correctly (unobservable) Variable

$X$  : possibly incorrect measurement of  $X$

\* - Notation (HK)

$X, Z$  : correctly (unobservable) Variable

$X^*, Z^*$ : Corresponding possibly incorrect measurements

## 1.2 One sample binary

### Model for misclassification

$X$  : true binary variable, gold standard

$X^*$  : observed value of  $X$ , surrogate

$$P(X^* = 1|X = 1) = \pi_{11} \text{ (Sensitivity)}$$

$$P(X^* = 0|X = 0) = \pi_{00} \text{ (Specificity)}$$

$$P(X^* = 0|X = 1) = 1 - \pi_{11} = \pi_{01}$$

$$P(X^* = 1|X = 0) = 1 - \pi_{00} = \pi_{10}$$

→ misclassification 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  $P(X = 1)$

We use  $\frac{1}{n} \sum_{i=1}^n X_i^*$

$$P(X^* = 1) = \pi_{11}P(X = 1) + \pi_{10}P(X = 0)$$

$$P(X^* = 1) - P(X = 1) = \pi_{10}P(X = 0) - \pi_{01}P(X = 1)$$

→ Examples:

**No bias** if  $P(X = 1) = \frac{1}{2}$  and  $\pi_{00} = \pi_{11}$

**Neg. bias** if  $P(X = 1) = 0.9$  and  $\pi_{00} = \pi_{11} = 0.9$

→ Bias =  $-0.1 \cdot 0.9 + 0.1 \cdot 0.1 = -0.08$

**Pos. bias** if  $P(X = 1) = 0.8$  and  $\pi_{11} = 0.99, \pi_{00} = 0.9$

→ Bias =  $-0.01 \cdot 0.8 + 0.1 \cdot 0.9 = 0.01$

## Effect of Misclassification

### Everything can happen

dependent on  $\pi_{11}$ ,  $\pi_{00}$  and  $P(X = 1)$ .

If  $\pi_{00} = \pi_{11}$  (in most times unrealistic) then

$$\text{Bias} = \pi_{00}(1 - 2P(X = 1))$$

$$P(X = 1) > 0.5 \implies \text{bias} < 0$$

$$P(X = 1) < 0.5 \implies \text{bias} > 0$$

### Attenuation towards 0.5

## Correction

### Idea: Solve the bias equation

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

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

### Assumptions

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

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

## Multinomial case

$X$  is multinomial with categories  $1, \dots, k$ .

$X^*$  is observed

The error model is given by the misclassification Matrix

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

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

$$\begin{aligned} P_X &= (p_{x1}, \dots, p_{xk})' \\ P_{X^*} &= (p_{x1}^*, \dots, p_{xk}^*)' \\ P_{X^*} &= \Pi * P_X \end{aligned}$$

# The matrix method

The correction method is given by

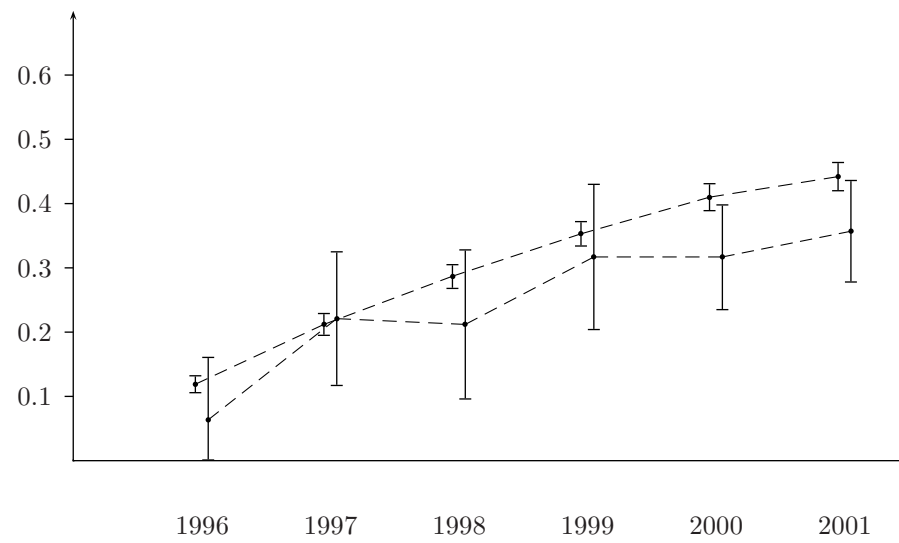
$$\hat{P}_X := \Pi^{-1} * \hat{P}_{X^*} \quad (1)$$

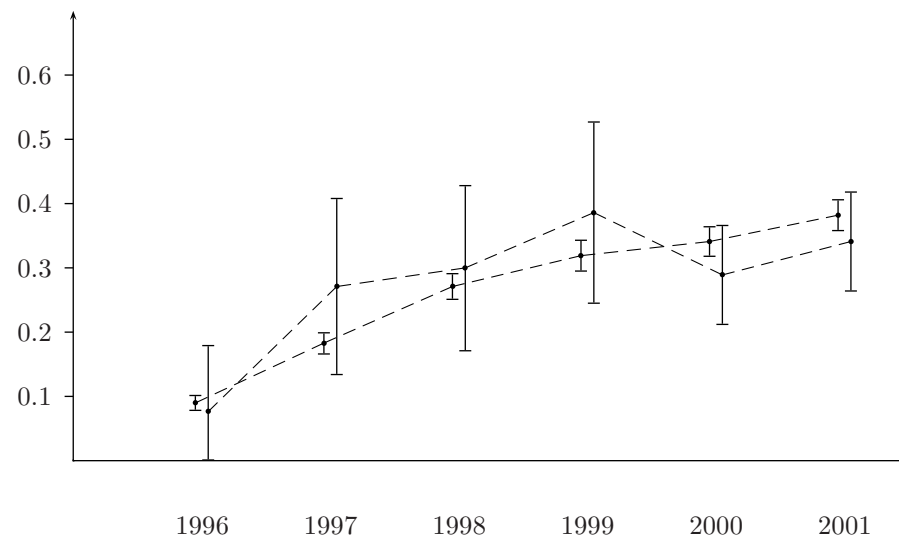
Properties

- Misclassification 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  $\Pi$ , 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





# 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

$P(X = 1|X^* = 1)$       (positive predicted value)

$P(X = 0|X^* = 0)$       (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 :

$$P(X_1^* = 0, X_2^* = 0) = P(X = 0) * \pi_{00}^2 + P(X = 1) * \pi_{01}^2)$$

$$P(X_1^* = 1, X_2^* = 1) = P(X = 0)\pi_{10}^2 + P(X = 1)\pi_{11}^2$$

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

$$P(X_1^* = 1, X_2^* = 0) = P(X_1^* = 0, X_2^* = 1)$$

Two independent equations, but three unknown parameters!

⇒ 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 (Heid, HK et al., work in progress)

# Kappa Statistics

Basic idea : Evaluate agreement and adjust for agreement by chance

Measuring agreement:

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

	$X_1^* = 0$	$X_1^* = 1$		$X_1^* = 0$	$X_1^* = 1$
$X_2^* = 0$	10	2	$X_2^* = 0$	5	2
$X_2^* = 1$	2	0	$X_2^* = 1$	2	5

Same proportion of agreement, but different situation !!

## Definition of Kappa

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

$$P_e = \frac{n_{0.}n_{.0}}{n^2} + \frac{n_{1.}n_{.1}}{n^2}$$

$$\kappa = (P_o - P_e)/(1 - P_e)$$

	$X_1^* = 0$	$X_1^* = 1$		$X_1^* = 0$	$X_1^* = 1$
$X_2^* = 0$	10	2	$X_2^* = 0$	5	2
$X_2^* = 1$	2	0	$X_2^* = 1$	2	5

$$\kappa < 0$$

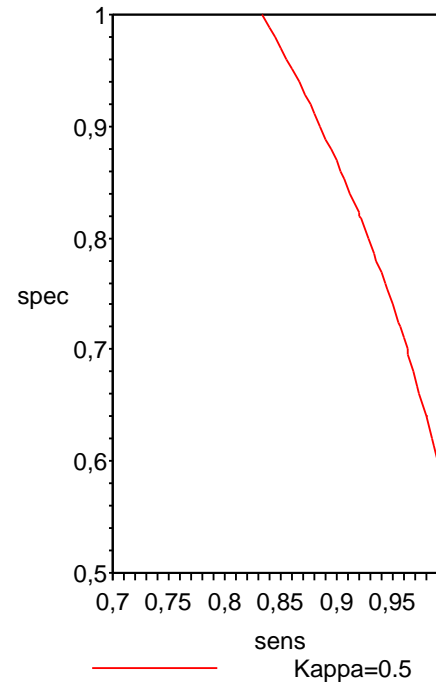
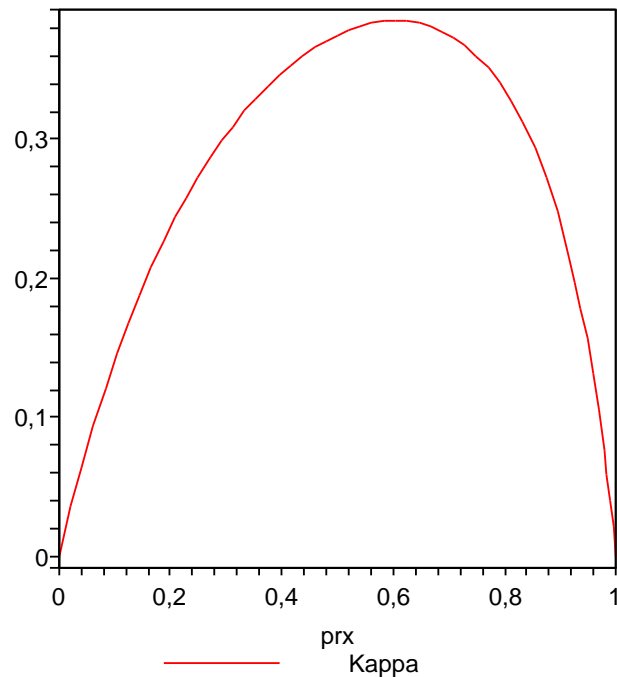
$$\kappa = 0.428$$

## Kappa and MC-Matrix

Kappa depends on the MC-Matrix and marginal distribution  $P(X=1)$

Fixed MC-Matrix  $\pi_{00} = 0.9, \pi_{11} = 0.7$  (1)

Sensitivity and specificity which result in  $\kappa = 0.5$  for  $P(X = 1) = 0.2(r)$



## 1.3 Bivariate analysis

### 2\*2 Tables, misclassification in disease status:

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

Model for misclassification:

$$\pi_{110} = P(Y^* = 1 | Y = 1, X = 0)$$

$$\pi_{111} = P(Y^* = 1 | Y = 1, X = 1)$$

$$\pi_{100} = P(Y^* = 1 | Y = 0, X = 0)$$

$$\pi_{101} = P(Y^* = 1 | Y = 0, X = 1)$$

**Non differential** misclassification if

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

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:

$$\pi_{110} = P(X^* = 1 | X = 1, Y = 0)$$

$$\pi_{111} = P(X^* = 1 | X = 1, Y = 1)$$

$$\pi_{100} = P(X^* = 1 | X = 0, Y = 0)$$

$$\pi_{101} = P(X^* = 1 | X = 0, Y = 1)$$

**Non differential** misclassification if

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

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

$$\begin{aligned} P(Y = 1, X^* = x^*) &= \sum_x P(Y = 1, X^* = x^*, X = x) \\ &= \sum_x P(Y = 1|X^* = x^*, X = x) * P(X^* = x^*, X = x) \\ &= \sum_x P(Y = 1|X = x) * P(X^* = x^*|X = x) * P(X = x) \end{aligned}$$

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

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

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

Bias to 1 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

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

Misclassification possibly on

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

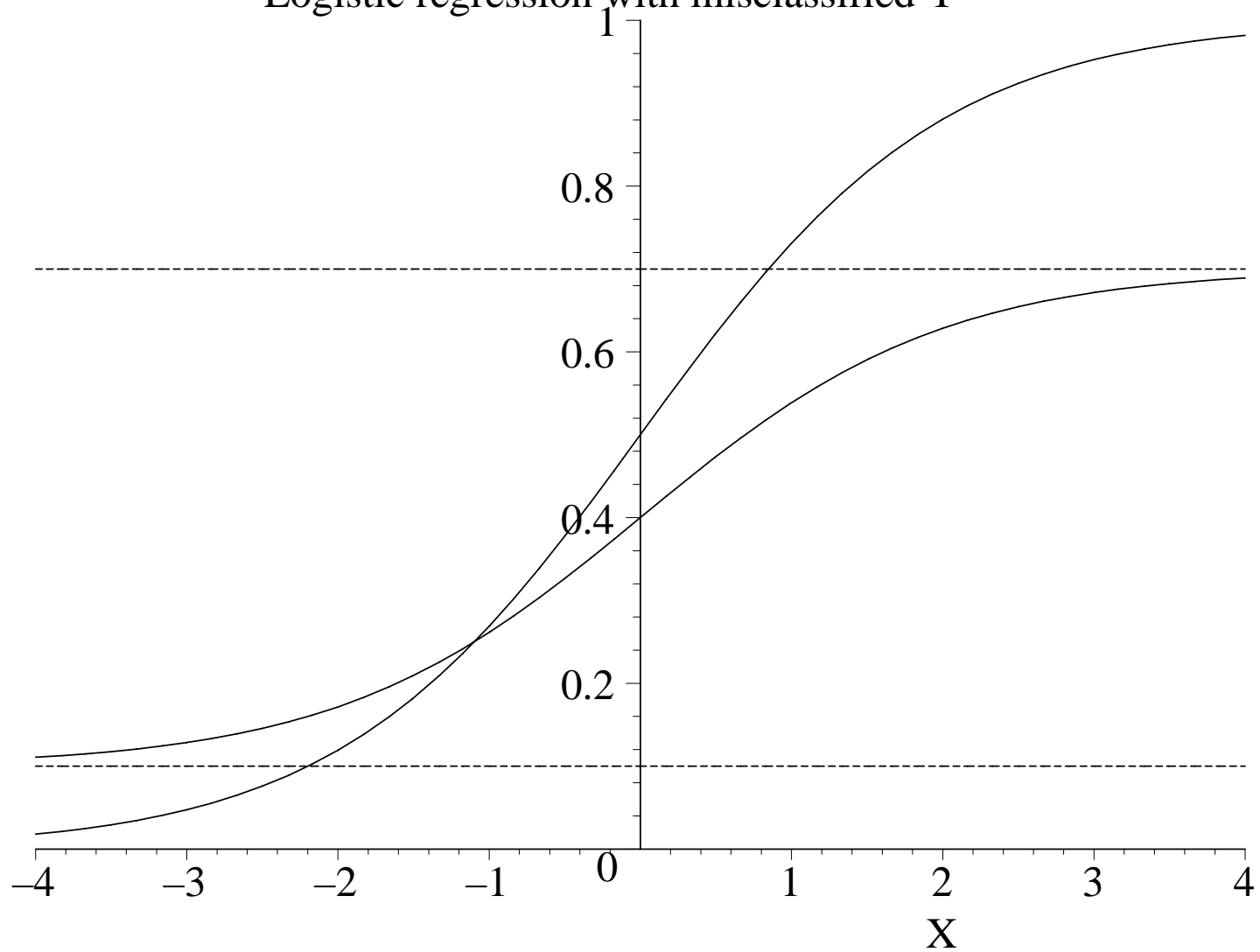
## Handling misclassification in $Y$ in binary regression

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

We observe  $Y^*$  instead of  $Y$  with misclassification matrix  $\Pi$

$$\begin{aligned}P(Y^* = 1|X) &= \pi_{11}G(x'\beta) + (1 - \pi_{10})(1 - G(x'\beta)) = H(x'\beta) \\H(t) &= \pi_{11}G(t) + (1 - \pi_{00})G(t)\end{aligned}$$

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:

$$E(Y|X^* = 0) = P(X = 0|X^* = 0) * \mu_0 + P(X = 1|X^* = 0) * \mu_1$$

$$E(y|X^* = 1) = P(X = 0|X^* = 1) * \mu_0 + P(X = 1|X^* = 1) * \mu_1$$

These equations can be solved for  $\mu_1$  and  $\mu_2$ , when MC Matrix and  $P(X=0)$  is known

Matrix Method

# Likelihood

$$\begin{aligned}L(Y = 1|X^* = x^*) &= \sum_x L(Y|X^* = x^*, X = x) \\&= \sum_x L(Y|X^* = x^*, X = x) * P(X^* = x^*, X = x) \\&= \sum_x L(Y|X = x) * P(X^* = x^*|X = x) * P(X = x)\end{aligned}$$

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 →

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

# Outlook

- Use of validation data
- Latent class analysis

# MISCLASSIFICATION AND MEASUREMENT ERROR IN REGRESSION MODELS

## Part 2

Helmut Küchenhoff  
Statistical Consulting Unit  
Ludwig-Maximilians-Universität München

Padova  
2./3.10.2007

## Outline 2. Measurement error: Models and effect

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

## 2. Measurement error Introduction

**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)

# 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"

# Examples

- Framingham heart study, Munich blood pressure study:  
Blood pressure, long term average  
Single measurement
- Munich bronchitis study:  
Average Occupational dust exposure  
Single measurement, expert ratings
- Silicosis study:  
Occupational dust exposure  
Job exposure matrix
- 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

## 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  $X$

$$X_i^* = X_i + U_i \quad (U_i, X_i) \text{ indep.}$$

$$E(U_i) = 0$$

$$V(U_i) = \sigma_U^2$$

$$U_i \sim N(0, \sigma_U^2)$$

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

<b>Statistical point of view:</b>	Accuracy :	Mean square error
	Validity :	Bias $E(U)$
	Reliability:	Measurement error variance $\sigma_U^2$

## Reliability measures

Two measurements

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

Assuming independence of the measurement errors  $U_{ij}$

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

$$R = \frac{\text{Var}(X_i)}{\text{Var}(X_{ij}^*)}$$

$$\text{Cor}(X_{i1}^*, X_{i2}^*) = \frac{\text{Cov}(X_{i1}^*, X_{i2}^*)}{\sqrt{\text{Var}(X_{i1}^*) * \text{Var}(X_{i2}^*)}} = R$$

$$\text{Cor}(X_{i1}^*, X_i) = \frac{\text{Cov}(X_{i1}^*, X_i)}{\sqrt{\text{Var}(X_{i1}^*) * \text{Var}(X_i)}} = \sqrt{R}$$

# 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:

Use

$$\text{Corr}(X_{i1}^*, X_{i2}^*)$$

Use

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

## 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

$\tau_j$  : 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

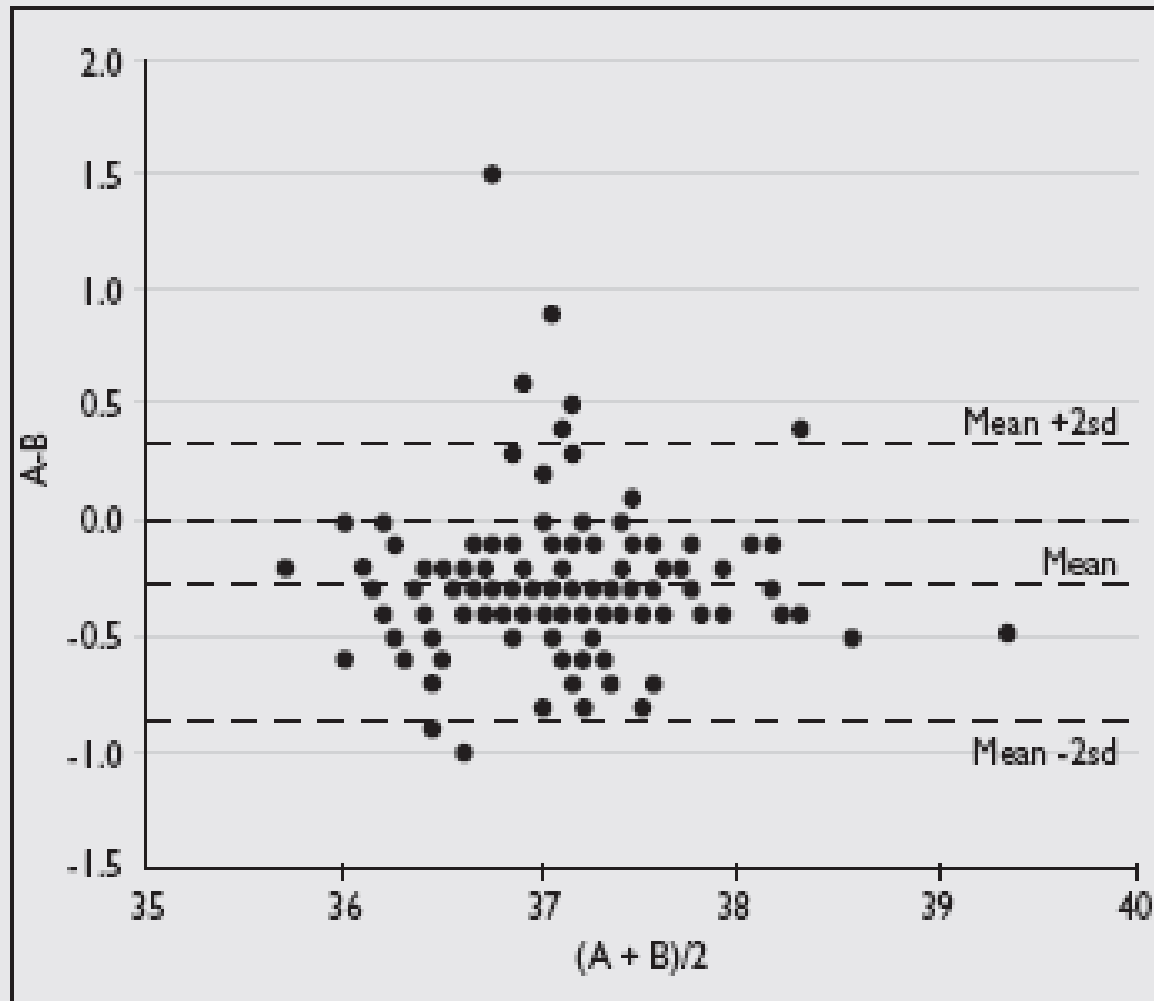
# Blend Altman Plots

Main Idea: Explore relationship between measurement error and true value

Data: Two types of measurement

Plot difference between two measurements and the mean

Fig. 6 Bland Altman plot.



# Approaches for Assessment of agreement

Choudhary and Ng (Biometrics 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 nonparametric (Splines) regression models
- Bootstrap and approximations

Useful for assessment, but correction methods cannot be derived

## Additive Berkson-error

$$X_i = X_i^* + U_i \quad (U_i, X_i^*) \text{ indep.}$$

$$E(U_i) = 0$$

$$V(U_i) = \sigma_U^2$$

$$U_i \sim N(0, \sigma_U^2)$$

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

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

Note that in the Classical additive case

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

## Multiplicative measurement error

$$X_i^* = X_i * U_i \quad (U_i, X_i) \text{ indep.}$$

Classical

$$X_i = X_i^* * U_i \quad (U_i, X_i^*) \text{ indep.}$$

Berkson

$$E(U_i) = 1$$

$$U_i \sim \text{Lognormal}$$

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

# Measurement error in response

## Simple linear regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y^* = Y + U \text{ additive measurement error}$$

→

$$Y^* = \beta_0 + \beta_1 X + \varepsilon + U$$

New equation error:  $\varepsilon + U$

Assumption :  $U$  and  $X$  independent,  $U$  and  $\varepsilon$  independent

→ Higher variance of  $\varepsilon$

→ 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  $\beta_1$

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 differential measurement error relates to the response:

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

For  $Y$  there is no further information in  $U$  or  $W$  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

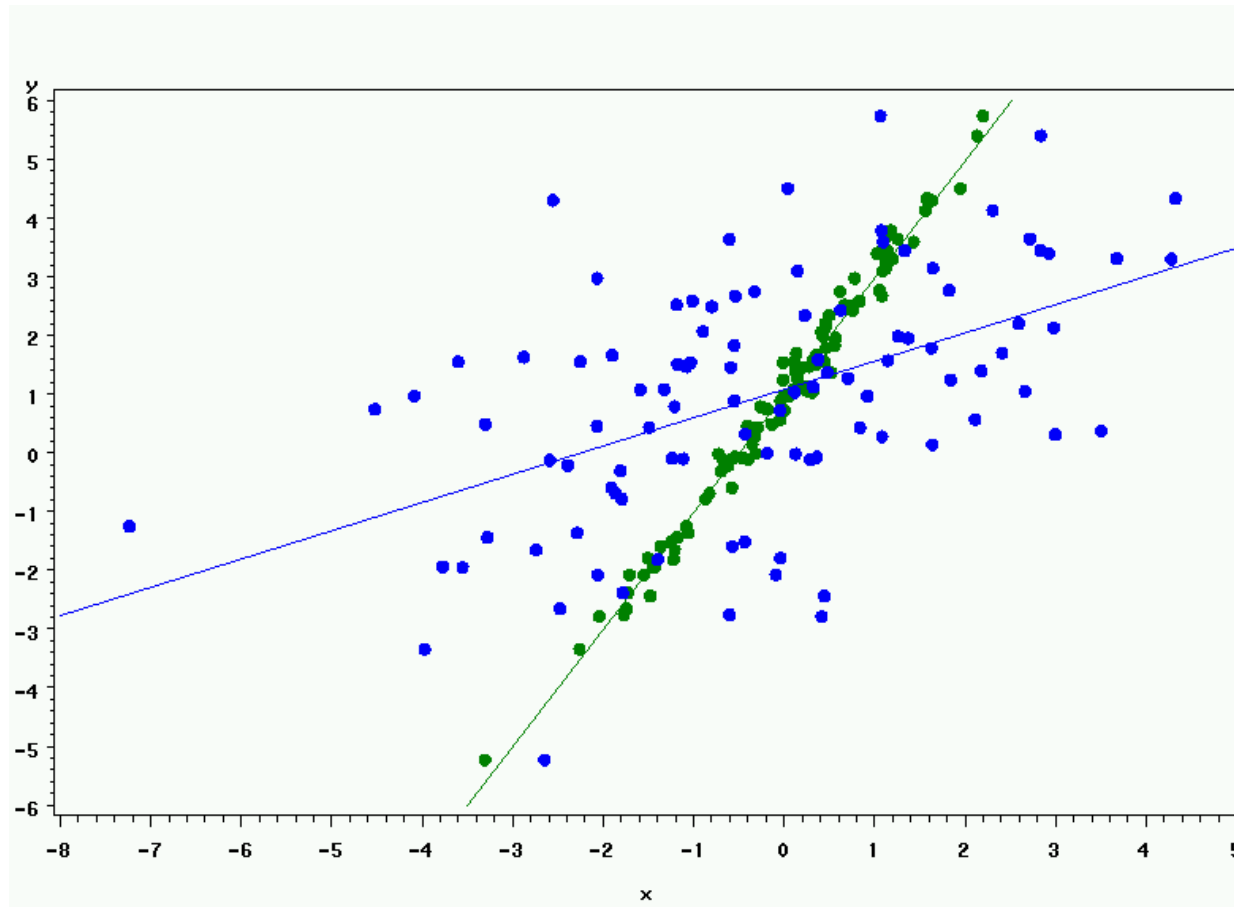
## Simple linear regression

We assume a classical non differential additive normal measurement error

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

## SAS-Simulation for linear m.e. model

```
/* simulation of x ~ N(0,1) */  
data sim ;  
do i=1 to 100; x=rannor(137);  
output;  
end;  
run; /* simulation of Y = 1+2*x + epsilon*/  
/* simulation of the surrogate with additive m.e */  
data sim ;  
set sim; su=2; /* measurement error std*/  
y= 1+2*x+0.3*rannor(123); w= x+su*rannor (167) ;  
run;  
/* Plot options green for the correct and blue for the surrogates */  
symbol1 c = green V = dot; symbol2 c = blue V = dot;  
proc gplot data = sim;  
symbol i=none; plot y*x y*w /overlay; /* scatter*/  
symbol1 i=r; /* regression lines */ symbol2 i=r;
```



*Figure 1: Effect of additive measurement error on linear regression*

## The observed model in linear regression

$$E(Y|X^*) = \beta_0 + \beta_1 E(X|W)$$

Assuming  $X \sim N(\mu_x, \sigma_x^2)$ , the observed model is:

$$E(Y|X^*) = \beta_0^* + \beta_1^* X^*$$

$$\beta_1^* = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \beta_1$$

$$\beta_0^* = \beta_0 + \left(1 - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right) \beta_1 \mu_x$$

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

- The observed model is still a linear regression !

- Attenuation of  $\beta_1$  by the factor  $\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$   
"Reliability ratio"
- Loss of precision (higher error term)

# Identification

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

$$(\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_u^2, \sigma_\epsilon^2) \text{ and } (\beta_0^*, \beta_1^*, \mu_x, \sigma_x^2 + \sigma_u^2, 0, \sigma_\epsilon)$$

yield the identical distributions of  $(Y, X^*)$ .  $\implies$  The model parameters are not identifiable

We need extra information, e.g

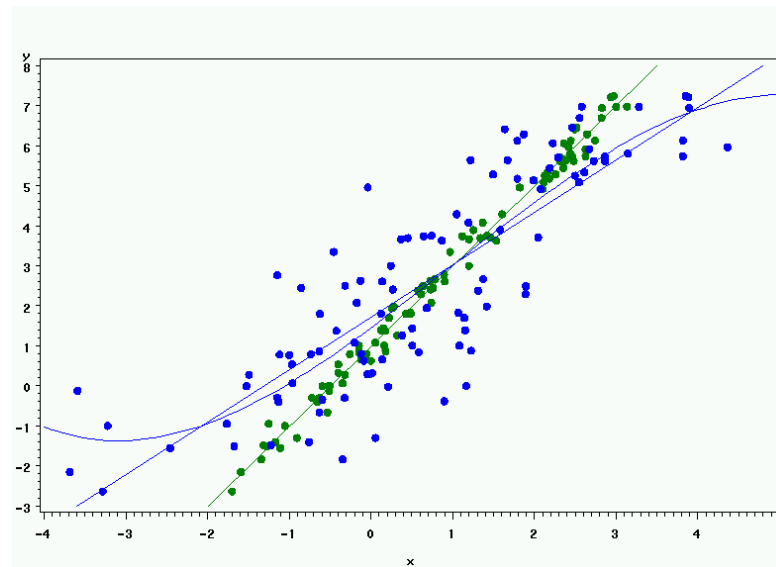
- $\sigma_u$  is known or can be estimated

- $\sigma_u/\sigma_\epsilon$  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 :

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

For the intercept:

$$\hat{\beta}_{0n} = \bar{Y} - \beta_{1n}\bar{X}^*$$

$$\begin{aligned}\text{plim}(\hat{\beta}_{0n}) &= \mu_y + \beta_1 * \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} * \mu_x \\ &= \beta_0 + \beta_1 * \left(1 - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right) * \mu_x\end{aligned}$$

## Naive LS- estimation (2)

For the residual variance:

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

## Multiple linear regression

The generalization from 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

$$\frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}$$

$\sigma_{x|z}^2$  is residual variance from regressing  $X$  on  $Z$

- $\hat{\beta}_{zn}$  is also biased

$$\hat{\beta}_{zn} \longrightarrow \beta_z + \left(1 - \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}\right) \beta_x \gamma_z$$

$\gamma_z$  is regression coefficient when regressing  $X$  on  $Z$

## Correction for attenuation

We have a first method: Solve the bias equation:

$$\hat{\beta}_1 = \hat{\beta}_{1n} \frac{\sigma_x^2 + \sigma_u^2}{\sigma_x^2}$$

$$\hat{\beta}_1 = \hat{\beta}_{1n} \frac{S_{x^*}^2}{S_{x^*}^2 - \sigma_u^2}$$

$$\hat{\beta}_0 = \hat{\beta}_{0n} - \hat{\beta}_1 \left( \frac{S_{x^*}^2 - \sigma_u^2}{S_{x^*}^2} \right) \bar{W}$$

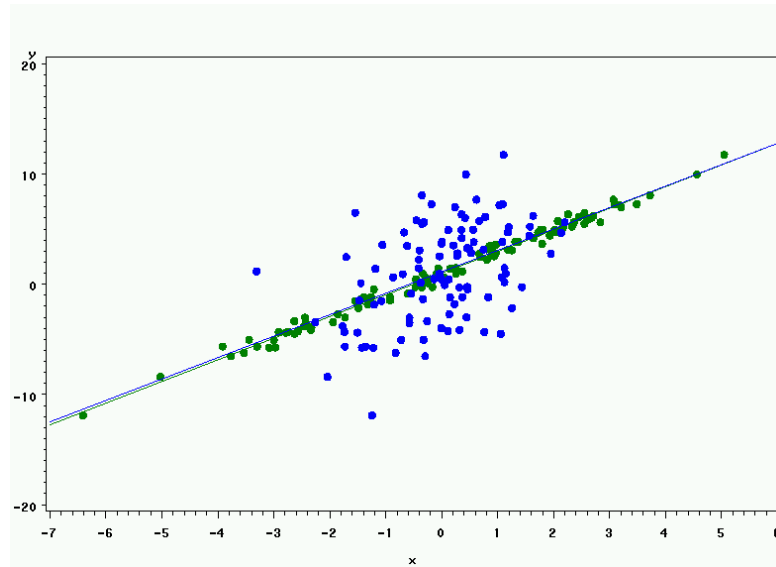
Correction by reliability ratio.

$V(\hat{\beta}_1) > V(\beta_{1n})$  Bias Variance trade off

## Berkson-Error in simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$X = X^* + U, \quad U, (X^*, Y) \text{ indep.}, \quad E(U) = 0$$



## Observed Model

$$E(Y|X^*) = \beta_0 + \beta_1 X^*$$

$$V(Y|X^*) = \sigma_\epsilon^2 + \beta_1 * \sigma_u^2$$

- Regression model with identical  $\beta$
- Measurement error ignorable
- Loss of precision

# Binary Regression

Logistic with additive non differential measurement error

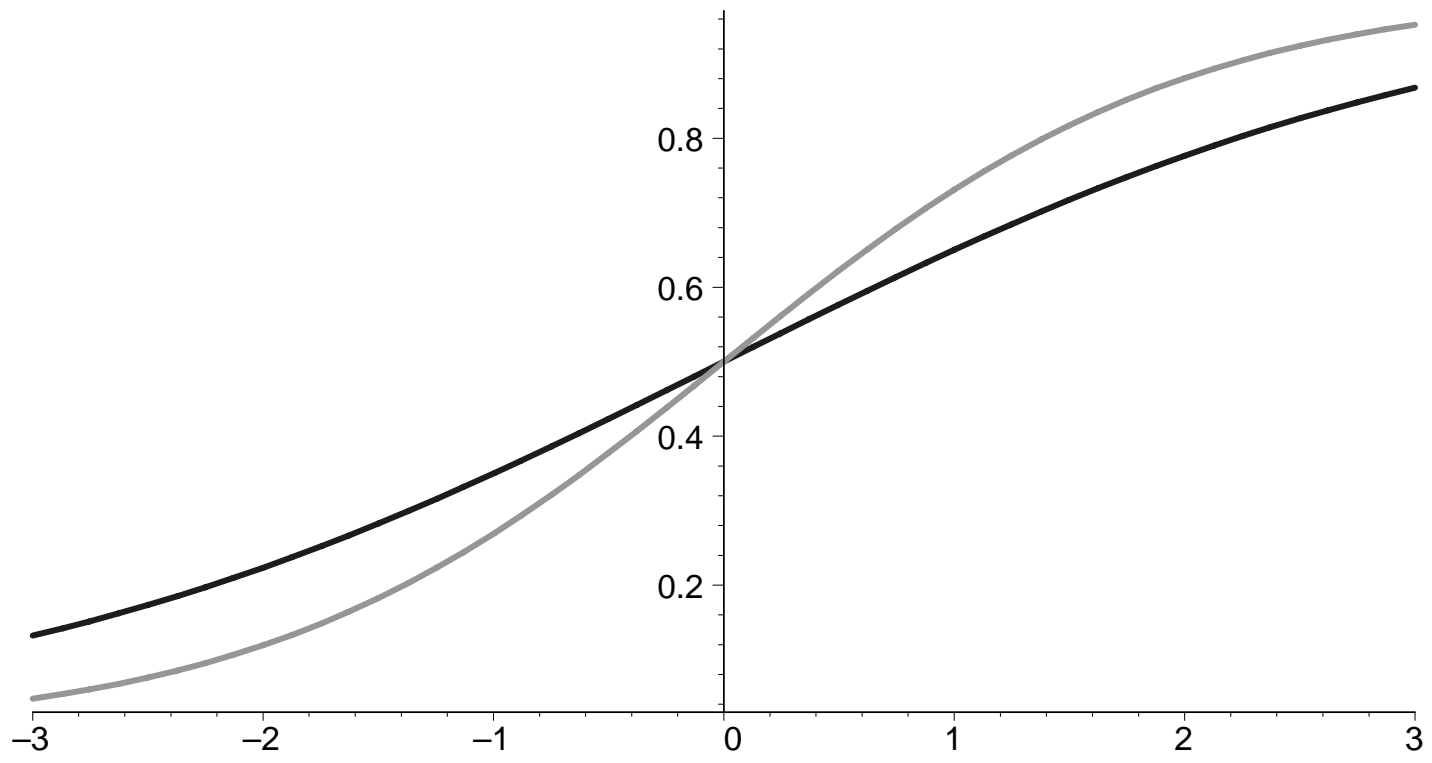
$$\begin{aligned}P(Y = 1|X) &= G(\beta_0 + \beta_1 X) \\G(t) &= (1 + \exp(-t))^{-1} \\X^* &= X + U\end{aligned}$$

**Observed model:**

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

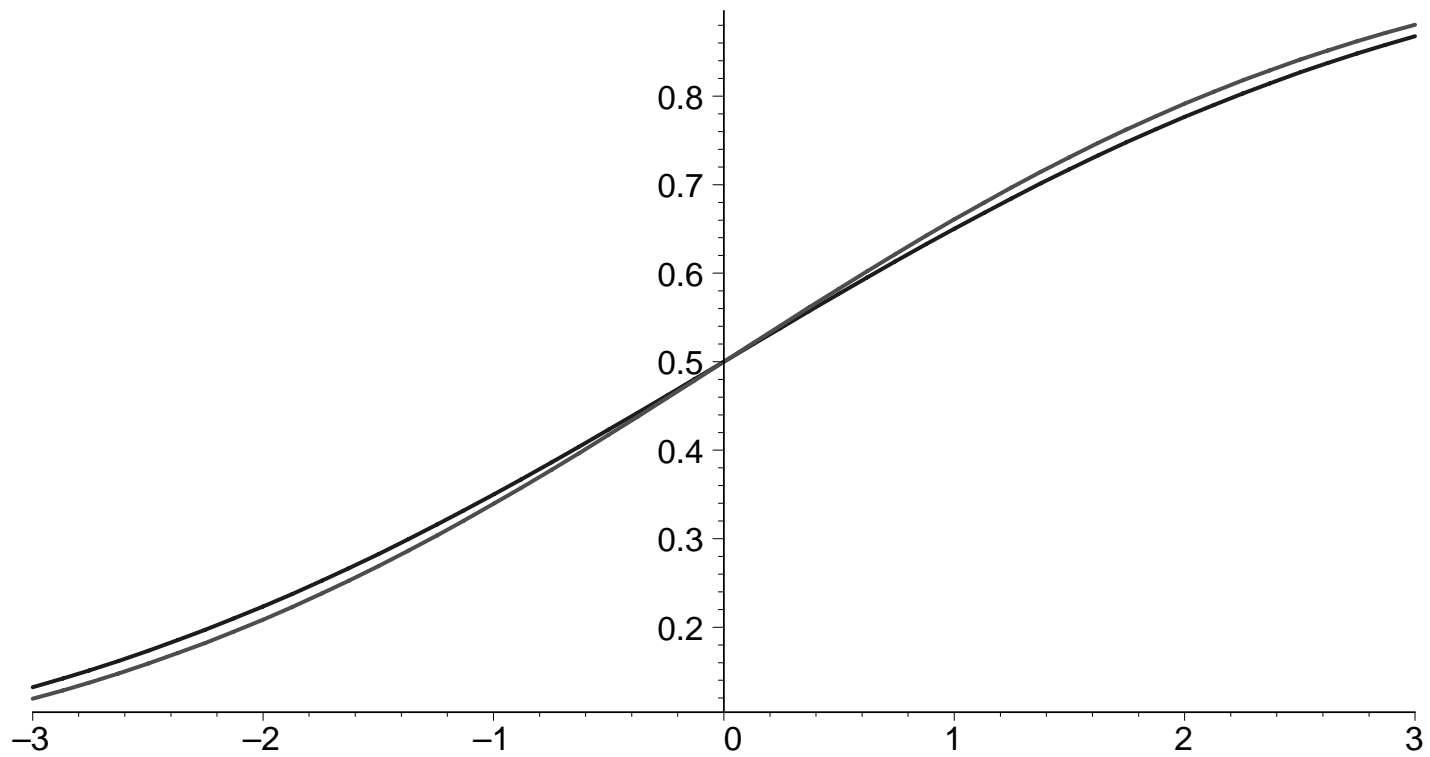
If we have additive measurement error and  $X$  and  $U$  are normal then  $X|X^*$  is also normal

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

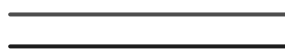


Legend

— true  
— observed



Legend



observed  
lin approx

## Probit Model

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

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

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  $\sqrt{1 + v\beta_1^2 v h_*^{-2}}$

## Effect of measurement error in survival

Example: Uranium miner study (HK, Langner , Bender, LDA, 2007)

- 60 000 male workers
- Date of birth
- Job history (job periods with exact dates)
- Response: cancer mortality
- Radiation exposure by means of job exposure matrix
- Multiplicative or additive Berkson error

## True and observed survivor function

Cox regression model: Survivor function:  $S_{true}(t, X) = P(T > t|X)$   
 Hazard:  $h_{true}(t, x) = \exp(X\beta)h_0(t)$

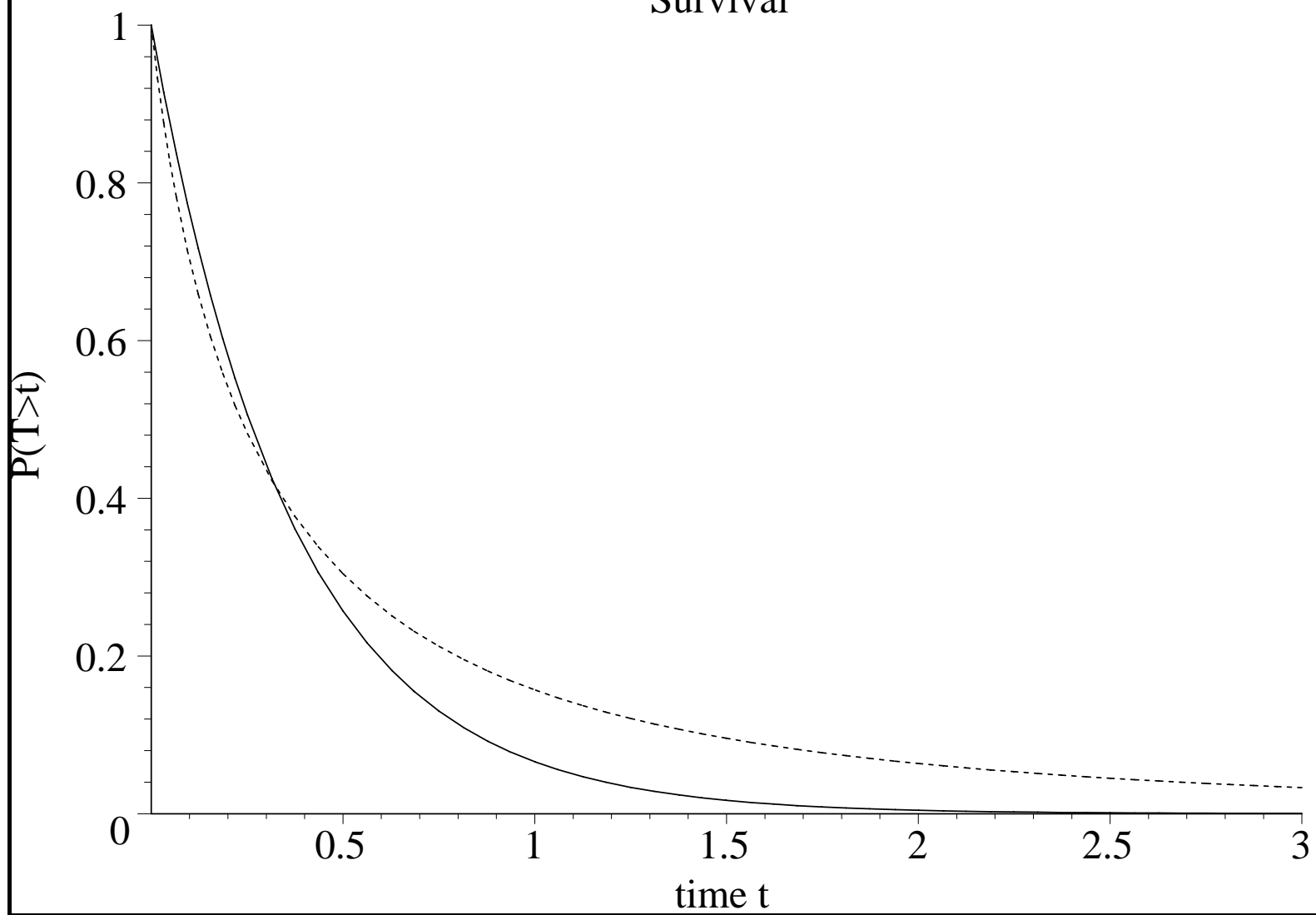
For an additive non differential Berkson error U we get

$$\begin{aligned} S_{obs}(t, X^*) &:= P(T \geq t|X^*) \\ &= \int S_{true}(t, X^* + u) f_u(u) du \\ h_{obs}(t, X^*) &= \frac{\partial}{\partial t} S_{obs}(t, X^*) / S_{obs}(t, X^*) \end{aligned}$$

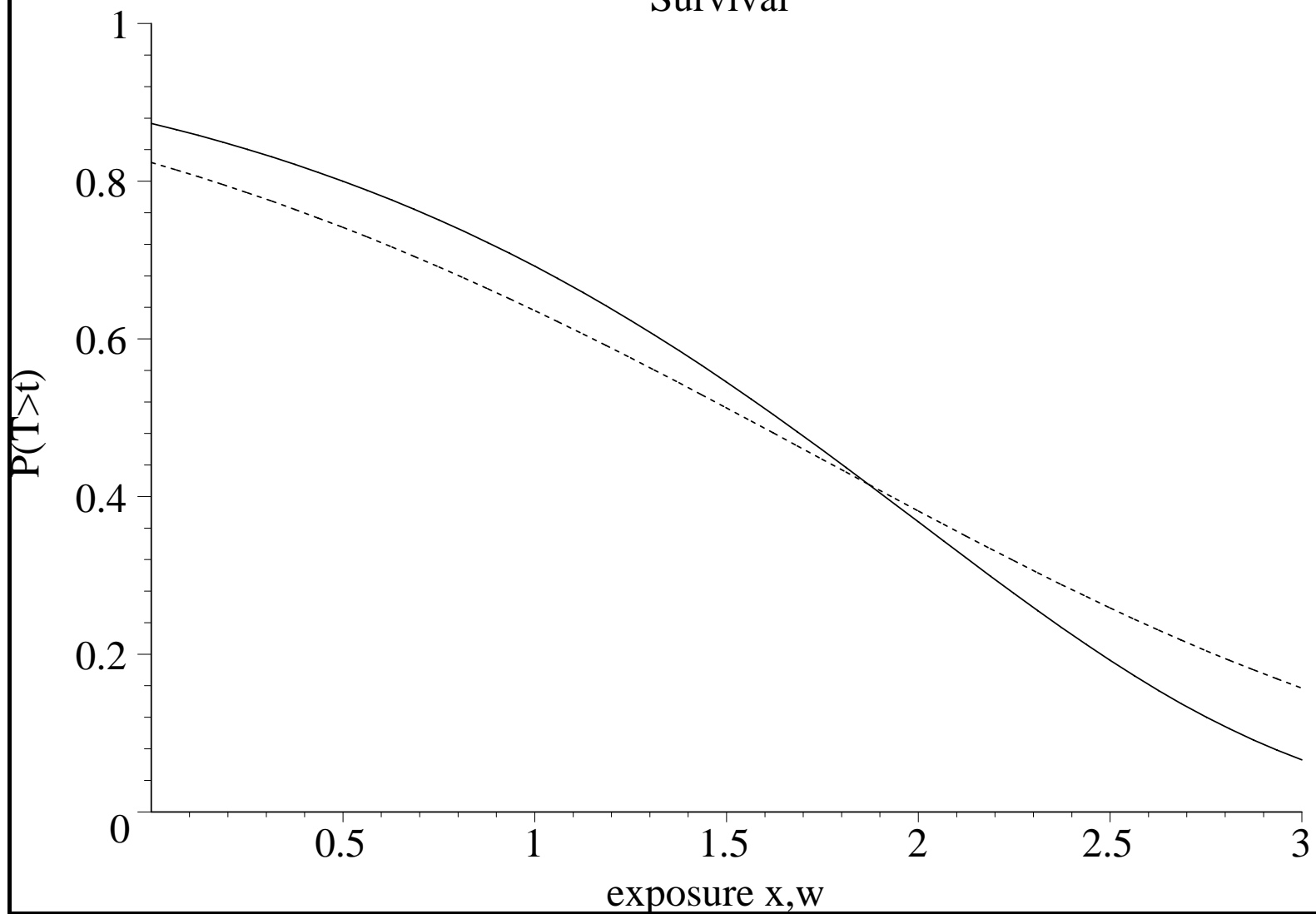
$$h_{obs}(t, X^*) = \int h_{true}(t, X^* + U) f_u(U) du = \exp(X^* \beta + \sigma_u^2 / 2)$$

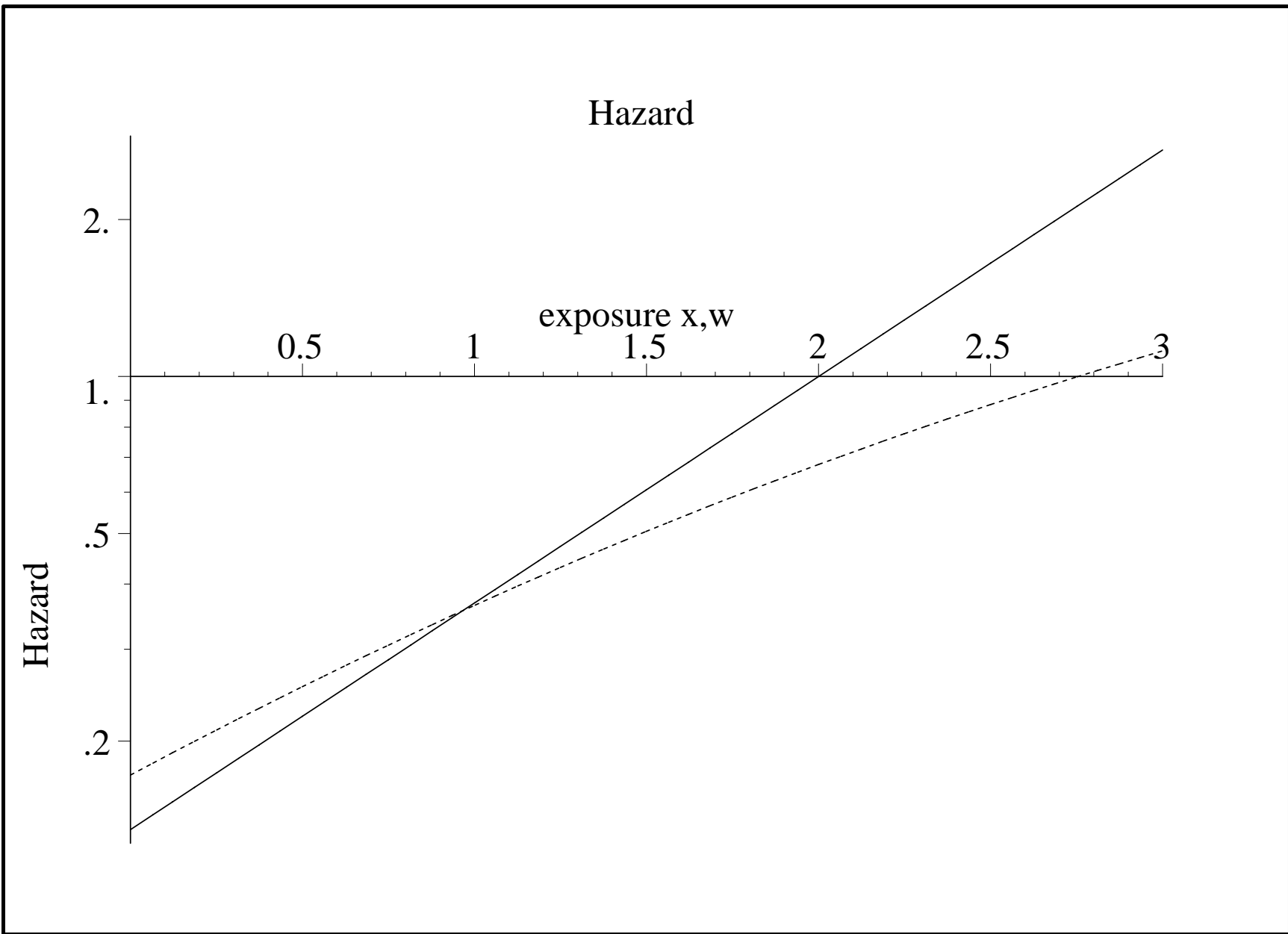
holds only for the rare disease assumption

# Survival

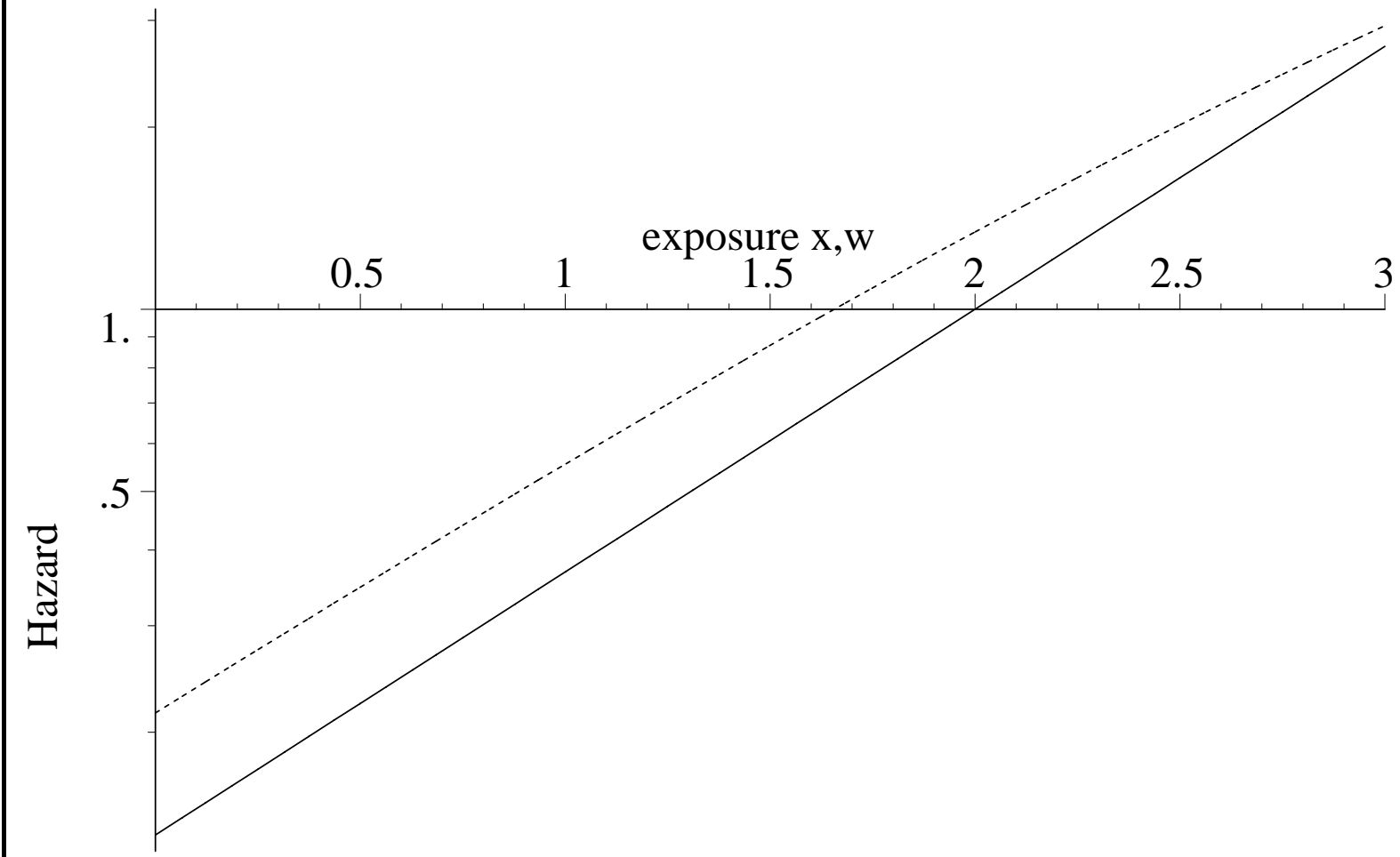


# Survival

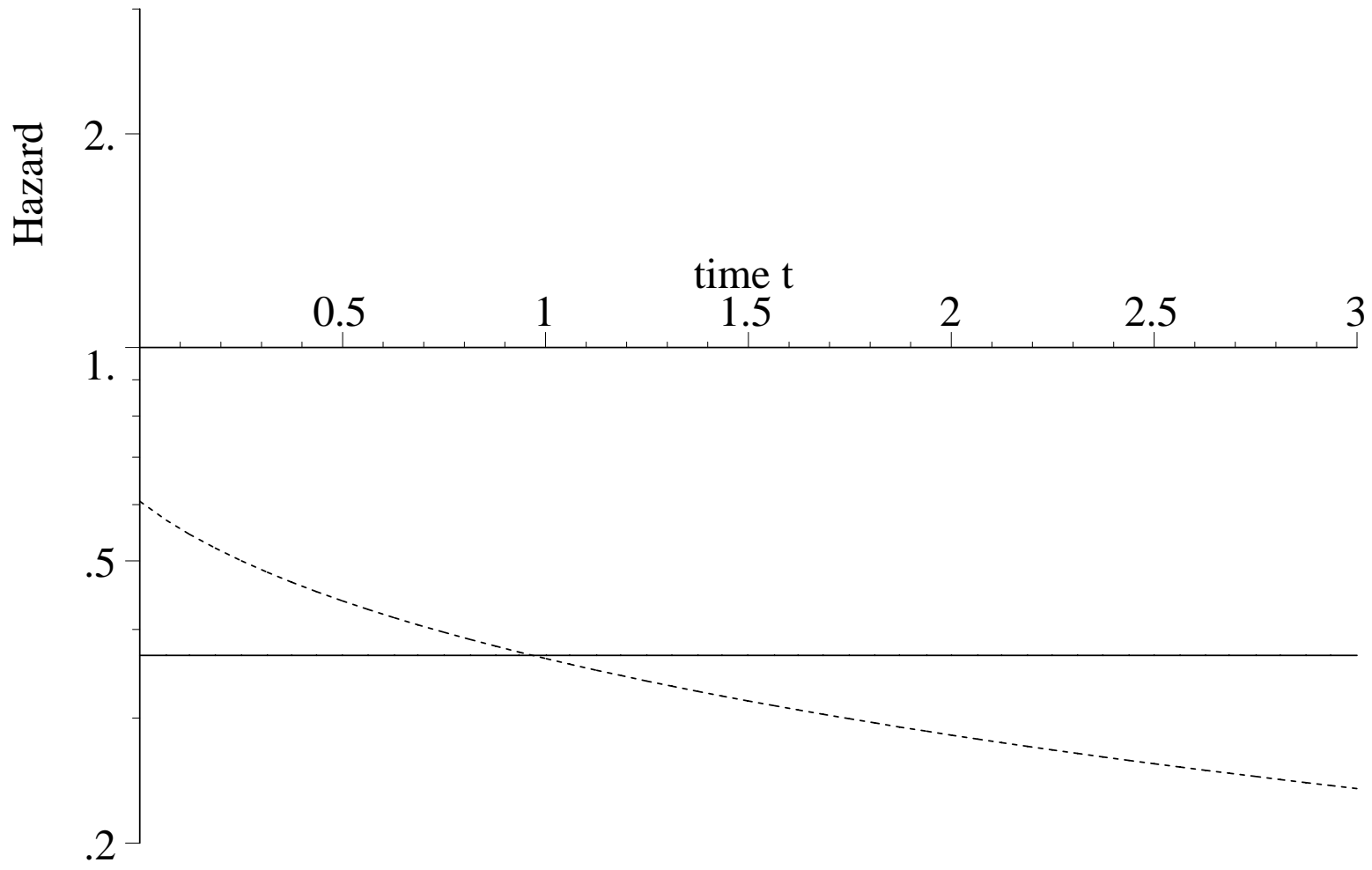


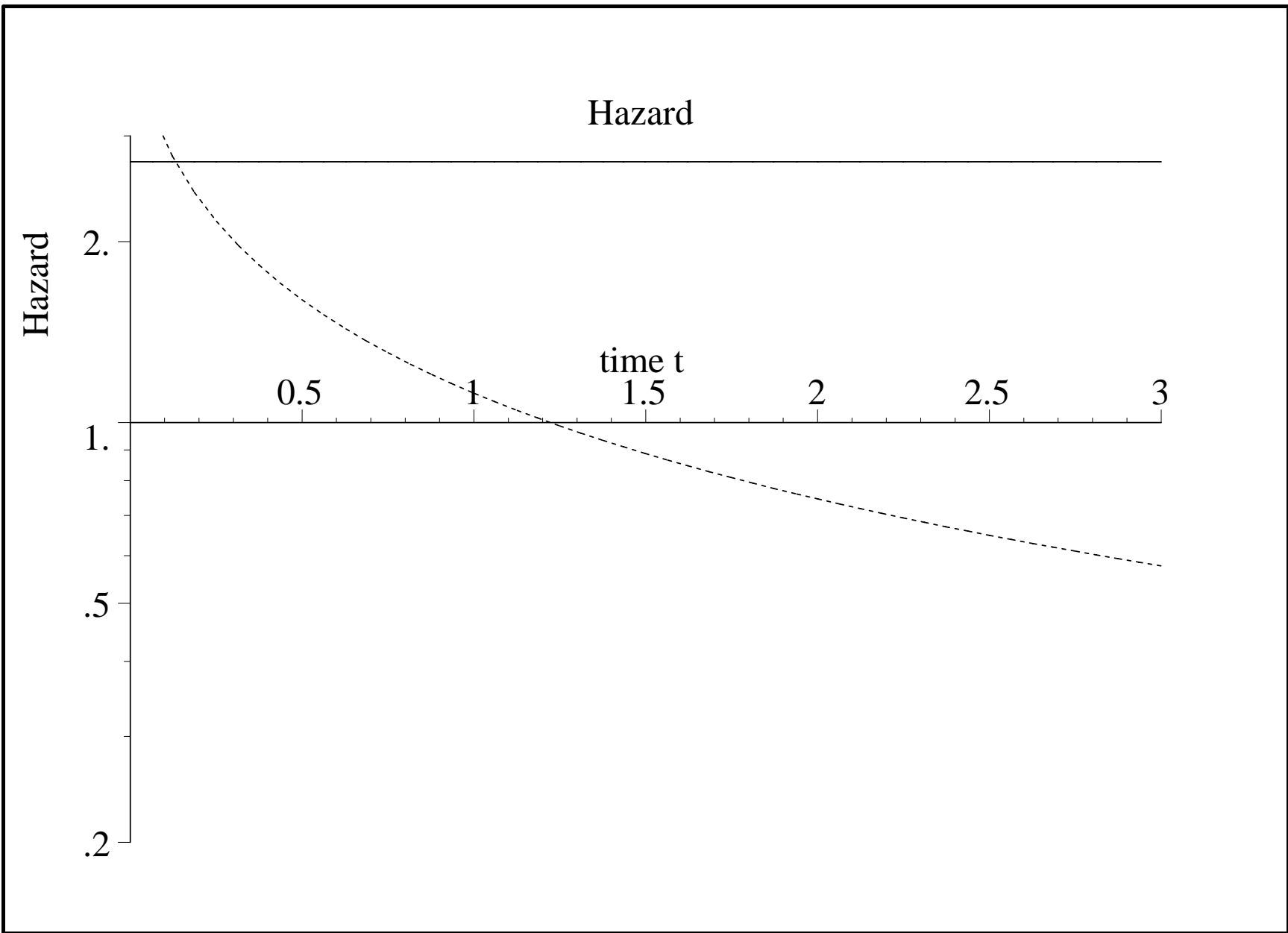


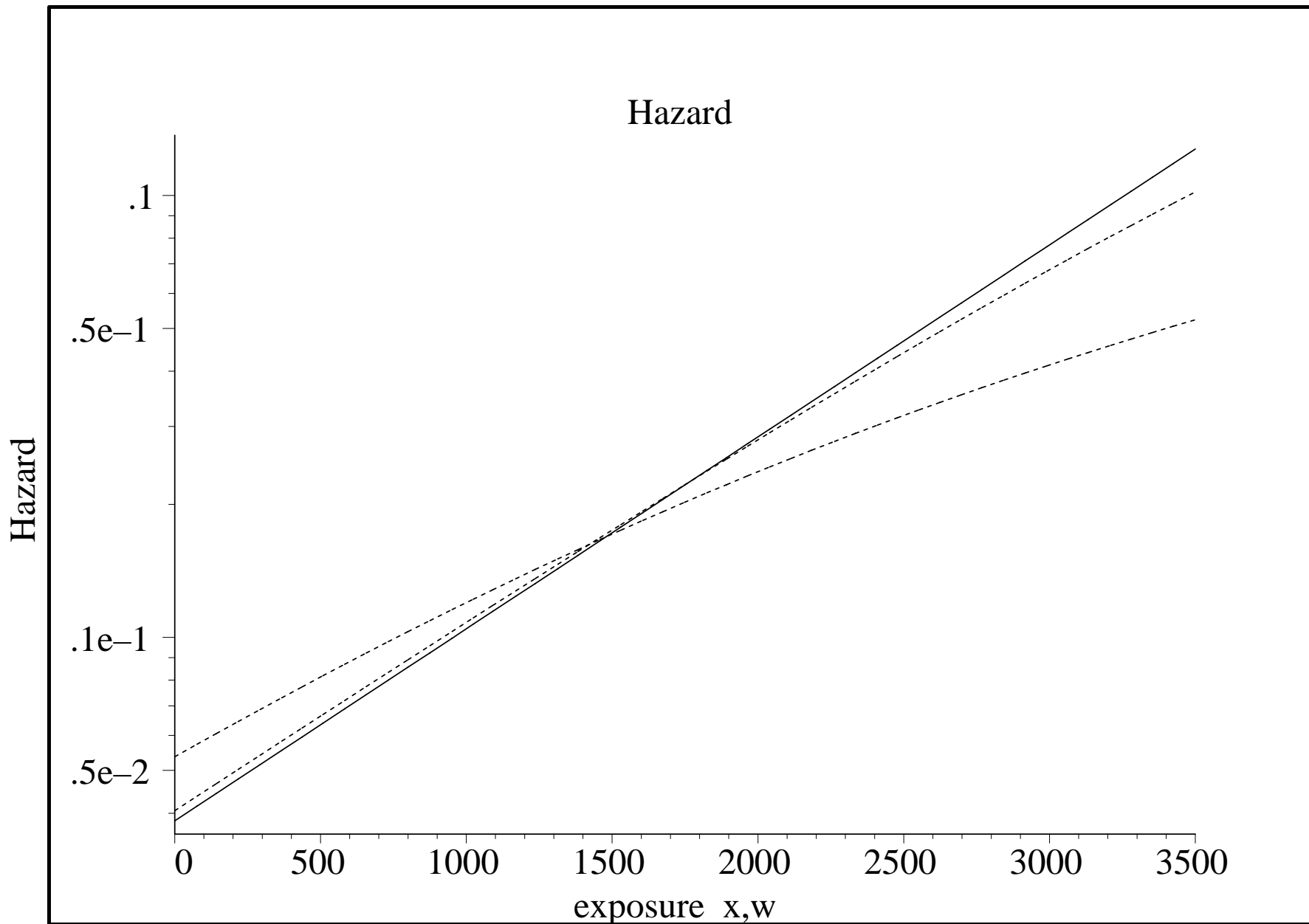
# Hazard

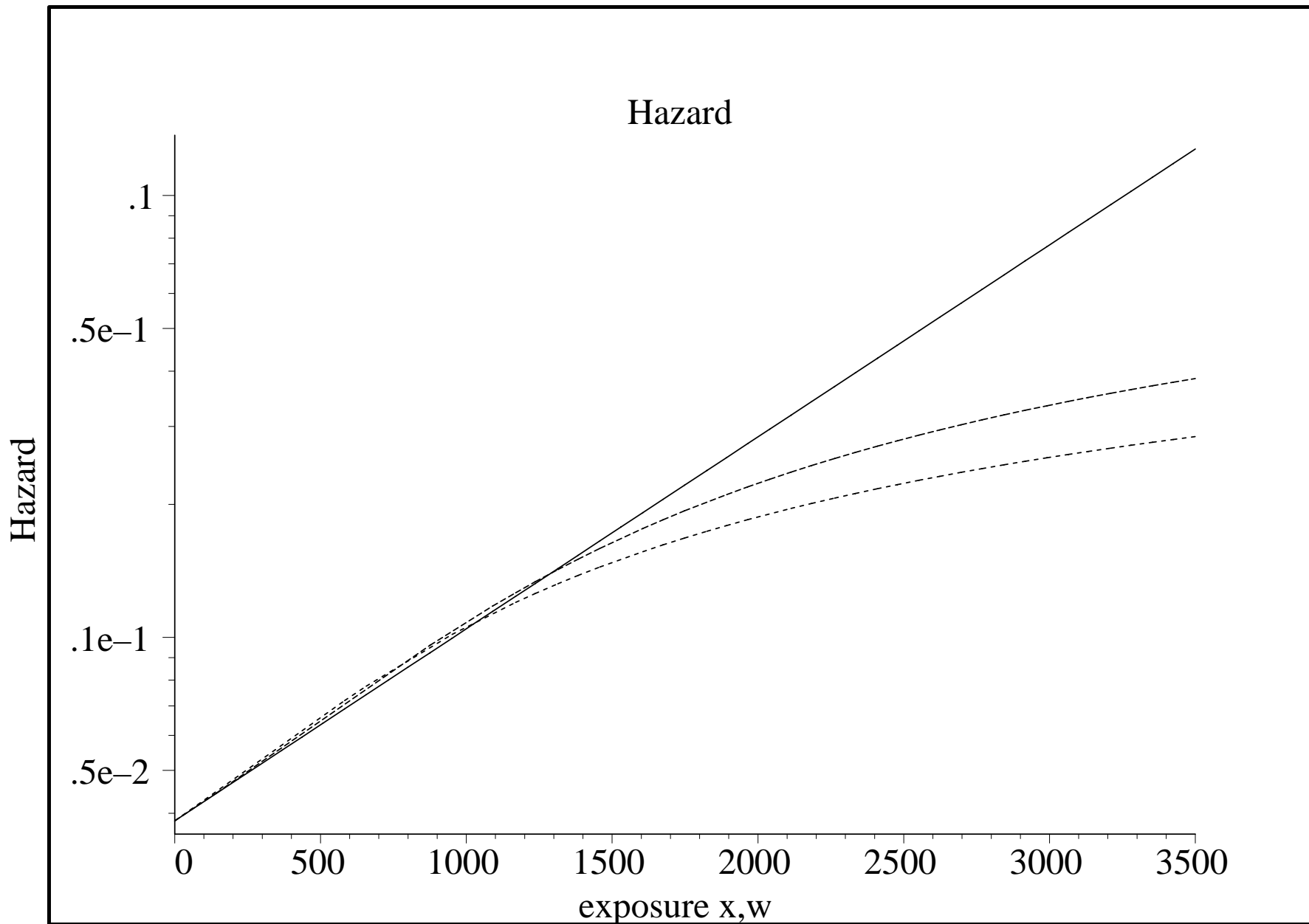


# Hazard









## Effects of measurement error

- No effect for rare disease
- Attenuation

## Effects of measurement error

- Baseline hazard different
- Proportional hazard assumption does not hold in the observed model

Calculations for the situation of the GUM study

# MISCLASSIFICATION AND MEASUREMENT ERROR IN REGRESSION MODELS

## Part 3

Helmut Küchenhoff  
Statistical Consulting Unit  
Ludwig-Maximilians-Universität München

Padova  
2./3.10.2007

### 3. Methods

- Functional and structural
- Correction and method of moments and orthogonal regression
- Regression calibration
- Likelihood
- Quasi likelihood
- Bayes
- Corrected score

## Functional and structural

- **Functional:**

$X$  fixed unknown constants

No assumptions about the distribution of  $X$

- **Structural:**  $X$  latent random variable

Use assumptions about the distribution of  $X$

## Direct correction

- Find

$$p \lim \hat{\beta}_n = \beta^* = f(\beta, \gamma)$$

- Then

$$\hat{\beta} := \hat{f}^{-1}(\hat{\beta}_n)$$

- Correction for attenuation in linear model
- General case Kü (1997)

## Method of moments

Moments of observed data can be estimated  
Solve moments equations  
Simple linear regression:

$$\mu_{X^*} = \mu_X$$

$$\mu_Y = \beta_0 + \beta_1 \mu_x$$

$$\sigma_{x^*}^2 = \sigma_x^2 + \sigma_u^2$$

$$\sigma_y^2 = \beta_1^2 \sigma_x^2 + \sigma_\epsilon^2$$

$$\sigma_{yx^*} = \beta_1 * \sigma_x^2$$

# Orthogonal regression, total least squares

In linear Regression with classical additive measurement error:

- Assume  $\frac{\sigma_\epsilon^2}{\sigma_u^2} = \eta$  is known,

e. g. no equation error and  $\sigma_\epsilon$  is measurement error in  $Y$ . Minimize

$$\sum_{i=1}^n \{(Y_i - \beta_0 - \beta_1 X_i)^2 + \eta(X_i^* - X_i)^2\}$$

in  $(\beta_0, \beta_1, X_1, X_2, \dots, X_n)$ .

- Total least squares, Van Huffel (1997)

- Technical symmetric applications
- In other applications a problem, assumption of no equation error not realistic

## Regression calibration

This simple method has been widely applied. It was suggested by different authors: Rosner et al. (1989) Carroll and Stefanski(1990)

1. Find a model for  $E(X|X^*, Z)$  by validation data or replication
2. Replace the unobserved  $X$  by estimate  $E(X|X^*, Z)$  in the main model
3. Adjust variance estimates by bootstrap or asymptotic methods
  - Good method in many practical situations
  - Calibration data can be incorporated
  - Problems in highly nonlinear models

## Regression calibration

- Berkson case:  $E(X|X^*) = X^* \longrightarrow$   
Naive estimation = Regression calibration
- Classical : Linear regression  $X$  on  $X^*$

$$E(X|X^*) = \frac{\sigma_x^2}{\sigma_w^2} * X^* + \mu_X * \left(1 - \frac{\sigma_x^2}{\sigma_w^2}\right)$$

Correction for attenuation in linear model

# Survival

For Cox Model and rare disease assumption appropriate

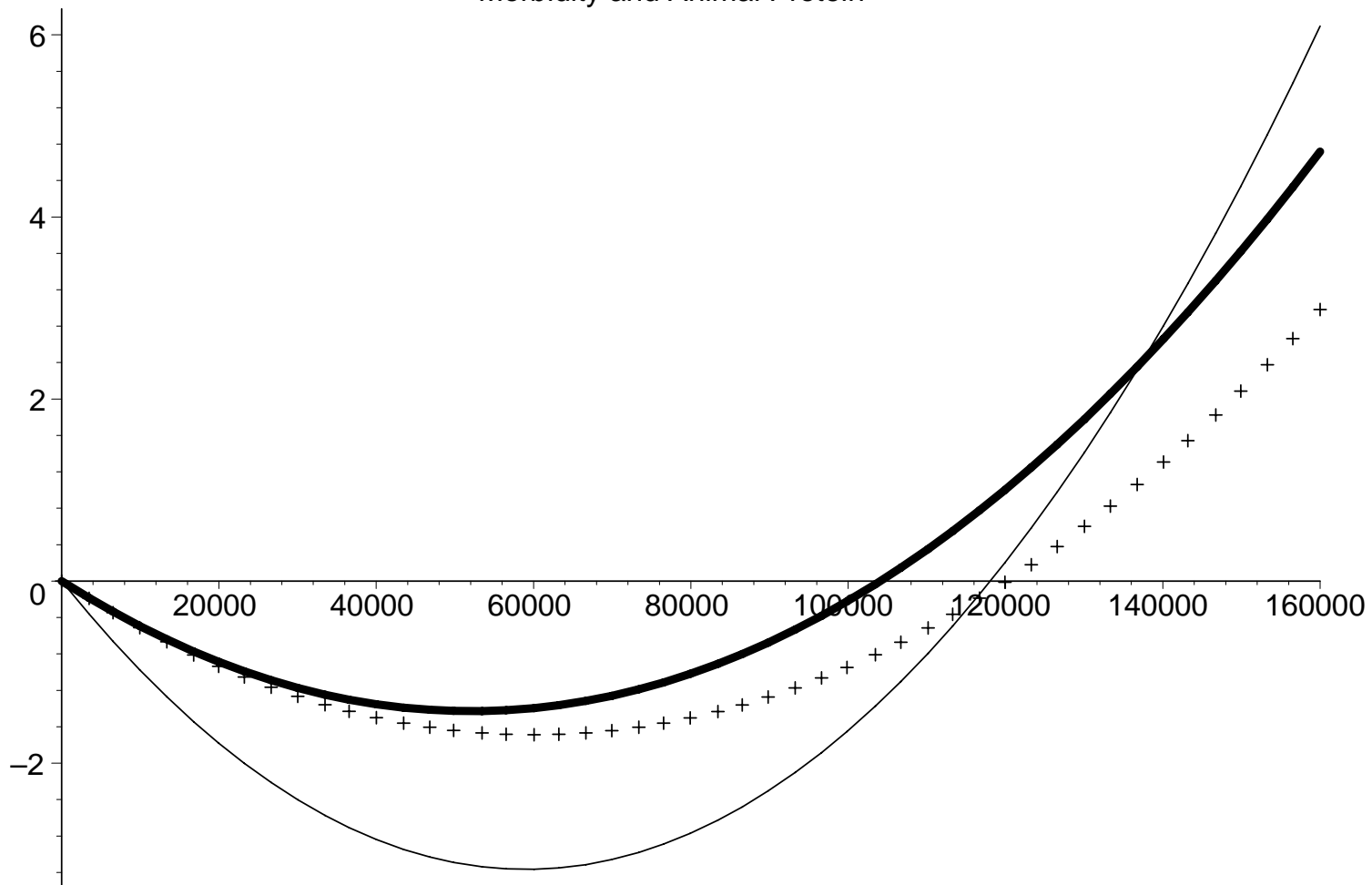
**Example: MONICA study, Augustin(2002)**

- CHD and fat intake
- Cox-Regression
- Quadratic model
- Classical additive measurement error
- Heteroscedastic measurement error
- Replication (7 days) for estimating measurement error variance

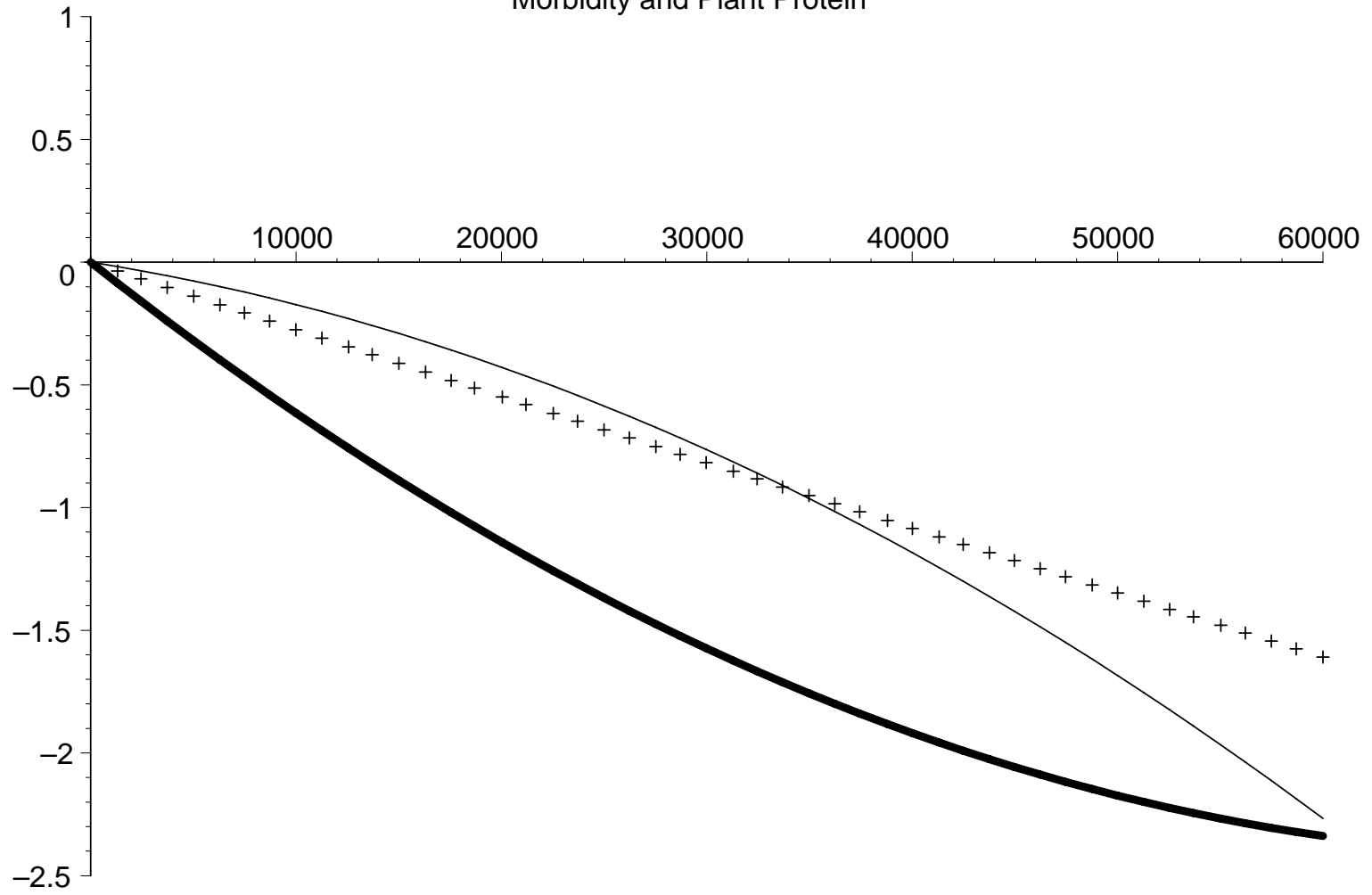
# Results

**Correction by regression calibration, animal protein**

Morbidity and Animal Protein



# Morbidity and Plant Protein



Results differ for assumption of homoscedastic and heteroscedastic measurement error

# Likelihood methods

- Standard inference can be done with standard errors and likelihood ratio tests
- Efficiency
- Combination of different data types are possible
- Sometimes more accurate than approximations
- Difficult to calculate
- Software not available

- Parametric model for the unobserved predictor necessary
- Robustness to strong parametric assumptions

## The classical error likelihood

Main model	$[Y   X, Z, \beta]$
Error model	$[X^*   X, \eta]$
Exposure model	$[X   Z, \lambda]$

$$[\mathbf{Y}, \mathbf{X}^* | \mathbf{Z}, \theta] = \prod_{i=1}^n \int [y_i | x, z_i, \zeta] [x_i^* | x, \eta] [x | z_i, \lambda] d\mu(x),$$

where  $\theta = (\zeta, \eta, \lambda)$

- Evaluation by numerical integration

## Berkson likelihood

3 components, but the third component contains no information

Main model	$[Y   X, Z, \zeta]$
Error model	$[X   X^*, \eta]$
Exposure model	$[X^*   Z, \lambda],$

$$[\mathbf{Y}, \mathbf{X}^* | \mathbf{Z}, \theta] = \prod_{i=1}^n \int [y_i | x, z_i, \zeta] [x | w_i, \eta] [w_i | z_i, \lambda] d\mu(x)$$

$$[\mathbf{Y}, \mathbf{X}^* | \mathbf{Z}, \theta] = \prod_{i=1}^n \int [y_i | x, z_i, \zeta] [x | w_i, \eta] d\mu(x) * const$$

**The Berkson likelihood does not depend on the exposure model**

## Quasi likelihood

If calculation of the likelihood is too complicated use

$$E(Y|X^*, Z) = \int g(X, Z) f_{x|x^*} dx$$

$$V(Y|X^*, Z) = \int v(X, Z) f_{x|x^*} dx + Var[g(X, Z)|X^*]$$

This can be done e.g. for exponential g:

- Poisson regression model
- Parametric survival

# Bayes

Richardson and Green (2002)

- Evaluation by MCMC techniques
- Conditional independence assumptions on the three models parts as seen in the likelihood approach
- The latent variable  $X$  is treated an unknown parameter
- Different data types can be combined
- Prior distributions for the error model
- Flexible handling of the exposure model

## Corrected Score

Functional method,  $X$  fixed constants, no distribution assumptions  
Estimating (score ) equations

$$E(\psi_{true}(Y, X, \beta)) = 0$$
$$\sum_{i=1}^n \psi_{true}(Y_i, X_i, \beta) = 0$$

Find corrected score function  $\psi_{CS}$  with

$$E_u(\psi_{CS}(Y, X^*, \beta)) = \psi_{true}(Y, X, \beta)$$

## Corrected Score: Examples

$$\psi_{CS} = (Y - \beta_0 - \beta_1 X^*)(1 \quad X^*)^T + (0 \quad \sigma_u^2 \beta)^T$$

Corrected Score equations are available for polynomial Regression, Poisson and for GLMs misclassification

General simulation method by Stefanski

## Case study : Occupational Dust and chronic bronchitis

Kü/Carroll (1997) and Gössl /Kü(2001)

**Research question:** Relationship between occupational dust and chronic bronchitis

Data form N=1246 workers:

$X$ :  $\log(1+\text{average occupational dust exposure})$

$Y$ : Chronic bronchitis (CBR)

$X^*$ : Measurements and expert ratings

$Z_1$ : Smoking

$Z_2$ : Duration of exposure

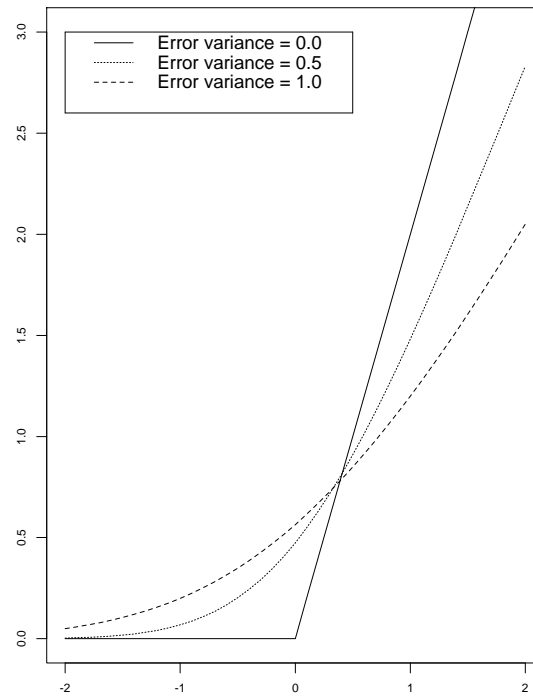
No validation or replication data available!

## The Model

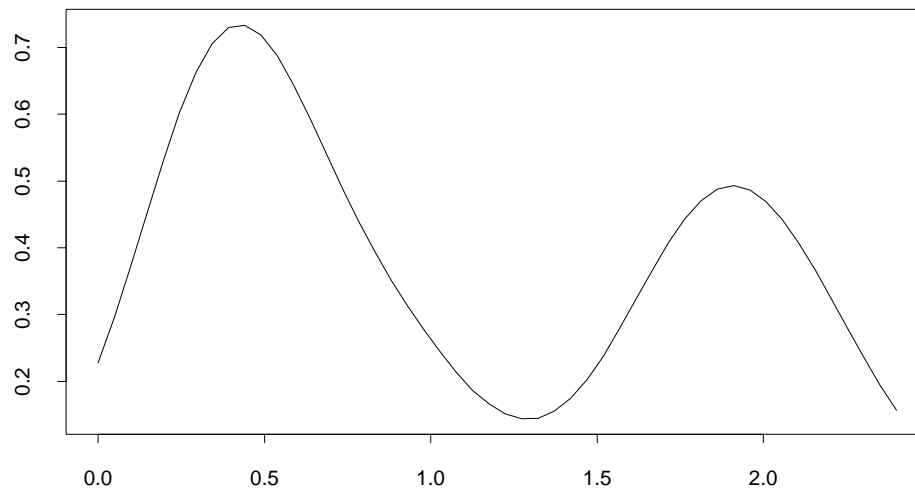
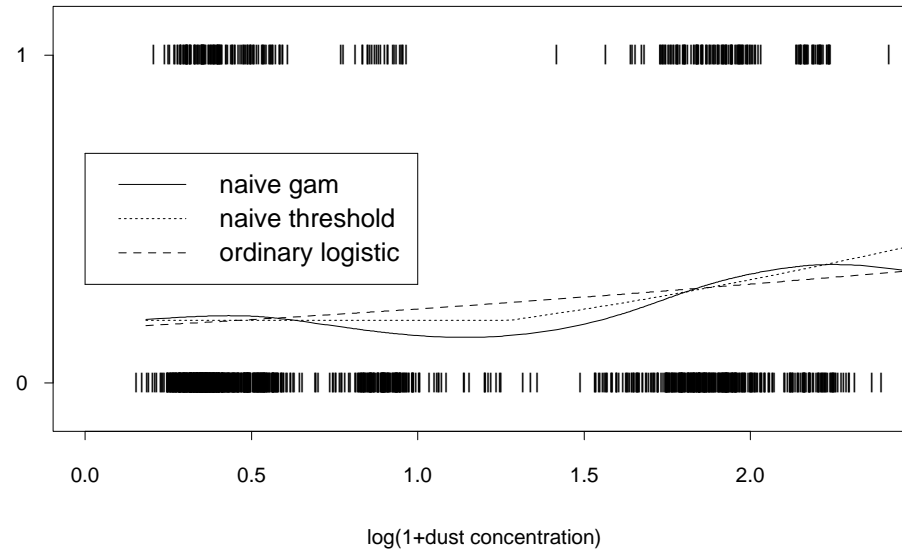
Segmented logistic regression an unknown threshold limiting value (TLV)  $\tau$

$$P(Y = 1|X = x, Z = z) = G(z'\beta_{k-} + \beta_k(x - \tau)_+),$$
$$(x - \tau)_+ = \max(0, x - \tau).$$

# Effect of measurement error



### Munich Data



# Likelihood

- Probit approximation
- Calculation of the integrals
- Assumption of a mixture of two normals for the exposure model
- Fixed additive measurement error

# Regression calibration

- Assumption of a mixture of two normals for the exposure model
- Fixed additive measurement error

## Results

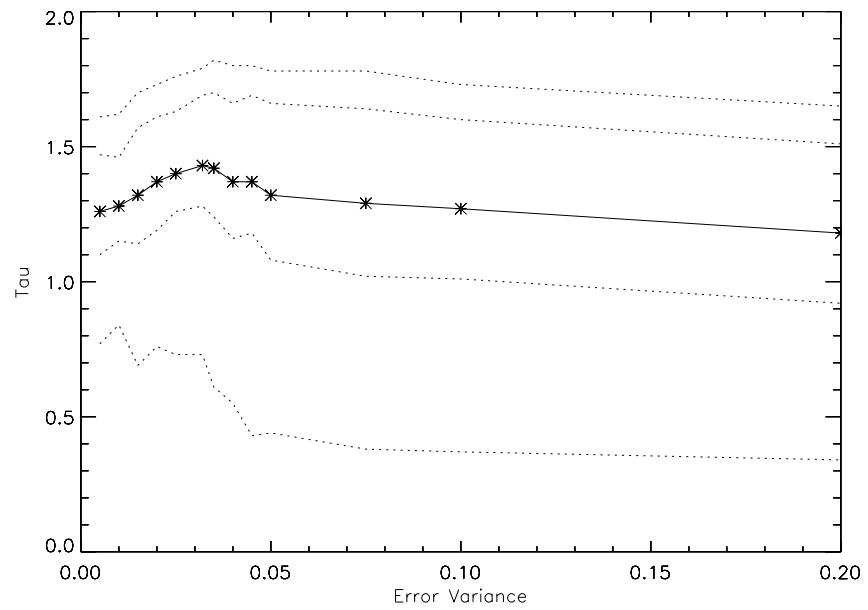
Method	TLV- $\tau_0$	Nom s. e.	boot s.e.
Naive	1.27	.41	.24
Pseudo-MLE	1.76	.17	.21
Regression Calibration	1.75	.12	.19

Table 1: *Estimated TLV in the Munich data, when  $\sigma_u^2 = 0.035$ .*

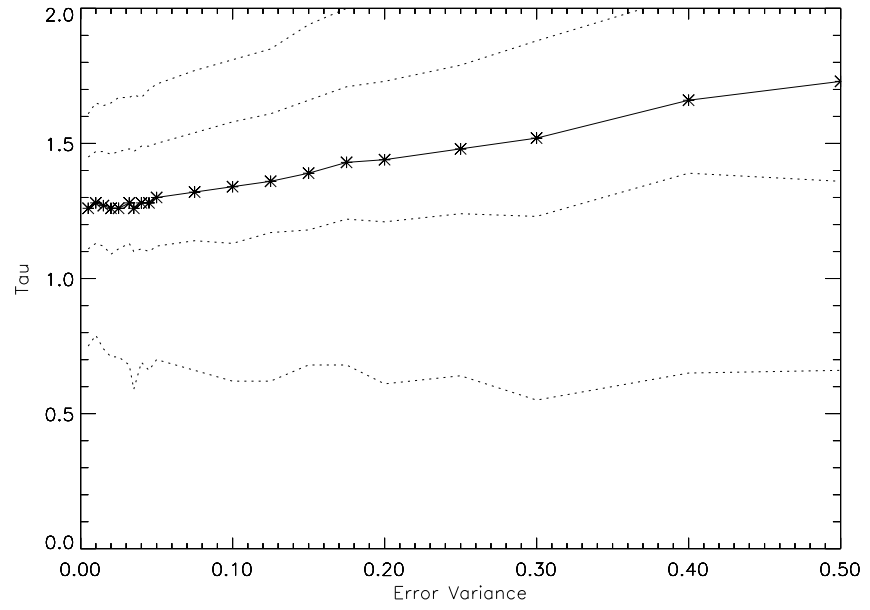
# Bayes

- Fixed additive measurement error: Sensitivity analysis
- flat priori for measurement error :No convergence
- Assumption of mixture of normals for exposure model
- Both models Berkson and additive

# Results: Estimation of TLV



class.



Berkson

# Conclusions

1. Measurement model essential: High Difference between Effect of Berkson and classical measurement error in most cases!
2. Additive classical non differential measurement error leads to attenuation
3. Many methods available
4. Regression calibration works in many cases
5. ML should be taken into account for Berkson error
6. Bayesian analysis is useful especially if model structure is complex

# MISCLASSIFICATION AND MEASUREMENT ERROR IN REGRESSION MODELS

## Part 4

Helmut Küchenhoff  
Statistical Consulting Unit  
Ludwig-Maximilians-Universität München

Padova  
2./3.10.2007

## 2. GENERAL SIMEX IDEA

### Linear regression with additive measurement error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (i = 1, \dots, n)$$

$$\text{Var}(X_i) = \sigma_x^2 \ \& \ \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

$$X_i^* = X_i + \sigma U_i \quad \text{with } (U_i, X_i, \varepsilon_i) \text{ independent}$$

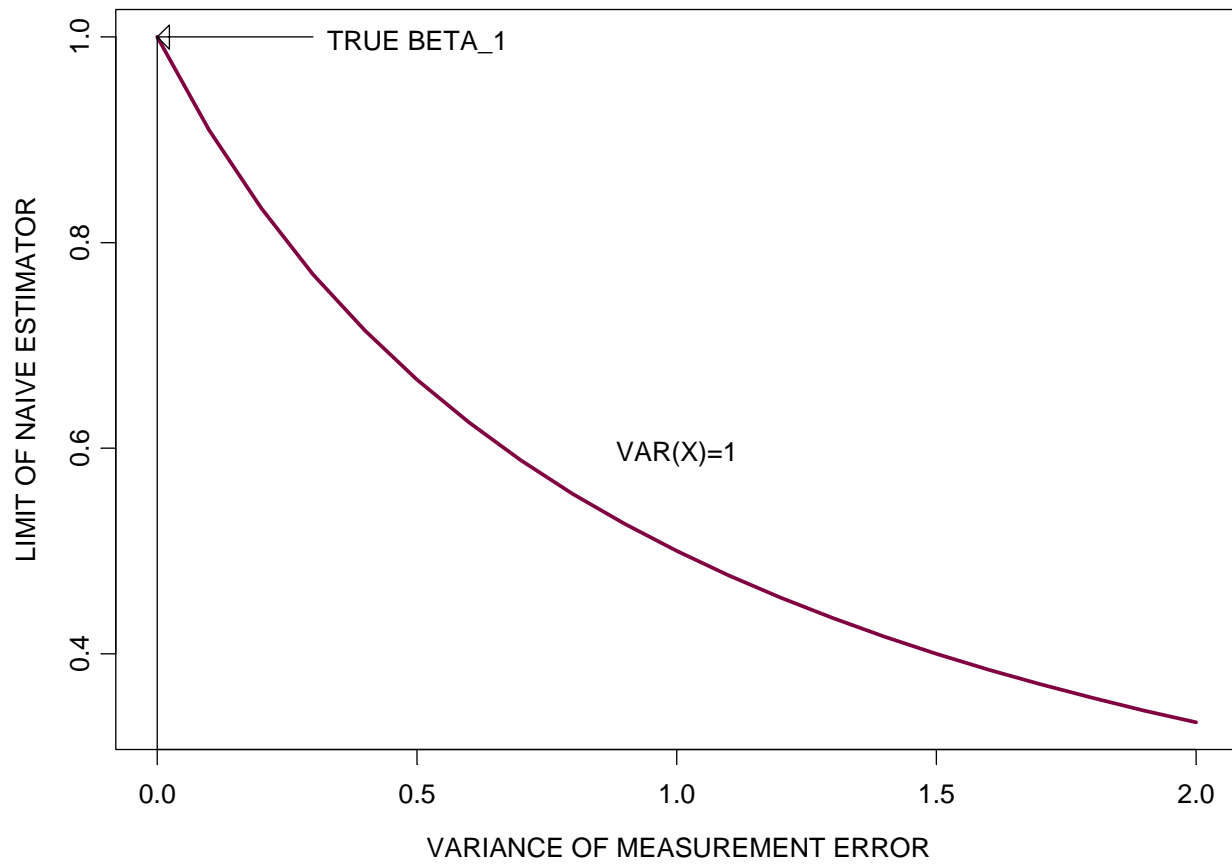
$$U_i \sim N(0, 1)$$

- Ignoring measurement error ( $U_i$ )  $\Rightarrow$  **naïve estimation** in  $Y_i = \beta_0^* + \beta_1^* X_i^* + \varepsilon_i^*$

$$\beta_1^* = \text{plim} \hat{\beta}_{\text{naive}} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma^2} \beta_1$$

$\Rightarrow$  Attenuation increases with measurement error variance

# LINEAR REGRESSION



## SIMEX idea (Cook & Stefanski, 1994)

- Assume
  - $\sigma$  is **known**
  - Observe  $(Y_i, X_i^*, Z_i)_{i=1}^n$  instead of  $(Y_i, X_i, Z_i)_{i=1}^n$
- **SIM**ulation step: generate more measurement error + calculate naïve estimators
  1. **Simulate pseudo-data**  $X_{b,i}^*(\lambda) = X_i^* + \sqrt{\lambda} \sigma U_{b,i}$  for a fixed grid  $\lambda_0 (\equiv 0), \lambda_1, \lambda_2, \dots, \lambda_m$   
 $\Rightarrow \text{Var}(X_{b,i}^*(\lambda)) = \sigma_X^2 + (1 + \lambda)\sigma^2$
  2. **Do this B times** ( $b=1, \dots, B$ )
  3. **Calculate mean:**  $\hat{\beta}(\lambda_k) = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{\text{NAIVE}} \left[ (Y_i, X_{b,i}^*(\lambda_k), Z_i)_{i=1}^n \right]$

- **EX**trapolation step: extrapolate back to  $\lambda = -1$  to estimate  $\beta$ 
  1. **Fit parametrically** relation  $(\lambda_k, \hat{\beta}(\lambda_k))$  ( $k = 0, \dots, m$ )
  2. **Find**  $\hat{\beta}_{\text{SIMEX}} \equiv \hat{\beta}(-1)$  for all regression coefficients

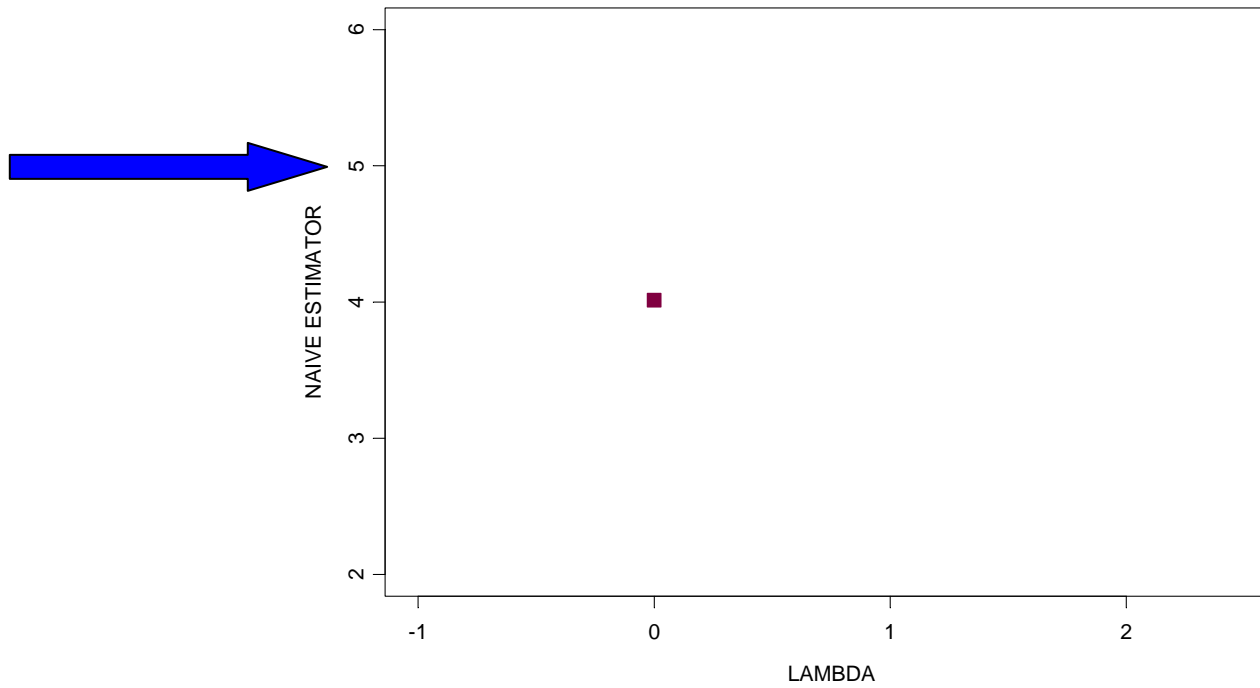
**Example**

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \dots, 200)$$

$$X_i \sim N(0, 2^2) \text{ \& } \varepsilon_i \sim N(0, 1)$$

$$Y_i = \beta_0^* + \beta_1^*(X_i + U_i) + \varepsilon_i$$

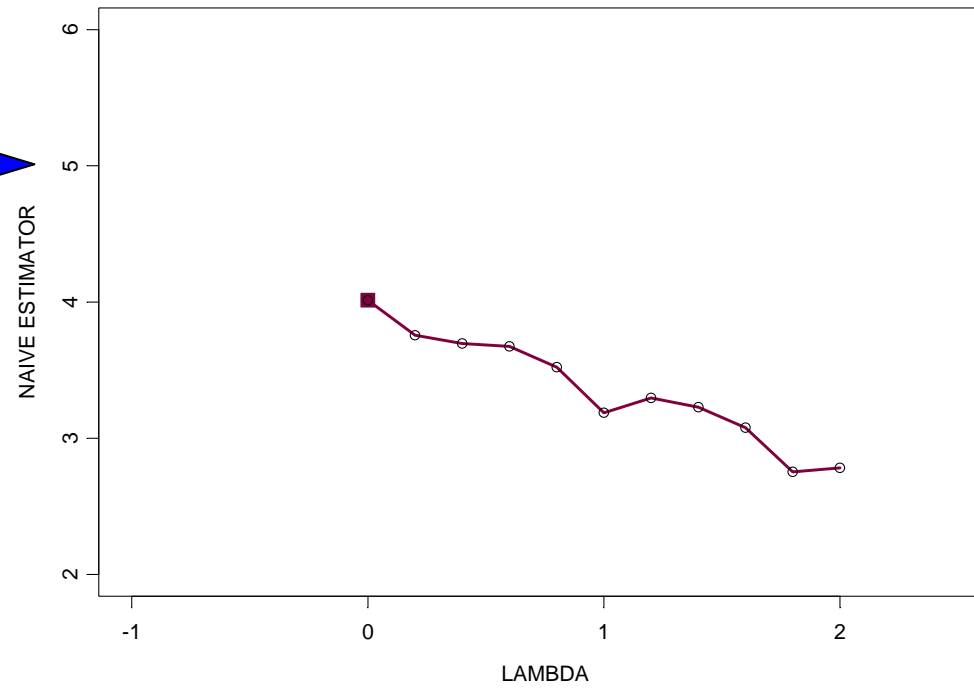
$$U_i \sim N(0, 1)$$



**Example**

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \dots, 200) \quad X_i \sim N(0, 2^2) \text{ \& } \varepsilon_i \sim N(0, 1)$$

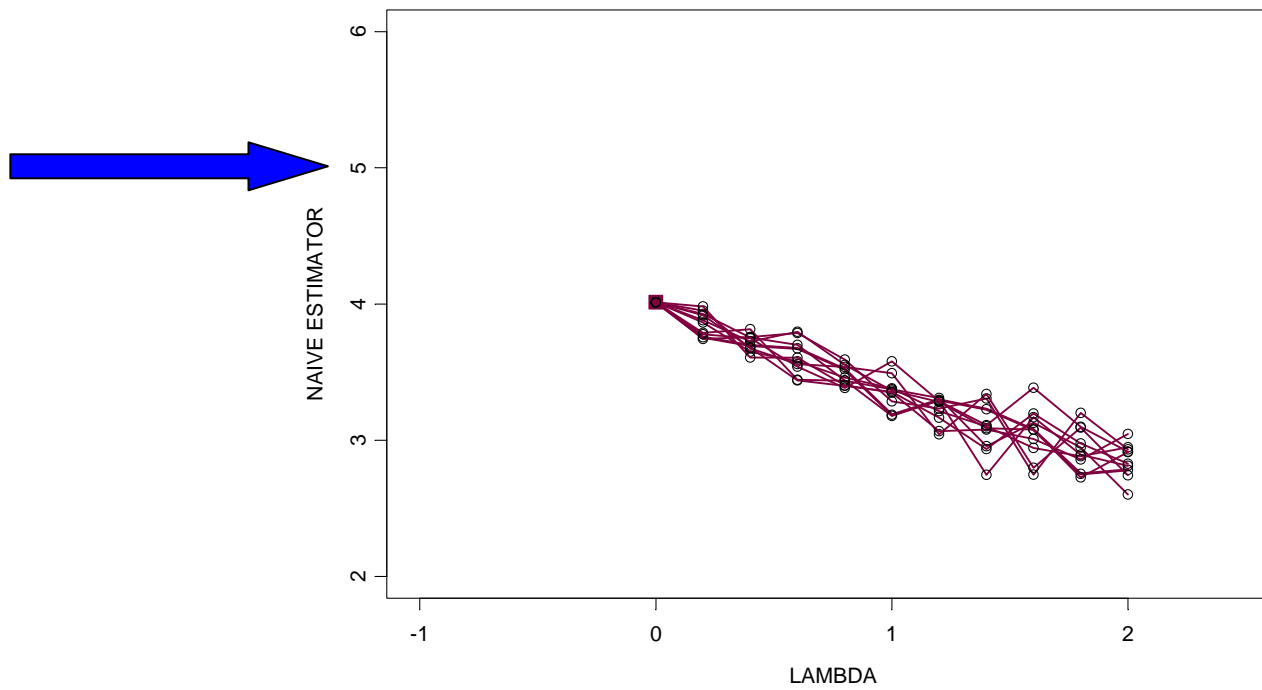
$$Y_i = \beta_0^* + \beta_1^*(X_i + U_i) + \varepsilon_i \quad U_i \sim N(0, 1)$$



**Example**

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \dots, 200) \quad X_i \sim N(0, 2^2) \text{ \& } \varepsilon_i \sim N(0, 1)$$

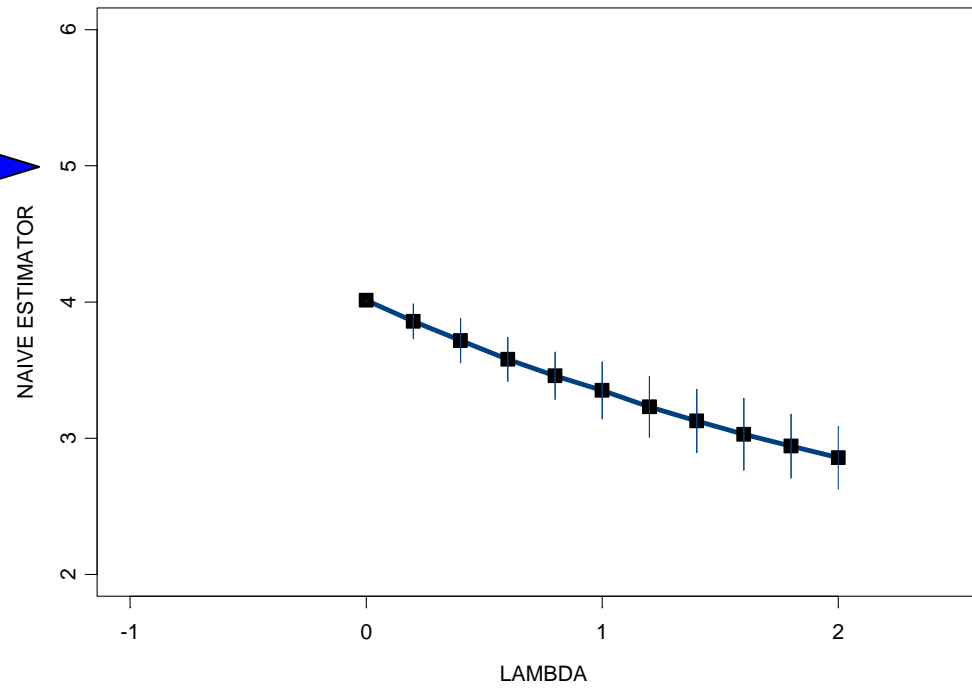
$$Y_i = \beta_0^* + \beta_1^*(X_i + U_i) + \varepsilon_i \quad U_i \sim N(0, 1)$$



**Example**

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \dots, 200) \quad X_i \sim N(0, 2^2) \text{ \& } \varepsilon_i \sim N(0, 1)$$

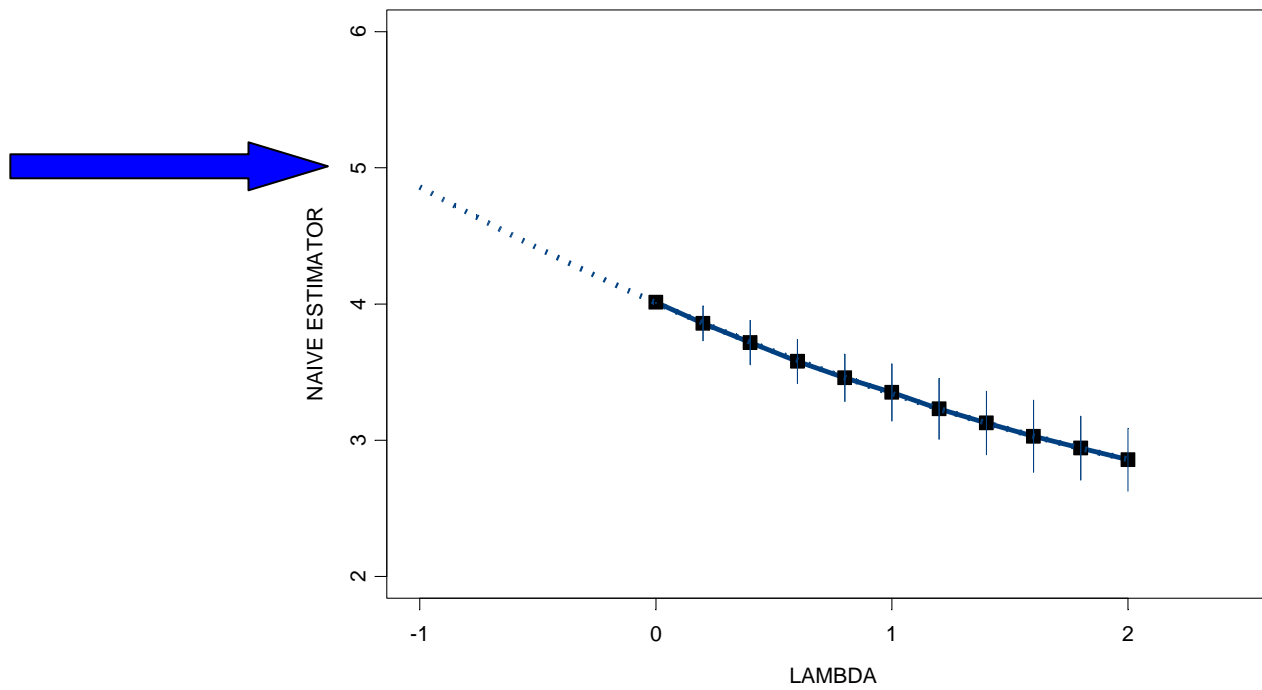
$$Y_i = \beta_0^* + \beta_1^*(X_i + U_i) + \varepsilon_i \quad U_i \sim N(0, 1)$$



**Example**

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \dots, 200) \quad X_i \sim N(0, 2^2) \text{ \& } \varepsilon_i \sim N(0, 1)$$

$$Y_i = \beta_0^* + \beta_1^*(X_i + U_i) + \varepsilon_i \quad U_i \sim N(0, 1)$$



**Example**

$$Y_i = 1 + 5X_i + \varepsilon_i \quad (i = 1, \dots, 200) \quad X_i \sim N(0, 2^2) \text{ \& } \varepsilon_i \sim N(0, 1)$$
$$Y_i = \beta_0^* + \beta_1^*(X_i + U_i) + \varepsilon_i \quad U_i \sim N(0, 1)$$

Average of extrapolated estimate =  $\hat{\beta}_{1, \text{SIMEX}} = 4.86$

## Extrapolation functions

Linear :  $g(\lambda) = \gamma_0 + \gamma_1 \lambda$

Quadratic :  $g(\lambda) = \gamma_0 + \gamma_1 \lambda + \gamma_2 * \lambda^2$

Nonlinear :  $g(\lambda) = \gamma_1 + \frac{\gamma_2}{\gamma_3 * \lambda}$

- Nonlinear is motivated by linear regression
- Quadratic works fine in many examples
- Motivation by Taylor Series expansions

## Variance estimation

- Delta method (Carroll et al.(1996))
- For known error variance the variance can be also be estimated by extrapolation (Stefanski and Cook (1995))
- Bootstrap (computer intensive)

## Case study : Occupational Dust and chronic bronchitis

Kü/Carroll (1997) and Gössl /Kü(2001)

**Research question:** Relationship between occupational dust and chronic bronchitis

Data form N=1246 workers:

$X$ :  $\log(1+\text{average occupational dust exposure})$

$Y$ : Chronic bronchitis (CBR)

$X^*$ : Measurements and expert ratings

$Z_1$ : Smoking

$Z_2$ : Duration of exposure

## Results for the TLV

Method	TLV- $\tau_0$	Nom s. e.	boot s.e.
Naive	1.27	.41	.24
Pseudo-MLE	1.76	.17	.21
Regression Calibration	1.75	.12	.19
simex: linear	1.37	.23	.23
simex: quadratic	1.40	.23	.34
simex: nonlinear	1.40	.23	.86

# Misclassification SIMEX

General Regression model with misclassification matrix  $\Pi$

$$\begin{aligned}\beta^*(\Pi) &:= p \lim \hat{\beta}_{naive} \\ \beta^*(I_{k \times k}) &= \beta\end{aligned}$$

Problem:  $\beta^*(\Pi)$  is a function of a matrix.

We define:

$$\lambda \rightarrow \beta^*(\Pi^\lambda)$$

$\Pi^\lambda$  is defined by  $\Pi^0 = I_{k \times k}$ ,  $\Pi^{n+1} = \Pi^n * \Pi$  for  $\lambda = 0, 1, 2, \dots$

$$\Pi^\lambda := E \Lambda^\lambda E^{-1}$$

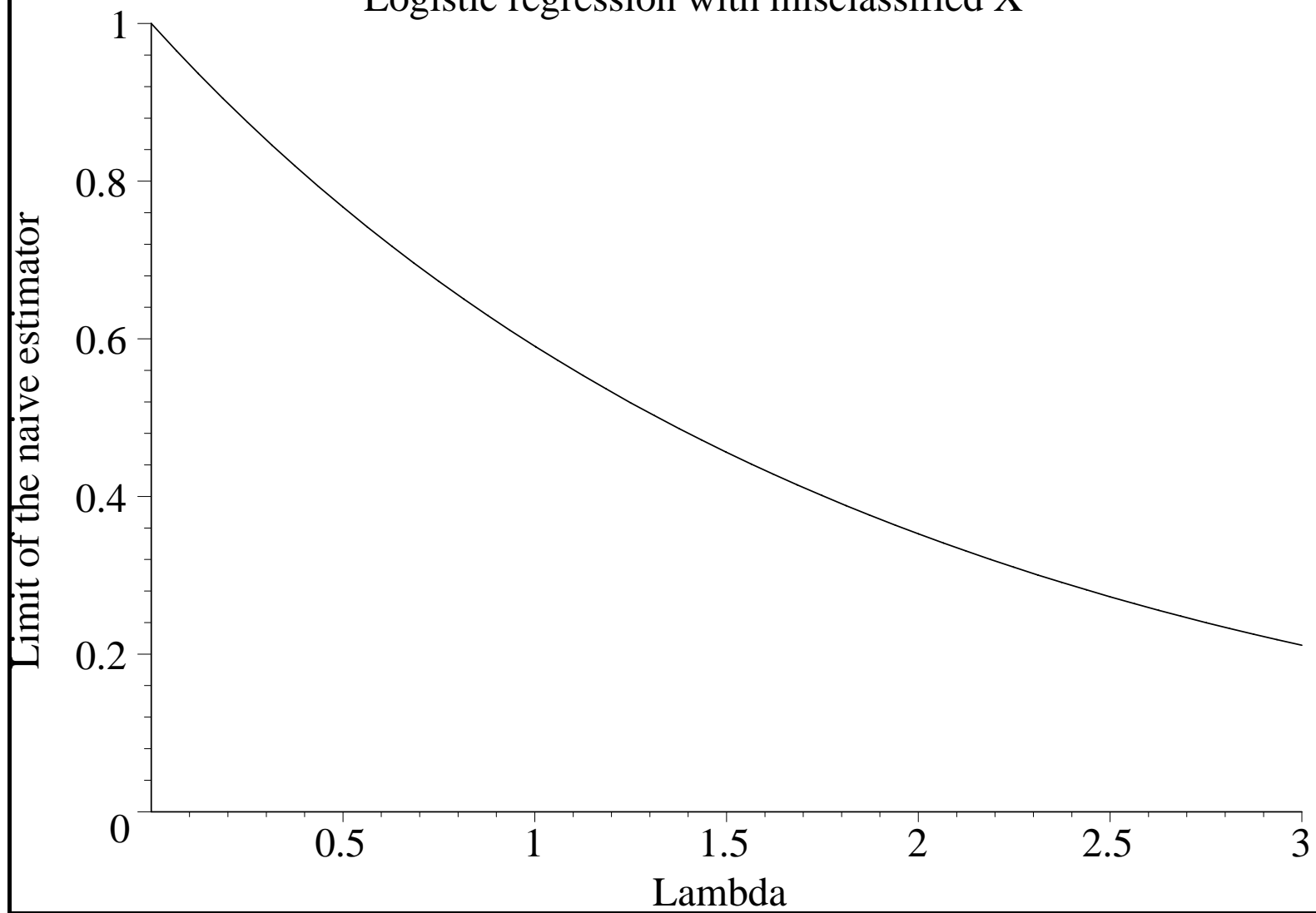
$$E := \text{Matrix of eigenvectors}$$

$$\Lambda := \text{Diagonal matrix of eigenvalues}$$

# Parameter Estimation in Relationship to the amount of misclassification

Logistic regression with misclassified  $X$  ( $\pi_{11} = \pi_{00} = 0.8$ )

# Logistic regression with misclassified X



## Properties of the function $\beta^*(\Pi^\lambda)$

- $\beta^*(\Pi^0) = \beta$
- differentiable

If  $X^*$  is misclassified in relation to  $X$  by MC-matrix  $\Pi$   
 $X^*(\lambda)$ , is misclassified in relation to  $X^*$  by MC-matrix  
 $\Pi^\lambda$ ,

$\Rightarrow$

$X^*(\lambda)$  is misclassified in relation to  $X$  by MC-matrix  
 $\Pi^{\lambda+1}$

# The MC-SIMEX Procedure

Data  $(Y_i, X_i^*, Z_i)_{i=1}^n$ ,

$X^*$  is observed instead of  $X$  with MC-matrix  $\Pi$

Naive estimator:  $\hat{\beta}_{naive}[(Y_i, X_i^*, Z_i)_{i=1}^n]$ .

## Simulation

For a fixed grid  $\lambda_1 \dots \lambda_m$   $B$  new pseudo data are generated by

$$X_{b,i}^*(\lambda_k) := MC[\Pi_k^\lambda](X_i^*), \quad i = 1, \dots, n; \quad b = 1, \dots, B;$$

There  $MC[M](X_i^*)$  is simulated from  $X_i^*$  using the misclassification matrix  $M$ .

$$\hat{\beta}(\lambda_k) := B^{-1} \sum_{b=1}^B \hat{\beta}_{na} [(Y_i, X_{b,i}^*(\lambda_k), Z_i)_{i=1}^n], \quad k = 1, \dots, m.$$

# Extrapolation

Parametric model:

$$\beta(\Pi^\lambda) = \mathcal{G}(\lambda, \Gamma) = \gamma_0 + \gamma_1 \lambda + \gamma_2 \lambda^2$$

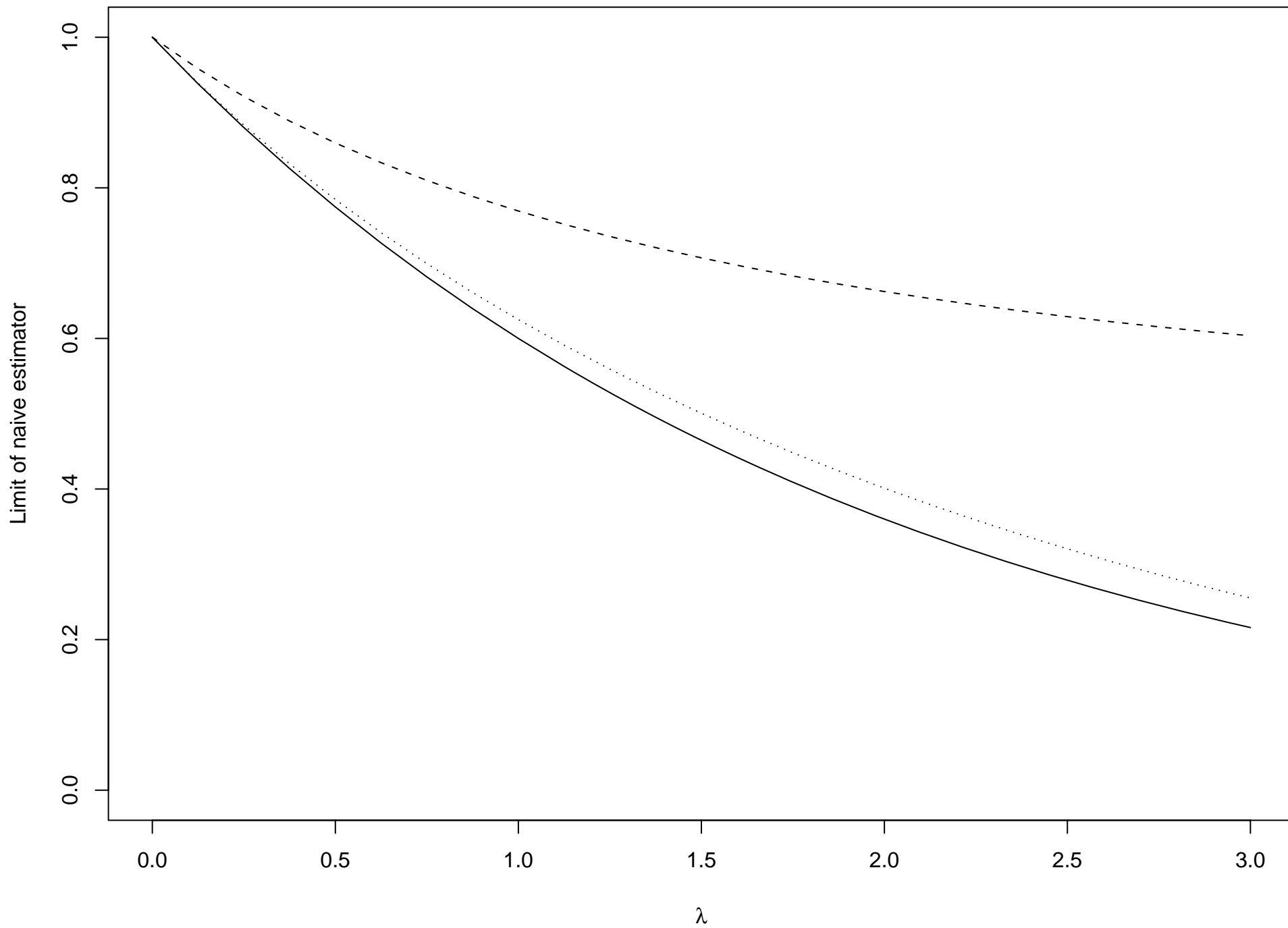
Fit by least squares from data  $[\lambda_k, \hat{\beta}(\lambda_k)]_{k=0}^m$ .

$$\hat{\beta}_{SIMEX} := \mathcal{G}(-1, \hat{\Gamma})$$

# Extrapolation function

Calculate true function  $\beta(\Pi)$  in several examples and simulation studies

- Funktion monotonic
- Quadratische Extrapolation suitable
- Loglinear Extrapolant



# Delta Method Variance Estimation

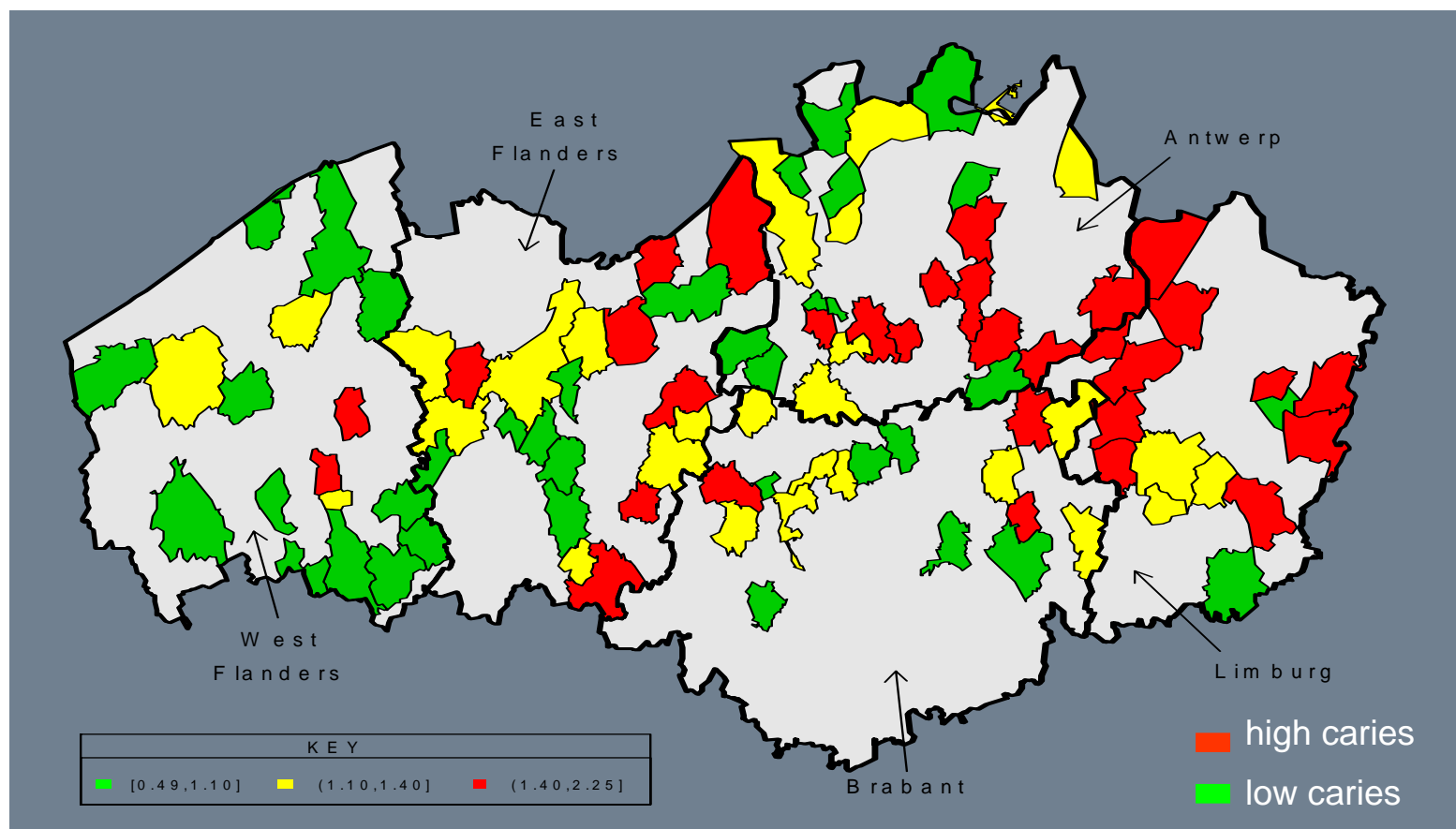
- All Estimators are solution of (biased) estimating equations
- Asymptotic expansions on averages of different estimating equations
- Extrapolation is a differentiable transformation
- Estimation of misclassification matrix can be included

## APPLICATION TO THE SIGNAL TANDMOBIEL STUDY<sup>®</sup>

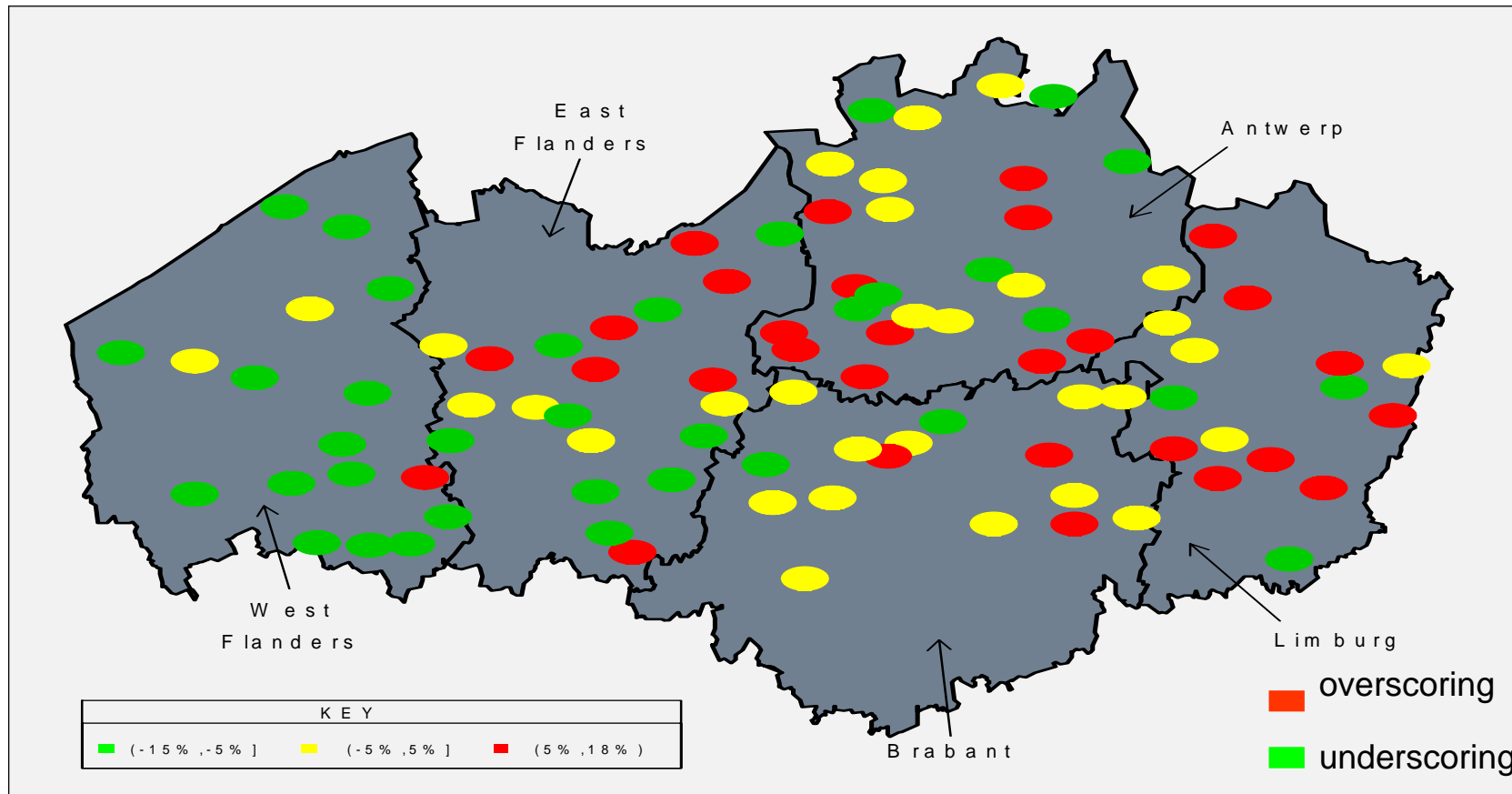
- Oral health study involving 4468 children in Flanders (Belgium)
- Children were examined annually by one of 16 dental examiners
- Binary response  $Y=1$  if tooth is decayed, filled or extracted due to caries
- GEE analysis for caries (combined response & individual teeth) on 4 first molars as a function of covariates
- **Questions:**
  - **East-West gradient in caries experience on the first 4 molars?**
  - **Does the trend remain the same in time?**

**But:** dental examiners showed high & different misclassification ⇔ benchmark scorer

**STS:** East-West trend in caries experience (1<sup>st</sup> year's cross-sectional results)



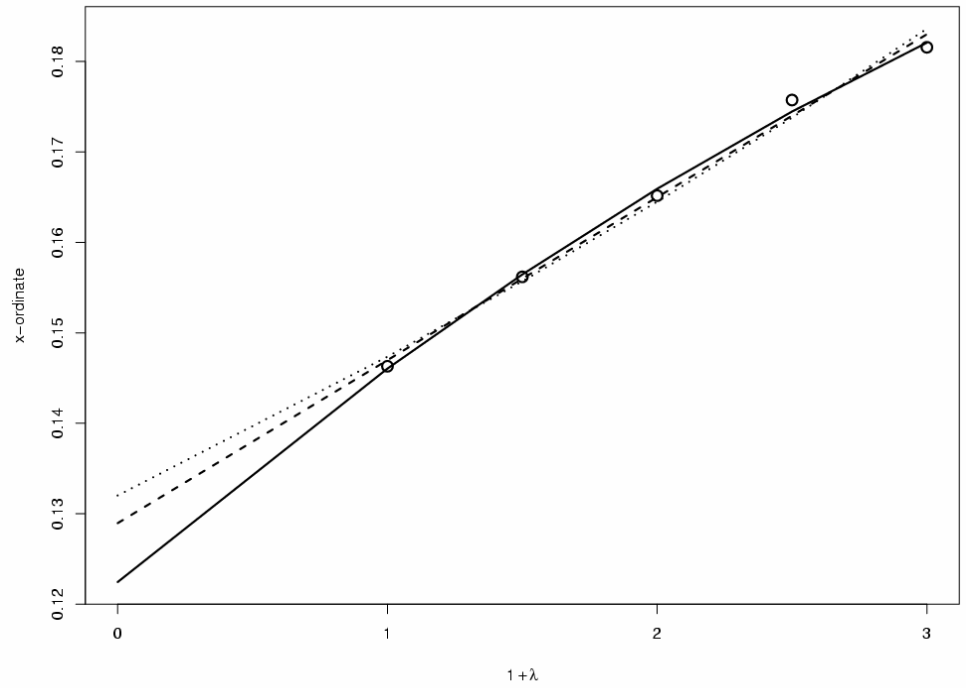
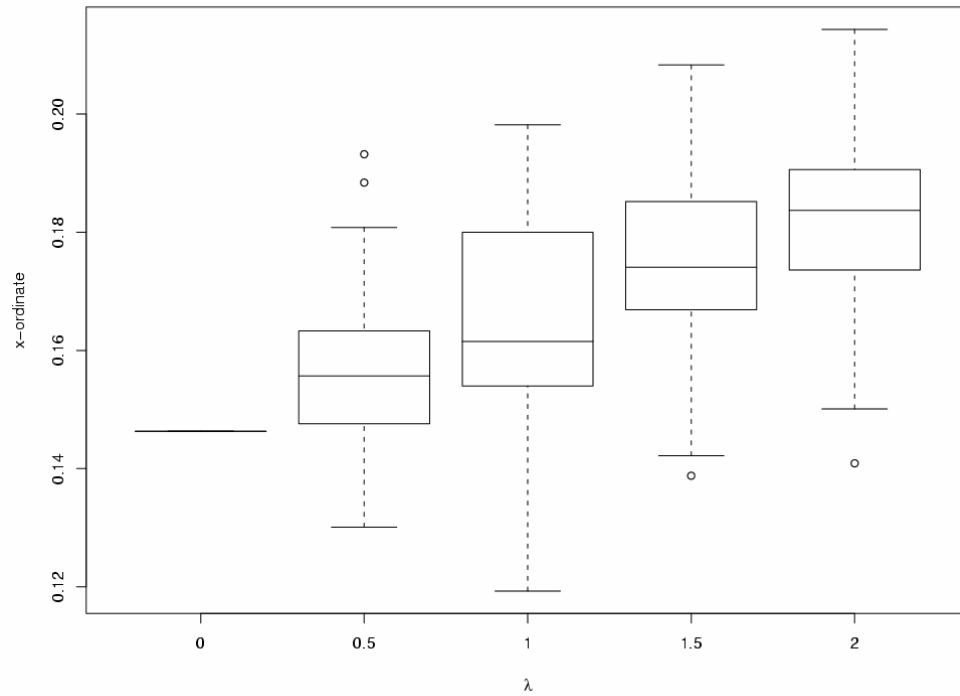
**STS:** Dental examiners are active in restricted geographical areas



⇒ East-West gradient?

## Results SIMEX approach (individual teeth)

- X-coordinate



## Remarks

- Existence of  $\Pi^\lambda$  for  $\lambda < 1$  has to be checked
- If  $\beta$  vector then use MC-SIMEX for every component
- The procedure also works for misclassified  $Y$  or more general cases
- $\hat{\beta}_{SIMEX}$  is consistent, if the extrapolating function is correctly specified.
- In general MC-SIMEX is approximately consistent, if  $\mathcal{G}(\lambda, \Gamma)$  is a good approximation of  $\beta^*(\Pi^\lambda)$ .

## Results SIMEX approach (individual teeth)

- East-West gradient confirmed
- East-West gradient increases over the years

# Software

- R-Package available (W. Lederer)
- Flexible statement for the main model
- Misclassification and additive measurement error
- Graphic display for the results

Lederer, Kü R-news (2006)

# Summary

- Very general computer intensive method
- Illustration of the effect of misclassification
- MC in  $X, Y$  or both, differential MC etc. can be handled
- Misclassification known or can be estimated by validation data