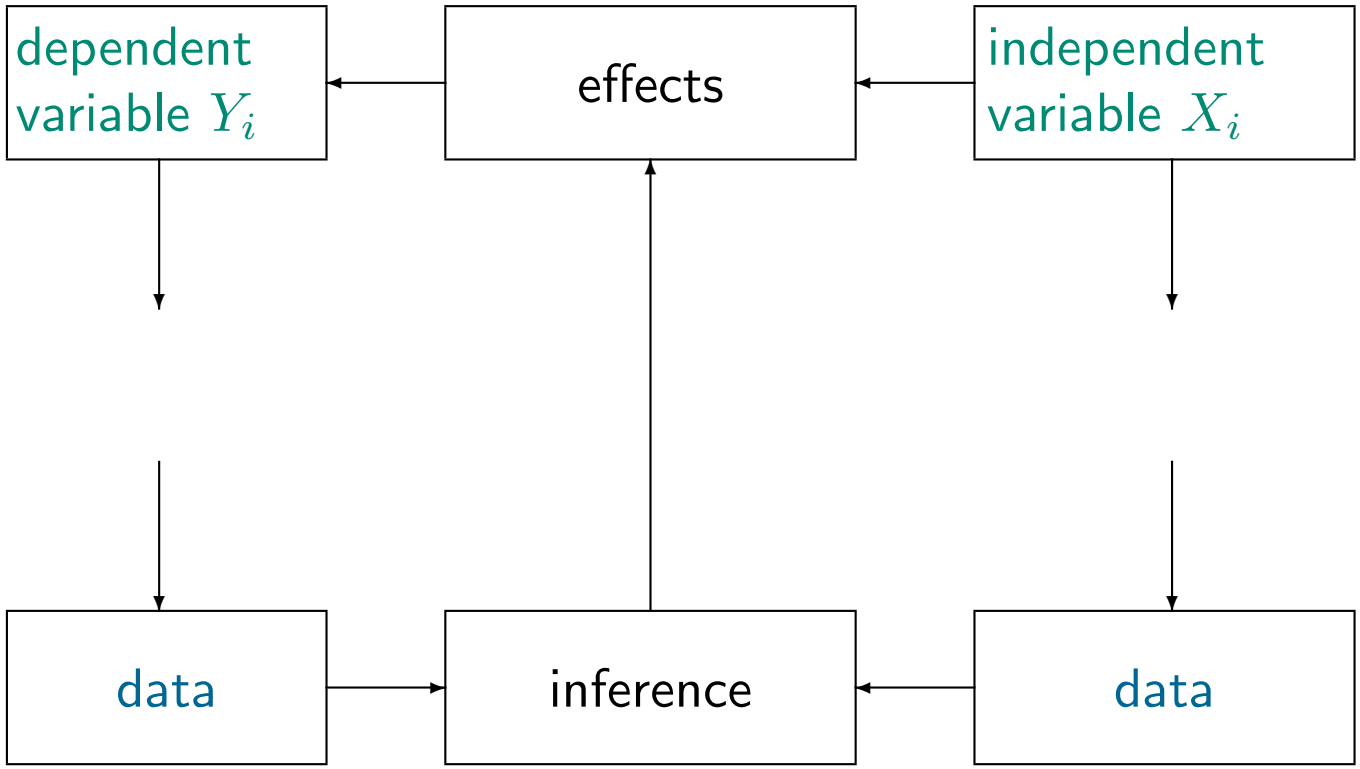
KTT und einige Überlegungen zu "Messfehlermodellen"

Thomas Augustin

Department of Statistics
Ludwig-Maximilians University Munich (LMU)

1. Background

Applied Statistics: Learning from data by sophisticated models
Complex relationships between variables



Often the relationship between variables and data is complex, too:

* Often variables of interest (gold standard) are not ascertainable.

* Only proxy variables (surrogates) are available instead.

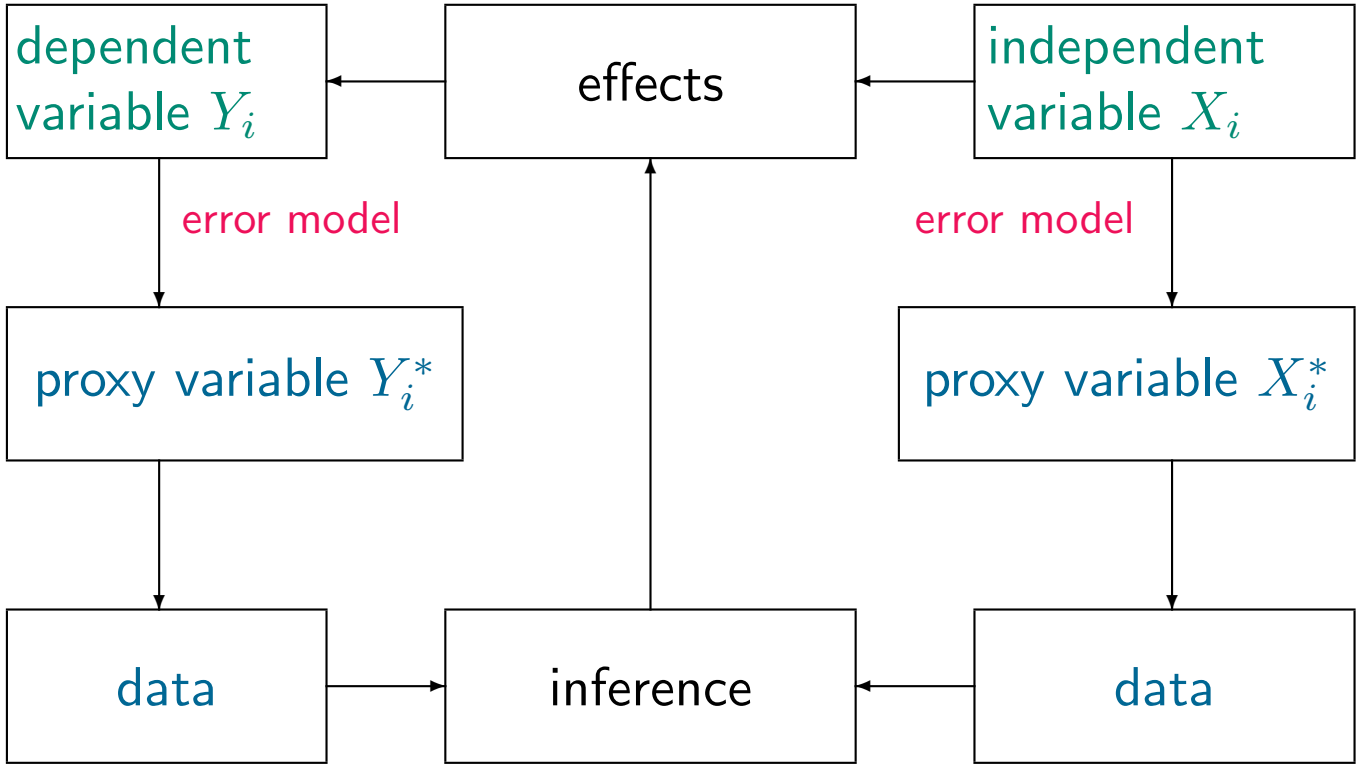
Typical examples: Measurement Error

- Error-prone measurements of true quantities

  * error in technical devices
  * indirect measurement
  * response effects (e.g. heaping)
  * use of aggregated quantities (JEM), averaged values, imputation, rough estimates etc.
  * anonymization of data by deliberate contamination
  * scores as estimates for abilities (CTT, IRT)

- Operationalization of complex constructs; latent variables/traits

  * long term quantities: permanent income,
  * importance of a patent
  * extent of motivation, degree of costumer satisfaction
  * severeness of undernutrition

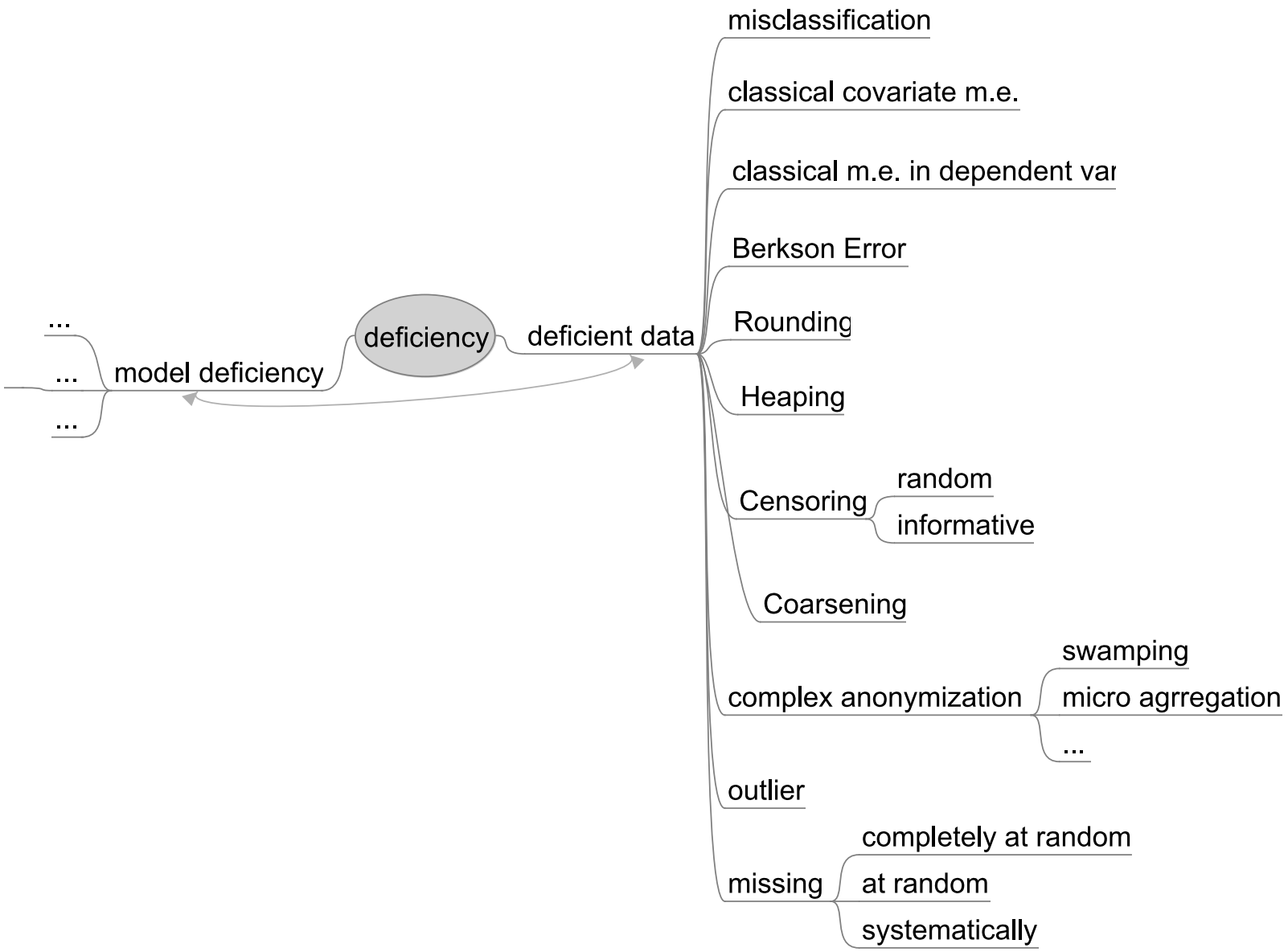

dependent variable Y_i
effects
independent variable X_i
error model
error model
proxy variable Y_i^*
proxy variable X_i^*
data
inference
data

Notation

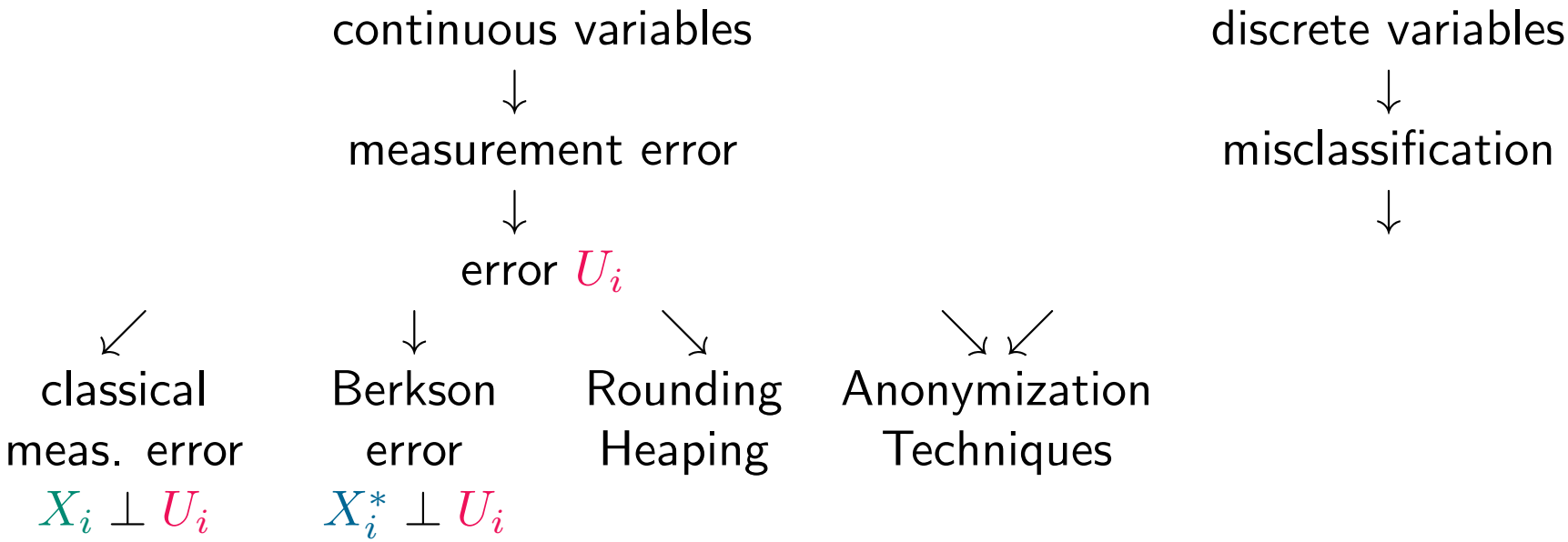
We have to distinguish between true (correctly measured) variable and its (possible incorrect) measurement, i.e. between the gold standard and the corresponding surrogate.

* - Notation (here)
X, Z : (unobservable) variable, gold standard
X^*, Z^*: corresponding possibly incorrect measurements analogously: Y, Y^* and T, T^*

Spinney of Deficiencies



Terminology



Typical Examples: Classical Measurement Error ($X_i \perp U_i$)

- error in measurement device, e.g., radon
- cheaper or faster measurement, e.g., surrogate marker
- latent variables, e.g., long term protein intake
- complex constructs, e.g., quality of life

Typical Examples: Berkson Error ($X_i^* \perp U_i$)

- Experimental design: X^* target value, X truly absorbed value
- Aggregated data: e.g., X^* mean exposure, X true individual exposure, JEM
- Often mixture of classical and Berkson error, e.g., radon studies

Additive m.e.: classical versus Berkson error

Note that, while in the classical additive case, due to $X^* \perp U$,

$$\begin{aligned} \mathbb{E}(U) = 0 \Rightarrow \mathbb{E}(X^*|X) &= X \\ \mathbb{V}(X^*) &= \mathbb{V}(X) + \mathbb{V}(U) \\ \mathbb{V}(X^*) &> \mathbb{V}(X), \end{aligned}$$

in the Berkson case, due to $X^* \perp U$,

$$\begin{aligned} \mathbb{E}(U) = 0 \Rightarrow \mathbb{E}(X|X^*) &= X^* \\ \mathbb{V}(X) &= \mathbb{V}(X^*) + \mathbb{V}(U) \\ \mathbb{V}(X) &> \mathbb{V}(X^*) \end{aligned}$$

Typical Examples: Rounding and Heaping

- abnormal concentration at certain "attractive numbers "
- Wright and Bray (2003, JRSS D): study in foetal medicine: ultrasound scan, measurement of fluid behind the neck (thickness of the nuchal translucency)
- duration data are commonly collected in a retrospective way:
  strong memory effects when time spans have to be remembered, e.g., Skinner & Humphreys (1999, Lifetime Data Analysis)
- Holt et al (1991, Biemer et al(eds.)): age at menarche
- heaping in episode / spell-based designs: Torelli & Trivellato (1993, J. Econometrics):
  concentration of values of unemployment duration at multiples of six ("identification problem": heaping versus effect of different levels of

compensation)
strong dependence of the bias on the DGP

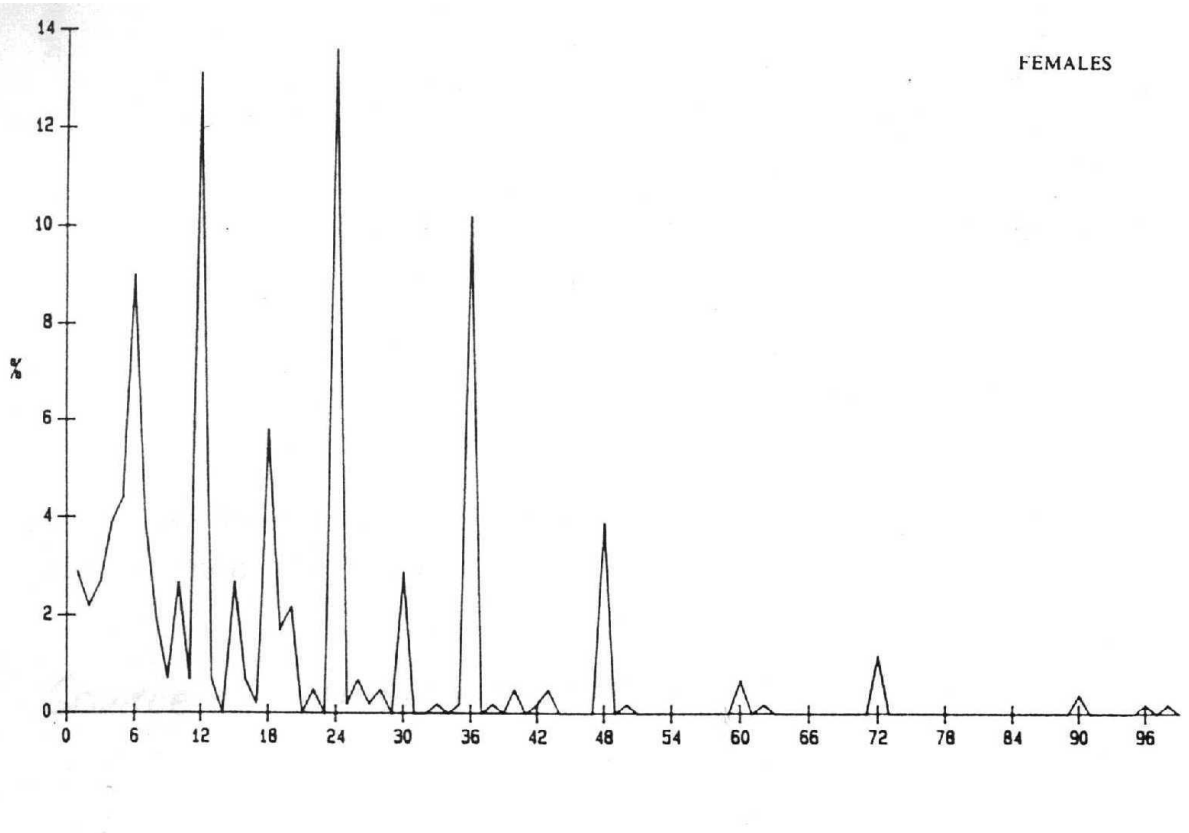


Fig. 2.1 Percentage distribution of unemployed individuals aged 14–29 years by reported unemployment duration (in months) at initial survey, Italian LFS, matched data for Lombardy, 1986.I–II: males, $N = 267$ and females, $N = 411$.

- heaping in calendar-based designs (German socio-economic panel SOEP)
  * distorted values for entry and leave of state of unemployment
  * bias analysis under simplified assumptions: Augustin & Wolff (2004, Stat.Papers)
  * simulation study (with data constellation based on the SOEP): Wolff & Augustin (2003, ASTA);Jürgens (2007, JRSS A)

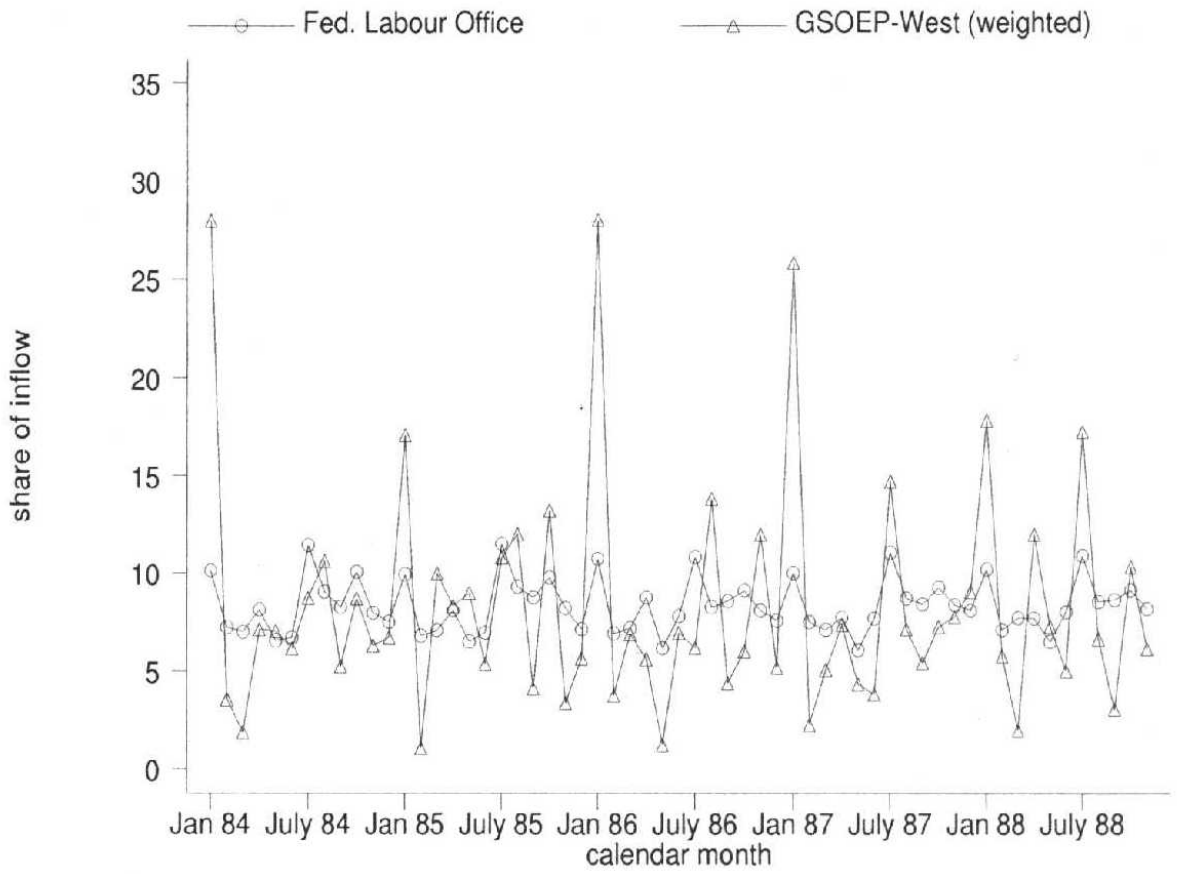
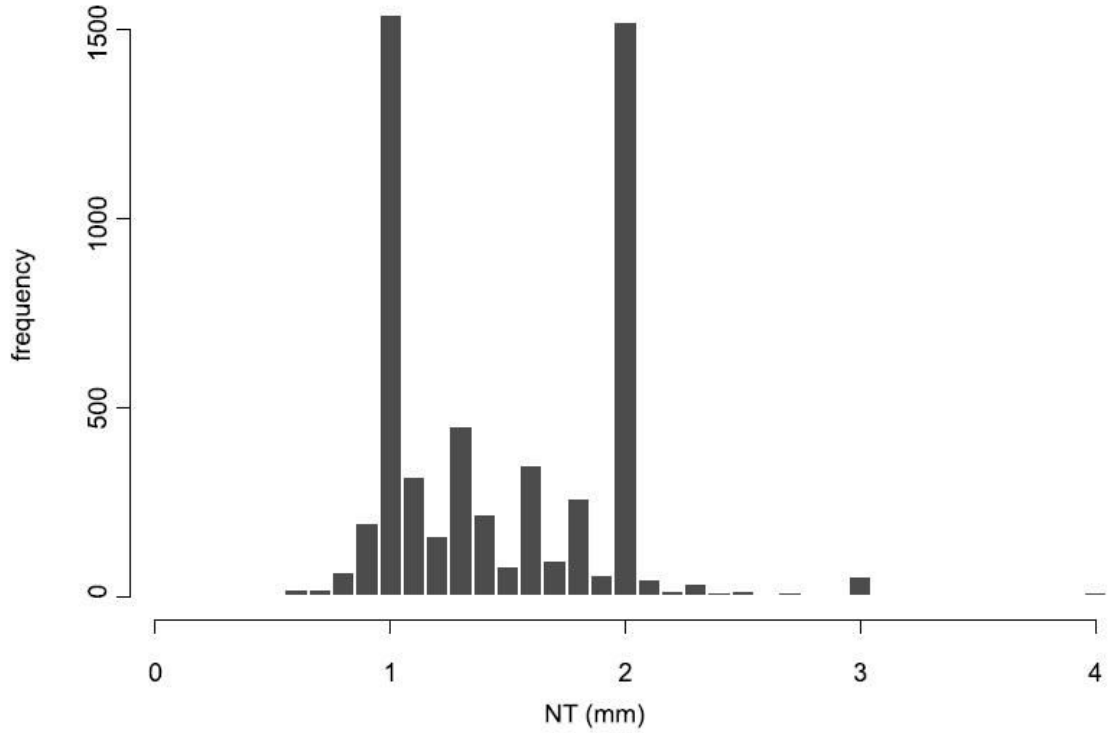


FIGURE 1. Proportion of the annual inflow into registered unemployment in each calendar month – Women, West-Germany.


frequency
0
500
1000
1500
0
1
2
3
4
NT (mm)

Typical Examples: Misclassification

- 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

Anonymisation Techniques

- Recent trend in official statistics: public use files (Statistisches Bundesamt, 2005, Statistik und Wissenschaft, Bd. 4) and big economic research institutes (IAB, Nuremberg)
- Error mechanism known!
- Distortion by classical measurement error or misclassification
- Often other techniques, e.g. micro-aggregation (Schmid, Schneeweiß, Küchenhoff (2007) Stat. Neerl.; Schmid, (2007, Diss. LMU))
- Growing importance in biometrics as well
- Discussion on "privacy preserving data mining"

The Fundamental Model of Classical Testing Theory (FMCTT)

$$\text{Measurement} = \text{True Value} + \text{Error}$$

$$X_i^*[j] = X_i[j] + U_i[j], \quad i = 1, \ldots, n, \; j = 1, \ldots, p$$

Assumptions on the distribution

$$\begin{aligned} \mathbb{E}(U_i[j]) &= 0 && \text{[A1.1]} \\ Var(U_i[j]) &= \sigma_j^2 && \text{[A1.2]} \\ U_i[j] &\sim N(0, \sigma_j^2) && \text{[A1.3]} \end{aligned}$$

Independence Assumptions "$\perp$" (Uncorrelatedness)

$$\begin{aligned} U_i[j] &\perp X_i[j] && \text{[A2.1]} \\ U_{i_1}[j] &\perp U_{i_2}[j] \quad i_1 \neq i_2 && \text{[A2.2]} \\ U_i[j_1] &\perp U_i[j_2] \quad j_1 \neq j_2 && \text{[A2.3]} \\ U_{i_1}[j_1] &\perp X_{i_2}[j_2] \quad i_1 \neq i_2; \; j_1 \neq j_2 && \text{[A2.4]} \end{aligned}$$

The meaningfullness of many well-known measures strongly depends on the FMCTT.

- reliability $Rel(X, X^*)$
- relationship between $Rel(X, X^*)$ and $Corr(X, X^*)$
- Splitt-Half-Reliability
- Spearman-Brown formula
- Cronbach's alpha

The triple whammy effect of measurement error in regression models

- bias
- masking of features
- loss of power

Note: Also the parameter estimates of exactly measured covariates (e.g. gender) may be affected.

Carroll, Ruppert, Stefanski, Crainiceanu (2006, Chap.H.)

- **classical error**: "attenuation "

Results

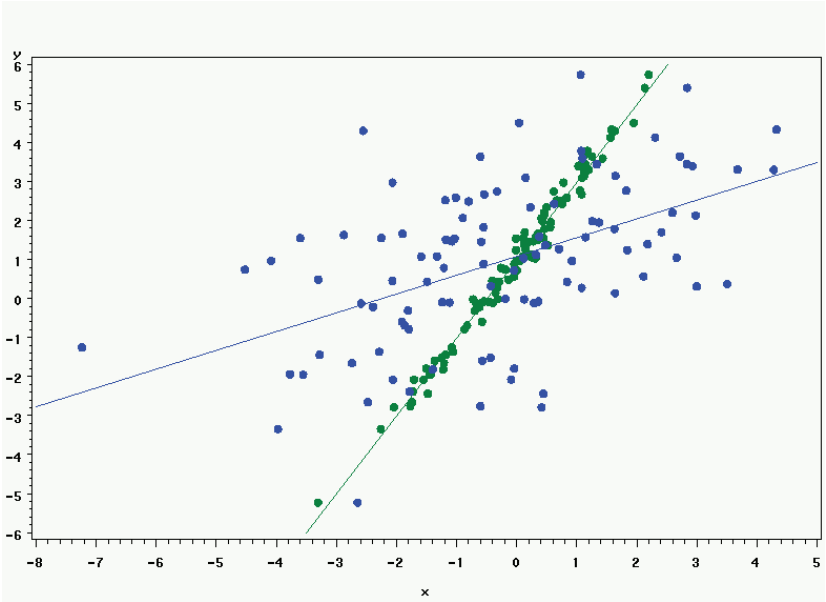


Figure 1: Effect of additive measurement error on linear regression

A first glance at the fundamental problem

Note that typically, even if

$$\mathbb{E}(X^*) = \mathbb{E}(X)$$

then

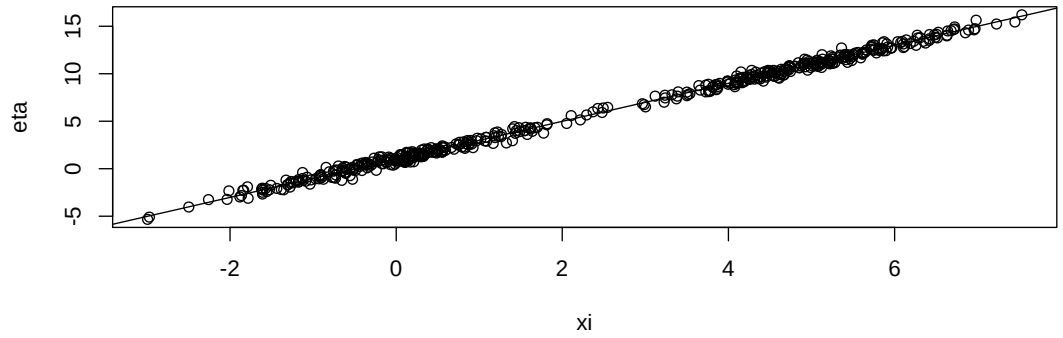
$$\mathbb{E}(X^*)^r) \neq \mathbb{E}(X^r), \quad r > 1.$$

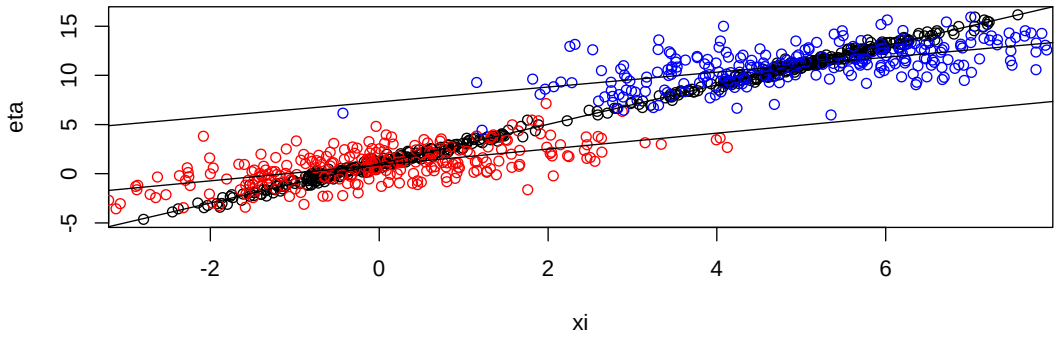
Potential consequences of neglecting the difference between latent variables and their surrogates:
severely biased estimation of all parameters

A cautionary note on the multiple case

$$Y = \beta_0 + \beta_X \cdot X + \beta_Z \cdot Z + \epsilon$$

Remember: also parameter estimates of error free variables may be affected.


eta
-5
0
5
10
15
-2
0
2
4
6
xi


eta
15
10
5
0
-5
-2
0
2
4
6
xi

- General methods

  + Regression Calibration: gives a first impression
  + SIMEX, allows handling of misclassification and measurement error

- Model-Specific Methods

  + Likelihood
  + Bayes
  + Corrected Score
  + Quasi-Likelihood

  + model-specific (programming effort versus gain in efficiency)
  + allows understanding of sources of the bias
  + promising for different extensions

- But note: almost no assumption is as wrong as assuming no error

- More work on more general error models is urgently needed:
  * flexible error distribution
  * covariate dependent measurement error
  * mixture of Berkson and classical error structure
  * heaping
  * anonymisation techniques

- correlation of measurement error with
  * true value
  * other covariates
  * dependent variable
  * m. e. of other variables

- Assumptions in the error model have a strong influence
  $\rightarrow$ assess carefully
  $\rightarrow$ perform sensitivity analysis (*sets* of models: imprecise probabilities partial identification)

Regression calibration — a universal first approach

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

1. Basic idea: estimate unknown X from the available data X^* and Z
2. Find a model for $E(X|X^*, Z)$ typically by validation or replication data
3. Replace the unobserved X by an estimate $E(X|X^*, Z)$ in the main model
4. Adjust variance estimates by bootstrap or asymptotic methods

- In general inconsistent estimator, but substantial bias reduction, except in highly nonlinear models
- Reasonable method in many practical situations
- Correction only concerns the data matrix, not the estimation routine itself: Standard software can be used for estimation!
- Calibration data can easily be incorporated
- Easy handling of nonlinear functions of the covariates

Differential and nondifferential measurement error

Assumption of nondifferential measurement error relates to the response:

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

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:

$$[Y, X^*, X] = [Y| \underbrace{X^*}_{\text{delete}}, X] \cdot [X^*|X] \cdot [X] = [Y|X][X^*|X][X]$$

Estimating Functions

- Crucial technical idea: Do not investigate estimators directly but the equations producing the estimators

$$\text{estimator} = \text{root}(\text{function}(\text{ObservedData}, \text{Parameter}))$$

$$\hat{\vartheta} = \text{root}(\psi(\mathbf{X}, \mathbf{Y}, \vartheta))$$

- Estimator is not systematically biased when
  * in the average this was the right decision,
  * i.e. when the true value is indeed the root of the expected value of the function, i.e.

$$\mathbb{E}_\vartheta(\psi(\mathbf{X}, \mathbf{Y}, \vartheta)) = 0$$

  Then $\psi(\cdot)$ is called **unbiased estimating function**.

- For the moment classical covariate measurement error only

$$X_i^* = X_i + U_i\,, \quad X_i \perp U_i\,.$$

- Note that typically, even if $\mathbb{E}(X^*) = \mathbb{E}(X)$ then $\mathbb{E}((X^*)^r) \neq \mathbb{E}(X^r), \quad r > 1.$

- Therefore *naive estimation* by simply replacing $\mathbf{X}$ with $\mathbf{X}^*$, leads in general to

$$\left|\mathbb{E}_{p_\vartheta}\left(\psi^{\mathbf{X}}\left(\mathbf{Y};\mathbf{X}^*;\vartheta\right)\right)\right| \geq a > 0\,,$$

resulting in inconsistent estimators. For instance,

$$\mathbb{E}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 \cdot X_i^*)\begin{pmatrix}1\\ X_i{}^*\end{pmatrix}\right) \neq \mathbb{E}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 \cdot X_i)\begin{pmatrix}1\\ X_i\end{pmatrix}\right) = 0$$

- Measurement error correction: Find an estimating function $\psi^{X^*}(\mathbf{Y}, \mathbf{X}^*, \vartheta)$ in the error prone data with

$$\mathbb{E}_{p_\vartheta}\psi^{X^*}(\mathbf{Y};\mathbf{X}^*;\vartheta) = \mathbf{0}.$$

$$
\begin{aligned}
\mathbb{E}\left(\Psi_{naive}\right) &= \mathbb{E}\left(\Psi_{ideal} + \underbrace{\left(\Psi_{naive} - \Psi_{ideal}\right)}_{\text{"}Rest\text{"}}\right) \\
&= \mathbb{E}\left(\Psi_{ideal} + Rest\right) \\
&= \underbrace{\mathbb{E}\left(\Psi_{ideal}\right)}_{0} + \mathbb{E}\left(Rest\right)
\end{aligned}
$$

$$
\mathbb{E}\left(\Psi_{naive}\right) - \mathbb{E}\left(Rest\right) = 0
$$

$$
\underbrace{\mathbb{E}\left(\Psi_{naive} - \left(\mathbb{E}\left(Rest\right)\right)\right)}_{\text{unbiased estimating function}} = 0
$$

- often:

$$\Psi_{ideal} = \sum_{i=1}^{n} \Psi_{i,ideal}$$

$$\Psi_{naive} = \sum_{i=1}^{n} \Psi_{i,naive}$$

$$Rest = \sum_{i=1}^{n} Rest_i$$

working on the "i-level" $\Rightarrow$ handling of heteroscedastic measurement error directly possible

- More generally, find $g(\cdot)$ such that

$$\mathbb{E}(\underbrace{\mathbb{E}(g(\Psi_{naive})|\text{ideal data})}) = \mathbb{E}(g(\Psi_{naive})) = 0 = \mathbb{E}(\underbrace{\Psi_{ideal}})$$

- $g(\cdot)$ with

$$\mathbb{E}g(\Psi_{naive}|\text{ideal data}) = \Psi_{ideal}$$

corrected score function

(Stefanski (1989, Comm.Stat.Theory.Meth.),
Nakamura (1990, Biometrika)).

References:

- T. Augustin aand H. Küchenhoff (2008): Measurement Error and Misclassification in Regression: Basics and some Recent Developments. Tutorial at Lifestat 2008: Statistics and Life Sciences: Perspectives and Challenges — 54. Biometrisches Kolloquium (mirrored at: www.stat.uni-muenchen.de/∼ helmut/Texte/Lifestat_TA_HK.pdf)

- Carroll, RJ., Ruppert, D., Stefanski, LA and Crainiceanu, CM. (2006). Measurement Error in Nonlinear Models. A Modern Perspective. Chapman & Hall, Boca Raton. 2nd edition.

- Gustafson, P. (2004). Measurement Error and Misclassification in Statistics and Epidemiology. Impacts and Bayesian Correction, CRC Press, Boca Raton.