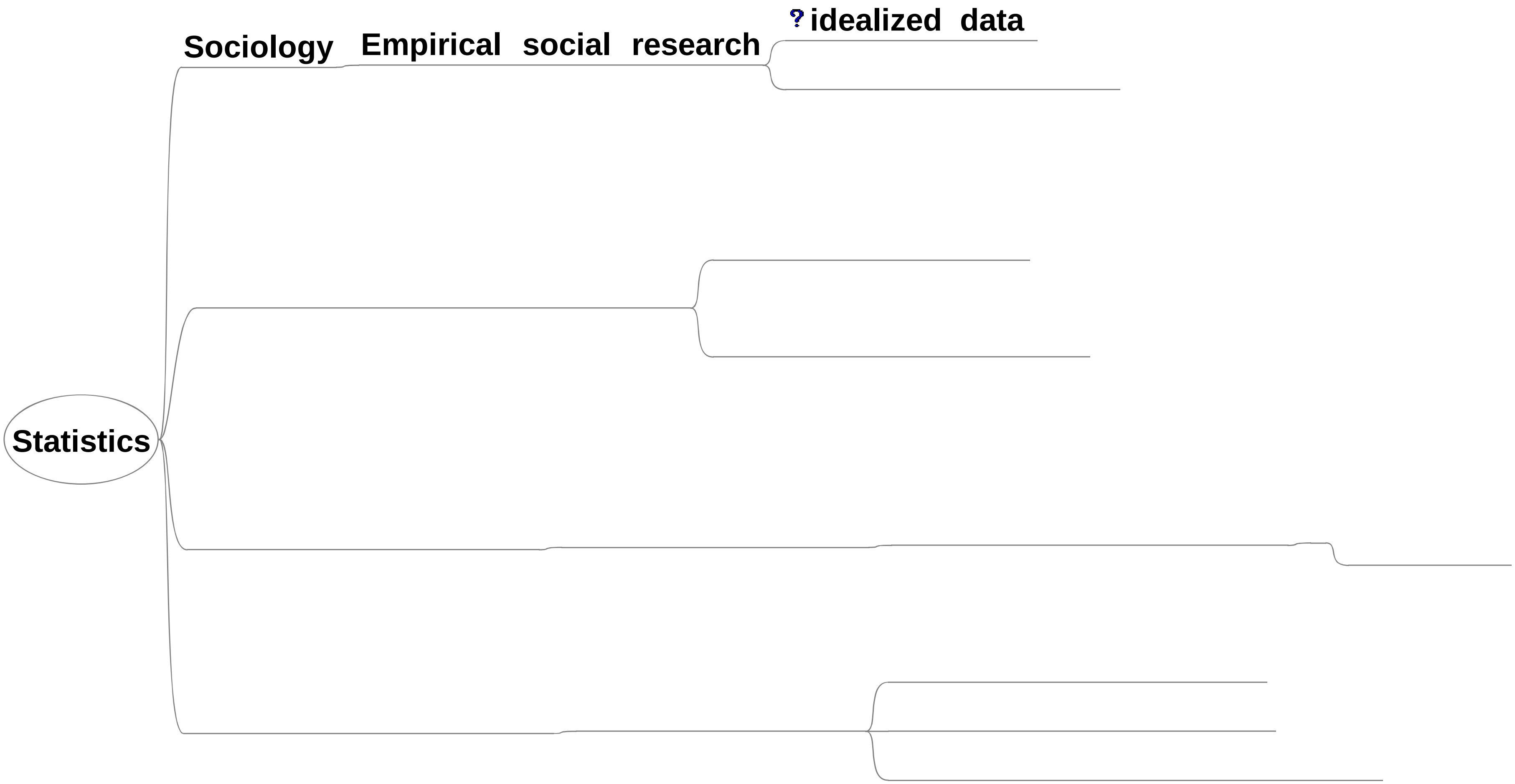
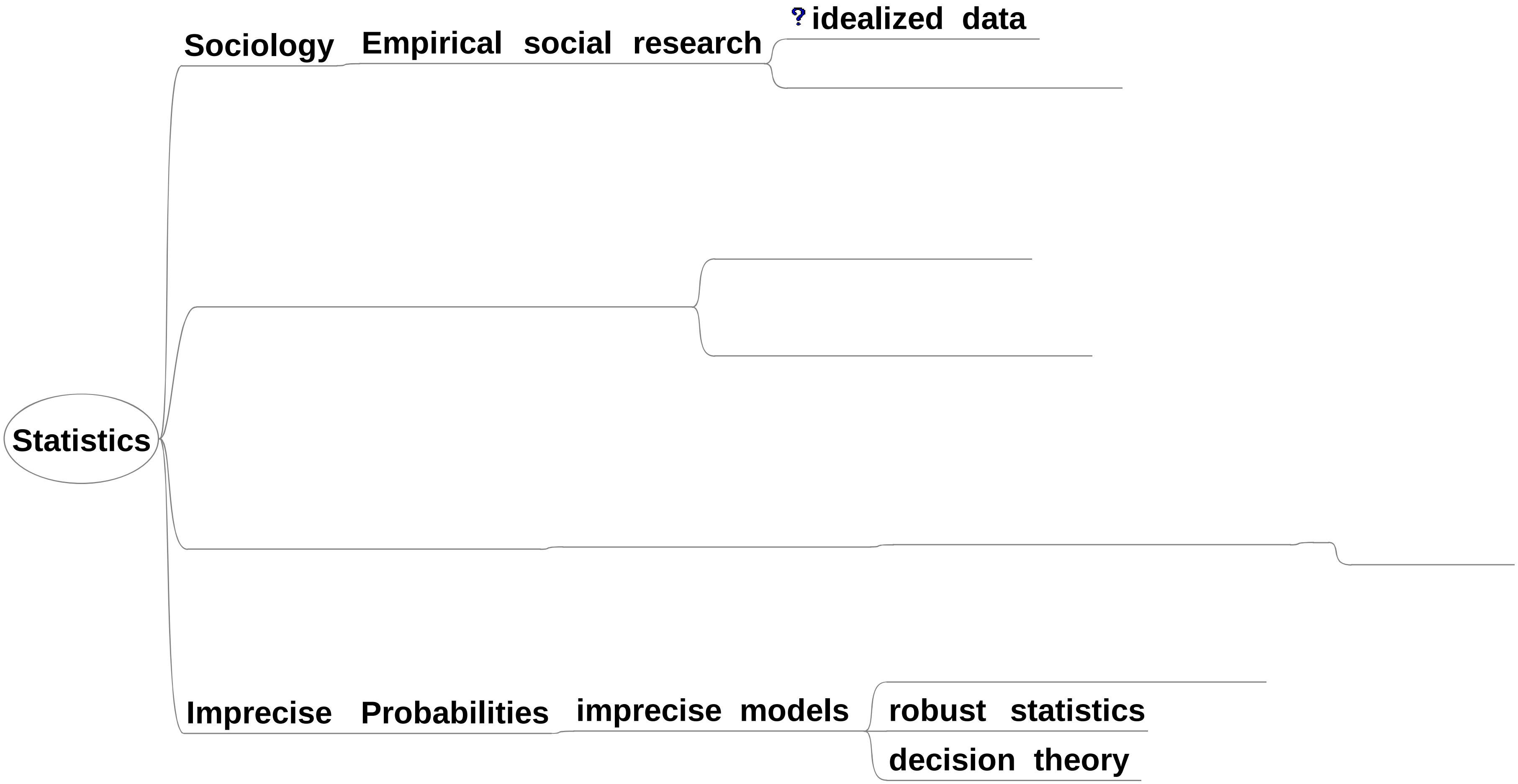
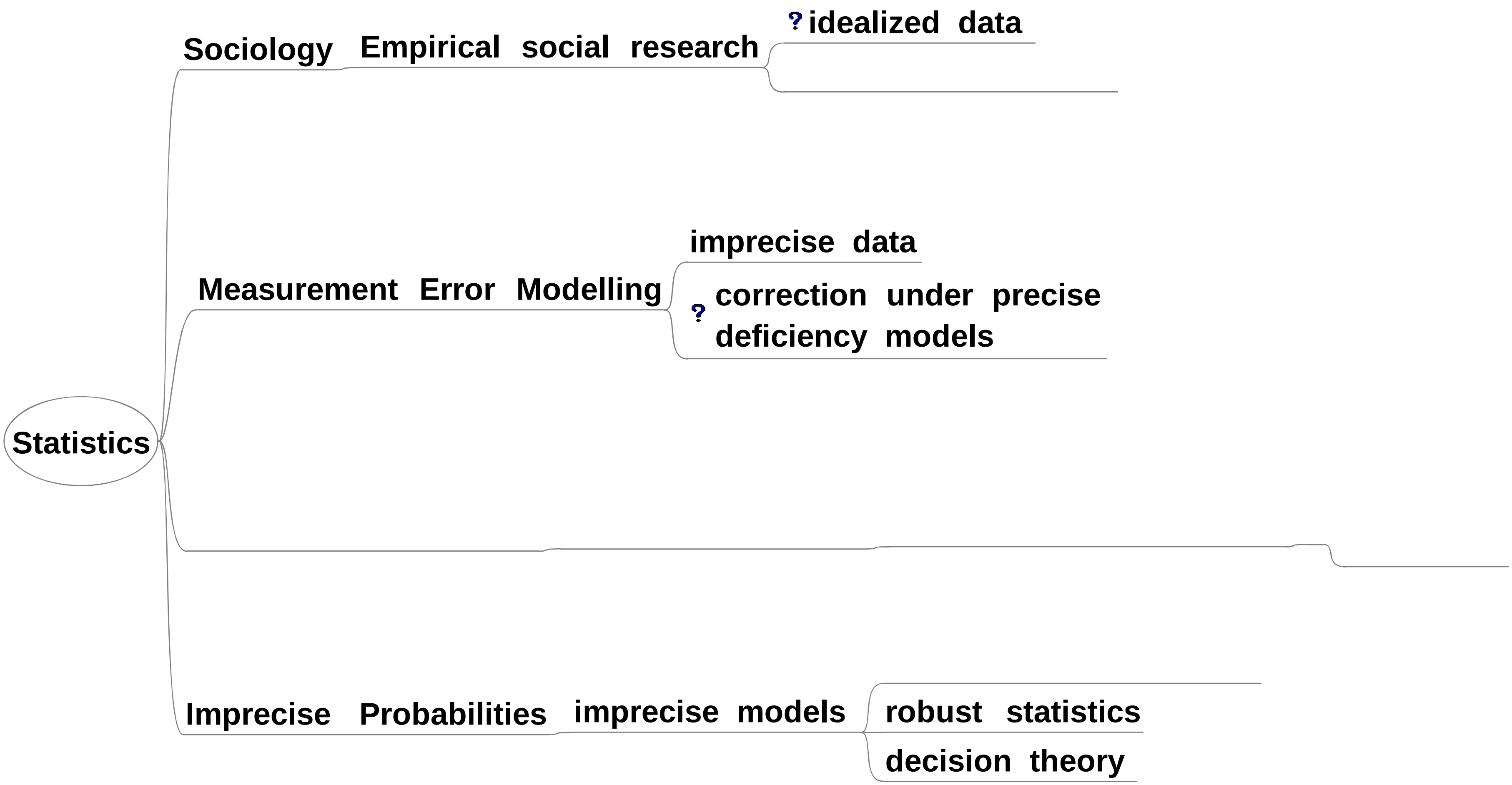
Imprecise Measurement Error Models – Towards a More Reliable Analysis of Messy Data

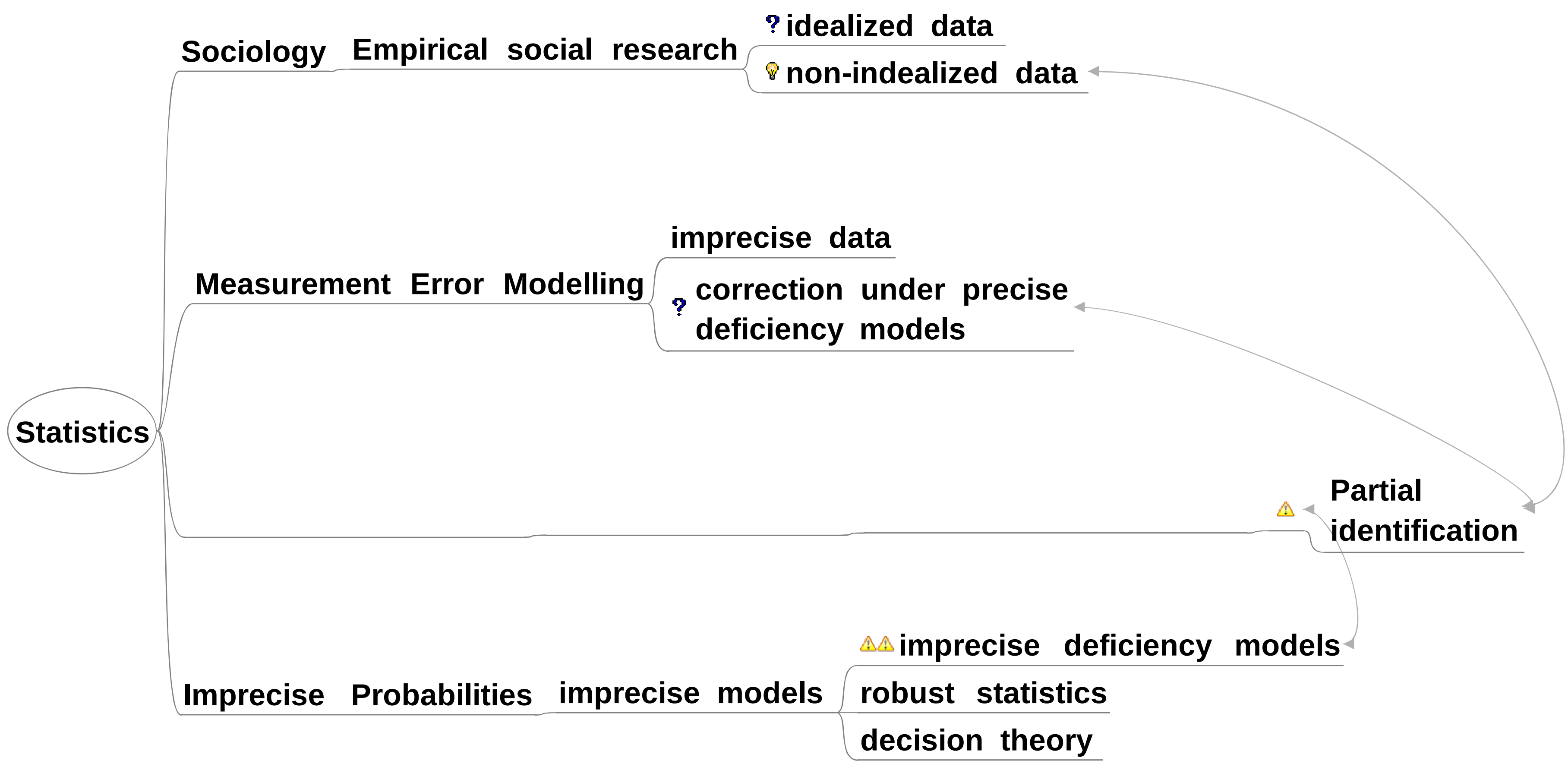
Thomas Augustin

Department of Statistics,
Ludwig-Maximilians Universität (LMU)
Munich, Germany


Statistics
Sociology
Empirical social research
? idealized data


Statistics
Sociology
Empirical social research
? idealized data
Imprecise Probabilities
imprecise models
robust statistics
decision theory


Statistics
Sociology
Empirical social research
? idealized data
Measurement Error Modelling
imprecise data
? correction under precise deficiency models
Imprecise Probabilities
imprecise models
robust statistics
decision theory


Statistics
Sociology
Empirical social research
idealized data
non-indealized data
Measurement Error Modelling
imprecise data
correction under precise deficiency models
Partial identification
Imprecise Probabilities
imprecise models
imprecise deficiency models
robust statistics
decision theory

Sketch of the Argument

- Omnipresence of measurement error
- Severe bias in statistical analysis when neglecting it
- Powerful correction methods based on the "classical model of testing theory"

  construct unbiased estimating functions $\rightarrow$ zero expectation $\rightarrow$ consistency

- The underlying assumptions are very restrictive, and rarely satisfied in social surveys
- Relax assumptions: imprecise measurement error model

  construct unbiased sets of estimating functions $\rightarrow$ zero expectation for one element of the credal set $\rightarrow$ credal consistency

Outline

1. **Measurement Error**

2. **Measurement Error Correction based on Precise Error Models**

   2.1 Classical Measurement Error Modelling

   2.2 Unbiased Estimating Equations and Corrected Score Functions for Classical Measurement Error

3. **Overcoming the Dogma of Precision in Deficiency Models**

   3.1 Credal Deficiency Models as Imprecise Measurement Error Models

   3.2 Credal Consistency of Set-Valued Estimators

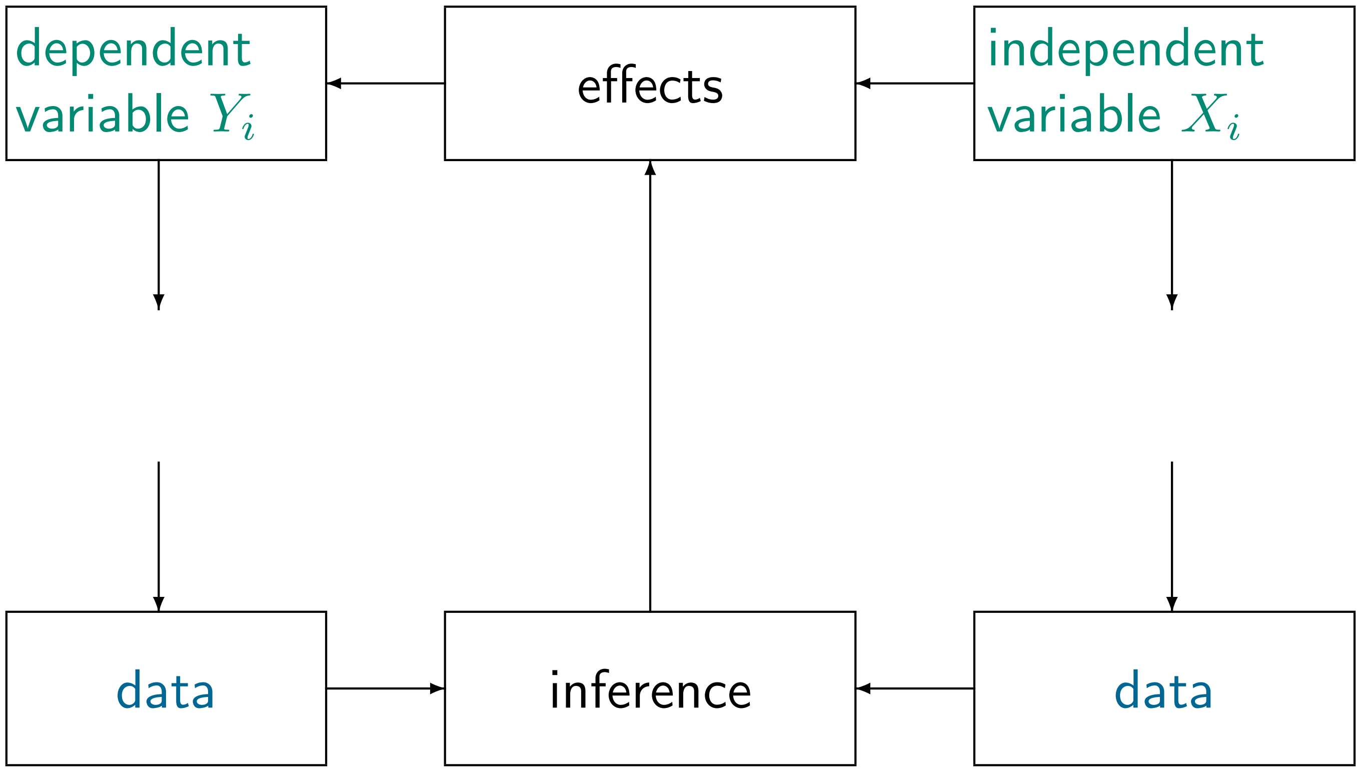
   3.3 Minimal and Complete Sets of Unbiased Estimating Functions under Imprecise Measurement Error Models

1. Measurement Error

1. Measurement Error

Applied Statistics (statistical modelling): Learning from data by sophisticated models

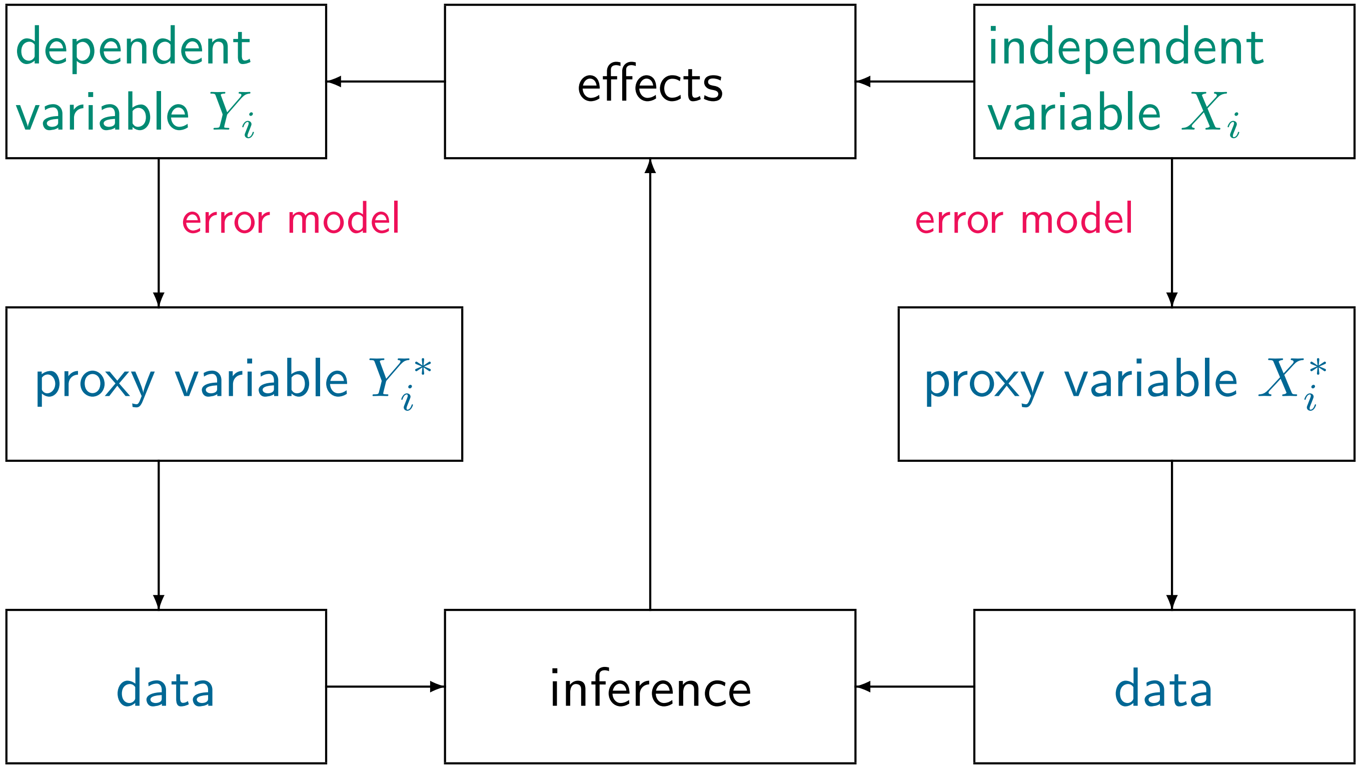
Complex relationships between variables



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.



*** -Notation (here)**

X, Z : (unobservable) variable, gold standard

X^*, Z^*: corresponding possibly incorrect measurements

analogously: Y, Y^* and T, T^*

Typical examples: Measurement Error

- Error-prone measurements of true quantities
  * error in technical devices
  * indirect measurement
  * response effects
  * use of aggregated quantities, averaged values, imputation, rough estimates etc.
  * anonymization of data by deliberate contamination

- Measured indicators of complex constructs; latent variables
  * long term quantities: long term protein intake, long term blood pressure
  * permanent income
  * importance of a patent
  * extent of motivation, degree of costumer satisfaction
  * severeness of undernutrition

Note:

- 'Measurement error' and 'misclassification' are not just a matter of sloppiness.
- Latent variables are eo ipso not exactly measurable.
- "Almost all economic variables are measured with error. [...] Unfortunately, the statistical consequences of errors in explanatory variables are severe."
  (Davidson and Mackinnon (1993), Estimation and Inference in Econometrics.)
- In nonlinear models, the later statement may apply (!?) to the dependent variable, too. (Dependence on the DGP: Torelli & Trivellato (1993, J. Econometrics))

The triple whammy effect of measurement error
Carroll, Ruppert, Stefanski, Crainiceanu (2006, Chap.H.)

- bias
- masking of features
- loss of power

- **classical error**: "attenuation "

Results

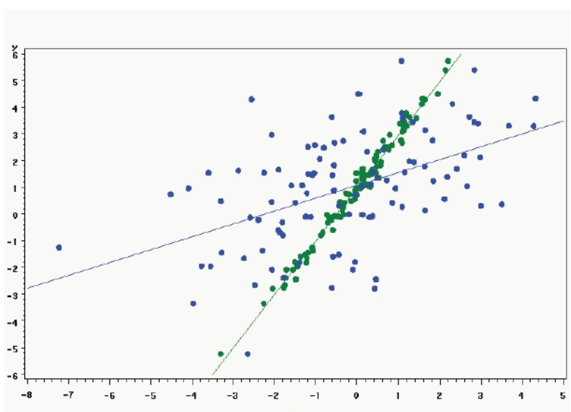


Figure 1: Effect of additive measurement error on linear regression

2. Measurement Error Correction based on Precise Error Models

2.1. Classical Measurement Error Modelling

Terminology

continuous variables → measurement error → error U_i

discrete variables → misclassification →

error U_i ↙ ↓ ↘ ↘

misclassification ↙

classical meas. error	Berkson error	Rounding Heaping	Anonymization Techniques
$X_i \perp U_i$	$X_i^* \perp U_i$		

The Classical Model of Testing Theory

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

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

Assumptions on the distribution

$\mathbb{E}(U_i)$	$=$	0	[A1.1]
$Var(U_i)$	$=$	$\sigma_i^2\ (\equiv \sigma)$	[A1.2]
U_i	$\sim$	$N(0, \sigma^2)$	[A1.3]

Independence Assumptions "$\perp$" (Uncorrelatedness)

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

2.2 Unbiased Estimating Equations and

Corrected Score Functions for Classical Measurement Error

Full Bayes-Inference in Flexible Models

- Sample latent variables in a hierarchial setting: MCMC; graphical modelling
- Berry, Carroll & Ruppert (2002, JASA).
- Kneib, Brezger & Crainiceanu (2010, Fahrmeir Fest.)
- Rummel, Augustin & Küechenhoff (2010, Biometrics)

The Basic Idea: Estimating Functions

- Idea: Do not investigate estimators directly but the equations producing estimators

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

- 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

- Frame the problem in terms of *unbiased estimating functions (score functions)* for the parameter ϑ

$$s^{\mathbf{X}}(\mathbf{Y};\mathbf{X};\vartheta) \quad \text{such that} \quad \mathbb{E}_{\vartheta}(s^{\mathbf{X}}(\mathbf{Y};\mathbf{X};\vartheta)) = 0$$

  at the true parameter value ϑ

- Concept contains as special cases

  * maximum likelihood estimators
  * least squares estimators in linear regression
  * quasi-likelihood estimators (McCullagh, 1981, Ann. Stat.; 1990, Cox Fest.)
  * M-estimators (Huber, 1981, Wiley) ;
  * Godambe (1991, Oxford UP).
  * GMM-estimators (Wansbeek & Meijer, 2000, Elsevier)

- Under mild regularity conditions still

  * consistency
  * and asymptotic normality.

- 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

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

  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 $s^{X^*}(\mathbf{Y},\mathbf{X}^*,\vartheta)$ in the error prone data with

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

The technical argument condensed:

on the construction of unbiased estimating equations under measurement error

- ϑ true parameter value
- Ideal estimating function: $\psi^{X,Y}(\mathbf{X}, \mathbf{Y}, \vartheta)$
- Naive estimating function: $\psi^{sic!\ X,Y}(\mathbf{X}^*, \mathbf{Y}^*, \vartheta)$
- Find $\psi^{X^*,Y^*}(\mathbf{X}^*, \mathbf{Y}^*, \vartheta)$ such that

$$\mathbb{E}_\vartheta\left(\psi^{X^*,Y^*}(\mathbf{X}^*, \mathbf{Y}^*, \vartheta)\right) \stackrel{!}{=} 0 \qquad (*)$$

- Idea: use the ideal score function as a building block!
- Try $\psi^{X^*,Y^*}(\mathbf{X}^*, \mathbf{Y}^*, \vartheta) = f(\psi^{X,Y}(\mathbf{X}^*, \mathbf{Y}^*, \vartheta))$ for some appropriate $f(\cdot)$

- In general, $\psi^{X^*,Y^*}(\cdot)$, i.e. $f(\cdot)$, can not be determined directly.

- Note that, since $\mathbb{E}_\vartheta\left(\psi^{X,Y})(X,Y,\vartheta)\right) = 0$, $(*)$ is equivalent to

$$\mathbb{E}_\vartheta\left(\psi^{X^*,Y^*}(\mathbf{X}^*,\mathbf{Y}^*,\vartheta)\right) = \mathbb{E}_\vartheta\left(\psi^{X,Y}(\mathbf{X},\mathbf{Y}^*,\vartheta)\right)$$

- Look at the expected difference between $\psi^{X^*,Y^*}(\cdot)$ and $\psi^{X,Y}(\cdot)$.

- Try to break $\psi^{X,Y}(\mathbf{X},\mathbf{Y},\vartheta)$ into „additive pieces“, and handle it piece by piece

- Typically, $\psi(\cdot)$ has the form $\psi(X, Y, \vartheta) = \frac{1}{n}\sum_{i=1}^{n} \psi_i(X_i, Y_i, \vartheta)$, and there are representations such that, for $i = 1, \ldots, n$,

$$\psi_i(X_i, Y_i, \vartheta) = \sum_{j=1}^{s} g_j(X_i, Y_i, \vartheta).$$

- Then try to find $f_1(\cdot), \ldots, f_s(\cdot)$ such that

$$\mathbb{E}_\vartheta \left(f_j(g_j(X_i^*, \mathbf{Y}_\mathbf{i}^*, \vartheta))\right) = \mathbb{E}_\vartheta \left(g_j(X_i, \mathbf{Y}_\mathbf{i}, \vartheta)\right) \quad (**)$$

- (conditionally/locally) corrected score functions (for covariate measurement error: Nakamura (1990, Biometrika), Stefanski (1989, Comm. Stat. Theory Meth.))

  Try to find $f_1(\cdot), \ldots, f_s(\cdot)$ such that

$$\mathbb{E}\vartheta \left(f_j(g_j(X_i^*, \mathbf{Y}_\mathbf{i}^*, \vartheta))|X_i, \mathbf{Y}_i\right) = g_j(X_i, \mathbf{Y}_\mathbf{i}, \vartheta) \quad (**),$$

  then the law of iterated expectation leads to (**).

- Sometimes indirect proceeding: (globally or locally) corrected log-likelihood $\mathrm{l}^{X^*}(\mathbf{Y}, \mathbf{X}, \vartheta)$ with

$$\mathbb{E}(\mathrm{l}^{X^*}(\mathbf{Y}, \mathbf{X}^*, \vartheta)|\mathbf{X}, \mathbf{Y}) = \mathrm{l}^{X}(\mathbf{Y}, \mathbf{X}, \vartheta).$$

or

$$\mathbb{E}\left(\mathrm{l}^{X^*}(\mathbf{Y}, \mathbf{X}^*, \vartheta)\right) = \mathbb{E}\left(\mathrm{l}^{X}(\mathbf{Y}, \mathbf{X}, \vartheta)\right).$$

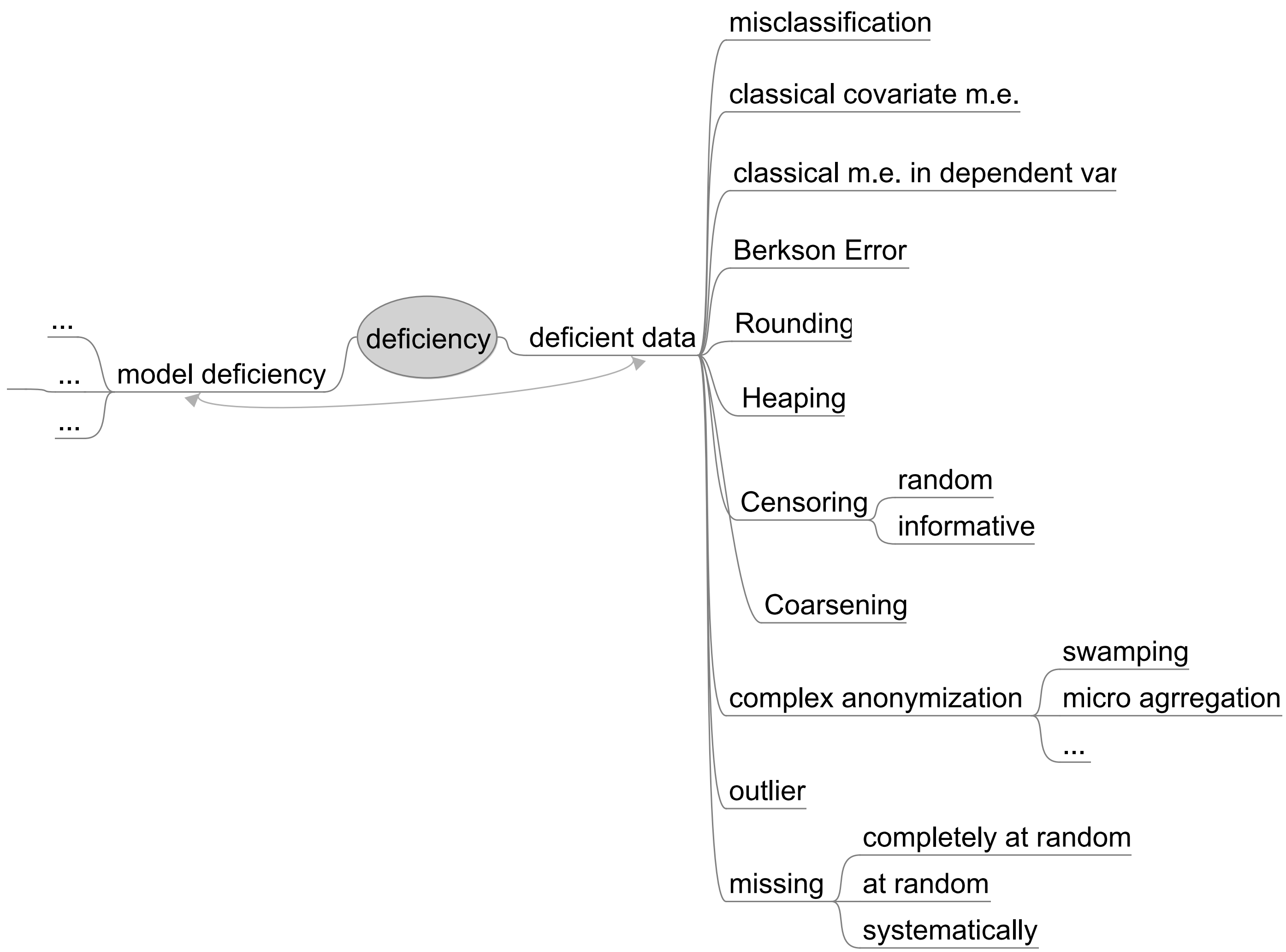
- Same techniques as before
  - * piece by piece
  - * globally or locally

- Under mild regularity conditions unbiased estimating function by taking the derivative with respect to ϑ.

+ Functional method: no (unjustified !?) assumptions on the distr. of X
+ Successful for generalized linear models, polynomial regression, etc. (Survey: Schneeweiß & Augustin, 2006, ASTA; Hübler & Frohn (eds.); Cox model: Augustin (2004, Scand. J. Stat.))
– Numerical difficulties for small samples
– Handling of transformations (e.g. $\ln X$) complicated or impossible
– Non-existence of corrected score functions for some models
+ Extensions to misclassification (Akazawa, Kinukawa, Nakamura, 1998, J. Jap. Stat. Soc.; Zucker, Spiegelman, 2008, Stat. Med.)
+ Quite general error distribution can often be handled (only existing moment generating function needed); this includes deliberate contamination for privacy protection (Bleninger & Augustin)
+ Extension to dependent variable ⇒ unified understanding of censoring and measurement error (Pötter & Augustin)
+ Extension to Berkson Error (Wallner & Augustin)
+ Extension to rounding (Felderer, Müller, Schneeweiß, Wiencierz & Augustin)

3. Overcoming the Dogma of Precision in Deficiency Models

3.1 Credal Deficiency Model as Imprecise Measurement Error Models



Manski's Law of Decreasing Credibility

Reliability !? Credibility ?

> "The credibility of inference decreases with the strength of the assumptions maintained." (Manski (2003, p. 1))

Manski's Law of Decreasing Credibility

Reliability !? Credibility ?

> "The credibility of inference decreases with the strength of the assumptions maintained." (Manski (2003, p. 1))

Identifying assumptions in measurement error models Very strong assumptions needed to ensure identifiability = precise solution

- Measurement error model must be known precisely
  - type of error, in particular assumptions on (conditional) independence
    - independence of true value
    - independence of other covariates
    - independence of other measurements
  - type of error distribution
  - moments of error distribution

- validation studies typically not available

Reliable Inference Instead of Overprecision!

- Make more „realistic“ assumption and let the data speak for themselves!
- Consider the *set* of *all* models that maybe compatible with the data (and then add successively additional assumptions, if desirable)
- The results may be imprecise, but are more reliable for sure
- **The extent of imprecision is related to the data quality!**
- As a welcome by-product: clarification of the implication of certain assumptions
- parallel developments (missing data; transfer to measurement error context!)
  * economics: *partial identification:* e.g., Manski (2003, Springer)
  * biometrics: *systematic sensitiviy analysis:* e.g., Vansteelandt, Goetghebeur, Kenword, Molenberghs (2006, Stat. Sinica)
- current developments, e.g.,
  * Cheng, Small (2006, JRSSB)
  * Henmi, Copas, Eguchi (2007, Biometrics)
  * Stoye (2009, Econometrica)
  * Gustafson & Greenland (2009, Stat. Science)
- Kleyer (2009, MSc.); Kunz, Augustin, Küchenhoff (2010, TR)

Credal Deficiency Models

Different types of deficiency can be expressed

- Measurement error problems
- Misclassification
- If $\mathcal{Y}^* \subseteqq \mathcal{P}(\mathcal{Y}) \times \{0,1\}$: coarsening, rounding, censoring, missing data
- Outliers

credal set: convex set of traditional probability distributions

$$
\begin{array}{rclcl}
[Y|X,\vartheta] & \in & \mathcal{P}_{Y|X,\vartheta} & & \\
[Y^*|X,Y] & \in & \mathcal{P}_{Y^*|X,Y} & \text{or} & [Y|X,Y] \in P_{Y|Y^*,X} \\
[X^*|X,Y] & \in & \mathcal{P}_{X^*|X,Y} & \text{or} & [X|X^*,Y] \in P_{X|X^*,Y}
\end{array}
$$

Credal Estimation

- Natural idea: *sets* of traditional models $\longrightarrow$ *sets* of traditional estimators

- Construct estimators $\widehat{\Theta} \subseteq \mathbb{R}^p$, i.e. set of plausible parameter values, appropriately reflecting the ambiguity (non-stochastic uncertainty, ignorance) in the credal set $\mathcal{P}$.

- $\widehat{\Theta}$ "small" if and only if (!) $\mathcal{P}$ is "small"

  * Usual point estimator as the border case of precise probabilistic information
  * Connection to Manski's (2003) *identification regions* and Vansteelandt, Goetghebeur, Kenward & Molenberghs (Stat Sinica, 2006) *ignorance regions*.

- Construction of unbiased sets of estimating functions

- Credal consistency

3.2 Credal Consistency

- $\left(\widehat{\Theta}^{(n)}\right)_{n\in\mathbb{N}} \subseteq \mathbb{R}^p$ is called *credally consistent* (with respect to the credal set $\mathcal{P}_\vartheta$) if $\forall \vartheta \in \Theta$:

$$\forall p \in \mathcal{P}_\vartheta \; \exists \; \left(\hat{\vartheta}_p^{(n)}\right)_{n\in\mathbb{N}} \in \left(\widehat{\Theta}^{(n)}\right)_{n\in\mathbb{N}} : \operatorname*{plim}_{n\to\infty} \hat{\vartheta}_p^{(n)} = \vartheta.$$

3.3 Construction of Credally Consistent Estimators

- Transfer the framework of unbiased estimating functions

- A set Ψ of estimating functions is called

  * *unbiased* (with respect to the credal set $\mathcal{P}_\vartheta$) if for all ϑ:

$$\forall \psi \in \Psi \ \exists p_{\psi,\vartheta} \in \mathcal{P}_\vartheta : \quad \mathbb{E}_{p_{\psi,\vartheta}}(\Psi) = 0$$

  * *complete* (with respect to the credal set $\mathcal{P}_\vartheta$) if for all ϑ:

$$\forall p \in \mathcal{P}_\vartheta \ \exists \psi_{p,\vartheta} \in \Psi : \quad \mathbb{E}_p(\psi_{p,\vartheta}) = 0.$$

- A complete and unbiased set ψ of estimating functions is called *minimal* if there is no complete and unbiased set of estimating functions $\tilde{\Psi} \subset \Psi$.

Construction of Minimal Consistent Estimators

Define for some set Ψ of estimating functions

$$\widehat{\Theta}_\Psi = \left\{ \hat{\vartheta} \middle| \, \hat{\vartheta} \text{ is root of } \psi, \, \psi \in \Psi \right\}.$$

Under the usual regularity conditions (in particular unique root for every ψ)

- Ψ unbiased and complete $\Rightarrow \widehat{\Theta}_\Psi$ credally consistent

3.4 Examples

- *Imprecise sampling model*: neighborhood model $\mathcal{P}_{Y|X,\vartheta}$ around some ideal central distribution $p_{Y|X,\vartheta}$
  Let ψ be an unbiased estimation function for $p_{Y|X,\vartheta}$. Then (if well defined)
  $$\Psi = \left\{\psi^* \middle|\, \psi^* = \psi - \mathbb{E}_p(\psi),\, p \in \mathcal{P}_{Y|X,\vartheta}\right\}$$
  is unbiased and complete.

- *Imprecise measurement error model*, e.g. $\mathcal{P}_{X^*|X,Y}$:
  $\Psi = \left\{\psi \middle|\, \psi \text{ is corrected score function for some } p \in \mathcal{P}_{X^*|X,Y}\right\}$ is unbiased and complete.

- Construction of confidence regeions:
  * union of traditional confidence regions
  * can often be improved (Vansteelandt, Goetghebeur, Kenward & Molenberghs (Stat Sinica, 2006), Stoye (2009, Econometrica)).

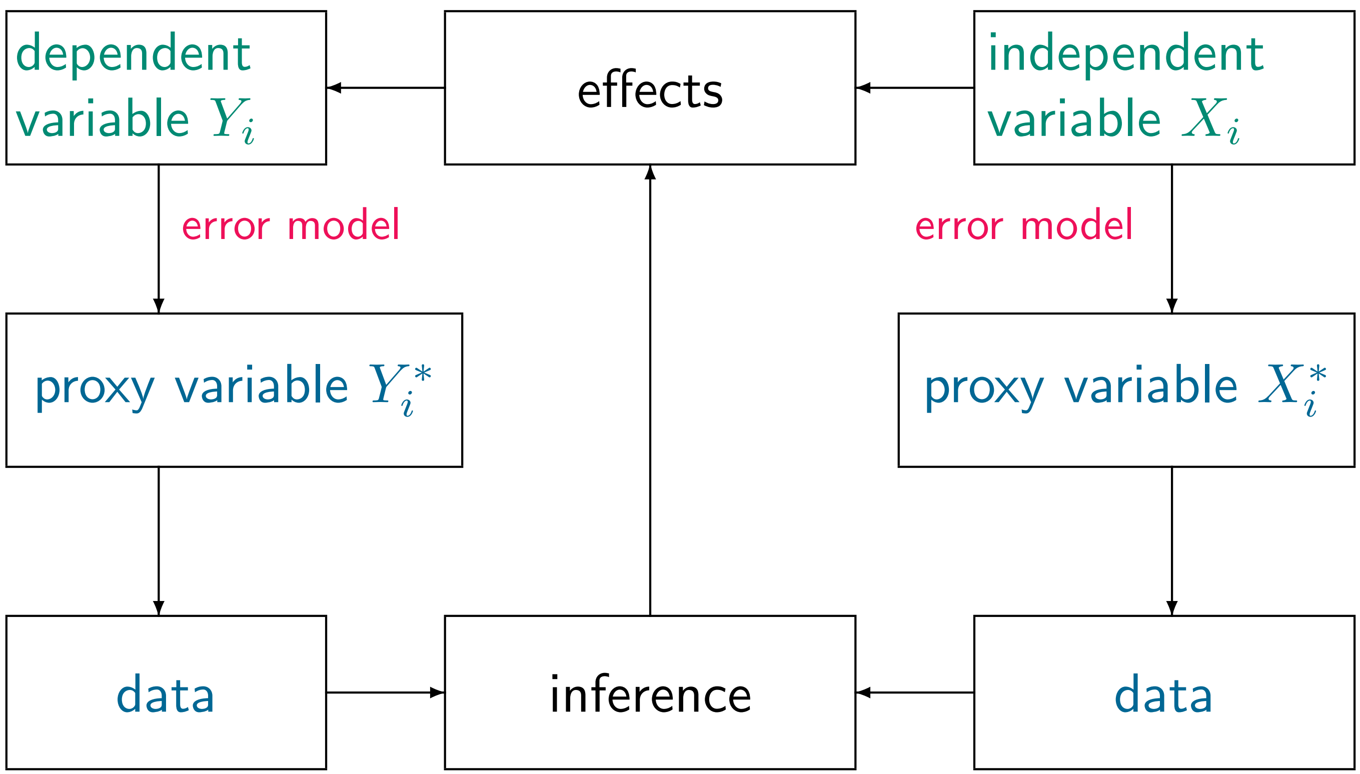
Sketch of the Argument

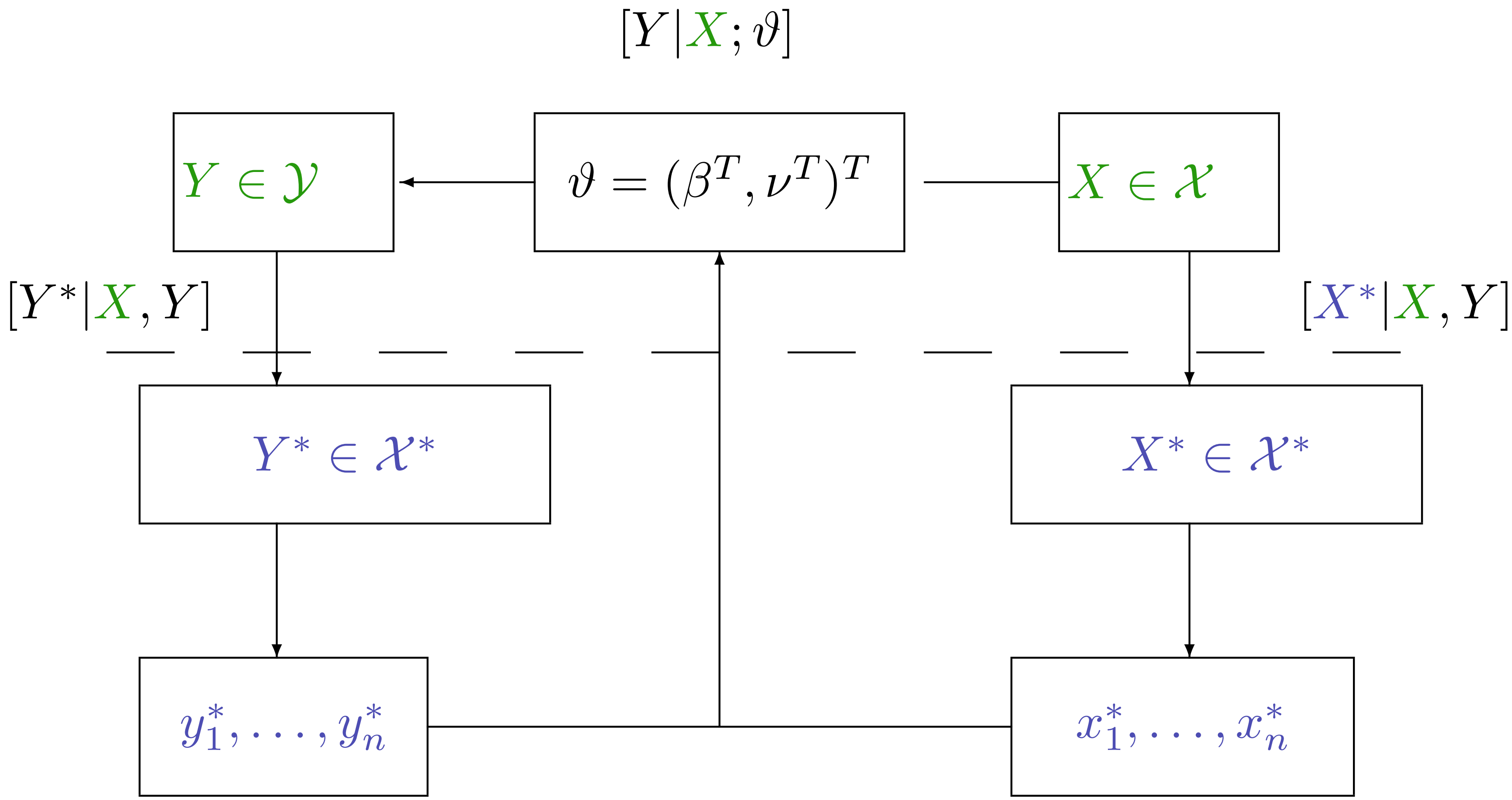
- Omnipresence of measurement error
- Severe bias in statistical analysis when neglecting it
- Powerful correction methods based on the "classical model of testing theory"

  construct unbiased estimating functions $\rightarrow$ zero expectation $\rightarrow$ consistency

- The underlying assumptions are very restrictive, and rarely satisfied in social surveys
- Relax assumptions: imprecise measurement error model

  construct unbiased sets of estimating functions $\rightarrow$ zero expectation for one element of the credal set $\rightarrow$ credal consistency


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


$[Y|X;\vartheta]$
$Y \in \mathcal{Y}$
$\vartheta = (\beta^T, \nu^T)^T$
$X \in \mathcal{X}$
$[Y^*|X, Y]$
$[X^*|X, Y]$
$Y^* \in \mathcal{X}^*$
$X^* \in \mathcal{X}^*$
$y_1^*, \ldots, y_n^*$
$x_1^*, \ldots, x_n^*$