Making Sense of Noisy Data:
Why and How?

Grace Y. Yi

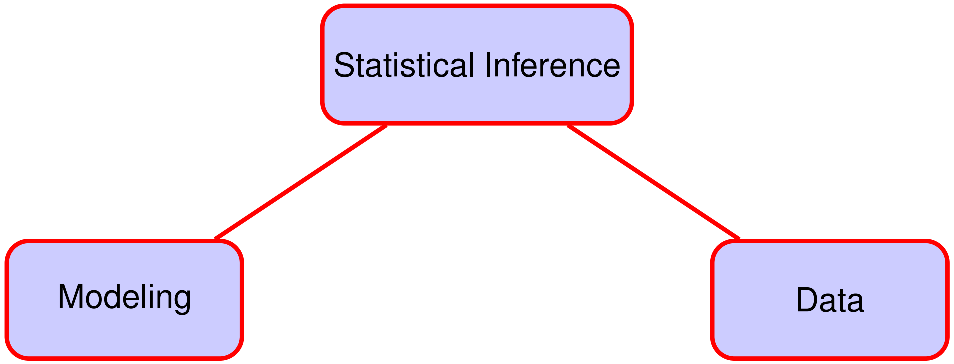
Canada Research Chair in Data Science (Tier 1)

Department of Statistical and Actuarial Sciences

Department of Computer Science

University of Western Ontario

Statistical Science

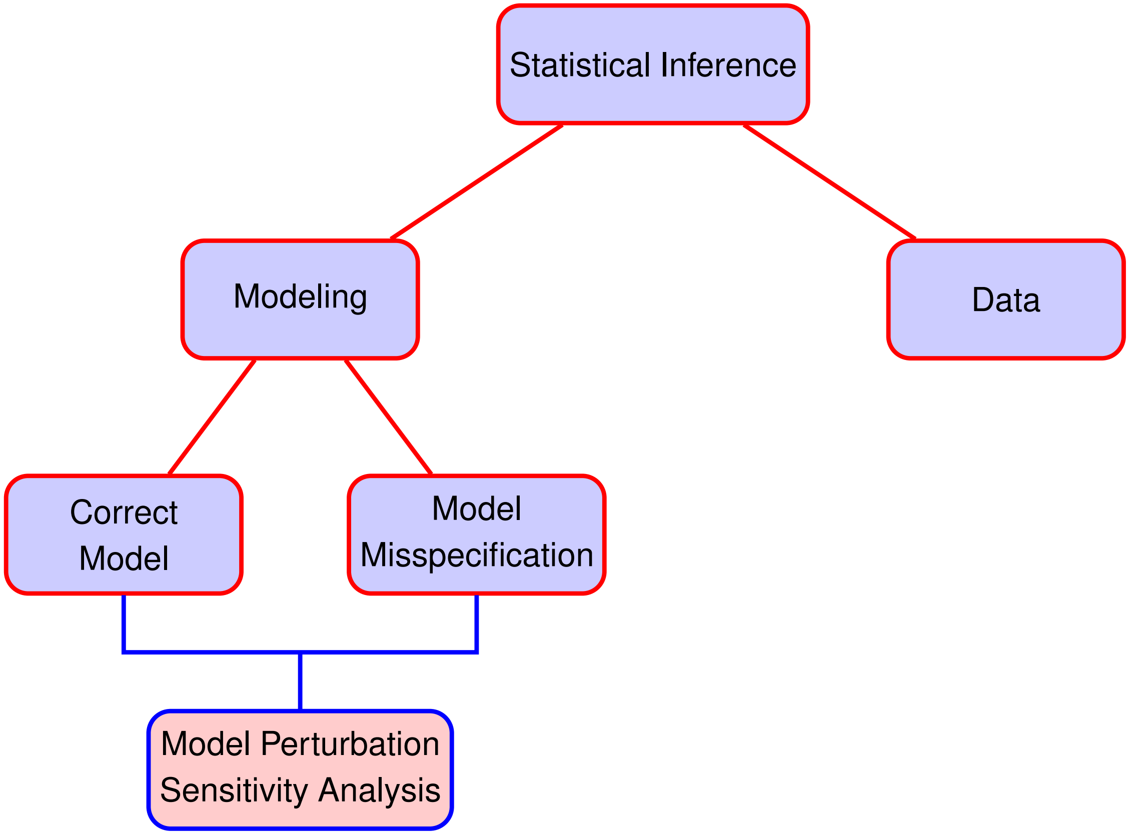


Statistical Inference

Modeling

Data

Statistical Science



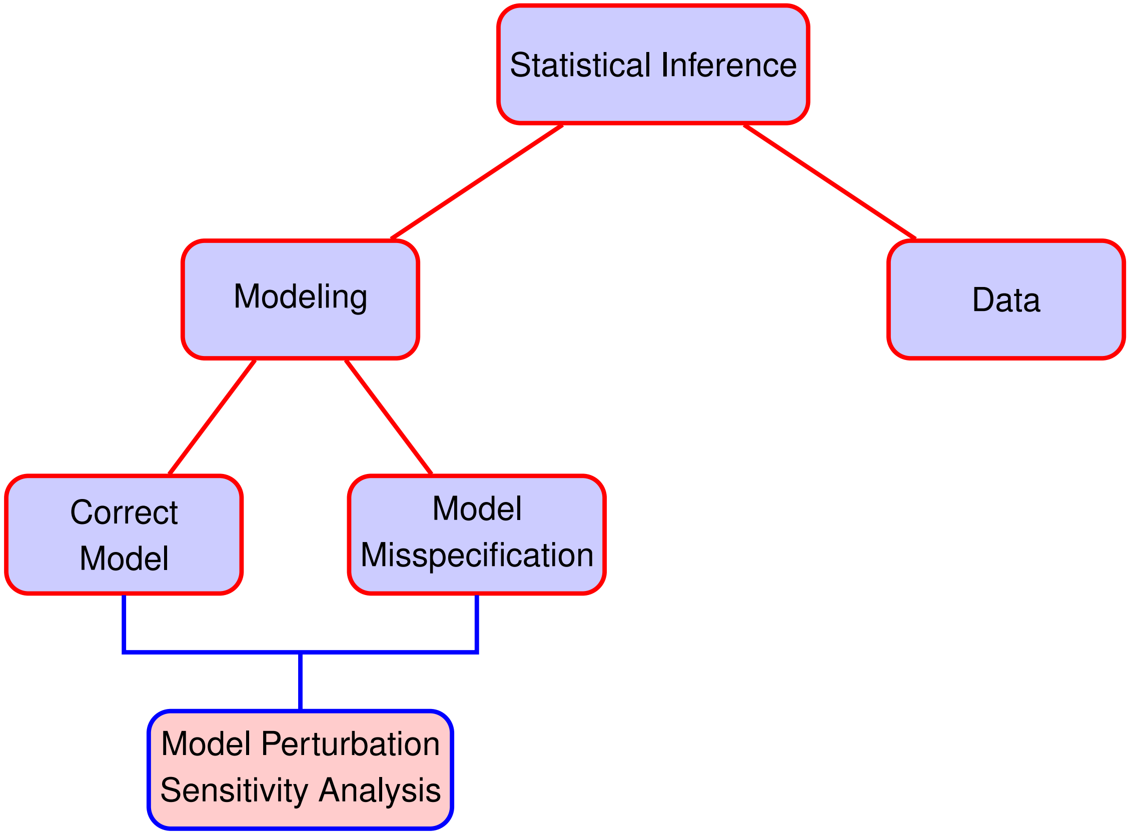
Classification of Data

- Use the collection method of data
  - by design
  - by observation

- Use the size of data
  - “small" data
  - “big" data

- Use the quality of data
  - “good" quality: complete and no error
  - “bad" quality: incomplete and error-prone

Statistical Science


Statistical Inference

Modeling

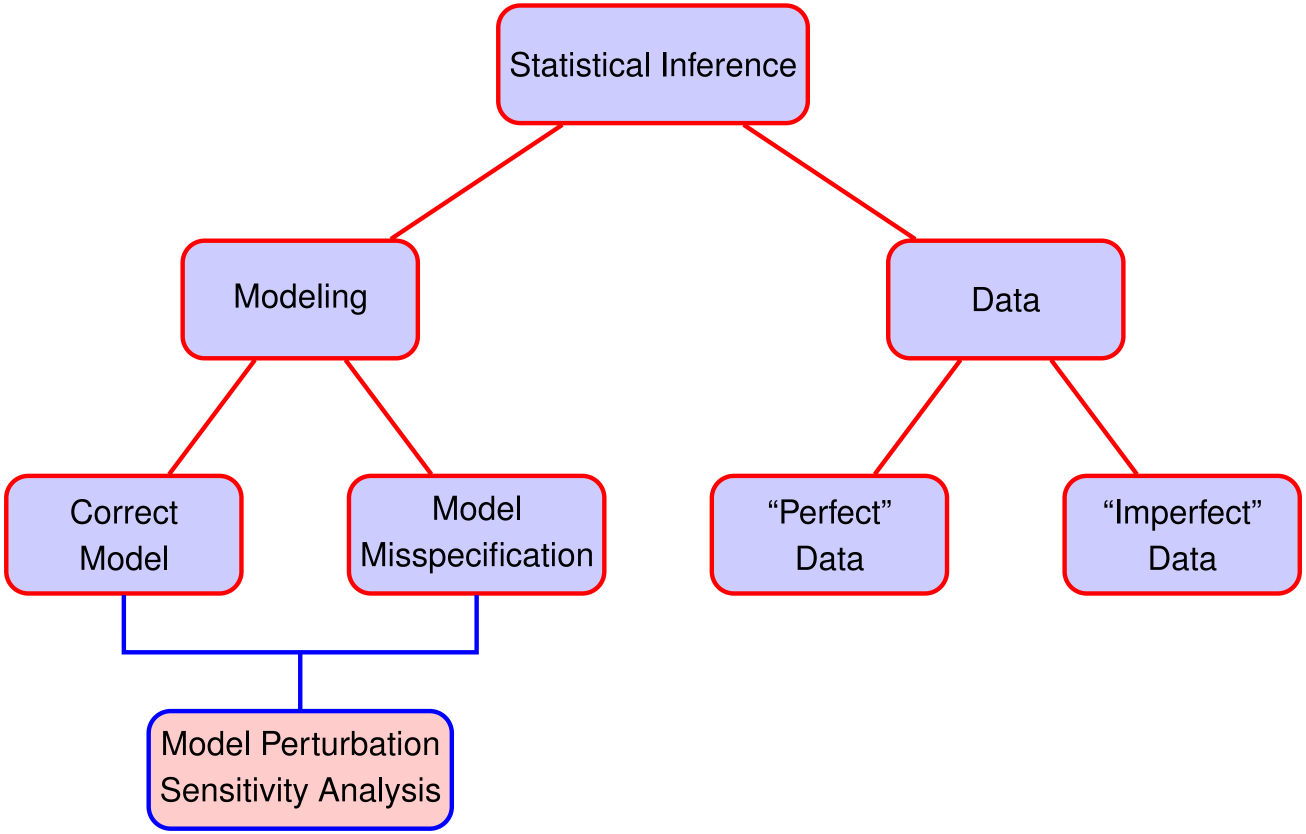
Data

Correct Model

Model Misspecification

Model Perturbation Sensitivity Analysis

Statistical Science



Statistical Inference

Modeling

Data

Correct Model

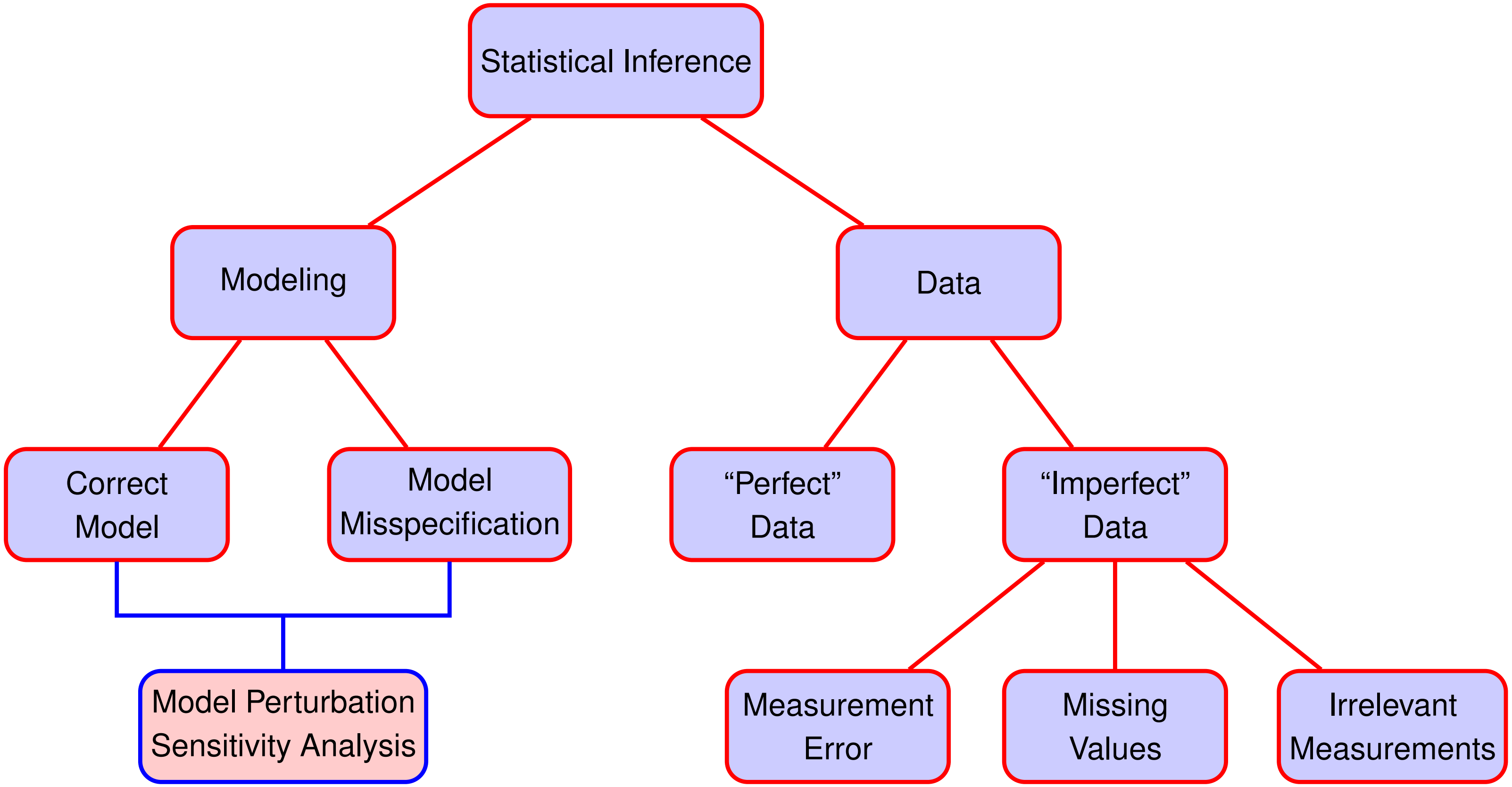
Model Misspecification

“Perfect” Data

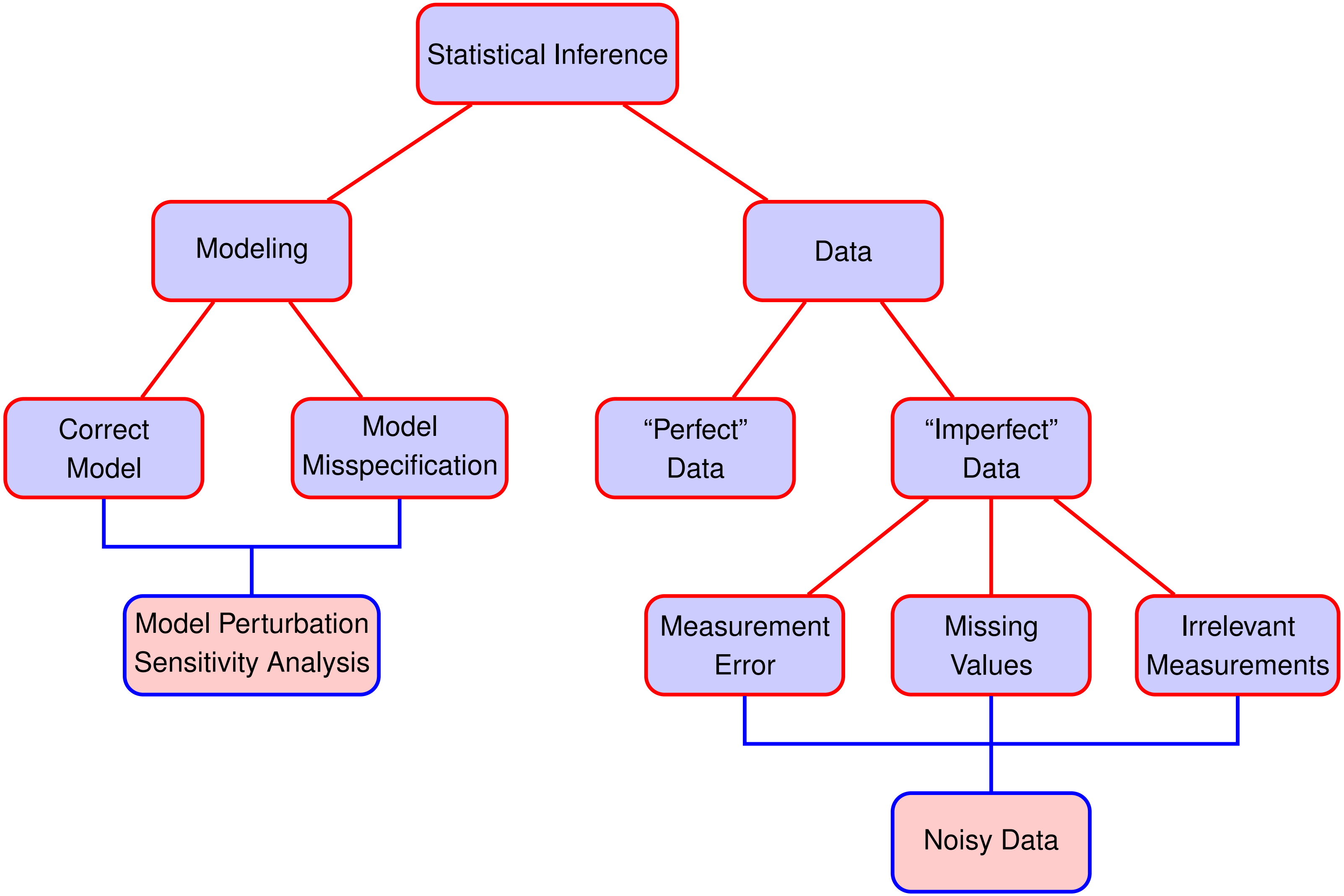
“Imperfect” Data

Model Perturbation Sensitivity Analysis

Statistical Science



Statistical Science



Statistical Inference

Modeling

Data

Correct Model

Model Misspecification

"Perfect" Data

"Imperfect" Data

Model Perturbation Sensitivity Analysis

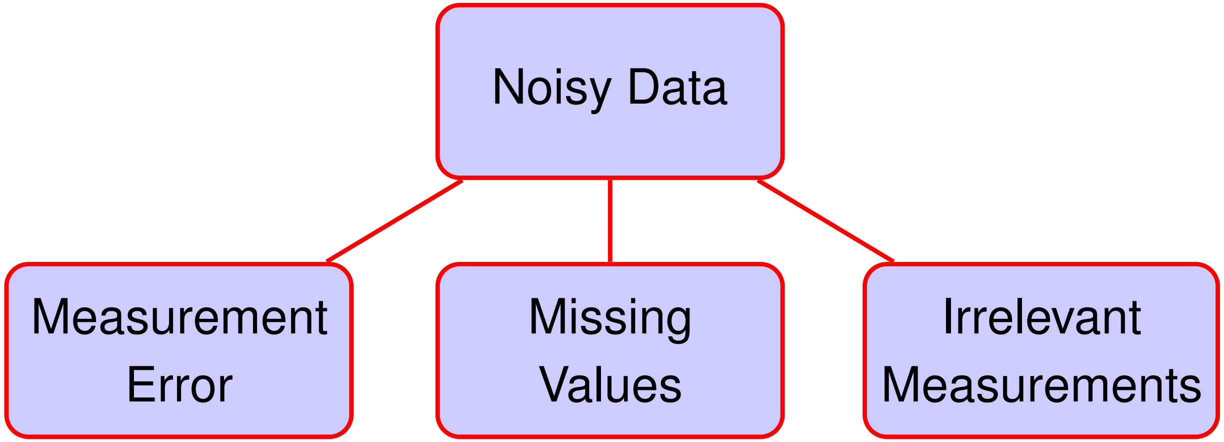
Measurement Error

Missing Values

Irrelevant Measurements

Noisy Data

Outline



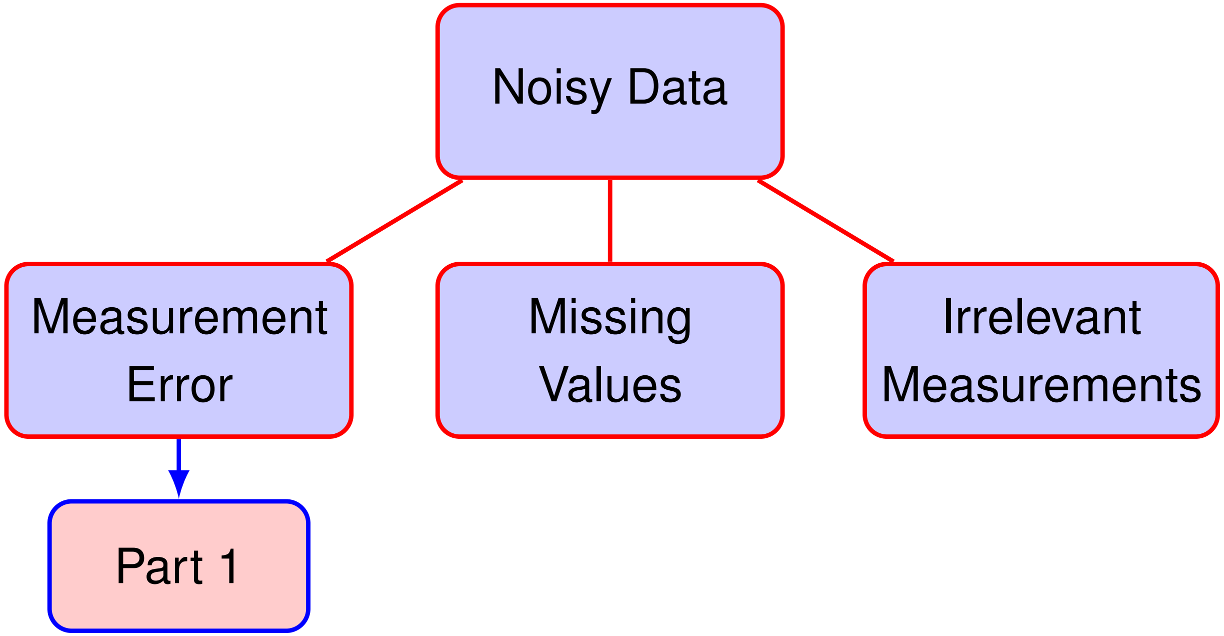
Noisy Data

Measurement Error

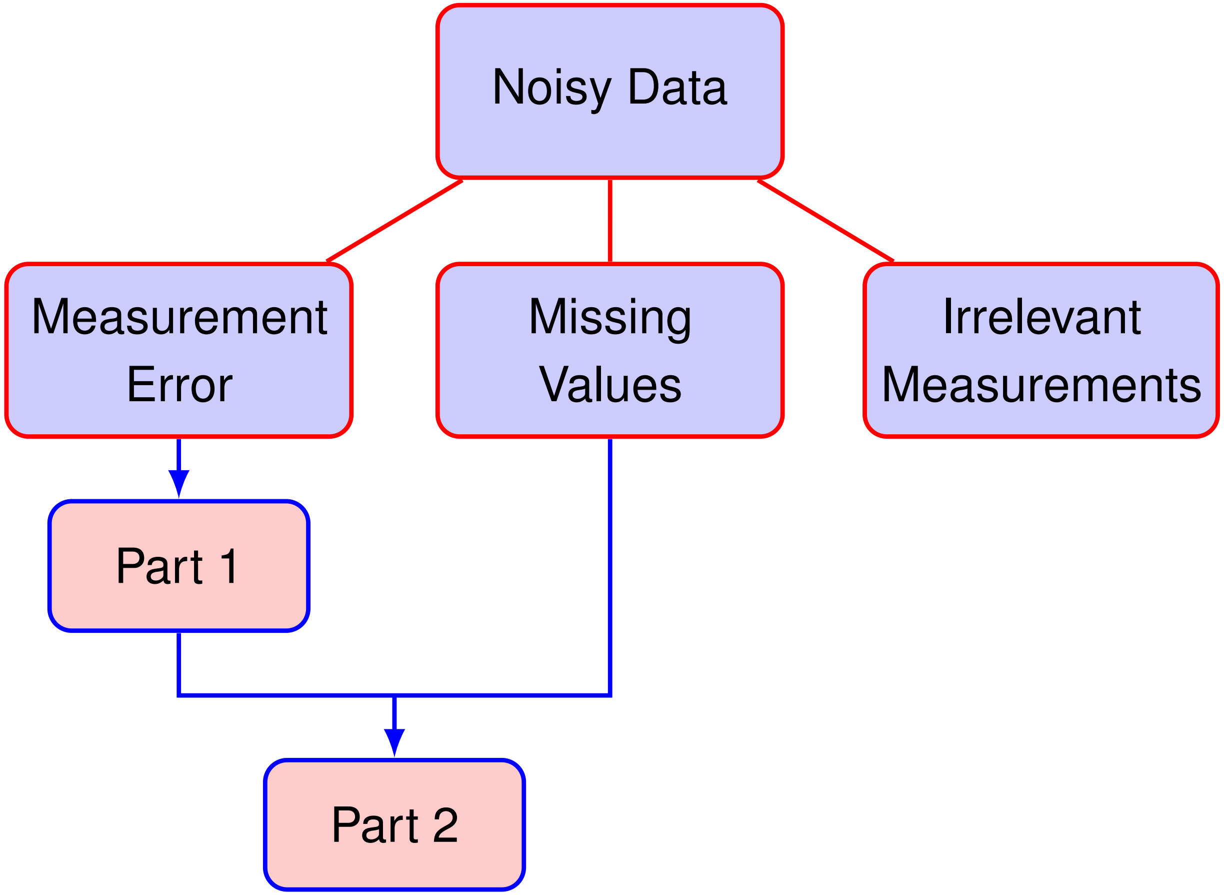
Missing Values

Irrelevant Measurements

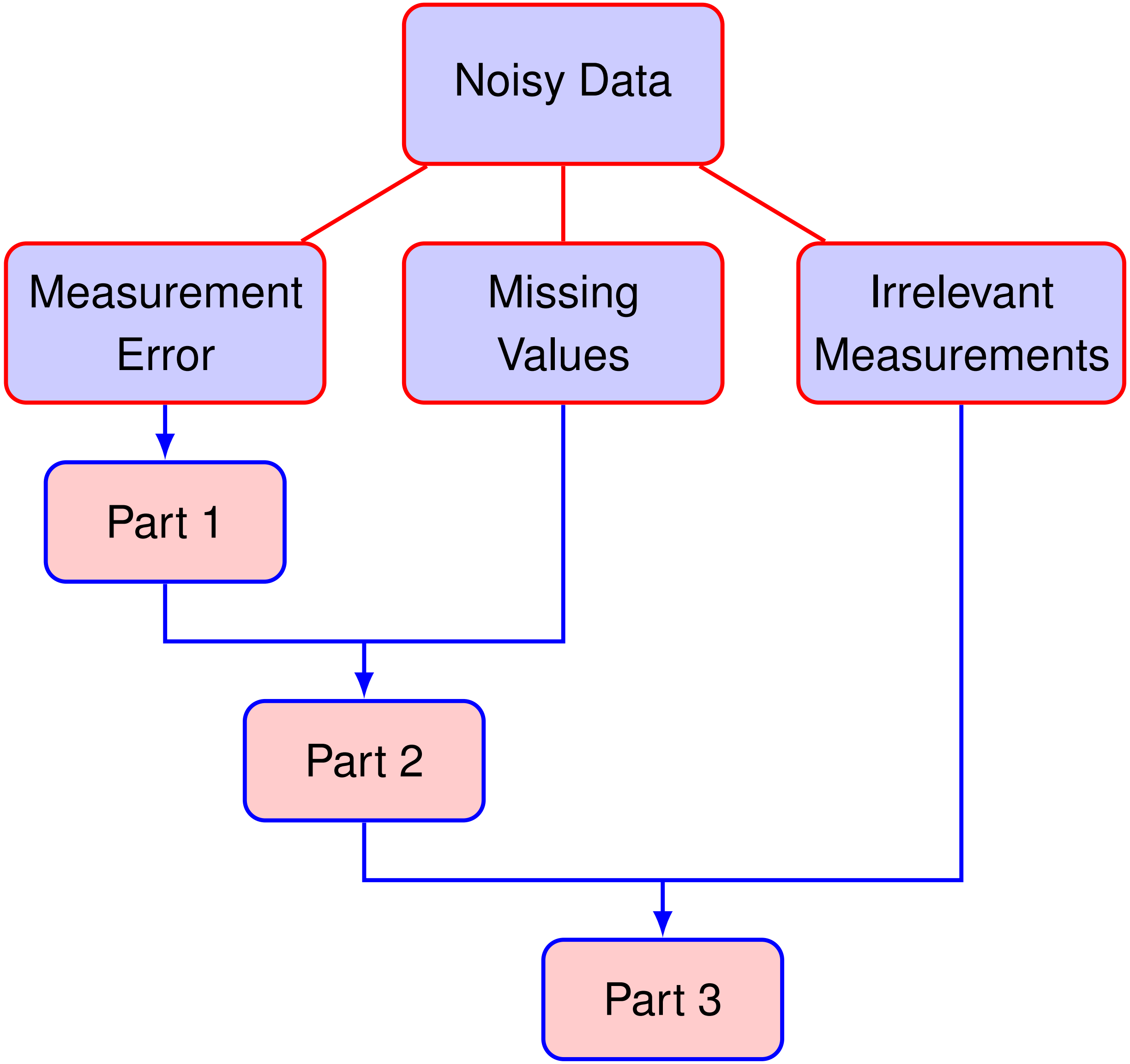
Outline


Noisy Data
Measurement Error
Missing Values
Irrelevant Measurements
Part 1

Outline


Noisy Data
Measurement Error
Missing Values
Irrelevant Measurements
Part 1
Part 2

Outline


Noisy Data
Measurement Error
Missing Values
Irrelevant Measurements
Part 1
Part 2
Part 3

Part 1: Noisy Data - Measurement Error

Part 1

Buzas (2019)

Brakenhoff et al. (2018). *Journal of Clinical Epidemiology*, 98, 89-97.
“Measurement error is often neglected in medical literature: a systematic review".

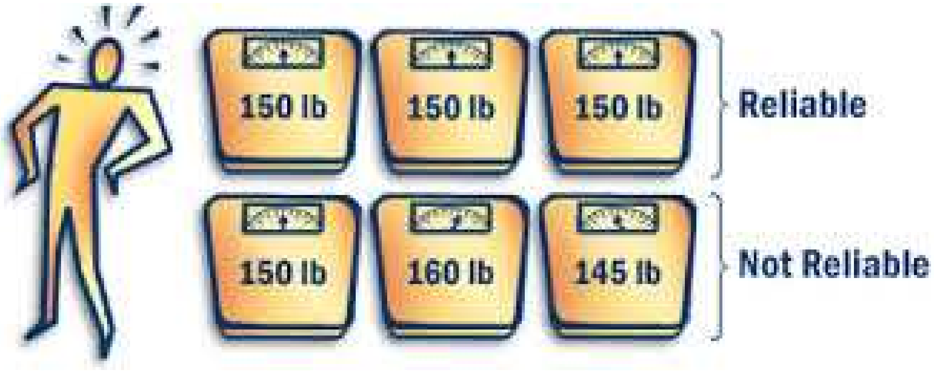
Key Findings

- Almost half (247/565) of research studies published in 2016 in top 12 general medicine and epidemiology journals mention measurement error.
- 7% of the 247 (18/247) did something about it (investigated measurement error effects or did a measurement error analysis).

Conclusions

- “More guidance and tutorials seem necessary to assist the applied researchers with the assessment of the type and amount of measurement error as well as the steps that can subsequently be taken to minimize its impact on the studied relations."
- “Given the increasing use of data not originally intended for medical or epidemiological research, we anticipate that the use and understanding of measurement error analyses and corrections will become increasingly important in the near future."

Measurement Error Examples and Sources



Example 1 - Cost Concern

Example (Case-Control Study, Carroll et al. 1993)

- Interest:
  association between invasive cervical cancer (Y) and exposure to herpes simplex virus type 2 (HSV-2) (X)

- Exposure to HSV-2 was assessed by
  - a refined western blot procedure (X)
  - a less accurate western blot procedure (X^*)

  for cases ($Y = 1$) and controls ($Y = 0$)

Issue: X is only directly observed for less than 6% of the subjects.

Example 2 - Protection of Privacy

Example (Survey Data, Hwang 1986)

- Interest: the relationship between energy consumption and housing characteristics
- 5,979 households: randomly selected from U.S.; yearly energy consumed by a chosen family was reported yearly energy was reported
- household conditions: # of windows, enclosed heated area, inches of wall insulation, roof insulation, floor insulation etc.
- - family income
  - whether there were persons staying in the house during the day
  - whether there were certain major appliances
- - geographic region index
  - local weather conditions
- Complexion

  some X_j are not reported but instead:

  $$X_j^* = X_j \cdot e_j$$

  is reported where e_j follows a given distribution

Example 3 - Reporting Error

Example (Survey Data, Bollinger 1998)

- Reporting errors are typical in survey data. For example, it was found that
  - Response error is negatively related to earnings, there is more measurement error among men than women;
  - Overreporting of income is concentrated in the lower end of the income distribution for men. High overreporting of income for low-income men is driven by about 10% of the reporters who grossly overreport their income;
  - For men a nonlinear relationship between reported earnings and true earnings existed but for women the relationship was linear;
  - Response error in income cannot be treated as additive white noise because of its relationship with gender and true earnings;
  - Measurement error is not related to age, education, and weeks worked.

Example 4 - Imaging Data

Prostate Cancer (e.g., Ward et al. 2012)

- It is the second most common type of cancer in men.
- Early diagnosis and confirmation of prostate cancer is crucial.

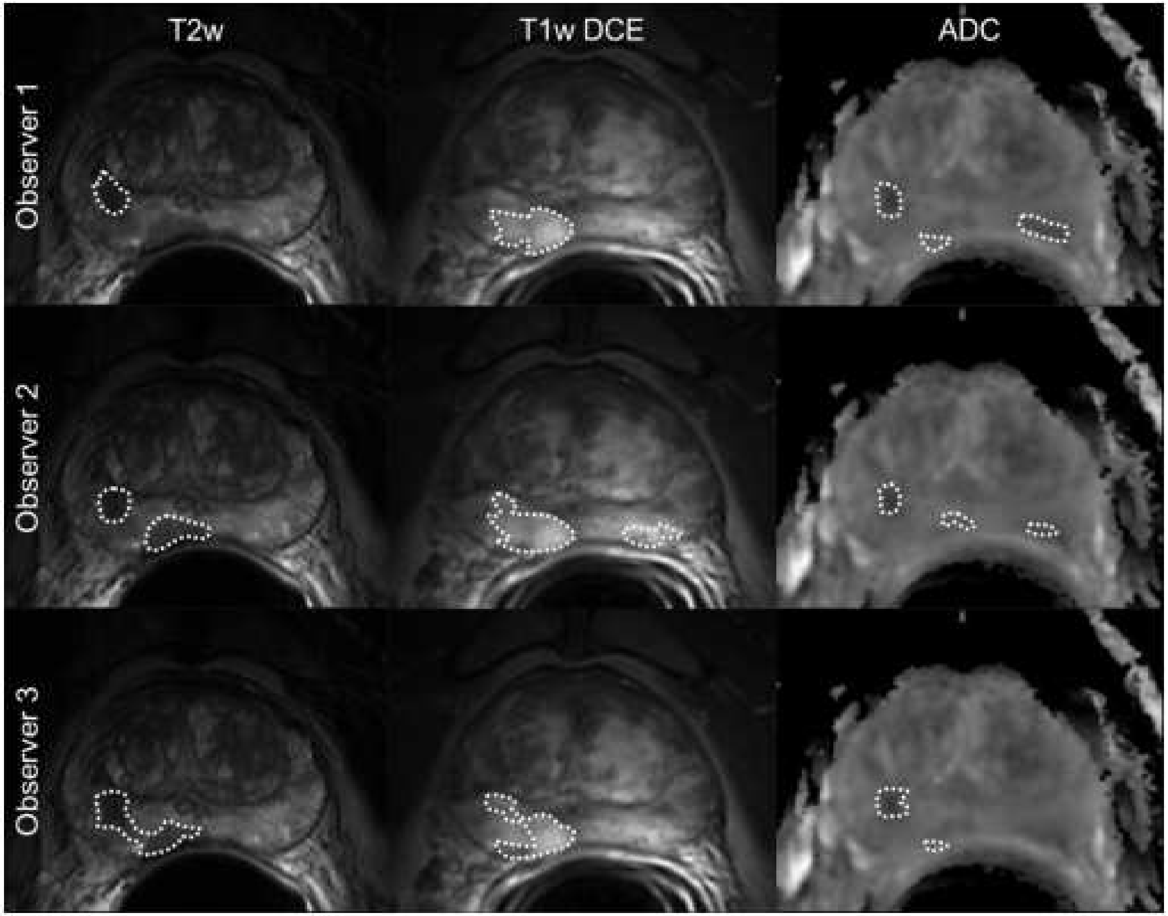
Diagnosis of Prostate Cancer

- Screening: Prostate-Specific Antigen (PSA)
- Diagnosis of prostate cancer
  - imaging contour
  - biopsy confirmation: 2D ultrasound guided biopsy

Issues

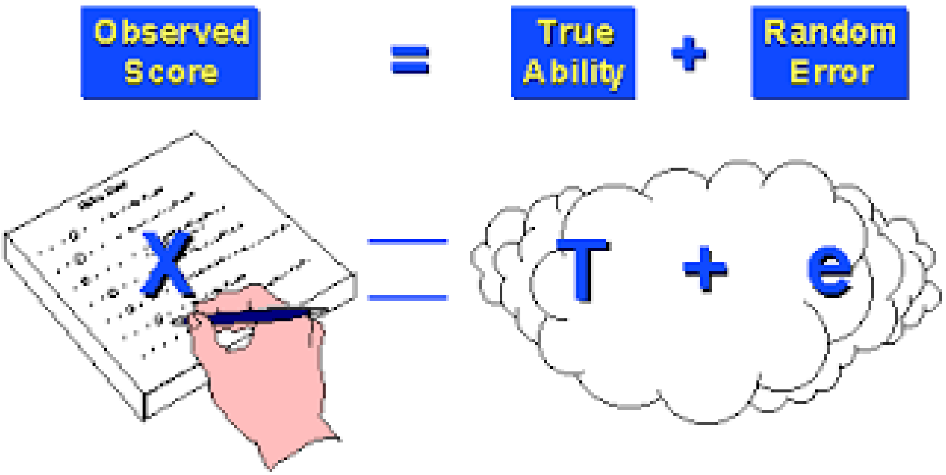
- The diagnosis process involves substantial variations.
  - The image diagnosis procedure depends on both doctors' experiences and the types of images.
  - Biopsy conformation: 2D ultrasound may not guide the needle to the right position precisely.

Example 4 - Imaging Data



Other Examples

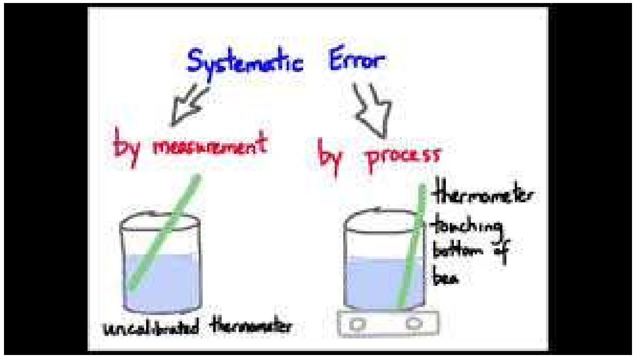
- measuring radiation dose
- measuring exposure to arsenic in drinking water, dust in the workplace, radon gas in the home and other environmental hazards
- The study of diet and disease has been a major motivation for studying measurement error problems.
  - In these studies, it is typical to measure diet via a self-report instrument, for example, a food frequency questionnaire (FFQ), or a 24-hour recall interview. It has been shown that these self-report instruments are only imperfect measures of long-term diary intakes, and hence that measurement error is a major concern.



Some Sources of Measurement Error (Yi 2017)

- Measurement error may refer to random noise, sampling error, or uncertainty/variation in the measuring process.
- Flawed or mismanaged data collection procedures result in imprecise measurements.
- Variables are not accurately measured due to reporting errors for sensitive questions.
- Variables are impossible to measure precisely.
- Variables represent averages of certain quantities over time.
- Variables may be manipulated artificially.
- Variables are too expensive or time consuming to measure precisely.

Measurement Error= reading error + biological variability + sampling error + others



Impact of Ignoring Measurement Error

Example 1 - Attenuation Effects

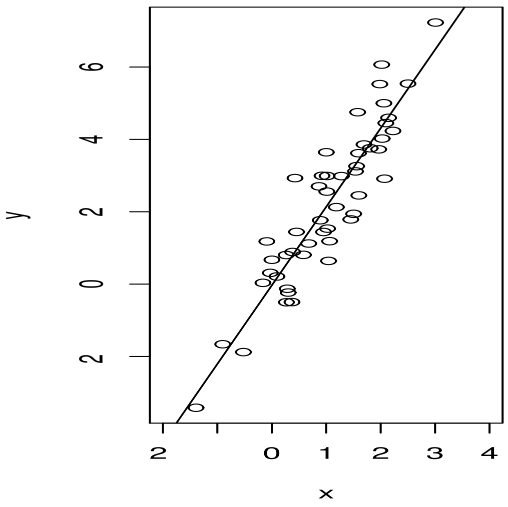
- Simple Linear Regression (Fuller 1987)

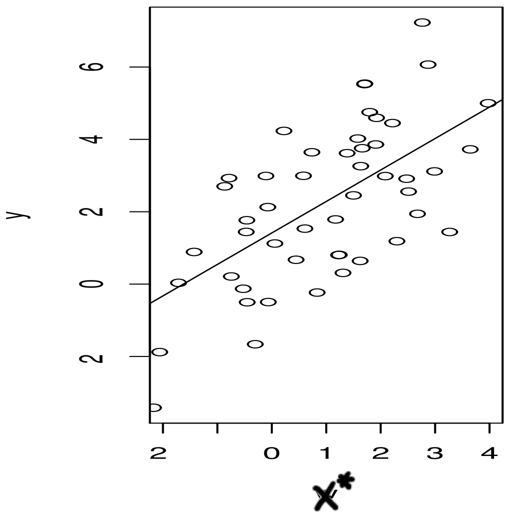
  $Y = \beta_0 + \beta_x X + \epsilon$ with $X \sim (\mu_x, \sigma_x^2)$, $\epsilon \sim (0, \sigma^2)$, indep.

  $X^* = X + e$ with $e \sim (0, \sigma_e^2)$

  If naively replacing X with X^*, then

  - $\beta_x^* = \left(\frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2}\right)\beta_x \quad \Rightarrow$ attenuation effect !
  - $\text{var}(Y|X^*) = \sigma^2 + \frac{\beta_x^2\sigma_e^2\sigma_x^2}{\sigma_x^2+\sigma_e^2}$





Example 2 - Opposite Evidence

Red Meat Intake in Relation to Colorectal Cancer Incidence in the Health Professionals Follow-up Study (HPFS)

- 49980 male health professionals who were free of cancer in 1986 were followed up to 2010 for colorectal cancer incidence. During this study period, 1281 individuals developed colorectal cancer.
- Outcome: $Y = 1$ if the participant has colorectal cancer and 0 otherwise
- Covariates:
  Z: precisely measured covariates
  X: binary indicator of red meat intake at the baseline
  - whether red meat intake was greater than 2 servings/week,

  - main study: X was assessed by the FFQ $\Longrightarrow X^*$
  - validation subsample: X was obtained from the DR $\Longrightarrow X$

- Estimation of specificity and sensitivity:

$$\widehat{P}(X^* = 0|X = 0) = 0.85$$
$$\widehat{P}(X^* = 1|X = 1) = 0.84$$

Example 2 - Opposite Evidence

Analysis

- Logistic Regression Model: $\mathrm{logit}\, P(Y = 1|X, Z) = \beta_0 + \beta_x X + \beta_z^{\mathsf{T}} Z$

- Inference Results (Yi, Yan, Liao and Spiegelman 2018)
  - Method 1: indicates moderate evidence for an increase of colorectal cancer risk in relation to red meat intake
  - Method 2: finds no evidence that red meat intake is associated with colorectal cancer

Example 2 - Opposite Evidence

Analysis

- Logistic Regression Model: $\mathrm{logit}\, P(Y = 1|X, Z) = \beta_0 + \beta_x X + \beta_z^{\mathsf{T}} Z$
- Inference Results (Yi, Yan, Liao and Spiegelman 2018)
  - Method 1: indicates moderate evidence for an increase of colorectal cancer risk in relation to red meat intake
  - Method 2: finds no evidence that red meat intake is associated with colorectal cancer

Remark

- Method 1 ignores the feature of measurement error in the data.
- Method 2 accommodates measurement error effects in inferential procedures.

Impact of Measurement Error

Remarks (Carroll et al. 2006; Yi 2017)

- Measurement error in covariates may
  - change the structure of the response model
  - cause bias in parameter estimation
  - lead to a loss of power for detecting interesting relationship among variables
  - mask the features of the data

- The effects of measurement error are very complex, depending on the form of
  - the inference method
  - the measurement error model
  - the response model
  - the association of the covariates

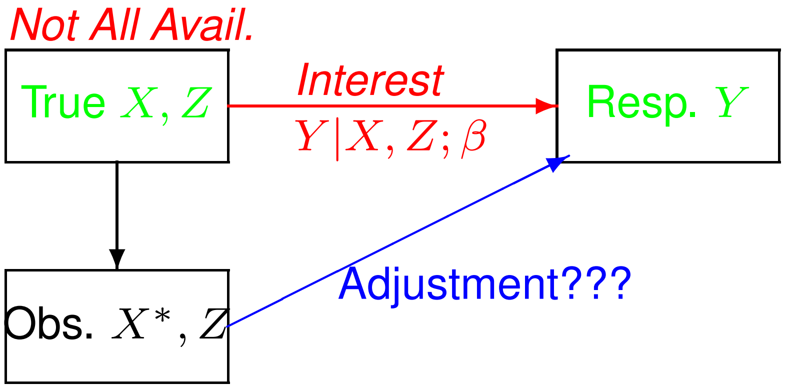
Framework of Handling Measurement Error

General Classification

Research on measurement error models may be categorized into three areas:

- Measurement Error / Misclassification in Covariates
- Measurement Error / Misclassification in Response
- Measurement Error / Misclassification in Covariates and Response

Objective



Traditional Analysis

- build a model between Y and X, Z: $f(y|x, z)$
- conditional analysis is often employed:
  e.g., regression $Y = \beta_0 + \beta_x^{\mathsf{T}} X + \beta_z^{\mathsf{T}} Z + \epsilon$
  no distributional assumption is made for covariates

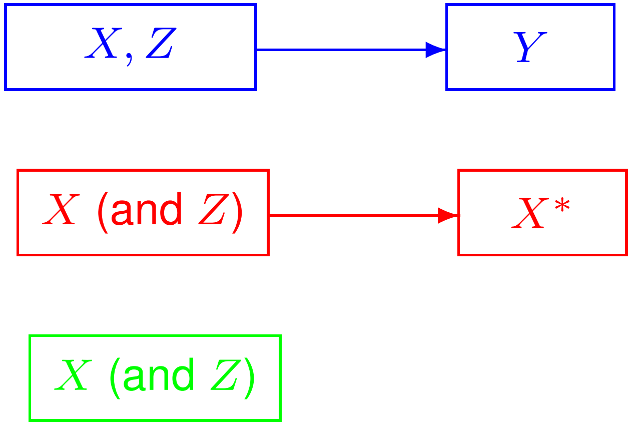
With Measurement Error

$$f(y, x, x^*, z) = f(x^*|y, x, z) f(y|x, z) f(x, z)$$

- do we need to worry how X and X^* are related?
- do we need to model distributions for covariates?
- what are the induced challenges and complications?

Modeling Strategy

- Structural Method: assuming a distribution of X

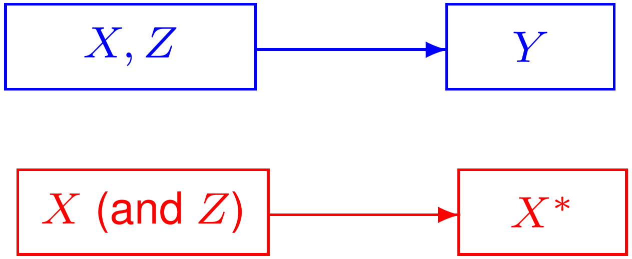


1. Model X, Z → Y

2. Model X (and Z) → X^*

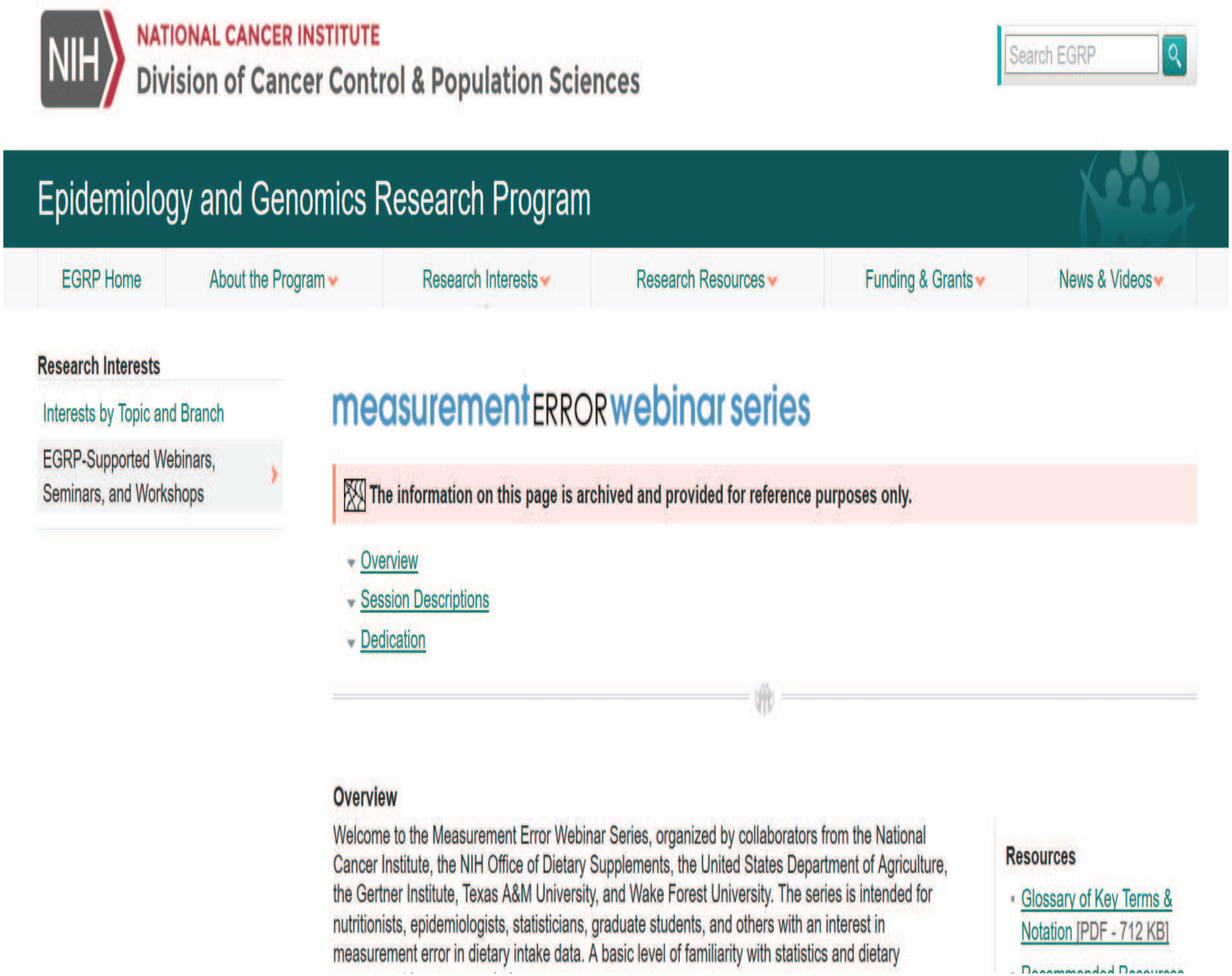
3. Model X (and Z)

- Functional Method: not assuming a distribution of X



1. Model X, Z → Y

2. Model X (and Z) → X^*


NIH
NATIONAL CANCER INSTITUTE
Division of Cancer Control & Population Sciences
Search EGRP
Epidemiology and Genomics Research Program
EGRP Home
About the Program
Research Interests
Research Resources
Funding & Grants
News & Videos
Research Interests
Interests by Topic and Branch
EGRP-Supported Webinars, Seminars, and Workshops

measurementERRORwebinar series

The information on this page is archived and provided for reference purposes only.

- Overview
- Session Descriptions
- Dedication

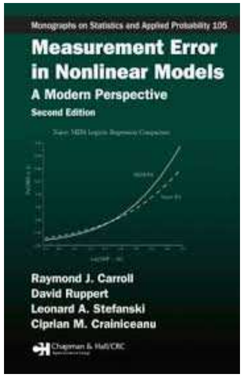
Overview

Welcome to the Measurement Error Webinar Series, organized by collaborators from the National Cancer Institute, the NIH Office of Dietary Supplements, the United States Department of Agriculture, the Gertner Institute, Texas A&M University, and Wake Forest University. The series is intended for nutritionists, epidemiologists, statisticians, graduate students, and others with an interest in measurement error in dietary intake data. A basic level of familiarity with statistics and dietary

Resources

- Glossary of Key Terms & Notation [PDF - 712 KB]
- Recommended Resources

Some Research Monographs


Monographs on Statistics and Applied Probability 105

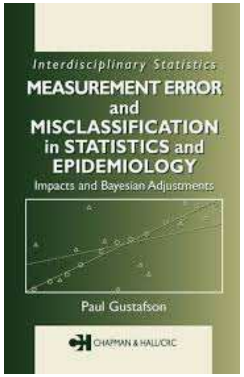
Measurement Error in Nonlinear Models

A Modern Perspective

Second Edition

Raymond J. Carroll
David Ruppert
Leonard A. Stefanski
Ciprian M. Crainiceanu

Chapman & Hall/CRC

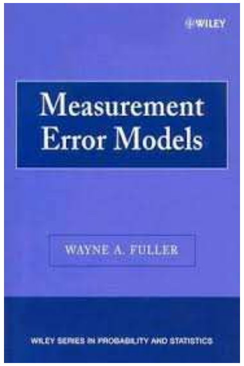

Interdisciplinary Statistics

MEASUREMENT ERROR and MISCLASSIFICATION in STATISTICS and EPIDEMIOLOGY

Impacts and Bayesian Adjustments

Paul Gustafson

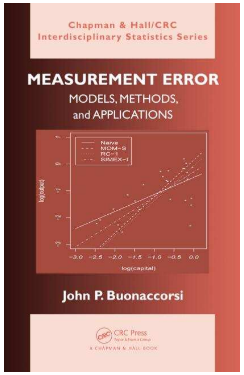
CHAPMAN & HALL/CRC


WILEY

Measurement Error Models

WAYNE A. FULLER

WILEY SERIES IN PROBABILITY AND STATISTICS

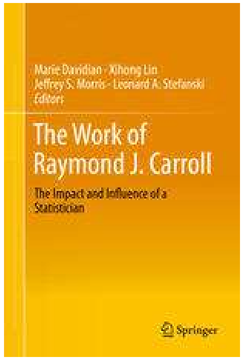

Chapman & Hall/CRC Interdisciplinary Statistics Series

MEASUREMENT ERROR

MODELS, METHODS, and APPLICATIONS

John P. Buonaccorsi

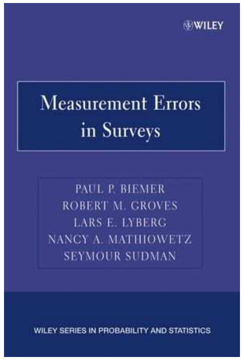
CRC Press
A CHAPMAN & HALL BOOK


Marie Davidian · Xihong Lin
Jeffrey S. Morris · Leonard A. Stefanski
Editors

The Work of Raymond J. Carroll

The Impact and Influence of a Statistician

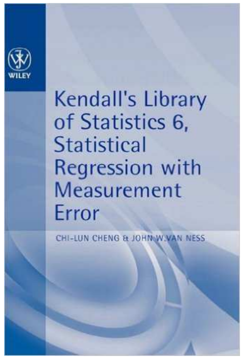
Springer


WILEY

Measurement Errors in Surveys

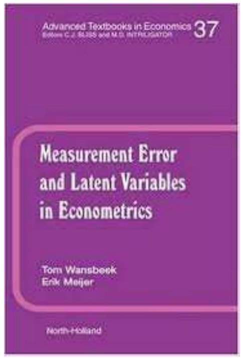
PAUL P. BIEMER
ROBERT M. GROVES
LARS E. LYBERG
NANCY A. MATHIOWETZ
SEYMOUR SUDMAN

WILEY SERIES IN PROBABILITY AND STATISTICS


WILEY

Kendall's Library of Statistics 6, Statistical Regression with Measurement Error

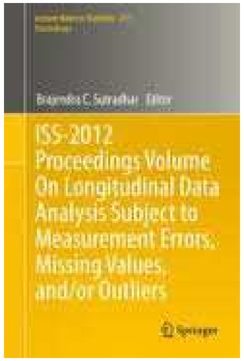
CHI-LUN CHENG & JOHN W.VAN NESS


Advanced Textbooks in Economics 37
Editors C.J. BLISS and M.D. INTRILIGATOR

Measurement Error and Latent Variables in Econometrics

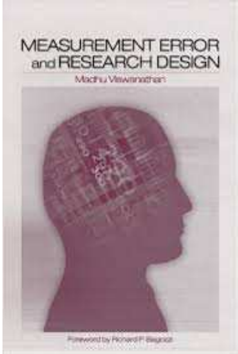
Tom Wansbeek
Erik Meijer

North-Holland


Brajendra C. Sutradhar Editor

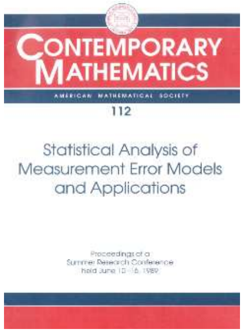
ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers

Springer


MEASUREMENT ERROR and RESEARCH DESIGN

Madhu Viswanathan

Foreword by Richard P. Bagozzi

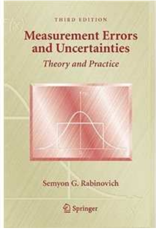

CONTEMPORARY MATHEMATICS

AMERICAN MATHEMATICAL SOCIETY

112

Statistical Analysis of Measurement Error Models and Applications

Proceedings of a Summer Research Conference held June 10–16, 1989

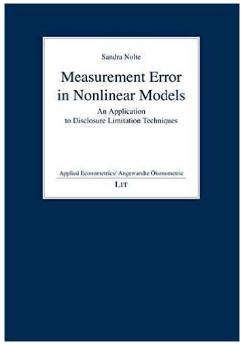

THIRD EDITION

Measurement Errors and Uncertainties

Theory and Practice

Semyon G. Rabinovich

Springer

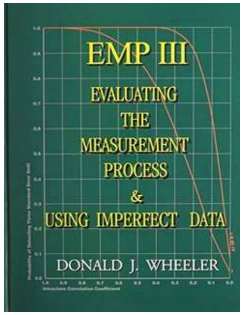

Sandra Nolte

Measurement Error in Nonlinear Models

An Application to Disclosure Limitation Techniques

Applied Econometrics/ Angewandte Ökonometrie

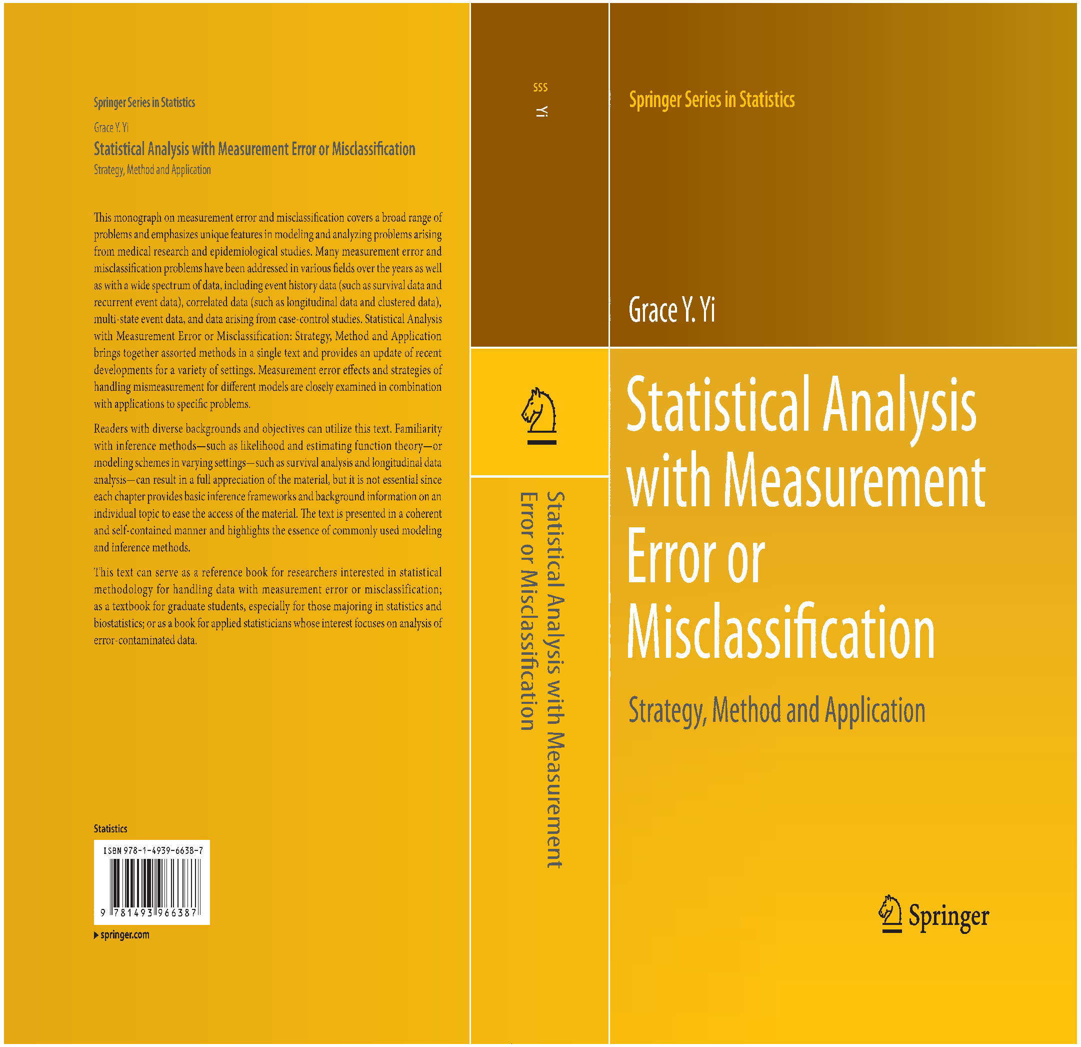
LIT


EMP III

EVALUATING THE MEASUREMENT PROCESS & USING IMPERFECT DATA

DONALD J. WHEELER

Shameless Promotion



Springer Series in Statistics

Grace Y. Yi

Statistical Analysis with Measurement Error or Misclassification

Strategy, Method and Application

This monograph on measurement error and misclassification covers a broad range of problems and emphasizes unique features in modeling and analyzing problems arising from medical research and epidemiological studies. Many measurement error and misclassification problems have been addressed in various fields over the years as well as with a wide spectrum of data, including event history data (such as survival data and recurrent event data), correlated data (such as longitudinal data and clustered data), multi-state event data, and data arising from case-control studies. Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application brings together assorted methods in a single text and provides an update of recent developments for a variety of settings. Measurement error effects and strategies of handling mismeasurement for different models are closely examined in combination with applications to specific problems.

Readers with diverse backgrounds and objectives can utilize this text. Familiarity with inference methods—such as likelihood and estimating function theory—or modeling schemes in varying settings—such as survival analysis and longitudinal data analysis—can result in a full appreciation of the material, but it is not essential since each chapter provides basic inference frameworks and background information on an individual topic to ease the access of the material. The text is presented in a coherent and self-contained manner and highlights the essence of commonly used modeling and inference methods.

This text can serve as a reference book for researchers interested in statistical methodology for handling data with measurement error or misclassification; as a textbook for graduate students, especially for those majoring in statistics and biostatistics; or as a book for applied statisticians whose interest focuses on analysis of error-contaminated data.

Statistics

ISBN 978-1-4939-6638-7

9 781493 966387

▸ springer.com

SSS

Yi

Statistical Analysis with Measurement Error or Misclassification

Springer Series in Statistics

Grace Y. Yi

Statistical Analysis with Measurement Error or Misclassification

Strategy, Method and Application

Springer

Summary and Take Home Messages

- When conducting statistical analysis,
  - the validity of modeling is a serious concern
  - on equal footing, the quality of data should also be taken into consideration
- While measurement error arises ubiquitously in applications and has been a longstanding concern in various fields, measurement error methods have not been always (in fact, not frequently at all) used in the situations that merit their use.
- There is the increasing use of data (e.g., EHR/EMR data, large scale administrative data) not originally designed for answering a scientific question, developing methods of handling faulty data becomes increasingly important.
- While we may attempt to collect good quality data by careful design, measurement error is inevitable. Understanding and correcting for measurement error effects are critical in conducting sensible data analysis.

Part 2: Noisy Data - Missing Value

Missing Data: Sources and Impact

Missing data arise in many applications, such as

- longitudinal studies
- survey sampling
- survival data
- clinical trials
- experimental design
- cancer research
- environmental studies
- ...

Missing data effects

- may yield seriously biased analysis results
- depend on missing data mechanisms
- depend on analysis methods

Handling Missing Value and Measurement Error Separately

- Comparisons from an Example

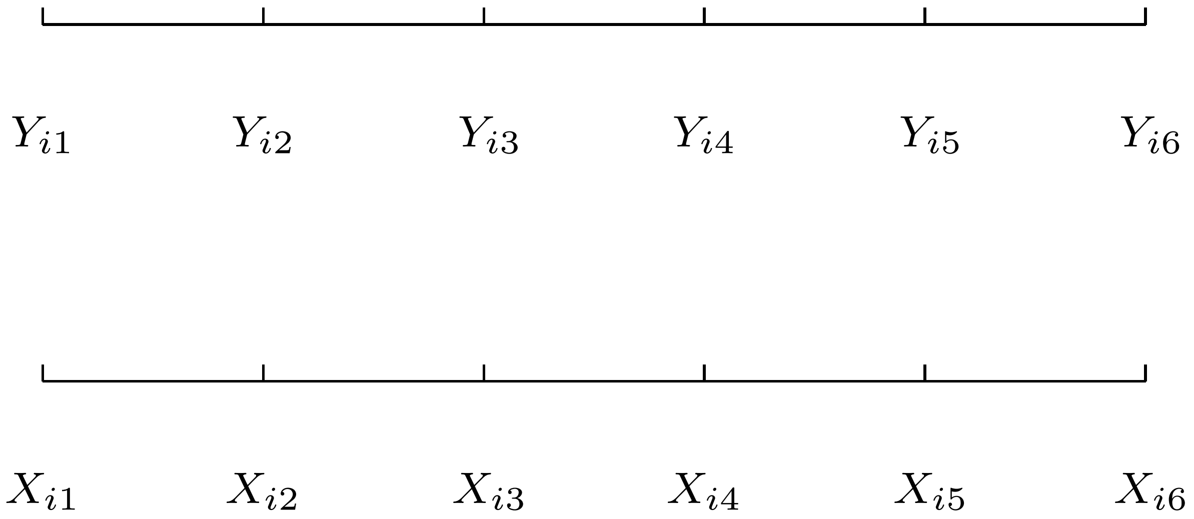
“Ideal” Longitudinal Data

Longitudinal Study

- individuals are followed over time
- a response Y with covariates X is recorded at each assessment

Data Features

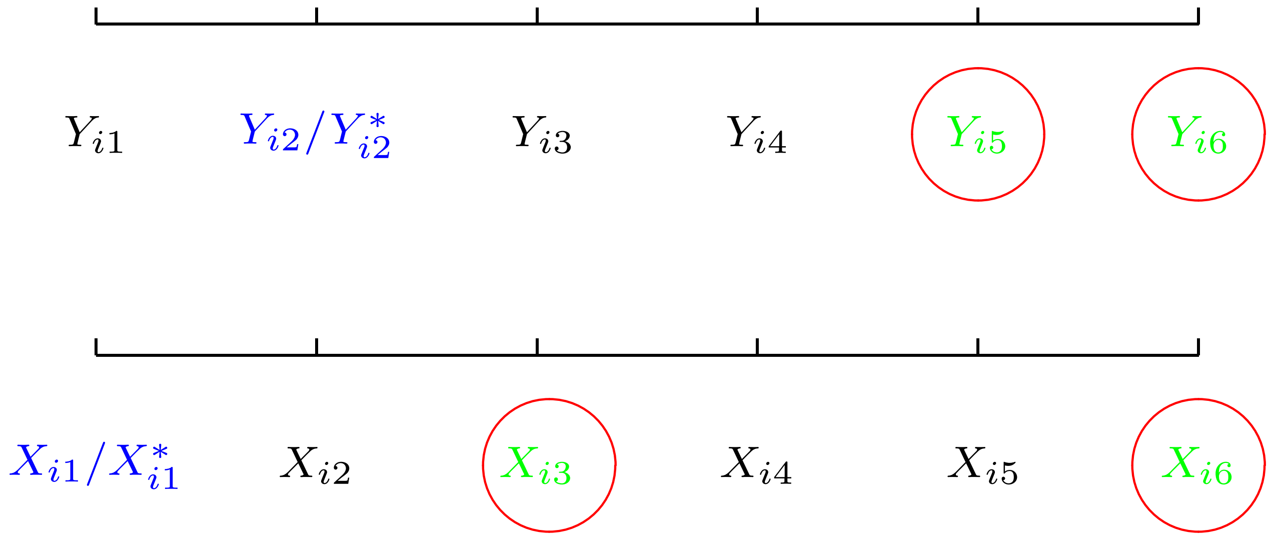
correlation among repeated measurements



Y_{i1} Y_{i2} Y_{i3} Y_{i4} Y_{i5} Y_{i6}

X_{i1} X_{i2} X_{i3} X_{i4} X_{i5} X_{i6}

Common Challenges

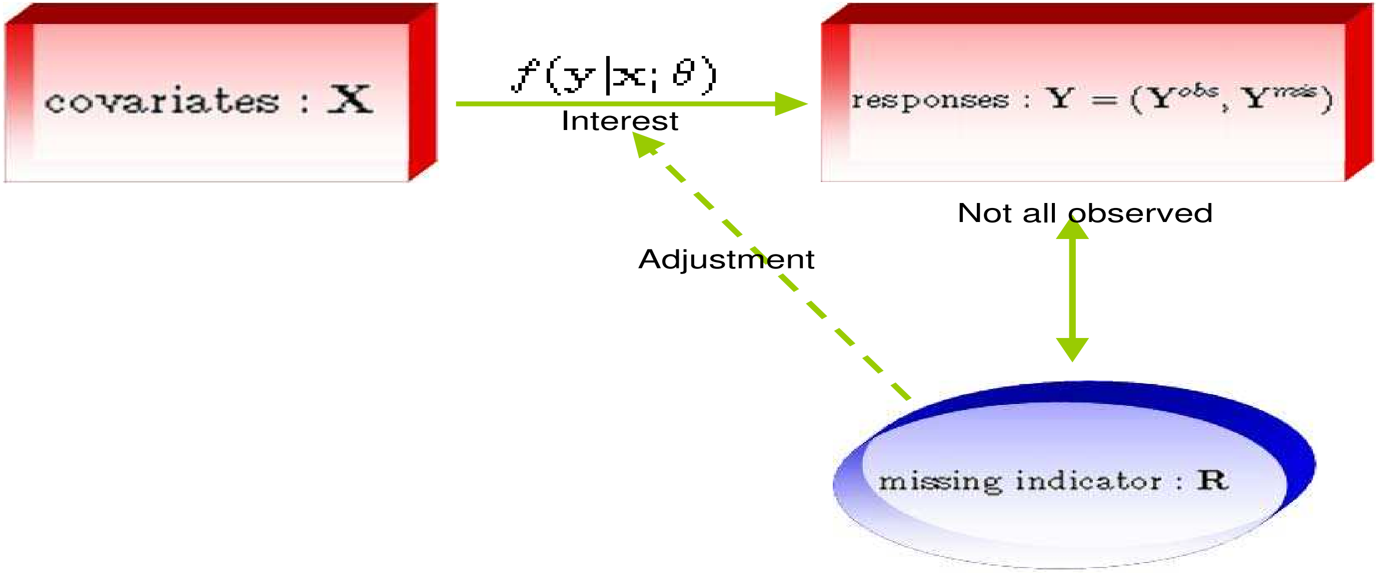


Y_{i1} Y_{i2}/Y_{i2}^* Y_{i3} Y_{i4} Y_{i5} Y_{i6}

X_{i1}/X_{i1}^* X_{i2} X_{i3} X_{i4} X_{i5} X_{i6}

- missing observations
  - missingness in response: R_{ij}^y - missingness indicator for Y_{ij}
  - missingness in covariates: R_{ij}^x - missingness indicator for X_{ij}
- measurement error
  - error in response: Y_{ij}^* - a surrogate/observed version of Y_{ij}
  - error in covariates: X_{ij}^* - a surrogate/observed version of X_{ij}

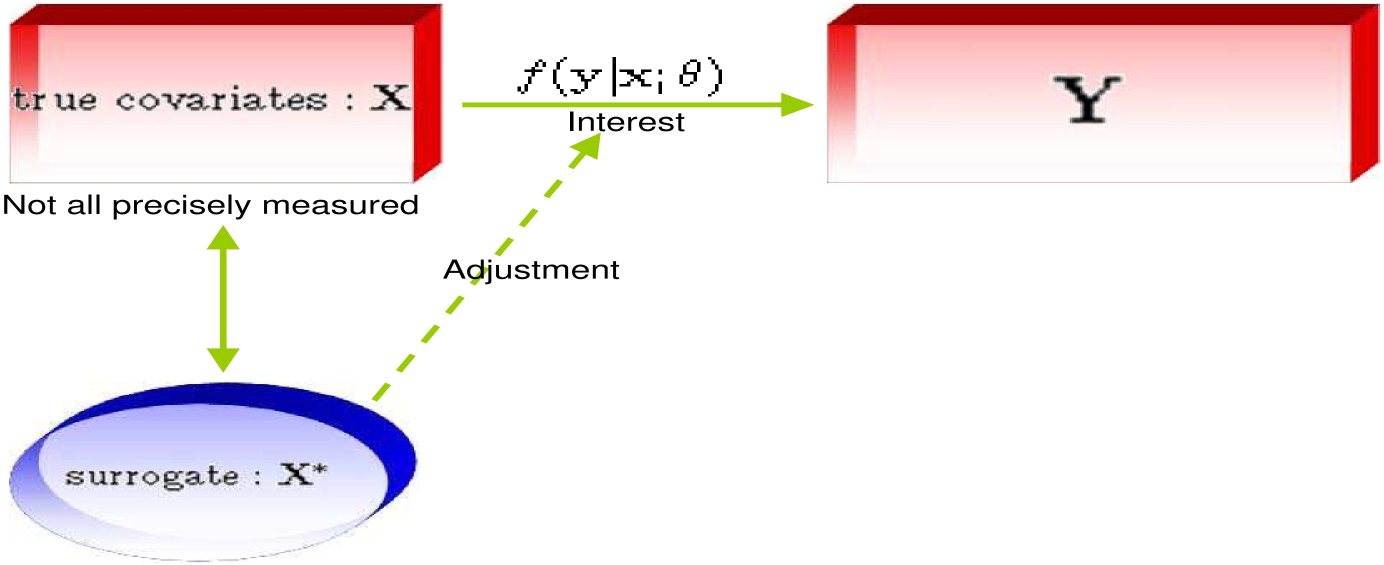
Accounting for Response Missingness Only


covariates : $\mathbf{X}$
$f(y|x;\theta)$
Interest
responses : $\mathbf{Y} = (\mathbf{Y}^{obs}, \mathbf{Y}^{mis})$
Not all observed
Adjustment
missing indicator : $\mathbf{R}$

Inference Framework

$$f(y, x, r)$$

Accounting for Covariate Error Only



Inference Framework

$$f(y, x, x^*)$$

Comparisons

Response Missingness

$$\begin{aligned} f(y,x,r) &= f(y,x|r)f(r) \\ &= f(y|x,r)f(x,r) \end{aligned}$$

Comparisons

Response Missingness

$$f(y,x,r) = f(y,x|r)f(r)$$

$$= f(y|x,r)f(x,r)$$

Covariate Measurement Error

$$f(y,x,x^*) = f(y,x|x^*)f(x^*)$$

$$= f(y|x)f(x,x^*)$$

assume: $y \perp x^*|x$

Accounting for Covariate Error Only

Response Missingness

$$f(y,x,r) = f(y,x|r)f(r)$$

$$= f(y|x,r)f(x,r)$$

- Pattern Mixture Model

Covariate Measurement Error

$$f(y,x,x^*) = f(y,x|x^*)f(x^*)$$

$$= f(y|x)f(x,x^*)$$

assume: $y \perp x^* | x$

- Nondifferential Measurement Error Mechanism

Accounting for Covariate Error Only

Response Missingness

$$f(y,x,r) = f(y,x|r)f(r)$$

$$= f(y|x,r)f(x,r)$$

- Pattern Mixture Model

Covariate Measurement Error

$$f(y,x,x^*) = f(y,x|x^*)f(x^*)$$

$$= f(y|x)f(x,x^*)$$

assume: $y \perp x^*|x$

- Nondifferential Measurement Error Mechanism

Accounting for Covariate Error Only

Response Missingness

$$f(y,x,r) = f(y,x|r)f(r)$$
$$= f(y|x,r)f(x,r)$$

- Pattern Mixture Model

- Alternative: Selection Model

$$f(y,x,r) = f(r|y,x)f(y,x)$$
$$= f(r|y,x)f(y|x)f(x)$$

Covariate Measurement Error

$$f(y,x,x^*) = f(y,x|x^*)f(x^*)$$
$$= f(y|x)f(x,x^*)$$

assume: $y \perp x^*|x$

- Nondifferential Measurement Error Mechanism

Accounting for Covariate Error Only

Response Missingness

$$f(y,x,r) = f(y,x|r)f(r)$$
$$= f(y|x,r)f(x,r)$$

- Pattern Mixture Model

- Alternative: Selection Model

$$f(y,x,r) = f(r|y,x)f(y,x)$$
$$= f(r|y,x)f(y|x)f(x)$$

Covariate Measurement Error

$$f(y,x,x^*) = f(y,x|x^*)f(x^*)$$
$$= f(y|x)f(x,x^*)$$

assume: $y \perp x^*|x$

- Nondifferential Measurement Error Mechanism

- No Assumption: Differential Error Mechanism

$$f(y,x,x^*) = f(x^*|y,x)f(y,x)$$
$$= f(x^*|y,x)f(y|x)f(x)$$

Accounting for Covariate Error Only

Response Missingness

$$f(y,x,r) = f(y,x|r)f(r)$$
$$= f(y|x,r)f(x,r)$$

- Pattern Mixture Model

- Alternative: Selection Model

$$f(y,x,r) = f(r|y,x)f(y,x)$$
$$= f(r|y,x)f(y|x)f(x)$$

- Missing Data Mechanism:

$$f(r|y,x) \overset{MNAR}{==} f(r|y^{mis},y^{obs},x)$$
$$f(r|y,x) \overset{MAR}{==} f(r|y^{obs},x)$$
$$f(r|y,x) \overset{MCAR}{==} f(r|x)$$

Covariate Measurement Error

$$f(y,x,x^*) = f(y,x|x^*)f(x^*)$$
$$= f(y|x)f(x,x^*)$$

assume: $y \perp x^*|x$

- Nondifferential Measurement Error Mechanism

- No Assumption: Differential Error Mechanism

$$f(y,x,x^*) = f(x^*|y,x)f(y,x)$$
$$= f(x^*|y,x)f(y|x)f(x)$$

Western

Comparisons

- Both contexts often assume parameter distinctness
- parameter identifiability could be an issue
- Mechanism is classified based on the relationship between response variable and the variable characterizing the specific feature:

Response Missingness	Covariates Error
MCAR : $R \perp (Y^{mis}, Y^{obs}) \mid X$	nondifferential : $Y \perp X^{*} \mid X$
MAR : $R \perp Y^{mis} \mid (Y^{obs}, X)$	
MNAR/NMAR : $R \not\perp Y^{mis} \mid (Y^{obs}, X)$	differential : $Y \not\perp X^{*} \mid X$

Handling Missing Value and Measurement Error Simultaneously

- Two Examples

Example 1: Longitudinal Data

What if both measurement error and missing observations are present in data?

- Example: Yi, Ma and Carroll (2011) examined a data set from Continuing Survey of Food Intake by Individuals.
  - The data set consists of repeated measurements for 1,737 individuals with 24-hour recall food intake interviews taken on four different days.
  - Information on age, vitamin A intake, vitamin C intake, total fat intake and total calorie intake is collected at each interview.
  - Goal: understand the relationship among the variables

- Findings
  - The consequence of ignoring the measurement error is attenuation of the covariate effects towards zero.
  - Ignoring the missingness results in slightly overestimating the covariate effects.

Example 1: Longitudinal Data

Question

What are covariate measurement error effects on missing response mechanisms?

Theorem (Yi, Ma and Carroll 2012)

If

$$P(R_{ij}=1 \mid \tilde{R}_{ij}, Y_i, X_i, Z_i) = P(R_{ij}=1 \mid \tilde{R}_{ij}, Y_i^{obs}, X_i, Z_i)$$

is true, it is not necessarily true that

$$P(R_{ij}=1 \mid \tilde{R}_{ij}, Y_i, X_i^*, Z_i) = P(R_{ij}=1 \mid \tilde{R}_{ij}, Y_i^{obs}, X_i^*, Z_i)$$

holds.

Message

the missingness process could be missing at random in X_i,
but not missing at random in X_i^*

Example 2: Causal Inference

Question: What is the interplay of measurement error and missingness?

Example (Lee et al. 2013)

- Study on the effectiveness of a perioperative smoking cessation program:
  - One hundred sixty-eight patients were equally randomized to either the treatment group or the control group, where the treatment group was assigned to the smoking cessation intervention and the control group received standard care.
- Data
  - Outcome:
    the smoking cessation status of a patient for previous 7 days at the 30-day follow-up postoperatively
  - Baseline Covariates:
    gender, age, body mass index, diabetes status, hypertension, chronic obstructive pulmonary disease (COPD), cigarettes per day, the number of years of smoking, and the exhaled carbon monoxide (CO) level.

Example 2: Causal Inference

Interest

quantifying the average treatment effect (ATE) of the intervention on quitting smoking

$$\tau_0 = E(Y_1) - E(Y_0)$$

Y_1: potential outcome that would have been observed had the subject been treated

Y_0: potential outcome that would have been observed had the subject been untreated

Remark

- Estimation of τ_0 cannot be obtained directly based on available measurements of outcome variables.
- Using the observed data allows us to estimate the difference of conditional mean outcomes between the treated and untreated groups, $E(Y|T=1) - E(Y|T=0)$, which differs from ATE τ_0 because of possible imbalance of X in the treated and untreated groups.

Example 2: Causal Inference

Propensity Score (e.g., Rosenbaum and Rubin 1983)

using the propensity score

$$e = P(T = 1|X)$$

to balance the distribution of X for the treated and untreated groups

Available Data

For subject i in a sample of size n:

X_i : pre-treatment covariates

T_i : observed binary treatment indicator

Y_i : observed outcome

Consistent Estimator (Rosenbaum 1998)

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\widehat{e}_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\widehat{e}_i}$$

Key assumptions: Y_i must be precisely measured and be observed!

Example 2: Causal Inference

Challenges of Noisy Data

- Missing Value: About 10.7% subjects have missing outcome measurements.
- Misclassification:
  Among those patients with complete observations, about 7.5% subjects misreport outcome measurements.

Naive Methods: Y is subject to missingness and misclassification.

- Method 1: Ignoring both misclassification and missingness $\Rightarrow$ $\hat{\tau}^{**}$
- Method 2: Ignoring missingness but accommodating misclassification effects $\Rightarrow$ $\hat{\tau}^{*}$
- Method 3: Ignoring misclassification but accommodating missingness effects $\Rightarrow$ $\tilde{\tau}^{*}$

Example 2: Causal Inference

Theorem (Shu and Yi 2019): when $p_{11} \neq p_{10}$,

$$\text{Bias}(\hat{\tau}^{**}) = (p_{11} - p_{10})\text{Bias}(\hat{\tau}^{*}) + \text{Bias}(\tilde{\tau}^{*})$$

where $p_{1k} = P(Y^* = 1|Y = k)$ for $k = 0, 1$

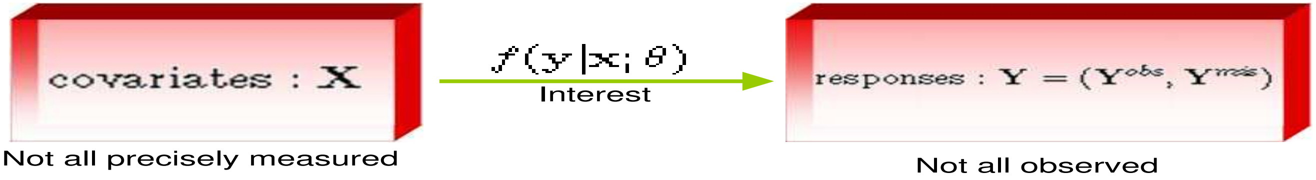
In terms of the absolute magnitude,

$$|\text{Bias}(\hat{\tau}^{**})| \leq |\text{Bias}(\hat{\tau}^{*})| + |\text{Bias}(\tilde{\tau}^{*})|$$

Remark

- It is possible that $|\text{Bias}(\hat{\tau}^{**})|$ can be smaller than either $|\text{Bias}(\hat{\tau}^{*})|$ or $|\text{Bias}(\tilde{\tau}^{*})|$.
- There are counter-intuitive situations where simultaneously ignoring both missingness and misclassification can perform better than merely ignoring one feature.

Accounting for 2 Features



covariates : $\mathbf{X}$

Not all precisely measured

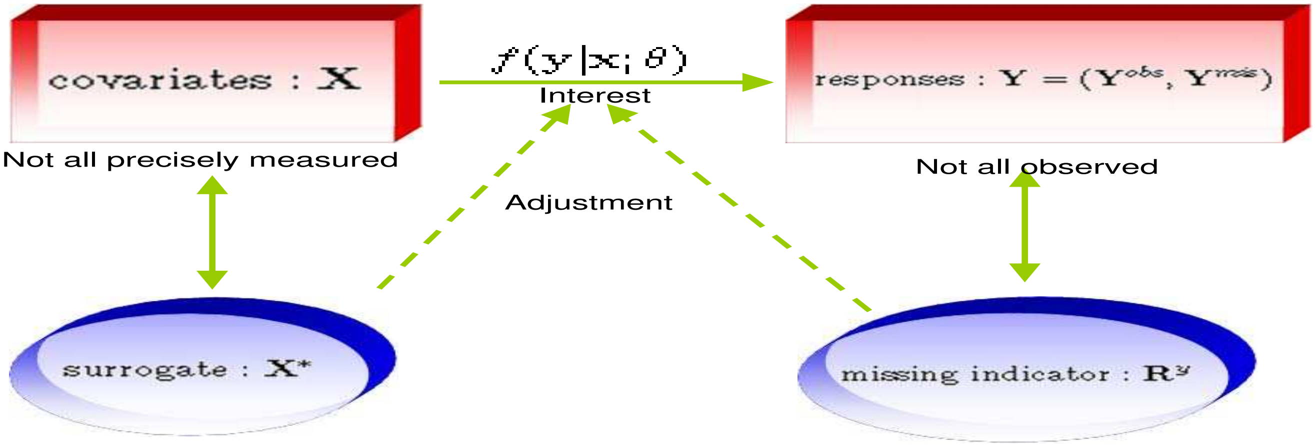
$f(y|\mathbf{x};\theta)$

Interest

responses : $\mathbf{Y} = (\mathbf{Y}^{obs}, \mathbf{Y}^{mis})$

Not all observed

Accounting for 2 Features


covariates : $\mathbf{X}$
Not all precisely measured
$f(y|x;\theta)$
Interest
responses : $\mathbf{Y} = (\mathbf{Y}^{obs}, \mathbf{Y}^{mis})$
Not all observed
Adjustment
surrogate : $\mathbf{X}^*$
missing indicator : $\mathbf{R}^y$

Inference Framework

$$f(y, x, x^*, r)$$

R Package “swgee" on CRAN

J. Xiong and G. Y. Yi (2019). swgee: An R Package for Analyzing Longitudinal Data with Response Missingness and Covariate Measurement Error. *The R Journal*, 11:1.

Summary and Take Home Messages

- There is certain similarity in handling data with measurement error alone and data with missing observations alone:
  - Extra modeling of a nuisance process is often needed:
    nonidentifiability is a typical concern.
  - Suitable modeling assumptions/mechanisms are usually employed:
    simplicity/generality/testability.

- When both measurement error and missing observations are present, the induced impacts are more complex than the effects induced from one feature.
  - These two features may interact in complex manners.
  - The effects can be counterintuitive:
    ignoring both features may sometimes better than ignoring one feature.

- Valid inference methods commonly requires introducing additional modeling for two nuisance processes:
  - the measurement error process
  - the missing data process

Part 3: Concurrent Features

measurement error, missing values, high dimensionality

Part 3

High Dimensionality and Irrelevant Measurements

High Dimensionality: $p \geq n$

- Contemporary technologies enable us to collect data of large volume and rich information.
- Common examples:
  - In disease classification using gene expression data, the number of gene expression profiles is in the order of tens of thousands.
  - In the study of protein-protein interactions, the number of features can be in the order of millions.
- Traditional methods break down, and dimensionality reduction is imperative.
  - feature screening
  - variable selection

High Dimensionality and Irrelevant Measurements

High Dimensionality: $p \geq n$

- Contemporary technologies enable us to collect data of large volume and rich information.
- Common examples:
  - In disease classification using gene expression data, the number of gene expression profiles is in the order of tens of thousands.
  - In the study of protein-protein interactions, the number of features can be in the order of millions.
- Traditional methods break down, and dimensionality reduction is imperative.
  - feature screening
  - variable selection

Challenge

- The impact of ignoring measurement error/missing values can be a lot more striking for big data than for small data: THE BIGGER, THE WORSE!

Introductory Example

NPHS Data

- The National Population Health Survey (NPHS) is a longitudinal study that was conducted every other year, beginning in 1994/1995.
- The questions for the NPHS include
  - many aspects of in-depth health information such as health status, use of health services, chronic conditions and activity restrictions
  - social background questions such as age, sex and income level, etc

Objective

understand how population health is influenced by multiple risk factors

Features

- Some variables, such as HUI, and INC, are subject to measurement error.
- Missing observations are present:
  14.1%, 24.7%, 37.4%, and 46.6% at cycles 2, 3, 4 and 5, respectively
- The dimension of the variables is high: some variables are unimportant.

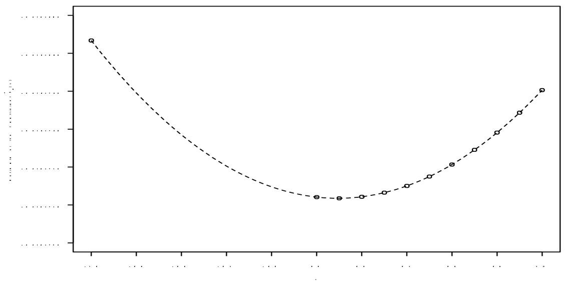
Yi, Tan and Li (2015): Key Ideas

Method of Yi, Tan and Li (2015)

- sequentially correct for biases induced from
  - missingness: re-weigh contributions from subjects by missingness probabilities
  - error: use simulation to portray the impact of varying degree of error
- simultaneously perform

  - estimation

  - variable selection



Remarks and Questions

Questions:

- Account for 3 features: need three additional processes
- Account for 4 features: need four additional processes
- ......

- Do we need to go that far?
- How to balance the complexity of modeling and interpretability of model parameters?
- Issues of model identifiability, model checking, sensitivity analyses, etc.?
- Develop inferential procedures that are statistically valid as well as computationally manageable?

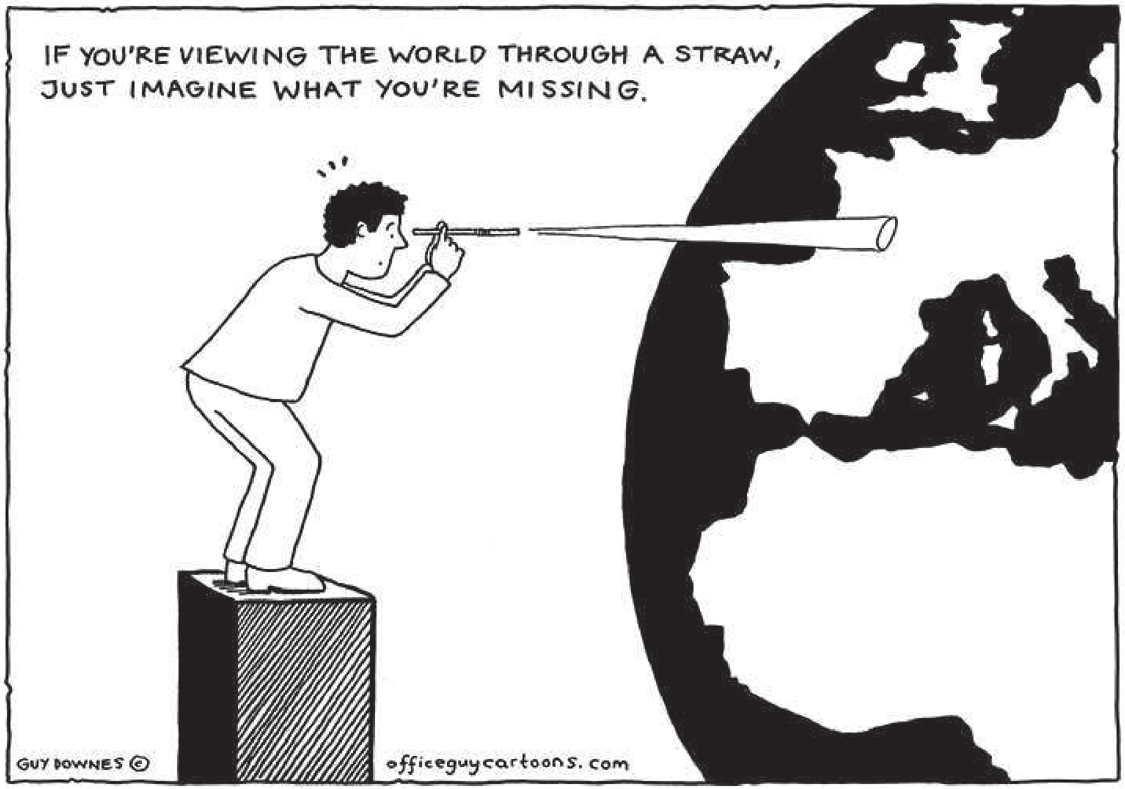
- All approaches rely on untestable assumptions about the relation between the measurement process and the dropout/error process. Hence, it is advisable to always perform a sensitivity analysis.
- “All statistical models are wrong, but some are useful" — George Box

Summary and Discussion

- What?
  - What noisy data do we examine?
  - Noisy data have features including
    - measurement error,
    - missing observations,
    - high dimensionality with inactive/unimportant variables
- Why?
  - Why do the features of noisy data matter?
  - Effects arising from noisy data are complex.
- How?
  - How to incorporate the features of noisy data ?
  - A case by case study is generally needed.
- What's left?

  Many research problems remain unexplored.

Summary and Discussion


IF YOU'RE VIEWING THE WORLD THROUGH A STRAW,
JUST IMAGINE WHAT YOU'RE MISSING.
GUY DOWNES ©
officeguycartoons.com

Take Home Messages

- In the era of big data, data is everywhere.
  - big information?
  - big mess?
  - big noise?
  - big opportunities?
  - big responsibilities?
  - big challenges?

Take Home Messages

- In the era of big data, data is everywhere.
  - big information?
  - big mess?
  - big noise?
  - big opportunities?
  - big responsibilities?
  - big challenges?

- Big dimension of the discipline

  - Data Science: intersection of Statistical Science, Computer Science, and others

Take Home Messages

- In the era of big data, data is everywhere.
  - big information?
  - big mess?
  - big noise?
  - big opportunities?
  - big responsibilities?
  - big challenges?

- Big dimension of the discipline
  - Data Science: intersection of Statistical Science, Computer Science, and others

- “Big data has arrived, but big insights have not" (Harford 2014)
  - statistical validity/efficiency
  - modeling complexity/validity
  - computation complexity/feasibility

Take Home Messages

- In the era of big data, data is everywhere.
  - big information?
  - big mess?
  - big noise?
  - big opportunities?
  - big responsibilities?
  - big challenges?

- Big dimension of the discipline
  - Data Science: intersection of Statistical Science, Computer Science, and others

- “Big data has arrived, but big insights have not" (Harford 2014)
  - statistical validity/efficiency
  - modeling complexity/validity
  - computation complexity/feasibility

- Issue often overlooked
  - EXAMINING DATA PROVENANCE AND QUALITY IS CRUCIAL!

Take Home Messages

CHECKLIST REPORT
Optimizing Data Quality for Big Data
By Philip Russom
Informatica
tdwi TRANSFORMING DATA WITH INTELLIGENCE™

Acknowledgements

- My Collaborators:

  Raymond Carroll
  Donna Spiegelman
  Wenqing He
  Yanyuan Ma
  Xiaomei Liao
  Xianming Tan
  Runze Li
  Juan Xiong
  Ying Yan
  Di Shu

- Funding Agency: NSERC

- References of Some Figures:
  https://www.google.ca