



Bundesamt
für Strahlenschutz



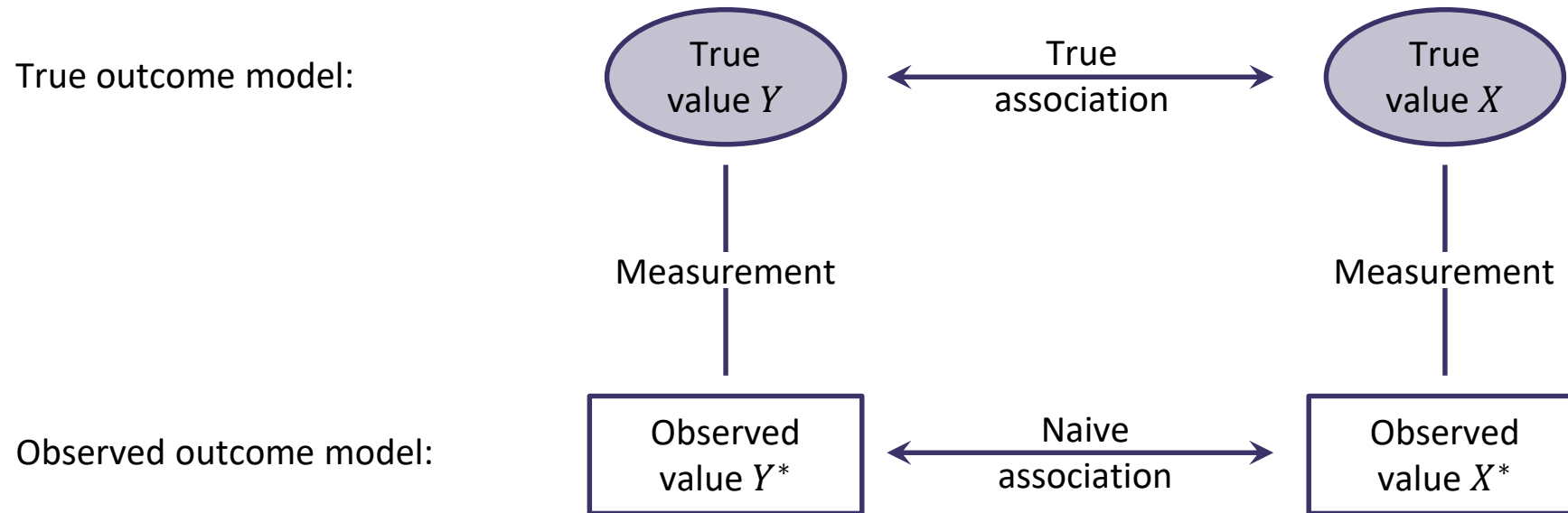
MEM-Explorer

Teaching tool for exploring measurement error and misclassification in statistical analyses

Veronika Deffner

TG 4 on measurement error and misclassification of the STRATOS Initiative

Analyses with erroneous observations



STRATOS TG 4: Measurement error and misclassification

Aims

1. Increase the **awareness** of the problems caused by measurement error and misclassification in statistical analyses
2. Remove barriers to use **statistical methods** that deal with such problems

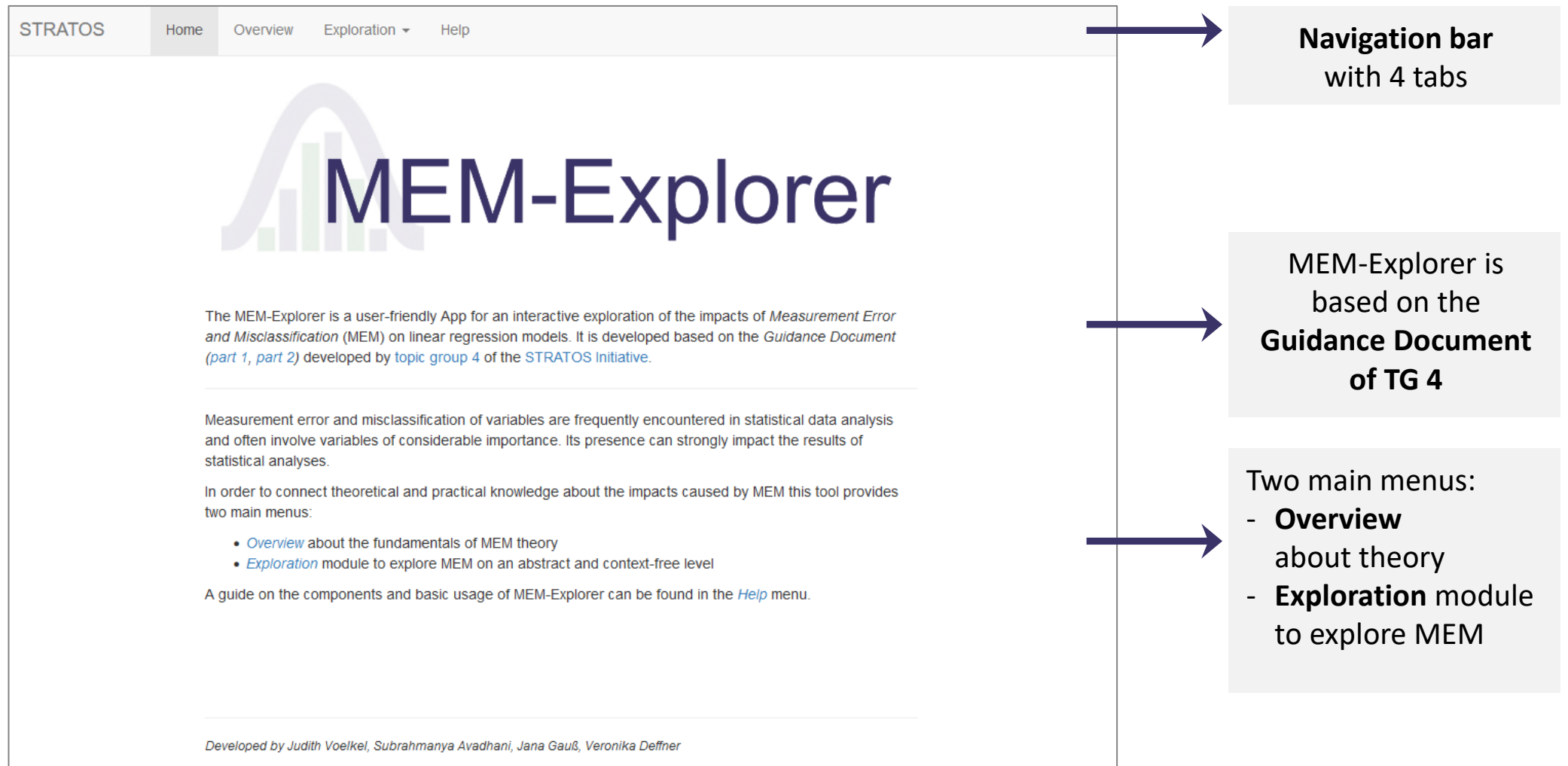
Activities

- ▶ Article on the **general relevance** of the topic (Wallace 2020)
- ▶ **Literature survey** about current practice (Shaw et al. 2018)
- ▶ **STRATOS guidance document:**
 - Part 1 – Basic theory and simple methods of adjustment (Keogh et al. 2020)
 - Part 2 – More complex methods of adjustment and advanced topics (Shaw et al. 2020)
- ▶ Presenting papers and workshops at **conferences**
- ▶ **Website:** <http://www.stratostg4.statistik.uni-muenchen.de>
- ▶ Interactive Shiny application “**MEM-Explorer**”:



<https://mem-explorer.shinyapps.io/MEMExplorer-v5/>

Home



The screenshot shows the MEM-Explorer web application interface. At the top, there is a navigation bar with four tabs: "STRATOS", "Home", "Overview", and "Exploration" (with a dropdown arrow), and "Help". The main content area features a large logo for "MEM-Explorer" with a stylized bar chart and a bell curve. Below the logo, there is a paragraph describing the application as a user-friendly tool for exploring measurement error and misclassification. This is followed by a paragraph explaining the importance of measurement error and misclassification in statistical data analysis. Then, there is a section titled "In order to connect theoretical and practical knowledge about the impacts caused by MEM this tool provides two main menus:" with a bulleted list of "Overview" and "Exploration". Below this, a sentence states that a guide on the components and basic usage can be found in the "Help" menu. At the bottom, there is a footer crediting the developers: "Developed by Judith Voelkel, Subrahmanya Avadhani, Jana Gauß, Veronika Deffner".

Navigation bar with 4 tabs

MEM-Explorer is based on the **Guidance Document of TG 4**

Two main menus:
- **Overview** about theory
- **Exploration** module to explore MEM

Developed by Judith Voelkel, Subrahmanya Avadhani, Jana Gauß, Veronika Deffner

Overview

STRATOS Home Overview Exploration Help

Theoretical overview of Measurement Error and Misclassification

This page serves as an overview about the fundamentals of MEM theory and is mainly based on the guidance document ([part 1](#), [part 2](#)) developed by STRATOS-Initiative.

Main Types of Error Impact on Regression Error Adjustment Sources

This section describes the main types of error that occur in measurements during observational studies.

Terminology

Two separate terms for errors in variables are distinguished:

- **measurement error** for error in continuous variables and
- **misclassification** for error in categorical variables.

Aim

The aim is to learn the regression relationship between a scalar outcome variable Y and covariates X .

Basic situation

X is measured with error, with the true value of X being unobserved. The error-prone observed variable is denoted by X^* .

Note: Mismeasurement can occur not only in covariates but also in the outcome variable Y . This is currently not considered in MEM-Explorer.

Requirements

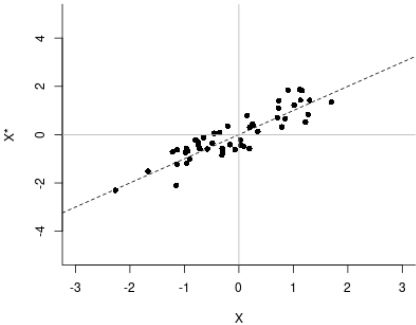
Learning the regression relationship between Y and X based on the observed covariates X^* requires knowledge about type and size of the error in X^* , i.e. the relationship of X and X^* has to be specified. This is called the **error model**.

Main Types of Error

Random vs. Systematic Classical vs. Linear vs. Berkson Additive vs. Multiplicative Differential vs. Non-differential

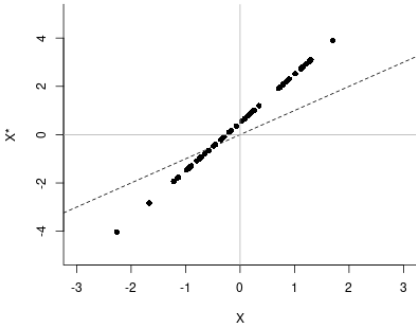
Random Error

Random or stochastic error occurs when X^* differs randomly from X . The following figure shows an example for random error.



Systematic Error

Systematic error occurs when X^* differs systematically from X . The following figure shows an example for systematic error with linear structure.



Basics for each section

Sections:

- Main Types of error
- Impact on Regression
- Error Adjustment

Subsections for each section

Overview

Main Types of Error

- Random vs. systematic
- Classical vs. linear vs. Berkson
 - Measurement error
 - Misclassification
- Additive vs. Multiplicative
- Differential vs. Non-differential

Impact on Regression

Error Adjustment

Main Types of Error

Random vs. Systematic

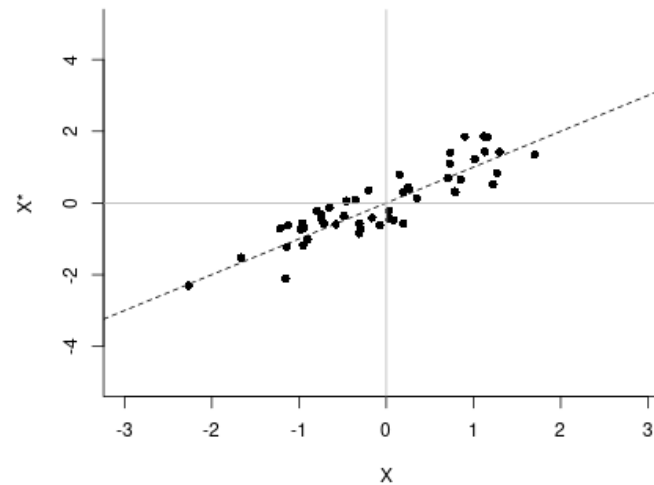
Classical vs. Linear vs. Berkson

Additive vs. Multiplicative

Differential vs. Non-differential

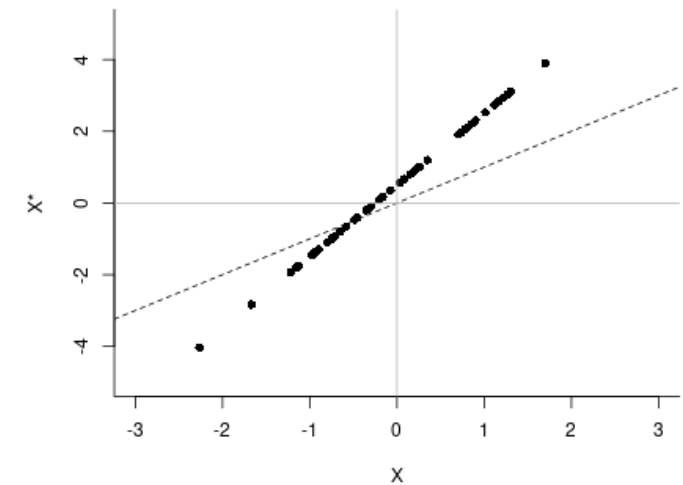
Random Error

Random or stochastic error occurs when X^* differs randomly from X . The following figure shows an example for random error.



Systematic Error

Systematic error occurs when X^* differs systematically from X . The following figure shows an example for systematic error with linear structure.



Overview

Main Types of Error

Impact on Regression

(a) Measurement error

- Classical vs. Linear vs. Berkson
- Impact on
Regression coefficient
Test of null hypothesis
Power

(b) Misclassification

- Classical vs. Berkson

Error Adjustment

Impact on Regression

(a) Measurement error

(b) Misclassification

Impacts of Measurement Errors on Linear Regression

Assume the following:

- X , X^* are continuous covariates.
- U is a random variable with mean 0.
- $U \perp X$ in linear and classical error models.
- $U \perp X^*$ in Berkson error model.
- Errors U are nondifferential with respect to outcome Y .
- ρ_{XX^*} is the correlation coefficient between X and X^* .

	Classical measurement error model $X^* = X + U$	Linear measurement error model $X^* = a_0 + a_X X + U$	Berkson error model $X = X^* + U$
Single covariate regression			
Regression coefficient	Underestimated $\lambda = \frac{\text{var}(X)}{\text{var}(X) + \text{var}(U)}$	Biased in either direction $\lambda = \frac{\alpha_X \text{var}(X)}{\alpha_X^2 \text{var}(X) + \text{var}(U)}$	Unbiased $\lambda = 1$
Test of null hypothesis	Valid	Valid	Valid
Power	Reduced → effective sample size reduced by approximately $\rho_{XX^*}^2 = \lambda$	Reduced → effective sample size reduced by approximately $\rho_{XX^*}^2$	Reduced → effective sample size reduced by approximately $\rho_{XX^*}^2$
Regression with multiple error prone covariates			

Overview

Main Types of Error

Impact on Regression

Error Adjustment

- SIMEX
 - Measurement error
 - Misclassification

Error Adjustment

SIMEX

The basic idea of SIMEX (Simulation extrapolation) is to add more error to the error prone covariate X^* , to see the impact this has on the parameter estimates, and then extrapolate back to the situation with no measurement error. This is equivalent to estimating the relationship between the measurement error variance $\text{var}(U)$ and the parameter estimates and to extrapolate back to the situation in which the error variance is 0.

(a) Measurement error

(b) Misclassification

Assume a *classical measurement error* model, that is $X^* = X + U$ (U is a random variable with mean 0 and $U \perp X$).

The theoretical relationship between the biased regression slope β_X^* , based on the regression of Y on X^* , can be described as a function of $\text{var}(U)$, $\beta_X^*(\text{var}(U))$. This function is estimated from simulated datasets obtained by adding further measurement error to X^* . To estimate β_X , the function $\beta_X^*(\text{var}(U))$ is extrapolated back to $\text{var}(U) = 0$ (as $\beta_X^*(0) = \beta_X$).

Let $\text{var}(U)$ be known or assumed, e.g. through comparison measurements.

- **Simulation step:**

For each value $s_1, \dots, s_m \geq 0$, B new pseudo-datasets are simulated by

$$X_{ib}^*(s_k) = X_i^* + \sqrt{s_k \text{var}(U)} U_{ibk}, \quad i = 1, \dots, n; b = 1, \dots, B; k = 1, \dots, m$$

where U_{ibk} are *iid* standard normal variables. The measurement error variance of $X_{ib}^*(s_k)$ is therefore $(1 + s_k)\text{var}(U)$.

For each pseudo-dataset, the unadjusted estimator based on Y and $X_{ib}^*(s_k)$ is calculated and the results are averaged over the B repetitions, which leads to $\hat{\beta}_{X_{s_k}}^*$.

- **Extrapolation step:**

We use a parametric approximation for the function $\beta_X^*((1 + s)\text{var}(U))$, denoted by $G(s, \Gamma)$, where Γ denotes the parameters of the function, e.g. in case of a quadratic approximation $G(s, \Gamma) = \gamma_0 + \gamma_1 s + \gamma_2 s^2$.

The parameters Γ are estimated by least squares and the estimated parametric function is then extrapolated to $s = -1$, the case of no measurement error, yielding the SIMEX estimator:

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

When β_X is a vector, the SIMEX-procedure can be applied separately for each component.

Exploration

STRATOS Home Overview Exploration Help

Measurement Error
Misclassification

Impacts of Measurement Error on Linear Regression

$$Y = \beta_0 + \mathbf{X}^T \beta_X + \epsilon,$$

$$\epsilon \sim N(0, \sigma_\epsilon^2 I_n),$$

$$\mathbf{X} = (X_1, \dots, X_p)^T, \beta_X = (\beta_{X1}, \dots, \beta_{Xp})^T$$

Error Model

Error form:
 Additive
 Multiplicative

Error type:
 Classical
 Linear
 Berkson

Variance of unbiased error of X:
 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Outcome Model

Sample size n:
 20 100 200 320 420 520 620 720 820 900 1,000

Number of covariates p:
 1

Intercept β_0 :
 -15 0 15

Coefficient β_1 :
 -15 0 15

Variance of error term ϵ :
 0 1 15

Covariate Distributions
 Normal distribution:
 $X \sim N(\mu_X, \sigma_X^2)$

Classical Error
 $X^* = X + U$

Density

X vs. X*

Legend: True value X, Measurement X*, Error U, (X, X*), Regression line, Identity X* = X

Exploration:

- **Measurement Error**
- **Misclassification**

Specification of **error model**

Subsections with output:

- Error Model
- Impact on Regression
- Error Adjustment
- Data Table

Specification of **sample size, outcome model and covariate distributions**

Exploration – Input

$$Y = \beta_0 + \mathbf{X}^T \beta_X + \epsilon,$$

$$\epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2 I_n),$$

$$\mathbf{X} = (X_1, \dots, X_p)^T, \beta_X = (\beta_{X_1}, \dots, \beta_{X_p})^T$$

Outcome Model

Sample size n:
Number of covariates p:
Intercept β_0 :
Coefficient β_X :
Variance of error term ϵ :

Covariate Distributions

Normal distribution:
 $X \sim N(\mu_X, \sigma_X^2)$

Log-normal distribution:
 $\log(X) \sim N(\mu_X, \sigma_X^2)$

Distribution X_i : **E(X_i):** **Var(X_i):**

normal

log-normal

Seed:

Error Model

Error form:

Additive
 Multiplicative

Error type:

Classical
 Linear
 Berkson

Variance of unbiased error of X_i :

Linear parameters for X_i :

α_0 :

α_X :

Exploration – Output

Error Model

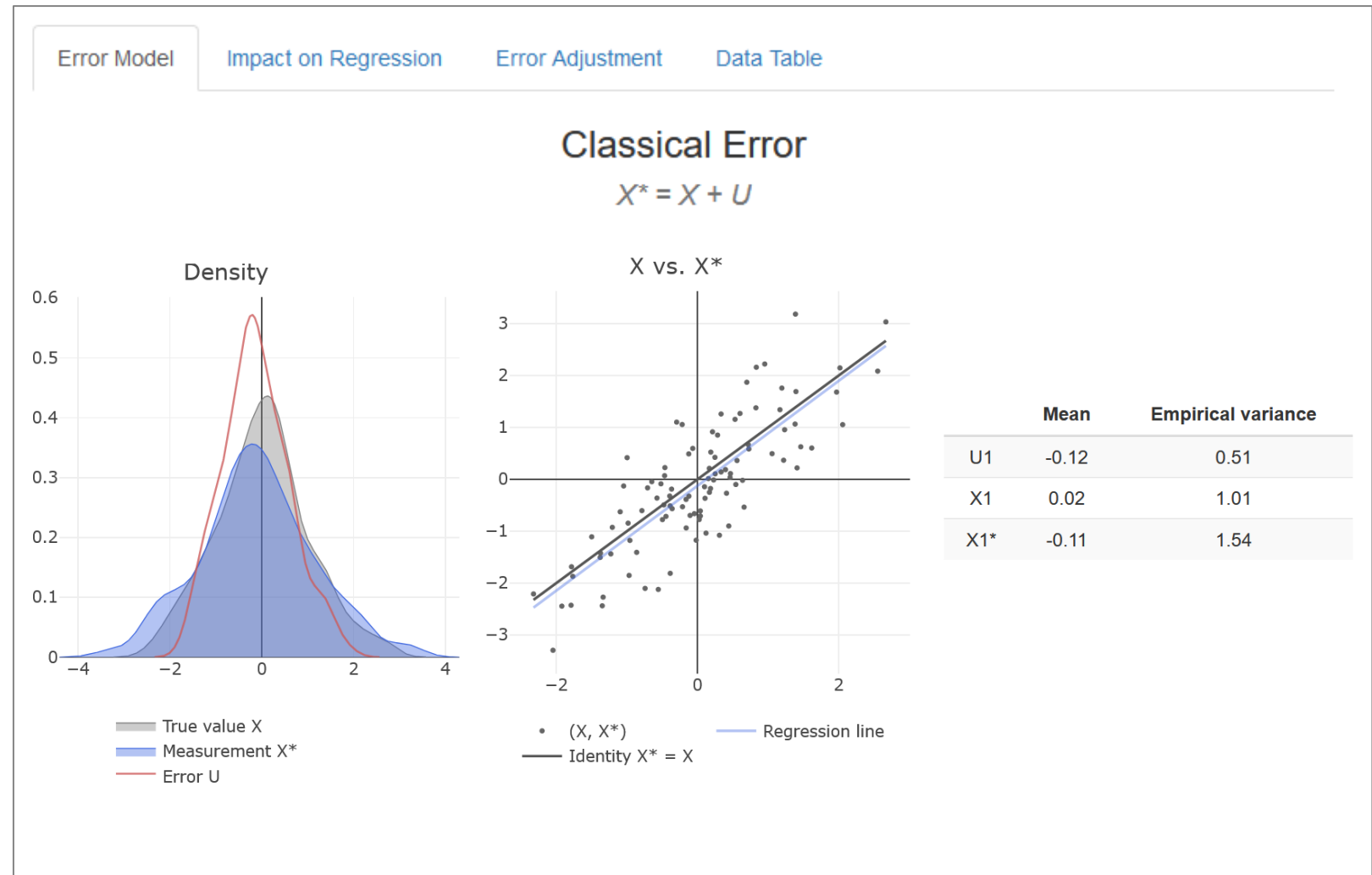
Empirical distributions:

- True covariate
- Error-prone covariate
- Measurement error

Impact on Regression

Error Adjustment

Data Table



Exploration – Output

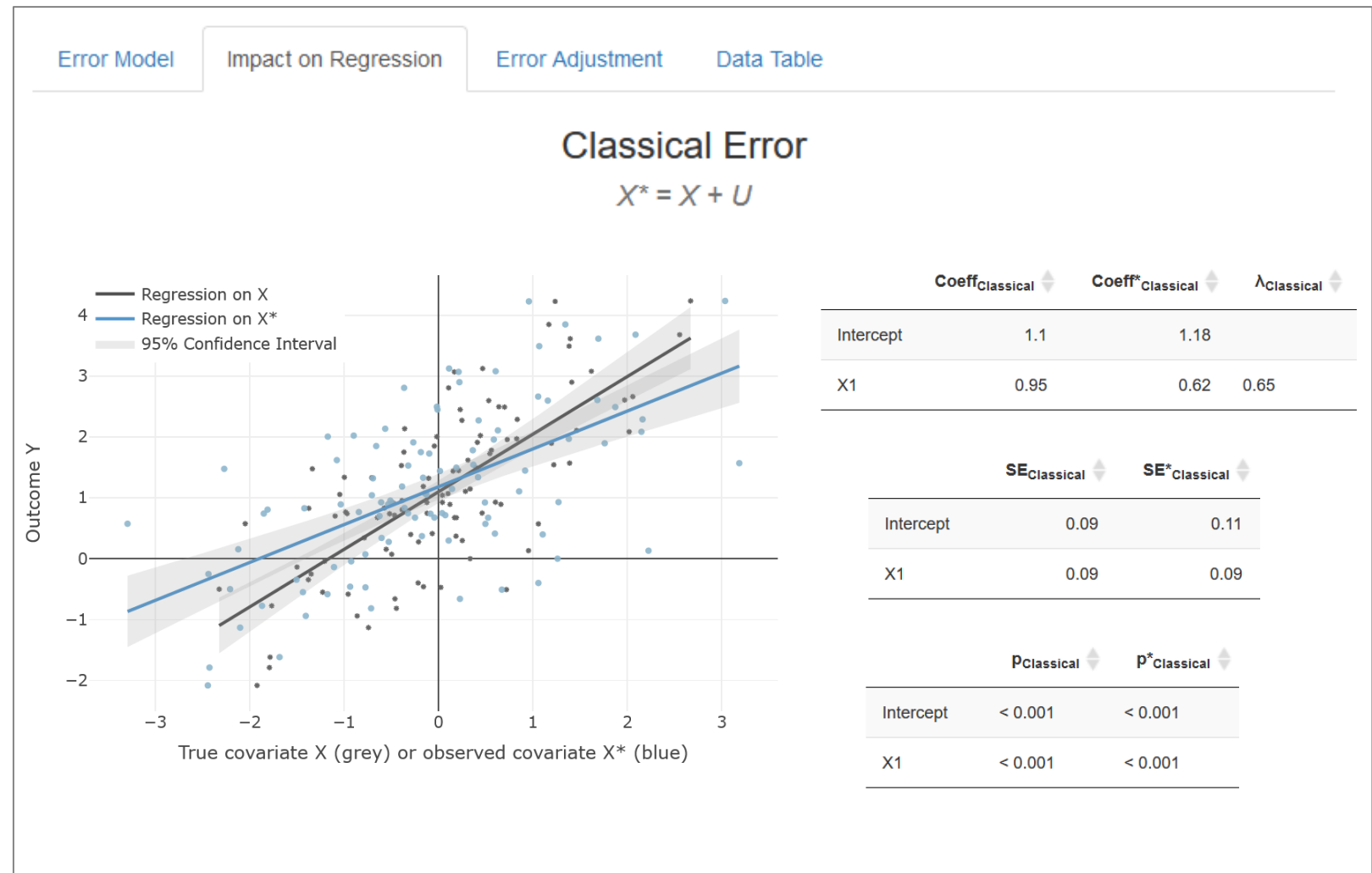
Error Model

Impact on Regression

Visualization of the impact of MEM on the regression model

Error Adjustment

Data Table



Exploration – Output

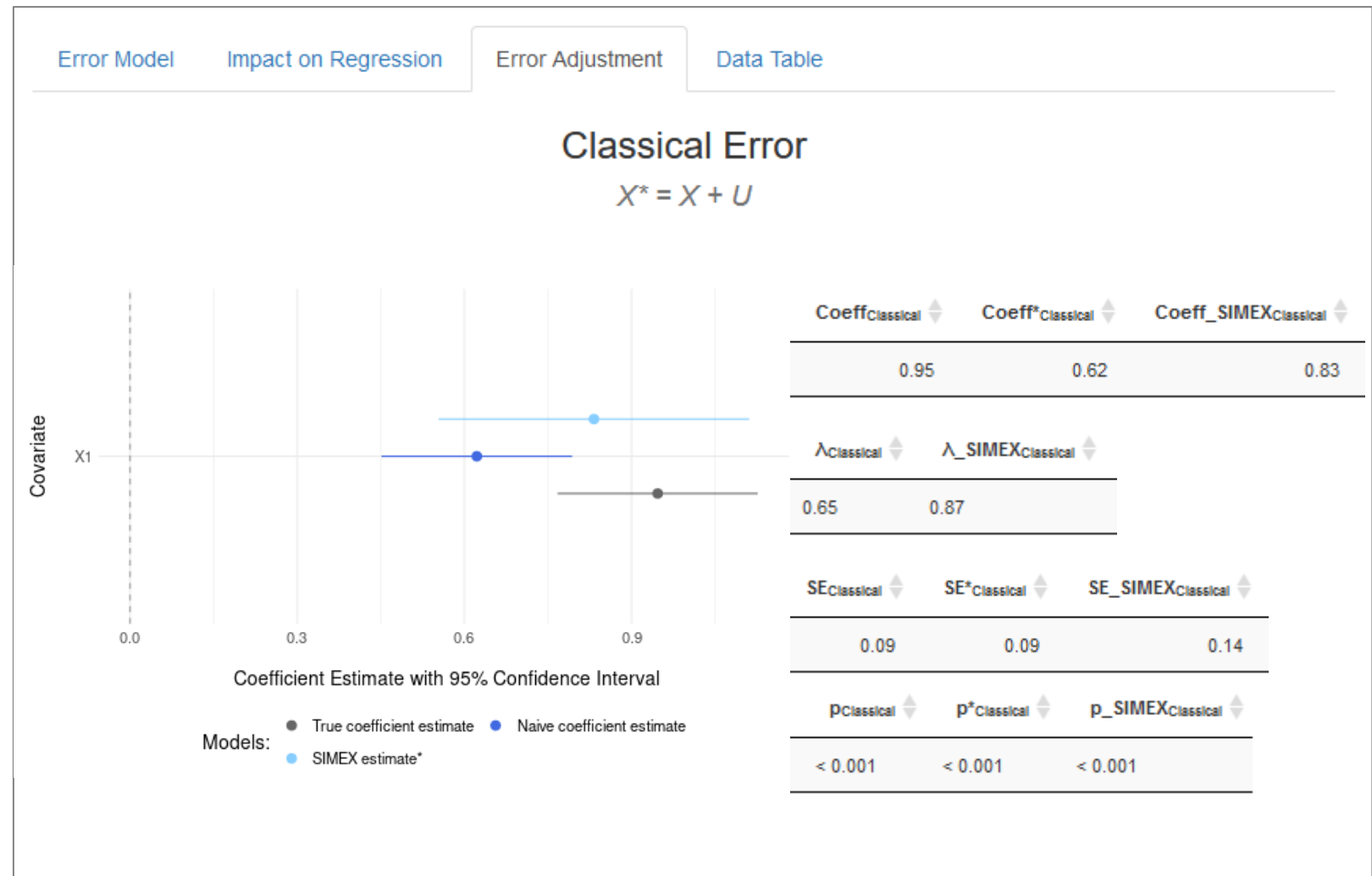
Error Model

Impact on Regression

Error Adjustment

Visualization of regression models using methods that adjust for bias in regression coefficients caused by MEM

Data Table



Exploration – Output

Error Model

Impact on Regression

Error Adjustment

Data Table


- Excerpt of simulated dataset
- Download

Error Model

Impact on Regression

Error Adjustment

Data Table

 Download Data Table

Header of simulated Data

y	x1.x	x1.u	x1.classical
-0.82	-0.45	-0.27	-0.71
-0.04	-1.21	0.28	-0.92
0.93	0.04	-0.65	-0.61
2.50	0.64	-0.66	-0.02
0.34	-0.79	0.18	-0.60
0.81	-0.39	-1.42	-1.81

Example

Sample size

$$n = 100$$

Outcome model

Simple linear regression model with

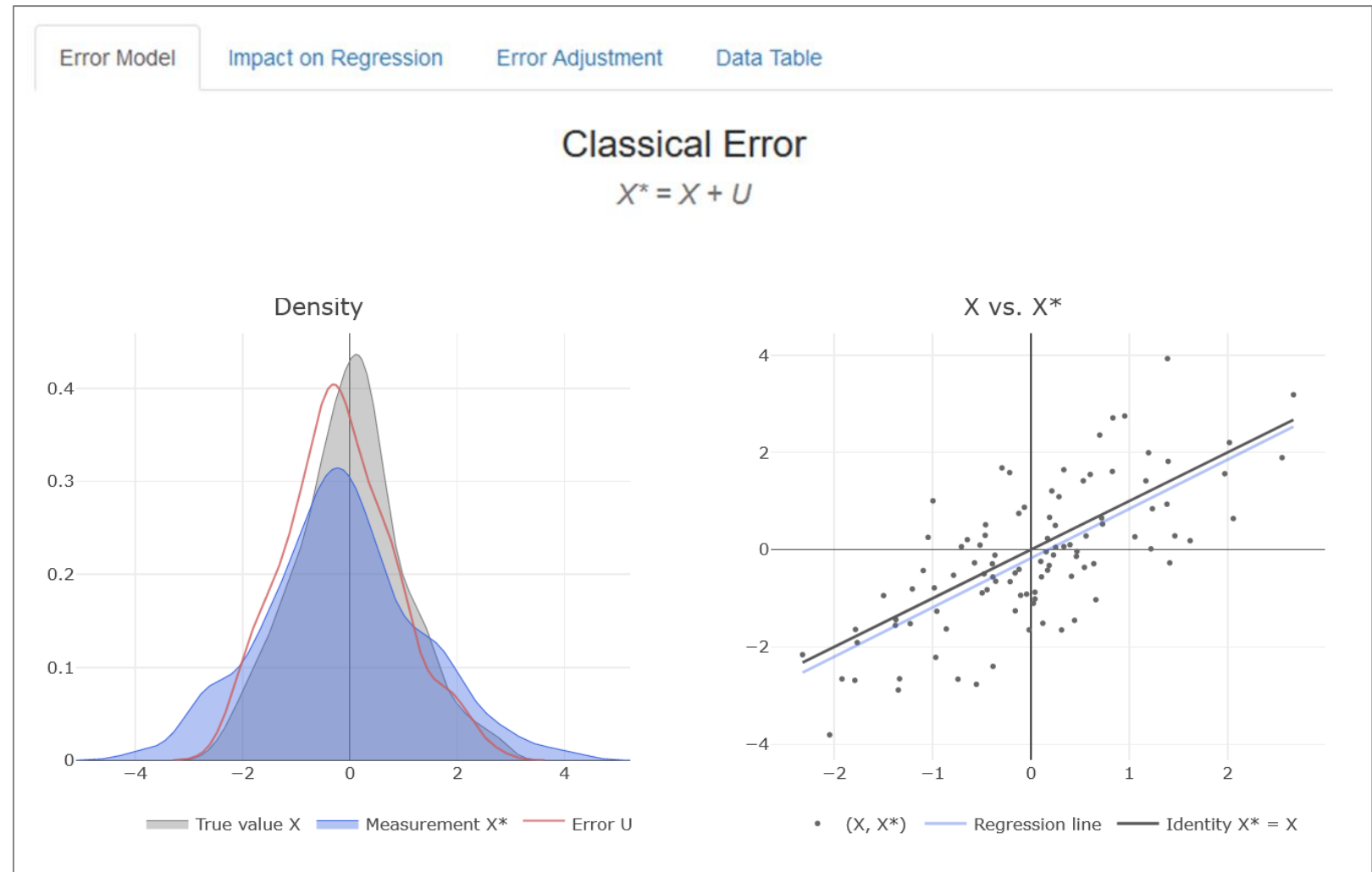
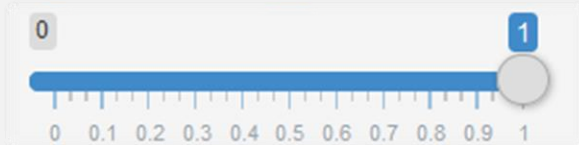
- coefficients $\beta_0 = 0$ and $\beta_1 = 1$
- error variance $\text{Var}(\varepsilon) = 1$

Covariate distribution

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

Error model

- Additive, classical error
- Error variance



Example

Sample size

$$n = 100$$

Outcome model

Simple linear regression model with

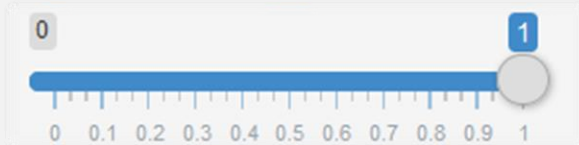
- coefficients $\beta_0 = 0$ and $\beta_1 = 1$
- error variance $\text{Var}(\varepsilon) = 1$

Covariate distribution

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

Error model

- Additive, classical error
- Error variance



Example

Sample size

$$n = 100$$

Outcome model

Simple linear regression model with

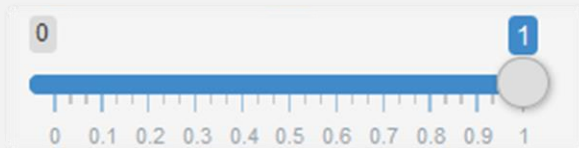
- coefficients $\beta_0 = 0$ and $\beta_1 = 1$
- error variance $\text{Var}(\varepsilon) = 1$

Covariate distribution

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

Error model

- Additive, classical error
- Error variance



Summary & outlook



- ▶ Shiny app for interactive exploration of measurement error and misclassification

- ▶ Find MEM-Explorer here: <https://mem-explorer.shinyapps.io/MEMExplorer-v5/>
- ▶ Ideas for improvement are welcome: vdeffner@bfs.de

Future work

- Options to specify parameters in MEM adjustment methods
- Integration of further MEM adjustment methods
- Measurement error in the outcome



**Bundesamt
für Strahlenschutz**



Impressum

Bundesamt für Strahlenschutz
Postfach 10 01 49
38201 Salzgitter

Tel.: +49 30 18333-0
Fax: +49 30 18333-1885
E-Mail: ePost@bfs.de

www.bfs.de

Kontakt für Rückfragen

Dr. Veronika Deffner
vdeffner@bfs.de
030/18333-2251