

Applied Error Modeling in Regression

An Introduction with Examples in R

Stefanie Muff and Lukas F. Keller

2018-12-21

Contents

Preface	5
1 Introduction	7
1.1 What is Error?	7
1.2 Why and When do I Have to Worry?	7
1.3 Organization and Take-Home Messages of This Book	7
1.4 Outlook	7
2 Types of Errors	9
2.1 Continuous Variables	9
2.2 Categorical and Count Variables	9
2.3 Differential vs Non-Differential Error	9
3 The Effects of Measurement Error	11
3.1 Classical Measurement Error	11
3.2 The Concept of Heritability, Regression to the Mean and Measurement Error	11
3.3 Berkson Measurement Error	11
3.4 Error in Categorical and Count Variables	11
3.5 Error in the response	11
4 Methods to Account for Measurement Error	13
4.1 Bayesian Methods	13
4.2 Simulation Extrapolation (SIMEX)	13
5 Linear Regression Models	15
6 Generalized Linear (Mixed) Models	17
6.1 Classical error	17
6.2 Berkson error	17
7 Survival Models	19
8 Advanced Topics	21
8.1 Misalignment Error in Spatial Data	21

Preface

This is a *first draft* of a book that deals with effects and cures of measurement error in variables of regression models. The aim of the book is not only to discuss a broad range of problems and biases that are induced by measurement error, but mainly to bridge the gap between theory and the applications. The idea is to provide a basic toolkit of methods to make error modeling accessible to a broad audience in the applied sciences. The many examples discussed and analyzed in the book all come with the associated R code.

Interestingly, the presence and effects of measurement error and misclassification in covariates and the response of regression models have been recognized already more than a century ago (see e.g. ...). Thanks to huge efforts of many researchers, the consequences of ignoring measurement error or misclassification are known in many settings, at least in theory. Moreover, a huge variety of methods to appropriately deal with measurement error exist, and several textbooks in statistics are devoted to the topic (Fuller, 1987; Gustafson, 2004; Carroll et al., 2006; Yi, 2017). Despite this, most – if not all – error modeling methods go largely unused. Why is this so? We can only hypothesize about the reasons, but the problem seems to have many facets. On one hand, measurement error is often nothing that seems worth paying attention to, and given that even most introductory textbooks in applied statistics do not discuss measurement error, it is not surprising that entire generations of young scientists get educated in statistics and data analysis without ever having heard of the problems it may cause. On the other hand, error modeling methods can quickly become very challenging. Unless the problem is a very standard case, it is often necessary to formulate a new model, and it may be all but obvious what the model should be, let alone how to implement an actual procedure to fit it. But even if the error model is relatively simple, like a standard classical measurement model in a covariate of a regression model (see Section 2.1), some extra-effort and more specialist software packages are required. As a consequence, the hurdle to get started with the proper handling of measurement error in data is much higher than for standard regression analyses.

If you are reading these lines, we assume that you either have a very specific measurement error problem at hand, or you would like to get a gentle introduction into the topic and its applications. ...

When we say “error”, we do not only mean actual mistakes in the data that are used to fit regression models.

Kind of uncertainty, noise or imprecision that are present in the data that we use to fit our models.

N. Breslow, *Lessons in Biostatistics* (2014) (Breslow, 2014) wrote

Obviously, [...] the *best* method of dealing with measurement error was to avoid it.

I say:

The *second best* method of dealing with measurement error is to properly account for it.

We might develop a package. In this case, the **package-to-be-developed** package can be installed from CRAN or Github:

```
install.packages("package-to-be-developed")
# or the development version
# devtools::install_github("stefaniemuff/package-to-be-developed")
```

Follow us on Twitter! @StefanieMuff @LukasFKeller

Chapter 1

Introduction

1.1 What is Error?

Explain that measurement error often comes in the shape of uncertainty, which is present in almost all data.

1.2 Why and When do I Have to Worry?

- Triple whammy of ME
- When is error a problem?
- Bias versus variance
- Is it sometimes better not to model the error?

It is surprising how many phenomena in statistics and its applications can be viewed through the measurement error lens. A prominent example is the concept of heritability in genetics and evolutionary biology, as we will explain in Section 3.2.

1.3 Organization and Take-Home Messages of This Book

1.4 Outlook

What we are going to do, outlook to chapters.

Chapter 2

Types of Errors

Before we can start to speak about the effects of measurement error (Chapter @ref{accounting}), or how to account for it (Chapter 4), we have to spend some time to understand what *kind* of error we are talking about.

Dichotomy into classical vs Berkson error, continuous vs categorical variables, differential vs non-differential error. Maybe more?

2.1 Continuous Variables

Two fundamentally different error types

2.1.1 Classical Measurement Error

2.1.2 Berkson Measurement Error

2.2 Categorical and Count Variables

2.3 Differential vs Non-Differential Error

Chapter 3

The Effects of Measurement Error

We will look into effects of ME in the linear regression case.

3.1 Classical Measurement Error

3.2 The Concept of Heritability, Regression to the Mean and Measurement Error

Geneticists, evolutionary biologists and animal breeders will be familiar with the concept of *heritability* (3.1).

$$h^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} \quad (3.1)$$

- Will use data in Figures 3.1 and 3.2 to explain regression to the mean
(Fuller, 1987; Galton, 1886)

3.3 Berkson Measurement Error

3.4 Error in Categorical and Count Variables

3.5 Error in the response

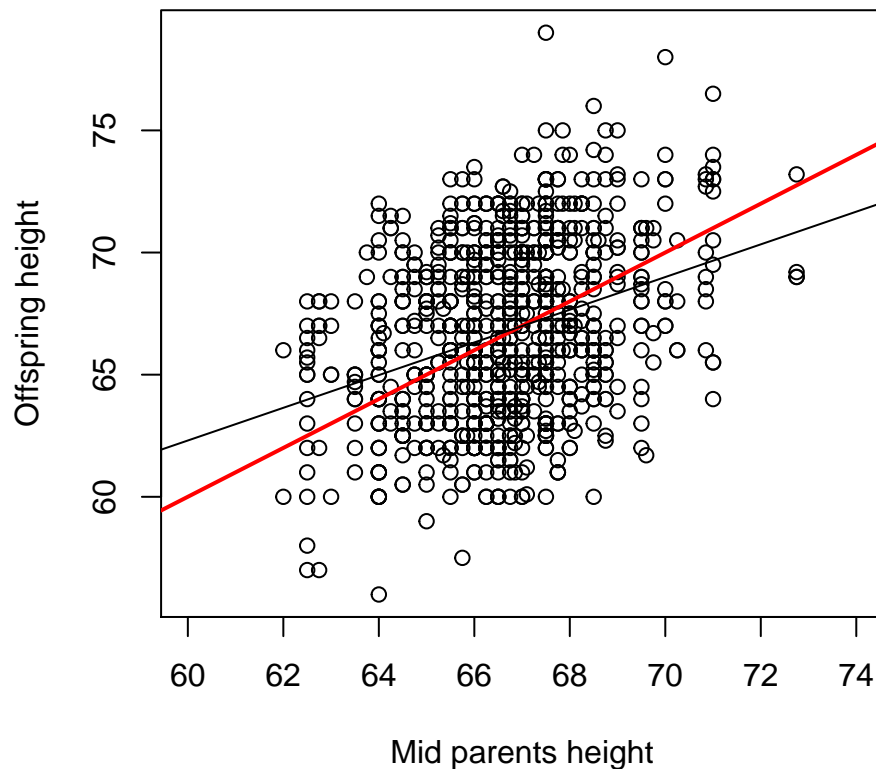


Figure 3.1: Data drawn from '<http://www.math.uah.edu/stat/data/Galton.txt>'

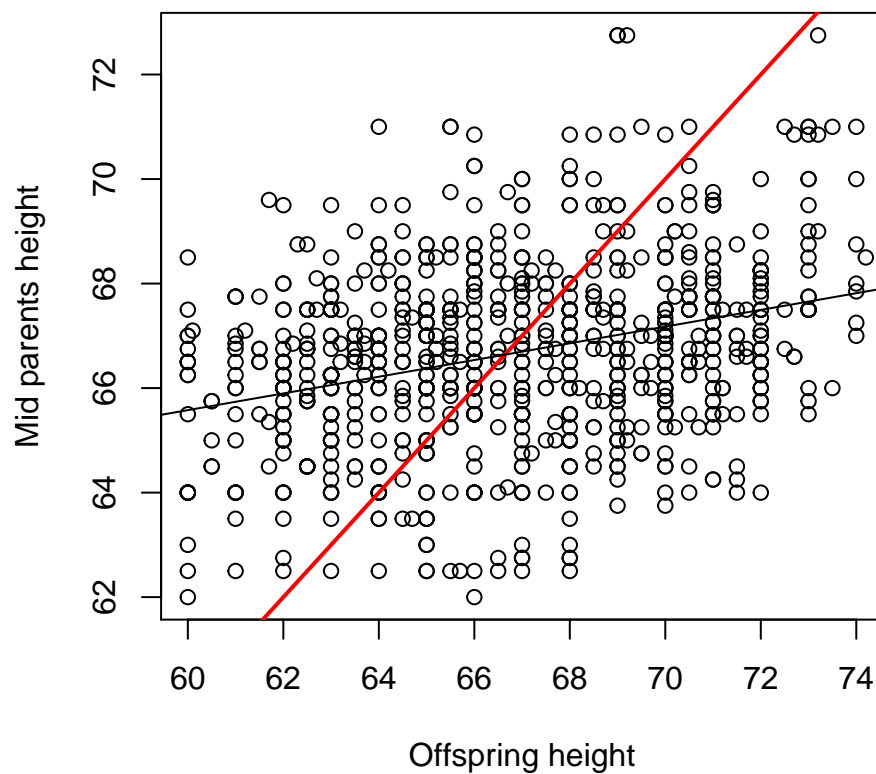


Figure 3.2: Data drawn from '<http://www.math.uah.edu/stat/data/Galton.txt>'

Chapter 4

Methods to Account for Measurement Error

4.1 Bayesian Methods

4.2 Simulation Extrapolation (SIMEX)

Chapter 5

Linear Regression Models

Chapter 6

Generalized Linear (Mixed) Models

6.1 Classical error

6.1.1 Error in a covariate

- Correlated covariates

6.1.2 Error in the response

6.2 Berkson error

6.2.1 Error in a covariate

6.2.2 Error in the response

Chapter 7

Survival Models

Chapter 8

Advanced Topics

8.1 Misalignment Error in Spatial Data

Bibliography

- Breslow, N. E. (2014). Lessons in biostatistics. In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J., editors, *Past, Present and Future of Statistical Science*, pages 335–347. CRC Press.
- Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, Boca Raton, 2 edition.
- Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.
- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.
- Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall/CRC, Boca Raton.
- Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer, New York.