Applied Error Modeling in Regression

An Introduction with Examples in R

Stefanie Muff and Lukas F. Keller

2019-05-14

Contents

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 modelig methods go largely unused. Why is this so? We can only hypothesize about the reasons, but the problem seems to have many factes. 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 hard 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.

We 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:

```r
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?

In our interactions with applied scientists, we often realize that ''error" is either regarded as something systematic, in the sense of a bias, potentially due to an inherent problem in the data collection process, or as the result of an erroneous step when data are handled and stored, like writing down a wrong number into the lab journal. However, the measurement error that we are talking about in this book is something much more universal. It should me understood as an *uncertainty* of measurements, which, so some degree, is present in virtually all data. To paraphrase Max Planck (in *The Meaning and Limits of Exact Science*, 1949):

> Measurement error is an uncertainty in our recording of Nature's answer to our questions

When measurement error is understood as the uncertainty that remains about an entity *after* a measurement of that entity was taken, it becomes obvious that measurement error may originate due to a wide variety of reasons, such as

- Measurement *imprecision* in the field or the lab, which may arise due to limited accuracy of an instrument, or because the targeted value is difficult to measure or volatile, for example blood pressure, body weight etc.
- *Incomplete* or *biased* observations, for example due to preferential sampling.
- Misclassification of categorical variables, e.g. when deciding if a disease is present or not.
- Rounding or digit preference.
- Temporal or spatial misalignments of observations, e.g. in interval samples GPS observations of telemetry studies.
- Measurement of an entity of interest by taking measurements on a substitute, e.g.
- Self-reporting of numbers, e.g. by patients in medicine or in food behaviour studies.

This is of course by no means a comprehensive list, and we are sure many readers could immediately come up with more examples. We put on the record that measurement errors should not, in generaly, merely be seen as ''mistakes", but as uncertainty that is inherently present due to the limitations of our ability to collect information in the real world.

It is surprising how many phenomena in statistics and its applications can be viewed through the measurement error lens. Prominant examples are the concept of heritability in genetics and evolutionary biology, as we will explain in Section 3.2, or the fact that omitting informative covariates in regression models is analogous to introducing a certain type of measurement error, the Berkson error, which we will introduce in Section 2.1.2.

1.2 Why and When do I Have to Worry?

An important question, and one that is central to this book, concerns the effect measurement error in the data, given that the aim is to estimate the parameters of a model. Measurement error has essentialy three effects, which Carroll et al. (2006) denote as the **Triple Whammy of Measurement Error**:

- Parameter estimators of statistical models are biased.
- It leads to loss of statistical power to detect relationships.
- Features of the data are masked in graphical analyses.

To be illustrated with simple and/or real examples for classical measurement error, with reference to Section 2.1.1.

1.2.1 When is Error a Problem?

This is a question practicioners often ask. It is, however, impossible to give general guidelines about when measurement error is ''severe" enough to affect parameters estimates and conclusions drawn from an analysis in a way that cannot be tolerated. A rather universal advice we can give is to simulate data from the model at hand, and check what happens to the estimated parameters, standard errors and - more broadly - the conclusions that are drawn when error of the type that is expected is artificially generated in the error-prone variables. Of course, this already requires a rough understanding of the error model at hand, but without such a model it is anyway elusive to hypothesize about the effect of the error.

Let us look at a very simple simulation setup for a linear regression model, with a covariate $x \sim \mathsf{N}(0,1)$ and a response y, generated according to the simple regression model

$$y = \beta_0 + \beta_x x + \epsilon \ , \quad \epsilon \sim \mathsf{N}(0,1) \ ,$$

with $\beta_0 = \beta_x = 1$.

```
set.seed(1234)
x <- rnorm(100)
y <- 1 + x + rnorm(100)
summary(lm(y~x))$coef
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 1.0371541  0.1049788 9.879657 2.208449e-16
## x           0.9739151  0.1037759 9.384789 2.620462e-15
```

Now let us add measurement error to the x covariate, namely by assuming we see a variable $w = x + u$, which is an error-prone version of x with measurement error $u \sim \mathsf{N}(0, \sigma_u^2)$ and measurement error variance $\sigma_u^2 = 1$. This is a so-called classical measurement error model (see Section 2.1.1), which is known to attenuate the estimated slope parameters, as can be verified by a simple simulation. Let us generate an error-prone version w and then regress y against w:

```
w <- x + rnorm(100)
summary(lm(y~w))$coef
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 0.8856732 0.11826999 7.488571 3.076746e-11
## w           0.5522517 0.08225006 6.714301 1.240119e-09
```

We see that that estimated slope for w, $\hat{\beta}_w = 0.55$ is much smaller than the actual slope $\hat{\beta}_x = 0.97$ that was estimated in the regression of y against x. This severe bias occurs because we have assumed that the error variance is as large as the actual variance of the true covariate ($\sigma_u^2 = \sigma_x^2 = 1$), which may not necessarily by realistic in practical applications. In any case, by playing with the error variance σ_u^2, the user can generate a feeling for the effect of the error at hand. A simulation for this simple linear regression model with classical error has been implemented in a Shiny app (see 3.1). Note, however, that the situation in applications is typically quite a bit more complicated.

When it comes to the decision whether one should worry or not, we strongly believe this depends on the context. Ideally, the user has pre-specified which effects can be tolerated *before* actually simulating the error, and it can then be decided whether the effect is negligible in that sense. Otherwise, the error must be accounted for by an appropriate error model. In that case, this book is for you!

1.2.2 Bias Versus Variance

Todo: Look at bias vs variance part in Carroll book, but then point out that the (typically) larger variance is also the more *honest* estimate, because error obscures information, and by ignoring it we pretend to be more certain about a (potentially biased) estimate than we really are.

1.2.3 Is it sometimes better not to model the error?

- If error is unknown, correction can go in the wrong direction
- Therefore it is crucial to find an error model
- Plan right during data collection process

1.3 Organization and Take-Home Messages of This Book

- How the book is organized
- Examples and R code (how will it be made available?)
- 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 3), or how to account for it (Chapter 4), we have to spend some time to understand whan *kind* of error we are talking about.

Dichotomy into classical vs Berkson error, continous vs categorial 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* (h^2), defined as the proportion of additive genetic variance (σ_A^2) to total phenotypic variance (σ_P^2) of all trait observations in a population of interest (3.1).

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2} \tag{3.1}$$

At its core, the above definition of heritability implies that the observed phenotype P of a trait of interest (e.g. body height) can be decomposed into the sum of two independent components, namely additive genetic effects A and an environmental component E, that is $P = A + E$, where A has variance σ_A^2 and E has variance σ_E^2. The independency assumption between A and E implies that phenotypic variance (σ_P^2) can be split into additive genetic variance (σ_A^2) and an environmental variance (σ_E^2).

- Explain why and how parent-offspring regression can be used to estimate h^2.
- Will use data in Figures 3.1 and 3.2 to explain regression to the mean.

The environmental component can be interpreted as the ‘’measurement error”, because it is the component that obscures the genetic merit.

(Fuller, 1987; Galton, 1886)

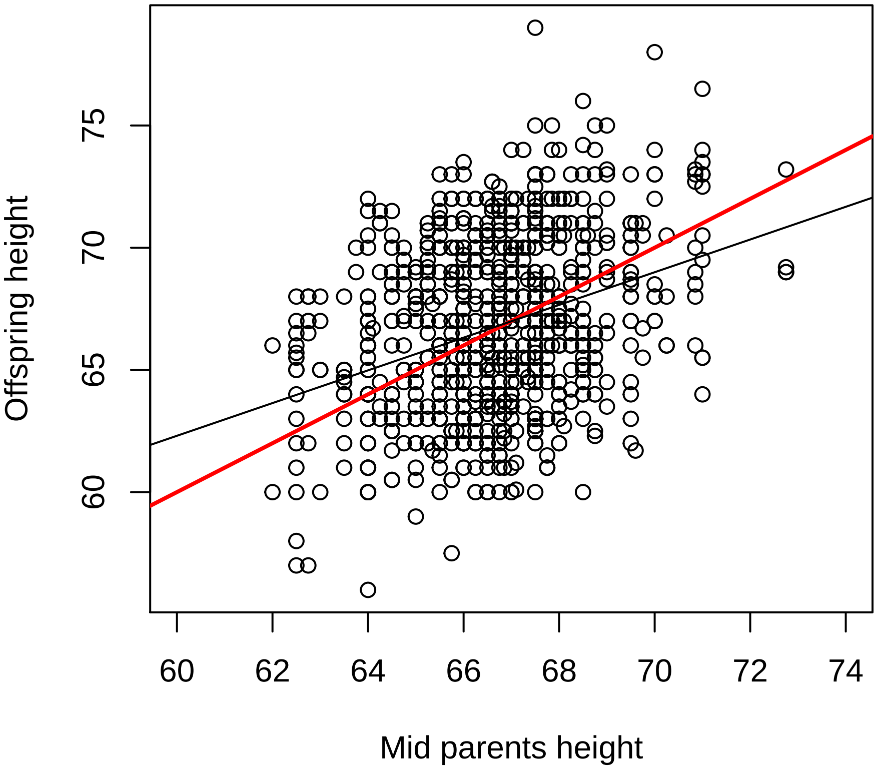

Figure 3.1: Data drawn from ‘http://www.math.uah.edu/stat/data/Galton.txt‘

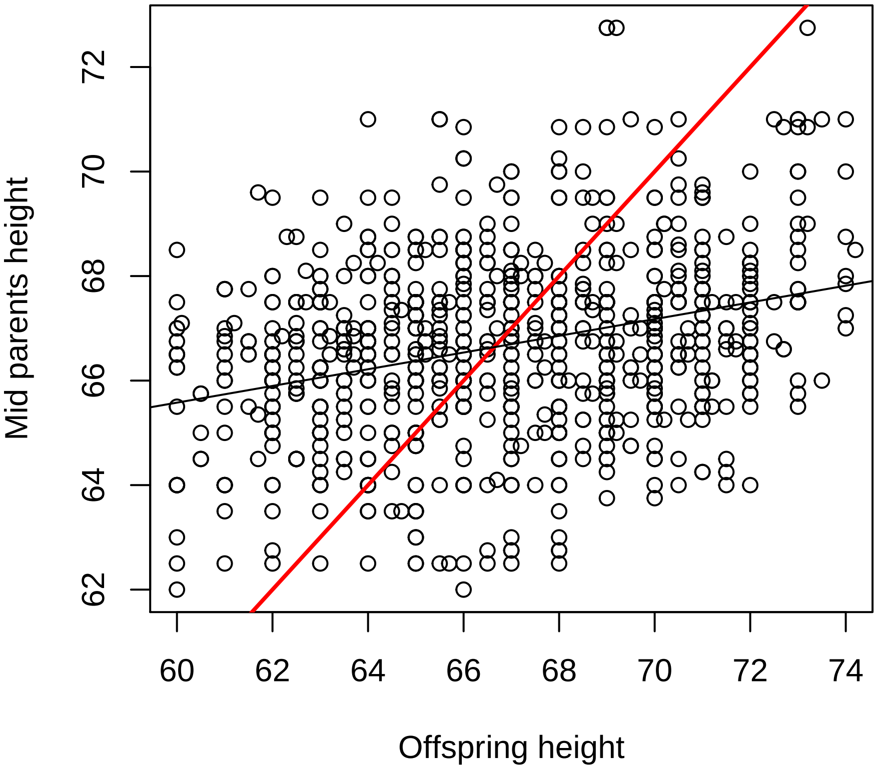

Figure 3.2: Data drawn from ‘http://www.math.uah.edu/stat/data/Galton.txt‘

3.3 Berkson Measurement Error

3.4 Error in Categorical and Count Variables

3.5 Error in the response

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 Rounding Error

8.2 Misalignement 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.