Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in Scienze Statistiche

Functional Methods for Measurement Error Correction in Astronomy: Application to a Simulated Framework and to Hubble's Law

Relatore: Prof.ssa Alessandra Rosalba Brazzale
Dipartimento di Scienze Statistiche

Laureando: Andrea Cappozzo
Matricola N 1055878

ANNO ACCADEMICO 2014/2015

I want to grow.
I want to be better.
You grow.
We all grow.
We're made to grow.
You either evolve or you disappear.

Tupac Shakur

Ringraziamenti

Il raggiungimento di questo traguardo è merito del sostegno e dell'appoggio di numerose persone.

Desidero ringraziare innanzitutto la mia relatrice, prof.ssa Alessandra Brazzale, le sue correzioni e osservazioni sono state fondamentali per la stesura di questo elaborato. Desidero ringraziarla anche per gli stimolanti dialoghi che abbiamo sostenuto riguardanti le possibili “strade da percorrere” una volta concluso il percorso universitario e per i suoi preziosi consigli a riguardo. Per la stessa ragione desidero ringraziare anche la prof.ssa Mariangela Guidolin ed il professor Renato Guseo, entrambi sono sempre stati disponibili nel dare una loro opinione ad ogni mio dubbio e curiosità.

Ringrazio l'“astronomo” e amico Samuel dal Dosso e il Dott. Simola Umberto per avermi sempre garantito il loro supporto durante la stesura della sezione della tesi relativa ai concetti più teorici dell' Astronomia, materia nella quale io ero inizialmente completamente impreparato.

Desidero ringraziare profondamente i miei compagni di università che hanno condiviso con me questi cinque anni di gioie e dolori a Scienze Statistiche, perchè questo percorso non mi ha donato solamente dei compagni con cui studiare e disperarsi per un esame, ma dei veri e propri amici che rimarranno tali anche quando non condivideremo più le stesse aule e gli stessi banchi.

Un ringraziamento va anche agli amici della “casetta”, del “sanse” e del “murialdo”, perchè tutti, chi prima e chi dopo, siete entrati a far parte della mia vita e l'avete resa migliore.

Ultimo ma non ultimo il ringraziamento piú sentito va alla mia famiglia ed in particolare ai miei genitori. Solamente la loro immancabile presenza, caratterizzata da una totale fiducia nei miei confronti, ha permesso il raggiungimento di questo traguardo.

Andrea Cappozzo

Contents

List of Figures

List of Tables

Introduction

What does measurement mean?

"Accurate and minute measurement seems to the non-scientific imagination, a less lofty and dignified work than looking for something new. But nearly all the grandest discoveries of science have been but the rewards of accurate measurement and patient long-continued labour in the minute sifting of numerical results". This quote by Baron William Thomson Kelvin (1824–1907) clearly summarizes how important measurement is for the scientific world, and thus for the continuous improvement of knowledge-based research. Nonetheless, even everyday experiences are influenced and affected by quantification: we weight the pasta before cooking it, we assess the good or bad performance of a firm via numerical indexes, we constantly monitor wristwatches and calendars for appointments and deadlines. We live in a world of measurements (Hand [21], 2005).

The classical definition of *measurement*, which is standard throughout the physical sciences, is *the determination or estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind* (Emerson [17], 2008). With *quantity* we refer to whichever attribute is possible to measure, at least in principle. For example, the statement"The Eiffel tower is 324 metres tall" expresses the height of the tower as the ratio of the unit (the metre in this case) to the height of the tower itself. The real number "324" is a real number in the strict mathematical sense of this term.

Mathematically expressed, a scientific measurement is:

$$Q = r \cdot [Q] \tag{1}$$

where Q is the magnitude of the quantity, r is a real number and $[Q]$ is a unit magnitude of the same kind. Literature subdivides measurement in two mutual categories: *representational measurement* and *pragmatic measurement.* The former relates to existing attributes of the objects, e.g. length, weight, blood concentration. On the contrary in the latter an attribute is defined by its measuring procedure, as there is no real existence of the attribute beyond its measurement. Examples of this type are pain score, intelligence, customer loyalty. These quantities are difficult to assess objectively and thus their statistical analysis results difficult and in some cases uncertain. An entire branch of statistics, the *latent variables theory*, deals with non directly observable variables which are inferred through a mathematical model from other variables that are directly measured (Loehlin [29], 1998).

In the following subsections two different measurement approaches will be presented. In particular, the *information theory* highlights a close connection between the concept of measurement and the concept of estimation; this relationship will then consequently lead to the introduction of the main topic of the present dissertation: *the measurement error theory.*

A step further: additive conjoint measurement

Many less intuitive and more complex definitions for *measurement* have been formulated, in order to clarify a concept that might not be so straightforward as it is thought at the first sight. In the representational theory, measurement is defined as *the correlation of numbers with entities that are not numbers* (Nagel [31], 1930). This concept does not express measurement as mere assignment of a value to the related entity, since it is based on correspondences or similarities between the structure of number systems and the structure of qualitative systems; this elaborate definition is referred as *additive conjoint measurement.*

A more statistical approach to measurement: information theory

Although the entire scientific world works and develops itself through the measurement of quantities, there is a specific field of human knowledge that would not exist if the necessity of measuring and processing information were not essential: Statistics. According to information theory, measurement is: "A set of observations that reduce uncertainty where the result is expressed as a quantity"(Hubbard [23], 2007). This definition implies that all data are inexact and random in nature. Therefore, the only purpose that someone can achieve in measuring a quantity is to try to diminish the uncertainty around the real sought value, though he will never be able to reach it. In practical terms, an initial guess for the real value of a quantity is made, and then, using various methods and instruments, the uncertainty in the value is systematically reduced until the size of uncertainty for the found value is small enough for it to be considered a fair "estimate" of the real target. Since all measurement is uncertain by definition, in information theory there is no clear distinction between the word "estimate" and "measurement". Actually, instead of assigning one value to each measurement, a range of values is considered. Every statistician would now undoubtedly appreciate the parallelism between point estimate and measurement, and between range of values and confidence intervals pictured by this approach.

We live in a world of mismeasurement

Paraphrasing the quote by Hand "we live in a world of measurements", it would be more correct to say that we live in a world of mismeasurement. As already highlighted in the previous Section, information theory states that every measurement is uncertain, and thus possibly wrong, by definition. Therefore, sampling a population in order to obtain an unbiased, consistent and efficient estimator for a parameter of interest is nothing but collecting biased information in which the bias enlarges as the sample size increases, since every statistical unit in the sample contains a certain level of uncertainty that augments in augmenting the sample size. This is clearly a far too pessimistic statement that goes against a fundamental statistical principle: the bigger the sample size the better the information acquired, in terms of inference on the parameters. Nonetheless it does contain a foundation of truth: if what we want to measure is not what we actually measure or it is not measured correctly then the inferential results will be wrong and they will lead to highly biased conclusions. Mathematically, this happens when X is the true variable that we want to measure but another variable X^* is measured instead. X^* is often called a *proxy* or *surrogate* variable, since it is to some extent similar to X but not equal to X. This is called a *measurement error* or a *errors-in-variables* issue. *Measurement error theory* will be discussed in detail in the first chapter of the present work (see §1.1).

Measurement error is a problem that afflicts every scientific framework (Carroll et al [9], 2012), nonetheless there are scientific fields in which using proxies instead of the true variables of interest is a habit, because obtaining the real measurement is either too expensive or actually impossible. This usually depends on the "size" of the research field considered. Epidemiology, biostatistics and genetics deal with microscopic sizes, which possess intrinsic variability and are affected by the inaccuracy of laboratory instruments and analyses. Moreover the retrieval costs of exact measures are usually high, thus cheaper solutions are normally utilized (Kipnis et al [27], 2003).

Likewise, the same problems arise when macroscopic sizes are considered. Kelly Brandon, one of the greatest experts in measurement error linked to astronomy and author of the paper "Measurement Errors Models in Astronomy" from which the idea for this thesis was born, states: "Measurement error is ubiquitous in astronomy"(Kelly [26], 2011). Astronomical data regard collecting passive observations of space objects, where, exploiting the functional relationship between wavelength, sky location and observational time permits the astronomers to directly measure the flux of an object. Nevertheless, the number of photons detected from an astronomical object is not deterministic, but it follows a Poisson process (Timmermann and Nowak [36], 1999), whereby the intrinsic nature of astronomical data makes measurement error unavoidable.

Aim of this thesis

The present work is about functional methods for measurement error correction in Astronomy. In his paper Kelly points out that astronomical data presents measurement errors that are large, skewed and exhibit multiple modes. The first part of this work attempts to understand how the functional methods described in Section 1.4 cope with measurement error stemming from different probability distribution.

Kelly continues arguing that the unceasing technological improvement permits to have data sets with millions of rows available on a daily basis. As an example, the Sloan Digital Sky Surveys (SDSS) telescope, a major multi-filter imaging and spectroscopic redshift survey located in New Mexico (US), has produced about 200 GB of data every night since 2000 (Fiegelson and Babu [18], 2012). This volume of data enormously enhances the amount of knowledge we may obtain from its analysis, but adequate computational power must be provided. Furthermore, methods for data mining of massive data sets do not include measurement error correction techniques. The second aim of the present work is to understand whether the measurement error impact on inferential results is influenced by the sample size considered.

The thesis is organised as follows. In Chapter 1 we provide an overview of the measurement error theory and of the functional methods for dealing with measurement error, with particular emphasis on linear regression models. In Chapter 2 we describe a simulation study, with which we analyse the behaviour of the functional methods for correction in coping with three different measurement error models. In Chapter 3 we provide a regression analysis for a real astronomical dataset, in which the covariate is affected by non-linear measurement error with heteroscedastic variance. In performing the aforementioned analysis, we develop and exploit a modified version of the BCES method; thus, in Chapter 4 we report a simulation study that proves the effectiveness of our newly-developed BCES technique in coping with the specific non-linear measurement error model encountered in Chapter 3.

In the appendices, we give some notions about the skew normal distribution (Appendix A) and some methodological details of the modifications we have made at the BCES approach in order to apply it to a non-linear measurement error situation (Appendix B). Finally, the R code for the functional methods implementation, for the simulation study described in Chapter 2 and 4, and for the data analysis presented in Chapter 3 is reported in Appendix C.

Chapter 1

Measurement Error Theory

1.1 Introduction

Measurement error is the deviation of the outcome of a measurement from the true value (Fuller [19], 1987). This issue is commonly referred to in Statistics as *measurement error* or *errors-in-variables* problems. There are many sources which can induce error in the measurement and data collection:

- low accuracy and precision of the instrument used in the analysis.
- researcher's oversight.
- use of surrogate variables (e.g. average exposure to pollution in a region where the study participant lives instead of individual exposure).
- definition itself of the problem investigated (e.g. long term average of daily salt intake).

Measurement error is a problem that affects, at various levels and extents, all scientific research. Therefore, due to this pervasive presence of measurement error, an enormous amount of literature has been developed which tries to better understand the problem and to find suitable solutions error-prone variables. In the following sections the main results regarding models, effects and methods for correction will be presented.

1.2 Models and effects

Consider a general regression model of a response Y on a predictor X with a set of parameters θ:

$$Y = f(X; \theta) + \varepsilon \tag{1.1}$$

The function $f(X; \theta)$ describes how the mean of Y depends on X as a function of the parameters θ. ε represents the error term of the regression model. Nevertheless, in a situation of measurement error in the regressor, X is not directly observed: a biased value X^* is collected. X is called the *true variable* whilst the biased version X^* is the *observed variable*. Therefore, the model that is estimated by the researcher is:

$$Y = f(X^*; \theta) + \varepsilon \tag{1.2}$$

Measurement error in covariates has four different effects:

- It causes biased estimates for the parameters θ, generally attenuating the regression slope in classical linear regression and biasing it toward zero. Therefore, trends between the response and the covariate will appear reduced.

- It leads to a loss of power, causing underestimation in the relationship among the variables of interest. When the covariate is crucially contaminated by measurement error, tests of significance might state that there is no relationship between the response and the covariate, even if this is not true.

- It smears out the features of the data, producing unclear graphical model analysis

- It biases the estimate of the residual variance of the model σ^2 upwards. Thus, the variance in the response about the regression will appear larger than it really is.

The first two items are called the *double whammy* of measurement error. (Carroll et al. [9], 2012). In order to deal with errors-in-variables problems literature offers a wide set of alternative models.

1.2.1 Functional vs structural

As we have already stated above in a regression context with presence of measurement error the regressors X are not directly observable. The first decision that has to be taken is whether the regressors X are considered fixed or random.

In *functional modeling* the Xs are considered as a set of fixed, unknown constants. It is possible to consider the regressors as a set of random variables either, in this case no, or only minimal, assumptions are performed about the distribution of the X (Carroll et al. [9], 2012). This type of modeling leads to methods of estimation and inference which are robust, because no assumptions about the distribution of the unobserved Xs are made. Even though the estimators are consistent *Functional modeling* is convenient if the analyst is not interested in the estimation of X. Since there are as many unknown regressors X_i as many observations i available in the sample, it would not be possible to obtain an estimation for the Xs. Thus, in most of the cases X^* are treated as fixed constant and the analysis will be conditioned on their values, as a standard practice in regression.

When a probability function, either parametric or non-parametric, is placed on the distribution of the random Xs we are in presence of a *structural model.* X is considered as a latent random variable and assumptions about the distribution of the Xs have to be made. Inevitably the resulting estimates and inferences performed will be influenced by the parametric or non-parametric model chosen; the analysis carried out in this way therefore will not be robust. On the other hand, likelihood based confidence intervals provided by structural approaches have proved to show better coverage properties with respect to asymptotic theory underlying functional models (Guolo [20], 2005).

Nowadays, the *structural* and the *functional modeling* approaches have moved closer to each other. The idea is to choose a flexible parametric model in order to increase the quality in terms of model robustness, keeping still the advantages of a parametric analysis. For further clarification, see Mallick, Hoffman & Carroll (Mallick et al [30], 2002) and Tsiatis & Ma (Tsiatis and Ma [38], 2004).

1.2.2 Classical vs Berkson

So far the functional form that links the *true variable* X and the *observed variable* X^* has not been discussed yet. The difference between the *classical model* and the *Berkson model* is about the nature of the relationship between the *true variable* and the *observed variable.*

In the *classical model* the conditional distribution of X^* given (Z, X) is modeled. The measured variable X^* is regressed on the unobserved X and observed predictors Z, where Z are covariates measured without error. Thus, the mathematical relationship for this type of model is:

$$X^* = f(X, Z; \gamma) + U, \qquad E(U|X, Z) = 0. \tag{1.3}$$

The error structure of U may be either homoscedastic or heteroscedastic (see §1.2.3). When the functional form of f is linear, that is the truth is measured with additive error, the model obtained is called *classical measurement error* (see §1.3.1.1).

When the unobserved variable X is regressed on the measured variable X^* and observed predictors Z the model obtained is called *Berkson model.* The *Berkson model* focuses on the distribution of X given (X^*,Z). The mathematical relationship for this type of model, unlike the *classical model* described above, is:

$$X = f(X^*, Z; \psi) + U, \qquad E(U|X^*, Z) = 0 \tag{1.4}$$

The simplest additive model following the Berkson relationship is described in Section 1.3.1.2.

Determining whether real data follow a *classical* or a *Berkson* specification is simple in practice. If an error-prone covariate (i.e. a regressor measured with error) is ineluctably measured uniquely to an individual, and specifically if the measurements can be replicated, then the preference should be *classical.* Examples of this situations are blood pressure measurements and daily fat intake. When we are interested in mean exposure of a region X^* instead of individual exposure X, that is, all people in a small group are given the same value of the error-prone covariate, then the *Berkson model* is more suitable. Dust exposure in a working place and given dose in a controlled experiment are examples of this type of situation.

Another important difference between the two models refers to the error component U. In the *classical model* the error U is independent of X, or at least $E(U|X) = 0$, while in the *Berkson model* U is independent of X^*, or at least $E(U|X^*) = 0$. Therefore for the *classical model* $Var(X^*) > Var(X)$ whilst $Var(X) > Var(X^*)$ for the *Berkson model.*

There is an interesting relationship that permits to switch the from *classical model* to the *Berkson model* using Bayes theorem:

$$f_{X|X^*}(x|x^*) = \frac{f_{X^*|X}(x^*|x) f_X(x)}{\int f_{X^*|X}(x^*|x) f_X(x)\, dx} \tag{1.5}$$

where f_X is the density of X, $f_{X^*|X}$ is the density of X^* given X and $f_{X|X^*}$ is the density of X given X^*. This formula is useful in *Regression-Calibration* (see §1.4.2) where a model for X given X^* is needed, but only a model for X^* given X is available.

1.2.3 Homoscedastic vs heteroscedastic

Whether we consider the *classical model* (1.3) or the *Berkson model* (1.4), we must decide the structure of the error component U. As it has already been stated, U is a random variable with mean zero. The question here is to decide whether U is an homoscedastic random variable or an heteroscedastic one. In the former case the variance structure of U is:

$$Var(U_i) = \sigma^2. \qquad \forall i = 1 \ldots N,$$

which implies that variance of the measurement error is the same for every observation i. This type of structure is useful when the measurement error is given by an imprecision of the instrument used for the data collection.

When the structure of the measurement error is more complex and the errors-in-variables are not only caused by the inaccuracy of the instrument, it is sometimes preferable to allow the variance of the error component to vary across the observations:

$$Var(U_i) = \sigma_i^2.$$

Heteroscedasticity is useful in handling measurement errors in areas where data have to be collected and consequently codified by a computer, such as Biostatistics and Astronomy, since many sources of error are involved (Kelly [26], 2011).

1.2.4 Differential vs nondifferential error

Section 1.2.2 analysed the different possible types of relations between the *true variable* X and the *observed variable* X^*. Nevertheless, nothing has been said about the relationship between X, X^* and the response variable Y nor between X, X^* and the predictors without error Z. *Nondifferential error* occurs when X^* does not incorporate information about Y other than is available in X and Z. Technically speaking, measurement error is *nondifferential* if the distribution of Y given (X, Z, X^*) depends only on (X, Z). As a result, X^* is conditionally independent of the response given the true covariates and it is said to be a surrogate. On the other hand, if X^* does contain information about Y other than is available in X and Z the model is called *differential* and X^* cannot be considered any longer a surrogate of X.

Generally, when true and observed covariates occur at a fixed point in space and time and the response is measured at a later time the analysis can reasonably be considered as having *nondifferential* measurement error. For example blood pressure on a special day is irrelevant for coronary heart disease if long-term average is known. Notwithstanding, there are situations in which this is not the case. In case-control or choice-based sampling studies, for example, firstly the response is obtained and subsequently the covariates are measured. This ordering of measurement often origins *differential* measurement error. Furthermore, if X^* is not simply a mismeasured version of X, but it acts as a type of proxy for X, then *differential* measurement error should be used in the analysis. For a real exposure-disease situation in which this occurs, see Satten & Kupper (Satten and Kupper [33], 1993).

It is also worth highlighting that whether X^* is a surrogate depends on the remaining variables Zs present in the model and on the types of the considered response. In order to better understand this phenomenon an algebraic example, taken from Carroll et al. (2012) is presented. Suppose to have a model in which Z has two components, $Z = (Z_1, Z_2)$ and X, $Z_1, \varepsilon_1, \varepsilon_2, U_1, U_2$ are mutually independent normal random variables with zero means. Define

$$Z_2 = X + \varepsilon_1 + U_1$$

$$Y = \beta_1 + \beta_{z1} Z_1 + \beta_{z2} Z_2 + \beta_x X + \varepsilon_2$$

$$X^* = X + \varepsilon_1 + U_2.$$

Due to joint normality it is easy to show that

$$E(Y|Z_1, Z_2, X, X^*) = E(Y|Z_1, Z_2, X,).$$

Thus X^* is a surrogate in the model containing both Z_1 and Z_2. Nonetheless,

$$E(Y|Z_1, X) = \beta_1 + \beta_{z1} Z_1 + (\beta_{z2} + \beta_x) X,$$

$$E(Y|Z_1, X, X^*) = E(Y|Z_1, X) + \beta_{z2} E(\varepsilon_1 | Z_1, X, X^*).$$

The last expectation is not equal to zero because X^* depends on ε_1. Thus, X^* would be a surrogate in the model that contains only Z_1 if and only if β_{z1} were equal to zero. In this example, since the measurement error $X^* - X$ is correlated with the covariate Z_2, the presence or the absence of Z_2 in the model determines whether or not X^* is a surrogate.

The advantage of *nondifferential* measurement error is that the parameters of the response models given the true covariates can generally be estimated, even though the true covariates are not observable. Conversely, with *differential* measurement error this is not the case: apart from a few special situations, the true covariate must be observed on some study subjects. This is the reason why *nondifferential* measurement errors are definitely more used in dealing with errors-in-variables than the *differential* ones.

1.3 Linear regression

In the previous section we provided a description of the models present in literature related to measurement error theory. This section will focus on the class of linear regression models. Firstly measurement error theory in simple linear models will be presented and consequently it will be generalized to multiple linear regression.

The effects of measurement error in linear regression are influenced by multiple factors: the level of error in the measurement; whether or not the predictor measured with error is univariate or multivariate and the regression model itself, being simple or multiple. These characteristics could lead to different inaccurate results in the analysis:

- biasing the slope estimate in the direction of zero. This bias is referred to in literature as *attenuation* or *attenuation to the null* (Fuller [19], 1987).

- observed data present relationships that do not occur in the error-free data.

- the direction of the relation, i.e., the sign of the estimated parameter, is reversed as compared to the case with no measurement error. This phenomenon leads to a complete misunderstanding of the relations between the variables in the model.

The effects of measurement error, and how they can be corrected, depend on the measurement error models chosen from the ones presented in the previous

Section. The most widely used models will be presented and analysed in the subsections below.

1.3.1 Simple linear regression

1.3.1.1 Classical nondifferential homoscedastic measurement error

Suppose to have a simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ in which the covariate X is measured with error. This means that the model estimated will actually be $Y = \beta_0 + \beta_1 X^* + \varepsilon$. When the measurement error component has the following characteristics:

$$X_i^* = X_i + U_i$$

$$E(U_i) = 0 \qquad Var(U_i) = \sigma_U^2 \tag{1.6}$$

the obtained model is called *classical error model*. It is the simplest additive model, notwithstanding it is the most widely used in practise. Being a linear model, it is possible to estimate the parameters β_0 and β_1 with the ordinary least squares (OLS) method. Indicating $plim(\cdot)$ the probability limit of a quantity, for the slope we obtain:

$$\hat{\beta}_1 = \frac{S_{YX^*}}{S_{X^*}^2}$$

$$plim(\hat{\beta}_1) = \frac{\sigma_{YX^*}}{\sigma_{X^*}^2} = \frac{\sigma_{YX}}{\sigma_X^2 + \sigma_U^2} = \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}, \tag{1.7}$$

while for the intercept we get:

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}^*$$

$$plim(\hat{\beta}_0) = \mu_Y + \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} \mu_X = \beta_0 + \beta_1 (1 - \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}) \mu_X, \tag{1.8}$$

and for the residual variance $Var(Y|X^*)$:

$$MSE = S_Y - \hat{\beta}_0 - \hat{\beta}_1 X^*$$

$$plim(MSE) = \sigma_\varepsilon^2 + \frac{\beta_1^2 \sigma_U^2 \sigma_X^2}{\sigma_X^2 + \sigma_U^2} = \sigma_\varepsilon^2 + \lambda \beta_1^2 \sigma_U^2. \tag{1.9}$$

In literature this is called *naive LS-estimation* because it does not take into account the presence of measurement error. It is visible from the formulas that the estimators obtained above are not consistent, because they do not converge to the real values β_0 and β_1. The quantity

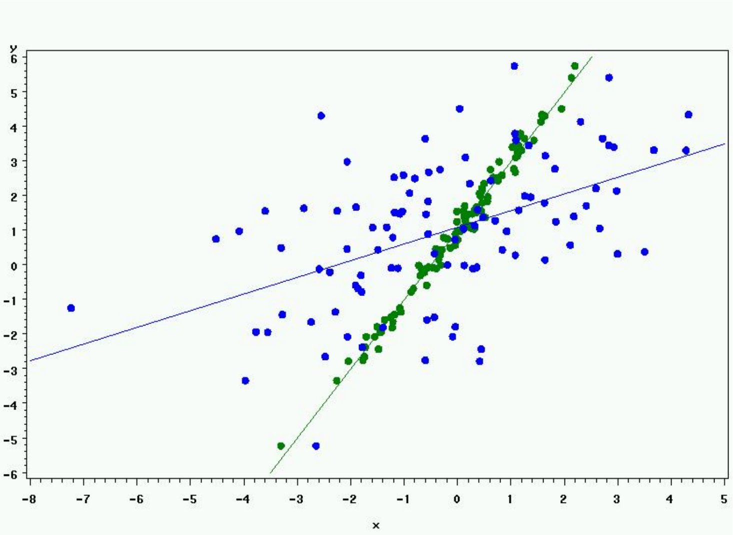


Figure 1.1: Effect of additive measurement error on linear regression. The green line and dots are for the true X data, while the blue line and dots are for the observed, error-prone X^* data. The slope to the true X data is steeper, and the variability about the line smaller.

$$\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} < 1 \tag{1.10}$$

is called the *reliability ratio* and, being smaller than 1, it causes an *attenuation to zero* of the estimate. The graph in Figure 1.1 shows an example of additive measurement error model. The green line and dots represent the true X data, while the blue line and dots represent the observed, error-prone X^* data. The slope to the true X data is steeper and the variability about the line is smaller. The variance of the *naive estimator* is smaller than the variance of the true-data estimator asymptotically if and only if

$$\frac{\beta_1^2 \sigma_X^2}{\sigma_X^2 + \sigma_U^2} < \frac{\sigma_\varepsilon^2}{\sigma_X^2}$$

which could occur when β_1^2 is small or either σ_ε^2 or σ_U^2 is large.

1.3.1.2 Berkson Error

A *Berkson error* structure for the measurement error component is defined as follow:

$$X_i = X_i^* + U_i, \qquad U_i \perp (X_i^*, Y_i), \quad E(U_i) = 0 \tag{1.11}$$

Therefore, it is straightforward to show that $E(X_i|X_i^*) = X_i^*$ which leads to

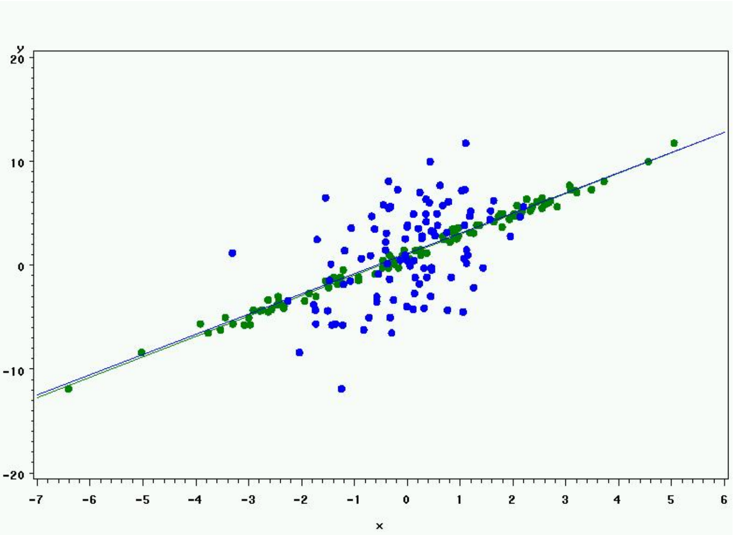


Figure 1.2: Effect of Berkson error on linear regression. The green line and dots are for the true X data, while the blue line and dots are for the observed, error-prone X^* data. Theory shows that the fit of Y to X is unbiased for the regression of Y to X. The two fits are similar.

$E(Y_i|X_i^*) = \beta_0 + \beta_1 X^*$. As a consequence, the naive estimator that regresses

Y_i on X_i^* is unbiased for β_0 and β_1.

This result is shown in Figure 1.2, in which it is possible to observe that the fit of Y_i to X_i^* (green line and points) is unbiased for the regression of Y_i on X_i (blue line and points), and the two fits are, in fact, similar.

In conclusion, linear models with *Berkson error* do not need a method for correction to be implemented in order to rectify the bias caused by measurement error.

1.3.1.3 Differential measurement error

In presence of *differential measurement error* the observed variable X^* contains more additional information about Y than is available only in X. This is the most troublesome type of error because bias correction bias requires the largest amount of additional information (Carroll et al. [9], 2012). The estimators for respectively the slope, the intercept and the residual variance in presence of *differential measurement error* in a simple linear model converge to the following quantities:

$$plim(\hat{\beta}_1) = \beta_1\left(\frac{\sigma_{XX^*}}{\sigma^2_{X^*}}\right) + \frac{\sigma_{\varepsilon X^*}}{\sigma^2_{X^*}} \tag{1.12}$$

$$plim(\hat{\beta}_0) = \beta_0 + \beta_1\mu_X - \frac{\beta_1\sigma_{XX^*} + \sigma_{\varepsilon X^*}}{\sigma^2_{X^*}}\mu_{X^*} \tag{1.13}$$

$$plim(MSE) = \sigma^2_\varepsilon + \beta_1^2\sigma^2_X - \frac{(\sigma_{XX^*} + \sigma_{\varepsilon X^*})^2}{\sigma^2_{X^*}} \tag{1.14}$$

It is worth noticing that, in order to estimate β_1 from the regression of Y on X^*, knowledge of or estimability of both the covariances σ_{XX^*} and $\sigma_{\varepsilon X^*}$ is necessary.

1.3.1.4 Nondifferential measurement error

When X^* does not add additional information to the regression of Y on the real predictor X, then X^* is called a *surrogate*. In presence of *nondifferential measurement error* in a simple linear model OLS estimation leads to the following results:

$$plim(\hat{\beta}_1) = \beta_1(\frac{\sigma_{XX^*}}{\sigma^2_{X^*}}) \tag{1.15}$$

$$plim(\hat{\beta}_0) = \beta_0 + \beta_1\mu_X - \frac{\beta_1\sigma_{XX^*}}{\sigma^2_{X^*}}\mu_{X^*} \tag{1.16}$$

$$plim(MSE) = \sigma^2_\varepsilon + \beta_1^2\sigma^2_X \tag{1.17}$$

As it is evident from Equation (1.15), only σ_{XX^*} has to be either known or estimated in order to recover β_1 from the regression of Y on X^*. Since a surrogate is always less informative than X, the residual variance $Var(Y|X^*)$ in the regression of Y on X^* is always greater than the residual variance σ^2 of the regression of Y on X.

1.3.1.5 Berkson/classical mixture measurement error

In Sections 1.3.1.1 and 1.3.1.2 both *classical* and *Berkson* errors have been discussed. Nevertheless, another situation in which both error components are present at the same time could be possible. Particularly, it is assumed that

$$X = L + U_b, \tag{1.18}$$

$$X^* = L + U_c. \tag{1.19}$$

This particular structure leads to a *classical error model* when $U_b = 0$, and to a *Berkson error* model when $U_c = 0$.

The mixture situation presents the problems of both *classical* and *Berkson* errors. The limits of convergence for this model are:

$$plim(\hat{\beta}_1) = \beta_1\frac{\sigma^2_L}{\sigma^2_L + \sigma^2_{U_c}} \tag{1.20}$$

$$plim(\hat{\beta}_0) = \beta_0 + \beta_1\mu_X(1 - \frac{\sigma^2_L}{\sigma^2_L + \sigma^2_{U_c}}) \tag{1.21}$$

$$plim(MSE) = \sigma^2_\varepsilon + \beta_1^2\sigma^2_X \tag{1.22}$$

where $\sigma^2_{U_c}$ and σ^2_L denote the *Berkson error* variance and the variance of the mixture component L respectively. It is worth noting that there is bias in the regression parameters when $\sigma^2_{U_c} > 0$, as in the *classical* model, because as it has already been stated in Subsection 1.3.1.2, the Berkson component does not introduce bias in parameter estimation.

1.3.1.6 Measurement error in the response variable

So far only measurement errors in covariates have been analysed. However, it may happen that the response variable Y is the variable measured with error: $Y_i^* = Y_i + U_i$. The model obtained will then be

$$Y^* = \beta_0 + \beta_1 X + \varepsilon + U, \tag{1.23}$$

where $\varepsilon + U$ is the new, larger, error term. Assuming the independence between U and X and and between U and ε, the measurement error in response does not cause bias in the estimation of the parameters, i.e., the OLS estimators are still consistent. Measurement error in the response only causes an increase in variability of the error term. It is therefore straightforward dealing with it as long as the error components are independent, which is almost always the case.

1.3.2 Multiple linear regression

1.3.2.1 Single covariate measured with error

In multiple linear regression the bias caused by measurement error is more tricky and difficult to treat, even for the *classical error* model. Suppose to have a multiple regression model:

$$Y = \beta_0 + \beta_1 X + \beta^T Z + \varepsilon \tag{1.24}$$

in which X is scalar and Z and β are column vectors. The measurement error structure is

$$X_i = X_i^* + U_i \qquad U_i \perp (X_i, Z_i) \qquad U_i \perp \varepsilon_i \tag{1.25}$$

The converge in probability of the estimator obtained from the *naive estimation* of the parameter β_1 is:

$$plim(\hat{\beta}_1) = \beta_1 \frac{\sigma^2_{X|Z}}{\sigma^2_{X^*|Z}} = \beta_1 \frac{\sigma^2_{X|Z}}{\sigma^2_{X|Z} + \sigma^2_U} = \beta_1 \lambda_1, \tag{1.26}$$

where $\sigma^2_{X^*|Z}$ and $\sigma^2_{X|Z}$ are the residual variances of the regression of X^* on Z and of X on Z, respectively. Note that, in general, λ_1 is smaller than the simple linear regression attenuation λ given by expression (1.10). $\lambda_1 = \lambda$ if and only if X and Z are uncorrelated This leads to an enhancement of the *attenuation to the null* in the multiple regression case. Moreover, the measurement error in X causes inconsistent estimation also for the parameters of the covariates measured without error, unless Z is independent of X. Carroll, Gallo and Gleser showed that (Carroll et al. [8] 1985)

$$plim(\hat{\beta}) = \beta + \beta_1(1 - \lambda_1)\Gamma \tag{1.27}$$

where Γ^T is the coefficient of Z in the regression $E(X|Z) = \Gamma_0 + \Gamma^T Z$.

As a significant example, in the particular case of analysis of covariance,

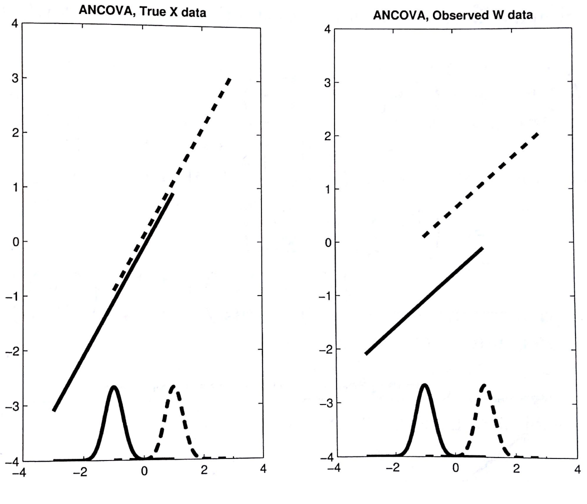


Figure 1.3: Effect of measurement error in an unbalanced analysis of covariance (taken from Carroll et al. 2012)

that is when Z is a categorical variable, measurement error in X can completely twist the results of the analysis. With respect to a two-group analysis of covariance, where Z is a treatment assignment variable, Carroll proved that the naive analysis can lead to observe a treatment effect when it actually does not exist or to note a positive effect when it is negative and vice versa (Carroll [6], 1989). Figure 1.3 highlights the previous statement. The left panel shows the (Y, X) fitted function, since the solid and the dotted line are close to each other there is no sign of treatment effect, even though the distribution of X in the two groups are very different, as can be seen at the bottom of the panel. The right panel shows the (Y, X^*) fitted function, in which there is measurement error in the continuous covariate. The error-in-variables *attenuates the mean in each group*, suggesting that there is a treatment effect, though if this is not true.

1.3.2.2 Multiple covariates measured with error

The model which defines the situation of multiple covariates measured with error is the following:

$$Y = \beta_0 + \beta_1^T X + \beta_2^T Z, \tag{1.28}$$

where X may consist of multiple predictors. The generalization from equation in (1.26) with X scalar is straightforward; using matrix calculation the naive ordinary least squares method leads to

$$plim \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} + \Sigma_{UU} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_X & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \tag{1.29}$$

in which Σ_{AB} defines the covariance matrix between random variables A and B. As it has already been underlined in the previous case with one error-prone variable, also in the case with multiple covariates measured with error the presence of errors-in-variables can bias the entire inference also with respect to the parameters of the error-free variables Z.

1.4 Methods for correction

1.4.1 Introduction

In the previous sections we discussed and analysed the main effects and models used in measurement error theory, with particular focus on linear regression. Furthermore, we presented the inference problems according to parameters estimation in having to do with error-prone variables. Methods and procedures which try to solve these problems will be introduced now.

As already presented in Section 1.2.1, two main categories of methods for correction can be identified: *functional methods* and *structural methods*. In the former approach little information on X is required, but a large number of parameters have to be estimated; on the contrary the latter method requires a less amount of parameters to be estimated, but both information on and validation of the exposure distribution is needed. The choice between a *functional* or a *structural* model usually depends on the assumptions made and on the form of the regression model (Guolo [20], 2005). The present work will focus on functional methods for correction, for an exhaustive and clear description of structural methods for correction see for example Carroll el al. (2012).

Generally, in order to avoid lack of identifiability of the parameters, additional information on the real variable X is needed. This additional information can either be *internal* in the form of subsets of the main data, also called primary dataset, or *external* in the form of independent studies. The additional data can be subdivided in three different types:

- *validation data*, in which the *gold standard* measurement is available, that is, a direct observation of X. Validation studies are very useful because they furnish a direct estimate of some error characteristics, as moments of the distribution. Nevertheless, exact measures of X might be really expensive and hard to obtain, therefore validation data are generally available only for a subset of the primary data.

- *replication data*, in which the same statistical unit is subjected to more replicated observations of X^*.

- *instrumental data*, in which another variable Z is observed in addition to X^*

In collecting additional information, an important aspect that has to be taken into account and monitored during the study design definition, is the trade-off between cost and information. Spiegelman (1994) presents a set of various principles and criteria useful for creating cost-efficient study designs in presence of mismeasured covariates (Spiegelman [34], 1994).

In the following chapter a set of functional methods, useful for obtaining consistent estimators also in presence of measurement error will be presented.

1.4.2 BCES method

In linear models the ordinary least squares estimates of the intercept, slope and residual variance are obtained from the sample moments of the data. Nevertheless, as shown in Section 1.3.1.1, in presence of measurement error in the covariate, the sample moments are biased estimates of the moments of the true distribution. Therefore, a straightforward method of handling measurement error in linear regression is to estimate the moments of the true value of the data and then to exploit these for estimating the regression parameters. The idea behind the bivariate correlated errors and intrinsic scatter (BCES) method is to use the real moments of the variables in order to correct for the bias in the parameters estimates due to measurement error. Firstly introduced by Akritas & Bershady (Akritas & Bershady [1], 1996), the BCES method is a direct generalization of the OLS estimator. It is generally applicable when both the covariate and the response present measurement error and even when the magnitude of the latter depends on the measurement (i.e., the measurement error has heteroscedastic variance).

For the sake of illustration, consider a simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ with additive measurement error in the covariate $X_i^* = X_i + U_i$. The variance of U_i can either be homoscedastic or heteroscedastic and it is assumed known, which is a fairly common assumption in all astronomical data sets. As shown in equation (1.29), *naive LS-estimation* produces an inconsistent estimator for β_1. The BCES estimator replaces the population moments with moment estimators from the observed data, that is

$$\hat{\beta}_1^{BCES} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i^* - \bar{X}^*)}{\sum_{i=1}^{n}(X_i^* - \bar{X}^*)^2 - \sum_{i=1}^{n} U_i^2} \tag{1.30}$$

$$\hat{\beta}_0^{BCES} = \bar{Y} - \hat{\beta}_1^{BCES}\bar{X}^* \tag{1.31}$$

As it is clearly visible from equation (1.30), the BCES estimator “debiases” the sample variance of X^* by subtracting the scatter due to the measurement error U. In order to calculate the variance of the BCES estimators, the following quantities need to be defined:

$$\xi_1 = \frac{(X^* - E(X^*))(Y - \beta_1 X^* - \beta_0) + \beta_1 U^2}{Var(X^*) - E(U^2)} \tag{1.32}$$

$$\xi_2 = Y - \beta_1 X^* - E(X^*)\xi_1 \tag{1.33}$$

Their estimates $\hat{\xi}_1$ and $\hat{\xi}_2$ are obtained by substituting the unknown quantities with the sample ones, and $\hat{\beta}_1^{BCES}$, $\hat{\beta}_0^{BCES}$ in place of β_1, β_0. The variance of $\hat{\beta}_1^{BCES}$ and $\hat{\beta}_0^{BCES}$ are then estimated by:

$$\hat{\sigma}_{\beta_1}^2 = \frac{1}{n^2}\sum_{i=1}^{n}(\hat{\xi}_{1i} - \bar{\hat{\xi}}_1)^2 \tag{1.34}$$

$$\hat{\sigma}_{\beta_0}^2 = \frac{1}{n^2}\sum_{i=1}^{n}(\hat{\xi}_{2i} - \bar{\hat{\xi}}_2)^2 \tag{1.35}$$

where $\bar{\hat{\xi}}_1$ and $\bar{\hat{\xi}}_2$ denote the arithmetic mean of $\hat{\xi}_1$ and $\hat{\xi}_2$. The example just described is a little bit less complex than the one used by Akritas & Bershady in their paper, in which measurement error afflicts both the covariate and the response. Since in both simulation experiments of Chapter 2 and in the real data problem analyzed in Chapter 3 only the covariate presents measurement error, a simplified version of the BCES estimator is applied. Thus, the regression reported was chosen in order to describe the theoretical situation encountered in the last part of the thesis.

Akritas & Bershady proved that the BCES estimator is asymptotically unbiased and its finite sample distribution is asymptotically normal (Akritas

& Bershady [1], 1996). The main advantage of the BCES estimator is that it does not make any assumptions about the distribution of the random variables present in the model. Therefore we can refer to the BCES estimator as a robust estimator. Nevertheless it loses precision when further information about the distribution of the measurement error or on the covariates is available, as it does not make any assumptions on the distribution of the variables. Furthermore, this estimator tends to be highly variable when the sample size is small and the measurement error is large. In conclusion, despite the robustness and its good behaviour in simple cases (see §2.2), when the sample size is small and the measurement errors produces a significant increase in the data variability, more stable estimators should be used.

1.4.3 Regression-Calibration

In the previous subsection the BCES method has been treated and its ability to adjust for the effects of errors-in-variables has been described. However, this method is feasible only with linear models or, more generally, with models whose parameters estimation has a close form. *Regression calibration*, initially suggested by Rosner, Willett and Spiegelman (Rosner et al [32], 1989) and successively modified by Carroll and Stefanski (Carroll & Stefanski [10], 1990), is simple and potentially applicable to any regression model, provided a sufficiently accurate approximation of the true values of the parameters. The basic idea of *Regression calibration* is to replace the *true variable* X by the regression of X on (Z, X^*), where Z and X^* represent respectively the error-free covariates and the error-prone observed variable. The fitted values obtained are consequently used to perform a standard analysis with the original model. This procedure can be described as an algorithm with three main steps:

1. Estimate the regression of X on (Z, X^*) with $m_X(Z, X^*, \gamma)$ which depends on the parameters γ. The estimations $\hat{\gamma}$ can be found using validation data or replications (Carroll et al. [9], 2012).

2. Replace the unobservable X by its estimate $m_X(Z, X^*, \hat{\gamma})$ in the main model and run a standard analysis to obtain the parameter estimates.

3. Adjust the estimate of the variance to account for the estimation of γ by using the bootstrap, jackknife or sandwich method.

Suppose that the model that has to be estimated is:

$$E(Y|Z,X) = m_Y(Z,X,\pi) \tag{1.36}$$

in which the mean of Y is regressed by (X,Z) for some unknown parameters π. Replacing the unobservable value X by its estimate $m_X(Z,X^*,\hat{\gamma})$ creates a modified model for the observed data, that will become:

$$E(Y|Z,X) \approx m_Y\{Z, m_X(Z,X^*,\gamma), \pi\}. \tag{1.37}$$

The *regression calibration model* obtained in (1.37) is an approximate working model for the observed data, which can be used for correcting for measurement error presence in regression models.

An example of how this procedure works in a simple case is presented. Consider the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$, where the covariate X is affected by measurement error. The first RC algorithm step is to provide a “debiased” version of the error-prone variable X^*. Let assume that a subset of the sample, in which all the X, X^* and Y variables were measured, is available. This is an *internal validation data* situation, in which the exact measure for X (gold standard) is observed for a small part of the considered sample. This situation is fairly common in medical statistics and biostatistics, where the measure of the true variable X could be feasible, but being the data collection task either too expensive or time-consuming, it is performed only for few cases amongst the entire dataset. It is less common in astronomy, where usually the considered variables present an intrinsic scatter and the corresponding true values cannot be directly observed. Using the internal validation data, the subsequent regression is created:

$$X = \gamma_0 + \gamma_1 X^* + \nu, \tag{1.38}$$

where the estimates $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are obtained using OLS. Next, in Step 2, for every X^* present in the dataset a “debiased” value $\hat{X}$ is calculated using the expression (1.38):

$$\hat{X} = \hat{\gamma}_0 + \hat{\gamma}_1 X^*. \tag{1.39}$$

The second part of Step 2 involves using the fitted values obtained from Equation (1.39) as a covariate for the original model, that becomes:

$$Y = \pi_0 + \pi_1 \hat{X} + \epsilon. \tag{1.40}$$

Again, the parameters are estimated using OLS. The estimated values $\hat{\pi}_0$ and $\hat{\pi}_1$ are the *regression calibration estimates* $\hat{\beta}_0^{RC}$ $\hat{\beta}_1^{RC}$ for the initial model. As previously stated, in Step 3 the standard errors for these estimates must be corrected in order to account for the fact that X is estimated in the previous step. Generally, a non-parametric bootstrap or jackknife is used (see §1.4.3.1 and §1.4.3.2)

As though the *regression calibration model* is a straightforward technique widely applied in empirical studies, it also has some drawbacks which have to be taken into consideration. X is not observed, therefore replacing its value by the estimate $m_X(Z, X^*, \hat{\gamma})$ cannot be done in an ordinary way. That is, additional data must somehow be provided in order to make the first step of the RC algorithm feasible. Literature offers many available procedures for this, see Carroll et al for a collection of possible solutions (Carroll et al. [9], 2012). The measurement errors have to be nondifferential with small variance and the model relating X to X^* has to be nearly homoschedastic and linear. If these assumptions are not satisfied, RC can be inefficient in reducing bias, especially in non linear models.

The following sections briefly present how to compute the standard errors for the RC estimates, via the bootstrap and the jackknife methods.

1.4.3.1 The bootstrap method for variance estimation

Bootstrap methods use a non-parametric approach to construct estimates based on a bootstrap sample of the data. Bootstrapping means to create additional data sets by re-sampling with replacement the original data B times (Efron [15], 1979). For every re-sampled data set, the estimate of interest is calculated, then, taking the average above all the B simulated data sets leads to the *bootstrap parameter estimate*. In particular, in a regression calibration context we are interested in estimating the RC standard errors of the parameters. To do that, the data used in (1.40) are bootstrapped and B RC estimates are computed. Let denote $\hat{\gamma}_b$ the bth RC estimate, with $b = 1 \ldots B$. The bootstrap RC variance estimator is

$$\hat{Var}(\hat{\gamma}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\gamma}_b - \hat{\gamma})^2, \qquad (1.41)$$

where $\hat{\gamma} = B^{-1}\sum_{b=1}^{B}\hat{\gamma}_b$. In general, bootstrap methods are robust, since no distributional assumption is made, and are easy to implement in whichever statistical software. Nonetheless, they are computationally intensive, since the estimation algorithm is performed B times, one for each re-sampled dataset. Bootstrap methods are useful when the theoretical distribution of a statistic of interest is complex or unknown, as in the RC method, where the standard error must account for estimation of $\bar{X}$ estimation.

1.4.3.2 The jackknife method for variance estimation

Likewise the bootstrap technique, the jackknife is a non-parametric method for computing estimates. It initially consists in the construction of the so-called jackknife samples. Let assume to have a dataset with N observations. In the ith jackknife sample, $i = 1 \ldots N$, every observation but the ith is included. This leads to obtain N jackknife samples, each with $N - 1$ observations. Define $\hat{\gamma}_{-i}$ as the RC estimator of γ computed from the ith jackknife data set. The jackknife variance estimator is

$$\hat{Var}(\hat{\gamma}) = \frac{N-1}{N}\sum_{i=1}^{N}(\hat{\gamma}_{-i} - \hat{\gamma}_.)^2, \qquad (1.42)$$

where $\hat{\gamma}_. = N^{-1}\sum_{i=1}^{N}\hat{\gamma}_{-i}$. The jackknife method is easy to implement and it is conservative, meaning that the real value of $Var(\hat{\gamma})$ is likely to be smaller than the variance jackknife estimation (Efron and Stein [16], 1981). For these reasons, the jackknife method is used to compute the RC standard errors in the simulation experiment of Chapter 2. The R code for the jackknife method can be found in Appendix C.1

1.4.4 Simulation-Extrapolation (SIMEX)

The *simulation extrapolation* (SIMEX) method is a simulation-based functional method that shares many properties with the *regression calibration* technique described above: it is easily applicable and widely used for its efficiency, even though the computational burden is larger. It makes no assumption on the distribution of the variables and it is specifically suitable to problem with additive measurement error and to any problems in which the measurement error structure can be generated on a computer via Monte Carlo methods (Carroll et al. [9], 2012).

First proposed by Cook & Stefanski (Cook and Stefanski [12], 1994) the basic idea behind the SIMEX technique is that the effect of measurement error can be determined and thus corrected via simulation. The method concerns in computing many *naive-estimates* by adding additional measurement error to the data: this generates a trend of measurement error-induced bias from which the case of no measurement error is extrapolated back. The SIMEX procedure for obtaining the bias-corrected estimates is developed in two different steps. Firstly, of the so called *SIMulation step*, measurement error is added increasingly to the original X^* values by simulation and the regression parameters obtained from this error-incrementing process are estimated. Secondly, of the *EXtrapolation step*, the relationship between the parameter estimates and the variances of the measurement error is modeled, in order to extrapolate the estimates back to the case of no measurement error.

1.4.4.1 Homoscedastic errors with Known variance

An example of an application of the SIMEX to the classical measurement error structure of Equation (1.6) will clarify the just described method. During simulation step M-1 additional data sets of increasingly larger measurement error $(1 + \lambda_m)\sigma_u^2$, where $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_M$, are simulated. In the SIMEX method an initial assumption is to consider σ_u^2 known or easily estimable, since it is needed for the generation of the simulated-error. For any $\lambda_m \geq 0$, define

$$X_b^*(\lambda_m) = X^* + \sqrt{\lambda_m} U_b, \qquad b = 1, \ldots, B, \tag{1.43}$$

where $\{U_b\}_{b=1}^B$ are the B mutually independent and identically distributed computer-generated pseudo errors.

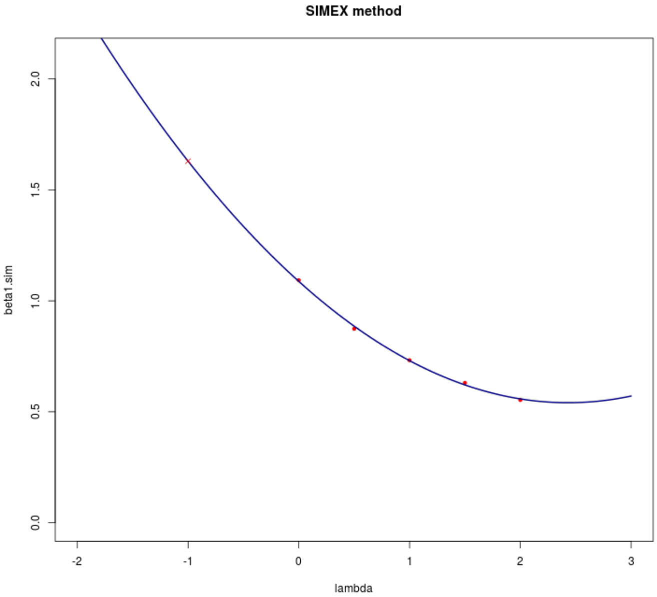


Figure 1.4: Example of the effect of the measurement error of size $(1+\lambda_m)\sigma_u^2$ on parameter estimate. The x-axis is λ, and the y-axis is the estimated coefficient. The SIMEX estimate is an extrapolation to $\lambda = -1$. The naive estimate occurs at $\lambda = 0$.

It is worth noticing that

$$var(X_b^*|X_i) = (1 + \lambda_m)\sigma_u^2 = (1 + \lambda_m)var(X_i^*|X_i), \tag{1.44}$$

which equals 0 when $\lambda_m = -1$: this is the key property of the pseudo data simulation. Consider a generic regression parameter θ that has to be estimated. For every data set the naive estimate $\hat{\theta}_b(\lambda_m)$ of θ is calculated and the average value of the B naive estimates is obtained

$$\hat{\theta}(\lambda_m) = \frac{\sum_{b=1}^{B} \hat{\theta}_b(\lambda_m)}{B}. \tag{1.45}$$

In the extrapolation step $\{\hat{\theta}(\lambda_m), \lambda_m\}_{m=1}^{M}$ is modeled as a function of λ_m for $\lambda_m \geq 0$ in order to extrapolate the fitted models back to $\lambda = -1$, that is, when the measurement error in the parameters is equal to 0.

A functional form for the extrapolant function has to be chosen. Generally literature suggests to use either a quadratic or a linear pattern (Carroll, Ruppert and Stefanski [7], 1995).

The SIMEX algorithm in case of a simple linear regression model is illustrated in Figure 1.4. The red points represents the estimates $\{\hat{\beta}_1(\lambda_m), \lambda_m\}_{m=1}^{M}$ whilst the red X shows the SIMEX estimate of the parameter obtained from a quadratic extrapolant function (blue line). The red dot in correspondence of $\lambda = 0$ represents the naive estimator, that is when no computer-generated error is added to the data.

1.4.4.2 Heteroscedastic errors with Known variance

The SIMEX method is useful for correcting for measurement error even when the measurement error structure presents heteroscedastic variance, with almost no complication in the procedure. Suppose that $X_i^* = X_i + U_i$, where U_i is a normal random variable with variance $\sigma_{u,i}^2$, and it is independent of X_i and Y_i. This heteroscedastic error structure provides a change in the simulated error generation, that will become:

$$X_{b,i}^*(\lambda_m) = X_i^* + \sqrt{\lambda_m} U_{b,i}, \qquad i = 1, \ldots, n, \qquad b = 1, \ldots, B, \tag{1.46}$$

where the pseudo errors $\{U_{b,i}\}_{i=1}^{n}$ are still mutually independent and independent of all the observed data. Nonetheless, in this situation the pseudo errors distribution varies amongst the observations: for every statistical unit $U_{b,i}$ follows a normal distribution with mean 0 and variance $\sigma_{u,i}^2$, that is, it is different for each i. Note that the conditional measurement error variance

$$var(X^*_{b,i}|X_i) = (1+\lambda_m)\sigma^2_{u,i} = (1+\lambda_m)var(X^*_i|X_i) \tag{1.47}$$

equals 0 when $\lambda = -1$, as in equation (1.44). Consequently the extrapolation step is done in exactly the same way as in the case of homoscedastic error.

A tricky part in the SIMEX procedure is to provide a reasonable estimation for the standard errors of the coefficients. This can be done either via the bootstrap or the sandwich method. The implementation of the former is straightforward, though it requires considerable computing time in order to be carried out. Primarily for this drawback, the sandwich method is used to obtain SIMEX standard errors. In the following section the procedure for computing the *SIMEX sandwich variance estimator* in presence of homoscedastic measurement error is described. For the case of heteroscedastic error, see Devanarayan [14] (1996).

1.4.4.3 Simulation-extrapolation variance estimation

The *SIMEX sandwich variance estimator* procedure was firstly implemented by Stefanski and Cook in 1995 (Stefanski and Cook [35], 1995). As already pointed out for the SIMEX estimates, this variance estimation method is applicable only when the measurement error variance σ^2_u is known.

Let us introduce a function T which denotes the estimator of the parameter θ under study. $T\{(Y_i, X^*_i)^n_1\}$ represents the naive estimator for the parameter θ. Consider a generic naive SIMEX estimate:

$$\hat{\theta}_b(\lambda) = T\{(Y_i, X^*_i + \sqrt{\lambda}U_{b,i})^n_1\}$$

and define

$$\hat{\theta}(\lambda) = E\{\hat{\theta}_b(\lambda)|(Y_i, X^*_i)^n_1\}. \tag{1.48}$$

The expectation in Equation (1.48) depends only on the distribution of $\{U_{b,i})^n_1\}$, since we condition on the observed data. $\hat{\theta}(\lambda)$ is obtained by considering the limit $B \to \infty$ of the average $\{\hat{\theta}_1(\lambda)+\cdots+\hat{\theta}_B(\lambda)\}/B$. An associated variance estimator is also introduced with the following notation:

$$T_{var}\{(Y_i, X_i^*)_1^n\} = \hat{var}(\hat{\theta}_{true}) = \hat{var}[T\{(Y_i, X_i^*)_1^n\}],$$

where $\hat{\theta}_{true}$ represents the "estimator" computed from the "true" data $(Y_i, X_i)_1^n$. Let us use τ^2 to denote the parameter $var(\hat{\theta}_{true})$, τ^2_{true} to denote the true variance estimator $T_{var}\{(Y_i, X_i)_1^n\}$ and τ^2_{naive} to denote the naive variance estimator $T_{var}\{(Y_i, X_i^*)_1^n\}$. Stefanski and Cook proved that

$$E\{\hat{\theta}_{simex}|(Y_i, X_i)_1^n\} \approx \hat{\theta}_{true}. \tag{1.49}$$

The approximation is due to both a large-sample approximation and the chosen extrapolant function (Stefanski and Cook [35], 1995). From equation (1.49) it follows that

$$var(\hat{\theta}_{simex}) \approx var(\hat{\theta}_{true}) + var(\hat{\theta}_{simex} - \hat{\theta}_{true}) \tag{1.50}$$

in which the variance of $\hat{\theta}_{simex}$ is decomposed into two different components: the former due to sampling variability $var(\hat{\theta}_{true}) = \tau^2$ and the latter due to measurement error variability $var(\hat{\theta}_{simex} - \hat{\theta}_{true})$. The former component can be estimated using the SIMEX variance estimate $\hat{\tau}^2(\lambda)$. $\hat{\tau}^2(\lambda)$ is calculated computing

$$\hat{\tau}_b^2(\lambda) = T_{var}[\{Y_i, X_{b,i}^*(\lambda)\}_1^n]$$

for each b, $b = 1, \ldots, B$ and then taking the mean. In order to obtain the second variance component, let us define the two quantities:

$$\Delta_b(\lambda) = \hat{\theta}_b(\lambda) - \hat{\theta}(\lambda), \quad b = 1, \ldots, B, \tag{1.51}$$

$$s^2_\Delta(\lambda) = (B-1)^{-1} \sum_{b=1}^{B} \Delta_b(\lambda)\Delta_b^T(\lambda). \tag{1.52}$$

The formula in (1.52) is the sample variance matrix of $\{\hat{\theta}_b(\lambda)\}_{b=1}^B$ and it represents an unbiased estimator for the conditional variance

$$var\{\hat{\theta}_b(\lambda) - \hat{\theta}(\lambda)|(Y_i, X_i^*)_1^n\}$$

for all $B > 1$.

Having estimated both the sampling variability variance $\hat{\tau}^2(\lambda)$ and the measurement error variance $s^2_\Delta(\lambda)$, the procedure terminates regressing the components of the difference $\hat{\tau}^2(\lambda) - s^2_\Delta(\lambda)$ on the λ values and extrapolating back to $\lambda = -1$; the fitted value obtained provides an estimate of $var(\hat{\theta}_{simex})$.

It is worth highlighting that the entire technique is approximate, meaning that it is valid only when the measurement error variance is small and the sample size is large (Carroll et al [9], 2012). Furthermore, it is not guaranteed that the variance so obtained is a positive number, since the extrapolation step does not put constraints on the parameter space.

Chapter 2

Simulation Study

2.1 Introduction

The present chapter describes a simulation study performed in a simple linear regression context with different types of measurement error. In particular, the aim of the simulations is to understand how the functional methods for correction described in the previous chapter cope with the mismeasured covariate and whether they can achieve a significantly improvement when making inference on the parameters. The real model used for the simulations is the following:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.1}$$

where $\beta_0 = 7$, $\beta_1 = 2$ and $\epsilon \sim N(0,1)$. x is randomly generated by a normal distribution with 0 mean and variance equal to 4. Nevertheless the true covariate x is not directly known: a mismeasured value $w = x + u$ is observed, where u represents the measurement error component. A classical error model structure (see §1.2.2) was therefore selected for the simulation. This is a reasonable choice since this hypothesis is often made when dealing with measurement error in an empirical framework (Carroll et al [9], 2012). Moreover, it has already been proved that correcting for Berkson measurement error is straightforward in the linear regression context (see §1.3.1.2).

For the simulation study three measurement error models were considered:

1. a normal distribution with 0 mean and variance equal to 4: $u \sim N(0, 4)$
2. a skew-normal distribution with 0 mean, variance equal to 4 and shape parameter α equal to 5: $u \sim SN(0, 4, 5)$. For a brief presentation of what a skew normal is and how it is generated, see Appendix A
3. a mixture of two normal distributions with variance equal to 1 and mean respectively equal to -2 and $+4$: $f_U = 0.5\phi(u+2) + 0.5\phi(u-4)$.

$R = 1000$ simulations were performed for each measurement error structure with three different sample sizes: $n = 100, n = 1.000$ and $n = 10.000$. The subsequent sections compare the results obtained using the different methods for correction described in Section 1.4. For each method two summary tables were created. The first presents some major summary statistics for the estimators of β_0 and β_1. Mean, median and standard deviation of the estimates were computed using the standard formulas $\bar{\theta} = R^{-1} \sum_{r=1}^{R} \theta_r$, $Me(\theta) = (\theta_{(R/2)} + \theta_{(R/2+1)})/2$ and $sd(\theta) = \sqrt{R^{-1} \sum_{r=1}^{R} (\theta_r - \bar{\theta})^2}$ respectively. The interquartile range was obtained subtracting the first quartile $Q_1 = (\theta_{(R/4)} + \theta_{(R/4+1)})/2$ from the third quartile $Q_3 = (\theta_{(3R/4)} + \theta_{(3R/4+1)})/2$ of the empirical distribution of the estimators. Bias was calculated using the formula $b = R^{-1} \sum_{r=1}^{R} (\hat{\theta}_r - \theta)$, in which θ represents the real value of the parameter, that is 7 for β_0 and 2 for β_1, as previously stated. The mean square error (MSE) was also computed, using the formula $MSE = \sqrt{R^{-1} \sum_{r=1}^{R} (\hat{\theta}_r - \theta)^2}$.

The second table illustrates the main inferential results extracted from the simulation. For each simulation the real coverage *Real*$(1 - \alpha)$ of two-tailed nominal $(1 - \alpha) = 0.95$ confidence intervals was calculated. *Real* R(1-α) and *Real* L(1- α) were computed in the same way, but refer to one-sided confidence intervals instead. Lastly, the mean interval length of nominal $(1 - \alpha) = 0.95$ confidence intervals was calculated. The real coverage of two-tailed and one-tailed confidence intervals with nominal confidence level $(1 - \alpha)$ equal to 0.90 and 0.99 are also provided for the RC, BCES and SIMEX methods.

The experiment was set up in order to better understand how differently the estimators behave in distinctive contexts of sample size and measurement error structure. In particular, we wanted to simulate a situation which is likely to encounter in coping with astronomical datasets that present measurement error in the variables. Our principal aim is to measure the effectiveness and efficiency of the different functional methods and whether they can be influenced by the sample size.

The simulations were performed using the R programming language (R Development Core Team [39], 2005), Version 3.0.2. The code used to implement the simulations can be found in Appendix C.5.

2.2 Normal measurement error

The first measurement error model considered is the normal distribution.

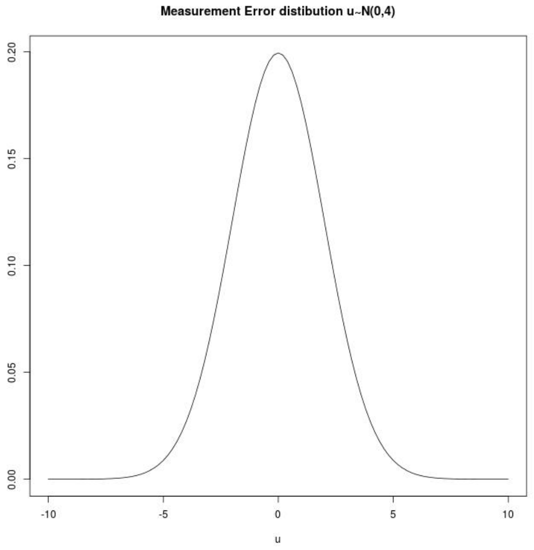


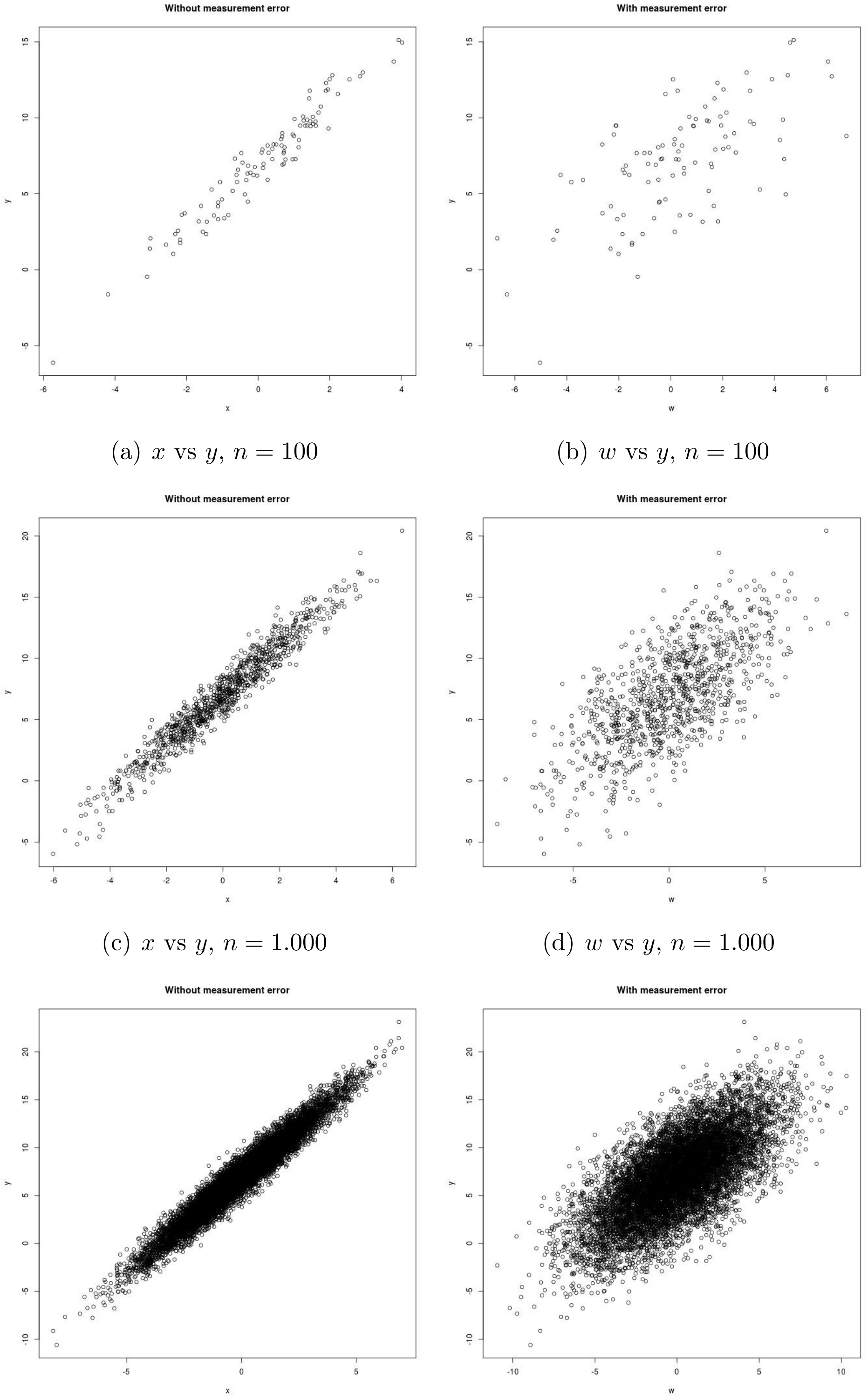
Figure 2.1: Measurement error $u \sim N(0,4)$. The graph represents the theoretical probability distribution of the normal measurement error u.

It represents the simplest type of measurement error and will be used as a benchmark for the other two, more complex, structures. We decided to set the considered variance for the measurement error distribution to a rather high value in this experiment. This is because the uncertainties in astronomical quantities are "large, skewed, or exhibit multiple modes"(Kelly [26], 2011). The simulation wants to reflect the real difficulties in working with these mismeasured quantities. In particular, in this first experiment the "large" aspect is put forward. Figure 2.1 shows the empirical density of a sample of size $n = 100$ from the measurement error distribution.

In each subsection the descriptive and inferential results obtained with the different methods for correction are presented while comparisons among the methods are made at the end of the section. Figure 2.2 graphically displays the measurement error effect in the relationship between y and the covariate, for the three sample sizes considered.

2.2.1 True model

The theoretical model in which the true x is used as a covariate is here summarized. The true model simulation was performed in order to understand how the OLS estimators would work if the true covariates were known. The-



(a) x vs y, $n = 100$

(b) w vs y, $n = 100$

(c) x vs y, $n = 1.000$

(d) w vs y, $n = 1.000$

(e) x vs y, $n = 10.000$

(f) w vs y, $n = 10.000$

Figure 2.2: Effect of normal measurement error $u \sim N(0, 4)$ in regression for three different sample sizes. The linear relationship between y and x is masked when a normal measurement error is added to the covariate x. The mismeasured points in the graphs on the right present more variability and smaller correlation.

ory states that under Gauss-Markov hypothesis, the OLS is the best — with the smallest value of MSE — linear unbiased estimator among the unbiased ones. Both the descriptive and inferential results in Table 2.1 and Table 2.2 highlight the truth of this statement. There is no significant difference between the various sample sizes analysed: all three present good results in terms of descriptive and inferential statistics pointing out how, in the absence of measurement error, a sample size of $n = 100$ is enough to yield high accuracy.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{TRUE}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	6.9927	1.9918	7.0006	2.0006	7.0016	2.0004
Median	6.9844	1.9882	6.9992	1.9988	7.0016	2.0005
Bias	0.0073	0.0082	-0.0006	-0.0006	-0.0016	-0.0004
St Dev	0.1030	0.0528	0.0310	0.0161	0.0092	0.0050
MSE	0.1033	0.0534	0.0310	0.0161	0.0093	0.0050
IQR	0.1478	0.0676	0.0441	0.0242	0.0117	0.0070

Table 2.1: Summary measures for the true model, the theoretical model obtained if the true x were observable.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{TRUE}$	β_0	β_1	β_0	β_1	β_0	β_1
Real	0.94	0.94	0.96	0.94	0.96	0.94
Real R	0.94	0.93	0.96	0.96	0.98	0.96
Real L	0.95	0.95	0.96	0.94	0.94	0.94
Average Length	0.40	0.20	0.12	0.06	0.04	0.02

Table 2.2: Inferential results for the true model with $1 - \alpha = .95$. The real coverages values are similar to the nominal ones, for both one-tailed and two-tailed confidence intervals.

2.2.2 Naive model

The naive analysis does not consider the presence of measurement error in the data: a simple regression model is fitted using the error-prone variable

w without any type of measurement error correction. As it is clearly visible from the summary Tables 2.3 and 2.4, the naive approach experiences a considerable bias of the estimator of β_1, which constantly underestimates the real value. An *attenuation-to-the-null* effect is undoubtedly present in this model. Increasing the sample size does not enhance the performance of the estimators because, as proved in Section 1.3.1.1, the naive estimator is inconsistent when the covariate is measured with error. Therefore, a correction technique is needed to improve the naive analysis.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{NAIVE}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	7.0019	1.0001	7.0060	0.9960	7.0014	1.0007
Median	7.0006	1.0008	7.0004	0.9973	7.0018	1.0002
Bias	-0.0019	0.9999	-0.0060	1.0040	-0.0014	0.9993
St. Dev	0.2959	0.1086	0.0960	0.0332	0.0292	0.0103
MSE	0.2959	1.0058	0.0962	1.0045	0.0292	0.9994
IQR	0.4091	0.1600	0.1258	0.0427	0.0387	0.0145

Table 2.3: Summary measures for the naive model in presence of normal measurement error. The analysis is performed without considering the presence of the measurement error.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{NAIVE}$	β_0	β_1	β_0	β_1	β_0	β_1
Real	0.96	0.00	0.95	0.00	0.95	0.00
Real R	0.96	0.00	0.95	0.00	0.94	0.00
Real L	0.96	1.00	0.94	1.00	0.95	1.00
Average Length	1.19	0.42	0.37	0.13	0.12	0.04

Table 2.4: Inferential results for the naive method with $1 - \alpha = 0.95$ in presence of normal measurement error.

2.2.3 Regression-Calibration

The regression calibration technique (RC) is the first attempt used to try to improve over the naive model. As already described in Section 1.4.3, in order to perform a RC analysis additional data must be provided. In this simulation experiment an internal validation dataset was used. In particular, from the original simulated dataset, a 10% of it was randomly extracted and used as a gold standard for performing the regression of x on w. To compute the standard errors of the RC estimates a jackknife approach (see §1.4.3.2) was used for this experiment. The predicted values from this regression were consequently treated as a new covariates in the original model. The analysis of the results of the simulation makes evident how Regression Calibration significantly improves the naive approach. $\hat{\beta}_1^{RC}$ is much closer to the real value 2 than the naive one. The RC estimators seem to slightly improve their accuracy in increasing the sample size, as it is underlined by the decreasing values of the bias in Table 2.5. Nonetheless, the improvement obtained by the RC approach is negligible considering the different orders of magnitude of the three sample sizes.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{RC}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	7.0863	2.2581	6.9973	2.0296	6.9982	2.0045
Median	7.0535	1.9820	6.9774	2.0216	7.0014	2.0029
Bias	-0.0863	-0.2581	0.0027	-0.0296	0.0018	-0.0045
St. Dev	1.0583	1.3321	0.2717	0.1738	0.0792	0.0572
MSE	1.0618	1.3569	0.2717	0.1763	0.0792	0.0574
IQR	1.1816	0.9301	0.3731	0.2100	0.1004	0.0820

Table 2.5: Summary measures for the RC model in presence of normal measurement error. The estimators improve their accuracy in increasing the sample size.

The RC estimates are not really acceptable taking into account the inferential results of Table 2.6. Actually, the real confidence level is far lower than the nominal 0.95 value, both for two-sided and one-sided confidence intervals. The two-sided confidence interval for β_1 does not include the true value 2 in almost half of the simulations. The same problem arises also considering lower and higher confidence levels. Nonetheless, as already stated at the beginning of this section, the measurement error considered in this simulation is large and therefore we are not expecting a well performed adjustment. Moreover, the validation data used in our regression calibration algorithm are only 10% of the total amount. Thus, we can consider the improvement made by the RC approach fairly acceptable compared to the naive estimate.

$\hat{\beta}_{RC}$	$n = 100$		$n = 1.000$		$n = 10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.47	0.41	0.45	0.54	0.49	0.41
Real R	0.73	0.64	0.69	0.76	0.68	0.68
Real L	0.65	0.67	0.68	0.67	0.69	0.65
Average Length	1.06	0.82	0.31	0.23	0.10	0.07
$(1-\alpha)$=0.95						
Real	0.54	0.46	0.50	0.62	0.54	0.53
Real R	0.78	0.68	0.73	0.81	0.73	0.71
Real L	0.69	0.73	0.71	0.73	0.76	0.69
Average Length	1.27	0.98	0.37	0.27	0.12	0.08
$(1-\alpha)$=0.99						
Real	0.59	0.58	0.64	0.72	0.70	0.66
Real R	0.81	0.74	0.80	0.86	0.82	0.82
Real L	0.74	0.80	0.79	0.83	0.81	0.80
Average Length	1.68	1.30	0.49	0.35	0.15	0.11

Table 2.6: Inferential results for the RC method in presence of normal measurement error with three different coverage levels. The real coverage levels are smaller than the nominal ones.

2.2.4 BCES

The Akritas & Bershady version of the bivariate correlated errors and intrinsic scatter (BCES) method (see §1.4.2) was originally developed for dealing with a linear regression that presents measurement error in both the response and the independent variable. Here a simplified version is used, since the response y is supposed to be an error-free variable. The correction obtained with the BCES method is the best among all the functional methods consid-

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{BCES}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	7.0107	2.0584	7.0044	1.9970	6.9983	2.0011
Median	7.0262	2.0283	6.9951	1.9922	6.9948	1.9990
Bias	-0.0107	-0.0584	-0.0044	0.0030	0.0017	-0.0011
St. Dev	0.4229	0.2434	0.1357	0.0658	0.0416	0.0219
MSE	0.4230	0.2503	0.1358	0.0659	0.0416	0.0219
IQR	0.5216	0.2929	0.1802	0.0808	0.0504	0.0264

Table 2.7: Summary measures for the BCES model in presence of normal measurement error. $\hat{\beta}_0^{BCES}$ and $\hat{\beta}_1^{BCES}$ are on average almost equal to the real intercept and slope values chosen for the simulation.

$\hat{\beta}_{BCES}$	$n = 100$		$n = 1.000$		$n = 10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.91	0.99	0.89	0.99	0.92	0.99
Real R	0.90	0.98	0.90	0.97	0.93	0.98
Real L	0.91	1.00	0.91	0.99	0.91	0.97
Average Length	1.43	1.25	0.43	0.37	0.14	0.12
$(1-\alpha)$=0.95						
Real	0.96	0.99	0.94	1.00	0.95	0.99
Real R	0.95	0.99	0.96	0.99	0.95	0.99
Real L	0.95	1.00	0.93	1.00	0.96	0.99
Average Length	1.71	1.49	0.51	0.44	0.16	0.14
$(1-\alpha)$=0.99						
Real	1.00	1.00	0.99	1.00	0.98	1.00
Real R	0.99	1.00	0.99	1.00	0.98	1.00
Real L	0.99	1.00	0.98	1.00	0.99	1.00
Average Length	2.26	1.98	0.68	0.58	0.21	0.18

Table 2.8: Inferential results for the BCES method in presence of normal measurement error with three different coverage levels.

ered in the normal measurement error model. The summary Table 2.7 points out how this method succeeds in nullifying the attenuation-to-the-null effect due to the presence of the error component u. Even with a sample size of 100 the BCES method performs extremely well: the $\hat{\beta}_1^{BCES}$ estimate is almost equal to the true value of $\beta_1 = 2$. Considering the inferential results in Table 2.8 it is worth noting that the real coverage level is even higher than the nominal one, for both one-sided and two-sided confidence intervals. This is mainly due to the precision of the point estimates and to the high values of $\hat{Var}(\hat{\beta}_1^{BCES})$ and $\hat{Var}(\hat{\beta}_0^{BCES})$, see Section 2.2.6 for further details.

2.2.5 SIMEX

The simulation extrapolation approach is the most computationally intensive method for correction amongst the ones analysed so far. As already presented in Section 1.4.4, the SIMEX method increasingly adds artificial measurement error of the same structure presumed for the real one which affects the data. Therefore, in this experiment the computer-generated pseudo errors $\{u_b\}_{b=1}^{B}$ have normal distribution with 0 mean and variance equal to 4, like the measurement error u. In empirical applications, choosing the correct distribution for the computer-generated errors is a delicate part in the simex algorithm. Generally, different distributions are used and then the one which is considered the best by the analysts is selected (Carroll et al [9], 2012). In Table 2.9 it is possible to notice that the SIMEX approach improves over the naive estimator, though it does not succeed in entirely nullifying the attenuation-to-the-null effect since the bias of $\hat{\beta}_1^{SIMEX}$ is still equal to 0.5 for all three sample sizes considered. The inferential results in Table 2.10 present an even worse scenario: almost 90% of the confidence intervals for β_1 do not contain the real value 2 when $n = 100$, and the percentage drops to 0 when we consider bigger sample sizes. This is because the variability of the estimators decrease in augmenting the sample size, but the point estimate $\hat{\beta}_1^{SIMEX}$ does not get closer to the real value 2. With this simulation we prove that the SIMEX technique does not improve its effectiveness in increasing the size of the sample.

	$n=100$		$n=1.000$		$n=10.000$	
$\hat{\beta}_{SIMEX}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	7.0061	1.4978	7.0052	1.4850	6.9999	1.4944
Median	7.0214	1.4993	7.0009	1.4899	6.9981	1.4946
Bias	-0.0061	0.5022	-0.0052	0.5150	0.0001	0.5056
St. Dev	0.3328	0.1825	0.1108	0.0539	0.0329	0.0163
MSE	0.3329	0.5343	0.1109	0.5178	0.0329	0.5059
IQR	0.4168	0.2368	0.1491	0.0767	0.0427	0.0233

Table 2.9: Summary measures for the SIMEX model in presence of normal measurement error. The value of $\hat{\beta}_1^{SIMEX}$ is closer to the real $\beta_1 = 2$ than $\hat{\beta}_1^{NAIVE}$, although the bias is still considerable.

$\hat{\beta}_{SIMEX}$	$n=100$		$n=1.000$		$n=10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.84	0.10	0.83	0.00	0.87	0.00
Real R	0.89	0.04	0.86	0.00	0.91	0.00
Real L	0.84	1.00	0.85	1.00	0.88	1.00
Average Length	0.95	0.47	0.30	0.15	0.10	0.05
$(1-\alpha)$=0.95						
Real	0.90	0.12	0.91	0.00	0.91	0.00
Real R	0.93	0.10	0.92	0.00	0.94	0.00
Real L	0.92	1.00	0.92	1.00	0.93	1.00
Average Length	1.14	0.56	0.36	0.17	0.11	0.06
$(1-\alpha)$=0.99						
Real	0.98	0.26	0.96	0.00	0.96	0.00
Real R	0.96	0.19	0.98	0.00	0.97	0.00
Real L	0.99	1.00	0.96	1.00	0.98	1.00
Average Length	1.51	0.74	0.47	0.23	0.15	0.07

Table 2.10: Inferential results for the SIMEX method in presence of normal measurement error with three different coverage levels. Weak real coverage level for β_1.

2.2.6 Methods comparison

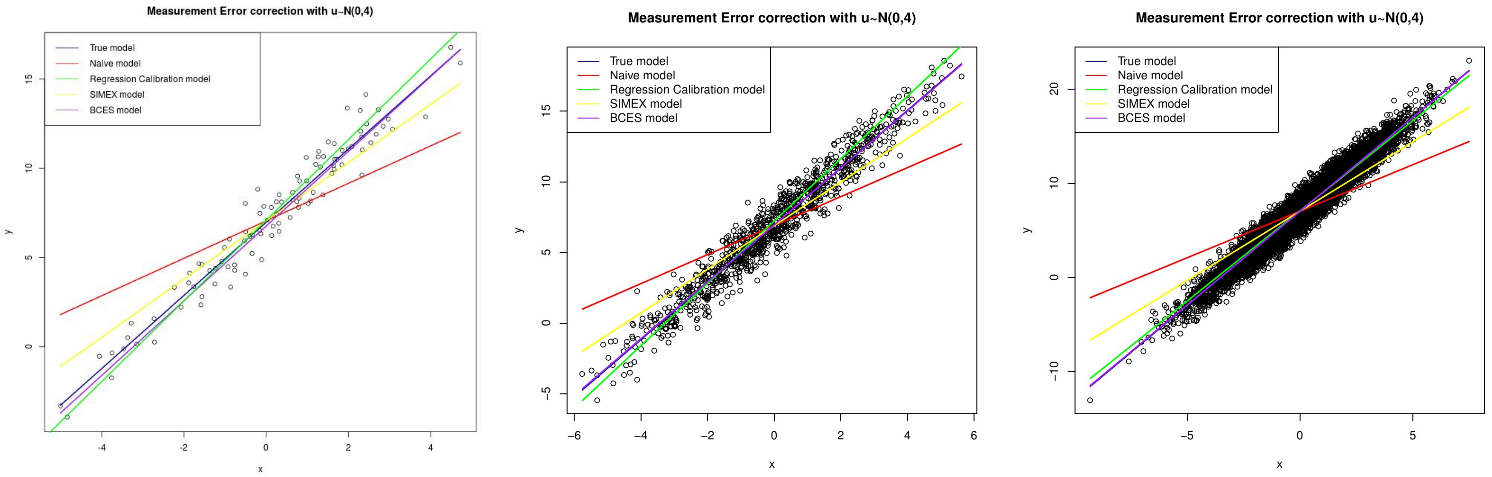
The experiment analysed so far could be seen as a “textbook case”, being a simple linear regression model with classical measurement error structure.

Nevertheless the results obtained are helpful and constructive for understanding the behaviour of the estimates even in more complex situations. As previously stated, the estimator which performs best is the BCES estimator by Akritas & Bershady, which works well when the measurement error structure is simple and symmetric. Moreover, belonging to the family of method-of-moments estimators, the BCES approach can be used only when the regression is linear; it is infeasible when the functional form is non-linear or cannot be linearised. The regression calibration method is simple, computationally not demanding and effective for almost every type of measurement error and functional regression form. The drawback is that additional valida-

	$\hat{\beta}_0$	$\hat{sd}(\hat{\beta}_0)$	$\hat{\beta}_1$	$\hat{sd}(\hat{\beta}_1)$
$n = 100$				
TRUE	6.9927	0.0997	1.9918	0.0509
NAIVE	7.0006	0.3001	1.0008	0.1062
RC	7.0863	0.3134	2.2581	0.2126
BCES	7.0107	0.4182	2.0584	0.3393
SIMEX	7.0061	0.2879	1.4978	0.1401
$n = 1.000$				
TRUE	7.0006	0.0317	2.0006	0.0158
NAIVE	7.0060	0.0950	0.9960	0.0334
RC	6.9973	0.0954	2.0296	0.0676
BCES	7.0044	0.1304	1.9970	0.1096
SIMEX	7.0052	0.0911	1.4850	0.0442
$n = 10.000$				
TRUE	7.0016	0.0100	2.0004	0.0050
NAIVE	7.0014	0.0301	1.0007	0.0106
RC	6.9982	0.0301	2.0045	0.0212
BCES	6.9983	0.0413	2.0011	0.0352
SIMEX	6.9999	0.0289	1.4944	0.0141

Table 2.11: Average values of the intercept, the slope and their standard errors for the normal measurement error model with three different sample sizes. $\hat{\beta}_0$, $\hat{sd}(\hat{\beta}_0)$, $\hat{\beta}_1$ and $\hat{sd}(\hat{\beta}_1)$ are calculated for each method for correction. The BCES method performs the best correction on average.

tion data must be available. Unlikely in Medicine and Genomics, obtaining validation data in Astronomy is never easy and most of the times impossible, being the astronomical quantities often derived from transformations of non-directly observed variables in which measurement error is already present (Kelly [26], 2011). The SIMEX approach is the most general and widely applicable functional method. It does not require additional data but it is computationally intensive. Moreover, the results obtained, even in the simplest case, do not entirely correct the attenuation-to-the-null effect caused by the u component. Of major interest is to compare how the estimators behave for different sample sizes. Table 2.11 reports the average values of the estimates and their standard deviations for β_0 and β_1, obtained with the true model, the naive model and the functional measurement error methods for correction, for the three considered sample sizes. As it can be seen from Table 2.11, there is basically no difference amongst the three experiments in terms of point estimate, whilst the standard errors obviously decrease in increasing the sample size. This proofs that when data are affected by classical measurement error the sample size does not affect the point estimate of the parameters. In Figure 2.3 the functional methods for correction are plotted together with the true and the naive models for the three different sample sizes. The BCES and the RC work well in all cases, whilst the SIMEX method is still affected by a slight attenuation-to-the-null effect. The latter fact reveals how the SIMEX method is not minimally affected by the sample size considered. As it is clearly visible from the graphs, when sample size increases more information is available for the analysis, nevertheless increasing the sample size means adding biased information due to measurement error: the addition of mismeasured observations does not compensate the lack of true measurements.



(a) x vs y + fitted models, $n = 100$

(b) x vs y + fitted models, $n = 1.000$

(c) x vs y + fitted models, $n = 10.000$

Figure 2.3: Measurement error models fitted to the real data x vs y. The graphs present the average behaviour of the correction techniques in presence of normal measurement error $u \sim N(0, 4)$, for the three sample sizes considered in the simulation.

2.3 Skew-Normal measurement error

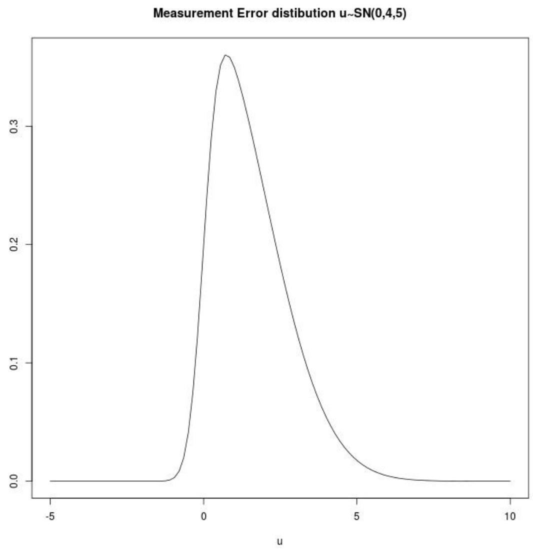


Figure 2.4: Measurement error $u \sim SN(0,4,5)$. The graph represents the theoretical probability distribution of the skew-normal measurement error u.

A skew-normal measurement error model was adopted for the second simulation. The peculiarity of the aforementioned model is the fact that the error added to the true covariate X is asymmetric, creating an unpredictable behaviour in the observed variable X^*. Namely, if the true variable is not affected by asymmetry, the measurement error will create either a positive or a negative skewness, depending on the nature of the skewness present in u. If the true variable X already presents skewness, the measurement error can either intensify it or hide it. For a unimodal distribution, negative skewness indicates that the tail on the left side of the probability density function is longer than the right side, conversely positive skewness indicates that the tail on the right side is longer than the left side. In order to simulate a measurement error that presents skewness, random values were generated from a skew-normal distribution, using the "sn" package developed for the R programming language (Azzalini [3], 2014). In particular, the u vector was generated from a skew-normal distribution with 0 mean, variance equal to 4 and shape parameter α equal to 5 (see Appendix A). As it is possible to see in Figure 2.4, the distribution of the measurement error u presents positive skewness. Since the true variable x was generated by a normal distribution, which is symmetric, the mismeasured variable x^* is skewed to the right. The effect of a skew measurement error in the covariate is shown in Figure 2.5. As it is graphically clearly visible, the positive skewness of u leads to increased values of the observed variable x^* with respect to the true variable x.

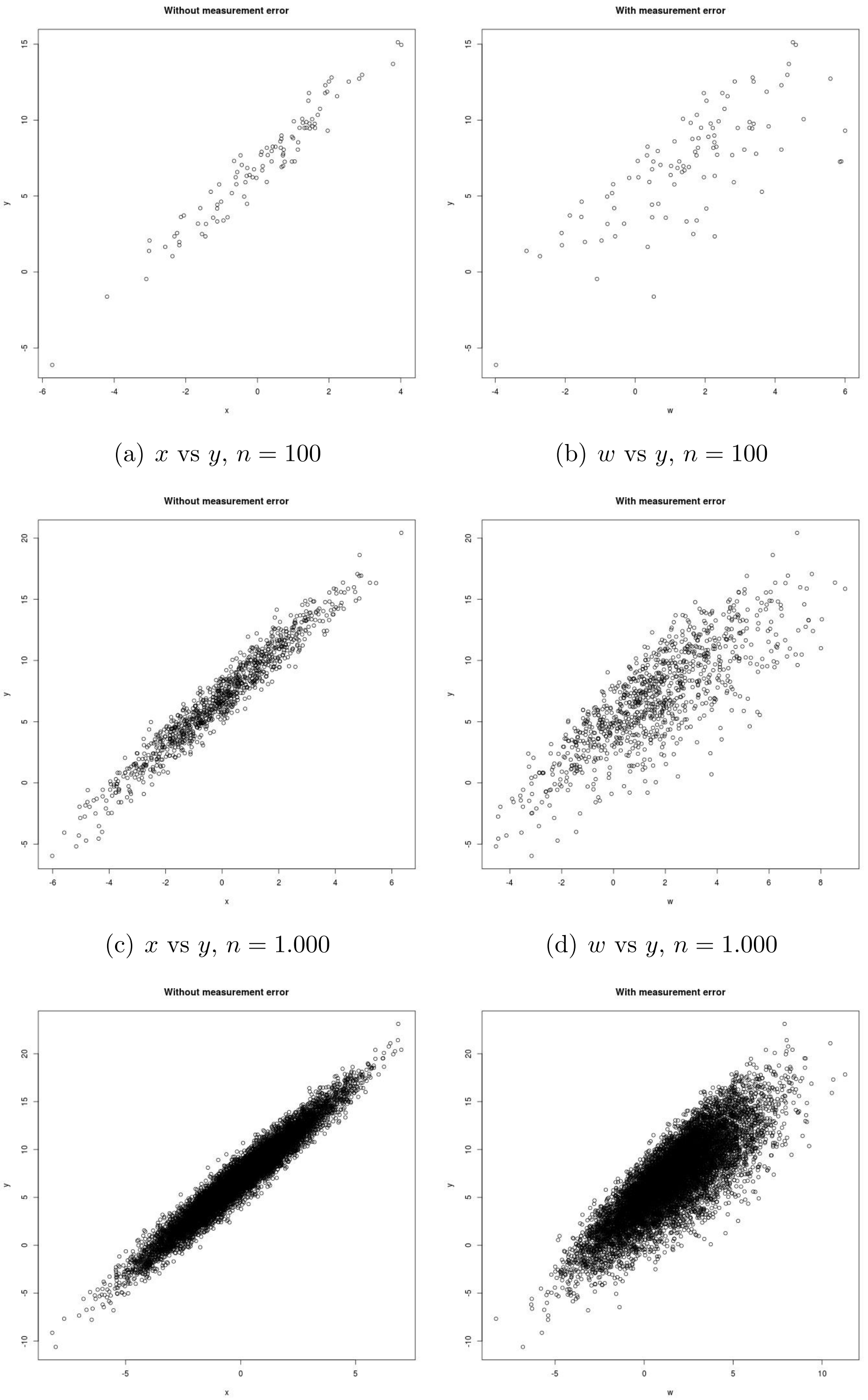
The present section is organized as the previous one: first each method for corrections is presented and analysed, and then, at the end, comparisons are made.

2.3.1 Naive model

The naive analysis does not count for the presence of measurement error and it estimates the parameters β_0 and β_1 as if x^* were the true variable. As it can be expected, the presence of measurement error completely biases inference on the parameters. Nevertheless, the behaviour in case of skew-normal measurement error is different compared to the one in Section 2.2.2. As it is highlighted in Table 2.12, in this case both the estimates of the intercept, $\hat{\beta}_0$, and of the slope, $\hat{\beta}_1$, exhibit significant bias. The *attenuation-to-the-null* effect seems to be slighter, however it is still significant. The inferential results in Table 2.13 reflect what already stated for the descriptive results. Likewise the case of normal measurement error, increasing the sample size in the skew-normal measurement error model does not help to enhance the quality of the naive approach.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{NAIVE}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	4.7333	1.4552	4.7454	1.4378	4.7463	1.4416
Median	4.7394	1.4639	4.7411	1.4347	4.7477	1.4414
Bias	2.2667	0.5448	2.2546	0.5622	2.2537	0.5584
St. Dev	0.2483	0.1012	0.0756	0.0335	0.0228	0.0111
MSE	2.2802	0.5541	2.2558	0.5632	2.2538	0.5585
IQR	0.3587	0.1322	0.1082	0.0409	0.0303	0.0148

Table 2.12: Summary measures for the naive model in presence of skew-normal measurement error. Neither the intercept nor the slope are estimated correctly.


Without measurement error

With measurement error

(a) x vs y, $n = 100$

(b) w vs y, $n = 100$

(c) x vs y, $n = 1.000$

(d) w vs y, $n = 1.000$

(e) x vs y, $n = 10.000$

(f) w vs y, $n = 10.000$

Figure 2.5: Effect of skew-normal measurement error $u \sim SN(0, 4, 5)$ in regression for three different sample sizes. The mismeasured value w presents an higher value than the true variable x, due to the presence of the skewed-to-the-right measurement error. The linear relationship between y and w is spread to the right hand-side of the graphs, hiding the real pattern of the model.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{NAIVE}$	β_0	β_1	β_0	β_1	β_0	β_1
Real	0.00	0.00	0.00	0.00	0.00	0.00
Real R	0.00	0.00	0.00	0.00	0.00	0.00
Real L	1.00	1.00	1.00	1.00	1.00	1.00
Average Length	1.11	0.39	0.35	0.12	0.11	0.04

Table 2.13: Inferential results for the naive method with $1 - \alpha = 0.95$ in presence of skew-normal measurement error. The real coverage levels are equal to 0, meaning that in none of the simulation the true values $\beta_0 = 7$ and $\beta_1 = 2$ are contained in the confidence intervals.

2.3.2 Regression-Calibration

In case of skew-normal measurement error, regression calibration is the technique which performs the best amongst the functional methods considered. Contrarily to the naive approach, the RC method provides estimates which are close to their real values. As it is shown in Table 2.14, it seems that the sample size does not influence the inference on parameters also in this case. Even with a sample of size $n = 100$ the regression calibration method leads to an almost perfect estimation, with the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ nearly equal to the real $\beta_0 = 7$ and $\beta_1 = 2$. The reason why the RC method performs so well is probably due to both the RC technique itself and the nature of the measurement error considered: the availability of 10% of the gold standard x is sufficient to perceive and thus to correct the asymmetry present in the measurement error model. The inferential results in Table 2.15 shows that the real coverage level is lower than the nominal one, for the three nominal coverage levels considered. Notwithstanding, the RC estimators can be considered an effective solution to cope with skew-normal measurement error; it is nevertheless worth highlighting again that some validation data must be provided in order to perform the aforementioned algorithm.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{RC}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	7.0071	2.2250	6.9762	1.9931	7.0014	1.9996
Median	6.9746	1.9785	6.9565	1.9928	6.9993	1.9987
Bias	-0.0071	-0.2250	0.0238	0.0069	-0.0014	0.0004
St. Dev	0.8495	1.6372	0.1898	0.1361	0.0628	0.0379
MSE	0.8495	1.6526	0.1913	0.1363	0.0629	0.0379
IQR	0.9625	0.5773	0.2483	0.1891	0.0980	0.0492

Table 2.14: Summary measures for the RC model in presence of skew-normal measurement error. $\hat{\beta}_0^{RC}$ and $\hat{\beta}_1^{RC}$ are on average really close to the real intercept and slope values chosen for the simulation.

$\hat{\beta}_{RC}$	$n = 100$		$n = 1.000$		$n = 10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.40	0.42	0.49	0.42	0.41	0.49
Real R	0.63	0.64	0.64	0.64	0.67	0.69
Real L	0.70	0.66	0.77	0.69	0.67	0.70
Average Length	0.81	0.54	0.24	0.15	0.08	0.05
$(1-\alpha)$=0.95						
Real	0.47	0.47	0.56	0.48	0.47	0.57
Real R	0.67	0.71	0.69	0.69	0.71	0.73
Real L	0.73	0.70	0.80	0.74	0.70	0.76
Average Length	0.97	0.65	0.29	0.18	0.09	0.06
$(1-\alpha)$=0.99						
Real	0.61	0.62	0.70	0.63	0.65	0.69
Real R	0.74	0.80	0.78	0.74	0.81	0.80
Real L	0.80	0.77	0.86	0.82	0.77	0.83
Average Length	1.28	0.86	0.38	0.24	0.12	0.07

Table 2.15: Inferential results for the RC method in presence of skew-normal measurement error with three different coverage levels. Although the real coverage levels are smaller than the nominal ones the real coverage levels provided by the RC estimator are the best amongst the functional methods considered for the skew-normal measurement error simulation.

2.3.3 BCES

With a skew-normal measurement error model the BCES method of moments does not perform as well as it does in Section 2.2.4. Table 2.16 shows that the BCES approach fails to correctly estimate the intercept of the linear model. Even though $\hat{\beta}_1^{BCES}$ still provides an effective estimate for β_1, the same cannot be said about the BCES estimator of β_0. Notably $\hat{\beta}_0^{BCES}$ underestimates the true value of the intercept and, once again, increasing the sample size does not bring any significant improvement to the performance of the estimator. The drawback of the BCES approach is that it cannot account for the asymmetric nature of the measurement error. In computing $\hat{\beta}_0^{BCES} = \bar{y} - \hat{\beta}_1^{BCES}\bar{x}^*$, the sample average of the observed variable $\bar{x}^*$ is higher than the true sample average $\bar{x}$ due to the positive skewness of u, which implies that $\hat{\beta}_0^{BCES}$ always miscalculates the value of β_0. Table 2.17 strengthens what we have already said commenting the descriptive results: the slope estimator behaves well in terms of real coverage, while the intercept estimator does not.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{BCES}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	3.8121	2.0449	3.8649	2.0019	3.8731	2.0002
Median	3.8555	2.0157	3.8532	2.0006	3.8701	1.9998
Bias	3.1879	-0.0449	3.1351	-0.0019	3.1269	-0.0002
St. Dev	0.3687	0.1536	0.1087	0.0422	0.0323	0.0130
MSE	3.2092	0.1600	3.1370	0.0423	3.1271	0.0130
IQR	0.5339	0.1760	0.1624	0.0595	0.0415	0.0173

Table 2.16: Summary measures for the BCES model in presence of skew-normal measurement error. Contrarily to what happens with a normal measurement error, with an asymmetric distribution $\hat{\beta}_0^{BCES}$ highly underestimates the true intercept $\beta_0 = 7$.

$\hat{\beta}_{BCES}$	$n = 100$		$n = 1.000$		$n = 10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.00	0.96	0.00	0.97	0.00	0.96
Real R	0.00	0.98	0.00	0.97	0.00	0.96
Real L	1.00	0.92	1.00	0.95	1.00	0.97
Average Length	1.11	0.63	0.34	0.20	0.11	0.06
$(1-\alpha)$=0.95						
Real	0.00	0.98	0.00	0.99	0.00	0.99
Real R	0.00	0.99	0.00	0.99	0.00	0.97
Real L	1.00	0.97	1.00	0.98	1.00	0.99
Average Length	1.33	0.75	0.41	0.23	0.13	0.07
$(1-\alpha)$=0.99						
Real	0.00	1.00	0.00	1.00	0.00	1.00
Real R	0.00	1.00	0.00	0.99	0.00	1.00
Real L	1.00	1.00	1.00	1.00	1.00	1.00
Average Length	1.76	1.00	0.54	0.31	0.17	0.10

Table 2.17: Inferential results for the BCES method in presence of normal measurement error with three different coverage levels. The true value of the intercept β_0 is never contained in the confidence intervals, no matter the coverage level considered.

2.3.4 SIMEX

As already reported in Section 2.2.5, the key factor for an effective application of the SIMEX method is to correctly choose the distribution of the computer-generated pseudo errors $\{u_b\}_{b=1}^{B}$. In an empirical framework many distributions for the artificial errors are taken into account and then the most realistic and effective is chosen. In a simulated framework one possible approach is to use the same distribution from which the measurement error was generated, that is a skew-normal distribution in our case. Nevertheless the attempt of using a skew-normal distribution for generating the pseudo errors $\{u_b\}_{b=1}^{B}$ led to an incongruence in the estimation of the parameters: the obtained variance estimator of $\hat{\beta}_0^{SIMEX}$ was a negative number! As already

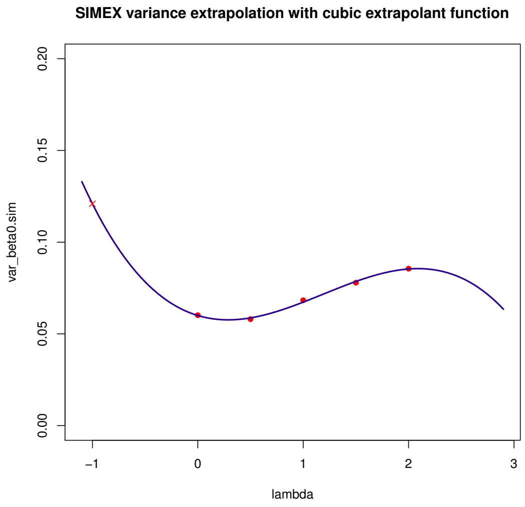


Figure 2.6: $\hat{Var}(\hat{\beta}_0)$ extrapolation step with cubic extrapolant function. A cubic component is needed to correctly fit the artificial variances generated in the simulation step.

pointed out in Section 1.4.4.3, this is not caused by an error in the simex algorithm, but the procedure simply does not assure that the number obtained will be non-negative. In order to avoid the aforementioned brawback a cubic extrapolant function was utilized for the variance component extrapolation. As it is clearly visible in Figure 2.6, a cubic function satisfactorily

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{SIMEX}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	7.3110	1.8799	7.2778	1.8629	7.2871	1.8664
Median	7.3230	1.8854	7.2726	1.8622	7.2839	1.8656
Bias	-0.3110	0.1201	-0.2778	0.1371	-0.2871	0.1336
St. Dev	0.2879	0.1488	0.0921	0.0483	0.0268	0.0158
MSE	0.4238	0.1912	0.2926	0.1454	0.2884	0.1346
IQR	0.4285	0.1788	0.1260	0.0672	0.0395	0.0186

Table 2.18: Summary measures for the SIMEX model in presence of skew-normal measurement error. The SIMEX method on average performs a good correction for both the intercept and the slope of the model.

interpolates the variances obtained in the simulation step and furthermore it avoids the nonpositivity problem which arose using a quadratic extrapolant function.

$\hat{\beta}_{SIMEX}$	$n = 100$		$n = 1.000$		$n = 10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.49	0.64	0.04	0.07	0.00	0.00
Real R	0.99	0.55	1.00	0.04	1.00	0.00
Real L	0.42	0.95	0.01	1.00	0.00	1.00
Average Length	0.62	0.36	0.20	0.11	0.06	0.04
$(1-\alpha)$=0.95						
Real	0.57	0.72	0.04	0.07	0.00	0.00
Real R	0.99	0.66	1.00	0.07	1.00	0.00
Real L	0.50	0.98	0.04	1.00	0.00	1.00
Average Length	0.74	0.42	0.24	0.13	0.08	0.04
$(1-\alpha)$=0.99						
Real	0.66	0.82	0.11	0.13	0.00	0.00
Real R	1.00	0.80	1.00	0.10	1.00	0.00
Real L	0.64	0.99	0.07	1.00	0.00	1.00
Average Length	0.98	0.56	0.32	0.18	0.10	0.06

Table 2.19: Inferential results for the SIMEX method in presence of skew-normal measurement error with three different coverage levels. The real coverage levels are smaller than the nominal ones.

The estimates obtained using the SIMEX method for coping with a skew-normal measurement error model are quite satisfactory. Table 2.18 highlights how the SIMEX approach provides estimations that are sufficiently close to the real values of β_0 and β_1. Even though $\hat{\beta}_0^{SIMEX}$ slightly overestimates the real intercept and $\hat{\beta}_1^{SIMEX}$ slightly underestimates the real slope, all in all the SIMEX approach works better with a skew-normal measurement error distribution than with a gaussian one, as seen in Section 2.2.5. The real coverage levels in Table 2.19 do not reflect the nominal ones. Apparently even if the point estimates are quite satisfactory the SIMEX procedure for estimating the standard errors of $\hat{\beta}_0^{SIMEX}$ and $\hat{\beta}_1^{SIMEX}$ underestimates their variability, which leads to short confidence intervals and, as a consequence, to real coverage values that are smaller than the nominal ones. A new result achieved with the present simulation is the discovery of the underestimation of the SIMEX estimators variability when the measurement error model is asymmetric.

2.3.5 Methods comparison

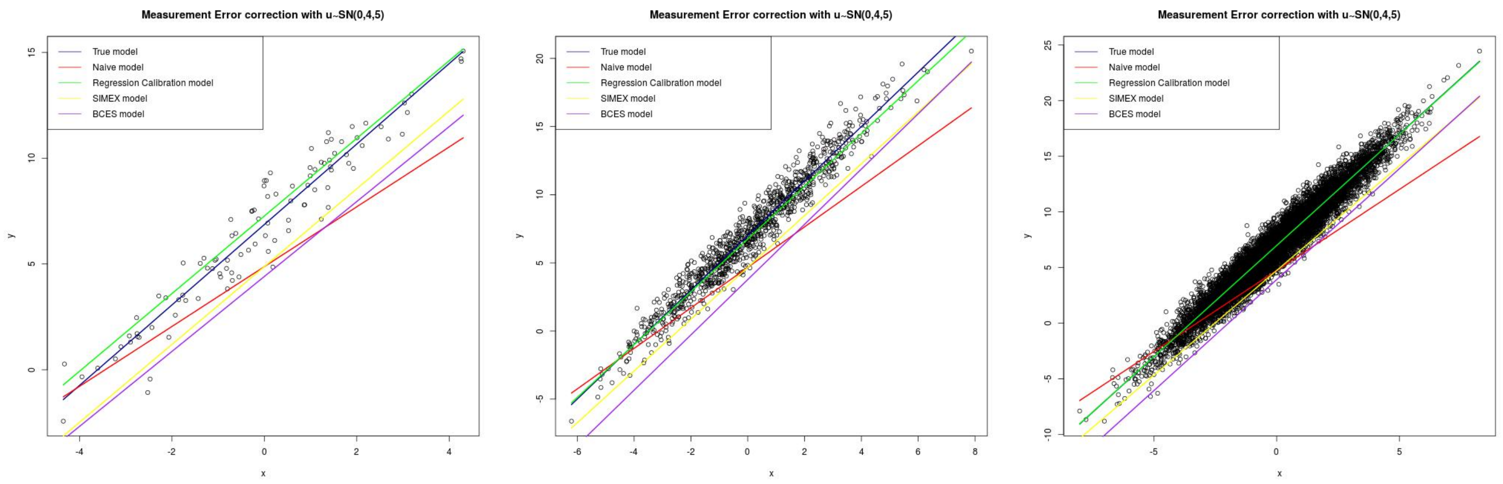
The results obtained for the simulated experiment of a skew-normal measurement error in linear regression are fairly interesting and in some ways unexpected. Contrarily to the outcomes reported in Section 2.2, the asymmetric nature of the error u involves a misestimation in both the intercept and the slope of the regression model. As a result, the naive model in Section 2.3.1 presents an attenuation for both the parameters. Moreover, the BCES method, which performs an optimal measurement error correction in the gaussian case, does not succeed in correctly estimating the intercept of the model when the measurement error is asymmetric. On the other hand, both the regression calibration and the SIMEX approach achieve the target of satisfactorily correcting for the skew-normal measurement error, in terms of point estimate. However, when we consider the inference provided by the aforementioned methods, none of them presents confidence intervals which reflect the nominal coverage level expected.

It is worth highlighting that both methods present limitations that must be taken into account in performing empirical measurement error correction. As already pointed out many times, RC approach needs further information in order to be feasible, although an internal validation data of only a 10% of the total amount was already sufficient to recognize and thus to account for the asymmetric measurement error behaviour. On the other hand, the SIMEX approach requires to previously know the measurement error variance and distribution in order to perform an effective correction. During the simulation process many distributions have been utilized for generating the artificial errors U, and some of them provided completely biased results. This is to emphasize once again how the performance of the SIMEX method is deeply affected by the chosen distribution for the generation of the pseudo-errors U. In simulated experiments providing and recognizing the most suitable distribution is fairly simple, it is not in coping with real data sets in which the measurement error nature is not known. As a consequence, many SIMEX algorithm applications could be required in order to find the most appropriate solution. Table 2.20 reports the estimators sample average and the estimators standard deviation for each functional method used in the simulation, with the three different sample sizes considered. As it has already been seen in Section 2.2.6 for the normal measurement error distribution, increasing the sample size does not produce significant improve-

	$\hat{\beta}_0$	$\hat{sd}(\hat{\beta}_0)$	$\hat{\beta}_1$	$\hat{sd}(\hat{\beta}_1)$
$n = 100$				
TRUE	7.0015	0.0997	2.0002	0.0498
NAIVE	4.7333	0.2793	1.4552	0.0983
RC	7.0071	0.2376	2.2250	0.1420
BCES	3.8121	0.3236	2.0449	0.1755
SIMEX	7.3110	0.1883	1.8799	0.1055
$n = 1.000$				
TRUE	7.0001	0.0315	1.9999	0.0158
NAIVE	4.7454	0.0887	1.4378	0.0315
RC	6.9762	0.0741	1.9931	0.0459
BCES	3.8649	0.1035	2.0019	0.0594
SIMEX	7.2778	0.0620	1.8629	0.0340
$n = 10.000$				
TRUE	7.0001	0.0100	1.9999	0.0050
NAIVE	4.7463	0.0281	1.4416	0.0099
RC	7.0014	0.0234	1.9996	0.0146
BCES	3.8731	0.0327	2.0002	0.0187
SIMEX	7.2871	0.0192	1.8664	0.0107

Table 2.20: Average values of the intercept, the slope and their standard errors for the skew-normal measurement error model with three different sample sizes. The RC method performs the best correction on average.

ment in coping with skew-normal measurement error either. The standard errors of the estimators naturally decrease when the observations number raises, nonetheless the point estimates remain basically the same; meaning that an increase in biased information acquisition does not directly produce a better inference on the parameters. In Figure 2.7 the different behaviour of the estimators is graphically presented, for each sample size. The graphs clearly highlight how the asymmetric nature of the measurement error attenuates the inference on the parameters. The RC model has an optimal fit to the real data, whilst both the simex and the BCES regression lines do not perfectly pass through the points mass. This behaviour is due to the positive skewness of the measurement error distribution, which causes the observed value x^* to be always larger than the correspondent true value x.



(a) x vs y + fitted models, $n = 100$

(b) x vs y + fitted models, $n = 1000$

(c) x vs y + fitted models, $n = 10000$

Figure 2.7: Measurement error models fitted to the real data x vs y. The graphs present the average behaviour of the correction techniques in presence of skew-normal measurement error $u \sim SN(0, 4, 5)$, for the three sample sizes considered in the simulation.

2.4 Normal mixture measurement error

The last simulation performed considers a normal mixture distribution for the measurement error model. A mixture distribution is the probability distribution of a random variable that is derived from a collection of other random variables. In our simulation two normal distributions, one with mean equal to -2 and the other with mean equal to 4, are added with weights equal to 0.5. Mathematically, this leads to the following expression:

$$f_U = 0.5\phi(u+2) + 0.5\phi(u-4) \tag{2.2}$$

in which the density function of the measurement error u is given by the stochastic mixture of the two previously defined normal distributions. An expression like the one in (2.2) is called *mixture density*. The two normals that are combined to form the mixture density are called the *mixture components*, and the probabilities (or weights) associated with each component are called the *mixture weights*. The case considered is an *equal-weighted mixture density*, since the weights are both equal to 0.5. As it is clearly visible in Figure 2.8, the obtained distribution is *bimodal*. As a consequence, the bias introduced by the presence of the measurement error can stochastically lead to either a highly-positive mismeasured value or a highly-negative mismeasured one. The Graphs in Figure 2.9 clarify the con-

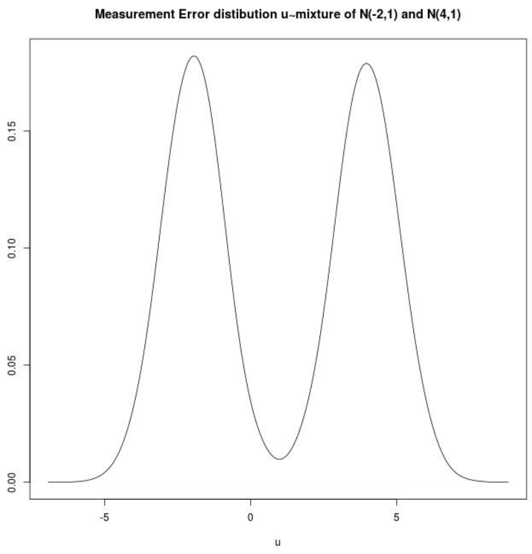


Figure 2.8: Measurement error $f_U = 0.5\phi(u+2) + 0.5\phi(u-4)$. The theoretical probability distribution chosen for this simulation is a mixture of normals with two different modes.

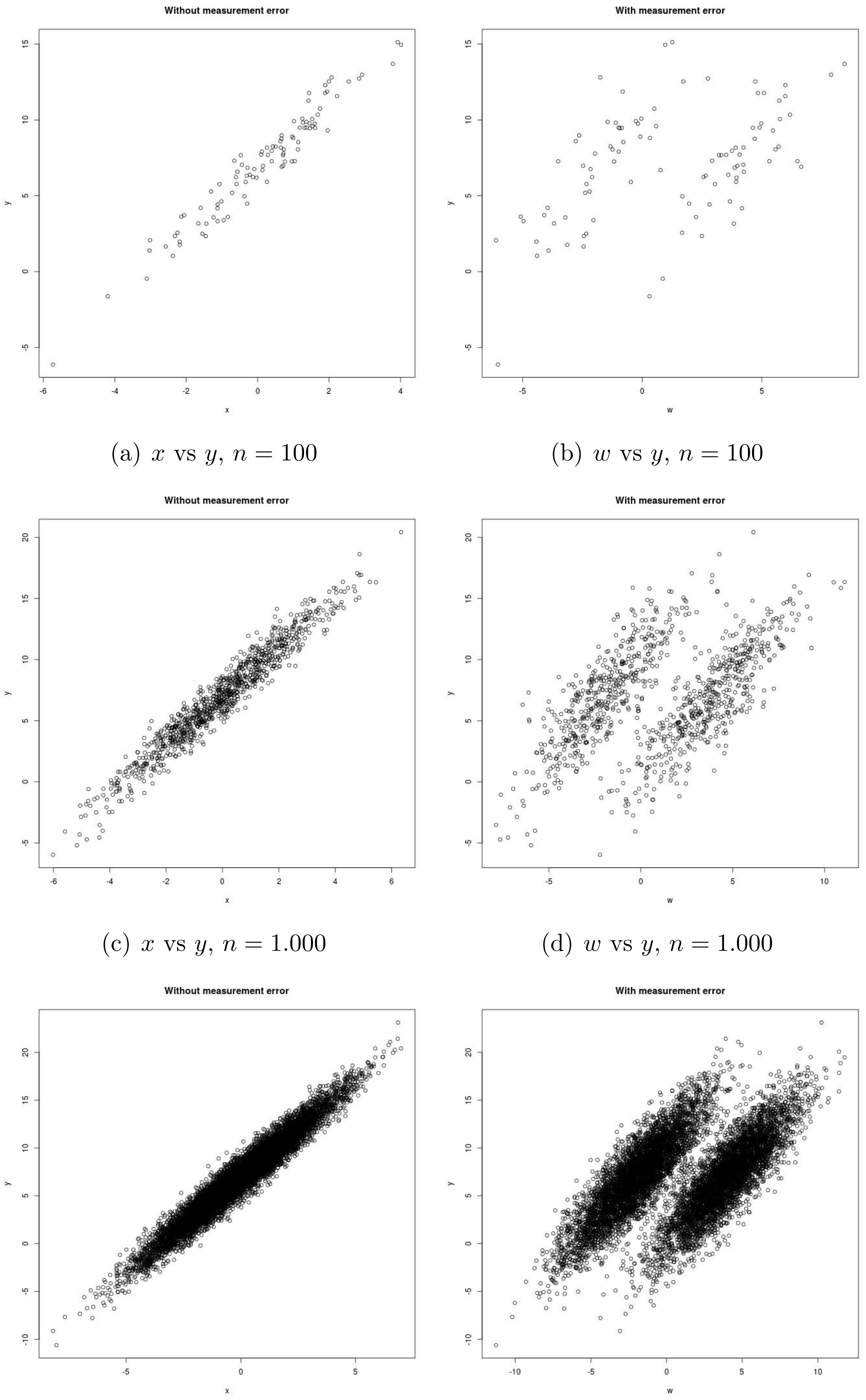
cept: the values observed for x^* follow two different linear patterns, neither of them is the original, error-free one.

The aim of the present simulation is to understand if, in presence of such a complex measurement error structure, the methods considered can perform an attempt of correction and statistically improve the inference on parameters. Even though a mixture density could seem an atypical measurement error structure, astronomical literature offers many examples in which mixture of normals are used to model measurement error densities (Kelly [25], 2007). The focus of the present work is on functional methods for correction, that are those methods in which no initial distributional hypotheses for the components are made (see §1.2.1). Therefore, the aim is to test the robustness of these approaches in coping with different measurement error structures, without taking into account the probability density of the latter one.

Likewise the previous sections, firstly the results for every model are presented and commented, and then comparisons are made.

2.4.1 Naive model

Neglecting the presence of the normal mixture measurement error leads to a naive model whose behaviour is the worst amongst the three measurement error structures considered. The bimodal distribution of u creates an underestimation of both the slope and the intercept of the model. In particular, $\hat{\beta}_1$ presents an high *attenuation-to-the-null* effect: its average value is about 0.56 for the three sample sizes considered. Table 2.21 presents a model that is completely unsatisfactory, with high bias and mean squared error for the slope $\hat{\beta}_1$. Table 2.22 shows a real coverage level equal to 0% for β_1, furthermore, the estimators are totally unaffected by the sample size considered. The present situation points out that using a naive regression when the measurement error distribution is complex would lead to a completely wrong inference on parameters. Therefore a method for correction must be used in order to enhance the performance of the naive model.



(a) x vs y, $n = 100$

(b) w vs y, $n = 100$

(c) x vs y, $n = 1.000$

(d) w vs y, $n = 1.000$

(e) x vs y, $n = 10.000$

(f) w vs y, $n = 10.000$

Figure 2.9: Effect of normal mixture measurement error $f_U = 0.5\phi(u+2) + 0.5\phi(u-4)$ in regression for three different sample sizes. The original linear relationship between y and x is split in two different patterns due to the bimodal measurement error distribution.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{NAIVE}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	6.4072	0.5690	6.4251	0.5711	6.4295	0.5714
Median	6.4100	0.5722	6.4184	0.5714	6.4287	0.5713
Bias	0.5928	1.4310	0.5749	1.4289	0.5705	1.4286
St. Dev	0.3715	0.0774	0.1121	0.0247	0.0356	0.0076
MSE	0.6996	1.4331	0.5857	1.4291	0.5716	1.4287
IQR	0.5265	0.1041	0.1502	0.0350	0.0466	0.0101

Table 2.21: Summary measures for the naive model in presence of normal mixture measurement error. The slope presents an high attenuation-to-the-null effect.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{NAIVE}$	β_0	β_1	β_0	β_1	β_0	β_1
Real	0.63	0.00	0.00	0.00	0.00	0.00
Real R	0.52	0.00	0.00	0.00	0.00	0.00
Real L	1.00	1.00	1.00	1.00	1.00	1.00
Average Length	1.45	0.38	0.45	0.12	0.14	0.04

Table 2.22: Inferential results for the naive method with $1 - \alpha = 0.95$ in presence of normal mixture measurement error. The real coverage levels for $\beta_0 = 7$ are lower than the nominal ones, whilst $\beta_1 = 2$ is contained in none of the $R = 1000$ simulated confidence intervals.

2.4.2 Regression-Calibration

The regression calibration is the only functional method which succeeded in effective correcting for the presence of the mixture of normals measurement error model. Likewise the skew-normal case presented in Section 2.3, the RC is the only approach that permits an improvement and thus an almost correct inference on parameters. Table 2.23 reports the sample mean and median of the estimators which are sufficiently close to the real value of β_0 and β_1. The RC technique applied to a normal mixture measurement error is the only case encountered in this simulation in which the sample size does have a significant effect in the performance of the estimator. In particular, Table 2.23 shows that the point estimate for β_1 is much closer to its real value and presents much less variability when $n = 1.000$ and $n = 10.000$ than when $n = 100$. Trying to understand the reason of this problem we discovered that the RC effectiveness depends on the "quality" of the additional data available for the analysis: when the additional information originates from a subset that is affected mainly by one of the two measurement error mixture component, the algorithm fails in performing an effectively correction. Since the mixture weights are both equal to 0.5 the aforementioned drawback is more likely to happen when the sample size is small, as a consequence the available gold standard could be strongly affected by only one measurement error mixture component and lead to biased results. However, this problem seldom happens when we consider bigger sample sizes.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{RC}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	7.0597	1.4205	6.9872	2.0550	7.0024	2.0100
Median	7.0301	2.0728	7.0151	2.0237	7.0012	2.0093
Bias	-0.0597	0.5795	0.0128	-0.0550	-0.0024	-0.0100
St. Dev	9.9691	14.0834	0.3587	0.2740	0.0989	0.0811
MSE	9.9693	14.0953	0.3589	0.2794	0.0989	0.0818
IQR	1.8797	1.4347	0.4894	0.2834	0.1321	0.1087

Table 2.23: Summary measures for the RC model in presence of normal mixture measurement error. $\hat{\beta}_0^{RC}$ and $\hat{\beta}_1^{RC}$ are slightly biased on average and they present high variability amongst the simulations.

As it has already happened several times, even though the descriptive results are fairly satisfactory, the real coverage level of confidence intervals for the true values of β_0 and β_1 are smaller than the nominal ones. Inferential results are summarized in Table 2.24, for three different coverage levels: 0.9, 0.95 and 0.99 respectively.

The assumption of having available a certain amount of the gold standard x is the main drawback of the regression calibration approach. However, it seems that the RC method is the only efficient solution amongst the functional methods when the measurement error model does not have a simple structure. Having additional information available permits to approximately deduct the distribution of the measurement error, and therefore to better correct for it. As already stated in Section 1.4.1, additional information is needed for parameters identification in certain models (i.e., RC) but also for providing a better measurement error correction when its distribution is complex.

$\hat{\beta}_{RC}$	$n = 100$		$n = 1.000$		$n = 10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.34	0.43	0.43	0.51	0.45	0.44
Real R	0.64	0.68	0.66	0.74	0.70	0.72
Real L	0.61	0.65	0.67	0.67	0.67	0.64
Average Length	1.95	2.07	0.37	0.29	0.12	0.09
$(1-\alpha)$=0.95						
Real	0.42	0.47	0.46	0.56	0.52	0.49
Real R	0.69	0.72	0.69	0.78	0.73	0.76
Real L	0.66	0.71	0.74	0.72	0.72	0.69
Average Length	2.33	2.48	0.44	0.35	0.14	0.11
$(1-\alpha)$=0.99						
Real	0.56	0.55	0.58	0.68	0.66	0.60
Real R	0.77	0.77	0.72	0.83	0.82	0.81
Real L	0.74	0.75	0.81	0.81	0.79	0.74
Average Length	3.09	3.28	0.58	0.46	0.18	0.14

Table 2.24: Inferential results for the RC method in presence of normal mixture measurement error with three different coverage levels. The real coverage levels are smaller than the nominal ones.

2.4.3 BCES

Likewise the skew-normal measurement error case in Section 2.3.3, the BCES estimator performs a good correction for the slope, but it fails in effectively estimating the intercept of the normal mixture measurement error model. As it is shown in Table 2.25, the BCES approach constantly underestimates the true value of the intercept, on the other hand it sufficiently correctly estimates the value of the slope β_1. The inferential results in Table 2.26 confirms what previously stated regarding the descriptive results: the real coverage levels for β_1 are comparable with the nominal ones, whilst the real coverage levels for β_0 are significantly lower than the real ones considered. The results are not affected by the sample size considered in the analysis.

All in all, the correction performed by the BCES approach can be considered fairly satisfactory. Neither additional data nor initial assumptions were required in order to perform the BCES algorithm. Thus the BCES method of moments can become useful in coping with an empirical linear regression with measurement error when no additional information is available. A real case of its application will be presented in Chapter 3.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{BCES}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	4.8001	2.1689	4.9792	2.0142	4.9964	2.0014
Median	4.9197	2.0136	4.9824	2.0007	4.9977	2.0003
Bias	2.1999	-0.1689	2.0208	-0.0142	2.0036	-0.0014
St. Dev	0.9596	0.6122	0.2266	0.1142	0.0721	0.0331
MSE	2.4001	0.6351	2.0335	0.1151	2.0049	0.0331
IQR	0.9912	0.4874	0.2961	0.1492	0.1003	0.0434

Table 2.25: Summary measures for the BCES model in presence of normal mixture measurement error. The slope is slightly overestimated on average, whilst the intercept is conspicuously underestimated by the BCES procedure.

$\hat{\beta}_{BCES}$	$n=100$		$n=1.000$		$n=10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.13	0.96	0.00	0.97	0.00	0.97
Real R	0.07	0.93	0.00	0.94	0.00	0.95
Real L	1.00	1.00	1.00	0.98	1.00	0.96
Average Length	3.48	2.32	0.83	0.48	0.26	0.15
$(1-\alpha)$=0.95						
Real	0.26	0.98	0.00	0.99	0.00	0.99
Real R	0.13	0.96	0.00	0.98	0.00	0.98
Real L	1.00	1.00	1.00	1.00	1.00	0.99
Average Length	4.16	2.77	0.99	0.58	0.31	0.18
$(1-\alpha)$=0.99						
Real	0.64	0.99	0.00	1.00	0.00	1.00
Real R	0.48	0.99	0.00	0.99	0.00	1.00
Real L	1.00	1.00	1.00	1.00	1.00	1.00
Average Length	5.51	3.66	1.30	0.76	0.40	0.23

Table 2.26: Inferential results for the BCES method in presence of normal mixture measurement error with three different coverage levels. The real confidence level for the intercept β_0 is lower than the nominal one.

2.4.4 SIMEX

Contrarily to what happened to the previously considered measurement error model structures, in the mixture of normal measurement error case the SIMEX application was not straightforward and many attempts have been done in order to find a solution that satisfactorily corrects for the error present in the covariate. Being the normal mixture a sophisticated measurement error distribution, some issues arose in choosing the distribution and the variance value of the artificially generated pseudo errors $\{u_b\}_{b=1}^{B}$ during the simulation step. The first attempt was to consider $\{u_b\}_{b=1}^{B}$ as generated by a mixture of normals distribution, namely the same used for generating the measurement error u. This solution led to a SIMEX algorithm that provided the same parameter estimations of the naive model presented in Section

2.4.1. Since the artificially generated bimodal errors did not implement an adequate correction, a normal distribution was chosen for $\{u_b\}_{b=1}^{B}$. The main drawback was that x^* presented high variability due to the nature of the measurement error distribution. Therefore, after many attempts, we have discovered that the best correction was obtained by setting the additional error variance σ_u^2 equal to 8^2. This value is certainly high, nevertheless it is needed in order to sufficiently take into account for the high variability of the measurement error distribution. However, as it is shown in Table 2.27, the correction performed by the SIMEX algorithm is far away of being perfect: it constantly underestimates the true value of both the intercept and the slope of the regression model. The bias and the MSE values are always considerable, no matter the sample size considered. The inferential results in Table 2.28 denote an inference on parameters in which the real coverage levels are far lower than the nominal ones. For example, with a nominal coverage level of 95% and $n = 100$, only in the 9% of the simulations the true value of β_1 belongs to the computed confidence interval.

Without any additional information required, the SIMEX approach still performs a significant improvement of the inference on parameters, if compared to the naive model. However, in order to perform a suitable estimation, the measurement error variance or at least information regarding measurement error variability has to be known, which is not always the case in dealing with real data applications.

	$n = 100$		$n = 1.000$		$n = 10.000$	
$\hat{\beta}_{SIMEX}$	β_0	β_1	β_0	β_1	β_0	β_1
Mean	5.5858	1.4327	5.5240	1.4576	5.5409	1.4522
Median	5.5846	1.4274	5.5341	1.4611	5.5472	1.4516
Bias	1.4142	0.5673	1.4760	0.5424	1.4591	0.5478
St. Dev	0.5057	0.1776	0.1560	0.0630	0.0500	0.0189
MSE	1.5019	0.5945	1.4843	0.5460	1.4600	0.5481
IQR	0.6749	0.2439	0.2230	0.0924	0.0708	0.0270

Table 2.27: Summary measures for the SIMEX model in presence of normal mixture measurement error. Both the slope and the intercept are on average underestimated by the SIMEX method.

$\hat{\beta}_{SIMEX}$	$n = 100$		$n = 1.000$		$n = 10.000$	
	β_0	β_1	β_0	β_1	β_0	β_1
$(1-\alpha)$=0.90						
Real	0.04	0.04	0.00	0.00	0.00	0.00
Real R	0.01	0.02	0.00	0.00	0.00	0.00
Real L	1.00	1.00	1.00	1.00	1.00	1.00
Average Length	1.06	0.54	0.32	0.17	0.10	0.05
$(1-\alpha)$=0.95						
Real	0.07	0.09	0.00	0.00	0.00	0.00
Real R	0.04	0.04	0.00	0.00	0.00	0.00
Real L	1.00	1.00	1.00	1.00	1.00	1.00
Average Length	1.26	0.64	0.38	0.20	0.12	0.06
$(1-\alpha)$=0.99						
Real	0.13	0.23	0.00	0.00	0.00	0.00
Real R	0.10	0.14	0.00	0.00	0.00	0.00
Real L	1.00	1.00	1.00	1.00	1.00	1.00
Average Length	1.67	0.85	0.50	0.26	0.16	0.08

Table 2.28: Inferential results for the SIMEX method in presence of normal mixture measurement error with three different coverage levels. The real coverage levels are greatly smaller than the nominal ones.

2.4.5 Methods comparison

As it could have been expected, the normal mixture is the measurement error distribution for which correcting for the presence of the mismeasured variable x^* creates major issues. None of the analysed functional methods provides a correction which is both efficient and stable at the same time. However, due to the sophisticated measurement error distribution, a perfect correction was not expected.

The Regression Calibration performs the best correction on average, nevertheless as we have already pointed out the estimators in the simulation present high variability, caused by the difference in quality of the gold standard considered for each algorithm (see §2.4.2). Moreover, it has not to be

	$\hat{\beta}_0$	$\hat{sd}(\hat{\beta}_0)$	$\hat{\beta}_1$	$\hat{sd}(\hat{\beta}_1)$
$n = 100$				
TRUE	6.9926	0.1000	1.9992	0.0506
NAIVE	6.4621	0.3710	0.5635	0.0968
RC	7.0597	0.3822	1.4205	0.2951
BCES	4.7869	0.8258	2.2066	0.5079
SIMEX	5.5858	0.3229	1.4327	0.1633
$n = 1.000$				
TRUE	7.0006	0.0315	2.0013	0.0158
NAIVE	6.4152	0.1156	0.5742	0.0297
RC	6.9872	0.1127	2.0550	0.0876
BCES	4.9749	0.2486	2.0027	0.1410
SIMEX	5.5240	0.0976	1.4576	0.0502
$n = 10.000$				
TRUE	7.0000	0.0100	2.0006	0.0050
NAIVE	6.4259	0.0365	0.5716	0.0094
RC	7.0024	0.0353	2.0100	0.0277
BCES	4.9871	0.0786	2.0030	0.0454
SIMEX	5.5409	0.0310	1.4522	0.0159

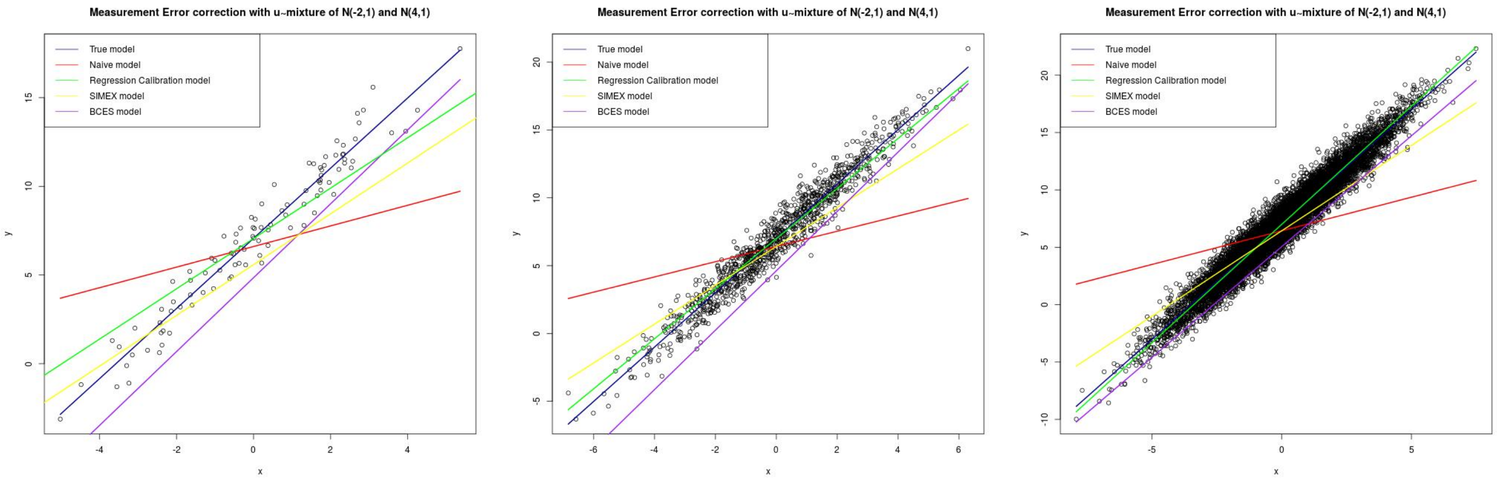
Table 2.29: Average values of the intercept, the slope and their standard errors for the normal mixture measurement error model with three different sample sizes. The RC method performs the best correction on average.

forgotten that the RC approach is feasible only when additional data are available, which is not always the case in real data application.

The BCES method performs a fairly good correction for the slope β_1, anyway it always underestimates the value of β_0. Amongst the functional methods analysed, the BCES leads on average to the worst estimation for the intercept β_0. Nonetheless, the slope estimator $\hat{\beta}_1^{RC}$ is almost equal in value to the OLS estimator $\hat{\beta}_1$ that would be obtained if the covariate were measured without error. Furthermore, no additional information is required for its usage, meaning that if the regression is linear and we are primarily interested in correctly estimating the parameter related to the variable measured with error the BCES estimator is a good alternative, also in presence of not banal measurement error distribution.

The SIMEX method performs a significant correction for both β_0 and β_1, even though none of the SIMEX estimators reach on average the parameters true values. Likewise the BCES estimator, also the SIMEX algorithm does not need additional data to be actuated, nevertheless a coherent measurement error variance needs to be specified in order to obtain a effective correction. The SIMEX approach is computationally intensive but it can be applied for correcting for measurement error presence in almost every type of regression, both linear and non-linear.

Table 2.29 summarizes the average mean and standard error obtained for each estimator in the simulation with normal mixture measurement error, with the three different sample size considered. As previously stated in Section 2.4.2, increasing the sample size produces a significant improvement only in the RC technique, whilst the other models are minimally effected. In Figure 2.10 the different approaches are graphically presented, for each sample size considered. The graphs highlight the correction performed by each estimator, it is clearly visible how every method succeeds in improving the general fit of the naive analysis.



(a) x vs y + fitted models, $n = 100$

(b) x vs y + fitted models, $n = 1000$

(c) x vs y + fitted models, $n = 10000$

Figure 2.10: Measurement error models fitted to the real data x vs y. The graphs present the average behaviour of the correction techniques in presence of normal mixture measurement error $f_U = 0.5\phi(u+2) + 0.5\phi(u-4)$, for the three sample sizes considered in the simulation.

2.5 Conclusions

In the present chapter a simple linear regression simulation study was performed in order to understand whether and how the RC, BCES and SIMEX methods effectively cope with different measurement error structures, and how the inference on parameters is influenced by the sample size. As previously defined, functional methods for dealing with measurement error are those methods in which no or few assumptions are made regarding the probability distribution of the involved quantities.

The aim of having simulated different types of measurement error was to clarify whether the correction performed by the aforementioned methods is robust, namely if the correction goodness remains the same in varying the measurement error distribution. In particular, we have chosen to simulate those measurement error structures that more likely are encountered in coping with astronomical data affected by uncertainties (Kelly [26], 2011). The simulation results highlight how the RC and the BCES are robust in general, even though the latter one presents some bias in estimating the intercept β_0 when the measurement error structure is more complex. On the other hand we discovered that the SIMEX approach does not present robustness, since its efficiency is highly influenced by the chosen distribution for the artificially simulated pseudo error $\{u_b\}_{b=1}^{B}$.

The aim of having considered different sample sizes was to understand whether the number of observations influences the inference on parameters in coping with measurement error. Of primarily interest was to find out whether the naive analysis improves in increasing the sample size, which means to understand whether a measurement error correction is needed when the sample size is extremely large. We have realized that the naive approach is not minimally affected by the size in the sample, this means that a measurement error correction is always needed when one or more covariates are affected by mis-measurement.

Having proved the necessity of methods for correction, the second aim was to understand whether the functional methods are influenced by the sample size. Once again, we assessed that the considered functional methods are only marginally influenced by the size in the sample, only the RC technique applied to the normal mixture measurement error turned out to be consistently influenced by it. The aforementioned results highlight the necessity of implementing measurement error algorithms that work also for massive astronomical datasets, since the measurement error presence cannot be omitted if we want to obtain correct estimates.

In the following chapter the functional methods applied in the simulation study will be used for coping with a linear regression study from a real astronomical dataset in which the covariate presents heteroscedastic measurement error.

Chapter 3

Hubble Data

3.1 Introduction

This chapter presents the analysis of a real astronomical data set in which one of the two variables considered is measured with error. The aim of the analysis is to apply the functional methods presented in Section 1.4 to a real measurement error regression problem. The dataset comes from the Surface Brightness Fluctuation (SBF) Survey of Galaxy Distances (Tonry et al [37], 2001). The SBF survey collects accurate measures of the distances from the nearby galaxies to the Earth: the aim of the data collection is to improve the knowledge of the local velocity field. The analysed dataset contains 280 observations of galaxies scattered throughout the sky; the data are available in digital form from *http://www.ifa.hawaii.edu/ jt/SBF*. Table 3.7 shows a preview of the entire dataset, presenting the variables of interest for 6 different galaxies.

The equation considered for the analysis is the famous relationship between the recessional velocity of a galaxy and its distance from the observer, known as Hubble's law:

$$v = H_0 D. \tag{3.1}$$

Hubble's law is a formula of observational cosmology stating that the velocities at which galaxies in the universe recede from each other is directly proportional to the distances between them (Hubble [24], 1953). Hubble's law is one of the pillars of the expanding space paradigm, since it mathematically represents the continuous expansion of the Universe. The motion of astronomical objects due only to this expansion is known as the Hubble flow. Even though attributed to Edwin Hubble, the law was first derived from the general relativity equations by Georges Lemaître, who proposed the theory of the expansion of the Universe and suggested an estimated value of the rate of expansion, the so-called Hubble constant H_0 (Lemaître [28], 1927). Many attempts of estimating the Hubble constant have been undertaken since 1927, the most recent estimation, dated June 6th 2014, provided a value for H_0 equal to $6,9 \times 10^{-5} \pm 0,7 \times 10^{-6}$ $km/s/pc$ (Bennett et al [4], 2014).

The velocity and the distance that appear in Hubble's law cannot be directly measured; they can only be derived from some directly observable quantities. Galaxy brightness provides information about the distance between the galaxy and the observer whilst the redshift provides a relation with the radiation spectrum of the galaxy. In Physics, redshift occurs when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum. The linear relationship between redshift and distance and the theoretical linear relation between recessional velocity and redshift leads to the straightforward mathematical formula in (3.1). For an extensive discussion on how these quantities are related, see Harrison (Harrison [22], 1993).

The remainder of the chapter is organized as follow. In Section 3.2 the statistical model used for the data analysis will be presented. Section 3.3 describes the naive analysis approach and underlines its limitations. Sections 3.4 and 3.5 present two functional methods for correcting for the measurement error present in the covariate. In the last Section comparisons between the methods are made and further research directions are presented.

3.2 The Hubble data model

The SBF survey provides the recessional velocity for each galaxy derived from its redshift together with its CMB [1] reference frame. The unit of measurement for v_{CMB} is km/s. The negative values of the recessional velocity for some galaxies is due to the fact that those galaxies were moving closer to the Earth when the data were collected. Galaxies which are getting closer to each other are defined to have a blueshift, which is a decrease in wavelength of electromagnetic waves, the opposite effect of the redshift.

On the other hand, the SBF survey does not provide a directly measure of the distance D, which has to be derived from the distance modulus $\mu = (m - M)$.

The distance modulus is a way of expressing distances used in Astronomy. It is calculated as the difference between the apparent magnitude m and the absolute magnitude M. The apparent magnitude of a celestial body is a measure of its brightness as seen from the Earth, adjusted to the value it would have without the presence of the atmosphere. On the other hand, the absolute magnitude is the measure of the intrinsic brightness celestial object. It is defined as the hypothetical apparent magnitude of an object at a standard luminosity distance of exactly 10.0 $parsecs$ from the observer, assuming no astronomical extinction of starlight.

A mathematical formula relates the distance modulus μ to the distance D of a celestial body from the observer:

$$log_{10} D = 1 + \frac{(m - M)}{5}. \tag{3.2}$$

An expression for computing the distance D given the distance modulus $(m - M)$ is obtained by inverting the relationship in Equation (3.2):

$$D = 10^{\frac{(m-M)}{5}+1}. \tag{3.3}$$

Thus, the statistical model applied to the data is a simple linear regression

[1] cosmic microwave background, which is the thermal radiation left over from the "Big Bang".

model:

$$v_{CMB} = \beta_0 + \beta_1 D + \varepsilon \tag{3.4}$$

where the parameter β_1 represents the Hubble constant H_0. Of primary interest is the correct estimation of β_1, in order to obtain an empirical confirmation of the theoretical value provided for H_0. Nonetheless, as previously stated, the involved quantities are not directly measured, therefore a measurement error structure is intrinsically present in the model. Hence, a way to correct for the measurement error presence must be provided in order to improve the naive estimation, which will likely be biased.

In the SBF sample additional information regarding the measured quantities is collected and can be exploited for enhancing the quality of the inference on the parameters. For the distance modulus $(m - M)$, all sources of error are summarized in a variable u, for which the standard deviation $\sigma_{u,i}$, $i = 1 \dots 280$, is provided for each galaxy. Due to the presence of measurement error, the observed covariate available for the analysis will then be:

$$D^* = 10^{\frac{(m-M)+u}{5}+1} \tag{3.5}$$

where D^* is a mismeasured quantity of the true unknown variable D. As it can be seen in Equation (3.5), the measurement error structure is non-linear for the considered model, since the error component appears in the exponential part of the formula. Furthermore, the measurement error u possesses an heteroscedastic variance, since the standard deviation σ_i varies for each observation.

No additional information for the radial velocity v_{CMB} is provided by the SBF survey, therefore the response variable in the model is assumed to be correctly measured.

In the end, the linear regression model that will be analysed is the following:

$$v_{CMB} = \beta_0 + \beta_1 10^{\frac{(m-M)+u}{5}+1} + \varepsilon \tag{3.6}$$

where the covariate presents a *non-linear measurement error structure with heteroscedastic variance.*

In the following sections we first perform the naive analysis and subsequently apply the BCES and the SIMEX methods in order to correct for the measurement error component.

3.3 Naive model

3.3.1 Preliminary analysis

In performing the naive analysis the additional information regarding the heteroscedastic nature of the measurement error u is not needed, since the covariate D^* is considered measured without errors. Initially a descriptive analysis of the involved quantities is performed and subsequently a linear model is fitted to the data. In Table 3.1 the main descriptive results for the response variable, namely the radial velocity v_{CMB}, and for the covariate, namely the distance D, are reported. The covariate D shows an high variability, mainly due to the presence of some galaxies which are really far away, as it can be seen from the boxplot in Figure 3.1.

	v_{CMB}	D
Minimum	-590.00	636795.52
Maximum	4939.00	52966344.39
1st Quartile	1153.50	15848931.92
3rd Quartile	2064.75	26791683.25
Mean	1607.44	21658062.15
St Dev	819.26	8917130.39
Median	1499.50	20989398.84
MAD	626.3985	8197213
Skewness	0.66	0.40
Kurtosis	1.82	0.55

Table 3.1: Descriptive statistics for the v_{CMB} and D variables

In Figure 3.2 the boxplot and the histogram of the response variable are also reported, the graphs show a slightly positive skewness, meaning that some galaxies present extreme values in terms of recessional velocity. The dots at the very bottom of the boxplot represent some galaxies that were hav-

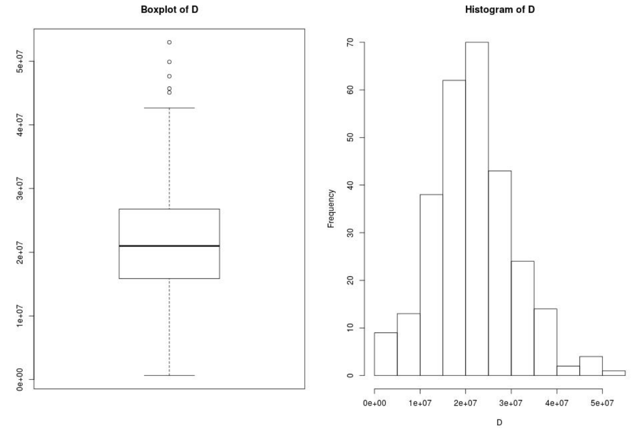


Figure 3.1: Boxplot and histogram of D

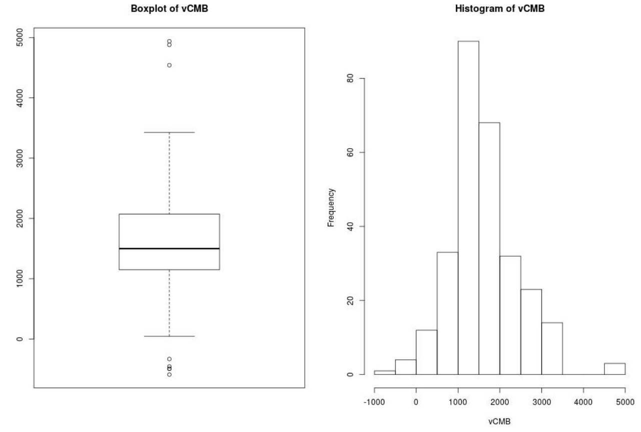


Figure 3.2: Boxplot and histogram of v_{CMB}

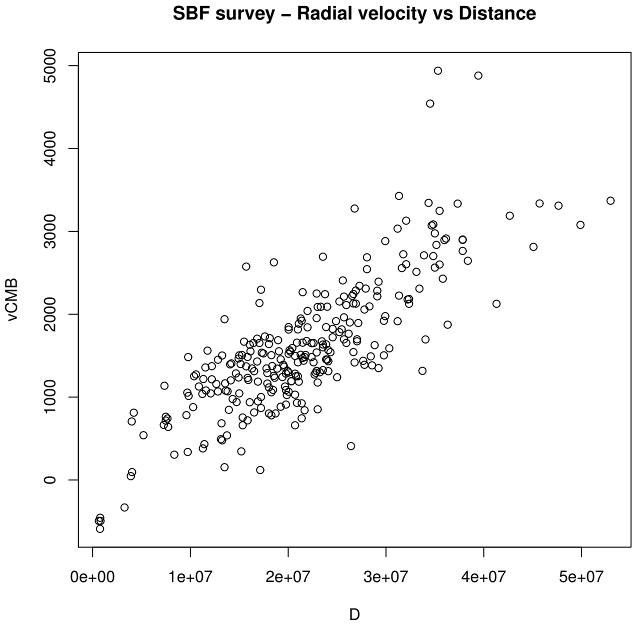


Figure 3.3: SBF survey - radial velocity vs distance

ing a blueshift when the data were collected; in other words their recessional velocity was negative because instead of receding they were approaching the Earth. Nonetheless, as it is underlined by the graph, only few galaxies of the total amount present this unexpected behaviour.

The scatter plot of radial velocity versus distance is reported in Figure 3.3: a linear pattern is clearly visible. In order to calculate the size of the linear relationship between v_{CMB} and D, the correlation coefficient is calculated:

$$\rho = \frac{Cov(v_{CMB}, D)}{\sqrt{Var(D)Var(v_{CMB})}} = 0.802. \tag{3.7}$$

The ρ value highlights a strong linear relationship between the variables. A test for evaluating the Pearson's product moment correlation coefficient provides a p-value smaller than 0.001, proving the significance of the index in Equation (3.7) for whichever value of α. It is therefore reasonable to proceed in fitting a simple linear model to the data.

3.3.2 Naive analysis

A simple linear model is fitted to the data, leading to the results summarized in Table 3.2.

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	10.9872	77.0186	0.14	0.8867
D	7.371e-05	3.289e-06	22.41	<0.001
			Residual standard error:	489.9
			R^2:	0.6437
			F-statistic:	502.2

Table 3.2: Summary output for the naive linear regression model. The analysis is performed without considering the presence of the measurement error of the variable D.

As it is visible from the table, the estimator for β_1 is significantly different from 0, with a p-value smaller than 0.01. Contrarily, the intercept is significantly equal to 0, as it was expected since the theoretical relationship $v = H_0 D$ in Hubble's law does not present an intercept. The coefficient of determination R^2 is equal to 0.6437, meaning that the fitted model explains almost 65% of the total variability amount present in the data.

Of primary interest is to to understand whether the estimate of the Hubble's constant provided by our naive analysis (i.e. $\hat{\beta}_1^{NAIVE}$) is statistically equal to the most recent estimate theorized by Bennett et al (2014). In order to perform the aforementioned test, a new linear model is fitted to the data using the *offset*[2] function:

$$vCMB = \beta_0 + H_0 D + \beta_1 D + \varepsilon \tag{3.8}$$

where H_0 is the value of the Hubble's constant provided by Bennett. The summary results for the model in Equation (3.8) are presented in Table 3.3. The slope $\hat{\beta}_1$ of the model in Equation (3.8) is no more statistically significant, meaning that the information provided by the $H_0 D$ component

[2] An offset is a term to be added to a linear predictor with known coefficient 1 rather than an estimated coefficient.

	Estimate	Std. Error	t value	Pr(>t)
(Intercept)	10.9872	77.0186	0.14	0.8867
D	4.712e-06	3.289e-06	1.43	0.1531
			Residual standard error:	489.9
			R^2:	0.6437
			F-statistic:	502.2

Table 3.3: Summary output for the naive linear regression model with offset. Neither the intercept nor the slope are statistically different from 0.

is already sufficient to explain part of the data variability, and then there is no need to insert an extra parameter β_1 in the model. The conclusion is that the H_0 value provided by Bennett and the estimate obtained with our model are statistically equal, when we do not take into account the measurement error present in the covariate D.

The analysis of the residuals from the fitted model is reported in Figure 3.4. The first graph represents the scatter plot of the residuals versus the fitted values: the residuals seem to have a good behaviour, with mean sufficiently equal to 0 and homogeneous variability. Three outliers are present in the dataset, as it is clearly visible in all 4 graphs, however this is not a serious issue since none of them is an influential point, being their leverage values less than 0.5, as it is reported in the forth graph. The graph in the bottom left represents a plot of approximate Cook statistics against leverage/(1-leverage). The SBF dataset presents some galaxies which are leverage points, meaning observations that have an extreme or outlying value in the independent variable D, as we have already pointed out in Section 3.3.1. The Q-Q plot in the top right graph does not present an ideal situation: realistically the residuals are not normally distributed.

If no information regarding the involved variables and their measurement error were available, the naive analysis developed so far could have been considered fairly acceptable. Nonetheless, it has been proved in Section 1.3.1.1 that when one of the covariates of a linear regression model is measured with error, the OLS estimator is inconsistent and provides biased inference on the parameters. Therefore, the obtained results must be corrected in

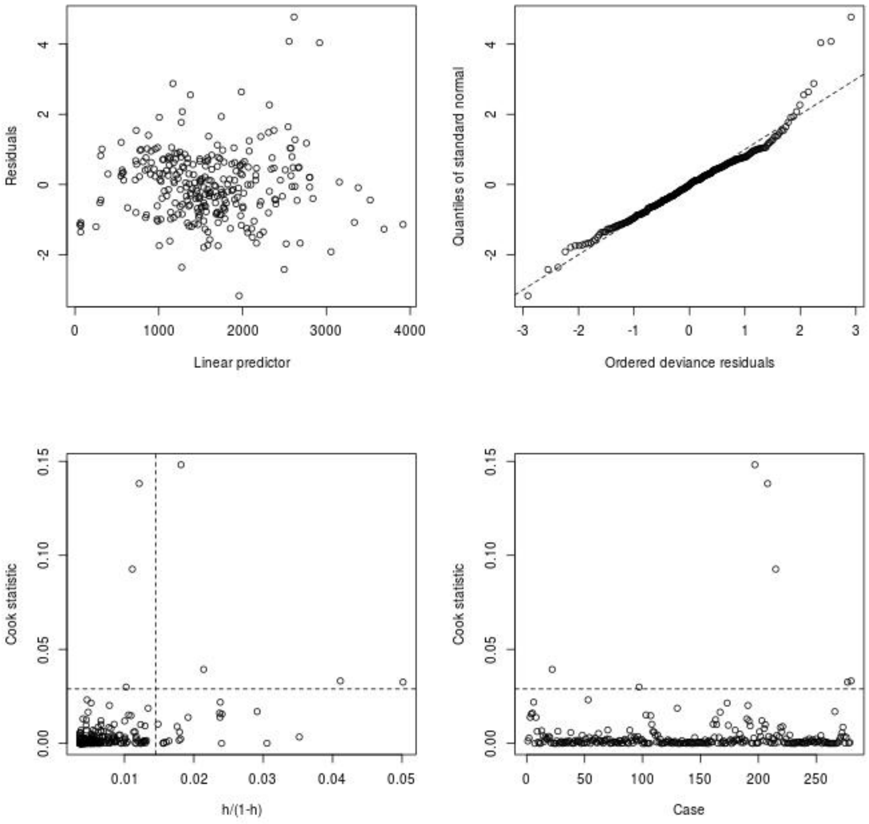


Figure 3.4: Residual plots for the naive model.

order to take into account the presence of a non-linear measurement error with heteroscedastic variance in the covariate D. In the following sections two functional techniques, namely BCES and SIMEX, will be applied to the original model.

3.4 BCES

As already presented in Section 1.4.2, the BCES approach is a widely applied technique belonging to the method of moments family; its effectiveness in coping with measurement error in linear regression has been proved in the simulation study presented in Chapter 3.

The SBF survey dataset presents an error structure which is slightly different compared to the classical measurement error $x_i^* = x_i + u_i$ considered in the simulation study. In particular, the measurement error u is heteroscedastic and non linear, as we underlined in Section 3.2. Therefore, some modifications in the BCES algorithm are necessary in order to take into account the more complex structure of the considered model. Mathematically, the mis-measured covariate D_i^* is represented by the following equation:

$$D_i^* = f(\mu_i + u_i) \tag{3.9}$$

where $f(\cdot)$ is the function in Equation (3.5) and μ_i is the distance modulus $(m_i - M_i)$. Considering the relationship in (3.9) it is not possible to provide a BCES estimation, since we cannot separate the variability due to the intrinsic scatter from the one due to measurement error and then subsequently correct for the latter one. Therefore, a good approximation is needed in order to separate the two variability sources. Using a Taylor series expansion, it is straightforward to prove that

$$f(\mu_i + u_i) \doteq f(\mu_i) + f'(\mu_i)u_i, \tag{3.10}$$

where the second addend $f'(\mu_i)u_i$ represents the new measurement error component, $u_i' = f'(\mu_i)u_i$. Using the linear approximation obtained through the Taylor series expansion in Equation (3.10), we developed a new BCES procedure for de-biasing the OLS estimator from the measurement error presence, which means to calculate the BCES estimates when the measurement error is non-linear. The expression for $\hat{\beta}_1^{BCES}$ and $\hat{\beta}_0^{BCES}$ are provided by the following formulas:

$$\hat{\beta}_1^{BCES} = \frac{\sum_{i=1}^{280}(D_i^* - \bar{D}^*)(v_{CMB,i} - \bar{v}_{CMB})}{\sum_{i=1}^{280}(D_i^* - \bar{D}^*)^2 - \sum_{i=1}^{280} f'(\mu_i^*)^2 \sum_{i=1}^{280}(u_i)^2} \tag{3.11}$$

$$\hat{\beta}_0^{BCES} = \bar{v}_{CMB} - \hat{\beta}_1^{BCES}\bar{D}^*. \tag{3.12}$$

Appendix B reports the technical procedure for obtaining the expression in Equation (3.11). Furthermore, we perform a simulation study in order to empirically prove the effectiveness in this particular case of our modification of the BCES technique. The results of the simulation study are reported in Chapter 4.

The summary of the BCES model is reported in Table 3.4. Contrarily to the naive model, $\hat{\beta}_0^{BCES}$ is significantly different from 0 at a 0.05 significance level, although its equality to 0 being accepted with $\alpha = 0.01$. $\hat{\beta}_1^{BCES}$ is highly significant, its equality to 0 would be rejected for whichever value of α. Hubble's constant (i.e. β_1) possesses a higher value when it is estimated with the BCES method than with the naive approach. This is likely due to the attenuation-to-the-null effect: the measurement error present in the covariate attenuates the slope estimate in linear models.

	Estimate	Std. Error	z value	Pr(>z)
(Intercept)	-1.941657e+02	9.628306e+01	-2.016614	0.04373584
D	8.318404e-05	4.591391e-06	18.117392	<0.001

Table 3.4: Summary output for the non-linear BCES model with heteroscedastic measurement error.

A test was performed in order to understand whether the estimate of H_0 provided by the BCES method (i.e. $\hat{\beta}_1^{BCES}$) is equal to the most recent value of Hubble's constant provided by Bennett. The p-value obtained for the described test is equal to 0.002, meaning that the null hypothesis of equality of the two values is rejected for whichever value of α. The BCES method provides a value for Hubble's constant which is statistically different from the very last estimate of it available in literature. This result is different compared to the one obtained considering the naive estimation for β_1 (see §3.3.2)

It is worth underlying that the entire procedure just described is approximate, meaning that the obtained BCES estimator will roughly possess the same properties of the one illustrated in Akritas & Bershady (1996).

Another last remark is about the independence assumption between μ_i and its measurement error u_i. For each galaxy the SBF survey provides only the measurement error standard deviation $\sigma_{u,i}$, with which it is not possible to compute the covariance between μ_i and u_i. Thus, the two quantities have been considered uncorrelated, although there is no theoretical guarantee that supports this assumption.

3.5 SIMEX

As underlined several times, the Simulation-Extrapolation technique is an highly flexible functional method that can be applied in contexts with different measurement error structures, with no or few modifications in the original algorithm. Contrarily to what it has been done in Section 3.4 with the BCES estimator, fitting a SIMEX model to the SBF dataset requires neither initial assumptions nor approximations in the original framework.

In the simulation step, additional heteroscedastic measurement errors are generated and added directly to the galaxy distance modulus, using the formula

$$w_{b,i}(\lambda) = 10^{(\mu_i + \sqrt{\lambda} U_{b,i})/5+1}, \qquad i = 1, \ldots, 280 \quad b = 1, \ldots, 1000. \tag{3.13}$$

The pseudo errors $\{U_{b,i}\}_{i=1}^{280}$ are mutually independent normal random variables with mean 0 and standard deviation equal to $\sigma_{u,i}$, provided by the SBF survey as an index of all sources of error for the distance modulus $(m_i - M_i)$. The remeasurement procedure in Equation (3.13) belongs to the SIMEX algorithms with heteroscedastic errors and known error variances; the extrapolation step for obtaining the SIMEX estimations of β_0 and β_1 is done in exactly the same way as described in Section 1.4.4. The summary results for the SIMEX model are shown in Table 3.5. Likewise the BCES model presented in the previous section, also the SIMEX algorithm reports an intercept which is significantly different from 0, contrarily to what is stated

	Estimate	Std. Error	z value	Pr(>z)
(Intercept)	-2.264771e+02	4.553398e+01	-4.973806	<0.001
D	8.542502e-05	2.299960e-06	18.693828	<0.001

Table 3.5: Summary output for the non-linear BCES model with heteroscedastic measurement error.

by Hubble's law. With the SIMEX approach the obtained p-value for β_0 is even lower than 0.01, leading to reject the null hypothesis for any nominal level of α. The incongruence of non-equality to 0 of the intercept is not a serious issue in terms of Hubble's law: the value of $\hat{\beta}_0^{SIMEX}$ is really small and it does not minimally affect the relation between the radial velocity and the distance. Furthermore, our primary interest is to provide an estimation of Hubble's constant (i.e. β_1) in which the measurement error present in the data is correctly modeled.

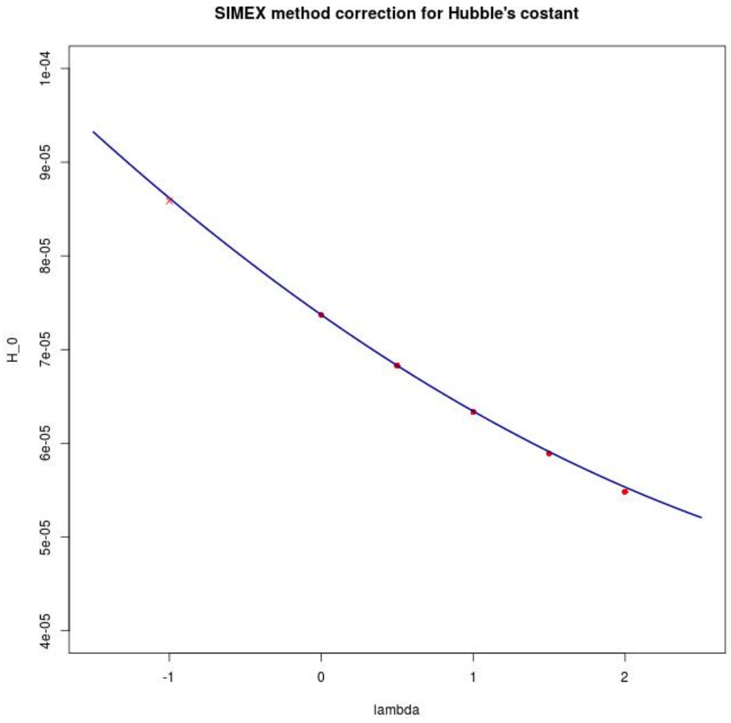


Figure 3.5: Simex method correction for Hubble's costant. The estimate provided by the SIMEX algorithm is statistically bigger than the one provided by the naive analysis.

The slope estimate is highly significant, $\hat{\beta}_1^{SIMEX}$ presents an even higher value than the β_1 estimation provided by the BCES method. A cubic extrapolant function has been fitted to the artificially remeasured values in order to provide the SIMEX estimator for H_0.

The astronomical interpretation is that Hubble's constant is underestimated by the naive model since the latter one does not take into account the attenuation-to-the-null-effect due to the measurement error component. Figure 3.5 reports the correction performed by the SIMEX method to Hubble's constant estimation.

Likewise what we have done for the BCES estimate of Hubble's constant, a test was performed for testing the equality between $\hat{\beta}_1^{SIMEX}$ and the most recent value of H_0 found in literature. The obtained p-value for this test is equal to 7.8×10^{-10}: the statistical equality between the involved quantities is rejected for whichever value of α. The conclusion is that the value of H_0 provided by the SIMEX method is statistically smaller than the value of H_0 provided by Bennett.

Astronomical possible reasons and implications regard this result are discussed in the following section.

3.6 Comparison and discussion

In the present chapter a real astronomical dataset was analysed. In particular, from each galaxy present in the SBF survey the recessional velocity and the distance modulus were used as variables in a simple linear regression model that represents the empirical formulation of well-known Hubble's law $v = H_0 D$. The particularity of the aforementioned model is that the covariate, namely the distance D, is affected by measurement error; for each observation its standard deviation is provided. As proved in Section 1.3.1.1 a mis-measured covariate in a linear regression model leads to a wrong inference for the parameters and particularly an underestimation of the real slope value β_1. Therefore, two of the three functional methods described in Chapter 1 were utilized in order to correct for the presence of measurement error. It was not possible to provide a Regression Calibration estimation since no additional information was available.

Figure 3.6 graphically reports the final result of the functional methods application. The SIMEX and the BCES lines are almost overlapping, meaning that the estimation provided by the two techniques is rather similar. On

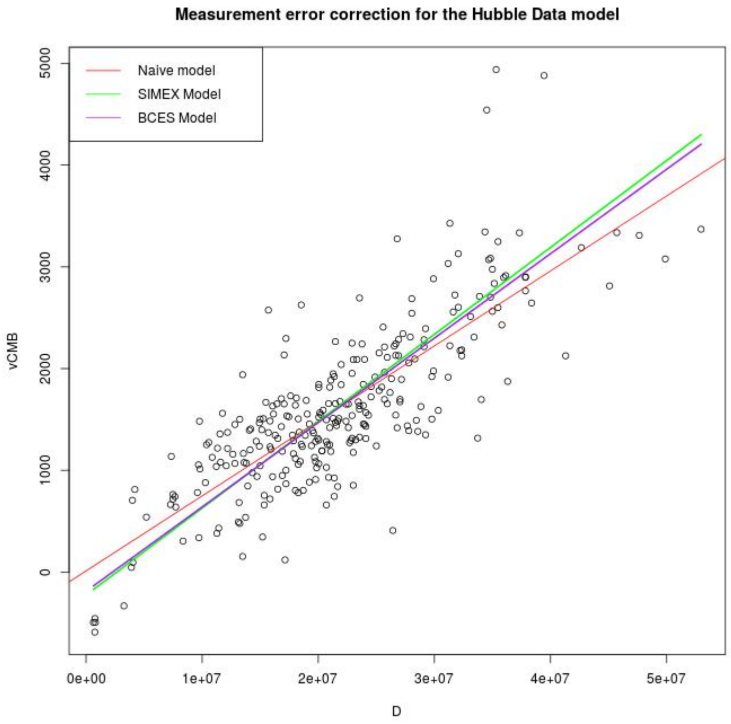


Figure 3.6: Measurement error correction for the Hubble data model, SBF survey. The SIMEX and the BCES line present a similar pattern, the naive model instead presents a smaller value for the slope.

the other hand, the slope estimate provided by the naive model leads to a line that is less inclined compared to the two ones which consider measurement error.

Table 3.6 reports confidence intervals for Hubble's constant (i.e., β_1) for the three considered methods with different coverage levels. Likewise the point estimate $\hat{\beta}_1^{BCES}$ and $\hat{\beta}_1^{SIMEX}$, also the confidence intervals provided by the two approaches are fairly similar. Nevertheless, since the BCES possesses higher standard deviation, the associated intervals are larger if compared to the SIMEX one. As it was expected, the values within the naive confidence intervals are significantly smaller than the ones inside the confidence intervals in which the measurement error has been corrected.

We performed two tests in order to prove the statistical equality between

NAIVE	Lower Limit	Upper limit
$(1-\alpha) = 0.90$	6.82834e-05	7.913996e-05
$(1-\alpha) = 0.95$	6.723685e-05	8.018651e-05
$(1-\alpha) = 0.99$	6.518081e-05	8.224255e-05
BCES	Lower Limit	Upper limit
$(1-\alpha) = 0.90$	7.566074e-05	9.025995e-05
$(1-\alpha) = 0.95$	7.426233e-05	9.165837e-05
$(1-\alpha) = 0.99$	7.152921e-05	9.439148e-05
SIMEX	Lower Limit	Upper limit
$(1-\alpha) = 0.90$	8.097646e-05	9.115925e-05
$(1-\alpha) = 0.95$	8.000108e-05	9.213463e-05
$(1-\alpha) = 0.99$	7.809477e-05	9.404094e-05

Table 3.6: Confidence intervals for Hubble's constant H_0, considering the naive, the BCES and the SIMEX analysis.

$\hat{\beta}_1^{BCES}$ and $\hat{\beta}_1^{SIMEX}$. Firstly we test $H_0 : \beta_1 = \hat{\beta}_1^{SIMEX}$ using the estimates provided by the BCES model. Consequently, we test $H_0 : \beta_1 = \hat{\beta}_1^{BCES}$ considering the estimates provided by the SIMEX model. In both cases the null hypothesis is accepted for whichever value of α, these results lead to the conclusion of $\hat{\beta}_1^{BCES}$ being statistically equal to $\hat{\beta}_1^{SIMEX}$.

The performed data analysis has also an astronomical interpretation regarding the estimation of Hubble's constant value over the years. From a statistical point of view it would be more correct to refer to Hubble's constant as the *Hubble's parameter*, since it measures the expansion rate of the Universe that changes with time. If the Universe is decelerating, Hubble's constant is decreasing. If Hubble's constant is increasing, the Universe is accelerating. As already underlined in the introduction, the very last estimate of Hubble's parameter was done in 2014, whilst the SBF Survey was conducted in 2001. Comparing the estimates obtained by our analysis ($H_0 = 8.296 \times 10^{-5} \pm 8.7 \times 10^{-6}$ with the BCES and $H_0 = 8.542 \times 10^{-5} \pm 6.1 \times 10^{-6}$ with the SIMEX) with the estimation provided by Bennett et al ($H_0 = 6,9 \times 10^{-5} \pm 0,7 \times 10^{-6}$) it is possible to note a diminution in the value of Hubble's constant. Therefore, the rational conclusion would be that the Universe is decelerating, but there is an intermediate regime in which the Universe is accelerating and Hubble's constant is decreasing; that is what the astronomers suppose is happening right now (Carroll et all [11], 1992). Since Hubble's law relates recessional velocity of a galaxy to its distance from the Earth, if increasing rate of the distance is higher than the decreasing rate of Hubble's constant then the recessional velocity can still augment.

In conclusion, it is worth highlighting how the obtained results may lead to further research related to this area. Since its first measurement attempt in 1927, Hubble's constant estimation seems to show a decreasing trend. Is this behaviour going to last? What would the consequences be if the Universe started to decelerate? There are still many unresolved issues related to this field. Fortunately the continuous technological improvement permits to have available increasingly large and complex amount of data, allowing the observational astronomers to constantly monitor the celestial bodies that surround us. Furthermore, the analysis performed in this chapter underlines the compelling necessity of considering the presence of possibly mismeasured variables.

The knowledge of advanced statistical methods for coping with massive datasets will therefore come more and more a skill that will have potential application in Astronomy and Astrophysics.

SBF DATA

Galaxy	RA	Dec	v_{CMB}	T	Grp	A_B	V-I	$\bar{m}_I$	M-m	σ_{M-m}	$\langle r \rangle$	Q	PD	$\bar{N}_I$
N7814	0.81	16.15	684	2	0	0.19	1.25	29.29	30.60	0.14	36	8.30	0.51	20.30
N0063	4.44	11.45	803	0	0	0.48	0.98	28.85	31.36	0.33	24	5.40	1.14	17.80
N0147	8.30	48.51	-456	-5	282	0.75	1.02	22.13	24.44	0.16	35	9.90	0.01	13.60
N0185	9.74	48.34	-494	-5	282	0.79	1.05	21.83	24.02	0.16	43	9.90	0.01	14.20
N0221	10.68	40.87	-494	-6	282	0.35	1.13	22.73	24.55	0.08	30	9.90	0.01	15.60
N0224	10.69	41.27	-590	3	282	0.35	1.23	23.03	24.40	0.08	51	9.90	0.01	21.60
N0274	12.76	-7.06	1390	-3	0	0.24	1.14	29.64	31.45	0.47	33	3.00	1.68	18.80
N0404	17.36	35.72	-332	-3	0	0.25	1.05	25.40	27.57	0.10	23	9.90	0.01	16.90
N0448	18.82	-1.62	1589	-3	0	0.26	1.13	30.58	32.41	0.35	18	4.30	1.41	19.30
N0524	21.20	9.54	2091	-1	0	0.36	1.22	30.48	31.90	0.20	64	2.60	1.65	21.60
N0584	22.84	-6.87	1566	-5	26	0.18	1.16	29.82	31.52	0.20	23	3.10	1.41	20.60
N0596	23.22	-7.03	1509	-4	26	0.16	1.14	29.89	31.69	0.10	31	3.10	1.51	20.30
N0636	24.78	-7.51	1504	-5	26	0.11	1.16	30.65	32.37	0.16	19	3.80	1.13	20.60
N0720	28.25	-13.74	1438	-5	0	0.07	1.21	30.76	32.21	0.17	32	4.00	1.09	21.70
N0821	32.09	11.00	1433	-5	0	0.47	1.20	30.38	31.91	0.17	30	5.30	1.00	20.90

Table 3.7: Data preview from the SBF Survey. The complete dataset can be found in digital form from *http://www.ifa.hawaii.edu/ jt/SBF*.

Chapter 4

BCES Method Simulation Study

4.1 Introduction

In the previous chapter we analysed a real astronomical dataset in which one variable is measured with error. Specifically, the covariate in the linear regression model of Equation 3.6 presents a non-linear measurement error with heteroscedastic variance. We applied two functional methods in order to correct for it, namely BCES and SIMEX. In particular, we developed and implemented a new version of the BCES method, since the original technique is applicable only to classical measurement error. More precisely, to separate the variability due to the intrinsic scatter from the one due to the measurement error we linearised the function in Equation 3.5 through a first order Taylor series expansion. The technical details we employed for obtaining our modified version of the BCES estimator are described in Appendix B.

In the present chapter we consider the same model structure we encountered in analysing the data from the SBF-survey and we assess via simulation the effectiveness of our BCES method in estimating Hubble's constant. Considering $R = 10.000$ replications, we focus on the estimation of β_1 comparing the results obtained by the true and the naive model with the estimations provided by our non-linear BCES method.

4.2 The simulated model

Since we want to assess the effectiveness of our BCES method in correctly estimating Hubble's constant, the theoretical model used for the simulation is the one considered in the previous chapter:

$$y = \beta_0 + \beta_1 10^{\frac{x+u}{5}+1} + \epsilon \tag{4.1}$$

where $\beta_0 = 0$, $\beta_1 = 2$ and $\epsilon \sim N(0, 10^{10})$. x is randomly generated by a normal distribution with mean equal to 30 and variance equal to 1.69. Both ϵ and x are coherently generated in order to maintain the same order of magnitude of the quantities involved in the Hubble data model. The measurement error u follows a normal distribution with 0 mean and heteroscedastic variance σ_i, which was generated by a chi-squared distribution with mean equal to 0.5. The sample size selected for the simulation is $n = 300$. The real value of the slope $\beta_1 = 2$ in the simulation does not reflect the order of magnitude of Hubble's constant; this is chosen on purpose in order to test the correctness of our method in estimating the slope of an arbitrary model whose structure is equal to the one in Equation (4.1).

Likewise Chapter 2, two summary tables are reported: the first table shows the descriptive results for each method whilst the second illustrates the main inferential results extracted from the simulation.

4.3 Simulation results

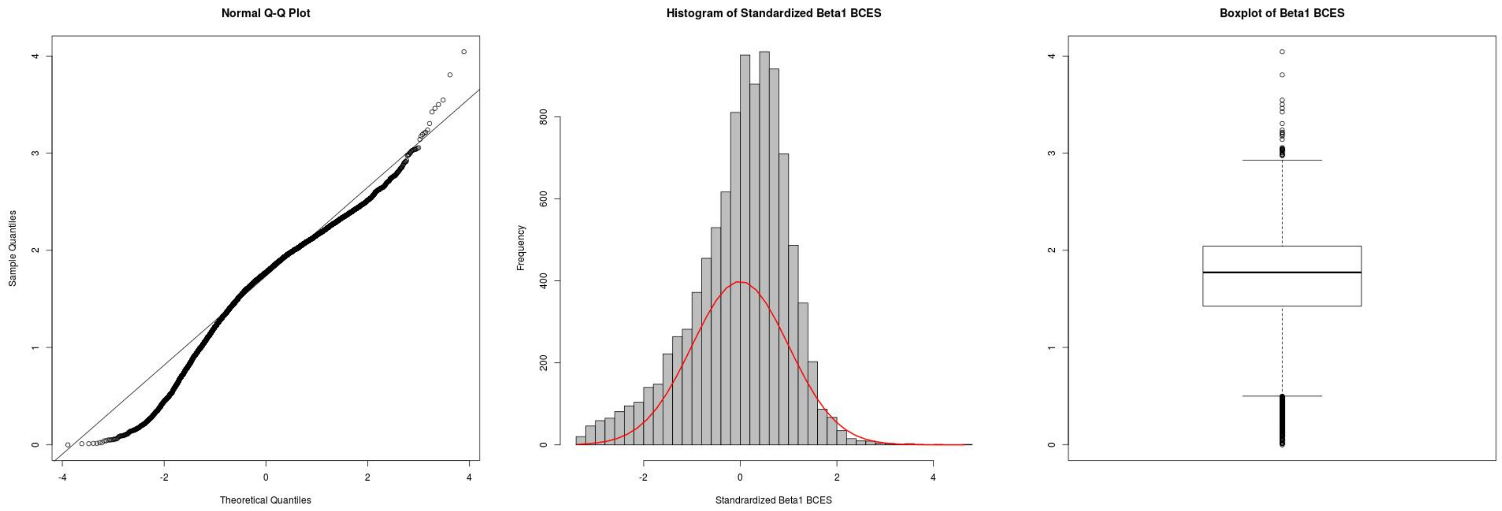
Of primary interest in the previous chapter was the correct estimation of Hubble's constant (i.e. β_1) from a model in which the covariate was measured with error. In this section we report the results of the simulation study where we use our modification of the BCES technique for estimating β_1. The effectiveness of our method is assessed comparing the estimates of β_1 with the ones obtained from the true and the naive model. The true model is obtained calculating an OLS estimator using the true variable $f(x)$ as a covariate, whilst the naive model performs an OLS estimation considering the mismeasured variable $f(x+u)$.

$\hat{\beta}_1$	TRUE	NAIVE	BCES
Mean	2.00	1.23	1.70
Median	2.00	1.23	1.79
Bias	<0.01	0.77	0.30
St. Dev	<0.01	0.33	0.49
MSE	<0.01	0.84	0.57
IQR	<0.01	0.41	0.61

Table 4.1: Summary measures of β_1 for the true, the naive and the BCES model. Our modifications of the BCES method succeeds in improving on the naive analysis.

Table 4.1 reports the summary measures obtained for β_1 fitting the true, the naive and our modified BCES model to the simulated data. Using the true model as a benchmark, the results point out that our modified version of the BCES technique succeeds in improving on the naive analysis. Both the mean and the median are closer to the real value $\beta_1 = 2$ than the naive model, meaning that the attenuation-to-the-null effect is partially corrected, even though it has not completely disappeared. Nonetheless a drawback of the BCES technique highlighted by the simulations is its variability: both the standard deviation and the interquartile range of $\hat{\beta}_1^{BCES}$ are considerably high.

Another aspect of interest is the approximation to the finite-sample distribution of our $\hat{\beta}_1^{BCES}$ estimator. In order to assess the validity of the central limit theorem in approximating the unknown distribution of $\hat{\beta}_1^{BCES}$ by a normal one, three graphs obtained from the simulation study are reported in Figure 4.1: the normal Q-Q plot, the histogram and the boxplot of $\hat{\beta}_1^{BCES}$. In particular, the normal Q-Q plot in graph (a) highlights that the unknown $\hat{\beta}_1^{BCES}$ distribution can be quite acceptably approximated by a normal one, even though two problems are clearly visible: the empirical distribution seems to be slightly asymmetric and, most of all, it is translated



(a) Q-Q plot of $\hat{\beta}_1^{BCES}$

(b) Histogram of standardized $\hat{\beta}_1^{BCES}$ + Standard Normal Distribution

(c) Boxplot of $\hat{\beta}_1^{BCES}$

Figure 4.1: Summary plots of $\hat{\beta}_1^{BCES}$ for $R = 10.000$ simulations: the normal Q-Q plot, the histogram and the boxplot are reported. The graphs seem to partially confirm the approximately normal finite-sample distribution of $\hat{\beta}_1^{BCES}$, although a symmetry problem is undoubtedly present.

to the left due to the attenuation-to-the-null effect caused by the measurement error. For the left-half of the distribution, the sample quantiles are always smaller than the theoretical ones. The same issue is visible also in the histogram of the standardized $\hat{\beta}_1^{BCES}$ values in graph (b). The boxplot in graph (c) underlines the presence of extreme values, — out of the boundaries of Q_1 and Q_3 — mainly in the left tail. All in all, a normal approximation can be considered sufficient in order to provide the required quantiles for calculating inference results, such as confidence intervals and statistical tests.

$\hat{\beta}_1$	TRUE	NAIVE	BCES
$1-\alpha=0.90$			
Real	0.90	0.00	0.91
Real R	0.90	0.00	0.92
Real L	0.90	1.00	0.95
Average Length	0.02	0.17	2.21
$1-\alpha=0.95$			
Real	0.95	0.00	0.92
Real R	0.95	0.00	0.94
Real L	0.95	1.00	0.96
Average Length	0.03	0.20	2.71
$1-\alpha=0.99$			
Real	0.99	0.00	0.96
Real R	0.99	0.00	0.96
Real L	0.99	1.00	0.99
Average Length	0.04	0.26	3.57

Table 4.2: Inferential results for β_1 for the true, the naive and the BCES model with three different coverage levels. The real coverage levels of our BCES technique reflects the nominal ones, nevertheless the average length of the confidence intervals is large.

The inferential results of β_1 for the true, the naive and the BCES model are reported in Table 4.2. Contrarily to the naive model, in which none of the simulated confidence intervals contains the real value $\beta_1 = 2$, our BCES method provides real coverage levels that reflect the nominal ones. The results in terms of inference on the parameter are fairly satisfactory for all three coverage levels considered. Nonetheless, as already underlined by the summary measures in Table 4.1 a problem that afflicts our BCES technique is the variability: the average length of the confidence intervals is large.

4.4 Conclusions

In the present chapter a simulation study was performed with the aim of empirically assessing the applicability of our modified version of the BCES method to a model that presents non-linear measurement error. In particular, our aim was to investigate the effectiveness of our approach in coping with the regression model encountered in the previous chapter while trying to estimate Hubble's constant. Therefore, the model chosen for the simulation is exactly the same we dealt with while analysing the dataset from the SBF-survey.

The results of the simulation underline the effective improvement made by our technique on the naive analysis. The non-linear BCES method weakens the attenuation-to-the-null effect and provides estimates that are closer to the real value of the slope β_1. We also confirmed the suitability of using a normal model for approximating the unknown distribution of $\hat{\beta}_1^{BCES}$.

Nonetheless, there remain some issues that require further investigation. The provided estimates present high variability, which lowers the precision of the point estimates and leads to confidence intervals whose average length is large. Furthermore, we investigated the validity of our modified method only in the specific case we were interested in. Further research is therefore needed in order either to prove its general applicability or to provide a widely-suitable modification of the original BCES version.

Discussion and final remarks

The thesis focuses on functional methods for correction of measurement error in Astronomy. In particular, we evaluated the applicability and the behaviour of these correction techniques when different measurement error structures and sample sizes are present.

Firstly, we implemented three functional methods for correction, namely BCES, RC and SIMEX, in the R programming language. Then, a simulation study was performed for a simple linear regression model, considering three different distributions for the measurement error: normal, skew-normal and normal mixture. We chose these particular structures in order to address the outstanding issues underlined by Brandon C. Kelly in his paper "Measurement Error Models in Astronomy" (2011). Specifically, he argues that the uncertainties in astronomical quantities are "large, skewed, or exhibit multiple modes"; our simulation analysis was driven by this statement.

Each simulation was repeated considering three different sample sizes. The results highlight that the correction techniques generally succeed in improving on the naive analysis. However, the functional methods we examined behave differently when assuming different measurement error models. The BCES method works extremely well when the measurement error model is simple and symmetric, nonetheless it provides misleading inferences when the measurement error distribution is more complex. The RC technique leads to satisfactory outcomes on average, although the simulation results are characterized by high variability and additional data must be provided. The SIMEX method always succeeded in adequately correcting for the attenuation-to-the-null effect; the computational burden is however large and the probability distribution for the artificially added errors must be correctly specified for its implementation. An interesting result is that none of the functional methods is greatly affected by the sample size considered. We noticed a very slight improvement in the correction techniques when the sample size increases, and almost no improvement in the naive model. Therefore, when a variable is measured with error, it is advisable to perform a measurement error correction, no matter the sample size involved.

Secondly, we analysed a real astronomical dataset in which the covariate is measured with error. The measurement error presents a non-linear structure and heteroscedastic variance; thus, we developed a new procedure for obtaining the BCES estimation of the parameters for this particular case. In particular, analysing the data collected in 2001 by the SBF-survey, we provided an estimate of Hubble's constant H_0 and we compared it with the last estimation available in literature, dated 2014. Our analysis proved that the two values are statistically different, which seems to indicate that Hubble's constant has decreased during the past decade. This result needs further investigation, one possible research direction is to monitor the trend of Hubble's constant whilst considering a measurement error model also for the radial velocity.

Thirdly, we empirically proved the effectiveness of our modified version of the BCES approach through a simulation study. The results underlined the actual validity of our method in coping with Hubble's data model. Nevertheless the simulation study was specifically tailored for our necessity, therefore we cannot assure its validity in treating a general case.

Finally, our study referred to situations in which only functional methods for correction have been applied. Additional research may consider the application of structural methods for correction in both the simulation study and in the analysis of the data provided by the SBF-survey. Alternatively, further investigation may regard the application of the analysed functional methods to a model in which more than one covariate is measured with error; more complex error structures should then be taken into account, as e.g.

correlated errors.

With the present work a statistical issue, i.e., the correct measurement of a quantity, has been presented and applied to a specific scientific field. The lesson learned is the compelling necessity of providing realistic statistical models in which the measurement error structure is correctly modeled. When the involved quantities cannot be directly measured, like in Astronomy, accounting for the measurement error present in the variables means providing correct inference and therefore real astronomical knowledge. A field in which mankind still has much to discover.

Appendix A

Skew-Normal Distribution

A.1 Introduction

In probability theory and statistics, the skew normal distribution is a continuous probability distribution that generalizes the normal distribution to allow for skewness. Being a manipulation of the most famous and widely used statistical distribution, it is not clear when its analytical form appeared for the first time. A paper by Birnbaum dated 1950 presents a mathematical formula which is equal to the modern SN (Skew-Normal) definition (Birnbaum [5], 1950). Nevertheless the first idea of extending the normal class of distributions in a constructive formulation via a population selection mechanism can be found in “Sulla rappresentazione analitica delle curve abnormali” (De Helguero [13], 1908). This appendix provides a brief introduction of the formulation and usage of the Skew-Normal distribution. For a detailed report regarding this topic see“The Skew-normal Distribution and Related Multivariate Families” (Azzalini [2], 2005)

A.2 Analytical construction

Consider a continuous random variable X having probability density function of the following form

$$f(x) = 2\phi(x)\Phi(\alpha x), \tag{A.1}$$

where $\phi(x)$ denotes the standard Normal (Gaussian) density function and $\Phi(\alpha x)$ its distribution function evaluated at point αx. The component α is called the *shape parameter* because it regulates the shape of the density function. As it is defined, the density $f(x)$ enjoys various interesting formal properties:

- $f(x)$ is equal to the Gaussian density function when the shape parameter $\alpha = 0$

- augmenting the absolute value of the shape parameter α the skewness of the distribution increases

- when $\alpha \to \infty$, the density converges to the commonly named half-normal (or folded normal) density function

- the sign of α indicates the skew direction of the distribution: left-skew when $\alpha > 0$ and right-skew when $\alpha < 0$ (see Figure A.1).

In order to obtain a representation of the Skew-Normal distribution location parameter ξ and scale parameter ω have to be added to the defined above random variable X. Therefore a linear transformation of X is considered:

$$Y = \xi + \omega X \tag{A.2}$$

Y is defined as a random variable Skew-Normally distributed with location parameter ξ, scale parameter ω and shape parameter α. The probability distribution function of Y is given by

$$f_Y(y|\xi, \omega^2, \alpha) = \frac{2}{\sqrt{\omega^2 + \alpha^2}}\phi(\frac{y-\xi}{\sqrt{\omega^2 + \alpha^2}})\Phi(\frac{\alpha}{\omega}\frac{y-\xi}{\sqrt{\omega^2 + \alpha^2}}) \tag{A.3}$$

A concise notation for this distribution is the following:

$$Y \sim SN(\xi, \omega^2, \alpha) \tag{A.4}$$

It is worth noticing that when $\alpha = 0$, a normal distribution $N \sim (\xi, \omega^2)$ is obtained. The following part presents some characteristic values of the random variable Y. Firstly, define the subsequent quantities:

$$\delta = \alpha/\sqrt{1+\alpha^2}$$

$$E(X) = \sqrt{2\pi}\delta$$

$$Var(X) = 1 - 2\delta^2/\pi$$

Having defined these quantities it is possible to retrieve the expected value, variance, skewness and kurtosis of a generic Skew-Normal random variable Y in the following form:

$$E(Y) = \xi + \omega\sqrt{2/\pi}\delta \tag{A.5}$$

$$Var(Y) = \omega^2(1 - 2\delta^2/\pi) \tag{A.6}$$

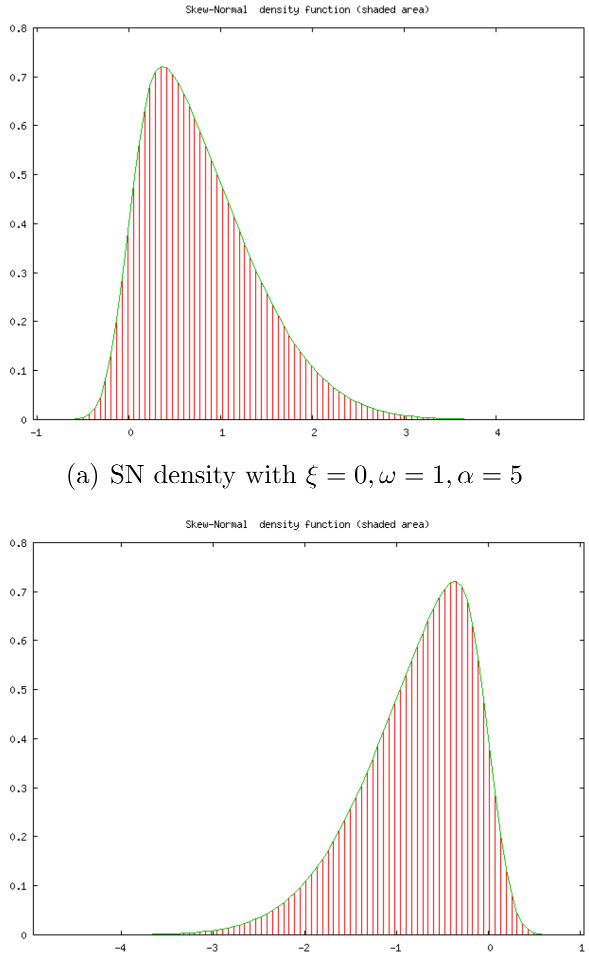
$$\gamma_1 = \frac{4-\pi}{2}\frac{E(X)^3}{Var(X)^{3/2}} \tag{A.7}$$

$$\gamma_2 = (2\pi - 3)\frac{E(X)^4}{Var(X)^2} \tag{A.8}$$

The distribution presented so far is used to fit data which are “normal-like” shaped but show a lack of symmetry. The Skew-Normal distribution family is a generalization of the Normal distribution family: it possesses the same relationship with the χ^2 distribution, that is, being X a generic $SN(0, 1, \alpha)$,

$$X^2 \sim \chi_1^2,$$

no matter the value of the shape parameter α. A multivariate generalization of the Skew-Normal distribution also exists, see Azzalini [2] 2005 for further details.



(a) SN density with $\xi = 0, \omega = 1, \alpha = 5$

(b) SN density with $\xi = 0, \omega = 1, \alpha = -5$

Figure A.1: Skew-Normal distributions with two different values of the parameter α

Appendix B

Technical details for the BCES method

In this appendix we present the newly developed methodological results achieved in the application of the BCES algorithm to a non-linear transformation of the covariate measured with error. The initial situation is a simple linear regression model with non-linear measurement error in the covariate, as encountered in SBF-survey of Chapter 3:

$$y_i = \beta_0 + \beta_1 f(x_i^*) + \varepsilon_i = \beta_0 + \beta_1 f(x_i + u_i) + \varepsilon_i, \tag{B.1}$$

in which the $f(\cdot)$ function is assumed known and non-linear. Being a method of moments estimator, the BCES approach is based on the fact that the parameters of Equation (B.1) are related to the moments of the bivariate distribution of $(Y, f(X))$ through:

$$\beta_1 = \frac{COV(Y, f(X))}{V(f(X))} \tag{B.2}$$

$$\beta_0 = E(Y) - \beta_1 E(f(X)). \tag{B.3}$$

Nonetheless, we do not observe realizations from the random variable $f(X)$; instead we observe realizations from the proxy random variable $f(X^*)$. The BCES methods replaces the expected moments of Equations (B.2) and (B.3) with moments estimators obtained from the observed data $(y, f(x^*))$. Assuming the independence between the measurement error U and the covariate

X, three approximated results are necessary in order to construct the BCES estimator for a non-linear measurement error in the covariate:

1. $E(f(X)) \doteq E(f(X^*))$
2. $V(f(X)) \doteq V(f(X^*)) - E(f'(X^*)^2)E(U^2)$
3. $COV(Y, f(X^*)) \doteq COV(Y, f(X))$.

Proof of relation 1:

$$\begin{aligned} E(f(X^*)) &= E(f(X+U)) \\ &\doteq E(f(X) + f'(X)U)) \\ &= E(E(f(X) + f'(X)U|X)) \\ &= E(f(X)) + f'(X)E(U|X)) \\ &= E(f(X)). \end{aligned}$$

Proof of relation 2:

$$\begin{aligned} E(f(X^*)^2) &\doteq E(E(f(X+U)^2|X)) \\ &= E(E(f(X)^2 + f'(X)^2U^2 + 2f(X)'f(X)U|X)) \\ &= E(f^2(X)) + E(f'^2(X)E(U^2|X)) \\ &= E(f^2(X)) + E(f'^2(X))E(U^2) \end{aligned}$$

from which it follows that $E(f^2(X)) = E(f(X^*)^2) - E(f'^2(X))E(U^2)$ and therefore, being $V(Z) = E(Z^2) - E(Z)^2$ for any random variable Z, we obtain the proof of relation 2.

Proof of relation 3:

$$\begin{aligned} COV(Y, f(X^*)) &= E(Yf(X+U)) - E(Y)E(f(X+U)) \\ &\doteq E(Y(f(X) + f'(X)U)) - E(Y)E(f(X) + f'(X)U) \\ &= E(Y(f(X)) - E(Y)E(f(X)) \\ &= COV(Y, f(X)). \end{aligned}$$

As we underlined at the beginning of Section 3.4, the obtained results hold approximately, since we have to perform a linear approximation in order to separate the variability due to the intrinsic scatter from the one due to measurement error. Furthermore, The correctness of using a linear approximation for the $f(X)$ function in the specific case of Hubble's data is validated through the simulation study performed in Chapter 4.

Using the just proved equations we can express the regression parameters β_0 and β_1 in terms of the expected moments of the observed data:

$$\beta_1 = \frac{COV(Y, f(X^*))}{V(f(X^*)) - E(f'(X^*)^2)E(U^2)} \tag{B.4}$$

$$\beta_0 = E(Y) - \beta_1 E(f(X^*)). \tag{B.5}$$

Thus, considering the sample moments of the observed data, we suggest the following extension for the BCES estimator in case of non-linear measurement error:

$$\hat{\beta}_1^{BCES} = \frac{\sum_{i=1}^{n}(f(x_i^*) - \overline{f(x^*)})(y_i - \overline{y})}{\sum_{i=1}^{n}(f(x_i)^* - \overline{f(x)^*})^2 - \sum_{i=1}^{n} f'(x_i^*)^2 \sum_{i=1}^{n}(u_i)^2} \tag{B.6}$$

$$\hat{\beta}_0^{BCES} = \overline{y} - \hat{\beta}_1^{BCES}\overline{f(X^*)}. \tag{B.7}$$

Likewise the original BCES approach, the variances of the estimators in Equations (B.6) and (B.7) are calculated by first defining the quantities

$$\xi_1 = \frac{(f(X^*) - E(f(X^*)))(Y - \beta_1 f(X^*) - \beta_0) + \beta_1 f'(X)^2 U^2}{V(f(X)^*) - E(f'(X)^2)E(U^2)} \tag{B.8}$$

$$\xi_2 = Y - \beta_1 f(X)^* - E(f(X)^*)\xi_1 \tag{B.9}$$

and then computing

$$\hat{\sigma}^2_{\beta_1} = \frac{1}{n^2}\sum_{i=1}^{n}(\hat{\xi}_{1i} - \bar{\hat{\xi}}_1)^2 \tag{B.10}$$

$$\hat{\sigma}^2_{\beta_0} = \frac{1}{n^2}\sum_{i=1}^{n}(\hat{\xi}_{2i} - \bar{\hat{\xi}}_2)^2. \tag{B.11}$$

$\bar{\hat{\xi}}_1$ and $\bar{\hat{\xi}}_2$ denote the arithmetic average of $\hat{\xi}_1$ and $\hat{\xi}_2$ obtained by replacing the unknown moments of Equations (B.8) and (B.9) by the sample moments obtained from the data.

As our new method was proposed and validated only with respect to the specific model analysed in Chapter 3. More applications considering different types of non-linear functions $f(X)$ must be inspected in order to assess its general applicability and effectiveness.

Appendix C

R Codes

C.1 Regression-Calibration

```
n=100
x <- rnorm(n,0,2)
u <- rnorm(n,0,2)
w <- x+u
y <-x +rnorm(n,0,1)
data<-data.frame(y,x,w)
true.model <- lm(y~x,data=data) # True model
naive.model <- lm(y~w, x=TRUE, data=data) # Naive model
plot(x,y)
plot(w,y)
# Estimate x.star via internal validation data
val.data<-data[sample(nrow(data), 0.1*nrow(data)),]
val.model<-lm(x~w,data=val.data)
x.star<-val.model$coefficient[1]+val.model$coefficient[2]*data$w
reg.calibration<-lm(y~x.star,x=TRUE)
plot(x,y)
abline(true.model,col="darkblue")
abline(reg.calibration,col ="red")
abline(naive.model,col = "green")
legend(min(x),max(y),legend=c("True Model","Regression Calibration","Naive Model")
   , col = c("darkblue","red","green"),lty=1)

# Adjust the resulting standard erros to account for the estimation of x.star,
# using bootstrap method
library(boot)
formula<-y~x.star
boot.reg.cal <- function(formula, databoot, indices){
   data <- databoot[indices,] # select obs. in bootstrap sample
   fit<-lm(formula,x=TRUE, data=data)
   coefficients(fit) # return coefficient vector
}
sd.boot <- boot(data=data, statistic=boot.reg.cal, R=2000, formula=y~x.star )
sd.boot
plot(sd.boot)

# Adjust the resulting standard erros to account for the estimation of x.star,
# using jackknife method
library(bootstrap)
DF<-data.frame(y,x.star)
model.lm <- formula(y ~ x.star)
theta <- function(x, xdata, coefficient){
   coef(lm(model.lm, data=xdata[x,]))[coefficient]
}
jackknife.apply <- function(x, xdata, coefs){
   sapply(coefs, function(coefficient)
   jackknife(x, theta, xdata=xdata, coefficient=coefficient), simplify=F)
}
results <- jackknife.apply(1:length(x.star), DF, c("(Intercept)", "x.star"))
results
```

C.2 BCES

```
## Notation:
## x1 true covariate
## y1 measured covariate
## x2 true response
## y2 measured response
## variance of e1 and e2 and their covariance are assumed known
n=100
x1 <- rnorm(n,0,2)
x2 <- 7+ 2*x1+rnorm(n,0,1)
e1<-rnorm(n,0,2)
e2<-rnorm(n,0,2)
V<-matrix(c(var(e1)*(length(e1)-1)/length(e1),cov(e1,e2),
   cov(e2,e1),var(e2)*(length(e2)-1)/length(e2)),2,2)
true.model <- lm(x2~x1) # True model
y1 <- x1+e1
y2 <- x2+e2
plot(y1,y2)
naive.model <- lm(y2~y1, x=TRUE)  # Naive model
# BCES estimator
beta1BCES=(cov(y1,y2)-V[1,2])/(var(y1)*(length(y1)-1)/length(y1)-V[1,1])
beta0BCES=mean(y2)-beta1BCES*mean(y1)
zeta1=((y1-mean(y1))*(y2-beta1BCES*y1-beta0BCES)+beta1BCES*V[1,1]-V[1,2])/
      (var(y1)*(length(y1)-1)/length(y1)-V[1,1])
zeta2=y2-beta1BCES*y1-mean(y1)*zeta1
```

```
# Beta1BCES variance estimation
varbeta1BCES=(var(zeta1)*(length(zeta1)-1)/length(zeta1))/n
# Beta0BCES variance estimation
varbeta0BCES=(var(zeta2)*(length(zeta2)-1)/length(zeta2))/n
varbeta1BCES
varbeta0BCES
BCES_fitted <- function(x) beta1BCES*x + beta0BCES
plot(x1,x2,main="Measurement Error correction")
abline(true.model, col="darkblue", lwd=2)
abline(naive.model, col="red", lwd=2)
curve(BCES_fitted, col="green", lwd=2, add=T)
legend("topleft",legend=c("True Model","Naive model","BCES Model") ,
      col = c("darkblue","red","green"),lty=1)
```

C.3 SIMEX

```
n=100
x <- rnorm(n,0,2)
u <- rnorm(n,0,2)
w <- x+u
y <-x +rnorm(n,0,1)
data_for_simulation<-data.frame(y,x,w)
true.model <- lm(y~x,data=data_for_simulation) # True model
naive.model <- lm(y~w, x=TRUE, data=data_for_simulation) #Naive model
sigma2u<-2
#Simulation step
B <- 100
sigma2u<-4 # Known measurement error variance
U=matrix(rnorm(B*length(w),0,sqrt(sigma2u)),B,length(w))
add.function<-function(lambda){
  w_bi=matrix(NA,B,length(w))
  for( k in 1:B ){
    for( i in 1:length(w)){
      w_bi[k,i]=w[i]+sqrt(lambda)*U[k,i]
    }
 }
 w_bi
}
w_bi0.5<-add.function(0.5) #lambda=0.5
w_bi1<-add.function(1)     #lambda=1
w_bi1.5<-add.function(1.5) #lambda=1.5
w_bi2<-add.function(2)     #lambda=2

# Estimate the coefficients (theta) for each lambda

est.theta.fun<-function(w.add){
  theta.k=matrix(NA,B,2)
  var.k <- matrix(NA, ncol=4, nrow=B)
  for( k in 1:B ){
    sim.model.k<-lm(y~w.add[k,],x=TRUE)
    theta.k[k,] <- c(coef(sim.model.k))
    var.k[k,1:2] <- vcov(sim.model.k)[1,]
    var.k[k,3:4] <- vcov(sim.model.k)[2,]
   }
  # Average of the B values of theta.K
  this.theta<- colMeans(theta.k)
  # SIMEX variance estimation, see appendix B.4.1 Carroll et al
  thetahat<-matrix(c(rep(this.theta[1],B),rep(this.theta[2],B)), B,2)
  deltab=theta.k-thetahat
  # ss function is used to calculate s^2_delta, see equation B.18 Carroll el al
  ss<-function(m){
    res<-matrix(NA,ncol(m),ncol(m))
    res<-crossprod(t(m[1,]))
    for(i in 2:nrow(m)){
     res<-res+m[i,]%*%t(m[i,])
    }
   return(res/(nrow(m)-1))
  }
 s2delta<-ss(deltab)
 # Variance component due to sampling variability
 tau2hat <- matrix(colMeans(var.k), ncol=2)
 # tau2hat-s2delta variance component due to measurement error variability
 return(list(this.theta, tau2hat-s2delta))
 }
sim.resultsw_bi0.5<-est.theta.fun(w_bi0.5) #lambda=0.5
sim.resultsw_bi1<-est.theta.fun(w_bi1)     #lambda=1
sim.resultsw_bi1.5<-est.theta.fun(w_bi1.5) #lambda=1.5
sim.resultsw_bi2<-est.theta.fun(w_bi2)     #lambda=2
# Vector that contains the simulated beta0 obtained in the previous step
beta0.sim<-c(naive.model$coefficient[1], sim.resultsw_bi0.5[[1]][1],
             sim.resultsw_bi1[[1]][1], sim.resultsw_bi1.5[[1]][1],
             sim.resultsw_bi2[[1]][1])
# Vector that contains the simulated beta1 obtained in the previous step
beta1.sim<-c(naive.model$coefficient[2], sim.resultsw_bi0.5[[1]][2],
             sim.resultsw_bi1[[1]][2], sim.resultsw_bi1.5[[1]][2],
             sim.resultsw_bi2[[1]][2])
var_beta0.sim<-c(vcov(naive.model)[1,1],sim.resultsw_bi0.5[[2]][1,1],
                 sim.resultsw_bi1[[2]][1,1], sim.resultsw_bi1.5[[2]][1,1],
                 sim.resultsw_bi2[[2]][1,1])
var_beta1.sim<-c(vcov(naive.model)[2,2],sim.resultsw_bi0.5[[2]][2,2],
                 sim.resultsw_bi1[[2]][2,2], sim.resultsw_bi1.5[[2]][2,2],
                 sim.resultsw_bi2[[2]][2,2])

#Extrapolation Step with quadratic extrapolant function
lambda <- c(0.0, 0.5, 1.0, 1.5, 2.0)
#SIMEX beta0,beta1
extr.fun.beta0<-lm(beta0.sim~lambda+I(lambda^2))
extr.fun.beta1<-lm(beta1.sim~lambda+I(lambda^2))
```

```
beta0.simex<-predict(extr.fun.beta0, newdata = data.frame(lambda = -1))
beta1.simex<-predict(extr.fun.beta1, newdata = data.frame(lambda = -1))
#SIMEX var(beta0), var(beta1)
extr.fun.var_beta0<-lm(var_beta0.sim~lambda+I(lambda^2))
extr.fun.var_beta1<-lm(var_beta1.sim~lambda+I(lambda^2))
var.beta0.simex<-predict(extr.fun.var_beta0, newdata = data.frame(lambda = -1))
var.beta1.simex<-predict(extr.fun.var_beta1, newdata = data.frame(lambda = -1))
#Simex plot for beta1
simex_function<- function(x) extr.fun.beta1$coefficient[3]*x^2 +
                 extr.fun.beta1$coefficient[2]*x + extr.fun.beta1$coefficient[1]
plot(lambda,beta1.sim,xlim=range(-2:3),ylim=range(0,1),main="SIMEX method")
curve(simex_function, col="darkblue", lwd=2, add=T )
points(-1,beta1.simex,pch=4)
#True, naive and SIMEX model's plot
simex_fitted <- function(s) beta1.simex*s + beta0.simex
plot(x,y)
abline(true.model,col="darkblue")
abline(naive.model,col = "red")
curve(simex_fitted, col="green", lwd=2, add=T)
legend("topleft",legend=c("True Model","Naive model","SIMEX Model") ,
      col = c("darkblue","red","green"),lty=1)
```

C.4 Heteroscedastic SIMEX

```
n=100
x <- rnorm(n,0,2)
u <- c(rep(NA,n))
sigma2ui<-rchisq(100,1)
for (i in 1:n){
 u[i]=rnorm(1,0,sqrt(sigma2ui[i]))
 }
w <- x+u
y <-x +rnorm(n,0,1)
data_for_simulation<-data.frame(y,x,w)
true.model <- lm(y~x,data=data_for_simulation) # True model
naive.model <- lm(y~w, x=TRUE, data=data_for_simulation) # Naive model
# Simulation step
B <- 100
# sigma2ui Known heteroscedastic measurement error variance
U<-matrix(NA,B,length(w))
for(i in 1:n){
 U[,i]=rnorm(B,0,sqrt(sigma2ui[i]))
 }
add.function<-function(lambda){
  w_bi=matrix(NA,B,length(w))
  for( k in 1:B ){
    for( i in 1:length(w)){
      w_bi[k,i]=w[i]+sqrt(lambda)*U[k,i]
    }
 }
 w_bi
}
w_bi0.5<-add.function(0.5) #lambda=0.5
w_bi1<-add.function(1)     #lambda=1
w_bi1.5<-add.function(1.5) #lambda=1.5
w_bi2<-add.function(2)     #lambda=2

# Estimate the coefficients (theta) for each lambda

est.theta.fun<-function(w.add){
  theta.k=matrix(NA,B,2)
  var.k <- matrix(NA, ncol=4, nrow=B)
  for( k in 1:B ){
    sim.model.k<-lm(y~w.add[k,],x=TRUE)
    theta.k[k,] <- c(coef(sim.model.k))
    var.k[k,1:2] <- vcov(sim.model.k)[1,]
    var.k[k,3:4] <- vcov(sim.model.k)[2,]
   }
  # Average of the B values of theta.K
  this.theta<- colMeans(theta.k)
  # SIMEX variance estimation, see appendix B.4.1 Carroll et al
  thetahat<-matrix(c(rep(this.theta[1],B),rep(this.theta[2],B)), B,2)
  deltab=theta.k-thetahat
  # ss function is used to calculate s^2_delta, see equation B.18 Carroll el al
  ss<-function(m){
    res<-matrix(NA,ncol(m),ncol(m))
    res<-crossprod(t(m[1,]))
    for(i in 2:nrow(m)){
     res<-res+m[i,]%*%t(m[i,])
    }
   return(res/(nrow(m)-1))
  }
 s2delta<-ss(deltab)
 # Variance component due to sampling variability
 tau2hat <- matrix(colMeans(var.k), ncol=2)
 # tau2hat-s2delta variance component due to measurement error variability
 return(list(this.theta, tau2hat-s2delta))
 }
sim.resultsw_bi0.5<-est.theta.fun(w_bi0.5) #lambda=0.5
sim.resultsw_bi1<-est.theta.fun(w_bi1)     #lambda=1
sim.resultsw_bi1.5<-est.theta.fun(w_bi1.5) #lambda=1.5
sim.resultsw_bi2<-est.theta.fun(w_bi2)     #lambda=2
# Vector that contains the simulated beta0 obtained in the previous step
beta0.sim<-c(naive.model$coefficient[1], sim.resultsw_bi0.5[[1]][1],
             sim.resultsw_bi1[[1]][1], sim.resultsw_bi1.5[[1]][1],
             sim.resultsw_bi2[[1]][1])
# Vector that contains the simulated beta1 obtained in the previous step
beta1.sim<-c(naive.model$coefficient[2], sim.resultsw_bi0.5[[1]][2],
          sim.resultsw_bi1[[1]][2], sim.resultsw_bi1.5[[1]][2],
          sim.resultsw_bi2[[1]][2])
var_beta0.sim<-c(vcov(naive.model)[1,1],sim.resultsw_bi0.5[[2]][1,1],
             sim.resultsw_bi1[[2]][1,1], sim.resultsw_bi1.5[[2]][1,1],
             sim.resultsw_bi2[[2]][1,1])
var_beta1.sim<-c(vcov(naive.model)[2,2],sim.resultsw_bi0.5[[2]][2,2],
             sim.resultsw_bi1[[2]][2,2], sim.resultsw_bi1.5[[2]][2,2],
             sim.resultsw_bi2[[2]][2,2])

#Extrapolation Step with quadratic extrapolant function
lambda <- c(0.0, 0.5, 1.0, 1.5, 2.0)
#SIMEX beta0,beta1
extr.fun.beta0<-lm(beta0.sim~lambda+I(lambda^2))
extr.fun.beta1<-lm(beta1.sim~lambda+I(lambda^2))
beta0.simex<-predict(extr.fun.beta0, newdata = data.frame(lambda = -1))
beta1.simex<-predict(extr.fun.beta1, newdata = data.frame(lambda = -1))
#SIMEX var(beta0), var(beta1)
extr.fun.var_beta0<-lm(var_beta0.sim~lambda+I(lambda^2))
extr.fun.var_beta1<-lm(var_beta1.sim~lambda+I(lambda^2))
var.beta0.simex<-predict(extr.fun.var_beta0, newdata = data.frame(lambda = -1))
var.beta1.simex<-predict(extr.fun.var_beta1, newdata = data.frame(lambda = -1))
#Simex plot for beta1
simex_function<- function(x) extr.fun.beta1$coefficient[3]*x^2 +
             extr.fun.beta1$coefficient[2]*x + extr.fun.beta1$coefficient[1]
plot(lambda,beta1.sim,xlim=range(-2:3),ylim=range(0,1),main="SIMEX method")
curve(simex_function, col="darkblue", lwd=2, add=T )
points(-1,beta1.simex,pch=4)
#True, naive and SIMEX model's plot
simex_fitted <- function(s) beta1.simex*s + beta0.simex
plot(x,y)
abline(true.model,col="darkblue")
abline(naive.model,col = "red")
curve(simex_fitted, col="green", lwd=2, add=T)
legend("topleft",legend=c("True Model","Naive model","SIMEX Model") ,
     col = c("darkblue","red","green"),lty=1)
```

C.5 Measurement error simulation

C.5.1 Models generation

```
################ NORMAL MEASUREMENT ERROR  u~norm(0,2^2)
n=100 #n=1000, n=10000 three different sample sizes
simfun<- function(n=100,a=7,b=2) { #n=1000, n=10000
  x <- rnorm(n,0,2)
  e<-rnorm(n,0,1)
  y <- a + b*x + e
  u <- rnorm(n,0,2)
```

```
   w <- x+u
   d=data.frame(x,y,w,u)
   # Validation data needed for RC model
   val.data<-d[sample(nrow(d), 0.1*nrow(d)), ]
   val.model<-lm(x~w,data=val.data)
   x.star<-val.model$coefficient[1]+
           val.model$coefficient[2]*d$w
   d<-cbind(d,x.star)
}
################ SKEW-NORMAL MEASUREMENT ERROR  u~skew-norm(0,2^2,5)
n=100 #n=1000, n=10000 three different sample sizes
simfun<- function(n=100,a=7,b=2) { #n=1000, n=10000
   library(stats4)
   library(sn)
   x <- rnorm(n,0,2)
   e<-rnorm(n,0,1)
   y <- a + b*x + e
   u <- rsn(n,0,2,5)
   w <- x+u
   d=data.frame(x,y,w,u)
   val.data<-d[sample(nrow(d), 0.1*nrow(d)), ]
   val.model<-lm(x~w,data=val.data)
   x.star<-val.model$coefficient[1]+val.model$coefficient[2]*d$w
   d<-cbind(d,x.star)
}
################ NORMAL MIXTURE MEASUREMENT ERROR  U=0.5*phi(u+2)+0.5*phi(u-4)
n=100 #n=1000, n=10000 three different sample sizes
simfun<- function(n=100,a=7,b=2) {   #n=1000, n=10000
   x <- rnorm(n,0,2)
   e<-rnorm(n,0,1)
   y <- a + b*x + e
   unif =runif(n)
   u = rep(0,n)
      for(i in 1:n){
          if(unif[i]<.5){
        u[i] = rnorm(1,-2,1)
      }else {
        u[i] = rnorm(1,4,1)
    }
   }
   w <- x+u
   d=data.frame(x,y,w,u)
   val.data<-d[sample(nrow(d), 0.1*nrow(d)), ] # It is needed for RC model
   val.model<-lm(x~w,data=val.data)
   x.star<-val.model$coefficient[1]+val.model$coefficient[2]*d$w
   d<-cbind(d,x.star)
}
```

C.5.2 Results collection

```
library(xtable)
library(bootstrap)
library(plyr)
error_correction_fun <- function(d) {
   true=coef(lm(y˜x,data=d)) # True parameters estimation
   true_sd=c(sqrt(vcov(lm(y˜x,data=d))[1,1]),
           sqrt(vcov(lm(y˜x,data=d))[2,2]))
   naive=coef(lm(y˜w,data=d)) # Naive parameters estimation
   naive_sd=c(sqrt(vcov(lm(y˜w,data=d))[1,1]),
            sqrt(vcov(lm(y˜w,data=d))[2,2]))
   reg.cal=coef(lm(y˜x.star,data=d)) # RC parameters estimation

   model.lm <- formula(y ˜ x.star, data=d)
   theta <- function(x, xdata, coefficient){
              coef(lm(model.lm, data=xdata[x,]))[coefficient]
              }
   jackknife.apply <- function(x, xdata, coefs){
       sapply(coefs,
       function(coefficient) jackknife(x, theta, xdata=xdata,
                                       coefficient=coefficient),
        simplify=F)
        }
   results <- jackknife.apply(1:length(d$x.star), d, c("(Intercept)", "x.star"))
   re.cal_sd=c(results$'(Intercept)'$jack.se,results$'x.star'$jack.se)
   # BCES parameters estimation
   V<-matrix(c(var(d$u)*(length(d$u)-1)/length(d$u),0,0,0),2,2)
   beta1BCES=(cov(d$w,d$y)-V[1,2])/(var(d$w)*(length(d$w)-1)/length(d$w)-V[1,1])
   beta0BCES=mean(d$y)-beta1BCES*mean(d$w)
   BCES=cbind(beta0BCES,beta1BCES)
   zeta1=((d$w-mean(d$w))*(d$y-beta1BCES*d$w-beta0BCES)+beta1BCES*V[1,1]-V[1,2])/
         (var(d$w)*(length(d$w)-1)/length(d$w)-V[1,1])
   zeta2=d$y-beta1BCES*d$w-mean(d$w)*zeta1
   varbeta1BCES=(var(zeta1)*(length(zeta1)-1)/length(zeta1))/length(d$y)
   varbeta0BCES=(var(zeta2)*(length(zeta2)-1)/length(zeta2))/length(d$y)
   BCES_sd=c(sqrt(varbeta0BCES),sqrt(varbeta1BCES))
   B=1000 # SIMEX parameters estimation
   sigma2u=2ˆ2
   U=matrix(rnorm(B*length(d$w),0,sqrt(sigma2u)),B,length(d$w))
   add.function<-function(lambda){
      w_bi=matrix(NA,B,length(d$w))
      for( k in 1:B ){
        for( i in 1:length(d$w)){
         w_bi[k,i]=d$w[i]+sqrt(lambda)*U[k,i]
       }
     }
    w_bi
   }
   w_bi0<-add.function(0) #lambda=0
```

```
w_bi0.5<-add.function(0.5) #lambda=0.5
w_bi1<-add.function(1)     #lambda=1
w_bi1.5<-add.function(1.5) #lambda=1.5
w_bi2<-add.function(2)     #lambda=2
est.theta.fun<-function(w.add){
  theta.k=matrix(NA,B,2)
  var.k <- matrix(NA, ncol=4, nrow=B)
  for( k in 1:B ){
    sim.model.k<-lm(d$y~w.add[k,],x=TRUE)
    theta.k[k,] <- c(coef(sim.model.k))
    var.k[k,1:2] <- vcov(sim.model.k)[1,]
    var.k[k,3:4] <- vcov(sim.model.k)[2,]
   }
  this.theta<- colMeans(theta.k)
  thetahat<-matrix(c(rep(this.theta[1],B),rep(this.theta[2],B)), B,2)
  deltab=theta.k-thetahat
  ss<-function(m){
     res<-matrix(NA,ncol(m),ncol(m))
     res<-crossprod(t(m[1,]))
     for(i in 2:nrow(m)){
        res<-res+m[i,]%*%t(m[i,])
       }
      return(res/(nrow(m)-1))
     }
    s2delta<-ss(deltab)
    tau2hat <- matrix( colMeans(var.k), ncol=2)
    return(list(this.theta, tau2hat-s2delta))
   }
sim.resultsw_bi0<-est.theta.fun(w_bi0)    #lambda=0
sim.resultsw_bi0.5<-est.theta.fun(w_bi0.5) #lambda=0.5
sim.resultsw_bi1<-est.theta.fun(w_bi1)     #lambda=1
sim.resultsw_bi1.5<-est.theta.fun(w_bi1.5) #lambda=1.5
sim.resultsw_bi2<-est.theta.fun(w_bi2)     #lambda=2
beta0.sim<-c(sim.resultsw_bi0[[1]][1], sim.resultsw_bi0.5[[1]][1],
           sim.resultsw_bi1[[1]][1],sim.resultsw_bi1.5[[1]][1],
           sim.resultsw_bi2[[1]][1])
beta1.sim<-c(sim.resultsw_bi0[[1]][2], sim.resultsw_bi0.5[[1]][2],
           sim.resultsw_bi1[[1]][2],sim.resultsw_bi1.5[[1]][2],
           sim.resultsw_bi2[[1]][2])
var_beta0.sim<-c(sim.resultsw_bi0[[2]][1,1],sim.resultsw_bi0.5[[2]][1,1],
               sim.resultsw_bi1[[2]][1,1],sim.resultsw_bi1.5[[2]][1,1],
               sim.resultsw_bi2[[2]][1,1])

var_beta1.sim<-c(sim.resultsw_bi0[[2]][2,2],sim.resultsw_bi0.5[[2]][2,2],
               sim.resultsw_bi1[[2]][2,2],sim.resultsw_bi1.5[[2]][2,2],
               sim.resultsw_bi2[[2]][2,2])
lambda <- c(0.0, 0.5, 1.0, 1.5, 2.0)
extr.fun.beta0<-lm(beta0.sim~lambda+I(lambda^2))
extr.fun.beta1<-lm(beta1.sim~lambda+I(lambda^2))
beta0.simex<-predict(extr.fun.beta0, newdata = data.frame(lambda = -1))
```

```
   beta1.simex<-predict(extr.fun.beta1, newdata = data.frame(lambda = -1))
   extr.fun.var_beta0<-lm(var_beta0.sim~lambda+I(lambda^2))
   extr.fun.var_beta1<-lm(var_beta1.sim~lambda+I(lambda^2))
   var.beta0.simex<-predict(extr.fun.var_beta0, newdata = data.frame(lambda = -1))
   var.beta1.simex<-predict(extr.fun.var_beta1, newdata = data.frame(lambda = -1))
   simex=cbind(beta0.simex,beta1.simex)
   simex_sd=c(sqrt(var.beta0.simex),sqrt(var.beta1.simex))
   return(list(true,naive,reg.cal,BCES,simex,true_sd,
          naive_sd,re.cal_sd,BCES_sd,simex_sd))
}
a=7 # True Intercept
b=2 # True slope
nsim=1000
# Replication of the simulated framework
sim_results<-raply(nsim,error_correction_fun (simfun()))
#Results organization
summary.measures=function(i){
   res=matrix(NA,nsim,2)
   for(k in 1:nsim){
     res[k,]=c(sim_results[,i][[k]][1],sim_results[,i][[k]][2])
     }
   list(Mean=colMeans(res),Median=c(median(res[,1]),median(res[,2])),
        Standard_Deviation=c(sd(res[,1]),sd(res[,2])),
        bias=c(a-mean(res[,1]),b-mean(res[,2])),
        Interquartile_Range=c(IQR(res[,1]),IQR(res[,2])),
        MSE=c(mean(res[,1]-a)^2,mean(res[,2]-b)^2))
}
true_summary.measures<- summary.measures(1)
naive_summary.measures<-summary.measures(2)
re.cal_summary.measures<-summary.measures(3)
BCES_summary.measures<-summary.measures(4)
simex_summary.measures<-summary.measures(5)
true_sd_summary.measures<- summary.measures(6)
naive_sd_summary.measures<-summary.measures(7)
re.cal_sd_summary.measures<-summary.measures(8)
BCES_sd_summary.measures<-summary.measures(9)
simex_sd_summary.measures<-summary.measures(10)
# Table 1 creation for each estimator
#ct1<-c("Mean","Median","St.Deviation","Bias","Int.Range","MSE")
ct1_true_beta0=c(true_summary.measures$Mean[1],
               true_summary.measures$Median[1],
               true_summary.measures$Standard_Deviation[1],
               true_summary.measures$bias[1],
               true_summary.measures$Interquartile_Range[1],
               true_summary.measures$MSE[1])
ct1_true_beta1=c(true_summary.measures$Mean[2],
               true_summary.measures$Median[2],
               true_summary.measures$Standard_Deviation[2],
               true_summary.measures$bias[2],
               true_summary.measures$Interquartile_Range[2],
               true_summary.measures$MSE[2])
ct1_naive_beta0=c(naive_summary.measures$Mean[1],
               naive_summary.measures$Median[1],
               naive_summary.measures$Standard_Deviation[1],
               naive_summary.measures$bias[1],
               naive_summary.measures$Interquartile_Range[1],
               naive_summary.measures$MSE[1])
ct1_naive_beta1=c(naive_summary.measures$Mean[2],
               naive_summary.measures$Median[2],
               naive_summary.measures$Standard_Deviation[2],
               naive_summary.measures$bias[2],
               naive_summary.measures$Interquartile_Range[2],
               naive_summary.measures$MSE[2])
ct1_re.cal_beta0=c(re.cal_summary.measures$Mean[1],
               re.cal_summary.measures$Median[1],
               re.cal_summary.measures$Standard_Deviation[1],
               re.cal_summary.measures$bias[1],
               re.cal_summary.measures$Interquartile_Range[1],
               re.cal_summary.measures$MSE[1])
ct1_re.cal_beta1=c(re.cal_summary.measures$Mean[2],
               re.cal_summary.measures$Median[2],
               re.cal_summary.measures$Standard_Deviation[2],
               re.cal_summary.measures$bias[2],
               re.cal_summary.measures$Interquartile_Range[2],
               re.cal_summary.measures$MSE[2])
ct1_BCES_beta0=c(BCES_summary.measures$Mean[1],
               BCES_summary.measures$Median[1],
               BCES_summary.measures$Standard_Deviation[1],
               BCES_summary.measures$bias[1],
               BCES_summary.measures$Interquartile_Range[1],
               BCES_summary.measures$MSE[1])
ct1_BCES_beta1=c(BCES_summary.measures$Mean[2],
               BCES_summary.measures$Median[2],
               BCES_summary.measures$Standard_Deviation[2],
               BCES_summary.measures$bias[2],
               BCES_summary.measures$Interquartile_Range[2],
               BCES_summary.measures$MSE[2])
ct1_simex_beta0=c(simex_summary.measures$Mean[1],
               simex_summary.measures$Median[1],
               simex_summary.measures$Standard_Deviation[1],
               simex_summary.measures$bias[1],
               simex_summary.measures$Interquartile_Range[1],
               simex_summary.measures$MSE[1])
ct1_simex_beta1=c(simex_summary.measures$Mean[2],
               simex_summary.measures$Median[2],
               simex_summary.measures$Standard_Deviation[2],
               simex_summary.measures$bias[2],
               simex_summary.measures$Interquartile_Range[2],
               simex_summary.measures$MSE[2])
true_t1<-cbind(ct1_true_beta0,ct1_true_beta1)
```

```
naive_t1<-cbind(ct1_naive_beta0,ct1_naive_beta1)
re.cal_t1<-cbind(ct1_re.cal_beta0,ct1_re.cal_beta1)
BCES_t1<-cbind(ct1_BCES_beta0,ct1_BCES_beta1)
simex_t1<-cbind(ct1_simex_beta0,ct1_simex_beta1)
xtable(true_t1,floating=FALSE,digits=c(rep(4,3)))
xtable(naive_t1,floating=FALSE,digits=c(rep(4,3)))
xtable(re.cal_t1,floating=FALSE,digits=c(rep(4,3)))
xtable(BCES_t1,floating=FALSE,digits=c(rep(4,3)))
xtable(simex_t1,floating=FALSE,digits=c(rep(4,3)))
#Inference results
summary.inf=function(i,beta,true_beta){
 upper.limit=rep(NA,nsim)
 lower.limit=rep(NA,nsim)
 Conf_Interval=matrix(NA,nsim,2)
 real.alpha=rep(NA,nsim)
 dx.alpha=rep(NA,nsim)
 sx.alpha=rep(NA,nsim)
 wald.test=rep(NA,nsim)
 p_value=rep(NA,nsim)
 for(k in 1:nsim){
  upper.limit[k]=sim_results[,i][[k]][beta]+
                 qt(0.975,n)*sim_results[,(i+5)][[k]][beta]
  lower.limit[k]=sim_results[,i][[k]][beta]-
                 qt(0.975,n)*sim_results[,(i+5)][[k]][beta]
  Conf_Interval[k,]=c(lower.limit[k],upper.limit[k])
  real.alpha[k]=true_beta>Conf_Interval[k,1] && true_beta<Conf_Interval[k,2]
  dx.alpha[k]=true_beta<sim_results[,i][[k]][beta]+
                 qt(0.95,n)*sim_results[,(i+5)][[k]][beta]
  sx.alpha[k]=true_beta>sim_results[,i][[k]][beta]-
                 qt(0.95,n)*sim_results[,(i+5)][[k]][beta]
  wald.test[k]=(sim_results[,i][[k]][beta]-
                 true_beta)/sim_results[,(i+5)][[k]][beta]
  p_value[k]=2*min(pt(wald.test[k],n),1-pt(wald.test[k],n))
  }
  list(Mean_Conf_Interval=colMeans(Conf_Interval),
                H0_accepted=sum(real.alpha=="TRUE"),
                H0_rejected=sum(real.alpha=="FALSE"),
                Real_alpha=1-(sum(real.alpha=="TRUE")/
                                  length(real.alpha)),
  Dx_alpha=1-(sum(dx.alpha=="TRUE")/length(real.alpha)),
  Sx_alpha=1-(sum(sx.alpha=="TRUE")/length(real.alpha)),
  Mean_Interval_length=mean(Conf_Interval[,2]-
                       Conf_Interval[,1]),
  mean_Wald.test=mean(wald.test),mean_p_value=mean(p_value))
}
#Inference results organization
summary.inf_true.beta0=summary.inf(1,1,a)
summary.inf_true.beta1=summary.inf(1,2,b)
summary.inf_naive.beta0=summary.inf(2,1,a)
summary.inf_naive.beta1=summary.inf(2,2,b)
summary.inf_re.cal.beta0=summary.inf(3,1,a)
summary.inf_re.cal.beta1=summary.inf(3,2,b)
summary.inf_BCES.beta0=summary.inf(4,1,a)
summary.inf_BCES.beta1=summary.inf(4,2,b)
summary.inf_simex.beta0=summary.inf(5,1,a)
summary.inf_simex.beta1=summary.inf(5,2,b)
# Table 2 creation for each estimator
#ct3<-c("Real_alpha","Real_Dx_alpha","Real_Sx_alpha","H0_accepted","H0_rejected")
ct3_true_beta0=c(summary.inf_true.beta0$Real_alpha,
                 summary.inf_true.beta0$Dx_alpha,
                 summary.inf_true.beta0$Sx_alpha,
               summary.inf_true.beta0$Mean_Conf_Interval)
ct3_true_beta1=c(summary.inf_true.beta1$Real_alpha,
                 summary.inf_true.beta1$Dx_alpha,
                 summary.inf_true.beta1$Sx_alpha,
                summary.inf_true.beta1$Mean_Conf_Interval)
ct3_naive_beta0=c(summary.inf_naive.beta0$Real_alpha,
                  summary.inf_naive.beta0$Dx_alpha,
                  summary.inf_naive.beta0$Sx_alpha,
                  summary.inf_naive.beta0$Mean_Conf_Interval)
ct3_naive_beta1=c(summary.inf_naive.beta1$Real_alpha,
                  summary.inf_naive.beta1$Dx_alpha,
                  summary.inf_naive.beta1$Sx_alpha,
                  summary.inf_naive.beta1$Mean_Conf_Interval)
ct3_re.cal_beta0=c(summary.inf_re.cal.beta0$Real_alpha,
                   summary.inf_re.cal.beta0$Dx_alpha,
                   summary.inf_re.cal.beta0$Sx_alpha,
                   summary.inf_re.cal.beta0$Mean_Conf_Interval)
ct3_re.cal_beta1=c(summary.inf_re.cal.beta1$Real_alpha,
                   summary.inf_re.cal.beta1$Dx_alpha,
                   summary.inf_re.cal.beta1$Sx_alpha,
                   summary.inf_re.cal.beta1$Mean_Conf_Interval)
ct3_BCES_beta0=c(summary.inf_BCES.beta0$Real_alpha,
                 summary.inf_BCES.beta0$Dx_alpha,
                 summary.inf_BCES.beta0$Sx_alpha,
                 summary.inf_BCES.beta0$Mean_Conf_Interval)
ct3_BCES_beta1=c(summary.inf_BCES.beta1$Real_alpha,
                 summary.inf_BCES.beta1$Dx_alpha,
                 summary.inf_BCES.beta1$Sx_alpha,
                 summary.inf_BCES.beta1$Mean_Conf_Interval)
ct3_simex_beta0=c(summary.inf_simex.beta0$Real_alpha,
                  summary.inf_simex.beta0$Dx_alpha,
                  summary.inf_simex.beta0$Sx_alpha,
                  summary.inf_simex.beta0$Mean_Conf_Interval)
ct3_simex_beta1=c(summary.inf_simex.beta1$Real_alpha,
                  summary.inf_simex.beta1$Dx_alpha,
                  summary.inf_simex.beta1$Sx_alpha,
                  summary.inf_simex.beta1$Mean_Conf_Interval)
true_t3<-cbind(ct3_true_beta0,ct3_true_beta1)
naive_t3<-cbind(ct3_naive_beta0,ct3_naive_beta1)
```

```
re.cal_t3<-cbind(ct3_re.cal_beta0,ct3_re.cal_beta1)
BCES_t3<-cbind(ct3_BCES_beta0,ct3_BCES_beta1)
simex_t3<-cbind(ct3_simex_beta0,ct3_simex_beta1)
xtable(true_t3,floating=FALSE,digits=c(rep(2,3)))
xtable(naive_t3,floating=FALSE,digits=c(rep(2,3)))
xtable(re.cal_t3,floating=FALSE,digits=c(rep(2,3)))
xtable(BCES_t3,floating=FALSE,digits=c(rep(2,3)))
xtable(simex_t3,floating=FALSE,digits=c(rep(2,3)))
# Table 4 creation for each estimator
#Error_Model<-c("TRUE","NAIVE","RC","BCES","SIMEX")
beta0<-c(true_summary.measures[[2]][1],
         naive_summary.measures[[2]][1],
         re.cal_summary.measures[[2]][1],
         BCES_summary.measures[[2]][1],
         simex_summary.measures[[2]][1])
beta1<-c(true_summary.measures[[2]][2],
         naive_summary.measures[[2]][2],
         re.cal_summary.measures[[2]][2],
         BCES_summary.measures[[2]][2],
         simex_summary.measures[[2]][2])
sd_beta0<-c(true_sd_summary.measures[[2]][1],
            naive_sd_summary.measures[[2]][1],
            re.cal_sd_summary.measures[[2]][1],
            BCES_sd_summary.measures[[2]][1],
            simex_sd_summary.measures[[2]][1])
sd_beta1<-c(true_sd_summary.measures[[2]][2],
            naive_sd_summary.measures[[2]][2],
            re.cal_sd_summary.measures[[2]][2],
            BCES_sd_summary.measures[[2]][2],
            simex_sd_summary.measures[[2]][2])
t4<-cbind(beta0,sd_beta0,beta1,sd_beta1)
xtable(t4,floating=FALSE,digits=c(rep(4,5)))
xtable(t4,floating=FALSE,digits=c(rep(4,5)))
```

C.6 Hubble data analysis

```
## Hubble Data "The SBF Survey of Galaxy Distances"
hubbleData<-read.table(file.choose(),col.names=c("Galaxy","RA","Dec",
           "vCMB","morphT","Grp","AB","V-I","V-I_1",
           "V-I_2","mI","mI1","mI2","DM","DMerr",
           "r","r1","r2","Q","PD","NI"))
## Hubble Law: vCMB=H0*D
## vCMB=beta0 + beta1*D
library(fBasics)
library (boot)
library(xtable)
H0=6.9*10^(-5) # Most recent H0 estimation (Bennett, 2014)
# Explorative data analysis
```

```
HData<-hubbleData[,c("vCMB","DM","DMerr")]
xtable(head(HData),floating=FALSE,digits=c(rep(4,4)))
attach(HData)
basicStats(vCMB)
D<-10^(DM/5+1)
basicStats(D)
Derr<-10^(DMerr/5+1)
par(mfrow=c(1,2))
boxplot(D, main="Boxplot of D")
hist(D)
par(mfrow=c(1,2))
boxplot(vCMB, main="Boxplot of vCMB")
hist(vCMB)
cor(vCMB,D)
cor.test(vCMB,D)
plot(D,vCMB,main="SBF survey - Radial velocity vs Distance")
naive.model=lm(vCMB~D,data=HData)
summary(naive.model)
xtable(summary(naive.model))
naive.model.offset=lm(vCMB~D+offset(H0*D),data=HData)
summary(naive.model.offset)
xtable(summary(naive.model.offset))
confint(naive.model,2,.90)
confint(naive.model,2,.95)
confint(naive.model,2,.99)
plot(naive.model)
naive.model.glm=glm(vCMB~D,data=HData)
glm.diag.plots(naive.model.glm)

## Simex correction for measurement errors
B <- 1000
sigma2ui<-DMerr^2 #known heteroscedastic measurement error variance
U<-matrix(NA,B,length(D))
for(i in 1:length(D)){
 U[,i]=rnorm(B,0,sqrt(sigma2ui[i]))
 }
add.function<-function(lambda){
  w_bi=matrix(NA,B,length(D))
  for( k in 1:B ){
    for( i in 1:length(D)){
      w_bi[k,i]=10^((DM[i]+sqrt(lambda)*U[k,i])/5+1)
    }
 }
 w_bi
}
w_bi0.5<-add.function(0.5) #lambda=0.5
w_bi1<-add.function(1)     #lambda=1
w_bi1.5<-add.function(1.5) #lambda=1.5
w_bi2<-add.function(2)     #lambda=2
```

```
#Estimate the coefficients (theta) for each lambda

est.theta.fun<-function(w.add){
  theta.k=matrix(NA,B,2)
  var.k <- matrix(NA, ncol=4, nrow=B)
  for( k in 1:B ){
    sim.model.k<-lm(vCMB~w.add[k,],x=TRUE)
    theta.k[k,] <- c(coef(sim.model.k))
    var.k[k,1:2] <- vcov(sim.model.k)[1,]
    var.k[k,3:4] <- vcov(sim.model.k)[2,]
   }
  # average of the B values of theta.K
  this.theta<- colMeans(theta.k)
  thetahat<- matrix(c(rep(this.theta[1],B),rep(this.theta[2],B)), B,2)
  deltab=theta.k-thetahat
  ss<-function(m){
    res<-matrix(NA,ncol(m),ncol(m))
    res<-crossprod(t(m[1,]))
    for(i in 2:nrow(m)){
     res<-res+m[i,]%*%t(m[i,])
    }
   return(res/(nrow(m)-1))
  }
 s2delta<-ss(deltab)
 tau2hat <- matrix(colMeans(var.k), ncol=2)
 return(list(this.theta, tau2hat-s2delta))
 }
sim.resultsw_bi0.5<-est.theta.fun(w_bi0.5) #lambda=0.5
sim.resultsw_bi1<-est.theta.fun(w_bi1)     #lambda=1
sim.resultsw_bi1.5<-est.theta.fun(w_bi1.5) #lambda=1.5
sim.resultsw_bi2<-est.theta.fun(w_bi2)     #lambda=2
beta0.sim<-c(naive.model$coefficient[1], sim.resultsw_bi0.5[[1]][1],
            sim.resultsw_bi1[[1]][1], sim.resultsw_bi1.5[[1]][1],
             sim.resultsw_bi2[[1]][1])
beta1.sim<-c(naive.model$coefficient[2], sim.resultsw_bi0.5[[1]][2],
            sim.resultsw_bi1[[1]][2], sim.resultsw_bi1.5[[1]][2],
             sim.resultsw_bi2[[1]][2])
var_beta0.sim<-c(vcov(naive.model)[1,1],sim.resultsw_bi0.5[[2]][1,1],
             sim.resultsw_bi1[[2]][1,1], sim.resultsw_bi1.5[[2]][1,1],
             sim.resultsw_bi2[[2]][1,1])
var_beta1.sim<-c(vcov(naive.model)[2,2],sim.resultsw_bi0.5[[2]][2,2],
             sim.resultsw_bi1[[2]][2,2], sim.resultsw_bi1.5[[2]][2,2],
             sim.resultsw_bi2[[2]][2,2])

#Extrapolation Step
lambda <- c(0.0, 0.5, 1.0, 1.5, 2.0)
#SIMEX beta0,beta1
extr.fun.beta0<-lm(beta0.sim~lambda+I(lambda^2)+I(lambda^3))
extr.fun.beta1<-lm(beta1.sim~lambda+I(lambda^2)+I(lambda^3))
summary(extr.fun.beta0)
summary(extr.fun.beta1)
beta0.simex<-predict(extr.fun.beta0, newdata = data.frame(lambda = -1))
beta1.simex<-predict(extr.fun.beta1, newdata = data.frame(lambda = -1))
beta0.simex
beta1.simex
#SIMEX var(beta0), var(beta1)
extr.fun.var_beta0<-lm(var_beta0.sim~lambda+I(lambda^2)+I(lambda^3))
extr.fun.var_beta1<-lm(var_beta1.sim~lambda+I(lambda^2)+I(lambda^3))
summary(extr.fun.var_beta0)
summary(extr.fun.var_beta1)
var.beta0.simex<-predict(extr.fun.var_beta0, newdata = data.frame(lambda = -1))
var.beta1.simex<-predict(extr.fun.var_beta1, newdata = data.frame(lambda = -1))
var.beta0.simex
var.beta1.simex
#Simex plot for beta1
simex_function<- function(x) extr.fun.beta1$coefficient[3]*x^2
               +extr.fun.beta1$coefficient[2]*x + extr.fun.beta1$coefficient[1]
plot(lambda,beta1.sim,main="SIMEX method correction for Hubble’s costant",
       xlim=range(-1.5:3),ylim=range(4*10^(-5),10*10^(-5)), ylab="H_0",pch=16,col="red")
curve(simex_function, col="darkblue", lwd=2, add=T )
points(-1,beta1.simex,pch=4,col="red")

# Collection of SIMEX results
SIMEX_HD=c(beta0.simex, beta1.simex)
SIMEX_HD_sd=c(sqrt(var.beta0.simex),sqrt(var.beta1.simex))
zvalueSIMEX=SIMEX_HD/SIMEX_HD_sd
pvalueSIMEXbeta0=2*min(pnorm(zvalueSIMEX[1]),1-pnorm(zvalueSIMEX[1]))
pvalueSIMEXbeta1=2*min(pnorm(zvalueSIMEX[2]),1-pnorm(zvalueSIMEX[2]))

beta1SIMEX_inf_0.9=beta1.simex-qnorm(.95)*sqrt(var.beta1.simex)
beta1SIMEX_sup_0.9=beta1.simex+qnorm(.95)*sqrt(var.beta1.simex)
c(beta1SIMEX_inf_0.9,beta1SIMEX_sup_0.9)
beta1SIMEX_inf_0.95=beta1.simex-qnorm(.975)*sqrt(var.beta1.simex)
beta1SIMEX_sup_0.95=beta1.simex+qnorm(.975)*sqrt(var.beta1.simex)
c(beta1SIMEX_inf_0.95,beta1SIMEX_sup_0.95)
beta1SIMEX_inf_0.99=beta1.simex-qnorm(.995)*sqrt(var.beta1.simex)
beta1SIMEX_sup_0.99=beta1.simex+qnorm(.995)*sqrt(var.beta1.simex)
c(beta1SIMEX_inf_0.99,beta1SIMEX_sup_0.99)

#H_0: beta1=H0
z_SIMEX.H0=(beta1.simex-H0)/(sqrt(var.beta1.simex))
z_SIMEX.H0
p_SIMEX.H0=2*min(pnorm(z_SIMEX.H0),1-pnorm(z_SIMEX.H0))
p_SIMEX.H0
#H_0: beta1=naive.model$coef[2]
z_SIMEX.naive=(beta1.simex-naive.model$coef[2])/(sqrt(var.beta1.simex))
z_SIMEX.naive
p_SIMEX.naive=2*min(pnorm(z_SIMEX.naive),1-pnorm(z_SIMEX.naive))
p_SIMEX.naive
```

```
## BCES estimator
dDM=(log(10)*10^(DM/5+1))/5
varfx=mean(D^2)-mean(dDM^2)*mean(DMerr^2)-mean(D)^2
V_i=(dDM^2)*(DMerr^2)
beta1BCES=cov(vCMB,D)/varfx # non-linear BCES estimation for beta_1
beta0BCES=mean(vCMB)-beta1BCES*mean(D) # non-linear BCES estimation for beta_0
zeta1=((D-mean(D))*(vCMB-beta1BCES*D-beta0BCES)+beta1BCES*V_i)/
      ((var(D)*(length(D)-1)/length(D))-mean(V_i))
zeta2=vCMB-beta1BCES*D-mean(D)*zeta1
# Beta1BCES variance estimation
varbeta1BCES=(var(zeta1)*(length(zeta1)-1)/length(zeta1))/length(D)
# Beta0BCES variance estimation
varbeta0BCES=(var(zeta2)*(length(zeta2)-1)/length(zeta2))/length(D)

# Collection of BCES results
BCES_HD=c(beta0BCES, beta1BCES)
BCES_HD_sd=c(sqrt(varbeta0BCES),sqrt(varbeta1BCES))
zvalueBCES=BCES_HD/BCES_HD_sd
pvalueBCESbeta0=2*min(pnorm(zvalueBCES[1]),1-pnorm(zvalueBCES[1]))
pvalueBCESbeta1=2*min(pnorm(zvalueBCES[2]),1-pnorm(zvalueBCES[2]))

beta1BCES_inf_0.9=beta1BCES-qnorm(.95)*sqrt(varbeta1BCES)
beta1BCES_sup_0.9=beta1BCES+qnorm(.95)*sqrt(varbeta1BCES)
c(beta1BCES_inf_0.9,beta1BCES_sup_0.9)
beta1BCES_inf_0.95=beta1BCES-qnorm(.975)*sqrt(varbeta1BCES)
beta1BCES_sup_0.95=beta1BCES+qnorm(.975)*sqrt(varbeta1BCES)
c(beta1BCES_inf_0.95,beta1BCES_sup_0.95)
beta1BCES_inf_0.99=beta1BCES-qnorm(.995)*sqrt(varbeta1BCES)
beta1BCES_sup_0.99=beta1BCES+qnorm(.995)*sqrt(varbeta1BCES)
c(beta1BCES_inf_0.99,beta1BCES_sup_0.99)

BCES_fitted <- function(x) beta1BCES*x + beta0BCES

#H_0: beta1=H0
z_BCES.H0=(beta1BCES-H0)/(sqrt(varbeta1BCES))
z_BCES.H0
p_BCES.H0=2*min(pnorm(z_BCES.H0),1-pnorm(z_BCES.H0))
p_BCES.H0

#H_0: beta1=naive.model$coef[2]
z_BCES.naive=(beta1BCES-naive.model$coef[2])/(sqrt(varbeta1BCES))
z_BCES.naive
p_BCES.naive=2*min(pnorm(z_BCES.naive),1-pnorm(z_BCES.naive))
p_BCES.naive


## SIMEX vs BCES
#H_0= beta1=beta1BCES
z_SIMEX.bces=(beta1.simex-beta1BCES)/(sqrt(var.beta1.simex))
z_SIMEX.bces
```

```
p_SIMEX.bces=2*min(pnorm(z_SIMEX.bces),1-pnorm(z_SIMEX.bces))
p_SIMEX.bces
#H_0= beta1=beta1SIMEX
z_BCES.simex=(beta1BCES-beta1.simex)/(sqrt(varbeta1BCES))
z_BCES.simex
p_BCES.simex=2*min(pnorm(z_BCES.simex),1-pnorm(z_BCES.simex))
p_BCES.simex

## Naive, BCES and SIMEX model's plot
simex_fitted <- function(s) beta1.simex*s +beta0.simex
plot(D,vCMB, main="Measurement error correction for the Hubble Data model")
abline(naive.model,col = "red")
curve(simex_fitted, col="green", lwd=2, add=T)
curve(BCES_fitted, col="purple", lwd=2, add=T)
legend("topleft",legend=c("Naive model","SIMEX Model","BCES Model") ,
                  col = c("red","green","purple"),lty=1)
list(coef(naive.model), c(beta0.simex,beta1.simex), c(beta0BCES, beta1BCES))
```

C.7 Non-linear BCES simulation

```
## y=beta0 + beta1*10^(x/5+1)+epsilon
library(xtable)
n=300 #sample size
b=2
a=0
simfun<- function(n=300,a=0,b=2) {
   x=rnorm(n, 30, 1.3)
   fx=10^(x/5+1)
   e<-rnorm(n,0,1)
   y <- a + b*fx + e
   u <- c(rep(NA,n))
   sigma2ui<-rchisq(n,0.5)
   for (i in 1:n){
       u[i]=rnorm(1,0,sqrt(sigma2ui[i]))
      }
   w <- x+u
   fw=10^(w/5+1)
   d=data.frame(x,y,w,u,fw,fx)
}
error_correction_fun <- function(d) {
   true=coef(lm(y~fx,data=d)) # True parameters estimation
   true_sd=c(sqrt(vcov(lm(y~fx,data=d))[1,1]),sqrt(vcov(lm(y~fx,data=d))[2,2]))
   naive=coef(lm(y~fw,data=d)) # Naive parameters estimation
   naive_sd=c(sqrt(vcov(lm(y~fw,data=d))[1,1]),sqrt(vcov(lm(y~fw,data=d))[2,2]))
   dfx=(log(10)/5*10^(d$x/5+1))
   dfw=(log(10)/5*10^(d$w/5+1))
   V_i=(dfw^2)*(d$u^2)
varfx=mean(d$fw^2)-mean(d$fw)^2-mean(dfw^2)*mean(d$u^2)
beta1BCES=cov(d$y,d$fw)/varfx
beta0BCES=mean(d$y)-beta1BCES*mean(d$fw)
zeta1=((d$fw-mean(d$fw))*(d$y-beta1BCES*d$fw-beta0BCES)+beta1BCES*V_i)/
      ((var(d$fw)*(length(d$fw)-1)/length(d$fw))-mean(V_i))
zeta2=d$y-beta1BCES*d$fw-mean(d$fw)*zeta1
# Beta1BCES variance estimation
varbeta1BCES=(var(zeta1)*(length(zeta1)-1)/length(zeta1))/length(d$fw)
# Beta0BCES variance estimation
varbeta0BCES=(var(zeta2)*(length(zeta2)-1)/length(zeta2))/length(d$fw)
BCES=cbind(beta0BCES,beta1BCES)
 BCES_sd=c(sqrt(varbeta0BCES),sqrt(varbeta1BCES))
 return(list(true,naive,BCES,true_sd,naive_sd,BCES_sd))
}
library(plyr)
nsim=1000
sim_results<-raply(nsim,error_correction_fun (simfun()))
beta1BCESal_1=rep(NA,nsim)
for(k in 1:nsim){
     beta1BCESal_1[k]=c(sim_results[,3][[k]][2])
     }
beta1BCESal_1
boxplot(beta1BCESal_1,main="Boxplot of Beta1 BCES")
qqnorm(beta1BCESal_1)
qqline(beta1BCESal_1)
stan.beta1BCESal_1=(beta1BCESal_1-mean(beta1BCESal_1))/sd(beta1BCESal_1)
stan.h=hist(stan.beta1BCESal_1,breaks=30,col="grey",
       xlab="Standrardized Beta1 BCES",
       main="Histogram of Standardized Beta1 BCES")
xfit<-seq(min(stan.beta1BCESal_1),max(stan.beta1BCESal_1),length=40)
yfit<-dnorm(xfit)
yfit <- yfit*diff(h$mids[1:2])*length(beta1BCESal_1)
lines(xfit, yfit, col="red", lwd=2)
summary.measures=function(i){
   res=matrix(NA,nsim,2)
   for(k in 1:nsim){
     res[k,]=c(sim_results[,i][[k]][1],sim_results[,i][[k]][2])
     }
   list(Mean=colMeans(res),Median=c(median(res[,1]),median(res[,2])),
        Standard_Deviation=c(sd(res[,1]),sd(res[,2])),
        bias=c(a-mean(res[,1]),b-mean(res[,2])),
        Interquartile_Range=c(IQR(res[,1]),IQR(res[,2])),
        MSE=c(mean(res[,1]-a)^2,mean(res[,2]-b)^2))
}
true_summary.measures<- summary.measures(1)
naive_summary.measures<-summary.measures(2)
BCES_summary.measures<-summary.measures(3)
true_sd_summary.measures<- summary.measures(4)
naive_sd_summary.measures<-summary.measures(5)
BCES_sd_summary.measures<-summary.measures(6)
# Table 1 creation for each estimator
```

```
#ct1<-c("Mean","Median","St.Deviation","Bias","Int.Range","MSE")
ct1_true_beta1=c(true_summary.measures$Mean[2],
               true_summary.measures$Median[2],
               true_summary.measures$Standard_Deviation[2],
               true_summary.measures$bias[2],
               true_summary.measures$Interquartile_Range[2],
               true_summary.measures$MSE[2])
ct1_naive_beta1=c(naive_summary.measures$Mean[2],
               naive_summary.measures$Median[2],
               naive_summary.measures$Standard_Deviation[2],
               naive_summary.measures$bias[2],
               naive_summary.measures$Interquartile_Range[2],
               naive_summary.measures$MSE[2])
ct1_BCES_beta1=c(BCES_summary.measures$Mean[2],
               BCES_summary.measures$Median[2],
               BCES_summary.measures$Standard_Deviation[2],
               BCES_summary.measures$bias[2],
               BCES_summary.measures$Interquartile_Range[2],
               BCES_summary.measures$MSE[2])
## Inference results 1-alpha=.90 #1-alpha=.95 1-alpha=.99
summary.inf=function(i,beta,true_beta){
 upper.limit=rep(NA,nsim)
 lower.limit=rep(NA,nsim)
 Conf_Interval=matrix(NA,nsim,2)
 real.alpha=rep(NA,nsim)
 dx.alpha=rep(NA,nsim)
 sx.alpha=rep(NA,nsim)
 wald.test=rep(NA,nsim)
 p_value=rep(NA,nsim)
 for(k in 1:nsim){
  upper.limit[k]=sim_results[,i][[k]][beta]+
                 qnorm(0.975,n)*sim_results[,(i+5)][[k]][beta]
  lower.limit[k]=sim_results[,i][[k]][beta]-
                 qnorm(0.975,n)*sim_results[,(i+5)][[k]][beta]
  Conf_Interval[k,]=c(lower.limit[k],upper.limit[k])
  real.alpha[k]=true_beta>Conf_Interval[k,1] && true_beta<Conf_Interval[k,2]
  dx.alpha[k]=true_beta<sim_results[,i][[k]][beta]+
                 qnorm(0.95,n)*sim_results[,(i+5)][[k]][beta]
  sx.alpha[k]=true_beta>sim_results[,i][[k]][beta]-
                 qnorm(0.95,n)*sim_results[,(i+5)][[k]][beta]
  wald.test[k]=(sim_results[,i][[k]][beta]-
                 true_beta)/sim_results[,(i+5)][[k]][beta]
  p_value[k]=2*min(pt(wald.test[k],n),1-pt(wald.test[k],n))
  }
  list(Mean_Conf_Interval=colMeans(Conf_Interval),
                H0_accepted=sum(real.alpha=="TRUE"),
                H0_rejected=sum(real.alpha=="FALSE"),
                Real_alpha=1-(sum(real.alpha=="TRUE")/
                                  length(real.alpha)),
  Dx_alpha=1-(sum(dx.alpha=="TRUE")/length(real.alpha)),
  Sx_alpha=1-(sum(sx.alpha=="TRUE")/length(real.alpha)),
  Mean_Interval_length=mean(Conf_Interval[,2]-
                       Conf_Interval[,1]),
  mean_Wald.test=mean(wald.test),mean_p_value=mean(p_value))
}
#Inference results organization
summary.inf_true.beta1=summary.inf(1,2,b)
summary.inf_naive.beta1=summary.inf(2,2,b)
summary.inf_BCES.beta1=summary.inf(3,2,b)
ct3_true_beta1=c(summary.inf_true.beta1$Real_alpha,
                 summary.inf_true.beta1$Dx_alpha,
                 summary.inf_true.beta1$Sx_alpha,
                 summary.inf_true.beta1$Mean_Conf_Interval)
ct3_naive_beta1=c(summary.inf_naive.beta1$Real_alpha,
                  summary.inf_naive.beta1$Dx_alpha,
                  summary.inf_naive.beta1$Sx_alpha,
                  summary.inf_naive.beta1$Mean_Conf_Interval)
ct3_BCES_beta1=c(summary.inf_BCES.beta1$Real_alpha,
                 summary.inf_BCES.beta1$Dx_alpha,
                 summary.inf_BCES.beta1$Sx_alpha,
                 summary.inf_BCES.beta1$Mean_Conf_Interval)

##TABLE 1
true_t1<-ct1_true_beta1_100
naive_t1<-ct1_naive_beta1_100
BCES_t1<-ct1_BCES_beta1_100
xtable(true_t1,floating=FALSE,digits=c(rep(4,2)))
xtable(naive_t1,floating=FALSE,digits=c(rep(4,2)))
xtable(BCES_t1,floating=FALSE,digits=c(rep(4,2)))

## TABLE 2
#1-alpha=.95 #1-alpha=.90 #1-alpha=.99
true_t3<-ct3_true_beta1_100
naive_t3<-ct3_naive_beta1_100
BCES_t3<-ct3_BCES_beta1_100
xtable(true_t3,floating=FALSE,digits=c(rep(2,2)))
xtable(naive_t3,floating=FALSE,digits=c(rep(2,2)))
xtable(BCES_t3,floating=FALSE,digits=c(rep(2,2)))
```

Bibliography

[1] Akritas, M. G., and Bershady, M. A. Linear regression for astronomical data with measurement errors and intrinsic scatter. *Astrophys.J. 470* (1996), 706.

[2] Azzalini, A. The skew-normal distribution and related multivariate families*. *Scandinavian Journal of Statistics 32*, 2 (2005), 159–188.

[3] Azzalini, A. *The R package sn: The skew-normal and skew-t distributions (version 1.1-2).* Universitá di Padova, Italia, 2014.

[4] Bennett, C., Larson, D., Weiland, J., and Hinshaw, G. The 1% concordance hubble constant. *The Astrophysical Journal 794*, 2 (2014), 135.

[5] Birnbaum, Z. W. Effect of linear truncation on a multinormal population. *Ann. Math. Statist. 21*, 2 (06 1950), 272–279.

[6] Caroll, R. J. Covariance analysis in generalized linear measurement error models. *Statistics in Medicine 8*, 9 (1989), 1075–1093.

[7] Carroll, R., Ruppert, D., and Stefanski, L. Nonlinear measurement error models. *Monographs on Statistics and Applied Probability.(Chapman and Hall, New York) Volume 63* (1995).

[8] Carroll, R. J., Gallo, R. P., and Gleser, L. J. Comparison of least squares and error-in-variables regression, with special reference

to randomized analysis of covariance. *Journal of American Statistical Association* (1985).

[9] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. *Measurement error in nonlinear models: a modern perspective.* CRC press, 2012.

[10] Carroll, R. J., and Stefanski, L. A. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association 85*, 411 (1990), 652–663.

[11] Carroll, S. M., Press, W. H., and Turner, E. L. The cosmological constant. *Annual review of astronomy and astrophysics 30* (1992), 499–542.

[12] Cook, J. R., and Stefanski, L. A. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association 89*, 428 (1994), 1314–1328.

[13] De Helguero, F. *Sulla rappresentazione analitica delle statistiche abnormali.* 1908.

[14] Devanarayan, V. *Simulation extrapolation methods for heteroscedastic Measurement Error Models with replicate measurements.* PhD thesis, North Carolina State University., 1996.

[15] Efron, B. Bootstrap methods: another look at the jackknife. *The annals of Statistics* (1979), 1–26.

[16] Efron, B., and Stein, C. The jackknife estimate of variance. *The Annals of Statistics* (1981), 586–596.

[17] Emerson, W. H. On quantity calculus and units of measurement. *Metrologia 45* (Apr. 2008), 134–138.

[18] Feigelson, E. D., and Babu, G. J. Big data in astronomy. *Significance 9*, 4 (2012), 22–25.

[19] Fuller, W. A. *Measurement error models*, vol. 305. John Wiley & Sons, 2009.

[20] Guolo, A. *Measurement Error Correction Techniques in Matched Case-Control Studies.* PhD thesis, Universita' degli Studi di Padova, Dipartimento di Scienze Statistiche, 2005.

[21] Hand, D. J. Size mattershow measurement defines our world. *Significance 2*, 2 (2005), 81–83.

[22] Harrison, E. The redshift-distance and velocity-distance laws. *The Astrophysical Journal 403* (1993), 28–31.

[23] Hubbard, D. *How to measure anything. Hoboken.* NJ: Wiley, 2007.

[24] Hubble, E. P. The law of red shifts (george darwin lecture). *Monthly Notices of the Royal Astronomical Society 113* (1953), 658.

[25] Kelly, B. C. Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal 665*, 2 (2007), 1489.

[26] Kelly, B. C. Measurement error in astronomy. *Harvard-Smithsonian Center for Astrophysics* (2011).

[27] Kipnis, V., Subar, A. F., Midthune, D., Freedman, L. S., Ballard-Barbash, R., Troiano, R. P., Bingham, S., Schoeller, D. A., Schatzkin, A., and Carroll, R. J. Structure of dietary measurement error: results of the open biomarker study. *American Journal of Epidemiology 158*, 1 (2003), 14–21.

[28] Lemaître, G. Un univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques. In *Annales de la Société scientifique de Bruxelles* (1927), vol. 47, pp. 49–59.

[29] Loehlin, J. C. *Latent variable models: An introduction to factor, path, and structural analysis* . Lawrence Erlbaum Associates Publishers, 1998.

[30] Mallick, B., Hoffman, F. O., and Carroll, R. J. Semiparametric regression modeling with mixtures of berkson and classical error, with application to fallout from the nevada test site. *Biometrics 58*, 1 (2002), 13–20.

[31] Nagel, E. *On the Logic of Measurement*. Columbia university, 1930.

[32] Rosner, B., Willett, W., and Spiegelman, D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in medicine 8*, 9 (1989), 1051–1069.

[33] Satten, G. A., and Kupper, L. L. Inferences about exposure-disease associations using probability-of-exposure information. *Journal of the American Statistical Association 88*, 421 (1993), 200–208.

[34] Spiegelman, D. Cost-efficient study designs for relative risk modeling with covariate measurement error. *Journal of Statistical Planning and Inference 42*, 1 (1994), 187–208.

[35] Stefanski, L., and Cook, J. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association 90*, 432 (1995), 1247–1256.

[36] Timmermann, K. E., and Nowak, R. D. Multiscale modeling and estimation of poisson processes with application to photon-limited imaging. *Information Theory, IEEE Transactions on 45*, 3 (1999), 846–862.

[37] Tonry, J. L., Dressler, A., Blakeslee, J. P., Ajhar, E. A., Fletcher, A. B., Luppino, G. A., Metzger, M. R., and Moore, C. B. The sbf survey of galaxy distances. iv. sbf magnitudes, colors, and distances. *The Astrophysical Journal 546*, 2 (2001), 681.

[38] Tsiatis, A. A., and Ma, Y. Locally efficient semiparametric estimators for functional measurement error models. *Biometrika 91*, 4 (2004), 835–848.

[39] Venables, W. N., and Smith, D. M. the r development core team. *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics* (2005).