A Robust Bayes Factor for the Evaluation of Informative Hypotheses with respect to Linear Regression Coefficients

Utrecht University

Methodology and Statistics for the Behavioural, Biomedical and Social sciences

Author:

Nicolás Pérez
(Student number: 5824176)

Supervisors:

Prof. Dr. Herbert Hoijtink

Abstract

In this paper we introduce a robust Bayes factor for informative hypotheses testing (IHT) in the context of linear regression models. We first introduce the concept of robust estimates for the linear regression coefficients, and subsequently relate it to the Bayes factor as a valuable tool for IHT. The paper focuses on comparing the performance of two robust Bayes factors with a non-robust Bayes factor based on common ordinary least squares estimates. Comparisons are made using a simulation study where outlier presence and heteroscedasticity conditions are manipulated. With the simulation results we show the advantage of robust Bayes factors over non-robust Bayes factors and propose to investigate further in the definition of criteria to select between the two robust Bayes factors.

Keywords: Bayes factor, Robust Estimation, Multiple Linear Regression, Informative Hypotheses Testing

1 Introduction

Multiple linear regression is often used to model the relationship between a dependent variable and a set of predictors. Several approaches are available to make inferences with respect to the parameters of the regression model (Draper and Smith, 1998). Standard null hypothesis testing (NHT) is considered as the standard for this kind of inferences and relies on the falsification of the null hypothesis H_0 by means of the p-value.

Newer approaches consider the evaluation of informative hypotheses with respect to the coefficients. Informative hypotheses use equality and inequality constraints amongst the regression parameters such that they represent the researcher's expectations. The result is a set of competing hypotheses that can be evaluated by means of the Bayes factors (Hoijtink,2012). An example of informative hypothesis testing (IHT) is the prediction of income from the socioeconomic status SES and the IQ of a person, e.g., is the effect of IQ larger than the effect of SES, and moreover, are both effects positive H_1:β_{IQ}>β_{SES}>0?

The current standard for Bayesian evaluation of informative hypotheses uses the so called "Approximate adjusted fractional Bayes factors" (Gu, Hoijtink, and Mulder, 2017). The procedure, as implemented in the "Bain" R-package (https://informative-hypotheses.sites.uu.nl/software/bain/), constitutes an efficient and flexible way to perform IHT for almost any statistical model since it only needs the parameter estimates, their variance-covariance matrix and the sample size to compute the Bayes factor (BF).

Even though the ordinary least squares (OLS) estimation procedure for multiple linear regression models is not distributionally dependent, most of the common hypothesis tests rely on the existence of normally distributed residuals in the population. This implies absence of outliers and homoscedasticity of the residuals. Several studies have been done regarding the assumptions of the linear regression model and the impact when they are violated. Violations usually imply that the coefficient estimates are biased or that their standard errors are under- or overestimated. For example, as mentioned by Wilcox (2017), the presence of outliers may not only bias the estimates by shifting the regression line to adjust for the extreme residual value, but they could also imply a larger standard error of the associated coefficient (because of the increase of the residual sum of squares). The case is similar when strong heteroscedastic residuals occur; they do not always affect the estimates themselves but they often lead to biased estimates of the standard error of the coefficients. These conditions will also affect the Bayes factors obtained with the Bain package, since the estimation procedure is based on the parameter estimates, their covariance matrix and the sample size.

Robust estimation represents a solution for situations when the data do not fulfill the requirements and assumptions of the underlying model. Although robust statistics have been studied in depth and are available in software, most of these techniques are still rarely used (Wilcox, 2017). In this paper, two robust Bayes factors (BF_{MM} and BF_{LMROB}) are developed for IHT in the context of linear regression models. We investigate their behaviour and we compare it to the Bayes factors based on OLS estimates (BF_{OLS}) when the assumptions of absence of outliers and homoscedasticity of the residuals are violated.

In Section 2 of this paper, we introduce the multiple regression model and its assumptions, and we focus on the problems that usually arise when the data contains outliers and when the residuals show a heteroscedastic behaviour. In Section 3, a definition of robustness is given, and we present the chosen approach for robust estimation of the linear regression coefficients through what is known as M and MM-estimators. In Section 4 we introduce the robust Bayes factor as a tool for IHT in the context of linear regression models. In Section 5 we present the design of the simulation study. In Section 6 we present the results of the simulation study comparing the BF_{OLS}, the BF_{MM} and the BF_{LMROB} in different scenarios that include variations in the proportion of outliers, the type of outliers and the presence of heteroscedastic residuals. Section 7 contains an example of robust Bayesian IHT for linear regression coefficients in a data set taken from the psychology reproducibility project of the Open Science Framework (OSF). This paper concludes with a discussion in Section 8.

2 The Linear Regression Model

Regression analysis is a technique in which researchers want to investigate the relationship between an outcome variable (often called the dependent variable) and a set of observed variables. Different goals may be achieved by using this technique depending on the specific situation that is being studied. Inferential analysis focuses on understanding, from a theoretical perspective, how the predictors are related to the outcome variable. On the other hand, in predictive analysis less focus is given to the theory and the main goal is to come up with the best possible prediction of the outcome variable; this is, obtaining a reliable forecast value for the outcome variable for a set of observations that has not been observed before. In this paper we will focus on inferential analysis of regression models in the form of IHT by means of the Bayes factor.

2.1 Multiple linear regression, model and assumptions

In multiple linear regression models the outcome variable is modeled as a linear combination of the predictors. The simplicity of this model and the high interpretability that is attached to it, is the main reason why the approach is so well suited for inferential analysis in the social sciences (James, Witten, Hastie, and Tibshirani, 2013). In terms of notation, we usually denote the predictors as $\mathbf{X} = \left(X_{(1)}, X_{(2)}, ...X_{(p)}, ..., X_{(P)}\right)$ where $X_{(p)}$ represents a column vector $(x_{1p}, x_{2p}, ...x_{Np})$ with all the N observations of variable p. In following Sections $X_{(0)}$ may be encountered and denotes a 1xN vector with all entries equal to one (Draper and Smith, 1998). The regression coefficients are denoted by $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p, ..., \beta_P)$ and the outcome variable as $\mathbf{y} = (y_1, y_2, ..., y_N)$. A regression model with P predictors is given by the following equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + ... + \beta_P x_{iP} + e_i, \quad \text{for } i = 1, ..., N \quad , \tag{1}$$

$$e_i \sim \mathcal{N}(0, \sigma_e^2), \tag{2}$$

that is, the residuals are normally distributed with mean $u_e = 0$ and variance σ_e^2. The goal is to estimate the regression coefficients $\boldsymbol{\beta}$ using the observed data. It is natural to address the problem from the perspective of selecting the coefficient estimates $\hat{\boldsymbol{\beta}}$ that lead to the best possible prediction $\hat{y}_i$ for subject i given its observed values $X_i = (x_{i1}, x_{i2}, ..., x_{ip}, ..., x_{iP})$:

$$\hat{y}_i = f(X_i, \hat{\boldsymbol{\beta}}) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip} + ... + \hat{\beta}_P x_{iP} \quad , \quad \text{for } i = 1, ..., N. \tag{3}$$

OLS estimates are obtained by choosing $\hat{\boldsymbol{\beta}}$ such that the sum of the squared residuals (e_i) is minimized, that is:

$$\min \sum_{i=1}^{N} e_i^2 = \min \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad , \quad \text{for } i = 1, ..., N. \tag{4}$$

Though many other residual functions could be selected, the one used in OLS is highly practical in the sense that it allows to get an analytical solution for the estimation of the regression coefficients. More details on the procedure can be found in Draper and Smith (1998) and James et al. (2013).

When using OLS estimates, several statistical inferential tests are available such as the significance test of the coefficients, or the model significance test. However, these statistical tests and also, for example, confidence intervals for the regression coefficients, are based on the adequacy of the assumptions of the regression model specified in Equation 1. The most important assumptions are related to the distribution of the residuals and what is known as the Gauss-Markov conditions (Wooldridge, 2012, p.59), which (amongst others) implies absence of outliers in the data and the homoscedasticity of the residuals.

2.2 Outliers

A non formal definition of an outlier is an observation that deviates from the rest of the values that it is part of; that is, an observed (sampled) value that is not likely to come from the population. Several statistical criteria exist to define an outlier and for the purpose of this paper we will use the interquartile range approach defined by Tukey (1977) in his proposal for Box and Whisker plots.

In the context of linear regression it is common to make a distinction between outliers in the X space, the Y space and the XY space. Even if specific terms are often used for outliers in a particular space, in this paper we use the term outlier in general since they represent a similar behaviour and they only differ on the space where they are embedded (and their potential effect which is explained later). Outliers in the three spaces are presented in Figure 1 for a multiple regression with one predictor. As will be elaborated below, regions A, B, and C, represent outliers in X, Y, and XY space, respectively.

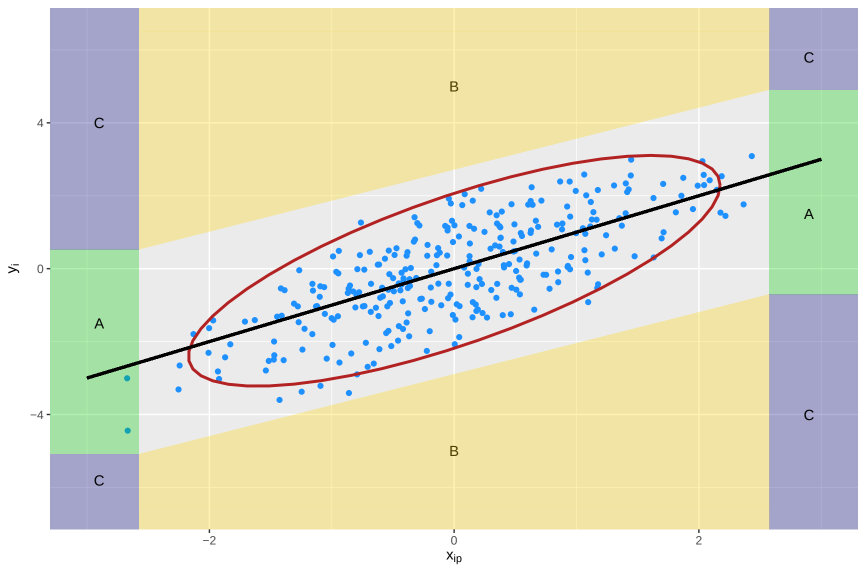


Figure 1: Outlier scenario characterization in a simple linear regression context.

Even though Figure 1 represents a simple linear regression, the three types of outliers can be easily generalized to multiple linear regression. The main issue with outliers in the context of regression is their capability to shift the regression line, since the estimates of the coefficients are obtained by minimizing a residual function which turns to be sensible to extreme values (Wilcox, 2017). In other words, outliers may lead to biased regression coefficients. However, outliers do not always shift the regression line, but in some cases they do affect the estimates of the standard error of the regression coefficients (usually larger). Both conditions, bias and over-estimation of the standard errors of the coefficients, represent a problem when hypotheses are tested in multiple linear regression models since they lead to inferences that are different from the inferences that would be obtained if the outlier is not included in the analysis. More specifically, such situation could lead to a biased Bayes factor and therefore possibly to wrong inferences in the IHT context.

2.2.1 Outliers in the X space

This type of outlier is represented by the green colored areas in Figure 1 ("A" regions). Given the interquartile range approach mentioned before, an observation in the $p-th$ predictor $X_{(p)}$ can be defined as an outlier if:

$$x_{ip} < Q_{1(p)} - 1.5(Q_{3(p)} - Q_{1(p)}) \quad \text{or} \quad x_{ip} > Q_{3(p)} + 1.5(Q_{3(p)} - Q_{1(p)}), \tag{5}$$

for $i \in \{1, ..., N\}$, $p \in \{1, ..., P\}$, and where $Q_{1(p)}$ corresponds to the first quartile of the distribution of $X_{(p)}$ and $Q_{3(p)}$ denotes the third quartile of the distribution of $X_{(p)}$. The difference between the third and the first quartile $(Q_{3(p)} - Q_{1(p)})$ is usually called the interquartile range of variable p $(IQR_{(p)})$. It is important to notice that different definitions of multivariate outliers exist in the literature, however, in the context of this paper we call an observation a multivariate outlier if it fulfills the conditions in Equation 5 in two or more variables.

2.2.2 Outliers in the Y space

The yellow areas ("B" regions) represent the Y space outliers in the context of multiple linear regression. This type of outlier is defined with respect to the residuals; this is, if the regression line is known (given a set of coefficients $\hat{\boldsymbol{\beta}}$), observed values of the variable y with an error term that lies outside the interquartile range of the residuals $(IQR_{(e)})$ can be considered as outliers in the Y space. Formally, an observation i in the dependent variable y is defined as an outlier in the Y space if:

$$e_i < Q_{1(e)} - 1.5(Q_{3(e)} - Q_{1(e)}) \quad \text{or} \quad e_i > Q_{3(e)} + 1.5(Q_{3(e)} - Q_{1(e)}), \tag{6}$$

for $i \in \{1, ..., N\}$, and where $Q_{1(e)}$ corresponds to the first quartile of the distribution of the residual and $Q_{3(e)}$ denotes the third quartile of their distribution.

2.2.3 Outliers in the XY space

Though many combinations of outliers in the XY space could be considered, for the purpose of this paper we will consider XY outliers as observations that combine a general small value in one or more of the predictors and a rather large residual, or vice-versa. These outliers are represented in the graph with blue areas ("C" regions). Mathematically, we refer once again to the interquartile range definition of outliers but now fulfilling conditions in both spaces. We consider an observation (x_{ip}, y_i) an outlier in the XY space if any of the following conditions is met:

$$x_{ip} < Q_{1(p)} - 1.5(Q_{3(p)} - Q_{1(p)}) \quad \text{and} \quad e_i > Q_{3(e)} + 1.5(Q_{3(e)} - Q_{1(e)}), \tag{7}$$

$$x_{ip} < Q_{1(p)} - 1.5(Q_{3(p)} - Q_{1(p)}) \quad \text{and} \quad e_i < Q_{1(e)} - 1.5(Q_{3(e)} - Q_{1(e)}), \tag{8}$$

$$x_{ip} > Q_{3(p)} + 1.5(Q_{3(p)} - Q_{1(p)}) \quad \text{and} \quad e_i > Q_{3(e)} + 1.5(Q_{3(e)} - Q_{1(e)}), \tag{9}$$

$$x_{ip} > Q_{3(p)} + 1.5(Q_{3(p)} - Q_{1(p)}) \quad \text{and} \quad e_i < Q_{1(e)} - 1.5(Q_{3(e)} - Q_{1(e)}), \tag{10}$$

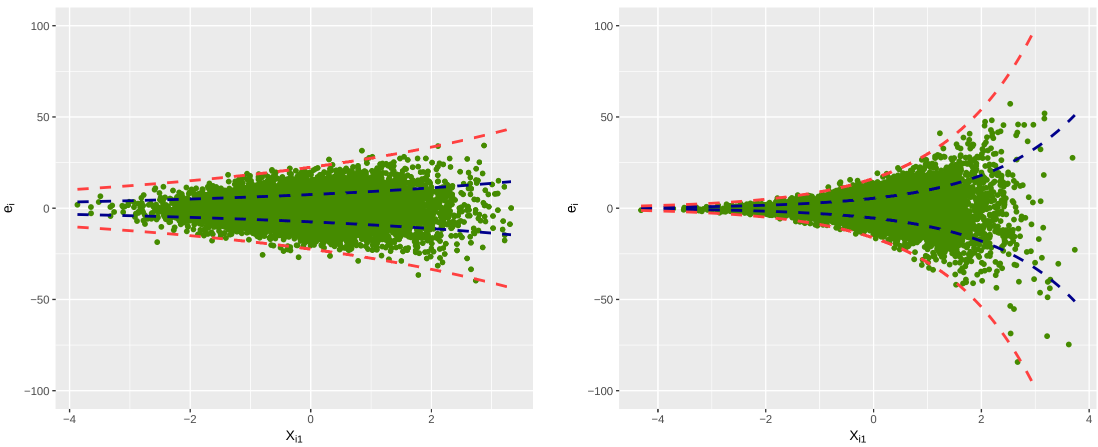
for $i \in \{1, ..., N\}$, $j \in \{1, ..., P\}$, and where $Q_{1(p)}$, $Q_{3(p)}$, $Q_{1(e)}$ and $Q_{3(e)}$ have the same meaning as given in Equations 5 and 6. For this kind of outliers we also consider the possibility of an observation fulfilling the condition on the X space for multiple predictors at the same time.

2.3 Heteroscedasticity

The second assumption that derives from considering a normally distributed error is that the residuals are independently and identically distributed. This means that the error term is not related to any of the variables and that the residual variance σ_e^2 is homogeneous across the different observations. The formal statement of this condition in the linear regression model is:

$$\sigma^2_{e|X_{(p)}=x} = \sigma^2_e \quad , \quad for \quad p = 1, ..., P. \tag{11}$$

When this condition is not fulfilled, we talk about a model with heteroscedastic (non homogeneous) residuals. In this paper we focus on heteroscedastic behaviour commonly known as "funnel" patterns. This patterns imply that there is an association between the measurement error and one of the variables that it is being measured. These kind of cases are encountered, for example, in situations dealing with income/expense data, where usually higher incomes tend to vary more than lower ones. In such a situation it is expected to find a residual with the mentioned behaviour (low error for low values of the predictors and vice-versa). Figure 2 presents a situation with a weak funnel heteroscedastic residual and a situation with a strong funnel heteroscedastic residual.



(a) Weak Funnel Heteroscedastic Residual

(b) Strong Funnel Heteroscedastic Residual

Figure 2: Weak and strong funnel heteroscedastic residuals. Blue lines represent one standard deviation confidence intervals and red lines represent three standard deviation confidence intervals

In this paper we will consider funnel heteroscedastic residuals in relation to one predictor only. When a model contains heteroscedastic residuals, it is known to have a systematic error in the predictions (under/over estimation) given that the residual will increase (or decrease) as a predictor increases (or decreases). However, in the context of hypothesis testing, the main issue with heteroscedasticity is that it directly violates one of the Gauss-Markov conditions which in turn could lead to biased estimates of the parameters and their covariance matrix. In such case, the analysis would result in a biased Bayes factor, and therefore to biased inferences in IHT.

3 Robust Estimation

Fortunately, the situations presented in Section 2 can be handled by the use of robust estimation. Even if the two cases present different problems, they are both tightly linked to the definition of residuals and the way they are included in the estimation of the coefficients. In this Section we will explain the definition of robustness of an estimator and we will also present a suitable approach for robust estimation in multiple linear regression models, for cases in which outliers and heteroscedasticity are present.

We start with a simple definition by calling an estimator robust if the estimates are unaffected or vary slightly (relative to the parameters in the reference population which contains no outliers and no heteroscedasticity as elaborated later in Section 5) when changes occur in the distribution of the input values that are used to compute it (Wilcox, 2017). Further definitions about robustness are given in the literature and the interested reader is referred to Huber (1981) and Wilcox (2017). A more specific definition is often used relating it to two mathematical concepts; quantitative robustness and infinitesimal robustness (Huber, 1981; Ruckstuhl, 2016; Wilcox, 2017). In the remaining part of this Section we explain the importance of these concepts and how they relate to robust estimation in the linear regression context, specifically why they are relevant for the definition of MM-estimators.

3.1 Quantitative Robustness

To define "quantitative robustness" we use the so called breakdown point, which refers to the maximal amount (as a proportion of the total sample size) of inconsistent observations (in our context represented by outliers or heteroscedastic residuals) that can be present in the data before the estimation completely breaks or renders implausible estimates (Ruckstuhl, 2016; Wilcox, 2017). A classic example to understand quantitative robustness is obtained when considering the population mean, which breakdown point is 0. That is, it can proven that if a single (large enough) outlier is added to the data, the mean estimate will be completely inaccurate. Therefore, the maximum proportion of admitted inconsistent observations for the mean example is 0. The breakdown point in linear regression behaves similarly; the reader is referred to Figure 1 to consider a case were an additional observation is added in one of the regions denoted by "C". It is clear that the regression line would be affected and biased relative to the original regression line. Given the relationship between outliers and the breakdown point, we prefer estimators with a high breakdown point, since they would be able to handle a larger amount of outliers. MM-estimation is based on the principle of obtaining estimates with high breakdown points (which theoretically can not be higher that 0.5 for any estimator).

3.2 Infinitesimal Robustness

"Infinitesimal robustness" is a measure of sensibility on the estimator, that is, what is the effect of an outlier (also referred as polluted observation x^*) on the estimator value. To understand what x^* represents, suppose the following example where $X_1 = \{5, 3, 7, 8, 9, 11, 1, 6\}$. We define $x^* = 33$, which represent a large value that was somehow obtained (e.g. typo or measurement error) but that in reality was a small observation in X_1. Therefore a new X_1 is given by $X_1 = \{5, x^*, 7, 8, 9, 11, 1, 6\} = \{5, 33, 7, 8, 9, 11, 1, 6\}$. Fitting a simple regression with and without x^* would result in different estimates. An estimator has infinitesimal robustness if for increasing x^* (tending to infinity), the residual function (cf. Equation 4) maintains a constant value. Robust estimators with infinitesimal robustness are preferred since they are robust to the magnitude of the outliers. This is also relevant if the residuals show strong funnel behaviour since funnels may contain very extreme outliers. For further information the interested reader is referred to Huber (1981) and Wilcox (2017).

3.3 Robust estimation in linear regression

3.3.1 M-estimators using OLS residual function

M-estimators, also known as maximum likelihood type estimators, were one of the first approaches to robust estimation in regression models and were studied by Huber (1981) when considering how to deal with outliers in these models. M-estimators generalize the idea of finding a set of parameter estimates (regression coefficients in our case) that minimizes a function of the residuals denoted by $\rho_M(\boldsymbol{e})$. The general form of this is:

$$\min \rho_M(\boldsymbol{e}) = \min \sum_{i=1}^{N} \rho_M(e_i) \tag{12}$$

where:

$$e_i = f(X_i, y_i, \hat{\boldsymbol{\beta}}) = y_i - (\hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p x_{ip}) \quad , \quad \text{for } i = 1, ..., N. \tag{13}$$

Using this definition it is possible to show that OLS estimates are a specific form of M-estimators, where the chosen function is the sum of the squared residuals (Figure 3-A):

$$\min \sum_{i=1}^{N} \rho_M(e_i) = \min \sum_{i=1}^{N} e_i^2 \quad , \quad \text{for } i = 1, ..., N. \tag{14}$$

Many approaches can be used to estimate the regression parameters using Equation 12. The most common one uses differential calculus:

$$\frac{d}{d\beta_p} \sum_{i=1}^{N} \rho_M(e_i) = 0 \quad , \quad \text{for } p = 0, 1, ..., P. \tag{15}$$

It can be proved that for OLS estimation, the previous system of equations is equivalent to solving what is known as the system of normal equations (Draper and Smith, 1998, p. 23):

$$\sum_{i=1}^{N} x_{ip} e_i = 0 \quad , \quad \text{for } p = 0, 1, ..., P. \tag{16}$$

3.3.2 M-estimators with other residual functions

The whole idea of M-estimators is based on the fact that using flexible residual functions would allow to penalize and even ignore very extreme residuals. Figure 3 shows the behaviour of different residual functions that are used to estimate the parameters of regression models, there it can be seen how sensitive the squared residual is when large errors are present in the data. For example, a residual of 5 renders a contribution of 25 to Equation 13 compared to contributions of 5.82 for the Huber (Figure 3-B) and 3.65 for the bisquare (Figure 3-C) residual functions.

As explained by Huber (1981) and Yohai (1987), $\rho_M(\boldsymbol{e})$ should have the some additional properties to make the estimation robust. These are: 1) $\rho_M(0) = 0$, 2) $\rho_M(e_i)$ is symmetric, 3) $\rho_M(e_i)$ is non-negative, 4) $\rho_M(e_i)$ is monotonic such that $\rho_M(e_i) < \rho_M(e_{i'})$ if $|e_i| < |e_{i'}|$, and 5) $\psi_M(e_i) = (d/d\beta_p)\,\rho_M(e_i)$ is continuous. Even if many functions meet these conditions, some of them have been studied in depth and are known to be useful such as the Huber or the bisquare function (Ruckstuhl, 2016).

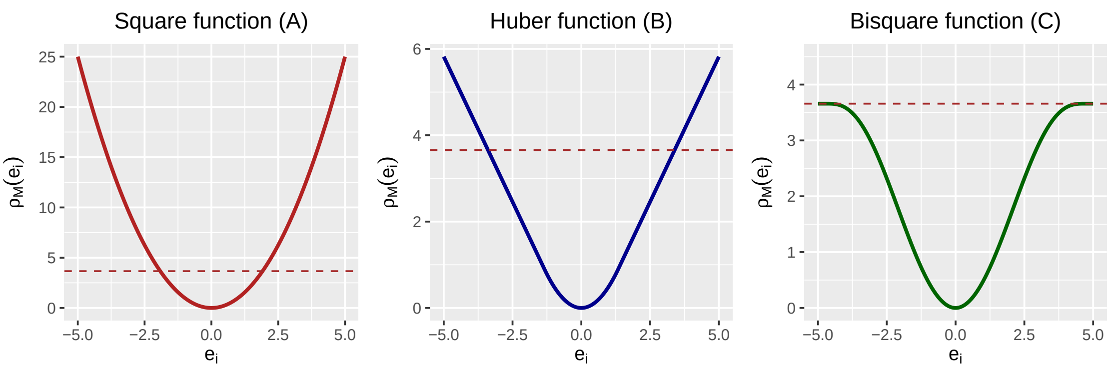


Figure 3: Comparison between the three mostly used residual functions for M estimators.

Notice that for large outliers, the Huber function is always increasing whereas the bisquare takes a maximum value after a cutoff point. The formal definition of the bisquare function is given by:

$$\rho(e_i) = \begin{cases} \frac{k^2}{6}\left(1 - \left(1 - \left(\frac{e_i}{k}\right)^2\right)^3\right) & |e_i| \le k \\ \frac{k^2}{6} & |e_i| > k \end{cases}, \tag{17}$$

where k is known as the rejection point and represents the value after which residuals are given the maximum value in the residual function. A common choice for k in the context of M-estimators for regression models is to use a value of 4.685 (Maronna, Martin, and Yohai, 2006) which leads to 95% efficiency if the residuals are normally distributed.

To obtain M-estimators in the context of linear regression model, the procedure that is commonly used is called Iterative Reweighted Least Squares (IRLS). This starts with the general form of M-estimators in Equation (12) in which the residuals are redefined to be scale-equivariant (see Wilcox, 2017, Chapter 2) :

$$\min \rho_M(\boldsymbol{e'}) = \min \sum_{i=1}^{N} \rho_M(e'_i),$$

where $e'_i = e_i/\hat{\sigma}_e$ and $\hat{\sigma}_e$ is an estimate of the residual standard deviation σ_e. A common choice for $\hat{\sigma}_e$ is the adjusted Mean Absolute Deviation (MAD) since it is a robust estimate for the standard deviation (Susanti, Pratiwi, Sulistijowati, and Liana, 2014; Ruckstuhl, 2016):

$$\hat{\sigma}_e = \frac{median|\, e_i - median\,(e_i)\,|}{0.6745}. \tag{18}$$

Like before, estimates of the regression coefficients can be obtained by solving the following system of equations:

$$\sum_{i=1}^{N} \frac{d}{d\beta_p} \rho_M\,(e'_i) = 0 \quad , \quad \text{for } p = 0, 1, ..., P. \tag{19}$$

Using $\psi_M\,(\mathbf{e'}) = (d/d\beta_p)\,\rho_M\,(\mathbf{e'})$, a solution for such system is given by the counterpart of the normal equations as presented in Equation 16:

$$\sum_{i=1}^{N} \psi_M\,(e'_i)\,x_{ip} = 0 \quad , \quad \text{for } p = 0, 1, ..., P. \tag{20}$$

Defining weights as:

$$w_i = \frac{\psi_M\,(e'_i)}{e'_i} \quad , \quad \text{for } i = 1, ..., N, \tag{21}$$

and solving for $\psi_M\,(e'_i)$, renders the general form of the normal equations for the estimation of the regression parameters as presented by Maronna et al. (2006):

$$\sum_{i=1}^{N} w_i e'_i x_{ip} = 0 \quad , \quad \text{for } p = 0, 1, ..., P. \tag{22}$$

In Equation 22 the weights are dependent on the estimates of the regression coefficients $\hat{\boldsymbol{\beta}}$ and the residual standard deviation ($\hat{\sigma}_e$). To obtain these estimates Iteratively Reweighted Least Squares is used. The general algorithm as explained by Ruckstuhl (2016) can be found in Appendix A.

3.3.3 MM-estimators

As elaborated by Maronna et al. (2006), M-estimators suffer from low efficiency when compared to OLS, this is that the standard error of the estimates are over-estimated. This represents a problem in IHT using Bayes factors as it uses the robust estimates and their covariance matrix. For this reason (and some additional ones as the possibility of M-estimators having a low breakdown point), Yohai (1987) proposed an approach that would lead to robust estimates based on M-estimators but that would have higher efficiency. These estimators are known as MM-estimators (Modified M-estimators) and they can be obtained by using the IRLS described for M-estimators, with a redefinition in the way that the residual variance is estimated.

The general idea of the MM-estimators is to estimate the regression parameters using iteratively updated robust estimates of the residual variance. In the literature these are known as S-estimators (standing for scale estimators) and these are the reason why MM-estimators are referred to as MS-estimators or M-estimators with S-initialization (Wilcox, 2017). S-estimates are the equivalent of M-estimates for scale measures that asymptotically achieve a breakdown point of 0.5. As is elaborated Maronna et al. (2006), this can be achieved by replacing Equation 18 by an iterative estimation for the residual variance given by:

$$\frac{1}{N}\sum_{i=1}^{N}\rho_S\left(\frac{e_i}{\hat{\sigma}_e}\right)=\delta \tag{23}$$

where in each iteration the residuals are recomputed using the current M-estimates of the regression parameters and where the minimum of δ and $(1-\delta)$ corresponds to the maximal breakdown point of the scale estimator. For MM-estimators δ is chosen to be 0.5 to attain a maximal breakdown point (any other choice results in a smaller value for the breakdown point).

Equation 23, is the counterpart of ρ_M for scale estimators. Several studies exist regarding the functions that should be used and how their parameters can be tuned to attain the desired efficiency levels (Maronna et al., 2006). For MM-estimates in regression models a common choice is to use the bisquare function with a k value of 1.56 which leads to an efficiency of 95% (if the residuals are normally distributed). One important fact about MM-estimators is that by trying to attain higher efficiency levels (that is, decreasing the variance of the estimators), often a small bias is introduced in the estimates of the regression coefficients. However, this is not problematic in the sense that MM-estimators are highly efficient and have an asymptotic breakdown point of 0.5.

MM-estimators can be obtained using an iterative procedure that aims for robust estimates of the regression parameters by combining M-estimation for the regression coefficients with S-estimation for the residual standard deviation, in such a way that the final estimates have a high breakdown point and are also highly efficient. Details on how to compute MM-estimators can be found in Appendix B. More detailed information about the procedure and properties of these estimators can be found in Maronna et al. (2006), Susanti et al., 2014 (2014) and Yohai (1987).

3.3.4 Robust Linear Regression in R

All the analysis in this paper were performed using R Version 3.4.3 (R Core Team, 2017). Many packages are available to perform robust regression in R allowing for great flexibility and different approaches for the researcher. However, in this paper we focus on the use of the "rlm" function from the "MASS" package Version 7.3-47 (Venables and Ripley, 2002) and the "lmRob" function from the "robust" package Version 0.4-18 (Wang et al., 2017).

The reason to choose the "rlm" function is that it allows the use of MM-estimators which are widely studied and have strong foundations in robust estimation theory. Moreover, this procedure also allows to use different parameterizations for the residual functions and estimation algorithms. For this paper we used MM-estimators with the parameters encountered often in the literature, that is, using bisquare functions for both location an scale estimates and using a k_M=4.685 for the M-estimate of location in the Equation 19 function and k_S=1.548 in Equation 23. These parameters result in robust estimates in regression models that have a high breakdown point and high efficiency (95% under normality of the residuals) (Maronna et al., 2006; Wilcox, 2017).

Despite that no detailed documentation about the theoretical background (apart from the package help file) was found for the "lmRob" function, we still decided to use it given its wide popularity in the field and also due to the reputation of its contributors (e.g. Yohai). It is important to highlight that this function is based on the use of different robust estimation techniques to obtain estimates with high efficiency and a high breakdown point but no specific information could be found about which specific kind of estimators and/or procedures are used.

4 Using Bayes factors fro IHT

Different approaches are available when researchers want to evaluate their hypotheses. For a long time, NHT by the means of p-values was considered the gold standard. However, recent studies have shown some of the downsides of the before-mentioned approach. The latter includes the fact that the rejection in NHT does not necessarily imply that the alternative hypothesis is true (Wagenmakers, 2007) as well as the fact that the choice of the significance level is somehow arbitrary and lacks foundation (Rosnow and Rosenthal ,1989). Moreover, it is increasingly realized that the significance level leads to questionable research practices, p-hacking, and straight out researcher misconduct in the quest for significant results (Ioannidis, 2005 ; Masicampo and Lalande, 2012).

As elaborated by Hoijtink (2012) NHT can be replaced by IHT and instead of the p-value, the Bayes factor can be used to evaluate a set of competing hypotheses. With this approach, more complex hypotheses than the ones investigated through NHT can be investigated by defining sets of inequality and/or equality constraints among the parameters of interest (in this paper, regression coefficients).

Nevertheless, given the (relative) recentness of this approach, few studies have addressed topics related to how this method is affected when the model assumptions are not met (van Rossum, van de Schoot, and Hoijtink, 2013). Specifically, in linear regression, even though IHT has been studied, not much is known in terms of how outliers and heteroscedasticity affect the results of the analyses.

In this Section we address the development of a robust Bayes factor for IHT in multiple linear regression models, based on the latest procedure for Bayes factor computation, known as the Approximate Adjusted Fractional Bayes Factor (AAFBF). To properly understand the AAFBF as presented by Gu et al. (2017), we first have to give a formal definition of informative hypotheses:

$$H_s : \mathbf{R_{s_0}}\boldsymbol{\beta} = \mathbf{r_{s_0}} \qquad , \qquad \mathbf{R_{s_1}}\boldsymbol{\beta} > \mathbf{r_{s_1}} \qquad , \tag{24}$$

where $\mathbf{R_{s_0}}$ and $\mathbf{R_{s_1}}$ represent the equality and inequality constraint matrices, and $\mathbf{r_{s_0}}$ and $\mathbf{r_{s_1}}$ represent the constant vectors that jointly define H_s. In the context of a linear regression model with P predictors, N_{s_0} equality constraints and N_{s_1} inequality constraints, the hypothesis H_s can be written in matrix notation as:

$$\mathbf{R_{s_0}}\boldsymbol{\beta} = \begin{bmatrix} R_{s_0\,11} & \dots & R_{s_0\,1P} \\ \dots & \dots & \dots \\ R_{s_0\,N_{s_0}1} & \dots & R_{s_0\,N_{s_0}P} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_P \end{bmatrix} = \begin{bmatrix} r_{s_0\,1} \\ \dots \\ r_{s_0\,N_{s_0}} \end{bmatrix},$$

$$\mathbf{R_{s_1}}\boldsymbol{\beta} = \begin{bmatrix} R_{s_1\,11} & \dots & R_{s_1\,1P} \\ \dots & \dots & \dots \\ R_{s_1\,N_{s_1}1} & \dots & R_{s_1\,N_{s_1}P} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \dots \\ \beta_P \end{bmatrix} > \begin{bmatrix} r_{s_1\,1} \\ \dots \\ r_{s_1\,N_{s_1}} \end{bmatrix}.$$

It is common to test an informative hypothesis versus the unconstrained hypothesis:

$$H_u : \boldsymbol{\beta}, \tag{25}$$

that is, there are no constraints on $\boldsymbol{\beta}$. As pointed out by Gu et al. (2017), under such conditions, the AAFBF of the constrained hypothesis (H_s) against the unconstrained hypothesis H_u can be computed as:

$$AAFBF_{su} = \frac{f_s}{c_s} \qquad , \tag{26}$$

where c_s is a measure of the complexity of the informative hypothesis H_s relative to the unconstrained hypothesis H_u, and f_s represents a measure of the support that the data provides in favor of hypothesis H_s relative to the unconstrained hypothesis H_u. Note that the complexity in the Bayes factor can also be seen as a measure of the parameter space size under the constraints relative to the unconstrained space size. For the purpose of this paper it suffices to recall the mathematical definition of fit f_s and complexity c_s as presented by Gu et al. (2017):

$$f_s = \int_{\boldsymbol{\beta}\in H_s} \mathcal{N}\left(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\right) d\boldsymbol{\beta}, \tag{27}$$

where, $\hat{\boldsymbol{\beta}}$ denotes the estimates of the regression coefficients and and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ their covariance matrix. The most important fact to note is that the fit depends only on the choice H_s, the regression parameter estimates $\hat{\boldsymbol{\beta}}$ and the corresponding variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$. We propose a robust estimate for the Bayes factor by using robust estimates of the regression coefficients and a robust estimate of their variance-covariance matrix. The complexity c_s given by Gu et al. 2017 is:

$$c_s = \int_{\boldsymbol{\beta} \in H_s} \mathcal{N}\left(\boldsymbol{\beta}|\boldsymbol{\beta}_b, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}/b\right) d\boldsymbol{\beta} \qquad , \tag{28}$$

where $\boldsymbol{\beta_b}$ represents the adjusted prior mean given by

$$\boldsymbol{\beta_b} \in \{\boldsymbol{\beta}|\mathbf{R_{s0}}\boldsymbol{\beta} = \mathbf{r_{s0}}, \mathbf{R_{s1}}\boldsymbol{\beta} > \mathbf{r_{s1}}\} \tag{29}$$

and

$$b = J/N \qquad , \tag{30}$$

where N corresponds to the sample size of the data set was used for the estimation of the regression parameters. Note that J is the row-rank (independent rows) of the constraints matrix $\mathbf{R} = [\mathbf{R}_{1_0}, \mathbf{R}_{1_1}, \mathbf{R}_{2_0}, \mathbf{R}_{2_1}, ..., \mathbf{R}_{s_0}, \mathbf{R}_{s_1}, ..., \mathbf{R}_{S_0}, \mathbf{R}_{S_1}]$. For the details of the derivation of the AAFBF and the measures of complexity and fit, we refer the reader to Gu, Mulder, Deković, and Hoijtink (2014) and Gu et al. (2017).

The Bayes factor represents a measure of the support in the data for pairs of competing hypotheses (Gu et al., 2017). More specifically, a Bayes factor of 4 represents that the evidence that the data provides in favor of H_s is 4 times larger than the evidence for H_u. To compare two or more competing informative hypotheses we will use Posterior Model Probabilities (PMPs) which are based on equal prior model probabilities and contain the same information as the Bayes factors:

$$PMP_s = \frac{BF_{su}}{\sum_{l=1}^{S} 1 + BF_{lu}} \qquad , \tag{31}$$

where $S+1$ is the number of competing informative hypotheses ($H_1, H_2, ..., H_l, ..., H_S$ plus the unconstrained hypothesis H_u) that are being tested. PMPs are logical probabilities, that is, numbers on a zero to one scale that add up to one, that can be interpreted as the probability of each hypothesis conditional on the data. In other words, how likely is each hypothesis given the estimates of the regression coefficients, their variance-covariance matrix, and the sample size. In the following Sections we describe and compare the results of a simulation study using robust and non-robust Bayes factors. We will use BF_{OLS}, BF_{MM} and BF_{LMROB} to denote the Bayes factors (and PMPs) obtained using regression estimates obtained through OLS, MM-estimation ("rlm") and the "lmRob" function, respectively.

5 Simulations

To begin, we first define a linear regression model in which all the assumptions are met and where three different reference populations are defined (see parameter setup in Table 1). Subsequently, we investigate the effect of violating the assumptions of absence of outliers and heteroscedasticity under different scenarios.

In the simulations we will use a multiple linear regression model with three predictors ($P = 3$) which have zero means and the following covariance matrix (this is, the predictors are standardized):

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.00 & 0.11 & 0.08 \\ 0.11 & 1.00 & 0.14 \\ 0.08 & 0.14 & 1.00 \end{bmatrix}$$

.

The proportion of variance explained at the population level is fixed at R^2=0.35 and the regression coefficients and residual variance (displayed in Table 1) are chosen accordingly. In each cell of the simulation design (displayed in Table 2) 1000 data sets with N=100 are simulated. Details on how the parameter values in Table 1 are chosen and how data sets are simulated can be found in Appendix C. In the simulations the following hypotheses will be considered (see also Table 1 and Table 2)

$$H_1 : \beta_1 = \beta_2 = \beta_3 = 0 \quad , \tag{32}$$

$$H_2 : \beta_1 > \beta_2 > \beta_3 > 0 \quad , \tag{33}$$

$$H_3 : \beta_1 > 0, \beta_2 > 0, \beta_3 < 0 \quad , \tag{34}$$

and

$$H_u : \beta_1, \beta_2, \beta_3. \tag{35}$$

Later in this paper, these will be referred as "Null", "Ordered Positive", "Directional" and "Unconstrained" respectively. It is important to point out that mutual comparison of regression coefficients as in Equations 32-35 only makes sense if the corresponding variables are measured on the same scale. This is arranged in our simulation study by generating standard normal data for each of the three predictors.

Table 1: Choices of β's for the different informative hypotheses

	Null	Ordered Positive	Directional
	$H_1 : \beta_1 = \beta_2 = \beta_3$	$H_2 : \beta_1 > \beta_2 > \beta_3 > 0$	$H_3 : \beta_1 > 0, \beta_2 > 0, \beta_3 < 0$
β_0	3	5	5
β_1	0	4	4
β_2	0	3	3
β_3	0	2	-2
R^2	0	0.35	0.35
σ_e	1.857	8.016	7.298

Using these informative hypotheses and recalling the assumptions of interest discussed in Section 2, we use the simulation design displayed in Table 2. We will not combine all the factors (outlier space, outlier amount, outlier size and heteroscedastic residual behaviour) in a full factorial design. Instead we will explore the main effect of these factors in combinations that give a clear idea on the impact of violations when the BF_{OLS}, the BF_{MM}, and the BF_{LMROB} are used for IHT.

Table 2: Simulation scenarios outline

	H_1	H_2	H_3
1) Base Scenario			
All assumptions met	Scenario 1	Scenario 2	Scenario 3
2) Outlier Type			
Outliers in X space (10% - 1.5 to 3 IQR)	Scenario 4	Scenario 5	Scenario 6
Outliers in Y space (10% - 1.5 to 3 IQR)	Scenario 7	Scenario 8	Scenario 9
Outliers in XY space (10% - 1.5 to 3 IQR)	Scenario 10	Scenario 11	Scenario 12
3) Amount of outliers (XY space)			
15% of observations (1.5 to 3 IQR)	Scenario 13	Scenario 14	Scenario 15
20% of observations (1.5 to 3 IQR)	Scenario 16	Scenario 17	Scenario 18
25% of observations (1.5 to 3 IQR)	Scenario 19	Scenario 20	Scenario 21
4) Outlier Size (10% outliers in XY)			
Large (2.5 to 5 IQR)	Scenario 22	Scenario 23	Scenario 24
5) Heteroscedastic residual			
Weak funnel	Scenario 25	Scenario 26	Scenario 27
Strong funnel	Scenario 28	Sceanrio 29	Scenario 30

Scenarios 4-30 are obtained by modifying the reference populations to violate the assumptions of absence of outliers and heteroscedastic residuals as given in Section 2. The general method for outlier generation is described below:

- **Outliers in the X space:** using the definition in Section 2, we randomly sample observations (x_{ip}) and replace their values with new small values that satisfy the first condition in Equation 5. New values are obtained by taking the first quantile ($Q_{1(p)}$) of the variable and subtracting the $IQR_{(p)}$ multiplied by a random value between 1.5 and 3 (this range denotes the outlier size that is modified in Scenarios 22-24).
- **Outliers in the Y:** outliers in this space are generated in a similar manner but using the residuals instead. Random observations are sampled and their residual is replaced by a new residual that is obtained by taking the third quantile ($Q_{3(e)}$) of the residuals and adding its $IQR_{(e)}$ multiplied by a random value between 1.5 and 3 (here again, this range denotes the outlier size that is modified in Scenarios 22-24).
- **Outliers in the XY:** outliers in the XY space are generated by combining both procedures; first generating low values for the X space and subsequently increasing their residuals as described before.

For the generation of heteroscedastic residuals we use the procedure described in Appendix D and with values α_1=0.2 for a weak funnel behaviour and α_1=0.8. More details about each scenario can be found on the scripts provided in `https://github.com/pereznic/robustBF`.

6 Simulation Results

We focus on three different measures that allow us to evaluate the performance of the estimators in terms of how accurate they are and how well they perform in terms of Bayesian IHT. The chosen measures are: 1) Bias of the estimates, 2) Coverage probabilities, and 3) Proportion of times the true hypothesis had the largest PMP of the four hypotheses under consideration. The bias of the estimate is defined as (James et al., 2013):

$$bias(\hat{\beta}_p) = \frac{\sum_{t=1}^{1000} \hat{\beta}_{pt}}{1000} - \beta_p \qquad \text{for } p = 0, 1, 2, 3 \qquad , \tag{36}$$

where $\hat{\beta}_{pt}$ represents the estimate of β_p in simulated data set t and β_p represents the regression coefficient for variable p in the reference population. Coverage probabilities are the percentage of cases where the true parameter (in this case the regression coefficient from the reference population) falls in the 95% confidence interval of the parameter estimate:

$$CP(\hat{\beta}_p) = \frac{\left(\sum_{t=1}^{1000} a_{pt}\right)}{T} \qquad \text{for } p = 0, 1, 2, 3 \qquad , \tag{37}$$

where:

$$a_{pt} = \begin{cases} 1 & \text{if} \quad \hat{\beta}_{pt} - 1.96 \times \sqrt{\hat{\boldsymbol{\Sigma}}_{\beta_{pp_t}}} < \beta_p < \hat{\beta}_{pt} + 1.96 \times \sqrt{\hat{\boldsymbol{\Sigma}}_{\beta_{pp_t}}} \\ 0 & \text{otherwise,} \end{cases}$$

in which $\hat{\boldsymbol{\Sigma}}_{\beta_{pp_t}}$ represents the p-th diagonal element of $\hat{\boldsymbol{\Sigma}}_{\beta}$, that is, the standard error of the p-th regression coefficient in simulated dataset t. Even though the previous measures are helpful to assess the performance of the robust estimation, they do not serve the purpose of evaluating their performance in a Bayesian IHT context. Therefore, the third measure is the proportion of simulated data sets rendering the highest PMP for the true hypothesis:

$$P(H_s) = \frac{\sum_{t=1}^{1000} q_{ts}}{1000} \tag{38}$$

where:

$$q_{ts} = \begin{cases} 1 & \text{if} \quad PMP_{ts} = \max\left(PMP_{t1}, PMP_{t2}, PMP_{t3}, PMP_{tu}\right) \\ 0 & \text{otherwise.} \end{cases}$$

Figure 4 summarizes the bias of each regression coefficient for the different scenarios. First thing to notice is that the bias in the different scenarios behaves similarly, independent of the true hypothesis that is being modeled (except for β_3 under H_3). For scenarios in which there are outliers (Scenarios 4-24), robust estimates outperform OLS in terms of bias, providing more accurate estimates (smaller bias). Note also that LMROB estimates are less biased than the MM-estimates. Also, when 25% of the data are outliers in the XY space (Scenarios 19, 20 and 21), MM-estimates have a significant decrease in their performance (larger bias) compared to the LMROB estimates. The factor outlier size seems to have no influence on the bias for the MM and LMOROB estimates; as can be seen, the bias in Scenarios 10, 11 and 12 is practically the same as the bias in Scenarios 22, 23 and 24 (respectively). Finally, the presence of heteroscedastic residuals (weak and strong) do not have a noticeable effect in terms of the bias of the estimates (Scenarios 25-30) for either OLS, MM or LMROB. To sum up, robust estimators behave better than non-robust methods providing smaller bias in the estimates as well as being more stable when outliers are present.

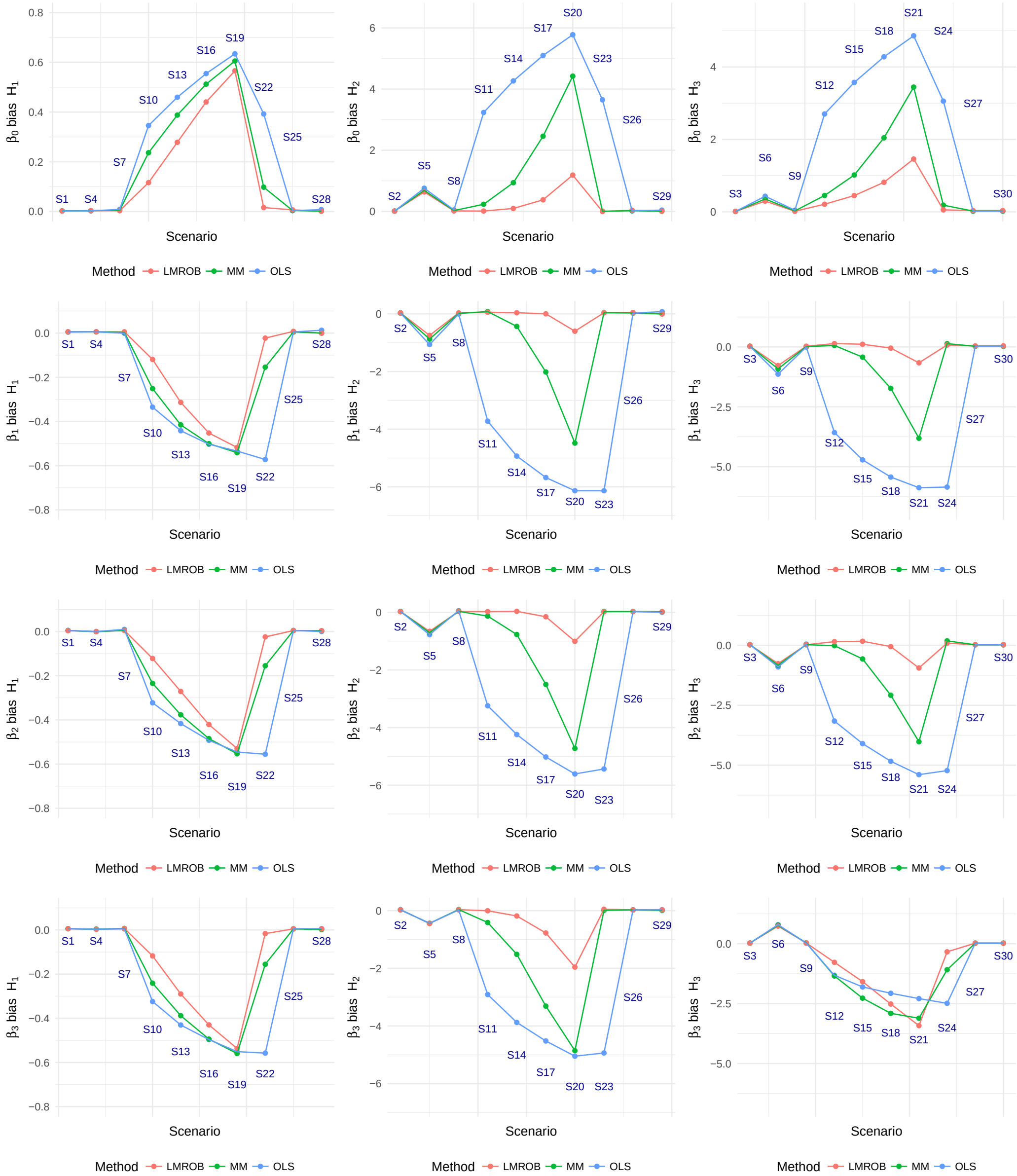


Figure 4: Bias of the estimate of each regression coefficient under H_s

Figure 5 shows the coverage probabilities for each regression coefficient under the different scenarios. We can observe a behaviour analogous to the results found with respect to the bias in

Figure 4. The coverage probabilities of MM and LMROB outperform those of OLS in most of the cases. This is very clear for H_2 and H_3, where for OLS the coverage probabilities decrease dramatically when outliers in the XY space are present (Scenarios 13-24). It is also clear that LMROB outperforms the MM-estimates when a considerable amount of outliers in the XY space is present in the data (Scenarios 14-21). However it is also important to note that when the percentage of outliers is smaller or equal to 15% (Scenarios 1-15) both robust estimators provide much more accurate results than the OLS method. The size of the outliers (Scenarios 22-24) does not have a considerable impact on the performance of the two robust estimation methods in terms of the coverage probabilities. The latter is not true for OLS since the residual function for this estimator is not bounded. Finally, the presence of a heteroscedastic residual (Scenarios 25-30) does not have a particular impact in the estimates obtained with either OLS, MM or LMROB. Previous results encourage the use of robust estimators since the results for coverage probabilities are higher and more stable (when outliers are present) than the ones obtained with non-robust estimators.

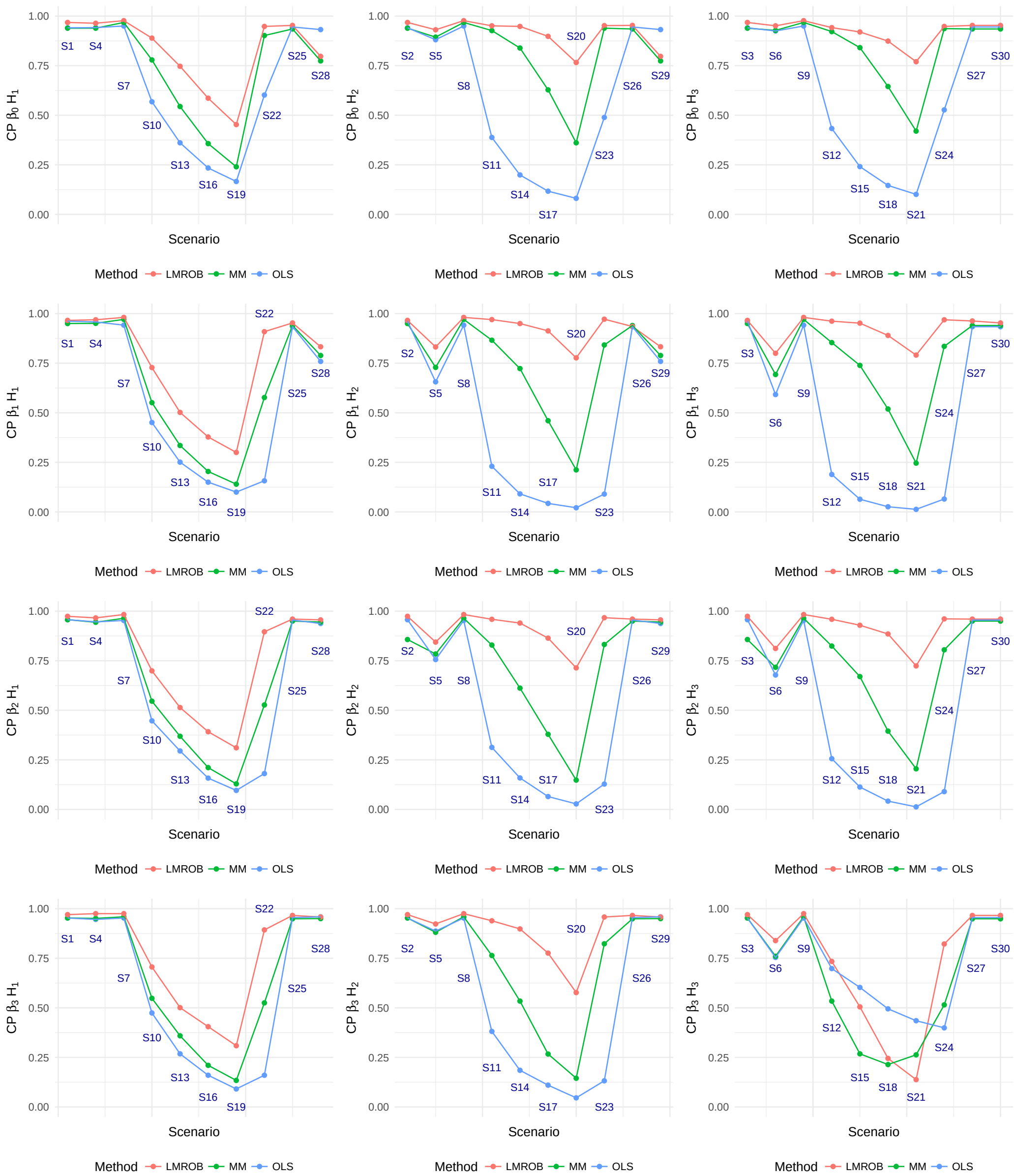


Figure 5: Coverage Probabilities of the confidence intervals for each regression coefficient under H_s

Figure 6 shows the proportion of data sets (in each scenario) in which the highest PMP corresponds to the true hypothesis. As can be seen, considering the three different hypotheses, the

BF_{LMROB} and BF_{MM} outperform the BF_{OLS} since they have higher proportion of correctly classified data sets. It can also be seen that in most of the cases, the proportion of correct classifications is higher for the BF_{LMROB} than for the BF_{MM}. Note that the performance of BF_{OLS} varies a lot, performing extremely well when H_1 is true and extremely bad when H_3 is true. The previous is related to the fact that Bayes factors are sensitive to the choice of the investigated hypotheses; it might well be that a combination of biased estimates and small coverage probabilities leads to an unreasonably high support for H_1. This indicates that the BF_{OLS} is not the most reliable option and should be used carefully. Note as well that the results obtained with the OLS estimates are considerably low and unstable in the scenarios where informative hypotheses are present (H_2 and H_3) which makes them even less reliable for IHT research.

Focusing on the robust Bayes factors we first must notice that the BF_{MM} behaves differently than the BF_{LMROB} in two specific conditions. The first condition is when the outlier size increases and the "Null" hypothesis is true. In this case, the BF_{MM} has an important performance decrease compared to the BF_{LMROB}. By looking at the bias and coverage probabilities we can see that the reason for this is the combination of a high bias presence and a small coverage probability. Furthermore, as explained by Maronna et al. (2006), MM-estimators are known to induce some level of bias as a trade-off for a better estimation of the variance. In this case, the effect of larger outliers ends up in biased coefficients with small confidence intervals, leading to a decrease in the performance of the Bayes factors for the "Null" hypothesis. The second case is when the "Directional" hypothesis (H_3) is true and the amount of outliers in the XY space reaches a level of 20% of the observations (or more). Once again, we recall the bias and variance trade-off that is present in MM-estimators; Figure 4 shows an increased bias for the MM-estimates compared to the ones obtained in the LMROB.

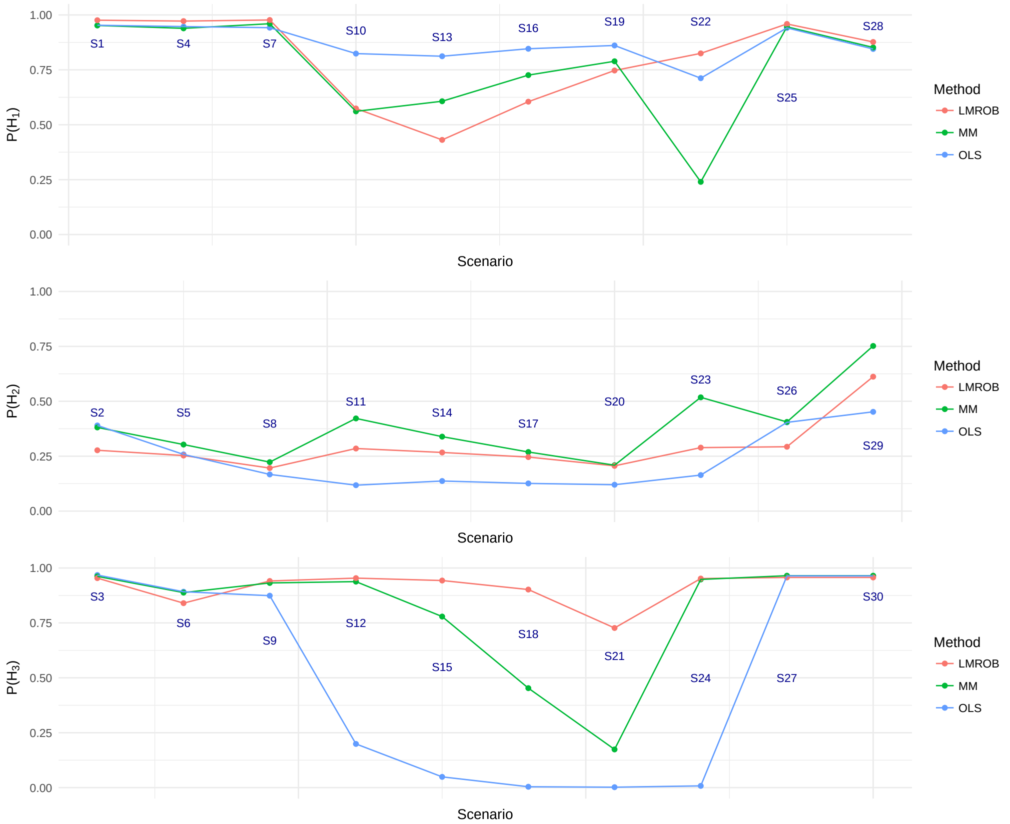


Figure 6: Proportion of data sets where the true hypothesis H_s obtained the highest PMP (results are not displayed for H_u given the focus of the paper)

It is clear that robust estimates should be preferred giving their better performance in terms of bias, coverage probabilities and proportion of right classifications. It should also be noted that the three measures are tightly related; better estimators have less bias and a better approximation of the residual variance which in the ends translates into more accurate information for the computation of the Bayes factor and the PMPs. Previous results allow us to recommend the use of LMROB estimation. It should be noted too, that the performance of LMROB it is not uniformly the best and that the others estimators may perform better in specific cases. However, since the true hypothesis is never known, LMROB still represents the best option.

7 Example: Life Satisfaction

In this Section we explore the relation between people's life satisfaction, people having what they want and people wanting what they have, using a data set retrieved from the psychology reproducibility project of The Open Science Framework (OSF). The data was collected by Seibel et al. (2016) in an attempt to replicate the original study by Larsen and McKibban (2008) "Is happiness having what you want, wanting what you have, or both?". The replication study followed the methodology used in the original paper and was performed in the Netherlands. A total of N=238 participants were considered in the study. The data and processing script were found in the archive of the project in the OSF website (https://osf.io/5dx4v/). For this example we included the constructs (based on self reported questionnaires) "Having what you want" (HW) and "Wanting what you have" (WH) as predictors for "Life Satisfaction" (LS). We also added an interaction term (HWWH) of the predictors, supposing that if both conditions are satisfied then the impact on the outcome variable would be even bigger. A summary of the data is provided in Table 3

Table 3: Summary statistics for the example data

	LS	WH	HW	HWWH
Minimum	1.000	-4.444	-6.333	-5.541
Maximum	7.000	1.925	1.935	2.381
Mean	4.862	0.000	0.000	0.000
Standard deviation	1.202	1.000	1.000	1.000

The regression model that we will use is:

$$\text{LS}_i = \beta_0 + \beta_1 \text{WH}_i + \beta_2 \text{HW}_i + \beta_3 \text{HWWH}_i + e_i, \quad \text{for } i = 1, ..., N. \tag{39}$$

From a theoretical perspective we want to investigate the following hypotheses:

- $H_1 :\ \beta_1 = \beta_2 = \beta_3 = 0$, this is, none of the predictors (WH,HW or HWWH) are relevant to predict participant's life satisfaction.
- $H_2 :\ \beta_3 > \beta_2 > \beta_1 > 0$, this is, the effect of the interaction term (HWWH) is larger than the individual effects, and at the same time, having what you want has a higher impact on life satisfaction than wanting what you have.
- $H_3 :\ \beta_1 > 0, \beta_2 > 0, \beta_3 > 0$, this is, The effect of all terms are positive (WH,HW and HWWH) but no specific ordering is expected in the coefficients.
- $H_u : \beta_1, \beta_2, \beta_3$, included as a fail hypothesis, that is, if all the previous hypotheses are wrong, then the data will favor this hypothesis.

Note that the previous hypotheses are equivalent to the ones presented in Equations 32-35 with a small adjustment in H_3 since we expect all coefficients to be larger than zero. As mentioned before, testing hypotheses in which the coefficients are mutually compared only makes sense if the predictors are measured in the same scale. We achieved this by standardizing each of the predictors to have a zero mean and standard deviation of one, as seen in Table 3. Subsequently, we proceed with the computation of the Posterior Model Probabilities PMPs with the robust BF_{LMROB}. The coefficients estimates and their variance-covariance matrix used as input for the BF_{LMROB} are:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} -0.107 \\ 0.212 \\ 0.569 \end{pmatrix} \quad , \quad \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \begin{bmatrix} 0.583 & 0.623 & -0.918 \\ 0.623 & 0.679 & -0.992 \\ -0.918 & -0.992 & 1.461 \end{bmatrix}.$$

Computing the AAFBF with the previous input and the competing hypotheses mentioned before, we obtained the results summarized in Table 4. Using the BF_{LMROB} we found strong evidence in support for H_3, this is that all coefficients are positive (with no specific ordering).

Table 4: Posterior Model Probabilities for the example data under the three different Bayes factors

	H_1	H_2	H_3	H_u
$PMP\ BF_{LMROB}$	0.024	0.064	0.891	0.021

8 Discussion

In this paper we showed, by means of a simulation study, the superiority of robust Bayes factors for IHT when compared to non-robust methods. We also showed that none of the methods is universally the best but the collected evidence suggests that the robust methods should be preferred over the non-robust ones. It is also important to note that the application of robust Bayes factors constitutes an easily implementable tool that should be used more often in the context of IHT for linear regression models (and potentially in other types of model where the AAFBF can be also applied).

Going into more detail about Bayes factors and their behaviour, we first see that (over all scenarios) the robust Bayes factors, BF_{LMROB} and BF_{MM}, perform better than the non-robust Bayes factor BF_{OLS}. We found particularly worrying that the BF_{OLS} performs extremely well in the "Null" hypothesis scenarios, and also performs extremely bad in the other hypotheses scenarios. The latter allow us to conclude that the BF_{OLS} is not a reliable option (compared to the other), specially if any of the model assumptions is violated. It should also be noted that the BF_{LMROB} over-performs the BF_{MM} (specially if violations are strongly violated) and should thus be preferred amongst the three estimation procedures.

We also found important to point that the robust methods investigated in this paper are just specific cases that have been used in the context of linear regression. However, several literature exist and many other approaches exist for robust estimation in different models. Other violations in the linear regression model, as well as combinations of violations, may already have specific procedures available to overcome them. For such situations we refer readers to Huber (1981) and Wilcox (2017).

Form a practical perspective, as shown in the example, using the robust Bayes factor is an easy an intuitive approach. Application of robust Bayes factors for IHT in the context of linear regression is thus encouraged. Moreover, under ideal conditions when all the assumptions are met, robust methods should perform as good as OLS estimate do in such situations, making them potentially a standard for research in such models.

As a final comment, it would be interesting to understand in detail how does LMROB performs the estimation, and which techniques and algorithms are involved in the package. In this sense, an open question in the paper is: What are the criteria that LMROB takes into account to use a specific type of estimation for the measures of scale and location in the context of linear regression coefficients? Understanding, researching and retrieving more information about this kind of criteria would be a valuable argument to promote the use of robust estimation procedures and the BF_{LMROB} specifically.

References

Draper, N. R., & Smith, H. (1998). *Applied regression analysis (wiley series in probability and statistics).* New York: John John Wiley & Sons, Ltd.

Gu, X., Hoijtink, H., & Mulder, J. (2017). Approximated adjusted fractional bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*. doi: 10.1111/bmsp.12110

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511–527. doi: 10.1037/met0000017

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists.* Boca Raton: Chapman and Hall/CRC.

Huber, P. J. (1981). *Robust statistics.* New York: John John Wiley & Sons, Ltd.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi: 10.1371/journal.pmed.0020124

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* New York: Springer. doi: 10.1007/978-1-4614-7138-7

Larsen, J., & McKibban, A. (2008). Is happiness having what you want, wanting what you have, or both? *Psychological Science*, *19*(4), 371-377. doi: 10.1111/j.1467-9280.2008.02095.x

Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006). *Robust statistics: Theory and methods.* New York: John John Wiley & Sons, Ltd.

Masicampo, E., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. doi: 10.1080/17470218.2012.711335

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276–1284. doi: 10.1037/0003-066x.44.10.1276

Ruckstuhl, A. (2016). *Robust fitting of parametric models based on m-estimation.* IDP Institute of Data Analyisis and Process Design, ZHAW Zurich University of Applied Sciences in Winterthur. Retrieved from https://www.ethz.ch/content/dam/ethz/special-interest/math/statistics/sfs/Education/Advanced%20Studies%20in%20Applied%20Statistics/course-material/robust-nonlinear/robstat16E.pdf

Seibel, L., Vermue, M., van Dooren, R., Kolorz, F. M., Cillessen, L., & Krause, R. W. (2016, Aug). *Replication of larsen & mckibban (psych science, 2008, exp. 2).* Open Science Framework. Retrieved from osf.io/5dx4v

Susanti, Y., Pratiwi, H., Sulistijowati, S., & Liana, T. (2014). M estimation, s estimation, and mm estimation in robust regression. *International Journal of Pure and Apllied Mathematics*, *91*(3). doi: 10.12732/ijpam.v91i3.7

Tukey, J. W. (1977). *Exploratory data analysis.* Massachusetts: Addison-Wesley.

van Rossum, M., van de Schoot, R., & Hoijtink, H. (2013). "is the hypothesis correct" or "is it not". *Methodology*, *9*(1), 13–22. doi: 10.1027/1614-2241/a000050

Venables, W., & Ripley, B. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer.

Wagenmakers, E.-J. (2007, 01). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*(5), 779–804. doi: 10.3758/BF03194105

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., ... Konis., K. (2017). robust: Port of the s+ "robust library" [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=robust (R package version 0.4-18)

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing (4th edition).* Cambridge, MA: Academic Press, an imprint of Elsevier.

Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach (upper level economics titles).* Mason (USA): South-Western College Pub.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, *15*(2), 642-656. doi: 10.1214/aos/1176350366

Appendix A - M estimation algorithm

1. Start with OLS estimation for the $\hat{\boldsymbol{\beta}}^{(0)}$ coefficients at Step $t = 0$, and use the adjusted MAD in Equation 18 to estimate the residual standard deviation $\hat{\sigma}_e^{(0)}$.

2. A. Compute the residuals at Step t using Equation 18

$$e_i^{(t)} = y_i - (\hat{\beta}_0^{(t)} + \sum_{p=1}^{P} \hat{\beta}_p^{(t)} X_{ip}) \quad , \quad \text{for } i = 1, 2, ..., N$$

   B. Compute the weights in Step t using Equation 21

$$w_i^{(t)} = \frac{\psi_M \left(e_i'^{(t)} \right)}{e_i'^{(t)}} \quad , \quad \text{for } i = 1, 2, ..., N.$$

3. Solve the system of equations given by the weighted residuals in equation 22 to estimate the coefficients $\hat{\boldsymbol{\beta}}^{(t+1)}$ in Step $t + 1$.

4. Compute the residuals for Step $t + 1$ using Equation 13

$$e_i^{(t+1)} = y_i - (\hat{\beta}_0^{(t+1)} + \sum_{p=1}^{P} \hat{\beta}_p^{(t+1)} X_{ip}) \quad , \quad \text{for } i = 1, 2, ..., N,$$

   and compute the estimate for the residual standard deviation in $t + 1$ using Equation 18

$$\hat{\sigma}_e^{(t+1)} = \frac{median \left| e_i^{(t+1)} - median \left(e_i^{(t+1)} \right) \right|}{0.6745} \quad , \quad \text{for } i = 1, 2, ..., N.$$

5. Compute the absolute percentual change for each coefficient in Step t and $t + 1$ as

$$AV_p = \left| \frac{\hat{\beta}_p^{(t+1)} - \hat{\beta}_p^{(t)}}{\hat{\beta}_p^{(t)}} \right| \quad , \quad \text{for: } p = 0, 1, 2, ..., p, ..., P$$

6. Repeat from Step 2.B if any $AV_p > C$ where C is a defined tolerance margin for the variation of coefficients between two iterations. If $AV_p <= C$ is satisfied for all P coefficients or $t + 1 > I$ (where I is the maximal amount of iterations), stop the algorithm and keep $\hat{\boldsymbol{\beta}}^{(t+1)}$ and $\hat{\sigma}_e^{(t+1)}$ as the M-estimates of the coefficients and the residual standard deviation. For the purpose of this paper we use a value of 100 iterations and an accuracy value of C=0.001 (default value).

Appendix B - MM estimation algorithm

1. Start with OLS estimation for the $\hat{\boldsymbol{\beta}}^{(0)}$ coefficients at Step $t = 0$, and use the adjusted MAD in Equation 18 to estimate the residual standard deviation $\hat{\sigma}_e^{(0)}$.

2. A. Compute the residuals at Step t using Equation 13

$$e_i^{(t)} = y_i - (\hat{\beta}_0^{(t)} + \sum_{p=1}^{P} \hat{\beta}_p^{(t)} X_{ip}) \quad , \quad \text{for } i = 1, 2, ..., N$$

   B. Compute the weights in Step t using Equation 21

$$w_i^{(t)} = \frac{\psi_M \left(e_i'^{(t)} \right)}{e_i'^{(t)}} \quad , \quad \text{for } i = 1, 2, ..., N.$$

3. Solve the system of equations given by the weighted residuals in Equation 22 to estimate the coefficients $\hat{\boldsymbol{\beta}}^{(t+1)}$ in Step $t + 1$.

4. Compute the residuals for Step $t + 1$ using Equation 13

$$e_i^{(t+1)} = y_i - (\hat{\beta}_0^{(t+1)} + \sum_{p=1}^{P} \hat{\beta}_p^{(t+1)} X_{ip}) \quad , \quad \text{for } i = 1, 2, ..., N$$

   and compute the S estimate for the residual standard deviation at Step $t + 1$ using the definition in Equation 23 (in this part the residuals of the S estimator are used)

$$\frac{1}{N} \sum_{i=1}^{N} \rho_S \left(\frac{e_i}{\hat{\sigma}_e^{(t+1)}} \right) = \delta.$$

5. Compute the absolute percentual change for each coefficient in Step t and $t + 1$ as

$$AV_p = \left| \frac{\hat{\beta}_p^{(t+1)} - \hat{\beta}_p^{(t)}}{\hat{\beta}_p^{(t)}} \right| \quad , \quad \text{for: } p = 0, 1, 2, ..., p, ..., P$$

6. Repeat from Step 2.B, if any $AV_p > C$ where C is a defined tolerance margin for the variation of coefficients between two iterations. If $AV_p <= C$ is satisfied for all P coefficients or $t+1 > I$ (where I is the maximal amount of iterations), stop the algorithm and keep $\hat{\boldsymbol{\beta}}^{(t+1)}$ and $\hat{\sigma}_e^{(t+1)}$ as the MM-estimates of the coefficients and the residual standard deviation. For the purpose of this paper we use a value of 100 iterations and an accuracy value of C=0.001 (default value).

Appendix C - Simulating multiple linear regression data

For any scenario in the simulation 1000 data sets are simulated considering a multiple regression model where 3 predictors $(X_{(1)}, X_{(2)}, X_{(3)})$ are sampled of a multivariate normal distribution:

$$\begin{pmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma} \right] \qquad ,$$

The sampling of the predictors is done using the "mrvnorm" function from the "MASS" package in R (Venables & Ripley, 2002) with (Σ) denoting the variance-covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.00 & 0.11 & 0.08 \\ 0.11 & 1.00 & 0.14 \\ 0.08 & 0.14 & 1.00 \end{bmatrix}$$

Given a set of sampled values of x_{i1}, x_{i2} and x_{i3}, the outcome variable is defined by the multiple regression model presented in Equation 1:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i, \quad \text{for } i = 1, 2, ..., N \quad , \tag{40}$$

where:

$$e_i \sim N(0, \sigma_e^2).$$

A general procedure to sample values of the outcome variable can be defined for any set of coefficients. For this, we consider the problem of obtaining an expression for σ_e^2 that depends on the values of β's, Σ and R^2. To do so, let us begin by expressing the variance of the outcome variable as follows:

$$Var(\mathbf{y}) = Var(\beta_0 + \sum_{j=1}^{P} \beta_j X_{(j)} + e)$$

Such expression can be reordered as:

$$Var(\mathbf{y}) = \sum_{j=1}^{P} \beta_j^2 Var(X_{(j)}) + 2 \sum_{j=1}^{P} \sum_{k=1}^{j-1} \beta_j \beta_k Cov(X_{(i)}, X_{(j)}) + \sigma_e^2.$$

Recalling that the R^2 expresses proportion of variance explained by the model over the total variance, we can define

$$R^2 = \frac{\sum_{j=1}^{P} \beta_j^2 Var(X_{(j)}) + 2 \sum_{j=1}^{P} \sum_{k=1}^{j-1} \beta_j \beta_k Cov(X_{(i)}, X_{(j)})}{\sum_{j=1}^{P} \beta_j^2 Var(X_{(j)}) + 2 \sum_{j=1}^{P} \sum_{k=1}^{j-1} \beta_j \beta_k Cov(X_{(i)}, X_{(j)}) + \sigma_e^2}.$$

Rearranging the terms and solving for the variance of the residuals we obtain:

$$\sigma_e^2 = \frac{(1 - R^2)}{R^2} \left(\sum_{j=1}^{P} \beta_j^2 Var(X_{(j)}) + 2 \sum_{j=1}^{P} \sum_{k=1}^{j-1} \beta_j \beta_k Cov(X_{(i)}, X_{(j)}) \right). \tag{41}$$

The R code for the simulation of data is available in `https://github.com/pereznic/robustBF`. The β's used in the simulations were chosen based in such way that they are in agreement with the hypotheses used in the simulation design. For Scenarios under the "Null" hypothesis we used the simplification $\sigma_e^2 = (1 - R^2)/R^2$.

Appendix D - Modeling Heteroscedasticity

For our simulation study we chose to focus on funnel shaped residuals as described in Section 2. Similarly to what was done in Appendix C, in this case we had to establish a sampling procedure for the residuals that was depending on the choice of the model parameters (coefficients values $\boldsymbol{\beta}$, correlation amongst the predictors Σ, explained proportion of variance R^2) but also depending on the specific observed value in the first variable $X_{(1)}$. To do so, we define the residuals as:

$$e_i \sim \mathcal{N}(0, \sigma_{ei}^2)$$

where:

$$log(\sigma_{ei}^2) = \alpha_0 + \alpha_1 x_{i1}.$$

Using the previous definition and considering that we want a model in which we can modify the degree of heteroscedasticity (which depend on α_1, we created a procedure to estimate the values of α_0 when all the other parameters have fixed values (funnel parameter α_1,coefficients values $\boldsymbol{\beta}$, correlation amongst the predictors Σ and explained proportion of variance R^2). The latter was done by simulating data with combinations of α_0 and α_1 and subsequently modeling α_0 as a function of the residual variance σ_e and α_1. Using some calculus and regression models, we obtained the following estimation:

$$\alpha_0 = \frac{log(\sigma_e)}{2} - \alpha_1^2$$

in our simulation the weak funnel behaviour is represented by setting α_1=0.2 and for the strong funnel behaviour we used α_1=0.8