Tests for Differences between Least Squares and Robust Regression Parameter Estimates and Related Topics

Tatiana A. Maravina

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

R. Douglas Martin, Chair

Eric W. Zivot

Peter D. Hoff

Program Authorized to Offer Degree:

Statistics

University of Washington

Abstract

Tests for Differences between Least Squares and Robust Regression Parameter Estimates and Related Topics

Tatiana A. Maravina

Chair of the Supervisory Committee:
Professor R. Douglas Martin
Department of Statistics

At the present time there is no well accepted test for comparing least squares and robust linear regression coefficient estimates. To fill this gap we propose and demonstrate the efficacy of two Wald-like statistical tests for the above purposes, using for robust regression the class of MM-estimators. The tests are designed to detect significant differences between least squares and robust estimates due to both inefficiency of least squares under fat-tailed non-normality and significantly larger biases of least squares relative to robust regression coefficient estimators under bias inducing distributions. The asymptotic normality of the test statistics is established and the finite sample level and power of the tests are evaluated by Monte Carlo, with the latter yielding promising results. The first part of our research focuses on the LS and robust regression slope estimators, both of which are consistent under skewed error distributions. A second part of the research focuses on intercept estimation, in which case there is a need to adjust for some bias in the robust MM-intercept estimator under skewed error distributions. An interesting by-product of our research is that use of the slowly re-descending Tukey bisquare loss function leads to better test performance than the rapidly re-descending min-max bias optimal loss function.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express sincere gratitude to the University of Washington's Department of Statistics for its extended long-term support, and especially to my advisor, Professor R. Douglas Martin, for his guidance and vast reserve of patience and knowledge. I also want to express appreciation to McKinley Capital Management and Dr. John B. Guerard, Jr. for their financial support during the early stages of my research. This thesis would have never been completed without the devotion and unconditional love of my parents, family and friends. Special thanks to my husband for his support and patience. Finally, I would like to thank the Applied Statistics Group at Boeing for allowing me as much flexibility as needed to work on my dissertation and for continuous inspiration and encouragement.

Dedication

To my Mom

1. Introduction

Linear regression models are typically fitted using least squares (LS) estimates of the coefficients, either ordinary least squares (OLS), weighted least squares (WLS), or generalized least squares (GLS). Each of these types of LS estimates are simple, widely available in software packages, and blessed by being best linear unbiased estimates (BLUE) under commonly made assumptions, where 'best' means a minimum covariance matrix[1] of the estimates. The latter insures that each individual BLUE has minimum variance. In addition, LS estimates are the best among both linear and nonlinear estimates when errors are normally distributed. However, it is well known that LS estimates for linear regression models are quite non-robust. In the data-oriented sense, they can be very adversely distorted by just a few outliers in a sample. In the statistical sense, LS estimates can suffer from a substantial loss of efficiency under deviations from normality, i.e., they can have much larger variances than minimum attainable variances. Furthermore, under some types of deviations from normality LS estimates will be biased even in large sample sizes (i.e., asymptotically).

Fortunately, a number of robust alternatives to LS estimates exist that suffer relatively little from severe inefficiency and bias. See for example the books by Huber (1981), Huber and Ronchetti (2009), Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Rousseeuw and Leroy (1987), Maronna, Martin, and Yohai (2006), and the references therein. Specific types of outlier-robust regression methods are implemented in commercial statistical software programs such as SAS and STATA, as well as in the open source R (R Core Team, 2012). Regression M-estimates of one form or another are perhaps the most widely available types.

Statistical inference methods for robust regression coefficients, such as robust t-tests, F-tests, robust R-squared and robust model selection criteria, have been available in the literature for many years. Nonetheless, one seldom sees research papers that report statistical inference results for the robust regression, and it is fair to say that the primary use of robust regression to date has been for diagnostic purposes. In this regard, Tukey (1979) stated long ago: "It is perfectly proper to use both classical and

[1] In the sense that the difference between a covariance matrix of any other unbiased estimator and the minimum covariance matrix is positive semi-definite.

robust/resistant methods routinely, and only worry when they differ enough to matter. But when they differ, you should think hard." This is good advice that leaves open the question of how much is "enough" in "when they differ enough", and so it is desirable to have a test statistic whose rejection region defines "enough". If such a test statistic has reliable level and adequate power, then acceptance of an appropriately defined null hypothesis would lead a user who routinely computes both LS and robust regressions to take comfort in the LS results. On the other hand, rejection of the null hypothesis would support reliance on the robust regression estimate and associated robust inferences. Unfortunately, there does not at present exist a well-accepted statistical test for determining whether or not a LS regression estimator differs significantly from a robust estimator. In this dissertation, we fill this gap for linear regression, for slope and intercepts separately.

The rest of the dissertation is organized as follows. Chapter 1 provides a brief introduction to the robust regression estimation theory, including MM estimator and its main properties. Chapter 2 proposes two tests, T1 and T2, for significant differences between the LS and robust MM estimators of the regression slopes, treating intercept as a nuisance parameter. Monte Carlo simulation study is performed to assess performance of these tests in finite samples. Chapter 3 presents further in-depth analysis of the T1 and T2 tests that explains certain peculiar behaviors discovered by Monte Carlo simulations in Chapter 2. Chapter 3 also compares T2 test to a third test, T3, which is an extension of the existing overall bias test of Yohai, Stahel, and Zamar (1991) implemented in the R robust library (Wang et al., 2012). Chapter 4 investigates behavior of the MM regression estimator under skewed fat-tailed errors, where due to skewness an MM estimator based on a symmetric loss function may be inefficient for the slopes (while still being consistent) and be biased for an intercept. A simple method for reducing MM intercept bias is studied. Chapter 5 comes back to the testing and discusses tests for differences between two intercept estimators.

1.1 Robust M Regression Estimation

We consider estimation in the linear regression model

$$y_i = \boldsymbol{x_i'\beta_0} + \epsilon_i, \qquad i = 1 \ldots n \tag{1.1}$$

under the assumption that the observed data $\boldsymbol{z_i'} = (y_i, \boldsymbol{x_i'}),\ i = 1, \cdots, n$, consists of independent and identically distributed (i.i.d.) random variables.

We assume that there is always an additional intercept term α_0 and that it is included in model (1.1) as part of the error term ϵ_i rather than as a component of $\boldsymbol{\beta_0}$. Specifically, it is assumed that the $\boldsymbol{x_i}$ do not contain a column of 1's. Then if the true model contains an intercept, then α_0 will capture it plus part of the error term. We note that α_0 is often viewed as a nuisance parameter. Define column vectors $\boldsymbol{x}^*$ and $\boldsymbol{\theta}$ as

$$\boldsymbol{x}^{*\prime} = (1, \boldsymbol{x}') \text{ and } \boldsymbol{\theta}' = (\alpha, \boldsymbol{\beta}') \tag{1.2}$$

Definition 1: Relative efficiency of a multi-dimensional unbiased parameter estimator $\widehat{\boldsymbol{\theta}}_\mathbf{2}$ with covariance matrix $\boldsymbol{V_2}$ relative to an unbiased estimator $\widehat{\boldsymbol{\theta}}_1$ with covariance matrix $\boldsymbol{V_1}$ can be defined as $REFF(\widehat{\boldsymbol{\theta}}_\mathbf{2}, \widehat{\boldsymbol{\theta}}_\mathbf{1}) \equiv \min_{\boldsymbol{c}\neq 0} \frac{\boldsymbol{c'V_1c}}{\boldsymbol{c'V_2c}} = \lambda_1(\boldsymbol{V_2^{-1}V_1})$ where $\lambda_1(\boldsymbol{M})$ denotes the largest eigenvalue of the matrix $\boldsymbol{M}$. The asymptotic (absolute) efficiency of $\widehat{\boldsymbol{\theta}}_\mathbf{2}$ is defined as $EFF(\widehat{\boldsymbol{\theta}}_\mathbf{2}) \equiv REFF(\widehat{\boldsymbol{\theta}}_\mathbf{2}, \widehat{\boldsymbol{\theta}}_\mathbf{1})$, where $\widehat{\boldsymbol{\theta}}_1$ is the MLE, and $\boldsymbol{V_2}$ and $\boldsymbol{V_1}$ are the corresponding asymptotic covariance matrices. In many situations $\boldsymbol{V_2} = a\boldsymbol{V_1}$ where a is a constant, and then the efficiency is simply $1/a$. See Maronna et al. (2006), pp. 70-71.

Definition 2: Asymptotic bias of a scalar estimator $\hat{\theta}_E$ is the difference between its' asymptotic value θ_E and the true parameter value θ_0: $Bias(\hat{\theta}_E) = \theta_E - \theta_0$.

Definition 3: Asymptotic bias of a multivariate estimator $\widehat{\boldsymbol{\theta}}_\boldsymbol{E}$ is often defined as $Bias(\boldsymbol{\theta_E}) = \sqrt{(\boldsymbol{\theta_E} - \boldsymbol{\theta_0})'\boldsymbol{V}_{x^*}(\boldsymbol{\theta_E} - \boldsymbol{\theta_0})}$ where $\boldsymbol{\theta_E}$ is the asymptotic value of an estimator, i.e. $\widehat{\boldsymbol{\theta}}_\boldsymbol{E} \rightarrow_p \boldsymbol{\theta_E}$ (where $\rightarrow_p$ denotes convergence in probability) and $\boldsymbol{\theta_0}$ is the true parameter value. See Section 5.9 in Maronna et al. (2006).

We consider the following distinct situations with respect to the joint distribution of the data and the corresponding behavior of the least squares estimator and robust regression MM-estimators that will be defined shortly in Section 1.4.

1. The $\boldsymbol{z_i'} = (y_i, \boldsymbol{x_i'}),\ i = 1, \cdots, n$, have a joint distribution

$$F_0(\boldsymbol{x}, y) = G(\boldsymbol{x})F_\epsilon(y - \boldsymbol{x}'\boldsymbol{\beta_0}). \quad (1.3)$$

where the factored form above assumes that for each i, $\boldsymbol{x_i}$ and ϵ_i are independent.

a. F_ϵ in model (1.3) is a normal distribution with mean α_0: The ordinary least squares (LS) estimator of $(\alpha_0, \boldsymbol{\beta_0'})$ is consistent and fully efficient, and the MM-estimators are consistent but inefficient with large sample efficiencies that can be set as close to 100% as one likes (use of 90% and 95% normal distribution efficiency is common).

b. F_ϵ in model (1.3) is a symmetric non-normal distribution with mean α_0, fat-tails and finite variance: The LS estimator is consistent but can have an efficiency arbitrarily close to zero, and the M-estimators are consistent and can have high efficiencies.

c. F_ϵ in model (1.3) is fat-tailed and asymmetric with non-zero mean $\alpha_0 = E(\epsilon)$: The LS estimator is consistent for $(\alpha_0, \boldsymbol{\beta_0'})$ but can have an efficiency arbitrarily close to zero, and the M-estimators of $\boldsymbol{\beta_0}$ are consistent and have high efficiency but M-intercept estimator is typically biased for α_0.

2. In order to allow for general types of $(y_i, \boldsymbol{x}_i)$ outliers that do not conform to (1.1) and (1.3) we take the common approach in robustness studies of using a broad family of mixture distributions

$$F(\boldsymbol{x}, y) = (1 - \gamma)F_0(\boldsymbol{x}, y) + \gamma H(\boldsymbol{x}, y) \quad (1.4)$$

where F_0 is given by (1.3), the mixing parameter γ is positive and often relatively small, e.g., in the range .01 to .1, and H is unrestricted. This family of models is motivated by the empirical evidence that most of the time the data are generated by the nominal distribution F_0 but with small probability γ the data comes from another distribution H that can generate a wide variety of outlier types. For this model the LS estimator of $\boldsymbol{\theta_0}$ can be not only highly inefficient but also badly biased for some outlier generating distributions H. So the goal in the context of this model is to obtain robust estimates of the

parameters $\boldsymbol{\theta}_\mathbf{0}' = (\alpha_0, \boldsymbol{\beta}_\mathbf{0}')$ of F_0 that have relatively small bias and high efficiency relative to the LS estimator.

An important class of robust estimators is the class of maximum-likelihood type M-estimators introduced by Huber (1964) for estimating location and by Huber (1973) for regression. A regression M-estimate of $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_c \left(\frac{y_i - \boldsymbol{x}_i^{*\prime}\boldsymbol{\theta}}{\hat{\sigma}} \right) \tag{1.5}$$

where $\hat{\sigma}$ is a robust scale estimate of the residuals and ρ_c is the loss function. The parameter c is a tuning parameter used to control the normal distribution efficiency of the estimate. With $\psi_c = \rho_c^{'}$ the resulting $\widehat{\boldsymbol{\theta}}$ satisfies the local minimum condition

$$\sum_{i=1}^{n} \boldsymbol{x}_\boldsymbol{i}^{*} \psi_c \left(\frac{y_i - \boldsymbol{x}_i^{*\prime}\widehat{\boldsymbol{\theta}}}{\hat{\sigma}} \right) = \mathbf{0}. \tag{1.6}$$

(Throughout the dissertation, a prime on a scalar-valued function, e.g. ρ_c' , denotes its derivative, otherwise it denotes transposition)

Solution to (1.6) may not be unique as (1.5) may have several local minima and one may wish to choose the global minimum as the MM-estimator.

Note that (1.6) can be re-written as

$$\widehat{\boldsymbol{\theta}} = \left[\sum_{i=1}^{n} w_i \boldsymbol{x}_\boldsymbol{i}^{*} \boldsymbol{x}_\boldsymbol{i}^{*\prime} \right]^{-1} \sum_{i=1}^{n} w_i \boldsymbol{x}_\boldsymbol{i}^{*} y_i \tag{1.7}$$

where $w_i = w_c\big((y_i - \boldsymbol{x}_i^{*\prime}\widehat{\boldsymbol{\theta}})/\hat{\sigma}\big)$ and w_c is the weight function

$$w_c(u) = \begin{cases} \dfrac{\psi_c(u)}{u} & \text{if } u \neq 0 \\ \rho''(0) & \text{if } u = 0 \end{cases} \tag{1.8}$$

Equation (1.7) provides an intuitive interpretation of the M-estimates as a weighted least-squares estimate with data dependent weights.

The LS estimate is a special case of (1.5) and (1.6) for which the loss function ρ_c is quadratic, ψ_c is linear and w_c is a constant. In this case the rapid quadratic growth rate of ρ_c and the linearity of ψ_c results in a lack of robustness of LS toward outliers. The least-absolute deviations (LAD) estimate is another special case in which ρ_c is the absolute value function. The well-known Huber loss function $\rho_{H,c}$ is a compromise between the LS and LAD estimate in which ρ_c is quadratic in a central region (-c,c) and is the absolute value function outside of this region. Note that both the LAD and Huber estimates are based on an unbounded loss function, albeit one that grows only linearly for large arguments. It turns out that in order to obtain bias robustness toward outliers one needs to use a bounded loss function ρ_c (Martin, Yohai, & Zamar, 1989). The following definition is used in various theoretical robustness studies, e.g., see Maronna et. al. (2006) and references therein.

Definition 4: A bounded ρ-function is a function $\rho(t)$ that is a continuous non-decreasing function of $|t|$, such that $\rho(0) = 0$, $\rho(\infty) = 1$ and $\rho(v) < 1$ implies that $\rho(u) < \rho(v)$ for $|u| < |v|$.

1.2 Optimal and Bisquare Loss Functions

Two choices of a bounded loss function that we describe and study in detail subsequently are the well-known Tukey bisquare function, and the optimal robust loss function of Yohai and Zamar (1997) or the very similar function of Svarc, Yohai, and Zamar (2002). The analytic expressions for the bisquare and optimal ρ- and ψ- functions are as follows:

- Bisquare

$$\rho_c(r) = \begin{cases} 1 - \left(1 - \left(\frac{r}{c}\right)^2\right)^3 & , \left|\frac{r}{c}\right| \leq 1 \\ 1 & , \left|\frac{r}{c}\right| > 1 \end{cases}$$

$$\psi_c(r) = \frac{6}{c^2} r \left(1 - \left(\frac{r}{c}\right)^2\right)^2 I[|r| \leq c]$$

These are the scaled bisquare ρ and ψ functions that are computed by the functions rho.weight(…, ips=2) and psi.weight(…, ips=2) in R robust library that satisfy scaling requirement in Definition 4, namely $\rho(\infty) = 1$. Figure 1, however, plots unscaled versions $\frac{c^2}{6}\rho_c(r)$ and $\frac{c^2}{6}\psi_c(r)$.

- Optimal (polynomial approximation)

The below ρ_c is a good piecewise polynomial approximation to an optimal bias robust M-estimate loss function discovered by Svarc et al. (2002) for the case of 95% normal distribution efficiency. lmRob and lmrob functions from R robust and robustbase libraries respectively (Wang et al., 2012; Rousseeuw et al., 2012) use this form for other normal distribution efficiencies even though the constants should be adjusted somewhat for efficiencies different than 95%.

$$\rho_c(r) = \begin{cases} 3.25k^2 & , \left|\frac{r}{k}\right| > 3 \\ k^2\left[1.792 + h_1\left(\frac{r}{k}\right)^2 + h_2\left(\frac{r}{k}\right)^4 + h_3\left(\frac{r}{k}\right)^6 + h_4\left(\frac{r}{k}\right)^8\right] & , 2 < \left|\frac{r}{k}\right| \leq 3 \\ 0.5r^2 & , \left|\frac{r}{k}\right| < 2 \end{cases}$$

$$\psi_c(r) = \begin{cases} 0 & , \left|\frac{r}{k}\right| > 3 \\ k\left[g_1\left(\frac{r}{k}\right) + g_2\left(\frac{r}{k}\right)^3 + g_3\left(\frac{r}{k}\right)^5 + g_4\left(\frac{r}{k}\right)^7\right] & , 2 < \left|\frac{r}{k}\right| \leq 3 \\ r & , \left|\frac{r}{k}\right| < 2 \end{cases}$$

where

$$k = \frac{c}{3}$$

$$g_1 = -1.944; \;\; h_1 = \frac{g_1}{2}$$

$$g_2 = 1.728; \;\; h_2 = \frac{g_2}{4}$$

$$g_3 = -0.312; \;\; h_3 = \frac{g_3}{6}$$

$$g_4 = 0.016; \;\; h_4 = \frac{g_4}{8}$$

The optimal robust loss functions minimize the maximum large sample bias over distribution families of the type (1.3) and (1.4) subject to constraining the normal distribution efficiency, e.g., to be a suitably high value such as 90% or 95%. The resulting small loss in normal distribution efficiency is a small insurance premium paid to obtain robustness toward outlier-generating non-normal distributions. Versions of the bisquare loss functions and approximating versions of the optimal robust loss functions are shown in the top row of Figure 1 for normal distribution efficiencies of 85%, 90%, 95% and 99%, and the corresponding psi-functions are shown in the bottom row. The optimal and bisquare weight functions are shown in Figure 2 for the same four normal distribution efficiencies. These are all of the rejection type, i.e., they have values of zero outside central regions (-*c*,*c*) where the choice of rejection points $\pm c$ determines the normal distribution efficiency.

The optimal functions in Figure 1 and Figure 2 look rather different than the bisquare functions in three important respects: (i) the optimal psi-functions remain nearly linear on a large fraction of their support interval (-*c*,*c*), (ii) the optimal psi-functions re-descend to zero faster than the bisquare functions, and (iii) the optimal psi-functions have rejection points $\pm c$ that are smaller than those for the bisquare function, the more so the higher the normal distribution efficiency. There is little discussion in the literature about these differences, and some practitioners may prefer to use the Tukey bisquare function based on its familiarity in spite of the theoretical justification of the optimal loss function. The difference between the two may be negligible in the context of using robust regression for exploratory data analysis and outlier diagnostics. However, the differences turn out to be important in hypothesis testing (see Chapters 2 and 3 as well as Koller & Stahel, 2011)

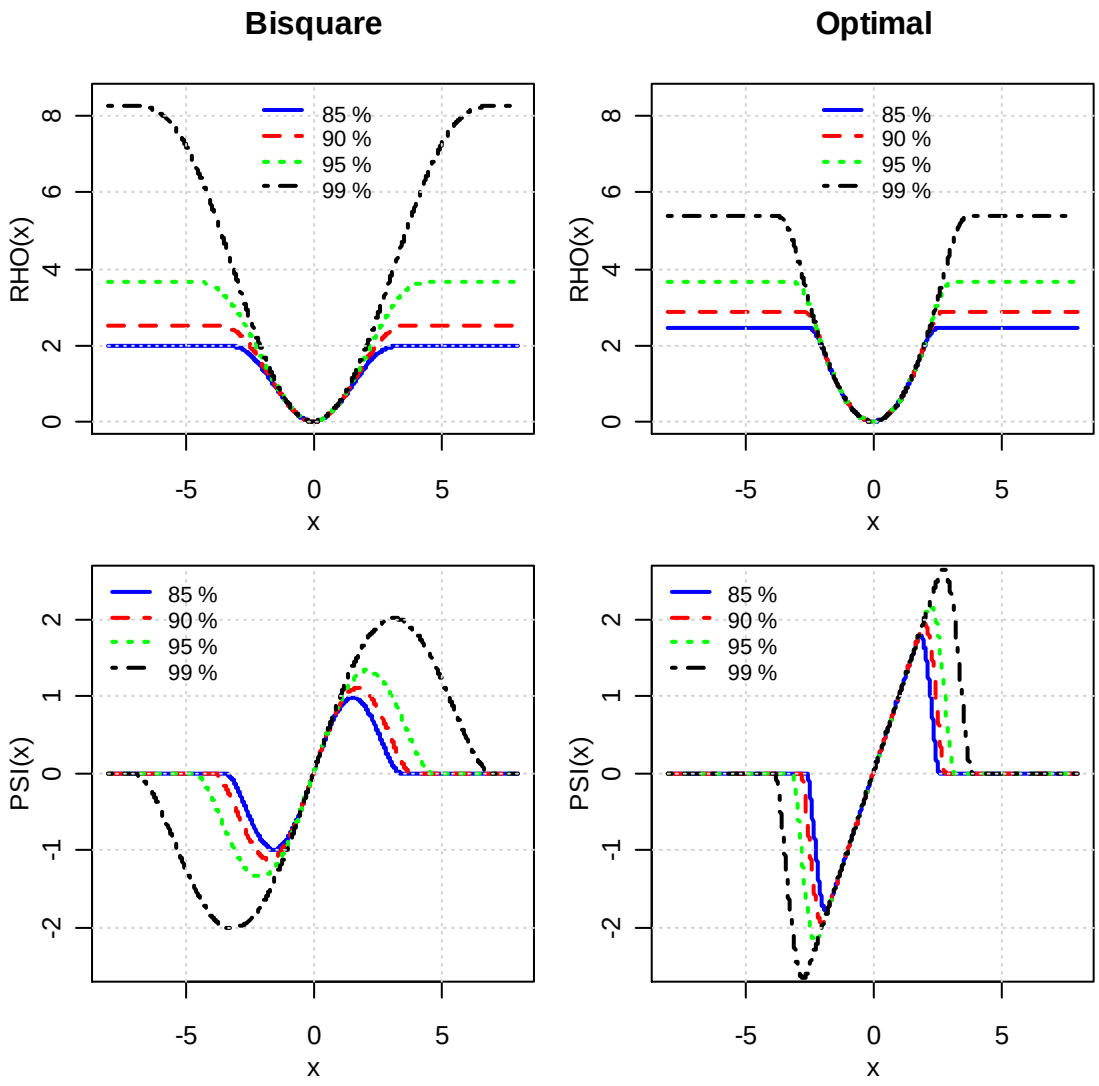

Figure 1. Bisquare (left) and optimal (right) rho- and psi- functions for four normal distribution efficiencies.

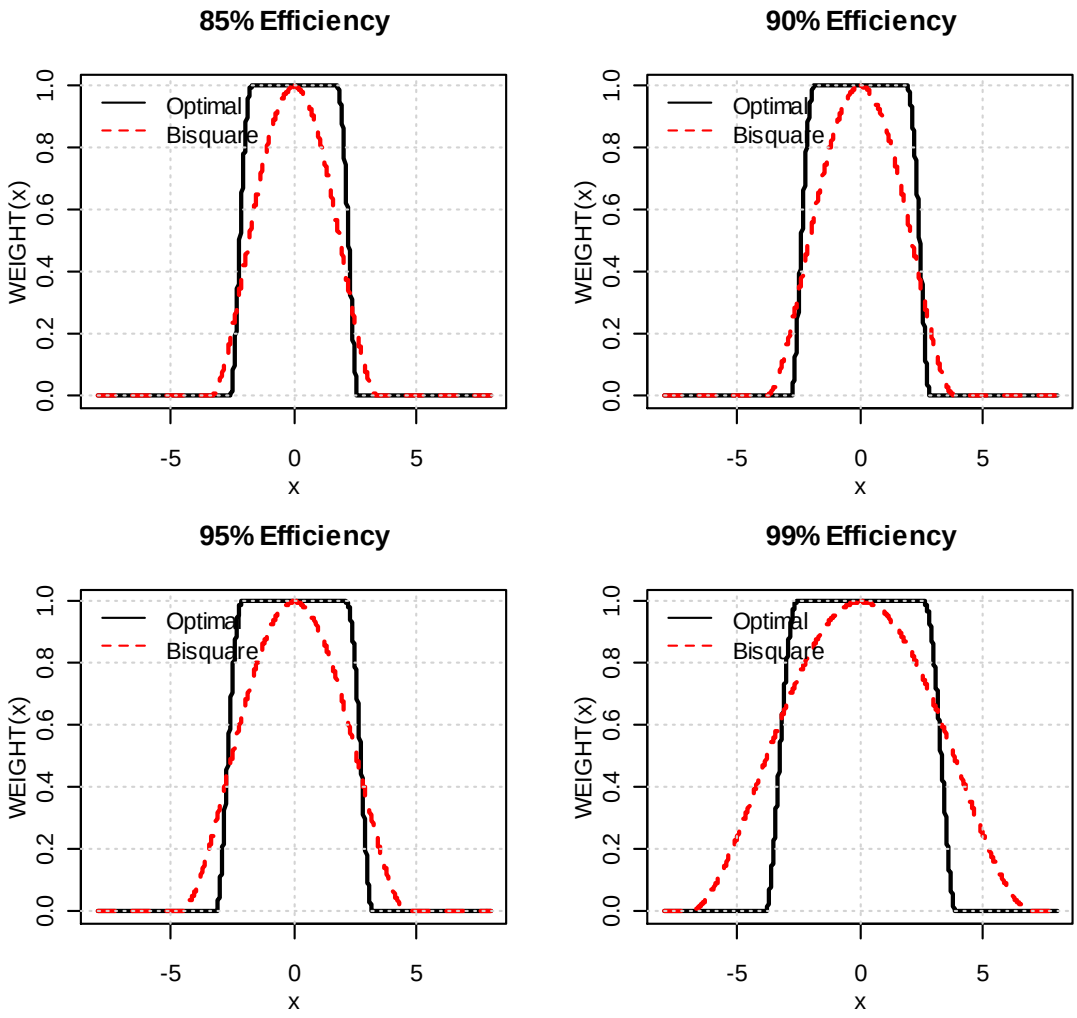

Figure 2. Bisquare and optimal weight functions (standardized to be equal to one at zero) for four normal distribution efficiencies.

For reader convenience, the values of constant c and corresponding fractions of residuals rejected under normality by bisquare and optimal (polynomial approximation) psi-functions are listed in Table 1 for the four normal distribution efficiencies 85%, 90%, 95% and 99%. Notice the fairly dramatic difference in the fraction of residuals rejected for the two psi-functions.

Table 1. *c* values and fractions of outliers rejected under normality for the bisquare and optimal loss functions for four normal distribution efficiencies.

		Normal Distribution Efficiency			
		85%	90%	95%	99%
Bisquare	c	3.44	3.88	4.68	7.04
	fraction of outliers (%)	0.058%	0.010%	0.0003%	$2 \cdot 10^{-10}$%
Optimal	c	2.604	2.832	3.18	3.87
	fraction of outliers (%)	0.921%	0.463%	0.147%	0.011%

1.3 The True Optimal Loss Function and its Polynomial Approximation

The actual optimal psi function discovered by Yohai and Zamar (1997) has the form

$$\psi_c(r) = sgn(r)\left(-\frac{\phi'(|r|) + c}{\phi(|r|)}\right)^+$$

where $\phi(r)$ is a standard normal density, and $t^+ = \max(t, 0)$ denotes the positive part of t. For $c = 0$ we have the LS estimate: $\psi_{LS}(r) = r$. See Section 5.9.1 in Maronna et al. (2006) and the original paper for more details.

The true optimal functions are compared to the polynomial approximations in Figure 3, which displays loss, psi- and weight functions for normal distribution efficiencies 85%, 90%, 95% and 99%. Due to symmetry the functions are only plotted on a positive line. The functions are normalized so that the maximum weight is equal to one. Figure 3 shows that unlike the polynomial approximation, the true optimal weight function decreases as residuals approach 0. This reveals an important feature of the true optimal estimator and raises a question of whether 'inliers' should be down-weighted in addition to any 'outliers'.

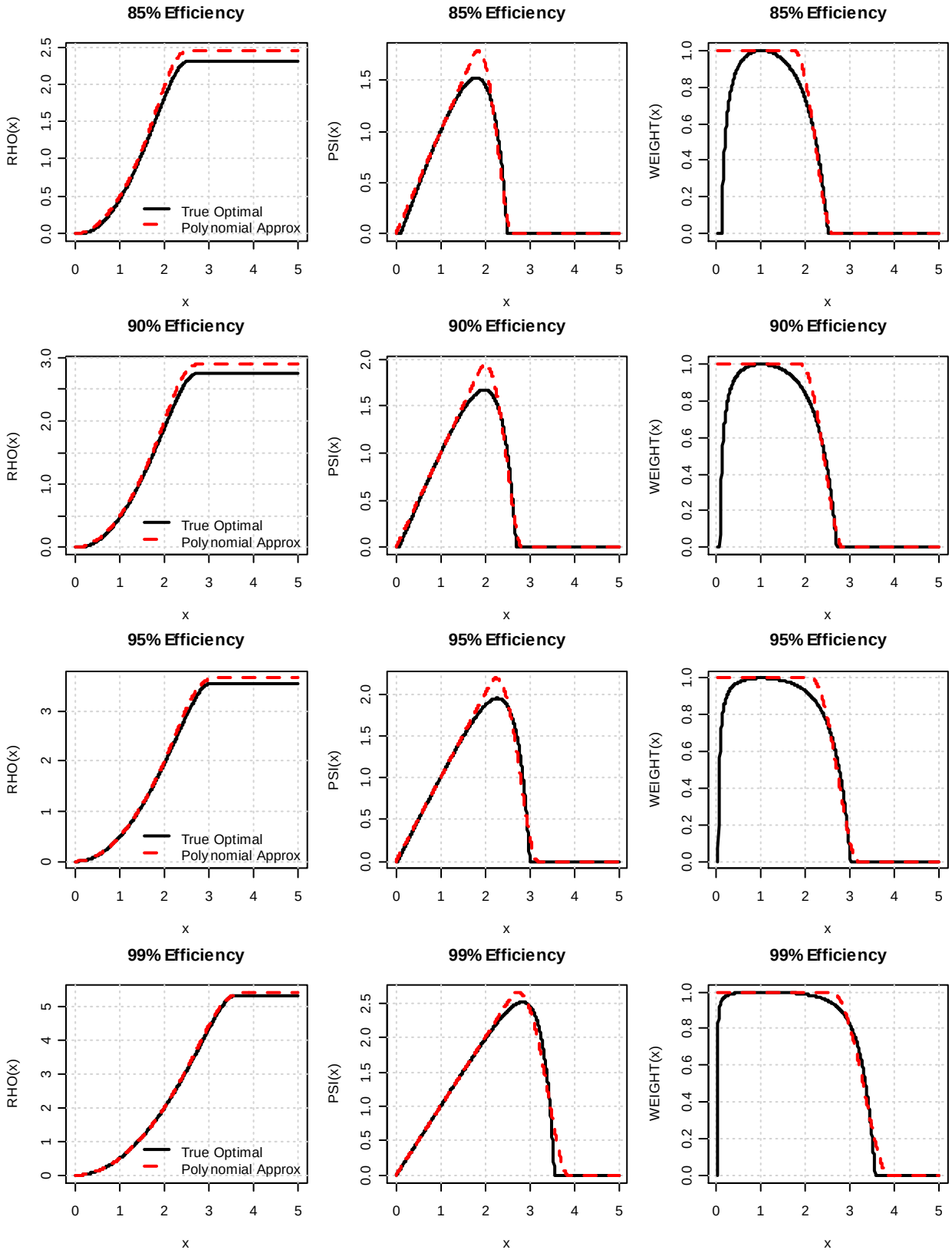

Figure 3. 95% efficiency polynomial approximations and exact optimal rho-, psi- and weight functions.

1.4 MM Regression Estimates and Initial S-Estimates

Definition 5: The breakdown point (BP) of an estimate is defined as the smallest fraction of contamination that can cause the estimator to take on values arbitrarily far from its value for outlier free data. For example, moving a single data value to $\pm\infty$ cause the sample mean to move to $\pm\infty$, i.e, BP = 0 for the sample mean. On the other hand, the sample median tolerates up to 50% of arbitrarily large outliers before it can be moved to $\pm\infty$, and, therefore BP = .5 for the sample median. The BP of the LS regression estimates is zero. See Hodges (1967) and Hampel (1971) for the introduction of the concept of breakdown point in robust statistics. Table 1 in Genton (2003) provides a nice summary of the breakdown point definitions available in the literature.

A special M-estimate, called an MM-estimate, that insures both high breakdown point (BP) and high user defined efficiency at normal distribution was proposed by Yohai (1987). An effective computational procedure for MM-estimates was developed in Yohai et al. (1991) and consists of the following key steps:

1) Compute an initial robust S-estimate $\widehat{\boldsymbol{\theta}}_0$ with high breakdown point, e.g. BP=0.5, but low normal distribution efficiency as follows:

   For any $\boldsymbol{\theta}$ let $s_c(\boldsymbol{\theta})$ be the solution of

$$\frac{1}{n-p}\sum_{i=1}^{n}\rho_c\left(\frac{y_i-\boldsymbol{x}_i^{*\prime}\boldsymbol{\theta}}{s}\right)=b\cdot\max_u\rho_c(u)$$

   The regression S-estimate of $\boldsymbol{\theta}$ is a value $\widehat{\boldsymbol{\theta}}_1$ that minimizes $s_{c_1}(\boldsymbol{\theta})$:

$$\widehat{\boldsymbol{\theta}}_1 = argmin_\theta\, s_{c_1}(\boldsymbol{\theta}) \qquad (1.9)$$

2) Compute a robust scale $\hat{\sigma}_1 = s_{c_1}(\widehat{\boldsymbol{\theta}}_1)$ based on the residuals in step 1.

3) The final MM-estimate $\widehat{\boldsymbol{\theta}}_2$ is then obtained as a local minimum of (1.5) nearest to $\widehat{\boldsymbol{\theta}}_1$, where the loss function is now ρ_{c_2} with $c_2 > c_1$ chosen to yield a user-specified 'high' normal distribution efficiency.

These S- and MM- estimates are regression, affine and scale equivariant[2].

In general, the breakdown point of an S-estimate $\widehat{\boldsymbol{\theta}}_1$ in (1.9) and of $\hat{\sigma}_1$ in step 2) is given by $BP = \min(b, 1-b)$. As we focus on the standard goal of achieving a "high" breakdown point, where "high" is taken to mean one-half, we set $b = 0.5$. By selecting c_1 such that $E_{F_\epsilon}\rho_{c_1}((\epsilon - \alpha_0)/\sigma) = b \cdot \max_u \rho_{c_1}(u)$ with F_ϵ being $N(\alpha_0, \sigma^2)$ one ensures consistency of $\hat{\sigma}_1$ for the standard deviation σ under normal distribution. For the bisquare loss function the corresponding value of c_1 is 1.548. For the optimal loss function the corresponding value of c_1 is 0.4047.

Unfortunately, the regression S-estimates defined by (1.9) with a smooth loss function cannot simultaneously have a high BP and high efficiency at normal distribution. Selection of the constants b and c_1 to achieve high BP leads to low efficiency. On the other hand, selection of the constants b and c_1 to yield high efficiency of the S-estimate results in poor breakdown point. For example, an initial estimate with the bisquare loss function ρ_{c_1} with $b = 0.5$ and $c_1 = 1.548$ has a BP=0.5, but normal distribution efficiency of only 0.287[3]. The final MM estimate inherits the breakdown point of an initial estimate because of the re-descending psi-function ψ_{c_2} and the fixed scale $\hat{\sigma}_1$. The efficiency is then improved by the choice $c_2 > c_1$.

Solution to (1.9) is traditionally found via random re-sampling algorithms. See, for example, Marazzi (1993), Salibian-Barrera and Yohai (2006) and Section 5.6 in Maronna et al. (2006). Once the initial estimates and scale are computed, the iteratively re-weighted least squares algorithm with fixed scale is used to compute the final MM estimate.

Going forward we will often drop c in the subscript of the loss and psi functions and for brevity write ρ_1 (ψ_1) and ρ_2 (ψ_2) to denote loss (psi-) functions for the initial S and final MM estimates respectively. In

---

[2] Estimate $\widehat{\boldsymbol{\theta}}$ is regression equivariant if $\widehat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y} + \boldsymbol{X\gamma}) = \widehat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y}) + \boldsymbol{\gamma}$ for all $\boldsymbol{\gamma} \in \boldsymbol{R^p}$.
Estimate $\widehat{\boldsymbol{\theta}}$ is scale equivariant if $\widehat{\boldsymbol{\theta}}(\boldsymbol{X}, \lambda\boldsymbol{y}) = \lambda\widehat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y})$ for all $\lambda \in \boldsymbol{R}$.
Estimate $\widehat{\boldsymbol{\theta}}$ is affine equivariant if $\widehat{\boldsymbol{\theta}}(\boldsymbol{XA}, \boldsymbol{y}) = \boldsymbol{A^{-1}}\widehat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y})$ for all nonsingular $p \times p$ matrices $\boldsymbol{A}$.

[3] The normal distribution efficiency of an initial estimate with the optimal loss function ρ_{c_1} with $b = 0.5$ and $c_1 = 0.4047$ is only 0.243.

addition to using subscripts 1 and 2 we will also use subscripts S and MM to denote initial S and final MM estimates.

1.5 Consistency and Asymptotic Normality Results

Various consistency and asymptotic normality results are available in the statistics literature for the S-estimators and MM-estimators described above. See, for example, Rousseeuw and Yohai (1984) for S-estimates and Yohai (1987) for MM-estimates under nominal distribution model $F_0(\boldsymbol{x}, y)$. See also Salibian-Barrera and Omelka (2010) for uniform consistency and asymptotic normality of the S- and MM-estimators over contamination neighborhoods of a nominal distribution model $F_0(\boldsymbol{x}, y)$. For the purposes of this dissertation we will make use of the recent general asymptotic results of Fasano, Maronna, Sued, and Yohai (2012) (henceforth referred to as FMSY). While we consider the i.i.d. assumption of Section 1.1 throughout, FMSY consider both linear and non-linear regression and show consistency and asymptotic normality of S- and MM-estimators under more general assumptions of $\{(\boldsymbol{x_i}, y_i), i \geq 1\}$ being identically distributed, but not necessarily independent. They still assume that for each i, $\boldsymbol{x_i}$ and ϵ_i are independent.

With regard to estimator consistency FMSY yield the following results:

1) The robust scale estimator $\hat{\sigma}_1$ is a consistent estimator of

$$\sigma = \min_{\boldsymbol{\theta}} s_1(\boldsymbol{\theta}) \tag{1.10}$$

with $s_1(\boldsymbol{\theta})$ defined as a solution to

$$E_F \rho_1\left(\frac{y - \boldsymbol{x}^{*\prime}\boldsymbol{\theta}}{s}\right) = b \cdot \max_u \rho_1(u) \tag{1.11}$$

2) The initial S-estimator $\widehat{\boldsymbol{\theta}}_S = (\hat{\alpha}_S, \widehat{\boldsymbol{\beta}}_{\boldsymbol{S}})$ is a consistent estimator of

$$\boldsymbol{\theta_S} = (\alpha_S, \boldsymbol{\beta_S'})' = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, s_1(\boldsymbol{\theta}) \tag{1.12}$$

It was shown by Rousseeuw and Yohai (1984) that S-estimate is in fact an M-estimate with the loss function ρ_1 and residual scale σ as defined in (1.10) and (1.11):

$$\boldsymbol{\theta_S} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, E_F \rho_1 \left(\frac{y - \boldsymbol{x}^{*\prime}\boldsymbol{\theta}}{\sigma} \right) \tag{1.13}$$

3) The final MM estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}} = (\hat{\alpha}_{MM}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}})$ is a consistent estimator of

$$\boldsymbol{\theta_{MM}} = (\alpha_{MM}, \boldsymbol{\beta'_{MM}})' = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, E_F \rho_2 \left(\frac{y - \boldsymbol{x}^{*\prime}\boldsymbol{\theta}}{\sigma} \right) \tag{1.14}$$

In general, the minimum in the above equations might be attained at more than one value of $\boldsymbol{\theta_S}$ and $\boldsymbol{\theta_{MM}}$. However, FMSY show that the solutions are unique under the condition that the error distribution is strongly unimodal[4] and that the loss functions ρ_1 and ρ_2 satisfy reasonable conditions that are met by the bisquare and optimal loss functions, as well as by the polynomial approximations to the optimal functions.

Under model (1.3) and rather general conditions discussed in Appendix A2 allowing asymmetry of the errors distribution F_ϵ, FMSY show that:

1) The S-slope estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}$ and the MM-slope estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$ are both consistent estimators to the true slope parameter $\boldsymbol{\beta_0}$, i.e. $\boldsymbol{\beta_S} = \boldsymbol{\beta_0}$ and $\boldsymbol{\beta_{MM}} = \boldsymbol{\beta_0}$

2) The S-intercept estimator $\hat{\alpha}_S$ and MM-intercept estimator $\hat{\alpha}_{MM}$ are consistent for α_S and α_{MM} defined by the centering equations

$$\alpha_S = \underset{t}{\operatorname{argmin}}\, s_0(t) = \underset{\mathrm{t}}{\operatorname{argmin}}\, E_{F_\epsilon} \rho_1 \left(\frac{\epsilon - t}{\sigma} \right) \tag{1.15}$$

[4] The density f is strongly unimodal if there exists a such that $f(t)$ is nondecreasing for $t < a$, nonincreasing for $t > a$, and f has a unique maximum at $t = a$.

$$\alpha_{MM} = \underset{\mathrm{t}}{\operatorname{argmin}}\, E_{F_\epsilon}\rho_2\left(\frac{\epsilon - t}{\sigma}\right) \tag{1.16}$$

where $\sigma = \min\ s_1(t)$ with $s_1(t)$ defined as a solution to

$$E_{F_\epsilon}\rho_1\left(\frac{\epsilon - t}{s(t)}\right) = b \tag{1.17}$$

3) Moreover, if F_ϵ is symmetric around α_0 then S- and MM-intercept estimators are consistent estimators of α_0, i.e. $\alpha_S = \alpha_0$ and $\alpha_{MM} = \alpha_0$. The asymptotic bias of the intercept estimators in the case of asymmetric error distributions F_ϵ will be discussed in detail in Chapter 4.

We want to note that the consistency results given by equations (1.10) through (1.14) are rather general and with certain regularity conditions also hold under contamination neighborhood models of type (1.4) (Salibian-Barrera & Omelka, 2010). In the presence of contamination, however, $\boldsymbol{\beta_S}$ as defined by (1.12) and $\boldsymbol{\beta_{MM}}$ as defined by (1.14) may not be equal to $\boldsymbol{\beta_0}$, i.e. may be biased.

Under model (1.3) with finite variances the least-squares (LS) estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{LS}} = \left(\hat{\alpha}_{LS}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}\right)$ is a consistent estimator of $\boldsymbol{\theta_{LS}} = (\alpha_{LS}, \boldsymbol{\beta_{LS}})$ with $\boldsymbol{\beta_{LS}} = \boldsymbol{\beta_0}$ and $\alpha_{LS} = E\epsilon$.

FMSY also provide general conditions that allow for asymmetry of F_ϵ under which the S-estimator and MM-estimator are asymptotically normally distributed. Their results are as follows.

Let

$$\boldsymbol{V}_{\boldsymbol{x}^*} \equiv E\left(\boldsymbol{x}^*\boldsymbol{x}^{*\prime}\right) = \begin{bmatrix} 1 & \boldsymbol{\mu}_{\boldsymbol{x}}' \\ \boldsymbol{\mu}_{\boldsymbol{x}} & \boldsymbol{E}(\boldsymbol{x}\boldsymbol{x}') \end{bmatrix} \tag{1.18}$$

where $\boldsymbol{\mu}_{\boldsymbol{x}} \equiv E(\boldsymbol{x})$ and $\boldsymbol{C}_{\boldsymbol{x}} \equiv \boldsymbol{Var}(\boldsymbol{x})$ denote the vector of expected values of $\boldsymbol{x}$ and covariance matrix of $\boldsymbol{x}$ respectively. Then

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{MM} - \boldsymbol{\theta_{MM}}\right) \rightarrow N(\mathbf{0}, \boldsymbol{\Sigma_{MM}}) \tag{1.19}$$

where

. $$\boldsymbol{\Sigma}_{MM} = \sigma^2 \tau \boldsymbol{V}_{\boldsymbol{x}^*}^{-\mathbf{1}} + \sigma^2 \vartheta_{MM} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{1.20}$$

. $$\tau = E_{F_\epsilon} \left\{ \left(\frac{\psi_2\left(\frac{u_{MM}}{\sigma}\right)}{E_{F_\epsilon} \psi_2'\left(\frac{u_{MM}}{\sigma}\right)} \right)^2 \right\} \tag{1.21}$$

$$u_{MM} = \epsilon - \alpha_{MM} \tag{1.22}$$

α_{MM} is defined by (1.16) and ϑ_{MM} has a complicated expression given in Chapter 4 which is focused on robust estimation of the intercept. It is important to note that when the error term distribution F_ϵ is symmetric then $\vartheta_{MM} = 0$ and (1.20) reduces to the well-known expression given by Yohai (1987). Similar results can be obtained for the S regression estimators.

Note that the asymptotic variance of the MM slope estimator is given by the lower right $p \times p$ block of the first term in (1.20). From (1.20) and the matrix inversion formula for a partitioned matrix we have

$$\boldsymbol{V}_{\boldsymbol{x}^*}^{-1} = \begin{bmatrix} 1 + \boldsymbol{\mu}_{\boldsymbol{x}}' \boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}} \boldsymbol{\mu}_{\boldsymbol{x}} & -\boldsymbol{\mu}_{\boldsymbol{x}}' \boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}} \\ -\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}} \boldsymbol{\mu}_{\boldsymbol{x}} & \boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}} \end{bmatrix} \tag{1.23}$$

It follows that under both symmetric and asymmetric error distributions the asymptotic variance of the MM slope estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$ is given by $\sigma^2 \tau \boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}$. It should be noted that FMSY obtain the above asymptotic variance result for the slopes under more general conditions than our i.i.d. assumption, namely under the condition that $\{\epsilon_i : i \geq 1\}$ is stationary and ergodic and that $\{x_i, i \geq 1\}$ are i.i.d. and independent of $\{\epsilon_i : i \geq 1\}$.

Under model (1.3) the LS estimator is also asymptotically normal with symmetric or asymmetric F_ϵ with finite second moment:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{LS} - \boldsymbol{\theta}_{\boldsymbol{LS}}\right) \rightarrow N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{LS}} = var(\epsilon) \boldsymbol{V}_{\boldsymbol{x}^*}^{-\mathbf{1}}) \tag{1.24}$$

1.5.1 MM-Estimator Efficiency at Normal Distributions

As follows from (1.19) the asymptotic variance of the MM-estimator under model (1.3) with F_ϵ being a $N(\alpha_0, \sigma^2)$ distribution is

$$\boldsymbol{V_{MM}} = \tau\sigma^2\boldsymbol{V_{x^*}^{-1}} \tag{1.25}$$

where $\boldsymbol{V_{x^*}^{-1}}$ is given in (1.23) and by the choice of the tuning constant c_1 the initial scale estimate $\hat{\sigma}_1$ converges to σ, standard deviation of the error term ϵ.

The asymptotic covariance matrix of the LS estimator under $N\ (\alpha_0, \sigma^2)$ errors is

$$\boldsymbol{V_{LS}} = \sigma^2\boldsymbol{V_{x^*}^{-1}} \tag{1.26}$$

Since the asymptotic covariance matrices of the MM-estimator and LS estimator differ only by the scalar factor τ, the large sample normal distribution efficiency of the MM-estimate is simply

$$EFF = EFF\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}\right) = \tau^{-1} = \frac{\left(E\psi_2'(u)\right)^2}{E\psi_2^2(u)} \tag{1.27}$$

where u is a standard normal random variable. Note that under normality this results in the convenient relationship that we will use in Chapter 2:

$$\boldsymbol{V_{LS}} = EFF \cdot \boldsymbol{V_{MM}}. \tag{1.28}$$

1.5.2 Standard Errors

In order to obtain finite sample approximations to the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$ one needs to compute estimates of τ, σ and $\boldsymbol{C_x}$ from (1.19). We use a method of doing so proposed by Yohai et al. (1991) and implemented in the function lmRob in the R robust package.

We report MM standard errors as returned by the lmRob function in R robust package. In particular, the three components of the covariance matrix, $\sigma^2\tau\boldsymbol{C}_x^{-1}$, are estimated as follows.

The scale parameter σ is estimated by the initial scale estimate $\hat{\sigma}_1$.

The parameter τ is estimated by

$$\hat{\tau} = \frac{n}{n-p} ave_i \left\{ \left(\frac{\psi_2(r_i/\hat{\sigma}_1)}{ave_j \psi_2'(r_j/\hat{\sigma}_1)} \right)^2 \right\} \tag{1.29}$$

where r_i are residuals from the final MM fit and the factor $\frac{1}{n-p}$ is used to recapture the classical formula for OLS with $\psi(u) = u$. See Section 7.6 in Huber and Ronchetti (2009) or Chapter 7 of Hampel et al. (1986).

Yohai et al. (1991) proposed the following robust estimate of $\boldsymbol{V}_{\boldsymbol{x}^*}$:

$$\widehat{\boldsymbol{V}}_{\boldsymbol{x}^*} = \frac{ave_i\{\boldsymbol{x}_i^* \boldsymbol{x}_i^{*\prime} w_i\}}{ave_i\, w_i} \tag{1.30}$$

The robust weights w_i are needed to down-weight the influence of high-leverage $\boldsymbol{x}_i$ outliers when estimating the covariance matrix $\boldsymbol{V}_{\boldsymbol{x}}$. lmRob uses robust weights corresponding to the initial S-estimates. $\widehat{\boldsymbol{C}}_x^{-1}$ is the corresponding subset of the $\widehat{\boldsymbol{V}}_{\boldsymbol{x}^*}^{-1}$. The $\widehat{\boldsymbol{V}}_{\boldsymbol{x}^*}$ is a consistent estimator of $\boldsymbol{V}_{\boldsymbol{x}^*}$. Thus, by Slutsky theorem, $\widehat{\boldsymbol{V}}_{\boldsymbol{x}^*}^{-\boldsymbol{1}}$ is a consistent estimator of $\boldsymbol{V}_{\boldsymbol{x}^*}^{-\boldsymbol{1}}$ and, consequently, $\widehat{\boldsymbol{C}}_x^{-1}$ is a consistent estimator of $\boldsymbol{C}_{\boldsymbol{x}}^{-\boldsymbol{1}}$.

2. Tests for Differences between Least Squares and Robust MM Regression Slope Coefficients

2.1 Introduction

In this chapter, we propose and study the performance of two Wald-like tests of differences between LS and robust MM slope estimators in a linear regression using composite null and alternative hypotheses. The null hypothesis for the first test (T1) is that regression errors are normally distributed. The alternative hypothesis for T1 consists of outlier-generating non-normal error distributions, as well as of more general types of bias-inducing joint distributions for predictors and response variables. The null hypothesis for the second test (T2) consists of normal and outlier-generating non-normal error distributions. The composite alternative for T2 consists of general types of bias-inducing joint distributions for predictors and response variables. Rejection of the null hypothesis for T1 can occur due to inefficiency of LS estimator alone, as well as due to both inefficiency and bias. Rejection of the null hypothesis for T2 occurs when the LS estimator has a sufficiently larger bias than the robust estimator.

The remainder of the chapter is organized as follows. Section 2.2 presents the test statistics and discusses the families of the null and alternative hypothesis distributions. Analytical power for the first test, T1, under some of the alternatives is given in Section 2.3. Monte Carlo simulations of validity of level and power for both tests are presented in Section 2.4 for the case of a simple linear regression. Univariate and multivariate empirical examples are presented in Section 2.5. Section 2.6 concludes the chapter by summarizing the main results and pointing out several open questions. Most of the technical details are deferred to Appendix A2.

2.2 Test Statistics, Null Hypotheses and Alternatives

In this section we propose two tests T1 and T2 for determining whether or not there is a significant difference between the LS estimate and an MM-estimate of the type described in Chapter 1. The tests are motivated by the models (1.3) and (1.4) for the distribution of the data $(y_i, \boldsymbol{x}_i)$. Large sample levels of the

tests T1 and T2 are based on asymptotic normal distribution results that we present. Since the intercept in the linear regression model is often a nuisance parameter of no great importance, we defer testing for differences between LS and MM-estimates of the intercept to Chapter 5 and focus on testing for differences in LS and MM-estimates of slopes vector $\boldsymbol{\beta}$.

2.2.1 Test T1 and its Null and Alternative Hypotheses

The test T1 is designed to test the following composite hypothesis H1 versus one of the following two alternatives K1 and K2:

H1: The distribution of the data $(y_i, \boldsymbol{x}_i)$ is given by (1.3) with F_ϵ a normal distribution.

K1: The distribution of the data $(y_i, \boldsymbol{x}_i)$ is given by (1.3) with F_ϵ a symmetric or skewed non-normal distribution

K2: The distribution of the data $(y_i, \boldsymbol{x}_i)$ is given by (1.4) with a $\boldsymbol{\beta}$-bias inducing distribution $H(\boldsymbol{x}, y)$.

Under K1 the LS estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}$ is an unbiased estimator of $\boldsymbol{\beta_0}$ but it can be highly inefficient for fat-tailed F_ϵ. We note that many distributions (1.4) under K2 result in bias for both the LS estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}$ and the MM-estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$. By virtue of the design of the latter its bias will typically be smaller than that of the former, sometimes considerably so.

A motivation for the test T1 that we introduce below may be found in the case of estimating the parameter β in a simple linear regression. Let $\hat{\beta}_{LS}$ and $\hat{\beta}_{MM}$ be LS and robust MM estimators of β in finite sample sizes. Under H1 $\hat{\beta}_{LS}$ is fully efficient but the robust estimator is by design somewhat inefficient. It is known that the efficiency of an inefficient estimator is equal to the squared correlation between the inefficient estimator and an efficient estimator (see, for example, Lehmann and Casella (1998), Theorem 4.8). Thus for $\hat{\beta}_{LS}$ and $\hat{\beta}_{MM}$ under the normality assumption of H1 we have:

$$\rho^2_{MM,LS} = EFF = \frac{Var(\hat{\beta}_{LS})}{Var(\hat{\beta}_{MM})} \tag{2.1}$$

It follows that under H1:

$$
\begin{aligned}
Var(\hat{\beta}_{LS} - \hat{\beta}_{MM}) &= Var(\hat{\beta}_{LS}) - 2cov(\hat{\beta}_{LS}, \hat{\beta}_{MM}) + Var(\hat{\beta}_{MM}) \\
&= Var(\hat{\beta}_{LS}) - 2\rho_{MM,LS}\sqrt{Var(\hat{\beta}_{LS})Var(\hat{\beta}_{MM})} + Var(\hat{\beta}_{MM}) \\
&= \rho^2_{MM,LS}Var(\hat{\beta}_{MM}) - 2\rho_{MM,LS}\sqrt{\rho^2_{MM,LS}Var(\hat{\beta}_{MM})Var(\hat{\beta}_{MM})} + Var(\hat{\beta}_{MM}) \\
&= \rho^2_{MM,LS}Var(\hat{\beta}_{MM}) - 2\rho^2_{MM,LS}Var(\hat{\beta}_{MM}) + Var(\hat{\beta}_{MM}) \\
&= (1 - \rho^2_{MM,LS})Var(\hat{\beta}_{MM}) \\
&= (1 - EFF)Var(\hat{\beta}_{MM}).
\end{aligned}
$$

In view of (1.28) the above expression may be written in the following alternative form:

$$Var(\hat{\beta}_{LS} - \hat{\beta}_{MM}) = Var(\hat{\beta}_{MM}) - Var(\hat{\beta}_{LS}).$$

A multi-parameter large-sample version of the above result was obtained by Hausman (1978) in his classic paper on specification tests.[5] Hausman's Corollary 2.6 to Lemma 2.1 states that the asymptotic covariance matrix of the difference between two consistent and asymptotically normal estimators, one of which is asymptotically efficient and the other is inefficient, is equal to the covariance matrix of the inefficient estimator minus the minimum covariance matrix. Thus in our case under H1 we have the following asymptotic covariance matrices relationship:

$$Var(\widehat{\boldsymbol{\beta}}_{MM} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}) = \boldsymbol{V_{MM}} - \boldsymbol{V_{LS}}. \tag{2.2}$$

In view of the asymptotic result (1.28) we have

$$\boldsymbol{V_{diff}} \equiv \boldsymbol{V_{MM}} - \boldsymbol{V_{LS}} = (1 - EFF)\boldsymbol{V_{MM}} = (1 - EFF)\tau\sigma^2\boldsymbol{C_x^{-1}} \tag{2.3}$$

Note that (2.3) holds only under H1 because in that case LS is fully efficient and the M-estimator is inefficient. A result analogous to (2.2), namely $Var(\widehat{\boldsymbol{\beta}}_{MM} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}) = \boldsymbol{V}_{LS} - \boldsymbol{V_{MM}}$, will hold when the MM-estimator is a maximum likelihood estimator (MLE) for a non-normal errors distribution and is therefore

[5] We thank Professor Eric Zivot for pointing out this reference.

asymptotically efficient, but the LS estimator is inefficient. Since there is seldom an obvious choice of non-normal distribution MLE to use, we do not pursue this possibility.

Hausman's results show that asymptotically $\boldsymbol{V_{MM} - V_{LS}}$ is non-negative definite under normality. However, this result does not hold under non-normal errors, and furthermore positive semi-definiteness of the finite sample estimate of the form $\widehat{\widehat{\boldsymbol{V}}}_{diff} = \widehat{\boldsymbol{V}}_{MM} - \widehat{\boldsymbol{V}}_{LS}$ is not guaranteed even under a normal errors distribution. However, since EFF is less than one, the estimate $\widehat{\boldsymbol{V}}_{diff} = (1 - EFF)\widehat{\boldsymbol{V}}_{\boldsymbol{MM}}$ is positive definite in the usual situation where the estimate $\widehat{\boldsymbol{V}}_{\boldsymbol{MM}}$ is positive definite[6].

By combining LS and MM regression coefficient estimates with a covariance matrix estimate $\widehat{\boldsymbol{V}}_{MM}$ and specified normal distribution efficiency EFF of the MM-estimator, one can construct two types of test statistics. The results (2.2) and (2.3) suggest the following:

1) A joint test statistic for any subset of *K* coefficients:

$$\begin{aligned} T_{1K} &= \left(\widehat{\boldsymbol{\beta}}_{MM}^{(K)} - \widehat{\boldsymbol{\beta}}_{LS}^{(K)}\right)' \left(\frac{(1-EFF)}{n}\widehat{\boldsymbol{V}}_{MM}^{(K)}\right)^{-1} \left(\widehat{\boldsymbol{\beta}}_{MM}^{(K)} - \widehat{\boldsymbol{\beta}}_{LS}^{(K)}\right) \\ &= \left(\widehat{\boldsymbol{\beta}}_{MM}^{(K)} - \widehat{\boldsymbol{\beta}}_{LS}^{(K)}\right)' \left(\frac{(1-EFF)}{n}\hat{\tau}\hat{\sigma}_1{}^2\widehat{\boldsymbol{C}}_{x,(K)}^{-1}\right)^{-1} \left(\widehat{\boldsymbol{\beta}}_{MM}^{(K)} - \widehat{\boldsymbol{\beta}}_{LS}^{(K)}\right) \end{aligned} \tag{2.4}$$

where $\widehat{\boldsymbol{\beta}}_{MM}^{(K)}$, $\widehat{\boldsymbol{\beta}}_{LS}^{(K)}$, $\widehat{\boldsymbol{V}}_{MM}^{(K)}$ and $\widehat{\boldsymbol{C}}_{x,(K)}^{-1}$ are the corresponding subsets of $\widehat{\boldsymbol{\beta}}_{MM}$, $\widehat{\boldsymbol{\beta}}_{LS}$, $\widehat{\boldsymbol{V}}_{MM}$ and $\widehat{\boldsymbol{C}}_{\boldsymbol{x}}^{-\boldsymbol{1}}$.

2) A test statistic for any individual coefficient:

$$T_{1i} = \frac{\hat{\beta}_{MM,i} - \hat{\beta}_{LS,i}}{\sqrt{1-EFF}\cdot se\left(\hat{\beta}_{MM,i}\right)} \tag{2.5}$$

where

$$se\left(\hat{\beta}_{MM,i}\right) = \sqrt{\frac{1}{n}\hat{\tau}\hat{\sigma}_1{}^2\hat{C}_{x,ii}^{-1}} \tag{2.6}$$

[6] In principal one might also use the estimate $\widehat{\boldsymbol{V}}_{diff} = (EFF^{-1} - 1)\widehat{\boldsymbol{V}}_{LS}$. While this estimate should result in decent accuracy of level in finite sample sizes, it is likely that it will result in lower power under non-normal alternatives of type K1 or K2.

with $\hat{C}_{x,ii}^{-1}$ equal to the i-th diagonal element of $\widehat{\boldsymbol{C}}_{\boldsymbol{x}}^{-\mathbf{1}}$.

It is expected that under H1 the statistic T_{1K} will have approximately a chi-squared distribution with *K* degrees of freedom and the statistic T_{1i} will have approximately a standard normal distribution. The extent to which use of such an approximation is valid is explored in Section 2.4.

Note that rejection of T1 does not tell us whether rejection occurred due to inefficiency of LS under K1, or bias of LS being larger than that of the MM estimator under K2, or both inefficiency and bias.

2.2.2 Test T2 and its Null and Alternative Hypotheses

In view of the above uncertainty as to the cause of rejection when using T1, it is desirable to work with a null hypothesis that allows skewed non-normality, for which neither LS nor MM-estimators are biased, and an alternative hypothesis that allows bias in both estimators. Thus test T2 is designed to test the previous K2 alternative versus the null hypothesis H2:

H2: The distribution of the data $(y_i, \boldsymbol{x}_i)$ is given by (1.3) with a normal or possibly skewed non-normal error term distribution F_ϵ.

K2: The distribution of the data $(y_i, \boldsymbol{x}_i)$ is given by (1.4) with a $\boldsymbol{\beta}$-bias inducing distribution $H(\boldsymbol{x}, y)$.

The idea is that under K2 the LS bias will typically be larger than that of the MM-estimator and if this difference is large enough it should be reliably detectable.

We show in Appendix A2 that under H2 an asymptotic normality argument yields

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{LS} - \widehat{\boldsymbol{\beta}}_{MM}\right) \rightarrow \boldsymbol{N}(\mathbf{0}, \delta_{LS,MM}^2 \boldsymbol{C}_x^{-1}) \tag{2.7}$$

where $\boldsymbol{C}_{\boldsymbol{x}} = Var(\boldsymbol{x})$ is positive definite and

$$\delta_{LS,MM}^2 = E\left\{\left(u_{LS} - \frac{\sigma\psi_2\left(\frac{u_{MM}}{\sigma}\right)}{E\psi_2'\left(\frac{u_{MM}}{\sigma}\right)}\right)^2\right\} \tag{2.8}$$

with u_{MM} defined in (1.22) and

$$u_{LS} = \epsilon - \alpha_{LS} = \epsilon - E\epsilon \tag{2.9}$$

We focus for a moment on the expression under the more restrictive null hypothesis H1 of Section 2.2.1, in which case $u_{LS} = u_{MM} = u$, where u is a normal random variable with mean 0 and standard deviation σ. It is easy to show that $E\left(u\frac{\sigma\psi_2(u/\sigma)}{E\psi_2'(u/\sigma)}\right) = \sigma^2$ under H1 and with τ defined by (1.21) basic algebra yields

$$\delta_{LS,MM}^2 \equiv E\left\{\left(u - \frac{\sigma\psi_2\left(\frac{u}{\sigma}\right)}{E\psi_2'\left(\frac{u}{\sigma}\right)}\right)^2\right\} = (\tau - 1)\sigma^2. \tag{2.10}$$

Then use of (1.27) gives

$$Var\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}\right) = \delta_{LS,MM}^2\, \boldsymbol{C_x^{-1}} = (1 - \tau^{-1})\tau\sigma^2\boldsymbol{C_x^{-1}} = (1 - EFF)\boldsymbol{V_{MM}}.$$

Thus, under H1 the asymptotic covariance in (2.7) is the same as (2.3). However this is not the case under H2.

We propose combining LS and MM regression coefficient estimates with estimates $\widehat{\delta_{LS,MM}^2}$ and $\widehat{\boldsymbol{C}}_{\boldsymbol{x}}$ to obtain the following two types of test statistics:

1) A joint test statistic for any subset of *K* coefficients:

$$T_{2K} = \left(\widehat{\boldsymbol{\beta}}_{MM}^{(K)} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}^{(K)}\right)' \left(\frac{\widehat{\delta_{LS,MM}^2}}{n}\,\widehat{\boldsymbol{C}}_{x,(K)}^{-1}\right)^{-1} \left(\widehat{\boldsymbol{\beta}}_{MM}^{(K)} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}^{(K)}\right) \tag{2.11}$$

2) A test statistic for any individual coefficient:

$$T_{2i} = \frac{\hat{\beta}_{MM,i} - \hat{\beta}_{LS,i}}{\sqrt{\frac{1}{n}\widehat{\delta_{LS,MM}^2}\,\hat{C}_{x,ii}^{-1}}} \tag{2.12}$$

Under H2 the statistic T_{2K} will have an approximate chi-square distribution with *K* degrees of freedom and T_{2i} will have an approximate standard normal distribution.

We use a robust estimate of $\boldsymbol{C_x}$ proposed by Yohai et al. (1991) and described in Chapter 1. We estimate $\delta^2_{LS,MM}$ as

$$\widehat{\delta^2_{LS,MM}} = ave_i\left(r_i^{LS} - \frac{\hat{\sigma}_1\psi_2\left(r_i^{MM}/\hat{\sigma}_1\right)}{ave_j\psi_2^{'}\left(r_j^{MM}/\hat{\sigma}_1\right)}\right)^2 \tag{2.13}$$

where $r_i^{LS} = y_i - \hat{\alpha}_{LS} - \boldsymbol{x_i'\widehat{\beta}_{LS}}$ and $r_i^{MM} = y_i - \hat{\alpha}_{MM} - \boldsymbol{x_i'\widehat{\beta}_{MM}}$ denote the least-squares and robust residuals.

2.3 Asymptotic Power of Test T1 under K1

In general, it may not be possible to analytically compute the asymptotic power of the tests T1 and T2 under the broad alternative K2. However, straightforward arguments given below show that the large-sample power of a level α test T1 under non-normal distributions F_ϵ in K1 is given by

$$\Pr\left(|Z| \geq \kappa \cdot q_{1-\frac{\alpha}{2}}\right) \tag{2.14}$$

where Z is a standard normal random variable, q_α is the α-quantile of a standard normal distribution, $\kappa = \sqrt{\frac{(1-EFF)\tau\sigma^2}{\delta^2_{LS,MM}}}$ with $\sigma \equiv plim\ \hat{\sigma}_1$, and τ and $\delta^2_{LS,MM}$ defined in (1.21) and (2.8) under non-normal F_ϵ.

To establish (2.14) one just needs to note that for F_ϵ from K1 we have $\sqrt{n}\ (\hat{\beta}_{MM,i} - \hat{\beta}_{LS,i}) \rightarrow N\left(0, \delta^2_{LS,MM}C^{-1}_{x,ii}\right)$. Then since $\hat{\tau}$, defined in (1.29), along with $\hat{\sigma}_1$ and $\hat{C}^{-1}_{x,ii}$ are consistent estimators of τ, σ and $C^{-1}_{x,ii}$ respectively, Slutsky's theorem gives:

$$T_{1i} \equiv \frac{\sqrt{n}(\hat{\beta}_{MM,i} - \hat{\beta}_{LS,i})}{\sqrt{(1-EFF)\hat{\tau}\hat{\sigma}_1{}^2\hat{C}^{-1}_{x,ii}}} \rightarrow N\left(0, \frac{\delta^2_{LS,MM}}{(1-EFF)\tau\sigma^2}\right).$$

Noting that we reject when $|T_{1i}| \geq q_{1-\frac{\alpha}{2}}$ gives (2.14).

Under the normality hypothesis H1, (2.10) with $EFF = \tau^{-1}$ gives $\kappa = 1$ and then (2.14) gives the level of the test. Note that the asymptotic power will be greater than the level but less than one whenever $0 < \kappa < 1$.

Figure 4 shows the asymptotic power of test T1 under a standard t-distribution versus degrees of freedom for normal distribution efficiencies of 85%, 90%, 95% and 99%. Figure 5 shows power results versus the parameter μ in the asymmetric normal mixture distribution $(1-\gamma)N(0,1) + \gamma N(\mu, 0.25^2)$ with $\gamma = 0.02$ for the same four normal distribution efficiencies as in Figure 4. The values of σ and α_{MM} were calculated by numerically solving equations (1.16) and (1.17). The values of τ and $\delta^2_{LS,MM}$ were calculated via numerical integration according to equations (1.21) and (2.8).

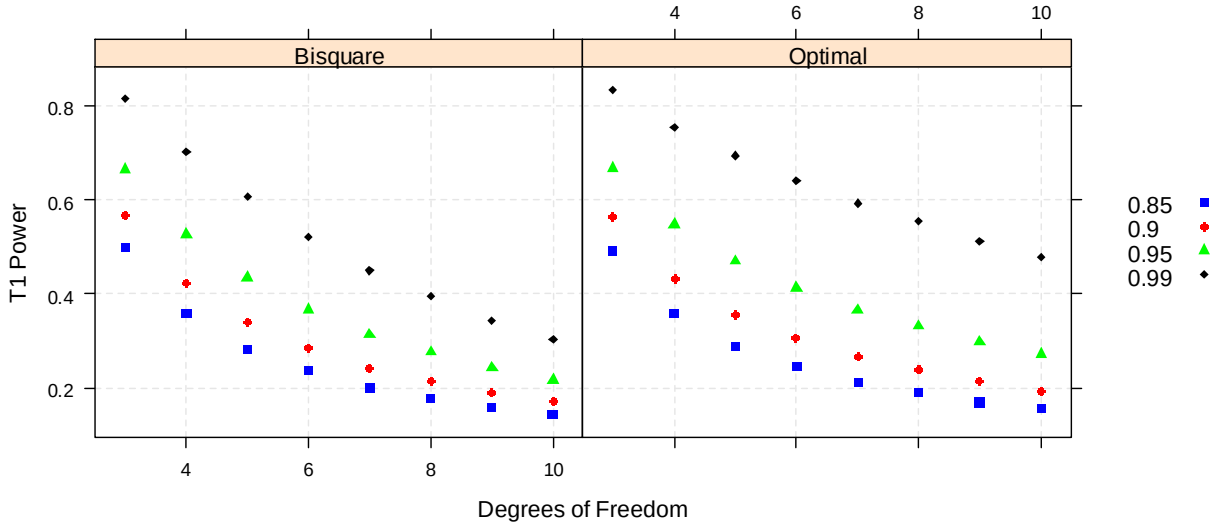

Figure 4. Asymptotic power of T1 vs degrees of freedom as calculated from (2.14) for t distributed errors.

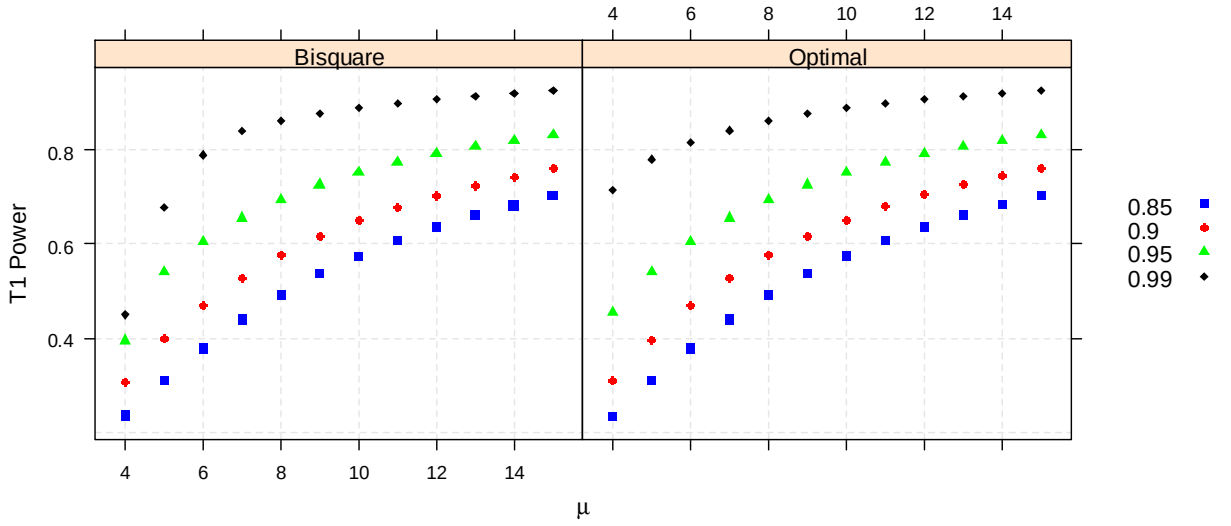

Figure 5. Asymptotic power of T1 vs μ as calculated from (2.14) for errors following a mixture distribution $(1-\gamma)N(0,1)+\gamma N(\mu,0.25^2)$ with $\gamma = 0.02$.

Both plots reveal that the greater the degree of non-normality, i.e., the smaller the degrees of freedom or the larger the μ, the greater the power.

In general, for distributions that are not too far from normality the power of T1 will be less than one, but the power will approach one as non-normality becomes arbitrarily "large". To see this, one just needs to show that $\delta^2_{LS,MM}$ can become arbitrarily large under arbitrarily "large" non-normality while at the same time $(1-EFF)\tau\sigma^2$ remains bounded, resulting in $\kappa \to 0$. To see that δ^2 can become arbitrarily large note that

$$\begin{aligned}\delta^2_{LS,MM} &= E\left\{\left(u_{LS}-\frac{\sigma\psi_2\left(\frac{u_{MM}}{\sigma}\right)}{E\psi_2'\left(\frac{u_{MM}}{\sigma}\right)}\right)^2\right\} \\ &= E(u^2_{LS})+\frac{\sigma}{E\psi_2'\left(\frac{u_{MM}}{\sigma}\right)}E\left(\psi_2\left(\frac{u_{MM}}{\sigma}\right)\left(\frac{\sigma\psi_2\left(\frac{u_{MM}}{\sigma}\right)}{E\psi_2'\left(\frac{u_{MM}}{\sigma}\right)}-2u_{LS}\right)\right)\end{aligned} \tag{2.15}$$

Since $\psi_c(r)$ are zero for r outside the interval $(-c, c)$, the asymptotic value σ of the robust scale estimator remains bounded, as does the second term in the above expression[7]. However, the term $E(u_{LS}^2) = E((\epsilon - E\epsilon)^2)$ is equal to the variance of ϵ and can be arbitrarily large for arbitrarily large non-normality. For example:

- If F_ϵ is a standard t-distribution with ν degrees of freedom, then $var(\epsilon) = \frac{\nu}{\nu-2} \to \infty$ as $\nu \searrow 2$.
- If F_ϵ is an asymmetric normal mixture distribution $(1-\gamma)N(0,1) + \gamma N(\mu, 0.25^2)$, then $var(\epsilon) = \gamma \cdot 0.25^2 + (1-\gamma)(1+\gamma\mu^2) \to \infty$ as $\mu \to \infty$.

Finally, since $\psi_c(r)$ is zero for r outside the $(-c, c)$ interval it follows that τ defined by (1.21) is bounded, and consequently, the numerator in κ is bounded while the denominator can become arbitrarily large as the non-normality becomes arbitrarily large.

There is relatively little difference between the two loss functions under substantial non-normality, i.e. for small degrees of freedom or large μ. However, for 99% normal distribution efficiency the power for the optimal loss function is noticeably larger than that of the bisquare loss function as the error distribution gets closer to normality, i.e., as the degrees of freedom increase or μ gets smaller.

In general, the greater the normal distribution efficiency of the MM-estimator the greater the power. This is because under non-normal errors when EFF increases $\delta_{LS,MM}^2$ tends to decrease slower than does $(1 - EFF)\tau$, leading to smaller values of κ and consequently higher T1 power. This is illustrated in Figure 6, which plots κ^2 versus EFF for t distributed errors. Note that triangles correspond to normal errors, and, as expected, $\delta_{LS,MM}^2$ is equal to $(1 - EFF)\tau\sigma^2$ regardless of normal distribution efficiency.

---

[7] Noting that $u_{LS} = \epsilon - E\epsilon$ and $u_{MM} = \epsilon - \alpha_{MM}$, we write

$$\psi_2(u_{MM}/\sigma)\left(\frac{\sigma\psi_2(u_{MM}/\sigma)}{E\psi_2'(u_{MM}/\sigma)} - 2u_{LS}\right) = \psi_2((\epsilon - \alpha_{MM})/\sigma)\left(\frac{\sigma\psi_2((\epsilon - \alpha_{MM})/\sigma)}{E\psi_2'((\epsilon - \alpha_{MM})/\sigma)} - 2\epsilon - 2E\epsilon\right)$$

As long as $|E(\epsilon)| < \infty$, the function is finite when $\epsilon \in (\alpha_{MM} - \sigma \cdot c, \alpha_{MM} + \sigma \cdot c)$ and is zero outside that interval. Thus, as long as $|\alpha_{MM}| < \infty$, the expectation in the second term in (2.15) is an integral of a finite function over a finite range and, thus, finite.

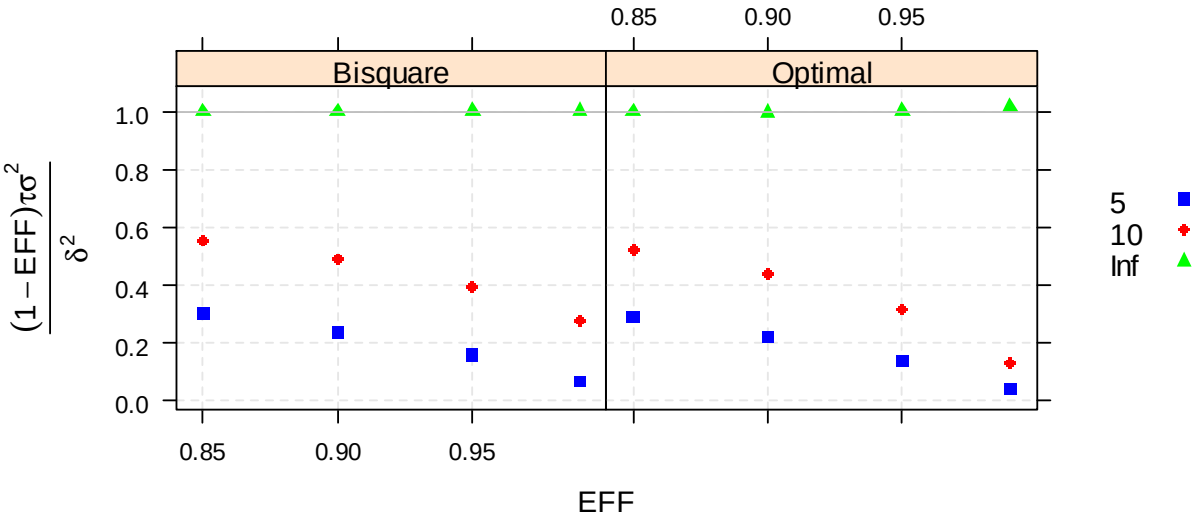

Figure 6. κ^2 vs EFF for t distributed errors with degrees of freedom equal to 5, 10 and ∞.

2.4 Monte Carlo Simulations

In order to understand the comparative finite sample behavior of the level and power of our tests T1 and T2 as a function of the choice of bisquare versus optimal loss functions at various normal distribution efficiencies, we carried out a number of Monte Carlo simulation studies with large-sample significance level α set at 0.05. As approximations to the finite sample level and power of the tests we calculate Monte Carlo rejection rates, i.e. the proportion of times out of N replicates that a given hypothesis was rejected. Because of the relative complexity of the comparative analysis we focus on the slope coefficient β in a simple linear regression model $y_i = \beta_0 x_i + \epsilon_i = \beta_0 x_i + \alpha_0 + u_i$, $i = 1, \cdots, n$. Simulations were conducted in R using lmRob function from the R robust library. The test statistic T1 given by (2.5) and (2.6) is easily computed from the output of lmRob and the standard R least-squares fitting function lm. For T2 given by (2.12) and (2.13) we wrote additional R code that is available upon request.

2.4.1 Distribution Models

We assume independent and identically distributed (i.i.d.) random x_i that are independent of i.i.d. errors u_i for the first four models below. We generate samples from the following distributions for the errors u_i:

Model 1: Standard normal, which is included in H1 and H2

Model 2: Standard-t with 5 degrees of freedom, which is included in H2 and K1

Model 3: Skewed-t with skewness parameter $\lambda = 1$ and 5 degrees of freedom, as implemented in the *R* package sn (Azzalini, 2011), which is included in H2 and K1

Model 4: Asymmetric two-term conditional normal mixture that is included in H2 and K1:

$(1-\gamma)N(0,1^2) + \gamma N(\mu, 0.25^2)$ where we condition by setting the number of observations from the second component to be $\lfloor \gamma n \rfloor$, with γ ranging from 0.01 to 0.1 and $\mu = 4, 7,$ and 14. In this case large positive residual outliers occur independently of x_i.

Model 5: Asymmetric two-term conditional joint normal mixture for x_i and u_i that is included in K2:

$\binom{x_1}{u_1}, \dots, \binom{x_n}{u_n}$ are i.i.d. $(1-\gamma)N\left(\binom{0}{0}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + \gamma N\left(\binom{2}{\mu}, \begin{bmatrix} 0.25^2 & 0 \\ 0 & 0.25^2 \end{bmatrix}\right)$, where we condition on the number of 'outliers' from the second component to be $\lfloor \gamma n \rfloor$, with γ ranging from 0.01 to 0.1 and $\mu = 4, 7,$ and 14. In this case the mixture model is such that large positive residuals occur for large values of x_i, and result in biased LS and robust estimates of β, with the bias of the latter being much smaller.

We carry out the conditioning in Models 4 and 5 as follows. We first generate $x_1 \dots x_n$ as i.i.d. $N(0, 1^2)$ and $u_1 \dots u_n$ as i.i.d. $N(0, 1^2)$. Then in Model 4 we randomly select $\lfloor \gamma n \rfloor$ observations and replace corresponding u_i with i.i.d. $N(\mu \gg 0, 0.25^2)$. In Model 5 we also replace the corresponding x_i with i.i.d. $N(2,\ 0.25^2)$. As a result the reported null hypotheses rejection rates are not confounded by the randomness of the outlier fraction in each sample. The corresponding unconditional rejection rates for these two models can be easily obtained from conditional rejection rates as $\sum_{i=0}^{\infty} RR_i \cdot p_i$, where RR_i is a conditional rejection rate when the number of outliers is equal to i, and p_i is the probability that the number of outliers is equal to i. For example, for $\gamma = 0.02$ about 13.3% $(p_0 = 0.133)$ of the samples of size 100 will have no outliers, 27.1% $(p_1 = 0.271)$ will have exactly one outlier, 27.3% $(p_2 = 0.273)$ will have exactly two outliers, and 32.3% of the samples will have three or more outliers.

We note that for $\mu = 14$ nearly all outliers get rejected by the robust estimators for both bisquare and optimal loss function. On the other hand, when $\mu = 4$ the observations from the mixture component are

closer to the rejection points of the optimal psi function at high normal distribution efficiencies of 95% and higher, and, therefore, not so many observations are rejected as 'outliers'.

We set $\alpha_0 = 0$, $\beta_0 = 1$. For models 1, 2, and 3 we generated $10{,}000$ replicates. Models 4 and 5 include many combinations of the parameters μ and γ, and for each such combination we generated $1{,}000$ replicates.[8] We used sample size n ranging from 35 to 500.

2.4.2 Results

Model 1 (normal distribution errors). The four panels in Figure 7 display H1 normal distribution Monte Carlo level versus sample size of the two tests T1 and T2 for both the bisquare and optimal loss functions, and for the four normal distribution efficiencies (85%, 90%, 95% and 99%).

T1 Test Performance: The actual level of T1 is larger than the nominal significance level of .05 for both the bisquare and optimal loss functions, with the exception of the bisquare at sample size 500. Furthermore, there is a substantial spread in the T1 rejection rates across the four efficiencies for both loss functions. As the sample size increases the spread diminishes relatively quickly for the bisquare choice with a consistent ordering of the values with respect to normal distribution efficiencies. The spread decreases more slowly for the optimal psi and has a very erratic pattern with respect to normal distribution efficiencies.

T2 Test Performance: The bottom left panel of Figure 7 reveals that the level of test T2 is right on target for the bisquare loss function for all sample sizes regardless of normal distribution efficiency. By way of contrast, the level performance of the optimal loss function in the bottom right panel of Figure 7 is very poor, particularly at higher normal distribution efficiencies, and is only consistently on target for samples sizes at least 250 with normal distribution efficiencies of 85% and 90%.

In Chapter 3 we explain the reason for some of the peculiar behavior of T1 and T2 in Figure 7.

---

[8] Standard errors for the Monte Carlo level and power estimates can be obtained from the estimation theory of a binomial proportion p. In particular, using classical standard errors, namely $\sqrt{\hat{p}(1-\hat{p})/N}$, we see that the standard errors are reasonably small even at $N = 1,000$ replicates. The standard errors of the Monte Carlo level, i.e. when $\hat{p} \approx 0.05$, are approx. 0.0069 for $N = 1{,}000$ and 0.0022 for $N = 10{,}000$. The standard error of the Monte Carlo power is the largest when $\hat{p} = 0.5$, and in this case is equal to 0.0158 for $n = 1,000$ and equal to 0.005 for $n = 10{,}000$.

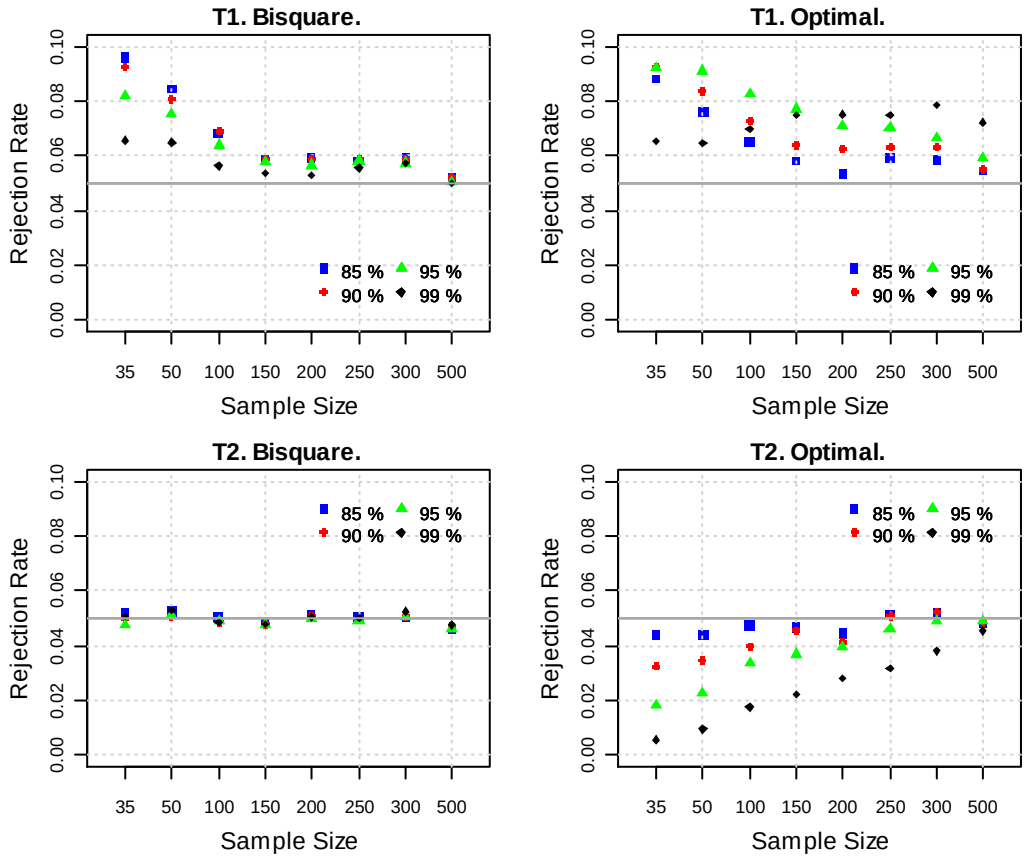

Figure 7. Model 1. Level of T1 (top) and T2 (bottom) test statistics for slope β in a simple linear regression under normal residuals. Grey horizontal line is at large-sample significance level of 0.05.

Model 2 (symmetric t-distribution errors). Results for a symmetric t-distribution with five degrees of freedom are displayed in Figure 8.

T1 Test Performance: The symmetric t-distribution with five degrees of freedom is in the alternative hypothesis K1 for T1 and thus one would hope for high power results. For both choices of loss function the power of T1 indeed increases with increasing sample size and with normal distribution efficiency. Furthermore, for the both loss functions the power of T1 for sample size 500 is close to the estimated asymptotic value for each of the four efficiencies in Figure 4. Recall from the discussion of Section 2.3 that because both the LS and robust estimates are consistent for beta under K1, the asymptotic power of

T1 will be less than 1. It is therefore not surprising to see that the power of T1 is less than one for the largest sample sizes in Figure 8.

T2 Test Performance: Since the symmetric t-distribution is in the null hypothesis H2, it is comforting that the T2 levels for the bisquare loss function are reasonably close to the large-sample significance level of .05 for all normal distribution efficiencies. The T2 levels for the optimal loss function are also reasonably on target except for sample sizes less than 100 and efficiencies of 95% and 99%.

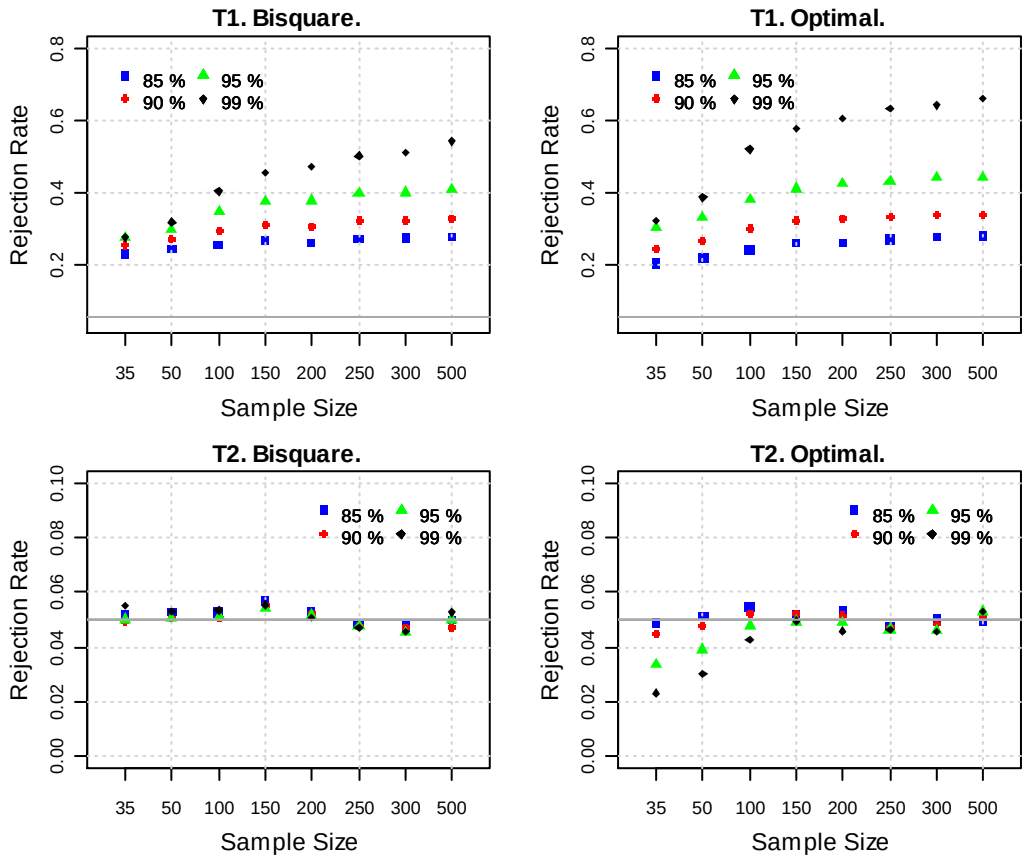

Figure 8. Model 2. Rejection rates of the T1 (top) and T2 (bottom) test statistics for the slope β in a simple linear regression under symmetric t_5 residuals. Grey horizontal line is at large-sample significance level of 0.05.

Model 3 (skewed t-distribution errors). Rejection rates for skewed t-distribution with five degrees of freedom are presented in Figure 9. The results in Figure 9 are very similar to those in Figure 8 and most

of the above comments for the symmetric t-distribution apply here. Thus, moderate skewness appears to have little impact on T1 power and T2 level, while tail fatness has a large impact. As tail fatness increases, the T1 power increases (c.f. Figure 4 in Section 2.3) and the T2 level behavior for the optimal psi improves (e.g., compare Figure 7 and Figure 8).

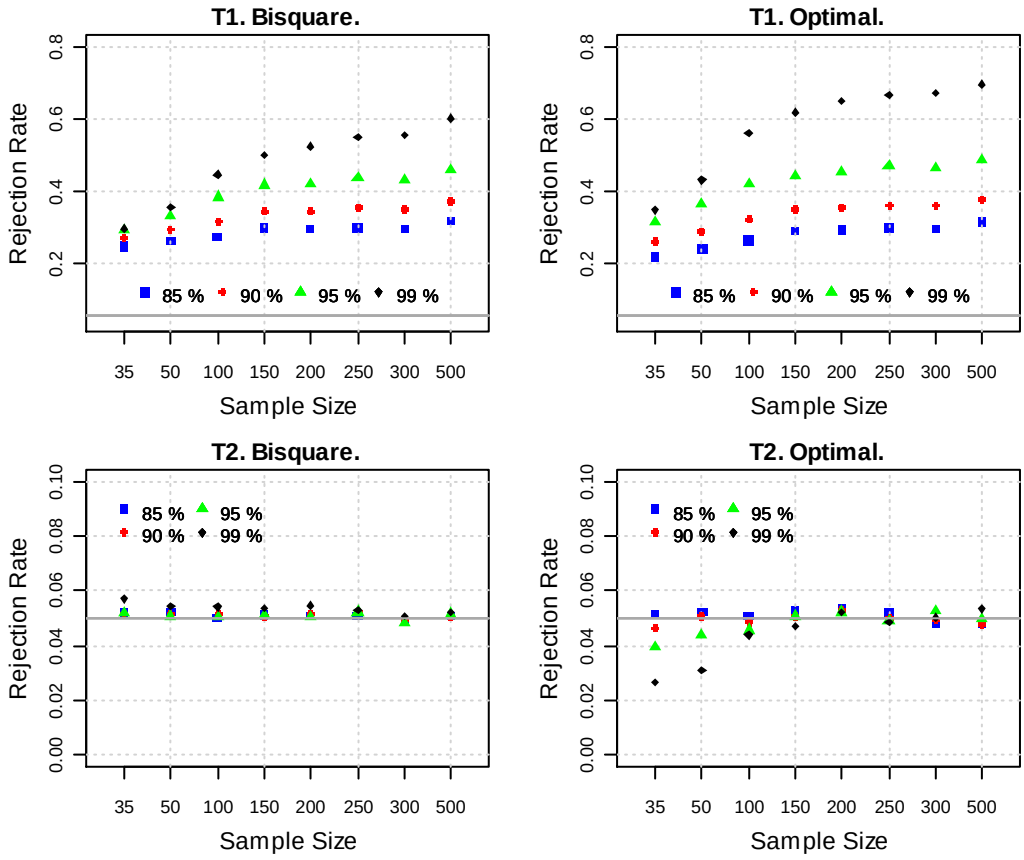

Figure 9. Model 3. Rejection rates of the T1 (top) and T2 (bottom) test statistics for the slope β in a simple linear regression under skewed t_5 residuals. Grey horizontal line is at large-sample significance level of 0.05.

Model 4 (asymmetric normal mixture distribution errors). Figure 10 displays Monte Carlo rejection rates versus sample size and normal distribution efficiency for mixing proportions 0.02, 0.04 and 0.06. The square, diamond and triangle symbols represent μ values of 4, 7 and 14 respectively.

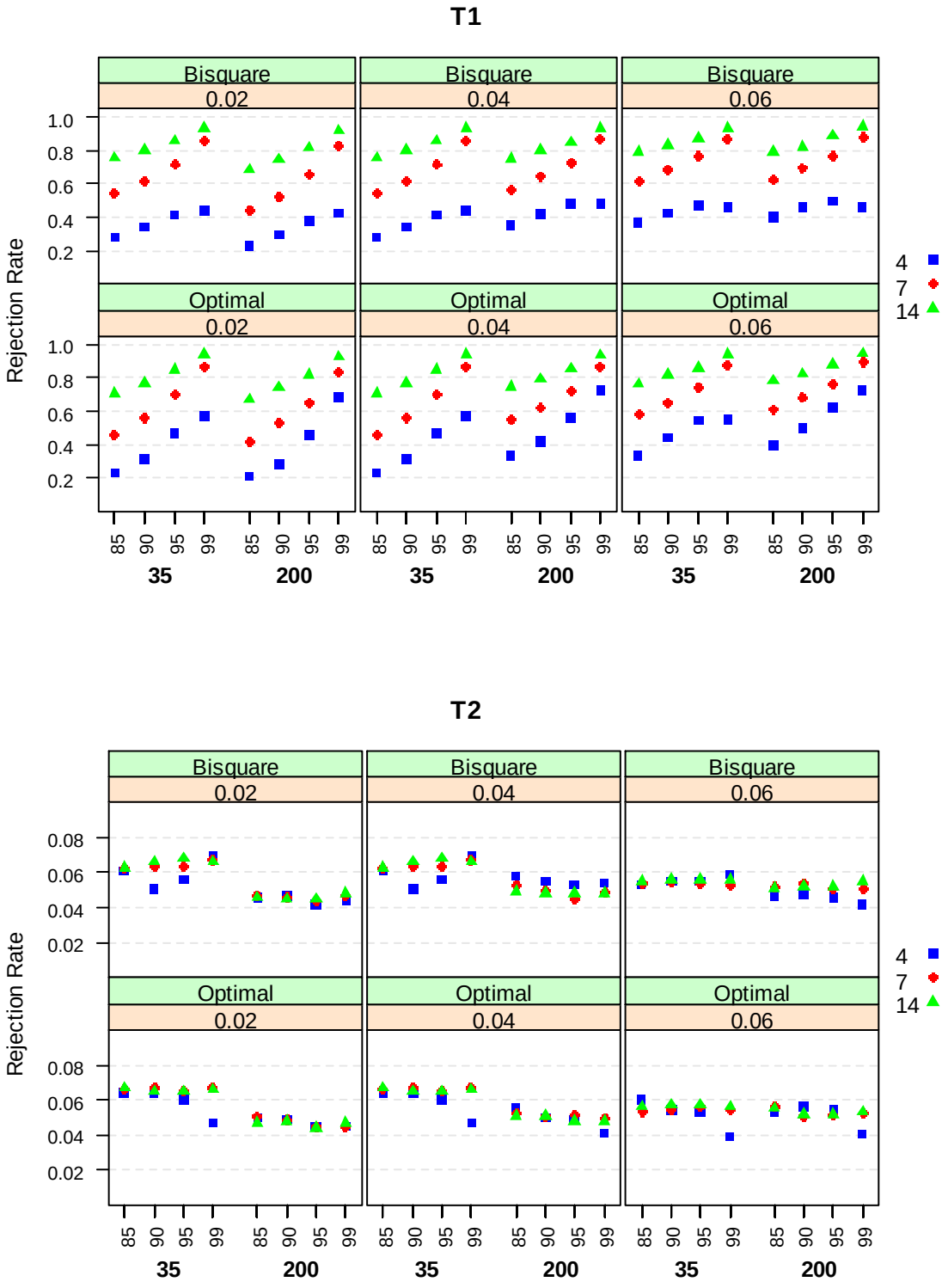

Figure 10. Model 4. Rejection rates of T1 and T2 test statistics for slope β in a simple linear regression under asymmetric residual contamination. Blue squares correspond to mild outliers $N(\mu = 4, 0.25^2)$, red diamonds – to medium outliers $N(\mu = 7, 0.25^2)$ and green triangles – to gross outliers $N(\mu = 14, 0.25^2)$.

T1 Test Performance: The normal mixture distribution is in the alternative K1 for the test T1 and so we expect its power to increase with increasing values of μ, and this is indeed the case. Interestingly, there is virtually no sample size effect, with the power at the small sample size 35 being almost the same as at 200. We believe that this is partially due to adopting a conditional mixture, i.e. fixing the number of outliers in a sample. As in the case of T1 for the previous two models, the power typically increases with increasing normal distribution efficiency. Except for the case of higher normal distribution efficiencies and $\mu = 4$, Figure 10 does not provide a strong argument in favor of choosing one of the two loss functions over the other. However for efficiencies of 95% and higher the power performance for $\mu = 4$ would lead one to choose the optimal loss function.

T2 Test Performance: In the case of T2 the normal mixture distribution used here is in the null hypothesis H2 and so one is interested in how close the level of T2 is to the large-sample level of .05. The empirical results in the bottom panels of Figure 10 are reasonably encouraging, being somewhat too large for sample size 35, more on target for sample size 200, and a bit erratic for 99% efficiency and $\mu = 4$.

Model 5 (bivariate normal mixture distribution). The results for Model 5 displayed in Figure 11 are strikingly different than those for Model 4 in Figure 10. Since Model 5 is in the alternative K2 one might hope for reasonable power for both tests T1 and T2, the more so the larger the sample size and the larger μ.

T1 Test Performance: The results for T1 are rather striking for two reasons. First of all, the power is essentially 100% for all samples sizes and all normal distribution efficiencies for $\mu = 7 \text{ and } 14$. Secondly, the behavior of the power for $\mu = 4$ is quite complex. For example, at each normal distribution efficiency the power mostly increases with sample size as anticipated for each value of γ. But the behavior of power with respect to increasing efficiency at each sample size is complex and depends on γ. For example, when $\gamma = .02 \text{ or } .04$ and the sample size is either 35 or 50 the power at first either increases or is constant with increasing efficiency and then decreases, while for $\gamma = .06$ the power decreases. This behavior is partially explained in Chapter 3. We also want to note that for $\mu = 4$ the bisquare tends to

have higher power than the optimal loss function when the normal distribution efficiency is 99%, and also a bit for 95%. This is rather opposite to what we observed under Model 4 where for $\mu = 4$ the optimal loss function has higher power than the bisquare for higher normal distribution efficiencies.

T2 Test Performance: The power of T2 increases with sample size for all values of μ and γ, reaching 100% at $\gamma = .04 \text{ and } .06$ except for the case of $\mu = 4$ and 99% normal distribution efficiency. What is notable is that the T2 power is considerably lower than that of T1 for sample sizes 35 and 50 at $\gamma = .02$ and at sample size 35 for $\gamma = .04$. This is a price paid for allowing non-normal distributions in the null hypothesis H2.

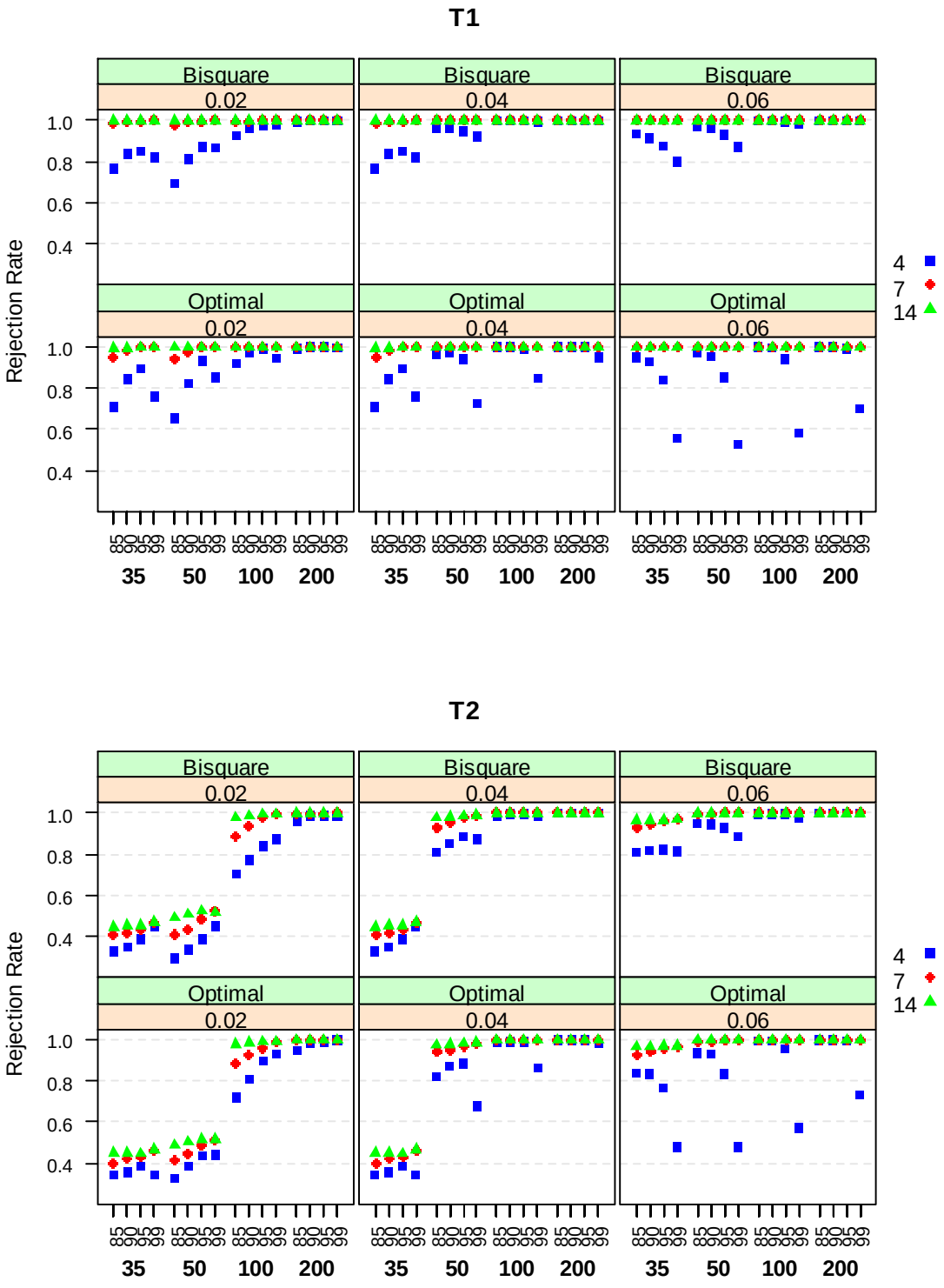

Figure 11. Model 5. Power of the T1 and T2 test statistics for slope β in a simple linear regression under bivariate asymmetric contamination. Blue squares correspond to mild outliers $N(\mu = 4, 0.25^2)$, red diamonds – to medium outliers $N(\mu = 7, 0.25^2)$ and green triangles – to gross outliers $N(\mu = 14, 0.25^2)$.

Summary of Monte Carlo Results.

Level: The levels of the T1 test in small to moderate sample sizes are consistently above the nominal significance level for both bisquare and optimal psi-functions. Compared to the bisquare, the optimal psi-function requires much larger sample sizes for the T1 finite sample level reach its asymptotic value. See Figure 7.

The finite sample levels of the T2 test with the bisquare psi-function are remarkably close to the nominal significance level for all sample sizes considered in the study. This is not the case for the T2 test with the optimal psi, especially at high normal distribution efficiencies such as 90% and above, in which case the levels of the T2 test in small to moderate sample sizes are substantially below the nominal significance level. See Figure 7 through Figure 10.

Chapter 3 explains these behaviors.

Power: In general, the asymptotic and finite sample power of the T1 test under alternative K1 increases with normal distribution efficiency of the robust estimate, but stays below one because both LS and MM estimators are consistent for $\boldsymbol{\beta}$. The choice of the psi-function makes little difference when departures from normality are large, e.g. at degrees of freedom of 4 or less in Models 2 and 3, or μ larger than 6 in Model 4. At small departures from normality under alternative K1, the T1 power was the largest for the optimal psi-function at 99% efficiency, and increase in power was substantial compared to lower efficiencies as well as to the bisquare psi at 99% efficiency, both asymptotically and in finite samples.

Typically, the finite sample power of the T1 and T2 tests under alternative K2 increases with normal distribution efficiency of the robust estimate, but asymptotically equals to one for all efficiencies. The exception is an alternative K2 with large proportion of a 'borderline' contamination, in which case the T1 and T2 rejection rate, even though asymptotically equal to one, in finite samples may drop at higher normal distribution efficiencies. For example, see Model 5 with $\mu = 4$ in Figure 11.

The standard errors in the T1 test tend to be smaller than that in the T2. Consequently, under K2 the T1 test typically has higher power than T2 unless sample size is large enough so that the power for both tests is essentially one. This is a price paid for allowing non-normal distributions in the null hypothesis H2.

2.5 Empirical Examples

In this section we present several empirical examples. The first example in subsection 2.5.1 illustrates application of the proposed test statistics in finance for determining significant differences between the classical and robust estimates of stock beta from a single factor market model. The two multivariate regression examples in subsections 2.5.2 through 2.5.3 use datasets discussed in Chapter 3 of Rousseeuw and Leroy (1987), namely the artificial Hawkins-Bradu-Kass data and the aircraft data.

2.5.1 Finance Single Factor Model Beta

The single factor model beta of a set of asset returns is the slope coefficient in a regression of the asset returns on market returns. Beta plays a central role in the capital asset pricing model (CAPM) (Sharpe, 1964) and is one of the most widely known and widely used measures of the market risk of an asset and the asset's excess returns over the risk-free rate. Figure 12 shows a scatter plot of the AAR Corp (ticker AIR) stock weekly returns versus the weekly returns of the CRSP (www.crsp.com) value-weighted market index during a 2 year period from June 2008 to June 2010. All returns are in excess of risk-free rate. The red dashed line is for the least squares fit and black solid line is for the robust MM regression fit using a bisquare loss function at 95% normal distribution efficiency. There are positive AIR return outliers that occurred at the same time as two large positive market returns. These two data points drive LS slope up, but do not affect the robust fit because they are rejected by the robust estimator.

The standard errors (SE) and p-values for the T2 and T1 tests for the difference in the two slopes are reported in Table 2. The T1 SE by definition is just a fraction of the MM beta SE, namely $\sqrt{1-0.95}\cdot 0.14 \approx 0.22 \cdot 0.14 = 0.031$ (see eq. (2.5)). The observed difference in the two slopes is very unlikely to occur due to inefficiency of the robust estimates under normality (T1 p-value is less than 0.0001). The T2 SE is equal to 0.16, which is larger than the MM beta SE of 0.14, but is smaller than the LS beta SE of 0.19. Nonetheless, the T2 p-value of 0.004 is significant at the 1% level, suggesting that the OLS beta of 1.56 is significantly biased upward. The robust beta of 1.1 better describes stock and market return relationship for majority of the time points, with a very small fraction of the AIR returns being outliers

relative to the overall pattern. It should be noted that the difference in two betas of 0.46 would be of practical significance to most investors.

Differences between LS and robust betas are very common, as is revealed in Bailer, Maravina, and Martin (2012). We highly recommend routine use of robust regression betas along with their standard errors and the test statistics T2 p-values as a complement to the LS estimates of asset returns provided by many financial data service providers (e.g., ValueLine, Barra, Bloomberg, Capital IQ, Datastream, Ibbotson, Google Finance, and others). Acceptance of the null hypotheses H2 of no significant differences between the robust and LS betas would give investors an extra comfort in making their decisions based on the classical LS beta estimates. On the other hand, large significant differences should alert analysts to investigate the returns data closely to determine which beta are the most useful guide to investment decisions.

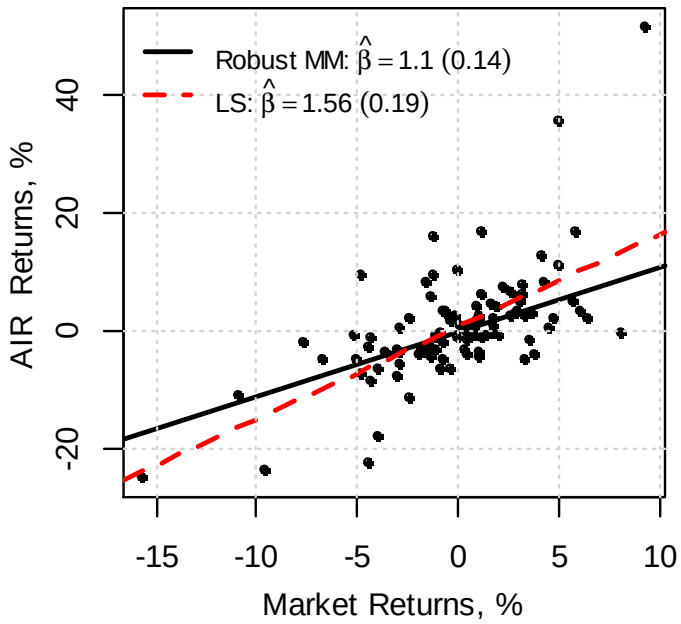

Figure 12. Scatter plot of the AIR and market weekly returns in excess of the risk free rate with the fitted least-squares and robust lines.

Table 2. T2 and T1 test statistics for the difference between LS and robust beta estimates for the AIR example.

	T2		T1	
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$	SE	p-value	SE	p-value
0.46	0.16	0.004	0.031	0.000

2.5.2 Hawkins-Bradu-Kass Data

This artificial data set generated by Hawkins, Bradu, and Kass (1984) was discussed in Example 1 on p.93 of Rousseeuw and Leroy (1987). The data set consists of 75 observations with one response and three explanatory variables. Unfortunately, the true model is not explicitly stated by the authors, but close examination of the data suggests the following data generation process:

$$y_i = 0 + \lambda_i + 0 \cdot X1_i + 0 \cdot X2_i + 0 \cdot X3_i + e_i \tag{2.16}$$

where e_i are $i.i.d.N(0, \sigma^2)$ with $\sigma < 1$ and

a) For $i = 1 \ldots 10$: $\lambda_i = 10$ and $X1, X2, X3$ have values close to 10, 20 and 30 respectively

b) For $i = 11, \ldots, 14$: $\lambda_i = 0$ and $X1, X2, X3$ have values close to 10, 20 and 30 respectively

c) For $i = 15, \ldots, 75$: $\lambda_i = 0$ and $X1, X2, X3$ are relatively uniformly distributed between 0 and 3.5

Therefore, the first ten observations contain ten highly influential leverage outliers that are approximately co-located, as well as four non-influential leverage points for which *X1*, *X2* and *X3* are outlying but the corresponding response values y_i fit the model quite well.

Table 3 presents a summary of the fitted LS and robust regressions and the T1 and T2 tests for the differences between the LS and robust individual coefficients. In particular, we used bisquare loss function at 95% efficiency, but results for the optimal loss function are virtually the same. None of the MM regression coefficients are significantly different from zero, which agrees quite well with what we believe to be the true model. This is in contrast to LS for which the coefficient for X2 is significantly negative, the coefficient for X3 is significantly positive, while the coefficient for X1 is non-significant. Thus, it is not surprising that our T2 test indicates significant differences between the LS and MM coefficients for X2 and X3, but not for X1. Note once again the large difference between the T2 and T1 standard errors, which

follows from the fact that latter are by definition a small fraction of the MM standard errors. With such small standard errors T1 inevitably yields p-values that are too small, quite unrealistically so in this case.

Table 3 Regression results for the Hawkins-Bradu-Kass example.

	LS			MM				T2		T1	
	Coef	SE	P-value	Coef	SE	P-value	$\widehat{\boldsymbol{\beta}}_{LS} - \widehat{\boldsymbol{\beta}}_{MM}$	SE	P-value	SE	P-value
	-0.388	0.417	0.355	-0.190	0.115	0.103					
X1	0.239	0.263	0.365	0.085	0.073	0.247	0.154	0.238	0.518	0.016	0.000
X2	-0.335	0.155	0.034	0.041	0.044	0.353	-0.376	0.143	0.009	0.010	0.000
X3	0.383	0.129	0.004	-0.054	0.039	0.168	0.437	0.126	0.001	0.009	0.000

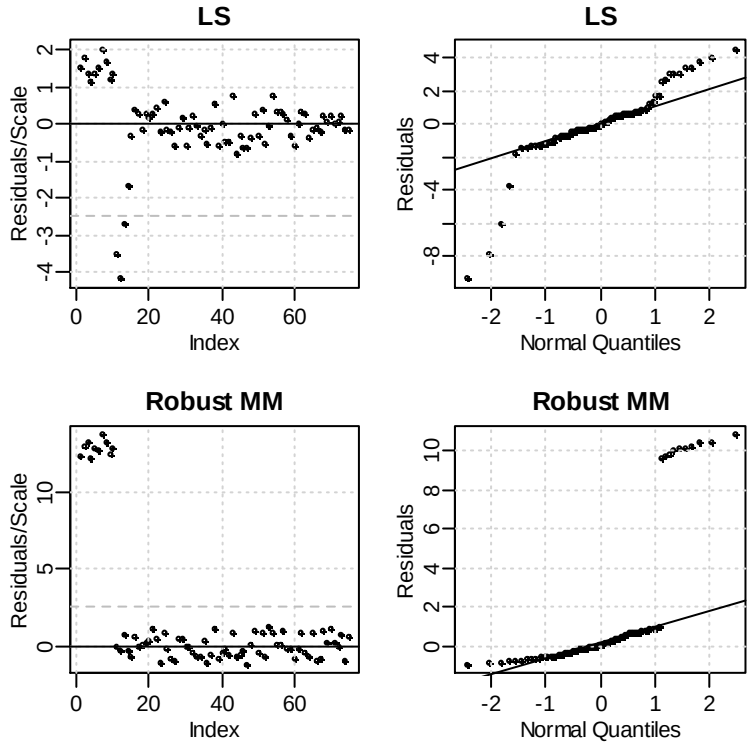

Figure 13. Residual plots for the Hawkins-Bradu-Kass example. Residual scale estimates are 2.3 for OLS and 0.8 for the robust regression.

The residuals plots in lower panels of Figure 13 show that the robust MM regression fit correctly identifies the first 10 observations as gross outliers. On the other hand the LS fit identifies the good leverage

observations 11, 12 and 13 as outliers based on their scaled residuals falling outside the ± 2.5 band, and these are actually good observations according to the model (2.16).

Removing bad leverage outliers (observations 1 through 10) leads to both LS and MM estimates to be non-significant. The differences are also non-significant for all three coefficients according to both T2 and T1.

2.5.3 Aircraft Data

This dataset, presented in Table 22 on p.154 of Rousseeuw and Leroy (1987) was originally provided by Gray (1985). It contains information on 23 single-engine aircrafts built over the years 1947-1979. The dependent variable is cost (in units of $100,000), and the explanatory variables are aspect ratio (X1), lift-to-drag ratio (X2), weight of the plane (in pounds) (X3), and maximal thrust (X4). We divided X3 and X4 by 1,000 so that they are of similar magnitude as X1 and X2.

The results for the optimal and bisquare MM estimates at 95% efficiency lead to similar conclusions so we present only the latter. Coefficient estimates and test statistics are listed in Table 4 and the residuals plots are displayed in Figure 14. Both the robust and LS estimates result in all coefficients being significant at the 5% level, and in some cases at the 1% level. However, LS fails to identify outliers while the robust regression identifies observation 22 and possibly observation 16 as outliers. These outliers significantly bias the LS coefficients for X3 and X4, as indicated by the fact that the T2 test finds the differences between the corresponding LS and robust estimates to be highly significant (p-values ≤ 0.01). Removing the two outlying observations and repeating the analysis leads to non-significant T2 and T1 p-values for all coefficients (all values are higher than 0.2 except one p-value of .08). Once again T1 has standard errors that are much too small, resulting in all differences in LS and robust coefficient estimates being significant instead of just those for X3 and X4.

Table 4. Regression results for the aircraft example.

	LS			MM				T2		T1	
	Coef	SE	P-value	Coef	SE	P-value	$\widehat{\boldsymbol{\beta}}_{LS} - \widehat{\boldsymbol{\beta}}_{MM}$	SE	P-value	SE	P-value
	-3.79	10.12	0.71	6.14	6.80	0.38					
X1	-3.85	1.76	0.04	-3.23	1.14	0.01	-0.62	1.16	0.59	0.25	0.01
X2	2.49	1.19	0.05	1.67	0.78	0.05	0.82	0.79	0.30	0.17	0.00
X3	3.50	0.48	0.00	1.92	0.45	0.00	1.58	0.45	0.00	0.10	0.00
X4	-1.95	0.50	0.00	-0.93	0.39	0.03	-1.02	0.39	0.01	0.09	0.00

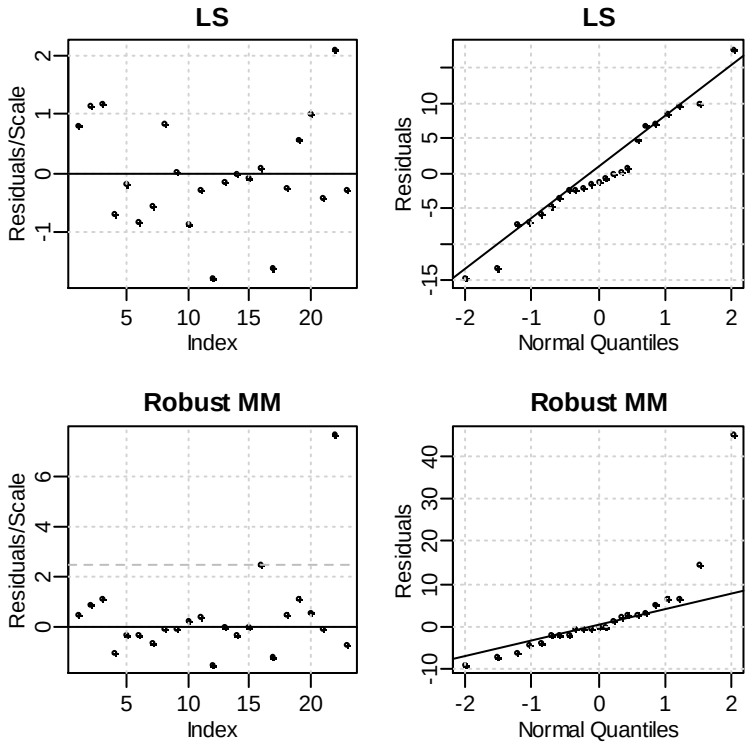

Figure 14. Residual plots for the aircraft example. Residual scale estimates are 8.4 for LS and 5.9 for the robust regression.

2.6 Summary and Discussion

This chapter proposes two new statistical tests, T1 and T2, for detecting differences between the LS and robust MM regression slope estimators in a linear regression, thereby filling a gap in diagnostic tools for comparing LS and robust regression. Test T1 is designed to detect significant differences between LS

and MM regression slope estimators due to inefficiency of the LS estimator under non-normal, fat-tailed error distributions or due to both LS inefficiency and large bias relative to the MM estimator. Test T2 is designed to detect cases when the LS estimator has significantly larger bias than the robust estimator. Thus, a rejection of T1 and no rejection of T2 would suggest inefficiency alone without differences in the biases of LS and MM estimators. A rejection in either of the two tests would support the use of the robust regression estimate instead of the LS estimate, and would signal the need for a careful examination of the data for influential outliers. The latter are not always evident with a LS fit, but are often revealed with a robust model fit.

T1 is based on an application of the famous result in econometrics by Hausman (1978). T2 is based on a null hypothesis asymptotic normal distribution of the differences between the LS and MM regression estimators that we have derived.

Through the Monte Carlo simulation study, we have shown that the choice of the loss function for the robust estimate greatly impacts the finite sample behavior of the T1 and T2 test statistics, a result which seemed surprising at first. In particular, in small to moderate sample sizes an optimal loss function with normal distribution efficiencies of 90% and higher results in incorrect levels of T2. This is in sharp contrast to the adequate finite sample levels of T2 with a bisquare loss function. The reason for this difference turns out to be that a bisquare psi-function has a slowly re-descending character, while an optimal psi function has a very rapidly decreasing character. A similar conclusion was found recently by Koller and Stahel (2011) on statistical inference criteria for choosing psi-functions for robust regression M-estimators. Even though the hypothesis testing problem in Koller and Stahel (2011) is different from the one we are focused on, the authors also conclude that “minimizing the maximum asymptotic bias, which entails a quickly re-descending psi function, can lead to poor inference properties, e.g., the levels of tests are not as claimed when p/n is not very small.”[9] This should not be surprising in view of the warning by Hampel (Hampel et al., 1986) long ago that one should not use a psi-function that has too high of a “local shift sensitivity”, e.g., it should not re-descend too rapidly. It seems that this was good advice in the

[9] Koller and Stahel (2011) focused on constructing robust confidence intervals and hypothesis tests for the regression coefficients when sample size is small, meaning a large ratio p/n of the number of parameters p relative to sample size n.

context of statistical inference for robust regression, though it is less relevant for controlling maximum bias. Overall, for testing for inefficiency and/or bias in a least-squares estimator, we recommend using an MM bisquare estimator with normal distribution efficiencies of 85% to 95%.

Steps for future research include, but are not limited to: a) derive a small-sample distribution or finite-sample correction for T1 to ensure proper test levels; b) obtain theoretical power results for T1 and T2 under alternatives K2, which might be possible using results of Salibian-Barrera and Omelka (2010); b) propose robust estimates of the T2 standard errors that are consistent under H2, but possibly lead to higher power under K2 (we note that estimators of the T2 standard errors considered in this chapter are not robust); c) study behavior of the tests for multi-parameter regression models; d) investigate viability of the tests in small sample sizes, i.e. with a small n/p ratio.

Appendix A2: T2 Asymptotic Result

In this section we derive the asymptotic normal distribution of the differences between the LS and MM regression estimators under distribution models (1.3), which include both symmetric and asymmetric errors. Our results are a straightforward extension of the results of Fasano, Maronna, Sued, and Yohai (2012) (henceforth referred to as FMSY) that cover not only the case of i.i.d. data $(y_i, \boldsymbol{x_i'}),\ i = 1, \cdots, n$ that is our focus, but also certain classes of stationary and serially correlated observations.

As a preliminary step we provide an overview of some key aspects of the FMSY results as a stepping stone to our results. We start by listing the substantial number of conditions that are used in the different sections of FMSY in order to establish Fisher consistency of the M and S estimators (FMSY Section 3), weak continuity of M and S regression functionals (FMSY Section 4), weak differentiability of the estimator functionals (FMSY Section 5), and asymptotic normality of MM-estimators (FMSY Section 6). Let $\hat{\sigma}$, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}} = \left(\hat{\alpha}_S, \widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}\right)$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}} = \left(\hat{\alpha}_{MM}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}\right)$ be S-scale and S- and MM-regression estimators as defined in Chapter 1 by eq.(1.5) and (1.9). Let σ, $\boldsymbol{\theta}_{\boldsymbol{S}} = (\alpha_S, \boldsymbol{\beta}_{\boldsymbol{MM}})$ and $\boldsymbol{\theta}_{\boldsymbol{MM}} = (\alpha_{MM}, \boldsymbol{\beta}_{\boldsymbol{MM}})$ be solutions to (1.10), (1.12) and (1.14), and let $\widehat{\boldsymbol{\eta}}' \equiv \left(\widehat{\boldsymbol{\theta}}'_{MM}, \widehat{\boldsymbol{\theta}}'_S, \hat{\sigma}\right)$ and $\boldsymbol{\eta}' \equiv (\boldsymbol{\theta}'_{MM}, \boldsymbol{\theta}'_S, \sigma)$ be the corresponding joint vectors.

(A1) [FMSY Condition 1] vector $\boldsymbol{x}$ is not concentrated on any hyperplane, i.e. for all $\boldsymbol{b} \neq \boldsymbol{0}$ and for all a we have $P_{G_x}(\boldsymbol{x}'\boldsymbol{b} = a) < 1$

(A2) [FMSY Condition 2] ρ_1 and ρ_2 are bounded loss functions such that $\rho_2(u) \leq \rho_1(u)$

(A3) F_ϵ has strongly unimodal density

(A4) ρ_1 and ρ_2 satisfy [FMSY Condition 3], i.e. the function ρ is such that for some $m > 0$, $\rho(u) = 1$ if $|u| \geq m$ and $\log(1-\rho)$ is concave on $(-m, m)$

(A5) [FMSY, Theorem 6 - (iii), condition b)] $\sigma > 0$ and $b < 1 - c(F_0)$ where $c(F_0) = \sup\{P_{F_0}(\boldsymbol{x}'\boldsymbol{b} + a = 0) : \boldsymbol{b} \in R^p, a \in R, \boldsymbol{b} \neq \boldsymbol{0}\}$

(A6) [FMSY Condition 7] ρ_1 and ρ_2 are twice continuously differentiable with ψ_1 and ψ_2 being the derivatives of ρ_1 and ρ_2, respectively

(A7) [FMSY Condition 9] $E_G\left||\boldsymbol{x}|\right|^2 < \infty$

(A8) [FMSY, Theorem 6 - (iv)] $\boldsymbol{C_x}$ is invertible

(A9) [FMSY, Theorem 6 - (iv)] $a_{MM} = E_{F_\epsilon}\psi_2'(u_{MM}/\sigma) \neq 0$, $a_S = E_{F_\epsilon}\psi_1'(u_S/\sigma) \neq 0$, $d_s = E_{F_\epsilon}\{\psi_1(u_S/\sigma)(u_S/\sigma)\} \neq 0$

FMSY showed that (A1) - (A5) are sufficient conditions for the consistency of the robust scale and S- and MM- regression estimators, and that (A1)-(A9) are sufficient conditions for the joint asymptotic normality. In particular,

- Under (A1), (A3) and (A4) the solutions to (1.10), (1.12) and (1.14) are unique and $\boldsymbol{\beta_S} = \boldsymbol{\beta_{MM}} = \boldsymbol{\beta_0}$
- Under (A1) - (A5) $\hat{\boldsymbol{\eta}}' = \left(\hat{\boldsymbol{\theta}}'_{\boldsymbol{MM}}, \hat{\boldsymbol{\theta}}'_{\boldsymbol{S}}, \hat{\sigma}\right)'$ is consistent to $\boldsymbol{\eta}' = (\boldsymbol{\theta}'_{MM}, \boldsymbol{\theta}'_S, \sigma)$. Moreover, $\boldsymbol{\eta}$ satisfies $E_{F_0}\boldsymbol{\Psi}(\boldsymbol{z}, \boldsymbol{\eta}) = \boldsymbol{0}$ with

$$\boldsymbol{\Psi}(\boldsymbol{z},\boldsymbol{\eta}) = \begin{pmatrix} \psi_2\left(\dfrac{y-\boldsymbol{x}^{*\prime}\boldsymbol{\theta}_{MM}}{\sigma}\right)\boldsymbol{x}^* \\ \psi_1\left(\dfrac{y-\boldsymbol{x}^{*\prime}\boldsymbol{\theta}_{S}}{\sigma}\right)\boldsymbol{x}^* \\ \rho_1\left(\dfrac{y-\boldsymbol{x}^{*\prime}\boldsymbol{\theta}_{S}}{\sigma}\right)-b \end{pmatrix} \tag{A2.1}$$

where ψ_2 is the psi-function for the final step of the MM estimate while ρ_1 and $\psi_1 = \rho_1'$ are the loss function and corresponding psi function for the initial S-estimate.

- Under (A1) - (A9)

$$\sqrt{n}\,(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\,) \to N(\mathbf{0},\boldsymbol{\Sigma})$$

with

$$\boldsymbol{\Sigma} = E_{F_0}\{\boldsymbol{I}_{\widehat{\boldsymbol{\eta}},F_0}\boldsymbol{I}'_{\widehat{\boldsymbol{\eta}},F_0}\}$$

where $\boldsymbol{I}_{\widehat{\boldsymbol{\eta}},F_0}$ is the influence function of $\widehat{\boldsymbol{\eta}}$ at F_0. Since $\widehat{\boldsymbol{\eta}}$ is an M functional its' influence function is given by $\boldsymbol{I}_{\widehat{\boldsymbol{\eta}},F_0} = -\boldsymbol{D}^{-\mathbf{1}}\boldsymbol{\Psi}(\boldsymbol{z},\boldsymbol{\eta})$ (FMSY eq. (5.9)) with $\boldsymbol{D} = E_{F_0}\dot{\boldsymbol{\Psi}}(\boldsymbol{z},\boldsymbol{\eta})$. Thus, the joint asymptotic covariance matrix of $\widehat{\boldsymbol{\eta}}$ can also be written as

$$\boldsymbol{\Sigma} = \boldsymbol{D}^{-\mathbf{1}}\boldsymbol{\Omega}\boldsymbol{D}'^{-\mathbf{1}} \tag{A2.2}$$

with $\boldsymbol{\Omega} = E_{F_0}\{\boldsymbol{\Psi}(\boldsymbol{z},\boldsymbol{\eta}\,)\boldsymbol{\Psi}'(\boldsymbol{z},\boldsymbol{\eta})\}$.

For the consistency and asymptotic normality of the LS estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{LS}}$ we need an extra assumption

(A.10) F_ϵ has finite second moment.

Under (A.1), (A.7), (A.8) and (A.10) the LS estimator is consistent to $\boldsymbol{\theta}_{\boldsymbol{LS}} = (E\epsilon, \boldsymbol{\beta}_{\mathbf{0}})$ and asymptotically normal with the same covariance matrix structure as that of the M-estimators, namely $\boldsymbol{D}_{LS}^{-\mathbf{1}}\boldsymbol{\Omega}_{LS}\boldsymbol{D}'^{-\mathbf{1}}_{LS}$ with $\boldsymbol{\Psi}_{\boldsymbol{LS}}(\boldsymbol{z},\boldsymbol{\theta}_{\boldsymbol{LS}}) = (y - \boldsymbol{x}^{*\prime}\boldsymbol{\theta}_{LS})\boldsymbol{x}^*$.

We can now move to the derivation of the joint asymptotic distribution of $(\widehat{\boldsymbol{\theta}}_{\boldsymbol{LS}}, \widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}})$. Since $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}$ depends on the robust scale and, consequently, on the initial S-estimate it is important to consider all these estimators together. We append the joint vector $\widehat{\boldsymbol{\eta}}' \equiv (\widehat{\boldsymbol{\theta}}'_{MM}, \widehat{\boldsymbol{\theta}}'_{S}, \hat{\sigma})$ with $\widehat{\boldsymbol{\theta}}_{\boldsymbol{LS}}$ so that from now on $\widehat{\boldsymbol{\eta}}' = (\widehat{\boldsymbol{\theta}}'_{LS}, \widehat{\boldsymbol{\theta}}'_{MM}, \widehat{\boldsymbol{\theta}}'_{S}, \hat{\sigma})$ and

$$\boldsymbol{\Psi}(\boldsymbol{z}, \boldsymbol{\eta}) = \begin{pmatrix} (y - \boldsymbol{x}^{*\prime}\boldsymbol{\theta}_{LS})\boldsymbol{x}^{*} \\ \psi_2\left(\frac{y - \boldsymbol{x}^{*\prime}\boldsymbol{\theta}_{MM}}{\sigma}\right)\boldsymbol{x}^{*} \\ \psi_1\left(\frac{y - \boldsymbol{x}^{*\prime}\boldsymbol{\theta}_{S}}{\sigma}\right)\boldsymbol{x}^{*} \\ \rho_1\left(\frac{y - \boldsymbol{x}^{*\prime}\boldsymbol{\theta}_{S}}{\sigma}\right) - b \end{pmatrix} \tag{A2.3}$$

Given the asymptotic results for the M-estimators one may expect that under (A.1) - (A.10) the combined vector $\widehat{\boldsymbol{\eta}}' = (\widehat{\boldsymbol{\theta}}'_{LS}, \widehat{\boldsymbol{\theta}}'_{MM}, \widehat{\boldsymbol{\theta}}'_{S}, \hat{\sigma})$ is also asymptotically normal:

$$\sqrt{n}\,(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\,) \rightarrow N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{A2.4}$$

with $\boldsymbol{\eta}' = (\boldsymbol{\theta}'_{LS}, \boldsymbol{\theta}'_{MM}, \boldsymbol{\theta}'_{S}, \sigma)$ and

$$\boldsymbol{\Sigma} = \boldsymbol{D}^{-1}\boldsymbol{\Omega}\boldsymbol{D}'^{-1} \tag{A2.5}$$

where $\boldsymbol{D} = E_{F_0}\dot{\boldsymbol{\Psi}}(\boldsymbol{z}, \boldsymbol{\eta})$, $\boldsymbol{\Omega} = E_{F_0}\{\boldsymbol{\Psi}(\boldsymbol{z}, \boldsymbol{\eta}\,)\boldsymbol{\Psi}'(\boldsymbol{z}, \boldsymbol{\eta})\}$. The formal proof of (A2.4) is straightforward and is deferred to the end of this section, Meanwhile we show calculations of the covariance matrix between $\widehat{\boldsymbol{\theta}}_{\boldsymbol{LS}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}$, which, in turn, leads to the asymptotic result (2.7) for the difference $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$.

In the calculations below we will use the fact that asymptotic values of $\widehat{\boldsymbol{\eta}}$ satisfy

$$E_{F_0}\boldsymbol{\Psi}(\boldsymbol{z}, \boldsymbol{\eta}) = \mathbf{0} \tag{A2.6}$$

Denote

$$a_{MM} = E_{F_\epsilon}\psi_2'\left(\frac{u_{MM}}{\sigma}\right) \qquad a_S = E_{F_\epsilon}\psi_1'\left(\frac{u_S}{\sigma}\right)$$

$$e_{MM} = E_{F_\epsilon}\left\{\psi_2'\left(\frac{u_{MM}}{\sigma}\right)\frac{u_{MM}}{\sigma}\right\} \qquad e_S = E_{F_\epsilon}\left\{\psi_1'\left(\frac{u_S}{\sigma}\right)\frac{u_S}{\sigma}\right\}$$

$$f_{MM,S} = E_{F_\epsilon}\left\{\psi_2\left(\frac{u_{MM}}{\sigma}\right)\psi_1\left(\frac{u_S}{\sigma}\right)\right\} \qquad d_S = E_{F_\epsilon}\left\{\psi_1\left(\frac{u_S}{\sigma}\right)\frac{u_S}{\sigma}\right\}$$

$$f_{MM} = E_{F_\epsilon}\left\{\psi_2^2\left(\frac{u_{MM}}{\sigma}\right)\right\} \qquad f_S = E_{F_\epsilon}\left\{\psi_1^2\left(\frac{u_S}{\sigma}\right)\right\} \tag{A2.7}$$

$$f_{MM,LS} = E_{F_\epsilon}\left\{\psi_2\left(\frac{u_{MM}}{\sigma}\right)u_{LS}\right\} \qquad f_{S,LS} = E_{F_\epsilon}\left\{\psi_1\left(\frac{u_S}{\sigma}\right)u_{LS}\right\}$$

$$g_{MM} = E_{F_\epsilon}\left\{\left(\rho_1\left(\frac{u_S}{\sigma}\right) - b\right)\psi_2\left(\frac{u_{MM}}{\sigma}\right)\right\} \qquad g_S = E_{F_\epsilon}\left\{\left(\rho_1\left(\frac{u_S}{\sigma}\right) - b\right)\psi_1\left(\frac{u_S}{\sigma}\right)\right\}$$

$$= E_{F_\epsilon}\left\{\rho_1\left(\frac{u_S}{\sigma}\right)\psi_2\left(\frac{u_{MM}}{\sigma}\right)\right\} \qquad = E_{F_\epsilon}\left\{\rho_1\left(\frac{u_S}{\sigma}\right)\psi_1\left(\frac{u_S}{\sigma}\right)\right\}$$

$$g_{LS} = E_{F_\epsilon}\left\{\left(\rho_1\left(\frac{u_S}{\sigma}\right) - b\right)u_{LS}\right\} = E_{F_\epsilon}\left\{\rho_1\left(\frac{u_S}{\sigma}\right)u_{LS}\right\}$$

As before let $\boldsymbol{V}_{\boldsymbol{x}^*} = E(\boldsymbol{x}^*\boldsymbol{x}^{*\prime})$ and $\boldsymbol{\mu}'_{\boldsymbol{x}^*} = (\mathbf{1}, \boldsymbol{\mu}'_{\boldsymbol{x}})$**.** Since $\boldsymbol{x}$ and errors ϵ are assumed to be independent we get

$$\boldsymbol{\Omega} = E_{F_0}\{\boldsymbol{\Psi}(\boldsymbol{z},\boldsymbol{\eta})\boldsymbol{\Psi}'(\boldsymbol{z},\boldsymbol{\eta})\} = \begin{bmatrix} Eu_{LS}^2\boldsymbol{V}_{\boldsymbol{x}^*} & f_{MM,LS}\boldsymbol{V}_{\boldsymbol{x}^*} & f_{S,LS}\boldsymbol{V}_{\boldsymbol{x}^*} & g_{LS}\boldsymbol{\mu}_{x^*} \\ f_{MM,LS}\boldsymbol{V}_{\boldsymbol{x}^*} & f_{MM}\boldsymbol{V}_{\boldsymbol{x}^*} & f_{MM,S}\boldsymbol{V}_{\boldsymbol{x}^*} & g_{MM}\boldsymbol{\mu}_{x^*} \\ f_{S,LS}\boldsymbol{V}_{\boldsymbol{x}^*} & f_{MM,S}\boldsymbol{V}_{\boldsymbol{x}^*} & f_S\boldsymbol{V}_{\boldsymbol{x}^*} & g_S\boldsymbol{\mu}_{x^*} \\ g_{LS}\boldsymbol{\mu}'_{\boldsymbol{x}^*} & g_{MM}\boldsymbol{\mu}'_{\boldsymbol{x}^*} & g_S\boldsymbol{\mu}'_{\boldsymbol{x}^*} & E\left\{\rho_1^2\left(\frac{u_S}{\sigma}\right) - b^2\right\} \end{bmatrix} \tag{A2.8}$$

Note that above we used the fact that $E\{(\rho_1(u_S/\sigma) - b)^2\} = E\{\rho_1^2(u_S/\sigma) - b^2\}$ which follows from (A2.6). Next, compute

$$\dot{\boldsymbol{\Psi}}(\boldsymbol{z},\boldsymbol{\eta}) = \left\{\dot{\Psi}_{jk}(z,\eta) = \frac{\partial\Psi_j}{\partial\eta_k}\right\} = -\frac{1}{\sigma}\begin{bmatrix} \sigma\boldsymbol{x}^*\boldsymbol{x}^{*\prime} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \psi_2'\left(\frac{u_{MM}}{\sigma}\right)\boldsymbol{x}^*\boldsymbol{x}^{*\prime} & \mathbf{0} & \psi_2'\left(\frac{u_{MM}}{\sigma}\right)\boldsymbol{x}^*\frac{u_{MM}}{\sigma} \\ \mathbf{0} & \mathbf{0} & \psi_1'\left(\frac{u_S}{\sigma}\right)\boldsymbol{x}^*\boldsymbol{x}^{*\prime} & \psi_1'\left(\frac{u_S}{\sigma}\right)\boldsymbol{x}^*\frac{u_S}{\sigma} \\ \mathbf{0} & \mathbf{0} & \psi_1\left(\frac{u_S}{\sigma}\right)\boldsymbol{x}^{*\prime} & \psi_1\left(\frac{u_S}{\sigma}\right)\frac{u_S}{\sigma} \end{bmatrix}$$

From (A2.6) it follows that $E\psi_1(u_S/\sigma) = 0$. Thus,

$$\boldsymbol{D} = E_{F_0}\boldsymbol{\Psi}(\boldsymbol{z},\boldsymbol{\eta}) = -\frac{1}{\sigma}\begin{bmatrix}\sigma\boldsymbol{V}_{\boldsymbol{x}^*} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & a_{MM}\boldsymbol{V}_{\boldsymbol{x}^*} & \mathbf{0} & e_{MM}\boldsymbol{\mu}_{x^*}\\ \mathbf{0} & \mathbf{0} & a_S\boldsymbol{V}_{\boldsymbol{x}^*} & e_S\boldsymbol{\mu}_{x^*}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & d_S\end{bmatrix}$$

which implies

$$\boldsymbol{D}^{-\mathbf{1}} = \begin{Bmatrix}\text{inverse of a}\\ \text{partitioned matrix}\end{Bmatrix} = -\sigma\begin{bmatrix}\frac{1}{\sigma}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \frac{1}{a_{MM}}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1} & \mathbf{0} & -\frac{e_{MM}}{a_{MM}d_S}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1}\boldsymbol{\mu}_{x^*}\\ \mathbf{0} & \mathbf{0} & \frac{1}{a_S}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1} & -\frac{e_S}{a_S d_S}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1}\boldsymbol{\mu}_{x^*}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1}{d_s}\end{bmatrix} \qquad \text{(A2.9)}$$

We note that the (MM, S, σ) submatrix in (A2.9) with some change in notation coincides with the expression of D_0^{-1} in FMSY section 7.4.1.

Let partition $\boldsymbol{\Sigma}$ into blocks corresponding to LS, MM, S and σ subsets of $\boldsymbol{\eta}$ as follows

$$\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{\mathrm{LS}} & \boldsymbol{\Sigma}_{\mathrm{LS,MM}} & \boldsymbol{\Sigma}_{\mathrm{LS,S}} & \boldsymbol{\Sigma}_{\mathrm{LS},\sigma}\\ \boldsymbol{\Sigma}'_{\mathrm{LS,MM}} & \boldsymbol{\Sigma}_{\mathrm{MM}} & \boldsymbol{\Sigma}_{\mathrm{MM,S}} & \boldsymbol{\Sigma}_{\mathrm{MM},\sigma}\\ \boldsymbol{\Sigma}'_{\mathrm{LS,S}} & \boldsymbol{\Sigma}'_{\mathrm{MM,S}} & \boldsymbol{\Sigma}_{\mathrm{S}} & \boldsymbol{\Sigma}_{\mathrm{S},\sigma}\\ \boldsymbol{\Sigma}'_{\mathrm{LS},\sigma} & \boldsymbol{\Sigma}'_{\mathrm{MM},\sigma} & \boldsymbol{\Sigma}'_{\mathrm{S},\sigma} & \boldsymbol{\Sigma}_{\sigma}\end{bmatrix}$$

Plugging (A2.8) and (A2.9) into (A2.5) we compute

$$\boldsymbol{\Sigma}_{LS} = E(u_{LS}^2)\boldsymbol{V}_{\boldsymbol{x}^*}^{-1}$$

$$\boldsymbol{\Sigma}_{MM} = \sigma^2\frac{f_{MM}}{a_{MM}^2}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1} + \sigma^2\frac{e_{MM}}{a_{MM}^2 d_s}\left(E\left(\rho_1^2\left(\frac{u_S}{\sigma}\right) - b^2\right)\frac{e_{MM}}{d_s} - 2g_{MM}\right)\boldsymbol{V}_{\boldsymbol{x}^*}^{-1}\boldsymbol{\mu}_{x^*}\boldsymbol{\mu}'_{x^*}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1} \qquad \text{(A2.10)}$$

$$\boldsymbol{\Sigma}_{LS,MM} = \sigma\frac{f_{MM,LS}}{a_{MM}}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1} - \sigma\frac{e_{MM}g_{LS}}{a_{MM}d_S}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1}\boldsymbol{\mu}_{x^*}\boldsymbol{\mu}'_{x^*}\boldsymbol{V}_{\boldsymbol{x}^*}^{-1}$$

The above expressions can be simplified further by noting that from expression (1.23) for $\boldsymbol{V}_{x^*}^{-1}$ it follows that

$$\boldsymbol{V}_{x^*}^{-1}\boldsymbol{\mu}_{x^*} = \begin{bmatrix} 1+\boldsymbol{\mu}_x^{'}\boldsymbol{C}_x^{-1}\boldsymbol{\mu}_x & -\boldsymbol{\mu}_x^{'}\boldsymbol{C}_x^{-1} \\ -\boldsymbol{C}_x^{-1}\boldsymbol{\mu}_x & \boldsymbol{C}_x^{-1} \end{bmatrix}\begin{pmatrix} 1 \\ \boldsymbol{\mu}_x \end{pmatrix} = \begin{pmatrix} 1+\boldsymbol{\mu}_x^{'}\boldsymbol{C}_x^{-1}\boldsymbol{\mu}_x-\boldsymbol{\mu}_x^{'}\boldsymbol{C}_x^{-1}\boldsymbol{\mu}_x \\ -\boldsymbol{C}_x^{-1}\boldsymbol{\mu}_x+\boldsymbol{C}_x^{-1}\boldsymbol{\mu}_x \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{0} \end{pmatrix} \tag{A2.11}$$

and

$$\boldsymbol{V}_{x^*}^{-1}\boldsymbol{\mu}_{x^*}\boldsymbol{\mu}_{x^*}^{'}\boldsymbol{V}_{x^*}^{-1} = \begin{pmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}$$

We note that the resulting expression for $\boldsymbol{\Sigma_{MM}}$ coincides with the asymptotic variance expression one gets from FMSY Theorem 6, part (vii) using influence function expressions given by FMSY equations (6.11) and (6.12).

We are finally ready to state the main asymptotic result for the differences between the LS and MM regression estimators:

$$\sqrt{n}\left(\{\widehat{\boldsymbol{\theta}}_{LS}-\widehat{\boldsymbol{\theta}}_{MM}\}-\{\boldsymbol{\theta}_{LS}-\boldsymbol{\theta}_{MM}\}\right) \rightarrow N(\boldsymbol{0},\boldsymbol{V}) \tag{A2.12}$$

where

$$\boldsymbol{V} = E\left\{\left(u_{LS}-\frac{\sigma\psi_2\left(\frac{u_{MM}}{\sigma}\right)}{E\psi_2'\left(\frac{u_{MM}}{\sigma}\right)}\right)^2\right\}\begin{bmatrix} 1+\boldsymbol{\mu}_x^{'}\boldsymbol{C}_x^{-1}\boldsymbol{\mu}_x & -\boldsymbol{\mu}_x^{'}\boldsymbol{C}_x^{-1} \\ -\boldsymbol{C}_x^{-1}\boldsymbol{\mu}_x & \boldsymbol{C}_x^{-1} \end{bmatrix} + \omega_{LS,MM}\begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \tag{A2.13}$$

and

$$\omega_{LS,MM} = \sigma^2\frac{e_{MM}}{a_{MM}d_s}\left(2\frac{g_{LS}}{\sigma}-2\frac{g_{MM}}{a_{MM}}+\frac{E\left(\rho_1^2\left(\frac{u_S}{\sigma}\right)-b^2\right)e_{MM}}{a_{MM}d_s}\right) \tag{A2.14}$$

with e_{MM}, a_{MM}, d_s , g_{MM} and g_{LS} are defined in (A2.7), u_{MM} and u_{LS} are defined in (1.22) and (2.9) and

$$u_S = \epsilon - \alpha_S \tag{A2.15}$$

Expression (A2.13) follows from the fact that $\boldsymbol{V} = \boldsymbol{\Sigma}_{LS} - \boldsymbol{\Sigma}_{LS,MM} - \boldsymbol{\Sigma}'_{LS,MM} + \boldsymbol{\Sigma}_{MM}$ and expressions (A2.10) and (A2.11).

The asymptotic result for the differences between the LS and MM estimators of the slopes is given by the lower $p \times p$ sub-matrix of (A2.13) and is the same under symmetric and asymmetric error distributions F_ϵ. This is, however, not the case for the intercept:

a) If F_ϵ is symmetric around α_0 then $\alpha_{LS} = \alpha_{MM} = \alpha_0$, $e_{MM} = E_{F_\epsilon}\{\psi_2^{'}(u_{MM}/\sigma)\, u_{MM}/\sigma\} = 0$ and, therefore, the extra term in (A2.13) disappears, i.e. $\omega_{LS,MM} = 0$

b) If F_ϵ is asymmetric then, in general, $\alpha_{LS} \neq \alpha_{MM}$ and $\omega_{LS,MM} \neq 0$

Proof of Joint Asymptotic Normality

Under the assumptions (A.1) – (A.10) vector $\widehat{\boldsymbol{\eta}} \equiv \left(\widehat{\boldsymbol{\theta}}'_{\boldsymbol{LS}}, \widehat{\boldsymbol{\theta}}'_{\boldsymbol{MM}}, \widehat{\boldsymbol{\theta}}'_{\boldsymbol{S}}, \widehat{\boldsymbol{\sigma}}\right)$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{z_i}, \widehat{\boldsymbol{\eta}}) = \boldsymbol{0}, \tag{A2.16}$$

vector $\boldsymbol{\eta}$ satisfies (A2.6) and $\widehat{\boldsymbol{\eta}} \rightarrow_p \boldsymbol{\eta}$.

Following FMSY (proof of Theorem 4) we use the Mean Value Theorem[10] to write

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{z_i}, \boldsymbol{\eta}) + \boldsymbol{D_n}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{z_i}, \widehat{\boldsymbol{\eta}}) = \boldsymbol{0}$$

[10] Note that the mean value theorem is essentially used only for the S- and MM- parts of Ψ while for the LS part the result is exact, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi_{LS}}(\boldsymbol{z_i}, \widehat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i^{*\prime}}\widehat{\boldsymbol{\theta}}_{LS}\right)\boldsymbol{x_i^*} = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \boldsymbol{x_i^{*\prime}}\boldsymbol{\theta}_{LS}\right)\boldsymbol{x_i^*} + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i^{*\prime}}(\boldsymbol{\theta}_{LS} - \widehat{\boldsymbol{\theta}}_{LS})\boldsymbol{x_i^*}$$
$$= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi_{LS}}(\boldsymbol{z_i}, \boldsymbol{\theta_{LS}}) + \boldsymbol{D_{n,LS}}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{LS}} - \boldsymbol{\theta}_{LS})$$

where $\boldsymbol{D_n} = \frac{1}{n}\sum_{i=1}^{n} \dot{\boldsymbol{\Psi}}(\boldsymbol{z}, \boldsymbol{\eta}_n^*)$ with $\boldsymbol{\eta}_n^* \to \boldsymbol{\eta}$.

Since for large n, $\boldsymbol{D_n}$ is non-singular, we may write

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} = -\boldsymbol{D_n^{-1}} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{z_i}, \boldsymbol{\eta}) = -\boldsymbol{D^{-1}} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{z_i}, \boldsymbol{\eta}) + \left(\boldsymbol{D^{-1}} - \boldsymbol{D_n^{-1}}\right) \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{z_i}, \boldsymbol{\eta}) \qquad \text{(A2.17)}$$

Let $\boldsymbol{I}(\boldsymbol{z_i}, \boldsymbol{\eta}) = -\boldsymbol{D^{-1}}\boldsymbol{\Psi}(\boldsymbol{z_i}, \boldsymbol{\eta})$. For $i = 1, 2, \ldots$, the vectors $\boldsymbol{\Psi}(\boldsymbol{z_i}, \boldsymbol{\eta})$ are i.i.d. with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$. The multivariate central limit theorem implies

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I}(\boldsymbol{z_i}, \boldsymbol{\eta})\right\} \to_d N(\mathbf{0}, \boldsymbol{D^{-1}\Omega D'^{-1}})$$

To complete the proof we need to show that the second term in (A2.17) multiplied by $\sqrt{n}$ is $\boldsymbol{o_p}(\mathbf{1})$, i.e. converges in probability to $\mathbf{0}$. This can be decomposed into two separate problems, one for LS and one for (MM, S, σ), since the (LS, MM), (LS,S) and (LS, σ) blocks of $\boldsymbol{D^{-1}}$ and $\boldsymbol{D_n^{-1}}$ are $\mathbf{0}$ so that

$$\sqrt{n}\left(\boldsymbol{D^{-1}} - \boldsymbol{D_n^{-1}}\right) \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{z_i}, \boldsymbol{\eta}) = \sqrt{n} \begin{pmatrix} \left(\left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x_j^* x_j^{*\prime}}\right)^{-1} - \boldsymbol{V_{x^*}^{-1}}\right) \frac{1}{n}\sum_{i=1}^{n} \left(y - \boldsymbol{x_i^{*\prime}\theta}_{LS}\right) \boldsymbol{x_i^*} \\ \left(\boldsymbol{D_{MM,S,\sigma}^{-1}} - \boldsymbol{D_{n,MM,S,\sigma}^{-1}}\right) \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi_{MM,S,\sigma}}\left(\boldsymbol{z_i}, \boldsymbol{\eta}_{MM,S,\sigma}\right) \end{pmatrix}$$

For (MM, S, σ) part, the result was proved by FMSY (see Theorem 4, Theorem 5 and the proof of Theorem 6, part (iii)).

For LS part, the result follows from the standard arguments of the multivariate central limit theorem and law of large numbers:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\left(y - \boldsymbol{x_i^{*\prime}\theta}_{LS}\right)\boldsymbol{x_i^*}\right) \to \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Omega_{LS}}) \text{ (CLT)},$$

$$\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x_j^* x_j^{*\prime}} \rightarrow_p \boldsymbol{V_{x^*}} \quad \text{(LLN)}$$

Hence, by continuous mapping theorem $\left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x_j^* x_j^{*\prime}}\right)^{-1} \rightarrow_p \boldsymbol{V_{x^*}^{-1}}$ provided $\boldsymbol{V_{x^*}}$ is non-singular and the product of $\boldsymbol{O_p(1)} \cdot \boldsymbol{o_p(1)}$ is $\boldsymbol{o_p(1)}$**.** Q.E.D.

3. Tests for Differences between Least Squares and Robust Regression Estimators: Further Analysis

3.1 Introduction

The goal of this chapter is threefold. First, we follow up on the Monte Carlo simulation results from Chapter 2, which revealed several unexpected properties of tests T1 and T2 in finite samples. We present an empirical explanation of some of these behaviors. Second, we explore whether using low normal distribution efficiency in T1 and T2 could be beneficial, i.e. whether it could lead to higher power while maintaining proper level. To this end, we repeat Monte Carlo simulations for the five models from Chapter 2 for the MM estimators with normal distribution efficiencies of 70%, 50% and 29%. This is an intermediate step before turning to the last goal of this chapter, which is to study the performance of the Yohai, Stahel and Zamar test of bias (Yohai et al., 1991) and to compare it to our test T2. The test was originally developed by these authors to detect a larger overall bias in the final MM estimator as compared to the highly robust, but low efficiency initial S-estimator. The test can easily be extended for testing of overall bias in the LS estimator versus that of the initial S-estimator.

The rest of the chapter is organized as follows. Section 3.2 explores in more detail the differences in the null distributions of the two test statistics, T1 and T2, under normal errors in finite samples. Section 3.3 investigates some peculiar behavior of the power of the two tests under the asymmetric bivariate contamination Model 5 of Chapter 2. Section 3.4 presents Monte Carlo simulation results for T1 and T2 with a bisquare loss function at low normal distribution efficiencies. Section 3.5 discusses a third test of overall bias, T3, and results of the Monte Carlo simulations. In section 3.6, we conclude the chapter with a summary and discussion. Technical details are deferred to Appendix A3.

3.2 T1 and T2 Null Distributions in Finite Samples

In this section we explain the important differences in the finite sample performance of the T1 and T2 tests under normal residuals which we observed in Figure 7 in Chapter 2. Figure 7 shows Monte Carlo

rejection rates (based on 10 000 replicates) versus sample size of the two tests T1 and T2 when errors are i.i.d. standard normal.

In Chapter 2 we showed that when F_ϵ is a normal distribution then $EFF = 1/\tau$ and $\delta^2_{LS,MM} = (\tau - 1)\sigma^2 = (1 - EFF)\tau\sigma^2$, i.e. the asymptotic variance in (2.7) is equal to the asymptotic variance in (2.3), as it should be. Thus, in finite samples under normality one may expect similar behavior of the $\widehat{\delta^2_{LS,MM}}$ and $(1 - EFF)\hat{\tau}\hat{\sigma}_1^2$ and, consequently, of the T1 and T2 test statistics. Surprisingly, Figure 7 clearly suggests that this is not the case.

It turns out that the ad hoc application of the Hausman result (2.3) ignores the data based dependency between the LS and MM estimates. This subtle, but important aspect of the finite sample distribution of the differences between the two regression estimates is automatically incorporated in the asymptotic normal distribution approximation (2.7). All graphical illustrations in this section are based on the Monte Carlo simulations with bisquare MM estimates at 95% efficiency. We focused on two sample sizes, 'moderate' n=100 and 'large' n=500, and generated 50 000 replicates of the datasets with normal residuals as per Model 1 from Chapter 2. The shape of the optimal loss function leads to numerical issue with the T2 test statistic which is typically not observed for the bisquare loss function. This will be discussed at the end of this section.

First we note that in spite of the asymptotic equivalence of T1 and T2 under normality, their finite sample behaviors are quite different in that unlike the latter the former is failing to have a normal distribution in finite samples. This is the case in spite of reasonably good finite sample normal distribution approximations for both the LS and robust estimates themselves. Figure 15 illustrates this claim quite clearly.

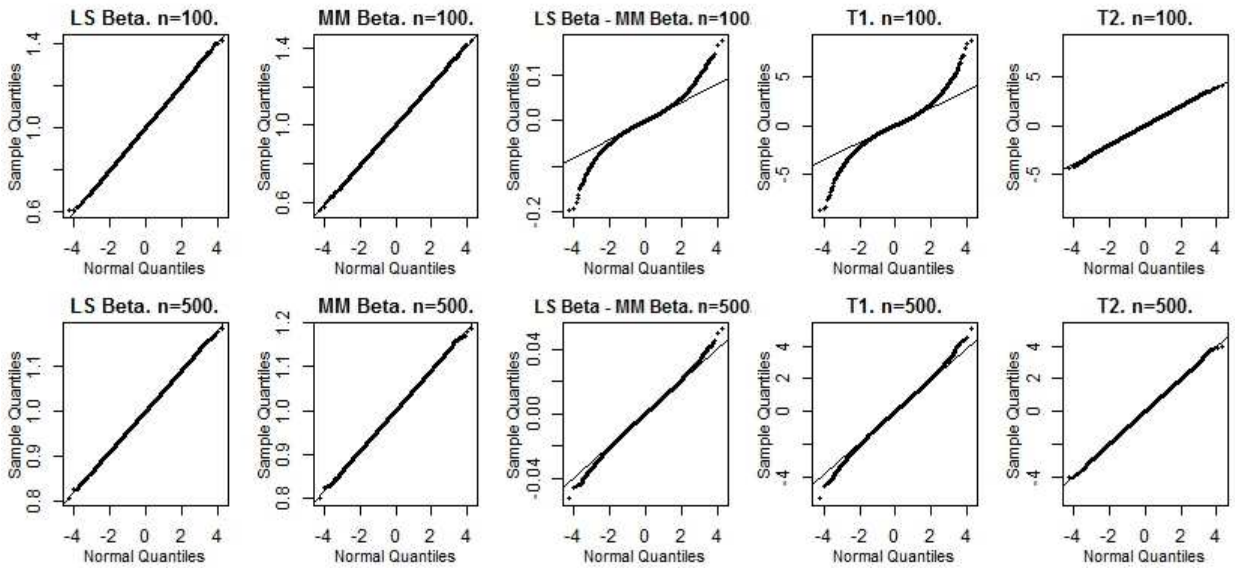

Figure 15. Normal quantile-quantile plots of LS and robust estimates of beta, of the differences between LS and robust betas and of T1 and T2 test statistics for sample sizes n=100 (top row) and n=500 (bottom row). Simple linear regression with normal errors. Bisquare at 95% efficiency.

The only difference between the two test statistics is that the scalar multipliers in the standard errors are different, namely $(1 - EFF)\hat{\tau}\hat{\sigma}_1^2$ and $\hat{\delta}_{LS,MM}$ for T1 and T2 respectively. These two scalar multipliers have radically different finite sample distributions under normal residuals. Their histograms are displayed in Figure 16. Even at n=500 the distribution of $\hat{\delta}_{LS,MM}$ has substantially wider spread than that of the T1 scalar multiplier and is asymmetric. We explain these observed behaviors subsequently.

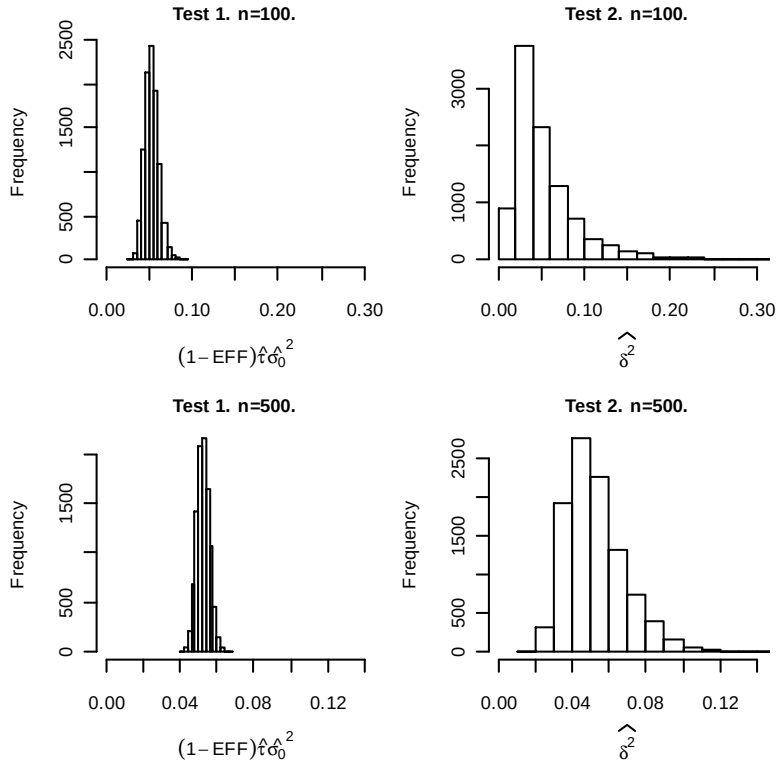

Figure 16. Histograms of the T1 and T2 scalar multipliers for sample sizes n=100 (top row) and n=500 (bottom row). Simple linear regression with normal errors. Bisquare at 95% efficiency.

The reason for the above differences between T1 and T2 arise because the finite sample correlations between LS and MM betas depend on the data configuration. In particular, the data configuration will affect the distribution of the MM residual weights[11] in a sample. In samples where all the weights are one or close to one the correlation between the LS and MM beta will be high leading to very small differences between the two estimates. Sometimes, especially with the optimal loss function, this difference can be exceedingly small, e.g. 10^{-10}, which can result in a serious problem with T2 that we describe later. On the other hand, as the proportion of smaller weights in the sample increases the correlation between the LS and MM estimates will be smaller, and their differences will exhibit higher variability. As a result of the sample to sample variation in the data configurations, we conjecture that the unconditional distribution of the beta differences is likely to be a scale mixture of normal distributions with a large number of

[11] Robustness weights for M-estimates are defined in Chapter 1.

components, e.g., in the limit something like a t-distribution. See, for example, the normal qq-plots of the differences between LS and bisquare betas in Figure 15.

T1 standard errors, however, are based on the assumption that correlation between the LS and robust betas is equal to $\sqrt{EFF}$. Therefore, T1 denominator doesn't depend on the data configurations[12] and the non-normal distribution of T1 in a finite sample is a direct consequence of the non-normal distribution of the beta differences, i.e., of its numerator. We notice that sample size needs to be extremely large, e.g. over 500, for the correlation dependence to become negligible.

On the contrary, $\widehat{\delta^2_{LS,MM}}$ exhibits similar strong dependence on the data configurations as beta differences. The ratio of the differences and the estimated δ^2 tends to be 'stabilized' by such correlated behavior of the two what results in a more normal distribution of T2. An exception is when the values of $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ and $\widehat{\delta^2_{LS,MM}}$ are nearly zero, which may occur for certain loss functions (optimal in particular, but not the bisquare) as we discuss at the end of the section.

Figure 17 to Figure 22 illustrate the claims for bisquare robust estimates at 95% efficiency and sample size n=100. Data configuration in a sample of size n is proxied by an average weight, i.e. $\frac{1}{n}\sum_{i=1}^{n} w_i$, where w_i are the final bisquare weights scaled to be equal to 1 for a zero residual. The histogram of the average weights is shown in Figure 17. Note that the expected weight under normal errors, i.e. the average weight in an infinitely large sample, is .9152 while the average weight in our simulated samples of size n=100 varied from 0.8454 to 0.9395.

[12] There may still be some weak dependence via the $\hat{\tau}$ and $\hat{\sigma}_1$ (e.g., see Figure 17)

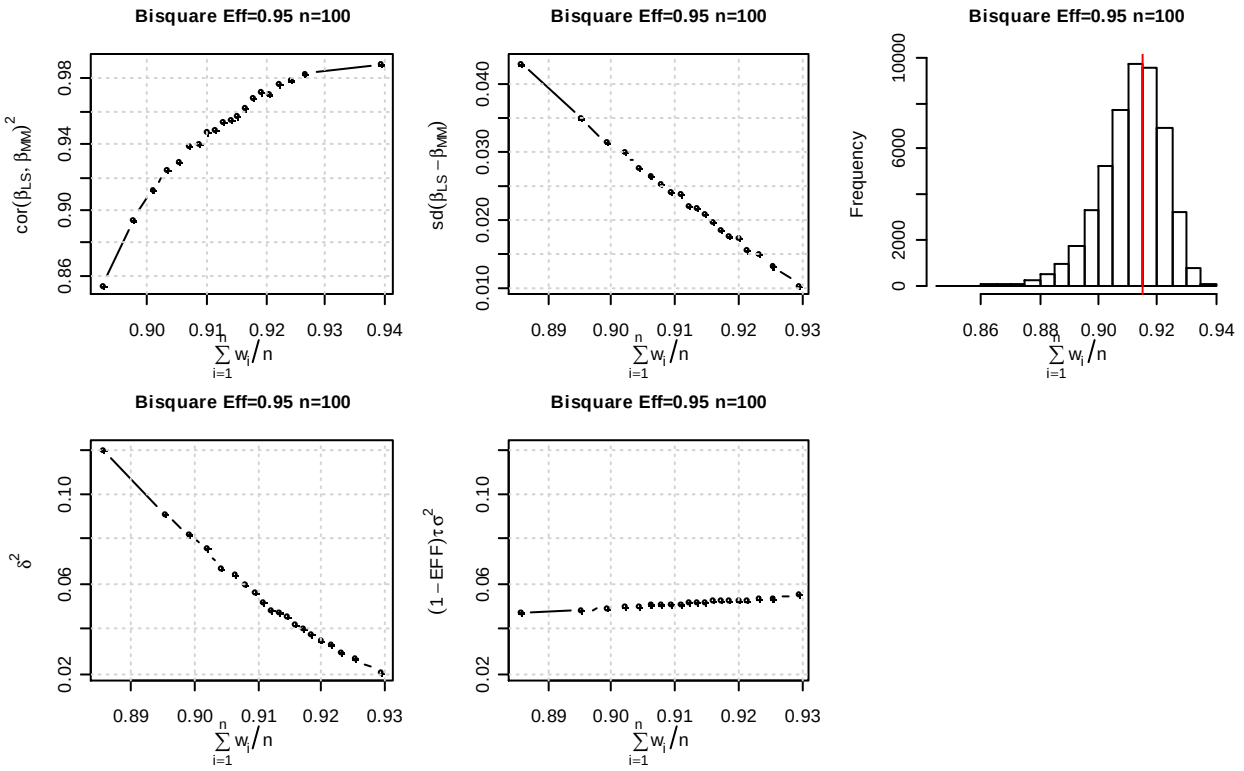

Figure 17. Squared correlation between the LS and robust betas, standard deviation of the differences, scalar multipliers in T2 and T1 standard errors vs average weight.

For further illustrations the datasets are split into 20 equally-sized groups based on the sample quantiles of the average weights. For each group we compute squared correlation between the LS and bisquare estimates, standard deviations of the differences, average $\widehat{\delta^2_{LS,MM}}$ and average $(1 - EFF)\hat{\tau}\hat{\sigma}_1^2$. These are plotted in Figure 17 against the mean average weight of the corresponding group. Recall that T1 is based on the assumption that the squared correlation is equal to the normal distribution efficiency, i.e. $\rho^2 = 0.95$. While this is the case for the samples with the average weight close to 0.9152 (aka 'ideal normal samples'), the correlation is substantially different in other, 'less normal', samples. For example, the squared correlation in Figure 17 varies from 0.85 to nearly 0.99. As expected, the correlation increases with the average weight. The $\widehat{\delta^2_{LS,MM}}$ decreases as average weight gets closer to one, similarly to the standard deviation of the differences. This is not the case for the T1 standard errors multiplier. Note that the largest $sd(\hat{\beta}_{LS} - \hat{\beta}_{MM}) \approx 0.04$ is four times the smallest $sd(\hat{\beta}_{LS} - \hat{\beta}_{MM}) \approx 0.01$.

Next we present a set of trellis displays (Becker, Cleveland, & Shyu, 1996; Sarkar, 2008) conditioning on the data configurations as proxied by the average weight categories. They show conditional distributions of the following quantities: beta differences, the T2 and T1 standard errors as well as the T2 and T1 test statistics themselves. Figure 18 and Figure 19 illustrate how distributions of the numerator and denominator of the T2 test statistic change in concordance with each other. Small average weight (top left panel) indicates substantial proportion of 'large outlying' residuals and, hence, we observe larger variability of the beta differences. This is accompanied by larger location and wider spread of the distribution of the $\widehat{\delta^2_{LS,MM}}$ and, consequently, of the T2 standard errors. As average weight increases the spread of the beta differences gets smaller and, at the same time, the corresponding T2 standard errors also tend to be smaller. As a result, the conditional distributions of the T2 test statistic (Figure 21) are fairly close to standard normal on all panels, i.e. the scale mixing effect is weak and, thus, the unconditional distribution is close to standard normal. On the contrary, the conditional distributions of the T1 standard errors (Figure 20) look quite similar across panels. The conditional distributions of the T1 test statistics (Figure 22) exhibit much wider range of scales than those of T2 suggesting a strong scale mixing effect that leads to non-normal unconditional distribution.

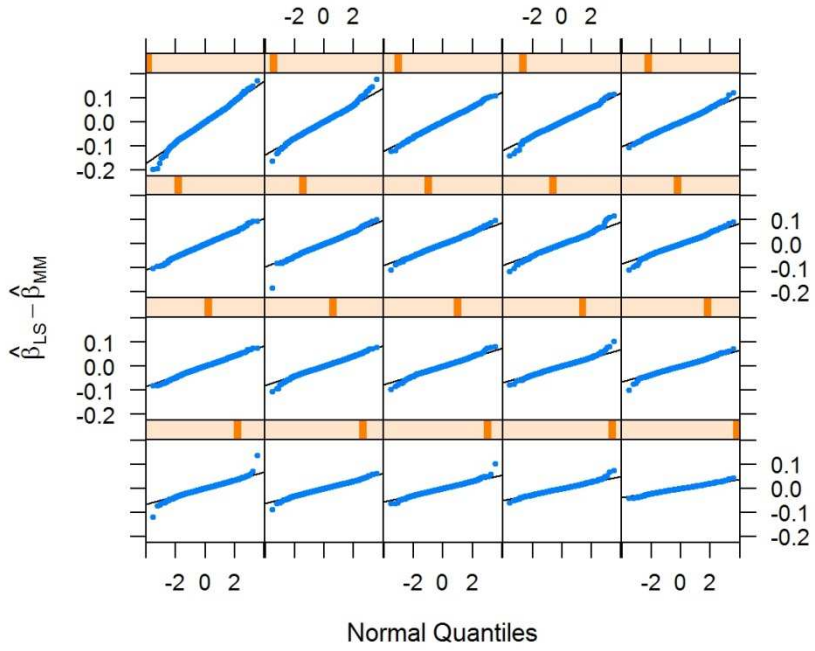

Figure 18. Beta differences conditional on the average weight category. Bisquare at 95% efficiency, n=100.

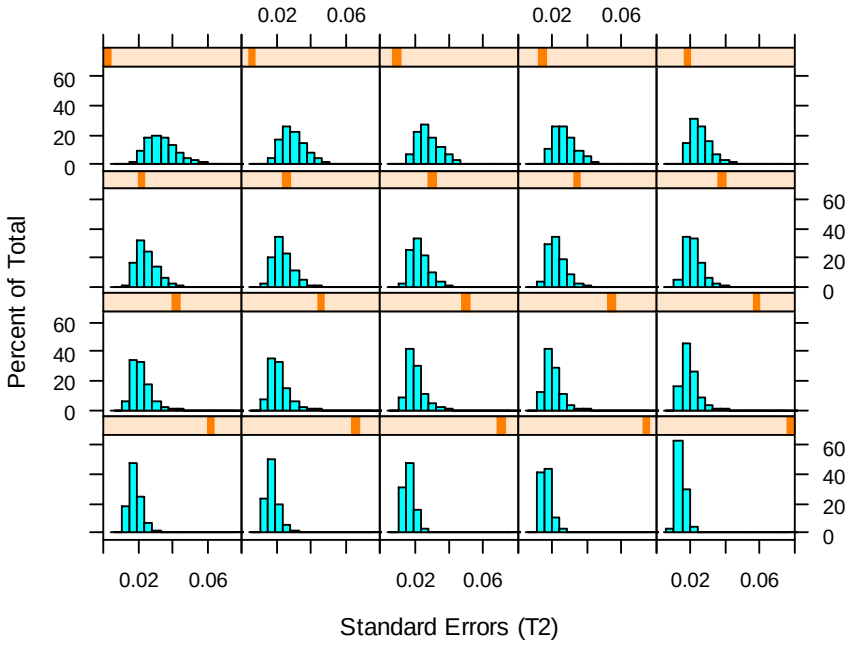

Figure 19. T2 standard errors conditional on the average weight category. Bisquare at 95% efficiency, n=100.

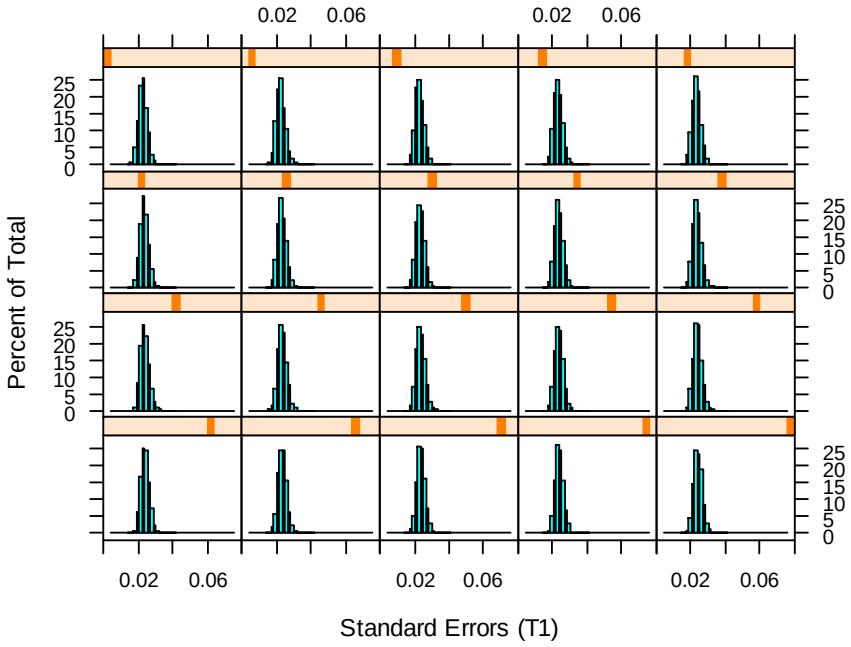

Figure 20. T1 standard errors conditional on the average weight category. Bisquare at 95% efficiency, n=100.

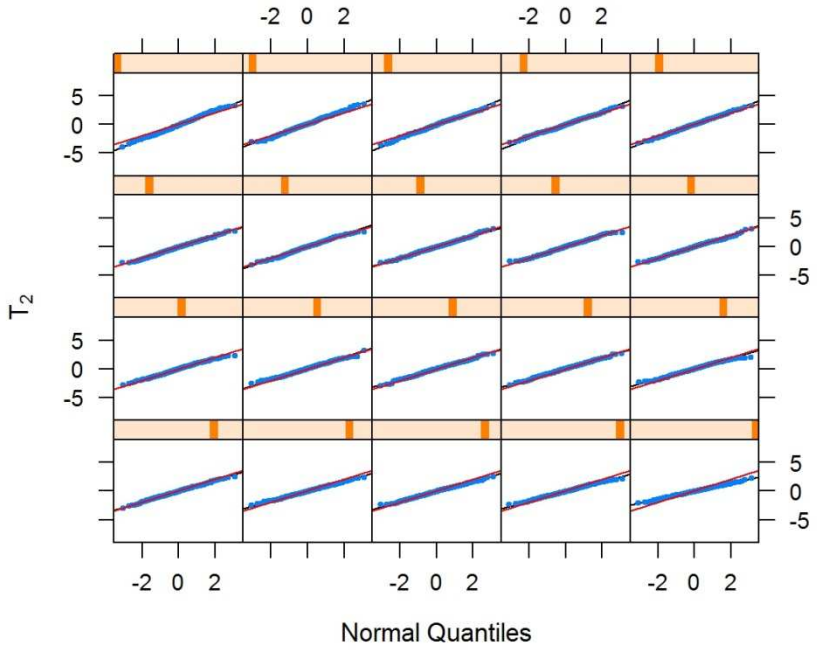

Figure 21. T2 test statistic conditional on the average weight category. Red line is a 45 degrees line. Bisquare at 95% efficiency, n=100.

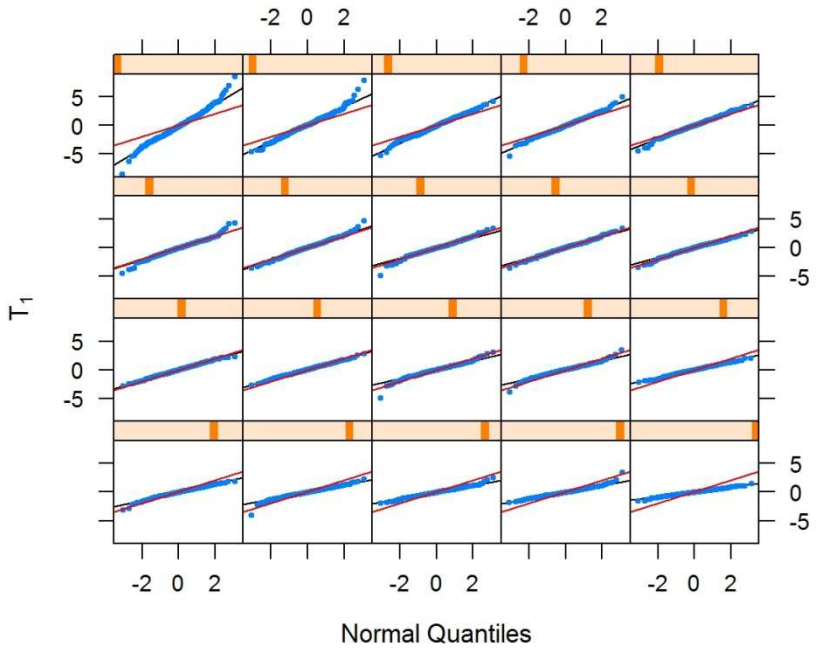

Figure 22. T1 test statistic conditional on the average weight category. Red line is a 45 degrees line. Bisquare at 95% efficiency, n=100.

While the dependence of $\widehat{\delta^2_{LS,MM}}$ on data configurations in finite samples is beneficial for the bisquare it creates a numerical problem for the optimal loss function in small to moderate sample sizes, especially at higher normal distribution efficiencies. In this case extremely small values of beta differences and T2 standard errors can occur and T2 acts like a 0/0 indeterminate. Recall that optimal psi-function is linear in the middle region (Figure 1, bottom right). When all residuals in a given sample fall inside that region, i.e. all robust weights are 1, then $\psi'(u) = 1$ and $\widehat{\delta^2_{LS,MM}} = ave_i\{(r_i^{LS} - r_i^{MM})^2\}$. In this case the MM and LS estimates are almost identical[13] leading to the two residual vectors being very similar and $\widehat{\delta^2_{LS,MM}}$ also being almost zero. Dividing nearly zero $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ by nearly zero standard errors is numerically unstable and most of the time leads to extreme values of the test statistic and overestimated rejection rates (see Figure 23). To deal with this issue we propose replacing all beta differences that are below a numerical precision threshold tol with exact 0 and treat them as non-significant. Also, any $\widehat{\delta^2_{LS,MM}} < tol$ is replaced with tol. The two cases tend to occur simultaneously. We used this method uniformly for T2 throughout the dissertation with $tol = 10^{-10}$. Setting tol to 10^{-8} or to 10^{-15} or even lower had no visible effect on the rejection rates. We want to note that poor behavior of the level of T2 under normality is unavoidable whether one does thresholding or not (Figure 7 vs Figure 23). However, it seems more reasonable to us to set a threshold and give user a warning message whenever T2 test statistic value cannot be trusted because of extremely low values of the numerator and denominator.

[13] Theoretically, when all weights are 1 and scale σ is known, the MM and LS estimates are exactly the same. In practice, however, they often differ by some non-negligible amount since $\widehat{\boldsymbol{\beta}}_{MM}$ is computed via numerical optimization algorithm that requires an estimate of σ and runs only till certain tolerance threshold is achieved.

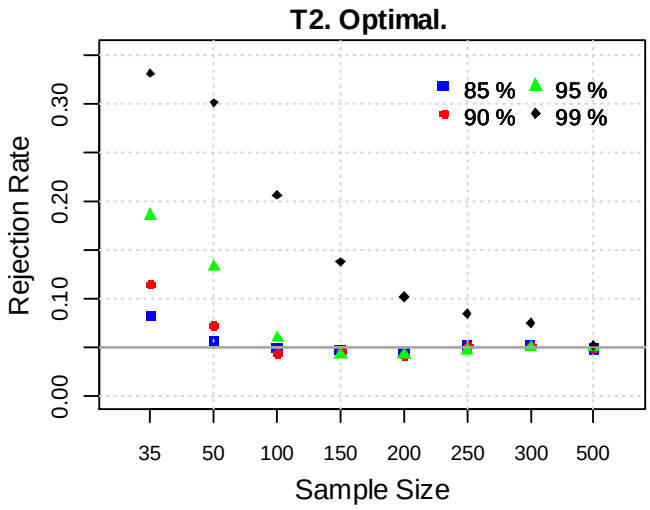

Figure 23. Model 1. Level of T2 test statistic for slope β in a simple linear regression under normal residuals without thresholding. Grey horizontal line marks nominal significance level of 0.05.

Table 5 illustrates how often the problem with T2 occurs for the optimal robust estimates and normally distributed errors. The proportion of samples with nearly zero beta differences grows drastically with normal distribution EFF, e.g. 0.03 for 85% vs 0.6 for 99% efficiency at n=50, and decreases as sample size grows[14]. This jump explains the gap observed in Figure 7 between the poor behavior of T2 level for high efficiencies and acceptable performance for low efficiencies. The number of samples with thresholding was substantially smaller for non-normal error models from Chapter 2 since those models are more likely to generate residuals that are down-weighted by the optimal MM estimator. This explains much better behavior of T2 level with optimal loss function observed in Figure 8 through Figure 10 in Chapter 2. In our simulations the extremely small values never occurred for the bisquare loss function. It may be not surprising given that the bisquare weight is equal to one only for an exactly zero residual and smoothly down-weights all non-zero residuals (see Figure 2).

[14] The higher the normal distribution efficiency EFF the smaller the probability of down-weighting a data point sampled from a normal distribution by the optimal psi-function: the probabilities are .0826, .0590, .0340 and only .0099 for 85%, 90%, 95% and 99% efficiency respectively. The thresholding occurred primarily in the samples with all weights equal to 1, but sometimes in the samples with a few weights slightly below 1.

Table 5. Number of samples (out of 10,000) with the absolute differences between LS beta and optimal MM beta below tolerance, i.e. with $\left|\hat{\beta}_{OLS} - \hat{\beta}_{MM}\right| < 10^{-10}$.

	Normal Distribution Efficiency			
n	85%	90%	95%	99%
35	755	1625	3439	6721
50	270	772	2232	6041
100	9	62	516	3864
150	1	3	121	2514
200	0	0	34	1638
250	0	0	12	1045
300	0	0	2	697
500	0	0	0	130

3.3 The Effect of the Mixture Component with Standard Deviation .25 when $\mu = 4$

In this section we take a closer look at the differences between LS and MM slope estimates under asymmetric bivariate contamination Model 5 from Chapter 2. We discuss the effect of the mixture component with $\mu = 4$ on the power of T2 and note that the results for T1 are quite similar.

Figure 24 displays histograms of the differences $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ and the resulting test statistic T2 for bisquare and optimal psi-functions at 99% and 85% normal distribution efficiencies. It is evident that at 85% normal distribution efficiency the distributions of the differences $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ and the resulting test statistic T2 are virtually the same irrespective of the choice of bisquare versus optimal psi-functions. However there are significant differences in the results for the bisquare and optimal psi-functions at 99% efficiency. We now explain the reasons for these behaviors.

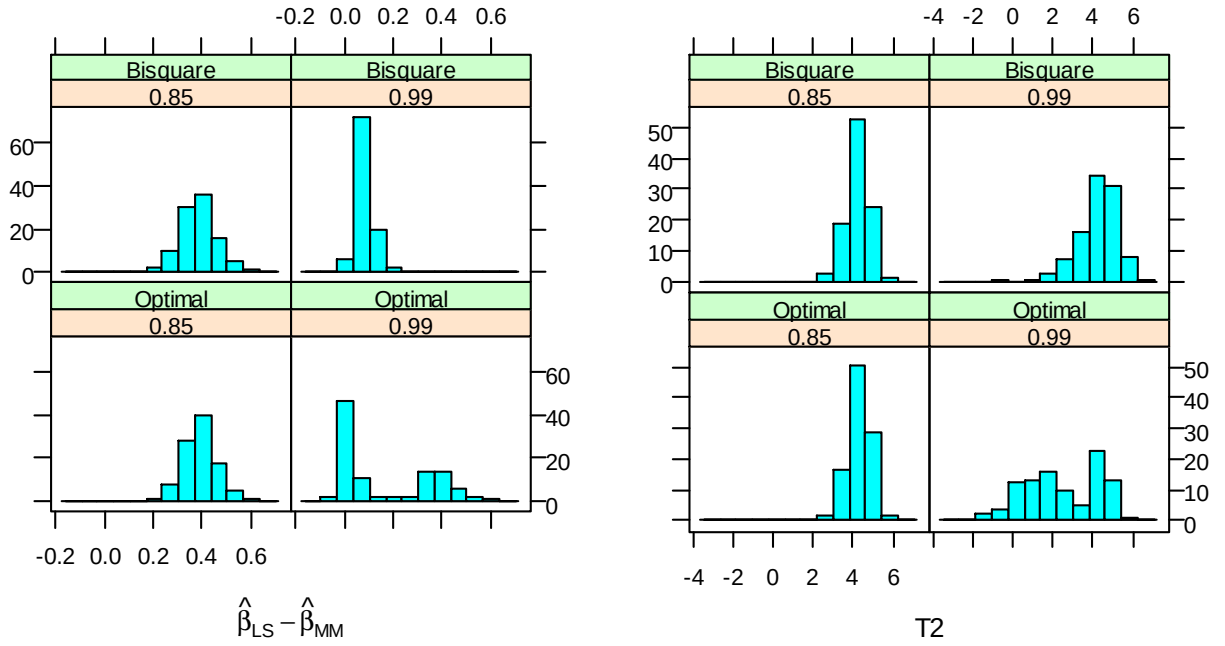

Figure 24. Distribution of the beta estimate differences (left) and T2 test statistic (right) under Model 5 with $\mu = 4$ and $\gamma = 0.06$ for sample size n=100 and two normal distribution efficiencies 85% and 99%.

First we consider the optimal psi-function for 99% normal distribution efficiency in which case residuals larger than 3.87 in magnitude are rejected as outliers. Thus when $\mu = 4$ and the standard deviation is .25 the mixture component $N(\mu = 4, 0.25^2)$ of Model 5 generates data that are borderline outliers that are sometimes rejected and sometimes not rejected by the optimal psi-function. This fact along with the rapid descent of the optimal psi-function results in non-normal bi-modality of the finite-sample distributions of the differences $\hat{\beta}_{LS} - \hat{\beta}_{MM}$, and consequently of the T1 and T2 test statistics. This is reflected in the bottom right panels of Figure 24 where the histograms of $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ and T2 under Model 5 with $\mu = 4$, $\gamma = 0.06$ and sample size n=100 are shown.

Now consider the case of the optimal psi-function with 85% normal distribution efficiency, for which residuals larger than 2.604 in magnitude are rejected. In this case essentially all of the data values generated by the mixture component are much larger than the rejection point and hence rejected. Consequently the corresponding histogram in the lower left panel of Figure 25 no longer exhibits bi-modality and reflects a reasonably normal distribution.

Now we turn to the bisquare psi-functions. At 99% normal distribution efficiency residuals larger in magnitude than 7.04 are rejected. So for 99% efficiency essentially all of the data generated by the mixture component $N(\mu = 4, 0.25^2)$ are not rejected, but somewhat down-weighted. Thus one does not expect bi-modality due to some residuals being rejected and some receiving high weight as in the case of the optimal psi-function at 99% efficiency. Indeed the histogram in upper right panel of Figure 24 is consistent with this expectation.

For the bisquare psi-function at 85% efficiency all residuals larger than 3.44 are rejected. In this case most of the residuals generated by this mixture component are rejected, and again one does not expect bi-modality and this is confirmed in the upper left panels of Figure 24. The reason for the similarity in results for the bisquare and optimal psi-functions at 85% efficiency in Figure 24 is that in this case both psi-functions reject most of the residuals generated by the mixture component.

It is interesting and not surprising to note that the differences $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ in Figure 24 (left) are much smaller for 99% efficiency than for 85% efficiency. This is explained by the fact that for the mixture component $N(\mu = 4, 0.25^2)$ residuals are somewhat down-weighted, but hardly ever rejected by the bisquare psi at 99% efficiency while essentially all such residuals are rejected at 85% efficiency. As a result, the bisquare estimates at 99% efficiency tend to be closer to the least squares estimates than those at 85%. Nonetheless, the top row of the right hand panel of Figure 24 reveals that at 99% efficiency and sample size n=100 nearly all the differences $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ are statistically significant in spite of their small values.

We also point out that for the bisquare psi-function there is unlikely to be bi-modality in the distributions of the differences $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ even if the mixture component generates residuals near the rejection point. For example, in the case of the mixture component $N(\mu = 7, 0.25^2)$ and 99% efficiency, the slow descent of the psi-function will result in about half of the residuals being rejected and half being smoothly down-weighted. This is in sharp contrast to the optimal psi-function in a similar context, as, for example in the case of the mixture component $N(\mu = 4, 0.25^2)$ and 99% efficiency we described earlier, where half the residuals would be rejected and the other half would be given large weight.

With the above differences in T2 test statistic behavior for the bisquare and optimal psi functions in mind, we return to Figure 11 from Chapter 2. It shows Monte Carlo rejection rates versus sample size of the two tests T1 and T2 under asymmetric bivariate contamination Model 5 from Chapter 2. For both T1 and T2 in that figure, there is a peculiar large drop in power for the optimal psi-function for $\mu = 4$ and $\gamma = .06$ and all samples sizes when the efficiency increases through 90% to 99%. The effect is also evident for smaller values of gamma to various degrees for T1 and T2. By way of contrast a drop in power as efficiency approaches 99% is only very weakly evident for the bisquare psi-function in Figure 11.

As an explanation for this behavior we provide Figure 25 showing the scatter-plots of the differences between the LS and MM beta estimates at 99% versus 85% efficiency under the same Model 5 parameters as before, namely $\mu = 4, \gamma = 0.06$ and sample size n=100. The scatter plots for the optimal psi-function at 99% efficiency reveal a bivariate bimodal distribution that is consistent with the bi-modal histograms in Figure 24. For the optimal psi-function estimates in Figure 25 there are many samples for which the difference $\hat{\beta}_{LS} - \hat{\beta}_{MM}$ is large and statistically significant at 85% efficiency, but is nearly zero and non-significant at 99% efficiency. These samples constitute the majority of the left component of the corresponding bi-modal histogram in Figure 24. This behavior results in a loss of power for the optimal psi-function estimates at 99% efficiency. On the other hand, use of the bisquare psi-function based estimates does not result in much loss of power at 99% efficiency.

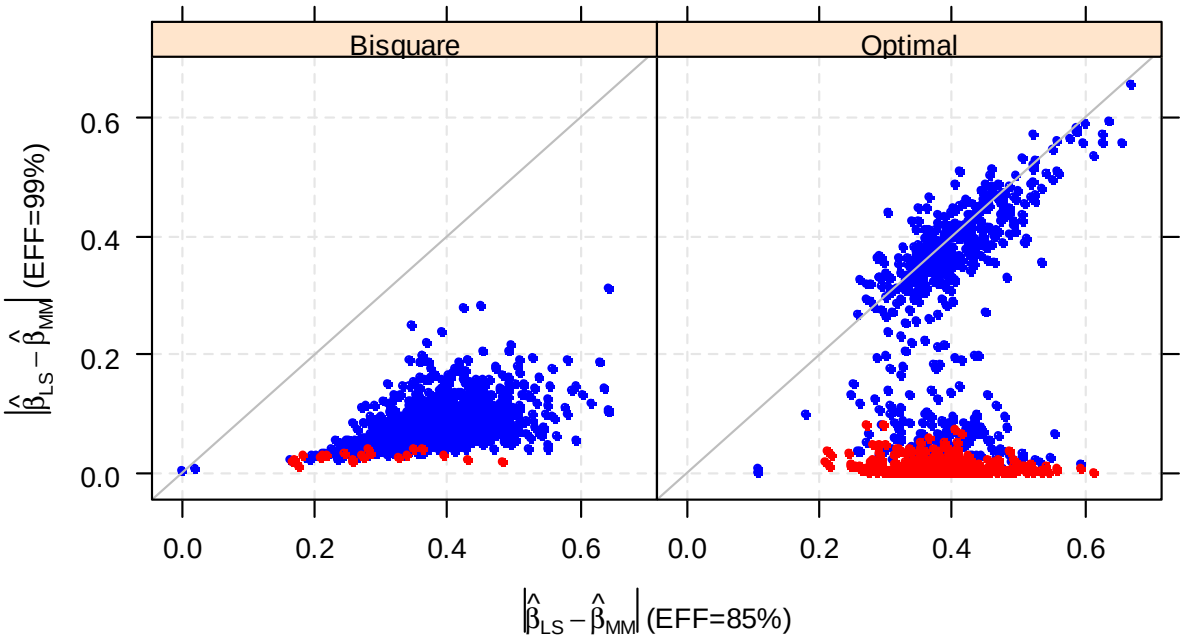

Figure 25. Scatter-plot of absolute differences between LS and MM beta estimates at 99% vs 85% normal distribution efficiency under Model 5 with $\gamma = 0.06$ and $\mu = 4$. Sample size n=100. Red points correspond to samples with significant T2 (EFF=85%), but non-significant T2 (EFF=99%).

3.4 T1 and T2 Tests using MM-Estimators with Low Normal Distribution Efficiency

The normal distribution efficiencies of the MM-estimators that were considered in Chapter 2, namely 85%, 90%, 95% and 99%, are perhaps the four values most commonly used in practice. We now examine the performance of the T1 and T2 tests for the MM-estimators with normal distribution efficiencies below 85%.

Figure 26 through Figure 30 display results of the Monte Carlo simulation study for the MM-estimators with efficiencies of 29%, 50%, 70% and 95% for the same five models as in Chapter 2. Normal distribution efficiency of 95% is included as a reference point. Normal distribution efficiency of 29% was selected because in this case the bisquare MM-estimator coincides with the highly robust initial S-estimator (see Section 1.4 in Chapter 1). For brevity we only focus on the bisquare loss function.

Model 1 (normal distribution errors). The simulation results under normal errors are shown in Figure 26 and reveal apparent problems with the finite sample levels of both T1 and T2 tests at low normal distribution efficiencies.

T1 Test Performance: The patterns in the left panel of Figure 26 are consistent with those observed in Chapter 2 (Figure 7, top left panel). The actual level of T1 is larger than the nominal significance level, more so the lower the normal distribution efficiency and the smaller the sample size. The T1 level at sample size 500 is fairly close to the large sample significance level of 0.05 for all four normal distribution efficiencies.

T2 Test Performance: A completely new pattern not seen in Chapter 2 emerges from the right panel of Figure 26. At normal distribution efficiencies between 85% and 99% the T2 level stayed quite on target even at the lowest sample size (Figure 7, top right panel). At low normal distribution efficiencies, however, we see T2 rejection rates larger than the nominal significance level, more so the lower the efficiency and the smaller the sample size. Note that the T2 level at 29% efficiency is still noticeably larger than 0.05 even at sample size 500.

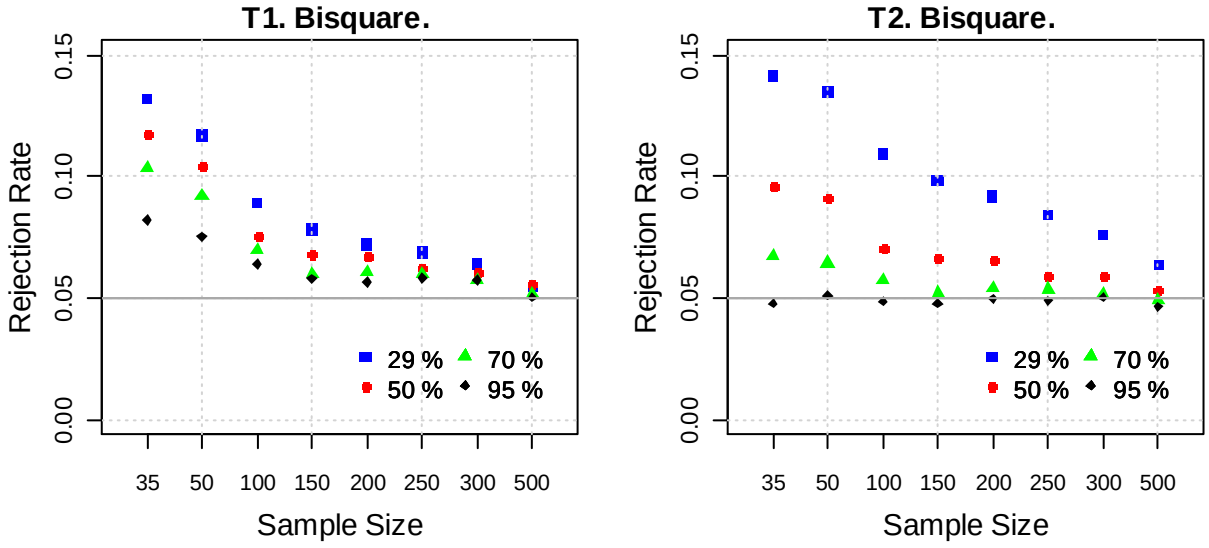

Figure 26. Model 1. Level of the T1(left) and T2(right) test statistics for the slope β in a simple linear regression under normal residuals. Grey horizontal line is at large-sample significance level of 0.05.

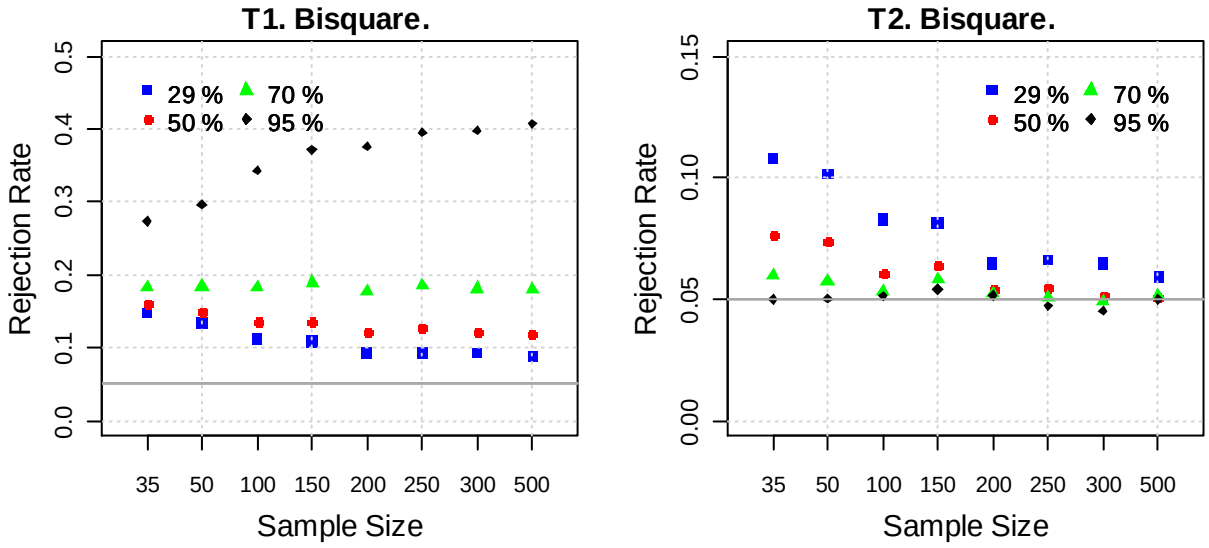

Figure 27. Model 2. Rejection rates of the T1 (left) and T2 (right) test statistics for the slope β in a simple linear regression under symmetric t_5 residuals. Grey horizontal line is at large-sample significance level of 0.05.

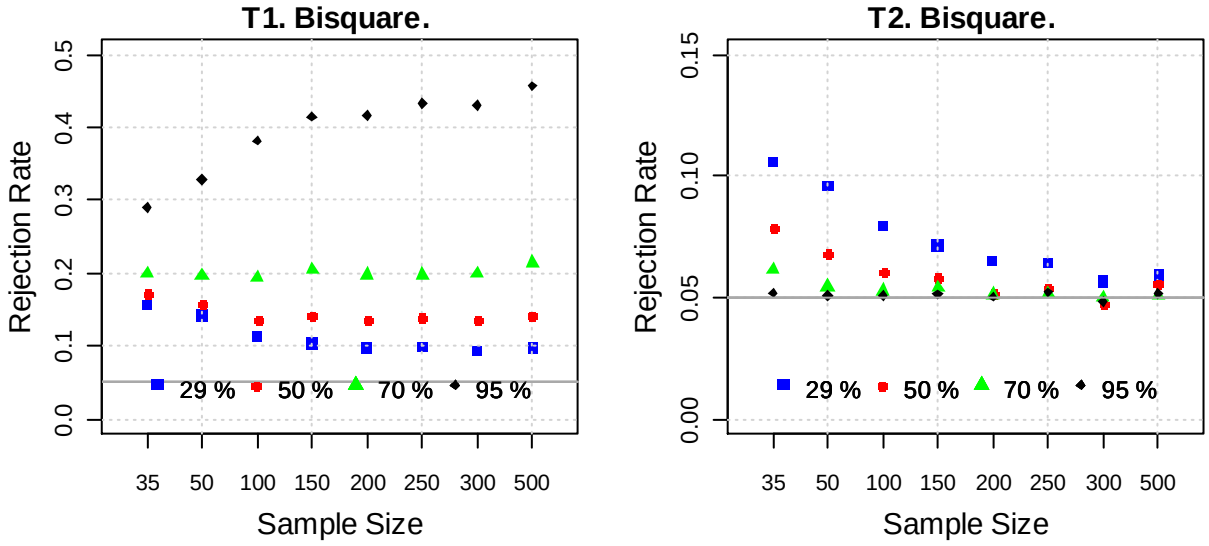

Figure 28. Model 3. Rejection rates of the T1 (left) and T2 (right) test statistics for the slope β in a simple linear regression under skewed t_5 residuals. Grey horizontal line is at large-sample significance level of 0.05.

Model 2 (symmetric t-distribution errors). The simulation results for a symmetric t-distribution with five degrees of freedom are displayed in Figure 27.

T1 Test Performance: The symmetric t-distribution with five degrees of freedom is in the alternative hypothesis K1 for T1. The left panel of Figure 27 shows that the lower the normal distribution efficiency the lower the rejection rate, a pattern that is consistent with that observed in Chapter 2 (Figure 8, bottom right). What is peculiar is a small decrease in the rejection rate as sample size grows for the 29% and 50% efficiencies.

T2 Test Performance: The symmetric t-distribution is in the null hypothesis H2 for T2. The results are very similar to those in Figure 26 for normal errors. The T2 level at low normal distribution efficiencies even though somewhat lower than that under normal errors can still be substantially above the nominal significance of 0.05. For example, consider efficiency of 29% and n=50, in which case the T2 rejection rates are 10% under t_5 errors and 13% under normal errors.

Model 3 (skewed t-distribution errors). The results for skewed t-distribution with five degrees of freedom are presented in Figure 28. They are very similar to those in Figure 27 and, thus, most of the above comments for the symmetric t-distribution apply here.

Model 4 (asymmetric normal mixture distribution errors).

Figure 29 displays Monte Carlo rejection rates versus sample size and normal distribution efficiency for mixing proportions 0.02, 0.04 and 0.06. The square, diamond and triangle symbols represent μ values of 4, 7 and 14 respectively.

T1 Test Performance: The normal mixture distribution is in the alternative K1 for the test T1 and so one is hoping for a large power. The patterns seen in the top row of Figure 29 are consistent with those observed in Chapter 2 (Figure 10, top) and most of the comments for Figure 10 apply here. The power of T1 at low normal distribution efficiencies such as 29%, 50% and 70% is consistently lower than that at 95% efficiency, sometimes substantially so.

T2 Test Performance: The normal mixture distribution is in the null hypothesis H2 for the test T2. The rejection rates for low efficiency MM estimators, especially 29% efficiency, are larger than nominal significance level of 0.05.

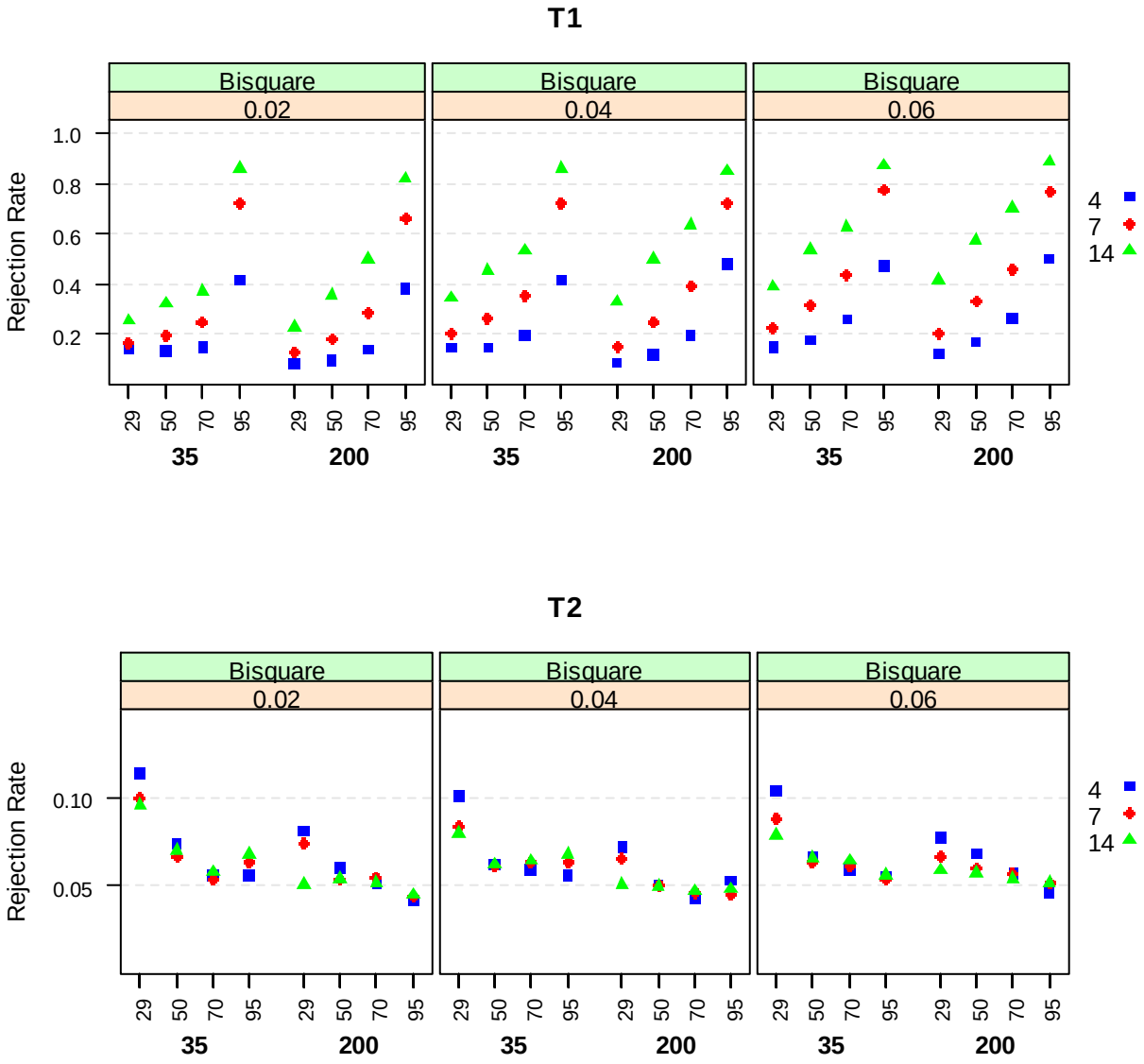

Figure 29. Model 4. Rejection rates of the T1 and T2 test statistics for slope β in a simple linear regression under asymmetric residual contamination. Blue squares correspond to mild outliers $N(\mu = 4, 0.25^2)$, red diamonds – to medium outliers $N(\mu = 7, 0.25^2)$ and green triangles – to gross outliers $N(\mu = 14, 0.25^2)$.

Model 5 (bivariate normal mixture distribution). The results for Model 5, which is in the alternative K2 for both T1 and T2, are displayed in Figure 30. Note low power for the MM estimators at low normal distribution efficiencies for both T1 and T2.

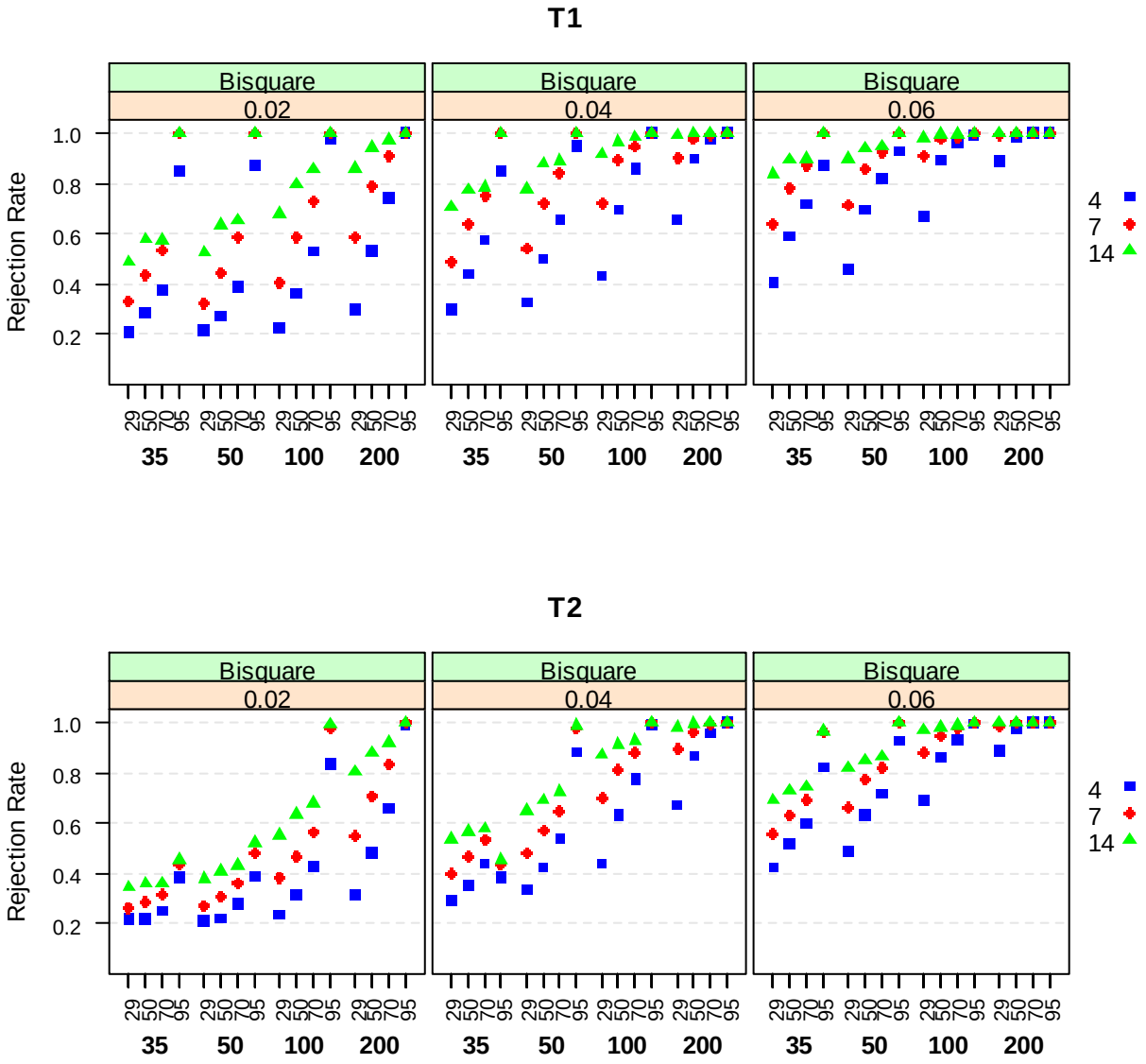

Figure 30. Model 5. Rejection rates of the T1 and T2 test statistics for slope β in a simple linear regression under bivariate asymmetric residual contamination. Blue squares correspond to mild outliers $N(\mu = 4, 0.25^2)$, red diamonds – to medium outliers $N(\mu = 7, 0.25^2)$ and green triangles – to gross outliers $N(\mu = 14, 0.25^2)$.

3.5 Yohai, Stahel and Zamar Test of Bias

Yohai et al. (1991) proposed a test to detect a larger overall bias in the final high efficiency MM-estimator $\widehat{\boldsymbol{\theta}}_{MM}$ relative to the more bias robust but much less efficient initial S-estimator $\widehat{\boldsymbol{\theta}}_S$ (Rousseeuw & Yohai, 1984)[15], which, according to the first authors, may happen under 'heavy contamination'. The "YSZ" test statistic, given in Section 3.5.1, is a studentized version of the difference between the scale M-estimate $\hat{\sigma}_2$ of the residuals obtained from $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}$ and the scale M-estimate $\hat{\sigma}_1$ of the residuals obtained from the initial S-estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}$. The difference between the two scale estimates converges to zero when both the initial and final estimators are consistent, which happens when the error distribution is symmetric, and the difference converges to a positive value when the asymptotic bias of $\widehat{\boldsymbol{\theta}}_{MM}$ is different from $\widehat{\boldsymbol{\theta}}_S$'s bias. The "YSZ" test statistic was shown by the first authors above to have an asymptotic χ_p^2 distribution under the null hypothesis of regression model (1.3) with a symmetric error distribution F_ϵ.

While the YSZ test was originally derived to test for a larger bias in the final MM- estimator relative to the initial S-estimator, it can be easily modified to test for a larger bias in the LS-estimator relative to the initial S-estimator by noting that $\psi_{LS}(r) = r$. We call the resulting test statistic T3.

Since the YSZ test T3 is a test of "overall" bias it cannot be used to test for bias in subsets of coefficients or in individual coefficients as is the case for our proposed Wald like tests T1 and T2 from Chapter 2. Simulation experiments in Section 3.5.2 reveal that T3 test for comparing LS with the low normal distribution efficiency initial S-estimator performs quite poorly with respect to achieving target test levels.

3.5.1 The YSZ Test Statistic

The bias test statistic as proposed by Yohai et al. (1991) is a studentized version of the difference between the scale of residuals obtained from the final MM-estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}$ and the one obtained from the initial S-estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}$:

[15] Asymptotic normal distribution efficiency of an S-estimator with 0.5 breakdown point is not larger than 33%. Numerical computation yields that the normal distribution efficiency of the S-estimator based on the bisquare psi-function with $c_0 = 1.548.$ and $b = 0.5$ is 0.29. See Maronna et al. (2006).

$$T_{YSZ} = n\frac{2(\hat{\sigma}_2 - \hat{\sigma}_1)}{\hat{v}_1 \widehat{\delta^2_{MM,S}}/\hat{\sigma}_1} \sim \chi^2_p \tag{3.1}$$

where

$\psi_1(z)$ and $\psi_2(z)$ are the initial and final estimate psi-functions respectively;

$\hat{\sigma}_1 = s_1(\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}})$ and $\hat{\sigma}_2 = s_1(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}})$ are the M-estimates of scale based on the initial and final coefficient estimates respectively;

$r_i^S = y_i - \boldsymbol{x_i^{*\prime}}\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}$ are the residuals based on the initial coefficient estimates;

$$\hat{v}_1 = \frac{ave_i\, \psi_1'(r_i^S/\hat{\sigma}_1)}{ave_i\{\psi_1(r_i^S/\hat{\sigma}_1) r_i^S/\hat{\sigma}_1\}}$$

$$\widehat{\delta^2_{MM,S}} = ave_i\left\{\left(\frac{\hat{\sigma}_1\psi_2(r_i^S/\hat{\sigma}_1)}{ave_j\psi_2'(r_j^S/\hat{\sigma}_1)} - \frac{\hat{\sigma}_1\psi_1(r_i^S/\hat{\sigma}_1)}{ave_j\psi_1'(r_j^S/\hat{\sigma}_1)}\right)^2\right\} \tag{3.2}$$

We use subscript (MM,S) to distinguish $\delta^2_{MM,S}$ from $\delta^2_{LS,MM}$ (2.8) in test T2.

Asymptotically, T_{YSZ} has χ^2_p distribution under the null hypothesis of model (1.3) with a symmetric error distribution F_ϵ. The derivation is given in Appendix A3.

Comments:

- Derivation of the null distribution of the test statistic relies heavily on using S-estimate as an initial estimate, more precisely on the fact that $\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}$ minimizes $s_1(\boldsymbol{\theta})$.

- Derivation of the null distribution of the test statistic relies on the fact that $\hat{\sigma}_2$ is an M estimate of scale of residuals obtained from $\widehat{\boldsymbol{\theta}}_{MM}$, i.e. $\hat{\sigma}_2 = s_1(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}})$. It is important to point out the subscript 0 in $s_1(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}})$.

- The test can be applied to detect overall bias in the LS estimates compared to the initial S-estimate. The resulting test statistic is

$$T_3 = n\frac{2(\hat{\sigma}_2 - \hat{\sigma}_1)}{\hat{v}_1\widehat{\delta^2_{LS,S}}/\hat{\sigma}_1} \sim \chi^2_p \tag{3.3}$$

where $\hat{v}_1$ and $\hat{\sigma}_1$ are as before, $\hat{\sigma}_2 = s_2(\widehat{\boldsymbol{\theta}_{\boldsymbol{LS}}})$ and

$$\widehat{\delta^2_{LS,S}} = ave_i\left\{\left(r_i^S - \frac{\hat{\sigma}_1\psi_1(r_i^S/\hat{\sigma}_1)}{ave_j\psi_1^{'}(r_j^S/\hat{\sigma}_1)}\right)^2\right\} \tag{3.4}$$

Note that $\widehat{\delta^2_{LS,S}}$ is the same as $\widehat{\delta^2_{MM,S}}$ with ψ_2 replaced by the "least-squares" psi- function $\psi_2(r) = \psi_{LS}(r) = r$. We will collectively call both T_{YSZ} (3.1) and T_3 (3.3) a T3 test.

- Alternative estimates of $\delta^2_{MM,S}$ and $\delta^2_{LS,S}$ can be considered as we discuss in Appendix A3.

- The YSZ test T3 and test T2 are based on similar asymptotic normality results. Instead of using the Wald type chi-square test statistic, however, Yohai et al. (1991) derive an approximation that does not require estimation of $\boldsymbol{V}_{\boldsymbol{x}^*}$. See Appendix A3 for more details.

3.5.2 Monte Carlo Simulations

In Chapter 2 we studied finite sample performance of the tests T1 and T2 in a simple linear regression under five data models. We now repeat the Monte Carlo simulation study for the new test T3. The T3 test of overall bias (3.1) in the final MM estimator as compared to the highly robust initial S-estimator is implemented in R package robust. In the simulation experiments for the final estimates we consider MM-estimates with bisquare and optimal loss functions at normal distribution efficiencies of 85%, 90%, 95% and 99%. The initial S-estimate is always based on the bisquare loss function ρ_1 with $c_1 = 1.548$ regardless of the choice of the loss function of the final estimate. This initial S-estimator has normal distribution efficiency of 29%.

We wrote additional code to implement T3 test for overall bias (3.3) in the LS estimator, which we label as 100% efficiency in Figure 31 through Figure 35 below.

Model 1 (normal distribution errors). Figure 31 displays Monte Carlo level of the test T3 versus sample size under normally distributed errors. Surprisingly, the T3 finite-sample level behavior is quite poor. At a given sample size there is a wide spread in rejection rates across different efficiencies for the bisquare loss function, but there is almost no such spread for the optimal loss function. The spread decreases as sample size grows and finite-sample levels converge to the large sample significance level of 0.05. Rejection rates exhibit a peculiar trend, which is similar for the two loss functions: the actual level is considerably below the nominal significance level of 0.05 at sample size 35, increases to values somewhat above 0.05 at sample size 200 and then decreases, getting closer to 0.05 for sample size 500, the largest sample size considered in our simulations.

Model 2 (symmetric t-distribution errors). Results for a symmetric t-distribution with five degrees of freedom are displayed in Figure 32. The symmetric t-distribution is in the null hypothesis for T3. Similarly to Figure 31 (Model 1), there is a wide spread in rejection rates across efficiencies for the bisquare loss function, but not for the optimal loss function. Rejection rates are much lower than the nominal significance level of 0.05 at sample size 35, then gradually increase with sample size till they get closer to 0.05 in large samples (n=150 to 300, depending on loss function and efficiency).

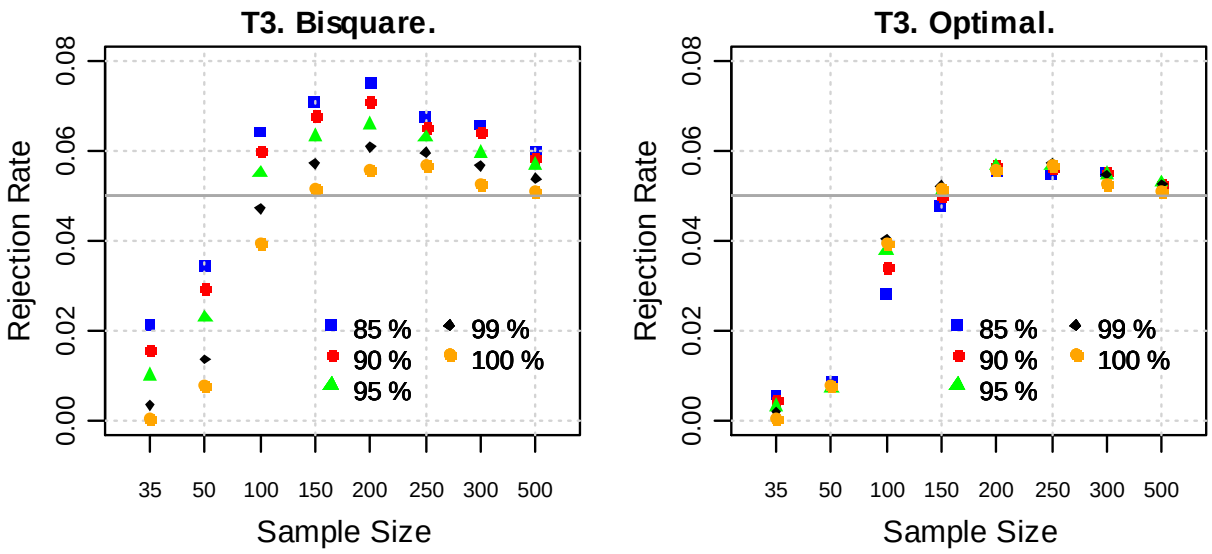

Figure 31. Model 1. Level of the YSZ T3 test for the overall bias in a simple linear regression under normal residuals. Orange points correspond to the extension of the YSZ test to LS estimates. Grey horizontal line marks nominal significance level of 0.05.

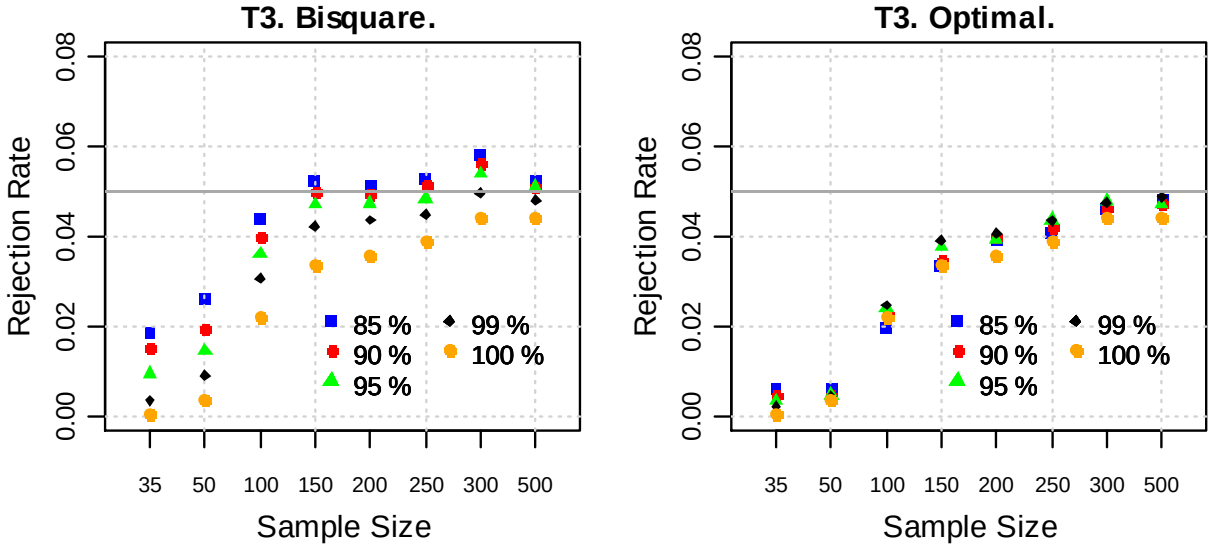

Figure 32. Model 2. Level of the YSZ T3 test for the overall bias in a simple linear regression under symmetric t_5 residuals. Orange points correspond to the extension of the YSZ test to LS estimates. Grey horizontal line marks nominal significance level of 0.05.

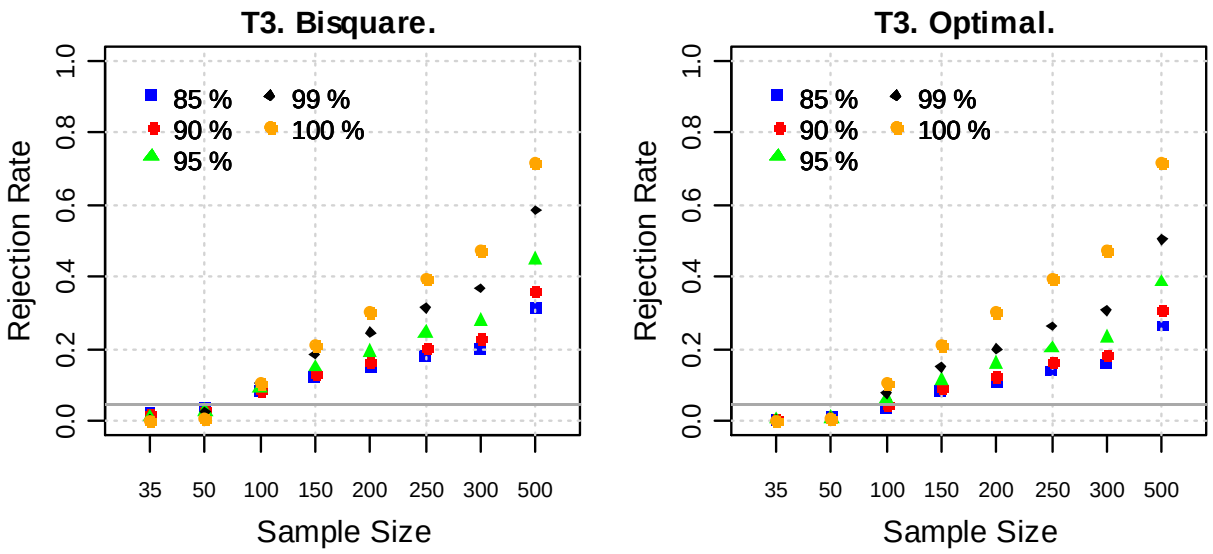

Figure 33. Model 3. Rejection rates of the YSZ T3 test for the overall bias in a simple linear regression under skewed t_5 residuals. Orange points correspond to the extension of the YSZ test to LS estimates. Grey horizontal line marks nominal significance level of 0.05.

Model 3 (skewed t-distribution errors). Rejection rates for skewed t-distribution with five degrees of freedom are presented in Figure 33. This is an asymmetric distribution and is in the alternative hypothesis for T3. For a given sample size the T3 rejection rate increases with the normal distribution efficiency of the MM estimator so that LS estimator has the largest rejection rate. The T3 power increases with sample size.

It should be noted that under Model 3 all three estimators, namely MM, S and LS, are consistent for the slope β and, thus, rejection in T3 is expected to be driven by the differences in the intercept biases. In this case the LS intercept is known to be a consistent estimator of $\alpha_0 = E(\epsilon)$. The S intercept, on the other hand, is a consistent estimator of a solution to (1.15) which, in general, has no intuitive interpretation as it coincides with neither a location parameter of a skewed-t density nor a certain quantile.

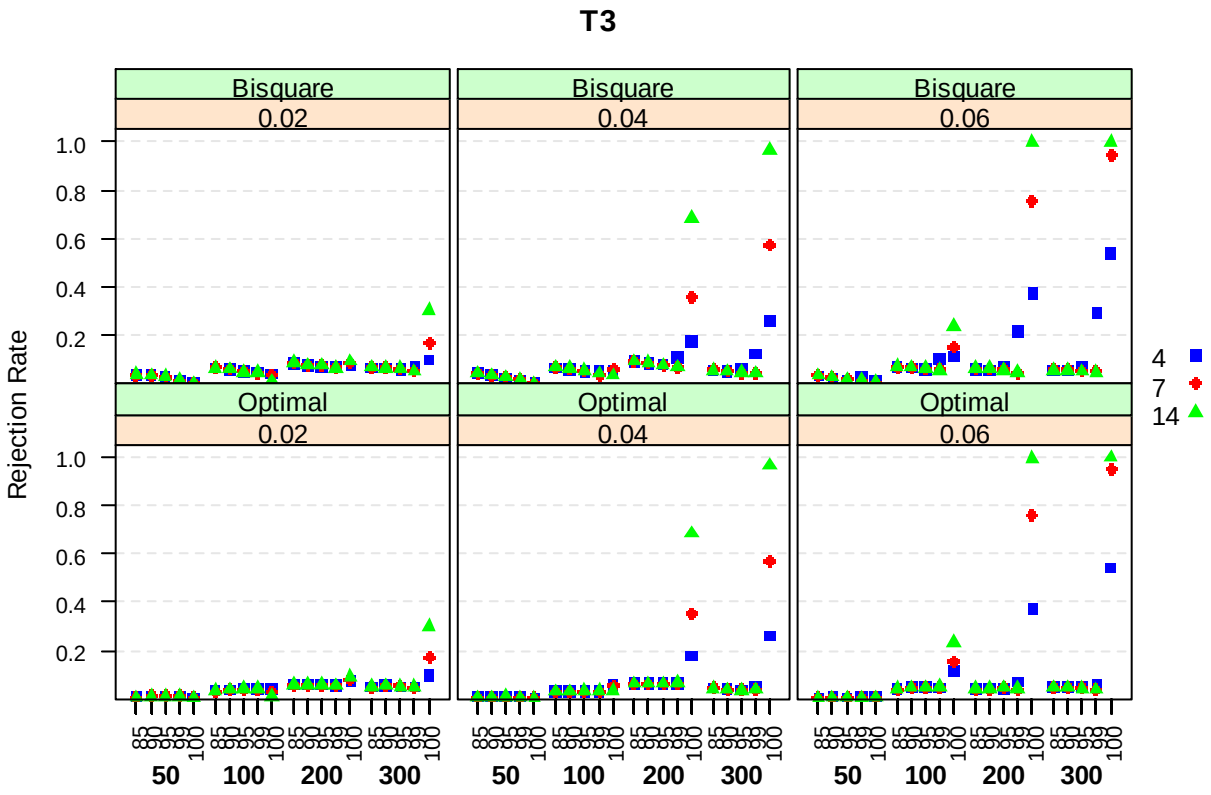

Figure 34. Model 4. Rejection rate for the YSZ T3 test for the overall bias in a simple linear regression under asymmetric residual contamination.

Model 4 (asymmetric normal mixture distribution errors). Figure 34 displays Monte Carlo rejection rates of T3 versus sample size and normal distribution efficiency for mixing proportions (γ) 0.02, 0.04 and 0.06. The square, diamond and triangle symbols represent μ values of 4, 7 and 14 respectively. The error distribution is asymmetric and, therefore, this model is in the alternative hypothesis for T3. Under Model 4 all three estimators, namely MM, S and LS, are consistent for the slope β and, thus, rejection in T3 is expected to be driven by the differences in the intercept biases.

Rejection rates for the MM estimators with the optimal loss function stay low even at large sample sizes for all four normal distribution efficiencies. This is not surprising as in this case the asymptotic values of the MM and S intercept estimators are virtually the same, and approximately 0 (the mean of the main mixture component). Similarly, rejection rates are low for the MM estimators with the bisquare loss

function. The exception is 99% efficiency, in which case the rejection rates are somewhat larger than the nominal significance level 0.05 for $\mu = 4$, $\gamma = 0.04$ and 0.06 and sample sizes 200 and 300. In these cases the asymptotic values of the bisquare intercept are 0.027 for $\gamma = 0.04$ and 0.05 for $\gamma = 0.06$ and their differences with 0 (the asymptotic value of an S-estimator) become detectable with sample sizes 200 and larger.

The asymptotic value of the LS intercept is $\beta_{0,LS} = E\epsilon = (1-\gamma)0 + \gamma\mu = \gamma\mu$, which is, for example, 0.16 and 0.24 for $\mu = 4$, $\gamma = 0.04$ and $\gamma = 0.06$ respectively. Since $\beta_{0,LS} > \beta_{0,S}$ we expect T3 for LS to have power. The difference $\beta_{0,LS} - \beta_{0,S}$ increases with μ and γ and so does the rejection rate for LS in Figure 34. The T3 rejection rate, however, is above the nominal significance level of 0.05 only at $n = 300$ when $\gamma = 0.02$, at $n = 200,\ 300$ when $\gamma = 0.04$ and at $n = 100, 200, 300$ when $\gamma = 0.06$.

Model 5 (bivariate normal mixture distribution). The results for Model 5, which is in the alternative hypothesis for T3, are displayed in Figure 35. The contamination component of Model 5 is such that large positive residual 'outliers' occur for large values of x_i, and result in biased LS and robust estimates of β, with the bias of the latter being much smaller. In fact, in most cases the asymptotic values of the MM and S estimators are indistinguishably close to $\beta_0 = 0$ and $\beta = 1$ (see Table 6) and, therefore, we see rejection rates near the nominal significance level of 0.05 even at sample size $n = 300$. The exception is $\mu = 4$ and bisquare psi-function at 95% and 99% efficiency. Since now both intercept and slope estimators are biased it may not be surprising to see higher power in Figure 35 as compared to Figure 34 (Model 4) for LS and MM bisquare at 95% and 99% efficiency.

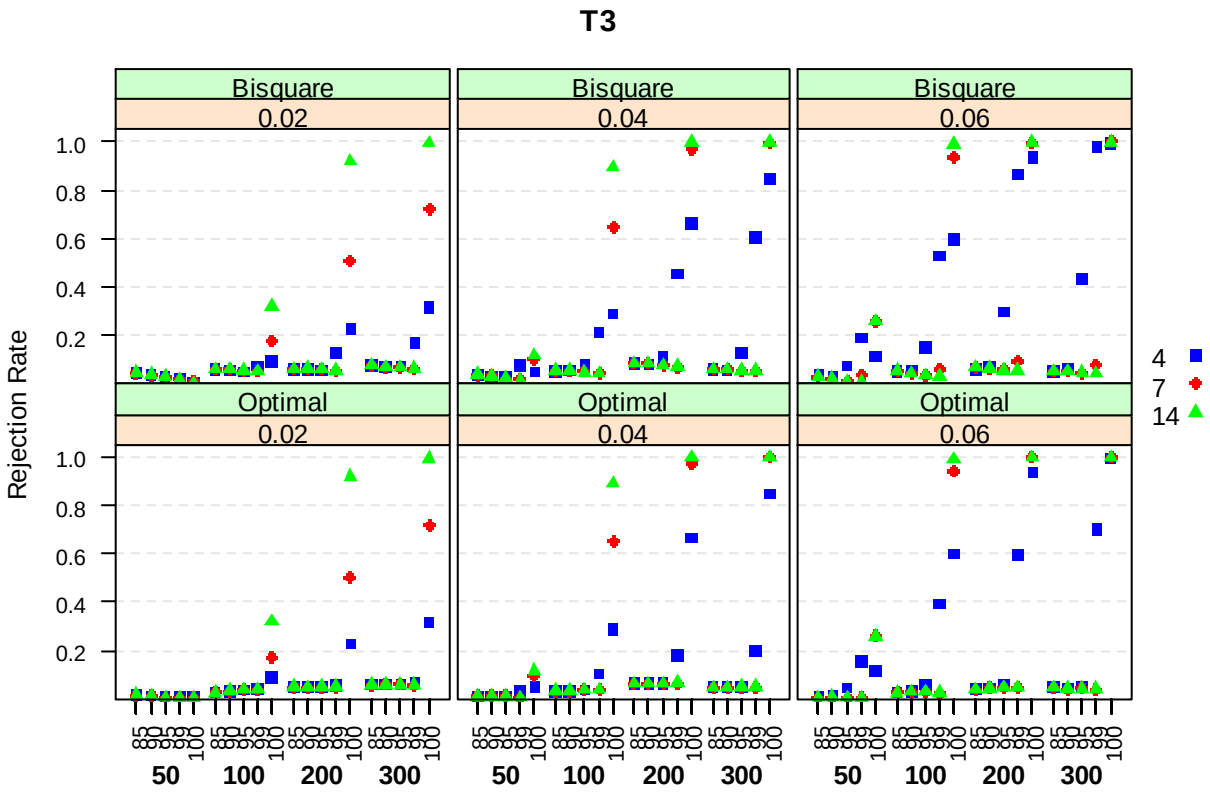

Figure 35. Model 5. Rejection rate for the YSZ T3 test for the overall bias in a simple linear regression under bivariate asymmetric residual contamination.

Table 6. Average of the intercept and slope estimates for sample size n=300 in the simulations for Model 5 (bivariate normal mixture distribution).

γ	μ	Efficiency at Normal Distribution	S Bisquare 29%	MM Bisquare 95%	MM Optimal 95%	LS 100%
0.02	4	Intercept	0.003	0.013	0.001	0.075
		Slope	1.00	1.02	1.00	1.15
	7	Intercept	0.003	0.001	0.001	0.130
		Slope	1.00	1.00	1.00	1.26
	14	Intercept	0.003	0.001	0.001	0.260
		Slope	1.00	1.00	1.00	1.52
0.04	4	Intercept	-0.002	0.032	0.001	0.138
		Slope	1.00	1.06	1.00	1.28
	7	Intercept	-0.002	0.000	0.000	0.241
		Slope	1.00	1.00	1.00	1.48
	14	Intercept	-0.002	0.000	0.000	0.482
		Slope	1.00	1.00	1.00	1.97
0.06	4	Intercept	-0.001	0.070	0.003	0.195
		Slope	1.00	1.14	1.00	1.39
	7	Intercept	-0.001	0.002	0.002	0.341
		Slope	1.00	1.00	1.00	1.68
	14	Intercept	-0.001	0.002	0.002	0.680
		Slope	1.00	1.00	1.00	2.36

3.6 Summary and Discussion

The T1 and T2 test statistics are standardized differences of the LS and robust MM estimates. T1 uses difference of large sample covariance matrices under normal distribution, while T2 uses correct asymptotic variance of the difference $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$ under a broader null hypothesis consisting of normal and non-normal error distributions. Under normality, the two variances coincide asymptotically. Nonetheless, the corresponding standard errors exhibit a rather different behavior in finite samples.

T1 standard errors ignore the data-based dependency between the LS and MM estimates. This results in a non-normal null distribution of the T1 test statistic in small samples. It is of interest and a topic of future

research to find an exact (or a good approximate) finite-sample null distribution of the T1 test statistic that would ensure proper levels of T1 test in small samples.

T2 standard errors, on the other hand, incorporate most of the specific finite-sample dependencies between the LS and MM estimates. This results in a close to normal null distribution of the T2 test statistic in small samples. Through simulations, we showed that the standard normal distribution approximation for T2 test statistic works very well for a bisquare loss function with normal distribution efficiency between 85% and 99%, even in small samples. However, the standard normal approximation is no longer appropriate at low normal distribution efficiencies of 50% or below in small to moderate sample sizes as in this case it takes longer for the asymptotic theory to 'kick in'. The standard normal approximation also does not work very well for a fast re-descending optimal loss function. Therefore, we do not recommend using the latter for testing with T2, at least until a proper correction is discovered. Derivation of such a correction for T2 is still an open question, complicated by the fact that it depends not only on a shape of the loss function and a tuning constant value, but also on a distribution of the regression errors - be it normal or a certain non-normal distribution from T2 null hypothesis.

Our simulation study revealed poor level behavior of the dispersion test T3, both of the original test proposed by Yohai et al. (1991) and of its extension for a LS. T3 is designed to detect larger overall bias in LS or final MM estimator as compared to highly robust, but low efficiency initial S-estimator. As such, T3 may not be very useful since it does not allow testing for a subset of the coefficients.

Appendix A3: Distribution of the Differences between LS and S Regression Estimators

In this section we derive the asymptotic distribution of the differences between the final MM and initial S regression estimators as well as between the LS and S regression estimators under distribution models (1.3), which include both symmetric and skewed errors. We then present the derivation of the "YSZ" test (3.1) of overall bias in the final MM regression estimator as it appears in Yohai et al. (1991) and show how to extend it to the case of the least squares.

Asymptotic Distribution of the Differences between MM and Initial S Regression Estimators

Consistent with the definitions and notation from Chapter 1 let $\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}$ denote the initial S and final MM estimates in linear regression (1.1) with asymptotic values $\boldsymbol{\theta}_{\boldsymbol{S}}$ and $\boldsymbol{\theta}_{\boldsymbol{MM}}$ respectively. From Appendix A2 in Chapter 2 (eq. (A2.8), (A2.9) and (A2.5)) it follows that under model (1.3) and rather general conditions

$$\sqrt{n}\left(\left\{\widehat{\boldsymbol{\theta}}_{MM}-\widehat{\boldsymbol{\theta}}_{S}\right\}-\left\{\boldsymbol{\theta}_{MM}-\boldsymbol{\theta}_{S}\right\}\right)\rightarrow N\left(\mathbf{0},\delta_{MM,S}^{2}\boldsymbol{V}_{\boldsymbol{x}^{*}}^{-\mathbf{1}}+\omega_{MM,S}\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right) \tag{A3.1}$$

where

$$\delta_{MM,S}^{2}=E\left\{\left(\frac{\sigma\psi_{2}\left(\frac{u_{MM}}{\sigma}\right)}{E\psi_{2}'\left(\frac{u_{MM}}{\sigma}\right)}-\frac{\sigma\psi_{1}\left(\frac{u_{S}}{\sigma}\right)}{E\psi_{1}'\left(\frac{u_{S}}{\sigma}\right)}\right)^{2}\right\} \tag{A3.2}$$

with u_S and u_{MM} defined in (A2.15) and (1.22) and

$$\omega_{MM,S}=\sigma^{2}\left\{\left(\frac{e_{MM}}{a_{MM}}-\frac{e_{S}}{a_{S}}\right)^{2}\frac{E\left(\rho_{1}^{2}\left(\frac{u_{S}}{\sigma}\right)-b^{2}\right)}{d_{s}^{2}}+\frac{2}{d_{s}}\left(\frac{e_{MM}}{a_{MM}}-\frac{e_{S}}{a_{S}}\right)\left(\frac{g_{S}}{a_{S}}-\frac{g_{MM}}{a_{MM}}\right)\right\} \tag{A3.3}$$

Derivation of the YSZ Test Statistic

When F_ϵ is symmetric around α_0 then both estimators are consistent estimators of $\boldsymbol{\theta_0}=(\alpha_0,\boldsymbol{\beta_0})$ so that $u_S=u_{MM}=u=\epsilon-\alpha_0$. Moreover, $\omega_{MM,S}$ is zero and (A3.1) reduces to the asymptotic result at the beginning of Section 4 of Yohai et al. (1991). Thus, under symmetric F_ϵ

$$\frac{n}{\delta_{MM,S}^{2}}\left(\widehat{\boldsymbol{\theta}}_{MM}-\widehat{\boldsymbol{\theta}}_{S}\right)'\boldsymbol{V}_{\boldsymbol{x}^{*}}\left(\widehat{\boldsymbol{\theta}}_{MM}-\widehat{\boldsymbol{\theta}}_{S}\right)\rightarrow\chi_{p}^{2} \tag{A3.4}$$

The derivation steps below are as they are given in Yohai et al. (1991), Section 4.

Let $s_1(\boldsymbol{\theta})$ be the solution of $ave_i\rho_1\left(\frac{y_i-\boldsymbol{\theta}'x_i^*}{s}\right)=b=0.5$[16]. Because by definition of a regression S-estimate $\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}$ minimizes $s_1(\boldsymbol{\theta})$, a standard Taylor series expansion yields

$$s_1\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}\right)-s_1\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}\right)\sim\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}_{MM}-\widehat{\boldsymbol{\theta}}_S\right)'s_1''\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}\right)\left(\widehat{\boldsymbol{\theta}}_{MM}-\widehat{\boldsymbol{\theta}}_S\right)$$

where $s_1''(\boldsymbol{\theta})=\partial^2 s_1(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}$. The derivative of $s_1(\boldsymbol{\theta})$ can be shown to be

$$s_1'(\boldsymbol{\theta})=-\frac{ave_i\left\{\psi_1\left(\frac{r_i(\boldsymbol{\theta})}{\sigma}\right)\boldsymbol{x_i^*}\right\}}{ave_i\left\{\psi_1\left(\frac{r_i(\boldsymbol{\theta})}{\sigma}\right)\frac{r_i(\boldsymbol{\theta})}{\sigma}\right\}}$$

where $r_i(\boldsymbol{\theta})=y_i-\boldsymbol{x_i^{*\prime}\theta}$. Differentiating this expression with respect to $\boldsymbol{\theta}$, evaluating at $\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}$ and replacing averages with expectations leads to $s_1''\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}\right)\sim v_1\boldsymbol{V}_{x^*}/\sigma$, where

$$v_1=\frac{E\psi_1'\left(\frac{r}{\sigma}\right)}{E\left\{\psi_1\left(\frac{r}{\sigma}\right)\frac{r}{\sigma}\right\}}$$

Therefore,

$$n\left\{s_1\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MM}}\right)-s_1\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{S}}\right)\right\}\sim\frac{n}{2}\frac{v_1}{\sigma}\left(\widehat{\boldsymbol{\theta}}_{MM}-\widehat{\boldsymbol{\theta}}_S\right)'\boldsymbol{V}_{x^*}\left(\widehat{\boldsymbol{\theta}}_{MM}-\widehat{\boldsymbol{\theta}}_S\right)$$

and asymptotic distribution of the test statistic T3 (3.1) results from (A3.4), because “estimation of the constants involved will not affect the asymptotic distribution”.

Asymptotic Distribution of the Differences between LS and Robust S Regression Estimators

Following the derivation steps in Appendix A2 in Chapter 2 and using (A2.8), (A2.9) and (A2.5) it is straightforward to show that under model (1.3) and rather general conditions

[16] This is to follow Yohai et al. (1991). In R packages robust and robustbase, $ave_i\{\}$ is replaced with $\frac{n}{n-p}ave_i\{\}$ and that is how S estimate was defined in Chapter 1. It is also assumed that loss functions are normalized to have maximum value of one, i.e. they conform to Definition 4.

$$\sqrt{n}\left(\{\widehat{\boldsymbol{\theta}}_{LS}-\widehat{\boldsymbol{\theta}}_{S}\}-\{\boldsymbol{\theta}_{LS}-\boldsymbol{\theta}_{S}\}\right)\rightarrow N\left(\mathbf{0},\delta_{LS,S}^{2}\boldsymbol{V}_{\boldsymbol{x}^{*}}^{-\mathbf{1}}+\omega_{LS,S}\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right) \tag{A3.5}$$

where

$$\delta_{LS,S}^{2}=E\left\{\left(u_{LS}-\frac{\sigma\psi_{1}\left(\frac{u_{S}}{\sigma}\right)}{E\psi_{1}'\left(\frac{u_{S}}{\sigma}\right)}\right)^{2}\right\} \tag{A3.6}$$

with u_S and u_{LS} defined in (A2.15) and (1.22) and

$$\omega_{LS,S}=\sigma^{2}\frac{e_{S}}{a_{S}d_{s}}\left(2\frac{g_{LS}}{\sigma}-2\frac{g_{S}}{a_{S}}+\frac{E\left(\rho_{1}^{2}\left(\frac{u_{S}}{\sigma}\right)-b^{2}\right)e_{S}}{a_{S}d_{s}}\right) \tag{A3.7}$$

with the constants defined in (A2.7).

When F_ϵ is symmetric around α_0 then both estimators are consistent estimators of $\boldsymbol{\theta_0}=(\alpha_0,\boldsymbol{\beta_0})$, $u_S=u_{LS}=u=\epsilon-\alpha_0$, $\omega_{LS,S}$ is zero and, thus,

$$\frac{n}{\delta_{LS,S}^{2}}\left(\widehat{\boldsymbol{\theta}}_{LS}-\widehat{\boldsymbol{\theta}}_{S}\right)'\boldsymbol{V}_{\boldsymbol{x}^{*}}\left(\widehat{\boldsymbol{\theta}}_{LS}-\widehat{\boldsymbol{\theta}}_{S}\right)\rightarrow\chi_{p}^{2} \tag{A3.8}$$

One can now repeat the derivation steps from the previous subsection to get (3.3).

Estimation of δ^2

There are several estimators of $\delta_{MM,S}^2$ and $\delta_{LS,S}^2$ that are consistent under model (1.3) with symmetric F_ϵ.

The estimates (3.2) and (3.4) follow Yohai et al. (1991) and use only residuals from the initial S estimate.

One may as well use the following estimates:

$$\widehat{\delta_{MM,S}^{2}}=ave_{i}\left\{\left(\frac{\hat{\sigma}_{1}\psi_{2}\left(r_{i}^{MM}/\hat{\sigma}_{1}\right)}{ave_{j}\psi_{2}'\left(r_{j}^{MM}/\hat{\sigma}_{1}\right)}-\frac{\hat{\sigma}_{1}\psi_{1}\left(r_{i}^{S}/\hat{\sigma}_{1}\right)}{ave_{j}\psi_{1}'\left(r_{j}^{S}/\hat{\sigma}_{1}\right)}\right)^{2}\right\} \tag{A3.9}$$

and

$$\widehat{\delta_{LS,S}^2} = ave_i\left\{\left(r_i^{LS} - \frac{\hat{\sigma}_1\psi_1\left(r_i^S/\hat{\sigma}_1\right)}{ave_j\psi_1^{'}\left(r_j^S/\hat{\sigma}_1\right)}\right)^2\right\} \quad \text{(A3.10)}$$

Note that (A3.10) uses both least-squares r_i^{LS} and S-estimate residuals r_i^S in contrast to (3.4) that uses only S-estimate residuals r_i^S. This way $\widehat{\delta_{LS,S}^2}$ in (A3.10) is a consistent estimator of $\delta_{LS,S}^2$ under asymmetric error distributions. When error distribution is symmetric then $u_{LS} = u_S$ and, thus, $\widehat{\delta_{LS,S}^2}$ in (3.4) is a consistent estimator of $\delta_{LS,S}$ under the null hypothesis of the YSZ test T3.

4. MM Regression Estimator under Skewed Fat-Tailed Error Distributions

4.1 Introduction

Consider a regression model (1.1) where the observed data $\boldsymbol{z_i'} = (y_i, \boldsymbol{x}_i'),\ i = 1, \cdots, n$, consists of independent and identically distributed random variables with a joint distribution (1.3), which assumes that $\boldsymbol{x_i}$ and $\boldsymbol{\epsilon_i}$ are independent for each i. A practitioner who is familiar with the robust MM regression described in Chapter 1 knows that when the error distribution is symmetric, the tuning constants can be chosen so that MM estimates: (1) are consistent, (2) have a user-specified high normal distribution efficiency, and (3) have a high efficiency relative to least squares when the errors are fat-tailed. What is not so well known and not very well understood is the behavior of the MM regression estimates when error distribution is skewed as well as fat-tailed.

It is not uncommon to see skewed error distributions in applications and, unfortunately, there are situations where data transformations to achieve symmetry make little sense or are not possible. Moreover, for the ease of interpretation, a data analyst may still prefer to work with the data on the original scale. In finance, for example, it is often of interest to model asset returns as a linear function of various macro-economic factors, market indices and other risk factors. Such regression models are called "factor models" in finance, and the intercept of such models is of great interest as it represents possible excess return referred to in the literature as "alpha". Asset returns, however, are known to exhibit skewness as well as kurtosis. Continuously compounded returns (a.k.a. log-returns or geometric returns) can be viewed as a symmetrizing log transformation of the discrete (a.k.a. arithmetic) returns, but both skewness and kurtosis still persist. In this context, a practitioner would want to know whether robust MM regression would work well for the intercept estimates as well as slopes.

We know that MM slopes estimates are consistent even when the error distribution is skewed (see comments to this effect in Chapter 1 and mathematical details in Appendix A2). On the other hand, it is an open question whether or not MM slope estimates have high absolute efficiency and high efficiency relative to least squares for skewed and fat-tailed non-normal distributions. In this chapter, we show that the answer to this question is in the affirmative.

When the error distribution is skewed, it turns out that the MM intercept estimator is biased and does not fully absorb a non-zero mean of the errors, unlike least squares where the intercept absorbs any non-zero mean component of a skewed error distribution. On one hand, a biased MM intercept estimator results in a biased conditional-mean predictor. On the other hand, an estimator with a small bias may have smaller mean squared error (MSE) than the unbiased and would therefore be preferred to the unbiased estimator. In this chapter, we examine the extent to which the MM intercept bias is significant for skewed fat-tailed error distributions and whether or not there is a simple method of removing the bias.

Among the existing proposals for the robust estimation in a linear regression with asymmetric errors, we want to mention the work of Bianco, Ben, and Yohai (2005) and Marazzi and Yohai (2004). Following the robustness literature, both papers focus on obtaining intercept and slope estimates that have high absolute efficiency and are consistent under the main model (1.3) $F_0(\boldsymbol{x}, y) = G(\boldsymbol{x})F_\epsilon(y - \boldsymbol{x}'\boldsymbol{\beta_0})$ with asymmetric F_ϵ , yet are robust in the presence of contamination under a general mixture model (1.4) $F(\boldsymbol{x}, y) = (1-\gamma)F_0(\boldsymbol{x}, y) + \gamma H(\boldsymbol{x}, y)$. Bianco et al. (2005) proposed a natural extension of the MM-estimates of Yohai (1987) to the case where error distribution F_ϵ comes from an exponential family. Their main idea is to replace the ordinary residuals with the deviance residuals. Such MM-estimates are consistent; also, they can simultaneously obtain high user specified asymptotic efficiency and have a maximum breakdown point of one half. This method, unfortunately, is not suitable for the skewed t-distributions that we are focused on as they do not belong to an exponential family. Marazzi and Yohai (2004) considered a simpler location-scale model for the error distributions and introduced a truncated maximum likelihood estimator, or a TML-estimator. The TML-estimator consists of three steps: first, low efficiency and high breakdown point initial S estimates are computed and are corrected for bias due to skewness; then, observations that are unlikely under the estimated model are rejected; finally, the maximum likelihood estimate is computed with the retained observations. The rejection rule in the second step is based on a cut-off parameter that can be tuned to attain a desired efficiency while maintaining the breakdown point of the initial estimator. The TML-estimator can be applied to the skewed t-distributions with the fixed skewness and degrees of freedom. In practice, the skewness and degrees of freedom are rarely known a priori and, in fact, these are the two parameters of the skewed t-distributions that are so troublesome to estimate. See the following Section 4.3.2 and references therein for a discussion of the

problems with maximum likelihood estimation (MLE) of all the parameters of a skewed-t regression. The task of finding an estimator that is highly efficient and consistent under skewed fat-tailed distributions with unknown skewness and degrees of freedom and yet robust to bias-inducing contamination of type (1.4) is very difficult. In this chapter, we investigate when the loss in efficiency of the MM slopes and the bias of the MM intercept become large enough to warrant concerns about the use of the MM estimator.

The rest of the chapter is organized as follows. Section 4.2 briefly describes one particular family of fat-tailed skewed t-distributions due to Azzalini and Capitanio (2003) that seems to be gaining popularity in practice. Section 4.3 is focused on the slope estimates. It explores the asymptotic and finite-sample efficiency of the MM and LS slope estimators under skewed-t distributed errors for a wide range of the degrees of freedom and skewness parameter values. For conciseness, this chapter only considers a bisquare loss function. Intercept estimators are studied in Section 4.4, where we focus on skewed-t distributed errors and compare mean-squared-error (MSE) efficiencies of the MM and LS intercept estimators as well as investigate the percent contribution of the MM intercept bias to its MSE. In the same section, we also propose a simple bias-corrected MM ("c-MM") estimator to reduce MM intercept bias and derive its standard errors. The performance of the c-MM intercept estimator is evaluated under skewed-t errors in the same manner as that of the MM and LS intercepts. Section 4.5 takes a closer look at the simulation results. Section 4.6 discusses results of an additional simulation study under contamination mixture Models 4 and 5 of Chapter 2. It investigates efficiency and bias of skewed-t and symmetric-t maximum likelihood estimators relative to MM in a situation where the tail structure differs from the underlying likelihood function. Section 4.7 contains empirical examples. Summary and discussion in Section 4.8 conclude the chapter. Technical details are deferred to Appendix A4.

4.2 Skewed-t Distribution

There exist numerous approaches in the literature for generalizing a symmetric t-distribution to model skewness. See, for example, Hansen (1994), Theodossiou (1998), Fernandez and Steel (1998, 1999), Branco and Dey (2001), Azzalini and Capitanio (2003), Sahu, Dey, and Branco (2003), Arellano-Valle and Genton (2010), Jones and Faddy (2003), Aas and Haff, (2006). Extensive review of the similarities and

differences between the various skewed fat-tailed distribution models is beyond the scope of our work. A good discussion of a few families of the skewed t-distributions can be found in Aas and Haff (2006). See also the review paper by Azzalini (2005) and a book edited by Genton (2004).

In our work we decided to focus on the skewed-t distribution treated by Azzalini and Capitanio (2003). This distribution has a relatively simple form with an intuitive stochastic representation, and, furthermore, methods for fitting the distribution, computing quantiles and probabilities and sampling from it are available in the R package sn (Azzalini, 2011). The following definitions and results may be found in the above reference.

Definition 6: A random variable Y follows a univariate skewed-t distribution with location parameter ξ, scale parameter ω, skewness parameter λ and degrees of freedom ν if its probability density is

$$2t_1(y|\xi,\omega,\nu)T_1\left(\lambda\frac{y-\xi}{\omega}\left(\frac{\nu+1}{\left(\frac{y-\xi}{\omega}\right)^2+\nu}\right)^{\frac{1}{2}};\ \nu+1\right) \tag{4.1}$$

where $T_1(.;\nu)$ and $t_1(.;\nu)$ are the cumulative distribution and probability density functions, respectively, of a univariate *t* random variable with ν degrees of freedom, i.e.

$$t_1(y|\xi,\omega,\nu)=\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}\omega}\left(1+\frac{1}{\nu}\left(\frac{y-\xi}{\omega}\right)^2\right)^{-(\nu+1)/2}$$

where $\Gamma(.)$ is the gamma function $\Gamma(z)=\int_0^\infty e^{-t}t^{z-1}dt$.

We use $Y\sim ST(\xi,\omega^2,\lambda,\nu)$ to denote this distribution. One interesting property of this distribution is that $Pr(|Y-\xi|>a)$ is the same regardless of the value of the skewness parameter λ.

Skewness is sometimes described by the alternative parameter

$$\delta=\frac{\lambda}{\sqrt{1+\lambda^2}}. \tag{4.2}$$

As λ ranges from $-\infty$ to $+\infty$, δ ranges from -1 to 1, and therefore δ might be more intuitive and easier to interpret than λ. When $\lambda = 0$ ($\delta = 0$) the skewed-t density reduces to the density of the classical symmetric t-distribution. When $\lambda = \pm\infty$ ($\delta = \pm 1$) the skewed-t density results in a half-t distribution on the positive or negative real line.

Let

$$b_\nu = \left(\frac{\nu}{\pi}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \tag{4.3}$$

The central moments of a $ST(\xi, \omega^2, \lambda, \nu)$ random variable were obtained by Azzalini and Capitanio (2003) - see also Arellano-Valle and Azzalini (2011) - and are as follows:

$$\mu = E(Y) = \xi + \omega\delta b_\nu, \qquad \nu > 1$$

$$Var(Y) = \omega^2\left(\frac{\nu}{\nu-2} - b_\nu^2\delta^2\right), \qquad \nu > 2$$

$$E\{(Y-\mu)^3\} = \omega^3 b_\nu \delta\left[\frac{3\nu}{(\nu-3)(\nu-2)} - \delta^2\left(\frac{\nu}{\nu-3} - 2b_\nu^2\right)\right], \qquad \nu > 3 \tag{4.4}$$

$$E\{(Y-\mu)^4\} = \omega^4\left[\frac{3\nu^2}{(\nu-2)(\nu-4)} - 6\delta^2 b_\nu^2\frac{\nu(\nu-1)}{(\nu-2)(\nu-3)} + \delta^4 b_\nu^2\left(\frac{4\nu}{\nu-3} - 3b_\nu^2\right)\right], \qquad \nu > 4$$

The skewed-t random variable $Y \sim ST(\xi, \omega^2, \lambda, \nu)$ can be represented as a scale mixture of skewed-normal and gamma random variables

$$Y = \xi + \frac{Z}{\sqrt{U}} \tag{4.5}$$

where $Z \sim SN(0, \omega^2, \lambda)$ and is independent of $U \sim Gamma(\nu/2, \nu/2)$ or, equivalently, $U \sim \chi_\nu^2/\nu$.

The $ST(0,1,\lambda,\nu)$ density functions are plotted in Figure 36 for λ equal to 0.5, 1, 2 and 3 ($\delta \approx$ 0.45, 0.71, 0.9 and 0.95) and degrees of freedom ν equal to 3, 5 and ∞. Several qq-plots of skewed-t versus normal distribution quantiles are shown in Figure 37.

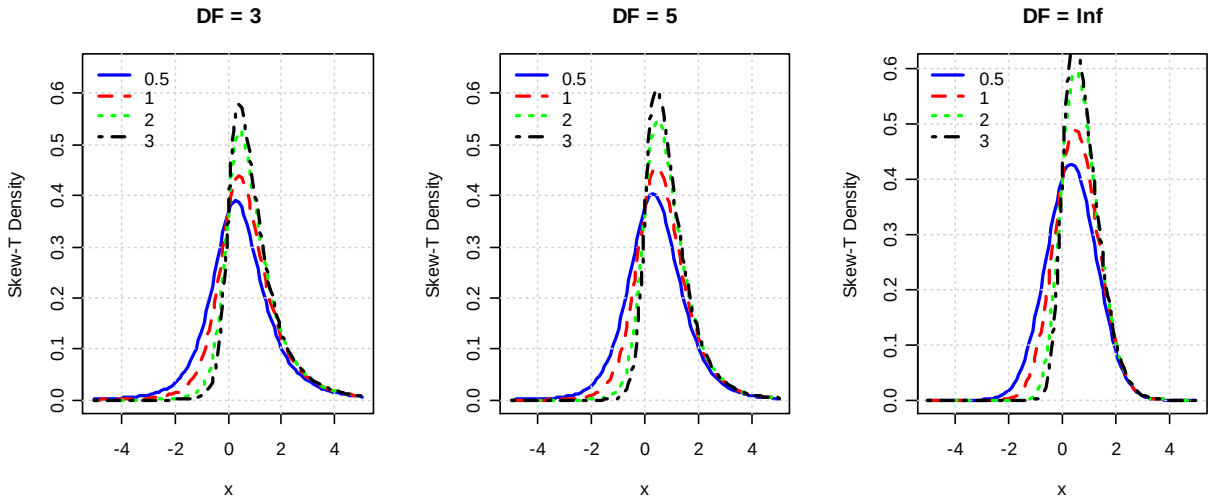

Figure 36. Skewed-t distribution density functions. Location parameter is zero, scale parameter is one.

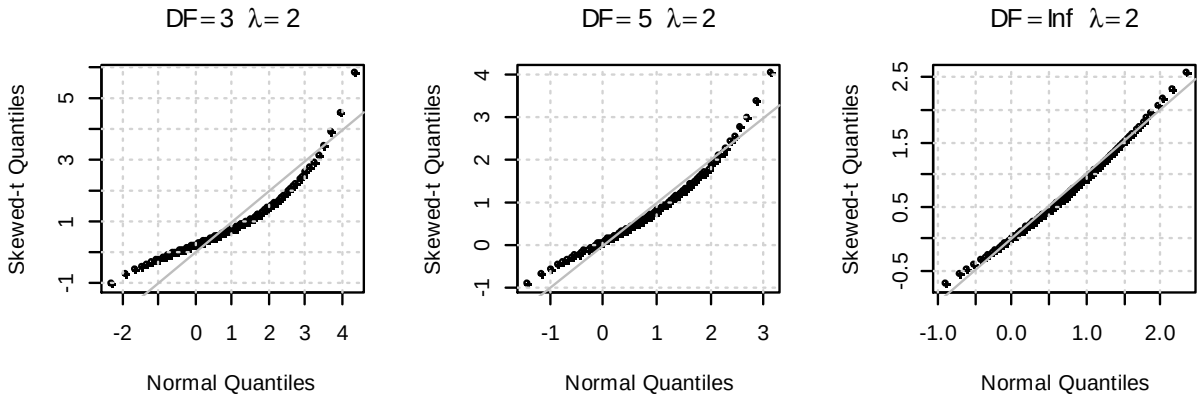

Figure 37. Skewed-t distribution quantiles versus normal distribution quantiles.

In Figure 38 we plot the skewness coefficient $\gamma_1 = \gamma_1(\delta, \nu) = E\{(Y-\mu)^3\}/Var(Y)^{3/2}$ versus values of λ (left-hand figure) and δ (right-hand figure) to facilitate interpretation of the skewness parameters λ and δ. Note that the skewness coefficient does not depend on the values of the location and scale parameters, but does depend on the degrees of freedom ν. Figure 38 shows four curves corresponding to ν equal to 4, 5, 10 and ∞. We start at four degrees of freedom because the 3rd moment and consequently the skewness coefficient do not exist for $\nu \leq 3$. The $ST(\xi, \omega^2, \lambda, \nu)$ distribution with $\nu = \infty$ reduces to the skewed-normal distribution of Azzalini and Capitanio (1999) which we denote $SN(\xi, \omega^2, \lambda)$.

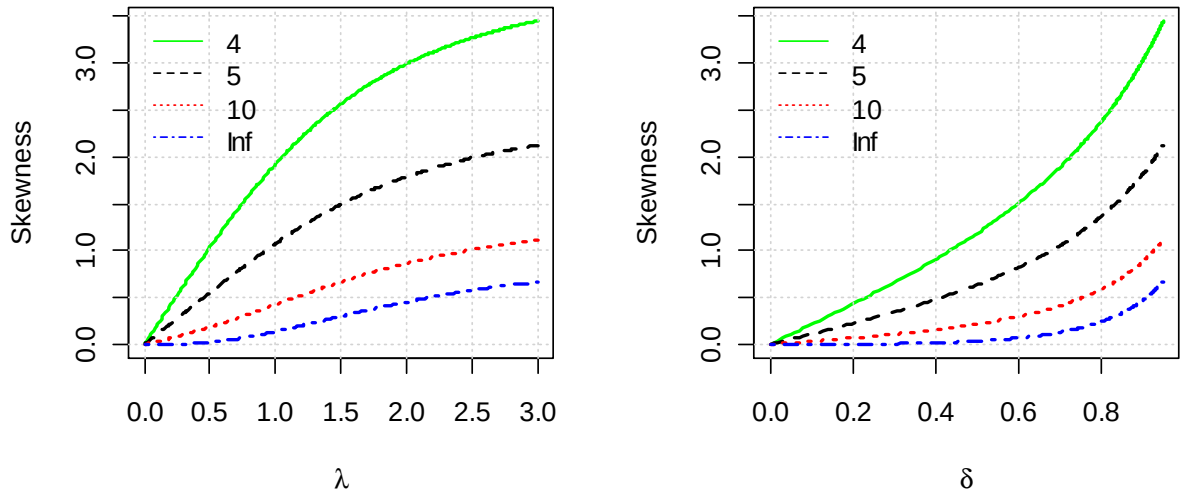

Figure 38. Skewness coefficient vs values of the skewed-t skewness parameter λ (left) and δ (right).

According to Bulmer (1979) 'as a rough guide we may consider a distribution with a skewness greater than 1 in absolute value as highly skewed, a distribution with a skewness between 0.5 and 1 as moderately skewed, and a distribution with a skewness between 0 and 0.5 as fairly symmetrical" (p. 63). Using this guideline, a skewed-t distribution with 5 degrees of freedom and $0.5 \leq \lambda \leq 1$ would be considered as moderately skewed and with $\lambda > 1$ would be considered as highly skewed. A skewed-t distribution with 10 degrees of freedom would be considered moderately skewed when $1 \leq \lambda \leq 2.5$ and highly skewed when $\lambda > 2.5$.

4.3 Regression Slopes under Skewed-t Errors

In this section we investigate efficiency of the MM and LS regression slope estimators under skewed-t distributed errors, asymptotically and in finite samples.

4.3.1 Asymptotic Efficiency

When errors ϵ follow a skewed-t distribution $ST(\xi, \omega^2, \lambda, \nu)$, MM and ST maximum-likelihood slope estimators are consistent estimators of the true regression slopes $\boldsymbol{\beta_0}$. The same is true of the LS estimator under a finite variance assumption on the errors and the independent variables. The asymptotic variances of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{ST}}$ are respectively

$$\begin{aligned} \boldsymbol{V_{LS}} &= \kappa_{LS}\boldsymbol{C_x^{-1}} \\ \boldsymbol{V_{MM}} &= \kappa_{MM}\boldsymbol{C_x^{-1}} \\ \boldsymbol{V_{ST}} &= \kappa_{ST}\boldsymbol{C_x^{-1}} \end{aligned} \tag{4.6}$$

The κ multipliers in the above expressions are as follows:

- $\kappa_{LS} = var(\epsilon)$ with variance formula given in (4.4)
- $\kappa_{MM} = \sigma^2\tau$ where $\tau = E_{F_\epsilon}\left\{\left(\frac{\psi_2((\epsilon-\alpha_{MM})/\sigma)}{E_{F_\epsilon}\psi_2'((\epsilon-\alpha_{MM})/\sigma)}\right)^2\right\}$ (same as eq. (1.21)) and σ is the asymptotic value of the MM initial scale estimator as defined in (1.17)
-

$$\kappa_{ST} = \frac{1}{E(f(u;\omega,\lambda,\nu)^2)} \tag{4.7}$$

where expectation is taken with respect to $u \equiv \epsilon - \xi$ and f is a complicated function defined in (A4.3) in Appendix A4. It should be noted that under the symmetric Student's t-distribution the asymptotic variance of the ST MLE $\widehat{\boldsymbol{\beta}}_{\boldsymbol{ST}}$ (computed at $\lambda = 0$) coincides with that of the T MLE $\widehat{\boldsymbol{\beta}}_{\boldsymbol{T}}$, i.e. κ_{ST} computed at $\lambda = 0$ is equal to $\omega^2\,(\nu+3)/(\nu+1)$.

The common structure of the three asymptotic variances in (4.6) allows simple expressions of the asymptotic efficiencies (see Definition 1 in Chapter 1) of the LS and MM slope estimators at skewed-t distribution, which are κ_{ST}/κ_{LS} and κ_{ST}/κ_{MM} respectively, and are independent of the distribution of the

predictor variables $\boldsymbol{x}$. Furthermore, the ratios κ_{ST}/κ_{LS} and κ_{ST}/κ_{MM} do not depend on the location ξ and scale ω parameters.

The asymptotic efficiencies of the LS and MM bisquare slope estimators at skewed-t distribution are plotted in Figure 39 versus the values of the shape parameter λ. The four panels correspond to the degrees of freedom ν equal to 3, 5, 10 and 15. All mathematical expectations involved in the definitions of κ_{MM} and κ_{ST} are computed via numerical intergration using the R function integrate. The values of σ and α_{MM} are also computed numerically according to (1.17) and (1.16) using the R functions uniroot and optimize.

When $\lambda = 0$ the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{ST}}$ is equal to that of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{T}}$, and hence in Figure 39 at $\lambda = 0$ we see the absolute efficiencies of the MM and LS regression estimators under a symmetric Student's t-distribution. The values of these efficiencies are also reported in Table 7. Our calculations confirm the claim that when errors follow a symmetric fat-tailed distribution, the MM estimates have high absolute efficiency while LS estimates are very inefficient. For example, let's compare MM bisquare at 95% normal distribution efficiency (green triangles) and LS (orange dots) estimators. Their efficiencies are 95.5% (MM) vs 50% (LS) at symmetric t-distribution with 3 degrees of freedom, and 98.4% (MM) vs 80% (LS) at Student's t-distribution with 5 degrees of freedom. Even at 10 degrees of freedom the asymptotic variance of the MM estimator is still somewhat smaller than that of the LS (efficiencies of 98.5% vs 94.5%).

Table 7. Asymptotic efficiency of the LS and MM bisquare slope estimators under symmetric t-distribution.

	MM				LS
DF	0.85	0.9	0.95	0.99	
3	95.4%	96.3%	95.5%	88.1%	50.0%
5	94.1%	96.8%	98.4%	95.5%	80.0%
10	91.0%	95.0%	98.5%	99.4%	94.5%
15	89.4%	93.7%	97.8%	99.9%	97.5%

Next, as degree of skewness grows both MM and LS efficiencies go down. At $\nu = 3$ and $\nu = 5$ the decrease of the MM and LS efficiencies is almost in parallel. Hence, a prominent advantage of the robust

MM estimates over LS estimates that was observed under a symmetric heavy-tailed distribution is preserved when skewness is present. At $\nu = 10$ and $\nu = 15$ LS efficiency appears to decline slightly faster than that of the MM estimates, particularly at $|\lambda| > 1$.

Now, let's look at the values of the absolute asymptotic efficiencies themselves. The bisquare MM estimator at 95% normal distribution efficiency (green triangles) is overall a very good choice since it has:

- Efficiencies of 90%, 78% and 66% at skewed-t distribution with 3 degrees of freedom and skewness parameter λ equal to 1, 2 and 3 respectively (compare to the LS efficiency of 42%, 31% and 23%)
- Efficiencies of 95%, 84% and 72% at skewed-t distribution with 5 degrees of freedom and skewness parameter λ equal to 1, 2 and 3 respectively (compare to the LS efficiency of 74%, 59% and 48%)
- Efficiencies of 96%, 88% and 76% at skewed-t distribution with 15 degrees of freedom and skewness parameter λ equal to 1, 2 and 3 respectively (compare to the LS efficiency of 95%, 83% and 70%)

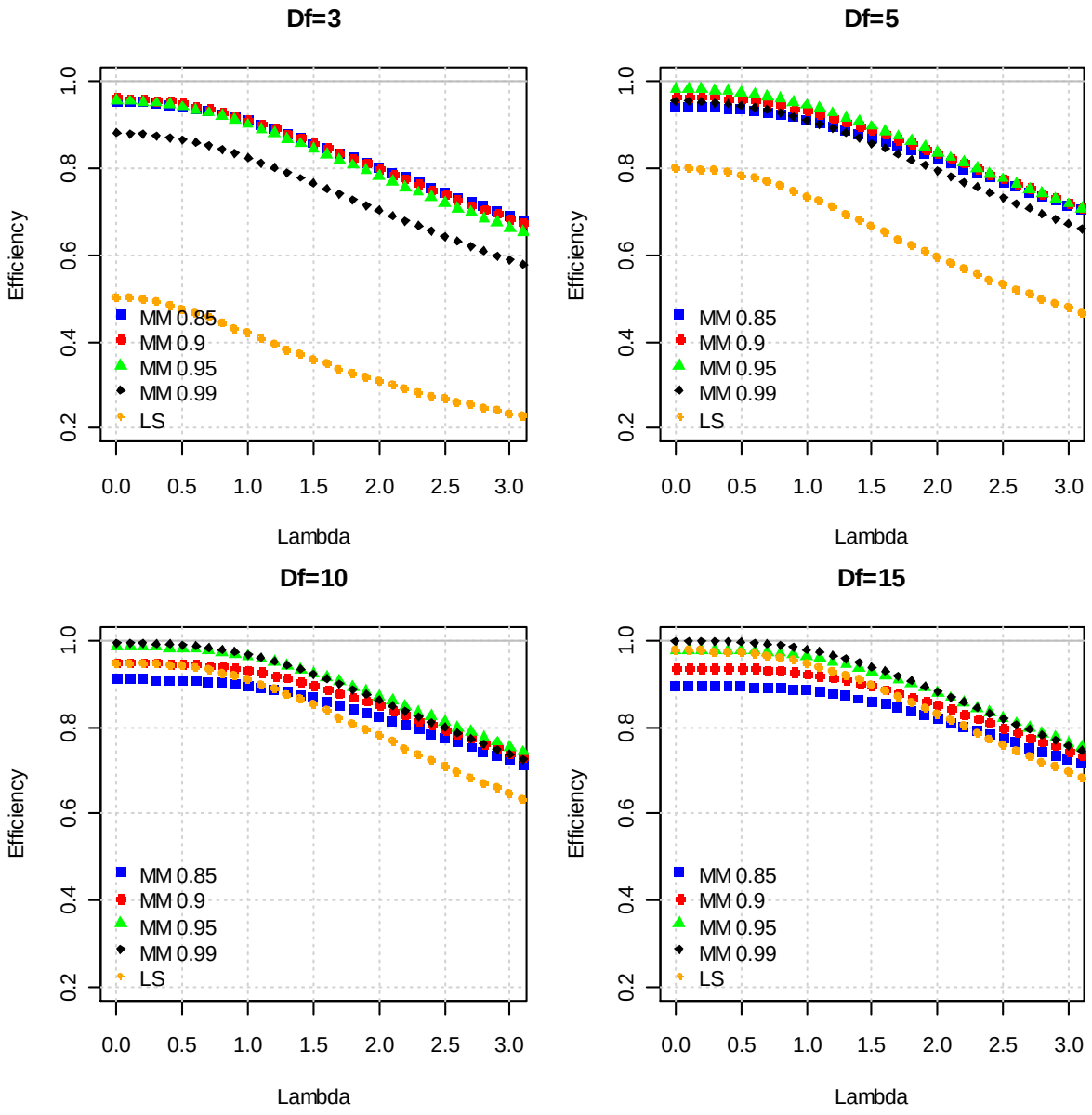

Figure 39. Asymptotic efficiency of the MM (Bisquare, at 85%, 90%, 95% and 99% normal distribution efficiency) and LS slope estimators at skewed-t distribution.

4.3.2 Monte Carlo Simulations

Efficiencies of the LS and MM regression slope estimators in finite samples may differ substantially from the asymptotic efficiencies displayed in Figure 39. To see whether or not this is the case we perform the following simulation study.

We focus on a simple linear regression model $y_i = \beta_0 x_i + \epsilon_i$, $i = 1, \cdots, n$ with independent and identically distributed (i.i.d.) random x_i that are independent of i.i.d. errors ϵ_i. The errors ϵ_i follow $ST(0, \omega^2, \lambda, \nu)$ distribution and x_i follow a normal $N(0, \sigma_x^2)$ distribution. We set $\beta_0 = 1$, $\sigma_x = 2$, $\omega = 1$ and consider the following grid of the values of the skewness and degrees of freedom parameters: $\lambda \in (0, 0.3, 1, 3)$ by $\nu \in (3, 5, 10)$. For each combination of λ and ν we generate Nsim=10,000 datasets with n=25, 50, 100, and 200. For each dataset, we fit simple linear regression using the following methods: LS, MM, ST MLE and symmetric T MLE. We used R package robustbase to fit MM regressions, R package sn to fit ST MLE, and R package hett to fit symmetric t MLE (Taylor, 2012). In addition, we performed simulations under symmetric-t and standard normal error distributions, which are in fact nested in the ST family with $\lambda = 0$. The results of the simulation study will be presented after we describe numerical issues with the ST maximum likelihood regression estimation.

ST Maximum-Likelihood Estimation is not Trust-Worthy

It is known that the maximum likelihood estimation for ST (and SN) models can lead to estimates of the skewness parameter on the boundary of the parameter space, which is infinity in case of λ (Azzalini & Capitanio, 1999; Azzalini & Capitanio, 2003; Azzalini & Genton, 2008). Indeed, large values of the maximum-likelihood estimates of λ, often of the order 10^8 and higher, occured for a substantial number of the simulated samples. Table 8 shows the percent of time $\hat{\lambda}_{ST}$ was larger than 100 for each combination of the sample size, true value of the degrees of freedom ν and shape parameter λ. This proportion increases with λ and decreases with sample size n and can be as high as 44% ($n = 25, \nu = 3, \lambda = 3$). The degrees of freedom appear to have little to no effect on the fraction of times the lambda estimate is abnormally large. Fortunately, the issue with estimating the skewness parameter λ of a skewed-t distribution appears to have little impact on the ST MLEs of the regression slopes. It does, however, affect the ST MLE of the intercept (defined as expected value of the error term).

Table 8. Percent of time (out of 10 000 simulation iterations) the ST MLE of λ was larger than 100.

$\hat{\lambda}_{ST} > 100$		Skewness Parameter ($\boldsymbol{\lambda}$)		
DF($\boldsymbol{\nu}$)	$\boldsymbol{n}$	0.3	1	3
3	25	9.97%	17.90%	44.47%
	50	0.75%	2.57%	16.42%
	100	0.00%	0.05%	1.62%
	200	0.00%	0.00%	0.04%
5	25	10.44%	16.97%	40.84%
	50	0.96%	2.70%	15.42%
	100	0.01%	0.02%	1.39%
	200	0.00%	0.00%	0.02%
10	25	10.46%	15.23%	37.75%
	50	1.15%	2.40%	13.02%
	100	0.03%	0.04%	1.35%
	200	0.00%	0.01%	0.01%

Currently three approaches to dealing with the unboundness of the maximum likelihood estimator of the shape parameter exist in the literature. A modified maximum likelihood method (MMLE), proposed by Sartori (2006) - see also Firth (1993) - addresses this situation for SN models and ST models with fixed degrees of freedom. This method consists of first regular maximum likelihood estimation and subsequent re-estimation of the shape parameter using a modified score function. It is available in the R sn package (functions sn.mmle and st.mmle). For a ST distribution with unknown degrees of freedom, however, an additional difficulty arises as ML estimates of the degrees of freedom may also be on the boundary, either at 0 or ∞, similarly to the case of the symmetric Student's t-distribution (see, e.g., Fernandez & Steel, 1999 and Fonseca, Ferreira, & Migon, 2008). Azzalini and Genton (2008) suggested another, deviance based, approach which consists of replacing the MLE of (λ, ν), in samples where it occurred on the boundary, by the smallest value (λ^*, ν^*) such that $H_0: (\lambda, \nu) = (\lambda^*, \nu^*)$ is not rejected by a likelihood ratio test statistics based on the χ^2_2 distribution at a fixed level, say 0.1. However, simulation results in Azzalini and Genton (2008) showed that this procedure provides only a partial solution to the problem. Very recently Branco, Genton, and Liseo (2012) provided an objective Bayesian solution to the problem for a univariate skewed-t distribution by considering Jeffrey's prior for the shape parameter. Development of the methods for overcoming the practical issues with the skewed-t maximum likelihood estimation is still

an active research area. Almost none of the existing proposals have been made available to a regular practitioner via standard statistical software (R in particular). Since comparing all existing solutions to this problem is not the goal of this chapter the ST results presented below are based on MLE as implemented in R package sn[17].

Also, a few times the ST MLE's of the error scale ω were nearly zero as summarized in Table 9. The scale estimates $\hat{\omega}_{ST}$ were either larger than 0.01 or of the order 10^{-13} and below. We take the latter as an indicator of a numerical failure of the ML method so that in this case ML estimates for all other parameters should be considered unreliable. Indeed, whenever $\hat{\omega}_{ST}$ was nearly zero, the corresponding values of $\hat{\beta}_{ST}$ appeared to be gross outliers. For example, at $\nu = 3$ and $\lambda = 3$ among 10 000 simulated datasets of sample size n=100 two had $\hat{\omega}_{ST}$ of the order 10^{-17} and 10^{-18} with the corresponding $\hat{\beta}_{ST}$ equal to -2.5 and -3.5. For the remaining 9998 simulations the scale MLE's $\hat{\omega}_{ST}$ ranged from 0.46 to 1.91, and the slope MLE's $\hat{\beta}_{ST}$ ranged from 0.86 to 1.14. It is important to note that neither LS nor MM slope estimators failed for these two datasets and had estimate values $\hat{\beta}_{LS} = 1.31\ (1.01)$ and $\hat{\beta}_{MM} = 0.82\ (1.08)$, respectively.

Table 9. Percent of time (out of 10 000 simulation iterations) the ST ML estimate of error scale ω was of the order 10^{-13} and below.

$\hat{\omega}_{ST}$ is nearly zero		Skewness Parameter (λ)		
DF(ν)	n	0.3	1	3
3	25	0.00%	0.01%	0.01%
	50	0.03%	0.00%	0.00%
	100	0.02%	0.02%	0.02%
	200	0.00%	0.00%	0.03%
5	25	0.01%	0.00%	0.00%
	50, 100, 200	0.00%	0.00%	0.00%
10	25, 50, 100, 200	0.00%	0.00%	0.00%

[17] In fact, our attempt to use Sartori's MMLE (as available in sn package) and Azzalini and Genton (2008) deviance approach (our own implementation) did not result in a substantial (if any) improvement.

Furthermore, Table 10 shows how often in our simulations the ST ML estimates of the degrees of freedom $\hat{\nu}_{ST}$ were less than one. Recall that the mean of a skewed-t distribution (4.4) is only defined for the degrees of freedom greater than one, and therefore the ST MLE of the intercept is meaningless whenever $\hat{\nu}_{ST} \leq 1$. Not surprisingly, the lower the true degrees of freedom and the smaller the sample size the higher the chance for the estimated degrees of freedom being less than one.

Table 10. Percent of time (out of 10 000 simulation iterations) the ST ML estimate of ν was less than 1.

	$\hat{\nu}_{ST} < 1$	Skewness Parameter (λ)		
DF(ν)	n	0.3	1	3
3	25	1.42%	1.62%	1.60%
	50	0.07%	0.02%	0.02%
	100	0.02%	0.02%	0.02%
	200	0.00%	0.00%	0.03%
5	25	0.56%	0.51%	0.54%
	50, 100, 200	0.00%	0.00%	0.00%
10	25	0.20%	0.22%	0.22%
	50, 100, 200	0.00%	0.00%	0.00%

Simulation Results

Table 11 displays the finite sample Monte Carlo variance of the slope estimators from the simulated data sets multiplied by the corresponding sample size *n* for sample sizes 25, 50, 100 and 200, along with the asymptotic variance (marked n=Inf). The datasets with a nearly zero ST scale MLE $\hat{\omega}_{ST}$ (see Table 9) were removed before calculating summaries in Table 11. The deletion did not visibly change MM and LS variances, but substantially affected sample variances of the ST slope estimates for the ν, λ and n combinations in which deletion occurred. For example, for $\nu = 3, \lambda = 3$ and n=100 the ST sample variance dropped from 0.454 (when based on all 10 000 $\hat{\beta}_{ST}$ estimates) to just 0.133 (if based on 9998 $\hat{\beta}_{ST}$ estimates). The asymptotic variance tends to underestimate finite sample variability of the slope estimators, but as sample size increases the approximation gets better as expected. The magnitude of the underestimation for ST MLE's is much larger than that of the MM and LS estimators. For example, the ratio of the sample variance at n=25 to the asymptotic variance varies from 1.5 ($\nu = 10, \lambda = 0.3$) to 1.9 ($\nu = 3, \lambda = 3$) for ST MLEs, but only from 1.1 to 1.4 for the MM and LS. As a result, the finite sample

efficiency of the MM and LS slope estimators could be much larger than their asymptotic efficiency depicted in Figure 39. In fact, the MM slope estimators can have smaller variance compared to the ST MLEs in small to moderate sample sizes under mild to moderate skewness. Such cases are highlighted with bold font in Table 11. Figure 40, which we describe subsequently, illustrates this claim graphically.

Table 11. Variance of the slope estimators in a simple linear regression: Monte Carlo simulations versus asymptotic approximation. Bold highlights cases where the sample variances of the MM or LS slope estimates were less than or equal to the sample variance of the skewed-t MLEs.

		Skewness Parameter (λ)								
		0.3			1			3		
DF (ν)	n	MM	LS	ST	MM	LS	ST	MM	LS	ST
3	25	**0.474**	0.828	0.633	**0.356**	0.650	0.444	**0.217**	0.651	0.217
	50	**0.413**	0.738	0.442	**0.308**	0.639	0.320	0.189	0.474	0.161
	100	**0.386**	0.734	0.386	0.291	0.668	0.277	0.178	0.473	0.133
	200	0.384	0.725	0.374	0.283	0.602	0.258	0.173	0.503	0.118
	Inf	0.375	0.725	0.356	0.278	0.598	0.251	0.168	0.476	0.112
5	25	**0.382**	**0.456**	0.502	**0.286**	**0.345**	0.359	**0.178**	0.243	0.181
	50	**0.353**	0.421	0.388	**0.260**	0.327	0.280	0.153	0.226	0.140
	100	**0.335**	0.411	0.343	**0.243**	0.312	0.244	0.150	0.220	0.118
	200	**0.325**	0.409	0.327	0.243	0.309	0.234	0.145	0.219	0.111
	Inf	0.322	0.398	0.316	0.236	0.304	0.224	0.142	0.214	0.102
10	25	**0.333**	**0.333**	0.428	**0.245**	**0.246**	0.306	**0.155**	0.169	0.162
	50	**0.310**	**0.322**	0.338	**0.223**	**0.231**	0.240	0.137	0.155	0.131
	100	**0.297**	0.310	0.304	**0.214**	0.225	0.217	0.130	0.150	0.107
	200	**0.294**	0.304	0.295	0.211	0.223	0.208	0.123	0.144	0.097
	Inf	0.285	0.297	0.280	0.207	0.219	0.199	0.124	0.144	0.093

For each ν, λ and n combination in Table 10 and each slope estimator E we compute mean squared error from Nsim simulated datasets as follows

$$MSE_E(n;\nu,\lambda) = \frac{1}{Nsim}\sum_{j=1}^{Nsim}\left(\hat{\beta}_{E,j}(n;\nu,\lambda) - \beta_0\right)^2$$

with $\beta_0 = 1$, the true value of the slope. Consistent with Table 11, the datasets that resulted in nearly zero ST error scale estimates (see Table 9) were removed before calculating MSE so that Nsim above ranges from 9 998 to 10 000. We call the ratio

$$\frac{MSE_{ST}(n;\nu,\lambda)}{MSE_{E}\ (n;\nu,\lambda)}$$

the Monte Carlo finite sample efficiency of the slope estimator E under ST errors. Figure 40 displays these finite sample efficiencies for the LS, MM (bisquare at 95% normal distribution efficiency) and symmetric T maximum likelihood estimates for the slope in a simple linear regression.

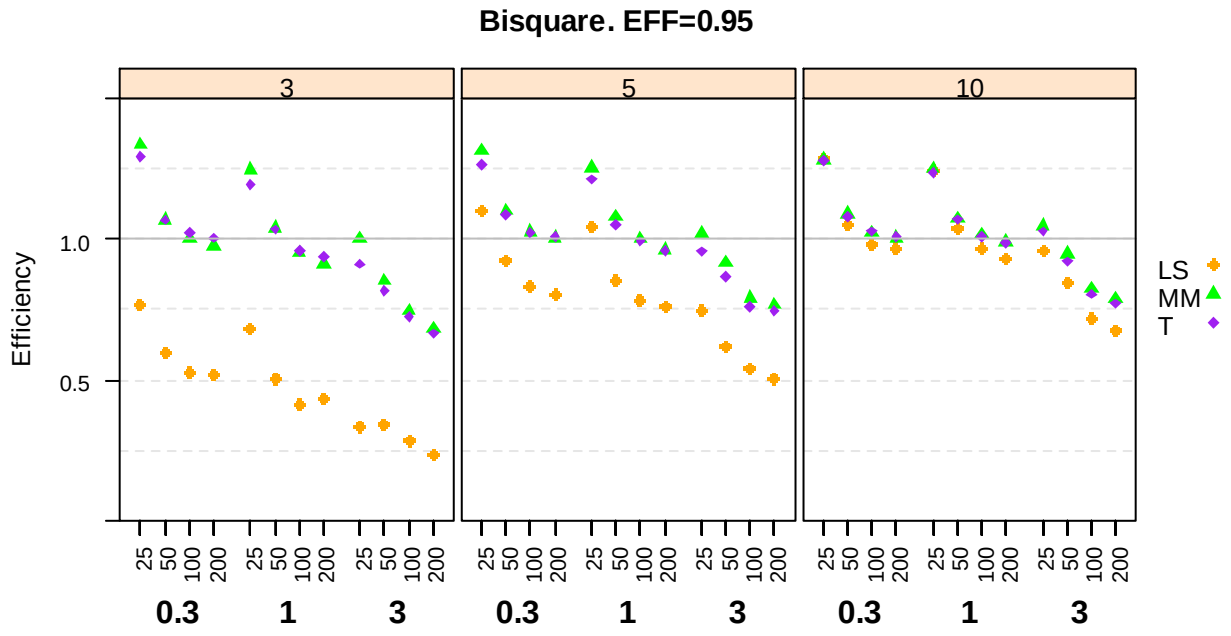

Figure 40. Monte Carlo finite sample efficiency of the LS, MM and symmetric T maximum likelihood slope estimates relative to ST MLE for df=3, 5, and 10, skewness parameter $\lambda = 0.3, 1,$ and 3, and sample sizes n=25, 50, 100 and 200.

From Figure 40 we have the following conclusions:

1) With regard to the LS and MM efficiencies in finite sample versus asymptotic efficiencies:
    a. Efficiencies at n=200 are very close to the corresponding asymptotic values from Figure 39.
    b. As sample size decreases, both LS and MM efficiencies increase and have differences between n=25 and n=200 that are relatively large. For example, at $\lambda = 3$ and df=3 the MM estimate efficiency is about 100% at n=25 compared to only 65% asymptotic efficiency.

c. In finite samples MM and LS efficiency can be larger than one. For example, for the MM estimate this is the case at sample sizes n=25 and 50, $\lambda = 0.3$ and 1, and df equal to 3, 5 and 10 (i.e. on all three panels).

2) With regard to MM efficiency versus LS efficiency:
   a. At 10 degrees of freedom the difference between LS and MM efficiencies is small.
   b. As the tails of the error distribution get fatter, i.e., as the degrees of freedom decrease, that gap between the LS and MM efficiency increases. At df=5 and even more so at df=3, the MM slope estimator is far more efficient than LS regardless of the degree of skewness, and the gain in efficiency appears to be similar across all sample sizes n.

3) With regard to efficiency of the symmetric T MLE versus MM efficiency
   a. The symmetric T maximum likelihood slope estimator has finite sample efficiencies almost identical to those of the MM estimator. It should be noted that when it comes to the degrees of freedom estimation, symmetric Student's T MLEs exhibit the same problem as ST MLEs: in our simulations it was not uncommon to see extremely large degrees of freedom estimates as well as very low values in a few cases.

Finally, we present simulation results for two symmetric distributions: standard Student-t with 3 degrees of freedom and standard normal (marked DF=Inf). Table 12 displays variance of the slope estimates from the simulated data sets multiplied by the corresponding sample size n along with the asymptotic variance which is marked as n=Inf. The corresponding results for the intercept will be discussed later in Section 4.4. Summaries in Table 12 are based on 10 000 simulation runs with 13 cases with nearly zero ST scale estimates excluded (ST scale estimate was either below 10^{-13} or larger than 0.03).

For the t-distribution with 3 degrees of freedom MM has high efficiency while LS is very inefficient, both in finite samples and asymptotically. For small sample sizes n = 25 and 50 the MM had 102% and 100% efficiency relative to the T MLE, and even higher relative to the ST MLE, thereby outperforming the T MLE slightly and the ST MLE substantially (about 30% larger variance for ST MLE at n = 25). For large

sample sizes the MM had 96%, 95% and 95% efficiency at n = 200, 500, and Inf (asymptotically) respectively relative to the T MLE and 97%, 95% and 95% at n = 200, 500, and Inf respectively relative to the ST MLE.

At normal distribution MM asymptotic efficiency is 95% by the choice of the tuning constant and in our simulations MM efficiency relative to LS varied from 94% to 96%. In small sample sizes T MLE has slightly larger variance compared to LS, which is a penalty for estimating DF. The penalty disappears in large sample sizes with the asymptotic variances of $\hat{\beta}_{LS}$ and $\hat{\beta}_{T}$ being equal.

Table 12. Variance of the slope estimators under symmetric t-distribution errors.

DF (ν)	n	MM	LS	T	ST
3	25	0.506	0.849	0.516	0.653
	50	0.427	0.770	0.428	0.464
	100	0.409	0.740	0.398	0.405
	200	0.392	0.749	0.377	0.381
	500	0.398	0.743	0.377	0.379
	Inf	0.393	0.750	0.375	0.375
Inf	25	0.306	0.286	0.299	0.384
	50	0.283	0.266	0.272	0.297
	100	0.270	0.255	0.258	0.266
	200	0.271	0.257	0.259	0.262
	500	0.263	0.252	0.252	0.254
	Inf	0.263	0.250	0.250	0.250

Note much larger variance of the ST slope estimates as compared to both MM estimator and T MLE for n=25, and larger variance for n=50. This is a penalty for estimating skewness parameter even if its' true value is zero. The penalty disappears in large sample sizes with the asymptotic variances of $\hat{\beta}_{ST}$ and $\hat{\beta}_{T}$ being equal, i.e., no asymptotic penalty for the ST slope estimator due to estimating a skewness parameter.

Overall, the simulation study in this section suggests that unless a severe skewness is anticipated, e.g., skewness parameter value larger than 3, or the sample size is larger than roughly 100, it will not be very

beneficial to use ST MLE of the slopes. For the ST skewness parameter values of 1 and below and sample sizes of 50 and below the T MLE and MM slope estimates actually exhibited much smaller variability compared to that of the ST MLE for both very heavy tails arising with 3 to 5 degrees of freedom and moderately heavy tails arising with10 degrees of freedom.

4.4 Regression Intercept under Skewed-t Errors

We now switch our attention from the regression slopes to an intercept. It is desirable for an intercept to account for skewness in the error term in a manner that has a meaningful interpretation. One way to do this is to define the intercept parameter as the mean value of the error term:

$$\alpha_0 = E\epsilon \tag{4.8}$$

This definition is consistent with use of the conditional expected value $E(y|\boldsymbol{x})$ of y given $\boldsymbol{x}$ as predictor of a new observation y under the linear regression model (1.1) and (1.3) with errors independent of predictors, for in that case $E(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta_0} + \alpha_0$.

Other possible definitions of α_0 include taking it to be the mode or median of the error distribution, which have the advantage of existing for very fat tailed error distributions, e.g., t or skewed-t distribution with $\nu \leq 1$ (see, for example, Azzalini an Genton (2008) or Arellano-Valle and Azzalini (2011)). None-the-less there is good motivation for wanting a robust estimate of the mean of an asymmetric error distribution so we henceforth focus on the definition (4.8) of the intercept as the mean value of the error term, tacitly assuming that mean value of the error distribution exists.

Under symmetric error distribution whose mean value exists, both LS and MM intercept estimators are consistent for α_0. Under asymmetry the LS intercept can absorb any non-zero mean of the errors and is therefore still an unbiased estimator of α_0. However, under asymmetric errors the MM intercept estimator will typically fail to be a consistent estimator of α_0. Moreover, the asymptotic value of the MM intercept is given by a solution of an equation that involves the expected value of a non-linear transformation of ϵ (see Chapter 1, eq (1.16)), which lacks a convenient interpretation in the presence of skewness. This

prompted us to study the extent of the bias of the MM intercept, and to propose a simple "corrected" MM-intercept estimator that we call a c-MM intercept estimator. In line with our choice of the error distributions in Section 4.3 we evaluate the performance of the LS, MM and c-MM intercept estimators under skewed-t errors using both asymptotic approximations and finite sample Monte Carlo studies. In particular we

a) Compare the mean-squared errors (MSE) of the MM, c-MM, LS and ST ML estimators for a range of sample sizes.
b) Study the percent contribution of bias to the MSE of the MM and c-MM intercept estimators for a range of sample sizes.

While our numerical investigation is focused on the skewed-t distribution, much of the discussion applies to other asymmetric error distributions.

4.4.1 MM Intercept and c-MM Intercept Estimators

Recall from (1.16) in Chapter 1 that under the regression model (1.1) and (1.3) the MM intercept $\hat{\alpha}_{MM}$ is a consistent estimator of the parameter α_{MM} whose expression is

$$\alpha_{MM} = \underset{t}{\operatorname{argmin}}\, E_{F_\epsilon}\rho_2((\epsilon - t)/\sigma\,).$$

It is important to note the following three cases with respect to the symmetry or asymmetry of F_ϵ:

1) If F_ϵ is symmetric around α_0 then $\alpha_0 = E(\epsilon)$ and $\alpha_{MM} = \alpha_0$, i.e. the asymptotic bias of the MM intercept estimator is 0: $ABias_{MM} \equiv \alpha_{MM} - \alpha_0 = 0$
2) If F_ϵ is positively skewed then $\alpha_{MM} < E(\epsilon) = \alpha_0$, i.e. the asymptotic bias of the MM intercept estimator is negative: $ABias_{MM} \equiv \alpha_{MM} - \alpha_0 < 0$
3) If F_ϵ is negatively skewed then $\alpha_{MM} > E(\epsilon) = \alpha_0$, i.e. the asymptotic bias of the MM intercept estimator is positive: $ABias_{MM} \equiv \alpha_{MM} - \alpha_0 > 0$

Intuitively, the above behavior in (2) and (3) follows from the symmetric shape of the loss function of the MM estimator. Positive (negative) skewness in the underlying distribution results in a larger portion of the

right (left) tail data being down-weighted by the MM estimator relative to the left (right) tail data, which makes MM intercept underestimate (overestimate) the mean of the error distribution.

Under general conditions given in Chapter 1, the MM intercept estimator $\hat{\alpha}_{MM}$ is asymptotically normal with asymptotic variance

$$V_{\hat{\alpha}_{MM}} = \kappa_{MM}\left(1 + \boldsymbol{\mu}_{\boldsymbol{x}}'\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{x}}\right) + \sigma^2\vartheta_{MM} \tag{4.9}$$

where $\kappa_{MM} = \sigma^2\tau = \sigma^2 f_{MM}/a_{MM}^2$ is the same as for the slopes in (4.6) and

$$\vartheta_{MM} = \frac{e_{MM}}{a_{MM}^2 d_s}\left(E\left(\rho_1^2\left(\frac{u_S}{\sigma}\right) - b^2\right)\frac{e_{MM}}{d_s} - 2g_{MM}\right) \tag{4.10}$$

with f_{MM}, a_{MM}, e_{MM}, g_{MM} and d_S given in (A2.7) and also restated below for readers convenience

$$f_{MM} = E_{F_\epsilon}\left\{\psi_2^2\left(\frac{u_{MM}}{\sigma}\right)\right\} \qquad a_{MM} = E_{F_\epsilon}\psi_2'\left(\frac{u_{MM}}{\sigma}\right)$$

$$e_{MM} = E_{F_\epsilon}\left\{\psi_2'\left(\frac{u_{MM}}{\sigma}\right)\frac{u_{MM}}{\sigma}\right\} \qquad d_S = E_{F_\epsilon}\left\{\psi_1\left(\frac{u_S}{\sigma}\right)\frac{u_S}{\sigma}\right\}$$

$$g_{MM} = E_{F_\epsilon}\left\{\left(\rho_1\left(\frac{u_S}{\sigma}\right) - b\right)\psi_2\left(\frac{u_{MM}}{\sigma}\right)\right\} = E_{F_\epsilon}\left\{\rho_1\left(\frac{u_S}{\sigma}\right)\psi_2\left(\frac{u_{MM}}{\sigma}\right)\right\}$$

The term ϑ_{MM} represents the effect of the estimation of the error scale on the variance of the MM intercept. Note that $\vartheta_{MM} = 0$ when error term distribution F_ϵ is symmetric as in this case symmetry in the loss function ρ_2 yields $e_{MM} = 0$.

The corrected MM intercept (c-MM)

Since we are taking the intercept to be the mean of an error term that has a skewed distribution our problem is to find a reasonably simple robust estimate of the mean of a skewed distribution. This is not the same as estimating the location parameter of a symmetric fat-tailed distribution, and in general estimation of the mean of a skewed distribution is a difficult problem. See, for example, Marazzi and Barbati (2002) for a review and references to related research results. The methods treated in this literature are focused on distributions that reduce to a location and scale parameter model after a suitable

nonlinear transformation. As such they apply to the case of a skewed t-distribution that we are focused on only when skewness and degrees of freedom are known, which is not the case in practice. There may not be a satisfactory theoretical way to specify a good robust estimator of a mean of a skewed distribution when degree of skewness and tail fatness are unknown. A simple and intuitive robust estimator that we propose, namely a corrected MM intercept (c-MM) estimator, is designed to reduce the MM intercept bias in estimating the mean of a skewed t-distribution while maintaining robustness toward gross outliers.

The estimation steps are:

1) First, estimate the regression slopes with an MM regression with a user-specified normal distribution efficiency (e.g., 95%), including an intercept term in the model but treating it as a nuisance parameter.
2) Compute the residuals from step 1, excluding the intercept, i.e., $r_i = y_i - \boldsymbol{x_i'}\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$, and estimate α_0 as the location parameter of r_i using an M-estimator

$$\hat{\alpha}_{cMM} = \underset{t}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_c \left(\frac{r_i - t}{\hat{\sigma}_1} \right), \tag{4.11}$$

where $\hat{\sigma}_1$ is a residual scale estimate from the MM fit in step 1 and ρ_c is a re-descending loss function (e.g., bisquare or optimal) with the tuning constant c selected to be much larger than c_2 in step 1.

The following points about the c-MM estimate are to be noted:

- By using a re-descending loss function ρ_c we keep the breakdown point of one half of the initial S-estimator.
- It is not desirable to use a large value of the tuning constant c_2 in step 1 as that would lead to low efficiency of the MM slope estimates. Figure 39 suggests that a value of c_2 corresponding to a 95% normal distribution efficiency is a good choice. The idea is to use larger tuning constant value only for an intercept and only if skewness is anticipated and intercept bias is of concern.

- In section 4.4.2 we will see that for certain degrees of tail fatness, skewness and sample sizes the corrected intercept estimator $\hat{\alpha}_{cMM}$ has smaller mean squared error than the original MM intercept estimator as well as the LS intercept estimator.
- The choice of a robust location estimator in step 2 is not limited to an M-estimator. Another attractive option, for example, is a trimmed mean with a very small percentage of trimming, such as 1% or 2%.
- Whether the asymptotic theory standard errors are known or not (and we provide asymptotic variance only for a c-MM intercept as defined by (4.11), but not for a trimmed mean), the finite sample inference can also be done with the help of a bootstrap. See, for example, Salibian-Barrera and Zamar (2002) for a fast bootstrap method for an MM regression.

c-MM Asymptotic Distribution

Since the residuals from an MM-estimate in step (1) above will behave asymptotically like the true errors in the regression model one can show (see Appendix A4) that $\hat{\alpha}_{cMM}$ is a consistent estimator of

$$\alpha_{cMM} = \underset{t}{\operatorname{argmin}} E_{F_\epsilon} \rho_c \left(\frac{\epsilon - t}{\sigma} \right). \tag{4.12}$$

Furthermore we show in Appendix A4 that $\hat{\alpha}_{cMM}$ is asymptotically normal with variance

$$V_{\hat{\alpha}_{cMM}} = \sigma^2 \left(\frac{f_{cMM}}{a_{cMM}^2} + \frac{f_{MM}}{a_{MM}^2} \boldsymbol{\mu}_{\boldsymbol{x}}' \boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}} \boldsymbol{\mu}_{\boldsymbol{x}} \right) + \sigma^2 \vartheta_{cMM} \tag{4.13}$$

where

$$\vartheta_{cMM} = \frac{e_{cMM}}{a_{cMM}^2 d_s} \left(E\left(\rho_1^2 \left(\frac{u_S}{\sigma} \right) - b^2 \right) \frac{e_{cMM}}{d_s} - 2 g_{cMM} \right) \tag{4.14}$$

$$u_c = \epsilon - \alpha_{cMM}$$

$$a_{cMM} = E_{F_\epsilon}\psi_c'\left(\frac{u_c}{\sigma}\right)$$

$$e_{cMM} = E_{F_\epsilon}\left\{\psi_c'\left(\frac{u_c}{\sigma}\right)\frac{u_c}{\sigma}\right\} \tag{4.15}$$

$$f_{cMM} = E_{F_\epsilon}\left\{\psi_c^2\left(\frac{u_c}{\sigma}\right)\right\}$$

$$g_{cMM} = E_{F_\epsilon}\left\{\left(\rho_1\left(\frac{u_S}{\sigma}\right) - b\right)\psi_c\left(\frac{u_c}{\sigma}\right)\right\} = E_{F_\epsilon}\left\{\rho_1\left(\frac{u_S}{\sigma}\right)\psi_c\left(\frac{u_c}{\sigma}\right)\right\}$$

with ψ_c the derivative of ρ_c and f_{MM}, a_{MM}, and d_s are given by (A2.7). We note that ϑ_{cMM} is equal to zero for a symmetric error distribution.

The asymptotic variance (4.13) of a c-MM intercept estimator can also be written as an asymptotic variance (4.9) of an MM intercept estimator plus a correction term as follows

$$V_{\hat{\alpha}_{cMM}} = V_{\hat{\alpha}_{MM}} + \sigma^2\left(\frac{f_{cMM}}{a_{cMM}^2} - \frac{f_{MM}}{a_{MM}^2} + \vartheta_{cMM} - \vartheta_{MM}\right)$$

Note that when $\psi_c = \psi_2$ then c-MM intercept estimate is the same as the original MM intercept and the two asymptotic variance expressions (4.13) and (4.9) coincide.

Selection of the Tuning Constant c

Recall that *c* defines the rejection region of the robust estimator. Hence, one option for choosing c is to assume a certain degree of tail fatness, say the worst anticipated, and consider the probability of observing data points that are likely to be rejected by a robust estimator $\hat{\alpha}_{cMM}$. Then one can choose c for which such probability is comfortably small ensuring that only gross outlying observations that are rarely observed under assumed error distribution are rejected. Conveniently, for the case of a skewed-t distribution of Azzalini and Capitanio (2003) described in Section 4.2 this probability does not depend on the shape parameter.

Note that $Pr\left(\frac{|Y-\xi|}{\omega} \geq c\right)$ for $Y \sim ST(\xi, \omega^2, \lambda, \nu)$ depends only on the degrees of freedom ν and c, and we show this probability for the four values of ν and the four values of c in Table 13. To put these numbers into prospective and to get an idea of how often there will be at least one point rejected in a sample of size n=100 we include Table 14, which shows probabilities of at least one observation being more than c scales away from the location parameter in an i.i.d. sample from a skewed-t distribution. Note that these are not exactly the probabilities of a data point being rejected by a robust M-estimator as the latter involves the asymptotic values of the robust intercept and scale which depend on the loss function. Therefore, the numbers in Table 13 and Table 14 must be viewed just as a crude reference.

We will consider c=15 and the bisquare loss function in the bias and mean squared error studies in the next sections. The choice of c=15 may seem to be too extreme for a fast re-descending optimal psi, but it is not so extreme for a slow re-descending bisquare psi function. Recall the differences in the shapes of the bisquare and optimal psi-functions discussed in Chapter 1 and that for the bisquare psi c=7.04 leads to a 99% efficiency at a normal distribution.

Table 13. $Pr(|Y| \geq c)$ where $Y \sim ST(0, 1, \lambda, \nu)$.

		Degrees of Freedom			
		3	5	10	Inf
c	5	0.015392	0.004105	0.000537	5.7E-07
	10	0.002128	0.000171	0.000002	7.6E-24
	15	0.000643	0.000024	3.5E-08	3.7E-51
	20	0.000273	0.000006	2.1E-09	2.8E-89

Table 14. $\Pr(\sum_{i=1}^{100} I[|Y_i| \geq c] \geq 1)$ for $Y_1, \ldots, Y_{100}$ i.i.d. from $ST(0,1,\lambda,\nu)$. $I[z]$ is a 0/1 indicator function.

		Degrees of Freedom			
		3	5	10	Inf
c	5	0.788009	0.337223	0.052329	0.000057
	10	0.191898	0.016951	0.000159	0.000000
	15	0.062307	0.002382	0.000003	0.000000
	20	0.026954	0.000577	0.000000	0.000000

∞MM Estimator

In this subsection we discuss the pros and cons of using the average of the MM residuals as an intercept estimate in step 2. We call it an ∞MM estimate as it can be viewed as the limiting case of a *c-MM* estimate with $\psi_\infty(z) = z$. The asymptotic variance can be obtained formally from (4.13) by checking that $\vartheta_{cMM} = 0$ for $\psi_\infty(z) = z$. The result is

$$V_{\hat{\alpha}_{\infty MM}} = var(\epsilon) + \sigma^2 \frac{f_{MM}}{a_{MM}^2} \boldsymbol{\mu_x'} \boldsymbol{C_x^{-1}} \boldsymbol{\mu_x} \tag{4.16}$$

Similarly to the LS intercept estimator the ∞MM estimator is consistent for $\alpha_0 = E\epsilon$, but not robust and among other things has BP=0. While the ∞MM intercept itself is no better than the LS intercept, we claim that combining it with the more efficient MM slope estimates can lead to better predictions of $E(y|\boldsymbol{x})$ at $\boldsymbol{x}$ values further from $\boldsymbol{\mu_x}$, the center of the predictor vector**.** The formal explanation of this statement is below.

Consider two estimators of the conditional mean $E(y|\boldsymbol{x}) = \alpha_0 + \boldsymbol{x}'\beta_0$:

LS) $\hat{E}(y|\boldsymbol{x})_{LS} = \hat{\alpha}_{LS} + \boldsymbol{x}'\widehat{\boldsymbol{\beta}}_{\boldsymbol{LS}}$ and

∞MM) $\hat{E}(y|\boldsymbol{x})_{\infty MM} = \hat{\alpha}_{\infty MM} + \boldsymbol{x}'\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$

Using the following expression for $cov(\hat{\alpha}_{\infty MM}, \widehat{\boldsymbol{\beta}}_{MM}) = -\frac{f_{MM}}{a_{MM}^2} \boldsymbol{C_x^{-1}} \boldsymbol{\mu_x}$ derived in the Appendix A4 we have

$$Var\left(\hat{E}(y|\boldsymbol{x})_{LS}\right) = var(\epsilon) + var(\epsilon)(\boldsymbol{x} - \boldsymbol{\mu_x})' \boldsymbol{C_x^{-1}} (\boldsymbol{x} - \boldsymbol{\mu_x})$$

$$Var\left(\hat{E}(y|\boldsymbol{x})_{\infty MM}\right) = var(\epsilon) + \kappa_{MM}(\boldsymbol{x} - \boldsymbol{\mu_x})' \boldsymbol{C_x^{-1}} (\boldsymbol{x} - \boldsymbol{\mu_x})$$

$$Var\left(\hat{E}(y|\boldsymbol{x})_{LS}\right) - Var\left(\hat{E}(y|\boldsymbol{x})_{\infty MM}\right) = (\kappa_{LS} - \kappa_{MM})(\boldsymbol{x} - \boldsymbol{\mu_x})' \boldsymbol{C_x^{-1}} (\boldsymbol{x} - \boldsymbol{\mu_x})$$

Since $\boldsymbol{C_x}$ is positive definite then $(\boldsymbol{x} - \boldsymbol{\mu_x})' \boldsymbol{C_x^{-1}} (\boldsymbol{x} - \boldsymbol{\mu_x}) \geq 0$ and is equal to 0 if and only if $\boldsymbol{x} = \boldsymbol{\mu_x}$**.**

Hence,

- When $\boldsymbol{x} = \boldsymbol{\mu_x}$ the two asymptotic variances are equal regardless of the error distribution. Moreover, the minimum of both variances is attained at $\boldsymbol{x} = \boldsymbol{\mu_x}$.
- When $\boldsymbol{x} \neq \boldsymbol{\mu_x}$, the variance of $\hat{E}(y|\boldsymbol{x})_{\infty MM}$ is smaller than that of the LS estimator whenever $\kappa_{LS} > \kappa_{MM}$. The difference in the variances of the two conditional means is larger the farther the predictor variables vector $\boldsymbol{x}$ is from its mean $\boldsymbol{\mu_x}$**.** Recall from Figure 39 that κ_{LS} is substantially larger than κ_{MM} for fat-tailed error distributions. At the same time under normal errors even though κ_{MM} is larger than κ_{LS} the difference is small as determined by the choice of a tuning constant of the MM estimator.

Hence, the smallest variance of the fitted values $\hat{y}_{LS}(\boldsymbol{x})$ and $\hat{y}_{\infty MM}(\boldsymbol{x})$ is the same and occurs at $\boldsymbol{x} = \boldsymbol{\mu_x}$**.** As we move further from $\boldsymbol{\mu_x}$ $\hat{y}_{\infty MM}(\boldsymbol{x})$ exhibits less variability than $\hat{y}_{LS}(\boldsymbol{x})$ under non-normality because of more efficient slope estimates. These observations apply to the intercept estimators themselves by noting that intercept is interpreted as $E(y|\boldsymbol{x} = \boldsymbol{0})$.

With this in mind we want to comment on a common practice of centering predictor variables before fitting a regression. Centering indeed facilitates interpretation and improves precision of the intercept estimates and in this case the LS and ∞MM intercepts have the same variance. While centering helps with the intercept estimates the conditional mean estimates $\hat{y}_{\infty MM}(\boldsymbol{x})$ could be better than $\hat{y}_{LS}(\boldsymbol{x})$ as $\boldsymbol{x}$ moves further from its mean. This will be illustrated with simulations in Section 4.5.

4.4.2 Bias and Mean-Squared-Error Efficiency

In this section we consider a regression with an error term following a skewed-t distribution $F_\epsilon \sim ST(\xi, \omega^2, \lambda, \nu)$ so according to (4.4) the true intercept is equal to

$$\alpha_0 = E\epsilon = \xi + \omega\delta b_\nu \tag{4.17}$$

with δ and b_ν defined in (4.2) and (4.3).The term $\omega\delta b_\nu$ reflects the skewness of the error distribution. We are interested in comparing the following intercept estimators: $\hat{\alpha}_{LS}$, $\hat{\alpha}_{MM}$, $\hat{\alpha}_{cMM}$, and the skewed-t

maximum likelihood estimate $\hat{\alpha}_{ST}$. The latter is computed by plugging into (4.17) the MLEs of the corresponding parameters.

Efficiency

Since now MM and c-MM intercept estimators can be biased, we define relative efficiency of $\hat{\alpha}_1$ to $\hat{\alpha}_2$ as the ratio of the mean squared errors:

$$REFF(\hat{\alpha}_1, \hat{\alpha}_2; n) = \frac{MSE_{\hat{\alpha}_2}(n)}{MSE_{\hat{\alpha}_1}(n)} \tag{4.18}$$

where $MSE_{\hat{\alpha}}(n)$ denotes the mean squared error of an estimator $\hat{\alpha}$ at sample size n. In what follows we compute the efficiencies of the LS, MM and c-MM estimators relative to a ST MLE.

The asymptotic normality results for the various estimates suggest that a finite sample estimate $\hat{\alpha}_E$ is approximately normal $N(\alpha_E, V_E/n)$ and, therefore,

$$MSE_{\hat{\alpha}_E}(n) = E(\hat{\alpha}_E - \alpha_0)^2 \approx Bias_E^2 + V_E/n \tag{4.19}$$

with $Bias_E = \alpha_E - \alpha_0$. The extent to which these approximations work in finite samples will be checked by Monte Carlo simulations in Section 4.4.3.

In Figure 41 we plot $REFF(\hat{\alpha}_{LS}, \hat{\alpha}_{ST}; n)$, $REFF(\hat{\alpha}_{MM}, \hat{\alpha}_{ST}; n)$ and $REFF(\hat{\alpha}_{cMM}, \hat{\alpha}_{ST}; n)$ with the mean squared errors computed from the asymptotic approximation (4.19) as follows:

- The ST MLE $\hat{\alpha}_{ST}$ is a consistent estimator of $\alpha_0 = E\epsilon$ and is asymptotically normal with a complicated variance expression V_{ST} given in Appendix A4.
- When $var(\epsilon) < \infty$ the LS intercept $\hat{\alpha}_{LS}$ is also a consistent estimator of $\alpha_0 = E\epsilon$ with the asymptotic variance

$$V_{LS} = var(\epsilon)\left(1 + \boldsymbol{\mu}_{\boldsymbol{x}}' \boldsymbol{C}_{\boldsymbol{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}}\right) \tag{4.20}$$

Note that since $Bias_{LS} = Bias_{ST} = 0$ then $REFF(\hat{\alpha}_{LS}, \hat{\alpha}_{ST}; n)$ computed via asymptotic approximation does not depend on sample size n and represents asymptotic efficiency.

- The asymptotic values of the MM and c-MM intercept estimators were obtained numerically as $\text{argmin}_t E_{F_\epsilon}\rho((\xi - t)/\sigma)$ with the corresponding loss function ρ, that is as solutions to (1.16) and (4.12). In particular, we used bisquare psi functions with c=4.68 (95% normal distribution efficiency) and with c=15 for the MM and c-MM estimators respectively. The asymptotic variances of the $\hat{\alpha}_{MM}$ and $\hat{\alpha}_{cMM}$ were computed from the equations (4.9) and (4.13) in Section 4.4.1. Since in the presence of skewness $Bias_{MM} = \alpha_{MM} - \alpha_0$ and $Bias_{cMM} = \alpha_{cMM} - \alpha_0$ are non-zero then the mean-squared-error relative efficiencies $REFF(\hat{\alpha}_{MM}, \hat{\alpha}_{ST}; n)$ and $REFF(\hat{\alpha}_{cMM}, \hat{\alpha}_{ST}; n)$ will depend on sample size n.

The asymptotic variances of the intercept estimators depend on $\boldsymbol{\mu}_x' \boldsymbol{C}_x^{-1} \boldsymbol{\mu}_x$ and, unfortunately, this term does not cancel when taking the ratio of the mean squared errors. In Figure 41 we assume that all predictors are centered at $\mathbf{0}$ so that $\boldsymbol{\mu}_x' \boldsymbol{C}_x^{-1} \boldsymbol{\mu}_x = 0$. Note that in this case $MSE_{\hat{\alpha}_{\infty MM}} = MSE_{\hat{\alpha}_{LS}}$ and, therefore, we do not consider ∞MM estimator separately in this section.

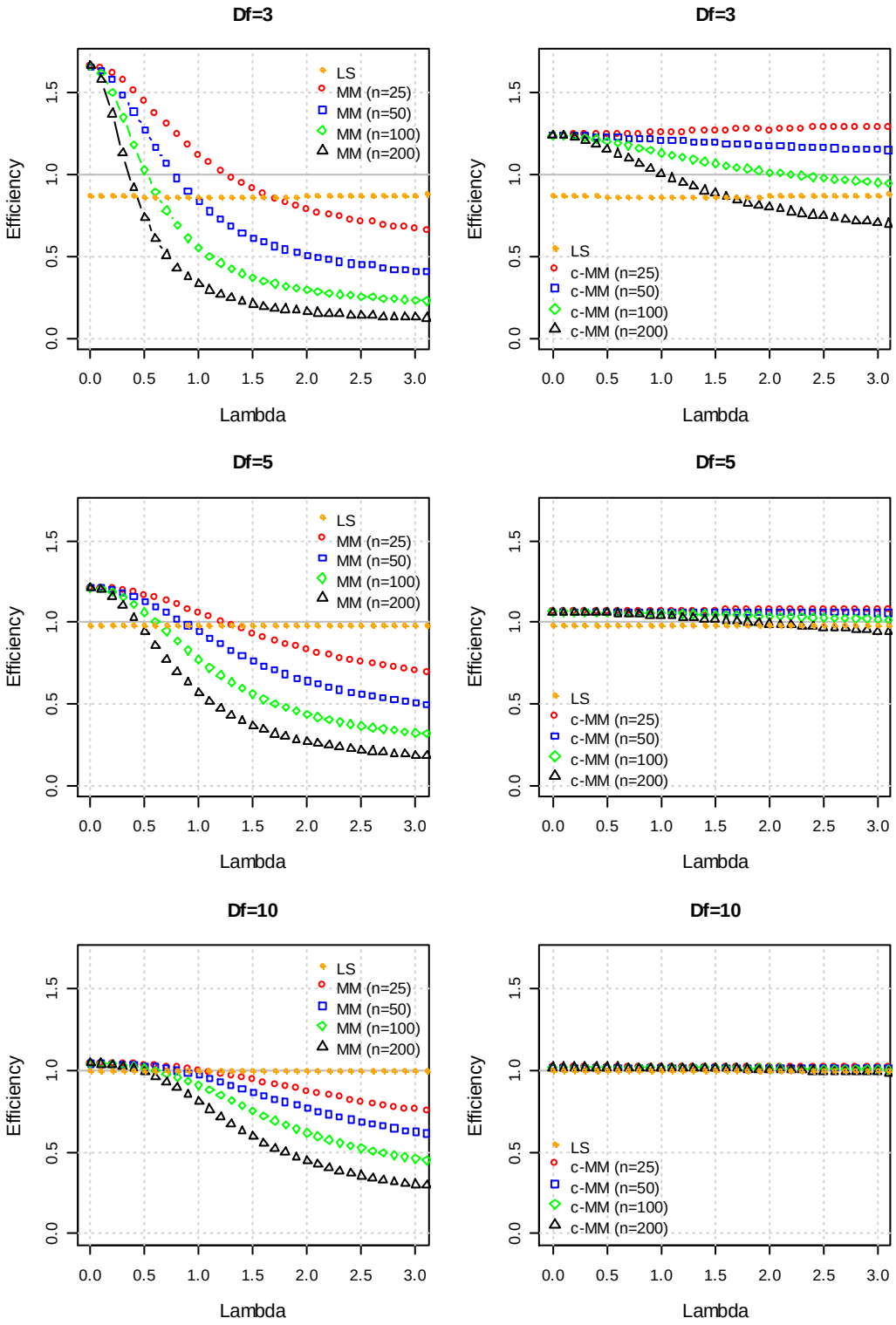

Figure 41. Efficiency of the LS, MM (left) and c-MM (right) intercept estimators relative to the skewed-t MLE of α when $\mu_x = 0$ versus the absolute value of lambda (shown simply as lambda). Bisquare loss functions with c=4.68 (95% normal distribution efficiency) and with c=15 were used for the MM and c-MM estimators respectively.

$\underline{\lambda = 0}$

We first discuss the case of $\lambda = 0$ when asymptotic bias of all considered intercept estimators, including MM and c-MM, is zero and relative efficiency is just a ratio of the asymptotic variances which does not depend on sample size. Unlike the ST MLE of the slopes the ST MLE of the intercept does not attain the Cramer Rao lower bound thereby allowing relative efficiencies to take values larger than one. Indeed, MM intercept can have much smaller variance than that of the ST MLE, more so the lower the degrees of freedom. For example, for the bisquare MM intercept estimator with c=4.68 (95% normal distribution efficiency) the relative efficiency $REFF(\hat{\alpha}_{MM}, \hat{\alpha}_{ST})$ is 166%, 121% and 104% at three, five and ten degrees of freedom respectively. The relative efficiency of the LS intercept $REFF(\hat{\alpha}_{LS}, \hat{\alpha}_{ST})$ even though less than 100% is still fairly large and is equal to 87%, 98.6% and 99.9% at three, five and ten degrees of freedom respectively.

Note that the absolute asymptotic efficiencies of the MM and LS intercept estimators under symmetric t-distribution are smaller than what is seen in Figure 41 at $\lambda = 0$. In particular, under a symmetric t-distribution the asymptotic variance expressions of the LS, MM and T MLE of the intercept have the same form of $\kappa(1 + \mu_x' C_x^{-1} \mu_x)$ so that absolute efficiencies of the LS and MM estimators are simply κ_T/κ_{LS} and κ_T/κ_{MM}. The constants κ are the same as for the slopes and are as follows: $\kappa_{LS} = var(\epsilon) = \frac{\nu}{\nu-2}\omega^2$ (the well-known expression for the variance of a t-distribution), $\kappa_{MM} = \sigma^2\tau$ (as before in eq. (4.6)) and $\kappa_T = \frac{\nu+3}{\nu+1}\omega^2$. That is, under a symmetric t-distribution the absolute efficiencies of the LS and MM intercept estimators are the same as that of the slopes and, therefore, can be found in Figure 39.

$\underline{\lambda \neq 0}$

Next we discuss the results in Figure 41 for $\lambda \neq 0$. Surprisingly, as λ grows we observe virtually no decline in efficiency of the LS intercept relative to ST MLE, and recall from our previous discussion that this efficiency is rather high, e.g. 87% for df=3 and over 99% for df=10. The MM and c-MM intercept estimators are now biased and depending on degree of skewness and sample size can have MSE that is

either smaller or larger than that of the ST MLE and LS. The relative mean-squared-error efficiency of the MM intercept is shown in the left panels of Figure 41. It decreases with $|\lambda|$ at a faster rate the larger the sample size and the smaller the degrees of freedom. For example, at sample size 100

- $REFF(\hat{\alpha}_{MM}, \hat{\alpha}_{ST})$ is over 100% when in absolute value lambda is under 0.6-0.7 depending on the degrees of freedom
- $REFF(\hat{\alpha}_{MM}, \hat{\alpha}_{ST})$ is under 50% when in absolute value lambda is over 1.2 for df=3, over 1.8 for df=5 and over 2.7 for df=10.

The relative mean-squared-error efficiency of the c-MM intercept is shown in the right column in Figure 41. Efficiency of the c-MM estimator is similar to (in the case of 10 d.o.f.) and less than (for 3 and 5 d.o.f.) that of the MM for mild skewness (lambda values closer to zero). But the efficiency of the c-MM estimator can be substantially higher than that of the MM for larger skewness (lambda values further from zero). In particularly the c-MM estimator has close to 100% efficiency for all values of lambda and sample sizes shown for the case of 5 and 10 d.o.f. and has mostly larger than 100% efficiency for 3 d.o.f. and sample sizes not larger than 100.

Bias

We now investigate the magnitude of the biases of the MM and c-MM intercept estimators and start by describing the scale and location equivariance properties of the MM and c-MM intercept estimators under skew-t errors. We write $\sigma(\xi, \omega, \lambda, \nu)$, $\alpha_{MM}(\xi, \omega, \lambda, \nu)$ and $Bias_{MM}(\xi, \omega, \lambda, \nu)$ to emphasize that the corresponding asymptotic values are computed when $F_\epsilon \sim ST(\xi, \omega^2, \lambda, \nu)$.

Recall from Chapter 1 that

$$\alpha_{MM} = \underset{t}{argmin}\, E_{F_\epsilon} \rho_2 \left(\frac{\epsilon - t}{\sigma} \right)$$

$$\alpha_S = \underset{t}{argmin}\, s_0(t) = \underset{t}{argmin}\, E_{F_\epsilon} \rho_1 \left(\frac{\epsilon - t}{\sigma} \right)$$

where $\sigma = min\ s_1(t)$ with $s_1(t)$ defined as a solution to $E_{F_\epsilon}\rho_1((\epsilon - t)/s(t)) = b$. By noting that $\epsilon = \xi + \omega\epsilon_{0,1}$ with $\epsilon \sim ST(\xi, \omega^2, \lambda, \nu)$ and $\epsilon_{0,1} \sim ST(0, 1, \lambda, \nu)$ it is easy to see that

$$\sigma(\xi, \omega, \lambda, \nu) = \omega\sigma(0, 1, \lambda, \nu)$$

$$\alpha_{MM}(\xi, \omega, \lambda, \nu) = \xi + \omega\alpha_{MM}(0, 1, \lambda, \nu) \quad (4.21)$$

Absolute Bias

Recalling that the true intercept is $\alpha_0 = \xi + \omega\delta b_\nu$ it follows from (4.21) that the bias of the MM (and c-MM) intercept does not depend on the location parameter ξ and is scale equivariant, i.e.

$$Bias_{MM}(\xi, \omega, \lambda, \nu) \equiv \alpha_{MM}(\xi, \omega, \lambda, \nu) - \alpha_0 = \omega[\alpha_{MM}(0, 1, \lambda, \nu) - \delta b_\nu] = \omega Bias_{MM}(0,1, \lambda, \nu)$$

$Bias_{MM}(0,1, \lambda, \nu)$ as a function of λ is shown in Figure 42 (left) for degrees of freedom ν equal to 3, 5 and 10 (different panels) and several values of the tuning parameter c (different symbols within a panel). Bias in Figure 42 is negative because for positive lambda (skewness) MM and c-MM intercepts underestimate true mean. Naturally, the larger the value of a tuning constant c the smaller the bias. As skewness grows the bias increases, at a slower rate the larger the value of a tuning constant c.

Relative Bias

Instead of looking at the absolute bias it is, perhaps, more meaningful to look at a ratio

$$\frac{Bias_{MM}(\xi, \omega, \lambda, \nu)}{\omega\delta b_\nu} = \frac{Bias_{MM}(0,1, \lambda, \nu)}{\delta b_\nu}$$

which can be interpreted as a proportion of a non-zero skewness component of the mean that is not absorbed by MM intercept. The ratio conveniently does not depend on location and scale parameters and is shown in Figure 42 (right). The ratio increases (in absolute value) with lambda, but at a surprisingly slow rate. For example, at five degrees of freedom the MM bisquare intercept estimator at 95% normal distribution efficiency (green triangles) does not capture about 10% of the skewness component of the

mean when $\lambda = 0.5$ and about 15% when $\lambda = 3$. For a c-MM bisquare intercept with c=15 at five degrees of freedom the proportion goes down to just 1-2%. Overall, percent bias of a c-MM estimate is almost negligible until 3 degrees of freedom, where it is still pretty small, e.g. 7.5% at df=3 and $\lambda = 3$.

It should be noted that the value of the bias relative to the true intercept

$$\frac{Bias_{MM}(\xi, \omega, \lambda, \nu)}{\alpha_0} = \frac{Bias_{MM}(0,1, \lambda, \nu)}{\xi + \omega\delta b_\nu}$$

tends to 0 as $|\xi| \to \infty$. In other words, when location parameter $|\xi|$ is large the bias of the MM estimator would often be rather small compared to the true intercept to warrant any practical concerns. This is illustrated in Section 4.6 with the Australian Institute of Sports example.

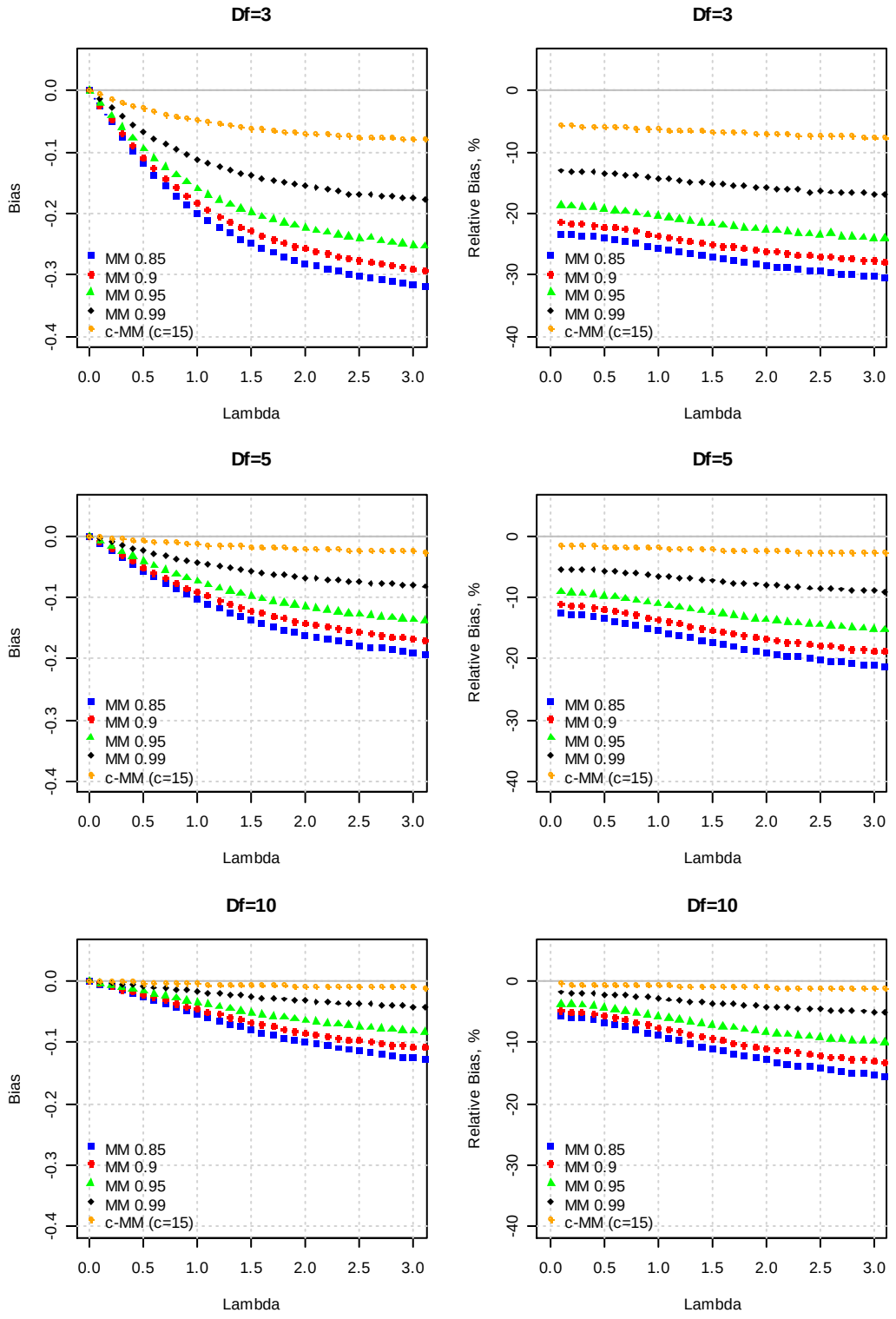

Figure 42. Left column: Asymptotic bias of the MM and c-MM intercepts. Right column: Asymptotic bias of the MM and c-MM intercepts relative to error skewness component $\omega\delta b_v$. MM are based on bisquare loss function at 85% (c=3.44), 90% (c=3.88), 95% (c=4.68) and 99% (c=7.04) normal distribution efficiencies and c-MM is based on bisquare loss function with c=15.

Bias Contribution to Mean Squared Error

Recall from (4.9) that asymptotic variance of an MM intercept estimator is equal to

$$\sigma^2\tau\left(1+\boldsymbol{\mu}_{\boldsymbol{x}}'\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{x}}\right)+\sigma^2\vartheta_{MM}$$

From scale and location equivariance properties (4.21) it follows that f_{MM}, a_{MM}, e_{MM}, g_{MM} and d_S do not depend on the location and scale parameters and, consequently, neither do τ and ϑ_{MM}. Hence,

$$V_{MM}(\xi,\omega,\lambda,\nu)=\omega^2V_{MM}(0,1,\lambda,\nu)$$

and

$$MSE_{MM}(\xi,\omega,\lambda,\nu)=\omega^2MSE_{MM}(0,1,\lambda,\nu)$$

Therefore, a ratio $Bias_{MM}^2(\xi,\omega,\lambda,\nu)/MSE_{MM}(\xi,\omega,\lambda,\nu)$ does not depend on ξ and ω. These ratios are displayed in Figure 43 for the MM (left) and c-MM (right) intercept estimators as a function of λ for degrees of freedom ν equal to 3, 5 and 10 (different panels) and four sample sizes n=25, 50, 100 and 200 (different symbols within a panel). When this ratio is small, say under 20%, then one may argue that bias is dominated by the standard errors of an estimate and therefore is less of a concern. For a c-MM intercept this is the case for all considered values of lambda and sample sizes for 5 and 10 degrees of freedom, and for lambda under 2.5, 1.5 and 1 for sample sizes 50, 100 and 200 respectively and 3 degrees of freedom. Since c-MM intercept has smaller bias and larger variance than that of an MM intercept it is not surprising that percentages for a c-MM intercept in Figure 43 are much smaller than that for an MM intercept.

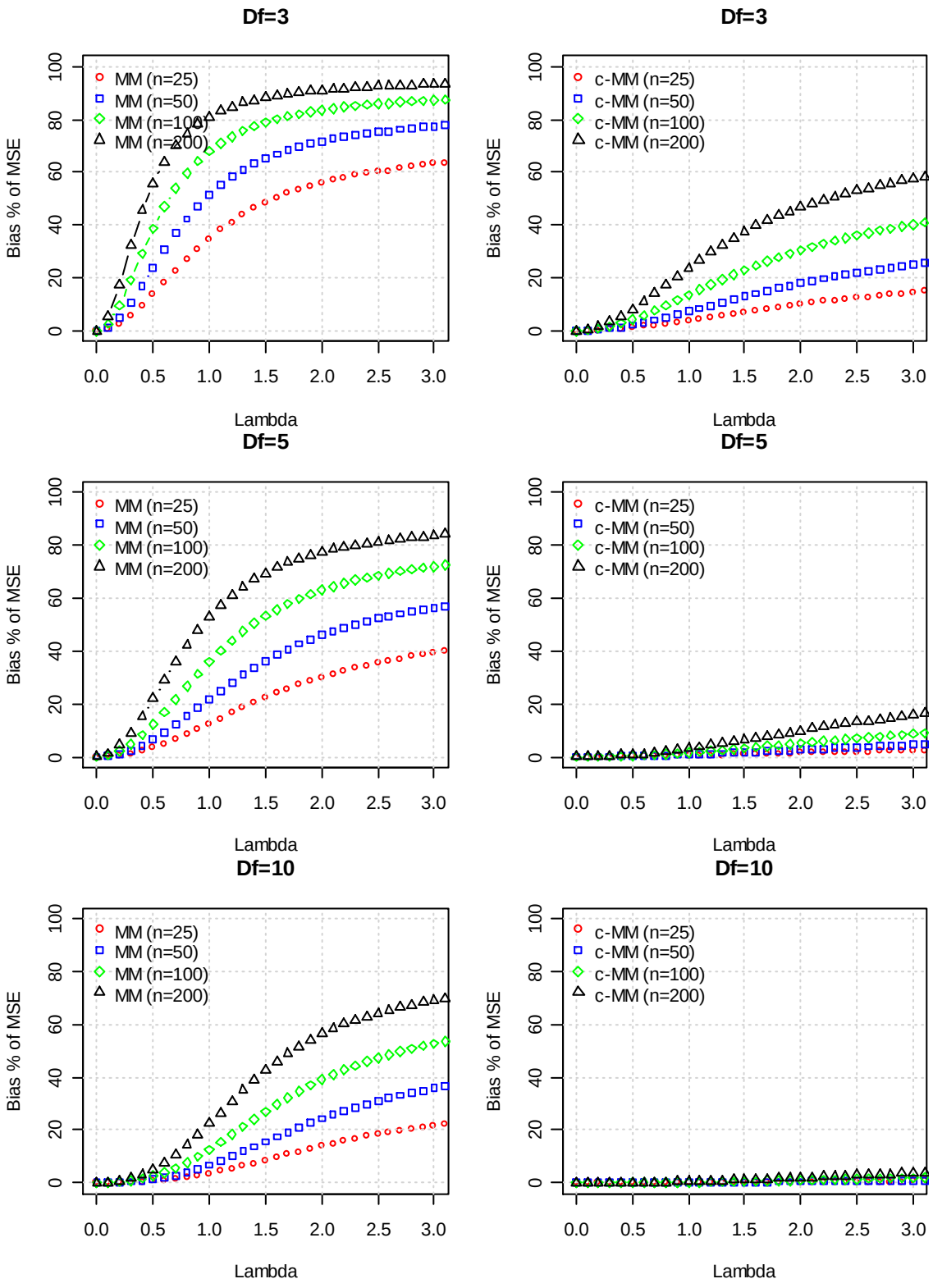

Figure 43. $Bias^2/MSE(n)$ for MM (left) and c-MM (right) intercept estimators when $\mu_x = 0$. Bisquare loss functions with c=4.68 (95% normal distribution efficiency) and with c=15 were used for the MM and c-MM estimators respectively.

At the end we point out a straightforward way of correcting the intercept bias when skewness parameter λ (and, consequently, δ) and degrees of freedom ν are known. From (4.21) it follows that

$$\tilde{\alpha}_{MM} = \hat{\alpha}_{MM} + \frac{\hat{\sigma}_{MM}}{\sigma(0,1,\lambda,\nu)} \left(\delta b_\nu - \alpha_{MM}(0,1,\lambda,\nu)\right)$$

is a consistent estimator of α_0. Under mild or moderate skewness, say λ no larger than 2, the resulting MM slopes and the above $\tilde{\alpha}_{MM}$ will be unbiased and will have fairly large efficiency (see Figure 39). For highly skewed data, however, when efficiency of an MM estimator is low one may consider a TML estimator due to Marazzi and Yohai (2004) which we briefly described in the introduction Section 4.1.

4.4.3 Monte Carlo Simulations

Mean-squared-error efficiencies of the LS, MM and c-MM regression intercept estimators in finite samples may differ drastically from the asymptotic efficiencies displayed in Figure 41. To see whether this is the case we perform a simulation study. The simulation setup is the same as that for the slopes and is described in Section 4.3.2.

Table 15 displays the finite sample Monte Carlo variance of the intercept estimators from the simulated data sets multiplied by the corresponding sample size n for sample sizes 25, 50, 100 and 200, along with the asymptotic variance (marked n=Inf). The datasets with a nearly zero ST scale MLE $\hat{\omega}_{ST}$ (see Table 9) and with ST degrees of freedom MLE $\hat{\nu}_{ST}$ less than one (see Table 10) were removed before calculating summaries in Table 15. Due to bias of the MM and c-MM intercept estimators Table 15 does not provide a fair comparison of the estimators and is primarily included for verification of our algebra and R code.

Table 15. Variance of the intercept estimators in a simple linear regression: Monte Carlo simulations vs asymptotic approximation. MM is bisquare at 95% normal distribution efficiency. c-MM is bisquare with c=15. Bold highlights cases where ST MLE breaks down as indicated by exceptionally large variance.

		Skewness Parameter (λ)											
		0.3				1				3			
DF	n	MM	cMM	LS	ST	MM	cMM	LS	ST	MM	cMM	LS	ST
3	25	1.69	2.14	2.86	**799.74**	1.36	1.72	2.36	**36.34**	1.02	1.26	1.86	**139.82**
	50	1.56	2.01	2.84	**5.06**	1.31	1.69	2.48	**3.13**	0.97	1.22	1.88	**3.22**
	100	1.55	2.04	2.94	2.67	1.26	1.67	2.50	2.29	0.95	1.18	1.83	1.88
	200	1.54	2.08	2.92	2.64	1.20	1.56	2.31	2.09	0.93	1.14	1.92	1.70
	Inf	1.51	2.02	2.90	2.52	1.20	1.58	2.39	2.06	0.92	1.15	1.91	1.67
5	25	1.42	1.55	1.67	**9.82**	1.06	1.15	1.26	**12.27**	0.75	0.79	0.88	**583.10**
	50	1.32	1.48	1.60	1.66	1.02	1.12	1.23	1.26	0.74	0.76	0.86	0.91
	100	1.32	1.49	1.61	1.60	1.02	1.14	1.25	1.24	0.75	0.78	0.87	0.87
	200	1.31	1.47	1.58	1.57	0.97	1.09	1.19	1.18	0.74	0.78	0.87	0.86
	Inf	1.29	1.47	1.59	1.57	0.99	1.11	1.22	1.19	0.72	0.76	0.86	0.84
10	25	1.24	1.23	1.25	**1.97**	0.89	0.89	0.90	**1.67**	0.63	0.60	0.61	**11.12**
	50	1.15	1.18	1.20	1.20	0.87	0.88	0.90	0.91	0.61	0.58	0.59	0.60
	100	1.14	1.16	1.18	1.18	0.85	0.87	0.89	0.89	0.60	0.57	0.59	0.59
	200	1.15	1.16	1.18	1.18	0.84	0.85	0.87	0.86	0.59	0.55	0.57	0.57
	Inf	1.14	1.17	1.19	1.19	0.84	0.86	0.88	0.87	0.59	0.56	0.58	0.57

From Table 15 we make the following observations:

- Asymptotic variance tends to underestimate finite sample variance, but as sample size increases the approximation gets better, as expected.
- The ST intercept MLE breaks at sample size n=25 as indicated by huge variances (highlighted in bold font). It should be noted that in addition to large variance the ST intercept estimates also exhibit noticeable bias in small samples, i.e. the averages of the ST intercept estimates from simulations were substantially different from the true intercept values (numbers are not shown, see Section 4.5 for partial illustration for the case of n=50). All this is likely a consequence of the problem with the ST MLE of the skewness parameter λ described in section 4.3.2 (see Table 8).
- As skewness increases the variances of all estimators go down
- As degrees of freedom increase the variances of all estimators go down, as one expects.

- For df=3 the c-MM variance is larger than MM variance, but is still noticeably smaller than that of LS and ST MLE. For df=5 the relationship is the same, but is of a slightly smaller magnitude. For df=10 all estimators have comparable variances except for breakdown of the ST MLE at n=25.

To account for the bias-variance trade-off the estimators are compared using finite sample mean-squared-error efficiencies. For each ν, λ and n combination and each intercept estimator E we compute mean squared error from Nsim simulated datasets as follows

$$MSE_E(n;\nu,\lambda) = \frac{1}{Nsim}\sum_{j=1}^{Nsim}\left(\hat{\alpha}_{E,j}(n;\nu,\lambda) - \alpha_0(\nu,\lambda)\right)^2$$

where α_0 is the true value of the intercept under the corresponding skewed-t distribution of the errors. Consistent with Table 15, the datasets that resulted in nearly zero ST error scale estimates (see Table 9) and in ST degrees of freedom estimate less than one (see Table 10) were removed before calculating MSE so that Nsim above ranges from 9 838 to 10 000. We call the ratio

$$\frac{MSE_{ST}(n;\nu,\lambda)}{MSE_E\ (n;\nu,\lambda)}$$

the Monte Carlo finite sample MSE efficiency of the intercept estimator E under ST errors. Figure 44 displays these finite sample efficiencies for the OLS, MM (bisquare at 95% normal distribution efficiency) and c-MM (bisquare with c=15) intercept estimators in a simple linear regression with $\mu_x = 0$. The results for sample size n=25 are not included in Figure 44 due to breakdown of the ST MLE.

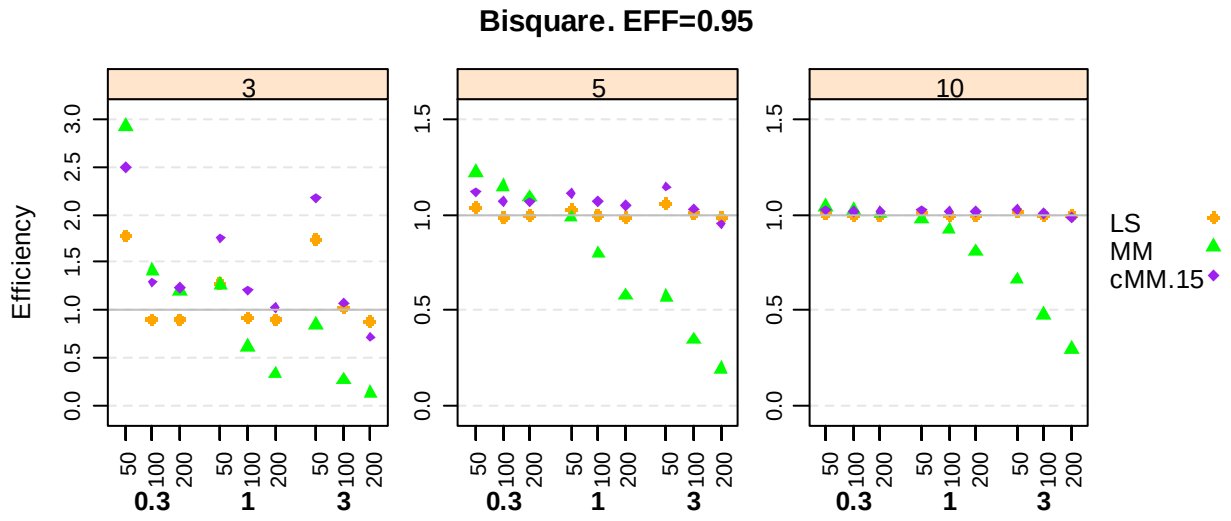

Figure 44. Monte Carlo finite sample efficiency of the LS, MM and cMM intercept estimators relative to ST MLE for df=3, 5, and 10, skewness parameter $\lambda = 0.3, 1,$ and 3, and sample sizes n= 50, 100 and 200.

From Figure 44 we make the following conclusions:

1) With regard to a simulated finite sample efficiency versus asymptotic approximation:
   a. For the most part the relative efficiency values in Figure 44 are reasonably close to those calculated from an asymptotic theory in Figure 41.
   b. Table 16 provides comparisons of the finite sample and asymptotic efficiencies for the cases of 3 and 5 degrees of freedom.
   c. The most noticeable exception is perhaps the case of df=3 (very heavy tails) and n=50 ('small' sample size), where Monte Carlo relative efficiencies in Figure 44 are much larger than their asymptotic counterparts in Figure 41 for all three estimators (MM, c-MM and LS). These cases are highlighted with bold in Table 16. The reason for such large discrepancies is an issue with ST MLE in small sample sizes, where it is known to be difficult to obtain good estimates of skewness and degrees of freedom (see Section 4.3.2).

Table 16. Comparison of the asymptotic and simulated mean-squared-error relative efficiencies of the intercept estimators for selected values of the degrees of freedom (ν), skewness parameter (λ) and sample size (n).

ν	λ	n	$REFF(\hat{\alpha}_{MM}, \hat{\alpha}_{ST})$		$REFF(\hat{\alpha}_{cMM}, \hat{\alpha}_{ST})$		$REFF(\hat{\alpha}_{LS}, \hat{\alpha}_{ST})$	
			Asym Figure 41	Sim Figure 44	Asym Figure 41	Sim Figure 44	Asym Figure 41	Sim Figure 44
3	0.3	50	**149%**	**293%**	**124%**	**249%**	**87%**	**179%**
	1	50	**83%**	**127%**	**121%**	**175%**	**86%**	**127%**
	3	50	**41%**	**85%**	**115%**	**217%**	**88%**	**173%**
	0.3	100	135%	141%	123%	129%	87%	91%
	1	100	55%	61%	113%	120%	86%	92%
	3	100	23%	27%	95%	107%	88%	103%
5	0.3	50	118%	122%	106%	112%	99%	104%
	1	50	94%	98%	106%	111%	98%	102%
	3	50	51%	57%	106%	115%	98%	106%
	0.3	100	115%	115%	106%	107%	99%	99%
	1	100	77%	80%	106%	107%	98%	99%
	3	100	32%	34%	102%	103%	98%	100%

2) With regard to c-MM versus MM versus LS efficiencies

   a. MM dominates c-MM only for small skewness ($\lambda = 0.3$) and df=3 and 5. At larger skewness ($\lambda = 1$ and 3) MM mean-squared-error efficiency is rather small and decreases dramatically with sample size and becomes quite inferior to c-MM. Such poor behavior of the MM intercept is due to larger bias under larger degree of skewness.

   b. c-MM dominates LS for df=3 and 5 and is comparable to LS for df=10 where efficiencies of both estimators are close to one for all three values of lambda.

   c. Efficiency of LS intercept is substantially larger than one for n=50 and df=3 and larger than one by a small amount for n=50 and df=5, in both cases for all three values of lambda. In other cases efficiency of LS stays above 90% and is very close to 100% for df=5 and 10.

Finally, we present simulation results for two symmetric distributions: Standard Student-t with 3 degrees of freedom and standard normal (marked DF=Inf). Table 17 displays Monte Carlo variances of the

intercept estimator multiplied by the corresponding sample size n, along with the asymptotic variance which is marked as n=Inf. Summaries in Table 17 are based on 10,000 simulation runs. There were 525 replicates with either nearly zero ST scale estimates or DF estimates less than one, and these were removed before calculating summaries.

Table 17. Variances of the intercept estimators under symmetric t-distribution errors.

DF	n	MM	MM.15	LS	T	ST
3	50	1.622	2.129	2.950	1.609	3.107
	100	1.602	2.141	3.028	1.559	2.736
	200	1.599	2.096	2.957	1.537	2.632
	500	1.627	2.125	2.939	1.550	2.637
	Inf	1.571	2.104	3.000	1.500	2.612
Inf	50	1.098	1.037	1.036	1.061	1.047
	100	1.054	1.004	1.003	1.011	1.006
	200	1.077	1.024	1.024	1.030	1.024
	500	1.038	0.981	0.980	0.983	0.980
	Inf	1.053	1.0005	1.000	1.000	1.000

For the t-distribution with 3 degrees of freedom MM has high efficiency while LS is very inefficient, both in finite samples and asymptotically. MM intercept efficiency relative to T MLE varied from 99% at n=25 to 95% (n=500 and asymptotically). Note nearly twice as large variance of the ST MLE as compared to both MM and T MLE. This 'extra variance' penalty for ST MLE of the intercept does not disappear as sample size increases with ST MLE asymptotic efficiency (relative to T MLE) being only 57%. This is in contrast to the slopes where asymptotically T and ST MLE have the same variance under symmetric distributions. The 'extra variance' penalty for ST MLE of the intercept, however, decreases as degrees of freedom increase so that at normal distribution ST MLE of the intercept is asymptotically fully efficient.

At normal distribution MM asymptotic efficiency is 95% by the choice of the tuning constant and in our simulations MM efficiency relative to LS varied from 94.3% to 95.2%. In small sample sizes T MLE has slightly larger variance compared to LS, which is a penalty for estimating DF. Similar to the slopes, the penalty disappears in large sample sizes with asymptotic variances of $\hat{\alpha}_{LS}$ and $\hat{\alpha}_{T}$ being equal.

4.5 Monte Carlo Simulations: A Closer Look

To further illustrate differences in bias and variability of the various intercept and slope estimators under skewed-t errors we take a closer look at the simulation results for two cases:

1) $\nu = 3, \lambda = 1$ and sample size n=50 in Figure 45 and Table 17,
2) $\nu = 3, \lambda = 1$ and sample size n=200 in Figure 46 and Table 18.

These are the same simulation runs as before where we generated $Nsim = 10\,000$ datasets from a simple linear regression model $y_i = \beta_0 x_i + \epsilon_i$ with $\epsilon_i \sim i.i.d. ST(0, 1, \lambda, \nu)$, $x_i \sim i.i.d. N(0, 2^2)$ and $\beta_0 = 1$. The true value of the intercept parameter α_0 is $E(\epsilon) \approx 0.780$. Figure 45 and Figure 46 display boxplots of six intercept estimators (LS, MM, c-MM, T MLE, SN MLE and ST MLE) on the left panel and of the corresponding five slope estimators (LS, MM, T MLE, SN MLE and ST MLE) on the right panel. Bisquare loss functions with c=4.68 (95% normal distribution efficiency) and with c=15 were used for the MM and c-MM estimates respectively.

4.5.1 n=50

Two datasets had skewed-t degrees of freedom estimate less than one and were removed before displaying results in Figure 45 for the intercepts, but not for the slopes.

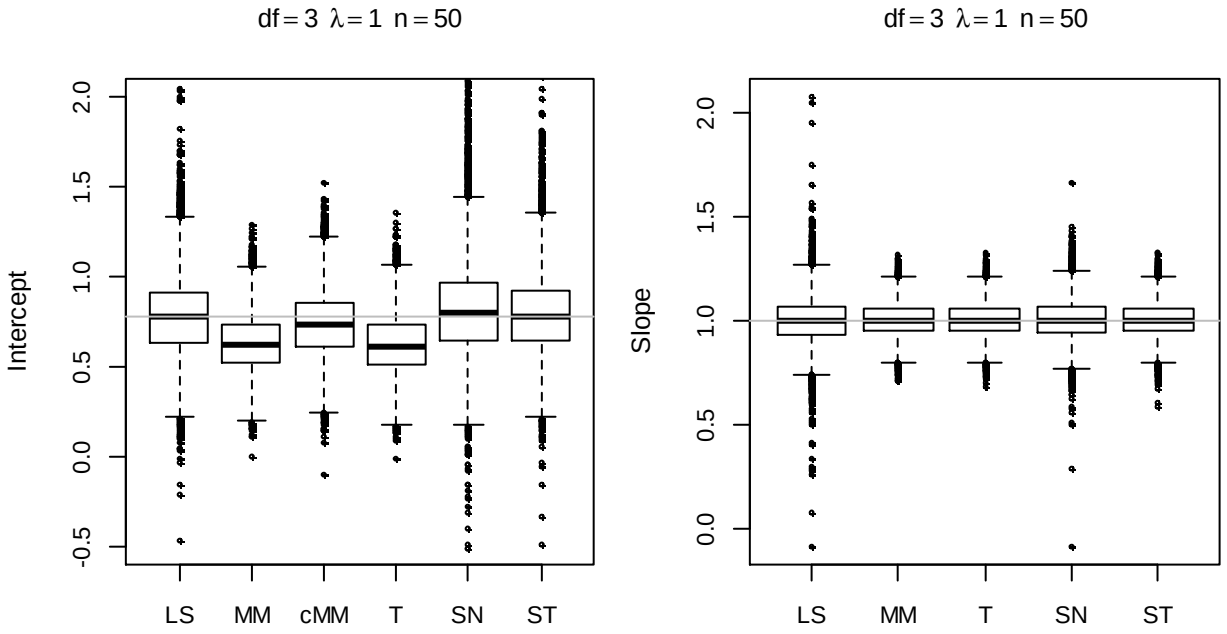

Figure 45. Sample size n=50. Boxplots of the intercept and slope estimates under skewed-t errors.

Table 18. Sample size n=50. Summary statistics associated with Figure 45. True intercept $\alpha_0 \approx 0.780$.

		LS	MM	c-MM (c=15)	∞MM	T	SN	ST
Intercept	Mean	0.783	0.627	0.734	0.782	0.619	0.835	0.797
	Bias	0.003	-0.153	-0.046	0.003	-0.160	0.055	0.017
	Std Dev	0.223	0.162	0.184	0.222	0.166	0.341	0.250
	MSE	0.050	0.050	0.036	0.049	0.053	0.119	0.063
	Rel Eff	127%	127%	175%	128%	118%	53%	100%
Slope	Mean	1.003	1.002			1.002	1.002	1.002
	Std Dev	0.113	0.078			0.079	0.096	0.080
	Rel Eff	50%	104%			103%	69%	100%
$\hat{y}(x=2)$	Mean	2.788	2.631	2.738	2.786	2.624	2.839	2.801
	Bias	0.008	-0.149	-0.042	0.007	-0.156	0.059	0.021
	Std Dev	0.317	0.227	0.243	0.273	0.230	0.394	0.299
	MSE	0.101	0.074	0.061	0.074	0.077	0.159	0.090
	Rel Eff	89%	122%	148%	121%	116%	57%	100%
$\hat{y}(x=4)$	Mean	4.793	4.635	4.742	4.790	4.628	4.843	4.805
	Bias	0.014	-0.145	-0.038	0.011	-0.152	0.063	0.025
	Std Dev	0.504	0.356	0.366	0.386	0.358	0.518	0.410
	MSE	0.254	0.148	0.135	0.149	0.151	0.272	0.168
	Rel Eff	66%	114%	125%	113%	112%	62%	100%

Intercept. For the MM estimates the squared bias constitutes 47% of the MSE while for the c-MM the squared bias is only 6% of the MSE. The numbers are very well in line with Figure 43. Indeed, from the boxplot one clearly sees that MM intercept bias is noticeable as compared to the variability of the estimates. Despite such large bias the MSE_{MM} of 0.05 is equal to MSE_{LS} and is smaller than MSE_{ST} of 0.063. Note that the sample mean for ST MLE is 0.797 which is somewhat larger than the true intercept α_0 of 0.78. The c-MM bias is barely noticeable on a boxplot and MSE_{cMM} of 0.036 is the smallest among all considered estimators. Since the mean of x is zero the variance of ∞MM intercept estimator is similar to that of LS.

Slope. The sample means for all estimators are reasonably close to one suggesting that all considered estimators are unbiased. MM and ST variances are comparable, while LS has very large variance, nearly twice as large as that of MM.

Conditional Expectation. The advantages of using a highly efficient MM slope estimates over highly inefficient LS slopes become apparent when we look at the fitted values $\hat{y}(x)$ one ($x = 2$) and two ($x = 4$) standard deviations away from the mean of the predictors. Efficiency of $\hat{y}_{\infty MM}$ (relative to $\hat{y}_{ST}$) is 121% at $x = 2$ and 114% at $x = 4$, while efficiency of $\hat{y}_{LS}$ is only 89% at $x = 2$ and 66% at $x = 4$. Even though biased, $\hat{y}_{cMM}$ has the smallest mean squared error among all considered estimators. Also note that efficiency of $\hat{y}_{MM}$ (relative to $\hat{y}_{ST}$) even though smaller than that of $\hat{y}_{cMM}$ is still larger than 100%.

For both intercept and slope T MLE performs similar to MM while SN MLE fails badly, especially for intercept.

4.5.2 n=200

All summaries in Table 19 and boxplots in Figure 46 are based on all 10 000 simulated datasets.

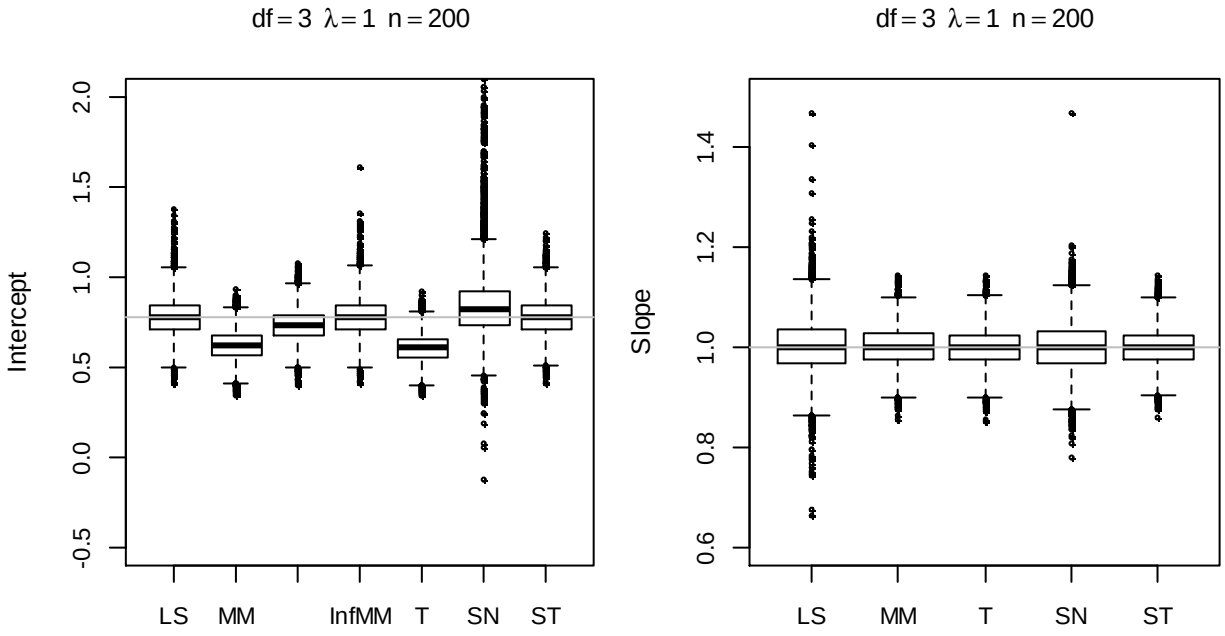

Figure 46. Sample size n=200. Boxplots of the intercept and slope estimates under skewed-t errors.

Table 19. Sample size n=200. Summary statistics associated with Figure 46. True intercept $\alpha_0 \approx 0.780$.

		LS	MM	c-MM (c=15)	∞MM	T	SN	ST
Intercept	Mean	0.779	0.621	0.731	0.779	0.607	0.851	0.781
	Bias	-0.001	-0.159	-0.049	-0.001	-0.172	0.071	0.001
	Std Dev	0.107	0.077	0.088	0.108	0.078	0.203	0.102
	MSE	0.012	0.031	0.010	0.012	0.036	0.046	0.010
	Rel Eff	91%	34%	103%	90%	29%	23%	100%
Slope	Mean	1.000	1.000			1.000	1.000	1.000
	Std Dev	0.055	0.038			0.037	0.048	0.036
	Rel Eff	43%	91%			94%	57%	100%
$\hat{y}(x=2)$	Mean	2.779	2.702	2.731	2.779	2.607	2.850	2.781
	Bias	-0.001	-0.078	-0.049	-0.001	-0.173	0.071	0.001
	Std Dev	0.155	0.112	0.116	0.131	0.107	0.225	0.125
	MSE	0.024	0.019	0.016	0.017	0.041	0.056	0.016
	Rel Eff	65%	83%	99%	91%	38%	28%	100%
$\hat{y}(x=4)$	Mean	4.779	4.621	4.731	4.779	4.607	4.850	4.780
	Bias	0.000	-0.159	-0.049	-0.001	-0.173	0.071	0.001
	Std Dev	0.246	0.169	0.174	0.184	0.167	0.280	0.176
	MSE	0.060	0.054	0.033	0.034	0.058	0.083	0.031
	Rel Eff	51%	58%	95%	91%	54%	37%	100%

Intercept. At sample size 200 the squared bias of the MM estimates constitutes 81% of the MSE and the squared bias of the c-MM is 23% of its' MSE. This is larger than what we saw for sample size 50 and the numbers are again very well in line with Figure 43. Bias in MM now starts to prevail leading to MSE_{MM} that is nearly three times larger than that of LS, c-MM and ST. c-MM is comparable to ST, and slightly better than LS. Note that the mean of the ST intercept estimates is 0.781 which is reasonably close to the true α_0 of 0.78, unlike sample size 50 where it was biased upward. ∞MM intercept performs similar to LS as a consequence of $\mu_x = 0$.

Slope. The sample means for all estimators are reasonably close to one suggesting that all considered estimators are unbiased. MM mean-squared-error efficiency relative to ST MLE is 91%, while LS efficiency is only 43%.

Conditional Expectation. The advantages of using a highly efficient MM slope estimates over highly inefficient LS slopes become apparent when we look at the fitted values $\hat{y}(x)$ one ($x = 2$) and two ($x = 4$) standard deviations away from the mean of the predictors. Efficiency of $\hat{y}_{\infty MM}$ (relative to $\hat{y}_{ST}$) is 91% at $x = 2$ and $x = 4$, while efficiency of $\hat{y}_{LS}$ is only 65% at $x = 2$ and 51% at $x = 4$. Domination of bias in an MM intercept at sample size 200 results in poor performance of $\hat{y}_{MM}$. Performance of $\hat{y}_{cMM}$, however, is reasonably good, with 99% and 95% efficiency (relative to $\hat{y}_{ST}$) at $x = 2$ and $x = 4$.

As was the case for sample size 50 in Figure 45, at sample size 200 T MLE performs similar to MM for both intercept and slope while SN MLE fails badly, especially for intercept.

4.6 Simulation Study for Mixture Models 4 and 5 from Chapter 2

MLEs based on heavy-tailed and possibly skewed distributions constitute a parametric approach to robust estimation. Hence, one may ask about efficiency of such MLEs when tails of the error distribution have a different structure than assumed by the parametric model. In this section we address this question by conducting a simulation study for an asymmetric normal mixture Model 4 of Chapter 2. It assumes that regression errors follow an asymmetric two-term normal mixture distribution

$(1-\gamma)N(0,1^2)+\gamma N(\mu,0.25^2)$.

One may also be interested in the bias robustness properties of the skewed-t MLE under bias-inducing mixture models (2). In this section we address this question by conducting a simulation study for a joint normal mixture Model 5 of Chapter 2, which is as follows

$\binom{x_1}{u_1},\ldots,\binom{x_n}{u_n}$ are i.i.d. $(1-\gamma)N\left(\binom{0}{0},\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)+\gamma N\left(\binom{2}{\mu},\begin{bmatrix}0.25^2 & 0\\ 0 & 0.25^2\end{bmatrix}\right)$

The four estimators considered in the study are LS, MM, T MLE and ST MLE. The simulation results are presented in a set of trellis plots where bias, variance and mean squared errors of the estimators are compared for contamination proportion γ equal to 0.02, 0.04 and 0.06, contamination location μ equal to 4, 7 and 14, and four sample sizes n equal to 50, 100, 200 and 500.

4.6.1 Slope

The simulation study in this section demonstrates that MM slope estimator, which is designed with the mixture models in mind, has the smallest bias and the smallest mean squared error among the four considered estimators. The T and ST MLE, even though inferior to MM, still offer a substantial improvement over the LS.

Model 4 (Asymmetric Normal Mixture Distribution Errors)

All four estimators are consistent for a regression slope. The observed finite sample bias in our simulations was close to zero for all estimators for all sample sizes so that comparison of the mean squared errors leads to the same results as comparison of the variances alone. Figure 47displays variance efficiency of the LS, T MLE and ST MLE of the slopes relative to MM. The choice of MM estimator as a reference point is motivated by the fact that MM estimator is the main focus of this dissertation.

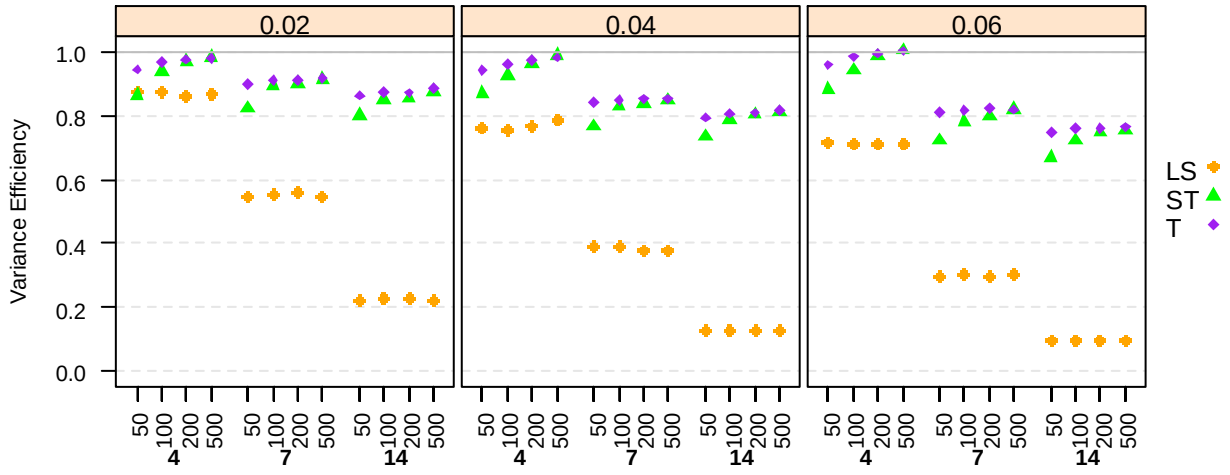

Figure 47. Efficiency of the slope estimators relative to MM under mixture Model 4 from Chapter 2.

From Figure 47 it is apparent that MM has the smallest variance among the four estimators as all relative efficiencies are below one. LS is highly inefficient, more so the larger the location μ of the contamination component. The T and ST MLE offer a substantial advantage over the LS even though they are still somewhat inferior to MM. The ST MLE has a similar variance as T MLE except for small sample size n=50 where it has noticeably larger variance.

The above summaries are based on 10 000 simulation runs with one run excluded due to numerical failure of ST MLE (ST scale estimate of the order 10^{-14}).

Model 5 (Bivariate Normal Mixture Distribution)

The true slope in this case is equal to one, a slope of the main mixture component. The biases of the estimators are plotted in Figure 48. Figure 49 and Figure 50 show variance and MSE efficiency of LS, ST MLE and T MLE relative to MM slope estimator. The summaries are based on 10 000 simulation runs with one run excluded due to numerical failure of ST MLE (ST scale estimate of the order 10^{-14}).

In Figure 48, MM has the smallest bias, which is non-zero for $\mu = 4$ and is almost zero for $\mu = 7$ and $\mu = 14$ by the virtue of the location of the contamination mixture relative to the rejection region of the MM estimator. MM estimator also has the smallest variance except for $\mu = 4$ where LS has the smallest

variance. The bias of the LS estimator is the largest. T and ST MLEs have substantially smaller bias as compared to LS, but noticeably larger bias as compared to MM.

All relative efficiencies in Figure 50 are less than one suggesting that MM estimator has the smallest mean squared error. As sample size increases the bias starts dominating and relative efficiencies go down.

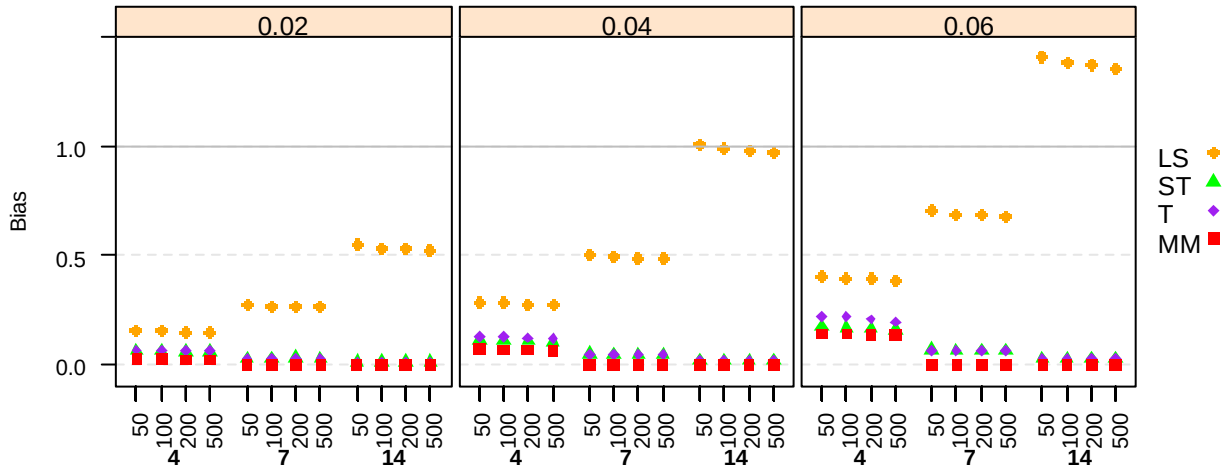

Figure 48. Bias of the slope estimators. Bivariate mixture Model 5 of Chapter 2.

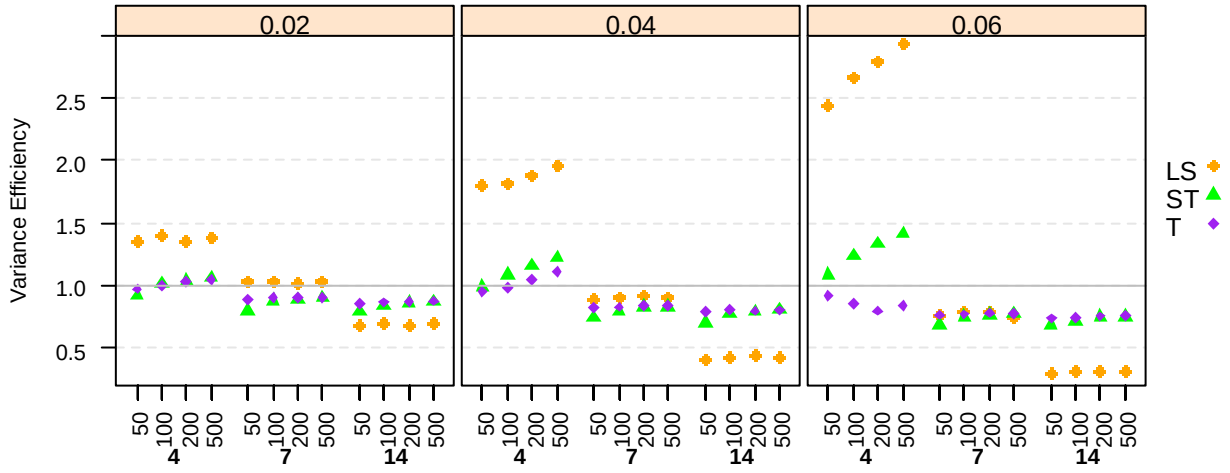

Figure 49. Variance efficiency of the slope estimators relative to MM. Bivariate mixture Model 5 of Chapter 2.

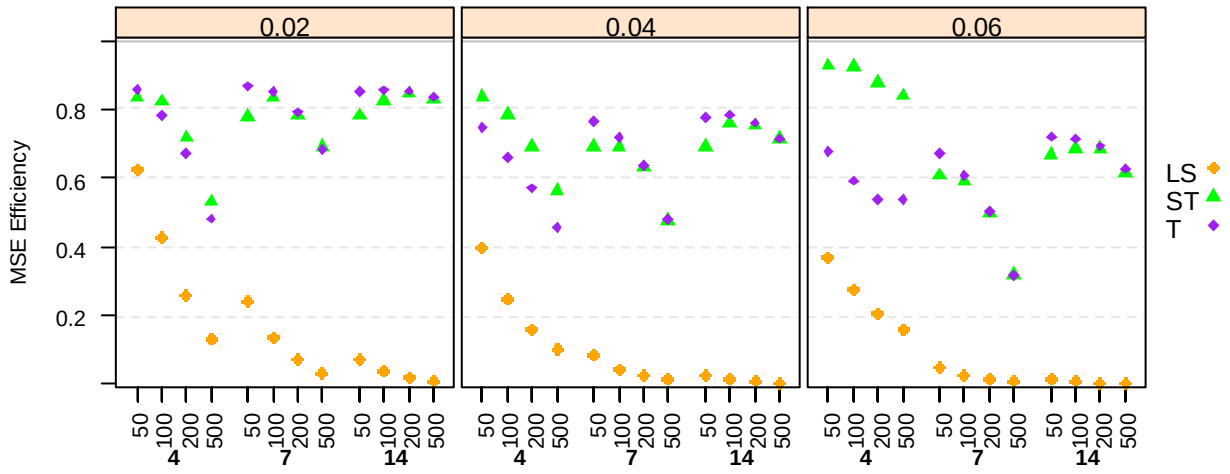

Figure 50. MSE efficiency of the slope estimators relative to MM. Bivariate mixture Model 5 from Chapter 2.

4.6.2 Intercept

The discussion in this section assumes that the true intercept parameter is defined by a mean of the main mixture component, which is zero for Models 4 and 5. In this context the MM intercept estimator has the smallest and often nearly zero bias. The ST and LS intercept estimators on the contrary exhibit large bias as they attempt to absorb asymmetry in the errors created by the contamination mixture component. Large bias leads to extremely low mean-squared-error efficiency of ST and LS relative to MM as seen in the simulation results that we present below. The symmetric T maximum likelihood intercept estimate has small bias and variance, both just a little larger than those of MM resulting in relative mean-squared-error efficiency of 70% to 90%.

Model 4 (Asymmetric Normal Mixture Distribution Errors)

The true intercept in this case is zero, a mean of the main mixture component. The biases of the estimators are plotted in Figure 51. Figure 52 and Figure 53 show variance and MSE efficiency of LS, ST MLE and T MLE relative to MM intercept estimator. All summaries are based on 10 000 simulation runs. There were 86 replicates with DF estimates less than one, and these were removed before calculating summaries. It is worth noting that all these cases occurred for sample size n=50 (except one case of n=100), once again highlighting a small sample size issue with ST maximum likelihood estimation.

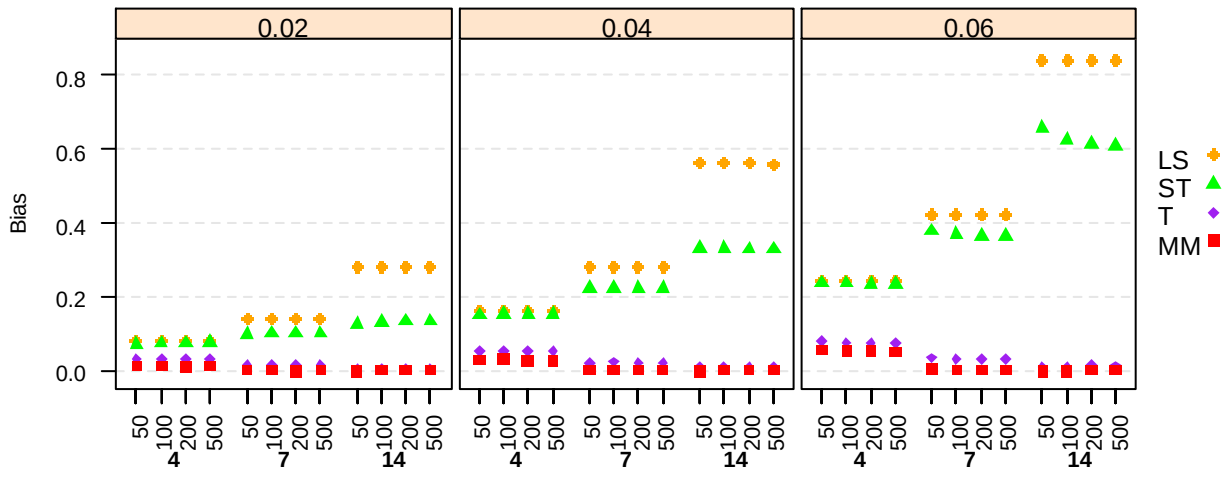

Figure 51. Bias of the intercept estimators under mixture Model 4 from Chapter 2.

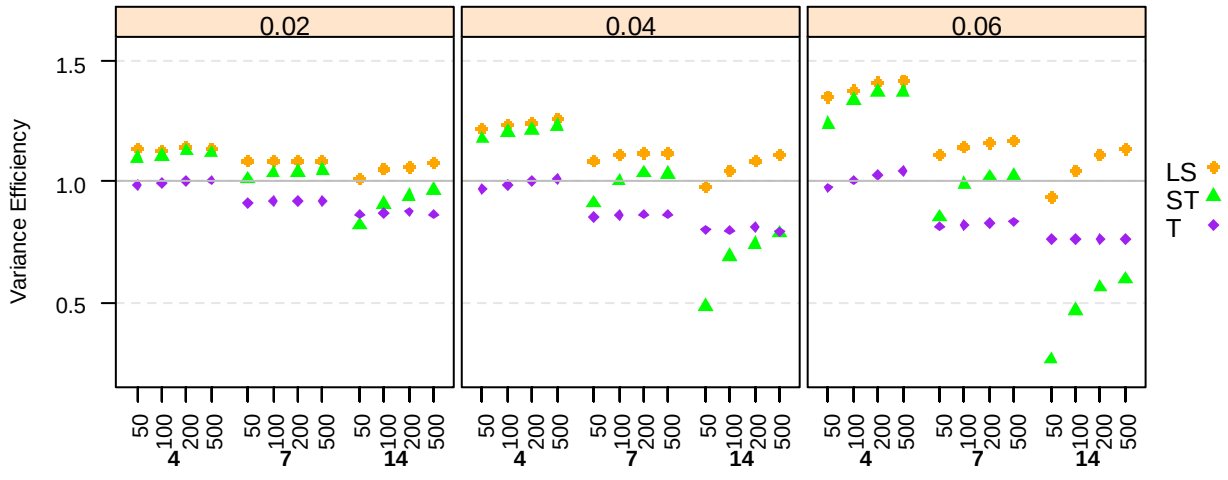

Figure 52. Variance efficiency of the intercept estimators relative to MM under mixture Model 4 from Chapter 2.

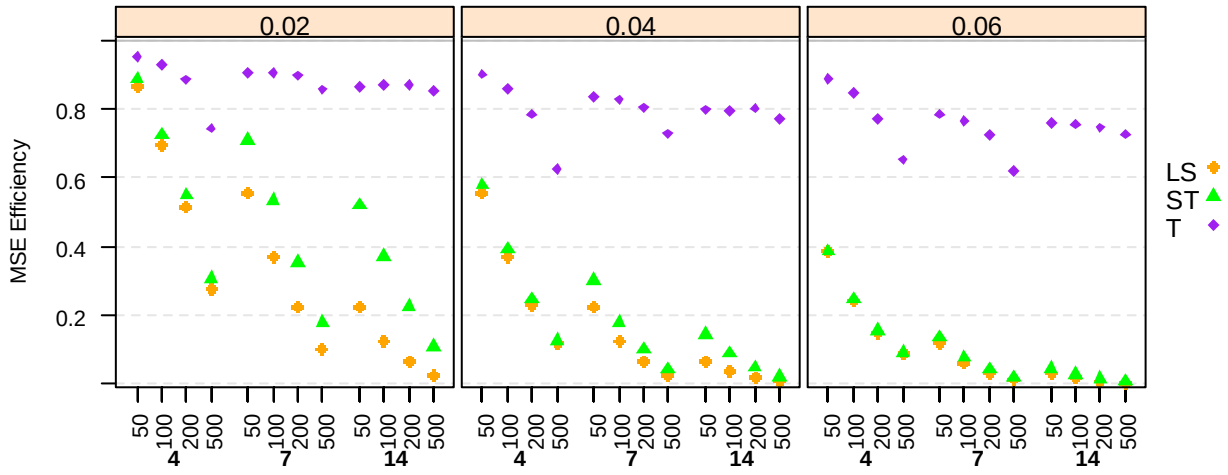

Figure 53. MSE efficiency of the intercept estimators relative to MM under mixture Model 4 from Chapter 2.

All relative efficiencies in Figure 53 are less than one meaning that MM estimator has the smallest mean squared error, a result that is mostly due to MM intercept having the smallest bias among all four

considered estimators. The MM bias is non-zero for $\mu = 4$ and is almost zero for $\mu = 7$ and $\mu = 14$ by the virtue of the location of the contamination mixture relative to the rejection region of the MM estimator.

The ST MLE and LS have similar mean-squared-error performance, which is very poor as a result of large bias in both estimators. LS is estimating a mean of the error term including contamination component, which is equal to $\gamma\mu$. The mean of the ST MLE is slightly less than $\gamma\mu$, but is substantially larger than zero (see Figure 51).

The mean squared error performance of T MLE is better than that of the LS and ST MLE as it has substantially smaller bias. The bias of the T MLE is nevertheless noticeably larger than that of the MM intercept estimator.

Model 5 (Bivariate Normal Mixture Distribution)

The true intercept in this case is equal to zero, a mean of the main mixture component. The biases of the estimators are plotted in Figure 54. Figure 55 and Figure 56 show variance and MSE efficiency of LS, ST MLE and T MLE relative to MM slope estimator respectively. The summaries are based on 10 000 simulation runs. There were 27 replicates with either nearly zero ST scale estimate (ST scale estimate was either below 10^{-13}, or larger than 0.03) or DF estimates below one, and these were removed before calculating summaries.

All relative efficiencies in Figure 56 are less than one meaning that MM estimator has the smallest mean squared error, a result that is mostly due to MM intercept having the smallest bias among all four considered estimators (see Figure 54). The mean squared error performance of the LS and ST MLE is similar as they both have large biases. The performance of T MLE is better than that of the LS and ST MLE as it has substantially smaller bias. The bias of the T MLE is nevertheless noticeably larger than that of the MM intercept estimator.

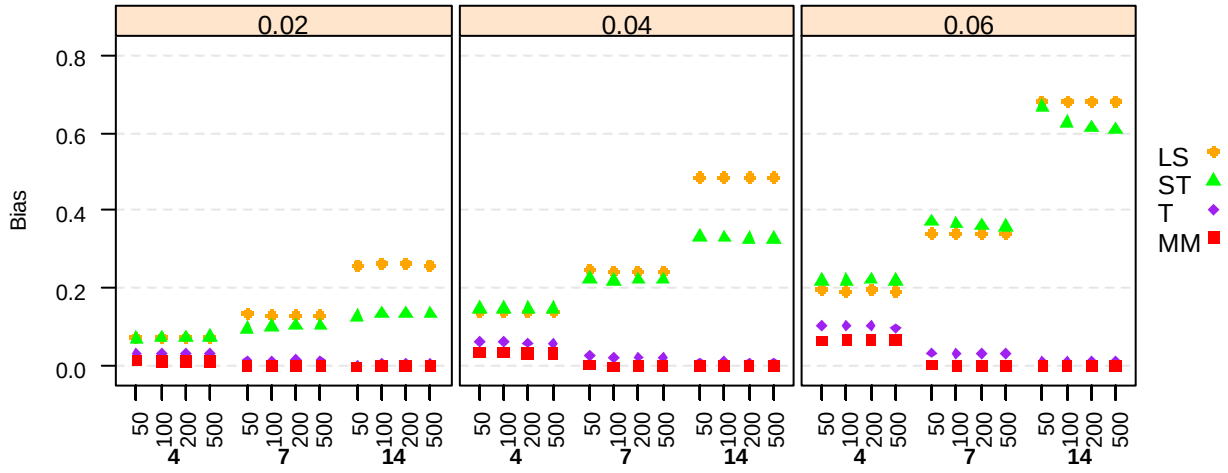

Figure 54. Bias of the intercept estimators. Bivariate mixture Model 5 of Chapter 2.

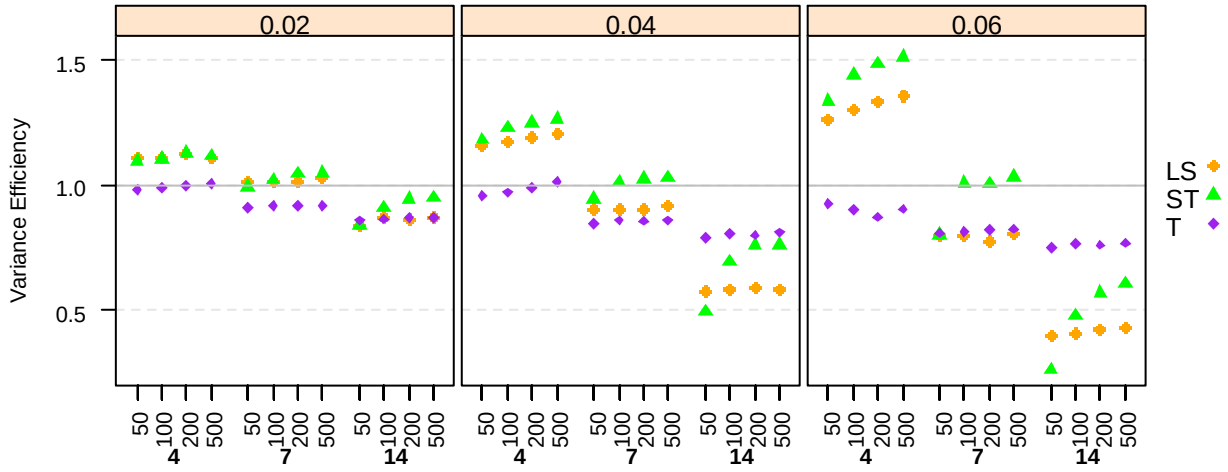

Figure 55. Variance efficiency of the intercept estimators relative to MM. Bivariate mixture Model 5 from Chapter 2.

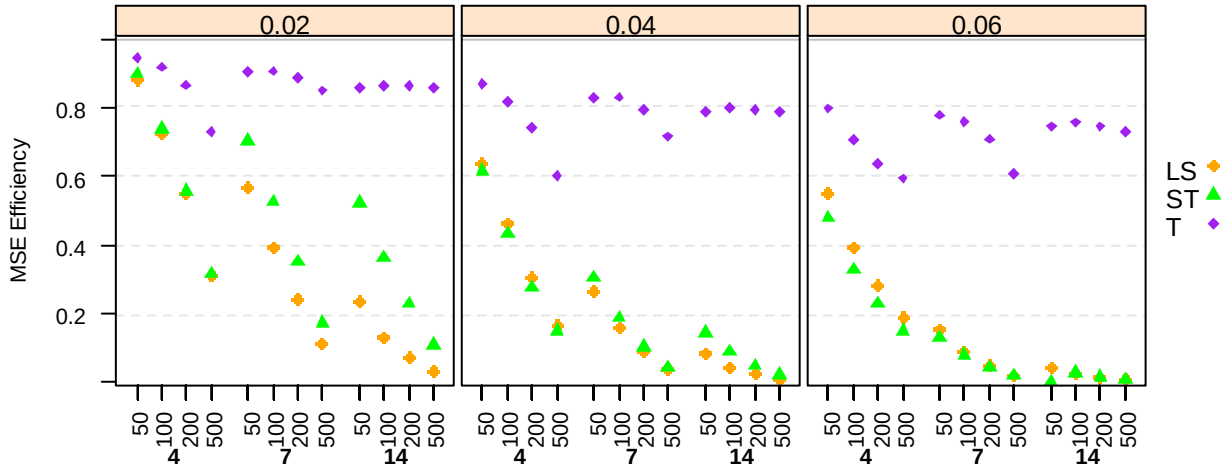

Figure 56. MSE efficiency of the intercept estimators relative to MM. Bivariate mixture Model 5 from Chapter 2.

4.7 Empirical Examples

4.7.1 Single Factor Market Model and MM Intercept Bias

In finance the terms 'alpha' and 'beta' arise from the use of linear regression, $r_{it} = \beta_i r_{Mt} + \epsilon_{it}$, to break asset return r_{it} into a component that is perfectly correlated with the market, $\beta_i r_{Mt}$, and an uncorrelated or residual component, ϵ_{it} (Grinold & Kahn, 1999). The regression slope is called asset's market beta. Beta plays a central role in the capital asset pricing model (CAPM) (Sharpe, 1964) and is one of the most widely known and widely used measures of the market risk of an asset. The mean of the residual component $E\epsilon_{it}$, or intercept, is often called Jensen's alpha (Jensen, 1968). Alpha determines abnormal return of an asset over a theoretical expected return predicted by the CAPM, $\beta_i \mu_M$. Investors favor positive skewness and are constantly seeking investments with high positive alpha.

There exists a considerable body of empirical finance literature on robust estimation of such linear regression models of asset returns (often referred to in finance as "factor models"). These studies were motivated by the knowledge that returns sometimes have fat-tailed and skewed non-normal distributions, and that an alternative to least-squares might therefore perform better. The focus of most of this empirical

work, however, has been on betas and LS inefficiency due to fat tails. See, for example, Bailer et al. (2012) , McDonald, Michelfelder, and Theodossiou (2010), Martin and Simin (1999, 2003), Bowie and Bradfield (1998), Fong(1997), Mills and Coutts (1996), Draper and Paudyal (1995), Chan and Lakonishok (1992), Cornell and Dietrich (1978), Sharpe (1971). We found only three robust 'alpha' studies, namely Hoorelbeke, Dewachter, and Smedts (2005), Bailer (2005) and McDonald, Michelfelder, and Theodossiou (2009), and only the latter directly addressed skewness in the residual returns. In particular, McDonald et al. (2009) modeled regression errors with a skewed fat-tailed distribution, obtaining parameter estimates via maximum likelihood. Even though the family of a skewed fat-tailed distribution used in McDonald et al. (2009) is different from that of Azzalini and Capitanio (2003) considered in our work, the implications are the same: a) alpha estimates based on symmetric fat-tailed probability distributions are biased downwards in positively skewed data and upwards in negatively skewed data; b) the size of the bias is directly related to the extent of skewness and kurtosis. Analysis by McDonald et al. (2009) of more than six thousand stocks from the CRSP database revealed that majority of the stocks (79%) have positively skewed CAPM residuals. The CAPM intercept bias due to skewness, measured as a difference between intercepts estimated using the skewed and corresponding symmetric likelihood specifications, is statistically significant at the 0.05 level for 67% of the stocks. These findings are important as "…biased alpha can lead to erroneous decisions on stock valuation, portfolio selection, and mutual fund investment evaluation. Moreover, stocks with biased alphas can lead to biased and inefficient portfolios…" (McDonald et al., 2009, pp.295-296).

This section presents a short empirical study following McDonald et al. (2009). The main goal is to assess extent of the MM intercept bias due to skewness in estimating Jensen's alpha in a large universe of publicly traded US stocks.

Data

The data for this study is from the Center of Research in Securities Prices (CRSP) Database. Following McDonald et al. (2009) we focus on all common stocks that were listed between January 1, 1995 and December 31, 2004 on the NYSE, AMEX and NASDAQ exchanges with at least four years of data (1 000

trading-day returns). This resulted in a universe of 6930 stocks. The times series sample sizes for individual stocks range between1003 and 2519 daily observations.

For each stock regression estimates of alpha and beta were computed using LS, ST MLE (based on skewed-t distribution of Azzalini and Capitanio (2003) and robust MM with bisquare psi-function at 95% normal distribution efficiency. In these regressions, r_{it} and r_{Mt} are stock and market returns in excess of a risk-free rate. The daily risk-free rates were obtained from the 'Fama/French Factors' daily dataset from Kenneth French data library (French, 2012). Stock returns are daily holding period total returns from the CRSP database. The CRSP value-weighted NYSE/AMEX/NASDAQ composite index was used as the market proxy.

We excluded stocks with grossly low liquidity as indicated by 30% or more zero returns. This resulted in excluding 356 stocks. In addition, we eliminated stocks for which the skewed-t degrees of freedom estimates were 1.5 or less. This resulted in excluding an additional 139 stocks. The final analysis universe is 6435 stocks.

Non-Normality of Stock Returns

Analysis of the regression residuals in Figure 57 and Figure 58 reveals that their distributions are highly non-normal in that they are both leptokurtic and skewed. For an overwhelming majority of the stocks the skewness is positive.

Figure 57 displays kernel densities (across all stocks) of the skewness and excess kurtosis coefficients of the MM residuals. The graphs for LS look the same and are not shown. There is a striking difference between skewness and kurtosis calculated using all residuals (labeled 'classical') and after trimming 0.5% of the smallest and 0.5% of the largest observations (labeled 'trimmed'). The values of the classical estimates appear to be unrealistically large and are driven by just a few extremely gross outliers. Nonetheless, even when trimming a tiny fraction of outliers, the resulting outlier robust skewness and kurtosis estimators reveal frequent significant skewness and kurtosis. For example, as a rough guide to judge statistical significance of a sample skewness one may use $\sqrt{6/n}$ for the standard errors, where n is the sample size. In our case a skewness outside the -0.2 to 0.2 range can roughly be deemed

significantly different from zero. For about 0.3% of stocks the trimmed skewness was below -0.2 and for about 79% of stocks it was above 0.2 (see Table 20). Approximate standard errors of a sample kurtosis coefficient are equal to $\sqrt{24/n}$ and range between 0.1 and 0.16 for the sample sizes at hand. For 99% of the stocks the trimmed excess kurtosis was above 0.5 (see Table 21).

Figure 58 displays characteristics of the skewed-t regression fits. The presence of fat tails and mostly positive skewness is evident from the histograms of the DF and skewness parameter λ estimates. The average DF estimate is 3 and for 83% of the stocks the estimated DF range between 2 and 4. The λ values are positive for 98.2% of the stocks. The λ values are positive and statistically significant at 0.05 level for 61% of the stocks. For the most part λ ranges between 0 and 0.75 when DF are larger than 2 and between -0.1 and 0.25 otherwise. Interestingly, the scatter plot of the ST skewness and location parameters in the bottom center in Figure 58 reveals a negative correlation between the two. There is, however, no strong correlation between estimated λ and alpha, the mean of the error distribution.

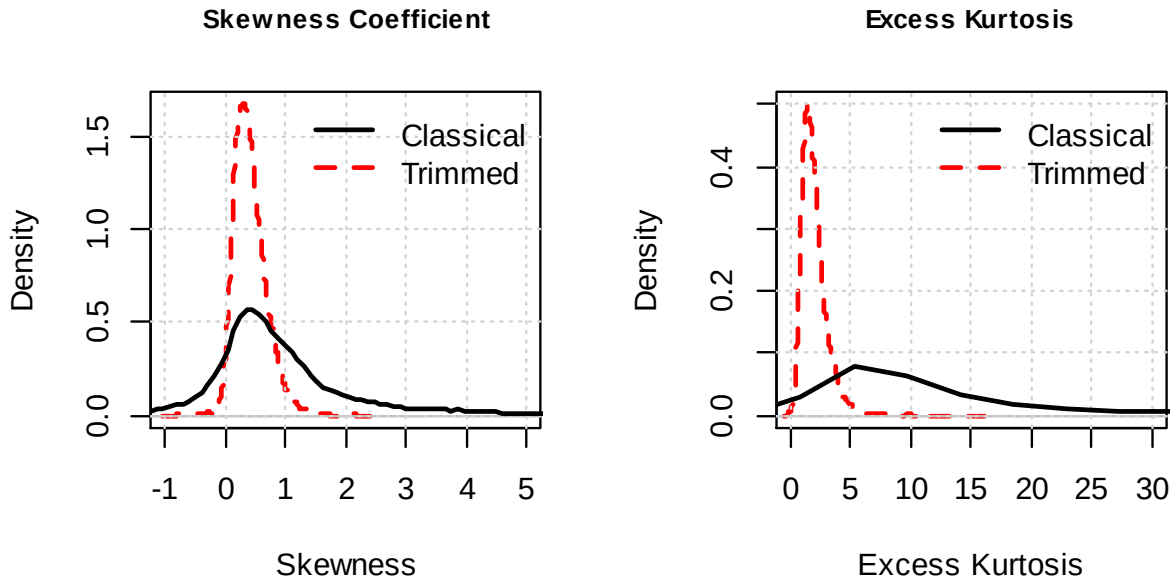

Figure 57. Kernel densities of cross-section distributions of the skewness (left) and excess kurtosis (right) of the MM residuals.

Table 20. Skewness coefficient of the residuals.

	LS		MM	
	< -0.2	>0.2	< -0.2	>0.2
Classical Skewness	9%	78%	9%	78%
Trimmed Skewness	0.4%	78%	0.3%	79%
McDonald et Al	9%	79%	NA	NA

Table 21. Excess kurtosis coefficient of the residuals.

	>4		>0.5	
	LS	MM	LS	MM
Classical Excess Kurtosis	88%	89%	1	1
Trimmed Excess Kurtosis	5%	5%	99%	99%
McDonald et Al	88%	NA	NA	NA

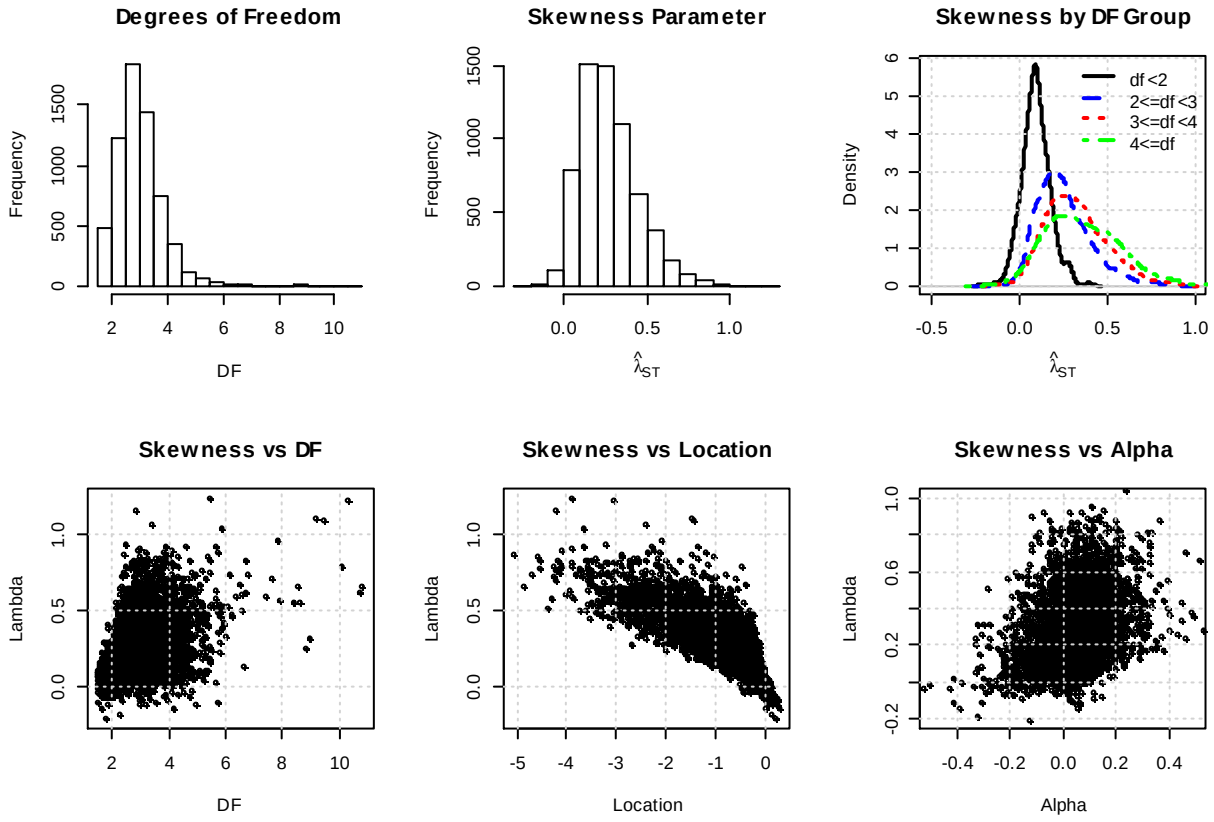

Figure 58. Characteristics of the skewed-t regression fits.

MM Intercept Bias

In view of the importance of Jensen's alpha as defined earlier, the intercept parameter of interest is taken to be the mean of the error term. ST intercept is calculated from the ST parameter MLE's using equation (4.17) for the mean of a skewed-t distribution.

Various alpha estimates are compared in Figure 59 with the scatter-plots and in Figure 60 with the boxplots of pair-wise differences. There are large differences between MM and LS, and between MM and ST intercept estimates. For vast majority of the stocks the differences are negative suggesting that MM intercept may be underestimating stock's alpha. This result is not surprising in the light of positive skewness in the regression errors described earlier. By comparing c-MM (c=20) intercept with ST MLE one may conclude that c-MM corrects most of the bias in the MM intercept estimate.

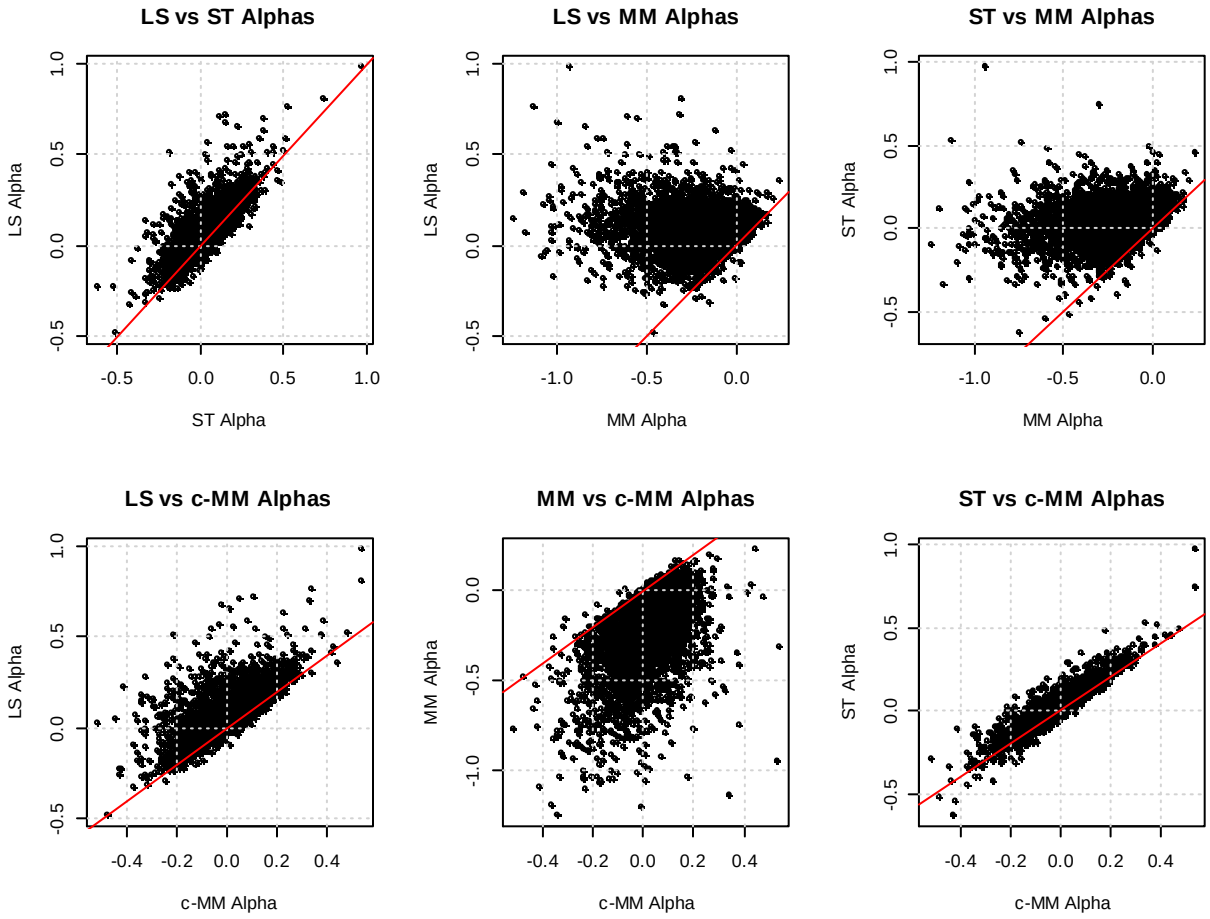

Figure 59. Scatter plots of alpha estimates: LS, ST, MM and c-MM with c=20.

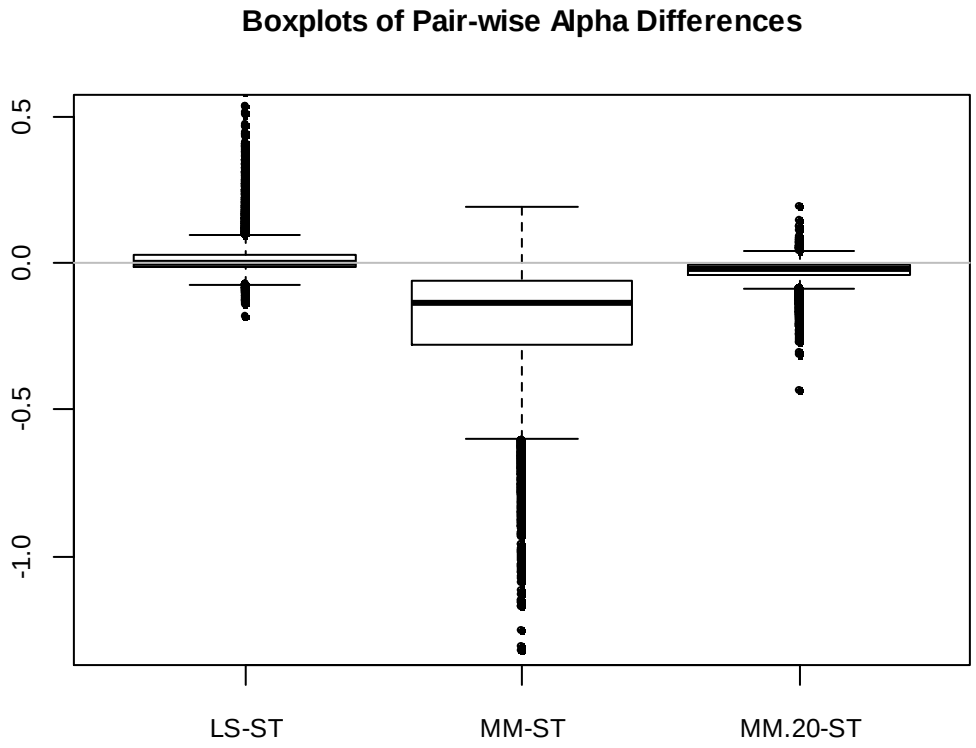

Figure 60. Boxplots of pair-wise differences between LS, MM, c-MM (c=20) and ST alpha estimates.

We further illustrate MM intercept bias by comparing it with the c-MM intercept estimates, for c=15, 20 and ∞. Figure 61 (left) shows kernel densities of the differences between c-MM and MM intercept estimates for all 6435 stocks from the analysis universe. For 98% of the stocks the differences are positive (for all three values of c), a consequence of positively skewed regression residuals. Median differences are 0.1%, 0.11% and 0.14% for c=15, 20 and ∞ respectively. The lower (upper) quartiles of the differences are 0.06% (0.3%), 0.05% (0.24%) and 0.05% (0.21%) for c=15, 20 and ∞ respectively. Figure 61 (right) shows kernel densities of the corresponding test statistics for the differences. The tests employed here are similar in nature to the test T2 from Chapter 2 and will be described in details in Chapter 5. Rejection in the test occurs when one of the two intercept estimators has larger asymptotic bias, which is, for example, a case under skewed errors. The vertical dashed lines mark rejection regions at 0.05 and 0.01 level. The three curves in Figure 61 (right) overlay suggesting that the proportions of stocks with statistically significance differences between c-MM and MM alpha estimates are the same for

all three values of c. These proportions, along with the one for the LS – MM difference, are reported in Table 22. We conclude that about 66% of the stocks have statistically significant bias in MM intercept at 0.05 level and about 49% of the stocks have statistically significant bias at 0.01 level. These results are quite consistent with the findings in Table 5 in McDonald et al. (2009), where the authors compared skewed-t with the symmetric-t intercept estimates using log-likelihood ratio test and found that the difference was significant at 0.05 (0.01) level for 66.6% (51.1%) of the stocks.

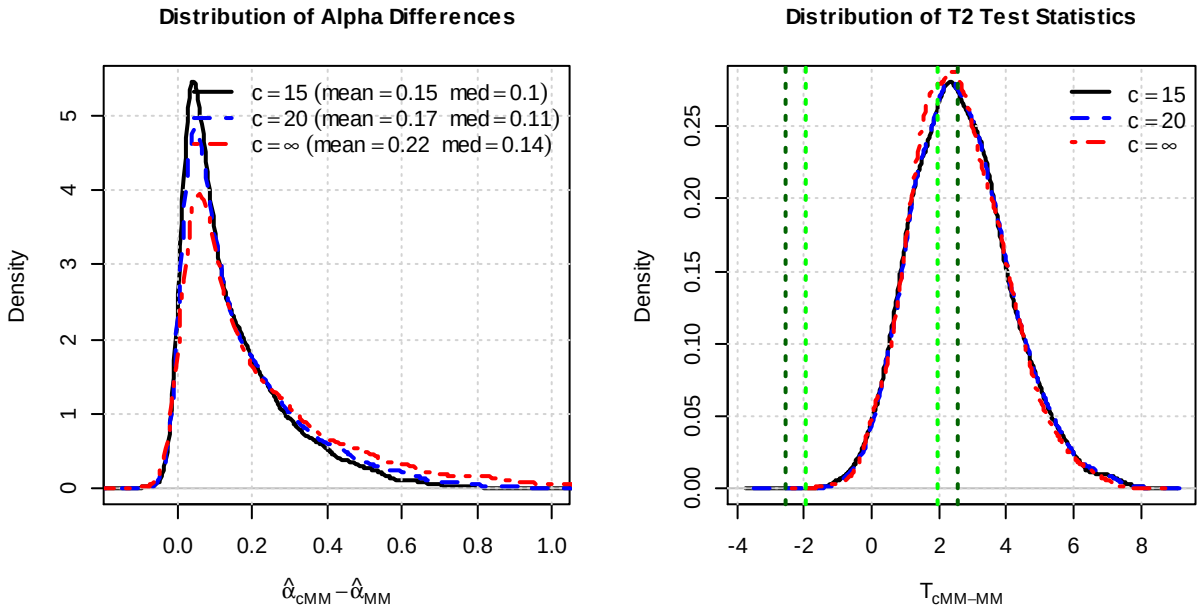

Figure 61. Kernel densities of the pair-wise differences between c-MM and MM alpha estimates (left) and of the corresponding T2 test statistics for the alpha differences (right). Dashed vertical lines mark test critical values at 0.05 and 0.01 significance levels.

Table 22. Statistically significant differences in estimated alphas (per T2 test).

	at 0.05 level	at 0.01 level
$LS - MM$	63.9%	46.4%
$\infty MM - MM$	65.1%	47.5%
$cMM - MM$; c=20	67%	49.7%
$cMM - MM$; c=15	66.5%	49%

It is also interesting to note the behavior of the different beta estimates. LS, MM and ST MLE are compared in Figure 62 with the scatter-plots and in Figure 63 with the boxplots of the pair-wise differences. MM and ST betas are similar to each other, but substantially different from LS beta. This result is not surprising since MM and ST MLE are both robust to outliers in the data.

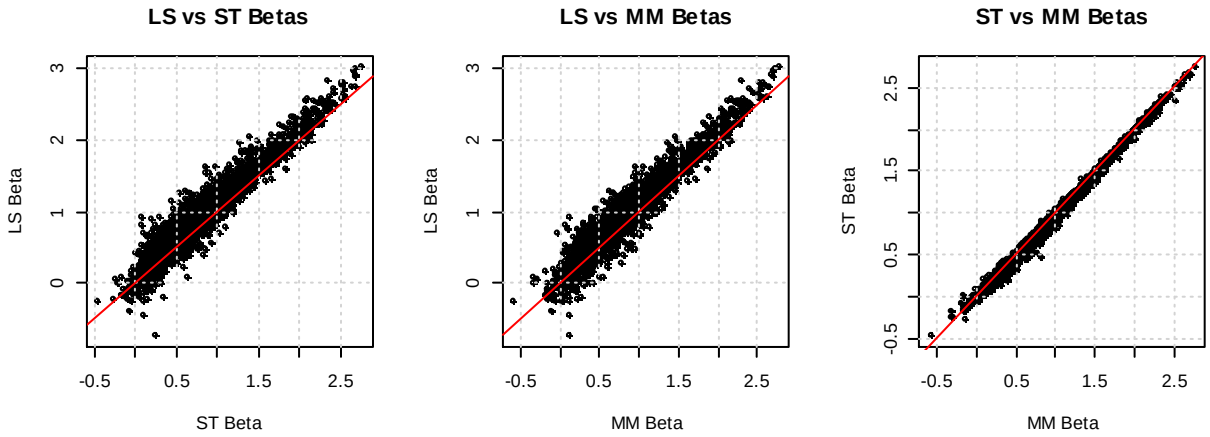

Figure 62. Scatter plots of beta estimates. LS, ST MLE and MM.

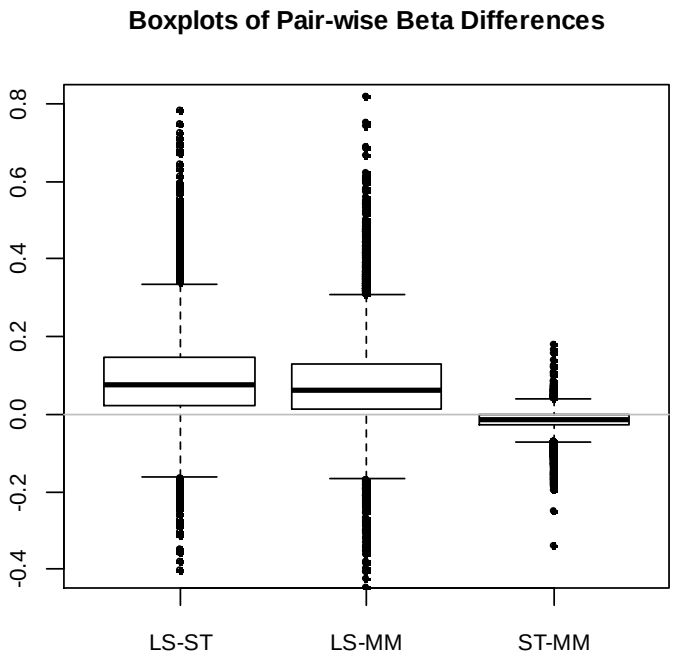

Figure 63. Boxplots of pair-wise differences between different beta estimates.

4.7.2 Australian Institute of Sport Data

The Australian Institute of Sport (AIS) data was described in Cook and Weisberg (1994) and became popular in the literature on skewed-t distributions (Azzalini & Capitanio, 2003; Azzalini & Genton, 2008; Arrelano-Valle & Genton, 2010; Marchenko, 2010; and others). The dataset is available in R package *sn* and contains information on several biomedical variables for 102 male and 100 female athletes.

We consider modeling lean body mass[18] (lbm) as a linear function of height (Ht) and weight (Wt) for the male athletes. Figure 64 shows the matrix scatter plot of the variables of interest. The predictor variables, Ht and Wt, are centered before fitting a regression model so that intercept is interpreted as an average lbm for a male athlete of average weight and height. The results of the OLS, MM and ST MLE regression fits are presented in Table 23. Bisquare psi-functions with c=4.68 (95% normal distribution efficiency) and c=15 were used for the MM and c-MM intercept estimates respectively.

The difference between LS and MM slope estimates for Ht is small and non-significant according to both T1 (p-value=0.88) and T2 (p-value=0.96) while the difference for Wt is somewhat larger and is significant in T1 (p-value=0.001), but not in T2 (p-value=0.19). Note that the estimated ST degrees of freedom is 2.3 so the observed results could be a consequence of inefficiency of LS slopes under a fat-tailed error distribution. We note that ST MLE slope estimate for Ht is somewhat smaller than LS and MM while ST MLE slope estimate for Wt is somewhat larger.

The MM intercept estimate is equal to 75.06 and appears to overstate average lbm for a male athlete as a consequence of negative skewness ($\hat{\lambda}_{ST} = -1.62$, p-value=0.013) and heavy tails ($\hat{\nu}_{ST} = 2.3$) of the regression errors. The ST intercept of 74.58 and OLS intercept of 74.66 are lower than MM intercept, though the differences of 0.4 and 0.48 are rather small relative to the intercept values themselves (only about 0.5%) and are likely negligible and not of practical significance.

---

[18] $Lean\ Body\ Mass = Body\ Weight \cdot \frac{100\% - Body\ Fat\ \%}{100\%}$, i.e. body weight minus fat content; composed of bones, muscles and other nonfat tissue.

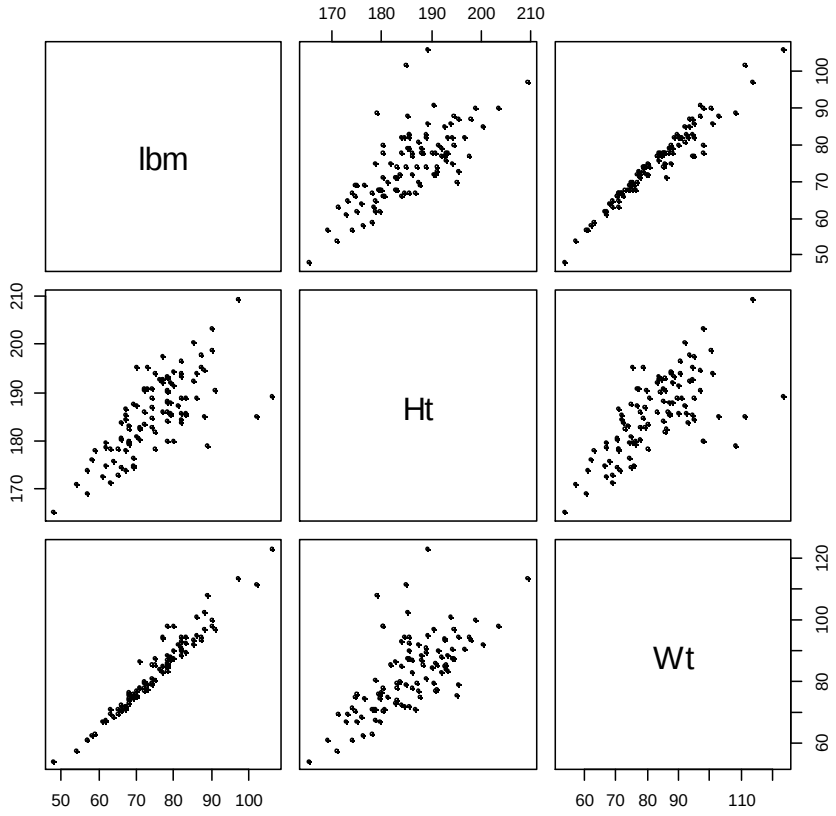

Figure 64. Scatter plots of lean body mass (lbm, kg), height (Ht, cm) and weight (Wt, kg) for 102 male athletes from Australian Institute of Sports dataset.

Since Ht and Wt variables are centered, the ∞MM intercept estimate value coincides with that of the LS intercept and has similar standard error. For both $LS - MM$ and $\infty MM - MM$ the difference of -0.4 is statistically significant with p-value equal to 0.02. If we do not center Ht and Wt then LS intercept is equal to -11.5 with a standard error of 5.9 while ∞MM intercept is equal to -13.5 and has smaller standard error of 3.8. Note that the difference between ∞MM and MM and its p-value do not change (equal to -0.4 with p-value of 0.02) while the difference between LS and MM becomes larger and statistically non-significant (equal to 1.6 with p-value of 0.74). This is an illustration of a propagation of the differences between slope estimates and their uncertainties as we move further from the mean of the predictor vector as was discussed in Section 4.4.1.

The c-MM intercept estimate is equal to 74.72. Even though in this example it does not offer any advantage over ∞MM or LS intercepts note that it corrects most of the bias of the MM intercept.

Table 23. Regression results for the AIS example.

Sex=Male		Estimate	Std Errors	T statistic	P value	
LS	Intercept	74.657	0.223	334.672	0.000	***
	Ht	0.148	0.038	3.901	0.000	***
	Wt	0.710	0.024	29.301	0.000	***
MM	Intercept	75.062	0.227	330.957	0.000	***
	Ht	0.147	0.046	3.169	0.002	**
	Wt	0.738	0.039	19.141	0.000	***
ST	Intercept	74.582				
	Ht	0.115	0.035	3.304	0.001	***
	Wt	0.766	0.028	26.964	0.000	***
	location	76.239	0.374	203.862	0.000	***
	scale	1.529	0.319	4.794	0.000	***
	shape	-1.616	0.650	-2.485	0.013	*
	df	2.290	0.612	3.744	0.000	***
LS-MM (T1)	Intercept[19]	-0.405	0.051	-7.990	0.000	***
	Ht	0.002	0.010	0.157	0.875	
	Wt	-0.028	0.009	-3.241	0.001	***
LS-MM (T2)	Intercept	-0.405	0.175	-2.319	0.020	*
	Ht	0.002	0.032	0.051	0.960	
	Wt	-0.028	0.021	-1.311	0.190	
$cMM\ (c = 15)$	Intercept	74.724	0.212	352.574	0.000	***
$cMM - MM$	Intercept	-0.338	0.150	-2.252	0.024	*
∞MM	Intercept	74.657	0.222	335.865	0.000	***
$\infty MM - MM$	Intercept	-0.405	0.176	-2.306	0.021	*

[19] In contrast to T2 test for the slopes that was introduced in Chapter 2 the null hypothesis for the intercepts is a symmetric error distribution. The tests for the differences between two intercept estimators will be discussed in more details in Chapter 5.

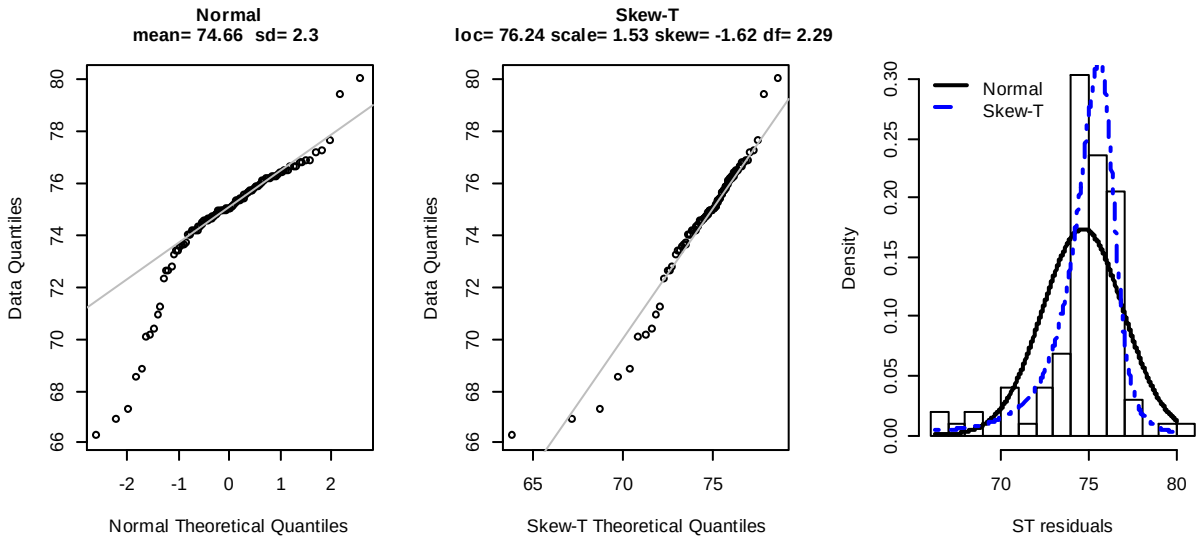

Figure 65. Residual plots for the AIS example.

4.8 Summary and Discussion

This chapter studied efficiency and bias of the least-squares (LS) and robust MM estimators in a regression with skewed-t distributed errors. This choice of error distributions was motivated by the fact that modeling errors with a fat-tailed and possibly skewed distribution (instead of a classical normal MLE) is a popular parametric approach to robust regression estimation. MM robust regression, on the other hand, is a non-parametric method and its performance relative to skewed-t MLEs was yet to be revealed. Our key conclusions are as follows.

The MM regression estimator can simultaneously achieve high asymptotic efficiency (e.g., over 95%) under both normal and symmetric fat-tailed errors. This is in sharp contrast to LS: while being fully efficient under normal errors, LS has very low efficiency under symmetric fat-tailed errors (e.g., 50% under t-distribution with 3 degrees of freedom). The statement is true for both the intercept and the slopes.

Under skewed errors, the behavior of the MM slopes differs from that of the MM intercept. MM slopes are consistent, but they lose some of their high asymptotic efficiency that is observed under symmetric errors. The decrease in efficiency is not substantial for small and moderate degree of skewness, and so is of

concern for highly skewed data only. In either case, efficiency of the MM estimator of the slopes is much higher than that of the LS, more so the fatter the tails. MM intercept, on the other hand, typically has smaller variance than a skewed-t MLE, but is biased: it overestimates the true mean of an error term that is negatively skewed and underestimates the mean in case of positive skewness. Under mild skewness, the mean-squared-error of the MM intercept is still smaller than that of the skewed-t MLE, as well as of the LS, in small and moderate sample sizes. For moderate and high skewness, however, MM bias quickly becomes an issue. Taking the mean of the MM residuals to be a new intercept estimate (called ∞MM) corrects the bias. Asymptotic variance efficiencies of the ∞MM and LS intercepts under skewed-t distributions are surprisingly high. Similarly to LS, however, the ∞MM intercept is sensitive to the presence of gross outliers and has a breakdown point of zero.

Estimating skewness when in fact it is zero results in a penalty in terms of extra finite-sample variance of the skewed-t MLE of the regression slopes and intercept. The penalty disappears in large samples (i.e. asymptotically) for the slopes, but not for the intercept. In general, in small samples estimation of the degrees of freedom and skewness parameters is unreliable and contributes to numerical instability and a large variance of the skewed-t MLEs of the slopes and intercept. Therefore, it is suggested that the skewed-T MLE should not be used unless there are reasons to expect at least moderate skewness in the data and unless large sample sizes are available.

In this work we focused on comparison of the MM regression estimators with a skewed-t MLE under skewed-t error distributions. A natural next step is to compare robustness properties of the MM and the skewed-t MLE under contamination models of type (1.4) where, for example, a main component is from a skewed-t distribution while contamination component creates outliers inconsistent with the main component. A difficult part of the task is to distinguish between the outliers generated by a contamination component that needs to be ignored and the genuine observations that constitute an essential part of the skewness that needs to be accounted for by an intercept estimate. The TML method of Marazzi and Yohai (2004) offers a solution only for a case when the skewness and degrees of freedom parameters of the main component are known. The robust c-MM intercept proposed in this chapter is an ad-hoc way to alleviate the issue with the MM intercept bias in a very general setting when degrees of skewness and tail

fatness are unknown. The c-MM intercept is an M-location estimator with a larger value of a tuning constant *c* than that of an original MM fit. This way, c-MM has a substantially smaller bias than MM intercept, but is still robust to gross outliers and has a high breakdown point. Under mild skewness and heavy tails, the c-MM intercept is inferior to MM due to its much larger variance. Under moderate skewness, however, the c-MM intercept outperforms MM, LS and skewed-t MLE in up to rather large sample sizes. Hence, we argue that the MM estimator, possibly with a corrected intercept, could be a good universal method for fat-tailed distributions with up to moderate skewness - particularly in small to moderate sample sizes where fitting a skewed-t distribution is difficult. Unfortunately, when skewness is high[20], MM slopes become very inefficient and intercept bias becomes large so that neither MM nor c-MM is satisfactory.

It is important to note that MLEs based on symmetric distributions also result in unbiased slopes but biased intercept. We observed similar behavior of the MM and symmetric-t MLE under skewed-t errors.

[20] While degrees of freedom and sample size are also the deciding factors in efficiency or inefficiency of MM and c-MM estimators, a rough guide for a skewed-t distribution of Azzalini and Capitanio (2003) is an absolute value of a shape parameter λ larger than 3.

Appendix A4:

Asymptotic Variance of ST MLE

Linear regression model (1.1) and (1.3) with $F_{\epsilon}\sim ST(\xi,\omega^2,\lambda,\nu)$ implies that $Y\sim ST(\xi+\boldsymbol{x}'\boldsymbol{\beta},\omega^2,\lambda,\nu)$.

Let

$$D=\frac{1}{\omega^2} \qquad \eta=\frac{\lambda}{\omega}$$

$$u=\epsilon-\xi \qquad Q=\frac{u^2}{\omega^2}=Du^2 \tag{A4.1}$$

$$L=\frac{\lambda u}{\omega}=\eta u \qquad t(L,Q,\nu)=L\left(\frac{\nu+1}{Q+\nu}\right)^{\frac{1}{2}}$$

Following Azzalini and Capitanio (2003) define a vector of working parameters $\boldsymbol{\vartheta}'=(\xi,\boldsymbol{\beta}',D,\eta,\nu)$. The asymptotic covariance matrix of the skew-t maximum likelihood estimator $\widehat{\boldsymbol{\vartheta}}_{\boldsymbol{ST}}$ is equal to the inverse of the expected information matrix:

$$\boldsymbol{V}_{\boldsymbol{\vartheta}}=\left(E\left\{\frac{\partial l}{\partial\boldsymbol{\vartheta}}\frac{\partial l'}{\partial\boldsymbol{\vartheta}}\right\}\right)^{-1} \tag{A4.2}$$

where l is the log-likelihood function

$$l(y;\,\boldsymbol{\vartheta})=\log 2+\frac{1}{2}\log|D|+\log g_1(Q,\nu)+\log T_1(t(L,Q,\nu);\nu+1)$$

The components of the score vector

$$\frac{\partial l}{\partial\boldsymbol{\vartheta}}=\left(\frac{\partial l}{\partial\xi},\frac{\partial l}{\partial\boldsymbol{\beta}},\frac{\partial l}{\partial D},\frac{\partial l}{\partial\eta},\frac{\partial l}{\partial\nu}\right)'$$

are as follows:

$$\frac{\partial l}{\partial \xi} = \left(-2\left(\tilde{g}_Q + \tilde{T}_1 * \dot{t}_Q\right)\frac{u}{\omega^2} - \tilde{T}_1 \dot{t}_L \eta\right) \equiv f(u;\omega,\lambda,\nu)$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = f(u;\omega,\lambda,\nu)\boldsymbol{x}$$

$$\frac{\partial l}{\partial D} = u^2\left(\tilde{g}_Q + \tilde{T}_1 * \dot{t}_Q\right) + \frac{1}{2}D^{-1} \tag{A4.3}$$

$$\frac{\partial l}{\partial \eta} = \tilde{T}_1 \dot{t}_L u$$

$$\frac{\partial l}{\partial \nu} = \frac{\partial \log g_1}{\partial \nu} + \frac{\partial \log T_1(t;\nu+1)}{\partial \nu}$$

where

$$\tilde{g}_Q \equiv \frac{\partial \log g_1(Q,\nu)}{\partial Q} = -\frac{\nu+1}{2\nu}\left(1+\frac{Q}{\nu}\right)^{-1}$$

$$\tilde{T}_1 \equiv \frac{\partial \log T_1(t;\nu+1)}{\partial t} = T_1(t;\nu+1)^{-1} t_1(t;\nu+1)$$

$$\dot{t}_L \equiv \frac{\partial t(L,Q,\nu)}{\partial L} = \left(\frac{\nu+1}{Q+\nu}\right)^{\frac{1}{2}} \tag{A4.4}$$

$$\dot{t}_Q \equiv \frac{\partial t(L,Q,\nu)}{\partial Q} = -\frac{L(\nu+1)^{\frac{1}{2}}}{2(Q+\nu)^{\frac{3}{2}}}$$

$$\frac{\partial \log g_1}{\partial \nu} = \frac{1}{2}\left(\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right) - \frac{1}{\nu} + \frac{(\nu+1)Q}{\nu^2\left(1+\frac{Q}{\nu}\right)} - \log\left(1+\frac{Q}{\nu}\right)\right)$$

with

$$\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$$

denoting the digamma function. Expression for $\frac{\partial \log T_1(t(L,Q,\nu);\,\nu+1)}{\partial \nu}$ is not given above as it appears intractable and must be evaluated numerically.

Using the matrix inversion formula of a partitioned matrix and equation (A2.11) which states that $\boldsymbol{V}_{x^*}^{-1}\boldsymbol{\mu}_{x^*} = \begin{pmatrix}1\\ \mathbf{0}\end{pmatrix}$ we obtain

$$\boldsymbol{V}_{\boldsymbol{\vartheta}} = \begin{pmatrix} \boldsymbol{V}_{(\xi,\boldsymbol{\beta})} & \boldsymbol{V}_{(\xi,\boldsymbol{\beta})\times(D,\eta,\nu)} \\ \boldsymbol{V}'_{(\xi,\boldsymbol{\beta})\times(D,\eta,\nu)} & \boldsymbol{V}_{(D,\eta,\nu)} \end{pmatrix}$$

where

$$\boldsymbol{V}_{(D,\eta,\nu)} \equiv \begin{pmatrix} V_D & V_{D\times\eta} & V_{D\times\nu} \\ V_{D\times\eta} & V_\eta & V_{\eta\times\nu} \\ V_{D\times\nu} & V_{\eta\times\nu} & V_\nu \end{pmatrix} = \left[\begin{pmatrix} S_{D^2} & S_{D\eta} & S_{D\nu} \\ S_{D\eta} & S_{\eta^2} & S_{\eta\nu} \\ S_{D\nu} & S_{\eta\nu} & S_{\nu^2} \end{pmatrix} - \frac{1}{S_{\xi^2}} \begin{pmatrix} S_{\xi D}^2 & S_{\xi D}S_{\xi\eta} & S_{\xi D}S_{\xi\nu} \\ S_{\xi\eta}S_{\xi D} & S_{\xi\eta}^2 & S_{\xi\eta}S_{\xi\nu} \\ S_{\xi\nu}S_{\xi D} & S_{\xi\nu}S_{f\eta} & S_{\xi\nu}^2 \end{pmatrix} \right]^{-1}$$

$$\boldsymbol{V}_{(\xi,\boldsymbol{\beta})\times(D,\eta,\nu)} = -\frac{1}{S_{\xi^2}} \begin{pmatrix} S_{\xi D} & S_{\xi\eta} & S_{\xi\nu} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{V}_{(D,\eta,\nu)} \equiv \begin{pmatrix} V_{\xi\times D} & V_{\xi\times\eta} & V_{\xi\times\nu} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \qquad \text{(A4.5)}$$

$$\boldsymbol{V}_{(\xi,\boldsymbol{\beta})} = \frac{1}{S_{\xi^2}} \boldsymbol{V}_{\boldsymbol{x}^*}^{-\mathbf{1}} - \frac{1}{S_{\xi^2}} \left(S_{\xi D}V_{\xi\times D} + S_{\xi\eta}V_{\xi\times\eta} + S_{\xi\nu}V_{\xi\times\nu} \right) \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

and $S_{..}$ is the expectation of the product of the corresponding components of the score vector, e.g. $S_{D\eta} = E\left(\frac{\partial l}{\partial D}\frac{\partial l}{\partial \eta}\right)$, $S_{\xi^2} = E\left(\frac{\partial l}{\partial \xi}\frac{\partial l}{\partial \xi}\right)$, etc.

The expectation in (A4.2) is computed via numerical integration using R function integrate.

Asymptotic variance of the ST MLE of the slopes

From (A4.5) it follows that the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_{ST}$ is

$$\kappa_{ST}\boldsymbol{C_x^{-1}}$$

where

$$\kappa_{ST} = \frac{1}{E(f(u;\omega,\lambda,\nu)^2)}$$

and f is defined in (A4.3).

We note that

1) From (A4.5) it also follows that the skewed-t maximum likelihood estimator of the slopes $\boldsymbol{\beta}$ is uncorrelated with the estimators of D, η, and ν
2) When $\boldsymbol{\mu_x} = \mathbf{0}$ then $\widehat{\boldsymbol{\beta}}_{\boldsymbol{ST}}$ is also uncorrelated with $\hat{\xi}_{ST}$, the MLE of the ST location parameter ξ. This result follows from (A4.5) and expression (1.23) for the inverse of $\boldsymbol{V_{x^*}}$ which is $\boldsymbol{V_{x^*}^{-1}} = \begin{bmatrix} 1+\boldsymbol{\mu_x'C_x^{-1}\mu_x} & -\boldsymbol{\mu_x'C_x^{-1}} \\ -\boldsymbol{C_x^{-1}\mu_x} & \boldsymbol{C_x^{-1}} \end{bmatrix}$.
3) When $\lambda = 0$ then $\kappa_{ST} = \frac{\nu+3}{\nu+1}\omega^2$ so that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{ST}}$ has the same asymptotic variance as symmetric T MLE of $\boldsymbol{\beta}$.

Asymptotic variance of the ST MLE of the intercept

The intercept is equal to $\alpha = \xi + \omega\delta b_\nu = \xi + b_\nu \frac{\eta D^{-1}}{\sqrt{1+\eta^2 D^{-1}}}$ where δ and b_ν are defined in (4.2) and (4.3).

The maximum likelihood estimate $\hat{\alpha}_{ST}$ is computed by plugging in the corresponding elements of $\widehat{\boldsymbol{\vartheta}}_{\boldsymbol{ST}}$.

The asymptotic variance of $\hat{\alpha}_{ST}$ can be obtained by delta-method:

$$\left(\frac{\partial\alpha}{\partial\boldsymbol{\vartheta}}\right)' \boldsymbol{V_\vartheta} \frac{\partial\alpha}{\partial\boldsymbol{\vartheta}} \qquad \text{(A4.6)}$$

where

$$\frac{\partial \alpha}{\partial \boldsymbol{\vartheta}} = \left(\frac{\partial \alpha}{\partial \xi}, \frac{\partial \alpha}{\partial \boldsymbol{\beta}}, \frac{\partial \alpha}{\partial D}, \frac{\partial \alpha}{\partial \eta}, \frac{\partial \alpha}{\partial \nu}\right)'$$

$$\frac{\partial \alpha}{\partial \xi} = 1$$

$$\frac{\partial \alpha}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

$$\frac{\partial \alpha}{\partial D} = -b_\nu \eta \frac{1 + \frac{1}{2}\eta^2 D^{-1}}{D^2(1 + \eta^2 D^{-1})^{\frac{3}{2}}} = -b_\nu \lambda \omega^3 \left(1 + \frac{1}{2}\lambda^2\right)(1 + \lambda^2)^{-\frac{3}{2}}$$

$$\frac{\partial \alpha}{\partial \eta} = b_\nu D^{-1} \frac{1}{(1 + \eta^2 D^{-1})^{\frac{3}{2}}} = b_\nu \omega^2 (1 + \lambda^2)^{-\frac{3}{2}} \tag{A4.7}$$

$$\frac{\partial \alpha}{\partial \nu} = \frac{\eta D^{-1}}{\sqrt{1 + \eta^2 D^{-1}}} b_\nu \frac{1}{2}\left(\frac{1}{\nu} + \psi\left(\frac{\nu - 1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right) = b_\nu \omega \frac{\lambda}{\sqrt{1 + \lambda^2}} \frac{1}{2}\left(\frac{1}{\nu} + \psi\left(\frac{\nu - 1}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right)$$

with

$$\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$$

denoting the digamma function.

We note that when $\lambda = 0$ the asymptotic variance $\hat{\alpha}_{ST}$ is larger than the asymptotic variance of the symmetric T MLE of location. This is in contrast to the slopes, where ST MLE and T MLE asymptotically have equal variance when $\lambda = 0$.

c-MM Asymptotic Result

In this section we derive the asymptotic distribution of the c-MM intercept. Consider a regression model (1.1) where the observed data $\boldsymbol{z_i'} = (y_i, \boldsymbol{x}_i'),\ i = 1, \cdots, n$, consists of i.i.d. random variables with a joint distribution F_0 (1.3), which assumes that for each i, $\boldsymbol{x_i}$ and $\boldsymbol{\epsilon_i}$ are independent.

Following the same steps as in Appendix A2 in Chapter 2 we write $\widehat{\boldsymbol{\eta}}' = \left(\hat{\alpha}_{cMM}, \hat{\alpha}_{MM}, \widehat{\boldsymbol{\beta}}_{MM}', \hat{\alpha}_S, \widehat{\boldsymbol{\beta}}_S', \hat{\sigma}\right)$, $\boldsymbol{\eta}' = (\alpha_{cMM}, \alpha_{MM},\ \boldsymbol{\beta}_{MM}',\ \alpha_S, \boldsymbol{\beta}_S', \sigma)$ and

$$\boldsymbol{\Psi}(\boldsymbol{z}, \boldsymbol{\eta}) = \begin{pmatrix} \psi_c\left(\dfrac{y - \alpha_{cMM} - \boldsymbol{x}'\boldsymbol{\beta}_{MM}}{\sigma}\right) \\ \psi_2\left(\dfrac{y - \alpha_{MM} - \boldsymbol{x}'\boldsymbol{\beta}_{MM}}{\sigma}\right)\boldsymbol{x}^* \\ \psi_1\left(\dfrac{y - \alpha_S - \boldsymbol{x}'\boldsymbol{\beta}_S}{\sigma}\right)\boldsymbol{x}^* \\ \rho_1\left(\dfrac{y - \alpha_S - \boldsymbol{x}'\boldsymbol{\beta}_S}{\sigma}\right) - b \end{pmatrix} \tag{A4.8}$$

where ψ_c is the psi-function for the c-MM intercept estimate, ψ_2 is the psi-function for the final step of the MM estimate while ρ_1 and $\psi_1 = \rho_1'$ are the loss and corresponding psi function for the initial S-estimate.

Consistency and asymptotic normality of $\widehat{\boldsymbol{\eta}}$ follow directly from the FMSY results for the M-functionals. The conditions under which the asymptotic results hold are assumptions (A1) – (A9) from Appendix A2 in Chapter 2 and their natural extension to a ρ_c loss function for a c-MM intercept. Thus, under rather general conditions $\widehat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$ and is asymptotically normal

$$\sqrt{n}\,(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\,) \to N(\mathbf{0}, \boldsymbol{\Sigma}) \tag{A4.9}$$

with

$$\boldsymbol{\Sigma} = \boldsymbol{D}^{-1}\boldsymbol{\Omega}\boldsymbol{D}'^{-1} \tag{A4.10}$$

where $\boldsymbol{D} = E_{F_0}\boldsymbol{\Psi}(\boldsymbol{z}, \boldsymbol{\eta})$ and $\boldsymbol{\Omega} = E_{F_0}\{\boldsymbol{\Psi}(\boldsymbol{z}, \boldsymbol{\eta}\,)\boldsymbol{\Psi}'(\boldsymbol{z}, \boldsymbol{\eta})\}$

$$a_{cMM} = E_{F_\epsilon}\psi_c'\left(\frac{u_c}{\sigma}\right)$$

$$e_{cMM} = E_{F_\epsilon}\left\{\psi_c'\left(\frac{u_c}{\sigma}\right)\frac{u_c}{\sigma}\right\}$$

$$f_{cMM} = E_{F_\epsilon}\left\{\psi_c^2\left(\frac{u_c}{\sigma}\right)\right\}$$

$$g_{cMM} = E_{F_\epsilon}\left\{\left(\rho_1\left(\frac{u_S}{\sigma}\right) - b\right)\psi_c\left(\frac{u_c}{\sigma}\right)\right\} = E_{F_\epsilon}\left\{\rho_1\left(\frac{u_S}{\sigma}\right)\psi_c\left(\frac{u_c}{\sigma}\right)\right\}$$

Direct calculations show

$$\boldsymbol{\Omega} = E_{F_0}\{\boldsymbol{\Psi}(\boldsymbol{z},\boldsymbol{\eta})\boldsymbol{\Psi}'(\boldsymbol{z},\boldsymbol{\eta})\} = \begin{bmatrix} f_{cMM} & f_{cMM,MM}\boldsymbol{\mu}_{\boldsymbol{x}^*}' & f_{cMM,S}\boldsymbol{\mu}_{\boldsymbol{x}^*}' & g_{cMM} \\ f_{cMM,MM}\boldsymbol{\mu}_{\boldsymbol{x}} & f_{MM}\boldsymbol{V}_{\boldsymbol{x}^*} & f_{MM,S}\boldsymbol{V}_{\boldsymbol{x}^*} & g_{MM}\boldsymbol{\mu}_{x^*} \\ f_{cMM,S}\boldsymbol{\mu}_{\boldsymbol{x}} & f_{MM,S}\boldsymbol{V}_{\boldsymbol{x}^*} & f_S\boldsymbol{V}_{\boldsymbol{x}^*} & g_S\boldsymbol{\mu}_{x^*} \\ g_{cMM} & g_{MM}\boldsymbol{\mu}_{\boldsymbol{x}^*}' & g_S\boldsymbol{\mu}_{\boldsymbol{x}^*}' & E\left\{\rho_1^2\left(\frac{u_S}{\sigma}\right) - b^2\right\} \end{bmatrix} \quad \text{(A4.11)}$$

where

$$f_{cMM,MM} = E_{F_\epsilon}\left\{\psi_c\left(\frac{u_c}{\sigma}\right)\psi_1\left(\frac{u_{MM}}{\sigma}\right)\right\} \qquad f_{cMM,S} = E_{F_\epsilon}\left\{\psi_c\left(\frac{u_c}{\sigma}\right)\psi_1\left(\frac{u_S}{\sigma}\right)\right\} \quad \text{(A4.12)}$$

and other constants are given in (4.15) and in (A2.7).

$$\dot{\boldsymbol{\Psi}}(\boldsymbol{z},\boldsymbol{\eta}) = \left\{\dot{\Psi}_{jk}(z,\eta) = \frac{\partial\Psi_j}{\partial\eta_k}\right\} = -\frac{1}{\sigma}\begin{bmatrix} \psi_c'\left(\frac{u_c}{\sigma}\right) & 0 & \psi_c'\left(\frac{u_c}{\sigma}\right)\boldsymbol{\mu}_{\boldsymbol{x}}' & \mathbf{0} & \psi_c'\left(\frac{u_c}{\sigma}\right)\frac{u_c}{\sigma} \\ \mathbf{0} & \psi_2'\left(\frac{u_{MM}}{\sigma}\right)\boldsymbol{x}^*\boldsymbol{x}^{*\prime} & \mathbf{0} & \psi_2'\left(\frac{u_{MM}}{\sigma}\right)\boldsymbol{x}^*\frac{u_{MM}}{\sigma} \\ \mathbf{0} & \mathbf{0} & \psi_1'\left(\frac{u_S}{\sigma}\right)\boldsymbol{x}^*\boldsymbol{x}^{*\prime} & \psi_1'\left(\frac{u_S}{\sigma}\right)\boldsymbol{x}^*\frac{u_S}{\sigma} \\ \mathbf{0} & \mathbf{0} & \psi_1\left(\frac{u_S}{\sigma}\right)\boldsymbol{x}^{*\prime} & \psi_1\left(\frac{u_S}{\sigma}\right)\frac{u_S}{\sigma} \end{bmatrix}$$

Since $E\psi_1\left(\frac{u_S}{\sigma}\right) = 0$ we get

$$\boldsymbol{D} = E_{F_0}\dot{\boldsymbol{\Psi}}(\boldsymbol{z},\boldsymbol{\eta}) = -\frac{1}{\sigma}\begin{bmatrix} a_{cMM} & 0 & a_{cMM}\boldsymbol{\mu}_{\boldsymbol{x}}' & \mathbf{0} & e_{cMM} \\ \mathbf{0} & a_{MM}\boldsymbol{V}_{\boldsymbol{x}^*} & \mathbf{0} & e_{MM}\boldsymbol{\mu}_{x^*} \\ \mathbf{0} & \mathbf{0} & a_S\boldsymbol{V}_{\boldsymbol{x}^*} & e_S\boldsymbol{\mu}_{x^*} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & d_S \end{bmatrix}$$

which implies

$$\boldsymbol{D^{-1}} = \begin{Bmatrix} \text{inverse of a} \\ \text{partitioned matrix} \end{Bmatrix}$$

$$= -\sigma \begin{bmatrix} \frac{1}{a_{cMM}} & -(0 \quad \boldsymbol{\mu_x'}) \frac{1}{a_{MM}} \boldsymbol{V_{x^*}^{-1}} & \mathbf{0} & -\frac{e_{cMM}}{a_{cMM} d_S} \\ \mathbf{0} & \frac{1}{a_{MM}} \boldsymbol{V_{x^*}^{-1}} & \mathbf{0} & -\frac{e_{MM}}{a_{MM} d_S} \boldsymbol{V_{x^*}^{-1}} \boldsymbol{\mu_{x^*}} \\ \mathbf{0} & \mathbf{0} & \frac{1}{a_S} \boldsymbol{V_{x^*}^{-1}} & -\frac{e_S}{a_S d_S} \boldsymbol{V_{x^*}^{-1}} \boldsymbol{\mu_{x^*}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1}{d_S} \end{bmatrix} \tag{A4.13}$$

In (A4.13) we used the fact that $(0 \quad \boldsymbol{\mu_x'})\, \boldsymbol{V_{x^*}^{-1}} = (-\boldsymbol{\mu_x' C_x^{-1} \mu_x} \quad \boldsymbol{\mu_x' C_x^{-1}})$ and $(0 \quad \boldsymbol{\mu_x'})\, \boldsymbol{V_{x^*}^{-1} \mu_{x^*}} = 0$.

Plugging (A4.11) and (A4.13) into (A4.10) we compute $V_{\hat{\alpha}_{cMM}}$ in (4.13)

and

$$cov\left(\hat{\alpha}_{cMM}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}\right) = -\frac{f_{MM}}{a_{MM}^2} \boldsymbol{C_x^{-1}} \boldsymbol{\mu_x}$$

5. Tests for Differences between Intercept Estimators

5.1 Introduction

While being treated as a nuisance parameter in tests T1 and T2 in Chapter 2, an intercept becomes the primary focus of this chapter. One of the reasons for separating the intercept from the slopes is their different behavior under skewed error distributions where the MM estimator is still consistent for the slopes, but is inconsistent for an intercept (for more details, see Chapter 4). As a result, the composite null hypothesis for the T2 test for the difference between the LS and robust MM intercept estimators differs from that for the slopes and consists of normal and symmetric non-normal distributions. The composite alternative hypothesis of this test consists of skewed non-normal distributions and of general types of bias-inducing joint distributions for predictors and response variables. Rejection of the null hypothesis for the T2 test occurs when one of the estimators has a sufficiently larger bias than the other. Additionally, we study a T2 type test of differences between a corrected MM intercept (c-MM) from Chapter 4 and an MM intercept.

The null hypothesis for T1 for an intercept is the same as that for the slopes and consists of normally distributed regression errors. Rejection of the null hypothesis in T1 can occur due to inefficiency of LS estimator or due to bias in either of the estimators. The performance of the T1 test for an intercept is similar to that for the slopes described in Chapter 2 and is not considered in this chapter.

In Chapter 2, we found an optimal loss function to be inferior to a bisquare one when it comes to comparing regression slopes. The same holds for an intercept so this chapter focuses solely on a bisquare loss function to keep the amount of the presented material within reasonable size.

The remainder of the chapter follows the steps in Chapter 2 closely and is organized as follows. Sections 5.2 and 5.3 present the test statistics and discuss the families of the null and alternative hypothesis distributions. Analytical power for the tests under some of the alternatives is given in Section 5.4. Monte Carlo simulations of validity of level and power of the tests are presented in Section 5.5 for the case of a simple linear regression. For an empirical example a reader is referred to Section 4.7 in Chapter 4.

Section 5.6 concludes the chapter by summarizing the main results. Technical details are deferred to Appendix A5.

5.2 Test for Differences between LS and MM Intercept Estimators

The test is designed to test the following composite hypothesis H0 versus one of the following two alternatives K1 and K2:

H0: The distribution of the data $(y_i, \boldsymbol{x}_i)$ is given by (1.3) with symmetric error term distribution F_ϵ

K1: The distribution of the data $(y_i, \boldsymbol{x}_i)$ is given by (1.3) with asymmetric error term distribution F_ϵ

K2: The distribution of the data $(y_i, \boldsymbol{x}_i)$ is given by (1.4) with an α_0-bias inducing distribution $H(\boldsymbol{x}, y)$.

Under H0 not only LS and MM slopes are consistent estimators of $\boldsymbol{\beta_0}$, but also both LS and MM intercepts are consistent for α_0, center of symmetry of F_ϵ. Under K1 and K2, however, the asymptotic values of the LS and MM intercept estimators are different and if the difference is large enough it should be detectable.

It was shown in Appendix A2 (Chapter 2) that under H0 the following asymptotic normality result holds

$$\sqrt{n}\,(\hat{\alpha}_{MM} - \hat{\alpha}_{LS}) \rightarrow \boldsymbol{N}\left(0, \delta^2_{LS,MM}\left(1 + \boldsymbol{\mu}_{\boldsymbol{x}}^{'}\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{x}}\right)\right) \tag{5.1}$$

with $\delta^2_{LS,MM}$ defined in (2.8).

By combining LS and MM regression coefficient estimates with estimates $\widehat{\delta^2_{LS,MM}}$ (2.13) and $\widehat{\boldsymbol{V}}_{\boldsymbol{x}^*}$ (1.30) one obtains the following test statistic:

$$T_{LS,MM} = \frac{\hat{\alpha}_{MM} - \hat{\alpha}_{LS}}{\sqrt{\frac{1}{n}\widehat{\delta^2_{LS,MM}}\hat{V}^{-1}_{x^*,11}}} \tag{5.2}$$

where $\hat{V}^{-1}_{x^*,11}$ is equal to the 1-st diagonal element of $\widehat{\boldsymbol{V}}^{-\mathbf{1}}_{\boldsymbol{x}^*}$.

Under H0 the $T_{LS,MM}$ will have an approximate standard normal distribution.

5.3 Test for Differences between c-MM and MM Intercept Estimators

The test is designed to test the same composite null hypothesis H0 versus the alternatives K1 and K2 as test $T_{LS,MM}$ in the previous section.

From the joint asymptotic normality result in Appendix A4 (Chapter 4) it follows that under H0

$$\sqrt{n}\,(\hat{\alpha}_{cMM} - \hat{\alpha}_{MM}) \rightarrow \boldsymbol{N}\left(0, \delta^2_{cMM,MM}\right) \tag{5.3}$$

with

$$\delta^2_{cMM,MM} = E\left\{\left(\frac{\sigma\psi_c\left(\frac{u_c}{\sigma}\right)}{E\psi'_c\left(\frac{u_c}{\sigma}\right)} - \frac{\sigma\psi_{MM}\left(\frac{u_{MM}}{\sigma}\right)}{E\psi'_{MM}\left(\frac{u_{MM}}{\sigma}\right)}\right)^2\right\} \tag{5.4}$$

Note that there is no $\boldsymbol{\mu}'_{\boldsymbol{x}}\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{x}}$ term in (5.3), which means that distribution of the differences between c-MM and MM intercept estimators does not depend on the distribution of the predictors $\boldsymbol{x}$. This important feature is a consequence of the two intercept estimates being based on the same slope estimates $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$. This is in contrast to the variance of the differences between MM and LS intercepts (5.1) , where part of the variability is due to the differences in the MM and LS slope estimates.

By combining c-MM and MM regression coefficient estimates with estimate $\widehat{\delta^2_{cMM,MM}}$ one obtains the following test statistic:

$$T_{cMM,MM} = \frac{\hat{\alpha}_{MM} - \hat{\alpha}_{cMM}}{\sqrt{\frac{1}{n}\widehat{\delta^2_{cMM,MM}}}} \tag{5.5}$$

We estimate $\delta^2_{cMM,MM}$ as

$$\widehat{\delta^2_{cMM,MM}} = ave_i\left(\frac{\hat{\sigma}_1\psi_c\left(r_i^{cMM}/\hat{\sigma}_1\right)}{ave_j\psi_c'\left(r_j^{cMM}/\hat{\sigma}_1\right)} - \frac{\hat{\sigma}_1\psi_2\left(r_i^{MM}/\hat{\sigma}_1\right)}{ave_j\psi_2'\left(r_j^{MM}/\hat{\sigma}_1\right)}\right)^2 \tag{5.6}$$

where $r_i^{cMM} = y_i - \hat{\alpha}_{cMM} - \boldsymbol{x_i'}\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$ and $r_i^{MM} = y_i - \hat{\alpha}_{MM} - \boldsymbol{x_i'}\widehat{\boldsymbol{\beta}}_{\boldsymbol{MM}}$ denote the c-MM and robust MM residuals.

Under H0 the $T_{cMM,MM}$ will have an approximate standard normal distribution.

5.4 Power under Alternative K1

From the joint asymptotic normality result in Appendix A4 (Chapter 4) it follows that under K1:

$$\sqrt{n}\left(\hat{\alpha}_{MM} - \hat{\alpha}_{cMM} - (\alpha_{MM} - \alpha_{cMM})\right) \rightarrow \boldsymbol{N}\left(0, \delta^2_{cMM,MM} + \omega_{cMM,MM}\right) \tag{5.7}$$

where $\omega_{cMM,MM}$ has a long expression given by (A5.2) in Appendix A5.

Since $\hat{\delta}^2_{cMM,MM}$ is a consistent estimator of $\delta^2_{cMM,MM}$ we have $\frac{\hat{\delta}^2_{cMM,MM}}{\delta_{cMM,MM}} \rightarrow 1$ and by Slutsky's theorem

$$\frac{\sqrt{n}\left(\hat{\alpha}_{MM} - \hat{\alpha}_{cMM} - (\alpha_{MM} - \alpha_{cMM})\right)}{\sqrt{\delta^2_{cMM,MM}}\sqrt{\frac{\hat{\delta}^2_{cMM,MM}}{\delta^2_{cMM,MM}}}} \rightarrow \boldsymbol{N}\left(0, \frac{\delta^2_{cMM,MM} + \omega_{cMM,MM}}{\delta^2_{cMM,MM}}\right) \tag{5.8}$$

Thus, under K1

$$T_{cMM,MM} \sim N\left(\frac{\sqrt{n}(\alpha_{MM} - \alpha_{cMM})}{\sqrt{\delta^2_{cMM,MM}}}, \frac{\delta^2_{cMM,MM} + \omega_{cMM,MM}}{\delta^2_{cMM,MM}}\right) \tag{5.9}$$

Similarly, for LS from asymptotic result (A2.12) we have that under K1

$$T_{LS,MM} \sim N\left(\frac{\sqrt{n}(\alpha_{MM}-\alpha_{LS})}{\sqrt{\delta^2_{LS,MM}\left(1+\boldsymbol{\mu}_{\boldsymbol{x}}'\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{x}}\right)}}, \frac{\delta^2_{LS,MM}\left(1+\boldsymbol{\mu}_{\boldsymbol{x}}'\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{x}}\right)+\omega_{LS,MM}}{\delta^2_{LS,MM}\left(1+\boldsymbol{\mu}_{\boldsymbol{x}}'\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{x}}\right)}\right) \quad (5.10)$$

with $\omega_{LS,MM}$ defined in (A2.14).

The power is now straightforward after noting that we reject when $|T_{cMM,MM}| \geq q_{1-\alpha/2}$ (or $|T_{LS,MM}| \geq q_{1-\alpha/2}$) , where $q_{1-\alpha/2}$ is a $(1-\alpha/2)$-quantile of standard normal distribution. Comparing equations (5.9) and (5.10) one sees that $T_{cMM,MM}$ power does not depend on the distribution of the predictors $\boldsymbol{x}$, but $T_{LS,MM}$ power depends on the mean and covariance of the predictors $\boldsymbol{x}$ via the term $\boldsymbol{\mu}_{\boldsymbol{x}}'\boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}}\boldsymbol{\mu}_{\boldsymbol{x}}$. Our simulations confirm this result. Specifically, the variances of the differences between c-MM and MM intercepts were the same for predictor $\boldsymbol{x}$ with mean equal to zero and mean equal to one. The differences between LS and MM intercepts, on the other hand, had larger variance in the case of a non-zero mean compared to a zero mean.

In Figure 66 we consider errors following a skewed-t distribution of Azzalini and Capitanio (2003) and plot the $T_{LS,MM}$ power as a function of sample size for five degrees of freedom and two values of the skewness parameter lambda, $\lambda = 1$ (top row) and $\lambda = 3$ (bottom row). In this case an MM intercept is a biased estimator of the mean of the error term in contrast to an unbiased LS intercept. As normal distribution efficiency of the MM estimate increases its bias gets smaller and so does the rejection rate in the test for the differences between LS and MM intercepts. Also, the larger the skewness the larger the MM intercept bias and the larger the rejection rate. Note much larger rejection rate for the case of the predictors with a zero mean (left column) versus a non-zero mean (right column) and see our discussion on the effect of the centering of the predictor vector in Section 4.4.1. The power increases with sample size and is equal to one asymptotically.

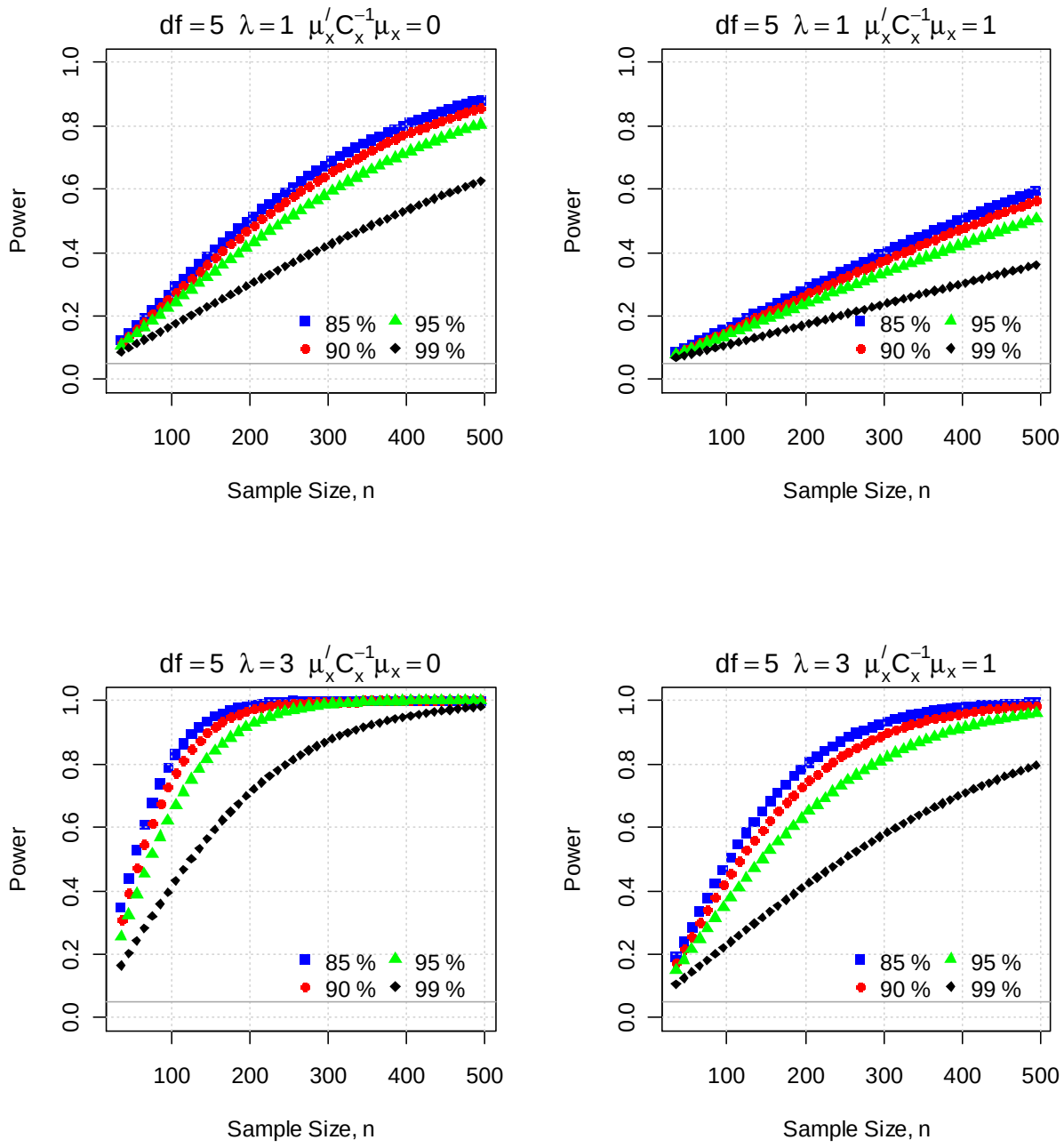

Figure 66. $T_{LS,MM}$ power under skewed-t error distributions.

We now look at the power of the $T_{cMM,MM}$ with the bisquare psi functions with c=4.68 (95% normal distribution efficiency) and c=15 for the MM and c-MM intercepts respectively. Recall that this power does

not depend on $\boldsymbol{\mu_x' C_x^{-1} \mu_x}$**.** The results for a skewed-t distribution with 5 degrees of freedom are shown in Figure 67 where power is plotted versus skewness parameter lambda for a set of sample sizes from 50 to 500. In this case the $T_{cMM,MM}$ power is very similar to that of $T_{LS,MM}$ when $\boldsymbol{\mu_x = 0}$**.**

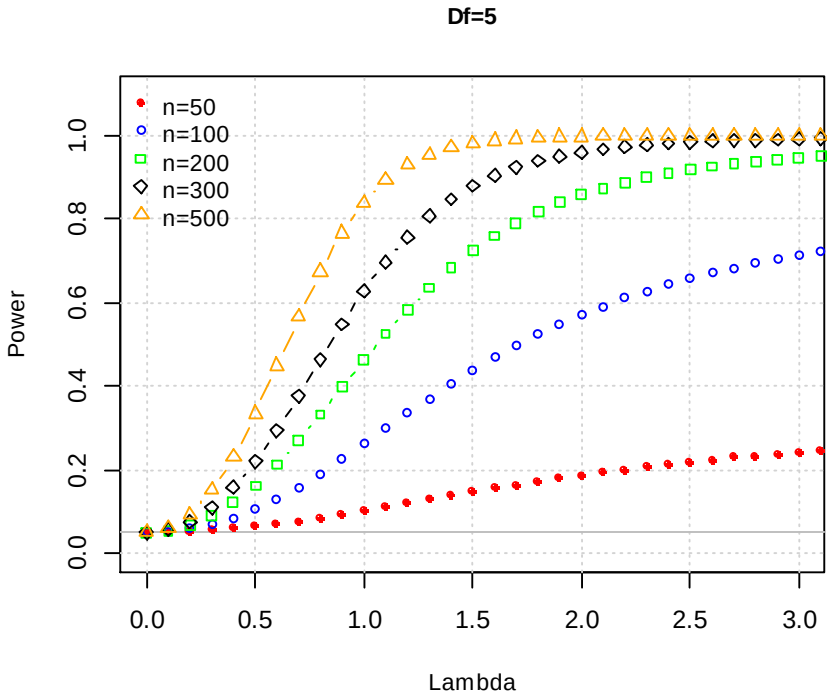

Figure 67. $T_{cMM,MM}$ power under skewed-t error distribution with 5 degrees of freedom as a function of a skewness parameter lambda for five values of sample sizes. MM and c-MM intercepts are based on a bisquare psi-function with c=4.68 (95% normal distribution efficiency) and c=15 respectively.

Figure 66 and Figure 67 are based on the normal distribution approximations (5.9) and (5.10). In Section 5.5 we check via simulations how this approximation works for the case of df=5 and $\lambda = 1$.

$\boldsymbol{T_{\infty MM,MM}}$ *versus* $\boldsymbol{T_{LS,MM}}$

By ∞MM intercept estimate in Chapter 4 we denoted an arithmetic average of the MM residuals. The asymptotic variance of the differences between ∞MM and MM estimators can be obtained from (5.7) with $\psi_\infty(z) = z$. Recall that under K1 both ∞MM and LS intercepts are consistent so that we have $\alpha_{\infty MM} = \alpha_{LS} = \alpha_0 = E\epsilon,\ u_c = u_{LS} = \epsilon - E\epsilon,\ u_{MM} = \epsilon - \alpha_{MM},$

$$\delta^2_{LS,MM} = \delta^2_{\infty MM,MM} = E\left\{\left(\epsilon - E\epsilon\ - \frac{\sigma\psi_2((\epsilon-\alpha_{MM})/\sigma)}{a_{MM}}\right)^2\right\}$$

and

$$\omega_{LS,MM} = \omega_{\infty MM,MM} = \sigma^2\left\{\left(\frac{e_{MM}}{a_{MM}}\right)^2 \frac{E\left(\rho_1^2\left(\frac{u_S}{\sigma}\right) - b^2\right)}{d_s^2} + \frac{2}{d_s}\frac{e_{MM}}{a_{MM}}\left(E\left(\rho_1\left(\frac{u_S}{\sigma}\right)\frac{(\epsilon - E\epsilon)}{\sigma}\right) - \frac{g_{MM}}{a_{MM}}\right)\right\}$$

With this result in mind we look back at the equations (5.9) and (5.10) to conclude that when $\boldsymbol{\mu_x} = \mathbf{0}$ the tests $T_{LS,MM}$ and $T_{\infty MM,MM}$ have the same power. Therefore, the left column in Figure 66 also shows the $T_{\infty MM,MM}$ power under skewed-t errors. When the mean of the predictor vector is non-zero the $T_{\infty MM,MM}$ power does not change (as compared to $\boldsymbol{\mu_x} = \mathbf{0}$) while $T_{LS,MM}$ power would typically be smaller. For example, compare the right column in Figure 66 which shows $T_{LS,MM}$ power when $\boldsymbol{\mu_x'C_x^{-1}\mu_x} = 1$ with the left column that shows the power when $\boldsymbol{\mu_x} = \mathbf{0}$**.** This suggests that one may prefer to use $T_{\infty MM,MM}$ rather than $T_{LS,MM}$ and not worry about centering the predictors.

5.5 Monte Carlo Simulations

In order to understand the finite sample behavior of the level and power of our tests $T_{LS,MM}$ and $T_{cMM,MM}$ we carried out a number of Monte Carlo simulation studies with large-sample significance level α set at 0.05. We consider intercept in a simple linear regression and the five distribution models from Section 2.4 in Chapter 2 with the only difference of using a non-zero mean for a distribution of the predictor x:

- In Model 1 through 4 we use $x \sim Normal(1,1)$
- In Model 5 we use $Normal(1,1)$ for the main mixture component and $Normal(3, 0.25^2)$ for the contamination mixture component.

Simulations were conducted in R using lmRob function from the R robust library. For the test statistics we wrote additional R code that is available upon request.

5.5.1 $T_{LS,MM}$ and $T_{\infty MM,MM}$

Figure 68 through Figure 72 show simulation results for the tests $T_{\infty MM,MM}$ and $T_{LS,MM}$. A non-zero mean of x was chosen to illustrate the differences between $T_{\infty MM,MM}$ whose power does not depend on the distribution of the predictors and $T_{LS,MM}$ whose power depends on $\boldsymbol{\mu_x' C_x^{-1} \mu_x}$**.** We also ran simulations for x with a standard normal distribution and found out that for all five models the $T_{LS,MM}$ and $T_{\infty MM,MM}$ had similar rejection rates which are similar to those of $T_{\infty MM,MM}$ for $\mu_x = 1$ that are shown in Figure 68 through Figure 72 below.

Model 1 (normal distribution errors) and Model 2 (t-distribution errors). Normal and symmetric t-distribution with five degrees of freedom are in the null hypothesis H0. Figure 68 and Figure 69 display Monte Carlo level versus sample size for the bisquare loss function and for the four normal distribution efficiencies of 85%, 90%, 95% and 99%. The $T_{\infty MM,MM}$ rejection rate is noticeably below 0.05 in sample sizes below 200 (and in case of a t-distribution and 99% efficiency even in sample size 500). The $T_{LS,MM}$ rejection rate is close to the nominal significance level of 0.05 for all sample sizes and all efficiencies. The latter is the effect of a non-zero mean of x. The $T_{LS,MM}$ rejection rate for x with zero mean exhibits the same instability as observed for $T_{\infty MM,MM}$.

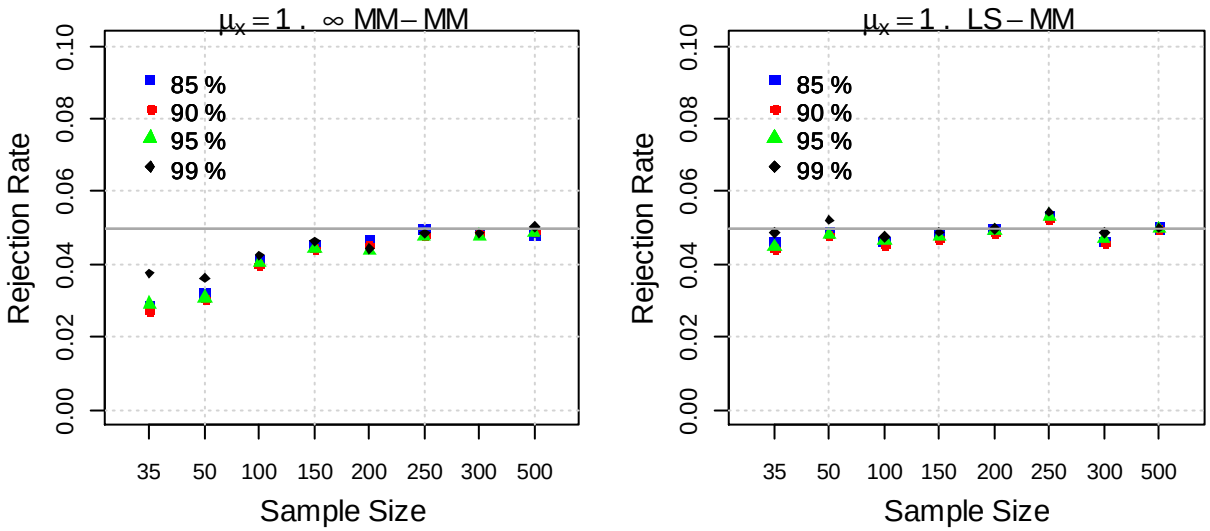

Figure 68. Model 1. Level of the $T_{\infty MM,MM}$ (left) and $T_{LS,MM}$ (right) test statistics for intercept in a simple linear regression under normal residuals. Grey horizontal line is at large-sample significance level of 0.05.

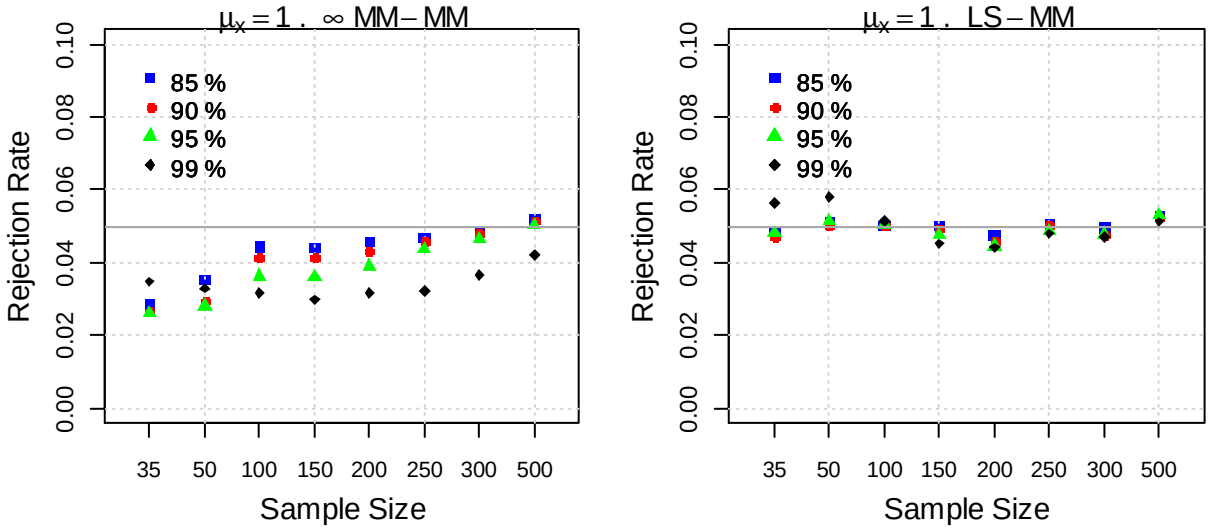

Figure 69. Model 2. Level of the $T_{\infty MM,MM}$ (left) and $T_{LS,MM}$ (right) test statistics for the intercept in a simple linear regression under symmetric t_5 residuals. Grey horizontal line is at large-sample significance level of 0.05.

Model 3 (skewed t-distribution errors). Rejection rates for skewed t-distribution with five degrees of freedom are presented in Figure 70. This distribution is in the alternative K1 and, thus, it is not surprising to see that both tests have power which increases with sample size.

The standard errors of $\hat{\alpha}_{LS} - \hat{\alpha}_{MM}$ tend to be larger than those of $\hat{\alpha}_{\infty MM} - \hat{\alpha}_{MM}$ due to an extra positive term ($\boldsymbol{\mu}_x' \boldsymbol{C}_x^{-1} \boldsymbol{\mu}_x$) in the variance. As a result, the $T_{LS,MM}$ power in Figure 70 is smaller than $T_{\infty MM,MM}$ power.

Low power of the tests for smaller sample sizes is in itself an indicator of an MM intercept bias being small compared to variability of the estimates.

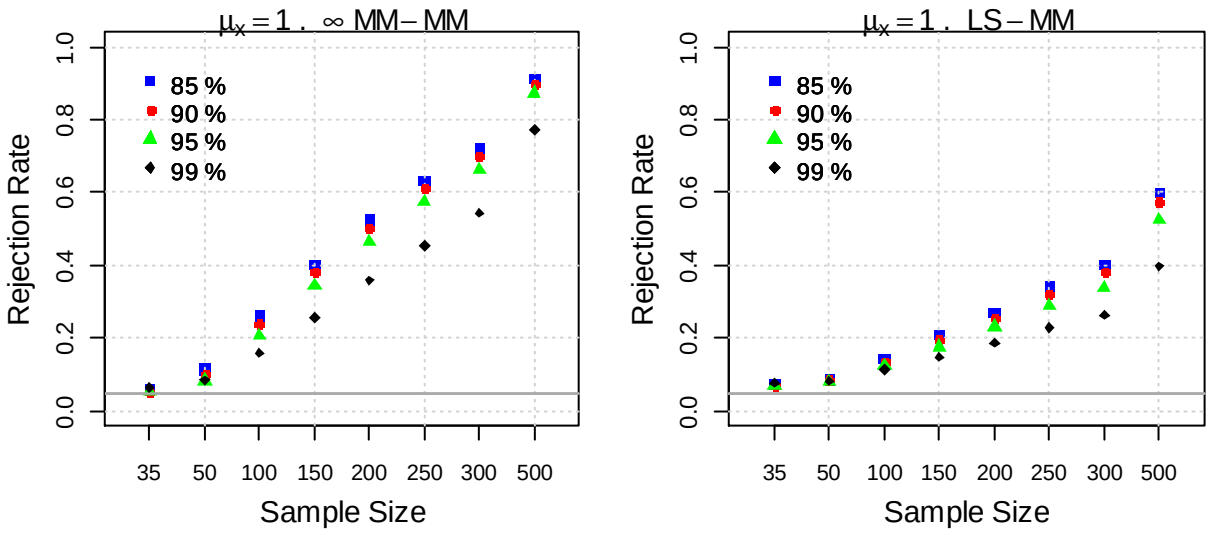

Figure 70. Model 3. Rejection rates of the $T_{\infty MM,MM}$ (left) and $T_{LS,MM}$ (right) test statistics for the intercept in a simple linear regression under skewed t_5 residuals. Grey horizontal line is at large-sample significance level of 0.05.

Model 4 (asymmetric normal mixture distribution errors) and Model 5 (bivariate normal mixture distribution).

Figure 71 and Figure 72 display Monte Carlo rejection rates versus sample size and normal distribution efficiency for mixing proportions 0.02, 0.04 and 0.06. The square, diamond and triangle symbols represent μ values of 4, 7 and 14 respectively. Model 4 and Model 5 are in the alternative K1 and K2 respectively. In sample sizes 50 and under the rejection rates for both tests are very low. In larger sample sizes the $T_{LS,MM}$ power is lower than that of $T_{\infty MM,MM}$. This is again a consequence of a non-zero mean of the predictor x. For x from a normal distribution with zero mean the rejection rates for both tests are visibly indistinguishable from each other and are similar to those seen for $T_{\infty MM,MM}$ in Figure 71 and Figure 72.

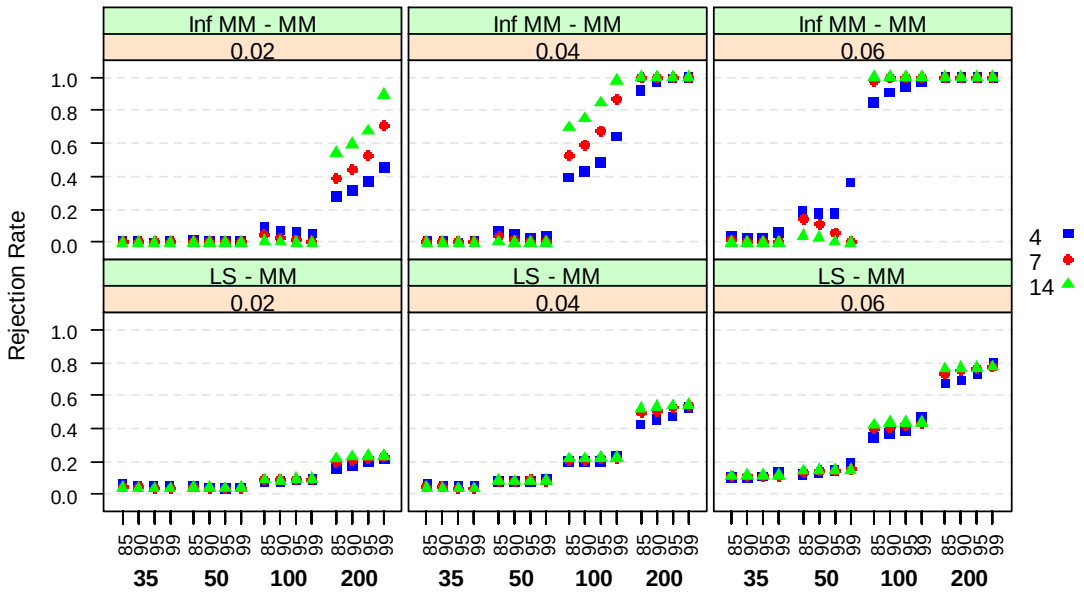

Figure 71. Model 4. Rejection rates of the $T_{\infty MM,MM}$ (top) and $T_{LS,MM}$ (bottom) test statistics for intercept in a simple linear regression under asymmetric residual contamination. Blue squares correspond to mild outliers $N(\mu = 4, 0.25^2)$, red diamonds – to medium outliers $N(\mu = 7, 0.25^2)$ and green triangles – to gross outliers $N(\mu = 14, 0.25^2)$.

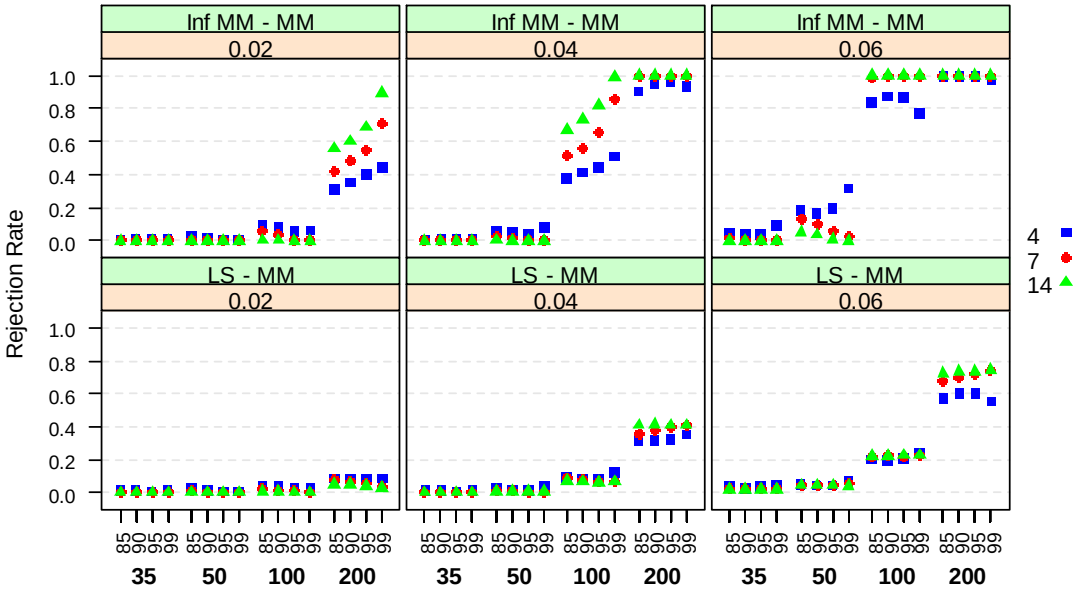

Figure 72. Model 5. Rejection rates of the $T_{\infty MM,MM}$ (top) and $T_{LS,MM}$ (bottom) test statistics for intercept in a simple linear regression under bivariate asymmetric contamination. Blue squares correspond to mild outliers $N(\mu = 4, 0.25^2)$, red diamonds – to medium outliers $N(\mu = 7, 0.25^2)$ and green triangles – to gross outliers $N(\mu = 14, 0.25^2)$.

5.5.2 $T_{cMM,MM}$

In this section we present simulation results for a test for differences in c-MM and MM intercepts. We consider bisquare loss function with c=4.68 (95% normal distribution efficiency) for an MM estimator and c=15 for a c-MM estimator. A test for differences between ∞MM and MM intercepts is used as a baseline for comparison.

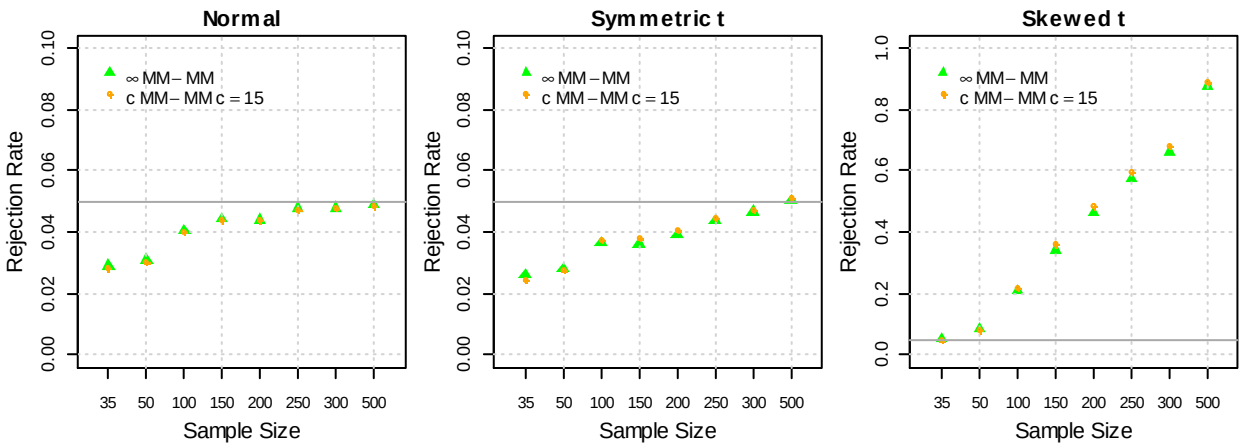

Figure 73. Rejection rates of the $T_{\infty MM,MM}$ (green triangles) and $T_{cMM,MM}$ (orange dots) test statistics for the intercept in a simple linear regression under normal (Model 1), symmetric t_5 (Model 2) and skewed t_5 (Model 3) residuals. Grey horizontal line is at large-sample significance level of 0.05.

Results for Model 1 (normal), Model 2 (symmetric t_5) and Model 3 (skewed t_5) are shown in Figure 73. For all sample sizes the rejection rates in $T_{cMM,MM}$ test are virtually the same as those in $T_{\infty MM,MM}$ test. Normal and symmetric t_5 errors are in a null hypothesis H0. Monte Carlo level is close to 0.05 in large sample sizes (250 and over for normal errors; 300 and over for t_5 errors), but is, unfortunately, below 0.05 in small to moderate sample sizes.

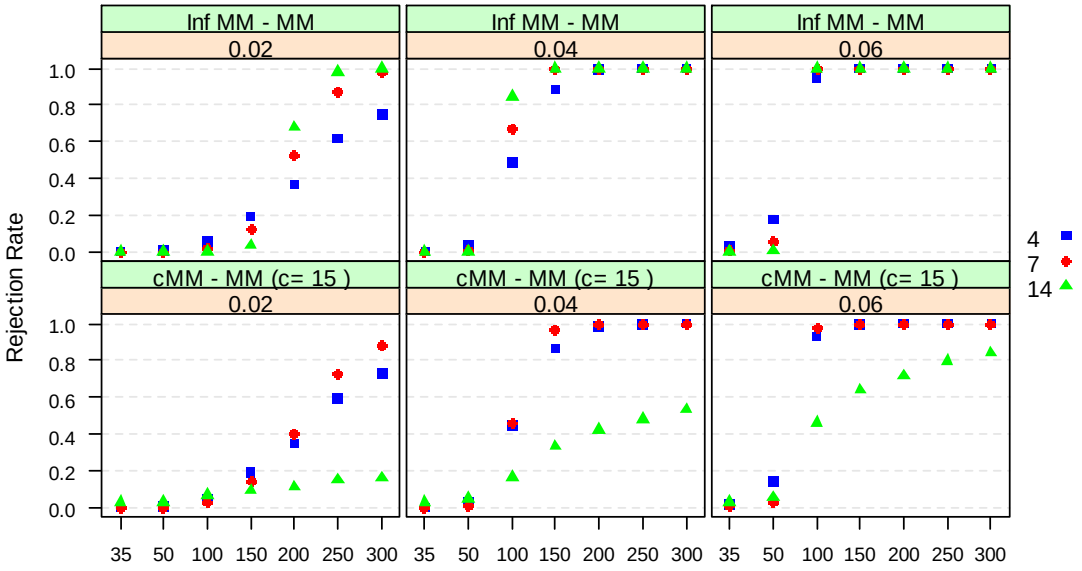

Figure 74. Model 4. Rejection rates of the $T_{\infty MM,MM}$ (top) and $T_{LS,MM}$ (bottom) test statistics for intercept in a simple linear regression under asymmetric residual contamination. Blue squares correspond to mild outliers $N(\mu = 4, 0.25^2)$, red diamonds – to medium outliers $N(\mu = 7, 0.25^2)$ and green triangles – to gross outliers $N(\mu = 14, 0.25^2)$.

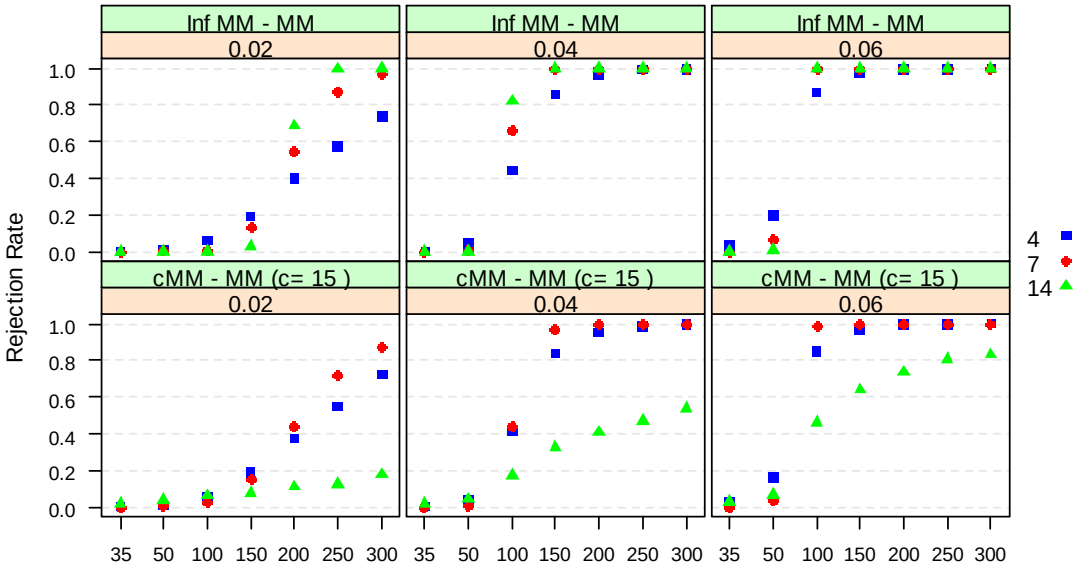

Figure 75. Model 5. Rejection rates of the $T_{\infty MM,MM}$ (top) and $T_{LS,MM}$ (bottom) test statistics for intercept in a simple linear regression under bivariate asymmetric contamination. Blue squares correspond to mild outliers $N(\mu = 4, 0.25^2)$, red diamonds – to medium outliers $N(\mu = 7, 0.25^2)$ and green triangles – to gross outliers $N(\mu = 14, 0.25^2)$.

Results for Model 4 (asymmetric normal mixture distribution) and Model 5 (bivariate normal mixture distribution) are shown in Figure 74 and Figure 75 respectively. Model 4 and Model 5 are in the alternative K1 and K2 respectively. The most noticeable feature is a much lower rejection rate in $T_{cMM,MM}$ for $\mu = 14$ as compared to $T_{\infty MM,MM}$. Observations generated by a mixture component $N(14, 0.25^2)$ are close to the rejection region of a c-MM estimator with c=15 and, therefore, have very low weights. Consequently, the bias of a c-MM estimator is much lower at $\mu = 14$ than it is for our mixture models with $\mu = 4$ and $\mu = 7$. Hence, lower power.

5.6 Summary and Discussion

This chapter completes our work on statistical tests for comparison of the LS and robust MM regression estimators by proposing a test for the differences in two intercepts. The proposed test $T_{LS,MM}$ is a Wald-type test based on a derived asymptotic distribution of the differences between LS and MM intercept estimators under symmetric error distribution. Test $T_{LS,MM}$ has the same form as test T2 for the slopes in Chapter 2. Rejection in test $T_{LS,MM}$ occurs due to one of the estimators having a sufficiently larger bias. It is recommended to center predictors before computing this test statistic to improve power.

Test $T_{cMM,MM}$ for differences between a corrected c-MM intercept (defined in Chapter 4) and MM intercept was also proposed and studied. This is also a Wald-type test which shares the same composite null and alternative hypotheses with a test $T_{LS,MM}$. An interesting special case is a $T_{\infty MM,MM}$ test for a non-robust ∞MM intercept. Since $T_{cMM,MM}$ power does not depend on location and covariance of the vector of predictors, it is not necessary to center predictors. When all predictors have zero mean, the $T_{\infty MM,MM}$ power is the same as that of $T_{LS,MM}$, and is typically larger for predictors with a non-zero mean.

Unfortunately, these tests alone do not tell us if the observed large difference in two intercept estimates is caused by an MM bias due to, for example, skewness of the main component or by an LS bias due to contamination. Note that we would want to capture the former (any skewness of the main component), but ignore the latter (outliers from contamination component). Nonetheless, these tests may serve as an additional diagnostic tool where rejection in a test warrants extra examination of the data.

Appendix A5: Asymptotic Distribution of the Differences between Intercept Estimators

From the joint asymptotic normality result for $(\hat{\alpha}_{cMM}, \hat{\alpha}_{MM})$ in Appendix A4 in Chapter 4 it follows that the difference $\hat{\alpha}_{cMM} - \hat{\alpha}_{MM}$ is also asymptotically normal with the variance

$$V_{\hat{\alpha}_{cMM}} - 2cov(\hat{\alpha}_{cMM}, \hat{\alpha}_{MM}) + V_{\hat{\alpha}_{MM}}$$

Straightforward calculations give

$$\begin{aligned} &cov(\hat{\alpha}_{cMM}, \hat{\alpha}_{MM}) \\ &\quad = \sigma^2 \left\{ \frac{f_{cMM,MM}}{a_{MM} a_{cMM}} - \frac{e_{MM} g_{cMM}}{a_{MM} a_{cMM} d_S} + \frac{f_{MM}}{a_{MM}^2} \boldsymbol{\mu}_{\boldsymbol{x}}' \boldsymbol{C}_{\boldsymbol{x}}^{-\mathbf{1}} \boldsymbol{\mu}_{\boldsymbol{x}} - \frac{g_{MM} e_{cMM}}{a_{MM} a_{cMM} d_S} \right. \\ &\quad \left. + \frac{e_{MM} e_{cMM}}{a_{MM} a_{cMM} d_S^2} E\left(\rho_0^2 \left(\frac{u_S}{\sigma} \right) - b^2 \right) \right\} \end{aligned}$$

and

$$\sqrt{n}\,(\hat{\alpha}_{MM} - \hat{\alpha}_{cMM} - (\alpha_{MM} - \alpha_{cMM})) \rightarrow \boldsymbol{N}\left(0, \delta_{cMM,MM}^2 + \omega_{cMM,MM}\right) \tag{A5.1}$$

where $\delta_{cMM,MM}^2$ is defined in (5.4) and

$$\begin{aligned} \omega_{cMM,MM} = \sigma^2 \left\{ \left(\frac{e_{MM}}{a_{MM}} - \frac{e_{cMM}}{a_{cMM}} \right)^2 \frac{E\left(\rho_1^2 \left(\frac{u_S}{\sigma} \right) - b^2 \right)}{d_s^2} \right. \\ \left. + \frac{2}{d_s} \left(\frac{e_{MM}}{a_{MM}} - \frac{e_{cMM}}{a_{cMM}} \right) \left(\frac{g_{cMM}}{a_{cMM}} - \frac{g_{MM}}{a_{MM}} \right) \right\} \end{aligned} \tag{A5.2}$$

We note that under symmetric F_ϵ the extra term $\omega_{cMM,MM}$ is equal to zero since symmetric loss function leads to $e_{MM} = e_{cMM} = 0$.

Bibliography

Aas, K., & Haff, I. H. (2006). The Generalized Hyperbolic Skew Student's t-Distribution. *Journal of Financial Econometrics, 4*(2), 275-309.

Arellano-Valle, R. B., & Azzalini, A. (2011). The Centred Parameterization and Related Quantities of the Skew-t Distribution. *Journal of Multivariate Analsysis, 113*, 73–90.

Arellano-Valle, R. B., & Genton, M. G. (2010). Multivariate Extended Skew-t Distributions and Related Families. *International Journal of Statistics, LXVIII*(3), 201-234.

Azzalini, A. (2005). The Skew-Normal Distribution and Related Multivariate Families (with discussion by Marc G. Genton and a rejoinder by the author). *Scandinavian Journal of Statistics*(32), 159-200.

Azzalini, A. (2011). sn: The skew-normal and skew-t distributions. R package version 0.4-17. Available from http://CRAN.R-project.org/package=sn.

Azzalini, A., & Capitanio, A. (1999). Statistical Applications of the Multivariate Skew-normal Distribution. *J. Roy. Statist. Soc. B, 61*(3), 579–602.

Azzalini, A., & Capitanio, A. (2003). Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew-t Distribution. *J. Roy. Statist. Soc. B, 65*, 367–389.

Azzalini, A., & Genton, M. G. (2008). Robust Likelihood Methods based on the Skew-t and Related Distributions. *International Statistical Review*(76), 106-129.

Bailer, H. M. (2005). *Robust Estimation of Factor Models in Finance.* PhD Dissertation, University of Washington, Seattle.

Bailer, H. M., Maravina, T. A., & Martin, R. D. (2012). Robust Betas in Asset Management. In B. Scherer, & K. Winston (Eds.), *The Oxford Handbook of Quantitative Asset Management* (pp. 203-242). Oxford University Press.

Becker, R. A., Cleveland, W. S., & Shyu, M. J. (1996). The Visual Design and Control of Trellis Graphics Displays. *Journal of Computational and Graphical Statistics, 5*(2), 123-156.

Bianco, A. M., Ben, M. G., & Yohai, V. J. (2005). Robust Estimation for Linear Regression with Asymmetric Errors. *The Canadian Journal of Statistics, 33*(4), 511-528.

Bowie, D. C., & Bradfield, D. J. (1998). Robust Estimation of Beta Coefficients: Evidence from a Small Stock Market. *Journal of Business Finance & Accounting, 25*(3&4), 439-454.

Branco, M. D., Genton, M. G., & Liseo, B. (2012). Objective Bayesian Analysis of Skew-t Distributions. *Scandinavian Journal of Statistics, (to appear)*.

Branco, M., & Dey, D. K. (2001). A General Class of Multivariate Skew Elliptical Distributions. *Journal of Multivariate Analysis*(79), 99-113.

Bulmer, M. (1979). *Principles of Statistics.* Dover Publications.

Chan, L. K., & Lakonishok, J. (1992). Robust Measurement of Beta Risk. *Journal of Financial and Quantitative Analysis, 27*(2), 265-282.

Cook, R. D., & Weisberg, S. (1994). *An Introduction to Regression Graphics.* John Wiley.

Cornell, B., & Dietrich, J. K. (1978). Mean-Absolute-Deviation versus Least-Squares Regression Estimation of Beta Coefficients. *Journal of Financial and Quantitative Analysis, 13*, 123-131.

Draper, P., & Paudyal, K. (1995). Empirical Irregularities in the Estimation of Beta: the Impact of Alternative Estimation Assumptions and Procedures. *Journal of Business Finance & Accounting, 22*(1), 157-177.

Fasano, M. V., Maronna, R. A., Sued, M., & Yohai, V. J. (2012). Continuity and Differentiability of Regression M Functionals. *To appear in Bernoulli*.

Fernandez, C., & Steel, M. F. (1998, March). On Bayesian Modeling of Fat Tails and Skewness. *Journal of the American Statistical Association, 93*(441), 359-371.

Fernandez, C., & Steel, M. F. (1999). Multivariate Student-t Regression Models: Pitfalls and Inference. *Biometrika*(86), 153-167.

Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika*(80), 27-38.

Fong, W. M. (1997). Robust Beta Estimation: Some Empirical Evidence. *Review of Financial Economics, 6*(2), 167-186.

Fonseca, T. C., Ferreira, M. A., & Migon, H. S. (2008). Objective Bayesian Analysis for the Student-t Regression Model. *Biometrika, 95*, 325-333.

French, K. R. (2012). Kenneth French Data Library. Available from http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (last accessed in December 2012).

Genton, M. G. (2003). Breakdown-point for Spatially and Temporally Correlated Observations. In R. Dutter, P. Filzmoser, U. Gather, & P. J. Rousseeuw (Eds.), *Developments in Robust Statistics, International Conference on Robust Statistics 2001* (pp. 148-159). Heidelberg: Springer.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality.* Chapman & Hall.

Gray, J. B. (1985). Graphics for Regression Diagnostic. *American Statistical Association Proceedings of the Statistical Computing Section*, 102-107.

Grinold, R., & Kahn, R. (1999). *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk* (2nd ed.). New York: McGraw-Hill.

Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *Ann. Math. Statist, 42*, 1887-1896.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* New York: John Wiley & Sons, Inc.

Hansen, B. (1994). Autoregressive Conditional Density Estimation. *International Economic Review*(35), 705-730.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica, 46*(6), 1251-1271.

Hawkins, D. M., Bradu, D., & Kass, G. V. (1984). Location of Several Outliers in Multiple-Regression Data Using Elemental Sets. *Technometrics, 26*(3), 197-208.

Hodges, J. L. (1967). Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location. *Proceedings of the Fifth Berkeley Symp. Math. Statist. and Probab., 1*, pp. 163-168.

Hoorelbeke, D., Dewachter, H., & Smedts, K. (2005). Robust Estimation of Jensen's Alpha. *ICORS.* Available from http://www.stat.jyu.fi/icors2005/icorsabstracts/hoorelbeke.pdf (last accessed in December 2012).

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics, 35*, 73-101.

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics, 1*, 799-821.

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

Huber, P. J., & Ronchetti, E. M. (2009). *Robust Statistic* (2nd ed.). John Wiley & Sons.

Jensen, M. C. (1968). The Performance of Mutual Funds in the Period 1945-1964. *Journal of Finance, 23*(2), 389-416.

Jones, M. C., & Faddy, M. J. (2003). A Skew Extension of the t-Distribution with Applications. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 65*(1), 159-174.

Koller, M., & Stahel, W. A. (2011). Sharpening Wald-Type Inference in Robust Regression for Small Samples. *Computational Statistics and Data Analysis, 55*(8), 2504–2515.

Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer.

Marazzi, A. (1993). *Algorithms, Routines, and S Functions for Robust Statistics.* Pacific Grove, CA: Wadsworth & Brooks/Cole.

Marazzi, A., & Barbati, G. (2002). Robust Parametric Means of Asymmetric Distributions: Estimation and Testing. *Estadistica, 54*, 47-72.

Marazzi, A., & Yohai, V. J. (2004). Adaptively Truncated Maximum Likelihood Regression with Asymmetric Errors. *Journal of Statistical Planning and Inference*(121), 271-291.

Marchenko, Y. V. (2010). *Multivariate Skew-t Distributions in Econometrics and Environmetrics.* PhD Dissertation, Texas A&M University.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods.* John Wiley & Sons, Ltd.

Martin, R. D., & Simin, T. (1999). Robust Estimation of Beta. *Technical Report No. 350, University of Washington, Department of Statistics.* Available from http://www.stat.washington.edu/research/reports/1999/tr350.pdf (last accessed in December 2012).

Martin, R. D., & Simin, T. T. (2003). Outlier-Resistant Estimates of Beta. *Financial Analysts Journal, 59*, 56-69.

Martin, R. D., Yohai, V. J., & Zamar, R. H. (1989). Min-Max Bias Robust Regression. *The Annals of Statistics, 17*, 1608-1630.

McDonald, J. B., Michelfelder, R. A., & Theodossiou, P. (2009). Robust Regression Estimation Methods and Intercept Bias: A Capital Asset Pricing Model Application. *Multinational Finance Journal, 13*(3/4), 293-321.

McDonald, J. B., Michelfelder, R. A., & Theodossiou, P. (2010). Robust Estimation with Flexible Parametric Distributions: Estimation of Utility Stock Betas. *Quantitative Finance, 10*(4), 375-387.

Mills, T. C., & Coutts, A. J. (1996). Misspecification Testings and Robust Estimation of the Market Model: Estimating Betas for the FT-SE Industry Baskets. *The European Journal of Finance, 2*, 319-331.

R Core Team. (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* New York: John Wiley & Sons, Inc.

Rousseeuw, P. J., & Yohai, V. J. (1984). Robust Regression by Means of S-Estimators. In F. W. Härdle, & R. D. Martin (Eds.), *Robust and Nonlinear Time Series, Lecture Notes in Statistics 26* (pp. 256-272). New York: Springer.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., & Maechler, M. (2012). robustbase: Basic Robust Statistics. R package version 0.9-2. Available from http://CRAN.R-project.org/package=robustbase.

Sahu, S. K., Dey, D. K., & Branco, M. D. (2003). A New Class of Multivariate Skew Distributions with Applications to Bayesian Regression Models. *The Canadian Journal of Statistics, 31*(2), 129-150.

Salibian-Barrera, M., & Omelka, M. (2010). Uniform Asymptotics for S- and MM-Regression Estimators. *Annals of the Institute of Statistical Mathematics, 62*, 897-927.

Salibian-Barrera, M., & Yohai, V. J. (2006). A Fast Algorithm for S-Regression Estimates. *Journal of Computational and Graphical Statistics, 15*(2), 414–427.

Salibian-Barrera, M., & Zamar, R. H. (2002). Bootstrapping Robust Estimates of Regression. *The Annals of Statistics, 30*(2), 556-582.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R.* Springer.

Sartori, N. (2006). Bias Prevention of Maximum Likelihood Estimates for Scalar Skew Normal and Skew t Distributions. *Journal of Statistical Planning and Inference*(136), 4259 – 4275.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *Journal of Finance, 19*, 425-442.

Sharpe, W. F. (1971). Mean-Absolute-Deviation Characteristic Lines for Securities and Portfolios. *Management Science, 18*(2), B1-B13.

Svarc, M., Yohai, V. J., & Zamar, R. H. (2002). Optimal Bias-Robust M-Estimates of Regression. *Statistical Data Analysis Based on the L1 Norm and Related Methods*, 191-200. (Y. Dodge, Ed.) Basle: Birkhäuser.

Taylor, J. (2012). hett: Heteroscedastic t-regression. R package version 0.3-1. Available from CRAN.R-project.org/package=hett.

Theodossiou, P. (1998). Financial Data and the Skewed Generalized T Distribution. *Management Science, 44*(12), 1650-1661.

Tukey, J. W. (1979). Robust Techniques for the User. In R. L. Launer, & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 103-106). New York: Academic Press.

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., . . . Konis, K. (2012). robust: Insightful Robust Library. R package version 0.3-19. Available from http://CRAN.R-project.org/package=robust.

Yohai, V. J. (1987). High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics, 15*(2), 642-656.

Yohai, V. J., & Zamar, R. H. (1997). Optimal Locally Robust M-Estimates of Regression. *Journal of Statistical Planning and Inference, 57*, 73-92.

Yohai, V. J., Stahel, W. A., & Zamar, R. H. (1991). A Procedure for Robust Estimation and Inference in Linear Regression. In W. Stahel, & S. Weisberg (Eds.), *Directions in Robust Statistics and Diagnostics (Part II)* (pp. 365-374). New York: Springer.

Vita

Tatiana A. Maravina was born in Russia. In 2004, she earned specialist degree (equivalent to M.S.) with honors in Mathematical Statistics at M.V. Lomonosov Moscow State University, Russia. In 2012, she graduated with a Doctor of Philosophy in Statistics from the University of Washington in Seattle, WA.