1 Brief introduction

While the effect of outliers on the parameter estimates of a normal linear regression has been studied, it is of interest to investigate the effect of outliers and contamination of the error terms on the variability of the parameter estimates. Should outliers and contamination have significant effects on the variances of the estimates, then the values of often used test statistics such as $t-$ and Wald-statistics, the corresponding significance levels of the estimates, and ultimate conclusions drawn based on these outcomes can change drastically. Hence, circumventing such problems is of interest, since it is crucial for actors such as policymakers, organizations and scientists to minimize the probability of drawing incorrect conclusions and/or making undesirable policy decisions. With this paper we aim to investigate the effects on the variability of parameter estimates through inclusion of outliers and contamination of the error terms by heavy-tailed distribution by means of a simulation experiment.

2 Data

We implement a simulation study in which we estimate our models on $n = 1000$ data points in each replication. We randomly generate a set of explanatory variables (x_1, x_2, x_3) from a zero-mean multivariate normal distribution with $cov(x_i, x_j) = (1/2)^{|i-j|}$ for $i = 1, 2, 3$. Furthermore, we generate the linearly dependent variable y from a normal distribution with mean $x'\beta = x_1 + 0.5x_2 + 0.25x_3$ and covariance matrix $\Sigma_{ij} = cov(x_i, x_j)$, with the aforementioned covariances. Subsequently, we consider two types of distortions of the simulated data and we examine their effects on the variability of the parameter estimates $(\beta_1, \beta_2, \beta_3)$. These two types concern outliers, and contamination of the error terms by a heavy-tailed distribution.

Contamination with bad leverage points
First, we introduce contamination with outliers in the form of bad leverage points. That is, with probability $\epsilon = 0.05$ original data points of the above described uncontaminated data set $(y, x_1, x_2, x_3) := (y-2, x_1+2, x_2+2, x_3+2)$.

Contamination with t-distribution
Next, we introduce contamination of the error terms by a heavy-tailed distribution in the form of a $t(3)$-distribution. Error terms are contaminated with probability $\epsilon = 0.05$.

3 Methodology

We make use of a set of estimators to study the effect of outliers on the variability of the parameter estimates in the case of a normal linear regression. As a benchmark model we apply the Ordinary Least Squares (OLS) regression,

which is expected to produce strongly inflated estimates of the parameters' variability in case of contamination. Subsequently, we consider the Mallows-type M-estimator and the MM-estimator, which are established in order to provide more robust estimates. We investigate the performance of our estimators by replicating the analysis $R = 500$ times.

3.1 Mallows-type M-estimator

The Mallows-type M-estimator is obtained by solving the following equation:

$$\int \Psi(z;T)dF = 0 \tag{1}$$

where $\Psi(z;T) = \psi_H(y - x'\beta)w(x)$ is the score function with $w(x)$ representing the weighting function based on the Mahalanobis distance which bounds the influence of leverage outliers, while ψ, which is the Huber function, takes care of the vertical outliers.

3.2 MM-estimator

The MM-estimating comprises of the following two steps:

1. S-estimation In this step the residual scale estimate $\hat{\sigma}_S$ is retained where $\hat{\beta}_S = \arg\min_b \min \hat{\sigma}_M^2(b)$
2. M-estimation In this step the residual scale $\hat{\sigma}_S$ is fixed and is the obtained by solving $\frac{1}{n}\sum_{i=1}^{n} \rho(\frac{x_i}{\hat{\sigma}_S}) = \delta$ where δ is equal to the expected value of the loss function.

4 Results

4.1 Uncontaminated data

Figure 1 shows four boxplots containing all estimators on the uncontaminated data belonging to $\beta_0, \beta_1, \beta_2, \beta_3$, respectively. In Figure 1 through 3, 1 expresses the OLS estimator, 2 expresses the Mallows-type M-estimator and 3 expresses the MM-estimator. It can be concluded that the MM estimator produces the most accurate estimates of the parameters. In particular, the variability of the β_2 and β_3 estimates is visibly smaller for MM than for OLS and Mallows-type M. Moreover, the boxplots for OLS and Mallows-type M are of similar scale.

4.2 Contaminated data with t-distribution

Figure 2 shows four boxplots containing all estimators on the contaminated data with t-distribution on the error terms belonging to $\beta_0, \beta_1, \beta_2, \beta_3$, respectively. It can be concluded that all estimators consistently estimate the parameters despite the contamination. Furthermore, it is observed that, especially for β_1, β_2

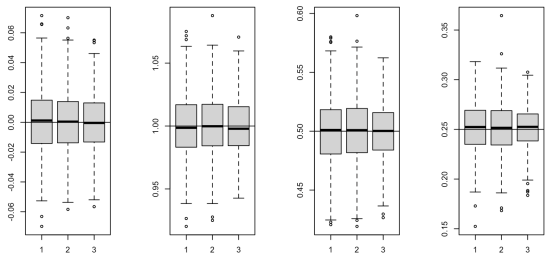

Figure 1: Boxplots for beta parameters for the uncontaminated data

and β_3, the estimated variances of the parameters clearly increase for Mallows-type M compared to OLS. An opposite conclusion can be drawn regarding MM, which produces more ambiguous results: whereas the boxplots for β_0 and β_2 indicate more precise estimates of MM than of OLS, the boxplots of β_1 and β_3 imply slightly less precise estimates.

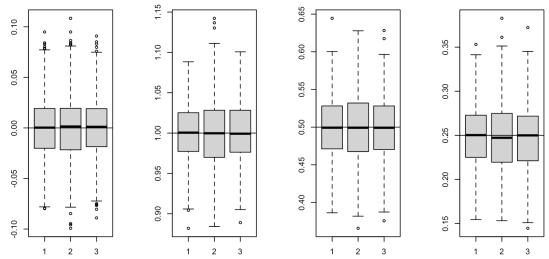

Figure 2: Boxplots for beta parameters for the contaminated data with t-distribution.

4.3 Contaminated data with bad leverage points

Figure 3 displays four boxplots containing the aforementioned estimators on the contaminated data with bad leverage points belonging to $\beta_0, \beta_1, \beta_2, \beta_3$, respectively. It is evident that the OLS estimator fails to correctly estimate the parameters, since the medians of the OLS estimates are considerably distant from the true values of the parameters. It is also observable that the Mallows-type M-estimator is more accurate in estimating the parameters, while the MM-estimator is the most precise in estimating the parameters. It can also be observed that the MM-estimates have the smallest variance, followed by Mallows-type M, and ultimately OLS.

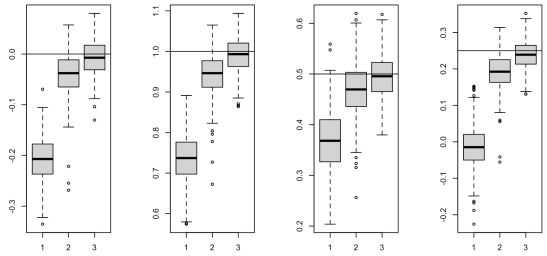

Figure 3: Boxplots for beta parameters for the contaminated data with bad leverage points.

4.4 Uncontaminated versus Contaminated

Figures 4 through 6 in the Appendix compare the variability of the estimated parameters for each of the estimation methods by displaying boxplots for results based on uncontaminated data and contaminated data where the error terms are contaminated with the t-distribution. Likewise, Figures 7 through 9 show these results but based on contaminated data with bad leverage points. It is clear that across both types of contamination, all estimation methods, and all parameters the variability of the parameters increases when the data set is contaminated with respect to when the data set stays uncontaminated. Moreover, contamination with the t-distribution does not influence the consistent estimation of all three estimation methods, whereas contamination with bad leverage points leads to highly significantly inconsistent estimates of OLS (Figure 7) and Mallows-type M (Figure 8).

Table 1 summarizes the averages of the estimated coefficient, their standard errors and significance levels based on the null hypothesis that the estimated parameters are equal to the true parameters. From Table 1 we can induce a slightly different conclusion: the OLS coefficients are significantly different from their true values, whereas Mallows-type M still produces unbiased results. However, with respect to the uncontaminated data, t-statistics increased a bit albeit without any change on p-values and hence conclusions about significance of the values.

	Uncontaminated				Contamination with t-distribution				Contamination with bad leverage			
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
OLS	0.002	1.000	0.501	0.249	-0.000	1.000	0.498	0.251	-0.208	0.733***	0.364**	-0.010***
	(0.021)	(0.025)	(0.028)	(0.025)	(0.038)	(0.043)	(0.049)	(0.043)	(0.046)	(0.052)	(0.059)	(0.052)
Mallows-type M	0.002	1.000	0.501	0.239	0.000	1.000	0.497	0.251	-0.041	0.938	0.469	0.193
	(0.032)	(0.038)	(0.042)	(0.038)	(0.035)	(0.043)	(0.048)	(0.043)	(0.035)	(0.042)	(0.048)	(0.042)
MM	0.001	1.000	0.500	0.250	0.000	1.001	0.496	0.251	-0.007	0.987	0.494	0.242
	(0.019)	(0.022)	(0.024)	(0.022)	(0.038)	(0.038)	(0.042)	(0.038)	(0.034)	(0.040)	(0.044)	(0.040)

Table 1: Simulation results for both uncontaminated and contaminated data sets, each of the three estimation methods, and all parameters.

5 Appendix

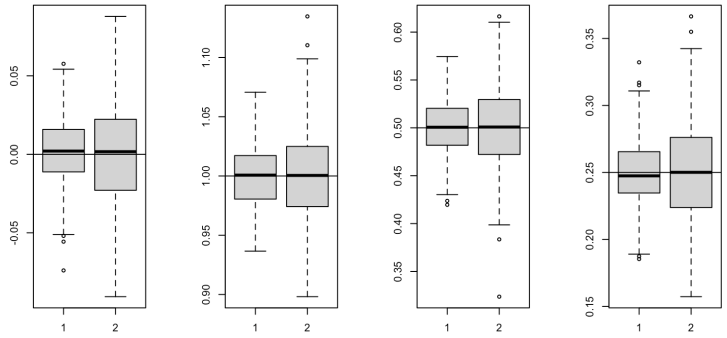

Figure 4: Boxplots for beta parameters for OLS to compare uncontaminated versus contaminated with t-distribution.

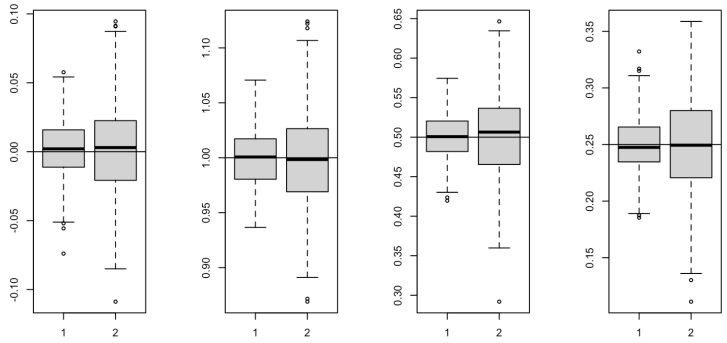

Figure 5: Boxplots for beta parameters for Mallows-type M to compare uncontaminated versus contaminated with t-distribution.

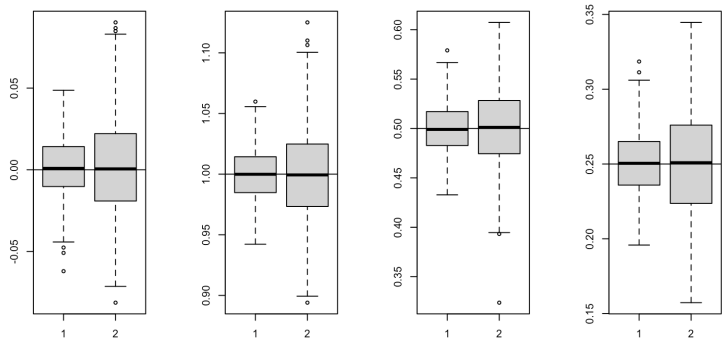

Figure 6: Boxplots for beta parameters for MM to compare uncontaminated versus contaminated with t-distribution.

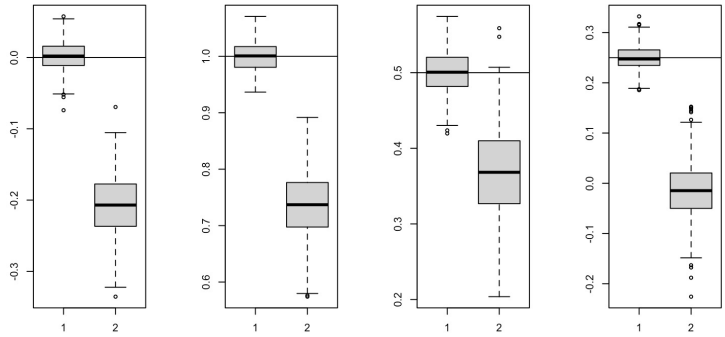

Figure 7: Boxplots for beta parameters for OLS to compare uncontaminated versus contaminated with bad leverage points.

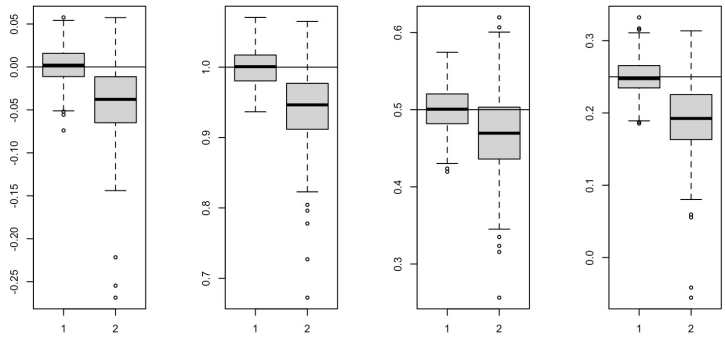

Figure 8: Boxplots for beta parameters for Mallows-type M to compare uncontaminated versus contaminated with bad leverage points.

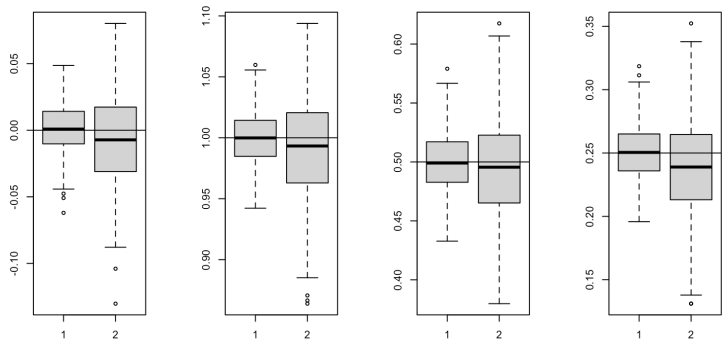

Figure 9: Boxplots for beta parameters for MM to compare uncontaminated versus contaminated with bad leverage points.

```r
install.packages("mvtnorm")
library("mvtnorm")
install.packages("robustbase")
library("robustbase")
install.packages("MASS")
library("MASS")

# Control Parameters
set.seed(123)
R <- 500
n <- 1000
mu <- rep(0,3)
sgmX <- matrix(c(1,0.5, 0.25, 0.5, 1, 0.5, 0.25, 0.5, 1),3,3)
bt <- c(1, 0.5, 0.25)

# Uncontaminated data
resultsUncontaminated <- replicate(R, {
  # Generate data
  x <- rmvnorm(n, mean = mu, sigma = sgmX)
  LP <- x%*%bt
  y <- rnorm(n, LP, sgmX)

  # Ordinary Least Squares (fit, param., var.)
  fit.ols <- lm(y~x)
  coef.ols <- coef(lm(y~x))
  var.ols <- coefficients(summary(fit.ols))[,2]
  tstat.ols <- (coef.ols-c(0,bt))/var.ols
  sign.ols <- coefficients(summary(fit.ols))[,4]

  # Mallows-type Estimator (fit, param., var.)
  fit.mallowsm_est = glmrob(y~x, family = gaussian(link = identity), method = "Mqle", weights.on.x = "hat")
  coef.mallowsm_est <- coefficients(summary(fit.mallowsm_est))[,1]
  var.mallowsm_est <- coefficients(summary(fit.mallowsm_est))[,2]
  tstat.mallowsm_est <- (coef.mallowsm_est - c(0,bt))/var.mallowsm_est
  sign.mallowsm_est <- coefficients(summary(fit.mallowsm_est))[,4]
```

Figure 10: R code part 1.

```r
  # MM-estimator
  fit.mm_est <- lmrob(y~x)
  coef.mm_est <- coefficients(summary(fit.mm_est))[,1]
  var.mm_est <- coefficients(summary(fit.mm_est))[,2]
  tstat.mm_est <- (coef.mm_est - c(0,bt))/var.mm_est
  sign.mm_est <- coefficients(summary(fit.mm_est))[,4]

  c(coef.ols, var.ols, tstat.ols, sign.ols, coef.mallowsm_est, var.mallowsm_est, tstat.mallowsm_est, sign.mallowsm_est, coef.mm_est, var.mm_est, tstat.mm_est, sign.m
})
estim_var <- rowMeans(resultsUncontaminated)
par(mfrow = c(1, 4))
boxplot(resultsUncontaminated[1,], resultsUncontaminated[17,], resultsUncontaminated[33,]); abline(h = 0)
boxplot(resultsUncontaminated[2,], resultsUncontaminated[18,], resultsUncontaminated[34,]); abline(h = bt[1])
boxplot(resultsUncontaminated[3,], resultsUncontaminated[19,], resultsUncontaminated[35,]); abline(h = bt[2])
boxplot(resultsUncontaminated[4,], resultsUncontaminated[20,], resultsUncontaminated[36,]); abline(h = bt[3])
```

Figure 11: R code part 2.

```r
# Contaminated data
resultsContaminated <- replicate(R, {
  # Generate data
  x <- rmvnorm(n, mean = mu, sigma = sgmX)
  LP <- x%*%bt
  ei <- rnorm(n,0,1)
  eps <- rbinom(n, 1, 0.05) # Contamination with probability 0.05

  # OPTION 1: CONTAMINATE DATA WITH T-DISTRIBUTION ERROR
  eps <- rbinom(n, 1, 0.05) # Contamination with probability 0.05
  ei <- (1-eps)*ei + eps*rt(n,3) # Simulate contamination from t distribution with 3 d.o.f.
  y_contam <- LP + ei
  x_contam <- x

  # OPTION 2: OUTLIERS (BAD LEVERAGE POINTS)
  eps <- rbinom(n, 1, 0.05) # Contamination with probability 0.05
  y <- LP + ei
  y_contam <- (1-eps)*y + eps*(y-2)
  x_contam <- (1-eps)*x + eps*(x+2)

  # Ordinary Least Squares (fit, param., var.)
  fit.ols <- lm(y_contam~x_contam)
  coef.ols <- coef(lm(y_contam~x_contam))
  var.ols <- coefficients(summary(fit.ols))[,2]
  tstat.ols <- (coef.ols-c(0,bt))/var.ols
  sign.ols <- coefficients(summary(fit.ols))[,4]

  # Mallows-type Estimator (fit, param., var.)
  fit.mallowsm_est = glmrob(y_contam~x_contam, family = gaussian(link = identity), method = "Mqle", weights.on.x = "covMcd")
  coef.mallowsm_est <- coefficients(summary(fit.mallowsm_est))[,1]
  var.mallowsm_est <- coefficients(summary(fit.mallowsm_est))[,2]
  tstat.mallowsm_est <- (coef.mallowsm_est - c(0,bt))/var.mallowsm_est
  sign.mallowsm_est <- coefficients(summary(fit.mallowsm_est))[,4]
```

Figure 12: R code part 3.

```r
  # MM-estimator
  fit.mm_est <- lmrob(y_contam~x_contam)
  coef.mm_est <- coefficients(summary(fit.mm_est))[,1]
  var.mm_est <- coefficients(summary(fit.mm_est))[,2]
  tstat.mm_est <- (coef.mm_est - c(0,bt))/var.mm_est
  sign.mm_est <- coefficients(summary(fit.mm_est))[,4]

  c(coef.ols, var.ols, tstat.ols, sign.ols, coef.mallowsm_est, var.mallowsm_est, tstat.mallowsm_est, sign.mallowsm_est, coef.mm_est, var.mm_est, tstat.mm_est, sign.m
})
estim_var <- rowMeans(resultsContaminated)
par(mfrow = c(1, 4))
boxplot(resultsContaminated[1,], resultsContaminated[17,], resultsContaminated[33,]); abline(h = 0)
boxplot(resultsContaminated[2,], resultsContaminated[18,], resultsContaminated[34,]); abline(h = bt[1])
boxplot(resultsContaminated[3,], resultsContaminated[19,], resultsContaminated[35,]); abline(h = bt[2])
boxplot(resultsContaminated[4,], resultsContaminated[20,], resultsContaminated[36,]); abline(h = bt[3])
```

Figure 13: R code part 4.