ASM Sc. J., 12, Special Issue 1, 2019 for IQRAC2018, 227-238

Robust Parameter Estimation for Fixed Effect Panel Data Model in the Presence of Heteroscedasticity and High Leverage Points

Muhammad Sani *[1,3], Habshah Midi[1,2], and Jayanthi Arasan[2]

[1] *Institute for Mathematical Research, Universiti Putra Malaysia Serdang, Malaysia*
[2] *Department of Mathematics, Faculty of Science, UPM Serdang, Malaysia*
[3] *Department of Mathematical Sciences and IT, Federal University Dutsinma, Katsina State, Nigeria*

* *Corresponding author: habshah@upm.edu.my*

In the presence of unknown heteroscedasticity structure and anomalous observations such as High Leverage Points (HLPs), the variance-covariance matrix of the ordinary least squares (OLS) estimator become bias and inconsistent in linear as well as in fixed effect (FE) panel data model. As a remedial measure, we propose Robust Heteroscedasticity Consistent Covariance Matrix (RHCCM) estimator based on Weighted Least Square in panel data model. In the proposed methods, weights are determined from HLPs detection methods so that the effect of HLPs can be minimized by assigning lower weights to HLPs. The numerical examples and simulation results indicate that the proposed RHCCM based on Fast Modified generalized Residuals (FMGt) offers substantial improvement over some existing estimators.

Keywords: HCCM, heteroscedasticity, high leverage point, Ordinary least squares, weighted least squares.

I. Introduction

Fixed effect (FE) panel data model is used when the individual specific effect (unobserved time invariant effects) is correlated with the explanatory variables (Víšek, 2015). The ordinary least squares (OLS) method is used to estimate the parameters of the model after the data transformation by mean centering (demeaned transformation). However, the OLS can strongly be biased and inconsistent in the presence of high leverage points (HLPs) and heteroscedasticity (unequal variances of the errors). HLPs referred to observations that are far away from the majority of the data points in X-direction. Many researches are available regarding outlying observations problem in FE panel data regression model, such as (Bakar and Midi, 2015, Bramati and Croux, 2007, Maronna et al., 2006, Verardi and Wagner, 2011). Their methods only addressed the problem of outliers but the combined problem of outliers particularly HLPs and heteroscedasticity in FE model is still missing in the literature.

Recently, least weighted squares (LWS) proposed by Víšek (2015) is used to estimate the model with FE in the presence of outliers in panel data. The shortcoming of this method is that, it make used of classical centering (mean centering) method and when there exist heteroscedasticity of unknown form it is inefficient and produces large variances which lead to inconsistency of the Variance-covariance matrix. Similarly, problem of heteroscedasticity in linear regression has been addressed by many researchers (Greene, 2007, Habshah et al., 2009, Rana et al., 2012, White, 1980). The heteroscedasticity consistent covariance matrix (HCCM) estimator denoted by HC0 proposed by White (1980) is used to remedy the problem heteroscedasticity of unknown form. There are many versions of HC0 estimator proposed by MacKinnon and White (1985), Cribari-Neto (2004) and Cribari-Neto et al. (2007) denoted by HC1, HC3, HC4 and HC5 respectively. Similarly, when heteroscedasticity comes along with the presence of HLPs in the data set, the HCCM estimator is bias which tends to perform poorly by providing unreliable parameter estimates.

The estimation strategy used for a model with heteroscedasticity of unknown form in the presence of HLPs as suggested by Furno (1996) is to perform ordinary least squares (OLS) estimation, and then employed a robust HCCM estimator which used residuals from weighted least squares (WLS) instead of OLS. The shortcoming of this method is that, the weighting method (hat matrix) used is inefficient as it suffers from masking and swamping effect (Habshah et al., 2009). This motivated us to propose weighting methods which are more efficient than hat matrix, in order to remedy the effects of HLPs and heteroscedasticity in FE panel data model.

In this study three robust weighting methods based on HLPs detection measures were used for robust HCCM estimator in FE panel data regression model. In this article, a more efficient robust weighting technique is used in order to successfully down weight the HLPs in a data set. The weights are based on; Robust mahalanobis distance based on minimum volume ellipsoid (RMD(MVE)), Diagnostic robust generalized potential based on index set equality (DRGP(ISE)) and proposed fast modified generalized studentized residuals (FMGt). Similarly, MM-centering method for data transformation will be employed to reduce the effects of HLPs.

Section 2 introduced the least weighted Square (LWS) estimation method. Section 3 explained classical HCCM and robust HCCM estimator. Section 4 described the propose robust estimation methods. The simulation study is presented in Section 5 and the real data examples are given in Section 6. Section 7 presents the conclusion.

II. Least Weighted Squares (LWS) Estimator

Víšek (2015) proposed an estimation technique for the fixed effect (FE) panel data model termed least weighted squares (LWS). Consider a FE model as,

$$y_{it} = \alpha_i + x'_{it}\beta + e_{it} \tag{1}$$

where, $i = 1, 2, \ldots, n$, $t = 1, 2, \ldots, T$, y_{it} are the response variables, x_{it} is the kth explanatory variables, α_i is the unobserved time-invariant effects and e_{it} is the error term that is assumed to be normal, uncorrelated across individual units and time. Also, $cov\,(x_{it},\ \alpha_i) \neq 0$ and α_i is usually eliminated when the data is transformed by demeaned transformation within each time series by mean given as:

$$(y_{it} - \overline{y}_{i.}) = (x_{it} - \overline{x}_{i.})\,\beta + e_{it} \tag{2}$$

where, $\overline{y}_{i.} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\overline{x}_{i.} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$ and Equation (2) becomes;

$$\tilde{y}_{it} = \tilde{x}_{it}\beta + e_{it} \tag{3}$$

where, $\tilde{y}_{it} = y_{it} - \overline{y}_{i}.$ and $\tilde{x}_{it} = x_{it} - \overline{x}_{i}.$ The least weighted squares (LWS) estimator proposed by Víšek (2015) is obtained by first computing the residual of the $(i,t)^{th}$observations from Equation (1) as:

$$r_{it} = y_{it} - x'_{it}\beta.$$

By denoting the qth squared residuals order statistic by $r^2_{(q)}(\beta)$, where $q = 1,2,\ldots,nT$ so that,

$$r^2_{(1)}(\beta) \leq r^2_{(2)}(\beta) \leq \cdots \leq r^2_{(n\ T)}(\beta),$$

followed by minimizing the weighted sum of squares residuals as,

$$WSS\,(\beta, \overrightarrow{w}) = \sum_{q=1}^{nT} w_q(y_q - x_q\beta)^2$$

where, weight (wq) is defined as $w_q \in [0,\ 1]$ for q = 1,2,..., nT for more details see (Víšek, 2015). The LWS estimator is given as:

$$\widehat{\beta}\,(LWS) = \ \underset{\beta}{argmin} \sum_{q=1}^{nT} w_q \; r_q^2\,(\beta)$$

$$= (X'WX)^{-1} X'Wy$$

where, W is a diagonal matrix of w_q. The absolute residuals ($r_{(it)}\,(\beta)'\,s$) value distribution function denoted by $F_{\beta}^{n}(r)$, its derivation shows that $\widehat{\beta}\,(LWS)$ is among the solutions of the normal equations, see Víšek (2011)

$$\sum_{i=1}^{n}\sum_{t=1}^{T} w\ \left(F_{\beta}^{nT}\left(|r_{it}\,(\beta)|\right)\right)\ x_{it}(y_{it} - \tilde{x}_{it}\beta) = 0$$

It has been proven that $\widehat{\beta}\,(LWS)$ is consistent under certain assumptions.

III. HCCM and Robust HCCM Estimators

White (1980) proposed the heteroscedasticity consistent covariance matrix (HCCM) estimator known as HC0 where he replaced σ_q^2 with $\widehat{e}_q^2$ in the covariance matrix of $\widehat{\beta}$ as:

$$HC0 = (x'x)^{-1} x'\hat{\Phi}_0 x(x'x)^{-1} \qquad (4)$$

where, $\hat{\Phi}_0$= diag $\widehat{\mathrm{e}}_q^2\}$. Different adjustment of HC0 was done by many researchers which give rise to HC1, HC2 and HC3. They are generally biased for small sample sizes, see (Furno, 1996, Hausman and Palmer, 2012, Lima et al., 2009). This research will only focus on HC4 and HC5. The HC4 proposed by Cribari-Neto (2004) was build under HC3, which is defined as follows:

$$HC4 = (x'x)^{-1} x'\hat{\Phi}_4 x(x'x)^{-1} \qquad (5)$$

where, $\hat{\Phi}_4$= diag $\left\{\frac{\widehat{e}_q^2}{(1-h_q)^{\delta_{\mathrm{q}}}}\right\}$ for $q = 1,\ \ldots, nT$ with δ_{q}= min$\left\{4,\ \frac{h_q}{h}\right\}$, which control the discount factor of theq^{th} squared residuals, given by the ratio of h_q and the average of h_q's (h). Since $0 < 1 - h_q < 1$ and $\delta_q > 0$ it implies that $0 < (1-h_q)^{\delta_q} < 1$. The larger h_q is relative to h, the more the HC4 discount factor inflates the q^{th} squared residual. The truncation at 4 amounts to twice what is used in the definition of HC3; that is, $\delta_{\mathrm{q}} = 4$ when $hq > 4h = 4p/nT$. The result obtained by Cribari-Neto (2004) suggested HC4 inference in finite sample size relative to HC3.

Similarly, another modification of the exponent $(1-h_q)$ of HC4 was proposed by Cribari-Neto et al. (2007) to control the level of maximal leverage. The estimator is termed HC5, defined as:

$$HC5 = (x'x)^{-1} x'\hat{\Phi}_5 x(x'x)^{-1} \qquad (6)$$

where the value of $\hat{\Phi}_5$= diag $\left\{\frac{\widehat{e}_q^2}{\sqrt{(1-\mathrm{h_q})^{\alpha_{\mathrm{q}}}}}\right\}$ for $q = 1,\ \ldots, nT$ with the quantity α_{q}= min$\left\{\frac{h_q}{h},\ \max\left\{4,\ \frac{\mathrm{k}h_{\max}}{h}\right\}\right\}$, which determine the amount of increased of the q^{th} squared residual, given by the ratio between $\mathrm{h_{max}}$ (maximal leverage) and h (mean leverage value of $\mathrm{h_q}$'s). When $\frac{h_{\mathrm{q}}}{h} \leq 4$ it follows that $\alpha_q = \frac{h_q}{h}$. Also, since $0 < 1 - h_q < 1$ and $\alpha_q > 0$, it implies that $0 < (1-h_q)^{\alpha_q} < 1$ and k is a constant ranges between $0 < k < 1$ where HC5 reduces to HC4 when k = 0. K value was suggested to be chosen as 0.7 by Cribari-Neto et al. (2007) following his simulation result that leads to efficient quasi-t inference.

Furno (1996) suggested using weighted least squares (WLS) regression residuals instead of OLS residuals used by White (1980) in HC0 estimator. The weight is based on the hat matrix (h_{i}) and the robust (weighted) version of HC0 is defined as:

$$HC0_W = (x'Wx)^{-1} x'W\widehat{\Phi}_{0\mathrm{w}} Wx(x'Wx)^{-1} \qquad (7)$$

where, W is an $n \times n$ diagonal matrix with,

$$w_q = \min\left(1,\quad c/h_{\mathrm{q}}\right), \qquad (8)$$

and c is the cutoff point, $c = \frac{1.5p}{nT}$, p is the number of parameters together with the intercept, $\widehat{\Phi}_{0\mathrm{w}}$= diag $\{\widetilde{e}_q^2\}$ with $\widetilde{e}_q$ being the qth residuals from weighted least squares. Note that, nonleveraged observations are weighted by 1 and leveraged observations are weighted by $(c/\mathrm{h_q})$ to reduce their intensity and w_q is considered as the weight, so that the WLS estimator of β is,

$$\widetilde{\beta} = (X'WX)^{-1}X'WY.$$

The robust HCCM estimators for the HC4 and HC5 based on Furno's weighting method are $HC4_W$ and $HC5_W$ define as:

$$HC4_W = (x'Wx)^{-1}x'W\widehat{\Phi}_{4\mathrm{w}}Wx(x'Wx)^{-1} \tag{9}$$

where, $\hat{\Phi}_{4\mathrm{w}}= \mathrm{diag}\left\{\frac{\widetilde{e}_q^2}{(1-h_{\mathrm{q}}^*)^{\delta_{\mathrm{q}}^*}}\right\}$ for $q = 1, \ldots, nT$ with $\delta_{\mathrm{q}}^*= \min\left\{4, \frac{h_{\mathrm{q}}^*}{h^*}\right\}$, and h_{q}^* is the q^{th} diagonal elements of the weighted hat matrix $H_w = \sqrt{W}x(x'Wx)^{-1}x'\sqrt{W}$. And,

$$HC5_W = (x'Wx)^{-1}x'W\widehat{\Phi}_{5\mathrm{w}}Wx(x'Wx)^{-1} \tag{10}$$

where the values of $\hat{\Phi}_{5\mathrm{w}}= \mathrm{diag}\left\{\frac{\widetilde{e}_q^2}{\sqrt{(1-h_{\mathrm{q}}^*)^{\alpha_{\mathrm{q}}^*}}}\right\}$ for $q = 1, \ldots, nT$ with the exponent $\alpha_{\mathrm{q}}^*= \min\left\{\frac{h_{\mathrm{q}}^*}{h^*}, \max\left\{4, \frac{\mathrm{kh^*}_{\max}}{h^*}\right\}\right\}$. The Furno's weighted least square method described here denoted by WLS_F will be applied to the transformed FE panel data model in Equation (3) and obtain an estimate which is based on Furno's weighting method.

IV. Proposed Robust HCCM Estimators

The idea of Furno (1996) was employed to formulate new RHCCM estimator base on HLPs identification weighting method; Robust mahalanobis distance based on minimum volume ellipsoid (RMD(MVE)), diagnostic robust generalized potential based on index set equality (DRGP(ISE)) and fast modified generalized studentized residualsl (FMGt). These methods are more efficient than the hat matrix used by Furno for the identification of HLPs (Habshah et al., 2009). We also employed robust centering method (MM-centering) for the data tranformation instead of classical mean-centering used by Víšek (2015) in order to reduce the effect of HLPs.

MM-centering is a robust technique of data transformation proposed by Bakar and Midi (2015), the centering involves transformation of panel data within each time series. The usual mean centering is highly sensitive to HLPs (Bramati and Croux, 2007). As an alternative, the centering using median is put forward. However, for uncontaminated data median centering was found to have low efficiency than the mean and causes non linearity to the transformed data (Maronna et al., 2006). This affects the efficiency of robust estimators when there is no HLPs. The MM-centering method is introduced in order to bring back linearity into the transformed data and also to provide high efficiency.

The MM-estimate of location was originally proposed by Yohai (1987). The goal was to produce a high efficiency and high breakdown point estimator. The procedure is given by: Firstly, compute S-estimates of location and covariance to obtain the preliminary scale estimate, $\hat{\sigma}_{\mathrm{n}}$:

$$\hat{\sigma}_n = \min_t s_n\ (t).$$

The scale estimate is obtained by minimizing an M-estimate of scale. Thus, the initial estimate has a high breakdown point of 50% and it is also equivariance. The location S-estimate $\hat{\mu}_{\mathrm{n}}$ is defined by:

$$\hat{\mu}_n = \arg\min_t s_n\ (t).$$

Tukey's Bisquare weight function was employed to down weight HLPs, where ρ is given as:

$$\rho\left(x\right) = \begin{cases} \frac{1}{6c^4}x^6 - \frac{1}{2c^2}x^4 + \frac{1}{2}x^2, & if\ |x| \leq c \\ \frac{c^2}{6}, & if\ |x| > c \end{cases}$$

where c is the tuning constant. At Gaussian errors, the initial estimator has only 28.7% efficiency (Ruppert, 1992). By fixing the scale estimate, the shape and location are re-estimated by a high efficient M-estimator, this provides 95% efficiency in a central model (Ruppert, 1992). This procedure of data transformation by the MM-estimate of location is called MM centering. The demeaned transformed data within each time series by MM centering is now given as:

$$y_{it} - \hat{\mu}_{mm}\left\{y_{it}\right\} = \left(x_{it}^{(k)} - \hat{\mu}_{mm}\left\{x_{it}^{(k)}\right\}\right)\beta + e_{it} \quad (11)$$

for $1 \leq t \leq T$, $1 \leq i \leq n$, and $1 \leq k \leq K$, where $x^{(k)}$ is the k^{th} explanatory variables.

A. Robust HCCM Estimator based on RMD(MVE)

Mahalanobis (1936) introduced a diagnostic measure of the deviation of an observation from its center named Mahalanobis Distance (MD), in which the independent variables of the q^{th} observations are presented as $x_q = (1,\ x_{q1},\ x_{q2}, \ldots,\ x_{qk}) = (1,\ R_q)$ so that $R_q = (x_{q1},\ x_{q2}, \ldots,\ x_{qk})$ will be k-dimensional row vector, where the mean and covariance matrix vector are $\overline{R} = \frac{1}{nT}\sum_{q=1}^{nT} R_q$ and $CV = \frac{1}{nT-1}\sum_{q=1}^{nT}\left(R_q - \overline{R}\right)'(R_q - \overline{R})$ respectively. The MD for the q^{th} points is given as:

$$MD_q = \sqrt{\left(R_q - \overline{R}\right)'(CV)^{-1}(R_q - \overline{R})} \quad (12)$$

where, $q = 1,\ 2, \ldots, nT$. The average vector $\overline{R}$ and covariance matrix CV in Equation (12) are not robust and easily affected by HLPs. Rousseeuw (1984) recommended using Robust Mahalanobis Distance (RMD) as in Equation (12) with slight modification where the $\overline{R}$ and CV are obtained from minimum volume ellipsoid (MVE). The corresponding CV is provided by the ellipsoid and multiplied by a suitable factor in order to obtain consistency. We suggest the cut off value of RMD as in Equation (13).

$$cd = \text{median}\left(RMD_q\right) + 3.\text{MAD}(RMD_q) \quad (13)$$

where, MAD stands for median absolute deviation. The weight obtained by this RMD(MVE) method is given by:

$$w_{qr} = \min\left(1,\quad cd/RMD_q\right), \quad (14)$$

so that HLPs are weighted by (cd/RMD_q) and non leverage by 1. To obtain the RHCCM estimator based on RMD(MVE) weighting method denoted by WLS_{RMD}, we replace Equation (8) by (14) and follow Furno's robust HCCM estimation method as discussed in Section (4.1).

B. Robust HCCM Estimator based on DRGP(ISE)

Lim and Midi (2016) proposed diagnostic robust generalized potential (DRGP) based on index set equality (ISE) in order to reduce the effect of swamping/masking and computional complexity of DRGP based on minimum volume ellipsoid (MVE). The ISE was developed from the fast minimum covariance determinant (MCD). ISE was tested and found to execute highly faster in the estimation of robust estimator of scale and location. Thus, ISE has faster running time compared to MVE (Lim and Midi, 2016). The DRGP(ISE) consist of two steps whereby RMD based on ISE is utilised to detect the suspected HLPs.
The generalized potential $(\hat{p}_q)$ is employed on the second step to check all the suspected identified HLPs, those possess a low leverage point will be put back to the 'R' group. This technique continued until all points of the 'D' group has been checked to confirm whether they can be referred as HLPs. The DRGP(ISE) denoted as $DRGP_q$ is defined as follows:

$$DRGP_q = \begin{cases} h_q^{(-D)} & for\ q \in D \\ \frac{h_q^{(-D)}}{1-h_q^{(-D)}} & for\ q\ \in R \end{cases} \quad (15)$$

The cut-off point for $DRGP_q$ is given by:

$$CDRGP_q = median(DRGP_q) + 3\ Q_{nT}\ (DRGP_q) \quad (16)$$

where $Q_{nT}{=}c\{|x_q{-}x_j\ |; < j\}_{(k)}$ is a pair wise order statistic for all distance proposed by Rousseeuw and Van Zomeren (1990) where $k = {}^{h}_{.}C_2 \approx {}^{h}_{.}C_2/4$ and $h = [nT/2] + 1$. They make used of c = 2.2219, as this value will provide Q_{nT} a consistent estimator for gaussian data. If some identified $DRGP_q$ did not exceed $CDRGP_q$ then, the case with the least $DRGP_q$ will return to the estimation subset for re-computation of $DRGP_q$. The values of diagnostic robust generalized potential (DRGP) based on final 'D' set is the DRGP(ISE) represented by $DRGP_q$ and the 'D' points will be declared as HLPs. Now, the DRGP(ISE) weight can be obtained as follows:

$$w_{qd} = min(1, CDRGP_q/DRGP_q) \quad (17)$$

where, the HLPs are weighted by $(CDRGP_q/DRGP_q)$ and non leverage by 1. We also replace Equation (8) by (17) and employed RHCCM estimation methods of Furno as discussed in Section (4.1) to obtain the RHCCM estimator based on DRGP(ISE) weighting method denoted by WLS_{DRGP}.

C. Robust HCCM Estimator based on FMGt

We suspect that the proposed RHCCM based on DRGP(ISE) weighting method will be affected by good HLPs because the RMD(ISE) only detect HLPs and unable to classify observations into good and bad HLPs. Hence good observations will be given low weight and the efficiency of the RHCCM estimator tends to decrease as the number of good leverage points increases. As such we propose to firstly classify observations into regular, good and bad HLPs, and vertical outliers, before formulating weighting method.

The proposed classification method is similar to Alguraibawi et al. (2015) with slight modification whereby the modified generalized studentized residual (MGtq) is established based on the Reweighted Least Squares (RLS) and DRGP(ISE) as initial estimates to make computation very fast compared to DRGP(MVE). Subsequently, the MGt_q is given by,

$$MGt_q = \begin{cases} \dfrac{\widehat{e}_{q(R^*)}}{\widehat{\sigma}_{R^*-1}\sqrt{1-h^{**}_{q(R^*)}}}, & for\ q \in R^* \\ \dfrac{\widehat{e}_{q(R^*)}}{\widehat{\sigma}_{R^*}\sqrt{1+h^{**}_{q(R^*)}}}, & for\ q \notin R^* \end{cases} \quad (18)$$

where $\widehat{e}_{q(R^*)}$, $\widehat{\sigma}_{(R^*)}$ are the OLS residuals and residuals standard error for remaining set R, respectively, once the R^* set is identified based on RLS and DRGP(ISE). The cut off point (C_{MGtq}) is calculated as follows:

$$C_{MGtq} = median\,(MGtq) + cMAD(MGtq) \quad (19)$$

The classification of observations into the four categories is called fast modified generalized studentized residuals (FMGt). The classification scheme is as follows:

1. Regular observation (RO): An observation is declared as regular observation if ;

$|MGtq| \leq C_{MGtq}$ and $|DRGPq| \leq C_{DRGPq}$

2. Vertical outlying observation (VO): An observation is declared as VO if ;

$|MGtq| > C_{MGtq}$ and $|DRGPq| \leq C_{DRGPq}$

3. Good leverage observation (GLO): An observation is declared as GLO if ;

$|MGtq| \leq C_{MGtq}$ and $|DRGPq| > C_{DRGPq}$

4. Bad leverage observation (BLO): An observation is declared as BLO if ;

$|MGtq| > C_{MGtq}$ and $|DRGPq| > C_{DRGPq}$

Once the bad HLPs and vertical outliers are identified, the weight is defined as:

$$w_{qr} = \min(1, \quad CMGt_q/MGt_q), \quad (20)$$

The RHCCM estimator based on FMGt weight denoted as WLS_{FMGt} is then formulated by replacing Equation (8) by (20). It is important to note that the proposed WLS_{FMGt} is expected to be more efficient because the weighting function is based only on VO and BLOs.

V. Simulation Study

Monte Carlo simulation is used to evaluate the performances of proposed weighting method (WLS_{RMD},WLS_{DRGP},WLS_{FMGt}) in FE panel data regression model. The response variable is set base on Equation (1) where $e_{it} \sim$

$N\left(0,\sigma_e^2\right)$, $\alpha_i \sim N\left(0,\ 5\right)$ and the vector of coefficients β equal to a vector of ones. Following Víšek (2015) the independent variables $(x_{it1}, x_{it2}, x_{it3})$ are generated from standard normal distribution. Three sample sizes $n =$ 5, 10, 15 with the corresponding $t = 10$, 15 20 were replicated twice to form $n = 10$, 20, 30 and $t = 20$, 30, 40 respectively, in order to create heteroscedasticity. The skedastic function is defined as $\sigma_e^2 = \exp\{c_1 x_{it1}\}$ (Lima et al., 2009) where the value of $c_1 = 0.45$ was chosen such that $\lambda \approx 87.5$ and will be constant among the sample sizes. The strength (degree) of heteroscedasticity is measured by $\lambda = \max\left(\sigma_e^2\right)/\min(\sigma_e^2)$ whereby, for homoscedasticity $\lambda = 1$.

The data were contaminated by introducing high leverage points (HLPs) which are randomly generated from $N\left(1,10\right)$, at 0%, 10% and 20% contamination level with R=1000 replications. The most efficient and best method is the one with lowest bias, lowest standard error of the estimate, lowest standard error of HC4 and HC5. Results from Table (1) to (3) show the performance of the proposed methods (WLS$_{RMD}$, WLS$_{DRGP}$ and WLS$_{FMGt}$) and the existing methods (LWS and WLS$_F$), at different sample sizes and HLPs contamination level. The results show that all the proposed methods were more efficient than the existing methods, by providing smaller bias, and standard error of HC4 and HC5. However, justification based on standard error of the estimates here is improper and inefficient, as the structure of heteroscedasticity is unknown. Therefore, the estimation will be based on the HCCM estimator (HC4 and HC5). The results of these two methods are close to each other. The HC4 and HC5 based on WLS$_{FMGt}$ was found to be the best method due to the smaller value of bias, and standard error of HC4 and HC5. Figures 1-3 show the performance of all the methods at 20% HLPs contamination level with different sample sizes, where WLS$_{FMGt}$ is the best followed by WLS$_{DRGP}$, WLS$_{RMD}$, WLS$_F$, and finally LWS.

Table 1: Simulation result of panel data estimates for n=10,t=20

Con. Level	Estimator		Bias	SE of Estimates	Stand. Error	
					HC4	HC5
0 % HLPs	LWS	b_1	0.0006	0.1476	0.0444	0.0444
		b_2	0.0021	0.1480	0.0468	0.0468
		b_3	0.0216	0.1460	0.0454	0.0454
	WLS $_F$	b_1	0.0024	0.1342	0.0185	0.0185
		b_2	0.0059	0.1344	0.0187	0.0187
		b_3	0.0234	0.1334	0.0184	0.0184
	WLS $_{RMD}$	b_1	0.0021	0.1305	0.0197	0.0197
		b_2	0.0064	0.1308	0.0203	0.0203
		b_3	0.0245	0.1297	0.0196	0.0196
	WLS $_{DRGP}$	b_1	0.0019	0.1317	0.0190	0.0190
		b_2	0.0062	0.1319	0.0194	0.0194
		b_3	0.0243	0.1309	0.0189	0.0189
	WLS $_{FMGt}$	b_1	0.0044	0.1252	0.0174	0.0174
		b_2	0.0062	0.1255	0.0179	0.0179
		b_3	0.0241	0.1245	0.0174	0.0174
10% HLPs	LWS	b_1	0.2832	0.3172	0.2323	0.2323
		b_2	0.3042	0.3095	0.2087	0.2087
		b_3	0.2437	0.3111	0.2297	0.2297
	WLS $_F$	b_1	0.2284	0.2782	0.0841	0.0841
		b_2	0.2660	0.2738	0.0774	0.0774
		b_3	0.2310	0.2745	0.0755	0.0755
	WLS $_{RMD}$	b_1	0.2590	0.2733	0.0915	0.0915
		b_2	0.2999	0.2694	0.0855	0.0855
		b_3	0.2646	0.2690	0.0842	0.0842
	WLS $_{DRGP}$	b_1	0.1088	0.2709	0.0826	0.0826
		b_2	0.1523	0.2660	0.0756	0.0756
		b_3	0.1152	0.2668	0.0747	0.0747
	WLS $_{FMGt}$	b_1	0.0025	0.2017	0.0478	0.0479
		b_2	0.0302	0.2180	0.0448	0.0448
		b_3	0.0285	0.2186	0.0453	0.0453
20% HLPs	LWS	b_1	0.1941	0.3640	0.2395	0.2395
		b_2	0.1739	0.3559	0.2495	0.2495
		b_3	0.1830	0.3550	0.2549	0.2549
	WLS $_F$	b_1	0.1023	0.3228	0.0860	0.0860
		b_2	0.1142	0.3231	0.0874	0.0874
		b_3	0.1351	0.3131	0.0879	0.0879
	WLS $_{RMD}$	b_1	0.1041	0.2901	0.0849	0.0849
		b_2	0.1213	0.2921	0.0861	0.0861
		b_3	0.1395	0.2927	0.0867	0.0867
	WLS $_{DRGP}$	b_1	0.0714	0.2720	0.0649	0.0649
		b_2	0.0755	0.2749	0.0667	0.0667
		b_3	0.0628	0.2556	0.0670	0.0670
	WLS $_{FMGt}$	b_1	0.0131	0.2198	0.0516	0.0516
		b_2	0.0304	0.2144	0.0530	0.0530
		b_3	0.0358	0.2018	0.0539	0.0539

Table 2: Simulation result of panel data estimates for n=20,t=30

Con. Level	Estimator		Bias	SE of Estimates	Stand. Error	
					HC4	HC5
0% HLPs	LWS	b_1	0.0116	0.0979	0.0166	0.0166
		b_2	0.0012	0.0981	0.0165	0.0165
		b_3	0.0157	0.0977	0.0162	0.0162
	WLS_F	b_1	0.0143	0.0902	0.0081	0.0081
		b_2	0.0031	0.0902	0.0081	0.0081
		b_3	0.0176	0.0900	0.0081	0.0081
	WLS_{RMD}	b_1	0.0155	0.0874	0.0084	0.0084
		b_2	0.0038	0.0875	0.0084	0.0084
		b_3	0.0172	0.0874	0.0084	0.0084
	WLS_{DRGP}	b_1	0.0152	0.0883	0.0082	0.0082
		b_2	0.0035	0.0883	0.0082	0.0082
		b_3	0.0172	0.0881	0.0082	0.0082
	WLS_{FMGt}	b_1	0.0167	0.0838	0.0074	0.0074
		b_2	0.0047	0.0839	0.0073	0.0073
		b_3	0.0181	0.0838	0.0073	0.0073
10% HLPs	LWS	b_1	0.2585	0.3840	0.2433	0.2433
		b_2	0.2363	0.3221	0.2424	0.2424
		b_3	0.1920	0.3210	0.2422	0.2422
	WLS_F	b_1	0.2236	0.2817	0.1210	0.1210
		b_2	0.2080	0.2963	0.1209	0.1209
		b_3	0.1871	0.2799	0.1205	0.1205
	WLS_{RMD}	b_1	0.2140	0.2785	0.1021	0.1021
		b_2	0.2017	0.2738	0.1020	0.1020
		b_3	0.1780	0.2741	0.1020	0.1020
	WLS_{DRGP}	b_1	0.1276	0.2597	0.0770	0.0770
		b_2	0.1221	0.2598	0.0765	0.0765
		b_3	0.1025	0.2598	0.0762	0.0762
	WLS_{FMGt}	b_1	0.0211	0.1937	0.0360	0.0360
		b_2	0.0148	0.1943	0.0354	0.0354
		b_3	0.0384	0.1944	0.0352	0.0352
20% HLPs	LWS	b_1	0.1511	0.3715	0.2766	0.2766
		b_2	0.1786	0.3871	0.2756	0.2756
		b_3	0.1751	0.3858	0.2793	0.2793
	WLS_F	b_1	0.1267	0.3521	0.0942	0.0942
		b_2	0.1294	0.3520	0.0948	0.0948
		b_3	0.1210	0.3527	0.0941	0.0941
	WLS_{RMD}	b_1	0.0930	0.2987	0.0927	0.0927
		b_2	0.0875	0.2987	0.0920	0.0920
		b_3	0.0901	0.2950	0.0924	0.0924
	WLS_{DRGP}	b_1	0.0618	0.2500	0.0593	0.0593
		b_2	0.0522	0.2500	0.0593	0.0593
		b_3	0.0627	0.2506	0.0593	0.0593
	WLS_{FMGt}	b_1	0.0201	0.2079	0.0485	0.0485
		b_2	0.0143	0.2079	0.0484	0.0484
		b_3	0.0363	0.2085	0.0484	0.0484

Table 3: Simulation result of panel data estimates for n=30,t=40

Con. Level	Estimator		Bias	SE of Estimates	Stand. Error	
					HC4	HC5
0% HLPs	LWS	b_1	0.0063	0.0939	0.0126	0.0113
		b_2	0.0120	0.0937	0.0126	0.0113
		b_3	0.0015	0.0938	0.0127	0.0113
	WLS_F	b_1	0.0045	0.0927	0.0107	0.0107
		b_2	0.0120	0.0926	0.0107	0.0107
		b_3	0.0021	0.0927	0.0107	0.0107
	WLS_{RMD}	b_1	0.0072	0.0922	0.0107	0.0107
		b_2	0.0146	0.0921	0.0107	0.0107
		b_3	0.0047	0.0922	0.0107	0.0107
	WLS_{DRGP}	b_1	0.0072	0.0923	0.0107	0.0107
		b_2	0.0146	0.0922	0.0101	0.0101
		b_3	0.0046	0.0923	0.0107	0.0107
	WLS_{FMGt}	b_1	0.0071	0.0914	0.0105	0.0105
		b_2	0.0151	0.0913	0.0105	0.0105
		b_3	0.0143	0.0914	0.0105	0.0105
10% HLPs	LWS	b_1	0.2976	0.3424	0.2318	0.2318
		b_2	0.3481	0.3423	0.2317	0.2317
		b_3	0.3741	0.3421	0.2325	0.2325
	WLS_F	b_1	0.2686	0.2876	0.1582	0.1582
		b_2	0.2792	0.2819	0.1592	0.1592
		b_3	0.2908	0.2832	0.1608	0.1608
	WLS_{RMD}	b_1	0.2070	0.2685	0.1168	0.1168
		b_2	0.2202	0.2619	0.1170	0.1170
		b_3	0.2329	0.2634	0.1170	0.1170
	WLS_{DRGP}	b_1	0.1510	0.2125	0.0839	0.0839
		b_2	0.1645	0.2244	0.0856	0.0856
		b_3	0.1737	0.2451	0.0856	0.0856
	WLS_{FMGt}	b_1	0.0330	0.1990	0.0420	0.0420
		b_2	0.0208	0.1986	0.0416	0.0416
		b_3	0.0283	0.1988	0.0425	0.0425
20% HLPs	LWS	b_1	0.1855	0.3935	0.2324	0.2324
		b_2	0.2048	0.3984	0.2336	0.2336
		b_3	0.1908	0.3906	0.2331	0.2331
	WLS_F	b_1	0.1082	0.2966	0.1179	0.1179
		b_2	0.1023	0.2979	0.1184	0.1184
		b_3	0.1064	0.2979	0.1182	0.1182
	WLS_{RMD}	b_1	0.0786	0.2732	0.8175	0.8175
		b_2	0.0783	0.2729	0.8180	0.8180
		b_3	0.0708	0.2833	0.8179	0.8179
	WLS_{DRGP}	b_1	0.0417	0.2330	0.5751	0.5751
		b_2	0.0423	0.2342	0.5804	0.5804
		b_3	0.0265	0.2343	0.5793	0.5793
	WLS_{FMGt}	b_1	0.0137	0.2160	0.0523	0.0523
		b_2	0.0103	0.2170	0.0512	0.0512
		b_3	0.0246	0.2173	0.0513	0.0513

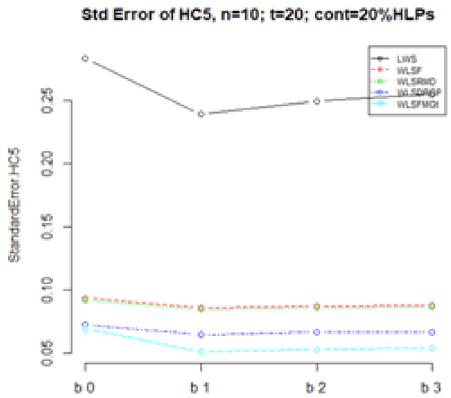

Figure 1: plot of HC5 SE for n=10,t=20

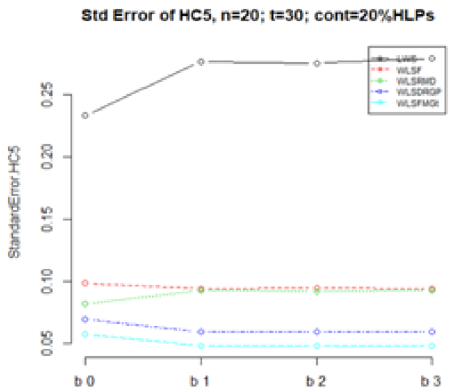

Figure 2: plot of HC5 SE for n=20,t=30

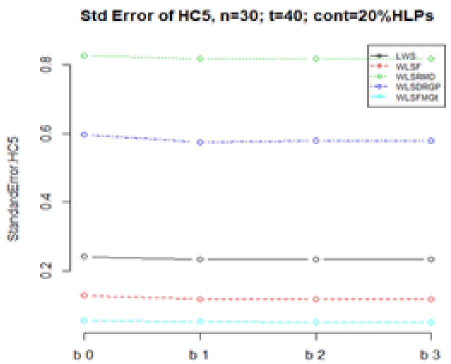

Figure 3: plot of HC5 SE for n=30,t=40

VI. Numerical Examples

In this section, the proposed robust methods (WLS_{RMD}, WLS_{DRGP} and WLS_{FMGt}) and the existing robust methods (LWS, WLS_F) will be applied to a real panel data set in order to evaluate their performances. Firstly, we consider a set of grunfeld investment data containing 200 observations taken from Kleiber and Zeileis (2008). The data represent investment of 10 firms over 20 years (1935 – 1954), with investment as the response variable. Value of firms and value of the firm's capital stock are treated as explanatory variables. This data set was diagnosed using FMGt with DRGP and found that it contains 46 outlying observations, where 19 of them are GLO, the remaining 27 are VO and BLO as shown in Figure 4. However, there is a presence of heteroscedasticity in this data set due to the funnel shape produce by a plot in Figure 5.

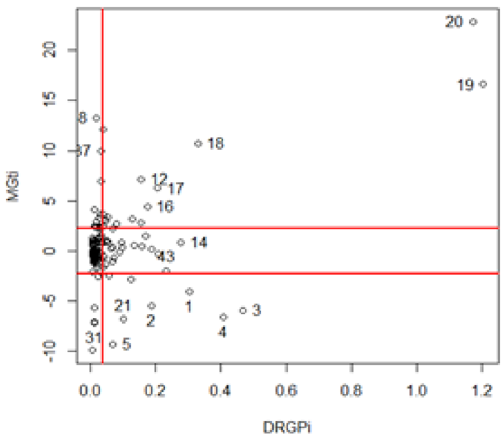

Figure 4: plot of MGTi vs DRGPi for grunfeld data

Table 4 presents the result of proposed and existing methods in grunfeld investment data set. The result indicates that WLS_{FMGt} is the most efficient method as the method provides the lowest standard error of HC4 and HC5, lowest standard error of the estimates.

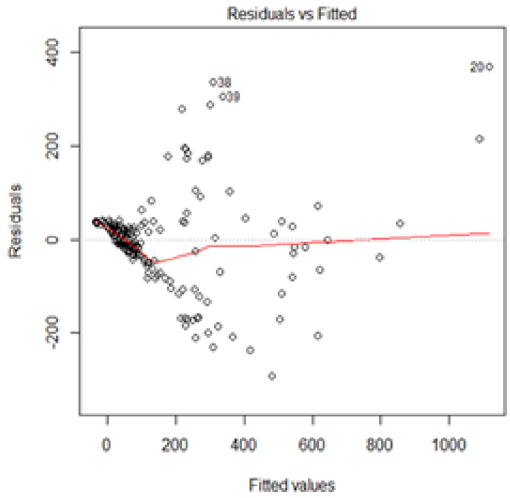

Figure 5: plot of residuals vs fitted value for grunfeld data

Table 4: Regression estimates for the grunfeld investment data set

Estimator		Coeff. of Estimates	SE of Estimates	Stand. Error	
				HC4	HC5
LWS	b_1	0.1027	0.0153	0.0005	0.0005
	b_2	0.1991	0.0198	0.0028	0.0028
WLS_F	b_1	0.0908	0.0142	0.0003	0.0003
	b_2	0.2337	0.0203	0.0014	0.0014
WLS_{RMD}	b_1	0.0993	0.0132	0.0004	0.0004
	b_2	0.2611	0.0191	0.0016	0.0017
WLS_{DRGP}	b_1	0.0882	0.0133	0.0003	0.0003
	b_2	0.2198	0.0182	0.0009	0.0009
WLS_{FMGt}	b_1	0.0928	0.0084	0.0002	0.0002
	b_2	0.1970	0.0131	0.0005	0.0006

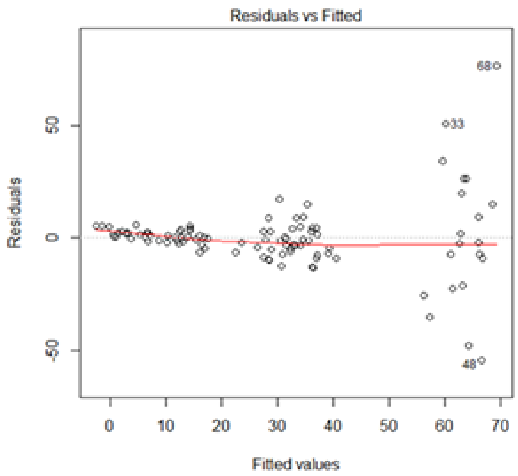

Figure 6: plot of residuals vs fitted value for artificial data

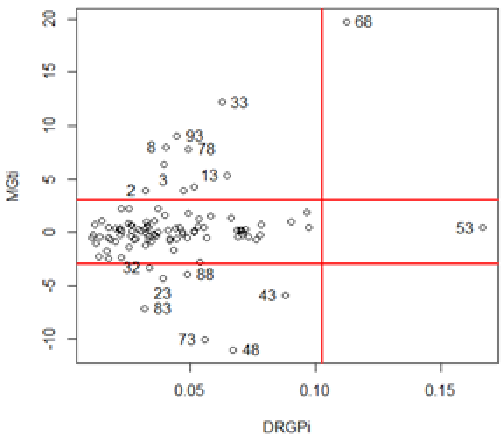

Figure 7: plot of MGTi vs DRGPi for artificial data

The second example is an artificial panel data set, generated according to Bramati and Croux (2007) simulation method consisting 100 observations for 5 individuals observed over a period of 20 years. The response variable is generated according to Equation (1) with $e_{it} \sim N\left(0, \sigma_e^2\right)$, $\alpha_i \sim U\,(0,\ 20)$ where $\sigma_e^2 = \exp\{c_1 x_1\}$ with $c_1 = 0.65$ (Lima et al., 2009). The vector of coefficients β equal to a vector of ones. The three independent variables are generated from standard normal distribution. Figures 6 and 7 indicate the presence of heteroscedasticity and outlying observations in the data respectively.

Table 5 presents the result of artificial panel data set. The result indicates that the proposed method (WLS_{FMGt}) is the best and most efficient method as it gives the lowest standard error of HC4 and HC5, lowest standard error of the estimates, followed by WLS_{DRGP}, WLS_{RMD}, WLS_F, and lastly LWS.

Table 5: Regression estimates for the artificial panel data set

Estimator		Coeff. of Estimates	SE of Estimates	Stand. Error	
				HC4	HC5
LWS	b_1	3.5807	1.7731	15.3064	15.3064
	b_2	3.0539	1.7671	10.8533	10.8533
	b_3	-3.3994	1.8554	34.5642	34.5642
WLS_F	b_1	1.4244	1.4235	1.2206	1.2206
	b_2	1.8017	1.4079	2.2439	2.2439
	b_3	0.1837	1.3412	2.5540	2.5540
WLS_{RMD}	b_1	1.8430	1.4664	2.1272	2.1272
	b_2	2.1277	1.4289	2.4837	2.4837
	b_3	-0.4986	1.3625	4.4673	4.4673
WLS_{DRGP}	b_1	1.7388	1.4672	1.8703	1.8703
	b_2	2.0739	1.4265	2.4104	2.4104
	b_3	-0.3580	1.3508	3.8956	3.8956
WLS_{FMGt}	b_1	0.8436	0.7290	0.2403	0.2403
	b_2	1.5130	0.7249	0.2479	0.2479
	b_3	0.8268	0.7034	0.2447	0.2447

VII. Conclusion

The main focus of this study is to develop a reliable estimation method for FE panel data model for rectifying the problem of heteroscedasticity in the presence of HLPs. The performance of the LWS estimator is very poor. The RHCCM based on hat matrix weighting method, i.e WLS_F also not very efficient as it suffers from swamping and masking effect. In this study, we propose robust estimation methods in FE panel data model which employ residuals from weighted least squares (WLS) based on high leverage points detection measure (RMD, DRGP and FMGt) weighting methods in computing robust heteroscedasticity consistent covariance matrix (RHCCM) estimator. The results based on both simulation and numerical examples signify that the proposed RHCCM estimator based on FMGT outperformed the existing methods (LWS, WLS_F) and other proposed methods by providing the least bias and least standard errors of HC4 and HC5.

References

[1] Alguraibawi, M., Midi, H., and Imon, A. (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*, 2015.

[2] Bakar, N. M. A. and Midi, H. (2015). Robust centering in the fixed effect panel data model. *Pakistan Journal of Statistics*, 31(1).

[3] Bramati, M. C. and Croux, C. (2007). Robust estimators for the fixed effects panel data model. *The econometrics journal*, 10(3):521–540.

[4] Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, 45(2):215–233.

[5] Cribari-Neto, F., Souza, T. C., and Vasconcellos, K. L. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics—Theory and Methods*, 36(10):1877–1888.

[6] Furno, M. (1996). Small sample behavior of a robust heteroskedasticity consistent covariance matrix estimator. *Journal of Statistical Computation and Simulation*, 54(1-3):115–128.

[7] Greene, W. H. (2007). Econometric analysis.

[8] Habshah, M., Norazan, M., and Rahmatullah Imon, A. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, 36(5):507–520.

[9] Hausman, J. and Palmer, C. (2012). Heteroskedasticity-robust inference in finite samples. *Economics Letters*, 116(2):232–235.

[10] Kleiber, C. and Zeileis, A. (2008). Applied econometrics with r. springer-verlag.

[11] Lim, H. A. and Midi, H. (2016). Diagnostic robust generalized potential based on index set equality (drgp (ise)) for the identification of high leverage points in linear model. *Computational Statistics*, 31(3):859–877.

[12] Lima, V. M., Souza, T. C., Cribari-Neto, F., and Fernandes, G. B. (2009). Heteroskedasticity-robust inference in linear regressions. *Communications in Statistics-Simulation and Computation*, 39(1):194–206.

[13] MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

[14] Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

[15] Maronna, R., Martin, R. D., and Yohai, V. (2006). *Robust statistics*, volume 1. John Wiley & Sons, Chichester. ISBN.

[16] Rana, S., Midi, H., and Imon, A. (2012). Robust wild bootstrap for stabilizing the variance of parameter estimates in heteroscedastic regression models in the presence of outliers. *Mathematical Problems in Engineering*, 2012.

[17] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

[18] Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

[19] Ruppert, D. (1992). Computing s estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*, 1(3):253–270.

[20] Verardi, V. and Wagner, J. (2011). Robust estimation of linear fixed effects panel data models with an application to the exporter productivity premium. *Jahrbücher für Nationalökonomie und Statistik*, 231(4):546–557.

[21] Víšek, J. Á. (2011). Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, 47(2):179–206.

[22] Víšek, J. Á. (2015). Estimating the model with fixed and random effects by a robust method. *Methodology and Computing in Applied Probability*, 17(4):999–1014.

[23] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

[24] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, pages 642–656.