Journal of Basics and Applied Sciences Research (JOBASR)
ISSN print: 3026-9091
Volume 1(1) December 2023
DOI: https://doi.org/10.33003/jobasr-2023-v1i1-8

Robust Hausman Pretest for Panel Data Model in the Presence of Heteroscedasticity and Influential Observations

Muhammad Sani[1]*, Shamsuddeen Suleiman[1], Mustapha Isyaku[2]
[1]Department of Statistics, Federal University Dutsin-Ma Katsina State, Nigeria.
[2]Depatment of Mathematics, Federal University Dutsin-Ma Katsina State, Nigeria.
*Corresponding Authors Email: sanimksoro@gmail.com

Keywords:
Hausman pretest,
Panel data,
Heteroscedasticity.

ABSTRACT
The classical Hausman pretest (HT) is used to specified the right model between random and fixed effect panel data models. However, in the presence of heteroscedastic error variances and influential observations (IOs) in the data set, it may not correctly identify the right model. Therefore, this motivated us to proposed a new method termed Robust Hausman Test (RHT_{FIID}) which employed residuals from weighted least square (WLS) instead of OLS in the construction of heteroscedasticity consistent covariance matrix (HCCM) estimator. The weighting method is based on an efficient High Leverage Points (HLPs) detection method called Fast Improvised Influential Distance (FIID) which down weight only vertical outliers and bad HLPs. The good HLPs were allowed in the estimation as they might contribute to the precision of the estimate. The result indicates that the new proposed RHT_{FIID} outperformed the existing classical Hausman pretest by identifying the right model with and without heteroscedasticity and influential observations.

INTRODUCTION

In panel data analysis the daunting question of which model to choose between fixed and random effect has become subject of discussion. This is not easy as it might seem (Baltagi, 2008). In fact, the fixed versus random effects issue has generated a hot debate in the biometrics and statistics literature since 1960's. Some researchers were in support of the fixed effects (FE) model (Mundlak, 1961; Wallace and Hussain, 1969). However, Balestra and Nerlove (1966) were advocates of the random effect (RE) model. Later, Hausman (1978) proposed a specification test which is based on the difference between the fixed and random effects estimators. The classical Hausman pretest is used to choose between random and fixed effect panel data models. However, in the presence of heteroscedastic error variances and influential observations (IOs) in the data set, it may not correctly identify the right model. Hausman test (HT) is a common tool, used in many articles and textbooks in Econometrics (see Wooldridge, 2002; Hsiao, 2003; Baltagi, 2005; Greene, 2008; Muhammad et al. 2019).

In this paper, a robust Hausman test is proposed based on HLPs/IOs detection measure of Habshah et al. (2021) called Fast Improvised Influential Distance (FIID) and Robust Heteroscedasticity Consistent Covariance Matrix (RHCCM) estimator. The performance of the newly proposed robust method is assessed by some well-known real data sets and Monte Carlo simulation

MATERIAL AND METHOD

Consider a panel data model as,

$$y_{it} = u_i + x_{it}'\beta + \varepsilon_{it}, \quad i = 1,2,\dots,n \quad \text{and} \quad t = 1,2,\dots,T \qquad (1)$$

where, y_{it} is the response variable, x_{it} is the explanatory variable, u_i is the unobserved time-invariant effects and ε_{it} is the error term (idiosyncratic error) that is assumed to be normal, uncorrelated across individual units and time.

Hausman (1978) proposed a pretest for panel data model in order to decide whether the subsequent inference will be carried out using the fixed effects model or the random effects model. If the HT rejects the null hypothesis of no correlation between the explanatory variables and unobserved time invariant effect then the fixed effects model is chosen for subsequent inference, otherwise the random effects model is chosen. The test is based on the difference between the vectors of the coefficient of the estimates. The choice of the appropriate model is based on information about the individual-specific components and the exogeneity of the explanatory variables.

The null and alternative hypotheses are:

H_0: The appropriate model is Random Effects. There is no correlation between the error term and the explanatory variables in the panel data model.
$Cov(u_i, x_{it}) = 0$

H_1: The appropriate model is Fixed Effects. The correlation between the error term and the explanatory variables in the panel data model is statistically significant.
$Cov(u_i, x_{it}) \neq 0$

Table 1: Properties of random and fixed effect models estimators

Model	Hypothesis	H_0 is true	H_1 is true
Random Effect (RE)	$H_0: Cov(u_i, x_{it}) = 0$ Exogeneity	Consistent Efficient	Inconsistent
Fixed Effect (FE)	$H_1: Cov(u_i, x_{it}) \neq 0$ Endogeneity	Consistent Inefficient	Consistent

Hausman (1978) proved that the conventional Hausman statistic (HT) is asymptotically Chi-square distribution with p degree of freedom, where p is the number of regressors. The null hypothesis is rejected when the test statistic HT exceeds the Chi-square value at a given value of significant level. Meaning that, the null hypothesis will be rejected when the p-value is less than α level of significant. The HT is defined as,

$$HT = (\hat{\beta}_{RE} - \hat{\beta}_{FE})'(var(\hat{\beta}_{RE}) - var(\hat{\beta}_{FE}))^{-1}(\hat{\beta}_{RE} - \hat{\beta}_{FE}) \sim \chi^2_{(p)} \quad (2)$$

where, p is the number of explanatory variables, $\hat{\beta}_{RE}$ and $\hat{\beta}_{FE}$ are the classical random and fixed effect estimates. The variances of $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ estimates are obtained as follows;

$Var(\hat{\beta}_{FE}) = \hat{\sigma}_u^2 (X'X)^{-1}$
where $\hat{\sigma}_u^2 = \hat{u}^2/(nT - n - p)$ and $\hat{u} = y_{it} - x_{it}\hat{\beta}_{FE}$ (OLS residuals of the demeaned transformed data using Mean-centering).

$Var(\hat{\beta}_{RE}) = \hat{\sigma}_\varepsilon^2 (\tilde{X}'\tilde{X})^{-1}$
where$\hat{\sigma}_\varepsilon^2 = \hat{\sigma}_\tau^2 - \hat{\sigma}_u^2$,$\hat{\sigma}_\tau^2 = \hat{\tau}^2/(nT - k)$ and $\hat{\tau} = y_{it} - \tilde{x}_{it}\hat{\beta}_{RE}$ (OLS idiosyncratic residuals of the partially demeaned transformed data using Mean-centering)

Proposed Robust Hausman Test

It has been noticed that the HT is based on the OLS estimates of both FE and RE models. Maronna (2006) pointed out that in the presence HLPs the OLS method failed to correctly estimate the parameters which makes the HT test statistic very sensitive and easily affected by HLPs. To remedy this problem, the robust FE and RE estimates based on FIID proposed by Habshah et al. (2021) will be used in the construction of the proposed Robust Hausman test denoted by RHT_{FIID}. In the RHT_{FIID} test, all the OLS estimates are replaced by weighted least squares (WLS) estimates based on FIID weighting method (WLS_{FIID}). The proposed RHT_{FIID} test is given by,

$$HT_{FMGt} = \left(\hat{\beta}_{FIID}(RE) - \hat{\beta}_{FIID}(FE)\right)' \left(var\left(\hat{\beta}_{FIID}(RE)\right) - var\left(\hat{\beta}_{FIID}(FE)\right)\right)^{-1} \left(\hat{\beta}_{FIID}(RE) - \hat{\beta}_{FIID}(FE)\right) \sim \chi^2_{(p)} \quad (3)$$

where, p is the number of explanatory variables, $\hat{\beta}_{FIID}(RE)$ and $\hat{\beta}_{FIID}(FE)$ are the robust random and fixed effects estimate. Moreover, the variances of $\hat{\beta}_{FIID}(FE)$ and $\hat{\beta}_{FIID}(RE)$ estimates are obtained based on RHCCM estimator (HC5) as follows;

$$Var\left(\hat{\beta}_{FIID}(FE)\right) = \text{diag}\left\{(X'WX)^{-1}X'W\widehat{\Phi}_{5w}WX(X'WX)^{-1}\right\}$$

where, W is the FIID weight function, $\widehat{\Phi}_{5w} = \text{diag}\left\{\frac{\hat{u}_i^2}{\sqrt{(1-h_i^*)^{\alpha_i^*}}}\right\}$ for $i = 1, 2, \ldots, nT$ with $\alpha_i^* = \min\left\{\frac{h_i^*}{h^*}, \max\left\{4, \frac{kh^*_{max}}{h^*}\right\}\right\}$. $\hat{u}$ and X here are the OLS residuals of the demeaned transformed data and transform explanatory variable using MM-centering respectively.

$$Var\left(\hat{\beta}_{FIID}(RE)\right) = \text{diag}\left\{(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\ddot{\Phi}_{5w}W\tilde{X}(\tilde{X}'W\tilde{X})^{-1}\right\}$$

where, W is the FIID weight function, $\ddot{\Phi}_{5w} = \text{diag}\left\{\frac{\hat{\varepsilon}_i^2}{\sqrt{(1-h_i^*)^{\alpha_i^*}}}\right\}$ for $i = 1, 2, \ldots, nT$ with $\alpha_i^* = \min\left\{\frac{h_i^*}{h^*}, \max\left\{4, \frac{kh^*_{max}}{h^*}\right\}\right\}$. $\hat{\varepsilon}$ and $\tilde{X}$ here are the OLS idiosyncratic residuals of the partially demeaned transformed data and transform explanatory variable using MM-centering.

Distribution of Proposed RHT_{FIID} Statistic

The distribution of RHT_{FIID} test is hard to proof. It is anticipated to approximately follow same distribution as HT (Chi-square distribution with p degree of freedom). The RHT_{FIID} statistic will be verified to have similar distribution as HT statistic. This will make it more convenient for RHT_{FIID} test to be comparable with the HT test in specifying the right model for panel data as used by Muhammad et al. (2019) for Robust Whites Test.

Moreover, the null hypothesis of no correlation between the explanatory variables and unobserved time invariant effect is rejected if the RHT_{FIID} statistic exceeds the Chi-

square value with *p* degree of freedom at a given value of significance level where *p* is the number of regressors.

We hypothesized that the RHT_{FIID} test asymptotically follows Chi-square with *p* degree of freedom. To verify this distribution, consider a panel data regression model given in eqt (1) with three independent variables. The data was generated in the same way as Muhammad et al. (2021). Four sample sizes were considered $n = 10$, 15, 20, 25 with corresponding $t = 15$, 20, 25, and 30. Followed by estimating the FE and RE models using classical OLS and robust WLS_{FIID} for both conventional HT and $RHT_{FIID,}$ respectively. The distribution for the comparison will be Chi-square distribution with 3 degree of freedom (since *p*=3 number of regressors).

The test statistic HT and RHT_{FIID} are computed for each sample size. Figure 1 shows the Cumulative Density Function (CDF) plot for Chi-square with 3 degree of freedom and CDF plots for HT and RHT based on their Lagrange Multiplier for some sample sizes. It can be observed that the CDF of HT and RHT are following the CDF of Chi-square distribution with 3 degree of freedom. The plots show that both HT and RHT statistics are asymptotically following Chi-squared distribution with 3 degree of freedom.

The mean and variance of the Lagrange Multiplier for HT and RHT_{FIID} had also been computed for each sample. The process is repeated for 1,000 times and the values of their mean and variance are calculated. If HT and RHT_{FIID} statistics follow Chi-square distribution with 3 degree of freedom, their mean and variance shall be equal to 3 and 6, respectively. The results are shown in Table 2. It can be seen very clearly that the mean and variance of both HT and RHT_{FIID} statistics are reasonably closed to 3 and 6, respectively. The values of mean and variance are getting closer to the expected mean and variance with the increase of sample size. This finding supports that the HT and RHT_{FIID} statistics are asymptotically following the Chi-square distribution with 3 degree of freedom.

The scientific method of Cramer-von Mises one sample test (Choulakian et al., 1994) is applied to verify that both HT and RHT statistics follow Chi-square distribution with 3 degree of freedom. Let $x_1, x_2, \dots, x_n$ be the observed values in increasing order. The Cramèr-von Mises one sample test statistic is given by,

$$T = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(x_i)\right]^2$$

Where $F(x_i)$ is the cumulative distribution of the hypothesized function. The 5, 10, 15, 20,.., until 95 percentiles (constant increase of 5 percentile) are obtained for the Lagrange Multiplier based on HT and RHT_{FIID} for each sample size. The Cramer-von Mises one sample test is then carried out for each sample size. The null hypothesis of no difference between HT and RHT_{FIID} statistics for Chi-square distribution with 3 degree of freedom will be rejected if the *p*-value is less than 0.05 significance level. The results are presented in Table 3.

In addition, the most powerful Anderson-Darling test (Rahman et al., 2006) is applied to reaffirm that both HT and RHT_{FIID} statistics follow Chi-square distribution with 3 degree of freedom. Again, let $x_1, x_2, \dots, x_n$ be the observed values in increasing order. The Anderson-Darling test statistic is given by,

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1).[\ln F(x_i) + \ln(1 - F(x_{n-i+1})]$$

where$F(x_i)$ is the cumulative distribution of the hypothesized function. The same procedure discussed in Cramer-von Mises one sample test above is followed to get the 5 to 95 percentiles of the Lagrange Multiplier based on HT and RHT_{FIID} test for the same sample sizes and perform the Anderson-Darling test. The null hypothesis is that HT and RHT_{FIID} statistics follow Chi-square distribution with 3 degree of freedom. The null hypothesis will be rejected if the *p*-value is less than 0.05 significance level. The findings are exhibited in Table 4. It is very encouraging to see that all the *p*-values are greater than 0.05 significance level thorough the studies. This finding reinsures that HT and RHT_{FIID} statistics are following Chi-square distribution with 3 degree of freedom.

Simulation Study and Real Data Examples

In this paper, we used several real data sets and Monte Carlo simulation to evaluate the performance of the proposed robust Hausman test (RHT_{FIID}) and the existing Hausman test (HT).

Monte Carlo Simulation Studies

A Monte Carlo simulation based on RE model has been generated. The percentage of null rejection rate has been computed. In this case the method that has the lowest percentage of null rejection is considered the best.

Consider the panel data model given by eqt (1). Three explanatory variables $(x_{it1}, x_{it2}, x_{it3})$ and $u_i$were generated from standard normal distribution. The true parameters $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$. Three sample sizes $n =$ 5, 10 and 15 with the corresponding $t = 10$, 15 and 20 were replicated twice to form $n = 10$, 20, 30 and $t = 20$, 30, 40 respectively, in order to create heteroscedasticity. The strength (degree) of heteroscedasticity is measured by $\lambda = \max(\sigma_\varepsilon^2) / \min(\sigma_\varepsilon^2)$. Following the idea of (Lima et. al. 2009) the skedastic function is defined as $\sigma_\varepsilon^2 = \exp\{c_1 x_{it1} + c_1 x_{it2}\}$, $c_1$can be chosen between 0 and 1. For $c_1 = 0.25$ the value of $\lambda = 12.46$ and for $c_1 = 0.65$ the value of $\lambda = 56.75$. The value of λ indicate the degree of the heteroscedasticity in the data, whereby for homoscedasticity $\lambda = 1$. Regular observations for both *X* and *y* were replaced with data points generated from

normal distribution $N(10,1)$ at 1%, 2%, 3%, 4% and 5% contamination level for all the sample sizes at the average of 2000 replications.

Real Data Examples

Airline Data set: A panel data set for six airline firms was used, taken from Greene (2008). The data set contained 90 observations for the period (1970 to 1984) with response variable cost and three predictor variables (output, fuel price, and load factor). A multiple linear panel data model was constructed to study the efficiency in production of airline services.

The airline data was modified by inflating the explanatory variables of the first observation by 10 in order to create 1 HLP contamination in the data set. Secondly, the 30th observation was also inflated in the same way in order to form 2 HLPs contamination in the data.

Grunfeld data set

The first 5 firms of Grunfeld (1958) data set containing 100 observations were used. This data set was taken from Kleiber and Zeileis (2008) contains 20 annual observations for 11 US firms for the years 1935–1954. Three variables real gross investment (invest) as response variable, real value of the firm (value) and real value of the capital stock (capital) as explanatory variables were observed. These data have been used in many textbooks of econometrics.

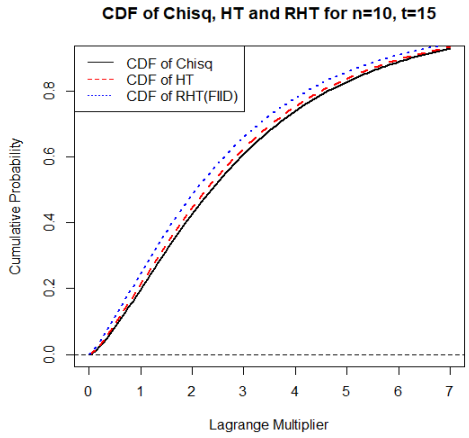
We modified the Grunfeld data by inflating the explanatory variables of observation 19 by 100 to create 1 bad HLP in the data set, and we also inflate observation 20 in the same way as observation 19 making another data set with 2 HLPs.

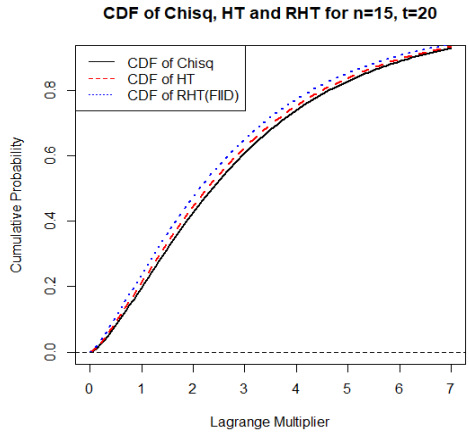
Artificial data set

An artificial heteroscedastic panel data set with n=6 and t=20 number of observations was generated. A sample of n=3 and t=10 were replicated twice to form n=6 and t=20 respectively. The response variable is generated in the same way as eqt (1), the explanatory variables and individual effect (u_i) are generated from normal distribution $N(10,1)$ and $N(0,1)$, respectively. The true parameters $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$. The skedastic function is defined as $\sigma_\varepsilon^2 = \exp\{c_1 x_{it1}\}$ (Lima et al. 2009) where the value of $c_1 = 0.75$ was chosen such that $\lambda \approx 42$. The value of λ indicate the degree of the heteroscedasticity in the data, whereby for homoscedasticity, $\lambda = 1$. The strength (degree) of heteroscedasticity is measured by $\lambda = \max(\sigma_\varepsilon^2) / \min(\sigma_\varepsilon^2)$.

The artificial heteroscedastic panel data set has been modified by introducing bad HLPs, the first observation was inflated by 10 for all the explanatory variables to form a data set with 1 bad HLP contamination. Also, the last observation was inflated similar to the first observation to form a data set with 2 HLPs contamination

RESULT AND DISCUSSION





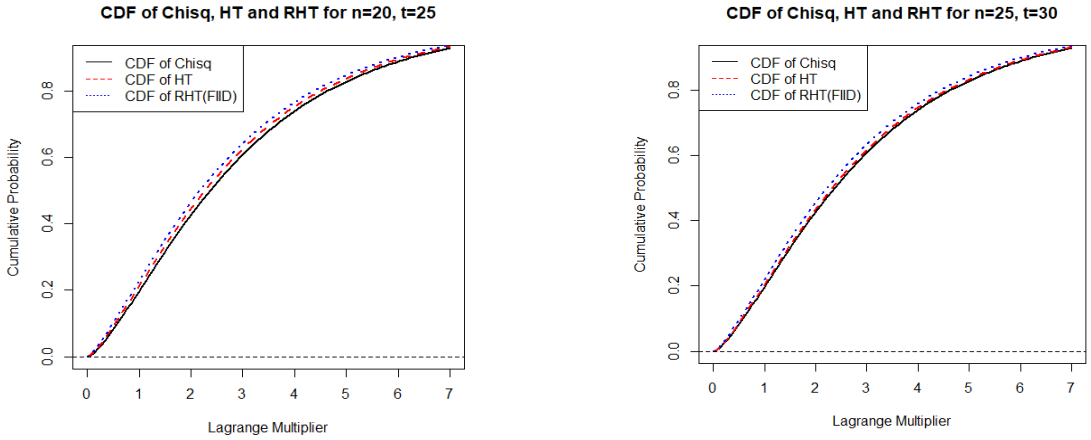


Figure 1: CDF of Chi-square, HT and RHT for some sample sizes

The coefficient of determination R^2 was computed when the Lagrange Multiplier of HT and RHT_{FIID} tests are regressed versus the theoretical Chi-square. This coefficient is used to measure the quality of the fitted model (Richard and Dean, 2002). The results are presented in Table 2. The high value of R^2 indicates that HT and RHT_{FIID} statistics follow Chi-square distribution with 3 degree of freedom.

Table 2: Mean and Variance of HT and RHT_{FIID} Statistic, R^2 of Test Statistic of HT and RHT and Theoretical Chi-Square

Tests	Values	Samples			
		n=10 t=15	***n=15 t=20***	***n=20 t=25***	***n=25 t=30***
HT	Mean	3.1097	3.0651	3.0101	3.0201
	Variance	6.1209	5.9347	5.9204	6.0302
	R^2	0.9797	0.9817	0.9823	0.9799
RHT_{FIID}	Mean	3.1095	3.0395	3.0197	3.0198
	Variance	6.1221	5.9815	5.9210	5.9209
	R^2	0.9795	0.9809	0.9796	0.9798

It is very interesting to see that all the *p*-values are greater than 0.05 significance level for all the sample sizes examined. This finding shows that HT and RHT_{FIID} statistics are following Chi-square distribution with 3 degree of freedom.

Table 3: Cramer-von Mises One Sample Test for Testing the Distribution of HT and RHT_{FIID} Statistics

Test	Cramèr-von Mises	Samples			
		n=10 t=15	*n=15 t=20*	*n=20 t=25*	*n=25 t=30*
HT	*T*	0.0381	0.0550	0.0365	0.0872
	p-values	0.6983	0.4218	0.7284	0.1563
RHT_{FIID}	*T*	0.0301	0.0282	0.0554	0.0712
	p-values	0.8346	0.8611	0.4159	0.2552

Table 4: Anderson-Darling Test for Testing the Distribution of HT and RHT_{FIID} Statistics

Test	Anderson-Darling	Samples			
		n=10 *t=15*	*n=15* *t=20*	*n=20* *t=25*	*n=25* *t=30*
HT	A^2	0.4138	0.5414	0.6343	0.7089
	p-values	0.8658	0.6623	0.5611	0.4956
RHT_{FIID}	A^2	0.4134	0.5326	0.6474	0.7242
	p-values	0.8669	0.6733	0.5488	0.4836

Table 5 to 7 exhibit the performance of the proposed methods (RHT_{FIID}) and the existing methods (HT) in a simulated heteroscedastic random effect panel data with different sample sizes and HLPs contamination level. The results show that all the proposed RHT_{FIID} was more efficient than the existing methods, by providing smallest percentage of null rejection, which indicate the RHT_{FIID} fail to reject the null hypothesis of random effect model is appropriate.

Table 5: Percentage of Null rejection rates of Hausman test for Random Effect simulated panel data n=10 t=20

HLPs	$\lambda = 12.46$		$\lambda = 56.75$	
Con.	HT	RHT_{FIID}	HT	RHT_{FIID}
0%	0.00	0.00	0.00	0.00
1%	1.35	$< 10^{-6}$	1.70	$< 10^{-6}$
2%	2.10	$< 10^{-6}$	3.15	$< 10^{-6}$
3%	5.61	$< 10^{-6}$	5.25	0.05
4%	10.65	0.05	10.05	0.10
5%	17.75	0.10	18.30	0.10

Table 6: Percentage of Null rejection rates of Hausman test for Random Effect simulated panel data n=20 t=30

HLPs	$\lambda = 12.46$		$\lambda = 56.75$	
Con.	HT	RHT_{FIID}	HT	RHT_{FIID}
0%	0.00	0.00	0.00	0.00
1%	1.20	$< 10^{-6}$	2.05	$< 10^{-6}$
2%	2.05	$< 10^{-6}$	1.90	$< 10^{-6}$
3%	2.85	$< 10^{-6}$	2.95	$< 10^{-6}$
4%	3.85	$< 10^{-6}$	4.20	0.05
5%	8.00	0.05	9.55	$< 10^{-6}$

Table 7: Percentage of Null rejection rates of Hausman test for Random Effect simulated panel data n=30 t=40

HLPs	$\lambda = 12.46$		$\lambda = 56.75$	
Con.	HT	RHT_{FIID}	HT	RHT_{FIID}
0%	0.00	0.00	0.00	0.00
1%	0.85	$< 10^{-6}$	1.10	$< 10^{-6}$
2%	1.05	$< 10^{-6}$	1.60	$< 10^{-6}$
3%	2.15	$< 10^{-6}$	2.10	$< 10^{-6}$
4%	2.95	$< 10^{-6}$	2.50	$< 10^{-6}$
5%	3.80	$< 10^{-6}$	4.85	$< 10^{-6}$

Figure 2 and 3 indicates the presence of heteroscedasticity and Good Leverage Points GLPs (observation 28) in the data. The existence of funnel shape in Figure 2 indicates the presence heteroscedasticity in the data.

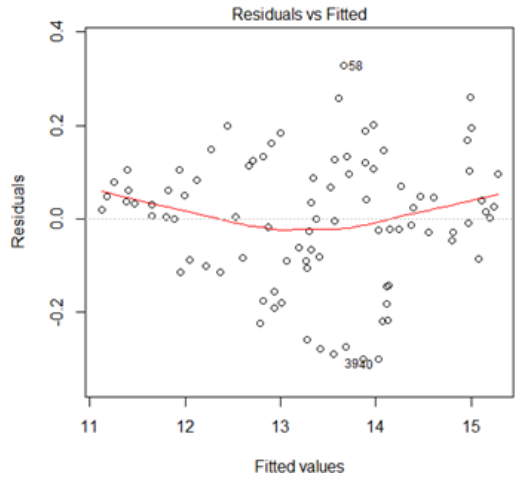


Figure 2: Plot of pooled OLS residuals versus fitted

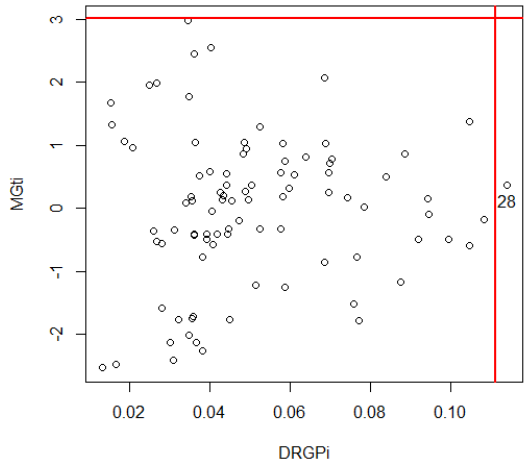


Figure 3: Plot of FIID for airline data values for airline data

The result from Table 8 show that both HT and RHT_{FIID} fail to reject H_0 in the absence of HLPs, but in the presence of 1 HLPs and 2 HLPs the HT reject H_0. This indicates that the proposed method (RHT_{FIID}) was found to be the best and the most resistance against the effect of HLPs as it provides the same result (fail to reject H_0) for both original and modified data sets (with HLPs).

Table 8: Power of Hausman tests for original and modified Airline data set, n = 6, t = 15, p = 3. The critical value: $\chi^2_{(3,\ \alpha=0.05)} = 7.818$

Tests	**Original data**		**Modified data with (1 HLP)**		**Modified data with (2 HLPs)**	
	Value of statistic	P-value	Value of statistic	P-value	Value of statistic	P-value
HT	1.6150[a]	0.6560	8.3462[b]	0.0394	67.4527[b]	1.4e-14
RHT_{FIID}	1.0841[a]	**0.7809**	0.6372[a]	**0.8878**	0.6455[a]	**0.8859**

Note: a=fail to reject Ho, b=reject Ho and $\alpha = 0.05$

Figure 4 and 5 indicates the presence of heteroscedasticity and GLPs (observation 41, 42, 67, 68, 73, 74) in the data. The presence of funnel shape in Figure 4 shows that heteroscedasticity exist in the data.

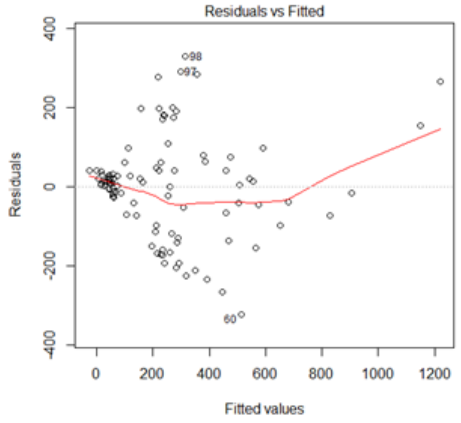


Figure 4 Plot of pooled OLS residuals versus fitted values for grunfeld data

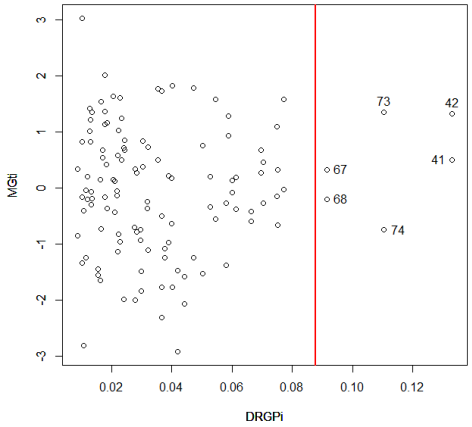


Figure 5: Plot of FIID for grunfeld data

Table 9 shows that, the proposed RHT_{FIID} provide the same result for both original and modified contaminated data set with almost equal p-value, but the HT in the presence of HLPs reject H_0 while in the absence of HLPs (original data) fail to reject H_0.

Table 9: Power of Hausman tests for original and modified first 5 firms of Grunfeld data, n = 5, t = 20, p = 2. The critical value: $\chi^2_{(2,\alpha=0.05)} = 5.991$

Tests	Original data		Modified data with 1 HLP		Modified data with 2 HLPs	
	Value of statistic	P-value	Value of statistic	P-value	Value of statistic	P-value
HT	0.3962[a]	0.8202	38.723[b]	3.9e-09	49.939[b]	2.1e-09
RHT_{FIID}	0.4835[a]	**0.7852**	0.4972[a]	**0.7798**	0.3843[a]	**0.8252**

Note: a=fail to reject Ho, b=reject Ho and $\alpha = 0.05$

Figure 6 indicates the presence of heteroscedasticity due to presence of funnel shape produced in the plot. Similarly, Figure 7 indicates the presence of GLPs (41, 42, 67, 68, 74) and IO (73) in the data set.

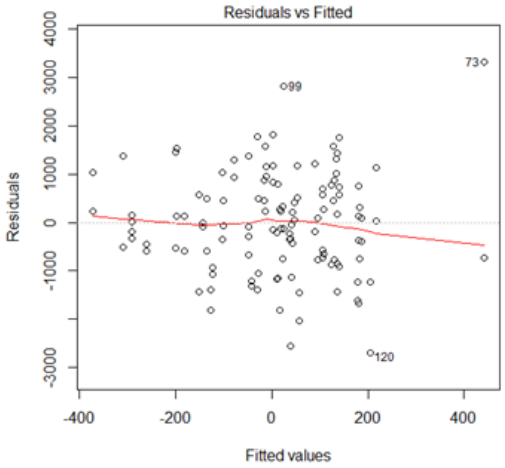


Figure 6: Plot of pooled OLS residuals versus fitted values for artificial data

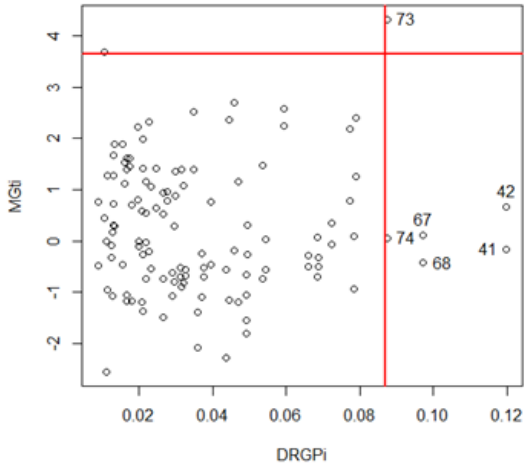


Figure 7: Plot of FIID for artificial data

Table 10 presents the result of both the original and modified data sets which clearly indicates that RHT_{FIID} method outperformed the existing method by providing same result for both original and modified contaminated data sets.

Table 10: Power of Hausman tests for original and modified Artificial data set n = 6, t = 20, p = 3. The Critical Value: $\chi^2_{(3,\ \alpha=0.05)} = 7.818$

Tests	Original data		Modified data with 1 HLP		Modified data with 2 HLPs	
	Value of statistic	P-value	Value of statistic	P-value	Value of statistic	P-value
HT	1.0e+17[b]	0.0000	-4.6e+12[a]	1.0000	-7.8e+12[a]	1.0000
RHT_{FIID}	3.1e+14[b]	**0.0000**	3.2e+14[b]	**0.0000**	3.4e+14[b]	**0.0000**

Note: a=fail to reject Ho, b=reject Ho and $\alpha = 0.05$

CONCLUSION

The Hausman pretest is used to choose between models in panel data studies, which specifies whether random or fixed effects panel data model should be used. The test examines the presence of endogeneity in panel data model. The use of panel data gives considerable advantages over the time series or cross-sectional data, but the appropriate model to be used is of great importance for obtaining consistent and efficient results. However, presence of heteroscedasticity and a single high leverage point in a data set can mislead the result of Hausman pretest. Therefore, this paper provides a robust method for a Hausman pretest in the presence of heteroscedasticity of unknown form and Influential observations (IOs). The new proposed method (RHT_{FIID}) used residuals from weighted least square (WLS) instead

of OLS residuals in the construction of heteroscedasticity consistent covariance matrix (HCCM) estimator. The weighting method used is based on very efficient HLPs detection measure (FIID), which down weight only vertical outliers and bad HLPs. The good HLPs were allowed in the estimation as they might contribute to the precision of the estimate. The result indicates that the new proposed RHT_{FIID} outperformed the existing HT method and indicates it resistivity to effect of HLPs and heteroscedasticity. This good performance of the proposed method is due the efficiency of FIID weighting method which has less swamping and masking effect.

REFERENCES

Baltagi, B. (2008). *Econometric analysis of panel data*, John Wiley & Sons.

Baltagi B.H. (2005). *The Econometrics of Panel Data*. John Wiley & Sons, New York.

Balestra P. and Nerlove M. (1966) Pooling cross-section and time series data in the estimation of a dynamic model: the demand for natural gas. *Econometrica* 34: 585-612

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994). Cramér-von Mises statistics for discrete distributions. *Canadian Journal of Statistics*, 22(1), 125-137.

Greene, W. (2008) *Econometric Analysis*; New York: Pearson

Grunfeld Y. (1958) *The Determinants of Corporate Investment.* Unpublished Ph.D. dissertation, Department of Economics, University of Chicago

Habshah Midi, Muhammad Sani Shelan Saied Ismaeel (2021) Fast Improvised Influential Distance for the Identification of Influential Observations in Multiple Linear Regression. *Sains Malaysiana*, 50 (7) (2021): 2085-2094 http://doi.org/10.17576/jsm-2021-5007-22

Hausman J.A. (1978) Specification tests in econometrics, *Econometrica* 46, 1251–1271

Hsiao C. (2003) *Analysis of Panel Data*, 2nd edition. Cambridge University Press.

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

Maronna, R. A., et al. (2006). "Wiley Series in Probability and Statistics." Robust Statistics: Theory and Methods: 404-414

Muhammad Sani, Habshah Midi & Jayanthi Arasan (2019) Robust Parameter Estimation for Fixed Effect Panel Data Model in the Presence of Heteroscedasticity and High Leverage Points, *ASM Science. Journal 12, Special Issue 1, 2019 for IQRAC2018, 227-238*

Muhammad Sani, Shamsuddeen Suleiman & Baoku Ismail G (2020) Robust Parameter Estimation for Random Effect Panel Data Model In The Presence Of Heteroscedasticity And Influential Observations, *FUDMA journal of sciences*, 4(4): 561- 569, ISSN; 2645 – 2944

Muhammad S., Habshah M. & Babangida I.B. (2019) Robust White's Test For Heteroscedasticity Detection In Linear Regression *FUDMA journal of sciences*, 3(2): 173- 178, ISSN; 2645 – 2944

Mundlak Y. (1961) Empirical production function free of management bias, *Journal of Farm Economics*43, 44–56

Rahman, M., Pearson, L. M., and Heien, H. C. (2006). A modified anderson-darling test for uniformity. *Bulletin of the Malaysian Mathematical Sciences Society*, 29(1).

Richard, A. J., and Dean, W. W. (2002). *Applied multivariate statistical analysis*. London: Prenticee Hall, 265

Wallace, T.D. and Hussain A. (1969). The use of error components models in combining cross-section and time-series data, *Econometrica*37, 55–72

Wooldridge, J.M., (2002) *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, London.