ADDIS ABABA UNIVERSITY
OFFICE OF GRADUATE PROGRAM
FUCULITY OF SCIENCE
DEPARTMENT OF STATISTICS

A COMPARATIVE SIMULATION STUDY OF THE HETEROSCEDASTICITY CONSISTENT COVARIANCE MATRIX ESTIMATORS IN THE LINEAR REGRESSION MODEL

BY
YEGNANEW ALEM

A Thesis Submitted to the Office of Graduate Program of Addis Ababa University in Partial fulfillment of the Requirement for the Degree of Master of Science in Statistics

July, 2008

ADDIS ABABA UNIVERSITY

OFFICE OF GRADUATE PROGRAM

FUCULITY OF SCIENCE

DEPARTMENT OF STATISTICS

A COMPARATIVE SIMULATION STUDY OF THE HETEROSCEDASTICITY CONSISTENT COVARIANCE MATRIX ESTIMATORS IN THE LINEAR REGRESSION MODEL

BY

YEGNANEW ALEM

Approved by the Board of Examiners:

Sileshi Fanta	
Department Head	Signature
Butte Gotu	
Internal Examiner	Signature
Emmanuel G. Yohannes	
External Examiner	Signature

Addis Ababa, Ethiopia

ACKNOWLEDGEMENT

I would like to extend my deepest appreciation to my advisor Olusanya E.Olubusoye (Ph.D) without whose contributions I would have had a lesser product.

My sincere gratitude goes to my family especially to Abeba, for their persistent love and encouragement.

My special thanks go to Dr. Emmanuel G.Yohannes and Ato Milion Atsbeha who helped me in writing the code for simulation using Matlab 7.0.

I would also like to thank all of my friends typically of which Ato Amare, Ato Eshetu, Ato workalemaw and Ato Solomon to their many excellent suggestions.

Most of all I want to express my gratitude to my Heavenly Father for granting me the opportunity to study at this university and for giving me the strength to push past my own limitations.

ACRONYMS

. HCCM: Heteroscedasticity Consistent Covariance Matrix

. CLRM: Classical Linear Regression Model

. BLUE: Best Linear Unbiased Estimators

. OLSCM: Ordinary Least Square Covariance Matrix

. HCSEs: Heteroscedasticity Consistent Standard Errors

. GLS: Generalized Least Square

. WLS: Weighted Least Square

. LM: Lagrange Multiplier

ABSTRACT

In the context of econometric methods of estimation the variances of OLS estimates derived under the assumption of homoscedasticity are not consistent when there is heteroscedasticity and their use can lead to incorrect inferences. Thus, this paper sets out to examine the performance of several modified versions of heteroscedasticity consistent covariance matrix (HCCM) estimator (namely HCO, HC1, HC2, and HC3) of White (1980) and white and Mackinnon (1985) over a range of sample sizes. Most applications that use HCCM appear to rely on HCO, yet tests based on the other HCCM estimators are found to be consistent even in the presence of heteroscedasticity of an unknown form. Based on Monte Carlo experiments which compare the performance of the t statistic, it was found out that HC2 and HC3 estimators precisely out perform the others in small samples. In particular HC3 estimator for samples of size less than 100 was found to be better than the other HCCM estimators; when samples are 250 or larger, other versions of the HCCM can be used. Added to that, it was cost advantageous to employ HC3 instead of ordinary least square covariance matrix (OLSCM) even when there is little evidence of hetreoscedasticity.

Key words

White estimator, Monte Carlo Simulation, Linear Regression, Heteroscedasticity

TABLE OF CONTENTS

CHAPTER ONE

1. INTRODUCTION

This chapter is divided into six sections. The first section gives the background of the study, the second section is about statement of the problem, the third section gives the objectives of the study, the fourth section states research questions, the fifth section explains the limitation of the study and the final section is about significance of the study.

1.1 Background of the Study

Applied econometricians extensively use the linear regression model. Together with its numerous generalizations, it constitutes the foundation of most empirical work in econometrics.

Despite this fact, little is known about properties of inferences made from this model when standard assumptions are violated. In particular, classical techniques require one to assume that the error terms have a constant variance.

This assumption is often not very plausible. Nevertheless, a way of consistently estimating the variance-covariance matrix of ordinary least squares estimates in the face of heteroskedasticity of unknown form is available (e.g. see White (1980)).

This heteroskedasticity-consistent covariance matrix estimator allows one to make valid inferences provided the sample size is sufficiently large.

By the assumption of the classical normal linear regression model, we have

$$E(\varepsilon_i^2) = \sigma^2 \text{ for all i.}$$

Since the mean of ε_i is assumed to be zero, we can write,

Var $(\varepsilon_i) = \sigma^2$ for all i.

This feature of the regression disturbances is known as homoscedasticity. It implies that the variance of the disturbance is constant for all observations.

This assumptions may not be trouble some for models involving observations on aggregates over time, since the values of the explanatory variable are typically of similar order of magnitude at all points of observation, and the same is true of the values of the dependant variable. For example, in an aggregate consumption function, the level of consumption in recent years is of a similar order of magnitude as the level of consumption twenty years ago, and the same is true of income.

Unless there are some special circumstances, the assumption of homoscedasticity in aggregate models seems plausible. However, when we are dealing with micro economic data, the observations may involve substantial differences in magnitude, as, for example, in the case of data on income and expenditure of individual families.

Here the assumptions of homoscedsticity are not very plausible on a priori grounds since we would expect less variation in consumption for low- income families than for high-income families. At low levels of income, the average level of consumption is low, and variation around this level is restricted: consumption cannot fall too far below the average level because this might mean starvation, and it cannot rise too far above the average because the asset and the credit position do not allow it.

These constraints are likely to be less binding at higher income levels. Empirical evidence suggests that these priori considerations are in accord with actual behavior. The appropriate model in this and other similar cases is then one with heteroscedastic disturbances. Regression disturbances whose variances are not constant across observations are **heteroscedastic.** Heteroscedasticty arises in numerous applications, in both cross-section and time series data. However, it is most commonly expected in cross-sectional data. For example, even after accounting for firm sizes, we expect to observe

greater variation in the profits of large firms than in those of small ones. The variance of profits might also depend on product diversification, research and development expenditure, and industry characteristics and therefore might vary across firms of similar sizes.

When analyzing family spending patterns, we find that there is greater variation in expenditure on certain commodity groups among high income families than low ones due to the greater discretion allowed by higher incomes.

In the heteroscedastic regression model,

$$\text{Var}(\varepsilon_i) = \sigma_i^2, \qquad i=1,\ldots,n$$

We continue to assume that the disturbances are pair wise uncorrelated.

That is, $E(\varepsilon_i \varepsilon_j) = 0, \qquad (i \neq j)$

When heteroscedasticity alone occurs, there are n + k unknown parameters; n unknown variances and k elements in β vector. Without some additional assumptions, estimation from n sample points is clearly impossible. Additional assumptions are usually made about the disturbance process.

The ordinary least squares (OLS) linear regression model is widely used throughout the physical, natural and social sciences. In OLS linear regression, a vector of regression coefficients, **β**, in a model of the form

$$Y = X\beta + \varepsilon$$

is estimated, where Y is a column vector of dependant variable to be estimated, X is a matrix of predictor variable values, and ε is a vector of errors. The elements in β provide information about a predictor variable's unique or partial relationship with the dependant variable, controlling for the other predictor variables.

It is well known that the presence of heteroscedasticity in the disturbances of an otherwise properly specified linear model leads to consistent but inefficient parameters estimates and inconsistent covariance matrix estimates. As a result, faulty inferences will be drawn when testing statistical hypotheses in the presence of heteroscedasticity.

Researchers often are interested in testing the null hypothesis that a specific element in β is zero, or forming a confidence interval (CI) around the estimate. It is well known that such inferential methods assume homoscedasticity in the errors. Violations of homoscedasticity can yield hypothesis tests that fail to keep false rejections at the nominal level, or confidence intervals that are either too narrow or too wide. Given that homoscedasticity is often an unrealistic assumption or clearly violated based on the data available, the researcher should be sensitive to if and how his or her results may be affected by heteroscedasticity.

Based on the work of Long and Ervin (2000), several HCCM methods of estimating the standard error of regression coefficient that can be used if the researcher is concerned about the effects of heteroscedasticity on hypothesis tests and confidence intervals. All of the HCCM methods described are based on an approximation of the variance covariance matrix of the estimated regression coefficients using the square of the residuals (e_i^2, where $e_i = y_i - x_i\hat{\beta}$ from an OLS linear regression.

The four HCCM methods HCO, HC1, HC2 and HC3 differ in how those squared residuals are used in the estimation processes. Once the heteroscedasticity-consistent covariance matrix (HCCM) is estimated, the standard errors for the regression coefficients are simply the square root of the diagonal elements of the HCCM. Since covariance matrix estimators are most frequently used to construct test statistics, we focus on the behavior of t statistic constructed using these different estimators.

In this study, we tried to investigate the performance of the HCCM methods over a range of sample sizes. In other words, we wished to empirically assess their performance.

1.2. Statement of the Problem

Since covariance matrix estimators are used to compute test statistics, we wish to empirically assess the small and large sample performance of the four HCCM and OLSCM t statistics to test hypotheses that particular element of β assume specified values.

1.3. Objectives of the Study

1.3.1 General Objectives:

- To compare the performance of various HCCM estimators over a range of sample sizes.

1.3.2 Specific Objectives:

- To generate disturbances with four different structures of heteroscedasticity and at different sample sizes.
- To apply the OLSCM and the four HCCM methods to the simulated data generated under the above condition.
- To evaluate the performance of the OLSCM and the HCCM methods over a range of sample sizes.

1.4. Research Questions

This paper will answer the following research questions:

1. Which HCCM performs better for sample sizes: 10, 25, 50, 100, 250, 500, 600 and 1000?
2. What is the consequence of using HCCM tests when there is no heteroscedasticity?
3. What is the consequence of using OLSCM test when there is heteroscedasticity?

1.5 Limitation of the Study

While no Monte Carlo simulation can cover all variations that might influence the properties of the statistics being studied, our simulations explore a medium range of statistics that are common in cross- sectional data.

1.6 Significance of the Study

It is well known that violations of homoscedasticity can yield hypothesis tests that fail to keep false rejections at the nominal level, or confidence intervals that are either too narrow or too wide. Given that homoscedasticity is often unrealistic assumption or clearly violated based on the data available, the researcher should be sensitive to if and how his and her results may be affected by heteroscedasticity.

The use of HCCM allows a researcher to easily avoid the adverse effects of heteroscedasticity even when nothing is known about the form of heteroscedasticity.

The HCCM provides a consistent estimator of the covariance matrix of the regression coefficients in the presence of heteroscedasticity of unknown form.

This is particularly useful when a suitable variance- stabilizing transformation or weights cannot be found, or weights cannot be estimated for use in GLS.

In this paper we have added our own contributions based on Monte Carlo simulation and recommended that HC2 or HC3 should be used in favor of HC0.

CHAPTER TWO

2. THEORETICAL FRAMEWORK AND LITERATURE REVIEW

In order to fully consider each of the estimators discussed in the previous chapter, the theoretical framework and literature review detailing these methods will be examined. This chapter is divided into five sections. The first section is about the nature of heteroscedasticity, the second section discuses the consequences of heteroscedasticity, the third section is about detecting heteroscedasticity, the fourth section is controlling for heteroscedasticity and estimation of $Cov(\hat{\beta})$, and the final section is a review of previous simulation studies involving the HCCM estimator.

2.1 The Nature of Heteroscedasticity

To illustrate the nature of heteroscedasticity, recall the two-variable regression model of consumption:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \qquad (2.1.1)$$

where Y_i = monthly household expenditure; and X_i = monthly household income. Now, previously we assumed, at least implicitly, that the population error term, ε_i, is homoscedastic, that is the variance of each ε_i, conditional on X_i, is some constant equal to σ^2 , or $E(\varepsilon_i^2) = \sigma^2$.

Intuitively, what this says is that the spread, or dispersion, of monthly expenditures among low-income households is the same as among high income households. But maybe this assumption is a little strong, for high income households have more scope for choice about the disposition of their income and thus more choices about their consumption behavior. Thus, we might expect a greater dispersion of monthly expenditures for higher-income households than for lower-income households. In this

case, the stochastic disturbance term, ε_i, is not homoscedastic, but rather heteroscedastic, with the variance of Y_i increases as X_i increases, that is $E(\varepsilon_i^2) = \sigma_i^2$ for all i.

Now in general, there are several reasons why stochastic disturbance terms may be heteroscedastic, five of which are:

1. Learning by doing: As people learn, their errors of behavior become smaller over time. In this case σ_i^2 would be expected to decrease. For example, consider the relationship between productivity (wages) and experience.
2. Scope for choice: As incomes grow, people have more discretionary income and thus more choices about consumption and saving behavior. Similarly, companies with larger profits are generally expected to show greater variability in their dividend policies than companies with lower profits.
3. Improvement in data collecting techniques: As data collecting techniques improve, σ_i^2 is likely to decrease. Thus, institutions and organizations that have sophisticated data processing equipment are likely to commit fewer errors.
4. Outliers: An outlying observation, or outlier, is an observation that is much different in relation to the other observations in the sample. The inclusion or exclusion of such an observation, especially if the sample size is small, can substantially alter the results of regression analysis.
5. Misspecified regression models: Very often what looks like heteroscedasticity may be due to the fact that some important variables are omitted from the model.

2.2 The Consequences of Heteroscedasticity

Suppose our data on household consumption expenditure is plagued by heteroscedasticity. What happens to our OLS estimators and their variances? To answer this question, let us just focus on the slope coefficient β_1 in our two-variable model of household consumption:

$$\hat{\beta}_1 = \frac{\sum y_i x_i}{\sum x_i^2} \qquad (2.2.1)$$

where, $y_i = Y_i - \overline{Y}$ and $x_i = X_i - \overline{X}$

. While the coefficient estimate is unaffected, its variance is:

$$Var\left(\hat{\beta}_1\right) = \frac{\sum x_i^2 \sigma_i^2}{\left(\sum x_i^2\right)^2}, \qquad (2.2.2)$$

where had the data not been plagued by heteroscedasticity, the variance would be:

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum x_i^2}, \qquad (2.2.3)$$

Hence, the consequences of heteroscedasticity are as follows:

1. OLS estimators and forecasts based on them remain unbiased and consistent.
2. However, OLS estimators are no longer BLUE because they are no longer efficient. As a result forecasts will be inefficient.
3. Because regression coefficient variances and covariances are biased and inconsistent, tests of hypothesis, that is *t*-tests and *F* tests, are invalid.

In short, if we persist in using the usual testing procedures despite heteroscedasticity, whatever conclusions we draw or inferences we make may be very misleading.

2.3 Detecting Heteroscedasticity

2.3.1 Informal Methods

The problem of heteroscedasticity is likely to be more common in cross-sectional than in time series data. In cross-sectional data, one usually deals with members of a population at a given point in time, such as individual consumers and their families, firms, industries, or geographical subdivisions such as state, country, city, etc. Moreover, these members may be different sizes, such as small, medium, or large firms or low, medium or high income. In time series data, on the other hand, the variables tend to be of similar orders of magnitude, because one generally collects the data for the same entity over a period of time.

1. Nature of the problem

Very often the nature of the problem under consideration suggests whether heteroscedasticity is likely to be encountered. In cross-sectional data involving heterogeneous units, heteroscedasticity may be the rule rather than the exception. As a rule of thumb, the greater is the degree of heterogeneity in a sample, the more likely heteroscedasticity will be present.

2. Graphical Methods

If there is no a priori or empirical information about the nature of heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no heteroscedasticity and then do a postmortem examination of the residual squared e_i^2 to see if they exhibit any systematic pattern. Although e_i^2 are not the same thing as ε_i^2, they can be used as proxies especially if the sample size is sufficiently large. To carry out this informal method, simply scatter plot e_i^2 against $\hat{Y}_i$ or one or more of the explanatory variables, or both.

2.3.2. Formal Methods

1. White's Heteroskedasticity Test

This is a test for heteroskedasticity in the residuals from a least squares regression (White, 1980). Ordinary least squares estimates are consistent in the presence heteroskedasticity, but the conventional computed standard errors are no longer valid. If you find evidence of heteroskedasticity, you should either choose the robust standard errors option to correct the standard errors or you should model the heteroskedasticity to obtain more efficient estimates using weighted least squares.

White's test is a test of the null hypothesis of no heteroskedasticity against heteroskedasticity of some unknown general form. The test statistic is computed by an auxiliary regression, where we regress the squared residuals on all possible (non redundant) cross products of the regressors. For example, suppose we estimated the following regression:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i \qquad (2.3.1)$$

The test statistic is then based on the auxiliary regression:

$$e_i^2 = \alpha_0 + \alpha_1 X_i + \alpha_2 Z_i + \alpha_3 X_i^2 + \alpha_4 Z_i^2 + \alpha_5 X_i Z_i + v_i \qquad (2.3.2)$$

For example the software Eviews reports two test statistics from the test regression. The F-statistic is an omitted variable test for the joint significance of all cross products, excluding the constant. It is presented for comparison purposes. The $n.R^2$ statistic is White's test statistic, computed as the number of observations times the centered from the test regression. The exact finite sample distribution of the F-statistic under Ho is not known, but White's test statistic is asymptotically distributed as a χ^2 with degrees of freedom equal to the number of slope coefficients (excluding the constant) in the test regression. White also describes this approach as a general test for model

misspecification, since the null hypothesis underlying the test assumes that the errors are both homoskedastic and independent of the regressors, and that the linear specification of the model is correct. Failure of any one of these conditions could lead to a significant test statistic. Conversely, a non-significant test statistic implies that none of the three conditions is violated.

a. Features of White's test

- Does not require any prior knowledge about the source of heteroscedasticity.
- It is a large sample Lagrange Multiplier (LM) test
- Does not depend on the normality of population errors.

2. Park, Glesjer, and Breusch-Pagan-Godfrey Tests

All three of these tests are similar, so we will address them as a group. Like White's test, each of these tests are LM tests and thus follow the same general procedure. Given the following regression model, carry out the following steps:

Steps

i. Given the data, estimate the regression model and obtain the residuals.

ii. Next, estimate the following auxiliary regression models and obtain their R^2s.

Park Test

$$\log e_i^2 = \alpha_o + \alpha_1 \log Z_{1i} + \alpha_2 \log Z_{2i} + - - - + \alpha_p \log Z_{pi} + v_i \qquad (2.3.3)$$

Glesjer Test

$$\left|e_i^2\right| = \alpha_o + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + - - - + \alpha_p Z_{pi} + v_i \qquad (2.3.4)$$

Breusch-Pagan-Godfrey Test

$$\tilde{e}_i^2 = \alpha_o + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + - - - + \alpha_p Z_{pi} + v_i \ , \qquad \tilde{e}_i^2 = \frac{e_i^2}{\sum e_i^2 / n} \qquad (2.3.5)$$

where in each auxiliary regression, the *Zs* may be some or all of the *Xs*.

iii. Compute the LM test statistic: Under the null hypothesis that there is no heteroscedasticity, it can be shown that the sample size *n* times the R^2 obtained from the auxiliary regressions asymptotically follows the chi-square distribution with degrees of freedom equal to the number of regressors (not including the constant term) in the auxiliary regression. That is,

$$n.R^2 \sim \chi^2 \qquad (2.3.6)$$

It is important to note that the test statistics originally proposed by Park and Glesjer are Wald test statistics, and the test statistic originally suggested in the Breusch-Pagan-Godfrey Test is one-half of the auxiliary regression's explained sum of squares, distributed as chi-square with *p* degrees of freedom. However, as pointed out by Gujarati (1995), since all of these tests are simply large-sample tests, they are all operationally equivalent to the LM test.

iv. Perform the LM test by comparing $n.R^2$ to the chi-square critical value$\chi^2_{\alpha, p}$. If $n.R^2 > \chi^2_{\alpha, p}$, the conclusion is that there is heteroscedasticity. If $n.R^2 < \chi^2_{\alpha, p}$, there is no heteroscedasticity, which is to say that

$$\alpha_1 = \alpha_2 = - - - = \alpha_p = 0$$

b. Features of the Park, Glesjer, and Breusch-Pagan-Godfrey Tests

• They all require knowledge about the source of heteroscedasticity that is the Z variables are known to be responsible for the heteroscedasticity.

• They are all, in essence, large sample Lagrange Multiplier (LM) tests

• In the Park test, the error term in the auxiliary regression may not satisfy the CLRM assumptions and may be heteroscedastic itself. In the Glejser test, the error term is nonzero, is serially correlated, and is ironically heteroscedastic. In the Breusch-Pagan-Godfrey test, the error term is quite sensitive to the normality assumption in small samples.

3. Goldfeld-Quandt test

An alternative and popular test for heteroscedasticity works from the intuition of the problem: If population errors are homoscedastic, and thus share the same variance over all observations, then the variance of residuals from a part of the sample observations should be equal to the variance of residuals from another part of the sample observations. Thus, a "natural" approach to testing for heteroscedasticity would be to perform and *F*-test for the equality of residual variances, where the F-statistic is simply the ratio of two sample variances. Like the other tests we have discussed, this one too involves a number of steps. Consider the following regression model:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + - - - + \beta_k X_{ki} + \varepsilon_i \qquad (2.3.7)$$

Steps

i. Identify a variable to which the population error variance is related. For the sake of illustration, we will assume that X_1 is related to var (ε_i) positively.

ii. Order or rank the observations according to the values of X_1, beginning with the lowest X value.

iii. Omit c central observations, where c is specified a priori, and divide the remaining n - c observations into two groups each of $(n - c)/2$ observations.

The choice of c, for the most part, is arbitrary; however, as a rule of thumb will usually lie between one-sixth and one-third of the total observations.

iv. Run separate regressions on the first $(n - c)/2$ observations and the last $(n - c)/2$ observations, and obtain the respective residual sum of squares: ESS_1 representing the residual sum of squares from the regression corresponding to the smaller X_1 values (the small variance group) and ESS_2 that from the larger X_2 values (the large variance group).

v. Compute the F-statistic

$$F = \frac{ESS_2 / d.f}{ESS_1 / d.f} \qquad (2.3.8)$$

where

$$d.f. = \frac{n - c - 2(k+1)}{2}$$

and k is the number of estimated slope coefficients.

vi. Perform the F-test. If ε_i are normally distributed, and if the homoscedasticity assumption is valid, then it can be shown that F follows the F distribution with degrees of freedom in both the numerator and denominator. If $F > F_\alpha$, then we can reject the hypothesis of homoscedasticity, otherwise we cannot reject the hypothesis of homoscedasticity.

Features of the Goldfeld-Quandt test

• The success of this test depends importantly on the value of c and on identifying the correct X variable with which to order the observations.

• This test can not accommodate situations where the combination of several variables is the source of heteroscedasticity. In this case, because no single variable is the cause of the problem, the Goldfeld-Quandt test will likely conclude that no heteroscedasticity exists when in fact it does.

2.4. HCCM for the Linear Regression Model

This paper deals exclusively with the linear regression model of the form

$$y = X\beta + \varepsilon \qquad (2.4.1)$$

where y is an $n \times 1$ vector of observations on a dependent variable, X is taken as a non-stochastic $n \times k$ matrix of observations on independent variables (assumed to be of full rank) and ε is an $n \times 1$ vector of disturbances.

$E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \Omega$, a positive definite matrix.

The ordinary least squares estimator for this model is

$$\hat{\beta} = (X'X)^{-1} X'Y \qquad (2.4.2)$$

which is best linear unbiased and has covariance matrix:

$$Cov(\hat{\beta}) = (X'X)^{-1} X\Omega X (X'X)^{-1} \qquad (2.4.3)$$

Suppose we assume the errors are homoscedastic,

That is, $E(\varepsilon\varepsilon') = \sigma^2 I_n \qquad (2.4.4)$

Then equation (1.2.3) simplifies to:

$$Cov(\hat{\beta}) = \sigma^2 (X'X)^{-1} \qquad (2.4.5)$$

This can be conventionally estimated as .

$$Cov(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1} \quad , \quad \text{where } \hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/_{n-k} \text{ and } \varepsilon = \left(I_n - X(X'X)^{-1}X'\right)Y$$

(2.4.6)

Since X is non-stochastic and ε is normally distributed, inferences in finite samples can then be based on the t or F distribution. If e_i denotes the OLS residuals, $y_i - X_i\hat{\beta}$, where y_i denote the ith observation on the dependant variable and X_i denote the ith row of the X matrix, then the OLS covariance matrix estimate is given by:

$$OLSCM = \frac{\sum e_i^2}{n-k}(X'X)^{-1} \qquad (2.4.7)$$

where n is the sample size and k is the size of thc β vector.

This is a classical technique that requires one to assume that the error terms have a constant variance (homoscedasticity). In this case, an OLSCM estimator is appropriate for hypothesis testing and computing confidence intervals. This assumption is often not very plausible. If the regression disturbance is heteroscedastic, we have $E(\varepsilon_i^2) = \sigma_i^2$, for all i.

This implies that the variance of the disturbance may vary from observation to observation, and we want to know how this behavior of the variance affects the properties of the least square estimators of the regression coefficients. A heteroscedasticity-consistent covariance matrix estimator (hereafter, HCCM) which allows one to estimate (2.4.3) consistently under general condition is

$$HCO = (X'X)^{-1} X'\hat{\Omega}X (X'X)^{-1} \qquad (2.4.8)$$

where $\hat{\Omega} = diag(e_1^2, e_2^2, \ldots, e_n^2)$

The square roots of the elements on the principal diagonal of HCO are the estimated standard errors of the OLS coefficients. They are often referred to as heteroscedasticity-consistent standard errors (HCSEs). The usual t and F tests are now only valid asymptotically. As shown by White (1980) and others, HCO is a consistent estimator of $Cov(\hat{\beta})$ in the presence of heteroscedasticity of an unknown form.

A number of studies have sought to improve on the White estimator for OLS. The asymptotic properties of the estimator are unambiguous, but its usefulness in small samples is open to question. In other words, the White procedure has large sample validity. It may not work very well in finite samples. The possible problems stem from the general result that the squared OLS residuals tend to under estimate the squares of the true disturbances [that is why we use $1/(n-k)$ rather than $1/n$ in computing s^2].

The end result is that in small samples, at least as suggested by some Monte Carlo studies [e.g., Mackinnon and White (1985)], the White estimator is a bit too optimistic and its asymptotic t ratios are a little too large. Davidson and Mackinnon (1993) suggest a number of fixes, which include some evidence about corrections to e_i^2 that can improve finite sample performance. The simplest correction, is to replace e_i^2 by, $\frac{n}{n-k}e_i^2$. This makes degrees of freedom correction that inflates each residual by a factor of $\sqrt{\frac{n}{n-k}}$. In other words it is a correction that includes scaling up the end result by a factor $\frac{n}{n-k}$. This yields the modified estimator, which is known as HC1:

$$HC1 = \frac{n}{n-k}(X'X)^{-1}X'diag[e_i^2]X(X'X)^{-1} = \frac{n}{n-k}HCO \qquad (2.4.9)$$

This correction is similar to the one conventionally used to obtain unbiased estimator of σ^2. Recall that $\hat{\Omega}$ in equation (1.2.8) is based on the OLS residuals e, not the error ε. Even if the errors are homoscedastic, the residuals may not be. A better correction is to use the squared residual scaled by its true variance $e_i^2/(1-h_{ii})$, instead of e_i^2, where $h_{ii} = x_i(X'X)^{-1}x_i'$. The rational of this correction may be shown as follows. It is known that

$$\mathbf{e} = \mathbf{My}$$

where $M = I - \mathbf{X}(X'X)^{-1}X'$, which is a symmetric, idempotent matrix with the properties

$$MX = \mathbf{0} \quad \text{and} \quad \mathbf{M\varepsilon} = \mathbf{e}$$

By assuming homoscedasticity, the variance matrix of the OLS residual vector is

$$E(\mathbf{ee}') = E(M\varepsilon\varepsilon'M) = \sigma^2 M \qquad (2.4.10)$$

The i^{th} element on the principal diagonal of the matrices in equation (2.4.10) gives

$$E(e_i^2) = \sigma^2(1-x_i(X'X)^{-1}x_i') = Var(e_i) \neq \sigma^2 \qquad (2.4.11)$$

The mean squared residual thus underestimates σ^2, which suggests the second correction given in this paragraph. The term m_{ii} is the ith diagonal element of the matrix $X(X'X)^{-1}X'$. Equation (2.4.11) suggests that while e_i^2 is a biased estimator of σ_i^2, $e_i^2/(1-h_{ii})$ will be less biased. Thus, Horn and Duncan (1975) suggest using

$$\tilde{\sigma}_i^2 = e_i^2/(1-h_{ii}) \qquad (2.4.12)$$

as an almost unbiased estimate for σ_i^2. Following their approach, Mackinnon and White (1985) propose the estimator and refer to this estimator as HC2:

$$HC2 = \left(X'X\right)^{-1} X\tilde{\Omega}X\left(X'X\right)^{-1} \qquad (2.4.13)$$

where $\tilde{\Omega} = diag\left(\tilde{\sigma}_1^2, \tilde{\sigma}_2^2, \ldots, \tilde{\sigma}_n^2\right)$

It is immediate from (2.4.11) that HC2 will be unbiased when the e_i are in fact homoscedastic.

Mackinnon and white (1985) also suggest a third correction, $e_i^2 / (1-h_{ii})^2$ as an approximation to an estimator based on the "Jackknife" technique, but their advocacy of this estimator is much weaker than that of the other two. They refer to this covariance matrix estimator as HC3. To get this estimator, replace e_i^2 by $e_i^2 / (1-h_{ii})^2$ as

$$HC3 = \left(X'X\right)^{-1} X'diag\left[e_i^2 / (1-h_{ii})^2\right] X\left(X'X\right)^{-1} \qquad (2.4.14)$$

It is evident that HC3 is asymptotically equivalent to HCO, HC1, and HC2, since $1/n$ times the middle factor clearly converges to $1/n$ times X'ΩX.

Dividing e_i^2 by $(1-h_{ii})^2$ further inflates e_i^2, which is thought to adjust for the "the over influence" of observations with large variances.

2.5 Controlling for Heteroscedasticity and Estimation of $Cov(\hat{\beta})$: Remedial measures

Heteroscedasticity does not remove the unbiasedness and consistency properties of the OLS estimators, but they are no longer efficient, not even asymptotically (i.e., large sample size). This lack of efficiency makes the usual hypothesis testing procedure of

dubious value. Therefore, remedial measures may be called for. Actually there are no hard and fast rules for detecting heteroscedasticity, only a few rules of thumb. Now given that we have detected heteroscedasticity, what in the world can we do about it? In other words, is there a remedial measure, which will permit our regression models to comply with the CLRM assumptions and thus obtain BLUE estimates? Fortunately, there are two approaches to remediation: When σ_i^2 is known and when σ_i^2 is not known.

2.5.1. The Method of Generalized (Weighted) Least Squares: When σ_i^2 is Known.

Ideally, we would like to devise the estimating scheme in such a manner that observations coming from populations with greater variability are given less weight than those coming from populations with smaller variability. Unfortunately, the usual OLS method does not follow this strategy and therefore does not make use of the "information" contained in the unequal variability of the dependent variable: It simply assigns equal weight or importance to each observation.

But a method of estimation, known as generalized least squares (GLS), takes such information into account explicitly and is therefore capable of producing estimators that are BLUE. To see how this is accomplished, let us recall the general three-variable regression model:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \qquad (2.5.1)$$

Now let's assume that the heteroscedastic variance σ_i^2 is known. Dividing our regression model through by σ_i, we obtain

$$\frac{Y_i}{\sigma_i} = \beta_o \frac{1}{\sigma_i} + \beta_1 \frac{X_{1i}}{\sigma_i} + \beta_2 \frac{X_{2i}}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i} \qquad (2.5.2)$$

which for the sake of exposition we express as

$$Y_i^* = \beta^* X_{oi}^* + \beta_1^* X_{1i}^* + \beta_2^* X_{2i}^* + \varepsilon_i^* \qquad (2.5.3)$$

where the starred or transformed variables are the original variables divided by the known σ_i , and $X_{oi}^* = \frac{1}{\sigma_i}$. I have starred the regression coefficients to distinguish them from the usual OLS parameters. Now the reason why we transform the variables this way is because the transformed error terms, ε_i^* are homoscedastic! To see this, notice the following:

$$Var(\varepsilon_i^*) = E(\varepsilon_i^{*2}) = E\left(\frac{\varepsilon_i}{\sigma_i}\right)^2$$

$$= \frac{1}{\sigma_i^2} E(\varepsilon_i^2) \qquad \text{since } E(\varepsilon_i^2) = \sigma_i^2 \qquad (2.5.4)$$

$$= \frac{1}{\sigma_i^2} \sigma_i^2$$

$$= 1$$

this is constant! Since with the transformed variables ε_i^* satisfies the CLRM assumptions, if we apply OLS to the transformed model, it will produce estimators that are BLUE. In other words, $\hat{\beta}_j^*$ for $j = 0, 1, 2$ are BLUE, but $\hat{\beta}_j$ are not. This procedure of transforming the original variables in such a way that the transformed variables satisfy the assumptions of the classical model and then applying OLS to them is known as the method of Generalized Least Squares, and the estimators produced by this method are titled GLS estimators. This case of GLS is also called Weighted Least Squares (WLS) where the variable transformations are simply a "weighting" of the original variables by the factor $w_i = \frac{1}{\sigma_i}$. Notice that the transformed model no longer includes a constant term. Hence, under GLS, the R^2 cannot be interpreted. While in practice σ_i^2 is almost never

known, if it is known then it will take on a particular form, specifically it will likely be proportional to one or more of the explanatory variables. For example, let us suppose that σ_i is proportional to X_1, that is

$$Var(\varepsilon_i) = \sigma_i^2 = \sigma^2 X_{1i}^2 \quad \text{or equivalently} \quad \sigma_i = \sigma X_{1i} \tag{2.5.5}$$

then transforming the original model by multiplying each variable (including the constant) by the weight $w_i = \frac{1}{X_{1i}}$ yields

$$\frac{Y_i}{X_{1i}} = \beta_o \frac{1}{X_{1i}} + \beta_1 \frac{X_{1i}}{X_{1i}} + \beta_2 \frac{X_{2i}}{X_{1i}} + \frac{\varepsilon_i}{X_{1i}}$$

$$Y_i^* = \beta_o^* X_{oi}^* + \beta_1^* + \beta_2^* X_{2i}^* + \varepsilon_i^* \quad \text{(Includes a constant term),}$$

where

$$Var\ (\varepsilon_i^*) = E\ (\varepsilon_i^2)^2 = E\left(\sigma \frac{\varepsilon_i}{\sigma_i}\right)^2 \tag{2.5.6}$$
$$= \frac{\sigma^2}{\sigma_i^2} E\ (\varepsilon_i^2) = \sigma^2 \frac{\sigma_i^2}{\sigma_i^2} = \sigma^2$$

2.5.2. When σ_i^2 is not known

Well given that σ_i^2 is rarely ever known, is there a way of obtaining consistent estimates of the variances and covariances of OLS estimators if there is heteroscedasticity? The answer is yes and the remedy is titled White's HCSEs (also known as HCCM). White (1980) introduced a consistent estimator for the situation in which $\boldsymbol{\Omega}$ is diagonal but possibly heteroscedastic. This method uses the estimate

$$Cov(\hat{\beta}) = (X'X)^{-1} X'\hat{\Omega}X (X'X)^{-1} \tag{2.5.7}$$

Where

$$\hat{\Omega} = \begin{pmatrix} (y_1 - x'_1\hat{\beta})^2 & 0 & \cdots & 0 \\ 0 & (y_2 - x'_2\hat{\beta})^2 & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \\ 0 & 0 & \cdots & (y_n - x'_n\hat{\beta})^2 \end{pmatrix} \tag{2.5.8}$$

Notice that this estimate of **Ω** does not allow for correlation among observations. This estimate has several names, including the sandwich estimator, the consistent variance estimator, the robust estimator, and the White's estimator. In this thesis, it will be referred to as the White estimator.

The White estimator is biased, but it is consistent (Kauermann 2001). One of the benefits of this estimator is that it does not rely on a specific formal model of the heteroscedasticity for its consistency (White 1980). The White estimator is computed using ordinary least squares residuals, which tend to be too small (White 1985). Kauermann (2001) provides an option to reduce bias when using the White estimator; namely, substitution of

$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{(1-h_{ii})^{1/2}} \tag{2.5.9}$$

for

$$\hat{\varepsilon}_i = (y_i - x_i\hat{\beta})(y_i - x_i\hat{\beta})' \tag{2.5.10}$$

where m_{ii} is the ith diagonal element of the hat matrix

$$H = X (X'X)^{-1}X' \qquad (2.5.11)$$

A generalization of the White estimator is often used when $\boldsymbol{\Omega}$ is block-diagonal, with blocks ($\boldsymbol{\Omega}_i$) of equal size. One can use the least squares estimator (2.4.2) with

$$Cov(\hat{\beta}) = (X'X)^{-1} X' \hat{\Omega}_b X (X'X)^{-1} \qquad (2.5.12)$$

Where

$$\hat{\Omega}b = \begin{pmatrix} (y_1 - x_1\hat{\beta})(y_1 - x_1\hat{\beta})' & 0 & \dots & 0 \\ 0 & (y_2 - x_2\hat{\beta})(y_2 - x_2\hat{\beta})' & \dots & 0 \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & & \\ 0 & 0 & \dots & (y_n - x_n\hat{\beta})(y_n - x_n\hat{\beta})' \end{pmatrix} \qquad (2.5.13)$$

2.6 Review of Simulation Studies of the HCCM Estimator

Few simulation studies have been done to investigate the properties of the HCCM estimator (e.g. White 1980, Mackinnon and White 1985, Davidson and Mackinnon 1993, and Long et al. 2000).

2.6.1 White, 1980

White (1980) has presented general conditions under which a consistent estimator of the OLS parameter covariance matrix can be obtained, regardless of the presence of

heteroscedasticity in the disturbances of a properly specified linear model. Since this estimator does not require a formal modeling of the structure of the heteroscedasticity and since it requires only the regressors and the estimated least squares residuals for its computation, the estimator that is given by

$$\begin{aligned} HCO &= \left(X'X\right)^{-1} X'\hat{\Omega}X\left(X'X\right)^{-1} \\ &= \left(X'X\right)^{-1} X'diag\left(e_i^2\right)X\left(X'X\right)^{-1} \end{aligned} \tag{2.6.1}$$

should have wide applicability. Additional conditions are given which allow the investigator to test directly for the presence of heteroscedasticity. If found, elimination of the heteroscedasticity by a more careful modeling of the stochastic structure of the model can yield improved estimator efficiency. According to White (1980), one had either to model heteroscedasticity correctly or suffers the consequences.

The fact that the covariance matrix estimator and heteroscedasticity test given by White (1980) do not require formal modeling of the heteroscedastic structure is a great convenience, but it does not relieve the investigator of the burden of carefully specifying his/her models. Instead, we hoped that the statistics presented here would enable researchers to be even more careful in specifying and estimating economic models.

Thus, when a formal model for heteroscedasticity is available, application of the tools presented by White (1980) will allow one to check the validity of this model, and undertake further modeling if indicated. But even when heteroscedasticity cannot be completely eliminated, the heteroscedasticity covariance matrix of equation (2.4.7) allows correct inferences and confidence intervals to be obtained.

White (1980) introduced this idea to econometricians and derived the asymptotically justified form of the HCCM known as HCO. HCO is the most commonly used form of the HCCM and is referred to as the White's estimator. As shown by White (1980) and others, HCO is a consistent estimator of $Cov\left(\hat{\beta}\right)$in the presence of heteroscedasticity of an unknown form.

White has shown that this estimate can be performed so that asymptotically valid (i.e., large-sample) statistical inferences can be made about the true parameter values. Nowadays, several computer packages present White's heteroscedasticity corrected variances and standard errors along with the usual OLS variances and standard errors. Incidentally, White's heteroscedasticity corrected standard errors are also known as robust standard errors. White's estimator is consistent under both homoscedasticity and heteroscedasticity of unknown form, but it can be quite biased when the sample size is small.

2.6.2 White and MacKinnon, 1985

White and MacKinnon (1985) considered three alternative estimators designed to improve the small sample properties of HCO. They examined several modified versions of the heteroscedasticity- consistent covariance matrix estimator of White (1980) on the basis of sampling experiments which compare the performance of t statistics. They found that one estimator, based on the jackknife, performs better in small samples than the rest. In every single case, the standard deviation of the t-statistics based on HC1 exceeded that for HC2, which in turn exceeded that for HC3. Since there was certainly no tendency for HC3 to have too small variances, this implies that HC3 is the covariance matrix estimator of choice. The difference between HC1 and HC3 is often striking, and the difference between HCO and HC3 would, of course, be even more striking. The usual OLS covariance estimator can be very seriously misleading in the presence of heteroscedasticity. When it is, HC3 is also likely to be misleading if the sample size is small, but much less so than OLS. When there is no heteroscedasticity, all the HCCM estimators are less reliable than OLS, but HC3 does not seem to be much less reliable. They studied the properties of several alternative tests for heteroscedasticity. They found that they often lack power to detect damaging levels of it. This fact, together with other results, suggests that it may wise to use HC3 when there is little evidence of heteroscedasticity.

2.6.3 Davidson and MacKinnon, 1993

Davidson and MacKinnon 1993 raised the following two crucial questions:

1. What is the difference between the conventional and the correct estimates of the standard errors of the OLS coefficients?
2. What is the difference between the correct OLS standard errors and the GLS standard errors?

They provide some Monte Carlo evidence on these questions. They considered the following simple model,

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i, \qquad \textbf{(2.6.2)}$$

They assumed that $\beta_1=1$, $\beta_2=1$ and $\varepsilon_i \sim N(0, X_i^{\alpha})$, with n=100, x_i uniformly distributed between 0 and 1, and α is a parameter that takes on various values. As the last expression shows, they assumed that the error variance is heteroscedastic and is related to the value of the regressor X with power α.

If, for example, $\alpha=1$, the error variance is proportional to the value of X; if $\alpha=2$, the error variance is proportional to the square of the value of X, and so on. The inefficiency increases substantially with α. Based on 20,000 replications and allowing for various values for α, they obtain the standard errors of the two regression coefficients using OLS, OLS allowing for heteroscedasticity, and GLS.

The most striking feature of this result is that OLS, with or without corrections for heteroscedasticity, consistently overestimates the true standard error obtained by the (correct) GLS procedure, especially for large values of α, thus establishing the superiority of GLS. These results also show that if we do not use GLS and rely on OLS-allowing for or not allowing for heteroscedasticity – the picture is mixed. The usual OLS

standard errors are either too large (for the intercept) or generally too small (for the slope coefficient) in relation to those obtained by OLS allowing for heteroscedasticity.

The message is clear: in the presence of heteroscedasticity, use GLS. However, in practice it is not always easy to apply GLS. The White procedure has large-sample validity. It may not work very well in finite samples.

2.6.4 Long et al., 2000

White and MacKinnon (1985) proposed three alternatives to reduce the bias of the Standard HCCM Estimator (2.5.7). They did this after noticing that the original HCCM estimator introduced by White (1980) did not take into account that ordinary least squares residuals tended to be too small. The Standard HCCM Estimator and three alternatives introduced to reduce bias are given by HC1, HC2 and HC3. Each of these three alternatives is designed to reduce the bias of $\tilde{\Omega}_b$(2.5.12). The properties of these three alternatives are described by White and MacKinnon (1985) and applied to data from various experiments. They ran a simulation study to compare the HCCM estimator HCO to the three alternatives HC1, HC2 and HC3. They also tested the ordinary least squares estimator with the usual estimate for the covariance matrix OLSCM. They tested each estimator under varying degrees of heteroscedasticity. They also varied the sample sizes from 25, 50, 100, 250, 500 and 1000. They generated 100,000 observations for each independent variable. They next generated random errors with the desired error structure. Then, with these errors, they computed the dependent variables. Random samples of the proper size were drawn without replacement. Each of these combinations were run using the Standard HCCM Estimator HCO, the three alternatives to the Standard HCCM Estimator HC1, HC2 and HC3, and the ordinary least squares estimator OLSCM. Since the ordinary least squares estimator is designed for homoscedastic data, it had the best properties for homoscedastic data. The properties for the HC3, were nearly the same as the properties of the ordinary least squares estimator, even at the smallest sample size (n = 25). The tests for the other three estimators had varying degrees of distortion for sample sizes $n \leq 100$. All of the tests had nearly the same properties for $n \geq 250$. Under milder

forms of heteroscedasticity, the ordinary least squares estimator worked well for all sample sizes. With more extreme forms of heteroscedasticity, the ordinary least squares estimator performed increasingly worse as the sample size increased. When $n \geq 500$, the four different estimators performed similarly. Long et al. (2000) concluded that the estimator HC3 worked better than all of the other estimators under both heteroscedasticity and homoscedasticity. Long et al. (2000) proposed that the estimator HC3 should always be used.

CHAPTER THREE

3. METHODOLOGY AND DATA GENERATION

This chapter consists of three different sections. The first section is about Monte Carlo experiments; the second section is data structures, the last section highlights the procedure for data generation.

3.1 Monte Carlo Experiments

There are many things that faster computers have made possible in recent years. For scientists, engineers, statisticians, managers, investors, and others, computers have made it possible to create **models that simulate reality** and aid in **making predictions**. One of the methods for simulating real systems is the ability to take in to account randomness by investigating hundreds of thousands of different scenarios. The results are then compiled and used to make decisions. This is what Monte Carlo simulation is all about.

In this study, Monte Carlo simulation is used to examine the behavior of tests using the ordinary least square covariance matrix (OLSCM) and the four versions of the HCCM presented above. Our experiments simulate a variety of data and error structures that are likely to be encountered in cross-sectional research. To this end, we considered errors that were normal. The independent variables were constructed with a variety of distributions, including uniform, chi-square, normal, binomial, and binary. Finally, a variety of different forms and degrees of heteroscedasticity were considered.

Each simulation involved the following steps:

1. Independent variables: 10,000 observations for two independent variables were constructed and used for each experiment. The independent variables were constructed to include a few distributions.

2. Errors: Four error structures were chosen to represent common types of homoscedasticity and heteroscedasticity. 10,000 observations were generated for each error type.
3. Dependent variables: The dependent variable was constructed as a linear combination of two independent variables plus the error term. The combination of the independent variables, the error, and the dependent variable made up the population for each structure.
4. Simulations: From each population, a random sample without replacement was drawn. Since a different random sample is used for each replication, the design matrix will vary. Regressions were estimated and tests of hypotheses were computed for each sample. This was done 10,000 times each for sample sizes of 10, 25, 50, 100, 250,500, 600 and 1,000.
5. Summary: The results were summarized across the 10,000 replications for each sample size from each population.

Details of our simulations are now given as follows:

In all of our experiments, we utilized the following model:

$$Y_i = 1 + 1X_{2i} + 1X_{3i} + \varepsilon_i \qquad i = 1, 2 \ldots n \qquad (3.1.1)$$

Where characteristics of the x's and ε's varied to simulate data typically found in cross sectional research. The independent variables have a variety of distributions, including uniform, chi-square, etc. suppose we consider a sample of size N and it is given ($X_1, X_2, \ldots, X_N$) which remain the same in all replications. We also fixed the values of the parameters (β, σ^2). That is, $\beta_1=\beta_2=\beta_3=1.0$ and $\sigma^2=1.0$. Therefore, $\varepsilon_i \sim N(0, \sigma_i^2)$. The null hypothesis under test is Ho: $\beta_1=\beta_2=\beta_3=1.0$ and the study was conducted under the null hypothesis and using normal errors.

Heteroscedasticity is introduced by allowing the variance of the errors to depend on the independent variables in three ways corresponding to structures 1 to 3 as shown by table 3.1. This resulted in 3-heteroscedastic error structures that represent few degrees and

types of heteroscedasticity that might be found in practice. There were **one hundred sixty** sets of experiments, in each of which the ε_i were chosen differently.

Table 3.1: Error Structures Used in the Simulations.

Structure	Heteroscedasticity function
0	$\varepsilon_i = \varepsilon_i^*$
1	$\varepsilon_i = X_{2i}\varepsilon_i^*$
2	$\varepsilon_i = (X_{3i} + 1.5)\varepsilon_i^*$
3	$\varepsilon_i = X_{2i}\,(X_{3i} + 2.5)\varepsilon_i^*$

Note: ε^* has a z distribution

It is more interesting to ask what proportion of the time each of the test statistics exceeds certain critical values. In other words, we compared the nominal significance level to the proportion of times that the correct Ho was rejected over the 10,000 replications at a given sample size. The critical values we chose were the 5% and 1% levels; absolute critical values for the standard normal at these levels are 1.96 and 2.576, respectively. Since the findings were similar only results for the 5% level are presented.

The obvious way to estimate these rejection frequencies is to use this estimator

$\hat{q} = \frac{R}{N}$, where R is the observed number of rejections and N is the number of replications (here 10,000). A consistent estimate of the variance of this estimator is

$$\hat{q}\,(1 - \hat{q})/N \qquad (3.1.2)$$

Davidson and MacKinnon (1981) have proposed this technique for doing so, McKinnon and White (1985) used it based on jackknife resampling technique, which we utilize here too. For details, see Davidson and MacKinnon (1981).

3.2 Data Structures

The first step was to generate observations for two independent random variables with the following distributions. With a sample size N, Δ_2 is uniformly distributed between 0 and 1 and Δ_3 is from χ^2 distribution with one degree of freedom or from standard normal distribution.

Long and Ervin (2000), used independent variables from different distributions directly from Binomial, Binary, Uniform, etc. But here we used some linear combinations of variables from uniform distributions and standard normal distributions to confirm the consistency of the HCCM estimators.

We have also included standardized variables as a regressor to see the effect of using such kind of variables as independent variables.

That is,

$\Delta_2 \sim$ Uniform (0, 1)

$\Delta_3 \sim \chi^2$ 1df or $\Delta_3 \sim N(0, 1)$ (3.2.1)

The Δ's were combined to construct the two independent variables:

$X_2 = 1+\Delta_2$

$X_3 = 2\Delta_2 + 0.3\Delta_3$ (3.2.2)

3.3 Data Generation

3.3.1 Simulation: It can be used as a means for extension of data collection from empirical studies. A simulation model is developed, based on the data from experiments, and new data is generated from the simulation model. It is argued that the data collection

process is the most crucial stage in the model building process. This is primarily due to the influence that data has in providing accurate simulation results. Data collection is an extremely time consuming process predominantly because the task is manually orientated. Hence, automating this process of data collection would be extremely advantageous. We, therefore, in this study have used simulation as a means of data collection from empirical results. For each error structure in table 3.1 and combination of types of variables, we ran simulations as follows: 10,000 observations for the independent variables (x's) were constructed. Random errors ε were generated according to the error structure being evaluated. These were used to construct the dependant variable y according to equation (3.1.1).

Each experiment involved 10,000 replications, and there were one hundred sixty experiments in all (for each of n=10, 25, 50, 100, 250, 500, 600 and 1,000). All results are based on 10,000 replications. For each of the β_i, we calculated five test statistics of the hypothesis that β_i equals its value. These statistics, denoted by OLSCM, HCO, HC1, HC2, and HC3, utilize the covariance matrices after which they are named. For each experiment, regressions were estimated and hypothesis tests were computed for each sample at each sample size. The estimates of the β's, standard error of coefficients, t-statistics and probabilities using the OLSCM and the four HCCMs were saved. These simulations were used to evaluate two situations in which the HCCM might be used. First, we examine the consequences of using a HCCM based test when errors are homoscedastic; second, we compare results using OLSCM tests and the HCCM tests when there is heteroscedasticity. This will be seen in the fourth chapter.

3.3.2 Code for Simulation: Beside to this we created a program by writing any sequence of Matlab 7.0 commands in a text.

Note: In the **Appendix A** we provide Monte Carlo simulation code in Matlab 7.0 that implements the OLSCM and the four HCCM methods in the linear regression model.

CHAPTER FOUR

4. ANALYSIS AND PRESENTATION OF RESULTS

This chapter can be divided into three sections. The first section deals about results of experiments, the second section gives homoscedastic errors and the last section reveals heteroscedastic errors.

4.1 Results of Experiments

Calculating the empirical size of t-tests for the β parameters assesses each method of computing the covariance matrix. For size, we compare the nominal significance level to the proportion of times that the correct H_0 is rejected over the 10,000 replications at a given sample size. The true hypothesis is H_0: $\beta_k = \beta_k^*$, where β_k*is the population value determined from the regression based on the entire N observations. We test the null hypothesis Ho with the t-statistic, which is given by

$$t = \frac{\hat{\beta} - 1}{\hat{S}(\hat{\beta})}$$

and the squared t-statistic, which is asymptotically χ^2 (1):

$\tau = t^2 = \left(\frac{\hat{\beta} - 1}{\hat{s}(\hat{\beta})} \right)^2$, where $\hat{\beta}$ is the ordinary least square estimate and $(\hat{s}(\hat{\beta}))^2$ the second diagonal element of the covariance matrix estimator. The realizations of the statistics τ are used to calculate a P-value at the nominal level α=0.05. Many statistical procedures work best when applied to variables that follow normal distribution. The t tests assume that variables follow a normal distribution.

This assumption may not be critical here in our case since we have checked the normality test before we use the t test. For the sake of convenience we presented one result of normality test in the **APPENDIX B** when N=50. In this case the P.value at the 5% level of significance is 0.932280. This tells us the given variable is from normal distribution. The safer alternative is employing a nonparametric test that does not assume normality which is χ^2 test. We have checked this also and our results still consistent.

The number of replications is R=10,000. For β_1, β_2, and β_3, we assess size by testing the true hypothesis: H_0: $\beta_k = \beta_k^*$. For the tables and figures below, we consider H_0: $\beta_k = \beta_k^*$. Size curves for each coefficient were also computed and are summarized where appropriate. While size was examined at the .05 and .10 nominal levels, the findings were similar so only results for the .05 level are presented here. The tables and figures speak by them selves, but we will discuss a few points of interest.

4.2 Homoscedastic Errors

In deriving ordinary least squares (OLS) estimates, we made the assumptions that the residuals ε_i were identically distributed with mean zero and equal variance σ^2, (i.e. $Var(\varepsilon_i) = E(\varepsilon_i^{\ 2}) = \sigma^2$). This assumption of equal variance as we mentioned earlier is known as homosedasticity (which means equal scatter). The variance σ^2 is a measure of dispersion of the residuals ε_i around their mean zero.

Equivalently, it is a measure of dispersion of the observed value of the dependant variable (Y) around the regression line $\beta_1 + \beta_2 X_2 + \ldots \beta_k X_k$. Homoscedasticity means that the dispersion is the same across all observations. In the first set the experiment ε_i were NID (0, 1), so that the OLS *t* statistic is appropriate.

The object here is to see how costly it is to use the various heteroskedasticity-consistent estimators when there is in fact no heteroskedasticity. Table 4.2 presents the proportion of times that the null hypotheses is rejected using tests based on the standard OLSCM and each type of HCCM when there is no heteroscedasticity over a range of sample sizes.

From table 4.2, it is clear that using HCO or HC1 when there is in fact no heteroscedasticity and the sample size is small could easily lead to serious errors of inference. Their worst performance was for β_3 when n=10. Here in this case OLSCM did always perform well when the sample size was small and there was no substantial heteroscedasticity.

The usual OLSCM t statistic would reject the null hypothesis 5% of the time at the nominal 5% level, HCO would incorrectly reject the null hypothesis 26.4% of the time, HC1 would reject the null hypothesis 20.1% of the time, HC2 would reject it 9.5% of the time and HC3 would reject it 4% of the time. Hence using HC3 is almost as reliable as using OLSCM.

Table 4.2: The proportion of times that the null hypothesis is rejected using tests based on the standard OLSCM and each type of HCCM when there is no heteroscedasticity.

Coeff.	Estimator	10	25	50	100	250	500	600	1000
β_1	OLSCM	0.048	0.053	0.049	0.052	0.052	0.050	0.050	0.052
	HCO	0.119	0.096	0.069	0.060	0.052	0.056	0.055	0.053
	HC1	0.076	0.077	0.062	0.057	0.051	0.053	0.052	0.048
	HC2	0.072	0.063	0.057	0.053	0.051	0.051	0.050	0.053
	HC3	0.030	0.047	0.049	0.048	0.050	0.049	0.047	0.049
β_2	OLSCM	0.047	0.051	0.047	0.051	0.052	0.051	0.054	0.051
	HCO	0.151	0.094	0.073	0.061	0.052	0.051	0.046	0.049
	HC1	0.101	0.078	0.064	0.057	0.051	0.049	0.051	0.051
	HC2	0.080	0.066	0.056	0.050	0.054	0.052	0.050	0.052
	HC3	0.034	0.047	0.047	0.047	0.053	0.051	0.052	0.050
β_3	OLSCM	0.050	0.051	0.048	0.048	0.050	0.050	0.051	0.050
	HCO	0.264	0.084	0.074	0.062	0.053	0.052	0.048	0.052
	HC1	0.201	0.067	0.067	0.058	0.051	0.052	0.048	0.049
	HC2	0.095	0.066	0.055	0.049	0.054	0.051	0.052	0.052
	HC3	0.040	0.046	0.045	0.044	0.052	0.050	0.050	0.048

Figure 4.2 uses results from a population with homoscedastic normal errors to illustrate our findings. The horizontal axis indicates the size of the sample used in the simulation; the vertical axis indicates the proportion of times that Ho was rejected out of 10,000 replications.

Figure 4.2a: Plots of B1

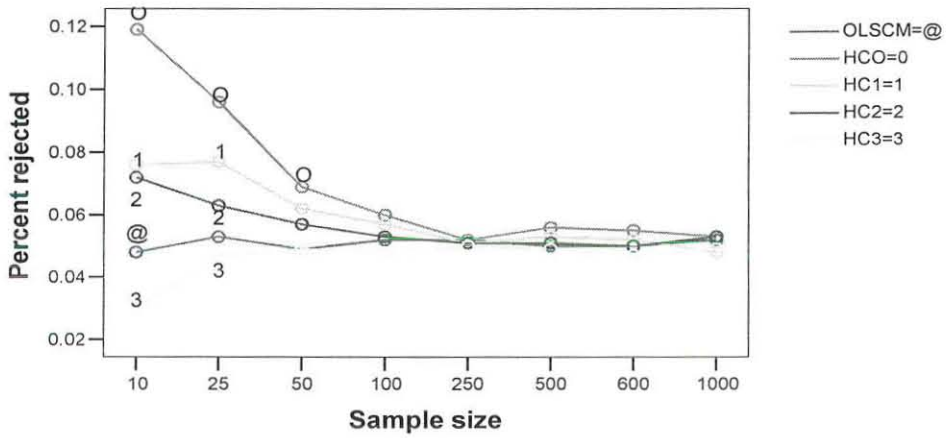


Figure 4.2b: Plots of B2

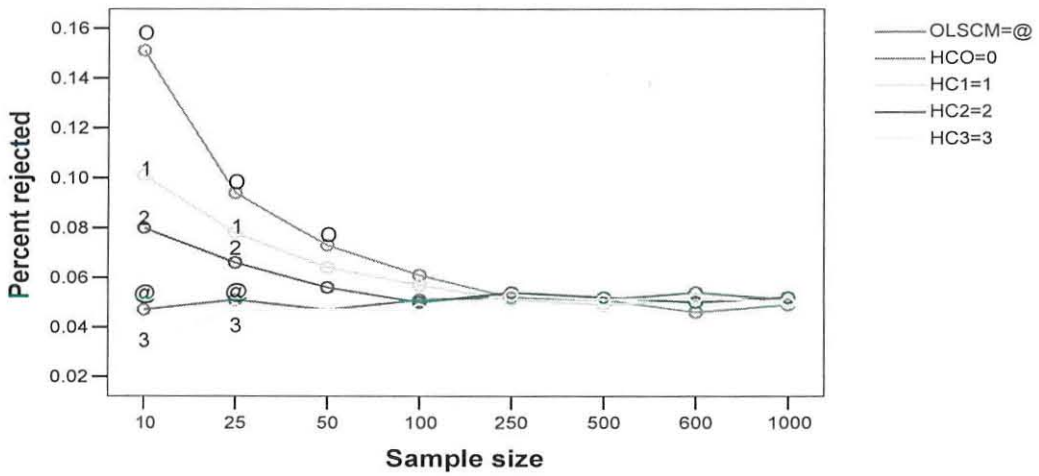


Figure 4.2c: Plot of B3

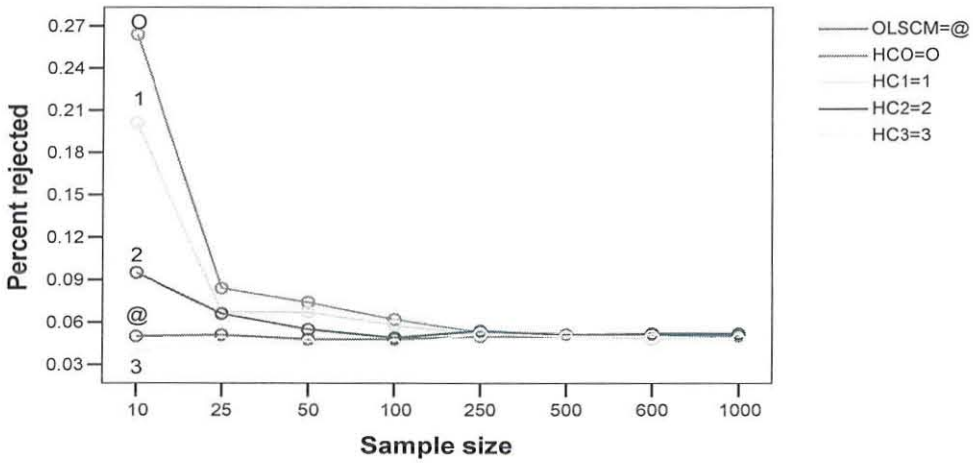


Figure 4.2: Plots the size distortion that is estimated null rejection frequencies at the nominal 5% level, against different sample sizes.

The key findings in these figures are:

1. Since the errors are homoscedastic, OLSCM tests have the best size properties.
2. For N<50, the size properties of HC3 tests are nearly as good as those for OLSCM tests, while tests based on HCO and HC1 have large size distortion.
3. When N≥250, there is very little distortion introduced by using any of the HCCM- based tests as well as OLSCM based test when there is no heteroscedasticity.

4.3 Heteroscedastic Errors

Heteroscedasticity is a violation of our assumptions about the error term, which has adverse implications for least squares estimation. But we do not know the errors, but proxy them with the residuals. The residuals are a function of our model specifications. Heteroscedasticity of the error term implies that we cannot use the estimates of the standard errors of the regression coefficients computed on the basis of the standard formulae derived from the classical regression model.

Furthermore, the least squares estimators are inefficient, but unbiased. But this prediction is not the most reliable (since it is in efficient) among all linear predictions and we cannot make any statement about the uncertainty (confidence interval, hypothesis tests) of the predictions based on the standard errors computed according to the formulae under the assumption of homoscedasticity.

Why not modify the formulae accordingly to allow for heteroscedastic errors? In principle, this is possible. Nevertheless, to do so we need to be able to state the nature of the heteroscedasticity of the error term. This is exactly what we try to do. Let us now consider each error structures used in the simulations.

4.3.1 Heteroscedasticity function: $\varepsilon_i = x_{2i}\varepsilon_i^*$

This expression shows that the error is heteroscedastic and is related to the value of the regressor x_{2i}. In this case, the error variance is proportional to the square of the value of x_{2i}. This error structure has milder forms of heteroscedasticity. Based on 10,000 replications, we obtain the rejection frequencies of the three regression coefficients using OLSCM and the four HCCM estimators. We quote these results by table 4.3.1

Table 4.3.1: The rejection frequencies of the three regression coefficients using OLSCM and the four HCCM estimators.

Coeff.	Estimator	10	25	50	100	250	500	600	1000
β_1	OLSCM	0.106	0.056	0.030	0.030	0.037	0.042	0.040	0.035
	HCO	0.119	0.087	0.070	0.056	0.057	0.058	0.054	0.053
	HC1	0.070	0.067	0.062	0.051	0.055	0.054	0.054	0.052
	HC2	0.057	0.056	0.052	0.052	0.053	0.054	0.052	0.051
	HC3	0.024	0.042	0.042	0.046	0.051	0.049	0.049	0.050
β_2	OLSCM	0.117	0.055	0.029	0.030	0.051	0.053	0.049	0.042
	HCO	0.111	0.086	0.074	0.055	0.055	0.054	0.054	0.053
	HC1	0.067	0.069	0.067	0.051	0.053	0.055	0.053	0.052
	HC2	0.074	0.059	0.050	0.054	0.052	0.052	0.053	0.052
	HC3	0.032	0.042	0.041	0.048	0.052	0.050	0.051	0.049
β_3	OLSCM	0.138	0.043	0.036	0.027	0.080	0.081	0.078	0.066
	HCO	0.097	0.077	0.087	0.058	0.055	0.056	0.056	0.059
	HC1	0.051	0.061	0.079	0.055	0.053	0.054	0.053	0.055
	HC2	0.120	0.072	0.053	0.056	0.053	0.053	0.051	0.052
	HC3	0.052	0.051	0.043	0.051	0.049	0.051	0.052	0.050

Figure 4.3.1 Plots the size properties of each test when the errors have a normal distribution for heteroscedasticity function that is given above.

Figure 4.3.1a: Plot of B1

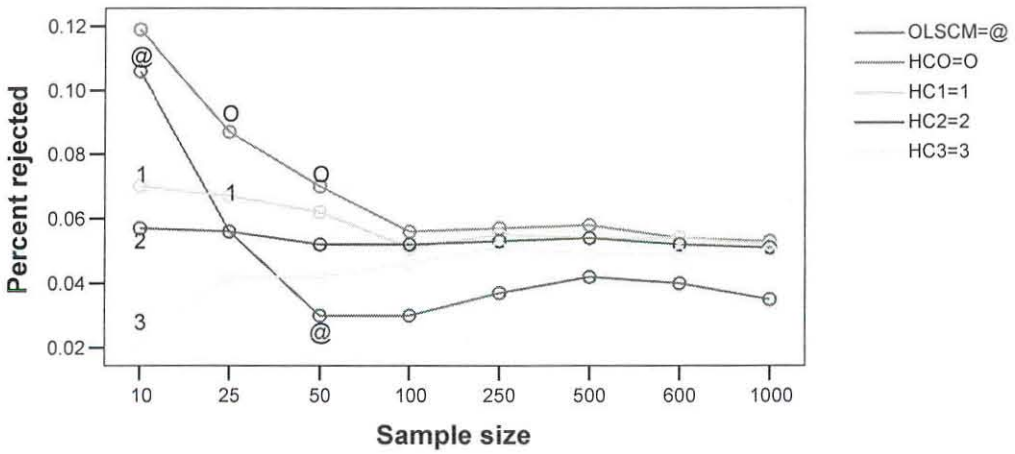


Figure 4.3.1b: Plot of B2

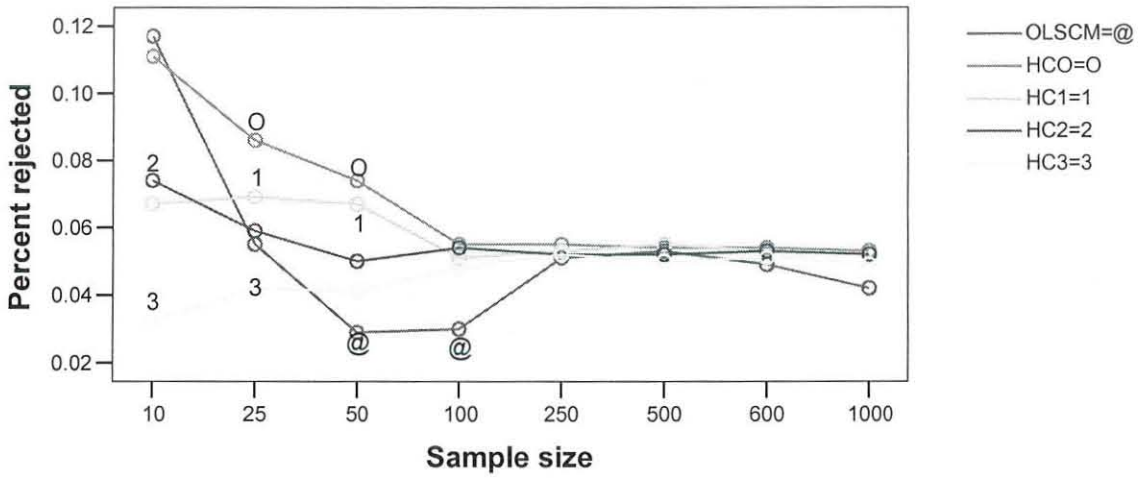


Figure 4.3.1c: Plot of B3

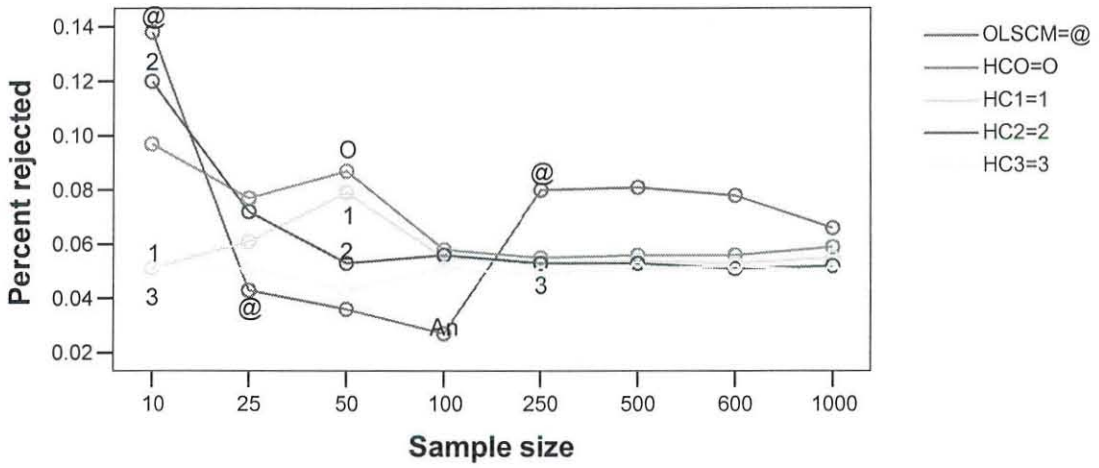


Figure 4.3.1: Plots of size of t- test of β_1, β_2 and β_3 for normal errors with heteroscedasticity function: $\varepsilon_i = x_{2i}\varepsilon_i^*$

The key findings in these figures are:

1. For N≤50, tests based on all tests seem to have large size distortion. But HC2 seems relatively good.
2. Tests based on OLSCM do have large size distortion. This size distortion does not vanish even for large sample size.
3. When N≥100, there is very little distortion introduced by using any of the HCCM-based tests when there is heteroscedasticity

4.3.2 Heteroscedasticity function: $\varepsilon_i = (x_{3i} + 1.5)\varepsilon_i^*$

The assumption about the nature of heteroscedasticity in this case is: $\varepsilon_i = (x_{3i} + 1.5)\varepsilon_i^*$. The specification that we make is that the error is related to the explanatory variable and some constant. This error structure has a moderate amount of heteroscedasticity. Table 4.3.2 presents the empirical level of the t statistics based on the standard OLSCM and the four HCCM estimators. We quote these results as shown below.

Table 4.3.2: The empirical level of the t statistics based on the standard OLSCM and the four HCCM estimators.

Coeff.	Estimator	10	25	50	100	250	500	600	1000
β_1	OLSCM	0.021	0.054	0.037	0.048	0.047	0.040	0.041	0.047
	HCO	0.132	0.097	0.064	0.058	0.054	0.054	0.053	0.049
	HC1	0.084	0.080	0.057	0.055	0.052	0.048	0.048	0.052
	HC2	0.079	0.059	0.057	0.051	0.051	0.049	0.051	0.053
	HC3	0.037	0.043	0.048	0.047	0.048	0.049	0.049	0.051
β_2	OLSCM	0.018	0.054	0.039	0.052	0.050	0.043	0.043	0.050
	HCO	0.139	0.094	0.065	0.061	0.056	0.054	0.054	0.053
	HC1	0.087	0.076	0.057	0.058	0.054	0.050	0.048	0.052
	HC2	0.079	0.057	0.057	0.051	0.052	0.053	0.052	0.049
	HC3	0.038	0.041	0.048	0.046	0.049	0.052	0.050	0.051
β_3	OLSCM	0.014	0.057	0.045	0.060	0.062	0.051	0.049	0.051
	HCO	0.143	0.088	0.068	0.060	0.058	0.056	0.052	0.053
	HC1	0.092	0.069	0.061	0.057	0.057	0.054	0.054	0.053
	HC2	0.089	0.063	0.056	0.055	0.049	0.052	0.052	0.050
	HC3	0.040	0.044	0.047	0.049	0.047	0.049	0.049	0.051

Figure 4.3.2 plots the size properties of each test when the errors have a normal distribution with heteroscedasticity function that is given above.

Figure 4.3.2a: Plot of B1

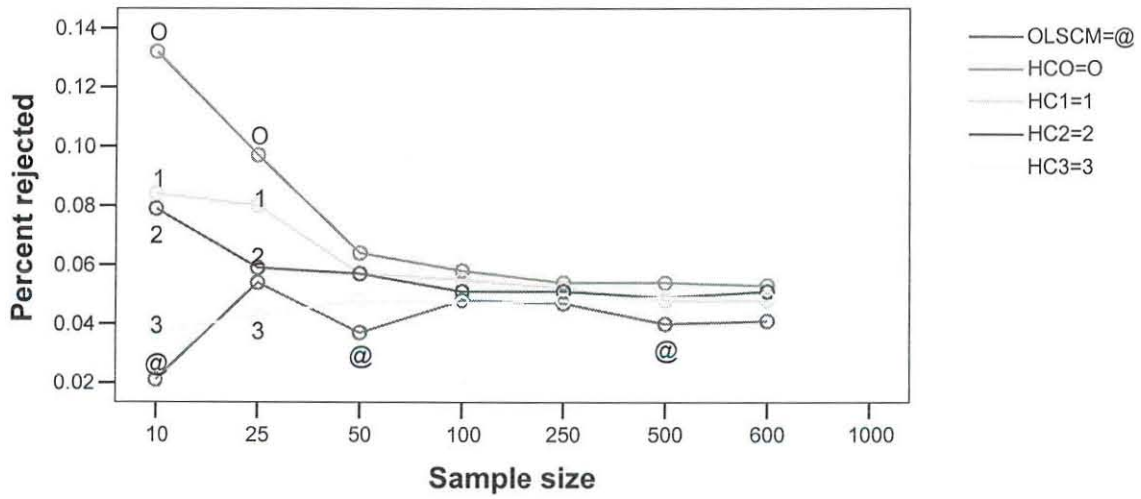


Figure 4.3.2b: Plot of B2

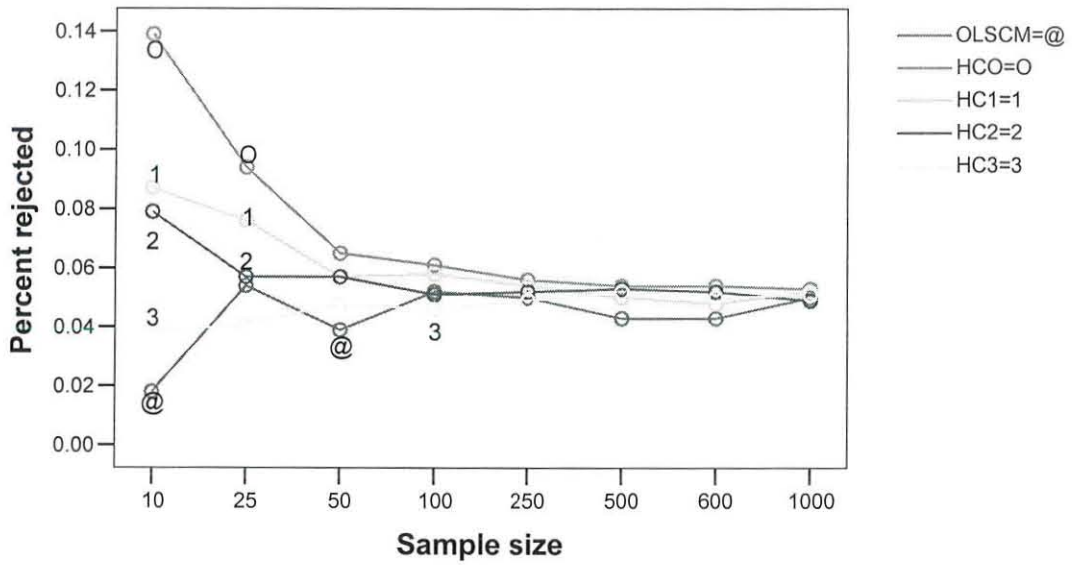


Figure 4.3.2c: Plots of B3

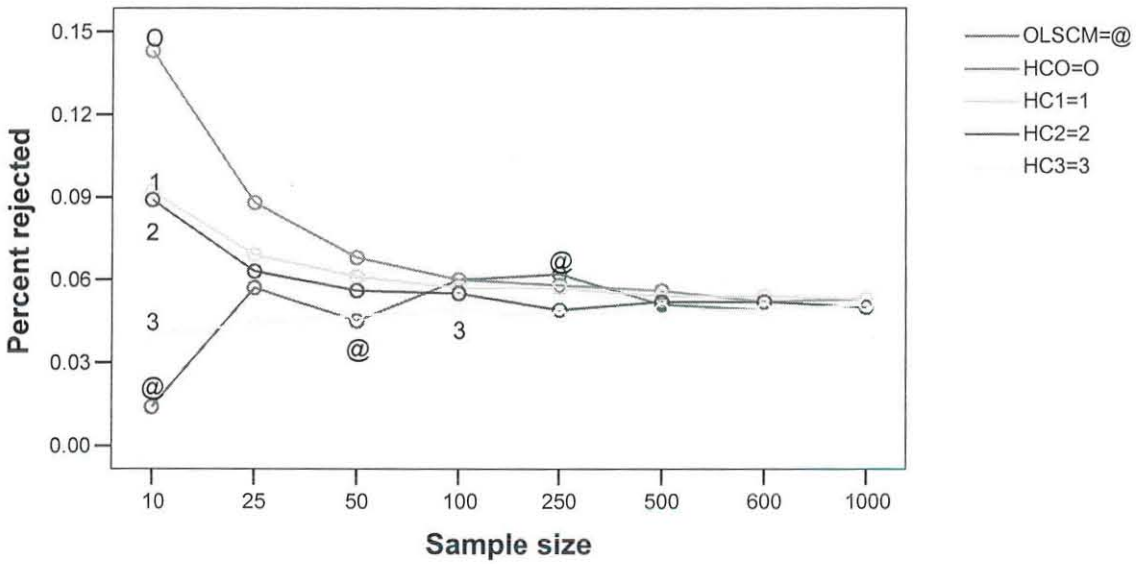


Figure 4.3.2: Plots of size of t- test of β_1, β_2 and β_3 for normal errors with heteroscedasticity function: $\varepsilon_i = (x_{3i} + 1.5)\varepsilon_i^*$

The key findings in these figures are:

1. For N=10, all tests do have size distortions. But HC3 seems relatively better.
2. For N=50, tests based on HC1, HC2 and HC3 perform well while those based on OLSCM and HCO indicate relatively large size distortion.
3. For N≤600, tests based on OLSCM do not have a clear pattern about size distortion.
4. For N=1000, rejection frequencies based on all tests converge to the nominal 5% level of significance.

4.3.3 Heteroscedasticity function: $\varepsilon_i = x_{2i}(x_{3i} + 2.5)\varepsilon_i^*$

This is more extreme forms of heteroscedasticity. Table 4.3.3 presents the proportion of times that the null hypothesis is rejected using tests based on the standard OLSCM and each type of HCCM.

Table 4.3.3: The proportion of times that the null hypothesis is rejected using tests based on the standard OLSCM and each type of HCCM.

Coeff.	Estimator	10	25	50	100	250	500	600	1000
β_1	OLSCM	0.032	0.027	0.054	0.053	0.041	0.039	0.041	0.044
	HCO	0.149	0.077	0.068	0.061	0.052	0.047	0.048	0.046
	HC1	0.096	0.058	0.060	0.057	0.051	0.048	0.052	0.053
	HC2	0.075	0.050	0.056	0.047	0.050	0.053	0.052	0.051
	HC3	0.039	0.035	0.048	0.043	0.048	0.049	0.049	0.050
β_2	OLSCM	0.029	0.035	0.059	0.060	0.048	0.047	0.048	0.050
	HCO	0.40	0.086	0.070	0.062	0.053	0.049	0.046	0.052
	HC1	0.091	0.068	0.063	0.059	0.052	0.053	0.052	0.053
	HC2	0.072	0.050	0.057	0.049	0.049	0.052	0.052	0.051
	HC3	0.035	0.035	0.048	0.043	0.047	0.049	0.048	0.049
β_3	OLSCM	0.011	0.046	0.055	0.061	0.051	0.055	0.054	0.052
	HCO	0.118	0.100	0.067	0.071	0.055	0.054	0.055	0.052
	HC1	0.073	0.083	0.059	0.068	0.053	0.052	0.053	0.051
	HC2	0.048	0.060	0.052	0.049	0.049	0.051	0.051	0.049
	HC3	0.018	0.044	0.043	0.043	0.048	0.050	0.049	0.051

When there was an extreme form of heteroscedasticity and the sample size was small, even HC3 did not always perform well. As it can be seen from table 4.3.3 for β_3 when n=10; tests based on the OLSCM estimator would reject the null hypothesis 1.1% of the time at the nominal 5% level, HCO would reject the null hypothesis 11.8% of the time, HC1 would reject it 7.3% of the time, HC2 would reject the null 4.8% of the time where as HC3 would reject it 1.8% of the time

Figure 4.3.3 plots the size properties of each test statistic when the errors have a normal distribution with the heteroscedasticity function: $\varepsilon_i = x_{2i}(x_{3i} + 2.5)\varepsilon_i^*$. This error structure (error structure 3 in table 3.1) has extreme forms of heteroscedasticity.

Figure 4.3.3a: Plot of B1

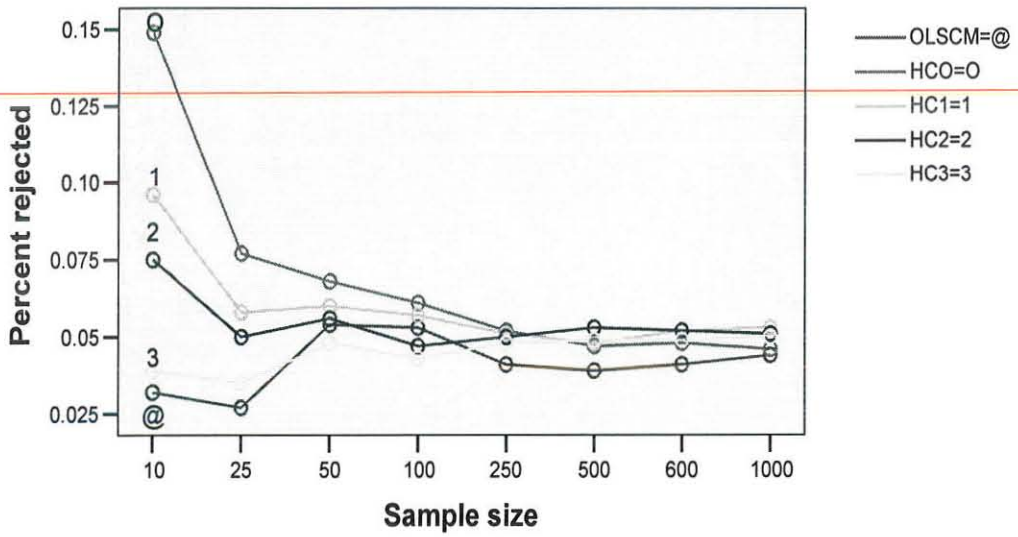


Figure 4.3.3b: Plot of B2

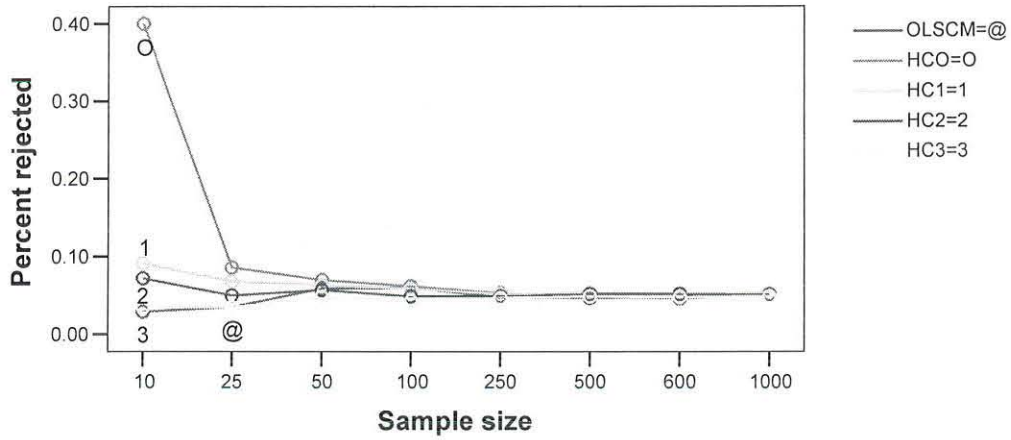


Figure 4.3.3c: Plot of B3

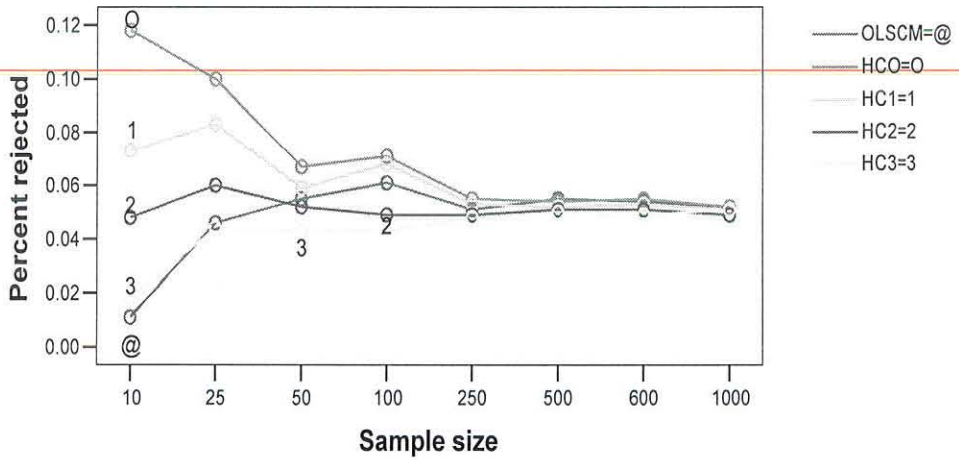


Figure 4.3.3: Plots of size of t- test of β_1, β_2 and β_3 for normal errors with heteroscedasticity function: $\varepsilon_i = x_{2i}(x_{3i} + 2.5)\varepsilon_i^*$

The findings in these figures are for the given error structure. The key findings are:

1. For N=10, tests based on all estimators show a large size distortion.
2. For N=25, HC2 test out performs all others.
3. For N≥250, all HCCM tests perform well for all coefficients.
4. For N=1000, all rejection frequencies are close to the nominal 5% level of significance. Hence the results from all types of tests are indistinguishable.
5. Heteroscedasticity does not affect tests of each coefficient to the same degree.
6. This experiment shows that, for N<50 HCO can be even more misleading than the conventional OLSCM which ignores the possibility of heteroscedasticity.

CHAPTER FIVE

5. SUMMARY, CONCLUSION AND RECOMMENDATION

This chapter can be divided into two sections. The first section deals about conclusions and recommendations, and the second section mentions possible areas of further research.

5.1 Conclusions and Recommendations

Cross sectional data display some form of heteroskedasticity. It is common practice to still use the OLSCM estimator of the vector of regression parameters, since it remains unbiased and consistent.

Its covariance matrix, however, has to be consistently estimated in order for inference to be performed. From the preceding discussion it is clear that heteroscedasticity is potentially a serious problem and the researcher needs to know whether it is present in a given situation.

If its presence is detected, then one can take corrective action, such as using HCCM tests. In this study, we have explored the asymptotically justified versions of the HCCM and the standard OLSCM tests in the linear regression model.

While no Monte Carlo can represent all possible error structures that can be encountered in practice, the consistency of our results across the four error structures adds credence to our suggestions for using HCCM and OLSCM based tests in the linear regression model. Our results lead us to the following conclusions.

1. If there is a priori reason to suspect that there is heteroscedasticity from a matter of intuition, educated guess work, prior empirical evidence or sheer speculation, HCCM based tests should be used.

2. Since the cost of using HC3, instead of OLSCM when heteroskedasticity is absent, is apparently not very great (see table 4.2), it would seem wise to employ *t* statistic based on HC2 and HC3 even when there is little evidence of heteroskedasticity.

3. For homoscedastic data, all the HCCM tests are less reliable than OLSCM test. But HC3 test performs as good as OLSCM test.

4. We have examined the performance of three modified versions of White's (1980) HCCM estimator. In the presence of heteroscedasticity, among the HCCM tests, HC2 and HC3 consistently out perform the other HCCM estimators generally.

5. In the presence of heteroscedasticity and for $N \leq 250$, HC2 and HC3 tests should be used; when $N \geq 250$, other versions of the HCCM can be used.

5.2 Further Research

During the course of this study, I as well as my advisor came up with many ideas. Only a few of these ideas were implemented because the rest were beyond the scope of this study. These ideas are introduced here as possible areas of further research into the topic of HCCM estimators.

As we have tried to show in the analysis part of the experiments, HC3 did not always perform well when the sample size was small and there was substantial heteroscedasticity. Hence, it may be interesting to see how to improve the finite sample properties of this estimator.

. REFERENCES:

[1]. Andrew F. Hayes (2003), "Heteroscedasticity Consistent Standard Error Estimates for the Linear Regression Model: SPSS and SAS Implementation," The Ohio State University, Columbus, Ohio.

[2]. Chanadan Mukherjee, Howard White and Marc Wuyts "Econometrics and Analysis for Developing Countries," London and New York.

[3]. C.R.Rao (1970), "Estimation of Heteroscedastic Variances in Linear Regression Models," Journal of the American Statistical Association. 46, 234-239

[4]. Davidson and Mackinnon (1993), "Estimation and Inference in Econometrics," Oxford University Press.

[5]. Francisco Cribari- Neto (2000), "Improved Heteroscedasticity-Consistent Covariance Matrix Estimators," Biometrika, Printed in Great Britain.46, 213-234

[6]. Eicker, F. (1963), Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," The Annals of Mathematical Statistics 34, 447-456.

[7]. Greene, W.H. (1997), "Econometric Analysis (5th Ed.),"Upper Saddle River, NJ: Prentice Hall.

[8]. Gujarati, D.N. (1995), "Basic Econometrics (4th Ed.)," New York: McGraw-Hill.

[9]. Gujarati, D.N. (2004), "Basic Econometrics (4th Ed.)," New York: McGraw-Hill.

[10]. H.Glejser, (1969), "A New Test for Homoscedasticity," Journal of the American Statistical Association. 66, 416-423

[11]. Horn, S.D., R.A. Horn, and D.B. Duncan. (1975)," Estimating Heteroscedastic Variances in Linear Model, "Journal of the American Statistical Association, 70,380-385.

[12]. Jack Johnston (1983)," Econometric Methods." University of California, Irvine.

[13]. Jan Kmenta (1971), "Elements of Econometrics," New York: McGraw-Hill.

[14]. Kauermann, G. and Carroll, R. (2001), A Note on the Efficiency of Sandwich Covariance Matrix Estimation," Journal of the American Statistical Association, 96, 1387-1396

[15]. Long, J.S and Ervin, L. (2000), "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," The American Statistician, 54, 217-224

[16]. Michel Hurd (1979), "Estimation in Truncated Sample When There is Heteroscedasticity," State University of New York

[17]. Natalie Johnson (2007)," A Comparative Simulation Study of Robust Estimators of Standard Errors," Brigham Young University.

[18]. Peter Schmidt (1976),"Econometrics," New York: McGraw-Hill.

[19]. Ramu Ramanthan (1986), "Introductory Econometrics with Applications," University of California, San Diego

[20]. White, H. (1980), "A Heteroscedastic Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedsticity" Ecnometrica, 48,817-838

[21]. White, H. and MacKinnon, J. G. (1985), "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," Journal of Econometrics.29, 53-57

APPENDIX A: Monte Carlo Simulation Code in Matlab

A.1 OLSCM Estimator Code for Simulation in Matlab

```
function [SS, rejfreq] = hetrobustestimatorOLSCM;
% [SS, rejfreq] = hetrobustestimatorOLSCM;
% This program calculates the rejection frequencies (of incorrectly rejecting
% the null hypothesis) when OLSCM is used.
T = 1000;
gen1 = rand(T,1);
gen2 = randn(T,1);
x1 = 1+gen1;
x2 = 2*gen1 + 0.3*gen2;
SS = [];
for i = 1:10000
   %err = randn(T,1);    %(homoscedastic errors)
   %errnor = randn(T,1);     % Heteroscedastic errors (type 1)
   %err = diag(errnor)*x2;
   %errnor = randn(T,1);       % Heteroscedastic errors (type 2)
   %for i = 1:T
    %   err(i) = (x2(i) + 1.5)*errnor(i);
   %end
      %err = err';
```

```
errnor = randn(T,1);        % Heteroscedastic errors (type 3)
 for i = 1:T
    err(i) = x1(i)*(x2(i) + 2.5)*errnor(i);
 end
   err = err';
   X = [ones(T,1) x1 x2];
y =  ones(T,1) + x1 + x2 + err;
teta = inv(X'*X)*X'*y;
fit = X*teta;
res =  y - fit;
RSS = res'*res;
df = length(y)-3;
var =  RSS/df;
cm = var*inv(X'*X);
se = sqrt(diag(cm));
 for j = 1:length(X(1,:))
    sig(j) = (teta(j) - 1)/se(j);
    pval(j) =  2*(1- tcdf(abs(sig(j)),df));
    critval(j) = tinv(0.975,df);
    if abs(sig(j)) > critval(j)
        H(j) = 1;
    else
```

```
            H(j) = 0;
        end
    end
    SSp = [H(1) H(2) H(3)];
    SS = [SS;SSp];
    clear SSp err y teta fit res RSS var cm se sig pval critval H errnor
end
rejfreq1 = sum(SS(:,1))/length(SS(:,1));
rejfreq2 = sum(SS(:,2))/length(SS(:,1));
rejfreq3 = sum(SS(:,3))/length(SS(:,1));
rejfreq = [rejfreq1 rejfreq2 rejfreq3];
disp('**********************************************************************');
fprintf('rejection frequency for T =  %10f\n',T);
disp('  ____________________________________________');
disp('    for OLSCM test    ');
disp('  ____________________________________________');
disp('    beta 1      beta 2     beta 3');
disp('  ____________________________________________');
fprintf('%10.3f  %10.3f  %10.3f\n',rejfreq1,rejfreq2,rejfreq3);
```

A.2 HCO and HC1 Estimator Code for Simulation in Matlab

```
function [SS,SS1, rejfreqHC0,rejfreqHC1] = hetrobustestimatorHC01;
% [SS,SS1, rejfreqHC0,rejfreqHC1] = hetrobustestimatorHC01;
```

```
% This program calculates the rejection frequencies (of incorrectly rejecting
% the null hypothesis) when HC0 and HC1 are used.
T = 10;
gen1 = rand(T,1);
gen2 = randn(T,1);
x1 = 1+gen1;
x2 = 2*gen1 + 0.3*gen2;
SS = [];
SS1 = [];
for i = 1:10000
  %err = randn(T,1);   %(homoscedastic errors)
    %errnor = randn(T,1);    % Heteroscedastic errors (type 1)
  %err = diag(errnor)*x2;
 errnor = randn(T,1);       % Heteroscedastic errors (type 2)
  for i = 1:T
    err(i) = (x2(i) + 1.5)*errnor(i);
  end
  err = err';
   %errnor = randn(T,1);       % Heteroscedastic errors (type 3)
  %for i = 1:T
   %   err(i) = x1(i)*(x2(i) + 2.5)*errnor(i);
  %end
```

```
%err = err';
X = [ones(T,1) x1 x2];
y =  ones(T,1) + x1 + x2 + err;
teta = inv(X'*X)*X'*y;
fit = X*teta;
res =  y - fit;
df = length(y)-3;
erertran = res*res';
omega = diag(diag(erertran));
cm = inv(X'*X)*X'*omega*X*inv(X'*X);
cm1 = (T*cm)/(T-3);
se = sqrt(diag(cm));
se1 = sqrt(diag(cm1));
   for j = 1:length(X(1,:))
    sig(j) = (teta(j) - 1)/se(j);
    pval(j) =  2*(1- tcdf(abs(sig(j)),df));
    critval(j) = tinv(0.975,df);
    if abs(sig(j)) > critval(j)
        H(j) = 1;
    else
        H(j) = 0;
    end
```

```
    end
    SSp = [H(1) H(2) H(3)];
    SS = [SS;SSp];
     for j = 1:length(X(1,:))
       sig1(j) = (teta(j) - 1)/se1(j);
       pval1(j) =  2*(1- tcdf(abs(sig1(j)),df));
       critval1(j) = tinv(0.975,df);
       if abs(sig1(j)) > critval1(j)
           G(j) = 1;
       else
           G(j) = 0;
       end
    end
    SSp1 = [G(1) G(2) G(3)];
    SS1 = [SS1;SSp1];
    clear SSp SSp1 err errnor y teta fit res erertran omega cm cm1 se se1 sig sig1
    clear pval pval1 critval critval1 H G
end
 rejfreqHC01 = sum(SS(:,1))/length(SS(:,1));
rejfreqHC02 = sum(SS(:,2))/length(SS(:,1));
rejfreqHC03 = sum(SS(:,3))/length(SS(:,1));
rejfreqHC0 = [rejfreqHC01 rejfreqHC02 rejfreqHC03];
```

```
rejfreqHC11 = sum(SS1(:,1))/length(SS1(:,1));

rejfreqHC12 = sum(SS1(:,2))/length(SS1(:,1));

rejfreqHC13 = sum(SS1(:,3))/length(SS1(:,1));

rejfreqHC1 = [rejfreqHC11 rejfreqHC12 rejfreqHC13];

disp('**************************************************************');

fprintf('rejection frequency for T =  %10f\n',T);

disp('  _____________________________________________');

disp('   for HC0 test   ');

disp('  _____________________________________________');

disp('   beta 1     beta 2     beta 3');

disp('  _____________________________________________');

fprintf('%10.3f  %10.3f  %10.3f\n',rejfreqHC01,rejfreqHC02,rejfreqHC03);

disp('  _____________________________________________');

disp('**************************************************************');

disp('   for HC1 test   ');

disp('  _____________________________________________');

disp('   beta 1     beta 2     beta 3');

disp('  _____________________________________________');

fprintf('%10.3f  %10.3f  %10.3f\n',rejfreqHC11,rejfreqHC12,rejfreqHC13);
```

A.5 HC2 and HC3 Estimator Code for Simulation in Matlab

```
function [SS,SS1, rejfreqHC2,rejfreqHC3] = hetrobustestimatorHC23;

% [SS,SS1, rejfreqHC2,rejfreqHC3] = hetrobustestimatorHC23;
```

```
% This program calculates the rejection frequencies (of incorrectly rejecting
% the null hypothesis) when HC2 and HC3 are used.
T = 100;
gen1 = rand(T,1);
gen2 = randn(T,1);
x1 = 1+gen1;
x2 = 2*gen1 + 0.3*gen2;
 X = [ones(T,1) x1 x2];
AA = ones(T,1);
   M = diag(AA) - X*inv(X'*X)*X';
   MM = diag(diag(M));
   MM1 = MM*MM;
   KK = inv(MM);
   KK1 = inv(MM1);
SS = [];
SS1 = [];
for i = 1:10000
   %err = randn(T,1);    %(homoscedastic errors)
    errnor = randn(T,1);     % Heteroscedastic errors (type 1)
   err = diag(errnor)*x2;
    %errnor = randn(T,1);       % Heteroscedastic errors (type 2)
   %for i = 1:T
```

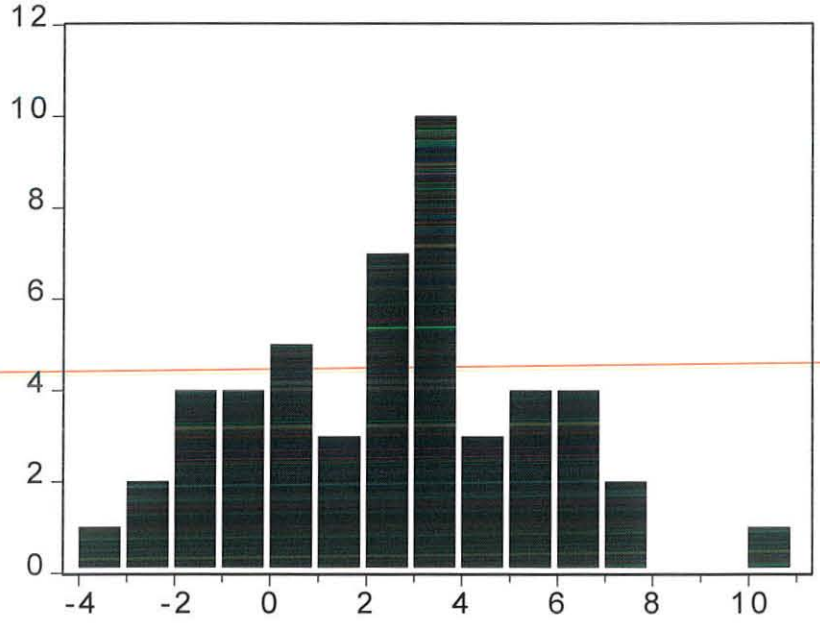
```
%   err(i) = (x2(i) + 1.5)*errnor(i);
%end
 %err = err';
%errnor = randn(T,1);        % Heteroscedastic errors (type 3)
%for i = 1:T
  % err(i) = x1(i)*(x2(i) + 2.5)*errnor(i);
%end
 %err = err';
 y =  ones(T,1) + x1 + x2 + err;
 teta = inv(X'*X)*X'*y;
fit = X*teta;
res =  y - fit;
df = length(y)-3;
erertran = res*res';
omega = diag(diag(erertran))*KK;
omega1 = diag(diag(erertran))*KK1;
cm = inv(X'*X)*X'*omega*X*inv(X'*X);
cm1 = inv(X'*X)*X'*omega1*X*inv(X'*X);
se = sqrt(diag(cm));
se1 = sqrt(diag(cm1));
  for j = 1:length(X(1,:))
    sig(j) = (teta(j) - 1)/se(j);
```

```
    pval(j) =  2*(1- tcdf(abs(sig(j)),df));
    critval(j) = tinv(0.975,df);
    if abs(sig(j)) > critval(j)
        H(j) = 1;
    else
        H(j) = 0;
    end
end
SSp = [H(1) H(2) H(3)];
SS = [SS;SSp];
  for j = 1:length(X(1,:))
    sig1(j) = (teta(j) - 1)/se1(j);
    pval1(j) =  2*(1- tcdf(abs(sig1(j)),df));
    critval1(j) = tinv(0.975,df);
    if abs(sig1(j)) > critval1(j)
        G(j) = 1;
    else
        G(j) = 0;
    end
end
SSp1 = [G(1) G(2) G(3)];
SS1 = [SS1;SSp1];
```

```
   clear SSp SSp1 err errnor y teta fit res erertran omega omega1
   clear cm cm1 se se1 sig sig1 pval pval1 critval critval1 H G
end
 rejfreqHC21 = sum(SS(:,1))/length(SS(:,1));
rejfreqHC22 = sum(SS(:,2))/length(SS(:,1));
rejfreqHC23 = sum(SS(:,3))/length(SS(:,1));
rejfreqHC2 = [rejfreqHC21 rejfreqHC22 rejfreqHC23];
rejfreqHC31 = sum(SS1(:,1))/length(SS1(:,1));
rejfreqHC32 = sum(SS1(:,2))/length(SS1(:,1));
rejfreqHC33 = sum(SS1(:,3))/length(SS1(:,1));
rejfreqHC3 = [rejfreqHC31 rejfreqHC32 rejfreqHC33];
disp('**************************************************************');
fprintf('rejection frequency for T =  %10f\n',T);
disp('  _____');
disp('   for HC2 test   ');
disp('  _____');
disp('   beta 1     beta 2     beta 3');
disp('  _____');
fprintf('%10.3f  %10.3f  %10.3f\n',rejfreqHC21,rejfreqHC22,rejfreqHC23);
disp('  _____');
disp('**************************************************************');
disp('   for HC3 test   ');
```

```
disp('  _____');
disp('    beta 1      beta 2      beta 3');
disp('  _____');
fprintf('%10.3f  %10.3f  %10.3f\n',rejfreqHC31,rejfreqHC32,rejfreqHC33);
```

APENDIX B. Plot of Y



Series: Y	
Sample 1 50	
Observations 50	
Mean	2.644097
Median	2.958014
Maximum	10.68513
Minimum	-3.876087
Std. Dev.	3.049242
Skewness	0.059361
Kurtosis	2.769321
Jarque-Bera	0.140224
Probability	0.932289

Declaration

I, the undersigned, declare that this thesis is my original work, has not been presented for degrees in any other university and all sources of material used for the thesis have been duly acknowledged.

Declared By:

Name: Yegnanew Alem

Signature: ----------

Place: Faculty of Science, Addis Ababa University

Date: June, 2008

This thesis has been submitted for examination with my approval as a university advisor.

Name: Olusanya E. Olubusoye (Ph.D)

Signature: ----------

Date: June, 2008