

Inference for linear regression models in the presence of heteroskedasticity

SW Rheeder
21768889

Dissertation submitted in partial fulfilment of the requirements
for the degree *Magister Scientiae* in *Statistics* at the
Potchefstroom Campus of the North-West University

Supervisor: Dr L Santana

May 2016

Acknowledgements

I would like to thank my parents Johan Rheeder and Riana Rheeder for the financial and emotional support throughout my tertiary education. Without your support and sacrifices I would not be where I am today.

At the North-West university I would like to express my sincerest thanks to:

- Dr. Leonard Santana for his guidance and insight into linear regression models and the bootstrap, and also for his assistance with the compilation of this text using $\text{\LaTeX}$ and the R code for the simulation study.
- Prof. James Allison for his assistance with bursaries in order pursue my post graduate degrees in Statistics.
- Prof. Jan Swanepoel and Prof. Corrie Swanepoel for their insight into the fields of non-parametric statistics and probability theory.

“You are caught in the current of unceasing change. Your life is a ripple in it. Every moment of your conscious life links the infinite past with the infinite future. Take part in both and you will not find the present empty.” – Oswald Spengler

Inference for linear regression models in the presence of heteroskedasticity

Smartreik Wessel Rheeder

School of Computer, Statistical, and Mathematical Sciences
Statistics Department
North-West University, Potchefstroom Campus, North-West Province
May, 2016

Keywords: Bootstrap, Heteroskedasticity, Heteroskedastic consistent covariance estimators, Hypothesis testing, Linear regression, Power of the test, Size of the test.

ABSTRACT

The linear regression model is a highly versatile tool with which one can model a response variable in terms of one or more predictor variables. The classical linear model is based on six primary theoretical assumptions. In this study the main focus is on the assumption of “homoskedasticity” and the violation thereof, called “heteroskedasticity”. Heteroskedasticity refers to the property where the error terms in the linear model do not all have the same variance. This text explores methods for conducting hypothesis tests of regression coefficients in the presence of heteroskedasticity. Various approaches for performing inference in the presence of heteroskedasticity are investigated, including tests that incorporate heteroskedasticity consistent covariance matrix estimators (HCCMEs), the wild bootstrap, and a newly proposed method based on a modification of the bootstrap residuals approach. A simulation study was conducted to compare the proposed new modified bootstrap residuals approach to the wild bootstrap in terms of size and power to determine if the new approach had any merit. The impact on the wild bootstrap’s size and power was also investigated when using different residual transformations, HCCMEs and ancillary distributions. It was found that when large leverage values were present in small homoskedastic samples, the sizes of the tests were highly elevated (tending to over-reject); the sizes of the test decreased to the specified significance level when the heteroskedasticity increased. The power of tests based on data with one or more large leverage values were more powerful than those based on data with small to moderate leverages. From the result in the study it is clear that the newly proposed modified bootstrap residuals approach is competitive compared to the wild bootstrap approach. We thus recommend that this new method be studied in more (theoretical) detail as a possible future research project.

OPSOMMING

Die lineêre regressiemodel is 'n veelsydige tegniek waarmee 'n responsveranderlike in terme van een of meer voorspeller veranderlikes gemodelleer kan word. Die klassieke lineêre model is gebaseer op ses hoof teoretiese aannames. Die primêre fokus in hierdie studie sal op die aanname van “homoskedastisiteit” wees, en meer spesifiek op die afwyking vanaf hierdie aanname, naamlik “heteroskedastisiteit.” Heteroskedastisiteit verwys na die eienskap waar die foutterme in die lineêre model nie almal dieselfde variansie het nie. In hierdie studie word metodes verken vir die uitvoering van hipotesetoetse vir regressiekoëffisiënte in die teenwoordigheid van heteroskedastisiteit. Verskillende benaderings vir die uitvoering van inferensie in die teenwoordigheid van heteroskedastisiteit word ondersoek, insluitend toetse wat heteroskedastiese konsekwente kovariansiematriksberamers (HKKMB's) inkorporeer, die wildeskoenlus, en 'n nuwe voorgestelde metode wat gebaseer is op 'n gewysigde weergawe van die skoenlusresidue benadering. 'n Simulasiestudie is uitgevoer om die voorgestelde nuwe gewysigde skoenlusresidue benadering te vergelyk met die wildeskoenlus in terme van toetsgrootte en onderskeidingsvermoë om sodanig te bepaal of die nuwe benadering enige meriete het. Die impak van verskillende transformasies op die residue, HKKMB's, en aanvullende verdelings op die toetsgrootte en onderskeidingsvermoë van die wildeskoenlus toetse is ook ondersoek. Daar is gevind dat, wanneer groothefboomwaardes in klein homoskedastiese steekproewe teenwoordig is, die toetsgrootte baie liberaal is (die tendens was om te maklik te verwerp), maar met 'n toename in heteroskedastisiteit, het die toetsgrootte afgeneem tot by die gespesifiseerde betekenispeil. Die onderskeidingsvermoë van toetse wat gebaseer is op datastelle waarin daar groot hefboomwaardes voorkom, het 'n groter onderskeidingsvermoë gehad as dié wat gebaseer is op data met klein tot matige hefboomwaardes. Uit die resultate in die studie blyk dit dat die skoenlus residubenadering goed meeding met die wildeskoenlus benadering. Ons stel dus voor dat die nuwe metode in meer (teoretiese) detail bestudeer word as 'n toekomstige navorsingsprojek.

Table of Contents

1	Introduction	1
2	Linear regression models	4
2.1	Introduction	4
2.2	Classical linear model	8
2.2.1	Assumptions of the classical linear model	9
2.3	Ordinary least squares (OLS)	11
2.4	Finite sample properties of the ordinary least squares estimator	13
2.4.1	The OLS estimator $\hat{\beta}$ is unbiased for β	13
2.4.2	The variance/covariance matrix of $\hat{\beta}$	14
2.4.3	Gauss Markov theorem	14
2.4.4	Estimator for $Var(\hat{\beta})$	16
2.4.5	Hypothesis testing	18
2.5	Large sample properties of the ordinary least squares estimator	19
2.5.1	$\hat{\beta}$ is a consistent estimator for β	19
2.5.2	Asymptotic distribution of the ordinary least squares estimator	22
2.6	Generalised linear model	25
2.7	Generalised least squares (GLS)	28
2.7.1	Singular value decomposition and the generalised least squares estimator	28
2.7.2	Properties of the generalised least squares estimator	30
2.8	Heteroskedasticity and remedial measures	32
2.8.1	Weighted least squares (WLS)	33
2.8.2	Feasible generalised least squares (FGLS)	34
2.8.3	Heteroskedastic consistent covariance matrix estimators (HCCMEs)	35
3	The Bootstrap	41
3.1	An introduction to the Bootstrap	42
3.2	Bootstrap hypothesis testing	45

3.3	Bootstrap regression	51
3.3.1	Bootstrap residuals approach	53
3.3.2	Bootstrap pairs approach	55
3.3.3	Wild bootstrap/External bootstrap	56
3.4	Bootstrap regression hypothesis testing	63
3.4.1	Hypothesis testing with homoskedastic errors: Residuals approach	64
3.4.2	Hypothesis testing with homoskedastic or heteroskedastic errors: Pairs approach	69
3.4.3	Hypothesis testing with heteroskedastic errors: Wild bootstrap approach .	71
4	Proposed new bootstrap approach for dealing with heteroskedasticity . . .	75
4.1	Modified bootstrap residuals approach for dealing with heteroskedasticity	75
4.2	Other models for the heteroskedastic structure	80
4.2.1	Park test	80
4.2.2	Glejser test	81
4.2.3	Goldfeld-Quandt test	81
4.2.4	Breusch-Pagan-Godfrey test	82
4.2.5	White's general test	82
4.2.6	Koenker-Bassett test	82
5	Simulation study	83
5.1	Heteroskedastic data generation procedure	83
5.1.1	Data generation design for this simulation study	85
5.2	Hypothesis used and test statistics for this simulation study	87
5.2.1	Algorithms used for this simulation study	88
5.3	Monte Carlo approximation of size and power of a test	91
5.4	Simulation results – Sizes of the tests	95
5.4.1	Remarks on the sizes of tests using lognormal predictors	96
5.4.2	Remarks on the sizes of tests using uniform predictors	97
5.5	Simulation results – Powers of selected tests	113
5.5.1	Remarks on the power of the test for this simulation study	114
6	Conclusion and future considerations	119
6.1	Conclusions from the simulation study	120
6.2	Future considerations	121

Bibliography	122
------------------------	-----

List of Figures

2.1	Functional relationship (left) and statistical relationship (right) in two dimensions - single response and predictor variable.	4
2.2	Functional relationship in three dimensions shifted to a statistical relationship in two dimensions.	5
2.3	First ever regression curve from Galton's sweet pea data, adapted from Pearson (1930).	8
3.1	Different "worlds".	43
3.2	Accuracy and precision illustration.	44
5.1	Examples of homoskedasticity and heteroskedasticity produced using model (5.1) with a sample size of $n = 160$	86
5.2	Size of the test for lognormal predictors and a sample size of $n = 20$	99
5.3	Size of the test for uniform predictors and a sample size of $n = 20$	100
5.4	Size of the test for lognormal predictors and a sample size of $n = 40$	101
5.5	Size of the test for uniform predictors and a sample size of $n = 40$	102
5.6	Size of the test for lognormal predictors and a sample size of $n = 80$	103
5.7	Size of the test for uniform predictors and a sample size of $n = 80$	104
5.8	Size of the test for lognormal predictors and a sample size of $n = 160$	105
5.9	Size of the test for uniform predictors and a sample size of $n = 160$	106
5.10	Size of the test for lognormal predictors and a sample size of $n = 320$	107
5.11	Size of the test for uniform predictors and a sample size of $n = 320$	108
5.12	Size of the test for lognormal predictors and a sample size of $n = 640$	109
5.13	Size of the test for uniform predictors and a sample size of $n = 640$	110
5.14	Size of the test for lognormal predictors and a sample size of $n = 1\,280$	111
5.15	Size of the test for uniform predictors and a sample size of $n = 1\,280$	112
5.16	Power for HC3 bootstrap methods, $n = 20$ and lognormal predictors.	115
5.17	Power for HC3 bootstrap methods, $n = 20$ and uniform predictors.	115
5.18	Power for HC3 bootstrap methods, $n = 40$ and lognormal predictors.	116
5.19	Power for HC3 bootstrap methods, $n = 40$ and uniform predictors.	116

5.20	Power for HC3 bootstrap methods, $n = 80$ and lognormal predictors.	117
5.21	Power for HC3 bootstrap methods, $n = 80$ and uniform predictors.	117
5.22	Power for HC3 bootstrap methods, $n = 160$ and lognormal predictors.	118
5.23	Power for HC3 bootstrap methods, $n = 160$ and uniform predictors.	118

List of Tables

2.1	Proposed estimators for $\overline{\mathbf{V}}_n$	40
5.1	Different wild bootstrap residual transformations, HCCMEs and ancillary distributions used with corresponding shorthand notation.	94

CHAPTER 1

Introduction

The linear regression model is a highly versatile and practical statistical model which can be used to model and predict a response variable using one or more predictors variables (see e.g., Greene, 2003; Kutner, Nachtheim, Neter and Li, 2005). The model is fitted using a method known as ordinary least squares (OLS), which minimises the squared residuals to ensure the best fit. In order to confirm if the fitted linear model is appropriate some statistical inference regarding the regression coefficients are needed. However, in order to conduct these tests model assumptions are required.

There are six primary assumptions to the classical linear regression model, but the one that is most central to this text is known as “homoskedasticity”. Homoskedasticity refers to the property where all the error terms in the linear model have the same variance across observations. On the other hand, when the variance of the error terms are not the same we get what is known as “heteroskedasticity”¹.

The presence of heteroskedasticity is of concern because the majority of the statistical theory related to inference for linear models was developed for the homoskedastic case. Homoskedasticity is assumed because it ensures that the Gauss-Markov theorem is satisfied. Specifically, the Gauss-Markov theorem states that the OLS estimator for the regression coefficients is the minimum variance linear unbiased estimator under the classical model assumptions. If the data exhibits heteroskedasticity the OLS estimator may not retain the minimum variance property and the estimates may be inconsistent (Greene, 2003; Davidson and MacKinnon, 2004). A modification to the linear model is required to retain this property, however it requires making more assumptions which are generally not feasible. White (1980) showed that the OLS estimator for the regression

¹The etymology of the two words are derived from the Greek words *homo* and *hetero*, which mean *the same* and *different*, respectively, and the word *skedastikos* which means *able to disperse*, which has its root in the words *skedannynai* or *skedannumi*, meaning *to disperse* (Merriam-Webster Dictionary, 2015). Throughout the text we used the spelling convention suggested by McCullough (1985), i.e., by using a “k” rather than a “c”.

coefficients remains unbiased and consistent under heteroskedasticity, and also proved that the regression coefficients estimator's variance/covariance matrix could be consistently estimated. It was shown that a consistent estimator could be obtained by simply consistently estimating a low dimensional matrix quantity which appears within the variance/covariance matrix expression. The problem of estimating the variance/covariance matrix of the estimators thus reduced to estimating a matrix of a much lower dimension than what was originally thought required. These consistent estimates became known as heteroskedastic consistent covariance matrix estimators (HCCMEs). Simulation evidence (MacKinnon and White, 1985; Long and Ervin, 1998) for these estimators suggest that they do not work well for small sample sizes and so bootstrap techniques were considered as a potential solution for alleviating these distortions associated with small sample sizes.

The bootstrap is a modifiable tool that allows one to study distributional properties of statistics and test statistics by approximating these properties with a Monte Carlo simulation approach (Efron and Tibshirani, 1993). This approach can be employed with the HCCMEs to construct tests for regression coefficients in the presence of heteroskedasticity. The first regression bootstrap approach was the residuals based approach (Efron, 1979), however this approach requires the assumption of homoskedasticity to hold in order for it to function correctly. Another approach, called the pairs bootstrap (Freedman, 1981), was suggested for dealing with heteroskedasticity, but simulation evidence suggested that it did not fare as well as another approach called the wild bootstrap. The wild bootstrap approach is a heteroskedasticity robust approach that preserves the heteroskedastic structure of the original data (Wu, 1986; Liu, 1988; Mammen, 1993).

In this text we shall explore and also measure the performance of these bootstrap approaches, specifically in the context of hypothesis tests for regression coefficients. Attention will be given to tests that combine the wild bootstrap approach and HCCMEs. In addition, these tests will serve as a benchmark against which a newly proposed bootstrap test will be compared. The new test is a modification of the bootstrap residuals approach and was created to remedy the shortcoming of the residual approach by modelling the heteroskedasticity with a simple model. The rationale is to maintain the original structure in the bootstrap samples and correct for it by means of a suitable HCCME. This can be achieved by modelling the heteroskedastic structure using the squared OLS residuals, the fitted values of which serve as estimators for the variances. Using

these estimators one can remove the heteroskedastic structure from the residuals, resample these residuals to form bootstrap samples, and then correct the bootstrap samples by reintroducing the heteroskedastic structure back into them. Finally, test statistics can be calculated and used to conduct inference.

The performance of these tests were measured by conducting an extensive simulation study whereby the size and power of the tests were compared to one another.

The remainder of this text is distributed as follows: Chapter 2 introduces and develops the theory for linear models and hypothesis testing, and the quandary of heteroskedasticity as well as its effect on hypothesis testing is discussed. The chapter is concluded with a few suggestions on how heteroskedastic robust inference might be achieved. In Chapter 3 the highly versatile method known as the bootstrap is introduced and the application thereof to hypothesis testing is illustrated by means of considering a simple example. The different bootstrap approaches for linear regression models are covered and the way in which they can be applied to hypothesis testing for regression coefficients are presented, along with appropriate algorithms. The new modified bootstrap residuals approach is proposed in Chapter 4 and wraps up with a few suggestions from the literature on how to model heteroskedastic structures. Chapter 5 discusses the details around the simulation study and presents the results thereof with a few remarks. The final chapter, Chapter 6, provides final comments on the simulation study and concludes with a few considerations for future research.

CHAPTER 2

Linear regression models

2.1 Introduction

A linear regression model is a statistical model for the linear relationship between a single response ¹ variable, denoted by Y and one or more predictor ² variables denoted by X . The linear relationship that is modelled is useful for predicting the response variable using specific values of the predictor variables. There exists two types of relationships: functional and statistical. Functional relationships are, in general, of the form $Y = f(X)$, where $f(\cdot)$ denotes some known mathematical function. A statistical relationship has the form $Y = f(X) + \varepsilon$, where ε represents some stochastic error term and $f(\cdot)$ again denotes some known mathematical function. Figure 2.1, illustrates these two types of relationships using a linear form for $f(\cdot)$.

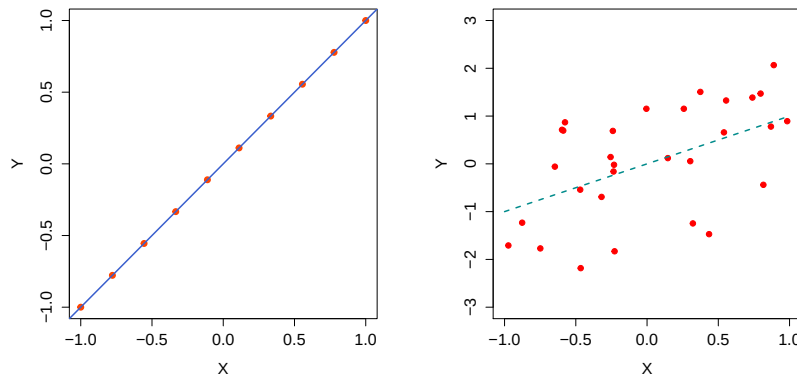


Fig. 2.1: Functional relationship (left) and statistical relationship (right) in two dimensions - single response and predictor variable.

On the left side we see the functional or “perfect” relationship between X and Y , i.e., the observed values (orange points) are perfectly located on the blue line. On the right, the points still follow

¹Also known as either an outcome or dependent variable.

²Also known as an independent variable.

the same general tendency as the functional relationship, but one sees that the points are now “noisy”, i.e., the red points are scattered around the fitted dashed blue line. The reason for the scattering can potentially be explained by considering a higher dimensional problem as shown in Figure 2.2.

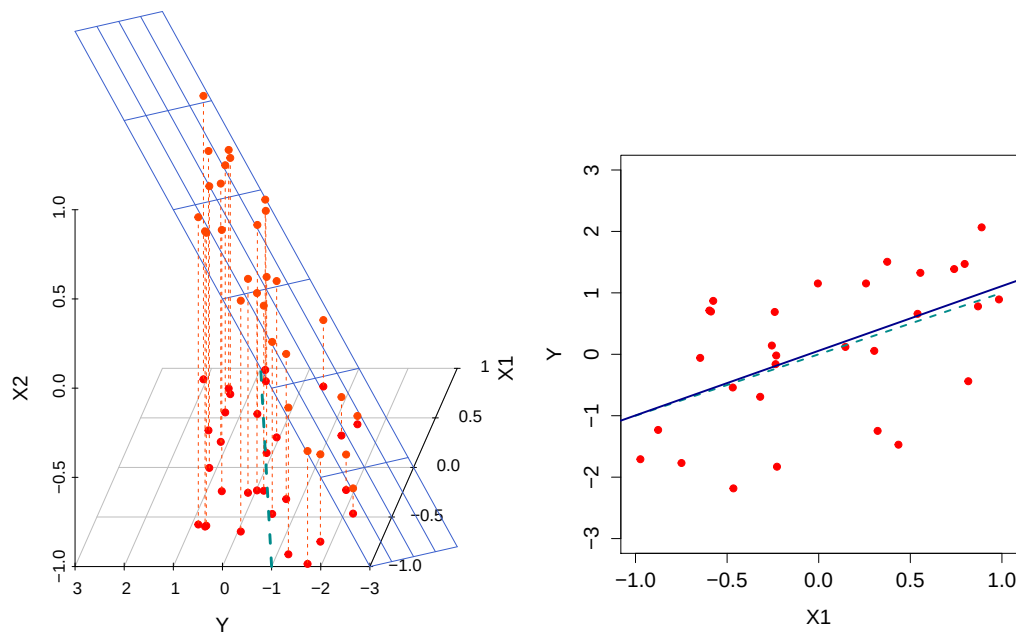


Fig. 2.2: Functional relationship in three dimensions shifted to a statistical relationship in two dimensions.

On the left of Figure 2.2 we see that, in three dimensions, the points lie perfectly on the plane $X_1 + 2X_2 = Y$. The graph on the right shows the projection of these same points onto the (X_1, Y) plane where we again see that the relationship is “noisy” since the information in X_2 is no longer considered. This example illustrates how the two concepts of functional and statistical relationships can be linked together in order to provide a basic motivation for the use of linear regression models. However, the motivation, which is dependent on three assumptions (Pratt and Schlaifer, 1985), is somewhat simplistic and serves only as an illustration rather than a rigorous theory. The three assumptions are:

- the combined effect of the predictor variables that have been omitted is independent of the predictor variables already included in the model,
- the combined effect of the predictor variables that have been omitted is independent across

observations, and

- the combined effect of the predictor variables that have been omitted has expectation zero.

Returning then to the points shown in Figure 2.2, we see that the graph on the left is a functional relationship of the form $Y = f(X_1, X_2) = 1X_1 + 2X_2$. We notice that all the data points (orange dots) lie on this surface and are thus representative of the previously mentioned “perfect” fit. When we set out to model a response variable using a linear regression model, we hope to model that perfect fit and thus explain how it is influenced by the predictor variables completely. However, that is not possible due to the experimenter not knowing beforehand which predictors are required to provide such a fit and/or due to other nuances such as measurement error. To better understand a statistical relationship, suppose the experimenter excluded the predictor X_2 from their model, this is represented by projecting the orange points onto the (X_1, Y) space, where the red dots now represent the data. The graph on the right is a two dimensional representation of this projection, and shows the “true” line, $Y = f(X_1) = X_1$, given by the dashed blue line versus the fitted regression line given by the solid blue line. By excluding the predictor X_2 , the hope of a perfect fit is not possible any more, and it is clear that some kind of “noise” has been introduced into the data. The noise in this example clearly stems from the fact that X_2 was omitted and needs to be accounted for by including an error term such that the relationship is now represented by $Y = 1X_1 + \varepsilon$. Therefore, the noise represents the information lost by excluding the predictor X_2 from the linear model.

A linear regression or statistical model is useful for investigating two important components of a statistical relationship (Kutner et al., 2005).

1. How the response variable Y varies with the predictor variable/s X in a systematic manner.
2. The scattering of observations around the statistical relation curve.

These components manifest in a regression model as:

1. A probability distribution of the response Y for each value of the predictor X .
2. The mean values of the response probability distribution varies in a systematic manner with the predictors.

The most basic linear regression model with one predictor variable called a “simple linear regression” model, will now be more formally defined as:

$$\begin{aligned} Y_i &= f(X_{i,1}) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i,1} + \varepsilon_i, \forall i = 1, 2, \dots, n. \end{aligned} \tag{2.1}$$

Equation (2.1) consists of both a functional $f(X_{i,1})$ and random/stochastic ε_i component, together they form the statistical relation. Since ε_i is random, Y_i will also be random as a result. The basic linear regression model and the concept of correlation was the inspiration of Francis Galton (Stanton, 2001), although the mathematical theory was developed by Karl Pearson. Galton’s interest in genetics and heredity led to the development of his theory of regression to the mean that he formalised as the simple linear regression model. The question around heredity, i.e., the transmission of genetic characteristics from one generation to the next, prompted his initial investigation by considering the cress plant and due to the nonlinear nature of the regression did not make any progress in discovering the new statistical tool (Pearson, 1930). He then chose the sweet pea because it was a self-fertilising organism and so the characteristics from mother to daughter plant could be studied without the contribution from another parent. This was done by distributing packets of seeds to nine friends who were then asked to germinate them. Each packet contained seven different sized sweet pea seeds (ten of each size). Unfortunately, two crops failed and he only received seven batches of new seeds which totalled $7 \times 7 \times 10 = 490$ seeds. The connection was made after the diameters of the mother and daughter seeds were plotted against each other, see Figure 2.3. He also stumbled upon what is known as homoskedasticity when he reported *“I was certainly astonished to find the family variability of the produce of the little seeds to be equal to that of the big ones; but so it was, and I thankfully accept the fact; for if it had been otherwise, I cannot imagine, from theoretical considerations, how the typical problem could be solved.”* The seemingly simple idea of regression grew to encompass multiple predictors and a host of other modifications.

In this Chapter we begin by introducing the basics behind the classical linear regression model in Section 2.2 by formally defining the model and stating the model assumptions thereof. We then consider estimating the classical linear regression models’ regression coefficients by using the

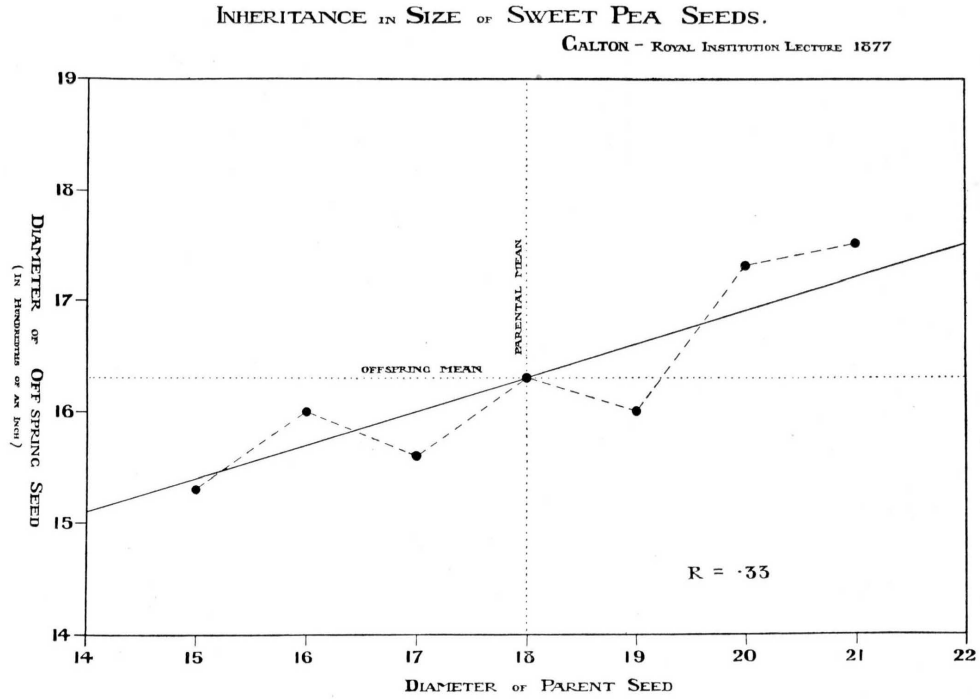


Fig. 2.3: First ever regression curve from Galton's sweet pea data, adapted from Pearson (1930).

method known as ordinary least squares (OLS) and discuss the finite and asymptotic properties thereof in Section 2.3. In order to relax some of the model assumptions, the generalised linear model is introduced in Section 2.6 as an intuitive extension of the classical model, followed by estimation using generalised least squares (GLS) in Section 2.7. In order to try and remedy the problems associated with heteroskedasticity when conducting inference the Sections 2.8.1 , 2.8.2 and 2.8.3 will provide an overview of different methods.

2.2 Classical linear model

Our investigation into linear regression models starts with the most commonly defined linear regression model known as the classical linear regression model. The classical multiple linear regression model (see for instance Greene, 2003; Kutner et al., 2005) based on $k - 1$ predictor variables is defined as:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{k-1} X_{i,k-1} + \varepsilon_i ; i = 1, 2, \dots, n$$

or in vector and matrix notation,

$$\underset{(n \times 1)}{\mathbf{Y}} = \underset{(n \times k)}{\mathbf{X}} \underset{(k \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}, \quad (2.2)$$

which translates into

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,k-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,k-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where

$\mathbf{Y}$ is an $(n \times 1)$ random response vector,

$\mathbf{X}$ is an $(n \times k)$ matrix with k independent variable column vectors, where $k - 1$ denotes the number predictor variables in the model,

$\boldsymbol{\beta}$ is a $(k \times 1)$ vector of unknown regression coefficients, assumed to be nonrandom, and

$\boldsymbol{\varepsilon}$ is an $(n \times 1)$ random error term vector, with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, $\forall i = 1, 2, \dots, n$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$, $\forall i \neq j$; $i, j = 1, 2, \dots, n$. The error terms are typically assumed to be multivariate normally distributed, i.e., $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\mathbf{I}_n$ is an $(n \times n)$ identity matrix.

This model may be regarded as the most “simple” model, but it is widely used in practice because of the convenience when making inferences. However, the convenience comes at a price: a range of model assumptions are required to hold true for finite and asymptotic results, these will be presented in Sections 2.4 and 2.5, respectively. Next the classical model assumptions will be introduced, followed by the method of least squares estimation for the regression coefficients (Section 2.3).

2.2.1 Assumptions of the classical linear model

The assumptions in the classical linear regression model include the following:

Assumption 1: Linearity

Linearity of the regression model (2.2) specifies a linear relationship between the response variable Y and the predictor variables $X_1, X_2, \dots, X_{k-1}$.

Assumption 2: The design matrix is a full rank matrix

The design matrix $\mathbf{X}$ must have full column rank to ensure that the predictor variables are not linearly dependent; a condition which is pertinent to the estimation of the regression coefficients and is required for the existence of the solution.

Assumption 3: Exogeneity of the predictor variables

The predictors do not contain any useful information for predicting the error terms, i.e. $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$.

Assumption 4: Homoskedasticity and uncorrelated errors

Each of the error terms ε_i has finite variance σ^2 and is uncorrelated with the other error terms ε_j , $i \neq j$; $i, j = 1, 2, \dots, n$; i.e., $Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$. Therefore, $Var(\boldsymbol{\varepsilon})$ contains σ^2 on the diagonal and 0 on the off-diagonal entries of the variance/covariance matrix. Note that the term “homoskedasticity” refers to the property where all the error terms have the same variance; this is in contrast with “heteroskedasticity” where the variances of the errors may not necessarily all be the same.

Assumption 5: The types of predictor variables

The design matrix $\mathbf{X}$ may be fixed or random/stochastic, however it is generated by some mechanism which is unrelated to the errors $\boldsymbol{\varepsilon}$. When the values of predictor variables are chosen by the experimenter, the design matrix will be fixed and as a result viewed as a known constant in the probability distribution of the response variable Y_i . On the other

hand if the predictors are observed and thus random, the joint distribution of $\mathbf{X}_i$ and Y_i is of concern when considering assumptions (1)–(4).

Assumption 6: Normality

The error terms have a multivariate normal distribution, $\varepsilon|\mathbf{X} \stackrel{i.i.d.}{\sim} N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. This assumption is not necessary for estimation, but serves as a convenience when deriving the distributions of test statistics for inferential purposes.

2.3 Ordinary least squares (OLS)

Now that the classical linear model and its assumptions have been defined, we turn our attention to estimation of the unknown regression coefficients for the classical model in Equation (2.2). The most commonly used method to estimate the unknown quantities in the regression coefficient vector β is the method of least squares, but other methods such as maximum likelihood, method of moments, Gauss-Newton, lasso, elastic net, ridge regression, least angle regression, orthogonal matching pursuit and others exist in the literature (Hoerl and Kennard, 1970; Davidson and MacKinnon, 1993; Tibshirani, 1996; Greene, 2003; Efron, Hastie, Johnstone and Tibshirani, 2004; Davidson and MacKinnon, 2004; Kutner et al., 2005; Zou and Hastie, 2005; Johnson and Wichern, 2007).

To begin the discussion around estimation using the method of least squares, an important distinction should be made between the unknown regression quantities β and ε_i and their sample estimates denoted by $\hat{\beta}$ and e_i , respectively. The unknown population regression function is given by

$$E[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$$

whereas its estimate is denoted by,

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}.$$

The random error term of the i^{th} data point is given by

$$\varepsilon_i = Y_i - \mathbf{X}_i\boldsymbol{\beta},$$

with $\mathbf{X}_i$ the i^{th} row of the design matrix $\mathbf{X}$. The sample version of ε_i is denoted by

$$e_i = Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}},$$

and is known as the i^{th} residual. The method of ordinary least squares estimates the unknown regression coefficient values $\boldsymbol{\beta}$ in the classical linear regression model (2.2) by minimising the sum of the squared residuals. The squared sum of the residuals can be written as follows to solve for $\boldsymbol{\beta}$ such that an expression for its estimator can be found,

$$\begin{aligned}\mathbf{Q}(\boldsymbol{\beta}) &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta},\end{aligned}$$

since $\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} = (\mathbf{Y}'\mathbf{X}\boldsymbol{\beta})' = \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y}$, i.e. the transpose of a scalar is again the scalar.

To obtain the values of $\boldsymbol{\beta}$ that minimise $\mathbf{Q}(\boldsymbol{\beta})$, we first need to determine the derivative of $\mathbf{Q}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$,

$$\frac{\partial \mathbf{Q}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

Setting the derivative equal to $\mathbf{0}$ and assuming $\mathbf{X}$ satisfies the full column rank assumption, we obtain:

$$\begin{aligned}-2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{0} \\ \Rightarrow \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.\end{aligned}\tag{2.3}$$

This estimator is known as the ordinary least squares estimator, which is unique and minimises the sum of squared residuals.

2.4 Finite sample properties of the ordinary least squares estimator

In this section we consider the statistical properties of the classical linear regression model that are valid for any sample size n . We start by proving that the OLS estimator given in Equation (2.3) is unbiased, and then derive an expression for its variance/covariance matrix. The OLS estimator is then shown to be the minimum variance linear unbiased estimator by means of the well known Gauss-Markov theorem. Finally, we derive an estimator for the variance/covariance matrix of the OLS estimator, and conclude the section with hypothesis testing using the finite sample properties and the normality assumption.

2.4.1 The OLS estimator $\hat{\beta}$ is unbiased for β

The regression coefficients estimator, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ exists if and only if the design matrix $\mathbf{X}$ has full column rank, i.e. $\mathbf{X}'\mathbf{X}$ is invertible. The OLS estimator can be rewritten as:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon.\end{aligned}\tag{2.4}$$

Taking the expectation on both sides, yields:

$$\begin{aligned}E(\hat{\beta}) &= E(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon) \\ &= \beta + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon) \\ &= \beta,\end{aligned}\tag{2.5}$$

since the expectation in (2.5) is zero due to $E(\varepsilon) = \mathbf{0}$ if the $\mathbf{X}$ s are fixed. In the case where the $\mathbf{X}$ s are random, we can use the exogeneity assumption from Section 2.2.1, and note that $\hat{\beta}$ is conditionally unbiased given $\mathbf{X}$, i.e., $E(\varepsilon|\mathbf{X}) = \mathbf{0}$.

2.4.2 The variance/covariance matrix of $\hat{\beta}$

The expression for the variance/covariance matrix of the OLS estimator is given by:

$$\begin{aligned}
 Var(\hat{\beta}) &= E((\hat{\beta} - \beta)(\hat{\beta} - \beta)') \\
 &= E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon)'] \\
 &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
 &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
 \end{aligned} \tag{2.6}$$

The unbiased estimator for this quantity will be derived in Section 2.4.4.

2.4.3 Gauss Markov theorem

The Gauss Markov theorem states that the ordinary least squares estimator $\hat{\beta}$ is efficient in the sense that its variance/covariance matrix $Var(\hat{\beta})$ is smaller than or equal to the variance/covariance class of linear unbiased estimators for β , i.e., $Var(\tilde{\beta}) \geq Var(\hat{\beta})$, where $\tilde{\beta}$ is some other unbiased linear estimator for β .

Proof. Because $\tilde{\beta}$ was defined to be an unbiased linear estimator it may be expressed as the linear combination $\tilde{\beta} = \mathbf{C}\mathbf{Y}$. For $\tilde{\beta}$ to be unbiased, $E(\tilde{\beta})$ must equal β . However, we know that $E(\mathbf{C}\varepsilon) = \mathbf{0}$ (since, by assumption $E(\varepsilon) = \mathbf{0}$, and therefore $E(\mathbf{C}\varepsilon) = \mathbf{0}$), so it is required that $\mathbf{C}\mathbf{X} = \mathbf{I}_k$ in the expression $\tilde{\beta} = \mathbf{C}\mathbf{Y} = \mathbf{C}(\mathbf{X}\beta + \varepsilon) = \mathbf{C}\mathbf{X}\beta + \mathbf{C}\varepsilon$ for $\tilde{\beta}$ to be unbiased. Now, if we define $\mathbf{C}$ as $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}$, then in order for $\mathbf{C}\mathbf{X} = \mathbf{I}_k$ we must have $\mathbf{D}\mathbf{X} = \mathbf{0}$, as shown below:

$$\begin{aligned}
 \mathbf{C} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D} \\
 \iff \mathbf{D} &= \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.
 \end{aligned}$$

Therefore it follows that,

$$\begin{aligned}
\mathbf{DX} &= \mathbf{CX} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\
&= \mathbf{CX} - \mathbf{I}_k, \text{ but } \mathbf{CX} = \mathbf{I}_k \text{ if } \tilde{\boldsymbol{\beta}} \text{ is unbiased} \\
&= \mathbf{0}.
\end{aligned}$$

We also note that $\mathbf{D}\boldsymbol{\varepsilon} = \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$:

$$\begin{aligned}
\mathbf{D}\boldsymbol{\varepsilon} &= (\mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{CY} - \mathbf{CX}\boldsymbol{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} + \mathbf{I}_k\boldsymbol{\beta}, \text{ since } \mathbf{CX} = \mathbf{I}_k \\
&= \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} - \hat{\boldsymbol{\beta}} + \boldsymbol{\beta} \\
&= \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}.
\end{aligned} \tag{2.7}$$

Next we shall derive an expression for the covariance between $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})$ which will be used to prove the final result:

$$\begin{aligned}
E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})') &= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})' \\
&= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})' \\
&= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})' \\
&= E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})(\mathbf{D}\boldsymbol{\varepsilon})', \text{ from (2.7) above} \\
&= E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{D}' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{D}' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}' = \mathbf{0}, \text{ because } \mathbf{DX} = \mathbf{X}'\mathbf{D}' = (\mathbf{DX})' = \mathbf{0}.
\end{aligned} \tag{2.8}$$

Note that this result also implies that the covariance between $\hat{\boldsymbol{\beta}}$ and $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})$ is $\mathbf{0}$. The term $\tilde{\boldsymbol{\beta}}$ can be written as:

$$\tilde{\boldsymbol{\beta}} = \mathbf{CY} = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})\mathbf{Y} = \hat{\boldsymbol{\beta}} + \mathbf{DY},$$

and so it follows that the variance/covariance matrix of $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is given by:

$$\begin{aligned}
Var(\tilde{\beta} - \beta) &= Var(\tilde{\beta} - \beta + \hat{\beta} - \hat{\beta}) \\
&= Var((\hat{\beta} - \beta) + (\tilde{\beta} - \hat{\beta})) \\
&= Var((\hat{\beta} - \beta)) + Var((\tilde{\beta} - \hat{\beta})), \text{ by (2.8)} \\
&= Var(\hat{\beta}) + Var(\mathbf{D}\mathbf{Y}) \\
&= Var(\hat{\beta}) + Var(\mathbf{D}\boldsymbol{\varepsilon}).
\end{aligned}$$

Finally, we have that the difference between the variance/covariance matrix of $\hat{\beta}$ and $\tilde{\beta}$ can be expressed as:

$$\begin{aligned}
Var(\tilde{\beta}) - Var(\hat{\beta}) &= Var(\mathbf{D}\boldsymbol{\varepsilon}) = \mathbf{D}Var(\boldsymbol{\varepsilon})\mathbf{D}' \\
&= \sigma^2\mathbf{D}\mathbf{D}',
\end{aligned}$$

which is a variance/covariance matrix, and implies that it is a positive semi-definite matrix. Therefore, $Var(\tilde{\beta}) \geq Var(\hat{\beta})$, and the ordinary least squares estimator $\hat{\beta}$ for the classical model (2.2) is the minimum variance linear unbiased estimator for β in the class of linear unbiased estimators. $\square$

2.4.4 Estimator for $Var(\hat{\beta})$

Before we are able to tests hypotheses around the regression coefficients an estimator for the variance of the ordinary least squares estimator $\hat{\beta}$ is required. Thus we would like to obtain an estimator for $Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Since σ^2 is the only unknown and is the expected value of ε_i^2 , and e_i serves as an estimate of ε_i , it may seem natural to estimate σ^2 by:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2.$$

However it is not that simple: consider the residuals vector,

$$\begin{aligned}
\mathbf{e} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\
&= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\
&= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\varepsilon} \\
&= (\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon},
\end{aligned}$$

where $\mathbf{I}_n$ denotes $(n \times n)$ identity matrices, and $\mathbf{H}$ is the idempotent hat/projection matrix. The sum of squared residuals now become,

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{H})'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}, \quad (2.9)$$

and has expectation,

$$\begin{aligned}
E(\mathbf{e}'\mathbf{e}) &= E(\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}) \\
&= \text{tr}E(\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}) = E(\text{tr}\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}) \\
&= E(\text{tr}(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \text{tr}(\mathbf{I}_n - \mathbf{H})E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\
&= \sigma^2 \text{tr}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \sigma^2 [\text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')] \\
&= \sigma^2 [\text{tr}(\mathbf{I}_n) - \text{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X})] \\
&= \sigma^2 [\text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{I}_k)] \\
&= \sigma^2(n - k).
\end{aligned}$$

Thus the unbiased estimator for σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n - k} \sum_{i=1}^n e_i^2 = \frac{\mathbf{e}'\mathbf{e}}{n - k}.$$

Finally the estimator for $Var(\hat{\beta})$ is given by:

$$\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1},$$

where the j^{th} diagonal element is the variance of the j^{th} estimated regression coefficient $\hat{\beta}_j$, denoted by $\widehat{Var}(\hat{\beta}_j) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})_{jj}^{-1}$.

2.4.5 Hypothesis testing

For the first time the normality assumption of the error terms will be invoked and is necessary for deriving the distribution of test statistics which will be used for inference concerning the regression coefficients. The normality of $\hat{\beta}$ follows from the fact that these estimators are a linear function of the normally distributed errors, see Equation (2.4). Throughout this text we would like to test hypotheses of the form:

$$H_0 : \beta_j = \beta_{j0} \quad \text{vs.} \quad H_A : \beta_j \neq \beta_{j0}, \quad j = 1, 2, \dots, k-1,$$

where β_{j0} is the hypothesised value of the regression coefficient for β_j , but we typically would choose $\beta_{j0} = 0$. This hypothesis can be tested using the pivotal test statistic,

$$z_j = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})_{jj}^{-1}}} \sim N(0, 1), \quad (2.10)$$

which will have a normal distribution with mean zero and variance one under the null hypothesis, having satisfied all the assumptions stipulated for the classical linear model (2.2). Note that $(\mathbf{X}'\mathbf{X})_{jj}^{-1}$ denotes the j^{th} diagonal element of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. Since σ^2 will rarely be known, it needs to be estimated using $\hat{\sigma}^2$. The test statistic then becomes:

$$t_j = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})_{jj}^{-1}}} \sim t_{n-k} \quad (2.11)$$

which has a t distribution with $n - k$ degrees of freedom under the null hypothesis. To derive the distribution of the test statistic (2.11) first consider,

$$\frac{(n - k)\hat{\sigma}^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}(n - k)}{\sigma^2(n - k)} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)'(\mathbf{I}_n - \mathbf{H})\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \chi_{n-k}^2. \quad (2.12)$$

The quantity in (2.12) is an idempotent quadratic form in a standard normal vector $(\boldsymbol{\varepsilon}/\sigma)$, and since $\boldsymbol{\varepsilon}$ is normally distributed this quantity follows a chi-squared distribution with $\text{rank}(\mathbf{I}_n - \mathbf{H}) = \text{tr}(\mathbf{I}_n - \mathbf{H}) = n - k$ degrees of freedom (see Theorem B.8 of Greene, 2003). Therefore,

$$t_j = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})_{jj}^{-1}}} = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})_{jj}^{-1}}} \bigg/ \sqrt{\frac{(n - k)\hat{\sigma}^2}{\sigma^2} / (n - k)}.$$

The quantity t_j thus follows a t_{n-k} distribution under the null hypothesis because, from (2.10), the numerator follows a standard normal distribution and, from (2.12), the denominator is the square root of a chi-squared random variable divided by its own degrees of freedom.

2.5 Large sample properties of the ordinary least squares estimator

In the finite sample properties section we derived the exact mean, variance and distributions for test statistics of the ordinary least squares estimator under the assumption of normally distributed errors and independent observations. However, the classical linear regression model in Equation (2.2) with normally distributed errors and independent observations is a special case. This section will focus on relaxing these two assumptions and deriving the asymptotic properties of the ordinary least squares estimator when the errors are not necessarily normally distributed. To begin, we shall show that the OLS estimator is consistent and then derive its asymptotic distribution.

2.5.1 $\hat{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\beta}$

We shall show that the OLS estimator is consistent for $\boldsymbol{\beta}$ by using only the first four assumptions in Section 2.2.1.

Proof. We find that by rewriting the OLS estimator as:

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
&= \beta + \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \left(\frac{\mathbf{X}'\varepsilon}{n}\right),
\end{aligned}$$

and then introducing the probability limit ³ we obtain the following result:

$$\text{plim}_{n \rightarrow \infty} \hat{\beta} = \beta + \text{plim}_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \text{plim}_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\varepsilon}{n}\right). \quad (2.13)$$

To proceed we are required to make the following assumption:

Assumption 7: Consistency of $\mathbf{X}'\mathbf{X}/n$

$$\text{plim}_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) = \mathbf{Q},$$

where $\mathbf{Q}$ is a finite, deterministic, positive definite matrix.

Theorem: Slutsky's theorem

Let $\mathbf{X}_n$, $\mathbf{X}$ and $\mathbf{Y}_n$ be random vectors and $\mathbf{c}$ a vector of constants. If $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{c}$ and $g(\mathbf{X}_n)$ is a continuous function, then:

1.

$$\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{d} \mathbf{X} + \mathbf{c}$$

$$\mathbf{X}_n \mathbf{Y}_n \xrightarrow{d} \mathbf{X} \mathbf{c}$$

$$\mathbf{X}_n \mathbf{Y}_n^{-1} \xrightarrow{d} \mathbf{X} \mathbf{c}^{-1},$$

³The notation plim is used to denote the *probability limit*. That is, if $\{A_n\}$ denotes a sequence of random variables, and A denotes a random variable, then for every $\epsilon > 0$, we have that A_n *converges in probability* to A if $\text{plim}_{n \rightarrow \infty} A_n = A$ or, equivalently, $\lim_{n \rightarrow \infty} P(|A_n - A| \geq \epsilon) = 0$.

provided $\mathbf{c}$ is invertible. It is also true for $\mathbf{Y}_n$ and $\mathbf{c}$ constants and matrices.

2.

$$g(\mathbf{X}_n) \xrightarrow{d} g(\mathbf{X})$$

Now, if $\mathbf{Q}^{-1}$ exists by Slutsky's theorem, i.e. the inverse of a nonsingular matrix is a continuous function of the elements of the matrix, then Equation (2.13) becomes,

$$\text{plim}_{n \rightarrow \infty} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{Q}^{-1} \text{plim}_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} \right).$$

If we could prove that $\text{plim}_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n} \right) = \mathbf{0}$, then we can conclude consistency of the estimator. The proof of this result involves recalling the exogeneity assumption and by using the law of iterative expectations to obtain:

$$\begin{aligned} E \left[\frac{1}{n} \mathbf{X}'\boldsymbol{\varepsilon} \right] &= \frac{1}{n} E[E[\mathbf{X}'\boldsymbol{\varepsilon}|\mathbf{X}]] \\ &= \frac{1}{n} E[\mathbf{X}'E[\boldsymbol{\varepsilon}|\mathbf{X}]] = \frac{1}{n} \cdot \mathbf{0} = \mathbf{0}. \end{aligned}$$

Now, the variance of the quantity $\frac{\mathbf{X}'\boldsymbol{\varepsilon}}{n}$ is:

$$\begin{aligned} \text{Var} \left[\frac{1}{n} \mathbf{X}'\boldsymbol{\varepsilon} \right] &= \text{Var} \left[E \left[\frac{1}{n} \mathbf{X}'\boldsymbol{\varepsilon} | \mathbf{X} \right] \right] + E \left[\text{Var} \left[\frac{1}{n} \mathbf{X}'\boldsymbol{\varepsilon} | \mathbf{X} \right] \right] \\ &= \text{Var}[\mathbf{0}] + E \left[E \left[\frac{1}{n^2} \mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X} | \mathbf{X} \right] \right] \\ &= E \left[\frac{1}{n^2} \mathbf{X}' E \left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X} \right] \mathbf{X} \right] \\ &= E \left[\frac{1}{n^2} \mathbf{X}' \sigma^2 \mathbf{I}_n \mathbf{X} \right] \\ &= \frac{\sigma^2}{n} E \left[\frac{\mathbf{X}'\mathbf{X}}{n} \right]. \end{aligned}$$

Now, using Assumption 7:

$$\begin{aligned} \lim_{n \rightarrow \infty} \text{Var} \left[\frac{1}{n} \mathbf{X}'\boldsymbol{\varepsilon} \right] &= \lim_{n \rightarrow \infty} \frac{\sigma^2}{n} E \left[\frac{\mathbf{X}'\mathbf{X}}{n} \right] \\ &= 0 \cdot \mathbf{Q} = \mathbf{0}. \end{aligned}$$

Since the mean is equal to zero and the variance converges to zero, the quantity $\frac{1}{n} \mathbf{X}'\boldsymbol{\varepsilon}$ converges

in mean square to zero, which in turn implies convergence in probability to zero. This result concludes the proof for consistency of the OLS estimator, i.e.,

$$\begin{aligned}\text{plim}_{n \rightarrow \infty} \hat{\beta} &= \beta + \text{plim}_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \text{plim}_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\varepsilon}{n} \right) \\ &= \beta + \mathbf{Q}^{-1} \cdot \mathbf{0} = \beta.\end{aligned}$$

□

Thus the ordinary least squares estimator under only the first four assumptions combined with Assumption 7 provides a consistent estimator for β in the classical linear regression model (2.2).

2.5.2 Asymptotic distribution of the ordinary least squares estimator

Next we would like to derive the asymptotic distribution of $\hat{\beta}$.

Proof. The asymptotic distribution of $\hat{\beta}$ can be obtained by first considering the limiting distribution of:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left(\frac{\mathbf{X}'\varepsilon}{\sqrt{n}} \right). \quad (2.14)$$

If Assumption 7 is valid and the limiting distribution of the random vector ε exists, then the limiting distribution on the right hand side of Equation (2.14) is equivalent to the distribution of:

$$\left[\text{plim}_{n \rightarrow \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \right] \left(\frac{\mathbf{X}'\varepsilon}{\sqrt{n}} \right) = \mathbf{Q}^{-1} \left(\frac{\mathbf{X}'\varepsilon}{\sqrt{n}} \right).$$

Thus to find the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta)$, we now only need to find the limiting distribution of:

$$\left(\frac{\mathbf{X}'\varepsilon}{\sqrt{n}} \right) = \sqrt{n} \left(\frac{1}{n} \mathbf{X}'\varepsilon - \frac{1}{n} E[\mathbf{X}'\varepsilon] \right), \quad (2.15)$$

where $E[\mathbf{X}'\varepsilon] = \mathbf{0}$.

Theorem: Multivariate Lindberg-Feller central limit theorem

Suppose that $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are a sample of random vectors such that $E(\mathbf{X}_i) = \boldsymbol{\mu}_i$, $Var(\mathbf{X}_i) = \mathbf{Q}_i$, and all mixed third moments of the multivariate distribution are finite.

Let

$$\bar{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\mu}_i,$$

and

$$\bar{\mathbf{Q}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{Q}_i.$$

If we assume that

$$\lim_{n \rightarrow \infty} \bar{\mathbf{Q}}_n = \mathbf{Q},$$

where $\mathbf{Q}$ is a finite, positive definite matrix, and that for all i ,

$$\lim_{n \rightarrow \infty} (n \bar{\mathbf{Q}}_n)^{-1} \mathbf{Q}_i = \lim_{n \rightarrow \infty} \left(\sum_{i=1}^n \mathbf{Q}_i \right)^{-1} \mathbf{Q}_i,$$

then

$$\sqrt{n} (\bar{\mathbf{X}}_n - \bar{\boldsymbol{\mu}}_n) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q}),$$

where $\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$.

In order to obtain a limiting distribution for Equation (2.15) the multivariate Lindberg-Feller central limit theorem will be employed. Since $\frac{1}{n} \mathbf{X}' \boldsymbol{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \varepsilon_i$ is an average of n independent random vectors $\mathbf{X}_i \varepsilon_i$, with mean vector $\mathbf{0}$ and variance/covariance matrix,

$$Var[\mathbf{X}_i \varepsilon_i] = E[\mathbf{X}_i \varepsilon_i \varepsilon_i' \mathbf{X}_i'] = \sigma^2 E[\mathbf{X}_i \mathbf{X}_i'] = \sigma^2 \mathbf{Q}_i, \quad i = 1, 2, \dots, n,$$

where $\mathbf{Q}_i = E(\mathbf{X}_i \mathbf{X}_i')$ and $\mathbf{X}_i$ is the i^{th} row vector in the design matrix $\mathbf{X}$. The variance/covariance matrix of $\frac{1}{\sqrt{n}} \mathbf{X}' \boldsymbol{\varepsilon}$ will then be given by:

$$Var \left[\frac{1}{\sqrt{n}} \mathbf{X}' \boldsymbol{\varepsilon} \right] = \sigma^2 \frac{1}{n} [\mathbf{Q}_1 + \mathbf{Q}_2 + \dots + \mathbf{Q}_n] = \sigma^2 \bar{\mathbf{Q}}_n.$$

If Assumption 7 holds, i.e., the predictors are well behaved and the sum is not dominated by

any particular term, then:

$$\lim_{n \rightarrow \infty} \bar{\mathbf{Q}}_n = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E(\mathbf{X}_i \mathbf{X}_i') = \lim_{n \rightarrow \infty} E\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) = \mathbf{Q},$$

and we obtain:

$$\lim_{n \rightarrow \infty} \sigma^2 \bar{\mathbf{Q}}_n = \sigma^2 \mathbf{Q}.$$

The multivariate Lindberg-Feller central limit may be applied to the vectors $\mathbf{X}_i \varepsilon_i$ since they are independent, have mean vectors $\mathbf{0}$ and variance/covariance matrices $\sigma^2 \mathbf{Q}_i < \infty$, and thus converges in distribution to a multivariate normal distribution:

$$\left(\frac{\mathbf{X}'\varepsilon}{\sqrt{n}}\right) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}),$$

then multiplying by $\mathbf{Q}^{-1}$ to obtain:

$$\mathbf{Q}^{-1} \left(\frac{\mathbf{X}'\varepsilon}{\sqrt{n}}\right) \xrightarrow{d} N(\mathbf{Q}^{-1}\mathbf{0}, \mathbf{Q}^{-1}\sigma^2 \mathbf{Q} \mathbf{Q}^{-1})$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}).$$

□

The asymptotic distribution of $\hat{\boldsymbol{\beta}}$ with independent observations is given by:

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \frac{\sigma^2}{n} \mathbf{Q}^{-1}\right).$$

In practice one would estimate $\frac{1}{n} \mathbf{Q}^{-1}$ with $(\mathbf{X}'\mathbf{X})^{-1}$, and σ^2 with $\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k}$. If the errors ε are normally distributed as in the classical linear regression model (2.2), then the exact distribution of $\hat{\boldsymbol{\beta}}$ is $N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$ for every sample, and holds asymptotically as well. The results that we have proved in this section only requires that the predictors are well behaved (as stated in Assumption 7) and that the observations are independent for the OLS estimator to have an asymptotic normal distribution having applied the multivariate Lindberg-Feller central limit theorem.

2.6 Generalised linear model

We now consider relaxing the homoskedasticity assumption (Assumption 4 in Section 2.2.1), resulting in errors that may or may not be heteroskedastic and/or correlated with each other (autocorrelation). The generalised linear model with $k - 1$ predictor variables allows for these different occurrences and is defined as:

$$\underset{(n \times 1)}{\mathbf{Y}} = \underset{(n \times k)}{\mathbf{X}} \underset{(k \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}, \quad (2.16)$$

where

$\mathbf{Y}$ is a random response vector,

$\mathbf{X}$ is a matrix with columns corresponding to the $k - 1$ predictor variable vectors, and typically the first column vector is a vector of ones to accommodate the intercept term,

$\boldsymbol{\beta}$ is a vector of unknown regression coefficients, assumed to be nonrandom,

$\boldsymbol{\varepsilon}$ is a random error term vector, with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\boldsymbol{\Omega} = \boldsymbol{\Sigma}$, and is a normally distributed error vector, i.e., $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{\Omega})$. $\boldsymbol{\Omega}$ is assumed to be a real symmetric positive definite matrix.

The general nature of the generalised linear regression model allows for variance/covariance structures such as:

Heteroskedastic and independent As with the classical linear model's covariance structure, this next covariance matrix has uncorrelated error terms, but the variances are now allowed to differ:

$$\sigma^2\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_{11} & 0 & \cdots & 0 \\ 0 & \omega_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

Here the variances of the individual errors ε_i are not necessarily all equal to one another, but their covariances are zero, i.e., $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2\omega_{ii}$ and $Cov(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j; i, j = 1, 2, \dots, n$.

Homoskedastic and dependent–autocorrelation An example where the error terms are both homoskedastic and dependent is when there is an autocorrelation structure in the covariance matrix such as in the one defined below:

$$\sigma^2 \mathbf{\Omega} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}.$$

Here we find that the individual errors ε_i are all the same, but their covariances differ, i.e., $Var(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \rho_{|i-j|} \forall i \neq j; i, j = 1, 2, \dots, n$.

Heteroskedastic and dependent–unstructured covariances Finally, the case where the covariances and variances of the error terms are allowed to differ is also accommodated by this model:

$$\sigma^2 \mathbf{\Omega} = \sigma^2 \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{bmatrix}.$$

Here we find that the variance of the individual errors ε_i differ, but the covariances also differ, i.e., $Var(\varepsilon_i) = \sigma_{ii}^2$ and $Cov(\varepsilon_i, \varepsilon_j) = \sigma_{ij}^2 \forall i \neq j; i, j = 1, 2, \dots, n$.

If the regression coefficients are to be estimated efficiently by means of OLS, the error terms must be uncorrelated and have the exact same variance. These assumptions were also required to prove the Gauss-Markov theorem in Section 2.4.3.

As stated in the Introduction, one of the aims of this text is to investigate the theoretical implications heteroskedasticity has on the regression coefficient estimators. In the previous sections we proved that the OLS estimator $\hat{\beta}$, is unbiased, consistent, has minimum variance among all other unbiased estimators (efficient), is normally distributed if the errors are normally distributed, and

is asymptotically normally distributed. We therefore know that the OLS estimator is best in its class when the assumptions hold, but what happens if they are not satisfied as is the case in the generalised linear model in Equation (2.16)? The properties that do hold include the following:

- The OLS estimator remains unbiased if the exogeneity assumption from the classical model is maintained.
- The OLS estimator remains consistent if the matrices $\frac{\mathbf{X}'\mathbf{X}}{n}$ and $\frac{\mathbf{X}'\mathbf{\Omega}\mathbf{X}}{n}$ are finite positive definite matrices as $n \rightarrow \infty$.
- The OLS estimator retains its normal distribution,

$$\hat{\boldsymbol{\beta}} \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}),$$

if the errors are normally distributed, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{\Omega})$.

- The OLS estimator retains its asymptotic normal distribution,

$$\hat{\boldsymbol{\beta}} \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}),$$

the proof in Section 2.5.2 was general enough to encompass the heteroskedastic error case.

However, it turns out that when the homoskedasticity assumption is violated, the OLS estimator of the variance/covariance of the regression coefficient estimates will either be too large or too small. Unfortunately no clear rules exist on which of these two situations might occur in practice (Hayes and Cai, 2007). The solution to this particular problem lies in using the *Generalised Least Squares* (GLS) estimator to obtain efficient estimates and is achieved by transforming the regression model (2.16) in order to satisfy the Gauss Markov theorem — this is done by ensuring that $\sigma^2\mathbf{\Omega}$ in the generalised model reduces to $\sigma^2\mathbf{I}_n$ after transforming it. The OLS theory developed for the classical linear model can then be applied to the transformed model.

2.7 Generalised least squares (GLS)

In the classic linear regression model the OLS procedure was sufficient for estimating the regression coefficients because the assumptions of homoskedasticity and independence were satisfied, however the generalised linear model does not satisfy these conditions, so now a generalised least squares procedure will be introduced. The estimator will be derived and the properties thereof discussed.

To begin the discussion on the derivation of the GLS estimator we note that the variance/covariance matrix $\mathbf{\Omega}$ of the errors was defined to be a known real symmetric positive definite matrix in the generalised linear model (2.16), and so it may be decomposed using Cholesky's decomposition (see, for example, Davidson and MacKinnon, 2004), Crout's decomposition (see, for example, Davidson and MacKinnon, 2004) or singular value decomposition (see, for example, Greene, 2003; Johnson and Wichern, 2007). The decomposition allows one to determine a transformation that, after transformation, the response has the property that the variance can be expressed as $\sigma^2\mathbf{\Omega} = \sigma^2\mathbf{I}_n$; this is done so that one may exploit the efficiency of OLS estimators (as stated in the Gauss Markov theorem) for the GLS estimator. The singular value method is discussed here, but the theoretical results are also valid for Cholesky's and Crout's decomposition.

2.7.1 Singular value decomposition and the generalised least squares estimator

The singular value decomposition (see, for example, Greene, 2003; Johnson and Wichern, 2007) of some symmetric positive definite matrix, say, $\mathbf{\Omega}$ is given by:

$$\mathbf{\Omega} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}',$$

where $\mathbf{C}$ is an orthonormal matrix⁴ of eigenvalues (columns) of $\mathbf{\Omega}$, and $\mathbf{\Lambda}$ is a diagonal matrix containing eigenvalues, λ_i , of the matrix $\mathbf{\Omega}$. Note that $\mathbf{\Lambda}^{\frac{1}{2}}$ will denote a diagonal matrix with diagonal elements $\sqrt{\lambda_i}$ and $\mathbf{\Lambda}^{-\frac{1}{2}}$ will be a diagonal matrix with diagonal elements $\sqrt{\frac{1}{\lambda_i}}$.

⁴Note that an orthonormal matrix $\mathbf{C}$ has the property: $\mathbf{C}^{-1} = \mathbf{C}'$.

The matrix $\mathbf{\Omega}$ can now be written as $\mathbf{\Omega} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}' = \mathbf{C}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{C}'$, and then it follows that

$$\mathbf{\Omega}^{-1} = (\mathbf{C}\mathbf{\Lambda}\mathbf{C}')^{-1} = \mathbf{C}\mathbf{\Lambda}^{-1}\mathbf{C}' = \mathbf{C}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{C}' = \mathbf{C}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda}^{-\frac{1}{2}'}\mathbf{C}' = \mathbf{P}'\mathbf{P},$$

where $\mathbf{P}' := \mathbf{C}\mathbf{\Lambda}^{-\frac{1}{2}}$. By premultiplying the generalised linear model (2.16) by $\mathbf{P}$, we obtain a linear regression model that satisfies the conditions required to satisfy the Gauss Markov theorem,

$$\mathbf{P}\mathbf{Y} = \mathbf{P}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} = \mathbf{Y}_* = \mathbf{X}_*\boldsymbol{\beta} + \boldsymbol{\varepsilon}_*, \quad (2.17)$$

where $\mathbf{Y}_* = \mathbf{P}\mathbf{Y}$, $\mathbf{X}_* = \mathbf{P}\mathbf{X}$, and $\boldsymbol{\varepsilon}_* = \mathbf{P}\boldsymbol{\varepsilon}$. This model has the following properties:

- The expected value of the error terms is zero:

$$E(\boldsymbol{\varepsilon}_*) = E(\mathbf{P}\boldsymbol{\varepsilon}) = \mathbf{P}E(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

- The variance/covariance of the error terms has the form $\sigma^2\mathbf{I}_n$:

$$\begin{aligned} Var(\boldsymbol{\varepsilon}_*) &= E(\boldsymbol{\varepsilon}_*\boldsymbol{\varepsilon}_*') = E(\mathbf{P}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{P}') = \mathbf{P}E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{P}' \\ &= \mathbf{P}\sigma^2\mathbf{\Omega}\mathbf{P}' \\ &= \sigma^2\mathbf{P}\mathbf{C}\mathbf{\Lambda}\mathbf{C}'\mathbf{P}' \\ &= \sigma^2\mathbf{\Lambda}^{-\frac{1}{2}'}\mathbf{C}'\mathbf{C}\mathbf{\Lambda}\mathbf{C}'\mathbf{C}\mathbf{\Lambda}^{-\frac{1}{2}} \\ &= \sigma^2\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{-\frac{1}{2}} = \sigma^2\mathbf{I}_n. \end{aligned}$$

Since the premultiplied model in Equation (2.17) now has the same properties as the classical model in Equation (2.2), the OLS estimator for $\boldsymbol{\beta}$ may be used to determine an expression for the generalised least squares estimator $\hat{\boldsymbol{\beta}}_{GLS}$. By substituting the transformed model into the OLS estimator $\hat{\boldsymbol{\beta}}$ we find the expression for $\hat{\boldsymbol{\beta}}_{GLS}$, i.e.,

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{Y}_* = (\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{Y} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{Y}. \quad (2.18)$$

This estimator is known as the generalised least squares estimator or Aitken's (1935) estimator for β .

As mentioned earlier, the results in this section could have been derived using either Cholesky's decomposition or Crout's decomposition (Davidson and MacKinnon, 2004), since both methods decompose the inverse of the symmetric positive definite matrix Ω into the product of component matrices, i.e., $\Omega^{-1} = \Psi'\Psi$, where Ψ is defined to be a triangular matrix for these two decompositions. The classical linear model is then premultiplied to yield $\Psi'Y = \Psi'X\beta + \Psi'\varepsilon$ or $Y_* = X_*\beta + \varepsilon_*$, and the same theoretical results follow using different mathematical arguments.

Next we turn our attention towards the finite sample and asymptotic properties of the GLS estimator (2.18) and show that it has the same properties as the OLS estimator.

2.7.2 Properties of the generalised least squares estimator

The GLS estimator $\hat{\beta}_{GLS}$ shares a number of good properties with the OLS estimator $\hat{\beta}$. In the sections below we shall highlight the similarities between these estimators in terms of their properties and assumptions required.

Expected value of the generalised least squares estimator

In order to prove that the GLS estimator $\hat{\beta}_{GLS}$ is unbiased we require the assumption of exogeneity for the transformed predictors, i.e., $E(\varepsilon_*|X_*) = \mathbf{0}$. This assumption is similar to the one we made for the classical linear model, except now using the transformed design matrix and errors.

$$\begin{aligned}
E(\hat{\beta}_{GLS}|X_*) &= E[(X_*'X_*)^{-1}X_*'Y_*|X_*] \\
&= E[(X_*'X_*)^{-1}X_*'(X_*\beta + \varepsilon_*)|X_*] \\
&= E[(X_*'X_*)^{-1}(X_*'X_*)\beta + (X_*'X_*)^{-1}X_*'\varepsilon_*|X_*] \\
&= \beta + E[(X_*'X_*)^{-1}X_*'\varepsilon_*|X_*] \\
&= \beta + (X_*'X_*)^{-1}X_*'E[\varepsilon_*|X_*] = \beta.
\end{aligned}$$

By making the exogeneity assumption for the transformed model, i.e., the transformed predictors and errors are uncorrelated, we notice that the GLS estimator remains unbiased and is not affected by the violation of the homoskedasticity assumption.

Estimator for $Var(\hat{\beta}_{GLS})$

To find an expression for the variance/covariance matrix of the GLS estimator we substitute the transformed model into the expression for the OLS variance/covariance matrix defined in Equation (2.6):

$$Var(\hat{\beta}_{GLS}|\mathbf{X}_*) = \sigma^2(\mathbf{X}'_*\mathbf{X}_*)^{-1} = \sigma^2(\mathbf{X}'\mathbf{P}'\mathbf{P}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}.$$

The only difference between the GLS and OLS variance/covariance expression is $\mathbf{\Omega}^{-1}$ between the two design matrices. Analogous to the OLS case, the unbiased estimator for σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^n e_{*i}^2 = \frac{\mathbf{e}'_*\mathbf{e}_*}{n-k}.$$

The estimator for $Var(\hat{\beta}_{GLS})$ is then given by:

$$\widehat{Var}(\hat{\beta}_{GLS}) = \hat{\sigma}^2(\mathbf{X}'_*\mathbf{X}_*)^{-1} = \hat{\sigma}^2(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1},$$

where the j^{th} diagonal element is the variance of the j^{th} estimated regression coefficient. Note that the quantity $\mathbf{\Omega}^{-1}$ in the expression above is assumed to be known and that we shall be able to calculate $\widehat{Var}(\hat{\beta}_{GLS})$ using sample data. However, $\mathbf{\Omega}^{-1}$ will rarely be known and so the estimation of this quantity will need to be addressed. A brief discussion of how to estimate this quantity is presented in Section 2.8.2.

Gauss Markov theorem

The GLS estimator $\hat{\beta}_{GLS}$ is the minimum variance linear unbiased estimator of β in the generalised linear regression model (2.16); this result is obtained by direct application of the Gauss Markov theorem.

Consistency

The GLS estimator is consistent if we assume $\mathbf{X}_*$ and $\boldsymbol{\varepsilon}_*$ to be uncorrelated, and if $\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}'_* \mathbf{X}_* = \text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X} = \mathbf{Q}_*$, where $\mathbf{Q}_*$ is a positive definite matrix. The proof of this result is analogous to the one stated for the OLS case in Section 2.5.1.

2.8 Heteroskedasticity and remedial measures

One of the main assumptions in the classical linear model (2.2) is that of homoskedasticity, i.e., $\text{Var}(\varepsilon_i) = \sigma^2 \forall i = 1, 2, \dots, n$, which is in contrast with heteroskedasticity, when the variances differ across observation, i.e., $\text{Var}(\varepsilon_i) = \sigma_i^2, i = 1, 2, \dots, n$. Heteroskedasticity, which commonly occurs in cross-sectional and time-series data (Greene, 2003), does not affect estimation of the regression coefficients, but it does influence the estimation of the variance/covariance matrix of the regression coefficient estimates. The ultimate result is that inference related to the regression coefficients suffers.

Throughout the rest of the dissertation we shall consider techniques for remedying the problems associated with inference for regression coefficients in the presence of heteroskedasticity. These techniques rely on suitable estimates of the variance/covariance matrix of the estimated regression coefficients. To facilitate exposition, the following generalised linear model with heteroskedastic and independent errors will be considered:

$$\underset{(n \times 1)}{\mathbf{Y}} = \underset{(n \times k)}{\mathbf{X}} \underset{(k \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}, \quad (2.19)$$

where the errors $\boldsymbol{\varepsilon}$ have the properties $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{\Omega} = \boldsymbol{\Sigma}$ defined as:

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

2.8.1 Weighted least squares (WLS)

In the previous sections we have shown that the OLS estimator for the generalised linear model remains unbiased and consistent, however the Gauss Markov theorem is no longer satisfied and as a result will no longer have minimum variance. The estimator for the variance of the GLS estimator contains n variance components, for each of the n observations. To ensure that the estimator has minimum variance, we would need to assign more weight to observation that have a smaller variance component and less weight to observations with higher variances. The rationale is that observation with less variance provide more reliable information about the regression function and vice versa (Kutner et al., 2005). Weighted least squares builds on the generalised linear model by assuming that $\mathbf{\Omega}^{-1} = \mathbf{W}^{-1}$ contained inside the GLS estimator (2.18) is a diagonal matrix, where the i^{th} diagonal element is $\frac{1}{\omega_i}$, i.e. $\text{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2 \omega_i$. Let $\mathbf{W}$ be defined to be the diagonal matrix,

$$\mathbf{W} = \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} \text{ and its inverse } \mathbf{W}^{-1} = \begin{bmatrix} \frac{1}{\omega_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\omega_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\omega_n} \end{bmatrix}.$$

Premultiply the generalised linear model (2.16) by $\mathbf{P} = \mathbf{W}^{-\frac{1}{2}}$ to obtain:

$$\mathbf{PY} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{P}\boldsymbol{\varepsilon} = \mathbf{Y}_w = \mathbf{X}_w\boldsymbol{\beta} + \boldsymbol{\varepsilon}_w. \quad (2.20)$$

Now use OLS estimation on the transformed model (2.20), to obtain the weighted least squares estimator:

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'_w \mathbf{X}_w)^{-1} \mathbf{X}'_w \mathbf{Y}_w = (\mathbf{X}' \mathbf{P}' \mathbf{P} \mathbf{X})^{-1} \mathbf{X}' \mathbf{P}' \mathbf{P} \mathbf{Y} = (\mathbf{X}' \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{-1} \mathbf{Y}.$$

Weighted least squares is naive in the sense that the weights are assumed to be known, which will rarely if ever be the case, but it illustrates a useful means to correct for heteroskedasticity. Next we shall look at a few cases where the known weights assumption will be relaxed to correct for heteroskedasticity.

2.8.2 Feasible generalised least squares (FGLS)

In order to use generalised least squares and weighted least squares, the $\mathbf{\Omega}$ in the model given by Equation (2.19) needs to be known — this is however rarely, if ever, the case. Feasible least squares addresses this issue by making the assumption that $\mathbf{\Omega}$ depends in a known way on a vector of unknown parameters $\boldsymbol{\theta}$, i.e., $\mathbf{\Omega} = \mathbf{\Omega}(\boldsymbol{\theta})$. Then, if $\boldsymbol{\theta}$ can be consistently estimated by $\hat{\boldsymbol{\theta}}$, the estimator for the variance/covariance matrix is given by $\hat{\mathbf{\Omega}} = \mathbf{\Omega}(\hat{\boldsymbol{\theta}})$. This estimator can be shown to be asymptotically efficient if we assume that a consistent estimator for $\boldsymbol{\theta}$ (but not necessarily an efficient one) is used (Greene, 2003). However the efficiency of $\hat{\mathbf{\Omega}}$ may not be present when used with small samples and the OLS variance/covariance matrix could provide a better results if the heteroskedasticity is not too severe.

Feasible generalised least squares may be used to find a solution for the weighted least squares approach by modelling the weights ω_i in the variance covariance matrix $\mathbf{\Omega} = \mathbf{W}$. The steps required for calculating this feasible weighted least squares estimate (Davidson and MacKinnon, 2004) is summarised:

1. Based on the parametrisation of $\mathbf{\Omega}(\boldsymbol{\theta})$ define a consistent estimator for $\boldsymbol{\theta}$.
2. Calculate $\hat{\mathbf{\Omega}} = \mathbf{\Omega}(\hat{\boldsymbol{\theta}}, \mathbf{X})$.
3. Decompose the matrix $\hat{\mathbf{\Omega}}$ into the form $\hat{\mathbf{\Omega}}^{-1} = \hat{\mathbf{P}}\hat{\mathbf{P}}'$.
4. Transform the model using $\hat{\mathbf{P}}\mathbf{Y} = \hat{\mathbf{P}}\mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{P}}\boldsymbol{\varepsilon}$.
5. Calculate OLS estimates using the transformed model to obtain the feasible generalised least squares estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{Y}.$$

Feasible generalised least squares may be used to find a solution for the weighted least squares approach by modelling the weights ω_i in the variance/covariance matrix $\mathbf{\Omega} = \mathbf{W}$. The steps required for calculating this feasible weighted least squares estimate (Davidson and MacKinnon, 2004) is summarised by considering the following example of a linear regression model $Y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$, where the variance of the errors have the following functional form $E(\varepsilon_i^2) = e^{\mathbf{Z}_i\boldsymbol{\theta}}$. The

regression coefficient vector β still has dimension k and the unknown parameter vector θ has dimension l , where $l \leq k$. The design matrices $\mathbf{X}$ and $\mathbf{Z}$ both have the same row dimension n , but $\mathbf{Z}$ is comprised of a subset or all of the predictors contained in $\mathbf{X}$ to produce a matrix of dimension $(n \times l)$.

1. Based on the parametrisation of $\Omega(\theta) = \mathbf{W}(\theta)$ define a consistent estimator $\widehat{\mathbf{W}} = \mathbf{W}(\widehat{\theta}, \mathbf{Z})$, i.e., $\widehat{\omega}_i = e^{\mathbf{Z}_i \widehat{\theta}}$, where $\log(e_i^2) = \mathbf{Z}_i \theta + v_i$ is used to find $\widehat{\theta}$ by means of OLS.
2. Transform the model into the form $\widehat{\mathbf{P}}\mathbf{Y} = \widehat{\mathbf{P}}\mathbf{X}\beta + \widehat{\mathbf{P}}\varepsilon$, where $\widehat{\mathbf{P}} = \widehat{\mathbf{W}}^{-\frac{1}{2}}$ is a square matrix with $\frac{1}{\sqrt{\widehat{\omega}_i}}$ on the diagonal and zeros on the off diagonal.
3. Calculate OLS estimates of β using the transformed model in the previous step to obtain the feasible weighted least squares regression coefficient estimates,

$$\widehat{\beta}_{FWLS} = (\mathbf{X}'\widehat{\mathbf{W}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{-1}\mathbf{Y}.$$

It can be shown that, under some suitable regularity conditions pertaining to the predictor variables and the errors, the feasible generalised least squares estimator $\widehat{\beta}_{FGLS}$ is consistent and is asymptotically equivalent to the generalised least squares estimator $\widehat{\beta}_{GLS}$ (Davidson and MacKinnon, 2004; Amemiya, 1973).

2.8.3 Heteroskedastic consistent covariance matrix estimators (HCCMEs)

Up to now methods for dealing with heteroskedasticity have assumed that the structure is known or can be modelled, allowing one to correct for heteroskedasticity when conducting inference. If the structure and element values of Σ are completely unknown, all the previous methods will not suffice for correct inference (even for feasible generalised least squares). We now assume the structure is of a completely unknown form. To begin the discussion on finding a solution we consider the generalised linear model (2.19) with unknown heteroskedastic errors. The variance/covariance matrix of $\widehat{\beta} - \beta$, where $\widehat{\beta}$ is the OLS estimator is then given by:

$$\begin{aligned} Var(\widehat{\beta} - \beta) &= E[(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \end{aligned}$$

$$\begin{aligned}
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},
\end{aligned}$$

this matrix is known as a sandwich covariance matrix, since $\mathbf{X}'\Sigma\mathbf{X}$ is “sandwiched” between the two matrices $(\mathbf{X}'\mathbf{X})^{-1}$. This sandwich covariance matrix form is commonly associated with inefficient least squares estimators (Davidson and MacKinnon, 2004), since observations with lower variance convey more information regarding the regression coefficients than those with higher variances. Ideally, observations with lower variance should be assigned more weight and those with higher variance less weight, to obtain a more efficient estimator. If the diagonal elements σ_i^2 of the variance/covariance matrix Σ were known we could solve the sandwich covariance matrix and proceed with inference using generalised least squares. However, the diagonal elements are unknown and form a total of n variance components, this characteristic makes estimation difficult since only one data point is available to estimate each variance component. If we recall in the OLS case n observations were used to estimate the single variance component σ^2 so this problem was not of concern. Fortunately, White (1980) found a solution for the sandwich covariance matrices by proving that estimating the k variance components in the matrix $\mathbf{X}'\Sigma\mathbf{X}$ was sufficient to consistently estimate the variance of $\hat{\beta}$, rather than the n elements in Σ .

An outline of this important result is presented, and starts by considering the required assumptions:

ASSUMPTION 1: *The linear regression model is defined as in (2.19), where the elements in the sequence $\{\mathbf{X}_i, \varepsilon_i\}_{i=1}^n$ are independent but not identically distributed, uncorrelated and satisfy $E(\mathbf{X}_i'\varepsilon_i) = 0$. Y_i and $\mathbf{X}_i$ are observable and ε_i is unobservable.*

The assumptions allow for heteroskedasticity of the form $E(\varepsilon_i^2|\mathbf{X}_i) = g(\mathbf{X}_i)$, where g is a known function, possibly parametric in nature. Moreover, the predictors may be of a stochastic or fixed nature.

ASSUMPTION 2: *There exist finite positive constants δ and ψ , such that $E(|\varepsilon_i|^2)^{1+\delta} < \psi$ and $E(|X_{ij}X_{il}|^{1+\delta}) < \psi \ \forall \ i = 1, 2, \dots, n$ and $j, l = 1, 2, \dots, k$. In addition, the average covariance*

matrix of the predictors $\bar{\mathbf{M}}_n = \frac{1}{n} \sum_{i=1}^n E(\mathbf{X}_i' \mathbf{X}_i)$ is nonsingular, such that $\bar{\mathbf{M}}_n > \delta > 0$, as $n \rightarrow \infty$. Thus the error variances and the elements of $\bar{\mathbf{M}}_n$ are uniformly bounded. If the predictors are fixed, then the additional assumption is made that they are also uniformly bounded. Moreover, $\bar{\mathbf{M}}_n$ is not required to converge to any particular limit.

ASSUMPTION 3: *There exist finite positive constants δ and ψ , such that $E(|\varepsilon_i^2 X_{ij} X_{il}|^{1+\delta}) < \psi$ $\forall i = 1, 2, \dots, n$ and $j, l = 1, 2, \dots, k$. In addition, the average covariance matrix*

$$\bar{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^n E(\varepsilon_i^2 \mathbf{X}_i' \mathbf{X}_i)$$

is nonsingular, such that $\bar{\mathbf{V}}_n > \delta > 0$, as $n \rightarrow \infty$. The assumptions ensure that the inverse of $\bar{\mathbf{V}}_n$ is uniformly bounded.

ASSUMPTION 4: *There exist finite positive constants δ and ψ , such that $E(|X_{ip}^2 X_{ij} X_{il}|^{1+\delta}) < \psi$ $\forall i = 1, 2, \dots, n$ and $p, j, l = 1, 2, \dots, k$.*

Given the first two assumptions, it can be proved that the OLS estimator $\hat{\beta}$ exists almost surely by ensuring that all inverted matrices exist in the expression of the estimator and $\hat{\beta} \xrightarrow{a.s.} \beta$, as $n \rightarrow \infty$.⁵

Next, consider finding a consistent estimator for the unknown quantity $\bar{\mathbf{M}}_n^{-1} \bar{\mathbf{V}}_n \bar{\mathbf{M}}_n^{-1}$, which is plagued by the uncertainty around estimating the quantity $\bar{\mathbf{V}}_n$. By considering the fixed predictor case $\bar{\mathbf{V}}_n = \frac{1}{n} \mathbf{X}' \Sigma \mathbf{X}$, we return to the problem of estimating n variance components in Σ with only one observation for each. By rewriting the unknown quantity

$$\bar{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \mathbf{X}_i' \mathbf{X}_i = \frac{1}{n} \sum_{i=1}^n E(\varepsilon_i^2 \mathbf{X}_i' \mathbf{X}_i),$$

one notices that it is an average of expected values. MacKinnon (2013) stated that econometricians were perplexed with the issue due to them assuming that each expectation required

⁵The notation $\xrightarrow{a.s.}$ is used to denote *almost sure convergence*. That is, if $\{A_n\}$ denotes a sequence of random variables, and A denotes a random variable, then A_n converges almost surely to A if $P(\lim_{n \rightarrow \infty} A_n = A) = 1$.

estimation before the average could be calculated despite early hints in the literature by Eicker (1963), Eicker (1967) and Hinkley (1977) that it might be possible to simply estimate the entire average rather than the individual components. This is where White's (1980) great breakthrough made headway by showing that estimation of the individual elements was not required for heteroskedastic consistent covariance matrix estimation.

Since $\hat{\beta}$ is consistent for β , e_i will be consistent for ε_i and e_i^2 will then also be consistent for ε_i^2 .

As a result,

$$\hat{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^n E(e_i^2 \mathbf{X}_i' \mathbf{X}_i)$$

is a natural estimator for $\bar{\mathbf{V}}_n$, and White (1980) proves under the four assumptions that,

$$|\hat{\mathbf{V}}_n - \bar{\mathbf{V}}_n| \xrightarrow{a.s.} 0 \text{ and } |(\mathbf{X}'\mathbf{X}/n)^{-1} \hat{\mathbf{V}}_n (\mathbf{X}'\mathbf{X}/n)^{-1} - \bar{\mathbf{M}}_n^{-1} \bar{\mathbf{V}}_n \bar{\mathbf{M}}_n^{-1}| \xrightarrow{a.s.} 0.$$

Thus $\hat{\mathbf{V}}_n$ is shown to consistently estimate $\bar{\mathbf{V}}_n$ despite the concerns discussed.

The second part concludes by presenting the expression for the HCCME and the consistency thereof. After White (1980) formally proved the consistency of his estimator, various other estimators for $\hat{\mathbf{V}}_n$ have been suggested and are summarised in Table 2.1 along with their respective shorthand notations.

Different HCCMEs

White's (1980) estimator is commonly defined as HC0 in the literature. Later three new modified versions were proposed by MacKinnon and White (1985). The first modification, denoted HC1, tries to correct for the small values that would result from squaring the residuals; it corrects the squared residuals similarly to the variance/covariance OLS estimator (Hinkley, 1977) in that they are multiplied by $n/(n-k)$ degrees of freedom. The second modification inspired by Horn, Horn and Duncan (1975), involves dividing each squared residuals by $(1-h_i)$, where h_i is the i^{th} diagonal element in the hat/projection matrix $\mathbf{H}$. This estimator is commonly known as HC2, and has the property that it will be unbiased under homoskedasticity where HC1 only

has this property for balanced design cases, i.e., when $h_i = k/n$ (MacKinnon, 2013). The third modification was based on the jackknife and is biased upwards when the errors are homoskedastic and will be denoted by HC3. Two other HCCMEs for dealing with high and low leverage value data were suggested years later and involve increasing (decreasing) the i^{th} residual when h_i is large (small) relative to the mean of leverages, k/n . The two methods can be found in Cribari-Neto (2004) and Cribari-Neto, Souza and Vasconcellos (2007), and are denoted by HC4 and HC5, respectively.

Which HCCME is best suited for heteroskedastic robust inference?

Simulation evidence by MacKinnon and White (1985) showed that the standard error for the quasi t -statistics based on HC1 exceeded that of HC2, and HC3 exceeded that of HC2. Another important result from that simulation study was that HC3 had similar performance to OLS when the data was homoskedastic ($n = 100$), however, tests based on HC0 and HC1 tended to over-reject severely when the sample size was small and homoskedastic. Even though HC3 did well compared to HC0, HC1 and HC2, it suffered from small sample distortions when $n \leq 50$. Later simulation evidence by Long and Ervin (1998) again confirmed that HC3 was the best HCCME to use for moderate to extreme heteroskedasticity, and even functions well in the presence of homoskedasticity. The authors highly recommend that if a researcher suspects that heteroskedasticity may be present, they should correct for it using a suitable HCCME, specifically HC3. It was suggested that for a sample size $n \leq 250$ HC3 should be used, and for $n \geq 500$ HC1 and HC2 may also be considered. However, it should be noted that HCCMEs are sensitive to large leverage values (Chesher and Austin, 1991; Long and Ervin, 1998) and was one of the features considered in the simulation study in Chapter 5 of this text.

Table 2.1: Proposed estimators for $\overline{\mathbf{V}}_n$.

Type	Estimator ($\widehat{\mathbf{V}}_n$)
HC0 (White, 1980)	$\frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{X}_i \mathbf{X}_i'$
HC1 (MacKinnon and White, 1985)	$\frac{1}{n} \sum_{i=1}^n \left(\frac{n}{n-k} \right) e_i^2 \mathbf{X}_i \mathbf{X}_i'$
HC2 (MacKinnon and White, 1985)	$\frac{1}{n} \sum_{i=1}^n \frac{e_i^2}{1-h_i} \mathbf{X}_i \mathbf{X}_i'$
HC3 (MacKinnon and White, 1985)	$\frac{1}{n} \sum_{i=1}^n \frac{e_i^2}{(1-h_i)^2} \mathbf{X}_i \mathbf{X}_i'$
HC4 (Cribari-Neto, 2004)	$\frac{1}{n} \sum_{i=1}^n \frac{e_i^2}{(1-h_i)^{\delta_i}} \mathbf{X}_i \mathbf{X}_i',$ $\delta_i = \min(4, \frac{nh_i}{p-1})$
HC5 (Cribari-Neto et al., 2007)	$\frac{1}{n} \sum_{i=1}^n \frac{e_i^2}{\sqrt{(1-h_i)^{\delta_i}}} \mathbf{X}_i \mathbf{X}_i',$ $\delta_i = \min(\frac{nh_i}{p-1}, \max(4, \frac{0.7nh_{max}}{p-1})),$ where $h_{max} = \max(h_i) \forall i = 1, 2, \dots, n$.

CHAPTER 3

The Bootstrap

The following two sections introduce the method known as the bootstrap, this method is shown to be highly versatile for the estimation of distributional properties for statistics (e.g. standard errors). The bootstrap's versatility is then expanded to the sampling distribution of test statistics, which play a central role in this text for the testing of hypotheses around the validity of model parameters (e.g. regression coefficients).

As discussed in the previous chapter, when the assumption of homoskedasticity is violated the estimated regression coefficients remain consistent and unbiased (White, 1980), however the estimated standard errors of the regression coefficients will be inconsistent in that they are either too large or small. Consequently, the ordinary least squares estimator is no longer the best in the class of linear unbiased estimators (Greene, 2003). This is an inconvenience when inference around the regression coefficients needs to be conducted due to the test statistics under the classical model (2.2) not having their familiar t or F asymptotic distributions; this ultimately leads to over-rejection or under-rejection of the null hypothesis. In Chapter 2 the following solutions were presented for dealing with heteroskedasticity for inference,

1. weighted least squares,
2. feasible generalised least squares, and
3. heteroskedasticity consistent covariance matrix estimators (HCCMEs).

The first two methods require the structure of heteroskedasticity to be known or to be able to be modelled, which is rarely if ever the case. The HCCMEs, on the other hand, do not function well when the sample size is small — see Section 2.8.3. HCCMEs do not make any assumptions about the structure of the heteroskedasticity, but suffer from sample size distortions (MacKinnon and

White, 1985), a possible solution could be to consider combining it with the bootstrap to try and limit these concerns when conducting inference. Since bootstrap methods work via simulation, it may prove handy in alleviating sample size distortions caused by small samples in inference combined with the desirable properties of HCCMEs.

Chapter 3 starts by introducing the concepts behind the bootstrap method in Section 3.1, and applies these concepts to a simple hypothesis testing example in Section 3.2. Section 3.3 introduces the different approaches available in the literature for bootstrap regression and Section 3.4 concludes the chapter by applying the bootstrap regression to test hypotheses concerning the regression coefficient and corresponding algorithms.

3.1 An introduction to the Bootstrap

In 1979, Efron published a landmark article where he suggested a general method known as the “Bootstrap”. To understand the rationale behind the bootstrap consider a prespecified random variable $\hat{\theta} := \theta(\mathbf{X}, F)$, where $\mathbf{X} = [X_1, X_2, \dots, X_n]$ denotes a random sample from an unknown probability distribution F . Since F is unknown, the distributional properties of $\hat{\theta}$ will consequently also be unknown. The bootstrap enables one to estimate these distributional properties by replacing the unknown distribution F with an empirical estimator $\hat{F}$. There are various estimators for F to choose from, but the most common is the empirical distribution function F_n , which assigns an equal weight ($1/n$) on each observation in the random sample $\mathbf{X}$,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x),$$

where

$$I(X \leq x) = \begin{cases} 0, & \text{if } X > x \\ 1, & \text{if } X \leq x. \end{cases}$$

Samples drawn with replacement from the random sample $\mathbf{X}$ are denoted $\mathbf{X}^* = [X_1^*, X_2^*, \dots, X_n^*]$ and are referred to as the bootstrap samples from F_n . One can then use the bootstrap samples to calculate what is known as the bootstrap statistic, $\hat{\theta}^* := \theta(\mathbf{X}^*, F_n)$. The distributional properties

of $\hat{\theta}$ can then be estimated using the bootstrap distributional properties of $\hat{\theta}^*$ (Efron, 1979). In practice, the distributional properties of $\hat{\theta}^*$ are often approximated via Monte Carlo simulation; the procedure involves repeatedly and independently sampling from F_n , say B times, and calculating the statistic each time to obtain $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$. In essence the bootstrap involves seeing the original sample data as a “population” from which samples may be drawn. The Monte Carlo procedure then exploits this by repeatedly drawing from this “real” world sample and calculating statistics of interest from the “bootstrap” world sample data. The sampling distribution of the “bootstrap” world statistic, which is an estimator for the sampling distribution of the “real” world statistic, is then approximated by using this Monte Carlo approach. Figure 3.1 illustrates the concept, by representing the original sample in the “real” world with its corresponding statistic $\hat{\theta}$. The bootstrap then manifests as the expansion into B bootstrap samples from the original and the corresponding statistics $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$ in the bootstrap world, which are ultimately used to approximate the distributional properties of $\hat{\theta}$.

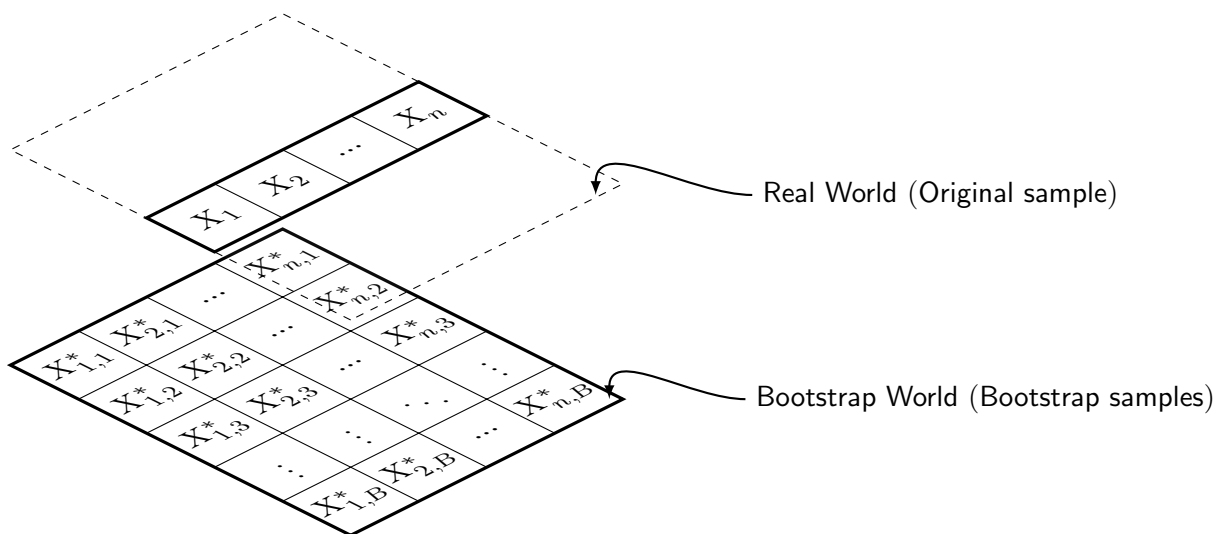


Fig. 3.1: Different “worlds”.

The most common distributional properties of interest calculated for the statistic $\hat{\theta}$ using the bootstrap is that of bias (accuracy) and standard error or variance (precision) (Efron and Tibshirani, 1993). To illustrate the concepts of accuracy and precision, consider the performance of a sport shooter during a stage where he needs to engage three targets, T1-T3 in the chest area with five shots each. The shooter draws his firearm at the sound of the buzzer and engages T1, from Figure (3.2) we notice his shots are both accurate and precise. The shooter then performs

a reload and engages T2, however due to bad trigger control his shots are low left, which is still precise but no longer accurate. The shooter then performs another reload and engages T3, due to the distance and his poor grip his shots are no longer accurate or precise.

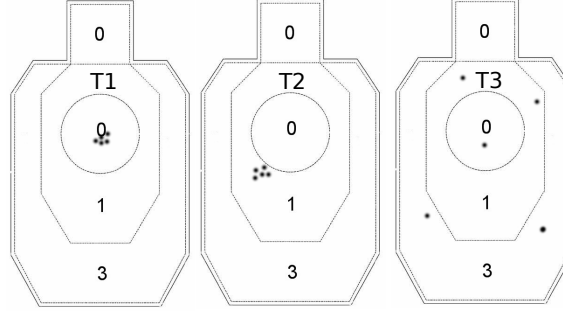


Fig. 3.2: Accuracy and precision illustration.

Consider the simple example of approximating the standard error (precision) of the statistic $\hat{\theta}$, given the sample $\mathbf{X} = [X_1, X_2, \dots, X_n] \stackrel{i.i.d.}{\sim} F$. Let the standard error of $\hat{\theta}$ be denoted by $\sigma_{SE} = \sqrt{Var(\hat{\theta})}$. The statistic $\hat{\theta}$ does not need to have an explicit form for the bootstrap to be applied, this is a major advantage of the bootstrap when faced with complex statistics which may not have explicit standard error expressions. Denote the ideal bootstrap estimate of the standard error by $\hat{\sigma}_{SE} = \sqrt{Var_*(\hat{\theta}^*)}$, where the operator Var_* denotes the conditional variance given the sample data $\{X_i\}_{i=1}^n$. The ideal bootstrap estimate is difficult to obtain using simulation since it would require the number of bootstrap replications to approach infinity ($B \rightarrow \infty$). Since an infinite number of replications would be impossible to calculate using a computer, the ideal bootstrap estimate of the standard error is approximated using a large, finite number of replications and will be denoted by $\hat{\sigma}_{SE_B}$. This Monte Carlo result approximates the ideal bootstrap estimate and converges to the bootstrap estimate as B approaches infinity. A possible algorithm for approximating the ideal bootstrap estimate $\hat{\sigma}_{SE}$ is.

Algorithm: Algorithm for approximating the ideal bootstrap estimate for standard error

1. Given the random sample $\mathbf{X} = [X_1, X_2, \dots, X_n] \stackrel{i.i.d.}{\sim} F$, generate the bootstrap sample $\mathbf{X}^* = [X_1^*, X_2^*, \dots, X_n^*]$ from the empirical distribution function (F_n) by sampling with replacement from $\mathbf{X}$.

2. Calculate the statistic $\hat{\theta}^* = \theta(\mathbf{X}^*, F_n)$ using the bootstrap sample from step (1).
3. Repeat steps (1) and (2) independently B -times to obtain $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$.
4. Finally calculate the Monte Carlo approximation:

$$\hat{\sigma}_{SE_B} = \sqrt{\frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}_\bullet^*)^2}, \text{ where } \hat{\theta}_\bullet^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*.$$

The algorithm serves as an illustration of how Monte Carlo simulations can be employed to obtain approximations for the ideal bootstrap estimates. Similarly one can also use this type of bootstrap Monte Carlo method to approximate estimators for bias and even confidence intervals — see Efron and Tibshirani (1993).

When conducting inference via hypothesis testing, the bootstrap can also be used to estimate the sampling distribution of a test statistic under the null hypothesis. The next section introduces these concepts in the context of a single sample hypothesis test, and the concept will be expanded to inference for linear regression model in the section thereafter.

3.2 Bootstrap hypothesis testing

In this section we shall consider composite hypotheses for some parameter θ of the form,

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_A : \theta \neq \theta_0.$$

Suppose that we use a test statistic denoted $T := T(\mathbf{X}, \theta_0)$, where the null distribution of T is potentially unknown. Hypothesis tests are conducted by comparing the value of a test statistic calculated from observed data to that of the quantities of the distribution of the test statistic if the null hypothesis were true (i.e. the null distribution). We reject the null hypothesis if the calculated test statistic lies within the extreme regions (tails) of the null distribution of the test statistic. In most cases the null distribution will only be known under certain model assumptions such as normality and/or independence of the observations. If the null distribution of the test statistic is unknown, parametric hypothesis testing of this kind will not be possible. To overcome

this problem, the bootstrap can be used to estimate the null distribution of the test statistic, but this cannot be done naively and a few considerations should be made as discussed in Hall and Wilson (1991) and Horowitz (1995).

1. The first of these considerations is that resampling should be done from a sample (real world sample) that reflects the null hypothesis,
2. secondly, the test statistic should be appropriately scaled using what is known as “bootstrap pivoting”.

Failing to comply with consideration 1 reduces the power of the test whereas not adhering to consideration 2 negatively affects coverage accuracy, i.e., does not provide the higher order asymptotic refinement the bootstrap possess for asymptotic test statistics.

To continue the discussion on hypothesis testing using the bootstrap we shall consider the example of testing for a single population mean, μ . The specific hypothesis we are now interested in testing is:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_A : \mu \neq \mu_0. \quad (3.1)$$

Application of the bootstrap for testing this hypothesis has been considered by a number of authors, Young (1988), Hall and Wilson (1991) and Efron and Tibshirani (1993). The classic test statistic that one normally uses to test this hypothesis is the standard t test statistic, given by:

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S_X}, \quad (3.2)$$

where

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ and } S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

If one assumes the data are *i.i.d.* from a normal distribution with mean μ and variance σ^2 , then this test statistic follows a t -distribution with $n - 1$ degrees of freedom. Denote the value of the test statistic calculated from the sample data by t , then the null hypothesis is rejected

$$\text{if } t \leq c\left(\frac{\alpha}{2}\right) \quad \text{or, if } t \geq d\left(\frac{\alpha}{2}\right)$$

where $c\left(\frac{\alpha}{2}\right)$ and $d\left(\frac{\alpha}{2}\right)$ are quantiles from the null distribution of T defined as

$$P\left(T \leq c\left(\frac{\alpha}{2}\right) \mid H_0\right) \approx \frac{\alpha}{2} \quad \text{and} \quad P\left(T \geq d\left(\frac{\alpha}{2}\right) \mid H_0\right) \approx \frac{\alpha}{2}.$$

Alternatively the null hypothesis is rejected if the p -value is less than the specified significance level α , where the p -value is defined as:

$$p\text{-value} = 2 \min \{P(T \geq t \mid H_0), P(T \leq t \mid H_0)\}.$$

In order to use the bootstrap for this test, we shall consider two approaches: one based on transforming the data and the other based on a model.

Transformation based approach for hypothesis testing Apply the following transformation to the data, $\{X_i\}_{i=1}^n$:

$$Y_i = X_i - \bar{X} + \mu_0, \quad i = 1, 2, \dots, n.$$

Note that the sample mean for $\{Y_i\}_{i=1}^n$ is equal to μ_0 . This transformation is done to ensure that the sample conforms to the statement made in the null hypothesis and is a direct application of the first consideration from Hall and Wilson (1991). The bootstrap procedure then proceeds by sampling with replacement from this new transformed variable to obtain the bootstrap sample $\{Y_i^*\}_{i=1}^n$, defined as

$$Y_i^* = X_i^* - \bar{X} + \mu_0, \quad i = 1, 2, \dots, n. \quad (3.3)$$

The null distribution of the test statistic T can then be estimated by considering the sampling distribution of the following bootstrap statistic

$$T^* = \frac{\sqrt{n}(\bar{Y}^* - \mu_0)}{S_{Y^*}},$$

where

$$\bar{Y}^* = \frac{1}{n} \sum_{i=1}^n Y_i^* \quad \text{and} \quad S_{Y^*}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i^* - \bar{Y}^*)^2.$$

If we substitute Equation (3.3) in the above expressions and note that $S_{X^*}^2$ is equal to $S_{Y^*}^2$, we get the following equivalent definition of T^*

$$T^* = \frac{\sqrt{n}(\bar{X}^* - \bar{X})}{S_{X^*}}.$$

To test the hypothesis we simply have to estimate the critical values $c(\alpha/2)$ and $d(\alpha/2)$ from the null distribution of the test statistic. These estimators, denoted $\hat{c}(\alpha/2)$ and $\hat{d}(\alpha/2)$ are obtained by solving the following expressions:

$$P^* \left(T^* \leq \hat{c} \left(\frac{\alpha}{2} \right) \mid H_0 \right) \approx \frac{\alpha}{2} \quad \text{and} \quad P^* \left(T^* \geq \hat{d} \left(\frac{\alpha}{2} \right) \mid H_0 \right) \approx \frac{\alpha}{2},$$

where P^* is the conditional probability operator given the sample data $\{X_i\}_{i=1}^n$.

Alternatively, one can also make use of the bootstrap p -value to test the hypothesis. This p -value is defined as:

$$p^* \text{-value} = 2 \min \{ P^* (T^* > t \mid H_0), P^* (T^* \leq t \mid H_0) \}.$$

The Monte Carlo approximation of these bootstrap critical values and bootstrap p -value is then given by the following algorithm.

Algorithm: Bootstrap hypothesis test for the hypothesis in (3.1) using the transformation based approach.

1. Given the sample data $\{X_i\}_{i=1}^n$ apply the transformation $Y_i = X_i - \bar{X} + \mu_0$.
2. Sample with replacement from the transformed data $\{Y_i\}_{i=1}^n$ to obtain the bootstrap sample $\{Y_i^*\}_{i=1}^n$, and then calculate the bootstrap test statistic T^* using this data.
3. Repeat step (2) independently B times to obtain the bootstrap replications $T_1^*, T_2^*, \dots, T_B^*$.
4. Determine the order statistics of the bootstrap replications, i.e., $T_{(1)}^* \leq T_{(2)}^* \leq \dots \leq T_{(B)}^*$.

5. The approximation for the bootstrap critical value estimators are given by,

$$\hat{c}_B\left(\frac{\alpha}{2}\right) = T_{(\lfloor B(\alpha/2) \rfloor)}^* \quad \text{and} \quad \hat{d}_B\left(\frac{\alpha}{2}\right) = T_{(\lfloor B(1-\alpha/2) \rfloor)}^*,$$

where $\lfloor \cdot \rfloor$ denotes the “floor” operator, i.e., $\lfloor x \rfloor$ is the largest integer value less than x . The algorithm for the Monte Carlo approximation of the bootstrap p -value is identical to the algorithm for the critical values described above, except that steps (4) and (5) are replaced with:

4. Approximate the p -value with:

$$p_B^* \text{-value} = 2 \min \left\{ \frac{1}{B} \sum_{b=1}^B I(T_b^* > t), \frac{1}{B} \sum_{b=1}^B I(T_b^* \leq t) \right\}.$$

Model based approach for hypothesis testing The transformation approach discussed in the previous subsection was done to ensure that the transformed data satisfied the null hypothesis by directly manipulating the sample data. We shall now consider a different approach that also attempts to enforce the conditions stated in the null hypothesis by assuming that the data generating process follows a known model. For the example of testing the hypothesis given in Equation (3.1) suppose that the data $\{X_i\}_{i=1}^n$ has mean μ and variance σ^2 and was generated from the following location/scale model:

$$X_i = \mu + \sigma \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (3.4)$$

where $\{\varepsilon_i\}_{i=1}^n$ are random *i.i.d.* errors with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = 1$, for all $i = 1, 2, \dots, n$. The first step then in the model based bootstrap approach is to estimate the parameter μ and σ^2 using $\bar{X}$ and S_X^2 , respectively. The estimated error terms can then be expressed as:

$$\hat{\varepsilon}_i = \frac{X_i - \bar{X}}{S_X}, \quad i = 1, 2, \dots, n.$$

The goal of this approach is, once again, to ensure that the bootstrap samples are drawn from a data generating process that satisfies the null hypothesis. The model based approach achieves

this by only resampling the estimated error terms (residuals) and using the estimated version of the model in Equation (3.4), where the null hypothesis has been enforced to generate bootstrap samples. If the resampled residuals, obtained by sampling with replacement from $\{\hat{\varepsilon}_i\}_{i=1}^n$, are denoted by $\{\hat{\varepsilon}_i^*\}_{i=1}^n$, then the bootstrap sample that would be generated under the null hypothesis is given by:

$$Y_i^* = \mu_0 + S_X \hat{\varepsilon}_i^*, \quad i = 1, 2, \dots, n. \quad (3.5)$$

Testing the hypothesis in Equation (3.1) continues as it did in the transformation based section, except that the definition of Y_i^* in Equation (3.5) is used instead of the one defined in Equation (3.3). The Monte Carlo approximation of these bootstrap critical values and bootstrap p -value is then similarly given by the following algorithm.

Algorithm: Bootstrap hypothesis test for the hypothesis in (3.1) using the model based approach.

1. Given the sample data $\{X_i\}_{i=1}^n$ estimate the parameters μ and σ^2 using $\bar{X}$ and S_X^2 , respectively.
2. Calculate the residuals using the formula:

$$\hat{\varepsilon}_i = \frac{X_i - \bar{X}}{S_X}, \quad i = 1, 2, \dots, n.$$

3. Sample with replacement from the residuals $\{\hat{\varepsilon}_i\}_{i=1}^n$ to obtain the bootstrap residuals denoted $\{\hat{\varepsilon}_i^*\}_{i=1}^n$.

4. Calculate

$$Y_i^* = \mu_0 + S_X \hat{\varepsilon}_i^*, \quad i = 1, 2, \dots, n,$$

and then calculate the bootstrap test statistic T^* using this data.

5. Repeat steps (3) and (4) independently B times to obtain the bootstrap replications $T_1^*, T_2^*, \dots, T_B^*$.

6. Determine the order statistics of the bootstrap replications, i.e., $T_{(1)}^* \leq T_{(2)}^* \leq \dots \leq$

$$T_{(B)}^*.$$

7. The approximations for the bootstrap critical value estimators are given by,

$$\hat{c}_B\left(\frac{\alpha}{2}\right) = T_{(\lfloor B(\alpha/2) \rfloor)}^* \quad \text{and} \quad \hat{d}_B\left(\frac{\alpha}{2}\right) = T_{(\lfloor B(1-\alpha/2) \rfloor)}^*.$$

Once again, the bootstrap p -value calculation is identical to the one used for the critical values above, except that steps (6) and (7) are replaced with:

6. Approximate the p -value with:

$$p_B^* \text{-value} = 2 \min \left\{ \frac{1}{B} \sum_{b=1}^B I(T_b^* > t), \frac{1}{B} \sum_{b=1}^B I(T_b^* \leq t) \right\}.$$

The concepts introduced here were used to demonstrate the application of the bootstrap to hypothesis testing for a single population mean. Next we shall further expand on the application by testing hypotheses for the regression coefficients of linear regression models.

3.3 Bootstrap regression

In Section 2.4.5 we noted that the distribution of the test statistic for testing the statistical significance of a regression coefficient followed a t distribution with $n - k$ degrees of freedom under the null hypothesis, if all of the OLS assumptions were valid. As previously noted, by violating the homoskedasticity assumption the OLS estimator of the variance/covariance of the regression coefficient estimates will either be too large or too small and, even though the normality assumption still remains valid, the inferences made will tend to over or under reject due to the distribution now being an approximation rather than an exact distribution. HCCMEs in Section 2.8.3 allowed us to correct for this inefficiency and to conduct inferences using the “exact” approach once more by substitution of a suitable HCCME in the place of the variance/covariance matrix derived in Section 2.4.4. Unfortunately, even though the variance/covariance can be estimated consistently, the distribution of these test statistics will still be unknown if the model assumptions are violated. In these cases, the bootstrap can be used to estimate the sampling distribution of these statistics.

The bootstrap applied to regression models mainly focusses on the approximation of standard errors, or rather the variance/covariance matrices of the estimated regression coefficients (MacKinnon, 2013), but can also be used to estimate the probability distributions of regression test statistics. Bootstrap standard errors of regression coefficient estimators are obtained by calculating the variation among estimated regression coefficients from the bootstrap samples. However, some of these approximations provide similar results and expressions for finite samples as those generated by HCCMEs, e.g., using the pairs bootstrap approach (discussed in Section 3.3.2) it can be shown using a delta method approximation that the bootstrap estimator for the variance/covariance estimator for the estimated regression coefficients $\hat{\beta}^*$ is given by the same expression as White's estimator – HC0 (Lancaster, 2006). MacKinnon (2013) states that when B is large enough the bootstrap variance estimate may possibly be better than HC0, however HC2 or HC3 should still outperform it. However, the method would be useful in cases where complex nonlinear models or complex nonlinear functions of the estimated regression coefficients are used, for which there exists no equivalent HCCME expressions.

In the context of hypothesis testing, bootstrapping regression models is perhaps more sensible. Since one can use the bootstrap to estimate the null distribution of test statistics, and thereby use it to test hypotheses around the significance of the regression coefficients. Bootstrap hypothesis testing is typically achieved by creating a data generating process (bootstrap DGP) from which a test can be constructed. Inference can then be conducted using a p -value which was calculated using the bootstrap null distribution of the test statistic under the bootstrap DGP.

In the next subsections we shall consider various approaches of applying the bootstrap to linear regression models in order to ultimately test hypothesis statements made concerning the regression coefficients. The first is known as the bootstrap residuals approach and is similar to the model based approach previously discussed in Section 3.2 for testing the population mean. A shortcoming of the bootstrap residuals approach is that it requires the assumption of homoskedasticity, and so two other approaches, namely the bootstrap pairs approach and the wild bootstrap approach, are introduced to remedy this shortcoming.

3.3.1 Bootstrap residuals approach

The first regression bootstrap approach was suggested by Efron (1979), and it may perhaps be considered the most intuitive bootstrap regression method since its basis is closely related to bootstrapping single sample problems. It is important to note that this bootstrap method does not allow for the errors to be heteroskedastic.

The classic linear regression model (2.2) can be written as:

$$Y_i = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i, i = 1, \dots, n,$$

where the error terms $\varepsilon_i \stackrel{i.i.d.}{\sim} F$, $i = 1, \dots, n$. The only assumption made is that the distribution F is centred around zero, e.g. around the mean or median. Since the response and predictors variables have been observed, the regression coefficients can be estimated using OLS,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The bootstrap sample is then generated using the following DGP,

$$Y_i^* = \mathbf{X}_i' \hat{\boldsymbol{\beta}} + e_i^*, i = 1, \dots, n, \quad (3.6)$$

where $\{e_i^*\}_{i=1}^n$ is vector of resampled residuals obtain by sampling with replacement from the OLS residuals $\{e_i\}_{i=1}^n$. The resampling may be done using the original residuals e_i or, if one suspects that the assumption of them being centred around zero is violated, it may be enforced by rather using the centred residuals $\check{e}_i = e_i - \frac{1}{n} \sum_{i=1}^n e_i$. Using $\mathbf{Y}^* = [Y_1^*, Y_2^*, \dots, Y_n^*]'$ the bootstrap sample regression coefficients can be estimated using OLS,

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^*. \quad (3.7)$$

The approximate bootstrap distribution of the bootstrap statistics $\hat{\boldsymbol{\beta}}^*$ can be obtained via a Monte Carlo simulation. This would entail independently repeating the above process, say B times, to obtain the bootstrap replicates $\hat{\boldsymbol{\beta}}_1^*, \hat{\boldsymbol{\beta}}_2^*, \dots, \hat{\boldsymbol{\beta}}_B^*$, which can then be used in the usual

way to approximate distributional properties of the bootstrap statistic.

The bootstrap estimates $\hat{\beta}^*$ have the following properties:

$$E_*(\hat{\beta}^*) = \hat{\beta} \text{ since } E_*(e_i^*) = 0 \ \forall \ i = 1, 2, \dots, n,$$

and

$$Var_*(\hat{\beta}^*) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var_*(\mathbf{Y}^*)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (3.8)$$

Although the expression for the variance $\hat{\sigma}^2$ in Equation (3.8) has a familiar form it should not be confused with the OLS sample version (which is typically expressed as $\frac{1}{n-k} \sum_{i=1}^n e_i^2$), but is instead given by

$$\hat{\sigma}^2 = Var_*(e_i^*) = \frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})^2, \quad (3.9)$$

and is downwardly biased (Chesher and Jewitt, 1987; Shao and Tu, 1995), since

$$E(\hat{\sigma}^2) = \left(1 - \frac{p_n}{n}\right)\sigma^2,$$

where $p_n = k + 1 - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_j \geq k > 0$. Efron (1982) suggested by resampling from $\frac{e_i - \bar{e}}{\sqrt{(n-k)/n}}$ the downward bias can be reduced, however it remains biased since $p_n > k$ if $\bar{e} \neq 0$. It is suggested in Shao and Tu (1995) that if resampling is done by using $\frac{e_i - \bar{e}}{\sqrt{(n-p_n)/n}}$, then

$$\hat{\sigma}^2 = \frac{1}{n - p_n} \sum_{i=1}^n (e_i - \bar{e})^2, \quad (3.10)$$

and if $\bar{e} = 0$ (i.e. if the intercept term is included in the linear regression model) then $p_n = k$ and the bootstrap variance estimator (3.10) is equal to the OLS sample version.

If the errors were found to be heteroskedastic the residuals bootstrap cannot be used and other approaches are required. The next approach, known as the pairs bootstrap, allows for this type of scenario and is the first of the heteroskedasticity robust bootstrap regression approaches.

3.3.2 Bootstrap pairs approach

The second bootstrap approach was suggested by Freedman (1981) to specifically deal with heteroskedastic errors by allowing the errors to have different variances of an unknown form unlike the residuals bootstrap. To overcome the problem of heteroskedasticity of unknown form he suggested to resample directly from the observed data $\mathbf{Y} = [Y_1, Y_2, \dots, Y_n]'$ and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n]'$. This can be done by forming the pairs $\mathbf{U}_i = (\mathbf{X}_i, Y_i)$, and assuming that these pairs have a multivariate distribution F (i.e. $\mathbf{U}_i \stackrel{i.i.d.}{\sim} F$), one would then create bootstrap samples by resampling the pairs $\{\mathbf{U}_i\}_{i=1}^n$ to get $\{\mathbf{U}_i^*\}_{i=1}^n$, and the bootstrap DGP is thus given by:

$$\mathbf{U}_i^* = (\mathbf{X}_i^*, Y_i^*), i = 1, \dots, n. \quad (3.11)$$

The bootstrap sample regression coefficients are then calculated by using the OLS estimator,

$$\hat{\beta}^* = (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{Y}^*.$$

An important theoretical consideration with the pairs bootstrap is that, even if the design matrix $\mathbf{X}$ has full column rank, it does not necessarily imply that it would be so for the bootstrap design matrix $\mathbf{X}^*$. If the bootstrap design matrix is found to not have full column rank, either a generalised inverse of $\mathbf{X}^{*'} \mathbf{X}^*$ must be used in calculating $\hat{\beta}^*$, e.g. Moore-Penrose pseudoinverse. Alternatively, one could adopt the more computationally efficient method suggested by Shao and Tu (1995):

$$\hat{\beta}^* = \begin{cases} (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{Y}^* & \text{if } \lambda^* \geq \frac{\lambda}{2} \\ \hat{\beta} & \text{otherwise,} \end{cases}$$

where λ and λ^* are equal to the smallest eigenvalues of $\mathbf{X}' \mathbf{X}$ and $\mathbf{X}^{*'} \mathbf{X}^*$, respectively. Because $P^*(\lambda^* < \frac{\lambda}{2}) \rightarrow 0$ at an exponential rate, the modification to the calculation of $\hat{\beta}^*$ is very minor and can be shown to have the properties,

$$E_*(\hat{\beta}^*) \approx \hat{\beta}$$

and

$$Var_*(\hat{\beta}^*) \approx (\mathbf{X}'\mathbf{X})^{-1} Var_* \left(\sum_{i=1}^n \mathbf{X}_i^* (Y_i^* - \mathbf{X}_i^{*'} \hat{\beta}) \right) (\mathbf{X}'\mathbf{X})^{-1} = \sum_{i=1}^n \mathbf{Z}_i \mathbf{Z}_i' e_i^2,$$

where E_* denotes the conditional expectation given the sample data $\{X_i\}_{i=1}^n$ and $\mathbf{Z}_i = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_i$.

The pairs bootstrap has been shown to be inferior in performance in various Monte Carlo simulations by Freedman (1981), Horowitz (1995), MacKinnon (2002), Flachaire (2005), Davidson and Flachaire (2008), and, additionally, is only valid if both the response and predictor variables are random which limits this approach to a smaller class of applications. The next heteroskedasticity consistent method for bootstrap regression is the wild bootstrap approach, and has improved performance over the pairs bootstrap approach.

3.3.3 Wild bootstrap/External bootstrap

The rationale behind the wild bootstrap was first proposed by Wu (1986) for dealing with heteroskedastic errors in regression models, and was based on the jackknife method in that the i^{th} fitted value was paired with the i^{th} residual, in contrast to the residuals based bootstrap where they do not necessarily have to correspond. The heteroskedastic linear regression model (2.19) can be written as:

$$Y_i = \mathbf{X}_i' \beta + \varepsilon_i, i = 1, \dots, n.$$

The wild bootstrap approach is similar to the residuals bootstrap in that bootstrap resampling mechanism is applied through the residuals rather than the original data as was the case with the pairs bootstrap. However, the wild bootstrap DGP differs from the residuals based bootstrap in that it attempts to retain the heteroskedastic structure present within the data. The wild bootstrap DGP is then given by,

$$Y_i^* = \mathbf{X}_i' \hat{\beta} + \frac{e_i}{\sqrt{1 - h_i}} t_i^*, i = 1, \dots, n, \quad (3.12)$$

where e_i is the i^{th} residual from OLS, $\hat{\boldsymbol{\beta}}$ is the OLS estimates of the regression coefficients, h_i is the i^{th} leverage value (i.e. the i^{th} diagonal element in the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$), and $\mathbf{t}^* = \{t_i^*\}_{i=1}^n$ are *i.i.d.* realisations from an ancillary distribution which satisfies the following assumptions:

$$E_*(t_i^*) = 0, \text{Var}_*(t_i^*) = E_*[(t_i^*)^2] = 1 \forall i \text{ and } \text{Cov}_*(t_i^*, t_j^*) = 0 \forall i \neq j. \quad (3.13)$$

The bootstrap regression coefficients are then estimated using OLS,

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^*.$$

This estimator has the following properties:

$$E_*(\hat{\boldsymbol{\beta}}^*) = \hat{\boldsymbol{\beta}} \text{ which follows as a consequence from assumptions (3.13),}$$

and

$$\text{Var}_*(\hat{\boldsymbol{\beta}}^*) = (\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \frac{e_i^2}{1-h_i} \text{Var}_*(t_i^*) \mathbf{X}_i \mathbf{X}_i' (\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \frac{e_i^2}{1-h_i} \mathbf{X}_i \mathbf{X}_i' (\mathbf{X}'\mathbf{X})^{-1}. \quad (3.14)$$

These estimators could then be used in hypothesis testing when constructing pivotal statistics. In fact, the bootstrap variance estimator (3.14) is equal to the heteroskedasticity robust weighted jackknife variance estimator proposed by Wu (1986), thus sharing the same desirable robustness property. Liu (1988) showed that by adding the additional assumption $E_*[(t_i^*)^3] = 0$ to assumptions (3.13) the bootstrap estimators would benefit from second order accuracy.

Moving forward we can now define a more general version of the wild bootstrap DGP as,

$$Y_i^* = \mathbf{X}_i' \hat{\boldsymbol{\beta}} + f(e_i) t_i^*, \quad i = 1, \dots, n, \quad (3.15)$$

where $f(e_i)$ is the i^{th} transformed residual, $\hat{\boldsymbol{\beta}}$ is the OLS estimates of the regression coefficients and $\mathbf{t}^* = \{t_i^*\}_{i=1}^n$ are once again *i.i.d.* realisations from an ancillary distribution which satisfies

the conditions given in (3.13). Higher order accuracy can be obtained by the inclusion of the third moment condition suggested by Liu (1988), however the next subsection will consider the limitations of imposing additional constraints on the moments of the ancillary distribution.

A note regarding the higher order accuracy of the wild bootstrap DGP

To ensure that the wild bootstrap works well, the residuals need to provide good approximations to the unknown error terms (Webb, 2014). A possible solution is to transform the residuals to ensure that the approximation is better. If the approximations are considered good, the bootstrap DGP's error terms should have mean zero and higher moments that are similar to that of the residuals.

MacKinnon (2013) noted a dilemma related to enforcing the first four moments of t_i^* 's distribution. This dilemma can be illustrated by attempting to impose the following moment conditions:

$$E(t_i^*) = 0, \quad E(t_i^{*2}) = 1, \quad E(t_i^{*3}) = 1, \quad E(t_i^{*4}) = 1. \quad (3.16)$$

It is impossible for any distribution to satisfy these conditions, since the outer product of the vector $[1 \ t^* \ t^{*2}]'$ for the random variable t^* with expected value 0 and variance 1 yields a 3×3 second moment matrix:

$$E \left(\begin{bmatrix} 1 \\ t^* \\ t^{*2} \end{bmatrix} \begin{bmatrix} 1 & t^* & t^{*2} \end{bmatrix} \right) = E \begin{bmatrix} 1 & t^* & t^{*2} \\ t^* & t^{*2} & t^{*3} \\ t^{*2} & t^{*3} & t^{*4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & \mu_3 \\ 1 & \mu_3 & \mu_4 \end{bmatrix}, \quad (3.17)$$

where μ_3 and μ_4 denote the third and fourth moments of t^* respectively. The second moment matrix (3.17)¹ is positive semidefinite (Lindsay, 1989a; Lindsay, 1989b), i.e., the determinant should be nonnegative. If we only impose restriction on the first two moments, i.e., $E(t_i^*) = 0$

¹The p^{th} moment matrix of the random variable X with distribution F is defined as:

$$\mathbf{M} = \begin{bmatrix} 1 & E(X) & E(X^2) & \dots & E(X^p) \\ E(X) & E(X^2) & E(X^3) & \dots & E(X^{p+1}) \\ E(X^2) & E(X^3) & E(X^4) & \dots & E(X^{p+2}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E(X^p) & E(X^{p+1}) & E(X^{p+2}) & \dots & E(X^{2p}) \end{bmatrix}.$$

and $E(t_i^{*2}) = 1$, then the determinant of the second moment matrix is given by $\mu_4 - \mu_3^2 - 1 \geq 0$. Now, if we would like to impose the third moment condition that $\mu_3 = 1$, then it necessarily implies that $\mu_4 \geq 2$. On the other hand, if we impose the fourth moment condition $\mu_4 = 1$, then it implies that $\mu_3 = 0$. Thus it is impossible for the distribution of t^* to satisfy *all* the conditions in (3.16), and so a trade-off must be made between having $\mu_3 = 1$, but $\mu_4 \geq 2$, or having $\mu_4 = 1$, but $\mu_3 = 0$.

Different ancillary distributions for t_i^*

Choosing a distribution to satisfy all four of the moment conditions stated in (3.16) was shown to be impossible and so a compromise needs to be made. In this subsection we shall encounter the different ancillary distributions suggested in the literature from which to generate t_i^* . The ancillary distributions range from simple discrete two point distributions to more complex continuous distributions, all of which attempt to roughly comply with the third and fourth moment conditions. To facilitate referencing of each ancillary distribution in the remainder of the text, the distributions will be designated by F1-F6.

Rademacher distribution (F1) The first distribution we consider is the simple Rademacher two point distribution, which has equal mass at each of the two points -1 and 1 . However, since the third moment of the distribution is zero, symmetry is enforced on the bootstrap error terms, and so it is suggested that one only uses this particular choice when the underlying distribution might be suspected of being symmetric in nature. The Rademacher distribution is:

$$t_i^* = \begin{cases} -1 & \text{w.p. } \frac{1}{2} \\ 1 & \text{w.p. } \frac{1}{2} \end{cases}$$

The moments of this distribution are given by,

$$E(t_i^*) = 0, \quad E(t_i^{*2}) = 1, \quad E(t_i^{*3}) = 0 \quad \text{and} \quad E(t_i^{*4}) = 1.$$

Note that the third moment indicates that the distribution is symmetric, which violates the third moment condition required for the second order accuracy enhancement (Liu, 1988).

Mammen (1993) two-point distribution (F2) This distribution satisfies all three moment conditions suggested by Liu (1988) and has the smallest possible fourth moment given the limitations on the distribution of t_i^* . Because this distribution satisfies the third moment condition the skewness of the bootstrap error terms is the same as the skewness of the residuals. Mammen's (1993) distribution is:

$$t_i^* = \begin{cases} \frac{-(\sqrt{5}-1)}{2} & \text{w.p. } \frac{(\sqrt{5}+1)}{2\sqrt{5}} \\ \frac{(\sqrt{5}+1)}{2} & \text{w.p. } \frac{(\sqrt{5}-1)}{2\sqrt{5}} \end{cases}$$

The moments of this distribution are,

$$E(t_i^*) = 0, \quad E(t_i^{*2}) = 1, \quad E(t_i^{*3}) = 1 \quad \text{and} \quad E(t_i^{*4}) = 2.$$

This distribution manages to satisfy the third moment condition, and achieves the minimum permissible value for the fourth moment.

Mammen (1993) continuous 1 distribution (F3) The next two distribution were suggested by Mammen (1993) and attempt to satisfy the four moment conditions by using a continuous distribution approach. However, we note that these distributions have rather large fourth moments relative to his two point distribution, but still satisfy the first three moment conditions. This distribution is defined as:

$$t_i^* = \frac{u_i}{\sqrt{2}} + \frac{1}{2}(w_i^2 - 1),$$

where u_i and w_i are independent standard normal random variables. The first continuous distribution we consider has moments given by,

$$E(t_i^*) = 0, \quad E(t_i^{*2}) = 1, \quad E(t_i^{*3}) = 1 \quad \text{and} \quad E(t_i^{*4}) = 6.$$

Note that the first three moments are satisfied, but the fourth is quite large.

Mammen (1993) continuous 2 distribution (F4) Mammen's (1993) second continuous distribution is defined as:

$$t_i^* = \left(\delta_1 + \frac{u_i}{\sqrt{2}} \right) \left(\delta_2 + \frac{w_i}{\sqrt{2}} \right) - \delta_1 \delta_2,$$

where u_i and w_i are independent standard normal random variables, $\delta_1 = \left(\frac{3}{4} + \frac{\sqrt{17}}{12} \right)$ and $\delta_2 = \left(\frac{3}{4} - \frac{\sqrt{17}}{12} \right)$. The second continuous distribution discussed by Mammen (1993) also satisfies the first three conditions and attempts to reduce the large fourth moment encountered above,

$$E(t_i^*) = 0, \quad E(t_i^{*2}) = 1, \quad E(t_i^{*3}) = 1 \quad \text{and} \quad E(t_i^{*4}) = 5.625.$$

Webb's (2014) distribution (F5) The following 6 point distribution was advocated for use in clustered wild bootstrap methods, and was shown to have an advantage over two point distributions in a Monte Carlo study by Webb (2014). The distribution is defined as:

$$t_i^* = \begin{cases} -\sqrt{1.5} & \text{w.p. } \frac{1}{6} \\ -1 & \text{w.p. } \frac{1}{6} \\ -\sqrt{0.5} & \text{w.p. } \frac{1}{6} \\ \sqrt{0.5} & \text{w.p. } \frac{1}{6} \\ 1 & \text{w.p. } \frac{1}{6} \\ \sqrt{1.5} & \text{w.p. } \frac{1}{6} \end{cases}.$$

This 6 point distribution has the following moments,

$$E(t_i^*) = 0, \quad E(t_i^{*2}) = 1, \quad E(t_i^{*3}) = 0 \quad \text{and} \quad E(t_i^{*4}) = \frac{7}{6}.$$

The distribution again imposes symmetry and has a slightly larger fourth moment to that of the Rademacher distribution.

Standard normal distribution (F6) Webb (2014) also suggested the use of the standard normal distribution for the clustered wild bootstrap, i.e.,

$$t_i^* \sim N(0, 1).$$

The standard normal distribution has the following well known four moments,

$$E(t_i^*) = 0, \quad E(t_i^{*2}) = 1, \quad E(t_i^{*3}) = 0 \quad \text{and} \quad E(t_i^{*4}) = 3.$$

Due to the plethora of distributions that can be used to generate values for $\{t_i^*\}_{i=1}^n$, choosing which distribution to use may seem overwhelming. The Rademacher and Mammen two point distributions are the most suggested and are extensively studied in the literature (Mammen, 1993; Flachaire, 2005; Davidson and Flachaire, 2008; MacKinnon, 2013). Their popularity is possibly due to their simplicity and desirable moments. A simulation study will be conducted in Chapter 5 using all of the different choices to determine if the use of more complex distributions provide any refinement over the simpler ones.

Different transformations ($f(\cdot)$) of the residuals (e_i)

We now consider three different forms of the transformations $f(\cdot)$ of the residuals e_i which can be used in the DGP described in Equation (3.15). To facilitate referencing in the remainder of the text, the three transformations will be designated by f1-f3.

- **Transformation f1:** The transformation is given by

$$\text{f1} := f(e_i) = \sqrt{\frac{n}{n-k}} e_i.$$

This method of transformation produces a variance/covariance matrix of the estimated regression coefficients similar to HC1.

- **Transformation f2:** The second transformation is related to the studentised residuals, and is given by:

$$\text{f2} := f(e_i) = \frac{e_i}{\sqrt{1-h_i}}.$$

This transformations produces a variance/covariance matrix similar to HC2.

- **Transformation f3:** The final transformation is a variation of the second transformation f2, and produces a variance/covariance matrix similar to HC3:

$$f3 := f(e_i) = \frac{e_i}{1 - h_i}.$$

It is interesting to note that the Monte Carlo approximation of the bootstrap covariance matrix estimator for $\hat{\beta}$ converges to an HCCME, i.e., the following quantity is equivalent to an HCCME for a given choice of the transformation f1, f2, or f3:

$$Var_*(\hat{\beta})_B = \frac{1}{B-1} \sum_{b=1}^B (\hat{\beta}_b^* - \bar{\beta}^*)(\hat{\beta}_b^* - \bar{\beta}^*)',$$

where $\hat{\beta}^*$ is the OLS estimate of the regression coefficients using a wild bootstrap sample and $\bar{\beta}^*$ is the average of the estimated bootstrap regression coefficients. If the transformations f1, f2 or f3 are used, the covariance matrix $Var_*(\hat{\beta})_B$ will converge to the HC1, HC2 or HC3 covariance matrices, respectively, as $B \rightarrow \infty$.

3.4 Bootstrap regression hypothesis testing

This subsection will deal with the application of the bootstrap to regression hypothesis testing for testing the significance of a linear regression model's coefficients. To apply the bootstrap to linear regression hypothesis testing via Monte Carlo simulation, start by applying a DGP to obtain appropriate bootstrap samples. These samples are then used to calculate B test statistics which can be used to approximate the sampling distribution of the test statistic. Finally, we are able to draw conclusions by comparing the observed sample test statistic to that of the critical values obtained from the sampling distribution generated by the bootstrap.

3.4.1 Hypothesis testing with homoskedastic errors: Residuals approach

In order to construct a test for one of the regression coefficients we consider the classical linear regression model with homoskedastic errors given by Equation (2.2),

$$\underset{(n \times 1)}{\mathbf{Y}} = \underset{(n \times k)}{\mathbf{X}} \underset{(k \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}},$$

where $Var(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i \neq j; i, j = 1, 2, \dots, n$. To test the significance for one of the linear regression coefficients, say β_j , define the hypothesis:

$$H_0 : \beta_j = \beta_j^0 \quad \text{vs.} \quad H_A : \beta_j \neq \beta_j^0, \quad (3.18)$$

for any $j = 0, 1, 2, \dots, k - 1$.

A way in which one would test this hypothesis using the residuals based bootstrap is to start by estimating the regression coefficients for the classical model by means of OLS. The residuals $\{e_i\}_{i=1}^n$ are also obtained through the OLS procedure, and these will be resampled to get $\{e_i^*\}_{i=1}^n$. The most important part of this procedure is to ensure that the null hypothesis is enforced in the bootstrap samples. This is accomplished by substituting the resampled residuals into a DGP that enforces the null hypothesis. The DGP will have the form:

$$\mathbf{Y}^{0*} = \mathbf{X}\hat{\boldsymbol{\beta}}^0 + \mathbf{e}^*,$$

where $\mathbf{Y}^{0*} = [Y_1^{0*}, Y_2^{0*}, \dots, Y_n^{0*}]'$ is the bootstrap sample generated under the null hypothesis, $\mathbf{X}$ is the original sample design matrix, $\hat{\boldsymbol{\beta}}^0$ is the vector of estimated regression coefficients where the j^{th} estimator is replaced by the hypothesised value β_j^0 , i.e., $\hat{\boldsymbol{\beta}}^0 = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \beta_j^0, \hat{\beta}_{j+1}, \dots, \hat{\beta}_{k-1}]'$, $j = 1, 2, \dots, k - 1$, and $\mathbf{e}^* = [e_1^*, e_2^*, \dots, e_n^*]'$. The test then follows by estimating the null distribution of the test statistic:

$$T_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\mathbf{S}_{jj}}}, \quad (3.19)$$

where $\mathbf{S}_{jj}$ is the j^{th} diagonal element of the matrix $\mathbf{S} = \widehat{Var}(\hat{\boldsymbol{\beta}}) = \frac{\mathbf{e}'\mathbf{e}}{n-k}(\mathbf{X}'\mathbf{X})^{-1}$. The bootstrap

estimator of this test statistic is given by:

$$T_j^* = \frac{\hat{\beta}_j^{0*} - \beta_j^0}{\sqrt{\mathbf{S}_{jj}^{0*}}}, \quad (3.20)$$

where $\hat{\beta}_j^{0*}$ is the j^{th} OLS estimate for the regression coefficient using the bootstrap sample $\mathbf{Y}^{0*}$, i.e., the j^{th} element of $\hat{\boldsymbol{\beta}}^{0*} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^{0*}$ and $\mathbf{S}_{jj}^{0*}$ is the j^{th} diagonal element of the matrix

$$\mathbf{S}^{0*} = \frac{\boldsymbol{\epsilon}^{0*'}\boldsymbol{\epsilon}^{0*}}{n - k}(\mathbf{X}'\mathbf{X})^{-1},$$

with $\boldsymbol{\epsilon}^{0*} = \mathbf{Y}^{0*} - \mathbf{X}\hat{\boldsymbol{\beta}}^{0*}$. If the assumption of homoskedasticity is deemed valid, the following algorithm could be used to test the hypothesis given in (3.18) with respect to the j^{th} regression coefficient.

Algorithm: Bootstrap residuals algorithm

1. Given the sample data $\mathbf{Y}$ and $\mathbf{X}$, estimate the regression coefficients ($\boldsymbol{\beta}$) using OLS:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Next, compute the residuals $\mathbf{e} = [e_1, e_2, \dots, e_n]'$ using:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

2. Centre the residuals to ensure the bootstrap residuals will reflect the model assumption of being centred around zero:

$$\check{\mathbf{e}} = \mathbf{e} - \mathbf{1}\bar{e},$$

where $\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i$ and $\mathbf{1}_{(n \times 1)}$ is a vector of ones.

3. Sample with replacement from $\check{\mathbf{e}} = [\check{e}_1, \check{e}_2, \dots, \check{e}_n]'$ to obtain the bootstrap residuals $\check{\mathbf{e}}^* = [\check{e}_1^*, \check{e}_2^*, \dots, \check{e}_n^*]'$.
4. Compute the bootstrap sample $\mathbf{Y}^{0*} = [Y_1^{0*}, Y_2^{0*}, \dots, Y_n^{0*}]'$ by substitution of the boot-

strap residuals $\check{\epsilon}^*$ into the DGP:

$$\mathbf{Y}^{0*} = \mathbf{X}\hat{\boldsymbol{\beta}}^0 + \check{\epsilon}^*,$$

where $\hat{\boldsymbol{\beta}}^0$ is the vector of estimated regression coefficient where the j^{th} estimator is replaced by the hypothesised value β_j^0 , i.e., $\hat{\boldsymbol{\beta}}^0 = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \beta_j^0, \hat{\beta}_{j+1}, \dots, \hat{\beta}_{k-1}]'$.

5. Compute the OLS estimates for the bootstrap regression coefficients using the bootstrap sample $\mathbf{Y}^{0*}$, i.e., $\hat{\boldsymbol{\beta}}^{0*} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^{0*} = [\hat{\beta}_0^{0*}, \hat{\beta}_1^{0*}, \dots, \hat{\beta}_{k-1}^{0*}]$.
6. Repeat steps (3)–(5) B times to get the bootstrap samples $\mathbf{Y}_1^{0*}, \mathbf{Y}_2^{0*}, \dots, \mathbf{Y}_B^{0*}$ and the corresponding bootstrap regression coefficient estimates $\hat{\boldsymbol{\beta}}_1^{0*}, \hat{\boldsymbol{\beta}}_2^{0*}, \dots, \hat{\boldsymbol{\beta}}_B^{0*}$.
7. For each of the B replication, calculate the bootstrap test statistic:

$$T_j^*(b) = \frac{\hat{\beta}_{jb}^{0*} - \hat{\beta}_j^0}{\sqrt{\mathbf{S}_{jj}^{0*}(b)}},$$

where $\mathbf{S}_{jj}^{0*}(b)$ is the j^{th} diagonal element in the matrix

$$\mathbf{S}^{0*}(b) = \frac{\boldsymbol{\epsilon}_b^{0*'} \boldsymbol{\epsilon}_b^{0*}}{n - k} (\mathbf{X}'\mathbf{X})^{-1},$$

with $\boldsymbol{\epsilon}_b^{0*} = \mathbf{Y}_b^{0*} - \mathbf{X}\hat{\boldsymbol{\beta}}_b^{0*}$, to get $T_j^*(1), T_j^*(2), \dots, T_j^*(B)$.

8. Compute the equal-tail bootstrap p -value by:

$$p_B^* \text{-value} = 2 \min \left(\frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_j^*(b) \leq T_j), \frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_j^*(b) > T_j) \right),$$

where T_j is defined as the sample test statistic in Equation (3.19).

Remark: Note that the numerator in the test statistic (3.20), $\hat{\beta}_j^{0*} - \beta_j^0$, requires the use of the bootstrap estimator $\hat{\boldsymbol{\beta}}^{0*}$ which was obtained by enforcing the null hypothesis in the bootstrap sample $\mathbf{Y}^{0*}$. However, it can be shown that this numerator is equivalent to $\hat{\beta}_j^* - \hat{\beta}_j$, which does not require a bootstrap sample that needs to be constructed under the null hypothesis. The following proof shows that this is indeed the case.

Proof. First note that the j^{th} element in $\hat{\boldsymbol{\beta}}^{0*} - \hat{\boldsymbol{\beta}}^0$ is $\hat{\beta}_j^{0*} - \beta_j^0$ and that the j^{th} element in $\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}$

is $\widehat{\beta}_j^* - \widehat{\beta}_j$. Therefore, we need to show that $\widehat{\beta}^{0*} - \widehat{\beta}^0 = \widehat{\beta}^* - \widehat{\beta}$. We define the quantities that appear in these expressions as follows:

- Define $\widehat{\beta}$ as:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

i.e., $\widehat{\beta}$ is the OLS estimate using the original sample.

- Define $\widehat{\beta}^0$ as:

$$\widehat{\beta}^0 = [\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_{j-1}, \beta_j^0, \widehat{\beta}_{j+1}, \dots, \widehat{\beta}_{k-1}]',$$

where β_j^0 is the value of the coefficient specified in the hypothesis (3.18) and $\widehat{\beta}_l$, $l = 0, 1, \dots, j-1, j+1, \dots, k-1$ are estimators for the corresponding regression coefficients. Note that for the purposes of hypothesis testing the way in which the estimators $\widehat{\beta}_l$, $l = 0, 1, \dots, j-1, j+1, \dots, k-1$ are obtained is *completely arbitrary* (as will be shown below) but typical choices include using the OLS estimators, $\widehat{\beta}$, of the model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, or by using the OLS estimators based on a model that conforms to the null hypothesis, i.e., $\tilde{\mathbf{Y}} = \mathbf{X}\beta^0 + \tilde{\varepsilon}$, where

$$\beta^0 = [\beta_0, \beta_1, \dots, \beta_{j-1}, \beta_j^0, \beta_{j+1}, \dots, \beta_{k-1}]'.$$

- Define $\widehat{\beta}^{0*}$ as:

$$\widehat{\beta}^{0*} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^{0*},$$

a $(k \times 1)$ vector of the estimated OLS regression coefficients using the the bootstrap sample $\mathbf{Y}^{0*} = \mathbf{X}\widehat{\beta}^0 + \mathbf{e}^*$ forced under the null hypothesis, where the bootstrap residuals vector $\mathbf{e}^*$ is generated by resampling from the unrestricted OLS residuals vector $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\beta}$.

- Define $\widehat{\beta}^*$ as:

$$\widehat{\beta}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^*,$$

where $\mathbf{Y}^* = \mathbf{X}\widehat{\beta} + \mathbf{e}^*$ and bootstrap residuals vector $\mathbf{e}^*$ is generated by resampling from the OLS residuals vector $\mathbf{e} = \mathbf{Y} - \mathbf{X}\widehat{\beta}$.

By expansion of $\widehat{\beta}^{0*} - \widehat{\beta}^0$ and substitution of the defined quantities above, we can prove the

difference vector is equal to $\hat{\beta}^* - \hat{\beta}$:

$$\begin{aligned}
\hat{\beta}^{0*} - \hat{\beta}^0 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^{0*} - \hat{\beta}^0 \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{X}\hat{\beta}^0 + \mathbf{e}^*] - \hat{\beta}^0 \\
&= \hat{\beta}^0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^* - \hat{\beta}^0 \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^*
\end{aligned} \tag{3.21}$$

Similarly, we can show that $\hat{\beta}^* - \hat{\beta}$ has the same form as (3.21).

$$\begin{aligned}
\hat{\beta}^* - \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^* - \hat{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{X}\hat{\beta} + \mathbf{e}^*] - \hat{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^* - \hat{\beta} \\
&= \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^* - \hat{\beta} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^*
\end{aligned} \tag{3.22}$$

It is clear from (3.21) and (3.22) that $\hat{\beta}^* - \hat{\beta} = \hat{\beta}^{0*} - \hat{\beta}^0$. □

This result has a number of implications:

- From (3.21), note that the choice of $\hat{\beta}^0$ is arbitrary since the term falls away entirely in the final expression.
- The “null model” used to generate the bootstrap samples when conducting bootstrap hypothesis testing must contain values that enforce the null hypothesis, but all other coefficients can be estimated arbitrarily. Also, the only strict requirement is that the bootstrap residuals used in the DGP must be drawn from the residuals from the original unrestricted model. In Shao and Tu (1995) a null model that is suggested for generating the bootstrap samples uses coefficient estimates from the reduced/null model but in de Beer and Swanepoel (1989), Paparoditis and Politis (2005) and Martin (2007) the estimated coefficients used in the DGP to generate the bootstrap samples are estimated using the full model. The question around which one of the null models should be used is often asked

but never really addressed. However, it turns out that it is unimportant, e.g., one may use $\widehat{\beta}$, of the model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, or one may use the OLS estimators, $\tilde{\beta}$, based on a model that conforms to the null hypothesis, i.e., $\tilde{\mathbf{Y}} = \mathbf{X}\beta^0 + \tilde{\epsilon}$, where

$$\beta^0 = [\beta_0, \beta_1, \dots, \beta_{j-1}, \beta_j^0, \beta_{j+1}, \dots, \beta_{k-1}]'.$$

- The bootstrap variance of the terms $\widehat{\beta}^* - \widehat{\beta}$ and $\widehat{\beta}^{0*} - \widehat{\beta}^0$ are identical and given by:

$$\begin{aligned} Var_*(\widehat{\beta}^* - \widehat{\beta}) &= Var_*(\widehat{\beta}^{0*} - \widehat{\beta}^0) = Var_*(\widehat{\beta}^*) \\ &= Var_*((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^*) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var_*(\mathbf{Y}^*)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var_*(\mathbf{X}\widehat{\beta} + \mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var_*(\mathbf{e}^*)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

3.4.2 Hypothesis testing with homoskedastic or heteroskedastic errors: Pairs approach

To conduct a test for one of the regression coefficients using the bootstrap pairs approach we will define the following linear regression model:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i, \quad \forall i = 1, 2, \dots, n,$$

where $Var(\varepsilon_i) = \sigma_i^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j; i, j = 1, 2, \dots, n$, however the case where $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2, \quad \forall i = 1, 2, \dots, n$ is also accommodated. The additional assumption that the pairs $(\mathbf{X}_i, Y_i)$ were generated from some unknown multivariate distribution F is also made, which limits the application of the pairs bootstrap to the class of models with random design matrices.

To test the hypothesis statement given by Equation (3.18), the pairs $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ are resampled to obtain the bootstrap pairs $\{(\mathbf{X}_i^*, Y_i^*)\}_{i=1}^n$. The following algorithm describes the process of testing the hypothesis using Monte Carlo.

Algorithm: Bootstrap pairs algorithm

1. Sample with replacement from the pairs $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)$ to obtain the bootstrap pairs sample $(\mathbf{X}_1^*, Y_1^*), (\mathbf{X}_2^*, Y_2^*), \dots, (\mathbf{X}_n^*, Y_n^*)$. Construct the bootstrap response vector $\mathbf{Y}^* = [Y_1^*, Y_2^*, \dots, Y_n^*]'$ and bootstrap design matrix $\mathbf{X}^* = [\mathbf{X}_1^*, \mathbf{X}_2^*, \dots, \mathbf{X}_n^*]'$

2. Estimate the bootstrap regression coefficients and variances using OLS:

$$\hat{\beta}^* = (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \mathbf{Y}^* \text{ and } \mathbf{S}^* = \widehat{Var}(\hat{\beta}^*) = \frac{\boldsymbol{\epsilon}^{*'} \boldsymbol{\epsilon}^*}{n - k} (\mathbf{X}^{*'} \mathbf{X}^*)^{-1},$$

where $\boldsymbol{\epsilon}^* = \mathbf{Y}^* - \mathbf{X}^* \hat{\beta}^*$.

3. Repeat steps (1) and (2) B times to obtain:

$$\hat{\beta}_1^*, \hat{\beta}_2^*, \dots, \hat{\beta}_B^* \text{ and } \mathbf{S}^*(1), \mathbf{S}^*(2), \dots, \mathbf{S}^*(B).$$

4. Calculate the B test statistics for the j^{th} regression coefficient $T_j^*(1), T_j^*(2), \dots, T_j^*(B)$, where

$$T_j^*(b) = \frac{\hat{\beta}_{jb}^* - \hat{\beta}_j}{\sqrt{\mathbf{S}_{jj}^*(b)}},$$

and $\mathbf{S}_{jj}^*(b)$ denotes the j^{th} diagonal element in $\mathbf{S}^*(b)$.

5. Compute the equal-tail bootstrap p -value by:

$$p_B^* \text{-value} = 2 \min \left(\frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_j^*(b) \leq T_j), \frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_j^*(b) > T_j) \right).$$

Another way by which the bootstrap pairs approach can be applied was suggested by Flachaire (1999), since the original pairs bootstrap DGP (3.11) does not explicitly enforce the null hypothesis. The specific hypothesis being tested here is

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_A : \beta_j \neq 0.$$

He proposed to modify the original bootstrap pairs DGP in order to ensure the null distribution is enforced. The modified DGP is defined as,

$$Y_i^* = \mathbf{X}_i^* \tilde{\boldsymbol{\beta}}^0 + \tilde{e}_i^*, \quad (3.23)$$

where $\tilde{\boldsymbol{\beta}}^0$ denotes estimated regression coefficients where the null hypothesis was enforced. In order to construct $\tilde{\boldsymbol{\beta}}^0$, we first need to estimate regression coefficients by means of OLS using a design matrix from which the j^{th} predictor has been omitted, $\mathbf{X}_{(j)}$, to obtain:

$$\tilde{\boldsymbol{\beta}}_{(j)} = \left(\mathbf{X}_{(j)}' \mathbf{X}_{(j)} \right)^{-1} \mathbf{X}_{(j)}' \mathbf{Y} = [\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_{j-1}, \tilde{\beta}_{j+1}, \dots, \tilde{\beta}_{k-1}]'.$$

We can then construct $\tilde{\boldsymbol{\beta}}^0$ by including the hypothesised value β_j^0 in the j^{th} position to get,

$$\tilde{\boldsymbol{\beta}}^0 = [\tilde{\beta}_0, \tilde{\beta}_1, \dots, \tilde{\beta}_{j-1}, \beta_j^0, \tilde{\beta}_{j+1}, \dots, \tilde{\beta}_{k-1}]'.$$

Now, obtain the residuals and leverage values under the alternative using OLS, i.e.,

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = [e_1, e_2, \dots, e_n]' \quad \text{and} \quad h_i = \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i',$$

and then calculate

$$\tilde{e}_i = \frac{e_i}{\sqrt{1-h_i}} - \frac{1}{n} \sum_{s=1}^n \frac{e_s}{\sqrt{1-h_s}}, \quad i = 1, \dots, n.$$

Next, resample from the pairs $\{(\mathbf{X}_i, \tilde{e}_i)\}_{i=1}^n$ to obtain the pairs $\{(\mathbf{X}_i^*, \tilde{e}_i^*)\}_{i=1}^n$. Finally, substitute $\{(\mathbf{X}_i^*, \tilde{e}_i^*)\}_{i=1}^n$ into the modified DGP given by (3.23) to obtain the bootstrap sample, $\{(\mathbf{X}_i^*, Y_i^*)\}_{i=1}^n$.

3.4.3 Hypothesis testing with heteroskedastic errors: Wild bootstrap approach

In order to construct a test for one of the regression coefficients using the wild bootstrap, we again consider the heteroskedastic linear regression model with heteroskedastic errors given by

Equation (2.19), where $Var(\varepsilon_i) = \sigma_i^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j; i, j = 1, 2, \dots, n$. To test the hypothesis in (3.18) will require the following heteroskedastic robust test statistic,

$$T_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_n(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}}}, \quad (3.24)$$

where $\hat{\beta}_j$ denotes the OLS estimate for the regression coefficient β_j ; $\left[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_n(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}$ denotes the j^{th} diagonal element of the heteroskedastic consistent covariance matrix estimator for $Var(\hat{\boldsymbol{\beta}})$, and β_j^0 is the hypothesised value as stated in the hypothesis statement in (3.18). Table 2.1 summarises the different HCCMEs which could be used for the estimator $\hat{\mathbf{V}}_n = \mathbf{X}'\hat{\boldsymbol{\Sigma}}\mathbf{X}$.

The bootstrap test statistic is similarly defined to that of Equation (3.24) and is given by,

$$T_j^* = \frac{\hat{\beta}_j^{0*} - \beta_j^0}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_n^*(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}}}, \quad (3.25)$$

where $\hat{\beta}_j^{0*}$ is the OLS estimate of the regression coefficient using the bootstrap sample $\mathbf{Y}^{0*}$ generated from the DGP

$$Y_i^{0*} = \mathbf{X}_i'\hat{\boldsymbol{\beta}}^0 + f(e_i)t_i^*, \quad i = 1, \dots, n,$$

where $f(\cdot)$ and t_i^* are defined as in Section 3.3.3, on pages 59 and 62, $\hat{\mathbf{V}}_n^* = \mathbf{X}'\hat{\boldsymbol{\Sigma}}^*\mathbf{X}$ is calculated using a HCCME based on the bootstrap residuals $\boldsymbol{\epsilon}^{0*} = \mathbf{Y}^{0*} - \mathbf{X}\hat{\boldsymbol{\beta}}^{0*}$ and β_j^0 is the hypothesised value as stated in the hypothesis statement in (3.18). Recall that $\hat{\boldsymbol{\beta}}^0$ was defined as:

$$\hat{\boldsymbol{\beta}}^0 = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \beta_j^0, \hat{\beta}_{j+1}, \dots, \hat{\beta}_{k-1}]',$$

where $\hat{\beta}_l$, $l = 0, 1, \dots, j-1, j+1, \dots, k-1$ are arbitrary estimators for the remaining regression coefficients.

An algorithm to test the hypothesis stated in (3.18) using the wild bootstrap approach is given below and concludes this chapter.

Algorithm: Algorithm for the wild bootstrap

1. Given the sample data $\mathbf{Y}$ and $\mathbf{X}$, estimate the regression coefficients (β) using OLS:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

and compute the residuals $\mathbf{e} = [e_1, e_2, \dots, e_n]'$ using:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

Also, calculate the sample test statistic T_j as defined in (3.24).

2. The wild bootstrap residuals $\mathbf{e}^* = [e_1^*, e_2^*, \dots, e_n^*]'$ are generated by:

$$e_i^* = t_i^* f(e_i),$$

where $\{t_i^*\}_{i=1}^n$ is an *i.i.d.* sequence that is independent of $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ and is generated from one of the ancillary distributions and $f(\cdot)$ is one of the residual transformations as defined in Section 3.3.3. Calculate the wild bootstrap sample $\mathbf{Y}^{0*}$ by using the wild bootstrap DGP,

$$Y_i^{0*} = \mathbf{X}_i' \hat{\beta}^0 + e_i^*, i = 1, 2, \dots, n.$$

3. Estimate the wild bootstrap regression coefficients and heteroskedastic consistent estimates using OLS:

$$\hat{\beta}^{0*} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^{0*} \text{ and } \hat{\mathbf{V}}_n^* = \mathbf{X}'\hat{\Sigma}^*\mathbf{X},$$

where $\epsilon^{0*} = \mathbf{Y}^{0*} - \mathbf{X}\hat{\beta}^{0*}$ are the residuals used to estimate the heteroskedastic consistent estimate chosen from Table 2.1.

4. Repeat steps (2) and (3) B times to obtain:

$$\hat{\beta}_1^{0*}, \hat{\beta}_2^{0*}, \dots, \hat{\beta}_B^{0*} \text{ and } \hat{\mathbf{V}}_{n,1}^*, \hat{\mathbf{V}}_{n,2}^*, \dots, \hat{\mathbf{V}}_{n,B}^*.$$

5. Calculate the wild bootstrap test statistic using the bootstrap sample and original design matrix $\mathbf{X}$:

$$T_j^* = \frac{\hat{\beta}_{jb}^{0*} - \beta_j^0}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_{n,b}^*(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}}},$$

where $\hat{\beta}_{jb}^{0*}$ is the j^{th} wild bootstrap OLS regression coefficient estimate, and the denominator is the j^{th} diagonal element of the heteroskedastic consistent covariance estimate, whereby $\hat{\mathbf{V}}_{n,b}^*$ was calculated using the same form of the HCCME in step (3).

6. Repeat steps (2) – (5) B times to obtain $T_j^*(1), T_j^*(2), \dots, T_j^*(B)$.

7. Compute the equal-tail bootstrap p -value by:

$$p_B^* \text{-value} = 2 \min \left(\frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_j^*(b) \leq T_j), \frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_j^*(b) > T_j) \right).$$

CHAPTER 4

Proposed new bootstrap approach for dealing with heteroskedasticity

In addition to using HCCMEs, and the wild bootstrap to conduct hypothesis tests in the presence of heteroskedasticity, this chapter will propose a modification of the residuals approach whereby the heteroskedastic structure is estimated using a linear approximation. This approach may seem trivial but it is new to the literature and the aim is to investigate the performance thereof relative to established methods via a Monte Carlo study in order to determine if the approach has any merit.

4.1 Modified bootstrap residuals approach for dealing with heteroskedasticity

A new approach is suggested to overcome the limitations associated with using the bootstrap residuals approach in a heteroskedastic errors setting. The reason for the modification of the bootstrap residuals approach is to ensure that when the residuals are resampled that the original heteroskedastic structure is still present. The problem with resampling from the OLS residuals is that the original heteroskedastic structure is not retained within the bootstrap samples.

The proposed procedure consists of two separate stages:

1. Estimate the heteroskedastic structure by fitting a linear regression model using the original predictors and with the squared OLS residuals as the response variable. The idea is to estimate the individual variances using this model and then rescale the OLS residuals to have unit variance; in so doing we remove the heteroskedastic structure from the sample residuals.

2. Next, resample from the scaled residuals and then reintroduce the heteroskedastic structure.

The resulting bootstrap samples will retain the heteroskedastic structure of the original data and will also possess sufficient randomness.

We begin by defining the heteroskedastic linear regression model (2.19) as:

$$Y_i = \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i, i = 1, \dots, n,$$

where the error terms ε_i are heteroskedastic, i.e., $Var(\varepsilon_i) = \sigma_i^2 \forall i = 1, 2, \dots, n$, and $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$. We shall assume that the heteroskedastic structure is defined to be of the general form:

$$\sigma_i^2 = m(\mathbf{X}_i, \boldsymbol{\gamma}), \quad (4.1)$$

where $m(\cdot)$ denotes some function of the original predictors $\mathbf{X}_i$ and unknown parameters $\boldsymbol{\gamma}$. However, since σ_i^2 is unknown it requires an estimator. In Section 2.4.4 we noted that σ^2 is the expected value of ε_i^2 , and argued that since e_i serves as an estimate for ε_i , e_i^2 could naturally serve as an estimate for ε_i^2 . Similarly, e_i^2 will now act as a rough estimate for σ_i^2 in order to model the heteroskedastic structure.

Estimate the linear regression model using OLS to obtain the residuals e_i ,

$$e_i = Y_i - \mathbf{X}_i' \hat{\boldsymbol{\beta}}, i = 1, \dots, n.$$

Then proceed to square the OLS residuals $\{e_i^2\}_{i=1}^n$ and use them as a proxy for the variances in the heteroskedasticity expression given by (4.1). The parameters $\boldsymbol{\gamma}$ in the heteroskedastic model are then estimated by a suitable procedure, and the resulting estimators S_i^2 are obtained:

$$S_i^2 \equiv \hat{e}_i^2 = m(\mathbf{X}_i, \hat{\boldsymbol{\gamma}}), i = 1, \dots, n, \quad (4.2)$$

specifically we shall only consider the following linear model:

$$S_i^2 = \mathbf{X}_i' \hat{\boldsymbol{\gamma}}.$$

After modelling the heteroskedastic structure we would like to perform a residual based bootstrap, however as noted in Section 3.3.1 this is not possible if the errors are not homoskedastic. The structure modelled in Equation (4.2) will now be used to “remove” the heteroskedasticity structure by calculating,

$$\check{e}_i = \frac{e_i}{\sqrt{S_i^2}} = \frac{e_i}{S_i}, i = 1, \dots, n.$$

Sampling with replacement from the scaled “homoskedastic” sample $\{\check{e}_i\}_{i=1}^n$, we obtain the sample $\{\check{e}_i^*\}_{i=1}^n$. Now the heteroskedastic structure is reintroduced into the resampled homoskedastic sample $\{\check{e}_i^*\}_{i=1}^n$ by calculating,

$$\check{e}_i^* = \check{e}_i^* S_i, i = 1, \dots, n.$$

We now finally have the bootstrap residuals for this approach. The bootstrap sample can then be calculated using the DGP,

$$Y_i^* = \mathbf{X}_i' \hat{\boldsymbol{\beta}} + \check{e}_i^*, i = 1, \dots, n, \quad (4.3)$$

from which we can calculate the bootstrap regression coefficients $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^*$. A typical Monte Carlo procedure would then entail independently repeating the above process, say B times, to obtain the bootstrap replicates $\hat{\boldsymbol{\beta}}_1^*, \hat{\boldsymbol{\beta}}_2^*, \dots, \hat{\boldsymbol{\beta}}_B^*$, which can be used to approximate distributional properties of the bootstrap statistic.

On the other hand, if one would like to test the hypothesis:

$$H_0 : \beta_j = \beta_j^0 \quad \text{vs.} \quad H_A : \beta_j \neq \beta_j^0,$$

for any $j = 0, 1, 2, \dots, k-1$. The way in which to proceed using the modified residuals bootstrap approach would be to use the DGP:

$$Y_i^* = \mathbf{X}_i' \hat{\boldsymbol{\beta}}^0 + \epsilon_i^*, \quad i = 1, \dots, n,$$

where $\hat{\boldsymbol{\beta}}^0 = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \beta_j^0, \hat{\beta}_{j+1}, \dots, \hat{\beta}_{k-1}]'$. The Monte Carlo algorithm for testing this hypothesis is given below.

Algorithm: Modified bootstrap residual algorithm

1. Fit the linear regression model using OLS to obtain the regression coefficients estimates,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

and the resulting residuals

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = [e_1, e_2, \dots, e_n]'$$

Also calculate the test statistic

$$T_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1} \hat{\mathbf{V}}_n (\mathbf{X}'\mathbf{X})^{-1} \right]_{jj}}},$$

where $\hat{\mathbf{V}}_n = \mathbf{X}'\hat{\boldsymbol{\Sigma}}\mathbf{X}$ was calculated using one of the HCCMEs in Table 2.1 with the OLS residuals $\{e_i\}_{i=1}^n$.

2. Model the heteroskedastic structure by using the squared residuals,

$$S_i^2 \equiv \hat{e}_i^2 = \hat{\gamma}_0 + \hat{\gamma}_1 X_{i1} + \hat{\gamma}_2 X_{i2} + \dots + \hat{\gamma}_{k-1} X_{ik-1},$$

then “remove” the heteroskedasticity from the residuals using the square root of the

fitted values via the equation:

$$\check{e}_i = \frac{e_i}{\sqrt{S_i^2}} = \frac{e_i}{S_i}, i = 1, 2, \dots, n.$$

3. Sample with replacement from the homoskedastic residuals $\{\check{e}_i\}_{i=1}^n$ to obtain the sample $\{\check{e}_i^*\}_{i=1}^n$. Then reintroduce the heteroskedasticity:

$$\check{\epsilon}_i^* = \check{e}_i^* \hat{S}_i, i = 1, \dots, n,$$

to obtain the bootstrap residuals. The bootstrap sample under the null hypothesis was then calculated using the DGP:

$$Y_i^{0*} = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_{j-1} X_{ij-1} + \beta_j^0 X_{ij} + \hat{\beta}_{j+1} X_{ij+1} + \dots + \hat{\beta}_{k-1} X_{ik-1} + \check{\epsilon}_i^*,$$

for $i = 1, \dots, n$.

4. Compute the OLS estimates for the bootstrap regression coefficients using the bootstrap sample $\mathbf{Y}^{0*}$, i.e., $\hat{\beta}^{0*} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^{0*} = [\hat{\beta}_0^{0*}, \hat{\beta}_1^{0*}, \dots, \hat{\beta}_{k-1}^{0*}]$.
5. Repeat steps (3) and (4) B times to get the bootstrap samples $\mathbf{Y}_1^{0*}, \mathbf{Y}_2^{0*}, \dots, \mathbf{Y}_B^{0*}$ and the corresponding bootstrap regression coefficient estimates $\hat{\beta}_1^{0*}, \hat{\beta}_2^{0*}, \dots, \hat{\beta}_B^{0*}$.
6. Calculate the bootstrap test statistic:

$$T_j^*(b) = \frac{\hat{\beta}_{j,b}^{0*} - \beta_j^0}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_{n,b}^*(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}}},$$

where $\hat{\mathbf{V}}_{n,b}^* = \mathbf{X}'\hat{\Sigma}_b^*\mathbf{X}$ was calculated using one of the HCCMEs in Table 2.1 with the residuals $\check{\epsilon}^{0*} = \mathbf{Y}^{0*} - \mathbf{X}\hat{\beta}^{0*}$ to obtain $T_j^*(1), T_j^*(2), \dots, T_j^*(B)$.

7. Compute the equal-tail bootstrap p -value

$$p_B^* \text{-value} = 2 \min \left(\frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_j^*(b) \leq T_j), \frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_j^*(b) > T_j) \right).$$

Remark: For the purposes for this text, we shall only consider a linear model for the function $m(\cdot)$ in Equation (4.2) due to the following reasons:

- Simplicity of implementation since a user merely applies an additional linear model,
- a linear approximation can serve as a good approximation to more complex functions e.g. Taylor series expansion, and
- if a simple linear model can be shown to work well then other more complex methods may be considered for future research.

Other approaches for estimating the heteroskedastic structures in the Equation (4.2) are commonly found in the literature as tests for heteroskedasticity, and in the next section we will briefly discuss these models.

4.2 Other models for the heteroskedastic structure

In many tests for heteroskedasticity a model based approach is employed to test for the significance of model parameters to determine if a heteroskedastic structure is possibly present. In this section a few different models used in tests for heteroskedasticity are introduced which might prove useful for future research when applied to the modified bootstrap residuals approach discussed in the previous section.

4.2.1 Park test

The Park test (1966) proposes that heteroskedasticity may be modelled using the functional form:

$$\sigma_i^2 = \sigma^2 X_i^\gamma \exp(v_i)$$

or

$$\ln(\sigma_i^2) = \ln(\sigma^2) + \gamma \ln(X_i) + v_i,$$

where $\ln(\sigma^2)$ denotes a constant term, $\gamma \ln(X_i)$ is the functional form of the heteroskedasticity and v_i denotes an error term. The squared residuals e_i^2 are used as a substitute for the error

variance for estimation. Note that since one applies the natural logarithm, the predictors are required to be positive.

4.2.2 Glejser test

Glejser (1969) suggested that heteroskedasticity may be modelled by a number of different forms, but uses the absolute value of the residuals $|e_i|$ rather than the squared residuals as substitute for estimation of the following models:

$$\begin{aligned}\sigma_i^2 &= \gamma_0 + \gamma_1 X_i + v_i, \\ \sigma_i^2 &= \gamma_0 + \gamma_1 \sqrt{X_i} + v_i, \\ \sigma_i^2 &= \gamma_0 + \gamma_1 \frac{1}{X_i} + v_i, \\ \sigma_i^2 &= \gamma_0 + \gamma_1 \frac{1}{\sqrt{X_i}} + v_i, \\ \sigma_i^2 &= \sqrt{\gamma_0 + \gamma_1 X_i} + v_i, \text{ and} \\ \sigma_i^2 &= \sqrt{\gamma_0 + \gamma_1 X_i^2} + v_i.\end{aligned}$$

Again we notice that most of these models require nonnegative or nonzero predictors, with the exception of the first.

4.2.3 Goldfeld-Quandt test

The Goldfeld-Quandt test (1965) assumes that the error variances are positively related to one of the predictor variables and are modelled by:

$$\sigma_i^2 = \sigma^2 X_i^2,$$

where σ^2 is a constant. The Goldfeld-Quandt test therefore assumes that the error variances σ_i^2 are proportional to the square of a predictor variable X . This methods does not use a substitute for the error variances σ_i^2 but rather a ranking approach to test if the proposed model is true – see Gujarati (2004) for more details.

4.2.4 Breusch-Pagan-Godfrey test

The Breusch-Pagan-Godfrey test (1979) models the error variances as:

$$\sigma_i^2 = f(\gamma_0 + \gamma_1 Z_{i1} + \cdots + \gamma_{k-1} Z_{ik-1}),$$

where Z_j denotes fixed predictors, which may include all or some of the predictors depending on their nature and $f(\cdot)$ denotes some function. Note that the modified bootstrap residuals approach described in this Chapter uses this approach and chooses $f(\cdot)$ as the identity function. The substitute for the error variances σ_i^2 is given by:

$$u_i = \frac{e_i^2}{\frac{1}{n} \sum_{i=1}^n e_i^2}.$$

4.2.5 White's general test

White's (1980) test involves modelling the error variances σ_i^2 by employing the original predictors, their squared values and cross products thereof. If we consider a two predictor model the errors variance are defined to be:

$$\sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i1}^2 + \gamma_4 X_{i2}^2 + \gamma_5 X_{i1} X_{i2} + v_i.$$

Again the squared residuals are used as a proxy for the error variance σ_i^2 , and proceeds by estimating the model above by means of OLS.

4.2.6 Koenker-Bassett test

The Koenker-Bassett test (1982) is different from all the other tests in that it does not use the predictor to model the error variances σ_i^2 , but rather uses the squared fitted values $\hat{Y}$ from the fitted OLS model, i.e.,

$$\sigma_i^2 = \gamma_0 + \gamma_1 \hat{Y}_i^2 + v_i.$$

The model can then be estimated using OLS and the squared residuals as the proxy for σ_i^2 .

CHAPTER 5

Simulation study

In order to compare the performance of the tests described in Chapters 3 and 4, a Monte Carlo simulation study was conducted where data was generated under structured heteroskedasticity. A multiple linear regression model was fitted to the generated data and a hypothesis test was conducted for a regression coefficient parameter. Each of the various bootstrap tests introduced in Chapters 3 and 4 for testing hypotheses relating to the regression coefficients in the presence of heteroskedasticity were compared, and their performance in terms of size and power were evaluated.

The chapter starts with Section 5.1 which introduces the heteroskedastic data generating process and specific settings for the simulation study. Section 5.2 consists of defining the hypothesis of interest and corresponding test statistic calculated for this simulation study. Next, in Section 5.3 algorithms that outline how size and power were approximated for the various bootstrap tests are introduced. Finally, Section 5.4 presents the simulation results from the extended simulation study with a brief overview of the noteworthy features from the size and power output.

5.1 Heteroskedastic data generation procedure

The Monte Carlo simulations for this study involved the generation of stochastic and independent predictor variables which, when combined with a generated heteroskedastic error structure, formed a response variable. The error structure was generated according to the convention proposed by MacKinnon (2002) and used by others (Hausman and Palmer, 2012; MacKinnon, 2013; Politis and Poulis, 2014). This model for the heteroskedastic structure contains a parameter which allowed for adjustment of the level of heteroskedasticity present in the errors. After the heteroskedastic data were generated, each wild bootstrap approach from Chapter 3

and the modified bootstrap residuals approach from Chapter 4 were applied to the same array of heteroskedastic data for the purpose of comparison. This design allowed for the study of statistical significance across a range of differing heteroskedasticity levels. Further, it allowed one to pinpoint a specific heteroskedasticity level where a subset of tests achieved the specified significance level ($\alpha = 0.05$) under the null hypothesis, thus allowing for a fair comparison of the power of these tests. Comparing power among test that do not achieve the same significance level under the null hypothesis would typically distort the comparison.

The data generation model can be formally represented by:

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + u_i, \quad u_i = \sigma_i \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \quad (5.1)$$

where p denotes the number of predictor variables, each of which are independent standard lognormal random variables. The standard deviations of the errors u_i are given by

$$\sigma_i = z(\lambda) \left(\beta_0 + \sum_{j=1}^p \beta_j X_{ij} \right)^\lambda,$$

where $z(\lambda)$ denotes a scaling factor that ensures that the average variance of u_i is equal to one. The parameter λ is used to adjust the level of heteroskedasticity and does not, on average, change the variance of u_i (MacKinnon, 2013). The suggested range for λ is $0 \leq \lambda \leq 2$, where $\lambda = 0$ represents homoskedasticity and $\lambda \gg 1$ represents extreme heteroskedasticity.

Unfortunately, the specifics around constructing or choosing the scaling factor $z(\lambda)$ was not included in any of the articles consulted (MacKinnon, 2002; Hausman and Palmer, 2012; MacKinnon, 2013; Politis and Poulis, 2014), and so, to ensure that the condition would be met, the following scaling factor was devised:

$$z(\lambda) = \sqrt{\frac{1}{\frac{1}{n} \sum_{i=1}^n S_i^2}}, \quad \text{where} \quad S_i = \left(\beta_0 + \sum_{j=1}^3 \beta_j X_{ij} \right)^\lambda. \quad (5.2)$$

Each of the predictor variables in Equation (5.1) is independently generated either from a standard lognormal distribution (to ensure that heteroskedastic robust inference is as challenging as possible — since most HCCMEs are sensitive to large leverage values), or from a uniform distribution (a less challenging situation). Generation from the standard lognormal distribution will produce predictors containing a few high leverage values (h_i) when the sample size is small, and the highest leverage value¹ h_{max} will decrease slowly only as $n \rightarrow \infty$.

5.1.1 Data generation design for this simulation study

A three predictor variable data generation model was chosen for investigation due to the computational costs incurred when attempting all possible combinations of wild bootstrap test approaches and HCCMEs that can appear in the test statistic. The model is then:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i, \quad i = 1, 2, \dots, n, \quad (5.3)$$

where the errors were defined as $u_i = \sigma_i \varepsilon_i$, with

$$\sigma_i = z(\lambda)(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3})^\lambda,$$

where $\varepsilon_i \sim N(0, 1)$, and the fixed regression coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ specified as constants in the simulation study. The scaling factor $z(\lambda)$ that was used in the simulation is given by,

$$z(\lambda) = \sqrt{\frac{1}{\frac{1}{n} \sum_{i=1}^n S_i^2}}, \quad \text{where} \quad S_i = (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3})^\lambda.$$

The predictor variables that were used in this simulation study were independently generated from either a standard lognormal or standard uniform distribution. As previously noted the motivation for using lognormal predictors is that when the sample size is small the maximum leverage value h_{max} will typically be fairly large, thus the sensitivity of the HCCMEs to these

¹The term h_{max} is defined as $h_{max} = \max(h_i) \forall i = 1, 2, \dots, n$.

values could be investigated. The standard uniform predictors were included as well to investigate the performance of the bootstrap approaches without any extreme leverage values.

Figure 5.1 is a simple illustration of the form of the heteroskedasticity that one can expect by varying the parameter λ in the data generation model (5.1) for a simple linear regression model. The predictor variable was generated from both a standard lognormal and standard uniform distribution. Note that for $\lambda = 0$ we obtain a homoskedastic structure and for $\lambda = 1$ we start to see a distinct heteroskedastic structure. The most extreme heteroskedastic structures possible is present when $\lambda = 2$. In addition, note that the data using the lognormal predictor produced a small number of extreme values (high leverages), whereas these kinds of extreme values did not appear for the uniform predictor.

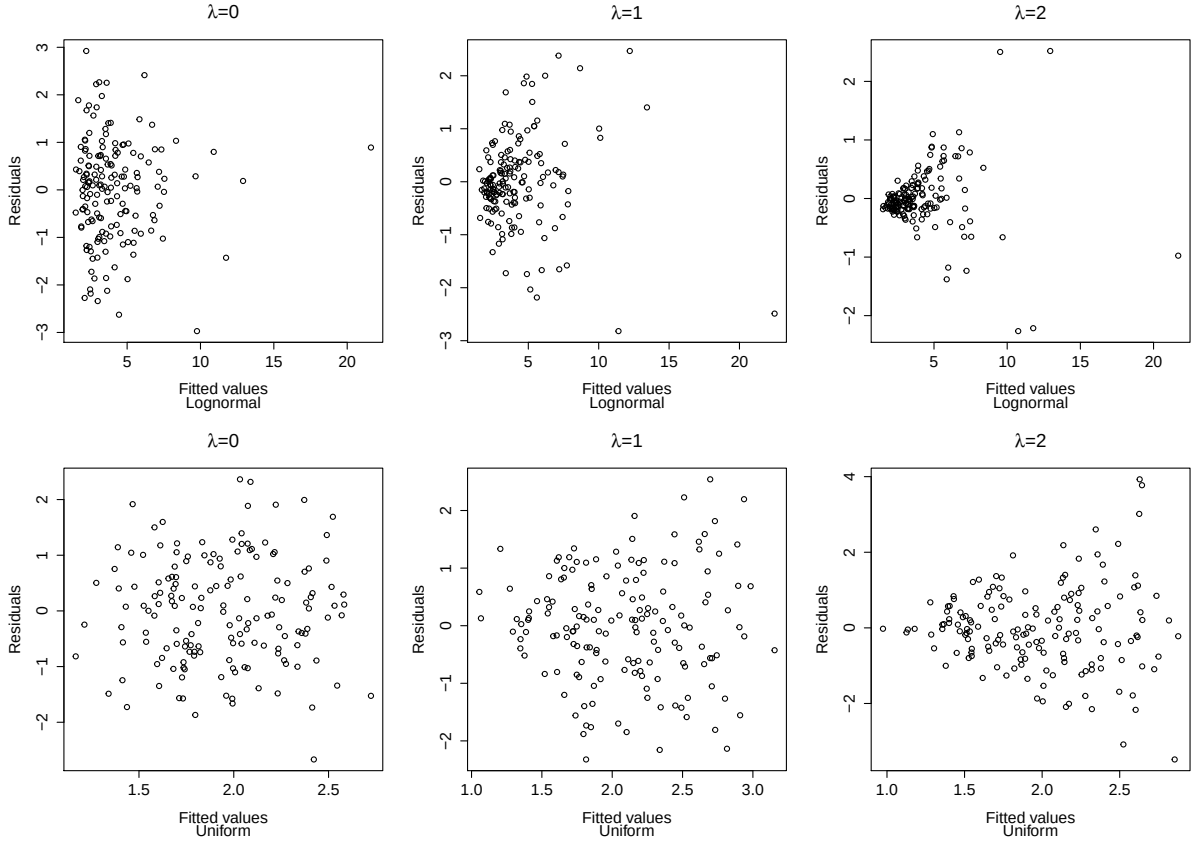


Fig. 5.1: Examples of homoskedasticity and heteroskedasticity produced using model (5.1) with a sample size of $n = 160$.

Data generation settings To generate heteroskedastic data in each of the iterations of the Monte Carlo simulations discussed in the next section, the model in (5.3) was used with different

combinations of the following specifications:

- Set the regression coefficients to $\{\beta_0, \beta_1, \beta_2, \beta_3\} = \{1, 1, 1, 0\}$,
- choose sample sizes of $n = \{20, 40, 80, 160, 320, 640, 1\,280\}$,
- independently generate all three predictor variables (X_{i1}, X_{i2}, X_{i3}) from either a standard lognormal distribution or standard uniform distribution,
- choose the heteroskedasticity parameter λ on the grid of values $\lambda = \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$, and
- the number of Monte Carlo replications was set to $MC = 10\,000$.

The data used in the Monte Carlo simulations will be generated by the specific settings stated above. In the next section we shall discuss the specifics around bootstrap tests and the various bootstrap approaches which used this heteroskedastic data to test a hypothesis concerning one of the regression coefficients.

5.2 Hypothesis used and test statistics for this simulation study

This section will deal with the various procedures that were conducted in the simulation study for testing a hypothesis for one of the regression coefficients. We begin by first defining the hypothesis statement that was tested by all the bootstrap approaches, and then providing the corresponding test statistics which were used.

All the different bootstrap approaches considered in this simulation study tested the following hypothesis around the regression coefficient associated with X_3 ,

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_A : \beta_3 \neq 0, \quad (5.4)$$

for the multiple linear regression model given by,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + u_i, \quad i = 1, 2, \dots, n.$$

Note that this simulation set-up had two parts, the first was used to approximate the size of the test and so we generated the data by setting $\beta_3 = 0$; the null hypothesis in (5.4) should thus not be rejected. The second part involved the calculation of power, and there we generated the value of β_3 according to a small set of possible “alternative” values, e.g. for $n = 80$ and uniform predictors, β_3 was chosen on a grid of values starting at -2.2 in increments of 0.275 up to 2.2.

The wild bootstrap approach in Chapter 3 and the modified bootstrap residuals approach from Chapter 4 were implemented to test the hypothesis in (5.4). This involved first calculating the following sample test statistic using the generated sample data:

$$T_3 = \frac{\hat{\beta}_3 - \beta_3^0}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_n(\mathbf{X}'\mathbf{X})^{-1}\right]_{33}}},$$

where $\hat{\beta}_3$ is the regression coefficient that was estimated using OLS, $\hat{\mathbf{V}}_n$ is a heteroskedasticity consistent estimate using a HCCME with the OLS residuals $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, and $\beta_3^0 = 0$. The corresponding bootstrap test statistics were then defined as:

$$T_3^* = \frac{\hat{\beta}_3^{0*} - \beta_3^0}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_n^*(\mathbf{X}'\mathbf{X})^{-1}\right]_{33}}},$$

where $\hat{\beta}_3^{0*}$ is the regression coefficient estimated using OLS with the bootstrap sample $\mathbf{Y}^{0*}$ that was generated under the null distribution, $\hat{\mathbf{V}}_n^*$ is a heteroskedasticity consistent estimate using a HCCME with the OLS residuals $\boldsymbol{\epsilon}^{0*} = \mathbf{Y}^{0*} - \mathbf{X}\hat{\boldsymbol{\beta}}^{0*}$, and $\beta_3^0 = 0$.

In the next subsection the algorithms for conducting the hypothesis test using the wild bootstrap approaches and modified bootstrap residuals approach will be presented.

5.2.1 Algorithms used for this simulation study

The wild bootstrap test statistics and corresponding p -value were calculated using the following algorithm.

Algorithm: Implemented wild bootstrap algorithm

1. Fit the linear regression model using OLS to obtain the regression coefficients estimates,

$$\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]',$$

and the resulting residuals,

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = [e_1, e_2, \dots, e_n]'$$

Also calculate the sample test statistic

$$T_3 = \frac{\hat{\beta}_3}{\sqrt{[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_n(\mathbf{X}'\mathbf{X})^{-1}]_{33}}},$$

where $\hat{\mathbf{V}}_n = \mathbf{X}'\hat{\boldsymbol{\Sigma}}\mathbf{X}$ was calculated using one of the HCCMEs in Table 5.1 with the OLS residuals $\mathbf{e}$.

2. Generate the wild bootstrap sample $\mathbf{Y}^{0*}$ using the wild bootstrap DGP that enforces the null hypothesis:

$$Y_i^{0*} = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + f(e_i)t_i^*, i = 1, \dots, n.$$

The various transformations $f(\cdot)$ and ancillary distributions t_i^* used are summarised in Table 5.1.

3. Calculate the wild bootstrap test statistic by estimating the regression coefficients using OLS with the wild bootstrap sample $\mathbf{Y}^{0*}$ and the original design matrix $\mathbf{X}$,

$$T_3^* = \frac{\hat{\beta}_3^{0*}}{\sqrt{[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_n^*(\mathbf{X}'\mathbf{X})^{-1}]_{33}}},$$

where $\hat{\mathbf{V}}_n^* = \mathbf{X}'\hat{\boldsymbol{\Sigma}}^*\mathbf{X}$ was calculated using one of the HCCMEs in Table 5.1 with the

OLS residuals $\epsilon^{0*} = \mathbf{Y}^{0*} - \mathbf{X}\hat{\beta}^{0*}$.

4. Repeat steps (2) and (3) B times to obtain: $T_3^*(1), T_3^*(2), \dots, T_3^*(B)$.

5. Compute the equal-tail bootstrap p -value:

$$p_B^* \text{-value} = 2 \min \left(\frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_3^*(b) \leq T_3), \frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_3^*(b) > T_3) \right).$$

The modified bootstrap residuals approach was applied using the algorithm.

Algorithm: Newly proposed modified bootstrap residuals approach algorithm

1. Fit the linear regression model using OLS to obtain the regression coefficients estimates,

$$\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]'$$

and the resulting residuals

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta} = [e_1, e_2, \dots, e_n]'$$

Also calculate the test statistic

$$T_3 = \frac{\hat{\beta}_3}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1} \hat{\mathbf{V}}_n (\mathbf{X}'\mathbf{X})^{-1} \right]_{33}}},$$

where $\hat{\mathbf{V}}_n = \mathbf{X}'\hat{\Sigma}\mathbf{X}$ was calculated using one of the HCCMEs in Table 5.1 with the OLS residuals $\mathbf{e}$.

2. Model the heteroskedastic structure by using the squared residuals,

$$S_i^2 \equiv e_i^2 = \hat{\gamma}_0 + \hat{\gamma}_1 X_{i1} + \hat{\gamma}_2 X_{i2} + \hat{\gamma}_3 X_{i3},$$

then “remove” the heteroskedasticity from the OLS residuals using the square root of

the fitted values via the equation:

$$\check{e}_i = \frac{e_i}{\sqrt{S_i^2}} = \frac{e_i}{S_i}, i = 1, 2, \dots, n.$$

3. Sample with replacement from the homoskedastic residuals $\{\check{e}_i\}_{i=1}^n$ to obtain the sample $\{\check{e}_i^*\}_{i=1}^n$. Then reintroduce the heteroskedasticity:

$$\check{\epsilon}_i^* = \check{e}_i^* S_i, i = 1, \dots, n,$$

to obtain the bootstrap residuals. The bootstrap sample under the null hypothesis was then calculated using the DGP:

$$Y_i^{0*} = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \check{\epsilon}_i^*, i = 1, \dots, n.$$

4. Calculate the bootstrap test statistic by again estimating the regression coefficients using OLS using the bootstrap sample $\mathbf{Y}^{0*}$ and the original design matrix $\mathbf{X}$,

$$T_3^* = \frac{\hat{\beta}_3^{0*}}{\sqrt{\left[(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{V}}_n^*(\mathbf{X}'\mathbf{X})^{-1}\right]_{33}}},$$

where $\hat{\mathbf{V}}_n^* = \mathbf{X}'\hat{\Sigma}^*\mathbf{X}$ was calculated using one of the HCCMEs in Table 5.1 with the OLS residuals $\epsilon^{0*} = \mathbf{Y}^{0*} - \mathbf{X}\hat{\beta}^{0*}$.

5. Repeat steps (3) and (4) B times to obtain $T_3^*(1), T_3^*(2), \dots, T_3^*(B)$.
6. Compute the equal-tail bootstrap p -value

$$p_B^* \text{-value} = 2 \min \left(\frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_3^*(b) \leq T_3), \frac{1}{B} \sum_{b=1}^B \mathbf{I}(T_3^*(b) > T_3) \right).$$

5.3 Monte Carlo approximation of size and power of a test

In order to evaluate the performance of the different bootstrap approaches that were used in the Monte Carlo study, the size and the power of a test was investigated. The size and power of a

test are defined as the following two probabilities:

- **The size of a test:** The probability of rejecting the null hypothesis when the null hypothesis is true, i.e., the probability of a type I error, given by

$$P(\text{reject } H_0 \mid H_0 \text{ is true}).$$

- **The power of a test:** The probability of rejecting the null hypothesis when the alternative is true, i.e., the probability of not obtaining a type II error, given by

$$P(\text{reject } H_0 \mid H_A \text{ is true}).$$

The Monte Carlo approximation of these two quantities involves generating multiple samples from a theoretical distribution, and then evaluating the test for each of these replications. The way in which the data should be generated will differ depending on whether you are trying to approximate the size or the power of a test.

- For the case of approximating the size of the test the data are generated in each Monte Carlo replication such that it satisfies the *null hypothesis*.
- When approximating the power of a test the data are generated in such a way that it satisfies a specific setting of the *alternative hypothesis*.

The procedure follows by calculating the test statistic from these data, evaluating the validity of the null hypothesis (i.e., either rejecting or not rejecting the null hypothesis), and then determining the proportion of the Monte Carlo replications that resulted in the rejection of the null hypothesis.

An ideal test would produce a size of the test that is equal to the specified significance level $\alpha = 0.05$, and also provide a level of power which is high across different settings of the alternative hypothesis.

The criteria for selecting the “best” test in the presence of heteroskedasticity followed the guideline above by investigating the size of the different bootstrap tests (the number of bootstrap

replications was set to $B = 500$) and identifying a point where a subset of the tests achieved the same significance level. Since these tests all had similar rejection frequencies at the specific point of λ they could be compared in a fair manner. The test with the highest power across alternative settings was then considered the “best” for heteroskedasticity robust inference. All simulations were conducted using the number of Monte Carlos set to $MC = 10\,000$ with R (2015) statistical software on an Intel® Core™ i7-4790 CPU @3.60GHz and 16 gigabytes of system memory, with additional packages `lmtest` (Zeileis and Hothorn, 2002), `sandwich` (Zeileis, 2004) and `snowfall` (Knaus, 2013).

Table 5.1: Different wild bootstrap residual transformations, HCCMEs and ancillary distributions used with corresponding shorthand notation.

Name	Formula	Shorthand notation
Choice of transformation		
HC2 transformation	$f2(u_i) = \frac{u_i}{\sqrt{1-h_i}}$	f2
HC3 transformation	$f3(u_i) = \frac{u_i}{1-h_i}$	f3
Choice of HCCME		
HC2	$\hat{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^n \frac{u_i^2}{1-h_i} \mathbf{X}_i \mathbf{X}_i'$	HC2
HC3	$\hat{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^n \frac{u_i^2}{(1-h_i)^2} \mathbf{X}_i \mathbf{X}_i'$	HC3
Choice of ancillary distribution		
Rademacher	$t_i^* = \begin{cases} -1 & \text{w.p. } \frac{1}{2} \\ 1 & \text{w.p. } \frac{1}{2} \end{cases}$	F1
Mammen's two-point	$t_i^* = \begin{cases} \frac{-(\sqrt{5}-1)}{2} & \text{w.p. } \frac{(\sqrt{5}+1)}{2\sqrt{5}} \\ \frac{(\sqrt{5}+1)}{2} & \text{w.p. } \frac{(\sqrt{5}-1)}{2\sqrt{5}} \end{cases}$	F2
Mammen's continuous 1	$t_i^* = \frac{u_i}{\sqrt{2}} + \frac{1}{2}(w_i^2 - 1)$ u_i and w_i independent standard normals	F3
Mammen's continuous 2	$t_i^* = \left(\delta_1 + \frac{u_i}{\sqrt{2}}\right)\left(\delta_2 + \frac{w_i}{\sqrt{2}}\right) - \delta_1\delta_2,$ $\delta_1 = \left(\frac{3}{4} + \frac{\sqrt{17}}{12}\right)$ and $\delta_2 = \left(\frac{3}{4} - \frac{\sqrt{17}}{12}\right)$ u_i and w_i independent standard normals	F4
Webb	$t_i^* = \begin{cases} -\sqrt{1.5} & \text{w.p. } \frac{1}{6} \\ -1 & \text{w.p. } \frac{1}{6} \\ -\sqrt{0.5} & \text{w.p. } \frac{1}{6} \\ \sqrt{0.5} & \text{w.p. } \frac{1}{6} \\ 1 & \text{w.p. } \frac{1}{6} \\ \sqrt{1.5} & \text{w.p. } \frac{1}{6} \end{cases}$	F5
Standard normal	$t_i^* \sim N(0, 1)$	F6

5.4 Simulation results – Sizes of the tests

Figures 5.2 – 5.15 illustrate the size of the tests for the different bootstrap approaches across a grid of heteroskedasticity levels λ . The figures include a legend at the top right corner which represents a shorthand notation of the form “f●HC●F●” and is used to indicate the type of residual transformation (f●) applied, the HCCME used (HC●) and the ancillary distribution (F●) used to generate the wild bootstrap samples. The different options are given in Table 5.1, and for the purpose of the simulation study made use of the following combinations:

- f2 HC2 F1–6
- f2 HC3 F1–6
- f3 HC2 F1–6
- f3 HC3 F1–6

No article could be found that had used all the ancillary distributions in a simulation study, typically articles only make use of the two point distributions by Rademacher and Mammen (MacKinnon, 2002; Flachaire, 2005; Davidson and Flachaire, 2008; Cribari-Neto and Lima, 2009; Hausman and Palmer, 2012; MacKinnon, 2013; Politis and Poulis, 2014). For this reason a few extra ancillary distribution were incorporated into this simulation study in order to determine if any of the others may outperform the established two distributions in the literature.

Previous simulation evidence by Davidson and Flachaire (2008), suggest that the performance of asymptotic tests based on HC3 performed best relative to the other HCCMEs. Also, Hausman and Palmer (2012) indicated that using HC2 and HC3 with a wild bootstrap test statistic performs well relative to other HCCMEs especially in small samples (the two methods provide similar results). For this reason HC3 was used as a benchmark to test the performance of the modified bootstrap residuals approach against. HC2 was also included since HC3 is a slightly modified version of it.

The transformations that were used in this simulation study, i.e., f2 and f3, were chosen on the basis of the suggestion that the transformations $f(\cdot)$ should be paired with their corresponding

HCCME for best results (Davidson and Flachaire, 2008), and supported by previous simulation studies (see, e.g., MacKinnon, 2013).

Thus for this simulation study the two established HCCMEs, HC2 and HC3 were chosen to correct for heteroskedasticity with their corresponding transformations f2 and f3. Combinations of the HCCMEs and residuals transformations were also investigated.

Finally, the results for tests based on the modified bootstrap residuals are also presented so that they can be compared to the other wild bootstrap tests. The next subsection will highlight a few remarks around the size of the test for this simulation study. The subsection will be split into two parts, the first part will discuss the results relating to lognormal predictors, i.e., models with a few large leverage values, and the second part will discuss the models with uniform predictors, i.e., models with small to moderate leverage values.

5.4.1 Remarks on the sizes of tests using lognormal predictors

In the lognormal case we noticed that for the sample sizes $n = 20, 40, 80, 160$ (Figures 5.2, 5.4, 5.6 and 5.8) the tests always over-rejected, attaining sizes of 10% and upwards, when the data was homoskedastic ($\lambda = 0$). This over-rejection was most pronounced for the transformation f2 in smaller samples, and which only achieved the significance level for $n = 40, 80$ at the most extreme heteroskedasticity setting $\lambda = 2$.

All the rejection frequencies in these lognormal predictor graphs (Figures 5.2, 5.4, 5.6, 5.8, 5.10, 5.12 and 5.14) follow a similar trend: for values of λ close to zero (homoskedasticity) the tests tended to wildly over-reject, but as the heteroskedasticity was increased, by incrementally increasing the value of λ , the rejection frequencies decreased towards the specified significance level of $\alpha = 0.05$. Each test eventually achieved the specified significance level at a particular value of $\lambda > 0$, but tended to continue decreasing for increasing values of λ and thus under-reject for these more extreme heteroskedastic structures.

Simulations involving the residuals transformation f3 appeared to produce lower sizes of the test when compared to simulations involving the residual transformation f2. This lowering of the rejection frequency trend curve became less pronounced from the sample size $n = 80$ and upward.

Concerning the ancillary distributions, it was noted that F1, F5 and F6 produced similar rejection frequency curves, but were always lower relative to those produced by the other distributions. As the sample size increased these curves became flatter and shifted towards the specified significance level. The worst performer was F2, even though it had a similar trend to the other distributions it always tended to over-reject relative to the others. Also, in Figure 5.14, as the heteroskedasticity became extreme we found that the tests based on F2 started to over-reject again.

As with the wild bootstrap approaches discussed above, the rejection frequency graphs of the tests based on the modified bootstrap residuals also had the same general trend of over-rejecting when the data was homoskedastic ($\lambda = 0$), but started to approach the specified significance level as the heteroskedasticity increased. It was interesting to note that the modified approach showed an improvement over the wild bootstrap approaches by not over-rejecting as severely for $n \geq 40$. In addition, the same phenomena of the rejection probability flattening out observed for the wild bootstrap approaches was also observed for the modified approach, but at a faster rate. For smaller sample sizes it appeared that the modified bootstrap residuals approach employing the HC3 estimator performed the best, but as the sample size increased the performance of HC1 tended to be better by not under-rejecting as the heteroskedasticity became extreme.

5.4.2 Remarks on the sizes of tests using uniform predictors

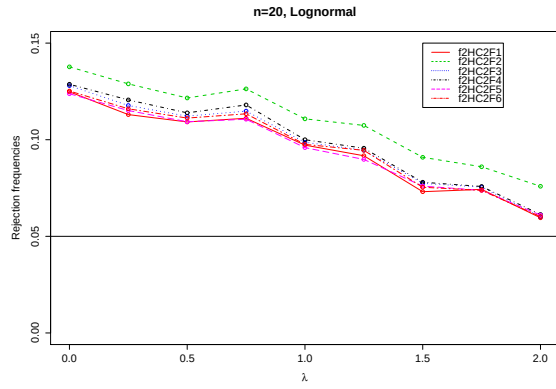
In the case of uniform predictors we considered a data generation model that did not produce large leverage values. The lack of extreme leverages means that the HCCMEs (and the associated test statistics) calculated would not be unduly distorted. This had a profound effect on the trend of the rejection frequencies; it produced a trend that was constant over the whole range of heteroskedasticity levels, see Figures 5.3, 5.5, 5.7, 5.9, 5.11, 5.13 and 5.15. This was in stark contrast to the lognormal case where tests tended to over-reject, decrease towards the desired significance level, and then ultimately under-reject for the most extreme heteroskedasticity levels.

For the wild bootstrap tests we noticed that the ancillary distributions F1 and F5 produced sizes of the test closest to the specified significance level for all the sample sizes. For sample sizes $n \geq 80$, all the tests began to produce similar results, i.e., the sizes of the tests all began to migrate towards each other and also to the specified significance level. From sample size $n = 160$

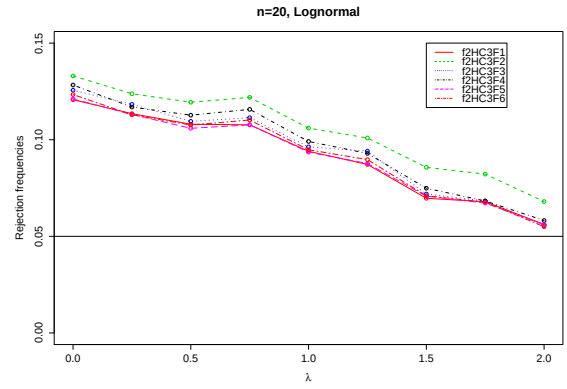
onwards, all the wild bootstrap tests performed equally well and appeared to have achieved the desired significance level. Note that the ancillary distribution F6 (standard normal distribution) converged to the size values of the two best ancillary distributions (F1 and F5) slightly faster than the distributions F2, F3 and F4, as the sample size increased.

The modified bootstrap residuals approach worked well for all sample sizes and achieved the desired significance level for each of the sample sizes considered. Indeed, the modified approach outperformed all other wild bootstrap approaches for samples $n \leq 40$, but the distinction grew less pronounced as the sample sizes were increased (since all the tests approximately achieved the desired significance level for large sample sizes). The excellent performance of this modified bootstrap residuals approach could possibly be attributed to the manner in which the data was generated: the modified bootstrap residuals approach would logically favour a structured heteroskedastic data set whereas the other approaches were designed for unstructured heteroskedastic data sets.

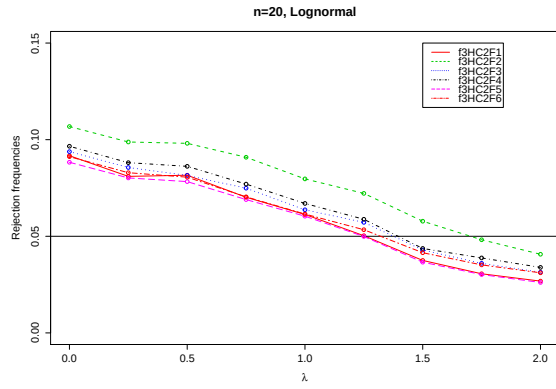
The choice of which HCCME to use is inconsequential for both the modified bootstrap residuals approach and wild bootstrap when the uniform predictors were used, since the rejection frequencies were found not to be affected by the choice for any sample size.



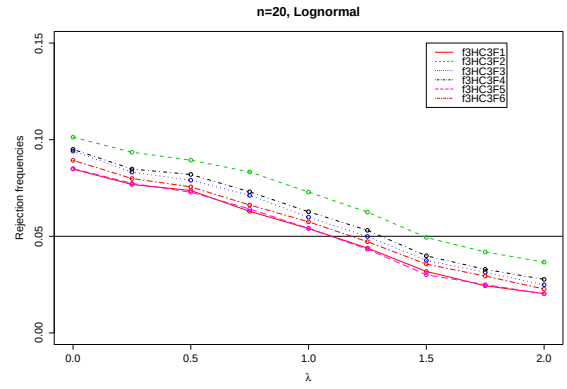
(a) Wild bootstrap HC2 t -test rejection frequencies with f2 residuals.



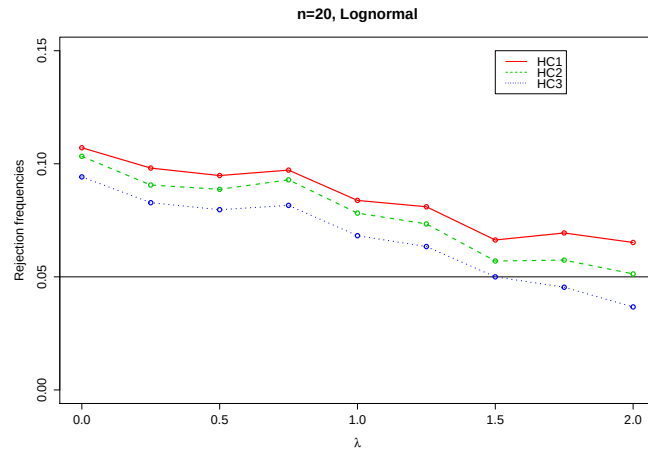
(b) Wild bootstrap HC3 t -test rejection frequencies with f2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f3 residuals.

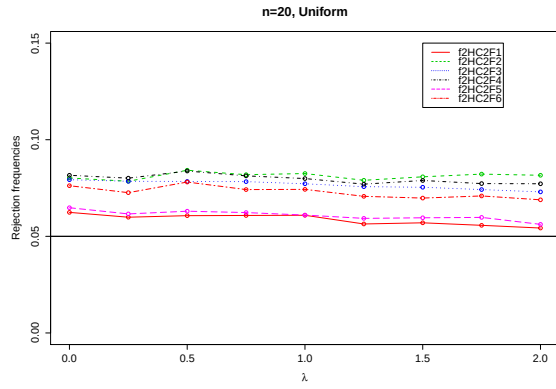


(d) Wild bootstrap HC3 t -test rejection frequencies with f3 residuals.

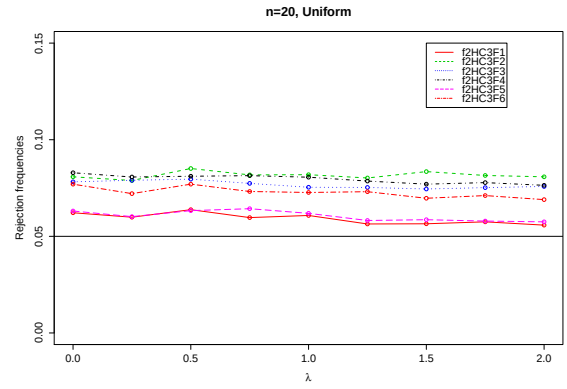


(e) Modified bootstrap residuals approach t -test rejection frequencies

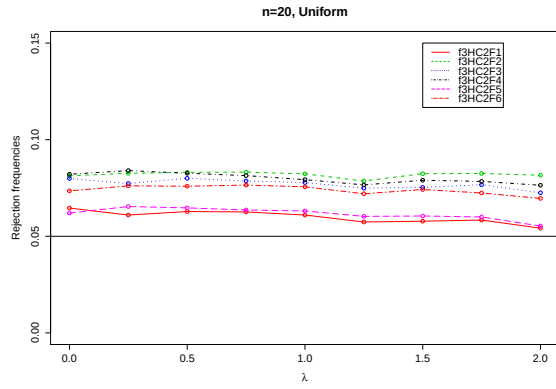
Fig. 5.2: Size of the test for lognormal predictors and a sample size of $n = 20$.



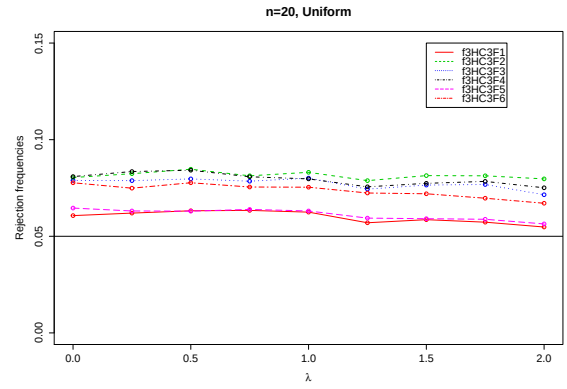
(a) Wild bootstrap HC2 t -test rejection frequencies with f2 residuals.



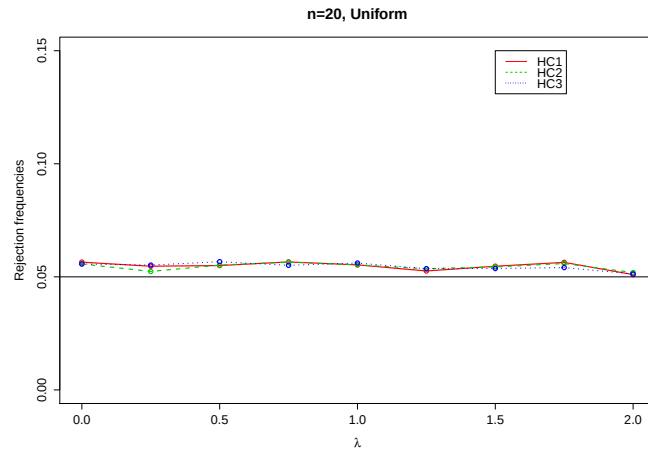
(b) Wild bootstrap HC3 t -test rejection frequencies with f2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f3 residuals.

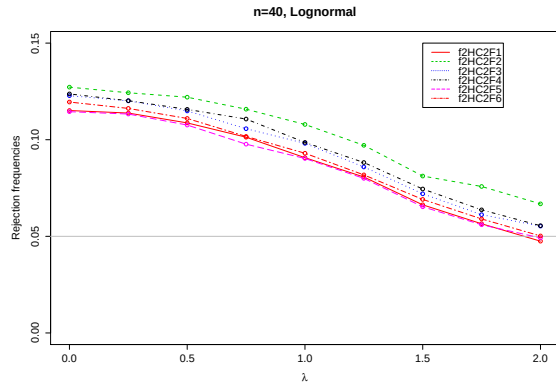


(d) Wild bootstrap HC3 t -test rejection frequencies with f3 residuals.

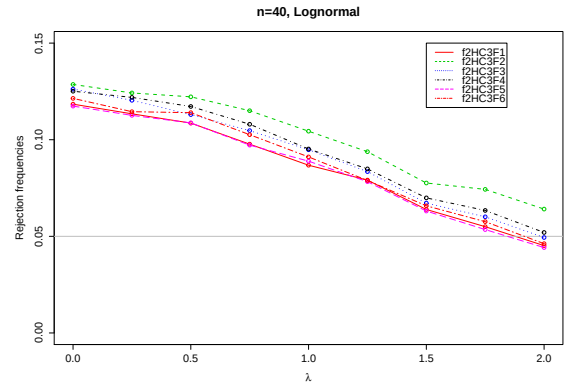


(e) Modified bootstrap residuals approach t -test rejection frequencies

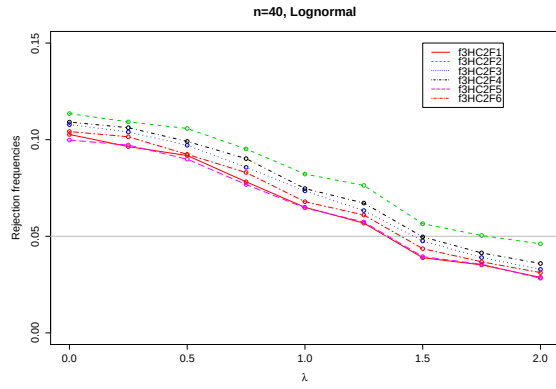
Fig. 5.3: Size of the test for uniform predictors and a sample size of $n = 20$.



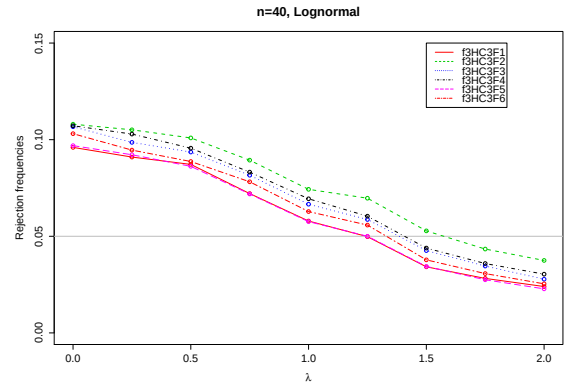
(a) Wild bootstrap HC2 t -test rejection frequencies with f2 residuals.



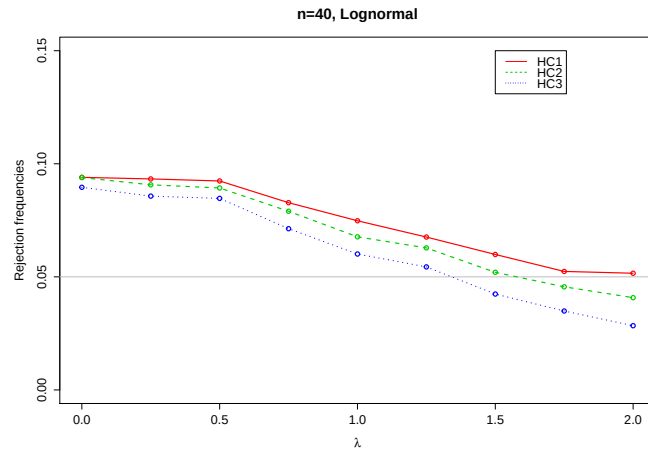
(b) Wild bootstrap HC3 t -test rejection frequencies with f2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f3 residuals.



(d) Wild bootstrap HC3 t -test rejection frequencies with f3 residuals.



(e) Modified bootstrap residuals approach t -test rejection frequencies

Fig. 5.4: Size of the test for lognormal predictors and a sample size of $n = 40$.

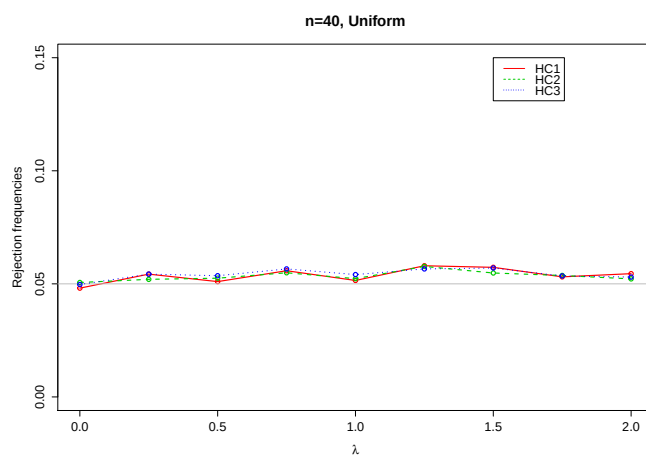
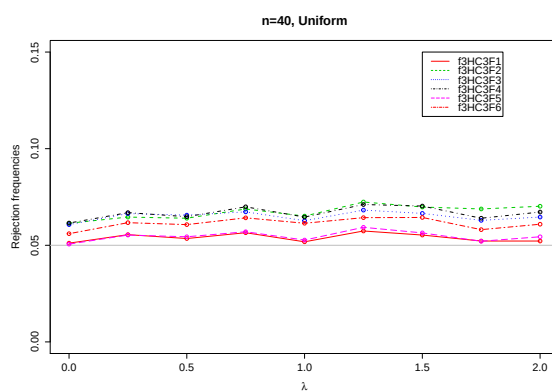
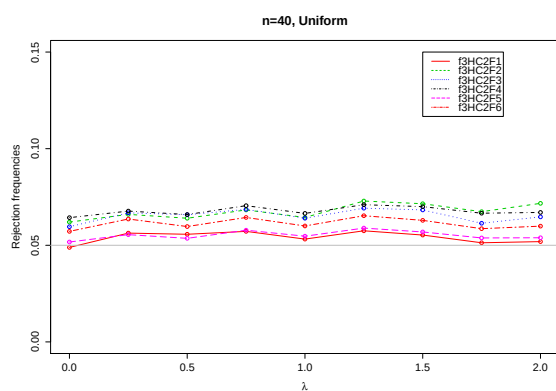
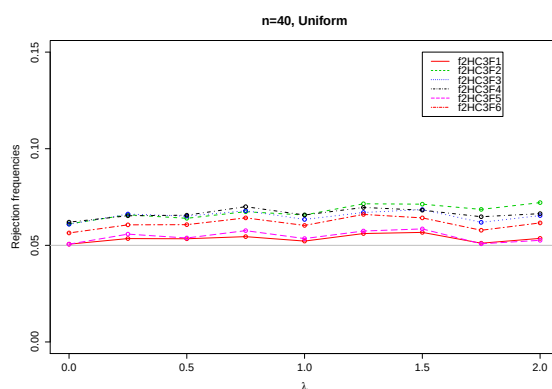
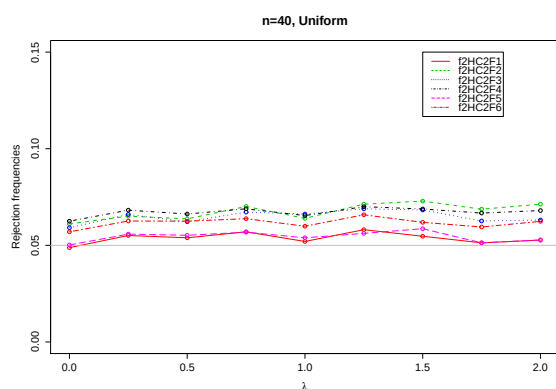
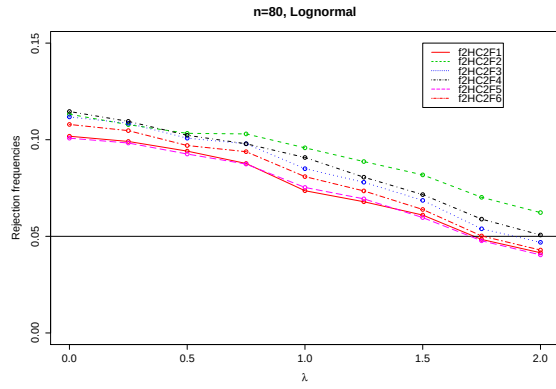
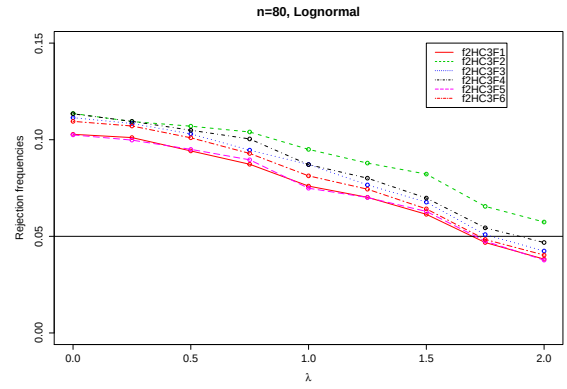


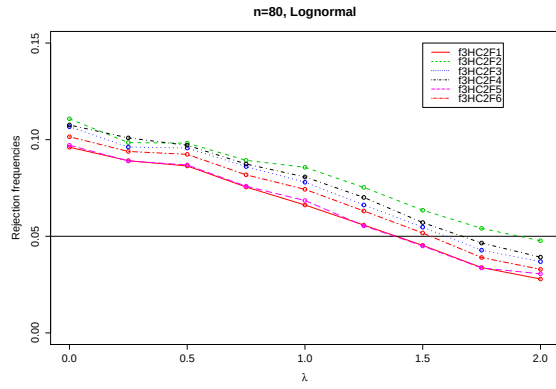
Fig. 5.5: Size of the test for uniform predictors and a sample size of $n = 40$.



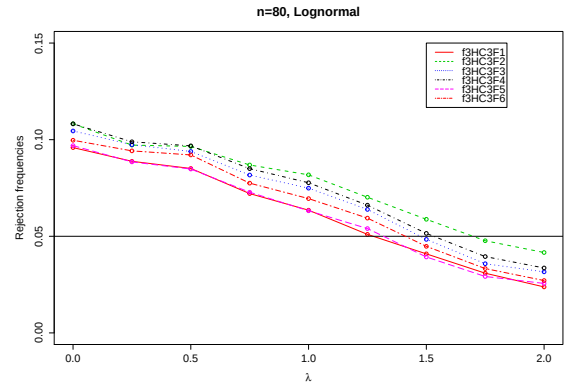
(a) Wild bootstrap HC2 t -test rejection frequencies with f2 residuals.



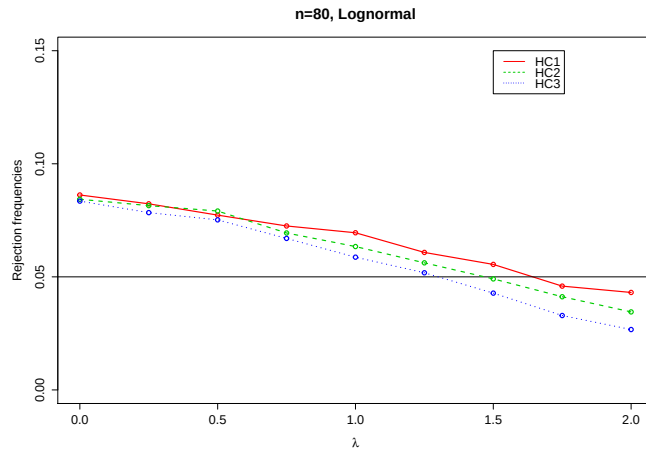
(b) Wild bootstrap HC3 t -test rejection frequencies with f2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f3 residuals.



(d) Wild bootstrap HC3 t -test rejection frequencies with f3 residuals.



(e) Modified bootstrap residuals approach t -test rejection frequencies

Fig. 5.6: Size of the test for lognormal predictors and a sample size of $n = 80$.

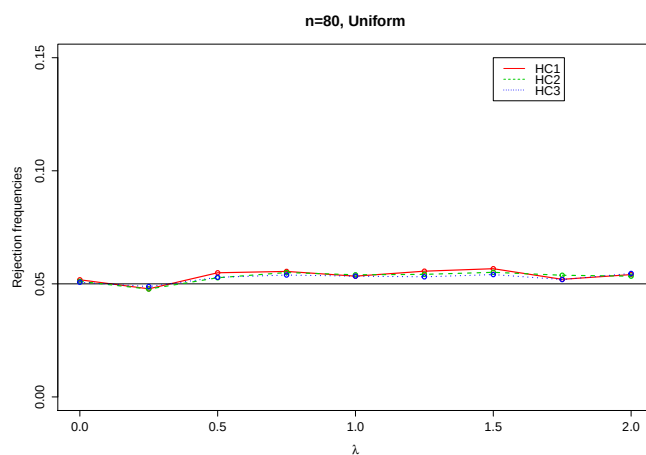
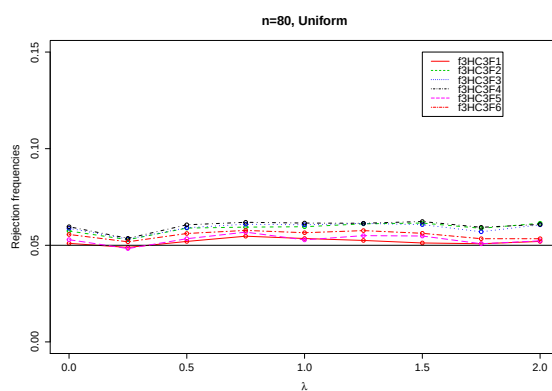
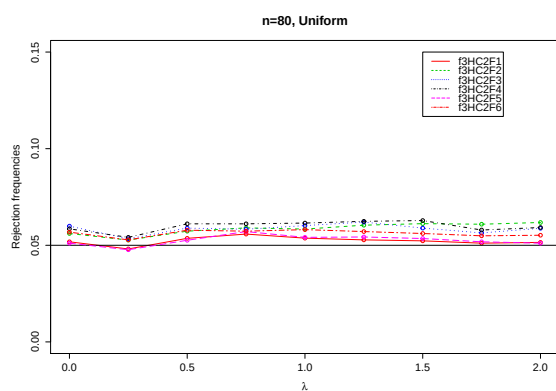
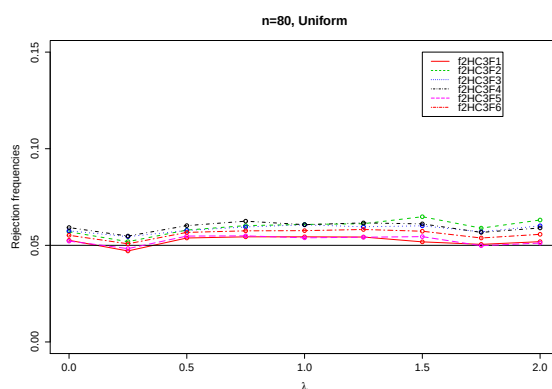
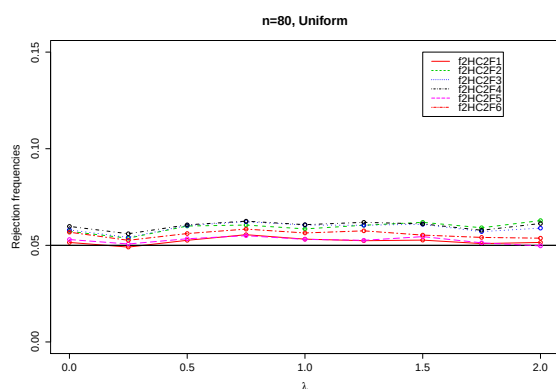
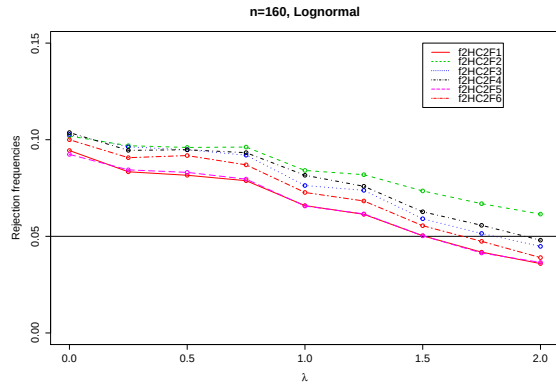
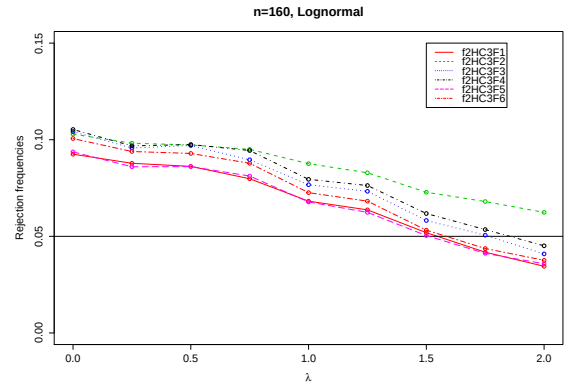


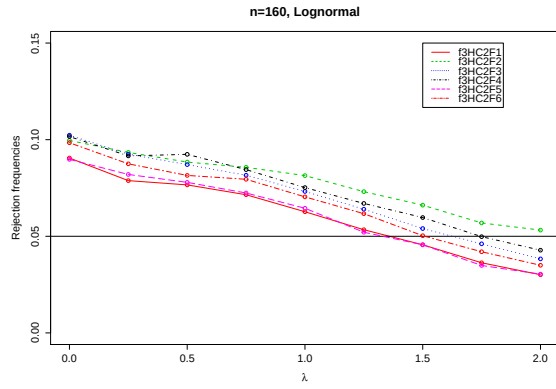
Fig. 5.7: Size of the test for uniform predictors and a sample size of $n = 80$.



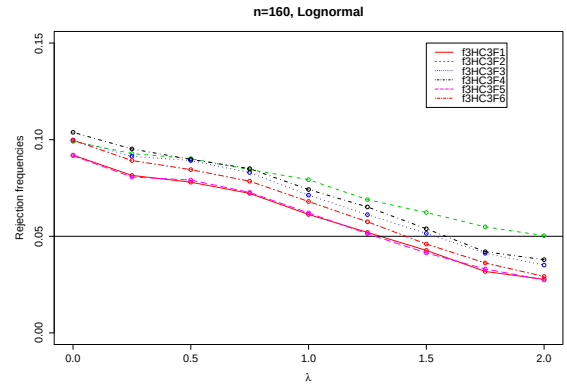
(a) Wild bootstrap HC2 t -test rejection frequencies with f2 residuals.



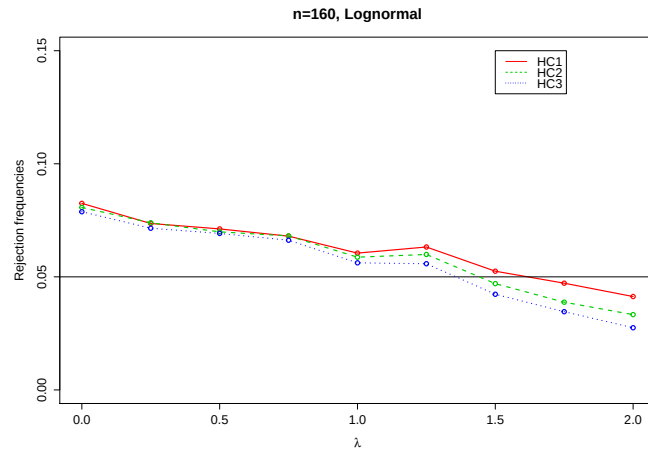
(b) Wild bootstrap HC3 t -test rejection frequencies with f2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f3 residuals.

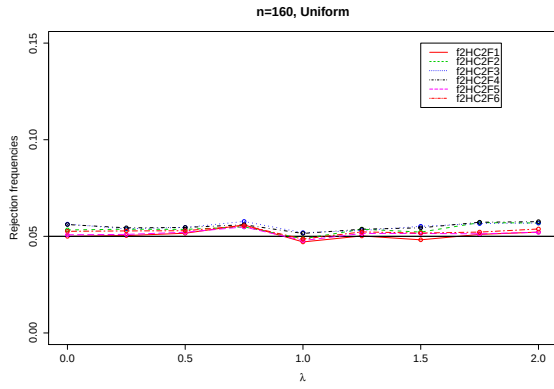


(d) Wild bootstrap HC3 t -test rejection frequencies with f3 residuals.

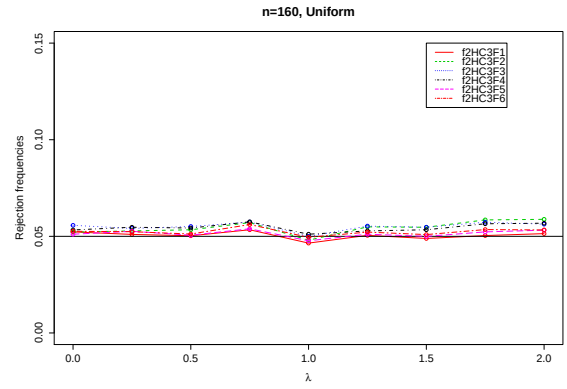


(e) Modified bootstrap residuals approach t -test rejection frequencies

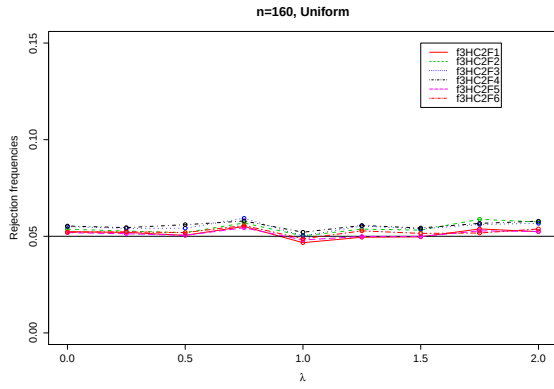
Fig. 5.8: Size of the test for lognormal predictors and a sample size of $n = 160$.



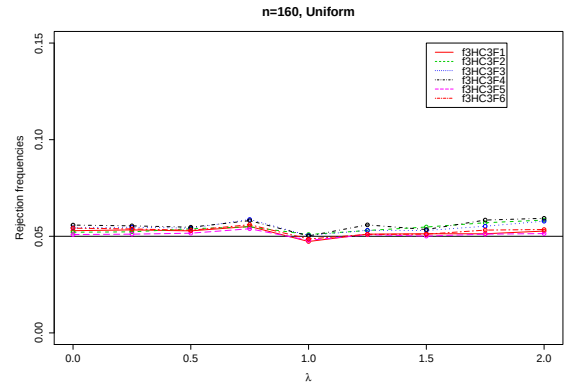
(a) Wild bootstrap HC2 t -test rejection frequencies with f_2 residuals.



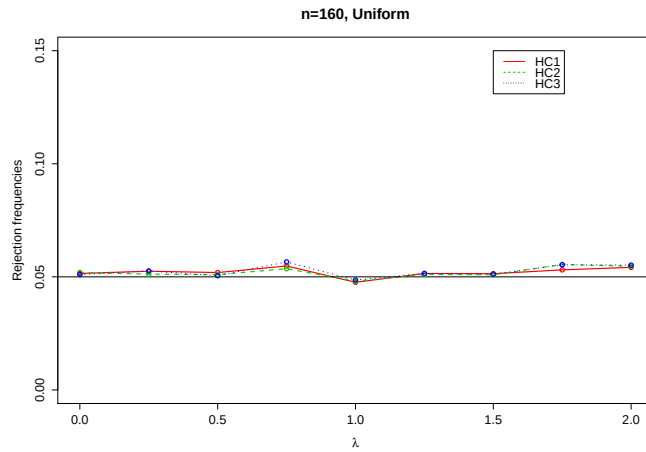
(b) Wild bootstrap HC3 t -test rejection frequencies with f_2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f_3 residuals.



(d) Wild bootstrap HC3 t -test rejection frequencies with f_3 residuals.



(e) Modified bootstrap residuals approach t -test rejection frequencies

Fig. 5.9: Size of the test for uniform predictors and a sample size of $n = 160$.

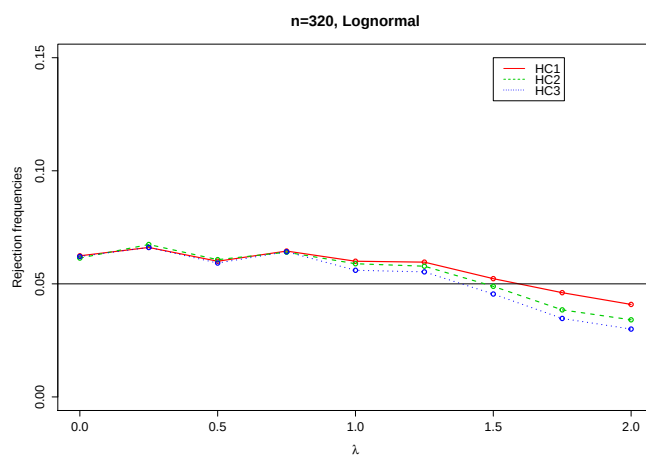
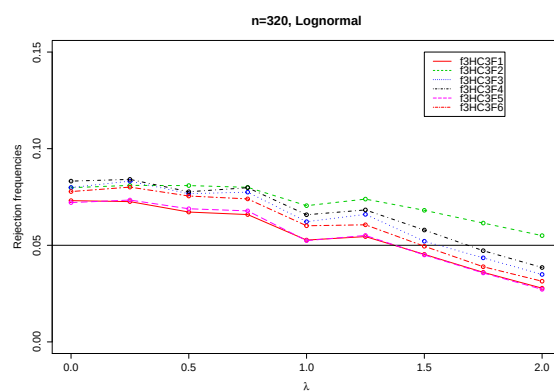
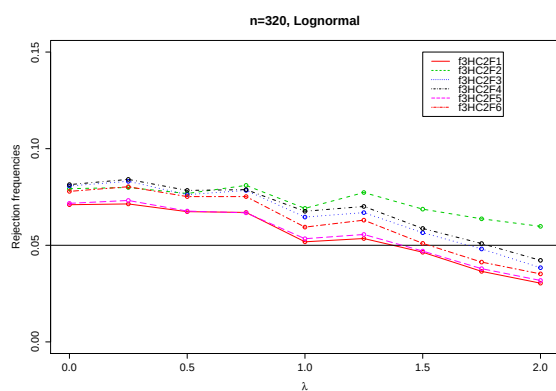
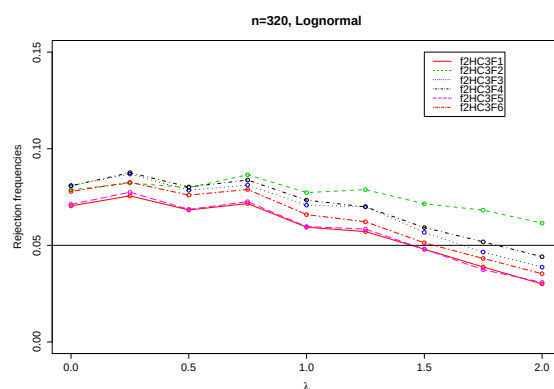
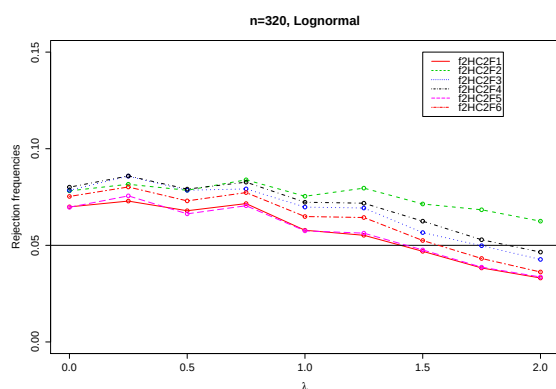
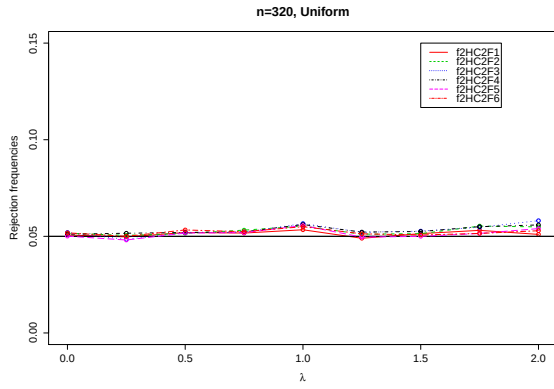
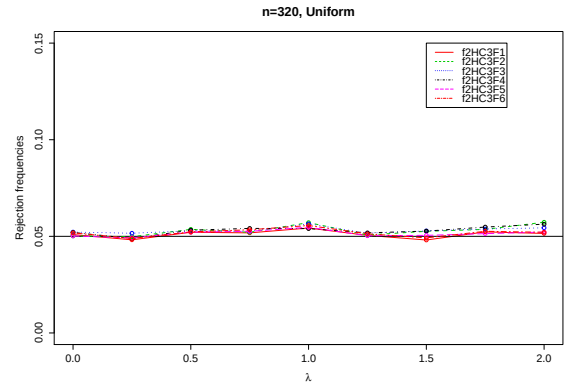


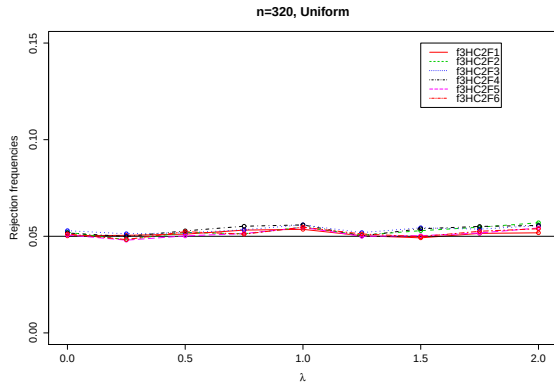
Fig. 5.10: Size of the test for lognormal predictors and a sample size of $n = 320$.



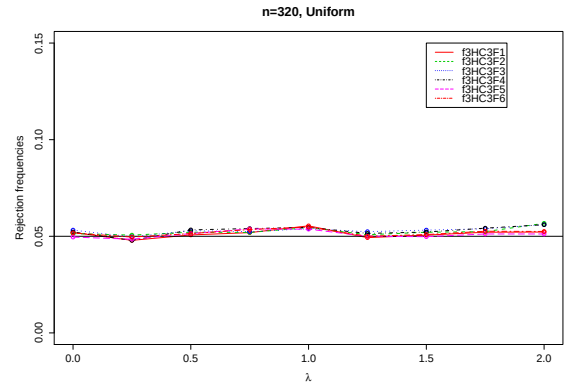
(a) Wild bootstrap HC2 t -test rejection frequencies with f_2 residuals.



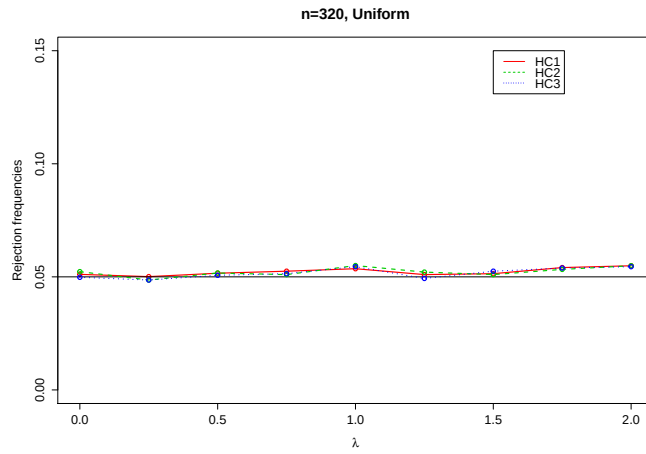
(b) Wild bootstrap HC3 t -test rejection frequencies with f_2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f_3 residuals.

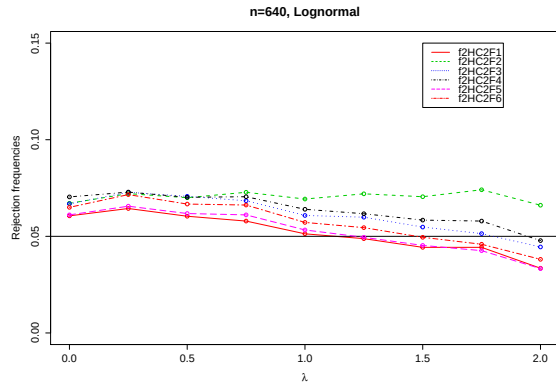


(d) Wild bootstrap HC3 t -test rejection frequencies with f_3 residuals.

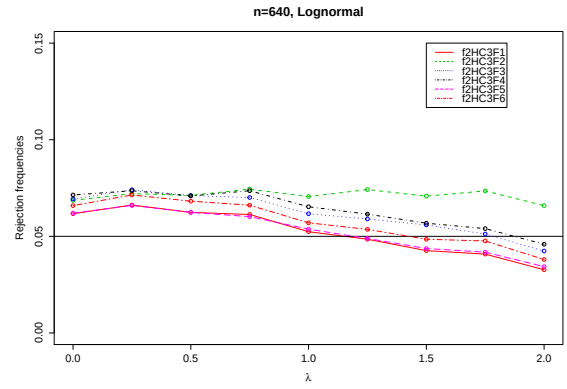


(e) Modified bootstrap residuals approach t -test rejection frequencies

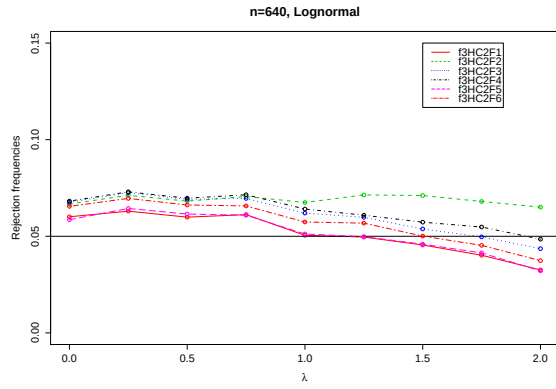
Fig. 5.11: Size of the test for uniform predictors and a sample size of $n = 320$.



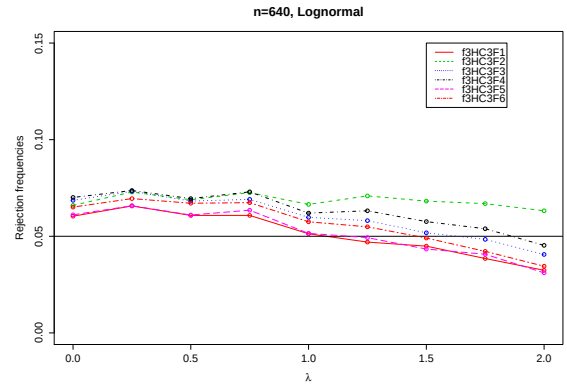
(a) Wild bootstrap HC2 t -test rejection frequencies with f2 residuals.



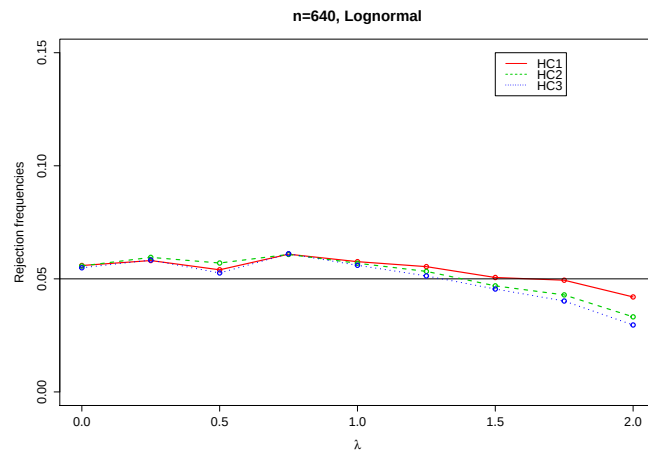
(b) Wild bootstrap HC3 t -test rejection frequencies with f2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f3 residuals.

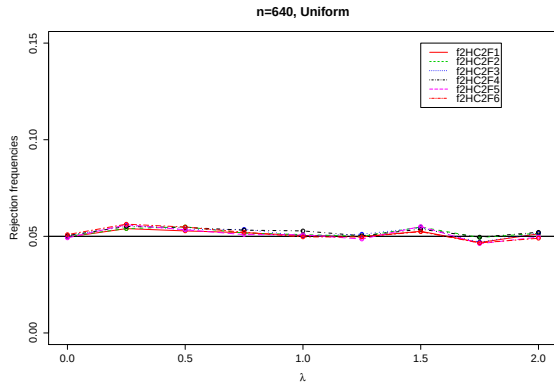


(d) Wild bootstrap HC3 t -test rejection frequencies with f3 residuals.

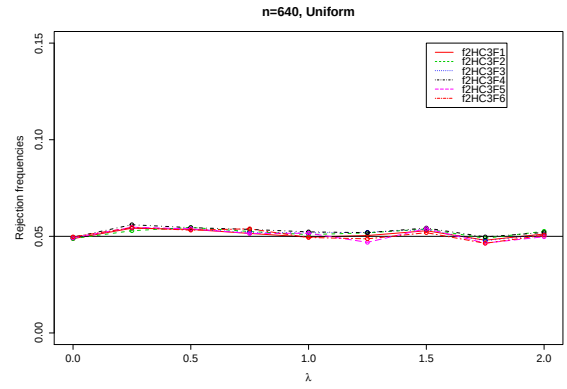


(e) Modified bootstrap residuals approach t -test rejection frequencies

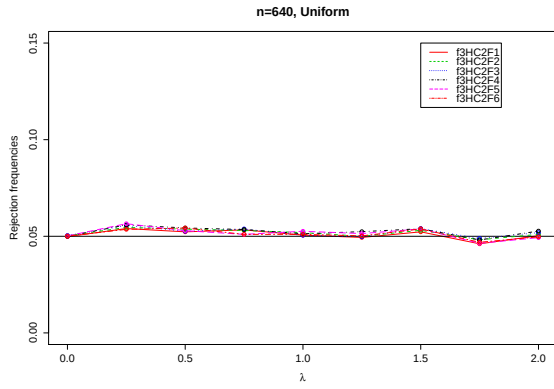
Fig. 5.12: Size of the test for lognormal predictors and a sample size of $n = 640$.



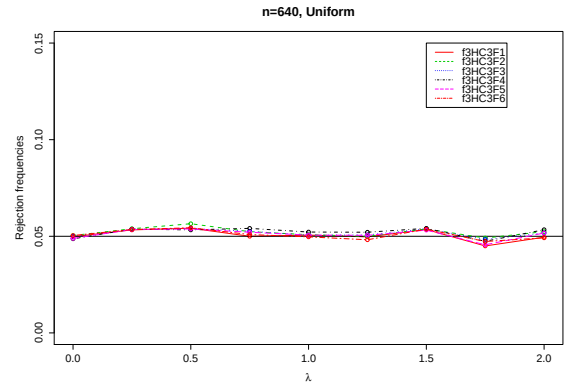
(a) Wild bootstrap HC2 t -test rejection frequencies with f2 residuals.



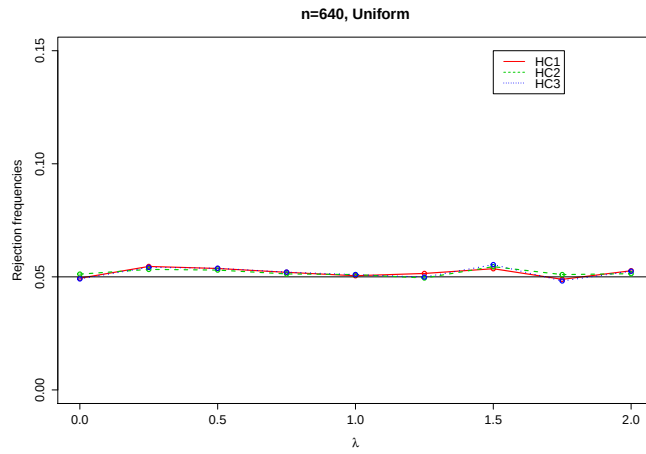
(b) Wild bootstrap HC3 t -test rejection frequencies with f2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f3 residuals.

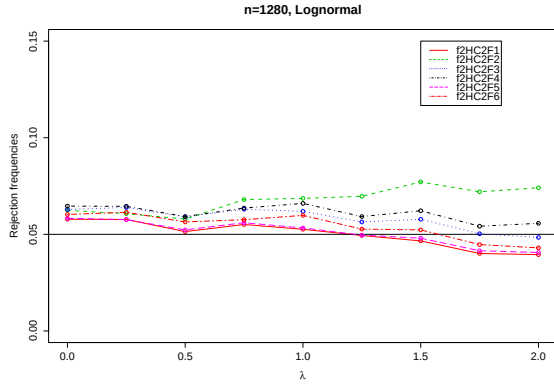


(d) Wild bootstrap HC3 t -test rejection frequencies with f3 residuals.

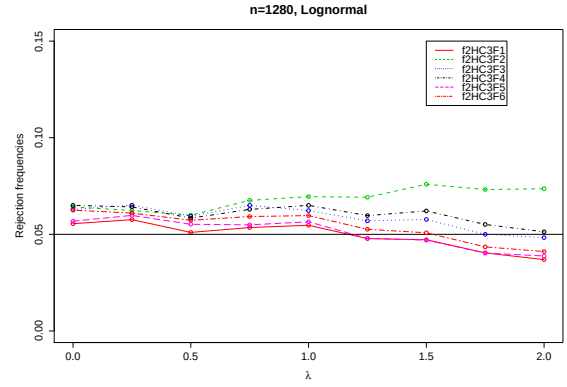


(e) Modified bootstrap residuals approach t -test rejection frequencies

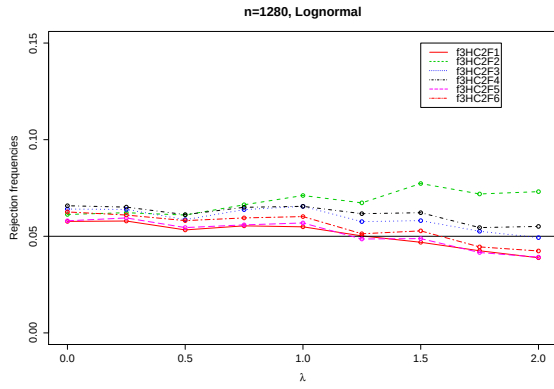
Fig. 5.13: Size of the test for uniform predictors and a sample size of $n = 640$.



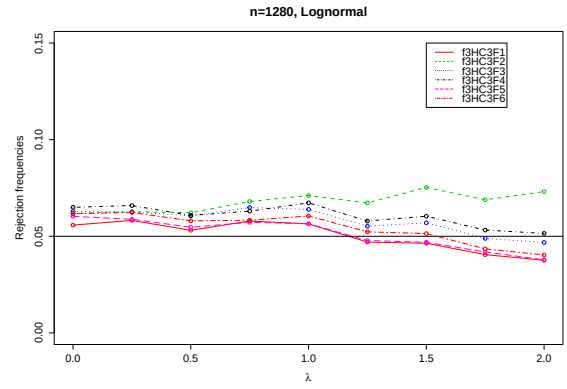
(a) Wild bootstrap HC2 t -test rejection frequencies with f2 residuals.



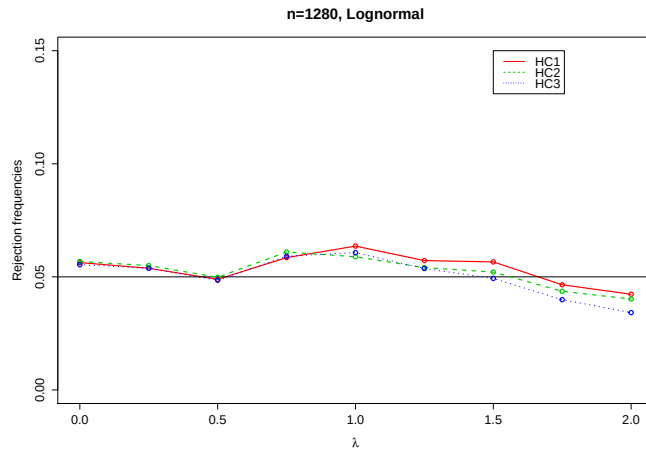
(b) Wild bootstrap HC3 t -test rejection frequencies with f2 residuals.



(c) Wild bootstrap HC2 t -test rejection frequencies with f3 residuals.



(d) Wild bootstrap HC3 t -test rejection frequencies with f3 residuals.



(e) Modified bootstrap residuals approach t -test rejection frequencies

Fig. 5.14: Size of the test for lognormal predictors and a sample size of $n = 1280$.

5.5 Simulation results – Powers of selected tests

Figures 5.16 – 5.23 illustrate the powers for a subset of bootstrap tests over a range of values within the set of values described by the alternative hypothesis, i.e., for a range of values $\beta_3 \neq 0$. Due to the computational costs involved with the calculation of each alternative, only a subset of the many possible combinations of settings used in the size calculations was used, specifically we used those settings where the bootstrap tests achieved the specified significance level. The heteroskedasticity setting that most commonly produced significance level close to $\alpha = 0.05$ was where $\lambda = 1.25$, and so all power calculation presented here were calculated with this level of heteroskedasticity. This was done to ensure that, at the very least, a fair comparison could be made between the power curves for the subset of bootstrap tests considered. Below is the list of bootstrap tests that were investigated in terms of their power performance, along with a short motivation for each one:

- **f3HC3F1 and f3HC3F5:** As was noted in the previous subsection, the wild bootstrap approach with ancillary distributions F1 and F5 performed the best when combined with the heteroskedasticity consistent covariance matrix estimator HC3 and residuals transformation f3 in small samples.
- **f3HC3F2:** Additionally, the wild bootstrap approach using F2 was included for the sake of interest since it is a skew distribution and the worst performer in terms of size.
- **Modified bootstrap residuals using HC3 (newHC3):** For the modified bootstrap residuals approach HC3 was selected because it had reasonable size performance for small sample sizes and is of primary interest in this study.

It was decided that the smaller sample sizes would be investigated because, specifically for the case where lognormal predictors are created, the data would contain a small number of very large leverage values. By generating the data in this fashion it would ensure that inference was as difficult as possible. All power simulations were conducted with $MC = 10\,000$ and $B = 500$.

5.5.1 Remarks on the power of the test for this simulation study

It is interesting to note that the lognormal predictors had much higher power relative to that of the uniform predictors for all the sample sizes. Naturally, the power curves (for both types of predictors) become steeper as the sample size increases.

For the sample size $n = 20$ the modified bootstrap residuals test (newHC3) did not achieve the specified significance level in the case of lognormal predictors (Figure 5.16), but the wild bootstrap approaches with the distributions F1 and F5 did. Moreover, for the uniform predictors the modified approach achieved the specified significance level, whereas the wild bootstrap approaches did not (Figure 5.17).

In the case of the uniform predictors with $n \geq 40$ (Figures 5.19, 5.21, and 5.23), all the tests performed equally well by achieving the specified significance level and having the same power. However, the wild bootstrap approach with ancillary distribution F2 only had these features for $n \geq 80$ onwards, and still did not achieve the specified significance level for $n = 40$. The remainder of the remarks will focus on aspects relating to the lognormal predictors.

For $n = 40$ in Figures 5.18 and 5.19, the modified bootstrap residuals approach almost achieved the specified significance level and, interestingly, had a similar power curve to that of the wild bootstrap with ancillary distribution F2. The wild bootstrap approaches with F1 and F5 achieved the specified significance level and had slightly less power compared to the previously mentioned approaches, however due to the other methods not attaining the same significance level the comparison would not be fair.

In Figures 5.20 and 5.21, where $n = 80$, the modified bootstrap residuals approach achieved the significance level and had a slightly better power curve to that of the wild bootstrap approaches with ancillary distributions F1 and F5. Again the wild bootstrap approach with ancillary distribution F2 did not achieve the significance level and so it was disregarded.

For $n = 160$, the wild bootstrap approach with ancillary distribution F2 did not achieve the significance level and so it was again disregarded. The other approaches attained the specified significance level with the modified bootstrap residuals approach still retaining the slightly better power curve overall, see Figures 5.22 and 5.23.

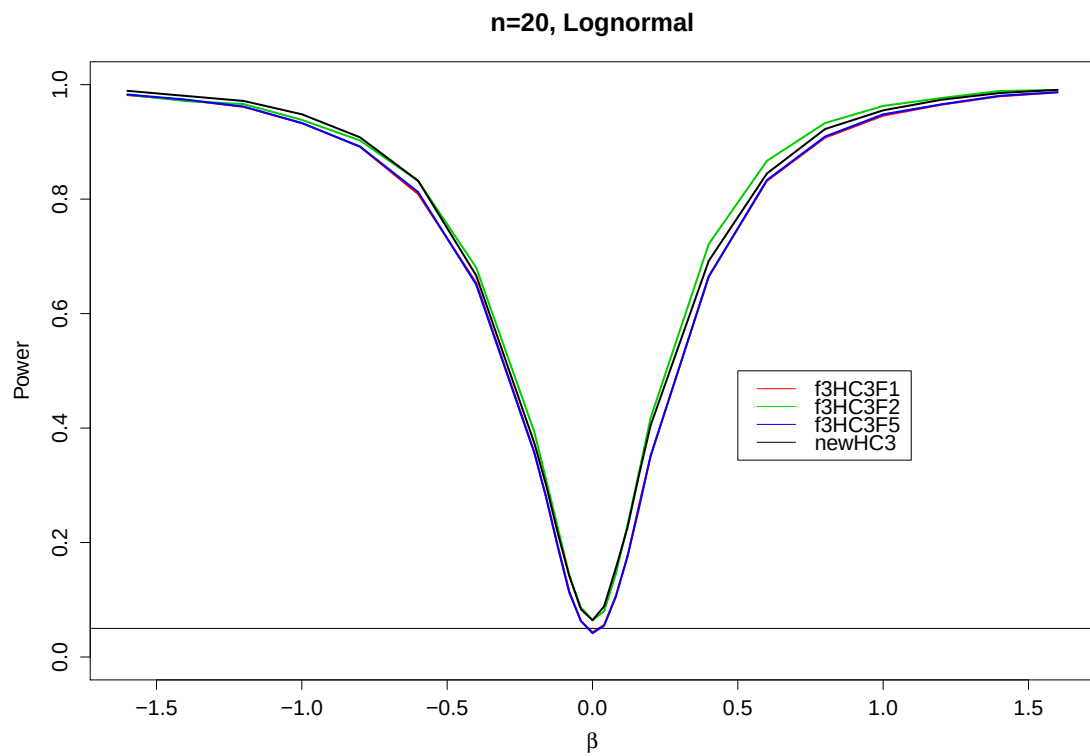


Fig. 5.16: Power for HC3 bootstrap methods, $n = 20$ and lognormal predictors.

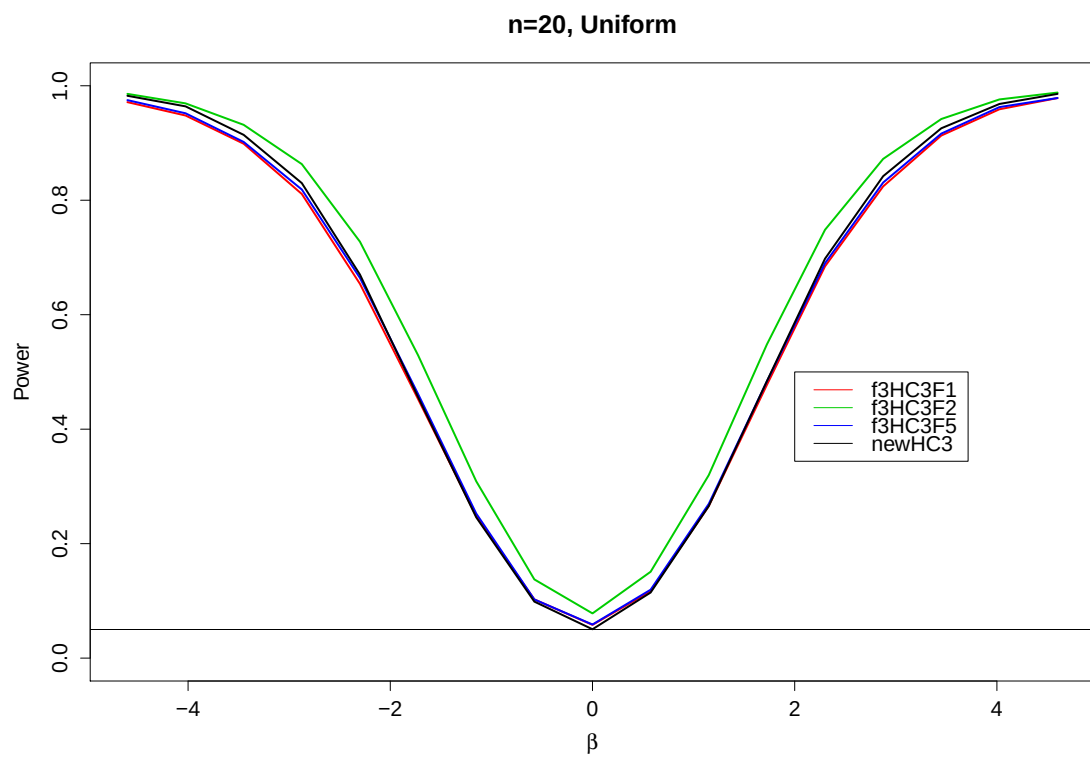


Fig. 5.17: Power for HC3 bootstrap methods, $n = 20$ and uniform predictors.

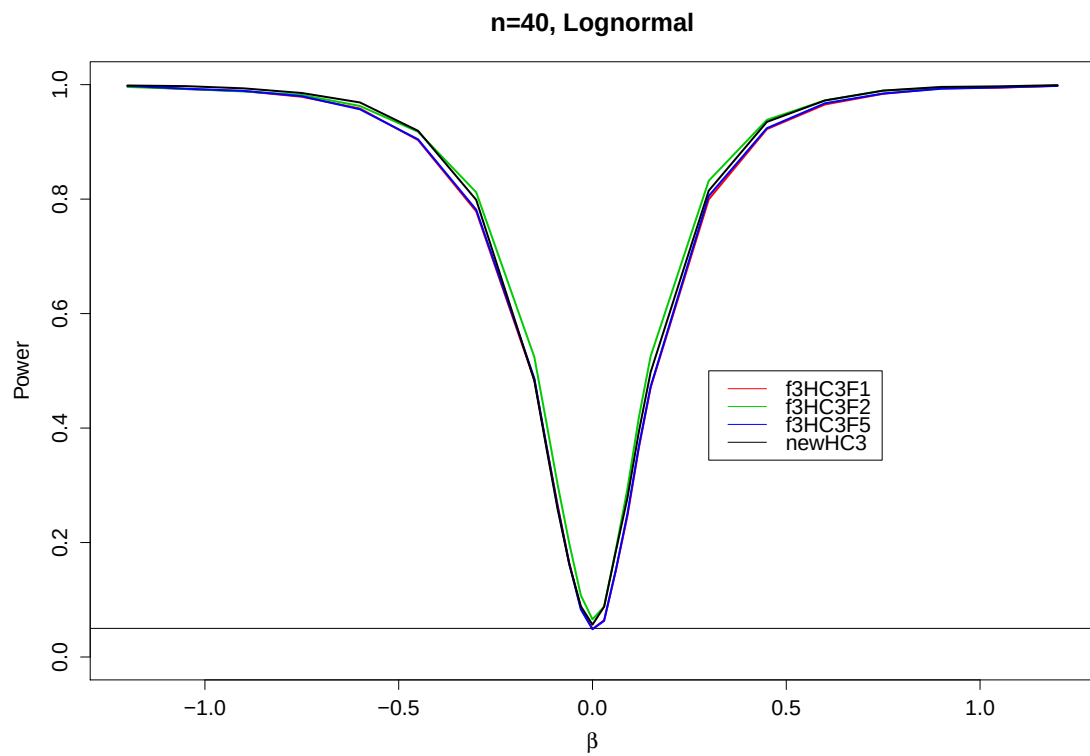


Fig. 5.18: Power for HC3 bootstrap methods, $n = 40$ and lognormal predictors.

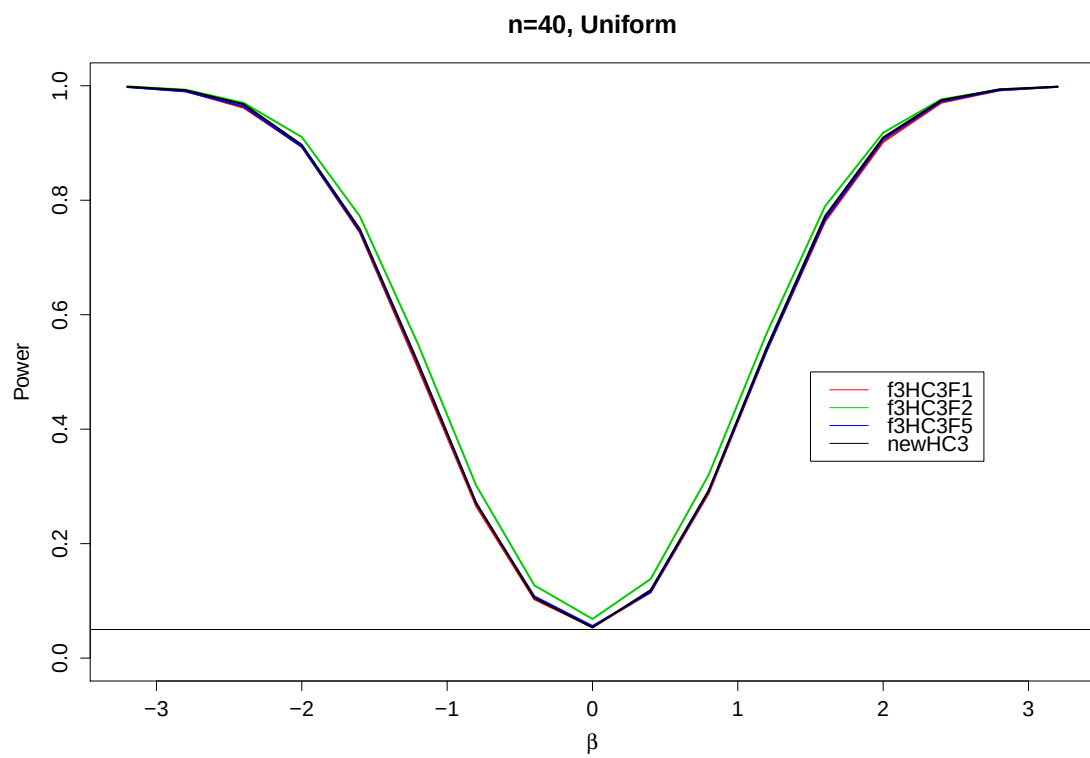


Fig. 5.19: Power for HC3 bootstrap methods, $n = 40$ and uniform predictors.

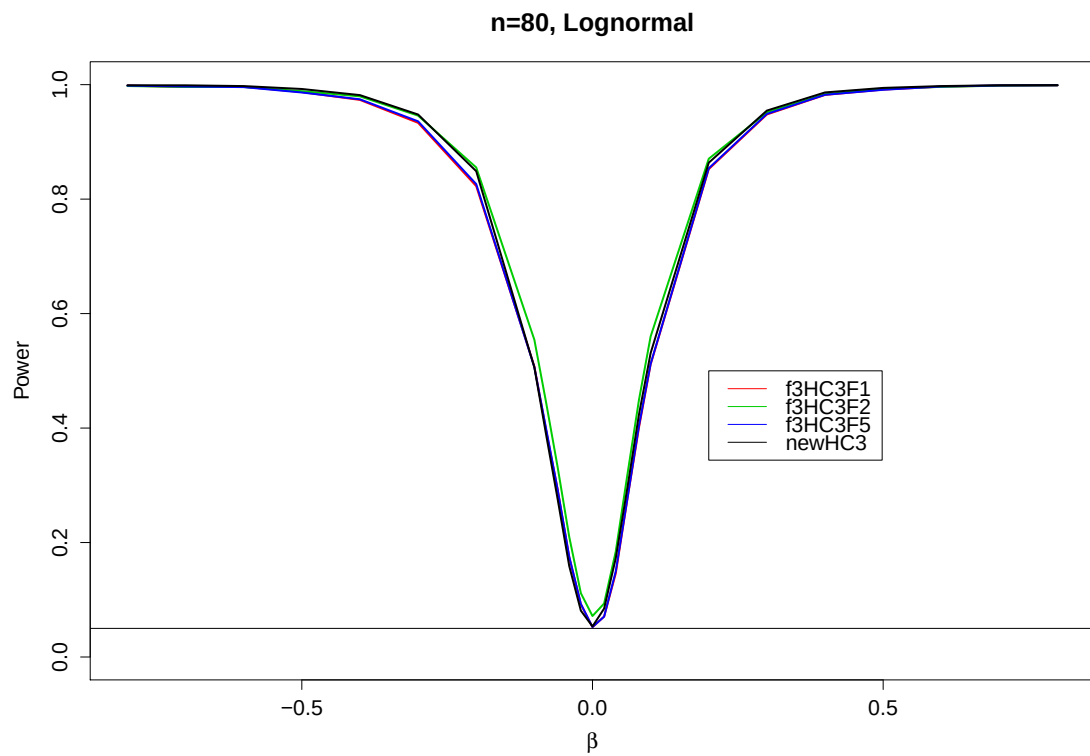


Fig. 5.20: Power for HC3 bootstrap methods, $n = 80$ and lognormal predictors.

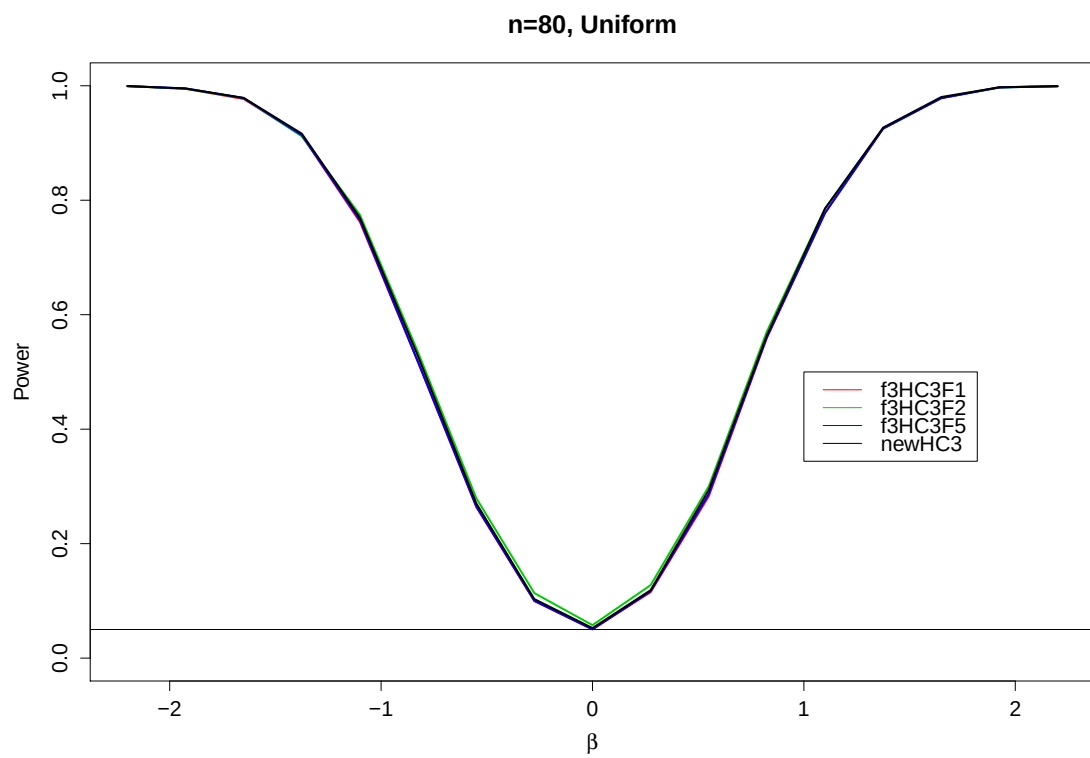


Fig. 5.21: Power for HC3 bootstrap methods, $n = 80$ and uniform predictors.

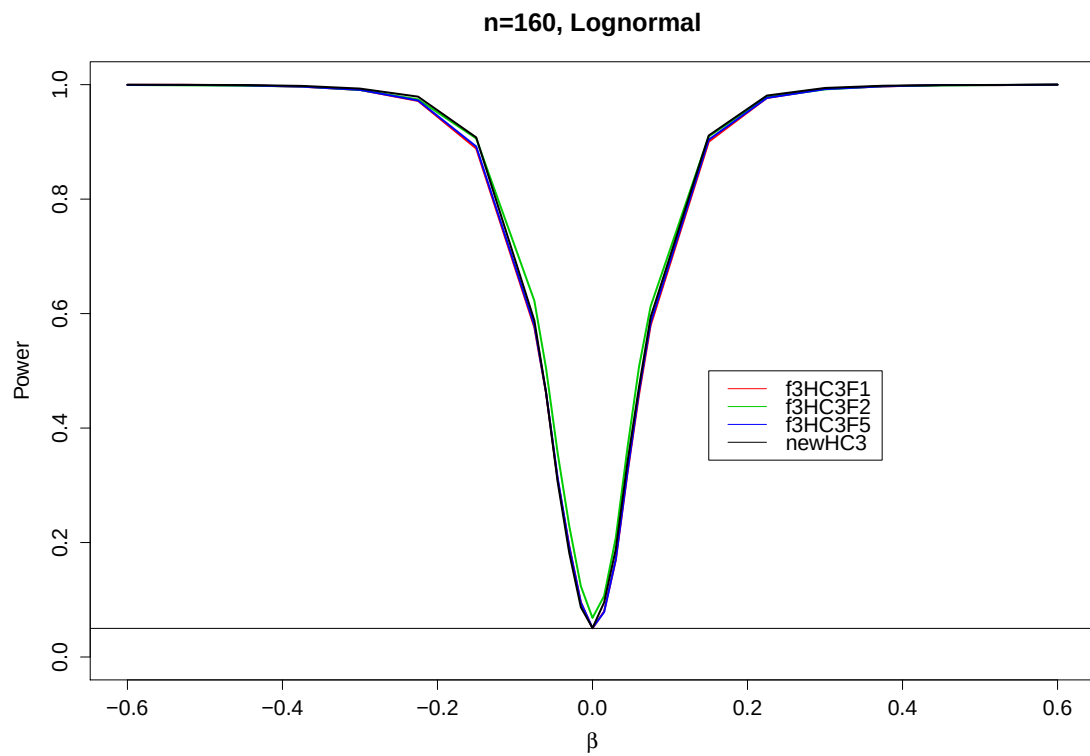


Fig. 5.22: Power for HC3 bootstrap methods, $n = 160$ and lognormal predictors.

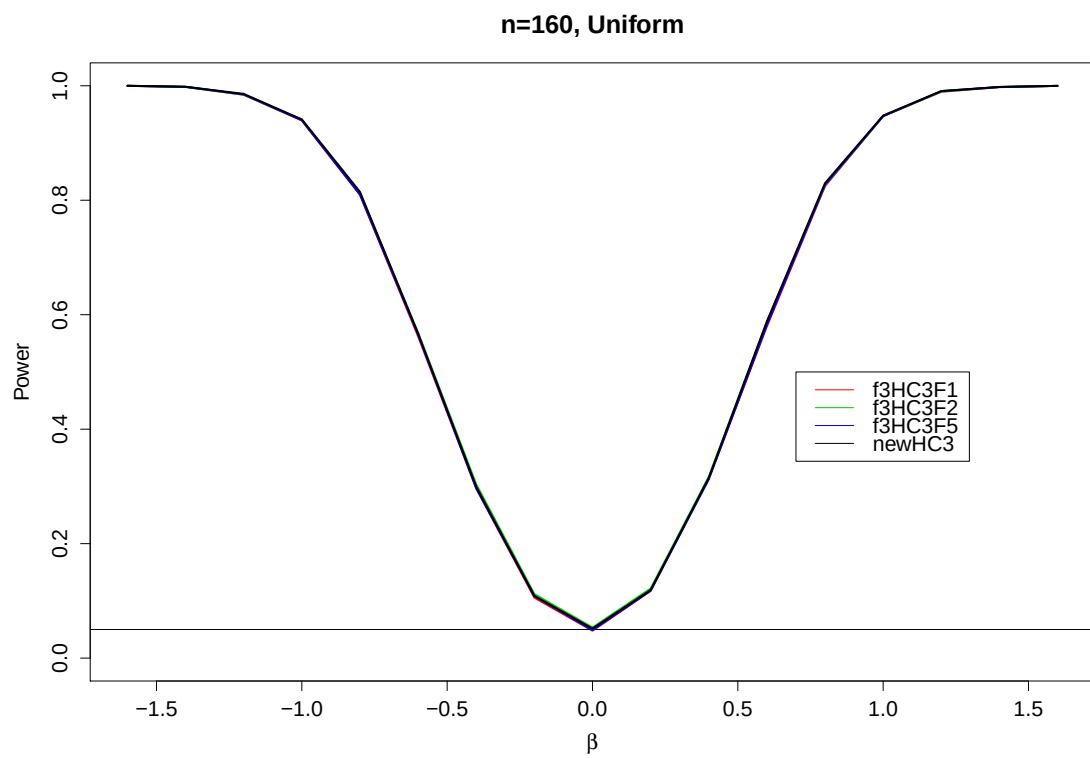


Fig. 5.23: Power for HC3 bootstrap methods, $n = 160$ and uniform predictors.

CHAPTER 6

Conclusion and future considerations

Throughout this text remedial measures were presented for tests that related to hypothesis tests concerning the regression coefficients when the data exhibited heteroskedasticity. Heteroskedasticity was shown to have an impact on the distribution of a test's test statistic, in that the regression coefficient's variance/covariance estimate may not have been consistent, thus possibly leading to under-rejection or over-rejection of the null hypothesis. Tests involving HCCMEs were the first methods that showed promise as a remedial measure for heteroskedasticity by not requiring the structure of the heteroskedasticity to be known. Unfortunately, HCCMEs do not fare well when the sample sizes are too small. In order to alleviate this problem the bootstrap was considered as a possible means to correct for these size distortions. The wild bootstrap is regarded as the best bootstrap based heteroskedasticity robust inference method for regression in the literature.

In the study presented here the wild bootstrap served as the benchmark against which a proposed new modified bootstrap residuals approach could be compared to determine if it had any merit. Size and power of the test were used as measures of performance to determine the following: What impact the different residual transformations $f_{\bullet}$, HCCMEs $HC_{\bullet}$ and ancillary distributions $F_{\bullet}$ had on the size and power of the tests and, moreover, if the performance of the proposed modified bootstrap residuals approach had comparable performance to the wild bootstrap approach. The next two sections will conclude this text by providing an analysis of the simulation results and a few suggestions for future research.

6.1 Conclusions from the simulation study

The most striking results from the simulation study were:

- When large leverage values were present (small samples and lognormal predictors), the sizes of the tests were extremely distorted (in most cases they over-rejected by 5 percent or more than the specified significance level) when the data was homoskedastic, but these sizes eventually decreased monotonically towards the specified significance level as increasing heteroskedasticity was introduced. The implication is that the practitioner would need to exercise caution in cases where data had large leverages and was also homoskedastic. They will need to perform detailed diagnostics before attempting to use these tests.
- The two transformations used for the residuals, f2 and f3, did not seem to have a significant impact on the size results when used with either HC2 or HC3.
- When considering the new modified bootstrap residuals approach using small samples and lognormal predictors, HC3 provided rejection frequencies closest to the specified significance level, but as the sample size increased it turned out that HC1 showed better results. It is recommended that HC1 is used for large samples and HC3 for small samples.
- The modified bootstrap residuals approach showed promising results versus the wild bootstrap methods by using the structured heteroskedastic data, but this may be attributed to the way in which the data was generated. A quick diagnostic test to check if the method is suitable in practice might be to fit a linear model to the squared residuals and determine goodness-of-fit measures like adjusted R^2 . If the model fits well then the approach would conceivably also work well.
- If the practitioner would like to achieve the specified significance level in the case of lognormal predictors using small sample sizes, then he/she is confronted by an array of options for implementing the wild bootstrap approach, i.e., the choice of HCCME, residual transformation and ancillary distribution. This forces the practitioner to identify the level of heteroskedasticity and leverage values present which is unnecessarily complex and may discourage its use in practice. However, on the bright side, these tests were more powerful than the tests that used the uniform predictors.

6.2 Future considerations

The study presented here gave some valuable insight into the wild bootstrap tests using HCCMEs, but it also opened up a host of new questions that could serve as future topics of research. We discuss a few of these below.

- The performance of the modified bootstrap residuals approach with unstructured heteroskedasticity using the linear approximation was shown to be excellent in this study, but better insight into the practicality of the approach could be achieved if additional settings for the heteroskedasticity structure were considered. Since the errors were generated from a normal distribution, the ancillary distributions which took symmetry into account could have been the reason for their excellent performance. Generating for asymmetric distributions might reveal interesting results. This would provide further insight on which ancillary distribution to use with the wild bootstrap.
- Other means of modelling the heteroskedastic structures when applying the modified bootstrap residuals approach, as suggested in Chapter 4, may provide a better refinement over the linear approximation. However, this would limit the usefulness of the approach to a smaller audience.
- If more comprehensive simulation study evidence suggests that the modified method may have merit, the asymptotic properties and theory of this approach can be developed.

Bibliography

- Aitken, A. C. (1935). On least squares and linear combinations of observations, *Proceedings of the Royal Statistical Society* **55**: 42–48.
- Amemiya, T. (1973). Generalized least squares with an estimated autocovariance matrix, *Econometrica* **41**: 723–732.
- Breusch, T. and Pagan, A. (1979). A simple test for heteroskedasticity and random coefficient variation, *Econometrica* **47**: 1287–1294.
- Chesher, A. and Austin, G. (1991). The finite-sample distributions of heteroskedasticity robust Wald statistics, *Journal of Econometrics* **47**: 153–173.
- Chesher, A. and Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator, *Econometrica* **55**: 1217–1222.
- Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form, *Journal of Computational Statistics and Data Analysis* **45**: 215–233.
- Cribari-Neto, F. and Lima, M. (2009). Heteroskedasticity-consistent interval estimators, *Journal of Statistical Computation and Simulation* **79**: 787–803.
- Cribari-Neto, F., Souza, T. C. and Vasconcellos, K. (2007). Inference under heteroskedasticity and leveraged data, *Communications in Statistics: Theory and Methods* **36**: 1977–1988. also see Erratum (2008), **37**, 3329–3330.
- Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last, *Journal of Econometrics* **146**(1): 162–169.
- Davidson, R. and MacKinnon, J. (1993). *Estimation and Inference in Econometrics*, Oxford University Press.

- Davidson, R. and MacKinnon, J. (2004). *Econometric Theory and Methods*, Oxford University Press.
- de Beer, C. F. and Swanepoel, J. W. H. (1989). A modified Durbin-Watson test for serial correlation in multiple regression under nonnormality using the bootstrap, *Journal of Statistical Computation and Simulation* **33**: 75–81.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Annals of Statistics* **7**(1): 1–26.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for industrial and applied mathematics. Philadelphia, Pennsylvania.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *The Annals of Statistics* **32**(2): 407–451.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall.
- Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions, *Annals of Mathematical Statistics* **34**: 447–456.
- Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors, *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed. L. M. Le Cam and J. Neyman **1**: 59–82. University of California Press.
- Flachaire, E. (1999). A better way to bootstrap pairs, *Economics Letters* **64**(3): 257–262.
- Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap, *Computational Statistics and Data analysis* **49**(2): 361–376.
- Freedman, D. A. (1981). Bootstrap regression models, *Annals of Statistics* **9**(6): 1218–1228.
- Glejser, H. (1969). A new test for heteroskedasticity, *Journal of the American Statistical Association* **64**: 316–323.
- Goldfeld, S. M. and Quandt, R. E. (1965). Some tests for homoskedasticity, *American Statistical Association* **60**(310): 539–547.

- Greene, W. (2003). *Econometric Analysis*, 5th edn, Prentice Hall.
- Gujarati, D. (2004). *Basic Econometrics*, Tata McGraw Hill.
- Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing, *Biometrics* **47**: 179–192.
- Hausman, J. and Palmer, C. (2012). Heteroskedasticity-robust inference in finite samples, *Economic Letters* **116**(2): 232–235.
- Hayes, A. and Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation, *Journal of Behaviour Research Methods* **39**: 709–722.
- Hinkley, D. V. (1977). Jackknifing in unbalanced situations, *Technometrics* **19**: 285–292.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**(1): 55–67.
- Horn, S., Horn, R. A. and Duncan, D. (1975). Estimating heteroskedastic variances in linear models, *Journal of the American Statistical Association* **70**: 380–385.
- Horowitz, J. L. (1995). Bootstrap methods in econometrics: Theory and numerical performance, *Symposium on Bootstrap Methods in Econometrics, 7th World Congress of the Econometric Society*.
- Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th edn, Pearson international.
- Knaus, J. (2013). *Snowfall: Easier cluster computing (based on snow)*. R package version 1.84-6 <http://CRAN.R-project.org/package=snowfall>.
- Koenker, R. and Bassett, G. (1982). Robust test for heteroskedasticity based on regression quantiles, *Econometrica* **50**: 43–61.
- Kutner, M., Nachtheim, C., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*, 5th edn, McGraw Hill.

- Lancaster, T. (2006). A note on bootstraps and robustness. Brown University Department of Economics Working Paper No. 2006-06.
- Lindsay, B. G. (1989a). Moment matrices: Applications in mixtures, *Annals of Statistics* **17**(2): 722–740.
- Lindsay, B. G. (1989b). On the determinants of moment matrices, *Annals of Statistics* **17**(2): 711–721.
- Liu, R. Y. (1988). Bootstrap procedures under some non i.i.d. models, *Annals of Statistics* **16**(4): 1696–1708.
- Long, J. S. and Ervin, L. H. (1998). Correcting for heteroskedasticity consistent standard errors in the linear regression model: Small sample consideration. Indiana University.
- MacKinnon, J. G. (2002). Bootstrap inference in Econometrics, *Canadian Journal of Econometrics* **35**: 615–645.
- MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference, in X. Chen and N. R. Swanson (eds), *Recent Advances and Future Directions in Causality, Prediction, and Specification analysis*, Springer, pp. 437–462. Essays in Honor of Halbert L. White Jr.
- MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties, *Journal of Econometrics* **29**: 305–325.
- Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models, *Annals of Statistics* **21**: 255–285.
- Martin, M. A. (2007). Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties, *Computational Statistics and Data Analysis* **51**: 6321–6342.
- McCullough, J. H. (1985). Miscellanea on heteros*edasticity, *Econometrica* **53**(2): 483.
- Merriam-Webster Dictionary (2015). www.merriam-webster.com. Date of access: 12 November 2015.

- Paparoditis, E. and Politis, D. N. (2005). Bootstrap hypothesis testing in regression models, *Statistics and Probability Letters* **74**: 356–365.
- Park, R. E. (1966). Estimation with heteroskedastic error terms, *Econometrica* **34**(4): 888.
- Pearson, K. (1930). *The Life, Letters and Labors of Francis Galton*, Cambridge University press.
- Politis, D. N. and Poulis, S. (2014). Heteroskedastic linear regression: steps towards adaptivity, efficiency, and robustness, *Topics in Nonparametric Statistics*, Vol. 74, Springer, pp. 283–297. Springer Proceedings in Mathematics & Statistics.
- Pratt, J. and Schlaifer, R. (1985). On the nature of discovery of structure, *Journal of the American Statistical Association* **79**: 9–21.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*, Springer series in Statistics, Springer.
- Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors, *Journal of Statistical Education* **9**(3).
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1): 267–288.
- Webb, M. D. (2014). Reworking wild bootstrap based inference for clustered errors. Queen’s Economics Department Working Paper No. 1315.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and direct test for heteroskedasticity, *Econometrica* **48**(4): 817–838.
- Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis, *Annals of Statistics* **14**(4): 1261–1295.
- Young, A. (1988). Resampling tests of statistical hypotheses, in D. Edwards and N. E. Raun (eds), *Proceedings in Computational Statistics 8th Symposium held in Copenhagen*, Physica-Verlag HD, pp. 233–238.

- Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators, *Journal of Statistical Software* **11**(10). <http://www.jstatsoft.org/v11/i10/>.
- Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships, *R News* **2**(3). <http://CRAN.R-project.org/doc/Rnews/>.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B* **67**(2): 301–320.