Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
A Diagnostic for the General Linear Model: An Application to Time Series

by

Carl Sullivan
Thesis submitted for the degree of Doctor of Philosophy

Department of Statistics
Trinity College Dublin

Supervisor: Prof. John Haslett,
Department of Statistics, Trinity College Dublin

October, 2001
Declaration

This thesis is entirely the author's own work. The contents of this thesis have not been submitted as an exercise for a degree at any other University. The contents of this thesis may be lent or copied without the author's consent by the Library at Trinity College Dublin.

Signature of Author

Carl Sullivan
Summary

An outlier is an observation which is thought to be unusual. The detection of such extreme values is an important issue. Developing a model based on data containing even a single outlier can seriously bias population inferences. The valuable role marginal and conditional residuals play in assuring model robustness is well established in the context of general linear models. The purpose of this thesis is to explore the potential of a statistic, developed for general linear models which incorporates both marginal and conditional residuals, as a diagnostic tool for time series and longitudinal data.

Theory is developed which enables this statistic to be applied in a wider range of situations. Empirical evidence is presented which displays the diagnostic's ability at detecting a wide variety of aberrant behaviour in times series. Particular attention is paid to single outliers and groups of outlying points. These methods generalize for multi-dimensional variables. Detecting influential observations is examined as well.

Computationally cheap approximations and bounds are proposed for this diagnostic and empirical evidence provides a measure of accuracy.
Acknowledgements

I would like to thank my parents for encouraging my pursuit of knowledge and I promise to get a job soon. A special thanks to my supervisor, John Haslett, for all his time and enthusiasm.

Thanks Deazer for being a friend and putting up with my thursday rants. Thanks for the camping experience Vince. I'd like to thank Cassian for all the games of pool and snooker, and yes I'm in the lead frame-wise. Thanks for the reindeer meat and the Finnish vodka Niko. Rome was great and Teramo was even better thanks Aldo (My Master). Greek culture seems surprisingly similar to Irish culture someday Georgos I will make it to Greece. Thanks Gilbert and Jan for the experience that was Keele University and of course the fine English Ales as well. Finally, I'd like to thank Cathal, Clare, Breedette, Ronnie, Pete, Eoin, Ingelin and Eleisa for making postgraduate life a little more interesting.

"I can't understand mathematics without a real problem. I've learned more mathematics than is good for any one person in a lifetime...But it's certainly true that I love mathematics..." Diaconis [19].
Contents

1 Introduction .................................................. 1
  1.1 Summary of research ........................................ 1
  1.2 Extensions in Original Field ............................ 2
  1.3 Computationally Cheap Approximations ............. 3
  1.4 Application to Time Series ............................. 4
  1.5 Diagnostics for Longitudinal Data .................... 5
  1.6 Mahalanobis Distance Approach ....................... 5

2 Contribution Methodology ................................. 7
  2.1 Outlier Diagnostic ......................................... 7
  2.2 Residuals ...................................................... 8
    2.2.1 Marginal Residuals ....................................... 8
    2.2.2 Conditional Residuals ................................. 10
  2.3 Contributions ............................................... 11
  2.4 Standardized Contributions ............................ 13
2.5 Applications ....................................................................................... 14
   2.5.1 Contributions and Simple Linear Models ....................... 15
   2.5.2 Contributions and General Linear Models .................... 20
2.6 Unusual Negative Contributions ........................................................23
   2.6.1 Upper Bound ........................................................................... 23
   2.6.2 Lower Bound ....................................................................... 24
   2.6.3 Application of Bounds ...................................................... 25
2.7 General Linear Models to Time Series ...............................................27

3 Approximating Standardized Contributions .............................................. 28
   3.1 Computational Efficiency ....................................................... 28
   3.2 Influence and Ranges ................................................................ 29
   3.3 Subsets ...................................................................................... 33
   3.4 Fixed effects are known ......................................................... 36
   3.5 Fixed effects are unknown .................................................... 40
      3.5.1 Independent Observations ............................................. 40
      3.5.2 General Covariance Structure ........................................44
   3.6 A Cheap Approximation of $\tau_A$ ........................................... 48
   3.7 Conclusions ............................................................................. 50

4 Time Series Analysis ............................................................................ 52
   4.1 Contributions and Time Series .............................................. 52
List of Tables

2.1 Analysis of Forbes' Data ......................................................... 16
2.2 Analysis of Huber's Data ........................................................ 18
2.3 Analysis of Ice-cream Data ...................................................... 22
2.4 Negative Contributions of Ice-cream data ............................... 26
3.1 $\phi_A$ and $\gamma_A$ associated with Forbes' Data ....................... 43
3.2 $\phi_A$ and $\gamma_A$ associated with Huber's Data ....................... 44
4.1 Diagnostic Results for a Single Additive Outlier ....................... 65
4.2 Diagnostic Results for a Single Innovation Outlier ..................... 69
4.3 Diagnostic Results for a Patch of Additive Outliers .................... 73
4.4 Unusual Contributions for Nile Data ......................................... 78
5.1 Contributions for Individual with Single Outlying Observation ........ 85
5.2 Contributions for Outlying Individual ....................................... 88
5.3 Observations with Unusual Contributions in Zerbe's Data .......... 91
5.4 Unusual Standardized Residuals of Observations associated with Outlying Individuals ...................................................... 93
6.1 $M_i$ Analysis of a Single Additive Outlier ................................. 101
6.2 $M_i$ Analysis of a Single Innovation Outlier ........................................ 102
6.3 $M_i$ Analysis of a Patch of Additive Outliers .................................... 103
6.4 Unusual $M_i$ values for Nile Data ....................................................... 104
List of Figures

2.1 Forbes’ Data ................................................................. 15
2.2 Huber’s Data ............................................................... 18
2.3 Ice-cream Data ............................................................ 21
3.1 $\phi_A$ and $\gamma_A$ associated with AR(1) processes ......... 39
3.2 $\phi_A$ and $\gamma_A$ associated with Ice-cream Data ........... 47
3.3 $T_A$ and $\tau_A$ associated with Ice-cream Data ............... 47
3.4 $\tau_A$ and $\tau_A^*$ associated with Ice-cream Data ........... 50
4.1 Single Additive Outlier .................................................. 63
4.2 Rate of Detection for An Additive Outlier ....................... 65
4.3 Single Innovation Outlier .............................................. 68
4.4 Rate of Detection for An Innovation Outlier ..................... 70
4.5 Patch of Additive Outliers .......................................... 72
4.6 Rate of Detection for a Patch of Additive Outliers .......... 74
4.7 Rate of Detection for a Patch of Additive Outliers using $DV(i)$ 75
4.8 Annual Flow Volume of the Nile .................................... 76
4.9 Level Shift Diagnostics ................................................ 79
4.10 Level Shift Diagnostics with the year 1913 Omitted .......... 79

5.1 Single Outlying Observation ............................................. 84
5.2 Single Outlying Individual ............................................... 87
5.3 Zerbe's Data ................................................................. 89
5.4 Expected Plasma Inorganic Phosphate ............................. 90
5.5 $\tau_A$ versus $D_A$ for Individuals ................................. 92
5.6 $\phi_A$ and $\gamma_A$ associated with an Individual ............... 94
5.7 $T_A$, $\tau_A^*$ and $\tau_A$ associated with an Individual ....... 95

6.1 P values of $M(A)$ ............................................................. 105
Chapter 1

Introduction

1.1 Summary of research

Detecting and explaining observations that are outlying has generated considerable material in books and journals. New developments in theory and application are continuously occurring. The detection of outliers continues to remain an important issue. As new models are developed and applied, an understanding of their robustness and sensitivity to unusual observations is important. A single outlier or just a small subset of unusual observations may greatly influence the inferences made about the entire population. Outliers may also provide evidence of model inadequacies.

This thesis focuses on the viability of using a statistic, developed for general linear models by Haslett and Hayes [29], to detect unusual observations
in time series and longitudinal data. Theoretical developments presented in this thesis allows the application of this statistic to a wider range of situations. Empirical evidence leads to the development of computationally cheap approximations and bounds for this diagnostic.

This statistic is referred to as the contribution of a single datum or collection of data points to the lack of fit. The contribution statistic has an additive property, detailed in chapter 2, which makes it highly computationally efficient compared to existing diagnostics. Remarkably this additive property is valid even when observations are correlated such as those common to time series and longitudinal data. It is this property which instigated the research contained in this thesis.

1.2 Extensions in Original Field

Chapter 2 of this thesis defines, explores and extends this statistic in its original field of development. In the case of simple linear models, $Y \sim N(X\beta, \sigma^2 I)$, the relationship between internally studentized residuals and the standardized version of this statistic is established. Clearly, in this case, the standardized contributions have a simple and easily interpretable meaning. The contributions and their standardized versions are always positive if $Y \sim N(X\beta, \sigma^2 I)$. 
When $Y \sim N(X\beta, V)$ the contributions and the standardized contributions may take negative values. Considerable effort is spent in Haslett and Hayes [29] to show how unusual observations can be detected by examining the contributions of the data points to the lack of fit. However all of these unusual observations generate large positive contributions. The authors were unable to establish what information lies in the negative tail of the contributions’ distributions. Indeed, no methods were even proposed to detect unusually large negative contributions.

The problem of large negative contributions is first detailed in chapter 2. Empirical evidence in chapter 4 indicates a type of outlier found in time series that can generate large negative contributions. Theory is developed in chapter 2 to enable the detection of unusual observations which generate unusually large negative contributions. These theoretical developments make the contribution diagnostics a far more comprehensive methodology for outlier detection.

1.3 Computationally Cheap Approximations

Obtaining standardized contributions can be considerably more computationally intensive than just calculating the contributions. This burden grows as the size of the subset to which the standardized contribution corresponds
increases. To standardize a contribution two parameters must be evaluated. One parameter is computationally cheaper than the other hence it is more desirable to work with this easily obtained parameter if possible.

Chapter 3 focuses on the relationship between these two parameters. Theory is derived which establishes when both these parameters are equal. Empirical evidence suggests situations when they are approximately the same. Their relationship to subset size is also explored. As a result of this analysis the conditions under which the contributions are themselves good approximations of their standardized versions are established. For cases when the contributions differ greatly from their standardized contributions computationally cheap approximations are proposed.

1.4 Application to Time Series

Serial correlation is fundamental to time series, hence any diagnostics applied to these series must be able to utilize the correlation structure. Haslett and Hayes [29] go to great lengths to emphasize the ability of the contribution methodology to detect outliers when observations are correlated; in fact all the examples presented in their paper assume non-independence. These examples along with the examples and theory in chapter 2 justify exploring the contribution approach with respect to time series analysis.
Chapter 4 reviews existing diagnostics in time series analysis and compares them to the contribution approach. Empirical evidence illustrates the ability of contribution diagnostics at detecting additive and innovation outliers, see Fox [23]. The detection of unusual subsets and level shifts are also explored.

1.5 Diagnostics for Longitudinal Data

The contribution methodology is applied to longitudinal data in chapter 5. A very general class of models is considered in Chi and Reinsel [10] which allow for fixed effects, random effects and for the errors to have a serial correlation structure. Empirical evidence is presented of the standardized contribution diagnostic's ability at detecting both outlying observations and individuals for such models. The data set analyzed in Chi and Reinsel [10] provides further empirical evidence of the accuracy of the computationally cheap approximations put forward in chapter 3 for the standardized contributions.

1.6 Mahalanobis Distance Approach

An alternative methodology to the contribution approach is put forward in chapter 6. A new statistic is proposed based on the same residuals which
generate the contributions. This statistic is geometrically more interpretive than the contribution approach. Its relationship to the contribution statistic is theoretical established.

Examples are presented which illustrates the ability of this new diagnostic at detecting a range of unusual behaviour experience by time series. This diagnostic is not as computationally efficient as the contribution approach and numerical difficulties can arise due to singular covariance matrices.

Chapter 2

2.1 Outlier Diagnostic

This chapter examines, in the context of general linear models, the construction and use of an outlier diagnostic referred to as the contribution of a deviant collection of data points to the lack of fit, see Hardast and Nace (1997). A standardized version is available and is simply known as the standardized contribution. The properties of contributions and their standardized versions are first explored for simple linear models \( Y \sim N(XB, \sigma^2 I) \). This exploration reveals the relationship between studentized residuals and standardized contributions. Unlike simple linear models, general linear models (G.L.M.’s) can generate negative contributions. Theory is developed which enables the detection of deviating points corresponding
Chapter 2

Contribution Methodology

2.1 Outlier Diagnostic

This chapter examines, in the context of general linear models, the generation and use of an outlier diagnostic referred to as the contribution of a datum or collection of data points to the lack of fit, see Haslett and Hayes [29]. A standardized version is available and is simply known as the standardized contribution. The properties of contributions and their standardized versions are first explored for simple linear models $Y \sim N(X\beta, \sigma^2 I)$. This exploration reveals the relationship between studentized residuals and standardized contributions. Unlike simple linear models, general linear models $Y \sim N(X\beta, V)$ can generate negative contributions. Theory is developed in this chapter which enables the detection of outlying points corresponding to
large negative contributions. These new developments aid in the interpretation and range of applications of the contribution methodology.

2.2 Residuals

The concept of a residual is fundamental to applied statistics. In its simplest form it can be defined as:

\[ \text{residual} = \text{observation} - \text{fitted value}. \]

Given a set of observations \( Y \) and a set of estimates \( \hat{Y} \) for these observations the vector of residuals are given by \( Y - \hat{Y} \). Clearly the values of the residuals depends on how \( \hat{Y} \) was obtained. Thus different estimates of \( Y \) will generate different residuals. By examining residuals the appropriateness of models can be assessed, the strength of underlying assumptions of normality or homoscedasticity can be gauged and outlying observations can be detected.

Two different types of residuals are used to construct the contribution statistic and they are referred to as marginal and conditional residuals. Both these residuals are reviewed in the following subsections.

2.2.1 Marginal Residuals

Given a set of observations \( Y \sim N(X\beta,V) \), where \( \beta \) and \( V \) are known then the set of marginal residuals are defined as \( e = Y - E[Y] = Y - X\beta \). The
marginal residual associated with the data point \( i \) is denoted \( e_i \) and \( e_A \) corresponds to the marginal residuals of the set of observations \( A \).

In many situations, \( \beta \), commonly referred to as the fixed effects of the model, is unknown and has to be estimated. The marginal residuals are now defined as \( \hat{e} = Y - X\hat{\beta} \) where \( \hat{\beta} \) is an estimate for the unknown parameter \( \beta \). For general linear models the classical estimate for \( \beta \) is given by 
\[
\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y = BY.
\]

It is also common for the covariance matrix of \( Y \), denoted \( V \), to be unknown. The covariance matrix \( V \) can be written as \( V = \sigma^2 R(\Theta) \) where \( \sigma^2 \) is associated with the variance of individual components of \( Y \) and \( R(\Theta) \) describes the correlation structure associated with \( Y \). For general linear models the classical estimate for \( \sigma^2 \) is given by 
\[
\hat{\sigma}^2 = \hat{\varepsilon}^T R(\Theta)^{-1} \hat{\varepsilon} / (n - p)
\]
where \( p \) is the rank of \( X \) and \( n \) is the number of observations in \( Y \). The correlation parameters used to construct \( R(\Theta) \) are denoted by \( \Theta \) and if they are unknown they can be estimated by maximum likelihood or restricted maximum likelihood. For the moment, it is assumed \( V \) is known or an estimate is available. In this chapter, the covariance matrix \( V \) or its estimate will simply be referred to as \( V \).

Previously, the marginal residuals were defined as \( \hat{e} = Y - X\hat{\beta} \). The marginal residuals can be re-expressed as \( \hat{e} = VQY \) where \( Q = V^{-1} - P \), and \( P = V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} \). The covariance matrix of \( \hat{e} \) is defined
as $\text{var}(e) = VQV = G$. The variance of a single marginal residual $e_i$ is given by the $i^{th}$ diagonal element of the matrix $G$ hence $g_{ii} = \text{var}(e_i)$. The variance of $e_A$ follows and is given by $G_{AA} = \text{var}(e_A)$. This representation can be found in Haslett and Hayes [29].

2.2.2 Conditional Residuals

The conditional residual, $\tilde{e}(i)$, for an observation $y_i$ is defined as:

$$\tilde{e}(i) = y_i - \tilde{y}(i)$$

where $\tilde{y}(i) = E[y_i | y_i; \hat{\beta}(i), V]$. $\tilde{y}(i)$ is the best linear unbiased predictor (BLUP), see Robinson [49], of $y_i$ given the observations $y_i = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. $\hat{\beta}(i)$ is used to denote the estimate of the fixed effects $\beta$ omitting the observation $y_i$. However the estimate of the variance structure $V$ is based on the entire data set. The conditional residuals associated with a subset $A$ is given by:

$$\tilde{e}(A) = y_A - \tilde{y}(A)$$

where $\tilde{y}(A) = E[y_A | y_A; \hat{\beta}(A), V]$. The expected value of the conditional residuals is zero and their variance is $D_A = (Q_{AA})^{-1}$, see Haslett and Hayes [29].

The following formula

$$D_A^{-1}\tilde{e}(A) = (V^{-1}e)_A$$
shows the relationship between the marginal and conditional residuals. It is interesting to note that Haslett and Hayes [29] showed this relationship in the context of general linear models and that de Jong and Penzer [18] proved it in a time series context. In time series analysis it is common for the conditional residual to be referred to as a cross-validation residual, see Christensen [11] section 2.2, or as a leave-\( \kappa_A \)-out residual in Bruce and Martin [8]. However, leave-\( \kappa_A \)-out residuals as defined in Bruce and Martin [8] involve the re-estimation of the covariance structure with \( \kappa_A \) observations removed whereas in this thesis \( V \) is not re-estimated.

### 2.3 Contributions

The contributions arise from the decomposition of the lack of fit which is denoted \( S \) and is defined as \( S = (Y - X\hat{\beta})^T V^{-1}(Y - X\hat{\beta}) = \hat{e}^T V^{-1}\hat{e} \). It is possible to express \( S \) in terms of both marginal and conditional residuals

\[
S = \sum_{i=1}^{n} d_i^{-1} \hat{e}_i \hat{e}_i^{(i)} = \sum_{i=1}^{n} T_i
\]

where \( d_i = \text{var}(\hat{e}_i) \) and \( T_i = d_i^{-1} \hat{e}_i \hat{e}_i^{(i)} \) is referred to as the contribution of the data point \( i \) to the lack of fit \( S \). This decomposition follows from the relationship \( d_i^{-1} \hat{e}_i^{(i)} = (V^{-1}\hat{e})_i \). If \( \hat{e}_i \) and \( \hat{e}_i^{(i)} \) are not of the same sign then \( T_i \) will be negative. It is also possible to express \( S \) as the sum of contributions for any partitioning of the data set. If the data set has \( n \) observations then let
\( I = \{1, 2, \ldots, n\} \) denote the indexing set. For time series a natural indexing set exists, however for general linear models an artificial indexing may have to be defined. \( I \) maybe partitioned into an arbitrary number of subsets, \( I = \{ A_1, A_2, \ldots, A_r \} \) of dimensions \( \kappa_1, \kappa_2, \ldots, \kappa_r \) such that \( n = \sum_{i=1}^{r} \kappa_i \) and \( \kappa_i \leq n - p \) for all \( i \). This partitioning of the indexing set allows the lack of fit to be expressed as follows

\[
S = \sum_{i=1}^{r} T_{A_i}
\]

where \( T_{A_i} = \bar{e}_{A_i}^T D_{A_i}^{-1} \bar{e}_{(A_i)} \) is referred to as the contribution of the data subset \( A_i \) to the lack of fit \( S \). The contribution \( T_A \), of an arbitrary subset \( A \), can be expressed as

\[
T_A = \sum_{i \in A} T_i
\]

which is a summation of the individual contributions of each datum indexed by the set \( A \). Thus the contributions of any subset can be obtained by adding the contributions of each datum in the subset. Thus the contribution for any arbitrary subset can be generated with computational ease.

Haslett and Hayes [29] show that the contribution of a subset \( A \) can be expressed as follows

\[
T_A = \left( \gamma_A + \phi_A \right) v_1 - \left( \gamma_A - \phi_A \right) v_2
\]

where \( v_1 \) and \( v_2 \) are independent \( \chi^2_{\kappa_A} \) random variables. The expected value of \( T_A \) is denoted \( E[T_A] = \phi_A \) and due to the additive property of the contri-
butions $E[T_A] = \sum_{i \in A} E[T_i] = \sum_{i \in A} \phi_i = \sum_{i \in A} (VQ)_{ii}$, see Haslett and Hayes [29]. The variance of the contributions is given by $var(T_A) = \frac{\gamma_A^2 + \phi_A^2}{\kappa_A}$ and $\gamma_A$ is defined as follows

$$\gamma_A = \text{tr}\{ (G_{AA}^{\frac{1}{2}} D_A^{-1} G_{AA}^{\frac{1}{2}}) \} = \sum_{i \in A} \pi_i^{\frac{1}{2}}$$

where $\pi_i$ are the eigenvalues of $F = G_{AA}^{\frac{1}{2}} D_A^{-1} G_{AA}^{\frac{1}{2}}$, see Haslett and Hayes [29]. The properties of $\phi_A$ and $\gamma_A$ are explored in chapter 3.

Although the expected value of $T_A$ and its variance are available, calculating $P(T_A \leq t)$ where $t$ is an arbitrary real number can only be achieved by simulation or numerical approximation methods, see Mathai and Provost [41]. However, this representation of contributions allows the application of a transformation found in Mathai and Provost [41], section 4.6, which enables $T_A$ to be standardized to an approximately $\chi^2_{\kappa_A}$ random variable. This new statistic is referred to as the standardized contribution.

### 2.4 Standardized Contributions

Standardized contributions provide a means of judging whether or not the contribution of a single datum or set of observations is significantly different from the expected contribution. The standardized contributions of a set $A$
is denoted $\tau_A$ and is defined as

$$\tau_A = \kappa_A \left\{ \frac{(T_A - \phi_A)\sqrt{2}}{\sqrt{\gamma_A^2 + \phi_A^2}} + 1 \right\}$$

where $E[\tau_A] = \kappa_A$ and $\text{var}(\tau_A) = 2\kappa_A$. $\tau_A$ is approximately $\chi^2_{\kappa_A}$ distributed.

Unfortunately the standardizing contribution $\tau_A$ does not share the same additive property as $T_A$ thus

$$\tau_A \neq \sum_{i \in A} \tau_i.$$

As a result generating $\tau_A$ requires a lot more computations then $T_A$.

2.5 Applications

Three data sets are examined in this section. A simple linear model is assumed for two of these data sets while a general linear model is applied to the third. All the models require the estimation of two fixed effects: an intercept and a slope. Each of these examples contains only one outlier which is represented graphically by the symbol $\oplus$. However each outlier provides its own unique challenge. Leverage is not an issue for Forbes' data which can be found in Weisberg [50], and forms the first example. The second data set analyzed is Huber's data and leverage is a major concern, see Huber [30]. The Ice-cream data which can be found in Kadiyala [34] provides an example where leverage is not an issue but there is a very high correlation be-
between observations. The ability of the contributions to detect these unusual observations with this variety of challenges is shown.

2.5.1 Contributions and Simple Linear Models

It is assumed in this subsection that $Y \sim N(X\beta, \sigma^2 I)$. The first data set to be examined is Forbes’ data and is displayed in Figure 2.1. This data set was collected in the 1840s and 1850s by James D. Forbes, see Weisberg [50], and was used to estimate altitude by understanding the relationship between the boiling point of water and air pressure. It is used in Weisberg [50] as a means
Table 2.1: Analysis of Forbes' Data

<table>
<thead>
<tr>
<th>Index</th>
<th>$y_i$</th>
<th>$x_i$</th>
<th>$T_i$</th>
<th>$\tau_i$</th>
<th>$\hat{e}_i$</th>
<th>$\tilde{e}_{(i)}$</th>
<th>$h_{ii}$</th>
<th>$D_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>131.79</td>
<td>194.5</td>
<td>0.42</td>
<td>0.53</td>
<td>-0.25</td>
<td>-0.31</td>
<td>0.19</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>131.79</td>
<td>194.3</td>
<td>0.03</td>
<td>0.04</td>
<td>-0.07</td>
<td>-0.08</td>
<td>0.20</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>135.02</td>
<td>197.9</td>
<td>0.03</td>
<td>0.03</td>
<td>-0.06</td>
<td>-0.07</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>135.55</td>
<td>198.4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.10</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>136.46</td>
<td>199.4</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>0.04</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>6</td>
<td>136.83</td>
<td>199.9</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.04</td>
<td>-0.05</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>137.82</td>
<td>200.9</td>
<td>0.02</td>
<td>0.02</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>8</td>
<td>138.00</td>
<td>201.1</td>
<td>0.02</td>
<td>0.02</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>9</td>
<td>138.06</td>
<td>201.4</td>
<td>0.17</td>
<td>0.18</td>
<td>-0.16</td>
<td>-0.17</td>
<td>0.06</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>138.05</td>
<td>201.3</td>
<td>0.04</td>
<td>0.04</td>
<td>-0.08</td>
<td>-0.08</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>11</td>
<td>140.04</td>
<td>203.6</td>
<td>0.15</td>
<td>0.16</td>
<td>-0.15</td>
<td>-0.15</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>12</strong></td>
<td><strong>142.34</strong></td>
<td><strong>204.6</strong></td>
<td><strong>12.87</strong></td>
<td><strong>13.75</strong></td>
<td><strong>1.36</strong></td>
<td><strong>1.45</strong></td>
<td><strong>0.06</strong></td>
<td><strong>7.34</strong></td>
</tr>
<tr>
<td>13</td>
<td>145.47</td>
<td>209.5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.14</td>
<td>0.00</td>
</tr>
<tr>
<td>14</td>
<td>144.34</td>
<td>208.6</td>
<td>0.73</td>
<td>0.82</td>
<td>-0.32</td>
<td>-0.37</td>
<td>0.12</td>
<td>0.47</td>
</tr>
<tr>
<td>15</td>
<td>146.30</td>
<td>210.7</td>
<td>0.41</td>
<td>0.50</td>
<td>-0.24</td>
<td>-0.29</td>
<td>0.17</td>
<td>0.30</td>
</tr>
<tr>
<td>16</td>
<td>147.54</td>
<td>211.9</td>
<td>0.04</td>
<td>0.05</td>
<td>-0.08</td>
<td>-0.10</td>
<td>0.21</td>
<td>0.03</td>
</tr>
<tr>
<td>17</td>
<td>147.80</td>
<td>212.2</td>
<td>0.05</td>
<td>0.07</td>
<td>-0.09</td>
<td>-0.11</td>
<td>0.22</td>
<td>0.04</td>
</tr>
</tbody>
</table>
of introducing the concept of simple linear regression. The Forbes' data is used in Atkinson [2] to illustrate the detection and effects of an outlier on simple linear models. Observation 12 of this data is deemed to be unusual in Atkinson [2]. From Table 2.1 it is clear that the contribution associated with this observation is extremely large compared to the contributions generated by the other data points. The standardized contributions provides further evidence that observation 12 is outlying. Cook's distance denoted $D_i$, see Cook [13], reveals that this datum is also influential in the estimation of the fixed effects. The leverage of the $i^{\text{th}}$ observation is denoted $h_{ii}$ and from Table 2.1 it is clear no observation has an unusually large leverage.

The next data set explored is Huber's data and is displayed in Figure 2.2. This data is a synthetic example used in Huber [30] to illustrate the effects of points with high leverage on model selection and parameter estimation. This example differs from any application presented in Haslett and Hayes [29] as the issue of leverage arises. Cook's distance generates an exceptionally large value for observation 6, see Table 2.2. It is interesting to note that the contribution associated with this datum does not appear unusual when contrasted with the contributions of the other data points. However its standardized contribution indicates that this point is unusual. What causes this dramatic difference?

Answering this question involves simplifying $T_A$ and $\tau_A$. Haslett and
Figure 2.2: Huber’s Data

Table 2.2: Analysis of Huber’s Data

<table>
<thead>
<tr>
<th>Index</th>
<th>$y_i$</th>
<th>$x_i$</th>
<th>$T_i$</th>
<th>$\tau_i$</th>
<th>$\hat{e}_i$</th>
<th>$\tilde{e}(i)$</th>
<th>$h_{ii}$</th>
<th>$D_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.48</td>
<td>-4</td>
<td>1.81</td>
<td>2.55</td>
<td>2.09</td>
<td>2.94</td>
<td>0.29</td>
<td>1.79</td>
</tr>
<tr>
<td>2</td>
<td>0.73</td>
<td>-3</td>
<td>0.07</td>
<td>0.09</td>
<td>0.42</td>
<td>0.55</td>
<td>0.24</td>
<td>0.06</td>
</tr>
<tr>
<td>3</td>
<td>-0.04</td>
<td>-2</td>
<td>0.03</td>
<td>0.04</td>
<td>-0.27</td>
<td>-0.34</td>
<td>0.20</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>-1.44</td>
<td>-1</td>
<td>1.05</td>
<td>1.27</td>
<td>-1.59</td>
<td>-1.93</td>
<td>0.17</td>
<td>0.77</td>
</tr>
<tr>
<td>5</td>
<td>-1.32</td>
<td>0</td>
<td>0.8</td>
<td>0.96</td>
<td>-1.39</td>
<td>-1.67</td>
<td>0.17</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>0.00</td>
<td>10</td>
<td>0.23</td>
<td>3.62</td>
<td>0.75</td>
<td>11.64</td>
<td>0.93</td>
<td>28.21</td>
</tr>
</tbody>
</table>
Hayes [29] state that when \( Y \sim N(X\beta, \sigma^2 I) \) then the contributions of an arbitrary subset indexed by \( A \) can be expressed as

\[
T_A = \sum_{i \in A} \frac{e_i^2}{\hat{\sigma}^2}.
\]

The authors failed to simplify the standardized contributions. When \( V = \sigma^2 I \) then it follows, see Appendix A.1.3, that the standardized contribution of an arbitrary observation can be expressed as

\[
t_i = \frac{\hat{e}_i^2}{\sigma^2 (1 - h_{ii})}
\]

where \( h_{ii} \) is the leverage of the \( i^{th} \) data point. Thus \( t_i \), in this case, can be viewed as an internally studentized residual which is squared, see Atkinson [2]. This allows the Cook's distance corresponding to this observation to be expressed as

\[
D_i = \frac{t_i h_{ii}}{p(1 - h_{ii})}
\]

where \( p \) is the rank of the design matrix. In this situation, the Cook's distance of a single observation is a scalar multiple of \( t_i \). The standardized contribution of a subset of observations, \( A \), can be expressed as

\[
t_A = \frac{\kappa_A \sum_{i \in A} \hat{e}_i^2}{\sum_{i \in A} \sigma^2 (1 - h_{ii})} = \frac{\kappa_A T_A}{\kappa_A - \sum_{i \in A} h_{ii}}
\]

see Appendix A.1.4.

Obviously there can be a large difference between the contributions and
their standardized versions when leverage is an issue. It is also worth noting that both $T_A$ and $\tau_A$ depend on the square of the marginal residuals and this is the reason they can never be negative when $Y \sim N(X\beta, \sigma^2I)$.

### 2.5.2 Contributions and General Linear Models

It is assumed in this subsection that $Y \sim N(X\beta, V)$. The Ice-cream data set is displayed in Figure 2.3. This data set was collected in the 1950s in America and was used to explore the relationship between ice-cream consumption and such factors as weekly family income, the price of ice-cream and monthly mean temperatures, see Kadiyala [34]. A detailed analysis of this data is performed in Martin [40] with an emphasis on outlier detection. The same model is used for the analysis now performed on the Ice-cream data as that proposed in Martin [40]. This model was selected by Martin [40] on the grounds of parsimony and fit. The data is modeled with a straight line regression on the mean temperature and the errors, $\hat{e}_t$, are assume to be from an autoregressive process of order one

$$\hat{e}_t = \varphi \hat{e}_{t-1} + \varepsilon_t$$

where $\varphi = 0.75$ and $\sigma^2 = 0.00105$, see Diggle [21] section 2.4 for an introduction of autoregressive processes.

Observation 30 is deemed to be outlying in Martin [40] and it can be
seen in Table 2.3 that this observation has both an exceptionally high contribution and standardized contribution. This point is not influential, in fact there are no influential observations present in this data set. Leverage is not an issue.

What makes this data set different from the previous two examples is that some of the contributions and indeed some standardized contributions are negative. Haslett and Hayes [29] note that negative contributions may occur when observations are assumed to be correlated. However, they remain uncertain what, if any, information is to be obtained from these negative values.
Table 2.3: Analysis of Ice-cream Data

<table>
<thead>
<tr>
<th>Index</th>
<th>$y_i$</th>
<th>$x_i$</th>
<th>$T_i$</th>
<th>$\tau_i$</th>
<th>$\hat{e}_i$</th>
<th>$\hat{e}_i^{(t)}$</th>
<th>$h_{ii}$</th>
<th>$D_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.39</td>
<td>41</td>
<td>1.89</td>
<td>1.99</td>
<td>0.04</td>
<td>0.06</td>
<td>0.11</td>
<td>0.19</td>
</tr>
<tr>
<td>2</td>
<td>0.37</td>
<td>56</td>
<td>0.87</td>
<td>0.94</td>
<td>-0.02</td>
<td>-0.03</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.39</td>
<td>63</td>
<td>0.36</td>
<td>0.56</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>0.43</td>
<td>68</td>
<td>-0.20</td>
<td>0.14</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>0.41</td>
<td>69</td>
<td>-0.53</td>
<td>-0.11</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>0.34</td>
<td>65</td>
<td>2.92</td>
<td>2.44</td>
<td>-0.08</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>0.33</td>
<td>61</td>
<td>0.95</td>
<td>1.01</td>
<td>-0.08</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>0.29</td>
<td>47</td>
<td>1.69</td>
<td>1.52</td>
<td>-0.08</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>0.27</td>
<td>32</td>
<td>-0.42</td>
<td>-0.02</td>
<td>-0.05</td>
<td>0.01</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>0.26</td>
<td>24</td>
<td>0.27</td>
<td>0.53</td>
<td>-0.04</td>
<td>-0.00</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>11</td>
<td>0.29</td>
<td>28</td>
<td>0.04</td>
<td>0.27</td>
<td>-0.02</td>
<td>-0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>12</td>
<td>0.30</td>
<td>26</td>
<td>-0.02</td>
<td>0.28</td>
<td>-0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>13</td>
<td>0.33</td>
<td>32</td>
<td>0.28</td>
<td>0.50</td>
<td>0.01</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>14</td>
<td>0.32</td>
<td>40</td>
<td>1.09</td>
<td>1.10</td>
<td>-0.03</td>
<td>-0.03</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>15</td>
<td>0.38</td>
<td>55</td>
<td>-0.32</td>
<td>0.07</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>16</td>
<td>0.38</td>
<td>63</td>
<td>2.28</td>
<td>1.96</td>
<td>-0.04</td>
<td>-0.04</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>17</td>
<td>0.47</td>
<td>72</td>
<td>1.65</td>
<td>1.59</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>18</td>
<td>0.44</td>
<td>72</td>
<td>-0.02</td>
<td>0.28</td>
<td>-0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>19</td>
<td>0.39</td>
<td>67</td>
<td>0.70</td>
<td>0.81</td>
<td>-0.04</td>
<td>-0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>20</td>
<td>0.34</td>
<td>60</td>
<td>2.55</td>
<td>2.19</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>21</td>
<td>0.32</td>
<td>44</td>
<td>-0.64</td>
<td>-0.17</td>
<td>-0.04</td>
<td>0.01</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>22</td>
<td>0.31</td>
<td>40</td>
<td>0.13</td>
<td>0.38</td>
<td>-0.04</td>
<td>-0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>23</td>
<td>0.28</td>
<td>32</td>
<td>1.56</td>
<td>1.46</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>24</td>
<td>0.33</td>
<td>27</td>
<td>1.16</td>
<td>1.20</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>25</td>
<td>0.31</td>
<td>28</td>
<td>-0.01</td>
<td>0.28</td>
<td>0.00</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>26</td>
<td>0.36</td>
<td>33</td>
<td>1.13</td>
<td>1.14</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>27</td>
<td>0.38</td>
<td>41</td>
<td>-0.22</td>
<td>0.13</td>
<td>0.03</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>0.42</td>
<td>52</td>
<td>0.59</td>
<td>0.72</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>29</td>
<td>0.44</td>
<td>64</td>
<td>-1.28</td>
<td>-0.66</td>
<td>0.02</td>
<td>-0.05</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>30</td>
<td>0.55</td>
<td>71</td>
<td>9.55</td>
<td>9.46</td>
<td>0.11</td>
<td>0.10</td>
<td>0.08</td>
<td>0.38</td>
</tr>
</tbody>
</table>
There are ten negative contributions in Table 2.3 so one third of the data generate negative contributions. Observation 29 appears to have a large negative contribution compared to the other negative values. Concerns about observation 29 are mentioned in Martin [40] due to this datum generating a large negative studentized residual, approximately -2. Thus a large negative contribution appears to indicate a departure from the model. The detection of observations which generate large negative contributions is considered in the next section.

2.6 Unusual Negative Contributions

Calculating $P(T_A \leq -t)$ where $T_A$ is the contribution of an arbitrary subset indexed by $A$ and $t$ is a positive real number can only be achieved by simulations or numerical approximation methods, see section 2.2. A method is proposed, in this section, of detecting unusually large negative contributions by obtaining upper and lower bounds for $P(T_A \leq -t)$. These bounds can be easily generated.

2.6.1 Upper Bound

The following proposition shows how an upper bound for $P(T_A \leq -t)$ can be trivially achieved.
Proposition 2.6.1 Let \( t \in \mathbb{R} \) such that \( t \geq 0 \) then
\[
P(T_A \leq -t) \leq 1 - P(x < \frac{2t\kappa_A}{\gamma_A - \phi_A})
\]
where \( x \sim \chi^2_{\kappa_A} \).

Proof By definition
\[
T_A = \frac{\gamma_A + \phi_A}{2\kappa_A} v_1 - \frac{\gamma_A - \phi_A}{2\kappa_A} v_2
\]
where \( v_1 \sim \chi^2_{\kappa_A} \) and \( v_2 \sim \chi^2_{\kappa_A} \). Also by definition \( \gamma_A \geq 0 \) and \( \phi_A \geq 0 \). It follows that if \( T_A < 0 \) then \( \gamma_A > \phi_A \). Now
\[
P(T_A \leq -t) = P\left(\frac{\gamma_A + \phi_A}{2\kappa_A} v_1 - \frac{\gamma_A - \phi_A}{2\kappa_A} v_2 \leq -t\right)
\]
\[
\leq P\left(-\frac{\left(\gamma_A - \phi_A\right)}{2\kappa_A} v_2 \leq -t\right)
\]
\[
= P(v_2 \geq \frac{2t\kappa_A}{\gamma_A - \phi_A})
\]
\[
= 1 - P(v_2 \leq \frac{2t\kappa_A}{\gamma_A - \phi_A}).
\]

2.6.2 Lower Bound

The following proposition gives a lower bound for \( P(T_A \leq -t) \). Unfortunately this lower bound is only valid for subsets of size one or two. Finding lower bounds for subsets greater than two is an area of current research.

Proposition 2.6.2 Let \( t \in \mathbb{R} \) such that \( t \geq 0 \) then
\[
P(T_A \leq -t) \geq 1 - P(x < \frac{2t\kappa_A}{\gamma_A - \phi_A} + \frac{\gamma_A + \phi_A}{\gamma_A - \phi_A} \kappa_A)
\]

24
where $x \sim \chi^2_{\kappa_A}$ and $\kappa_A = 1$ or $\kappa_A = 2$.

**Proof** Let $v_1 \sim \chi^2_{\kappa_A}$ and $v_2 \sim \chi^2_{\kappa_A}$ then by definition

$$P(T_A \leq -t) = P\left(-\gamma_A - \phi_A v_2 \leq -t - \frac{\gamma_A + \phi_A}{2\kappa_A} v_1\right)$$

$$= \int_{v_1} P\left(-\gamma_A - \phi_A v_2 \leq -t - \frac{\gamma_A + \phi_A}{2\kappa_A} v_1\right) f(v_1) dv_1$$

$$= E_{v_1}[P(-\gamma_A - \phi_A v_2 \leq -t - \frac{\gamma_A + \phi_A}{2\kappa_A} v_1)].$$

Let

$$g(v_1) = P\left(-\frac{\gamma_A - \phi_A}{2\kappa_A} v_2 \leq -t - \frac{\gamma_A + \phi_A}{2\kappa_A} v_1\right)$$

note $g(v_1)$ is a convex function when $\kappa_A = 1$ or $\kappa_A = 2$, see appendix A.1.1.

Hence

$$P(T_A \leq -t) = E_{v_1}[g(v_1)]$$

$$\geq g(E_{v_1}[v_1]) \quad \text{by Jensen's Inequality, see Mood et al. [43]}$$

$$= P\left(-\frac{\gamma_A - \phi_A}{2\kappa_A} v_2 \leq -t - \frac{\gamma_A + \phi_A}{2\kappa_A} E_{v_1}[v_1]\right)$$

$$= P(v_2 \geq \frac{2t\kappa_A - \gamma_A - \phi_A}{\gamma_A - \phi_A} + \frac{\gamma_A + \phi_A}{\gamma_A - \phi_A} - \kappa_A)$$

$$= 1 - P(v_2 < \frac{2t\kappa_A - \gamma_A - \phi_A}{\gamma_A - \phi_A} + \frac{\gamma_A + \phi_A}{\gamma_A - \phi_A} - \kappa_A).$$

## 2.6.3 Application of Bounds

These bounds are used to examine the negative contributions generated by the Ice-Cream data set. From Table 2.4 it can be seen that as the negative
Table 2.4: Negative Contributions of Ice-cream data

<table>
<thead>
<tr>
<th>Index</th>
<th>$T_i$</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>-0.20</td>
<td>0.04</td>
<td>0.45</td>
</tr>
<tr>
<td>5</td>
<td>-0.53</td>
<td>0.02</td>
<td>0.21</td>
</tr>
<tr>
<td>9</td>
<td>-0.42</td>
<td>0.03</td>
<td>0.27</td>
</tr>
<tr>
<td>12</td>
<td>-0.02</td>
<td>0.05</td>
<td>0.82</td>
</tr>
<tr>
<td>15</td>
<td>-0.32</td>
<td>0.03</td>
<td>0.35</td>
</tr>
<tr>
<td>18</td>
<td>-0.02</td>
<td>0.05</td>
<td>0.83</td>
</tr>
<tr>
<td>21</td>
<td>-0.64</td>
<td>0.02</td>
<td>0.18</td>
</tr>
<tr>
<td>25</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.84</td>
</tr>
<tr>
<td>27</td>
<td>-0.22</td>
<td>0.04</td>
<td>0.43</td>
</tr>
<tr>
<td>29</td>
<td>-1.28</td>
<td>0.01</td>
<td>0.06</td>
</tr>
</tbody>
</table>

values increase in magnitude the difference between the bounds decreases.

In fact $\forall \kappa_A \in \{1, 2, 3, \ldots\}$

\[
\lim_{t \to \infty} \left| P\left(t < \frac{2t\kappa_A}{\gamma_A - \phi_A} \right) - P\left(t < \frac{2t\kappa_A}{\gamma_A - \phi_A} \right) \right| = 0.
\]

The probability of observing a negative contribution of larger magnitude than observation 29 is between 0.01 and 0.06. This range indicates that although the contribution of observation 29 is not exceptionally unusual there is still some evidence that observation 29 deviates from the model.
When the Ice-Cream data is re-analyzed with the outlying observation 30 omitted then observation 29 generates a positive contribution close to zero.

In Martin [40] when the analysis is redone with observation 30 removed then observation 29 generates a positive studentized residual close to zero. This shows that contributions can suffer from smearing, see Bruce and Martin [8], where smearing is used to define non-outlying observations being detected as outlying due to the influence of outlying points elsewhere.

2.7 General Linear Models to Time Series

Each unusual observation in the three data sets was detected by the standardized contributions. These examples provide further empirical evidence of the ability of this approach at detecting unusual observations in simple and general linear models.

The success of this method to detect outlying points in the presence of high serial correlation opens the way forward to application in the field of time series analysis. It is now possible to detect unusually large negative contributions. The importance of this development with respect to time series analysis is explored in chapter 4. It has also been established that contributions and standardized contributions suffer from smearing so care must be taken to avoid the false detection of outliers.
Chapter 3

Approximating Standardized Contributions

3.1 Computational Efficiency

The additive property of the contributions, \( T_A = \sum_{i \in A} T_i \), makes them more desirable to work with from a computational viewpoint than their non-additive standardized versions. Thus for computational efficiency it would be very useful to know when \( T_A \) is a good approximation for \( \tau_A \). The relationship between \( T_A \) and \( \tau_A \) is determined by the parameters \( \phi_A \) and \( \gamma_A \). It is these parameters which indicate the models for which \( T_A \) is a good approximation for \( \tau_A \) and for which it is not. When \( T_A \) is not a good approximation for \( \tau_A \) an understanding of the relationship between \( \phi_A \) and \( \gamma_A \) leads to a
computationally cheap approximation.

Except for models where observations are assumed to be independent, the parameter $\gamma_A$ is computationally more difficult to obtain than the parameter $\phi_A$. This computational difference arises from the fact that $\phi_A = \sum_{i \in A} \phi_i$; thus $\phi_A$ can be obtained for arbitrary subsets by adding each $\phi_i$ corresponding to the observations in the set $A$. The parameter $\gamma_A$ does not share this additive property except when observations are assumed to be independent; without this assumption there appears to be no simple relationship between $\gamma_A$ and $\{\gamma_i | i \in A\}$. However this chapter provides evidence that in many situations $\phi_A$ is a good approximation for $\gamma_A$. This approximation enables the p value associated with $T_A$ to be used as a bound for the p value of $\tau_A$.

The following sections in this chapter focus on the parameters $\phi_A$ and $\gamma_A$. Theoretical exploration, along with computations of these parameters, gives insight into the relationship between $T_A$ and $\tau_A$. The result of this work is computationally cheap approximations and bounds for $\tau_A$ based on $T_A$.

### 3.2 Influence and Ranges

For observations where homoscedasticity is assumed $\phi_A$ and $\gamma_A$ are independent of the variance of the observations. It is the correlation structure of the model and the leverages associated with the observations which determine
the values taken by \( \phi_A \) and \( \gamma_A \). This can easily be seen in the case of the parameter \( \phi_A \); by definition

\[
\phi_A = \text{tr}((VQ)_{AA}) = \text{tr}((I - X(X^{-1}V^{-1}X)^{-1}X^TV^{-1})_{AA})
\]

\[
= \text{tr}((I - X(X^{-1}R^{-1}X)^{-1}X^TR^{-1})_{AA})
\]

since \( V = \sigma^2 R \), where \( R \) is a correlation matrix. It can be shown in a similar manner that \( \gamma_A \) is independent of the variance of observations. If the variance of the data is not constant then the parameters \( \phi_A \) and \( \gamma_A \) will both be influenced by the variance of the observations.

The ranges of the parameters \( \phi_A \) and \( \gamma_A \) will now be defined. From proposition 3.2.1 it is established that, for any general linear model, the parameter \( \phi_A \) has a limited range from zero to \( \kappa_A \). Proposition 3.2.2 proves that \( \gamma_A \geq 0 \). The example after proposition 3.2.2 provides evidence that any upper bound of \( \gamma_A \) would depend on the correlation structure of the observations. From the example presented in this section a correlation structure can be constructed such that \( \gamma_A \) is exceptionally large. As a result \( T_A \) has the potential to be very different from \( \tau_A \).

**Proposition 3.2.1** The parameter \( \phi_i \) has the following range

\[
0 \leq \phi_i \leq 1.
\]
In the general case \( \phi_A \) has the following range

\[
0 \leq \phi_A \leq \kappa_A
\]

where \( \kappa_A \) is the size of the subset \( A \).

**Proof** By definition

\[
\phi_i = (VQ)_{ii}
\]

where

\[
VQ = I - X(X^T V^{-1} X)^{-1} X^T V^{-1} = I - H.
\]

Both \( VQ \) and \( H \) are idempotent matrices. Idempotent matrices have eigenvalues equal to one or zero, see Graybill [26] section 12.3. Since the eigenvalues of \( VQ \) and \( H \) are either one or zero it follows that both \( VQ \) and \( H \) are non-negative definite matrices, see Graybill [26] section 12.2. Since \( H \) is non-negative definite then \( h_{ii} \geq 0 \). But \( \phi_i = 1 - h_{ii} \geq 0 \) hence \( 0 \leq h_{ii} \leq 1 \). Thus \( 0 \leq \phi_i \leq 1 \). It immediately follows that \( 0 \leq \phi_A \leq \kappa_A \).

**Proposition 3.2.2** The parameter \( \gamma_A \) has the following range

\[
\gamma_A \geq 0
\]

provided \( D_A \) is positive definite.

**Proof** By definition

\[
\gamma_A = tr(F^{\frac{1}{2}})
\]
where \( F = G_{AA}^{\frac{1}{2}} D_A^{-1} G_{AA}^{\frac{1}{2}} \). By definition \( D_A \) is a symmetric positive definite matrix thus \( D_A^{-1} \) is positive definite matrix. It follows that \( G_{AA}^{\frac{1}{2}} D_A^{-1} G_{AA}^{\frac{1}{2}} \) is a non-negative definite matrix. It immediately follows that

\[ \gamma_A \geq 0. \]

**Example** Consider a set of \( n \) observations such that \( Y \sim N(0, V) \) where the covariance matrix is defined as follows

\[
V = \frac{\sigma^2}{1 - \varphi^2} \begin{pmatrix}
1 & \varphi & \varphi^2 & \ldots & \varphi^{n-1} \\
\varphi & 1 & \varphi & \ldots & \varphi^{n-2} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\varphi^{n-1} & \ldots & \varphi^2 & \varphi & 1
\end{pmatrix}
\]

where \( |\varphi| < 1 \). This type of model is used to describe an autoregressive process of order one, \( AR(1) \) see Box and Jenkins [7]. This process is stationary provided \( |\varphi| < 1 \), see Harvey [27], which implies that the variance of observations are constant. This model has a known mean hence no fixed effects have to be estimated. For this model the marginal residual associated with observation \( i \) is denoted \( e_i \) and the corresponding conditional residual is represented by \( e_{i(i)} \). Recall the notation \( \hat{e}_i \) and \( \bar{e}_{i(i)} \), defined in the previous chapter, is used to indicate the use of fixed effects estimates in calculating marginal and conditional residuals. The variance of any marginal residual associated with this model is given by \( \text{var}(e_i) = \frac{\sigma^2}{1 - \varphi^2} \) where \( i \in \{1, \ldots, n\} \).
The variances of the corresponding conditional residuals are given by

\[ \text{var}(e_i) = \begin{cases} \sigma^2, & \text{if } i = 1 \text{ or } n \\ \frac{\sigma^2}{1+\varphi^2}, & \text{otherwise.} \end{cases} \]

Haslett and Hayes [29] show that \( \gamma_i = \frac{\text{var}(e_i)}{\text{var}(e_{(1)})} \) thus for this type of model

\[ \gamma_i = \begin{cases} \frac{1}{\sqrt{1-\varphi^2}}, & \text{if } i = 1 \text{ or } n \\ \frac{\sqrt{1+\varphi^2}}{1-\varphi^2}, & \text{otherwise.} \end{cases} \]

Clearly the value \( \gamma_i \) takes depends on the correlation parameter \( \varphi \); as \( |\varphi| \to 1 \) then \( \gamma_i \to \infty \).

### 3.3 Subsets

This section is concerned with understanding the relationship between \( \phi_A \) and \( \phi_B \) when \( A \) is a subset of \( B \), \( A \subseteq B \). The relationship between \( \gamma_A \) and \( \gamma_B \) is also examined. First the values of \( \phi_* \) and \( \gamma_* \) are established where \( * \) represents a set which indexes the entire data.

**Proposition 3.3.1** Given a set of \( n \) observations such that \( Y \sim N(X\beta, V) \) and \( p \) is the rank of the design matrix \( X \) then

\[ \phi_* = n - p. \]

**Proof** Since \( * \) indexes the entire data set then

\[ \phi_* = \text{tr}(VQ) \]

33
Proposition 3.3.2 Given a set of $n$ observations such that $Y \sim N(X\beta, V)$ and $p$ is the rank of the design matrix $X$ then

$$\gamma_* = n - p.$$  

Proof By definition

$$\gamma_* = tr(F^{\frac{1}{2}})$$

where $F = G^{\frac{1}{2}}D^{-1}G^{\frac{1}{2}}$. When the entire data set is indexed the resulting $F$ matrix is idempotent, see appendix A.1.2, thus

$$\gamma_* = tr(F).$$

From the definition of $F$ then

$$\gamma_* = tr(G^{\frac{1}{2}}D^{-1}G^{\frac{1}{2}})$$

$$= tr(G^{\frac{1}{2}}QG^{\frac{1}{2}})$$

$$= tr(GQ)$$

$$= tr(VQVQ)$$
\[ = \text{tr}(VQ) \]
\[ = n - p. \]

**Proposition 3.3.3** Provided that \( A \) is a subset of \( B \), \( A \subseteq B \), then

\[ \phi_B \geq \phi_A. \]

**Proof** The set \( B \) can be expressed as the union of two mutually exclusive subsets

\[ B = (B \cap A) \cup (B \cap \bar{A}) \]

where \( \bar{A} \) is the complement of \( A \). However \( A \subseteq B \) hence

\[ B = A \cup (B \cap \bar{A}). \]

It follows that

\[ \phi_B = \phi_A + \sum_{i \in (B \cap A)} \phi_i. \]

From proposition 3.1.1 \( \phi_i \geq 0 \) thus

\[ \phi_B \geq \phi_A. \]

From proposition 3.3.1 and proposition 3.3.3 it can be concluded for any arbitrary subset \( A \) that \( \phi_A \) is bound above by \( n - p \). Also given a sequence of subsets such that \( A_1 \subseteq A_2 \subseteq \ldots \subseteq \cdot \) then \( \phi_{A_1} \leq \phi_{A_2} \leq \ldots \leq \phi_\ast \). The parameters \( \gamma_A \) and \( \gamma_B \) do not share the same relationship as \( \phi_A \) and \( \phi_B \). This is shown by counter-example. Returning to the example in the previous
section it was established for that model \( \gamma_i \to \infty \) as \( |\varphi| \to 1 \). This model has no fixed effects. Applying proposition 3.3.2 then \( \gamma_* = n \). Thus \( \varphi \) can be sufficiently large to ensure \( \gamma_i > \gamma_* \). For example assume a data set with a hundred observations, \( n = 100 \) and \( \varphi = .99999 \) then \( \gamma_1 = 223.6 \) thus \( \gamma_1 > \gamma_* \).

However if \( \varphi = .9 \) then \( \gamma_1 = 2.29 \) hence \( \gamma_* > \gamma_1 \). Unlike the monotone increasing relationship between \( \phi_A \) and \( \phi_B \), \( \phi_B \geq \phi_A \) when \( B \supseteq A \), the relationship between \( \gamma_A \) and \( \gamma_B \) depends on the correlation of observations.

### 3.4 Fixed effects are known

For models where the fixed effects are assumed to be known proposition 3.4.1 establishes that \( \phi_A = \kappa_A \) and that \( \gamma_A \geq \phi_A \). Proposition 3.4.2 uses these properties to establish for such models that \( T_A \) can be used as a bound for the p value of \( \tau_A \) when \( A \) indexes a set which \( \tau_A \) detects as outlying.

**Proposition 3.4.1** Given a set of data \( Y \sim N(X\beta, V) \) such that the fixed effects, \( \beta \), are known then

\[
\phi_A = \kappa_A \text{ and } \gamma_A \geq \phi_A.
\]

**Proof** Since \( \beta \) is known then \( Q = V^{-1} \) hence

\[
\phi_A = \text{tr}((VQ)_{AA}) = \text{tr}(I_A)
\]

Hence \( \gamma_A = \text{tr}(I_A) \geq \phi_A \).
Proposition 3.3.1 and proposition 3.3.2 established that if $A$ indexes the entire data set, $A = \bullet$ then $\gamma_\bullet = \phi_\bullet$. Arbitrary subsets are now considered which do not index the entire data set. By definition

$$\gamma_A = \text{tr}(F^{1/2})$$

where $F = G_A^{1/2}D_A^{-1}G_A^{1/2}$. Since $\beta$ is known $G_A = (VQV)_{AA} = V_{AA}$ and $D_A^{-1} = Q_{AA} = (V^{-1})_{AA}$. It is possible to express $(V^{-1})_{AA}$ as follows

$$(V^{-1})_{AA} = V_{AA}^{-1} + V_{AA}^{-1}V_{AA}V_{AA}^{-1}V_{AA}V_{AA}^{-1}.$$ 

Hence

$$F = V_{AA}^{1/2}V_{AA}^{-1}V_{AA}^{1/2} + V_{AA}^{1/2}V_{AA}^{-1}V_{AA}^{1/2}V_{AA}^{-1}V_{AA}V_{AA}^{-1}V_{AA}^{1/2}.$$ 

$$= I_A + V_{AA}^{-1/2}V_{AA}V_{AA}^{-1}V_{AA}V_{AA}^{-1/2}$$

where $C = V_{AA}^{-1/2}V_{AA}V_{AA}^{-1}V_{AA}V_{AA}^{-1}$ is a non-negative definitive matrix. Thus all the eigenvalues, $\lambda$, of $C$ are non-negative. Note

$$F = I_A + C$$ 

$$= I_A + B\Lambda B^T$$ where $B$ comprises the eigenvectors of $C$

$$= B(I_A + \Lambda)B^T.$$ 

It follows that the eigenvalues of $F$ are of the form $1 + \lambda_i$ where $\lambda_i \geq 0$. Hence $\gamma_A = \text{tr}(F^{1/2}) = \sum_{i\in A} \sqrt{1 + \lambda_i}$. Thus $\gamma_A \geq \phi_A$.  

37
Proposition 3.4.2 Given a set of data \( Y \sim N(X\beta, V) \) such that the fixed effects, \( \beta \), are known and let \( A \) index a set such that \( P(x \geq \tau_A) < 0.05 \) where 
\[ x \sim \chi^2_{\kappa_A} \] then 
\[ T_A \geq \tau_A. \]

Proof If \( P(x \geq \tau_A) < 0.05 \) then \( \tau_A > \kappa_A \) since \( \kappa_A \) is the mean of \( \chi^2_{\kappa_A} \). By definition 
\[ \tau_A = \kappa_A \left\{ \frac{(T_A - \phi_A)\sqrt{2}}{\phi_A^2 + \gamma_A^2} + 1 \right\} \]
thus in order for \( \tau_A > \kappa_A \) then \( T_A - \phi_A > 0 \). Since \( T_A - \phi_A > 0 \) and from proposition 3.4.1 \( \gamma_A \geq \phi_A \) then 
\[ \tau_A \leq \kappa_A \left\{ \frac{(T_A - \phi_A)\sqrt{2}}{\phi_A^2 + \gamma_A^2} + 1 \right\} \]
\[ = \kappa_A \left\{ \frac{(T_A - \kappa_A)\sqrt{2}}{\kappa_A^2 + \gamma_A^2} + 1 \right\} \quad \text{since } \phi_A = \kappa_A \text{ from proposition 3.4.1} \]
\[ = T_A. \]

Although it has been shown in proposition 3.4.1 that \( \gamma_A \geq \phi_A \) it remains unclear how much greater. This difference would indicate how good an estimate \( T_A \) is for \( \tau_A \). This question was explored by computation. Obviously it is not possible to examine every conceivable collection of subsets even for relatively small data sets so attention is focused on contiguous blocks of observations. Three \( AR(1) \) models where \( \varphi \) took the value 0.3, 0.6 and 0.9 were examined. It was assumed that the mean of each process was known
Figure 3.1: $\phi_A$ and $\gamma_A$ associated with $AR(1)$ processes

and equal to zero. Figure 3.1 is a plot of $\phi_A$ and $\gamma_A$, corresponding to the $AR(1)$ processes, against the subset $A$ where if $A = 10$ then this corresponds to a contiguous block of observations stretching from the first observation to the tenth observation. Figure 3.1 contains four curves where three of these curves correspond to $\gamma_A$ when $\varphi = 0.3$, $\varphi = 0.6$ and $\varphi = 0.9$. The fourth curve in Figure 3.1 is of $\phi_A = \kappa_A$. It is clear from Figure 3.1 that $\phi_A \approx \gamma_A$ over a considerable range of $\varphi$ and subset sizes. This suggests that $\phi_A \approx \gamma_A$ for models where observations are assumed not to be extremely correlated.

An immediate consequence of these calculations is that $T_A$ will be a very
good approximation of $\tau_A$ for models where the fixed effects are known. It also provides evidence that for such models the difference between $T_A$ and $\tau_A$ is not very sensitive to changes in the correlation parameters.

### 3.5 Fixed effects are unknown

In chapter 2 the relationship between $T_i$ and $\tau_i$ was explored and it was shown that if leverage is an issue then $T_i$ can be quite different from $\tau_i$. The data sets used to illustrate this relationship are revisited in this section. These sets of data will be used to explore the relationship between $T_A$ and $\tau_A$ where $A$ is a subset of observations. This exploration starts by examining the values that $\phi_A$ and $\gamma_A$ take when fixed effects are unknown and must be estimated.

#### 3.5.1 Independent Observations

Proposition 3.5.1 shows that when the observations are independent and identically distributed then $\gamma_A = \phi_A$. In this situation, these parameters can be expressed in terms of leverages.

**Proposition 3.5.1** Given a set of data $Y \sim N(X\beta, \sigma^2 I)$ then

$$\gamma_A = \phi_A.$$
Proof By definition

\[ VQ = I - X(X^TV^{-1}X)^{-1}X^TV^{-1} \]

\[ = I - X(X^TX)^{-1}X^T \quad \text{since } V = \sigma^2 I. \]

By definition

\[ \phi_A = tr((VQ)_{AA}) \]

\[ = \sum_{i \in A} (1 - h_{ii}) \]

\[ = \kappa_A - \sum_{i \in A} h_{ii} \]

where \( H = X(X^TX)^{-1}X^T \) and \( h_{ii} \) is the \( i^{th} \) diagonal element of \( H \). The term \( h_{ii} \) is known as the leverage of the \( i^{th} \) observation.

By definition

\[ \gamma_A = tr(F^\frac{1}{2}) \]

where \( F = G_{AA}^{\frac{1}{2}}D_A^{-1}G_{AA}^{\frac{1}{2}} \). But \( G = VQV = \sigma^2 Q \) hence \( G_{AA} = \sigma^2 Q_{AA} \). Since \( D_A^{-1} = Q_{AA} \) it follows that \( F = \sigma^2 Q_{AA} \). Thus

\[ \gamma_A = tr(\sigma^2 Q_{AA}) \]

\[ = \sum_{i \in A} (1 - h_{ii}) \]

\[ = \kappa_A - \sum_{i \in A} h_{ii} \]

The proof immediately follows.
The first data set encountered in chapter 2 is Forbes' data and this data set contains one unusual observation indexed by the number 12. There are no points with a high leverage. From Table 3.1 it can be seen that \( \gamma_i = \phi_i \approx \frac{n-p}{n} = .88 \) where \( n = 17 \) is the number of observations and \( p = 2 \) is the number of fixed effects estimated. The values of \( \phi_A \) where \( A = \{1\} \), \( A = \{1, 2\} \), \( A = \{1, 2, 3\} \), etc are presented in Table 3.1. It is clear that \( \phi_A \approx \frac{\kappa_A(n-p)}{n} = .88 \ast \kappa_A \). Although \( T_A \) does not increment up as smoothly as \( \phi_A \) it is clear that \( T_A \) is a good estimate of \( \tau_A \).

The second data set analyzed in chapter 2 is Huber's data and this data set contains one unusual observation indexed by the number 6. This unusual point has a very high leverage. From Table 3.2 it is not always the case that \( \phi_i \) is approximated by \( \frac{n-p}{n} \) and subsequently \( \phi_A \) can be quite different from \( \kappa_A \ast \frac{n-p}{n} \). The contribution associated with observation six is considerably different from its standardized version. The non-standardized contribution is not sufficient to detect this unusual observation.

It is deduced from proposition 3.4.1 and both these examples that \( T_A \) is not a suitable estimate for \( \tau_A \) when leverage is a concern. This is due to the fact that \( \phi_A \) and subsequently \( \gamma_A \) are considerably different from \( \kappa_A \). As a result when these parameters are used to standardize \( T_A \) they cause \( \tau_A \) to be dramatically different. For independent observations calculating \( \tau_A \) is not very computationally intensive since \( \gamma_A = \phi_A \) and \( \phi_A \) is relatively easy to
Table 3.1: $\phi_A$ and $\gamma_A$ associated with Forbes' Data

<table>
<thead>
<tr>
<th>Index</th>
<th>$\gamma_i = \phi_i$</th>
<th>$T_i$</th>
<th>$\tau_i$</th>
<th>$\gamma_A = \phi_A$</th>
<th>$T_A$</th>
<th>$\tau_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.81</td>
<td>0.42</td>
<td>0.53</td>
<td>0.81</td>
<td>0.42</td>
<td>0.53</td>
</tr>
<tr>
<td>2</td>
<td>0.80</td>
<td>0.03</td>
<td>0.04</td>
<td>1.61</td>
<td>0.46</td>
<td>0.57</td>
</tr>
<tr>
<td>3</td>
<td>0.89</td>
<td>0.03</td>
<td>0.03</td>
<td>2.51</td>
<td>0.48</td>
<td>0.58</td>
</tr>
<tr>
<td>4</td>
<td>0.90</td>
<td>0.00</td>
<td>0.00</td>
<td>3.41</td>
<td>0.48</td>
<td>0.57</td>
</tr>
<tr>
<td>5</td>
<td>0.92</td>
<td>0.01</td>
<td>0.01</td>
<td>4.32</td>
<td>0.49</td>
<td>0.57</td>
</tr>
<tr>
<td>6</td>
<td>0.92</td>
<td>0.01</td>
<td>0.01</td>
<td>5.24</td>
<td>0.51</td>
<td>0.58</td>
</tr>
<tr>
<td>7</td>
<td>0.93</td>
<td>0.02</td>
<td>0.02</td>
<td>6.17</td>
<td>0.52</td>
<td>0.59</td>
</tr>
<tr>
<td>8</td>
<td>0.93</td>
<td>0.02</td>
<td>0.02</td>
<td>7.11</td>
<td>0.54</td>
<td>0.61</td>
</tr>
<tr>
<td>9</td>
<td>0.94</td>
<td>0.17</td>
<td>0.18</td>
<td>8.04</td>
<td>0.71</td>
<td>0.80</td>
</tr>
<tr>
<td>10</td>
<td>0.94</td>
<td>0.04</td>
<td>0.04</td>
<td>8.98</td>
<td>0.75</td>
<td>0.84</td>
</tr>
<tr>
<td>11</td>
<td>0.94</td>
<td>0.15</td>
<td>0.16</td>
<td>9.92</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>12</td>
<td>0.94</td>
<td>12.87</td>
<td>13.75</td>
<td>10.86</td>
<td>13.77</td>
<td>15.22</td>
</tr>
<tr>
<td>13</td>
<td>0.86</td>
<td>0.00</td>
<td>0.00</td>
<td>11.72</td>
<td>13.77</td>
<td>15.27</td>
</tr>
<tr>
<td>14</td>
<td>0.88</td>
<td>0.73</td>
<td>0.82</td>
<td>12.64</td>
<td>14.49</td>
<td>16.10</td>
</tr>
<tr>
<td>15</td>
<td>0.83</td>
<td>0.41</td>
<td>0.50</td>
<td>13.43</td>
<td>14.91</td>
<td>16.65</td>
</tr>
<tr>
<td>16</td>
<td>0.79</td>
<td>0.04</td>
<td>0.05</td>
<td>14.22</td>
<td>14.95</td>
<td>16.82</td>
</tr>
<tr>
<td>17</td>
<td>0.78</td>
<td>0.05</td>
<td>0.07</td>
<td>15.00</td>
<td>15.00</td>
<td>17.00</td>
</tr>
</tbody>
</table>

43
Table 3.2: $\phi_A$ and $\gamma_A$ associated with Huber’s Data

<table>
<thead>
<tr>
<th>Index</th>
<th>$\gamma_i = \phi_i$</th>
<th>$T_i$</th>
<th>$\tau_i$</th>
<th>$\gamma_A = \phi_A$</th>
<th>$T_A$</th>
<th>$\tau_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.71</td>
<td>1.81</td>
<td>2.55</td>
<td>0.71</td>
<td>1.81</td>
<td>2.55</td>
</tr>
<tr>
<td>2</td>
<td>0.76</td>
<td>0.07</td>
<td>0.09</td>
<td>1.47</td>
<td>1.88</td>
<td>2.56</td>
</tr>
<tr>
<td>3</td>
<td>0.8</td>
<td>0.03</td>
<td>0.04</td>
<td>2.27</td>
<td>1.91</td>
<td>2.53</td>
</tr>
<tr>
<td>4</td>
<td>0.83</td>
<td>1.05</td>
<td>1.27</td>
<td>3.1</td>
<td>2.96</td>
<td>3.82</td>
</tr>
<tr>
<td>5</td>
<td>0.83</td>
<td>0.8</td>
<td>0.96</td>
<td>3.93</td>
<td>3.76</td>
<td>4.78</td>
</tr>
<tr>
<td>6</td>
<td>0.07</td>
<td>0.23</td>
<td>3.62</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
</tbody>
</table>

obtain. However when the observations are correlated obtaining $\tau_A$ increases in complexity. The next section proposes a cheap approximation for $\tau_A$ when observations are correlated and the model has unknown fixed effects. This approximation is valid even when leverage may be an issue.

### 3.5.2 General Covariance Structure

In this section general linear models are considered in which the fixed effects are unknown and a non-independent correlation structure is assumed. Proposition 3.5.2 establishes that $\gamma_i \geq \phi_i$ where $i$ indexes a single observation. It is conjectured that $\gamma_A \geq \phi_A$ for arbitrary subsets, but whether or not this is true remains an area of research for such models. Computations are
presented which add credibility to this conjecture. A computationally cheap approximation of \( \tau_A \) is proposed as a result of this conjecture.

**Proposition 3.5.2** Given a set of observations \( Y \sim N(X\beta,V) \) then

\[
\gamma_i \geq \phi_i
\]

where \( i \) indexes a single observation of \( Y \).

**Proof** Let \( M_i \) be the covariance matrix of the vector \( (\hat{e}_i, \hat{e}_{(i)}) \) which is defined as follows

\[
M_i = \begin{pmatrix}
    g_{ii} & d_i \phi_i \\
    d_i \phi_i & d_i
\end{pmatrix}
\]

where \( g_{ii} = \text{var}(\hat{e}_i) \) and \( d_i = \text{var}(\hat{e}_{(i)}) \). The \( \text{cov}(\hat{e}_i, \hat{e}_{(i)}) = d_i \phi_i \), a general proof of this covariance is found in Appendix B.1.1. \( M_i \) is a non-negative definite matrix thus its determinant must be greater than or equal to zero, \( \det M_i \geq 0 \).

\[
\det M_i = g_{ii}d_i - d_i^2 \phi_i^2 = g_{ii}d_i(1 - \frac{d_i \phi_i^2}{g_{ii}})
\]

By definition \( g_{ii} = \text{var}(\hat{e}_i) \geq 0 \) and \( d_i = \text{var}(\hat{e}_{(i)}) \geq 0 \) thus in order for \( \det M_i \geq 0 \) then \( 1 - \frac{d_i \phi_i^2}{g_{ii}} \geq 0 \). But

\[
1 - \frac{d_i \phi_i^2}{g_{ii}} = 1 - \frac{\phi_i^2}{\gamma_i^2}
\]

45
hence

\[ 1 \geq \frac{\phi_i^2}{\gamma_i^2} \]

\[ \gamma_i^2 \geq \phi_i^2. \]

From proposition 3.1.1 and 3.1.2 \( \gamma_i \geq 0 \) and \( \phi_i \geq 0 \) thus

\[ \gamma_i \geq \phi_i. \]

The third data set presented in chapter 2 is referred to as the Ice-cream data set. The model for this data assumes two unknown fixed effects and that its observations are correlated; this correlation structure is described by an autoregressive process of order one where the correlation parameter \( \varphi = 0.75 \). This data set is used to explore the relationship between \( \phi_A \) and \( \gamma_A \) for models which have unknown fixed effects and general correlation structures. This exploration in turn gives insight into the relationship between \( T_A \) and \( \tau_A \).

Attention is focused on contiguous blocks of observations starting from the first observation. Figure 3.2 contains plots of both \( \phi_A \) and \( \gamma_A \) against subset sizes. The x-axis in Figure 3.2 is labeled subset size and can be interpreted as follows: if subset size takes the value 10 then this corresponds to \( A = \{1, \ldots, 10\} \). From Figure 3.2 it is clear that \( \gamma_A \approx \phi_A \), \( \gamma_A \geq \phi_A \) and that \( \kappa_A \) is a good approximation for both these parameters.

Since \( \gamma_A \approx \phi_A \) and \( \phi_A \approx \kappa_A \) then \( T_A \) will be a good approximation of \( \tau_A \). Figure 3.3 compares \( T_A \) and \( \tau_A \), it is clear that for such subsets \( T_A \) is a
Figure 3.2: $\phi_A$ and $\gamma_A$ associated with Ice-cream Data

![Graph showing $\phi_A$ and $\gamma_A$ vs. Subset Size]

$\gamma_A = \frac{\phi_A \cdot A}{r_A}$

Figure 3.3: $T_A$ and $\tau_A$ associated with Ice-cream Data

![Graph showing $T_A$ and $\tau_A$ vs. Subset Size]
good approximation of $\tau_A$. An alternative approximation of $\tau_A$ is proposed in the next section which is computationally efficient and empirical evidence is presented of its superiority over $T_A$.

### 3.6 A Cheap Approximation of $\tau_A$

Previously, theory and examples have shown in order for $T_A$ to be a good approximation for $\tau_A$ then two conditions must hold: $\phi_A \approx \gamma_A$ and $\phi_A \approx \kappa_A$. An alternative statistic $\tau_A^*$ is proposed which requires only that $\phi_A \approx \gamma_A$.

This new statistic is defined as follows

$$
\tau_A^* = \frac{\kappa_A}{\phi_A} T_A.
$$

The statistic $\tau_A^*$ does not share the same additive property as $T_A$,

$$
\tau_A^* \neq \sum_{i \in A} \tau_i^*.
$$

However $\tau_A^*$ relies only on parameters which can easily be obtained for arbitrary subsets since

$$
T_A = \sum_{i \in A} T_i \text{ and } \phi_A = \sum_{i \in A} \phi_i
$$

as a result $\tau_A^*$ is computationally cheap to obtain. When $Y \sim N(X\beta, \sigma^2I)$ then proposition 3.5.1 states that $\phi_A = \gamma_A$ hence $\tau_A = \tau_A^*$ since

$$
\tau_A = \kappa_A \left\{ \frac{T_A - \phi_A}{\phi_A} + 1 \right\} = \frac{\kappa_A}{\phi_A} T_A.
$$
This suggests that $\tau_A^*$ deals better with observations with high leverage than $T_A$. The following proposition shows that $\tau_i^*$ can be used as a bound for $\tau_i$; this has yet to be generalized for arbitrary subsets.

**Proposition 3.6.1** Given a set of data $Y \sim N(X\beta, V)$ and let $i$ index an observation such that $P(x \geq \tau_i) < 0.05$ where $x \sim \chi^2_1$ then

$$\tau_i^* \geq \tau_i$$

where $\tau_i^* = \frac{T_i}{\phi_i}$.

**Proof** If $P(x \geq \tau_i) < 0.05$ then $\tau_i > 1$ since 1 is the mean of $\chi^2_1$. By definition

$$\tau_i = \frac{(T_i - \phi_i)\sqrt{2}}{\sqrt{\phi_i^2 + \gamma_i^2}} + 1$$

thus in order for $\tau_i > 1$ then $T_i - \phi_i > 0$. Since $T_i - \phi_i > 0$ and from proposition 3.5.2 $\gamma_i \geq \phi_i$ then

$$\tau_i \leq \frac{(T_i - \phi_i)\sqrt{2}}{\sqrt{\phi_i^2 + \phi_i^2}} + 1$$

$$= \frac{T_i - \phi_i}{\phi_i} + 1$$

$$= \frac{T_i}{\phi_i}$$

$$= \tau_i^*$$.  

Although for subsets of data it has not been theoretically confirmed that $\gamma_A \geq \phi_A$ computations suggest this to be true; recall Figure 3.2. It is also clear
from Figure 3.2 that $\phi_A$ is a good approximation for $\gamma_A$ and subsequently $\tau_A^*$ is a good approximation for $\tau_A$. Figure 3.4 compares $\tau_A$ and $\tau_A^*$, certainly for this example $\tau_A^*$ is considerably better than $T_A$ at estimating $\tau_A$. $\tau_A^*$ is such a good approximation for $\tau_A$ in this example that in Figure 3.4 the dashed line representing $\tau_A^*$ lies almost perfectly on top of the line representing $\tau_A$.

3.7 Conclusions

There is plenty of evidence from both a theoretical and computational aspect that for many models $T_A$ is a good approximation for $\tau_A$. In fact proposition 3.4.2 shows for a particular class of models that $T_A$ can be used to find a lower
bound for the p values associated with $\tau_A$. Thus the additive property of $T_A$
can be utilized which allows computational cheap diagnostics to be performed
on arbitrary subsets of observations. It is also clear when data points have
a high leverage $T_A$ can be quite different from $\tau_A$; in such situations it is
necessary to calculate $\tau_A$ or use the computationally cheap approximation
$\tau^*_A$.

4.1 Contributions and Time Series

This chapter intends to assess the ability of the methods in
Hasselt and Hameleers (2006) and the new developments in detection
complementary combination of time series which are hypo-
test statistics inside deletion of data points or that are used for
analysis and model fitting. All the methods have been widely variety of methods, and the methods are described in detail in the
paper that is described in Brats and Martin (2010). We refer to the
authors. The strategy proposed in this thesis is comparable to the one
found in Brats and Martin [8]. The final example presented involves a non-stationary process. This example shows that the
proposed in Hasselt and Hameleers [79] and their extensions can be
Chapter 4

Time Series Analysis

4.1 Contributions and Time Series

This chapter empirically assesses the ability of the diagnostics proposed in Haslett and Hayes [29] and the new developments in chapters 2 and 3 to detect aberrant behaviour in a class of time series which are known as stationary equally spaced ARMA$(p,q)$ processes, see Box and Jenkins [7]. A wide variety of unusual behaviour that may occur to ARMA$(p,q)$ processes is described in Bruce and Martin [8] and a diagnostic is suggested by its authors. The strategy proposed in this thesis is compared to the approach found in Bruce and Martin [8]. The final example presented in this chapter involves a non-stationary process. This example shows that the methods proposed in Haslett and Hayes [29] and their extensions can be applied to...
non-stationary data. Thus the contribution methodology can be applied to the more general class of \(ARIMA(p, d, q)\) processes, see Box and Jenkins [7].

The next section introduces the prediction error decomposition, see Harvey [27], which is a fundamental tool in time series analysis. The importance of this decomposition is explained and its influence on the development of time series diagnostics is reviewed.

4.2 Prediction Error Decomposition

So far this thesis has assumed that the correlation structure of the fitted models are known or an estimate is available. For situations in time series where the correlation parameters have to be estimated the prediction error decomposition is invaluable. This decomposition enables the likelihood of complex time series to be expressed in a more manageable form. This form reduces the computational burden associated with maximum likelihood estimates; specifically the calculation of correlation parameters which are estimated by non-linear optimization, see Jones [31]. The following example illustrates how the prediction error decomposition works.

Let \(Y = \{y_1, \ldots, y_n\}\) be a vector of \(n\) observations with the following distribution \(Y \sim N(\Theta, V(\Theta))\) where \(\Theta\) denotes the set of correlation parameters which the covariance matrix of \(Y\) depends on. If \(\Theta\) is unknown an estimate,
\( \hat{\Theta} \), can be obtained by maximum likelihood, see Cox [15]. Maximizing the function

\[
L(\Theta; Y) = \frac{1}{\sqrt{(2\pi)^n|V(\Theta)|}} e^{-\frac{1}{2}(Y^tV(\Theta)^{-1}Y)}
\]

with respect to \( \Theta \) will obtain these estimates. \(|V(\Theta)|\) denotes the determinant of the matrix \( V(\Theta) \). A log transform may be performed on \( L(\Theta; Y) \) in order to simplify the calculation of these estimates, see Cox [15]. A log transform of \( L(\Theta; Y) \) yields

\[
\log L(\Theta; Y) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V(\Theta)| - \frac{1}{2}(Y^tV(\Theta)^{-1}Y).
\]

The same estimate for \( \Theta \) is obtained whether \( L(\Theta; Y) \) or \( \log L(\Theta; Y) \) is maximized with respect to \( \Theta \).

Maximizing \( \log L(\Theta; Y) \), in its current form, with respect to \( \Theta \) is computationally intensive and careful restrictions must be placed on the parameter space which is searched. The computational burden arises from the inversion of the \( V(\Theta) \) matrix. Searching the parameter space must be done in a manner which avoids the generation of non-positive definite matrices, see Jones [33]. Singular matrices may also be a concern. As a result of these issues a more appropriate representation of the log likelihood is sought. This representation is achieved using the prediction error decomposition.

It is possible to express \( L(\Theta; Y) \) as follows

\[
L(\Theta; Y) = \prod_{t=2}^{n} L(\Theta; y_t|y_{t-1}, \ldots, y_1) L(\Theta; y_1).
\]
This decomposition arises from the following axiom

\[ P(y_1, y_2) = P(y_2|y_1)P(y_1). \]

If \( Y \) is a series of observations from a temporal stochastic process then a natural ordering of the observations exists. It is this order which indicates the appropriate decomposition of the joint probability density function of \( Y \).

The log transform of \( L(\Theta; Y) \) yields

\[
\log L(\Theta; Y) = \sum_{t=2}^{n} \log L(\Theta; y_t|y_{t-1}, \ldots, y_1) + \log L(\Theta; y_1).
\]

In the case where \( Y \sim N(\Theta, V(\Theta)) \) then

\[
\log L(\Theta; y_1) = \log \frac{1}{\sqrt{2\pi f_1}} e^{-\frac{(y_1 - \hat{y}_1)^2}{2f_1}}
= \log \frac{1}{\sqrt{2\pi f_1}} e^{-\frac{y_1^2}{2f_1}}
= -\frac{y_1^2}{2f_1} - \frac{1}{2} \log 2\pi - \frac{1}{2} \log f_1
\]

where \( \hat{y}_1 = E[y_1|\hat{\Theta}] = 0 \) denotes the best linear unbiased predictor of the observation \( y_1 \) given the estimate \( \hat{\Theta} \) which is obtained using the entire data set \( \{y_1, \ldots, y_n\} \). \( f_1 \) denotes the variance of \( y_1 \).

Note that

\[
\log L(\Theta; y_t|y_{t-1}, \ldots, y_1) = \log \frac{1}{\sqrt{2\pi f_t}} e^{-\frac{(y_t - \hat{y}_t|y_{t-1})^2}{2f_t}}
= -\frac{(y_t - \hat{y}_t|y_{t-1})^2}{2f_t} - \frac{1}{2} \log 2\pi - \frac{1}{2} \log f_t
\]
where \( \hat{y}_{t|t-1} = E[y_t|y_{t-1}, \ldots, y_1, \hat{\Theta}] \) denotes the best linear unbiased predictor of the observation \( y_t \) given the observations \( \{y_1, \ldots, y_{t-1}\} \) and the estimate \( \hat{\Theta} \). \( f_t \) denotes the variance of \( y_t - \hat{y}_{t|t-1} \). Re-expressing \( \log L(\Theta; Y) \) yields

\[
\log L(\Theta; Y) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} \log f_t - \frac{1}{2} \sum_{t=2}^{n} \frac{(y_t - \hat{y}_{t|t-1})^2}{2f_t}
\]

If \( f_t \) and \( \hat{y}_{t|t-1} \) were known then maximizing this version of \( \log L(\Theta; Y) \) would be considerably less computationally intensive than the original form of the \( \log L(\Theta; Y) \) which requires the inversion of an \( n \times n \) matrix. However \( f_t \) and \( \hat{y}_{t|t-1} \) must be calculated. Fortunately, the Kalman Filter is an efficient algorithm for obtaining these values, see Singpurwalla [42] and Harvey [28]. Recent extensions by de Jong [17] and Proietti [46] have enabled the Kalman Filter to be used as an efficient algorithm for obtaining conditional residuals. As a result the Kalman Filter can be used to generate \( T_i \); however obtaining \( \tau_i \) using the Kalman Filter has not yet been achieved.

A by-product of the standard Kalman Filter is the generation of one-step ahead prediction errors, \( y_t - \hat{y}_{t|t-1} \), and their respective variances \( f_t \). The availability of these residuals and the assumption of independence makes these residuals desirable to work with. The one-step predictor error divided by its standard deviation has an approximate t-distribution thus a t-test can be performed on each residual as an approximate way of checking for observations which depart from the model. These residuals are also used
in Diggle [21] and Box and Jenkins [7] to ensure the appropriate model is selected. Due to the availability and properties of these residuals it is not surprising that diagnostics were developed based on these residuals. The $DV(A)$ diagnostic proposed in Bruce and Martin [8] is a function of the variance of these residuals. This statistic and other common time series diagnostics are reviewed in the next section.

4.3 Time Series Diagnostic Review

Attention is focused on diagnostics which are applied to the general class of time series models known as $ARMA(p,q)$ processes. These processes are defined as follows

$$y_t = \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

where the set of random variables $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent and identically distributed such that $\varepsilon_t \sim N(0, \sigma^2)$ $\forall t \in \{1, \ldots, n\}$. The set of random variables $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are often referred to in the time series literature as the innovations of the process, see Fox [23]. Their variance $\sigma^2$ is referred to as the variance of the innovations.
4.3.1 Reviewing DV(A)

The following statistic

\[ DV(A) = \frac{n}{2} \left( \frac{\hat{\sigma}_e^2(\hat{\Theta})}{\hat{\sigma}_e^2(\hat{\Theta}(A))} - 1 \right)^2 \]

is proposed in Bruce and Martin [8] as a diagnostic for ARMA\((p, q)\) processes and indeed generalizes for ARIMA\((p, d, q)\) processes. \(\hat{\sigma}_e^2(\hat{\Theta})\) denotes the estimated variance of the innovations and \(\hat{\Theta}\) denotes a vector of maximum likelihood estimates for the correlation parameters \(\Theta = \{\varphi_1, \ldots, \varphi_p, \theta_1, \ldots, \theta_q\}\).

\(\hat{\sigma}_e^2(\hat{\Theta}(A))\) denotes the estimate of the innovation variance when the subset \(A\) is omitted where \(\hat{\Theta}(A)\) is the maximum likelihood estimate for \(\Theta\) omitting \(A\). The reference distribution for \(DV(A)\) is a chi-squared distribution with one degree of freedom, \(\chi^2\), and a subset \(A\) is considered by Bruce and Martin [8] to be outlying if the p value associated with \(DV(A)\) is less than 0.5.

Unlike the contribution approach the \(DV(A)\) statistic re-estimates the correlation parameters every time an observation or set of observations is tested for unusual behaviour. Thus the \(DV(A)\) statistic is considerably more computationally expensive than the contribution statistics. However Bruce and Martin [8] recommend the re-estimation of the correlation parameters as outliers can bias correlation parameters and subsequently bias the test statistic. This problem is mentioned in Peña [45] with an example of biasing caused to estimates of AR(1) correlation parameters. The detection of
observations which are over-influential on correlation parameter estimates is discussed in LeFrançois [38] and Peña [45]. However LeFrançois [38] warns that measures of influence are meant to be used to detect over-influential observations and not used to detect outliers, as that is not their purpose.

Although outliers can affect correlation estimates the examples analyzed in this chapter can be detected without resorting to the re-estimation of correlation parameters. This suggests that the damage done to the correlation parameters is not a major concern for the detection of outliers when using the contribution approach.

4.3.2 Likelihood Ratio

Two types of outliers known as additive and innovative outliers are considered in Fox [23]. The author recommends a likelihood ratio approach. These methods consist of a likelihood ratio of the hypotheses that an observation is outlying in a particular manner against the null that the observation is not outlying. A separate test is proposed for each type of outlier. Chang et al. [9] extended this work by proposing alternative diagnostics based on likelihood ratios including a test to distinguish between additive and innovation outliers. Muirhead [44] also considers this problem using a likelihood ratio approach. Ljung [39] reviews the leave-$K_A$-out diagnostic proposed in Bruce and Martin
and establishes the relationship between the likelihood ratio approach and leave-$\kappa_A$-out diagnostics. For the detection of additive outliers Ljung [39] shows that likelihood ratio tests are nearly equivalence to deletion diagnostics based on $\hat{\sigma}_e^2$ such as the $DV(A)$ statistic. Bruce and Martin [8] provide simulations which show that their diagnostic performs better at detecting additive outliers then innovation outliers. Ljung [39] recommends using the statistic

$$\hat{\nu}_t^* = \frac{\hat{\epsilon}_t}{\hat{\sigma}_e(i)}$$

over deletion based statistics and provides simulations of its superiority at detecting innovation outliers. Note $\hat{\nu}_t^*$ is approximately $t_{n-1}$ distributed.

### 4.3.3 Leave-$\kappa_A$-Out Diagnostics

For general linear models leave-$\kappa_A$-out diagnostics are used to detect influential observations where an observation is deemed influential if it is unusually important in determining the estimation of fixed effects. Diagnostics for detecting these observations such as Cook’s distance, Cook’s modified distance and DFBETA can be found in Cook [13], Atkinson [2] and Belsley et al. [6] respectively. These diagnostics compare estimates of the fixed effects using the entire data set to estimates obtained when a single observation or subset of observations are omitted from the analysis. Belsley et al. [6] notes that
outlying points can have a considerable impact on variance estimates and
subsequently recommends re-estimating the variance parameters with the
suspect observations removed. Andrews and Pregibon [1] propose a statistic
which examines both the effect an observation has on variance estimates and
its leverage. This statistic is reviewed in Cook and Weisberg [14], section 4.2.
A selection of these model diagnostics are extended in Barrett and Ling [5]
for general linear models and special cases such as mixed models are covered
in Banerjee and Frees [4].

For time series analysis leave-$k_A$-out diagnostics can refer to diagnostics
used to detect observations which are not necessarily influential on fixed ef-
facts estimates but which may bias variance or correlation estimates. The ex-
amples presented in Bruce and Martin [8] assume the fixed effects are known
or can be removed by differencing. In de Jong and Penzer [18] concerns are
mentioned about the computational aspects of leave-$k_A$-out diagnostics and
situations in which they are not appropriate. As a result of these concerns
a new diagnostic is proposed in de Jong and Penzer [18] with an example of
the ability of this diagnostic at detecting level shifts. It is considerably more
computationally efficient than $DV(A)$ since it does not re-estimate any pa-
rameters obtained by maximum likelihood. Later in this chapter the ability
of the contribution approach to detect level shifts will be illustrated with this
example.
4.4 A Single Outlier

The purpose of the next two examples is to display the ability of the contribution methodology to detect a single additive or innovative outlier when observations are stationary, equally spaced and serially correlated. The contribution methodology does not distinguish between these different types of outliers nor does the $DV(A)$ diagnostic.

4.4.1 An Additive Outlier

Consider the set of observations $\{y_1, \ldots, y_n\}$ with an additive outlier at time $k$ then

$$y_t = \begin{cases} 
  u_t & t \neq k \\
  u_t + \Delta & t = k 
\end{cases}$$

where $u_t$ is the underlying process of the data which is unaffected by the outlier and $\Delta$ is the effect of the additive outlier. When a time series contains an additive outlier, its presence affects no other observation in the series.

The first example presented in Bruce and Martin [8] is of a stationary first order autoregressive process with a single additive outlier present. This model is achieved by letting

$$u_t = \varphi u_{t-1} + \varepsilon_t$$
where $\varphi$ is a correlation parameter with the constraint $|\varphi| < 1$ which ensures stationarity. The set of random variables $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent and identically distributed such that $\varepsilon_t \sim N(0, \sigma^2_\varepsilon) \forall t \in \{1, \ldots, n\}$. The parameters $\varphi$ and $\sigma^2$ are given the same values as those in Bruce and Martin [8] hence $\varphi = 0.4$ and $\sigma^2_\varepsilon = 1$. It is assumed that the mean of the process generating $u_t$ is known and equal to zero. Using this model a series containing one hundred observations, $n = 100$, was simulated using the computer package Splus, see Splus [48]. A single additive outlier such that $\Delta = +4$ is placed at the time point 28. This series is displayed in Figure 4.1 and the outlier is represented by the symbol $\varnothing$. An $AR(1)$ model was fitted to the data and
maximum likelihood estimates for $\varphi$ and $\sigma^2_\epsilon$ were obtained. The correlation parameter $\varphi$ was estimated at $\hat{\varphi} = 0.28$ and the variance parameter $\sigma^2_\epsilon$ was estimated at $\hat{\sigma}^2_\epsilon = 1.02$. Assuming an $AR(1)$ model with these estimates the standardized contributions $\tau_i$ were calculated for each observation in the series. Table 4.1 contains the diagnostic analysis performed on the observations around and including observation 28. Observation 28 generates a large positive contribution this is due to the fact that observation has both a large positive marginal residual and conditional residual. From Table 4.1 it is clear that both $\tau_i$ and the diagnostic DV(i) proposed in Bruce and Martin [8] detect the additive outlier. Neither statistic encounters problems with smearing in this example. This is due to the very low estimate of the correlation parameter $\hat{\varphi} = 0.28$.

In order to explore issues concerning smearing this experiment was repeated 300 times and the proportion of observations correctly and falsely identified was calculated. This study was carried out on three $AR(1)$ processes with a single additive outlier of $\Delta = -4$ at location 28. The correlation parameter $\varphi$ took the values 0.3, 0.6, and 0.9, and the variance parameter $\sigma^2_\epsilon$ was set at $\sigma^2_\epsilon = 1$ for all three processes. These experiments are summarized in Figure 4.2. In this figure the results corresponding to $\varphi = 0.3$ are represented by $-$, $-$ represents $\varphi = 0.6$ and $\cdots$ represents $\varphi = 0.9$. From Figure 4.2 it can be seen that the rate of false detection of outliers either side
Table 4.1: Diagnostic Results for a Single Additive Outlier

<table>
<thead>
<tr>
<th>Index</th>
<th>$y_i$</th>
<th>$T_i$</th>
<th>P value of $\tau_i$</th>
<th>P value of $DV(i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>1.12</td>
<td>1.16</td>
<td>0.28</td>
<td>1.00</td>
</tr>
<tr>
<td>27</td>
<td>-0.46</td>
<td>0.80</td>
<td>0.37</td>
<td>0.86</td>
</tr>
<tr>
<td>28</td>
<td>3.37</td>
<td>12.53</td>
<td>0.00</td>
<td>0.31</td>
</tr>
<tr>
<td>29</td>
<td>-0.10</td>
<td>0.08</td>
<td>0.74</td>
<td>0.97</td>
</tr>
<tr>
<td>30</td>
<td>-1.05</td>
<td>1.65</td>
<td>0.20</td>
<td>0.92</td>
</tr>
</tbody>
</table>

4.4.2 An Innovative Outlier

Depending on the actual correlation a single innovation (or the error term) can have a substantial influence on the rate of detection of an additive outlier $\Delta$. In the example, the outlier is identified approximately 95% of the time, even when $\psi = 0.3$. For $\psi = 0.3$ the rate of detection of the outlier is similar in all cases. However, this increases as $\psi$ increases. For $\psi = 0.9$ the detection rate increases to almost 100%.

Figure 4.2: Rate of Detection for An Additive Outlier

65
of the outlying observation increases as $\varphi$ increases in value. For $\varphi = 0.3$ the observations 27 and 29 are detected about 20% of the time; when $\varphi = 0.9$ this increases to about 40%. There is also some evidence to suggest that the rate of correct identification of the outlier at the location 28 is dependent on the correlation parameter however this effect appears quite minor. Even when $\varphi = 0.9$ a single additive outlier of $\Delta = +4$ at this location is identified approximately 90% of the time. For $\varphi = 0.3$ and $\varphi = 0.6$ the additive outlier is identified approximately 95% of the time.

4.4.2 An Innovative Outlier

Depending on the serial correlation a single innovative outlier has the potential to influence all future observations although in the stationary case its influence on an observation decreases the further away it is from that observation. The second example encountered in Bruce and Martin [8] is a first order autoregressive process, of length a hundred, with a single innovation outlier. This can be described by the following model

$$y_t = \varphi y_{t-1} + u_t$$

and

$$u_t = \begin{cases} 
\epsilon_t & t \neq k \\
\epsilon_t + \Delta & t = k 
\end{cases}$$
where $\varphi$ is a correlation parameter and an innovation outlier, $\Delta$, is introduced at time $k$. The set of random variables $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent and identically distributed such that $\varepsilon_t \sim N(0, \sigma^2_\varepsilon) \forall t \in \{1, \ldots, n\}$. The parameters $\varphi$ and $\sigma^2_\varepsilon$ are given the values $\varphi = 0.4$ and $\sigma^2_\varepsilon = 1$ in Bruce and Martin [8]. A single innovation outlier of $\Delta = +4$ is placed at the time point 28. Due to the small magnitude of $\varphi$ the effect of the single innovation outlier rapidly becomes negligible, for all intents and purposes this example could be viewed as a first order autoregressive model with a single additive outlier at time point 28. The $DV(i)$ diagnostic only detects one unusual point which is observation 28. This simulation is duplicated here except the parameter $\varphi$ is given the value $\varphi = 0.8$ and an innovation outlier of $\Delta = +8$ is introduced at time 28. Due to the high value of $\varphi$ the effect of the innovation outlier is non-negligible on a number of future observations. The data is displayed in Figure 4.3 and the outlier is represented by the symbol $\oplus$. From Table 4.2 it is clear that the $DV(i)$ diagnostic detects only two unusual observations indexed by 27 and 28. The contribution approach detects six unusual observations at time 27, 28, 29, 30, 32 and 33. Both methods falsely identify observation 27. Note the terms Lower and Upper correspond to lower and upper bounds for the probability of negative contributions.

This experiment was repeated 300 times and the proportion of observations correctly and falsely identified was calculated. This study was carried
Figure 4.3: Single Innovation Outlier
Table 4.2: Diagnostic Results for a Single Innovation Outlier

<table>
<thead>
<tr>
<th>Index</th>
<th>$y_i$</th>
<th>$T_i$</th>
<th>P value of $\tau_i$</th>
<th>P value of $DV(i)$</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>-1.09</td>
<td>-1.11</td>
<td>-</td>
<td>0.99</td>
<td>0.04</td>
<td>0.20</td>
</tr>
<tr>
<td>27</td>
<td>-1.48</td>
<td>6.81</td>
<td>0.04</td>
<td>0.14</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>28</td>
<td>5.81</td>
<td>27.20</td>
<td>0.00</td>
<td>0.13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>29</td>
<td>5.40</td>
<td>6.22</td>
<td>0.05</td>
<td>0.99</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>30</td>
<td>3.23</td>
<td>-5.05</td>
<td>-</td>
<td>0.94</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>31</td>
<td>3.80</td>
<td>-1.17</td>
<td>-</td>
<td>0.95</td>
<td>0.04</td>
<td>0.19</td>
</tr>
<tr>
<td>32</td>
<td>5.02</td>
<td>12.07</td>
<td>0.01</td>
<td>0.78</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>33</td>
<td>2.35</td>
<td>-3.46</td>
<td>-</td>
<td>0.95</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>34</td>
<td>2.25</td>
<td>1.26</td>
<td>0.28</td>
<td>0.96</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
out on four $AR(1)$ processes with a single innovation of $\Delta = +8$ at location 28. The correlation parameter $\varphi$ took the values $0.3, 0.6, 0.8$ and $0.9$, and the variance parameter $\sigma_{\varepsilon}^2$ was set at $\sigma_{\varepsilon}^2 = 1$ for all three processes. These experiments are summarized in Figure 4.4. In this figure the results corresponding to $\varphi = 0.3$ are represented by — , — — represents $\varphi = 0.6$, - - represents $\varphi = 0.8$ and ⋯ represents $\varphi = 0.9$. The innovation outlier at location 28 is detected above 95% of the time for all four processes. From Figure 4.4 it can be seen that the rate of false detection of outliers either side of the outlying observation is asymmetric and increases as $\varphi$ increases in value. This asymmetry is caused by the effect of the innovation outlier on observations after location 28. As $\varphi \to 1$ then the effect of the shock intro-
duced at location 28 on future observations is similar to a level shift. This may explain the smearing behaviour experienced by $\tau_i$ as $\varphi \to 1$ in Figure 4.4. The detection of level shifts is explored later in this chapter.

4.5 Subsets of Unusual Observations

The third example presented in Bruce and Martin [8] is of a stationary first order moving average process, see Box and Jenkins [7], with a contiguous patch of additive outliers and a single isolated additive outlier. Due to the remoteness of this single additive outlier relative to the patch of unusual observations its presence does not mask, see Atkinson [3], or smear any other additive outlier. For these reasons the simulation was duplicated with the single additive outlier removed.

Let $A$ be a set which indexes the patch of additive outliers then

\[ y_t = \begin{cases} 
  u_t & t \notin A \\
  u_t + \Delta_t & t \in A 
\end{cases} \]

where \( \{y_1, \ldots, y_n\} \) is the set observations, \( \{u_1, \ldots, u_n\} \) is a first order moving average process and $\Delta_t$ is the effect of an additive outlier on observation $y_t$.

The process $u_t$ is generated by

\[ u_t = \theta \varepsilon_{t-1} + \varepsilon_t \]
where $|\theta| < 1$, to ensure stationarity, and $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent and identically distributed variables such that $\varepsilon_t \sim N(0, \sigma^2_t) \ \forall t \in \{1, \ldots, n\}$. The correlation parameter, $\theta$, was given the value $\theta = 0.5$ and the variance of the innovations, $\sigma^2_\varepsilon$, was set at $\sigma^2_\varepsilon = 1$. One hundred observations from a first order moving average process were generated using these parameters. A patch of additive outliers were introduced at the times 60, 61 and 62 by adding +4 to the observations at these points. The data is displayed in Figure 4.5 and the outliers are represented by $\oplus$. The results of the diagnostic analysis is displayed in Table 4.3. The contribution approach detects the entire patch of additive outliers and falsely identifies adjacent points at each end of the
### Table 4.3: Diagnostic Results for a Patch of Additive Outliers

<table>
<thead>
<tr>
<th>Index</th>
<th>$y_i$</th>
<th>$T_i$</th>
<th>P value of $\tau_i$</th>
<th>P value of $DV(i)$</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>-1.14</td>
<td>-1.60</td>
<td>-</td>
<td>0.72</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>60</td>
<td>3.05</td>
<td>13.13</td>
<td>0.00</td>
<td>0.58</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>61</td>
<td>4.73</td>
<td>28.64</td>
<td>0.00</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>62</td>
<td>5.89</td>
<td>33.45</td>
<td>0.00</td>
<td>0.04</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>63</td>
<td>-1.33</td>
<td>-2.09</td>
<td>-</td>
<td>0.33</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

patch. The $DV(i)$ diagnostic detects observation 62 and falsely identifies observation 63.

This example is the first case in which a subset of observations are unusual as opposed to a single datum. Fortunately the contribution approach detected this entire subset by examining the observations singly; however the contribution methodology has the ability to deal with subsets as well. The $DV(i)$ diagnostic has a similar property and examples are presented in Bruce and Martin [8] of its effectiveness at detecting subsets of unusual observations.

This experiment was repeated 300 times and the proportion of observations correctly and falsely identified was calculated. This study was carried out on three $MA(1)$ processes with three additive outliers of $\Delta = +4$ placed
at locations 60, 61 and 62. The correlation parameter $\theta$ took the values 0.3, 0.6 and 0.9, and the variance parameter $\sigma^2_\varepsilon$ was set at $\sigma^2_\varepsilon = 1$ for all three processes. These experiments are summarized in Figure 4.6. In this figure the results corresponding to $\theta = 0.3$ are represented by $-$, $-$ represents $\theta = 0.6$ and $\cdots$ represents $\theta = 0.9$. From Figure 4.6 it is clear that irrespective of the value that the correlation parameter $\theta$ takes examining the contributions of single observations is sufficient to detect this entire patch of outliers. For the three processes the outliers are detected at the locations 60, 61 and 62 about 95% of the time. The false detection of outliers due to smearing grows quite rapidly as $\theta \to 1$. Figure 4.6 suggests that using the contribution methodology to detect patches of additive outliers is more effective than the
Figure 4.7: Rate of Detection for a Patch of Additive Outliers using $DV(i)$

$DV(i)$ diagnostic. It is shown in Bruce and Martin [8] that single examination of observations using $DV(i)$ performs quite poorly at detecting patches of unusual observations and the authors recommend additional subset examination. In Figure 4.7 the performance of the $DV(i)$ statistic at detecting this patch of additive outliers can be viewed. It is clear that the ability of the $DV(i)$ statistic to detect this patch of additive outliers depends on the correlation parameter $\theta$. This simulation shows that the $DV(i)$ diagnostic suffers less from smearing than the standardized contributions $r_i$. However, masking appears to be a greater issue for $DV(i)$ than $r_i$.

The unusual behaviour presented so far could all be detected by examining the contributions of single observations. An example is provided in the
4.6 Level Shifts

In this section the detection of level shifts are considered. These subsets of unusual behaviour present situations where examining the contributions of single observations can fail to detect these shifts. An appropriate strategy is developed for their detection and an example is presented.

Figure 4.8 is a plot of the Nile data set which can be found in Cobb [12]. This data set consists of the annual flow volume of the Nile River at Aswan for the years 1871 to 1970. The primary interest in this data set arises from
the belief that there is a level shift around the year 1900. In Cobb [12] it is noted that a dam began operation on the River in 1902 and also that rainfall records indicate a change around this period.

This data set is used in de Jong and Penzer [18] to illustrate the ability of various diagnostics to detect aberrant behaviour in time series. For the Nile data, the following model is assumed in de Jong and Penzer [18]

\[ y_t = x_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \]

\[ x_t = x_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma^2) \]

where \( \sigma^2 = 15,099 \) and \( \sigma^2 = 1,469.2 \). It is assumed that \( \varepsilon_t \) and \( \eta_t \) are uncorrelated. This model is a random walk plus noise which is non-stationary and hence its likelihood function is not defined. A conditional likelihood function is recommend in de Jong and Penzer [18] to obtain the estimates \( \sigma^2 \) and \( \sigma^2 \). This is achieved by conditioning on the first observation such that \( x_0 \) is assumed to be known and set to be the flow volume at year 1871. Thus \( \{y_1, \ldots, y_n\} \) corresponds to the data collected from years 1872 to 1970 and hence diagnostics are carried out only on these years.

The analysis of the Nile data set in de Jong and Penzer [18] indicates two outliers and one level shift. The outliers are detected at 1877 and 1913, and the level shift is placed at 1899. Assuming the same model proposed in de Jong and Penzer [18] the contribution of each data point was generated
Table 4.4: Unusual Contributions for Nile Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Flow Volume</th>
<th>$T_i$</th>
<th>P value of $\tau_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1877</td>
<td>813</td>
<td>5.75</td>
<td>0.03</td>
</tr>
<tr>
<td>1913</td>
<td>456</td>
<td>15.10</td>
<td>0.00</td>
</tr>
</tbody>
</table>

and examined, only two observations are deemed unusual: the observations at 1877 and 1913. Table 4.4 displays this information. Singly examining the contribution of each data point gives no indication of a level shift. Checking the contributions corresponding to subsets of size 2, 3, 4 and 5 also proved fruitless. The following strategy is proposed to check for a level shift: a running sum of the contributions and their corresponding p values is generated backwards from year 1970. This information is displayed in Figure 4.9 and this graph indicates that the largest unusual contiguous block of observations stretches from 1913 to 1970. The year 1913 was identified as outlying by examining observations singly hence concerns arise that the observation at 1913 might be masking or smearing the detection of the level shift.

This analysis was redone with the observation at the year 1913 removed and as a result the subset stretching from 1901 to 1970 is the largest contiguous block of observations with a contribution which has a p value below 0.05, see Figure 4.10. This analysis places the level shift around the same
Figure 4.9: Level Shift Diagnostics

![Figure 4.9: Level Shift Diagnostics](image)

4.7 Conclusions

The examples provided in this chapter provide concrete evidence of the ability of the contribution diagnostics such as a diagnostic test for time series. A wide variety of unusual behaviors in regression data and time series is revealed by contribution diagnostics. The extent to which the contribution statistic $r_A$ suffers from skewing and masking has been explored.

Figure 4.10: Level Shift Diagnostics with the year 1913 Omitted

![Figure 4.10: Level Shift Diagnostics](image)
year as de Jong and Penzer [18] and Cobb [12]. It is clear that the outlier present at 1913 masked the extent of the level shift.

4.7 Conclusions

The examples presented in this chapter provide empirical evidence of the ability of the contribution approach as a diagnostic tool for time series. A wide variety of unusual behaviour in time series can be detected by contribution diagnostics such as additive and innovative outliers, patches of outliers and level shifts. The extent to which the contribution statistic $r_A$ suffers from smearing and masking has been explored.

The contribution diagnostics have been successfully used to analysis regression data and time series. A natural extension is the application of the contribution methodology to multilevel data (also called hierarchical data), see Goldstein [24]. The contribution methodology is used in Chapter 5 to detect unusual behaviour in repeated measures which is a type of multilevel data.
Chapter 5

Longitudinal Models

5.1 Laird-Ware Models

In this chapter the contribution methodology is applied to repeated measures or longitudinal data, see Diggle et al. [20] and Crowder and Hand [16]. Attention is focused on longitudinal data with random effects, see Laird and Ware [37], and later more general versions are explored which allow the errors to come from $ARMA(p, q)$ processes, see Jones [33].

For repeated measures interest lies in not just detecting unusual observations but also unusual individuals. Thus any diagnostic approach applied to repeated measures should encompass both concerns. The contribution methodology appears to be a very natural means of addressing both these concerns due to the fact that the contribution of an individual is simply the
sum of the contributions of its observations. Empirical evidence is provided of the ability of the contribution diagnostics to detect both outlying observations and individuals. Note that, in the usual terminology of multilevel modelling, observations are level 1 units and individuals are level 2 units, see Goldstein and Woodhouse [25] pages 13-26.

Later in this chapter, the approximations for standardization contributions defined in chapter 3 are examined for a repeated measures study performed in Chi and Reinsel [10]. From this analysis further evidence is obtained of the appropriateness of $T_A$ and $\tau_A^*$ as approximations for $\tau_A$.

5.2 Random Effects

5.2.1 Outlying Observations

The exploration of the contribution methodology as a diagnostic tool for repeated measures starts with the simple model

$$y_{ij} = \eta_i + \varepsilon_{ij}$$

where $\text{var}(\eta_i) = \sigma^2_\eta$ and $\text{var}(\varepsilon_{ij}) = \sigma^2_\varepsilon \ \forall \ i \in \{1, \ldots, I\}$ and $\forall \ j \in \{1, \ldots, J\}$. The random effect associated with the $i^{th}$ individual is denoted by $\eta_i$ and it is assumed that individuals are independent. The individuals in this repeated measures model are indexed by $i$ and the number of individuals is denoted by $I$. The observations of each individual are indexed by $j$ and each individual has $J$ observations. The random variables $\varepsilon_{ij}$ are assumed to be
independent.

Simpler interpretations for the contributions and standardized contributions associated with this model were sought. If $A$ indexes an arbitrary subset of observations then $\phi_A = \kappa_A$, see proposition 3.4.1, however no meaningful interpretation of $\gamma_A$ could be found, see Appendix A.2, except when $A$ indexes an entire individual then $\gamma_A = J$ subsequently $T_A = \tau_A = e_A^T V_A^{-1} e_A$.

Even for this relatively simple model finding a meaningful interpretation of $T_A$ and $\tau_A$ was not possible. Since theoretical exploration of $T_A$ and $\tau_A$ proved less than fruitful simulations were used to confirm the ability of contribution methodology as a diagnostic tool for repeated measures.

5.2.1 Outlying Observations

In this subsection a simulation is presented showing the standardized contribution's ability at detecting a single outlying observation in a repeated measures data set. Using the same model from the previous section $y_{ij} = \eta_i + \varepsilon_{ij}$

where $\text{var}(\eta_i) = \sigma^2_{\eta}$ and $\text{var}(\varepsilon_{ij}) = \sigma^2_{\varepsilon}$ then $I = 10$ individuals were simulated each containing $J = 10$ observations. For this simulation $\sigma^2_{\eta} = 1$ and $\sigma^2_{\varepsilon} = 1$. A shock of $\Delta = +4$ was given to the first observation belonging to the $10^{th}$
The data generated by this model are displayed in Figure 5.1 and the outlying observation is indicated by the symbol $\Theta$.

Given this simulated data and assuming $y_{ij} = \eta_i + \varepsilon_{ij}$ $\forall i \in \{1, \ldots, I\}$ and $\forall j \in \{1, \ldots, J\}$, the parameters $\sigma^2_\eta$ and $\sigma^2_\varepsilon$ were estimated by maximum likelihood. The following values were obtained $\hat{\sigma}^2_\eta = 1.35$ and $\hat{\sigma}^2_\varepsilon = 1.02$. Assuming the same model for $y_{ij}$ and using these maximum likelihood estimates the contribution methodology was applied to this data set. The standardized contributions associated with the individual, $i = 10$, containing
the single outlying observation, $j = 1$, are summarized in Table 5.1. The standardized contribution associated with the outlying observation indicates this observation is unusual, see Table 5.1. Note the terms Lower and Upper correspond to lower and upper bounds for the probability of negative contributions.

Table 5.1 also contains the standardized contributions associated with contiguous blocks of observations such that $A = \{1\}$, $A = \{1, 2\}$, \ldots, $A = \{1, \ldots, 10\}$. The single outlying observations at $j = 1$ has a considerable smearing effect on the standardized contributions associated with these subsets. In fact, the individual is deemed by its standardized contribution to be
outlying. Thus care must be taken when using the contribution methodology that an individual is not classified as unusual when the individual contains a few unusual observations or indeed a single observation which deviates from the assumed model.

5.2.2 Outlying Individuals

In this subsection a simulation is presented showing the contribution's ability at detecting a single outlying individual in a repeated measures data set. Using the same model from the previous section

\[ y_{ij} = \eta_i + \epsilon_{ij} \]

where \( \text{var}(\eta_i) = \sigma^2_\eta \) and \( \text{var}(\epsilon_{ij}) = \sigma^2_\epsilon \) then \( I = 10 \) individuals were simulated each containing \( J = 10 \) observations. For this simulation \( \sigma^2_\eta = 1 \) and \( \sigma^2_\epsilon = 1 \). A shock of \( \Delta = +4 \) was given to the random effect associated with the 10\(^{th}\) individual thus the new model is

\[ y_{ij} = \begin{cases} \eta_i + \Delta + \epsilon_{ij} & i = 10 \\ \eta_i + \epsilon_{ij} & i \neq 10. \end{cases} \]

The data generated by this model is displayed in Figure 5.2 and the observations of this outlying individual are represented by the symbol \( \oplus \).

Given the simulated data and assuming \( y_{ij} = \eta_i + \epsilon_{ij} \) \( \forall i \in \{1, \ldots, I\} \) and \( \forall j \in \{1, \ldots, J\} \), the parameters \( \sigma^2_\eta \) and \( \sigma^2_\epsilon \) were estimated by maximum
likelihood. The following values were obtained $\hat{\sigma}_\eta^2 = 3.36$ and $\hat{\sigma}_e^2 = 1.16$.

Assuming the same model for $y_{ij}$ and using these maximum likelihood estimates the contributions methodology was applied to this data set. The standardized contributions associated with the outlying individual, $i = 10$, are summarized in Table 5.2. Singly examining the observations in this data set indicates that 7 out of the 10 observations which make up this individual are deemed by their standardized contributions to be unusual, see Table 5.2.

The standardized contributions associated with contiguous blocks of observations such that $A = \{1\}, A = \{1, 2\}, \ldots, A = \{1, \ldots, 10\}$ are also displayed in Table 5.2 and each subset is deemed by their standardized contributions as unusual.
### Table 5.2: Contributions for Outlying Individual

<table>
<thead>
<tr>
<th>$j$</th>
<th>$\tau_{10,j}$</th>
<th>P value of $\tau_{10,j}$</th>
<th>Lower</th>
<th>Upper</th>
<th>$\tau_{10,A}$</th>
<th>P value of $\tau_{10,A}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.90</td>
<td>0.03</td>
<td>-</td>
<td>-</td>
<td>4.90</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>1.57</td>
<td>0.21</td>
<td>-</td>
<td>-</td>
<td>6.88</td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>7.54</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
<td>15.74</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>-2.54</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
<td>13.03</td>
<td>0.01</td>
</tr>
<tr>
<td>5</td>
<td>8.56</td>
<td>0.00</td>
<td>-</td>
<td>-</td>
<td>23.87</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>3.40</td>
<td>0.07</td>
<td>-</td>
<td>-</td>
<td>28.63</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>-1.33</td>
<td>-</td>
<td>0.01</td>
<td>0.02</td>
<td>27.23</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>-2.29</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
<td>24.33</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>-2.34</td>
<td>-</td>
<td>0.00</td>
<td>0.00</td>
<td>21.12</td>
<td>0.01</td>
</tr>
<tr>
<td>10</td>
<td>1.56</td>
<td>0.21</td>
<td>-</td>
<td>-</td>
<td>23.56</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### 5.3 Random Effects and $AR(1)$ Errors

Models for longitudinal data which have fixed effects, random effects and $AR(1)$ errors are explored in Chi and Reinsel [10]. A test is proposed in their paper which given a model with fixed and random effects determines whether the errors should be modeled by an autoregressive structure rather than assuming independence. As a result they propose a model with three fixed effects, a random intercept and $AR(1)$ errors to model data found in Zerbe [51]. This data is displayed in Figure 5.3 and consists of plasma inorganic phosphate (mg/dl) measurements determined from 8 blood samples at 0, .5 hours after a glucose tolerance test.
Figure 5.3: Zerbe’s Data

, 1, 1.5, 2, 3, 4 and 5 hours after a glucose tolerance test. This study was carried out on 20 obese patients. The model for patient $i$ is

$$y_i = X_i \beta + C_i \eta_i + u_i, \quad i = 1, \ldots, 20,$$

where $y_i$ is a $8 \times 1$ vector of observations, $X_i$ is the $8 \times 3$ design matrix for the mean vector of patient $i$ and has the form

$$X_i = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.5 & 1 & 1.5 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 \end{pmatrix}^T,$$

$\beta$ is the $3 \times 1$ population fixed effect parameter vector, $C_i = (1, \ldots, 1)$ is the $8 \times 1$ design matrix for the random effect of patient $i$, $\eta_i$ is the random effect associated with patient $i$ where $\eta_i \sim N(0, \sigma^2_\eta)$ $\forall i \in \{1, \ldots, 20\}$, $u_i$ is the $8 \times 1$
vector of within-individual errors whose components are assumed to follow the AR(1) model

\[ u_{ij} = \varphi u_{ij-1} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_e^2) \]

\( \forall j \in \{1, \ldots, 8\} \). Under this model, Chi and Reinsel [10] obtained the following parameter estimates: \( \hat{\beta}^T = (4.506, -0.677, 0.286) \), \( \sigma_n^2 = 0.2696 \), \( \sigma_e^2 = 0.1265 \) and \( \varphi = 0.6973 \). Using the estimate \( \hat{\beta} \) for the population fixed effects then the expected value of an obese patient’s plasma inorganic phosphate (mg/dl) over time was calculated and these values are displayed in Figure 5.4. The ability of their model to predict the plasma inorganic phosphate curves of individuals is examined in their paper. However no diagnostics are presented in Chi and Reinsel [10] which indicate whether influential
Table 5.3: Observations with Unusual Contributions in Zerbe’s Data

<table>
<thead>
<tr>
<th>Patient</th>
<th>Time (Hour)</th>
<th>$T_{ij}$</th>
<th>$\tau_{ij}$</th>
<th>P value of $\tau_i$</th>
<th>Upper</th>
<th>Lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>8.78</td>
<td>5.25</td>
<td>0.02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>-2.19</td>
<td>-1.11</td>
<td>-</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>-4.17</td>
<td>-1.84</td>
<td>-</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>10.11</td>
<td>5.97</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{1}{2}$</td>
<td>11.11</td>
<td>6.52</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>10.82</td>
<td>6.35</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>16.00</td>
<td>9.32</td>
<td>0.00</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>8.02</td>
<td>5.72</td>
<td>0.02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>$1\frac{1}{2}$</td>
<td>-4.44</td>
<td>-1.96</td>
<td>-</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>8.49</td>
<td>5.16</td>
<td>0.02</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>6.60</td>
<td>4.06</td>
<td>0.04</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>-2.06</td>
<td>-1.02</td>
<td>-</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>6.34</td>
<td>3.91</td>
<td>0.04</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

individuals or outlying observations are present.

Using their proposed model the contribution methodology was applied to this data set. The unusual standardized contributions generated by single observations are presented in Table 5.3. These unusual observations are highlighted in Figure 5.3 by the symbol $\Theta$. From this graph and Table 5.3 it is clear that no time point is a particular focus for observations with unusual contributions.

Influence diagnostics for these models can be found in Banerjee and Frees.
[4]. These diagnostics can be viewed as extensions of Cook's distance and are concerned with detecting individuals which are over influential in the estimation of fixed effects. When these diagnostics were applied to this data set no influential individuals were detected. Three individuals were detected as unusual by their standardized contributions. Figure 5.5 is a plot of the standardized contributions of individuals, \( \tau_A \), against their Cook's distance, \( D_A \). The three unusual individuals are represented in Figure 5.5 by the symbol \( \oplus \). In chapter 2 it was theoretically established that for independent observations their standardized contributions are linearly related to the corresponding Cook's distance. From Figure 5.5 the relationship between standardized contributions and Cook's distance for independent individuals
Table 5.4: Unusual Standardized Residuals of Observations associated with Outlying Individuals

<table>
<thead>
<tr>
<th>Patient</th>
<th>Time (Hour)</th>
<th>$\hat{z}_{ij}$</th>
<th>$\bar{z}_{(ij)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>$\frac{1}{2}$</td>
<td>2.84</td>
<td>3.91</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>-3.19</td>
<td>-3.39</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>5.31</td>
<td>3.06</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>5.77</td>
<td>1.42</td>
</tr>
<tr>
<td>17</td>
<td>$\frac{1}{2}$</td>
<td>2.06</td>
<td>-2.15</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>3.93</td>
<td>2.19</td>
</tr>
</tbody>
</table>

is not as trivial. Individual 6 has the highest value for $\tau_A$ of any individual but is not deemed unusual by Cook’s distance, in fact it does not even correspond to the individual with the highest value for Cook’s distance. Individual 6 contains two observations which are extremely outlying both marginally and conditionally. The other two unusual individuals, indexed by 11 and 17, contain at least one observation which is considerably outlying marginally and not necessarily in the conditional sense, see Table 5.4. Each of these individuals is classified as outlying due to the contributions of a few of its observations and these unusual observations can be found in Table 5.3.

It is shown in Appendix A.3.1 that for repeated measures where individuals are assumed to be independent and the set $A$ indexes an individual then $\phi_A = \frac{n-p}{I}$ where $I$ is the number of individuals and $p$ is the rank of
Figure 5.6: $\phi_A$ and $\gamma_A$ associated with an Individual

the design matrix. Figure 5.6 is a plot of $\gamma_A$ and $\phi_A$ associated with an individual against subset size. It is clear from this plot that $\gamma_A \geq \phi_A$. This provides further evidence that $\gamma_A \geq \phi_A$ irrespective of the subset $A$. When $A$ indexes all the observations of an individual then Figure 5.6 shows that for this model $\gamma_A = \phi_A$ however this remains to be theoretically confirmed.

When $\gamma_A = \phi_A$ then $\tau_A$ simplifies down to the approximation $\tau^*_A = \frac{r_A T_A}{\phi_A}$. For this model $\phi_A = \frac{n-p}{T}$ and provided $n \gg p$ then $\frac{r_A}{\phi_A} \approx 1$. Figure 5.7 is a plot of the statistics $T_A$, $\tau^*_A$ and $\tau_A$ associated with the 1st patient. Visually very little difference can be seen between the various statistics. The theoretical developments found in chapter 3 explain this similarity since $\gamma_A \approx \phi_A$ and $\phi_A \approx \kappa_A$. 

94
5.4 Conclusions

The examples presented in this chapter give a favourable indication of the contribution methodology as a diagnostic approach for detecting unusual behaviour in repeated measures. The ability of contribution statistics to detect both outlying observations and individuals provides evidence that this methodology can be successfully applied to repeated measures. All the examples presented in this chapter have observations which are equally spaced, as a result issues concerning leverage do not arise. For such models $T_A$ is a good approximation for $\tau_A$.

Applying the contribution methodology to irregular spaced observations appears a natural area of further research. The success of the contribution
diagnostics to detect outlying observations with high leverage, which was exam­ined in chapter 2, is a source of encouragement. The statistic $\tau^*_A$ should also prove valuable as a computationally efficient diagnostic for such models. Methods for fitting Laird-Ware models with serially correlated errors to irregular spaced observations can be found in Jones and Ackerson [32]. Extensions of these approaches are considered in Reeves and MacKenzie [47] however neither Jones and Ackerson [32] or Reeves and MacKenzie [47] performed any diagnostic analysis. It is worth noting that multilevel analysis of repeated measures, as in Goldstein [24], does not require the level 1 units (the observations) to be regularly spaced.
Chapter 6

Alternative Methodology

6.1 Dual Residuals

The previous chapters in this thesis are concerned with theoretically extending a diagnostic proposed for general linear models and applying this diagnostic with its extensions to time series. This statistic is defined in chapter 2 and is composed of two residuals known as marginal and conditional residuals. Haslett and Hayes [29] claim that these residuals have a dual role in establishing departures from the assumed model. The contribution statistic and its standardized form is viewed as one way of encapsulating the information contained in both residuals. An alternative method of combining both residuals is in the form of a Mahalanobis distance, see Krzanowski [36] section 8.2. This new distance is defined in the next section and empirical evidence
is provided of its ability to detect outlying observations in time series.

6.2 Mahalanobis Distance

The Mahalanobis distance of \((\hat{e}_i, \hat{e}(i))\) from zero, denoted \(M_i\), is defined as follows

\[
M_i = \left( \begin{array}{c} \hat{e}_i \\ \hat{e}(i) \end{array} \right)^T \left( \begin{array}{cc} g_{ii} & d_i \phi_i \\ d_i \phi_i & d_i \end{array} \right)^{-1} \left( \begin{array}{c} \hat{e}_i \\ \hat{e}(i) \end{array} \right)
\]

Clearly \(M_i\) is related to the contribution approach as \(M_i\) depends on the contribution \(T_i\) however there appears to be no simple interpretation of this relationship. \(M_i\) is approximately \(\chi^2\) distributed and it is always the case that \(M_i \geq 0\). Attention is restricted to observations which generate values of \(M_i\) such that their corresponding p values are less than 0.05. A general version of \(M_i\) for arbitrary subsets, denoted \(M_A\), can be found in appendix B. It follows that \(M_A \geq 0\) and that \(M_A\) is approximately \(\chi^2_{2\kappa_A}\) distributed.

In chapter 2 it is shown when \(Y \sim N(X \beta, \sigma^2 I)\) then the standardized contributions, \(\tau_i\), are studentized residuals to be squared. In this situation the vector \((\hat{e}_i, \hat{e}(i))\) can be expressed as \((\hat{e}_i, \frac{\hat{e}_i}{1 - h_{ii}})\) where \(h_{ii}\) denotes the lever-
age of the $i^{th}$ observation. The covariance matrix of this vector is singular. However $e_i g_{ii}^{-1} \hat{e}_i = \frac{e_i^2}{\sigma^2 (1 - h_{ii})} = z_{(i)}^2$ and $\hat{e}_{(i)} d_{i}^{-1} \hat{e}_{(i)} = \frac{\hat{e}_{(i)}^2}{\sigma^2 (1 - h_{ii})} = z_{(i)}^2$. Thus when observations are assumed to be independent the Mahalanobis distance of $\hat{e}_i$ or $\hat{e}_{(i)}$ from 0 corresponds to the square of the studentized residual associated with the $i^{th}$ observation. As a result there appears to be no reason to combine both marginal and conditional residuals in the form of the $M_i$ statistic for models where $Y \sim N(X\beta, \sigma^2 I)$ such as those associated with Huber's data and Forbes' data.

6.3 Additive Property

The additive property of the contributions $T_A = \sum_{i \in A} T_i$ makes then computationally more appealing to work with than standardized contributions since $\tau_A \neq \sum_{i \in A} \tau_i$. Chapter 3 establishes situations in which $T_A \approx \tau_A$ and how to get cheap approximations to $\tau_A$ when $T_A$ is not a good approximation. The statistic $M_A \neq \sum_{i \in A} M_i$ but understanding the relationship between $M_A$ and $\sum_{i \in A} M_i$ maybe fruitful from a computational viewpoint. This relationship is not explored as the goal of this chapter is to establish that this new statistic can detect unusual observations experienced by time series; computational issues can be addressed by further analysis.
6.4 Application of $M_i$ Statistic

The examples presented in chapter 4 are revisited; the same models are assumed for these data sets as those previously proposed. These examples provide empirical evidence of the ability of the $M_i$ statistic as a diagnostic tool for time series. This diagnostic successfully detects both single outliers and groups of outlying observations.

6.4.1 Single Outliers

The first example presented in chapter 4 is a stationary equally spaced autoregressive process of order one which has a single additive outlier present. From Table 6.1 it is clear that the statistic $M_i$ detects the additive outlier present at the point indexed by 28. It is also evident due to the false detection of observation 27 as outlying that the $M_i$ diagnostic suffers from smearing. Smearing is also a problem for both the contribution statistic $\tau_A$ and the diagnostic $DV(A)$.

The next example analyzed is a stationary equally spaced autoregressive process of order one which has a single innovation outlier present. From Table 6.2 it is clear that the $M_i$ correctly identifies the outlier at observation 28 and numerous other outlying points after this observation. Again $M_i$ falsely detects observation 27 as outlying.
Table 6.1: \( M_i \) Analysis of a Single Additive Outlier

<table>
<thead>
<tr>
<th>Index</th>
<th>( y_i )</th>
<th>( M_i )</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>1.12</td>
<td>1.13</td>
<td>0.57</td>
</tr>
<tr>
<td>27</td>
<td>-0.46</td>
<td>11.22</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>3.37</td>
<td>13.51</td>
<td>0.00</td>
</tr>
<tr>
<td>29</td>
<td>-0.10</td>
<td>2.76</td>
<td>0.25</td>
</tr>
<tr>
<td>30</td>
<td>-1.05</td>
<td>3.42</td>
<td>0.18</td>
</tr>
</tbody>
</table>

These examples illustrate the ability of the \( M_i \) statistic to detect single outlying points. Both additive and innovation outliers are successfully detected. It is also apparent that this diagnostic suffers from smearing. It is worth noting that this diagnostic detects these unusual observations without re-estimating variance or correlation parameters so is computationally cheaper than the diagnostic \( DV(A) \).

6.4.2 Groups of Outliers

The third example presented in chapter 4 is of a first order moving average process with a contiguous block of three additive outliers which are indexed by 60, 61 and 62. This patch of outliers is identified by examining the value \( M_i \) takes for each singly observation, see Table 6.3. Just like the contribution
Table 6.2: $M_i$ Analysis of a Single Innovation Outlier

<table>
<thead>
<tr>
<th>Index</th>
<th>$y_i$</th>
<th>$M_i$</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>-1.09</td>
<td>1.85</td>
<td>0.40</td>
</tr>
<tr>
<td>27</td>
<td>-1.48</td>
<td>18.88</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>5.81</td>
<td><strong>19.17</strong></td>
<td>0.00</td>
</tr>
<tr>
<td>29</td>
<td>5.40</td>
<td>6.36</td>
<td>0.04</td>
</tr>
<tr>
<td>30</td>
<td>3.23</td>
<td>7.48</td>
<td>0.02</td>
</tr>
<tr>
<td>31</td>
<td>3.80</td>
<td>4.46</td>
<td>0.11</td>
</tr>
<tr>
<td>32</td>
<td>5.02</td>
<td>7.24</td>
<td>0.03</td>
</tr>
<tr>
<td>33</td>
<td>2.35</td>
<td>5.20</td>
<td>0.07</td>
</tr>
<tr>
<td>34</td>
<td>2.25</td>
<td>1.11</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Due to the nature of the $M_i$ statistic a detection of a contiguous block of outliers is not possible. The Nile level shifts the Nile data provides an appropriate data set. Examination of observations singly reveals three data points which generate a value for $M_i$ which has a p-value less than the critical value of 0.05, see Table 6.4. The contribution statistic $y_i$ in chapter 4 as well as the statistic proposed in de Jong and Peazer [18].
Table 6.3: $M_i$ Analysis of a Patch of Additive Outliers

<table>
<thead>
<tr>
<th>Index</th>
<th>$y_i$</th>
<th>$M_i$</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>0.30</td>
<td>1.38</td>
<td>0.50</td>
</tr>
<tr>
<td>59</td>
<td>-1.14</td>
<td>14.80</td>
<td>0.00</td>
</tr>
<tr>
<td>60</td>
<td>3.05</td>
<td>35.49</td>
<td>0.00</td>
</tr>
<tr>
<td>61</td>
<td>4.73</td>
<td>66.22</td>
<td>0.00</td>
</tr>
<tr>
<td>62</td>
<td>5.89</td>
<td>48.73</td>
<td>0.00</td>
</tr>
<tr>
<td>63</td>
<td>-1.33</td>
<td>19.00</td>
<td>0.00</td>
</tr>
<tr>
<td>64</td>
<td>0.02</td>
<td>0.14</td>
<td>0.93</td>
</tr>
</tbody>
</table>

approach it is unnecessary to examine subsets of observations in order to detect these unusual points. The adjacent observations at either end of the patch of outliers which are indexed by 59 and 63 are falsely detected, see Table 6.3, again smearing is a concern.

Due to the success of the $M_i$ statistic at detecting a contiguous block of outliers a natural step is to test its ability at detecting level shifts; the Nile data provides an appropriate data set. Examination of observations singly reveals three data points which generate a value for $M_i$ which has a p value less than the critical value of 0.05, see Table 6.4. The contribution statistic $\tau_i$ in chapter 4 as well as the statistic proposed in de Jong and Penzer [18]
identified the first two observations in Table 6.4 as outlying.

Examining the values for $M_i$ generated by single observations gives no indication of a level shift. Thus $M_A$ was generated for sets containing 2, 3, 4 and 5 observations but this failed to identify a level shift. Then $M_A$ was calculated for contiguous blocks of observations such that $A$ indexes a block of observations starting at the final observation (indexed by the year 1970) and running backwards. Unfortunately when the set $A$ indexes more then ten observations numerical difficulties occur due to the covariance matrix of $(\hat{e}_A, \hat{e}_{(A)})$ approaching singularity. As a result of these numerical difficulties using $M_A$ to detect the level shift was abandoned. Since the covariance matrix of $(\hat{e}_A, \hat{e}_{(A)})$ rapidly tends to singularity this implies that $\hat{e}_A \approx \hat{e}_{(A)}$ when $A$ indexes more then ten observations. As a result the diagnostic $M_{(A)} = \hat{e}_{(A)}^T D_A^{-1} \hat{e}_{(A)}$ which is approximation $\chi^2_{\Delta A}$ distribution is proposed instead of $M_A$. Alternatively a diagnostic could be generated using $\hat{e}_A$ and

<table>
<thead>
<tr>
<th>Year</th>
<th>Volume Flow</th>
<th>$M_i$</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1877</td>
<td>813</td>
<td>6.31</td>
<td>0.04</td>
</tr>
<tr>
<td>1913</td>
<td>456</td>
<td>10.13</td>
<td>0.01</td>
</tr>
<tr>
<td>1916</td>
<td>1120</td>
<td>6.21</td>
<td>0.04</td>
</tr>
</tbody>
</table>
should roughly contain the same information since for this example $\hat{e}_A \approx \hat{e}(A)$ when $A$ indexes a block of ten observations or more. $M(A)$ was calculated for contiguous blocks of observations such that $A$ indexes a block of observations starting at the final observation (indexed by the year 1970) and running backwards. The p values associated with $M(A)$ are displayed in Figure 6.1: for example the p value indexed by the year 1940 corresponds to the p value of $M(A)$ where $A$ indexes the set of observations $\{y_{1940}, \ldots, y_{1970}\}$. The p values of $M(A)$ indicate that the subset of observations stretching inclusively from 1917 to 1970 are outlying. Due to concerns that $M(A)$ may be susceptible to
masking the observation at the year 1913 was removed and the Nile data was re-analyzed. Again $M_{(A)}$ indicates that the set of observations inclusively from 1917 to 1970 are outlying; any larger contiguous block of observations then \{y_{1917}, \ldots, y_{1970}\} is not detected by $M_{(A)}$ as outlying. This result puts the level shift occurring at the year 1917 but the diagnostic proposed in de Jong [18] detects the level shift at the year 1899 and the contribution approach places the level shift at the year 1900. Kraus [35] notes that in 1898 "... changes of the rainfall regime can be observed in the records of most tropical stations at places as far apart as Northern Australia and Central America". This lends weight to the belief that the volume flow rate of the Nile experiences a level shift around 1898.

### 6.5 Conclusions

These examples illustrate the potential of the $M_i$ statistic as a diagnostic tool for time series. This statistic successfully detects a variety of unusual behaviour. However using this approach to detect outlying subsets is computationally expensive due to the inverting of covariance matrices. Numerical difficulties encountered due to singular matrices also makes this approach less appealing then the contribution methodology.
Chapter 7

Further Research

7.1 Possible Theoretical Extensions

As a result of the theoretical develops in Chapters 2 and 3, two possible areas of further theoretical research arise. Both of these research topics involve the generalization of existing proofs.

In Chapter 2, bounds were developed for examining suspiciously large negative contributions. However they are only valid for subsets of size two, see section 2.3. A more general proof for arbitrary subsets would be good for completeness.

In Chapter 3, the relationship between the parameters $\phi_A$ and $\gamma_A$ is explored. It is conjectured that for arbitrary subsets $\phi_A \geq \gamma_A$ for models where $Y \sim N(X\beta, V)$ such that the fixed effects, $\beta$, are unknown and the covari-
ance matrix, $V$, is non-diagonal. If this conjecture is proven to be true this would enable the use of $T_A$ as a computationally cheap bound for $\tau_A$ for a general class of models, recall section 3.6, page 49.

### 7.2 Conclusions

This thesis has shown theoretically and empirically the ability of the contribution methodology as a diagnostic tool for detecting observations and groups of observations which deviate from a wide range of models. The extensions proposed in this chapter would particularly encourage the use of the contribution diagnostics for detecting outlying subsets.
Appendix A

Additional Proofs

A.1 Propositions used in Chapter 2

Proposition A.1.1 The function

\[ g(x) = P\left(-\frac{\gamma_A - \phi_A}{2\kappa_A} v \leq -t - \frac{\gamma_A + \phi_A}{2\kappa_A} x \right) \]

where \( v \sim \chi^2_{\kappa_A} \) and \( x \sim \chi^2_{\kappa_A} \) is convex when \( \kappa_A = 1 \) or \( \kappa_A = 2 \).

Proof In this case

\[
\begin{align*}
g(x) &= P\left(-\frac{\gamma_A - \phi_A}{2\kappa_A} v \leq -t - \frac{\gamma_A + \phi_A}{2\kappa_A} x \right) \\
&= P\left(\frac{\gamma_A - \phi_A}{2\kappa_A} v \geq t + \frac{\gamma_A + \phi_A}{2\kappa_A} x \right) \\
&= P\left(v \geq \frac{2t\kappa_A}{\gamma_A - \phi_A} + \frac{\gamma_A + \phi_A}{\gamma_A - \phi_A} x \right) \\
&= 1 - P\left(v \leq \frac{2t\kappa_A}{\gamma_A - \phi_A} + \frac{\gamma_A + \phi_A}{\gamma_A - \phi_A} x \right)
\end{align*}
\]
hence

\[
\frac{dg(x)}{dx} = -\frac{\gamma_A + \phi_A}{\gamma_A - \phi_A} f\left(\frac{2t\kappa_A}{\gamma_A - \phi_A} + \frac{\gamma_A + \phi_A}{\gamma_A - \phi_A} x\right)
\]

where \(f(.)\) denotes the probability density function and

\[
\frac{d^2g(x)}{dx^2} = -\left(\frac{\gamma_A + \phi_A}{\gamma_A - \phi_A}\right)^2 \frac{d}{dx} f\left(\frac{2t\kappa_A}{\gamma_A - \phi_A} + \frac{\gamma_A + \phi_A}{\gamma_A - \phi_A} x\right).
\]

If \(\frac{df(x)}{dx} < 0\) then \(\frac{d^2g(x)}{dx^2} > 0\).

By definition if \(\frac{d^2g(x)}{dx^2} > 0\) \(\forall x \in (0, \infty)\) then \(g(x)\) is a convex function on the range \(x \in (0, \infty)\), see Finney et al. [22] section 3.3. Thus if \(\frac{df(x)}{dx} < 0\) \(\forall x \in (0, \infty)\) then \(g(x)\) is a convex function on the range \(x \in (0, \infty)\).

The p.d.f of a \(\chi^2\) distribution with \(\kappa_A\) degrees of freedom is given by

\[
f(x) = \frac{x^{(\kappa_A-2)/2}e^{-(x/2)}}{2^{\kappa_A/2}\Gamma(\kappa_A/2)} \quad x \geq 0.
\]

It follows that \(\frac{df(x)}{dx} < 0\) \(\forall x \in (0, \infty)\) if \(\frac{d}{dx}(x^{(\kappa_A-2)/2}e^{-(x/2)}) < 0\) \(\forall x \in (0, \infty)\).

The case when \(\kappa_A = 1\) then

\[
\frac{d}{dx}(x^{(\kappa_A-2)/2}e^{-(x/2)}) = \frac{d}{dx}(x^{-1/2}e^{-x/2}) = \frac{-e^{-x/2}}{2}(x^{-1/2} + x^{-3/2}) < 0 \quad \forall x \in (0, \infty).
\]

Thus \(g(x)\) is a convex function on the range \(x \in (0, \infty)\) when \(\kappa_A = 1\). The case when \(\kappa_A = 2\) then

\[
\frac{d}{dx}(x^{(\kappa_A-2)/2}e^{-(x/2)}) = \frac{d}{dx}(e^{-x/2})
\]
Thus \( g(x) \) is a convex function on the range \( x \in (0, \infty) \) when \( k = 2 \).

Proposition A.1.2 Let \( Y \sim N(X\beta, V) \) and \( Q = V^{-1} - V^{-1}X(X'\sqrt{V} X)^{-1}X'\sqrt{V} \) then the matrix

\[
F = (VQV)^{\frac{1}{2}}Q(VQV)^{\frac{1}{2}}
\]

is idempotent.

Proof

\[
F^2 = (VQV)^{\frac{1}{2}}Q(VQV)^{\frac{1}{2}}(VQV)^{\frac{1}{2}}Q(VQV)^{\frac{1}{2}}
\]

\[
= (VQV)^{\frac{1}{2}}QVQVQ(VQV)^{\frac{1}{2}}
\]

but

\[
QVQ = Q
\]

hence

\[
F^2 = (VQV)^{\frac{1}{2}}QVQ(VQV)^{\frac{1}{2}}
\]

\[
= (VQV)^{\frac{1}{2}}Q(VQV)^{\frac{1}{2}}
\]

\[
= F.
\]

It immediately follows that \( F \) is idempotent.
Proposition A.1.3 Given a set of data $Y \sim N(X\beta, \sigma^2 I)$ then the standardized contributions $\tau_i$ of a single observation, $i$, is given by

$$\tau_i = \frac{\hat{e}_i^2}{\delta^2(1 - h_{ii})}.$$ 

Thus $\tau_i$ is an internally studentized residual squared.

Proof The standardized contribution of $T_i$ is given by

$$\tau_i = \frac{(\hat{z}_i\bar{z}_{(i)} - \varphi_i)\sqrt{2}}{\sqrt{1 + \varphi_i^2}} + 1$$

where $\varphi_i = (VQ)_{ii}/u_i$ and $u_i = (g_{ii} q_{ii})^{\frac{1}{2}}$. $G = VQV$ which for the case when $V = \sigma^2 I$ can be expressed as $G = \sigma^4 Q$ hence $g_{ii} = \sigma^4 q_{ii}$. It follows that $u_i = (\sigma^4 q_{ii} q_{ii})^{\frac{1}{2}} = \sigma^2 q_{ii}$. Thus $\varphi_i = \sigma^2 q_{ii}/\sigma^2 q_{ii} = 1$. Thus

$$\tau_i = \frac{(\hat{z}_i\bar{z}_{(i)} - 1)\sqrt{2}}{\sqrt{1 + 1}} + 1$$

$$= \hat{z}_i\bar{z}_{(i)}$$

$$= \frac{\hat{e}_i}{\sqrt{g_{ii}}} \frac{\bar{e}_{(i)}}{\sqrt{d_i}}$$

$$= \frac{\hat{e}_i}{\sqrt{g_{ii}}} \frac{\sqrt{q_{ii}}}{\sqrt{q_{ii}}} \bar{e}_{(i)}$$

$$= \frac{\hat{e}_i}{\sqrt{\sigma^4 q_{ii}}} \bar{e}_{(i)}$$

$$= \frac{\hat{e}_i}{\sigma^2} \bar{e}_{(i)}$$

$$= \frac{1}{\sigma^2} \bar{e}_{(i)}$$

$$= \frac{\hat{e}_i}{\sigma^2} q_{ii}^{-1} \sigma^{-2} \hat{e}_i$$

$$= \frac{\hat{e}_i^2}{\sigma^4 q_{ii}}$$

$$= \frac{\hat{e}_i^2}{\sigma^2(1 - h_{ii})}$$
\[
\hat{e}_i \quad \text{is an internally studentized residual.} \]

**Proposition A.1.4** Given a set of data \( Y \sim N(X\beta, \sigma^2 I) \) then the standardized contributions \( \tau_A \) of a set of observations, indexed by the set \( A \), is given by

\[
\tau_A = \frac{\kappa_A \sum_{i \in A} \hat{e}_i^2}{\sum_{i \in A} \hat{\sigma}^2(1 - h_{ii})}.
\]

**Proof** The standardized version of \( T_A \) is given by

\[
\tau_A = \kappa_A \left\{ \frac{(T_A - \phi_A)\sqrt{2}}{\sqrt{\gamma_A^2 + \phi_A^2}} + 1 \right\}.
\]

Since \( V = \sigma^2 I \) then \( \phi_A = \gamma_A^2 \). It follows that

\[
\tau_A = \kappa_A \left\{ \frac{T_A}{\phi_A} \right\} \frac{\kappa_A \sum_{i \in A} \hat{e}_i^2}{\sum_{i \in A} \hat{\sigma}^2(1 - h_{ii})}.
\]
A.2 Compound Symmetry and $\gamma_A$

Proposition A.2.1 Given a set of observations $Y \sim N(0, V)$ where

$$V = \begin{pmatrix}
1 & \rho & \rho & \cdots & \rho \\
\rho & 1 & \rho & \cdots & \rho \\
\rho & \rho & 1 & \cdots & \rho \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \rho & \cdots & 1
\end{pmatrix}$$

then $\forall i \in \{1, \ldots, n\}$

$$\gamma_i = \sqrt{1 + \frac{\rho^2(n - 1)}{(1 - \rho)(n\rho - \rho + 1)}}.$$

Proof By definition $V_{ii} = 1$ hence $V_{ii}^{-\frac{1}{2}} = 1$. Let $V_{ii} = (\rho, \ldots, \rho)$ and $V_{ii} = (\rho, \ldots, \rho)^T$. By definition $V_{ii} = V_{ii} - V_{ii} V_{ii}^{-1} V_{ii}$ thus

$$V_{ii} = \begin{pmatrix}
1 & \rho & \cdots & \rho \\
\rho & 1 & \cdots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \cdots & 1
\end{pmatrix} - \begin{pmatrix}
\rho \\
\rho \\
\vdots \\
\rho
\end{pmatrix} \left( \begin{pmatrix}
\rho \\
\rho \\
\vdots \\
\rho
\end{pmatrix} \right)^T = \begin{pmatrix}
1 & \rho & \cdots & \rho \\
\rho & 1 & \cdots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \cdots & 1
\end{pmatrix} - \begin{pmatrix}
\rho^2 & \rho^2 & \cdots & \rho^2 \\
\rho^2 & \rho^2 & \cdots & \rho^2 \\
\vdots & \vdots & \ddots & \vdots \\
\rho^2 & \rho^2 & \cdots & \rho^2
\end{pmatrix}.$$
The inverse of $V_{ij}$, see Jones [33] section 1.5, is defined as

$$V_{ij}^{-1} = \begin{pmatrix}
\frac{1}{1-\rho} - \frac{\rho}{(1-\rho)(np-\rho+1)} & -\frac{\rho}{(1-\rho)(np-\rho+1)} & \cdots & -\frac{\rho}{(1-\rho)(np-\rho+1)} \\
-\frac{\rho}{(1-\rho)(np-\rho+1)} & \frac{1}{1-\rho} - \frac{\rho}{(1-\rho)(np-\rho+1)} & \cdots & -\frac{\rho}{(1-\rho)(np-\rho+1)} \\
\vdots & \vdots & \ddots & \vdots \\
-\frac{\rho}{(1-\rho)(np-\rho+1)} & -\frac{\rho}{(1-\rho)(np-\rho+1)} & \cdots & \frac{1}{1-\rho} - \frac{\rho}{(1-\rho)(np-\rho+1)}
\end{pmatrix}$$

hence

$$V_{ii} V_{ij}^{-1} V_{ii} = \rho^2 \left( \frac{n-1}{1-\rho} - \frac{\rho(n-1)^2}{(1-\rho)(np-\rho+1)} \right)$$

and

$$V_{ii}^{-\frac{1}{2}} V_{ii} V_{ij}^{-1} V_{ii} V_{ii}^{-\frac{1}{2}} = \rho^2 \left( \frac{n-1}{1-\rho} - \frac{\rho(n-1)^2}{(1-\rho)(np-\rho+1)} \right).$$

Let $F = 1 + V_{ii}^{-\frac{1}{2}} V_{ii} V_{ij}^{-1} V_{ii} V_{ii}^{-\frac{1}{2}}$ then

$$F = 1 + \rho^2 \left( \frac{n-1}{1-\rho} - \frac{\rho(n-1)^2}{(1-\rho)(np-\rho+1)} \right) = 1 + \frac{\rho^2(n-1)}{(1-\rho)(np-\rho+1)}.$$

From proposition 3.4.1 $\gamma_i = \sqrt{F}$ hence

$$\gamma_i = \sqrt{\frac{\rho^2(n-1)}{(1-\rho)(np-\rho+1)}}.$$
Note that $\gamma_i$ depends on both $n$ and $\rho$ hence

$$\lim_{\rho \to +1} \gamma_i = \infty$$

and

$$\lim_{n \to \infty} \frac{\rho^2(n-1)}{(1-\rho)(n\rho-\rho+1)} = \frac{\rho}{1-\rho} \lim_{n \to \infty} \frac{n-1}{n-1+\frac{1}{\rho}} = \frac{\rho}{1-\rho}$$

thus

$$\lim_{n \to \infty} \gamma_i = \frac{1}{\sqrt{1-\rho}}.$$

### A.3 $\phi_A$ for Experimental Units

**Proposition A.3.1** Consider a repeated measures study where the experimental units are assumed to be independent and identically distributed. Each experimental unit has the same number of measures taken at the same time points. It is assumed $Y \sim N(X\beta, V)$ where $V$ is block diagonal such that each block corresponds to an experimental unit. The design matrix $X$ has a rank of $p$. Then

$$\phi_{A_i} = \frac{n-p}{n_e}$$

where $A_i$ indexes the set of observations corresponding to the $i^{th}$ experimental unit and $n_e$ is number of observations indexed by $A_i$. $n$ denotes the number of observations in the entire data set.
Proof

\[ \phi. = tr(VQ) \]

\[ = \sum_{i \in e} tr((VQ)_{A_i A_i}) \]

\[ = n_e tr((VQ)_{A_i A_i}) \]

where \( e \) indexes the experimental units. Proposition 3.3.1 establishes that \( \phi. = n - p \) hence

\[ \phi_{A_i} = tr((VQ)_{A_i A_i}) = \frac{n - p}{n_e}. \]

The mahalanobis distance of \((\hat{e}_A, \hat{e}_{(A)})\) from zero is defined as follows.

\[ M_A = \begin{pmatrix} \hat{e}_A \\ \hat{e}_{(A)} \end{pmatrix}^T \begin{pmatrix} G_A & (VQ)_{A_A} D_A \\ D_A^T (VQ)_{A_A} & D_A \end{pmatrix}^{-1} \begin{pmatrix} \hat{e}_A \\ \hat{e}_{(A)} \end{pmatrix} \]

\( M_A \) is approximately \( \chi^2_{k_A} \) distribution. Unfortunately \( M_A \) can not be easily expressed in terms of contributions.

B.1 Covariance of \( \hat{e}_A \) and \( \hat{e}_{(A)} \)

The following Lemma can be found in Hastett and Hayes [29].

Lemma B.1.1. Partitioned conditional residual \( \hat{e}_{(A)} \). Let \( P \) be a partition of the labels \( (1, \ldots, n) \) into \( r \) blocks \( A_1, A_2, \ldots, A_r \) of size \( \kappa_1, \kappa_2, \ldots, \kappa_r \) such that \( n = \sum \kappa_i \) and \( \kappa_i \leq n - p \) for all \( i \). Let \( \hat{e}_{(A)} \) be the \( \kappa_i \)-dimensional conditional
Appendix B

General Form of $M_A$

The mahalanobis distance of $(\hat{e}_A, \bar{e}_{(A)})$ from zero is defined as follows

$$M_A = \begin{pmatrix} \hat{e}_A \\ \bar{e}_{(A)} \end{pmatrix}^T \begin{pmatrix} \begin{pmatrix} G_{AA} & (VQ)_{AA}D_A \\ D_A(VQ)_{AA}^T & D_A \end{pmatrix}^{-1} & \end{pmatrix} \begin{pmatrix} \hat{e}_A \\ \bar{e}_{(A)} \end{pmatrix}.$$

$M_A$ is approximately $\chi^2_{2\kappa_A}$ distribution. Unfortunately $M_A$ can not be easily expressed in terms of contributions.

B.1 Covariance of $\hat{e}_A$ and $\bar{e}_{(A)}$

The following Lemma can be found in Haslett and Hayes [29].

Lemma B.1.1 Partitioned conditional residual $\hat{e}_{(P)}$. Let $P$ be a partition of the labels $(1,\ldots,n)$ into $r$ blocks $A_1,A_2,\ldots,A_r$ of size $\kappa_1,\kappa_2,\ldots,\kappa_r$ such that $n = \sum_i \kappa_i$ and $\kappa_i \leq n - p$, all $i$. Let $\hat{e}_{(P)}$ be the $\kappa_i$-dimensional conditional
residual. Define \( \hat{e}_{(P)} \) to be the (stacked) vector based on \( (\hat{e}_{(A_1)}, \ldots, \hat{e}_{(A_r)}) \).

Then

\[
D_P^{-1} \hat{e}_{(A)} = V^{-1} \hat{e}
\]

where \( D_P \) is block diagonal with blocks \( (D_{A_1}, \ldots, D_{A_r}). \)

From this Lemma, proposition B.1.1 establishes the covariance of \( (\hat{e}_A, \hat{e}_{(A)}) \).

**Proposition B.1.1**

\[
\text{Cov}(\hat{e}_A, \hat{e}_{(A)}) = (VQ)_{AA} D_A
\]

**Proof** Let \( P \) be a partition such that \( P = (A, \overline{A}) \) where \( \overline{A} \) is the complement of \( A \).

\[
\text{Cov}(\hat{e}, \hat{e}_{(P)}) = \text{Cov}(\hat{e}, D_P V^{-1} \hat{e})
\]

\[
= E[\hat{e}\hat{e}^T V^{-1} D_P]
\]

\[
= E[\hat{e}\hat{e}^T] V^{-1} D_P
\]

\[
= VQVV^{-1} D_P
\]

\[
= VQ D_P.
\]

It follows that

\[
\text{Cov}(\hat{e}_A, \hat{e}_{(A)}) = (VQ D_P)_{AA}
\]

\[
= (VQ)_{AA} D_A.
\]

When \( A \) indexes a single observation, \( A = i \), then

\[
\text{Cov}(\hat{e}_i, \hat{e}_{(i)}) = (VQ)_{ii} d_i = \phi_i d_i.
\]
Bibliography


