DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

JACKKNIFE-AFTER-BOOTSTRAP METHOD AS DIAGNOSTIC TOOL IN GENERALIZED LINEAR MODELS

by
Ufuk BEYAZTAŞ

July, 2016
İZMİR

JACKKNIFE-AFTER-BOOTSTRAP METHOD AS DIAGNOSTIC TOOL IN GENERALIZED LINEAR MODELS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Statistics

by
Ufuk BEYAZTAŞ

July, 2016
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"JACKKNIFE-AFTER-BOOTSTRAP METHOD AS DIAGNOSTIC TOOL IN GENERALIZED LINEAR MODELS"** completed by **UFUK BEYAZTAŞ** under supervision of **PROF. DR. AYLİN ALIN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

.....

Prof. Dr. Aylin ALIN

Supervisor

.....

Assoc. Prof. Dr. A. Fırat ÖZDEMİR

Thesis Committee Member

.....

Prof. Dr. Ali Kemal ŞEHİRLİOĞLU

Thesis Committee Member

.....

Prof. Dr. Meral Çetin

Examining Committee Member

.....

Doç. Dr. Hakan Savaş SAZAK

Examining Committee Member

Prof. Dr. Ayşe OKUR

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

The words alone cannot be express the thank to my supervisor Prof. Dr. Aylin ALIN, for her direction, motivation, assistance and guidance. Her recommendations and suggestions have been invaluable for my dissertation and my academic career, and this dissertation would not have been possible without the support of her. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would like to thank Assist. Prof. Dr. Soutir BANDYOPADHYAY for his support and comments on my works during my visit to Lehigh University in USA. I also wish to thank to my committee members Assoc. Prof. Dr. A. Fırat ÖZDEMİR and Prof. Dr. Ali Kemal ŞEHİRLİOĞLU for their helpful comments which improved my dissertation significantly. My sincere thanks goes to Prof. Dr. Serdar KURT for the continuous support of my Ph.D study and to Tufan ALIN for his help in improving R coding in the simulation study. Besides, I am deeply grateful to Prof. Dr. Michael A. MARTIN for his contributions on my studies. I would also like to thank my computer. I couldn't have done it without you!

I appreciate the financial support of the Scientific and Technological Research Council of TURKEY (TUBITAK) during my researches in USA (Grant no: 1059B141500288).

A special thank goes to my wife, Beste, for her love, support, and taking care of me. Last but not the least, I want to express my deepest love and thanks to my parents for their understanding and endless love, through the duration of my studies.

Ufuk BEYAZTAŞ

JACKKNIFE-AFTER-BOOTSTRAP METHOD AS DIAGNOSTIC TOOL IN GENERALIZED LINEAR MODELS

ABSTRACT

The detection and evaluation of influential observations are critical aspects of data analysis in the context of linear regression models. There are lots of measures proposed to flag these observations. The idea behind these measures is to compare a feature of the model fit obtained from the full data set with the one obtained from reduced set not including corresponding point. Observations whose removal causes major changes in the analysis are termed influential with respect to the feature of the model fit under consideration. The assessment of whether the change in the model resulting from the inclusion/exclusion of each respective data point is major is usually based on cut-off values obtained from asymptotic approximations. However, the asymptotic approximations used for these diagnostic measures suffer both because the null distributions of these quantities are very complex and the approximations tend to be poor when sample sizes are small. Resampling methods, e.g. bootstrap, can be used to overcome these problems.

In this thesis, several resampling based methods are proposed to detect influential observations in linear and binary logistic regression models. Performances of the proposed methods have been compared with the traditional methods for several influence measures by both real world examples and simulation studies. Our results reveal that under a variety of scenarios, our proposed methods provide more accurate and reliable results, and they are more robust to masking effects.

Keywords: Bootstrap, Influential observation, Masking, Regression diagnostics, Robustness, Swamping.

GENELLEŞTİRİLMİŞ DOĞRUSAL MODELLERDE SORUN TANIMLAMA ARACI OLARAK JACKKNIFE DEN SONRA BOOTSTRAP YÖNTEMİ

ÖZ

Doğrusal regresyon modelleri ile yapılan veri analizlerinde etkin gözlemlerin belirlenmesi ve değerlendirilmesi kritik bir öneme sahiptir. Bu gözlemleri belirlemek için bir çok ölçü belirlenmiştir. Bu ölçülerin arkasındaki fikir, tüm gözlemlerin olduğu veri setinden elde edilen model tahmininin bir özelliği ile ilgilenilen gözlemin olmadığı indirgenmiş veri setinden elde edilen özelliğin karşılaştırılmasına dayanır. Çıkartıldığında analiz sonuçları üzerinde büyük değişimlere sebep olan gözlemler etkin gözlem olarak adlandırılır. Her bir gözlemin modele eklenmesi veya çıkartılmasından kaynaklanan değişimin büyüklüğünün değerlendirilmesi, asimptotik yaklaşımlardan elde edilen eşik değerlere dayanır. Fakat, bu ölçüler için kullanılan asimptotik yaklaşımlar, niceliklerin dağılımlarının kompleks yapıda olmasından ve yaklaşımların küçük örneklem genişliklerinde zayıf olmasından dolayı yetersizdir. Bu problemlerin üstesinden gelmek için bootstrap gibi yeniden örnekleme yöntemleri kullanılabilir.

Bu tezde, doğrusal ve ikili lojistik regresyon modellerinde etkin gözlemlerin belirlenmesi için yeniden örneklemeye dayalı çeşitli yöntemler önerilmiştir. Önerilen yöntemlerin performansları gerçek dünya verileri ve simülasyon çalışmaları aracılığı ile çeşitli ölçüler için klasik yöntemler ile karşılaştırılmıştır. Elde ettiğimiz sonuçlar, çeşitli senaryolar altında, önerilen yöntemlerin daha doğru ve güvenilir sonuçlar ürettiğini ve maskeleme etkisine karşı daha dirençli olduğunu göstermektedir. Önerdiğimiz yöntemlerin diğer tüm doğrusal ve doğrusal olmayan regresyon modellerine genişletilebileceği not edilmelidir.

Anahtar kelimeler: Bootstrap, Etkin gözlem, Masking, Regresyon tanıları, Dirençlik, Swamping.

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 Introduction

The detection and evaluation of influential observations and outliers are critical aspect of data analysis. There are lots of measures proposed to flag these observations. The idea behind these measures is to compare a feature of the model fit obtained from the full data set with the one obtained from reduced set not including corresponding point. Observations whose removal causes major changes in the analysis are termed influential with respect to the feature of the model fit under consideration. The assessment of whether the change in the model resulting from the inclusion/exclusion of each respective data point is major is usually based on cut-off values obtained from asymptotic approximations. However, the asymptotic approximations used for these diagnostic measures suffer both because the null distributions of these quantities are very complex and the approximations tend to be poor when sample sizes are small.

As computing power has exploded over the intervening decades, computer-intensive methods such as Efron (1979)'s bootstrap have also been widely used in data analysis. In the present context, bootstrap method is used to estimate both the distribution of the diagnostic measure and the cut-off values. Since bootstrap distribution is obtained by taking random samples from original sample with replacement, it is sensitive to unusual observations. What we need is to have a sampling distribution free from any effect of these observations. Hence, Efron (1992)'s Jackknife-after-Bootstrap (JaB) technique has been proposed by Martin and Roberts (2010) and Beyaztas and Alin (2013) as a method for detecting influence in regression models through refining the cut-off values for the common influence diagnostics used in linear regression models. Later, Beyaztas and Alin (2014a) proposed sufficient JaB which reduces the computational burden about 30%. Also, Beyaztas et al. (2014) proposed a robust version of the JaB method. Moreover, Beyaztas and Alin (2014b) proposed delete-2 JaB method to detect influential

observations when the data have multiple influential data points with masking and swamping effects. Furthermore, Beyaztas and Alin (2014c) used the JaB method to detect influential observations in binary logistic regression model. Following subsections include detailed information about bootstrap and JaB methods.

1.2 The Bootstrap

The bootstrap, which was proposed by Bradley Efron (1979) and further developed by Efron and Tibshirani (1993), is an one of the most important tool of modern statistical analysis. It establishes a general framework for simulation based statistical inference. There are two types of bootstrap methods: parametric and nonparametric. Our interest will be on nonparametric bootstrap. From now, we will simply call it bootstrap. The main goal of the bootstrap method is; to estimate the standard errors, bias and other measures of a statistic, and approximate the sampling distribution by resampling with replacement from the original sample. The most useful references about theory and applications of bootstrap are Efron and Tibshirani (1993), Davison and Hinkley (2005), and Hall (1995). In the bootstrap method, bootstrap resamples of the data are obtained by random sampling with replacement from the original data set, and these resamples are assumed to be independent and identically distributed (i.i.d.).

Let $Y_1,Y_2,...,Y_n$ be the i.i.d. random samples from unknown distribution F with parameter θ. The data $Y_1,Y_2,...,Y_n$ is used to estimate θ; $\hat{\theta} = \hat{\theta}(Y_1, Y_2, ..., Y_n)$. Generally, we are interested in the distribution of $\hat{\theta}$ in order to provide standard errors, to construct confidence intervals, or to perform test of hypothesis. Using random samples taken from a population, we estimate the population parameter θ whereas in the bootstrap context, we try to estimate the parameter of the sampling distribution. That is, our population is now the original sample, and now we estimate the parameter of the sampling distribution $\hat{\theta}$. The general bootstrap idea is given step by step as follows;

- Let Y_1^*, Y_2^*,..., Y_n^* be the generated bootstrap resamples with replacement from the original sample $Y_1, Y_2, ..., Y_n$.
- Let $\hat{\theta}^*$ be the bootstrap estimates of $\hat{\theta}$.
- The first two steps are repeated for B times, say B = 1000, and B values of $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*$ are obtained.

The empirical distribution of $\hat{\theta}^*$ is used to approximate the sampling distribution of $\hat{\theta}$.

1.3 Jackknife-after-Bootstrap

Jackknife-after-Bootstrap method was proposed by Bradley Efron (1992) for estimating the standard errors and bias of a statistic. He described the idea behind the JaB method as follows: a sample of size n from $y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_n$ has the same distribution as a bootstrap sample from $y_1, y_2, ..., y_n$ in which none of the bootstrap values equals y_i. This method requires about e times more resamples than regular bootstrap. For example, for any data set, if we want to determine whether an individual data point is influential or not, and to obtain 1000 resamples without this individual data point, about $1000e \approx 3000$ resamples are required. Then, these 1000 resamples are used to construct the sampling distribution, and to determine the influence cut-offs. The algorithm of JaB method for detection of influential observations proposed by Martin and Roberts (2010) can be described as follows;

- Let θ_i be the diagnostic statistic that we study. The appropriate model is fitted for original data set, and the θ_i for i= 1, 2, …, n are calculated.
- Construct B re-samples with replacement from the original data set.
- For each data point within these B resamples, get a subset of the samples which do not contain that data point, so there are B/e re-samples obtained for each data point. Calculate the n values of θ_i, i = 1, 2, …, n, for each of these resamples, so nB/e values of θ_i are obtained. Collect all nB/e values of θ into a single vector.

- Suitable quantiles (say 2.5% and 97.5%) of this generated bootstrap distribution are determined. Percentiles of this distribution are then compared to the original θ_i, $i = 1, 2, \ldots, n$, values to flag the points as influential or not.

The steps 1-4 are repeated *M* times. Then, the average and standard deviation for the number of flagged points for all these *M* simulations can be calculated. It should be noted that this algorithm runs only once for the real data.

As described by Martin and Roberts (2010), the rationale behind this approach is to generate a "null" bootstrap distribution of θ under the hypothesis that the *ith* data point is not influential. They propose that since the *ith* data point is not present in any of the resamples from which this bootstrap distribution is generated, it cannot exert influence, and thus the distribution generated is free from the influence of this point.

1.4 Influence Measures in Linear Regression Models

The linear regression model used with influence measures throughout this thesis is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1,2,\ldots,n \tag{1.1}$$

This can be written in matrix form as

$$Y = X\beta + \varepsilon \tag{1.2}$$

where, Y is an $n \times 1$ column vector for response variable, X is an $n \times p$ ($p=k+1$) fixed full-rank design matrix, β is an $p \times 1$ vector of unknown parameters including β_0, and ε is an $n \times 1$ error vector with zero mean and unknown variance σ^2. Using the method of least squares with the multiple linear regression model (1.1) we have;

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.3}$$

$$\hat{Y} = X\hat{\beta} = PY \tag{1.4}$$

$$Var(\hat{Y}) = \sigma^2 P \tag{1.5}$$

$$Var(e) = \sigma^2 (I - P) \tag{1.6}$$

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \tag{1.7}$$

$$P = X(X^T X)^{-1} X^T \tag{1.8}$$

$$e = Y - \hat{Y} = (I - P)Y \tag{1.9}$$

$$\hat{\sigma}^2 = \frac{e^T e}{n - p} \tag{1.10}$$

These quantities can be influenced by one or a group of observations, but all observations do not have the same impact over the least square regression outputs. For this reason, identification of influential observations is an important part of regression analysis, and this process is required to make a good inference. To identify influential observations, as mentioned above, several methods have been proposed.

Before examining these methods, we want to determine what is meant by influence. An influential observation is the one which, either individually or together with several observations, has a demonstrably larger impact on the calculated values of various model features than is the case for other observations (Belsley et al. 1980). Existing diagnostic statistics explore the impact of the observations in various way. In general, the influence measures can be classified as follows;

- Measures based on the prediction matrix
- Measures based on the volume of confidence ellipsoids
- Measures based on influence functions, and
- Measures based on partial inference.

The rest of this subsection describes the influence measures used in this thesis. For more information about these measures and other measures, see Chatterjee and Hadi (1986).

1.4.1 t-star Statistic

Generally, the least squared residual for the *ith* observation can be found as;

$$e_i = y_i - x_i \hat{\beta} \tag{1.11}$$

where x_i is the *ith* row of *X*. The standard error for this residual is

$$\sigma_{e_i} = \frac{e_i}{\sigma\sqrt{1-p_i}} \tag{1.12}$$

where p_i is the *ith* diagonal element of *P* given with (2.6). Two special cases of (1.12) are:

$$t_i = \sigma_{e_i} = \frac{e_i}{\hat{\sigma}\sqrt{1-p_i}} \tag{1.13}$$

where $\hat{\sigma}$ is defined in (1.10), and

$$t_i^* \equiv \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-p_i}} \tag{1.14}$$

where

$$\begin{aligned}\hat{\sigma}_{(i)}^2 &= \frac{Y_{(i)}^T(I-P_{(i)})Y_{(i)}}{(n-p-1)} \\ &= \frac{(n-p)\hat{\sigma}^2}{(n-p-1)} - \frac{e_i^2}{(n-p-1)(1-p_i)}\end{aligned} \tag{1.15}$$

So, equivalently the t_i^* statistic can be computed as follows;

$$t_i^* = t_i\sqrt{\frac{(n-p-1)}{(n-p-t_i^2)}} \tag{1.16}$$

This measure is based on residuals with and without *ith* observation, and is distributed approximately *t-distribution* with (*n-p-1*) degrees of freedom. That is the cut-off points for this measure approximately are $t_{\alpha/2,(N-p-1)}$.

1.4.2 The Likelihood Distance

Let $L(\hat{\beta})$ and $L(\hat{\beta}_{(i)})$ be the log likelihood functions at $\hat{\beta}$ and $\hat{\beta}_{(i)}$, respectively. A measure of the influence of the *ith* observation on $\hat{\beta}$ can be based on the distance between $L(\hat{\beta})$ and $L(\hat{\beta}_{(i)})$ (Cahtterjee and Hadi, 1986). The likelihood distance defined by Cook and Weisberg (1982) is

$$\begin{aligned}LD_i &= 2\left|L(\hat{\beta}) - L(\hat{\beta}_{(i)}\right| \\ &= n\log\left[\left(\frac{n}{n-1}\right)\frac{n-p-1}{t_i^{*2}+n-p-1}\right] + \frac{t_i^{*2}(n-1)}{(1-p_i)(n-p-1)} - 1\end{aligned} \tag{1.17}$$

This influence measure is based on the change in volume of confidence ellipsoids with and without the *ith* observation. The likelihood distance is related to the asymptotic confidence region, $\{\ \beta : 2\left[L(\hat{\beta}) - L(\beta) \right] \le \chi^2_{\alpha, p+1}\ \}$ where $\chi^2_{\alpha, p+1}$ is the upper α point of the χ^2 distribution with (*p+1*) degrees of freedom (Chatterjee and Hadi, 1986). Hence, LD_i is compared to χ^2_{p+1}.

1.4.3 Welsch's Distance

Welsch's Distance is based on the idea of influence function introduced by Hampel (1986, 1974) with and without ith observation,

$IF_i(x_i; y_i; F; T)$

$$\lim_{\varepsilon \to \infty} \frac{T\left[(1-\varepsilon)F + \varepsilon \delta_{x_i, y_i} \right] - T\left[F \right]}{\varepsilon} \tag{1.18}$$

where $T(\cdot)$ is a vector-valued statistic, and is based on a random sample from the cumulative distribution function (cdf) of *F*, δ_{x_i, y_i} is the kronecher delta function which takes value of 1 at x_i, y_i and 0 otherwise. IF_i measures the change in T caused by adding x_i, y_i to a very large sample. For a finite sample, several approximations exist including empirical influence curve, the sample influence curve and the sensitivity curve.

Let $\hat{F}$ be the empirical distribution function based on the full sample, and $\hat{F}_{(i)}$ be the empirical distribution function when the *ith* observation is omitted. The empirical influence curve (*EIC*) is

$$\begin{aligned} EIC_i &= (n-1)(X_{(i)}^T X_{(i)})^{-1} x_i^T (y_i - x_i \hat{\beta}_{(i)}) \\ &= (n-1)(X^T X)^{-1} x_i^T \frac{e_i}{(1-p_i)^2} \end{aligned} \tag{1.19}$$

where

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)} \tag{1.20}$$

is the estimate of β when the *ith* observation is omitted. Eg. (1.19) is obtained by replacing $\hat{F}_i$ by F and $T(\hat{F}_i)$. Omitting the limit in (1.18) and taking $F = \hat{F}$, $T(\hat{F}) = \hat{\beta}$, $\varepsilon = -1/(N-1)$ gives the following formula for the sample influence curve.

$$\begin{aligned} SIC_i &= (N-1)(X^T X)^{-1} x_i^T (y_i - x_i \hat{\beta}_{(i)}) \\ &= (N-1)(X^T X)^{-1} x_i^T \frac{e_i}{(1-p_i)} \end{aligned} \tag{1.21}$$

On the other hand, setting $F = \hat{F}_{(i)}$, $T(\hat{F}_{(i)}) = \hat{\beta}_{(i)}$, and $\varepsilon = 1/N$ yields the sensitivity curve (*SC*).

$$SC_i = N(X^T X)^{-1} x_i^T \frac{e_i}{1-p_i} \tag{1.22}$$

To be able to order the observations in a meaningful way, IF_i vector must be normalized. The class of norms which are location/scale invariant is given by

$$D_i(M;c) = \frac{(IF_i)^T M (IF_i)}{c} \tag{1.23}$$

for any appropriate choice of *M* and *c* Chatterjee and Hadi (1986). If $D_i(M;c)$ is large it means that *ith* observation has strong influence on estimated coefficients relative to *M* and *c*. Using (1.19) to approximate (1.18) and setting $M = X_{(i)}^T X_{(i)}$ and $c = (n-1)\hat{\sigma}_{(i)}^2$, (1.23) becomes the Welsch Distance.

$$\begin{aligned} W_i^2 D_i(X_{(i)}^T X_{(i)}; (n-1)\hat{\sigma}_{(i)}^2) \\ = (n-1) t_i^{*2} \frac{p_i}{(1-p_i)^2} \end{aligned} \tag{1.24}$$

Welsch (1982) suggested using W_i as a diagnostic tool and, *n > 15*, using $3\sqrt{p}$ as a cut-off point for W_i. Equivalently

$$W_i = WK_i \sqrt{\frac{n-1}{1-p_i}} \tag{1.25}$$

where $WK_i = |t_i^*| \sqrt{p_i/(1-p_i)}$ also known as $DFFITS_i$.

1.4.4 Modified Cook's Distance

The measure is the modified version of the Cook's Distance proposed by Cook (1977).

$$C_i^* = \sqrt{D_i(X^T X; \frac{p(n-1)^2}{n-p}\hat{\sigma}^2_{(i)}} = \left|t_i^*\right|\sqrt{\frac{n-p}{p}\frac{p_i}{1-p_i}} = WK_i\sqrt{\frac{n-p}{p}} \tag{1.26}$$

The cut-off point for this measure is defined as $2\sqrt{\frac{n-p}{n}}$.

1.5 An Overview of Generalized Linear Models

Generalized Linear Models (GLM) which is the combination of various statistical models first introduced by Nelder and Weddernburn (1972). This method is an extension of Linear Regression Model (LRM). In LRM, several assumptions such as normality, homoscedasticity and linearity should be provided. On the other hand, for generalized linear models, there is no need to do such assumptions to represent a mathematical way of measuring the relationship between a response variable and a set of independent variables. In GLM, the response variable (Y) may belong to any exponential family instead of normal distribution. This part is called as the random component of GLM. This family includes the Gaussian, Bernoulli, binomial, Poisson, geometric, negative binomial, exponential, gamma, chi-square, Weibull and Drichlet distributions. Instead of homoscedasticity assumption, variance function is used to explain how the variation depend on the value of the mean.

In GLM context,

$$\eta = X\beta = \sum_{j=1}^{p} x_j \beta_j \tag{1.27}$$

is called as the systematic component or linear predictor of the model. The link function in Equation (1.28) is used to build a relation, which is linear in parameters, between the conditional mean of the response and the explanatory variables. Link function is determined by the probability distribution of Y.

$$g(\mu) = \eta = X\beta \tag{1.28}$$

It is assumed that a random variable *Y* given *X* depending on parameter $x^T\beta$ is a member of exponential family if its probability density function can be written in the following form

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{1.29}$$

for some specific functions $a(.), b(.), c(.)$ and canonical parameter θ. ϕ is a dispersion parameter, and we assumed that it is a positive constant, $\phi > 0$. In (1.29), $b(.)$ and $c(.)$ depend on the distribution, and generally $a(\phi)$ has the form $a(\phi) = \phi/\omega$ where known prior weights ω are also positive, $\omega > 0$. In matrix notation, each component of *Y* is assumed to be a member of the exponential family.

The exponential family form of the most commonly used distributions are as follows;

- Normal distribution

$$f_Y = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} = \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\}$$

where $\theta = \mu$, $a(\phi) = \sigma^2$, $b(\theta) = \frac{\mu^2}{2}$ and $c(y,\phi) = -\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right\}$. Here, the link function is the "identity" link so that $\theta = \mu$.

- Poisson distribution

$$f_Y = \frac{e^{-\mu}\mu^y}{y!} = \exp\{y\ln\mu - \mu - \ln(y!)\}$$

where $\theta = \mu\ln$, $a(\phi) = 1$, $b(\theta) = \mu = e^\theta$ and $c(y,\phi) = -\ln(y!)$. The link function is the "log" link.

- Binomial distribution

$$f_Y = \binom{n}{y}\pi^y(1-\pi)^{n-y} = \exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right) + n\ln(1-\pi) + \ln\binom{n}{y}\right\}$$

Here, $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$ where $\pi = \frac{e^\theta}{1-e^\theta}$, $a(\phi) = 1$, $b(\theta) = n\ln(1+e^\theta)$ and $c(y,\phi) = \ln\binom{n}{y}$. The link function is the "logit" link.

- Gamma distribution

$$f_Y = \frac{\beta^\alpha}{\gamma(\alpha)} y^{\alpha-1}e^{-\beta y} = \exp\left\{-\beta y + \ln\beta^\alpha + (\alpha-1)\ln y - \ln\gamma(\alpha)\right\}$$
$$= \exp\left\{\alpha\left[y\left(-\frac{\beta}{\alpha}\right) + \ln\beta\right] + (\alpha-1)\ln y - \ln\gamma(\alpha)\right\}$$

where $\theta = -\frac{\beta}{\alpha}$, $a(\phi) = \phi = \frac{1}{\alpha}$, $b(\theta) = -\ln\beta = -\ln(-\alpha\theta)$ and $c(y,\phi) = (\alpha-1)\ln y - \ln\gamma(\alpha)$. The link function is the "reciprocal" link.

The rest of this dissertation is organized as follows. In the next chapter we present robust JaB method for linear regression models. Chapter 3 describes delete-2 JaB method to evaluate influential observations in linear regression models. Chapter 4 extends the JaB method to binary logistic regression models. Concluding remarks are given in Chapter 5.

CHAPTER TWO
ROBUST BCa-JaB METHOD AS A DIAGNOSTIC TOOL FOR LINEAR REGRESSION MODELS

2.1 Introduction

In this chapter, we propose using robust versions of Efron (1987)'s bias-corrected and accelerated (BCa) bootstrap confidence intervals to improve the performance of the JaB method. Briefly, the rationale behind the use of the JaB method in this context is its construction based on deleting observations one at a time, with cut-offs based on the bootstrap distribution of the influence measure calculated using the JaB approach which allows the consideration of bootstrap samples that are free of the influence of a particular data point. We will refer here to the JaB method which uses as estimated cut-off values the $\alpha/2\,th$ and $(1-\alpha/2)\,th$ quantiles of the generated JaB distribution as conventional JaB. In the nomenclature of bootstrap confidence intervals, conventional JaB cut-offs are based on a simple percentile-method approach. Martin and Roberts (2010) and Beyaztas and Alin (2013) showed that conventional JaB has several advantages over the traditional asymptotic cut-offs for the measures considered in this dissertation. First, the conventional JaB method does not impose symmetric cut-offs, while the traditional cutoffs, based on normal theory, are symmetric. The asymmetry allowed by conventional JaB is advantageous particularly in small sample situations and when the underlying data distributions are very skewed. Nevertheless, percentile-method confidence limits have well-known deficiencies, a common criticism being that they tend to be too short, and that they can thus often under-cover significantly. In the present context, this under coverage will manifest as a tendency to fail to flag points that may, in fact, be unusual. Bias-corrected and accelerated bootstrap confidence intervals were proposed by Efron (1987) to improve the performance of percentile-method bootstrap intervals. Unlike percentile-method confidence limits, BCa interval endpoints account properly for bias and skewness, leading to second-order correct confidence limits $\hat{\theta}^{[\alpha]}$ (i.e. $\hat{\theta}^{[\alpha]} = \hat{\theta}^{EXACT[\alpha]} + O_p(n^{-1})$, where $\hat{\theta}^{EXACT[\alpha]}$ is an exact confidence limit) in the rich smooth functions of means model setting of which the cases considered here are

examples (for detailed information on the smooth functions of means model, see Hall, 1992, p.52-53), and thus to reduced coverage error for the resultant intervals over percentile-method intervals. Moreover, BCa intervals retain the useful property of percentile-method intervals of being transformation-respecting, a feature not enjoyed by other second-order correct approaches, such as the percentile-*t* approach. Further, the percentile-*t* approach, which also has better coverage properties than percentile-method intervals, suffers in the present context because of the difficulty in stably estimating the variance of the influence diagnostics, particularly in small sample settings.

The detection of influential data points in regression is a difficult problem, especially in the presence of *masking* effects. If there are several unusual data points that are located *en masse* far away from the bulk of the data, these points mask one another. In this case, typically used diagnostic measures fail to detect these points as influential. In general, one needs a diagnostic measure based on *multiple*-case deletion techniques to identify these points successfully. Here, we extend the work of Martin and Roberts (2010) and Beyaztas and Alin (2013), and propose replacing percentile-method based influence cut-offs with cut-offs developed using BCa confidence limits within the JaB approach. We further adjust the BCa cut-offs by a calibration of the nominal coverage probability to make them more robust to masking effects.

2.1 The Robust BCa-JaB Method

Efron (1992) described the rationale underlying the JaB method as follows: a sample of size *n* from $y_1, y_2, ..., y_{i-1,} y_{i+1}, ..., y_n$ drawn at random with replacement has the same distribution as a bootstrap sample from $y_1, y_2, ..., y_n$ in which none of the bootstrapped values equals y_i. In order to generate a specified number of such resamples, this method requires about *e* times the number of resamples than for a regular bootstrap because the probability of a "regular" resample not containing a specific data point is, roughly, e^{-1}. For example, for any data set, if we want to determine whether an individual data point is influential or not, to obtain 1000

resamples without this particular data point, about $1000e \approx 3000$ resamples are required.

Then, these 1000 resamples may be used to construct the JaB-sampling distribution of, say, an influence diagnostic, and thus to determine the influence cut-offs. Let $\hat{\theta}$ be the estimator of the parameter, θ, the influence measure under consideration. As described by Martin and Roberts (2010), the rationale behind this approach is to generate a "null" bootstrap distribution of $\hat{\theta}$ under the hypothesis that the *i'*th data point is not influential. They noted that, since the *i'*th data point is not present in any of the resamples from which the JaB-bootstrap distribution is generated, it cannot exert influence, and thus the sampling distribution generated is free from the influence of this point. The percentiles of this sampling distribution were thus used as cut-offs to gauge how extreme the observed influence measure associated with that data point was in the context of that influence-free null distribution.

The BCa method introduced by Efron (1987) is an automatic algorithm for producing highly accurate bootstrap confidence limits (see also DiCiccio and Efron (1996)). The method adjusts percentile-method bootstrap confidence limits based on the notional existence of a normalizing transformation on the estimator of θ, $\hat{\phi} = m(\hat{\theta})$, for which $\frac{\hat{\varphi} - \varphi}{\sigma_{\hat{\varphi}}} \sim N(-z_0, 1)$ holds, where $\sigma_{\phi} = 1 + a\phi$, z_0 is the bias correction parameter and a is the acceleration constant. For detailed information about BCa intervals, see DiCiccio and Efron (1996) and Efron (1987). Suppose θ is the parameter of interest, $\hat{\theta}$ is an estimate of θ based on the original data, $\hat{q}^*$ is the bootstrap version of $\hat{\theta}$ from a bootstrap resample, and *B* is the number of bootstrap replications. Let $\hat{G}(c)$ in Equation (2.1) be the cumulative distribution function (cdf) for $\hat{q}^*$ based on *B* bootstrap replications, $\hat{\theta}_b^*$, which in the present context is the influence measure calculated from *b*'th bootstrap resample:

$$\hat{G}(c) = \#\{\hat{\theta}_b^* < c\} / B \tag{2.1}$$

The α'th BCa quantile is calculated as

$$\hat{\theta}^{BCa[\alpha]} = \hat{G}^{-1}\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{\alpha}}{1-\hat{a}(\hat{z}_0 + z^{\alpha})} \right) \tag{2.2}$$

where $\hat{G}^{-1}$ is the inverse of the cdf given in Equation (2.1), and $P\left(Z \le z^{\alpha}\right) = \alpha$ for Z which is N(0,1). The quantities $\hat{z}_0$ and $\hat{a}$, calculated as in Equations (2.3) and (2.4), below, are the estimates of the bias-correction parameter and the acceleration constant, respectively.

$$\hat{z}_0 = \Phi^{-1}\left\{ \frac{\#(\hat{\theta}_b^* < \hat{\theta})}{B} \right\} \tag{2.3}$$

$$\hat{a} = \frac{\sum_{\substack{i,j=1 \\ i \neq j}}^{n} (\bar{\hat{\theta}}_{(-i)} - \hat{\theta}_{(-i),j})^3}{6\left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^{n} (\bar{\hat{\theta}}_{(-i)} - \hat{\theta}_{(-i),j})^2 \right\}^{3/2}} \tag{2.4}$$

The estimates $\hat{\theta}_{(-i),j}$ and $\bar{\hat{\theta}}_{(-i)}$ in Equation (4) are, respectively, the values of $\hat{\theta}$ for the j'th observation calculated without the i'th data point, and the mean of all such $\hat{\theta}_{(-i),j}$ values. We will have (n – 1) diagnostic measures $\hat{\theta}_{(-i),j}$ for each of $i = 1, \ldots,$ n and hence, in total, we will have $n(n-1)$ $\hat{\theta}_{(-i),j}$ values The acceleration constant, $\hat{a}$, is a measure of how quickly the standard error of the transformed estimator is changing on the normalized scale. For instance, for the simulation study in this chapter, this parameter was calculated as 0.101 when two deliberately-inserted influential observations were included in the data set with n = 50. In practice, this means that even though we may think $\tilde{\phi}$ is 1.645 standard errors to the right of the normalized estimator $\hat{\phi}$ of $\phi = m(\theta)$, if $\tilde{f} = \hat{f} + 1.645s_{\hat{f}}$, with $\hat{a}$ = 0.101 and $\sigma_{\tilde{\phi}} = (1+1.645\hat{a})\sigma_{\hat{\phi}} = 1.166145$, so $\tilde{\phi}$ is actually 1.645/1.166145 ≈ 1.410 standard errors to the right of $\hat{\phi}$. Similarly, this result turns out to be about 1.519 rather than 1.645 standard errors away from $\hat{\phi}$ for the case of three additional influential observations. When the bootstrap sampling distribution is symmetric and $\hat{\theta}$ is an

unbiased estimator of θ, Equation (2.3) will equal zero. When both Equations (2.3) and (2.4) are zero, Equation (2.2) becomes simply the α'th cut-off for a percentile-method bootstrap confidence interval (and in that case, the percentile method is itself second-order correct). Of course, in most practical circumstances, Equations (2.3) and (2.4) are not zero, and the BCa method corrects the percentile method endpoints to be second-order correct. Note that explicit knowledge of the normalizing transformation, *m*, is not required to implement the BCa method, only the notion that such a transformation exists, at least approximately.

Constructing BCa confidence limits in the context of JaB sampling distributions of influence diagnostic measures is somewhat different from the usual form of BCa in a straightforward parameter estimation context. In that latter context, the bootstrap distribution comprises *B* resampled quantities, while in our JaB context, there are approximately n^2B/e resampled diagnostic statistics (see BCa-JaB algorithm, below, for details). Usually, the estimate of the bias correction parameter $\hat{z}_0$ given in Equation (2.3) is for one statistic. However, in our context we deal with *n* different statistics which are the diagnostic measures calculated corresponding to each of the *n* deleted observations. As a result, the mean of these *n* measures for the original sample, $\bar{\hat{\theta}}$, replaces $\hat{\theta}$ in the usual form of Equation (2.3). With an unbounded influence function, the mean may be very sensitive to unusual data points. To counter this issue in calculating $\hat{z}_0$, we further propose using a robust 20%-trimmed mean, $\bar{\hat{\theta}}_{20\%trim}$ (see Rousseeuw and Leroy (1987), for more detailed information on trimmed means and choice of trimming proportion) rather than a regular average. Let $\hat{z}_0^R$ and $\hat{a}^R$, respectively, represent the robust estimates of the bias correction parameter and acceleration constant that result from using 20%-trimmed rather than regular means.

These estimators are calculated as follows:

$$\hat{z}_0^R = \Phi^{-1}\left\{\frac{\#\{\hat{\theta}^*_{i,b} < \bar{\hat{\theta}}_{20\%trim}\}}{\#S}\right\} \quad i = 1, \ldots, n \text{ and } b = 1, \ldots, B \tag{2.5}$$

$$\hat{a}^R = \frac{\sum_{\substack{i,j=1 \\ i\neq j}}^{n} (\hat{\bar{\theta}}_{(-i)20\%trim} - \hat{\theta}_{(-i),j})^3}{6\left\{\sum_{\substack{i,j=1 \\ i\neq j}}^{n} (\hat{\bar{\theta}}_{(-i)20\%trim} - \hat{\theta}_{(-i),j})^2\right\}^{3/2}} \tag{2.6}$$

where $\hat{\theta}^*_{b,i}$ is the *i*'th influence diagnostic measure estimated from the *b*'th bootstrap resample and #*S* is the total number of bootstrap influence diagnostics $\hat{\theta}^*_{b,i}$ calculated, which is *nB*. Further, we have made an additional modification to the α'th BCa quantile given in Equation (2.2) in which symmetric z^{α} quantiles are used (say ± 1.96, for 2.5%'th and 97.5%'th standard normal quantiles). Using these symmetric z^{α} quantiles yields second-order correct confidence limits that are inadequate to flag the influential observations. Instead of using these symmetric values, we propose a calibration based on an *adjusted* BCa-level, $z^{BCa[\alpha]}$, where BCa(α) is calculated as

$$BC_a(\alpha) = \Phi\left(\hat{z}_0^R + \frac{\hat{z}_0^R + z^{\alpha}}{1-\hat{a}^R(\hat{z}_0^R + z^{\alpha})}\right) \tag{2.7}$$

from the original data set before proceeding with the JaB algorithm. Let $\hat{G}(c)_{JaB}$ and $\hat{G}^{-1}_{JaB}$ represent the JaB cdf and its inverse, respectively. After determining BCa(α) as given in Equation (2.7), the robust-BCa (RBCa) α'th quantile for the JaB sampling distribution is computed as follows.

$$\hat{\theta}^{RBCa-JaB[\alpha]} = \hat{G}^{-1}_{JaB}\Phi\left(\hat{z}_0^{JaBR} + \frac{\hat{z}_0^{JaBR} + z^{BCa[\alpha]}}{1-\hat{a}^R(\hat{z}_0^{JaBR} + z^{BCa[\alpha]})}\right) \tag{2.8}$$

The estimate of the bias correction parameter for the JaB sampling distribution, $\hat{z}_0^{JaBR}$, is calculated as in Equation (2.5) except with $\hat{\theta}^*_{i,b}$ replaced by $\hat{\theta}^{JaB*}_{i,b}$, which is the estimate of influence diagnostic for the *i*'th observation obtained from the *b*'th JaB resample, and #*S* will be n^2B/e instead of *nB*. Lower and upper limits $\hat{\theta}^{RBCa-JaB[\alpha/2]}$ and $\hat{\theta}^{RBCa-JaB[1-\alpha/2]}$ are then calculated. These limits are used as cut-off values for the relevant influence measures to flag influential observations.

The principal benefit of considering using BCa endpoints in the bootstrap construction rather than simple percentile-method cutoffs is to correct for the well-known under coverage suffered by percentile-method bootstrap confidence intervals. The BCa approach, by comparison, is easy to compute and provides the same asymptotic benefit as the percentile-*t* method while enjoying the feature of transformation equivariance as previously mentioned. The use of robust measures within the BCa construction is designed to reduce the impact of unusual points in constructing the diagnostic measures designed to reveal such points, and the 20%-trimmed means used in the construction remain root-*n* consistent estimators of the underlying parameters because the trimming proportion is fixed, so the second-order correctness property of BCa method calculated with the quantities given in Equations (2.5) and (2.6) is retained. We now briefly discuss some asymptotics that describe some of the properties of the calibrated robust BCa method we propose.

Hall (1988) developed asymptotic expansions for the bias correction z_0 and acceleration constant α in terms of polynomials p_i and q_i within Edgeworth expansions for the distributions of the $n^{1/2}(\hat{\theta}-\theta)/\sigma$ and $n^{1/2}(\hat{\theta}-\theta)/\hat{\sigma}$, respectively in the context of the smooth function model. Following Hall (1988)'s notation,

$$z_0 = n^{-1/2} p_1(0) + O(n^{-1}) \tag{2.9}$$

where p_1 is the first polynomial in the Edgeworth expansion for the distribution function of the standardized quantity $n^{1/2}(\hat{\theta}-\theta)/\sigma$, and

$$a = n^{-1/2}(z^{\alpha})^{-2}\{p_1(z^{\alpha}) + q_1(z^{\alpha}) - 2p_1(0)\} \tag{2.10}$$

where q_1 is the first polynomial in the Edgeworth expansion for the distribution function of the studentized quantity $n^{1/2}(\hat{\theta}-\theta)/\hat{\sigma}$. Note that the polynomial $\{p_1(z^{\alpha}) + q_1(z^{\alpha}) - 2p_1(0)\}$ has no constant or linear term since $p_1(0) = q_1(0)$ and p_1 and q_1 are even, quadratic polynomials. As a result of these observations, neither the bias correction z_0 nor the acceleration constant a depend on z^{α}. As a result, replacing z_0 and a with asymptotically consistent estimators (designated generically

here with a hat, and noting that the robust estimators used in the construction of the bias correction and acceleration constant estimates are asymptotically consistent, and that the diagnostic measures under discussion are each smooth functions of means), we see

$$z^{BCa[\alpha]} = \hat{z}_0^R + \frac{\hat{z}_0^R + z^\alpha}{1-\hat{a}^R(\hat{z}_0^R + z^\alpha)} = z^\alpha + 2\hat{z}_0^R + \hat{a}^R(z^\alpha)^2 + O_p(n^{-1})$$
$$= z^\alpha + n^{-1/2}\{\hat{p}_1(z^\alpha) + \hat{q}_1(z^\alpha)\} + O_p(n^{-1}) \qquad (2.11)$$

where $\hat{p}_1$ and $\hat{q}_1$ are the first polynomials in the respective Edgeworth expansions for the distribution functions of $n^{1/2}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$ and $n^{1/2}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$. Hence,

$$BCa(\alpha) = \Phi\left(\hat{z}_0^R + \frac{\hat{z}_0^R + z^\alpha}{1-\hat{a}^R(\hat{z}_0^R + z^\alpha)}\right)$$
$$= \alpha + n^{-1/2}\{\hat{p}_1(z^\alpha) + \hat{q}_1(z^\alpha)\}\varphi(z^\alpha) + O_p(n^{-1})$$
$$= \alpha + O_p(n^{-1/2}). \qquad (2.12)$$

Thus, the calibration we propose with Equation (2.7) is of order $n^{-1/2}$. Now we give the following theorem for asymptotic results regarding the correctness of the proposed calibrated robust quantile $\hat{\theta}^{RBCa[\alpha]}$ and the coverage errors of the one- and two-sided intervals

$$\hat{I}_{RBCa(1-sided)}(\alpha) = (-\infty, \hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\alpha]})$$

$$\hat{I}_{RBCa(2-sided)}(\alpha) = (\hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\frac{1-\alpha}{2}]}, \hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\frac{1+\alpha}{2}]})$$

The results, given for the regular bootstrap sampling distribution, also hold for JaB sampling distribution since both distributions are equivalent (see Efron (1992), Lemma 1).

Theorem:

(i) $\hat{\theta}^{RBCa[\alpha]}$ is first-order correct for $\hat{\theta}^{BCa[\alpha]}$; that is, $\hat{\theta}^{RBCa[\alpha]}$ differs from $\hat{\theta}^{BCa[\alpha]}$ by terms of asymptotic order $n^{-1/2}$.

(ii) The one-sided interval $\hat{I}_{RBCa(1-sided)}(\alpha) = (-\infty, \hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\alpha]})$ is first-order accurate.

(iii) The equal-tailed two-sided interval

$$\hat{I}_{RBCa(2-sided)}(\alpha) = (\hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\frac{1-\alpha}{2}]}, \hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\frac{1+\alpha}{2}]})$$

is second-order accurate.

Proof:

(i) Let $z^{RBCa[\alpha]}$ be defined as follows:

$$z^{RBCa[\alpha]} = \hat{z}_0^R + \frac{\hat{z}_0^R + z^{BCa[\alpha]}}{1 - \hat{a}^R(\hat{z}_0^R + z^{BCa[\alpha]})} = z^{BCa[\alpha]} + 2\hat{z}_0^R + \hat{a}^R(z^{BCa[\alpha]})^2.$$

Using Equation (11), we get $z^{RBCa[\alpha]} = z^{\alpha} + 2n^{-1/2}\{\hat{p}_1(z^{\alpha}) + \hat{q}_1(z^{\alpha})\} + O_p(n^{-1})$. By Cornish-Fisher expansion, we have the following result which completes the proof.

$$\begin{aligned}\hat{\theta}^{RBCa[\alpha]} &= z^{RBCa[\alpha]} - n^{-1/2}\hat{p}_1(z^{RBCa[\alpha]}) + O_p(n^{-1}) \\ &= z^{\alpha} + n^{-1/2}(\hat{p}_1(z^{\alpha}) + \hat{q}_1(z^{\alpha})) + n^{-1/2}\hat{q}_1(z^{\alpha}) + O_p(n^{-1}) \\ &= \hat{\theta}^{BCa[\alpha]} + O_p(n^{-1/2}),\end{aligned}$$

where

$$\hat{\theta}^{BCa[\alpha]} = z^{BCa[\alpha]} - n^{-1/2}\hat{p}_1(z^{BCa[\alpha]}) + O_p(n^{-1}) = z^{\alpha} + n^{-1/2}\hat{q}_1(z^{\alpha}) + O_p(n^{-1}).$$

(ii)

$$\begin{aligned}\hat{I}_{RBCa(1-sided)}(\alpha) &= P(\theta \le \hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\alpha]}) = P\left(\frac{n^{1/2}(\hat{\theta} - \theta)}{\hat{\sigma}} \ge -\hat{\theta}^{RBCa[\alpha]}\right) \\ &= 1 - P\left(\frac{n^{1/2}(\hat{\theta} - \theta)}{\hat{\sigma}} \le -\hat{\theta}^{RBCa[\alpha]}\right) = 1 - \hat{K}(-\hat{\theta}^{RBCa[\alpha]})\end{aligned}$$

where $\hat{K}(x) = \Phi(x) + n^{-1/2}\hat{q}_1(x)\phi(x) + O_p(n^{-1})$ is the Edgeworth expansion for the distribution of the studentized quantity $n^{1/2}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$. Then, we get the following result which proves that $\hat{I}_{RBCa(1-sided)}(\alpha)$ is first-order accurate:

$$
\begin{aligned}
P(\theta \le \hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\alpha]}) &= 1 - \Phi(-\hat{\theta}^{RBCa[\alpha]}) - n^{-1/2}\hat{q}_1(-\hat{\theta}^{RBCa[\alpha]})\phi(-\hat{\theta}^{RBCa[\alpha]}) + O_p(n^{-1}) \\
&= 1 - \Phi(-\{z^{\alpha} + n^{-1/2}(\hat{p}_1(z^{\alpha}) + \hat{q}_1(z^{\alpha})) + n^{-1/2}\hat{q}_1(z^{\alpha})\}) \\
&\quad - n^{-1/2}\hat{q}_1(\hat{\theta}^{RBCa[\alpha]})\phi(z^{\alpha}) + O_p(n^{-1}) \\
&= \alpha + n^{-1/2}(\hat{p}_1(z^{\alpha}) + \hat{q}_1(z^{\alpha}))\phi(z^{\alpha}) + O_p(n^{-1}) \\
&= \alpha + O_p(n^{-1/2}).
\end{aligned}
$$

(iii)

$$
\begin{aligned}
\hat{I}_{RBCa(2-sided)}(\alpha) &= P\left(\hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\frac{1-\alpha}{2}]} \le \theta \le \hat{\theta} + n^{-1/2}\hat{\sigma}\hat{\theta}^{RBCa[\frac{1+\alpha}{2}]}\right) \\
&= P\left(\frac{n^{1/2}(\hat{\theta}-\theta)}{\hat{\sigma}} \ge -\hat{\theta}^{RBCa[\frac{1+\alpha}{2}]}\right) - P\left(\frac{n^{1/2}(\hat{\theta}-\theta)}{\hat{\sigma}} \ge -\hat{\theta}^{RBCa[\frac{1-\alpha}{2}]}\right) \\
&= 1 - P\left(\frac{n^{1/2}(\hat{\theta}-\theta)}{\hat{\sigma}} \le -\hat{\theta}^{RBCa[\frac{1+\alpha}{2}]}\right) - 1 + P\left(\frac{n^{1/2}(\hat{\theta}-\theta)}{\hat{\sigma}} \le -\hat{\theta}^{RBCa[\frac{1-\alpha}{2}]}\right) \\
&= \hat{K}\left(-\hat{\theta}^{RBCa[\frac{1-\alpha}{2}]}\right) - \hat{K}\left(-\hat{\theta}^{RBCa[\frac{1+\alpha}{2}]}\right)
\end{aligned}
$$

Since, $z = z^{(1+\alpha)/2} = -z^{(1-\alpha)/2}$, we have

$$
\begin{aligned}
\hat{I}_{RBCa(2-sided)}(\alpha) &= \Phi(z - n^{-1/2}(\hat{p}_1(z) + 2\hat{q}_1(z))) + n^{-1/2}\hat{q}_1(z)\phi(z) \\
&\quad - \Phi(-z - n^{-1/2}(\hat{p}_1(-z) + 2\hat{q}_1(-z))) - n^{-1/2}\hat{q}_1(-z)\phi(-z) + O_p(n^{-1}) \\
&= \Phi(z) - n^{-1/2}(\hat{p}_1(z) + 2\hat{q}_1(z))\phi(z) + n^{-1/2}\hat{q}_1(z)\phi(z) \\
&\quad - \Phi(-z) + n^{-1/2}(\hat{p}_1(z) + 2\hat{q}_1(z))\phi(z) - n^{-1/2}\hat{q}_1(z)\phi(z) + O_p(n^{-1}) \\
&= \alpha + O_p(n^{-1}).
\end{aligned}
$$

Note that $\hat{p}_1(z)$, $\hat{q}_1(z)$ and $\phi(z)$ are even functions.

Even though our proposed method is first-order correct and it yields a first-order accurate one-sided interval compared to the usual BCa method, which is second-order correct, it still maintains second-order accuracy for two-sided equal tailed intervals which is the context that we use to determine our cut-offs. Moreover, it is empirically more successful in flagging the points under masking effects. Below is the algorithm of the proposed RBCa-JaB method.

Let $\hat{\theta}_i$ represent the diagnostic statistic for *i*'th observation under consideration, being one of DFFITS, Welsch's distance and modified Cook's distance in our case.

Step 1: Fit the appropriate model to the original data set, and calculate the measure $\hat{\theta}_i$ for each of *n* observations for $i = 1, 2, \ldots, n$.

Step 2: *C*alculate $\hat{\theta}_{(-i),j}$ for $i, j = 1, 2, \ldots, n$ for $i \neq j$. Note that for each *i*, we will have $(n-1)$ measures giving $n(n-1)$ measures in total for the *n* data points. Then, collect all these $n(n-1)$ values of $\hat{q}_{(-i),j}$ into a single column vector and calculate the 20% trimmed mean $\bar{\hat{\theta}}_{20\%trim}$.

Step 3: Calculate the robust acceleration constant $\hat{a}^R$ as in Equation (2.6) using $\bar{\hat{\theta}}_{20\%trim}$ obtained in Step 2. It should be noted this value is calculated using all $n(n-1)$ measures.

Step 4: Draw $B = 2000$ (say) resamples with replacement from the original data set and calculate $\hat{\theta}_i^*$ for $i = 1, 2, \ldots, n$ for each of the *B* resamples. Then, collect all *nB* measures into a vector $\hat{\theta}^*$

Step 4: Calculate the robust bias correction parameter $\hat{z}_0^R$ as in Equation (2.5) using the trimmed mean calculated in Step 2. Here, #*S* equals *nB*.

Step 5: Calculate adjusted BCa$(\alpha / 2)$ and BCa$(1 - \alpha / 2)$ using Equation (2.7). Now we are ready to begin the "jackknife after bootstrap" component of the calculations for our proposed Robust JaB method.

Step 6: Draw $B = 3100$ resamples (i.e., sufficient to obtain about $B = 1000$ for each deleted data point) with replacement from the original data set.

Step 7: For each data point (denoted *i*) within these *B* resamples, obtain the subset of the resamples without each individual data point in turn, so there are $B_i \approx \frac{B}{e} \approx 1000$ resamples obtained for each data point. Calculate *n* values of $\hat{\theta}_i^{JaB*}$ for each of these resamples, so $nB_i \approx nB/e$ values of $\hat{\theta}_i^{JaB*}$ in total,

are obtained. Collect these nB_i values of $\hat{\theta}_i^{JaB*}$ into a single vector, denoted here as $\hat{\theta}^{JaB*}$.

Step 8: Calculate the bias-correction parameter $\hat{z}_0^{JaBR}$ as in Equation (2.5), for $\hat{\theta}^{JaB*}$ using the trimmed mean obtained in Step 2. Note that for these calculations, $\#S = n\sum_{i=1}^{n} B_i \approx n^2 B / e$.

Step 9: Calculate the RBCa confidence limits of the generated JaB distribution, $\hat{\theta}^{RBCa-JaB[\alpha/2]}$ and $\hat{\theta}^{RBCa-JaB[1-\alpha/2]}$, given in Equation (2.8) using $\hat{z}_0^{JaBR}$ and $\hat{a}^R$ calculated in Steps 8 and 3, respectively. These quantiles are then compared to the original $\hat{\theta}_i$, i = 1, 2, …, n, values to gauge the extent to which the points are influential or not.

Steps 1 to 9 are repeated M times to form the simulation. Of course, it should be noted that the algorithm runs only once for a real data set.

2.3 Numerical Results

The performance of the methods was assessed using three real data sets and two simulated scenarios. All calculations were carried out using R 2.15.2 on an Intel Core i7-2670QM 2.20 GHz PC. It should be noted that in all tables for the real world data and simulation study, the values in parentheses below the RBCa-JaB cut-offs are the adjusted quantiles corresponding to Equation (2.7).

2.3.1 Real Data Examples

2.3.1.1 Life Cycle Savings Data

The life cycle savings data for 50 countries are explained by four explanatory variables. According to Belsley et al. (1980), the set includes influential observations that were determined by various diagnostic measures described in that monograph. Noted influential points corresponded to countries such as Japan (point 23), Zambia

(46), Libya (49), Canada (6), Chile (7), South Rhodesia (37), and the United States (44).

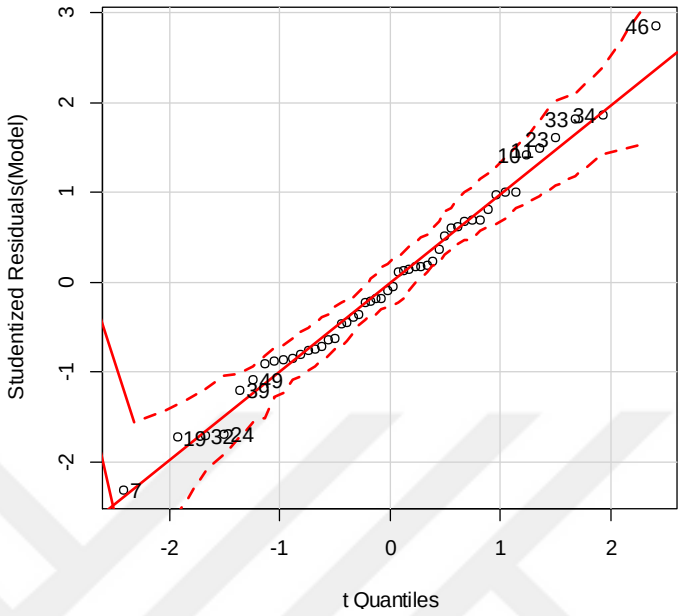

Figure 2.1 Normal quantile plot for life cycle savings data

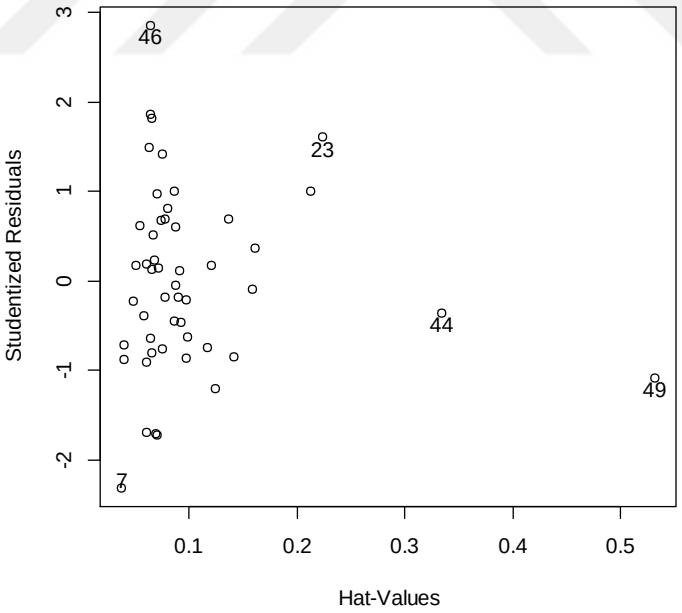

Figure 2.2 Influence plot for life cycle savings data

In this chapter, we will study this well-known data set using conventional JaB influence cut-offs and our proposed RBCa-JaB cut-offs. As evident from the normal quantile plot (Figure 2.1), some points seem unusual, but no point extends beyond

the dotted boundaries suggesting that the normality assumption remains tenable. In this case, the JaB distribution was not overly affected by these points, and from Table 2.1, it can be observed that the cut-offs are very similar for both versions of JaB. In other words, the performance of both methods was roughly the same in terms of detecting influential points. According to the regression influence plot presented as Figure 2.2, points 23, 46, 49 and 7 appear influential, but both JaB versions were unable to flag point 7 (Chile). Arguably, of course, the traditional, non-bootstrap based cut-offs have acted conservatively in flagging point 7, and by imposing cut-offs more adapted to the data, the JaB methods have made a sound decision.

Table 2.1 Regression influence diagnostics for life cycle savings data, n=50, p=5

Method		**Welsch's distance**	**Modified Cook's distance**	**DFFITS**
Conventional JaB				
	Low cut-off	-4.477	-1.772	-0.590
	High cut-off	5.122	2.052	0.685
	Points below	49	49	49
	Points above	23, 46	23, 46	23, 46
RBCa -JaB				
	Low cut-off	-4.451 (2.54%)	-1.797 (2.37%)	-0.599 (2.37%)
	High cut-off	5.152 (97.54%)	2.028 (97.36%)	0.676 (97.36%)
	Points below	49	49	49
	Points above	23, 46	23, 46	23, 46

2.3.1.2 The Hertzsprıng-Russell Diagram of Star Cluster Data

For this dataset, there is one explanatory variable to explain the logarithm of light intensity of a star. Points 7, 11, 14, 20, 30 and 34 appear as if they may be influential when we examine Figure 2.4. Our results are shown in Table 2.2. In contrast to the previous example, the cut-off values are demonstrably not similar for both versions of JaB. This difference occurred for this example because of the bias associated with the relatively large proportion of influential observations. As observed from Figure 2.4, points 11, 20, 30 and 34 are located in a line of points next to one another, and the influence of each of these observations has an effect on the next one, and vice versa. That is, this data set exhibits a pronounced masking effect, a phenomenon that was well examined by Martin and Roberts (2010). These data points also are at the extreme in terms of their response values – see Figure 2.3. Despite these problems,

RBCa-JaB was able to detect points 14, 34, 11, 20, 30 as influential, whereas traditional single-case-deletion methods fail because of the pronounced masking effect. Multiple-case deletion techniques do, of course, detect the points, but the question then arises as to what size of group must be deleted in order to detect all of the points reliably. On the other hand, conventional JaB is able to flag only the first two data points, making the success of our robust BCa-JAB method even more noteworthy.

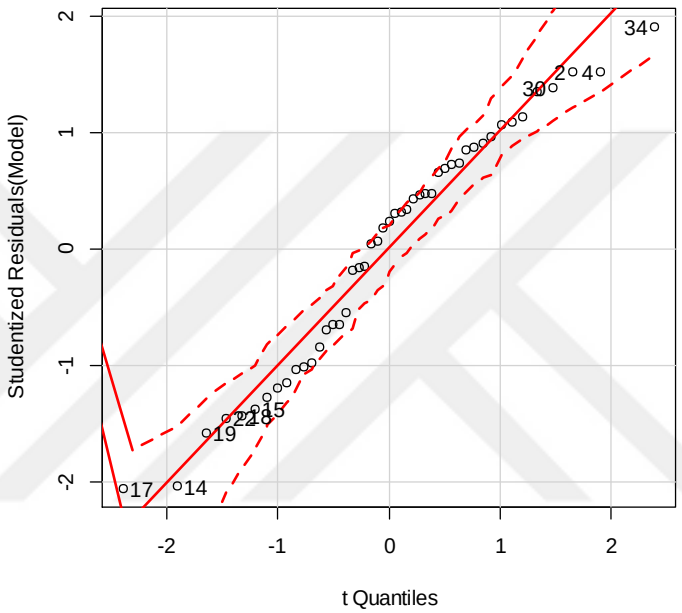

Figure 2.3 Normal quantile plot for Hertzsprung-Russell diagram of star cluster data

Table 2.2 Regression influence diagnostics for Hertzsprung-Russell diagram of the star cluster data, n=47, p=2

Method		**Welsch's distance**	**Modified Cook's distance**	**DFFITS**
Conventional JaB				
	Low cut-off	-2.395	-1.639	-0.345
	High cut-off	5.419	3.397	0.716
	Points below	14	14	14
	Points above	34	34	34
RBCa -JaB				
	Low cut-off	-2.986 (0.85%)	-2.029 (0.86%)	-0.427 (0.86%)
	High cut-off	2.622 (93.99%)	1.746 (94.03%)	0.368 (94.03%)
	Points below	14	14	14
	Points above	11, 20, 30, 34	20, 30, 34	20, 30, 34

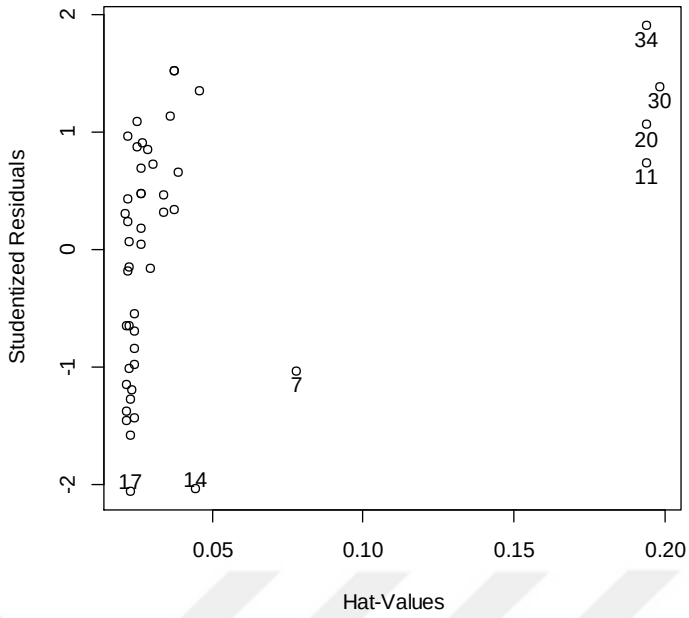

Figure 2.4 Influence plot for Hertzsprung-Russell diagram of star cluster data

2.3.1.3 Soil Evaporation Data

This data set given by Freund (1979) is also available in the "*Teaching Demos*" R package. There are 46 observations and 10 explanatory variables. Two points, 2 and 33, appear to materially impact the normal quantile plot (see Figure 2.5), and RBCa-JaB adjusts the cut-off points to the right to handle this problem. Points 2, 31, 32 and 41 appear unusual, and potentially influential, from Figure 2.6. Our results for this data set are presented in Table 2.3. As in the previous example, RBCa-JaB is more effective than the conventional JaB method for all influence measures we considered. This data set also suffers a serious masking problem, with points 2 and 41 masking the impact of the other data points. This masking seriously impacts the conventional JaB method, to the extent that it is unable to adequately detect any but one of the problematic data points. Nevertheless, our proposed method is successful in detecting *all* of the suspect points for all measures, overcoming the serious masking effect present in this data set.

Table 2.4 includes the estimates of the acceleration constants and the calibrated coverage probabilities for each of the three examples. The calibrated coverage

probabilities are calculated as differences of adjusted upper and lower RBCa-JaB quantiles given in Tables 2.1 to 2.3.

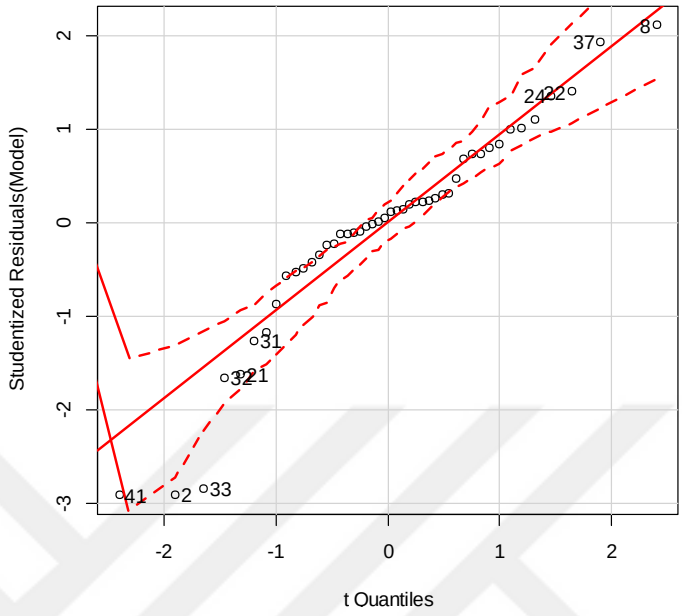

Figure 2.5 Normal quantile plot for soil evaporation data

Table 2.3 Regression influence diagnostics for the soil evaporation data, n=46, p=11

Method		**Welsch's distance**	**Modified Cook's distance**	**DFFITS**
Conventional JaB				
	Low cut-off	-20.176	-3.992	-2.238
	High cut-off	10.578	2.296	1.287
	Points below	31	None	None
	Points above	None	None	None
RBCa -JaB				
	Low cut-off	-12.632 (4.33%)	-2.749 (4.22%)	-1.541 (4.22%)
	High cut-off	13.795 (98.69%)	2.859 (98.63%)	1.602 (98.63%)
	Points below	2, 31, 32, 41	2, 31, 32, 41	2, 31, 32, 41
	Points above	None	None	None

2.3.2 Simulation Study

A simulation study was conducted based on the design of Martin and Roberts (2010). We considered the cases $(n, p) = (20, 2)$ (small sample), $(n, p) = (50, 5)$ (large sample), and three error distributions: normal ($N(0,0.5625)$), $t_{(3)}$ (heavy-tailed), and centered lognormal ($1.5\left[\exp\{N(0,0.5625)\} - \exp(\frac{1}{2})\right]$, skewed). The

"true" regression models are assumed as $Y = 1 + 2X_1 + 4X_2 + 3X_3 + 2X_4 + \varepsilon$ for the large sample cases and $Y = 1 + 2X + \varepsilon$ for the small sample cases, respectively. For each simulated case, X was generated as i.i.d. $N(2, 1)$ variates and ε was generated with one of the three error distributions mentioned above. We considered two scenarios for large sample cases, one of which included two deliberately inserted data points, whereas the other included three influential points. The influential points $(x_2, y) = (10, 10)$ were deliberately inserted twice when the data had two additional influential observations, and $(x_2, y) = (0, 0)$ was also added as a third influential observation for the second scenario. Small sample cases included only two deliberately inserted data points. Points $(x, y) = (5, 2)$ were added twice into the data set each time. For each influence measure, $M = 500$ simulations were performed. The nominal α was taken as 0.05. The results are presented in Tables 2.5 and 2.6 for large sample scenarios and in Table 8 for small sample scenarios. The average number of points flagged as influential for all simulations is recorded as "Average no. of points" in the tables. For the deliberately inserted influential data points, these results are given as "Av. no of deliberately inf. points". Standard errors are given in brackets.

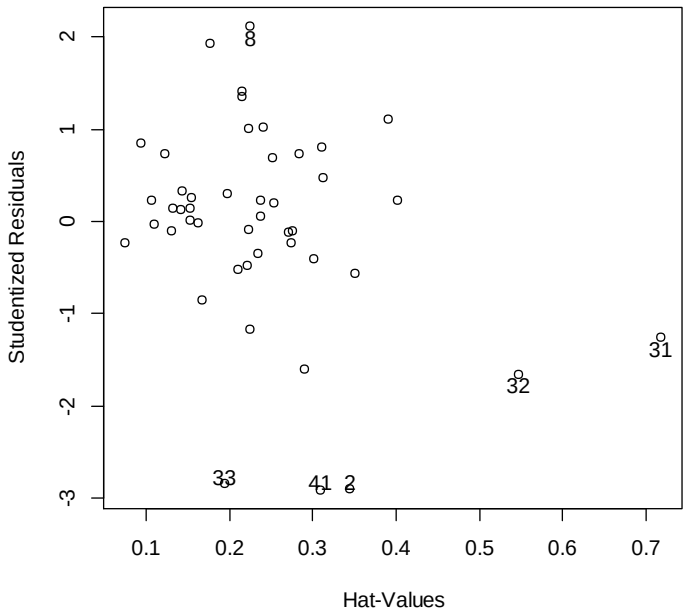

Figure 2.6 Influence quantile plot for soil evaporation data

Table 2.4 Acceleration constant, $\hat{a}$, values and calibrated coverage probabilities for real-data examples for all influence measures considered

	Welsch's distance		**Modified Cook's distance**		**DFFITS**	
Example	$\hat{a}$	Coverage	$\hat{a}$	Coverage	$\hat{a}$	Coverage
E.1 Life Cycle Savings Data	0.00368	95.00%	0.00061	94.99%	0.00061	94.99%
E.2 Star Data	-0.00633	93.14%	-0.00541	93.17%	-0.00541	93.17%
E.3 Soil Evaporation Data	0.00847	94.36%	0.00641	94.41%	0.00641	94.41%

For both scenarios and sample sizes, the structure of the JaB distribution is materially distorted by the deliberately inserted influential data points. The skewness induced in the data sets caused by the inserted points is to the left, and the JaB method adjusted the cut-offs accordingly to handle this problem. However, the RBCa-JaB method was able to refine and calibrate the adjustments made by the conventional JaB method on the cut-offs to correct the JaB bootstrap confidence limits so that we ended up with slightly more symmetric cut-offs. This refinement has resulted in considerably better observed performance by the robust BCa-JaB over the conventional JaB method for all distributions of errors considered, especially for the small sample size cases. Since the added influential observations are more disruptive in small samples, the JaB distribution for $n = 20$ is more distorted by those included points compared to the large sample scenarios. As such, we observed that the RBCa-JaB method worked better for the small sample size cases. The calibrated coverage probabilities of the proposed method were marginally less than the nominal coverage of 95%, but they were still within acceptable limits despite the small sample sizes involved and reflected a significant improvement over the undercoverage typical for percentile-method intervals – see, for example, Dale (1986).

2.4 Discussion and Conclusion

In this chapter, we proposed replacing percentile-method bootstrap confidence limits with BCa bootstrap confidence limits within the JaB method proposed by Efron (1992). However, we made two further adjustments to the BCa cut-off formula originally conceived: first, we sought to make the estimation of the bias-correction

parameter in the BCa algorithm more robust against outlying or influential observations by calculating the bias-correction parameter estimate using a 20% trimmed mean; and second, we calibrated the nominal level of the standard normal quantiles within the BCa formula using the estimated BCa level based on the bias-correction and acceleration constant adjustments. As previously described by Martin and Roberts (2010) and Beyaztas and Alin (2013), the conventional JaB method has been shown already to possess key advantages over traditional methods, since it automatically takes into account the underlying distribution structure in setting cut-offs that are not necessarily symmetric. Our proposed adjustments in making the JaB method more robust by adjusting the confidence limits used in determining the relevant JaB cut-offs has been shown in this paper to produce significantly improved outcomes over the conventional JaB method. Even though our BCa-level adjustment came at an apparent theoretical cost of sacrificing second-order correctness of the resultant interval endpoints, the intervals themselves remained second-order accurate, and the choice appears to have been justified by the excellent empirical performance of the proposed method in a variety of difficult real-world data sets and consistently within our simulation study. In particular, our proposed method has been shown to work well even in the presence of pronounced masking effects, a feature that improves markedly on the performance of the regular JaB approach, which itself is better than traditional approaches. Normally, in our experience, overcoming such masking effects has required the use of multiple-case-deletion methods, so it is particularly noteworthy that the adjustments we have proposed to the JaB methodology has resulted in such good performance. Of course, it must be acknowledged that the notion of influence is tied closely to the adequacy or correctness of the underlying assumed model, and that if a different model were pursued – for example, one that allowed for curvature of other effects such as heteroscedasticity – then the points detected with respect to the linear model we have assumed may no longer appear influential or even particularly unusual. That said, the use of linear models in scientific endeavor is widespread and the detection of influential points remains an important practical problem for scientists seeking to understand their data.

Table 2.5 Simulation results, n=50, p=5 with two inf. obs. for all distribution of errors.

Distribution of errors		**Normal**			**t(3)**			**Log-normal**		
Method		**Welsch's distance**	**Modified Cook's distance**	**DFFITS**	**Welsch's distance**	**Modified Cook's distance**	**DFFITS**	**Welsch's distance**	**Modified Cook's distance**	**DFFITS**
Conventional JaB										
	Low cut-off	-9.336	-3.433	-1.144	-9.257	-3.400	-1.133	-8.936	-3.284	-1.094
	High cut-off	4.760	1.901	0.633	4.805	1.919	0.639	4.984	1.995	0.665
	Average no. of points	3.208	3.178	3.178	3.178	3.208	3.208	3.050	3.03	3.03
	(SD)	(0.660)	(0.674)	(0.674)	(0.641)	(0.646)	(0.646)	(0.663)	(0.667)	(0.667)
	Av. no of deliberately inf. points	1.981	1.973	1.973	1.990	1.984	1.984	1.978	1.956	1.956
RBCa -JaB										
	Low cut-off	-3.785 (6.51%)	-1.586 (6.48%)	-0.528 (6.48%)	-3.956 (7.45%)	-1.687 (6.72%)	-0.562 (6.72%)	-3.850 (7.46%)	-1.668 (6.56%)	-0.556 (6.56%)
	High cut-off	10.620 (99.53%)	3.647 (99.45%)	1.215 (99.45%)	9.717 (99.80%)	3.285 (99.65%)	1.095 (99.65%)	9.985 (99.64%)	3.293 (99.42%)	1.097 (99.42%)
	Average no. of points	4.075	3.848	3.848	3.790	3.51	3.51	4.05	3.742	3.742
	(SD)	(1.166)	(1.098)	(1.098)	(1.080)	(0.9632)	(0.9632)	(1.210)	(1.126)	(1.126)
	Av. no of deliberately inf. points	1.997	1.993	1.993	1.996	1.994	1.994	1.998	1.992	1.992

Table 2.6 Simulation results, n=50, p=5 with three inf. obs. for all distribution of errors.

Distribution of errors		**Normal**			**t(3)**			**Log-normal**		
Method		**Welsch's distance**	**Modified Cook's distance**	**DFFITS**	**Welsch's distance**	**Modified Cook's distance**	**DFFITS**	**Welsch's distance**	**Modified Cook's distance**	**DFFITS**
Conventional JaB										
	Low cut-off	-11.897	-4.480	-1.493	-11.950	-4.507	-1.502	-11.107	-4.195	-1.398
	High cut-off	4.249	1.695	0.565	4.196	1.675	0.558	4.558	1.818	0.606
	Average no. of points	3.354	3.452	3.452	3.342	3.450	3.450	3.272	3.322	3.322
	(SD)	(0.799)	(0.772)	(0.772)	(0.801)	(0.839)	(0.839)	(0.786)	(0.807)	(0.807)
	Av. no of deliberately inf. points	2.338	2.432	2.432	2.432	2.544	2.544	2.280	2.340	2.340
RBCa -JaB										
	Low cut-off	-6.712 (5.36%)	-2.974 (4.65%)	-0.991 (4.65%)	-6.735 (5.08%)	-2.972 (4.41%)	-0.990 (4.41%)	-5.603 (5.78%)	-2.470 (5.07%)	-0.823 (5.07%)
	High cut-off	6.093 (99.14%)	2.158 (98.83%)	0.719 (98.83%)	5.829 (99.03%)	2.089 (98.70%)	0.696 (98.70%)	6.943 (99.23%)	2.461 (98.96%)	0.820 (98.96%)
	Average no. of points	3.403	3.422	3.422	3.330	3.374	3.374	3.478	3.406	3.406
	(SD)	(0.633)	(0.632)	(0.632)	(0.581)	(0.592)	(0.592)	(0.717)	(0.640)	(0.640)
	Av. no of deliberately inf. points	2.911	2.875	2.875	2.928	2.910	2.910	2.946	2.918	2.918

Table 2.7 Simulation results, n=20, p=2 with two inf. obs. for all distribution of errors.

Distribution of errors		**Normal**			**t(3)**			**Log-normal**		
Method		**Welsch's distance**	**Modified Cook's distance**	**DFFITS**	**Welsch's distance**	**Modified Cook's distance**	**DFFITS**	**Welsch's distance**	**Modified Cook's distance**	**DFFITS**
Conventional JaB										
	Low cut-off	-8.061	-4.682	-1.560	-7.829	-4.544	-1.514	-6.724	-3.918	-1.306
	High cut-off	2.865	1.851	0.617	2.920	1.888	0.629	3.300	2.143	0.714
	Average no. of points	0.820	0.968	0.968	0.598	0.716	0.716	0.652	0.650	0.650
	(SD)	(0.881)	(1.018)	(1.018)	(0.696)	(0.858)	(0.858)	(0.592)	(0.613)	(0.613)
	Av. no of deliberately inf. points	0.212	0.364	0.364	0.080	0.192	0.192	0.040	0.044	0.044
RBCa -JaB										
	Low cut-off	-6.166 (4.96%)	-3.720 (4.75%)	-1.240 (4.75%)	-5.602 (5.13%)	-3.428 (4.91%)	-1.142 (4.91%)	-4.804 (5.16%)	-2.944 (4.95%)	-0.981 (4.95%)
	High cut-off	4.134 (98.66%)	2.503 (98.54%)	0.834 (98.54%)	4.131 (98.77%)	2.520 (98.65%)	0.840 (98.65%)	4.505 (98.60%)	2.777 (98.50%)	0.925 (98.50%)
	Average no. of points	1.960	2.004	2.004	1.958	1.974	1.974	1.936	1.932	1.932
	(SD)	(0.758)	(0.738)	(0.738)	(0.813)	(0.799)	(0.799)	(0.810)	(0.825)	(0.825)
	Av. no of deliberately inf. points	1.620	1.636	1.636	1.620	1.628	1.628	1.528	1.500	1.500

CHAPTER THREE
DELETE-2 JACKKNIFE-AFTER-BOOTSTRAP IN REGRESSION

3.1 Introduction

In this chapter, we propose Delete-2 Jackknife-after-Bootstrap (D-2 JaB) method to refine the cut-offs when the data have multiple influential data points with masking and swamping effects. we only concentrate on Cook's distance because of its poor performance against the masking and swamping effects. The points with masking effect mask the influence of the other points while the ones with swamping effect cause the good data points incorrectly diagnosed as influential. We may have only one data point masking the effect of another or multiple-point interactions which can only be detected through deletion of multiple points that may cause swamping effect. The masking problem is a topic of considerable recent research interest (See Martin et al. (2010) and references therein).

The JaB sampling distribution is obtained from bootstrap samples where none of the values equal to the *i*th data point. Hence, the cut-off values will be free of the influence of this particular data point. This method itself is based on single-case deletion and it may not be successful under masking or/and swamping effects when it is used on any diagnostics. Hence, we propose an approach to the use of Cook's distance that involves applying Cook's distance to Delete-2 Jacknife after Bootstrap (D-2 JaB) resamples. Through this approach, the sampling distribution as well as the cut-off values for Cook's distance will be free of the effect of any influential pair of points. Deleting a d ($d \geq 2$) group of observations requires $C_{n,d} = n!/\{d!\,(n-d)!\}$ different reduced samples, meaning too much computational burden and time. Hence, we only focused on d = 2 (delete-2) case which is enough for our proposed method to flag masked observations.

Martin et al. (2010) proposed a new approach to the use of single-case deletion diagnostics that involves applying these diagnostics to delete-2 and delete-3 jacknife replicates of the data. Below is the single-case deleted Cook's distance formula for

*i*th observation applied to the delete-2 jackknife replicates when both *k*th and *s*th data points for $k = 1, \dots, n-1, = k+1, \dots, n$, $k \neq s$ are removed.

$$\hat{\theta}_{(i)}^{(k,s)} = \frac{\left(\hat{\beta}^{(k,s)} - \hat{\beta}_{(i)}^{(k,s)}\right)^T \left(X^T X\right)\left(\hat{\beta}^{(k,s)} - \hat{\beta}_{(i)}^{(k,s)}\right)}{p\hat{\sigma}^2} \tag{3.1}$$

$\hat{\beta}^{(k,s)}$ and $\hat{\beta}_{(i)}^{(k,s)}$, respectively, are the least squares estimates for the regression coefficient obtained from the data sets where (*k*,*s*) th and (*k*,s,*i*) th data points missing. $\hat{\sigma}^2$ is the maximum likelihood estimate for the error variance. The cut-off point for the Cook's distance is chosen as 1 or the median value of corresponding *F* distribution. In this study, the observations are flagged as influential if its corresponding Cook's distance value is greater than 1. However, this approximation is ineffective to detect influential points under the presence of the points causing masking or/and swamping effect.

3.2 D-2 JaB Method

Following Lemma is a generalization of the Efron (1992)'s idea.

Lemma

A sample of size *n* from $y_1, y_2, \dots, y_{k-1}, y_{k+1}, \dots, y_{s-1}, y_{s+1}, \dots, y_n$ drawn at random with replacement has the same distribution as a bootstrap sample from $y_1, y_2, \dots, y_n$ in which none of the bootstrap values equals y_k and y_s for $k = 1, \dots, n-1$, $s = k+1, \dots, n$, $k \neq s$.

The proof of lemma follows directly Efron (1992)'s *Lemma 1*. Our method requires about e^2 times the number resamples for a regular bootstrap because the probability of a "regular" resample not containing a pair of (*k, s*) th data points is, roughly, e^{-2}. For example, for any data set, if we want to determine whether an individual data point is influential or not, to obtain 1000 resamples without a pair data points, about $1000 \times e^2 \approx 7400$ resamples are required. Then, these 1000 resamples may be used to estimate the D-2 JaB-sampling distribution of Cook's distance measure and the influence cut-offs. The rationale behind this approach is to

generate a "null" bootstrap distribution of Cook's distance under the hypothesis that the joint influence of observations *k* and *s* on the other observations is insignificant. Since the *k*th and *s*th data points are not present in any of the resamples from which the D-2 JaB distribution is generated, they cannot exert influence, and thus the sampling distribution generated is free from the influence of these points. The percentiles of this sampling distribution were thus used as cut-offs to gauge how extreme the observed influence measure associated with that data point were in the context of that influence-free null distribution. The algorithm of D-2 JaB can be described briefly as follows:

Step 1. Fit the appropriate model to the original data set, and calculate the single-case deleted Cook's distance measure for *i*th data point, $\hat{\theta}_{(i)}$ for $i = 1,2, \dots, n,$ for full data set as

$$\hat{\theta}_{(i)} = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)^T (X^T X)\left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{p\hat{\sigma}^2}$$

where $\hat{\beta}$ and $\hat{\beta}_{(i)}$ are the least squares estimates obtained from full set and reduced set without *i*th point, respectively.

Step 2. Draw *B* resamples with replacement from the original data set.

Step 3. For a specific pair of (*k*,*s*)*th* data points, say (1,2), obtain the subset of the resamples without corresponding pair within these *B* resamples.

Step 4. Calculate $\hat{\theta}^*_{(i)}, i = 1, \dots, n$ as defined in Step 1 for each of these resamples, and collect these values into a single vector, say $\widehat{\theta}^*$.

Step 5. Repeat Steps 3 and 4 for each pair of (*k*, *s*) for $= 1, \dots, n-1$, $s = k + 1, \dots, n$, $k \neq s$.

Step 6. Determine the suitable quantiles (say 2.5% and 97.5%) of this generated bootstrap distribution including approximately $nC_{n,2}\, B/e^2$ values.

Step 7. Compare these quantiles with each of the original $\hat{\theta}_{(i)}$ for $i = 1,2, \dots, n$ values calculated in Step 1 to check if the points are influential or not.

For calculating Cook's distance based on Martin et al. (2010) approach, we construct the jackknife sample by removing the (*k*, *s*)*th* (*k*≠ *s*) data points from the

original data set, and obtain the usual Cook's distance based on single-case deletion technique for remaining *n*-2 observations as in Equation (3.1). For this reduced data set, the point is flagged as influential if its Cook's distance value exceeds the cut-off 1.

Martin et al. (2010) considered two different percentages to determine if any point is influential. The first one given in Equation (3.2) is calculated to see how unusual a data point appears to be.

$$dr = \frac{\#d}{C_{n-1,2}} \tag{3.2}$$

$\#d$ is the detection number of a specific data point in $C_{n-1,2} = (n-1)!/\{2!\,(n-3)!\}$ different reduced data sets without the corresponding pair. According to Martin et al. (2010), this overall percentage will typically be large for clearly influential, unmasked data points, but may remain small for an unusual point masked by another data point, as the masking point will be present in many of the jackknife replicates as well. The second percentage is calculated to reveal points that we consider to be masked by another point. Martin et al. (2010), considered (*n*-2) jackknife replicates that do not contain the potentially masking data point in question so that they could use this as a measure of how unusual each data point is when the potentially masking data point is excluded. If a low *dr* becomes significantly larger among replicates not containing a certain data point, this is considered as an evidence of single-point masking.

3.3 Numerical Study

We studied the performance of our proposed method with four well-known data sets which are used in many papers in this context and a designed simulation study. All calculations have been done under the assumption that the linear regression model is the correct model, and all the models satisfy the homoscedasticity assumption. For all of real data examples, 8000 resamples were created from the original data set, so that roughly 1000 resamples without each pair of binary data points were produced for each pair. The calculations were obtained using R 2.15 (see

Appendix for the R codes). The values in the brackets in the Tables are the detection rates (*dr*) of flagged influential observations calculated as in Equation (3.2).

3.3.1 Real Data Examples

3.3.1.1 Life Cycle Savings Data

As visible from the normal quantile plot (Figure 2.1), some points seem unusual, but no point extends beyond the dotted boundaries suggesting that the normality assumption remains tenable. According to the regression influence plot presented as Figure 3.1, points 23, 46, 49 and 7 appear influential. In this plot, the areas of the circles represent the observations proportional to Cook's distance. It seems that there is no masking effect between the individual data points since there is no interlocking circle in this figure. On the other hand, discarding two observations at a time causes swamping effect for this data set. Table 3.1 shows that D-2 jackknife Cook's distance flagged points 47 and 48 which are not influential while Cook's distance with the refined cut-offs obtained by D-2 JaB method successfully flagged three of the influential data points.

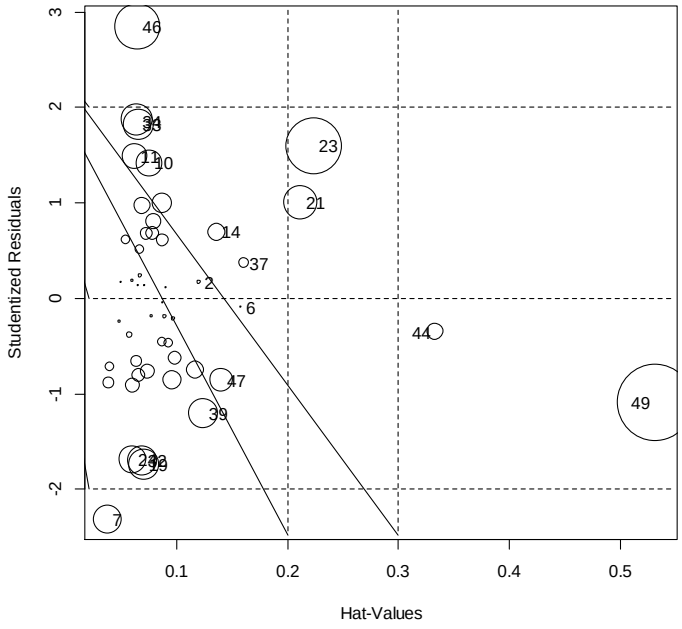

Figure 3.1Influence plot for life cycle savings data

Table 3.1 Regression influence diagnostics for life cycle savings data, $n = 50$, $p = 5$

Method		
D-2 Jacknife Cook's distance		
	Cut-off	1.000
	Influential points	47 ($dr = 0.00387$)
		48 ($dr = 0.00129$)
D-2 JaB Cook's distance		
	Cut-off	0.077
	Influential points	23, 46, 49

3.3.1.2 The Hertzsprung-Russell Diagram of Star Cluster Data

The description of this data set is given in Chapter 2. Our results for this data set are presented in Table 3.2. Normality assumption is satisfied for this data set (see Figure 2.3), and as observed from Figure 3.2, points 11, 20, 30 and 34 are located in a line of points next to one another, and the influence of each of these observations has an effect on the next one, and vice versa. That is, this data set exhibits a pronounced masking effect, a phenomenon that was well examined by Rousseeuw and Leroy (1987). Despite these problems, Cook's distance with cut-offs from D-2 JaB method was able to detect points 14, 20, 30, 34 as influential, whereas the opponent traditional D-2 Jackknife Cook's distance not only fails to catch these data points but also flags some good points as influential because of the swamping effect, such as point 13.

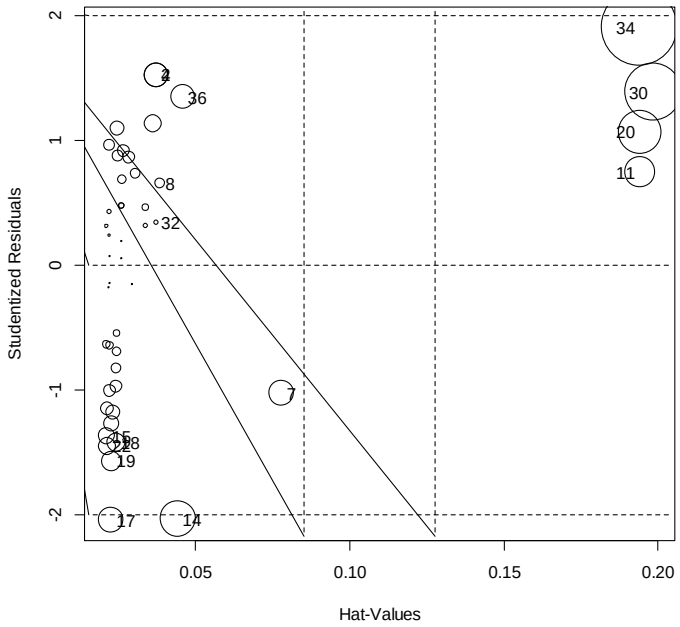

Figure 3.2 Influence plot for Hertzsprung-Russell diagram of star cluster data

Table 3.2 Regression influence diagnostics for Hertzprung-Russell diagram of star cluster data, n = 47, p = 2

Method	Cut-off	
D-2 Jacknife Cook's distance		1.000
		10 (dr = 0.0058)
		11 (dr = 0.0395)
	Influential points	12 (dr = 0.0761)
		13 (dr = 0.4216)
		14 (dr = 0.9736)
D-2 JaB Cook's distance		0.085
	Influential points	14, 20, 30, 34

3.3.1.3 Soil Evaporation Data

This data set given by Freund (1979) is also available in the "*Teaching Demos*" R package. There are 46 observations and 10 explanatory variables. Two points, 2 and 33, appear to materially impact the normal quantile plot (see Figure 3.3). Points 2, 31, 32 and 41 appear unusual, and potentially influential, from Figure 3.4. Our results for this data set are presented in Table 3.3. This data set also suffers a serious masking problem, with points 2 and 41 masking the impact of the other data points. This masking seriously affects the D-2 jackknife Cook's distance so that it is unable to detect problematic data points. Nevertheless, our proposed method is successful in detecting all of the suspicious points, overcoming the serious masking problem and not causing swamping effect.

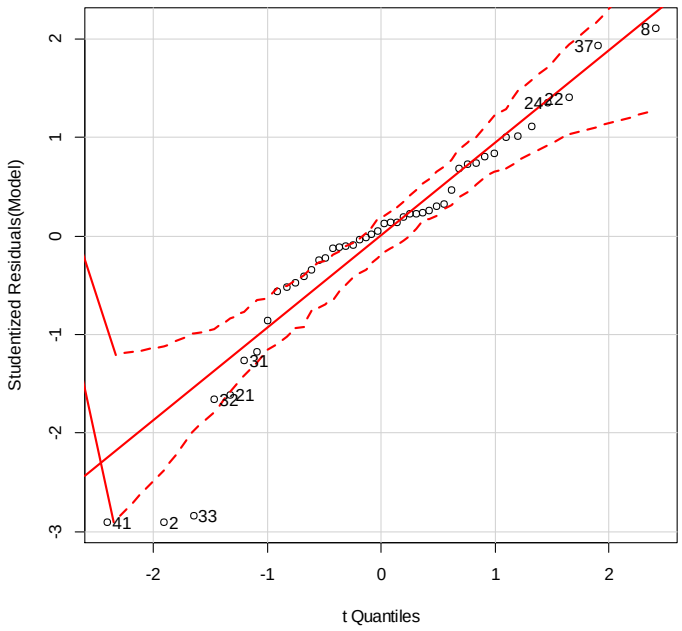

Figure 3.3 Normal quantile plot for soil evaporation data

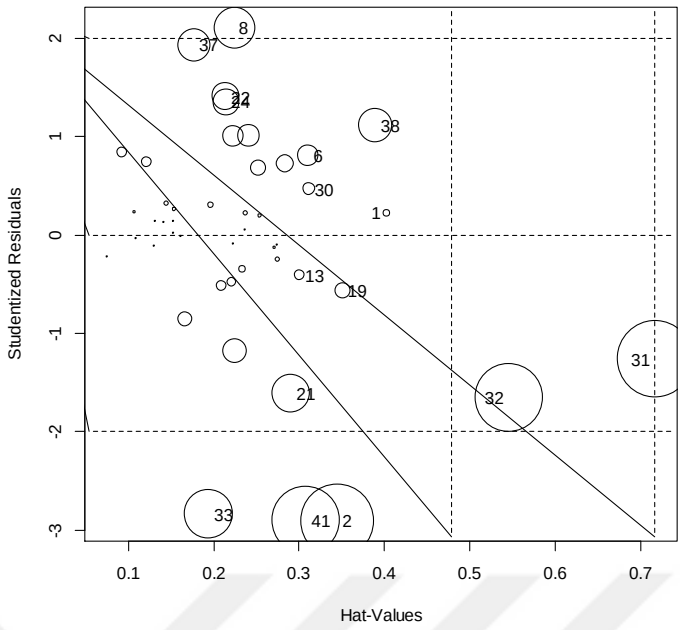

Figure 3.4 Influence plot for soil evaporation data

Table 3.3 Regression influence diagnostics for the soil evaporation data, $n = 46$, $p = 11$

Method		
D-2 Jacknife Cook's distance		
	Cut-off	1.000
	Influential points	30 ($dr = 0.0015$)
		32 ($dr = 0.0015$)
D-2 JaB Cook's distance		
	Cut-off	0.221
	Influential points	2, 31, 32, 41

3.3.1.4 Health Club Data

The Health Club data given by Chatterjee and Hadi (1988) consists of health records of 30 employees who were regular members of a given company's health club. For this data set, the dependent variable was modelled on four explanatory variables. According to Chatterjee and Hadi (1988), point 28 is influential based on influence curve, point 30 is outlying in the residuals, and point 23 is outlying in the explanatory variables. Points 28 and 30 seem influential (see Figures 3.5 and 3.6). According to our results presented in Table 3.4, D-2 jacknife Cook's distance is unsuccessful to identify those points while our proposed method can.

Table 3.4 Regression influence diagnostics for the health club data, n = 30, p = 5

Method		
D-2 Jacknife Cook's distance		
	Cut-off	1.000
	Influential points	None
D-2 JaB Cook's distance		
	Cut-off	0.205
	Influential points	28, 30

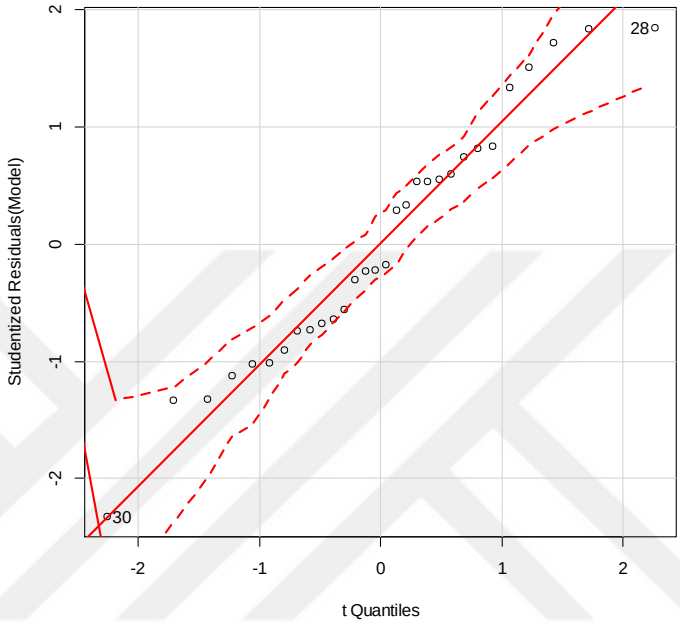

Figure 3.5 Normal quantile plot for health club data

3.4 Simulation Study

Simulation study is based on the design of Beyaztas et al. (2013) where considered cases are (n, p) = (20, 2) (small sample), (n, p) = (50, 5) (large sample) and three error distributions: normal ($N(0,0.5625)$), $t_{(3)}$ (heavy-tailed), and centered lognormal $(1.5\left[\exp\{N(0,0.5625)\}-\exp(\frac{1}{2})\right]$, skewed). The "true" regression models are assumed as $Y = 1 + 2X_1 + 4X_2 + 3X_3 + 2X_4 + \varepsilon$ for the large sample cases, and $Y = 1 + 2X + \varepsilon$ for the small sample cases, respectively. For each simulated case, X was generated as i.i.d. $N(2, 1)$ variates and ε was generated with one of the three error distributions mentioned above. For small sample, the first two observations were replaced by the points (x, y) = (5, 2). For large sample, on the other hand, the points {(10, 10), (10, 10), (0, 0)} took place of the first three data points . M = 500

simulations were performed for each scenario with the nominal significance level of α = 0.05. The proportion of inserted influential data points detected in all 500 simulations are presented in Tables 3.5 and 3.6 for small and large samples, respectively, where the significant success of the proposed method under considered scenarios is obvious.

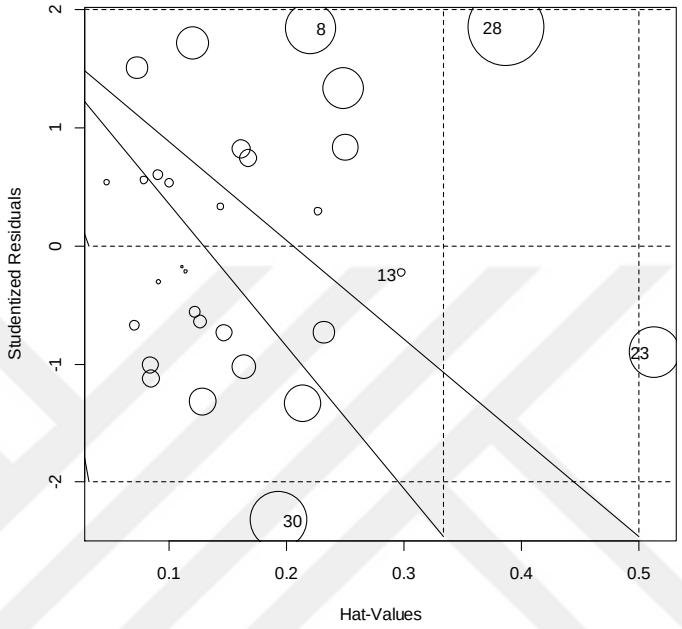

Figure 3.6 Influence plot for health club data

Table 3.5 Simulation results, n=20, p=2 with two inf. obs. for all distribution of errors

		Distribution of errors		
		Normal	**t(3)**	**Log-normal**
Method				
D-2 Jacknife Cook's distance				
	Cut-off	1.000	1.000	1.000
	Detection proportion for point 1	0.193	0.193	0.185
	Detection proportion for point 2	0.001	0.001	0.001
D-2 JaB Cook's distance				
	Cut-off	0.473	0.453	0.339
	Detection proportion for point 1	1.000	1.000	0.906
	Detection proportion for point 2	1.000	1.000	0.906

Table 3.6 Simulation results, n=50, p=5 with three inf. obs. for all distribution of errors

		Distribution of errors		
		Normal	**t(3)**	**Log-normal**
Method				
D-2 Jacknife Cook's distance				
	Cut-off	1.000	1.000	1.000
	Detection proportion for point 1	0.532	0.532	0.471
	Detection proportion for point 2	0.448	0.396	0.445
	Detection proportion for point 3	0.024	0.027	0.011
D-2 JaB Cook's distance				
	Cut-off	0.154	0.137	0.154
	Detection proportion for point 1	0.996	0.996	0.992
	Detection proportion for point 2	0.988	0.996	1.000
	Detection proportion for point 3	0.964	0.996	1.000

3.4 Conclusion and Discussion

In this study, we propose D-2 JaB method to refine the cut-offs for single case deleted Cook's distance under masking problem. The results for all the data sets and simulation study show that with the refined cut-offs which are robust masking and swamping effects, Cook's distance successfully flag influential observations. The computing time gets longer as we increase the units to be deleted (d). However, for $d = 2$ case which is enough for the proposed method to detect influential observations, computing time is around 4.5 minutes. Finally, it should be noted that all the calculations in this paper are performed under the assumption that linear model form is correct. Otherwise, as pointed out by a referee, the points flagged may not really be influential under the model that allows curvature.

We only worked on refining the cut-offs for Cook's distance because of its sensitivity against masking and swamping effects. However, this estimation method can also be used on other influence measures to improve their performance. Moreover, it can be combined with some methods such as cluster based bounded influence regression proposed by Lawrence et al. (2013) to detect unusual data points.

CHAPTER FOUR
JACKKNIFE-AFTER-BOOTSTRAP AS LOGISTIC REGRESSION DIAGNOSTIC TOOL

4.1 Introduction

This chapter extends JaB method to the logistic regression analysis to provide an alternative for the traditional diagnostic methods. Logistic regression, also known as logit model, is one of the most useful statistical modeling techniques to describe the relationship between a qualitative response variable and one or more explanatory variables. As in all mathematical models, identification and evaluation of unusual data points is a critical part of the logistic regression analysis. In binary logistic regression where two possible values 0 and 1 are available for response variable, each error term can take only two values causing the error terms to be non-normally distributed with heteroscedasticity. Because of these problems, detection of influential observations and outliers in binary logistic regression are problematic and difficult compared to linear regression. Commonly used principles for linear regression models to identify the unusual data points are also available for logistic regression models with some differences. The early studies in this context were conducted by Pregibon (1981) who proposed diagnostics using one-step approximation. See Hosmer and Lemeshow (2000) for an overview of research into logistic regression diagnostics. In this study, we choose the standardized Pearson residuals, Cook's distance, change in the Pearson chi-square and change in the deviance statistics as our diagnostic measures.

4.2 Model and the Measures

Suppose a binary response random variable *y* follows a Bernoulli distribution with probability π, and let $x_i = (x_{1i}, x_{2i}, ..., x_{pi})$ be a vector of *p* explanatory variables for the *i*th observation. The conditional probability for success $P(y_i = 1 | x_{1i}, x_{2i}, ..., x_{pi})$ is defined as follows;

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_{1i}+...+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+...+\beta_p x_{pi}}} \quad ; \; j = 1,2,...,n \tag{4.1}$$

where $\beta_0, \beta_1, ..., \beta_p$ are unknown parameters. In this situation, the conditional mean is bounded with 0 and 1 for all values of x_i since $E(y_i) = \pi(x_i)$. Logit transformation given in Equation (4.2) is used to overcome this problem.

$$g(x_i) = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} \quad ; \; j = 1,2,...,n \tag{4.2}$$

Since the error terms in the logistic regression model $Y = \pi(x) + \varepsilon$ take only two values, they are not normally distributed. Hence, instead of ordinary least squares estimation maximum likelihood estimation (MLE) method based on iteratively reweighted least square is used to estimate the parameters. But, the MLE method is very sensitive to outlying and influential data points. Therefore it is very important to detect these points.

In the logistic regression, diagnostics for influential observations are generally built on two residual types. The first is Pearson residual given in Equation (4.3).

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{4.3}$$

Since these residuals do not have a unit variance, the better approach is standardized Pearson residuals calculated using the following weighted hat matrix.

$$H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2} \tag{4.4}$$

where V is an $n \times n$ diagonal matrix of $\hat{\pi}_i(1-\hat{\pi}_i)$, and X is the $n \times (p+1)$ design matrix. According to Pregibon (1981), an internally Studentized residual can be obtained by dividing the Pearson residual by the square root of $1-h_{ii}$ where h_{ii} is the ith diagonal element of H.

$$sr_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}} = \frac{ri}{\sqrt{1-h_{ii}}} \tag{4.5}$$

The threshold value for this measure is α *th* quantile of standard normal distribution. The second type of residual is the deviance residual which shows the contribution of each point to the likelihood statistic. For logistic regression, it equals to the Equation (4.6).

$$d_i = sign(Y_i - \hat{\pi}(x_i))\left\{-2\left[Y_i \ln\left(\frac{Y_i}{\hat{\pi}(x_i)}\right) + (1-Y_i)\ln\left(\frac{1-Y_i}{1-\hat{\pi}(x_i)}\right)\right]\right\} \tag{4.6}$$

Change in the deviance and change in the Pearson chi-square statistics given in Equations (4.7) and (4.8) are important measures to detect influential measures. The former is based on d_i, while the latter is based on sr_i.

$$\Delta D_i = d_i^2 + \frac{r_i^2 h_{ii}}{1-h_{ii}} = \frac{d_i^2}{1-h_{ii}} \tag{4.7}$$

$$\Delta\chi_i^2 = \frac{r_i^2}{1-h_{ii}} = sr^2 \tag{4.8}$$

Each of these measures has chi-square distribution with one degree of freedom. Cook's distance is another influence measure used in this study. Based on the Pregibon's one step approximation, Cook's distance is calculated as

$$\Delta\hat{\beta}_j = (\hat{\beta} - \hat{\beta}_{(-i)})^T (X^T V X)(\hat{\beta} - \hat{\beta}_{(-i)}) = \frac{sr_i^2 h_{ii}}{1-h_{ii}} \tag{4.9}$$

where $\hat{\beta}_{(-i)}$ is the estimated regression coefficient when the *i*th observation is deleted. The cut-off point for the Cook's distance is chosen as 1 or the median value of corresponding *F* distribution. In our study, the observations are flagged as influential if its corresponding Cook's distance value is greater than 1. Even though the measures given in Equations (4.7-4.9) are used to detect influential observations, they measure different effects.

4.3 Algorithm of the JaB Method

The algorithm of JaB method for detection of influential observations in logistic regression can be described as follows:

Step 1. Let θ_i be the diagnostic statistic under study. The appropriate model is fitted for original data set, and the θ_i for *i*= 1, 2, …, *n* are calculated.

Step 2. Construct *B* resamples with replacement from the original data set.

Step 3. For each data point within these *B* resamples, get a subset of the samples which do not contain that data point, so there are *B/e* resamples obtained for each

data point. Calculate the n values of θ_i, $i = 1, 2, \ldots, n$, for each of these resample, so nB/e values of θ_i are obtained. Collect all nB/e values of θ into a single vector.

Step 4. Suitable quantiles (say 2.5% and 97.5%) of this generated bootstrap distribution are determined. Percentiles of this distribution are then compared to the original θ_i, $i = 1, 2, \ldots, n$, values to flag the points as influential or not.

4.3 Numerical Results

4.3.1 Real-World Examples

4.3.1.1 Finney's Data on Vasoconstriction in the Skin of the Digits.

In the original data set from Finney (1947) the response y is the occurrence ($y = 1$) or non-occurrence ($y = 0$) of vaso constriction in the skin of the digits of a subject after he or she inhaled a certain volume of air at a certain rate. This data set was used by Pregibon (1981) to illustrate the performance of diagnostic measures used for detecting influential observations. It has been reported that there are two outliers: points 4 and 18. We also used the same data to study the performance of our proposed method.

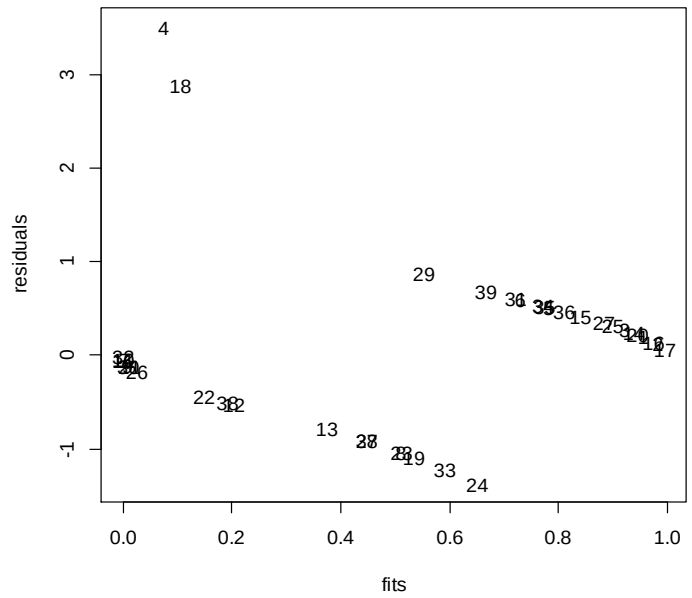

Figure 4.1Plot of fitted values versus residuals for Finney's data on vasoconstriction in the skin of the digits

Figure 4.1 shows the plot of fitted values versus residuals. It is very obvious that points 4 and 18 are far from the bulk of data. According to the results given in Table 4.1, almost all measures detect these observations as influential. Traditional and JaB versions of change in the deviance and chi-square statistics show the same performance. For standardized Pearson residuals JaB flagged one extra point 24. For Cook's distance, while JaB can flag both points, traditional method did not flag point 18. There are two problems with this data set which may be the reason for the inadequate behavior of Cook's distance. The first one is the masking effect of point 4 on point 18. Lawrence (1995) proposed a measure to determine the masking effect of a point on another for Cook's distance in linear regression. When we adopt that measure for logistic regression, we calculate the value of 1.3473, meaning point 18 is masked by point 4. This value means that the influence of point 18 increases 1.3473 times when point 4 is deleted. Another problem with the data set is the fitted probabilities for points 4 and 18. These probabilities are at the boundary. As pointed out by Hosmer and Lemesow (2000), this causes small leverages because of the related weights. Even though other diagnostic measures for those points are large, Cook's distance will probably be small. The results reveal that using JaB solves these problems for this data set.

Table 4.1 Results for Finney's data on vasoconstriction in the skin of the digits

Method		**Standardized Pearson Residuals**	**Cook's Distance**	**Pearson chi-square**	**Deviance**
Traditional					
	Low Cut-off	-1.960			
	High Cut-off	1.960	1.000	3.840	3.840
	Influential points	4, 18	4	4, 18	4, 18
JaB					
	Low Cut-off	-1.392			
	High Cut-off	2.445	0.354	2.934	2.846
	Influential points	4, 18, 24	4, 18	4, 18	4, 18

4.3.1.2 Modified Brown Data

The data set given in Brown (1980) includes 53 prostate cancer patients. The response and explanatory variables are the lymph node involvement and the level of

acid phosphates in the blood serum, respectively. Ryan (1997) reports that this set contains one influential observation: point 24. The data was modified by Imon and Hadi (2008) by adding two influential observations: points 54 and 55.

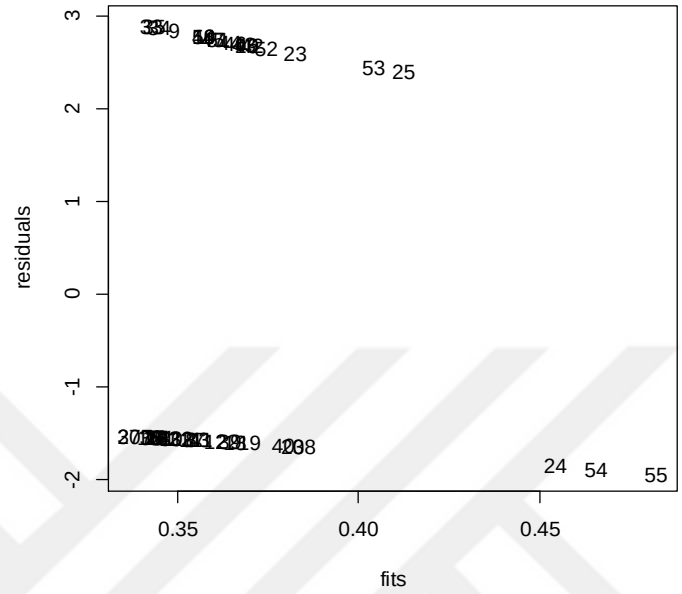

Figure 4.2 Plot of fitted values versus residuals for modified Brown data

Figure 4.2 shows that these three points deform the overall structure of the data. Even though they seem to be well explained by the model, they are separated from other observations. This pattern supports the notion that these observations would be influential. Our results given in Table 4.2 are compatible with the plot. These points have not been detected by both versions of standardized residuals, change in the chi-square and deviance statistics since they seem to be well explained by the model. As in the previous example this data set has masking problem which is the reason for inadequate behavior of traditional Cook's distance. According to the conditional masking measure proposed by Lawrence (1981), when we remove point 54, the Cook's distance value of point 24 increases 2.3041 times. This increment is 2.2435 times when we remove point 55. The change is more effective when points 54 and 55 are both removed such that all the measures can detect point 24. Being able to detect these points under severe masking proves the robustness of the JaB Cook's distance to masking effect.

Table 4.2 Results for modified Brown data

Method		Standardized Pearson Residuals	Cook's Distance	Pearson chi-square	Deviance
Traditional					
	Low Cut-off	-1960			
	High Cut-off	1.960	1.000	3.840	3.840
	Influential points	None	None	None	None
JaB					
	Low Cut-off	-1.099			
	High Cut-off	1.649	0.111	2.474	2.522
	Influential points	55	24, 25, 54, 55	None	None

4.3.1.3 Modified Kyphosis Data

The kyphosis data have 81 measurements for children who have had corrective spinal surgery. The dependent variable "Kyphosis" represents if a kyphosis (a type of deformation) is present or absent after the operation. There are two explanatory variables, the number of vertebrae involved and the number of the first (topmost) vertebra operated on. For the original data set, only point 43 has a large influence on the model fit. The original six observations (points 10, 11, 23, 40, 46 and 77) were deliberately changed by Nurunnabi et al (2010) to be influential. Our results for this modified set are presented with Table 4.3.

Except for Cook's distance, both methods perform equally for all diagnostic measures, and detect these points which seem to be not well explained by the model. According to Figure 4.3, point 43 is separated from other points which may affect the model fit. Even though we would expect it would be detected by Cook's distance, it was not. The reason for that point to go unnoticed by Cook's distance is its fitted probability (0.83) which is close to the boundary. We have the same type of problem as in the first example for which JaB version of Cook's distance is not affected.

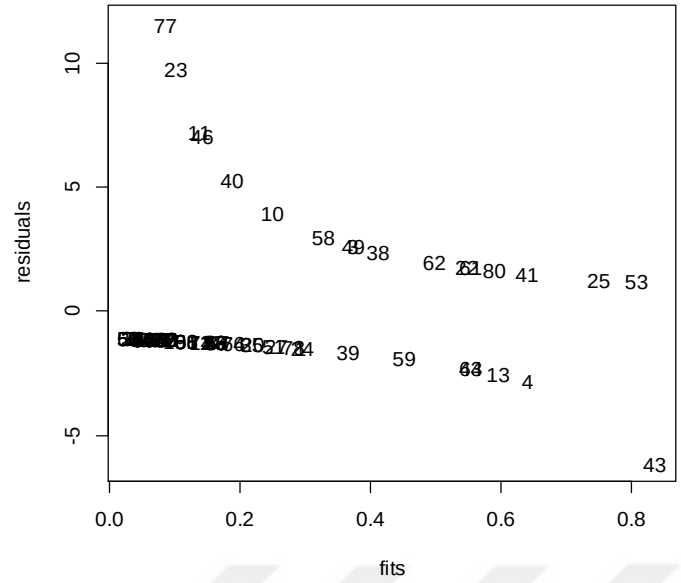

Figure 4.3 Plot of fitted values versus residuals for modified kyphosis data

Table 4.3 Results for modified kyphosis data

Method		**Standardized Pearson Residuals**	**Cook's Distance**	**Pearson chi-square**	**Deviance**
Traditional					
	Low Cut-off	-1.960			
	High Cut-off	1.960	1.000	3.840	3.840
	Influential points	11, 23, 40, 43, 46, 77	None	11, 23, 40, 43, 46, 77	11, 23, 43, 46, 77
JaB					
	Low Cut-off	-1.379			
	High Cut-off	2.673	0.203	4.846	3.612
	Influential points	4, 23, 43, 77	43	11, 23, 43, 46, 77	11, 23, 43, 46, 77

4.3.1.4 Coronary Heart Disease Data

Coronary heart disease (CHD) data set relating age to the presence or absence of coronary disease is one of the most used data in logistic regression analysis. Consider the heuristic CHD data have 100 observations. As seen in Figure 4.4, points 5, 16, 23, 91 and 97 tend to spoil the overall impression of the data, and they might be influential. Our results in Table 4.4 reveal that traditional and JaB methods show nearly same performance except for Cook's distance. For this measure, while JaB can flag all mentioned observations as influential, traditional Cook's distance did not flag

any point. The reason for this is the same as in the previous example. Because of the fitted probabilities being close to the boundaries, the leverage values for these points tend to be small causing small Cook's distances.

Table 4.4 Results for CHD data

Method		Standardized Pearson Residuals	Cook's Distance	Pearson chi-square	Deviance
Traditional					
	Low Cut-off	-1.960			
	High Cut-off	1.960	1.000	3.840	3.840
	Influential points	5, 16, 23, 91, 97	None	5, 16, 23, 91, 97	5, 16, 97
JaB					
	Low Cut-off	-1.804			
	High Cut-off	2.183	0.093	3.909	3.220
	Influential points	5, 16, 91, 97	5, 16, 23, 91, 97	5, 16, 23, 91, 97	5, 16, 23, 91, 97

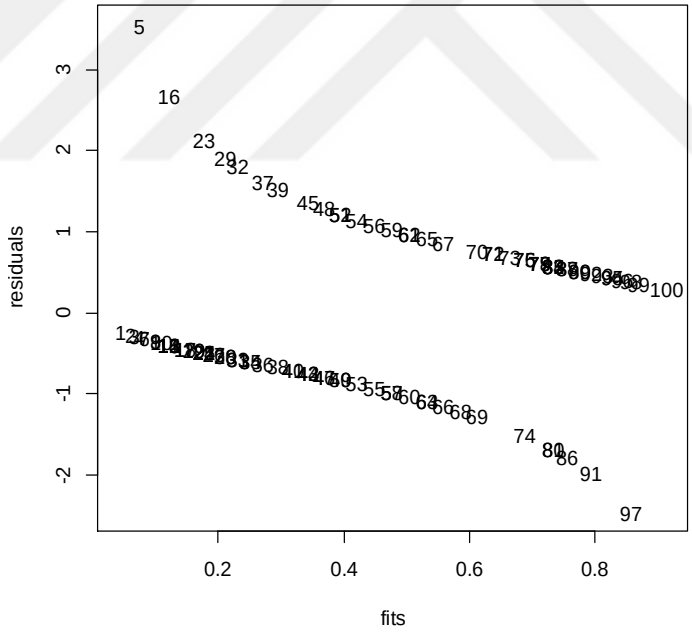

Figure 4.4 Plot of fitted values versus residuals for CHD data

4.4 Simulation Results

A simulation study was conducted to further explore the performances of the measures, and to strengthen the conclusions of the real data examples. The sample sizes were chosen as 30, 50 and 100. The simulated data consist of a dependent

variable and an explanatory variable. For all the sample sizes, half of the explanatory variable X were generated from Uniform (10, 100), and the other half were generated from Uniform (50, 500). The dependent variable Y were generated from Binomial ($n = n/2$, $p = 0.2$) for the first half of explanatory variable, and Binomial ($n = n/2$, $p = 0.8$) for the second half of explanatory variable. That is, the Y values tend to have value 1 for larger X values, and value 0 for smaller X values. The probability parameter $p = 0.2$ and $p = 0.8$ show that the data set is likely to have potential influential observation about 20% on the Y space. In addition, we changed the last values of X and Y as 495 and 0, respectively so that this value will be significantly influential on the regression fits, and the masking effect will appear between this deliberately inserted influential observation and other potential influential observations. This design was repeated for $M = 500$ times, and each time we controlled whether this observation was flagged as influential or not. For each sample size, $B = 3100$ resamples were drawn in each resampling operation so that for each data point, roughly 1000 resamples without that point were produced. The results are given in Tables 4.5 - 4.7 for sample size $n = 30$, 50 and 100, respectively. The average number of points flagged as influential for each simulation is recorded as "Average no. of points" in the tables. For deliberately inserted data point, the detection rate for all simulations is recorded as "Percent of times point identified". The standard errors for Low - High cut-offs and for flagged influential observations are given in brackets below.

JaB performed better than traditional method for all diagnostic measures and for all sample sizes by means of the average number of points flagged and detection rate of deliberately inserted observation. Tables show that detection rate of traditional Cook's distance is always very low compared to other measures. But, the improvement gained using JaB is so significant that it even performs the best with smaller standard errors compared to other diagnostic statistics. As in real world examples, JaB is more robust to masking effect than traditional method for all diagnostics used especially for Cook's distance.

Table 4.5 Simulation results where $n = 30$

Method	Standardized Pearson Residuals	Cook's Distance	Pearson chi-square	Deviance
Traditional				
Average no. of points	1.056	0.552	1.056	0.705
(SE)	(0.014)	(0.009)	(0.014)	(0.011)
Percent of times point identified	0.739	0.539	0.739	0.616
JaB				
Low cut-off	-1.956			
(SE)	(0.005)			
High cut-off	1.713	0.304	3.272	3.002
(SE)	(0.003)	(0.001)	(0.010)	(0.005)
Average no. of points	1.075	1.508	1.240	1.281
(SE)	(0.012)	(0.009)	(0.13)	(0.012)
Percent of times point identified	0.803	0.977	0.951	0.861

Table 4.6 Simulation results where $n = 50$

Method	Standardized Pearson Residuals	Cook's Distance	Pearson chi-square	Deviance
Traditional				
Average no. of points	1.815	0.289	1.815	1.105
(SE)	(0.020)	(0.009)	(0.020)	(0.015)
Percent of times point identified	0.846	0.267	0.846	0.728
JaB				
Low cut-off	-2.039			
(SE)	(0.005)			
High cut-off	1.704	0.166	3.269	2.949
(SE)	(0.003)	(0.001)	(0.010)	(0.005)
Average no. of points	1.890	2.513	2.197	2.219
(SE)	(0.015)	(0.012)	(0.015)	(0.014)
Percent of times point identified	0.947	1.000	0.995	0.960

Table 4.7 Simulation results where n = 100

Method	Standardized Pearson Residuals	Cook's Distance	Pearson chi-square	Deviance
Traditional				
Average no. of points	3.785	0.037	3.785	2.233
(SE)	(0.025)	(0.003)	(0.025)	(0.024)
Percent of times point identified	0.971	0.037	0.971	0.831
JaB				
Low cut-off	-2.189			
(SE)	(0.005)			
High cut-off	1.653	0.078	3.243	2.910
(SE)	(0.002)	(0.001)	(0.008)	(0.004)
Average no. of points	4.420	4.831	4.607	4.598
(SE)	(0.0024)	(0.014)	(0.017)	(0.018)
Percent of times point identified	1.000	1.000	1.000	1.000

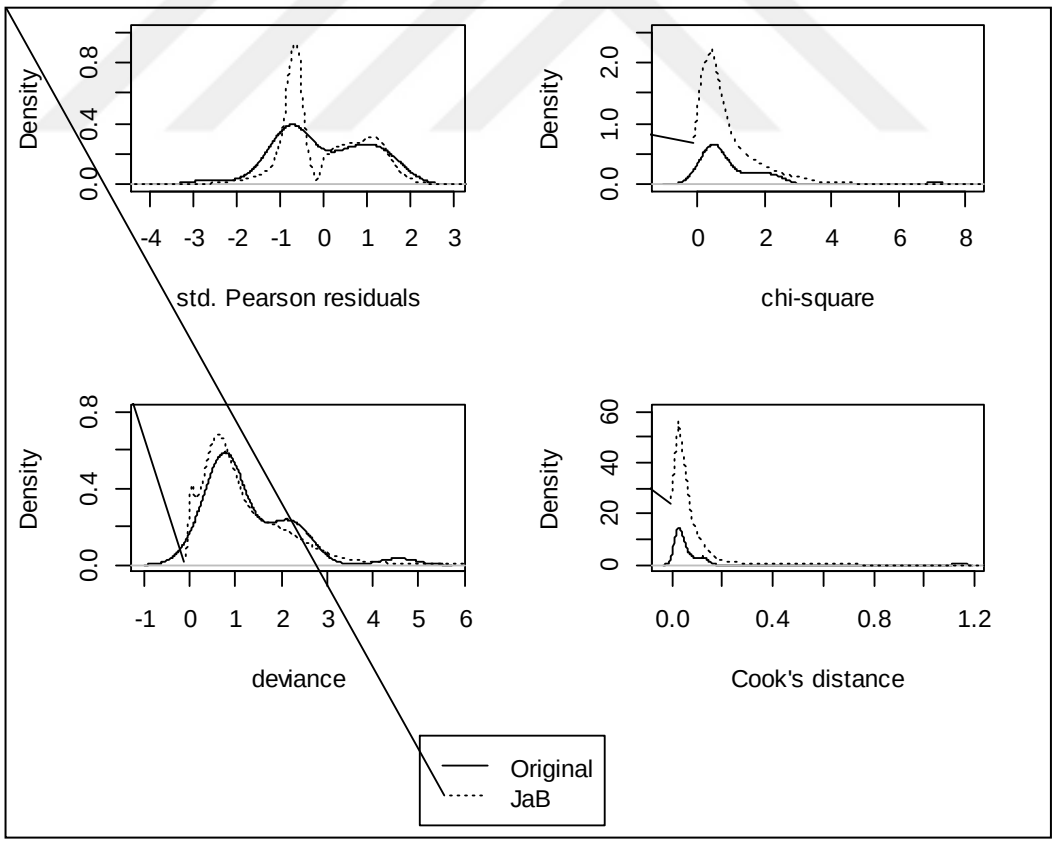

Figure 4.5 Density plots of JaB and original influence measures for the sample size n = 30

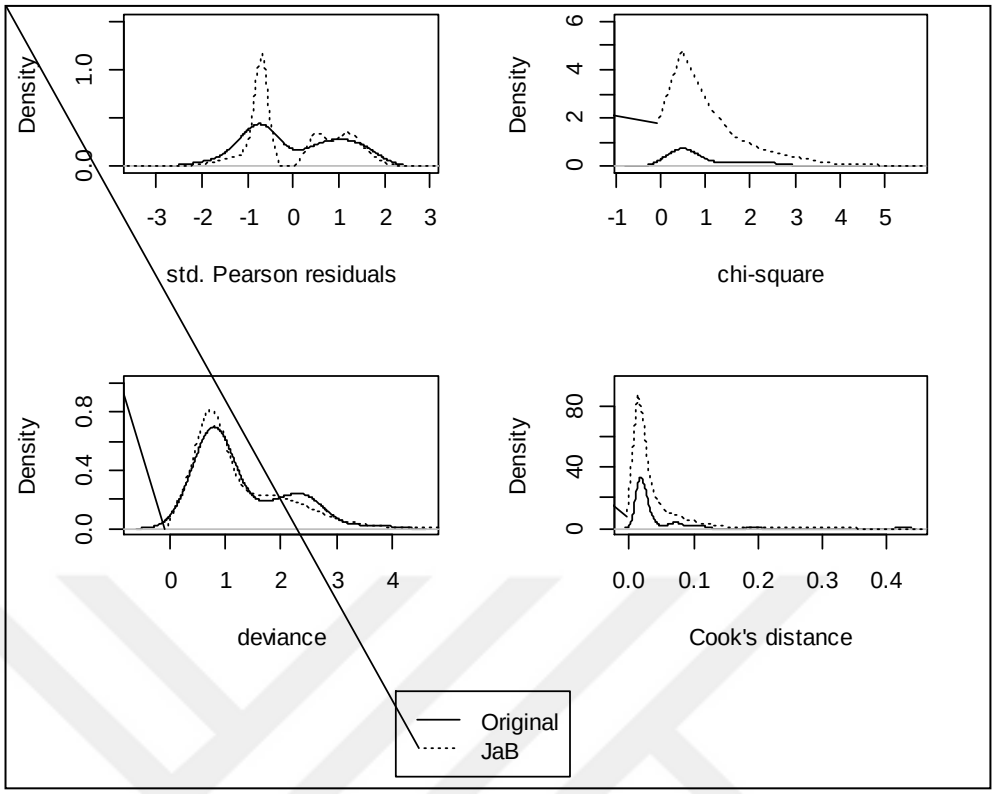

Figure 4.6 Density plots of JaB and original influence measures for the sample size n = 50

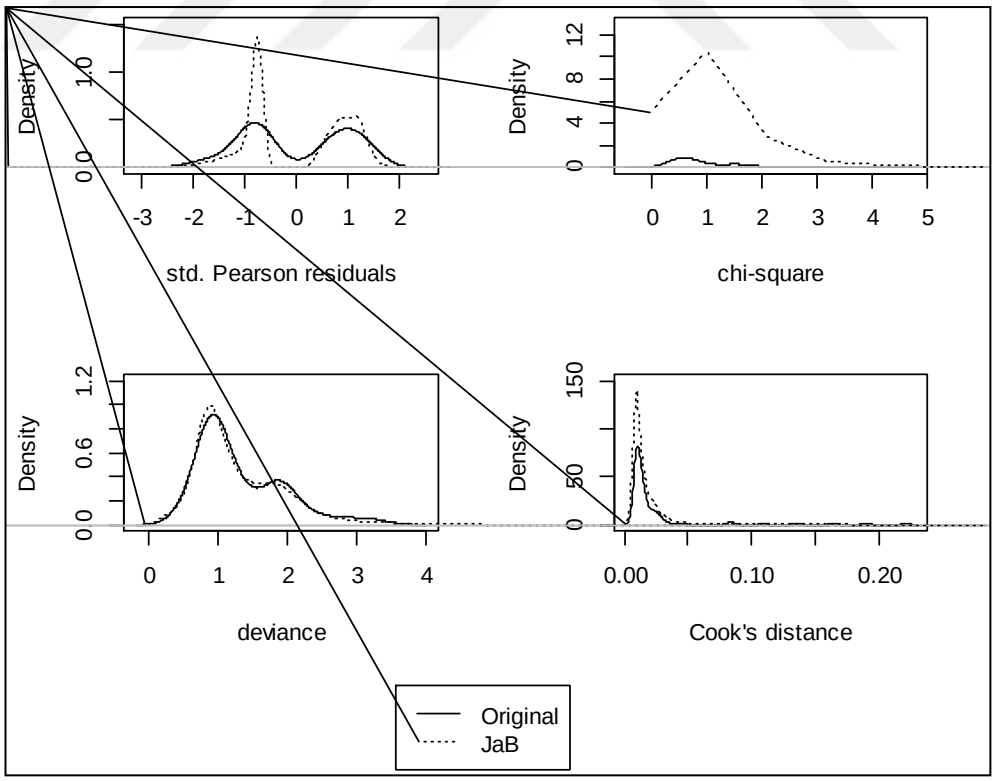

Figure 4.7 Density plots of JaB and original influence measures for the sample size n = 100

Figures 4.5-4.7 illustrate both JaB sampling distributions and the distributions of the true values of the measures obtained from the sample. Under the smallest sample size, the JaB sampling distribution for change in the deviance statistic fits better to the distribution of its true values (see Figure 4.5). While JaB version of Cook's distance's results are better than other statistics for n = 30, its performance as the sampling distribution approximation is poor. But, as it is clearly seen from Figures 4.6 and 4.7, the performance gets better as the sample size increases.

4.4 Conclusion and Discussion

In this study, we extend JaB method to logistic regression for four diagnostic measures: Standardize Pearson residuals, change in the deviance, change in the Pearson chi-square, and Cook's distance. While Cook's distance shows the effect on the model fit, the others show the points not well explained by the model. For all measures, JaB performs better than traditional versions, but it is the most effective for Cook's distance. The traditional version of Cook's distance does not generally behave well for logistic regression models. There may be two possible reasons for that. The first is the fitted probability of the corresponding point. If this probability is at the boundary its leverage will not act as a distance and, unlike the expected, it will be small causing Cook's distance not to detect the point. The other problem is masking. According to our results Cook's distance is the one most affected by masking which is quite a challenging not only for logistic regression but also for all members of generalized linear models. Both the real world data results and simulations show that if we use the cut-offs from the JaB sampling distribution, Cook's distance will be robust to those problems. Its good behavior under JaB is definitely based on the calculation of cut-offs. As described by Martin and Roberts (2010), the rationale behind JaB is to generate a "null" bootstrap distribution of the corresponding parameter under the hypothesis that the *i*th data point is not influential. As they propose since the *i*th data point is not present in any of the resamples from which the bootstrap distribution is generated, it cannot exert influence, and thus the distribution generated is free from its influence. This is similar to the conditional influence proposed by Lawrence (1995). Conditional

influence shows the real effect of the point after deleting the other point causing masking. It seems that calculating cut-offs free from influential point not only makes Cook's distance robust to masking effect but also makes it robust to the problem of fitted probabilities at the boundary. As a solution to masking effect, Cook and Weisberg (1982) proposed the Cook's distance calculated by deleting group of observations. However, with JaB there is no need for group deletion.

For change in the Pearson chi-square and change in the deviance statistics, chi-squared distribution assumption is valid for the grouped observations where the number of replicates is very large. For our data sets, this assumption is invalid since we have ungrouped points. Using the cut-off from the χ_1^2 , however, gives some idea about the size of the points. But, we don't need such a distributional assumption for JaB which makes it very suitable for the ungrouped observations.

Even though all the measures studied in this paper are used to detect influential observations, they measure different effects. To detect influential observations, it is not wise to lay our decision on one measure. Instead, more measures should be calculated to understand if these points are the ones not well explained by the model or the ones which affect model fit. The accessibility of the traditional versions of these measures through all statistical packages would be the reason to prefer them. However, considering the pros of the proposed method and the increasing technology, it is worth trying new approach.

CHAPTER FIVE
CONCLUSION

Our main purpose in this dissertation is to develop resampling based methods to identify influential observations in generalized linear models. Traditional methods are successful for identification of influential observations under the assumptions of large sample theory and normal distribution. But, in case of non–normal error distributions or in case of small sample size, these methods may not be sufficient since they always use the same quantity as a cut-off point with the same sample sizes, irrespective of what might be known or suspected about the data generating process. An influential observation arising from a certain underlying distribution does not have to be influential with respect to other underlying distributions, or a non-influential observation for a certain underlying distribution may be influential with respect to other underlying distributions. Hence, the observations detected as influential by the traditional methods under the different distributions may not be reasonable. In addition, the cut-offs approximated by the large sample theory may not be accurate for small samples. Besides, traditional methods fail to provide satisfactory results to flag actual influential observations because of two possible phenomena called masking and swamping effect. One of the remedy to overcome these problems is to use well known resampling methods, e.g., the jackknife-after-bootstrap.

The JaB method has a number of advantages compared to traditional methods. For instance, JaB tries to approximate the sampling distribution and calculates the cut-off points, regardless of sample size. It allows for asymmetry in the error distribution. Moreover, it combines the model information with the values of the diagnostic statistics to approximate the sampling distribution, while the traditional methods do not take into account the model information when the cut-offs are calculated. All of these advantages make JaB a good alternative against the traditional methods.

Throughout this dissertation, we first proposed robust BCa JaB method which combines the BCa bootstrap method with JaB. In this method, we used robust estimators and calibrated standard normal quantiles for the calculations of BCa confidence limits. This calibrated quantile is based on the significance level calculated from the BCa method (BCa-level). Even though our BCa-level adjustment came at an apparent theoretical cost of sacrificing second-order correctness of the resultant interval endpoints, the intervals themselves remained second-order accurate, and the choice appears to have been justified by the excellent empirical performance of the proposed method in a variety of difficult real-world data sets and consistently within our simulation study. In particular, our proposed method has been shown to work well even in the presence of masking effects.

Secondly, we proposed delete-2 JaB method to refine the cut-offs for single case deleted Cook's distance under the masking problem. The results for all the data sets and simulation study show that with the refined cut-offs which are robust to both masking and swamping effects, Cook's distance successfully flag influential observations. We have only worked on refining the cut-offs for Cook's distance because of its sensitivity against masking and swamping effects. However, this estimation method can also be used on other influence measures to improve their performance.

Finally, we extended JaB method to detect influential observations in binary logistic regression model for four diagnostic measures: Standardize Pearson residuals, change in the deviance, change in the Pearson chi-square, and Cook's distance. Both the real world data results and simulations show that, for all measures, our proposed method performs better than traditional versions, but it is the most effective for Cook's distance.

As a summary, in this dissertation, we have developed JaB methods to identify influential observations and applied it to the linear and binary logistic regression models. It should be noted that our proposed methods can also be extended to all the other linear and non-linear generalized regression models.

REFERENCES

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics*. New York: Wiley.

Beyaztas, U., & Alin, A. (2013). Jackknife-after-bootstrap method for detection of influential observations in linear regression models. *Communication in Statistics – Simulation and Computation*, *42*, 1256-1267.

Beyaztas, U., Alin, A., & Martin, M. A. (2014). Robust BCa-JaB method as a diagnostic tool for linear regression models. *Journal of Applied Statistics*, *41*, 1593-1610.

Beyaztas, U., & Alin, A. (2014a). Sufficient jackknife-after-bootstrap method for detection of influential observations in linear regression models. *Statistical Papers*, *55*, 1001-1018.

Beyaztas, U., & Alin, A. (2014b). Delete-2 jackknife-after-bootstrap in regression. *Quality and Reliability Engineering International*, *30*, 993-1002.

Beyaztas, U., & Alin, A. (2014c). Jackknife-after-bootstrap as logistic regression diagnostic tool. *Communication in Statistics – Simulation and Computation*, *43*, 2047-2060.

Brown, B.W. (1980). *Prediction analysis in binary data, in Biostatistics Casebook*. New York: Wiley.

Chatterjee, S., & Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, *1*, 379-416.

Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity analysis in linear Regression*. USA: Wiley.

Cook, R.D., & Weisberg, S. (1992). *Residuals and influence in regression*. New York: Chapman & Hall.

Dale, J. R. (1986). Asymptotic normality goodness of fit statistics for sparse product multinomials. *Journal of the Royal Statistical Society Series B, 41,* 48-59.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*, 189-228.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1-26.

Efron, B. (1987). better bootstrap confidence intervals. *Journal of the American Statistical Association, 82,* 171-185.

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of Royal statistical Society, 54,* 83-127.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Finney, D. J. (1947). The estimation from individual records of the relationship between dose and quantile response. *Biometrika, 34,* 320-334.

Freund, R. J. (1979). Multicollinearity etc.: Some "new" examples. *American Statistical Association Proceedings of the Statistical Computing Section,* 111-112.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics, 16,* 927-953.

Hall, P. (1992). *The bootstrap and edgeworth expansion*. Verlag: Springer.

Hosmer, D.W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Imon, A. M. H. R., & Hadi, A. S. (2008). Identification of multiple outliers in logistic regression. *Communication in Statistics – Theory and Methods*, *37*, 1697-1709.

Lawrence, D. E., Birch, J. B., & Chen, Y. (2014). Cluster-based bounded influence regression. *Quality and Reliability Engineering International*, *30*, 97-109.

Martin, M. A., Roberts, S., & Zheng, L. (2010). Delete-2 and delete-3 jackknife procedures for unmasking in regression. *Australian & New Zealand Journal of Statistics*, *52*, 45-60.

Martin, M. A., & Roberts, S. (2010). Jackknife-after-bootstrap regression influence diagnostics. *Journal of Nonparametric Statistics*, *22*, 257-269.

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, *135*, 370-384.

Nurunnabi, A. A. M., Imon, A. H. M. R., & Nasser, M. (2010). Identification of multiple influential observations in logistic regression. *Journal of Applied Statistics*, *37*, 1605-1624.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, *9*, 705-724.

Ryan, T. P. (1997). *Modern regression methods*. New York: Wiley.

Rousseeuw, P. J., & Leroy A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Waltham: Elsevier.

APPENDICES

```
# R codes of Delete-2 Jackknife-after-Bootstrap method for Cook's distance

# For Soil Evaporation data

library(TeachingDemos)

data(evap)

attach(evap)


y <-  Evap

n <- length(y)

B <- 8000

alpha <- 0.05


X <- cbind(1,as.matrix(evap[,4:13]))

Y <- matrix(y, ncol=1)

index <- matrix(c(1:n), ncol=1)

Design.data <- cbind(X, Y, index)

index.original <- c(index)


B.cap <- solve(crossprod(X)) %*% crossprod(X, Y)

P <- X %*% solve(crossprod(X)) %*% t(X)

Y.cap <- P %*% Y

e <- Y - Y.cap

dX <- nrow(X) - ncol(X)

var.cap <- crossprod(e) / (dX)

ei <- as.vector(Y - X %*% B.cap)

pi <- diag(P)
```

```
var.cap.i <- (((dX) * var.cap)/(dX - 1)) -
(ei^2/((dX - 1) * (1 - pi)))
ti <- ei / sqrt(var.cap * (1 - pi))
Ci <- (pi * ti^2) / (ncol(X) * (1 - pi))

boot.vector <- vector("list", )
boot.index <- vector("list", )

for(boot in 1:B) {
        data <- Design.data[sample(n,n,replace=TRUE),]
        dataX <- data[,1:ncol(X)]
        dataY <- data[,ncol(X)+1]
        index.bootstrap <- data[,ncol(X)+2]
        index.simulation <- c(index.bootstrap)

        s.c.dX<-solve(crossprod(dataX))
        B.cap.simulation <- s.c.dX %*% crossprod(dataX, dataY)
        P.simulation <- dataX %*% s.c.dX %*% t(dataX)
        Y.cap.simulation <- P.simulation %*% dataY
        e.simulation <- dataY - Y.cap.simulation
        dX.simulation <- nrow(dataX) - ncol(dataX)
        var.cap.simulation <- crossprod(e.simulation) / (dX.simulation)
        ei.simulation <- as.vector(dataY - dataX %*% B.cap.simulation)
        pi.simulation <- diag(P.simulation)
        var.cap.i.simulation <- (((dX.simulation) * var.cap.simulation)/(dX.simulation - 1)) -
        (ei.simulation^2/((dX.simulation - 1) * (1 - pi.simulation)))
        ti.simulation <- ei.simulation / sqrt(var.cap.simulation * (1 - pi.simulation))
```

```
        Ci.simulation <- (pi.simulation * ti.simulation^2) / (ncol(dataX) * (1 -
        pi.simulation))


        boot.vector[[boot]] <- Ci.simulation

        boot.index[[boot]] <- index.simulation

}

finalresult <- vector("list", )

result <- vector("list", )


for (j in 1:n-1) {

        l=j+1

        for (k in l:n) {

                for( i in 1: B) {


                result[[i]] <- list(outCi.simulation = as.numeric(boot.vector[[i]]),

                influ.obs = !(index.original[j] %in% as.numeric(boot.index[[i]])) &
                !(index.original[k] %in% as.numeric(boot.index[[i]])))

                }

        i.obs <- sapply(result, function(x) {x$influ.obs})

        no.i.obs <- which(i.obs == TRUE)

        data.no.i.obs <- result[no.i.obs]

        D.2.JaB.data <- sapply(data.no.i.obs, function(x) {x$outCi.simulation})

        finalresult <- c(finalresult, list(D.2.JaB.data))

        }

}

data.JaB <- c(unlist(finalresult))

percentile.JaB <- quantile(data.JaB, 1-alpha)

which(Ci > percentile.JaB)
```