DIAGNOSTIC AND INFLUENCE MEASURES IN LINEAR REGRESSION

by

Carlos Wong

B.Sc. Simon Fraser University 1987

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

of

Simon Fraser University

SIMON FRASER UNIVERSITY

August, 1992

APPROVAL

Name: Carlos Wong

Degree: Master of Science

Title of Project: **Diagnostic and Influence Measures in Linear Regression**

Examining Committee: Dr. Alistair H. Lachlan

Chair

____

Dr. Charmaine Dean, Senior Supervisor

____

Dr. Richard Lockhart

____

Dr. Tim Swartz

____

Dr. David Eaves, External Examiner

Date Approved: August 5, 1992

Title of Thesis/[✓ Project]/Extended Essay

Diagnostic and Influence Measures in

Linear Regression

Author:

(signature)

Carlos A. Weng

(name)

Aug. 5, 1992

(date)

Abstract

Linear regression is a simple, powerful, often reasonable tool for modelling the dependence of a response variable upon other factors or conditions. However, incorrect inferences concerning parameters in the model may result if the underlying assumptions are not met. This project considers diagnostic techniques for model checking in linear regression, and influence measures for identifying observations which severely affect the results of the analysis.

These techniques are illustrated by applying them to a Children's Aid Society Expenditures Data. The response variable in this data set is the per child capita expenditure in 44 Ontario counties and districts for Children's Aid Societies in 1980. The dependence of the expenditures by the Children's Aid Society on sixteen socioeconomic factors is investigated. Linear models that fit the data reasonably well are identified.

Acknowledgements

I would like to give my thanks to my senior supervisor Dr. C. Dean for all the assistance, suggestions, dedication and guidance given me throughout this project. I am especially grateful for her advice, both academic and other, and immeasurable encouragement at difficult times.

I would also like to extend my gratitude to my committee members, Dr. T. Swartz, Dr. R. Lockhart and Dr. D. Eaves, for their invaluable suggestions and kindness. They have also been resourceful and available throughout my course of study. My appreciation is offered to Dr. Swartz for always having an open door whenever I need his advice.

I am also thankful to all the faculty members of the Department of Mathematics and Statistics and especially Dr. M. A. Stephens for his expertise and friendliness.

Finally, I would like to thank all the departmental staff. A special thanks goes to Sylvia for her patience and helpfulness.

Contents

List of Tables

List of Figures

CHAPTER 1
INTRODUCTION

Linear Regression methods have been used for a long time as a tool for the investigation of the dependence of a response variable upon various conditions. There are many reasons for its popularity. It is easily understood and computationally simple. Because of the linear structure, the mathematics involved in fitting and making inferences in a linear model is tractable and simple. Most important of all, linear regression very often provides an adequate approximation to the underlying model and it can be a powerful tool when used properly. It can be applied not only to situations where linear dependencies exist, but to a variety of other situations.

Like many scientific methods, linear regression is applicable only if certain assumptions are satisfied, else incorrect inferences may result. This project describes test statistics and graphical procedures useful for checking the assumptions of the linear regression model.

1.1 Basic Assumptions and the Theory of Linear Regression Models

The linear regression model that will be used throughout this project is:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1, \ldots, n.$$

This can be written compactly in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is an $n \times 1$ observable random vector,

$\mathbf{X}$ is an $n \times p$ fixed design matrix of known constants describing the conditions upon which $\mathbf{Y}$ depends,

$\boldsymbol{\beta}$ is an $p \times 1$ vector of unknown parameters,

and $\boldsymbol{\varepsilon}$ is an $n \times 1$ unobservable random error vector.

Very often, the model contains a constant term so, for example, $x_{i1} = 1$.

It is assumed here that $rank(\mathbf{X}) = p$. It is also assumed that the means of the y_i's can be expressed as linear functions of the unknown parameters $\beta_1, \ldots, \beta_p$, hence the name linear model. The usual additional assumption imposed on the model is that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. This simple looking expression implies

(i) that the ε_i have zero mean,

(ii) that the ε_i are independently distributed; that is, the observed y_i are independent of each other,

(iii) homoscedasticity; that is, the variance of y is the same regardless of the values of $x_1, \ldots, x_p$ at which the observation is taken, and

(iv) that the y_i are normally distributed.

The following is a brief review of the inference associated with the linear regression model. (See, for example, Christensen [1987], Draper and Smith [1981], and Graybill

[1976].)

The least squares estimate of $\boldsymbol{\beta}$, usually denoted by $\hat{\boldsymbol{\beta}}$, is defined to be the value of $\boldsymbol{\beta}$ that minimizes the quantity $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Since $rank(\mathbf{X}) = p$, $\mathbf{X}'\mathbf{X}$ is nonsingular and is invertible. In this case, the least squares estimate of $\boldsymbol{\beta}$ is unique and unbiased and is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Consequently, all linear functions of $\boldsymbol{\beta}$, $\boldsymbol{\lambda}'\boldsymbol{\beta}$, are estimable † with estimate $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$. In particular, the estimate of $\mathbf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ is $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. From the fact that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, it can be shown that $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ is the unique best linear unbiased estimate of $\boldsymbol{\lambda}'\boldsymbol{\beta}$, provided $\sigma^2 > 0$. Assuming that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, then $\boldsymbol{\Lambda}'\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\Lambda}'\boldsymbol{\beta}, \sigma^2\boldsymbol{\Lambda}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\Lambda})$ for any matrix $\boldsymbol{\Lambda}$.

An unbiased estimate of σ^2 is given by the quadratic form $\hat{\sigma}^2 = \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}}{n-p}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The matrix $\mathbf{H}$ is sometimes called the hat matrix. It is also known as the projection matrix onto $\mathbf{X}$ since $\mathbf{HY}$ is the orthogonal projection of $\mathbf{Y}$ onto the space spanned by the columns of $\mathbf{X}$. (In general, we will use $C(\mathbf{Z})$ to denote the space spanned by the columns of $\mathbf{Z}$ and $\mathbf{H_Z}$ to denote $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ for any matrix or column vector $\mathbf{Z}$.) Assuming $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, it can be shown that $\hat{\sigma}^2$ is a minimum variance unbiased estimate of σ^2 and that $\frac{\hat{\sigma}^2(n-p)}{\sigma^2} = \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}}{\sigma^2} \sim \chi^2(n-p)$.

Let $\mathbf{Z}$ be an $n \times r$ matrix such that $C(\mathbf{Z}) \subseteq C(\mathbf{X})$ with $rank(\mathbf{Z}) = r \leq p$. Assume that the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is being considered. Then the validity of the reduced model $\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ can be tested with the statistic

$$F = \frac{\mathbf{Y}'(\mathbf{H}-\mathbf{H_Z})\mathbf{Y}/rank(\mathbf{H}-\mathbf{H_Z})}{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}/rank(\mathbf{I}-\mathbf{H})} = \frac{\mathbf{Y}'(\mathbf{H}-\mathbf{H_Z})\mathbf{Y}/(p-r)}{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}/(n-p)}.$$

The distribution of F is $\mathcal{F}(p-r, n-p, \boldsymbol{\beta}'\mathbf{X}'(\mathbf{H}-\mathbf{H_Z})\mathbf{X}\boldsymbol{\beta}/2\sigma^2)$. If the reduced model is correct, the noncentrality parameter reduces to 0. Since a nonzero noncentrality parameter

† $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable if $\boldsymbol{\lambda}' = \boldsymbol{\rho}'\mathbf{X}$ for some vector $\boldsymbol{\rho}$.

shifts the $\mathcal{F}$ distribution to the right, we reject the null hypothesis H_0 : *the reduced model is correct* with a significance level of α if $F > \mathcal{F}(1-\alpha; p-r, n-p, 0)$. The hypothesis $H_0 : \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_k} = 0$ with $i_1, \ldots, i_k \in \{1, \ldots, p\}$ can be tested by setting $\mathbf{Z}$ to be the resulting matrix after the $i_1, \ldots, i_k$ columns of $\mathbf{X}$ are deleted.

1.2 Problems with Linear Regression Fitting

Although linear regressions are easily performed, it is not so easy to justify the correctness of the conclusions that might be drawn from them. A deviation from the assumed model may alter the results substantially. Worst of all, many of these problems may go unnoticed unless further detailed analyses are performed.

One deviation from the linear model is that the dependence of the response variable upon the independent variables is not linear. As a result, the theory developed for the linear model, and hence the conclusion drawn from it, is irrelevant and incorrect. This problem can sometimes be solved by transformation of the independent or dependent variables (Atkinson [1985]). Another departure from the model is that $E(\mathbf{Y}) \notin C(\mathbf{X})$. In some situations where $E(\mathbf{Y})$ depends on the values $x_1, \ldots, x_p$ in a non-linear manner, a transformation may linearize the problem if the transformed variates are normally distributed.

A commonly violated assumption is $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. In this case, the tests and confidence regions constructed by the ordinary least squares method will be incorrect. If the deviation is such that $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$ where $\mathbf{V}$ is some known positive definite matrix, then generalized least squares is used instead. (For a discussion of generalized least squares, see Graybill [1976].) A special case of this occurs when $\mathbf{V}$ is diagonal, so a weighted regression analysis is appropriate. Another cause of incorrect confidence regions and tests is the violation of the normality assumption. Neter, Wasserman, and Kutner [1990; Section 4.3] state that small departures from normality do not create serious problems, but large departures, however, do.

Conclusions can also be affected by the presence of outliers – extreme observations that are significantly 'different' from the rest of the data set; these observation arouse suspicion about the validity of the underlying distribution. Outliers can have undue influence on the regression line. They also may cause σ^2 to be overestimated. Sometimes outliers are the result of mistakes in encoding or recording the data, and the model can be rectified by correcting the mistakes or by discarding the observations associated with the outliers. However, such explanations may not be available. Outliers may contain valuable information; discarding them in this situation is inappropriate.

Multicollinearity in the design matrix itself may also cause problems. This occurs when some of the columns of $\mathbf{X}$ are linearly or nearly linearly dependent to each other. An exact linear dependency among the columns of $\mathbf{X}$ implies that $\mathbf{X}'\mathbf{X}$ is noninvertible and that $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is not necessarily estimable. In particular, the least squares estimate of $\boldsymbol{\beta}$ is no longer unique. Fortunately, the columns of $\mathbf{X}$ seldom exhibit exact linear dependencies. However, nearly linear dependencies are not rare. In these cases, the variance of the estimates of some linear functions of $\boldsymbol{\beta}$ may be inflated.

Influential observations may also give rise to misleading results. These are data points located in such a way that a change in their values will affect the regression line (i.e. $\hat{\boldsymbol{\beta}}$) substantially, regardless of the fact that they might make up only a very small portion of the data. It is important to identify influential observations and to note their influence. Influential observations deserve further attention from the experimenter.

1.3 Purpose and Outline of this Project

The purpose of this project is to present and investigate statistical methods that are used to detect the various departures from the assumptions underlying the linear model, as discussed in the last section. The basic theory of residuals will be presented in Section 2.1. Diagnostic procedures based on residuals will be discussed in subsequent sections of Chapter 2. In particular, methods for detecting outliers will be discussed in Section 2.2; how to determine whether the addition of further independent variables significantly improves the model is discussed in Section 2.3; tests for non-normal errors are given in Section 2.4, and methods for the detection of some special cases of heteroscedasticity in Section 2.5. Techniques for identifying influential observations will be treated in the last section of Chapter 2. A discussion of transformations, variable selection and multicollinearity will be covered in the three sections of Chapter 3 in that order. A data set on Children's Aid Society expenditures will be introduced in Section 4.1. The rest of Chapter 4 will be devoted to an analysis of this data with special consideration to illustration of the statistical methods presented in Chapters 2 and 3. Conclusive remarks are given in Chapter 5.

CHAPTER 2
DIAGNOSTICS AND INFLUENCE ANALYSIS

Most plots and tests in regression analysis are designed to either (i) criticize the model fitted, or (ii) criticize or examine any abnormality of the data. Procedures that aim for the former and latter tasks are called (Weisberg [1983]) diagnostics analysis and influence analysis.

Diagnostic procedures for model criticism are usually statistics or plots designed to check the various assumptions imposed upon the model. The following summarizes the properties that Weisberg [1983] proposed a good diagnostic should have.

(D1) The behavior of a diagnostic procedure should be known, at least approximately, both under the assumed model and other models with preferably only one assumption modified.

(D2) The diagnostics can be derived by parameterizing the assumptions so that the problem of criticism can be investigated with significance test.

(D3) Diagnostic methods should not be computationally intensive, with respect to current computing facilities.

(D4) Diagnostics should be graphical or have graphical equivalents.

(D5) Diagnostic procedures should suggest remedial action.

The idea of influence analysis is to study the changes in the outcome of the regression when small perturbations are introduced in the data. As with regression diagnostics, Weisberg [1983] also proposed that a good influence measure possess certain properties as summarized below.

(I1) The perturbation scheme should be well defined (eg. the deletion of a case).

(I2) Influence measures must reter to some specific aspect of the problem. They must measure something interesting.

(I3) Influence measures should depend on the sample at hand.

(I4) If a vector norm is used to summarize influential information provided by a vector, (a) it should possess desirable statistical properties, (b) it should depend on the specific aspect of the analysis of interest, and (c) the resulting values should be calibrated with respect to some external reference.

However, seldom does a diagnostic or influence procedure possess all the above properties. In this chapter, we will present some diagnostic methods that criticize different aspects of the model and possess some of the properties in D1 to D5. Since most diagnostics are functions of the residuals, the first section will be devoted to the theory of residuals. Statistical procedures concerning the identification of outliers, the significance of the addition of more independent variables to the current model, checking for non-normality of errors, and non-constant variance of errors will be presented in Section 2.2 through Section 2.5 in that order. Some influence procedures that possess some of the properties in I1 to I4 are discussed in the last section.

2.1 Residuals

The ordinary residual is defined by

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \text{ with } \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})).$$

The usual use of the residuals is to check for violations of a given standard regression model as described in Section 1.1. However, since $Var(e_i) = \sigma^2(1 - h_{ii})$ (where h_{ii}, usually called the leverage, is the i^{th} diagonal entry of $\mathbf{H}$), the ordinary residuals have heteroscedastic variance. Thus, the assumption of homoscedastic variance of the errors may not be properly checked by looking at a plot of the ordinary residuals. Furthermore, Christensen [1987; Chapter 13] noted that since some normality tests are sensitive to inequality of variances, using the ordinary residuals may lead to a non-normal conclusion about the errors even though they are actually normally distributed. For these reasons, it is more appropriate to use the studentized residuals (also called standardized residuals):

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \qquad \text{where} \qquad \hat{\sigma}^2 = \frac{SSE}{n-p} = \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}}{n-p}.$$

It can be shown that, assuming that $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $r_i^2/(n-p)$ has a Beta distribution with parameter $1/2$ and $(n-p-1)/2$. (See Cook [1982] and Ellenberg [1973].) It follows that $E[r_i] = 0$ and $Var[r_i] = 1$. Moreover, Cook [1982], applying results from Ellenberg [1973], showed that $Cov[r_i, r_j] = -h_{ij}/[(1-h_{ii})(1-h_{jj})]^{1/2}$ for $i \neq j$. The r_i's are therefore not independent of each other. The last statement is clear from the fact that $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$.

An alternative is to use the standardized predicted residuals derived as follows. Let $\mathbf{Y}_{[i]}$ and $\mathbf{X}_{[i]}$ be, respectively, the corresponding $\mathbf{Y}$ and $\mathbf{X}$ with the i^{th} case deleted.

Similarly, let $\boldsymbol{\beta}_{[i]}$, $\mathbf{H}_{[i]}$, $SSE_{[i]}$, and $\hat{\sigma}^2_{[i]}$ be defined

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{[i]} &= (\mathbf{X}'_{[i]}\mathbf{X}_{[i]})^{-1}\mathbf{X}'_{[i]}\mathbf{Y}_{[i]}, \\
\mathbf{H}_{[i]} &= \mathbf{X}_{[i]}(\mathbf{X}'_{[i]}\mathbf{X}_{[i]})^{-1}\mathbf{X}'_{[i]}, \\
SSE_{[i]} &= \mathbf{Y}'_{[i]}(\mathbf{I}-\mathbf{H}_{[i]})\mathbf{Y}_{[i]}, \\
\hat{\sigma}^2_{[i]} &= \frac{SSE_{[i]}}{n-p-1}.
\end{aligned}$$

Note that updating formula are available so that the above statistics can be obtained without actually deleting any cases and performing another regression. (Some of the updating formulae are given by Cook and Weisberg [1982] and Atkinson [1985] without proofs. I provide detailed proofs of them below.) First notice that, if $\mathbf{x}'_i$ represents the i^{th} row of $\mathbf{X}$,

$$(\mathbf{X}'_{[i]}\mathbf{X}_{[i]})_{jk} = \sum_{\substack{l=1\\ l\neq i}}^{n} x_{lj}x_{lk} = \sum_{l=1}^{n} x_{lj}x_{lk} - x_{ij}x_{ik} = (\mathbf{X}'\mathbf{X}-\mathbf{x}_i\mathbf{x}'_i)_{jk}$$

$$\Rightarrow \quad \mathbf{X}'_{[i]}\mathbf{X}_{[i]} = \mathbf{X}'\mathbf{X}-\mathbf{x}_i\mathbf{x}'_i$$

and, similarly, $\mathbf{X}'_{[i]}\mathbf{Y}_{[i]} = \mathbf{X}'\mathbf{Y}-\mathbf{x}_i y_i$. Therefore,

$$\begin{aligned}
&(\mathbf{X}'_{[i]}\mathbf{X}_{[i]})\left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}\right) \\
&= (\mathbf{X}'\mathbf{X}-\mathbf{x}_i\mathbf{x}'_i)\left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}\right) \\
&= (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i} - \mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1} \\
&\quad - \frac{\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i} \\
&= \mathbf{I} - \mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1} + \frac{\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i} \\
&= \mathbf{I} - \mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1} + \frac{\mathbf{x}_i[1-\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i]\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i} \\
&= \mathbf{I} - \mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{I}.
\end{aligned}$$

That is, $(\mathbf{X}'_{[i]}\mathbf{X}_{[i]})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}$. Then

$$
\begin{aligned}
(1)\qquad \hat{\boldsymbol\beta}_{[i]} &= (\mathbf{X}'_{[i]}\mathbf{X}_{[i]})^{-1}\mathbf{X}_{[i]}\mathbf{Y}_{[i]} \\
&= \left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}\right)(\mathbf{X}'\mathbf{Y} - \mathbf{x}_i y_i) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}}{1-h_{ii}} \\
&\quad - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i}{1-h_{ii}} \\
&= \hat{\boldsymbol\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i(1-h_{ii}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'\hat{\boldsymbol\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i h_{ii} y_i}{1-h_{ii}} \\
&= \hat{\boldsymbol\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i h_{ii} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \hat{y}_i + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i h_{ii} y_i}{1-h_{ii}} \\
&= \hat{\boldsymbol\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}.
\end{aligned}
$$

$$
\begin{aligned}
(2)\qquad \hat{y}_{[i]j} &= \mathbf{x}_j'\hat{\boldsymbol\beta}_{[i]} = \mathbf{x}_j'\left(\hat{\boldsymbol\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}\right) \\
&= \hat{y}_j - \frac{\mathbf{x}_j'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}} = \hat{y}_j - \frac{h_{ij}e_i}{1-h_{ii}}.
\end{aligned}
$$

$$
\begin{aligned}
(3)\quad SSE_{[i]} &= \mathbf{Y}'_{[i]}(\mathbf{I} - \mathbf{H}_{[i]})\mathbf{Y}_{[i]} \\
&= \mathbf{Y}'_{[i]}(\mathbf{I} - \mathbf{X}_{[i]}(\mathbf{X}'_{[i]}\mathbf{X}_{[i]})^{-1}\mathbf{X}'_{[i]})\mathbf{Y}_{[i]} \\
&= \mathbf{Y}'_{[i]}\mathbf{Y}_{[i]} - \mathbf{Y}'_{[i]}\mathbf{X}_{[i]}(\mathbf{X}'_{[i]}\mathbf{X}_{[i]})^{-1}\mathbf{X}'_{[i]}\mathbf{Y}_{[i]} \\
&= \mathbf{Y}'\mathbf{Y} - y_i^2 - \mathbf{Y}'_{[i]}\mathbf{X}_{[i]}\hat{\boldsymbol\beta}_{[i]} \\
&= \mathbf{Y}'\mathbf{Y} - y_i^2 - (\mathbf{Y}'\mathbf{X} - y_i\mathbf{x}_i')\left(\hat{\boldsymbol\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}\right) \\
&= \mathbf{Y}'\mathbf{Y} - y_i^2 - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol\beta} + \frac{\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}} + y_i\mathbf{x}_i'\hat{\boldsymbol\beta} - \frac{y_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}} \\
&= \mathbf{Y}'\mathbf{Y} - y_i^2 - \mathbf{Y}'\mathbf{H}\mathbf{Y} + \frac{\mathbf{x}_i'\hat{\boldsymbol\beta} e_i}{1-h_{ii}} + y_i\mathbf{x}_i'\hat{\boldsymbol\beta} - \frac{y_i h_{ii} e_i}{1-h_{ii}} \\
&= \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} - \frac{y_i^2(1-h_{ii}) - \hat{y}_i e_i - y_i\hat{y}_i(1-h_{ii}) + y_i h_{ii} e_i}{1-h_{ii}}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y} - \frac{y_i^2 - y_i^2 h_{ii} - \hat{y}_i y_i + \hat{y}_i^2 - y_i \hat{y}_i + y_i \hat{y}_i h_{ii} + y_i^2 h_{ii} - y_i \hat{y}_i h_{ii}}{1-h_{ii}} \\
&= \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y} - \frac{(y_i - \hat{y}_i)^2}{1-h_{ii}} \\
&= SSE - \frac{e_i^2}{1-h_{ii}}.
\end{aligned}
$$

Now let $e_{[i]} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{[i]}$, which is sometimes called the PRESS residual. With the help of the updating formula, we have

$$e_{[i]} = y_i - \hat{y}_i + \frac{h_{ii}e_i}{1-h_{ii}} = \frac{e_i}{1-h_{ii}},$$

and therefore

$$E[e_{[i]}] = 0 \qquad \text{and} \qquad Var[e_{[i]}] = \frac{\sigma^2(1-h_{ii})}{(1-h_{ii})^2} = \frac{\sigma^2}{1-h_{ii}}.$$

Based on $e_{[i]}$, the standardized predicted residual is defined to be

$$
\begin{aligned}
r_{[i]} &= \frac{e_{[i]}}{\sqrt{\widehat{Var}[e_{[i]}]}} = \frac{e_{[i]}}{\sqrt{\hat{\sigma}_{[i]}^2/(1-h_{ii})}} \\
&= \frac{e_i}{\sqrt{\hat{\sigma}_{[i]}^2(1-h_{ii})}}, \qquad \text{where} \qquad \hat{\sigma}_{[i]}^2 = \frac{SSE_{[i]}}{n-p-1} = \frac{\mathbf{Y}_{[i]}'(\mathbf{I}-\mathbf{H}_{[i]})\mathbf{Y}_{[i]}}{n-p-1}.
\end{aligned}
$$

The distribution of $r_{[i]}$ can be derived as follows. Since $e_i \sim \mathcal{N}(0, \sigma^2(1-h_{ii}))$, $\frac{e_{[i]}}{\sigma/\sqrt{1-h_{ii}}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}} \sim \mathcal{N}(0,1)$. Furthermore, $\mathbf{Y}_{[i]}'(\mathbf{I}-\mathbf{H}_{[i]})\mathbf{Y}_{[i]}/\sigma^2 \sim \chi^2(n-p-1)$. Since $\hat{\sigma}_{[i]}^2$ does not depend on y_i and $\hat{\sigma}_{[i]}^2$ can be shown to be independent of $\hat{\beta}_{[i]}$, $\hat{\sigma}_{[i]}^2$ is independent of $e_{[i]} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{[i]}$. Therefore,

$$r_{[i]} = \frac{\frac{e_{[i]}}{\sigma/\sqrt{1-h_{ii}}}}{\sqrt{\frac{\mathbf{Y}_{[i]}'(\mathbf{I}-\mathbf{H}_{[i]})\mathbf{Y}_{[i]}}{\sigma^2(n-p-1)}}} \sim t(n-p-1).$$

For this reason, $r_{[i]}$ is also called the studentized predicted residual and is sometimes denoted by t_i.

Note that

$$\begin{aligned}
r_{[i]} &= \frac{e_{[i]}}{\sqrt{\frac{SSE_{[i]}}{(n-p-1)(1-h_{ii})}}} = \sqrt{n-p-1}\,\frac{e_i/(1-h_{ii})}{\sqrt{\frac{SSE-\frac{e_i^2}{1-h_{ii}}}{1-h_{ii}}}} \\
&= \sqrt{n-p-1}\,\frac{e_i}{\sqrt{1-h_{ii}}\sqrt{SSE-\frac{e_i^2}{1-h_{ii}}}} \\
&= \frac{e_i}{\sqrt{1-h_{ii}}\sqrt{\frac{SSE}{n-p}}}\,\frac{\sqrt{n-p-1}}{\sqrt{n-p-\frac{e_i^2}{\frac{SSE}{n-p}(1-h_{ii})}}} \\
&= r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}.
\end{aligned}$$

Therefore, if the studentized residual r_i is available, $r_{[i]}$ can be easily calculated from the above formula. The advantage of using $r_{[i]}$ over r_i is that the distribution of the former is known exactly to be $t(n-p-1)$, which is asymptotically normal. On the other hand, since the distribution of $r_i^2/(n-p)$ is Beta, the distribution of r_i is non-normal. For this reason, we intuitively expect $r_{[i]}$ to reflect ε_i (which is assumed to be normally distributed) better than r_i does. Furthermore, note that the only computational difference between $r_{[i]}$ and r_i is that $r_{[i]}$ uses $\hat{\sigma}_{[i]}^2$ to estimate σ^2 while r_i uses $\hat{\sigma}^2$. Using $r_{[i]}$ instead of r_i in diagnostical methods should lead to more accurate and unbiased results because the estimate $\hat{\sigma}_{[i]}^2$ does not depend on e_i and therefore $\hat{\sigma}_{[i]}^2$ should better estimate σ^2 when e_i is large.

As a final remark, note that, as the r_i's, the $r_{[i]}$'s are correlated.

2.2 Outliers

One task in model criticism is to check for outliers, observations y_i that do not fit the linear model. This is usually done with statistics that are functions of the residuals since residuals contain the information not explained by the fitted line. The common ones are: $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$, $\epsilon_{[i]} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{[i]}$, and $r_{[i]} = t_i = \frac{e_i}{\hat{\sigma}_{[i]}\sqrt{1-h_{ii}}}$. Any of these statistics can be used to detect outliers by plotting the statistic against $\hat{y}_i$, and then picking out points that are far from zero and the other points. It is better not to use e_i or $\epsilon_{[i]}$ since they are not standardized and results may be complicated by the fact that the e_i or $\epsilon_{[i]}$ have different variances. The reason for plotting residuals against $\hat{y}_i$, instead of y_i, is that e_i, and hence the above statistics, are not independent of y_i. In fact, if h_{ij} is small for $i \neq j$, then

$$\begin{aligned} e_i &= y_i - \sum_{j=1}^{n} h_{ij} y_i \\ &\approx y_i(1 - h_{ii}) \\ &= y_i \left[1 - \frac{h_{ii} \sum_{j=1}^{n} y_j^2}{\sum_{j=1}^{n} y_j^2} \right] \\ &\approx y_i \left(1 - \frac{\mathbf{Y'HY}}{\mathbf{Y'Y}} \right) \\ &= y_i(1 - R^2). \end{aligned}$$

Among the four statistics, the one that is most suitable to perform a significant test is t_i since it is standardized and has a known distribution. Since a large value of t_i indicates the possibility of an outlier, we reject the null hypothesis H_0 : *the i^{th} observation is not an outlier* at a significance level of α if $|t_i| > t(1 - \alpha/2; n - p - 1)$.

In the above test, we have assumed that i is known. However, in most cases, there is no way of knowing at which case will the outlier occur, if any. The natural thing to do is then to let i be the case number of the observation that yields the largest $|t_i|$ and test for the possibility of the i^{th} case being an outlier. Though the distribution of $max|t_i|$ is not clear, we can find an upper bound for the $(1 - \alpha)^{th}$ percentage point by using the

Bonferroni inequality and the fact that $t_i \sim t(n-p-1)$:

$$\begin{aligned}
&P[max|t_i| > t(1-\alpha/2n; n-p-1)] \\
&= P[|t_i| > t(1-\alpha/2n; n-p-1) \; for \; some \; i] \\
&= P\left\{\bigcup_{i=1}^{n}[|t_i| > t(1-\alpha/2n; n-p-1)]\right\} \\
&\le \sum_{i=1}^{n} P[|t_i| > t(1-\alpha/2n; n-p-1)] \\
&= \sum_{i=1}^{n} \alpha/n = \alpha.
\end{aligned}$$

The test for a single outlier will then reject if $max|t_i| > t(1-\alpha/2n; n-p-1)$ with a maximum type I error of α. The test for an outlier can be turned into a test for a parameter in a linear model. Suppose that the i^{th} case is suspected as being an outlier. Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{d}_i\phi + \boldsymbol{\varepsilon},$$

where $\mathbf{d}_i$ is an $n \times 1$ column vector with a 1 in the i^{th} element and 0 everywhere else. This model can also be rewritten as

$$\begin{aligned}
\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{d}_i\phi - \mathbf{H}\mathbf{d}_i\phi + \mathbf{H}\mathbf{d}_i\phi + \boldsymbol{\varepsilon} \\
&= \mathbf{X}(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{d}_i\phi) + (\mathbf{I}-\mathbf{H})\mathbf{d}_i\phi + \boldsymbol{\varepsilon} \\
&= \mathbf{X}\boldsymbol{\gamma} + (\mathbf{I}-\mathbf{H})\mathbf{d}_i\phi + \boldsymbol{\varepsilon}, \qquad \text{where} \quad \boldsymbol{\gamma} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{d}_i\phi.
\end{aligned}$$

Note that $\mathbf{X}$ and $(\mathbf{I}-\mathbf{H})\mathbf{d}_i$ are orthogonal to each other. Now let $\mathbf{Z} = [\mathbf{X}, (\mathbf{I}-\mathbf{H})\mathbf{d}_i]$, then the usual F test for $\phi = 0$ is

$$\begin{aligned} F &= \frac{\mathbf{Y}'(\mathbf{H_Z} - \mathbf{H})\mathbf{Y}/1}{\mathbf{Y}'(\mathbf{I} - \mathbf{H_Z})\mathbf{Y}/(n-p-1)} \\ &= \frac{\mathbf{Y}'\mathbf{H}_{(\mathbf{I}-\mathbf{H})\mathbf{d}_i}\mathbf{Y}}{\mathbf{Y}'(\mathbf{I} - \mathbf{H_Z})\mathbf{Y}}(n-p-1) \\ &= \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{d}_i[\mathbf{d}_i'(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\mathbf{d}_i]^{-1}\mathbf{d}_i'(\mathbf{I}-\mathbf{H})\mathbf{Y}}{\mathbf{Y}'(\mathbf{I}-\mathbf{H}-\mathbf{H}_{(\mathbf{I}-\mathbf{H})\mathbf{d}_i})\mathbf{Y}}(n-p-1) \\ &= \frac{e_i^2/(1-h_{ii})}{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y} - e_i^2/(1-h_{ii})}(n-p-1) \\ &= \frac{e_i^2/(1-h_{ii})}{\left(SSE - \frac{e_i^2}{1-h_{ii}}\right)/(n-p-1)} \\ &= \frac{e_i^2/(1-h_{ii})}{SSE_{[i]}/(n-p-1)} \\ &= \frac{e_i^2}{\hat{\sigma}_{[i]}^2(1-h_{ii})} = t_i^2, \end{aligned}$$

and F has an $\mathcal{F}(1, n-p-1)$ distribution with noncentrality of

$$\begin{aligned} &(\mathbf{X}\gamma + (\mathbf{I}-\mathbf{H})\mathbf{d}_i\phi)'(\mathbf{H_Z}-\mathbf{H})(\mathbf{X}\gamma + (\mathbf{I}-\mathbf{H})\mathbf{d}_i\phi)/(2\sigma^2) \\ &= \phi\mathbf{d}_i'(\mathbf{I}-\mathbf{H})(\mathbf{H_Z}-\mathbf{H})(\mathbf{I}-\mathbf{H})\mathbf{d}_i\phi/(2\sigma^2), \qquad \text{since} \qquad \mathbf{H_Z X} = \mathbf{HX} \\ &= \phi\mathbf{d}_i'(\mathbf{H_Z}-\mathbf{H})\mathbf{d}_i\phi/(2\sigma^2), \qquad \text{since} \qquad \mathbf{H_Z H} = \mathbf{HH} \\ &= \phi\mathbf{d}_i'\frac{(\mathbf{I}-\mathbf{H})\mathbf{d}_i\mathbf{d}_i'(\mathbf{I}-\mathbf{H})}{2\sigma^2(1-h_{ii})}\mathbf{d}_i\phi \\ &= \phi\frac{(1-h_{ii})(1-h_{ii})}{2\sigma^2(1-h_{ii})}\phi \\ &= \frac{\phi^2(1-h_{ii})}{2\sigma^2}. \end{aligned}$$

Consequently, $sgn(e_i)\sqrt{F} \sim t\left(n-p-1, \frac{\phi\sqrt{1-h_{ii}}}{\sigma}\right)$ has the same distribution as t_i under $H_0 : \phi = 0$. Testing whether the i^{th} case is an outlier is therefore the same as testing $H_0 : \phi = 0$.

Since the noncentrality is small when h_{ii} is close to 1, the test does not have much power in this situation. In other words, it is hard to detect outliers when the corresponding observations are influential. (See section 2.6 for a discussion on influential observation.) This agrees with the fact that influential points will pull the regression line in their direction and hence will reduce the value of the associated residuals.

2.3 Inclusion of Additional Variables

It is often desired to see whether the addition of an independent variable to the linear model will improve the fit of the model significantly. To be general, assume that q independent variables are added to the base model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

so that the expanded model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is an $n \times p$ matrix of rank p, $\mathbf{Z}$ is an $n \times q$ matrix of rank q, $\boldsymbol{\beta}$ is an $p \times 1$ column vector, $\boldsymbol{\gamma}$ is an $q \times 1$ column vector and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Let $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ be the usual linear estimate of $\boldsymbol{\beta}$ and $\tilde{\mathbf{e}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ be the usual residual vector in the base model. Also, let $\mathbf{W} = (\mathbf{I} - \mathbf{H})\mathbf{Z}$.

To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in the expanded model, we rewrite the model as

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{H})\mathbf{Z}\boldsymbol{\gamma} + \mathbf{H}\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}) + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.\end{aligned}$$

Since $\mathbf{X}'(\mathbf{I} - \mathbf{H})\mathbf{Z} = (\mathbf{X} - \mathbf{X})\mathbf{Z} = \mathbf{0}$, $\mathbf{X}$ and $\mathbf{W}$ are orthogonal matrices, so the estimate of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}$ are simply given by the projections of $\mathbf{Y}$ onto $C(\mathbf{W})$ and $C(\mathbf{X})$

respectively:

$$\begin{aligned}\hat{\gamma} &= (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}\\ &= (\mathbf{W}'\mathbf{W})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{Y}\\ &= (\mathbf{W}'\mathbf{W})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{Y}\\ &= (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\tilde{\mathbf{e}},\end{aligned}$$

and

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

$$\Rightarrow \qquad \hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\gamma}.$$

The covariance matrix for $\hat{\gamma}$ and the residual vector of the expanded model are then given, respectively, by

$$\begin{aligned}Cov[\hat{\gamma}] &= (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'Cov[\mathbf{Y}]\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\\ &= \sigma^2(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\\ &= \sigma^2(\mathbf{W}'\mathbf{W})^{-1}\\ &= \sigma^2(\mathbf{Z}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{Z})^{-1}\\ &= \sigma^2(\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z})^{-1},\end{aligned}$$

and

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}}\\ &= \mathbf{Y} - (\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\gamma})\\ &= \mathbf{Y} - \mathbf{X}(\tilde{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\gamma}) - \mathbf{Z}\hat{\gamma}\\ &= \mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\gamma} - \mathbf{Z}\hat{\gamma}\\ &= \tilde{\mathbf{e}} - (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Z}\hat{\gamma}\\ &= \tilde{\mathbf{e}} - \mathbf{W}\hat{\gamma}.\end{aligned}$$

To see whether the addition of the extra variables improves the fit of the model, we test $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ with

$$\begin{aligned} F &= \frac{\mathbf{Y}'(\mathbf{H}_{[\mathbf{X},\mathbf{Z}]} - \mathbf{H})\mathbf{Y}/rank(\mathbf{H}_{[\mathbf{X},\mathbf{Z}]} - \mathbf{H})}{\mathbf{Y}'(\mathbf{I} - \mathbf{H}_{[\mathbf{X},\mathbf{Z}]})\mathbf{Y}/rank(\mathbf{I} - \mathbf{H}_{[\mathbf{X},\mathbf{Z}]})} \\ &= \frac{\mathbf{Y}'(\mathbf{H} + \mathbf{H}_{(\mathbf{I}-\mathbf{H})\mathbf{Z}} - \mathbf{H})\mathbf{Y}/rank(\mathbf{H}_{(\mathbf{I}-\mathbf{H})\mathbf{Z}})}{\mathbf{Y}'(\mathbf{I} - \mathbf{H} - \mathbf{H}_{(\mathbf{I}-\mathbf{H})\mathbf{Z}})\mathbf{Y}/rank(\mathbf{I} - \mathbf{H} - \mathbf{H}_{(\mathbf{I}-\mathbf{H})\mathbf{Z}})} \\ &= \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Z}[\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z}]^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{Y}/rank(\mathbf{W})}{(\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y} - \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Z}[\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z}]^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Y})/(n-p-rank(\mathbf{W}))} \\ &= \frac{\tilde{\mathbf{e}}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\tilde{\mathbf{e}}}{\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \tilde{\mathbf{e}}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\tilde{\mathbf{e}}} \frac{n-p-q}{q}, \end{aligned}$$

which has a central $\mathcal{F}(q, n-p-q)$ distribution under the base model.

Now suppose that $q = 1$, then the expanded model reduces to

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\phi + \boldsymbol{\varepsilon}$$

with

$$\begin{aligned} \hat{\phi} &= [\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z}]^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Y} \\ &= \frac{\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Y}}{\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z}}, \end{aligned}$$

and we reject $H_0 : \phi = 0$ with a significance level of α if $F = \frac{(\tilde{\mathbf{e}}'\mathbf{W})^2(n-p-1)}{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}\mathbf{W}'\mathbf{W}-(\tilde{\mathbf{e}}'\mathbf{W})^2} > \mathcal{F}(1-\alpha; 1, n-p-1)$ and conclude that the added variable does improve the fit of the model significantly. Here we have assumed that $\mathbf{Z} \notin \mathbf{C}(\mathbf{X})$. Graphically, the significance of ϕ can be detected by the ***added variable plot*** of $\tilde{\mathbf{e}}$ vs. $(\mathbf{I}-\mathbf{H})\mathbf{Z}$. Since

$$\begin{aligned} &\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\phi + \boldsymbol{\varepsilon} \\ \Rightarrow &(\mathbf{I}-\mathbf{H})\mathbf{Y} = (\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}-\mathbf{H})\phi\mathbf{Z} + (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon} \\ \Rightarrow &\tilde{\mathbf{e}} = \phi(\mathbf{I}-\mathbf{H})\mathbf{Z} + (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon} \\ \Rightarrow &E[\tilde{\mathbf{e}}] = \phi(\mathbf{I}-\mathbf{H})\mathbf{Z}, \end{aligned}$$

a plot of $\tilde{\mathbf{e}}$ vs. $(\mathbf{I}-\mathbf{H})\mathbf{Z}$ should reveal a straight line with slope ϕ, which can be estimated by $\hat{\phi}$, in expectation. The variability of the slope can be estimated by $\widehat{Var}(\hat{\phi}) =$

$[\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z}]^{-1}\hat{\sigma}^2 = \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{H}-\mathbf{H}_\mathbf{W})\mathbf{Y}}{\mathbf{Z}'(\mathbf{I}-\mathbf{H})\mathbf{Z}(n-p-1)}$. Added variable plots are helpful since they allow one to see whether the F statistic, on which we may base our decision, is influenced by isolated points.†

Note that, instead of using a new constructed variable, $\mathbf{Z}$ can be one of the variables in the model. This allows a check on the significance of the variables already in the model. Let $\mathbf{X}_i$ denote the i^{th} column of $\mathbf{X}$ and $\mathbf{H}_i$ the projection matrix onto the space spanned by the vectors $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_p$. Then we can apply the above technique to the model

$$\mathbf{Y} = [\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_p]\boldsymbol{\beta} + \mathbf{X}_i\phi + \boldsymbol{\varepsilon},$$

and plot $\tilde{\mathbf{e}} = (\mathbf{I}-\mathbf{H}_i)\mathbf{Y}$ vs. $(\mathbf{I}-\mathbf{H}_i)\mathbf{X}_i$ to see whether the variable x_i should be included in the model. A straight line graph with a nonzero slope indicates that x_i should be kept in the model. These kinds of added variable plots are called *partial leverage regression plots.*

2.4 Normality

The assumption of normality is important primarily for prediction; it is also important because many statistics used to test various aspects of the model assume the normality of the distribution of the error, ε_i. To construct a test, first suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are *i.i.d.* $\mathcal{N}(0, \sigma^2)$ so that $\varepsilon_1/\sigma, \ldots, \varepsilon_n/\sigma$ are *i.i.d.* $\mathcal{N}(0, 1)$. Also let $\mathbf{\Phi}$ denote the distribution function of the standard normal with ϕ representing the corresponding density function.

† One of the examiners of this project has pointed out that these plots can also probably be used to identify whether the candidate variable should be transformed for better fit.

Now let $U_{(1)}, \ldots, U_{(n)}$ be the order statistics of n *i.i.d.* $\mathcal{U}(0,1)$ random observations. Then

$$f_{U_{(1)},\ldots,U_{(n)}}(u_{(1)},\ldots,u_{(n)}) = n!.$$

Transforming the $U_{(i)}$'s by $Z_{(i)} = \Phi^{-1}(U_{(i)})$, we have $U_{(i)} = \Phi(Z_{(i)})$ and, therefore,

$$\frac{\partial U_{(i)}}{\partial Z_{(j)}} = \begin{cases} \phi(Z_{(i)}) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Therefore, $f_{Z_{(i)},\ldots,Z_{(n)}}(z_{(1)},\ldots,z_{(n)}) = n! \prod_{i=1}^{n} \phi(z_{(i)})$ so $Z_{(1)},\ldots,Z_{(n)}$ represent the order statistics of n *i.i.d.* $\mathcal{N}(0,1)$ random variables. If $\varepsilon_{(1)},\ldots,\varepsilon_{(n)}$ are the order statistics of $\varepsilon_1,\ldots,\varepsilon_n$, then

$$\begin{aligned} E[\varepsilon_{(i)}/\sigma] &= E[Z_{(i)}] \\ &= E[\Phi^{-1}(U_{(i)})] \\ &\approx \Phi^{-1}\left[\frac{i-3/8}{n+1/4}\right], \qquad for \qquad n \geq 5. \end{aligned}$$

(See Blom [1958] for a proof of the last approximation.) Consequently, a plot of the ordered standardized residuals $r_{(i)} = \frac{e_{(i)}}{\sigma\sqrt{1-h_{ii}}}$ vs. $\Phi^{-1}\left[\frac{i-3/8}{n+1/4}\right]$ should resemble a straight line with slope 1. If the plot is not linear, the assumption of normality may be violated. A test statistic that is closely related to this graphical procedure is the Shapiro and Wilk statistic:

$$W = \frac{(E[\mathbf{Z}']\mathbf{V}^{-1}\boldsymbol{\varepsilon}/\sigma)^2}{(E[\mathbf{Z}']\mathbf{V}^{-2}E[\mathbf{Z}])\sum_{i=1}^{n}(\varepsilon_{(i)}-\bar{\varepsilon})^2/\sigma^2},$$

where $\mathbf{Z}' = [Z_{(1)},\ldots,Z_{(n)}]$ and $\mathbf{V}$ is the variance-covariance matrix of $Z_{(i)}$. Since $\mathbf{V}$ cannot be computed easily, an approximation to the W test statistic, the square of the sample correlation coefficient between $E[\mathbf{Z}]$ and $\boldsymbol{\varepsilon}/\sigma$, is often used instead:

$$\begin{aligned} W' &= \frac{\left(\sum_{i=1}^{n}(E[Z_{(i)}] - \overline{E[Z_{(i)}]})(\varepsilon_{(i)}-\bar{\varepsilon})/\sigma\right)^2}{\sum_{i=1}^{n}(E[Z_{(i)}] - \overline{E[Z_{(i)}]})^2 \sum_{i=1}^{n}(\varepsilon_{(i)}-\bar{\varepsilon})^2/\sigma^2} \\ &= \frac{\left(\sum_{i=1}^{n} E[Z_{(i)}](\varepsilon_{(i)}-\bar{\varepsilon})/\sigma\right)^2}{\sum_{i=1}^{n} E^2[Z_{(i)}] \sum_{i=1}^{n}(\varepsilon_{(i)}-\bar{\varepsilon})^2/\sigma^2} \\ &= \frac{\left(\sum_{i=1}^{n} E[Z_{(i)}]\varepsilon_{(i)}/\sigma\right)^2}{\sum_{i=1}^{n} E^2[Z_{(i)}] \sum_{i=1}^{n}(\varepsilon_{(i)}-\bar{\varepsilon})^2/\sigma^2}. \end{aligned}$$

Since $\varepsilon_{(i)}/\sigma$ is not observable, it is usually replaced by $r_{(i)}$. If the plot of $r_{(i)}$ vs $E[Z_{(i)}]$ reveals a straight line graph, $r_{(i)}$ and $E[Z_{(i)}]$ are then highly correlated and W' should be large. Therefore, we reject $H_0 : \varepsilon_i \sim i.i.d.\mathcal{N}(0, \sigma^2)$, against the arbitrary alternative of non-normal errors, if W' is small. Percentage points for the distribution of W' are given in Weisberg [1974].

2.5 Heteroscedasticity

Another important assumption in the linear model that should be investigated is the assumption of constant variance of the errors. First, suppose that all other assumptions of the linear model hold except for this one so that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{W})$, where $\mathbf{W}$ has diagonal elements w_i and zero off-diagonal entries. Since ε_i is not observable, the problem of heteroscedasticity is usually investigated with functions of the residual e_i instead. Even if $w_i = 1$, e_i has nonconstant variance, $(1 - h_{ii})\sigma^2$, so to check the constant variance assumption, the standardized residual r_i should be used instead. A graphical procedure plots r_i vs. $\hat{y}_i$ or against the observed values $\mathbf{X}_{ij}$ of any independent variable x_j. An improvement suggested by Cook and Weisberg [1983] is to use r_i^2 instead of r_i, especially when the sample size is small; this has the effect of doubling the sample size since the pattern suggested by the negative residuals and that of the positive residuals are now superimposed to give a single pattern. Another improvement is to use $(1 - h_{ii})\hat{y}_i$ and $(1 - h_{ii})\mathbf{X}_{ij}$ in place of $\hat{y}_i$ and $\mathbf{X}_{ij}$. A plot with non zero slope then suggests that the variance is a function of the independent variable being plotted against. A wedged-shaped graph indicates that the variance is a monotonic function of the independent variable.

Statistical tests for homoscedasticity are more complicated. The idea of the score test presented below can be extended to include more general, twice differentiable functions $w_i = f(\mathbf{z}_i, \boldsymbol{\lambda})$, where $\mathbf{z}_i$ is a known $a \times 1$ vector not necessarily chosen from the design matrix and $\boldsymbol{\lambda}$ is a $b \times 1$ vector of unknown parameters. For simplicity, we only consider the special cases where $a = b = 1, f(z_i, \lambda) = e^{\lambda z_i}$ with $z_i = E(y_i)$ and $z_i = \mathbf{X}_{ij}$. That is, we will be testing the hypothesis of constant variance of the errors (i.e. $\lambda = 0$) against the alternative that the variance depends exponentially on the mean response or the independent variables (i.e. $\lambda \neq 0$).

A test statistic for testing whether the variance depends on the mean response in the form $w_i = e^{\lambda E(y_i)}$ is

$$S_1 = \frac{\left[\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}_i})(e_i^2/\hat{\sigma}^2 - 1)\right]^2}{2\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}_i})^2}$$

which has an asymptotic $\chi^2(1)$ distribution under $H_0 : \lambda = 0$. (See Cook and Weisberg [1983].) Similarly, the test statistic for testing whether the variance depends on the j^{th} independent variable in the form $w_i = e^{\lambda \mathbf{X}_{ij}}$ is

$$S_2 = \frac{\left[\sum_{i=1}^{n}(\mathbf{X}_{ij} - \overline{\mathbf{X}_{ij}})(e_i^2/\hat{\sigma}^2 - 1)\right]^2}{2\sum_{i=1}^{n}(\mathbf{X}_{ij} - \overline{\mathbf{X}_{ij}})^2}$$

where the average $\overline{\mathbf{X}_{ij}}$ is taken over the observed values of the j^{th} independent variable. The asymptotic distribution of S_2 is also $\chi^2(1)$ under the hypothesis $H_0 : \lambda = 0$. It should be noted that using the chi-squared approximation for small sample size will in general lead to a conservative test.

2.6 Influence

The analysis of residuals allows one to check the fit of a regression line on a set of data. However, it does not allow the user to assess the sensitivity of the results against modifications of the data set. In particular, none of the procedures discussed in previous sections are designed to detect the presence of influential observations, those whose deletion will lead to a dramatic change in the regression estimates. One way to detect an influential observation is to compare the difference in the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{[i]}$; that is, to compare the difference in estimating $\boldsymbol{\beta}$ with and without the observation being investigated. One measure of this difference is provided by the sample influence curve for the parameter $\boldsymbol{\beta}$

$$\mathbf{SIC}_i = (n-1)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]}).$$

(See Mallows [1975] for details.) Since $\mathbf{SIC}_i$ is a vector- valued function, it is difficult to compare the $\mathbf{SIC}_i$'s for different observations (i.e. different value of i). Hence, we can instead consider norms of $\mathbf{SIC}_i$ that have the form

$$D_i(\mathbf{M}, c) = \frac{(\mathbf{SIC}_i)'\mathbf{M}(\mathbf{SIC}_i)}{c(n-1)^2} = \frac{(\hat{\boldsymbol{\beta}}_{[i]} - \hat{\boldsymbol{\beta}})'\mathbf{M}(\hat{\boldsymbol{\beta}}_{[i]} - \hat{\boldsymbol{\beta}})}{c},$$

where $\mathbf{M}$ is a $p \times p$ symmetric, positive (semi-) definite matrix and c is a positive scale factor. For any fixed $\mathbf{M}$ and c, contours of D_i are ellipsoid in p-dimension with $\hat{\boldsymbol{\beta}}$ (or $\hat{\boldsymbol{\beta}}_{[i]}$) as the center. For an influential observation, $\hat{\boldsymbol{\beta}}_{[i]}$ would be 'far away' from $\hat{\boldsymbol{\beta}}$ and hence D_i would be large.

To compute D_i, Cook [1977] suggested using $\mathbf{M} = \mathbf{X}'\mathbf{X}$ and $c = p\hat{\sigma}^2$. In this case, D_i is called the Cook's distance

$$\begin{aligned} C_i &= \frac{(\hat{\boldsymbol{\beta}}_{[i]} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}}_{[i]} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} \\ &= \frac{(\hat{\mathbf{Y}}_{[i]} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{[i]} - \hat{\mathbf{Y}})}{p\hat{\sigma}^2}. \end{aligned}$$

Therefore, C_i is also a summary of how far apart are the predicted values with and without the i^{th} observation. A high C_i value ($C_i \geq 1$; see Cook [1982]) then indicates that the i^{th} observation is influential in the sense that deleting it from the data set will alter the estimated mean responses significantly. Since $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1-h_{ii}}$, C_i can be written as

$$\begin{aligned} C_i &= \frac{e_i \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{p\hat{\sigma}^2(1-h_{ii})^2} \\ &= \frac{e_i^2}{p\hat{\sigma}^2(1-h_{ii})^2}\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = \frac{1}{p}\frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})}\frac{h_{ii}}{1-h_{ii}} \\ &= \frac{1}{p}r_i^2\frac{h_{ii}}{1-h_{ii}}. \end{aligned}$$

It is clear that C_i will be large if h_{ii} is close to 1 (unless r_i is very close to zero) or if r_i is large. The increasing function $P_i = \frac{h_{ii}}{1-h_{ii}}$ is called the ***potential.*** It is interesting to note that (i) $P_i = \mathbf{x}_i'(\mathbf{X}_{[i]}'\mathbf{X}_{[i]})^{-1}\mathbf{x}_i$, which is a measure of distance relative to the ellipsoids defined by $(\mathbf{X}_{[i]}'\mathbf{X}_{[i]})^{-1}$, (ii) $\hat{y}_i = (1-h_{ii})\hat{y}_{[i]i} + h_{ii}y_i$, so that P_i is the quotient of the weights of $\hat{y}_{[i]i}$ and y_i respectively, and (iii) $P_i \propto \sum_{j=1}^n Var(\hat{y}_{[i]j}) - \sum_{j=1}^n Var(\hat{y}_j)$, which is the difference of the total variance of the estimated mean values with and without the i^{th} case.

Another measure, known as the $DFFITS$, uses $c = \hat{\sigma}_{[i]}^2$ instead of $p\hat{\sigma}^2$; its square is defined by

$$\begin{aligned} DFFITS_i^2 &= \frac{(\hat{\boldsymbol{\beta}}_{[i]} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}}_{[i]} - \hat{\boldsymbol{\beta}})}{\hat{\sigma}_{[i]}^2} \\ &= pC_i\frac{\hat{\sigma}^2}{\hat{\sigma}_{[i]}^2} \\ &= \left[\frac{r_i^2\hat{\sigma}^2}{\hat{\sigma}_{[i]}^2}\right]\frac{h_{ii}}{1-h_{ii}} \\ &= r_{[i]}^2\frac{h_{ii}}{1-h_{ii}}. \end{aligned}$$

It should be clear that $DFFITS_i^2$ is essentially the same as C_i except that $DFFITS_i^2$ gives more weight to outliers, where $r_{[i]} > r_i > 1$. One shortcoming in using $\hat{\sigma}_{[i]}^2$ instead of

$\hat{\sigma}^2$ is that the shape of the contours of $DFFITS_i^2$ (still ellipsoidal) now depends on i. This makes comparison between $DFFITS_i^2$ for different observations less meaningful since the distances from $\hat{\boldsymbol{\beta}}_{[i]}$ and $\hat{\boldsymbol{\beta}}_{[j]}$ $(i \neq j)$ to $\hat{\boldsymbol{\beta}}$ are now on different scales.

Another related measure is the Mahalanobis distance, which measures the distance of a random vector to the middle of its distribution. (See Christensen [1987].) Let $\mathbf{z}$ be a random (column) vector with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{U}$; the squared Mahalanobis distance is defined by

$$D^2 = (\mathbf{z} - \boldsymbol{\mu})'\mathbf{U}^{-1}(\mathbf{z} - \boldsymbol{\mu}).$$

Even though the rows of $\mathbf{X}$ are not random vectors, we can still apply the idea of the Mahalanobis distance to find out how extreme a particular covariate vector, $\mathbf{x}_i$, is. Estimating $\boldsymbol{\mu}$ by $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}_n$ (where $\mathbf{1}_n$ is a column vector with n 1's) and $\mathbf{U}$ by $\mathbf{S} = \frac{1}{n-1}\left[\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i' - n\overline{\mathbf{x}_i\mathbf{x}_i'}\right] = \frac{1}{n-1}\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n^n\right)\mathbf{X}$ (where $\mathbf{1}_n^n$ is an $n \times n$ matrix with all elements being 1), an estimate of the Mahalanobis distance for the i^{th} data point is

$$\begin{aligned}
\hat{D}_i^2 &= (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) \\
&= (n-1)\left(\mathbf{x}_i - \frac{1}{n}\mathbf{X}'\mathbf{1}_n\right)'\left[\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n^n\right)\mathbf{X}\right]^{-1}\left(\mathbf{x}_i - \frac{1}{n}\mathbf{X}'\mathbf{1}_n\right),
\end{aligned}$$

which is the i^{th} diagonal element of

$$\begin{aligned}
&(n-1)\left(\mathbf{X}' - \frac{1}{n}\mathbf{X}'\mathbf{1}_n\right)'\left[\mathbf{X}'\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n^n\right)\mathbf{X}\right]^{-1}\left(\mathbf{X}' - \frac{1}{n}\mathbf{X}'\mathbf{1}_n\right) \\
&= (n-1)\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}' \qquad \text{where} \qquad \mathbf{K} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n^n\right)\mathbf{X} \\
&= (n-1)\begin{pmatrix} \textit{projection matrix onto the orthogonal complement} \\ \textit{of the column space of } \mathbf{1}_n \textit{ with respect to } \mathbf{X} \end{pmatrix} \\
&= (n-1)\left(\mathbf{H} - \frac{1}{n}\mathbf{1}_n^n\right).
\end{aligned}$$

This has diagonal element $(n-1)\left(h_{ii} - \frac{1}{n}\right)$. Therefore, $\hat{D}_i^2 = (n-1)\left(h_{ii} - \frac{1}{n}\right)$. Note that $DFFITS_i^2$ is an approximate Mahalanobis distance between $\hat{\boldsymbol{\beta}}_{[i]}$ and $\boldsymbol{\beta}$.

A quick way to detect influential cases (influential in the sense that they are far from the center of the data set) is then to pick out cases with high h_{ii} values. This can be done by plotting h_{ii} vs. case number and identifying those with larger leverages. Since $e_{[i]} = \frac{e_i}{1-h_{ii}} \approx e_i$ for small h_{ii}, a plot of $e_{[i]}$ vs. e_i should form a straight line with slope 1. Any points that significantly fall away from such a line may be considered as having high leverage. However, this method may not be appropriate for small n. Since h_{ii} has a lower bound of $1/n$, the slope of the graph of $e_{[i]}$ vs. e_i (assuming it is straight) can be quite different from 1 for small n. Since $\sum_{i=1}^{n} h_{ii} = p$, the average value of h_{ii} is p/n. This suggests a guiding line with a slope $\frac{1}{1-p/n} = \frac{n}{n-p}$ rather than 1. Note that the identification of influential points should be followed with a discussion of how they are expected to sway the analysis and a discussion with the experimenter as to why they arise.

CHAPTER 3
TRANSFORMATION, VARIABLE SELECTION, AND MULTICOLLINEARITY

The first section of this chapter is devoted to a discussion of transformations on the response variable to make the errors (more nearly) normally distributed. Although we will not be discussing transformations on the independent variables, it should be mentioned that they are equally important. Also omitted is a discussion of joint modelling of the mean and variance of the response variable via generalized linear models. The second section deals with techniques for identifying important factors (independent variables) that explain the variation in the response variable. The computational aspect of one such technique will be discussed in the third section. The fourth section discusses the problem of multicollinearity and procedures for detecting the existence of multicollinearity.

3.1 Transformation

It is not always possible to satisfy all the assumptions of the linear model in the original scale of the responses y_i. Sometimes the problem can be resolved by transforming the y_i by a nonlinear function such that the new transformed observations would then satisfy these assumptions. Mathematically, we write

$$z_i = g(y_i) = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i.$$

For example, if the standard deviation of y increases as $\mu = E(y)$ increases, then, to the first order of $y - \mu$,

$$\begin{aligned} ln\ y &\approx ln\ \mu + \left.\frac{\partial ln\ y}{\partial y}\right|_{y=\mu} \cdot (y - \mu) \\ &= \mu + \frac{y-\mu}{\mu} \\ \Rightarrow Var(ln\ y) &\approx \frac{1}{\mu^2}\sigma_y^2, \end{aligned}$$

and hence the transformed response $z_i = ln\ y_i$ may have a stable variance. The *ln* transformation is a special case of the family of transformations considered by Box and Cox [1964]:

$$g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ ln\ y & \text{if } \lambda = 0 \end{cases},$$

where λ is a constant to be determined. Note that $g(y)$ is a continuous function with respect to λ since, by l'Hôpital's rule,

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \lim_{\lambda \to 0} \frac{y^\lambda ln\ y}{1} = ln\ y.$$

A regression model usually includes a constant term. In this case, the transformed model can be assumed to have the form

$$y_i^\lambda = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

since

$$\begin{aligned} g(y_i) &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \\ \Rightarrow \frac{y_i^\lambda - 1}{\lambda} &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \\ \Rightarrow y_i^\lambda &= 1 + \mathbf{x}_i'(\boldsymbol{\beta}\lambda) + \lambda\varepsilon_i, \end{aligned}$$

and 1 can be absorbed by the constant term. The procedure is then to estimate λ and to test whether it is significantly different from 1; if so, a transformation is in the order. One of the methods of estimating λ is the likelihood ratio method. Since $g(\mathbf{Y}) = [g(y_1), \ldots, g(y_n)]'$ is assumed to be $\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ distributed, its density function is:

$$f(g(\mathbf{Y})) = (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}(g(\mathbf{Y}) - \mathbf{X}\boldsymbol{\beta})'(g(\mathbf{Y}) - \mathbf{X}\boldsymbol{\beta})\right\}$$

$$\Rightarrow f(\mathbf{Y}) = (2\pi\sigma^2)^{-n/2} exp\left\{-\frac{1}{2\sigma^2}\left(\frac{\mathbf{Y}^\lambda - 1}{\lambda} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\frac{\mathbf{Y}^\lambda - 1}{\lambda} - \mathbf{X}\boldsymbol{\beta}\right)\right\} J,$$

where

$$J = \left|\frac{\partial(g(y_1), \ldots, g(y_n))}{\partial(y_1, \ldots, y_n)}\right| = \prod_{i=1}^{n}\left|\frac{\partial g(y_i)}{\partial y_i}\right|,$$

and

$$\mathbf{Y}^\lambda - 1 = [y_1^\lambda - 1, \ldots, y_n^\lambda - 1]'.$$

Chen [1991] pointed out that the density function $f(\mathbf{Y})$ is not proper in the sense that it does not integrate to 1. Fortunately, after adjusting the linear model to remedy this defect, he found that approaches with and without the adjustment 'will lead to practically the same parameter estimates for a given set of data' (Chen [1991], Section 5.2.1).

An estimate of λ is then found by maximizing the log likelihood

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma, \lambda) &= ln\ f(\mathbf{Y}) \\ &= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln\ \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{y_i^\lambda - 1}{\lambda} - \mathbf{x}_i'\boldsymbol{\beta}\right)^2 + ln\ J.\end{aligned}$$

Note that, for fixed λ, $ln\ J$ is a constant and so maximizing $\ell(\boldsymbol{\beta}, \sigma, \lambda)$ can be viewed as a least squares problem with response $\frac{y_i^\lambda - 1}{\lambda}$. Therefore, for a fixed λ, the maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2 are, respectively,

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

and $$\tilde{\sigma}^2 = \frac{1}{n}[g(\mathbf{Y}')(\mathbf{I} - \mathbf{H})g(\mathbf{Y})].$$

With $\boldsymbol{\beta}$ and σ^2 estimated by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$, the log likelihood then becomes

$$\begin{aligned}
\ell(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2, \lambda) &= -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln\ \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2} n\tilde{\sigma}^2 + ln\ J \\
&= -\frac{n}{2} ln\ \tilde{\sigma}^2 + ln\ J + constant \\
&= -\frac{n}{2} ln \left[\sum_{i=1}^{n} \left(\frac{y_i^\lambda - 1}{\lambda} - \mathbf{x}_i' \tilde{\boldsymbol{\beta}} \right)^2 \right] + \sum_{i=1}^{n} ln \left| \frac{\partial g(y_i)}{\partial y_i} \right| + constant.
\end{aligned}$$

The maximum likelihood estimate, $\hat{\lambda}$, is the value of λ that maximizes the above log likelihood. This is usually done by evaluating $\ell(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2, \lambda)$ at selected points λ_j over a reasonable range, say -2 to 2. The λ_j which yields the maximum log likelihood in this set can be treated as $\hat{\lambda}$. Accuracy can be increased by "fine tuning" the values of λ_j at which the log likelihood is evaluated. However, since the values of λ are often rounded to values such as $-2, -1, -1/2, 0, 1/3, 1/2, 1, 2$, for the sake of convenience, ease of interpretation, or physical reason associated with the problem at hand, it is thus usually not necessary to estimate λ to a high accuracy.

An approximate $100(1-\alpha)\%$ confidence interval for λ is given by the values of λ_0 such that

$$2[\ell(\hat{\lambda}) - \ell(\lambda_0)] \leq \chi^2(1-\alpha; 1);$$

this is based on the asymptotic property of the likelihood ratio test. The likelihood ratio test can be used to test the hypothesis $H_0 : \lambda = 1$, whose rejection indicates the need for a transformation.

A graphical aid to check on the need for transformation can be derived from the following. First, following Atkinson [1985], we normalize the power transformation by the geometric mean of the response values so that the transformation is:

$$G(y_i) = \begin{cases} \frac{y_i^\lambda - 1}{\lambda \tilde{y}^{\lambda-1}} & \text{if } \lambda \neq 0 \\ \tilde{y} ln\ y_i & \text{if } \lambda = 0 \end{cases},$$

where $\tilde{y} = (\prod_{i=1}^{n} y_i)^{1/n}$. Now assume that, for some value λ, the model

$$G(\mathbf{Y};\lambda) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is 'correct'. (Of course, the value of λ that makes $G(\mathbf{Y};\lambda)$ normally distributed is, in general, not the same as the one that makes $g(\mathbf{Y};\lambda)$ normally distributed. However, it should be mentioned that, asymptotically, $G(\mathbf{Y};\lambda)$ and $g(\mathbf{Y};\lambda)$ are identically distributed.) Expanding $G(\mathbf{Y};\lambda)$ about some point $\lambda = \lambda_0$ and ignoring terms after the first order, we have

$$G(\mathbf{Y};\lambda) \approx G(\mathbf{Y};\lambda_0) + (\lambda - \lambda_0)\frac{\partial G(\mathbf{Y};\lambda)}{\partial\lambda}\bigg|_{\lambda=\lambda_0}.$$

The model can then be written approximately as

$$G(\mathbf{Y};\lambda_0) = \mathbf{X}\boldsymbol{\beta} - (\lambda - \lambda_0)\frac{\partial G(\mathbf{Y};\lambda_0)}{\partial\lambda} + \boldsymbol{\varepsilon}.$$

The significance of $\lambda - \lambda_0$ can be checked by the added variable plot (see Section 2.3) of $\tilde{\mathbf{e}} = (\mathbf{I} - \mathbf{H})G(\mathbf{Y};\lambda_0)$ vs. $\mathbf{w} = (\mathbf{I} - \mathbf{H})\frac{\partial G(\mathbf{Y};\lambda)}{\partial\lambda}\Big|_{\lambda=\lambda_0}$. This should yield a straight line graph with slope $-(\lambda - \lambda_0)$. Therefore, if no transformation is needed, the added variable plot for $\lambda_0 = 1$ should display, approximately, a horizontal line.

The advantage of the added variable plot is that it shows not only the need for a transformation, but also whether such a need is dictated by the whole set of data or just by one or a few cases. In the latter case, all but a few data points would lie around a horizontal line. This information is very useful because sometimes the untransformed model is perfectly all right after the deletion of an outlying case. (For an example, see Example 8 in Chapter 6 of Atkinson [1985].) In contrast, the likelihood test statistic is a single summary value pooling information possessed by the data.

Very often one simply wants to see whether a transformation of the response variable is necessary; that is, to see whether $\lambda = 1$. In this situation, estimating λ is not the prime directive and therefore using the likelihood ratio test may be too time-consuming for such a simple task. An alternative is to use the usual F test on the expanded model

$$G(\mathbf{Y}; \lambda_0) = \mathbf{X}\boldsymbol{\beta} - (\lambda - \lambda_0)\frac{\partial G(\mathbf{Y}; \lambda_0)}{\partial \lambda} + \boldsymbol{\varepsilon}.$$

As mentioned Section 2.3, the F test for $\gamma = \lambda - \lambda_0 = 0 \Leftrightarrow \lambda = \lambda_0$ is

$$F = \frac{(\tilde{\mathbf{e}}'\mathbf{w})^2(n - p - 1)}{(\tilde{\mathbf{e}}'\tilde{\mathbf{e}})(\mathbf{w}'\mathbf{w}) - (\tilde{\mathbf{e}}'\mathbf{w})^2}$$

and we reject $H_0 : \lambda = \lambda_0$ with a significance level of α if $F > \mathcal{F}(1 - \alpha; 1, n - p - 1)$. (Atkinson [1985] mentioned that the equivalent score test $T_p(\lambda_0) = sgn(-\tilde{\mathbf{e}}'\mathbf{w})\sqrt{F}$ is an approximation to the likelihood test for the hypothesis $\lambda = \lambda_0$.)

3.2 Variable Selection

One of the major goals in regression is to find out the factors that strongly influence the values of the response variable. It is therefore desirable to cut down the number of independent variables by deleting those which do not contribute significantly. Besides, although it is no longer a heavy job to perform a regression based on many variables, the inclusion of irrelevant independent variables will result in complicated models which are difficult to interpret. Furthermore, deleting variables has some desirable statistical properties as discussed below.

Consider the model

$$\mathbf{Y} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon},$$

where $\mathbf{X}_p$ (with rank p) and $\mathbf{X}_r$ (with rank r) form a partition of $\mathbf{X}$ ($rank(\mathbf{X}) = q$) so that $\mathbf{X} = [\mathbf{X}_p\ \mathbf{X}_r]$, and similarly $\boldsymbol{\beta}' = [\boldsymbol{\beta}'_p\ \boldsymbol{\beta}'_r]$. It is assumed that the variables in $\mathbf{X}_r$ are the ones we are attempting to delete; the reduced model is then

$$\mathbf{Y} = \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}.$$

Let $\hat{\boldsymbol{\beta}}' = [\hat{\boldsymbol{\beta}}'_p\ \hat{\boldsymbol{\beta}}'_r]$ and $\tilde{\boldsymbol{\beta}}_p$ be, respectively, the usual least square estimate of the $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}_p$ in the full and reduced models. Also, let $\hat{\sigma}^2$ and $\tilde{\sigma}^2_p$ be the residual mean squares for the two models; that is,

$$\hat{\sigma}^2 = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}/(n - p - r),$$

$$\tilde{\sigma}^2_p = \mathbf{Y}'(\mathbf{I} - \mathbf{X}_p(\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p)\mathbf{Y}/(n - p).$$

Then, from Section 2.3, we have

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

$$\Rightarrow E(\hat{\boldsymbol{\beta}}_p) = E(\tilde{\boldsymbol{\beta}}_p) - (\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p\mathbf{X}_r E(\hat{\boldsymbol{\beta}}_r),$$

and, if we assume the full model is correct,

$$E(\tilde{\boldsymbol{\beta}}_p) = \boldsymbol{\beta}_p + (\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p\mathbf{X}_r\boldsymbol{\beta}_r.$$

If $\boldsymbol{\beta}_r = 0$, the resulting estimate $\tilde{\boldsymbol{\beta}}_p$ is unbiased. If $\boldsymbol{\beta}_r \neq 0$, $\tilde{\boldsymbol{\beta}}_p$ is only unbiased if $\mathbf{X}'_p\mathbf{X}_r = 0$; that is, the columns in $\mathbf{X}_p$ are orthogonal to those in $\mathbf{X}_r$.

Another desirable property is that $Cov(\hat{\boldsymbol{\beta}}_p) - Cov(\tilde{\boldsymbol{\beta}}_p)$ is positive semidefinite, which I prove below. Note first that

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}'^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}'^{-1} & (\mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

If we let

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{pmatrix} = (\mathbf{X}'\mathbf{X}) = \begin{pmatrix} \mathbf{X}_p'\mathbf{X}_p & \mathbf{X}_p'\mathbf{X}_r \\ \mathbf{X}_r'\mathbf{X}_p & \mathbf{X}_r'\mathbf{X}_r \end{pmatrix},$$

then

$$\begin{aligned} Cov\hat{\boldsymbol{\beta}} &= Cov\begin{pmatrix} \hat{\boldsymbol{\beta}}_p \\ \hat{\boldsymbol{\beta}}_r \end{pmatrix} = Cov[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}'^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}'^{-1} & (\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix} \end{aligned}$$

$$\Rightarrow Cov(\hat{\boldsymbol{\beta}}_p) = \sigma^2[\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}'^{-1}]$$

$$\begin{aligned} \Rightarrow Cov(\hat{\boldsymbol{\beta}}_p) - Cov(\tilde{\boldsymbol{\beta}}_p) &= \sigma^2[(\mathbf{X}_p'\mathbf{X}_p)^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}'^{-1}] - \sigma^2(\mathbf{X}_p'\mathbf{X}_p)^{-1} \\ &= \sigma^2\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}'^{-1}. \end{aligned}$$

Now, for any column vector **a**,

$$\begin{aligned} &\mathbf{a}'[Cov(\hat{\boldsymbol{\beta}}_p) - Cov(\tilde{\boldsymbol{\beta}}_p)]\mathbf{a} \\ &= \sigma^2\mathbf{a}'\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}'\mathbf{A}'^{-1}\mathbf{a} \\ &= \mathbf{b}'(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{b}, \qquad \text{where} \qquad \mathbf{b}' = \mathbf{a}'\mathbf{A}^{-1}\mathbf{B} \\ &\geq 0. \end{aligned}$$

The last inequality holds since $(\mathbf{D}-\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1}$ is the covariance matrix of $\hat{\boldsymbol{\beta}}_r$ and hence positive (semi-) definite. This concludes the required proof.

The advantage of $Cov(\hat{\boldsymbol{\beta}}_p) - Cov(\tilde{\boldsymbol{\beta}}_p)$ being positive semidefinite can be realized if we look at the variance of the estimate of $\boldsymbol{\lambda}'\boldsymbol{\beta}_p$. Consider any linear function of $\boldsymbol{\beta}_p$, $\phi = \boldsymbol{\lambda}'\boldsymbol{\beta}_p$. (Note that ϕ is estimable for any vector $\boldsymbol{\lambda}$ because $\boldsymbol{\beta}_p$ itself is estimable.) Let $\hat{\phi} = \boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}_p$ and $\tilde{\phi} = \boldsymbol{\lambda}'\tilde{\boldsymbol{\beta}}_p$ be the estimate of ϕ based on $\hat{\boldsymbol{\beta}}_p$ and $\tilde{\boldsymbol{\beta}}_p$ respectively. Then

$$\begin{aligned} Var(\hat{\phi}) - Var(\tilde{\phi}) &= \boldsymbol{\lambda}'Cov(\hat{\boldsymbol{\beta}}_p)\boldsymbol{\lambda} - \boldsymbol{\lambda}'Cov(\tilde{\boldsymbol{\beta}}_p)\boldsymbol{\lambda} \\ &= \boldsymbol{\lambda}'[Cov(\hat{\boldsymbol{\beta}}_p) - Cov(\tilde{\boldsymbol{\beta}}_p)]\boldsymbol{\lambda} \\ &\geq 0. \end{aligned}$$

That is, estimating linear combinations of $\boldsymbol{\beta}_p$ using $\tilde{\boldsymbol{\beta}}_p$ rather than $\hat{\boldsymbol{\beta}}_p$ yields more precise results. Consequently, confidence intervals for ϕ will be narrower if they are constructed using $\tilde{\boldsymbol{\beta}}_p$ rather than $\hat{\boldsymbol{\beta}}_p$. Of course, even though $Cov(\hat{\boldsymbol{\beta}}_p) - Cov(\tilde{\boldsymbol{\beta}}_p)$ is positive semidefinite regardless of whether $\tilde{\boldsymbol{\beta}}_p$ is unbiased, the advantages just discussed are desirable only if $\tilde{\boldsymbol{\beta}}_p$ is unbiased.

The remaining question is then how to decide which variables may be deleted.† Some common procedures designed for such a purpose are forward regression, backward regression and stepwise regression (a combination of forward and backward regression). Between forward and backward regression, the latter is more appropriate since it starts with a full model and eliminates only those variables that are not significant; however the resulting models can be complicated. Among the three, stepwise regression seems to be the most acceptable and widely used method. However, since all three methods process only one variable at time, valuable information provided by certain combinations of the independent variables can easily be missed, and the resulting model may be far from the best. Another shortcoming of these methods is that they only give a single model – they do not provide the second best or other alternative models for further consideration and decision; in many situations involving observational studies, there may not be a single best subset but several good ones. An alternative is to consider all possible regressions, provided that the number of variables is not very large (say, ≤ 15). This method considers

† Note that the experimenter may be able to suggest which variables are expected to be of foremost importance in influencing the response. Sometimes, variables may be grouped into primary and secondary categories of potential importance. In any case, variable selection methods are useful.

the effect of each of the 2^q possible linear combinations of the independent variables on the response and allows one to select the best subset (or the first, say, ten best ones) based on some predefined criteria. One common criterion for this method is the C_p criterion.

Before stating the criterion, we first derive the C_p statistic. Consider the mean square error for the fitted value $\tilde{y}_i = \mathbf{x}'_{pi}\tilde{\boldsymbol{\beta}}_p$ (where $\mathbf{x}'_{pi}$ denotes the i^{th} row of $\mathbf{X}_p$) defined by

$$\begin{aligned}
&E[\tilde{y}_i - E(y_i)]^2 \\
&= E\{[\tilde{y}_i - E(\tilde{y}_i)] + [E(\tilde{y}_i) - E(y_i)]\}^2 \\
&= E[\tilde{y}_i - E(y_i)]^2 + 2E[\tilde{y}_i - E(\tilde{y}_i)][E(\tilde{y}_i) - E(y_i)] + [E(\tilde{y}_i) - E(y_i)]^2 \\
&= Var(\tilde{y}_i) + [E(\tilde{y}_i) - E(y_i)]^2.
\end{aligned}$$

The (scaled) total mean square error for all n fitted values is then

$$\Gamma_p = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left\{Var(\tilde{y}_i) + [E(\tilde{y}_i) - E(y_i)]^2\right\},$$

where the subscript p indicates that the statistic is calculated from a model with p variables employed. Assuming the full model is correct, $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$. Then, in matrix form,

$$\begin{aligned}
\Gamma_p &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}[Var(\mathbf{x}'_{pi}(\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p\mathbf{Y})] + [E(\tilde{\mathbf{Y}}) - E(\mathbf{Y})]'[E(\tilde{\mathbf{Y}}) - E(\mathbf{Y})]\right\} \\
&= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}[\mathbf{x}'_{pi}(\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p\sigma^2\mathbf{X}_p(\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{x}_{pi}] + [E(\mathbf{H}_{\mathbf{X}_p}\mathbf{Y}) - E(\mathbf{Y})]'[E(\mathbf{H}_{\mathbf{X}_p}\mathbf{Y}) - E(\mathbf{Y})]\right\} \\
&= \sum_{i=1}^{n}[\mathbf{x}'_{pi}(\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{x}_{pi}] + \frac{1}{\sigma^2}[\mathbf{H}_{\mathbf{X}_p}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}]'[\mathbf{H}_{\mathbf{X}_p}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}] \\
&= trace[\mathbf{H}_{\mathbf{X}_p}] + \frac{1}{\sigma^2}\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{X}\boldsymbol{\beta} \\
&= p + \frac{1}{\sigma^2}\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{X}\boldsymbol{\beta}.
\end{aligned}$$

Note that

$$\begin{aligned}
E(\mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}) &= trace[(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\sigma^2\mathbf{I}] + E(\mathbf{Y}')(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})E(\mathbf{Y}) \\
&= \sigma^2(n - p) + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{X}\boldsymbol{\beta}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\Gamma_p &= p + \frac{1}{\sigma^2}[E(\mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}) - \sigma^2(n-p)] \\ &= \frac{E[\mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}]}{\sigma^2} + 2p - n.\end{aligned}$$

Since Γ_p contains unknowns parameters, it is usually estimated by

$$C_p = \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}}{\hat{\sigma}^2} + 2p - n.$$

(Discussions of Γ_p and C_p can be found in Christensen [1987], Daniel and Wood [1980], Hocking [1976], and Neter *et al.* [1985].)

It should be clear that, for a correct reduced model, the total mean squared error, and hence the C_p value, should be small. It is therefore possible to use C_p as a guideline for selecting the best subset. The procedure for variable selection is to perform 2^q regressions using different subsets of the independent variables. For each regression, the value of C_p is calculated. The better models have small C_p, with $C_p \approx p$. Often, especially when q is large, more than one C_p may satisfy the above criterion. Note that we do not simply choose the subset associated with the smallest C_p as the best subset. The reason is that

$$\begin{aligned}E[C_p|E(\tilde{\mathbf{Y}}) = E(\mathbf{Y})] &= E\left[\frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}}{\hat{\sigma}^2} + 2p - n \middle| E(\tilde{\mathbf{Y}}) = E(\mathbf{Y})\right] \\ &\approx \Gamma_p|E(\tilde{\mathbf{Y}}) = E(\mathbf{Y}) \\ &= p.\end{aligned}$$

Therefore, if the bias in $\tilde{\mathbf{Y}}$ is small, we expect C_p to be close to p.

Graphically, variable selection can be done by plotting C_p vs. p together with the line $C_p = p$. The points with small C_p values that are relatively close to the line are usually chosen as the better subsets. To get an idea of the variation in the C_p vs. p plot, consider, again, the full model

$$\mathbf{Y} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}.$$

The test statistic for $\boldsymbol{\beta}_r = \mathbf{0}$ is

$$\begin{aligned}F &= \frac{\mathbf{Y}'(\mathbf{H} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}}{r\hat{\sigma}^2} \\ &= \frac{-\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}}{r\hat{\sigma}^2} \\ &= \frac{-(n-p+r)\hat{\sigma}^2}{r\hat{\sigma}^2} + \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}}{r\hat{\sigma}^2}\end{aligned}$$

$$\Rightarrow rF = -n + p - r + C_p - 2p + n$$

$$\Rightarrow r(F-1) = C_p - p$$

$$\Rightarrow Var(C_p - p) = r^2 Var(F).$$

Since $F \sim \mathcal{F}(r, n-q)$ (recall that q is the number of independent variables in the full model) under $H_0 : \boldsymbol{\beta}_r = \mathbf{0}$,

$$Var(C_p - p) = r^2 \frac{2(n-q)^2(r+n-q-2)}{r(n-q-2)^2(n-q-4)} \qquad for\ n-q > 4$$

$$\Rightarrow \sqrt{Var(C_p - p)} = \frac{n-q}{n-q-2}\sqrt{\frac{2r(r+n-q-2)}{n-q-4}} \qquad for\ n-q > 4.$$

The table below gives values of $\sqrt{Var(C_p - p)}$ for various values of r and $q - n$.

$n-q \setminus r$	2	4	6	8	10	12
5	7.5	12	17.3	22.1	26.9	31.6
10	3.2	5	6.6	8.2	9.7	11.2
15	2.7	4.1	5.3	6.4	7.5	8.5
20	2.5	3.7	4.7	5.7	6.6	7.5
25	2.4	3.5	4.4	5.3	6.1	6.9
30	2.3	3.4	4.2	5.0	5.8	6.5
35	2.3	3.3	4.1	4.9	5.6	6.3

As seen from the table, the variation can be quite large as compared to the value of p, which is usually less than 15. This makes the problem of selecting the best subset more difficult since points that are close to the line $C_p = p$ statistically (in the sense that their distances to the line are within, say, one standard deviation of $C_p - p$) may not appear so to the human eye.

3.3 Computational Aspects of Variable Selection

Even with a computer, performing 2^q regression is still a tremendous amount of work and may take a long time for $q \geq 10$. Since most selection criteria are functions of $(\mathbf{X}'_p\mathbf{X}_p)^{-1}$ or the error sum of squares ($ESS_p = \mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y}$), the task of finding the best subset can be speeded up if there is a quicker way of finding $(\mathbf{X}'_p\mathbf{X}_p)^{-1}$ and/or ESS_p for all 2^q regressions. One such method uses the *SWEEP* operator (cf. Goodnight [1979]). Given an $n \times n$ matrix $\mathbf{A}$, the *SWEEP* operator on $\mathbf{A}$ is defined by the following algorithm.

1. Let $b = a_{kk}$ and divide row k of $\mathbf{A}$ by b.

2. For each row $i \neq k$, let $c_i = a_{ik}$ and add $(-a_{ik})$*(row k) to row i.

3. Set a_{kk} to $1/b$ and $a_{ik} (i \neq k)$ to $-c_i/b$.

In matrix form, this is

$$S_k\mathbf{A} = \begin{array}{c|ccccc|}
 & & j & & k & \\
 & & \vdots & & \vdots & \\
i & \cdots & a_{ij} - \frac{a_{kj}a_{ik}}{a_{kk}} & \cdots & -\frac{a_{ik}}{a_{kk}} & \cdots \\
 & & \vdots & & \vdots & \\
k & \cdots & \frac{a_{kj}}{a_{kk}} & \cdots & \frac{1}{a_{kk}} & \cdots \\
 & & \vdots & & \vdots &
\end{array}.$$

If we now perform another *SWEEP* operation, S_l, we will have

$S_lS_k\mathbf{A}$

$$
= \begin{array}{cc}
 & \begin{array}{ccccccc} & j & & l & & k & \end{array} \\
\begin{array}{c} \\ i \\ \\ l \\ \\ k \\ \\ \end{array} &
\left(\begin{array}{ccccccc}
 & \vdots & & \vdots & & \vdots & \\
\cdots & a_{ij}-\frac{a_{kj}a_{ik}}{a_{kk}} - \frac{\left(a_{il}-\frac{a_{kl}a_{ik}}{a_{kk}}\right)\left(a_{lj}-\frac{a_{kj}a_{lk}}{a_{kk}}\right)}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots & -\frac{a_{il}-\frac{a_{kl}a_{ik}}{a_{kk}}}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots & -\frac{a_{ik}}{a_{kk}}-\frac{\left(-\frac{a_{lk}}{a_{kk}}\right)\left(a_{il}-\frac{a_{kl}a_{ik}}{a_{kk}}\right)}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots \\
 & \vdots & & \vdots & & \vdots & \\
\cdots & -\frac{a_{lj}-\frac{a_{kj}a_{lk}}{a_{kk}}}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots & \frac{1}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots & \frac{-\frac{a_{lk}}{a_{kk}}}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots \\
 & \vdots & & \vdots & & \vdots & \\
\cdots & -\frac{a_{kj}}{a_{kk}}-\frac{\left(-\frac{a_{kl}}{a_{kk}}\right)\left(a_{lj}-\frac{a_{kj}a_{lk}}{a_{kk}}\right)}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots & \frac{-\frac{a_{kl}}{a_{kk}}}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots & \frac{1}{a_{kk}}-\frac{\left(-\frac{a_{lk}}{a_{kk}}\right)\left(\frac{a_{kl}}{a_{kk}}\right)}{a_{ll}-\frac{a_{kl}a_{lk}}{a_{kk}}} & \cdots \\
 & \vdots & & \vdots & & \vdots &
\end{array}\right)
\end{array}
$$

$$
= \begin{array}{cc}
 & \begin{array}{ccccccc} & j & & l & & k & \end{array} \\
\begin{array}{c} \\ i \\ \\ l \\ \\ k \\ \\ \end{array} &
\left(\begin{array}{ccccccc}
 & \vdots & & \vdots & & \vdots & \\
\cdots & a_{ij}-\frac{a_{ij}a_{kj}a_{ll}+a_{il}a_{lj}a_{kk}-a_{il}a_{kj}a_{lk}-a_{ik}a_{lj}a_{kl}}{a_{ll}a_{kk}-a_{kl}a_{lk}} & \cdots & -\frac{a_{il}a_{kk}-a_{kl}a_{ik}}{a_{ll}a_{kk}-a_{kl}a_{lk}} & \cdots & -\frac{a_{il}a_{kk}-a_{kl}a_{ik}}{a_{ll}a_{kk}-a_{kl}a_{lk}} & \cdots \\
 & \vdots & & \vdots & & \vdots & \\
\cdots & \frac{a_{lj}a_{kk}-a_{kj}a_{lk}}{a_{kk}a_{ll}-a_{kl}a_{lk}} & \cdots & \frac{a_{kk}}{a_{kk}a_{ll}-a_{kl}a_{lk}} & \cdots & -\frac{a_{lk}}{a_{ll}a_{kk}-a_{kl}a_{lk}} & \cdots \\
 & \vdots & & \vdots & & \vdots & \\
\cdots & \frac{a_{kj}a_{ll}-a_{lj}a_{kl}}{a_{kk}a_{ll}-a_{kl}a_{lk}} & \cdots & -\frac{a_{kl}}{a_{kk}a_{ll}-a_{kl}a_{lk}} & \cdots & \frac{a_{ll}}{a_{ll}a_{kk}-a_{kl}a_{lk}} & \cdots \\
 & \vdots & & \vdots & & \vdots &
\end{array}\right)
\end{array}.
$$

It should be obvious from the last matrix that interchanging the subscripts l and k (including the one used to indicate the row and column) will leave the matrix unchanged.

That is

(P1) $S_l S_k \mathbf{A} = S_k S_l \mathbf{A}$.

Furthermore, by performing another S_k operation on the matrix $S_k\mathbf{A}$, we can see that

$$S_kS_k\mathbf{A} = \begin{array}{cc} & \begin{array}{ccccc} & j & & k & \end{array} \\ \begin{array}{c} \\ i \\ \\ k \\ \\ \end{array} & \begin{pmatrix} & \vdots & & \vdots & \\ \cdots & a_{ij} - \frac{a_{kj}a_{ik}}{a_{kk}} - \frac{(-a_{ik}/a_{kk})(a_{kj}/a_{kk})}{1/a_{kk}} & \cdots & -\frac{-a_{ik}/a_{kk}}{1/a_{kk}} & \cdots \\ & \vdots & & \vdots & \\ \cdots & \frac{a_{kj}/a_{kk}}{1/a_{kk}} & \cdots & \frac{1}{1/a_{kk}} & \cdots \\ & \vdots & & \vdots & \end{pmatrix} \end{array}$$

$$= \begin{array}{cc} & \begin{array}{ccccc} & j & & k & \end{array} \\ \begin{array}{c} \\ i \\ \\ k \\ \\ \end{array} & \begin{pmatrix} & \vdots & & \vdots & \\ \cdots & a_{ij} & \cdots & a_{ik} & \cdots \\ & \vdots & & \vdots & \\ \cdots & a_{kj} & \cdots & a_{kk} & \cdots \\ & \vdots & & \vdots & \end{pmatrix} \end{array}.$$

This proves

(P2) $S_kS_k\mathbf{A} = \mathbf{A}$.

Although not straightforward, it can be shown that,

(P3) If $\mathbf{A}$ is transformed by $S_{i_1}, S_{i_2}, \ldots, S_{i_k}$ successively to a matrix $\mathbf{B}$, the $k \times k$ submatrix of $\mathbf{B}$ indexed by $\{i_1, \ldots, i_k\}$ is the inverse of the corresponding $k \times k$ submatrix of $\mathbf{A}$ indexed by $\{i_1, \ldots, i_k\}$.

Finally, note that, after a S_k operation on $\mathbf{A}$, the elements in the k^{th} column can affect only themselves in further *SWEEP* operations. That is, if we are to perform a S_i on $S_k\mathbf{A}$, the resulting values of the elements that do not fall in the k^{th} column will be the same whether or not we carried out step 3 of the algorithm when we performed $S_k\mathbf{A}$. We therefore conclude that

(P4) Let $S_{i_1}, \ldots, S_{i_k}$ be the operation performed on $\mathbf{A}$ so far. (Without loss of generality, we can assume that none of the subscripts is the same due to property P2.) The matrix obtained after deleting the columns indexed by $\{i_1, \ldots, i_k\}$ would have been the same had we carried out only steps 1 and 2 of the algorithm; this is essentially Gauss-Jordan elimination.

To use the *SWEEP* on variable selection, consider the model

$$\mathbf{Y} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon},$$

and augment the matrix $\mathbf{X}'\mathbf{X}$ to

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \hline \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{array}\right] = \left[\begin{array}{c|c|c} \mathbf{X}_p'\mathbf{X}_p & \mathbf{X}_p'\mathbf{X}_r & \mathbf{X}_p'\mathbf{Y} \\ \hline \mathbf{X}_r'\mathbf{X}_p & \mathbf{X}_r'\mathbf{X}_r & \mathbf{X}_r'\mathbf{Y} \\ \hline \mathbf{Y}'\mathbf{X}_p & \mathbf{Y}'\mathbf{X}_r & \mathbf{Y}'\mathbf{Y} \end{array}\right].$$

Now let $\mathbf{B} = S_p, \ldots, S_1\mathbf{A}$. As far as the submatrix $\mathbf{X}_p'\mathbf{Y}$ and $\mathbf{Y}'\mathbf{Y}$ are concerned, these operations are the same as performing the Gauss-Jordan elimination on the first p rows. (This follows from P4.) Therefore

$$\mathbf{B} = \left[\begin{array}{c|c|c} \cdots & \cdots & (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{Y} \\ \hline \cdots & \cdots & \cdots \\ \hline \cdots & \cdots & \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{Y} \end{array}\right].$$

Moreover, P3 tells us that the submatrix $\mathbf{X}_p'\mathbf{X}_p$ will be reduced to $(\mathbf{X}_p'\mathbf{X}_p)^{-1}$ and $\mathbf{B}$ thus becomes

$$\mathbf{B} = \left[\begin{array}{c|c|c} (\mathbf{X}_p'\mathbf{X}_p)^{-1} & \cdots & (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{Y} \\ \hline \cdots & \cdots & \cdots \\ \hline \cdots & \cdots & \mathbf{Y}'(\mathbf{I} - \mathbf{H}_{\mathbf{X}_p})\mathbf{Y} \end{array}\right]$$

$$= \left[\begin{array}{c|c|c} (\mathbf{X}_p'\mathbf{X}_p)^{-1} & \cdots & \tilde{\boldsymbol{\beta}}_p \\ \hline \cdots & \cdots & \\ \hline \cdots & \cdots & ESS_p \end{array}\right].$$

(For simplicity, we have used the first p rows of the matrix $\mathbf{A}$. The results still hold if we apply the $SWEEP$ operators to rows chosen according to some other schemes. For example, if $rank(\mathbf{X}) = q$ and we apply $S_{i_1}, \ldots, S_{i_k}$ to $\mathbf{A}$ to obtain $\mathbf{B} = S_{i_k}, \ldots, S_{i_1}\mathbf{A}$, we will have (i) $b_{i_j,q+1} = \tilde{\boldsymbol{\beta}}_{i_j}$ for $1 \leq j \leq k$, (ii) $b_{q+1,q+1} = ESS_k$, and (iii) the submatrix formed by $b_{i_j,i_l}, 1 \leq j, l \leq k$, is the inverse of the submatrix formed by $a_{i_j,i_l}, 1 \leq j, l \leq k$,.) Applying the $SWEEP$ operator successively, we can then obtain $(\mathbf{X}_p'\mathbf{X}_p)^{-1}$, $\tilde{\boldsymbol{\beta}}_p$, and ESS_p for all possible subsets. This is illustrated in the table below, where we have 3 regressor variables and we want to fit all $2^3 - 1 = 7$ models.

Step	*Operator*	*Resulting Matrix*	*Variables in Model*
0		$\mathbf{A}$	*none*
1	S_1	$S_1\mathbf{A}$	x_1
2	S_2	$S_2S_1\mathbf{A}$	x_1, x_2
3	S_1	$S_1S_2S_1\mathbf{A} = S_2S_1S_1\mathbf{A} = S_2\mathbf{A}$	x_2
4	S_3	$S_3S_2\mathbf{A}$	x_2, x_3
5	S_2	$S_2S_3S_2\mathbf{A} = S_3S_2S_2\mathbf{A} = S_3\mathbf{A}$	x_3
6	S_1	$S_1S_3\mathbf{A}$	x_1, x_3
7	S_2	$S_2S_1S_3\mathbf{A}$	x_1, x_2, x_3

3.4 Multicollinearity

Consider again the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The most fruitful statistical inference results from a well designed experiment where the columns of $\mathbf{X}$ are orthogonal. Observational studies are not rare, however, especially in medical statistics, and in these studies $\mathbf{X}$ is generally non-orthogonal. Multicollinearity refers to the fact that the columns of the design matrix $\mathbf{X}$ are not linearly independent. The effect of this is that some linear functions of $\boldsymbol{\beta}$ do not have unique estimates. Fortunately, it is unusual in regression analysis that columns of $\mathbf{X}$ will exhibit exact linear dependency. However, near multicollinearity does occur occasionally.

Since $\mathbf{X}'\mathbf{X}$ is a real symmetric matrix, it can be decomposed as $\mathbf{X}'\mathbf{X} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$ such that $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1, \ldots, \lambda_p$ being the eigenvalues of $\mathbf{X}'\mathbf{X}$ and $\mathbf{P}$ being the matrix whose columns, $\mathbf{v}_1, \ldots, \mathbf{v}_p$, are orthonormal eigenvectors corresponding to $\lambda_1, \ldots, \lambda_p$. Suppose we want to estimate $\boldsymbol{\rho}'\mathbf{P}'\boldsymbol{\beta}$ for some vector $\boldsymbol{\rho}$. The least squares estimate is

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

with variance

$$\begin{aligned}
Var(\boldsymbol{\rho}'\mathbf{P}'\hat{\boldsymbol{\beta}}) &= \boldsymbol{\rho}'\mathbf{P}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{P}\boldsymbol{\rho} \\
&= \sigma^2\boldsymbol{\rho}'\mathbf{P}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{P}\boldsymbol{\rho} \\
&= \sigma^2\boldsymbol{\rho}'\boldsymbol{\Lambda}^{-1}\boldsymbol{\rho} \\
&= \sigma^2\boldsymbol{\rho}'\begin{pmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\lambda_p} \end{pmatrix}\boldsymbol{\rho}.
\end{aligned}$$

From the last expression, it is obvious that the variance of the estimate would be large if any of the λ_i's is small but ρ_i (the i^{th} component of $\boldsymbol{\rho}$) is not zero or compatible to $\sqrt{\lambda_i}$. In other words, linear functions of $\boldsymbol{\beta}$ in the form $\boldsymbol{\rho}'\mathbf{P}'\boldsymbol{\beta}$ will be accurately estimated only if $\boldsymbol{\rho}'\mathbf{P}'$ is a linear combination of eigenvectors of $\mathbf{X}'\mathbf{X}$ associated with relatively large eigenvalues.

The last statement can be generalized to any linear function of $\boldsymbol{\beta}$ since, for any vector c,

$$\begin{aligned} \mathbf{c}'\boldsymbol{\beta} &= \mathbf{c}'\mathbf{P}\mathbf{P}'\boldsymbol{\beta} \\ &= \boldsymbol{\rho}'\mathbf{P}'\boldsymbol{\beta}, \qquad \text{where} \qquad \boldsymbol{\rho}' = \mathbf{c}'\mathbf{P}. \end{aligned}$$

Therefore, all linear functions of $\boldsymbol{\beta}$ can be written in the form $\boldsymbol{\rho}'\mathbf{P}'\boldsymbol{\beta}$. Consequently, for any $\mathbf{c}$, $\mathbf{c}'\boldsymbol{\beta}$ can be accurately estimated only if $\mathbf{c}$ can be approximated by a linear combination of the eigenvectors corresponding to relatively large eigenvalues. In particular, the estimate of the j^{th} component of $\boldsymbol{\beta}$, $\hat{\beta}_j = \mathbf{d}_j\hat{\boldsymbol{\beta}}$ (where $\mathbf{d}_j$ is a vector with its j^{th} component equals 1 and the rest zeroes), and the prediction of the future response value at $\mathbf{x}_0$, $\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$, will not be precise if the projections of $\mathbf{d}_j$ and $\mathbf{x}_0$ onto the space spanned by the eigenvectors associated with small eigenvalues is not much smaller in magnitude than their projections onto the space spanned by the eigenvectors associated with larger eigenvalues. (See Silvey [1969] for a more detailed discussion on imprecise estimation caused by multicollinearity.)

The problem now is to devise a way to detect collinearity. First of all, given the matrix $\mathbf{X}$, we should scale the columns so that they are compatible in length. The reason for doing so is that it should make no difference to the resulting model whether, say the mass of an object, is measured in kg or g. However, it makes a difference numerically and affects the procedures that we are going to discuss. One common way to scale $\mathbf{X}$ is to normalize its columns. Another way centers the independent variables first so that the model is (assuming there is a constant term)

$$E(y_i) = \beta_0 + (x_{i1} - \bar{x}_1)\beta_1' + \cdots + (x_{i,p-1} - \bar{x}_{p-1})\beta_{p-1}',$$

$$\text{where} \qquad \bar{x}_j = \frac{1}{n}\sum_{1=1}^{n} x_{ij}$$

$$\Rightarrow E(\mathbf{Y}) = \beta_0 + \mathbf{X}_c\boldsymbol{\beta}_c.$$

The independent variables are then standardized so that

$$E(y_i) = \beta_0 + \frac{x_{i1} - \bar{x}_1}{\sqrt{\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)^2}}\beta_1^* + \cdots + \frac{x_{i,p-1} - \bar{x}_{p-1}}{\sqrt{\sum_{j=1}^{n}(x_{j,p-1} - \bar{x}_{p-1})^2}}\beta_{p-1}^*$$

$$\Rightarrow E(\mathbf{Y}) = \beta_0\mathbf{1}_n + \mathbf{X}_s\boldsymbol{\beta}_s.$$

Multicollinearity analysis can then be applied to the standardized matrix $\mathbf{X}_s$. (To follow the convention for models with a constant term, we have used subscripts $0,\ldots,p-1$ instead of $1,\ldots,p$ for the p components of $\boldsymbol{\beta}$. We have also use x_i instead of x_{i+1} to denote the variable corresponding to the $i+1^{th}$ column of the matrix $\mathbf{X}$.)

One way to detect multicollinearity is by means of the variance inflation factor, VIF (see Neter *et al.* [1985]), which is defined to be

$$VIF_i = (\mathbf{R}^{-1})_{ii} \qquad \text{for} \qquad 1 \le i \le p-1,$$

where $\mathbf{R}$ is the correlation matrix for $x_1,\ldots,x_{p-1}$ in the standardized model. Note that, for the standardized model $\mathbf{Y} = \beta_0\mathbf{1}_n + \mathbf{X}_s\boldsymbol{\beta}_s + \boldsymbol{\varepsilon}$ with $Cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$,

$$\begin{aligned}
&Var(\hat{\boldsymbol{\beta}}_s)\\
=&Var[(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{Y}]\\
=&(\mathbf{X}_s'\mathbf{X}_s)^{-1}\sigma^2\mathbf{I}\mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\\
=&\sigma^2(\mathbf{X}_s'\mathbf{X}_s)^{-1}
\end{aligned}$$

$$=\sigma^2\left[\begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2}} & \cdots & \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2}} \\ \vdots & \ddots & \vdots \\ \frac{x_{1.p-1}-\bar{x}_{p-1}}{\sqrt{\sum_{j=1}^{n}(x_{j.p-1}-\bar{x}_{p-1})^2}} & \cdots & \frac{x_{n.p-1}-\bar{x}_{p-1}}{\sqrt{\sum_{j=1}^{n}(x_{j,p-1}-\bar{x}_{p-1})^2}} \end{pmatrix}\begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2}} & \cdots & \frac{x_{1,p-1}-\bar{x}_{p-1}}{\sqrt{\sum_{j=1}^{n}(x_{j,p-1}-\bar{x}_{p-1})^2}} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2}} & \cdots & \frac{x_{n,p-1}-\bar{x}_{p-1}}{\sqrt{\sum_{j=1}^{n}(x_{j,p-1}-\bar{x}_{p-1})^2}} \end{pmatrix}\right]^{-1}$$

$$=\sigma^2\begin{pmatrix} \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2} & \cdots & \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i,p-1}-\bar{x}_{p-1})}{\sqrt{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2\sum_{j=1}^{n}(x_{j,p-1}-\bar{x}_{p-1})^2}} \\ \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sqrt{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2\sum_{j=1}^{n}(x_{j2}-\bar{x}_2)^2}} & \cdots & \frac{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(x_{i,p-1}-\bar{x}_{p-1})}{\sqrt{\sum_{j=1}^{n}(x_{j2}-\bar{x}_2)^2\sum_{j=1}^{n}(x_{j,p-1}-\bar{x}_{p-1})^2}} \\ \vdots & \ddots & \vdots \\ \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i,p-1}-\bar{x}_{p-1})}{\sqrt{\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2\sum_{j=1}^{n}(x_{j,p-1}-\bar{x}_{p-1})^2}} & \cdots & \frac{\sum_{i=1}^{n}(x_{i,p-1}-\bar{x}_{p-1})^2}{\sum_{j=1}^{n}(x_{j,p-1}-\bar{x}_{p-1})^2} \end{pmatrix}^{-1}$$

$$=\sigma^2\begin{pmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{12} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,p-1} & r_{2,p-1} & \cdots & r_{p-1,p-1} \end{pmatrix}^{-1}$$

$$=\sigma^2\mathbf{R}^{-1}.$$

Therefore,

$$Var[i^{th}\ component\ of\ \hat{\boldsymbol{\beta}}_s]=\sigma^2 VIF_i,$$

and it will be large if VIF_i is large, hence the name variance inflation factor.

It is possible to show that†

$$VIF_i=\frac{1}{1-R_i^2},$$

† See Appendix A for proof.

where R_i is the coefficient of multiple determination when the i^{th} column of $\mathbf{X}_s$ is regressed upon the other $p-2$ columns. If a linear relation exists between x_i and some other variables in the model, R_i will be near 1 and VIF_i will be large. Therefore, a large VIF_i value indicates a problem of multicollinearity. A common practice is to take $\max_{1 \leq i \leq p-1} VIF_i > 10$ as an indication of the existence of near dependency among the columns $\mathbf{X}_s$.

Belsley *et al.* [1980] noted that a shortcoming of the use of the VIF is that it cannot distinguish the difference between one linear dependency and the coexistence of several linear dependencies. For example, high values of $VIF_1, \ldots, VIF_4$ can result from a single linear relationship between x_1, x_2, x_3 and x_4, or two linear relations, one between x_1 and x_2 and one between x_3 and x_4.

A better method for detecting multicollinearity is based on the singular values of $\mathbf{X}$. (See Belsley *et al.* [1980].) By the Singular-Value Decomposition, the $p \times p$ matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{UDV}',$$

where $\mathbf{U}$ is an $n \times p$ matrix whose columns are orthonormal eigenvectors of $\mathbf{XX}'$,

$\mathbf{D} = diag(\sqrt{\lambda_i})$ is an $p \times p$ matrix, and

$\mathbf{V}$ is an $p \times p$ matrix whose columns are orthonormal eigenvectors of $\mathbf{X}'\mathbf{X}$.

The quantities $\mu_i = \sqrt{\lambda_i}$, $1 \leq i \leq p$, are called the singular values of $\mathbf{X}$. Since $\mathbf{V}$ is an orthogonal matrix,

$$\mathbf{XV} = \mathbf{UDV}'\mathbf{V} = \mathbf{UD}.$$

If μ_i is near zero, then

$$\mathbf{X}_1 v_{1i} + \mathbf{X}_2 v_{2i} + \cdots + \mathbf{X}_p v_{pi} = \mu_i \mathbf{U}_i \approx \mathbf{0},$$

where $\mathbf{X}_j$ and $\mathbf{U}_j$ represent the j^{th} column of $\mathbf{X}$ and $\mathbf{U}$ respectively, and v_{jk} denotes the jk^{th} entry of $\mathbf{V}$. Therefore, each small value of μ_i corresponds to a near linear dependency between the columns of $\mathbf{X}$. This agrees with the fact that small λ_i are problematic, as discussed in the beginning of this section. Note that if there are r small μ_i, there will be r near linear dependencies between the columns of $\mathbf{X}$. To see how small should μ_i be for it to be considered problematic, Belsley *et al* recommended using the *condition index*

$$\eta_i = \frac{\max_{1 \leq j \leq p} \mu_j}{\mu_i}$$

instead of μ_i. Since $\sum_{i=1}^{p} \mu_i^2 = \sum_{i=1}^{p} \lambda_i = Trace(\mathbf{X}'\mathbf{X}) = p$ (the last equality holds because columns of $\mathbf{X}$ are assumed to be normalized), not all μ_i can be small simultaneously. For this reason, a small μ_i will result in a large η_i. By their experience, weak dependencies are associated with η_i around 10, whereas moderate to strong relations are associated with η_i of 30 to 100.

Once a linear dependency is detected, the independent variables that are involved can be found by using the variance-decomposition proportion defined below. Since

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}')^{-1} = \sigma^2\mathbf{P}\boldsymbol{\Lambda}^{-1}\mathbf{P}' \\ &= \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}', \end{aligned}$$

we have $Var(\hat{\beta}_i) = \sigma^2 \sum_{j=1}^{p} \frac{v_{ij}^2}{\mu_j^2}$. It is then obvious that the quantity, called the variance decomposition proportion, defined by

$$\pi_{ji} = \frac{v_{ij}^2/\mu_j^2}{\sum_{j=1}^{p} v_{ij}^2/\mu_j^2},$$

represents the proportion of the variance of $\hat{\beta}_i$ associated with μ_j. For a high value of η_i, the variables $x_{j_1}, x_{j_2}, \ldots, x_{j_r}$ $(j_1, \ldots, j_r \in \{1, \ldots, p\})$ are considered to be nearly linearly dependent if π_{ij_k} is large for $1 \leq k \leq r$. Belsley *et al* used $\pi_{ij_k} > 0.5$ as a judgement for large π_{ij_k}. Such variables can be easily picked out with the help of a variance-decomposition proportion table like the one shown below.

μ	$Var(\hat{\beta}_1)$	$\cdots$	$Var(\hat{\beta}_p)$	η
μ_1	π_{11}	$\cdots$	π_{1p}	η_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
μ_p	π_{p1}	$\cdots$	π_{pp}	η_p

A few remarks are in order here:

(1) A linear relation must involve at least two variables; a single high π_{ji} value at any row cannot be used to establish linear dependence.

(2) The involvement of the variates in two or more linear relations may be confounded if the associated condition indexes are roughly equal in magnitude.

To clarify this point, I considered a hypothetical example where I related four variables in the following manner: $x_1 + x_2 \approx 0$, $x_3 + x_4 \approx 0$ (i.e. $\sum_{i=1}^{n}(x_{1i} + x_{2i})^2$ and $\sum_{i=1}^{n}(x_{3i} + x_{4i})^2$ are small). The resulting table may look somewhat like

μ_i	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	$Var(\hat{\beta}_3)$	$Var(\hat{\beta}_4)$	η_i
1.48	0.02	0.01	0.01	0.04	1
1.34	0.03	0.01	0.01	0.01	1.1
0.06	0.35	0.28	0.53	0.55	25
0.05	0.60	0.70	0.45	0.40	30

even though it is expected to be something like

μ_i	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	$Var(\hat{\beta}_3)$	$Var(\hat{\beta}_4)$	η_i
1.48	0.02	0.01	0.01	0.04	1
1.34	0.01	0.01	0.01	0.01	1.1
0.06	0.02	0.02	0.90	0.89	25
0.05	0.95	0.96	0.08	0.06	30

Although it is still clear that two linear relations exist (two η_i are considerably greater than 15), the variables that are involved in each dependency can no longer be definitely picked out.

(3) When a variable is involved in two or more linear dependencies, its involvement in the weak one may be masked by that in the strong one.

For example, suppose $x_1 + x_2 \approx 0$ and $x_2 + x_3 + x_4 \approx 0$ where it is assumed that the latter linear relation is much stronger than the first one. Then the following table may result.

μ_i	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	$Var(\hat{\beta}_3)$	$Var(\hat{\beta}_4)$	η_i
1.449	0.031	0.003	0.015	0.020	1
1.378	0.042	0.002	0.010	0.020	1.05
0.020	0.925	0.055	0.005	0.010	72
0.005	0.002	0.940	0.970	0.950	290

Although x_2 is involved in a linear relation with x_1, its involvement in such a relation did not show up in the table because its effect is being dominated by its involvement in the stronger relation with x_3 and x_4 where $\eta_i = 290$. This problem can sometimes be overcome by picking an involved variable from each dependency and regressing them on the remaining variables. For the example above, we can pick x_3 from the stronger dependency and x_1 from the weaker one and regress each on x_2 and x_4; the regression coefficient of x_2 should be significant for the regression of x_1 on x_2 and x_4, indicating a relationship between x_1 and x_2.

The advantage of using condition index over the VIF is obvious. To obtain all the VIF's, a matrix multiplication and an inversion are required. To find all η_i, a singular value decomposition and some divisions are required. With some more arithmetic, all variance decomposition proportions can be found. With a computer, both procedures can be performed in seconds. However, the information provided by the η_i and π_{ji} are more valuable than that by the VIF_i.

We shall now briefly describe the Ridge Regression, an alternative that can be used in place of the usual linear regression when multicollinearity is detected. The idea is to use $\mathbf{X}'\mathbf{X} + k\mathbf{I}$ instead of $\mathbf{X}'\mathbf{X}$ in the estimate of $\boldsymbol{\beta}$ so that, instead of $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, we have

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

where k is a non-negative constant to be determined. Returning to our original problem in estimating $\boldsymbol{\rho}'\mathbf{P}'\boldsymbol{\beta}$ (see the beginning of this section), we can now use $\boldsymbol{\rho}'\mathbf{P}'\hat{\boldsymbol{\beta}}_R$ instead. It can be shown that the mean square error for $\boldsymbol{\rho}'\mathbf{P}'\hat{\boldsymbol{\beta}}_R$ is

$$\begin{aligned}
&E[(\boldsymbol{\rho}'\mathbf{P}'\hat{\boldsymbol{\beta}}_R - \boldsymbol{\rho}'\mathbf{P}'\boldsymbol{\beta})^2] \\
&= Var(\boldsymbol{\rho}'\mathbf{P}'\hat{\boldsymbol{\beta}}_R) + [E(\boldsymbol{\rho}'\mathbf{P}'\hat{\boldsymbol{\beta}}_R) - \boldsymbol{\rho}'\mathbf{P}'\boldsymbol{\beta}]^2 \\
&= Variance \ + \ Bias \\
&= \sigma^2\boldsymbol{\rho}' \begin{pmatrix} \frac{\lambda_1}{(\lambda_1+k)^2} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{\lambda_p}{(\lambda_p+k)^2} \end{pmatrix} \boldsymbol{\rho} + \left[\boldsymbol{\rho}' \begin{pmatrix} \frac{-k}{\lambda_1+k} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{-k}{\lambda_p+k} \end{pmatrix} \mathbf{P}'\boldsymbol{\beta} \right]^2 .
\end{aligned}$$

Note that the problem of inflated variances in the presence of small eigenvalues can be eliminated by choosing a value of k that is considerably larger than 0. However, there is a trade off: a non-zero value of k creates bias. Nevertheless, it can be shown that there exists $k > 0$ such that the mean square error is smaller than that given by the least square estimate ($k = 0$). Unfortunately, such a k is generally unknown. A method commonly used in determining the value of k to use (say k_0) is the *ridge trace*, a simultaneous plot of the estimated regression coefficients against k; k_0 is usually chosen as the value of k beyond which the graph looks flat. (Ridge regression has been discussed by Hoerl and Kennard [1970], Neter *et al.*, and Smith and Campbell [1980], for example.)

CHAPTER 4
ANALYSIS OF CHILDREN'S AID SOCIETY EXPENDITURES DATA

4.1 Children's Aid Society Expenditures Data

The data set that we investigate deals with the per child capita expenditure by the Children's Aid Society in 44 Ontario counties and districts. The Children's Aid Society is interested in determining how this expenditures is related to the sixteen variables listed below.

x_1 : proportion of population whose mother tongue is not English or French.

x_2 : proportion of children less than 18 who are from single parent families.

x_3 : proportion of tax returns from the two lowest categories.

x_4 : proportion of GWA beneficiaries.

x_5 : proportion of legal aid cases.

x_6 : migration rate outside of municipality.

x_7 : infant mortality rate.

x_8 : criminal code offense rate.

x_9 : Juvenile Delinquent Act offense rate.

x_{10} : number of doctors/1,000 population.

x_{11} : proportion of population of (N.A.) Indian descent.

x_{12} : proportion of tenant occupied dwellings.

x_{13} : proportion of population from large families.

x_{14} : proportion of population with grade 8 education or less.

x_{15} : dependency ratio.

x_{16} : rate of incidence of births to unmarried mothers.

All rates are measured in incidents per 100 population. The data is given in Appendix B where y represents the response variable,

$y =$ per child capita expenditure (i.e. total amount spent divided by number of children in the region) on Children's Aid Society services in 1980, in dollars.

This data is analyzed in the following sections. Outlying and influential cases are discussed, as well as the problem of multicollinearity in the design matrix and the identification of the important factors influencing the response. Many of the diagnostic procedures presented in the previous two chapters will be illustrated here. For the convenience of the reader, these are listed below and their previous references are indicated.

Table 4.1 Diagnostics discussed in the analysis of the Children's Aid Society data.

	First Reference	
Statistic/Procedure	*Section*	*page*
t test for outliers	2.2	14
F test for a single added variable	2.3	19
Added variable plot	2.3	19
Normal plot	2.4	21
W' test for normality	2.4	21
Cook's distance, C_i	2.6	24
$DFFITS^2$	2.6	25
Mahalanobis distance, $\hat{D}_i^2$	2.6	26
Leverage, h_{ii}	2.6	27
$e_{[i]}$ vs. e_i plot	2.6	27
Added variable plot for transformation	3.1	32
F test for transformation	3.1	33
C_p	3.2	38
VIF_i	3.3	47
Condition number, η_i	3.3	50
Variance-decomposition proportion table	3.3	50

4.2 Analysis

The objective of this section is to identify important factors influencing the response y, and to simultaneously demonstrate some of the techniques discussed in previous chapters. The software MINITAB will be used as a tool for fitting the regression line and providing basic statistics. Analyses based on procedures discussed in previous chapters are provided by a program written by the author.

4.2.1 Identification of potentially outlying and influential cases

The estimated regression line from the full model is

$$\begin{aligned} E(y) = & -89.9 + 26.0x_1 + 355x_2 - 3.6x_3 + 188x_4 - 347x_5 - 297x_6 \\ & - 411x_7 - 92x_8 + 1646x_9 + 4760x_{10} + 173x_{11} + 84.0x_{12} \\ & - 48x_{13} + 90.9x_{14} + 204x_{15} + 58x_{16}. \end{aligned}$$

Table 4.2 gives the estimated coefficients, their standard errors, the significance level associated with a test that the variable may singly be excluded from the model (p), and the variance inflation factors (VIF); the analysis of variance table is also given. Notice that some of the coefficients are different in sign than perhaps would initially be expected; for example, the coefficient of x_3 is negative. Comments on the variance inflation factors will be given in the following section. Figure 4.1 and Figure 4.2, plots of e_i and r_i against $\hat{y}_i$, show the difference between using the ordinary residuals and the standardized residuals. Comparing the relative positions of the points, the two plots are similar; the greatest location change occurs for cases 40 and 43, which are labelled on the plots. Table 4.3 gives information on some potentially influential points. These cases can also be identified from Figure 4.3-4.6, which plot h_{ii}, $\hat{D}_i^2$, C_i, and $DFFITS_i^2$ against i. As expected, h_{ii} and $\hat{D}_i^2$ provide very similar results. With leverage values in the proximity of 0.9, both cases 6 and

Table 4.2 Estimates from fitting a full model to the Children's Aid Society Data

Variable	Coefficient	Standard Error	p	VIF
Constant	-89.93	91.48	0.334	
X1	25.97	54.26	0.636	3.0
X2	355.3	222.3	0.122	4.1
X3	-3.57	23.91	0.883	1.5
X4	188.0	404.5	0.646	4.2
X5	-346.9	546.0	0.531	1.9
X6	-297.2	139.3	0.042	1.3
X7	-410.9	658.8	0.538	2.0
X8	-91.6	436.4	0.835	4.1
X9	1646	1139	0.160	3.0
X10	4760	11786	0.689	2.8
X11	173.0	190.0	0.371	10.6
X12	83.96	64.36	0.203	5.6
X13	-47.9	223.7	0.832	7.9
X14	90.95	82.64	0.281	4.9
X15	204.3	249.9	0.421	6.7
X16	57.6	228.4	0.803	18.6

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	16	13510.4	844.4	5.24	0.000
Error	27	4354.5	161.3		
Total	43	17865.0			

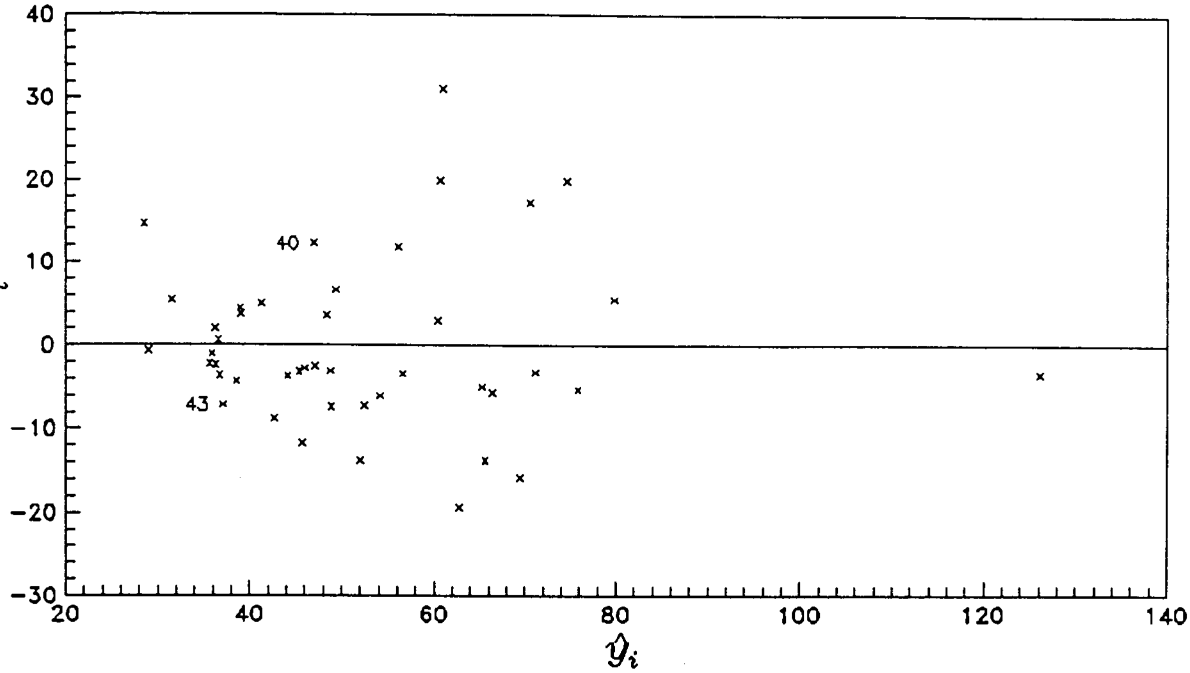

Figure 4.1 Plot of ordinary residuals, e_i, vs. fitted mean values, $\hat{y}_i$.

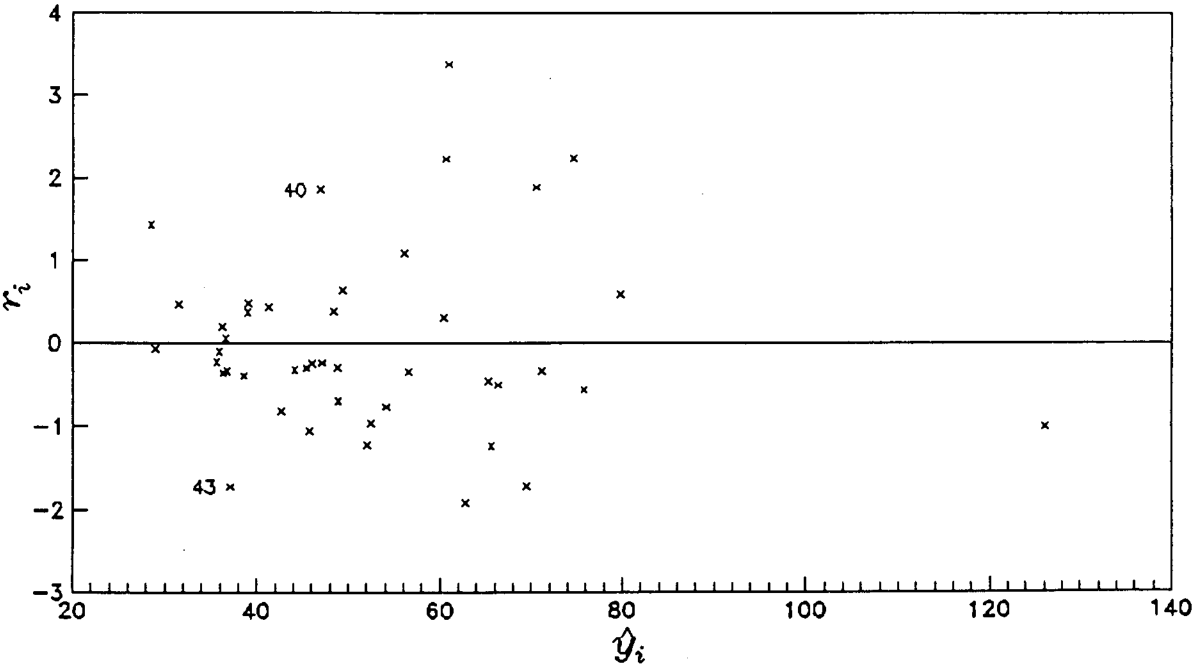

Figure 4.2 Plot of standardized residuals, r_i, vs. fitted mean values, $\hat{y}_i$.

Table 4.3 Statistics concerning potentially outlying and influential cases.

Case Number	District	Leverage	PRESS Residual	Cook's Distance	$DFFITS^2$
6	Kenora	0.9263	1.0040	0.7445	12.6606
7	Rainy River	0.4865	1.9888	0.1987	3.7469
9	Cochrane	0.4542	0.5773	0.0167	0.2774
31	York	0.7313	0.3481	0.0201	0.3298
32	Toronto	0.5094	2.4286	0.3049	6.1245
33	Frontenac	0.6501	0.9539	0.0998	1.6906
39	Ottawa-Carleton	0.4745	4.3688	0.6070	17.2317
40	Prescott & Russel	0.7339	1.9572	0.5625	10.5649
43	Durham	0.8943	1.7914	1.4770	27.1627

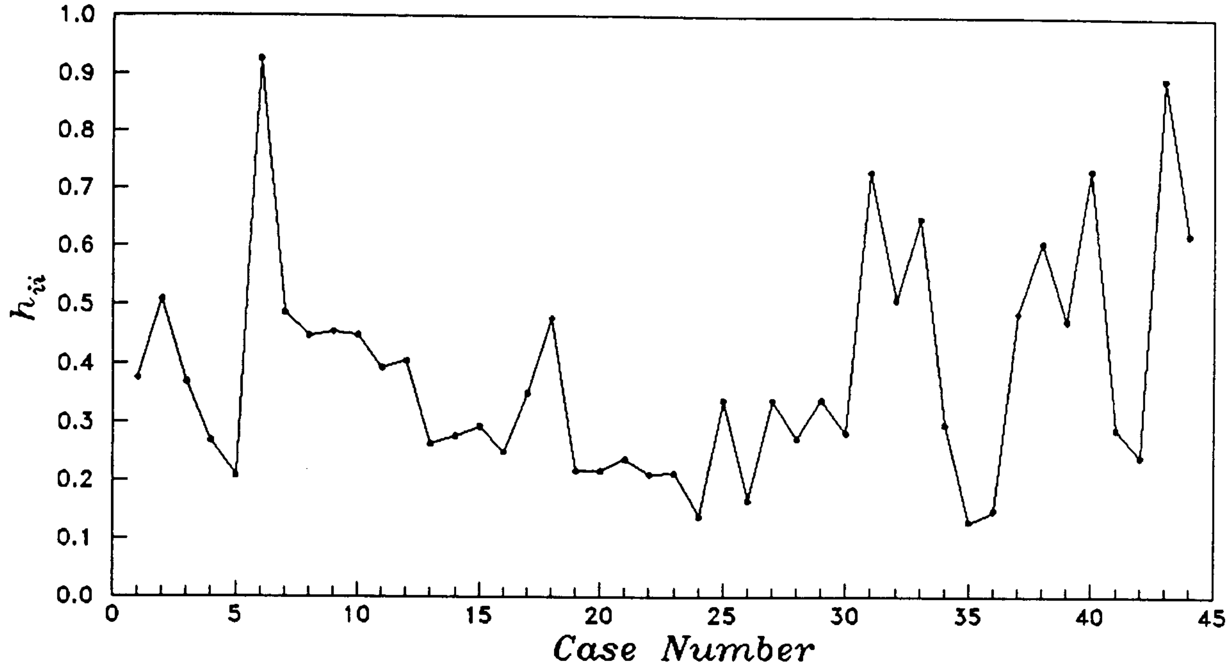

Figure 4.3 Plot of leverage, h_{ii}, vs. case number.

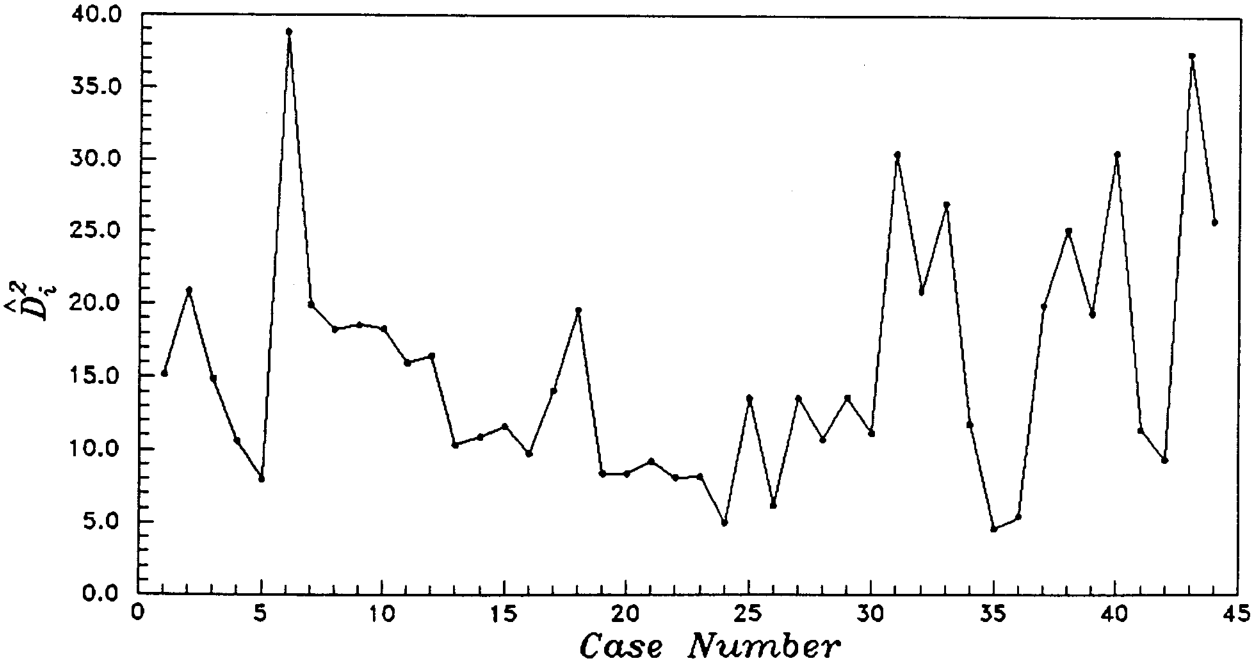

Figure 4.4 Plot of $\hat{D}_i^2$ (estimate of Mahalanobis distance) vs. case number.

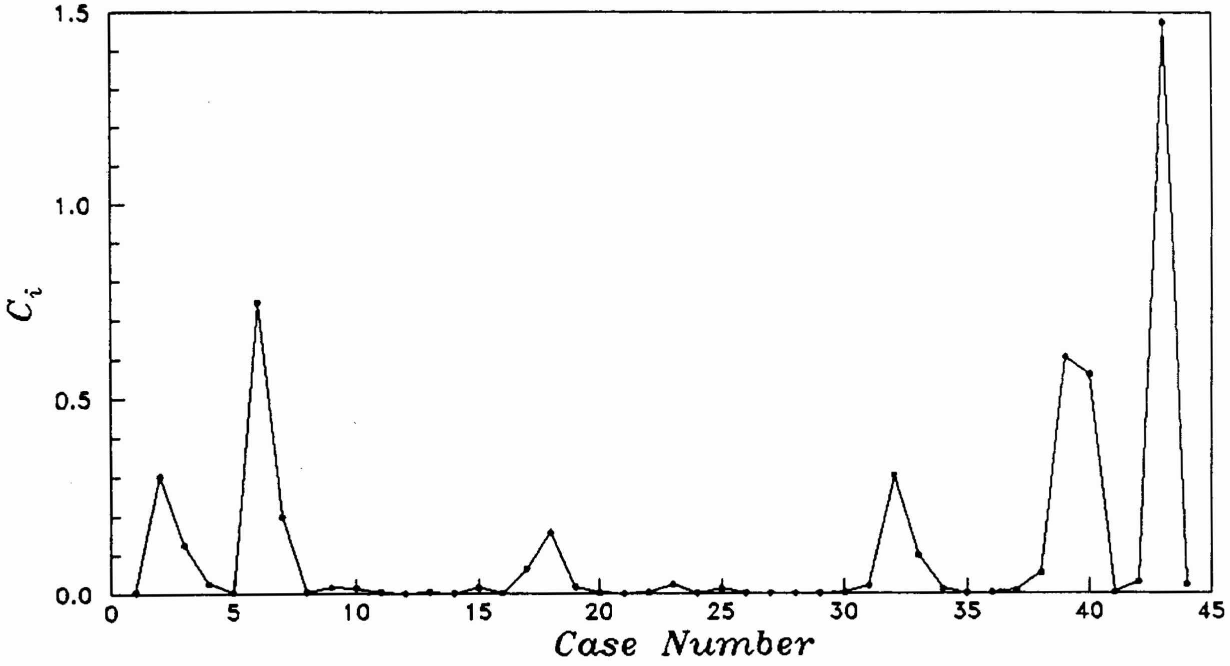

Figure 4.5 Plot of Cook's distance, C_i, vs. case number.

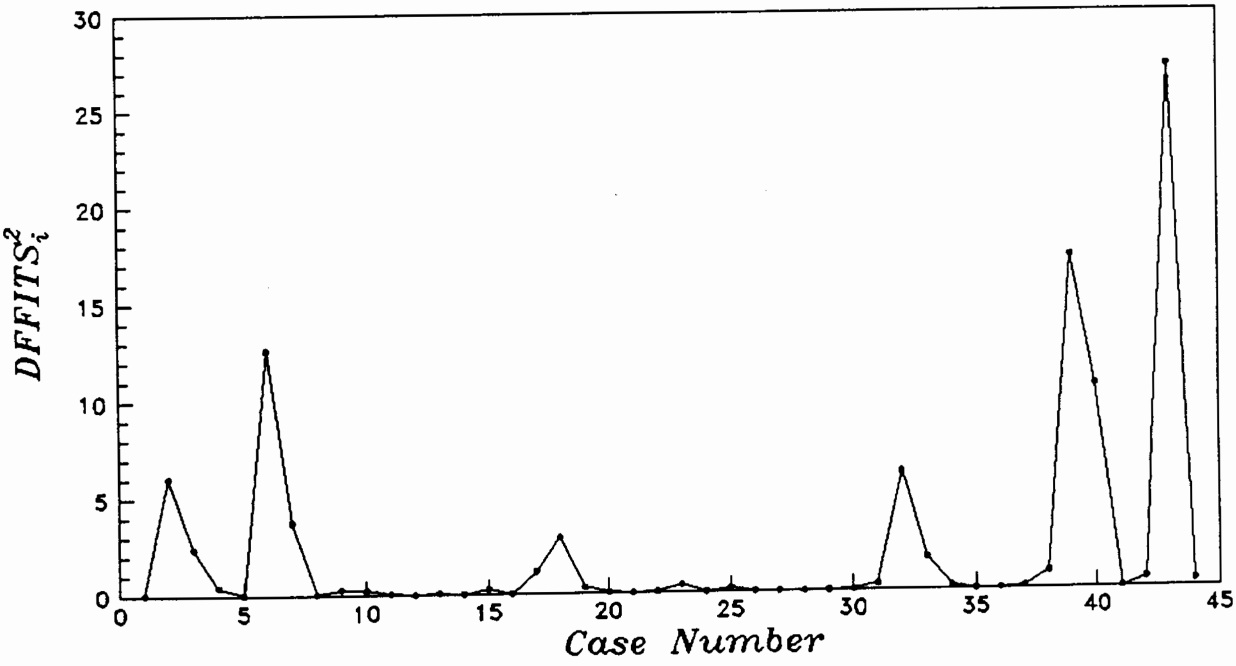

Figure 4.6 Plot of $DFFITS_i^2$ vs. case number.

43 should be considered influential since they are far from the center of the data. The next two most influential cases are 31 and 40, having leverage values larger than 0.7. Observation of the Cook's distance values (see Figure 4.5) also indicates that cases 6 and 43 are influential, with the latter case being more so. Detailed analysis indicates that case 6 has values of x_1, x_4, x_9, x_{11}, and x_{16} far outside the range of the other cases (see Appendix C). With C_i values slightly larger than 0.5, cases 39 and 40 are slightly influential. The graph of $DFFITS_i^2$ gives very similar results as that of C_i except that in Figure 4.6 case 39 stands out more because case 39 is an outlier (see the discussion in the second paragraph on p.25). Similar conclusions concerning which cases are influential can be drawn from the plot of $e_{[i]}$ vs. e_i in Figure 4.7. Cases 6 and 43 distinctly fall outside the linear trend outlined by the other data points. Note that cases 6, 40, and 43 are identified by all three (h_{ii}, C_i, and $e_{[i]}$ vs. e_i) graphs. It should also be mentioned that the potentially influential cases 7, 9, 32, and 33 included in Table 4.3 are not recognized as highly influential in these three graphs.

Figure 4.8 shows a plot of the absolute standardized PRESS residual, $|t_i|$, against case number. With $|t_{39}| = max|t_i| = 4.369$, the Bonferroni test is significant with a p-value less than 0.008. We conclude that case 39 is much different from the rest of the data and can be considered as an outlier. It is not unlikely that case 39 should behave differently from the other districts since it corresponds to the Ottawa-Carleton region.

4.2.2 The problem of multicollinearity

The correlation matrix of the independent variables is given Table 4.4. There are slight correlations between x_8 and x_{16}, x_9 and x_{11}, x_{13} and x_{14}, x_{12} and x_{15}, and x_{11} and

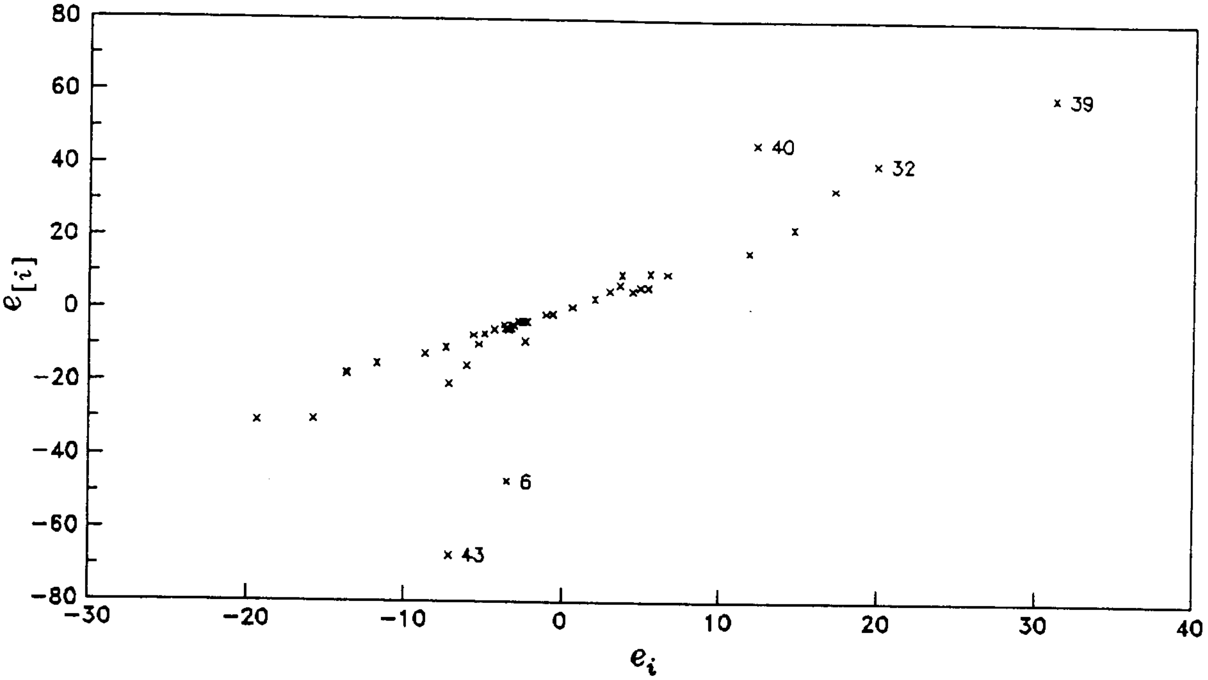

Figure 4.7 Plot of PRESS residuals, $e_{[i]}$, vs. ordinary residuals, e_i.

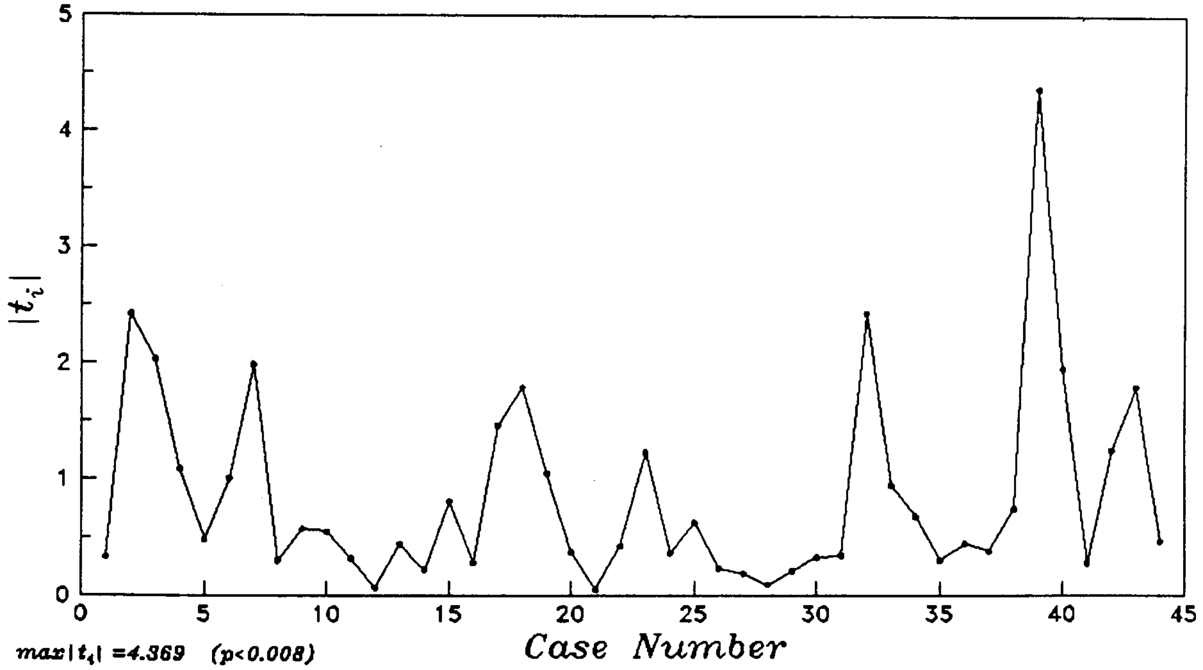

Figure 4.8 Plot of absolute standardized PRESS residuals, $|t_i|$, vs. case number. Included is the value of the maximun $|t_i|$; the quantity in parenthesis gives the significance level for testing whether the observation associated with the largest $|t_i|$ is an outlier.

Table 4.4 Correlation matrix of independent variables. Larger values are identified by asterisks.

	x1	x2	x3	x4	x5	x6	x7	x8
x2	0.254							
x3	-0.102	-0.050						
x4	0.256	0.289	0.175					
x5	0.220	0.487	-0.121	0.354				
x6	-0.211	-0.145	0.099	-0.072	-0.145			
x7	0.292	0.244	0.137	0.543	0.141	-0.022		
x8	0.322	0.277	0.045	0.469	0.396	-0.194	0.323	
x9	0.234	-0.018	0.257	0.496	0.280	0.134	0.458	0.437
x10	0.214	0.453	-0.081	0.210	0.349	-0.207	0.236	-0.023
x11	0.413	0.071	0.097	0.673	0.236	-0.023	0.516	0.523
x12	0.474	0.576	-0.106	0.194	0.261	-0.230	0.154	-0.008
x13	-0.256	-0.130	0.120	0.381	0.008	0.282	0.283	0.196
x14	-0.130	0.026	0.144	0.362	0.061	0.280	0.229	0.244
x15	-0.363	-0.447	0.091	-0.025	-0.219	0.286	0.032	0.127
x16	0.368	0.457	0.146	0.645	0.422	-0.094	0.572	0.723*

	x9	x10	x11	x12	x13	x14	x15
x10	-0.029						
x11	0.735*	0.027					
x12	-0.118	0.678	-0.040				
x13	0.407	-0.367	0.394	-0.329			
x14	0.281	-0.375	0.270	-0.336	0.775*		
x15	0.225	-0.602	0.210	-0.792*	0.609	0.646	
x16	0.604	0.104	0.780*	0.164	0.489	0.403	0.087

x_{16}. Note further that two VIF_i values are greater than 10 (see Table 4.2). The variance decomposition proportions are given in Table 4.5, and these provide a deeper insight into the problem of multicollinearity. There are three condition numbers that are significantly greater than 30, indicating that there are three strong linear dependencies. The strongest relation has a condition number of 255, and it involves the constant term and x_{15}, with x_{12} and x_{14} playing minor roles. The linear relation with condition number 94 seems to involve the variables x_2, x_8, x_{13}, and x_{16}; a regression of x_{13} upon other variables confirms that these are the major variables that enter the relation. Since all variance decomposition proportions for the third strongest dependency, with condition number 52, are less than 0.45, it is not clear which variables are involved. It is likely that their involvements in this relation is either masked by their involvement in the stronger relations or confounded with relations having compatible condition numbers, 32 and 35.

4.2.3 Selecting a transformation

Since the results of the rest of the analysis may be affected significantly by influential or outlying cases, we delete them from the data set initially and consider re-including them later. The cases deleted are 6, 40. 43, and 39 as the first three seem to be the most influential and the last is an outlier. Table 4.6 gives estimates from fitting a regression without these observations. Comparing the regression equations with and without these cases, we note a significant change in the regression coefficients for $x_1, x_3, x_8, x_{10}, x_{12}$, and x_{16}. The sign of $\hat{\beta}_3$ is now meaningful, but that of $\hat{\beta}_{16}$ is not. Note the significant decrease in the variance estimate (from 161.3 to 65.49). With these four points deleted, case 31 stands out as an influential point while case 32 is an outlier.

Table 4.5 Variance decomposition proportion table. Larger values of condition number are identified by asterisks.

Singular value	Const.	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	Condition Number
3.848	0.0000	0.0003	0.0000	0.0005	0.0002	0.0003	0.0002	0.0002	0.0001	0.0004	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	1
1.010	0.0000	0.0007	0.0002	0.0011	0.0024	0.0001	0.0014	0.0001	0.0000	0.0092	0.0007	0.0511	0.0002	0.0000	0.0001	0.0000	0.0002	4
0.571	0.0000	0.1096	0.0003	0.0917	0.0000	0.0121	0.0089	0.0000	0.0000	0.0240	0.0055	0.0000	0.0019	0.0004	0.0008	0.0000	0.0000	7
0.428	0.0000	0.1736	0.0004	0.2010	0.0274	0.0999	0.0008	0.0003	0.0003	0.0042	0.0017	0.0000	0.0001	0.0002	0.0003	0.0000	0.0006	9
0.392	0.0000	0.0111	0.0001	0.2561	0.1012	0.0003	0.0499	0.0031	0.0007	0.0479	0.0127	0.0016	0.0010	0.0008	0.0017	0.0001	0.0000	10
0.364	0.0000	0.0002	0.0000	0.0519	0.0308	0.2015	0.0076	0.0194	0.0000	0.2609	0.0020	0.0183	0.0000	0.0010	0.0027	0.0000	0.0002	11
0.308	0.0000	0.0240	0.0003	0.0735	0.0007	0.0651	0.0245	0.0801	0.0137	0.1318	0.0904	0.0000	0.0039	0.0002	0.0034	0.0000	0.0078	12
0.271	0.0000	0.1084	0.0032	0.0311	0.3149	0.0026	0.0122	0.0000	0.0011	0.0947	0.0578	0.1225	0.0015	0.0000	0.0024	0.0000	0.0097	14
0.262	0.0000	0.0004	0.0001	0.0253	0.0158	0.0197	0.1736	0.3394	0.0074	0.0620	0.0172	0.0988	0.0017	0.0000	0.0001	0.0000	0.0101	15
0.210	0.0002	0.0107	0.0002	0.0608	0.0265	0.3347	0.2262	0.3335	0.0236	0.2202	0.0153	0.0149	0.0144	0.0012	0.0002	0.0002	0.0001	18
0.182	0.0011	0.0045	0.0245	0.0107	0.0058	0.0894	0.2451	0.0378	0.0031	0.0406	0.0264	0.0598	0.0243	0.0015	0.0045	0.0018	0.0721	21
0.161	0.0000	0.0056	0.0010	0.0024	0.0262	0.0269	0.1964	0.0055	0.2637	0.0078	0.0109	0.0034	0.0252	0.0178	0.0420	0.0000	0.0020	24
0.015	0.9547	0.0278	0.0203	0.0026	0.1020	0.0044	0.0006	0.0144	0.0175	0.0103	0.0007	0.0689	0.2410	0.0030	0.2756	0.9563	0.0051	255*
0.041	0.0373	0.1075	0.5006	0.1732	0.1678	0.0005	0.0114	0.0178	0.4839	0.0261	0.0534	0.2440	0.0107	0.8410	0.0575	0.0373	0.7148	94*
0.075	0.0039	0.2784	0.0140	0.0127	0.1760	0.0214	0.0000	0.0758	0.1679	0.0026	0.0537	0.2263	0.4185	0.0931	0.4486	0.0029	0.1173	52*
0.120	0.0026	0.1221	0.1039	0.0002	0.0010	0.0184	0.0186	0.0561	0.0020	0.0010	0.6035	0.0873	0.0837	0.0021	0.0887	0.0012	0.0485	32
0.110	0.0000	0.0153	0.3308	0.0051	0.0012	0.1026	0.0225	0.0165	0.0151	0.0562	0.0481	0.0029	0.1717	0.0375	0.0713	0.0001	0.0114	35

Table 4.6 Estimates from fitting a full model to the Children's Aid Society Data with cases 6, 39, 40, and 43 deleted.

```
The regression equation is
E(Y) = - 130 + 139 x1 + 380 x2 + 112 x3 + 178 x4 - 404 x5 - 195 x6
       - 898 x7 + 317 x8 + 895 x9 + 19613 x10 + 377 x11 - 18.2 x12
       - 73 x13 + 75.8 x14 + 277 x15 - 80 x16


                            Standard
Variable    Coefficient       Error            p

Constant        -130.19       64.30        0.055
x1               139.47       42.04        0.003
x2                380.2       168.7        0.034
x3               111.69       42.92        0.016
x4                177.9       406.1        0.665
x5               -404.3       368.8        0.284
x6              -195.12       93.11        0.047
x7               -897.8       448.4        0.057
x8                317.2       299.7        0.301
x9                894.9       862.7        0.310
x10               19613        8109        0.024
x11               376.6       176.7        0.044
x12              -18.17       48.30        0.710
x13               -72.6       144.2        0.619
x14               75.77       66.39        0.265
x15               276.6       169.3        0.116
x16               -80.1       152.9        0.606


Analysis of Variance

SOURCE        DF          SS           MS          F        p
Regression    16     9038.51       564.91       8.63    0.000
Error         23     1506.28        65.49
Total         39    10544.79
```

Figure 4.9 gives a plot of $r_{(i)}$ vs. $E[z_{(i)}]$. The W' statistic for testing normality has a value of 0.957, which is almost significant at the 0.10 significance level. Also, the plot does not resemble a straight line. Figure 4.10 is an added variable plot for detecting whether a transformation is required (see Section 3.1). The F statistic associated with this plot has a large value of 23.218 with corresponding significance level less than 0.0001, and the graph certainly seems to follow a linear trend. Note the striking difference in significance levels corresponding to these two tests.

We consider a Box-Cox transformation to normalize the response. To estimate the parameter λ in the Box-Cox transformation, we plot maximized log likelihood for various λ in Figure 4.11. The maximum likelihood estimate of λ is approximately -0.7. Note that $\lambda = -1.0$ is well within the 95% confidence interval and, for convenience, we will therefore use the inverse transformation to normalize the data.

Figure 4.12 is a plot of $r_{(i)}$ against $E[z_{(i)}]$ for the transformed model with all variables included, while Figure 4.13 gives an added variable plot for checking whether this transformation is satisfactory. These two plots can be contrasted to the corresponding plots for the untransformed model (Figure 4.9 and 4.10). It is apparent that the transformed model follows a normal distribution more closely. The W' statistic corresponding to Figure 4.9 is 0.983 and the corresponding significance level is well above the 50 percent point. Since $max|t_i| = 2.071$, there are no apparent outliers.

4.2.4 Variable Selection

In this section we attempt to identify the variables which most highly influence the response. First, we consider all possible regressions to the transformed response.

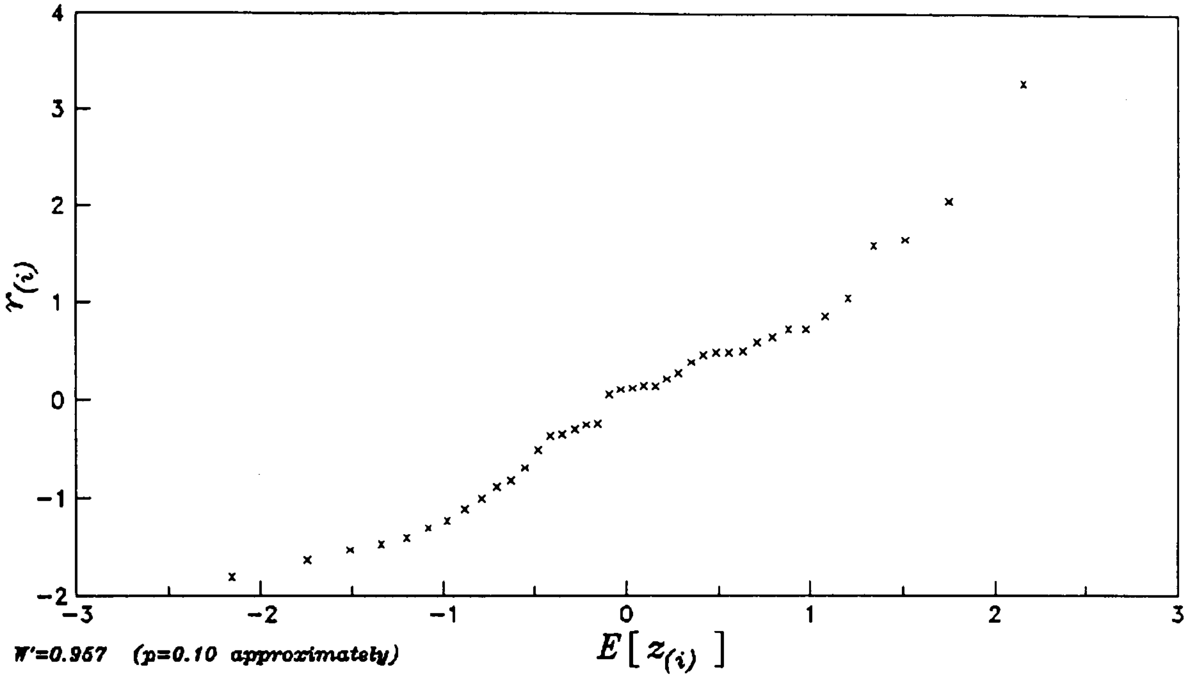

Figure 4.9 Plot of ordered standardized residuals, $r_{(i)}$, vs. expected values of normal order statistics, $E[z_{(i)}]$. Included is the associated test statistic for normality, W'; the quantity in parenthesis gives the significance level for the test.

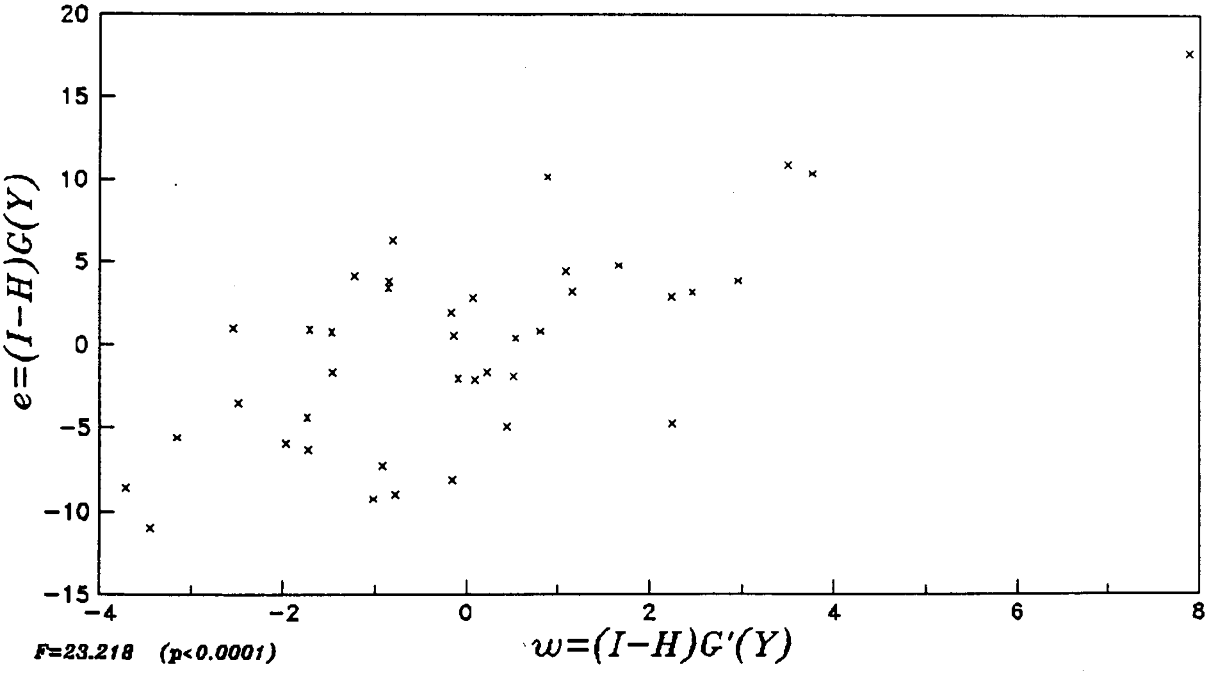

Figure 4.10 Added variable plot for detecting the need for a transformation ($\lambda_0 = 1$) for the full model. Included is the associated F test statistic for testing that a transformationis required; the quantity in parenthesis gives the significance level for the test.

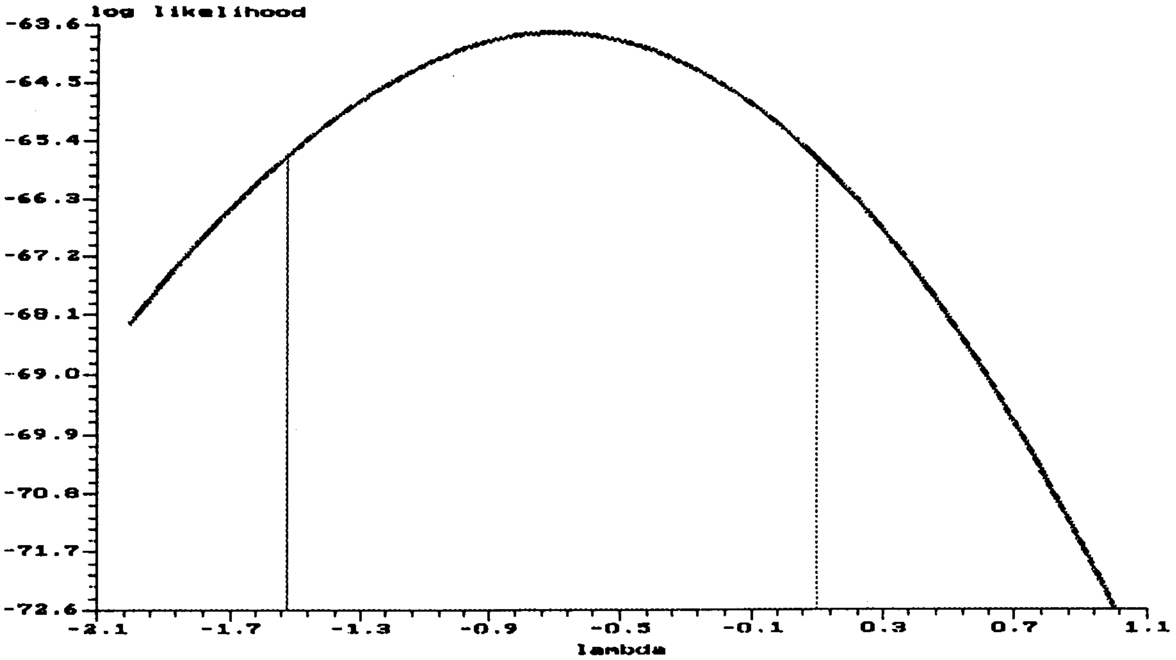

Figure 4.11 Plot of maximum log likelihood vs. λ. The vertical lines form an approximate 95 percent confidence interval for λ.

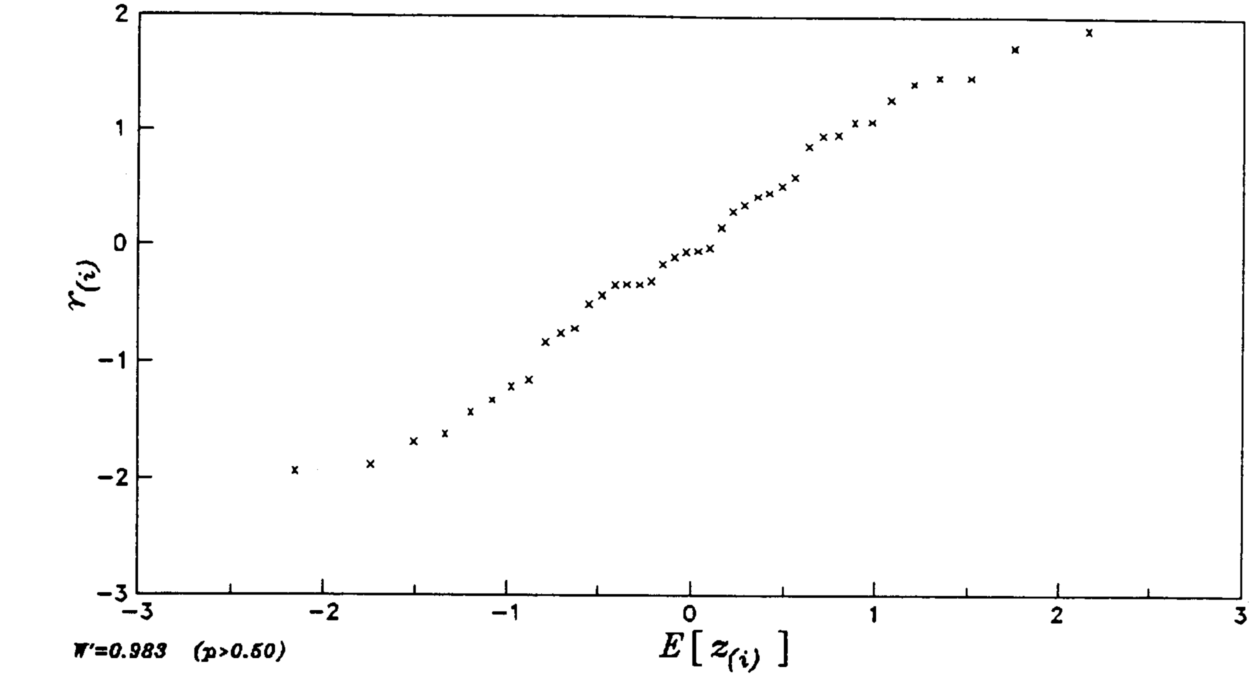

Figure 4.12 Plot of ordered standardized residuals, $r_{(i)}$, vs. expected values of normal order statistics, $E[z_{(i)}]$, for the transformed model. Included is the associated test statistic for normality, W'; the quantity in parenthesis gives the significance level for the test.

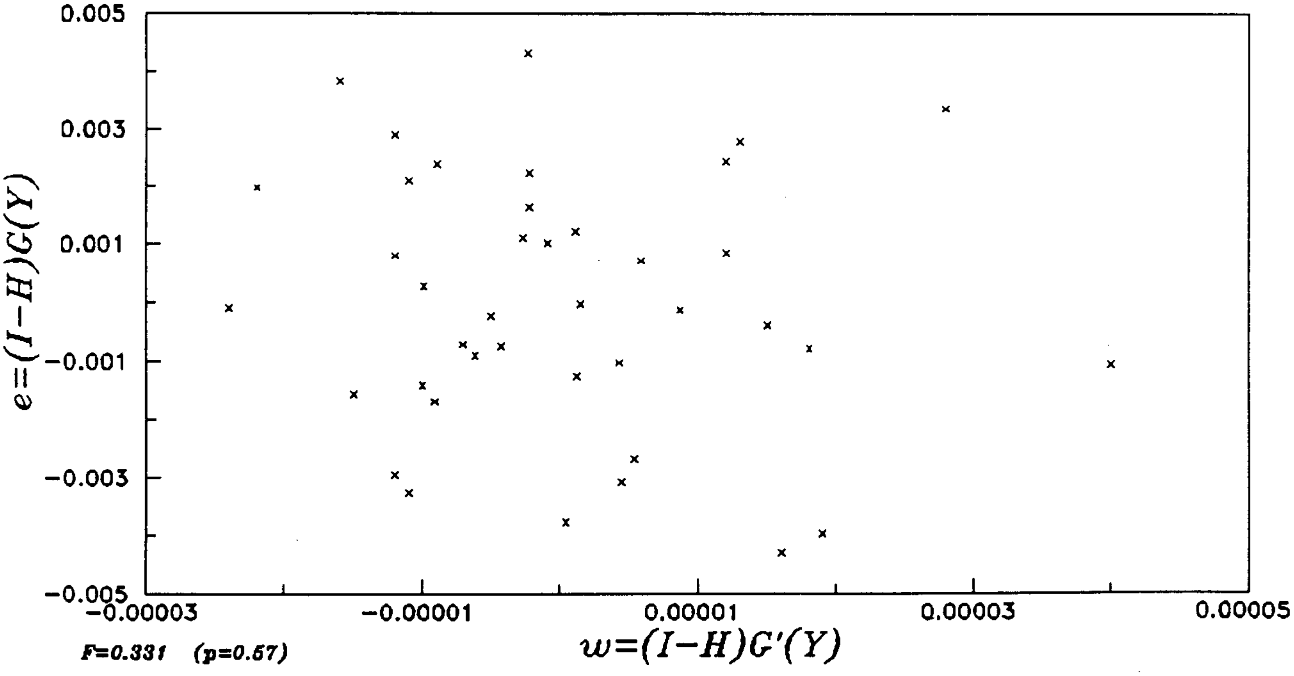

Figure 4.13 Added variable plot for detecting the need for a further transformation ($\lambda_0 = 1$) for the transformed model. Included is the associated F test statistic for testing that a transformation is required; the quantity in parenthesis gives the significance level for the test.

Reducing the number of explanatory variables in the model may remove some of the problems encountered above, especially multicollinearity. All submodels contain a constant term. Many of the models with low C_p values include $x_1, x_2, x_3, x_6, x_7, x_{11}, x_{14}$, and x_{15}, with x_1, x_2, x_3 and x_{11} included most often.

Based on the C_p statistic and with a consideration of parsimony, eight reasonable submodels are shown in Table 4.7 together with some of their respective diagnostic and influence statistics. Testing for normality using the W' statistic shows that all models have errors which are reasonably approximated by the normal distribution. The last four models provide small $C_p - p$ values. Note that x_1, x_2, x_3, and x_{11} are common to all these models.

To illustrate the technique of added variable plots and to test the importance of x_2, an added variable plot of x_2 for model 1 is displayed in Figure 4.14. Both the linear trend and the highly significant F value strongly suggest keeping x_2 in the model. Figure 4.15 shows the added variable plot of x_4 for the same model (with x_2 included). With a more or less random pattern in the plot and a non-significant F value, adding x_4 to the model does not significantly improve the model.

4.2.5 A working model

Before comparing the models in Table 4.7, we should return to the (four) influential and outlying cases that we have deleted and try to re-include them. Out of the four cases, only case 40 can be re-included in models 1, 2, 3, 4, and 6 without significantly altering the estimates. Cases 6 and 43 are still highly influential; similarly, case 39 is again an outlying value.

Table 4.7 Models selected from all possible regressions on the transformed response.

Model	Variables included	C_p-p	Maximum Cook's Distance		Maximum $\lvert t_i \rvert$		Maximum
			i	C_i	i	$\lvert t_i \rvert$	Codition Number
1	$x_1x_2x_3x_6x_{11}x_{14}$	5.694	9	0.83	2	2.389	23
2	$x_1x_2x_3x_6x_7x_{11}x_{14}$	3.805	9	0.81	2	2.668	24
3	$x_1x_2x_3x_6x_7x_{10}x_{11}x_{14}$	1.079	9	0.82	2	2.756	30
4	$x_1x_2x_3x_4x_6x_7x_{10}x_{11}x_{14}$	1.574	9	0.98	2	2.636	32
5	$x_1x_2x_3x_7x_{10}x_{11}x_{14}x_{15}$	0.433	9	0.30	30	2.982	159
6	$x_1x_2x_3x_6x_7x_{10}x_{11}x_{14}x_{15}$	-1.618	9	0.69	30	2.530	170
7	$x_1x_2x_3x_4x_7x_{10}x_{11}x_{15}$	-0.634	9	0.69	30	2.883	131
8	$x_1x_2x_3x_4x_6x_7x_{10}x_{11}x_{15}$	-1.112	9	1.10	30	2.612	138

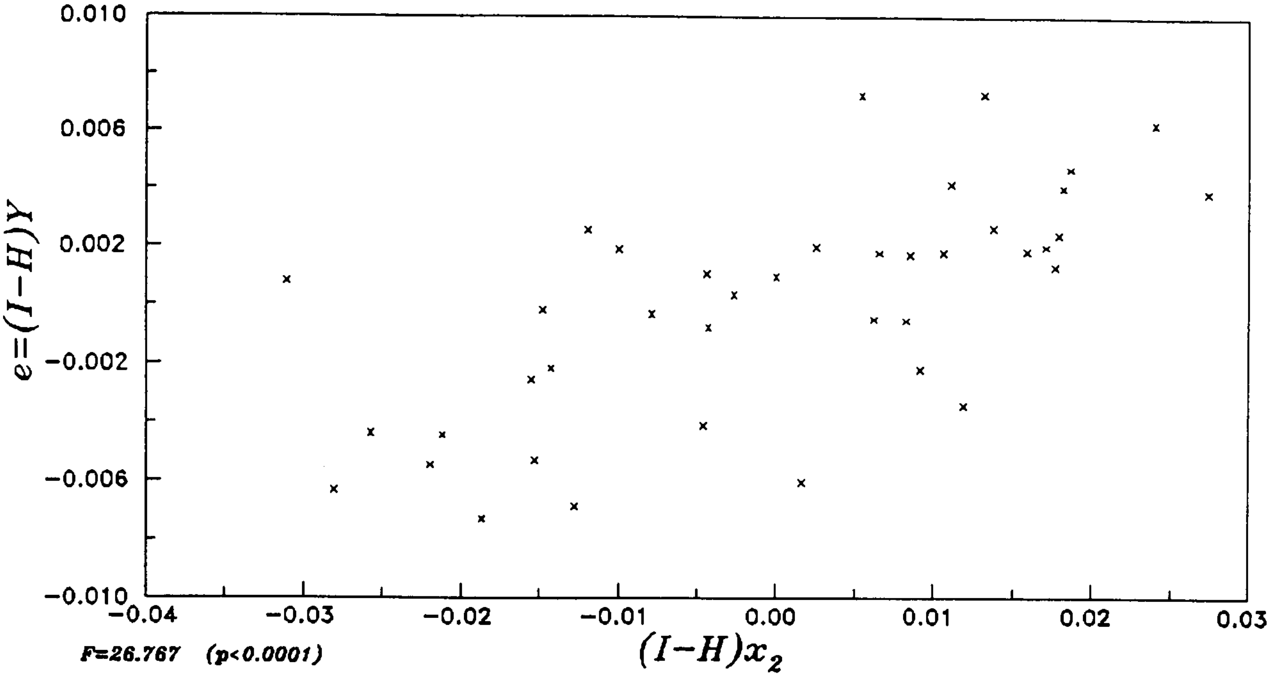

Figure 4.14 Added variable plot for adding x_2 to the model with x_1, x_3, x_6, x_{11}, and x_{14}. The associated F test statistic for including x_2 in the modelis is stated; the quantity in parenthesis gives the significance level for the test.

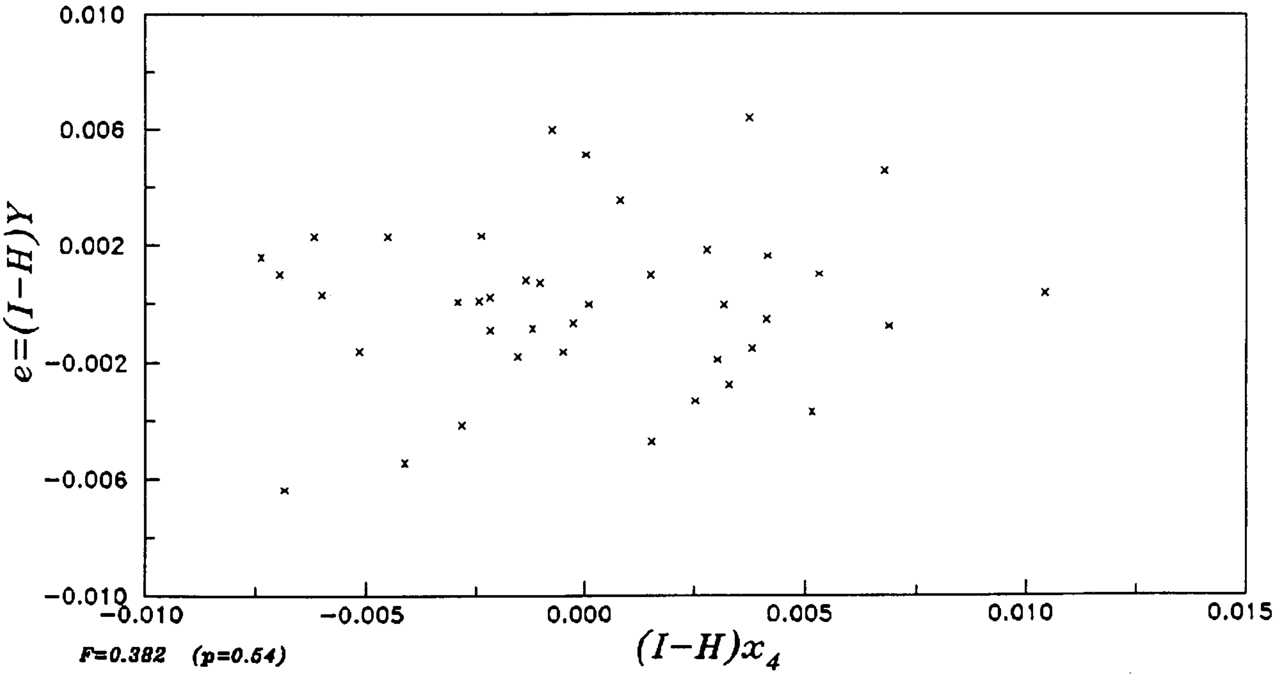

Figure 4.15 Added variable plot for adding x_4 to the model with $x_1, x_2, x_3, x_6, x_{11}$, and x_{14}. The associated F test statistic for including x_4 in the model is stated; the quantity in parenthesis gives the significance level for the test.

From the previous analysis, the factors important in explaining y are x_1, x_2, x_3, x_4, x_6, x_7, x_{10}, x_{11}, x_{14}, x_{15}. All of the models in Table 4.7 provide a good fit to the data. Models 3, 5, 6, 7 and 8 all have small $C_p - p$ values; model 3 may be preferable overall since x_{15} is highly correlated with the constant term. Details of the fitted model with cases 6, 39, and 43 deleted are given in Table 4.8. Figure 4.16 and 4.17 give, respectively, the added variable plots of x_4 and x_{15} for this model. The F test for including x_4 is significant, but the plot indicates that it is affected by an isolated case. Figure 4.17 shows an opposite situation. The F test for including x_{15} is non-significant but the plot reveals that it might be influenced by case 40.

4.2.6 Discussion

The variables that seem to be important in explaining the variation in y are:

x_1 : proportion of population whose mother tongue is not English or French.

x_2 : proportion of children less than 18 who are from single parent families.

x_3 : proportion of tax returns from the two lowest categories.

x_6 : migration rate outside of municipality.

x_7 : infant mortality rate.

x_{10} : number of doctors/1,000 population.

x_{11} : proportion of population of (N.A.) Indian descent. families.

x_{14} : proportion of population with grade 8 education or less.

Note, however, that some of the other independent variables originally considered are correlated with those listed above, as discussed previously. In addition, some of the districts have values of independent variables very different from the rest of the cases. Inclusion of

Table 4.8 Estimates from fitting a submodel to the transformed Children's Aid Society data with cases 6, 39, and 43 deleted.

```
The regression equation is
E[y(-1)] = 0.946 + 0.0197 x1 + 0.118 x2 + 0.0181 x3 - 0.0771 x6
           - 0.234 x7 + 5.10 x10 + 0.109 x11 + 0.0693 x14

Predictor        Coef        Stdev          p
Constant     0.945954     0.004263      0.000
x1            0.01968      0.01003      0.058
x2            0.11824      0.03609      0.003
x3            0.01809      0.01255      0.159
x6           -0.07709      0.03247      0.024
x7            -0.2343       0.1478      0.123
x10             5.098        2.389      0.041
x11           0.10931      0.04388      0.018
x14           0.06933      0.01436      0.000


Analysis of Variance

SOURCE       DF          SS          MS          F        p
Regression    8  0.00115363  0.00014420      15.02    0.000
Error        32  0.00030720  0.00000960
Total        40  0.00146083
```

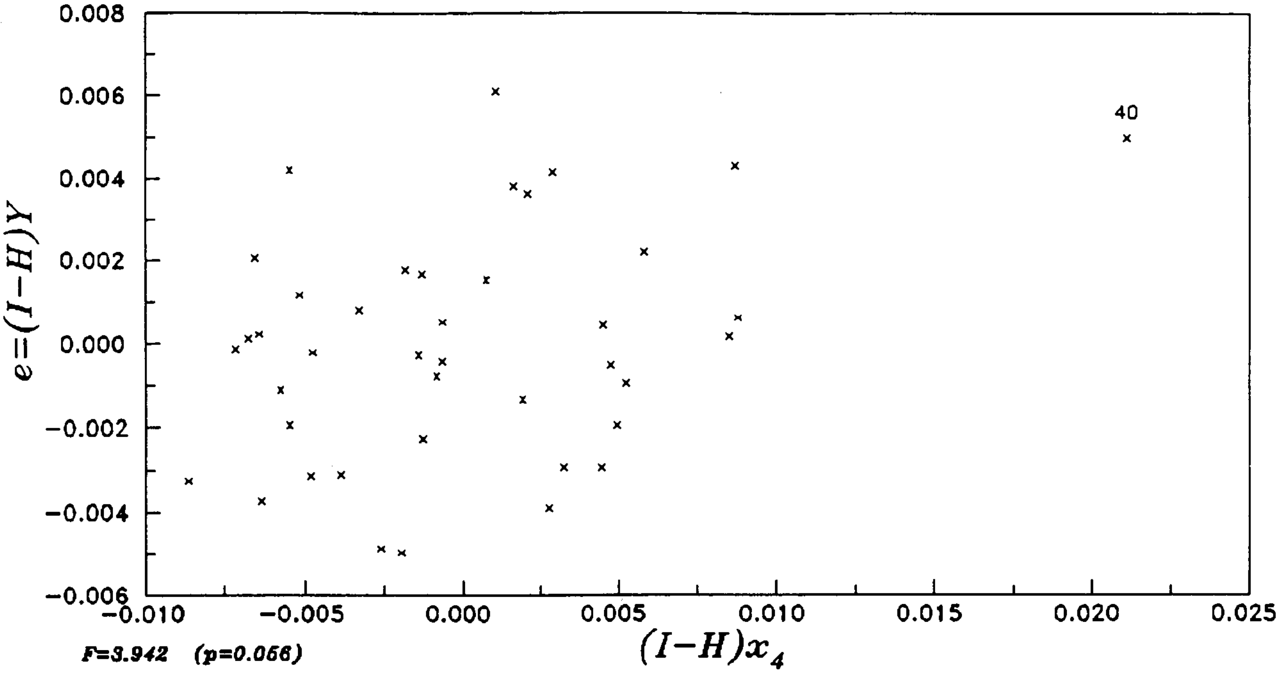

Figure 4.16 Added variable plot for adding x_4 to the model with x_1, x_2, x_3, x_6, x_7, x_{10}, x_{11}, and x_{14}. The associated F test statistic for including x_4 in the model is stated; the quantity in parenthesis gives the significance level for the test.

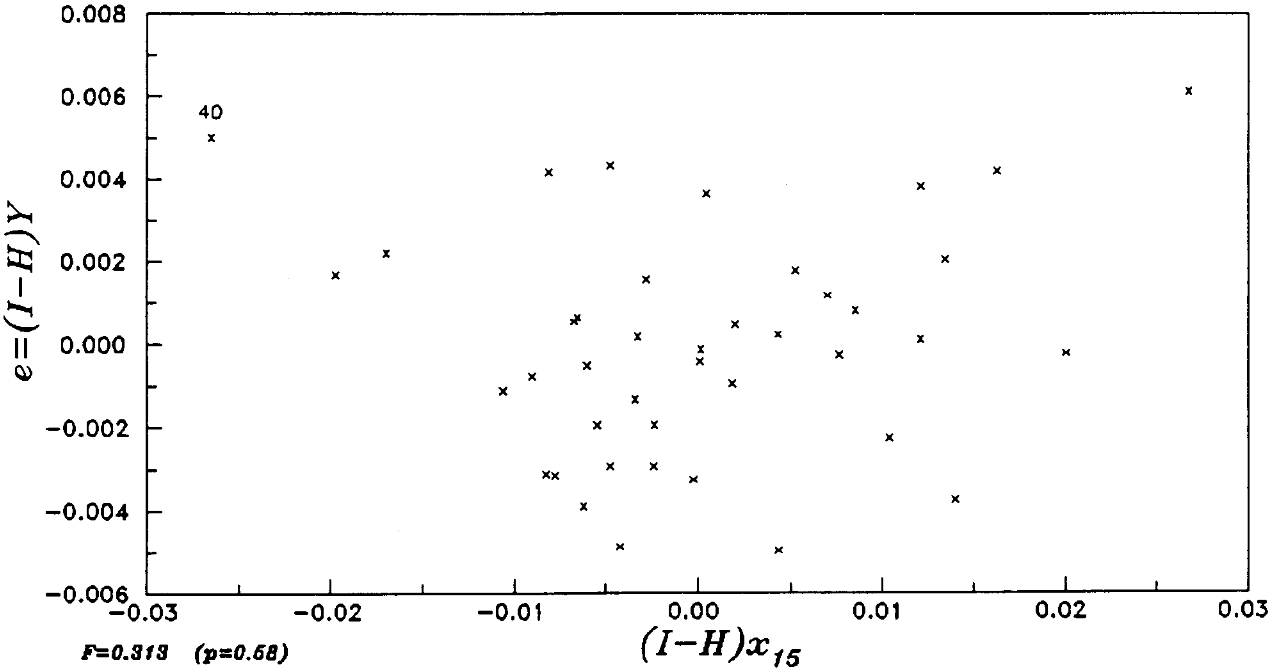

Figure 4.17 Added variable plot for adding x_{15} to the model with x_1, x_2, x_3, x_6, x_7, x_{10}, x_{11}, and x_{14}. The associated F test statistic for including x_{15} in the model is stated; the quantity in parenthesis gives the significance level for the test.

these observations would have biasing effect on the analysis. These districts are:

1. Kenora.

2. Prescott and Russel.

3. Durham.

4. Ottawa-Carleton.

The fourth of the districts above is also an outlying case since it does not fit into any of the reasonable submodels discussed in Section 4.2.4. Note that Kenora is the largest district. Also, the per child capita expenditure in the Ottawa-Carleton region is far higher than that predicted by any of the regression models studied.

CHAPTER 5
CONCLUSION

After fitting a regression line, it is important to consider the goodness of fit of the model. The procedures discussed in this project can be helpful in this regard. Note that sometimes two statistics can give very similar results, as in the case of the Cook's distance and $DFFITS^2$, or leverages and Mahalanobis distance. Considering Cook's distance and $DFFITS^2$, Cook's distance may be preferable over $DFFITS^2$ because the former statistic is invariant under nonsingular linear transformations and can be calibrated by comparison to confidence contours for $\hat{\boldsymbol{\beta}}$. Likewise, h_{ii} may be preferred over D_i since its bounded range provides an easy way to measure the 'strength' of influence of an observation.

Graphs are powerful analytical tools and they are easily understood. Some statistics, such as the S test for heteroscedasticity and the F test for determining whether a transformation is required, can be misleading if not presented together with their graphical equivalence or graphical procedures that are designed for similar purposes; the latter methods can help to identify whether a significant result is caused by a few cases or the data set as a whole.

Care should be taken when using some of the statistical procedures presented as often they require some underlying assumptions, and violation of these assumptions may lead to misleading interpretations. For example, non-constant variances may offset the W' test for normality to indicate that the errors are non- normal when they are actually normally distributed. Similarly, all F tests discussed rely on normality of the errors.

Finally, it should be mentioned that there are many other statistical procedures that are not considered here, but may be useful or necessary in some situations. Some examples are the test for lack of fit when repeated measurements are available, a consideration of transformations of independent variables, or transformations of the response values other than the Box-Cox transformation, and the assessment of influence when more than one case is deleted. Some recent work in regression goodness of fit include diagnostics for measurement-error models (Carroll and Spiegelman [1992]) and diagnostics for assessing the influence of individual cases on the estimation of the parameter in the Box-Cox transformation model (Tsai and Wu [1992]).

APPENDIX A

Proof of $VIF_i = \frac{1}{1-R_i^2}$

Consider the standardized model

$$E(\mathbf{Y}) = \beta_0\mathbf{1}_n + \mathbf{X}_s\boldsymbol{\beta}_s,$$

and define VIF_i as in Section 3.4 by

$$VIF_i = (\mathbf{R}^{-1})_{ii} \qquad \text{for} \qquad 1 \leq i \leq p-1,$$

where $\mathbf{R}$ is the correlation matrix for $x_1, \ldots, x_{p-1}$. Proving that $VIF_i = \frac{1}{1-R_i^2}$ is therefore the same as proving that the ii^{th} entry of the inverse of $\mathbf{R} = \mathbf{X}_s'\mathbf{X}_s$ is the same as $\frac{1}{1-R_i^2}$, where R_i is the coefficient of multiple determination when the i^{th} column of $\mathbf{X}_s$ is regressed upon the other $p-2$ columns. Without lost of generality, we can assume $i = 1$ and partition $\mathbf{X}_s$ as

$$\mathbf{X}_s'\mathbf{X}_s = \begin{bmatrix} \mathbf{U'U} & \mathbf{U'V} \\ \mathbf{V'U} & \mathbf{V'V} \end{bmatrix},$$

where $\mathbf{U}$ is the first column of $\mathbf{X}_s$ and $\mathbf{V}$ is $\mathbf{X}_s$ without the first column. Using the fact that

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} [\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C}]^{-1} & -\mathbf{A}^{-1}\mathbf{B}[\mathbf{D}-\mathbf{CA}^{-1}\mathbf{B}]^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}[\mathbf{A}-\mathbf{BD}^{-1}\mathbf{C}]^{-1} & [\mathbf{D}-\mathbf{CA}^{-1}\mathbf{B}]^{-1} \end{pmatrix},$$

we would have

$$(\mathbf{X}_s'\mathbf{X}_s)_{11}^{-1} = [\mathbf{U'U} - \mathbf{U'V}(\mathbf{V'V})^{-1}\mathbf{V'U}]^{-1}$$

$$
\begin{aligned}
&= \frac{1}{\mathbf{U}'\mathbf{U}\left(1 - \frac{\mathbf{U}'\mathbf{H}_{\mathbf{V}}\mathbf{U}}{\mathbf{U}'\mathbf{U}}\right)} \\
&= \frac{1}{\mathbf{U}'\mathbf{U}(1 - R_1^2)} \\
&= \frac{1}{1 - R_1^2}.
\end{aligned}
$$

The last equality follows from the fact that $\mathbf{X}_s$ is a standardized matrix.

APPENDIX B

Children's Aid Society Expenditures Data

Area	y	x1	x2	x3	x4	x5	x6	x7
1 Algoma	53.15	0.1330	0.1070	0.1558	0.0239	0.0145	0.0460	0.0110
2 Muskoka	80.55	0.0490	0.0970	0.2225	0.0222	0.0136	0.0270	0.0126
3 Nipissing	43.39	0.0490	0.1010	0.1805	0.0246	0.0180	0.0390	0.0124
4 Parry Sound	67.84	0.0430	0.1110	0.2431	0.0140	0.0152	0.0550	0.0073
5 Sudbury	60.80	0.1150	0.1120	0.1760	0.0311	0.0130	0.0580	0.0153
6 Kenora	122.62	0.2880	0.0880	0.2098	0.0579	0.0177	0.0600	0.0223
7 Rainy River	87.77	0.1290	0.0960	0.2383	0.0160	0.0142	0.0520	0.0196
8 Thunder Bay	63.29	0.0590	0.1120	0.2758	0.0225	0.0099	0.0620	0.0213
9 Cochrane	85.25	0.0860	0.0930	0.4330	0.0244	0.0141	0.0420	0.0154
10 Timiskaming	70.55	0.1970	0.1060	0.1599	0.0171	0.0114	0.0380	0.0172
11 Brant	67.97	0.1040	0.1130	0.1682	0.0211	0.0085	0.0370	0.0115
12 Halton	28.31	0.1120	0.0650	0.1268	0.0053	0.0075	0.0220	0.0125
13 Hamilton–Wentworth	60.31	0.1920	0.1190	0.1619	0.0243	0.0200	0.0440	0.0124
14 Niagara	44.66	0.1660	0.1040	0.1639	0.0193	0.0092	0.0600	0.0110
15 Elgin	33.99	0.1130	0.0920	0.2041	0.0071	0.0093	0.0800	0.0157
16 Haldimand–Norfolk	42.21	0.1180	0.0720	0.1968	0.0098	0.0087	0.0540	0.0104
17 Huron	43.11	0.0570	0.0560	0.2027	0.0046	0.0036	0.0820	0.0094
18 Middlesex	53.64	0.1200	0.1130	0.1503	0.0157	0.0072	0.0400	0.0107
19 Oxford	33.95	0.0970	0.0790	0.1612	0.0123	0.0033	0.0590	0.0122
20 Perth	34.26	0.0810	0.0680	0.1457	0.0062	0.0031	0.0650	0.0122
21 Bruce	37.21	0.0520	0.0650	0.1717	0.0123	0.0056	0.0750	0.0099
22 Grey	46.19	0.0540	0.0840	0.2115	0.0154	0.0124	0.0750	0.0141
23 Waterloo	38.31	0.1780	0.0920	0.1460	0.0154	0.0120	0.0590	0.0097
24 Wellington	43.36	0.1100	0.0780	0.1767	0.0115	0.0080	0.0600	0.0087
25 Essex	55.96	0.1740	0.1070	0.1726	0.0143	0.0149	0.0730	0.0107
26 Kent	43.21	0.0840	0.0970	0.1596	0.0090	0.0096	0.0850	0.0110
27 Lambton	38.24	0.0810	0.0860	0.1385	0.0157	0.0079	0.0800	0.0103
28 Dufferin	34.86	0.0670	0.0710	0.1134	0.0071	0.0104	0.0390	0.0082
29 Peel	33.36	0.1770	0.0700	0.1028	0.0060	0.0086	0.0200	0.0096
30 Simcoe	33.13	0.0700	0.0960	0.1795	0.0113	0.0182	0.0820	0.0104
31 York	34.00	0.1210	0.0700	0.1628	0.0065	0.0165	0.0500	0.0093
32 Toronto	94.51	0.2900	0.1210	0.1407	0.0191	0.0161	0.0470	0.0107
33 Frontenac	45.26	0.0850	0.1020	0.1894	0.0216	0.0236	0.0560	0.0130
34 Hastings	41.51	0.0430	0.1000	0.1843	0.0247	0.0174	0.0740	0.0127
35 Leeds & Grenville	40.45	0.0500	0.0900	0.1540	0.0133	0.0085	0.0500	0.0075
36 Lennox & Addington	36.95	0.0370	0.0750	0.1564	0.0161	0.0109	0.0610	0.0075
37 Prince Edward	51.93	0.0350	0.0920	0.2098	0.0082	0.0180	0.0530	0.0000
38 Lanark	48.04	0.0290	0.1130	0.1607	0.0198	0.0099	0.0590	0.0192
39 Ottawa–Carleton	92.05	0.1050	0.1210	0.1353	0.0158	0.0096	0.0500	0.0120
40 Prescott & Russel	59.16	0.0230	0.0660	0.1826	0.0432	0.0100	0.0550	0.0152
41 Renfrew	45.71	0.0660	0.0870	0.2188	0.0124	0.0076	0.0760	0.0065
42 Str., Dund. & Glen.	51.95	0.0420	0.1090	0.1905	0.0149	0.0116	0.0430	0.0076
43 Durham	29.98	0.1020	0.0780	0.7351	0.0205	0.0027	0.0620	0.0115
44 Northumberland	42.73	0.0470	0.0720	0.1762	0.0168	0.0025	0.0640	0.0046

x8	x9	x10	x11	x12	x13	x14	x15	x16
0.0495	0.0011	0.0008	0.0297	0.2678	0.1540	0.2593	0.3520	0.1328
0.0462	0.0040	0.0010	0.0130	0.2135	0.1166	0.2532	0.3882	0.0941
0.0459	0.0048	0.0007	0.0215	0.3259	0.1760	0.2560	0.3638	0.1255
0.0413	0.0045	0.0006	0.0088	0.1880	0.1541	0.3131	0.3809	0.0976
0.0431	0.0038	0.0010	0.0206	0.3496	0.1617	0.2803	0.3546	0.1074
0.0598	0.0180	0.0010	0.2100	0.2823	0.1827	0.3179	0.3838	0.2212
0.0595	0.0084	0.0008	0.0703	0.2306	0.1715	0.2775	0.3805	0.1887
0.0387	0.0068	0.0009	0.0330	0.2813	0.1429	0.2705	0.3443	0.1214
0.0424	0.0083	0.0008	0.0365	0.3478	0.1812	0.3390	0.3711	0.1391
0.0518	0.0056	0.0007	0.0075	0.3153	0.1669	0.3320	0.3772	0.1355
0.0349	0.0022	0.0008	0.0465	0.2982	0.1185	0.2550	0.3593	0.0928
0.0275	0.0017	0.0010	0.0023	0.2819	0.1064	0.1290	0.3366	0.0372
0.0402	0.0022	0.0011	0.0042	0.3793	0.1038	0.2514	0.3334	0.0916
0.0328	0.0013	0.0008	0.0026	0.2662	0.1208	0.2439	0.3523	0.0732
0.0301	0.0054	0.0006	0.0029	0.2618	0.1399	0.2602	0.3804	0.0646
0.0333	0.0024	0.0006	0.0138	0.2625	0.1447	0.3051	0.3699	0.0750
0.0175	0.0031	0.0007	0.0016	0.2083	0.1662	0.2889	0.3991	0.0339
0.0321	0.0046	0.0016	0.0085	0.3997	0.1084	0.1827	0.3361	0.0886
0.0291	0.0036	0.0007	0.0022	0.2699	0.1354	0.2727	0.3673	0.0699
0.0299	0.0028	0.0008	0.0010	0.2632	0.1547	0.2817	0.3758	0.0706
0.0385	0.0038	0.0007	0.0193	0.2010	0.1609	0.2890	0.3934	0.0605
0.0323	0.0044	0.0008	0.0015	0.2361	0.1354	0.2978	0.3814	0.0612
0.0395	0.0022	0.0009	0.0016	0.3855	0.1126	0.2434	0.3442	0.0691
0.0295	0.0023	0.0008	0.0017	0.3131	0.1413	0.2189	0.3586	0.0538
0.0319	0.0035	0.0007	0.0024	0.2879	0.1426	0.2520	0.3640	0.0826
0.0327	0.0045	0.0008	0.0049	0.2758	0.1400	0.2767	0.3712	0.0825
0.0287	0.0042	0.0006	0.0235	0.2678	0.1358	0.2075	0.3473	0.0834
0.0342	0.0024	0.0007	0.0024	0.2017	0.1182	0.2149	0.3873	0.0348
0.0361	0.0012	0.0007	0.0015	0.3055	0.0964	0.1438	0.3380	0.0481
0.0472	0.0042	0.0009	0.0053	0.2541	0.1268	0.2372	0.3638	0.0915
0.0384	0.0122	0.0007	0.0031	0.2069	0.1107	0.1676	0.3504	0.0466
0.0483	0.0037	0.0013	0.0027	0.4890	0.0907	0.2299	0.3130	0.0872
0.0323	0.0033	0.0020	0.0030	0.4288	0.1199	0.1866	0.3230	0.0865
0.0402	0.0056	0.0008	0.0138	0.3158	0.1386	0.2367	0.3576	0.0910
0.0354	0.0020	0.0010	0.0018	0.2622	0.1316	0.2233	0.3658	0.0674
0.0396	0.0023	0.0007	0.0049	0.2082	0.1381	0.2372	0.3740	0.0560
0.0323	0.0044	0.0007	0.0039	0.2299	0.1431	0.2973	0.3630	0.0692
0.0418	0.0043	0.0012	0.0024	0.2734	0.1414	0.2767	0.3626	0.0698
0.0251	0.0029	0.0013	0.0022	0.5036	0.1172	0.1348	0.3092	0.0805
0.0297	0.0050	0.0007	0.0011	0.2362	0.1875	0.3560	0.3701	0.0620
0.0323	0.0066	0.0008	0.0063	0.2728	0.1677	0.2884	0.3614	0.0816
0.0328	0.0044	0.0007	0.0072	0.2977	0.1606	0.2931	0.3682	0.0966
0.0337	0.0051	0.0008	0.0042	0.2622	0.1113	0.1958	0.3506	0.0618
0.0570	0.0041	0.0005	0.0040	0.2519	0.1347	0.2411	0.3701	0.0795

APPENDIX C

Standardized Residual Plots

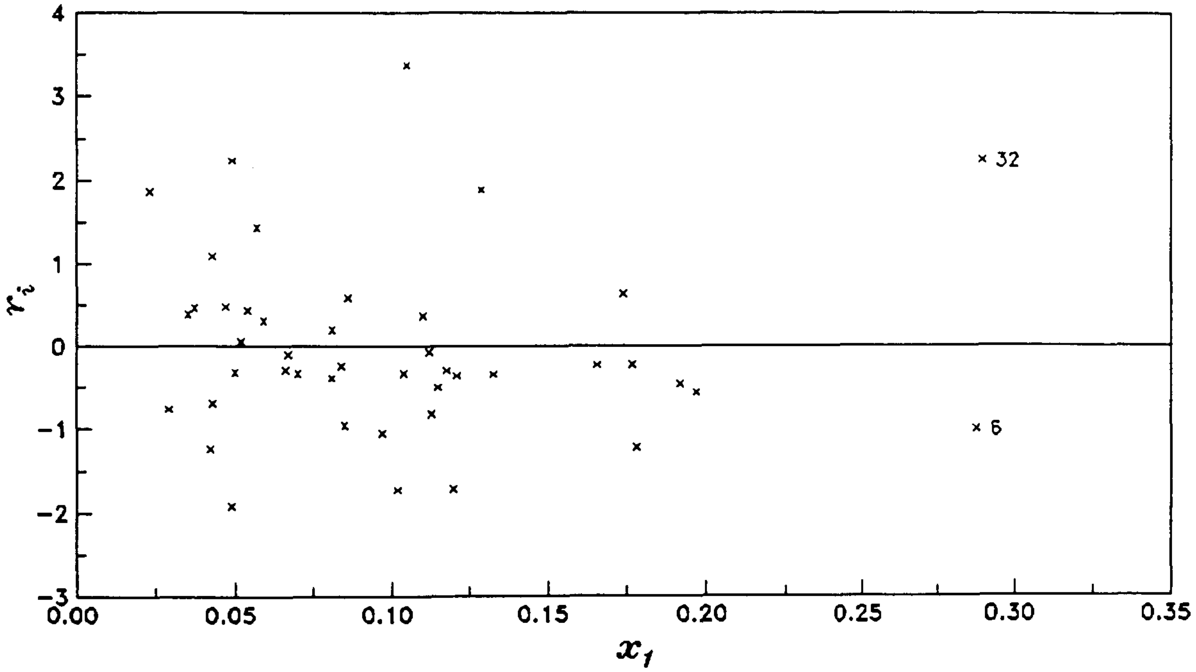

Figure C.1 Plot of standardized residuals, r_i, vs. proportion of population whose mother tongue is not English or French, x_1.

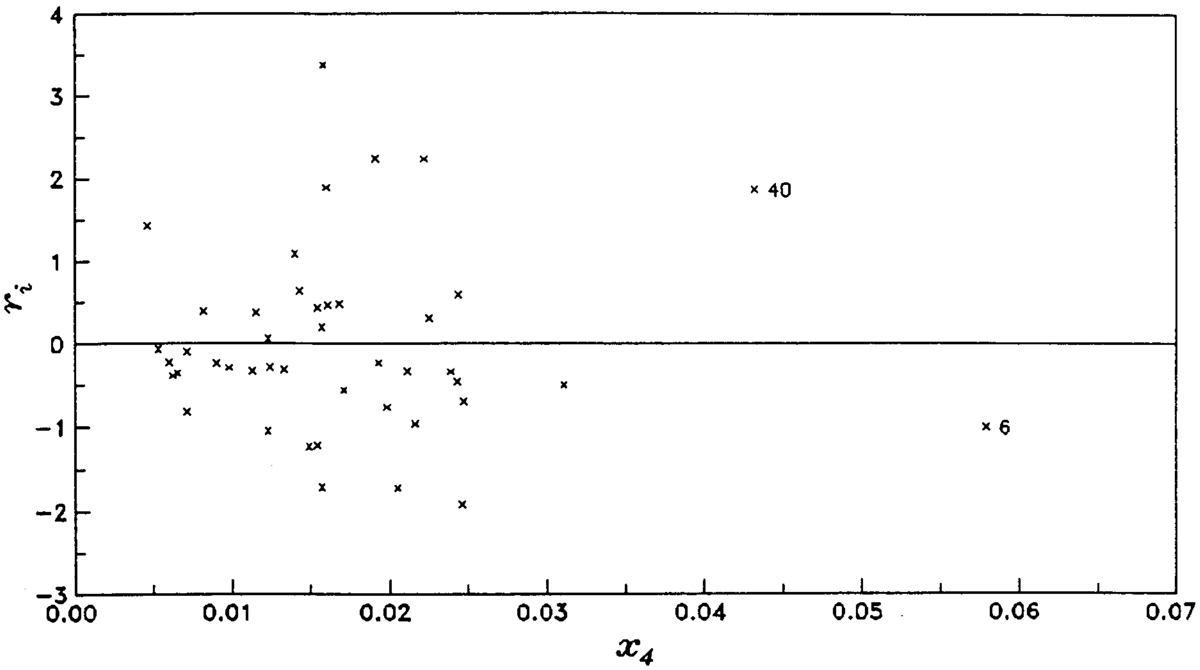

Figure C.2 Plot of standardized residuals, r_i, vs. proportion of GWA beneficiaries, x_4.

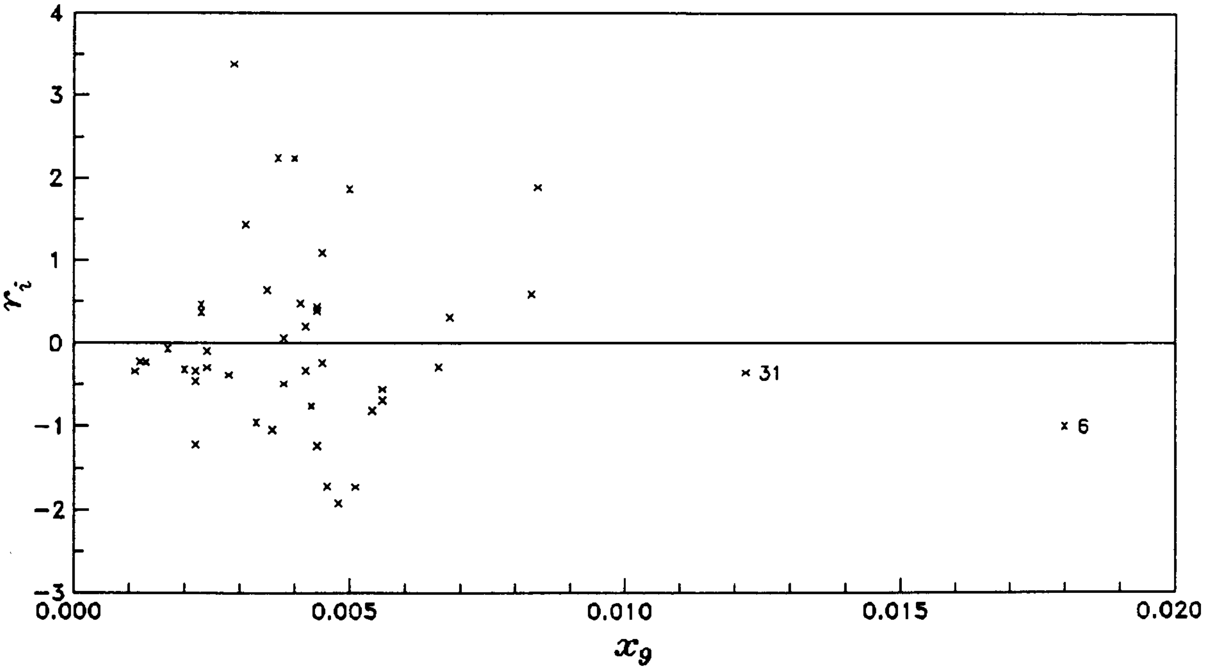

Figure C.3 Plot of standardized residuals, r_i, vs. JDA offense rate, x_9.

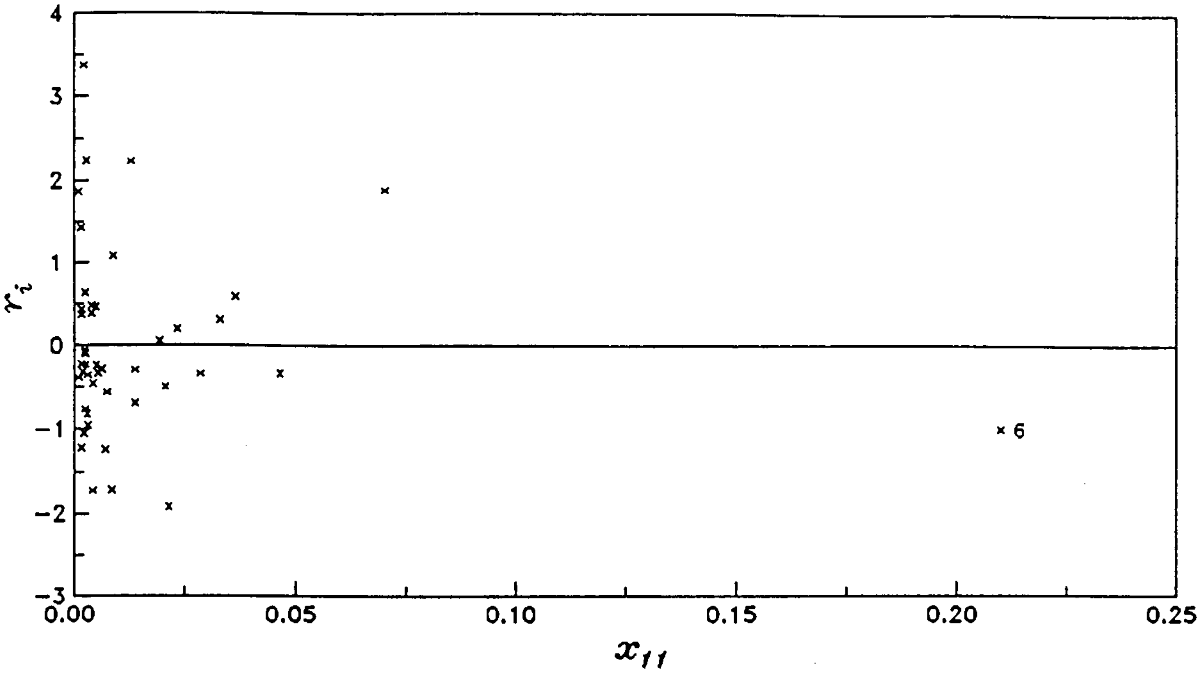

Figure C.4 Plot of standardized residuals, r_i, vs. proportion of population of (N.A.) Indian descent, x_{11}.

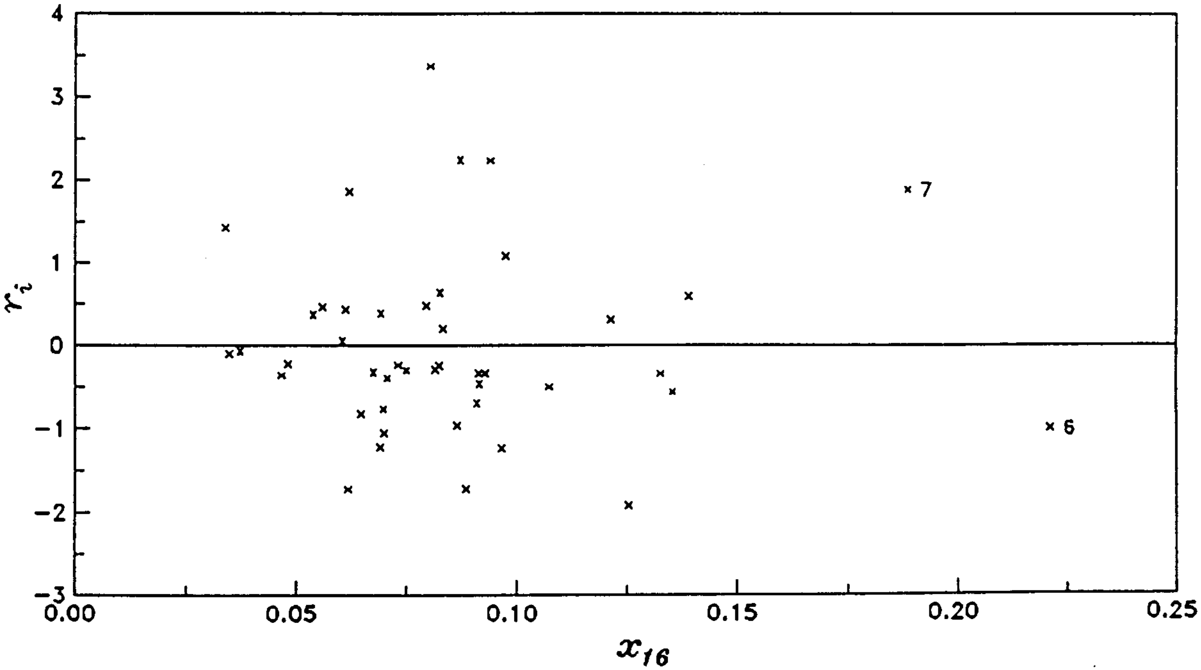

Figure C.5 Plot of standardized residuals, r_i, vs. rate of incidence of births to unmarried mothers, x_{16}.

References

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An introduction to graphical methods of diagnostic regression analysis.* Clarendon Press, Oxford.

Beckman, R. J. and Cook, R. D. (1983). Outlier. *Technometrics*, **25**, 119-149.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley and Sons, New York.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates.* John Wiley and Sons, New York.

Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, **26**, 211-246.

Carroll, R. J. and Spiegelman, C. H. (1992). Diagnostics for Nonlinearity and Heteroscedasticity in Errors-in-Variables Regression. *Technometrics*, **34**, 186-196.

Chen, G. M. (1991). *Empirical Processes Based on Regression Residuals theory and Applications. P.hd thesis.* Simon Fraser University, Burnaby.

Christensen, R. (1987). *Plane Answers to Complex Questions: The Theory of Linear Models.* Springer-Verlag, New York.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

Cook, R. D. and Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, **70**, 1-10.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*, Second Edition. John Wiley and Sons, New York.

Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*, Second Edition. John Wiley and Sons, New York.

Ellenberg, J. H. (1973). The joint distribution of the standardized least squares residuals from a general linear regression. *Journal of the American Statistical Association*, **68**, 941-943.

Goodnight, J. H. (1979). A Tutorial on the SWEEP Operator. *The American Statistician*, **33**, 149-158.

Graybill, F. A. (1976). *Theory and Application of the Linear Model.* Duxbury Press, Massachusetts.

Hocking, R. R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, **32**, 1-49.

Hocking, R. R. (1983). Developments in Linear Regression Methodology: 1959-1982. *Technometrics*, **25**, 219-244.

Hoerl, A. E. and Kennard. R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67.

Mallows, C. L. (1975). On some topics in robustness. Unpublished Bell Telephone Laboratories report, Murray Hill, N.J.

Neter, J., Wasserman, W. and Kutner, M. H. (1990). *Applied Linear Statistical Models*, Third Edition. Richard Irwin, Inc., Illinois.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1986). *Numerical Recipes: The Art of Scientific Computing.* Cambridge, New York.

Silvey, S. D. (1969). Multicollinearity and Imprecise Estimation. *Journal of the Royal Statistical Society, Series B*, **31**, 539-552.

Smith, G. and Campbell, F. (1980). A Critique of Some Regression Methods. *Journal of the American Statistical Association*, **75**, 74-102.

Tsai, C. L. and Wu, X. (1992). Transformation-Model Diagnostics. *Technometrics*, **34**, 197-202.