1-31-2013

Contributions to linear models : lack-of-fit test and linear model with singular covariance matrices

Yong Lin

Yong Lin

Candidate

Department of Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Ronald Christensen , Chairperson

Edward Bedrick

Erik Erhardt

Michael Sonkson

CONTRIBUTIONS TO LINEAR MODELS
Lack-of-fit tests
and
linear model with singular covariance matrices

by

Yong Lin

B.A. in Statistics, Henan University of Finance and Law
2007

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico
Albuquerque, New Mexico

December, 2012

Dedication

To my Parents, Heng, Yan, and Jonathan.

Acknowledgments

I would like to thank my dissertation advisor Dr. Ronald Christensen for his patience, encouragement, advice and guidance in my academic research. I have learned so much from him and enjoyed working with him as well as his lectures and jokes.

I would also like to thank Dr. Edward Bedrick, Dr. Erik Erhardt, Dr. Michael Sonksen for being my dissertation committee and giving me valuable suggestion, Dr. Yan Lu for her classes and valuable advice, which I believe will benefit me all my life, Dr. Gabriel Huerta for many wonderful classes and the opportunity to work in the statistics clinic, Dr. Erik Erhardt for his wonderful class, being a good role model and his help on my employment opportunities, Dr. Huining Kang for his detailed help and guidance, as well as the opportunity to work with him as an R.A and Post-doctoral fellow, Dr. Michele Guindani for his advice and support, Dr. Margo Collier for being a good collaborator as well as her support as an R.A.

Also I would like to thank my parents and my brother who have been so supportive of my Ph.D education, my wife for being there all the time no matter what, my son Jonathan who brings so much joy to my life, and also my uncle who gave me endless help and advice during my time in graduate school. They are unreplaceable treasures in my life.

Finally, I want to thank my friends who make my life colorful and enjoyable, and the Mathematics and Statistics Department who gave me the opportunity of pursuing a Ph.D education.

Contributions to linear models

Lack-of-fit tests and linear model with singular covariance matrices

by

Yong Lin

B.S. in Statistics, Henan University of Finance and Law, 2007

Ph.D, Statistics, University of New Mexico, 2012

Abstract

Linear models are statistical models that are linear in their parameters. This class of models include traditional regression, ANOVA, ACOVA, mixed models and even many time series models. They can be extended into generalized linear models in which case the parameters are still linear, but they are not linearly associated with the dependent variables. This dissertation contributes in two directions.

First, it proposes and studies new lack-of-fit tests. Su and Wei (1991) proposed a lack-of-fit test based on partial sums of residuals. They computed P values using an unusual bootstrapping simulation. However, the simulation can not be performed for even moderate numbers of predictor variables because it is prohibitively time consuming. I examine the nature of their bootstrap simulation and argue that it reduces the power of Su and Wei's test. I modify their test for linear models and propose two lack-of-fit tests based on partial sums of residuals. I find the non-normal limiting distributions for both tests and small sample corrections that enable more precise calculation of 0.05 cut-offs. Empirical sizes and powers are studied for both tests in small samples.

In the second contribution, I studied the linear model with singular covariance matrix. In these models, frequently there exists estimable functions of $\boldsymbol{\beta}$ that are known with probability 1. Traditional methods of analysis employ a psuedo-covariance matrix that gives BLUEs and tests that are appropriate for the actual covariance matrix $\boldsymbol{V}$. Contrary to traditional methods of adjusting $\boldsymbol{V}$, I decompose $\boldsymbol{\beta}$ into known and unknown parts and adjust $\boldsymbol{X}$ to allow estimation and testing of the unknown part of $\boldsymbol{\beta}$. Specifically, I adjust this model, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, to get an equivalent model, $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 = \boldsymbol{X}_v\gamma + \boldsymbol{e}$, where $\boldsymbol{X}\boldsymbol{\beta}_0$ is a known vector, then perform estimation and tests on this equivalent model. The equivalence of the models is studied.

KEY WORD: Linear model; Lack-of-fit test; Partial sums of residuals; Asymptotic distribution; Singular covariance matrix.

Contents

List of Figures

Chapter 1

Introduction

1.1 Notation

A standard linear model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \boldsymbol{0}, \qquad Cov(\boldsymbol{e}) = \sigma^2\boldsymbol{I}, \tag{1.1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of observable random values, $\boldsymbol{X}$ is an $n \times p$ known model matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{e}$ is an $n \times 1$ vector of independent, unobservable errors.

Bold face math characters stand for matrices and vectors. For any matrix $\boldsymbol{A}$, $C(\boldsymbol{A})$ denotes the column space of $\boldsymbol{A}$ and $r(\boldsymbol{A})$ is the rank of $\boldsymbol{A}$; $\boldsymbol{A}^-$ and $\boldsymbol{A}^+$ denote a generalized inverse and the Moore-Penrose pseudoinverse of $\boldsymbol{A}$ respectively. $\boldsymbol{M}_{\boldsymbol{A}}$ is the perpendicular projection operator onto the column space $C(\boldsymbol{A})$. SSE and MSE denote the sum of squared error and the mean squared errors of a particular model, respectively. $\boldsymbol{I}$ is an $n \times n$ identity matrix. $\boldsymbol{J}_k$ is vector with k 1s and $\boldsymbol{J}_k^k$ is a $k \times k$ matrix of 1s.

1.2 Overview of the problems

Linear models are models that are linear in their parameters, such as regression, ANOVA, ACOVA, Mixed models, even many time series models. I am interested in two particular problems in linear model theory. They are lack-of-fit testing and linear models with singular covariance matrices. This section briefly introduces the background of the problems and the contributions of this dissertation.

1.2.1 Lack-of-fit tests

Lack-of-fit tests, also known as goodness-of-fit tests, are techniques to check whether the proposed models have valid mean structures. The classical approaches extend model (1.1) to a model

$$\boldsymbol{Y} = \tilde{\boldsymbol{X}}\boldsymbol{\gamma} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \boldsymbol{0}, \qquad Cov(\boldsymbol{e}) = \sigma^2\boldsymbol{I}, \tag{1.2}$$

such that $C(\boldsymbol{X}) \subset C(\tilde{\boldsymbol{X}})$. Then a test statistic (usually a F test) is constructed to compare the differences between model (1.1) and the extended model (1.2). Three classical approaches to this problem are clustering, partitioning and smooth tests. Clustering lack-of-fit tests extend $C(\boldsymbol{X})$ to $C(\tilde{\boldsymbol{X}})$ by clustering data into groups in which covariates are exact replications or near replications, partitioning method divides the data into subsets and fit models like (1.1) to each subset, whereas smooth tests expands $C(\boldsymbol{X})$ to $C(\tilde{\boldsymbol{X}})$ by incorporating smooth functions of the predictors. These method are discussed in more detail in the Chapter 2.

Su and Wei (1991), on the other hand, proposed an alternative approach. They revisited the idea of examining residuals. In linear models, their test statistic considers a process of summing residuals over successively larger covariate space. This dissertation proposes and examines two lack-of-fit tests that are based on Su and Wei's test.

1.2.2 Linear models with singular covariance matrices

The general Gauss-Markov model is

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \boldsymbol{0}, \quad Cov(\boldsymbol{e}) = \sigma^2 \boldsymbol{V}. \tag{1.3}$$

Here, no assumptions are made on $r(\boldsymbol{X})$. If $\boldsymbol{V}$ is singular and $C(\boldsymbol{X}) \subset C(\boldsymbol{V})$, estimation and tests can be obtained using the same formulas as if $\boldsymbol{V}$ is nonsingular. This dissertation focuses on the most difficult cases wherein $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$ which forces $\boldsymbol{V}$ to be singular. Under this condition, there exist nontrivial linear functions of $\boldsymbol{Q'X\beta}$ that are known with probability 1 (perfectly) where $C(\boldsymbol{Q}) = C(\boldsymbol{V})^{\perp}$.

To treat $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, traditional methods obtain estimates and tests by replacing model (1.3) with a model involving a pseudo covariance matrix

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \boldsymbol{0}, \quad Cov(\boldsymbol{e}) = \sigma^2 \boldsymbol{T}, \tag{1.4}$$

where $\boldsymbol{T} = \boldsymbol{V} + \boldsymbol{XUX'}$ for some nonnegative definite $\boldsymbol{U}$ such that $C(\boldsymbol{X}) \subset C(\boldsymbol{T})$.

Different from the traditional methods, this dissertation proposes a more intuitive approach by decomposing $\boldsymbol{\beta}$ into the sum of two orthogonal parts, $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_0$ is known. The unknown component of $\boldsymbol{X\beta}$ is shown to be $\boldsymbol{X\beta}_1 \equiv \boldsymbol{X}_v\boldsymbol{\gamma}$, where $C(\boldsymbol{X}_v) = C(\boldsymbol{X}) \cap C(\boldsymbol{V})$. Replace model (1.3) with

$$\boldsymbol{Y} - \boldsymbol{X\beta}_0 = \boldsymbol{X}_v\boldsymbol{\gamma} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = \boldsymbol{0}, \quad \textbf{Cov}(\boldsymbol{e}) = \boldsymbol{\sigma^2 V}, \tag{1.5}$$

for which $C(\boldsymbol{X}_v) \subset C(\boldsymbol{V})$ and $\boldsymbol{X\beta}_0$ is known with probability 1. Then estimation and tests are obtained under model (1.5) for which the simplifying assumption $C(\boldsymbol{X}_v) \subset C(\boldsymbol{V})$ holds. This dissertation shows that this alternative method also provides the usual estimates and tests.

1.3 Outline of the dissertation

Chapter 2, presents a review of classical approaches to lack-of-fit tests. Su and Wei (1991)'s test is examined in detail.

In Chapter 3, the two new lack-of-fit tests are proposed. Their large sample distributions and small sample adjustments are studied. Also, simulation studies of the sizes and powers are provided.

In Chapter 4, methods for dealing with linear models having singular covariance matrices are reviewed and our new method to estimation and testing in this condition is developed.

Chapter 5 outlines anticipated future work.

Chapter 2

Review of lack-of-fit techniques

In this dissertation, I am interested in the lack-of-fit test proposed by Su and Wei (1991) that sums residuals over successively larger covariate space. This approach is different from the classical approaches that involve extending null model (1.1) to a larger model (1.2).

In this Chapter, Su and Wei's test is reviewed in detail. For the completeness of this dissertation, I begin with the reviews of two initial works of clustering and smooth tests.

2.1 Classical approaches to lack-of-fit testing

Sun (2010) reviewed classical lack-of-fit tests. They are Fisher (1922)'s test, Neyman (1937)'s test, Green(1971)'s test, Shillington(1979)'s test, Neill and Johnson(1985)'s test, Christensen (1989)'s test, Joglekar, Schuenemeyer and LaRiccia (1989)'s test, Christensen (1991)'s test, Eubank and Hart(1992)'s test, Aerts Claeskens and Hart (2000)'s test, Fan and Huang (2001)'s test and Su and Yang (2006)'s test. I present a brief review of Fisher's exact test and Neyman's smooth test. Reviews of Utts (1982) and Eubank and Spiegelman (1990) are in the Appendix C to complement Sun's work.

2.1.1 Fisher's exact replicates test

Fisher (1922) proposed a lack-of-fit test for simple linear regression in which the covariate has exact replicates. Considering n pairs of observations x and y, we suppose that there are k distinct values in x, i.e., k clusters of x's, and the number of observations for which $x = x_i$ is n_i for $i = 1, ..., k$. Obviously, $n = \sum_{i=1}^{k} n_i$. A simple regression model for these data can be written as

$$y_{ij} = [1x_i] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + e_{ij}, \tag{2.1}$$

where y_{ij} and e_{ij} are the response and error associated with jth replicate in ith cluster. In each cluster, write

$$\boldsymbol{Y}_i = [y_{i1}, ..., y_{in_i}]', \quad \boldsymbol{X}_i = [\boldsymbol{J}_{n_i}, x_i \boldsymbol{J}_{n_i}], \quad \text{and} \quad \boldsymbol{e}_i = [e_{i1}, ..., e_{in_1}]'.$$

Model (2.1) can be formulated as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.2}$$

with

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_k \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_k \end{bmatrix}, \quad \text{and} \quad \boldsymbol{e} = \begin{bmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_k \end{bmatrix}.$$

To check the adequacy of model (2.2), Fisher considered obtaining a more accurate fit and comparing the result with the fit from model (2.2). To begin with, he assumed the mean for each cluster is known and fitted model like (2.2) to each cluster,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\gamma}_i + \boldsymbol{e}_i, \tag{2.3}$$

where $\boldsymbol{\gamma}_i = [\gamma_0, \gamma_1]'$. Since $C(\boldsymbol{X}_i) = C(\boldsymbol{J}_{ni})$, model (2.3) is equivalent to

$$\boldsymbol{Y}_i = \mu_i \boldsymbol{J}_{n_i} + \boldsymbol{e}_i. \tag{2.4}$$

where μ_i the group mean of ith cluster. The estimate of μ_i from least square estimate is simply $\bar{y}_i = \boldsymbol{J}'_{n_i}\boldsymbol{Y}_i/n_i$. Standardizing $\bar{y}_i$ to $z_i = \sqrt{n}(\bar{y}_i - \mu_i)/\sigma$ gives us k i.i.d. standard normal random variables and

$$\sum_{i=1}^{k} z_i^2 = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \mu_i)^2}{\sigma^2} \sim \chi^2_{(k)}.$$

Here, σ^2 can be estimated by the MSE from model (2.4). The sum of squares of lack-of-fit of the model (2.2) can be measured using $\sum_{i=1}^{k} n_i(\bar{y}_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the estimate of the kth cluster mean from model (2.2). Fisher's test statistic is

$$\chi^2 = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \hat{y}_i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-k)} \xrightarrow{\mathcal{L}} \chi^2_{(k-1)}.$$

Later on, the degree of freedom $k-1$ is further corrected with $k-2$ and Fisher's χ^2 test is transferred into a F statistic which follows an exact F distribution,

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \hat{y}_i)^2/k-2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-k)} \sim F_{(k-2,n-k)}. \tag{2.5}$$

Sun (2010) pointed out that Fisher compared the original regression model with a largest model that is equivalent to a one-way ANOVA model with k treatment. With the same row structure of the original data, he formulated Fisher's exact test in term of testing model (2.2) against

$$\boldsymbol{Y} = \tilde{\boldsymbol{X}}\boldsymbol{\gamma} + \boldsymbol{e}, \tag{2.6}$$

where

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{J}_{n_1} & 0 & \cdots & 0 & 0 \\ 0 & \boldsymbol{J}_{n_2} & \cdots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & \boldsymbol{J}_{n_{k-1}} & 0 \\ 0 & 0 & \cdots & 0 & \boldsymbol{J}_{n_k} \end{bmatrix}.$$

Clearly $C(\boldsymbol{X}) \subset C(\tilde{\boldsymbol{X}})$. The SSEs of the two models are:

$$SSE(2.2) = \boldsymbol{Y'(I - M_X)Y} \quad \text{and} \quad SSE(2.6) = \boldsymbol{Y'(I - M_{\tilde{X}})Y},$$

and the F statistic is

$$F = \frac{[SSE(2.2) - SSE(2.6)]/[r(\tilde{\boldsymbol{X}}) - r(\boldsymbol{X})]}{[SSE(2.6)]/[n - r(\tilde{\boldsymbol{X}})]} = \frac{\boldsymbol{Y'(M_{\tilde{X}} - M_X)Y}/(k-2)}{\boldsymbol{Y'(I - M_{\tilde{X}})Y}/(n-k)}.$$

After some algebra, it can be shown that

$$\boldsymbol{Y'(M_{\tilde{X}} - M_X)Y} = \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{y}_i)^2 \quad \text{and} \quad \boldsymbol{Y'(I - M_X)Y} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2.$$

Hence, Fisher's exact test is equivalent to the classical F test for testing model (2.6) against model (2.2). When exact replicates are not available, people proposed methods that partition data into near replicates or subsets within which the data has the same features. Actually, the clustering method is a special case of partitioning where each subset is a cluster. Green (1971); Utts (1982); Miller, Neill and Sherfey (1988); Joglekar, Schuenemeyer and LaRiccia (1989); Christensen (1989, 1991) and Su and Yang (2006) all suggested different methods of partitioning data.

2.1.2 Neyman's smooth test

Neyman (1937), proposed a lack-of-fit test for a completely specified distribution with the attractive feature of focusing power towards smooth alternatives. Although his original test is less directly related to linear models, it is closely related to my future works. For completeness, it is necessary to start from the original test. The original test considers testing a probability density function (PDF) $f(x, \boldsymbol{\beta})$, where $\boldsymbol{\beta} = [\beta_1, ..., \beta_q]'$ is a $q \times 1$ vector of parameters. The alternative PDF is defined as

$$f^*(x, \boldsymbol{\beta}, \boldsymbol{\theta}) = C(\boldsymbol{\theta}, \boldsymbol{\beta}) \exp\left\{\sum_{j=1}^{k} \theta_j h_j(x; \boldsymbol{\beta})\right\} f(x; \boldsymbol{\beta}), \tag{2.7}$$

where $\boldsymbol{\theta} = [\theta_1, ..., \theta_k]'$ is a vector of k parameters, $C(\boldsymbol{\theta}, \boldsymbol{\beta})$ is a constant that normalizes $f^*(x, \boldsymbol{\beta}, \boldsymbol{\theta})$ to a PDF and the $h_j(x; \boldsymbol{\beta})$s are a set of Legendre polynomials functions. The parameter k is a small integer and these smooth functions $h_j(x; \boldsymbol{\beta})$s are chosen subjectively to maximize the power of testing a certain class of alternative hypothesis. The lack-of-fit test is equivalent to testing

$$H_0 : \boldsymbol{\theta} = \mathbf{0}, \quad \text{against} \quad H_a : \boldsymbol{\theta} \neq \mathbf{0}.$$

Let $l(\hat{\boldsymbol{\theta}})$ and $l(\mathbf{0})$ be the log-likelihoods under the MLE and null hypothesis respectively. The above hypothesis can be tested using the likelihood ratio test,

$$2(l(\hat{\boldsymbol{\theta}}) - l(\mathbf{0})) \xrightarrow{\mathcal{L}} \chi^2_{(k)}.$$

Alternatively, Neyman recommended to estimate the sample mean of $h_j(x_i, \boldsymbol{\beta})$s by

$$\bar{h}_j = \frac{1}{n} \sum_{i=1}^{n} h_j(x_i, \hat{\boldsymbol{\beta}})$$

and test the above hypothesis using

$$\Phi^2 = n \sum_{j=1}^{k} \bar{h}_j^2 \xrightarrow{\mathcal{L}} \chi^2_{(k)}.$$

Although the original Neyman (1937)'s test is not designed for linear regressions, its idea is well applicable. Sun (2010) considered a simple linear regression in the form of model (1.1) with $\boldsymbol{X} = [\boldsymbol{J}_n, \boldsymbol{x}]$, where $\boldsymbol{x} = [x_1, ..., x_n]'$. The more general alternative distribution is obtained by extending model (1.1) to model (1.2) with

$$\tilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{H}_k] \quad \text{and} \quad \boldsymbol{\gamma} = [\boldsymbol{\beta}, \boldsymbol{\delta}]',$$

where $\boldsymbol{H}_k$ is an $n \times k$ matrix of smooth functions, $\boldsymbol{\beta} = [\beta_0, \beta_1]'$, $\boldsymbol{\delta}$ is a vector of k unknown parameters and

$$\boldsymbol{H}_k = \begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \cdots & \varphi_k(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \cdots & \varphi_k(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_1(x_n) & \varphi_2(x_n) & \cdots & \varphi_k(x_n) \end{bmatrix}.$$

Here $\varphi_i(x)$s are known and fixed functions from $\mathbf{R} \to \mathbf{R}$. Theoretically, the alternative density function approaches the true density function when $k = \infty$. However, in application k is chosen from the set $\{1, \ldots, n-2\}$. To emphasize the smooth functions in (1.2), rewrite it as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{H}_k\boldsymbol{\gamma}_k + \boldsymbol{e}. \tag{2.8}$$

The lack-of-fit test can be performed using the usually F test as in Fisher's exact test. Let $r(\boldsymbol{X}, \boldsymbol{H}_k) = r_1$ and $r(\boldsymbol{X}) = r_2$.

$$F = \frac{[SSE(1.1) - SSE(2.8)]/(r_1 - r_2)}{SSE(2.8)/(n - r_1)} = \frac{\boldsymbol{Y}'(\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{H}_k} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{Y}/(r_1 - r_2)}{\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X},\boldsymbol{H}_k})\boldsymbol{Y}/(n - r_1)},$$

which follows $F(r_1 - r_2, n - r_1)$ under H_0. The large values of F indicates the inadequacy of model (1.1). Sun (2010) discussed the case of simple linear regression, but this method can be easily extended into linear models with multiple covariates. The key point is how to choose the amount k and type of the smooth functions $\varphi_i(\cdot)$s. Apparently these choices have crucial effects on the performance of the test. For example, Fourier series are powerful in detecting the lack-of-fit of periodic terms, while Wavelets are good at picking up the local lack-of-fit in the proposed models. Originally, Neyman suggested k to be a predetermined integer. Here, with a fixed k, the test is consistent against some but not all alternatives, so later on, people become more interested in data driven versions of k. This data driven k is used to control the over fitting or smoothness of the extended model. Alternatively, people maximize k and introduce additional smooth parameter to control the smoothness of the extended model. Eubank and Spiegelman (1990); Eubank and Hart (1992); Aerts, Claeskens and Hart (2000); Simonoff and Tsai (1999); Fan and Huang (2001) and Christensen (2010) all proposed different approaches to extend Neyman's test to linear regression, but they all involve comparing model (1.1) with model (2.8).

2.2 Lack-of-fit techniques using partial sum of residuals

Su and Wei (1991) proposed a lack-of-fit test based on measuring the difference between the partial sum process of observed residuals and that of null model. Large difference between the two processes provides evidence to reject H_0. In application, the P value of their test is calculated by comparing the empirical process and bootstrap approximated null process. Stute, Manteiga and Quindimil (1998); Lin, Wei and Ying (2002); Hosmer and Hjort (2002); Stute, Thies and Zhu (1998); Diebolt and Zuber (1999); Koul and Stute (1999); Koul, Baillie, and Surgailis (2004) all adopted similar approaches to different problems.

Besides S-W's test there are other tests that use partial sums, for example, Fan and Huang (2001)'s test and Christensen and Sun (2010)'s test but they are based on a different idea going back to Cramér-Von Mises's (CVM's) lack-of-fit test of density functions. Fan (1996) examined the inefficiency of CVM's test and proposed an alternative approach based on power consideration. His idea involves transforming the problem of testing density functions to testing the mean vector of a high dimension multivariate normal distribution. To shrink its dimension, he proposed a partial-sum type test. Later on Fan and Huang (2001) extended Fan (1996)'s method to regression models by testing $E(\hat{\boldsymbol{e}}) = \mathbf{0}$, which is also the mean vector of a multivariate normal distribution. Christensen and Sun (2010) examined F-H's test under the linear models and recast their approach to classical smooth tests.

In this Section, S-W's test and bootstrap simulation method are examined under the context of linear models. A review of Lin, Wei and Ying (2002) is presented in the Appendix C. C-S's test and F-H's are related to my future research but less related to my proposed test, so these reviews are postponed to the Appendix C. The remaining literature related to S-W's method either study the mathematical aspect of S-W's partial sum process or applying S-W's test to more complicate problems. Thus these reviews are not shown.

2.2.1 Su and Wei's test

Su and Wei (1991) proposed a lack-of-fit test for generalized linear models based on summing residuals over successively larger subsets of the covariate space. Specifically, for linear model (1.1) they consider the random process in p dimensions,

$$W_n(\boldsymbol{t}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_n] I(\boldsymbol{x}_i \leq \boldsymbol{t}),$$

where y_i and $\boldsymbol{x}_i'$ are the ith rows of $\boldsymbol{Y}_n$ and $\boldsymbol{X}_n$, respectively, $\hat{\boldsymbol{\beta}}_n$ is the least square estimate (LSE) of $\boldsymbol{\beta}$, $\boldsymbol{t} = (t_1, ..., t_p)' \in \mathbf{R}^p$, I is the indicator function, the inequality $\boldsymbol{x}_i \leq \boldsymbol{t}$ stands for $x_{ij} \leq t_j$ for all $j = 1, ..., p$. S-W proposed the lack-of-fit test statistic

$$G_n = \sup_{\boldsymbol{t} \in \mathbf{R}^p} |W_n(\boldsymbol{t})| .$$

Under H_0, the process $W_n(\boldsymbol{t})$ is expected to fluctuate around 0. Getting a large value of G_n provides evidence to reject model (1.1).

$W_n(\boldsymbol{t})$ is a jump process, so the suprema occurs at one of the finite number of jumps. Moreover, the indicator function in $W_n(\boldsymbol{t})$ creates a partial ordering of the residuals. G_n is the maximum value of $W_n(\boldsymbol{t})$ taken over all the full orderings that agree with the partial ordering. To calculate the statistic G_n, S-W replace $\sup_{\boldsymbol{t} \in \mathbf{R}^d} |W_n(\boldsymbol{t})|$ by $\max_{\boldsymbol{t} \in S} |W_n(\boldsymbol{t})|$ where S is the product of the sets S_k for $k = 1, \ldots, p$, where S_k consists of all elements in the kth column of $\boldsymbol{X}_n$. S contains as many as n^p vectors. To evaluate the significance of G_n, S-W proposed a bootstrapping method to simulate the null distribution of G_n, and then obtain the P values. To better understand their methods, write

$$W_n(\boldsymbol{t}) = \frac{1}{\sqrt{n}} \boldsymbol{J}' \boldsymbol{I}(\boldsymbol{t}) (\boldsymbol{Y}_n - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}_n) = \frac{1}{\sqrt{n}} \boldsymbol{J}' \boldsymbol{I}(\boldsymbol{t}) \hat{e}_n,$$

where $\boldsymbol{I}(\boldsymbol{t})$ is a diagonal matrix with ith diagonal element equal to $I(\boldsymbol{x}_i \leq \boldsymbol{t})$. To approximate the process $W(\boldsymbol{t})$, S-W bootstrap a response vector $\boldsymbol{Y}_n^s$ as

$$\boldsymbol{Y}_n^s = \boldsymbol{X}_n\hat{\boldsymbol{\beta}}_n + \boldsymbol{D}(\boldsymbol{Z}^s)\hat{\boldsymbol{e}}_n,$$

where $\boldsymbol{D}(\boldsymbol{Z}^s)$ is a diagonal matrix whose elements are a random sample of standard normals.

The approximated process replaces the original data $\boldsymbol{Y}_n$ with bootstrapped data $\boldsymbol{Y}_n^s$, and $\hat{\boldsymbol{\beta}}_n$ with $\hat{\boldsymbol{\beta}}_n^s$, the LSE of the parameter based on $\boldsymbol{Y}_n^s$ and $\boldsymbol{X}_n$. Specifically,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n^s &= (\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n'\boldsymbol{Y}_n^s \\ &= \hat{\boldsymbol{\beta}}_n + (\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n\boldsymbol{D}(\boldsymbol{Z}^s)(\hat{\boldsymbol{e}}_n).\end{aligned}$$

The approximated process $W_n^s(\boldsymbol{t})$ can be expressed as:

$$\begin{aligned}W_n^s(\boldsymbol{t}) &= \frac{1}{\sqrt{n}}\boldsymbol{J}'\boldsymbol{I}(\boldsymbol{t})(\boldsymbol{Y}_n^s - \boldsymbol{X}_n\hat{\boldsymbol{\beta}}_n^s) \\ &= \frac{1}{\sqrt{n}}\boldsymbol{J}'\boldsymbol{I}(\boldsymbol{t})\left[\boldsymbol{D}(\boldsymbol{Z}^s)\hat{\boldsymbol{e}}_n - \boldsymbol{M}_{\boldsymbol{X}_n}\boldsymbol{D}(\boldsymbol{Z}^s)\hat{\boldsymbol{e}}_n\right] \\ &= \frac{1}{\sqrt{n}}\boldsymbol{J}'\boldsymbol{I}(\boldsymbol{t})(I - \boldsymbol{M}_{\boldsymbol{X}_n})\boldsymbol{D}(\boldsymbol{Z}^s)\hat{\boldsymbol{e}}_n.\end{aligned}$$

After some algebra, we get the same result as in their paper,

$$W_n^s(\boldsymbol{t}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i^s(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n)\left\{I(\boldsymbol{x}_i \leq t) - \left\{\sum_{i=1}^{n}\boldsymbol{x}_i' I(\boldsymbol{x}_i \leq t)\right\}(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i')^{-1}\boldsymbol{x}_i\right\}. \tag{2.9}$$

Note that,

$$W_n(\boldsymbol{t}) = \frac{1}{\sqrt{n}}\boldsymbol{J}'\boldsymbol{I}(\boldsymbol{t})(I - \boldsymbol{M}_{\boldsymbol{X}_n})\hat{\boldsymbol{e}}_n.$$

Here $W_n^s(\boldsymbol{t})$ approximates $W_n(\boldsymbol{t})$ by introducing the multiplicative random white noise $\boldsymbol{D}(\boldsymbol{Z}^s)$. Stute (1997) shows $W_n(\boldsymbol{t}) \xrightarrow{\mathcal{L}} \sigma W(\boldsymbol{t})$, where $W(\boldsymbol{t})$ is a normal process with mean 0. The bootstrap relies on having $W_n^s(\boldsymbol{t}) \xrightarrow{\mathcal{L}} \sigma W(\boldsymbol{t})$. Hence, one can infer $G_n \xrightarrow{\mathcal{L}} \sigma G$

and $G_n^s \xrightarrow{\mathcal{L}} G$, where $G_n^s = \sup_{\boldsymbol{t}\in\mathbf{R}^d} |W_n^s(\boldsymbol{t})|$. Finally, simulates from G_n^s approximate simulations from σG. The P value of the test statistic G_n is computed by evaluating the percentage of G_n^s values exceeding G_n.

It is useful to examine how these ideas work in a simplified special case. Consider $V_1, V_2, \ldots, V_n$, independent identical distributed (iid) with $E(V_i) = \mu$, $Var(V_i) = \sigma^2$ and $Z_i, Z_2, \ldots, Z_n$ iid with $E(Z_i) = 0$, $Var(Z_i) = 1$. In the null case of $\mu = 0$,

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} V_i \xrightarrow{\mathcal{L}} N(0,1).$$

Multiplying by the white noise variable, $\sum_{i=1}^{n} Z_i V_i / \sigma\sqrt{n} \xrightarrow{\mathcal{L}} N(0,1)$. This follows because the Z_iV_is are iid with $E(Z_iV_i) = 0$, $Var(Z_iV_i) = \sigma^2$. However, in the non-null case,

$$\frac{1}{\sigma\sqrt{n}} E\left[\sum_{i=2}^{n} V_i\right] \to \infty, \quad Var\left[\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} V_i\right] = 1,$$

whereas

$$\frac{1}{\sigma\sqrt{n}} E\left[\sum_{i=1}^{n} Z_i V_i\right] = 0, \quad Var\left[\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} Z_i V_i\right] = 1 + \mu^2/\sigma^2.$$

Not only does the bootstrap contain more noise than using the asymptotic distribution because it approximates σG by G_n^s under the null model but under the alternative it simulates a process with more variability than σG, thus reducing its power. More over, S-W's simulation can be prohibitively time consuming even for moderate sized regression problems. Recall S-W's partial ordering, for data with 46 observations, $p = 3$, and 500 simulation samples, my R program for the test took as much as 2 hours to perform.

Numerous authors have examined aspects of the partial sum process $W_n(\boldsymbol{t})$. Stute (1997) carefully studied the asymptotic convergence of $W_n(\boldsymbol{t})$ to a tied down Gaussian process. (With an intercept in the model, the sum of all residuals is zero.) However, these results provide no relief from the computational burden. Stute, Manteiga and Quindimil (1998) approximated the process using a bootstrap simulation method different from S-W's but with results similar to S-W's bootstrap method. Lin, Wei and Ying (2002) studied the process but used a total ordering determined by one variable. They extended the S-W test by considering different methods of partitioning and other ways of aggregating residuals such as moving averages. Hosmer and Hjort (2002) proposed a related test for logistic regression using a weighted partial sum of residuals over partitions of the estimated logits. For other related tests, see, Stute, Thies and Zhu, (1998); Diebolt and Zuber (1999); Koul and Stute (1999); Koul, Baillie and Surgailis (2004). One similarity is that these papers all approximate the partial sum process by a Gaussian process and typically report P values using simulations.

Chapter 3

Lack-of-fit test using partial sum of residuals

In this Chapter, I propose two new lack-of-fit tests based on modifications of S-W's test and study their asymptotic distributions. Small sample adjustments to the proposed tests are in Section 3.2. Simulation studies of the power and sizes of the tests are presented in Section 3.3 and Section 3.4.

In this Chapter and the related proof in the Appendix A, which emphasizes the asymptotic theory, I add the subscript n to all key matrices in model (1.1). For example, $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{e}$, $\boldsymbol{I}$, $\boldsymbol{\beta}$ are replaced with $\boldsymbol{X}_n$, $\boldsymbol{Y}_n$, $\boldsymbol{e}_n$, $\boldsymbol{I}_n$ and $\boldsymbol{\beta}_n$, respectively.

3.1 My proposed tests

S-W's simulation technique is based on $W_n(\boldsymbol{t})$ converging to a limiting process that clearly depends on the variance σ^2 and I show in Section 2.2.1 that this approximation, mixing σ^2

into $W_n(\boldsymbol{t})$, lowers the power of the test. Instead I estimate σ^2 separately, standardize G_n and find an asymptotic distribution directly.

To obtain asymptotic results, the range of the maximum needs to converge to infinity more slowly than the sample size, so I maximize the partial sums between 1 and $\tilde{n}$ where $\tilde{n} \equiv \tilde{n}(n) \leq n$ goes to infinity as n goes to infinity. This is a pretty standard device, see Fan and Huang (2001) and Christensen and Sun (2010). I also standardize G_n using $\hat{\sigma}_n$, a consistent estimate of σ with $\hat{\sigma}_n^2/\sigma^2 - 1 = O_p(n^{-1/2})$. Under these conditions, S-W's test statistic reduces to

$$T_n = \frac{1}{\sqrt{\tilde{n}}} \max_{1\leq m\leq \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} \right|. \tag{3.1}$$

In addition, I propose an alternative test statistic that puts more weight on terms lower in the ordering,

$$P_n = \max_{1\leq m\leq \tilde{n}} \frac{1}{\sqrt{m}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} \right|. \tag{3.2}$$

The difference between T_n and P_n is that T_n averages the partial sum of residuals using $\tilde{n}$ while P_n uses the number of residuals in the partial sum. The factor $1/\sqrt{m}$ in P_n puts more weight on the terms with lower orderings, so when the lack of fit occurs in the first few residuals of the partial sums, T_n is expected to be less sensitive than P_n. The asymptotic properties of the proposed tests are studied in the following two subsections.

Note that, y_is and $\boldsymbol{x}_i$s in the proposed statistics are after ordering. My method of imposing total ordering on the $\boldsymbol{x}_i$s and maximize partial sums of the residuals are postponed to Section 3.3.

3.1.1 First test

The test statistic T_n presented in (3.1) uses the residuals. Replacing them with the independent homoscedastic errors, and replacing $\hat{\sigma}_n$ with σ, it is relatively easy to find the asymptotic distribution of

$$\tilde{T}_n \equiv \frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{e_i}{\sigma} \right|,$$

for any choice of $\tilde{n} \le n$ that goes to infinity. The main problem is that the residuals $\hat{e}_i = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_n$ used in T_n are dependent. A lesser problem is that σ is unknown. To obtain convergence, the residual dependence needs to be mild and the estimate of σ needs to be consistent. Specifically, I assume:

Conditions

(a) $\frac{1}{n}\boldsymbol{X}_n'\boldsymbol{X}_n$ converges in probability to $\boldsymbol{A}$, where $\boldsymbol{A}$ is some positive define matrix.

(b) $\hat{\sigma}_n = \sigma + O_p(1/\sqrt{n})$.

Condition (a) holds if the $\boldsymbol{x}_i s$ are generated at random from a distribution with finite variances. If $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $Cov(\boldsymbol{x}_i) = \boldsymbol{\Sigma}$, then $E(\boldsymbol{x}_i \boldsymbol{x}_i') = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'$ and $\frac{1}{n}\boldsymbol{X_n'X_n} \xrightarrow{p} \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'$. Condition (b) holds if $\hat{\sigma}_n$ is a $\sqrt{n}$ consistent estimate of σ. Theoretically, any $\hat{\sigma}_n$ that converges in probability to σ will give the asymptotic distribution, but to get a faster convergence rate for the test statistic, $\hat{\sigma}_n$ needs to be root n convergence.

Condition (a) ensures that $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$ but the convergence needs to occur at a rate faster than the increase of the range of the maximum in T_n, so as to relieve the dependence problem. In the proof, I show it suffices to take $\tilde{n} = \lceil n / \log\log n^{1+\delta} \rceil$, for $\delta > 0$. The value $\tilde{n}$ restricts the number of residuals in the partial sum process. In practice, this restriction does not have much effect on the test statistic which is often dominated by the first few residuals of the partial sum. In simulations, I found that $\delta = 2$ has relatively slower convergence to the limiting distribution but improves the power for the test as compared to $\delta = 3$, the value used by F-H.

There is no obvious way of writing $T_n = u_n \tilde{T}_n + R_n$, where $u_n \xrightarrow{p} 1$ and $R_n \xrightarrow{p} 0$, which would be a simple way to show that T_n and $\tilde{T}_n$ have the same asymptotic distribution. Instead, I bound T_n by:

$$\tilde{T}_n - S_n(u) \le \frac{\hat{\sigma}_n}{\sigma} T_n \le \tilde{T}_n + S_n(v),$$

where $S_n(u) \equiv \left| \sum_{i=1}^{u} \boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right| / \sigma\sqrt{\tilde{n}}$, u and v are two numbers no greater than $\tilde{n}$. Under conditions (a) and (b), I show that both $S_n(u)$ and $S_n(v)$ converge in probability to 0 so that T_n has the same limiting distribution as $\tilde{T}_n$. I obtain

Theorem 1 *If conditions (a) and (b) in Section 3.1 are satisfied,* $T_n \xrightarrow{\mathcal{L}} T$, *where*

$$Pr[T < t] = \frac{4}{\pi} \sum_{m=0}^{\infty} \frac{(-1)^m}{2m+1} \exp(-(2m+1)^2 \pi^2 / 8t^2) \quad \text{for} \quad t > 0.$$

A detailed proof is in the Appendix A. In applications, it suffices to approximate the asymptotic distribution by summing m from 0 to 10.

3.1.2 Second test

To put more weight on terms lower in the ordering I replace $\tilde{n}$ in the divisor of T_n with the number of residuals in the partial sum,

$$P_n = \max_{1 \leq m \leq \tilde{n}} \frac{1}{\sqrt{m}} \left| \sum_{i=1}^{m} \frac{\hat{e}_i}{\hat{\sigma}_n} \right|.$$

I expect this test to be more sensitive to lack-of-fit at lower order terms. To obtain the limiting distribution, I normalize P_n as

$$Q_n = a_{\tilde{n}} P_n - b_{\tilde{n}},$$

where $a_{\tilde{n}} = \sqrt{2 \log \log \tilde{n}}$, and $b_{\tilde{n}} = (a_{\tilde{n}})^2 + \log a_{\tilde{n}} - \log(\sqrt{2\pi})$, with $\tilde{n} < n$ going to infinity. As before, replacing the residuals and $\hat{\sigma}_n$ with the independent homoscedastic errors and σ, it is easy to find the limiting distribution. Specifically, with $\tilde{P}_n = \max_{1 \leq m \leq \tilde{n}} \left| \sum_{i=1}^{m} e_i \right| / \sigma \sqrt{m}$, I can find the limiting distribution of

$$\tilde{Q}_n = a_{\tilde{n}} \tilde{P}_n - b_{\tilde{n}}.$$

To deal with the dependent $\hat{e}_i$s, restrict $\tilde{n}$ to $\lceil n/(\log \log n)^{2+\delta} \rceil$ for $\delta > 0$. In extensive simulations, several different choices of $\tilde{n}$s are compared and I found $\tilde{n} = \lceil n/(\log \log n)^3 \rceil$ most appropriate as it improves the empirical power while maintaining relatively fast convergence to the asymptotic distribution. To show that Q_n has the same asymptotic distribution as $\tilde{Q}_n$, write

$$\tilde{Q}_n - a_{\tilde{n}} \tilde{S}_n(u) \leq a_{\tilde{n}} \frac{\hat{\sigma}_n}{\sigma} P_n - b_{\tilde{n}} \leq \tilde{Q}_n + a_{\tilde{n}} \tilde{S}_n(v),$$

where now $\tilde{S}_n(u) = \left| \sum_{i=1}^{u} \boldsymbol{x}_i' (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right| / \sigma \sqrt{u}$ for any $u \leq \tilde{n}$. In the Appendix A, I show $a_{\tilde{n}} \tilde{S}_n(u) \xrightarrow{p} 0$, so $a_{\tilde{n}} \hat{\sigma}_n P_n / \sigma - b_{\tilde{n}}$ has the same limiting distribution as $\tilde{Q}_n$. I also establish that Q_n and $a_{\tilde{n}} \hat{\sigma}_n P_n / \sigma - b_{\tilde{n}}$ have the same asymptotic distribution.

Theorem 2 *If conditions (a) and (b) of Section 3.1 are satisfied, $Q_n \xrightarrow{\mathcal{L}} Q$ where $Pr[Q < t] = \exp\left[-\exp(-t)\right]$.*

The proof is in the Appendix A. The limiting distribution for Q_n is the same as for the tests given by F-H and C-S in a different problem, but requires slightly different normalizing constants.

Theorems 1 and 2 both require condition (a) which is violated in overparameterized ANOVA models. However, I am interested in

$$\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{\hat{e}_i}{\hat{\sigma}_n} \right| \quad \text{and} \quad \max_{1 \le m \le \tilde{n}} \left| \frac{1}{\sqrt{\tilde{m}}} \sum_{i=1}^{m} \frac{\hat{e}_i}{\hat{\sigma}_n} \right|$$

and the residuals do not depend on whether the model is overparameterized. Since any linear model can be reparameterized into a regression model, if a regression version of the model satisfies conditions (a) and (b), the theorems hold. In particular, for a one way ANOVA, I need each group sample size N_i to have $N_i/n \to \delta_i$ with $0 < \delta_i < \infty$.

3.1.3 Estimation of σ^2

The validity of Theorems 1 and 2 requires condition (b). The MSE from model (1.1) satisfies this but the tests work better if the estimate of σ does not inflate too much when H_0 is false. With an ordering imposed on the data, C-S proposed a consistent estimate of σ^2 obtained from an extended model:

$$\boldsymbol{Y}_n = \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\Gamma}_k \boldsymbol{\gamma}_k + \boldsymbol{e},$$

where $\boldsymbol{\Gamma}_k$ is a $n \times k$ discrete Fourier (sine and cosine) transformation matrix and $\boldsymbol{\gamma}_k$ is a $k \times 1$ vector of unknown parameters. I adopt C-S's estimate of σ^2 which is the MSE from this extended model, specifically

$$\hat{\sigma}_n^2 = \boldsymbol{Y}_n^T (\boldsymbol{I}_n - \boldsymbol{M}_{\boldsymbol{X}_n, \boldsymbol{\Gamma}_k}) \boldsymbol{Y}_n \Big/ [n - r(\boldsymbol{X}_n, \boldsymbol{\Gamma}_k)].$$

C-S suggest using $k = \lceil n/10(\log\log n)^2 \rceil$. By Lemma 3 of C-S, $\hat{\sigma}_n^2 = \sigma + O_p(1/\sqrt{n})$ whenever $k/n \to c$ where $0 \le c < 1$. Simulations suggest that this provides my tests with higher power than using the MSE of model (1.1).

3.2 Small sample adjustments

The sizes of the tests depend on the quality of the asymptotic approximation. In both T_n and Q_n, the mild dependence in the residuals hurts the convergence rate. This can be improved by choosing an appropriate $\tilde{n}$. In simulations, I studied the empirical size and power of my tests with four different choices of $\tilde{n}$: $\tilde{n}_0 = \lceil n/(\log\log n)^{2.5}\rceil$, $\tilde{n}_1 = \lceil n/(\log\log n)^3\rceil$, $\tilde{n}_2 = \lceil n/(\log\log n)^{3.5}\rceil$ and $\tilde{n}_3 = \lceil n/(\log\log n)^4\rceil$. I concluded that $\tilde{n}_1$ is best for the power of my tests while achieving a reasonable rate of convergence. The convergence can be further improved by adding a constant c_n that converges to 0 as $n \to \infty$. Note that Q_n normalized P_n by multiplying $a_{\tilde{n}}$ and adding $b_{\tilde{n}}$, where $b_{\tilde{n}} \to \infty$. Applying Theorem 2.2 of Eicker (1979), if $a_{\tilde{n}}/a'_{\tilde{n}} \to 1$, $b_{\tilde{n}} - b'_{\tilde{n}} \to 0$ and $a_{\tilde{n}}P_n - b_{\tilde{n}} \xrightarrow{\mathcal{L}} Q$, then $a'_{\tilde{n}}P_n - b'_{\tilde{n}} \xrightarrow{\mathcal{L}} Q$. So to improve the convergence, I can modify either $a_{\tilde{n}}$ or $b_{\tilde{n}}$. From extensive simulation, I found that modifying $b_{\tilde{n}}$ is preferable to modifying $a_{\tilde{n}}$ and without changing $a_{\tilde{n}}$, I want $b'_{\tilde{n}}$ less than $b_{\tilde{n}}$. The modified Q_n after proper adjustment is $Q_n + b_{\tilde{n}} - b'_{\tilde{n}} \equiv Q_n + c_{2n}$. Using the same idea, a different constant c_{1n} is added to T_n. Based on the chosen $\tilde{n}$ and extensive simulations, I found an appropriate adjustment for my first test to be $T_n + c_{1n}$ where

$$c_{1n} = \log\left(\frac{\log(\delta_{1n})}{(\log\log n)^4} + 1\right)$$

and

$$\delta_{1n} = \begin{cases} 0.839 + 0.08q + 0.1545q^2 + 2.583\hat{n} & n < 40, \\ 0.190 + 2.995q - 4.71\log(q) + 2.788\hat{n} & 40 \le n < 100, \\ 5.504 + 0.828q + 2.834\hat{n} & n \ge 100. \end{cases}$$

Here $\hat{n} = (\log\log n)^3$, $q = r(\boldsymbol{X}) - 1$. As $n \to \infty$, $\hat{n} \to \infty$, and it is easy to show that $c_{1n} \to 0$.

My second test is adjusted as $Q_n + c_{2n}$, where

$$c_{2n} = \log\left(\frac{\log(\delta_{2n})}{(\log\log n)^{2.5}} + 1\right)$$

and

$$\delta_{2n} = \begin{cases} \log(0.977 + 0.599q + 0.212\tilde{q} + 1.16\hat{n} - 0.194q\hat{n}) & n < 55, \\ -205.89 + 55q + 7.19q^2 + 102.8 - 26.4q\hat{n} & 55 \leq n < 100, \\ -309.8 + 37.74q + 1.26q^2 + 98.5\hat{n} & 100 \leq n < 500, \\ -492.6 + 22q + 120.6\hat{n} & n \geq 500, \end{cases}$$

Clearly, as $n \to \infty$, $c_{2n} \to 0$.

Similar techniques have been used in other large sample tests. For a statistic similar to my Q_n, C-S adjusted their test by raising the additive normalizing constant $b_{\tilde{n}}$ by a additional power c in the $\log$ function. They found a sophisticated expression for this c through extensive simulation. Their method is essentially the same as adding a constant that converges to 0 as $n \to \infty$. F-H's test also includes a hidden adjustment for small samples.

3.3 Ordering

The ordering of the data can have a great influence on the power of test statistics based on partial sums of residuals. S-W's test is based on a partial ordering of residuals. Residuals associated with lower covariates values are lower in the partial ordering. Using this ordering method, S-W's test is powerful for detecting lack-of-fit that occurs at lower covariates values. This is a very flexible approach, especially when the model contains multiple covariates. For example, with two covariates, the supremum over this partial ordering will equal or exceed the maximum when ordering the residuals according to either of the two covariates marginally. However, this flexibility for multiple covariates is paid for in computational time which increases exponentially as the number of predictors increases.

The problem with the S-W simulation is the partial ordering of their data. If I completely order the data, their simulation method becomes a viable alternative to using the asymptotic distribution to define tests. However, as mentioned earlier, the validity of the simulation presupposes asymptotic convergence.

I first considered ordering the residuals according to the Mahablanobis distance of the predictor variables. Unfortunately, this ordering did not yield rapid convergence and the convergence rates differed for predictors from different distributions. I believe this is caused by using the sample mean as the center of the distribution. For example, with a single covariate, if x is from a symmetrical distribution, Mahablanobis distance assigns both tails of x lower orderings. However, if the data is right skewed, Mahablanobis distance primarily assigns data associated with high values of x to lower orderings. To alleviate this I considered centering at the midrange.

I tried several modifications of Mahablanobis distance with different center estimates and I found an effective ordering method that stabilized the convergence rates. First, identify a set of columns $\boldsymbol{X}_n^0$ that contains the covariates suspected of lack-of-fit. Compute the midranges of each column of $\boldsymbol{X}_n^0$, say $\boldsymbol{\eta} = (\eta_1, ..., \eta_k)'$. For the ith observation, define $d_i = (\boldsymbol{x}_i^0 - \boldsymbol{\eta})'\boldsymbol{S}^-(\boldsymbol{x}_i^0 - \boldsymbol{\eta})$, where $\boldsymbol{S}^-$ is the generalized inverse of the usual covariance matrix of $\boldsymbol{X}^0$. The observations are ordered from large to small values of d_i.

As in the proofs of Theorem 1 and Theorem 2, the convergence of T_n and Q_n depends partly on S_n and $\tilde{S}_n$ converging to 0, which in turn mildly depend on the behavior of the predictors. Through extensive simulations I found that with an appropriate choice of $\tilde{n}$ and this ordering method, the predictors' impact on the size of the test is negligible.

3.4 Simulations

In this Section I study the empirical sizes and powers of the proposed tests. Their empirical powers are compared with S-W's test in Section 2.2.1. The significant level is set to be

0.05. For fitting simple linear regression, the results are based on 6000 simulations and the sample size is $n = 64$. For multiple regression models, due to the extensive computation required by S-W's test, the results are based on 1800 simulations. I also took $n = 50$ and restricted the number of predictor variables to 2, so that $p = 3$. For Theorems 1 and 2 to hold, I need $\tilde{n} < \lceil n/(\log\log n) \rceil$, for T_n, and $\tilde{n} < \lceil n/(\log\log n)^2 \rceil$ for Q_n, with $\tilde{n} \to \infty$ as $n \to \infty$. The effects of different $\tilde{n}$s were also studied and representative results are presented. To simplify notation, denote T_n using $\tilde{n}_j$ from Section 3.2 as T_{nj} and similarly Q_n using $\tilde{n}_j$ as Q_{nj}. For S-W's test, I also present its performance using my complete ordering in additional to their partial ordering. As mentioned earlier the simulation results suggested that $\tilde{n}_1$ improves the performance of my tests.

The sizes of T_n and Q_n, after small sample adjustment, were studied separately by simulating predictors from various distributions and fitting data with various models. In Section 3.4.2 they are checked under 9 different sample sizes from $n = 48$ to $n = 200$. Instead of S-W's test, my tests' empirical sizes are compared with another two tests that have known exact or limiting distribution. Specifically, they are Fisher's exact test mentioned in Section 2.1.1 and Christensen and Sun (2010)'s first test. Results are based on 20000 simulations and the significance level is 5%. The sizes of my tests are shown with comparable size for C-S's test and Fisher's exact test.

My methods provide higher power and computation speed than S-W's test. With $p - 1$ predictors and m bootstrap samples, computing the P value of S-W's test involves $W_n(\boldsymbol{t})$ being evaluated $n^{p-1}(m + 1)$ times. For $n = 50$, $m = 500$, $p - 1 = 3$, $n^{p-1}(m + 1)$ is 6.26×10^7. My method requires $\tilde{n}_1 = \lceil n/(\log\log n)^3 \rceil$ evaluations. Comparing to S-W's test, the computation time for my tests are significantly reduced by adopting a full ordering method and using limiting distribution to evaluate P values.

In most cases, my two tests perform similarly. Since Q_n puts more weight on the residuals lower in the ordering, T_n is usually more powerful than Q_n when the lack-of-fit exists in a relatively long string of residuals of lower orders. On the other hand, when the lack-of-fit appears only in the first few residuals of lower order, Q_n is more powerful.

3.4.1 Power of the tests

To begin with, I examine lack-of-fit when fitting a simple linear regression $y_i = \beta_0 + x_i\beta_1 + \epsilon_i$. Neither S-W's test nor my tests require normality of the error terms, but for convenience the ϵ_is are simulated from $N(0, 2)$.

Through various simulations I found that $\tilde{n}_3$ yields similar results to $\tilde{n}_2$ and that $\tilde{n}_1$ yields similar results to $\tilde{n}_0$. Using $\tilde{n}_1$ is noticeably more powerful than using $\tilde{n}_3$. In the following examples, to simplify the graphs, I only show the empirical powers for $\tilde{n}_1$.

EXAMPLE 3.1. The independent variable x is simulated from a $N(0, 1)$, and the dependent variable y is drawn from the nonlinear model

$$y = 1 + \exp(\theta x) + \epsilon$$

Figure 3.1 shows that T_{n1} is more powerful than Q_{n1} or either version of S-W's tests. For example, at $\theta = 0.75$, T_{n1} is around 9% more powerful than S-W's test with my ordering and 39% more powerful than S-W's test with their original ordering. S-W's test with my ordering is more powerful than Q_{n1} which is more powerful than S-W's original test. At $\theta = 0.75$, Q_{n1} is about 13% less powerful than S-W's test with my ordering and 10% more powerful than S-W's original test.

EXAMPLE 3.2. The independent variable x is simulated from $N(0, 1)$, and the dependent variable y is drawn from a quadratic model that is linear in the parameter θ.

$$y = 1 + \theta x^2 + \epsilon.$$

The results of Example 3.2 are similar to Example 3.1 and not plotted. T_{n1} outperforms the other tests, S-W's test with my ordering is more powerful than Q_{n1}, and S-W's original

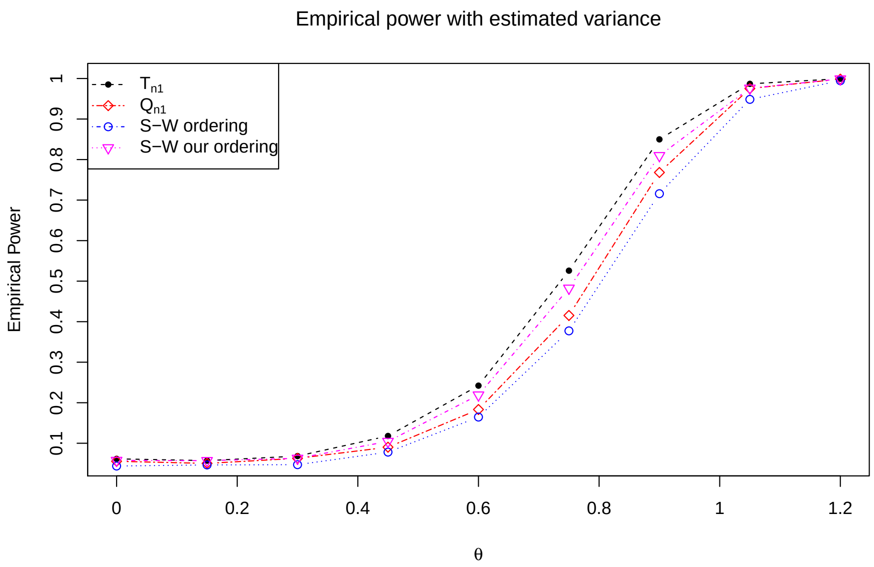

Figure 3.1: Lack-of-fit from Example 3.1

test is the weakest. The difference between T_{n1} and S-W's original test is larger in this example.

EXAMPLE 3.3. The independent variable x is simulated from a $U(-2, 2)$, and the dependent variable y is drawn from the linear model

$$y = 1 + 2x + \theta\cos(x) + \epsilon.$$

The results of Example 3.3 are a little different from Example 3.1. Here, the differences between using $\tilde{n}_1$ and $\tilde{n}_0$ and between using $\tilde{n}_3$ and $\tilde{n}_2$ are more apparent. T_{n1} and S-W's test with my ordering are the most powerful tests and they behave almost the same. Q_{n1}, and S-W's original test are much weaker and they behave nearly the same.

EXAMPLE 3.4. The independent variable x is simulated from a $U(0.1, 2)$ and the dependent variable y is drawn from

$$y = 1 + \frac{\theta}{x} + \epsilon.$$

In this example, different $\tilde{n}$s have little effect on T_n. For Q_n, $\tilde{n}_0$ and $\tilde{n}_1$ are more powerful than $\tilde{n}_2$ and $\tilde{n}_3$. Figure 3.2 shows that empirical powers for T_{n1} and Q_{n1} are close to each other and both outperform either version of S-W's tests. S-W's original test is still the least powerful. At $\theta = 0.8$, T_{n1} and Q_{n1} are about 25% more powerful than S-W's test with my ordering and 35% more powerful than S-W's original test.

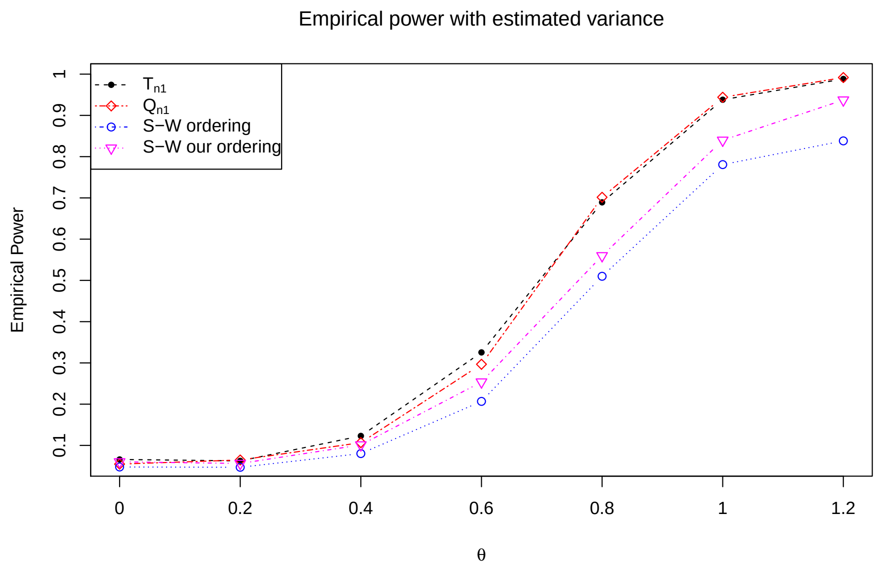

Figure 3.2: Lack-of-fit from Example 3.4

If x is simulated from $U(0.01, 2)$, the curvature of plotting y against x will be more apparent near 0, and the lack-of-fit will quickly appear at the first few data points. As expected, Q_{n1} is more powerful than T_{n1}. At $\theta = 0.2$, Q_{n1} is about 12% more powerful than T_{n1}. (Graph is not shown.)

Now compare the power of my tests with S-W's tests for fitting multiple regression models. The fitted model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Recall that S-W's original test is not viable for moderate to large p.

EXAMPLE 3.5. The independent variables x_1 and x_2 are independently simulated from $N(0, 1)$ and the dependent variable y is the same as Example 3.1,

$$y = 1 + \exp(\theta x_1) + \epsilon$$

and variable x_2 is independent of y. If I correctly identify the lack-of-fit as likely to come from x_1, Figure 3.3 shows that Q_{n1} is more powerful than T_{n1} which is much more powerful than either version of S-W's tests. Although adding a noisy predictor into the fitted model, Q_{n1} and T_{n1} are almost as powerful as in Example 3.1, compared to which Q_{n1} loses almost no power and T_{n1} mildly decreases in power. However, both versions of S-W's test have their powers significantly reduced.

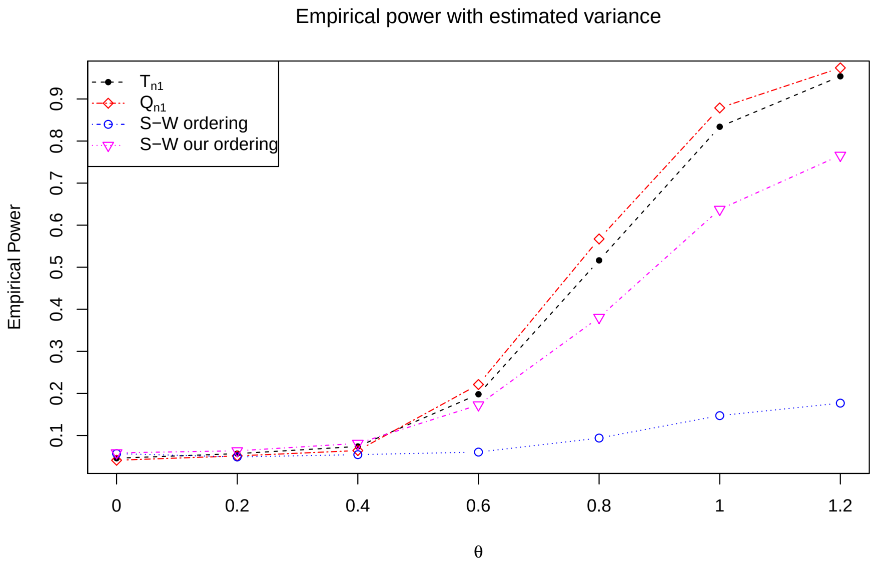

Figure 3.3: Lack-of-fit from Example 3.5: my ordering based on x_1

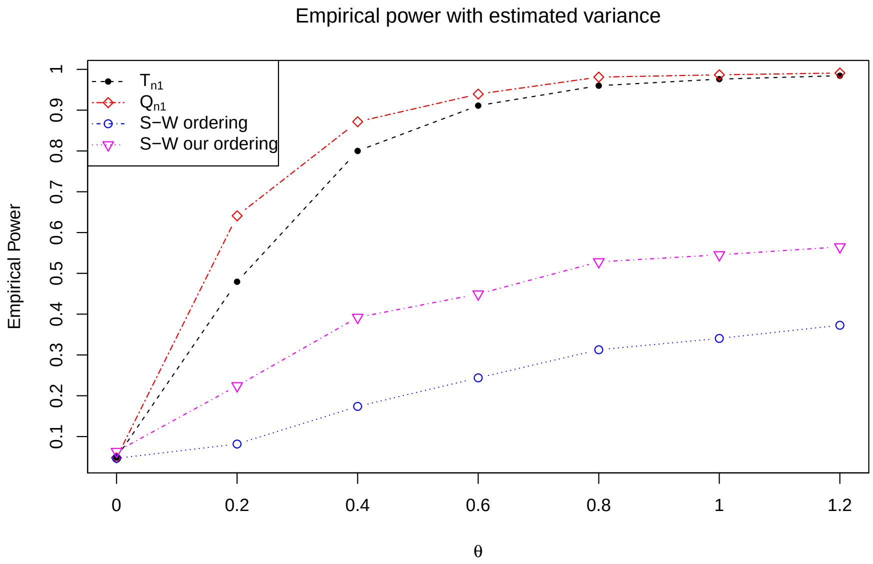

Figure 3.4: Lack-of-fit from Example 3.6: my ordering based on x_1

In fact, S-W's original test has almost no power. I believe this is due partly to the fact that S-W's ordering is no longer efficient when some of the predictors in the fitted model are weakly related to the dependent variable. Even for S-W's test with my ordering, the power reduction is large. If I order the data according to x_1 and x_2 jointly, my tests moderately decrease in power, but are still more powerful than S-W's tests. (Graph is not shown.) A similar example with x_1 and x_2 weakly dependent using my ordering based on both x_1 and x_2 is postponed to Example 3.7.

EXAMPLE 3.6. The variable x_1 is $U(0.01, 2)$ independent of x_2 which is $N(0, 1)$ and the dependent variable y is the same as in Example 3.4,

$$y = 1 + \frac{\theta}{x_1} + \epsilon.$$

I again identify the lack-of-fit as likely to come from x_1. Unlike Example 3.4, x_1 is simulated from $U(0.01, 2)$ instead of $U(0.1, 2)$. Figure 3.4 shows that T_{n1} is still more powerful than S-W's tests and the differences between my test T_{n1} and S-W's tests are dramatic. Q_{n1}

is clearly more powerful than T_{n1}, especially for $\theta < 0.4$. At $\theta = 0.2$, Q_{n1} is about 30% more powerful than T_{n1}, 260% more powerful than S-W's test with my ordering and 760% more powerful than S-W's original test.

EXAMPLE 3.7. The variables x_1 and x_2 are $N(0, 1)$ with a correlation 0.3 and the dependent variable y is as in Example 3.4. Here I correctly identify the lack-of-fit as likely to come from both x_1 and x_2 and use my ordering. Figure 3.5 shows the powers.

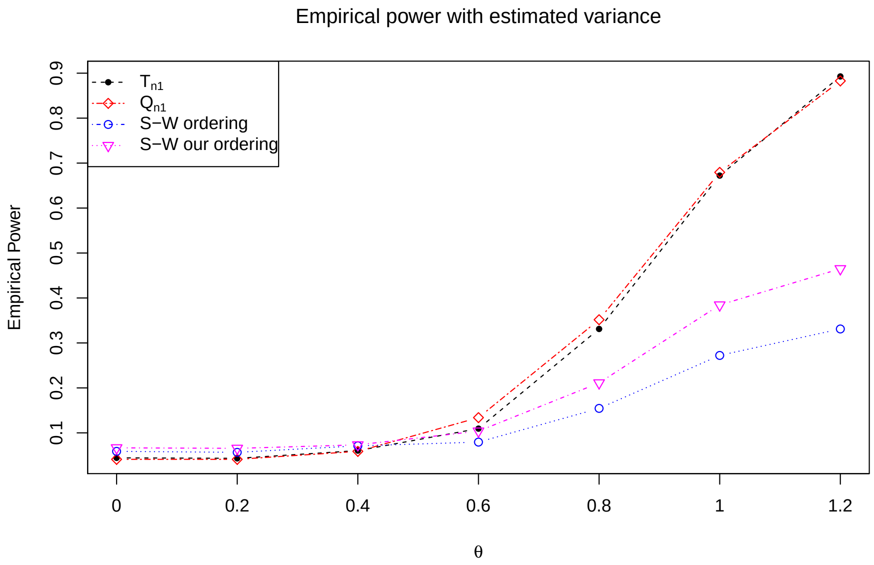

Figure 3.5: Lack-of-fit from Example 3.7: my ordering based on x_1 and x_2

Compared to Figure 3.3, Q_{n1}, T_{n1} and S-W's test with my ordering all decrease in power. S-W's original test on the other hand mildly increases in power. Q_{n1} and T_{n1} now behave close to each other and both are still more powerful than S-W's test with either ordering. S-W's original test remains the weakest. Note that, in this example the sizes for S-W's tests are inflated to around 0.06.

3.4.2 Size of the tests

This Section studies the empirical sizes of my proposed tests. In each of following examples, multiple cases are considered.

EXAMPLE 3.8. Independent variable x is simulated from $N(0,1)$. The fitted model is $y_i = \beta_0 + x_i\beta_1 + \epsilon_i$. The dependent variable are drawn from three distributions

$$(a) \quad y = 10 + \epsilon, \quad (b) \quad y = 1 + \epsilon, \quad \text{and} \quad (c) \quad y = 1 + 2x + \epsilon,$$

where ϵ_is are independently simulated from $N(0,2)$. The following three sets of graphs show that the performances of my two tests are similar to C-S's test 1 and Fisher's exact test and all of them are consistently around 0.05. Also, the three different scenarios almost give the same feature.

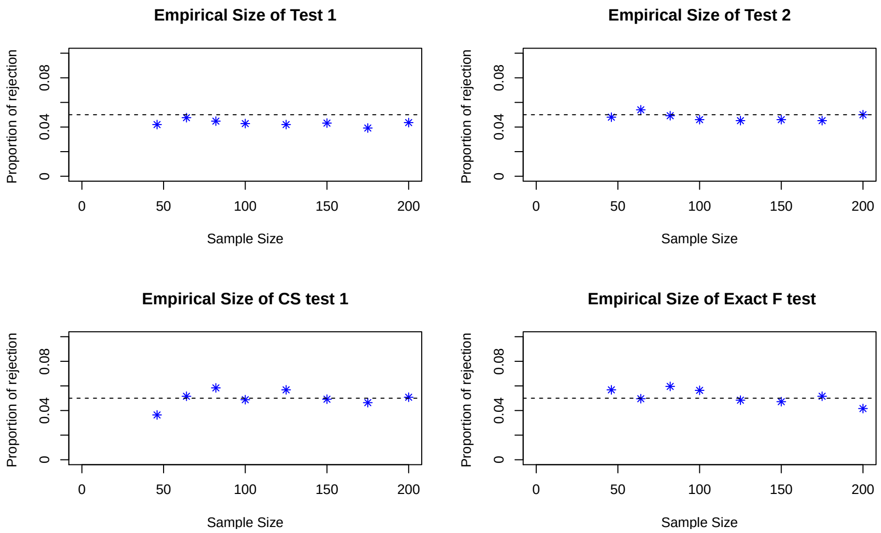

Figure 3.6: Sizes comparison from Example 3.8 (a)

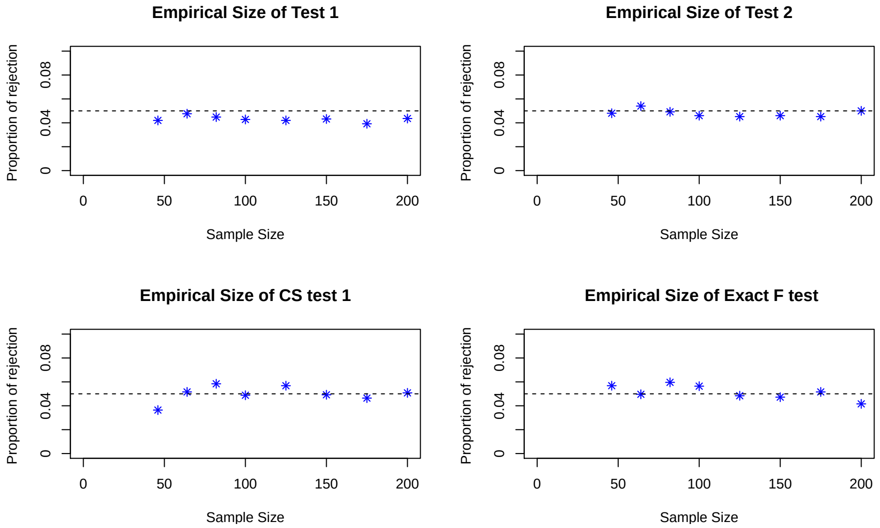

Figure 3.7: Sizes comparison from Example 3.8 (b)

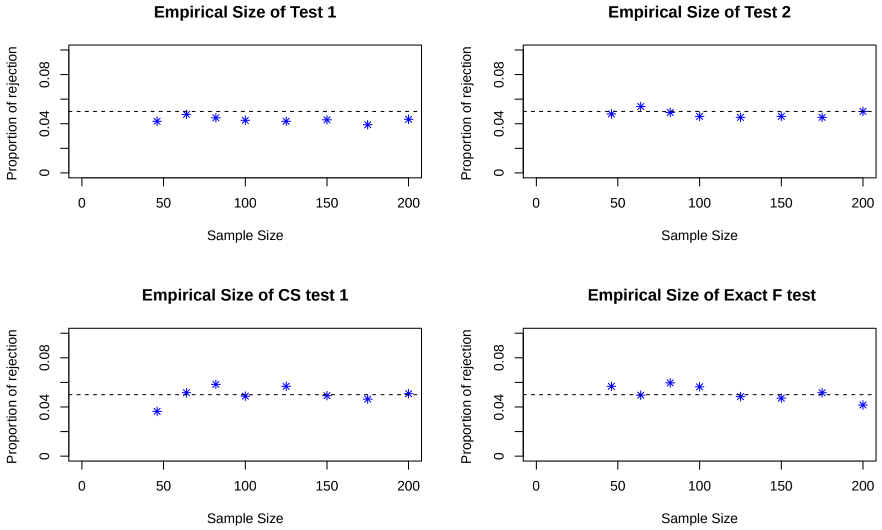

Figure 3.8: Sizes comparison from Example 3.8 (c)

EXAMPLE 3.9. The dependent variable is drawn from $y = 1 + 2x + \epsilon$. The fitted model is the same as in Example 3.8 and I modify the ways of simulating predictors. Three different distributions are considered. They are (a) $N(0, 10)$, (b) a strong skewed distribution with outliers $1/U(0.1, 2)$ and mild skewed distribution $\exp(5)$. Simulation results show that the size of my first proposed test is a little less than but close to 0.05 level. The empirical sizes for my second test are consistent for these three cases.

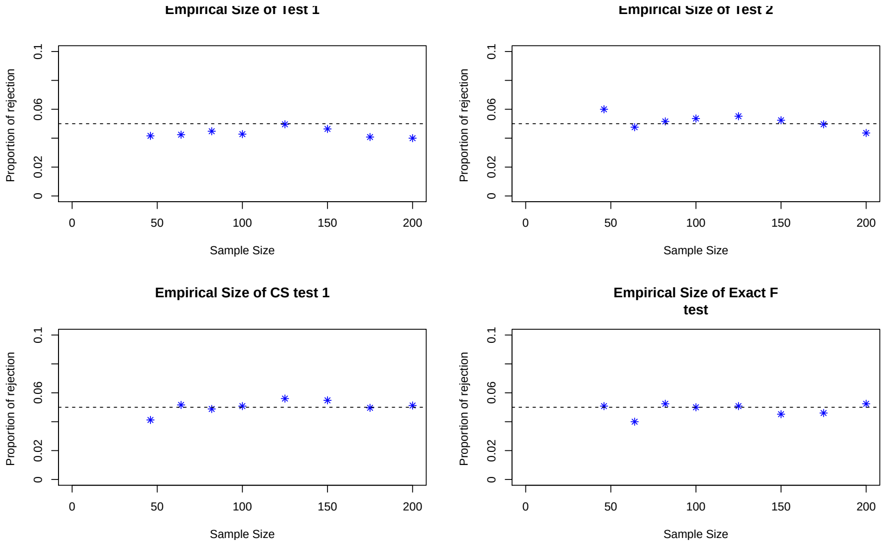

Figure 3.9: Sizes comparison from Example 3.9 (a)

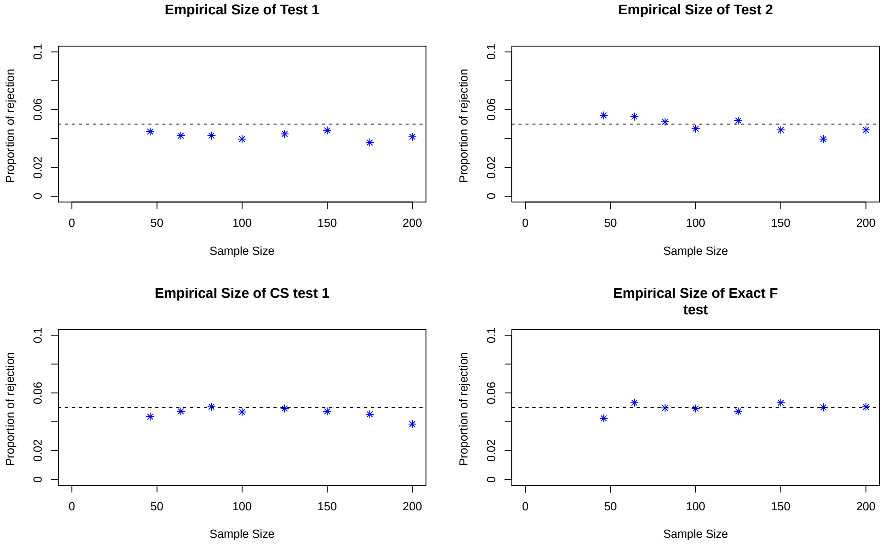

Figure 3.10: Sizes comparison from Example 3.9 (b)

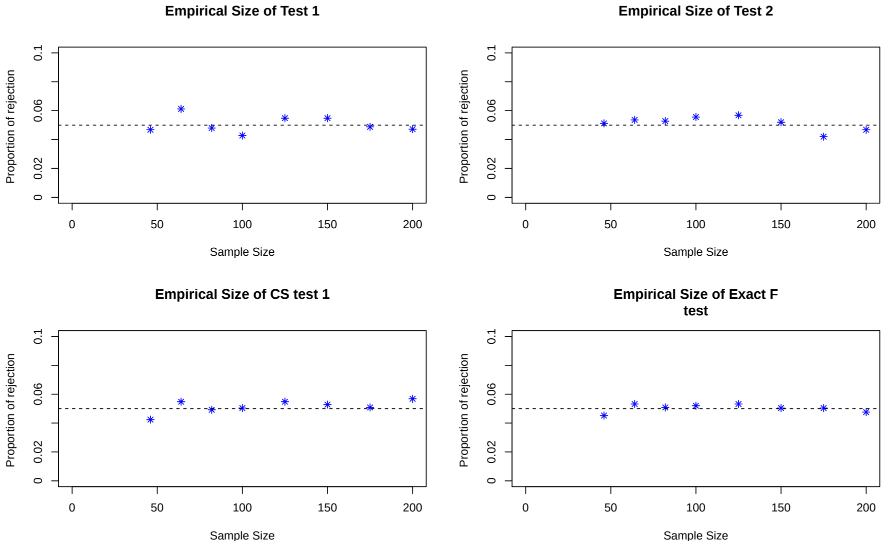

Figure 3.11: Sizes comparison from Example 3.9 (c)

EXAMPLE 3.10. Similar to Example 3.9, x is simulated from three distributions, (a) $N(0,1)$, (b) $\exp(5)$ and (c) $U(0,2)$.

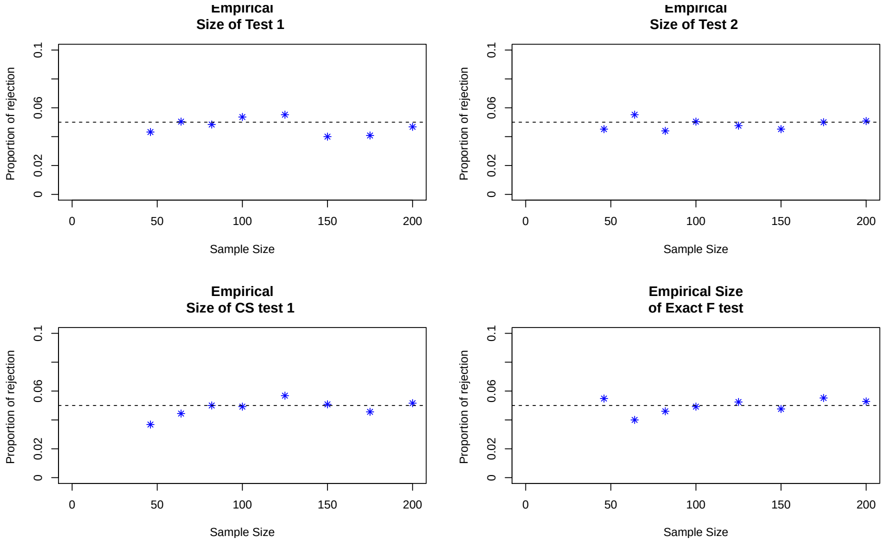

Figure 3.12: Sizes comparison from Example 3.10 (a)

The dependent variable is simulated from $y = 1 + \epsilon$. However, this time a over parameterized model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ is fitted to the data.

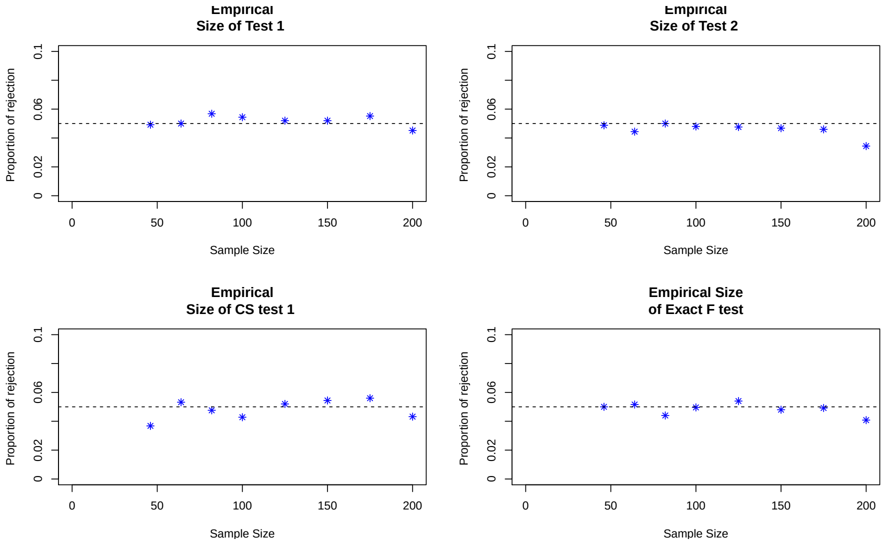

Figure 3.13: Sizes comparison from Example 3.10 (b)

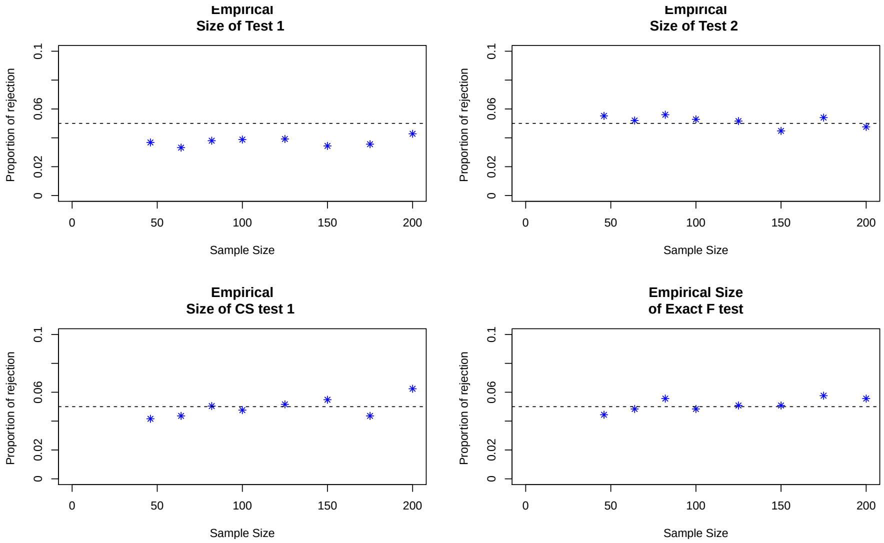

Figure 3.14: Sizes comparison from Example 3.10 (c)

This is an important concern; C-S pointed out that the size of F-H's test is greatly reduced when true model is a simple linear model but fitted with an additional quadratic term. The sizes for all four tests show more differences than the previous examples. Especially, in Example 3.10 (c), the size of my test 1 is a little below 0.04 but close to 0.05 when $n = 200$. In the meanwhile, the size of C-S's test 1 shows the same level of variations, whereas the sizes for my test 2 and Fisher's exact test are consistently around 0.05.

EXAMPLE 3.11. Use same setting as in 3.10, but the fitted model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$. The simulation result is similar to Example 3.8 and Example 3.9. (Graphs are not shown)

EXAMPLE 3.12. Keep the same setting as Example 3.11, but fit the data with a nonlinear function of x. That is $y_i = \beta_0 + \beta_1 x_i + \beta_2 / x_i$. The simulation result is similar to previous examples. (Graphs are not shown)

EXAMPLE 3.13. Continuation of Example 3.12, the fitted model is $y = \beta_0 + \beta_1 x + \beta_2 \cos(x)$. The simulation result is similar to previous examples. (Graphs are not shown)

EXAMPLE 3.14. Continuation of Example 3.12 with fitted model simulated from $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x + \beta_4 x^2$. The simulation result is similar to previous examples. My tests are consistent around 0.05. This time the results are shown for evidence.

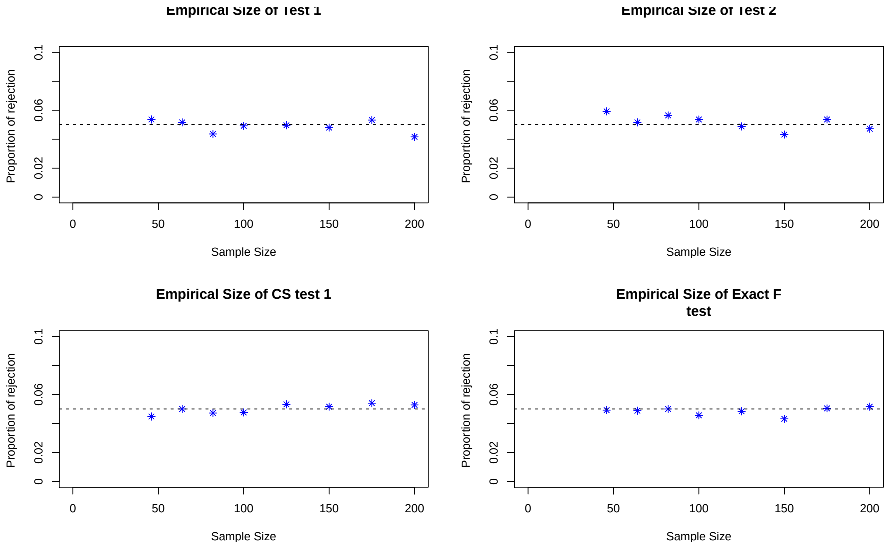

Figure 3.15: Sizes comparison from Example 3.14 (a)

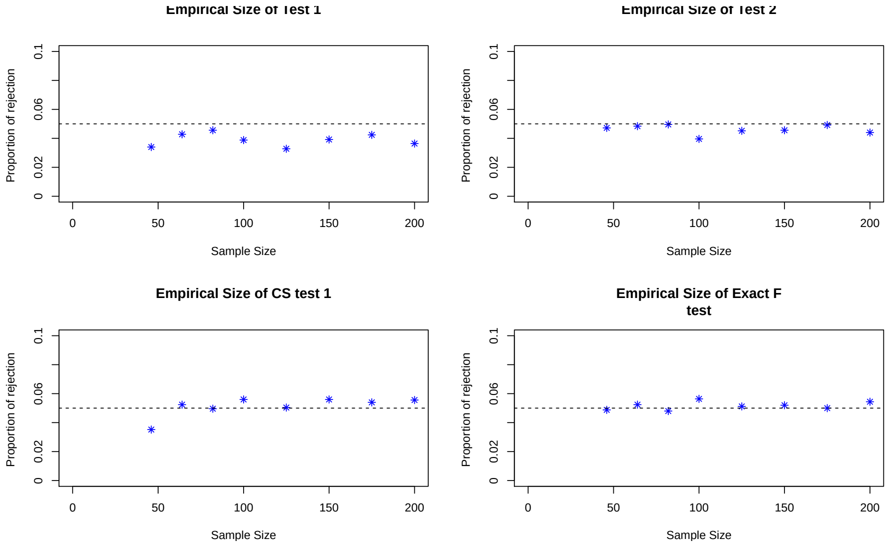

Figure 3.16: Sizes comparison from Example 3.14 (b)

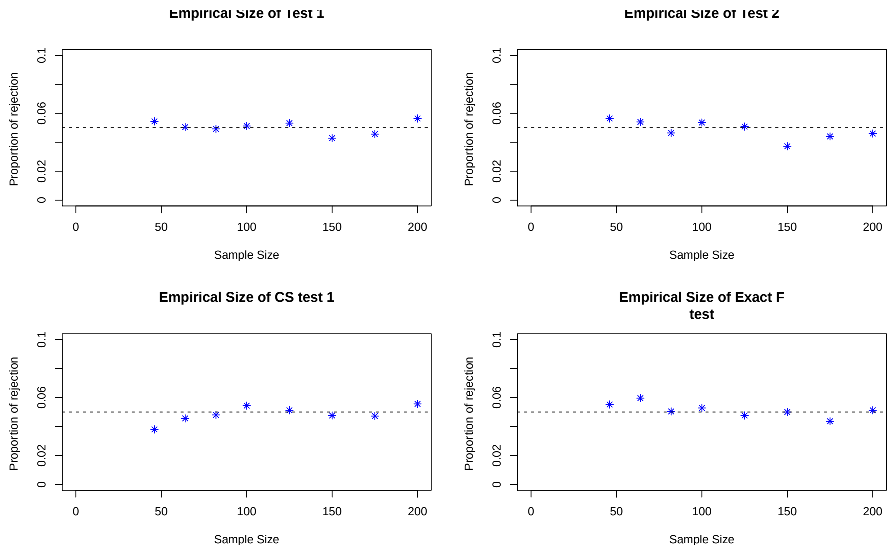

Figure 3.17: Sizes comparison from Example 3.14 (c)

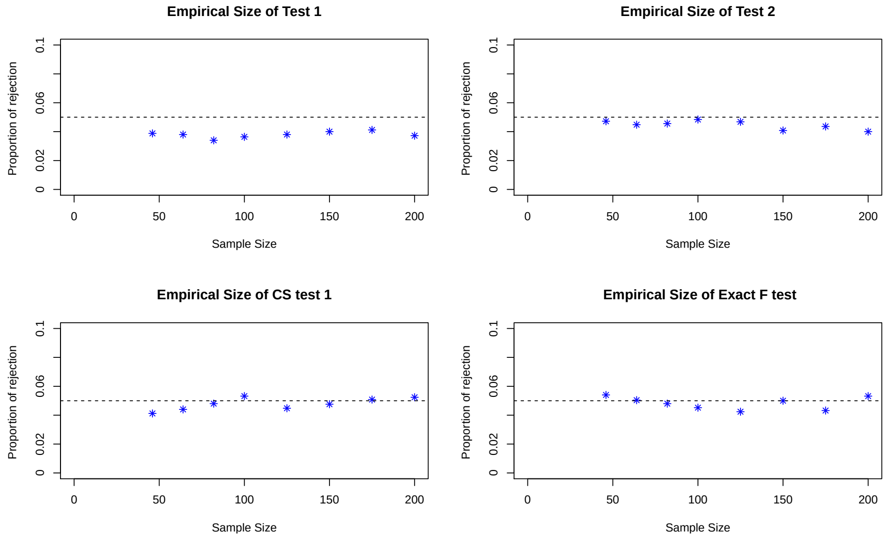

Figure 3.18: Sizes comparison from Example 3.15

EXAMPLE 3.15. Use the same setting as in Example 3.14 but the independent variable x is simulated from distributions with large variance. They are (a) $U(-200, 200)$, and (b) $N(0, 50)$. The results are similar in these two cases. Here I only show the result of case

(a). The figure shows that, the sizes of my test 1 is around 0.04 and the size of my test 2 is a about in between 0.05 and 0.04. They are not as consistent as Fisher's test and CS's test 1, but still acceptable.

3.5 Summary

I studied the limiting behavior of S-W's test under linear models and a complete ordering of the data. I proposed two tests based on using the asymptotic distributions of the maximized partial sum of residuals to check the goodness-of-fit of the fitted model. My tests do not rely on simulating the partial sum process to evaluate the P values. Instead, I found the limiting null distributions of my test statistics, and adjusted them for small sample accuracy. The empirical power studies show that my first proposed test has high power when the fitted model is a simple concave or convex function while the second test is powerful to detect lack-of-fit that occurs at the lower orderings. The size of my test are checked in various cases and shown to have consistency. I hope to extended my tests to generalized linear models.

Chapter 4

Linear models that allow perfect estimation

The Gauss-Markov model (1.3) with singular $\boldsymbol{V}$ allows perfect estimation which means a function of $\boldsymbol{\beta}$ can be learned with probability one. Since the initial work by Goldman and Zelen (1964), this problem has been studied by many authors, see, for example Albert (1973), Kreijger and Neudecker (1977); Rao (1967, 1968); Harville (1981); Puntanen and Styan (1989); Kempthorne (1989). In this Chapter, I proposed a new way of dealing with general Gauss-Markov models like (1.3) with $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$. Traditional approach is briefly reviewed Section 4.1. Section 4.2 outlines some historical background and discusses perfect estimation. Section 4.3 introduces methods for when only part of $\boldsymbol{X\beta}$ is perfectly estimable. In Section 4.4, I develop hypothesis tests based on the findings in Section 4.3. A general overview of the traditional approach is presented in the Appendix D.

4.1 The problem

The analysis of model (1.3) depends crucially on whether the column space of $\boldsymbol{X}$, $C(\boldsymbol{X})$, has the property $C(\boldsymbol{X}) \subset C(\boldsymbol{V})$ or $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$. When $C(\boldsymbol{X}) \subset C(\boldsymbol{V})$, with probability one that $\boldsymbol{Y} \in C(\boldsymbol{V})$ and the analysis can proceed almost as if $\boldsymbol{V}$ is positive definite. In particular, the BLUE of $\boldsymbol{X\beta}$ is simply

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

with probability one. Under multivariate normality, the test of an estimable hypothesis $\boldsymbol{\lambda}'\boldsymbol{\beta} = \boldsymbol{0}$, where $\boldsymbol{\lambda}' = \boldsymbol{\rho}'\boldsymbol{X}$ for some vector $\boldsymbol{\rho}$, or of an equivalent reduced model (cf. Christensen, 2002, Section 3.3) $\boldsymbol{Y} = \boldsymbol{X}_0\boldsymbol{\delta} + \boldsymbol{e}$ with $C(\boldsymbol{X}_0) \subset C(\boldsymbol{X})$ has F statistic

$$\frac{(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_0\hat{\boldsymbol{\delta}})'\boldsymbol{V}^{-}(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_0\hat{\boldsymbol{\delta}})/[r(\boldsymbol{X}) - r(\boldsymbol{X}_0)]}{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{V}^{-}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/[r(\boldsymbol{V}) - r(\boldsymbol{X})]}.$$

These results hold for any choice of the generalized inverse. Moreover, both results look like (and are) similar to results for $\boldsymbol{V}$ positive definite. One simply reduces the vector space in question to $C(\boldsymbol{V})$ rather than $\mathbf{R}^n$ except that observing $\boldsymbol{Y} \notin C(\boldsymbol{V})$ would cause us to reject model (1.3).

It turns out that equally simple formulae work when $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$. Let

$$\boldsymbol{T} = \boldsymbol{V} + \boldsymbol{XUX}'$$

where $\boldsymbol{U}$ is nonnegative definite having $C(\boldsymbol{X}) \subset C(\boldsymbol{T})$ so that $C(\boldsymbol{X}, \boldsymbol{V}) = C(\boldsymbol{T})$. In particular, $\boldsymbol{U} = \boldsymbol{I}$ gives such a matrix $\boldsymbol{T}$. Then the BLUE of $\boldsymbol{X\beta}$ in model (1.3) is

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{T}^{-}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{T}^{-}\boldsymbol{Y}$$

and the F statistic is

$$\frac{(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_0\hat{\boldsymbol{\delta}})'\boldsymbol{T}^{-}(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}_0\hat{\boldsymbol{\delta}})/[r(\boldsymbol{X}) - r(\boldsymbol{X}_0)]}{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{T}^{-}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/[r(\boldsymbol{T}) - r(\boldsymbol{X})]}.$$

While these formulae look simple, there is much more here than meets the eye. Why do such $\boldsymbol{T}$ matrices give the correct answers and what about the well known fact that when $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$ there are some linear functions of $\boldsymbol{X\beta}$ that are known perfectly, i.e., with probability one?

Suppose $\boldsymbol{a}$ is a vector with $\boldsymbol{a}'\boldsymbol{V} = 0$ but $\boldsymbol{a}'\boldsymbol{X} \neq \boldsymbol{0}$. With $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$ such vectors exist. From model (1.3), $\boldsymbol{a}'\boldsymbol{Y}$ has variance 0 so $\boldsymbol{a}'\boldsymbol{Y} = \boldsymbol{a}'\boldsymbol{X\beta}$ a.s., and $\boldsymbol{a}'\boldsymbol{X\beta}$ is known perfectly. Whereas in the model (1.4) used to obtain BLUEs and F statistics for model (1.3). The variance of $\boldsymbol{a}'\boldsymbol{Y}$ reduces to $\sigma^2\boldsymbol{a}'\boldsymbol{XUX}'\boldsymbol{a} > 0$. So in model (1.4), things that are perfectly estimated in model (1.3) are no longer perfectly estimated. Model (1.4) can be viewed as a model with two independent error terms, $\boldsymbol{e} = \boldsymbol{e}_1 + \boldsymbol{e}_2$, where $Cov(\boldsymbol{e}_1) = \sigma^2\boldsymbol{V}$ and $Cov(\boldsymbol{e}_2) = \sigma^2\boldsymbol{XUX}'$. For unbiased estimation $\boldsymbol{e}_2$ could just as well be zero because $Pr[\boldsymbol{e}_2 \in C(\boldsymbol{X})] = 1$ and variability in the estimation space $C(\boldsymbol{X})$ is ignored by unbiased estimates.

The approach I take to examine model (1.3) is unique. When $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, I isolate what can be known about $\boldsymbol{X\beta}$ perfectly and then fit a reduced model (1.5). This model contains all the relevant variability for estimating $\boldsymbol{X\beta}$. My approach emphasizes the role of perfect estimation and avoids the incorporation of an unintuitive pseudo-covariance matrix $\boldsymbol{T}$. I think my approach allows a clearer understanding of the issues involved in fitting linear models with $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$. In the following Sections of this chapter, I begin with some historical background and a discussion of perfect estimation. This leads us to fitting model (1.5) and showing that the simple intuitive procedures involved in fitting (1.5) lead to the appropriate procedures for fitting model (1.3) and agree with the simple but mysterious procedures of fitting model (1.4). Obviously if model (1.5) is equivalent to model (1.3) when $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, it better produce the same BLUEs and F tests as models (1.3) and (1.4).

One potential application of results on linear models with singular covariance matrices is for modeling the residuals of model (1.1). In that case, fitting model (1.1) with least squares gives residuals $\hat{\boldsymbol{e}} \equiv (\boldsymbol{I} - \boldsymbol{M_X})\boldsymbol{Y}$. A linear model for the residuals has

$$\hat{\boldsymbol{e}} = \boldsymbol{\Gamma\xi} + \tilde{e} \quad E(\tilde{\boldsymbol{e}}) = \boldsymbol{0}, \quad Cov(\tilde{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{M_X}).$$

This is essentially what Fan and Huang (2001) do when they apply Fourier transforms to the residuals of a linear model but they ignore the covariance structure and use least squares estimates. If $C(\boldsymbol{\Gamma}) \not\subset C(\boldsymbol{I} - \boldsymbol{M_X})$, the nontrivial parameter $\boldsymbol{M_X\Gamma\xi}$ is known with probability 1. (I generally prefer incorporating $\boldsymbol{\Gamma\xi}$ into model (1.1) to get model (2.8) rather than fitting models to the residuals.)

4.2 Historical review

The general Gauss-Markov model (1.3) has been well studied. Although most studies focus on nonsingular $\boldsymbol{V}$, the possibility of singular covariances still draws attention. Christensen (2002) classifies the linear model into 4 categories, depending on the assumptions made about $\boldsymbol{V}$:

(a) $\boldsymbol{V}$ is the identity matrix,

(b) $\boldsymbol{V}$ is positive definite,

(c) $\boldsymbol{V}$ is nonnegative definite with $C(\boldsymbol{X}) \subset C(\boldsymbol{V})$, and

(d) $\boldsymbol{V}$ is nonnegative definite.

These categories are increasingly general. Groß (2004) gave a survey of important results for linear models with possibly singular covariance matrixes. Rao (1967, 1968) provided necessary and sufficient conditions for the equivalence of the Ordinary Least Square Estimate (OLSE) and the Best Linear Unbiased Estimate (BLUE) of $\boldsymbol{X\beta}$. His two conditions are: $\boldsymbol{X'VZ} = \mathbf{0}$, or $\boldsymbol{V} = \boldsymbol{aI} + \boldsymbol{XBX'} + \boldsymbol{ZEZ'}$ in which $C(\boldsymbol{Z}) = C(\boldsymbol{X})^{\perp}$ and $\boldsymbol{B}$, $\boldsymbol{E}$ are symmetric with $\boldsymbol{a}, \boldsymbol{B}, \boldsymbol{E}$ chosen so that $\boldsymbol{V}$ is nonnegative definite. Kreijger and Neudecker (1977) introduced a method of linear estimation under the parameter restrictions determined by the singularity of $\boldsymbol{V}$. They use $\boldsymbol{G}' = (\boldsymbol{I} - \boldsymbol{VV}^{+})$. In this way, $\boldsymbol{G}'\boldsymbol{X\beta} = \boldsymbol{G}'\boldsymbol{Y}$ with probability 1 and $\boldsymbol{G}'\boldsymbol{X\beta}$ is nontrivial only if $\boldsymbol{G}'\boldsymbol{X} \neq \mathbf{0}$. Any other generalized inverse for $\boldsymbol{V}$, say $\boldsymbol{V}^{-}$ has the same two properties. Harville (1981) introduced ways to find minimum-variance unbiased estimates of estimable functions $\boldsymbol{\rho' X\beta}$ under singular covariance matrixes. He suggests finding a matrix $\boldsymbol{Q}$ where $\boldsymbol{Q'V} = \mathbf{0}$ and $r(\boldsymbol{Q}) = n - r(\boldsymbol{V})$, so that $\boldsymbol{b} = \boldsymbol{Q'Y} = \boldsymbol{Q'X\beta}$ with probability 1. Then he forces the estimator $\boldsymbol{c} + \boldsymbol{a'Y}$ of $\boldsymbol{\rho' X\beta}$ to depend on $\boldsymbol{b}$. With this method, he shows the proposed estimator is the minimum variance unbiased estimator of $\boldsymbol{\rho' X\beta}$. Puntanen and Scott (1996) showed that the BLUE of $\boldsymbol{X\beta}$ is $[\boldsymbol{I} - \boldsymbol{VN}(\boldsymbol{NVN})^{-}\boldsymbol{N}]\boldsymbol{Y}$ for any choice of general inverse $(\boldsymbol{NVN})^{-}$, where $\boldsymbol{V}$ can be deficient in rank, and $\boldsymbol{N}$ is the perpendicular projection operator onto $C(\boldsymbol{X})^{\perp}$. They also show that the covariance matrix of the BLUE of $\boldsymbol{X\beta}$ is $\mathbf{0}$ if and only if $C(\boldsymbol{V}) \cap C(\boldsymbol{X}) = \{\mathbf{0}\}$. Puntanen and Styan (1989) reviewed conditions for the OLSE to be BLUE, see also Christensen (1990), Harville (1990) and Kempthorne (1989). I focus on cases with $\boldsymbol{V}$ singular and $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$. My interest is in the role of perfect estimation, i.e., linear functions of $\boldsymbol{\beta}$ that are known with probability one, and how perfect estimation relates to estimating other functions of $\boldsymbol{\beta}$ and testing.

Theorem 3 *If $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, there exist nontrivial estimable functions that can be estimated perfectly. These functions are the linear functions of $\boldsymbol{Q}'\boldsymbol{X\beta}$ where $\boldsymbol{Q}$ is a full column rank matrix with $C(\boldsymbol{Q}) = C(\boldsymbol{V})^{\perp}$.*

This result is well known. The key feature is choosing $\boldsymbol{Q}$ with $C(\boldsymbol{Q}) = C(\boldsymbol{V})^{\perp}$. My $\boldsymbol{Q}'$ is obviously equivalent to Harville's (1981) $\boldsymbol{Q}'$ and Kreijger and Neudecker's (1977) $\boldsymbol{G}'$ matrix, except that I specify full column rank. Zyskind (1967) and Zyskind and Martin (1969) choose $\boldsymbol{Q}$ to be a matrix whose columns are orthonormal eigenvectors of $\boldsymbol{V}$ with respect to the eigenvalue 0. This leads to $\boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{V}}$.

The existence of perfectly estimable functions depends on the fact that $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, so that $\boldsymbol{Q}'\boldsymbol{X} \neq \boldsymbol{0}$. Actually, the contrapositive is more obvious, if $\boldsymbol{Q}'\boldsymbol{X} = \boldsymbol{0}$ then

$$C(\boldsymbol{X}) \subset C(\boldsymbol{Q})^{\perp} = \left[C(\boldsymbol{V})^{\perp}\right]^{\perp} = C(\boldsymbol{V}).$$

Because $\boldsymbol{Q}'\boldsymbol{X} \neq \boldsymbol{0}$, the estimable function $\boldsymbol{Q}'\boldsymbol{X}\boldsymbol{\beta}$ nontrivial.

Since $E[\boldsymbol{Q}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})] = \boldsymbol{0}$ and $Cov[\boldsymbol{Q}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})] = \boldsymbol{0}$, I have

$$\Pr[\boldsymbol{Q}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}] = 1 \quad \text{and} \quad \boldsymbol{Q}'\boldsymbol{Y} = \boldsymbol{Q}'\boldsymbol{X}\boldsymbol{\beta}, \text{a.s.}$$

Therefore, whenever $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, there exist nontrivial estimable functions of $\boldsymbol{\beta}$ that can be perfectly estimated. Moreover, with $Cov(\boldsymbol{Q}'\boldsymbol{Y}) = \sigma^2\boldsymbol{Q}'\boldsymbol{V}\boldsymbol{Q} = \boldsymbol{0}$, my choice of $\boldsymbol{Q}$ with full column rank implies that among linear functions of $\boldsymbol{Y}$ only linear functions of $\boldsymbol{Q}'\boldsymbol{Y}$ will have 0 covariance matrices, so only linear functions of $\boldsymbol{Q}'\boldsymbol{X}\boldsymbol{\beta}$ will be estimated perfectly.

My approach differs from the traditional approach that obtains estimates and tests by replacing the covariance matrix $\boldsymbol{V}$ with a pseudo-covariance matrix $\boldsymbol{T}$ such that $C(\boldsymbol{X}) \subset C(\boldsymbol{T})$. I appeal to what I think is a more intuitive method based on adjusting $\boldsymbol{X}$. After adjusting for the part of the analysis that is known perfectly, the remainder of the analysis is performed with the methods that apply when $C(\boldsymbol{X}) \subset C(\boldsymbol{V})$. My procedures provide an alternative way of dealing with difficult aspects of models when $\boldsymbol{V}$ is singular.

4.3 Partly Perfect Estimation

Puntanen and Scott (1996) show that the BLUE $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ has $Cov(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ if and only if $\boldsymbol{X}\boldsymbol{\beta}$ is known perfectly if and only if $C(\boldsymbol{X}) \cap C(\boldsymbol{V}) = \{\boldsymbol{0}\}$. Note that for nontrivial $\boldsymbol{X}$ and $\boldsymbol{V}$, $C(\boldsymbol{X}) \cap C(\boldsymbol{V}) = \{\boldsymbol{0}\}$ implies $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$.

When $C(\boldsymbol{X}) \cap C(\boldsymbol{V}) \neq \{\mathbf{0}\}$ and $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, $\boldsymbol{X\beta}$ cannot be known perfectly, but some functions of $\boldsymbol{X\beta}$ can be estimated perfectly. If this happens, write $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1$ with $\boldsymbol{\beta}_0 \in C(\boldsymbol{X'Q})$ and $\boldsymbol{\beta}_1 \perp C(\boldsymbol{X'Q})$. I show that $\boldsymbol{X\beta}_0$ is known, so that I need only estimate $\boldsymbol{X\beta}_1$ to learn everything about $\boldsymbol{X\beta}$. Since $C(\boldsymbol{X'Q}) = C(\boldsymbol{X'QQ'}) = C([\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{M_V})])$, the decomposition does not really depend on the choice of $\boldsymbol{Q}$.

In fact, $\boldsymbol{\beta}_0$ is known, not just $\boldsymbol{X\beta}_0$. By the definition of $\boldsymbol{\beta}_0$ as part of a unique orthogonal decomposition, with probability one,

$$\boldsymbol{\beta}_0 = \boldsymbol{M}_{\boldsymbol{X'Q}}\boldsymbol{\beta} = \boldsymbol{X'Q}[\boldsymbol{Q'XX'Q}]^{-}\boldsymbol{Q'X\beta} = \boldsymbol{X'Q}[\boldsymbol{Q'XX'Q}]^{-}\boldsymbol{Q'Y}.$$

Note that $\boldsymbol{\beta}_0$ is always identifiable (estimable) even though $\boldsymbol{\beta}$ may not be. Since the perpendicular projection operator does not depend on the choice of generalized inverse, neither does $\boldsymbol{\beta}_0$, and with probability one, it does not depend on $\boldsymbol{Q}$.

Let $\boldsymbol{X}_v$ satisfy $C(\boldsymbol{X}_v) = C(\boldsymbol{X}) \cap C(\boldsymbol{V})$. I now estimate $\boldsymbol{X\beta}_1$. Notice that $\boldsymbol{\beta}_1 \perp C(\boldsymbol{X'Q}) \Leftrightarrow \boldsymbol{Q'X\beta}_1 = 0 \Leftrightarrow \boldsymbol{X\beta}_1 \perp C(\boldsymbol{Q}) \Leftrightarrow \boldsymbol{X\beta}_1 \in C(\boldsymbol{V}) \Leftrightarrow \boldsymbol{X\beta}_1 \in C(\boldsymbol{X}_v) \Leftrightarrow \boldsymbol{X\beta}_1 = \boldsymbol{X}_v\boldsymbol{\gamma}$ for some $\boldsymbol{\gamma}$. Since $\boldsymbol{X\beta}_0$ is fixed and known, it follows that $E(\boldsymbol{Y} - \boldsymbol{X\beta}_0) = \boldsymbol{X\beta}_1 \in C(\boldsymbol{X}_v)$ and $Cov(\boldsymbol{Y} - \boldsymbol{X\beta}_0) = \sigma^2\boldsymbol{V}$, so I can estimate $\boldsymbol{X\beta}_1$ by fitting model (1.5), i.e.,

$$\boldsymbol{Y} - \boldsymbol{X\beta}_0 = \boldsymbol{X}_v\boldsymbol{\gamma} + \boldsymbol{e}, \quad E(\boldsymbol{e}) = 0, \quad Cov(\boldsymbol{e}) = \sigma^2\boldsymbol{V}.$$

The generalized least squares estimate and BLUE of $\boldsymbol{X}_v\boldsymbol{\gamma}$ is

$$\boldsymbol{X}_v\hat{\boldsymbol{\gamma}} = \boldsymbol{X}_v(\boldsymbol{X}_v'\boldsymbol{V}^{-}\boldsymbol{X}_v)^{-}\boldsymbol{X}_v'\boldsymbol{V}^{-}(\boldsymbol{Y} - \boldsymbol{X\beta}_0).$$

Under normality, tests for model (1.5) are also relatively easy to construct and all the nice properties of linear models like (1.3) under the condition $C(\boldsymbol{X}) \subset C(\boldsymbol{V})$ apply to model (1.5).

Intuitively, $\boldsymbol{P}'(\boldsymbol{X\beta}_0 + \boldsymbol{X}_v\hat{\boldsymbol{\gamma}})$ is a reasonable way to estimate an arbitrary estimable function $\boldsymbol{P'X\beta}$ in model (1.1). The following theorem, proven in the Appendix B, establishes that this is also the BLUE.

Theorem 4 *Every representation of a BLUE in model (1.3) for $\boldsymbol{\rho}'\boldsymbol{X\beta}$ uniquely determines a BLUE for $\boldsymbol{\rho}'\boldsymbol{X}_v\boldsymbol{\gamma}$ in model (1.5) and vice versa.*

Although the definition of $\boldsymbol{X}_v$ is simple, it is worth noting ways to find it. In some low-dimensional examples, finding $\boldsymbol{X}_v$ from $\boldsymbol{X}$ and $\boldsymbol{V}$ is fairly easy. Otherwise, $C(\boldsymbol{X}_v) = C(\boldsymbol{X}\tilde{\boldsymbol{U}})$, where $C(\tilde{\boldsymbol{U}}) = C[\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{M_V})]^\perp \equiv C(\boldsymbol{X}'\boldsymbol{Q})^\perp$ because $\boldsymbol{\beta}_1 \in C(\boldsymbol{X}'\boldsymbol{Q})^\perp = C(\tilde{\boldsymbol{U}}) \Leftrightarrow \boldsymbol{X\beta}_1 \in C(\boldsymbol{X}\tilde{\boldsymbol{U}})$. One direction is immediate. The other follows from writing $\boldsymbol{X\beta}_1 = \boldsymbol{X}\tilde{\boldsymbol{U}}\boldsymbol{\gamma}$ and premultiplying by $\boldsymbol{Q}'$, see also Rao and Mitra (1971, p.118). Note that $r(\boldsymbol{X}_v) = r(\boldsymbol{X}) + r(\boldsymbol{V}) - r(\boldsymbol{X}, \boldsymbol{V})$. Additionally, Christensen (2002 Section 10.4), establishes that a basis for $\boldsymbol{X}_v$ is any basis for the orthocomplement of $C(\boldsymbol{I} - \boldsymbol{M_X}, \boldsymbol{I} - \boldsymbol{M_V})$, i.e., $C(\boldsymbol{X}_v) = C(\boldsymbol{I} - \boldsymbol{M_X}, \boldsymbol{I} - \boldsymbol{M_V})^\perp$.

EXAMPLE 4.1. Consider a three sample model $y_i = \mu_i + \varepsilon_i$, $i = 1, 2, 3$ with correlated observations. The correlation between the first and the third observations is -1. The second observation is uncorrelated with the other two. The key matrices for model (1.3) are

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

Clearly $C(\boldsymbol{V})$ is a singular matrix and $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, so I can apply my method.

$$C(\boldsymbol{X}_v) = C\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix}\right)$$

and it is easy to find $\boldsymbol{Q}$ and $C(\boldsymbol{Q}'\boldsymbol{X})$ as

$$\boldsymbol{Q} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \qquad C(\boldsymbol{X}'\boldsymbol{Q}) = C\left(\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right).$$

It follows that with $\boldsymbol{\beta} = [\mu_1, \mu_2, \mu_3]'$,

$$\boldsymbol{\beta}_0 = \begin{bmatrix} \frac{\mu_1+\mu_3}{2} \\ 0 \\ \frac{\mu_1+\mu_3}{2} \end{bmatrix}, \qquad \boldsymbol{\beta}_1 = \begin{bmatrix} \frac{\mu_1-\mu_3}{2} \\ \mu_2 \\ \frac{\mu_3-\mu_1}{2} \end{bmatrix},$$

so

$$\boldsymbol{X}\boldsymbol{\beta}_0 = \begin{bmatrix} \frac{\mu_1+\mu_3}{2} \\ 0 \\ \frac{\mu_1+\mu_3}{2} \end{bmatrix} = \begin{bmatrix} \frac{y_1+y_3}{2} \\ 0 \\ \frac{y_1+y_3}{2} \end{bmatrix} \quad \text{a.s.,}$$

and model (1.5) becomes

$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 = \begin{bmatrix} \frac{y_1-y_3}{2} \\ y_2 \\ \frac{y_3-y_1}{2} \end{bmatrix} = \boldsymbol{X}_v\boldsymbol{\gamma} + \boldsymbol{e} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \boldsymbol{e}.$$

where

$$\gamma_1 \equiv \frac{\mu_1 - \mu_3}{2}, \quad \gamma_2 \equiv \mu_2.$$

Notice that now $C(\boldsymbol{X}_v) \subset C(\boldsymbol{V})$, so I can apply standard methods, e.g., Christensen (2002, p. 233), to get the BLUE of γ_1, γ_2 :

$$\hat{\gamma_1} = \frac{y_1 - y_3}{2}, \quad \hat{\gamma_2} = y_2.$$

According to Theorem 4, $(y_1 - y_3)/2$ and y_2 are, respectively, the BLUE of $(\mu_1 - \mu_3)/2$ and μ_2 in both the original model (1.3) and the adjusted model (1.5). Also $\boldsymbol{X}\hat{\boldsymbol{\beta}} \equiv \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{X}_v\hat{\boldsymbol{\gamma}} = [y_1\ y_2\ y_3]'$. Actually, in this example $C(\boldsymbol{V}\boldsymbol{X}) \subset C(\boldsymbol{X})$ so the least square estimate of $\boldsymbol{\mu}$ is the BLUE in the original model (1.3) which also implies $\hat{\mu_i} = y_i$. It is interesting that $\mu_1 + \mu_3$ can be estimated perfectly.

4.4 Testing a Linear Hypothesis

In model (1.3) with multivariate normal errors, when $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$ I can construct hypothesis tests using model (1.5). If $\boldsymbol{X\beta}$ is known, there is nothing left to test, so I restrict attention to general Gauss-Markov models that satisfy $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$ and $C(\boldsymbol{X}) \cap C(\boldsymbol{V}) \neq \{\mathbf{0}\}$.

Consider testing a vector estimable linear hypothesis $H_0 : \boldsymbol{P'X\beta} = \boldsymbol{d}$, under

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim N(\mathbf{0}, \sigma^2\boldsymbol{V}).$$

A standard way to test H_0 involves finding the corresponding reduced model which is

$$\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}} = \boldsymbol{X}_0\boldsymbol{\eta} + \boldsymbol{e}, \quad \boldsymbol{e} \sim N(\mathbf{0}, \sigma^2\boldsymbol{V}), \tag{4.1}$$

where $C(\boldsymbol{X}_0) \subset C(\boldsymbol{X})$ and $\boldsymbol{P'X}\tilde{\boldsymbol{b}} = \boldsymbol{d}$ for known $\tilde{\boldsymbol{b}}$, c.f. Christensen (2002, Section 3.3). Here I need to check $\boldsymbol{d} \in C(\boldsymbol{P'X})$ or the hypothesis makes no sense. I now go a step further to find the reduced model relative to model (1.5) that is determined by $\boldsymbol{P'X\beta} = \boldsymbol{d}$.

Under the conditions $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$ and $C(\boldsymbol{X}) \cap C(\boldsymbol{V}) \neq \{\mathbf{0}\}$, I adjust the model for forms that can be estimated perfectly. As in Section 4.2, write $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1$ with $\boldsymbol{\beta}_0 \in C(\boldsymbol{X'Q})$ and $\boldsymbol{\beta}_1 \perp C(\boldsymbol{X'Q})$, where $\boldsymbol{\beta}_0$ is known with probability 1. The null hypothesis can be rewritten $H_0 : \boldsymbol{P'X}(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1) = \boldsymbol{d}$ or $H_0 : \boldsymbol{P'X\beta}_1 = \boldsymbol{d} - \boldsymbol{P'X\beta}_0$. If $\boldsymbol{d} - \boldsymbol{P'X\beta}_0 \notin C(\boldsymbol{P'X}_v)$ then $\boldsymbol{P'X\beta} = \boldsymbol{d}$ must be false. Let $\boldsymbol{d} - \boldsymbol{P'X\beta}_0 \equiv \boldsymbol{k}$, so the null hypothesis is $H_0 : \boldsymbol{P'X\beta}_1 = \boldsymbol{k}$. Since $\boldsymbol{X\beta}_1 \equiv \boldsymbol{X}_v\gamma$, I have $\boldsymbol{P'X\beta}_1 \equiv \boldsymbol{P'X}_v\gamma$ and the null hypothesis becomes $\boldsymbol{P'X}_v\gamma = \boldsymbol{k}$ in model (1.5) which is clearly still estimable. The full model (1.5) reduces to

$$(\boldsymbol{Y} - \boldsymbol{X\beta}_0) - \boldsymbol{X}_v\boldsymbol{b} = \boldsymbol{W}\gamma_0 + \boldsymbol{e}, \tag{4.2}$$

where $\boldsymbol{b}$ is any solution to $\boldsymbol{P'X_v b} = \boldsymbol{k}$ and $C(\boldsymbol{W}) \subset C(\tilde{\boldsymbol{V}})$, specifically, $C(\boldsymbol{W}) = C(\boldsymbol{X}_v\boldsymbol{U})$ where $C(\boldsymbol{U}) = C(\boldsymbol{X}_v'\boldsymbol{P})^{\perp}$ as in Christensen (2002, p.233). With

$$\boldsymbol{A} = \boldsymbol{X}_v(\boldsymbol{X}_v'\boldsymbol{V}^-\boldsymbol{X}_v)^-\boldsymbol{X}_v'V^-, \quad \boldsymbol{A}_0 = \boldsymbol{W(W'V^-W)^-W'V^-},$$

I know that the BLUE of $\boldsymbol{X}_v\boldsymbol{\gamma}$ is $\boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)$ and the BLUE of $\boldsymbol{W}\boldsymbol{\gamma}_0$ is $\boldsymbol{A}_0(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0-\boldsymbol{X}_v\boldsymbol{b})$.

To test the reduced model assume multivariate normal errors and use the standard method of comparing sums of squared error (SSE). Define SSE and SSE_0 for models (1.5) and (4.2) as

$$\begin{aligned} SSE &= (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0)'(\boldsymbol{I}-\boldsymbol{A})'\boldsymbol{V}^-(\boldsymbol{I}-\boldsymbol{A})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0) \\ &= (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0-\boldsymbol{X}_v\boldsymbol{b})'(\boldsymbol{I}-\boldsymbol{A})'\boldsymbol{V}^-(\boldsymbol{I}-\boldsymbol{A})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0-\boldsymbol{X}_v\boldsymbol{b}) \end{aligned}$$

and

$$SSE_0 = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0-\boldsymbol{X}_v\boldsymbol{b})'(\boldsymbol{I}-\boldsymbol{A}_0)'\boldsymbol{V}^-(\boldsymbol{I}-\boldsymbol{A}_0)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0-\boldsymbol{X}_v b)$$

and denote

$$\boldsymbol{C} \equiv (\boldsymbol{I}-\boldsymbol{A})'\boldsymbol{V}^-(\boldsymbol{I}-\boldsymbol{A}); \qquad \boldsymbol{C_0} \equiv (\boldsymbol{I}-\boldsymbol{A}_0)'\boldsymbol{V}^-(\boldsymbol{I}-\boldsymbol{A}_0).$$

Under model (1.5), the usual test statistic has a noncentral F distribution,

$$F^* = \frac{(SSE_0-SSE)/tr[(\boldsymbol{C}_0-\boldsymbol{C})\boldsymbol{V}]}{SSE/tr(\boldsymbol{C}\boldsymbol{V})}, \tag{4.3}$$

and

$$F^* \sim F(tr[(\boldsymbol{C}_0-\boldsymbol{C})\boldsymbol{V}], tr(\boldsymbol{C}\boldsymbol{V}), \boldsymbol{\gamma}'\boldsymbol{X}_v'(\boldsymbol{C}_0-\boldsymbol{C})\boldsymbol{X}_v\boldsymbol{\gamma}/2\sigma^2)$$

whereas under H_0: $\boldsymbol{\gamma}'\boldsymbol{X}_v'(\boldsymbol{C}_0-\boldsymbol{C})\boldsymbol{X}_v\boldsymbol{\gamma} = \boldsymbol{0}$, a central F distribution is appropriate. If $\boldsymbol{Y} \notin C(\boldsymbol{X},\boldsymbol{V})$, the full model is wrong. In testing the null model, I also reject if $\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_0 \notin C(\boldsymbol{V})$. Finally, I establish that F^* can be computed directly from the model (1.5) BLUEs.

Theorem 5 *Under models (1.3) and (1.5) the distribution in (4.3) holds. Under H_0 : $\boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{d}$ or model (4.2),*

$$F^* = \frac{(SSE_0-SSE)/tr[(\boldsymbol{C}_0-\boldsymbol{C})\boldsymbol{V}]}{SSE/tr(\boldsymbol{C}\boldsymbol{V})} \sim F(tr[(\boldsymbol{C}_0-\boldsymbol{C})\boldsymbol{V}], tr(\boldsymbol{C}\boldsymbol{V}), 0)$$

which reduces to

$$F^* = \frac{(\boldsymbol{P}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[\boldsymbol{P}'\boldsymbol{X}_v(\boldsymbol{X}_v'\boldsymbol{V}^-\boldsymbol{X}_v)^-\boldsymbol{X}_v'\boldsymbol{P}]^-(\boldsymbol{P}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{d})/r(\boldsymbol{X}_v'\boldsymbol{P})}{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\hat{\gamma})'\boldsymbol{V}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\hat{\gamma})/[r(\boldsymbol{V}) - r(\boldsymbol{X}_v)]} \sim F(r(\boldsymbol{X}_v'\boldsymbol{P}), r(\boldsymbol{V}) - r(\boldsymbol{X}_v), 0).$$

PROOF: See Appendix B.

This testing method is equivalent to the method of constructing test statistics by replacing $\boldsymbol{V}$ with a pseudo-covariance matrix $\boldsymbol{T} = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{U}\boldsymbol{X}'$, where $C(\boldsymbol{X}) \subset C(\boldsymbol{T})$, cf. Christensen (2002, Section 10.3) and the Appendix B.

EXAMPLE 4.1. This is a two sample problem with the first two observations having variance 1 but the third has variance 0. The key matrices are

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

With $\boldsymbol{\beta} = [\mu_1, \mu_2]'$ and

$$\boldsymbol{Q}'\boldsymbol{X} = \begin{bmatrix} 0 & 1 \end{bmatrix},$$

let us consider an hypothesis test of $H_0 : \boldsymbol{\rho}'\boldsymbol{X}\boldsymbol{\beta} = d$, where $\boldsymbol{\rho}' = [\frac{1}{2}, \frac{1}{2}, -1]$. The full model (1.5) is

$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 = \begin{bmatrix} y_1 \\ y_2 \\ y_3 - y_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ 0 \end{bmatrix} = \boldsymbol{X}_v\boldsymbol{\gamma} + \boldsymbol{e} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\boldsymbol{\gamma} + \boldsymbol{e},$$

where $\boldsymbol{X}\boldsymbol{\beta}_1 = \boldsymbol{X}_v\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} \equiv \boldsymbol{\mu}_1$. The null hypothesis can be rewritten as $H_0 : \boldsymbol{\rho}'\boldsymbol{X}_v\boldsymbol{\gamma} = k$. Here $k = d - \boldsymbol{\rho}'\boldsymbol{X}\boldsymbol{\beta}_0 = d + y_3$. The reduced model (4.2) now becomes

$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b} = \boldsymbol{W}\boldsymbol{\gamma}_0 + \boldsymbol{e},$$

where

$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 = \begin{bmatrix} y_1 \\ y_2 \\ 0 \end{bmatrix}, \quad b = k, \quad \boldsymbol{W} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

After some algebra

$$\boldsymbol{A} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{A}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{C}_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

With $\boldsymbol{K} = [k, k, 0]'$,

$$SSE = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{K})'\boldsymbol{C}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{K}) = \frac{1}{2}(y_1 - y_2)^2$$

$$SSE_0 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{K})'\boldsymbol{C}_0(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{K}) = (y_1 - k)^2 + (y_2 - k)^2,$$

and

$$tr[(\boldsymbol{C}_0 - \boldsymbol{C})\boldsymbol{V}] = 1, \quad tr(\boldsymbol{C}\boldsymbol{V}) = 1.$$

Thus,

$$\begin{aligned} F^* &= \frac{(SSE_0 - SSE)/tr((\boldsymbol{C}_0 - \boldsymbol{C})\boldsymbol{V})}{SSE/tr(\boldsymbol{C}\boldsymbol{V})} = \frac{(y_1 - k)^2 + (y_2 - k)^2 - (y_1 - y_2)^2/2}{(y_1 - y_2)^2/2} \\ &= \frac{(y_1 + y_2 - 2k)^2/2}{(y_1 - y_2)^2/2}. \end{aligned}$$

Recall $k = d - \boldsymbol{\rho}'\boldsymbol{X}\boldsymbol{\beta}_0 = d + y_3$, so $F^* = (y_1 + y_2 - 2d - 2y_3)^2/(y_1 - y_2)^2$. If F^* is greater than the $1 - \alpha$ percentile of $F(1, 1, 0)$, H_0 is rejected at level α.

Alternatively, I can compute the test statistic using the BLUEs. It is easy to check

$$\left[\boldsymbol{\rho}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{d}\right] = \left(\frac{y_1 + y_2}{2} - y_3 - d\right) = \left(\frac{y_1 + y_2}{2} - k\right)$$

and

$$\left[\boldsymbol{\rho}'\boldsymbol{X}_v(\boldsymbol{X}_v'\boldsymbol{V}^-\boldsymbol{X}_v)^-\boldsymbol{X}_v'\rho\right]^- = \left[1 \cdot \frac{1}{2} \cdot 1\right]^- = 2,$$

so the numerator of F^* equals $(y_1 + y_2 - 2k)^2/2r(\boldsymbol{X}_v'\boldsymbol{\rho})$. Clearly the degrees of freedom involve $r(\boldsymbol{X}_v'\boldsymbol{\rho}) = 1, r(\boldsymbol{V}) = 2, r(\boldsymbol{X}_v) = 1$, which agree with the previous method. Thence the results of the two methods agree.

4.5 Summary

In general Gauss-Markov models like (1.3) with $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$, there exist estimable functions of $\boldsymbol{\beta}$ that are known with probability 1. Specifically these are linear functions of $\boldsymbol{Q'X\beta}$ where $\boldsymbol{Q}$ is a full column rank matrix with $C(\boldsymbol{Q}) = C(\boldsymbol{V})^{\perp}$. Many traditional methods for handling these problems adjust $\boldsymbol{V}$ by finding a matrix $\boldsymbol{T}$ to act as a pseudo-covariance matrix with $C(\boldsymbol{X}) \subset C(\boldsymbol{T})$ that gives BLUEs and tests that are appropriate for $\boldsymbol{V}$. Contrary to traditional methods of adjusting $\boldsymbol{V}$, I decompose $\boldsymbol{\beta}$ into known and unknown parts and adjust $\boldsymbol{X}$ to allow estimation and testing of the unknown part of $\boldsymbol{\beta}$. Specifically, I adjust model (1.3), $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, to get an equivalent model (1.5), $\boldsymbol{Y} - \boldsymbol{X\beta}_0 = \boldsymbol{X}_v\boldsymbol{\gamma} + \boldsymbol{e}$, where $\boldsymbol{X\beta}_0$ is a known vector, then perform estimation and tests on model (1.5). In the Appendix B I show the equivalence of model (1.5) and model (1.3) for estimation and testing.

Chapter 5

Future Work

I have two major directions of interests following this dissertation. My first direction is that tests studied in this dissertation can be extended into generalized linear models. There are three approaches. First approach is to extend my first proposed test T_{1n} directly to generalized linear models. The proof of asymptotic distribution is promising, but careful studies of small sample adjustments are required. The second approach is similar to F-H's method. I consider proposing test based on Fourier transformation of standardized residuals in generalized linear models. The third approach is to use a partial sum of the components in score test based on smooth test. The second direction is that the bootstrap method that S-W used to evaluate P values brings my attention to the questions that when and how bootstrap methods can be safely used.

A brief discussion of each future direction is presented in the following Sections.

5.1 Lack-of-fit test using generalized linear model residuals

The beauty of linear model lies in its simplicity in estimation and convenience of checking models using residuals. However, in generalized linear models, the parameters are usually estimated using likelihood, quasi-likelihood or by solving other estimating equations. The residuals are not prevalent used for model checking in generalized linear models. According to Pierce and Schafer (1986), they are not even clearly defined.

Pierce and Schafer (1986) discussed the appropriateness of using deviance-based residuals to provide insights of the model fittings such as examining effects of potential new covariates, or detecting nonlinear functions of existing covariates. Let $y_1, ..., y_n$ be a sample of independent response variables each with density $f(y_i, \theta_i)$ and link function $g(\cdot)$ such that $\theta_i = g(\boldsymbol{x}_i'\boldsymbol{\beta})$. Here $\boldsymbol{x}_i$ is a vector of p covariates associated with ith response y_i and $\boldsymbol{\beta}$ is a vector of p unknown coefficients. Three types of residuals approximately normal distributed are considered. They are linear residuals, transformed linear residuals and deviance contributions. Here, linear residuals standardize response variable with estimated mean and standard deviation and for ith observation, it is

$$r_i^l(y_i, \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) = \{y_i - \hat{\mu}(y_i)\}/\hat{\sigma}(y_i),$$

where $\mu(y_i) = E(y_i)$, and $\sigma(y_i)$ is the standard deviation of y_i. Transformed residuals first transform the response variables and then standardize them, that is

$$r_i^t(y_i, \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) = \{t(y_i) - \hat{\mu}(t(y_i))\}/\hat{\sigma}(t(y_i)),$$

where $t(\cdot)$ is a transformation function which is used to correct the skewness of y_is, so that residuals $r^t(y_i, \theta_i)$s converge faster. McCullaph and Nelder (1983) called them Anscombe residuals. The third, deviance contribution,

$$r_i^D(y_i, \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) = \mathrm{sgn}(y_i - \hat{\theta}_i)\{2d(y_i, \hat{\theta}_i)\}^{1/2},$$

where $\hat{\theta}_i$ is the maximum likelihood estimate of θ_i, and $d(y_i, \hat{\theta}_i)$ is the unite deviance. McCullagh and Nelder (1983) further adjusted deviance residuals to standard normal distribution. Specifically, that is

$$r_i^{AD}(y_i, \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) = r^D(y_i, \hat{\theta}_i) + \rho_3(\hat{\theta}_i)/6,$$

where $\rho_3(\theta_i) = E_{\theta_i}\{[(y_i - \hat{\mu}(y_i))/\hat{\sigma}(y_i)]^3\}$. The main idea is to apply discrete Fourier transformation to Anscombe residuals and adjusted deviance residuals. In this way, the transformed residuals compress lack-of-fit signals into lower frequencies, so that the proposed tests can perform well.

Denotes the normalized residuals (either r_i^{AD} or r_i^t) in a generalized linear model as $\tilde{e}_i$s for $i = 1, ..., n$. Applying Fourier transformation, we have

$$\hat{e}^*_{2j-1} = (2/n)^{1/2} \sum_{i=1}^{n} \cos(2\pi ij/n)\tilde{e}_i,$$

$$\hat{e}^*_{2j-1} = (2/n)^{1/2} \sum_{i=1}^{n} \sin(2\pi ij/n)\tilde{e}_i,$$

for $j = 1, ..., [n/2]$. My proposed test statistic is similar in form of Fan and Huang's statistic, that is

$$T_{f1} = \max_{1 \le m \le \tilde{n}} \frac{1}{\sqrt{2m\hat{\sigma}_e}^4} \sum_{i=1}^{m} (\hat{e}^{*2}_i - \hat{\sigma}_e^2).$$

where $\tilde{n} \to \infty$ as $n \to \infty$ and $\tilde{n}/n \to 0$. Note that here $\hat{\sigma}_e$ are standard deviation of the transformed residuals. After standardization and Fourier transformation, residuals $\hat{e}^*_i$s have mean of 0 and standard deviation 1. Hence, I might just replace $\hat{\sigma}_e$ with 1, but to make the test robustly converge I will also consider an appropriate estimate of it. After suitable normalization, I expect the statistic to have an extreme value distribution. However, a new set of proofs for asymptotic convergence need to be developed and $\tilde{e}_i$ can not be treated as if it is from standard normal distribution. For example, the asymptotical normality approximations for residuals in generalized linear models require two type asymptotic situations, the sample size $n \to \infty$ and an index $m \to \infty$. Here, the m could be the

the number of trials in a Binomial distribution, or Poisson mean, Gamma distribution's shape parameter. Because of the requirement for m convergence, I will first consider Binomial underlining distribution with fairly a large number of trials, say $m > 10$ and Poisson distribution with a fairly large mean. (Although the convergence requires $m \to \infty$, in application, a moderate m is enough.) In the end, I may even apply it to Bernoulli distribution, since in theory, to prove covergence, I only need the transformed residuals to be asymptotically independent, identical distributed with mean 0 and variance 1.

Besides convergence, to make test T_{f1} efficient, an appropriate ordering method needs to be proposed. Using the key matrices in linear model (2.8), F-H's test involves fitting model

$$\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{M_X})\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{H}$ is orthonormal matrix of Fourier series. F-H's test is efficient only if $(\boldsymbol{I} - \boldsymbol{M_X})\boldsymbol{H}$ is smooth. However, in generalized linear model, such clear form is not available. So additional studies on finding efficient ordering methods are required.

Finally, small sample performance of the proposed tests needs to be studied, especially for the cases with multiple predictors.

5.2 Smooth test using partial sums

Neyman's smooth test of density function is originally proposed in Neyman (1937) and further developed by many people over the years, for example Barton (1953, 1955, 1956); Hamdan (1962, 1963, 1964); Thomas and Pierce (1979); Bargal and Thomas (1983); Rayner and Best (1986, 1988, 1989). The ones that are close to my new ideas are by Rayner and Best who frequently adopt orthonormal functions to make components of score test identifiable and independent.

Denotes y_is for $i = 1, ..., n$ as independent random variables from distribution

$$f(y_i;\theta_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} - c(y_i, \phi_i)\right],$$

with unknown θ_i and known nuisance parameter ϕ_i. $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. This is a typical density function of exponential family. The mean of y_i, $\mu_i = \mu(y_i)$, is a function of θ_i which is linked to covariates by function $g(\cdot)$. That is

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}, \tag{5.1}$$

where $\boldsymbol{x}_i$ is a vector of p covariates associated with y_i and $\boldsymbol{\beta}$ is a vector of p unknown parameters. Using the idea of smooth test, this distribution function is embedded within an alternative density function

$$f_k(y_i, \boldsymbol{\eta}, \theta_i)C(\boldsymbol{\eta}, \theta_i)\exp\left\{\sum_{j=1}^{k}\eta_j h_j(y_i;\theta_i)\right\} f(y_i;\theta_i),$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_k)'$ is a vector of k parameters. $C(\boldsymbol{\eta}, \theta_i)$ is a normalizing constant and $h_j(y_i;\theta_i)$ is a set of orthonormal functions on $f(y_i;\theta_i)$, satisfying

$$\int_{-\infty}^{\infty} h_p(y_i, \theta_i)h_q(y_i, \theta_i)f(y_i, \theta_i)dy_i = \delta_{pq},$$

where $\delta_{pq} = 1$ if $p = q$ and 0 otherwise. Hypothesis of lack-of-fit test is

$$H_0 : \boldsymbol{\eta} = \mathbf{0}, \quad \text{against} \quad H_a : \boldsymbol{\eta} \neq \mathbf{0}.$$

To test it, one can use score test, which is based on the statistic

$$\hat{S} = \sum_{j=1}^{k}\hat{V}_j^2 \quad \text{with} \quad \hat{V}_j = \sum_{i=1}^{n}\frac{1}{\sqrt{n}}h_j(y_i, \hat{\theta}_i),$$

where $\hat{\theta}_i$ is the maximum likelihood estimate of θ_i. Recall that θ_is are linked with $\boldsymbol{\beta}$ by $g(\cdot)$, so the MLE of θ_i is estimated by plugging in the MLE of $\boldsymbol{\beta}$. The $\hat{V}_j$s for $j = 1, ..., k$ are components of $\hat{S}_k$. Intuitively large value of $\hat{S}_k$ provides evidence against the null hypothesis. For cases without nuisance parameter, as $n \to \infty$, $\hat{V}_j^2$ converges to $\chi^2_{(1)}$ and S_k converge to $\chi^2_{(k)}$.

The key point is to select k. Simulation studies show that for different choices of ks the score test may lead to different results. Rayner (1986) suggested performing multiple times of tests using different ks. However, this procedure involves multiple looks of the data without considering adjustments to significant levels. This may lead to inconsistent size and making the test less objective than it could be.

My ideas of resolving this issue are in two folds. Frist, using Rayner and Best (1986, 2009)'s (orthogonal polynomial function) $h_j(\cdot)s$, I can proposed test statistic based on partial sums of the normalized (with mean 0 and standard deviation 1) $\hat{V}_j$ or $\hat{V}_j^2$. Denote the mean and standard deviation of $\hat{V}_j$ as $\mu(\hat{V}_j)$ and $\sigma(\hat{V}_i)$. Similarly, for $\hat{V}_j^2$ they are $\mu(\hat{V}_j^2)$ and $\sigma(\hat{V}_j^2)$. My proposed test statistics could be

$$T_{f2} = \max_{1 \le k \le \tilde{n}} \frac{1}{\sqrt{k}} \sum_{j=1}^{k} (\hat{V}_j - \hat{\mu}(\hat{V}_j))/\sigma(\hat{V}_j)$$

and

$$T_{f3} = \max_{1 \le k \le \tilde{n}} \frac{1}{\sqrt{k}} \sum_{j=1}^{k} (\hat{V}_j^2 - \hat{\mu}(\hat{V}_j^2))/(\hat{\sigma}\hat{V}_j^2),$$

This is a promising approach as orthogonal functions of lower order have priority to be included in the partial sum and given higher weights. A power comparison between the two tests and Rayner and Best (1986, 2009)'s test using different ks will be performed.

An alternative approach is to use a different set of orthogonal functions, for example, Fourier series. However, difficulties exist in formulating such orthogonal functions. I will try to resolve it in a following study.

Regards to the difficulties of finding such orthonormal functions $h_j(\cdot)$s, I consider proposing an alternative test for testing the mean function when the link function is assumed correct. Formulate the link function (5.1) in matrices form

$$\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\beta}, \tag{5.2}$$

where

$$\boldsymbol{G} = \begin{bmatrix} g(y_1, \theta_1) \\ \vdots \\ g(y_n, \theta_n) \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}.$$

The the smooth alternatives is obtained by extending (5.2) to

$$\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{H}_k\boldsymbol{\gamma}_k, \tag{5.3}$$

where $\boldsymbol{H}_k$ is a $n \times k$ matrix of orthonormal Fourier series and $\boldsymbol{\gamma}_k$ is a vector of k unknown parameters. Lack-of-fit test is again transformed into testing

$$\boldsymbol{\gamma}_k = \mathbf{0} \quad \text{and} \quad \boldsymbol{\gamma}_k \neq \mathbf{0}.$$

Let $l_0 = l(\hat{\boldsymbol{\beta}})$ be the log-likelihood of model with mean function (5.2), and $l_k = l_k(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}_k})$ be the log-likelihood of model with mean function (5.3). For a fixed k, the above hypothesis can be tested using likelihood ratio method. The test statistic is

$$\tilde{\chi}_k^2 = 2(l_k(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}_k}) - l(\hat{\boldsymbol{\beta}})) \xrightarrow{\mathcal{L}} \chi^2_{(k)}.$$

Now, let $w_k = 2(l_k - l_{k-1})$, for $i = 1, ..., \tilde{n}$, where $\tilde{n}$ is function of n satisfying $\tilde{n} < n$. Here w_is for $i = 1, ..., \tilde{n}$ are asymptotically independent with a $\chi^2_{(1)}$ distribution. My proposed test statistic is

$$T_{f4} = \max_{1 \leq k \leq \tilde{n}} \frac{1}{\sqrt{k}} \sum_{i=1}^{k} (w_i - \hat{\mu}(w_i))/\hat{\sigma}(w_i).$$

After suitably normalization, the normalized tests are expected to have an extreme value distribution. I believe this is a natural extension of Christensen and Sun (2010)'s test to generalized linear models.

5.3 Bootstrapping confidence interval

The name of “bootstrap” is first brought into the realm of statistics by Efron (1979) with pointing out ideas of bootstrap methods had already been around for many years. Since then, there are a large amount of literatures focusing on understanding, improving and applying bootstrap methods to specific problems.

With the quick development of computing power and availability of statistical softerware such as R, S-plus, SAS and ect., bootstrap methods are more available to non-statistics researchers for calculating confidence intervals. The most popular types of bootstrap confidence intervals are the studentized bootstrap interval, the bootstrap percentile interval, bias correction bootstrap percentile (BC) and accelerate bootstrap percentile(BCa). However, when people talk about ”bootstrap”, most people think of percentile method and tend to use it without considering possible limitations. There are two situations under which people tend to use bootstrap confidence intervals. First, the study with small sample size. Second, complicate problems.

However, the dilemma is that bootstrap methods are based on large sample approximations. Once it is applied in small sample, the large sample condition is violated. So, when people apply bootstrap in small sample, it is usually more of a faith rather than actually applying it with consideration of the statistical mechanism and the limitation. Also, for complicate problems, especially when finding large sample approximation of the confidence interval is hopeless, bootstrap methods become life saving straws. Many applications simply re-sample the data and repeat the estimation procedures in each bootstrap sample without checking the large sample convergence. Although bootstrap method is robust in many cases, the validity of bootstrap in complicate problems should not be overlooked. Moreover, there are published literatures, for example Rigby (2009), treated bootstrap as sampling with replacement and commented “Where statistical modeling is applied, the CI should be estimated using bootstrap". Hence we feel it is necessary to clarify the concept of bootstrap methods and emphasize their limitations.

In this paper, we will start with a quick review of assumptions and limitations of each of the four prevalent methods. we will study the bootstrap accuracy in samples with sizes such as 5, 10 and 15. A comparison of the performances of the four methods will be provided in small sample. Also, we will reiterate the importance of checking the large sample convergence of the bootstrap methods when they are applied in complicate problems. Counter examples that use bootstrap methods to approximate P values of lack-of-fit test but do not produce asymptotic consistent results will be provided.

Appendix A

Proof of theorems for Chapter 3

A.1 Lemma1

To prove Theorem 1, I need the following lemma. Throughout conditions (a) and (b) of Section 3.1 are assumed

Lemma 1. $\sqrt{n} \parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \parallel / a_n$ is bounded a.s, where $a_n = \sqrt{2 \log \log n}$.

PROOF. To show $\sqrt{n} \parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \parallel / a_n$ is bounded, define $\boldsymbol{A}_n = \boldsymbol{X}'_n \boldsymbol{X}_n / n$. Under condition (a), $\boldsymbol{A}_n \rightarrow \boldsymbol{A}$ as $n \rightarrow \infty$. By Amemiya (1990), for $j \in \{1, 2, ...p\}$, $\lambda_j(n) \rightarrow \lambda_j$ as $n \rightarrow \infty$, where $\lambda_j(n)$ is the jth smallest eigenvalue of $\boldsymbol{A}_n$, and λ_j is the jth smallest eigenvalue of $\boldsymbol{A}$.

Since $\boldsymbol{A}$ is a positive definite matrix, all λ_js are finite positive values. Koval$'$ (2002) corollary 1, showed that if

$$\lim_{n \rightarrow \infty} n\lambda_1(n) = \infty, \quad \limsup_{n \rightarrow \infty} \frac{(n+1)\lambda_1(n+1)}{n\lambda_1(n)} < \infty,$$

$$\text{and} \quad \log\log n\lambda_1(n) \sim \log\log n\lambda_p(n), \quad \text{as} \quad n \to \infty,$$

it follows that

$$\limsup_{n\to\infty} \left(\frac{n\lambda_1(n)}{2\sigma^2 \log\log n\lambda_1(n)} \right)^{1/2} \parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \parallel = 1 \quad \text{with probability } 1,$$

It is easy to check that the first two conditions are met. The third is less apparent. To check the third condition,

$$\log\log n\lambda_1(n) \sim \log\log n\lambda_p(n) \quad \text{iff} \quad \frac{\log\log n\lambda_1(n)}{\log\log n\lambda_p(n)} - 1 \to 0$$

However,

$$\frac{\log\log n\lambda_1(n)}{\log\log n\lambda_p(n)} - 1 = \frac{\log\log n\lambda_1(n) - \log\log n\lambda_p(n)}{\log\log n\lambda_p(n)} = \frac{\log\left[\log n\lambda_1(n)/\log n\lambda_p(n)\right]}{\log\log n\lambda_p(n)}.$$

The last term on the right goes to 0 since the denominator goes to infinity, whereas the numerator goes to 0 because

$$\frac{\log n\lambda_1(n)}{\log n\lambda_p(n)} = \frac{\log n + \log \lambda_1(n)}{\log n + \log \lambda_p(n)} \to 1.$$

The three conditions are satisfied, hence

$$\limsup_{n\to\infty} \left(\frac{n\lambda_1(n)}{2 \log\log n\lambda_1(n)} \right)^{1/2} \parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \parallel = \sigma \quad \text{with probability } 1,$$

and $\sqrt{n} \parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \parallel / a_n$ is bounded a.s. because $\log n \sim \log n\lambda_1(n)$. □

A.2 Lemma 2

Lemma 2. Under assumption (a), $\left|\sum_{i=1}^{u(\tilde{n})} \boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\right| / \sigma\sqrt{\tilde{n}}$ converges in probability to 0 as $n \to \infty$, for any integers $u(\tilde{n}) \in \{1, 2..., \tilde{n}\}$, where $\tilde{n} = \lceil n/(\log\log n)^{1+\delta} \rceil$.

PROOF. By Cauchy- Schwartz,

$$\begin{aligned}
\frac{1}{\sqrt{\tilde{n}}}\left|\sum_{i=1}^{u(\tilde{n})}\frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right| &\leq \frac{1}{\sqrt{\tilde{n}}}\sum_{i=1}^{u(\tilde{n})}\left|\frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right| \\
&\leq \frac{1}{\sqrt{\tilde{n}}\sigma}\sum_{i=1}^{u(\tilde{n})} \parallel \boldsymbol{x}_i \parallel \cdot \parallel \hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta} \parallel \\
&\leq \frac{1}{\sqrt{\tilde{n}}\sigma}\sum_{i=1}^{\tilde{n}} \parallel \boldsymbol{x}_i \parallel \cdot \parallel \hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta} \parallel,
\end{aligned}$$

which is equivolent to

$$\left[\frac{\sqrt{2\tilde{n}}}{\sqrt{n}}\sqrt{\log\log n}\right]\left[\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \parallel \boldsymbol{x}_i \parallel\right]\left[\frac{\sqrt{n}}{\sqrt{2\sigma^2\log\log n}} \parallel \hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta} \parallel\right].$$

By assumption (a), $\sum_{i=1}^{n} \parallel \boldsymbol{x}_i \parallel^2 /n = tr(\boldsymbol{X}\boldsymbol{X}')/n$ which converges to $tr(\boldsymbol{A})$. It follows that $0 \leq \sum_{i=1}^{n} \parallel \boldsymbol{x}_i \parallel /n < \sum_{i=1}^{n}\max(1, \parallel \boldsymbol{x}_i \parallel^2)/n < \sum_{i=1}^{n}(1+ \parallel \boldsymbol{x}_i \parallel^2)/n \rightarrow 1+tr(\boldsymbol{A})$, so the second term converges to a constant. By Lemma 1, the third term $\sqrt{n} \parallel \hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta} \parallel /\sqrt{2\sigma^2\log\log n}$ is bounded. If the first term $\sqrt{2\tilde{n}\log\log n/n}$ converges to 0, as $n \rightarrow \infty$, the whole term converges to 0. If $\tilde{n} = n/(\log\log n)^{1+\delta}$ for $\delta > 0$, $\sqrt{2\tilde{n}\log\log n/n} = \sqrt{2}/\sqrt{(\log\log n)^\delta} \rightarrow 0$, as $n \rightarrow \infty$. □

A.3 Proof of Theorem 1

PROOF OF THEOREM 1. First suppose both $\boldsymbol{\beta}$ and σ are known, then $e_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ for $i \in \{1, ..., n\}$ are independently distributed with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$. By Erdös and Kac (1945), as $\tilde{n} \rightarrow \infty$,

$$\frac{1}{\sqrt{\tilde{n}}}\max_{1\leq m\leq\tilde{n}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right| \xrightarrow{\mathcal{L}} T,$$

where T has a distribution with cdf

$$Pr[T < t] = \frac{4}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m}{2m+1}\exp(-(2m+1)^2\pi^2/8t^2) \quad \text{for} \quad t > 0.$$

Let $\hat{\boldsymbol{\beta}}_n$ be the least square estimator of $\boldsymbol{\beta}$ and let

$$k \in \left\{ j : \left| \sum_{i=1}^{j} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma} \right| = \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma} \right| \right\},$$

so that,

$$\begin{aligned}
\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma} \right| &= \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{k} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma} \right| \\
&= \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{k} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} - \sum_{i=1}^{k} \frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \\
&\le \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{k} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \right| + \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{k} \frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \\
&\le \frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \right| + \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{k} \frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right|.
\end{aligned}$$

Also let

$$q \in \left\{ j : \left| \sum_{i=1}^{j} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \right| = \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \right| \right\}.$$

Then have

$$\begin{aligned}
\frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma} \right| &\ge \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{q} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma} \right| \\
&= \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{q} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} - \sum_{i=1}^{q} \frac{x_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \\
&\ge \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{q} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \right| - \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{q} \frac{x_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \\
&= \frac{1}{\sqrt{\tilde{n}}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \right| - \frac{1}{\sqrt{\tilde{n}}} \left| \sum_{i=1}^{q} \frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right|.
\end{aligned}$$

By Lemma 2, both $\left| \sum_{i=1}^{q} \boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right| / \sigma\sqrt{\tilde{n}}$ and $\left| \sum_{i=1}^{k} \boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \right| / \sigma\sqrt{\tilde{n}}$ converge in probability to 0 as $n \to \infty$, therefore $\max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n \right| / \sigma\sqrt{\tilde{n}}$ has the same limiting distribution as $\max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} y_i - \boldsymbol{x}_i'\boldsymbol{\beta} \right| / \sigma\sqrt{\tilde{n}}$.

Finally write

$$T_n = \frac{1}{\sqrt{\tilde{n}}} \max_{1 \leq m \leq \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_n}{\hat{\sigma}_n} \right| = \frac{\sigma}{\hat{\sigma}_n \sqrt{\tilde{n}}} \max_{1 \leq m \leq \tilde{n}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_n}{\sigma} \right| .$$

By condition (b), $\sigma / \hat{\sigma}_n \xrightarrow{p} 1$, hence $T_n \xrightarrow{\mathcal{L}} T$. □

Note that, in the proof, the order of the $\boldsymbol{x}_i$s irrelevant. Under the null model the asymptotic distribution of the test statistic depends on $\boldsymbol{x}_i$s only through assumption (a). Any permutation of the rows of $\boldsymbol{X}_n$ will not change the validity of assumption (a). However, simulation studies show that when sample sizes are small, the ordering affects the size of test statistics, hence my small sample adjustments.

A.4 Lemma 3

To prove Theorem 2, I need Lemma 3.

Lemma 3. If condition (a) is satisfied,

$$a_{\tilde{n}} \max_{1 \leq m \leq \tilde{n}} \left| \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \frac{\boldsymbol{x}_i' (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \xrightarrow{p} 0, \quad \text{as} \quad \tilde{n} {\rightarrow} \infty,$$

where $a_{\tilde{n}} = \sqrt{2 \log \log \tilde{n}}$ and $\tilde{n} = \lceil n / (\log \log n)^{2+\delta} \rceil$, for $\delta > 0$.

PROOF. Let

$$h \in \{ j : \sum_{i=1}^{j} \left| \boldsymbol{x}_i' (\hat{\beta} - \beta) \right| \, \| \, / \sqrt{j} = \max_{1 \leq m \leq \tilde{n}} \sum_{i=1}^{m} \left| \boldsymbol{x}_i' (\hat{\beta} - \beta) \right| / \sqrt{m} \}$$

. By Cauchy-Schwartz

$$\begin{aligned} a_{\tilde{n}} \max_{1\le m\le \tilde{n}} \left| \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| &= a_{\tilde{n}} \frac{1}{\sqrt{h}} \left| \sum_{i=1}^{h} \frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \\ &\le a_{\tilde{n}} \frac{1}{\sqrt{h}} \sum_{i=1}^{h} \left| \frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})}{\sigma} \right| \\ &\le \left[\frac{a_{\tilde{n}} a_n \sqrt{h}}{\sigma\sqrt{n}} \right] \left[\frac{1}{h} \sum_{i=1}^{h} \parallel \boldsymbol{x}_i \parallel \right] \left[\frac{\sqrt{n}}{a_n} \parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \parallel \right]. \end{aligned}$$

By Lemma 1, the third term $\sqrt{n} \parallel \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \parallel / a_n$ is bounded a.s. For the second term, if $h < \infty$, then $\sum_{i=1}^{h} \parallel \boldsymbol{x}_i \parallel / h$ is bounded or if $h \to \infty$, by assumption (a), $\sum_{i=1}^{h} \parallel \boldsymbol{x}_i \parallel / h$ is again bounded. If the first term converges to 0, then the right hand side converges in probability to 0. However, $a_n a_{\tilde{n}} \sqrt{h}/\sigma\sqrt{n} \le a_n^2 \sqrt{\tilde{n}}/\sigma\sqrt{n} = 2/\sigma(\log\log n)^{\delta/2} \to 0$, hence Lemma 3 is proved. □

A.5 Proof of Theorem 2

PROOF OF THEOREM 2. First suppose both $\boldsymbol{\beta}$ and σ are known. Applying the Darling-Erdös theorem (Darling and Erdös, 1956),

$$a_{\tilde{n}} \max_{1\le m\le \tilde{n}} \frac{1}{\sqrt{m}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma^2} \right| - b_{\tilde{n}} \xrightarrow{\mathcal{L}} Q \quad \text{as} \quad \tilde{n} \to \infty,$$

where $a_{\tilde{n}} = \sqrt{2\log\log\tilde{n}}$, $b_{\tilde{n}} = (a_{\tilde{n}})^2 + \log a_{\tilde{n}} - \log(\sqrt{2\pi})$, and Q has an extreme value distribution whose cdf is

$$Pr\left[Q < t\right] = \exp\left[-\exp(-t)\right].$$

Let

$$k^* \in \left\{ j : \frac{1}{\sqrt{j}} \left| \sum_{i=1}^{j} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma} \right| = \max_{1\le m\le \tilde{n}} \frac{1}{\sqrt{m}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma} \right| \right\}.$$

So that

$$\begin{aligned}
a_{\tilde{n}} \max_{1\le m\le \tilde{n}} \frac{1}{\sqrt{m}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma}\right| &= a_{\tilde{n}}\frac{1}{\sqrt{k^*}}\left|\sum_{i=1}^{k^*}\frac{y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma}\right| \\
&= a_{\tilde{n}}\frac{1}{\sqrt{k^*}}\left|\sum_{i=1}^{k^*}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}-\sum_{i=1}^{k^*}\frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right| \\
&\le a_{\tilde{n}}\frac{1}{\sqrt{k^*}}\left|\sum_{i=1}^{k^*}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right|+a_{\tilde{n}}\frac{1}{\sqrt{k^*}}\left|\sum_{i=1}^{k^*}\frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right| \\
&\le a_{\tilde{n}} \max_{1\le m\le \tilde{n}} \frac{1}{\sqrt{m}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right|+a_{\tilde{n}}\frac{1}{\sqrt{k^*}}\left|\sum_{i=1}^{k^*}\frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right|.
\end{aligned}$$

Now, let $\quad q^* \in \left\{ j : \frac{1}{\sqrt{j}}\left|\sum_{i=1}^{j}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right| = \max_{1\le m\le \tilde{n}}\frac{1}{\sqrt{m}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right|\right\}.$

I have

$$\begin{aligned}
a_{\tilde{n}} \max_{1\le m\le \tilde{n}} \frac{1}{\sqrt{m}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma}\right| &\ge a_{\tilde{n}}\frac{1}{\sqrt{q^*}}\left|\sum_{i=1}^{q^*}\frac{y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n}{\sigma}\right| \\
&= a_{\tilde{n}}\frac{1}{\sqrt{q^*}}\left|\sum_{i=1}^{q^*}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}-\sum_{i=1}^{q^*}\frac{x_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right| \\
&\ge a_{\tilde{n}}\frac{1}{\sqrt{q^*}}\left|\sum_{i=1}^{q^*}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right|-a_{\tilde{n}}\frac{1}{\sqrt{q^*}}\left|\sum_{i=1}^{q^*}\frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right| \\
&= a_{\tilde{n}} \max_{1\le m\le \tilde{n}} \frac{1}{\sqrt{m}}\left|\sum_{i=1}^{m}\frac{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right|-a_{\tilde{n}}\frac{1}{\sqrt{q^*}}\left|\sum_{i=1}^{q^*}\frac{\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})}{\sigma}\right|.
\end{aligned}$$

By Lemma 3, both

$$a_{\tilde{n}}\left|\sum_{i=1}^{q^*}\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\right|/\sigma\sqrt{q^*} \quad \text{and} \quad a_{\tilde{n}}\left|\sum_{i=1}^{k^*}\boldsymbol{x}_i'(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\right|/\sigma\sqrt{k^*}$$

converge in probability to 0, as $n \to \infty$, therefore

$$a_{\tilde{n}} \max_{1\le m\le \tilde{n}}\left|\sum_{i=1}^{m}\left(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_n\right)/\sigma\sqrt{m}\right|-b_{\tilde{n}} \quad \text{and} \quad a_{\tilde{n}} \max_{1\le m\le \tilde{n}}\left|\sum_{i=1}^{m}\left(y_i-\boldsymbol{x}_i'\boldsymbol{\beta}\right)/\sigma\sqrt{m}\right|-b_{\tilde{n}}$$

have the same limiting distribution. Also

$$Q_n = a_{\tilde{n}} \frac{\sigma}{\hat{\sigma}_n} \max_{1 \le m \le \tilde{n}} \frac{1}{\sqrt{m}} \left| \sum_{i=1}^{m} \frac{y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_n}{\sigma} \right| - b_{\tilde{n}}$$

can be written as

$$Q_n = \frac{\sigma}{\hat{\sigma}_n} \left(a_{\tilde{n}} \max_{1 \le m \le \tilde{n}} \left| \sum_{i=1}^{m} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_n \right) \Big/ \sigma \sqrt{m} \right| - b_{\tilde{n}} \right) + \left(\frac{\sigma}{\hat{\sigma}_n} b_{\tilde{n}} - b_{\tilde{n}} \right).$$

Under condition (b), $\sigma / \hat{\sigma}_n \xrightarrow{p} 1$ as $n \to \infty$, so that the first term converges in law to Q. The second term converges in probability to 0, because by condition (b) and the choice of $\tilde{n}$, $\sigma / \hat{\sigma}_n - 1$ converges to 0 faster than $b_{\tilde{n}}$ converges to infinity. □

Appendix B

Proof of theorems for Chpater 4

B.1 Proof of Theorem 4

PROOF OF THEOREM 4: Any linear unbiased estimate of $\boldsymbol{\rho}'\boldsymbol{X\beta}$ in model (1.3), say $\boldsymbol{a'Y}$, has $\boldsymbol{a}'\boldsymbol{X\beta} = \boldsymbol{\rho}'\boldsymbol{X\beta}$ for any $\boldsymbol{\beta}$, so $\boldsymbol{a}'\boldsymbol{X\beta}_0 = \boldsymbol{\rho}'\boldsymbol{X\beta}_0$ and $\boldsymbol{a}'(\boldsymbol{Y} - \boldsymbol{X\beta}_0)$ determines a linear unbiased estimate of $\boldsymbol{\rho}'\boldsymbol{X}_v\boldsymbol{\gamma}$ in model (1.5) with the same variance in model (1.5) as $\boldsymbol{a'Y}$ has in model (1.3). Thus the BLUE of $\boldsymbol{\rho}'\boldsymbol{X\beta}$ in model (1.3) must have variance no less than the BLUE of $\boldsymbol{\rho}'\boldsymbol{X}_v\boldsymbol{\gamma}$ in model (1.5).

Conversely, if $\boldsymbol{c}'(\boldsymbol{Y} - \boldsymbol{X\beta}_0)$ is unbiased for $\boldsymbol{\rho}'\boldsymbol{X}_v\boldsymbol{\gamma}$ in model (1.5), $\boldsymbol{c}'(\boldsymbol{Y} - \boldsymbol{X\beta}_0) + \boldsymbol{\rho}'\boldsymbol{X\beta}_0 \equiv \tilde{\boldsymbol{c}}'\boldsymbol{Y}$ is unbiased for $\boldsymbol{\rho}'\boldsymbol{X\beta}$ in model (1.3) and has the same variance, so the variance of the BLUE of $\boldsymbol{\rho}'\boldsymbol{X}_v\boldsymbol{\gamma}$ in model (1.5) must have variance no less than the BLUE of $\boldsymbol{\rho}'\boldsymbol{X\beta}$ in model (1.3).

Since the variance of the BLUE of $\boldsymbol{\rho}'\boldsymbol{X}_v\boldsymbol{\gamma}$ in model (1.5) equals the variance of the BLUE of $\boldsymbol{\rho}'\boldsymbol{X\beta}$ in model (1.3), finding the BLUE in model (1.5) determines the BLUE in model (1.3) and vice versa.

B.2 Proof of Theorem 5

PROOF OF THEOREM 5: To prove this, I first check that F^* follows a noncentral F distribution. Christensen (2002, Section 10.3) gives a procedure for testing models like (4.1) when $\tilde{\boldsymbol{b}} = \mathbf{0}$ using $\boldsymbol{T} = \boldsymbol{V} + \boldsymbol{XUX}'$. This can be generalized to models with $\tilde{\boldsymbol{b}} \neq \mathbf{0}$ just like other such tests, see Christensen (2002, Chapter 3). To test $\boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{d}$ or equivalently $\boldsymbol{P}'\boldsymbol{X}_v\gamma = \boldsymbol{k}$ in model (1.5), apply the procedure to model (4.2) by taking $\boldsymbol{T} = \boldsymbol{V}$. Standard arguments (modified in a manner similar to what follows) establish that the noncentrality parameter is 0 if and only if the null hypothesis (model) is true.

To show that F^* can be computed using BLUEs from model (1.3), show

(a) $SSE = \boldsymbol{(Y - X\beta_0 - X_v\hat{\gamma})'V^-(Y - X\beta_0 - X_v\hat{\gamma})}$

(b) $SSE_0 - SSE = (\boldsymbol{P}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[\boldsymbol{P}'\boldsymbol{X}_v(\boldsymbol{X}_v'\boldsymbol{V}^-\boldsymbol{X}_v)^-\boldsymbol{X}_v'\boldsymbol{P}]^-(\boldsymbol{P}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$

(c) $tr(\boldsymbol{CV}) = r(\boldsymbol{V}) - r(\boldsymbol{X}_v)$

(d) $tr(\boldsymbol{C}_0\boldsymbol{X}_v - \boldsymbol{CV}) = r(\boldsymbol{X}_v'\boldsymbol{P})$

First, define matrices based on Christensen (2002, p.232). Pick $\boldsymbol{E}, \boldsymbol{D}$ so that $\boldsymbol{VE} = \boldsymbol{ED}$. Here $\boldsymbol{D} = Diag(d_i)$, where the $d_i s$ are all positive eigenvalues of $\boldsymbol{V}$ and $r(\boldsymbol{V}) = m$. Also $\boldsymbol{E}$ is a matrix of orthonormal columns with the ith column an eigenvector corresponding to d_i. Define $\boldsymbol{D}^{1/2} \equiv Diag(\sqrt{d_i})$. Write

$$\tilde{\boldsymbol{Q}} \equiv \boldsymbol{ED}^{1/2} \quad \tilde{\boldsymbol{Q}}^- \equiv \boldsymbol{D}^{-1/2}\boldsymbol{E}'.$$

Useful facts are

$$(1)\ C(\boldsymbol{V}) = C(\boldsymbol{E}) = C(\tilde{\boldsymbol{Q}}), \quad (2)\ \boldsymbol{M_V = \tilde{Q}\tilde{Q}^-}, \quad (3)\ V = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}',$$

$$(4)\ \boldsymbol{V}^- = \tilde{\boldsymbol{Q}}^{-\prime}\tilde{\boldsymbol{Q}}^-, \quad (5)\ \tilde{\boldsymbol{Q}}^-\tilde{\boldsymbol{Q}} = I_m = \tilde{\boldsymbol{Q}}'\tilde{\boldsymbol{Q}}^{-\prime}.$$

Since $C(\boldsymbol{X}_v) \subset C(\boldsymbol{V})$ and $\boldsymbol{M}_{\boldsymbol{V}} = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^-$, I can write $\boldsymbol{X}_v = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v$. Let $\boldsymbol{b}$ be any solution to $\boldsymbol{P}'\boldsymbol{X}_v\boldsymbol{b} = \boldsymbol{d} - \boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta}_0$,

Proof of (a):

$$SSE = [(\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})]'\boldsymbol{V}^-[(\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})]$$

and recall that $\boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0)$ is $\boldsymbol{X}_v\hat{\boldsymbol{\gamma}}$, the BLUE of $\boldsymbol{X}_v\boldsymbol{\gamma}$, so

$$\begin{aligned}(\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b}) &= (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0) - (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{X}_v\boldsymbol{b} \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\hat{\boldsymbol{\gamma}}) - \boldsymbol{0}.\end{aligned}$$

Hence (a) is proved.

Proof of (b):

$$\begin{aligned}SSE &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})'(\boldsymbol{I} - \boldsymbol{A})'\boldsymbol{V}^-(\boldsymbol{I} - \boldsymbol{A})(Y - X\beta_0 - \boldsymbol{X}_v\boldsymbol{b}) \\ &= [(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})'(\boldsymbol{I} - \boldsymbol{A})'\tilde{\boldsymbol{Q}}^{-\prime}][\tilde{\boldsymbol{Q}}^-(\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \tilde{\boldsymbol{X}}\boldsymbol{b})] \\ &= [\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})]'(\boldsymbol{I} - \boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v})[\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})],\end{aligned}$$

The last equality follows because $\tilde{\boldsymbol{Q}}^-(\boldsymbol{I} - \boldsymbol{A}) = (\boldsymbol{I} - \boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v})\tilde{\boldsymbol{Q}}^-$. With the same argument,

$$SSE_0 = [\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})]'(\boldsymbol{I} - \boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{W}})[\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})].$$

As in Christensen (2002, Prop 3.3.2),

$$\begin{aligned}&SSE_0 - SSE \\ &\quad= [\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})]'(\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v} - \boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{W}})[\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})] \\ &\quad= [\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})]'(\boldsymbol{M}_{\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}'\boldsymbol{P}})[\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})] \\ &\quad= [\boldsymbol{P}'\tilde{\boldsymbol{Q}}\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})]'(\boldsymbol{P}'\tilde{\boldsymbol{Q}}\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}'\boldsymbol{P})^- \\ &\qquad\qquad\times[\boldsymbol{P}'\tilde{\boldsymbol{Q}}\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{0}} - \boldsymbol{X}_v\boldsymbol{b})].\end{aligned}$$

Consider the first and last terms, recalling that $\boldsymbol{d} - \boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta}_0 = \boldsymbol{P}'\boldsymbol{X}_v\boldsymbol{b}$,

$$\begin{aligned}
\boldsymbol{P}'\tilde{\boldsymbol{Q}}\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}^-(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b}) &= \boldsymbol{P}'\boldsymbol{M}_{\boldsymbol{V}}\boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b}) \\
&= \boldsymbol{P}'\boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{V}_0\boldsymbol{b}) \\
&= \boldsymbol{P}'\boldsymbol{V}_0\hat{\gamma} - (\boldsymbol{d} - \boldsymbol{P}'\boldsymbol{X}\boldsymbol{\beta}_0) \\
&= \boldsymbol{P}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{d}
\end{aligned}$$

and the middle term is

$$(\boldsymbol{P}'\tilde{\boldsymbol{Q}}\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}'\boldsymbol{P})^- = [\boldsymbol{P}'\boldsymbol{X}_v(\boldsymbol{X}'_{\boldsymbol{V}}\boldsymbol{V}^-\boldsymbol{X}_v)^-\boldsymbol{X}'_v\boldsymbol{P}]^-$$

so $SSE_0 - SSE = (\boldsymbol{P}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{d})'[\boldsymbol{P}'\boldsymbol{X}_v(\boldsymbol{X}'_v\boldsymbol{V}^-\boldsymbol{X}_v)^-\boldsymbol{X}'_v\boldsymbol{P}]^-(\boldsymbol{P}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{d})$.

Proof of (c): By Christensen (2002, Corollary 10.3.5) with $\boldsymbol{T} = \boldsymbol{V}$,

$$\begin{aligned}
tr[\boldsymbol{C}\boldsymbol{V}] &= tr[\boldsymbol{V}^-(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{V}] \\
&= tr[\tilde{\boldsymbol{Q}}^{-\prime}\tilde{\boldsymbol{Q}}^-(\boldsymbol{I} - \boldsymbol{A})\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}'] \\
&= tr[\tilde{\boldsymbol{Q}}'\tilde{\boldsymbol{Q}}^{-\prime}\tilde{\boldsymbol{Q}}^-(\boldsymbol{I} - \boldsymbol{A})\tilde{\boldsymbol{Q}}] \\
&= tr[\boldsymbol{I}_m - \boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}].
\end{aligned}$$

So $tr(\boldsymbol{C}\boldsymbol{V}) = tr[\boldsymbol{I}_m - \boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\tilde{\boldsymbol{X}}}] = m - r(\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v) = r(\boldsymbol{V}) - r(\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v)$. Note also that $r(\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v) = r(\boldsymbol{D}^{-1/2}\boldsymbol{E}'\boldsymbol{X}_v) = r(\boldsymbol{E}'\boldsymbol{X}_v) = r(\boldsymbol{E}\boldsymbol{E}'\boldsymbol{X}_v) = r(\boldsymbol{M}_{\boldsymbol{V}}\boldsymbol{X}_{\boldsymbol{v}}) = r(\boldsymbol{X}_v)$. Thus $tr(\boldsymbol{C}\boldsymbol{V}) = r(\boldsymbol{V}) - r(\boldsymbol{X}_v)$.

Proof of (d): Similar arguments give $tr(\boldsymbol{C}_0\boldsymbol{V}) = tr(\boldsymbol{I} - \boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{W}})$. So

$$tr(\boldsymbol{C}_0\boldsymbol{V}) - tr(\boldsymbol{C}\boldsymbol{V}) = tr(\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v} - \boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{W}}) = tr(\boldsymbol{M}_{\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}'\boldsymbol{P}})$$

and

$$tr(\boldsymbol{M}_{\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}'\boldsymbol{P}}) = r(\boldsymbol{M}_{\tilde{\boldsymbol{Q}}^-\boldsymbol{X}_v}\tilde{\boldsymbol{Q}}'\boldsymbol{P}) = r(\boldsymbol{X}'_v\tilde{\boldsymbol{Q}}^{-\prime}\tilde{\boldsymbol{Q}}'\boldsymbol{P}) = r(\boldsymbol{X}'_v\boldsymbol{M}_{\boldsymbol{V}}\boldsymbol{P}) = r(\boldsymbol{X}'_v\boldsymbol{P}).$$

Thus $tr(\boldsymbol{C}_0\boldsymbol{V} - \boldsymbol{C}\boldsymbol{V}) = r(\boldsymbol{X}'_v\boldsymbol{P})$. □

B.3 Proof equivolence of tests

EQUIVALENCE OF TESTS: Consider testing model (1.3) against its reduced model (4.1). The traditional method constructs a test statistic by defining

$$\tilde{\boldsymbol{A}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{T}^{-}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{T}^{-} \quad \text{and} \quad \tilde{\boldsymbol{A}}_0 = \boldsymbol{X}_0(\boldsymbol{X}_0'\boldsymbol{T}_0^{-}\boldsymbol{X}_0)^{-}\boldsymbol{X}_0'\boldsymbol{T}_0^{-},$$

where

$$\boldsymbol{T} = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}' \quad \text{and} \quad \boldsymbol{T}_0 = \boldsymbol{V} + \boldsymbol{X}_0\boldsymbol{B}_0\boldsymbol{X}_0'$$

for some nonnegative definite matrix $\boldsymbol{B}$ and $\boldsymbol{B}_0$ such that $C(\boldsymbol{X}) \subset C(\boldsymbol{T})$ and $C(\boldsymbol{X}_0) \subset C(\boldsymbol{T}_0)$. $\tilde{\boldsymbol{A}}\boldsymbol{Y}$ is the BLUE of $\boldsymbol{X}\boldsymbol{\beta}$ in model (1.3) and $\tilde{\boldsymbol{A}}_0(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}})$ is the BLUE of $\boldsymbol{X}_0\boldsymbol{\eta}$ in model (4.1). The SSE in model (1.3) and its reduced model are

$$SSE^1 = (\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}})'(\boldsymbol{I} - \tilde{\boldsymbol{A}})'\boldsymbol{T}^{-}(\boldsymbol{I} - \tilde{\boldsymbol{A}})(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}}),$$

$$SSE_0^1 = (\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}})'(\boldsymbol{I} - \tilde{\boldsymbol{A}}_0)'\boldsymbol{T}_0^{-}(\boldsymbol{I} - \tilde{\boldsymbol{A}}_0)(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}}).$$

Denote $\boldsymbol{C}^1 \equiv (\boldsymbol{I} - \tilde{\boldsymbol{A}})'\boldsymbol{T}^{-}(\boldsymbol{I} - \tilde{\boldsymbol{A}})$ and $\boldsymbol{C}_0^1 \equiv (\boldsymbol{I} - \tilde{\boldsymbol{A}}_0)'T^{-}(\boldsymbol{I} - \tilde{\boldsymbol{A}}_0)$.

To prove the equivalence of my testing method and this traditional method, it suffices to show (a) $SSE^1 = SSE$, (b) $SSE_0^1 = SSE_0$, (c) $tr(\boldsymbol{C}\boldsymbol{V}) = tr(\boldsymbol{C}^1\boldsymbol{T})$ and (d) $tr(\boldsymbol{C}_0\boldsymbol{V}) = tr(\boldsymbol{C}_0^1\boldsymbol{T})$.

Proof of (a): The SSE^1 for model (1.3) is, by Christensen (2002, Theorem 10.3.1),

$$\begin{aligned}(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}})'(\boldsymbol{I} - \tilde{\boldsymbol{A}})'\boldsymbol{T}^{-}(\boldsymbol{I} - \tilde{\boldsymbol{A}})(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}}) &= \boldsymbol{Y}'(\boldsymbol{I} - \tilde{\boldsymbol{A}})'\boldsymbol{T}^{-}(\boldsymbol{I} - \tilde{\boldsymbol{A}})\boldsymbol{Y} \\ &= \boldsymbol{Y}'(\boldsymbol{I} - \tilde{\boldsymbol{A}})'\boldsymbol{V}^{-}(\boldsymbol{I} - \tilde{\boldsymbol{A}})\boldsymbol{Y}\end{aligned}$$

where $\tilde{\boldsymbol{A}}\boldsymbol{Y}$ is the BLUE of $\boldsymbol{X}\boldsymbol{\beta}$. However, the BLUE of $X\beta$ is also

$$\tilde{\boldsymbol{A}}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{X}_v\hat{\gamma} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0),$$

so $(\boldsymbol{I} - \tilde{\boldsymbol{A}})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0)$ and $SSE^1 = SSE.$

Proof of (b): In $(\boldsymbol{I} - \tilde{\boldsymbol{A}}_0)(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}})$, $\boldsymbol{X}\tilde{\boldsymbol{b}} + \tilde{\boldsymbol{A}}_0(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}})$ is the BLUE of $E(\boldsymbol{Y})$ in model (4.1). In model (4.2) the BLUE of $E(\boldsymbol{Y})$ is $\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{X}_v\boldsymbol{b} + \boldsymbol{A}_0(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})$, so

$$(\boldsymbol{I} - \tilde{\boldsymbol{A}}_0)(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{b}}) = (\boldsymbol{I} - \boldsymbol{A}_0)(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_0 - \boldsymbol{X}_v\boldsymbol{b})$$

and $SSE_0^1 = SSE_0$.

Proof of (c): To show $tr(\boldsymbol{C}\boldsymbol{V}) = tr(\boldsymbol{C}^1\boldsymbol{T})$, it suffices to show $r(\boldsymbol{T}) - r(\boldsymbol{X}) = r(\boldsymbol{V}) - r(\boldsymbol{X}_v)$.

$$\begin{aligned} C(\boldsymbol{T}) &= C(\boldsymbol{X}, \boldsymbol{V}) \\ &= C(\boldsymbol{X}_v) \oplus C(\boldsymbol{X}_v)^{\perp}_{C(\boldsymbol{V})} \oplus C(\boldsymbol{X}_v)^{\perp}_{C(\boldsymbol{X})} \\ &= C(\boldsymbol{V}) \oplus C(\boldsymbol{X}_v)^{\perp}_{C(\boldsymbol{X})}. \end{aligned}$$

So $r(\boldsymbol{T}) = r(\boldsymbol{V}) + r(C(\boldsymbol{X}_v)^{\perp}_{C(\boldsymbol{X})})$. Also, $C(\boldsymbol{X}) = C(\boldsymbol{X}_v) \oplus C(\boldsymbol{X}_v)^{\perp}_{C(\boldsymbol{X})}$ and $r(\boldsymbol{X}) = r(\boldsymbol{X}_v) + r(C(\boldsymbol{X}_v)^{\perp}_{C(\boldsymbol{X})})$. Thus $r(\boldsymbol{T}) - r(\boldsymbol{X}) = r(\boldsymbol{V}) - r(\boldsymbol{X}_v)$.

Proof of (d) Arguments similar to (c) give $tr(\boldsymbol{C}_0\boldsymbol{V}) = tr(\boldsymbol{C}_0^1\boldsymbol{T})$.

Thus, the traditional testing method is equivalent to this testing method. □

Appendix C

Review of lack-of-fit tests

C.1 Utts's test

Utts (1982) proposed a lack-of-fit test, known as Rainbow test which can be used to check the adequacy of regression models with p covariates. Considering testing the lack-of-fit of model (1.1), his method involves fitting the"same model" twice, once respect to the full dataset of n observations and the other time to a sub dataset of n_1 observations. For $k = 1, 2$, in each dataset write

$$\boldsymbol{Y}_k = \begin{bmatrix} y_{k1} \\ \vdots \\ y_{kn_k} \end{bmatrix}, \quad \boldsymbol{X}_k = \begin{bmatrix} \boldsymbol{x}'_{k1} \\ \vdots \\ \boldsymbol{x}'_{kn_k} \end{bmatrix} \quad \text{and} \quad \boldsymbol{e}_k = \begin{bmatrix} e_{k1} \\ \vdots \\ e_{kn_k} \end{bmatrix},$$

where y_{ki}, $\boldsymbol{x}_{ki}$ and e_{ki} are the ith response, covariates and error corresponding to the kth subset data. Here $n_2 = n - n_1$. The model for the full dataset is

$$\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \boldsymbol{\beta}_1 + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{bmatrix} \tag{C.1}$$

and for the sub dataset, it is

$$\boldsymbol{Y}_1 = \boldsymbol{X}_1 \boldsymbol{\beta}_2 + \boldsymbol{e}_1. \tag{C.2}$$

Here both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of p unknown parameters. Let

$$\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2]', \quad \boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]' \quad \text{and} \quad \boldsymbol{e} = [\boldsymbol{e}_1, \boldsymbol{e}_2]',$$

the SSEs for model (C.1) and model (C.2) are

$$SSE(C.1) = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{Y} \quad \text{and} \quad SSE(C.2) = \boldsymbol{Y}_1'(\boldsymbol{I}_{n_1} - \boldsymbol{M}_{\boldsymbol{X}_1})\boldsymbol{Y}_1.$$

To test the lack-of-fit of model (C.1), Utts proposed an F statistic based on comparing model (C.1) and model (C.2). That is

$$F_1 = \frac{SSE(C.2) - SSE(C.1)/n_2}{SSE(C.2)/n_1 - p} \sim F_{(n_2,\ n_1 - p)}.$$

Large values of F_1 indicate inadequacy of model (C.1).

Utts's approach seems different from the classical methods that involve extending the null model to a larger model. However, it is exactly partitioning lack-of-fit test and the extended model is obtained by fitting the rest of dataset (n_2 observations) exactly. Consider such an extended model

$$\boldsymbol{Y} = \tilde{\boldsymbol{X}}\boldsymbol{\gamma} + \boldsymbol{e}, \tag{C.3}$$

where

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_1} \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 \end{bmatrix},$$

and $\boldsymbol{\beta}_3$ is a vector of n_2 unknown parameters. Clearly $C(\boldsymbol{X}) \subset C(\tilde{\boldsymbol{X}})$ and the SSE of model (C.3) is

$$SSE(C.3) = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M}_{\tilde{\boldsymbol{X}}})\boldsymbol{Y}.$$

The classical lack-of-fit test is

$$F_2 = \frac{SSE(C.3) - SSE(C.1)/n_2}{SSE(C.3)/n_1 - p} = \frac{\boldsymbol{Y}'(\boldsymbol{M}_{\tilde{\boldsymbol{X}}} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{Y}/n_2}{\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M}_{\tilde{\boldsymbol{X}}})\boldsymbol{Y}/n_1 - p} \sim F_{(n_2,\ n_1-p)}.$$

Note that

$$\boldsymbol{M}_{\tilde{\boldsymbol{X}}} = \begin{bmatrix} \boldsymbol{M}_{\boldsymbol{X}_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{n_2} \end{bmatrix} \quad \text{and} \quad \boldsymbol{I} - \boldsymbol{M}_{\tilde{\boldsymbol{X}}} = \begin{bmatrix} \boldsymbol{I}_{n_1} - \boldsymbol{M}_{\boldsymbol{X}_1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}.$$

It follows that

$$SSE(C.3) = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M}_{\tilde{\boldsymbol{X}}})\boldsymbol{Y} = \boldsymbol{Y}_1'(\boldsymbol{I}_{n_1} - \boldsymbol{M}_{\boldsymbol{X}_1})\boldsymbol{Y}_1 = SSE(C.2).$$

Hence, $F_1 = F_2$ and Utts's test is equivalent to the classical lack-of-fit test of extended model (C.3) against null model (C.1).

Utts's test involves partitioning the data into two parts. The key issue is how to partition the data. It apparently can affect the test's performance. In application, the author suggested partitioning the data according to the leverage scores. In this article, he used the half of dataset with lower leverage scores as sub dataset. Apparently, this method can be extended by considering other partitioning methods.

C.2 Eubank and Spiegelman's test

Eubank and Spiegelman (1990) proposed a lack-of-fit test for model (1.1) with

$$\boldsymbol{X} = [\boldsymbol{X_1, ..., X_n}]' \quad \text{and} \quad \boldsymbol{X}_i = [1, x_i]',$$

where $x_1 \leq x_2, ..., \leq x_n$. They considered an extended model, similar to model (2.8) with $\boldsymbol{H}$ satisfying

$$\boldsymbol{H}'\boldsymbol{H} = n\boldsymbol{I} \quad \text{and} \quad \boldsymbol{H}'\boldsymbol{X} = \boldsymbol{0}.$$

Here $\boldsymbol{H}/\sqrt{n}$ is an orthonomal matrix and $\boldsymbol{M_H} = \boldsymbol{H}\boldsymbol{H}'/n$. The lack-of-fit test of model (1.1) is performed through testing model (2.8) on

$$H_0 : \boldsymbol{\gamma} = \boldsymbol{0}, \quad \text{and} \quad H_a : \boldsymbol{\gamma} \neq \boldsymbol{0}. \tag{C.4}$$

With $\hat{\gamma} \equiv \boldsymbol{H}'\boldsymbol{Y}/n$, for a fixed k, (C.4) can be tested using a χ^2 test,

$$n\hat{\gamma}'\hat{\gamma} \sim \chi^2(k).$$

This test involves comparing model (1.1) and model (2.8). To see that write the sum of square error of model (2.8) as

$$\begin{aligned} SSE &= \boldsymbol{Y}'\left(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X},\boldsymbol{H}}\right)\boldsymbol{Y} \\ &= \boldsymbol{Y}'\left(\boldsymbol{I} - \boldsymbol{M_X} - \boldsymbol{M_H}\right)\boldsymbol{Y} \\ &= \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M_X})\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{H}'\boldsymbol{Y}/n. \end{aligned}$$

Here $\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M_X})\boldsymbol{Y}$ is the SSE of the model (1.1). Thus, $n\hat{\gamma}'\hat{\gamma} = \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{H}'\boldsymbol{Y}$ is the difference of SSEs between model (1.1) and model (2.8).

However, using a pre-specified k dimension $\boldsymbol{H}$ matrix, the test is only consistent against some but not all alternatives. The authors suggested a more general approach. They began with fitting residuals from model (1.1),

$$\hat{\boldsymbol{e}} = \boldsymbol{H}_2\boldsymbol{\gamma}_2 + \boldsymbol{e}_2, \tag{C.5}$$

where $\boldsymbol{H}_2$ is a $n \times k_2$ matrix of smooth functions and $t_{ij}(\cdot)$ is the ith row and jth column of $\boldsymbol{H}_2$. The hypothesis (C.4) is approximated by

$$H_0 : \boldsymbol{\gamma}_2 = \boldsymbol{0}, \quad \text{versus} \quad H_a : \boldsymbol{\gamma}_2 \neq \boldsymbol{0}, \tag{C.6}$$

where $\boldsymbol{\gamma}_2$ is a vector of k_2 parameters. Then a test is considered by standardizing $\hat{\gamma}_2'\hat{\gamma}_2$, that is

$$T^*_{ES} = n\left(\hat{\gamma}_2'\hat{\gamma}_2 - \sigma^2 p/n\right)/\sigma^2\sqrt{2p} \xrightarrow{\mathcal{L}} N(0,1) \tag{C.7}$$

Intuitively, the H_0 of (C.6) is rejected for large values of T^*_{ES}. To obtain the test that are consistent against all alternatives, one needs to let k_2 vary. The key point is to select the type of smooth functions and the value k_2, which is used to control the over fitting (smoothness) of the extended model (C.5). The authors, on the other hand, set k_2 to be as large as $n-2$ and control the smoothness of extended model by fitting residuals $\hat{e}_i$s with a cubic smooth spline, which minimizes

$$\sum_{i=1}^{n}(\hat{e}_i - r_\lambda(x_i))^2 + \lambda \int_0^1 r''_\lambda(x)^2 dx. \quad \lambda > 0, \tag{C.8}$$

Here,

$$r_\lambda(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \sum_{j=3}^{n} \boldsymbol{t}'_j \hat{\boldsymbol{e}} t_j(x_i)/n(1+\lambda\theta_j),$$

where $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1]'$ is LSE of $\boldsymbol{\beta}$ from model (1.1), $\boldsymbol{t}_j(\cdot)$s for $j = 3, ..., n$ are vectors of n natural spine basis functions with

$$\boldsymbol{t}_j = [t_j(x_1), ..., t_j(x_n)]' \quad \text{and} \quad \boldsymbol{t}'_j \boldsymbol{J}_n = \boldsymbol{t}'_j \boldsymbol{t}_i = n\delta_{ij},$$

where $\delta_{ij} = 1$ for $i = j$, 0 otherwise. Also, θ_js satisfy $0 < \theta_3 \leq ... \leq \theta_n$ and λ is a smooth parameter that controls the smoothness of the function $r_\lambda(\cdot)$. Defined the jth column of $\boldsymbol{H}_2$ as $\boldsymbol{t}_j/\sqrt{n(1+\lambda\theta_j)}$ and

$$\hat{\boldsymbol{e}} - \boldsymbol{r}_\lambda = (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{Y} - (\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{Y} + \boldsymbol{M}_{\boldsymbol{H}_2}(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{Y}).$$

Since $\boldsymbol{X}'\boldsymbol{H} = 0$,

$$\boldsymbol{M}_{\boldsymbol{H}_2}(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{Y} = \boldsymbol{M}_{\boldsymbol{H}_2}\boldsymbol{Y}$$

and $\boldsymbol{e}^s = \hat{\boldsymbol{e}} - \boldsymbol{r}_\lambda = (\boldsymbol{I} - 2\boldsymbol{M}_{\boldsymbol{X}} - \boldsymbol{M}_{\boldsymbol{H}_2})\boldsymbol{Y} = (\boldsymbol{Y} - 2\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{H}}_2\hat{\gamma})$. Hence, minimizing (C.8) not only penalize the smoothness of $r_\lambda(\cdot)$, but also force $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ to be a biased estimate. The fitted function of $r_\lambda(x_i)$ is

$$\hat{r}_\lambda(x_i) = \sum_{j=3}^{n} \boldsymbol{t}'_j \boldsymbol{r} t_j(x_i)$$

and the sum of square of fitted value is

$$\sum_{i=1}^{n} \hat{r}_{\lambda}^{2}(x_i) = n \sum_{j=3}^{n} \hat{\gamma}_j^2/(1+\lambda\theta_j) = \boldsymbol{t}_j' \hat{\boldsymbol{e}} t_j(x_i)/n(1+\lambda\theta_j).$$

The smooth spline version of test (C.7) is

$$T_{ES} = \left(\sum_{j=1}^{n} h_{\lambda}^{2}(x_j) - \sigma^2 \sum_{j=3}^{n}(1+\lambda\theta_j)^{-2}\right)/\sigma^2 \left(2\sum_{j=3}^{n}(1+\lambda\theta_j)^{-4}\right)^{1/2}. \quad \text{(C.9)}$$

The author showed that if $h(n) = 1/(n\lambda^{1/8})^{1/2}$,

$$T_{ES} \xrightarrow{\mathcal{L}} N(||g||^2(\sqrt{2}C)^{1/2}\sigma^2, 1)$$

as

$$n \to \infty, \quad \lambda \to 0 \quad \text{and} \quad n\lambda \to \infty.$$

Here x_js are assumed from a continuous positive density w. f and g are in Hilbert space $L_2(w)/\{1, x\}$, f and f' are absolutely continuous, f'' is square integrable and C is set to be

$$C = (\int_0^{\infty}(1+x^4)^{-4}dx)(\int_0^1 w(t)^{1/4}dt)/\pi.$$

In calculation,

$$\sum_{j=3}^{n}(1+\lambda\theta_j)^{-4} \quad \text{and} \quad \sum_{j=3}^{n}(1+\lambda\theta_j)^{-4}$$

are approximated by

$$((j-2)\pi)^4(\int_0^1 w(t)^{1/4}dt)^{-4},$$

θ_j is approximated by

$$((j-1.5)\pi)^4(\int_0^1 w(t)^{1/4}dt)^{-4}$$

and $\int_0^1 w(t)^{1/4}$ can be approximated using a density estimator. σ can be replace by a appropriate estimate using nonparametric techniques. The key point is to select λ. The author suggested using cross-validation. Also, as the author pointed out, the power of the test seems relatively insensitive to the choice of λ, but this statement is inconclusive.

C.3 Fan's test

Fan (1996) proposed a lack-of-fit test of density functions. Let x_is for $i = 1, ..., n$ be a iid sample from a specific distribution with unknown cumulative density function (CDF) $F(x)$. The hypothesis is

$$H_0 : F(x) = F_0(x), \quad \text{versus} \quad H_a : F(x) \neq F_0(x). \tag{C.10}$$

Intuitively, testing the adequacy of $F_0(x)$ can be performed by checking the distance between $F_0(x)$ and the empirical CDF $\hat{F}_n(x)$. Two popular tests based on empirical CDFs are Kolmogorove-Smirnov test (KS test), and the Cramér-Von Mises test (CVM test):

$$\hat{T}_{KS} = \sqrt{n} \sup_x |\hat{F}_n(x) - F_0(x)|,$$

and

$$\hat{T}_{CVM} = n \int \{\hat{F}_n(x) - F_0(x)\}^2 dF_0(x).$$

However, these procedures have been known as inefficient in detecting the local features of proposed models, for example, high frequency components and local bumps. In this paper, Fan showed the inefficiency of CVM test and proposed an alternative approach based on power consideration. He transformed the test of CDF into the test of the mean vector of multivariate normal distribution. In test (C.10), under H_0, $F_0(x)$ is uniformly distributed, whereas under H_a, $F_0(x)$ is not uniformly distributed as x is no longer from the distribution $F_0(\cdot)$. Hence (C.10) is equivalent to

$$H_0 : F(x) \sim \text{uniform(0, 1)} \quad \text{versus} \quad H_1 : F(x) \not\sim \text{uniform(0, 1)}. \tag{C.11}$$

Applying the Fourier transformation to $F(x)$ gives us,

$$\theta_{2j-1} = \int_0^1 \cos(2\pi jx) d(F(x)) \quad \text{and} \quad \theta_{2j} = \int_0^1 \sin(2\pi jx) d(F(x)),$$

for $j = 1, 2, ...\infty$. Apparently, if $F(x) = x$, we have $\theta_j = 0$. Hence, testing (C.11) is equivalent to testing

$$H_0 : \theta_j = 0, \quad \text{versus} \quad H_1 : \theta_j \neq 0, \tag{C.12}$$

for $j = 1, ...\infty$. Let $\hat{\theta}_j$ be the corresponding empirical estimate of θ_j for $j = 1, ...\infty$. Using the results from Eubank and Lariccia (1992), rewrite CVM test as

$$\hat{T}_{CVM} = \frac{n}{2\pi^2} \sum_{j=1}^{\infty} j^{-2}(\hat{\theta}_{2j-1}^2 + \hat{\theta}_{2j}^2). \tag{C.13}$$

It has been shown that these Fourier terms can efficiently transfer the empirical CDF into high frequency terms (larger j) and low frequency terms (smaller j). The term $1/j^2$ greatly weights down the high frequency terms, making the test procedure almost uses only the first few coefficients. Thus, Fan argued that this is why CVM test is ineffective, and proposed an alternative approach based on Neyman's most powerful test. Under H_0, one can show that these Fourier series are asymptotically independent and normally distributed, with

$$\hat{\theta}_j \xrightarrow{\mathcal{L}} N(\theta_j, n^{-1}),$$

for $j = 1, ..., N$, where $N/n \to 0$. This leads the problem to consider a multivariate normal distribution with $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, n^{-1}\boldsymbol{I}_N)$ and (C.12) is approximated by

$$H_0 : \boldsymbol{\theta} = 0, \quad \text{versus} \quad H_a : \boldsymbol{\theta} \neq 0. \tag{C.14}$$

The difficulty is that $\boldsymbol{\theta}$ is of high dimension and testing all dimensions of $\boldsymbol{\theta}$ is difficult. Then the key point is how to select a meaningful part of summation from equation (C.13) to boost the power of the test. The author proposed several approaches. They are adaptive Neyman's test, hard thresholding test and soft thresholding test.

In an easy scenario, where the specific alternative of (C.14) is $\boldsymbol{\theta} = \boldsymbol{\theta}_a$, as $m \to \infty$, the power of Neyman's test is approximately equal to

$$1 - \Phi\left(z_{1-\alpha} - \frac{1}{\sqrt{2m}} \sum_{j=1}^{m} \theta_{j0}^2\right),$$

where θ_{j0} is the jth component of θ_0. To achieve the highest power in the above function (maximize the above function) one needs to maximize $\frac{1}{\sqrt{2m}} \sum_{j=1}^{m} \theta_{j0}^2$. This in turn suggests selecting

$$\hat{m} = \underset{m:1\leq m\leq n}{\text{argmax}} \left\{ \frac{1}{\sqrt{2m}} \sum_{j=1}^{m} (\hat{\theta}_j^2 - 1) \right\}.$$

Note that $\frac{1}{\sqrt{2m}} \sum_{j=1}^{m} (\hat{\theta}_j^2 - 1)$ is a unbiased estimate of $\frac{1}{\sqrt{2m}} \sum_{j=1}^{m} \theta_{j0}^2$. Naturally, Fan's test takes the form of

$$T_{AN}^{*} = \max_{1\leq m\leq n} \left\{ \frac{1}{\sqrt{2m}} \sum_{j=1}^{m} (\hat{\theta}_j^2 - 1) \right\}$$

and H_0 is rejected for large values of T_{AN}^{*}.

By Darling Erdös theorem, $T_{AN} = a_n T_{AN}^{*} - b_n$, with $a_n = \sqrt{2 \log\log n}$ and $b_n = a_n^2 + \log a_n - 0.5 \log(4\pi)$, approximates an extreme value distribution,

$$P(T_{AN} < x) \xrightarrow{\mathcal{L}} \exp(\exp(-x)), \quad \text{as} \quad n \to \infty.$$

However, if the signals of CDF concentrate on high frequency terms, adaptive Neyman's test suffers a significant decrease in power. To improve it under this situation, Fan proposed a threshold (hard) value leading the test statistic to

$$\hat{T}_H^{*} = \sum_{j=1}^{n} \hat{\theta}_j^2 I(|\hat{\theta}_j| > \delta).$$

By Theorem 4 of Donoho and Johnstone (1994), δ needs to be close to $\sqrt{2 \log n}$. For better performance the author suggested using

$$\delta = \sqrt{2 \log(n a_n)}, \quad \text{with} \quad a_n = c(\log n)^{-d},$$

for some positive constants c and d. Normalizing $\hat{T}_H^{*}$ with its mean and variance, the testing procedure is

$$\hat{T}_H = \sigma_{n,H}^{-1}(\hat{T}_H^{*} - \mu_{n,H}) \xrightarrow{\mathcal{L}} N(0,1)$$

where

$$\mu_{n,H} = \sqrt{2/\pi} a_n^{-1} \delta(1 + \delta^{-2}) \quad \text{and} \quad \sigma_{n,H} = \sqrt{2/\pi} a_n^{-1} \delta^3 (1 + 3\delta^{-2})$$

are the asymptotic mean and variance of $\hat{T}_H^*$.

The soft-thresholding function is introduced by Bickel (1983). However, in practice the soft-thresholding having slow converging problems. So it is not introduced here. For more information, please check Bickel (1983) and Fan (1991).

Note that, both soft and hard thresholding methods are less related to the partial sum process that mentioned in this dissertation. However, it is necessary to add them as a counter part which makes up the inefficiency of adaptive Neyman's test.

C.4 Fan and Huang's test

Fan and Huang (2001) extended the Fan (1996)'s lack-of-fit test of density function to regression models. Under null model (1.1), the residuals vector is

$$\hat{\boldsymbol{e}} \sim N(\mathbf{0}, (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\sigma^2).$$

A quick and simple diagnostic is to plot residuals against each predictor for systematic departures from 0. Conditional on equal variance assumption, this technique can be more objectively presented by performing hypothesis test of

$$H_0 : E(\hat{\boldsymbol{e}}) = \mathbf{0}, \quad \text{versus} \quad H_a : E(\hat{\boldsymbol{e}}) \neq \mathbf{0}, \tag{C.15}$$

which is testing the mean of multivariate normal distribution. The issue is $E(\hat{\boldsymbol{e}})$ is a vector that has the same amount of parameters as sample size. Hence, it is impossible to test all dimensions. In order to make it work, the dimension of the test needs to be shrunken to a manageable number while reserving as much useful information as possible. Using the similar idea as in Fan (1996), they applied a discrete Fourier transformation to these residuals. Specifically, the Fourier transformed residuals are

$$\hat{e}_{2j-1}^* = (2/n)^{1/2} \sum_{i=1}^{n} \cos(2\pi ij/n)\hat{e}_i,$$

$$\hat{e}^*_{2j-1} = (2/n)^{1/2} \sum_{i=1}^{n} \sin(2\pi ij/n)\hat{e}_i,$$

for $j = 1, ..., [n/2]$. The test (C.15) is equivalent to

$$H_0 : E(\hat{\boldsymbol{e}}^*) = \mathbf{0}, \quad \text{versus} \quad H_a : E(\hat{\boldsymbol{e}}^*) \neq \mathbf{0}. \tag{C.16}$$

The high dimension problem still exist, but fortunately, the Fourier transformation may compresses useful signals into lower frequencies so that it is appropriate to only work with the low frequency terms. Using the test form in Fan (1996), the author offered an "objective" way of selecting the subset of testing parameters by defining their test as

$$T^*_{AN} = \max_{1\le m\le \tilde{n}} \frac{1}{\sqrt{2m\hat{\sigma}}^4} \sum_{i=1}^{m} (\hat{e}^{*2}_{\ i} - \hat{\sigma}^2).$$

According to Darling Erdös theorem, normalizing T^*_{AN} as

$$T_{AN} = a_{\tilde{n}} T^*_{AN} - b_{\tilde{n}},$$

with $a_{\tilde{n}} = \sqrt{2\log\log\tilde{n}}$ and $b_{\tilde{n}} = a^2_{\tilde{n}} + \log a_{\tilde{n}} - 0.5\log(4\pi)$,

$$Pr[T_{AN} < t] \xrightarrow{\mathcal{L}} \exp(-\exp(-t)).$$

The hidden benefit of Fourier transformation is the rate of convergency. To apply Darling Erdös theorem the terms in the partial sums are assumed to be independent. Fourier transformation orthogonalizes the residual vector so that $\hat{e}^*_i$s are more linear independent with each other than $\hat{e}_i$s. When lack-of-fit appears, one would expect both T^*_{AN} and T_{AN} to take large values.

F-H's test T^*_{AN} considers summing up to $\tilde{n} = n$ Fourier transformed residuals. However, in F-H's proof and simulation studies, they set $\tilde{n} = n/(\log\log n)^4$ as the upper bound and argued that in application it does not make much difference from using $\tilde{n} = n$. While partially agreed with this argument that the performance of the test is insensitive to mild changes in $\tilde{n}$, C-S pointed out that the limiting theorem does not apply if $\tilde{n} = n$. Additional discussion of $\tilde{n}$ is postponed to the next section.

C.5 Christensen and Sun's test

Christensen and Sun (2010) proposed tests by recasting F-H's test back to smooth test. They pointed out that F-H's method involves fitting a two-stage model. To see that, define $\mathbf{\Gamma}_m$ as a matrix generated by normalizing the columns of Φ, where $\Phi = [\phi_2, ..., \phi_m]$ are vectors of Fourier series with

$$\phi_{2q} = \left[\cos(2\pi q\frac{1}{n}), ..., \cos(2\pi q\frac{n}{n})\right]$$

$$\phi_{2q+1} = \left[\sin(2\pi q\frac{1}{n}), ..., \sin(2\pi q\frac{n}{n})\right].$$

$\phi_1 = [1, ..., 1]'$ is redundant in models with intercept. F-H's transformed residuals vector is

$$\hat{\boldsymbol{e}}^* = \mathbf{\Gamma}_m\hat{\boldsymbol{e}} = \mathbf{\Gamma}_m(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Note that the column ϕ_js are independent with each other. So we have $\mathbf{\Gamma}_m'\mathbf{\Gamma}_m = \boldsymbol{I}_m$ and define

$$\boldsymbol{M}_m \equiv \boldsymbol{M}_{\mathbf{\Gamma}_m} = \mathbf{\Gamma}_m\mathbf{\Gamma}_m'.$$

The key component in F-H's test, $\hat{\boldsymbol{e}}^{*\prime}\hat{\boldsymbol{e}}^*$, involves fitting residuals. That is

$$\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{\Gamma}_m\boldsymbol{\gamma}_m + \boldsymbol{e}. \tag{C.17}$$

The sum square error for this model is

$$SSE = \hat{\boldsymbol{e}}'(\boldsymbol{I} - \boldsymbol{M}_m)\hat{\boldsymbol{e}} = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} - \hat{\boldsymbol{e}}'\boldsymbol{M}_m\hat{\boldsymbol{e}}.$$

Here $\hat{\boldsymbol{e}}'\boldsymbol{M}_m\hat{\boldsymbol{e}} = \hat{\boldsymbol{e}}^{*\prime}\hat{\boldsymbol{e}}^*$ and $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$ is the SSE of the model (1.1). Hence $\hat{\boldsymbol{e}}^{*\prime}\hat{\boldsymbol{e}}^*$ is the difference between the SSE of the model (1.1) and the model (C.17). This SSE is the outcome of a two-stage fitting (first fit (1.1) to estimate $\hat{\boldsymbol{\beta}}$ and then fit model (C.17)). C-S proposed an alternative method which requires only one stage fitting of

$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} = \mathbf{\Gamma}_m\boldsymbol{\gamma}_m + \boldsymbol{e}. \tag{C.18}$$

They rewrote model (C.18) as

$$\boldsymbol{Y} = \boldsymbol{X\beta}_0 + (\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{\Gamma}_m\boldsymbol{\gamma} + \boldsymbol{e}.$$

The SSE of this model is

$$\begin{aligned} SSE &= \boldsymbol{Y'}(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}} - \boldsymbol{M}_{(\boldsymbol{I}-\boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{\Gamma}_m})\boldsymbol{Y} \\ &= \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{M}_{(\boldsymbol{I}-\boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{\Gamma}_m}\boldsymbol{Y}. \end{aligned}$$

Here $\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{Y}$ is the SSE of model (1.1) and $\boldsymbol{Y}'\boldsymbol{M}_{(\boldsymbol{I}-\boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{\Gamma}_m}\boldsymbol{Y}$ is the difference between the SSE of (1.1) and the model (C.18). This method is quite similar to the ones that compare the performance of full model and reduced model. Replacing $\sum_{i=1}^{m} \hat{e}^{*2}_{i}$ in F-H's test with $\boldsymbol{Y}'\boldsymbol{M}_{(\boldsymbol{I}-\boldsymbol{M}_X)\boldsymbol{\Gamma}_m}\boldsymbol{Y}$, C-S proposed their first test statistic

$$\hat{T}_{cs1,\tilde{n}} = \max_{2\le m\le\tilde{n}} \left\{ \sqrt{\frac{r_m}{2}} \frac{\boldsymbol{Y}'\boldsymbol{M}_{(1-\boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{\Gamma}_m}\boldsymbol{Y}/r_m - \hat{\sigma}^2}{\hat{\sigma}^2} \right\},$$

where $r_m \equiv r[(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{\Gamma}_m]$. Similar to F-H's test, after normalization with $a_{r_{\tilde{n}}} = \sqrt{2\log\log r_{\tilde{n}}}$ and $b_{r_{\tilde{n}}} = a^2_{r_{\tilde{n}}} + \log a_{r_{\tilde{n}}} - \log(2\sqrt{2\pi})$,

$$\hat{W}_{cs1} = a_{r_{\tilde{n}}}\hat{T}_{cs1,\tilde{n}} - b_{r_{\tilde{n}}},$$

approximates an extreme value distribution.

C-S's second approach is to directly estimate $\boldsymbol{\Gamma}'_m\boldsymbol{e}$. F-H estimated

$$\boldsymbol{\Gamma}'_m\boldsymbol{e} = \boldsymbol{\Gamma}'_m(\boldsymbol{Y} - \boldsymbol{X\beta})$$

by plugging in the least square estimate of $\boldsymbol{\beta}$. An alternative approach of improving F-H's test is through directly estimation of $\boldsymbol{e}^* = \boldsymbol{\Gamma}'_m\boldsymbol{e}$. C-S multiplied $\boldsymbol{\Gamma}'_m$ to the left of null model to get

$$\boldsymbol{\Gamma}'_m\boldsymbol{Y} = \boldsymbol{\Gamma}'_m\boldsymbol{X\beta} + \boldsymbol{\Gamma}'_m\boldsymbol{e}. \tag{C.19}$$

In this way $\boldsymbol{\Gamma}'_m\boldsymbol{e}$ can be estimated using least square method directly,

$$\tilde{\boldsymbol{e}}_m = (\boldsymbol{I}_m - \boldsymbol{M}_{\boldsymbol{\Gamma}'_m\boldsymbol{X}})\boldsymbol{\Gamma}'_m Y.$$

With some algebra steps, the sum of square is

$$\tilde{\boldsymbol{e}}_m' \tilde{\boldsymbol{e}}_m = \boldsymbol{Y}'(\boldsymbol{M}_m - \boldsymbol{M}_{\boldsymbol{M}_m \boldsymbol{X}})\boldsymbol{Y},$$

for $m = 1, 2, ..., \tilde{n}$, leading C-S's second test to

$$\tilde{T}_{cs2,\tilde{n}} = \max_{1 \le m \le \tilde{n}} \left\{ \sqrt{\frac{\tilde{r}_m}{2}} \frac{\boldsymbol{Y}'(\boldsymbol{M}_m - \boldsymbol{M}_{\boldsymbol{M}_m \boldsymbol{X}})\boldsymbol{Y}/\tilde{r}_m - \hat{\sigma}^2}{\hat{\sigma}^2} \right\},$$

where $\tilde{r}_m$ denotes the rank of $C(\boldsymbol{M}_m - \boldsymbol{M}_{\boldsymbol{M}_m \boldsymbol{X}})$. With the similar but different normalizing terms

$$a_{\tilde{r}_{\tilde{n}}} = \sqrt{2 \log \log(\tilde{r}_{\tilde{n}})} \quad \text{and} \quad b_{\tilde{r}_{\tilde{n}}} = a_{\tilde{r}_{\tilde{n}}}^2 + \log a_{\tilde{r}_{\tilde{n}}} - \log(2\sqrt{2\pi}),$$

the normalized test statistic

$$\tilde{W}_{cs2,\tilde{n}} = a_{\tilde{r}_{\tilde{n}}} \tilde{T}_{2,\tilde{n}} - b_{\tilde{r}_{\tilde{n}}}$$

converges to the same extreme value distribution. In application, different small sample adjustments are made to $b_{\tilde{n}}$ and $b_{\tilde{r}_{\tilde{n}}}$, so that the tests can achieve right size even in small sample settings.

There are also other differences between C-S's test and F-H's test. For example, the estimates of variance in F-H's test and C-S's tests are different, but these estimates are all root n convergence and chosen to optimize the tests' powers. The upper bounds for the number of partial sums $\tilde{n} \in [1, ..., n]$ are also different for C-S and F-H's test. Intuitively, large $\tilde{n}$ offers more flexibilities to the test. However, in application, different $\tilde{n}$s only make mild differences. F-H's test use $\tilde{n} = n/(\log \log n)^4$, whereas C-S argued that such a small $\tilde{n}$ put too much constrains on the test, so they adopt $n/(\log \log n)^3$. By extensive simulations studies, C-S concluded that their test 1 is usually of high power and test 2 is of high power only when the fitted model is simple linear regression.

A hidden issue for both F-H's test and C-S's test is their ordering methods. The goal of ordering is to make the residual sequence $\{e_i\}$ smooth so that large Fourier coefficients concentrate on low frequencies. For simple linear regression, one can just order data increasingly according to the covariate. However, for multivariate regression, it is hard to provide an unified solution. C-S's test 1 involves fitting model (C.18) and F-H's test involves fitting model (C.17). Hence for C-S's test 1, to achieve the optimal efficiency, $\boldsymbol{\Gamma}_m$ needs to be a smooth function, while for F-H's, it requires $(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{\Gamma}_m$ to be a smooth function. In application, F-H suggested using score variation,

$$S_{FH,i} = \lambda_1(\boldsymbol{\zeta}_1'\boldsymbol{x}_i)^2 + ... + \lambda_p(\boldsymbol{\zeta}_p'\boldsymbol{x}_i)^2 = \boldsymbol{x}_i'\boldsymbol{S}\boldsymbol{x}_i.$$

While C-S considered several ordering methods, for example, using Mahalanobis distance,

$$S_{CS_1,i} = \frac{1}{\lambda_1}(\boldsymbol{\zeta}_1'\boldsymbol{x}_i)^2 + ... + \frac{1}{\lambda_p}(\boldsymbol{\zeta}_p'\boldsymbol{x}_i)^2 = \boldsymbol{x}_i'\boldsymbol{S}^{-1}\boldsymbol{x}_i$$

or jth principle component

$$S_{CS_2,j} = \frac{1}{\lambda_j}(\boldsymbol{\zeta}_1'\boldsymbol{x}_i)^2.$$

Here $\boldsymbol{S}$ is the sample covariance matrix of $\boldsymbol{X}$, with λ_j and $\boldsymbol{\zeta}_j$ as the corresponding eigenvalues and eignevectors of jth covariate. C-S argued that F-H's test should consider making $(\boldsymbol{I} - \boldsymbol{M}_{\boldsymbol{X}})\boldsymbol{\Gamma}_m$ smooth, hence their ordering may be inappropriate. However, in application, it is hard to tell which one is the better ordering strategy. One certain thing is that the ordering should be taken according to a selection of covariates that are relevant to the response variable.

C.6 Lin, Wei, Ying's method

Lin, Wei, Ying (2002) considered graphical examination of specific model assumptions, for example, functional form of covariates. Although their method is not formally lack-of-fit test, it is an interesting extension of of S-W's test.

In linear model (1.1), using the definition in Section (2.2.1), S-W's test considered a process of partial sum of residuals

$$W_{sw}(\boldsymbol{t}) = \sum_{i=1}^{n} I(\boldsymbol{x}_i \leq \boldsymbol{t})\hat{e}_i.$$

The authors pointed out the graphical checking lack-of-fit using this process is difficult for visualization. Even if considering the partial sum process respect to a specific covariate, $W_{sw}(t)$ is dominated by small covariate values and the graph of this process provides little information about the cause of the lack-of-fit. On the other hand, lowess fit of the raw residual provides information about the cause of the possible lack-of-fit, but provides little guidance on it that if the departure is large enough to reject the adequacy assumption of the proposed model. Alternatively the author considered a moving sum process respect each of the covariate,

$$W_j(t, b) = \sum_{i=1}^{n} I(t - b \leq \boldsymbol{x}_{ij} \leq t)\hat{e}_i,$$

where b is some arbitrary number and $W_j(t, b)$ can take nonzero values for t between $\min_i x_{ij} + b$ and $\max_i x_{ij}$. In general, this process represents a sum of residuals with blocks of size b. For $b = \infty$, this process is equivalent to S-W's process with respect to jth covariate. To approximate the null process, they modified S-W's bootstrap process (2.9) by replacing the indication function $I(x_{ij} < t)$ with $I(t - b < x_{ij} < t)$, that is

$$\widehat{W}_j(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i^s \hat{e}_i \left\{ I(t - b \leq x_{ij} \leq t) - \{\sum_{i=1}^{n} x_{ij} I(t - b \leq x_{ij} \leq t)\}(\sum_{i=1}^{n} x_{ij}^2)^{-1} x_{ij} \right\},$$

where Z_i^s for $i = 1, ..., n$ is a random sample from $N(0, 1)$. P values are evaluated using S-W's approach. The author claimed that if b is chosen to be roughly in the range of the lower half of the covariate values, this test is slightly more powerful than S-W's test. Since the block b is fixed before the analysis, when the data are not evenly distributed along the covariate, this moving sum process contains more variation and does not mimic the lowess fit of raw residuals. The authors then proposed a moving average process in which the number of residuals in each block is further adjusted. That is

$$W_j^a(t,b) = \frac{\sqrt{n}\sum_{i=1}^n I(t-b \leq \boldsymbol{x}_{ij} \leq t)\hat{e}_i}{\sum_{i=1}^n I(t-b \leq \boldsymbol{x}_{ij} \leq t)},$$

for

$$\min_i x_{ij} + b \leq t \leq \max_i x_{ij}.$$

This moving average process is approximated by

$$\widehat{W}_j^a(t)/n^{-1}\sum_{i=1}^n I(t-b \leq \boldsymbol{x}_{ij} \leq t).$$

P values are evaluated using the same way as before. The author concluded that this process provides more information about lack-of-fit as it closely mimic the lowess fit of the residuals.

When use these graphic methods in exploring lack-of-fit of proposed model, the author recommend generate multiple graphs for different bs, for example large b provides information about global lack-of-fit whereas a small b is sensitive in identifying the local lack-of-fit. However, for formal lack-of-fit test, the authors suggested b to be pre-specified. Theoretically, b can be any specific constant or a data-dependent quantity that become constant as $n \to \infty$. The optimal choice of b is still an open problem.

Appendix D

Linear model with singular covariance matrix

D.1 Traditional method

Chapter 4, presents an alternative method of dealing with linear models with singular covariance matrices. The most relevant approach is the traditional method which involves replacing model (1.3) with model (1.4). This traditional method is detailed explained in Christensen (2010). In this section, a brief overview of important results and steps of traditional method is presented.

There are several important statements that lead to the traditional method of analysis on models with singular covariance matrices and $C(\boldsymbol{X}) \not\subset C(\boldsymbol{V})$: (a) if $\boldsymbol{V}$ is singular and $C(\boldsymbol{X}) \subset C(\boldsymbol{V})$, estimation and test can be carried out as if $C(\boldsymbol{X}) \subset C(\boldsymbol{V})$; (b) one can find a matrix $\boldsymbol{U}$ such that $C(\boldsymbol{X}) \subset C(\boldsymbol{T})$, where $\boldsymbol{T} = \boldsymbol{X}'U\boldsymbol{X}$; (c) the BLUEs from model (1.3) and model (1.4) are equivalent.

Statement (a) is shown in Christensen (2010) Theorem 10.1.2. Whereas statement

(b) can be easily proved by showing (I). $C(\boldsymbol{X}) \subset C(\boldsymbol{T}) \Leftrightarrow \boldsymbol{T}\boldsymbol{T}^{-}\boldsymbol{X} = \boldsymbol{X}$ and (II). $\boldsymbol{T} = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{X}^{-} \Rightarrow \boldsymbol{T}\boldsymbol{T}^{-}\boldsymbol{X} = \boldsymbol{X}$.

For (I), if $C(\boldsymbol{X}) \subset C(\boldsymbol{T})$, then $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{B}$ for some $\boldsymbol{B}$. It follows that

$$\boldsymbol{T}\boldsymbol{T}^{-}\boldsymbol{X} = \boldsymbol{T}\boldsymbol{T}^{-}\boldsymbol{T}\boldsymbol{B} = \boldsymbol{T}\boldsymbol{B} = \boldsymbol{X}.$$

The other side of the proof is self-explained. Hence (I) is proved. For (II), write $\boldsymbol{T}\boldsymbol{T}^{-}\boldsymbol{T} = \boldsymbol{T}$, so that

$$(\boldsymbol{I} - \boldsymbol{T}^{-}\boldsymbol{T})\boldsymbol{T} = \boldsymbol{0} \quad \text{and} \quad (\boldsymbol{I} - \boldsymbol{T}^{-}\boldsymbol{T})\boldsymbol{T}\boldsymbol{T}^{-} = \boldsymbol{0}.$$

Multiply the above two equation together, we have

$$\begin{aligned}
\boldsymbol{0} &= (\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})\boldsymbol{T}\boldsymbol{T}^{-}\boldsymbol{T}(\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})' \\
&= (\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})\boldsymbol{T}(\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})' \\
&= (\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})' + (\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})\boldsymbol{X}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})'.
\end{aligned}$$

Since the last two term is the sum of two nonnegative definite matrices, we have $(\boldsymbol{I} - \boldsymbol{T}\boldsymbol{T}^{-})\boldsymbol{X} = \boldsymbol{0}$. Thus (II) is shown. This is the Exercise 10.3 in Christensen (2010). Case (c) is the theorem of 10.1.3 in Christensen (2010).

References

[1] Aerts, M., Claeskens, G., and Hart, J. D. (2000), Testing Lack of Fit in Multiple Regression. *Biometrika,* 87, 405-424.

[2] Amemiya, Y. (1990). On the Convergence of the Ordered Roots of a Sequence of Determinantal Equations. *Linear Algebra and its Applications,* 127, 531-542.

[3] Bargal, A.I. (1986). Smooth Tests of Fit for Censored Gamma Samples. *Communication in Statistics,* 15, 537-549.

[4] Barton, D.E. (1953). On Neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. *Scandinavian Actuarial Journal,* 36, 24-63.

[5] Barton, D.E. (1955). A form of Neyman's Φ^2 test of goodness of fit applicable to grouped and discrete data. *Scandinavian Actuarial Journal,* 38, 1-16.

[6] Barton, D.E. (1956). Neyman's Φ^2 test of goodness of fit when the null hypothesis is composite. *Scandinavian Actuarial Journal,* 39, 216-245.

[7] Christensen, R. (1989). Lack-of-Fit Tests Based on Near or Exact Replicates, *The Annals of Statistics,* 17, 673-683.

[8] Christensen R (1990). Comment on Puntanen and Styan (1989). *The American Statistician,* 44, 191-192.

[9] Christensen, R. (1991). Small-Sample Characterizations of Near Replicate Lack-of-Fit Tests. *Journal of the American Statistical Association,* 86, 752- 756.

[10] Christensen R (2010) Plane answers to complex questions: the theory of linear models. Springer, New York.

[11] Christensen, R., and Sun, S.K, (2010), Alternative Goodness-of-Fit Tests for Linear Models, *Journal of the American Statistical Association*, 105, 291-301.

[12] Darling, D. A., and Erdös, P. (1956), A Limit Theorem for the Maximum of Normalized Sums of Independent Random Variables, *Duke Mathematical Journal,* 23, 143-155.

[13] Eubank, R. and Spiegelman, C. (1990), Testing the Goodness of-Fit of a Linear Model via Nonparametric Regression Techniques. *Journal of the American Statistical Association,* 85, 387-392.

[14] Eubank, R. L., and Hart, J. D. (1992). Testing Goodness-of-Fit in Regression via Order Selection Criteria, *The Annals of Statistics,* 20, 1412-1425.

[15] Eicker, F. (1979), The Asymptotic Distribution of the Suprema of the Standardized Empirical Processes, *The Annals of Statistics,* 7, 116-138.

[16] Erdös, P., and Kac M. (1946), On Certain Limit Theorems of the Theory of probability, *Bulletin of American Mathematical Society*, 52, 292-302.

[17] Fan, J.Q., and Huang L.S. (2001), Goodness-of-Fit Tests for Parametric Regression Models, *Journal of the American Statistical Association*, 96, 640-652.

[18] Fisher, R. A. (1922). The Goodness of Fit of Regression Formulae and the Distribution of Regression Coefficients. *Journal of the Royal Statistical So- ciety,* 85, 597-612.

[19] Green, J. R. (1971), Testing Departure from a Regression, Without Using Replication, *Technometrics,* 13, 609-615.

[20] Groß, J (2004), The General Gauss-Markov Model with Possibly Singular Dispersion Matrix. *Statistical Papers,* 45, 311-336.

[21] Hamdan, M.A. (1963). The Number and Width of Classes in the Chi-square Test. *Journal of American Statistics Association,* 58, 678-689.

[22] Hamdan, M.A. (1964), A Smooth Test of Goodness-of-Fit Based on the Walsh Functions. *Austrilia Journal of Statist,* 6, 130-136.

[23] Hamdan, M.A. (1974). The Use of Orthogonal Polynomials and Orthonormal Functions in the Calculation of the Noncentrality Parameter of Chi-squared, *Communication in Statistics,* 3, 157-166.

[24] Harville DA (1981), Unbiased and Minimum-Variance Unbiased Estimation of Estimable Functions for Fixed Linear Models with Arbitrary Covariance Structure. *Annals of Statistics* 9:633-637.

[25] Harville DA (1990), Comment on Puntanen and Styan (1989). *The American Statistician,* 44, 192.

[26] Hosmer, W. D., and Hjort, N. L., (2002), Goodness-of-fit Processes for Logistic Regression: Simulation Results, *Statistics in Medicine,* 21, 2723-2738.

[27] Koul, H.L., and Stute W (1999), Nonparametric Check for Time Series, *The Annals of Statistics,* 27, 204-236.

[28] Koul, H.L., Baillie, R.T, and Surgailis, D. (2004), "Regression Model Fitting with a Long Memory Covariate Process," *Econometric Theory,* 20, 485-512.

[29] Koval′, V.A. (2002), The Law of the Iterated Logarithm for Matrix-Normed Sums of Independent Random Variables and Its Applications, *Mathematical Notes*, 72, 331-336.

[30] Kempthorne O (1989) Comment on Puntanen and Styan (1989). *The American Statistician,* 43, 161-162.

[31] Kreijger RG, Neudecker H (1977) Exact linear restrictions on parameters in the general linear model with a singular covariance matrix. *Journal of American Statistics Association,* 72, 430-432.

[32] Joglekar, G., Schuenemeyer, J. H., and LaRiccia, V. (1989). Lack-of-Fit Testing When Replicates Are Not Available. *The American Statistican,* 43, 135-143.

[33] Lin, D.Y., Wei, L.J., and Ying, Z. (2002), Model-Checking Techniques Based on Cumulative Residuals, *Biometrics,* 58, 1-12.

[34] McCullaph, P. and Nelder, J. A. (1983). Generalized Linear Models. Chapman and Hall, London.

[35] Miller, F.R., Neill, J. W., and Sherfey, B. W. (1998). Maximin Clusters for Near Replicate Regression Lack of Fit Tests. *The Annals of Statistics,* 26, 1411-1433.

[36] Neill, J. W., and Johnson, D. E. (1985). Testing Linear Regression Function Adequacy without Replication. *The Annals of Statistics,* 13, 1482-1489.

[37] Neyman, J. (1937). Smooth Test for Goodness of Fit. *Skandinavisk Aktu- arietidskrift,* 20, 149-199.

[38] Pierce, D. and Schafer, D., Residuals in Generalized Linear Models, *Journal of the American Statistical Association,* 396,977-986.

[39] Puntanen S, Scott AJ (1996) Some further remarks on the singular linear model. *Linear Algebra Application,* 237/238, 313-327.

[40] Puntanen S, Styan GPH (1989) The Equality of The Ordinary Least Squares Estimator and The Best Linear Unbiased Estimator (with discussion). *The American Statistician,* 43, 153-164.

[41] Rao CR (1967) Least Squares Theory Using an Estimated Dispersion Matrix and Its Application to Measurement of Signals. Proc Fifth Berkeley Symp 1, 355-372. University of California Press.

[42] Rao CR (1968) A Note on A Previous Lemma in The Theory of Least Squares and Some Further Results. *Scandinavian Journal of Statistics,* 30, 259-266.

[43] Rao CR, Mitra SK (1971) Generalized Inverse of Matrices and Its Applications. Wiley, New York.

[44] Rayner, J.C.W. and Best, D.J. (1986). Neyman-Type Smooth Tests For Location-Scale Families. Biometrika 73, 437-446.

[45] Rayner,J .C.W. and Best, D.J. (1988). Smooth Tests of Goodness of Fit for Regulard Istributions. *Communication in Statistics,* 17, 3235-3267.

[46] Rayner, J.C.W. and Best, D.J. (1989). Smooth Tests of Goodness of Fit. Oxford University Press, New York.

[47] Rigby, A (2009) Statistical Recommendations for Papers Submitted to Developmental Medicine and Child Neurology. *Development Medicine and Child Neurology,* 52:299-304.

[48] Shillington, E. R. (1979). “Testing Lack of Fit in Regression without Replication,” *The Canadian Journal of Statistics*, 7:137-146.

[49] Stute, W. (1997) Nonparametric Model Checks for Regression, *The Annals of Statistics,* 25:613-641.

[50] Stute, W., Thies S., and Zhu L.X. (1998), Model Checks for Regression: An Innovation Process Approach, *The Annals of Statistics,* 26, 1916-1934.

[51] Stute, W., Gonzalez Manteiga, W. and Presedo Quindimil, M. (1998), Bootstrap Approximations in Model Checks for Regression, *Journal of the American Statistical Association,* 93, 141-149.

[52] Stute, W., and Zhu, L.X. (2002), Model Checks for Generalized Linear Models, *Scandinavian Journal of Statistics,* 29, 535-545.

[53] Su, J.Q., and Wei, L.T. (1991), A Lack-of-Fit Test for the Mean Function in a Generalized Linear Model, *Journal of the American Statistical Association,* 86, 420-426.

[54] Su, Z., and Yang, S. S. (2006). A Note on Lack-of-Fit Tests for Linear Models Without Replication. *Journal of the American Statistical Association,* 101, 205-210.

[55] Sun, S.K. (2010) Alternative Goodness-of-Fit for Linear Model. Dissertation.

[56] Thomas,D .R. and Pierce, D.A. (1979), Neyman'ss Mooth Goodness-of-Fit Test When the Hypothesis is Composite. *Journal American Statistics Association,* 74, 441-445.

[57] Utts, J. M. (1982). The Rainbow Test for Lack of Fit in Regression. *Communications in Statistics,* 11, 2801-2815.

[58] Zyskind G (1967) On Canonical Forms, Non-Negative Covariance Matrices and Best and Simple Least Square Linear Estimators in Linear Models. *The Annals of Mathematical Statistics,* 38, 1092-1109.

[59] Zyskind G, Martin FB (1969) On Best Linear Estimation and General Gauss-Markov Theorem in Linear Models with Arbitrary Nonegative Covariance Structure. *SIAM Journal on Applied Mathematics,* 17, 1190-1202.