2-9-2010

Alternative goodness-of-fit tests for linear models

Siu Kei Sun

Siu Kei Sun

Candidate

Mathematics and Statistics

Department

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

Approved by the Dissertation Committee:

, Chairperson

ALTERNATIVE GOODNESS-OF-FIT TESTS FOR LINEAR MODELS

BY

SIU KEI SUN

B.Sc., Mathematics, Hong Kong University of Science and Technology, 2002
M.Sc., Statistics, Hong Kong University of Science and Technology, 2004

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 2009

Dedication

To my parents, Wilson, Amy, Tony, Barry, and Kevin

Acknowledgments

I would like to express my sincere gratitude to my dissertation advisor Dr. Ronald Christensen for his encouragement, patience, and guidance throughout my Ph.D. studies. I have to thank him deeply from my heart for bringing me into the world of linear modeling, which is the field that I am now fascinated with. I enjoyed the time working with him in these years.

I would also like to thank Dr. Edward J. Bedrick, Dr. Gabriel Huerta, and Dr. James A. Ellison. Learning in their classes was a joy to me.

A special thanks to my best friend Alvaro Nosedal-Sanchez for his support and encouragement. Alvaro helped me a lot in these years, no matter on my study or on my life in Albuquerque.

Finally, I would like to thank my parents and my brothers for their endless support and love.

ALTERNATIVE GOODNESS-OF-FIT TESTS FOR LINEAR MODELS

BY

SIU KEI SUN

ABSTRACT OF 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 2009

ALTERNATIVE GOODNESS-OF-FIT TESTS FOR LINEAR MODELS

by

SIU KEI SUN

B.Sc., Mathematics, Hong Kong University of Science and Technology, 2002

M.Sc., Statistics, Hong Kong University of Science and Technology, 2004

Ph.D., Statistics, University of New Mexico, 2009

Abstract

Fan and Huang (2001) presented a goodness-of-fit test for linear models based on Fourier transformations of the residuals of the fitted model. We present two more theoretically appealing tests in which the Fourier transforms are incorporated into a fitted model. We show that when suitably normalized, the new test statistics have the same asymptotic distribution as Fan and Huang's test. We propose modifications to the asymptotic normalization constants to improve the small sample sizes of our tests while retaining their asymptotic distributions. Small sample sizes and powers are examined via simulations. Real data of short-leaf pines from Bruce and Schumacher (1935) are used to illustrate the performance of the proposed tests.

KEY WORDS: Fourier transforms, Lack-of-fit tests, Linear models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The problem

In linear models, it is assumed that the data can be approximated by a model of the form

$$Y = X\beta + \epsilon, \tag{1.1}$$

where Y is an $N \times 1$ vector of responses, X is an $N \times p$ matrix of known covariates, β is a $p \times 1$ vector of unknown regression parameters, ϵ is an $N \times 1$ vector of random errors and it is assumed that $\epsilon \sim N(0, \sigma^2 I)$, I is an $N \times N$ identity matrix. If there is lack-of-fit in model (1.1), it indicates that the mean structure of the responses can not be well approximated by the vector $X\beta$, i.e. $E(Y) \neq X\beta$. This dissertation presents two statistical procedures for testing the lack-of-fit in linear models.

1.2 Notation

For any matrix A, let $C(A)$ be the column space of A; $r(A)$ be the rank of A; A^- be the generalized inverse of A, and $M_A \equiv A(A^TA)^-A^T$ be the perpendicular

projection operator (ppo) onto the column space $C(A)$. All models will be in numbered equations, so let $SSE(n)$ and $MSE(n)$ be the sum of squares error and the mean squares error of model (n) respectively.

1.3 Quadratic forms and central χ^2 distribution

Let M be a ppo. Y^TMY is a random variable and is called a quadratic form. Under model (1.1), Christensen (2002, Chapter 1) provides

$$\frac{Y^TMY}{\sigma^2} \sim \chi^2\left(r(M), \frac{\beta^TX^TMX\beta}{2\sigma^2}\right),$$

where $r(M)$ is the degrees of freedom of the χ^2 distribution, and $\frac{\beta^TX^TMX\beta}{2\sigma^2}$ is a noncentrality parameter. If $MX = 0$, the noncentrality parameter is 0 and the χ^2 distribution is called a central χ^2 distribution.

When $\frac{Y^TMY}{\sigma^2}$ has a central χ^2 distribution with degrees of freedom $r(M)$,

$$E\left(\frac{Y^TMY}{r(M)}\right) = \sigma^2,$$

so under model (1.1), $MX = 0$ yields $\frac{Y^TMY}{r(M)}$ an unbiased estimate to σ^2.

1.4 Central F distribution

Let M_1 and M_2 be any $N \times N$ ppo's. If $M_1M_2 = 0$, the quadratic forms Y^TM_1Y and Y^TM_2Y are independent. Moreover under model (1.1) and $M_2X = 0$, Christensen (2002, Appendix C) gives

$$\frac{Y^TM_1Y/r(M_1)}{Y^TM_2Y/r(M_2)} \sim F\left(r(M_1), r(M_2), \frac{\beta^TX^TM_1X\beta}{2\sigma^2}\right),$$

where $r(M_1)$ and $r(M_2)$ are the degrees of freedom of the F distribution, and $\frac{\beta^TX^TM_1X\beta}{2\sigma^2}$ is a noncentrality parameter of the F distribution. If $M_1X = 0$, the noncentrality parameter becomes 0 and thus $\frac{Y^TM_1Y/r(M_1)}{Y^TM_2Y/r(M_2)}$ has a central F distribution.

1.5 Outline of the dissertation

In Chapter two, brief reviews are given of two classical approaches of lack-of-fit tests in the literature.

In Chapter three, Fan and Huang's (2001) test is placed into the context of linear models. Two new lack-of-fit tests are developed. The asymptotic results of the new tests are also provided in this chapter.

Chapter four provides several estimates of the variance and introduces the adjustments to the normalizing constants used in the proposed tests. The adjustments improve the power of the tests in small samples.

Chapter five presents simulation results. Testing lack-of-fit in simple regressions and multiple regressions are discussed. Comparisons among the proposed tests and Fan and Huang's test are provided.

In Chapter six, we apply the proposed tests on a data set from Bruce and Schumacher (1935). The data contain 70 observations of short-leaf pine on their volume in cubic feet, their girth in inches, and their height in feet. Several models are used for illustrations.

Chapter seven gives a summary on the thesis. Further research directions are also provided.

Chapter 2

Review of two classical approaches of lack-of-fit testing

There are two classical approaches in the study of lack-of-fit testing in linear regressions. First, Fisher (1922) provided what has become an exact F-test based on clustering the data into groups in which the covariates are exact replications. Generalizing Fisher's test to clusters of near-replicates, Green (1971), Shillington (1979), Neill and Johnson (1985), Christensen (1989, 1991), Joglekar, Schuenemeyer, and LaRiccia (1989), and Su and Yang (2006) all proposed lack-of-fit tests. Neyman (1937) provided a classical approach to testing the goodness-of-fit of a distribution, his smooth test. This procedure has been adapted to testing lack-of-fit in regression using ideas related to nonparametric regression and model selection. Eubank and Hart (1992), Aerts, Claeskens, and Hart (2000), and Fan and Huang (2001) all proposed tests following from Neyman's approach. Brief reviews of Fisher's test and Neyman's smooth test are provided in this chapter. Review on Fan and Huang's test is given in Chapter 3. Reviews of all other approaches are provided in Appendices B and C.

2.1 Clustering: Fisher's Test

Slutsky (1913) and Pearson (1916) applied the contingency table chi-squared goodness-of-fit test, to test regression curves. However, Fisher pointed out that Pearson's formula for the degrees of freedom was incorrect. Fisher (1922) proposed to test the goodness-of-fit of regression lines using his test for the goodness-of-fit of frequencies. In his article, the proposed test statistic was first claimed to have a χ^2 distribution. Later, it was shown that the χ^2 distribution supplies only an approximation. Eventually, the test proposed by Fisher took the form of an exact F-test.

Fisher investigated goodness-of-fit of simple linear regression on the basis of covariate replications. With 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, $i = 1, \ldots, k$. Then $\sum_{i=1}^{k} n_i = N$. The model is

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}, \tag{2.1}$$

where $j = 1, \ldots, n_i$. Within each cluster, let

$$Y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, X_i = \begin{bmatrix} 1 & x_i \\ \vdots & \vdots \\ 1 & x_i \end{bmatrix}, J_{n_i} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \text{and } \epsilon_i = \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{in_i} \end{bmatrix}, \tag{2.2}$$

where Y_i, J_{n_i}, and ϵ_i are $n_i \times 1$ vectors, X_i is an $n_i \times 2$ matrix. Let

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix}, X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}, \text{and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_k \end{bmatrix}, \tag{2.3}$$

where $\epsilon \sim N(0, \sigma^2 I)$ and I is an $N \times N$ identity matrix. Model (2.1) is written in matrix notation as

$$Y = X\beta + \epsilon, \tag{2.4}$$

with regression parameters $\beta^T = [\beta_0, \beta_1]$. We start with a discussion of the original idea of Fisher's test and then the modern convention of interpreting Fisher's test.

Fisher starts the problem with the assumption of known group means of the responses μ_i and a known common variance σ^2 for all clusters. If simple regression models are fitted within each cluster, say

$$Y_i = X_i\gamma_i + \epsilon_i, \tag{2.5}$$

where $\gamma_i^T = (\gamma_{0i}, \gamma_{1i})$, for $i = 1, \ldots, k$, since $C(X_i) = C(J_{n_i})$, model (2.5) is equivalent to

$$Y_i = J_{n_i}\mu_i + \epsilon_i, \tag{2.6}$$

where μ_i is the group mean in the i-th cluster. The least squares predictors of y_{ij} from model (2.6) are the group sample means $\bar{y}_i$. Then, with $E(\bar{y}_i) = \mu_i$, Fisher standardizes the group means to $z_i = \sqrt{n_i}(\bar{y}_i - \mu_i)$, so the z_i's are i.i.d. $N(0, \sigma^2)$ random variables. The sum of z_i^2 for all clusters is $\sum_{i=1}^{k} n_i(\bar{y}_i - \mu_i)^2$ and

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

is χ^2 distributed as the test statistic of the goodness-of-fit test in contingency tables.

We want to test the lack-of-fit of model (2.4). A natural sum of squares for lack-of-fit is estimating the group mean μ_i from model (2.1), say $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, Fisher suggests that the MSE of fitting model (2.6) for all clusters simultaneously can be used as $\hat{\sigma}^2$. Hence, the test statistic is

$$\chi^2 = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\mu}_i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(N-k)},$$

which has an asymptotic χ^2 distribution with $k-1$ degrees of freedom. In modern statistics, the degrees of freedom $k-1$ is further corrected to $k-2$ and incorporated into the test statistic

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\mu}_i)^2/(k-2)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(N-k)}. \tag{2.7}$$

The test statistic has an F distribution with degrees of freedom $(k-2, N-k)$.

Let

$$Z = \begin{bmatrix} J_{n_1} & 0 & 0 & 0 & 0 \\ 0 & J_{n_2} & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & J_{n_{k-1}} & 0 \\ 0 & 0 & 0 & 0 & J_{n_k} \end{bmatrix}. \tag{2.8}$$

Z is a block diagonal matrix used in ANOVA models. The ppo M_Z is also a block diagonal matrix, i.e.

$$M_Z = \begin{bmatrix} \frac{1}{n_1} J_{n_1}^{n_1} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{n_2} J_{n_2}^{n_2} & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{n_{k-1}} J_{n_{k-1}}^{n_{k-1}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{n_k} J_{n_k}^{n_k} \end{bmatrix},$$

where $J_{n_i}^{n_i}$ is an $n_i \times n_i$ matrix of ones. Obviously

$$(I - M_Z)Y = \begin{bmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1n_1} - \bar{y}_1 \\ \vdots \\ y_{k1} - \bar{y}_k \\ \vdots \\ y_{kn_k} - \bar{y}_k \end{bmatrix}.$$

The sum of squares in the denominator in (2.7) can be written as $Y^T(I - M_Z)Y$. Hence the denomiator in (2.7) is equivalent to the MSE of fitting the model

$$Y = Z\gamma + \epsilon, \tag{2.9}$$

i.e. $MSE(2.9) = Y^T(I - M_Z)Y/(N - k)$. Moreover since $M_X Y = X\hat{\beta}$ and

$$[X\hat{\beta}]^T = [(\hat{\beta}_0 + \hat{\beta}_1 x_1)J_{n_1}^T, (\hat{\beta}_0 + \hat{\beta}_1 x_2)J_{n_2}^T, \ldots, (\hat{\beta}_0 + \hat{\beta}_1 x_k)J_{n_k}^T],$$

we have

$$M_Z Y = \begin{bmatrix} \bar{y}_1 J_{n_1} \\ \bar{y}_2 J_{n_2} \\ \vdots \\ \bar{y}_k J_{n_k} \end{bmatrix} \quad \text{and} \quad M_X Y = \begin{bmatrix} \hat{\mu}_1 J_{n_1} \\ \hat{\mu}_2 J_{n_2} \\ \vdots \\ \hat{\mu}_k J_{n_k} \end{bmatrix}.$$

These yield

$$\begin{aligned} & [(M_Z - M_X)Y]^T[(M_Z - M_X)Y] \\ = \; & \left[(\bar{y}_1 - \hat{\mu}_1)J_{n_1}^T, (\bar{y}_2 - \hat{\mu}_2)J_{n_2}^T, \cdots, (\bar{y}_k - \hat{\mu}_k)J_{n_k}^T\right] \begin{bmatrix} (\bar{y}_1 - \hat{\mu}_1)J_{n_1} \\ (\bar{y}_2 - \hat{\mu}_2)J_{n_2} \\ \vdots \\ (\bar{y}_k - \hat{\mu}_k)J_{n_k} \end{bmatrix} \\ = \; & \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\mu}_i)^2. \end{aligned}$$

Since $C(X) \subset C(Z)$, $M_Z - M_X$ is a ppo and $[(M_Z - M_X)Y]^T[(M_Z - M_X)Y] = Y^T(M_Z - M_X)Y$. The numerator in (2.7) can be expressed as $Y^T(M_Z - M_X)Y/(k-2)$. Hence, (2.7) can be written as

$$F = \frac{Y^T(M_Z - M_X)Y/(k-2)}{Y^T(I - M_Z)Y/(N-k)}, \tag{2.10}$$

which is the classical F statistic used in lack-of-fit test when replicates are available.

A recent interpretation of the exact F-test is provided in Christensen (2002, Chapter 6). Since the rows in X are replicated with k clusters, $C(X) \leq k$. The design matrix Z of model (2.9) has the same row structure as X and achieves the largest possible rank of $C(X)$, i.e. $C(X) \subset C(Z)$ and $C(Z) = k$. Therefore, model (2.9) is regarded as the most general model with the same pattern of equal means that can be generated from model (2.4). If there exists a lack-of-fit in model (2.4) but the means remain constant within clusters, model (2.9) should give a better fit to the data. $SSE(2.9) = Y^T(I - M_Z)Y$ is called the sum of squares for pure error. The difference

$$SSE(2.4) - SSE(2.9) \;\; = \;\; Y^T(I - M_X)Y - Y^T(I - M_Z)Y$$

$$= \quad Y^T(M_Z - M_X)Y$$

is called the sum of squares for lack-of-fit. Fisher's exact F-test can be used in multiple regressions with k clusters of exact replicates by replacing $k-2$ with $k - r(X)$.

2.2 Smooth test: Neyman's Smooth Test

The use of the smooth test for goodness-of-fit was proposed by Neyman (1937). Neyman's smooth test was not designed for testing lack-of-fit in linear regression, but his idea applies. As mentioned earlier, when replicates are available, a most general model can be generated for testing lack-of-fit. If replicates are not available, no most general model exists. The smooth test proposes an alternative model that is more general than (2.4). Define

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \text{ and } H = \begin{bmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_n) \end{bmatrix}, \tag{2.11}$$

where $\epsilon \sim N(0, \sigma^2 I)$ and $h : \mathbb{R} \to \mathbb{R}$ is an unknown smooth function. Smooth functions are functions that can be differentiated infinitely many times. For testing lack-of-fit in the simple linear regression

$$Y = X\beta + \epsilon, \tag{2.12}$$

a more general model with a smooth function is

$$Y = X\beta + H + \epsilon. \tag{2.13}$$

The function $h(x)$ is an arbitrary smooth function in x. $h(x)$ can be described by a series expansion, i.e.

$$h(x) = \sum_{t=0}^{\infty} \theta_t \varphi_t(x), \tag{2.14}$$

where θ_t are unknown coefficients and the functions $\varphi_t(x)$ are known, fixed, and referred to as basis functions. Efromovich (1999, Chapter 2) gives some choices for $\varphi_t(x)$. In the series expansion of $h(x)$, infinite numbers of coefficients are involved. It is impossible to deal with infinite coefficients in practice. Therefore, a partial sum is used to approximate $h(x)$, i.e.

$$h_k(x) = \sum_{t=0}^{k} \theta_t \varphi_t(x), \tag{2.15}$$

for some integer k. Define

$$H_k = \begin{bmatrix} \varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_k(x_1) \\ \varphi_0(x_2) & \varphi_1(x_2) & \cdots & \varphi_k(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \cdots & \varphi_k(x_n) \end{bmatrix} \text{ and } \gamma_k = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}. \tag{2.16}$$

Model (2.13) can be approximated by

$$Y = X\beta + H_k\gamma_k + \epsilon. \tag{2.17}$$

Using model (2.17), we extend Neyman's smooth test to testing lack-of-fit in linear regression. As $C(X) \subset C(X, H_k)$, we propose a natural F-statistic for testing lack-of-fit of model (2.12), i.e.

$$F_k = \frac{Y^T(M_{H'_k})Y/r(H'_k)}{Y^T(I - M_X - M_{H'_k})Y/(n - r(X) - r(H'_k))}, \tag{2.18}$$

where $H'_k \equiv (I - M_X)H_k$ and F_k has a central F distribution with degrees of freedom $(r(H'_k), n - r(X) - r(H'_k))$ when (2.12) is true. Neyman suggests that in the smooth test, one fixed k should be used. If k is chosen depending on the data, for example we could choose the k that gives the smallest p-value among several k's, a new critical region must be introduced rather than using the critical region from the F distribution with degrees of freedom appropriate to the chosen k.

Eubank and Hart (1992), Aerts, Claeskens, and Hart (2000), and Fan and Huang (2001) all proposed tests by extending Neyman's smooth test to linear regression. Although they did not mention in their articles, all of their tests can

be interpreted by comparing models (2.12) and (2.17), with different criteria for choosing k, and hence different critical regions.

Chapter 3

Two new approaches to testing Lack-of-Fit

In this chapter, we place Fan and Huang's test into a linear model theory context that suggests two potential improvements. Section 3.1 provides a brief introduction to Fan and Huang's test. In Section 3.3 and 3.4, we apply the Darling-Erdős (1956) theorem to obtain the asymptotic distribution of our proposed test statistics.

3.1 Fan and Huang's Test:

Fan and Huang, henceforth referred to as FH, proposed a lack-of-fit test based on Fourier transforms. The null linear model is

$$y_i = x_i^T \beta + \epsilon_i,$$

$i = 1, \ldots, n$, where y_i is the dependent variable, x_i is a $p \times 1$ vector of known covariates, β is a $p \times 1$ vector of fixed unknown regression parameters, and the ϵ_i are independent $\epsilon_i \sim N(0, \sigma^2)$. FH compare this to a more general model

$$y_i = x_i^T \beta + h(x_{ij}) + \epsilon_i,$$

where x_{ij} is the jth component of x_i and $h(\cdot)$ is an arbitrary unknown smooth function.

Let

$$Y \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad X \equiv \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \equiv [\, X_1 \cdots X_p \,].$$

In matrix form the models are

$$Y = X\beta + \epsilon \tag{3.1}$$

and

$$Y = X\beta + H(X_j) + \epsilon, \tag{3.2}$$

respectively, with $H(X_j) \equiv [h(x_{1j}), \ldots, h(x_{nj})]^T$ and $\epsilon \sim N(0, \sigma^2 I_n)$.

FH use Fourier transforms. Define an $n \times 1$ vector $\psi_1 = [1, \ldots, 1]^T$. For any positive integer q, let

$$\psi_{2q} = \left[\cos\left(2\pi q \frac{1}{n}\right), \ldots, \cos\left(2\pi q \frac{n}{n}\right)\right]^T$$

and

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

Define $\Psi = [\psi_1, \ldots, \psi_n]$. Let Γ be the $n \times n$ orthogonal matrix generated by normalizing the columns of Ψ to have length one. The matrix Γ defines the discrete Fourier transformation, that is, for any vector w the Fourier transform is

$$w^* \equiv \Gamma^T w.$$

We will have occasion to use submatrices of Γ. For an even number m, let

$$\Psi_m = [\psi_2, \ldots, \psi_{m+1}]$$

and normalize the columns to give the corresponding submatrix of Γ, say, Γ_m. The vector ψ_1 is eliminated because it is just a vector of 1s and, for models with

an intercept, its use will be redundant. Γ_m will contain $m/2$ pairs of sine and cosine terms. Note that $\Gamma_m^T \Gamma_m = I_m$ and define

$$M_m \equiv \Gamma_m \Gamma_m^T = M_{\Gamma_m}.$$

For any vector w define a component vector of the Fourier transform

$$w_m^* = \Gamma_m^T w.$$

FH propose a test for lack of fit based on the Fourier transform of the model (3.1) residuals $\hat{\epsilon} \equiv Y - X\hat{\beta} = (I - M_X)Y$, that is,

$$\hat{\epsilon}^* = \Gamma^T \hat{\epsilon} \equiv \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{bmatrix}.$$

Letting $\hat{\sigma}^2$ be a consistent estimate of σ^2 with $\hat{\sigma}^2 = \sigma^2 + O_p(n^{-1/2})$, FH define the test statistic

$$T_{FH} = \max_{1 \le m \le \tilde{n}} \frac{1}{\sqrt{2m}\hat{\sigma}^2} \sum_{i=1}^{m} (v_i^{*2} - \hat{\sigma}^2).$$

FH use $\tilde{n} \equiv n$ to define their test but need to redefine $\tilde{n}$ to obtain asymptotic results. Note that with an intercept in the model, $v_1^* = 0$. The test statistic T_{FH} is normalized to

$$W_{FH} = a_{\tilde{n}} T_{FH} - b_{\tilde{n}},$$

where $a_{\tilde{n}} = \sqrt{2 \log \log \tilde{n}}$ and $b_{\tilde{n}} = a_{\tilde{n}}^2 + \log a_{\tilde{n}} - \log(2\sqrt{2\pi})$. Fan and Huang show that under model (3.1),

$$P(W_{FH} < x) \to \exp(-\exp(-x)) \qquad \text{as} \qquad n \to \infty. \tag{3.3}$$

3.2 Linear models and the FH test

FH's test is based on $\Gamma^T(Y - X\hat{\beta})$. In particular, it is based on finding the sum of squares for regression in

$$Y - X\hat{\beta} = \Gamma_m \gamma_m + e. \tag{3.4}$$

To see this note that the total sum of squares for model (3.4) is $SSE(3.1)$ and

$$\begin{aligned} SSE(3.4) &= \hat{\epsilon}^T(I - M_m)\hat{\epsilon} \\ &= \hat{\epsilon}^T\hat{\epsilon} - \hat{\epsilon}^T M_m \hat{\epsilon} \\ &= SSE(3.1) - \left[\Gamma_m^T\hat{\epsilon}\right]^T\left[\Gamma_m^T\hat{\epsilon}\right] \\ &= SSE(3.1) - \left[\hat{\epsilon}_m^*\right]^T\left[\hat{\epsilon}_m^*\right], \end{aligned}$$

so $\left[\hat{\epsilon}_m^*\right]^T\left[\hat{\epsilon}_m^*\right]$ is the sum of squares for regression in model (3.4).

To see that FH use $\left[\hat{\epsilon}_m^*\right]^T\left[\hat{\epsilon}_m^*\right]$ to test the adequacy of model (3.1), rewrite

$$\begin{aligned} T_{FH} &= \max_{1\le m\le\tilde{n}} \frac{1}{\sqrt{2m}\hat{\sigma}^2}\sum_{i=1}^{m}(v_i^{*2} - \hat{\sigma}^2) \\ &= \max_{1\le m\le\tilde{n}} \frac{1}{\sqrt{2m}}\frac{\sum_{i=1}^{m} v_i^{*2} - m\hat{\sigma}^2}{\hat{\sigma}^2} \\ &= \max_{1\le m\le\tilde{n}} \sqrt{\frac{m}{2}}\frac{\sum_{i=1}^{m} v_i^{*2}/m - \hat{\sigma}^2}{\hat{\sigma}^2} \\ &= \max_{1\le m\le\tilde{n}} \sqrt{\frac{m}{2}}\frac{\left[\hat{\epsilon}_m^*\right]^T\left[\hat{\epsilon}_m^*\right]/m - \hat{\sigma}^2}{\hat{\sigma}^2}. \end{aligned}$$

3.3 Approach one: Model comparison

Fitting model (3.4) involves using a two-stage fitting procedure to fit

$$Y - X\beta = \Gamma_m\gamma_m + \epsilon, \tag{3.5}$$

wherein β is first estimated from model (3.1) and then γ_m is fitted to model (3.4). However, model (3.5) can be fitted directly. Clearly, model (3.5) is equivalent to

$$Y = X\beta + \Gamma_m\gamma_m + \epsilon \tag{3.6}$$

and using results from analysis of covariance, for example Christensen (2002, Chapter 9), rewrite model (3.6) as

$$Y = X\beta_0 + (I - M_X)\Gamma_m\gamma + \epsilon, \tag{3.7}$$

where $\beta_0 \equiv \beta + (X^T X)^- X^T \Gamma_m \gamma$. The sum of squares error of models (3.6) and (3.7) are

$$\begin{aligned} SSE(3.6) &\equiv Y^T(I - M_X - M_{(I-M_X)\Gamma_m})Y \\ &= Y^T(I - M_X)Y - Y^T M_{(I-M_X)\Gamma_m} Y \\ &= SSE(3.1) - Y^T M_{(I-M_X)\Gamma_m} Y. \end{aligned}$$

When testing lack-of-fit by testing model (3.1) versus model (3.6), the sum of squares lack of fit is $Y^T M_{(I-M_X)\Gamma_m} Y$, so our first proposal is to replace $[\hat{\epsilon}_m^*]^T [\hat{\epsilon}_m^*]$ in FH's test by $Y^T M_{(I-M_X)\Gamma_m} Y$. Our first test statistic is

$$\widehat{T}_{1,\,\widetilde{n}} = \max_{1 \le m \le \widetilde{n}} \left\{ \sqrt{\frac{r_m}{2}} \frac{Y^T M_{(I-M_X)\Gamma_m} Y / r_m - \hat{\sigma}^2}{\hat{\sigma}^2} \right\},$$

where $r_m \equiv r\left[(I - M_X)\Gamma_m\right]$ and $\hat{\sigma}^2 = \sigma^2 + O_p(n^{-1/2})$.

FH define their test statistic with m ranging from 1 to n but in their proof of the asymptotic null distribution of T_{FH}, they use m ranging from 1 to $\widetilde{n} = \dfrac{n}{(\log\log n)^4}$. FH claim that the reduction on the range of m has little impact on the performance of their test statistic. We define our tests to agree with our asymptotic results which use

$$\widetilde{n} \equiv \left\lceil \frac{n}{(\log\log n)^{1+\delta}} \right\rceil \tag{3.8}$$

for $\delta > 0$.

As with FH, we appeal to the Darling-Erdős Theorem (Darling and Erdős, 1956) to show that the normalized test statistic converges to an extreme value distribution. We normalize our first test statistic as

$$\widehat{W}_{1,\,\widetilde{n}} = a_{r_{\widetilde{n}}} \widehat{T}_{1,\,\widetilde{n}} - b_{r_{\widetilde{n}}},$$

where $a_{r_{\widetilde{n}}} = \sqrt{2 \log\log r_{\widetilde{n}}}$ and $b_{r_{\widetilde{n}}} = a_{r_{\widetilde{n}}}^2 + \log a_{r_{\widetilde{n}}} - \log(2\sqrt{2\pi})$.

Theorem 1. *If* $\dfrac{\hat{\sigma}^2}{\sigma^2} - 1 = O_p\left(n^{-1/2}\right)$, *then*

$$Pr(\widehat{W}_{1,\,\widetilde{n}} < x) \to \exp(-\exp(-x)) \qquad as \quad n \to \infty.$$

The proof is given in the Appendix A at the end of the thesis.

3.4 Approach two: Direct estimation of $\epsilon_m^* \equiv \Gamma_m^T \epsilon$

FH estimated $\Gamma_m^T(Y - X\beta) \equiv \Gamma_m^T \epsilon$ using the least squares $\hat{\beta}$, to obtain $\hat{\epsilon}_m^* \equiv \Gamma_m^T \hat{\epsilon} = \Gamma_m^T(Y - X\hat{\beta})$. We propose an alternative method of direct estimation.

The goal is estimating $\epsilon_m^* \equiv \Gamma_m^T \epsilon$. Multiplying model (3.1) on the left by Γ_m^T gives

$$\Gamma_m^T Y = \Gamma_m^T X\beta + \Gamma_m^T \epsilon. \tag{3.9}$$

We can estimate $\Gamma_m^T \epsilon$ directly by using the least-square residuals from model (3.9), i.e.,

$$\tilde{\epsilon}_m^* \equiv (I_m - M_{\Gamma_m^T X})\Gamma_m^T Y.$$

Tests are based on the sum of squares,

$$\begin{aligned}
[\tilde{\epsilon}_m^*]^T [\tilde{\epsilon}_m^*] &= \left[(I_m - M_{\Gamma_m^T X})\Gamma_m^T Y\right]^T \left[(I_m - M_{\Gamma_m^T X})\Gamma_m^T Y\right] \\
&= Y^T \Gamma_m (I_m - M_{\Gamma_m^T X})\Gamma_m^T Y \\
&= Y^T (M_m - M_{M_m X}) Y. \qquad (3.10)
\end{aligned}$$

For $m = 1, 2, \ldots, \widetilde{n}$, let $\widetilde{r}_m$ denote the rank of $C(M_m - M_{M_m X})$. Our second proposed test statistic is

$$\widetilde{T}_{2,\widetilde{n}} = \max_{1 \le m \le \widetilde{n}} \left\{ \sqrt{\frac{\widetilde{r}_m}{2}} \frac{Y^T (M_m - M_{M_m X}) Y / \widetilde{r}_m - \hat{\sigma}^2}{\hat{\sigma}^2} \right\}.$$

This procedure has connections to Shillington's (1979) and Christensen's (1991) tests. The sum of squares $Y^T(M_m - M_{M_m X})Y$ is the difference between the sum of squares errors in the models:

$$\begin{aligned}
Y &= M_m X\beta + \epsilon, \\
Y &= \Gamma_m \gamma_m + \epsilon.
\end{aligned}$$

We normalize the test statistic as

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

where $a_{\widetilde{r}_{\widetilde{n}}} = \sqrt{2 \log\log(\widetilde{r}_{\widetilde{n}})}$ and $b_{\widetilde{r}_{\widetilde{n}}} = a_{\widetilde{r}_{\widetilde{n}}}^2 + \log a_{\widetilde{r}_{\widetilde{n}}} - \log(2\sqrt{2\pi})$.

Theorem 2. *If* $\dfrac{\hat{\sigma}^2}{\sigma^2} - 1 = O_p\left(n^{-1/2}\right)$, *then*

$$Pr(\widetilde{W}_{2,\tilde{n}} < x) \to \exp(-\exp(-x)) \qquad as \quad n \to \infty.$$

The proof is in the Appendix A.

Chapter 4

Estimating σ^2 and Adjustments

In Section 4.1, we discuss choices of $\hat{\sigma}^2$ for our proposed test statistics. In Section 4.2, we examine adjustments to the normalizing constants of the Darling-Erdős theorem to improve small sample behavior.

4.1 Estimating σ^2

We have not specified $\hat{\sigma}$ in any of the test statistics although the asymptotics require that $\hat{\sigma}^2$ satisfy $\hat{\sigma}^2 = \sigma^2 + O_p(n^{-1/2})$. FH suggest using the sample variance of $\{\hat{\epsilon}_i^*, i = K+1, \ldots, n\}$, that is, they use

$$\hat{\sigma}_0^2 = \frac{1}{n-K} \sum_{i=K+1}^{n} \hat{\epsilon}_{ni}^{*2} - \left\{ \frac{1}{n-K} \sum_{i=K+1}^{n} \hat{\epsilon}_{ni}^{*} \right\}^2,$$

where $\hat{\epsilon}_{ni}^*$ is the i-th entry of $\hat{\epsilon}_n^* \equiv \Gamma_n^T \hat{\epsilon}$, and $K \equiv K(n)$ is a constant that depends on n. FH use $K = [n/4]$ in their simulations. We use $\widetilde{n}$ from (3.8) instead of n in computing FH's test statistic, so we use $K = [\widetilde{n}/4]$. From simulation we found that FH's test works better when computed using $\widetilde{n}$ rather than n.

A natural estimate of σ^2 for use with $\hat{T}_{1,\tilde{n}}$ would be SSE(3.6) with $m = \tilde{n}$. However, in small samples this may not provide enough degrees of freedom. Another natural estimate is using the MSE of an intermediate model between model

(3.1) and model (3.6) with $m = \tilde{n}$, say,

$$Y = X\beta + \Gamma_K \gamma_K + \epsilon, \tag{4.1}$$

for $0 \leq K \leq \widetilde{n}$. Define

$$\hat{\sigma}_1^2 \equiv MSE(4.1) = \frac{Y^T(I - M_{X,\Gamma_K})Y}{n - r(X) - r_K}.$$

A natural variance estimate for $\hat{T}_{2,\tilde{n}}$ follows by analogy to Christensen (1991). Note that a one-way analysis of variance model $y_{ij} = \mu_i + \epsilon_{ij}$ can be written in matrix notation as

$$Y = Z\mu + \epsilon, \tag{4.2}$$

where Z, as defined in (2.8), is a design matrix providing a specific clustering to the responses. Christensen (1989) used the model

$$Y = X\beta + Z\mu + \epsilon \tag{4.3}$$

as the full model in order to construct an F-statistic for testing lack-of-fit. Christensen (1991) used the ideas of within-clusters orthogonal lack-of-fit and between-clusters orthogonal lack-of-fit to define and derive optimal tests. The orthogonal lack-of-fit space was characterized by writing the perpendicular projection operator onto $C(X)^\perp$ as the sum of mutually orthogonal projection operators involving the between-cluster $C(X)^\perp \cap C(Z)$ and within-cluster $C(X)^\perp \cap C(Z)^\perp$ spaces. The sum of squares lack-of-fit in the Christensen (1991) test has ppo $M_Z - M_{M_Z X}$ onto the column space $C(X)^\perp \cap C(Z)$ and the optimal variance estimate uses the ppo $(I - M_X) - (M_Z - M_{M_Z X})$.

A similar characterization of the orthogonal lack-of-fit space can be applied by replacing Z with Γ_K. The sum of squares for lack-of-fit in our second test is given in (3.10), which has the same structure as the Christensen (1991) test. With $m = K$, the corresponding estimate of σ^2 for our second test is

$$\hat{\sigma}_2^2 \equiv \frac{Y^T\left[(I - M_X) - (M_K - M_{M_K X})\right]Y}{n - r(X) - \widetilde{r}_K},$$

The most natural choice of K seems to be $K = \widetilde{n}$ but that may not be a good choice, especially for small samples. Although picking $K = \tilde{n}$ in model (4.1) in large samples should give the best approximation to the true model and the best estimate of σ^2, for smaller samples it may over fit the data. For small samples $\widetilde{n}$ is close to n so it may provide insufficient degrees of freedom for estimating σ^2. From Lemma 3 below, any K between 0 and $\widetilde{n}$ suffices to give the asymptotic null distribution on which the tests are based. Moreover, the intuition that suggests picking $K = \widetilde{n}$ is based on having $\hat{\sigma}^2$ independent of the constructs $W_{1,\widetilde{n}}$ and $W_{2,\widetilde{n}}$ used in the large sample proof, but independence is not a particularly relevant consideration after using $\hat{\sigma}^2$ to construct $\widehat{W}_{1,\widetilde{n}}$ and $\widetilde{W}_{2,\widetilde{n}}$.

The asymptotic distributions require $\hat{\sigma}^2 = \sigma^2 + O_p(n^{-1/2})$.

Lemma 3. *If* $\dfrac{K}{n} \to c$ *as* $n \to \infty$, *where* $0 \le c < 1$, *then* $\hat{\sigma}_i^2 = \sigma^2 + O_p(n^{-1/2})$ *for* $i = 1, 2$ *under model (3.1).*

The proof is in the Appendix A.

4.2 Adjustments to the test statistics

This section presents an adjustment to the test statistics that maintains their asymptotic distribution while improving sizes of the tests in small samples. In

$$\widehat{T}_{1,\widetilde{n}} = \max_{1 \le m \le \widetilde{n}} \left\{ \sqrt{\frac{r_m}{2}} \frac{Y^T M_{(I-M_X)\Gamma_m} Y / r_m - \hat{\sigma}^2}{\hat{\sigma}^2} \right\}.$$

we select the maximum from $\widetilde{n}$ objects. With $r_{\widetilde{n}} < \widetilde{n}$, there may be repetitions in these $\widetilde{n}$ objects. Moreover, forcing the $\sin(\cdot)$ and $\cos(\cdot)$ terms to appear in pairs reduces the number of terms computed by half. Both considerations seem to harm the rate of convergence to the asymptotic distribution of the test statistic. Our simulation results indicate slightly inflated test sizes in small samples relative to the asymptotic distribution.

The normalizing constants used in the Darling-Erdős theorem come from Lemma 3.10 in Darling and Erdős (1956). The lemma provides that if $2\mu(T)y = \log n$, where $y > 0$ and $\mu(T) = \dfrac{(2\pi)^{1/2}\exp\{\alpha^2/2\}}{\alpha}$, then for $n \to \infty$

$$T = (2\log\log n)^{1/2} + \frac{\log\log\log n}{2(2\log\log n)^{1/2}} - \frac{\log((4\pi)^{1/2}y)}{(2\log\log n)^{1/2}} + o\left((\log\log n)^{-1/2}\right).$$

T is the asymptotic solution to the equation $2\mu(T)y = \log n$. T also plays the role of our test statistics before standardization. We express the asymptotic result of the Darling-Erdős theorem in a different way, so we replace $\log((4\pi)^{1/2}y)$ by $-W + \log 2\sqrt{\pi}$ where W is the value our test statistics take after standardization. Thus

$$T = (2\log\log n)^{1/2} + \frac{\log\log\log n}{2(2\log\log n)^{1/2}} - \frac{\log 2\sqrt{\pi} - W}{(2\log\log n)^{1/2}} + o\left((\log\log n)^{-1/2}\right).$$

After some manipulation

$$\begin{aligned} W &= \sqrt{2\log\log n}\,T - \left(2\log\log n + \frac{\log\log\log n}{2} - \log(2\sqrt{\pi})\right) + o(1) \\ &\equiv a_n T - b_n + o(1). \end{aligned}$$

The simulated results indicate that if we keep this a_n, a larger b_n should be used in small samples. Furthermore, the adjustment should be larger as n is increased. The normalizing constant b_n involves the term $\log\log\log n$. We implement the adjustment on b_n by raising the power of n in $\log\log\log n$ to $c\log\log n$, where c is a constant and depends on the choice of K in estimating $\hat{\sigma}^2$. This defines a new constant $\bar{b}_n = 2\log\log n + \frac{\log\log\log\left(n^{c\log\log n}\right)}{2} - \log(2\sqrt{\pi})$. Since

$$\begin{aligned} & \lim_{n\to\infty}\left\{\log\log\log\left(n^{c\log\log n}\right) - \log\log\log n\right\} \\ = & \lim_{n\to\infty}\log\left\{\frac{\log\left((c\log\log n)(\log n)\right)}{\log\log n}\right\} \\ = & \lim_{n\to\infty}\log\left\{\frac{\log c + \log\log\log n + \log\log n}{\log\log n}\right\} \\ = & \ 0, \end{aligned}$$

$b_n - \bar{b}_n \to 0$ as $n \to \infty$, and the asymptotic distribution of W is unaffected by this adjustment. In applications to Theorems 1 and 2, n is replaced by $r_{\tilde{n}}$ and $\widetilde{r}_{\tilde{n}}$, respectively.

It is worth noting that FH expressed the sum of squares lack-of-fit as the sum of $\widetilde{n}$ asymptotically independent one degree of freedom χ^2 random variables. Our tests expressed the sum of squares lack-of-fit as the sum of exact small sample independent one degree of freedom χ^2 random variables. This difference constitutes, in small samples, a hidden adjustment implemented in FH's test statistics.

Chapter 5

Simulations

In this chapter, simulations are used to examine the empirical sizes and powers of our proposed tests, along with FH's test. The empirical powers of an exact F-test based on models (3.1) and (3.6) with $m = \dfrac{\widetilde{n}}{2}$ as the full model are also given for comparison. The simulations are based on models from FH or slight modifications. The results are based on 20000 simulations and the significance level is taken to be 5%. The sample size is $n = 64$ and the samples are ordered by a covariate. As mentioned earlier, the matrix Γ_m^T contains rows of $\cos(\cdot)$'s and $\sin(\cdot)$'s with identical arguments that appear in pairs.

Simulations not reported here compared all three test statistics as computed with all three estimates of σ^2, and a variety of sample sizes. Although there are natural relationships between the variance estimates and test statistics, we report the results of our proposed tests using $\hat{\sigma}_1^2$ as the estimate of σ^2. For FH, $\hat{\sigma}_0^2$ is used. It is worth noting that FH with $\hat{\sigma}_0^2$ provides better empirical powers than with $\hat{\sigma}_1^2$. Figures obtained from these simulations with the test statistics computed based on the true σ^2 are also provided as reference. In this case, the F-test is replaced by a χ^2-test.

We found that for these tests $\hat{\sigma}_1^2$ is a better estimator of σ^2 than $\hat{\sigma}_2^2$ when K is chosen to be $[\widetilde{n}/4]$, especially in small samples. When the fitted model is simple

linear regression, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ provide similar empirical sizes and powers in all tests. The difference becomes substantial in multiple regression, wherein the tests with $\hat{\sigma}_2^2$ provide relatively low powers.

FH's $\hat{\sigma}_0^2$ is computed from the sample variance of transformed residuals and their test statistic is built on the transformed residuals, so $\hat{\sigma}_0^2$ is a convenient choice of $\hat{\sigma}^2$ for them. Our test statistics are not constructed directly from residuals. Although the sum of squares lack-of-fit can be decomposed into residuals, this is not the genesis of our test statistics. Our simulated results indicate that, in multiple regression, the size of FH's test is far below 5% when the true σ^2 is used. But the problem of undersizing is ameliorated by the use of their estimate $\hat{\sigma}_0^2$. Thus $\hat{\sigma}_0^2$ may not be a good estimator for σ^2 but is a good choice for FH's test. When the true σ^2 is used, undersizing in our tests is not as serious as that in FH's test. Therefore, $\hat{\sigma}_0^2$ is not recommended for our proposed tests, it would make our tests oversized.

FH use $\widetilde{n} = \dfrac{n}{(\log\log n)^4}$ in the proof of the asymptotic distribution of their test statistic but $\widetilde{n} = n$ in their simulations. Not only does $\widetilde{n} = n$ violate the theory but we found that it gave relatively poor simulated results. Since $\widetilde{n} = \dfrac{n}{(\log\log n)^4}$ induces a relatively large reduction on the sample size, we used $\widetilde{n} = \dfrac{n}{(\log\log n)^2}$. All examples use this $\widetilde{n}$.

As mentioned earlier, the c in our proposed adjustment $\bar{b}_n$ depends on the choice of K. Moreover, the c's used in our first and second proposed tests are different. Define

$$\Omega(x) = \begin{cases} 0 & \text{if } x \leq 35 \\ \exp\left\{-\dfrac{1}{(x-35)^{0.1}}\right\} & \text{if } x > 35 \end{cases}.$$

We suggest $K = [\widetilde{n}/10]$ with $c = 2.1$ for our first proposed test and $K = [\widetilde{n}/4]$ with

$$c = \frac{11}{\log(r - 0.9)}\left(1 - \Omega\left(\frac{\widetilde{r}_{\widetilde{n}}}{r}\right)\right),$$

where $r \equiv r(X)$ and $\widetilde{r}_m$ is the rank of $C(M_m - M_{M_m X})$, $M_m \equiv M_{\Gamma_m}$, for our second proposed test. We use this c for Test 2 because when $\widetilde{r}_{\widetilde{n}}$ is much larger than r, the adjustment $\dfrac{11}{\log(r-0.9)}$ becomes too large for the test statistic. The constant 35 is obtained from extensive computer experiments. $K = [\widetilde{n}/4]$ is used in FH's test. Since our proposed adjustment depends on the sample size n, we provide the empirical sizes of the tests based on 5000 simulations and various sample sizes, $n = 48, 64, 80, 96, 128, 160, 192, 224, 256,$ to illustrate the little impact from the sample size on the sizes of our proposed tests under the adjustment.

For simplicity, we call our first proposed test "Test 1" our second proposed test "Test 2"; FH's test "FH"; and the exact F-test with $m = \widetilde{n}/2$ "F".

5.1 Simple regression model

5.1.1 Brief outline

We start with eight examples that test lack-of-fit when fitting linear models with an intercept and one predictor, i.e. $y_i = \beta_0 + x_{1i}\beta_1 + \epsilon_i$. The ϵ_i are i.i.d. $N(0, \sigma^2)$ random variables. The true model changes in each example. In the simulations, we choose $\sigma = 2$ instead of $\sigma = 1$ as in Fan and Huang because the larger variation amplifies the differences between the empirical powers among tests.

In Examples 1 and 2, the response variables are drawn from two models, respectively, such that the inverse of $E(Y)$ is linear in the parameter. The empirical powers for all four tests show different patterns in these examples but share the same property that a critical point is shown at the regression parameter $\theta = 1$.

Examples 3 and 4 present simulations when the response variables are drawn from even-order polynomials of x_1: a quadratic polynomial and a forth-order polynomial, respectively. The effect of raising the highest order term in the even-order polynomial on the empirical powers of the tests will be discussed.

The setup of the simulations in Examples 5 and 6 is similar to Examples 3 and 4 but the response variables are drawn from odd-order polynomials of x_1: a cubic polynomial and a fifth-order polynomial, respectively. The difference in the empirical powers of the tests between odd-order polynomial models and even-order polynomial models will be discussed via a comparison between Examples 3 and 6.

In Example 7, the response variable is drawn from a simple regression but we put a cosine transformation on the predictor. A comparison between testing lack-of-fit when fitting a simple linear regression and when fitting the polynomial regression, i.e. $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + \epsilon_i$, is provided.

In Example 8, a model which is not linear in the parameter is used to simulate the response variable. In addition to testing lack-of-fit when fitting a simple linear regression, we also consider fitting a model in which $C(X)$ contains the low frequency terms in $\Gamma_m \equiv [G_1, G_2, \ldots, G_m]$, i.e.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 g_{1i} + \beta_3 g_{2i} + \beta_4 g_{3i} + \beta_5 g_{4i} + \epsilon_i, \tag{5.1}$$

where g_{ji} is the i-th entry in G_j, is used in the comparison.

The larger models included in Examples 7 and 8 illustrate the price that FH pays for not adjusting Γ_m for the fitted X when $C(X)$ is adept at picking up low frequency terms in Γ_m.

5.1.2 Examples

Example 1. In this example, the predictor variable x_1 is sampled from a $N(0, 1)$. The response variable y is drawn from the model

$$y = \frac{10}{1 + \theta \exp(-2x_1)} + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.2}$$

The true model has a logistic structure when the regression parameter $\theta \neq 0$. Figure 5.1 provides empirical powers for all four tests based on $n = 64$ and the

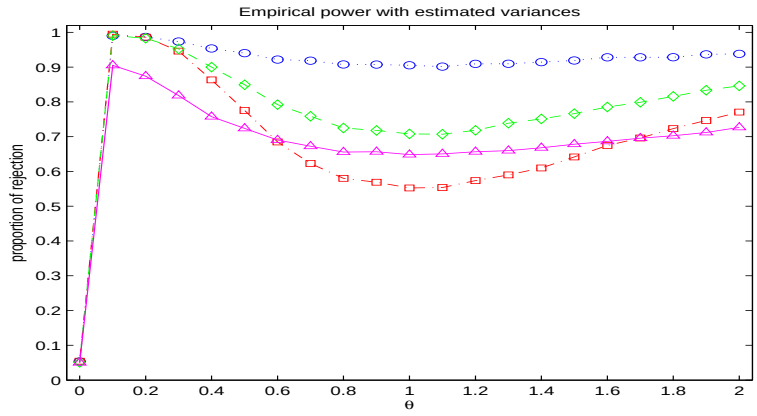

Figure 5.1: Powers for Example 1. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

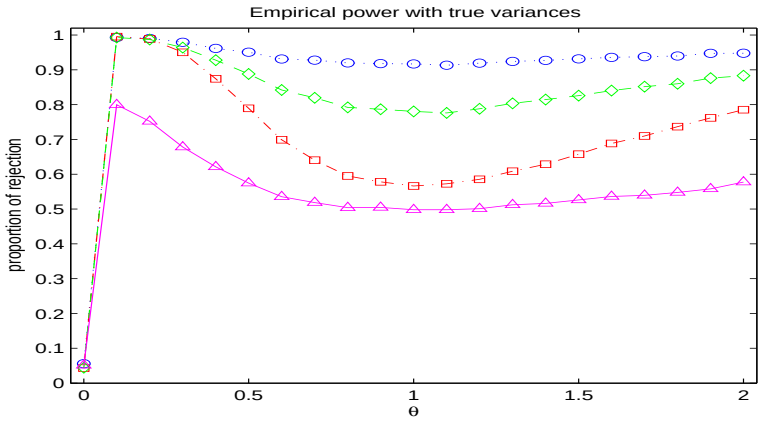

Figure 5.2: Powers for Example 1. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

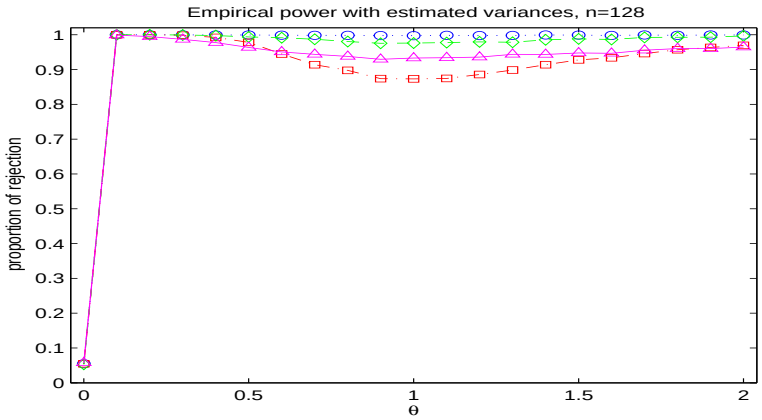

Figure 5.3: Powers for Example 1. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

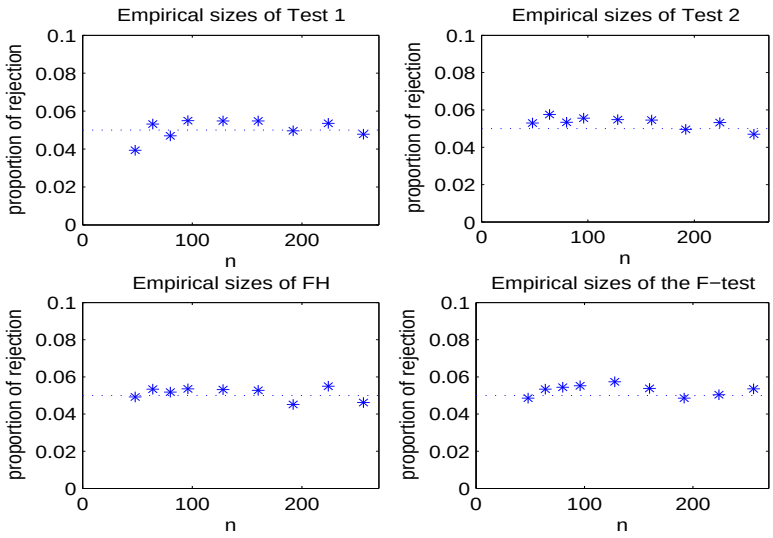

Figure 5.4: Empirical sizes for Example 1 under various sample sizes.

estimated variances $\hat{\sigma}_i^2$ are used. Test 1 far outperforms all other tests. It shows a far stabler and relatively high empirical power. FH works far poorer than Test 1 but still much better than the F statistic. Test 2 performs well when θ is close to 0 but drops dramatically when θ gets close to 1.

In Figure 5.2, using the true variance, the empirical powers of all tests are similar to those in Figure 5.1 except the F-test lost power when it is reduced to a χ^2-test.

Comparing Figures 5.1 and 5.3, when the sample size is doubled from $n = 64$ to $n = 128$, the sizes of our proposed tests remain 0.05 level and the empirical powers show similar patterns. Figure 5.4 shows the empirical sizes of all tests under various sample sizes. When the sample size is too small as $n = 48$, Test 1 is sightly undersized. All tests achieve the 0.05 significance level in all other samples.

Example 2. The predictor variable x_1 is sampled from a $N(0, 1)$. The response variable y is drawn from the model

$$y = \frac{1}{1 + \theta \cos(x_1)} + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.3}$$

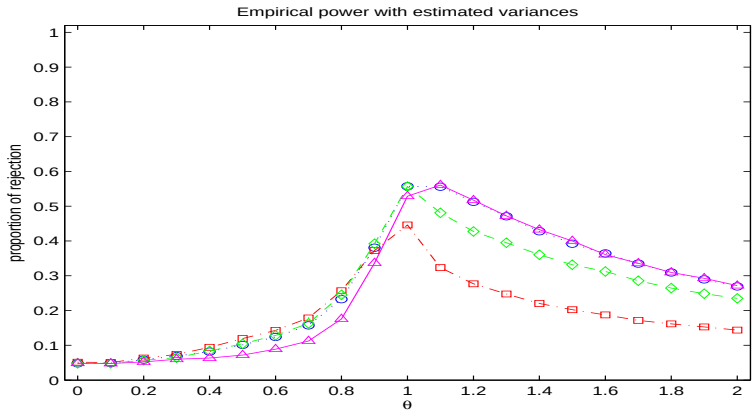

Figure 5.5: Powers for Example 2. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

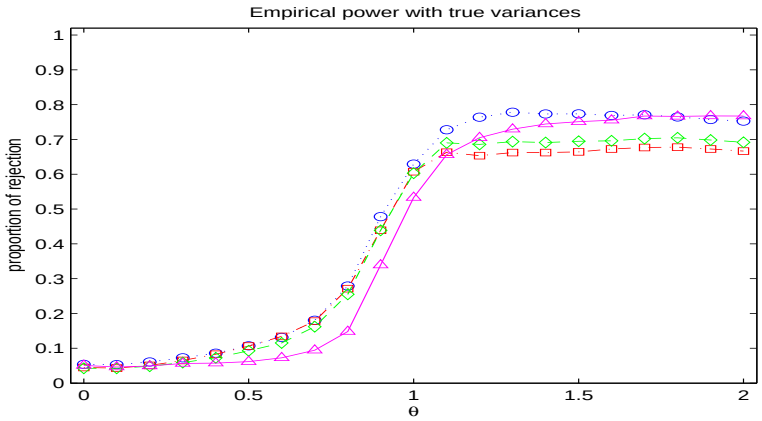

Figure 5.6: Powers for Example 2. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

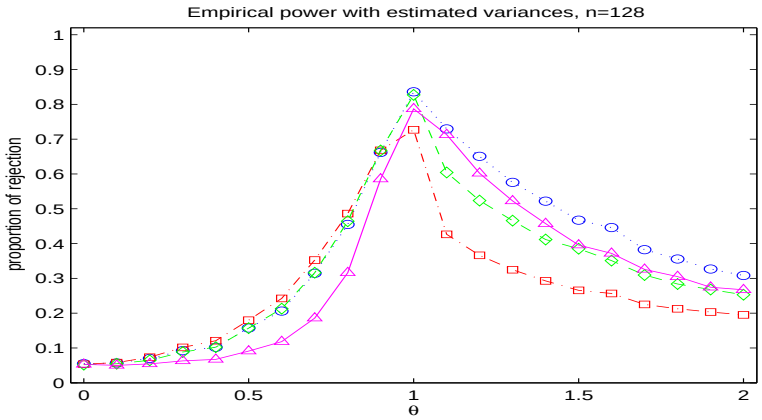

Figure 5.7: Powers for Example 2. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

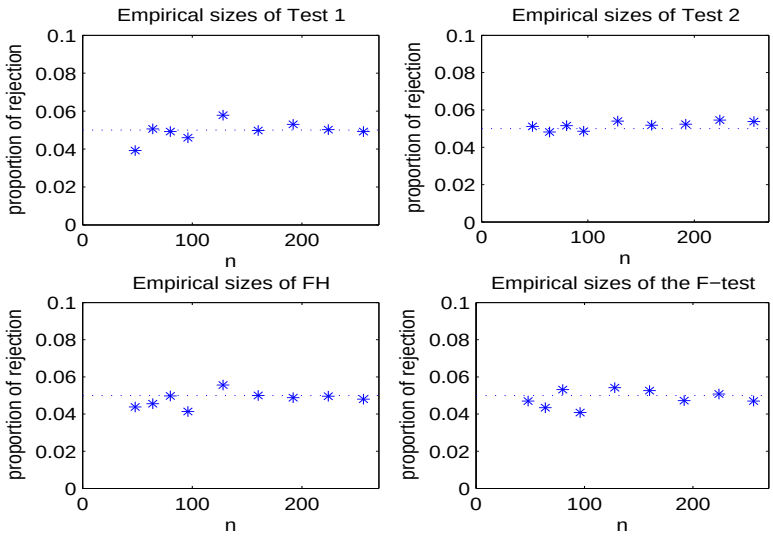

Figure 5.8: Empirical sizes for Example 2 under various sample sizes.

The results are presented in Figures 5.5 to 5.7. Note that although the powers in Figure 5.1 and Figure 5.5 show very different patterns, they have a similar nature, i.e. θ at around 1 to 1.1 gives a point of changing curvatures. When θ is close to 0, the tests with Neyman adaptive structure perform better than the F-test. FH and Test 2 lose power dramatically when θ is beyond 1 and thus, perform worse than F-test. Test 1 provides the best power with the θ in the range that we investigated. The improvement of Test 1 is more substantial when the size of the sample is doubled from 64 to 128. Figure 5.7 shows that the F-test lost power dramatically relative to other tests over the interval $\theta > 1$ when a larger sample is used.

Comparing Figures 5.5 and 5.6, the powers of all tests show very different patterns at $\theta > 1$. The simulations based on estimated variances show all tests keep losing power when θ gets beyond 1. When the true variance is used, all tests give slightly increasing powers when θ becomes larger. This suggests that the variance σ^2 cannot be well estimated by the estimators we used in this example.

To investigate the empirical sizes under various sample sizes, we use model (5.3)

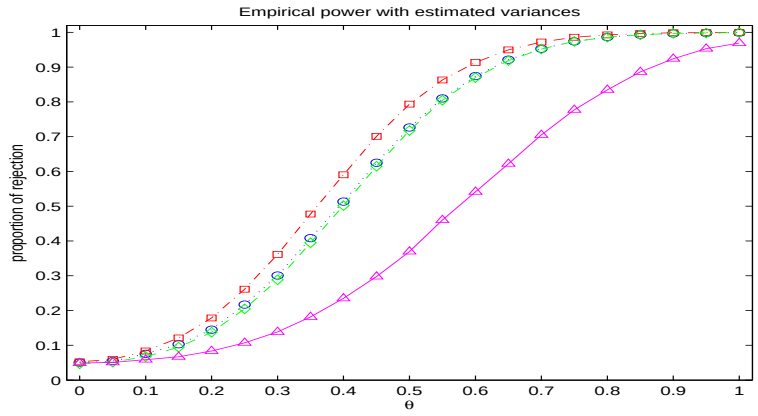

Figure 5.9: Powers for Example 3. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

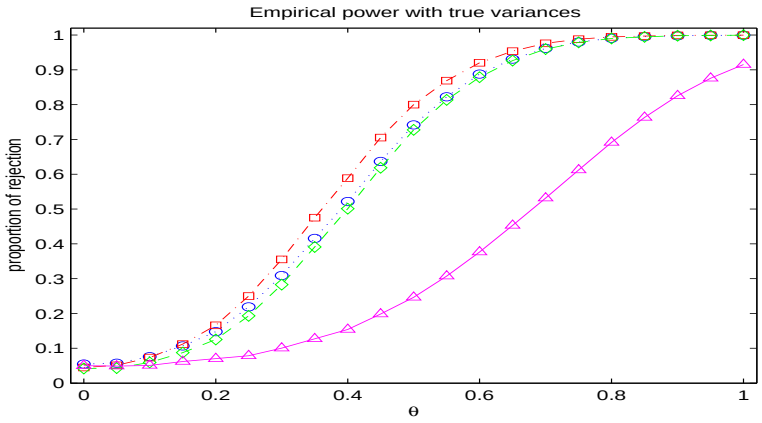

Figure 5.10: Powers for Example 3. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

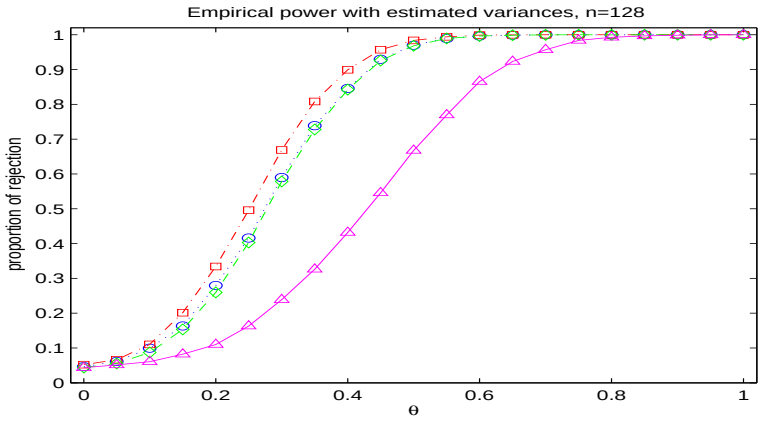

Figure 5.11: Powers for Example 3. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

with $\theta = 0$ to simulated the observations. Model (5.3) is reduced to

$$y = 1 + \epsilon. \tag{5.4}$$

From Figure 5.8, the pattern of the sizes are similar to Figure 5.4. Test 1 is slightly undersized when $n = 48$.

Model (5.5) in Example 3 and model (5.11) in Example 7 for testing lack-of-fit in simple linear regressions are the same as model (5.4) when $\theta = 0$. Therefore, illustrations of empirical sizes for Examples 3 and 7 are omitted.

Example 3. In this example, we simulate y from a quadratic model

$$y = 1 + \theta x_1^2 + \epsilon, \qquad \epsilon \sim N(0, 2^2), \tag{5.5}$$

for various values of θ and x_1 is sampled from a uniform $(-2, 2)$.

The results are given in Figures 5.9 to 5.11. Test 1 and FH perform close to each other in these figures. Test 2 makes a conspicuous advance over the other tests. In Figure 5.9, when θ is between 0.4 and 0.5, the empirical power of Test 2 is roughly 10% better than Test 1 and FH. All the tests outperform the exact F-test.

FH also include an exact F-test as a basis for comparison in their simulations based on model (5.5) but they test the simple linear regression against a quadratic regression. The full model in their exact F-test is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \epsilon_i. \tag{5.6}$$

Based on their simulated results, as should be expected, the F-test works slightly better than FH. Fan and Huang suggest that their test pays a large price to be more omnibus in nature, but our simulations suggest that the price may be very little. We found that our second test works as well as the F-test for a quadratic regression.

Example 4. The response variable y is drawn from a forth-order polynomial

$$y = 1 + 2x_1 + \theta x_1^4 + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.7}$$

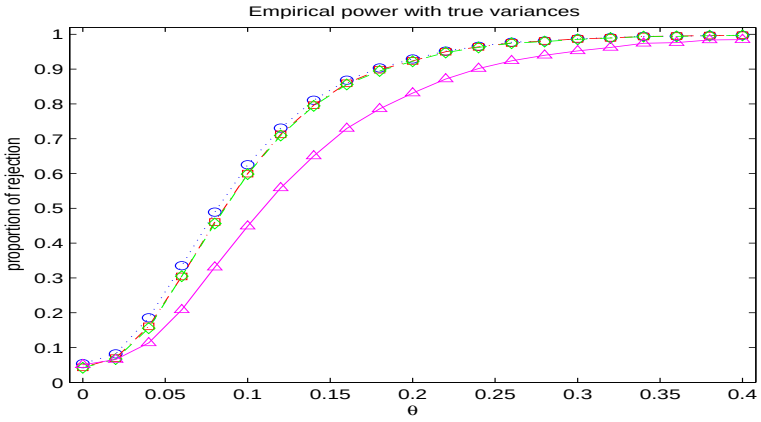

Figure 5.12: Powers for Example 4. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

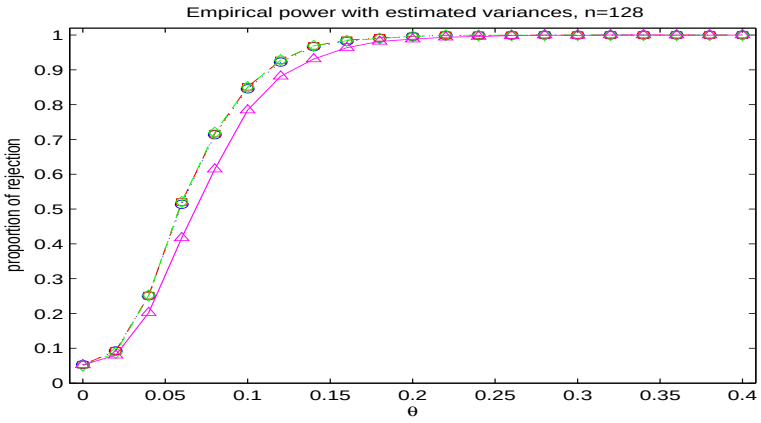

Figure 5.13: Powers for Example 4. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

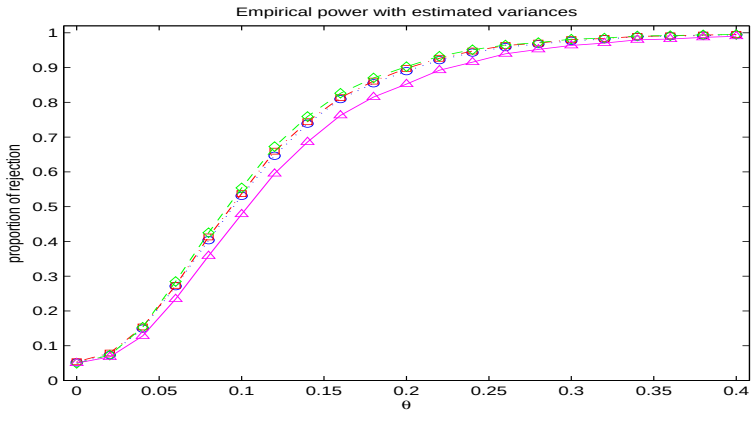

Figure 5.14: Powers for Example 4. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

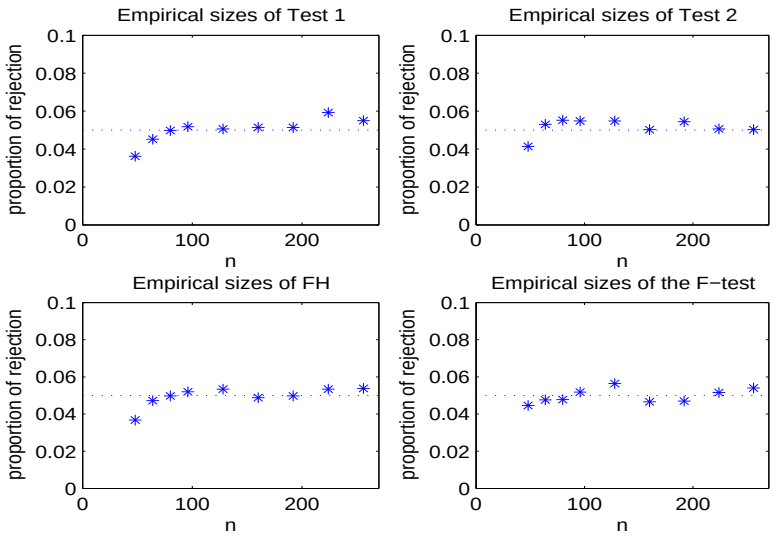

Figure 5.15: Empirical sizes for Example 4 under various sample sizes.

with x_1 sampled from a $N(0, 1)$.

The empirical powers of all tests are shown in Figures 5.14 to 5.13. The tests with Neyman adaptive structure perform as well as each other. Comparing Figures 5.9 and 5.14, when the highest-order term in the model is raised from 2 to 4, Test 1, Test 2 and FH are still more powerful than F-test, but the difference is not as substantial as before. Test 2 has an obvious loss of power relative to Test 1 and FH.

When $\theta = 0$, model (5.7) is reduced to

$$y = 1 + 2x_1 + \epsilon. \tag{5.8}$$

The same model is obtained from model (5.9) in Example 5 and model (5.10) in Example 6 when $\theta = 0$. The fitted model in Examples 4 to 6 are the same. Therefore, the graphs on the empirical sizes under various sample sizes in Examples 5 and 6 are omitted. From Figure 5.15, when $n = 48$, all adaptive tests are undersized. This indicates that the asymptotic distribution of the test statistics can not be achieved when the sample size is 48.

Example 5. The model used in this example is similar to those in Examples 3 and 4 such that the response variable y_i is simulated from a polynomial in x_{1i}.

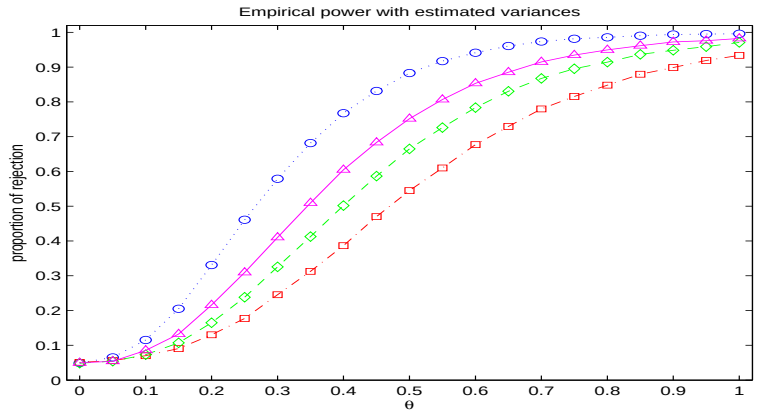

Figure 5.16: Powers for Example 5. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

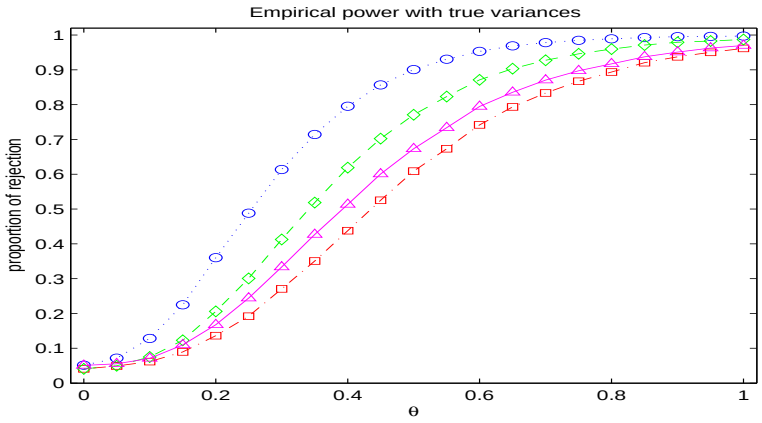

Figure 5.17: Powers for Example 5. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

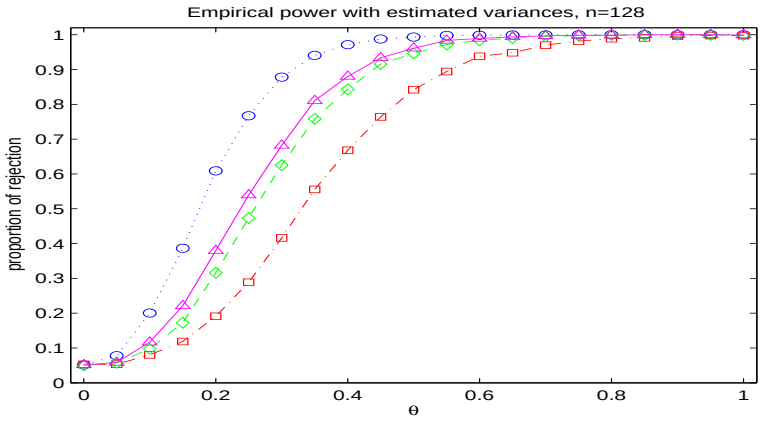

Figure 5.18: Powers for Example 5. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

Instead of the even-order polynomials in Examples 3 and 4, an odd-order polynomial is used. The covariate x_1 is sampled from a $N(0,1)$ and the response variable is drawn from a cubic polynomial

$$y = 1 + 2x_1 + \theta x_1^3 + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.9}$$

The patterns of the empirical powers are quite different from those in Examples 3 and 4. Figures 5.16 to 5.18 depict the results. In Figure 5.16, between $\theta = 0.3$ and $\theta = 0.5$, Test 1 is at least 25% more powerful than FH's test. FH is even less powerful than F-test. Test 2 performs worst among all four tests.

Similar to Example 1, when the estimated variances are replaced by the true variance, the χ^2-test has a dramatical loss in power relative to the F-test. As shown in Figure 5.17, FH is more powerful than the χ^2-test. Test 2 still performs worst when the true variance is used.

Example 6. In this example, the response variable y_i is simulated from a fifth-order polynomial in x_{1i}

$$y = 1 + 2x_1 + \theta x_1^5 + \epsilon, \qquad \epsilon \sim N(0, 2^2), \tag{5.10}$$

with the predictor x_1 sampled from a $N(0,1)$.

The results are shown in Figures 5.19 to 5.21. The patterns of the empirical powers are similar to those in Example 5. Test 1 shows a substantial advance over all other tests. Test 2 performs the worst among the tests. Comparing the figures with the figures in Example 5, the rise in the order of an odd-order polynomial has little impact on the performance of any tests except the power of the F-test is enhanced.

We want to investigate the difference on how the odd-order and even-order polynomials affect the performance of the tests. Figure 5.22 provides two scatterplots illustrating the data from Examples 3 and 6 with corresponding regression lines from the fitted model. The data from a quadratic polynomial do not provide an

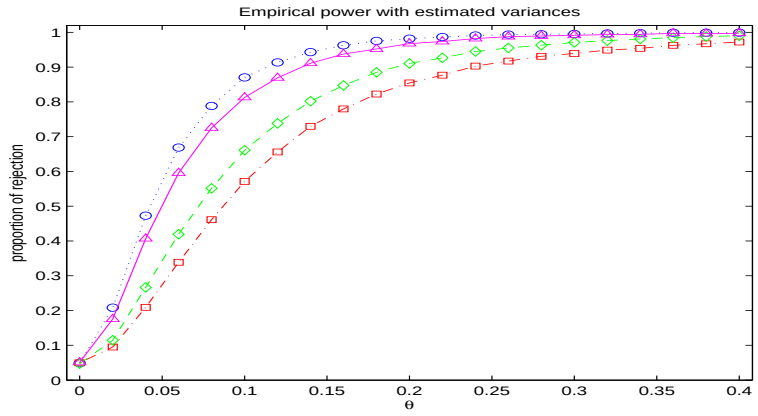

Figure 5.19: Powers for Example 6. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

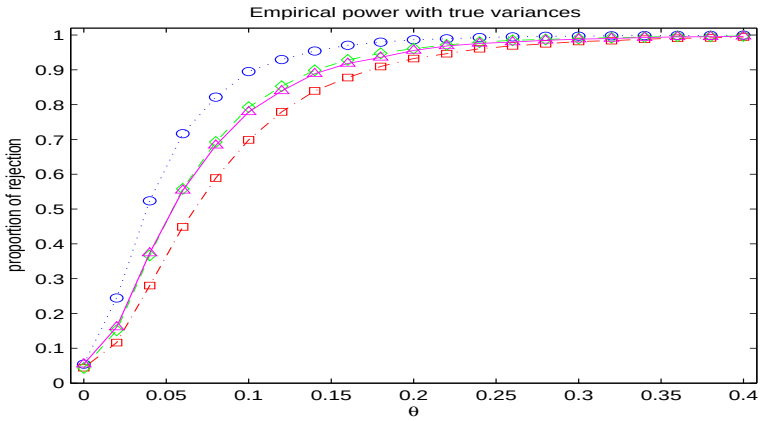

Figure 5.20: Powers for Example 6. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

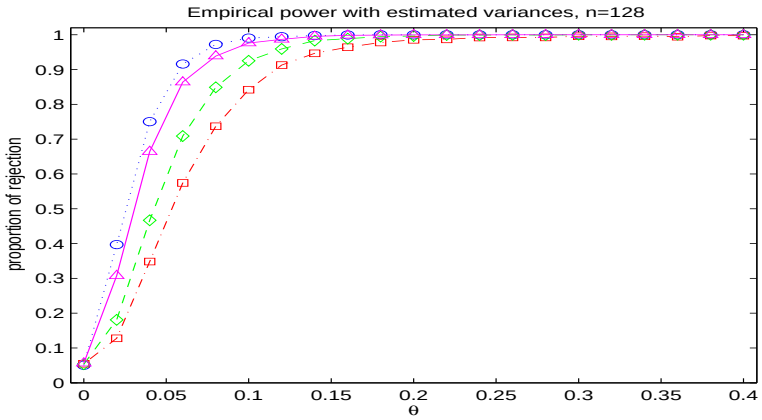

Figure 5.21: Powers for Example 6. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

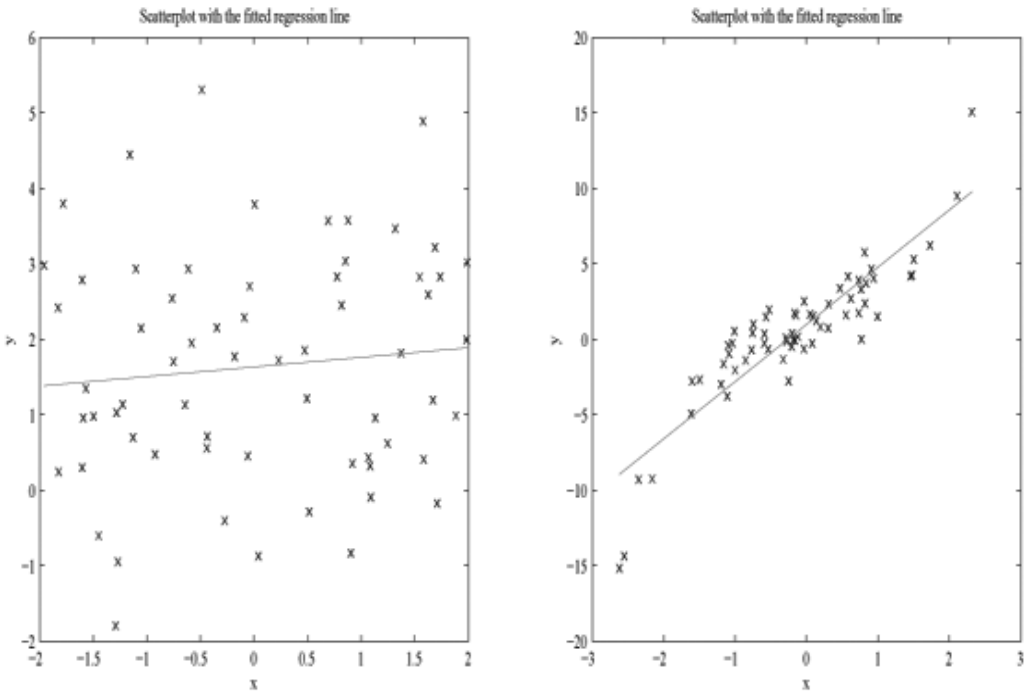

Figure 5.22: Left: Data from Example 2, model (5.5); Right: Data from Example 3, model (5.10).

obvious quadratic pattern. Recall that in this case, Test 2 is the most powerful among all tests; Test 1 and FH have similar powers. When the data show an obvious pattern, as those from the fifth-order polynomial, Test 1 outperforms all tests; Test 2 and FH have substantial loss of power. Figures 5.9 to 5.19 show similar results for both polynomials with an even order and also for both odd-order polynomials. The figures also indicate that the empirical power of the F-test approaches the power of Test 1 when the highest-order term in the polynomial is raised, no matter the highest-order term is an even-order term or an odd-order term.

Example 7. The covariate x_1 is sampled from a $N(0,1)$, and the response is drawn from

$$y = 1 + \theta \cos(x_1) + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.11}$$

Example 7 is similar to Example 3 in that both have $y_i = 1 + f(x_{1i})\theta + \epsilon_i$

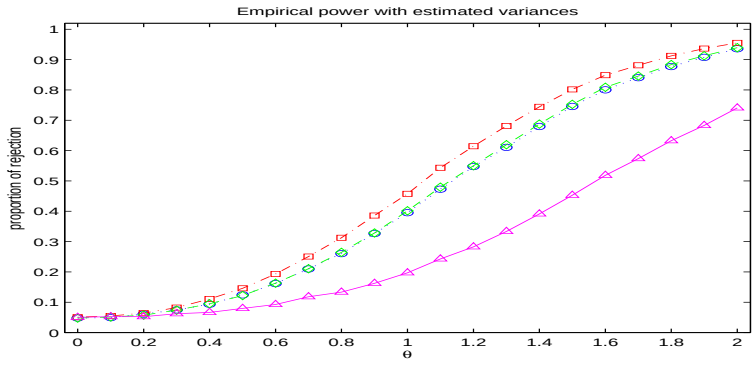

Figure 5.23: Powers for Example 7. $n = 64$ and estimated variances are used. Testing lack-of-fit for a simple linear model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

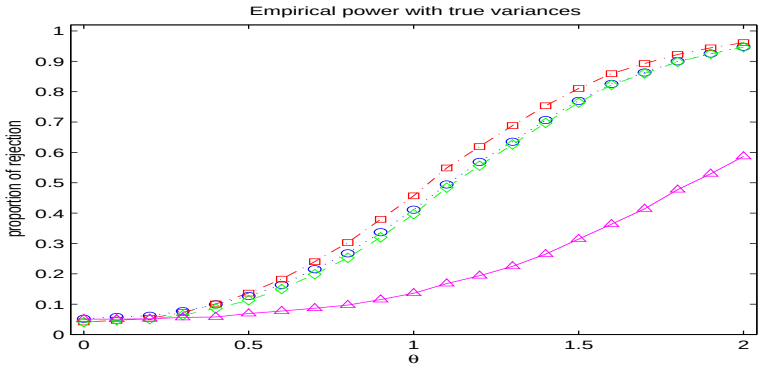

Figure 5.24: Powers for Example 7. $n = 64$ and the true variance is used. Testing lack-of-fit for a simple linear model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

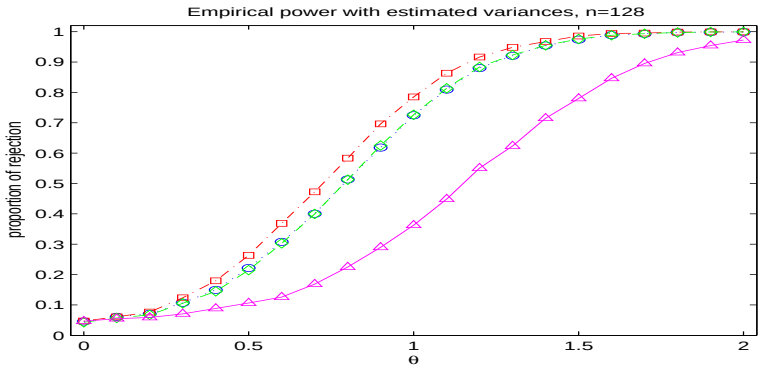

Figure 5.25: Powers for Example 7. $n = 128$ and estimated variances are used. Testing lack-of-fit for a simple linear model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

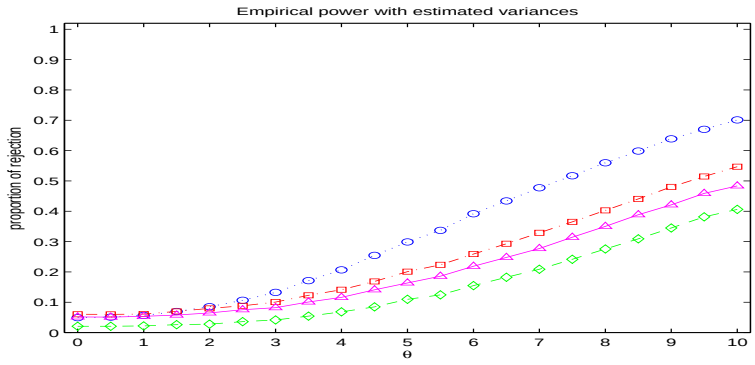

Figure 5.26: Powers for Example 7. $n = 64$ and estimated variances are used. Testing lack-of-fit for a cubic model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

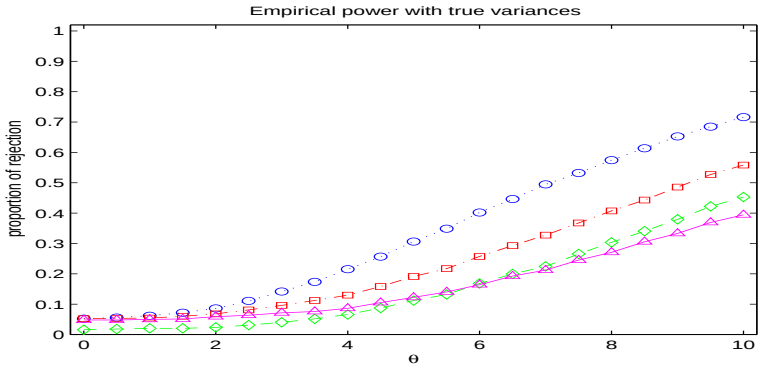

Figure 5.27: Powers for Example 7. $n = 64$ and the true variance is used. Testing lack-of-fit for a cubic model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

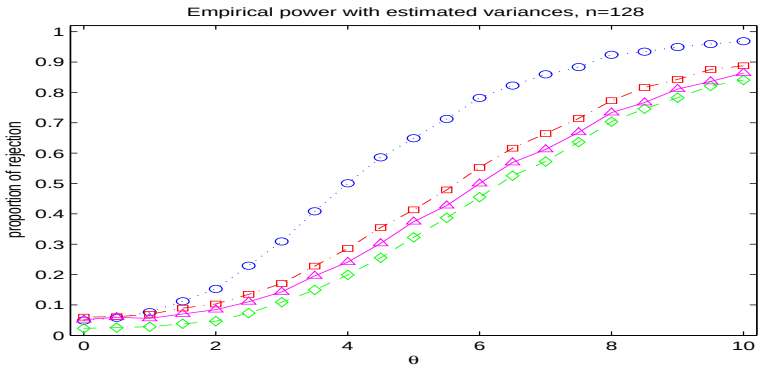

Figure 5.28: Powers for Example 7. $n = 128$ and estimated variances are used. Testing lack-of-fit for a cubic model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

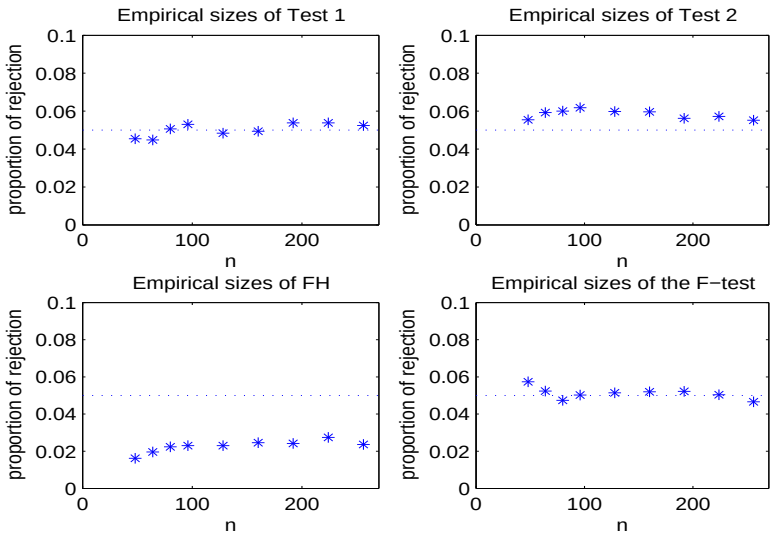

Figure 5.29: Empirical sizes for Example 7, testing lack-of-fit for a cubic model, under various sample sizes.

for some $f(\cdot)$. Example 7 replaces the quadratic structure of model (5.5) by a periodic structure. In this example, the lack-of-fit from fitting both a simple linear regression and a cubic model will be examined. In the case with a simple linear regression, the patterns of the empirical powers are similar to those in Figure 5.9. Figure 5.23 depicts the results. For simple linear regression, our second test is clearly superior.

When the lack-of-fit of a cubic model is tested, our Test 1 far outperforms all other tests. Figure 5.26 shows that at $\theta = 10$, Test 1 is 15%, 30%, and 20% more powerful than Test 2, FH, and the F-test respectively. Test 2 is more powerful than the F-test but FH works far poorer than the F-test. The increasing complexity of the tested model determines a matrix X such that the $C(X)$ is pretty adept at picking up low frequency terms in Γ_m. This makes the low frequency terms in Γ_m largely redundant. The FH statistic keeps these redundant terms in the test but gets little contribution from them making the test statistic relatively small. This forces down both the size and power of the test so that FH pays a large price for not adjusting Γ_m for the fitted X. Part of that price takes the form of a poor asymptotic approximation to the small sample null distribution. With

our small samples $n = 64$, the critical point from the asymptotic distribution of FH's test gives a size of 0.0211, which is far below the Test 1 size of 0.0493. Even when the sample size is doubled to $n = 128$, FH's test gives a size of 0.0230, which is still far below the Test 1 size of 0.0484. Figure 5.29 shows that Test 2 is slightly oversized and FH is far undersized. The oversize in Test 2 can be improved when the sample gets large. But FH does not seem to be improved by raising the sample size. Our proposed tests largely eliminate the overlap from our test statistics, thus our proposed tests maintain their size and relatively high powers in testing lack-of-fit for a polynomial model.

An even more extreme case arises if X contains low frequency sine and cosine terms. This will be illustrated in the next example.

Example 8. The covariate x_1 is sampled from a $N(0, 1)$, and the response variable is drawn from

$$y = 1 + \exp(\theta x_1) + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.12}$$

Two models are tested for lack-of-fit: a simple linear regression on x_1 and model (5.1). The empirical powers of all tests are shown in Figures 5.30 to 5.35. When a simple linear model is fitted, the tests with adaptive Neyman structure perform equally well and they are more powerful than the F-test.

When the low frequency sine and cosine terms are contained in the design matrix X of the tested model, FH has a substantial loss of power and the size of the test tends to be 0. Test 1, Test 2, and the F-test outperform FH. This is because the low frequency terms in Γ_m are completely redundant and the FH statistic adds zero for each redundant term, forcing the test statistic to be relatively small. Our tests correct for the low frequency terms being in the fitted model. Thus a good small sample approximation of the asymptotic null distribution is retained in our tests. In Figures 5.33, the empirical size of Test 1 and Test 2 are 0.0493 and 0.0500 respectively. As shown in Figures 5.35, when the sample size is doubled to $n = 128$, the improvement of our proposed tests becomes substantial. Our

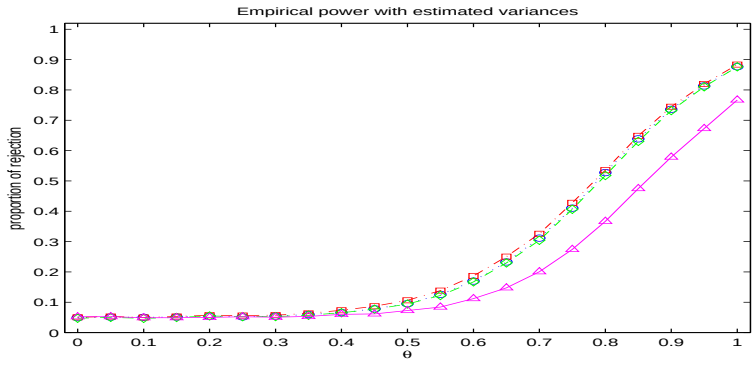

Figure 5.30: Powers for Example 8. $n = 64$ and estimated variances are used. Testing lack-of-fit for a simple linear model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

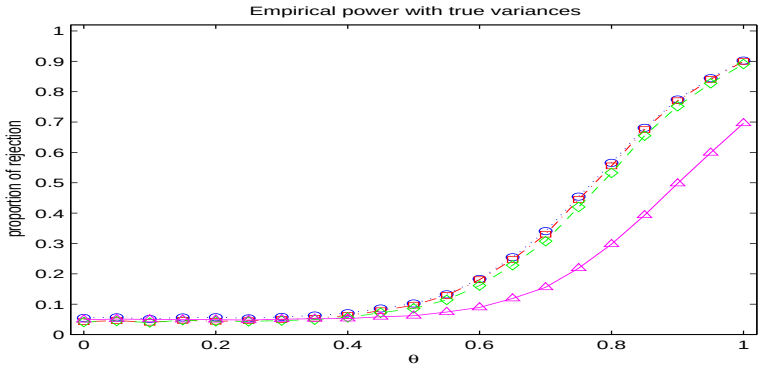

Figure 5.31: Powers for Example 8. $n = 64$ and the true variance is used. Testing lack-of-fit for a simple linear model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

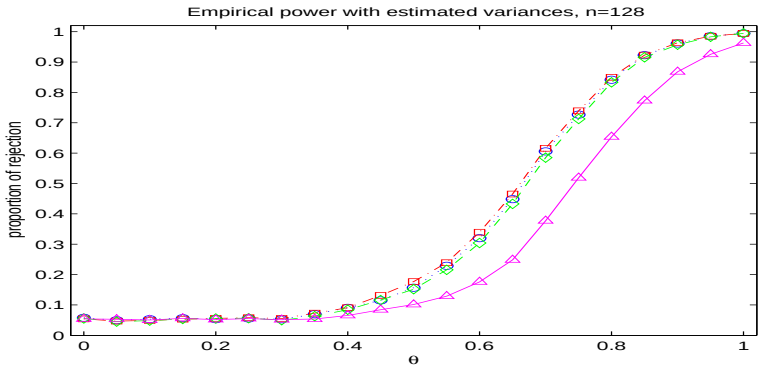

Figure 5.32: Powers for Example 8. $n = 128$ and estimated variances are used. Testing lack-of-fit for a simple linear model. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

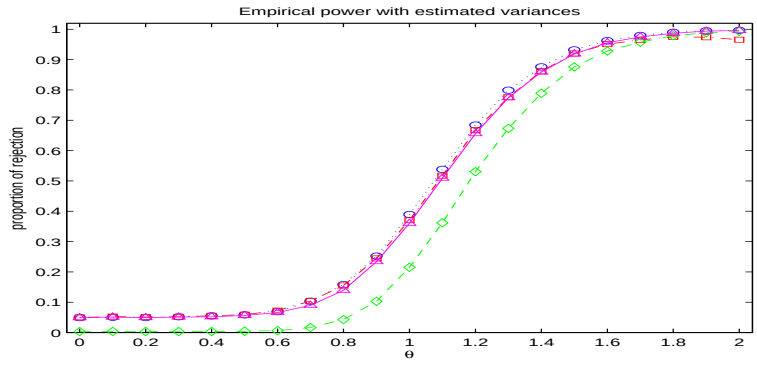

Figure 5.33: Powers for Example 8. $n = 64$ and estimated variances are used. Testing lack-of-fit for model (5.1). Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

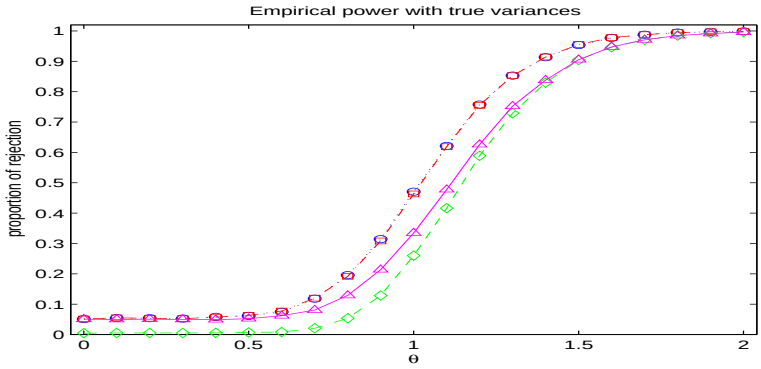

Figure 5.34: Powers for Example 8. $n = 64$ and the true variance is used. Testing lack-of-fit for model (5.1). Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

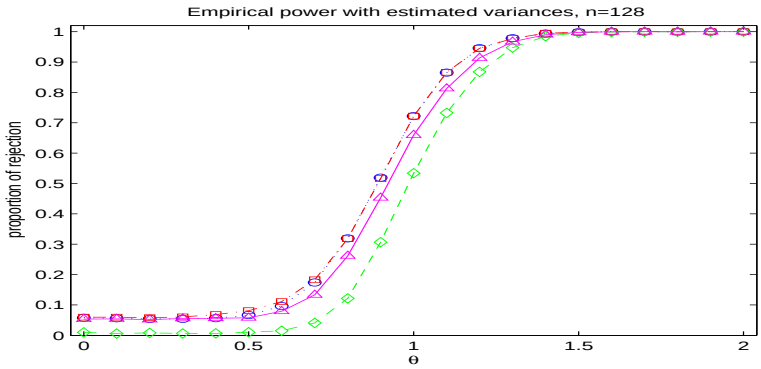

Figure 5.35: Powers for Example 8. $n = 128$ and estimated variances are used. Testing lack-of-fit for model (5.1). Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

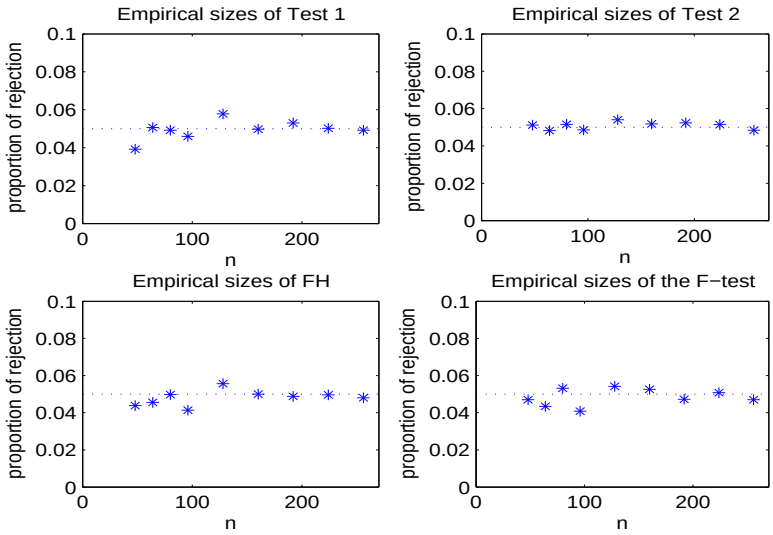

Figure 5.36: Empirical sizes for Example 8, testing lack-of-fit for a simple linear model, under various sample sizes.

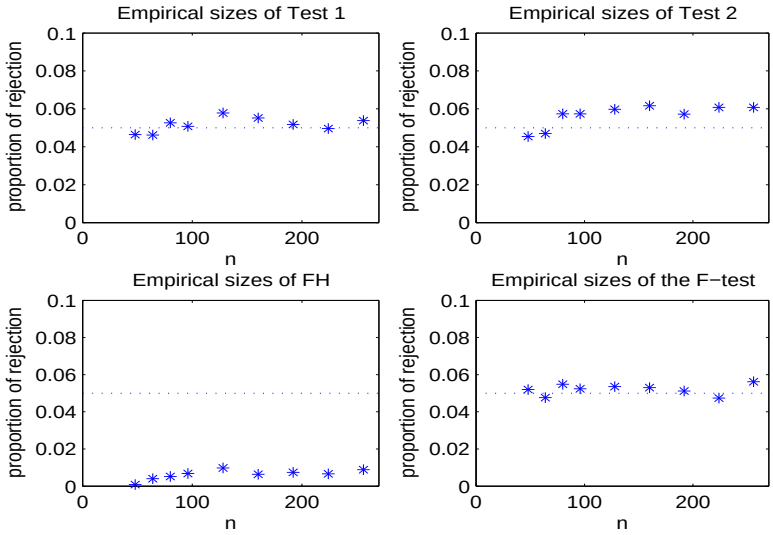

Figure 5.37: Empirical sizes for Example 8, testing lack-of-fit for model (5.1), under various sample sizes.

proposed tests work best among all four tests.

Figures 5.36 and 5.37 shows that the empirical sizes of FH drop dramatically when some of the low frequency terms in Γ_m are completely redundant. Comparing Figures 5.36 and 5.29, obviously the performance of FH for testing lack-of-fit in model (5.1) is much poorer than testing lack-of-fit in a cubic model.

5.1.3 Summary

Summarizing the results above, the F-test is very sensitive to the highest order term in the true model. As we mentioned earlier, the F-test approaches Test 1 when the order of the polynomial in the true model is raised. Our first proposed test outperforms all other tests when we are testing lack-of-fit in a model with one predictor in most situations. When the fitted model has a complicated structure such as the cubic polynomial in Example 7 and model (5.1) in Example 8, FH failed to achieve the asymptotic distribution in small samples. Test 2 outperforms Test 1 only in some situations that the data do not provides clear patterns of the underlying true model.

5.2 Multiple regression

5.2.1 Brief outline

We examine testing lack-of-fit in multiple regressions using seven examples.

In Examples 9 to 14, the fitted model has four predictors, x_1, x_2, x_3 and x_4. The covariates x_1, x_2, x_3 are standard normal with correlation 0.5 and x_4 is a Bernoulli with probability 0.4. x_4 is independent of x_1, x_2, and x_3. Following FH, the model being fitted in these examples is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{i4} + \epsilon_i. \tag{5.13}$$

We assume that the lack-of-fit is known to come from x_2, so we order the observations according to x_2, and the true model has the form

$$y_i = \beta_0 + \beta_1 x_{1i} + h(x_{2i}) + \beta_3 x_{3i} + \beta_4 x_{i4} + \epsilon_i. \tag{5.14}$$

In Example 9, function $h(\cdot)$ in model (5.14) is chosen to be a periodic function with the parameter θ dominating the period. The performance of the tests at different rates of oscillation are compared and discussed.

The true models involved in Examples 10 to 14 are multiple regression versions of the models in Section 5.1. In Example 10, we examine the tests by using a true model that is not linear in the parameter θ. Examples 11 to 14 use models that are linear in θ. The tests give similar power patterns in these examples except in Examples 13 and 14, in which an odd-order polynomial in x_2 is used. Moreover the performance of Test 1, FH, and the F-test in Examples 10 to 14 are consistent to their performance in the corresponding simple linear regressions in Section 5.1. The increasing number of predictors only makes a huge impact on our second proposed test.

In Example 15, the four covariates x_1 to x_4 defined above are involved in the model and we further add another four covariates, x_5, x_6, x_7, and x_8 into the model. These four covariates are i.i.d. standard normal random variables and they are independent of other covariates. The fitted model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \epsilon_i.$$

This example further illustrates the fact that the effect of the number of predictors on the performance of the tests is negligible.

5.2.2 Examples

Example 9. In this example, the response variable y is drawn from

$$y = x_1 + \cos(\theta x_2 \pi) + 2x_4 + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.15}$$

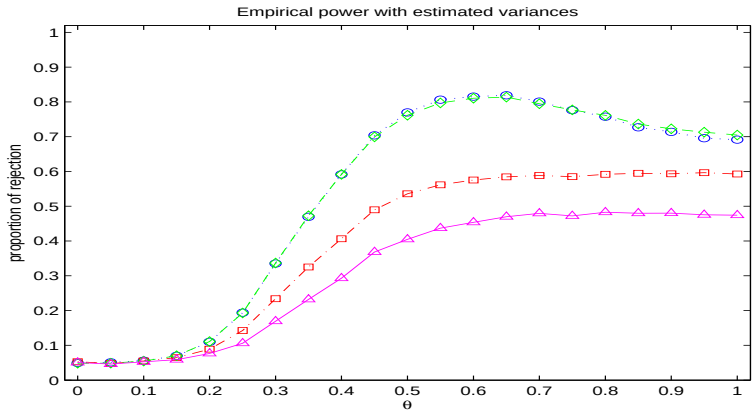

Figure 5.38: Powers for Example 9, for $0 \leq \theta \leq 1$. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

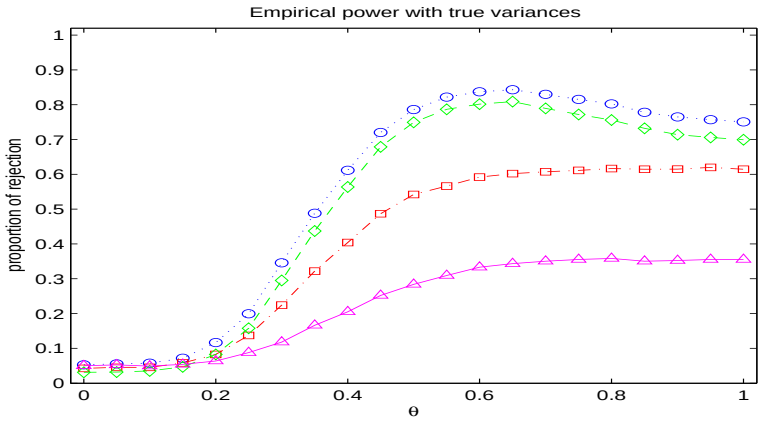

Figure 5.39: Powers for Example 9, for $0 \leq \theta \leq 1$. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

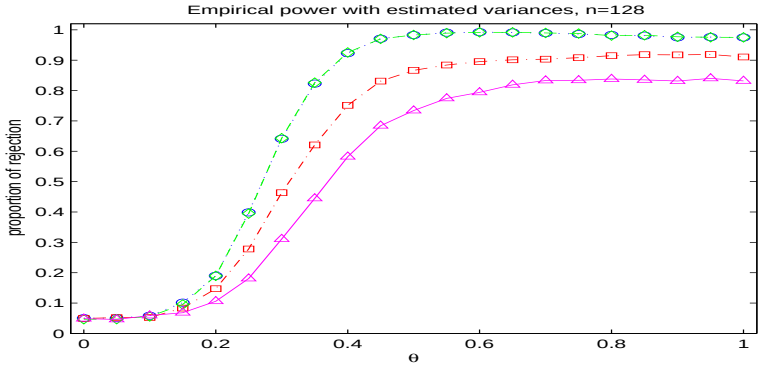

Figure 5.40: Powers for Example 9, for $0 \leq \theta \leq 1$. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

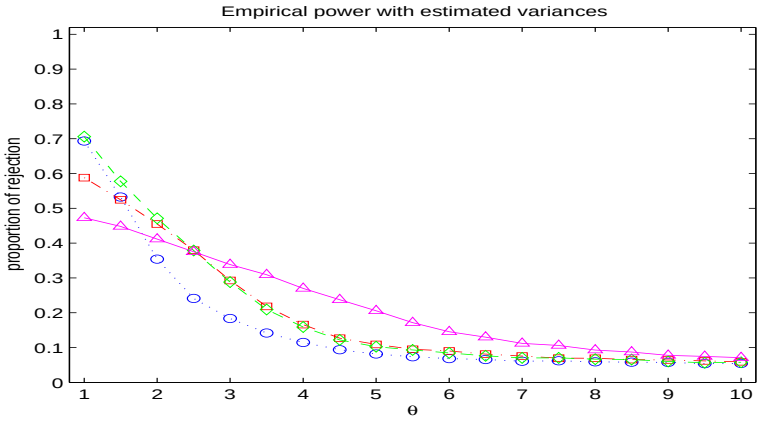

Figure 5.41: Powers for Example 9, for $\theta \geq 1$. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

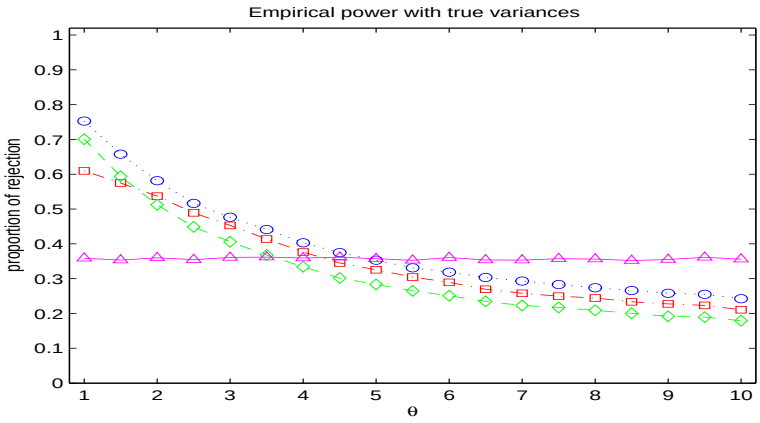

Figure 5.42: Powers for Example 9, for $\theta \geq 1$. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

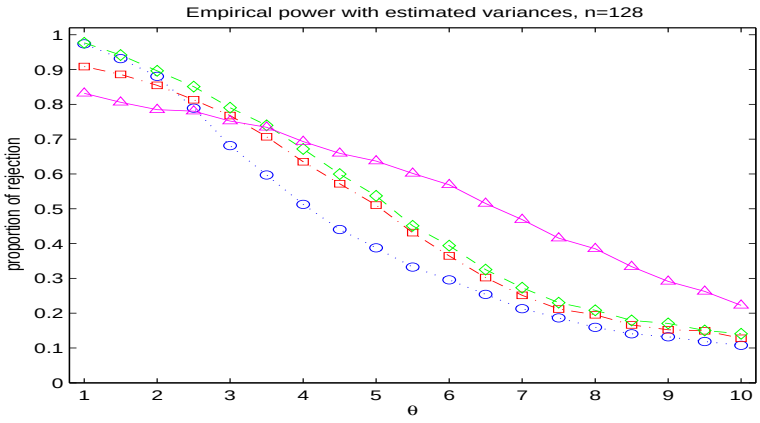

Figure 5.43: Powers for Example 9, for $\theta \geq 1$. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

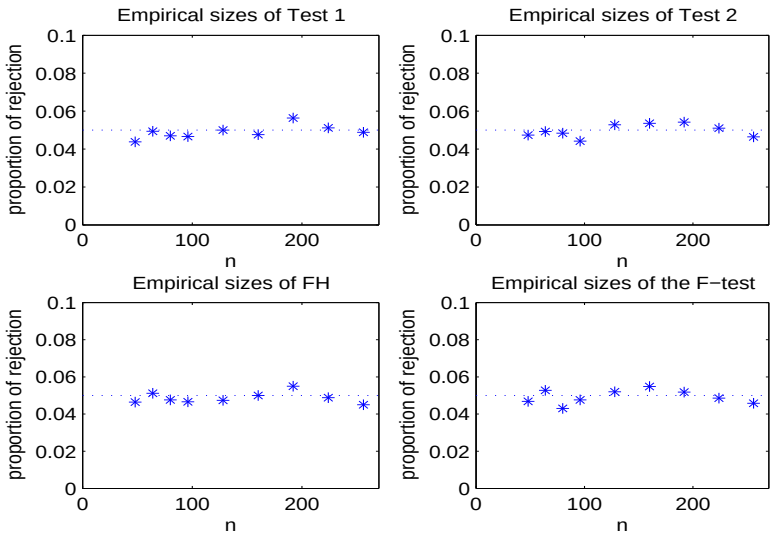

Figure 5.44: Empirical sizes for Example 9 under various sample sizes.

Figures 5.38 to 5.40 and Figures 5.41 to 5.43 depict the results for $0 \leq \theta \leq 1$ and $1 \leq \theta \leq 10$ respectively. For θ between 0 and 1, when the estimated variances are used, we cannot see any substantial difference among the empirical powers of Test 1 and FH and both of them outperform the other two tests. When the true σ^2 is used, Test 1 slightly outperforms FH. In both cases, Test 2 is more powerful than the F-test.

These results occur only if the true model is relatively smooth, i.e. $\theta \leq 2$. The empirical power of the tests drops dramatically when θ gets large so that oscillations are rapid. From Figures 5.41 to 5.43, we found that when θ is larger than 3, the F-test outperforms all the other tests. Figure 5.41 shows that Test 1 has a substantial loss of power relative to Test 2 and FH when θ is greater than 1. The differences among the tests are more substantial in a larger sample as shown in Figure 5.43. FH argue that the loss of power in adaptive tests is because when θ is large, it is difficult to estimate σ^2 well. It is worth noting that when σ^2 is known, FH works worst and Test 1 works best among the tests with adaptive Neyman structure for $\theta \geq 2$. The ranking on the performance of these tests are reversed when the estimated variances are used. Similar results were obtained from a similar simulation using a only x_2 in the fitted and true models.

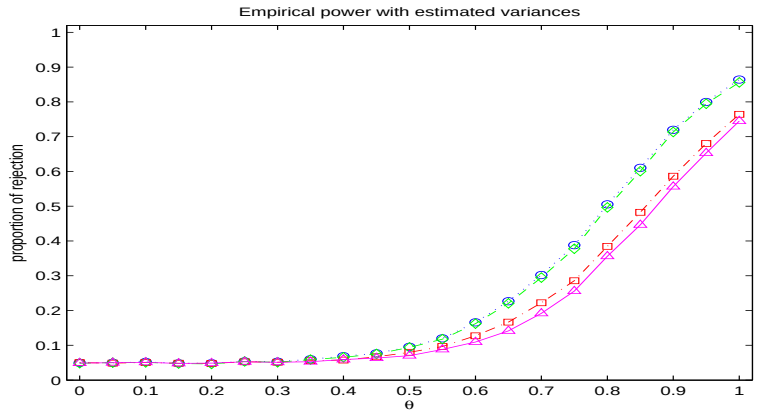

Figure 5.45: Powers for Example 10. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

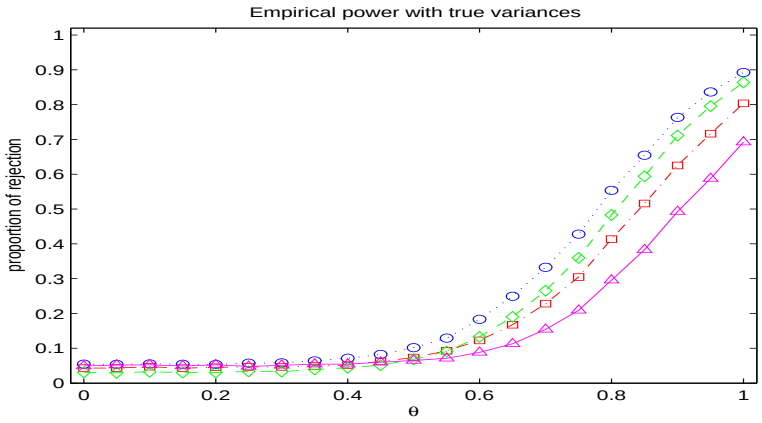

Figure 5.46: Powers for Example 10. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

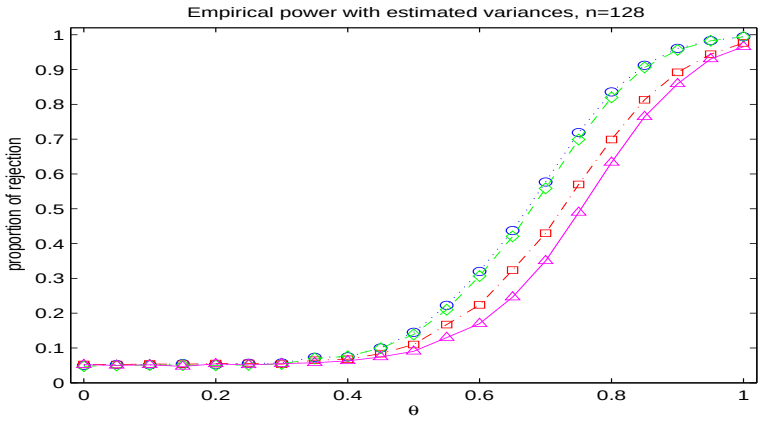

Figure 5.47: Powers for Example 10. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

To investigate the empirical sizes of the tests, we use $\theta = 0$ and model (5.15) is reduced to

$$y = 1 + x_1 + 2x_4 + \epsilon. \tag{5.16}$$

The model is equivalent to model (5.17) in Example 10 with $\theta = 0$. Therefore, investigation on the empirical sizes of the tests under various sample sizes is omitted in Example 10. In Figure 5.44, all tests provide stable empirical sizes around 0.05 in various sample sizes. The asymptotic distribution of the test statistics is achieved in small samples.

Example 10. The response variable y is simulated from model

$$y = x_1 + \exp(\theta x_2) + 2x_4 + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.17}$$

This is a multiple regression version of model (5.12) in Example 8.

Comparing the results of Figure 5.45 with Figure 5.30, the patterns of the empirical test powers look similar except for the power of our second proposed test. It is obvious that the power of Test 2 is reduced dramatically when there are more covariates involved in the model. In Example 8 with a simple linear regression as fitted model, Test 1, Test 2, and FH work equally well. With more covariates, Test 2 becomes the worst among the adaptive tests but still outperforms the F-test. The loss of power of Test 2 is more conspicuous in Examples 11 and 12.

Example 11. In this example, the response variable y is drawn from

$$y = x_1 + \theta \cos(x_2) + 2x_4 + \epsilon, \qquad \epsilon \sim N(0, 2^2), \tag{5.18}$$

which is a multiple regression version of model (5.11) in Example 7.

As shown in Figures 5.23 to 5.25, in Example 7 with a simple linear regression as fitted model, Test 2 was the most powerful of the tests. When more covariates are involved in th model, Test 2 becomes the worst among the adaptive tests. Comparing Figure 5.48 with Figure 5.23, for model with one predictor, Test 2 is roughly 5% more powerful than Test 1 and FH at θ around 1.2. When three more

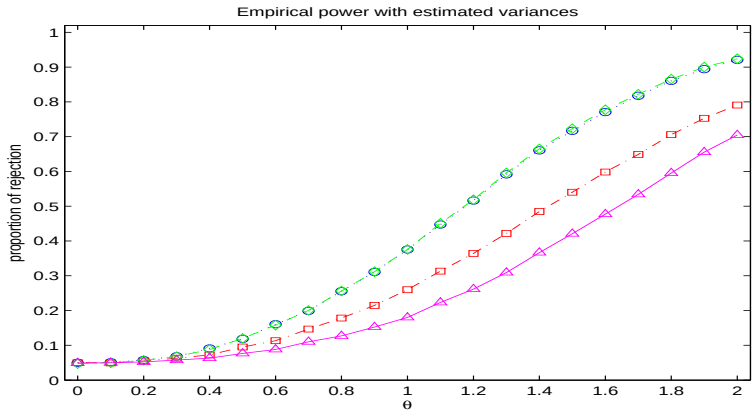

Figure 5.48: Powers for Example 11. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

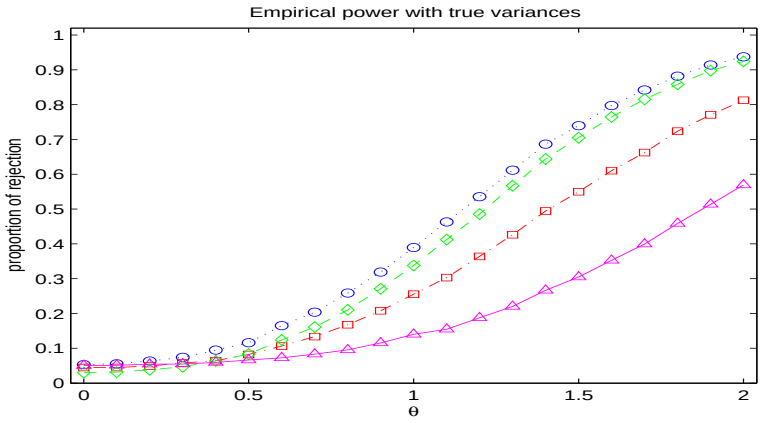

Figure 5.49: Powers for Example 11. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

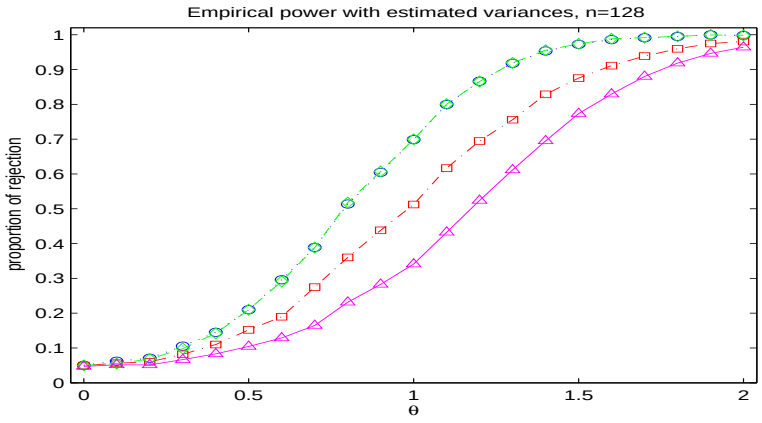

Figure 5.50: Powers for Example 11. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

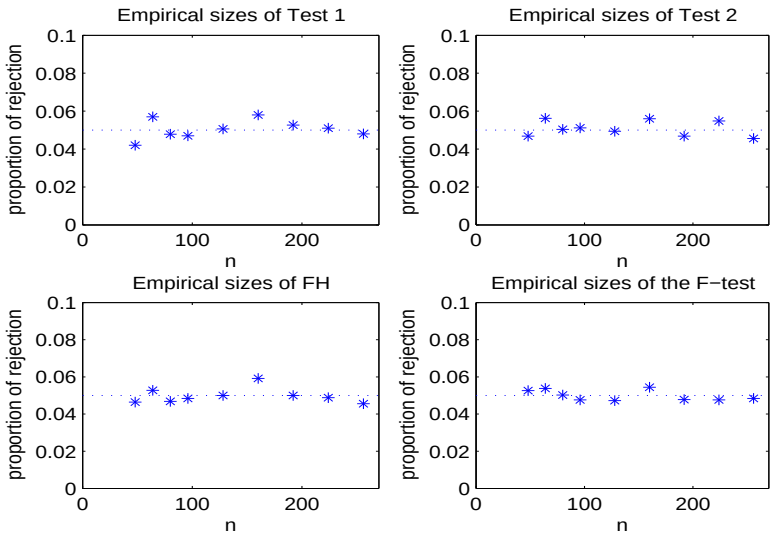

Figure 5.51: Empirical sizes for Example 11 under various sample sizes.

covariates are involved in the model, Test 2 becomes 20% less powerful than the other two tests for the same θ.

When $\theta = 0$, model (5.18) is reduced to

$$y = x_1 + 2x_4 + \epsilon. \tag{5.19}$$

The same model is obtained from model (5.20) in Example 12 and model (5.21) in Example 13 when $\theta = 0$. The plots of the empirical sizes of the tests under various sample sizes in Examples 12 and 13 are omitted. In Figure 5.51, all tests provide empirical sizes around 0.05, which is similar to those in Figure 5.44.

Example 12. The response variable y is drawn from model

$$y = x_1 + \theta x_2^2 + 2x_4 + \epsilon, \qquad \epsilon \sim N(0, 2^2), \tag{5.20}$$

which is a multiple regression version of model (5.5) in Example 3.

Comparing Figure 5.52 with Figure 5.9, similar characteristics as in Example 11 can be found. Together with the comparison between Figure 5.45 and Figure 5.30, we conclude that increasing the number of predictor variables in the model has a deleterious effect on the power of Test 2. Nevertheless, Test 2 still outperforms the F-test. The performance of Test 1 and FH do not depend on the number

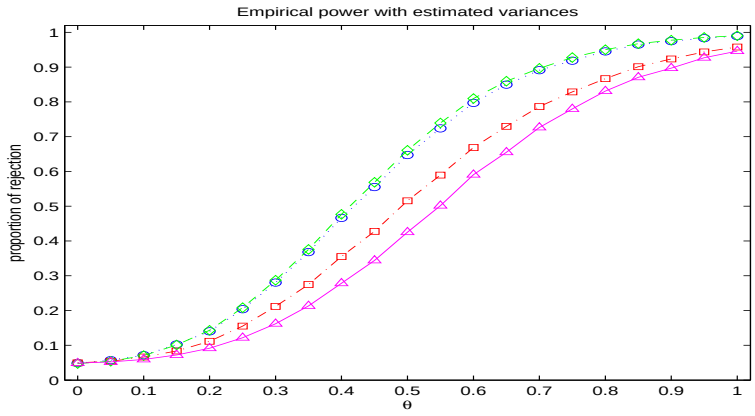

Figure 5.52: Powers for Example 12. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

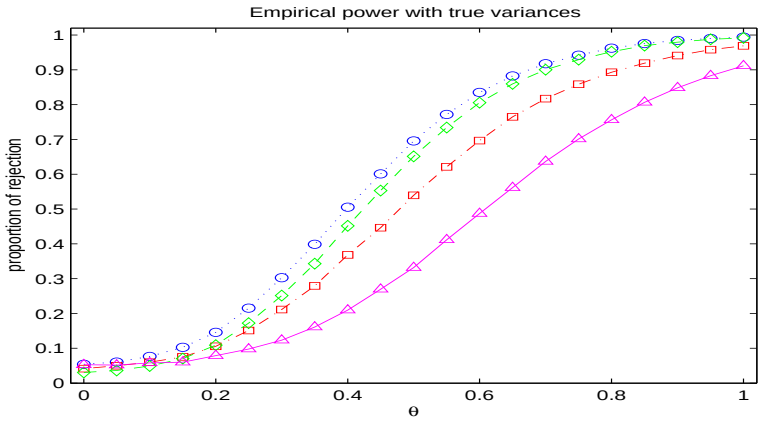

Figure 5.53: Powers for Example 12. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

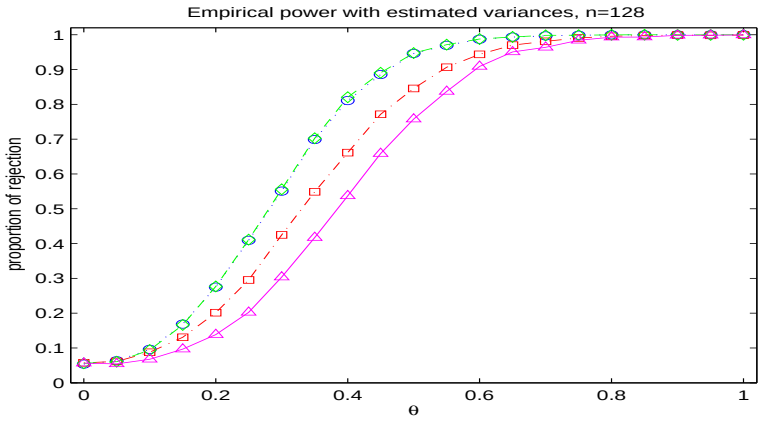

Figure 5.54: Powers for Example 12. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

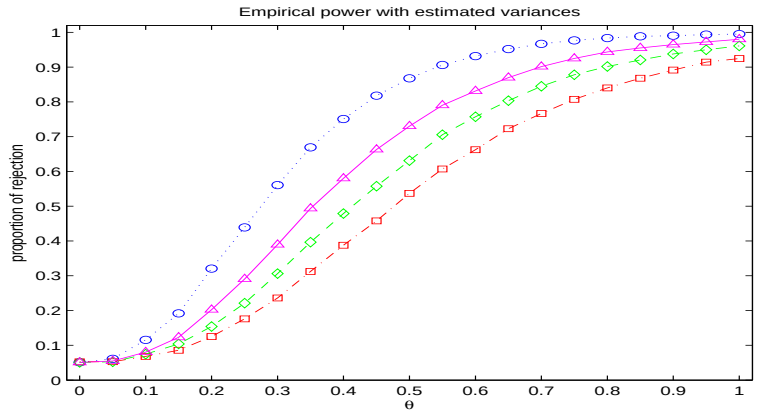

Figure 5.55: Powers for Example 13. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

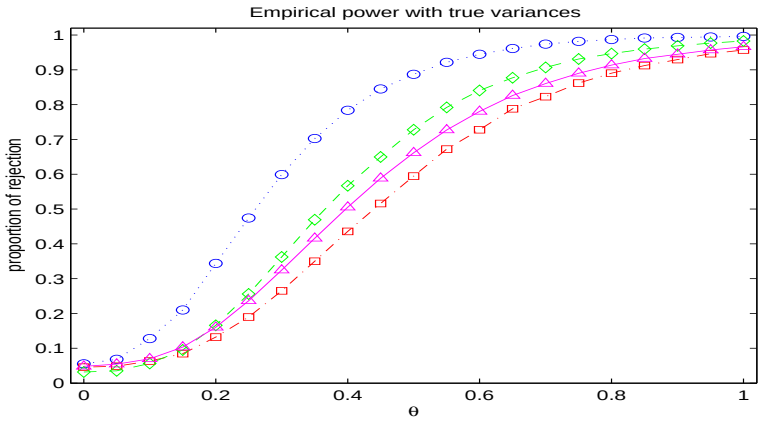

Figure 5.56: Powers for Example 13. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

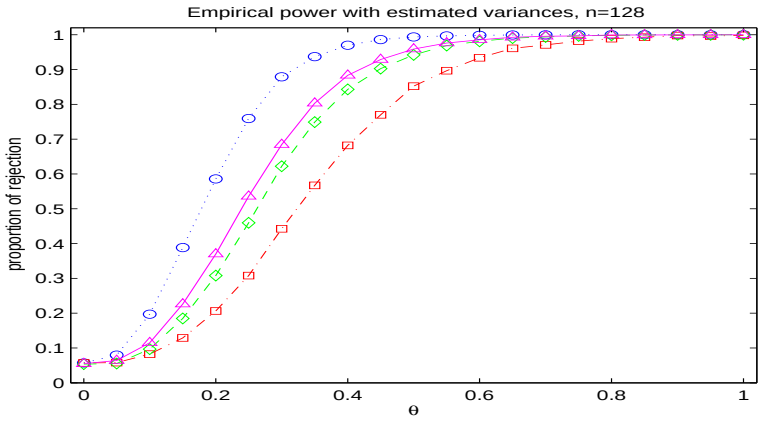

Figure 5.57: Powers for Example 13. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

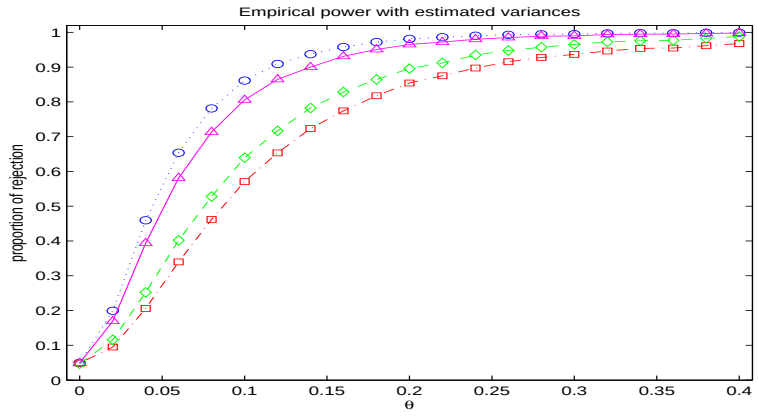

Figure 5.58: Powers for Example 14. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

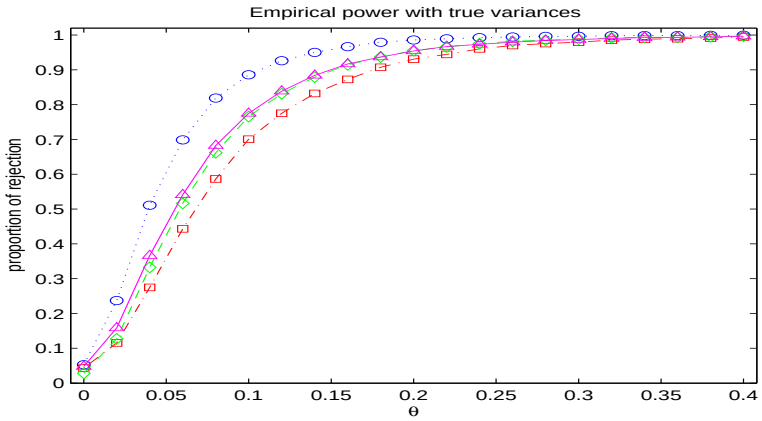

Figure 5.59: Powers for Example 14. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

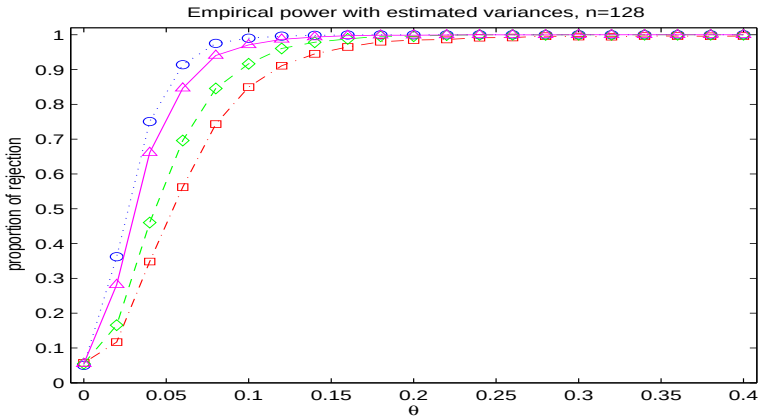

Figure 5.60: Powers for Example 14. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

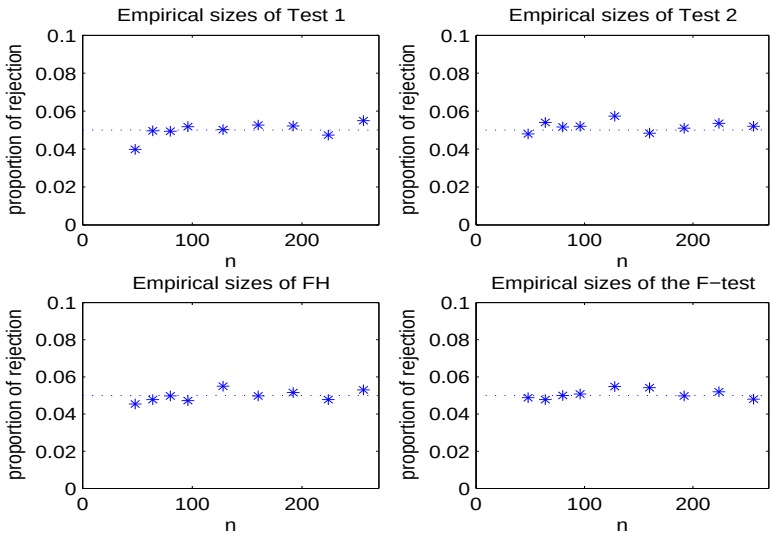

Figure 5.61: Empirical sizes for Example 14 under various sample sizes.

of predictors in the fitted model at all. The powers of Test 1 and FH's show no substantial difference.

Example 13. The responses are simulated from model

$$y = x_1 + \theta x_2^3 + 2x_4 + \epsilon, \qquad \epsilon \sim N(0, 2^2). \tag{5.21}$$

This model is an extension of model (5.9) in Example 5.

Figures 5.55 to 5.57 depict the results. Comparing Figure 5.55 and 5.16, there is little difference between the relative powers of the tests when fitting a simple linear regression and the corresponding multiple regression.

Example 14. In this example, the response variable is drawn from model

$$y = 1 + 2x_2 + \theta x_2^5 + 3x_4 + \epsilon, \qquad \epsilon \sim N(0, 2^2), \tag{5.22}$$

which is a multiple regression version of model (5.10) in Example 6.

The results are shown in Figures 5.58 to 5.60. Similar to Example 13, the comparison between Figure 5.58 and 5.19 indicates little difference between the relative powers of the tests when fitting a simple linear regression and the corresponding multiple regression. The relative powers of all tests seem unaffected

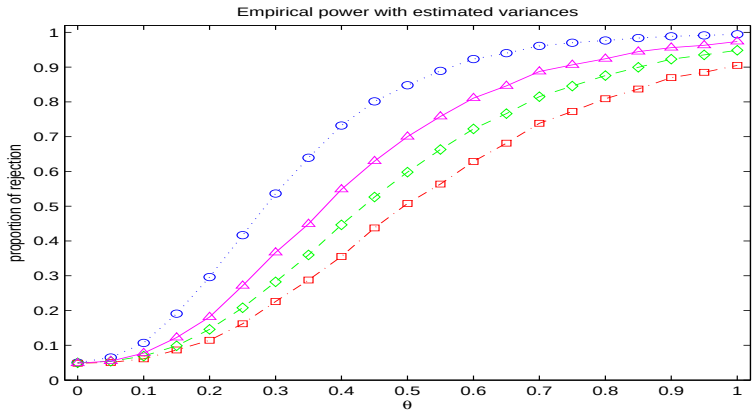

Figure 5.62: Powers for Example 15. $n = 64$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

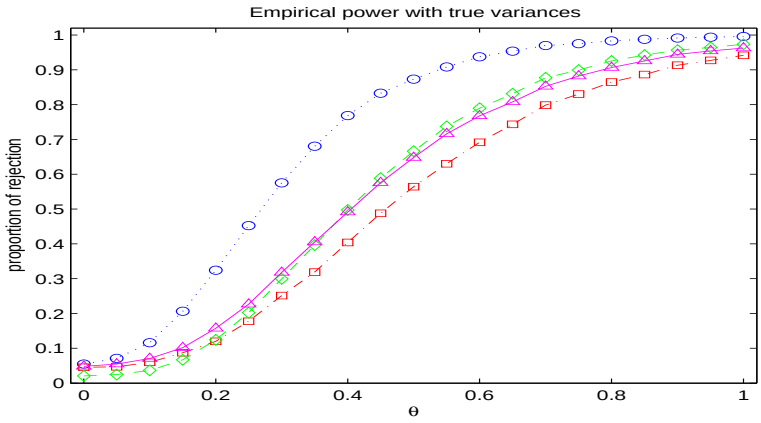

Figure 5.63: Powers for Example 15. $n = 64$ and the true variance is used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

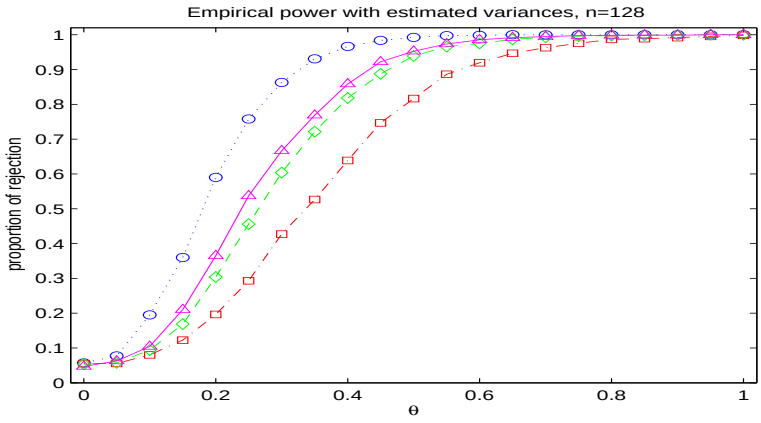

Figure 5.64: Powers for Example 15. $n = 128$ and estimated variances are used. Key: Circle, Test 1; Square, Test 2; Diamond, FH; Triangle, F-test.

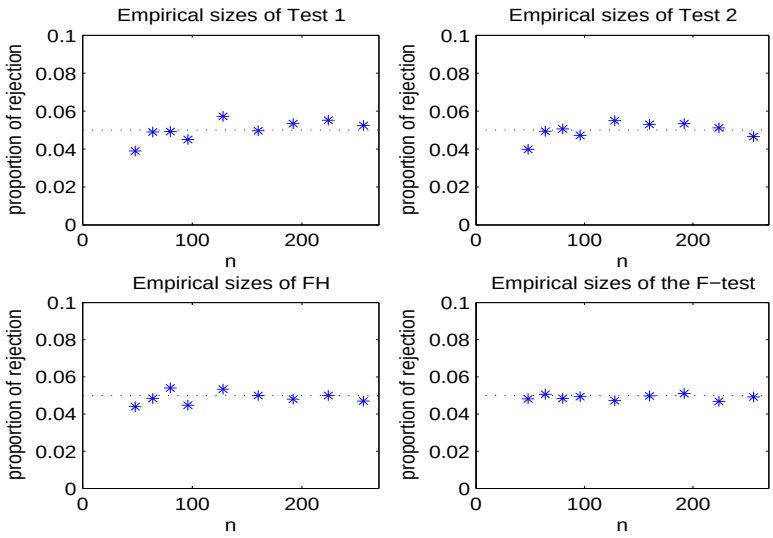

Figure 5.65: Empirical sizes for Example 15 under various sample sizes.

by the increasing number of regression parameters when data are drawn from a model with an odd-order polynomial. Figure 5.61 depicts the empirical sizes of the tests in various sample sizes.

Example 15. In this example, the response is drawn from

$$y = x_1 + \theta x_2^3 + 2x_4 + 3x_8 + \epsilon, \qquad \epsilon \sim N(0, 2^2) \tag{5.23}$$

which extends model (5.21) in Example 13.

Figure 5.62 shows a pattern similar to that in Figures 5.16 and 5.55. This example further illustrates the relative powers of all tests are not affected by the increasing number of regression parameters. Test 2 and FH still perform far worse than Test 1 and the F-test with Test 2 performing worst among all four tests. Figure 5.65 provides the plots of the empirical sizes of the tests against various sample sizes. Test 1 and Test 2 are slightly undersized in a small sample as $n = 48$ but perform well in other samples.

5.2.3 Summary

Summarizing all simulations in this section, the difference between Test 1 and FH is due more to the change of the highest order term in the polynomial in the true model than to the increasing number of regression parameters in the tested model. Test 2 is relatively less resistant to the increasing number of regression parameters in the tested model. It has a loss of power in many cases that we have investigated when the number of regression parameters in the tested model is increased. Since similar results on the empirical sizes of the tests are obtained in all examples in this section, the adjustment we make on our proposed tests performs well in various sample sizes.

Chapter 6

Application to Real Data

Table 6.1 contains 70 observations from Bruce and Schumacher (1935) on the volume, in cubic feet, of usable timber from short-leaf pine, together with two predictor variables x_1, the girth of each tree (the diameter at breast height), in inches and x_2, the height of the tree in feet. The aim is to find a formula for predicting volume from the girth and height. Figure 6.1 shows two scatterplots of the volume against the girth and height of the tree, respectively.

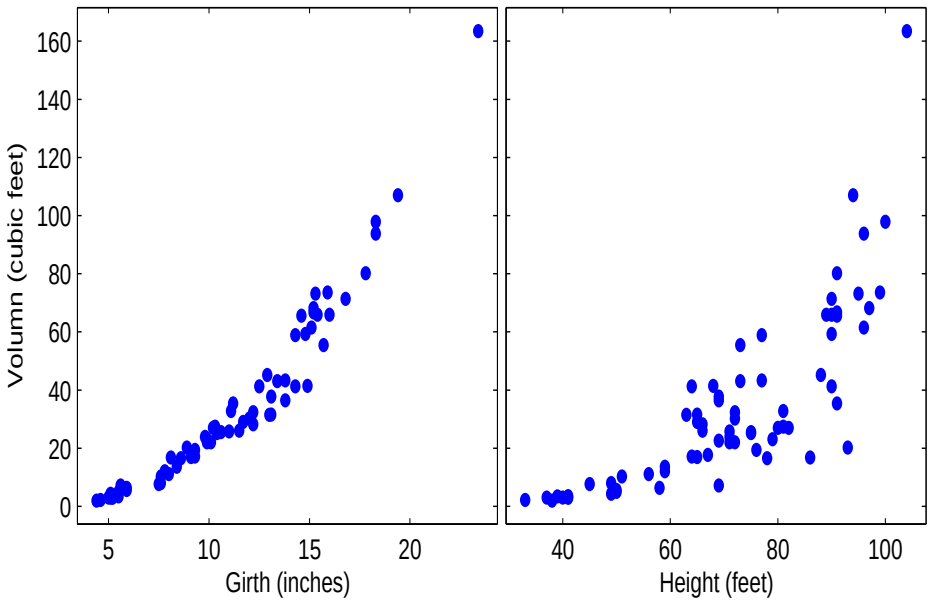

Figure 6.1: Scatterplots plotting y against x_1 and x_2 respectively.

Table 6.1: Short-leaf pine. The response y is the volume of the tree, x_1 is the girth and x_2 is the height.

Obs.	x_1	x_2	y	Obs.	x_1	x_2	y	Obs.	x_1	x_2	y
1	4.6	33	2.2	26	9.8	71	23.9	51	13.8	77	43.3
2	4.4	38	2.0	27	9.9	72	22.0	52	14.3	64	41.3
3	5.0	40	3.0	28	9.9	79	23.1	53	14.3	77	58.9
4	5.1	49	4.3	29	9.9	69	22.6	54	14.6	91	65.6
5	5.1	37	3.0	30	10.1	71	22.0	55	14.8	90	59.3
6	5.2	41	2.9	31	10.2	80	27.0	56	14.9	68	41.4
7	5.2	41	3.5	32	10.2	82	27.0	57	15.1	96	61.5
8	5.5	39	3.4	33	10.3	81	27.4	58	15.2	91	66.7
9	5.5	50	5.0	34	10.4	75	25.2	59	15.2	97	68.2
10	5.6	69	7.2	35	10.6	75	25.5	60	15.3	95	73.2
11	5.9	58	6.4	36	11.0	71	25.8	61	15.4	89	65.9
12	5.9	50	5.6	37	11.1	81	32.8	62	15.7	73	55.5
13	7.5	45	7.7	38	11.2	91	35.4	63	15.9	99	73.6
14	7.6	51	10.3	39	11.5	66	26.0	64	16.0	90	65.9
15	7.6	49	8.0	40	11.7	65	29.0	65	16.8	90	71.4
16	7.8	59	12.1	41	12.0	72	30.2	66	17.8	91	80.2
17	8.0	56	11.1	42	12.2	66	28.2	67	18.3	96	93.8
18	8.1	86	16.8	43	12.2	72	32.4	68	18.3	100	97.9
19	8.4	59	13.6	44	12.5	90	41.3	69	19.4	94	107.0
20	8.6	78	16.6	45	12.9	88	45.2	70	23.4	104	163.5
21	8.9	93	20.2	46	13.0	63	31.5				
22	9.1	65	17.0	47	13.1	69	37.8				
23	9.2	67	17.7	48	13.1	65	31.6				
24	9.3	76	19.4	49	13.4	73	43.1				
25	9.3	64	17.1	50	13.8	69	36.5				

Figure 6.1 shows the the volume of the tree has nonlinear marginal associations with the girth and height of the tree. We examine six models and test lack-of-fit in each of them. The models are:

$$
\begin{aligned}
&(1) \qquad y_i = \beta_{00} + \beta_{10} x_{i1} + \beta_{20} x_{i1}^2 + \epsilon_i, \\
&(2) \qquad y_i = \beta_{00} + \beta_{01} x_{i2} + \beta_{02} x_{i2}^2 + \epsilon_i, \\
&(3) \qquad \log(y_i) = \beta_{00} + \beta_{01} \log(x_{i1}) + \beta_{02} \log(x_{i2}) + \epsilon_i, \\
&(4) \qquad y_i = \beta_{00} + \beta_{10} x_{i1} + \beta_{20} x_{i1}^2 + \beta_{01} x_{i2} + \beta_{11} x_{i1} x_{i2} + \epsilon_i, \\
&(5) \qquad y_i = \beta_{00} + \beta_{10} x_{i1} + \beta_{01} x_{i2} + \beta_{02} x_{i2}^2 + \beta_{11} x_{i1} x_{i2} + \epsilon_i, \\
&(6) \qquad y_i = \beta_{00} + \beta_{10} x_{i1} + \beta_{20} x_{i1}^2 + \beta_{01} x_{i2} + \beta_{02} x_{i2}^2 + \beta_{11} x_{i1} x_{i2} + \epsilon_i.
\end{aligned}
$$

The lack-of-fit test statistics depend on ordering the observations. A good ordering should make the sequence $\{\epsilon_i\}$ as smooth as possible so that the large Fourier coefficients are concentrated on low frequencies. For the models with one covariate, the observations are ordered according to the covariate. For the models with both x_{i1} and x_{i2}, we considered four ordering methods. Let S be the sample covariance matrix of the covariates x_1 and x_2. Denote the ordered eigenvalues of S and their corresponding eigenvectors as λ_1 and λ_2, and ζ_1 and ζ_2 respectively. FH suggest using the sample score of variation of the observations to order the data. The sample score of variation of the i-th observation is given by

$$s_{1i} = \lambda_1 (\zeta_1^T x_i)^2 + \lambda_2 (\zeta_2^T x_i)^2 = x_i^T S x_i,$$

where $x_i^T = [x_{i1}, x_{i2}]$. We suggest an alternative to s_{1i},

$$s_{2i} = \frac{1}{\lambda_1} (\zeta_1^T x_i)^2 + \frac{1}{\lambda_2} (\zeta_2^T x_i)^2 = x_i^T S^{-1} x_i.$$

We also provide the results based on the ordering by the first principal component $s_{3i} \equiv \zeta_1^T x_i$,and the second principal component $s_{4i} \equiv \zeta_2^T x_i$. For the original data, s_{1i} is completely dominated by λ_1. Ordering by s_{3i} gives the same results as those by s_{1i}. Using $\log(x_1)$ and $\log(x_2)$ provides a covariance matter different from the original predictors, and hence completely different eigenvalues and eigenvectors.

Table 6.2: Test statistics and p-values of the tests to testing the lack-of-fit in model (1).

	Test 1	Test 2	FH	F
Test statistic	1.8547	3.0603	2.3342	1.3490
P-value	0.1449	0.0458	0.0923	0.2056

Table 6.3: Test statistics and p-values of the tests to testing the lack-of-fit in model (2).

	Test 1	Test 2	FH	F
Test statistic	10.6970	2.8110	1.9009	1.6627
P-value	0.0000	0.0584	0.1388	0.0859

Ordering by s_{3i} is no longer be equivalent to that by s_{1i} when the model is fitted to the log data. Therefore the results from ordering by s_{3i} are presented only for model (3). We use the four tests in Chapter 5 to evaluate lack-of-fit in the models. Since the F statistic has a distribution different from the other three test statistics. It is presented but not discussed in the comparisons among the test statistics. The p-value of the F statistic is used as a reference.

Intuitively, volumn of timber should depend on both x_1 and x_2. Models (1) and (2) only use one of the variables. The tests are presented in Tables 6.2 and 6.3. Test 2 is the only test to suggest lack-of-fit in model (1). Test 2 also suggests lack-of-fit in model (2) but Test 1 clearly detects lack-of-fit in model (2). Recall that in Chapter 5, the simulated results show that FH's test lost power dramatically when the tested model is a polynomial regression with order higher than 1. Presumably because low order polynomials mimic low order frequencies. The right panel of Figure 6.1 show a weaker quadratic association between the volume and the height of a tree relative to that between the volume and the girth of a tree.

In Tables 6.2 and 6.3, Test 1 and the F-test share similar characteristics. In Table 6.2, both tests give larger p-values relative to FH. The situation is opposite in Table 6.3, both tests give smaller p-values relative to FH. We suspect that

Table 6.4: Test statistics and p-values of the tests to testing the lack-of-fit in model (3).

Ordering		Test 1	Test 2	FH	F
s_{1i}	Test statistic	0.2583	1.6084	0.4529	1.5252
	P-value	0.5381	0.1814	0.4705	0.1272
s_{2i}	Test statistic	4.4603	1.8694	2.7399	1.5308
	P-value	0.0115	0.1429	0.0629	0.1252
s_{3i}	Test statistic	−0.0669	1.1933	0.5009	1.4322
	P-value	0.6567	0.2399	0.4545	0.1645
s_{4i}	Test statistic	2.8169	1.0917	3.3001	2.2642
	P-value	0.0580	0.2851	0.0362	0.0140

these may due to the relative smoothness of the unknown true error of the models and the residuals of the models. Suppose the true model is

$$Y = X\beta + H(X) + \epsilon.$$

Test 1 and the F-test are derived based on model

$$Y - X\beta = H(X) + \epsilon,$$

and then approximate the vector of lack-of-fit, $H(X)$, by Fourier series. FH derive their test based on model

$$Y - X\hat{\beta} = H(X) + e,$$

where $\hat{\beta}$ is a least squares estimate from fitting the tested model, and then approximate the residuals by Fourier series. Therefore under the true model, Test 1 and the F-test only need the function $H(X)$ to be a smooth function. But FH requires $(I - M_X)H(X)$ to be smooth. So the performance of these tests may rely on the ordered errors or fitted residuals being a relatively smooth function.

Models (3), (4), (5) and (6) illustrate testing of lack-of-fit with two distinct predictors. A natural approximation to the volume of a tree is the volume of a

Table 6.5: Test statistics and p-values of the tests to testing the lack-of-fit in model (4).

Ordering		Test 1	Test 2	FH	F
s_{1i}	Test statistic	-0.2009	-1.8063	-2.2599	0.8196
	P-value	0.7055	0.9977	0.9999	0.6579
s_{2i}	Test statistic	-0.0661	0.4140	-0.0390	1.4392
	P-value	0.6564	0.4837	0.6465	0.1632
s_{4i}	Test statistic	1.2465	-0.1598	-0.4728	1.5295
	P-value	0.2499	0.6907	0.7990	0.1274

Table 6.6: Test statistics and p-values of the tests to testing the lack-of-fit in model (5).

Ordering		Test 1	Test 2	FH	F
s_{1i}	Test statistic	0.3740	-1.0365	-0.9355	1.5736
	P-value	0.4974	0.9404	0.9218	0.1126
s_{2i}	Test statistic	3.9133	3.5791	1.9568	2.5217
	P-value	0.0198	0.0275	0.1318	0.0067
s_{4i}	Test statistic	4.8753	7.3913	6.4275	2.5184
	P-value	0.0076	0.0006	0.0016	0.0068

Table 6.7: Test statistics and p-values of the tests to testing the lack-of-fit in model (6).

Ordering		Test 1	Test 2	FH	F
s_{1i}	Test statistic	-0.5043	0.1161	-0.6918	1.2051
	P-value	0.8091	0.5895	0.8643	0.2987
s_{2i}	Test statistic	-0.4370	0.6321	-0.2049	1.2318
	P-value	0.7873	0.4123	0.7069	0.2799
s_{4i}	Test statistic	-2.4813	-1.7624	-2.8969	0.5842
	P-value	1.0000	0.9971	1.0000	0.8802

cylinder, which is proportional to the product of the squared radius and the height. Atkinson and Riani (2000) suggest that model (3) is an adequate model for the data. Table 6.4 summarizes the results from lack-of-fit testing. When the data are ordered by the sample score of variation or the first principal component, all tests agree with Atkinson and Riani's analysis and identify no lack-of-fit. However, Test 1 and FH give small p-values with the other two orderings. Both Test 1 and FH suggest that model (3) may not be an adequate model for the data, that is, the volume of a tree may not be well approximated by the volume of a cylinder.

Models (4) and (5) are sub-models of model (6). The multiple regressions contain the girth, the height, and the interaction between these two predictors. The squared girth, the squared height, and both squared girth and squared height are included in models (4), (5), and (6) respectively. Tables 6.5, 6.6, and 6.7 present the results. None of the tests find evidence for lack-of-fit in models (4) or (6). However, when the data are ordered in s_{2i}, all tests except FH identify lack-of-fit in model (5), which is the only model of the three without squared girth. Both Test 1 and Test 2 are good at picking up the need for the squared girth. The ordering criterion suggested by FH may not be suitable for this study. The ordering criterion we suggested, s_{2i}, provides relatively small p-values. Moreover ordering by s_{4i} provides uniformly small p-values.

Chapter 7

Summary and Conclusion

We have proposed two new lack of fit tests, found their asymptotic distributions, modified their standardization constants to improve small sample test sizes, and studied their powers. Our first proposed test typically has good power and often has the best power. Our second proposed test has the best power only when testing the lack-of-fit for some simple linear regressions.

Any smooth function can be approximated by its Fourier transform, so all of the proposed tests adapt to a large class of models for lack-of-fit. Moreover, the theory extends beyond Fourier transforms. The matrix Ψ defining the Fourier transform in Section 3.1 can be generalized. The properties of Ψ_m required in the proofs are the orthogonality and the hierarchical structure of Ψ_m for $m = 1, \ldots, n$. Orthogonality can be achieved by applying Gram-Schmidt. For example, in simple regression, we can extend Green's (1971) test with the technique used here by redefining Ψ_m based on polynomials rather than sines and cosines. The polynomials would have to be orthogonalized. Wavelets could also be used.

Moreover, the procedures can be applied to "near replicate" clusters of covariates. Suppose the indicator matrix Z in model (4.2) is redefined to Z_m such that m is the number of clusters. When the clustering has a hierarchical structure, our proposed test can be applied.

Further investigation on the ordering of the observations in multiple regressions will be done. From Chapter 6, we found that the ordering of the observations can have a huge impact on the performance of the smooth tests. Finding specific ordering procedures for each of our proposed tests could further enhance the performance these tests.

Appendix A: Proofs

The ideas and techniques used in the proofs of Theorems 1 and 2 are quite different from those in FH.

A.1 Proof of Theorem 1

To prove Theorems 1 and 2, we need two lemmas.

Lemma 4. *For any $r_n \geq 1$, define*

$$T_n = \max_{1\leq k\leq r_n}\left\{\frac{\sum_{i=1}^k[(o_i^TY)^2-\sigma^2]}{\sqrt{2k}\sigma^2}\right\} \quad \textit{and} \quad \widehat{T}_n = \max_{1\leq k\leq r_n}\left\{\frac{\sum_{i=1}^k[(o_i^TY)^2-\hat{\sigma}^2]}{\sqrt{2k}\hat{\sigma}^2}\right\}.$$

We then have

$$T_n = \frac{\hat{\sigma}^2}{\sigma^2}\widehat{T}_n + Q_n,$$

$$\textit{where } \min_{1\leq k\leq r_n}\left\{\left(\frac{\hat{\sigma}^2}{\sigma^2}-1\right)\sqrt{\frac{k}{2}}\right\} \leq Q_n \leq \max_{1\leq k\leq r_n}\left\{\left(\frac{\hat{\sigma}^2}{\sigma^2}-1\right)\sqrt{\frac{k}{2}}\right\}.$$

Proof of Lemma 4

From the expression for T_n,

$$\begin{aligned} T_n &= \max_{1\leq k\leq r_n}\left\{\frac{\sum_{i=1}^k[(o_i^TY)^2-\sigma^2]}{\sqrt{2k}\sigma^2}\right\} \\ &= \frac{\hat{\sigma}^2}{\sigma^2}\max_{1\leq k\leq r_n}\left\{\frac{\sum_{i=1}^k[(o_i^TY)^2-\hat{\sigma}^2]}{\sqrt{2k}\hat{\sigma}^2}+\left(1-\frac{\sigma^2}{\hat{\sigma}^2}\right)\sqrt{\frac{k}{2}}\right\}. \end{aligned}$$

Then,

$$\begin{aligned} T_n &\leq \frac{\hat{\sigma}^2}{\sigma^2} \max_{1\leq k\leq r_n} \left\{ \frac{\sum_{i=1}^{k}[(o_i^T Y)^2 - \hat{\sigma}^2]}{\sqrt{2k}\hat{\sigma}^2} \right\} + \frac{\hat{\sigma}^2}{\sigma^2} \max_{1\leq k\leq r_n} \left\{ \left(1 - \frac{\sigma^2}{\hat{\sigma}^2}\right) \sqrt{\frac{k}{2}} \right\} \\ &= \frac{\hat{\sigma}^2}{\sigma^2} \widehat{T}_n + \max_{1\leq k\leq r_n} \left\{ \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right) \sqrt{\frac{k}{2}} \right\}; \end{aligned}$$

and

$$\begin{aligned} T_n &\geq \frac{\hat{\sigma}^2}{\sigma^2} \max_{1\leq k\leq r_n} \left\{ \frac{\sum_{i=1}^{k}[(o_i^T Y)^2 - \hat{\sigma}^2]}{\sqrt{2k}\hat{\sigma}^2} \right\} + \frac{\hat{\sigma}^2}{\sigma^2} \min_{1\leq k\leq r_n} \left\{ \left(1 - \frac{\sigma^2}{\hat{\sigma}^2}\right) \sqrt{\frac{k}{2}} \right\} \\ &= \frac{\hat{\sigma}^2}{\sigma^2} \widehat{T}_n + \min_{1\leq k\leq r_n} \left\{ \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right) \sqrt{\frac{k}{2}} \right\}. \end{aligned}$$

Lemma 4 follows. □

Lemma 5. *For any* $\delta > 0$ *and* $1 \leq r_n \leq \dfrac{n}{(\log\log n)^{1+\delta}}$, *if* $\dfrac{\hat{\sigma}^2}{\sigma^2} - 1 = O_p(n^{-1/2})$, *then*

$$a_{r_n} \sup_{1\leq k\leq r_n} \left| \sqrt{\frac{k}{2}} \right| \left| \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right) \right| \xrightarrow{P} 0 \quad \textit{and} \quad b_{r_n} \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right) \xrightarrow{P} 0 \quad \textit{as} \quad n \to \infty.$$

Proof of Lemma 5

Obviously, $b_{r_n} = O(\log\log r_n)$. As $\dfrac{\hat{\sigma}^2}{\sigma^2} - 1 = O_p(n^{-1/2})$, $b_{r_n}\left(\dfrac{\hat{\sigma}^2}{\sigma^2} - 1\right) \xrightarrow{p} 0$ as $n \to \infty$. $1 \leq r_n \leq \dfrac{n}{(\log\log n)^{1+\delta}}$ and $\dfrac{\hat{\sigma}^2}{\sigma^2} - 1 = O_p(n^{-1/2})$ yield,

$$a_{r_n} \sup_{1\leq k\leq r_n} \left| \sqrt{\frac{k}{2}} \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1\right) \right| = a_{r_n} \sqrt{\frac{r_n}{2}} \left| \frac{\hat{\sigma}^2}{\sigma^2} - 1 \right| = O_p\left((\log\log n)^{-\delta/2}\right).$$

Thus, $a_{r_n} \sup_{1\leq k\leq r_n} \left| \sqrt{\dfrac{k}{2}} \left(\dfrac{\hat{\sigma}^2}{\sigma^2} - 1\right) \right| \xrightarrow{p} 0$ as $n \to \infty$ and Lemma 5 follows. □

Proof of Theorem 1

We begin by showing the result for known σ^2. With $Y \sim N(X\beta, \sigma^2 I)$,

$$\frac{Y^T M_{(I-M_X)\Gamma_m} Y}{\sigma^2} \sim \chi^2(r_m),$$

Obviously, $C[(I - M_X)\Gamma_m] \subseteq C[(I - M_X)\Gamma_{m+1}]$ for $m = 1, \ldots, \widetilde{n} - 1$. Use the Gram-Schmidt algorithm to orthonormalize the columns of $(I - M_X)\Gamma_{\widetilde{n}}$ and let

O_m be the matrix whose columns form an orthonormal basis for $C[(I - M_X)\Gamma_m]$. Write

$$O_m = [o_1, o_2, \ldots, o_{r_m}],$$

for $m = 1, \ldots, \widetilde{n}$. Since $M_{(I-M_X)\Gamma_m} = O_m O_m^T = \sum_{i=1}^{r_m} o_i o_i^T$,

$$\frac{Y^T M_{(I-M_X)\Gamma_m} Y}{\sigma^2} = \sum_{i=1}^{r_m} \frac{(o_i^T Y)^2}{\sigma^2}.$$

Define $\nu_i = (o_i^T Y)^2/\sigma^2$ so that under model (3.1), for $i = 1, \ldots, r_{\widetilde{n}}$, the ν_i's are i.i.d. $\chi^2(1)$. The usual standardization gives

$$U_i = \frac{\nu_i - 1}{\sqrt{2}}.$$

Define $S_k = \sum_{i=1}^{k} U_i$ and

$$T_{1,\widetilde{n}} = \max_{1 \le k \le r_{\widetilde{n}}} \frac{S_k}{\sqrt{k}} = \max_{1 \le k \le r_{\widetilde{n}}} \left\{ \frac{\sum_{i=1}^{k} [(o_i^T Y)^2 - \sigma^2]}{\sqrt{2k}\sigma^2} \right\}.$$

Define a standardized version of $T_{1,\widetilde{n}}$,

$$W_{1,\widetilde{n}} = a_{r_{\widetilde{n}}} T_{1,\widetilde{n}} - b_{r_{\widetilde{n}}}$$

with $a_{r_{\widetilde{n}}}$ and $b_{r_{\widetilde{n}}}$ defined earlier. Since $r_{\widetilde{n}} \to \infty$ as $n \to \infty$, and $E|\nu_i|^3 < \infty$, the Darling-Erdős theorem applies with the replacement of n by $r_{\widetilde{n}}$, so that

$$Pr(W_{1,\widetilde{n}} < x) \to \exp(-\exp(-x)) \qquad \text{as} \quad n \to \infty.$$

Incorporating the estimate of σ^2, define $\hat{\nu}_i = (o_i^T Y)^2/\hat{\sigma}^2$, $\widehat{U}_i = (\hat{\nu}_i - 1)/\sqrt{2}$, and $\widehat{S}_k = \widehat{U}_1 + \widehat{U}_2 + \cdots + \widehat{U}_k$. Thus our test statistic $\widehat{T}_{1,\widetilde{n}}$ can be expressed as

$$\widehat{T}_{1,\widetilde{n}} = \max_{1 \le k \le r_{\widetilde{n}}} \frac{\widehat{S}_k}{\sqrt{k}}.$$

By Lemma 4,

$$\begin{aligned} W_{1,\widetilde{n}} &= a_{r_{\widetilde{n}}} T_{1,\widetilde{n}} - b_{r_{\widetilde{n}}} \\ &= a_{r_{\widetilde{n}}} \left(\frac{\hat{\sigma}^2}{\sigma^2} \widehat{T}_{1,\widetilde{n}} + Q_{\widetilde{n}} \right) - b_{r_{\widetilde{n}}} \\ &= \frac{\hat{\sigma}^2}{\sigma^2} \widehat{W}_{1,\widetilde{n}} + a_{r_{\widetilde{n}}} Q_{\widetilde{n}} + b_{r_{\widetilde{n}}} \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1 \right). \end{aligned}$$

Recall that we have assumed $\frac{\hat{\sigma}^2}{\sigma^2} - 1 = O_p(n^{-1/2})$. (This assumption is equivalent to the Fan and Huang's assumption A2.) This assumption provides $\frac{\hat{\sigma}^2}{\sigma^2} \xrightarrow{p} 1$ as $n \to \infty$. Furthermore, $r_{\widetilde{n}} \leq \widetilde{n} = \frac{n}{(\log\log n)^{1+\delta}}$. By Lemma 5, $a_{r_{\widetilde{n}}} Q_{\widetilde{n}} \xrightarrow{p} 0$ and $b_{r_{\widetilde{n}}} \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1 \right) \xrightarrow{p} 0$. These imply $\widehat{W}_{1,\widetilde{n}}$ and $W_{1,\widetilde{n}}$ have the same asymptotic distribution. Thus,

$$Pr(\widehat{W}_{1,\widetilde{n}} < x) \to \exp(-\exp(-x)) \qquad \text{as} \quad n \to \infty.$$

□

A.2 Proof of Theorem 2

Before we prove Theorem 2, we need the following Lemma.

Lemma 6. *For* $m = 1, 2, \ldots, \widetilde{n} - 1$,

$$C(M_m - M_{M_m X}) \subseteq C(M_{m+1} - M_{M_{m+1} X}).$$

Proof of Lemma 6

Recall that $\Gamma_{\widetilde{n}} = [G_1, G_2, \ldots, G_{\widetilde{n}}]$. Let $\widehat{m}$ be the smallest integer such that $r(M_{\widehat{m}} X) = r(X)$. For any $m \geq \widehat{m}$, $C(M_m X) = C(M_{m+1} X)$. Thus, $M_{M_m X} = M_{M_{m+1} X}$ and the Lemma holds. For any $m \leq \widehat{m}$, we consider the followings. Note that

$$\begin{aligned} M_{m+1} &= \Gamma_{m+1} \Gamma_{m+1}^T \\ &= [\Gamma_m, G_{m+1}] [\Gamma_m, G_{m+1}]^T \\ &= \Gamma_m \Gamma_m^T + G_{m+1} G_{m+1}^T \\ &= M_m + M_{G_{m+1}}, \end{aligned}$$

where $M_{G_{m+1}}$ is the perpendicular projection operator onto $C(G_{m+1})$. Furthermore, $M_{m+1} X = \left(M_m + M_{G_{m+1}} \right) X = M_m X + M_{G_{m+1}} X$. Since $C(M_m X)$ and $C(M_{G_{m+1}} X)$ are orthogonal, this yields

$$M_{M_{m+1} X} = M_{M_m X} + M_{M_{G_{m+1}} X}.$$

Then,

$$\begin{aligned} M_{m+1} - M_{M_{m+1}X} &= M_m + M_{G_{m+1}} - (M_{M_mX} + M_{M_{G_{m+1}}X}) \\ &= (M_m - M_{M_mX}) + (M_{G_{m+1}} - M_{M_{G_{m+1}}X}). \end{aligned}$$

Obviously, $C(M_{G_{m+1}}X) \subseteq C(G_{m+1})$. Since G_{m+1} is a non-zero column vector, $r(G_{m+1}) = 1$. This implies $r(M_{G_{m+1}}X)$ can only be 0 or 1. If $r(M_{G_{m+1}}X) = 0$, we have $M_{M_{G_{m+1}}X} = 0$. Thus,

$$\begin{aligned} C(M_{m+1} - M_{M_{m+1}X}) &= C(M_m - M_{M_mX}, M_{G_{m+1}}) \\ &\supset C(M_m - M_{M_mX}); \end{aligned}$$

If $r(M_{G_{m+1}}X) = 1$, we have $M_{M_{G_{m+1}}X} = M_{G_{m+1}}$. And thus,

$$C(M_{m+1} - M_{M_{m+1}X}) = C(M_m - M_{M_mX}).$$

Lemma 6 follows. □

Using Lemma 6, the proof of Theorem 2 is similar to the proof of Theorem 1.

Proof of Theorem 2

Define $O^*_m = [o^*_1, o^*_2, \ldots, o^*_{\widetilde{r}_m}]$ be the matrix who's columns form an orthonormal basis of the column space $C(M_m - M_{M_mX})$. From the proof of Lemma 6, we have

$$\begin{aligned} & Y^T(M_{m+1} - M_{M_{m+1}X})Y \\ &= \begin{cases} Y^T(M_m - M_{M_mX})Y & \text{if } \widetilde{r}_{m+1} = \widetilde{r}_m \\ Y^T(M_m - M_{M_mX} + M_{G_{m+1}})Y & \text{if } \widetilde{r}_{m+1} = \widetilde{r}_m + 1 \end{cases} \\ &= \begin{cases} Y^T(O^*_m O^{*T}_m)Y & \text{if } \widetilde{r}_{m+1} = \widetilde{r}_m \\ Y^T(O^*_m O^{*T}_m)Y + Y^T G_{m+1} G^T_{m+1} Y & \text{if } \widetilde{r}_{m+1} = \widetilde{r}_m + 1 \end{cases} \\ &= \begin{cases} \sum_{i=1}^{\widetilde{r}_m} (o^{*T}_i Y)^2 & \text{if } \widetilde{r}_{m+1} = \widetilde{r}_m \\ \sum_{i=1}^{\widetilde{r}_m+1} (o^{*T}_i Y)^2 & \text{if } \widetilde{r}_{m+1} = \widetilde{r}_m + 1 \end{cases}, \end{aligned}$$

for $m = 1, 2, \ldots, \widetilde{n} - 1$, and define $o_{\widetilde{r}_m+1} \equiv G_{m+1}$ if $\widetilde{r}_{m+1} = \widetilde{r}_m + 1$. Denote $\nu_{2,i} = (o^{*T}_i Y)^2/\sigma^2$ and $\nu_{2,i}$'s are i.i.d. random variables following central χ^2-distributions with one degree of freedom for $i = 1, \ldots, \widetilde{r}_{\widetilde{n}}$ under model (3.1). What

followings will be similar to our earlier proof. Since $E(\nu_{2,i}) = 1$, and $Var(\nu_{2,i}) = 2$; we can normalize $\nu_{2,i}$ as

$$U_{2,i} = \frac{\nu_{2,i} - 1}{\sqrt{2}}.$$

Let $S_{2,k} = \sum_{i=1}^{k} U_{2,i}$ and $T_{2,\widetilde{n}} = \max_{1\le k\le \widetilde{r}_{\widetilde{n}}} \frac{S_{2,k}}{\sqrt{k}} = \max_{1\le k\le \widetilde{r}_{\widetilde{n}}} \left\{ \frac{\sum_{i=1}^{k}[(o_i^{*T}Y)^2 - \sigma^2]}{\sqrt{2k}\sigma^2} \right\}$. We define $\widetilde{\nu}_{2,i} = (o_i^{*T}Y)^2/\hat{\sigma}^2$, $\widetilde{U}_{2,i} = (\widetilde{\nu}_{2,i} - 1)/\sqrt{2}$, and $\widetilde{S}_{2,k} = \widetilde{U}_{2,1} + \widetilde{U}_{2,2} + \cdots + \widetilde{U}_{2,k}$. The test statistic we proposed, $\widetilde{T}_{2,\widetilde{n}}$, can be written as

$$\begin{aligned}
\widetilde{T}_{2,\widetilde{n}} &= \max_{1\le m\le \widetilde{n}} \left\{ \sqrt{\frac{\widetilde{r}_m}{2}} \frac{Y^T(M_m - M_{M_mX})Y/\widetilde{r}_m - \hat{\sigma}^2}{\hat{\sigma}^2} \right\} \\
&= \max_{m\in\mathbb{K}} \left\{ \sqrt{\frac{\widetilde{r}_m}{2}} \frac{Y^T(M_m - M_{M_mX})Y/\widetilde{r}_m - \hat{\sigma}^2}{\hat{\sigma}^2} \right\} \\
&= \max_{1\le k\le \widetilde{r}_{\widetilde{n}}} \left\{ \sqrt{\frac{k}{2}} \frac{\sum_{i=1}^{k}(o_i^{*T}Y)^2/k - \hat{\sigma}^2}{\hat{\sigma}^2} \right\} \\
&= \max_{1\le k\le \widetilde{r}_{\widetilde{n}}} \left\{ \frac{\sum_{i=1}^{k}[(o_i^{*T}Y)^2 - \hat{\sigma}^2]}{\sqrt{2k}\hat{\sigma}^2} \right\} \\
&= \max_{1\le k\le \widetilde{r}_{\widetilde{n}}} \frac{\widetilde{S}_{2,k}}{\sqrt{k}}.
\end{aligned}$$

The asymptotic distribution of $\widetilde{T}_{2,\widetilde{n}}$ can be obtained by considering the asymptotic distribution of $T_{2,\widetilde{n}}$. Since $\widetilde{r}_{\widetilde{n}} \ge \widetilde{n} - p$, we have $\widetilde{r}_{\widetilde{n}} \to \infty$ as $n \to \infty$. Furthermore, $E|\nu_{2,i}|^3 < \infty$, the Darling-Erdős theorem applies with the replacement of n by $\widetilde{r}_{\widetilde{n}}$. Define

$$W_{2,\widetilde{n}} = a_{\widetilde{r}_{\widetilde{n}}} T_{2,\widetilde{n}} - b_{\widetilde{r}_{\widetilde{n}}},$$

where $a_{\widetilde{r}_{\widetilde{n}}}$ and $b_{\widetilde{r}_{\widetilde{n}}}$ as defined earlier. We have

$$Pr(W_{2,\widetilde{n}} < x) \to \exp(-\exp(-x)) \qquad \text{as} \quad n \to \infty.$$

Similar to the proof of Theorem 1, by Lemma 4, we have

$$W_{2,\widetilde{n}} = \frac{\hat{\sigma}^2}{\sigma^2}\widetilde{W}_{2,\widetilde{n}} + a_{\widetilde{r}_{\widetilde{n}}} Q_{\widetilde{n}} + b_{\widetilde{r}_{\widetilde{n}}} \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1 \right).$$

As $\widetilde{r}_{\widetilde{n}} \le \widetilde{n} = \frac{n}{(\log\log n)^{1+\delta}}$, Lemma 5 gives $a_{\widetilde{r}_{\widetilde{n}}} Q_{\widetilde{n}} \xrightarrow{p} 0$ and $b_{\widetilde{r}_{\widetilde{n}}} \left(\frac{\hat{\sigma}^2}{\sigma^2} - 1 \right) \xrightarrow{p} 0$. Together with the assumption that $\frac{\hat{\sigma}^2}{\sigma^2} \xrightarrow{p} 1$ as $n \to \infty$, $\widetilde{W}_{2,\widetilde{n}}$ and $W_{2,\widetilde{n}}$ have the

same asymptotic distribution. Thus,

$$Pr(\widetilde{W}_{2,\tilde{n}} < x) \to \exp(-\exp(-x)) \qquad \text{as} \quad n \to \infty.$$

□

A.3 Proof of Lemma 3

Proof of Lemma 3

That $\hat{\sigma}_i^2 = \sigma^2 + O_p(n^{-1/2})$ is equivalent to $\sqrt{n}(\hat{\sigma}_i^2 - \sigma^2) = O_p(1)$. Both estimates take the form $Y^T(I - M_A)Y/(n - r(A))$ for a ppo M_A. It follows that under model (1) $Y^T(I - M_A)Y/\sigma^2 \sim \chi^2(n - r(A))$. The lemma is a direct application of Chebyshev's inequality using the variance of a χ^2. It is not completely obvious that $M_X + M_K - M_{M_K X}$ is a ppo.

Let us prove Lemma 3 holds for $\hat{\sigma}_i^2$. By Chebyshev's inequality, for any $\varepsilon > 0$, we have

$$\begin{aligned} Pr(\sqrt{n}|\hat{\sigma}_i^2 - \sigma^2| > \varepsilon) &\leq \frac{n}{\varepsilon^2} E(\hat{\sigma}_i^2 - \sigma^2)^2 \\ &= \frac{n}{\varepsilon^2} E\left(\frac{Y^T(I - M_A)Y}{n - r(A)} - \sigma^2\right)^2 \\ &= \frac{n\sigma^4}{\varepsilon^2(n - r(A))^2} E\left(\frac{Y^T(I - M_A)Y}{\sigma^2} - (n - r(A))\right)^2. \end{aligned}$$

Under model (3.1), $\dfrac{Y^T(I - M_A)Y}{\sigma^2} \sim \chi^2(n - r(A))$. This and $r(A) \leq K + p$ yield

$$\begin{aligned} Pr(\sqrt{n}|\hat{\sigma}_i^2 - \sigma^2| > \varepsilon) &\leq \frac{n\sigma^4}{\varepsilon^2(n - r(A))^2} Var\left(\frac{Y^T(I - M_A)Y}{\sigma^2}\right) \\ &= \frac{2n\sigma^4}{\varepsilon^2(n - r(A))} \\ &\leq \frac{2\sigma^4}{\varepsilon^2(1 - (K + p)/n)} \\ &\to \frac{2\sigma^4}{\varepsilon^2(1 - c)} \qquad \text{as } n \to \infty \\ &\to 0 \qquad \text{as } \varepsilon \to \infty. \end{aligned}$$

Lemma 3 holds for $\hat{\sigma}_1^2$. To complete the proof for $\hat{\sigma}_2^2$ it suffices to show $(I - M_X) - (M_K - M_{M_K X})$ is a ppo. Since $(X^T M_K X)^-$ is a generalized inverse of $(M_K X)^T (M_K X)$, we have

$$M_K X (X^T M_K X)^- X^T M_K X \;=\; M_K X.$$

This and its transpose yields

$$\begin{aligned} M_X (M_K - M_{M_K X}) &= M_X M_K - X(X^T X)^- X^T M_K X (X^T M_K X)^- X^T M_K \\ &= M_X M_K - X(X^T X)^- X^T M_K \\ &= M_X M_K - M_X M_K \\ &= 0. \end{aligned}$$

Therefore $C(M_K - M_{M_K X}) \subseteq C(I - M_X)$ and thus $(I - M_X) - (M_K - M_{M_K X})$ is a perpendicular projection operator. Similarly, under model (3.1),

$$\frac{Y^T \left[(I - M_X) - (M_K - M_{M_K X}) \right] Y}{\sigma^2} \sim \chi^2(n - r_3).$$

With this and the fact that $r_3 \leq K + p$, Lemma 3 holds for $\hat{\sigma}_2^2$. □

Appendix B: Review of Clustering Tests

B.1 Green's Test:

With one covariate x, Green investigates testing lack-of-fit for linear models

$$y_i = f(x_i)^T\beta + \epsilon_i, \tag{B.1}$$

where $f(x_i)$ is a $p \times 1$ vector of known functions of x_i, β is a vector of regression parameters, and ϵ_i are i.i.d. $N(0, \sigma^2)$ random variables for $i = 1, \ldots, N$. When replicates are available, as we mentioned at the end of last section, a most general model can be generated from the tested model and the exact F-test applies.

If replicates are not available, a most general model version of the tested model does not exist. Green proposed a test based on the idea of near-replicates. Near-replicates are cases grouped together to form clusters based on any specific criterion applied the covariates. After grouping the units into k clusters, index the units as ij, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$ and use the notation defined in (2.3) but re-defining

$$X_i^T = [x_{i1}, x_{i2}, \ldots, x_{in_i}], \text{ and } X^T = [X_1^T, X_2^T, \ldots, X_k^T].$$

Let $f(X_i)^T = [f(x_{i1})^T, f(x_{i2})^T, \ldots, f(x_{in_i})^T]$. Assume the true mean response is known from prior information, i.e. $E(Y)$. Let $\widetilde{X} \equiv f(X)$. Model (B.1) is written

in matrices as

$$Y = \widetilde{X}\beta + \epsilon.$$

If β is known, the difference between $E(Y)$ and $\widetilde{X}\beta$ can be used to group observations. Green suggests the clustering follows a criterion that $E(Y) - \widetilde{X}\beta$ will be well approximated by a qth-order polynomial in x within each cluster. In practice, $E(Y)$ is not known. We can use the residual plot to make decisions on clustering. The residuals from fitting a model may show some patterns in the residual plot. These patterns may be more conspicuous when they are broken up into several sub-patterns. Green's idea is breaking up these patterns into several sub-patterns in which each of the sub-patterns can be well approximated by a qth-order polynomial in x. Hence, the observations corresponding to each of the sub-patterns form a cluster. Other clustering criteria are proposed in Utts (1982), and Miller, Neill, and Sherfey (1998).

Although a most general version of the tested model does not exist, a more general model than model (B.1) can be constructed using Green's clustering criterion. Define $p(x_{ij})^T = [1, x_{ij}, x_{ij}^2, \ldots, x_{ij}^q]$. The mean structure of model (B.1) can differ from the true mean by a qth-order polynomial in x within each cluster. Thus, more general model version of model (B.1) can be written as

$$y_{ij} = f(x_{ij})^T\beta + p(x_{ij})^T\gamma_i + \epsilon_{ij}, \tag{B.2}$$

where γ_i is a $(q+1)\times 1$ vector of regression parameters and $p(x_{ij})\gamma_i$ is a qth-order polynomial in x_{ij}. Denote $P_i^T \equiv [p(x_{i1})^T, \ldots, p(x_{in_i})^T]$, and define

$$P = \begin{bmatrix} P_1 & 0 & 0 & 0 & 0 \\ 0 & P_2 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & P_{k-1} & 0 \\ 0 & 0 & 0 & 0 & P_k \end{bmatrix}. \tag{B.3}$$

Then models (B.1) and (B.2) can be written in matrix notation as

$$Y = \widetilde{X}\beta + \epsilon, \tag{B.4}$$

and

$$Y = \widetilde{X}\beta + P\gamma + \epsilon, \tag{B.5}$$

respectively. In model (B.5), γ is a $k(q+1) \times 1$ vector of regression parameters, i.e. $\gamma^T = [\gamma_1^T, \ldots, \gamma_k^T]$. Following Christensen (2002, Chapter 9), rewrite model (B.5) as

$$\begin{aligned} Y &= (I - M_P)\widetilde{X}\beta + P\gamma' + \epsilon \\ &\equiv X'\beta + P\gamma' + \epsilon, \end{aligned} \tag{B.6}$$

where $X' \equiv (I - M_P)\widetilde{X}$ so that $C(P)$ and $C(X')$ are orthogonal. The sum of squares error of model (B.6) is

$$SSE(B.6) = Y^T(I - M_P - M_{X'})Y.$$

Obviously, $C(\widetilde{X}) \subset C(X', P) = C(\widetilde{X}, P)$. An exact F-test can be applied to testing model (B.4) against the more general model (B.6). The sum of squares for pure error is chosen to be $SSE(B.6)$. The sum of squares for lack-of-fit is

$$\begin{aligned} SSE(B.4) - SSE(B.6) &= Y^T(I - M_{\widetilde{X}})Y - Y^T(I - M_P - M_{X'})Y \\ &= Y^T(M_P + M_{X'} - M_{\widetilde{X}})Y. \end{aligned}$$

The test statistic proposed by Green is

$$F = \frac{Y^T(M_P + M_{X'} - M_{\widetilde{X}})Y/(k(q+1) + r(X') - r(\widetilde{X}))}{Y^T(I - M_P - M_{X'})Y/(N - k(q+1) - r(X'))}. \tag{B.7}$$

The tested model is rejected at level α if F exceeds the critical value

$$F_{\alpha,\ k(q+1)+r(X')-r(\widetilde{X}),\ N-k(q+1)-r(X')}.$$

It is worth noting that if $f(x_i)^T = [1, x_i, \ldots, x_i^\iota]$, $\iota = 1, \ldots, q$, then $C(\widetilde{X}) \subseteq C(P)$. This yields $C(X')$ a zero vector. (B.7) becomes

$$F = \frac{Y^T(M_P - M_{\widetilde{X}})Y/(k(q+1) - r(\widetilde{X}))}{Y^T(I - M_P)Y/(N - k(q+1))}.$$

Green points out that the strength of the test depends on the quality of clustering, that is, the approximation to $E(Y) - \widetilde{X}\beta$ by a qth-order polynomial of x. For a

good approximation to $E(Y) - \widetilde{X}\beta$, the number of clusters k and the order of the polynomial q should be large enough. But large k and q also reduce the power of the test due to the loss in degrees of freedom. Therefore prior investigations on the choices of k and q are necessary before applying the test. Green also mentions that since the true mean $E(Y)$ is not known in practice, the x's in each cluster must be chosen in a narrow interval to retain the power of the test.

B.2 Shillington's Test:

Shillington proposed a test for lack-of-fit of a regression model by extending Fisher's exact F-test on multiple regressions to near-replicates setting. Suppose observations have been grouped into k clusters by any specific clustering procedure. There are $p-1$ covariates. We use the same notation defined in (2.2) and (2.3) but X_i is re-defined to a multivariate version as

$$X_i = \begin{bmatrix} 1 & x_{i11} & x_{i12} & \cdots & x_{i1,p-1} \\ 1 & x_{i21} & x_{i22} & \cdots & x_{i2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{in_i1} & x_{in_i2} & \cdots & x_{in_i,p-1} \end{bmatrix}, \tag{B.8}$$

for $i = 1, \ldots, k$. The tested model is

$$Y = X\beta + \epsilon, \tag{B.9}$$

where β is a $p \times 1$ vector of regression parameters.

When replicates are available, Fisher's proposed test statistic for lack-of-fit of a multiple regression model is

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\mu}_i)^2/(k - r(X))}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(N-k)}, \tag{B.10}$$

where $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{p-1} x_{i,p-1}$, $\hat{\beta}_j$'s are the least sqaures estimates from fitting model (B.9), and $\bar{y}_i$ is the i-th group mean. Shillington classifies the

sum of squares in (B.10) as the sum of squares between clusters and the sum of squares within clusters. The sum of squares in the numerator, measures the variations between the observed cluster mean and the cluster mean predicted by the fitted model (B.9) for each cluster, and is called the sum of squares between clusters. The sum of squares in the denominator, measures the variations between the observed response and the observed mean of its corresponding cluster, and is called the sum of squares within clusters.

When replicate is not available, let $\bar{y}_i$ be the mean response, $\bar{x}_i$ be a $p \times 1$ vector of mean covariates, and $\bar{\epsilon}_i$ be the mean error of the i-th cluster. Define

$$\bar{Y} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_k \end{bmatrix}, \bar{X} = \begin{bmatrix} \bar{x}_1^T \\ \bar{x}_2^T \\ \vdots \\ \bar{x}_k^T \end{bmatrix}, \bar{\epsilon} = \begin{bmatrix} \bar{\epsilon}_1 \\ \bar{\epsilon}_2 \\ \vdots \\ \bar{\epsilon}_k \end{bmatrix}.$$

Let V be a $k \times k$ matrix such that

$$V = \begin{bmatrix} \frac{1}{n_1} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{n_{k-1}} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{n_k} \end{bmatrix}.$$

Shillington considers the model

$$\bar{Y} = \bar{X}\beta + \bar{\epsilon}, \tag{B.11}$$

where $\bar{\epsilon} \sim N(0, \sigma^2 V)$, and computes the sum of squares between clusters by

$$SSE_B = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{x}_i^T \hat{\beta})^2,$$

where $\hat{\beta}$ is the weighted least squares estimate of β in model (B.11).

It is not easy to notice but Shillington's idea is similar to projecting model (B.9) onto $C(Z)$ where Z is defined as in (2.8), i.e.

$$M_Z Y = M_Z X\beta + M_Z \epsilon, \tag{B.12}$$

and uses $SSE(B.12)$ as the sum of squares between clusters. Note that $M_Z\epsilon \sim N(0, \sigma^2 M_Z)$ where M_Z is a singular matrix. As $C(M_Z X) \subset C(M_Z)$, Christensen (2002, Chapter 10) provides a least squares estimate of $M_Z X\beta$ in model (B.12) i.e.

$$M_Z X\hat{\beta} = M_Z X(X^T M_Z X)^- X^T M_Z Y = M_{M_Z X} Y.$$

Hence $SSE(B.12) = [M_Z Y - M_{M_Z X} Y]^T [M_Z Y - M_{M_Z X} Y] = Y^T (M_Z - M_{M_Z X}) Y$.

Similar to the arguments in Fisher's test, we have

$$M_Z Y = \begin{bmatrix} \bar{y}_1 J_{n_1} \\ \bar{y}_2 J_{n_2} \\ \vdots \\ \bar{y}_k J_{n_k} \end{bmatrix} \quad \text{and} \quad M_Z X = \begin{bmatrix} \bar{x}_1^T J_{n_1} \\ \bar{x}_2^T J_{n_2} \\ \vdots \\ \bar{x}_k^T J_{n_k} \end{bmatrix}.$$

These yield

$$\begin{aligned} & [M_Z Y - M_Z X\hat{\beta}]^T [M_Z Y - M_Z X\hat{\beta}] \\ = & \left[(\bar{y}_1 - \bar{x}_1^T\hat{\beta}) J_{n_1}^T, (\bar{y}_2 - \bar{x}_2^T\hat{\beta}) J_{n_2}^T, \cdots, (\bar{y}_k - \bar{x}_k^T\hat{\beta}) J_{n_k}^T \right] \begin{bmatrix} (\bar{y}_1 - \bar{x}_1^T\hat{\beta}) J_{n_1} \\ (\bar{y}_2 - \bar{x}_2^T\hat{\beta}) J_{n_2} \\ \vdots \\ (\bar{y}_k - \bar{x}_k^T\hat{\beta}) J_{n_k} \end{bmatrix} \\ = & \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{x}_i^T\hat{\beta})^2. \end{aligned}$$

Therefore, $SSE_B = [M_Z Y - M_Z X\hat{\beta}]^T [M_Z Y - M_Z X\hat{\beta}]$.

Let A be a $k \times N$ block diagonal matrix such that

$$A = \begin{bmatrix} \frac{1}{n_1} J_1^T & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{n_2} J_2^T & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{n_k} J_{k-1}^T & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{n_k} J_k^T \end{bmatrix}.$$

Model (B.11) is equivalent to

$$AY = AX\beta + A\epsilon.$$

Since V is nonsingular. Following Christensen (2002, Section 2.7) we have

$$AX\hat{\beta} = AX(X^TA^TV^{-1}AX)^-X^TA^TV^{-1}AY.$$

The fact $M_Z = A^TV^{-1}A$ yields

$$\begin{aligned} M_ZX\hat{\beta} &= A^TV^{-1}AX\hat{\beta} \\ &= A^TV^{-1}AX(X^TA^TV^{-1}AX)^-X^TA^TV^{-1}AY \\ &= M_ZX(X^TM_ZX)^-X^TM_ZY \\ &= M_{M_ZX}Y. \end{aligned}$$

Thus $SSE_B = [M_ZY - M_Z X\hat{\beta}]^T[M_ZY - M_Z X\hat{\beta}] = [M_ZY - M_{M_ZX}Y]^T[M_ZY - M_{M_ZX}Y] = SSE(B.12)$.

The sum of squares between clusters is the SSE of the model created by projecting model (B.9) onto $C(Z)$. Shillington uses the SSE of the model created by projecting model (B.9) onto $C(Z)^\perp$, i.e.

$$(I - M_Z)Y = (I - M_Z)X\beta + (I - M_Z)\epsilon, \tag{B.13}$$

as the sum of squares within clusters. The computation of $SSE(B.13)$ is similar to $SSE(B.12)$. Since $(I-M_Z)\epsilon \sim N(0, \sigma^2(I-M_Z))$ and $C((I-M_Z)X) \subset C(I-M_Z)$, a best linear unbiased estimate of $(I - M_Z)X\beta$ in model (B.13) is

$$\begin{aligned} (I - M_Z)X\hat{\beta} &= (I - M_Z)X(X^T(I - M_Z)X)^-X^T(I - M_Z)Y \\ &= M_{(I-M_Z)X}Y. \end{aligned}$$

Hence

$$\begin{aligned} SSE(B.13) &= [(I - M_Z)Y - M_{(I-M_Z)X}Y]^T[(I - M_Z)Y - M_{(I-M_Z)X}Y] \\ &= Y^T(I - M_Z - M_{(I-M_Z)X})Y \\ &\equiv SSE_W. \end{aligned}$$

It is obvious that SSE_B and SSE_W are independent. Under the assumption that model (B.9) is true, the test statistic,

$$F = \frac{Y^T(M_Z - M_{M_ZX})Y/(k - r(M_ZX))}{Y^T(I - M_Z - M_{(I-M_Z)X})Y/(N - k - r((I - M_Z)X))},$$

proposed by Shillington has a central F distribution with degrees of freedom $(k - r(M_Z X), N - k - r((I - M_Z)X))$.

When replicates are available, $C(X) \subset C(Z)$. Then $M_Z X = X$ and $(I - M_Z)X = 0$. Hence the test statistic becomes

$$F = \frac{Y^T(M_Z - M_X)Y/(k - r(X))}{Y^T(I - M_Z)Y/(N - k)}, \tag{B.14}$$

which matches with Fisher's exact F-test.

B.3 Neill and Johnson's Test:

To regulate a near-replicates setting to an exact-replicates setting, Shillington projects the tested model (B.9) onto $C(Z)$. Neill and Johnson deal with this issue in a different direction. Suppose the observations are grouped into k clusters and the covariates are close to each other within each cluster. Neill and Johnson treat the vector of covariates of each observation in the same cluster as a deviation from a known vector. Then the exact-replicates environment can be created as follows.

Let $\mu_i = [\mu_{i0}, \mu_{i1}, \ldots, \mu_{i,p-1}]$ be a $1 \times p$ vector of known constants for $i = 1, \ldots, k$. Define $\mu^T = [\mu_1^T, \mu_2^T, \ldots, \mu_k^T]$ and model (B.9) is equivalent to

$$\begin{aligned} Y &= X\beta - Z\mu\beta + Z\mu\beta + \epsilon \\ &= (X - Z\mu)\beta + Z\mu\beta + \epsilon. \end{aligned} \tag{B.15}$$

Let $Y' = Y - (X - Z\mu)\beta$. Model (B.15) can be written as

$$Y' = Z\mu\beta + \epsilon. \tag{B.16}$$

Assume Y' is observable. Obviously $C(Z\mu) \subset C(Z)$. Following the same idea as discussed at the end of section 2.1, the most general model generated from model (B.16) is

$$Y' = Z\gamma + \epsilon.$$

The exact F-test applies. The test statistic has the same form as (B.14) but Y and X are replaced by Y' and $Z\mu$ respectively, i.e.

$$F' = \frac{Y'^T(M_Z - M_{Z\mu})Y'/(k - r(Z\mu))}{Y'^T(I - M_Z)Y'/(N - k)}. \tag{B.17}$$

The distribution of F' is an F-distribution with degrees of freedom $(k-r(Z\mu), N-k)$.

In practice, Y' can not be observed. Define $\hat{Y}' \equiv Y - (X - Z\mu)\hat{\beta}$ be the estimate of Y' where $\hat{\beta}$ is the least-squares estimate of β from the tested model (B.9). Neill and Johnson suggest that if $Y' - \hat{Y}'$ converges to 0 with probability one as $N \to \infty$, $\hat{Y}'$ can be used to compute the test statistic, i.e.

$$\hat{F}' = \frac{\hat{Y}'^T(M_Z - M_{Z\mu})\hat{Y}'/(k - r(Z\mu))}{\hat{Y}'^T(I - M_Z)\hat{Y}'/(N - k)}, \tag{B.18}$$

where $\hat{F}'$ converges to F' in distribution as $N \to \infty$. Neill and Johnson show that if the constants μ_{i0} are chosen to be 1, and μ_{ij} are chosen to be the cluster mean of the j-th covariate in the i-th cluster for $i = 1, \ldots, k$ and $j = 1, \ldots, p-1$, i.e. $Z\mu = M_Z X$, then $Y' - \hat{Y}'$ converges to 0 with probability one as $N \to \infty$. The choice $Z\mu = M_Z X$ and the fact $M_Z - M_{M_Z X} = (I - M_{M_Z X})M_Z$ yield

$$\begin{aligned}
(M_Z - M_{M_Z X})\hat{Y}' &= (M_Z - M_{M_Z X})(Y - (X - M_Z X)\hat{\beta}) \\
&= (M_Z - M_{M_Z X})Y - (M_Z - M_{M_Z X})(I - M_Z)X\hat{\beta} \\
&= (M_Z - M_{M_Z X})Y - (I - M_{M_Z X})M_Z(I - M_Z)X\hat{\beta} \\
&= (M_Z - M_{M_Z X})Y.
\end{aligned}$$

Hence $\hat{Y}'^T(M_Z - M_{M_Z X})\hat{Y}' = Y^T(M_Z - M_{M_Z X})Y$. In fact, it is not necessary to estimate β to compute the sum of squares for lack-of-fit in Neill and Johnson's test. It is worth noting that Christensen later derives a test from Neill and Johnson's test to testing lack-of-fit for linear regressions along this way.

Similarly, we can find

$$\begin{aligned}
(I - M_Z)\hat{Y}' &= (I - M_Z)(Y - (X - M_Z X)\hat{\beta}) \\
&= (I - M_Z)(Y - X\hat{\beta}).
\end{aligned} \tag{B.19}$$

Note that the difference between Shillington's test, and Neill and Johnson's test is on the construction of the sum of squares for pure error. Shillington projects model (B.9) onto $C(I - M_Z)$ and uses the sum of squares error of the projected model as the sum of squares for pure error. (B.19) explains that Neill and Johnson project the residuals of model (B.9) onto $C(I - M_Z)$ and then compute the sum of squares for pure error by the sum of squares of the projected residuals. Expanding (B.19) gives $(I - M_Z)\hat{Y}' = (I - M_Z)(I - M_X)Y$. The sum of squares for pure error can be written as

$$\hat{Y}'^T(I - M_Z)\hat{Y}' = Y^T(I - M_X)(I - M_Z)(I - M_X)Y.$$

With the use of $Z\mu = M_Z X$, the test statistic is

$$F = \frac{Y^T(M_Z - M_{M_Z X})Y/(k - r(M_Z X))}{Y^T(I - M_X)(I - M_Z)(I - M_X)Y/(N - k)}.$$

Under the assumption that model (B.9), which is equivalent to model (B.16), is true, the denominator in (B.17) is an unbiased estimator of the σ^2. This can be shown by

$$\begin{aligned} E\left(\frac{Y'^T(I - M_Z)Y'}{N - k}\right) &= \sigma^2 + \frac{\beta^T\mu^T Z^T(I - M_Z)Z\mu\beta}{N - k} \\ &= \sigma^2. \end{aligned} \tag{B.20}$$

Since $Y'^T(I - M_Z)Y'/(N - k)$ is a nonnegative random variable. (B.20) implies

$$\frac{Y'^T(I - M_Z)Y'}{N - k} \to \sigma^2 \text{ in probability as } N \to \infty. \tag{B.21}$$

Under the same model assumption, Neill and Johnson further show

$$\frac{\hat{Y}'^T(I - M_Z)\hat{Y}'}{N - k} \to \sigma^2 \text{ in probability as } N \to \infty \tag{B.22}$$

when $Y' - \hat{Y}'$ converges to 0 with probability one as $N \to \infty$. As $F \equiv \hat{F}'$ under the choice of $Z\mu = M_Z X$, (B.21) and (B.22) yield

$$\frac{F'}{F} \to 1 \text{ in probability as } N \to \infty.$$

The asymptotic distribution of F is a central F-distribution with degrees of freedom $(k - r(M_Z X), N - k)$.

B.4 Christensen's Test (1989):

Recall that the tested model, model (B.9), can be written as model (B.16). With the choice of $Z\mu = M_Z X$, model (B.16) becomes

$$Y' = M_Z X\beta + \epsilon, \tag{B.23}$$

where $Y' = Y - (I - M_Z)X\beta$. Model

$$Y' = Z\gamma + \epsilon \tag{B.24}$$

is the most general model that can be generated from model (B.23). Since Y' is unobservable. Neill and Johnson suggest using the estimate $\hat{Y}' = Y - (I - M_Z)X\hat{\beta}$, where $\hat{\beta}$ is a vector of least-squares estimates of the regression parameters of model (B.9), instead of Y' to compute the test statistic. As shown in (B.18), $SSE(B.23)$ can be computed without estimating Y'. Christensen shows that the sum of squares for pure error can be found without knowing Y' by manipulating model (B.24).

Define $Y'' \equiv Y - X\beta$. Model (B.24) becomes

$$Y'' = Z\gamma - M_Z X\beta + \epsilon. \tag{B.25}$$

Since $C(Z, M_Z X) = C(Z)$. Model (B.25) is equivalent to

$$Y'' = Z\theta + \epsilon,$$

where $\theta \equiv (I_k\gamma - (Z^T Z)^- Z^T X\beta)$, and I_k is a $k \times k$ identity matrix. After the above manipulations, model (B.24) is re-written as

$$Y = X\beta + Z\theta + \epsilon. \tag{B.26}$$

Hence the sum of squares for pure error is $SSE(B.26)$. Following Christensen (2002, Chapter 9), $C(X, Z) = C(X, (I - M_X)Z) = C((I - M_Z)X, Z)$ yields

$$SSE(B.26) = Y^T(I - M_X - M_{(I-M_X)Z})Y, \tag{B.27}$$

or

$$SSE(B.26) = Y^T(I - M_Z - M_{(I-M_Z)X})Y.$$

It is worth noting that $SSE(B.26) = SSE(B.13)$, which is the sum of squares within clusters. By (B.27), the sum of squares for lack-of-fit is

$$\begin{aligned} SSE(B.9) - SSE(B.26) &= Y^T(I - M_X)Y - Y^T(I - M_X - M_{(I-M_X)Z})Y \\ &= Y^T(M_{(I-M_X)Z})Y. \end{aligned}$$

The test statistic proposed by Christensen is

$$F = \frac{Y^T(M_{(I-M_X)Z})Y/r((I - M_X)Z)}{Y^T(I - M_Z - M_{(I-M_Z)X})Y/(N - k - r((I - M_Z)X))},$$

which has a central F distribution with degrees of freedom $(r((I - M_X)Z), N - k - r((I - M_Z)X))$.

Recall that the sum of squares for lack-of-fit is the difference between $SSE(B.9)$ and $SSE(B.26)$, which are the sum of squares error of the tested model and the sum of squares within clusters defined in Section B.2 respectively. The residual vector from fitting model (B.9) is $(I - M_X)Y$. The error space is $C(I - M_X)$. Since

$$I - M_X = (I - M_X - M_{(I-M_X)Z}) + M_{(I-M_X)Z}. \tag{B.28}$$

Christensen partitions the error space into two orthogonal spaces $C(I - M_X - M_{(I-M_X)Z}) = C(I - M_Z - M_{(I-M_Z)X})$ and $C(M_{(I-M_X)Z})$, where $C(I - M_X - M_{(I-M_X)Z})$ is the space for the orthogonal lack of fit within clusters, and the remaining $C(M_{(I-M_X)Z})$ is the space for lack-of-fit. If the tested model is rejected only for small F values, Christensen shows that his test is a uniformly most powerful invariant (UMPI) test against the alternative that the orthogonal lack-of-fit lies within clusters.

B.5 Joglekar, Schuenemeyer, and LaRiccia's Test:

As mentioned in Section B.2, Shillington chooses the sum of squares within clusters, $SSE(B.13)$, as the sum of squares for pure error. When testing model (B.9) against the alternative that the lack-of-fit lies between clusters, Joglekar (1985) argues that $MSE(B.13)$ as an estimate of σ^2 has a large bias and hence the test statistic proposed by Shillington behaves unpredictably. We use the notation defined in and (2.3) but X is an $N \times p$ matrix of covariates. Let

$$A = \begin{bmatrix} X_1 & 0 & 0 & 0 \\ 0 & X_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & X_k \end{bmatrix}.$$

Joglekar, Schuenemeyer, and LaRiccia modify Shillington's test by replacing $MSE(B.13)$ with the MSE of model

$$Y = A\tau + \epsilon, \tag{B.29}$$

where τ is an $(N \times p) \times 1$ vector of regression parameters.

Referring to the discussion in Section B.4, $SSE(B.13)$ is the sum of squares error of model (B.26), i.e. $Y = X\beta + Z\theta + \epsilon$. It is obvious that $C(X, Z) \subset C(A)$. This implies model (B.29) is more general than model (B.26), and $(I - M_A)(M_Z - M_{M_ZX}) = 0$. Hence the sum of squares for pure error $SSE(B.29) = Y^T(I - M_A)Y$ and the sum of squares for lack-of-fit $SSE(B.26) = Y^T(M_Z - M_{M_ZX})Y$ are independent. The exact F-test applies and the test statistic is

$$F = \frac{Y^T(M_Z - M_{M_ZX})Y/(k - r(M_ZX))}{Y^T(I - M_A)Y/(N - r(A))}.$$

F has a central F distribution with degrees of freedom $(k - r(M_ZX), N - r(A))$.

Note that Joglekar, Schuenemeyer, and LaRiccia's test shows the same problem as in Green's test. Large k and p would reduce the power of the test due to the loss in degrees of freedom.

B.6 Christensen's Test (1991):

When testing model (B.9) against the alternative that the lack-of-fit lies between clusters, Joglekar finds that the mean squares for pure error, $MSE(B.13)$, used in Shillington's test does not perform well. Joglekar, Schuenemeyer, and LaRiccia modified Shillington's test along this way and proposed a test. Christensen also proposed a test to testing the lack-of-fit lies between clusters by modifying Shillington's test with a different choice in the sum of squares for pure error.

Shillington projects the tested model onto $C(Z)$ and use the sum of squares error of the projected model as the sum of squares for lack-of-fit; projects the tested model onto $C(Z)^{\perp}$ and use the sum of squares error of the projected model as the sum of squares for pure error. Christensen states that the error space of the tested model is not complete by these projections. Recall that the residual vector from fitting model (B.9) is $(I - M_X)Y$ and the error space is $C(I - M_X)$. Similar to (B.28),

$$\begin{aligned} I - M_X &= (I - M_Z - M_{(I-M_Z)X}) + (M_Z - M_{M_ZX}) \\ &\quad +(M_Z + M_{(I-M_Z)X} - M_X - M_Z - M_{M_ZX}) \\ &= (I - M_Z - M_{(I-M_Z)X}) + (M_Z - M_{M_ZX}) \\ &\quad +(M_{(I-M_Z)X} + M_{M_ZX} - M_X). \end{aligned}$$

Obviously $C(X) \subseteq C(M_ZX, (I - M_Z)X)$. Let $\rho \in C(M_ZX, (I - M_Z)X)$. ρ can be written as $M_ZXa + (I - M_Z)Xb$ for any $p \times 1$ vectors a, b. Then

$$\rho = M_ZXa + (I - M_Z)Xb = Xb + M_ZX(a - b) \in C(X, M_ZX).$$

Unless $C(X) \subseteq C(Z) \cup C(Z)^{\perp}$, $M_{(I-M_Z)X} + M_{M_ZX} - M_X$ is not 0. With the fact $M_Z - M_{(I-M_Z)X} = M_X - M_{(I-M_X)Z}$, it can be observed that

$$(I - M_Z - M_{(I-M_Z)X})(M_Z - M_{M_ZX}) = 0,$$

and

$$(I - M_Z - M_{(I-M_Z)X})(M_{(I-M_Z)X} + M_{M_ZX} - M_X) = 0.$$

We further show

$$\begin{aligned}(M_Z - M_{M_ZX})(M_{(I-M_Z)X} + M_{M_ZX} - M_X) &= -(M_Z - M_{M_ZX})M_X \\ &= -(I - M_{M_ZX})M_ZX(X^TX)^-X^T \\ &= 0.\end{aligned}$$

Hence the error space can be partitioned into three orthogonal spaces, i.e. $C(I - M_Z - M_{(I-M_Z)X})$, $C(M_Z - M_{M_ZX})$, and $C(M_{(I-M_Z)X} + M_{M_ZX} - M_X)$, where they are the spaces for the orthogonal lack of fit within clusters, the orthogonal lack of fit between clusters, and the remaining of the error space respectively.

The partitioning of the error space indicates that Shillington's use of $Y^T(M_Z - M_{M_ZX})Y$ and $Y^T(I - M_Z - M_{(I-M_Z)X})Y$ in the test statistic cannot cover the entire error space of the tested model. When the sum of squares for lack-of-fit is chosen to be the orthogonal lack of fit between clusters $Y^T(M_Z - M_{M_ZX})Y$, a natural choice in the sum of squares for pure error is the sum of squares from the other two spaces, i.e.

$$\begin{aligned}& Y^T(I - M_Z - M_{(I-M_Z)X} + M_{(I-M_Z)X} + M_{M_ZX} - M_X)Y \\ = \;& Y^T(I - M_X)Y - Y^T(M_Z - M_{M_ZX})Y \\ = \;& SSE(B.9) - SSE(B.12).\end{aligned}$$

The test statistic proposed by Christensen is

$$F = \frac{Y^T(M_Z - M_{M_ZX})Y/(k - r(M_ZX))}{Y^T(I - M_X - M_Z + M_{M_ZX})Y/(N - r(X) - k + r(M_ZX))},$$

where F has a central F distribution with degrees of freedom $(k - r(M_ZX), N - r(X) - (k - r(M_ZX)))$. Christensen shows that this test is UMPI for the alternative that the orthogonal lack-of-fit lies between clusters.

B.7 Su and Yang's Test:

Su and Yang proposed three lack-of-fit tests in multiple regression. These tests are classified as the overall lack-of-fit test, the between clusters lack-of-fit test, and

the within clusters lack-of-fit test respectively. Su and Yang discuss the problems in the usual setting of multiple regressions, i.e. model (B.9) is the tested model. But their tests would better be treated as the generalizations of Green's test. Therefore the tested model used in the following discussion is the multivariate version of model (B.4). Use $\widetilde{X}_i$ defined in (B.8) and let

$$f(\widetilde{X}_i) = \begin{bmatrix} f_1(x_{i11}) & f_2(x_{i12}) & \cdots & f_{p-1}(x_{i1,p-1}) \\ f_1(x_{i21}) & f_2(x_{i22}) & \cdots & f_{p-1}(x_{i2,p-1}) \\ \vdots & \vdots & \vdots & \vdots \\ f_1(x_{in_i1}) & f_2(x_{in_i2}) & \cdots & f_{p-1}(x_{in_i,p-1}) \end{bmatrix},$$

for $i = 1, \ldots, k$, where $f_j(x)$ is a function of x such that $f_j : \mathbb{R} \to \mathbb{R}$ for $j = 1, \ldots, p-1$. Define $\widehat{X}_i \equiv [J_{n_i}, f(\widetilde{X}_i)]$ and re-define $\widehat{X}^T \equiv [\widehat{X}_1^T, \widehat{X}_2^T, \ldots, \widehat{X}_k^T]$. The tested model is

$$Y = \widehat{X}\beta + \epsilon. \tag{B.30}$$

Let $P_j(x)\gamma_j$ be a q_j-th order polynomial of x for $j = 1, \ldots, p-1$, and

$$P_i = \begin{bmatrix} P_1(x_{i11})\gamma_{i1} & P_2(x_{i12})\gamma_{i2} & \cdots & P_{p-1}(x_{i1,p-1})\gamma_{i,p-1} \\ P_1(x_{i21})\gamma_{i1} & P_2(x_{i22})\gamma_{i2} & \cdots & P_{p-1}(x_{i2,p-1})\gamma_{i,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ P_1(x_{in_i1})\gamma_{i1} & P_2(x_{in_i2})\gamma_{i2} & \cdots & P_{p-1}(x_{in_i,p-1})\gamma_{i,p-1} \end{bmatrix}$$

for $i = 1, \ldots, k$. We further define a matrix P as in (B.3). Following Green's arguments, a more general model can be generated from model (B.30) as

$$Y = \widehat{X}\beta + P\gamma + \epsilon. \tag{B.31}$$

Since $C(Z) \subseteq C(P)$. It is worth noting that if model (B.9) is the test model, i.e. $\widehat{X} = X$, model (B.31) can be simplified as follows. When all q_j's are 0, model (B.31) is equivalent to $Y = X\beta + Z\gamma' + \epsilon$; when all $q_j \geq 1$, model (B.31) is equivalent to $Y = P\gamma' + \epsilon$. Let $X' = (I - M_P)\widehat{X}$. Following Christensen (2002, Chapter 9), model (B.31) is re-written as

$$\begin{aligned} Y &= (I - M_P)\widehat{X}\beta + P\gamma' + \epsilon \\ &= X'\beta + P\gamma' + \epsilon. \end{aligned}$$

Denote $M_0 = M_{\widehat{X}}$. The exact F test applied and the test statistic is

$$F_0 = \frac{Y^T(M_P + M_{X'} - M_0)Y/(r(P) + r(X') - r(\widehat{X}))}{Y^T(I - M_P - M_{X'})Y/(N - r(P) - r(X'))}.$$

This is an extension of Green's test and Su and Yang call this test the overall lack-of-fit test. F_0 has a central F distribution with degrees of freedom $(r(P) + r(X') - r(\widehat{X}), N - r(P) - r(X'))$.

The sum of squares for lack-of-fit in F_0 is $SSE(B.31) = Y^T(I - M_P - M_{X'})Y$. We can interpret $SSE(B.31)$ with the idea of model projection. If the tested model (B.30) is projected onto $C(I - M_P)$, the projected model is

$$(I - M_P)Y = (I - M_P)\widehat{X}\beta + (I - M_P)\epsilon, \tag{B.32}$$

where $(I - M_P)\epsilon \sim N(0, \sigma^2(I - M_P))$. $C((I - M_P)\widehat{X}) \subset C(I - M_P)$, Christensen (2002, Chapter 10) provides the best linear unbiased estimate of $(I - M_P)\widehat{X}\beta$ as

$$\begin{aligned}(I - M_P)\widehat{X}\hat{\beta} &= (I - M_P)\widehat{X}(\widehat{X}^T(I - M_P)\widehat{X})^-\widehat{X}^T(I - M_P)Y \\ &= M_{(I-M_P)\widehat{X}}Y \\ &= M_{X'}Y\end{aligned}$$

The sum of squares error of model (B.32) is

$$SSE(B.32) = Y^T(I - M_P - M_{X'})Y = SSE(B.31).$$

Su and Yang's overall lack-of-fit test and Christensen's (1989) test have the same moral but Su and Yang extend Z to P.

Following the same idea, Shillington's test can generalized if the sum of squares for lack-of-fit is replaced by the sum of squares errors of the model

$$M_P Y = M_P\widetilde{X} + M_P\epsilon,$$

where $(M_P)\epsilon \sim N(0, \sigma^2(M_P))$. Christensen's (1991) test can also be generalized along this way.

Unlike Christensen's idea on partitioning the error space of the tested model, Su and Yang only split the lack-of-fit space regarding to the overall lack-of-fit test, $C(M_P + M_{X'} - M_0) = C(M_{(I-M_0)P})$, into two. Since

$$M_{(I-M_0)P} = (M_{(I-M_0)P} - M_{(I-M_0)Z}) + M_{(I-M_0)Z}.$$

$C((I-M_0)Z) \subset C((I-M_0)P)$ gives $M_{(I-M_0)P} - M_{(I-M_0)Z}$ and $M_{(I-M_0)Z}$ are two ppo's onto two orthogonal spaces, $C(M_{(I-M_0)P} - M_{(I-M_0)Z})$ and $C(M_{(I-M_0)Z})$, respectively. The sum of squares $Y^T(M_{(I-M_0)Z})Y$ measures the lack-of-fit contributed by the common intercept. Su and Yang use $Y^T(M_{(I-M_0)Z})Y$ and propose the second test statistic

$$F_1 = \frac{Y^T(M_{(I-M_0)Z})Y/(r(M_{(I-M_0)Z}))}{Y^T(I - M_P - M_{X'})Y/(N - r(P) - r(X'))}.$$

Su and Yang call this test between clusters lack-of-fit test, but this "between clusters lack-of-fit" is different from that proposed by Christensen in moral. F_1 has a central F distribution with degrees of freedom $(r(M_{(I-M_0)Z}), N - r(P) - r(X'))$.

The sum of squares $Y^T(M_{(I-M_0)P} - M_{(I-M_0)Z})Y$ measures the lack-of-fit contributed by the common regression parameters for covariates in the tested model. The third test statistic proposed by Su and Yang is

$$F_2 = \frac{Y^T(M_{(I-M_0)P} - M_{(I-M_0)Z})Y/(r(M_{(I-M_0)P}) - r(M_{(I-M_0)Z}))}{Y^T(I - M_P - M_{X'})Y/(N - r(P) - r(X'))}.$$

Su and Yang name this test within clusters lack-of-fit test. Once again, this "within clusters lack-of-fit" is also different with that Christensen proposed. The test statistic F_2 has a central F distribution with degrees of freedom $(r(M_{(I-M_0)P})$ $-r(M_{(I-M_0)Z}), N - r(P) - r(X'))$.

Appendix C: Review of Smooth Tests

C.1 Eubank and Hart's Test:

Eubank and Hart investigate the lack-of-fit of univariate regression models. We use the notation defined in (2.11) and re-define

$$X = \begin{bmatrix} f_0(x_1) & f_1(x_1) & \cdots & f_{p-1}(x_1) \\ f_0(x_2) & f_1(x_2) & \cdots & f_{p-1}(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ f_0(x_n) & f_1(x_n) & \cdots & f_{p-1}(x_n) \end{bmatrix}, \tag{C.1}$$

where $\forall\ x_i \in [0,1]$ for $i = 1, \ldots, n$ and $f_j(x)$ are functions in x for $j = 0, 1, \ldots, p-1$. The tested model is

$$Y = X\beta + \epsilon, \tag{C.2}$$

where β is a $p \times 1$ vector of regression parameters. With the newly defined X, model (2.13) is an arbitrary alternative model. Note that $C(X) \subset C(X, H)$ and hence model (2.13) is more general then the tested model. As mentioned in Section (2.2), model (2.17) is a series approximation to model (2.13) that we can actually handle. Eubank and Hart proposed a test based on the alternative model

$$Y = X\beta + H_k\gamma_k + \epsilon, \tag{C.3}$$

where H_k is defined in (2.16).

Eubank and Hart assume that X is of full rank, i.e. $r(X) = p$, and the functions in the matrix H_k satisfy the orthogonal conditions

$$H_k^T H_k = nI_k \qquad \text{and} \qquad H_k^T X = 0.$$

Note that $\frac{1}{\sqrt{n}} H_k$ is an orthonormal matrix. Christensen (2002, Appendix B) gives the ppo onto $C(H_k)$ as $\frac{1}{n} H_k H_k^T$. From the above assumptions, $C(X, H_k)$ is of full rank. The ppo onto $C(X, H_k)$ is $M_X + M_{H_k} = M_X + \frac{H_k H_k^T}{n}$. The sum of squares error of model (C.3) is

$$\begin{aligned} SSE(C.3) &= Y^T \left(I - M_X - \frac{H_k H_k^T}{n} \right) Y \\ &= Y^T (I - M_X) Y - \frac{Y^T H_k H_k^T Y}{n} \\ &\equiv SSE(C.2) - n\hat{\gamma}_k^T \hat{\gamma}_k, \end{aligned}$$

where $\hat{\gamma}_k \equiv \frac{1}{n} H_k^T Y$.

A risk function is defined as

$$R(k) = E\left[\frac{1}{n} \sum_{j=1}^{n} (h(x_j) - \hat{h}_k(x_j))^2 \right],$$

where $\hat{h}_k(x_j)$ is the estimate of $h_k(x_j)$, and $h(x_j)$ and $h_k(x_j)$ are defined in (2.14) and (2.15) respectively. Recall that the true function $h(x)$ is approximated by the partial sum $h_k(x)$. Obviously, the risk function measures the errors from the partial sum approximation. Eubank and Hart derive a lack-of-fit test statistic by minimizing the risk function.

If σ^2 is known, Rice (1984) shows that the unbiased estimate of the risk function is

$$\begin{aligned} \hat{R}(k) &= \frac{1}{n} SSE(C.3) - \sigma^2 + \frac{2\sigma^2 r(X, H_k)}{n} \\ &= \frac{Y^T (I - M_X) Y}{n} - \hat{\gamma}_k^T \hat{\gamma}_k - \sigma^2 + \frac{2(p+k)\sigma^2}{n} \\ &= \frac{Y^T (I - M_X) Y}{n} - \frac{(n-2p)\sigma^2}{n} - \left[\hat{\gamma}_k^T \hat{\gamma}_k - \frac{2k\sigma^2}{n} \right]. \end{aligned} \tag{C.4}$$

The estimated risk function $\hat{R}$ is minimized by maximizing

$$\hat{\gamma}_k^T \hat{\gamma}_k - \frac{2k\sigma^2}{n}. \tag{C.5}$$

Eubank and Hart suggest that the maximizer $\tilde{k}$ of (C.5) can be used to test the lack-of-fit in model (C.2). Since σ^2 cannot be known in practice and the asymptotic distribution of $\tilde{k}$ does not give an explicit rejection region for any specific size α. The test statistic proposed by Eubank and Hart is $\hat{k}$ where $\hat{k}$ is the maximizer of

$$g(k) = \begin{cases} 0, & k = 0, \\ \tilde{\gamma}_k^T \tilde{\gamma}_k - c_\alpha k \hat{\sigma}^2/n, & k = 1, \ldots, n-p, \end{cases} \tag{C.6}$$

in which $\hat{\sigma}^2$ is any consistent estimator of σ^2 and c_α is chosen so that $P(\hat{k} = 0) = 1 - \alpha$ under the tested model. Model (C.2) is rejected if $\hat{k} \geq 1$.

Define Z_j be a random variable having χ^2 distribution with degrees of freedom j. Eubank and Hart provide an approximation to c_α in (C.6) by solving the equation

$$1 - \alpha = \exp\left\{ -\sum_{j=1}^{\infty} \frac{P(Z_j > j c_\alpha)}{j} \right\}. \tag{C.7}$$

Let $p_0 = 1$ and $p_s = \sum_{(\theta_1, \ldots, \theta_s) \in C_s} \left\{ \prod_{j=1}^{s} \frac{1}{\theta_r!} \left[\frac{P(Z_j > j c_\alpha}{j} \right]^{\theta_r} \right\}$ for $s = 1, \ldots, n-p$, where C_s is the set of all s-tuples $(\theta_1, \ldots, \theta_s)$ of integers such that $\theta_1 + 2\theta_2 + \cdots + s\theta_s = s$. Eubank and Hart also prove the asymptotic distribution of the test statistic $\hat{k}$. Under the assumptions that $\hat{\sigma}^2 \to \sigma^2$ in probability, c_α is the solution of (C.7), and $\max_{1 \leq j \leq n-p} \sup_x |\varphi_j(x)| \leq C$ for some constant C that is independent of n, then

$$P(\hat{k} = s) \to p_s(1 - \alpha) \quad \text{as } n \to \infty \quad \text{for} \quad s = 0, 1, \ldots.$$

C.2 Aerts, Claeskens, and Hart's Test:

Use the notation defined in (2.11) and X as defined in (C.1) without the restriction on the range of x_i's. Aerts, Claeskens, and Hart, henceforth referred to as ACH, consider lack-of-fit in model (C.2). Model (C.3) is referred to as an approximation

to the more general model of model (C.2). Eubank and Hart use k that minimizes the risk function as a test statistic for lack-of-fit in model (C.2). ACH build up test statistics by the score function.

ACH assume X is of full rank, and the functions in X and H_k satisfy the orthogonal condition

$$H_k^T H_k = nI_k \qquad \text{and} \qquad H_k^T X = 0.$$

Define $\hat{\gamma}_k = \frac{1}{n} H_k^T Y$. The score statistic is given by

$$\begin{aligned} S_k &= \frac{SSE(C.2) - SSE(C.3)}{SSE(C.2)} \\ &= \frac{Y^T M_{H_k} Y}{Y^T (I - M_X) Y} \\ &= \frac{Y^T M_{H_k} Y/n}{Y^T (I - M_X) Y/n} \\ &= \frac{\hat{\gamma}_k^T \hat{\gamma}_k}{\hat{\sigma}^2}, \end{aligned}$$

where $\hat{\sigma}^2 = \dfrac{Y^T(I - M_X)Y}{n}$ is the maximum likelihood estimate of σ^2 under model (C.2). ACH introduce the penalized score criterion,

$$SIC(k, C_n) = S_k - C_n k, \tag{C.8}$$

where S_0 is defined to be 0 and C_n is a constant greater than 1. $SIC(k, C_n)$ is used to choose $\hat{k}$ where $\hat{k}$ maximizes $SIC(k, C_n)$. Note that $SIC(k, 2)$ and $SIC(k, \log n)$ are the AIC and BIC used in model selection in regressions respectively.

Let Z_j be iid χ^2 random variables with degrees of freedom 1 for $j = 1, \ldots, n-p$. Define $V_0 = 0$ and $V_k = \sum_{j=1}^{k} Z_j$ for $k = 1, \ldots, n-p$, $\tilde{k}$ be the maximizer of $V_k - 2k$, $\hat{k}_1$ be the maximizer of $SIC(k, 2)$, and $\hat{k}_2$ be the maximizer of $SIC(k, \log n)$. ACH proposed five test statistics. Each test statistic has a specific asymptotic distribution. The test statistics are:

$$T_1 = S_{\hat{k}_1}, \qquad T_2 = S_{\hat{k}_2}, \qquad T_3 = \frac{S_{\hat{k}_1} - \hat{k}_1}{\max(1, \hat{k}_1^{1/2})},$$

$$T_4 = \max_{1 \le k \le n-p} \frac{S_k}{k}, \qquad \text{and} \qquad T_5 = SIC(\hat{k}_1, 2).$$

ACH show that, under the assumption that model (C.2) is true,

$$T_1 \to V_{\tilde{k}}, \qquad T_2 \to V_1, \qquad T_3 \to \frac{V_{\tilde{k}} - \tilde{k}}{\max(1, \tilde{k}^{1/2})},$$
$$T_4 \to \max_{1 \le k \le n-p} \frac{V_k}{k}, \qquad \text{and} \qquad T_5 \to V_{\tilde{k}} - 2\tilde{k},$$

in distribution as $n \to \infty$. Note that the test statistic T_2 has an asymptotic χ^2-distribution with one degree of freedom. There is no tractable distribution for any other test statistics.

If σ^2 is known, (C.8) can be written as

$$\begin{aligned} SIC(k, C_n) &= \frac{\hat{\gamma}_k^T \hat{\gamma}_k}{\sigma^2} - C_n k \\ &= \frac{1}{\sigma^2} \left(\hat{\gamma}_k^T \hat{\gamma}_k - C_n k \sigma^2 \right). \end{aligned}$$

It is worth noting that maximizing $SIC(k, C_n)$ is equivalent to maximizing $\hat{\gamma}_k^T \hat{\gamma}_k - C_n k \sigma^2$. Then the maximizer of $SIC(k, 2/n)$ is identical to the maximizer of the risk function in (C.4). ACH's test is consistent to Eubank and Hart's testing procedures when σ^2 is known. Since the MLE $\hat{\sigma}^2$ is a consistent estimate of σ^2, the result in ACH's test is also consistent to Eubank and Hart's test when $\hat{\sigma}^2$ is used.

Bibliography

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

[2] Atkinson, A., and Riani, M. (2000). *Robust Diagnostic Regression Analysis*, Springer-Verlag, New York.

[3] Bruce, D. and Schumacher, F. X. (1935). *Forest Mensuration*, McGraw-Hill, New York.

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

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

[6] Christensen, R. (2001). *Advanced Linear Modeling: Multivariate, Time Series, ans Spatial Data; Nonparametric Regression, and Response Surface Maximization*, Second Edition. Springer-Verlag, New York.

[7] Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer-Verlag, New York.

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

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

[10] Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer-Verlag, New York.

[11] 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.

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

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

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

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

[16] 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.

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

[18] Neyman, J. (1937). "Smooth Test for Goodness of Fit," *Skandinavisk Aktuarietidskrift*, 20, 149-199.

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

[20] Rice, J. (1984). "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics*, 12, 1215-1230.

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

[22] 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.

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