A SIMULATION COMPARISON OF CLUSTER BASED LACK OF FIT TESTS

by

ZHIWEI SUN

B.S., South China Agricultural University, 1996
M.S., South China Agricultural University, 1999

A REPORT

Submitted in partial fulfillment of the requirements for the degree

Master of Science

Department of Statistics
College of Arts and Science

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2007

Approved by:

Major Professor
James W. Neill

Abstract

Cluster based lack of fit tests for linear regression models are generally effective in detecting model inadequacy due to between- or within-cluster lack of fit. Typically, lack of fit exists as a combination of these two pure types, and can be extremely difficult to detect depending on the nature of the mixture. Su and Yang (2006) and Miller and Neill (2007) have proposed lack of fit tests which address this problem. Based on a simulation comparison of the two testing procedures, it is concluded that the Su and Yang test can be expected to be effective when the true model is locally well approximated within each specified cluster and the lack of fit is not due to an unspecified predictor variable. The Miller and Neill test accommodates a broader alternative, which can thus result in comparatively lower but effective power. However, the latter test demonstrated the ability to detect model inadequacy when the lack of fit was a function of an unspecified predictor variable and does not require a specified clustering for implementation.

Table of Contents

List of Figures

List of Tables

Acknowledgements

There are a number of people I would like to thank for their help in carrying out the work described in this report. I am especially grateful to Dr. James W. Neill, my major professor, for proposing the report topic and for his guidance, patience and support that he has given me. This report would not have been completed without his assistance.

Special thanks to my committee members, Dr. Haiyan wang and Dr. Weixing Song, for their kind help. Also, I wish to express my appreciation to all of the faculty in the Department of Statistics.

Finally, I would like to especially thank my wife Yali for her continuous support and encouragement throughout my studies.

CHAPTER 1 - Introduction

In parametric linear regression models, lack of fit is said to exist when the proposed model has an incorrectly specified mean structure. The traditional regression lack of fit test for models with replication was given by Fisher (1922). This test is appropriate for detecting model inadequacy which generalizes the mean structure of the proposed model. In particular, if the true model is embedded in a general alternative given by the one-way ANOVA model then the traditional test is useful for detecting such lack of fit. Such lack of fit is referred to as between cluster lack of fit, and for example, may involve the need for higher order polynomial terms. For the common circumstance in which replicate measurements are not obtained, there is a set of methods for testing lack of fit which is based on mimicking the traditional lack of fit. With these methods, rows of the design matrix that are nearly replicates are identified in order to construct alternative full models. Tests for lack of fit using near replicates include the work by Green (1971), Breiman and Meisel (1976), Atwood and Ryan (1977), Lyons and Proctor (1977), Shillington (1979), Daniel and Wood (1980), Utts (1982), Neill and Johnson (1985), Joglekar, Schuenemeyer, and LaRiccia (1989) and Christensen (1989, 1991).

Christensen’s approach is of particular interest since the lack of fit space was characterized as a sum of orthogonal subspaces with corresponding optimal tests. In particular, Christensen characterized lack of fit as existing between clusters of near replicates, within clusters, or as a combination of these two pure types. However, Christensen’s tests are uniformly most powerful invariant only for the specified grouping of the data into near replicates. Indeed, Christensen (1991) noted that an important fundamental problem in nonreplicated lack of fit testing was the lack of an optimal strategy for grouping observations into efficacious clusters. Miller, Neill and Sherfey (1998, 1999) then presented a statistically principled framework within which to study the selection of near replicates for use with Christensen’s tests. The methodology is based on a maximin power criterion that incorporates nearness considerations to cluster observations into groups of near replicates. Their work constructs a single clustering for the nonreplicated case that is optimal for detecting between-cluster lack of fit.

Christensen (2003) has noted that significantly small values of the traditional F-statistic may indicate the presence of lack of fit due to features that are in fact not part of the proposed model. Such lack of fit is referred to as within-cluster lack of fit, and for example, may involve a trend in time within each group of replicates whenever the replicates are observed in a time sequence. Typically, lack of fit exists as a combination of these two pure types and can be extremely difficult to detect depending on the nature of the mixture. Christensen (2003) also showed that the traditional F-statistic can become large or small because the assumed covariance structure is incorrect, even when the mean structure of the proposed model is correct.

As noted by Christensen (1989, 1991, 2002), the suggested lack of fit tests based on near replicates may also be unable to detect mixtures of the two pure types. Since these tests reduce to the traditional test when the clusters consist of exact replicates, such performance is not unexpected. Recently, Su and Yang (2006) and Miller and Neill (2007) have suggested lack of fit tests for the case of nonreplication that may be useful for detecting all of the above types of model inadequacies, including mixtures. These tests assume normal errors and that the covariance structure is not misspecified. Cluster-based tests for assessing independence and variance function specification are given by Christensen and Bedrick (1997) and Bedrick (2000).

Su and Yang (2006) assume that clusters of near replicate observations have been identified. Given such a clustering, the authors construct a full model that contains the proposed model, and which is intended to be able to approximate the true model locally in each cluster. In particular, the constructed full model depends on functions of all of the independent variables of the proposed model, which for example, may include powers and cross-products of the specified predictive variables. This test thus contains Christensen’s (1989) test and the test proposed by Atwood and Ryan (1977) based on the partition method, which is also discussed by Christensen (2002). This test may work well whenever any model inadequacies are not due to unspecified predictor variables. Also, when exact replicates are available, this test reduces to the traditional test, which has been shown not to be effective in detecting mixtures of between- and within-cluster lack of fit.

Miller and Neill (2007) developed a lack of fit testing procedure based on families of groupings of the observations. For models with replication, the possible groupings are inherently determined by the row structure of the design matrix. The use of groupings embeds the one-way ANOVA model in more general models which provides tests which can be effective in detecting mixtures of the two pure types of lack of fit. Since the efficacy of a particular choice of grouping is a function of the unobservable lack of fit, several such tests are considered, each based on a different grouping of the observations, and the multiple testing approach of Baraud et al. (2003) is followed. More generally, the preceding testing procedure based on families of groupings was extended to the case of nonreplication. For this case, the authors proposed that such families be determined by linear orders on the predictor variables based on disjoint parallel tubes in predictor space. Test statistics follow the clustering-based lack of fit tests given by Christensen (1989, 1991), by considering the groupings as determining special types of clusterings. In order to detect general lack of fit, several such tests are again considered, each based on a different grouping of the observations, and the multiple testing approach given by Baraud et al. (2003) is followed.

For completeness, it is remarked that several other approaches for assessing the existence of lack of fit have been proposed. In particular, the use of nonparametric regression techniques to test the adequacy of parametric regression models is discussed by Hart (1997), Aerts et al. (2000), Eubank et al. (2005) and the references therein. Also, the use of partial sum residual empirical processes has been suggested by Khmaladze and Koul (2004) and the references therein to assess goodness of fit in regression models. Finally, graphical methods used for checking model adequacy were given by Cook and Weisburg (1997), for example.

This report reviews the tests of Su and Yang (2006) and Miller and Neill (2007) in Chapter 2. Chapter 3 presents some simulation results which compare these two lack of fit testing procedures, the results of which are given in Chapter 4.

Some notation that is used in the report is now introduced. Let $\mathcal{V} \subset \mathcal{R}^n$ be a vector space and let $\mathcal{U} \subset \mathcal{V}$ be a subspace. We denote by $\mathcal{U}_{\mathcal{V}}^{\perp}$ the orthogonal complement of $\mathcal{U}$ with respect to $\mathcal{V}$. If $\mathcal{V} = \mathcal{R}^n$ then we simply write $\mathcal{U}^{\perp}$. Let $P_{\mathcal{V}}$ denote the orthogonal projection operator onto $\mathcal{V}$, and let $\dim \mathcal{V}$ represent the dimension of $\mathcal{V}$. For $\upsilon \in \mathcal{R}^n$, $\| \upsilon \|^2$ is the squared Euclidean length $\sum_{i=1}^{n} \upsilon_i^2$ of υ. For a matrix A, A' denotes the transpose of A and $C(A)$ is the linear subspace of $\mathcal{R}^n$ generated by the columns of A. The orthogonal projection matrix for projecting onto $C(A)$ can be computed as $A(A'A)^{-}A'$ where $(A'A)^{-}$ represents any generalized inverse of $A'A$. Let $F_{r,s}$ denote the central F distribution with r numerator and s denominator degrees of freedom. We write $F_{r,s}(\alpha)$ for the 100αth percentile of a $F_{r,s}$ distribution and let $\bar{F}_{r,s}(u) = \Pr(F > u)$ where $F \sim F_{r,s}$, so that $\bar{F}_{r,s}^{-1}(\alpha) = F_{r,s}(1-\alpha)$.

In order to specify a grouping of n observations into c groups, we use an $n \times c$ matrix Z which contains indicator variables for the groups. That is, Z contains only zeros and ones, and the nonzero values in the jth column of Z correspond to the observations in the jth group, $j = 1, ..., c$. A grouping determined by such a Z matrix will also be called a clustering or partition of the observations. For example, with $n = 11$ and $c = 3$, to indicate that observations 1, 2, 3 and 4 belong to group 1, that observation 5, 6 and 7 belong to group 2 and that the observations 8, 9, 10 and 11 belong to group 3, Z has the form

$$Z' = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

CHAPTER 2 - Lack of Fit Tests

A common approach to testing a proposed linear regression model $Y = X\beta + e$ for lack of fit involves testing the model against a constructed full model $Y = X_*\beta + e$ with $C(X) \subset C(X_*)$. Here, X is a nonrandom $n \times p$ matrix of predictor variables, $\beta \in \mathcal{R}^p$ is an unknown parameter vector and $e \sim N(0, \sigma^2 I_n)$ is an $n-$dimensional random error vector with unknown $\sigma^2 > 0$. As noted by Christensen (2002), there are few theoretical guidelines for choosing X_*, and hence the constructed full model. With no other variables available, the choice of X_* is necessarily based on those variables represented in the proposed model matrix X. The challenge is to define X_* so that the constructed full model provides not only a sufficiently general alternative which includes the true data generating model, but also leads to a test with sufficient power to detect lack of fit i.e. when $E(Y) \neq X\beta$.

For examples of specific models representing the general types of lack of fit discussed in Section 1, consider testing the adequacy of a simple linear regression model with replication given by

$$y_{ij} = \beta_0 + \beta_1 x_i + e_{ij} \tag{1}$$

for $i = 1, \cdots, c$ and $j = 1, \cdots, N$, when the underlying true model has the form

$$y_{ij} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \gamma \cos(\omega t_j) + \delta t_j + e_{ij} \tag{2}$$

where $t_1 < t_2 < \ldots < t_N$. For different choices of the parameters, model (2) (except for the cosine term) was utilized by Christensen (2003) to exemplify various types of model inadequacy. In particular, with $\beta_2 \neq 0$ and $\gamma = \delta = 0$ lack of fit exists in the between-cluster subspace $B = C(X)^\perp \cap C(Z)$ and represents traditional polynomial lack of fit between the groups of replicates. If $\beta_2 = 0$ and either $\gamma \neq 0$ or $\delta \neq 0$ then lack of fit exists in the within-cluster subspace W= $C(X)^\perp \cap C(Z)^\perp$ and represents a trend and/or periodic lack of fit within the groups of replicates. Finally, if $\beta_2 \neq 0$ and either $\gamma \neq 0$ or $\delta \neq 0$ then lack of fit exists in both subspaces B and W as a mixture of the two pure types. In the preceding, there are replicated rows of X and Z is a matrix of indicator variables having the same row structure as X (i.e. Z contains only zeros and ones, and the nonzero values in a column of Z correspond to a grouping or clustering of like rows of X). The interpretation for the case of near replication generalizes the preceding concepts.

2.1 Su and Yang Test

Su and Yang (2006) proposed an overall lack of fit test, along with tests for detecting between- and within-cluster lack of fit. The overall lack of fit test is based on a constructed full model which is intended to approximate the true model locally within each cluster, according to a specified clustering of the observations. In particular, the authors model the j^{th} response in the i^{th} cluster as

$$y_{ij} = x'_{ij}\beta + w'_{ij}\alpha_i + e_{ij} \tag{3}$$

for $i = 1,...,c$ and $j = 1,...,n_i$.

Here x'_{ij} is a known $1 \times p$ vector of regression variables associated with the proposed model and w'_{ij} is a $1 \times q_i$ vector of regression variables determined as functions of the variables in x'_{ij}. For example, in addition to containing a 1 which functions as an intercept for cluster i, w'_{ij} may contain the first and second order powers of all of the predictor variables in x'_{ij}. If sufficiently large amounts of data are available within a cluster, then a more complex model may be fitted where the w'_{ij} may contain functions of the variables in x'_{ij} other than powers and cross-products, such as trigonometric functions and wavelets. The β and α_i are unknown parameter vectors associated with x'_{ij} and w'_{ij}, respectively.

In matrix form, the constructed full model is

$$Y = X\beta + W\alpha + e \tag{4}$$

where

$$X = [x'_{ij}]$$

is a $n \times p$ matrix with $n = \sum_{i=1}^{c} n_i$ and

$$W = \text{Diagonal}[W_1, W_2, \cdots, W_c]$$

is a $n \times q$ matrix with $q = \sum_{i=1}^{c} q_i$, where

$$W_i = [w'_{ij}]$$

is a $n_i \times q_i$ matrix and

$$a' = (a'_1, a'_2, \cdots, a'_c).$$

An F-statistic for testing the proposed model $Y = X\beta + e$ for lack of fit as compared to the constructed full model (4) is

$$F = \frac{\| (P_{C(X,W)} - P_{C(X)})Y \|^2 \big/ (\dim C(X,W) - \dim C(X))}{\| (I_n - P_{C(X,W)})Y \|^2 \big/ (n - \dim C(X,W))}.$$

Lack of fit is concluded at level a whenever observed

$$F > F_{\dim C(X,W) - \dim C(X), n - \dim C(X,W)}(1-\alpha).$$

As noted by Su and Yang, the success of the proposed overall test depends on how well model (4) approximates the true model locally within each cluster. The authors provide an analytical comparison of the overall test for lack of fit with the test presented by Christensen (1989) based on expected mean squares. The analysis explains why the proposed overall test may perform better than Christensen’s test, especially in the presence of within-cluster lack of fit.

If lack of fit is detected by the overall test, then Su and Yang suggest the following test statistics for detecting between- and within-cluster lack of fit

$$F = \frac{\| (P_{C(X,Z)} - P_{C(X)})Y \|^2 \big/ (\dim C(X,Z) - \dim C(X))}{\| (I_n - P_{C(X,W)})Y \|^2 \big/ (n - \dim C(X,W))}$$

and

$$F = \frac{\| (P_{C(X,W)} - P_{C(X,Z)})Y \|^2 \big/ (\dim C(X,W) - \dim C(X,Z))}{\| (I_n - P_{C(X,W)})Y \|^2 \big/ (n - \dim C(X,W))},$$

respectively, to determine whether there is a dominant pure type of lack of fit. The critical points associated with these test statistics are

$$F_{\dim C(X,Z) - \dim C(X), n - \dim C(X,W)}(1-\alpha)$$

and

$$F_{\dim C(X,W) - \dim C(X,Z), n - \dim C(X,W)}(1-\alpha),$$

respectively. The matrix Z in the preceding test statistics represents a specified clustering. Su and Yang suggest that a reasonable number of degrees of freedom for the F-tests should be a factor in determining such clusterings of the observations into near replicates and also in the choice of a W matrix. When there are insufficient data, the authors suggest other tests, such as those presented by Christensen (1989, 1991), may be preferable.

This report will focus on the overall lack of fit test and thus will involve model (4) as the constructed full model. Further, this report will consider only the case when p=1 for X and w'_{ij} = $(1, x_{ij}, x_{ij}^2)$ for the construction of W in the simulation study in this report.

2.2 Miller and Neill Test

Miller and Neill (2007) proposed a multiple testing procedure to test a proposed model $Y = X\beta + e$ for lack of fit. For the case of one predictor variable, the values of the predictor can be ordered to obtain useful groupings of the observations for detecting lack of fit in nonreplicated regression. In particular, we may consider groupings based on consecutive pairs, triples, quadruples, etc. along the ordered values of the predictor. Each such grouping determines a matrix $Z^{\#}$ of indicator variables and is used in the cluster-based regression lack of fit tests presented by Christensen (1989, 1991), with such grouping serving as the clusterings. In particular, the groupings provide function approximations to the underlying true regression, and hence variance estimators to be used in testing for lack of fit. Using the multiple testing approach of Baraud et al. (2003) with families of such groupings, Miller and Neill demonstrated that lack of fit involving low signal-to-noise ratios and high frequency misspecifications can be effectively detected.

The importance of ordering predictors, residuals or sequences of alternative models in the multiple regression setting, and the associated difficulties in testing for lack of fit, has been noted by Aerts et al. (1999, 2000 and 2004) and Fan and Huang (2001). For the purpose of forming families of groupings in higher dimension for use with Christensen's lack of fit tests, Miller and Neill proposed that such families be determined by linear orders on the predictors based on disjoint parallel tubes in predictor space. Kulasekera and Gallagher (2002) used a similar tube construction, along with smoothness conditions for a specified nonparametric regression surface, in order to obtain consistency and asymptotic normality of difference-based estimators of variance determined by such ordering. A more complete discussion for the case of predictors with dimension greater than one is given by Miller and Neill (2007). However, the focus of this report will be on the case of one predictor variable without replication.

Now let $Z^{\#}$ denote a matrix of indicator variables corresponding to a grouping determined by taking groups of consecutive predictors along the linear orders for predictors of dimension one as described in the preceding. Then the lack of fit space $C(X)^{\perp}$ can be written as

$$C(X)^{\perp} = B^{\#} \oplus W^{\#} \oplus S^{\#}$$

where $B^{\#} = C(X)^{\perp} \bigcap C(Z^{\#})$, $W^{\#} = C(X)^{\perp} \bigcap C(Z^{\#})^{\perp}$ and $S^{\#} = (B^{\#} \oplus W^{\#})^{\perp}_{C(X)^{\perp}}$. This decomposition of the lack of fit space follows Christensen (1991) with the clustering given by the grouping represented by $Z^{\#}$. Analogous to the case of replication, the first two subspaces in the above decomposition are called the (orthogonal) between- and within-cluster lack of fit subspaces, respectively, corresponding to a particular grouping. The third subspace $S^{\#}$ has relatively low dimension. Thus, for a specified grouping represented by $Z^{\#}$, Miller and Neill test the model given by

$$H_0 : Y = X\beta + e$$

for lack of fit by comparing it to the complementary full models (i.e. two alternative full models whose error spaces are complementary with respect to $C(X)^{\perp}$) specified by

$$H_a^{B^\#} : Y = X\theta + Q_{B^\#}\gamma + e$$

and

$$H_a^{W_E^\#} : Y = X\theta + Q_{W_E^\#}\gamma + e$$

where $W_E^\# = W^\# \oplus S^\#$ is the (extended) within-cluster lack of fit subspace, and where $Q_{B^\#}$ and $Q_{W_E^\#}$ are matrices such that $C(Q_{B^\#}) = B^\#$ and $C(Q_{W_E^\#}) = W_E^\#$. Thus, H_0 is rejected if

$$F_{B^\#} > F_{\dim B^\#, \dim W_E^\#}(1-\alpha)$$

or if the complementary test statistic

$$F_{W_E^\#} = 1/F_{B^\#} > F_{\dim W_E^\#, \dim B^\#}(1-\alpha)$$

where $F_{B^\#}$ is the likelihood ratio test statistic for testing H_0 versus $H_a^{B^\#}$ given by

$$F_{B^\#} = \frac{\| P_{B^\#} Y \|^2 / \dim B^\#}{\| P_{W_E^\#} Y \|^2 / \dim W_E^\#}.$$

We remark that especially for predictors of dimension greater than one the use of Christensen’s tests, based on most clusterings, is likely to involve mixtures of (orthogonal) between- and within-cluster lack of fit. Thus, it is important to be able to detect such. With this in mind, let Z be any grouping matrix such that $C(Z) \subset C(Z^\#)$ where $Z^\#$ is a matrix of indicator variables for the nonreplicated case of the sort described above. As shown by Miller and Neill (2007), when lack of fit exists as a combination of the two pure types as determined by Z, the expectation of the numerator in the $F_{B^\#}$ statistic includes a function of such mixture lack of fit. This allows tests based on suitable $Z^\#$ groupings to possess effective power to detect mixtures of Christensen’s (orthogonal) between- and within-cluster lack of fit. In addition, the use of $Z^\#$ groupings associated with alternative spaces of large dimension give complementary tests based on critical points which can provide effective power for detecting of lack of fit. For example, using groupings based on consecutive pairs according to the linear orders in the nonreplicated case, the degrees of freedom parameters are approximately equal, assuming $p \ll n/2$. In particular, if $\dim C(Z^{\#}) = n/2$ then (generically) $\dim B^{\#} = n/2 - p$ and $\dim W_E^{\#} = n/2$. (Alternatively, similar results are obtained for complementary tests based on the within- and (extended) between-cluster lack of fit subspaces $W^{\#}$ and $B_E^{\#} = B^{\#} \oplus S^{\#}$, respectively. In particular, (generically) $\dim W^{\#} = n/2 - p + 1$ and $\dim B_E^{\#} = n/2 - 1$ whenever $\dim C(Z^{\#}) = n/2$.) However, given the unknown nature of the underlying regression function, $Z^{\#}$ groupings associated with alternative subspaces of smaller dimension are also potentially useful. Thus, Miller and Neill also considered groupings based on consecutive triples, quadruples, quintuples, hextuples, etc. The use of alternative subspaces with various dimensions reflects the bias-variance tradeoff problem, as encountered in nonparametric smoothing (Hart (1997) and Wasserman (2006)). It is also analogous to the use of a family of bandwidths in the scale space approach to curve estimation as discussed in Chaudhuri and Marron (1999, 2000).

As noted previously, the efficacy of a particular choice of grouping of the observations depends on the unobservable lack of fit. Thus, to enhance the power to detect general lack of fit associated with the proposed model $Y = X\beta + e$, Miller and Neill considered a testing procedure which involves doing several pairs of complementary tests, each based on a different grouping contained in a specified family of such groupings. For the case of one predictor, the authors considered families of groupings based on consecutive pairs, triples, quadruples, etc. and use the multiple testing approach of Baraud et al. (2003) with corresponding complementary test statistics $F_{B^{\#}}$ and $F_{W_E^{\#}}$.

To describe a multiple testing procedure based on a family of groupings, let $\mathcal{G}$ denote the collection of groupings under consideration. Also, let the set of corresponding pairs of complementary lack of fit subspaces $\mathcal{S} = \{B^{\#}, W^{\#} : Z^{\#} \in \mathcal{G}\}$ be indexed by a set $\mathcal{M}$. That is, $\mathcal{S} = \{S_m : m \in \mathcal{M}\}$ where S_m is a $B^{\#}$ or a $W^{\#}$ for some $Z^{\#} \in \mathcal{G}$. Let $V_m = C(X) \oplus S_m$ and note that card $\mathcal{M} = 2$ card $\mathcal{G}$. Following Baraud et al. (2003), let

$$T_a = \sup_{m \in \mathcal{M}} \left(\frac{\| P_{S_m} Y \|^2 / D_m}{\| (I - P_{V_m}) Y \|^2 / N_m} - \bar{F}^{-1}_{D_m, N_m}(a_m) \right)$$

and reject H_0 whenever T_a >0. In the preceding, $D_m = \dim S_m$, $N_m = \dim V_m^{\perp}$ and $\{a_m : m \in \mathcal{M}\}$ is a collection of numbers in (0, 1) such that $\Pr_{H_0}(T_a > 0) \leq a$. Note that with $e \sim N(0, \sigma^2 I_n)$, this multiple testing procedure rejects the adequacy of the proposed model if the F-statistic for testing H_0 against $H_{a,m} : E(Y) \in V_m$ is significant at level a_m for some $m \in \mathcal{M}$. Further, as stated in Baraud et al., if a_n denotes the a-quantile of the random variable

$$T_n = \inf_{m \in \mathcal{M}} \bar{F}_{D_m, N_m}(\mathcal{R}_m),$$

where

$$\mathcal{R}_m = \frac{\| P_{S_m} e \|^2 / D_m}{\| e - P_{V_m} e \|^2 / N_m},$$

then the choice of $a_m = a_n$ for all $m \in \mathcal{M}$ ensures that $\Pr_{H_0}(T_a > 0) \leq a$.

As in Baraud et al. (2003) and Miller and Neill (2007), in this report simulation (using Gaussian errors) is used to determine the value of a_n. Alternatively, a choice of a_m such that $\sum_{m \in \mathcal{M}} a_m = \alpha$ provides a conservative testing procedure with level at most equal to α according to the Bonferonni inequality. However, with $\alpha_m = \alpha / \text{card } \mathcal{M}$, this alternative choice for α_m was shown by Baraud et al. to be more robust with respect to departures from Gaussian errors.

CHAPTER 3 - Simulation Studies

The lack of fit testing procedures proposed by Su and Yang (2006) and Miller and Neill (2007) are compared with various data generating models and parameter settings in the following simulation studies. In all of the studies, the null model to be tested for lack of fit is the univariate regression model $y = \beta_0 + \beta_1 x + e,$ where the errors were assumed to be independent and identically distributed according to $N(0, \sigma^2)$. The errors for the data generating models were randomly generated according to $N(0,1)$. The empirical power for each case considered is based on 2000 simulated datasets corresponding to each parametric setting of a particular data generating model. The significance level was set at $\alpha = .05$ for all cases. In the following, the parameter Nsize refers to the groupings based on consecutive pairs, triples, quadruples, etc. along the ordered values of the predictor as required for the Miller and Neill test. Also, the required clusterings for the Su and Yang test were determined by the R functions *cutree* and *hclust* to create hierarchical clustering groups corresponding to specified numbers of clusters. The R codes used for all calculations in the simulations are collected in the Appendix to this report. Finally, the results for the Su and Yang test will be labeled with SY, while the results for the Miller and Neill test will be labeled with MN for brevity.

3.1 The First Simulation Study

For the first simulation study, $n = 25$ observations were generated according to the true model $y = \beta_1 x + \beta_2 \sin(4x) + e$, where $\beta_1 = 1$, and β_2 = 0, 0.8, 1.6, 2.4, 3.2. Also, x takes the values 0.0, 0.2, 0.4, 0.6, 0.8, 2.0, 2.2, 2.4, 2.6, 2.8, 4.0, 4.2, 4.4, 4.6, 4.8, 6.0, 6.2, 6.4, 6.6, 6.8, 8.0, 8.2, 8.4, 8.6, 8.8. This data generating model was also used by Su and Yang (2006), where values for the predictor x possess a clear clustering pattern. In particular, the $n = 25$ values can be readily partitioned into $c = 5$ clusters. A scatter plot of a typical simulated dataset generated by true model, along with the fitted null model and true model regression function are given in Figure 3.1. The empirical powers for the SY and MN testing procedures for various parameters are listed in Table 3.1 and plotted in Figure 3.2.

Figure 3.1 A scatter plot of the data generated from the model $y = x + \beta_2 \sin(4x) + e$, along with the fitted null model and true regression curve.

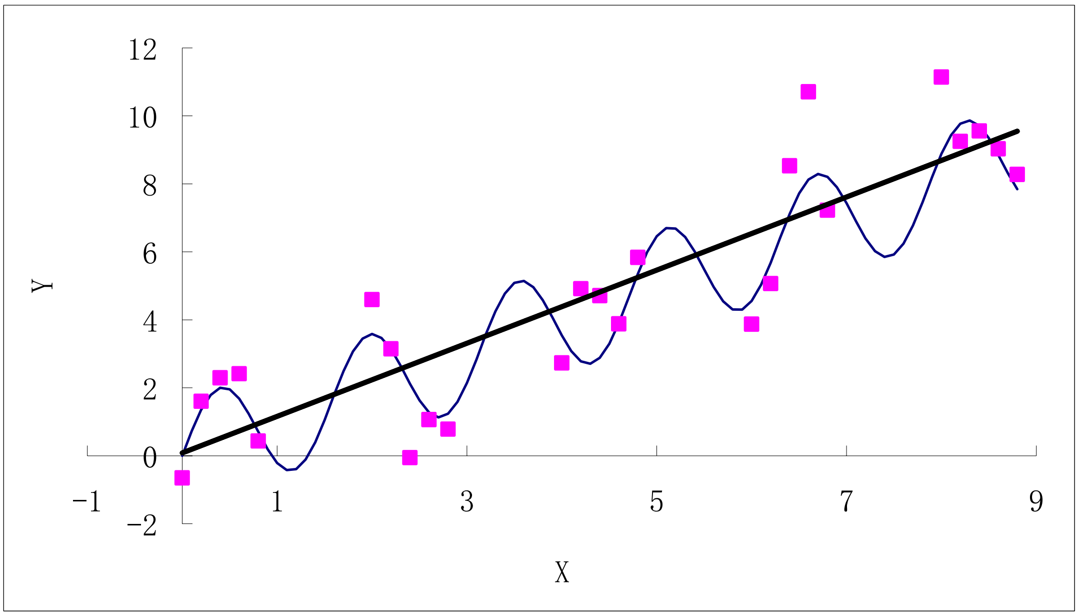
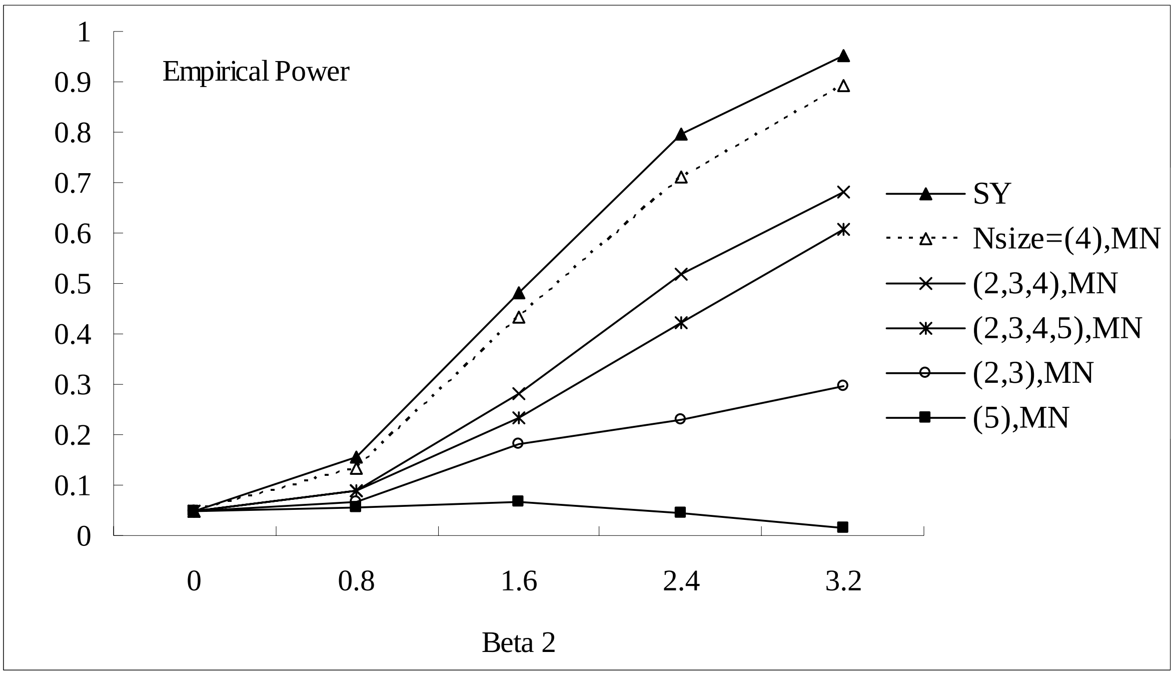


Table 3.1 Empirical Power for the SY and MN tests with data generated from the model $y = x + \beta_2 \sin(4x) + e$.

β_2	MN test					SY test
	Nsize=(2,3)	(2,3,4)	(2,3,4,5)	(4)	(5)	
0	.0405	.0575	.0435	.054	.053	.047
.8	.0655	.09	.088	.133	.054	.155
1.6	.18	.28	.235	.433	.065	.4815
2.4	.231	.52	.424	.7125	.045	.798
3.2	.297	.681	.606	.893	.0165	.9515

Figure 3.2 Empirical Power for the SY and MN tests with data generated from the model $y = x + \beta_2 \sin(4x) + e$.



With a clear pattern of clustering and lack of fit that is a function of only the specified predictor, the SY test has effective power for the selected parameter settings in this particular data generating model. Empirical powers for the MN test were computed for several settings of the Nsize parameter corresponding to several families of groupings. Each setting of Nsize provides a family of alternative lack of fit subspaces with various dimensions. Since the efficacy of a particular choice of grouping depends on the unobservable lack of fit, a family of groupings with sufficient breadth of dimensions is generally required for a more powerful testing procedure. This is reflected in the sensitivity to the Nsize settings for the MN test. Since the MN test accommodates a broader alternative than the SY test, the empirical powers for the MN test are comparatively lower than those for the SY test for this particular simulation.

In order to compare the SY and MN tests for detecting lack of fit with different frequency components, data was generated according to the true model $y = \beta_1 x + \beta_2 \sin(x) + e$ where the parameters β_1 and β_2 were chosen as before, along with the same values of the predictor x. A scatter plot of a typical simulated dataset generated by this true model, along with the fitted null model and true model regression function are given in Figure 3.3. The empirical powers for the SY and MN testing procedures for various parameters are listed in Table 3.2 and plotted in Figure 3.4.

Figure 3.3 A scatter plot of data generated from the model $y = x + \beta_2 \sin(x) + e$, along with the fitted null model and true regression curve.

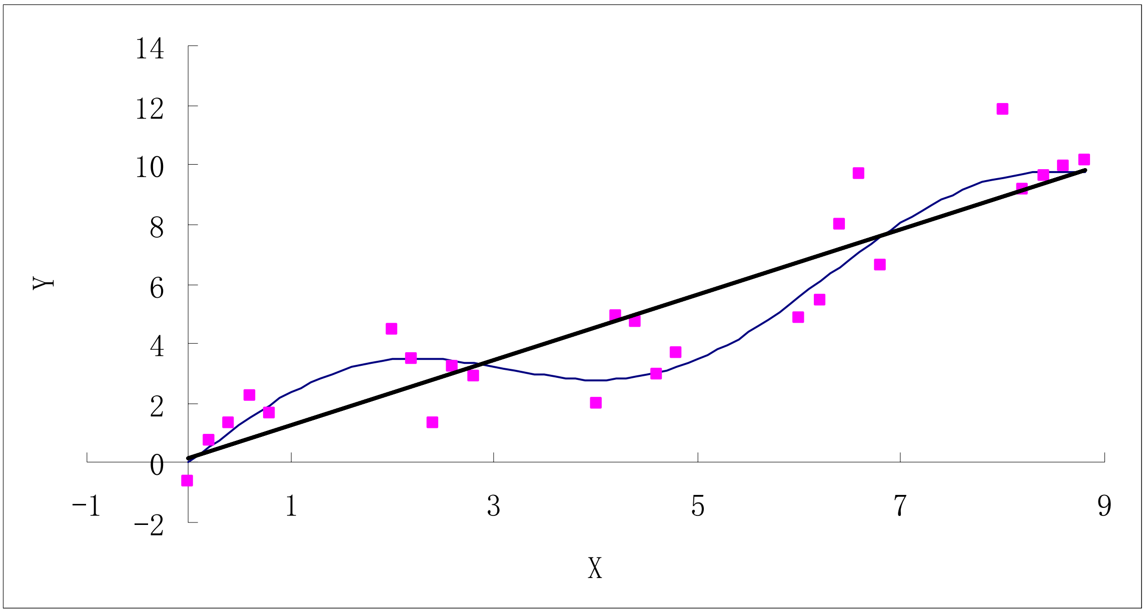
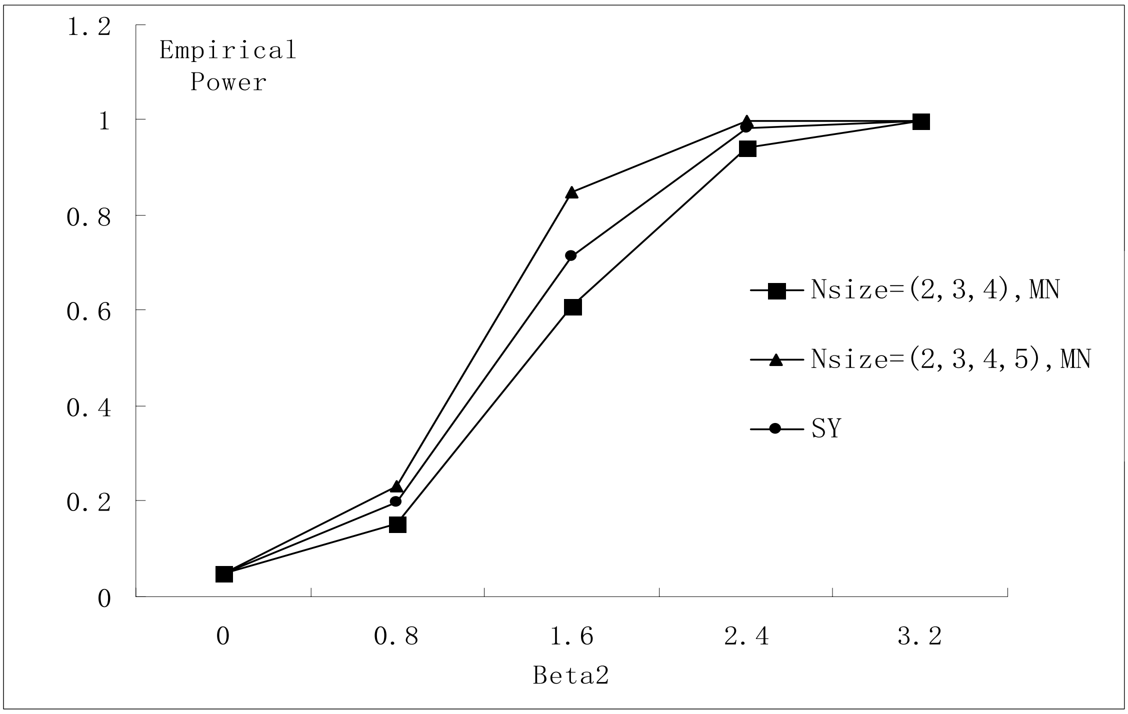


Table 3.2 Empirical Power for the SY and MN tests with data generated from the model $y = x + \beta_2 \sin(x) + e$.

β_2	MN test		SY test
	Nsize=(2,3,4)	Nsize=(2,3,4,5)	
0	.049	.055	.05
.8	.154	.2315	.197
1.6	.6095	.8475	.7154
2.4	.9425	.998	.984
3.2	.9965	1	1

Figure 3.4 Empirical Power for the SY and MN tests with data generated from the model $y = x + \beta_2 \sin(x) + e$.



As before, when there is a clear pattern of clustering and lack of fit is a function of only the specified predictor, the SY test has effective power. Empirical powers for the MN test with either setting of Nsize are reasonably effective as well for this particular simulation.

3.2 The Second Simulation Study

For the second simulation study, $n = 25$ observations were generated according to the true model $y = \beta_1 x + \beta_2 \sin(\omega x) + e$ as in the first simulation study, except that the values of the predictor variable x were taken from 0.352 to 8.8 at equally spaced intervals of length 0.352. Unlike the first simulation, the values of x do not exhibit a clear clustering structure. Since the SY test requires a specified clustering, several different clusterings were used for evaluation of the SY test. In particular, clusterings with 3, 4, 5 and 6 groups were used to construct the alternative model for the SY test. The empirical powers for the case when $\omega = 1$ are listed in

Table 3.3 and plotted in Figure 3.5. Both testing procedures exhibit comparable and effective power for this case. However, the results for the case when $\omega = 4$ indicate that the SY test may not necessarily perform well when a clear clustering structure is not present. The MN test with a family of groupings corresponding to Nsize = (2, 3, 4, 5) exhibits effective power for this simulation case. The empirical powers for this case are listed in Table 3.4 and plotted in Figure 3.6.

Table 3.3 Empirical Power for the SY and MN tests with data generated from the model $y = x + \beta_2 \sin(x) + e$.

β_2	MN test	SY test			
	Nsize=(2,3,4,5)	Clusters=3	Clusters=4	Clusters=5	Clusters=6
0	.0565	.0505	.0515	.0505	.054
.8	.228	.3455	.265	.2175	.1965
1.6	.835	.94	.8345	.76	.6905
2.4	.996	1	.996	.99	.9785
3.2	1	1	1	1	1

Figure 3.5 Empirical Power for the SY and MN tests with data generated from the model $y = x + \beta_2 \sin(x) + e$.

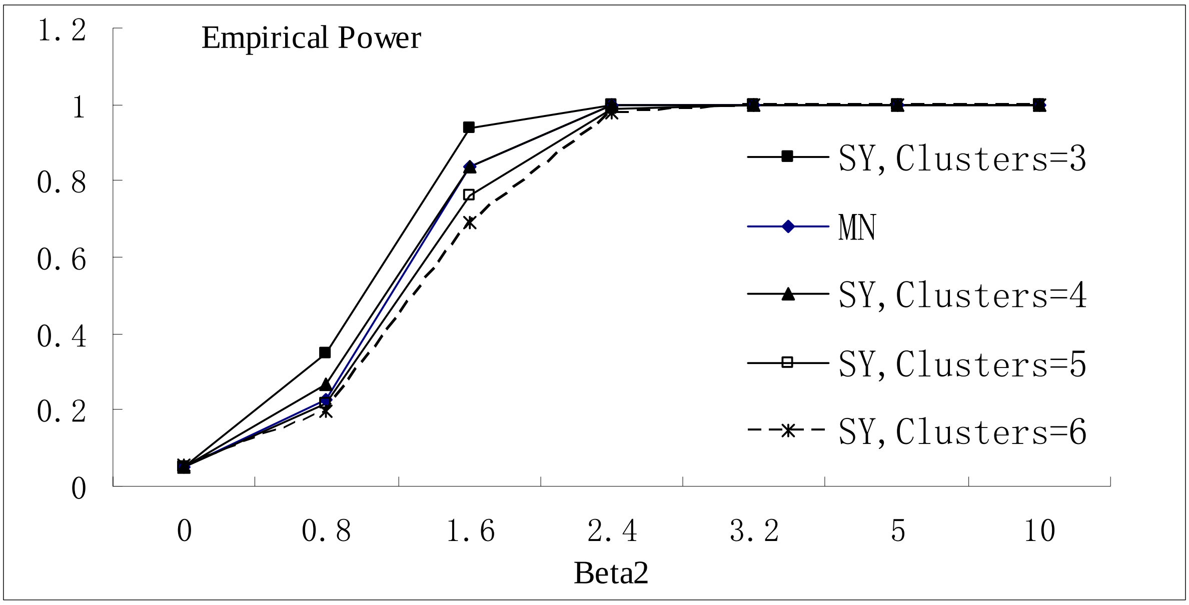
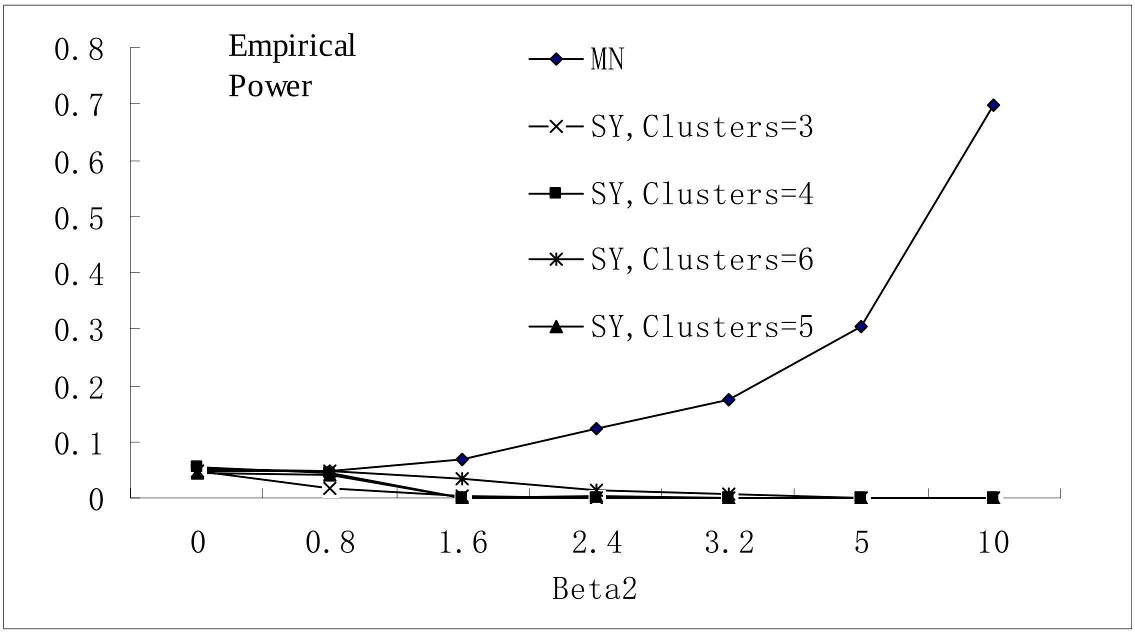


Table 3.4 Empirical Power for the SY and MN tests with data generated from the model $y = x + \beta_2 \sin(4x) + e$.

β_2	MN test	SY test			
	Nsize=(2,3,4,5)	Clusters =3	Clusters =4	Clusters =5	Clusters =6
0	.049	.047	.053	.046	.048
.8	.0465	.0185	.0455	.041	.0475
1.6	.07	.003	.0011	.0016	.0325
2.4	.124	0	0	.0035	.0145
3.2	.175	0	0	.0001	.006
5	.305	0	0	0	0
10	.6975	0	0	0	0

Figure 3.6 Empirical Power for the SY and MN tests with data generated from the model $y = x + \beta_2 \sin(4x) + e$.



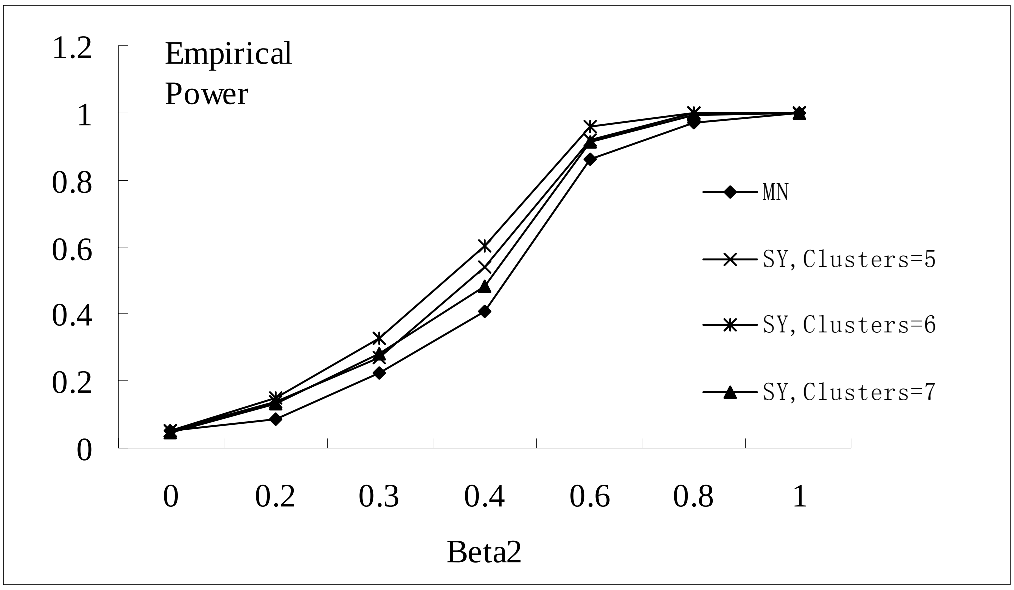
3.3 The Third Simulation Study

For the third simulation study, observations were generated according to the true model $y = 1 + \beta_2 x^2 + e$ where the predictor variable x takes values generated from the uniform distribution on (-2, 2). As in the second simulation study, the values of x do not exhibit a clear clustering structure so that clusterings with 5, 6 and 7 groups were used to construct the alternative model for the SY test for the case of $n = 64$ observations. To further investigate the effect of sample size on the performance of the MN test, the case of $n = 120, 180$ were also considered, each with the parameter Nsize = (2, 3, 4, 5). Since the lack of fit is due to the omission of a polynomial term in the specified predictor variable, the lack of fit is dominated by the between-cluster pure type. As a result, both testing procedures exhibit comparable and effective power for this particular simulation. The empirical power results are listed in Table 3.5 and plotted in Figure 3.7. Note that only the results for the case $n = 64$ are plotted.

Table 3.5 Empirical Power for the SY and MN tests with data generated from the model $y = 1 + \beta_2 x^2 + e$.

β_2	MN test			SY test		
	$n = 64$	$n = 120$	$n = 180$	$n = 64$		
	Nsize=(2,3,4,5)	Nsize=(2,3,4,5)	Nsize=(2,3,4,5)	Clusters=5	Clusters=6	Clusters=7
0	.0425	.052	.055	.0505	.0535	.0465
.2	.088	.1163	.152	.137	.148	.1305
.3	.223	.3268	.4235	.2715	.33	.2815
.4	.4097	.6595	.8185	.5425	.6025	.4845
.6	.8597	.989	1	.918	.9615	.914
.8	.9711	1	1	.998	.9995	.9955
1	.9985	1	1	.9995	1	1

Figure 3.7 Empirical Power for the SY and MN tests with data generated from the model $y = 1 + \beta_2 x^2 + e$.



3.4 The Fourth Simulation Study

For the fourth simulation study, observations were generated according to the true model

$$y = \frac{10}{1.0 + \beta_2 \exp(-2x)} + e,$$

where the predictor x takes values randomly generated from the $N(0,1)$ distribution. As in the second and third simulation studies, the values of x do not exhibit a clear clustering structure so that clusterings with 5, 6 and 7 groups were used to construct the alternative model for the SY test for the case of $n = 64$ observations. To further investigate the effect of sample size on the performance of the MN test, the cases of $n = 120, 180$ were also considered, each with the parameter Nsize = (2, 3, 4, 5). With the true model representing a logistic regression model, both testing procedures exhibit comparable and effective power, although the empirical power for the SY test is slightly higher than that for the MN test for comparable sample size. Note that both tests indicate some loss of power for increasing values of β_2 for this particular simulation. The empirical power results are listed in Table 3.6.

Table 3.6 Empirical Power for the SY and MN tests with data generated from the model $y = \frac{10}{1.0 + \beta_2 \exp(-2x)} + e.$

β_2	MN test			SY test		
	n=64	n=120	n=180	n=64		
	Nsize=(2,3,4,5)	Nsize=(2,3,4,5)	Nsize=(2,3,4,5)	Clusters=5	Clusters=6	Clusters=7
0	.055	.047	.0525	.052	.0535	.0485
.2	1	1	1	1	1	1
.4	.999	.999	1	1	.9995	.999
.6	.9845	.989	1	.999	.995	.987
.8	.9335	.9675	1	.9895	.977	.952
1	.868	.9605	1	.9735	.9445	.907

3.5 The Fifth Simulation Study

For the fifth simulation study, $n = 50$ observations were generated according to the true model $y = 5 + 3x_1 + \beta_2 x_2 + e$ where predictors x_1 and x_2 take predetermined values as indicated in Figure 3.8. The null model remains a simple linear regression model of y on x_1 in this case, but unlike the preceding simulation studies the true model depends on an unspecified predictor variable x_2. With only x_1 specified, there are $c = 5$ readily identified clusters. However, there are consistent trends across such clusters determined by the x_2 predictor variable, as can be seen in Figure 3.8. Based on a constructed alternative model using the $c = 5$ clusters determined by x_1, the SY test possesses extremely low power. However, the MN test with Nsize = (2, 3, 4, 5) exhibits effective power, even in the case when the true data generating model depends on an unspecified predictor and when such lack of fit involves consistent trends across the identifiable clusters determined only by the specified predictor. It is of interest to note that when the constructed alternative model for the SY test is based on the $c = 10$ clusters that can be readily identified in Figure 3.8 by using both predictors x_1 and x_2, the corresponding power values are clearly improved. Of course, such clustering would not be available to the experimenter in practice since x_2 is unknown. The empirical power results are listed in Table 3.7 and plotted in Figure 3.9.

Figure 3.8 Scatter plot for the predictor variables x_1 and x_2 for Simulation Study 5.

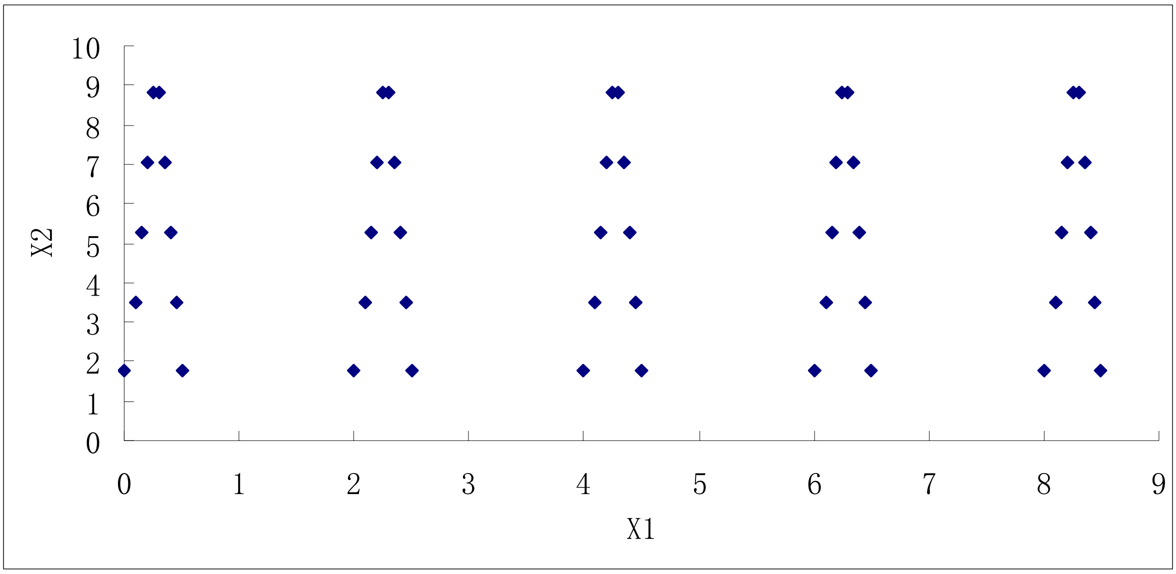
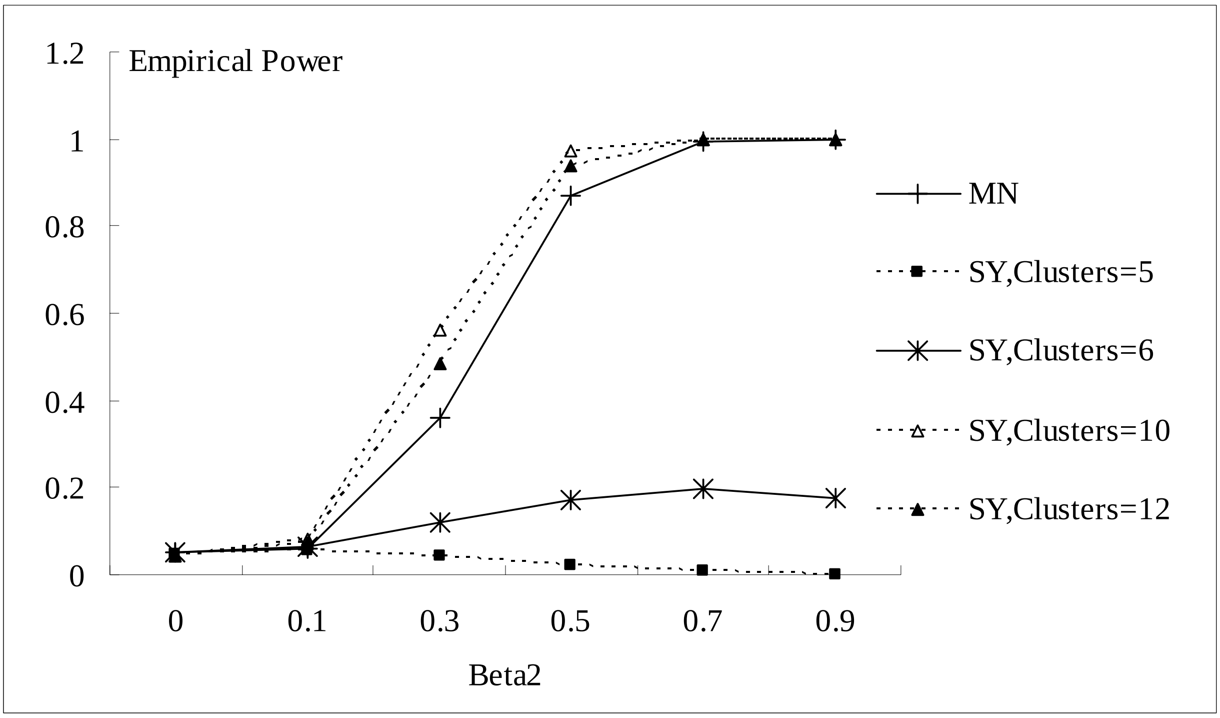


Table 3.7 Empirical Power for the SY and MN tests with data generated from the model $y = 5 + 3x_1 + \beta_2 x_2 + e$.

β_2	MN test	SY test			
	Nsize=(2,3,4,5)	Clusters=5	Clusters=6	Clusters=10	Clusters=12
0	.0445	.0485	.0513	.043	.0435
.1	.0585	.0543	.0623	.0825	.0742
.3	.36	.042	.119	.562	.485
.5	.871	.0205	.172	.971	.937
.7	.994	.008	.198	.9995	.998
.9	1	.002	.174	1	1

Figure 3.9 Empirical Power for the SY and MN tests with data generated from the model $y = 5 + 3x_1 + \beta_2 x_2 + e$.



3.6 The Sixth Simulation Study

For the sixth simulation study, $n = 80$ observations were generated according to the true model $y = 5 + 3x_1 + \beta_2 \cos(x_2) + e$ where the predictor x_1 takes sorted values generated from the uniform distribution on (0, 10) and the predictor x_2 takes predetermined values according to a repeating sequence between 0 and 10 at equally spaces intervals of 10/15. A typical simulated scatter plot of x_1 and x_2 values is given in Figure 3.10. The null model remains a simple linear regression model of y on x_1 in this case, and like the fifth simulation study the true model depends on an unspecified predictor variable x_2. With only x_1 specified, there is no clear clustering structure so that clusterings with various group sizes were used to construct the alternative model for the SY test. However, there are consistent trends across the values of x_1

determined by the x_2 predictor variable, as can be seen in Figure 3.10. Based on constructed alternative models using 5, 6 and 7 clusters on the unstructured values taken by x_1, the SY test possesses extremely low power. However, it is of interest to note that when constructed alternative models for the SY test are based on 12, 14, and 16 clusters, the corresponding power values are clearly improved. Of course, such clusterings would not necessarily be selected by the experimenter in practice since x_2 is unknown. In contrast, the MN test with Nsize = (2, 3, 4, 5) exhibits effective power, even in this case where the true data generating model depends on an unspecified predictor and the lack of fit involves consistent trends across unstructured values taken by the specified predictor. The empirical power results are listed in Table 3.8 and plotted in Figure 3.11.

Figure 3.10 Scatter plot for the predictor variables x_1 and x_2 for Simulation Study 6.

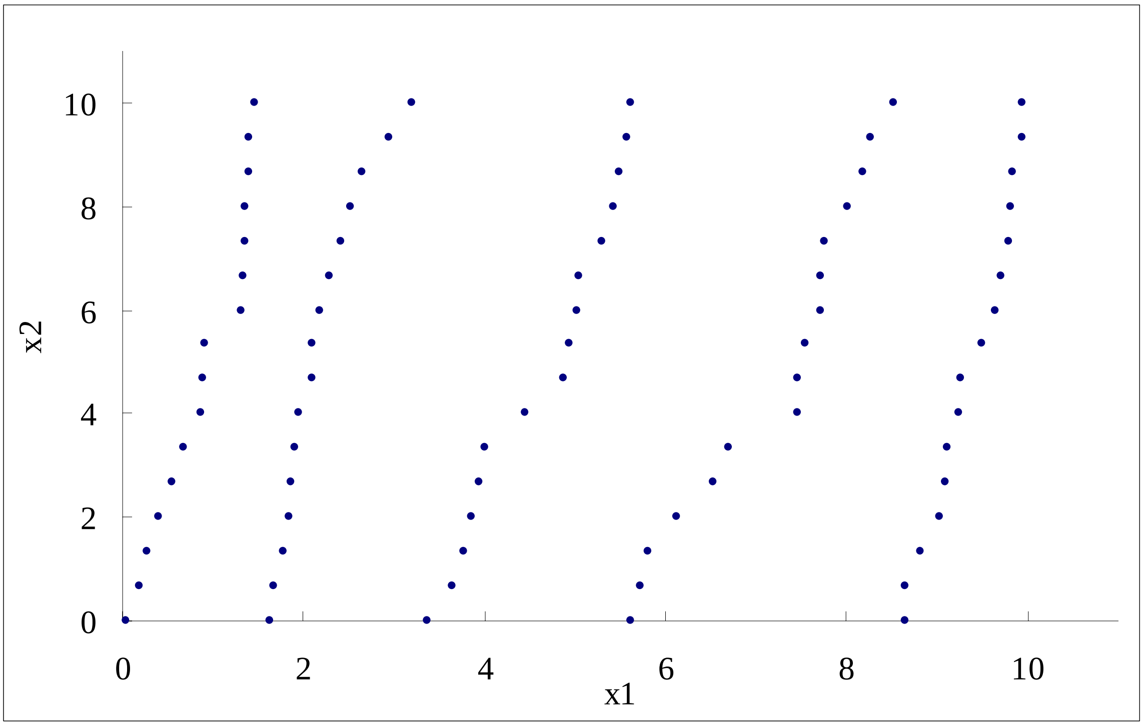
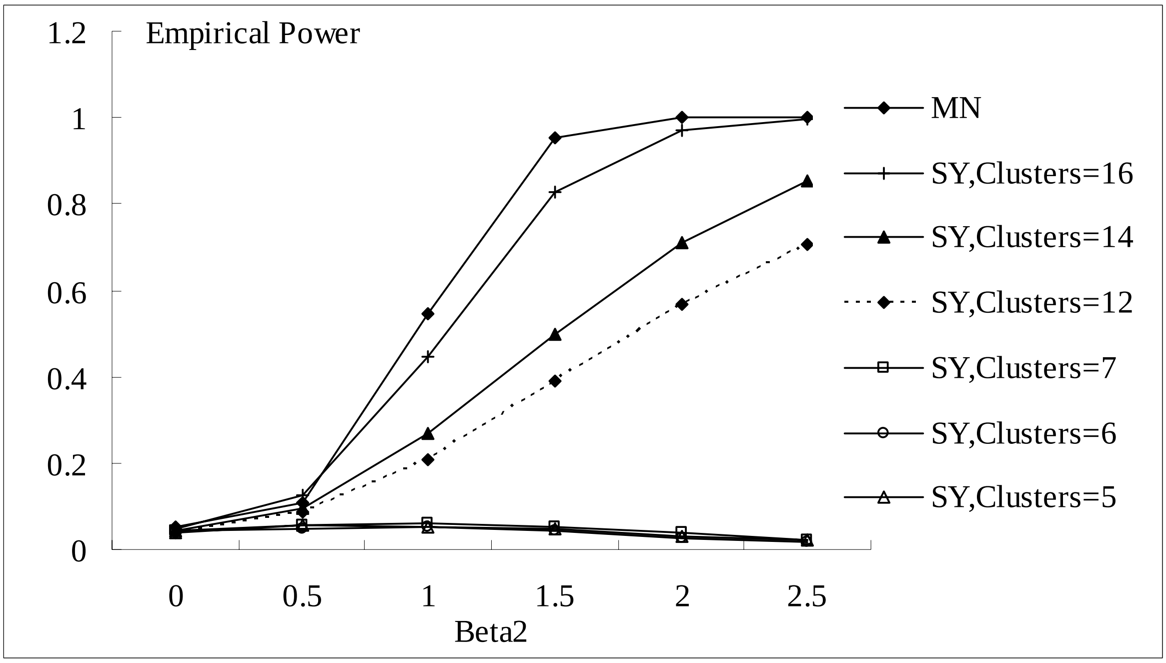


Table 3.8 Empirical Power for the *SY* and *MN* tests with data generated from the model $y = 5 + 3x_1 + \beta_2 \cos(x_2) + e$.

β_2	*MN* test	*SY* test					
	Nsize=(2,3,4,5)	Clusters=5	Clusters=6	Clusters=7	Clusters=12	Clusters=14	Clusters=16
0	.0535	.04	.043	.044	.0405	.0425	.0465
.5	.108	.0545	.0495	.0565	.087	.095	.124
1.0	.5455	.054	.0535	.0605	.21	.2705	.446
1.5	.954	.049	.042	.053	.3905	.498	.8255
2.0	1	.032	.026	.0375	.5665	.711	.97
2.5	1	.022	.0165	.0225	.704	.8555	.997

Figure 3.11 Empirical Power for the *SY* and *MN* tests with data generated from the model $y = 5 + 3x_1 + \beta_2 \cos(x_2) + e$.



CHAPTER 4 - Conclusion

The goal of this report is to first review two recently proposed cluster based regression lack of fit tests. These test procedures were presented by Su and Yang (2006) and Miller and Neill (2007), and address the problem of detecting lack of fit which may exist as a combination of the two pure types of between- and within-cluster lack of fit, and were discussed in Chapters 1 and 2. The second goal of this report is to make some comparisons between the two testing procedures, at least for the case of one specified predictor variable. The simulation studies presented in Chapter 3 indicate that the test proposed by Su and Yang is especially effective when the lack of fit is not due to an unspecified predictor variable and when there is a clear pattern of clustering in the specified predictor variable. The simulation studies also indicate that the test proposed by Miller and Neill (2007) is especially effective when the family of alternative lack of fit subspaces possesses sufficient breadth of dimensions. This test accommodates a broader alternative, which can thus result in comparatively lower but effective power. However, this test demonstrated an ability to detect model inadequacy when the lack of fit was a function of an unspecified predictor variable and does not require a specified clustering for implementation. Future comparisons would involve the case of more than one specified predictor variable.

References

Aerts, M., Claeskens, G. and Hart, J.D. (1999). Testing the Fit of a Parametric Function, Journal of the American Statistical Association, 94, 869-879.

Aerts, M., Claeskens, G. and Hart, J. D. (2000). Testing Lack of Fit in Multiple Regression, Biometrika, 87, 405-425.

Aerts, M., Claeskens, G. and Hart, J.D. (2004). Bayesian-Motivated Tests of Function Fit and Their Asymptotic Frequentist Properties, The Annals of Statistics, 32, 2580-2615.

Atwood, C., and Ryan, T. (1977). A Class of Tests for Lack of Fit to a Regression Model, unpublished manuscript.

Baraud, Y., Huet, S. and Laurent, B. (2003). Adaptive Tests of Linear Hypotheses by Model Selection, The Annals of Statistics, 31, 225-251.

Bedrick, E. (2000). Checking for Lack of Fit in Linear Models with Parametric Variance Functions, Technometrics, 42, 227-236.

Breiman, L. and Meisel, W. S. (1976). General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models, Journal of the American Statistical Association, 71 301-307.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for Exploration of Structures in Curves, Journal of the American Statistical Association, 94, 807–823.

Chaudhuri, P. and Marron, J. S. (2000). Scale Space View of Curve Estimation. The Annals of Statistics, 28, 408–428.

Christensen, R. (1989). Lack of Fit Based on Near or Exact Replicate, The Annals of Statistics, 17, 673-683.

Christensen, R. (1991). Small Sample Characterizations of Near Replicate Lack of Fit Tests, Journal of the American Statistical Association, 88, 752-756.

Christensen, R. (2002). Plane Answers to Complex Questions: The Theory of Linear Models, 3rd ed., Springer.

Christensen, R. (2003). Significantly Insignificant F Tests, The American Statistician, 57, 27-32.

Christensen, R. and Bedrick, E. (1997). Testing the Independence Assumption in Linear Models. Journal of the American Statistical Association, 92, 1006-1016.

Cook, R. D. and Weisberg, s. (1997). Graphics for Assessing the Adequacy of Regression Models, Journal of the American Statistical Association, 92, 490-499.

Daniel, C. and Wood, F. S. (1980). Fitting Equations to Data 2nd ed., Wiley.

Eubank, R. L., Chin-Shang L. and Wang, S. (2005). Testing Lack-of-Fit of Parametric Regression Models Using Nonparametric Regression Techniques, Statistica Sinica, 15, 135-152.

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

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

Green, J. R. (1971). Testing Departure From a Regression Without Using Replication, Technometrics, 13, 609-615.

Hart, J. D. (1997). Nonparametric Smoothing and Lack of Fit Tests, Springer.

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

Khmaladze, E. V. and Koul, H. L. (2004). Martingale Transforms Goodness-of-Fit Tests in Regression Models, The Annals of Statistics, 32, 995-1034.

Kulasekera, K. and Gallagher, C. (2002). Variance Estimation in Nonparametric Multiple Regression, Communications in Statistics, Part A- Theory and Methods, 31, 1373-1383.

Lyons, N. I. and Proctor, C. H. (1977). A Test for Regression Function Adequacy, Communications in Statistics, Part A- Theory and Methods, 6, 81-86.

Miller, F. R. and Neill, J. W. (2007). General Lack of Fit Tests Based on Families of Groupings, To appear in Journal of Statistical Planning and Inference.

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.

Miller, F. R., Neill, J. W. and Sherfey, B. W. (1999). Implementation of a Maximin Power Clustering Criterion to Select Near Replicates for Regression Lack of Fit Tests, Journal of the American Statistical Association, 94, 610-620.

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

Shillington, E.R. (1979). Testing Lack-of-Fit in Regression Without Replication. Canadian Journal of Statistics, 7, 137-146.

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

Utts, J. M. (1982). The Rainbow Test for Lack of Fit in Regression, Communications in Statistics, Part A- Theory and Methods, 11, 2801-2815.

Wasserman L. (2006). All of Nonparametric Statistics, Springer.

Appendix A - R code for Simulation Study

```
##################################################################################
# The following R codes were used in the first, third and sixth simulation studies to compare
# Miller and Neill’s lack of fit test based on a family of groupings and Su and Yang’s overall
#lack of fit test method. The R codes for the other simulation studies are omitted due to
#similarities.
###################################################################################
```

Appendix A-1 R Code for Yang’s overall test for the First Simulation Study

```
source("ppo.s")
source("basis.s")
source("CZmat.s")
source("CXmat.s")
options(echo=F)
n<-25
ysim<- 2000
alpha<- .05
stdev<-1
onesn<- matrix(rep(1,n),byrow=F)
x1<-
matrix(c(0,.2,.4,.6,.8,2.0,2.2,2.4,2.6,2.8,4.0,4.2,4.4,4.6,4.8,6.0,6.2,6.4,6.6,6.8,8.0,8.2,8.4,8.6,8.8),
byrow=T)
x2<-sin(4*x1)
X<-x1
#################################
#predetermine the size for each cluster
```

```
###############################
Blksize<-c(5,5,5,5,5)
xperm<-rep(1:n)
C<- vector("list",1)
N<-0
for (i in 1:length(Blksize)){
    if(i==1) {N=Blksize[i]}
    if(i>1){N=N+Blksize[i]}
    start<-N-Blksize[i]+1
    C[[length(C)+1]]<-xperm[start:N]
    }
 C[1]<- NULL
 Cmat<- CZmat(C,n)
 Xmat<-CXmat(C,x1,n)
############################
 ## Get XW matrix
 ############################
 XW0<-lapply(as.list(1:length(Blksize)),function(i,Cmat,Xmat)
      {
          C1<-matrix(Cmat[,i],ncol=1)
          X1<-matrix(Xmat[,i],ncol=1)
          X12<-X1**2
          Wi<-cbind(C1,X1,X12)
           return(Wi)},Cmat,Xmat)

XW1<-matrix(unlist(XW0),n, 3*length(Blksize))
XW<-cbind(x1,XW1)

beta<-c(0,0.8,1.6,2.4,3.2)
for (L in 1:length(beta)){
   beta2<-beta[L]
```

```
     ytrue<-x1+beta2*x2
    YTRUE<-matrix(ytrue,n,byrow=T)
    nreject<-matrix(100)
    for (k in 1:ysim){
        error<- matrix(rnorm(n,mean=0,sd=stdev),n,byrow=T)
        Y<- YTRUE+error

        ############
        #   get ssex and dfssex
        #############
       dfssex<-basis( diag(n)-ppo(X))
       ssex<-t(Y)%*%(diag(n)-ppo(X))%*%Y

       dfssexw<-basis(diag(n)-ppo(XW))
       ssexw<-t(Y)%*%(diag(n)-ppo(XW))%*%Y

       ssnum<-ssex-ssexw
       dfnum<-dfssex-dfssexw
       msnum<-ssnum/dfnum
       ssden<-ssexw
       dfden<-dfssexw
       msden<-ssden/dfden
       F<-msnum/msden
       Fc<-qf(1-alpha,dfnum,dfden)
       nreject[k]<-F>Fc
      }
# ******* get test power
cat(" Simpower is \n")
print(simpower)
}
```

Appendix A-2 R code for the Third Simulation Study

```
source("ppo.s")
source("basis.s")
source("CZmat.s")
source("CXmat.s")
source("quantnorep.s")
options(echo=F)
n<-64
stdev<- 1
MIN<--2
MAX<-2

Nsize<- c(2,3,4,5)
ysim<- 2000
ansim<- 10000
alpha<- .05
thetaL<-c(0,.2,.4,.6,.8,1.0)

x<- matrix(runif(n,MIN,MAX),byrow=F)
xsq<-x**2
onesn<- matrix(rep(1,n),byrow=F)

X<- cbind(onesn,x)
error<- matrix(rnorm(n*ysim,mean=0,sd=stdev),ysim,n,byrow=T)
for (L in 1:length(thetaL)){                    #calculate for different theta

##############################
##  based on Neill's method
#############################
theta<-thetaL[L]
```

```
#cat("theta is \n")
#print(theta)

lof<-theta*xsq
LOF<- matrix(lof,ysim,n,byrow=T)

Y<- LOF+error
xperm<- order(x)

Clist<- lapply(as.list(1:length(Nsize)),function(i,xperm,Nsize)
     {Ci<- vector("list",1)
      N<- Nsize[i]
      Blk<- floor(length(xperm)/N)
      start1<- 1

      if(1<=Blk)
               {for(j in 1:Blk)
                       {mj<- ((j-1)*N)+start1
                        Ci[[length(Ci)+1]]<- xperm[mj:(mj+N-1)]}
               }
       resid<- length(xperm)-(Blk*N)

 if(resid>0)
 #exclude singletons     if(resid>1)
               {start2<- (Blk*N)+start1
                Ci[[length(Ci)+1]]<- xperm[start2:length(xperm)]}

       Ci[[1]]<- NULL
       return(Ci)
                },xperm,Nsize)
```

```
NC<- length(Clist)

Cmatlist<- lapply(Clist,function(C,n)
          {Cmat<- CZmat(C,n)
           return(Cmat)
                       },n)
MBlist<- lapply(Cmatlist,function(Cmat,X)
       {MB<- ppo(Cmat)-ppo(ppo(Cmat)%*%X)
        return(MB)
                 },X)
dfBlist<- lapply(MBlist,function(MB)
        {dfB<- basis(MB)

         return(dfB)
                   })
MXperp<- diag(n)-ppo(X)
dimCXperp<- basis(MXperp)
MWSlist<- lapply(MBlist,function(MB,MXperp)
        {MWS<- MXperp-MB
         return(MWS)
                   },MXperp)

dfWSlist<- lapply(as.list(1:NC),function(j,dfBlist,dimCXperp)
         {dfWS<- dimCXperp - dfBlist[[j]]

          return(dfWS)
                     },dfBlist,dimCXperp)

an<- quantnorep(n,ansim,NC,alpha,MBlist,MWSlist,dfBlist,dfWSlist)
Tvals<- apply(Y,1,function(y,NC,MBlist,MWSlist,dfBlist,dfWSlist,an)
      {
        bestcomp<- lapply(as.list(1:NC),function(j,y,an,MBlist,MWSlist,dfBlist,dfWSlist)
                {
                   MSB<- (sum((MBlist[[j]]%*%y)^2))/dfBlist[[j]]
                   MSWS<- (sum((MWSlist[[j]]%*%y)^2))/dfWSlist[[j]]

                   FB<- MSB/MSWS
                   FWS<- 1/FB

                   Bcpt<- qf(1-an,dfBlist[[j]],dfWSlist[[j]])
                   WScpt<- qf(1-an,dfWSlist[[j]],dfBlist[[j]])

                   diffB<- FB-as.numeric(Bcpt)
                   diffWS<- FWS-as.numeric(WScpt)

                   best<- max(diffB,diffWS)
                   return(best)
                           },y,an,MBlist,MWSlist,dfBlist,dfWSlist)

        Tval<- max(unlist(bestcomp))
        return(Tval)
                   },NC,MBlist,MWSlist,dfBlist,dfWSlist,an)
   nreject<- Tvals > 0
simpower<- mean(nreject)

cat("T simulated power based on Neill's method \n")
print(simpower)
#Tstats<- summary(Tvals)
#cat("T statistic summary\n")
#print(Tstats)
```

```
################################
## The following code is used to simulate for yang's overall test
## use the above x1, x2 and error matrix
###################################

#############################
# Cluster the dataset
# Then get the size for each cluster
#############################
cat(" Simpower based on Yang's test is \n")
cluster_n<-c(5,6,7,8,9,10,11,12,13,14,15,16)
x1<-matrix(sort(x),ncol=1)
x12<-x1**2
for (cl in 1:length(cluster_n)){
Blk<-cluster_n[cl]
cluster<-cutree(hclust(dist(x1),method="complete"),Blk)
xperm_yang<-rep(1:n)
Blksizelist<-vector("list",1)
for (i in 1:Blk) {
Blksizelist[length(Blksizelist)+1]<-sum(cluster==i)}
Blksizelist[1]<-NULL
Blksize<-unlist(Blksizelist)

C_yang<- vector("list",1)
N<-0
for (i in 1:length(Blksize)){
    if(i==1) {N=Blksize[i]}
    if(i>1){N=N+Blksize[i]}
    start<-N-Blksize[i]+1
    C_yang[[length(C_yang)+1]]<-xperm_yang[start:N]
   }
```

```
C_yang[1]<- NULL
 Cmat_yang<- CZmat(C_yang,n)
 Xmat_yang<-CXmat(C_yang,x1,n)

X_yang<-cbind(onesn,x1)

 ###########################
 ## Get XW matrix
 ###########################
 XW0<-lapply(as.list(1:length(Blksize)),function(i,Cmat_yang,Xmat_yang)
           {
                 C1<-matrix(Cmat_yang[,i],ncol=1)
                 X1<-matrix(Xmat_yang[,i],ncol=1)
                 X12<-X1**2
                 Wi<-cbind(C1,X1,X12)
                  return(Wi)},Cmat_yang,Xmat_yang)

XW1<-matrix(unlist(XW0),n, 3*length(Blksize))
XW<-cbind(X_yang,XW1)

 ytrue<-1+theta*x12
YTRUE<-matrix(ytrue,n,byrow=T)

 nreject_yang<-matrix(100)

    for (k in 1:ysim){
         error_yang<- matrix(unlist(error[k,]),nrow=n)
        Y_yang<- YTRUE+error_yang
        ############
        #   get ssex and dfssex
        #############
    dfssex<-basis( diag(n)-ppo(X_yang))
    ssex<-t(Y_yang)%*%(diag(n)-ppo(X_yang))%*%Y_yang

    dfssexw<-basis(diag(n)-ppo(XW))
    ssexw<-t(Y_yang)%*%(diag(n)-ppo(XW))%*%Y_yang

    ssnum<-ssex-ssexw
    dfnum<-dfssex-dfssexw
    msnum<-ssnum/dfnum
    ssden<-ssexw
    dfden<-dfssexw
    msden<-ssden/dfden
    F<-msnum/msden
    Fc<-qf(1-alpha,dfnum,dfden)
    nreject_yang[k]<-F>Fc
   }

# ******* get test power

simpower_yang<-mean(nreject_yang)
print(simpower_yang)
}
}
```

Appendix A-3 R code for the Six Simulation Study

```
rm(list=ls())
source("ppo.s")
source("basis.s")
source("CZmat.s")
source("quantnorep.s")
source("CXmat.s")
options(echo=F)

n<-80
ysim<- 2000
ansim<- 10000
alpha<- .05
stdev<-1
MIN<-0
MAX<-10

cluster_n<-c(5,6,7,8,9,10,11,12,13,14,15,16)
Nsize<- c(2,3,4,5)
#thetaL<-c(0,.5,1.0,1.5,2.0,2.5,3.2)
 thetaL<-c(0,0.5,.6,.7,.8)
onesn<- matrix(rep(1,n),byrow=F)

x1<-matrix(sort(runif(n,MIN,MAX)),ncol=1)
x20<-matrix(rep(seq(MIN,MAX,by=(MAX-MIN)/15),length=n),ncol=1)
x2<-cos(x20)
#x2<-sin(4*x20)
#x1<-matrix(sort(runif(n,MIN,MAX)),ncol=1)
#x20<-matrix(rep(seq(MIN,MAX,by=(MAX-MIN)/23),length=n),ncol=1)
#x2<-x20

error<- matrix(rnorm(n*ysim,mean=0,sd=stdev),ysim,n,byrow=T)
for (L in 1:length(thetaL)){                  #calculate for different theta
      theta<-thetaL[L]
      #cat("theta is \n")
      lof1<-5+3*x1+theta*x2
     ################################################
```

```
# The following code used for Neill's method
##################################################
 LOF1<- matrix(lof1,ysim,n,byrow=T)
 X_N<- cbind(onesn,x1)
 xperm_N<-order(x1)
 Y1<- LOF1+error
 Clist<- lapply(as.list(1:length(Nsize)),function(i,xperm_N,Nsize)
        {Ci<- vector("list",1)
          N<- Nsize[i]
          Blk_N<- floor(length(xperm_N)/N)
          start1<- 1

          if(1<=Blk_N)
              {for(j in 1:Blk_N)
                 {mj<- ((j-1)*N)+start1
                  Ci[[length(Ci)+1]]<- xperm_N[mj:(mj+N-1)]}
              }
         resid<- length(xperm_N)-(Blk_N*N)
         if(resid>0)
        #exclude singletons    if(resid>1)
              {start2<- (Blk_N*N)+start1
               Ci[[length(Ci)+1]]<- xperm_N[start2:length(xperm_N)]}
         Ci[[1]]<- NULL
         return(Ci)
          },xperm_N,Nsize)

NC<- length(Clist)

Cmatlist<- lapply(Clist,function(C,n)
      {Cmat<- CZmat(C,n)
       return(Cmat)
```

```
        },n)
MBlist<- lapply(Cmatlist,function(Cmat,X_N)
       {MB<- ppo(Cmat)-ppo(ppo(Cmat)%*%X_N)
        return(MB)
        },X_N)
dfBlist<- lapply(MBlist,function(MB)
      {dfB<- basis(MB)
       return(dfB)
       })
MXperp<- diag(n)-ppo(X_N)
dimCXperp<- basis(MXperp)

MWSlist<- lapply(MBlist,function(MB,MXperp)
      {MWS<- MXperp-MB
        return(MWS)
       },MXperp)

dfWSlist<- lapply(as.list(1:NC),function(j,dfBlist,dimCXperp)
     {dfWS<- dimCXperp - dfBlist[[j]]


      return(dfWS)
      },dfBlist,dimCXperp)

an<- quantnorep(n,ansim,NC,alpha,MBlist,MWSlist,dfBlist,dfWSlist)

Tvals1<- apply(Y1,1,function(y,NC,MBlist,MWSlist,dfBlist,dfWSlist,an)
        {
          bestcomp<- lapply(as.list(1:NC),function(j,y,an,MBlist,MWSlist,dfBlist,dfWSlist)
                {
                  MSB<- (sum((MBlist[[j]]%*%y)^2))/dfBlist[[j]]
                  MSWS<- (sum((MWSlist[[j]]%*%y)^2))/dfWSlist[[j]]
```

```
                 FB<- MSB/MSWS
               FWS<- 1/FB

               Bcpt<- qf(1-an,dfBlist[[j]],dfWSlist[[j]])
              WScpt<- qf(1-an,dfWSlist[[j]],dfBlist[[j]])

              diffB<- FB-as.numeric(Bcpt)
              diffWS<- FWS-as.numeric(WScpt)

              best<- max(diffB,diffWS)
              return(best)
             },y,an,MBlist,MWSlist,dfBlist,dfWSlist)

       Tval1<- max(unlist(bestcomp))
       return(Tval1)
      },NC,MBlist,MWSlist,dfBlist,dfWSlist,an)

nreject1_N<- Tvals1 > 0
simpower1_N<- mean(nreject1_N)
cat("T1 simulated power based on Neill's test  is \n")
print(simpower1_N)

#cat("T2 simulated power\n")
#print(simpower2)
#Tstats<- summary(Tvals)
#cat("T statistic summary\n")
#print(Tstats)
##################################
## The following code is used to simulate for yang's overall test
## use the above x1, x2 and error matrix
## for the same theta, try different number of clusters
```

```
####################################
cat(" Simpower based on Yang's test is \n")
cat("the numbers of clusters for Yang's method are\n")
print(cluster_n)
x1_Y<-x1                                        # x1 used for Yang's method
x12_Y<-x1_Y**2
Y_Y<-lof1                                      # Y used for Yang's method
for (cl in 1:length(cluster_n)){
   Blk_Y<-cluster_n[cl]
     #partition the dataset x1 into Blk_Y clusters
     cluster<-cutree(hclust(dist(x1),method="complete"),Blk_Y)

 ##############################
 # predetermine the size for each cluster
 #    and get  Z matrix and corresponding Xmatrix, they have the same rows and collumns
 ##############################
 xperm_Y<-rep(1:n)
 Blksizelist<-vector("list",1)
 for (i in 1:Blk_Y) {
    Blksizelist[length(Blksizelist)+1]<-sum(cluster==i)}
    Blksizelist[1]<-NULL
    Blksize<-unlist(Blksizelist)
    C_Y<- vector("list",1)
    N_Y<-0
    for (i in 1:length(Blksize)){
       if(i==1) {N_Y=Blksize[i]}
       if(i>1){N_Y=N_Y+Blksize[i]}
      start<-N_Y-Blksize[i]+1
      C_Y[[length(C_Y)+1]]<-xperm_Y[start:N_Y]
    }
    C_Y[1]<- NULL
    Cmat_Y<- CZmat(C_Y,n)
    Xmat_Y<-CXmat(C_Y,x1_Y,n)

    X_Y<-cbind(onesn,x1_Y)

    ###########################
    ## Get XW matrix
    ###########################
    XW0<-lapply(as.list(1:length(Blksize)),function(i,Cmat_Y,Xmat_Y)
       {
            C1_Y<-matrix(Cmat_Y[,i],ncol=1)
            X1_Y<-matrix(Xmat_Y[,i],ncol=1)
            X12_Y<-X1_Y**2
            Wi<-cbind(C1_Y,X1_Y,X12_Y)
             return(Wi)},Cmat_Y,Xmat_Y)

    XW1<-matrix(unlist(XW0),n, 3*length(Blksize))
    XW<-cbind(X_Y,XW1)
##################################
     # Use loop to calculate the nubmer of detecting the lack of fit
     # in ysim times
     ################################
     nreject_Y<-matrix(100)
   for (k in 1:ysim){
       error_Y<- matrix(unlist(error[k,]),nrow=n)
      Y_Ytrue<- Y_Y+error_Y
      ############
      #   get ssex and dfssex
      #############
     dfssex<-basis( diag(n)-ppo(X_Y))
     ssex<-t(Y_Ytrue)%*%(diag(n)-ppo(X_Y))%*%Y_Ytrue
```

```
        dfssexw<-basis(diag(n)-ppo(XW))
        ssexw<-t(Y_Ytrue)%*%(diag(n)-ppo(XW))%*%Y_Ytrue

        ssnum<-ssex-ssexw
        dfnum<-dfssex-dfssexw
        msnum<-ssnum/dfnum
        ssden<-ssexw
        dfden<-dfssexw
        msden<-ssden/dfden
        F<-msnum/msden
        Fc<-qf(1-alpha,dfnum,dfden)
        nreject_Y[k]<-F>Fc
       }
# ############### get test power
simpower_Y<-mean(nreject_Y)
#cat(" SSnum is\n ")
#print(ssnum)
#cat("SSden is\n")
#print(ssden)
#cat("dfnum is \n")
#print(dfnum)
#cat("dfden is \n")
#print(dfden)
#cat("F is \n")
#print(F)
print(simpower_Y)
}}
```

Appendix A-4 R code for the common function CXmat

```
CXmat<- function(cl,x,n)
       {CX<- matrix(rep(0,n*length(cl)),ncol=length(cl))

        for(j in 1:length(cl))
               {cverts<- matrix(rep(0,n),ncol=1)
                vertj<- cl[[j]]
                       for(k in 1:length(vertj))
                               {cverts[vertj[k],]<- x1[vertj[k]]}
                CX[,j]<- cverts}
        return(CX)}
```

Appendix A-5 R code for the common function CZmat

```
CZmat<- function(cl,n)
       {CZ<- matrix(rep(0,n*length(cl)),ncol=length(cl))

        for(j in 1:length(cl))
               {cverts<- matrix(rep(0,n),ncol=1)
                vertj<- cl[[j]]
                       for(k in 1:length(vertj))
                               {cverts[vertj[k],]<- 1}
                CZ[,j]<- cverts}
        return(CZ)}
```

Appendix A-6 R code for the common function basis

```
basis<- function(A)
{
 B<- A%*%t(A)
 m<- nrow(B)
 e<- eigen(B,symmetric=T)
 vals<- e$values
 vals<- unlist(vals)
 #cat("eigenvalues\n")
 #print(vals)
 vecs<- e$vectors
 nzvals<- vals[abs(vals)>1.0e-6]
 rankA<- length(nzvals)
 #basis<- vecs[1:m,1:k]
 #return(basis)
 return(rankA)
}
```

Appendix A-7 R code for the common function ppo

```
require(MASS)
ppo<- function(C)
{
 M<- C%*%ginv(t(C)%*%C)%*%t(C)
 return(M)
}
```

Appendix A-8 R code for the common function quantnorep

```
quantnorep<- function(n,ansim,NC,alpha,Mnumlist,Mdenlist,dfnumlist,dfdenlist)
        {
        error<- matrix(rnorm(n*ansim,mean=0,sd=1),ansim,n,byrow=T)
         infs<- apply(error,1,function(e,NC,Mnumlist,Mdenlist,dfnumlist,dfdenlist)
                 {
           Fpvals<- lapply(as.list(1:NC),function(j,e,Mnumlist,Mdenlist,dfnumlist,dfdenlist)
                       {
                        MSnum<- (sum((Mnumlist[[j]]%*%e)^2))/dfnumlist[[j]]
                        MSden<- (sum((Mdenlist[[j]]%*%e)^2))/dfdenlist[[j]]
```

```
            F<- MSnum/MSden
            Fpval<- 1-pf(F,dfnumlist[[j]],dfdenlist[[j]])
            return(Fpval)
                   },e,Mnumlist,Mdenlist,dfnumlist,dfdenlist)
  Fcompvals<- 1-unlist(Fpvals)
  pvals<- c(unlist(Fpvals),Fcompvals)

  inf<- min(pvals)
  return(inf)
        },NC,Mnumlist,Mdenlist,dfnumlist,dfdenlist)
an<- quantile(infs,alpha)
return(an)
     }
```