THE EGYPTIAN STATISTICAL JOURNAL
ISSR, CAIRO UNIV., VOL., 43, NO.2, 1999

An Exact Procedure for Testing for Lack-of-Fit Without Replicates

ELIES KOUIDER and HANFENG CHEN
United Arab Emirates University and Bowling Green State University

SUMMARY

There is a number of testing Lack-of-Fit procedures available in the literature, each being proposed to deal with a particular situation. While these existing tests might do a reasonable job, their disadvantages and drawbacks have also been acknowledged. In this paper, we propose an exact F-test that applies to any situation even if replicates or near replicates are not present. Its implementation is easy, and found to be extremely powerful compared with other tests. Its use is also developed when multidimensional covariates occur.

KEY WORDS: Regression, Lack-of-Fit, replicates, power.

1. INTRODUCTION

In the context of regression analysis, violation of the linear structure for the mean response may result in misleading or wrong conclusions.

In some instances, a straight line should not have been fitted because the response function may be quadratic in X or any other non-linear function in X. Suppose that we want to detect if lack-of-fit occurred. A graphical analysis of the residuals may reveal information regarding the potential violation of assumptions than any other techniques. Draper and Smith (1981) and Graybill (1976) discuss graphical approaches. The disadvantage of a graphical approach is also obvious. First, it is subject to visualizing and different observers may draw different conclusions on the same graph. Second, graphics become cumbersome or even impossible when multidimensional covariates occur. Therefore, a formal statistical inference about lack-of-fit is desirable and necessary sometimes.

A well-recognized and accepted method for testing lack-of-fit to linear regression in the case that exact replicates are present is the classical lack-of-fit test. This test is only applicable to exact replicates. Consequently, an experimenter is always advised to consider exact replicates in his design. Sometimes though, an experimenter cannot produce exact replicates even when he is establishing the design, for example in chemical kinetic experiments, or some historical data that cannot be reproduced. In these cases the researcher is faced with a harder problem to manage, since replications are not possible.

A lot of work has been done on near replicates mimicking the classical case. Neill and Johnson (1984) presented an informative and lengthy review. In particular, the works by Green (1971), Lyons and Proctor (1977), Shillington (1979), Daniel and Wood (1980), Draper and smith (1981) and Utts (1982) were cited. Recently, Neill and Johnson (1985, 1989), Christensen (1989, 1991), and Joglekar, Schuenemeyer and LaRiccia (1989) developed procedures in the case of no replicates but under existence of near replicates. The power of these tests is highly dependent on the choice of clusters.

Su and Wei (1991) proposed a new test using a supremum-type statistic, based on partial sums of residuals. The proposed test does not need a partition of the space of covariates to handle the case without replications. This test has good power and was shown by simulations to be even more powerful than the classical test in the case that replicates are present. The test though has a drawback if the hypothesized link function and the true link function intersect too often in the covariate space. Also for cases with high dimensional covariates and large sample sizes, a large computation time is required to calculate the observed test statistic value and the critical values as pointed out recently by Cheng and Wu (1994). Cheng and Wu in their paper, proposed a new test which seems to be as good as Su and Wei's test and their test needs less computations as claimed in their paper.

We suggest in this article a natural idea to detect lack-of-fit which is done by comparing the overall linear fit with piecewise linear fits. First we split the covariate space into two or more portions. We then calculate the piecewise linear fits. The exact lack-of-fit test is to test the equality of the coefficients of the piecewise linear fits. The advantage of this test for revealing occurrence of any lack-of-fit or functional misspecification associated with the deterministic portion of a proposed linear regression model is that it does not rely on a grouping method and does not require near replicates. This test is simple, easy to understand and has good power. Simulation reveals that it is even more powerful than the classical test statistic in the

case of exact replicates. In particular, we find that the new test is rather sensitive to detect a non-linear regression. Moreover this test is an exact F-test, which is an advantage since one uses the exact critical values obtained from the F-table. The test requires partitioning the covariate space. The issue of partitioning is discussed in Section 5.5

2. CLASSICAL LACK-OF-FIT TEST

An ideal experimental situation is that there are replicating observations available for analysis, i.e., at some sampling points, more than one response are independently obtained. In this case, lack-of-fit to a linear regression function can be easily tested in the light of analysis of variances. A linear regression model with replicates can be written as

$$Y_{ij} = \sum_{k=1}^{p} X_{ik}\beta_k + \varepsilon_{ij}, \qquad (2.1)$$

where i = 1,...,M, j = 1,...,n_i and $n_i > 1$for at least one i. Here n_i is the number of replicates at the covariate level (X_{i1},...,X_{ip}). For example, if two independent responses are collected at the same covariate vector (X_{11},...,X_{1p}), then $n_1 = 2$. Let the total number of observations be denoted by $N = \sum_{i=1}^{M} n_i$. For each fixed i, Y_{i1},...,Y_{in_i}, are n_i independent responses corresponding to the covariates X_{ik}, k = 1,...,p, and $\beta'=(\beta_1,...,\beta_p)$ is an unknown parameter vector. The random errors Y_{ij} are assumed to be independently, identically, and normally distributed with mean zero and unknown variance σ^2. Throughout this dissertation, it is assumed that M > p.

The classical lack-of-fit test compares model (2.1) against general alternatives

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad (2.2)$$

Where, i = 1,...,M, j = 1,...,n_i, and μ_i are any real numbers, but unknown. Here ε_{ij} are assumed the same as in the model (2.1). Given X, model (2.2) is reduced to model (2.1) when μ_i is chosen to be $\mu_i = \sum_k X_{ik}\beta_k$, i = 1,...,M.

Under model (2.2), σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{1}{N-M}\sum_{i=1}^{M}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i\cdot})^2 ; \qquad (2.3)$$

and under model (2.1), σ^2 can be estimated by

$$\hat{\sigma}_0^2 = \frac{1}{N-p}\sum_{i=1}^{M}\sum_{j=1}^{n_i}(Y_{ij} - \sum_{k=1}^{p} X_{ik}\hat{\beta}_k)^2 .$$

Here $\overline{Y_{i\cdot}} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ and $\hat{\beta}_k$ is the least squares estimate of β_k under model (2.1), k=1,...,p. Then the classical lack-of-fit test is an F-test with the following test statistic. Let SSR = (N - p)$\hat{\sigma}_0^2$ and SSE = (N - M)$\hat{\sigma}^2$, and define

$$F_c = \frac{(SSR - SSE)/(M-p)}{SSE/(N-M)} \qquad (2.4)$$

When model (2.1) is true, F_c follows an F-distribution with (M - p, N - M) degrees of freedom. Therefore, the F-test rejects the adequacy of model (2.1) at level

α in favor of model (2.2), whenever $F_c > F(\alpha; M-p, N-M)$, where $F(\alpha; M-p, N-M)$ is the $100(1-\alpha)\%$ percentile of the F-distribution with (M-p, N-M) degrees of freedom.

3. MODEL

Consider a full rank linear regression model

$$Y = X\beta + \varepsilon, \tag{3.1}$$

where $Y' = (Y_1,...,Y_n)$ is an n dimensional observable random vector and ε is a corresponding n dimensional unobservable random error vector. ε is assumed to be $N(0, \sigma^2 I)$ distributed, where 0 is an $n\times 1$ zero vector, I is an $n\times n$ identity matrix, and σ^2 is an unknown variance parameter. $\beta' = (\beta_0, \beta_1,..., \beta_{p-1})$ is a p dimensional vector of unknown parameters defined in $\Re^p$, and X is an $n\times p$ matrix with rank p.

A lack-of-fit test is to test model (2.1) against the general alternative model

$$Y = \mu(X) + \varepsilon, \tag{3.2}$$

where Y and ε are as defined before and μ is any function of X.

4. MOTIVATION

Our aim is to find a procedure that detects non-linearity of the mean function, i.e., to test

$$H_0: \mu(X) = X\beta$$
$$\text{versus} \tag{4.1}$$
$$H_a: \mu(X) \neq X\beta.$$

For simplicity let us consider the case of p = 2. Our procedure is to partition the covariate space into two portions. At the moment, let us assume that the two (m=2) portions are cut off at the midpoint of the range of X values or just near it. Suppose that the first group is of size n_1 and the second group is of size $n_2 = n - n_1$. Then model (3.1) is split as

$$Y_i = X_i\beta_i + \varepsilon_i, \; i = 1,2, \tag{4.2}$$

Where Y_i is $n_i\times 1$ vector, X_i is $n_i\times p$ matrix of rank p, β_i is $p\times 1$ vector, and ε_1, ε_2 are independently and normally distributed as $N(0, \sigma^2 I)$. If the mean response were linear in X, then the mean response of each portion would also be linear with coefficients, $\beta_1 = \beta_2$. On the other hand, if $\beta_1 \neq \beta_2$ then the mean response must be nonlinear over the whole range of X.

5. DEVELOPMENT OF THE TEST STATISTIC AND MAIN SESULTS

5.1 Development of the Test

To determine whether or not the two fitted lines are approximately the same, one would reduce the problem to testing

$$H_0: \beta_1 = \beta_2$$
$$\text{versus} \tag{5.1.1}$$
$$H_a: \beta_1 \neq \beta_2.$$

Although this model is not the same as the one described in (4.1), we can notice from the motivation and the development of the procedure that it is quite close to that.

To test (4.2) one would apply the least squares theory to the two (m=2) portions to get estimates of β_i:

$$\hat{\beta}_i = (X_i'X_i)^{-1}X_i'Y_i, \text{ for i = 1,2.} \qquad (5.1.2)$$

Under the null hypothesis in (4.1),

$$\hat{\beta}_1 \sim N(\beta, \sigma^2(X_1'X_1)^{-1})$$

$$\hat{\beta}_2 \sim N(\beta, \sigma^2(X_2'X_2)^{-1}),$$

and $\hat{\beta}_1$ is independent of $\hat{\beta}_2$ since Y_1 is independent of Y_2. It follows that under the null hypothesis,

$$\hat{\beta}_1 - \hat{\beta}_2 \sim N(\beta, \sigma^2[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]),$$

and so

$$[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]^{-\frac{1}{2}}(\hat{\beta}_1 - \hat{\beta}_2)/\sigma \sim N(0, I)$$

Thus under the null hypothesis,

$$(\hat{\beta}_1 - \hat{\beta}_2)'[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]^{-1}(\hat{\beta}_1 - \hat{\beta}_2)/\sigma^2 \sim \chi_p^2,$$

where χ_p^2 is the chi-square distribution with p degrees of freedom. Since σ^2 is unknown, we estimate it by the pooled sample variance

$$\hat{\sigma^2} = [S_1^2 + S_2^2]/(n-2p),$$

where

$$S_i^2 = Y_i'(I - X_i(X_i'X_i)^{-1}X_i')Y_i, \text{ i = 1,2.} \qquad (5.1.3)$$

Since

$$S_i^2/\sigma^2 \sim \chi^2(n_i - p), \text{ i = 1,2}$$

and clearly S_1^2 and S_2^2 are independent, we have

$$(n-2p)\hat{\sigma^2}/\sigma^2 \sim \chi^2(n-2p).$$

To establish an F-test, we need the independence of

$$[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]^{-\frac{1}{2}}(\hat{\beta}_1 - \hat{\beta}_2) \text{ and } S_1^2 + S_2^2.$$

5.2 Lemma

Under model (2.1) with the partition of (4.2),

$[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]^{-\frac{1}{2}}(\hat{\beta}_1 - \hat{\beta}_2)$ *and* $S_1^2 + S_2^2$ *are independent, where* $\hat{\beta}_i$ *and* S_i^2 *are defined as in (5.1.2) and (5.1.3), respectively.*

The exact F-test statistic is defined as

$$F_{ex} = \frac{(\hat{\beta}_1 - \hat{\beta}_2)'[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}]^{-1}(\hat{\beta}_1 - \hat{\beta}_2)/p}{(S_1^2 + S_2^2)/(n-2p)}$$

5.3 Theorem

When the null hypothesis in (2.1) is true, F_{ex} *is distributed as an F-distribution with* (p, n-2p) *degrees of freedom.*

By the theorem, we would reject the null hypothesis at level α whenever $F_{ex}>F(\alpha; p, n - 2p)$.

5.4 Generalizations and Matrix Notation

In general, model (5.1.1) can be expressed in a linear model by matrix notation. This expression will be convenient when we extend the test to a general consideration of partitions. Assume that we partition the observations into m portions according to the covariate values. Thus, we will be testing,

$$H_0: C\beta=0$$

versus

$$H_a: C\beta\neq 0.$$

The exact test for testing for lack-of-fit can be written as:

$$F_{ex}=\frac{Y'X(X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}C(X'X)^{-1}X'Y/[p(m-1)]}{Y'[I-X(X'X)^{-1}X']Y/(n-mp)},$$

where $\beta'=[\beta_{10},\beta_{11},\beta_{20},\beta_{21},\ldots,\beta_{m0},\beta_{m1}]$, and

$C=\begin{bmatrix}1&0&-1&0\\0&1&0&-1\end{bmatrix}$, when m=2, and $C=\begin{bmatrix}1&0&-1&0&0&0\\1&0&0&0&-1&0\\0&1&0&-1&0&0\\0&1&0&0&0&-1\end{bmatrix}$, when m=3,

and so on.

The vector of observations can be written as, $Y=\begin{bmatrix}Y_1\\\vdots\\Y_{n_1}\\\hdashline\vdots\\\hdashline Y_{n_1+\ldots+n_{m-1}+1}\\\vdots\\Y_{n_1+\ldots+n_m}\end{bmatrix}$,

and the X matrix, as $X=\begin{bmatrix}1&X_1&\cdots&0&0\\\vdots&\vdots&\ddots&\vdots&\vdots\\1&X_{n_1}&\cdots&0&0\\\hdashline\vdots&\vdots&\ddots&\vdots&\vdots\\\hdashline0&\cdots&0&1&X_{n_1+\ldots+n_{m-1}+1}\\\vdots&\ddots&\vdots&\vdots&\vdots\\0&\cdots&0&1&X_{n_1+\ldots+n_m}\end{bmatrix}$.

The proposed test rejects the adequacy of model (2.1) at level α in favor of model (2.2), whenever $F_{ex} > F(\alpha; p(m-1), n-mp)$, where $F(\alpha; p(m-1), n-mp)$ is the $100(1-\alpha)\%$ percentile of the F-distribution with (p(m-1), n-mp) degrees of freedom.

It is clear that the number of degrees of freedom in the denominator is decreasing by mp (multiple of p). Therefore, an increase in the number of portions reduces the power of the test at certain situations, say at a quadratic non-linear alternative. From this point of view, we suggest to have only at most two or three portions. In the course of simulation study not reported here, it shows that with two portions, the testing powers of quadratic and cubic alternatives are higher than with three. Of course, fewer alternatives can be detected when smaller number of portions is used instead. This is really a trade-off that an experimenter should make.

5.5 Partitioning the Covariate Space

When p = 2 we have a one-dimensional covariate space. To partition the covariate space, we would first arrange the X covariates in an increasing order, then the first n_1 of the X's sorted consist the first portion and the next $n_2 = n - n_1$ of the X's the second portion. It is noted that the partition is a matter of X values, so it does not affect the independence of the responses.

When p > 2, we have a multidimensional covariate space. Partitioning the covariate space becomes cumbersome, since we do not have a natural method to sort X vectors. For illustration and without loss of generality, let us consider the p = 3 case. In this case we have a two-dimensional covariate space.

One possible method to split the data is to fit a convex hall in the shape of a hyper ellipsoid. Find its major axis, and project each point perpendicularly on that axis. Then partitioning is implemented by these projections that are one-dimensional. Arrange these projections into an increasing order and the first portion consists of the n_1 points corresponding to the first n_1 values of the projections, and so on.

Using results from linear algebra and multivariate statistical analysis we can describe this method as follows. Let k=p-1 and write

$$V = \begin{bmatrix} v_{11} & \cdots & v_{k1} \\ \vdots & \ddots & \vdots \\ v_{1k} & \cdots & v_{kk} \end{bmatrix}, \text{ where } v_{ij} = \frac{1}{n-1}\sum_{l=1}^{n}(X_{il} - \overline{X_i})(X_{jl} - \overline{X_j}), \text{ with } \overline{X_t} = \frac{1}{n}\sum_{l=1}^{n} X_{lt}.$$

The matrix V comprises the information about the variance and covariance of X_{ij} values, when X_{ij}'s are deemed randomly designed.

Note that V is symmetric and non-negative definite, and so has non-negative eigen values. Suppose $\lambda_1, \ldots, \lambda_k$ are the eigen values of V. Let $\lambda^* = \max_{1 \le i \le k} \lambda_i$, and let e^* be its corresponding eigen vector with length one, called the major axis. The projection of covariate vector X_i along e^* is given by $X_i' e^*$, i.e., the inner product of X_i and e^*. Then we sort these projections, since those inner products are just values on a real line. Afterwards, the first partition is realized by picking n_1 points from the covariate space corresponding to the n_1 first sorted projections, and so on.

5.6 Remark

A partition of covariate values to construct the exact test F_{ex} test has been made free of the responses, so free of random errors. From a practical point of view, F_{ex} seems to be easy, as it is inevitable for one to refer to a scatter plot to make a preferred partition, especially in one-dimensional case where people often rely on a

scatter plot to make a diagnostic modeling analysis. For example, if the scatter plot shows a rainbow pattern with the vertex appearing around the lower quartile of the X values (assume dimension is one), then one may tend to split at the lower quartile and have a one third versus two thirds partition. As for this aspect, we do not criticize such a practice, but hope that practitioners keep in mind that the conclusions drawn from this test based on data-related partition of covariate values are conditional and the interpretations of the conclusions should be confined and applicable to the partition used.

6. DESCRIPTION OF SIMULATION AND COMPARISON OF F_{ex} AND F_c

Simulations are carried out to compare the powers of the exact lack-of-fit test and the classical lack-of-fit test. For the purpose of comparison, the same Monte Carlo samples were utilized for the model used in the simulation. In each comparison, 10,000 Monte Carlo trials were performed, so the bound on the estimated errors of simulated powers is 1%. The nominal significance level for all comparisons was 5%. All Monte Carlo data were generated through the RNLNL subroutine in the IMSL.

Simulation results are tabulated and put in the appendix. Entries in the table are the simulated testing powers of the suggested lack-of-fit test F_{ex} and the classical test F_c. The null hypothesis is $E(Y)=\beta_0+X$ and the alternative is $E(Y)=\beta_0+X+\beta_2X^2$. The standard deviation used is $\sigma=1$. The X values are -5 (1) 5 each being replicated three times. To perform the exact test F_{ex}, the covariate space was partitioned into, $n_1 = 15$ and $n_2 = 18$ for the first and the second portion respectively.

The exact test F_{ex} is compared with the classical test F_c only, because the classical test when exact replicates are present is accepted to be the most desirable test in practice. The suggested lack-of-fit test turns out to be twice as powerful as the classical test.

7. SUMMARY

The exact test was introduced and its F-distribution was derived. This test is applicable as showed under any circumstances, but can be deficient under certain alternatives. The test relies on splitting the covariate space into portions, and so its power depends on the portions chosen and their number. Simulation study was performed in the case of exact replicates to compare the powers of the suggested test with the classical lack-of-fit test. It was shown that the proposed test is more powerful than the classical lack-of-fit test.

8. APPENDIX

8.1 Proof of lemma

We use the result that if $Z \sim N(\mu,\sigma^2 I)$, then AZ and BZ are independent if and only if $A'B = 0$. Henceforth, $\hat{\beta}_i$ and S_i^2 are independent, for i=1,2. It is clear that $(\hat{\beta}_1, S_1^2)$ and $(\hat{\beta}_2, S_2^2)$ are independent since $(\hat{\beta}_i, S_i^2)$ depends on Y_i only, i = 1,2 and

Y_1 and Y_2 are independent. Thus, we conclude that $\hat{\beta}_1$, $\hat{\beta}_2$, S_1^2 and S_2^2 are independent. The lemma follows. ■

8.2 Proof of the theorem

Proof follows from the lemma ■

8.3 Table

β_2	Power of F_{ex}	Power of F_c
0.00	0.051	0.049
0.01	0.066	0.054
0.02	0.123	0.073
0.03	0.221	0.109
0.04	0.359	0.167
0.05	0.531	0.252
0.06	0.699	0.369
0.07	0.834	0.509
0.08	0.922	0.647
0.09	0.969	0.764
0.10	0.989	0.864
0.11	0.997	0.929
0.12	0.999	0.967
0.13	1.000	0.986
0.14	1.000	0.995
0.15	1.000	0.998
0.16	1.000	0.999
0.17	1.000	1.000

REFERENCES

Cheng, K. F. and Wu, J. W. (1994). Testing Goodness-of-Fit for a Parametric Family of Link Functions, *Journal of the American Statistical Association,* **89**, 657-664.

Christensen, R. (1989). Lack-of-Fit Tests Based on Near or Exact Replicates, *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,* **82**, 752-756.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis* (Wiley, New York, 2nd Ed.).

Daniel, C. and Wood, F.S. (1980). *Fitting Equations to Data* (Wiley, New York, 2nd Ed.).

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

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

Joglekar, G., Schuenemeyer, J. H., and LaRiccia, V. (1989). Lack-of-Fit Testing When Replicates Are Not Available, *The American Statistician,* **43**, 135-143.
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.
Neill, J. W. and Johnson, D. E. (1984). Testing for Lack-of-Fit in Regression—A Review, *Communications in Statistics, Part A—Theory Methods,* **13**, 485-511.
Neill, J. W. and Johnson, D. E. (1985). Testing Linear Regression Function Adequacy Without Replication, *The Annals of Statistics,* **13**, 1482-1489.
Neill, J. W. and Johnson, D. E. (1989). A Comparison of Some Lack-of-Fit Tests Based on Near Replicates, *Communications in Statistics, Part A—Theory Methods,* **18**, 3533-3570.
Shillington, E. R. (1979). Testing Lack-of-Fit in Regression Without Replication, *The Canadian Journal of Statistics,* 7, 137-146.
Su, J. Q. and Wei, L. J. (1991). A Lack-of-Fit Test for the Mean Function in a Generalized Linear Model, *Journal of the American Association,* **86**, 420-426.
Utts, J. M. (1982). The Rainbow Test for Lack-of-Fit in Regression, *Communications in Statistics,* Part A—Theory Methods, **11,** 2801-2815.

Department of Statistics
United Arab Emirates University
P.O. Box 17555
Al-Ain, U.A.E.
eliesk@uaeu.ac.ae

Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, OH 43403-0221, U.S.A.
hchen@math.bgsu.edu