Internat. J. Math. & Math. Sci.
VOL. 17 NO. 1 (1994) 173-180

TESTING THE STABILITY OF REGRESSION PARAMETERS WHEN SOME ADDITIONAL DATA SETS ARE AVAILABLE

R. RADHAKRISHNAN
DON R. ROBINSON

Department of Management and Quantitative Methods
Illinois State University
Normal, Illinois 61761

(Received June 2, 1992 and in revised form September 8, 1992)

ABSTRACT. We consider the problem of testing the stability of regression parameters in regression lines of different populations when some additional, but unidentified, data sets from those populations are available. The standard test (T_0) discards the additional data and tests the stability of the regression parameters using only the data sets from identified populations. We propose two test procedures (T_1 and T_2) utilizing all the available data, because the additional data may contain information about the parameters of the regression lines which are tested for stability. A power comparison among the tests is also presented. It is shown that T_1 always has larger power than T_0. In certain situations T_2 has the largest power.

KEY WORDS AND PHRASES: least squares estimates, regression parameters, power of the test.
1992 AMS SUBJECT CLASSIFICATION CODES: 62J05, 62J99

1. INTRODUCTION.

Consider the regression model

$$y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_i) + \epsilon_{ij}, \quad i=1,2,...,k, \quad j=1,2,...,n_i, \tag{1.1}$$

where the y_{ij} are observations on the response variable, the x_{ij} are observations on the predictor variable, α_i and β_i are the regression parameters, and the ϵ_{ij} are the error terms, which are unobserved random variables. It is assumed that the errors are independent, normally distributed random variables with mean 0 and common unknown variance σ^2. For the model, $\alpha_i + \beta_i(x_{ij} - \bar{x}_i)$ is the regression line of the variable y on the predictor variable x for the i^{th} group, α_i is the y-intercept when $x = \bar{x}_i$, and β_i is the slope. Suppose we have m ($m \leq k$) additional data sets corresponding to m regression lines whose model is given by

$$y_{ij} = \alpha_i + \beta_i(x_{ij} - \bar{x}_i) + \epsilon_{ij}, \quad i=k+1,...,k+m, \quad j=1,2,...,n_i. \tag{1.2}$$

We assume that the error terms ϵ_{ij} in model (1.2) are independent, normally distributed random variables with mean 0 and common unknown variance σ^2. It is further assumed that the m regression lines in model (1.2) are an unknown subset of the k regression lines in model (1.1).

However, we cannot identify the m regression lines associated with the additional data sets (y_{ij}, x_{ij}, $i=k+1,...,k+m$, $j=1,2,...,n_i$). We are interested in testing the null hypothesis $H_0: \alpha_1 = \alpha_2 = ... = \alpha_k; \beta_1 = \beta_2 = ... = \beta_k$ against H_a: either $\alpha_i \neq \alpha_{i'}$ or $\beta_i \neq \beta_{i'}$ for at least one pair (i,i'), where $i \neq i'$, $i,i'=1,2,...,k$, utilizing all the available data. The null hypothesis implies that all the k regression lines in model (1.1) are coincident whereas the alternative hypothesis is that at least two of the regression lines are different. The standard test (T_0) of H_0 against H_a using the k data sets (y_{ij}, x_{ij}, $i=1,2,...,k$, $j=1,2,...,n_i$) is well-known in the literature and has diverse applications. A biostatistician may be interested in testing the equivalence of regression lines for predicting the systolic blood pressure using age as the predictor variable for four social groups. A test for the stability of the regression parameters that generated the data sets is $H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4; \beta_1 = \beta_2 = \beta_3 = \beta_4$. If H_0 is true, we use a single regression line based upon the four data sets for predicting systolic blood pressure using age as the predictor variable, Klienbaum and Kupper [1]. An economist might be interested in testing the equivalence of multiple regression models for predicting the gross domestic product using labor and capital as predictor variables for different time periods, Maddala [2].

In this paper we consider two tests (T_1 and T_2) utilizing all the available data and make a power comparison between these two tests and the standard test which is based solely on the k data sets relating to the regression lines whose parameters are tested for stability. In Section 2 we determine least squares estimates of the regression parameters to obtain the test statistics for the problem. The noncentrality parameter of the tests is derived in Section 3. In Section 4 we derive our proposed tests, T_1 and T_2. We illustrate and compare the power of all three tests in Section 5.

2. LEAST SQUARES ESTIMATES. Consider the sum

$$\phi_0 = \sum_{i=1}^{k+m} \sum_{j=1}^{n_i} (y_{ij} - \hat{\alpha}_i - \hat{\beta}_i(x_{ij} - \bar{x}_i))^2, \tag{2.1}$$

where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the estimates of regression parameters α_i and β_i ($i=1,2,...,k+m$). The least squares estimates of the regression parameters are obtained by differentiating ϕ_0 partially with respect to $\hat{\alpha}_i$ and $\hat{\beta}_i$ and then solving the resulting normal equations for $\hat{\alpha}_i$ and $\hat{\beta}_i$. It can be shown that the least squares estimates of α_i and β_i are given by

$$\hat{\alpha}_i = \sum_{j=1}^{n_i} y_{ij}/n_i \tag{2.2}$$

and

$$\hat{\beta}_i = \sum_{j=1}^{n_i} y_{ij}(x_{ij} - \bar{x}_i) \Big/ \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2. \tag{2.3}$$

Then $R_0'^2$, the unconditional error sum of squares, is obtained by substituting $\hat{\alpha}_i$ and $\hat{\beta}_i$ given by (2.2) and (2.3) into ϕ_0. It can be shown that

$$R_0'^2 = \sum_{i=1}^{k+m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 - \sum_{i=1}^{k+m} \hat{\beta}_i^2 S_i^2, \tag{2.4}$$

where

$$S_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \quad i = 1,2,...,k+m. \tag{2.5}$$

The conditional error sum of squares under H_0 is obtained by minimizing

$$\phi_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\alpha} - \hat{\beta}(x_{ij} - \bar{x}_i))^2 + \sum_{i=k+1}^{k+m} \sum_{j=1}^{n_i} (y_{ij} - \hat{\alpha}_i - \hat{\beta}_i(x_{ij} - \bar{x}_i))^2 \tag{2.6}$$

with respect to $\hat{\alpha}, \hat{\beta}, \hat{\alpha}_i$, and $\hat{\beta}_i$ $(i = k+1,...,k+m)$.

The second sum on the right-hand side of (2.6) is minimized with respect to $\hat{\alpha}_i$ and $\hat{\beta}_i$ $(i = k+1,...,k+m)$ where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are defined, respectively, as in equations (2.2) and (2.3). The least squares estimates of the regression parameters α and β are given by

$$\hat{\alpha} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}/n = \bar{y} \tag{2.7}$$

$$\hat{\beta} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}(x_{ij} - \bar{x}_i) / \sum_{i=1}^{k} S_i^2 \tag{2.8}$$

where

$$n = \sum_{i=1}^{k} n_i. \tag{2.9}$$

The conditional error sum of squares under H_0 is

$$R_1^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 - \hat{\beta}^2 \sum_{i=1}^{k} S_i^2 + \sum_{i=k+1}^{k+m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 - \sum_{i=k+m}^{k+m} \hat{\beta}_i^2 S_i^2. \tag{2.10}$$

The sum of squares for testing the null hypothesis H_0 is

$$\begin{aligned} SSH_0 &= R_1^2 - R_0'^2 \\ &= \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k} \hat{\beta}_i^2 S_i^2 - \hat{\beta}^2 \sum_{i=1}^{k} S_i^2, \\ &= \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k} S_i^2 (\hat{\beta}_i - \hat{\beta})^2, \end{aligned} \tag{2.11}$$

where

$$\hat{\beta} = \sum_{i=1}^{k} S_i^2 \hat{\beta}_i / \sum_{i=1}^{k} S_i^2. \tag{2.12}$$

It is well-known in the literature that $R_0'^2/\sigma^2$ is distributed as chi-square with

$$n' = n + \sum_{i=k+1}^{k+m} n_i - 2(k+m) \tag{2.13}$$

degrees of freedom and SSH_0/σ^2 is distributed as noncentral chi-square with 2(k-1) degrees of freedom. When H_0 is true, SSH_0/σ^2 is distributed as chi-square with 2(k-1) degrees of freedom. Further, $R_0'^2$ and SSH_0 are independent; for example see Kshirsagar [3].

3. NONCENTRALITY PARAMETER. Here we derive the expected value of SSH_0 under the non-null case. It can be shown that

$$\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} n_i(\alpha_i - \bar{\alpha})^2 + \sum_{i=1}^{k} n_i(\bar{\epsilon}_i - \bar{\epsilon})^2 + 2\sum_{i=1}^{k} n_i(\alpha_i - \bar{\alpha})(\bar{\epsilon}_i - \bar{\epsilon}), \tag{3.1}$$

where

$$\bar{\alpha} = \sum_{i=1}^{k} n_i\alpha_i/n, \tag{3.2}$$

$$\bar{\epsilon}_i = \sum_{i=1}^{k} \epsilon_{ij}/n_i, \tag{3.3}$$

$$\bar{\epsilon} = \sum_{i=1}^{k} n_i\bar{\epsilon}_i/n. \tag{3.4}$$

Taking expectations of both sides of (3.1) we obtain

$$E\left(\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2\right) = \sum_{i=1}^{k} n_i(\alpha_i - \bar{\alpha})^2 + (k-1)\sigma^2. \tag{3.5}$$

Now

$$E\left(\sum_{i=1}^{k} \beta_i^2 S_i^2\right) = \sum_{i=1}^{k} S_i^2(\sigma^2/S_i^2 + \beta_i^2)$$

$$= k\sigma^2 + \sum_{i=1}^{k} \beta_i^2 S_i^2, \tag{3.6}$$

and

$$E\left(\bar{\beta}^2 \sum_{i=1}^{k} S_i^2\right) = \sum_{i=1}^{k} S_i^2\left(\sigma^2/\sum_{i=1}^{k} S_i^2 + \bar{\beta}^2\right)$$

$$= \sigma^2 + \bar{\beta}^2 \sum_{i=1}^{k} S_i^2, \tag{3.7}$$

where

$$\bar{\beta} = \sum_{i=1}^{k} S_i^2\beta_i/\sum_{i=1}^{k} S_i^2. \tag{3.8}$$

Using equations (3.5), (3.6), (3.7), and (2.11) we get

$$E(SSH_0) = 2(k-1)\sigma^2 + \sum_{i=1}^{k} n_i(\alpha_i - \bar{\alpha})^2 + \sum_{i=1}^{k} S_i^2(\beta_i - \bar{\beta})^2. \tag{3.9}$$

Since SSH_0/σ^2 is distributed as noncentral chi-square with 2(k-1) degrees of freedom, it follows from (3.9) that the noncentrality parameter is given by

$$\lambda = (1/\sigma^2)[\sum_{i=1}^{k} n_i(\alpha_i - \bar{\alpha})^2 + \sum_{i=1}^{k} S_i^2(\beta_i - \bar{\beta})^2]. \tag{3.10}$$

4. TEST PROCEDURES. The standard procedure for testing H_0 against H_a is an F-test based upon the test statistic

$$F_0 = (SSH_0/2(k-1))/(R_0^2/(n-2k)), \tag{4.1}$$

where

$$R_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 - \sum_{i=1}^{k} \hat{\beta}_i^2 S_i^2. \tag{4.2}$$

See, for example, Kshirsagar [3]. The above test rejects the null hypothesis H_0 if $F_0 > F_{\alpha,2(k-1),n-2k}$ and accepts H_0 otherwise, where F_{α,f_1,f_2} is the upper 100α percentile point of the F-distribution with f_1 numerator degrees of freedom (ndf) and f_2 denominator degrees of freedom (ddf). We note that the standard test is based upon the k data sets $(y_{ij}, x_{ij}, i=1,2,...,k, j=1,2,...,n_i)$ and discards the additional data $(y_{ij}, x_{ij}, i=k+1,...,k+m, j=1,2,...,n_i)$.

Consider the following test procedure (T_1). Reject H_0 if

$$F_1 = (SSH_0/2(k-1))/(R_0'^2/n') > F_{\alpha,2(k-1),n'} \tag{4.3}$$

and accept H_0 otherwise. A comparison between T_0 and T_1 shows that both have the same ndf but that the latter has larger ddf than T_0. We further note that T_1 is based upon all the available data. Under the non-null case, both test statistics have noncentral F-distributions with the same noncentrality parameter λ as in (3.10). Therefore F_1 will have larger power than F_0, Graybill [4].

When the m regression lines in (1.2) are an unidentified subset of the k regression lines in the model (1.1), testing H_0 against H_a is equivalent to testing H_0': the k+m regression lines are identical against H_a': at least two of them are different.

Following the procedure outlined in Section 2, it can be shown that the sum of squares for testing H_0' is

$$SSH_0' = \sum_{i=1}^{k+m} n_i(\hat{\alpha}_i - \hat{\alpha}')^2 + \sum_{i=1}^{k+m} S_i(\hat{\beta}_i - \hat{\beta}')^2, \tag{4.4}$$

where

$$\hat{\alpha}' = \sum_{i=1}^{k+m} n_i\bar{y}_i \Big/ \sum_{i=1}^{k+m} n_i, \tag{4.5}$$

$$\hat{\beta}' = \sum_{i=1}^{k+m} S_i^2\hat{\beta}_i \Big/ \sum_{i=1}^{k+m} S_i^2. \tag{4.6}$$

The sampling distribution of SSH_0'/σ^2, when H_a' is true, is noncentral chi-square with 2(k+m-1) degrees of freedom and noncentrality parameter

$$\lambda' = (1/\sigma^2)[\sum_{i=1}^{k+m} n_i(\alpha_i - \alpha)^2 + \sum_{i=1}^{k+m} S_i^2(\beta_i - \beta)^2], \tag{4.7}$$

where

$$\alpha = \sum_{i=1}^{k+m} n_i\alpha_i / \sum_{i=1}^{k+m} n_i, \tag{4.8}$$

and

$$\beta = \sum_{i=1}^{k+m} S_i^2\beta_i / \sum_{i=1}^{k+m} S_i^2 . \tag{4.9}$$

We note that SSH_0' can be obtained from SSH_0 by replacing k with k+m. When H_0' is true, the sampling distribution of SSH_0'/σ^2 is chi-square with 2(k+m-1) degrees of freedom. Further, SSH_0' and $R_0'^2$ are independent.

We use an F-test (T_2) to test H_0' against H_a' based upon the test statistic

$$F_2 = (SSH_0'/f_1')/(R_0'^2/n'), \tag{4.10}$$

where $f_1' = 2(k+m-1)$. We reject H_0' if $F_2 > F_{\alpha,f_1',n'}$ and accept H_0' otherwise. When H_a' is true, the sampling distribution of F_2 is noncentral F with f_1' ndf and n' ddf and noncentrality parameter λ'. We use noncentral F-distribution tables to compute the power of the tests. The next section illustrates and compares the power of these three tests.

5. POWER COMPARISONS OF THE TESTS. When H_0 and H_0' are not true the test statistics (4.1), (4.3), and (4.10) follow noncentral F-distributions. The non-null distributions of the test statistics F_0 and F_1 have the same noncentrality parameter, λ, defined in (3.10). The noncentrality parameter for the non-null distribution of F_2 is λ' as defined in (4.7). The ndf for both T_0 and T_1 is $f_1 = 2(k-1)$. For T_2 the ndf is $f_1' = 2(k+m-1)$. T_0 has ddf $f_2 = n-2k$, while the ddf for T_1 and T_2 is n' as defined in (2.13).

Tables 1, 2, and 3 illustrate the powers of T_0 and our proposed tests, T_1 and T_2. We chose $\alpha = 0.05$ and situations involving k = 4 regression lines. The number of data sets considered from unidentified populations is m, where $1 \le m \le k$. For simplicity we use equal sample sizes ($n_i = 10$) for the k identified populations and equal sample sizes (n_i^*) for the m unidentified populations. From our earlier notation $n_i^* = n_{k+i}$ (i = 1,...,m). Tables 1, 2, and 3 differ in the magnitude of n_i^*.

In the tables we denote the noncentrality parameter for the power of test T_i as λ_i. The power of each test is a function of λ_i and the relevant degrees of freedom. As indicated above, $\lambda_0 = \lambda_1$. For m = k, each λ_i is a specific value. For m < k, λ_0 and λ_1 are (the same) specific values, but λ_2 varies depending upon which unidentified populations produce the m data sets. For this reason we calculate the tests' powers for selected sets of k regression lines and values of S_i^2. The differences between the parameters of these lines together with S_i^2 and σ^2 affect λ_i. The parameters of the k lines, S_i^2, and σ^2 were chosen to produce the three values indicated for λ_0, so that the power of T_0 is about .25, .5, and .75. If T_0 has very small power, then additional data provide very little improvement. Conversely, when the power of T_0 is very large, there is little need for improvement with additional data.

Examinations of the tables produce the following observations. The powers of T_0 are the same in all three tables because this test ignores the additional data sets. For a given λ_0 (and λ_1)

the power of T_1 is always greater than the power of T_0 consistent with Graybill's conclusion [4] that for a given ndf the power of the test increases as the ddf increases. Also, in each table the power of T_1 increases as m increases, because the ndf remains at 2(k-1) while the ddf increases by n_i^*-2. Likewise, for each value of m the power of T_1 increases from Table 1 through Table 3 because the ddf increases as a result of the n_i^* increasing from 5 to 7 and finally to 10.

The power of T_2 is heavily influenced by the choice of regression lines for the additional data when m < k. In each table the power of T_2 does not consistently increase as m increases. The increases in λ_2 and ddf are sometimes offset by the increase in f_1' of 2m. For each value of m the power of T_2 generally increases from Table 1 through Table 3 because of the same increase in ddf as for T_1. But as Table 1 indicates, for small n_i^* relative to n_i the power of T_2 may be lower than the power of T_0, and is seldom much better than the power of T_1. In Table 2 when the n_i^* approaches n_i in size, improvements in the power of T_2 over the power of T_1 are noticeable. Table 3 indicates that when the n_i^* equal n_i, T_2 is superior to the power of T_1 except occasionally for small m.

6. APPLICATIONS. Using additional data from unidentified populations improves the power of the test for stability of the parameters in k regression lines. The only requirement is that the error terms of the regression lines from all populations have a common variance. The power of our proposed test, T_1, is always greater than the power of the standard test, T_0. If m, the number of data sets from unidentified populations, is close to k and if the n_i^* are near the n_i, then T_2 can produce a larger increase in the power than T_1. If m is small or if n_i^* is small relative to n_i, then T_1 may be a better choice than T_2.

m	noncentrality parameters λ_0, λ_1	noncentrality parameters λ_2	Power of the tests T_0	Power of the tests T_1	Power of the tests T_2
4	4.454	6.681	.2502	.2622	.2409
4	9.085	13.627	.5016	.5252	.5069
4	14.866	22.299	.7500	.7750	.7744
3	4.454	5.707 to 6.440	.2502	.2598	.2233 to .2518
3	9.085	12.012 to 12.770	.5016	.5205	.4803 to .5105
3	14.866	19.191 to 21.353	.7500	.7701	.7299 to .7858
2	4.454	4.895 to 6.240	.2502	.2570	.2115 to .2684
2	9.085	10.648 to 12.055	.5016	.5151	.4645 to .5243
2	14.866	16.601 to 20.564	.7500	.7645	.6944 to .8050
1	4.454	4.650 to 5.248	.2502	.2539	.2256 to .2536
1	9.085	9.784 to 10.405	.5016	.5089	.4738 to .5029
1	14.866	15.637 to 17.399	.7500	.7579	.7139 to .7687

Table 1
Power of the F-tests for $\alpha = .05$ $k = 4$ $n_i = 10$ $n_i^* = 5$

m	noncentrality parameters λ_0, λ_1	noncentrality parameters λ_2	Power of the tests T_0	T_1	T_2
4	4.454	7.572	.2502	.2674	.2836
4	9.085	15.444	.5016	.5353	.5914
4	14.866	25.273	.7500	.7851	.8522
3	4.454	6.178 to 7.228	.2502	.2643	.2481 to .2913
3	9.085	13.127 to 14.217	.5016	.5294	.5388 to .5810
3	14.866	20.826 to 23.919	.7500	.7792	.7874 to .8529
2	4.454	5.071 to 6.954	.2502	.2606	.2228 to .3057
2	9.085	11.287 to 13.243	.5016	.5221	.5014 to .5832
2	14.866	17.295 to 22.843	.7500	.7718	.7271 to .8616
1	4.454	4.717 to 5.519	.2502	.2560	.2310 to .2692
1	9.085	10.022 to 10.854	.5016	.5131	.4900 to .5287
1	14.866	15.900 to 18.261	.7500	.7624	.7283 to .7977

Table 2
Power of the F-tests for $\alpha = .05$ $k = 4$ $n_i = 10$ $n_i^* = 7$

m	noncentrality parameters λ_0, λ_1	noncentrality parameters λ_2	Power of the tests T_0	T_1	T_2
4	4.454	8.908	.2502	.2730	.3498
4	9.085	18.170	.5016	.5459	.7024
4	14.866	29.732	.7500	.7955	.9270
3	4.454	6.867 to 8.404	.2502	.2695	.2851 to .3522
3	9.085	14.776 to 16.373	.5016	.5393	.6189 to .6756
3	14.866	23.220 to 27.749	.7500	.7891	.8539 to .9207
2	4.454	5.336 to 8.026	.2502	.2650	.2395 to .3632
2	9.085	12.230 to 15.025	.5016	.5306	.5540 to .6643
2	14.866	18.336 to 26.262	.7500	.7805	.7703 to .9210
1	4.454	4.807 to 5.883	.2502	.2589	.2383 to .2909
1	9.085	10.343 to 11.461	.5016	.5188	.5119 to .5633
1	14.866	16.254 to 19.425	.7500	.7683	.7470 to .8330

Table 3
Power of the F-tests for $\alpha = .05$ $k = 4$ $n_i = 10$ $n_i^* = 10$

REFERENCES

[1] KLIENBAUM, D. G. and L. L. KUPPER, Applied Regression and Other Multivariable Methods, Chapter 13, PWS-Kent, Boston, MA, 1978.

[2] MADDALA, G. S., Introduction to Econometrics, Chapter 4, MacMillian, New York, NY, 1988.

[3] KSHIRSAGAR, A. M., A Course in Linear Models, Chapter 3, Marcel Dekker, New York, NY, 1983.

[4] GRAYBILL, F. A., Theory and Applications of Linear Models, Chapter 4, Duxbury Press, Boston, MA, 1976.