[illegible]

[illegible]

[illegible] NSC 88-2118-M-041-001

[illegible]

[illegible]

[illegible]

Abstract

To select among two competing normal linear models, if they are nested, the usual F test can be used, also the behaviors of the F test are available. If the two models are nonnested, adopt similar idea by considering the ratio of the squared lengths of the projection of the observations **y** onto the space corresponding to the violation of the two models, the numerator and the denominator of this ratio follow the noncentral chi-square distribution with some specific degrees of freedom and noncentrality parameter and they are dependent unless the model spaces are orthogonal. In this research, I explore the empirical structure of the unknown population which generates the ratio and also explore how the selection rule depends on the number of the observations **y** [illegible] the length of the true mean vector [illegible] the angle between the two model mean vectors and other affecting factors.

Keywords [illegible] noncentral chi-square, F test, nonnested.

[illegible]

[illegible]

é ß æ ê d ê ß ÿ ê á ó ¯ã â ß æ ã ß_..KK ó ã ã ã ê K

â ß_ê ÿ ó ê d

Let the true normal linear model be given by

$$\mathbf{Y} = \text{¯} + \text{£}, \quad \text{£} \sim \mathbf{N}\left(\text{M}, \text{±}^{0}\mathbf{I}\right) \text{ ß^}$$

and let the two competing models be referred to as:

Model A ß d $\mathbf{Y} = \text{¯}_{\wedge} + \text{£}, \quad \text{¯}_{\wedge} \in \mathbf{n}_{\mathbf{A}}$.

and

Model B ß d $\mathbf{Y} = \text{¯}_{\_} + \text{£}, \quad \text{¯}_{\mathbf{B}} \in \mathbf{n}_{\mathbf{B}}$.

If the above are two nonnested normal linear models, i.e., $\mathbf{n}_{\mathbf{A}} \cap \mathbf{n}_{\mathbf{B}} \neq \mathbf{n}_{\mathbf{A}}$ and $\mathbf{n}_{\mathrm{A}} \cap \mathbf{n}_{\mathrm{B}} \neq \mathbf{n}_{\mathrm{B}}$, here $\mathbf{n}_{\mathbf{A}} \cap \mathbf{n}_{\mathbf{B}} \neq \mathbf{n}_{\mathbf{A}}$ and $\mathbf{n}_{\mathrm{A}} \cap \mathbf{n}_{\mathrm{B}} \neq \mathbf{n}_{\mathrm{B}}$. Since Θ_A and Θ_B may overlap, After using orthogonal decomposition to remove the overlap subspace, model A becomes:

$$Y_{A+B} = \text{¦}_{A} + \text{£}, \quad \text{¦}_{\mathbf{A}} \in L_A;$$

model B becomes :

$$Y_{A+B} = \text{¦}_{B} + \text{£}, \quad \text{¦}_{B} \in L_B,$$

here L_A and L_B represent the model parameter spaces after the overlap subspace has been removed for model A and model B respectively, L_A and L_B only intersects at the origin point, i.e., $L_A \cap L_B = \{0\}$; y_{A+B} represents the projection of the sample observation vector $\mathbf{y}$ onto the combined parameter space $L_A \oplus L_B$. Note that the true variance is assumed to be the same as the variance under the candidate models and the true model, for easy sake, it will be taken as 1. Also the symmetric coordinate system for $L_A \oplus L_B$, suggested in Efron (1984), is adopted below. Let d_A and d_B represent the dimensions of L_A and L_B respectively, decompose L_A into d_A orthogonal one dimensional spaces, L_{A1}, L_{A2},…, and a similar decomposition of L_B into d_B orthogonal one dimensional spaces, L_{B1}, L_{B2},…. Note that L_{Ai} is orthogonal to L_{Bj} for $i \neq j$ and, L_{Ai} and L_{Bi} is the ith pair of the canonical variables.

Three factors were varied in the Monte Carlo simulations, the first factor is the length of the mean vector, which is l;the second factor is the angle between L_A and L_B,which is ~ ; the third factor is the degrees of freedom of the common error terms ß^which is p_E. The length l steps from 0 to 8 by increment 0.4, the angle ~ steps among ®/8, 3®/8 and ®/2, p_E steps among 5,10,20 and 50. The data Y_{A+B} were generated by adding the error terms £ to the true mean vector ¦ . The $n \times 1$ error terms are randomly independently generated 10000 times from the standard normal distributions and were stored in an $n \times 10000$ matrix, they were used repeatly in each step for the reason of eliminating the effects of variation from the error terms while doing the comparisons.

Assume ¦ is uniformly generated from the $d_A + d_B$ dimensional sphere. To select among two candidate models, the ratio of the squared lengths of the projection of $\mathbf{y}$ onto the space corresponding to the violation of the two models are computed, the selection rule can be written as

$$\text{Select A} \quad \text{if} \quad \frac{\left\|\mathbf{y}_{A^{\perp}}\right\|^2 + \left\|\mathbf{y}_{\mathrm{E}}\right\|^2}{\left\|\mathbf{y}_{\mathrm{B}^{\perp}}\right\|^2 + \left\|\mathbf{y}_{\mathrm{E}}\right\|^2} < c_1,$$

Here $\mathbf{y}_E$ represents the projection of $\mathbf{y}$ onto the error space and

$$\mathbf{y}_{A+B} = \mathbf{y}_{A^{\perp}} + \mathbf{y}_{A} = \mathbf{y}_{B^{\perp}} + \mathbf{y}_{B}$$

is the orthogonal decompositions.

Let $\mathbf{W} = \dfrac{\left\|\mathbf{y}_{A^{\perp}}\right\|^2 + \left\|\mathbf{y}_{\mathrm{E}}\right\|^2}{\left\|\mathbf{y}_{\mathrm{B}^{\perp}}\right\|^2 + \left\|\mathbf{y}_{\mathrm{E}}\right\|^2}$,

here the numerator of $\mathbf{W}$ follows the noncentral chi-square distribution with $d_B + p_E$ degrees of freedom and noncentrality parameter $\left\| ¦_{\mathrm{A}^{\perp}} \right\|^2$,the denominator of $\mathbf{W}$ follows the noncentral chi-square distribution with $d_A + p_E$ degrees of freedom and noncentrality parameter $\left\| ¦_{\mathrm{B}^{\perp}} \right\|^2$ and they are dependent unless $\mathbf{n_A} \perp \mathbf{n_B}$. Suppose $\mathbf{W}$ is generated from the unknown population distribution $\mathbf{F}$, since the structure of $\mathbf{F}$ is completely unknown, the empirical distribution function $\hat{\mathbf{F}}$ will replace the unknown F. Consider the case when $d_A = 1$ and $d_B = 2$, let ~ represent the angle between L_{A1} and L_{B1}, for fixed p_E, ~ and l ,the mean vector ¦ is uniformly generated 100 times from the combined parameter spaces, the data is generated by adding some error terms £, £ ~ $N(0, I)$, to each of the mean vector, and they are replicated 10000 times. It will be interesting to see how the probability $P(select\ A) = P(\mathbf{W} < \mathrm{c}_1)$ depends on ¦ for some fixed c_1 values. Especially to see how such probability depends on $\left\| ¦_{A^{\perp}} \right\|^2 - \left\| ¦_{B^{\perp}} \right\|^2$, which is the difference in the discrepancies due to approximation of model A and model B to the true operating model (Linhart and Zucchini (1986)). Note that negative such measurement indicate model A has less discrepancy due to approximation than model B, thus, negative region denotes A better region. Positive such measurement indicates model A has more discrepancy due to approximation than model B, thus positive region denotes B better region. If such measurement is zero, it indicates both models representing the true operating model equally well. This is the region that model A fits same as model B. Using such methodology for defining the better fitting is same as using the Kullback-leibler information number for measuring the discrepancy about the true model and the competing models. Let $f_A(\mathbf{y} \mid ¯_A)$ denote the likelihood p.d.f. of model A, $f_B(\mathbf{y} \mid ¯_B)$ denote the likelihood p.d.f. of model B. the Kullback-leibler information number of model A is less than the number of model B

iff $\mathrm{E}[\log f_A(\mathbf{y} \mid ¯_A)] > \mathrm{E}[\log f_B(\mathbf{y} \mid ¯_{\mathrm{B}})]$

iff $\left\| ¯ - ¯_A \right\|^2 < \left\| ¯ - ¯_B \right\|^2$

From the simulation results, some phenomena can be observed.

(1) The probability of selecting model A versus the quantity $\left\| ¦_{A^{\perp}} \right\|^2 - \left\| ¦_{B^{\perp}} \right\|^2$ is nonincreasing, for any p_E, ~ and l, it shows that when the signed magnitude of $\left\| ¦_{A^{\perp}} \right\|^2 - \left\| ¦_{B^{\perp}} \right\|^2$ gets bigger, which means when the mean vector gets closer to model B than model A, the probability $P(select\ A)$ is nonincreasing.

(2) To see the effect of l, for fixed p_E and ~, the plots of $P(select\ A)$ versus $\left\| ¦_{A^{\perp}} \right\|^2 - \left\| ¦_{B^{\perp}} \right\|^2$ for different value of l show that when l is very small, the difference of the distances from L_A and L_B is small too, thus no matter the mean vector is closer to L_A or L_B, the probability of selecting A remains about the same. When l increases, the selecting problem is easier since $\mathbf{y}$ will tend to belong to the region for selecting model A

or $\mathbf{y}$ will tend to belong to the region for selecting model B, therefore, $P(select\ A)$ increases on A better region and decreases on B better region. For large l, when $\|¦_{A^{\perp}}\|^2 << \|¦_{B^{\perp}}\|^2$, $P(select\ A)$ is approximately to be 1 and when $\|¦_{A^{\perp}}\|^2 >> \|¦_{B^{\perp}}\|^2$, $P(select\ A)$ is approximately to be 0.

(3) For fixed $\|\mathbf{y}_{A^{\perp}}\|^2$, $\|\mathbf{y}_{B^{\perp}}\|^2$ and $\|\mathbf{y}_E\|^2$, the selection rule can be written as: Select A if $\|\mathbf{y}_{A^{\perp}}\|^2 - c_1\|\mathbf{y}_{B^{\perp}}\|^2 + (1-c_1)\|\mathbf{y}_E\|^2 < 0$.

The left hand side of this above equation is a decreasing function of c_1, therefore, a large value of c_1 will give a higher probability of selecting model A.

(4) To see the effect of p_E, for fixed l and $\sim$, when c_1=1, the selection rule becomes :Select A if $\|\mathbf{y}_{A^{\perp}}\|^2 < \|\mathbf{y}_{B^{\perp}}\|^2$, of which does not cooperate the $\|\mathbf{y}_E\|^2$ term. Thus, all the plots are essentially the same with respect to different p_E. But when c_1>1, since c_1 is the penalty that been put on the projection of $\mathbf{y}$ onto the violation space of model B to increase the probability of selecting model A, as c_1 increases, $P(select\ A)$ increases, too. When c_1 value changes from 1 to higher than 1, the probability $P(select\ A)$ has bigger changing. For example, when $\sim = {}^{®}\!/_{8}$, l=1 and p_E=40, when c_1=1, the range of the probability is from 0.15 to 0.3, but when c_1=1.1, such range jumps to 0.82 to 0.95. In this case, the probability is very sensitive about the c_1 value. For fixed $\|\mathbf{y}_{A^{\perp}}\|^2$, $\|\mathbf{y}_{B^{\perp}}\|^2$ and c_1>1, the selection rule is equivalent to

Select A if $\dfrac{\|\mathbf{y}_{A^{\perp}}\|^2 - c_1\|\mathbf{y}_{B^{\perp}}\|^2}{c_1 - 1} < \|\mathbf{y}_E\|^2$,

Therefore, as p_E tends to infinity, $\|\mathbf{y}_E\|^2$ will tend to p_E, and $P(select\ A)$ will increase. Also, for fixed c_1>1, $P(select\ A)$ increases by p_E.

(5) In most of the applied problems, a simpler model is preferred. If this is the case, since all the plots indicate that $P(select\ A)$ is decreasing versus $\|¦_{A^{\perp}}\|^2 - \|¦_{B^{\perp}}\|^2$, there exists one special c_1 value, say c_1^*, such that $P(select\ A) = P(\mathbf{W} < c_1^*)$ is at least 0.5 when $\|¦_{A^{\perp}}\|^2 \le \|¦_{B^{\perp}}\|^2$. Whenever using another $c_1^{'}$> c_1^* value, there is a trade in the probabilities, that is, $P(select\ A)$ increases in A better region, but $P(select\ A)$ also increases in B better region. When $d_A = 1$ and $d_B = 1$, there exists a c_1^*=1, such that no matter what p_E, l or $\sim$ are, the probability of selecting model A is

$$P(\mathbf{W} < 1)\begin{Bmatrix} > 0.5 \text{ when } \|¦_{A^{\perp}}\|^2 < \|¦_{B^{\perp}}\|^2 \\ = 0.5 \text{ when } \|¦_{A^{\perp}}\|^2 = \|¦_{B^{\perp}}\|^2 \\ < 0.5 \text{ when } \|¦_{A^{\perp}}\|^2 > \|¦_{B^{\perp}}\|^2 \end{Bmatrix}.$$

When chose another $c_1^{'}$> c_1^*=1, $P(select\ A)$ increases in A better region, but it also increases in B better region. When $d_A = 1$ and $d_B = 2$, the c_1^* can be affected by several factors: p_E, l and $\sim$. The goal is to look for the minimum c_1 value such that when model A fits better than or equally to model B, the probability of selecting A is at least 0.5. But when $\|¦_{A^{\perp}}\|^2 - \|¦_{B^{\perp}}\|^2$ is fixed, the computer results show that the probability $P(select\ A)$ is not constant, the location of the mean vector actually causes a small difference in the probability. For example, when systematically generating 100 vectors on the three dimensional combined parameter space $L_A \oplus L_B$ sphere with same distance away from L_A and L_B space, it shows that for these points, most of the time, $P(select\ A)$ has a fixed pattern, which has minimum occurs at the point $(l\cos\frac{\sim}{2}, l\sin\frac{\sim}{2}, 0)$. Thus, this vector will be called "the least favorable point" among those vectors satisfying $\|¦_{A^\perp}\|^2 \le \|¦_{B^\perp}\|^2$ and having the minimum probability of selecting A. We will look forward a special c_1 value, say c_1^*, such that $P(select\ A)$ is 0.5 at this point, then using this c_1^*, the probability $P(select\ A)$ will at least 0.5 for all of the points satisfying $\|¦_{A^\perp}\|^2 \le \|¦_{B^\perp}\|^2$. To see how the c_1^* is affected by p_E, l and $\sim$, 10000 replications were simulated for each of the following steps. The length l was varied from 0 to 8 with 50 steps in between, the angle was varied among ®/8, 2®/8, 3®/8 and ®/2, p_E was shown 5 to 30 by increment 5. Several phenomena can be observed from the simulation results.

(a) For fixed p_E, when $\sim$ is small, c_1^* remains almost constant no matter how large l is. But for large angle $\sim$, c_1^* decreases by l. The reason is when $\sim$ is small, letting all points in the A better or equally better region to select A with at least probability 0.5 is not easy even l is large, since "the least favorable point" is half way between L_A and L_B, which is very close to each other when $\sim$ is small, thus, the penalty remains about the same even when l is large. But when $\sim$ is large, then as l increases, it is more and more easier to let "the least favorable point" tend to select A, thus, the penalty c_1^* value decreases.

(b) When l and $\sim$ are fixed, c_1^* decreases by p_E, since the larger the sample sizes is, the more easier to tell which of the model should be chosen, thus, c_1^* value decreases.

(c) When p_E is really large, c_1^* stays stable for $l \le 3$, and for $l > 3$, the c_1^* has a lightly changing in the value, with larger angle $\sim$ causing smaller c_1^* value.

One example was illustrated to explain the use of c_1^*. When $d_A = 1$, $d_B = 2$ and $\sim =$ ®/8 (which means the correlation between the first pair of the canonical variables is cos(®/8)), to choose a suitable c_1^* value for which $P(select\ A)$ is at least 0.5 when A actually fits better than B, when p_E is 10, the c_1^* value chosen to be used is about 1.18 with minor difference according to the length l, and when p_E is 20, the c_1^* value chosen to be used is about 1.09 with minor difference according to the length l. Choosing any value c_1' bigger than c_1^* will cause selecting model A more and selecting model B less.

â B_î đ ĥ ǧ«

[1] Akaike H. (1973). Information theory and an extension of the maximum likelihood principle.

In *proc. of the Second International Symposium on Information Theory,* B.N. Petrov and F. Czaki, eds. Akademiai Kiado, Budapest, 267-281.

[2] Akinson, A.C. (1970). A method of discriminating between models (with discussion). *J. Amer. Statist. Assoc.,* **32**, 323-345.

[3] Cox, D.R. (1962). Further results on tests of separate families of hypotheses . *J. Roy. Statist. Soc. Ser. B,* **24**, 406-424.

[4] Drer, A.R. (1974). Hypothesis testing procedures for separate families of hypotheses. *J. Amer. Statist. Assoc.,* **69**, 140-145.

[5] Efron,B. (1984). Comparing nonested linear models. *J. Amer. Statist. Assoc.,* **79**, 791-803.

[6] Fraser D.A.S. and R.J. Gebotys (1987). Nonnested linear models: a conditional confidence approach. *Canad. J. Statist.,* **15**, 375- 386.

[7] Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem on nuisance parameters. *Ann. Math. Statist.,* **11**, 271-283.

[8] Linhart, H. and W. Zucchini (1986). *Model Selection.* Wiley, New York.

[9] Pauler D.K. (1995). The Schwarz criterion and related methods for model selection in linear regression. Technical Report, **634.** Carnegie Mellon University, Pittsburg.

[10] Sawyer, K.R.(1983). Testing separate families of hypotheses : an information criterion. *J. Roy. Statis. Soc. Ser.,* **B**, 89-99.