ON TESTING A SUBSET OF REGRESSION PARAMETERS UNDER HETEROSCEDASTICITY

Miin-Jye Wen
Department of Statistics
National Cheng-Kung University
Tainan, Taiwan 70101

Hubert J. Chen
Department of Statistics
University of Georgia
Athens, GA 30602

Shun-Yi Chen
Department of Mathematics
Tamkang University
Tamsui, Taiwan 251

Key Words and Phrases: Regression analysis; heteroscedasticity; one-sample procedure; point estimation; hypothesis testing; partial test; analysis of variance.

ABSTRACT

Assuming a general linear model with unknown and possibly unequal normal error variances, the interest is to develop a one-sample procedure to handle the hypothesis testing on all, partial, or a subset of regression parameters. The sampling procedure is to split up each single sample of size $n_i$ at a controllable regressor’s data point into two portions, the first consisting of the $n_i - 1$ observations for initial estimation and the second consisting of the remaining one for overall use in the final estimation in order to define a weighted sample mean based on all sample observations at each data point. Then, the weighted sample mean is used to serve as a basis for parameter estimates and test statistics for a general linear regression model. It is found that the distributions of the test statistics based on the weighted sample means are completely independent of the unknown variances. This method can be applied to analysis of variance under various designs of experiments with unequal variances.

MSC: 62F03; 62J05; 62J10
1. Introduction

Assuming a general linear model with unknown and possibly unequal normal error variances, the interest is to develop a one-sample procedure to handle the statistical problems involving point estimation and hypothesis testing on all, partial, or a subset of regression parameters. The sampling procedure is to split up each single sample of size $n_i$ at a controllable regressor’s vector data point $X_i$ into two portions, the first consisting of the first (or randomly) $n_i - 1$ observations for initial estimation and the second consisting of the remaining one for the overall use in the final estimation. A linear combination of all sample observations at each data point is formulated, where its coefficients depend on the first portion’s sample information. As a result, the proposed linear combination is a weighted unbiased sample mean for the expected regression mean response at the given vector data point. The linear combination so formulated will be used to serve as a basis for parameter estimates and test statistics for testing the hypotheses about the parameters under a general linear regression model with unknown and unequal variances. The advantage of this proposed sampling procedure is that the distributions of the tests concerning the regression parameters are completely independent of the unknown variances. Consequently, the $p$-values and critical values for various tests about the hypotheses can be simulated by a computer program as given in an Appendix. This method can be applied to regression analysis including estimation, overall test, test for a subset of parameters, and a partial test for model selection whenever the quantitative or qualitative variables or their combination are used as predictors. In addition, the proposed inferential technique can also be applied to analysis of variance models for one-way layout, two-way layout with or without interactions, and three or higher-way layout models, under a completely randomized design, randomized block design, factorial design, Latin square design and so on. Thus, this procedure shall provide an alternative to practitioners who deal with statistical data analysis whenever they encounter heterogeneous error variances.
Consider the classical general linear regression model

\[ Y_{ij} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_{ij} \]

\[ = \mathbf{P}'_i \beta + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \]  

(1)

where \( Y_{ij} \) is a response variable, \( X_{i1}, \ldots, X_{ip} \) are predictor variables and \( \mathbf{P}'_i = (1, X_{i1}, \ldots, X_{ip}) \) is a known row vector of predictors at \( i^{th} \) \( X \)-data point, \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)' \) is an unknown column vector of regression parameters subject to \( (p + 1) \leq k \), and \( e_{ij} \)'s are independently distributed as normal \( N(0, \sigma_i^2) \) random errors, where the variances \( \sigma_i^2 \)'s are unknown and possibly unequal. It is also assumed that at each \( X \)-data point \( i \), there are \( n_i \) responses, \( Y_{i1}, \ldots, Y_{in_i} \) observed. In many experimental situations the assumption that the random errors \( e_{ij} \)'s have equal variances is not justifiable, and investigations have shown that inequality of the error variances can have a serious effect on the probabilities associated with inferences where the \( p \)-value of a test is inflated (e.g., see Bishop and Dudewicz (1978)). To eliminate such deficiencies, Bishop (1978) proposed a Stein-type (1945) two-stage sampling procedure such that the distributions of the test statistics are completely independent of the variances and that the power of the test and the length of an interval can be controlled at a desirable level. The two-stage procedure, however, is a design-oriented procedure which requires additional observations, could be large, at the second stage. As a result, it may not be practicable for the problem of data analysis in practice due to project termination, budget limitation, unavailability of additional samples and other cost factors. Therefore, in the paper, we develop a one-sample procedure without taking additional observations in the future, which yields statistics whose distributions are also completely independent of the unknown variances. Furthermore, this method can be readily applied in sampling from data warehouse in data mining. The basic method and point estimates of the regression parameters are derived accordingly in Section 2. In Section 3 the hypothesis testing problems concerning all slope parameters and a subset of regression parameters are, respectively, studied, and its application to analysis of variance is elaborated. In Section
4, a simple linear and a multiple regression models are given to illustrate the use of the one-sample procedure and a numerical is given for demonstrating the use of a simulation program REGTEST.SAS. Finally, in Section 5, a summary of the proposed method is given and related works are discussed.

2. One-Sample Procedure and Point Estimates

Under the model (1), the one-sample procedure for point estimation and hypothesis testing is described as follows:

At each $X$-data point $P'_i (i = 1, \ldots, k)$ take a random sample of $n_i (\geq 3)$ observations and use the first (or randomly) $n_i - 1$ observations, say $Y_{ij} (j = 1, \ldots, n_i - 1)$, to define the usual unbiased variance estimate $S^2_i$ of $\sigma^2_i$. Let, for $i = 1, \ldots, k$,

$$
U_i = \frac{1}{n_i} + \frac{1}{n_i} \sqrt{\frac{1}{n_i} - 1 \left[ \frac{S^2_{[k]}}{S^2_i} - 1 \right]}
$$

$$
V_i = \frac{1}{n_i} - \frac{1}{n_i} \sqrt{(n_i - 1) \left[ \frac{S^2_{[k]}}{S^2_i} - 1 \right]}
$$

be coefficients which satisfy the following conditions

$$(n_i - 1)U_i + V_i = 1,$$

$$(n_i - 1)U_i^2 + V_i^2 = S^2_{[k]} / n_i S^2_i,$$

where $S^2_{[k]}$ is the maximum of $S^2_1, \ldots, S^2_k$. Let the final weighted sample mean (including the remaining one last sample observation, $Y_{in_i}$) be defined by

$$
\bar{Y}_i = \sum_{j=1}^{n_i} W_{ij} Y_{ij}, \quad i = 1, \ldots, k,
$$

where

$$
W_{ij} = \begin{cases} 
U_i & \text{for } 1 \leq j \leq n_i - 1 \\
V_i & \text{for } j = n_i.
\end{cases}
$$

From expression (4), it is clear that the $i^{th}$ sample of size $n_i$ is split up into two portions, the first consisting of $n_i - 1$ observations and the second consisting of the remaining one,
which is the best allocation argued by Chen and Lam (1989) in their point estimation of the largest normal mean. They also explained that the weighted sample mean \( \tilde{Y}_i \) (with weights \( U_i \) and \( V_i \)) depends on \( S_i^2 \) with \( U_i \) being larger than \( 1/n_i \) and \( V_i \) being smaller than \( 1/n_i \), where the weight \( V_i \) could be negative. The first condition of (3) can guarantee that the resulting weighted sample mean \( \tilde{Y}_i \) is unbiased for the expected mean, \( E(Y_{ij}) \) at the \( X \)-data point \( P'_i \) and the second condition of (3) together with the first one produces the solution \( U_i \) and \( V_i \) (2). In the case where the error variances \( \sigma_i^2 \)'s are homogeneous, the coefficients \( U_i \) and \( V_i \) defined in (2) are close to \( 1/n_i \) and hence the weighted sample mean (4) is close to the usual unbiased sample mean. Not only is the weighted sample mean \( \tilde{Y}_i \) (4) unbiased, but also it is consistent for \( P'_i \beta \). In theory, no matter which \( (n_i - 1) \) observations were used in the first portion, the fundamental inference discussed below just works fine.

It is clear that for given \( S_i^2 \) \((i = 1, \ldots, k)\), the weighted mean response \( \tilde{Y}_i \) at the data point \( P'_i \) has a normal distribution, \( N(P'_i \beta, \sigma_i^2 \sum_{j=1}^{n_i} W_{ij}^2) \). Let

\[
T_i = \frac{\tilde{Y}_i - P'_i \beta}{S_i^2 / \sqrt{n_i}} = \frac{\tilde{Y}_i - P'_i \beta}{S_i \sqrt{\sum_{j=1}^{n_i} W_{ij}^2}} \tag{5}
\]

where the denominator in (5) has a nice property which can eliminate the influence of the unknown variances \( \sigma_i^2 \) in the probability distribution of the above random variable \( T_i \) used in the inference. We can use the same technique by Chen and Chen (1998) in their single-stage ANOVA problem to derive the distribution of \( T_i \)'s. Then, conditioning on \( S_i^2 = s_i^2 \) \((i = 1, \ldots, k)\), \( T_i \) has a conditional normal distribution \( N(0, \sigma_i^2 / s_i^2) \). Let \( h_i(s_i^2) \) be the density of \( S_i^2 \), which is distributed as \( (\sigma_i^2 / v_i)\chi^2_{v_i} \) with \( v_i = n_i - 2 \) df. Then, the marginal density of \( T_i \) is given by

\[
g_{v_i}(t_i) = \int_0^\infty N(0, \sigma_i^2 / s_i^2) h_i(s_i^2) ds_i^2
\]

\[
= \int_0^\infty \frac{s_i^2}{\sqrt{2\pi\sigma_i}} \exp\left\{ -\frac{s_i^2 t_i^2}{2\sigma_i^2} \right\} h_i(s_i^2) ds_i^2
\]

\[
= \frac{\Gamma((v_i + 1)/2)}{\Gamma(v_i/2)\sqrt{\pi v_i}} \left( 1 + \frac{t_i^2}{v_i} \right)^{-(v_i+1)/2}
\]

which is exactly the density of the Student’s t distribution with \( v_i \) df. Thus, if \( T_i \) has
a conditional normal distribution with a mean of zero and a variance of $\sigma_i^2/s_i^2$ and if $S_i^2$ is distributed as $(\sigma_i^2/v_i)\chi^2_{v_i}$, then $T_i$ has an unconditional Student’s $t$ distribution with $v_i$ df (denoted by $t_{v_i}$). This argument was also claimed by Stein (1945) without explicit proof. Furthermore, conditioning on $S_i^2 = s_i^2$ ($i = 1, \ldots, k$), $T_1, \ldots, T_k$ are independently distributed as normal $N(0, \sigma_i^2/s_i^2)$, $\ldots, N(0, \sigma_k^2/s_k^2)$, respectively. As $S_1^2, \ldots, S_k^2$ are independent r.v.’s with densities $h_i(s_i^2), \ldots, h_k(s_k^2)$, respectively, the joint density of $T_1, \ldots, T_k$ is obtained by integrating the j.p.d.f. of $T_1, \ldots, T_k, S_1^2, \ldots, S_k^2$ with respect to $S_1^2, \ldots, S_k^2$ as

$$g(t_1, \ldots, t_k) = \int_0^\infty \cdots \int_0^\infty \prod_{i=1}^k [N(0, \sigma_i^2/s_i^2)h_i(s_i^2)] \prod_{i=1}^k ds_i^2$$

$$= \prod_{i=1}^k \int_0^{\infty} N(0, \sigma_i^2/s_i^2)h_i(s_i^2)ds_i^2$$

$$= \prod_{i=1}^k g_i(t_i).$$

Thus, by noticing that the j.p.d.f. of $T_1, \ldots, T_k$ being equal to the product of their marginal p.d.f.’s, $T_1, \ldots, T_k$ have independent $t_{v_1}, \ldots, t_{v_k}$ distributions with $v_1, \ldots, v_k$ d.f., respectively.

Equation (5) can be rewritten as

$$\tilde{Y}_i = P_i' \beta + \epsilon_i, \ i = 1, \ldots, k,$$

(6)

where $\epsilon_i = S_i[k]/T_i/\sqrt{\nu_i}$. Or it can be written by the matrix form of a general linear model as

$$\tilde{Y} = X\beta + \epsilon$$

(7)

where

$$\tilde{Y} = \begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \vdots \\ \tilde{Y}_k \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{k1} & \ldots & X_{kp} \end{bmatrix}, \quad P_i = \begin{bmatrix} P_i' \\ \vdots \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{bmatrix}.$$

It is easy to see from (6) that conditioning on $S_1^2, \ldots, S_k^2$, the random errors $\epsilon_1, \ldots, \epsilon_k$ are uncorrelated with a mean of zero and variances $(S_i^2[k]/\nu_i)(\sigma_i^2/S_i^2), i = 1, \ldots, k$, respectively.
After taking expectation on \( \epsilon_i \), we have

\[
(i) \quad E\{E(\epsilon_i|S[k])\} = 0, \quad (ii) \quad E\{\text{cov}(\epsilon_i, \epsilon_j)|S[k]\} = 0, \quad i \neq j,
\]

and

\[
(iii) \quad \tau_i^2 = E\left\{\text{var}(\epsilon_i|S[k])\right\} = E\left\{\frac{S_k^2}{n_i} \cdot \frac{\sigma_i^2}{\tau_i^2}\right\} \approx E\left\{\frac{S_k^2}{\bar{n}} \cdot \frac{\sigma_i^2}{\tau_i^2}\right\}, \quad (8)
\]

where \( \bar{n} = [(\sum_{j=1}^{k} 1/n_j)/k]^{-1} \) is the harmonic mean of \( n_i \)'s. Although \( \tau_i^2 \) is a function of unknown variances, the ratio \( E\left\{\text{var}(\epsilon_i|S[k])\right\}/E\left\{\sigma_k^2/\bar{n}\right\} \) converges to one almost surely, where \( \sigma_k^2 \) is the largest of \( \sigma_i^2 \)'s, (Chen (1975)) as the common sample size \( n \) is large, \( n = n_1 = \ldots = n_k \) or the smallest sample size is large if \( n_i \) are not all equal.

It is reasonable to consider the usual least-squares estimate of \( \beta \) as an approximate point estimate, given by

\[
\hat{\beta} = (X'X)^{-1}X'\tilde{Y}, \quad (9)
\]

which will make the testing procedure possible in Section 3. Substituting \( \hat{\beta} \) in (9) for \( \beta \) in (6), we obtain the prediction equation for (1) given by

\[
\hat{Y}_i = P'_i \hat{\beta}, \quad i = 1, \ldots, k. \quad (10)
\]

Substituting \( \tilde{Y} \) in (7) into (9), we obtain

\[
\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon.
\]

Then, the expectation of \( \hat{\beta} \) is \( E(\hat{\beta}) = \beta \), and the variance-covariance matrix of \( \hat{\beta} \) is given by

\[
\Sigma \hat{\beta} = (X'X)^{-1}X'DX(X'X)^{-1},
\]

where \( D = \{\tau_i^2\} \) is a diagonal matrix of order \( k \). When \( n_1, \ldots, n_k \) are large (or equivalently \( \bar{n} \) is large),

\[
\tau_i^2 \approx E\left\{\frac{\sigma_k^2}{\bar{n}}\right\} = \tau^2. \quad (11)
\]
Then, asymptotically, the Gauss-Markovoff theorem applies and

\[ \Sigma \hat{\beta} \approx \tau^2 (X'X)^{-1}. \]  

(12)

The variance of \( \epsilon_i \) is asymptotically equal to a common unknown value \( \tau^2 \) which is a function of all variances, \( \sigma^2_1, \ldots, \sigma^2_k \). This provides the rationale of using the estimate in (9) as least squares estimate. In conclusion, the estimate of \( \beta \) in (9) are unbiased, strongly consistent and asymptotically having minimum variance-covariance matrix.

3. Some Commonly Used Hypothesis Testing

3.1. Testing All Slope Parameters

In the classical general linear regression analysis, one of the goals is to test the null hypothesis that all slope parameters are simultaneously equal to zero, or equivalently to test the overall null hypothesis \( H_0 : \beta_1 = \cdots = \beta_p = 0 \) against the alternative that at least one slope parameter is not equal to zero. It is assumed that for \( i = 1, \ldots, k \) the one-sample procedure has been conducted and that the final weighted mean response \( \tilde{Y}_i \) have been computed as in (4) and the one-sample linear regression model is given by (6). Let \( \tilde{Y}_i = \sum_{i=1}^{k} \tilde{Y}_i/k \), be the mean of \( \tilde{Y}_i \). We consider the following test statistic for \( H_0 \)

\[ \hat{F} = \sum_{i=1}^{k} \left( \frac{\tilde{Y}_i - \tilde{Y}_i}{S[k]/\sqrt{n_i}} \right)^2 \]  

which can be interpreted as the sum of squares due to the full regression model (1). Substituting the prediction equation (10) and the model (6) into (13), \( \hat{F} \) can be rewritten as

\[ \hat{F} = \sum_{i=1}^{k} \left( \frac{P'_{i}(\hat{\beta} - \beta) - (\tilde{Y}_i - \sum_{j=1}^{k} P'_{j}\beta/k) + P'_{i}\beta - \sum_{j=1}^{k} P'_{j}\beta/k}{S[k]/\sqrt{n_i}} \right)^2 \]  

\[ = \sum_{i=1}^{k} \left( \frac{P'_{i}(\hat{\beta} - \beta) - (\sum_{j=1}^{k} \tilde{Y}_j - P'_{j}\beta)/k + (P'_{i} - \bar{P}'_{i})\beta}{S[k]/\sqrt{n_i}} \right)^2 \]  

\[ = \sum_{i=1}^{k} \left( \frac{(P'_{i}(X'X)^{-1}X'\epsilon - \sum_{j=1}^{k} \epsilon_j/k + (P'_{i} - \bar{P}'_{i})\beta}{S[k]/\sqrt{n_i}} \right)^2 \]
\[
\sum_{i=1}^{k} \left( \frac{\sum_{j=1}^{k} b_{ij} \epsilon_j - \sum_{j=1}^{k} \epsilon_j / k + (\mathbf{P}'_i - \mathbf{P}') \beta}{S[k] / \sqrt{n_i}} \right)^2
\]

\[
= \sum_{i=1}^{k} \left( \sqrt{n_i} \left( \frac{\sum_{j=1}^{k} b_{ij} T_j}{\sqrt{n_j}} - \frac{1}{k} \sum_{j=1}^{k} T_j \frac{1}{\sqrt{n_j}} \right) + \frac{(\mathbf{P}'_i - \mathbf{P}') \beta}{S[k] / \sqrt{n_i}} \right)^2
\]

\[
= \sum_{i=1}^{k} \left( \sqrt{n_i} \sum_{j=1}^{k} \left( b_{ij} - \frac{1}{k} \right) T_j \frac{1}{\sqrt{n_j}} + \frac{(\mathbf{P}'_i - \mathbf{P}') \beta}{S[k] / \sqrt{n_i}} \right)^2
\]  (14)

where \( \{b_{ij}\} = \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \) and \( \mathbf{P}' = \sum_{j=1}^{k} \mathbf{P}'_j / k \).

Under the null hypothesis \( H_0 : \beta_1 = \ldots = \beta_p = 0 \), \( \tilde{F} \) in (14) is distributed as

\[
\tilde{F} = \sum_{i=1}^{k} \left( \sqrt{n_i} \sum_{j=1}^{k} \left( b_{ij} - \frac{1}{k} \right) T_j \frac{1}{\sqrt{n_j}} \right)^2
\]  (15)

In situation where all sample sizes, \( n_1, n_2, \ldots, n_k \) are equal to \( n \), the test statistic in (15) reduces to the simpler form

\[
\tilde{F} = \sum_{i=1}^{k} \left( \sum_{j=1}^{k} \left( b_{ij} - \frac{1}{k} \right) T_j \right)^2
\]  (16)

The null distribution of \( \tilde{F} \) in (15) is a quadratic function of linear combinations of independent Student’s t r.v.’s \( T_1, \ldots, T_k \) with df \( \nu_1, \ldots, \nu_k \), respectively. The coefficients \( b_{ij} \)'s in (15) and (16) are functions of known constants in the data (or design) matrix \( \mathbf{X} \). Thus, for a given set of data, we reject the null hypotheses \( H_0 : \beta_1 = \cdots = \beta_p = 0 \) if the \( p \)-value of the test statistic \( \tilde{F} \) in (13) (the probability that the r.v. \( \tilde{F} \) in (15) is larger than the computed test statistic \( \tilde{F} \) in (13)) is smaller than the significance level \( \alpha \), where the \( p \)-value of the test (13) can be obtained by a simulation program using SAS programming language given in Appendix named REGTEST.SAS.

3.2. Relation to One-Way ANOVA Model

We now show that the one-way layout analysis of variance model under a completely randomized design is a special case of the general linear regression model (1). Consider the one-way layout model

\[
Y_{ij} = \mu_i + \epsilon_{ij}
\]
\[
\hat{\mu} + \alpha_i + e_{ij}, \ i = 1, \ldots, k, \ j = 1, \ldots, n_i,
\]
where \( \hat{\mu} = \sum_{i=1}^{k} \mu_i / k \), \( \alpha_i = \mu_i - \hat{\mu} \), and \( e_{ij} \) are independent \( N(0, \sigma^2_i) \) random errors.

The above model can be rewritten as a general linear regression model in (1), or

\[
Y_{ij} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{i,k-1} + e_{ij},
\]

where the dummy variables are defined as

\[
X_1 = 1 \quad \text{if a } Y \text{ is in treatment 2}; \ X_1 = 0 \quad \text{otherwise},
\]
\[
X_2 = 1 \quad \text{if a } Y \text{ is in treatment 3}; \ X_2 = 0 \quad \text{otherwise},
\]
\[
\vdots
\]
\[
X_{k-1} = 1 \quad \text{if a } Y \text{ is in treatment } k; \ X_{k-1} = 0 \quad \text{otherwise}.
\]

For this model, we have \( \beta_0 = \mu_1, \ \beta_1 = \mu_2 - \mu_1, \ \beta_2 = \mu_3 - \mu_1, \ \ldots, \ \beta_{k-1} = \mu_k - \mu_1 \), and the design matrix (data matrix) is given by a square matrix \( X \) of order \( k \), where

\[
X = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
1 & 1 & 0 & \cdots & 0 & 0 \\
1 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & & & \ddots & & \vdots \\
1 & 0 & 0 & \cdots & 0 & 1
\end{bmatrix}.
\]

Furthermore, the matrix \( X(X'X)^{-1}X' \) - 1, an identity matrix of order \( k \) (This can happen in the one-way layout.) and the coefficient \( b_{ij} = 1 \) for \( i = j \) and \( b_{ij} = 0 \) for \( i \neq j \). Thus, the test statistic (14) becomes

\[
\tilde{F} = \frac{1}{k} \sum_{i=1}^{k} \left( T_i - \frac{T_i}{k} \sum_{j=1}^{k} \frac{T_j}{\sqrt{n_j}} + \frac{\mu_i - \hat{\mu}}{S[k]/\sqrt{n_i}} \right)^2
\]

and under the null hypothesis \( H_0 : \mu_1 = \cdots = \mu_k \), it further reduces to

\[
\tilde{F} = \sum_{i=1}^{k} \left( T_i - \frac{T_i}{k} \sum_{j=1}^{k} \frac{T_j}{\sqrt{n_j}} \right)^2
\]

which is the form of Chen and Chen (1998). The percentage points \( \hat{F}_\alpha \) of \( \tilde{F} \) in (18) under the one-way layout setting when sample sizes are all equal were calculated by Chen and Chen (1998). For the sample sizes being unequal, the \( p \)-value of the test statistic (13) for testing the treatment effects in an ANOVA can be obtained by the simulation program provided in Appendix named REGTEST.SAS.
3.3. Testing a subset of Slope Parameters

It is often necessary to test whether some subset of the predictors is not significantly important in the general linear regression model. This idea leads to the popular model selection procedure such as backward elimination, forward selection and stepwise selection in multiple regression analysis. In this section we consider a general partial test for testing whether a concerned subset of predictor variables has a significant contribution to the full model given in (1).

Without loss of generality, let the partial null hypothesis be

\[ H_0^* : \beta_{q+1} = \cdots = \beta_p = 0, \quad q < p, \] (19)

which means that none of the corresponding predictors \( X_{q+1}, \ldots, X_p \) has a contribution to model (1). Then, the reduced model relative to the full model (1) is given by

\[
Y_{ij} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + e_{ij} \\
= Q'_i \beta_R + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i,
\] (20)

where these responses \( Y_{ij} \) are the same as given in (1), \( Q'_i = (1, X_{i1}, \ldots, X_{iq}) \) is a \( 1 \times (q + 1) \) known vector of predictors, \( \beta_R = (\beta_0, \beta_1, \ldots, \beta_q)' \) is a reduced unknown \( 1 \times (q + 1) \) vector of regression parameters and \( e_{ij} \) are independent \( N(0, \sigma^2_i) \) r.v.'s. By using the one-sample procedure described in Section 2, similar to (6), we obtain

\[
\tilde{Y}_i = Q'_i \beta_R + e^*_i, \quad i = 1, \ldots, k,
\] (21)

where \( e^*_i = T^*_i S[k]/\sqrt{n_i} \), and

\[
T^*_i = \tilde{Y}_i - Q'_i \beta_R \frac{S[k]}{\sqrt{n_i}}, \quad i = 1, \ldots, k,
\] (22)

are independent Student t r.v.'s \( t_{\nu_1}, \ldots, t_{\nu_k} \) with df \( \nu_1, \ldots, \nu_k \), respectively. By least-squares method, we have the estimate of \( \beta_R \), similar to (9),

\[
\hat{\beta}_R = (X'_1 X_1)^{-1} X'_1 \tilde{Y} \\
= \beta_R + (X'_1 X_1)^{-1} X'_1 e^*.
\]
where

\[ \tilde{Y} = \begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \vdots \\ \tilde{Y}_k \end{bmatrix}, \quad X_1 = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1q} \\ 1 & X_{21} & \ldots & X_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{k1} & \ldots & X_{kq} \end{bmatrix}, \quad \beta_R = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}, \quad \epsilon^* = \begin{bmatrix} \epsilon_1^* \\ \epsilon_2^* \\ \vdots \\ \epsilon_k^* \end{bmatrix}. \]

Note that \( \tilde{Y}_i \) is the same one calculated using (4). The prediction equation is given by

\[ \tilde{Y}_i^* = Q_i^* \hat{\beta}_R, \quad i = 1, \ldots, k. \quad (23) \]

Consider the test statistic for \( H_0^* \)

\[ F_R = \sum_{i=1}^{k} \left( \frac{\tilde{Y}_i - \tilde{Y}_i^*}{S_{[k]}/\sqrt{m_i}} \right)^2, \quad (24) \]

where \( F_R \) can be interpreted as the sum of squares due to extra slope parameters, \( \beta_{q+1}, \ldots, \beta_p \).

Using prediction equations (10) and (23), \( F_R \) can be rewritten as

\[
\begin{aligned}
F_R &= \sum_{i=1}^{k} \left( \frac{P_i'\hat{\beta} - Q_i'\hat{\beta}_R}{S_{[k]}/\sqrt{m_i}} \right)^2 \\
&= \sum_{i=1}^{k} \left( \frac{P_i'(\hat{\beta} - \beta) - Q_i'(\hat{\beta}_R - \beta_R) + P_i'\beta - Q_i'\beta_R}{S_{[k]}/\sqrt{m_i}} \right)^2 \\
&= \sum_{i=1}^{k} \left( \frac{P_i'(X'X)^{-1}X'\epsilon - Q_i'(X_i'X_1)^{-1}X_i^*\epsilon^* + P_i'\beta - Q_i'\beta_R}{S_{[k]}/\sqrt{m_i}} \right)^2 \\
&= \sum_{i=1}^{k} \left( \frac{P_i'(X'X)^{-1}X'T_s - Q_i'(X_i'X_1)^{-1}X_i^*T^*_s}{1/\sqrt{m_i}} + \frac{R_i'\beta^*}{S_{[k]}/\sqrt{m_i}} \right)^2
\end{aligned}
\]

where \( R_i' = (X_{i,q+1}, \ldots, X_{ip}) \), \( \beta^* = (\beta_{q+1}, \ldots, \beta_p)' \), \( T_s = (T_1/\sqrt{m_1}, \ldots, T_k/\sqrt{m_k})' \), and \( T^*_s = (T^*_1/\sqrt{m_1}, \ldots, T^*_k/\sqrt{m_k})' \). We observe that the relationship between \( T_i \) and \( T_i^* \) is, using (5) and (22),

\[
T_i = \frac{\tilde{Y}_i - P_i'\beta}{S_{[k]}/\sqrt{m_i}}, \quad i = 1, \ldots, k
\]

\[
= T_i^* - \frac{(P_i'\beta - Q_i'\beta_R)}{S_{[k]}/\sqrt{m_i}}
\]

\[
= T_i^* - \frac{R_i'\beta^*}{S_{[k]}/\sqrt{m_i}}
\]
Under the null hypothesis $H_0^*: \beta^* = 0$, we obtain $T_i = T_i^*$, $i = 1, \ldots, k$, and the test statistic $\tilde{F}_R$ is distributed as

$$\tilde{F}_R = \sum_{i=1}^{k} \left( \frac{P_i'(X'X)^{-1}X'T_s - Q'_i(X_i'X_1)^{-1}X_1'T_s}{1/\sqrt{n_i}} \right)^2$$

$$= \sum_{i=1}^{k} \left( \frac{\sqrt{n_i} \left( P_i'(X'X)^{-1}X' - Q'_i(X_i'X_1)^{-1}X_1' \right) T_s}{\sqrt{n_i}} \right)^2$$

$$= \sum_{i=1}^{k} \left( \sqrt{n_i} \sum_{j=1}^{k} (b_{ij} - q_{ij}) \frac{T_j}{\sqrt{n_j}} \right)^2,$$  \hspace{1cm} (25)

where $\{q_{ij}\} = X_1(X_1'X_1)^{-1}X_1'$. The null hypothesis $H_0^*$ is rejected if the $p$-value that the r.v. $\tilde{F}_R$ in (25) is larger than the computed test statistic $\tilde{F}_R$ in (24) is smaller than the significance level $\alpha$. Since the null distribution of $\tilde{F}_R$ is a quadratic function of linear combinations of independent Student $t$ r.v.'s $t_{\nu_1}, \ldots, t_{\nu_k}$ with df $\nu_1, \ldots, \nu_k$, respectively, the $p$-value in (24) requires extensive computer simulation which is given in the Appendix named REGTEST.SAS.

It should be noted that in the case where $q = 0$, $H_0^* = H_0 : \beta_1 = \cdots = \beta_p = 0$ and $\tilde{Y}_i = \beta_0 + \epsilon^*_i$, we have $q_{ij} = 1/k$ for all $i, j$ and the test $\tilde{F}_R$ in (25) becomes $\tilde{F}$ in (15), the statistic (24) reduces to (13). Furthermore, the partial test can be applied to any subset of $(p + 1)$ regression parameters. In particular, for just testing one $\beta_i = 0$, it leads to a backward elimination or a forward selection in model selection problem.

### 3.4. Relation Between Partial Test And Two-Way ANOVA Test

The technique of a partial test for a general regression model (1) can be applied to two-way layout ANOVA model under a completely randomized design or a randomized block design described below. The model without interaction is given by

$$Y_{ijl} = \mu + \alpha_i + \tau_j + \epsilon_{ijl}$$

$$= \mu_{ij} + \epsilon_{ijl}, \quad i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad l = 1, \ldots, n_{ij},$$

where $\epsilon_{ijl}$ are assumed to be independent $N(0, \sigma_{ij}^2)$ random errors, $\mu_{ij} = \mu + \alpha_i + \tau_j$ with
µ being the overall mean, α_i is the effect due to level i of factor A and \( \tau_j \) is the effect due to level j of factor B subject to \( \sum_i \alpha_i = 0 \) and \( \sum_j \tau_j = 0 \).

For ease of explanation let us consider the case where factor A has three levels (I = 3) and factor B has two levels (J = 2). Now define the following dummy variables:

\[
X_1 = 1 \text{ if a } Y \text{ is in level 2 of factor A; } X_1 = 0 \text{ otherwise,}
\]

\[
X_2 = 1 \text{ if a } Y \text{ is in level 3 of factor A; } X_2 = 0 \text{ otherwise,}
\]

\[
X_3 = 1 \text{ if a } Y \text{ is in level 2 of factor B; } X_3 = 0 \text{ otherwise.}
\] (27)

After introducing the dummy variables \( X_1, X_2, \) and \( X_3 \), the two-way layout ANOVA model (26) can be rewritten as a general linear regression model as in (1), i.e.,

\[
Y_{ij} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i.
\]

In this case, \( k = 3 \times 2 = 6 \) rows, \( p = 3 + 2 - 2 = 3 \) predictors, \( n_1 = n_{11}, n_2 = n_{21}, n_3 = n_{31}, n_4 = n_{12}, n_5 = n_{22}, \) and \( n_6 = n_{32}. \) Note that the sub-indices for sample sizes are redefined in the regression model for simplicity.

The expected values between the regression model and the ANOVA model in parentheses are given in TABLE 1.

<table>
<thead>
<tr>
<th>Factor B</th>
<th>1 ( (\mu_{11}) )</th>
<th>2 ( (\mu_{12}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>( \beta_0 )</td>
<td>( \beta_0 + \beta_3 )</td>
</tr>
<tr>
<td>Factor A</td>
<td>2 ( (\mu_{21}) )</td>
<td>( \beta_0 + \beta_1 + \beta_3 )</td>
</tr>
<tr>
<td>3 ( (\mu_{31}) )</td>
<td>( \beta_0 + \beta_2 + \beta_3 )</td>
<td></td>
</tr>
</tbody>
</table>

By applying the one-sample procedure described in the beginning of Section 2, we have the regression model similar to (6)

\[
\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \ldots, 6.
\] (28)

The design matrix \( \mathbf{X} \) for the general linear regression model and the inverse of \( \mathbf{X}'\mathbf{X} \) are
given, respectively, by
\[
X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix} {\frac{3}{2}} & -{\frac{1}{2}} & -{\frac{1}{2}} & {\frac{1}{3}} \\ -{\frac{1}{2}} & 1 & {\frac{1}{2}} & 0 \\ -{\frac{1}{2}} & {\frac{1}{2}} & 1 & 0 \\ -{\frac{1}{3}} & 0 & 0 & {\frac{2}{3}} \end{bmatrix}.
\] (29)

For testing the null effects of factor B, the partial hypothesis is set up to be \(H_0^B: \beta_3 = 0\) which is equivalent to testing the null hypothesis for Factor B \(H_0^B: \tau_1 = \tau_2 = 0\) for the two-way layout ANOVA model. Then, under \(H_0^B\), the reduced regression model is given by
\[
\tilde{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \ldots, 6.
\] (30)

The reduced design matrix \(X_1\) and the inverse of \((X_1'X_1)^{-1}\) are given, respectively, by
\[
X_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad (X_1'X_1)^{-1} = \begin{bmatrix} 1 & -{\frac{1}{2}} & -{\frac{1}{2}} \\ -{\frac{1}{2}} & 1 & {\frac{1}{2}} \\ -{\frac{1}{2}} & {\frac{1}{2}} & 1 \end{bmatrix}.
\]

The matrices for the full model \(\{b_{ij}\} = X(X'X)^{-1}X'\) and for the reduced model \(\{q_{ij}\} = X_1(X_1'X_1)^{-1}X_1\) needed for calculating the statistic (25) are obtained as follows.
\[
\{b_{ij}\} = \begin{bmatrix} {\frac{2}{3}} & {\frac{1}{3}} & {\frac{1}{3}} & -{\frac{1}{6}} & -{\frac{1}{6}} \\ -{\frac{1}{6}} & {\frac{2}{3}} & {\frac{1}{6}} & -{\frac{1}{6}} & {\frac{1}{6}} \\ {\frac{1}{6}} & {\frac{1}{3}} & {\frac{2}{3}} & -{\frac{1}{6}} & -{\frac{1}{3}} \\ -{\frac{1}{3}} & -{\frac{1}{6}} & {\frac{2}{3}} & {\frac{1}{6}} & -{\frac{1}{3}} \\ -{\frac{1}{6}} & -{\frac{1}{3}} & -{\frac{1}{6}} & {\frac{2}{3}} & -{\frac{1}{3}} \end{bmatrix}, \quad \{q_{ij}\} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & {\frac{1}{2}} & 0 & 0 & {\frac{1}{2}} \\ 0 & 0 & {\frac{1}{2}} & 0 & 0 \\ 0 & {\frac{1}{2}} & 0 & 0 & {\frac{1}{2}} \\ 0 & 0 & {\frac{1}{2}} & 0 & 0 \end{bmatrix}.
\]

Substitute the matrix \(\{b_{ij}\}\) for the full model (28) and the matrix \(\{q_{ij}\}\) for the reduced model (30) into the formula (25) we can obtain the test statistic after simplification
\[
\tilde{F}_R^{(2)} = \sum_{i=1}^{6} \left(\sqrt{n_i} \sum_{j=1}^{6} (b_{ij} - q_{ij}) \frac{T_j}{\sqrt{n_j}}\right)^2
= \sum_{i=1}^{3} \sum_{j=1}^{2} \left(\sqrt{\frac{n_{ij}}{3}} \sum_{m=1}^{3} T_{mj} - \sqrt{\frac{m_{ij}}{6}} \sum_{m=1}^{3} \sum_{l=1}^{2} \frac{T_{ml}}{\sqrt{n_{ml}}}\right)^2.
\] (31)
where $T_{mj} (m = 1, 2, 3, j = 1, 2)$ are independent Student $t$ r.v.'s with $n_{mj} - 2$ degrees of freedom, respectively, and $T_1 = T_{11}$, $T_2 = T_{21}$, $T_3 = T_{31}$, $T_4 = T_{12}$, $T_5 = T_{22}$, $T_6 = T_{32}$. The test statistic $\tilde{F}_R^{(2)}$ in (31) turns out to be the same as that of (12) by Chen and Chen (1998) for testing the null effects of factor B in a two-way layout ANOVA model.

Similarly, for testing the null effects of factor A, the partial null hypothesis for the regression model is $H_{01}^A : \beta_1 = \beta_2 = 0$ which is equivalent to the null hypothesis $H_{00}^A : \alpha_1 = \alpha_2 = \alpha_3 = 0$ for the two-way layout ANOVA model. Under $H_{00}^A$, the reduced regression model is given by

$$\tilde{Y}_i = \beta_0 + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \ldots, 6.$$ (32)

The design matrix $X_1^*$ and the inverse of $(X_1^* X_1^*)$, and the matrix $\{q_{ij}^*\} = X_1^* (X_1^* X_1^*)^{-1} X_1^*$ are calculated as follows.

$$X_1^* = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \quad (X_1^* X_1^*)^{-1} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix},$$

and

$$q_{ij}^* = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$ (33)

Again substitute the matrix $\{b_{ij}\}$ for the full model (28) and the matrix $\{q_{ij}^*\}$ for the reduced model (32) into the formula (25), we can obtain the test statistic after simplification

$$\tilde{F}_R^{(1)} = \sum_{i=1}^{6} \left( \sum_{j=1}^{6} (b_{ij} - q_{ij}^*) \frac{T_j}{\sqrt{n_{ij}}} \right)^2$$

$$= \sum_{i=1}^{3} \sum_{j=1}^{2} \left( \frac{\sqrt{n_{ij}}}{2} \sum_{l=1}^{2} T_{il} - \frac{\sqrt{n_{ij}}}{6} \sum_{m=1}^{3} \sum_{l=1}^{2} T_{ml} \right)^2$$ (33)
which is the same test as that of (11) by Chen and Chen (1998) for testing the null effects of factor A in a two-way layout ANOVA model.

In general, if there are \( I \) levels of Factor A and \( J \) levels of Factor B, we will need to create \((I - 1)\) dummy variables for Factor A and \((J - 1)\) dummy variables for Factor B which produce \( p = I + J - 2 \) predictors, and the corresponding design matrix \( X \) will have \( k = I \times J \) rows and \((I + J - 2) + 1\) columns, whose elements consist of 0’s and 1’s similar to the \( 6 \times 4 \) design matrix \( X \) in the special case. Thus, the ANOVA tests can be performed by the partial test using a general linear regression model. In situation where the factor B is treated as a block in a randomized block design, the test procedure for testing null block effects \( H_0^2: \tau_1 = \tau_2 = 0 \) or equivalently \( H_0^2: \beta_3 = 0 \) in regression form is the same as \( \tilde{F}_{R}^{(2)} \) in a two-factor factorial design for factor B.

When the two-way layout ANOVA model under a completely randomized design has interaction terms, we have the model, for simplicity,

\[
y_{ijl} = \mu + \alpha_i + \tau_j + (\alpha \tau)_{ij} + e_{ijl}, \quad i = 1, 2, 3; \quad j = 1, 2; \quad l = 1, \ldots, n_{ij},
\]

and the corresponding equivalent full regression model by one-sample procedure is given by

\[
\tilde{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i3} + \beta_5 X_{i2} X_{i3} + \epsilon_i, \quad i = 1, \ldots, 6,
\]

where \( X_1, X_2, X_3 \) and \( e_{ij} \)'s are defined in a same manner as (27), the subscripts \( i, j \) are redefined as previously stated. The interaction terms \( \beta_4 X_{i1} X_{i3} \) and \( \beta_5 X_{i2} X_{i3} \) were thoroughly discussed by Ott (1993, p.897-).

The purpose is to test the null interaction effects between the two factors \( H_0^3: (\alpha \tau)_{ij} = 0 \) for all \( i, j \), for the ANOVA model (34), which is equivalent to testing \( H_0^3: \beta_4 = \beta_5 = 0 \) for the regression model (35), where the reduced model under \( H_0^3 \) is given by

\[
\tilde{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i.
\]

The design matrices for the full model (35) and reduced model (36) under \( H_0^3 \) are given,
respectively, by

\[
X = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1
\end{bmatrix},
X_1 = \begin{bmatrix}
1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 \\
1 & 1 & 0 & 1
\end{bmatrix}.
\]

The matrix \( \{b_{ij}\} = X(X'X)^{-1}X' = I \), an identity matrix of order 6 for the full model (35), and \( \{q_{ij}\} = X_1(X_1'X_1)^{-1}X_1' \) for the reduced model (36) under \( H_0^3 \) is

\[
q_{ij} = \begin{bmatrix}
\frac{2}{3} & \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & -\frac{1}{6} & -\frac{1}{6} \\
\frac{1}{6} & \frac{2}{3} & \frac{1}{6} & -\frac{1}{6} & \frac{1}{3} & -\frac{1}{6} \\
\frac{1}{5} & \frac{1}{5} & -\frac{1}{3} & \frac{2}{6} & -\frac{1}{5} & \frac{1}{3} \\
-\frac{1}{6} & \frac{1}{3} & \frac{1}{6} & -\frac{2}{6} & \frac{1}{3} & -\frac{1}{6} \\
\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}
\end{bmatrix}.
\]

Substitute the matrix \( \{b_{ij}\} \) for the full model (35) and the matrix \( \{q_{ij}\} \) for the reduced model (36) into (25), we can obtain the following test statistic after simplification

\[
\tilde{F}_R^{(3)} = \sum_{i=1}^{6} \left( \sqrt{n_i} \sum_{j=1}^{6} (b_{ij} - q_{ij}) \frac{T_j}{\sqrt{n_j}} \right)^2
\]

(37)

= \sum_{i=1}^{3} \sum_{j=1}^{2} \left( T_{ij} - \frac{\sqrt{m_{ij}}}{2} \sum_{l=1}^{2} \frac{T_{il}}{\sqrt{n_{il}}} - \frac{\sqrt{m_{ij}}}{3} \sum_{m=1}^{3} \frac{T_{mj}}{\sqrt{n_{mj}}} + \frac{\sqrt{m_{ij}}}{6} \sum_{m=1}^{3} \sum_{l=1}^{2} \frac{T_{ml}}{\sqrt{n_{ml}}} \right)^2,
\]

where the relation between \( T_{ml}, n_{ml}, \) and \( T_{ij}, n_j \) are defined as before. The test statistic \( \tilde{F}_R^{(3)} \) in (37) is the same as that of (13) for testing the null interaction terms for the two-way layout ANOVA model by Chen and Chen (1998). The result is valid for the model with \( J \) levels of factor A and \( J \) levels of factor when interaction terms are present. Tables of critical values for testing \( H_0^3 : (\alpha \tau)_{ij} = 0 \) for equal cell sample sizes, \( n_{ij} = n \), are given by Chen and Chen (1998). When \( n_{ij} \) are not all equal, the \( p \)-value for testing \( H_0^3 \) is given by our simulation program in Appendix named REGTEST.SAS.

The technique of partial test can also be applied to Latin square design, repeated measures design and other designs. These works are under investigation by the authors.

4. Examples
4.1. Simple Linear Regression

The important special case of general linear model is the simple linear regression:

\[ Y_{ij} = \beta_0 + \beta_1 X_i + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i \]

where \( \beta_0 \) and \( \beta_1 \) are unknown parameters, the \( X_i \)'s are known constants and \( e_{ij} \) are independently distributed as \( N(0, \sigma_i^2) \). Based on our one-sample procedure we may express our model in terms of the form (6) as

\[
\begin{bmatrix}
\tilde{Y}_1 \\
\tilde{Y}_2 \\
\vdots \\
\tilde{Y}_k
\end{bmatrix} =
\begin{bmatrix}
1 & X_1 \\
1 & X_2 \\
\vdots & \vdots \\
1 & X_k
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1
\end{bmatrix} + S_k
\begin{bmatrix}
T_1/\sqrt{n_1} \\
T_2/\sqrt{n_2} \\
\vdots \\
T_k/\sqrt{n_k}
\end{bmatrix}.
\]

Our least-squares estimates of \( \beta_0 \) and \( \beta_1 \) are

\[
\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(\tilde{Y}_i - \bar{Y})}{\sum(X_i - \bar{X})^2}
\]

and

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},
\]

where \( \bar{X} = \sum X_i / k \), and \( \bar{Y} = \sum \tilde{Y}_i / k \).

Now the interests to test the null hypothesis \( H_0 : \beta_1 = 0 \). Using (13) the test statistic is

\[
\tilde{F} = \sum_{i=1}^k \left( \frac{\tilde{Y}_i - \bar{Y}}{S_k/\sqrt{n_i}} \right)^2,
\]

under \( H_0 : \beta_1 = 0 \), the test statistic reduces to the r.v.

\[
\tilde{F} = \sum_{i=1}^k \left( \sqrt{n_i} \sum_{j=1}^k \left( b_{ij} - \frac{1}{k} \right) \frac{T_j}{\sqrt{n_j}} \right)^2
\]

where

\[
b_{ij} = \frac{\sum_l X_l^2/k - \bar{X}(X_i + X_j) + X_iX_j}{\sum_l (X_l - \bar{X})^2}
\]

and one can reject \( H_0 \) if and only if the p-value of the test is smaller than the level of significance \( \alpha \).

4.2. Multiple Regression with Two Predictor Variables
The simplest type of regression model relating the response variable $Y$ to a predictor variable $X$ is the one discussed in Section 4.1. But, not all data sets are adequately described by a model whose expectation is a straight line. When two predictor variables are involved, the multiple regression model can be written as

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_{ij}, \quad i = 1, \ldots, k, \ j = 1, \ldots, n,$$

where $\beta_0$, $\beta_1$, and $\beta_2$ are unknown parameters, the $X_i$’s and $e_{ij}$ are defined as Section 1.

Based on our one-sample procedure as of Section 2 we may express our model in terms of the form (6) as

$$
\begin{bmatrix}
\tilde{Y}_1 \\
\tilde{Y}_2 \\
\vdots \\
\tilde{Y}_k
\end{bmatrix}
= 
\begin{bmatrix}
1 & X_{11} & X_{12} \\
1 & X_{21} & X_{22} \\
\vdots & \vdots & \vdots \\
1 & X_{k1} & X_{k2}
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2
\end{bmatrix}
+ S[k]
\begin{bmatrix}
T_1/\sqrt{n_1} \\
T_2/\sqrt{n_2} \\
\vdots \\
T_k/\sqrt{n_k}
\end{bmatrix}.
$$

The matrix form of this model is analogue to equation (7). Our least-squares estimates of $\beta_0$, $\beta_1$, and $\beta_2$ are given, respectively, by

$$\hat{\beta}_1 = \frac{(\sum \tilde{y}_i x_{i1})(\sum x_{i2}^2) - (\sum \tilde{y}_i x_{i2})(\sum x_{i1} x_{i2})}{(\sum x_{i1}^2)(\sum x_{i2}^2) - (\sum x_{i1} x_{i2})},$$

$$\hat{\beta}_2 = \frac{(\sum \tilde{y}_i x_{i2})(\sum x_{i1}^2) - (\sum \tilde{y}_i x_{i1})(\sum x_{i1} x_{i2})}{(\sum x_{i1}^2)(\sum x_{i2}^2) - (\sum x_{i1} x_{i2})},$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_{i1} - \hat{\beta}_2 \bar{X}_{i2},$$

where the lowercase letters denote deviations from sample mean values (i.e., $x_{i1} = (X_{i1} - \bar{X}_{i1})$, $x_{i2} = (X_{i2} - \bar{X}_{i2})$ and $\tilde{y}_i = (\tilde{Y}_i - \bar{Y})$), the uppercase letters denote, respectively, by $\bar{X}_{i1} = \sum X_{i1}/k$, $\bar{X}_{i2} = \sum X_{i2}/k$, and $\bar{Y} = \sum \tilde{Y}_i/k$.

Firstly, the interest is to test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. Using (13) the test statistic is

$$\tilde{F} = \sum_{i=1}^{k} \left( \frac{\tilde{Y}_i - \bar{Y}}{S[k]/\sqrt{n_i}} \right)^2.$$
under $H_0 : \beta_1 = \beta_2 = 0$, the test statistic reduces to the r.v.

$$\tilde{F} = \sum_{i=1}^{k} \left( \frac{1}{\sqrt{n_i}} \sum_{j=1}^{k} \left( b_{ij} - \frac{1}{k} T_j \right) \frac{T_j}{\sqrt{n_j}} \right)^2,$$

where $\{b_{ij}\} = X(X'X)^{-1}X'$ and

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{k1} & X_{k2} \end{bmatrix} = \begin{bmatrix} P'_1 \\ P'_2 \\ \vdots \\ P'_k \end{bmatrix}.$$

One can reject $H_0$ if and only if the p-value of the test is smaller than the level of significance $\alpha$.

Secondly, the interest is to test the partial null hypothesis $H'_0 : \beta_2 = 0$ (or $\beta_1 = 0$). Then, the reduced model is given by

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + e_{ij} = Q'_i \beta_R + e_{ij}, \quad i = 1, \ldots, k, \ j = 1, \ldots, n_i,$$

where $Q'_i = (1, X_{i1})$, $\beta'_R = (\beta_0, \beta_1)$. Based on our one-sample procedure one may express the reduced model in terms of (21) becomes a simple linear regression as

$$\begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \vdots \\ \tilde{Y}_k \end{bmatrix} = \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{k1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + S[k] \begin{bmatrix} T'_1 / \sqrt{n_1} \\ T'_2 / \sqrt{n_2} \\ \vdots \\ T'_k / \sqrt{n_k} \end{bmatrix},$$

or

$$\tilde{Y}_i = X_1 \beta_R + \epsilon'_i, \quad i = 1, \ldots, k,$$

where $\tilde{Y}_i$ and $\epsilon'_i$ are defined in (21)-(22) and $X_1 = \begin{bmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{k1} \end{bmatrix}$. Under $H'_0 : \beta_2 = 0$, the test statistic reduces to the r.v.

$$\tilde{F} = \sum_{i=1}^{k} \left( \frac{1}{\sqrt{n_i}} \sum_{j=1}^{k} (b_{ij} - q_{ij}) \frac{T_j}{\sqrt{n_j}} \right)^2.$$
where \(\{b_{ij}\}\) is the one as defined for the full model and \(\{q_{ij}\} = X_1(X_1'X_1)^{-1}X_1'\) for the reduced model. Analogous to the simple linear regression and the first part of this Section, one can reject \(H_0\) if and only if the p-value of the test is smaller than the level of significance \(\alpha\).

4.3. Numerical Example

Suppose we wish to compare four treatment levels of a protein source which are to be fed to pigs and their weight gains (in lbs) over a period of six months can be recorded. For illustrative purpose, we have the following data generated from one-way layout ANOVA model (17) according to some specified values of the parameter \(\mu = (2, 4, 6, 10), \sigma = (1, 9, 4, 4)\) and sample sizes \(n_i\) in the experiment having four protein levels as given in TABLE 2.

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generated Observations</td>
<td>2.06973</td>
<td>1.8590</td>
<td>6.5216</td>
<td>6.2274</td>
</tr>
<tr>
<td></td>
<td>0.60371</td>
<td>2.3574</td>
<td>8.4193</td>
<td>9.3458</td>
</tr>
<tr>
<td></td>
<td>1.53612</td>
<td>5.2234</td>
<td>7.0669</td>
<td>12.6645</td>
</tr>
<tr>
<td></td>
<td>2.75949</td>
<td>6.5739</td>
<td>9.9989</td>
<td>8.6539</td>
</tr>
<tr>
<td></td>
<td>2.07399</td>
<td>10.6586</td>
<td>9.4182</td>
<td>19.2552</td>
</tr>
<tr>
<td></td>
<td>2.44852</td>
<td>-4.7587</td>
<td>6.2097</td>
<td>7.6564</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13.8181</td>
<td>8.5116</td>
<td>15.1630</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.9535</td>
<td>4.4013</td>
<td>11.9067</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-11.2055</td>
<td></td>
</tr>
<tr>
<td>Sample Size</td>
<td>6</td>
<td>12</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Sample Mean</td>
<td>1.9153</td>
<td>3.7159</td>
<td>7.5684</td>
<td>11.3591</td>
</tr>
<tr>
<td>Sample S.D.</td>
<td>0.7629</td>
<td>7.1166</td>
<td>1.8581</td>
<td>4.3134</td>
</tr>
</tbody>
</table>

The purpose is to test the claim that all protein levels have no difference in terms of their mean weight gains, i.e. to test the null hypothesis \(H_0 : \mu_1 = \cdots = \mu_4\) against the alternative that not all mean weight gains are equal.
First of all, by using the modified Levene test to check the equality of all population variances, we have an $F$-statistic of 3.88 with a $p$-value of 0.0186. Each of these data from four treatment levels also passed Shapiro-Wilks normality test with a $p$-value of greater than .55. Therefore, at 5% level of significance, we reject the equality of population variances and we can apply the one-sample procedure in Section 2 for testing the hypothesis $H_0 : \mu_1 = \cdots = \mu_4$.

The one-way layout model (17) can be converted, by the technique of introducing dummy variables, to a full regression model as

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_{ij}$$

where the $\beta$’s are expressed in terms of $\mu$’s, i.e., $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$, $\beta_2 = \mu_3 - \mu_1$, $\beta_3 = \mu_4 - \mu_1$ and the dummy variables $X$’s are similarly defined in Section 3.2. The design (data) matrix for the full regression model is given by

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

The hypothesis $H_0 : \mu_1 = \cdots = \mu_4$ for ANOVA model is equivalent to the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ for the full regression model as discussed in Section 3.1.

The test statistic for $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is $\tilde{F}$ given in (15) or equivalently (18). By REGTEST.SAS, the test statistic $\tilde{F}$ has a value of 16.87 and a $p$-value of 0.033 (by 1000 simulation runs); the critical value of $\tilde{F}$ is 14.03 at $\alpha = 0.05$ level of significance. Therefore, one can reject the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (or reject $\mu_1 = \cdots = \mu_4$). The frequency distribution of $\tilde{F}$ of (15) based on 1000 runs is given in Graph 1.
Furthermore, if the interest is to examine whether protein levels 1 and 4 can stimulate the same mean weight gains on pigs, then it leads to testing the hypothesis $H_0^* : \mu_1 = \mu_4$ for ANOVA model. By regression technique, the ANOVA model under $H_0^*$ can be converted to a reduced regression model with $\beta_3 = 0$ as

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_{ij}$$

whose reduced design (data) matrix is given by

$$X_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$ 

The test statistic for a partial test $H_0^* : \beta_3 = 0$ (or equivalently $H_0^* : \mu_1 = \mu_4$) is $\tilde{F}_R$ given in (24). By the second part of REGTEST.SAS, the test statistic $\tilde{F}_R$ in (24) has a value of 8.28 and a $p$-value of 0.04 (by 1000 simulation runs); the critical value of $\tilde{F}_R$ in (25) is 7.23 at $\alpha = 0.05$ level of significance. Therefore, one can reject the hypothesis $\beta_3 = 0$ (or reject $\mu_1 = \mu_4$). The frequency distribution $\tilde{F}_R$ of (24) based on 1000 runs is given in Graph 2.

24
5. Summary and Discussion

Assuming a general linear regression model with unknown and possibly unequal error variances, a one-sample procedure is employed to derive statistical inference on point estimation and hypothesis testing on all slope parameters or a subset of regression parameters. The sampling procedure is to split up each single sample of size $n_i \geq 3$ at the $i^{th}$ data point of regressors into two portions; the first consists of $n_i - 1$ observations for a preliminary estimation and the second consists of the remaining one for the overall use in constructing a weighted sample mean response at each given data point. Then, the weighted sample mean response is used as a basis to derive least squares estimates for the regression parameters and the corresponding prediction equation. When the goal is to test the null hypothesis that all slope parameters are equal to zero, or equivalently, the mean response is free of all predictor variables, a quadratic test statistic (13) is proposed to measure the contribution of regressors due to the full model and it is used to test the null hypothesis. It has been
found that the proposed test statistic is a quadratic function of independent Student’s t r.v’s under the null hypothesis, and whose distribution is independent of all unknown population variances. Following the regression technique, the one-way layout fixed-effects analysis of variance model is a special case, and the test statistic for testing all null slope parameters becomes the one for testing the hypothesis of all null treatment effects proposed by Chen and Chen (1998).

To examine whether a subset of predictor variables plays an important role in a linear regression model we generally need to perform a partial test (24) for a subset of regression parameters (19). Under the null hypothesis, the partial test becomes a function of quadratic form (25) whose distribution is completely independent of all unknown parameters. Such a partial test can be used as a tool to build a better model if only one parameter is tested at a time in a stepwise manner. Furthermore, the partial test can be readily applied to the two-way layout analysis of variance model using regression technique in order to test the hypothesis of null row effects on one factor, the hypothesis of null column effects on another factor as well as the null interaction effects as studied by Chen and Chen (1998) in their single-stage analysis of variance model. When an individual parameter is of interest, the technique of hypothesis testing (24)-(25) can also be used. All of the proposed test statistics are linear or quadratic functions of independent Student’s t r.v.’s under their corresponding null hypotheses. These tests (13), (15), (18), (24), (25), (31), (33), and (37) of respective hypotheses can be carried through using critical valued or the p-value technique, where these values of a given test can be simulated by Monte Carlo simulation for small samples provided in the Appendix or approximated by a large sample approximation.

The one-sample procedure is a data-analysis procedure for a linear model when its error variances are unequal and unknown. All statistical inference based on the proposed statistical procedure becomes possible as the distribution of the test statistic is completely independent of the unknown variances. Under the same assumption on the random errors, Bishop (1978) proposed a Stein-type (1945) two-stage sampling procedure such that the
distribution of his proposed test statistic is also completely independent of the variances and that the power of the test can be controlled at a desirable level. However, the two-stage sampling procedure is a design-oriented procedure which must require additional observations at the second stage in order to complete its experiment. As a result, it may not be practical for the problem of data analysis in situations where data had already been collected or in the situations where the two-stage sampling experiment was terminated earlier than required due to budget limitation, time shortage, unavailability of additional samples, or other factors. Furthermore, the one-sample procedure can be readily applied in sampling from a data warehouse in the area of data mining. Whenever such situations happen the proposed one-sample procedure can be readily applied to these type of data without hurting the workability of the statistical inference.

The relative merits between the one-sample and two-stage sampling procedure studied by Chen (2001) revealed that (i) if \( \frac{S_i^2}{n_i} = \frac{S_j^2}{n_j} \) at all regressor’s data points, the one-sample and two-stage procedures are equivalent in terms of level and power consideration, (ii) if the actual sample observations obtained at each data point are more (less) than the required ones, the one-sample procedure has a larger (smaller) power than that of the two-stage procedure at the same level of significance, and (iii) in other situations, the one-sample procedure could have a power larger than, smaller than or equal to that of the two-stage procedure depending on actual samples and true error variances. Therefore, it is recommended that the one-sample procedure be used in a data-analysis whenever the sampling data are available, and in situations where the two-stage sampling procedure is interrupted before its completion.

6. Acknowledgement

This research was supported by National Science Council NSC92-2119-M-006-007 and NSC93-2118-M-006-008, Taiwan, 2003-2005.
REFERENCES


%GLOBAL _DISK_; %GLOBAL _PRINT_; 
%LET _PRINT_=OFF; %LET _DISK_=ON; 
options formdlim = '-';

data _null_; /* Read number of slope.*/
  infile "H:\mjwen\datax.txt" obs=1; /* parameters (p)(full) */
  input p; call symput('p',trim(left(p)));
  cards; run;

data datax; /* read predictors X_ij */
  infile "H:\mjwen\datax.txt" firstobs=3;
  input x1-x&p; cards; run;
  proc print; title 'X-design or data matrix for full model';
  run;

data _null_; /* Read number of slope. */
  infile "H:\mjwen\datax.txt" obs=2; /* parameters (q)(reduced)*/
  input p; call symput('p',trim(left(p)));
  cards; run;

data datax1; /* Read predictors X1_ij. */
  infile "H:\mjwen\datax.txt" firstobs=3;
  input x1-x&p; cards; run;
  proc print; title 'X1-design or data matrix for reduced model';
  run;

data _null_; /* Read no. of populations.*/
  infile "H:\mjwen\datay.txt" obs=1; /* (k) of response Y_ij -- */
  input k; call symput('k',trim(left(k)));

APPENDIX

/*-----------------------------------------------------------------------*/
/* REGTEST.SAS "On Testing A Subset Of Regression Parameters Under */
/* Heteroscedasticity" by Wen-Chen-Chen, 07/12/2004. */
/* This SAS program finds the p-value and the critical value of a test */
/* for a null hypothesis H0: B1=B2= ... =0, i.e., all slope parameters */
/* being equal to zero, or a subset of regression parameters being zero. */
/* The predictor data X_ij’s are read from file "datax.txt", where the */
/* first row is the number of slope parameters (p) for a full model, */
/* the second row is the number of slope parameters for a reduced model */
/* and beginning the third row are the data matrix {X_ij}. */
/* The response data Y_ij’s are read from file "datay.txt", where the */
/* first row is the number of populations or data points (k) and the */
/* second row contains the sample sizes (n_i) at each point, and the */
/* third row and thereafter are input values by triplets, (i j y) in a */
/* continuous manner. The sas program was originally written by Shun-Yi */
/* Chen and then modified by Ping Shan Chen, 05/12/2004. */
/*-----------------------------------------------------------------------*/

%GLOBAL _DISK_; %GLOBAL _PRINT_; 
%LET _PRINT_=OFF; %LET _DISK_=ON; 
options formdlim = '-';

data _null_; /* Read number of slope.*/
  infile "H:\mjwen\datax.txt" obs=1; /* parameters (p)(full) */
  input p; call symput('p',trim(left(p)));
  cards; run;

data datax; /* read predictors X_ij */
  infile "H:\mjwen\datax.txt" firstobs=3;
  input x1-x&p; cards; run;
  proc print; title 'X-design or data matrix for full model';
  run;

data _null_; /* Read number of slope. */
  infile "H:\mjwen\datax.txt" obs=2; /* parameters (q)(reduced)*/
  input p; call symput('p',trim(left(p)));
  cards; run;

data datax1; /* Read predictors X1_ij. */
  infile "H:\mjwen\datax.txt" firstobs=3;
  input x1-x&p; cards; run;
  proc print; title 'X1-design or data matrix for reduced model';
  run;

data _null_; /* Read no. of populations.*/
  infile "H:\mjwen\datay.txt" obs=1; /* (k) of response Y_ij -- */
  input k; call symput('k',trim(left(k)));

29
cards; run;
data y1; /* Read sample size \( n_i \). */
   infile "H:\mjwen\datay.txt" firstobs=2 obs=2;
   input n1-n&k; cards; run;
proc print; title 'The sample sizes \( n(1)-n(k) \)'; run;
%macro y1; /* Convert \( n_i \) to global */
data _null_; set y1; /* variable. */
   %do i=1 %to &k;
      call symput("n&i",trim(left(n&i)));
   %end;
%mend; %y1;
data datay; /* Read response \( Y_{ij} \) in a continuous way.*/
   infile "H:\mjwen\datay.txt" firstobs=3;
   input pop j y @@; cards; run;
proc sort; by pop; run; /* This statement is needed for safety. */
proc print; title 'The response \( Y(i,j) \)'; run;
data datay; set datay;
   sim=1; run;
%macro y2; /* Take first \( n(i)-1 \) obsn. */
data tmp1; set datay;
   %do i=1 %to &k;
      if pop=&i THEN IF j Le &&n&i-1;
   %end;
%mend; %y2;
proc means data=tmp1 nway noprint; /* Compute variance for. */
   class sim pop;
   var y;
   output out=out1 mean=mean var=var; run;
proc means data=out1 nway noprint; /* Find maximum variance */
   class sim; var var;
   output out=out2 max=max_var; run; /* maximum variance */
data ttt2; merge out2 out1; by sim; /* Merge variance with. */
proc print; run;
%macro y3;
data ttt2; set ttt2; /* Compute weights U, V. */
   %do i=1 %to &k;
      if pop=&i THEN DO;
         U=(1+SQRT((max_var/var-1)/(&&N&i-1)))/&&N&i;
         V=(1-SQRT((max_var/var-1)*(&&N&i-1)))/&&n&i;END;
   %end;
   keep sim pop var max_var U V;
%mend; %y3;
proc print; title 'The coefficients of the weighted means in (2)';
run;
DATA TMP1; MERGE DATAY TTT2;  /* Merge with original */
   BY SIM POP;  /* data set. */
%MACRO Y4;
DATA TMP1; SET TMP1;  /* Assign weights U, V.*/
   %do i=1 %to &k;
      if pop=&i THEN DO;
         IF J LT &&n&i THEN W=U;
         IF J EQ &&n&i THEN W=V;  END;
   %end;
%MEND; %Y4;
DATA TMP1; SET TMP1;
   YY=Y*W;  RUN;
PROC MEANS DATA=TMP1 NOPRINT;  /* compute Y~(i.), Y~(..) */
   CLASS SIM MAX_VAR POP;
   OUTPUT OUT=OUT1 SUM=MEAN;  RUN;
DATA OUT1; SET OUT1;
   IF (MAX_VAR NE .) AND (SIM NE .);  RUN;
PROC SORT DATA=OUT1; BY SIM MAX_VAR POP;  RUN;
DATA allMEAN; SET OUT1; IF (POP EQ .);
   KEEP SIM MAX_VAR POP MEAN;  RUN;
DATA YMEAN; SET OUT1; IF (POP NE .);
   KEEP SIM MAX_VAR POP MEAN;  RUN;  /* Print Y~(i.) for each pop.*/
proc print; title 'The weighted mean Y~(i.) in (4)'; run;
data degn(drop=i);
   do i=1 to &k;
      one=1; output; end;
proc iml;
   use YMEAN;
   read all var{MEAN} into ytild;
   read all var{sim max_var pop} into ytemp;
   use degn;
   read all var{one} into one;
   use datax;
   read all into xx;
   x=one||xx;
   px=x*inv(x'*x)*(x');  /* Compute X(X'X)^(-1)X'.*/
   create px from px;
   append from px;
   pred=px*ytild;
   /* Compute Y_hat. */
   pred=ytemp||pred;
   create pred from pred;
   append from pred;
31
quit;
data pred; set pred;
  rename col1=sim; rename col2=max_var;
  rename col3=pop; rename col4=mean;
/*proc print data=pred; title ’The predicted value for full model’;
run;*/
data out1; set allmean pred;
by sim max_var pop; run;
%macro y5;
DATA OUT1; SET OUT1; /* Compute F~={\[Yhat-Y~(..)\]}^2 */
   BY SIM MAX_VAR POP; /* *sqrt(N)/S[k]}^2 */
   IF (POP= .) THEN MN_ALL=MEAN/&K;
   RETAIN MN_ALL;
%do i=1 %to &k;
   if pop=&i THEN CHANGE=(MEAN-MN_ALL)**2*&&N&i/MAX_VAR;
%end;
%mend; %y5;
DATA OUT1; SET OUT1;
   IF (POP NE .); RUN;
PROC MEANS DATA=OUT1 NWAY NOPRINT; /* Compute F~ (save as FT). */
   CLASS SIM;
   VAR CHANGE;
   OUTPUT OUT=OUT2 SUM=FT_1; RUN;
proc print data=out2; title ’The value of the test F~ in (13)’; run;
data _null_; set out2;
   call symput('FT', trim(left(FT_1))); run;
data final; /* To store final results */
   input col1; /* from macro regr. */
cards;
run;
/*/---------------------------------------------------------------*/
/* Simulation to find null distribution of F~ in (15). */
/*/ The number of simulation is set to 1,000 or more runs. */
/*/---------------------------------------------------------------*/
%macro regr;
%do simu=1 %to 1000;
DATA T&simu;
%do i=1 %to &k;
  tn=NORMAL(0)/sqrt(2*RANGAM(0,(&&n&i-2)/2)/(&&n&i-2))/sqrt(&&n&i);
  nn=&&n&i;
  OUTPUT;
%end;
proc iml;
use T&simu;
read all var{tn} into tn;
read all var{nn} into nn;
use px;
read all into px;
ptn=(px - 1/&k)*tn; /* ff is a value of null */
ff=(nn')*(ptn##2); /* distributionn of F~ */
create dat&simu from ff;
append from ff;
quit;
proc append base=final data=dat&simu;
run;
%end;
%mend; %regr;
data final; set final; rename col1 = F_1;
data final; set final; pval= (F_1 ge &FT); run;
*ALPHA=.05, k=4, WITH A CRITICAL VALUE=14.03 for (15);
*ANOVA COMPUTED F~=16.87 in (13) with a p_value=0.033;
proc summary data=final nway;
var pval;
output out=result mean=p_val;
run;
proc print; title 'The p-value of the test F~ in (13)'; run;
*-----------------------------------------------------------------*;
* This SAS program finds the critical values of F~ distribution *
* in (15) with k treatments and v(i) degrees of freedoms. *
* The program returns the critical values of the upper 20%, 10%, *
* 5%,2.5% and 2%. *
*-----------------------------------------------------------------*;
options ls = 76 ps = 60;
data chart; set final;
gooption reset=all ftext=swiss;
proc gchart;
vbar F_1/midpoints=0 to 30 by 1 space=0;
where F_1 <= 30;
title 'The histogram of the test statictic F~ in (15)'; run;
data final; set final;
proc chart;
hbar F_1 / type=percent midpoints=0.0 to 30 by 1;
title 'The Frequency Distribution for the F~ Statistic in (15)'; run;
PROC UNIVARIATE DATA=final NOPRINT;
VAR F_1;
PROC PRINT;
TITLE 'PERCENTAGE POINTS FOR THE DISTRIBUTION OF F^ in (15)';
RUN;

proc iml;
use YMEAN;
read all var{MEAN} into ytilde;
read all var{sim max_var pop} into ytemp;
use degn;
read all var{one} into one;
use datax1;
read all into xx1;
x1=one||xx1;
qx1=x1*inv(x1'x1)*(x1');  /* Compute X1(X1'X1)^(-1)X1'. */
create qx1 from qx1;
append from qx1;
predstar=qx1*ytilde;  /* Compute Yhat-Yhat_star. */
predstar=ytemp||predstar;
create predstar from predstar;
append from predstar;
quit;

data predstar; set predstar;
  rename col1=sim; rename col2=max_var;
  rename col3=pop; rename col4=meanstar;
/*proc print data=predstar; title 'The predicted value for reduced model';
run;*/
data out11; merge pred predstar;
by sim max_var pop; run;
%macro y6;
DATA out11; SET out11; /* compute F_R~={[Yhat-Yhatstar)}]. */
%do i=1 %to &k;
  if pop=&i THEN CHANGE2=(mean-meanstar)**2*&&N&i/MAX_VAR;
%end;
%mend; %y6;
DATA OUT11; SET OUT11;
PROC MEANS DATA=OUT11 NWAY NOPRINT;
  CLASS SIM;
  VAR CHANGE2;
  OUTPUT OUT=OUT22 SUM=FR_1;
RUN;
PROC PRINT data=OUT22;
title 'The value of the test F_R~ in (24)'; run;
data _null_; set out22;
call symput('FR', trim(left(FR_1))); run;
data final2; /* To store final results */ input col1; /* from macro regr */ cards; run;/*--------------------------------------------------------------*/ /* Simulation to find full distribution of F_R~ in (25). */ /* The number of simulation is set to 1,000 (or more) runs. *//*--------------------------------------------------------------*/ %macro regr; %do simu=1 %to 1000; DATA T&simu; %do i=1 %to &k; tn=NORMAL(0)/sqrt(2*RANGAM(0,(&&n&i-2)/2)/(&&n&i-2))/sqrt(&&n&i); nn=&&n&i; OUTPUT; %end; proc iml; use T&simu; read all var{tn} into tn; read all var{nn} into nn; use px; read all into px; use qx1; read all into qx1; ptn=(px - qx1)*tn; /* ff is a value of null */ ff=(nn')*(ptn##2); /* distributionn of F_R~ */ create dat&simu from ff ; append from ff ; quit; proc append base=final2 data=dat&simu; run; %end; %mend; %regr; data final2; set final2; rename col1 = FR_1; data final2; set final2; pval2= (col1 ge &FR); run; *ALPHA=.05, k=4, WITH A CRITICAL VALUE=7.23 for (25); *ANOVA COMPUTED F_R~=8.28 in (24) with a p_value=0.04; proc summary data=final2 nway; var pval2; output out=result mean=p_val; run; proc print; title 'The p-value of the test F_R~ in (24)'; run;
* This SAS program finds the critical values of F_R~ distribution *;
* with k treatments and v(i) degrees of freedoms in (25). *;
* The program returns the critical values of the upper 20%, 10%, *;
* 5%, 2.5% and 2%. *;
*-------------------------------------------------------------------*;
data chart2; set final2;
goption reset=all ftext=swiss;
proc gchart;
  vbar FR_1/levels=15 space=0;
  where FR_1 <= 15;
  title 'The histogram of the test statistic F_R~ in (25)'; run;
data final2; set final2;
proc chart;
  hbar FR_1 / type=percent midpoints=0.0 to 15 by 1;
  title 'The Frequency Distribution for the F_R~ Statistic in (25)';
run;
PROC UNIVARIATE DATA=final2 NOPRINT;
  VAR FR_1;
  OUTPUT OUT=OUT7 PCTLPTS=80 90 95 97.5 98 PCTLPRE=P;
PROC PRINT;
  TITLE 'PERCENTAGE POINTS FOR THE DISTRIBUTION OF F_R~ in (25)';
RUN;
/*----------------------------------------------------------------*/
/* design matrix X, datax.txt, for one-way with 4 treatments */
/* 3 */
/* 2 */
/* 0 0 0 */
/* 1 0 0 */
/* 0 1 0 */
/* 0 0 1 */
/* response data sequence datay.txt */
/* 4 */
/* 6 12 8 8 */
/* 1 1 2.06973 2 1 1.8590 3 1 6.5216 4 1 6.2274 */
/* 1 2 0.60371 2 2 2.3574 3 2 8.4193 4 2 9.3458 */
/* 1 3 1.53612 2 3 5.2234 3 3 7.0669 4 3 12.6645 */
/* 1 4 2.75949 2 4 6.5739 3 4 9.9889 4 4 8.6539 */
/* 1 5 2.07399 2 5 10.6586 3 5 9.4182 4 5 19.2552 */
/* 1 6 2.44852 2 6 -4.7587 3 6 6.2097 4 6 7.6564 */
/* 2 7 13.8181 3 7 8.5116 4 7 15.1630 */
/* 2 8 12.9535 3 8 4.4013 4 8 11.9067 */
/* 2 9 -11.2055 */
/*
2 10 2.3602 */
/*
2 11 4.0908 */
/*
2 12 0.6602 */
------------- */