A STUDY OF HOMOGENEITY AMONG REGRESSION RELATIONSHIPS

by

John P. Robinson

Thesis submitted to the Graduate Faculty of the
Virginia Polytechnic Institute
in candidacy for the degree of

MASTER OF SCIENCE

in

STATISTICS

APPROVED:

Director of Graduate Studies

Dean of Applied Science and Business Administration

APPROVED:

Head of Department

Supervisor

September, 1958

Blacksburg, Virginia

TABLE OF CONTENTS

		Page
I.	INTRODUCTION	3
II.	HOMOGENEITY TESTS FOR REGRESSION RELATIONSHIPS	18
III.	HOMOGENEITY TESTS ON INDIVIDUAL PARAMETERS	21
IV.	CHOICE OF THE ZERO-POINT FOR X_{ij}	29
V.	TWO-WAY CLASSIFICATION	36
	A. RANDOMIZED BLOCKS	
	B. INTERACTION IN A TWO-WAY CLASSIFICATION	
VI.	NUMERICAL EXAMPLE	50
VII.	MORE THAN ONE CONCOMITANT VARIATE	59
VIII.	SUMMARY	64
IX.	ACKNOWLEDGEMENTS	66
X.	BIBLIOGRAPHY	67
XI.	VITA	68
XII.	APPENDIX	69

CHAPTER I

INTRODUCTION

In a linear regression model,

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p,$$

it is assumed that some linear relationship exists between a dependent variable Y and a set of independent or concomitant variables X_1, X_2 ... X_p. In any such regression study we are interested in the behavior of the dependent variable for a given or fixed set of values of the independent variables. On the basis of a sample, least squares estimates a, b_1, b_2 ... b_p, can be made of the parameters α, β_1, β_2 ... β_p, respectively. From these estimates, the experimenter can evaluate the relative influence of each of the independent variables on the dependent variable and can ascertain the precision of his estimates. If these estimates are obtained with a high degree of precision, he should find them quite useful in predicting future values of Y for given value(s) of the independent variate(s), within the range of X values in the sample.

For example, we might wish to investigate how the number of new TV sets that will be sold (Y) may be affected by the single concomitant variate, price of the sets (X). This investigation is initiated by taking samples of TV

dealers in different locations (e.g., Boston, Detroit, San Francisco, etc.) or under different economic conditions (e.g., in times of prosperity or recession). In addition to evaluating the results of these regression studies individually (e.g., for each city), it is often of interest to investigate the possibility that the regressions, say from each city, are the same or, in other words that one underlying relationship between price and quantity sold can explain the true physical situation for all samples taken (e.g., for all cities).

In many situations, it may be preferable to perform an analysis allowing for variations under two classifications simultaneously (e.g., each regression in each city can then be studied under each economic condition). Moreover, under such a plan we could examine an "interaction" effect (e.g., between locations and conditions) causing possible variation among the regressions. The general subject of regressions in an array analagous to a two-way classification together with an explanation of the "interaction" concept will be further treated in Chapter V.

In situations where it has been found that the regression relationship varies significantly among the groups, it is of further interest to ascribe the variation to one or several of the individual parameters

involved. Many tests of homogeneity (i.e., equality) among the regression parameters are available and will be found to be extremely practical. We will discuss and illustrate the practicality and scope of such tests with a somewhat simplified and artificial example:

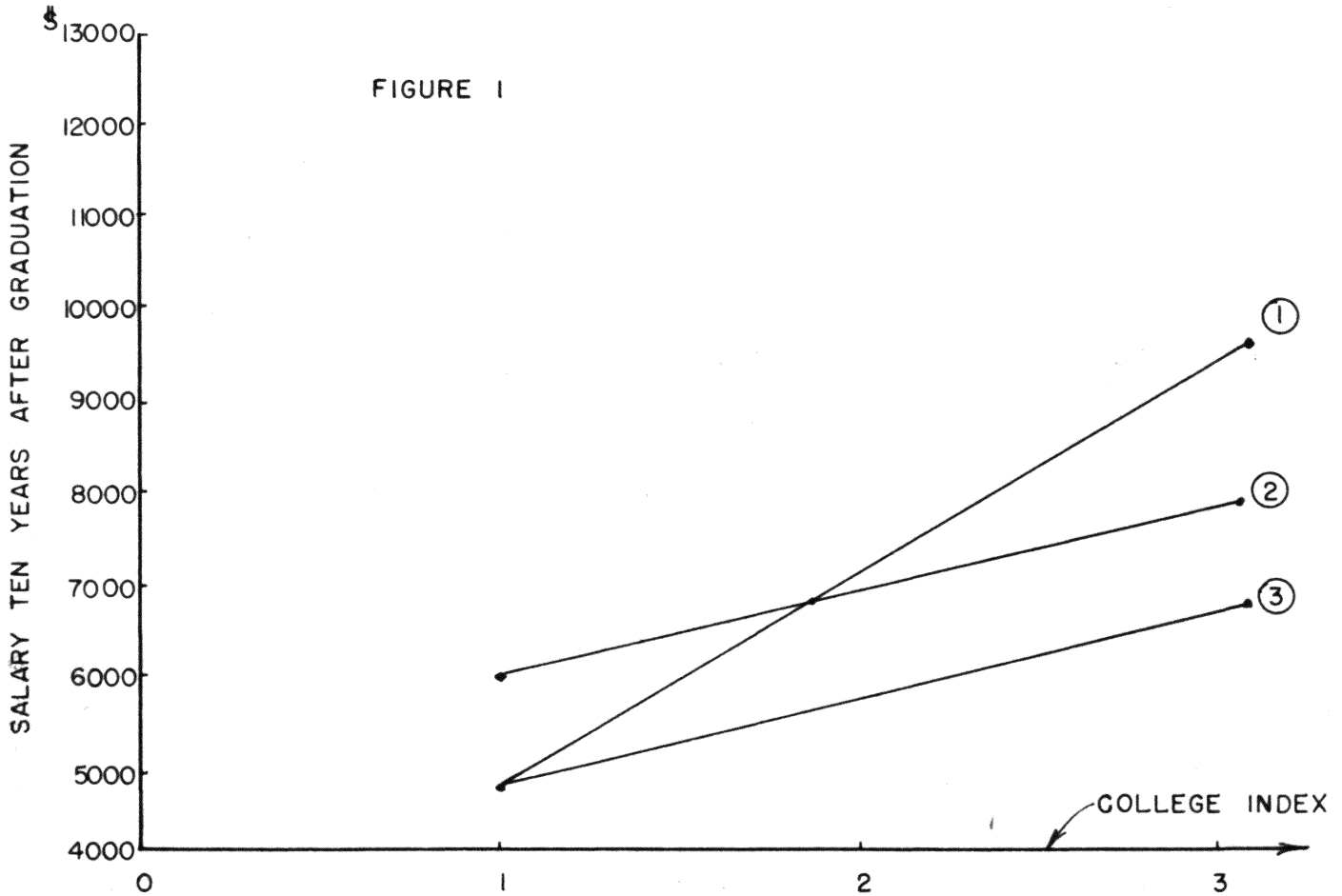
Suppose that we are about to make a regression study of the effect of a person's overall college index (or quality point average as a measure of academic achievement)[1] on his annual salary ten years after graduation. In order to study, in addition, the effect of college curricula on these salaries, we divide the sample of college graduates into three groups, according to programs taken in college: business, science and liberal arts. To control extraneous factors influencing our problem, as much as possible, we sample only those graduates who now work in fields directly related to their course of studies.

Say then that we expect, on the basis of previous information, that the three regression lines:

---

[1] For instance, let 3.0 college index be the equivalent of a straight A average, 2.0 be the equivalent of a B average and 1.0 be the equivalent of a C average.

FIGURE 1



HYPOTHETICAL REGRESSION LINES FOR A STUDY OF THE EFFECT OF A PERSON'S COLLEGE INDEX ON HIS SALARY TEN YEARS AFTER GRADUATION

$$(1)\quad \hat{Y}_B = a_B + b_B X \qquad \text{for business students}$$

$$(2)\quad \hat{Y}_S = a_S + b_S X \qquad \text{for science students}$$

$$(3)\quad \hat{Y}_A = a_A + b_A X \qquad \text{for liberal arts students}$$

will in general resemble those given in Figure 1. It should be noted that $\hat{Y}_B$ is the estimate of the true annual salary of business graduates, Y_B, for a given value of the concomitant variate, college index. Likewise, $\hat{Y}_S$ and $\hat{Y}_A$ are the estimates of this salary for science graduates and business graduates, respectively.

Let us further assume that, by the use of significance tests which will be presented in Chapter II of this paper, we have rejected the hypothesis that one regression line could explain the relation between college index and annual salary ten years after graduation for either business and science graduates, business and liberal arts graduates, or science and liberal arts graduates; that is, each regression line has been judged to be significantly different from any of the other lines.

We may expect, however, that the regression lines for science and liberal arts graduates will (as in Figure 1) be similar in one respect, that is, that they will be approximately parallel. We might then wish to test whether in fact they are parallel or, in other words, to test the

null hypothesis: $\beta_A = \beta_S$. If we fail to reject this hypothesis, two conclusions are possible. First, we can conclude that for science graduates and liberal arts graduates, college index has the same effect on annual salary ten years after graduation. Then, since the two regression lines are significantly different, we can also conclude that science graduates make significantly higher salaries ten years after graduation than do liberal arts graduates, regardless of academic achievement.

The reader who is familiar with statistical techniques will recognize that the above problem is frequently handled by the analysis of covariance. However, analysis of covariance assumes that the independent variate(s) exert the same influence on all values Y of the dependent variable for all groups (e.g., $\beta_1 = \beta_2 = \ldots = \beta_K$). The purpose of the analysis is then to test for differences in mean values of Y adjusted for the effect of the independent variate(s), since these effects are considered extraneous. The difference between the analysis of covariance and homogeneity tests of regressions is that the latter tests the assumption of equal coefficient (β) or coefficients (β_1, β_2 ... β_K) which the former assumes.

Now let us consider similarities between the hypothetical regression lines one and two and between lines one and three. We do not expect that either set of lines will be parallel as

in the case of science and liberal arts graduates. Are we to conclude that since neither pair is equivalent or parallel, that they have no similarity whatever? Or more precisely, do we infer that for no value of the concomitant variate (X) do the regression for business, science and liberal arts graduates have anything in common? From Figure 1, we see that we do not expect this; in fact, we expect the regression lines for business and science graduates to intersect at the point X = 1.8 and the lines for business and liberal arts graduates to intersect at the point X = 1. In terms of physical interpretation, business and science graduates who maintain what is approximately a B- average during college can expect to earn the same salary ten years after graduation, if actual results were similar to the ones in our fictitious example. The same is true for business and liberal arts graduates who maintain an even C average. We might then wish to test the hypothesis that the regression lines from our samples do intersect at these points.

If X = 1 were chosen as the origin of all X values, that is, if we subtracted 1.0 from each observed X, thus obtaining a new set of X*, we would have the two regression lines:

$$\text{i)} \quad \hat{Y}_B = a_B^* + b_B X^*$$

$$\text{ii)} \quad \hat{Y}_A = a_A^* + b_A X^*.$$

The test $\alpha_B = \alpha_A$ would then precisely be the test that the two regression lines intersect at the point X = 1 (X* = 0). In this case, the choice of X = 1, as the origin of our concomitant variate may be justified in that no student below this mark would graduate. However, in many, if not most, cases the choice of the origin is quite arbitrary.

We may then explore the possibility whether for some particular value(s) of the concomitant variate(s), all three regressions do not differ significantly. A test of this hypothesis would generally only be of interest, after we have rejected both of the hypotheses:

i) One regression for all data

ii) Equal slopes for all regression lines (one concomitant variate) or equal β effects for all regressions (more than one concomitant variate).

For example, we note from Figure 1 that we expect the three regression lines to be closest together between the points X = 1 and X = 2. If we could, by some procedure, find a value A of the concomitant, between X = 1 and X = 2, at which the lines have minimum separation,[1] and reject the hypothesis that the regression lines intersect at the point, then we must reject this hypothesis for all values of X. A

---

[1] The term "minimum separation" will be defined in Chapter IV.

further elaboration of this problem will be given in the discussion of models later in this chapter.

Let us review in more precise format, a step-by-step procedure to be used in a study of homogeneity for a set of k regressions:

STEP 1: Test the null hypothesis: Can one regression (α and β) be used for all the data? The test will be given in Chapter II. If we fail to reject this hypothesis, no further tests need be made. If we reject it, we proceed to

STEP 2: $H_o : \beta_1 = \beta_2 = \ldots = \beta_k$. The test will be given in Chapter III. If we fail to reject this hypothesis, we know that the concomitant variate has the same influence on the dependent variate for all k groups. Hence we are led to the conclusion that the dependent variate differs significantly for all groups (for this case, the method is, in essence, a covariance analysis). If we reject this hypothesis, we proceed to

STEP 3: where many possible tests may be made. If we expect the regression lines to intersect when $X = 0$, we can test $H_o : \alpha_1 = \alpha_2 = \ldots = \alpha_k$. If for each group the same set of X values are used or the mean value of X, $\bar{X}_i$, is the same, the test $H_o : \mu_1 = \mu_2 = \ldots = \mu_k$ will be of value. These

tests will be presented in Chapter III. Another appropriate test would be that the regression lines intersect at a point A. The choice of such a point, A, will be discussed below and also in Chapter IV. Other additional tests might be of interest, including a test of the hypothesis that the regression of group or sample means, $\bar{X}_i$, is linear. This test is given in Ostle [3].

It is mostly in connection with step 3 that we see the importance of selecting a proper model to represent the physical situation. The customary model used in regression studies is

$$Y_{ij} = \alpha_i + \beta_i X_{ij}, \qquad \begin{matrix} i = 1 \ldots k \\ j = 1 \ldots n_i \end{matrix}$$

where β_i is the slope of the true regression line and α_i is the value of Y at which the true regression line and the line X = 0 intersect for the i'th group. The main drawback of this model is that in most regression studies, the choice of the origin of all X values is completely arbitrary. There is no reason, therefore, to expect that each regression will have the same Y value at this point. Hence a test of the null hypothesis: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$, would be of little practical importance unless it is expected that the point X = 0 is indeed the true origin or intersection point for all regression lines. The model is perfectly acceptable, however, for making the tests outlined in steps 1 and 2 above.

Another model commonly used in regression analysis is

$$Y_{ij} = \mu_i + \beta_i(X_{ij} - \bar{X}_i) \qquad \begin{matrix} i = 1 \ldots k \\ j = 1 \ldots n_i \end{matrix}$$

where β_i is the same slope as in the model discussed in the above paragraph. In this model, however, the origin of all X values is shifted to the mean value, $\bar{X}_i$, of all X values in group i. The parameter, μ_i, then becomes the true mean of all Y values in this group. One advantage of this model is that the parameters, μ_i and β_i, are uncorrelated for all groups. For this reason, individual tests for each μ_i and β_i can be calculated independently and, incidently, much more easily than in the model discussed above in which α_i and β_i are correlated. Unless, however, the group means of the concomitant variate, $\bar{X}_i$, are equal, a test of the null hypothesis: $\mu_1 = \mu_2 = \ldots = \mu_k$ is of little useful consequence, since the differences among the μ_i could well be caused by differences among the $\bar{X}_i$, whose values are considered fixed. To alleviate, at least in part, the shortcomings of the two previous models, we may consider the following one:

$$Y_{ij} = \gamma_i + \beta_i(X_{ij} - A). \qquad \begin{matrix} i = 1 \ldots k \\ j = 1 \ldots n_i \end{matrix}$$

Under this model, the concomitant variate is adjusted to the same point of origin, A, for every group. This point may be chosen in such a way that, at X = A, the regressions have minimum separation, the parameters, γ_i, are then the values

of Y at this "ideal" X-origin. A rejection of the null hypothesis: $\gamma_1 = \gamma_2 = \ldots = \gamma_k$ would mean that for no value of the concomitant or independent variate do the regressions concur. The point $\overline{X}$, the mean of all X values in the study, is frequently chosen in lieu of A, since the latter is quite difficult to find and since the standard F test is only an approximation in the case where we estimate A from the same sample. A more elaborate discussion of this model together with a method of finding the point A will be the main consideration of Chapter IV.

Figures 2,3 and 4 point out the difference between the three models and the parameters α_i, μ_i, and μ_i, respectively.

The reader will note that we have been dealing entirely with the simple case of a single concomitant variate in our example in the step-by-step procedure and in the models just discussed. This has been done merely for purposes of illustration. For situations in which p concomitant variates are involved, the considerations illustrated in the example and the models are changed very slightly. The step procedure remains unchanged. Procedures for carrying out step 1 are discussed in Chapter II, however, procedures for carrying out steps 2 and 3 become much more complex, and will not be rigorously developed. With more than one concomitant variable, we may also wish to test the effects of certain concomitants or subgroups of concomitants for homogeneity.

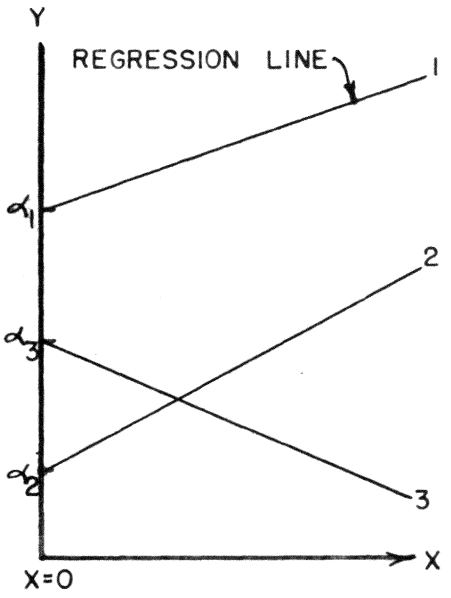


FIGURE 2

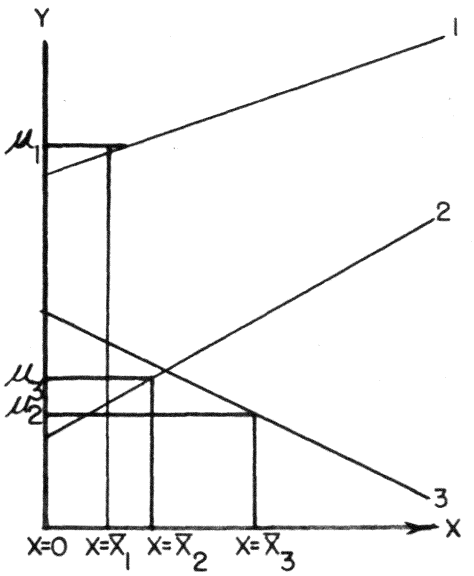


FIGURE 3

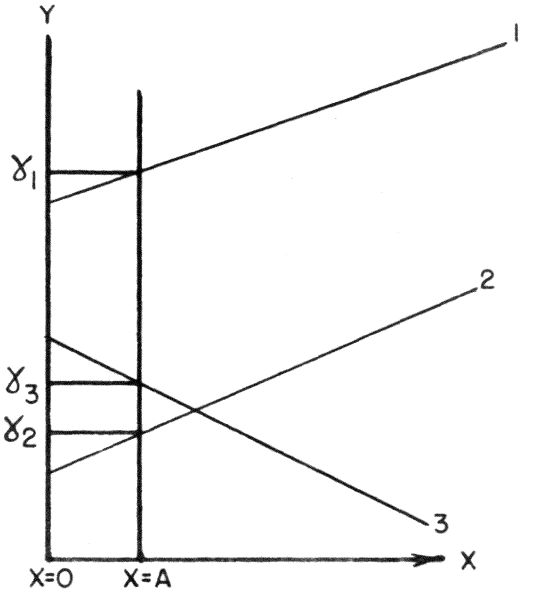


FIGURE 4

Most of the tests for regression lines in a one-way classification have been developed and can be found in Li [2] and Ostle [3]. For a more theoretical exposition of the techniques which were used to develop these tests, reference is made to Kempthorne [1].

Li (Chapter 19, pp. 344-378) develops these tests from a weighted sum of squares approach. Using the model,

$$Y_{ij} = \mu_i + \beta_i(X_{ij} - \bar{X}_i) + \epsilon_{ij},$$

the computing techniques for testing hypothesis of equal regression slopes (i.e., Ho: $\beta_1 = \beta_2 = \ldots = \beta_k$) and equal mean effects (i.e., Ho: $\mu_1 = \mu_2 = \ldots = \mu_k$) are given. An elaborate discussion of the model,

$$Y_{ij} = \gamma_i + \beta_i(X_{ij} - \bar{\bar{X}}) + \epsilon_{ij},$$

where the differences in values of the μ_i are adjusted to the value $\bar{\bar{X}}$, the mean value of all X values under study, is also presented. For this latter model, several extensions are presented (e.g., adjusted means in a randomized block experiment and extension to the relation between the analysis of covariance and factorial experiments).

Ostle (Chapter 6.11, pp. 133-138) sets up a breakdown of sums of squares in tabular form and gives the computing techniques in which the following hypotheses can be tested in a systematic manner:

i) Can one regression line be used for all the observations?

ii) Are the k regression slopes equal?

iii) Is the regression of means linear?

It can easily be shown that Ostle's test (ii) and the test given in Li's book are equivalent.

We will assume for all tests of significance to be given in the following chapters that the term ϵ_{ij} is normally and independently distributed with mean zero and common variance σ^2 for all observations.

It will be the purpose of this paper to present results found in [2] and [3] together with possible extentions, as in the case of the adjusted means concept described in step 3 of the step procedure. Special emphasis will also be placed on the subject of regressions in an array analagous to a two-way classification.

CHAPTER II

HOMOGENEITY TESTS FOR REGRESSION RELATIONSHIPS

The general question "Can one regression relation be used for all groups or samples?" (i.e., the double H_o: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$; $\beta_1 = \beta_2 = \ldots = \beta_k$, where the β_i are either scalars or vectors[1]) can be resolved by a least squares procedure which is the same regardless of the model postulated on the number of concomitant variates involved. For a description of this procedure, see for example Kempthorne [1]. For the case of k regressions arranged in any classification, with p independent variates in each regression, n_i observations in the i^{th} group, the test is

$$F_{(p+1)(k-1),\sum_{i=1}^{k}(n_i-1-p)} = \frac{\left[SSE_t - \sum_{i=1}^{k} SSE_i\right]/(p+1)(k-1)}{\sum_{i=1}^{k} SSE_i \Big/ \sum_{i=1}^{k}(n_i-1-p)}, \qquad (2.1)$$

where SSE_t is the residual or error sum of squares after fitting one general regression to all data, and SSE_i is the residual sum of squares after fitting a regression to

---

[1] There is one β_i for each group, if there is only one concomitant variate, and there is a set of regression coefficients, $\underline{\beta}_i$, for each group, if there are several concomitant variates.

the observations in group i. This F essentially tests whether the mean square error around one regression for all data is significantly larger than the mean square error around all individual regressions.

For the case of one concomitant variate, which is a special case of the general test given above, Ostle gives the test,

$$F_{2(k-1);\sum_{i=1}^{k}(n_i-2)} = \frac{\left\{SSy^2 - \frac{(SSxy)^2}{SSx^2} - \sum_{i=1}^{k}\left(Sy_i^2 - \frac{(Sx_iy_i)^2}{Sx_i^2}\right)\right\}/2(k-1)}{\sum_{i=1}^{k}\left(Sy_i^2 - \frac{(Sx_iy_i)^2}{Sx_i^2}\right)/\sum_{i=1}^{k}(n_i-2)} \qquad (2.2)$$

where for this formula and for all formulas given in this paper,

$$SSy^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}\right)^2}{\sum_{i=1}^{k} n_i}$$

$$SSx^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}\right)^2}{\sum_{i=1}^{k} n_i}$$

$$SSxy = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}Y_{ij} - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}\right)\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}\right)}{\sum_{i=1}^{k} n_i} \qquad (2.3)$$

$$Sy_i^2 = \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{(\sum_{j=1}^{n_i} Y_{ij})^2}{n_i}$$

$$Sx_i^2 = \sum_{j=1}^{n_i} X_{ij}^2 - \frac{(\sum_{j=1}^{n_i} X_{ij})^2}{n_i}$$

(2.3) CONT

$$Sx_iy_i = \sum_{j=1}^{n_i} X_{ij}Y_{ij} - \frac{(\sum_{j=1}^{n_i} X_{ij})(\sum_{j=1}^{n_i} Y_{ij})}{n_i} .$$

In this and similar relations throughout this thesis, lower case letters will indicate deviations from the means; for example,

$$x_{ij} = X_{ij} - \bar{X}_i$$

and (2.4)

$$y_{ij} = Y_{ij} - \bar{Y}_i .$$

CHAPTER III

HOMOGENEITY TESTS ON INDIVIDUAL PARAMETERS

The tests discussed in this chapter will be those outlined in steps 2 and 3 of the procedure presented in Chapter I. More specifically, we are concerned with tests of homogeneity of the parameters α_i, μ_i, and β_i in the two models:

1) $Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij}$, and

2) $Y_{ij} = \mu_i + \beta_i(X_{ij}-\bar{X}_i) + \epsilon_{ij} = \mu_i + \beta_i x_{ij} + \epsilon_{ij}$,

where for these models and all models in this paper, the term ϵ_{ij} is assumed to be normally, independently distributed with mean zero and variance σ^2. These tests will be derived by the well known general linear hypothesis technique which is outlined in the appendix. A very short description of the technique is as follows:

We start by setting up a "model" for all observations in terms of parameters to be estimated. The parameters are then estimated in such a way that the error variance becomes a minimum. An arbitrary hypothesis matrix, denoting the linear functions of the parameters which we wish to test for significance, is defined and the sum of squares due to this hypothesis is evaluated. The mean squares due to the hypothesis and due to error are found by dividing their respective sum of squares or quadratic forms by the appro-

priate number of degrees of freedom. If, by the use of the F-statistic, the mean square due to the hypothesis is judged to be significantly larger than the mean square due to error, then the corresponding hypothesis is rejected.

Consider the model,

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij}.$$

For the general case of k such regressions, n_i observations for each regression, the least squares equations are:

$$A'A\hat{\xi} = A'Y, \text{ i.e.,}$$

$$\begin{bmatrix} n_1 & \sum_{j=1}^{n_1} X_{1j} & 0 & 0 & \cdots & 0 & 0 \\ \sum_{j=1}^{n_1} X_{1j} & \sum_{j=1}^{n_1} X_{1j}^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & n_2 & \sum_{j=1}^{n_2} X_{2j} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & n_k & \sum_{k=1}^{n_k} X_{kj} \\ 0 & 0 & 0 & 0 & \cdots & \sum_{k=1}^{n_k} X_{kj} & \sum_{k=1}^{n_k} X_{kj}^2 \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \\ a_2 \\ \vdots \\ a_k \\ b_k \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \sum_{j=1}^{n_1} X_{1j} Y_{1j} \\ \sum_{j=1}^{n_2} Y_{2j} \\ \vdots \\ \sum_{j=1}^{n_k} Y_{kj} \\ \sum_{j=1}^{n_k} X_{kj} Y_{kj} \end{bmatrix}$$

(3.1)

The matrix, A'A, is non-singular and its inverse, $(A'A)^{-1}$ is:

$$\begin{bmatrix}
\frac{1}{n_1}+\frac{\bar{X}_1^2}{Sx_1^2} & -\frac{\bar{X}_1}{Sx_1^2} & 0 & 0 & \cdots & 0 & 0 \\
-\frac{\bar{X}_1}{Sx_1^2} & \frac{1}{Sx_1^2} & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \frac{1}{n_2}+\frac{\bar{X}_2^2}{Sx_2^2} & -\frac{\bar{X}_2}{Sx_2^2} & \cdots & 0 & 0 \\
0 & 0 & -\frac{\bar{X}_2}{Sx_2^2} & \frac{1}{Sx_2^2} & \cdots & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\
0 & 0 & 0 & 0 & \cdots & \frac{1}{n_k}+\frac{\bar{X}_k^2}{Sx_k^2} & -\frac{\bar{X}_k}{Sx_k^2} \\
0 & 0 & 0 & 0 & \cdots & -\frac{\bar{X}_k}{Sx_k^2} & \frac{1}{Sx_k^2}
\end{bmatrix} \qquad (3.2)$$

where Sx_i^2 is defined in equations 2.3 and

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}$$

The solution of the least squares equations is:

$$\hat{\underline{\xi}} = (A'A)^{-1}A'y = \begin{bmatrix} a_1 \\ b_1 \\ \vdots \\ a_k \\ b_k \end{bmatrix} = \begin{bmatrix} \bar{Y}_1 - \dfrac{Sx_1y_1}{Sx_1^2}\bar{X}_1 \\ \dfrac{Sx_1y_1}{Sx_1^2} \\ \vdots \\ \bar{Y}_k - \dfrac{Sx_ky_k}{Sx_k^2}\bar{X}_k \\ \dfrac{Sx_ky_k}{Sx_k^2} \end{bmatrix} \qquad (3.3)$$

where Sx_1y_1 is defined in equations 2.3 and,

$$\bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} .$$

The error sum of squares for this model (see appendix) is,

$$Y'Y - \hat{\underline{\xi}}'A'Y$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k}\left[\frac{(\sum_{j=1}^{n_i} Y_{ij})^2}{n_i} - b_i\left(\frac{\sum_{j=1}^{n_i} X_{ij}\sum_{j=1}^{n_i} Y_{ij}}{n_i} - \sum_{j=1}^{n_i} X_{ij}Y_{ij}\right)\right],$$

which reduces to

$$\sum_{i=1}^{k}\left\{Sy_i^2 - \frac{Sx_iy_i}{Sx_i^2}\right\} = \sum_{i=1}^{k} SSE_i \qquad (3.4)$$

with $\sum_{i=1}^{k} n_i - 2k$ degrees of freedom. This is the same error sum of squares as in equations 2.1 and 2.2.

1. To test for equality of slopes among the k regressions, we set up $H_o : \beta_1 = \beta_2 = \ldots = \beta_k$, which in matrix form is,

$$C\underline{\xi} = \underline{0}\ ,\ \text{i.e.,}$$

$$\underset{2(k-1)xk}{\begin{bmatrix} 0&1&0&-1&0&0&\ldots&0&0 \\ 0&1&0&0&0&-1&\ldots&0&0 \\ \cdot&\cdot&\cdot&\cdot&\cdot&\cdot&&\cdot&\cdot \\ \cdot&\cdot&\cdot&\cdot&\cdot&\cdot&&\cdot&\cdot \\ \cdot&\cdot&\cdot&\cdot&\cdot&\cdot&&\cdot&\cdot \\ 0&1&0&0&0&0&\ldots&0&-1 \end{bmatrix}} \underset{kx1}{\begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \\ \alpha_3 \\ \beta_3 \\ \vdots \\ \alpha_k \\ \beta_k \end{bmatrix}} = \underset{2(k-1)x1}{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}} \qquad (3.1.1)$$

The sum of squares due to this hypothesis is (see appendix)

$\hat{\underline{\xi}}'C'[C(A'A)^{-1}C']^{-1}C\hat{\underline{\xi}} =$

$$(b_1-b_2)^2(Sx_1^2+Sx_2^2) + (b_1-b_3)^2(Sx_1^2+Sx_3^2) + \ldots + (b_1-b_k)^2(Sx_1^2+Sx_k^2)$$

where $(A'A)^{-1}$ is shown in equation 3.2. This sum of squares reduces to

$$b_1^2Sx_1^2+b_2^2Sx_2^2+\ldots+b_k^2Sx_k^2 - \frac{(b_1Sx_1^2+b_2Sx_2^2+\ldots+b_kSx_k^2)^2}{Sx_1^2 + Sx_2^2 + \ldots + Sx_k^2}$$

$$= \frac{(Sx_1y_1)^2}{Sx_1^2} + \frac{(Sx_2y_2)^2}{Sx_2^2} + \ldots + \frac{(Sx_ky_k)^2}{Sx_k^2} - \frac{(\sum_{i=1}^{k} Sx_iy_i)^2}{\sum_{i=1}^{k} Sx_i^2} \qquad (3.1.2)$$

This sum of squares has, by step 8 in the appendix, (k-1) degrees of freedom, since there are (k-1) linearly independent rows in C. Since, under the null hypothesis, sum of squares, 3.1.2 and 3.4, are each distributed as independent Chi-squares, the test for homogeneity of regression slopes

becomes

$$F_{k-1;\sum_{i=1}^{k} n_i - 2k} = \frac{\left\{ \sum_{i=1}^{k} \left\{ \frac{(Sx_i y_i)^2}{Sx_i^2} \right\} - \frac{\left(\sum_{i=1}^{k} Sx_i y_i\right)^2}{\sum_{i=1}^{k} Sx_i^2} \right\} \Big/ k-1}{\left\{ \sum_{i=1}^{k} \left\{ Sy_i^2 - \frac{(Sx_i y_i)^2}{Sx_i} \right\} \right\} \Big/ \sum_{i=1}^{k} n_i - 2k} \qquad (3.1.3)$$

2. For testing equality of Y intercepts, the $H_o: \alpha_1 = \alpha_2 = \ldots = \alpha_k$ assumes the form,

$$C\underline{\xi} = \underline{0}$$

or

$$\underset{2(k-1)x2k}{\begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ 1 & 0 & 0 & 0 & 0 & 0 & \ldots & -1 & 0 \end{bmatrix}} \underset{2kxl}{\begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \\ \alpha_3 \\ \beta_3 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_k \\ \beta_k \end{bmatrix}} = \underset{2(k-1)xl}{\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 0 \end{bmatrix}} \qquad (3.2.1)$$

The sum of squares due to this hypothesis is

$$\hat{\underline{\xi}}' C' [C(A'A)^{-1} C'] C \hat{\underline{\xi}},$$

which using the definition of $(A'A)^{-1}$ given in 3.2 and expanding becomes,

$$\frac{(a_1 - a_2)^2}{v_1 + v_2} + \frac{(a_1 - a_3)^2}{v_1 + v_3} + \ldots + \frac{(a_1 - a_k)^2}{v_1 + v_k},$$

where

$$(3.2.a)\quad v_i = \frac{1}{n_i} + \frac{\bar{X}_i^2}{Sx_i^2} = \frac{Sx_i^2 + n_i X^2}{n_i Sx_i^2} .$$

This sum of squares, similar to the case of the test for equal β_i's, reduces to

$$\frac{a_1^2}{v_1} + \frac{a_2^2}{v_2} + \cdots + \frac{a_k^2}{v_3} - \frac{\left[\frac{a_1}{v_1} + \frac{a_2}{v_2} + \cdots + \frac{a_k}{v_k}\right]^2}{\frac{1}{v_1} + \frac{1}{v_2} + \cdots + \frac{1}{v_k}}$$

$$= \sum_{i=1}^{k} \frac{a_i^2}{v_i} - \frac{\left[\sum_{i=1}^{k} \frac{a_i}{v_i}\right]^2}{\sum_{i=1}^{k} \frac{1}{v_i}} \qquad (3.2.2)$$

Since again there are (k - 1) linearly independent rows in C_1 this sum of squares is distributed as $\chi^2\sigma^2$ with (k - 1) degrees of freedom. The error sum of squares and the sum of squares due to the null hypothesis are independent, therefore the test becomes

$$F_{k-1;\ \sum_{i=1}^{k} n_i - 2k} = \frac{\left\{ \sum_{i=1}^{k} \frac{a_i^2}{v_i} - \frac{\left[\sum_{i=1}^{k} \frac{a_i}{v_i}\right]^2}{\sum_{i=1}^{k} \frac{1}{v_i}} \right\} \Big/ k - 1}{\left(\sum_{i=1}^{k} Sy_i^2 - \frac{(Sx_i y_i)^2}{Sx_i^2} \right) \Big/ \sum_{i=1}^{k} n_i - 2k} \qquad (3.2.3)$$

It should be noted that sum of squares 3.1.2 and 3.2.2 will not add up to the numerator of the top expression of the F-ratio given by equation 2.2. This is due to the fact that

each (α_i, β_i)-pair is correlated.

3. For the model,

$$Y = \mu_i + \beta_i x_{ij} + \varepsilon_{ij},$$

the test for equality of all β's is given by equation 3.1.3. The sum of squares due to $H_o: \mu_1 = \mu_2 = \ldots = \mu_k$ is given in Li (Chapter 19.1), and is the usual sum of squares used in the numerator of the F-test in a one-way analysis of variance. The error sum of squares for this model is the same as given in equation 3.4. The resulting test statistic is:

$$F_{k-1;\ \sum_{i=1}^{k} n_i - 2k} = \frac{\sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{\bar{Y}})^2 / (k-1)}{\sum_{i=1}^{k} \left\{ Sy_i^2 - \frac{(Sx_i y_i)^2}{Sx_i^2} \right\} \Big/ \sum_{i=1}^{k} n_i - 2k} \qquad (3.3.1)$$

where

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}}{\sum_{i=1}^{k} n_i}$$

and

$$\bar{Y}_i = \frac{\sum_{i=1}^{n_i} Y_{ij}}{n_i} .$$

4. As discussed in Chapter 1, the test of equality of all μ_i's under the assumption that all β_i's are equal is the conventional one in a standard analysis of covariance.

CHAPTER IV

CHOICE OF ZERO-POINT FOR X_{ij}

The model,

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij},$$

has one severe limitation as discussed previously. If the origin of the concomitant variate is shifted (e.g., is measured in centigrades instead of Fahrenheit when the concomitant variate is temperature), the contrast among the α_i may be drastically changed. In fact, two regression lines of unequal slope will always intersect in one point. Further, if the measuring scale is arbitrary (as it often is), we may move the origin by defining X to be zero at this point, which, in the case of two regressions, would make $\alpha_1 = \alpha_2$. In the case of three or more regressions, we may choose, as the origin, that value of X at which the regression lines have minimum separation.

Our first attempt would be to find this point A as a function of the parameters only, i.e.

$$A = f(\alpha,\beta):$$

and to estimate A by

$$\hat{A} = f(\hat{\alpha},\hat{\beta}).$$

We may, for instance obtain, the mean of the expressions,

$$E(Y_{ij}) = \alpha_i + \beta_i X_{ij},$$

for each value of X_{ij} and form the squares of the distances from this mean for each of the k regression lines,

$$d^2 = \sum_{i=1}^{k} [E(Y_{ij}) - \frac{1}{k} \sum_{i=1}^{k} E(Y_{ij})]^2;$$

and then find that value of X which minimizes d^2. But this procedure would attach equal importance to each of the regression lines. In practice we may have only few observations on one regression line, and many on another, thus some weighting procedure seems more appropriate. We recommend choosing as $\hat{A}$ that point which minimizes the sums-of-squares due to the hypothesis that all intercepts are equal. This relation is presented in (4.2). The reason for choosing such an origin is that, if we are led to rejection of the hypothesis of equal γ_i's in the relation

$$E(Y_{ij}) = \gamma_i + \beta_i (X_{ij} - \hat{A}),$$

then we must reject the hypothesis of equality of γ_i^* in the relation

$$E(Y_{ij}) = \gamma_i^* + \beta_i (X_{ij} - A^*)$$

for any choice $A^* = \hat{A}$.

A simpler problem will be studied first. Assume that there exists one point (the same for all groups), X = A, at which the concomitant variate has the same influence on the observations in all groups. This point may be the mean of all observations of the concomitant variate (i.e.

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m_i} X_{ij}}{\sum_{i=1}^{k} n_i}$$ as discussed in Li, or better still, a

physical zero point (e.g. time = 0). It should be understood, however, that this point must be clearly definable in terms of the physical situation. With this definition of A, we set up the model,

$$Y_{ij} = \gamma_i + \beta_i (X_{ij} - A) + \epsilon_{ij}.$$

By straightforward least squares procedures, we may estimate

1. $b_i = \frac{Sx_i y_i}{Sx_i^2}$

2. $\hat{\gamma}_i = \bar{Y}_i - b_i(\bar{X}_i - A).$

It can be then readily shown that:

$$\text{var } (b_i) = \frac{\sigma^2}{Sx_i^2}$$

$$\text{var } (\hat{\gamma}_i) = [\frac{1}{n} + \frac{(\bar{X}_i - A)^2}{Sx_i^2}]\sigma^2$$

$$\text{and covar } (\hat{\gamma}_i, b_i) = -\frac{(\bar{X}_i - A)}{Sx_i^2}\sigma^2$$

which are the elements of $(A'A)^{-1}$. As in Chapter III, the sum of squares due to error is as in equation (3.4):

$$\sum_{i=1}^{k}\left\{Sy_i^2 - \frac{(Sx_iy_i)^2}{Sx_i^2}\right\} .$$

The test for equality of all β_i is given in equation 3.1.3.

The sum of squares due to the null hypothesis: $\gamma_1 = \gamma_2 = \cdots \gamma_k$ can be found by the heuristic argument of weighted regression.[1] The sum of squares due to H_o: $\alpha_1 = \alpha_2 = \cdots = \alpha_k$ (equation 3.2.2) is :

$$\sum_{i=1}^{k}\frac{a_i^2}{v_i} - \frac{\left(\sum_{i=1}^{k}\frac{a_i}{v_i}\right)^2}{\sum_{i=1}^{k}\frac{1}{v_i}}$$

where v_i is defined in 3.2.a. Analogously, the sum of squares due to the hypothesis of equal γ_i is

$$\sum_{i=1}^{k} z_i\hat{\gamma}_i^2 - \frac{\left[\sum_{i=1}^{k} z_i\hat{\gamma}_i\right]^2}{\sum_{i=1}^{k} z_i} \qquad (4.1)$$

[1] This formula was verified by the general linear hypothesis technique.

where the weighting factor, z_i is:

$$\frac{\frac{1}{n_i}}{\frac{1}{n_i} + \frac{(\bar{X}_i - A)^2}{Sx_i^2}} = \frac{n_i Sx_i^2}{Sx_i^2 + n_i(\bar{X}_i - A)^2} \quad .$$

If A is not known by physical considerations, we may place it at the point of minimum separation of the regression lines, which may be defined as follows:

Minimize, with respect to A, the sum of squares due to the null hypothesis of equal γ_i as stated above. Let,

$$Z = \sum_{i=1}^{k} z_i$$

$$(Z.\hat{\gamma}.) = \sum_{i=1}^{k} z_i \hat{\gamma}_i$$

$$\text{and } Q_i = \frac{\hat{\gamma}_i^2 - 2\hat{\gamma}_i (Z.\hat{\gamma}.) Z - (Z.\hat{\gamma}.)}{Z} \quad .$$

Differentiating (4.1) with respect to A and setting the derivative equal to zero, we obtain the following expression:

$$(4.2) \qquad \sum_{i=1}^{k} [b_i \left\{ z_i \hat{\gamma}_i - \frac{z_i}{Z}(Z.\hat{\gamma}.) \right\} + \frac{\bar{X}_i - \hat{A}}{Sx_i^2} z_i^2 Q_i] = 0$$

which we may solve for $\hat{A}$ by some iterative procedure (noting that $\hat{A}$ is contained in the $z_i, \hat{\gamma}_i$ and $(Z.\hat{\gamma}.)$. As a first guess of the value $\hat{A}$ satisfying (4.2), we may take $\bar{\bar{X}}$; then we can take a value somewhat smaller and larger and find out in which direction the correct solution differs from

our first guess. Continuing this way, we may find the appropriate value to any degree of accuracy.

After the determination of $\hat{A}$ satisfying (4.2), we may test the hypothesis: $\gamma_1 = \gamma_2 = \ldots = \gamma_k$ by an F-test with k-1 and $\sum_{i=1}^{k} n_i - 2k$ degrees of freedom, using $\frac{1}{k-1}$ times expression (4.1) as the numerator and $\frac{1}{\sum_{i=1}^{k} n_i - 2k}$ times expression (3.4) as denominator. It should be noted, however, that this test is not exact for, it can be seen from (4.2), the value $\hat{A}$ satisfying this relation is a function of the observations, Y_{ij}, and not only of the fixed concomitant variate, X_{ij}. This function is very complex and can be stated in implicit form only. The distribution of the statistic (4.1), is Chi-square only if the whole expression $(X_i - A)$ is fixed; to obtain an exact test, we would have to take into consideration the very difficult distribution of the random variable $\hat{A}$. Thus, if we perform the F-test recommended above, we treat $\hat{A}$ as if it were a concomitant variate, hence the approximate nature of this test.

If, however, we obtain the estimate of the ideal origin, $\hat{A}$, from one sample and use this estimate in a new sample, then A will be fixed for the new sample. Hence the proposed F-test will be exact.

Rejection of the above hypothesis indicates that, however, we may choose the arbitrary origin of X, the test of equal intercepts has to be rejected.

Tocher [5] gives the solution to a similar problem. Given that the regression line for the i'th group intersects the X-axis at a value $X_{0i} = \frac{-\alpha_i}{\beta_i}$, he then tests the null hypothesis: $X_{0i} = X_{0j}$ for all $i \neq j$. While his procedure in finding this test is similar to the one given above, it will be noted that his problem is different in that his hypothetical point X_0 is the common intersection point of all regression lines on the X-axis. Tocher's X_0 will be the same as A, above, only in the exceptional case when all regression lines intersect on the X-axis, which is too special for our present considerations. Williams [6] discusses further extensions of the special case studied by Tocher.

CHAPTER V

TWO-WAY CLASSIFICATION

A. Randomized Blocks

It frequently happens in any experiment which is analyzed by an analysis of variance technique that the pooling of the entire sample introduces a large variance due to the fact that some extraneous factors were left uncontrolled. To insure greater homogeneity, we usually subdivide the whole sample into smaller subsamples or "blocks". If every treatment is applied at least once to a unit within each block, the analysis is performed by the familiar "randomized block" technique.

A similar situation can also arise in our problem. In the illustration presented in Chapter I, some of the graduates may work in an urban area, where salaries are generally higher, some may work in rural communities, and others in mixed communities. Again, we would wish to study the influence of college performance on salaries, but we would now consider a two-way classification as shown in Figure 5.

There is no reason to assume interaction between our two classifications (curricula and community size ten years later) so that we would perform the analysis in the following manner, which we will illustrate for a 3 x 3 classification:

	URBAN	MIXED	RURAL
BUSINESS	$E(Y_{11m})=\mu_{11}+\beta_{11}x_{11m}$	$E(Y_{12m})=\mu_{12}+\beta_{12}x_{12m}$	$E(Y_{13m})=\mu_{13}+\beta_{13}x_{13m}$
SCIENCE	$E(Y_{21m})=\mu_{21}+\beta_{21}x_{21m}$	$E(Y_{22m})=\mu_{22}+\beta_{22}x_{22m}$	$E(Y_{23m})=\mu_{23}+\beta_{23}x_{23m}$
LIBERAL ARTS	$E(Y_{31m})=\mu_{31}+\beta_{31}x_{31m}$	$E(Y_{32m})=\mu_{32}+\beta_{32}x_{32m}$	$E(Y_{33m})=\mu_{33}+\beta_{33}x_{33m}$

FIGURE 5

MODEL FOR REGRESSIONS IN A TWO-WAY CLASSIFICATION

We will number our cells as in Figure 5 and in order to study the particular effects of treatments (curricula) adjusted for block (community size) effects, we redefine the model

$$Y_{ijm} = \mu_{ij} + \beta_{ij}(X_{ijm} - \bar{X}_{ij}) + \epsilon_{ij}, \quad \begin{matrix} i = 1, 2, 3 \\ j = 1, 2, 3 \\ m = 1 \ldots n_{ij} \end{matrix}$$

to be

$$Y_{ijm} = \mu + \mu_{i.} + \mu_{.j} + (\beta + \beta_{i.} + \beta_{.j})(X_{ijm} - \bar{X}_{ij}) + \epsilon_{ij}.$$

That is, we define each cell or group regression to be a linear combination of three components, namely, a general effect (μ and β), a row (treatment) effect ($\mu_{i.}$ and $\beta_{i.}$) and a column (block) effect $\mu_{.j}$ and $\beta_{.j}$). It should be emphasized that the parameters ($\mu_{i.}$, $\beta_{i.}$, $\mu_{.j}$, $\beta_{.j}$) are not parameters of the marginal regressions (e.g., the regression for <u>all</u> graduates who live in a rural community), but measure row and column effects for the regression in each cell.

For this design, the mathematical model (see appendix)

$$E(Y) = A\underline{\beta}$$

assumes the following form (letting $x_{ij} = X_{ijm} - \bar{X}_{ij}$):

$$\begin{bmatrix}
\underline{1} & \underline{1} & \underline{0} & \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{x}_{11} & \underline{x}_{11} & \underline{0} & \underline{0} & \underline{x}_{11} & \underline{0} & \underline{0} \\
\underline{1} & \underline{1} & \underline{0} & \underline{0} & \underline{0} & \underline{1} & \underline{0} & \underline{x}_{12} & \underline{x}_{12} & \underline{0} & \underline{0} & \underline{0} & \underline{x}_{12} & \underline{0} \\
\underline{1} & \underline{1} & \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{1} & \underline{x}_{13} & \underline{x}_{12} & \underline{0} & \underline{0} & \underline{0} & \underline{0} & \underline{x}_{13} \\
\underline{1} & \underline{0} & \underline{1} & \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{x}_{21} & \underline{0} & \underline{x}_{21} & \underline{0} & \underline{x}_{21} & \underline{0} & \underline{0} \\
\underline{1} & \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{1} & \underline{0} & \underline{x}_{22} & \underline{0} & \underline{x}_{22} & \underline{0} & \underline{0} & \underline{x}_{22} & \underline{0} \\
\underline{1} & \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{0} & \underline{1} & \underline{x}_{23} & \underline{0} & \underline{x}_{23} & \underline{0} & \underline{0} & \underline{0} & \underline{x}_{23} \\
\underline{1} & \underline{0} & \underline{0} & \underline{1} & \underline{1} & \underline{0} & \underline{0} & \underline{x}_{31} & \underline{0} & \underline{0} & \underline{x}_{31} & \underline{x}_{31} & \underline{0} & \underline{0} \\
\underline{1} & \underline{0} & \underline{0} & \underline{1} & \underline{0} & \underline{1} & \underline{0} & \underline{x}_{32} & \underline{0} & \underline{0} & \underline{x}_{32} & \underline{0} & \underline{x}_{32} & \underline{0} \\
\underline{1} & \underline{0} & \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{1} & \underline{x}_{33} & \underline{0} & \underline{0} & \underline{x}_{33} & \underline{0} & \underline{0} & \underline{x}_{33}
\end{bmatrix}
\begin{bmatrix}
\mu \\ \mu_{1\cdot} \\ \mu_{2\cdot} \\ \mu_{3\cdot} \\ \mu_{\cdot 1} \\ \mu_{\cdot 2} \\ \mu_{\cdot 3} \\ \beta \\ \beta_{1\cdot} \\ \beta_{2\cdot} \\ \beta_{3\cdot} \\ \beta_{\cdot 1} \\ \beta_{\cdot 2} \\ \beta_{\cdot 3}
\end{bmatrix}
= E
\begin{bmatrix}
\underline{Y}_{11} \\ \underline{Y}_{12} \\ \underline{Y}_{13} \\ \underline{Y}_{21} \\ \underline{Y}_{22} \\ \underline{Y}_{23} \\ \underline{Y}_{31} \\ \underline{Y}_{32} \\ \underline{Y}_{33}
\end{bmatrix}$$

where $\underline{x}_{ij}$ is the vector of all x-observations, corrected for the mean, in the cell (i, j); $\underline{Y}_{ij}$ is the corresponding vector of Y-observations and $\underline{1}$ is a vector which has all elements equal to unity, the number of elements being equal to the number of x and Y observations in each cell. Now, with this definition, the normal equations are:

$$\begin{bmatrix} N & 0 \\ 0 & S \end{bmatrix}
\begin{bmatrix} \hat{\underline{\mu}} \\ \hat{\underline{\beta}} \end{bmatrix}
= \begin{bmatrix} \underline{T}_y \\ \underline{T}_{xy} \end{bmatrix}$$

where

$$N = \begin{bmatrix} n_{..} & n_{1.} & n_{2.} & n_{3.} & n_{.1} & n_{.2} & n_{.3} \\ n_{1.} & n_{1.} & 0 & 0 & n_{11} & n_{12} & n_{13} \\ n_{2.} & 0 & n_{2.} & 0 & n_{21} & n_{22} & n_{23} \\ n_{3.} & 0 & 0 & n_{3.} & n_{31} & n_{32} & n_{33} \\ n_{.1} & n_{11} & n_{21} & n_{31} & n_{.1} & 0 & 0 \\ n_{.2} & n_{12} & n_{22} & n_{32} & 0 & n_{.2} & 0 \\ n_{.3} & n_{13} & n_{23} & n_{33} & 0 & 0 & n_{13} \end{bmatrix}$$

where $n_{i.} = \sum_{j=1}^{3} n_{ij}$, $n_{.j} = \sum_{i=1}^{3} n_{ij}$, and $n_{..} = \sum_{i=1}^{3} n_{i.} = \sum_{j=1}^{3} n_{.j}$

$$S = \begin{bmatrix} S_{..} & S_{1.} & S_{2.} & S_{3.} & S_{.1} & S_{.2} & S_{.3} \\ S_{1.} & S_{1.} & 0 & 0 & S_{11} & S_{12} & S_{13} \\ S_{2.} & 0 & S_{2.} & 0 & S_{21} & S_{22} & S_{23} \\ S_{3.} & 0 & 0 & S_{3.} & S_{31} & S_{32} & S_{33} \\ S_{.1} & S_{11} & S_{21} & S_{31} & S_{.1} & 0 & 0 \\ S_{.2} & S_{12} & S_{22} & S_{32} & 0 & S_{.2} & 0 \\ S_{.3} & S_{13} & S_{23} & S_{33} & 0 & 0 & S_{.3} \end{bmatrix}$$

where $Sx_{ij}^2 = \sum_{m=1}^{n_{ij}} (X_{ijm} - \bar{X}_{ij})^2$, $S_{i.} = \sum_{j=1}^{3} Sx_{ij}^2$, $S_{.j} = \sum_{i=1}^{3} Sx_{ij}^2$,

$S_{..} = \sum_{i=1}^{3} S_{i.} = \sum_{j=1}^{3} S_{.j}$.

$$\hat{\underline{\mu}}' = [\hat{\mu}, \hat{\mu}_{1.}, \hat{\mu}_{2.}, \hat{\mu}_{3.}, \hat{\mu}_{.1}, \hat{\mu}_{.2}, \hat{\mu}_{.3}]$$

$$\hat{\underline{\beta}}' = [\hat{\beta}, \hat{\beta}_{1.}, \hat{\beta}_{2.}, \hat{\beta}_{3.}, \hat{\beta}_{.1}, \hat{\beta}_{.2}, \hat{\beta}_{.3}]$$

$$\underline{T}_y' = [Y_{...}, Y_{1..}, Y_{2..}, Y_{3..}, Y_{.1.}, Y_{.2.}, Y_{.3.}]$$

where

$$Y_{i..} = \sum_{jm}\sum Y_{ijm}$$

$$Y_{.j.} = \sum_{i=1}^{3} \sum_{m=1}^{n_{ij}} Y_{ijm}$$

and

$$Y_{...} = \sum_{i=1}^{3} Y_{i..} = \sum_{j=1}^{3} Y_{.j.}$$

$$T'_{xy} =$$

$$\left[(Sxy)_{..},(Sxy)_{1.},(Sxy)_{2.},(Sxy)_{3.},(Sxy)_{.1},(Sxy)_{.2},(Sxy)_{.3}\right]$$

where

$$(Sx_{ij}y_{ij}) = \sum_{m=1}^{n_{ij}} (X_{ijm} - \bar{X}_{ij})Y_{ijm}$$

$$(Sxy)_{i.} = \sum_{j=1}^{3} Sx_{ij}y_{ij}$$

$$(Sxy)_{.j} = \sum_{i=1}^{3} Sx_{ij}y_{ij}$$

and

$$(Sxy)_{..} = \sum_{i=1}^{3}(Sxy)_{i.} = \sum_{j=1}^{3}(Sxy)_{.j}$$

Now, let us write:

$$N = \begin{bmatrix} n_{..} & n_{1.} & n_{2.} & n_{3.} & n_{.1} & n_{.2} & n_{.3} \\ n_{1.} & & & & & & \\ n_{2.} & & D_{n_{i.}} & & & M & \\ n_{3.} & & & & & & \\ n_{.1} & & & & & & \\ n_{.2} & & M' & & & D_{n_{.j}} & \\ n_{.3} & & & & & & \end{bmatrix}$$

and introduce the two constraints:

$$n_{1.}\mu_{1.} + n_{2.}\mu_{2.} + n_{3.}\mu_{3.} = 0$$

and

$$n_{.1}\mu_{.1} + n_{.2}\mu_{.2} + n_{.3}\mu_{.3} = 0$$

Now let

$$\underline{n}'_{i.} = [n_{1.}, n_{2.}, n_{3.}]$$

$$\underline{n}'_{.j} = [n_{.1}, n_{.2}, n_{.3}]$$

Then,

$$\hat{\mu} = \frac{1}{n_{..}} Y_{...}$$

and

$$D_{n_i.}\hat{\underline{\mu}}_{i.} + M\hat{\underline{\mu}}_{.j} = \underline{Y}_{i..} - \underline{n}_{i.}\hat{\mu}$$

$$M'\hat{\underline{\mu}}_{i.} + D_{n.j}\hat{\underline{\mu}}_{.j} = Y_{.j.} - \underline{n}_{.j}\hat{\mu}$$

Premultiplying (II) by $MD^{-1}_{n_{.j}}$ and subtracting the result from (I) we have

$$(D_{n_{i.}} - MD^{-1}_{n_{.j}}M')\hat{\underline{\mu}}_{i.} = \underline{Y}_{i..} - MD^{-1}_{n_{.j}}Y_{.j.}$$

where

$$\underline{\mu}'_{i.} = [\hat{\mu}_{1.}, \hat{\mu}_{2.}, \hat{\mu}_{3.}]$$

$$Y'_{i..} = \left[Y_{1..}, Y_{2..}, Y_{3..} \right]$$

and

$$Y'_{.j.} = \left[Y_{.1.}, Y_{.2.}, Y_{.3.} \right]$$

This can be written in full as

$$\begin{bmatrix} n_{1.} - \sum_{j=1}^{3} \frac{n_{1j}^2}{n_{.j}} & - \sum_{j=1}^{3} \frac{n_{1j}n_{2j}}{n_{.j}} & - \sum_{j=1}^{3} \frac{n_{1j}n_{3j}}{n_{.j}} \\ - \sum_{j=1}^{3} \frac{n_{1j}n_{2j}}{n_{.j}} & n_{2.} - \sum_{j=1}^{3} \frac{n_{2j}^2}{n_{.j}} & - \sum_{j=1}^{3} \frac{n_{2j}n_{3j}}{n_{.j}} \\ - \sum_{j=1}^{3} \frac{n_{1j}n_{3j}}{n_{.j}} & - \sum_{j=1}^{3} \frac{n_{2j}n_{3j}}{n_{.j}} & n_{3.} - \sum_{j=1}^{3} \frac{n_{3j}^2}{n_{.j}} \end{bmatrix} \begin{bmatrix} \hat{\mu}_{1.} \\ \hat{\mu}_{2.} \\ \hat{\mu}_{3.} \end{bmatrix} =$$

$$\begin{bmatrix} Y_{1..} - \sum_{j=1}^{3} \frac{n_{1j}Y_{.j.}}{n_{.j}} \\ Y_{2..} - \sum_{j=1}^{3} \frac{n_{2j}Y_{.j.}}{n_{.j}} \\ Y_{3..} - \sum_{j=1}^{3} \frac{n_{3j}Y_{.j.}}{n_{.j}} \end{bmatrix} \qquad (5.1)$$

A similar argument for the matrix S leads to equations for $\hat{\beta}_{1.}, \hat{\beta}_{2.}, \hat{\beta}_{3.}$:

$$\begin{bmatrix} S_{1.} - \sum_{j=1}^{3} \frac{Sx_{1j}^2}{S_{.j}} & - \sum_{j=1}^{3} \frac{Sx_{1j}Sx_{2j}}{S_{.j}} & - \sum_{j=1}^{3} \frac{Sx_{1j}Sx_{3j}}{S_{.j}} \\ - \sum_{j=1}^{3} \frac{Sx_{1j}Sx_{2j}}{S_{.j}} & S_{2.} - \sum_{j=1}^{3} \frac{Sx_{2j}^2}{S_{.j}} & - \sum_{j=1}^{3} \frac{Sx_{2j}Sx_{3j}}{S_{.j}} \\ - \sum_{j=1}^{3} \frac{Sx_{1j}Sx_{2j}}{S_{.j}} & - \sum_{j=1}^{3} \frac{Sx_{2j}Sx_{3j}}{S_{.j}} & S_{3.} - \sum_{j=1}^{3} \frac{Sx_{3j}^2}{S_{.j}} \end{bmatrix} \begin{bmatrix} \hat{\beta}_{1.} \\ \hat{\beta}_{2.} \\ \hat{\beta}_{3.} \end{bmatrix}$$

(5.2)

$$= \begin{bmatrix} (Sxy)_{1.} - \sum_{j=1}^{3} \frac{Sx_{1j}(Sxy)_{.j}}{S_{.j}} \\ (Sxy)_{2.} - \sum_{j=1}^{3} \frac{Sx_{2j}(Sxy)_{.j}}{S_{.j}} \\ (Sxy)_{3.} - \sum_{j=1}^{3} \frac{Sx_{3j}(Sxy)_{.j}}{S_{.j}} \end{bmatrix}$$

Both systems of equations (5.1) and (5.2) contain one singularity. This difficulty may be overcome in many ways. We may define one of the parameters to be a constant, solving the resulting 2 equations in 2 unknowns. Another possible solution is to introduce a dummy variate z into each equation (see e.g., [1]), define the sum of the parameters to be zero as another equation and solve the resulting non-singular 4 equations in 4 unknowns.

We have treated for this problem the special case of 3 x 3 classification. The solution for the r x c general case can be carried out by the same procedure, and the results for this case are straightforward generalizations of (5.1) and (5.2).

B. Interaction in a Two-Way Classification

The problem discussed in the A section of this chapter becomes more complicated when we wish to study the effect of interaction between the two classifications, as is done in a factorial experiment. The advantage of the factorial experiment over the ordinary randomized block design is that the latter assumes that the two factors act independently of one another while the former tests this assumption. That is, a factorial experiment assumes two types of treatments, A and B, instead of one treatment and extraneous effects (i.e., blocks). The test that the factors act independently of one another is the test of significance of the interaction effect mentioned above. An interaction effect, then, occurs when two (or more) factors combine to produce an added (either positive or negative) effect not due to either (or any) of them alone.

For the situation of regressions in a factorial arrangement, an examination of the many models which can be used would, for the most part, be a repetition of the discussion of the models given in the introductory chapter. For this reason, we will merely state that, for our purposes, the most advantageous model is:

$$Y_{ijm} = \mu + \mu_{i.} + \mu_{.j} + \mu^{ij} + (\beta + \beta_{i.} + \beta_{.j} + \beta^{ij})X_{ijm} + \epsilon_{ijm}. \qquad \begin{matrix} i = 1 \ldots r \\ j = 1 \ldots c \\ m = 1 \ldots n_{ij} \end{matrix}$$

As in section A, we assume the regression parameters in ij'th cell to be made up of components. Let μ_{ij} and β_{ij} be the parameters in the ij'th cell, i.e.,

$$Y_{ijm} = \mu_{ij} + \beta_{ij}(X_{ijm} - \bar{X}_{ij}) + \epsilon_{ijm}.$$

We then assume that

$$\mu_{ij} = \mu + \mu_{i.} + \mu_{.j} + \mu^{ij} \text{ and}$$

$$\beta_{ij} = \beta + \beta_{i.} + \beta_{.j} + \beta^{ij};$$

or in words, we assume that the regression parameters in each cell are composed of a general effect (μ, β), a row effect ($\mu_{i.}$, $\beta_{i.}$), a column effect ($\mu_{.j}$, $\beta_{.j}$) and an interaction effect (μ^{ij} , β^{ij}).

The chief advantage of this model over the one using α and β, as in all cases where μ and β are used, is that μ and β are uncorrelated. It also makes the tests of significance for equality of row contribution to slopes, for equality of column contribution to slopes and equality of "interaction between row and column contribution to slopes" easier to perform, as will be seen in Chapter VI.

This model has the same limitations as the one discussed earlier in this paper in that any test of equality among the μ or mean terms would be of little importance (except for the special case where $\bar{X}_{ij}$ is the same for all groups). Further, μ and β as defined above are not overall regression parameters but contributions to the regressions within cells or groups. Hence, tests of the hypotheses of

i) Equal marginal regression lines for all rows (columns)

ii) Equal marginal regression slopes for all rows (columns)

require reparametrization of the model, which we did not consider because these tests are quite artificial.

The normal equations for the estimation of parameters in this model are singular. We may remove this singularity by the introduction of constraints on the parameters or by the elimination of certain rows and columns in the $(A'A)$ matrix. For the case we are discussing this would amount to the elimination of one of the column effects, one of the row effects and $r + c - 1$ of the interaction effects. We will employ this latter device for the solution of our problem.

For the case of a two-way classification with r rows and c columns, Figure 6 represents the experimental design when these effects are eliminated.

Columns	1	2		c
Rows 1	$E(Y_{11m})=\mu+\beta x_{11m}$	$E(Y_{12m})=\mu+\mu_{.2}+(\beta+\beta_{.2})x_{12m}$	. . .	$E(Y_{1cm})=\mu+\mu_{.c}+(\beta+\beta_{.c})x_{1cm}$
2	$E(Y_{21m})=\mu+\mu_{2.}+(\beta+\beta_{2.})x_{21m}$	$E(Y_{22m})=\mu+\mu_{.2}+\mu_{2.}+\mu^{22}$ $+(\beta+\beta_{2.}+\beta_{.2}+\beta^{22})x_{22m}$	. . .	$E(Y_{2cm})=\mu+\mu_{.c}+\mu_{2.}+\mu^{2c}$ $+(\beta+\beta_{.c}+\beta_{2.}+\beta^{2c})x_{2cm}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
r	$E(Y_{r1m})=\mu+\mu_{r.}+(\beta+\beta_{r.})x_{r1m}$	$E(Y_{r2m})=\mu+\mu_{r.}+\mu_{.2}+\mu^{r2}$ $+(\beta+\beta_{r.}+\beta_{.2}+\beta^{r2})x_{r2m}$	. . .	$E(Y_{rcm})=\mu+\mu_{r.}+\mu_{.c}+\mu^{rc}$ $+(\beta+\beta_{r.}+\beta_{.c}+\beta^{rc})x_{rcm}$

FIGURE 6

TWO-WAY CLASSIFICATION REGRESSION MODEL WITH INTERACTION

The derivation of normal equations in this case is quite straightforward, but the general algebraic expressions are so cumbersome that we recommend inserting numerical values, once a specific design is given. An example for the case of two columns and two rows will be presented in the following chapter. It is of interest to note that, for this two-by-two case, with the ordinary (α, β) model, an 8 x 8 matrix must be inverted for the estimation of all parameters; with the (μ, β) model, only two 4 x 4 matrices must be inverted, saving considerable computing time.

CHAPTER VI

NUMERICAL EXAMPLE

An example to illustrate the more important techniques presented in Chapters II, III and V will be given below, using a subset of data collected for the Southern Regional Livestock Marketing Survey; see Stout [4]. Data on transactions involving two classes of livestock, beef steers and beef heifers, were obtained from two livestock auction markets, denoted Markets 17 and 21. The data consists of observations, for each sales lot on each of eight sampled auctions, on price per hundred pounds liveweight and market grade. Grade is coded in descending order from 1, highest quality prime, to 21, lowest quality canner. It is of interest to see how the regression of price on grade varies between the two classes and markets given above. A data summary of the 1909 lots used for our example is presented in Table 1.

The test that all four regression lines are the same is given by formula (2.2). From the data, we can easily find that

$$\sum_{i=1}^{4} n_i = 1909$$

$$SSy^2 = 16,671.6822,$$

$$SSx^2 = 14,537.14,$$

and $$SSxy = -7131.59.$$

	21 Steers	17 Steers	21 Heifers	17 Heifers
n_i	655.0000	609.0000	310.0000	335.0000
$\sum_j Y_{ij}$	10,667.5500	9,574.9500	4,320.0300	4,303.4500
$\sum_j X_{ij}$	6,129.0000	6,906.0000	3,222.0000	3,724.0000
$\sum_j Y_{ij}^2$	178,666.9927	154,422.6875	62,697.0065	57,367.3425
$\sum_j X_{ij}^2$	61,255.0000	83,026.0000	35,364.0000	44,028.0000
$\sum_j X_{ij}Y_{ij}$	97,733.5900	106,849.6000	43,830.9800	46,586.8500
Sy_i^2	4,931.6905	3,881.3616	2,494.8801	2,084.7100
Sx_i^2	3,904.4031	4,712.6404	1,875.9871	2,630.4598
Sx_iy_i	-2,085.3625	-1,729.3896	-1,069.4608	-1,252.0986

Table 1

SUMMARY OF DATA FOR NUMERICAL EXAMPLE

The error for one overall regression is given by:

$$SSy^2 - \frac{(SSxy)^2}{Sx^2} = 13,173.0867.$$

The error for each individual regression is given by

$$\sum_{i=1}^{4} \left(Sy_i^2 - \frac{(Sx_i y_i)^2}{Sx_i^2}\right) =$$

$$3246.7305 + 3817.8873 + 1885.2030 + 1488.7112$$

$$= 10,438.5320 .$$

The test is then,

$$F_{6,1901} = \frac{(13,173.0867 - 10,438.5320)/6}{10,438.5320/1901}$$

$$= \frac{455.7591}{5.4911} = 83.00 ,$$

which when compared with the corresponding tabular F-value at the 1% level, 2.81, is highly significant. We must therefore reject the null hypothesis of one regression line for all the data.

To test $H_o: \beta_1 = \beta_2 = \beta_3 = \beta_4$, we use the F-test given in equation (3.1.3.). For our numerical example, the test is,

$$F_{3,1901} = \frac{(2954.1102 - 2869.2305)/3}{10,438.5320/1901}$$

$$= \frac{28.2932}{5.4911} = 5.15 ,$$

which when compared with the tabular F-value, at the 1% level, 3.79 is significant. We must also reject the hypothesis of equal slopes for all groups.

The reader will note that the data can be arranged in a two-way classification as below in order to study variations of the regressions between markets and between types of beef.

FIGURE 7

REGRESSION MODEL FOR NUMERICAL EXAMPLE

Type \ Market	21	17
Steers	$E(Y_{11m}) = \mu + \beta x_{11m}$	$E(Y_{12m}) = \mu + \mu_{12} + (\beta + \beta_{.2})x_{12m}$
Heifers	$E(Y_{21m}) = \mu + \mu_{2.} + (\beta + \beta_{2.})x_{21m}$	$E(Y_{22m}) = \mu + \mu_{.2} + \mu_{2.} + \mu^{22} + (\beta + \beta_{.2} + \beta_{2.} + \beta^{22})x_{22m}$

For this design, the general least squares equations for the μ terms are:

$$\begin{bmatrix} \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} & n_{12}+n_{22} & n_{21}+n_{22} & n_{22} \\ n_{12}+n_{22} & n_{12}+n_{22} & n_{22} & n_{22} \\ n_{21}+n_{22} & n_{22} & n_{21}+n_{22} & n_{22} \\ n_{22} & n_{22} & n_{22} & n_{22} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mu}_{.2} \\ \hat{\mu}_{2.} \\ \hat{\mu}^{22} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{m=1}^{n_{ij}} Y_{ijm} \\ \sum_{i=1}^{2}\sum_{m=1}^{m_{ij}} Y_{i2m} \\ \sum_{j=1}^{2}\sum_{m=1}^{n_{ij}} Y_{2jm} \\ \sum_{m=1}^{n_{22}} Y_{22m} \end{bmatrix}$$

The general least squares equations for the β terms are:

$$\begin{bmatrix} \sum_{i=1}^{2}\sum_{i=1}^{2} Sx_{ij}^2 & Sx_{12}^2+Sx_{22}^2 & Sx_{21}^2+Sx_{22}^2 & Sx_{22}^2 \\ Sx_{12}^2+Sx_{22}^2 & Sx_{12}^2+Sx_{22}^2 & Sx_{22}^2 & Sx_{22}^2 \\ Sx_{21}^2+Sx_{22}^2 & Sx_{22}^2 & Sx_{21}^2+Sx_{22}^2 & Sx_{22}^2 \\ Sx_{22}^2 & Sx_{22}^2 & & Sx_{22}^2 \end{bmatrix} \begin{bmatrix} b \\ b_{.2} \\ b_{2.} \\ b^{22} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{2}\sum_{j=1}^{2} Sx_{ij}y_{ij} \\ \sum_{i=1}^{2} Sx_{i2}y_{i2} \\ \sum_{j=1}^{2} Sx_{2j}y_{2j} \\ Sx_{22}y_{22} \end{bmatrix}$$

For our numerical example, these equations are:

$$\underset{(6.1)}{\begin{bmatrix} 1909 & 944 & 645 & 325 \\ 944 & 944 & 335 & 335 \\ 645 & 335 & 645 & 335 \\ 335 & 335 & 335 & 335 \end{bmatrix}} \underset{A'A}{} \quad \underset{\hat{\xi}}{\begin{bmatrix} \hat{\mu} \\ \hat{\mu}_{.2} \\ \hat{\mu}_{2.} \\ \hat{\mu}^{22} \end{bmatrix}} = \underset{A'Y}{\begin{bmatrix} 28,865.98 \\ 13,878.40 \\ 8,623.48 \\ 4,303.45 \end{bmatrix}}$$

and

(6.2)

$$\begin{bmatrix} 13,123.4904 & 7343.1002 & 4506.4469 & 2630.4598 \\ 7,343.1002 & 7343.1002 & 2630.4598 & 2630.4598 \\ 4,506.4469 & 2630.4598 & 4506.4469 & 2630.4598 \\ 2,630.4598 & 2630.4598 & 2630.4598 & 2630.4598 \end{bmatrix} \begin{bmatrix} b \\ b_{.2} \\ b_{2.} \\ b^{22} \end{bmatrix} \quad \begin{bmatrix} -6136.3115 \\ -2981.4882 \\ -2321.5594 \\ -1252.0986 \end{bmatrix}$$

The inverse of (A'A) for μ terms (i.e. [6.1]) is:

$$(6.3)\qquad \begin{bmatrix} 1.526718 & -1.526718 & -1.526718 & 1.526718 \\ -1.526718 & 3.168754 & 1.526718 & -3.168754 \\ -1.526718 & 1.526718 & 4.752524 & -4.752524 \\ 1.526718 & -3.167854 & -4.752524 & 9.379635 \end{bmatrix} \times 10^{-3};$$

and the inverse of A'A for β terms is:

$$(6.4)\qquad \begin{bmatrix} 2.5612112 & -2.5612112 & -2.5612112 & 2.5612112 \\ -2.5612112 & 4.6831638 & 2.5612112 & -4.6831638 \\ -2.5612112 & 2.5612112 & 7.8917386 & -7.8917386 \\ 2.5612112 & 4.6831638 & 7.8917386 & 71.3815307 \end{bmatrix}$$

The values of the estimates of the parameters are:

$$(6.5)\qquad \hat{\xi}_1 = \begin{bmatrix} \hat{\mu} \\ \hat{\mu}_{.2} \\ \hat{\mu}_{2.} \\ \hat{\mu}_{22} \end{bmatrix} = \begin{bmatrix} 16.424 \\ -.849 \\ -2.488 \\ -.240 \end{bmatrix},$$

and

$$(6.6)\qquad \hat{\xi}_\beta = \begin{bmatrix} b \\ b_{.2} \\ b_{2.} \\ b_{22} \end{bmatrix} = \begin{bmatrix} -.367 \\ -.167 \\ -.203 \\ -.261 \end{bmatrix}.$$

To test whether the regression slopes remain the same between columns, we require a test of the hypothesis $\beta_{.2} = 0$. By use of step 5 of the general linear hypothesis technique,

as outlined in the appendix, the sum of squares due to this hypothesis is:

$$(6.7) \qquad \hat{\underline{\xi}}_{\beta}' C' [C(A'A)^{-1}C']^{-1} C \hat{\underline{\xi}}_{\beta}$$

where C, the hypothesis matrix, is:

$$(6.8) \qquad [\,0 \;\; 1 \;\; 0 \;\; 0\,].$$

Substituting the numerical values (6.4), (6.6) and (6.8) for $(A'A)^{-1}$, $\hat{\underline{\xi}}_{\beta}$ and C respectively, the sum of squares due to the null hypothesis is 59.5516 with one degree of freedom. The mean square error as given on page 52 is 5.4911. The resulting F-test gives

$$F_{1,1901} = \frac{59.5516}{5.4911} = 10.845 .$$

This value is significantly large to reject the hypothesis: $\beta_{.2} = 0$, when compared with the corresponding tabular value at the 1% probability level, 6.65.

The equivalent test for row slopes is the null hypothesis: $\beta_{2.} = 0$. For this hypothesis, the value of C, the hypothesis matrix, to be used in expression (6.7), is

$$[\,0 \;\; 0 \;\; 1 \;\; 0\,].$$

Upon insertion of the proper numerical values, the sum of squares due to this null hypothesis becomes 1.6422. The F-test gives

$$F_{1,1901} = \frac{1.6422}{5.4911} = .30$$

which is not large enough to reject the hypothesis: $\beta_{2.} = 0$ at the 5% level.

To test whether there is significant interaction contribution to slopes, we require a test of H_o: $\beta_{22} = 0$. In this case, the value of C is

$$[0 \quad 0 \quad 0 \quad 1].$$

The resulting sum of squares is 3.8573 with one degree of freedom. The resulting F-statistic,

$$F_{1,1901} = \frac{3.8573}{5.4911} = .70 ,$$

which when compared with the tabular value at the 5% level, is again not large enough to reject the null hypothesis.

Of course, the five tests described in this chapter are not statistically independent in any case because all F-ratios have the same denominator. Moreover the sums of squares due to hypotheses of equal row effects on regressions, column effects on regressions and interaction effects on regressions are not orthogonal. The arrangement of the tests used in this chapter, then, is not intended to be a recommendation for the procedures to be followed in a study such as ours, but rather was done to illustrate how each hypothesis may be tested for significance.

For simplicity, we decomposed

$$\mu_{ij} \text{ into } \mu + \mu_{i.} + \mu_{.j} + \mu^{ij}.$$

But a test of all $\mu_{i.} = 0$, all $\mu_{.j} = 0$ and all $\mu^{ij} = 0$ would be artificial. We may, however, obtain from

$$\hat{\mu},\ \hat{\mu}_{i.},\ \hat{\mu}_{.j},\ \hat{\mu}^{ij},\ \hat{\beta},\ \hat{\beta}_{i.},\ \hat{\beta}_{.j},\ \hat{\beta}^{ij}$$

the estimates of γ, $\gamma_{i.}$, $\gamma_{.j}$, γ^{ij}, if we know A. Thus, let $x'_{ijm} = x_{ijm} - A$. With this definition

$$\hat{\gamma} = \hat{\mu} - \hat{\beta}\bar{x}'_{11}$$

$$\hat{\gamma}_{i.} = \hat{\mu}_{i.} - \hat{\beta}_{i.}\bar{x}'_{1i} + \hat{\beta}(\bar{x}'_{11} - \bar{x}'_{1i})$$

$$\hat{\gamma}_{.j} = \hat{\mu}_{.j} - \hat{\beta}_{.j}\bar{x}'_{j1} + \hat{\beta}(\bar{x}'_{11} - \bar{x}'_{j1})$$

and

$$\hat{\gamma}^{ij} = \hat{\mu} + \hat{\mu}_{i.} + \hat{\mu}_{.j} + \hat{\mu}^{ij} + \hat{\beta}(\bar{x}'_{11} - \bar{x}'_{1i} - \bar{x}'_{j1}) - \hat{\beta}_{i.}\bar{x}'_{1i} - \hat{\beta}_{.j}\bar{x}'_{.j}\ .$$

To test, e. g., that all $\gamma_{i.}$ are equal to zero we may apply the general linear hypothesis technique with

$$C = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & (\bar{x}'_{11} - \bar{x}'_{12}) & -\bar{x}'_{12} & 0 & 0 \dots 0 \\ 0 & 0 & 1 & \dots & 0 & (\bar{x}'_{11} - \bar{x}'_{13}) & 0 & -\bar{x}'_{13} & 0 \dots 0 \\ 0 & 0 & 0 & \dots & 1 & (\bar{x}_{11} - \bar{x}'_{1C}) & 0 & 0 & 0 \dots \bar{x}'_{1C} \end{bmatrix}.$$

CHAPTER VII

MORE THAN ONE CONCOMITANT VARIATE

For the general multiple regression model:

$$Y_j^{(i)} = \beta_0^{(i)} + \beta_1^{(i)} X_{1j}^{(i)} + \beta_2^{(i)} X_{2j}^{(i)} + \ldots + \beta_p^{(i)} X_{pj}^{(i)} + \epsilon_j^{(i)},$$

(where i denotes differences between classifications and j denotes differences within classifications) as mentioned before, the nature of the considerations presented in the previous chapters are changed very little.

For testing all hypotheses on the above model, then, we will use the general linear hypothesis technique. For this technique, the error in all cases can be found by (see formula (2.1)).

$$(7.1) \qquad \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_j^{(i)2} - \hat{\underline{\beta}}' A' \underline{Y}$$

where $\hat{\underline{\beta}}' = [b_0^{(1)}, b_1^{(1)}, \ldots b_p^{(1)}, b_0^{(2)}, b_1^{(2)}, \ldots b_p^{(2)}, \ldots b_0^{(k)}, b_1^{(k)}, \ldots b_p^{(k)}]$.

$$A'\underline{y} = \begin{pmatrix} \sum_{j=1}^{n_1} Y_j(1) \\ \sum_{j=1}^{n_1} X_{1j}(1) Y_j(1) \\ \vdots \\ \sum_{j=1}^{n_1} X_{pj}(1) Y_j(1) \\ \sum_{j=1}^{n_2} Y_j(2) \\ \sum_{j=1}^{n_2} X_{1j}(2) Y_j(2) \\ \vdots \\ \sum_{j=1}^{n_2} X_{pj}(2) Y_j(2) \\ \vdots \\ \sum_{j=1}^{n_k} Y_j(k) \\ \sum_{j=1}^{n_k} X_{1j}(k) Y_j(k) \\ \vdots \\ \sum_{j=1}^{n_k} X_{pj}(k) Y_j(k) \end{pmatrix}$$

The sum of squares due to any null hypothesis, H_o: $C\underline{\xi}=0$ is

$$\hat{\underline{\xi}}' C' \left[C (A'A)^{-1} C'\right]^{-1} C \hat{\underline{\xi}}.$$

For this case,

$$\begin{bmatrix} A'A^{(1)} & 0 & \cdots & 0 \\ \cdot & A'A^{(2)} & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & A'A^{(k)} \end{bmatrix}$$

where

$$A'A^{(i)} = \begin{bmatrix} n_i & \sum_{j=1}^{n_i} X_{1j}^{(i)} & \sum_{j=1}^{n_i} X_{2j}^{(i)} & \cdots & \sum_{j=1}^{n_i} X_{pj}^{(i)} \\ \sum_{j=1}^{n_i} X_{1j}^{(i)} & \sum_{j=1}^{n_i} X_{1j}^{(i)2} & \sum_{j=1}^{n_i} X_{1j}^{(i)} X_{2j}^{(i)} & \cdots & \sum_{j=1}^{n_i} X_{1j}^{(i)} X_{pj}^{(i)} \\ \cdot & \cdot & & \cdots & \cdot \\ \cdot & \cdot & & \cdots & \cdot \\ \cdot & \cdot & & \cdots & \cdot \\ \sum_{j=1}^{n_i} X_{pj}^{(i)} & \sum_{j=1}^{n_i} X_{pj}^{(i)} X_{1j}^{(i)} & & \cdots & \sum_{j=1}^{n_i} X_{pj}^{(i)2} \end{bmatrix}.$$

All $(A'A)^{(i)}$ must be <u>numerically</u> inverted, hence no straightforward computational techniques are available as in Chapters II and III. For a test of H_o: $\beta_h^{(1)} = \beta_h^{(2)} = \ldots = \beta_h^{(k)}$ for all p values of h, we note that

$$\text{var } b_h^{(i)} = C_{hh}^{(i)} \sigma^2$$

where $C_{hh}^{(i)}$ is the element of the h'th row and h'th column of the inverse of $(A'A)^{(i)}$. The above test would be the test that each individual concomitant variate has the same influence for all classifications, which is the usual assumption in a standard analysis of covariance. The sum of squares due to this hypothesis is:

$$(7.2) \sum_{h=1}^{p} SS_h + \sum_{h=1}^{p} \sum_{\substack{g=1 \\ h \neq g}}^{p} SC_{hg},$$

with p(k-1) degrees of freedom, where

$$(7.3) \quad SS_h = \sum_{i=1}^{k} \frac{b_h^{(i)2}}{C_{hh}^{(i)}} - \frac{\left[\sum_{i=1}^{k} \frac{b_h^{(i)}}{C_{hh}^{(i)}}\right]^2}{\sum_{i=1}^{k} \frac{1}{C_{hh}^{(i)}}}$$

and

$$(7.4) \quad SC_{hg} = \sum_{i=1}^{k} \frac{b_h^{(i)} b_g^{(i)}}{C_{hg}^{(i)}} - \frac{\left[\sum_{i=1}^{k} \frac{b_h^{(i)}}{C_{hg}^{(i)}}\right]\left[\sum_{i=1}^{k} \frac{b_g^{(i)}}{C_{hg}^{(i)}}\right]}{\sum_{i=1}^{k} \frac{1}{C_{hg}^{(i)}}}$$

To test any particular subset of these $\beta_h^{(i)}$ for equality i.e., H_o: $\beta_h^{(1)} = \beta_h^{(2)} = \ldots = \beta_h^{(k)}$ for any q values of h($q \leq p$), the above formula (7.2) still holds. For such a hypothesis, we merely renumber the q concomitant variates of interest to run from 1 to q and substitute q for p in expression (7.2). The resulting sum of squares has q(k-1) degrees of freedom.

To test one particular $\beta_h^{(i)}$ for equality for all i (i.e., H_o: $\beta_h^{(1)} = \beta_h^{(2)} = \ldots = \beta_h^{(k)}$), the sum of squares due to the hypothesis is $\sum_{h=1}^{p} SS_h$ with (k-1) degrees of freedom. The term SS_h is defined in equation (7.3).

For all of the hypotheses described above, the error term to be used in the denominator of all F-tests is given by either of the expressions (2.1) or (7.1).

CHAPTER VIII

SUMMARY

This thesis discusses tests for several contrasts of the parameters $\alpha_i, \beta_{1i}, \ldots, \beta_{pi}$ in the model:

$$Y_{ij} = \alpha_i + \beta_{1i}X_{1ij} + \ldots + \beta_{pi}X_{pij} + \epsilon_{ij}, \quad \begin{matrix} i = 1 \ldots k \\ j = 1 \ldots n_i \end{matrix}$$

Methods which have been established for testing such homogeneity in the special case of one concomitant variate (X) are presented in the first three chapters of this paper, together with a general procedure to be followed in such a study. When, in this special case, it is found that there is a significant difference among the regression slopes (i.e., β_i), a problem arises regarding the selection of the proper origin for X. This problem is discussed in Chapter IV.

Chapter V deals with situations when the data for a regression study can be arranged in a two-way classification. Section A of this chapter considers situations when it is thought that a randomized block model best describes the physical situation. If there is reason to believe that there is some interaction effect between the classifications, the problem can be handled by techniques given in the B section of this chapter. Chapter VI contains a numerical example which illustrates the methods described in section B of Chapter V.

Chapter VII deals briefly with tests of homogeneity when p concomitant variates are involved.

IX ACKNOWLEDGEMENTS

The author wishes to express his thanks to Dr. R. J. Freund and Dr. Rolf Bargmann for their ideas and help in the organization of this paper. Thanks are also extended to and , , and who prepared the final typewritten copies.

BIBLIOGRAPHY

1. Kempthorne, Oscar, "Design and Analysis of Experiments", John Wiley and Sons Inc.; New York; 1952

2. Li, Jerome C. R., "Introduction to Statistical Inference", Edward Brothers Inc.; Ann Arbor, Michigan; 1957.

3. Ostle, Bernard, "Statistics in Research", Iowa State College Press, Ames, Iowa; 1956.

4. Stout, Roy G., "Marketing Cattle and Calves Through Southern Auctions"; Bulletin 48, Southern Co-operative Series, February 1957.

5. Tocher, K. D., "On the Concurrence of a Set of Regression Lines"; Biometrika, pp. 109-117, Vol. 39; 1952.

6. Williams, Evan J. "Tests of Significance for Concurrent Regression Lines"; Biometrika, pp. 297-305; Vol. 40; 1953.

The vita has been removed from the scanned document

APPENDIX

GENERAL LINEAR HYPOTHESIS TECHNIQUE

An exposition of the general linear hypothesis technique is presented in Kempthorne [1]. What follows below is an outline which summarizes in matrix form the important points of the technique. For the theoretical justification of each of these steps, the reader may consult [1].

1) Define the model: $E(Y) = A \underset{\sim}{\xi}$
   where A is the design matrix and $\underset{\sim}{\xi}$ is a vector of the parameters involved.

2) Set up the hypothesis to be tested in matrix form.
   $$\text{Under } H_o: \quad C \underset{\sim}{\xi} = 0$$
   $$H_A: \quad C \underset{\sim}{\xi} \neq 0,$$
   where C is an arbitrary "hypothesis" matrix denoting the linear functions of the parameters which we wish to test for significance.

3) Set up the normal or least-squares equations,
   $$(A'A) \hat{\underset{\sim}{\xi}} = A'\underline{Y} = \underline{T},$$
   in order to estimate $\underset{\sim}{\xi}$.

4) (a) If A'A is singular, introduce constraints on some function of $\underset{\sim}{\xi}$ by dropping columns in A

or adding rows in $\underline{\xi}$; thus obtain a new A'A which is non-singular.

(b) If A'A is non-singular, proceed to step 5.

5) The estimate of $\underline{\xi}$ is

$$\hat{\underline{\xi}} = (A'A)^{-1} A'\underline{y}.$$

6) The sum of squares due to the null hypothesis is then given by,

$$SS(H_o) = \hat{\underline{\xi}}'C'[C(A'A)^{-1}C']^{-1}C\hat{\underline{\xi}}.$$

7) The sum of squares due to error is given by

$$SS(\text{error}) = \underline{y}'\underline{y} - \hat{\underline{\xi}}'A'\underline{y}.$$

8) The sums of squares (6) and (7) are both distributed as independent Chi-squares; hence, the test of the null hypothesis is given by

$$F_{s,\ N-r} = \frac{SS(H_o) \ / \ s}{SS(\text{error}) \ / \ N-r},$$

where s is the number of independent rows in C or the number of independent parameters, r is the rank of the original A matrix and N is the total number of observations.

ABSTRACT

This thesis reviews and describes several methods of testing hypotheses which are assumed to be true in the standard analysis of covariance. In some instances, the usual assumption that regression lines or planes be parallel from group to group may not be justified. Tests of homogeneity which permit this conclusion have been studied in this thesis. If significant departure from this hypothesis is observed, adjustments on the concomitant variate or variates are required before an appropriate test can be made on contrasts between groups after elimination of the effects of concomitant variates. Some procedures for this analysis are reviewed and an additional one is given in this study.

Explicit results are stated for one-way classification and two-way classification with or without interaction, for the case of one concomitant variate. A brief outline is added describing the necessary modification of results if two or more concomitant variates are involved.

Tests and estimates were developed with the aid of the general linear hypothesis technique and, wherever possible, presented in a form comparable to standard sums-of-squares notation.