BIBLIOGRAPHIC DATA SHEET

1. CONTROL NUMBER
PN-AAJ-148

2. SUBJECT CLASSIFICATION (695)
PA00-0000-0000

3. TITLE AND SUBTITLE (240)
Linear models for WFS data

4. PERSONAL AUTHORS (100)
Little, R. J. A.

5. CORPORATE AUTHORS (101)
Int. Statistical Inst.

6. DOCUMENT DATE (110)
1980

7. NUMBER OF PAGES (120)
71p.

8. ARC NUMBER (170)
312.L778

9. REFERENCE ORGANIZATION (130)
ISI

10. SUPPLEMENTARY NOTES (500)

(In World Fertility Survey tech. bull. no. 9/Tech. 1282P)

11. ABSTRACT (950)

12. DESCRIPTORS (920)

Vital statistics
Models
Fertility survey
Statistical analysis
Demography
Sri Lanka

13. PROJECT NUMBER (150)
931054700

14. CONTRACT NO.(140)
AID/csd-3606

15. CONTRACT TYPE (141)

16. TYPE OF DOCUMENT (160)

312
L778

WORLD FERTILITY SURVEY

TECHNICAL BULLETINS

Linear Models for WFS Data

RODERICK J.A. LITTLE

Agency for International Development
Library
Room 105 SA-18
Washington, D.C. 20523

JUNE 1980 NO. 9/TECH. 1282P

CONTENTS

ACKNOWLEDGEMENTS		6
1.	INTRODUCTION	7
2.	CROSS-TABULATION AND DIRECT STANDARDIZATION	9
2.1	Introduction	9
2.2	One-Way Cross-Classifications	9
2.3	Two-Way Cross-Classifications	10
2.3.1	Interaction and Association	11
2.3.2	Direct Standardization	13
2.4	Three-Way Classifications	16
3.	ADDITIVE MODELS BETWEEN FACTORS: ANALYSIS OF VARIANCE AND MULTIPLE CLASSIFICATION ANALYSIS	21
3.1	Introduction	21
3.2	Multiple Classification Analysis	24
3.3	Analysis of Variance: Introduction	26
3.4	One-Way and Two-Way Analysis of Variance	27
3.5	Three-Way and Higher Tables	32
3.6	Refinements	34
4.	ANALYSIS OF COVARIANCE	39
4.1	Introduction	39
4.2	Analysis of Covariance	40
5.	MULTIPLE LINEAR REGRESSION	46
5.1	Introduction	46
5.2	Elements of Multiple Linear Regression	46
5.3	Treatment of Factors in Regression	49
5.3.1	A Single Factor	49
5.3.2	Two or More Factors	52
5.4	Covariates and Factors	55
5.5	Controlling the Order of Adjustment by Stepwise Regression	55
5.6	Interactions Between Factors and Covariates	58
6.	STRATEGIES FOR DETERMINING THE CHOICE OF VARIABLES IN THE REGRESSION	63
6.1	Introduction	63
6.2	The Causal Ordering and Total Effects	64
6.3	Examples	65
6.4	A Compromise Strategy	66
REFERENCES		67
TABLES		
2.1	Effect of a Variable from a One-Way Classification: Mean Number of Children Ever Born, by Respondent's Level of Education	10
2.2	Fitted Mean Parities from Additive Model	12
2.3	Effects of Education from Table D1 Expressed as Deviations from the Reference Category, No Schooling	14

2.4	Effects of Education from Table 2.2 Expressed as Deviations from the Reference Category, No Schooling	15
2.5	Examples of Additive Data Patterns for a 2x2x2 Table of Means	18
2.6	Distribution of Sample, by Age and by Age at Marriage	19
2.7	Standardization on Age and on Age at Marriage	19
3.1	Mean Number of Children Ever Born, by Marital Duration and by Level of Education: A) Observed Means, B) Fitted Means from Test Factor Standardization, C) Average Residuals	23
3.2	Mean Number of Children Ever Born, by Marital Duration and by Level of Education: A) Observed Means, B) Fitted Means from MCA, C) Average Residuals	24
3.3	Stem and Leaf Plots Comparing the Distribution of Absolute Residuals x 100 from Test Factor Standardization and from MCA	25
3.4	Multiple Classification Analysis of Parity, by Marital Duration and by Level of Education	27
3.5	One-Way ANOVA Table for a Factor A	28
3.6	One-Way Analysis of Variance of Parity, by Level of Education	28
3.7	Classical Two-Way Analysis of Variance of Parity, by Marital Duration and by Level of Education	30
3.8	Hierarchical Two-Way Analysis of Variance of Parity, by Marital Duration and by Level of Education	30
3.9	Hierarchical ANOVA Table of Parity on Marital Duration and Level of Education, for British Data in Table D3	30
3.10	Classical Three-Way Analysis of Variance of Parity, by Age, by Age at Marriage, and by Level of Education Treated as a Dichotomy	33
3.11	Hierarchical Three-Way Analysis of Variance of Parity, by Age, by Age at Marriage, and by Level of Education in Four Groups, with Three-Way Interactions Pooled with the Residual	34
3.12	Multiple Classification Analysis Corresponding to ANOVA on Table 3.11	35
3.13	Analysis of Variance with Age at Marriage and Age as a Joint Variable	35
3.14	Multiple Classification Analysis Corresponding to ANOVA in Table 3.13	36
3.15	A 2 x 2 Table of Proportions Additive on the Logit Scale	36
3.16	Weighted Analysis of Variance of Parity Divided by Marital Duration, by Marital Duration and by Level of Education	38
4.1	Analysis of Variance of Parity on Level of Education with Years Since First Marriage as a Covariate	43
4.2	Multiple Classification Analysis Corresponding to Table 4.1	43
4.3	Analysis of Variance of Parity by Level of Education, with Linear and Quadratic Terms of Marital Duration as Covariates	43
4.4	Multiple Classification Analysis Corresponding to Table 4.3	44
4.5	Summary Effects of Education, Expressed as Deviations from Overall Mean	44
4.6	Weighted Analysis of Variance of Parity Divided by Duration, by Respondent's Level of Education (LVED) and by Husband's Level of Education (HEDL), with Linear and Quadratic Terms in Years Since First Marriage and Age at First Marriage and the Product of Years Since First Marriage and Age at First Marriage as Covariates	45
4.7	Multiple Classification Analysis Corresponding to Table 4.6	45
5.1	Quadratic Regression of Parity on Years Since First Marriage	48

5.2	Regression of Parity on Level of Education Represented as a Set of Dummies	48
5.3	Regression of Parity on Marital Duration and Level of Education Represented as Sets of Dummy Variables	50
5.4	Analysis of Covariance Using Regression Program. Regression of Parity on Years Since First Marriage and Level of Education Represented as a Set of Dummy Variables	54
5.5	Regression of Parity on Linear and Quadratic Terms of Years Since First Marriage and Level of Education Represented as a Set of Dummy Variables	56
5.6	Per Cent of Currently Married, Non-Pregnant, 'Fecund' Women Currently Using an Efficient Method, by Region, Adjusted for Indicated Variables by Linear Regression	57
5.7	Analysis of Variance of Regression of PBYD with Interactions Added Hierarchically	59
5.8	Effects of Education from Regression of PBYD Including Interactions, Expressed as Deviations from Mean	61
D1	Mean Number of Children Ever Born, by Marital Duration (MGP6) and by Level of Education (LVED)	68
D2	Mean Number of Children Ever Born, by Age (AGP5), by Age at First Marriage (AMGP), and by Level of Education (LVED)	69
D3	Mean Number of Children Ever Born, by Marital Duration (MGP5) and by Level of Education (LVED)	71
D4	Means of Parity Divided by Marital Duration (PBYD), Weighted by Marital Duration, Cross-Classified by Marital Duration (MGP6) and by Level of Education (LVED)	72

FIGURES

2.1	Plots of Mean Parity, by Marital Duration and by Level of Education	12
4.1	Mean Number of Children Ever Born as a Function of Years Since First Marriage. Sri Lanka	39
4.2	Fitted Values from Analysis of Covariance of Mean Parity on Years Since First Marriage and Level of Education	41
5.1	Fitted Effects of Education as Quadratic Functions of Marital Duration, from Step 6 of Regression on PBYD in Table 5.8	61

ACKNOWLEDGEMENTS

This technical bulletin is based on lecture notes prepared for the UN/ESCAP/WFS Workshop on Use of Multivariate Techniques in Second-stage Analysis of World Fertility Survey Data held in Bangkok, Thailand, September to November 1979. The author acknowledges valuable comments on an earlier draft by John Casterline of WFS and John McDonald of the University of Washington, Seattle.

1. INTRODUCTION

This paper describes statistical methods based on the linear model in the context of applications to World Fertility Survey data. The common aim of all the methods presented is to relate the mean level of a quantitative variable Y to a set of other variables $X_1, \ldots, X_k$ measured in the survey. For example, Y may be a measure of fertility, such as the number of children ever born or births in the five years prior to the interview, or a measure of contraceptive use such as current use of a method. The variables $X_1, \ldots, X_k$ may be demographic controls such as age at interview or age at first marriage, or socioeconomic factors such as level of education, type of place of residence or ethnic group. The term *model* in the title means that the analysis is based on the imposition of a simplifying pattern or structure relating Y to the X's. The term *linear* relates to the fact that the equation defining the model is linear in the unknown parameters, a technical point that will be elucidated later. In practice a number of common statistical techniques are based on linear models, including direct standardization, multiple classification analysis, analysis of variance, analysis of covariance and linear regression, and these will be the principal subjects of the paper.

The linear model can also be described as the linear *regression* model. Although the term regression is sometimes limited to cases where the X's are interval-scaled, it can also be regarded as a generic term describing the whole family of methods. The technical justification of this viewpoint is delayed until chapter 5, but from the outset we use the terminology of regression to describe the variables. Thus Y is called the *regressand* variable (or simply the regressand) and the X's are called *regressor* variables (or simply regressors). Alternative terms for the regressand in the literature include the response variable and the dependent variable; alternative terms for regressors include predictor variables, independent variables and explanatory variables. Regressand and regressor are used here despite the initial difficulty of remembering which is which, because the other terms have potentially misleading implications.

Two types of regressors will also be distinguished. In the first part of the bulletin we shall be concerned with regressors which are categorical variables. Interval-scaled variables such as age are grouped into a relatively small number of categories, and the ordering between the categories is ignored. We described categorial regressors as *factors*. Later in the paper we shall treat interval-scaled regressors without grouping them into categories. The term *covariates* is reserved for regressors treated in this way. This terminology is not entirely satisfactory, but it is nevertheless convenient.

The basic strategy of the methods can be described as follows. Let us suppose we have N individuals in the sample, and let the suffix i denote the ith individual in the sample. Then the observed values of the regressand Y are

$$Y_1, Y_2, \ldots, Y_N, \text{ or } \{Y_i : i=1 \text{ to } N\}.$$

In an illustrative example used throughout the text, Y is the variable Number of Children Ever Born, also called simply parity, for a sample of N = 6810 ever-married women from the Sri Lanka Fertility Survey. Thus Y_i is the number of children ever born to the ith individual in the sample.

For simplicity we assume for the moment that the individuals in the sample are "representative" of the population, in the sense that they were selected by probability sampling and each individual in the population had an equal chance of being selected. In other words, we assume a self-weighted probability sample. In fact for the Sri Lanka case this assumption is not valid, since certain areas were sampled more heavily than others. For the analysis given here, individuals were assigned weights which are (a) inversely proportional to the probability of selection, and (b) normalized so that they sum to the number of observations in the sample.

The first step in an analysis is to estimate the mean value of Y for the population by the sample mean, $\bar{Y}$. For example, the mean parity for the Sri Lanka sample is $\bar{Y} = 3.94$ children. If all the

women in the sample had the same value, $\bar{Y}$, then this would be an adquate description of the data. In practice of course the values of Y vary among women in the sample. It is convenient to decompose each value of Y into the mean and the deviation from the mean, that is,

$$Y_i = \bar{Y} + (Y_i - \bar{Y}). \tag{1.1}$$

Then the *deviations* $Y_i - \bar{Y}$ represent fluctuations in the values of Y about the mean. The basic intent of the analysis is not to explain the average level of Y in the poulation. Rather we attempt to discover patterns or structure in the set of deviations $(Y - \bar{Y})$. That is, we attempt to "explain" the differences in the values of Y between individuals in terms of other variables measured in the survey.

It is not feasible to look at all the deviations in the sample, because of the large number of individuals in the sample. A powerful technique is to cross classify the mean values of Y by the factors expected to influence the deviations. For example, in the Sri Lanka example we may cross-classify the mean parities by demographic variables such as marital duration and age at marriage, socioeconomic variables such as education level, or geographical indicators such as urban-rural residence or region. The analysis of cross-tabulated means in the subject of the first two chapters of the text. Later more flexible methods for analyzing patterns in the deviations based on linear regression are considered.

Each analysis has an associated decomposition of the observed values Y into ***fitted values***, representing the simplified structure imposed on the data and ***residuals***, the deviations of the observed values from the fitted values, representing unexplained variation in the data. That is,

$$\text{observed} = \text{fit} + \text{residual}. \tag{1.2}$$

Thus in (1.1), the fitted values are all equal to the sample mean, and the residuals are the deviations from the mean. In a 1-way cross-tabulation, the fitted values are the means within each level of the cross-classifying factor, and the residuals are the deviations of the individual values in each category from the mean for that category. At the other extreme, the observed values are equal to the fitted values and the residuals are all zero.

In general, as more factors are included to account for the variations in the dependent variable, the "fit" component becomes more elaborate and the residual component has correspondingly less structure. One of the aims of the analysis is to find a model for the Y values which is a parsimonious description of the data and which leaves residuals which are relatively free of explainable patterns.

This general procedure should become clearer when applied to a specific set of analysis. We begin with the simplest, a one way cross-classification of means.

2. CROSS-TABULATION AND DIRECT STANDARDIZATION

2.1 Introduction

In this chapter we review some of the basic ideas underlying the analysis of cross-tabulations of means and sample sizes. These are analysed with the aid of the method of direct standardization, a simple technique for controlling categorical predictors which is familiar to most demographers. This method provides a convenient introduction to the statistical methods which are the principal subjects of this document.

2.2 One-Way Cross-Classifications

We begin with the simplest data structure, consisting of a single regressand variable Y and a single categorical regressor *(or factor)* X. For illustrative purpose we shall refer to the problem of assessing the relationship between education and fertility from a fertility survey of ever-married women. The first step in the process is taken in the following example.

Example 2.1 Data from the Sri Lanka Fertility Survey of 6810 women are available on the following variables:

Y = Number of Children Ever Born, otherwise called simply Parity.

X = Respondent's Educational level (LVED), with four categories:

LVED= 1 : No schooling
2 : 0-5 Years schooling
3 : 6-9 Years schooling
4 : 10 or more years schooling

The full data consists of the distribution of parities within each educational level. However, we suppose that interest is confined to a comparison of average parities, and hence we reduce the data to a one way cross-classification of the mean parity $\bar{y}_j$ for each educational level j, together with the sample sizes. This one way cross-classification is presented in (Table 2.1.a).

The mean parity for the 6810 women in the sample is 3.94. There are large differentials in mean parity by educational level, ranging from 2.30 for the higher educated group to 5.17 for the group with no education.

The *effects* of a categorical regressor X on a response Y consist of the differences in the mean of Y between categories of X.

There is no unique way of representing the effects of a categorical regressor; four common forms are shown in Table 2.1.c). The first alternative is to present all the pairwise differences between the category means in a triangle. This form has the merit a symmetry, but it is redundant since given any three pairwise differences involving all four categories the others can be calculated by addition or subtraction. The second and third methods of presentation express effects as deviations from the weighted and unweighted sample means. Note that the weighted mean depends on the distribution of the sample over the categories, and hence the effects expressed as deviations from the weighted mean are also sensitive to this distribution. This is not entirely satisfactory when comparing the effects of X for two populations with different distributions of X. In the final form of presentation in Table 2.1.c), effects are calculated as deviations from the mean of one category, the so-called *reference* category, here chosen as the group with no schooling.

Note that all alternatives give the same information about the effects of X, and we shall use them interchangeably according to convenience.

TABLE 2.1: Effects of a Variable from a One-Way Classification: Mean Number of Children Ever Born, by Respondent's Level of Education

		Educational Level				
		No Schooling	1–5 Years	6–9 Years	10 or More Years	Mean
a) Means		5.17	4.24	3.26	2.30	3.94
b) Sample Sizes		1512	2686	1704	908	
c) Effects of Education						
1. Expressed as Pairwise Differences:						
	None	–				
Educational	Primary	–0.93	–			
Level	Secondary	–1.91	–0.98	–		
	Higher	–2.87	–1.94	–0.96	–	
2. Expressed as Deviations from the Sample Mean (3.94)		1.23	0.30	–0.69	–1.64	
3. Expressed as Deviations from the Unweighted Mean (3.74)		1.43	0.50	–0.49	–1.44	
4. Expressed as Deviations from Reference Category: No schooling		–	–0.93	–1.91	–2.87	

Source: Sri Lanka Fertility Survey 1975.

The term "effect" as applied here has potential dangers, since it carries an unwarranted causal connotation. It is tempting to conclude from the one way cross-classification that the "effect" of secondary education has been to reduce mean parity, from 5.17 to 3.26. However, such an interpretation is clearly invalid, since the difference could be attributed to compositional effects of other factors correlated with education but unconnected with the educational process. The most easily recognisable of these are demographic factors such as age and age at marriage. Specifically, in developing countries more educated women tend to be younger and to marry later than average, and hence in a cross-sectional survey have had below average exposure to the risk of child-bearing. Thus the differentials in mean parity by educational level may be attributed to differentials in the distribution of marital duration between the education groups. These considerations lead naturally to higher way cross-tabulations which are the subjects of the next subsection.

2.3 Two-Way Cross-Classifications

We have noted that the effect of education in the one way classification of mean parities may be attributed to compositional effects of marital duration. To investigate this we cross-classify mean parity by Educational Level and Marital Duration. Table D1* displays the results of a cross-classification of Mean Parity by Educational Level and Years Since First Marriage in six categories (MGP6)

MGP6 = 1 = 0 - 4 years
2 = 5 - 9 years
3 = 10 -14 years
4 = 15 -19 years
5 = 20 -24 years
6 = 25+ years

* Tables of raw data are labelled D and appear after the text.

The first entry in each cell is the mean, the second entry is the sample count, and the third entry is the sample standard deviation.

2.3.1 Interaction and Association

Table D1 illustrates two important concepts relating to two way and higher way cross-classifications of means, namely, *interaction* and *association*. The term interaction refers to the first entries in the cells of Table D1, the Cell means. Two cross-classifying factors A & B are said to *interact* in their effect on a response if the effects of one factor vary according to the levels of the other factor. If there is no interaction, that is the effects of one factor are the same for all levels of the other factor, then the effects of A & B on the response are said to be additive.* We denote this additive structure by the symbol [A + B].

It will be useful for later developments to express the additive structure in symbols. Suppose that A has J levels and B has K levels. Let μ_{jk} and n_{jk} denote the mean and sample size for the cell with levels A = j and B - k of the factors, for j = 1 to J, k = 1 to K. The effects of A and B are additive if and only if the means μ_{jk} can be written in the form

$$\mu_{jk} = m + r_j + c_k \,, \; j = 1 \text{ to } J; \; k = 1 \text{ to } K \,, \tag{2.1}$$

where m is a constant, $\{r_j\colon\ j = 1 \text{ to } J\}$ are quantities defined for each level of j of the row factor A, and $\{c_k\colon\ k = 1 \text{ to } K\}$ are quantities defined for each level k of the column factor B.

To verify this equivalence, note that the effects of B within level j of A, expressed as deviations from the first category of B, can be written as

$$\mu_{jk} - \mu_{j1}, \; k = 2 \text{ to } K.$$

Substituting the right hand side of equation (2.1), we obtain

$$\begin{aligned} \mu_{jk} - \mu_{j1} &= (\mu + r_j + c_k) - (\mu + r_j + c_1) \\ &= c_k - c_1 \,. \end{aligned}$$

Now $c_k - c_1$ does not involve the row subscript j, and hence is the same for all levels of A. In other words, the effects of B within levels of A are the same for all levels of A, which is the definition of additivity.

The terms on the right hand side of equation (2.1) are not unique, in that different sets of m, $\{r_j\}$ and $\{c_k\}$ give the same set of means $\{\mu_{jk}\}$. For example, if we add a constant 2d to m, and subtract d from all the values $\{r_j\}$ and $\{c_k\}$, we obtain the same means: That is, if we define

$$m^* = m+2d \,, \; r_j^* = r_j - d \,, \; c_k^* = c_k - d \,,$$

then

$$m^* + r_j^* + c_k^* = (m+2d) + (r_j - d) + (c_k - d) = m + r_j + c_k = \mu_{jk} \,.$$

Thus restrictions are required to define the values m, $\{r_j\}$ and $\{c_k\}$ uniquely. There are two common choices, corresponding to the definition of m as the mean of the first cell of the table, or as the overall (weighted) mean of the table. In the first case, we set $m = \mu_{11}$ in equation (2.1) and obtain

$$\mu_{jk} = \mu_{11} + r_j + c_k \,, \; j = 1 \text{ to } J; \; k = 1 \text{ to } K \,. \tag{2.2}$$

Setting j = k = 1 in this equation, we have

$$\mu_{11} = \mu_{11} + r_1 + c_1 \,,$$

---

* Strictly speaking, the definition of additivity is related to the scale of measurement. This definition corresponds to additivity on the *linear* scale, or linear additivity. Other scales are discussed in section 3.5.

FIGURE 2.1: Plots of Mean Parity by Marital Duration and by Level of Education

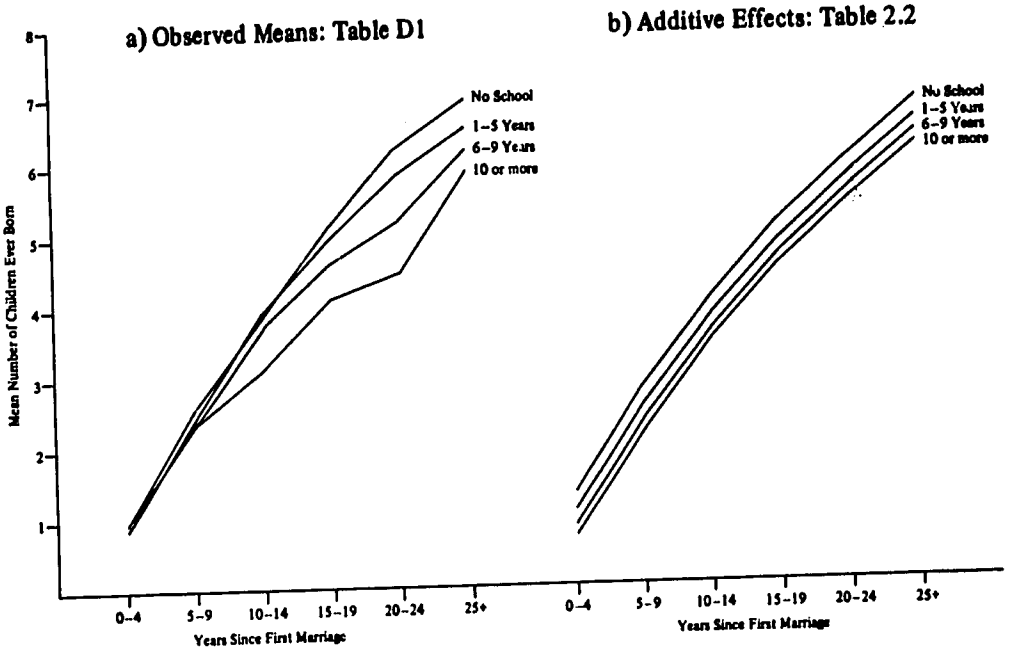


TABLE 2.2: Fitted Mean Parities from Additive Model

Variable averaged . . . NCEB.

		Lved				
Mgp6	Fitted Mean Count	No Schooling (1)	1–5 Years (2)	6–9 Years (3)	10 or More Years (4)	Total
0–4	1	1.31 112	1.07 376	0.86 442	0.71 351	.92 1280
5–9	2	3.78 172	2.54 442	2.33 362	2.18 255	2.44 1231
10–14	3	4.06 197	3.82 482	3.61 293	3.46 145	3.76 1118
15–19	4	5.11 239	4.87 461	4.66 262	4.51 95	4.83 1057
20–24	5	6.01 292	5.77 377	5.56 184	5.41 40	5.78 893
25 +	6	6.82 501	6.58 548	6.37 161	5.22 22	6.64 1231
Total		5.16 1512	4.24 2686	3.26 1704	2.30 908	3.94 6810

and hence the quantities $\{r_j\}$ and $\{c_k\}$ are defined so that the first row and column effects are zero, that is

$$r_1 = c_1 = 0. \tag{2.3}$$

We have already noted that $\{c_k - c_1\}$ represent the effects of B within levels of A, expressed as deviations from the first category of B. Thus if m is defined as in (2.2), $c_1 = 0$ and $\{c_k\}$ have this definition. Similarly $\{r_j\}$ represent the effects of A within levels of B, expressed as deviations from the first category of A.

If we set $m = \bar{\mu}$, the overall weighted mean of μ_{jk}, we obtain the alternative form

$$\mu_{jk} = \bar{\mu} + r_j + c_k\,,\ j = 1 \text{ to } J\,;\ \ k = 1 \text{ to } K\,, \tag{2.4}$$

and then it is easy to show that $\{r_j\}$ and $\{c_k\}$ can be restricted so that they average to zero over their respective marginal distributions of counts in the sample. That is, if $\{n_{j+}\,; j = 1 \text{ to } J\}$ is the marginal distribution of factor A and $\{n_{+k}; k = 1 \text{ to } K\}$ is the marginal distribution of factor B, then

$$\sum_{j=1}^{J} n_{j+}r_j = \sum_{k=1}^{K} n_{+k}c_k = 0. \tag{2.5}$$

In this case, the quantities $\{r_j\}$ and $\{c_k\}$ still represent the effects of one factor within levels of the other factor, but now the effects are expressed as deviations from the overall mean.

For the data in Table D1, the effects of education on parity are not the same for all levels of marital duration. For example, the difference in mean parities between LVED = 1 and LVED = 4 is .96-.92 = 0.04 of MGP6 = 1 and 6.92-5.97 = 0.95 for MGP6 = 6. Table 2.2 presents hypothetical data where the effects are additive. For example, the difference in mean parities between LVED = 1 and LVED = 4 is 0.60 for all levels of marital duration. A visual check on additivity can be obtained by plotting the cell means and joining the means of one factor for each level of the other factor, as in Figure 2.1. If the effects are additive, as in Table 2.2, the result is a set of parallel piecewise linear curves (Figure 2.1.b). Deviations from parallel lines indicate interactions (Figure 2.1.a).

The term *association* refers to the second entries in the cells of Table D1, the cell counts or sample sizes. These reflect the joint distribution of the classifying factors in the sample. Two cross-classifying factors are *associated* if the distribution of one factor varies according to the level of the other factor. If there is no association, that is the distribution of one factor is the same for all levels of the other factor, then the two factors are said to be *independently distributed*, or *orthogonal*.

For the data in Table D1 it is clear that MGP6 and LVED are associated. The distribution of education is not the same for all levels of marital duration, and reflects the historical increase in education between marriage cohorts. For example, 27 per cent of the cohort married less than five years have 10 or more years of education, compared with 2 per cent for women married 25 or more years.

The concepts of interaction and association should be carefully distinguished. Confusion often arises from variations in terminology. The term interaction is sometimes used in both contexts. Also, the term independence is occasionally used to refer to what is described here as additivity.

2.3.2 Direct Standardization

We now concentrate on the variable educational level. The penultimate row of Table D1 gives the one-way cross-classification of mean parities discussed in the previous section, unadjusted

TABLE 2.3: Effects of Education from Table D1 Expressed as Deviations from The Reference Category, No Schooling

		Educational Level			
		No Schooling	1–5 Years	6–9 Years	10 or More Years
a) Unadjusted		–	–0.93	–1.91	–2.87
b) Duration-Specific Effects					
	0–4	–	–0.08	–0.01	–0.04
	5–9	–	–0.08	–0.15	–0.15
Marriage	10–14	–	–0.04	–0.14	–0.73
Duration	15–19	–	–0.16	–0.52	–1.00
	20–24	–	–0.35	–1.00	–1.75
	25 +	–	–0.37	–0.69	–0.95
c) Adjusted for MGP6 by test factor standardization		–	–0.16	–0.39	–0.71

for marital duration. These are weighted averages of the means in each column, with weights given by the distribution of MGP6 for each level of LVED. Since the factors are associated, the set of weights varies between the columns. We can adjust for the different composition of marital duration of each educational group by averaging the means in each column with the *same* set of weights. This technique is known as *direct standardization.* The choice of weights, or the *standard distribution,* is somewhat arbitrary. A simple choice is to give equal weights to each cell, obtaining the unweighted column means. Alternatively, we may weight proportional to the distribution of the adjusted factor (MGP6) in the whole sample, a variant of the method known as *Test Factor Standardization.* Other choices are also possible.

In symbols, direct standardization involves calculating the standardized column means

$$\tilde{y}_k = \sum_j w_j \bar{y}_{jk},$$

where the summation is over the row j, $\bar{y}_{jk}$ is the mean for row j and column k, and w_j is the weight for row j.

Test factor standardization can be applied to the data in Table D1, giving the standardized education means in the last row of the table. These are interpreted as the predicted mean parities for each level of education if women in that category had the distribution of marital duration in the entire sample.

Effects of education, adjusted for marital duration, are obtained from the standardized means by subtraction. They are displayed in row c) of Table 2.3, in the form of deviations from the standardized mean for the reference category NO SCHOOLING. Row a) of Table 2.3 gives the unadjusted effects of education as in the last row of Table 2.1, and the next six rows of the table give the effects of education calculated separately within each marriage duration Table 2.4 gives the corresponding estimates for the artificial data in Table 2.2.

We can use these tables to illustrate the consequences of association and interaction on the effects in a two-way table. Firstly, note that for both tables, additive and non-additive, the adjusted effects in row c) in Tables 2.3 and 2.4 are different from the unadjusted effects in row a); in fact here the adjusted effects are considerably smaller, although in other examples they may be larger. This impact of adjustment occurs because the factors are *associated.* Turning

TABLE 2.4: Effects of Education from Table 2.2 Expressed as Deviations from the Reference Category, No Schooling

		Educational Level			
		No Schooling	1–5 Years	6–9 Years	10 or More Years
a) Unadjusted		–	–0.92	–1.90	–2.86
b) Duration-Specific Effects					
	0–4	–	–0.24	–0.45	–0.60
	5–9	–	–0.24	–0.45	–0.60
Marriage	10–14	–	–0.24	–0.45	–0.60
Duration	15–19	–	–0.24	–0.45	–0.60
	20–24	–	–0.24	–0.45	–0.60
	25	–	–0.24	–0.45	–0.60
c) Adjusted for MGP6 by test factor standardization		–	–0.24	–0.45	–0.60

this statement round, we obtain the following property: *Property 1: If the factors A and B are independent (that is, not associated), then the unadjusted and adjusted effects of either factor are equal.* This property is hardly surprising, since the point of standardization is to deal with the consequences of association on the unadjusted effects.

The second property concerns the effects in rows b) and c) and the consequence of interaction. It can readily be shown that the adjusted effects in row c) can be obtained by averaging the duration specific effects in row b) with respect to the standard distribution. For example, the adjusted effect for the 1-5 years category in Table 2.3 is

$$-0.16 = [(-.08)(1280)+(-.08)(1231)+(-.04)(1118)+(-.16)(1051)+ (-.35)(893)+(-.37)(1231)]/6810$$

This property holds for any two-way table:

Property 2: The effects of B adjusted for A by standardization are weighted averages of the effects of B within each level of A, with weights given by the standard distribution.

In the presence of interaction, the effects of B vary according to the level of A, as seen in Table 2.3. Hence Property 2 implies that the adjusted effects vary according to the choice of standard distribution. On the other hand, if the effects are additive, as in Table 2.4, the effects of B are the same within each level of A, and the adjusted effects are obtained by averaging values which are all equal. Hence this averaging clearly is not affected by the choice of weights, that is, the standard distribution. Thus we obtain *Property 3: If A and B are additive (that is, do not interact in their effects on Y) then the adjusted effects of B equal the effects of B within any level of A, for any choice of standard distribution.*

We can prove Property 2 with a little algebra. Let μ_{jk} be the mean for the cell with level j of factor A and level k of factor B. Then adjusting the means of B for factor A by standardization consists in choosing a standard distribution of factor A,

$$\{w_j : j = 1, \ldots, J\} \tag{2.6}$$

and averaging the means of A within each level of B using these weights; that is, forming

$$\tilde{\mu}_k(w) = \sum_{j=1}^{J} w_j \mu_{jk} . \tag{2.7}$$

The argument (w) is used to emphasize the dependence of the adjusted mean on the choice of standard. The adjusted effects of B, expressed as deviations from the reference category k = 1, are given by

$$\{\tilde{\mu}_k(w) - \tilde{\mu}_1(w) : k = 2, \ldots, K\}$$

Substituting equation (2.7), we obtain

$$\begin{aligned} \tilde{\mu}_k(w) - \tilde{\mu}_1(w) &= \sum_{j=1}^{J} w_j \mu_{jk} - \sum_{j=1}^{J} w_j \mu_{j1} \\ &= \sum_{j=1}^{J} w_j (\mu_{jk} - \mu_{j1}) , \end{aligned}$$

which is an average of effects of B within levels of A with weights w_j. This proves Property 2.

Now suppose that the effects of A and B are additive. Then we showed above that the means μ_{jk} can be written as

$$\mu_{jk} = m + r_j + c_k ,$$

and $\{c_k\}$ represent the effects of B within any level of A. Hence by Property 3, they also represent the effects of B adjusted for A, for any choice of standard. Thus *if an additive table of means with factors A, B is expressed in the form (2.1), then $\{r_j\}$ are the effects of A adjusted for B and $\{c_k\}$ are the effects of B adjusted for A.* Furthermore, *if the constant term m is defined as μ_{11},* as in equation (2.2), *then the effects are expressed as deviations from the first category. If m is defined as the overall mean, as in equation (2.4), the effects are expressed as deviations from the overall mean.*

Some authors argue that standardization is an appropriate method of summary only if the effects are approximately additive. If large interactions are present, the adjusted effects depend on the choice of standard, and the summarization of the effects within levels of the other factor involves a loss of information. In the present example, for instance, the degree of interaction is considerable, as evidenced in Table 2.3.b). Substantively speaking, the educational differentials are small for the first two marriage cohorts. Reduced fertility of women with 10 or more years of education emerges in the third marriage cohort and persists thereafter. The differentials between the other education groups emerges only for the last three marriage cohorts. All this information is lost if the adjusted effects of education are summarized by the single row of Table 2.3.c).

Despite this loss of information, standardization still illustrates an essential feature of the data with or without the presence of interactions. The comparison of unadjusted and adjusted effects demonstrates clearly the compositional effect of marriage duration on the average education differentials. Specifically the differentials by educational level are greatly reduced when marriage duration is controlled.

To summarize, if the cross-classifying factors are independent then rows a) and c) are equal. If the effects of the cross-classifying factors are additive, then rows b) and c) are equal; otherwise the rows of b) are different and row c) is a weighted average, which varies according to the choice of standard.

2.4 Three-Way Cross-Classifications

The concepts and methods of the previous section can be readily extended to three-way cross-classifications. As an example, we analyze the table resulting from replacing the single demographic control marital duration in Table D1 by the two demographic controls respondent's

age and respondent's age at first marriage. Table D2 gives the cross-classification of Mean Parity by Educational Level, Current Age (AGP5) in five groups:

AGP5:	1	=	15-24 years
	2	=	25-29 years
	3	=	30-34 years
	4	=	35-39 years
	5	=	40-49 years

and Age at First Marriage (AMGP) in four groups:

AMGP:	1	=	< 15 years
	2	=	15-19 years
	3	=	20-24 years
	4	=	25 + years

The definitions of independence and additivity in the two way table were unambiguous, but for the three way table various extensions are possible. Suppose we denote the three variables of the cross-classification by A, B and C. One possibility is to view the three way table as a set of two way tables, one for each level of one of the factors. For example we might consider the set of two way tables of B and C for each level of A. Then we apply definitions of additivity and independence to this set of two way tables. Thus, B and C are *conditionally additive* given A if the effects of B and C are additive for all these two way tables. That is, within each level of A, the effects of C are the same for all levels of B. This structure is denoted by [A(B+C)] or alternatively [AB + AC].

A second form of additivity is obtained by combining two of the three variables, say A and B, into a single joint variable (AB) consisting of all combinations of levels of A and B. Then the three way table of means can be considered as a single two way cross-classification by (AB) and C. The definitions of the previous section can be applied to this table. Thus (AB) and C are additive in their effect on the response if the effects of C are the same for all levels of the joint variable AB. This structure is denoted by [AB + C]. For example Table D2 might be regarded as a two-way cross-classification of mean parity by C = LVED and AB = (AGP5. AMGP), the variable consisting of all combinations of AGP5 and AMGP, with 5 x 4 levels. Additivity in this two-way table, denoted by [LVED+AGP5.AMGP], means that the effects of education are the same for all levels of Age and Age at Marriage, an implausible structure for the present data. The structure [AB+C] does not make any assumptions about the pattern of means cross-classified by A and B within each level of C. If we assume in addition that the effects of A and B are conditionally additive given C, then we obtain a stronger form of additivity. We say that the effects of A, B and C are additive on the response, and denote this structure by [A+B+C].

Examples of the patterns [A (B+C)], [AB+C] and [A+B+C] are given in Table 2.5, for a 2x2x2 table. The patterns can also be expressed in symbols. Let $\mu_{jk\ell}$ be the mean response for level A = j, B = k, and C = ℓ of the cross-classifying variables. The three structures [A (B+C)] [AB+C] and [A+B+C] correspond to the following forms for the cell means:

$$[A\,(B+C)] \quad : \quad \mu_{jk\ell} = m_j + c_{jk} + s_{j\ell} \quad \text{for all j, k and } \ell \qquad (2.2)$$

$$[AB+C] \quad : \quad \mu_{jk\ell} = m + c_{jk} + s_{\ell} \quad \text{for all j, k and } \ell \qquad (2.3)$$

$$[A+B+C] \quad : \quad \mu_{jk\ell} = m + r_j + c_k + s_{\ell} \quad \text{for all j, k and } \ell \qquad (2.4)$$

To illustrate the correspondence, consider the difference in means between levels ℓ and ℓ' of variable C, for A = j and B = k.

Equation (2.2) gives

$$\mu_{jk\ell} - \mu_{jk\ell'} = m_j + c_{jk} + s_{j\ell} - (m_j + c_{jk} + s_{j\ell'}) = s_{j\ell} - s_{j\ell'}$$

and hence this difference depends on the level of A, j, but not on the level of B, k. Hence the

TABLE 2.5: Examples of Additive Data Patterns for a 2x2x2 Table of Means

a) [A(B+C)] B and C conditionally additive given A

A	B
1	1
1	2
2	1
2	2

	C: 1	C: 2
	5 7	8 10
	9 8	10 9

b) [A B+C] AB and C additive

A	B
1	1
1	2
2	1
2	2

	C: 1	C: 2
	5 7	8 10
	9 8	12 11

c) [A+B+C] A, B and C additive

A	B
1	1
1	2
2	1
2	2

	C: 1	C: 2
	5 7	8 10
	9 11	12 14

effects of B and C within each level of A are additive, as required by the model [A(B+C)]. Equations (2.3) and (2.4) both give

$$\mu_{jk\ell} - \mu_{jk\ell'} = s_\ell - s_{\ell'}$$

which implies that the effect of C is the same for all levels of A and B. However equation (2.4) also gives

$$\mu_{jk\ell} - \mu_{j'k\ell} = r_j - r_{j'}, \quad \mu_{jk\ell} - \mu_{jk'\ell} = c_k - c_{k'},$$

which implies that the effects of A and B are also additive. This additional property is not shared by equation (2.3). There is clearly a *hierarchy* between the three data patterns in that [A+B+C] implies [AB+C] and [AB+C] implies [A(B+C)].*

The concept of association between the cell counts generalizes to the three way table in much the same way as that of additivity between the cell means. Thus B and C are *conditionally independent given A* if the distribution of B and C are independent within each level of A. The

* Other patterns exist for a three-way table. The variables A, B, and C can be permuted. One or more effects can be assumed equal to zero, leading to one or two way tabulations by summing over factors. Finally one model [AB +BC+CA], cannot be described in terms of two way tables. A full description of these models is given in another technical bulletin (Little, 1978).

TABLE 2.6: Distribution of Sample, by Age and by Age at Marriage

	AMGP			
AGP5	<15	15–19	20–24	25 +
15–24	114	688	285	0
25–29	165	474	520	138
30–34	175	502	305	240
35–39	199	455	331	218
40–49	330	873	489	310

TABLE 2.7: Standardization on Age and on Age at Marriage

	Educational Level			
	No Schooling	0–5 Years	6–9 Years	10 or More Years
Mean Parity Standardized for AGP5 and AMGP	4.13	3.96	3.78	3.10
Effects of Education	–	–0.17	–0.35	–1.03

joint variable AB and the variable C are independent if the distribution of C is the same for all levels of the joint factor AB. Finally, A, B and C are independently distributed if AB and C are independent and A and B are conditionally independent given C. These structures are important in the analysis of contingency tables, but are not considered in detail in the present context.

The method of direct standardization can be applied to calculate the effects of one factor, adjusted for the other two. For our example we are interested in calculating education effects adjusted for age and age at marriage. We shall once again use test factor standardization, applying the distribution of AGP5 and AMGP for the whole sample, given in Table 2.6, to the set of means for each educational level.

The present application illustrates a practical problem of the method which also has analytical consequences. The distribution of the sample over the cells of the three way table is not uniform, and some cells are empty. The four cells with AGP5 = 15-24 and AMGP = 25 + are empty because they are unobservable. Also two cells with LVED = 4, and AMGP = 1 are empty, because very few women with 10 or more years of education were married before age 15. In applying the standard in Table 2.6 to the data in Table D2, means are not required for the unobservable cells since they are given weight zero; however means are required for the empty cells with LVED = 4, AMGP = 1, since they are given positive weight in the standardized mean. Here the present values for the adjacent group with AMGP = 2 were imputed for these cells. This procedure introduces a small bias into the final estimates. More generally, the method of standardization can give unduly large weights to cell means which are based on very few observations, and hence have large variances. In more technical terms, it is a statistically inefficient method of calculating adjusted effects. Hence it should be used with caution when the sample sizes become small. In the next chapter we consider another method for calculating adjusted effects which is statistically optimal under certain conditions, namely multiple classification analysis.

We conclude this introductory section by presenting the results of applying test factor standardization to the education means. The adjusted means and effects expressed as deviations from the NO SCHOOLING group, are given in Table 2.7.

We observe that the adjusted effects are quite similar to those standardized for marital duration, given in Table 2.4.c). Hence it appears that in this case marital duration is a reasonable proxy for the demographic control of age and age at marriage.

Under what circumstances are the means of one factor, say C, adjusted for the other two factors, say A and B, an adequate summary of the effects of C? As in the previous section, the standardized effects are weighted average of the effects within each level of A and B, which are constant if and only if the effects AB and C are additive. Hence standardization is particularly appropriate when AB and C are additive (or, a fortiori, A, B and C are additive).

3. ADDITIVE MODELS BETWEEN FACTORS: ANALYSIS OF VARIANCE AND MULTIPLE CLASSIFICATION ANALYSIS

3.1 Introduction

Direct standardization is a simple and convenient method for calculating adjusted means and effects. However if cell sample sizes are small, because of limitations in the total sample size or because of the degree of cross-classification, the method is not entirely satisfactory; procedures are required to deal with empty cells, and the large sampling variances of cell means based on small numbers of observations are not taken into account in the final estimates. Also, as sample sizes diminish the statistical significance of effects may be called into question. Standardization does not supply estimates of the statistical significance of effects, and although these can be calculated, a more satisfactory approach is available, namely, analysis of variance.

In the first chapter we noted the decomposition of the observed values of the dependent variable into fitted values under some model, and residuals representing departures from the model. Before proceeding further it is convenient to relate the particular methods of the previous chapter to this general conceptual framework.

We began with the simplest model which summarizes the individual values of the dependent variable in a single number, the mean $\bar{y}$. The corresponding decomposition, given as equation (1.1), was

$$y_i = \bar{y} + (y_i - \bar{y}) \tag{3.1}$$

$$\text{observed} = \text{fit} + \text{residual}$$

The next step was to construct a one way cross-tabulation of the means of y by the factor LVED = Level of Education. To write down the corresponding decomposition, relabel the values so that y_{ij} is the parity for individual i within category j of level of education. Then the fitted value under the model is $\bar{y}_j$, the mean parity for education level j, and the decomposition is

$$y_{ij} = \bar{y}_j + (y_{ij} - \bar{y}_j). \tag{3.2}$$

$$\text{observed} = \text{fit} + \text{residual}$$

We denote the model underlying this decomposition by [LVED]. The third step was to further cross-tabulate by the factor MGP6 = Marital Duration. The decomposition corresponding to this cross-tabulation is

$$y_{ijk} = \bar{y}_{jk} + (y_{ijk} - \bar{y}_{jk}), \tag{3.3}$$

$$\text{observed} = \text{fit} + \text{residual}$$

where now y_{ijk} denotes the parity for individual i within the cell with educational level j, marital duration level k, and $\bar{y}_{jk}$ is the mean parity of individuals in this cell. We denote the model underlying this cross-tabulation by [LVEDxMGP6]. The extension to three-way tables is clear.

How does the technique of standardization fall into this scheme? Let us consider the case of the two-way table. We have noted the strong relationship between standardization and an *additive* structure for the cell means. Namely, if the means have an additive structure then the effects of one factor within levels of other factor are equal and given by the adjusted effects from standardization for any choice of standard. Moreover, the means inside the table can be constructed from the adjusted means found by the method. In fact, ***direct standardization corresponds to fitting an underlying additive model between the factors.***

To write the data decomposition, let $\bar{y}_{jk}$ be the mean for the cell with LVED = j and MCP6 = k.

Suppose we form adjusted means for *both* factors by standardizing with respect to standard distributions $\{w_j\}$ for education and $\{v_k\}$ for marital duration. That is, we form adjusted means

$$\tilde{y}_k = \sum_j w_j \bar{y}_{jk} \; ; \; \tilde{y}_j' = \sum_k v_k \bar{y}_{jk} .$$

Let m be the (weighted) average of the adjusted means of either factor, that is,

$$m = \sum_k v_k \tilde{y}_k = \sum_j w_j \tilde{y}_j' .$$

It is readily shown that both expressions for m are equal. Now let r_j and c_k be the adjusted *effects* of the factors, expressed as deviations from m. That is,

$$r_j = \tilde{y}_j - m \, , \, c_k = \tilde{y}_k - m .$$

Then the fitted values from direct standardization take the additive form

$$\hat{\mu}_{jk} = m + r_j + c_k . \tag{3.4}$$

What we have constructed is *the additive structure* $\{\mu_{jk}\}$ *which would have given the adjusted effects and mean that we have obtained by standardizing the observed means* $\bar{y}_{jk}$. The decomposition corresponding to the method is

$$y_{ijk} = \hat{\mu}_{jk} + (y_{ijk} - \hat{\mu}_{jk}) , \tag{3.5}$$

$$\text{observed} = \text{fit} + \text{residual}$$

where y_{ijk} is the individual parity defined as for equation (3.3), and μ_{jk} is given by equation (3.4). The decomposition corresponds to an additive model for the factors, which we write [LVED+MGP6].

Why should we construct fitted means $\{\hat{\mu}_{jk}\}$ by this elaborate procedure when the population means are readily estimated by the sample means $\{\bar{y}_{jk}\}$? There are three principal reasons. Firstly, the additive means are based on a small number of parameters, J+K-1 for a JxK table, and hence are more stable than the observed means in the table, particularly when the sample sizes are small. That is, fitting the additive model effectively smooths the observed means and reduces sampling fluctuations. Secondly, the additive model provides a summary of the table, the adjusted means, which have an appealing substantive interpretation. Thirdly, the deviations of the fitted means $\hat{\mu}_{jk}$ under the additive model from the sample means $\bar{y}_{jk}$ are convenient quantities for studying the pattern of interactions in the table. If the residuals in (3.5) are further decomposed in the form

$$y_{ijk} - \hat{\mu}_{jk} = \bar{y}_{jk} - \hat{\mu}_{jk} + y_{ijk} - \bar{y}_{jk} , \tag{3.6}$$

then the components $\bar{y}_{jk} - \hat{\mu}_{jk}$, are the average residuals in each cell, and represent deviations from the additive structure for the cell means, and the components $y_{ijk} - \bar{y}_{jk}$ measure within cell variations in the dependent variable.

The results of constructing the fitted values for the data in Table D1 are given in Table 3.1. The first entries in the body of the table are the cell sample means, the second entries are the fitted values, and the third entries are the residuals. The last column gives the adjusted means of MGP6 and the last row gives the adjusted means of LVED. The final entry, 3.88, is the value of m, the weighted average of the adjusted means. Note that this does *not* equal the overall sample mean for the data, 3.84, an unfortunate characteristic of any form of standardization. The form

TABLE 3.1: Mean Number of Children Ever Born, by Marital Duration and by Level of Education: A) Observed Means, B) Fitted Means from Test Factor Standardization, C) Average Residuals

MGP6 Marital Duration (Years)	Lved: Level of Education				
	No Schooling (1)	1–5 Years (2)	6–9 Years (3)	10 or More Years (4)	Adjusted Means of MGP6
0–4 (1)	.96 a) 1.18 b) –.22 c)	.88 1.02 –.14	.95 .79 –.16	.92 .47 .45	.92
5–9 (2)	2.54 2.71 –.17	2.46 2.55 –.09	2.39 2.32 .07	2.39 2.00 .39	2.45
10–14 (3)	3.87 4.01 –.14	3.91 3.85 .06	3.73 3.62 .11	3.14 3.30 –.16	3.75
15–19 (4)	5.13 5.06 .07	4.97 4.90 .07	4.61 4.67 –.06	4.13 4.35 –.22	4.80
20–24 (5)	6.22 5.86 .36	5.87 5.70 .17	5.22 5.47 –.25	4.47 5.15 –.68	5.60
25 + (6)	6.92 5.86 1.06	6.55 5.70 .85	6.23 5.47 .76	5.97 5.15 –.82	6.47
Adjusted Means of Lved	4.14	3.98	3.75	3.43	3.88

chosen is as before test factor standardization; thus the last row is the same as the last row of Table D1.

To obtain the fitted value for row j and column k, the adjusted means for row j and column k are summed and then the overall mean is subtracted.* For example, for row 2, column 3 we obtain the fitted value

$$\hat{\mu}_{23} = 2.45 + 3.75 - 3.88 = 2.32 .$$

Finally, the average residuals are calculated by subtracting the fitted means from the observed means in each cell.

If the data were exactly additive, the observed and fitted values would be equal and the average residuals would all be zero. The average residuals thus represent interaction effects between the factors. These show a systematic pattern; namely they tend to be negative in the north-west and south-east portions of the table, and positive in the north-east and south-west portions. This pattern arises from fitting average adjusted effects for education which are too large for low marital durations and not large enough for high marital durations.

* This is clearly equivalent to *adding* m to the sum of adjusted effects, $r_j + c_k$.

The average residuals are subject to certain restrictions; that is each row or column averages to zero over its respective standard distribution. Different standard distributions give different sets of residuals, although with similar patterns.

3.2 Multiple Classification Analysis

We now ask the question, is there a set of fitted values with an additive structure which fits the data better, in the sense of yielding smaller average residuals. The answer to this question is yes, and the method which finds (in fact, is defined to find) a best fitting additive structure is multiple classification analysis (MCA).

To define what is meant by "best", we need a criterion for measuring the fit. For MCA, this is given by the sum of squares of the average residuals, weighted by the sample size in each cell, that is

$$ss = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk} (\bar{y}_{jk} - \hat{\mu}_{jk})^2 \tag{3.7}$$

Thus for the two-way table, MCA calculates fitted values

$$\hat{\mu}_{jk} = m + \hat{r}_j + \hat{c}_k \tag{3.8}$$

TABLE 3.2: Mean Number of Children Ever Born, by Marital Duration and by Level of Education; A) Observed Means, B) Fitted Means from MCA, C) Average Residuals

	Lved: Level of Education				
MGP6 Marital Duration (Years)	No Schooling (1)	1–5 Years (2)	6–9 Years (3)	10 or More Years (4)	Adjusted Means of MGP6
0–4 (1)	.96 a) 1.31 b) –.35 c)	.88 1.07 –.19	.95 .86 +.09	.92 .71 .21	.92
5–9 (2)	2.54 2.78 –.24	2.46 2.54 –.08	2.39 2.33 .06	2.39 2.18 .21	2.49
10–14 (3)	3.87 4.06 –.19	3.91 3.82 .09	3.73 3.61 .12	3.14 3.46 –.32	3.77
15–14 (4)	5.13 5.11 .02	4.97 4.87 .10	4.61 4.66 –.05	4.13 4.51 –.38	4.82
20–24 (5)	6.22 6.01 .21	5.87 5.77 .10	5.22 5.56 –.34	4.47 5.41 –.94	5.72
25 + (6)	6.92 6.82 .10	6.55 6.58 –.03	6.23 6.37 –.14	5.97 6.22 –.25	6.53
Adjusted Means of Lved	4.23	3.99	3.78	3.63	3.94

TABLE 3.3: Stem and Leaf Plots* Comparing the Distribution of Absolute Residuals x 100 from Test Factor Standardization and from MCA

Test Factor Standardization								Multiple Classification Analysis						
						6	100–109							
							90–99	4						
					5	2	80–89							
						6	70–79							
						8	60–69							
							50–59							
						5	40–49							
					9	6	30–39	2	4	5	8			
				5	2	2	20–29	1	1	1	4	5		
7	7	6	6	4	4	1	10–19	0	0	0	2	4	9	9
	9	7	7	7	6	6	0–9	2	3	5	6	8	9	9

* Note: A *Stem and Leaf Plot* invented by J.W. Tukey, is a histogram on its side with the individual data values retained. The *stem* is a set of grouping invervals, here in the centre of the diagram, varying from 0-9 to 100-109. The *leaves* are the values falling in each interval, ordered and represented by their last digits. Thus the leaf for MCA corresponding to the 10-19 group.

0 0 0 2 4 9 9,

represents the set of absolute residuals 10, 10, 10, 12, 14, 19, 19. The outline of the plot gives the shape of the distribution. However, unlike a histogram the values are retained. For further details, see Tukey (1977).

which minimize the weighted sum of squared residuals, ss. The fitted values from MCA are given in Table 3.2, and are in fact the means given earlier in Table 2.2.

A comparison of the (average) residuals in Tables 3.1 and 3.2 reveals the improvement in the fit obtained by MCA. The distributions of the absolute residuals are compared in Table 3.3, and indicate six residuals with absolute values of more than forty from standardization, compared with only one from MCA.

The large residual from MCA illustrates an important property of MCA, and in particular the criterion sum of squares (3.7), namely that it weights the squared residuals by the sample size n_{jk} in each cell. Thus empty cells are given weight zero, that is are effectively ignored. Cells with small counts are given less weight than cells with large counts, which implies that the fitted values are allowed to deviate more from the observed means if the observed means are based on small samples and are thus subject to a high variance. This rather sensible property is not shared by the fitted values from standardization.

Compare, for example, the residuals from the cell with MGP6=5 and LVED=4, with n_{54}=40 observations, with the cell with MGP6=6 and LVED=1, with 501 observations. Standardization yields residuals of -.68 and 1.06 respectively for these cells. Thus the fit is worse for the cell with a well determined mean. MCA yields residuals of -.94 and .10 respectively. Thus the fit is worse than that from standardization for the cell with 40 observations, but is very good for the cell with 501 observations. The conclusion is that if the sample sizes are small enough for sampling error to be substantial, then MCA is a much more sensible way of determining the fitted values, and hence of calculating adjusted effects. Finally, the abnormally large negative residual from MCA for the cell with MGP6=5, LVED=4, suggests that the population mean for that cell may be underestimated by the sample mean, 5.41. An estimate of about 4.9 may be more reasonable on the basis of neighbouring residuals.

The adjusted means from MCA are presented in the margins of Table 3.2, and the last entry is their weighted average, m=3.84. Note that this equals the weighted mean of the sample. This no accident: it can be shown that the value of m which with the other parameters minimizes the

weighted sum of squares is always the weighted sample mean $\bar{y}$. Another way of saying this is that the adjusted effects, taken about the sample mean, average to zero. Comparison of the adjusted means of MGP6 and LVED in Table 3.2 and 3.3 indicate quite small differences, the largest being the means for the last education group (3.63 for MCA, 3.43 for standardization). Thus the methods differ more in the fitted values and residuals inside the table than in the average effects in the margins.

It is customary in MCA to present effects as deviations from the overall mean, rather than as deviations from a reference category as in Table 2.4. The output is given in the form of an MCA table. The table for the data in Table D1 is presented in Table 3.4. The first column of the table gives the sample counts for each category of the cross-classifying factors. The second column gives the set of unadjusted deviations from the overall mean and the fourth column gives the set of deviations for each factor, adjusted for the other factor. The third and fifth columns give summary measures of the effects, Eta and Beta, which are discussed in more detail later. For the factor Educational Level the Eta value for the unadjusted deviations is .32 and the Beta value for the adjusted deviations is much lower, .07, indicating the reduction in the effect of Educational Level when Duration is controlled. Note that by subtracting the deviations for LVED=2, 3 and 4 from the deviations for LVED=1, we obtain the unadjusted and adjusted effects in the form of Table 2.4 a) and c).

3.3 Analysis of Variance: Introduction

So far we have been concerned with estimating sets of effects for factors, unadjusted or adjusted for associated factors. We now describe a way of summarizing a set of effects in a single number, indicating the overall magnitude of the differences between the category means. There are two reasons for doing this. The first is to derive tests for the statistical significance of the effects, that is to find out whether the observed differences could be attributable to sampling fluctuations rather than real differences in the population means. The second reason is to allow a simple comparison of the effects of a particular variable when adjusted for a variety of the other factors. The first of these reasons is probably the more important of the two. The basic measures employed are the *sum of squares* (SS) for an effect, and a closely related quantity the *mean square* (MS). The method of calculation is called *analysis of variance.*

Analysis of Variance is closely related to the decomposition of the observed values of the response into fitted values and residuals. That is,

$$\text{observed}_i = \text{fit}_i + \text{residual}_i ,$$

where i is again a subscript denoting the individual. Squaring this equation and summing over individuals i, we obtain

$$\sum_i \text{observed}_i^2 = \sum_i (\text{fit}_i + \text{residual}_i)^2$$

$$= \sum_i \text{fit}_i^2 + \sum_i \text{residual}_i^2 + 2 \sum_i \text{fit}_i \,.\, \text{residual}_i \,.$$

It can be shown that if the fit component is calculated so that the sum of squared residuals is minimized then the last term on the right hand side is equal to zero. That is, we have

$$\sum_i \text{observed}_i^2 = \sum_i \text{fit}^2 + \sum_i \text{residual}_i^2 \,.$$

If the observed and fitted values are measured around the overall sample mean, we obtain the basic equation of analysis of variance. That is, if y_i is the observed value, $\hat{y}_i$ is the fitted value, and $y_i - \hat{y}_i$ is the residual, then

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \qquad (3.9)$$

The left hand side represents a summary measure of the variation of the individual values y_i

TABLE 3.4: Multiple Classification Analysis of Parity, by Marital Duration and by Level of Education

Grand Mean = 3.94							
		Unadjusted		Adjusted for Independents		Adjusted for Independents + Covariates	
Variable + Category	N	Dev'n	Eta	Dev'n	Beta	Dev'n	Beta
MGP6							
1 0–4	1280	−3.02		−2.92			
2 5–9	1231	−1.51		−1.45			
3 10–14	1118	−.19		−.17			
4 15–19	1057	.90		.88			
5 20–24	893	1.84		1.78			
6 25 +	1231	2.71		2.59			
			.70		.68		
LVED							
1 No Schooling	1512	1.23		.29			
2 1–5 Years	2686	.30		.05			
3 6–9 Years	1704	−.69		−.16			
4 10 + Years	908	−1.64		−.31			
			.32		.07		
Multiple R Squared					.497		
Multiple R					.705		

about the mean; in fact it is proportional to the variance. The first term on the right hand side summarizes the variation accounted for by the fitted values from the model, and the remainder summarizes the variation not accounted for in the model. All analysis of variance tables are constructed from decompositions of this simple form.

3.4 One-Way and Two-Way Analysis of Variance

For a one-way classification by factor A let y_{ij} be the value of the dependent variable for individual i within category j of the classification. The fitted values for individuals in category j are all equal to $\bar{y}_j$, the sample mean for that category, and the decomposition of the data is given by equation (3.2). The analysis of variance decomposition is thus

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (\bar{y}_j - \bar{y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_j)^2 .$$

$$SS_T = SS_A + SS_{RES}$$

The sum of squares SS_A measures the variation of the response between the categories of the factor A, and hence is the sum of squares associated with the factor A. The sum of squares SS_{RES} measures the variation of the response within the categories of the factor A, and is called the error or residual sum of squares. These statistics can be entered in a one-way ANOVA table, as in Table 3.5.

The first column of the ANOVA table indicates the source of variation, and the second column presents the sum of squares associated with each source. The third column gives the degrees of freedom, which equals the number of independent parameters associated with each source. If there are N observations and the factor A has J levels, there are N-1 degrees of freedom for SS_T (the number of observations less one degree of freedom for the grand mean), and this decomposes into J-1 degrees of freedom for the effects of A and N-J degrees of freedom for the residual. The next column gives the mean square for each source, defined as the sum of squares

TABLE 3.5: One-Way ANOVA Table for a Factor A.

Source of Variation	SS	df	MS	F	Significance of F
A	SS_A	J–1	$MS_A = SS_A/(J-1)$	MS_A/MS_{RES}	P
Residual	SS_{RES}	N–J	$MS_{RES} = SS_{RES}/(N-J)$	–	
Total	SS_T	N–1	$MS_T = SS_T/(N-1)$		

TABLE 3.6: One-Way Analysis of Variance of Parity, by Level of Education

Source of Variation	Sum of Squares	DF	Mean Square	F	Significanc of F
Main Effects	5747.380	3	1915.793	261.694	.000
LVED	5747.380	3	1915.793	261.694	.000
Explained	5747.383	3	1915.794	261.694	.000
Residual	49817.548	6805	7.321		
Total	55565.031	6808	8.162		

divided by the degrees of freedom. The magnitude of the effects of A can be compared with the average within cell variability by comparing MS_A with MS_{RES}.

Since the survey data are based on a sample of the population, the means for each category of A can differ when there is no difference in the means for the population from which the sample is drawn. Hence a test of statistical significance of the effects of A is desirable. This is achieved by the F-test in the last two columns of Table 3.5. If i) the data are a random sample from the population, ii) the variance of Y within each category of A is constant, iii) the population means in each category of A are in fact equal, and iv) the cell means are normally distributed, then

$$F = MS_A / MS_{RES}$$

has an F distribution with (J-1) and (N-J) degrees of freedom. The significance level of the F statistic is given in the last column of Table 3.5. For example, if $P \geqslant 0.05$ then the effects of A are not significant at the 5% level.

The one-way ANOVA for the data in Table 2.1 is presented in Table 3.6. The mean square for the effects of educational level is 1915.8, compared with a residual mean square of 7.32. These yields a highly significant F value of 261.7. That is, the difference in the unadjusted education means cannot be attributed to random fluctuations.

For a two-way table with factors A and B more than one decomposition of the total sum of squares is possible. Treating AB as a single factor, we obtain as before

$$SS_T = SS_{AB} + SS_{RES}$$

where SS_{AB} is the sum of squares for the joint factor (AB) and SS_{RES} is the residual sum of squares. Then SS_{AB} can be decomposed into SS_{A+B}, the sum of squares for main effects of A

and B assuming an additive model, and $SS_{A.B}$, the sum of squares for the interaction of A and B, adjusted for the main effects of A and B:

$$SS_{AB} = SS_{A+B} + SS_{A.B}.$$

Finally the sum of squares for the main effects, SS_{A+B} can be distributed between the effects of A and B in two ways.

$$SS_{A+B} = SS_A + SS_{B|A}$$

$$SS_{A+B} = SS_{A|B} + SS_B$$

where SS_A and SS_B are the sum of squares for the unadjusted effects of A and B, and $SS_{A|B}$ and $SS_{B|A}$ are the sum of squares for the effects of A adjusted for B and B adjusted for A. Hence the full decompositions are

$$SS_T = \underset{\text{A, unadjusted}}{SS_A} + \underset{\text{B, adjusted for A}}{SS_{B|A}} + \underset{\text{AB interaction, adjusted for A + B}}{SS_{A.B}} + \underset{\text{residual}}{SS_{RES}}$$

$$\text{or } SS_T = \underset{\text{B, unadjusted}}{SS_B} + \underset{\text{A, adjusted for B}}{SS_{A|B}} + \underset{\text{AB interaction, adjusted for A + B}}{SS_{A.B}} + \underset{\text{residual}}{SS_{RES}}$$

Note that when A and B are not associated, the adjusted and unadjusted effects of A or B are equal. Hence one would expect the sum of squares of these effects to be equal, and this is indeed the case. That is, $SS_A = SS_{A|B}$ and $SS_B = SS_{B|A}$, and the two alternative decompositions of SS_T are the same. A special case of this is *balanced* analysis of variance, where the cell sample sizes are all equal.

The sums of squares from these decompositions are presented in a two-way analysis of variance table. Two common modes of presentation, *classical* and *hierarchical*, are illustrated for the data of Table D1 in the ANOVA Tables 3.7 and 3.8. Both tables present sums of squares for the main effects, SS_{A+B}, the interactions $SS_{A.B}$, the total explained by AB, SS_{AB}, the residual, SS_E, and the total, SS_T, in the rows as indicated. In a *classical* analysis of variance (Table 3.7), the adjusted sums of squares for each of the main effects A and B, $SS_{A|B}$ and $SS_{B|A}$, are presented. Note that these do not add up to SS_{A+B}. For example in Table 3.7, A = MGP6, B = LVED and

$$SS_{MGP6|LVED} + SS_{LVED|MGP6} \neq SS_{MGP6+LVED}$$

$$218808 + 225.5 \neq 27628.2$$

In a *hierarchical* analysis of variance (Table 3.8) the sums of squares for the first effect specified on the ANOVA control card is unadjusted, and the second effect is adjusted. Thus if A is specified first, SS_A and $SS_{B|A}$ are presented, and if B is specified first, SS_B and $SS_{A|B}$ are presented. In both cases the sums of squares do add up to SS_{A+B}. For example, in Table 3.8,

TABLE 3.7: Classical Two-Way Analysis of Variance of Parity, by Marital Duration and by Level of Education

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	27628.219	8	3453.527	845.017	.000
MGP6	21880.840	5	4376.168	1070.770	.000
LVED	225.538	3	75.179	18.395	.000
2-Way Interactions	206.965	15	13.798	3.376	.000
MGP6 LVED	206.963	15	13.798	3.376	.000
Explained	27835.184	23	1210.225	296.121	.000
Residual	27729.848	6785	4.087		
Total	55565.031	6808	8.162		

TABLE 3.8: Hierarchical Two-Way Analysis of Variance of Parity, by Marital Duration and by Level of Education

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	27628.219	8	3453.527	845.017	.000
MGP6	27402.684	5	5480.537	1340.990	.000
LVED	225.535	3	75.178	18.395	.000
2-Way Interactions	206.965	15	13.798	3.376	.000
MGP6 LVED	206.963	15	13.798	3.376	.000
Explained	27835.184	23	1210.225	296.121	.000
Residual	27729.848	6785	4.087		
Total	55565.031	6808	8.162		

TABLE 3.9: Hierarchical ANOVA Table of Parity on Marital Duration and Level of Education, for British Data in Table D3.

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	2714.299	7	387.757	239.473	.000
MGP5	2641.761	5	528.352	326.303	.000
LVED	72.538	2	36.269	22.399	.000
2-Way Interactions	18.959	10	1.896	1.171	.305
MGP5 LVED	18.959	10	1.896	1.171	.305
Explained	2733.258	17	160.780	99.295	.000
Residual	8100.891	5003	1.619		
Total	10034.148	5020	2.158		

$$SS_{MGP6} + SS_{LVED|MGP6} = SS_{MGP6+LVED},$$

$$27402.7 + 225.5 = 27628.2$$

In SPSS, the hierarchical ANOVA is obtained by specifying OPTION 10 on the OPTIONS cards. We shall generally adopt the hierarchical form since the sums of squares add up correctly. Note, however, that for the two way table the classical ANOVA gives more information since both hierarchical ANOVAs can be derived from it by subtraction, but the reverse procedure is impossible.

As noted above, the additive parts of the ANOVA table are based on the same model as that used for Multiple Classification Analysis. In fact the Eta and Beta measures in the MCA Table are derived from sums of squares in the ANOVA table. Specifically the squares of ETA and BETA for an effect A are

$$ETA^2 = SS_A/SS_T \quad BETA^2 = SS_{A|B} / SS_T$$

and are interpreted as the proportions of the total variance explained by the unadjusted and adjusted effects of A, respectively.

Two way analysis of variance again allows us to test the statistical significance of effects. A test for the significance of the interactions is obtained by comparing the interaction mean square, $MS_{A.B}$, with the residual mean square, MS_{RES}. Specifically, the F statistic, $MS_{A.B}/MS_{RES}$, is compared with the tabulated F distribution with the degrees of freedom of the interaction A.B in the numerator and of the residual in the denominator. In SPSS output, F values are given in the penultimate column of the table, and the last column gives the P-value, the probability of obtaining a value of F higher than that observed under the null hypothesis that the interactions are zero in the population. For example, in Table 3.8, the F-value for the interactions is:

$$13.798/4.087 = 3.376,$$

which is highly significant, giving a P-value indistinguable from zero to three decimal places. Thus the additive model [A+B] does not fit the data, a result which confirms the visual inspection of the data given in Section 2.2.

Table 3.8 also gives F-values for the unadjusted effects of MGP6 (F=1340.990) and for the effects of LVED adjusted for MGP6 (F=18.395), both highly significant. The former confirms the obvious fact that differences in parity exist between duration groups. The latter the less obvious finding that significant educational differences persist after marital duration is controlled.

An important objection can be raised to testing for the significance of the adjusted effects of education in this example. As we noted in Chapter 2, the adjusted effects are uniquely defined only if the additive model [MGP6 + LVED] holds for the population means. However, this is never likely to be exactly correct in practice, and in the present example we have strong evidence that it is not the case, since the interactions are highly significant. Thus testing the significance of adjusted effects has little point when the interactions have been shown to be non-zero.

Nevertheless, even in the presence of interaction, the statistics for the main effects in the ANOVA table still have some value, as summary measures of the size of an average of the adjusted main effects. The effect of the control of marital duration on the education differentials is illustrated by the reduction of the mean square for education from $MS_{LVED} = 261.7$ (Table 3.6) to $MS_{LVED|MGP6} = 75.2$ (Table 3.7). The latter compares with the interactions mean square of 13.8 and the residual mean square of 4.09. Thus the average size of the adjusted main effects is considerably greater than the average size of the interactions, although the latter are statistically significant.

In general we would expect a significant interaction between these variables on a priori grounds, to the extent that differentials according to educational level emerge as marital duration increases. However this is not always the case. Equivalent data for the British survey (Table D3) indicate an initial differential by educational level in the first marriage duration group which is maintained at a similar level for successive marriage duration groups. The resulting analysis of variance, Table 3.9, reveals no evidence of a significant interaction effect. The issue of interactions is considered further in Section 3.5.1.

It should be pointed out that the assumptions underlying the F tests in this example are violated, and hence the significance levels can be regarded as at best approximate. The tests assume simple random sampling, whereas all WFS samples are based on complex sampling designs involving stratification and clustering. The effect of this on significance tests is largely unknown, although there are reasons to believe that for WFS data it is not critical. More important for the present example is the assumption that the variance of the response within the cells of the table is constant, (the assumption known as *homoscedasticity*). This is clearly untenable for the present example, since the variance of parity clearly increases with marital duration. This is confirmed by the sample standard deviations in Table D1. Other situations where the variance is not constant are data on binary responses, such as Current Use of Contraception (1 = Yes, 0 = No), where the cell means lie close to zero or one. In these cases the variances of the response decreases as the mean approaches the limiting values, zero and one.

For the present example the lack of homoscedasticity seriously distorts the significance levels, and some weighting of the individual observations is desirable. This is described in Section 3.5. However even after allowance is made for gross departures from the assumptions, we are rarely in a situation to interpret significance tests exactly. This does not mean that the statistical analysis is rendered useless, but rather that the statistics should be regarded as useful diagnostic measures derived from the data, and should not be used to construct strict 5%/95% cut off points for deciding whether an effect is present or not.

3.5 Three-Way and Higher Tables

The methods of multiple classification analysis and analysis of variance are particularly valuable for multiway tables involving three or more factors. Simple techniques such as standardization become awkward to apply because of empty cells, and it becomes increasingly advisable to use a statistical model to smooth the cell means.

For a three-way table of means with factors A, B, and C the analysis of variance is based on the decomposition

$$SS_T = SS_{ABC} + SS_{RES}$$

$$= SS_{A+B+C} + SS_{A.B+B.C+C.A} + SS_{A.B.C} + SS_{RES}$$

where SS_{RES} is the residual sum of squares, SS_{ABC} is the sum of squares for the joint variable (ABC), which decomposes into SS_{A+B+C}, the sum of squares for the main effects of A, B and C, $SS_{A.B+B.C+C.A}$, the sum of squares for the two-way interactions adjusted for the main effects, and $SS_{A.B.C}$, the sum of squares for the three-way interactions adjusted for the main effects and the two-way interactions. Note the distinction drawn here between interactions of different orders. Two-way interactions measure differences in the effect of one factor between levels of a second factor, averaged over the third factor. Three-way interactions measure differences in the two-way interactions between levels of the third factor. For higher way tables interactions between four or more factors are defined in a similar way.

High order interactions are a problem in two respects; they are hard to interpret, and they are numerous, particularly for cross-tabulations involving factors with more than two or three

TABLE 3.10: Classical Three-Way Analysis of Variance of Parity by Age, by Age at Marriage, and by Level of Education Treated as a Dichotomy

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	27413.930	8	3426.741	834.409	.000
AGP5	17964.504	4	4491.126	1093.586	.000
AMGP	7919.861	3	2639.954	642.827	.000
LVED	105.523	1	105.523	25.695	.000
2-Way Interactions	326.262	19	17.172	4.181	.000
AGP5 AMGP	210.000	12	17.500	4.261	.000
AGP5 LVED	49.702	4	12.425	3.026	.017
AMGP LVED	53.307	3	17.769	4.327	.005
3-Way Interactions	21.879	11	1.989	.484	.914
AGP5 AMPG LVED	21.879	11	1.989	.484	.914
Explained	27762.070	38	730.581	177.896	.000
Residual	27802.961	6770	4.107		
Total	55565.031	6808	8.162		

levels. For example, for a three-way table with factors A, B and C with J, K and L levels, respectively, there are (J-1) (K-1) (L-1) linearly independent three-way interactions. Thus for the data in Table D2 there are (5-1) (4-1) (4-1) = 36 of them. In fact, a full analysis of variance for these data was not possible in this analysis because of excessive space requirements in the computer.

Two solutions to the analysis of Table D2 are presented. The first is to group the factor level of education into two levels, 1 = less than 6 years of education, 2 = Six or more years of education. This allows the full analysis of variance table to be calculated and it is given in Table 3.10. The second solution is to set the three-way interactions equal to zero and to calculate only the main effects and two-way interactions. This is achieved in SPSS by specifying option 4 in the OPTIONS card, and results in the sum of squares $SS_{A.B.C.}$ being added to (or "pooled" with) the error sum of squares, SS_{RES}. The results from this analysis appear in Table 3.11.

The sums of squares SS_{A+B+C} and $SS_{A.B+B.C+C.A}$ are again decomposed into components for the separate factors. The sum of squares for A.B, A.C and B.C are adjusted for main effects and other two-way interactions. The sum of squares for A, B and C are presented hierarchically,

$$SS_{A}, \quad SS_{B|A}, \quad SS_{C|A+B},$$

if OPTIONS = 10 is specified, and otherwise in the classical manner, adjusted for other main effects:

$$SS_{A|B+C}, \qquad SS_{B|A+C}, \qquad SS_{C|A+B}.$$

Table 3.11 displays the ANOVA for the data in Table D2 in hierarchical form.

The full ANOVA of Table 3.10 reveals a mean square of only 1.99 for the 3-way interactions between AGP5, AMGP and LVED, compared with a residual mean square of 4.1. Hence there is no evidence of significant 3-way interactions and there is some justification in omitting them for the analysis of the original data.

TABLE 3.11: Hierarchical Three-Way Analysis of Variance of Parity, by Age, by Age at Marriage, and by Level of Education in Four Groups, with 3-Way Interactions Pooled with the Residual

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	27468.961	10	2746.898	670.515	.000
AGP5	17267.563	4	4316.891	1053.750	.000
AMGP	10040.844	3	3346.948	816.987	.000
LVED	160.555	3	53.518	13.064	.000
2-Way Interactions	381.930	33	11.574	2.825	.000
AGP5 AMGP	190.685	12	15.890	3.879	.000
AGP5 LVED	87.205	12	7.267	1.774	.047
AMGP LVED	79.827	9	8.870	2.165	.022
Explained	27850.891	43	647.695	158.102	.000
Residual	27714.141	6765	4.097		
Total	55565.031	6808	8.162		8.162

Table 3.11 indicates a large interaction between AGP5 and AMGP ($F_{12,6765} = 3.88$) and smaller but significant interactions between AGP5 and LVED (F= 1.77) and between AMGP and LVED (F= 2.17). The Multiple Classification Analysis for the data in Table D2 is presented in Table 3.12. The adjusted effects for each factor in this table are based on fitting the additive model [AGP5 + AMGP + LVED] to the data, and are adjusted for both the other factors. Thus in particular the effects for LVED are adjusted for age and age at marriage, and can be compared with the adjusted effects from standardization given in Table 2.7. Also, a comparison of Table 3.12 and 3.4 indicates that the control of AGP5 and AMGP has a similar effect on the differentials by educational level as the control of MGP6.

An alternative adjustment for Age and Age at Marriage which incorporates the two-way interactions between Age and Age at Marriage is to form the joint variable (AGP5*AMGP) and calculate the analysis of variance of NCEB on LVED and (AGP5.AMGP). This is equivalent to fitting the additive model [AGP5.AMGP+LVED]. The analysis of variance appears in Table 3.13 and the resulting MCA table is given in Table 3.14. This analysis is theoretically preferable to Table 3.12 since the included interactions are significant. However, the adjusted effects of Educational Level are not noticeably altered.

3.6 Refinements

Interactions are often important and interesting in their own right – for example, cross-tabular analysis of the Fiji Fertility Survey indicated an important interaction between Race and Educational Level, namely that after adjusting for suitable demographic controls, differentials in fertility by educational level were evident for Fijians of Indian race but not for indigenous Fijians. An additive model between Education and Race would ignore this interaction and simply calculate an average adjusted effect of education for both races.

Nevertheless, sometimes interactions are an artifact of the way in which variables are measured. A change of variable or the scale in which a variable is measured may eliminate interactions and lead to a simpler interpretation of the data.

A common example of this occurs with dichotomous responses, taking values 0 and 1. Here the mean of Y in a cell corresponds to the proportion of cases with Y = 1. If the proportions lie near zero or one, then linear additivity can conflict with the requirement that proportions lie

TABLE 3.12: Multiple Classification Analysis Corresponding to ANOVA on Table 3.11

Grand Mean = 3.94

Variable + Category	N	Unadjusted Dev'n	Unadjusted Eta	Adjusted for Independents Dev'n	Adjusted for Independents Beta	Adjusted for Independents + Covariates Dev'n	Adjusted for Independents + Covariates Beta
AGP5							
1 15–24	1088	−2.55		−2.92			
2 25–29	1295	−1.41		−1.29			
3 30–34	1221	−.14		.03			
4 35–39	1203	.95		1.06			
5 40–49	2003	1.81		1.76			
			.56		.58		
AMGP							
1 LT15	984	1.81		1.45			
2 15–19	2991	.48		.63			
3 20–24	1932	−.81		−.61			
4 25 +	903	−1.83		−2.38			
			.38		.40		
LVED							
1 No School	1512	1.23		.25			
2 1–5 Years	2686	.30		.04			
3 6–9 Years	1704	−.69		−.13			
4 10 + Years	908	−1.64		.29			
			.32		.06		
Multiple R Squared					.494		
Multiple R					.703		

TABLE 3.13: Analysis of Variance with Age at Marriage and Age as a Joint Variable

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	27707.293	21	1319.395	321.445	.000
AGP5.AMGP	27552.414	18	1530.690	372.923	.000
LVED	154.879	3	51.626	12.578	.000
Explained	27707.293	21	1319.395	321.445	.000
Residual	27857.738	6787	4.105		
Total	55565.031	6808	8.162		

TABLE 3.14: Multiple Classification Analysis Corresponding to ANOVA in Table 3.13

Grand Mean = 3.94		Unadjusted		Adjusted for Independents		Adjusted for Independents + Covariates	
Variable + Category	N	Dev'n	Eta	Dev'n	Beta	Dev'n	Beta
AGP5AMGP							
1	114	−1.13		−1.19			
2	165	.48		.41			
3	175	1.56		1.47			
4	200	2.62		2.48			
5	330	3.15		3.01			
6	688	−2.47		−2.46			
7	473	−.62		−.63			
8	502	.64		.61			
9	455	1.91		1.85			
10	874	2.55		2.48			
11	286	−3.30		−3.20			
12	521	−2.23		−2.14			
13	306	−.72		−.65			
14	330	.30		.36			
15	489	1.35		1.34			
17	136	−3.29		−3.14			
18	239	−2.30		−2.14			
19	218	−1.61		−1.46			
20	310	−.99		−.93			
			.70		.68		
LVED							
1 No School	1512	1.23		.25			
2 1–5 Years	2686	.30		.03			
3 6–9 Years	1704	−.69		−.12			
4 10 + Years	908	−1.64		−.29			
			.32		.06		
Multiple R Squared					.499		
Multiple R					.706		

TABLE 3.15: A 2 x 2 Table of Proportions Additive on the Logit Scale

a) observed proportions

	Factor A 1	Factor A 2
Factor B 1	.119	.269
2	.018	.047

b) logits

	Factor A 1	Factor A 2
Factor B 1	−2.0	−1.0
2	−4.0	−3.0

between the limits zero and one. Table 3.15a) shows a simple example for a 2x2 table. If the bottom left hand proportion is deleted, then the impossible value of -.103 (=.047+.119-.269) is required for the means to be additive on the linear scale. In fact the means in this table are additive on the *logit scale*. That is, if the proportions are converted to logits by the transformation

$$\text{logit } p = \log \quad [p/(1-p)] ,$$

then the logits are additive, as in Table 3.15b). The logit transformation stretches the scale at zero and one, thus inflating small differentials near these limiting values.

Hence a simple approach to the analysis of cross-tabulated proportions lying near zero or one is transform the observed proportions to logits and carry out a standard analysis of variance of these transformed values. A modification is required for observed proportions of zero or one, for which the logit is not defined: One possibility is to replace the observed zero proportions by $(2n)^{-1}$, and unit proportions by $1-(2n)^{-1}$, where n is the cell sample count. A more sophisticated procedure based on log linear models for contingency tables is described in Little (1978).

Returning to the response Y = Parity of our illustrative examples, we have noted that additive models between level of education and the demographic controls are unlikely because of the cumulative nature of the response over the life cycle – For the case of Sri Lanka there is clear evidence that interactions do exist. We now give two alternatives fertility measures for which additive models appear a priori more plausible.

$$\text{i) } Y = \log(\text{parity})$$

For countries where differences in mean parity according to a background variable increase with mean parity, it may be plausible that *percentage* (or *proportional*) differences in mean parity are the same for all levels of marital duration. This is equivalent to differences in the *logarithm* of mean parity being constant for all levels of marital duration, or additivity on the log scale. For, if μ_{jk} is the mean parity for marital duration level j, educational level k, and the proportional differences

$$\mu_{jk}/\mu_{jk'} , \quad k \neq k' ,$$

are the same for all values of duration j, then the differences in the log means,

$$\log \mu_{jk} - \log \mu_{jk'}$$

are also the same for all values of duration j*.

Thus interactions may be reduced by taking log (Parity) as the response. If logarithms are taken at the individual level, then some modification is required for women with zero parity, for which logarithms cannot be taken. One possibility is simply to restrict the analysis to women with one or more births. Another is to add a constant before taking logs (Hermalin and Mason, 1979). A more sophisticated procedure based on log-linear models which avoids this problem is presented in Little (1978).

$$\text{ii) } Y = \text{Parity / Marital Duration (P/D).}$$

An alternative approach is to postulate that differences in mean parity between categories of a background variable are proportional to marital duration. Thus if the response is defined as Parity divided by Marital Duration (P/D)** an additive model is obtained. A detailed discussion of this measure is given in Little (1977).

---

* Recall the basic property of the logarithm.

$$\log(\mu_1/\mu_2) = \log \mu_1 - \log \mu_2 \text{ for all } \mu_1, \mu_2.$$

** In fact, the variable is defined as 120xParity/(Months Since First Marriage), and, as such, measures births per ten years of marital duration.

TABLE 3.16: Weighted Analysis of Variance of Parity Divided by Marital Duration, by Marital Duration and by Level of Education

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	1165.521	8	145.690	94.858	.000
MGP6	1159.160	5	231.832	150.944	.000
LVED	81.090	3	27.030	17.599	.000
2-Way Interactions	61.361	15	4.091	2.663	.000
MGP6 LVED	61.361	15	4.091	2.663	.000
Explained	1226.881	23	53.343	34.731	.000
Residual	10313.412	6715	1.536		
Total	11540.293	6738	1.713		

We shall apply the second of these approaches to the Sri Lanka data. A straight analysis of variance of the mean values of P/D cross-classified by MGP5 and LVED is not recommended, because it gives equal weight to observations with low or high marital durations. Intuitively we would expect the values of P/D with small values of D to be much less stable, since they are highly sensitive to the timing of the early births. In statistical terminology, the variance of P/D is not constant for all values of D, and hence one of the main statistical assumptions of Analysis of Variance is not satisfied.

The solution is to carry out a *weighted* analysis, with weights inversely proportional to the variance. Here we assume that the variance of P/D is inversely proportional to D, and hence the weights are proportional to D. Thus each individual in the sample is given a weight proportional to her marital duration. This choice of weighting is particularly appropriate for the chosen measure of fertility. It results in weighted means of P/D which are simply ratios of cumulated births divided by cumulated exposure.

That is,

$$\sum_{\text{Subclass}} D_i (P_i/D_i) \Big/ \sum_{\text{Subclass}} D_i = \left(\sum_{\text{Subclass}} P_i\right) \Big/ \left(\sum_{\text{Subclass}} D_i\right)$$

The weighted two-way analysis of variance is presented in Table 3.16. Unfortunately in this case the transformation has not eliminated the interaction between marital duration and educational level; it yields a highly significant F-statistic of 2.663. Inspection of the weighted cross-classification of P/D by MGP6 and LVED, displayed in Table D4, reveals the reason. For low durations, the tempo of fertility (as measured by the response) is greater than average for highly educated women. At high durations, the reverse is the case. Hence here it appears that the interaction between duration and education is an inherent characteristic of the data rather than an artifact of the choice of response.

One technical difficulty in the weighted analysis needs attention. The degrees of freedom for residual in the ANOVA table, and hence the F statistics, are incorrect unless the weights are scaled so that they sum to the true number of observations, in this case 6559. The scaling here is not quite correct: the degrees of freedom for residual should be 6535 rather than 6715, as shown in Table 3.16. This inaccuracy is not serious enough to change the substantive interpretation of the table, but the table should be corrected in practice. Some statistical packages automatically scale the weights so that the degrees of freedom are correct, but this was not the case in the version of SPSS used here.

4. ANALYSIS OF COVARIANCE

4.1 Introduction

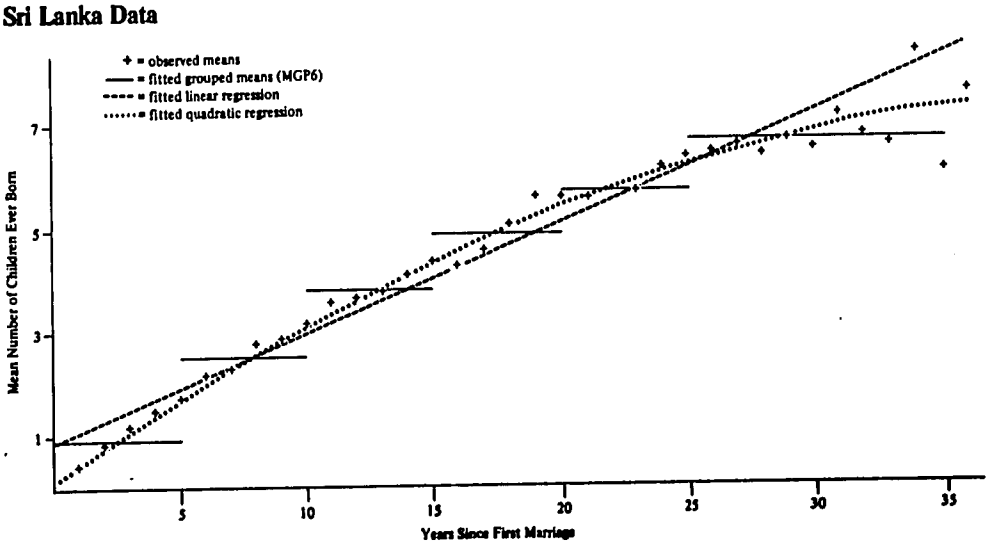
So far the independent variables in our examples have all been categorical in nature. The analyses could all be calculated using only the means, variances and sample counts within the cells of a cross-tabulation. From now on we consid methods which require individual level data. It should be stressed that this does not represent a change in the emphasis of the analysis from the "macro" to the "micro" level, since the statistics produced will still be averages or summaries based on the individual level data which lead to statements at the aggregate level.

A common characteristic of the techniques considered so far is that they require the *grouping* of interval scaled regressors, such as age and age at marriage, into a small number of categories. Also the treatment of the categorical variables does not take into account any ordering between the categories. The effects of a grouped variable are estimated in a way which implies that the mean response is constant within the ranges of each grouping, and jumps suddenly between groups. For example consider the relationship between parity and marital duration for the Sri Lanka data. The mean parities by single years of marital duration are plotted in Figure 4.1. The step function represents the relationship implied by replacing marital duration by the six categories MGP6. As a model this is clearly not ideal, since it is discontinuous and does not reflect the positive relationship between duration and parity within duration groups. A finer grouping would model the relationship more accurately, but this increases the number of parameters required to represent the effect (c-1 parameters for a variable with c groups). We have already encountered difficulties in estimating the number of parameters required for the three-way analysis of variance of Section 3.3; increasing the number of categories of marital duration makes problems like this even worse, and conflicts with the need for a parsimonious representation of effects.

A natural alternative treatment of the means in Figure 4.1 is to fit a smooth curve to the data. Two alternatives are shown. The first assumes a linear relationship. In symbols, the population mean parity μ at duration YSFM (Years Since First Marriage) is assumed to take the form

$$\mu(\text{YSFM}) = \alpha_0 + \alpha_1 \text{ YSFM} \tag{4.1}$$

FIGURE 4.1: Mean Number of Children Ever Born as a Function of Years Since First Marriage. Sri Lanka Data



Where α_1 is the slope, that is the increase in parity per year of marital duration, and α_0 is the intercept (which might be taken as zero). This equation does not model a decline in the slope as marital duration increases, which is expected on substantive grounds and also apparent in the observed means. Hence the second curve models a linear decline by including a quadratic term in marital duration. That is,

$$\mu(\text{YSFM}) = \alpha_0 + \alpha_1 \text{ YSFM} + \alpha_2 \text{ YSFM}^2. \tag{4.2}$$

Here the effects of duration are represented by two parameters, α_1, the slope (or more precisely the tangent to the curve) at YSFM = 0 and α_2, the rate of change of the slope with marital duration. The parameters in equations (4.1) and (4.2) may be estimated by least squares, using a linear regression program; further details are deferred until Chapter 5.

Inspection of Figure 4.1 indicates that the effects of duration are more closely modelled by the one or two parameters in the regression models (4.1) and (4.2) than by the five parameters of the grouped model.

As in analysis of variance, the effect of a variable in a regression has an associated sum of squares, which measures the variation in the response attributable to that variable. The relative effectiveness of the grouped model and the regression models in capturing the effect of marital duration on parity can be compared via the sum of squares for the effects of duration obtained from each fit. These are as follows:

Model	Marital Duration	Degrees of Freedom	Sum of Squares
Grouped	MGP6	5	27,403
(4.1)	YSFM	1	27,696
(4.2)	YSFM and YSFM^2	2	28,498

The sum of squares for the grouped model is taken from Table 3.8 for the regression models they are taken from Tables 4.1 and 4.3. (Despite the appearance of LVED in these tables, the effects of duration are not adjusted by education.) The superiority of the linear regression representation (4.1) over the grouped factor MGP6 is reflected by the greater sum of squares explained (27,696 as opposed to 27,403), achieved with four less degrees of freedom. The addition of the quadratic term in Model (4.2) further improves the fit by a significant amount.

In conclusion, polynomial regression models often provide a parsimonious method for summarizing the effects of interval scaled regressors.

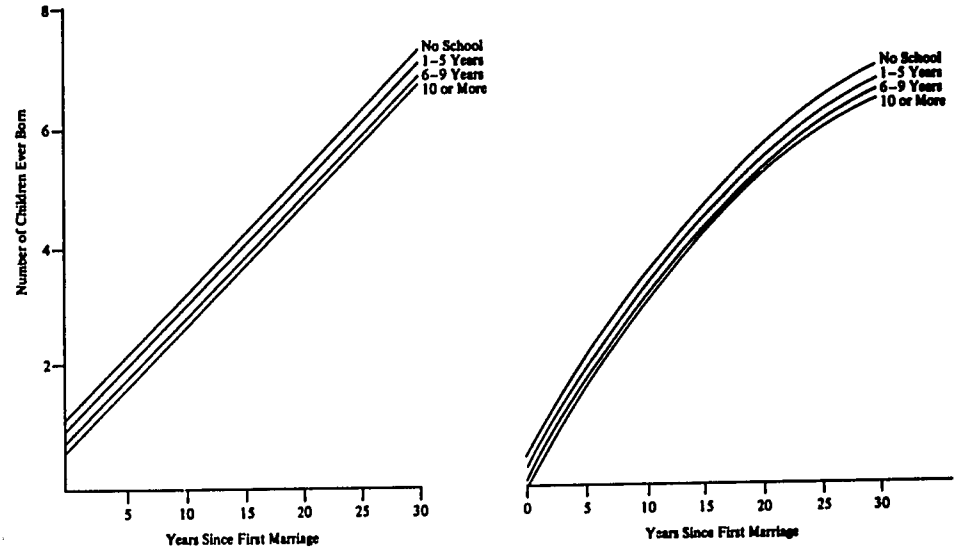
4.2 Analysis of Covariance

We now return to the problem of estimating the effects of education, adjusted for marital duration. The adjusted means of education from multiple classification analysis are imperfect in that they do not take into account differences in the distribution of marital duration between education groups within each level of MGP6. In other words, they also may be distorted by the representation of marital duration as a grouped variable. What is required is a method of adjustment which adopts the superior polynomial representation of marital duration described in the previous section. The technique which achieves this is *analysis of covariance.*

In the model of equation (4.1) a straight line was fitted to the plot of mean parity by marital duration for the whole sample. In analysis of covariance a separate straight line is fitted to the data for each education group. The results are plotted in Figure 4.2.a. The effects of education adjusted for marital duration are represented by the vertical displacements between the lines.

As with multiple classification analysis, the effects of education and marital duration are assumed to be additive in this method. Thus the fitted curves in Figure 4.2.a are constrained so that the vertical displacements between the lines are constant for all values of marital duration;

FIGURE 4.2: Fitted Values from Analysis of Covariance of Mean Parity on Years Since First Marriage and Level of Education



a) Linear Fit to Marital Duration

b) Quadratic Fit to Marital Duration

in other words, the lines are parallel. The effect of marital duration adjusted for education is the (common) slope of the regression lines. The fitted lines are obtained by finding values of the common slope and the intercepts for each education group so that the sum of squared deviations between the observed and fitted values is minimized.

The alternative analysis of covariance diagrammed in Figure 4.2.b assumes a quadratic relationship between parity and marital duration for each education group. Note that again the curves are parallel, reflecting the additivity assumption. This is achieved by constraining the linear and quadratic terms of the regression to be equal for each education group. The adjusted effects of education are again represented by the vertical displacements between the curves. Both the figures can be compared with the MCA representation in Figure 2.1.b.

We can also write down equations for the fitted means in each model. The MCA of the previous chapter was based on fitting a model where the mean parity μ_{jk} for marital duration group j, and level of education k takes the form

$$\mu_{jk} = \mu + \alpha_j + \beta_k . \tag{4.3}$$

If the effect of marital duration is modelled by linear regression, as in Figure 4.2.a, we obtain the analysis of covariance model which expresses the mean parity for marital duration YSFM and education level k in the form

$$\mu_k \text{ (YSFM)} = \mu + \alpha_1 \text{ YSFM} + \beta_k . \tag{4.4}$$

If the effect of marital duration is modelled by a quadratic regression, as in Figure 4.2.b, we obtain the more detailed model

$$\mu_k \text{ (YSFM)} = \mu + \alpha_1 \text{ YSFM} + \alpha_2 \text{ YSFM}^2 + \beta_k . \tag{4.5}$$

Note the similarity between (4.3), (4.4) and (4.5). All three equations express the fact that the effects are additive, that is, the effects of education are the same for all levels of marital duration. For example, under model (4.5) the difference in mean parity between educational levels k and k' for women with marital duration YSFM is

$$\mu_k(\text{YSFM})-\mu_{k'}(\text{YSFM}) = (\mu + \alpha_1 \text{ YSFM } + \alpha_2 \text{ YSFM}^2 + \beta_k) - (\mu + \alpha_1 \text{ YSFM} + \alpha_2 \text{YSFM}^2 + \beta_{k'})$$

$$= \beta_k - \beta_{k'},$$

and this is the same for all values of YSFM. The other models give the same result. Thus the adjusted effects of education are given by the parameters β_k. If the parameter β_1 is set equal to zero, then β_k represents the difference between educational level k and educational level 1, adjusted for the effect of marital duration. The only difference between the methods is the way in which marital duration is controlled.

Analysis of Covariance can be carried out using the SPSS Analysis of Variance program by specifying covariates on the ANOVA card. The results of fitting the model (4.4) with factor LVED and covariate YSFM are presented in Tables 4.1 and 4.2. The results for model (4.5) with covariates YSFM and YSFM^2 are given in Tables 4.3 and 4.4. Various options are available for the presentation of the ANOVA and MCA tables according to whether covariates are adjusted before or after the main effects of factors, and whether the sums of squares for the effects are presented in a classical or a hierarchical form.

Here the covariates are adjusted first, the default in SPSS. Thus in Tables 4.1 and 4.3 the sums of squares for the covariates are presented first and are not adjusted for the factor LVED. The sum of squares for LVED is presented next and is adjusted for the covariates. This sum of squares is slightly smaller (174.5) when the quadratic term YSFM^2 is included than otherwise (193.0). The adjusted education effects are presented in the MCA tables, Table 4.2 and Table 4.4. Finally Table 4.5 summarizes the adjusted and unadjusted effects of education on parity obtained by the various methods considered in this and the previous chapters.

The examples of Analysis of Covariance considered so far have been restricted to a single factor (LVED) and have not included interaction terms. Within the context of our Analysis of Variance program, interactions between covariates and interactions between factors can be included, but interactions between covariates and factors are not allowed. As an illustration, Tables 4.6 and 4.7 presents the results of a weighted analysis of covariance with response parity divided by marital duration (P/D), weights proportional to duration, factors respondents educational level (LVED) and husband's educational level (HEDL), with the same categories as LVED, and covariates representing the effects of marital duration and age at first marriage. The covariates consist of linear and quadratic terms in duration and age at first marriage (YSFM, YSFMSQ, AGFM, AGFMSQ), and the interaction represented by multiplying the individual values of years since first marriage (YSFM) and age at first marriage (AGFM), that is,

YSFMAGFM = YSFM x AGFM.

OPTION 10 was specified on the ANOVA card, which implies that i) Covariates are adjusted first, ii) factors are adjusted after the covariates, and iii) the sums of squares for covariates and the main effects of factors are presented in a hierarchical form. Thus, for example, the sum of squares for LVED in Table 4.6 (22.667) is adjusted for the covariates, and the sum of squares for HEDL (22.231) is adjusted for the covariates *and* LVED. The latter yields a highly significant F of 5.03, suggesting that husband's education has an effect on fertility after adjusting for respondent's education and the demographic variables marital duration and age at first marriage.

Two sets of interactions are presented in Table 4.6. The interactions between age at first marriage and marital duration are represented by the single product term YSFMAGFM, and have a

TABLE 4.1: Analysis of Variance of Parity on Level of Education with Years Since First Marriage as a Covariate

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Covariates	27696.445	1	27696.445	6809.126	.000
YSFM	27696.445	1	27696.445	6809.126	.000
Main Effects	192.992	3	64.331	15.816	.000
LVED	192.995	3	64.332	15.816	.000
Explained	27889.438	4	6972.359	1714.143	.000
Residual	27675.594	6804	4.068		
Total	55565.031	6808	8.162		

Covariate	Raw Regression Coefficient
YSFM	.214

TABLE 4.2: Multiple Classification Analysis Corresponding to Table 4.1

Grand Mean = 3.94

Variable + Category	N	Unadjusted Dev'n	Unadjusted Eta	Adjusted for Independents Dev'n	Adjusted for Independents Beta	Adjusted for Independents + Covariates Dev'n	Adjusted for Independents + Covariates Beta
LVED							
1 No School	1512	1.23				.23	
2 1–5 Years	2686	.30				.07	
3 6–9 Years	1704	–.69				–.14	
4 10 + Years	908	–1.64				–.32	
			.32				.06
Multiple R Squared							.502
Multiple R							.708

TABLE 4.3: Analysis of Variance of Parity by Level of Education with Linear and Quadratic Terms of Marital Duration as Covariates

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Covariates	28497.988	2	14248.994	3604.571	.000
YSFM	5289.371	1	5289.371	1338.053	.000
YSFMSO	801.545	1	801.545	202.767	.000
Main Effects	174.547	3	58.182	14.718	.000
LVED	174.547	3	58.182	14.718	.000
Explained	28672.535	5	5734.507	1450.659	.000
Residual	26892.496	6803	3.953		
Total	55565.031	6808	8.162		

Covariate	Raw Regression Coefficient
YSFM	.343
YSFMSQ	–.004

TABLE 4.4: Multiple Classification Analysis Corresponding to Table 4.3

Grand Mean = 3.94		Unadjusted		Adjusted for Independents		Adjusted for Independents + Covariates	
Variable + Category	N	Dev'n	Eta	Dev'n	Beta	Dev'n	Beta
LVED							
1 No School	1512	1.23				.25	
2 1–5 Years	2686	.30				.04	
3 6–9 Years	1704	–.69				–.15	
4 10 + Years	908	–1.64				–.27	
			.32				.06
Multiple R Squared							.516
Multiple R							.718

TABLE 4.5: Summary of Effects of Education, Expressed as Deviations from Overall Mean

		Effects of Education						
Control	Method of Adjustment	No Schooling	1–5 Years	6–9 Years	10 + Years	Mean	Sum of Squares of Effect	Source Tables
a. Unadjusted	–	1.23	.30	–.69	–1.64	3.94	5747.4	D1, 3.6
b. MGP6	Test Factor Standardization	.26	.10	–.13	–.45	3.88	–	D1, 3.1
MGP6	ANOVA, MCA	.29	.05	–.16	–.31	3.94	225.5	3.2, 3.4
YSFM	ANCOVA, MCA	.23	.07	–.14	–.32	3.94	193.0	4.1, 4.2
YSFM, $YSFM^2$	ANCOVA, MCA	.25	.04	–.15	–.27	3.94	174.5	4.3, 4.4
c. AGP5*AMGP	Test Factor Standardization	.29	.16	–.06	–.74	3.84	–	D2
AGP5, AMGP	ANOVA, MCA	.25	.04	–.13	–.29	3.94	160.6	3.11, 3.12
AGP5*AMGP	ANOVA, MCA	.25	.03	–.12	–.29	3.94	154.9	3.13, 3.14

highly significant F value, 29.07. Indeed, note that all the covariates contribute significantly to the fit. The interactions between respondent's education and husband's education are represented by the two-way interactions for LVED and HEDL, with 9 degrees of freedom. These yield a non-significant F value of 1.12, indicating no evidence of a significant effect.

This is a fairly elaborate model, but it fails to encorporate one important feature of the data, namely the interactions between education and the demographic controls, as evidenced in Table 3.16. Models which allow interactions between covariates and factors are treated in the next chapter.

TABLE 4.6: Weighted Analysis of Variance of Parity Divided by Duration by Respondent's Level of Education (LVED) and Husband's Level of Education (HEDL), with Linear and Quadratic Terms in Years Since First Marriage and Age at First Marriage and the Product of Years Since First Marriage and Age at First Marriage as Covariates

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Covariates	1585.090	5	317.018	215.281	.000
YSFM	1098.046	1	1098.046	745.663	.000
YSFMSQ	28.979	1	28,979	19.679	.000
AGFM	356.166	1	356.166	241.865	.000
AGFMSO	59.088	1	59.088	40.126	.000
YSFMAGFM	59.088	1	42.812	29.073	.000
Main Effects	44.909	6	7.485	5.083	.000
LVED	22.677	3	7.559	5.133	.002
HEDL	22.231	3	7.410	5.032	.002
2-Way Interactions	14.849	9	1.650	1.120	.344
LVED HEDL	14.849	9	1.650	1.120	.344
Explained	1644.848	20	82.242	55.849	.000
Residual	9953.146	6759	1.473		
Total	11597.994	6779	1.711		

Covariate	Raw Regression Coefficient
YSFM	–.049
YSFMSQ	.001
AGFM	–.059
AGSMSQ	–.003
YSFMAGFM	–.004

TABLE 4.7: Multiple Classification Analysis Corresponding to Table 4.6

Grand Mean = 2.69

Variable + Category	N	Unadjusted Dev'n	Unadjusted Eta	Adjusted for Independents Dev'n	Adjusted for Independents Beta	Adjusted for Independents + Covariates Dev'n	Adjusted for Independents + Covariates Beta
LVED							
No School	2003	–.03				.07	
2 1–5 Years	2873	–.00				–.01	
3 6–9 Years	1394	.02				–.07	
4 10 + Years	511	.09				.02	
			.02				.04
HEDL							
1 No Schooling	681	–.02				.04	
2 1–5 Years	3110	.00				.01	
3 6–9 Years	2214	.03				.04	
4 10 + Years	774	–.08				–.18	
			.03				.05
Multiple R Squared							.141
Multiple R							.375

5. MULTIPLE LINEAR REGRESSION

5.1 Introduction

In chapter 3 analysis of variance and multiple classification analysis were introduced as methods for analysing the effects of categorical regressors (factors) on a response. In Section 4.1 multiple regression was introduced as a method for calculating the effects of an interval scaled regressor, years since first marriage, on a response. Then in subsequent sections an extension of analysis of variance to include factors and covariates, analysis of covariance, was discussed.

This terminology, which developed for historical reasons, is not entirely appropriate. Analysis of variance is a general method which can be applied to problems with interval-scaled regressors with or without factors, as well as to problems involving factors only.

Indeed, ANOVA tables for Analysis of Covariance have been presented in Section 4.2. Furthermore, all the models fitted can be viewed as special cases of multiple linear regression. Thus in this chapter analyses of the previous chapters are replicated using a multiple regression program.

The perspective adopted in this chapter is more general and flexible than that of previous chapters, and includes models involving interactions between covariates and factors which lie outside the scope of Chapters 3 and 4. The basic method of analysis is multiple linear regression, which can be used to fit models involving the main effects of categorical and/or interval-scaled regressors, and specified interactions between them. Analysis of variance and multiple classification analysis are viewed as optional outputs which can be calculated from the basic output of the regression program. Analysis of variance decomposes the variance in the response into components for each effect; the ANOVA table is derived from the regression sums of squares from a sequence of regressions. Multiple classification analysis presents the effects of a categorical regressor from a regression which is additive between that factor and the other factors and covariates.

We begin by presenting the elements of multiple regression; this is sketched rather briefly and assumes some prior knowledge on the part of the reader. We then discuss how categorical regressors are treated within the context of the method, by the creation of dummy variables. Finally, we discuss how to encorporate interactions in the regression.

5.2 Elements of Multiple Linear Regression

The data for a regression analysis consist of values for each individual i of a response variable Y_i and a set of k regressors, $X_{i1}, \ldots, X_{ik}$. Multiple linear regression calculates a fitted value $\hat{Y}_i$ for each individual i which is a linear combination of the regressor values for that individual, that is takes the form

$$\hat{Y}_i = b_0 + \sum_{j=1}^{k} b_j X_{ij}, \tag{5.1}$$

and is as close as possible to the observed value, Y_i. Specifically, values of the *intercept* b_0 and the *slopes* $b_1, \ldots, b_k$ are chosen so that the fitted values minimize the sum of squared deviations

$$SS = \sum_i (Y_i - \hat{Y}_i)^2$$

or more generally, the weighted sum

$$SS_w = \sum_i w_i (Y_i - \hat{Y}_i)^2 \tag{5.2}$$

for some chosen set of weights, w_i. Associated with this calculation is a decomposition of the

response Y_i into the *fitted value*, $\hat{Y}_i$, and the *residual*, $Y_i - \hat{Y}_i$. If Y_i is measured as a deviation from the sample mean $\bar{Y}$, we have

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Squaring and summing over the (weighted) observations, the cross-product terms $(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$ sum to zero and we obtain the analysis of variance decomposition

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2 \tag{5.3}$$

$$SS_T = SS_{REG} + SS_{RES}$$

That is, the total corrected sum of squares, SS_T, with N - 1 degrees of freedom, decomposes into the sum of squares for the regression, SS_{REG}, with k degrees of freedom, and the residual sum of squares, SS_{RES}, with N-k-1 degrees of freedom. The basic output of regression is the set of regression co fficients, $b_0, b_1, \ldots, b_k$, and the ANOVA table based on the decomposition given in equation (5.3).

Note that according to equation (5.1), if a regressor X_{ij} is increased by one unit with the other regressors held fixed, the fitted value $\hat{Y}_i$ is increased by b_j. Thus b_j is interpreted as the effect on $\hat{Y}$ of increasing X_j by one unit with the other regressors controlled. Sometimes regression coefficients are calculated after standardizing the response and regressors to unit variance by dividing by their standard deviations. The resulting coefficients b_j' are called *standardized*, and are related to the coefficients b_j by the forumla

$$b_j' = b_j \, sd(X_j) / sd(Y),$$

where sd stands for standard deviation. The standardized coefficient b_j estimates the increase in $\hat{Y}$, measured in standard deviations of Y, obtained by increasing X_j by one standard deviation with the other regressors controlled. Standardized coefficients are labelled BETA in SPSS output.

Other statistics commonly presented are multiple R^2, defined as the proportion of the variance explained by the regression, SS_{REG}/SS_T, and the multiple R, the square root of this measure.

Also, of interest are the F statistic for the regression sum of squares, that is, the ratio of the regression mean square to the residual mean square, and standard errors for the regression coefficients. Statistical tests and confidence intervals based on these quantities are available under the following statistical model. The response values Y_i are assumed independently normally distributed with a mean linear in the X's, that is

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik}, \tag{5.4}$$

and variance inversely proportional to the weight w_i. If this model is true, then the overall statistical significance of the regression coefficients can be tested by comparing the F-statistic for the regression sum of squares with the tabulated F distribution with k and N-k-1 degrees of freedom. Also the significance of individual coefficients can be tested by comparing the individual F values with the tabulated F distribution with 1 and N-k-1 degrees of freedom. Alternatively, 95% confidence intervals may be obtained by subtracting plus or minus two standard errors from the estimated coefficients.

Table 5.1 gives the SPSS output for the regression of NCEB on YSFM and $YSFM^2$, discussed in Section 4.1 and plotted in Figure 4.1. The important quantities in this output can be identified as follows:

TABLE 5.1: Quadratic Regression of Farity on Years Since First Marriage

VARIABLE(S) ENTERED ON STEP NUMBER 2 . . YSFMSQ

MULTIPLE R	.71630
R SQUARE	.51309
ADJUSTED R SQUARE	.51294
STANDARD ERROR	1.99389

ANALYSIS OF VARIANCE	DF	SUM OF SQUARES	MEAN SQUARE	F
REGRESSION	2.	28516.43333	14258.21666	3586.42834
RESIDUAL	6807.	27061.93227	3.97560	

. . . . . VARIABLES IN THE EQUATION . . . . . .

VARIABLE	B	BETA	STD ERROR B	F
YSFM	.3408267	1.12240	.00939	1317.513
YSFMSQ	–.3940649E-02	–.43242	.00028	195.559
(CONSTANT)	.1311261			

. . . . . VARIABLES NOT IN THE EQUATION . . . . . .

VARIABLE	BETA IN	PARTIAL	TOLERANCE	F

TABLE 5.2: Regression of Parity on Level of Education Represented as a Set of Dummies

VARIABLE LIST 1
REGRESSION LIST 1

REGRESSIONS ON NCEB

DEPENDENT VARIABLE . . NCEB
VARIABLE(S) ENTERED ON STEP NUMBER 1 . . PRIM
RSEC
HIGH

MULTIPLE R	.32154
R SQUARE	.10339
ADJUSTED R SQUARE	.10299
STANDARD ERROR	2.70589

ANALYSIS OF VARIANCE	DF	SUM OF SQUARES	MEAN SQUARE	F
REGRESSION	3.	5746.02924	1915.34308	261.59370
RESIDUAL	6806.	49832.33636	7.32182	

. . . . . VARIABLES IN THE EQUATION . . . . . .

VARIABLE	B	BETA	STD ERROR B	F
PRIM	–.9254233	–.15832	.08700	113.151
RSEC	–1.909992	–.28959	.09560	399.150
HIGH	–2.863278	–.34068	.11361	635.143
(CONSTANT)	5.167174			

. . . . . VARIABLES NOT IN THE EQUATION . . . . . .

VARIABLE	BETA IN	PARTIAL	TOLERANCE	F
YSFM	.68196	.66669	.85690	5444.731
MG09	–.20829	–.21796	.98180	339.397
MG14	–.02162	–.02281	.99734	3.541
MG19	.12150	.12809	.99659	113.515
MG24	.21234	.22231	.98224	353.794
M25P	.38773	.39486	.92989	1256.992

$N = 6810$, $k = 2$.

Coefficients:	$b_0 = .1311, b_1 = .3408, b_2 = -.00394$
Standard Errors:	$se(b_1) = .00939, se(b_2) = .00028$
Standard Coefficients:	$b_1' = 1.122, b_2' = -.4324$
ANOVA:	$SS_{REG} = 28516.4, SS_{RES} = 27061.9, R^2 = .5131, R = .7163$

F - value for regression sum of squares = 3586.4

5.3 Treatment of Factors in Regression

5.3.1 A Single Factor

Interval scaled covariates, such as YSFM and AGFM, are introduced directly into a regression without recording. On the other hand, the treatment of categorical variables, and the interpretation of the regression coefficients obtained for them, requires more care.

We first consider a simple regression on a dichotomous variable, that is, a variable with two categories. For example, suppose that the response Y_i is the parity for individual i and X_i is a variable taking the value one for women with formal education and zero otherwise. The fitting equation (5.1), specifies that predicted mean parity $\hat{Y}_i$ takes the form

$$\hat{Y}_i = b_0 + b_1 X_i, \tag{5.5}$$

where b_1 is the slope and b_0 is the intercept. Hence the predicted mean parities for women with no education and for women with formal education are obtained by substituting $X_i = 0$ and $X_i = 1$ respectively in equation (5.5):

$$(\hat{Y}_i | X_i = 0) = b_0 ; (\hat{Y}_i | X_i = 1) = b_0 + b_1. \tag{5.6}$$

Hence b_0 is the predicted mean for individuals with $X_i = 0$ and b_1 is the difference in predicted means between individuals with $X_i = 1$ (that is, educated women) and individuals with $X_i = 0$ (that is, uneducated women). Note that the original interpretation of regression coefficients still remains. The slope b_1 represents the increase in the fitted mean obtained by changing $X = 0$ to $X = 1$, which is equivalent to switching from the uneducated to the educated group.

It comes as no surprise that in practice the fitted values (5.6) calculated by regression to minimize (5.2) are simply the weighted sample mean parities for the two groups. That is,

$$b_0 = \bar{Y}_0, b_0 + b_1 = \bar{Y}_1,$$

where $\bar{Y}_j$ is the (weighted) mean parity for women with $X_i = j$, $(j = 0,1)$. Hence in a sense the regression is equivalent to a simple cross-tabulation of the mean parities for the two groups.

Now consider a factor with $k>2$ groups. For example, let us consider the factor LVED with $k = 4$ groups, NO EDUCATION, 1-5 YEARS, 6-9 YEARS and 10 OR MORE YEARS. Suppose that these levels are coded 1 to 4 and the variable introduced into the regression as a covariate. Then the regression model will predict means for the four groups which are equally spaced. That is, if the intercept and slope are b_0 and b_1 respectively, the predicted means for the four groups are $b_0 + b_1$, $b_0 + 2b_1$, $b_0 + 3b_1$ and $b_0 + 4b_1$, and thus adjacent groups all differ by the quantity b_1. This procedure effectively assumes an ordering between the categories, which is justified for this variable but does not make sense for unordered factors such as, say, Religion. The imposition of equal spacing between the category means is often less desirable, and implies that the regression is not analogous to the cross-tabulation of mean parities, as was the case for a binary factor. We now give an alternative treatment of factors which does correspond to cross-tabulation in the simple case when a single factor is included in the regression.

TABLE 5.3: Regression of Parity on Marital Duration and Level of Education, Represented as Sets of Dummy Variables

REGRESSIONS ON NCEB

VARIABLE LIST 1
REGRESSION LIST 1

DEPENDENT VARIABLE.. NCEB
VARIABLE(S) ENTERED ON STEP NUMBER 2.. M25P
MG09
MG14
MG19
MG24

MULTIPLE R	.70505	ANALYSIS OF VARIANCE	DF	SUM OF SQUARES	MEAN SQUARE	F
R SQUARE	.49710	REGRESSION	8.	27627.74499	3453.46812	840.30466
ADJUSTED R SQUARE	.49650	RESIDUAL	6801.	27950.62061	4.10978	
STANDARD ERROR	2.02726					

. . . . . VARIABLES IN THE EQUATION . . . . . .

VARIABLE	B	BETA	STD ERROR B	F
PRIM	–.2380297	–.04072	.06604	12.991
RSEC	–.4495334	–.06816	.07456	36.350
HIGH	–.5951281	–.07081	.09067	43.078
M25P	5.511676	.74246	.08596	4111.369
MG09	1.465360	.19737	.08122	325.517
MG14	2.742686	.35560	.08410	1063.598
MG19	3.796765	.48123	.08602	1948.337
MG24	4.695667	.55494	.09138	2640.312
(CONSTANT)	1.311514			

. . . . . VARIABLES NOT IN THE EQUATION. . . . . .

VARIABLE	BETA IN	PARTIAL	TOLERANCE	F
YSFM	.64841	.17774	.03779	221.833

For a k-category variable, one category is selected and called the *reference* category. For each of the (k-1) other categories, a *dummy* or *indicator* variable is defined, taking value one for individuals falling in that category and zero otherwise. Here we choose NO SCHOOLING as the reference category, and define k-1 = 3 variables

$$\text{PRIM} = \begin{cases} 1, & \text{1-5 Years Education ;} \\ 0, & \text{Otherwise} \end{cases}$$

$$\text{RSEC} = \begin{cases} 1 & \text{,6-9 Years of Education} \\ 0 & \text{,Otherwise} \end{cases} ;$$

$$\text{HIGH} = \begin{cases} 1 & \text{,10 or More Years of Education} \\ 0 & \text{,Otherwise} \end{cases}$$

The factor is represented in the regression by the set of dummy variables defined thus, in this case PRIM, RSEC and HIGH.

To see the effect of this, note that the fitted values from this regression are

$$\hat{Y}_i = b_0 + b_1 \text{PRIM}_i + b_2\text{RSEC}_i + b_3\text{HIGH}_i ,$$

where PRIM_i, RSEC_i and HIGH_i are the values of PRIM, RSEC and HIGH for respondent i. For individuals with no education, $\text{PRIM}_i = \text{RSEC}_i = \text{HIGH}_i = 0$. Hence the predicted mean is

$$(\hat{Y}_i \,|\text{LVED} = 1) = b_0,$$

the intercept of the regression. For individuals with 1-5 years education, PRIM = 1 and RSEC = HIGH = 0. Hence the predicted mean is

$$(\hat{Y}_i \,|\text{LVED} = 2) = b_0 + b_1,$$

Similarly for the other categories of education we obtain predicted means

$$(\hat{Y}_i \,|\text{LVED} = 3) = b_0 + b_2, \quad (Y_i \,|\text{LVED} = 4) = b_0 + b_3.$$

Hence *the intercept b_0 is the fitted mean for the reference category,* and *the slope b_j is the difference in the fitted mean between category j+1 and the reference category.* These properties are of central importance in the interpretation of regressions with factors.

Once again, the fitted mean obtained from the regression are simply the (weighted) sample means within each category of the factor, and hence regression is here a rather unwieldy way of obtaining the cross-classification of means.

Table 5.2 gives the results of the regression of NCEB on the three dummy variables PRIM, RSEC and HIGH. The cross-tabulation of mean parity by Level of Education, given in Table 2.1.a), is reconstructed from the regression coefficients as follows:

$$\begin{aligned} b_0 &= 5.167, \\ b_0 + b_1 &= 5.167 - 0.925 = 4.242, \\ b_0 + b_2 &= 5.167 - 1.910 = 3.257, \\ b_0 + b_3 &= 5.167 - 2.863 = 2.304. \end{aligned}$$

Note that the ANOVA table corresponds to the analysis of variance of NCEB on LVED, given in Table 3.3. The two tables differ only because of rounding error. This correspondence is inevitable, because both analyses are based on the same fitted model.

5.3.2 Two or More Factors

Now suppose we add another factor to the regression. Following the procedure in Chapter 3, we include the factor marital duration with six levels, MGP6. As with LVED, this is represented in the regression by a set of dummy variables. A reference category, 0-4 years, is chosen, and dummy variables are defined for the other five year marriage groups, MG09, MG14, MG19, MG24 and M25P. If these are added to the regression equation, we obtain the output given in Table 5.3. The regression model here is

$$\hat{Y} = b_0 + b_1 \text{ PRIM} + b_2 \text{ RSEC} + b_3 \text{ HIGH} + b_4 \text{ MG09} + b_5 \text{ MG14} + b_6 \text{ MG19} + b_7 \text{ MG24} + b_8 \text{ M25P}. \tag{5.7}$$

Substituting the coefficients in Table 5.3 we obtain the equation

$$\hat{Y} = 1.31 - .24 \text{ PRIM} - .45 \text{ RSEC} - .60 \text{ HIGH} + 1.47 \text{ MG09} + 2.74 \text{ MG14} + 3.80 \text{ MG19} + 4.70 \text{ MG24} + 5.51 \text{ M25P} \tag{5.8}$$

From this expression we can derive fitted means for each level of LVED and MGP6. Let $\hat{Y}_{jk}$ denote the fitted mean for respondents with LVED = j and MGP6 = k. The reference categories correspond to no education, LVED = 1, and 0-4 years since first marriage, MGP6 = 1. Hence substituting zero for all the variables in (5.8), we obtain the fitted value for women with no education married less than five years:

$$Y_{11} = 1.31 .$$

To obtain the fitted value for women with 6-9 years of education married 15-19 years, we set RSEC = MG19 = 1 and the other variables equal to zero, obtaining

$$Y_{34} = 1.31 - .45 + 3.80 = 4.66$$

Other fitted values are derived in a similar manner. The fitted values obtained are identical to those obtained from the multiple classification analysis of NCEB with factors LVED and MGP6, given in Table 2.2. This can be verified by comparing the values of Y_{11} and Y_{34} with the corresponding values in that table. The correspondence arises because both models fit the same additive model for the two factors. To demonstrate it rigorously, note that the regression equation (5.7) can be re-written in the same form as the MCA model,

$$\hat{Y}_{jk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_k \ , \quad 1 \leq j \leq 4, \ 1 \leq k \leq 6 \ ,$$

by setting

$$\hat{\mu} = b_0, \ \hat{\alpha}_1 = 0, \ \hat{\alpha}_2 = b_1, \ \hat{\alpha}_3 = b_2, \ \hat{\alpha}_4 = b_2$$

$$\hat{\beta}_1 = 0, \hat{\beta}_2 = b_4, \hat{\beta}_3 = b_5, \hat{\beta}_4 = b_6, \hat{\beta}_5 = b_7, \hat{\beta}_6 = b_8.$$

The coefficients in (5.7) have the interpretation as deviations between the dummy variable category and the reference category, adjusted for the other factor. For example, b_1 = -.24 estimates the difference in mean parity between respondents with 1-5 years of schooling and respondents with no schooling, adjusted for marital duration. Thus the adjusted effects of education, expressed as deviations from the NO SCHOOLING group, are simply b_1 = -.24, b_2 = -.45 and b_3 = -.60. (Cf Table 2.4).

Suppose we wish to express these effects as deviations from the overall mean, μ = 3.84, as in the

MCA table. If d_1 is the deviation for the NO SCHOOLING group, then the deviations for the other groups are $d_1 + b_1$, $d_1 + b_2$ and $d_1 + b_3$ respectively. To calculate d_1, we exploit the fact that the average over the distribution of the factor in the sample of the category deviations is zero. That is, if p_j is the (weighted) proportion of the sample in category j, then

$$0 = \sum_{j=1}^{4} p_j d_j = d_1 + p_2 b_1 + p_3 b_2 + p_4 b_3,$$

and hence $d_1 = - p_2 b_1 - p_3 b_2 - p_4 b_3$.

If the means of the regressor variables are requested as part of the regression output, then these include the values of p_2, p_3 and p_4. For example, the mean of the dummy variable PRIM is simply p_2, the (weighted) proportion of individuals with primary education. In SPSS the means are obtained by specifying

STATISTICS 2

after the regression card. From this output, we obtain the weighted proportions.

$$p_2 = .3944, \; p_3 = .2502, \; p_4 = .1333.$$

Hence the deviation for the NO SCHOOLING category is

$$d_1 = -.3944\,(-.24) - .2502\,(-.45) - .1333\,(-.60) = .29,$$

and hence $d_1 = .29$, $d_2 = .05$, $d_3 = .16$, $d_4 = -.31$. These are identical to the adjusted deviations of LVED from multiple classification analysis, as given in Table 3.4. Finally, adjusted means for each category can be calculated by adding the overall mean to each deviation d_i.

It remains to draw analogies between the analyses of variance in Table 5.3 and Tables 3.7 and 3.8. Since the regression fits the additive model [MGP6 + LVED], the regression sum of squares (27627.7) corresponds to the sum of squares for the main effects in Table 3.7 and 3.8, the difference being rounding error. In the notation of Section 3, this is $SS_{MGP6 + LVED}$. The unadjusted sum of squares for LVED, SS_{LVED}, is given by the regression sum of squares from Table 5.2, viz 5746.0. Hence the amount added by introducing MGP6 is

$$SS_{MGP6|LVED} = SS_{MGP6+LVED} - SS_{LVED} = 27627.7-5746.0 = 21881.7,$$

which corresponds to the adjusted sum of squares for MGP6 in Table 3.7 (21880.8). The unadjusted sum of squares for MGP6 and the adjusted sum of squares for LVED cannot be obtained from the existing regressions, requiring a further regression of NCEB on MGP6 alone. Also the interaction sum of squares cannot be found without fitting a regression including the interactions between the factors. This involves calculating all the product terms between the dummy variables in each group,

PRIMMG09 = PRIM x MG09, PRIMMG14 = PRIM x MG14, . . . ,

HIGHM25P = HIGH x M25P

and adding them to the regression. The addition to the regression sum of squares when the terms are added will then be $SS_{LVED.MGP6} = 206.96$.

This reconstruction of the ANOVA table by a regression program is not recommended in practice, since forming the product terms is tedious and the calculations are automatically presented in the desired form by the ANOVA program*. However it is instructive, and illustrat-

* Not all ANOVA programs, however, present sufficient output of the effects of the model, and here regression can have an advantage.

TABLE 5.4: Analysis of Covariance Using Regression Program. Regression of Parity on Years Since First Marriage and Level of Education Represented as a Set of Dummy Variables

REGRESSIONS ON NCEB

VARIABLE LIST 1
REGRESSION LIST 2

DEPENDENT VARIABLE.. NCEB
VARIABLE(S) ENTERED ON STEP NUMBER 1 .. YSFM
PRIM
RSEC
HIGH

			DF	SUM OF SQUARES	MEAN SQUARE	F
MULTIPLE R	.70846	ANALYSIS OF VARIANCE				
R SQUARE	.50191	REGRESSION	4.	27895.38699	6973.84675	1714.30350
ADJUSTED R SQUARE	.50162	RESIDUAL	6805.	27682.97861	4.06804	
STANDARD ERROR	2.01694					

. . . . . VARIABLES IN THE EQUATION . . . . . .

VARIABLE	B	BETA	STD ERROR R	F
YSFM	.2071711	.68196	.00281	5444.731
PRIM	–.1666822	–.02852	.06566	6.445
RSEC	–.3788385	–.05744	.07422	26.054
HIGH	–.5564369	–.06621	.09027	37.995
(CONSTANT)	1.135904			

. . . . . VARIABLES NOT IN THE EQUATION. . . . . .

VARIABLE	BETA IN	PARTIAL	TOLERANCE	F
YSFMSQ	–.43582	–.16824	.07422	198.186

es the sort of calculations that are done to construct the analysis of variance. The main advantages of the regression program occur when interval-scaled covariates are present, as discussed in the next section.

5.4 Covariates and Factors

If the dummy variables representing the factor MGP6 are replaced by the interval-scaled variable YSFM in the regression, we obtain the analysis of covariance model, 4.4. The addition of the quadratic term YSFMSQ gives the fit for the model 4.5. Output from these regressions are given in Tables 5.4 and 5.5, and corresponds to the output obtained from the ANOVA program given in Tables 4.1 to 4.4. The adjusted effects of education are again contained in the regression coefficients for PRIM, RSEC and HIGH. The coefficients for YSFM and YSFMSQ, on the other hand, give the effects for marital duration adjusted for education.

The F-statistics for the individual coefficients deserve some comment. In Table 5.4, the F-statistic for YSFM (5444.7) indicates the obvious fact that the adjusted linear effect of marital duration is highly significant. The standard errors and F-statistics for the dummy variables estimate the precision with which the corresponding effects are measured, and are not without interest. However, note that for variables with more than two categories the set of values presented depends on the choice of reference category. Also the F-values of pairwise differences do not present a reliable picture of the *overall* significance of the factor. It is possible for an isolated pairwise difference to be significant even through a simultaneous test for equality of the category means is not significant. Conversely, a simultaneous test may yield a significant result even though none of the pairwise differences are significant for the choice of reference category adopted. A sensible strategy here is to test for equality of the category means, and to avoid interpreting individual differences unless this test is significant. The simultaneous test is based on the change in the regression sum of squares when the factor, represented by its set of dummies, is entered. Specifically, let SS_{added} and df_{added} be the sum of squares and degrees of freedom added by the factor, and let ms_{RES} and df_{RES} be the residual mean square and residual degrees of freedom after the factor is added. Then the statistic

$$F = \frac{SS_{added}/df_{added}}{ms_{RES}}$$

is compared with an F distribution on df_{added} and df_{RES} degrees of freedom. The test is illustrated in Section 5.6.

Finally, in Table 5.5 the F-statistic for YSFMSQ is 198.19, again highly significant, indicating that it improves the fit and is worthy of inclusion.

5.5 Controlling the Order of Adjustment by Stepwise Regression

In the ANOVA program, covariates are controlled before or after the factors, according to an option specified by the user. It is not possible to interleave covariates and factors. A more flexible way of ordering controls is to fit a set of regressions using a stepwise regression program. The basic approach is illustrated with an application from the WFS Illustrative Analysis on Socio-Economic Determinants of Contraceptive Use in Thailand (Cleland, Little, and Pitaktepsombati, 1979).

The response variable is a binary variable indicating current use of contraception (CUSE), taking values one for users and zero for non-users. The regressors consisted of two interval scaled variables, respondent's age at survey (AGE) and a standard of living index (STANDLIV); one binary factor, type of place of residence (TYPE OF PLACE) taking values one for urban and zero for rural; and four factors with more than two categories, number of living children (LIVCHILD), with nine categories, region (REGION), with five categories, husband's education (HEDUC), with four categories, and husband's occupation (HOCCUP), with five cate-

TABLE 5.5: Regression of Parity on Linear and Quadratic Terms of Years Since First Marriage and Level of Education Represented as a Set of Dummy Variables

REGRESSIONS ON NCEB

VARIABLE LIST 1
REGRESSION LIST 2

DEPENDENT VARIABLE.. NCEB
VARIABLE(S) ENTERED ON STEP NUMBER 2.. YSFMSQ

MULTIPLE R	.71834	ANALYSIS OF VARIANCE	DF	SUM OF SQUARES	MEAN SQUARE	F
R SQUARE	.51601	REGRESSION	5.	28678.90879	5735.78176	1450.81960
ADJUSTED R SQUARE	.51565	RESIDUAL	6804.	26899.45681	3.95348	
STANDARD ERROR	1.98834					

. . . . . VARIABLES IN THE EQUATION . . . . . .

VARIABLE	B	BETA	STD ERROR B	F
YSFM	.3349421	1.10256	.00949	1246.024
PRIM	–.2129170	–.03643	.06481	10.793
RSEC	–.4020727	–.06096	.07319	30.182
HIGH	–.5188914	–.06174	.08903	53.967
YSFMSQ	–.3977706E-02	–.43582	.00028	198.186
(CONSTANT)	.4882426			

. . . . . VARIABLES NOT IN THE EQUATION. . . . . .

VARIABLE	BETA IN	PARTIAL	TOLERANCE	F

ALL VARIABLES ARE IN THE EQUATION

gories. An analysis of particular interest concerned the effect on the regional differentials of adjusting for the other regressors. The analysis was based on the following set of regressions.

(1) CUSE ON REGION
(2) CUSE ON REGION, LIVCHILD
(3) CUSE ON REGION, LIVCHILD, AGE
(4) CUSE ON REGION, LIVCHILD, AGE, TYPE OF PLACE
(5) CUSE ON REGION, LIVCHILD, AGE, TYPE OF PLACE, HEDUC
(6) CUSE ON REGION, LIVCHILD, AGE, TYPE OF PLACE, HEDUC, HOCCUP
(7) CUSE ON REGION, LIVCHILD, AGE, TYPE OF PLACE, HEDUC, HOCCUP, STANDLIV.

The aim was to monitor the effects of region at each step and hence determine the impact of the correlated factors and covariates. The factors were represented by blocks of dummy variables, as in the previous section. The factor REGION is introduced first so that the first regression gives the unadjusted regional means. The order of introduction of the other variables is somewhat arbitrary and is discussed in more detail in the report of the analysis. Strategies for determining the order of adjustment are reviewed in Chapter 6 of this paper.

These regressions can be carried out in a stepwise regression program in a single run, by forcing the covariates or factors into the equation in the following order:

REGION, LIVCHILD, AGE, TYPE OF PLACE, HEDUC, HOCCUP, STANDLIV.

In SPSS this is achieved by giving the variables in each block even priority levels, 14, 12, 10, 8, 6, 4, 2, respectively. This use of stepwise regression should be contrasted with the more familiar form where the regressor included or rejected at each step is determined by levels of significance. This is not appropriate here, since the set of dummies for a factor needs to be included or excluded as a block.

From the resulting output, the coefficients for REGION are identified and converted to adjusted category means using the procedure of the previous section. The resulting summary of the

TABLE 5.6: Per Cent of Currently Married, Non-Pregnant, "Fecund" Women Currently Using An Efficient Method, by Region. Adjusted for Indicated Variables by Linear Regression

		Region of Residence						
Step	Controls	Bangkok	North	Northeast	South	Central	Mean	Added R
1	–	54.9	52.8	32.5	18.9	53.7	42.6	.253
2	LIVCHILD	55.6	53.4	31.8	19.1	53.6	42.6	.259
3	LIVCHILD, AGE	56.2	53.8	31.2	19.1	53.9	42.6	.263
4	LIVCHILD, AGE, TYPE OF PLACE	44.4	54.6	32.4	19.2	54.8	42.6	.253
5	LIVCHILD, AGE, TYPE OF PLACE, HEDUC	43.5	55.3	32.2	21.1	53.5	42.6	.245
6	LIVCHILD, AGE, TYPE OF PLACE, HEDUC, HOCCUP	43.2	54.9	33.5	21.5	51.8	42.6	.221
7	LIVCHILD, AGE, TYPE OF PLACE, HEDUC, HOCCUP, STANDLIV	40.1	55.0	34.9	22.1	50.1	42.6	.208
SAMPLE SIZE = 2141								
PER CENT DISTRIBUTION =		6.7	25.8	35.3	9.9	22.3		

Source: Little, Cleland and Pitaktepsombati (1979).

effects of region is presented in Table 5.6. For a dichotomous response, the fitted means are interpreted as the proportion taking value 1, that is in this case the proportion currently using contraception. In the table these proportions are converted to percentages by multiplying by 100. The first row of the table is unadjusted, and is simply a cross-tabulation of the regional means. As variables are introduced, the adjusted means tend to converge towards the overall mean of 42.6 indicating the effect of the composition of other variables on the regional differentials. However, even after all the other controls are included, the effects of region are still large and statistically significant, suggesting that other unmeasured factors are contributing to the regional disparity of contraceptive use. A fuller interpretation of the table is given in the original paper.

The last column of Table 5.6 deserves comment. The proportion of additional variance explained by region at each step is found as

$$\text{Added } R^2 = (SS_{REGION+OV} - SS_{OV}) / SS_T$$

where SS_{OV} is the sum of squares for the other variables, excluding region, $SS_{REGION+OV}$ is the sum of squares for region and the other variables, and SS_T is the total sum of squares about the mean. The square root of this measure is presented in the last column of the table; it is equivalent to the BETA measure in Multiple Classification Analysis. According to this measure, the effects of Region are reduced from .253 to .208 by the inclusion of the other controls, a reduction of some 20%. Such summary conclusions are of some interest, but cannot replace the detailed information in the body of the table.

5.6 Interactions Between Factors and Covariates

The last example of analysis of covariance in Section 4.2 involved the weighted regression of parity divided by marital duration (P/D) on factors LVED and HEDL and covariates consisting of linear and quadratic terms in years since first marriage (YSFM, YSFMSQ) and age at first marriage (AGFM, AGFMSQ), and the interaction formed by the product of the linear terms (YSFMAGFM). In that section it was pointed out that interactions between covariates and factors could not be modelled within the ANOVA procedure. In this section we model these interactions using the more general REGRESSION program.

For simplicity we shall restrict ourselves to one factor, LVED, represented in the regression by the three dummy variables PRIM, RSEC and HIGH, as before. The covariates for marital duration and age at first marriage are defined as above, that is

YSFM, YSFMSQ, AGFM, AGFMSQ, YSFM. AGFM.

Variables for the interactions between factors and covariates are defined by forming products of the covariates and dummy variables. The following are defined:

LVED.YSFM, LVED.YSFMSQ, LVED.AGFM, LVED.AGFMSQ, LVED.AGFM.YSFM.

Each of these terms involves three variables; for example, LVED.YSFM is represented by the three products PRIM x YSFM, RSEC x YSFM, and HIGH x YSFM. The list of interactions thus formed is not exhaustive, since it does not include three way interactions involving LVED and the quadratics YSFMSQ or AGFMSQ.

The data are analyzed by a stepwise regression, with variables added in the following steps:

1. LVED; 2. YSFM, YSFMSQ; 3. AGFM, AGFMSQ; 4. YSFM.AGFM;
5. LVED.YSFM; 6. LVED.YSFMSQ; 7. LVED.AGFM; 8. LVED.AGFMSQ;
9. LVED.YSFM.AGFM.

TABLE 5.7: Analysis of Variance of Regression of PBYD with Interactions Added Hierarchically

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(j)	(k)	(l)	(m)
		Regression		Residual			R^2	Added at Step			
Step	Variables Added	Sum of Squares	DF	Sum of Squares	DF	Mean Square		Sum of Squares	DF	R^2	F
1	LVED	6.19	3	11,230	6556	1.713	.0006	6.19	3	.0006	1.205
2	YSFM, YSFMSQ	1169.55	5	10,066	6554	1.536	.1041	1163.36	2	.1035	395.16
3	AGFM, AGFMSQ	1509.23	7	9,726	6552	1.485	.1343	339.68	2	.0302	115.38
4	YSFM.AGFM	1549.91	8	9,686	6551	1.479	.1369	40.68	1	.0026	27.64
5	LVED.YSFM	1564.34	11	9,671	6548	1.477	.1392	14.43	3	.0023	3.27
6	LVED.YSFMSQ	1596.02	14	9,640	6545	1.473	.1421	31.68	3	.0029	7.17
7	LVED. AGFM	1602.89	17	9,633	6542	1.472	.1427	6.87	3	.0006	1.56
8	LVED.AGFMSQ	1604.79	20	9,631	6539	1.473	.1428	1.90	3	.0001	.43
9	LVED.YSFM.AGFM	1614.58	23	9,621	6536	1.472	.1437	9.79	3	.0009	2.22

The analysis of variance from the regression at each step is presented in columns (c) to (g) of Table 5.7. The regression sum of squares and degrees of freedom appear in columns (c) and (d), and the sum of squares, degrees of freedom and mean square for the residual are given in columns (e), (f) and (g). From these values, the sum of squares and degrees of freedom added at each step can be derived by subtraction, and are given in columns (j) and (k). For example, the sum of squares added by LVED.YSFM at Step 5 is 1564.34 - 1549.91 = 14.43. Other statistics presented in the table are the regression R^2, in column (h), and the R^2 added at each step (or partial R^2), in column (1). The final column gives the F-statistic for the net effect of each term when it is added to the regression, obtained by dividing the mean square added at each step by the residual mean square 1.472 at the final step, viz, Step 9. This test differs slightly from that described in Section 5.4, in that the residual mean square is taken from the final step rather from the step at which the variable is added. Both tests are valid; in the chosen method the residual mean square is the same for the tests at each step, and thus the F-statistics are more directly comparable.

The following points emerge from this summary table:

1. The percentage of variance explained by all the variables is 14.4%. This is less than that obtained by regressions with parity as response, but the comparison is misleading: the response PBY D encorporates a partial control for marital duration in its definition, and hence a large explanatory factor in regressions on NCEB is discounted by the choice of response.

2. The introduction of LVED at Step 1 is not significant. However, the interpretation of differentials in PBYD is not clear unless these effects are adjusted for marital duration. Although this is not evident from the table, educational differentials emerge after Step 2, when this control is implemented.

3. The demographic controls YSFM, YSFMSQ, AGFM, AGFMSQ and YSFM.AGFM are highly significant, from Steps 3, 4 and 5. Inspection of the individual coefficients indicates that the quadratics YSFMSQ and AGFMSQ add significantly to the fit.

4. Significant interactions between education and marital duration emerge at Steps 5 and 6. Taken together, LVED.YSFM and LVED.YSFMSQ add a mean square of

$$(14.4 + 31.7) / 6 = 7.69,$$

compared with the residual mean square of 1.472. The nature of these interactions are described below.

5. The last three steps of the regression, taken together, do not add significantly to the fit, although the three-way interaction yields an F-value of 2.22. We shall not interpret these effects in subsequent analysis.

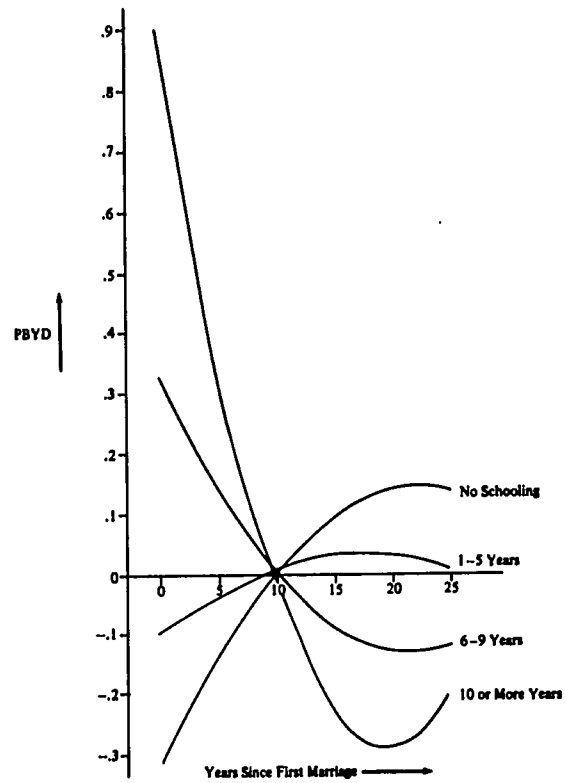
We now concentrate on the adjusted effects of educational level. These are presented in Table 5.8 for the first six steps of the stepwise regression, in the form of deviations from the mean. The first four steps involve models which are additive with respect to LVED (that is, involve no interactions with LVED). Thus the effects of education are found from the means and regression coefficients for the dummy variables PRIM, RSEC and HIGH, using the method described in Section 5.3.2.

Steps 5 and 6 include interactions between education and marital duration, and as a result the effects of education depend on the level of marital duration. The adjusted effect for each education category is obtained by subtracting two functions of marital duration and age at marriage, the fitted mean for the whole group obtained by substituting mean values for PRIM, RSEC and HIGH in the equation, and the fitted mean for the education category, obtained by substituting one or zero for the dummy variables as appropriate. For example, in Step 5 the fitted equation is:

TABLE 5.8: Effects of Education from Regression of PBYD Including Interactions, Expressed as Deviations From Mean

Step	Variables Added	No Schooling	1–5 Years	6–9 Years	10 or More Years
1	LVED	–.031	–.001	.016	.089
2	YSFM, YSFMSQ	.128	–.022	–.119	–.301
3	AGFM, AGFMSQ	.083	–.008	–.088	–.039
4	YSFM.AGFM	.083	–.002	–.087	–.074
5	YSFM.LVED	–.011 +.0048 YSFM	–.037 +.0024 YSFM	–.009 –.0034 YSFM	.274 –.0230 YSFM
6	YSFMSQ.LVED	–.310 +.0403 YSFM –.0009 $YSFM^2$	–.096 +.0142 YSFM –.0004 $YSFM^2$	.320 –.0427 YSFM +.0010 $YSFM^2$	.883 –.1208 YSFM +.0031 $YSFM^2$

FIGURE 5.1: Fitted Effects of Education as Quadratic Functions of Marital Duration, from Step 6 of Regression



$$\hat{Y} = -.026\ PRIM + .002\ RSEC + .285\ HIGH - .0024\ YSFM.PRIM - .0081\ YSFM.RSEC$$
$$-.0278\ YSFM.HIGH + o.t,$$

where o.t. represents other terms not involving PRIM, RSEC or HIGH, which cancel in the subtraction. The effect for the 1-5 Years Schooling Group is obtained by subtracting the fitted mean with PRIM = $\overline{PRIM}$, RSEC = $\overline{RSEC}$ and HIGH = $\overline{HIGH}$ from the fitted mean with PRIM = 1, RSEC = HIGH = 0. Substituting the observed means $\overline{PRIM}$ = .425, $\overline{RSEC}$ = .206 and $\overline{HIGH}$ = .075, we obtain for the adjusted effect

$$\{-.026 - .0024\ YSFM + o.t.\}$$
$$- \{(.026)(.425) + (.002)(.206) + (.285)(.075) + YSFM\ [(-.0024)(.425)$$
$$+ (-.0081)(.206) + (-.278)(.075)]\ + o.t.\}$$
$$= -.037 + .0024\ YSFM,$$

as seen in the table. The calculation in effect repeats the procedure for calculating the main effects of education, described in Section 5.3.2, for the interactions terms which include education.

We conclude by giving a substantive interpretation of Table 5.8. The unadjusted effects in Step 1 are relatively small and of limited substantive interest. When marital duration is controlled (Step 2), we note that the NO SCHOOLING group has the largest adjusted fertility tempo P/D, and the highest education group has the smallest, differing from the no schooling group by nearly one half a birth per ten years marriage duration. The intermediate education groups rank in the expected way. In Steps 3 and 4, we learn that a large part of the differential in fertility tempo is attributable to the quadratic effect of Age at Marriage, namely that the more educated women marry later and hence have a lower average tempo of fertility. Finally, Steps 5 and 6 indicate that the residual effects of education after adjusting for age at marriage and marital duration are specific to marital duration. Step 5 shows that recent marriage cohorts, the tempo of fertility is positively associated with education. Thereafter the differentials decline with marital duration, and for cohorts married ten or more years the pattern is reversed. Step 6 estimates quadratic relationships between the effects and marital duration. Effects from this step are plotted in Figure 5.1, which shows very clearly the cross-over between low and high durations of marriage.

The interpretation of these results is not easy for the present example since fertility in Sri Lanka is declining, and it is not possible to distinguish life cycle effects and trends in fertility. The separation of these components requires alternative measures of fertility, such as are used in the Sri Lanka illustrative analysis of cumulative fertility (Little and Perera, 1980). Nevertheless, the example does illustrate the formation and interpretation of interactions within a regression model.

6. STRATEGIES FOR DETERMINING THE CHOICE OF VARIABLES IN THE REGRESSION

6.1 Introduction

We have presented a flexible collection of methods for calculating adjusted effects of interval-scaled and categorical regressors, and for assessing their statistical significance. Given a set of regressors*, the most important issue facing the analyst applying these techniques systematically to data is which variables to control when calculating the effects of a regressor, and how to interpret the results substantively. A brief introduction to this topic is presented in this final chapter. Parts of the discussion are based on Little (1979).

Two extreme strategies encountered in the literature are:

a) to calculate the effects of all the regressors unadjusted, in the form of one-way cross-tabulations of means or univariate regressions, and

b) to calculate the effects of each variable adjusted for all other regressors in the study, using a single regression equation with all variables included.

The former method is clearly unsatisfactory, as noted in the early chapters of this bulletin. The latter method is not uncommon, but can lead to considerable problems when highly associated regressors are included. For example, it is quite possible that the adjusted effects of husband's and respondent's education are not significant when both are included in the regression, even though the effect of education as measured by either one alone is highly significant.Suppressing effects of this type are described in Gordon (1968).

A more illuminating approach is to consider what adjusted effects represent in the context of a causal ordering between the variables.

The definition of an adjusted effect is at first glance straightforward – it represents the average effect on the regressand of increasing the regressor by one unit, holding other variables in the regression fixed. Such statements have a descriptive value for the population under study, but they should not be regarded as a basis for causal inference. That is, it does not follow that if a policy maker was in fact able to create conditions in the population which led to an increase in the average value of a regressor, holding other regressors constant, then this increase would necessarily result in the increase in the mean regressand predicted by the model.

The potential absurdity of causal inferences of this kind is easily demonstrated. For example, the relationship between fertility and contraceptive use may be explored by a regression of number of children ever born on current use of contraception, adjusted for demographic and socio-economic controls according to taste. For many developing countries the resulting adjusted effect of contraceptive use is positive, that is the mean parity increases with level of use. The reason is that at early stages of a family planning program contraceptive use tends to be concentrated among women with large families. The implied causal inference is that the result of increasing the level of contraceptive use is to increase fertility, which is clearly absurd. Correct causal inferences about the relationship between contraceptive use and fertility require information about the timing of births and contraception for individuals in the sample.

A more subtle example concerns the relationship between education and fertility. Many countries show a negative relationship between level of formal schooling and fertility, after adjustment for controls for exposure to risk of childbearing. The extent to which this observed relationship can be used to infer than an emphasis on increasing education facilities will yield the predicted fertility decline is questionable. In the past education might be restricted to an elite group, and as education spreads, it affects different groups of the population. It is not

* The choice of regressors to be included in the study is an important issue which lies outside the scope of this technical bulletin.

necessarily true that the relationship between education and fertility will be the same for different cohorts of the population. It is not clear that formal schooling is a factor which directly affects the level of fertility, since the observed relationship between formal education and fertility could be caused by other factors which are associated with formal schooling, but would not be affected by an increase in the level of formal schooling. Finally, the effect of increasing education would presumably depend on the specific policies introduced to bring about the change.

Despite the evident difficulty in making direct causal inferences from the data, any analysis which goes beyond simple reporting of the means of variables is presumably trying to provide information which is ultimately causal in nature. Furthermore, causal analysis provides a valuable conceptual framework for deciding the specific question of which variables to control when analysing the effect of a variable in a regression. The most important aspects of this framework are now presented.

6.2 The Causal Ordering and Total Effects

We suppose that the regressor and regressand variables can be placed in a causal ordering

$$X_1 \longrightarrow X_2 \longrightarrow X_3 \longrightarrow \ldots \longrightarrow Y, \tag{2.1}$$

such that changes in the values of any variable can affect a variable later in the chain, but do not affect variables earlier in the chain. Two points require special emphasis here:

a) The causal ordering cannot be decided by an empirical analysis of the data, but must be based on prior theoretical knowledge of the population;

b) The specification of a causal ordering in effect rules out the possibility of circular causation between variables, where one variable both affects and is affected by another variable in the series. In the examples, we shall proceed under the assumption that at least a predominant direct of causal ordering can be established. In cases where this is not possible the interpretation of the data is much more difficult, and more complex analytical techniques than those discussed are required to disentangle relationships between the variables. See, for example, the non-recursive models discussed by Hood and Koopmans (1953).

Two general rules stem from this causal ordering:

Rule 1 The regressand variable Y, must be the last variable in the causal chain. In other words, variables causally posterior to the response should not be included.

Rule 2 In assessing the effect of any regressor variable X on a response, Y, all variables causally prior to X should be controlled.

To clarify these rules, consider a particular regressor variable X. We can represent the position of X in the causal chain as follows:

$$X_b \longrightarrow X \longrightarrow X_a \longrightarrow Y,$$

where X_b are the set of regressor variables prior to X, X_a are the set of regressor variables posterior to X, and the response Y is by rule 1 the last variable in the chain. Then Rule 2 states that the variables X_b should be controlled when calculating the effect of X on Y.

Rule 2 does not specify whether the regressor variables posterior to X, X_a, should be controlled. If none of these are controlled, the resulting effect of X is called the *total* effect. The total effect of a variable X on a response Y is the effect calculated with all regressor variables prior to X controlled and all regressor variables causally posterior to X not controlled. For a given causal ordering the total effect is the effect with the clearest substantive interpretation.

The rational is that changes in the distribution of X will not affect variables prior to X in the chain, but will affect variables posterior to X.

In addition to calculating the total effect, it is also possible to assess the extent to which the effect operates through changes in the intervening variables X_a in the causal chain. If the variables X_a are controlled, as well as X_b, we obtain the so-called *direct* effect of X on the response. The difference between the total effect and the direct effect is called the *indirect* effect of X on Y through X_a and represents the effect of X on Y operating through changes in the distribution of X_a. The most developed form of these decompositions occurs in the technique of recursive path analysis, which is described in another technical bulletin (Kendall and O'Muircheartaigh, 1977).

6.3 Examples

If a predominant direction of causation can be established, then the total effects of variables can be calculated for this ordering. In addition, the total effects can be decomposed into direct and indirect components if this the decomposition is of substantive interest. The following examples illustrate the method.

Example 1: X_1 = Respondent's age, X_2 = Education, X_3 = Age at marriage, Y = Parity. One plausible causal ordering is:

Age ⟶ Education ⟶ Age at marriage ⟶ Parity

Age is a cohort marker and fully exogenous to the other variables. To the extent that children are born after marriage, the response variable Parity does not affect the respondent's history up to marriage and hence can be considered causally posterior to education and age at marriage. The placement of education prior to age at marriage is less certain, and in some populations might reflect a predominant direction of causation. Although in some cases a respondent may terminate her education to get married, for the most part education has the effect of delaying age at marriage, and this is reflected in the chosen direction of causation between these variables. Given the ordering, the total effect of age on parity is unadjusted, the total effect of education on parity is adjusted for age, and the total effect of age at marriage on parity is adjusted for age and age at marriage. In practice, the effect of education often calculated is the direct effect adjusted for age *and* age at marriage. One practical reason for this is that with an ever-married sample the total effect is biased because of selection effects. However, the direct effect does not take into account the indirect effect of education operating through changes in age at marriage.

Example 2: X_1 = Marital duration, X_2 = Education, Y = Parity. Here the predominant causal ordering is:

Duration ⟶ Education ⟶ Parity

However the causal relationship between duration and education is not clear, because marriage duration includes components of age and age at marriage which, according to the previous example, are respectively prior and posterior to education. The total effect of education on parity in this system is obtained by controlling marital duration.

Example 3: X_1 = Age, X_2 = Age at marriage, X_3 = Current use of contraception, X_4 = Parity. Consider two causal orderings, with (a) Y = X_4, i.e. parity, as response and (b) Y = X_3, i.e. contraceptive use, as response:

(a) Age ⟶ Age at marriage ⟶ Contraceptive use ⟶ Parity

(b) Age ⟶ Age at marriage ⟶ Parity ⟶ Contraceptive use

The causal ordering between contraceptive use and parity in (a) seems plausible, as one expects

that contraceptive use affects the number of live births a woman has. However, in practice the predominant causal ordering is more likely to be (b), particularly in countries where family planning is of recent origin. That is, women with high parities are more likely to use contraception, and consequently parity is a major determinant of contraceptive use; although contraceptive use may have an inhibiting effect on parity, this effect is smaller in the initial stages of a family planning programme. The consequences of this circularity were noted in Section 6.1.

Example 4: X_1 = Age, X_2 = Age at marriage, X_3 = Education, X_4 = Desired family size, Y = Parity. Here the causal ordering is:

$$\text{Age} \longrightarrow \text{Education} \longrightarrow \text{Age at marriage} \longrightarrow \text{Desired family size} \longrightarrow \text{Parity}$$

seems plausible. However, in a real population the relationship between the last two variables is complicated to the extent that women tend to rationalize their stated desired family size on the basis of how many children they in fact have had. Thus, again, circular causation is a possibility which obscures the interpretation of the data.

6.4 A Compromise Strategy

As can be seen from the examples of the previous section, the principal difficulty of the proposed strategy is that in practice it is often hard to justify even an approximate causal ordering between the variables. Consequently a more flexible approach may be desirable, where the effects of a variable are calculated with a variety of controls. The extreme version of this strategy would be to calculate effects for all possible subsets of controls, but this soon produces an unpalatable amount of data. A compromise solution, which relies to some extent on a causal ordering but calculates a range of effects for each variable, has been adopted in two WFS Illustrative Analyses (Cleland, Little and Pitaktepsombati, 1979; Little and Perera, 1980). An ordering

$$X_1 \longrightarrow X_2 \longrightarrow \ldots \longrightarrow X_k \longrightarrow Y$$

is decided on causal or substantive grounds. For each variable X_j, the unadjusted effect is calculated first. Then other variables are added in (K-1) steps, according to the ordering obtained by moving X_j to the beginning of the sequence. At each step the adjusted effects of X_j are calculated. The results of this strategy for a single variable are shown in Table 5.6, and discussed in Section 5.5. The output is still dependent on the choice of ordering, but the method does provide information on the effects of each variable with a variety of controls, and as such illuminates some of the consequences of association between the regressors which is the principal motivation of these methods.

REFERENCES

REFERENCES IN TEXT

Cleland J.G., Little R.J.A. and Pitaktepsombati P. (1979). "Illustrative Analysis: Socio-economic determinants of contraceptive use in Thailand". *WFS Scientific Reports* series, No. 5, International Statistical Institute, Voorburg, Netherlands.

Dunnell K. (1976). *Family Formation 1976.* Office of Population Censuses and Surveys, Social Survey Division, HM Stationery Office, London.

Gordon R.A. (1968). "Issues in Multiple Regression". *American Journal of Sociology,* 3,592-601.

Hermalin A.I. and Mason W.M. (1979). *"A Strategy for Comparative Analysis of WFS Data, with Illustrative Examples,"* The Population Studies Centre, University of Michigan, May 1979.

Hood W.C. and Koopmans T.C. (eds) (1953). *Studies in Econometric Method.* Cowles Commission Monograph No. 14, Wiley Press.

Kendall M.G. and O'Muircheartaigh C.A. (1977). "Path analysis and model building". *WFS Technical Bulletin* series, No. 2, International Statistical Institute, Voorburg, Netherlands.

Little R.J.A. (1977). "Regression Models for Differentials in Fertility". 1. Parity as Response. *WFS Tech. Paper* 635.

(1978). "Generalized linear models for cross-classified data". *WFS Technical Bulletin* series, No. 5, International Statistical Institute, Voorburg, Netherlands.

(1979). "Linear Models and Path Analysis" From Regional Workshop on Techniques of Analysis of World Fertility Survey Data. *UN* Asian Population Studies series No. 44. Reprinted in *WFS Occasional Papers* series, International Statistical Institute, Voorburg, Netherlands.

and Perera S. (1980). "Illustrative Analysis. Socio-economic Differentials in Cumulative Fertility in Sri Lanka,. A Marriage Cohort Approach". *WFS Scientific Reports,* series, No. 12 (in press) International Statistical Institute, Voorburg, Netherlands.

Tukey J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley.

World Fertility Survey (1975). *"Sri Lanka 1975 - First Report".* Department of Census and Statistics, Colombo, Sri Lanka.

REFERENCES FOR FURTHER READING

Draper N. and Smith H. (1966). *Applied Regression Analysis,* John Wiley and Sons.

Nie N.H., Hull C.H., Jenkins J.G., Steinbrenner K. and Bent D.H. (1975). *SPSS, Second Edition.* McGraw-Hill.

Pullum T. (1978). "Standardisation". *WFS Technical Bulletin* series, No. 3, International Statistical Institute, Voorburg, Netherlands.

Scheffé H. (1959). *The Analysis of Variance.* John Wiley and Sons.

TABLE D1: Mean Number of Children Ever Born, by Marital Duration (MGP6) and by Level of Education (LVED). a) Means, b) Sample Sizes, c) Standard Deviations

	LVED				
MGP6	No Schooling (1)	1–5 Years (2)	6–9 Years (3)	10+ Years (4)	Row Total
0–4 (1)	.96 a) 112 b) .84 c)	.88 376 .76	.95 442 .78	.92 351 .77	.92 1280 .78
5–9 (2)	2.54 172 1.24	2.46 442 1.28	2.39 362 1.21	2.39 255 1.19	2.44 1231 1.23
10–14 (3)	3.87 197 1.67	3.91 482 1.72	3.73 293 1.49	3.14 145 1.47	3.76 1118 1.64
15–19 (4)	5.13 239 2.35	4.97 461 2.34	4.61 262 2.18	4.13 95 2.10	4.84 1057 2.30
20–24 (5)	6.22 292 2.62	5.87 377 2.38	5.22 184 2.64	4.47 40 2.11	5.79 893 2.54
25 + (6)	6.92 501 3.16	6.55 548 2.99	6.23 161 2.70	5.97 22 1.98	6.65 1231 3.02
Total	5.17 1512 3.10	4.24 2686 2.85	3.26 1704 2.47	2.30 908 1.85	3.94 6810 2.86
Standardized Means	4.14	3.98	3.75	3.43	3.88

Source: Special Tabulation Sri Lanka Fertility Survey 1975.

TABLE D2: Mean Number of Children Ever Born, by Age (AGP5), by Age at First Marriage (AMGP), and by Level of Education (LVED). a) Means and b) Sample Sizes

LVED = 1 NO SCHOOLING

	AMGP				
AGP5	<15 (1)	15–19 (2)	20–24 (3)	25+ (4)	Row Total
15–24 (1)	2.83 a)	1.65	.64	.00	1.75
	29 b)	101	19	0	149
25–29 (2)	4.39	3.26	1.91	.57	3.05
	43	103	55	7	207
30–34 (3)	5.32	4.88	3.41	2.30	4.66
	67	101	38	9	215
35–39 (4)	6.84	6.01	4.68	1.90	5.96
	111	139	41	13	304
40–49 (5)	7.34	6.77	5.86	2.95	6.45
	173	293	121	51	637
Total	6.28	5.18	4.18	2.50	5.17
	423	736	274	79	1512

LVED = 2 1 TO 5 YEARS

	AMGP				
AGP5	<15 (1)	15–19 (2)	20–24 (3)	25+ (4)	Row Total
15–24 (1)	2.93	1.41	.56	.00	1.49
	66	311	78	0	456
25–29 (2)	4.56	3.30	1.90	.45	3.00
	94	213	134	25	466
30–34 (3)	5.99	4.61	3.43	1.66	4.30
	80	258	105	48	490
35–39 (4)	6.31	5.86	4.50	2.64	5.23
	73	197	112	47	430
40–49 (5)	6.85	6.46	5.30	2.82	5.88
	132	425	202	85	844
Total	5.54	4.44	3.54	2.22	4.24
	446	1404	632	204	2686

TABLE D2: Mean Number of Children Ever Born, by Age, by Age at First Marriage, and by Level of Education (cont'd.)

LVED = 3 6–9 YEARS

AGPS	AMGP <15 (1)	15–19 (2)	20–24 (3)	25+ (4)	Row Total
15–24 (1)	2.62 14	1.53 234	.69 115	.00 0	1.31 364
25–29 (2)	3.97 28	3.47 125	1.65 184	.68 43	2.31 380
30–34 (3)	4.52 26	4.34 126	3.23 82	1.55 68	3.43 302
35–39 (4)	5.82 15	5.66 106	4.28 113	2.26 58	4.46 292
40–49 (5)	6.74 23	6.09 143	5.03 110	3.11 90	5.08 366
Total	4.78 106	3.83 735	2.79 604	2.11 259	3.26 1704

LVED = 4 10 OR MORE YEARS

AGPS	AMGP <15 (1)	15–19 (2)	20–24 (3)	25+ (4)	Row Total
15–24 (1)	1.58 5	1.24 42	.64 73	.00 0	.89 119
25–29 (2)	.00 0	3.15 33	1.56 147	.72 63	1.55 242
30–34 (3)	4.71 2	4.11 17	2.86 80	1.64 115	2.33 214
35–39 (4)	.00 0	5.42 13	3.48 65	2.29 100	2.95 177
40–49 (5)	5.55 2	5.88 12	4.59 56	2.91 84	3.79 155
Total	3.35 9	3.14 117	2.35 421	1.96 361	2.30 908

Source: Special Tabulation, Sri Lanka Fertility Survey 1975.

TABLE D3: Mean Number of Children Ever Born,* by Marital Duration (MGP5) and by Level of Education (LVED). a) Means, b) Sample Sizes

	LVED			
MGP5	Lower Secondary (1)	Upper Secondary (2)	Higher (3)	Row Total
0–4 (1)	.81 a) 365 b)	.58 338	.35 147	.64 850
5–9 (2)	1.71 555	1.50 371	1.36 121	1.60 1047
10–14 (3)	2.27 560	2.07 272	2.08 95	2.20 927
15–19 (4)	2.56 651	2.40 213	2.27 84	2.50 948
20–24 (5)	2.75 572	2.32 151	2.79 56	2.67 779
25 + (6)	2.83 391	2.44 68	2.45 11	2.76 470
Column Total	2.22 3094	1.66 1413	1.53 514	1.99 5021

Source: Special Tabulation from U.K. Family Formation Survey, 1976. See Dunnell (1976).
*Note that the sample base is restricted to ever-married women. Consequently the last two marriage groups are biased towards women who marry early.

TABLE D4: Means of Parity Divided by Marital Duration (PBYD), Weighted by Marital Duration, Cross-Classified by Marital Duration (MGP6) and by Level of Education (LVED)
a) Means, b) Years of Exposure, c) Standard Deviations

	LVED				
MGP6	No Schooling (1)	1–5 Years (2)	6–9 Years (3)	10+ Years (4)	Row Total
0–4 (1)	3.53 a) 303 b) 2.26 c)	3.58 897 2.19	4.08 1028 2.15	4.18 765 2.26	3.90 2993 2.22
5–9 (2)	3.42 1274 1.63	3.38 3219 1.64	3.33 2596 1.52	3.32 1837 1.50	3.36 8927 1.58
10–14 (3)	3.15 2419 1.29	3.12 6038 1.35	2.99 3660 1.18	2.55 1784 1.20	3.02 13901 1.29
15–19 (4)	2.95 4153 1.30	2.87 7969 1.34	2.67 4513 1.20	2.41 1633 1.19	2.80 18268 1.29
20–24 (5)	2.77 6562 1.14	2.61 8467 1.06	2.35 4093 1.19	2.00 899 .92	2.58 20020 1.13
25 + (6)	2.36 14700 1.08	2.27 15805 1.04	2.16 4634 .94	2.16 598 .76	2.29 35738 1.04
Total	2.66 29411 1.24	2.69 42395 1.30	2.70 20525 1.35	2.78 7515 1.51	2.69 99846 1.31

Source: Special Tabulation, Sri Lanka Fertility Survey 1975.