

Other Linear Models

CONTENTS

11.1	Introduction	579
11.2	The Dummy Variable Model	580
11.3	Unbalanced Data	584
11.4	Computer Implementation of the Dummy Variable Model	587
11.5	Models with Dummy and Interval Variables	589
11.6	Extensions to Other Models	601
11.7	Estimating Linear Combinations of Regression Parameters	602
11.8	Weighted Least Squares	606
11.9	Correlated Errors	610
11.10	Chapter Summary	613
11.11	Chapter Exercises	619

■ Example 11.1: Survival of Cancer Patients

Medical researchers are interested in determining how the survival time of cancer patients is affected by the grade of the tumor and age of the patient. The data in Table 11.1 show the result of a hospital study, showing the survival time of patients along with their ages and histological grade of their tumor. At first glance a regression of survival on grade and age would appear to be useful. However, histological grade is not strictly a numerical measure and is therefore more like a factor as used in analysis of variance models.

One of the topics of this chapter deals with the analysis of models that include both measured and categorical independent factors. The analysis is given in Section 11.10.

Table 11.1 Survival of Cancer Patients

OBS	SURVIVAL	AGE	GRADE	OBS	SURVIVAL	AGE	GRADE
1	69	46	1	41	94	31	3
2	80	68	1	42	92	60	3
3	78	58	1	43	1	80	3
4	85	32	1	44	45	84	3
5	30	59	1	45	13	83	3
6	63	61	1	46	8	65	3
7	73	52	1	47	25	80	3
8	100	57	1	48	13	73	3
9	73	65	1	49	72	63	3
10	96	58	1	50	63	62	4
11	72	60	1	51	31	61	4
12	4	61	2	52	16	63	4
13	83	62	2	53	108	58	4
14	38	75	2	54	20	66	4
15	107	64	2	55	1	73	4
16	72	53	2	56	70	82	4
17	82	66	2	57	6	69	4
18	70	62	2	58	13	60	4
19	105	62	2	59	69	67	4
20	8	86	2	60	53	53	4
21	68	67	2	61	13	61	4
22	8	58	2	62	80	56	4
23	67	91	2	63	103	79	4
24	96	58	2	64	63	67	4
25	91	61	2	65	67	63	4
26	90	66	2	66	67	82	4
27	97	68	2	67	75	70	4
28	89	60	2	68	43	72	4
29	89	77	2	69	78	74	4
30	90	44	2	70	41	68	4
31	66	68	2	71	77	61	4
32	100	68	2	72	1	46	4
33	88	42	2	73	64	69	4
34	64	57	2	74	7	62	4
35	30	66	2	75	31	48	4
36	87	59	2	76	18	78	4
37	28	57	2	77	6	60	4
38	73	67	2	78	2	50	4
39	60	70	3	79	1	80	4
40	67	64	3	80	32	62	4

11.1 INTRODUCTION

The linear model was first introduced in Chapter 6 for justifying the use of the F ratio from the analysis of variance for inferences on a set of means. That model was written

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

and relates the observed values of the response variable y to parameters representing differences among population means, denoted as $(\mu + \tau_i)$, and a random error, ε_{ij} .

The model is primarily used to make inferences about differences among the τ_i , the so-called treatment or factor level effects. The model was generalized in Chapters 9 and 10 by adding parameters representing different treatment factors and/or experimental conditions. For all such models, called **analysis of variance models**, the analysis is based on the partitioning of sums of squares and on using ratios of the resulting mean squares as test statistics for making inferences on parameters that represent the various means.

Another linear model, the **regression model**, was introduced in Chapter 7. This model relates the response variable y to an independent variable x using the relationship

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where the parameters β_0 and β_1 are the regression coefficients that specify the nature of the relationship and ε is the random error. This model was generalized in Chapter 8 to allow several independent variables. Again the inferences are based on the partitioning of sums of squares and computing F ratios using the corresponding mean squares. In the analysis of multiple regression models, it was also necessary to recognize the proper interpretations and inferences for partial coefficients, reflecting the effect of the corresponding independent variable over and above that provided by all the other variables in the model.

Although the analysis of variance and regression models appear to have different applications and somewhat different analysis procedures, both are examples of a very broad set of statistical models known as the **general linear model**. A comprehensive presentation of the general linear model is beyond the scope of this text; however, there are some very useful features that will be introduced in this chapter. These are

- the use of dummy variables to represent analysis of variance models as regression models,
- the use of dummy variables to analyze unbalanced factorial data sets,
- the use of models that contain both dummy and interval independent variables with emphasis on the analysis of covariance,