Advanced Applied Statistics

Jianqing Fan and Runze Li

Contents

1

Overview of Linear Models

Linear regression analysis is one of the most classical and commonly used techniques in statistics. A thorough understanding of linear regression is essential for those who want to become statisticians or data analysts in a variety of applied statistics fields. There are many excellent books in the topic of linear models. This chapter gives a brief overview of linear models. Details may refer to the text-books, such as, Neter, Kutner, Nachtsheim and Wasserman (1996) and Smith and Young (2000).

1.1 Linear Models

For given pairs of data $(X_i, Y_i),\ i = 1, \cdots, n$, one tries to fit a line through the data. In other words, the data are regarded as a sample

from the linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.1}$$

where ε is a random error with mean 0. An extension of model (1.1) arises when there are many x variables, resulting in the equation

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \tag{1.2}$$

In statistics, the X variables are often called explanatory variables , or independent variables , or covariates , while the Y variable is called the response variable, or the dependent variable. The β's are unknown parameters, which have to be estimated.

Linear model refers to the fact that the regression function, denoted by $m(x)$ and defined by

$$m(x) \equiv E(Y|X = x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

is linear in β's. For model (1.1), the linear assumption of regression function $m(x)$ on x is not always granted. One may try polynomial regression:

$$m(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p.$$

The polynomial regression is a linear model as it is linear in β's, though nonlinear in x.

Example 1.1 (**Environmental Data**) *There are 730 observations in the data set, collected in Hong Kong from January 1, 1994 to December 31, 1995. The data consist of daily measurements of levels of air pollutants and the number of total hospital admissions for circulatory and respiratory problems. Of interest is to investigate the*

association between the number of total hospital admissions and the levels of air pollutants.

We set the Y variable to be the number of total hospital admissions and the X variables the levels of air pollutants. Define

X_1*: the level of sulfur dioxide* ($\mu g/m^3$)*;*

X_2*: the level of nitrogen dioxide* ($\mu g/m^3$)*;*

X_3*: the level of dust* ($\mu g/m^3$)*.*

The following linear regression model may be fitted to the data,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon. \tag{1.3}$$

Linear models can be very complicated by introducing dummy variables and creating new variables. For example, one may want to consider seasonal effects in Example 1.1. Define the season indicators S_1, S_2, S_3 and S_4, which are dummy variables, for spring, summer, autumn and winter, respectively. Including the seasonal indicators in model (1.3) yields a new linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 S_1 + \beta_5 S_2 + \beta_6 S_3 + \varepsilon. \tag{1.4}$$

This model yields different intercepts for different seasons. The β_4 is the difference between the intercept for spring and the intercept for winter.

As an extension of model (1.4), one may consider the following model having different intercepts and slopes for different seasons

$$m(\mathbf{x}) = \beta_{0i} + \beta_{1i} x_1 + \beta_{2i} x_2 + \beta_{3i} x_3$$

for the i-th season, which can be rewritten as

$$m(\mathbf{x}) = \sum_{i=1}^{3}(\beta_{0,i}S_i + \beta_{1,i}S_iX_i).$$

One may also construct a more complicated model by including interaction terms.

1.2 Methods of Least Squares

Assume that $(X_{i1}, \cdots, X_{ip}, Y_i),\ i = 1, \cdots, n$, are a random sample from the model

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon, \tag{1.5}$$

where ε_i is a random error. Usually we assume that ε_i's are uncorrelated random variables with mean 0 and common variance σ^2. We may take $X_{i1} = 1$ for all i so that there is a constant term in model (1.5).

The method of least squares was advanced early in the nineteenth century by Gauss. It is crucial in the area of regression analysis. To find a "good" estimate of the regression coefficients $\beta_1, \cdots, \beta_p$, we employ the method of least squares. For the observations $(\mathbf{x}_i, Y_i)$, the method of least squares considers the deviation of Y_i from its expected value for each case

$$Y_i - \sum_{j=1}^{p} X_{ij}\beta_j.$$

The method of least squares is to minimize the sum of the n squared deviations:

$$S(\beta_1, \cdots, \beta_p) = \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2. \tag{1.6}$$

One might directly minimize (1.5), but a succinct expression may be derived using matrix notation. Denote

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \cdots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then model (1.5) can be rewritten in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.7}$$

The matrix $\mathbf{X}$ is known as the *design matrix* and is of crucial importance to the whole theory of linear regression analysis. The $S(\boldsymbol{\beta})$ in (1.6) indeed is

$$S(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Differentiating $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, we obtain the *normal equations*

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}, \tag{1.8}$$

which are a fundamental starting point for the analysis of linear models. If $\mathbf{X}^T\mathbf{X}$ is invertible, equation (1.8) yields an unbiased estimator of $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

as

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\beta}.$$

It also follows that

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

which involves an unknown parameter σ^2. An unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2}{n-p}.$$

This can be proved by some straightforward matrix algebra computation.

Denote $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, the fitted values of $\mathbf{y}$. Then

$$\widehat{\sigma}^2 = \frac{||\mathbf{y} - \widehat{\mathbf{y}}||^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2.$$

Gauss-Markov Theorem shows that $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$.

Now consider the testing problem

$$H_0 : C\boldsymbol{\beta} = \mathbf{h} \quad \text{versus} \quad H_1 : C\boldsymbol{\beta} \neq \mathbf{h},$$

where C is a constant matrix and $\mathbf{h}$ is a constant vector. Under H_0, the method of least squares is to minimize $S(\boldsymbol{\beta})$ with constraints $C\boldsymbol{\beta} = \mathbf{h}$. This leads to

$$\widehat{\boldsymbol{\beta}}_0 = \boldsymbol{\beta} - (\mathbf{X}^T\mathbf{X})^{-1}C\{C(\mathbf{X}^T\mathbf{X})^{-1}C^T\}^{-1}(C\widehat{\boldsymbol{\beta}} - \mathbf{h}).$$

The residual sum of squares, denoted by RSS_1 and RSS_0 for under the full model and the null hypothesis, are

$$\mathrm{RSS}_1 = ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2,$$

and

$$\mathrm{RSS}_0 = ||\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_0||^2.$$

The F-test for H_0 is based on the decomposition

$$\frac{\mathrm{RSS}_0}{\sigma^2} = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\sigma^2} + \frac{\mathrm{RSS}_1}{\sigma^2} \tag{1.9}$$

When ε_i are independent and identically distributed $N(0, \sigma^2)$, the left hand side of (1.9) has a χ^2_{n-p+q} distribution, where q is the rank of C, and the second term on the right-hand side has a χ^2_{n-p} distribution, when H_0 is true. Furthermore,

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/q}{\mathrm{RSS}_1/(n-p)}$$

has an $F_{q,n-p}$ distribution. The null hypothesis H_0 is rejected if the test statistics on the left hand side is large.

This information may be summarized in the ANalysis Of VAriance (ANOVA) table which follows.

TABLE 1.1. Analysis of Variance for a Multiple Regression

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE
Full	$\mathrm{SSR}_1 \equiv \sum(Y_i - \overline{Y})^2 - \mathrm{RSS}_1$	$p-1$	$\mathrm{SSR}_1/(p-1)$
Reduced	$\mathrm{SSR}_0 \equiv \sum(Y_i - \overline{Y})^2 - \mathrm{RSS}_0$	$p-1-q$	$\mathrm{SSR}_0/(p-1-q)$
Difference	$\mathrm{RSS}_0 - \mathrm{RSS}_1$	q	$(\mathrm{RSS}_0 - \mathrm{RSS}_1)/q$
Residuals	RSS_1	$n-p$	$\mathrm{RSS}_1/(n-p)$
Total	$\sum(Y_i - \overline{Y})^2$	$n-1$	

Each problem is only a specific problem of this. But for simplified models, such as one way ANOVA and two way ANOVA models, simpler formulas can be obtained. Further, for these specific models, other techniques can be used.

1.3 Maximum Likelihood Point of View

Consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.10}$$

When the functional form of the probability distribution of the error term is specified, estimates of the regression coefficients and the unknown parameter σ^2 can be obtained by the method of maximum likelihood. The method of maximum likelihood was first proposed by C.F. Gauss in 1821. However, the principal is usually credited to R. A. Fisher, who first studied the properties of the approach (See Fisher,1922). Essentially, the method of maximum likelihood consists of finding those values of unknown parameters which is "most likely" to have produced the sampled data. Consider the joint density of the sampled data as a function of the parameters for fixed data $(\mathbf{x}_i, Y_i), i = 1, \cdots, n$. This function is called the likelihood function and denoted by $\ell(\boldsymbol{\beta}, \sigma)$, The maximum likelihood estimator is the maximizer $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ of the likelihood function $\ell(\boldsymbol{\beta}, \sigma)$. That is

$$\ell(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}) = \max_{\boldsymbol{\beta}, \sigma} \{\ell(\boldsymbol{\beta}, \sigma)\}.$$

Assume that the error $\boldsymbol{\varepsilon}$ is a n-dimensional normal distribution $N_n(\mathbf{0}, \sigma^2 I_n)$, where I_n is the $n \times n$ identity matrix. The conditional likelihood of $\mathbf{y}$, given $\mathbf{X}$, is

$$\ell(\boldsymbol{\beta}, \sigma) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2}{2\sigma^2}\right\}.$$

The maximum likelihood estimate for $\boldsymbol{\beta}$ is the same as the least squares estimate of $\boldsymbol{\beta}$. However, the maximum likelihood estimate

for σ^2 is

$$\widehat{\sigma}^2 = \frac{\mathrm{RSS}_1}{n}$$

which is a biased estimator for σ^2.

In general, a maximum likelihood estimate is more efficient than a least squares estimate when the model is correct. On the other hand, the maximum likelihood method has to know the whole distribution of $\boldsymbol{\varepsilon}$, and hence it is less robust. Under the normality assumption, the least squares estimator for $\boldsymbol{\beta}$ is equivalent to the maximum likelihood estimator of $\boldsymbol{\beta}$. It is well known that the maximum likelihood estimator for $\boldsymbol{\beta}$ is the asymptotically most efficient estimator and the minimum variance unbiased estimator.

Consider the general testing problem

$$H_0 : C\boldsymbol{\beta} = \mathbf{h} \quad \text{versus} \quad H_1 : C\boldsymbol{\beta} \neq \mathbf{h}.$$

Under the null hypothesis,

$$\max_{\boldsymbol{\beta},\sigma,C\boldsymbol{\beta}=\mathbf{h}} \ell(\boldsymbol{\beta},\sigma) = \max_{\sigma} \ell(\widehat{\boldsymbol{\beta}}_0,\sigma) = \ell(\widehat{\boldsymbol{\beta}}_0,\widehat{\sigma}_0)$$

where $\widehat{\sigma}_0 = \mathrm{RSS}_0/n$.

Thus the log-likelihood ratio test is

$$\begin{aligned} \lambda &= 2\{\log \ell(\widehat{\boldsymbol{\beta}},\widehat{\sigma}) - \log \ell(\widehat{\boldsymbol{\beta}}_0,\widehat{\sigma}_0)\} \\ &= 2\log(\mathrm{RSS}_0/\mathrm{RSS}_1) \end{aligned}$$

which tends to a χ^2_q distribution as $n \to \infty$.

The H_0 is rejected when $\lambda \geq c$, which is equivalent to $\mathrm{RSS}_0/\mathrm{RSS}_1 > c_1$. This implies that F test is equivalent to the maximum likelihood ratio test.

1.4 Weighted and Generalized Least Squares

Consider the model

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1.11}$$

where the errors ε_i are independent with mean 0. In many situations it may not be reasonable to assume that the variances of the errors are the same for all levels $\mathbf{x}_i$ of the independent variables. However, we may be able to characterize the dependence of $\text{var}(\varepsilon_i)$ on $\mathbf{x}_i$ at least up to a multiplicative constant. That is, $\text{var}(\varepsilon_i) = \sigma^2 v_i$. The estimation of the regression coefficients in model (1.11) could be done by using the least squares estimator in Section 1.2 with equal error variances. These estimators are still unbiased and consistent for model (1.11) with unequal error variances. But they are no longer have minimum variance.

Let $Y_i^* = v_i^{-1/2} Y_i$, $X_{ij}^* = v_i^{-1/2} X_{ij}$, $\varepsilon_i^* = v_i^{-1/2}\varepsilon_i$. Then

$$Y_i^* = \sum_{i=1}^{p} X_{ij}^*\beta_j + \varepsilon_i^* \tag{1.12}$$

with $\text{var}(\varepsilon_i^*) = \sigma^2$. Apply the method of least squares to model (1.12),

$$||\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T W (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $W = \text{diag}(v_1, \cdots, v_n)$ is the weight matrix. It follows that the above weighted least squares estimate is the BLUE for $\boldsymbol{\beta}$.

In general, assume that

$$\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 \Sigma,$$

in which Σ is a known constant matrix. Denote by $\Sigma^{-1/2}$ the square root of Σ^{-1}, i.e., $(\Sigma^{-1/2})^T\Sigma^{-1/2} = \Sigma^{-1}$. Then

$$\text{var}(\Sigma^{-1/2}\boldsymbol{\varepsilon}) = \sigma^2 I_n.$$

Denote

$$\mathbf{y}^* = \Sigma^{-1/2}\mathbf{y}, \quad \mathbf{X}^* = \Sigma^{-1/2}\mathbf{X}, \quad \boldsymbol{\varepsilon}^* = \Sigma^{-1/2}\boldsymbol{\varepsilon}.$$

Then model (1.11) reduces to

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*. \tag{1.13}$$

Thus

$$||\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\Sigma^{-1}(\mathbf{x} - \mathbf{X}\boldsymbol{\beta}).$$

The least squares estimate for $\boldsymbol{\beta}$ based on model (1.12) is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{y}^* = (X^T\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\Sigma^{-1}\mathbf{y})$$

which is a generalized least squares estimate for $\boldsymbol{\beta}$ based on model (1.11). Therefore $\widehat{\boldsymbol{\beta}}$ is the BLUE based on $\mathbf{y}$.

Misspecification of Σ in the generalized least squares still gives us a root n consistent unbiased estimator. To see this, assume that $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2\Sigma_0$.

$$\text{E}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

and

$$\text{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\Sigma^{-1}\Sigma_0\Sigma^{-1}\mathbf{X})(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1} = O(\frac{1}{n}).$$

The ordinary least squares uses $\Sigma = I_n$, and it still gives us an unbiased estimate and $n^{-1/2}$ consistent estimator. However, when Σ is close to Σ_0 leads to an unbiased estimator with somewhat better performance.

1.5 An Example

In this section, we will illustrate theory of linear regression models by analyses of the environmental data set introduced in Section 1.1. Throughout this book, S-plus will be used to analyze all examples. Readers may refer Chambers and Hastie (1993) for details.

Example 1.1 (Continued) The scatter plots of the response and the three air pollutants are depicted in Figure 1.1. From Figure 1.1, it seems that there exists linear relationship between the covariates of NO_2 and dust. S-plus codes and outputs are displayed in the end of this section. From the outputs, we have estimated coefficients and their estimated standard errors, listed in Table 1.2.

TABLE 1.2. Estimated coefficients and standard errors for Example 1.1

	$\widehat{\beta}_i$	SE	t-value	p-value
(Intercept)	234.8353	5.5249	42.5049	0.0000
SO_2	-0.4490	0.1623	-2.7673	0.0058
NO_2	0.5226	0.1600	3.2654	0.0011
dust	0.1121	0.1165	0.9626	0.3361

TABLE 1.3. Estimated coefficients and standard errors for Example 1.1

	Df	SS	MS	F- Value
SO_2	1	336	335.62	0.14386
NO_2	1	88725	88725.33	38.03038
dust	1	2162	2161.81	0.92662
Error	726	1693766	2333.01	

S-plus also provides us an ANOVA table, shown in Table 1.3. The values of sums of squares are dependent on the order of terms added. For example, if we reverse the adding order of NO_2 and dust, then

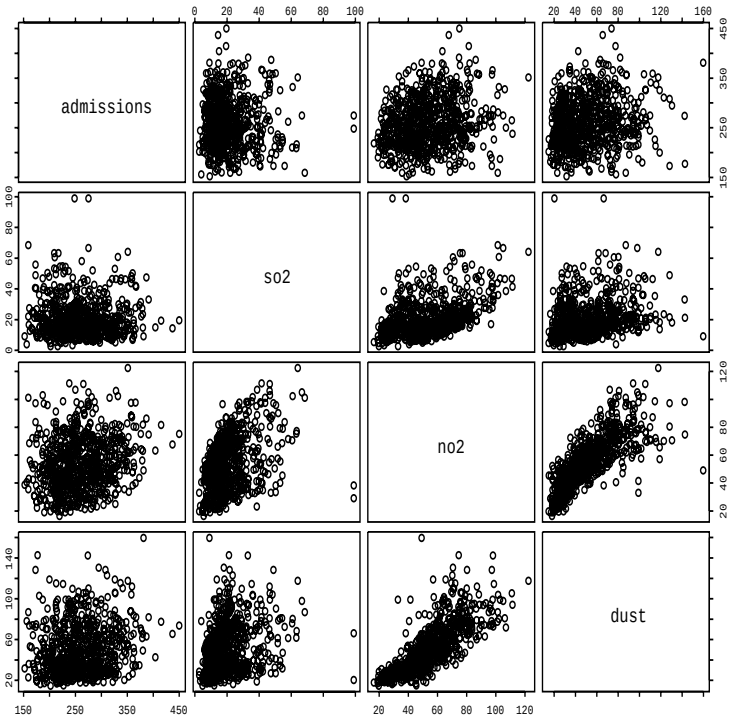

FIGURE 1.1. Pair-plot of the hospital admissions and air pollutants: SO_2, NO_2 and dust

TABLE 1.4. Estimated coefficients and standard errors for Example 1.1

	Df	SS	MS	F-value
SO_2	1	336	335.62	0.14386
dust	1	66011	66010.66	28.29418
NO_2	1	24876	24876.48	10.66282
Errors	726	1693766	2333.01	

the corresponding values of sum of squares is changed, see Table 1.4. We will explain this phenomenon later.

From Table 1.2, the p-value for t-test for the null hypothesis whether the coefficient of dust equals to zero is 0.3361. This indicates that the contribution of covariate dust is weak and one may delete it from

the model. This might be due to that the linear relationship between the covariates of NO_2 and dust. The multiple R^2 is only 0.0511. This implies that the three linear terms of air pollutants do not significant reduce the sum of squares. It suggests that more complicate models, such as model (1.6), should be considered. In fact, several authors have studied this data set using nonparametric regression techniques, see Fan and Zhang (1999) and Cai, Fan and Li (2000).

Appendix: S-plus codes for Example 1.1

```
>hk <- read.table("d:/rli/book/hkepd.dat",header=T)
>            # Read data in
>admissions <- hk[,15]
>    # set the response variable: hospital admissions
>so2 <- hk[,4]  # covariate: SO_2
>no2 <- hk[,5]  # covariate: NO_2
>dust <- hk[,6] # covariate: dust
>hosp <- data.frame(admissions,so2,no2,dust)
>    # create a data frame in Splus
>pairs(~admissions+so2+no2+dust) # pair-plots
>out <- lm(admissions~so2+no2+dust,hosp)
>     # use object lm()
>summary(out) # a summary


Call: lm(formula = admissions ~ so2 + no2 + dust, data = hosp)
Residuals:
    Min     1Q Median    3Q   Max
 -107.8 -32.63 -4.964 31.26 176.5

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  234.8353    5.5249    42.5049    0.0000
        so2   -0.4490    0.1623    -2.7673    0.0058
        no2    0.5226    0.1600     3.2654    0.0011
```

```
       dust    0.1121    0.1165     0.9626    0.3361

Residual standard error: 48.3 on 726 degrees of freedom
Multiple R-Squared: 0.05111
F-statistic: 13.03 on 3 and 726 degrees of freedom,
  the p-value is 2.678e-008

Correlation of Coefficients:
     (Intercept)     so2     no2
 so2 -0.1871
 no2 -0.5598     -0.3054
dust  0.0818      0.0591 -0.7613


>anova(out)  # Print ANOVA table


Analysis of Variance Table

Response: admissions

Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value     Pr(F)
      so2   1       336   335.62  0.14386 0.7045889
      no2   1     88725 88725.33 38.03038 0.0000000
     dust   1      2162  2161.81  0.92662 0.3360636
Residuals 726   1693766  2333.01


>out1 <- lm(admissions~so2+dust+no2,hosp)
>     # change the order of dust and no2
>anova(out)  # Print ANOVA table


Analysis of Variance Table

Response: admissions

Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value     Pr(F)
```

```
            so2   1       336   335.62  0.14386 0.7045889
            no2   1     88725 88725.33 38.03038 0.0000000
           dust   1      2162  2161.81  0.92662 0.3360636
Residuals 726   1693766  2333.01
```

2

One Way Analysis of Variance

In many scientific researches, there may be many factors having impact on the outcome of experiments. But there exists a very important factor, and the experimenters want to investigate its effect on response variable assuming that the other factors are fixed. In this chapter, we introduce analysis of one factor experiment, which is fundamental to analysis of experiment.

2.1 Complete Randomized Design

In many situations, we are interested in the dependence of the outcome variable upon the level of the treatment factor. For example, in dose range study, suppose that there are five levels of doses, A, B, C, D and E. The experimenter is interested in treatment effects. 33 mice are assigned at random into one of the five treatments.

Treatments	A	B	C	D	E	Total
No. of Mice	6	7	7	7	6	33

In this study, the variation of response consists of treatment variation and individual variation.

Let Y_{ij} be the response obtained on the jth observation of the ith treatment. Then one way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad j = 1, \cdots, n_i, \quad i = 1, \cdots, I, \tag{2.1}$$

where I is the number of treatments, and n_i is the number of observations to be taken on the ith treatment. In this model, μ is called *grand mean* , α_i *treatment effect* and ε_{ij} *individual variation* having mean 0. An alternative way of writing model (2.1) is to replace $\mu+\alpha_i$ by μ_i, so that the model becomes

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad j = 1, \cdots, n_i, \quad i = 1, \cdots, I. \tag{2.2}$$

The μ_i is the mean of the ith group. Model (2.1) contains $I + 1$ unknown parameters, while there are only I unknown parameters in model (2.2). In fact, the $I + 1$ unknown parameters $\mu, \alpha_1, \cdots, \alpha_I$ in model (2.1) are not identifiable. This means that there is no unique estimate of the $I + 1$ parameters in model (2.1). On the other hand, many experimenters prefer model (2.1) because μ and α_i have their own physical interpretation. The parameter μ is a constant, and usually is referred to the mean effect. The parameters α_i represents the positive or negative deviation of the response from the mean effect when the ith treatment is observed. Thus the deviation is called the "effect" on the response of the ith treatment.

To make the $I+1$ parameters in model (2.1) estimable, statisticians usually impose a constraint on the model. In many statistics textbooks, it is assumed that

$$\sum_{i=1}^{I} n_i \alpha_i = 0.$$

This reduces to $\sum_i \alpha_i = 0$ for balanced one way ANOVA models. Based on this assumption,

$$\mu = \bar{\mu}_\cdot \equiv \frac{1}{I} \sum_{i=1}^{I} \mu_i,$$

here, as usual, the subscript "$\cdot$" stands for "average", and

$$\alpha_i = \mu_i - \bar{\mu}_\cdot.$$

Under this assumption, α_i is the difference from the average treatment effect.

Different packages of statistical data analysis impose different constraints on the model. For example, in S-plus, it is assume that $\alpha_1 = 0$ with some option. Thus, μ is the effect of the first treatment and $\alpha_i = \mu_i - \mu_1$, the difference between the first treatment and the ith treatment. It is assumed that $\alpha_I = 0$ in SAS, so the constant μ is the effect of the I-th treatment, and $\alpha_i = \mu_i - \mu_I$, the difference between the ith treatment and the I-th treatment.

2.2 Estimation of Parameters

In this section, the method of least squares is used to estimate parameters in the one way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad j = 1, \cdots, n_i, \quad i = 1, \cdots, I, \tag{2.3}$$

where ε_{ij}'s are independent and identically distributed $N(0, \sigma^2)$. Denote $n = \sum_{i=1}^{I} n_i$. Here it is not required that $n_1 = \cdots = n_I$. This implies that we are studying an unbalanced one way ANOVA model, and balance one way ANOVA models are special cases thereof. Use matrix notation, the model can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{In_I} \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_I \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{In_I} \end{pmatrix}$$

In this model, the design matrix $\mathbf{X}$ is not full rank, so the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible, and the least squares estimator of $\boldsymbol{\beta}$ cannot be obtained via computing $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. To obtain the least

squares estimate for $\boldsymbol{\beta}$, the method of least squares is to minimize

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\mu-\alpha_i)^2. \tag{2.4}$$

Take derivative of (2.4) with respect to μ and α_i, it follows that the following $I+1$ equations

$$\overline{Y}_{\cdot\cdot}-\widehat{\mu}-\textstyle\sum_{i=1}^{I}(n_i/n)\widehat{\alpha}_i=0, \tag{2.5}$$

$$\overline{Y}_{i\cdot}-\widehat{\mu}-\widehat{\alpha}_i=0 \quad i=1,\cdots,I, \tag{2.6}$$

where $\overline{Y}_{\cdot\cdot}=\frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}Y_{ij}$. Among these $I+1$ equations, there are only I independent equations. Statisticians usually impose a constraint on the one way ANOVA model. A commonly used constraint is

$$\sum_{i=1}^{I}n_i\alpha_i=0.$$

Under this constraint,

$$\widehat{\mu}=\overline{Y}_{\cdot\cdot},$$

and

$$\widehat{\alpha}_i=\overline{Y}_{i\cdot}-\widehat{\mu}.$$

In S-plus, it is assumed that $\alpha_1=0$. Under this assumption, $\widehat{\mu}=\overline{Y}_{1\cdot}$, $\widehat{\alpha}_i=\overline{Y}_{i\cdot}-\overline{Y}_{1\cdot},\ i=2,\cdots,I$.

Though the parameters $\mu,\alpha_1,\cdots,\alpha_I$ are not identifiable, many parameters, such as $\alpha_1-\alpha_2$, are estimable without any constraints. Such a parameter is called a contrast. More generally, a contrast θ has the form

$$\theta=\sum_{i=1}^{I}c_i\alpha_i$$

with $\sum_{i=1}^{I} c_i = 0$ and can be estimated as

$$\widehat{\theta} = \sum_{i=1}^{I} c_i \widehat{\alpha}_i$$

which is independent of constraints.

The residual sum of squares for one way ANOVA model is

$$\text{RSS} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \widehat{\mu} - \widehat{\alpha}_i)^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2.$$

By simple algebra,

$$\text{RSS} \sim \sigma^2 \chi^2_{n-I}$$

and hence

$$\widehat{\sigma}^2 = \frac{\text{RSS}}{n - I}$$

is an unbiased estimator of σ^2.

Example 2.1 *(**Female Labor Supply**) This data set from East Germany was collected around 1994. It consists of 607 female salary. It is of interest to investigate the relationship between hourly earning and education level. This data set contains several related measurements. But here we take the hourly earning as the response variable and defined educational level as:*

- *Level A: more than 16 year education*
- *Level B: between 13 and 16 year education*
- *Level C: exact 12 year education*
- *Level D: less than 12 year education.*

Comparisons of means, medians and variances and box-plot for each treatment are depicted in Figure 2.1.

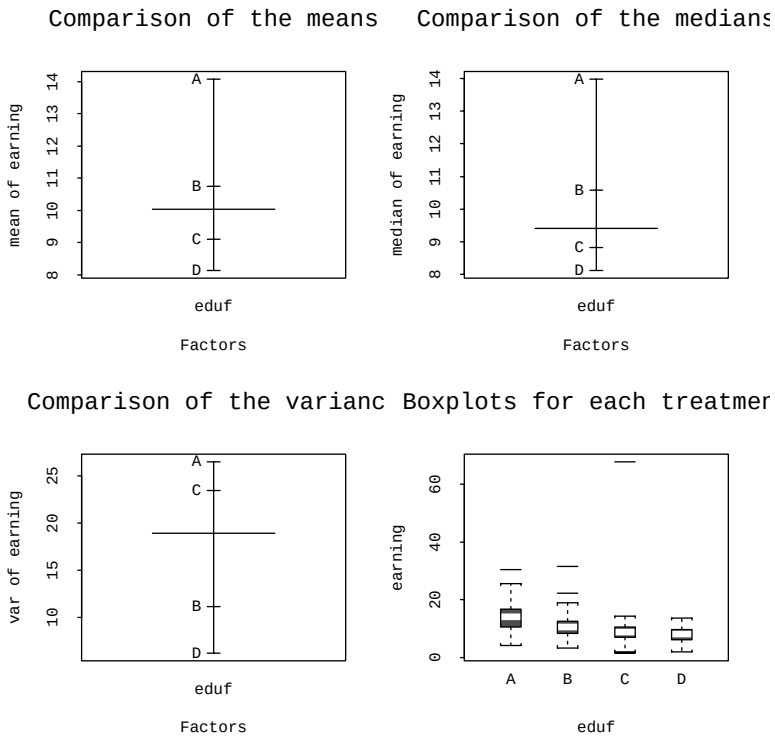

FIGURE 2.1. Comparisons of means, median and variances

The default contrast in S-plus is the Helmert polynomial. To compute the cell means easily, we set contrast to be treatment, in which it is assumed that $\alpha_1 = 0$.

(Intercept)	*edufB*	*edufC*	*edufD*
14.07384	*-3.328302*	*-4.959358*	*-5.938258*

From the output, we can compute the cell mean, such as, $\overline{Y}_{1\cdot} = 14.073$ *and* $\overline{Y}_{2\cdot} = 14.07 - 3.328$, *so on. Table 2.1 gives a summary of this data set.*

TABLE 2.1. Summary of Female Labor Supply Data

Level	Sample Size	$\overline{Y}_{i\cdot}$	Variance
A	70	14.07	26.507
B	207	10.75	11.177
B	200	9.11	23.459
C	130	8.14	6.248

If we impose the constraint $\sum_{i=1}^{4} n_i\alpha_i = 0$, *then we have*

$$\widehat{\mu} = \overline{Y}_{\cdot\cdot} = 10.52, \quad \widehat{\alpha}_1 = \overline{Y}_{1\cdot} - \overline{Y}_{\cdot\cdot} = 3.55,$$

$$\widehat{\alpha}_2 = \overline{Y}_{2\cdot} - \overline{Y}_{\cdot\cdot} = 0.23, \quad \widehat{\alpha}_3 = \overline{Y}_{3\cdot} - \overline{Y}_{\cdot\cdot} = -1.40,$$

$$\widehat{\alpha}_4 = \overline{Y}_{4\cdot} - \overline{Y}_{\cdot\cdot} = -2.38.$$

2.3 One Way ANOVA

In model (2.3), it is of interest to test whether or not there are treatment effects. Thus one may wish to test the hypothesis

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I$ versus $H_1 : \{$ at least two of the α_i's differ$\}$.

Although $\alpha_1, \cdots, \alpha_I$ are non-estimable parameters, the null hypothesis is equivalent to

$$H_0 : \alpha_2 - \alpha_1 = \cdots = \alpha_I - \alpha_1 = 0$$

which is a special linear hypothesis. Under the null hypothesis, one way ANOVA model reduces to

$$Y_{ij} = \mu + \varepsilon_{ij}.$$

Therefore the least square estimate for μ is

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{I}\sum_{i=1}^{n_i} Y_{ij} = \frac{1}{n}\sum_{i=1}^{I} n_i\overline{Y}_{i\cdot} = \overline{Y}_{\cdots}$$

Under H_0, the residual sum of squares

$$\begin{aligned} \mathrm{RSS}_0 &= \sum_i \sum_j (Y_{ij} - \overline{Y}_{..})^2 \\ &= \sum_i \sum_j (Y_{ij} - \overline{Y}_{i.})^2 + \sum_i n_i (\overline{Y}_{i.} - \overline{Y}_{..})^2. \end{aligned}$$

Under H_1, the residual sum of squares

$$\mathrm{RSS}_1 = \sum_i \sum_j (Y_{ij} - \overline{Y}_{i.})^2 \equiv \mathrm{SS}_\mathrm{E}.$$

Thus, the sum of the squares reduction due to treatment is

$$\mathrm{SS}_\mathrm{T} = \mathrm{RSS}_0 - \mathrm{RSS}_1 = \sum_i n_i (\overline{Y}_{i.} - \overline{Y}_{..})^2.$$

Hence one should reject H_0 at significance level α if

$$\frac{\mathrm{SS}_\mathrm{T}/(I-1)}{\mathrm{SS}_\mathrm{E}/(n-I)} \geq F_{I-1,\, n-I}(1-\alpha).$$

Denote F_{obs} the observed F statistic, then p-value is

$$p = P\{F_{I-1,\, n_I} > F_{\mathrm{obs}}\}.$$

The above result can conveniently summarized as one way ANOVA table in Table 2.2.

TABLE 2.2. Analysis of Variance for One Way ANOVA

Source	Sum Of Squares	D.F.	Mean Squares	F value
Treatment	SS_T	$I-1$	$\mathrm{SS}_\mathrm{T}/(I-1)$	$\frac{\mathrm{SS}_\mathrm{T}/(I-1)}{\mathrm{SS}_\mathrm{E}/(n-I)}$
Error	SS_E	$n-I$	$\mathrm{SS}_\mathrm{E}/(n-I)$	
Total	$\sum_i \sum_j (Y_{ij} - \overline{Y}_{..})^2$	$n-1$		

Example 2.1 (Continued) Now we investigate whether education makes any difference in hourly earning. Table 2.3 depicts the one way ANOVA table, summarized from S-plus outputs. The result shows highly statistical significant. This implies that education plays a role in hourly earning. A natural questions arises here is how important role the education plays. To answer this question, we can examine the treatment effect rather than the p-value. In this example, compared with the level A, the treatment effects are -3.33, -4.98 and -5.94 for levels B, C and D, respectively.

TABLE 2.3. ANOVA Table for Example 2.1

Source	SS	D.F.	MS	F value	p-value
Education	1884.8	3	628.3	39.44	0
Residuals	9605.6	603	15.9		
Total	11490.4	606			

2.4 Individual Contrasts

A function of the parameters of any model is said to be *estimable* if there exists a linear unbiased estimator $\widehat{\theta} = \sum_i \sum_j c_{ij} Y_{ij}$ of θ, that is $\mathrm{E}\,\widehat{\theta} = \theta$. For the one way ANOVA model,

$$\mathrm{E}\,\widehat{\theta} = \sum_i (\sum_j c_{ij})(\mu + \alpha_i) \equiv \sum_i c_i(\mu + \alpha_i).$$

Therefore a parameter $\theta = \sum_i c_i \alpha_i$ is estimable if and only if $\sum_i c_i = 0$. Any such function $\sum_i c_i \alpha_i$ for which $\sum_i c_i = 0$ is a contrast. For example, $\alpha_i - \alpha_{i'}$ is a contrast if $i \neq i'$.

Let $\theta = \sum_i c_i \alpha_i$ be a contrast, then θ can be rewritten as $\theta = \sum_i c_i(\mu + \alpha_i) = \sum_i c_i \mu_i$. Substitute μ_i by its estimator $\overline{Y}_{i\cdot}$. Then $\widehat{\theta} = \sum_i c_i \overline{Y}_{i\cdot}$ is the BLUE for θ and the MLE under the normal model.

Theorem 2.1 *(i) $\widehat{\theta}$ is BLUE for θ.*

(ii) If the random error ε_{ij}'s in model (2.1) are independent and identically distributed $N(0, \sigma^2)$, then

$$\widehat{\theta} \sim N(\theta, \sigma^2 \sum_{i=1}^{I} c_i^2/n_i).$$

Further $\widehat{\theta}$ is independent of $\widehat{\sigma}^2$.

Proof: (i) By straightforward algebra,

$$\mathrm{E}\,\widehat{\theta} = \sum_{i=1}^{n} c_i(\mu + \alpha_i) = \theta.$$

So $\widehat{\theta}$ is an unbiased estimator with variance

$$\mathrm{var}(\widehat{\theta}) = \sum_{i=1}^{n} c_i^2 \mathrm{var}(\overline{Y}_{i\cdot}) = \sum_{i=1}^{I} (c_i^2/n_i)\sigma^2.$$

For any linear unbiased estimator of θ,

$$\widehat{\theta}^* = \sum_{i=1}^{I} \sum_j d_{ij} Y_{ij},$$

we have

$$\mathrm{E}\,\widehat{\theta}^* = \sum_{i=1}^{I} \sum_j d_{ij}(\mu + \alpha_i).$$

Since $\widehat{\theta}^*$ is an unbiased estimator,

$$\sum_{i=1}^{I} (\sum_j d_{ij})\mu + \sum_{i=1}^{I} (\sum_j d_{ij})\alpha_i = \sum_{i=1}^{I} c_i \alpha_i.$$

This implies that $\sum_{i=1}^{I}\sum_j d_{ij} = 0$ and $c_i = \sum_j d_{ij}$. Thus $c_i^2 = (\sum_j d_{ij})^2$, so $c_i^2 \le n_i \sum_j d_{ij}^2$ by arithmetic-geometric inequality. Thus

$$\mathrm{var}(\widehat{\theta}^*) = \sum_{i=1}^{I}\sum_j d_{ij}^2\sigma^2 \ge \sum_{i=1}^{I}(c_i^2/n_i)\sigma^2 = \mathrm{var}(\widehat{\theta}).$$

This completes the proof of Part (i).

(ii) Since $\widehat{\theta}$ is a linear combination of Y_{ij}, if ε_{ij} is normally distributed, so is $\widehat{\theta}$. Thus

$$\widehat{\theta} \sim N(\theta, \sigma^2 \sum_{i=1}^{I} c_i^2/n_i)$$

by the proof of (i).

To show that $\widehat{\theta}$ is independent of $\widehat{\sigma}^2$. Note that $\widehat{\sigma}^2 = \frac{1}{n-I}\sum_i\sum_j(Y_{ij} - \overline{Y}_{i\cdot})^2$. It suffices to show that $\sum_j(Y_{ij} - \overline{Y}_{i\cdot})^2$ is independent of $\widehat{\theta} = \sum_i n_i\overline{Y}_{i\cdot}$. This follows obviously from the independence between the sample mean $\overline{Y}_{i\cdot}$ and sample variance $\sum_j(Y_{ij} - \overline{Y}_{i\cdot})^2$, since

$$\mathrm{cov}(\overline{Y}_{i\cdot}, Y_{ij} - \overline{Y}_{i\cdot}) = 0$$

by some algebra. This completes the proof of Part (ii).

By Theorem 2.1, it follows that

$$\frac{\widehat{\theta} - \theta}{\widehat{\sigma}\sqrt{\sum_i c_i^2/n_i}} \sim t_{n-I}.$$

Thus $(1-\alpha)100\%$ confidence interval is

$$\widehat{\theta} \pm t_{n-I}(1-\alpha/2)\widehat{\sigma}\sqrt{\sum_i c_i^2/n_i}.$$

For testing hypothesis $H_0 : \theta = \theta_0$, we use t-statistic

$$t_{\mathrm{obs}} = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}\sqrt{\sum_i c_i^2/n_i}}$$

and compute the corresponding p-value.

Example 2.1 (Continued) Let $\theta = \alpha_2 - \alpha_1$ be the difference in hourly earning between Education levels B and A. The θ is a contrast, and its estimate is

$$\widehat{\theta}_1 = -3.33 \qquad \text{and} \qquad \widehat{\sigma}^2 = 15.9.$$

Thus a 95% confidence interval for θ_1 is $-3.33 \pm 1.96 \times \sqrt{15.9} \times \sqrt{1/70 + 1/207} = -3.33 \pm 1.08$. Since the confidence interval does not contain 0, it is significant at level 0.05 to reject $H_0 : \theta = 0$, which is equivalent to a t-test. Let $\theta_2 = \frac{1}{3}(\alpha_2 + \alpha_3 + \alpha_4) - \alpha_1$. Then

$$\widehat{\theta}_2 = \frac{1}{3}(-3.33 - 4.96 - 5.94) = -4.74$$

with the standard error

$$\widehat{SE} = 3.99 \times \sqrt{\frac{1}{9}(\frac{1}{207} + \frac{1}{200} + \frac{1}{130}) + \frac{1}{70}} = 0.508.$$

Therefore a confidence interval for θ_2 is $-4.74 \pm 1.96 \times 0.508 = -4.74 \pm 1.00$.

In this section, we only studied individual contrasts. The result cannot be directly applied for multiple contrasts. For example, we cannot say that with 95% confidence that

$$\theta_1 \in -3.33 \pm 1.08 \qquad \text{and} \qquad \theta_2 \in -4.74 \pm 1.00.$$

We will study multiple comparison next section.

2.5 Methods of Multiple Comparisons

Suppose that a random interval S_j contains a parameter θ_j with probability $1-\alpha_j,\ j=1,\cdots,m$. Here θ_j may be a contrast or treatment mean. Frequently, we have to calculate the probability that all S_j occur simultaneously, or we have to determine the value of α_j such that the probability that all S_j occur simultaneously equal to $1-\alpha$ for some given α. This is a challenging problem in statistics, and many researches have been done to find ways around this problem. The developed techniques are known as *methods of multiple comparisons*, which are important concepts in the context of analysis of variance. In this section, we introduce three important multiple comparison methods: Bonferroni method, Scheffé method and Tukey method.

2.5.1 Bonferroni Method

Since $P\{\theta_j \in S_j\} = 1-\alpha_j$, thus simultaneously

$$\begin{aligned} & P\{\bigcap_{j=1}^{m}(\theta_j \in S_j)\} \\ = \ & 1 - P\{\bigcup_{j=1}^{m}(\theta_j \in \overline{S}_j)\} \\ \geq \ & 1 - \sum_{j=1}^{m} P\{\theta_j \in \overline{S}_j\} \\ \geq \ & 1 - \sum_{j=1}^{m} \alpha_j, \end{aligned}$$

where $\overline{S}_j$ is the complement of S_j. For given α, to construct a $100(1-\alpha)\%$ simultaneous confidence set, one has to choose appropriate α_i

such that $\sum_{j=1}^{m} \alpha_j = \alpha$. Then with probability $1 - \alpha$, the following events occur simultaneously

$$\theta_1 \in S_1, \cdots, \theta_m \in S_m.$$

Therefore $S_1, \cdots, S_m$ may be regarded as a $100(1-\alpha)\%$ simultaneous confidence set. In practice, one often takes $\alpha_1 = \cdots = \alpha_m = \alpha/m$. Therefore the larger m is, the wider the simultaneous confidence interval is. The probability inequality demonstrates that the Bonferroni method is conservative in that the width of the intervals can be unnecessarily wide.

Example 2.1 (Continued) Let $\theta_1 = \alpha_2 - \alpha_1$ and $\theta_2 = \frac{1}{3}(\alpha_2 + \alpha_3 + \alpha_4) - \alpha_1$. Now we construct a Bonferroni simultaneous confidence interval. Take $m = 2$, $\alpha = 5\%$. Since the degrees of freedom for residuals is 603, t-distribution with 603 degrees of freedom is well approximated by the standard normal distribution. Thus $z_{1-\alpha/4} = z_{0.9875} = 2.241$. A 95% Bonferroni simultaneous confidence intervals for $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are

$$-3.33 \pm 2.241 \times 0.552 = -3.33 \pm 1.24$$
$$-4.74 \pm 2.241 \times 0.508 = -4.74 \pm 1.14,$$

respectively.

Note that if the confidence set are independent, then

$$P\{\theta_1 \in S_1, \cdots, \theta_m \in S_m\} = \prod_{j=1}^{m}(1 - \alpha_j) \approx 1 - \sum_{j=1}^{m} \alpha_j$$

when α_j's are small. So the Bonferroni method is about right.

2.5.2 Scheffé method

The main drawback of the Bonferroni method of multiple comparisons is that the confidence intervals can become very wide if m is large. The Scheffé method can repair this drawback.

For any contrast, we have

$$\widehat{\theta} - \theta = \sum_{i=1}^{I} c_i(\overline{Y}_{i\cdot} - \alpha_i) = \sum_{i=1}^{I} c_i\{(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}) - (\alpha_i - \mu)\}$$

as $\sum_i c_i = 0$. Denoted by $\widehat{\tau}_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$ and $\tau_i = \alpha_i - \mu$. Then

$$\widehat{\theta} - \theta = \sum_{i=1}^{I} c_i(\widehat{\tau}_i - \tau_i).$$

By the Cauchy-Schwartz inequality

$$|\widehat{\theta} - \theta| \le \sqrt{\sum_{i=1}^{I} c_i^2/n_i} \cdot \sqrt{\sum_{i=1}^{I} n_i(\widehat{\tau}_i - \tau_i)^2}.$$

Thus for all $\mathbf{c} = (c_1, \cdots, c_I)$,

$$\left| \frac{\widehat{\theta} - \theta}{\widehat{\sigma}\sqrt{\sum_{i=1}^{I} c_i^2/n_i}} \right| \le \sqrt{\frac{\sum_{i=1}^{I} n_i(\widehat{\tau}_i - \tau_i)^2}{\widehat{\sigma}^2}}.$$

Note that $\sum_{i=1}^{I} n_i(\widehat{\tau}_i - \tau_i)^2 = \mathrm{RSS}_0 - \mathrm{RSS}_1$. Thus

$$\frac{\sum_{i=1}^{I} n_i(\widehat{\tau}_i - \tau)^2/(I-1)}{\widehat{\sigma}^2} \sim F_{I-1,n-I}.$$

Thus with probability $1 - \alpha$,

$$\frac{|\widehat{\theta} - \theta|}{\widehat{\sigma}\sqrt{\sum_{i=1}^{I} c_i^2/n_i}} \le \sqrt{(I-1)F_{I-1,n-I}(1-\alpha)}$$

for all $\mathbf{c}$. This is Scheffé simultaneous confidence interval.

Example 2.1 (Continued) Here we are interested in constructing Scheffé simultaneous confidence intervals for all contrasts. In this example, $I = 4$, and $n - I = 603$. So

$$F_{I-1,n-I}(1-\alpha) \approx \frac{\chi^2_{I-1}(1-\alpha)}{I-1} = \frac{1}{3}\chi^2_3(0.95) = 7.815/3.$$

For all $\mathbf{c}$, it holds that

$$\left| \frac{\sum_{i=1}^4 c_i(\overline{Y}_{i\cdot} - \alpha_i)}{\sqrt{\sum c_i^2/n_i}\widehat{\sigma}^2} \right| \leq \sqrt{7.815} = 2.796.$$

In particular, a 95% simultaneous confidence interval is

θ_1: $-3.33 \pm 2.796 * 0.552$

θ_2: $-4.74 \pm 2.796 * 0.508$

This is wider than Bonferron simultaneous confidence intervals because m here is small. Note that the Scheffé confidence interval holds for all possible contrasts.

2.5.3 Tukey Method

Scheffé method can be applied for any m estimable contrasts or functions of the parameters and gives shorter intervals than Bonferroni method if m is large. In practice, experimenters are interested in all pairwise contrasts $\alpha_i - \alpha_j,\ i \neq j,\ i, j = 1, \cdots, I$. In this situation, Tukey method gives the best solution.

Assume that $n_i = n$ for all i. Note that the normalized form of the estimator $\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}$

$$\frac{(\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}) - (\alpha_i - \alpha_j)}{\widehat{\sigma}\sqrt{2}/\sqrt{n}}$$

is independent of unknown parameters, and

$$\frac{(\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}) - (\alpha_i - \alpha_j)}{\widehat{\sigma}\sqrt{2}/\sqrt{n}} \leq \frac{\max_i(\overline{Y}_{i\cdot} - \alpha_i - \mu) - \min_i(\overline{Y}_{i\cdot} - \alpha_i - \mu)}{\widehat{\sigma}\sqrt{2}/\sqrt{n}}.$$

Denote

$$Q = \frac{\max_i(\overline{Y}_{i\cdot} - \alpha_i - \mu) - \min_i(\overline{Y}_{i\cdot} - \alpha_i - \mu)}{\widehat{\sigma}\sqrt{2}/\sqrt{n}}.$$

Note that $\sqrt{n}(\overline{Y}_{i\cdot} - \alpha_i - \mu)$ are independent and identically distributed $N(0, \sigma^2)$. The distribution of Q is called the Studentized range distribution with degree of freedom $(I, n - I)$. Let $q_{I,n-I}(1-\alpha)$ be α-upper quantile of Q. Then $1 - \alpha$ simultaneous confidence interval for all pairwise contrasts

$$\left| \frac{(\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}) - (\alpha_i - \alpha_j)}{\sqrt{2}\widehat{\sigma}} \right| \leq q_{I,n-I}(1-\alpha)/\sqrt{2};$$

namely

$$\alpha_i - \alpha_j \in \{\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot} \pm \widehat{\sigma} q_{I,n-I}(1-\alpha)/\sqrt{n}\}.$$

The above idea readily applies to unbalanced design. The distribution of Q will depend on $n_1, \cdots, n_I$, but independent of unknown parameters. See Hockberg and Tamhance (1987) for details. Critical values of $q_{I,n-I}$ can be easily simulated by Monte Carlo method. Most of the statistical softwares, such as S-plus and SAS, provide, the upper percentiles of Q. A upper percentils table of Q can also be found in Dean and Voss (1999).

Example 2.2 *(Comparisons among Bonferroni, Scheffé and Tukey methods)*

Suppose that $I = 5$, $n = 35$, $\alpha = 5\%$, *and the design is balanced. It is of interest to compare the width of simultaneous confidence in-*

tervals for all pairwise comparisons: $\alpha_i - \alpha_j$ *for* $i \neq j$*. Thus we have totally 10 pairs. The coefficient in the confidence interval*

$$\textit{estimated coefficient} \pm w \cdot \widehat{SE}$$

has the following values:

Bonferroni: $w_B = t_{30}(1 - 0.05/20) = 3.02$,

Scheffé: $w_S = \sqrt{4F_{4,30}(0.95)} = 3.28$,

Tukey: $w_T = \frac{1}{\sqrt{2}} q_{5,30}(0.95) = 2.91$.

Therefore, Tukey method gives us the best solution.

2.6 Model diagnostic

In the one way ANOVA model, it is assumed that

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where ε_{ij} are independent and identically distributed $N(0, \sigma^2)$. The estimates and tests derived in previous sections are obtained as if the model and assumptions are correct. In many practical problems the assumptions are in doubt. Therefore, we have to check model assumptions after obtaining the estimates. Thus, when we fit a data set by the one way ANOVA model, we have to check the following three assumptions:

1. $\{\varepsilon_{ij}\}_{j=1}^{n_i}$ are a random sample. In other words, there are no outliers among the data.

2. ε_{ij} are normally distributed.

3. The variances of ε_{ij} are constant.

Many statistical techniques, including graphical display, have been developed to check the assumptions for general linear regression models, and formed a branch of statistics, known as *model diagnostic*. In the section, we will introduce some useful model diagnostic methods for the one way ANOVA model.

2.6.1 Outlier detection

One important assumption made in the one way ANOVA model is that the model is appropriate for all of the data. In practice, it is common for one or more cases to have an observation response that does not seem to correspond to the model fitted to the data. Cases that do not follow the same model as the rest of the data are called *outliers*, and one important function of case analysis is to identify such cases. There is a vast literature on the methods for handling outliers, including two books: Barnett and Lewis (1978) and Hawkins (1981). For the one way ANOVA model, graphical tools and visual inspections of residuals are very useful for detecting outliers. For the one way ANOVA model, the residuals are

$$\widehat{\varepsilon}_{ij} = Y_{ij} - \overline{Y}_{i\cdot}, \qquad j = 1, \cdots, n_i.$$

A common graphical display is the scatter plot of the residuals against the level i. Figure 2.2 depicts the scatter plot of residuals

against the level i for the data set studied in Example 2.1. It can be seen from Figure 2.2 that there is an outlier in this data set.

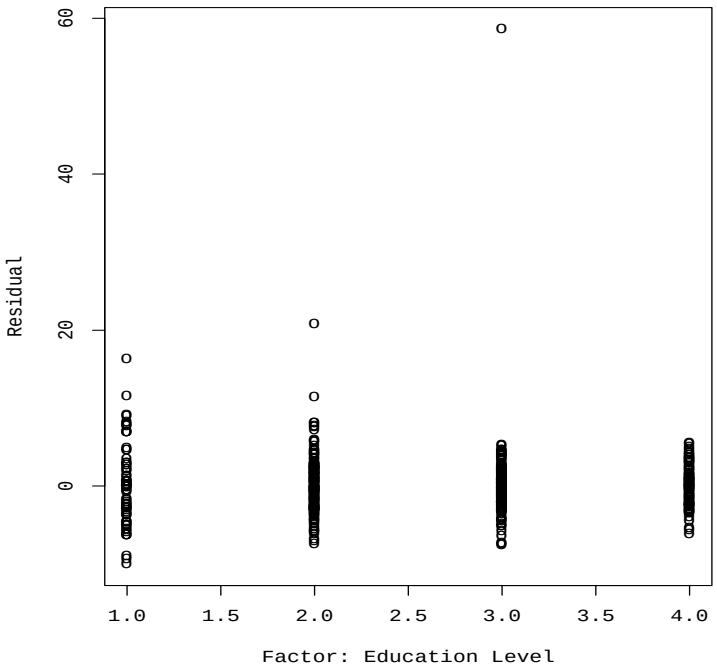

FIGURE 2.2. Scatter plot of residuals against the education levels: 1 stands for A, 2 for B, so on.

Box-plot contains more information. The box-plot not only display outliers but also the interquartiles for each level. As the interquartiles dividing by 0.6745 may be regarded as a robust estimator of standard deviation, Thus the box-plot may be used to check the first and the third assumptions at the same time. Figure 2.3 displays the boxplot of residuals for the labor supply data set in Example 2.1. From Figure 2.3, it can be seen that there are a few outliers in this data set.

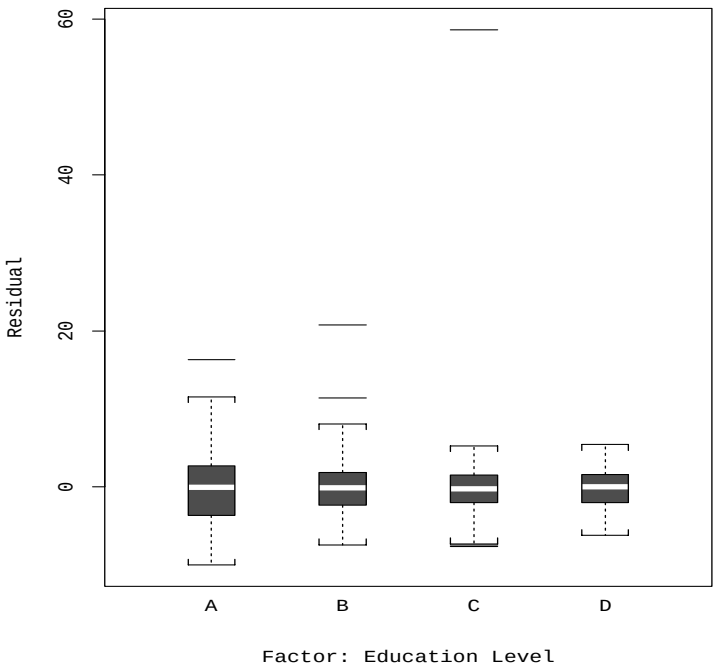

FIGURE 2.3. Boxplot of residuals against the education levels

The scatter plot of residuals against index is used to detect independence among the data. Figure 2.4 depicts the scatter plot of residuals against index. We did not find unusual pattern except a few outliers.

2.6.2 Testing of normality

Quantile to quantile (Q-Q) plot may be employed for test of normality. When the sample size in each group is small, we plot the residuals against the quantiles of standard normal distribution $\Phi^{-1}\{(i - 0.375)/(n + 0.25)\}$, denoted by ξ_i, $i = 1, \cdots, n$. When the sample

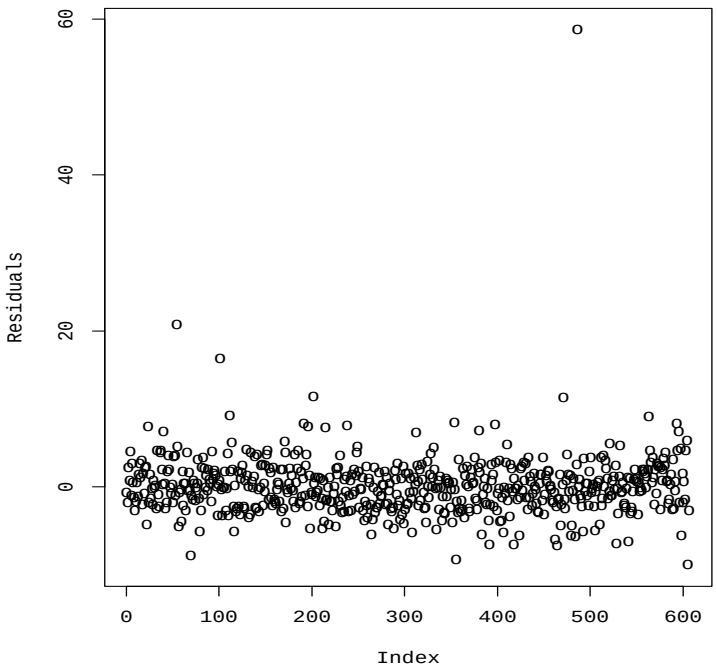

FIGURE 2.4. Boxplot of residuals against index

size is large, we may directly plot ε_{ij} against the quantiles of standard normal distribution. Figure 2.5 depicts the normal Q-Q plot of residuals in Example 2.1.

When the sample size in each group is large, we may estimate the density of Y's for each group by using nonparametric kernel estimation

$$\widehat{f}_{i,h}(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} K\left(\frac{Y_{ij} - x}{h}\right) /h,$$

where $K(x)$ may be chosen as Gaussian density and the bandwidth h may be taken as $1.06\widehat{\sigma} n^{-1/5}$. A estimated density function of the residuals in Example 2.1 is depicted in Figure 2.6. The density curve

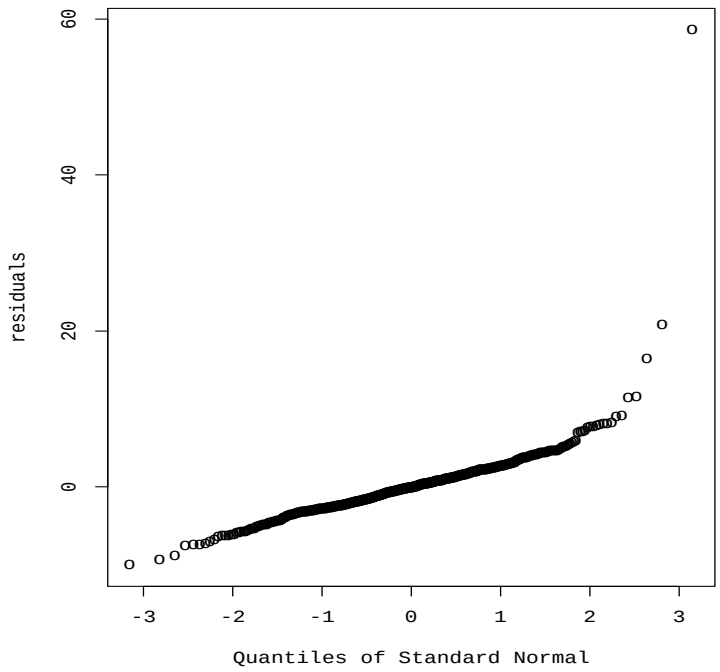

FIGURE 2.5. Q-Q plot of residuals

shows that the distribution of random errors looks like a normal distribution after excluding the outliers.

2.6.3 Equal variance

In the one way ANOVA model, it is assumed that $\text{var}(\varepsilon_{ij}) = \sigma^2$ for all cases in the data. This assumption is in doubt in many problems, as variance can depend on the response, or on the factor, or other factors, such as time of physical ordering. If nonconstant variance is diagnosed, but exact variance are unknown, we may use weighted least squares, with empirically chosen weights.

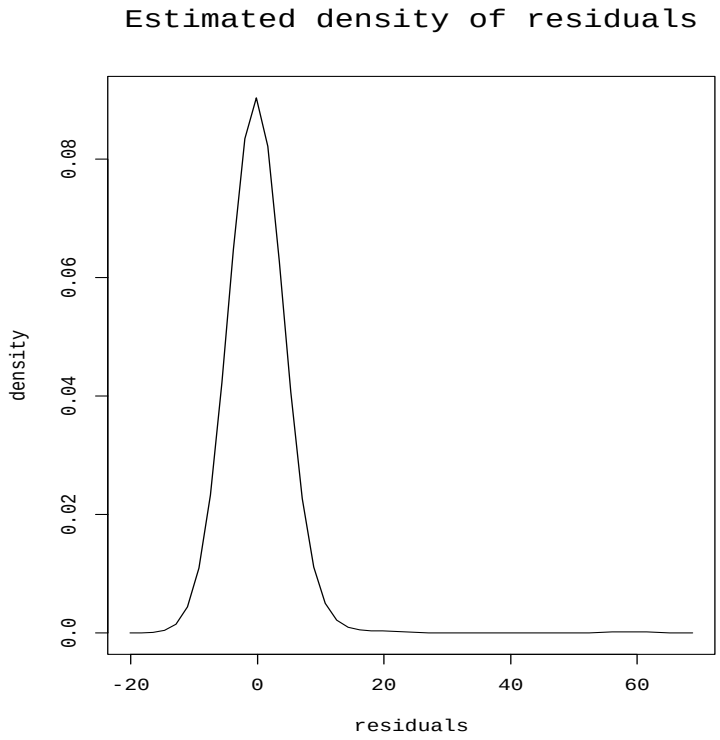

FIGURE 2.6. Estimated density function of residuals

When the sample size of each group is moderate, we may compute the sample variance for each group and visually inspect the difference.

Consider the nonconstant variance one way ANOVA model,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

with $\text{var}(\varepsilon_{ij}) = \sigma_i^2$. Testing equal variance is equivalent to testing the hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$$

based on the assumption that Y_{ij} are independent and identically distributed $N(\mu_i, \sigma_i^2)$. Likelihood ratio test can be constructed for testing the null hypothesis. It is easy to show that the maximum

likelihood estimate for $(\mu_i,\ \sigma_i^2)$ is $(\overline{Y}_{i\cdot}, \sum_j (Y_{ij} - \overline{Y}_{i\cdot})^2/n_i)$. Under the null hypothesis H_0, the maximum likelihood estimate for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2.$$

Therefore the logarithm of likelihood ratio test statistic is

$$T = \sum_{i=1}^{I} n_i \log\left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_i^2}\right)$$

which has a limiting distribution χ^2_{I-1} if $\min_i n_i \to \infty$. Therefore the rejection region is $\{T \geq \chi^2_{I-1}(1-\alpha)\}$ at significance level α.

Example 2.1 (Continued)

$$\begin{aligned} T = \quad & 70\log\frac{15.930}{26.507} + 207\log\frac{15.930}{11.177} \\ & +200\log\frac{15.930}{23.459} + 130\log 15.9306.248 = 81.747 \end{aligned}$$

with 3 degrees of freedom. We reject H_0. This suggests us to fit the data using weight least squares approach with weight for i-th group $\widehat{\sigma}_i^{-2}$.

2.7 S-plus codes for Example 2.1

```
>
>  labor <- read.table("d:/rli/book/labor.dat",header=T)
>  earning <- labor[,3]
>  edu <- labor[,5]
>  eduf <- rep("A",length(edu))
>  eduf[edu<16]<-"B"
```

```
>  eduf[edu<13]<-"C"
>  eduf[edu<12]<-"D"
>  laborf <-data.frame(earning,eduf)
>  postscript("d:/rli/book/figsa/ch2fig1.ps",width=4,
+      height=4,horizontal=F,pointsize=8)
>  par(mfrow=c(2,2))
>  plot.design(laborf)
>  title("Comparison of the means")
>  plot.design(laborf,fun=median)
>  title("Comparison of the medians")
>  plot.design(laborf,fun=var)
>  title("Comparison of the variance")
>  plot.factor(laborf)
>  title("Boxplots for each treatment")
>  dev.off()
 graphsheet
          2
>
>    # To fit the data by a linear model
>  lm(earning~eduf,laborf)
Call:
lm(formula = earning ~ eduf, data = laborf)

Coefficients:
 (Intercept)     eduf1     eduf2      eduf3
    10.51736 -1.664151 -1.098402 -0.7939263

Degrees of freedom: 607 total; 603 residual
Residual standard error: 3.991209
>
>  options()$contrast
          factor      ordered
 "contr.helmert" "contr.poly"
>
>   ## change default contrast
>  options(contrasts=c("contr.treatment"))
>  lm(earning~eduf)
Call:
lm(formula = earning ~ eduf)
```

```
Coefficients:
 (Intercept)     edufB     edufC     edufD
    14.07384 -3.328302 -4.959358 -5.938258

Degrees of freedom: 607 total; 603 residual
Residual standard error: 3.991209
>
>  options(contrasts=c("contr.sum"))
>  lm(earning~eduf)
Call:
lm(formula = earning ~ eduf)

Coefficients:
 (Intercept)    eduf1     eduf2     eduf3
    10.51736 3.556479 0.2281777 -1.402878

Degrees of freedom: 607 total; 603 residual
Residual standard error: 3.991209
>    ## To get an ANOVA table  ##
>  aov.labor <- aov(earning~eduf)
>  summary(aov.labor)
            Df Sum of Sq  Mean Sq  F Value Pr(F)
     eduf   3  1884.844 628.2813 39.44076     0
Residuals 603  9605.637  15.9297
>
>   ## To make a histogram, density and Q-Q plot ##
>  postscript("d:/rli/book/figsa/ch2fig2.ps",width=4,
+    height=5,horizontal=F,pointsize=8)
>  residuals <- resid(aov.labor)
>  edua <- rep(1,length(edu))
>  edua[edu<16]<-2
>  edua[edu<13]<-3
>  edua[edu<12]<-4
>  plot(edua,residuals,xlab='Factor: Education Level',
+  ylab='Residual')
>  dev.off()
 graphsheet
          2
>
>  postscript("d:/rli/book/figsa/ch2fig3.ps",width=4,
```

```
+    height=5,horizontal=F,pointsize=8)
>  labora <-data.frame(residuals,eduf)
>  plot.factor(labora,xlab='Factor: Education Level',
+  ylab='Residual')
>  dev.off()
 graphsheet
          2
>
>  postscript("d:/rli/book/figsa/ch2fig4.ps",width=4,
+    height=5,horizontal=F,pointsize=8)
>  qqnorm(residuals)
>  dev.off()
 graphsheet
          2
>
>  postscript("d:/rli/book/figsa/ch2fig5.ps",width=4,
+    height=5,horizontal=F,pointsize=8)
>  plot(density(residuals),type="l",xlab='residuals',
+  ylab='density')
>  title("Estimated density of residuals")
>  dev.off()
 graphsheet
          2
>
>  postscript("d:/rli/book/figsa/ch2fig6.ps",width=4,
+    height=5,horizontal=F,pointsize=8)
>  plot(residuals,xlab='Index',ylab='Residuals')
>  dev.off()
>
```

3

Two-way Layout Models

In the last chapter, we studied one-way ANOVA model in details. The one-way ANOVA model is appropriate when there exists a very important factor among many possible impacting factors. In practice, there may have many factors simultaneously playing important roles. In this situation, a multi-way layout model should be used. In this chapter, we studied two-way layout model. The techniques and methods introduced in this chapter can be directly extended to multi-way ANOVA models.

3.1 Two-way Layout Models

Example 3.1 *(Battery Experiment) In this example, we are interested in which type of non-rechargeable battery was the most economical. There are two treatment factors: duty and brand, each having*

TABLE 3.1. Data for Example 3.1

	Brand							
Duty	Name				Store			
Alkaline	611	537	542	593	923	794	827	898
Heavy	445	490	384	413	476	569	480	460

two levels. The response variable is the life per unit cost (min/dollar) . The data, extracted from Dean and Voss (1999), are listed in Table 3.1.

This example is a balanced design as the number of observations in each cell are the same. When the number of experiments in each cell are not the same, the corresponding experiment is called unbalanced design.

Statisticians usually label the treatment factors as A, B, and so on. In Example 3.1, every level of A is observed with every level of B, so the factor are crossed. In the analysis of two-way layout model, it is interested to check whether or not there is an interaction effect between the two treatment factors. Very frequently, some explanatory analysis can be done, such as interaction plot. An interactive plot may be constructed as follows: for each level of one treatment, plot the corresponding cell mean of the response against the levels of the other treatment and link them by line. Figure 3.1 depicts an interaction plot of *Battery Experiment* data. From Figure 3.1, The two lines are not parallel. This implies that there may be an interaction effect between *Duty* and *Brand*.

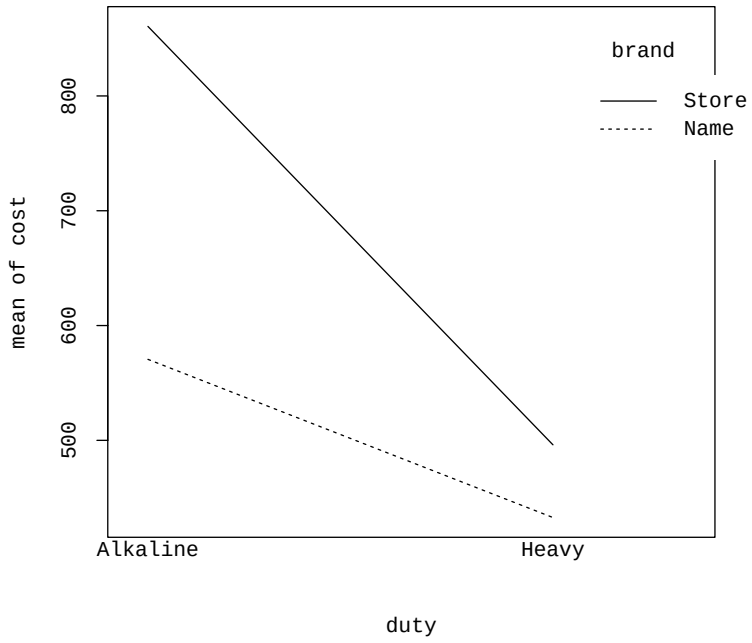

FIGURE 3.1. Interaction plot of battery experiments

One may examine whether or not the interaction effect is statistically significant. To this end, we have to fit the data by a two-way layout model. A two-way complete model is defined as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{3.1}$$

where μ is the grand mean, α_i and β_j are main effects corresponding factors A and B, γ_{ij} is the interaction between the factors A and B, and ε_{ijk} are individual variation and usually it is assumed that ε_{ijk}'s are independent and identically distributed $N(0, \sigma^2)$, which implies that the underlying model is a homoscedastic model. Occasionally, an experimenter has sufficient knowledge about the two treatment

factors being studying to state with reasonable certainty that the factors do not interact and that the lines in an interaction plot are parallel. This knowledge may be gleaned from previous similar experiments or from scientific facts about the treatment factors. If this is so, then the interaction term can be dropped from the two-way complete model, which reduces to the main effect model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \tag{3.2}$$

with ε_{ijk}'s are independent and identically distributed $N(0, \sigma^2)$. The main effect model is also called as two-way additive model, since the effect on the response of treatment combination is modeled as sum of the individual effects of the two factors. If an additive model is used when the factors really do interact, then inferences on main effects can be very misleading. Consequently, if the experimenter does not have reasonable knowledge about the interaction, then two-way complete model should be used.

3.2 Contrasts for Main Effects and Interactions

The two-way complete model can be rewritten as

$$Y_{ijk} = \mu + \tau_{ij} + \varepsilon_{ijk}$$

for $k = 1, \cdots, n_{ij}$, $i = 1, \cdots, I$, and $j = 1, \cdots, J$. By treating different combinations of A and B into one single factor with levels $\{(i, j) : i = 1, \cdots, I, j = 1, \cdots, J\}$, the model is equivalent to a one factor model. For example, the *battery experiment* can be regarded as

a one-factor experiment with 4 levels: Alkaline Name, Alkaline Store, Heavy Name and Heavy Store. Therefore the techniques for one-way layout model may continue to apply.

Theorem 3.1 *(i) A parameter θ is estimable if and only if $\theta = \sum_i \sum_j d_{ij}\tau_{ij}$, and $\sum_i \sum_j d_{ij} = 0$.*

(ii) The least squares estimate of θ is $\widehat{\theta} = \sum_i \sum_j d_{ij}\overline{Y}_{ij\cdot}$ and $var(\widehat{\theta}) = \sum_i \sum_j \frac{d_{ij}^2}{n_{ij}}\sigma^2$.

(iii) $\widehat{\sigma}^2 = \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij\cdot})^2 \sim \sigma^2 \chi^2_{n-IJ}$, where $n = \sum_i \sum_j n_{ij}$.

(iv) A $(1-\alpha)100\%$ confidence interval for θ is

$$\widehat{\theta} \pm t_{n-IJ}(1-\alpha/2)\sqrt{\sum_i \sum_j d_{ij}^2/n_{ij}}\widehat{\sigma}.$$

In the two-way complete model, there are many kinds of contrasts.

(i) Treatment contrasts: $\tau_{ij} - \tau_{kl}$ for $i \neq k$ and $j \neq l$.

(ii) Interaction contrasts:

$$(\tau_{ij} - \tau_{(i+1)j}) - (\tau_{ik} - \tau_{(i+1)k}) = (\gamma_{ij} - \gamma_{(i+1)j}) - (\gamma_{ik} - \gamma_{(i+1)k})$$

In general, interaction contrasts are always of form

$$\sum_i \sum_j d_{ij}\tau_{ij}$$

which is equal to $\sum_i \sum_j d_{ij}\gamma_{ij}$ with $\sum_i d_{ij} = 0$, for $j = 1, \cdots, J$ and $\sum_j d_{ij} = 0$ for $i = 1, \cdots, I$.

(iii) Contrasts of level B for each level of A: $\sum_j c_j \tau_{ij}$, with $\sum_j c_j = 0$, $i = 1, \cdots, I$. For example, $\tau_{ij} - \tau_{ik}$.

(iv) When interaction exists, main effect is hard to define. For convenience, we define main effect as

$$\alpha_i^* = \bar{\tau}_{i\cdot} = \alpha_i + \bar{\gamma}_{i\cdot}$$

for $i = 1, \cdots, I$. And a main effect contrast in A is

$$\sum_{i=1}^{I} c_i \alpha_i^*$$

with $\sum_{i=1}^{I} c_i = 0$, which is estimable.

Example 3.1 (*Continued*) It is easy obtained that $\overline{Y}_{11\cdot} = 570.75$, $\overline{Y}_{12\cdot} = 860.50$, $\overline{Y}_{21\cdot} = 433.00$, $\overline{Y}_{22\cdot} = 496.25$, $\mathrm{RSS} = \mathrm{SS}_E = 28412.52$ and $\widehat{\sigma} = \sqrt{\mathrm{SS}_E/12} = 48.66$.

Thus interaction contrast:

$$\theta_1 = \frac{1}{2}\{(\tau_{11} - \tau_{12}) - (\tau_{21} - \tau_{22})\} = \frac{1}{2}\{(\gamma_{11} - \gamma_{12}) - (\gamma_{21} - \gamma_{22})\}$$

has the least squares estimate

$$\widehat{\theta}_1 = -113.25$$

with estimated standard error

$$\widehat{\mathrm{SE}} = \sqrt{1/4 + 1/4 + 1/4 + 1/4}/2\widehat{\sigma} = \frac{\widehat{\sigma}}{2} = 24.33.$$

The 95% confidence interval for the interact contrast θ is

$$-113.25 \pm t_{12}(0.975) \times 24.33 = -113.25 \pm 53.02$$

which does not cover the origin. So we reject the null hypothesis $H_0 : \theta = 0$. Namely, there is an interaction between *Duty* and *Brand*. In fact, the observed t-statistic

$$t_{obs} = \frac{-113.25}{24.33}$$

with p-value 5.556×10^{-4} for testing the hypothesis

$$H_0 : \theta_1 = 0 \quad \text{versus} \quad H_1 : \theta_1 \neq 0.$$

The duty contrast (*Alkaline-Heavy duty*)

$$\theta_2 = \alpha_1^* - \alpha_2^* = \frac{1}{2}(\tau_{11} + \tau_{12} - \tau_{21} - \tau_{22})$$

has the least-squares estimate

$$\widehat{\theta}_2 = \overline{Y}_{1\cdot\cdot} - \overline{Y}_{2\cdot\cdot} = 251.00$$

and associated variance

$$\text{var}(\widehat{\theta}_2) = \text{var}(\overline{Y}_{1\cdot\cdot}) + \text{var}(\overline{Y}_{2\cdot\cdot}) = \frac{\sigma^2}{4}.$$

Hence

$$\widehat{\text{SE}}(\widehat{\theta}_2) = \frac{\widehat{\sigma}}{2} = 24.33.$$

Thus a 95% confidence interval for the duty contrast is $251 \pm 2.1799 \times 24.33$.

For one-sided test:

$$H_0 : \theta_2 = 0 \quad \text{versus} \quad H_1 : \theta_2 > 0.$$

The observed t-statistic is

$$t_{obs} = \frac{251}{24.33} = 10.32$$

with p-value 1.273×10^{-7}.

Similarly, the brand contrast (*Name-Store*) is

$$\theta_3 = \beta_1^* - \beta_2^*$$

has the least squares estimate

$$\widehat{\theta}_3 = \overline{Y}_{\cdot 1\cdot} - \overline{Y}_{\cdot 2\cdot} = -176.50$$

with

$$\text{var}(\widehat{\theta}_3) = \frac{\widehat{\sigma}}{2} = 24.33.$$

For hypothesis

$$H_0 : \theta_3 \geq -150 \quad \text{versus} \quad H_1 : \theta_3 < -150,$$

the observed t-statistic

$$t_{\text{obs}} = \frac{-176.50 + 150}{24.33} = -1.089$$

with p-value 0.1488, which implies that there is no strong enough evidence to conclude that *Alkaline* saves at least 150 min per dollar, comparing with the heavy duty ones.

3.3 Multiple Comparisons

In this section, we study the multiple comparisons in the two-way complete model (3.1). The multiple comparison techniques in the one-way ANOVA model can be applied to construct simultaneous confidence interval for contrasts.

Directly from the result for the one-way ANOVA model, for all linear contrasts of form

$$\sum_i \sum_j d_{ij}\tau_{ij} \quad \text{with} \quad \sum_i \sum_j d_{ij} = 0$$

the Scheffé simultaneous confidence interval is

$$\sum_i \sum_j d_{ij}\overline{Y}_{ij} \pm \sqrt{(IJ-1)F_{IJ-1,n-IJ}(1-\alpha)}\sqrt{\sum_i \sum_j d_{ij}^2}\widehat{\sigma}.$$

When I and J are moderate, the interval can be very wide. For example, $I = 3$, $J = 4$, and $\alpha = 5\%$, The multiplier

$$\begin{aligned} & \sqrt{(IJ-1)F_{(IJ-1),n-IJ}(1-\alpha)} = \sqrt{11F_{11,n-12}(0.95)} \\ \approx \quad & \sqrt{11\chi^2_{11}(0.95)/11} = 4.436 \end{aligned}$$

when n is large. So, the interval can be very wide. In most practical applications, we may only be interested in some of the contrasts. In this case, we can construct a shorter interval. For example, if we want to construct m linear contrasts $\sum_i \sum_j d_{ij}\tau_{ij}$, the multiplier can be replace by Bonferroni multiplier $w_B = t_{u-IJ}(1-\alpha/2m)$.

When $n_{ij} = K$, we are interested in main effect contrasts $\sum_i c_i\bar{\tau}_{i\cdot}$ with $\sum_i c_i = 0$. For this purpose, our problem is nearly the same as that based on the data

$$\overline{Y}_{i\cdot k} = \mu + \bar{\tau}_{i\cdot} + \bar{\varepsilon}_{i\cdot k},$$

where $\bar{\varepsilon}_{i\cdot k}$'s are independent and identically distributed $N(0, \sigma^2/J)$, which is a one-way ANOVA model. Therefore the simultaneous confidence interval for $\sum_i c_i\bar{\tau}_{i\cdot}$ has the form

$$\sum_i c_i\overline{Y}_{i\cdot\cdot} \pm w\sqrt{\sum_i c_i^2/(JK)}\widehat{\sigma},$$

where the multiplier w may be either Scheffé multiplier w_S for all linear contrasts, or Tukey multiplier w_T for pairwise contrasts, or

Bonferroni multiplier w_B, where

$$w_S = \sqrt{(I-1)F_{I-1,n-IJ}(1-\alpha)},$$
$$w_T = q_{I,n-IJ}(\alpha)/\sqrt{2},$$
$$w_B = t_{n-IJ}(1-\alpha/2m).$$

For example, $I = 3, J = 4$ and $\alpha = 5\%$, when n is large, $w_S \approx \sqrt{\chi^2_3(0.95)} = \sqrt{7.815} = 2.796 < 4.436$.

3.4 ANOVA for the Two-way Complete Model

When the two-way complete model is used, a natural question is whether or not the interaction between treatment factors A and B can be negligible. This leads us to study the testing hypothesis:

$$H_0^{AB} : \gamma_{ij} - \gamma_{il} = \gamma_{jk} - \gamma_{jl} \quad \text{for} \quad \text{all} \;\; i \neq j, k \neq l.$$

Under the null hypothesis H_0, the model becomes the main effect model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}.$$

Therefore the sum of squares values is

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \mu - \alpha_i - \beta_j)^2.$$

The normal equations are

$$\textstyle\sum_i \sum_j \sum_k (Y_{ijk} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j) = 0,$$
$$\textstyle\sum_i \sum_k (Y_{ijk} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j) = 0,$$
$$\textstyle\sum_j \sum_k (Y_{ijk} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j) = 0.$$

The normal equations can be re-arranged as

$$\sum_i \sum_j \sum_k Y_{ijk} - n\widehat{\mu} - \sum_i \widehat{\alpha}_i n_{iT} - \sum_j \widehat{\beta}_j n_{Tj} = 0,$$
$$\sum_i \sum_k Y_{ijk} - n_{iT}\widehat{\alpha}_i - \sum_j n_{ij}\widehat{\beta}_j = 0,$$
$$\sum_j \sum_k Y_{ijk} - n_{Tj} \sum_i n_{ij}\widehat{\alpha}_i - n_{Tj}\widehat{\beta}_j = 0,$$

where $n_{iT} = \sum_j n_{ij}$ and $n_{Tj} = \sum_i n_{ij}$. Thus

$$\overline{Y}_{...} - \widehat{\mu} - \tfrac{1}{n} \sum_i n_i \widehat{\alpha}_i - \tfrac{1}{n} \sum_j n_{Tj}\widehat{\beta}_j = 0,$$
$$\overline{Y}_{i..} - \widehat{\mu} - \widehat{\alpha}_i - \tfrac{1}{n_{iT}} \sum_j n_{ij}\widehat{\beta}_j = 0,$$
$$\overline{Y}_{.j.} - \widehat{\mu} - \tfrac{1}{n_{Tj}} \sum_i n_{ij}\widehat{\alpha}_i - \widehat{\beta}_j = 0.$$

There are only $(I+J-2)$ independent equations, but we have $I+J-1$ independent variables. Thus set constraints

$$\sum_{i=1}^{I} n_{iT}\widehat{\alpha}_i = 0 \quad \text{and} \quad \sum_j n_{Tj}\widehat{\beta}_j = 0.$$

Under the constraints, it follows that

$$\widehat{\mu} = \overline{Y}$$

and $\widehat{\alpha}_i$ and $\widehat{\beta}_j$ can be solved. Hence the residual sum of squares under the null hypothsis

$$\mathrm{RSS}_0 = \sum_i \sum_j \sum_k (Y_{ijk} - \widehat{\alpha}_i - \widehat{\beta}_j - \widehat{\mu})^2$$

and the sum of squares due to interaction

$$\mathrm{SS}_{AB} = \mathrm{RSS}_0 - \mathrm{RSS}_1.$$

where $\mathrm{RSS}_1 = \mathrm{SS}_E = \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij.})^2$. Therefore F-statistic is

$$F = \frac{\mathrm{SS}_{AB}/(I-1)(J-1)}{\mathrm{SS}_E/(n-I-J)}.$$

When designs in the two-way complete model are balanced, we can simplify the above formula. In this case, the constraints become $\sum_{i=1}^{I} \widehat{\alpha}_i = 0$ and $\sum_{j=1}^{J} \widehat{\beta}_j = 0$. Hence

$$\widehat{\alpha}_i = \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots},$$
$$\widehat{\beta}_j = \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}.$$

Consequently,

$$\begin{aligned} \mathrm{SS}_{AB} &= \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2 - \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{ij\cdot})^2 \\ &= K \sum_{i,j} (Y_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2 \end{aligned}$$

where $K = n_{ij}$ for all i as this is balanced design.

As defined in Section 3.2,

$$\alpha_i^* = \bar{\tau}_{i\cdot} = \alpha_i + \bar{\gamma}_{i\cdot}, \quad i = 1, \cdots, I.$$

Now we consider testing hypothesis concerning the main effects

$$H_0^A : \alpha_1^* = \alpha_2^* = \cdots = \alpha_I^*$$

with constraints $\sum_i n_{iT}\alpha_i = 0$, $\sum_j n_{Tj}\beta_j = 0$, $\sum_i n_{iT}\bar{\gamma}_{i\cdot} = 0$ for $j = 1, \cdots, J$ and $\sum_j n_{Tj}\bar{\gamma}_{ij} = 0$ for $i = 1, \cdots, I$.

For the two-way complete model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

with the constraints. H_0^A is equivalent to

$$H_0 : \alpha_i = 0, \quad \text{for} \quad i = 1, 2, \cdots, I.$$

Under the null hypothesis, and when designs are balanced

$$\widehat{\mu} = \overline{Y}_{...},\ \widehat{\beta}_j = (\overline{Y}_{.j.} - \overline{Y}_{...}) \text{ and } \widehat{\gamma}_{ij} = (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})$$

for $i = 1, \cdots, I$ and $j = 1, \cdots, J$. Therefore the residuals sum of squares under H_0^A,

$$\text{RSS}_0 = \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{ij.} - \overline{Y}_{i..} + \overline{Y}_{...})^2.$$

Hence, the sum of squared errors due to factor A is

$$\begin{aligned} \text{SS}_A &= \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{ij.} - \overline{Y}_{i..} + \overline{Y}_{...})^2 - \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{ij.})^2 \\ &= JK \sum_i (\overline{Y}_{i..} - \overline{Y}_{...})^2 \end{aligned}$$

The corresponding degrees of freedom is $I - 1$. Similarly, for testing the main effects of factor B

$$H_0^B : \beta_1^* = \beta_2^* = \cdots = \beta_J^*.$$

The sum of squared errors due to factor B is

$$\text{SS}_B = IK \sum_j (\overline{Y}_{.j.} - \overline{Y}_{...})^2.$$

Hence we have the following decompositions:

$$\begin{aligned} & \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{...})^2 \\ = & \sum_{i,j,k} \{(Y_{ijk} - \overline{Y}_{ij.}) + (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...}) \\ & + (\overline{Y}_{i..} - \overline{Y}_{...}) + (\overline{Y}_{.j.} - \overline{Y}_{...})\}^2 \\ = & \text{SS}_E + \text{SS}_{AB} + \text{SS}_A + \text{SS}_B \end{aligned}$$

Usually, we summarized the results for analysis of variance in the in the following two-way ANOVA table.

TABLE 3.2. Analysis of Variance for Two Way Complete Models

Source	D.F.	SS	MS	F value
Factor A	$I-1$	SS_A	$\text{SS}_A/(I-1)$	MS_A/MS_E
Factor B	$J-1$	SS_B	$\text{SS}_B/(J-1)$	MS_B/MS_E
Interactions	$(I-1)(J-1)$	SS_{AB}	$\text{SS}_{AB}/(I-1)(J-1)$	$\text{MS}_{AB}/\text{MS}_E$
Error	$n-IJ$	SS_E	$\text{SS}_E/(n-IJ)$	
Total	$n-1$			

Example 3.1 (**Continued**) Using S-plus, the following two-way ANOVA table can be extracted from the output.

TABLE 3.3. ANOVA table for *Battery Experiments*

Source	D.F.	SS	MS	F	p
Brand	1	124,609	124,609	52.6285	$1.008 * 10^{-5}$
Duty	1	252,004	252,004	106.4337	$2.555 * 10^{-7}$
Interactions	1	51,302	51,302	21.6675	$5.558 * 10^{-4}$
Error	12	28,412.5	2367.7		
Total	15				

From Table 3.3, the p-values indicate that all of the main effects and interactions are highly statistically significant.

Example 3.2 *(Female Labor Supply: revisited) Note that this an unbalanced design. So, the unconditional ANOVA decomposition does*

not hold. However, the following conditional decomposition hold

$$\begin{aligned}
& \sum_{ijk}(Y_{ij} - \overline{Y}_{...})^2 \\
= \; & RSS(\mu) \\
= \; & RSS(\mu) - RSS(A) + RSS(A,B) - RSS(A) \\
& + RSS(A*B) - RSS(A,B) + RSS(A*B) \\
= \; & SS(A) + SS(B|A) + SS(A*B|A,B) + SS_E
\end{aligned}$$

The first term $SS(A)$ is the sum of squares due to factor A, the second one $SS(B|A)$ is the sum of squares reduction due to factor B given the contributions of A. Similarly, the third one $SS(A*B|A,B)$ is the sum of squares reduction due to interaction given the contributions of A and B. We also can decompose in another way,

$$\begin{aligned}
\sum_i (Y_i - \overline{Y})^2 &= RSS(\mu) \\
&= SS(B) + SS(A|B) + SS(A*B|A,B) + SS_E
\end{aligned}$$

First the data set was fitted a two way layout model. From the S-plus outputs, the estimated contrasts are as follows:

Intercept	*C*	*EB*	*EC*	*ED*	*C*EB*	*C*EC*	*C*ED*
14.40	*-0.44*	*-3.87*	*-4.07*	*-5.97*	*0.76*	*-1.19*	*-0.68*

The two way ANOVA table is depicted in Table 3.4.

Now we add the factors into the model in different order. The decomposition of sum of squares is listed in Table 3.5. Compared with Table 3.4, the conditional contribution of factor child change a lot. The p-value reduces from 0.613 to 0.061. But the contribution of interaction terms does not change and is not significant at level

TABLE 3.4. ANOVA table for Example 3.2

Source	D.F.	SS	MS	F	p
Child	1	4.057	4.058	0.257	0.613
Education	3	1936.489	645.496	40.852	0.000
Interaction	3	85.213	28.404	1.798	0.146
Residuals	599	9464.721	15.8009		
Total	606				

0.05. This suggest us to fit the data set a main effect model, studied in next section.

TABLE 3.5. ANOVA table for Example 3.2

Source	D.F.	SS	MS	F	p
Eduf	3	1884.844	628.281	39.762	0.000
Child	1	55.703	55.703	3.52531	0.061
Interaction	3	85.213	28.404	1.798	0.146
Residuals	599	9464.721	15.8009		
Total	606				

3.5 Analysis of the Main Effect Model

In practice, the main effect model is also very useful. In fact, the interaction between factors A and B may be negligible when some prior knowledge about interactions or scientific reasons are available. In these situations, it is suggested to use the main effect model. In

this section, we consider the main effect model with balanced design

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

where $i = 1, \cdots, I$, $j = 1, \cdots, J$ and $k = 1, \cdots, K$. With constraints $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$, the least squares estimates for the parameters in the model are

$$\begin{aligned} \widehat{\mu} &= \overline{Y}_{\cdots}, \\ \widehat{\alpha}_i &= \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}, \\ \widehat{\beta}_j &= \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}. \end{aligned}$$

So for the estimable main effect contrasts:

$$\theta = \sum_i c_i \alpha_i$$

with $\sum_i c_i = 0$, the least squares estimate is

$$\sum_i c_i \widehat{\alpha}_i = \sum_i c_i \overline{Y}_{i\cdot\cdot}$$

with the associated variance

$$\mathrm{var}(\sum_i c_i \widehat{\alpha}_i) = \sum_i (c_i^2/JK)\sigma^2.$$

Similar formula applies to the contrast on the factor B. The sum of squared residuals is

$$\begin{aligned} \mathrm{RSS} &= \sum_{i,j,k} (Y_{i,j,k} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2 \\ &= \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{ij\cdot})^2 + \sum_{i,j,k} (Y_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2 \\ &= \mathrm{SS}_E + \mathrm{SS}_{AB} \end{aligned}$$

for a two-way layout model.

Using the theory for two-way ANOVA models, it follows that

$$\mathrm{SS}_E \sim \sigma^2 \chi^2_{IJK-IJ} \quad \text{and} \quad \mathrm{SS}_{AB} \sim \sigma^2 \chi^2_{(I-1)(J-1)}$$

Thus

$$\mathrm{RSS} \sim \sigma^2 \chi^2_{n-I-J+1}$$

and

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{n-I-J+1} \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots})^2 \\ &\sim \sigma^2 \chi^2_{n-I-J+1}/(n-I-J+1). \end{aligned}$$

The multiple comparisons for main effects continue to apply. For example, simultaneous intervals for $\sum_i c_i \alpha_i$ is

$$\sum_i c_i \overline{Y}_{i\cdot\cdot} \pm w \sqrt{\sum_i c_i^2 / JK \widehat{\sigma}},$$

where w is the critical coefficients. The w equals

$$w_B = t_{n-I-J+1}(1 - \frac{\alpha}{2m})$$

for Bonferroni type simultaneous interval, or

$$w_S = \sqrt{(I-1) F_{I-1,n-I-J+1}(1-\alpha)}$$

for Scheffé type simultaneous confidence interval, or

$$w_T = q_{I,n-I-J+1}(1 - \frac{1}{\alpha})/\sqrt{2}$$

for Tukey type simultaneous confidence interval.

For analysis of variance, consider testing

$$H_0^B : \beta_1 = \cdots = \beta_J = \beta.$$

The model becomes

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

which is a one-way layout model. Hence

$$\mathrm{RSS}_0 = \sum_{i,j,k} (Y_{ijk} - \overline{Y}_{i\cdot\cdot})^2.$$

The sum of squares due to the factor B is

$$\mathrm{SS}_B = \mathrm{RSS}_0 - \mathrm{RSS}_1 = \sum_{ijk} (\overline{Y}_{\cdots} - \overline{Y}_{\cdot j\cdot})^2 = IK \sum_{j} (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})^2.$$

Similarly, the sum of squares due to factor A is

$$\mathrm{SS}_A = JK \sum_{i} (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2.$$

Further, we have the following decomposition

$$\sum_{ijk} (Y_{ijk} - \overline{Y}_{\cdots})^2 = \mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_E$$

which yields to the following two-way ANOVA table without interactions.

TABLE 3.6. ANOVA Table for Two Way Main Effect Model

Source	D.F.	SS	MS	F value
Factor A	$I-1$	SS_A	$\mathrm{SS}_A/(I-1)$	$\mathrm{MS}_A/\mathrm{MS}_E$
Factor B	$J-1$	SS_B	$\mathrm{SS}_B/(J-1)$	$\mathrm{MS}_B/\mathrm{MS}_E$
Error	$n-I-J+1$	SS_E	$\mathrm{SS}_E/(n-I-J+1)$	
Total	$n-1$			

Note that for a balanced design, the estimators for μ, α_i and β_j are the same as those for the complete model. The only difference is the degrees of freedom in estimating σ.

Example 3.2 (Continued) As mentioned in Section 3.4, the interaction between child and education is not significant. So we may fit the data set by the main effect model

$$\text{earning} \sim \text{child} + \text{education}$$

The estimated contrasts are displayed as follows:

Intercept	child	edufB	edufC	edufD
14.575	-0.675	-3.373	-4.958	-6.258

The two way ANOVA table is depicted in Table 3.7.

TABLE 3.7. ANOVA table for Example 3.2

Source	D.F.	SS	MS	F	p
child	1	4.057	4.058	0.256	0.613
eduf	3	1936.489	645.496	40.690	0.000
Residuals	602	9549.934	15.864		
Total	606				

Table 3.8 shows that the decomposition of sum of squares are dependent on the order of covariates adding into the model.

TABLE 3.8. ANOVA table for Example 3.2

Source	D.F.	SS	MS	F	p
eduf	3	1884.844	628.281	39.605	0.000
child	1	55.703	55.703	3.511	0.061
Residuals	602	9549.934	15.864		
Total	606				

3.6 S-plus Codes

3.6.1 S-plus codes for Example 3.1

```
>
>data <- read.table("d:/rli/book/battery.dat",header=T)
>duty <- data[,1]
>brand <- data[,2]
>cost <- data[,3]
>postscript("d:/rli/book/figsa/ch3fig1.ps",width=4,
+  height=4,horizontal=F,points=8)
>interaction.plot(duty,brand,cost)
>dev.off()
>battery.aov <- aov(cost ~ duty*brand)
>summary(battery.aov)
>
            Df Sum of Sq  Mean Sq  F Value        Pr(F)
      duty  1  252004.0 252004.0 106.4337 0.0000002555
     brand  1  124609.0 124609.0  52.6285 0.0000100835
duty:brand  1   51302.2  51302.2  21.6675 0.0005558051
 Residuals 12   28412.5   2367.7

 summary(aov(cost~brand))
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
    brand  1  124609.0 124609.0 5.259052 0.03783034
Residuals 14  331718.8  23694.2


> summary(aov(cost~brand+duty))
```

```
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
    brand  1  124609.0 124609.0 20.32142 0.0005886884
     duty  1  252004.0 252004.0 41.09719 0.0000230424
Residuals 13   79714.8   6131.9


> summary(aov(cost~duty+brand))


          Df Sum of Sq  Mean Sq  F Value        Pr(F)
     duty  1  252004.0 252004.0 41.09719 0.0000230424
    brand  1  124609.0 124609.0 20.32142 0.0005886884
Residuals 13   79714.7   6131.9
```

3.6.2 S-plus codes for Example 3.2

```
>
>labor <- read.table("d:/rli/book/labor.dat",header=T)
>earning <- labor[,3]
>edu <- labor[,5]
>eduf <- rep("A",length(edu))
>eduf[edu<16]<-"B"
>eduf[edu<13]<-"C"
>eduf[edu<12]<-"D"
>child <- labor[,8]
>lm(earning~child*eduf)
>

Call:
lm(formula = earning ~ child * eduf)

Coefficients:
   (Intercept)     child     eduf1      eduf2      eduf3
childeduf1  childeduf2 childeduf3
    10.92663 -0.722068 -1.935618 -0.7122286 -0.8293598
0.3795355   -0.5229781 -0.1344074
```

```
Degrees of freedom: 607 total; 599 residual
Residual standard error: 3.975031


> options()$contrast


          factor      ordered
 "contr.helmert" "contr.poly"


> options(contrasts=c("contr.sum","contr.sum"))
> lm(earning~child*eduf)


Call:
lm(formula = earning ~ child * eduf)

Coefficients:
   (Intercept)     child    eduf1      eduf2      eduf3
childeduf1  childeduf2 childeduf3
    10.92663 -0.722068 3.477207 -0.3940298 -0.5950974
0.27785     1.036921 -0.9115487

Degrees of freedom: 607 total; 599 residual
Residual standard error: 3.975031


> options(contrasts=c("contr.treatment","contr.treatment"))
> lm(earning~child*eduf)


Call:
lm(formula = earning ~ child * eduf)

Coefficients:
   (Intercept)      child     edufB     edufC     edufD
childedufB  childedufC  childedufD
    14.40383 -0.4442179 -3.871236 -4.072304 -5.965286
0.7590709  -1.189399 -0.6810723
```

```
Degrees of freedom: 607 total; 599 residual
Residual standard error: 3.975031


> labor.aov <-aov(earning~child*eduf)
> summary(labor.aov)


             Df Sum of Sq  Mean Sq  F Value     Pr(F)
     child    1     4.057   4.0575  0.25679 0.6125209
      eduf    3  1936.489 645.4964 40.85195 0.0000000
child:eduf    3    85.213  28.4042  1.79764 0.1464194
 Residuals  599  9464.721  15.8009

> summary(aov(earning~eduf*child))

             Df Sum of Sq  Mean Sq  F Value     Pr(F)
      eduf    3  1884.844 628.2813 39.76245 0.0000000
     child    1    55.703  55.7030  3.52531 0.0609242
eduf:child    3    85.213  28.4042  1.79764 0.1464194
 Residuals  599  9464.721  15.8009


>lm(earning~child+eduf)


Call:
lm(formula = earning ~ child + eduf)

Coefficients:
 (Intercept)      child     edufB     edufC     edufD
    14.57527 -0.6749966 -3.373208 -4.957911 -6.257954

Degrees of freedom: 607 total; 602 residual
Residual standard error: 3.982923


>summary(aov(earning~child+eduf))
```

```
           Df Sum of Sq  Mean Sq  F Value     Pr(F)
    child   1     4.057   4.0575  0.25577 0.6132244
     eduf   3  1936.489 645.4964 40.69021 0.0000000
Residuals 602  9549.934  15.8637


>summary(aov(earning~eduf+child))

           Df Sum of Sq  Mean Sq  F Value      Pr(F)
     eduf   3  1884.844 628.2813 39.60502 0.00000000
    child   1    55.703  55.7030  3.51135 0.06143388
Residuals 602  9549.934  15.8637
>
```

4

Analysis of Covariance

4.1 Introduction

In previous two chapters, it is assumed that the source of variations are mainly due to one or more treatment factors. If confounding or nuisance factors are expected to be a major source of variations, they should be taken into account in the design and analysis of the experiment. For example, in the female labor supply data, in addition to the education level that affects the income, gender, age and job prestige should also affect the income. To control or adjust for this prior weight variability, we may make each group resemble with respect to gender, age and job prestige, and eliminate confounding factors in the stage of design experiments. An alternative approach is to adjust them by linear models in the stage of analysis of experiments.

Example 4.1 *(Female Labor Supply: Revisited) In Chapters 2 and 3, we used this data set to illustrate the one way ANOVA model and the two way layout model. We will illustrate the ANOCOVA model using this data set in this example. Here, we take the hourly earning as response variable again. It has been concluded in Example 2.1 that the levels of education affect the hourly earnings. It is known that hourly earnings are affected by other factors, such as age and job prestige. In Example 2.1, factors but the education factor are treated as confounding covariates. To see how the education factor affects the hourly earnings, a good model should also consider the confounding variables. In this example, we incorporate the factors of age and job prestige. Thus we consider the model*

$$earning_{ik} = \mu + eduf_i + \beta_1 age_{ik} + \beta_2 prestige_{ik} + \varepsilon_{ik}.$$

This model may be abbreviated as

$$earning \sim education + age + prestige$$

For this model, of interest is to investigate whether or not

(i) there exist any treatment effects;

(ii) there exists any age effects.

It is also of interest to estimate various contrasts.

4.2 Models

Consider an experiment conducted as a completely randomized design to compare the effects of the levels of ν treatments on a response

variable Y. Suppose that the response is also affected by a few nuisance covariates whose value $\mathbf{x}$ can be measured during or prior to the experiment. Further, suppose that there is a linear relationship between $E(Y)$ and $\mathbf{x}$, which can be examined by a scatter plot for each level of the factors. A comparison of the effects of the two treatments can be done by comparison of mean response at any value of $\mathbf{x}$.

The model that allows this type of analysis is analysis of covariance model. For simplicity, assume that there are one factor and p covariates. Then the analysis of covariance model is

$$\begin{aligned} Y_{ik} &= \mu + \alpha_i + \sum_{j=1}^{p} \beta_j x_{ijk} + \varepsilon_{ik}, \qquad (4.1) \\ & \qquad k = 1, \cdots, K_i, \quad i = 1, \cdots, I, \end{aligned}$$

where ε_{ik} are independent and identically distributed $N(0, \sigma^2)$.

In this model, the effect of the i-th treatment is modeled as α_i as usual. If there is more than one treatment factor, then it could be replaced by main effects and interactions. The value of covariates on the k-th time that treatment i is observed is written as $\mathbf{x}_{ik}$, and the linear relationship between the response and the covariates is modeled as $\mathbf{x}_{ik}^T \boldsymbol{\beta}$ as in a regression model. It is important for the analysis that treatments do not affect the value $\mathbf{x}_{ik}$ of the covariates, otherwise, comparison of treatment affects at a common $\mathbf{x}$-value would not be meaningful.

A common alternative form of the analysis of covariance model is

$$Y_{ik} = \mu + \alpha_i + \sum_{j=1}^{p} \beta_j (x_{ijk} - \bar{x}_{\cdot j \cdot}) + \varepsilon_{ik},$$

in which the covariate values have been "centered". This model is frequently used to reduce computational problem and is a little easier to work with in obtaining least squares estimates. The two models are equivalent for comparison of treatment effects. The slope parameter β has the same interpretation in both models.

In addition to the usual assumptions on the error variables, the analysis of covariance model (4.1) assumes that there exists a linear relationship between the covariate and the mean response, with the same slope for each treatment. It is appropriate to start by checking for model lack of fit.

Lack of fit can be investigated by plotting the residuals versus the covariates for each treatment on the same scale. If the plot looks nonlinear for any treatment, then a linear relationship between the response and covariate may not be adequate. If each plot does look linear, one can assess whether the slopes are comparable. A formal test of equality of slopes can be conducted by comparing the fit of the analysis of covariance model (4.1) with the fit of the corresponding model that does not require equal slopes, for which

$$Y_{ik} = \mu + \alpha_i + \sum_{j=1}^{p} \beta_{ij} x_{ijk} + \varepsilon_{ik}.$$

This model will reduce modeling bias, but will result in a wider confidence intervals as it uses more parameters, compared with the model (4.1). If there is no significant lack of fit of the model, then plots of the residuals versus run order, predicted values, and normal scores can be used to assess the assumptions of independence equal variance, and normality of the random error terms.

4.3 Least Squares Estimates

The least squares estimates of $\mu, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ may be derived via minimizing the sum of squared errors. As in the one way ANOVA model, the corresponding normal equations are not linearly independent. To obtain the least squares estimate a little easier, we rewrite model (4.1) as

$$Y_{ik} = \tau_i + \sum_{j=1}^{p} \beta_j (x_{ijk} - \bar{x}_{ij\cdot}) + \varepsilon_{ik}, \tag{4.2}$$

where $\tau_i = \mu + \alpha_i + \sum_{j=1}^{p} \beta_j \bar{x}_{ij\cdot}$. Express model (4.2) using matrix notation, it follows that

$$\mathbf{y} = \mathbf{A}\boldsymbol{\tau} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{A}$ is an $n \times I$ design matrix associated with a one way ANOVA model and $\mathbf{X}$ is an $n \times p$ design matrix associated with $(x_{ijk} - \bar{x}_{ij\cdot})$. Since the $\mathbf{X}$ was centered, $\mathbf{A}^T\mathbf{X} = 0$. Thus the least squares estimate for $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ is

$$\begin{aligned} \widehat{\tau}_i &= \overline{Y}_{i,}, \\ \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned}$$

To obtain the least squares estimate of μ and α's, one have to solve the equations

$$\widehat{\mu} + \widehat{\alpha}_i = \overline{Y}_{i\cdot} - \sum_{j=1}^{p} \widehat{\beta}_j \bar{x}_{ij\cdot}. \tag{4.3}$$

This is the same as the normal equations in a one way ANOVA model if we treat the term in right-hand side of (4.3) as "cell mean". It is clear that there are $I+1$ parameters, but there are only I equations. Therefore there is no unique solution for $\widehat{\mu}$ and $\widehat{\alpha}_i$'s. Like in the

analysis of one way layout model, one needs to impose a constraint on α_i's. Typically it is assumed that $\sum_{i=1}^{I} n_i \widehat{\alpha}_i = 0$. Under this assumption,

$$\begin{aligned} \widehat{\mu} &= \frac{1}{I}\sum_{i=1}^{I}(\widehat{\tau}_i - \sum_{j=1}^{p}\widehat{\beta}_j \bar{x}_{ij\cdot}), \\ \widehat{\alpha}_n &= \widehat{\tau}_i - \sum_{j=1}^{p}\widehat{\beta}_j(\bar{x}_{ij\cdot} - \widehat{\mu}). \end{aligned}$$

Different package of statistical analysis may impose a different constraint on α_i's. For example, SAS assumes that $\alpha_I = 0$.

4.4 Analysis of Covariance

In practice, it is of interest to examine whether or not there are differences among different treatments. For a completely randomized design and analysis of covariance model (4.1), a one way analysis of covariance is used to test the null hypothesis

$$H_0^{\alpha} : \alpha_1 = \cdots = \alpha_p = 0$$

against the alternative hypothesis H_A that at least two of the α_i differ. In this section, we study the analysis of covariance. For simplicity, we still assume that there is one factor and p-covariates $X_1, \cdots, X_p$.

Denote by $\mathrm{SS}_E (= RSS_1(T, \boldsymbol{\beta}))$ the residual sum of squares under the full model, i.e.

$$\mathrm{SS}_E = RSS_1(T, \boldsymbol{\beta}) = \sum_{i,k}(y_{ik} - \widehat{\mu} - \widehat{\alpha}_i - \sum_{j=1}^{p}\widehat{\beta}_j \bar{x}_{ijk})^2$$

and RSS_0 be the residual sum of squares under the null hypothesis,

$$RSS_0(\boldsymbol{\beta}) = \sum_{i,k}(y_{ik} - \widehat{\mu}^0 - \sum_{j=1}^{p} \widehat{\beta}_j^0 x_{ijk})^2,$$

where $\widehat{\mu}^0$ and $\widehat{\boldsymbol{\beta}}^0$ are the maximum likelihood estimate under H_0. Then, by the general linear model theory,

$$\mathrm{SS}_E \sim \sigma^2 \chi^2_{n-p-I}$$

and sum of squares due to the treatment is

$$SS_{T|\boldsymbol{\beta}} = RSS_0(\boldsymbol{\beta}) - RSS_1(T, \boldsymbol{\beta}) \sim \sigma^2 \chi^2_{I-1}.$$

Therefore, we reject H_0 if

$$F = \frac{SS_{T|\boldsymbol{\beta}}/(I-1)}{\mathrm{SS}_E/(n-p-I)}$$

is too large, comparing with $F_{I-1,n-p-I}(1-\alpha)$. Now consider testing the problem of no covariate effect

$$H_0^{\boldsymbol{\beta}} : \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

In this case, the model becomes the one way layout model and

$$RSS_0(T) = \sum_{i,k}(Y_{i,k} - \overline{Y}_{i\cdot})^2 \sim \chi^2_{n-I}.$$

Thus sum of squares due to covariates

$$SS(\boldsymbol{\beta}|T) = RSS(T, \boldsymbol{\beta}) - RSS_0(T) \sim \chi^2_p,$$

we reject $H_0^{\boldsymbol{\beta}}$ if

$$F = \frac{SS(\boldsymbol{\beta}|T)/p}{\mathrm{SS}_E/(n-p-I)}$$

is too large.

TABLE 4.1. Analysis of Covariance for p Linear Covariates

Source	D.F.	SS	MS	F value
Treatment	$I-1$	SS_T	$SS_T/(I-1)$	MS_T/MS_E
Covariate	p	$SS(\boldsymbol{\beta}\|T)$	$SS(\boldsymbol{\beta}\|T)/p$	$MS(\boldsymbol{\beta}\|T)/\text{MS}_E$
Error	$n-I-p$	SS_E	$\text{SS}_E/(n-I-p)$	
Total	$n-1$			

Since the parameters in the treatment and covariate are not orthogonal, $SS(\boldsymbol{\beta}|T)$ equals to the sum of squares reduction due to covariates, given the contributions of treatment. In fact, one can decompose the total of sum of squares in a more detailed form

$$\begin{aligned} RSS(\mu) &= RSS(\mu) - RSS(A) + RSS(A, X_1) - RSS(A) + \cdots \\ &\quad + RSS(A, X_1, \cdots, X_{p-1}) - RSS(A, X_1, \cdots, X_p) \\ &\quad + RSS(A, X_1, \cdots, X_p) \\ &= SS_T + SS_{\boldsymbol{\beta}} + SS_E. \end{aligned}$$

Interpretation: $RSS(A, X_1, \cdots, X_k) - RSS(A, X_1, \cdots, X_{k-1}) \sim \sigma^2\chi^2_1$, which holds if model $Y \sim A + X_1 + \cdots X_{k-1}$ is correct, is the sum of squares reduction due to X_k, given the contribution already in $A, X_1, \cdots, X_{k-1}$.

Example 4.1 (Continued) We fit the earning data with the following model

earning$\sim$ eduf+age+prestige

From S-plus, the following ANCOVA table is obtained. Thus age

TABLE 4.2. Analysis of Covariance for p Linear Covariates

Source	D.F.	SS	F value	p-value
eduf	3	1884.844	43.808	0.0000
age	1	32.268	2.250	0.1314
prestige	1	954.093	66.53	0.0000
Error	601	8619.276		

is not significant given eduf. Sum of squares due to age and prestige = 32.268+954.093=977.361 with degrees 2 of freedom. For testing $H_0^{\boldsymbol{\beta}} : \beta_1 = \beta_2 = 0$,

$$F = \frac{977.361/2}{8629.276/601} = 34.035$$

with degrees (2, 601) of freedom. The corresponding p-value equals to 0.

We refit the data by adding the prestige covariate first, then adding *eduf* and *age* one by one. The corresponding ANCOVA table is Table 4.3. The p-value for age is 0.31, which is not significant. Note that after the contributions of prestige, the contribution due to education is considerably reduced. However, it is still highly statistical significant.

TABLE 4.3. ANOCA Table for Example 4.1

Source	D.F.	SS	F value	p-value
prestige	1	2394.842	166.986	0.0000
eduf	3	461.556	10.852	$7 * 10^{-7}$
age	1	14.806	1.0324	0.31
Error	601	8619.276		

For testing $H_0^\alpha : \alpha_1 = \cdots = \alpha_4 = 0$ in the model

$$\text{earning}_{i,k} \sim \mu + \alpha_i + \beta(\text{prestige}_{i,k}) + \varepsilon_{i,k}.$$

The p-value is 7×10^{-7}.

Even though the sum of squares admits different ways of decomposition, the model is still the same and the estimated coefficients are the same. The estimated coefficients and their standard errors are summarized in Table 4.3 based on S-plus output. Note that it is assumed that $\alpha_1 = 0$ in S-plus.

TABLE 4.4. Estimated Contrasts and Standard Errors

	Est	SE	t-value	p-value
Intercept	6.7481	1.1958	5.6430	0
Prestige	0.1186	0.0145	8.1564	0
EdufB	-2.2353	0.5373	-4.3465	0
EdufC	-2.9164	0.5809	-5.0205	0
EdufD	-3.6508	0.6582	-5.5470	0
Age	0.0212	0.0209	1.0161	0.31

Compareed with the linear model

$$\text{earning} \sim \text{eduf}$$

,

the effect of education reduces a lot.

TABLE 4.5. Estimated Contrasts

Intercept	EdufB	EdufC	EdufD
14.07	-3.328	-4.959	-5.938

4.5 Treatment Contrasts and Confidence Intervals

The parameter $(\mu+\alpha_i)$ is the intercept of the i-th treatment in linear model (4.1), therefore it is estimable. Thus $\sum c_i(\mu+\alpha_i)$ is estimable. In particular, $\theta = \sum_{i=1}^{I} c_i\alpha_i$ with $\sum_{i=1}^{I} c_i = 0$ is estimable. From (4.3), the least squares estimates of the contrast θ is

$$\widehat{\theta} = \sum_{i=1}^{I} c_i(\overline{Y}_{i.} - \sum_{j=1}^{p} \widehat{\beta}_j \bar{x}_{ij.})$$

Its associated variance can be expressed as follows by noting $\mathbf{A}^T\mathbf{X} = 0$,

$$\begin{aligned} \mathrm{var}(\widehat{\theta}) &= \mathrm{var}(\sum_{i=1}^{I} c_i\widehat{\tau}_i - \sum_{i=1}^{I} c_i \sum_{j=1}^{p} \widehat{\beta}_j \bar{x}_{ij.}) \\ &= \sigma^2\{\mathbf{c}^T(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{c} + \mathbf{c}^T\overline{\mathbf{X}}(\overline{\mathbf{X}}^T\overline{\mathbf{X}})^{-1}\overline{\mathbf{X}}\mathbf{c}\}, \end{aligned}$$

where $\mathbf{A}$ and $\mathbf{X}$ were defined in Section 4.3 and $\overline{\mathbf{X}} = (\bar{x}_{ij.})_{I\times p}$.

Note that $\widehat{\sigma}$ and $\widehat{\theta}$ are independent

$$\frac{\widehat{\theta} - \theta}{\widehat{\mathrm{SE}}(\widehat{\theta})} \sim t_{n-I-p}.$$

Thus a $100(1-\alpha)\%$ confidence interval for θ is

$$\widehat{\theta} \pm t_{n-I-p}(1 - \frac{\alpha}{2})\widehat{\mathrm{SE}}(\widehat{\theta}).$$

Similarly, simultaneous confidence interval are of form

$$\sum_{i=1}^{I} c_i\alpha_i \in \sum_{i=1}^{I} c_i\widehat{\alpha}_i \pm w\widehat{\mathrm{SE}}(\widehat{\theta})$$

for the Bonferroni method $w = w_B = t_{n-I-p}(1 - \frac{\alpha}{2m})$ and for the Scheffé method $w = w_S = \sqrt{(I-1)F_{I-1,n-I-p}(1-\alpha)}$.

TABLE 4.6. Estimated Contrasts and Their Standard Errors

	Est	SE
EdufB= $\alpha_2 - \alpha_1$	-2.2353	0.5373
EdufC= $\alpha_3 - \alpha_1$	-2.9164	0.5809
EdufD= $\alpha_4 - \alpha_1$	-3.6508	0.6582

Example 4.1 (Continued)

The associated correlation matrix among all contrasts is as follows

	Intercept	prestige	EdufB	EdufC	EdufD
prestige	-0.6256				
EdufB	-0.4702	0.2235			
EdufC	-0.6205	0.4131	0.7504		
EdufD	-0.3965	0.4684	0.6802	0.7086	
Age	-0.6440	-0.0590	0.0036	0.0738	-0.2573

A 95% confidence interval for $\alpha_2 - \alpha_1$ is $-2.3353 \pm t_{601}(1 - \frac{0.05}{2}) \times 0.5373 = -2.3353 \pm 1.0531$. Because $\alpha_3 - \alpha_2 = (\alpha_3 - \alpha_1) - (\alpha_2 - \alpha_1)$, thus $\widehat{\alpha}_3 - \widehat{\alpha}_2 = -2.9164 - (-2.3353) = -0.5811$ and its variance

$$\text{var}(\widehat{\alpha}_3 - \widehat{\alpha}_2) = \text{var}\{(\widehat{\alpha}_3 - \widehat{\alpha}_1) - (\widehat{\alpha}_2 - \widehat{\alpha}_1)\}.$$

Thus

$$\widehat{\text{SE}} = \sqrt{0.5809^2 + 0.5373^2 - 2 \times 0.07504 \times 0.5809 \times 0.5373} = 0.3900.$$

Therefore a 95% confidence interval for $\alpha_3 - \alpha_2$ is $-0.5811 \pm 1.96 \times 0.3900 = 0.5811 \pm 0.7644$.

The coefficient w_B for all pairwise contrasts is $t_{601}(1 - 0.05/12) = 2.647$ as $m = \binom{4}{2} = 6$, and the coefficient for constructing all

contrasts using Scheffé's method $w_S = \sqrt{(4-1)F_{3,601}(1-0.05)} = 2.803$. Thus all pairwise treatment contrasts have the simultaneous Bonferroni conference intervals are as follows:

$$\alpha_2 - \alpha_1 : \quad -2.3353 \pm 2.647 \times 0.5373$$

$$\alpha_3 - \alpha_1 : \quad -2.9164 \pm 2.647 \times 0.5809$$

$$\alpha_4 - \alpha_1 : \quad -2.6508 \pm 2.647 \times 0.6582$$

$$\vdots$$

$$\alpha_4 - \alpha_3 : \quad -0.5811 \pm 2.647 \times 0.3900$$

Simultaneous confidence intervals for all linear contrasts may be constructed by replacing 2.647 by 2.803.

4.6 S-plus codes for Example 4.1

```
> labor <- read.table("d:/rli/book/labor.dat",header=T)
> earning <- labor[,3]
> edu <- labor[,5]
> eduf <- rep("A",length(edu))
> eduf[edu<16]<-"B"
> eduf[edu<13]<-"C"
> eduf[edu<12]<-"D"
> age <- labor[,2]
> prestige <- labor[,4]
> lm(earning ~ eduf + age + prestige)
Call:
lm(formula = earning ~ eduf + age + prestige)

Coefficients:
 (Intercept)    eduf1      eduf2      eduf3        age  prestige
    4.522455 -1.16764 -0.5829174 -0.4750492 0.02124989 0.1185623
```

```
Degrees of freedom: 607 total; 601 residual
Residual standard error: 3.787025

> lm(earning ~ age + prestige)

Call:
lm(formula = earning ~ age + prestige)

Coefficients:
 (Intercept)         age  prestige
    3.224799 0.003113607 0.1590534

Degrees of freedom: 607 total; 604 residual
Residual standard error: 3.880508

> summary(aov(earning ~ eduf + age + prestige))

            Df Sum of Sq  Mean Sq  F Value     Pr(F)
     eduf    3  1884.844 628.2813 43.80844 0.0000000
      age    1    32.268  32.2683  2.24999 0.1341406
 prestige    1   954.093 954.0929 66.52645 0.0000000
Residuals  601  8619.276  14.3416

> summary(aov(earning ~ age + eduf  + prestige))

            Df Sum of Sq  Mean Sq  F Value     Pr(F)
      age    1     2.484   2.4838  0.17319 0.6774415
     eduf    3  1914.628 638.2094 44.50071 0.0000000
 prestige    1   954.093 954.0929 66.52645 0.0000000
Residuals  601  8619.276  14.3416

> summary(aov(earning ~ prestige + eduf + age))

            Df Sum of Sq  Mean Sq  F Value     Pr(F)
 prestige    1  2394.842 2394.842 166.9862 0.0000000
     eduf    3   461.556  153.852  10.7277 0.0000007
      age    1    14.806   14.806   1.0324 0.3100069
Residuals  601  8619.276   14.342
```

5

Mixed Effects Models

5.1 Introduction

So far, the treat effects in models we studied is treated as *fixed*. That is, levels of treatment factors were specifically chosen. We have tested hypothesis about, and calculated confidence intervals for, comparisons in the effects of these particular treatment factor levels. These treatment effects are known as fixed effects, and models that contain only fixed effects are called fixed effects models.

In some situations, when the levels of a treatment factor is large, the levels that are used in the experiment can be regarded as random variables. Such treatment factor effects are called random effect, and the corresponding models are called random effects models. Furthermore models containing both fixed effects and random effects are called *mixed effects models*. The models were introduced by Hender-

son (1950) for genetics. See Robinson(1991) for an overview. In mixed effects models, we are not interested in just the levels that happen to be in the experiment. Rather, we are concerned with the variability of the effects of all the levels in the population. Consequently, random effects are handled somewhat differently from fixed effects. Here we give some brief discussions. See references books Longford (1993), Davidian and Giltinan (1995) for details. Some of examples of experiment involving random effects are given as follows.

Example 5.1 *This example is extracted from Dean and Voss (1999). Suppose that a manufacture of canned tomato soup wishes to reduce the variability in the thickness of the soup. Suppose that the most likely causes of the variability are the quality of the cornflour (corstarch) recerived from the supplier and the actions of the machine operators. Let us consider two different scenarios:*

Scenario 1: The machine operators are highly skilled and have been with the company for a long time. Thus, the most likely cause of variability is the equality of the cornflour delivered to the company. The treatment factor is cornflour, and its possible levels are all the possible batches of cornflour that the supplier could deliver. Theoretically, this is an infinite population of batches. We are interested not only in the batches of cornflour that have currently been delivered, but also in all those that might be delivered in the future. If we assume that the batches delivered to the company are a random sample from all batches that could be delivered, and if we take a random sample of delivered batches to be observed in the experiment, then the effect of

the cornflour on the thickness is a random effect and can be modeled by a random variable.

Scenario 2: It is known that the quality of the cornflour is extremely consistent, so the most likely cause of variability is due to the different actions of the machine operators. The company is large and machine operators change quite frequently. Consequently, those that are available to take part in the experiment are only a small sample of all operators employed by the company at present or that might be employed in the future. If we can assume that the operators available for the experiment are representative of the population, then we can assume that they are similar to a random sample from a very large population of possible operators, present and future. Since we would like to know about the variability of the entire population, we model the effect of the operators as random variables, and call them random effects.

Example 5.2 *In the last chapter, the earning data were fitted by a fixed effect model*

$$earning \sim eduf + prestige.$$

In this model, we regarded the variability of earnings comes from education and job prestige. It is likely that two persons who have the same education and job have quite different salary. This indicates the variability of earning might come from each individual. Now we treat each individual as a random draw from a population. In this situation, we may use the following mixed effect models

$$earning \sim eduf + \beta_0 + \beta_1 prestige_i + \alpha_{i0} + \alpha_{i1} prestige_i.$$

where $(\alpha_{i0}, \alpha_{i1}) \sim N(\mathbf{0}, D)$. *Compared with the fixed effect model, the mixed effect model is far more flexible. It allows different slopes and intercepts. Hence it reduces modeling bias substantially.*

Compared with the fixed effect model with different slopes and intercepts

$$earning \sim eduf + \beta_{i0} + \beta_{i1} prestige_i,$$

the latter is more flexible, but it has $2n+3+1$ *independent parameters. As a result, the parameters cannot be estimated with good accuracy. Indeed, they are not estimable for this example. On the other hand, the mixed effect model allows varying coefficients, but these coefficients are restricted to be a realization from a normal distribution. The total number of parameter is* $2+3+1+3=9$. *In conclusion, the random effects model enhance the flexibility of the model without introducing many parameters.*

5.2 Random Effects One Way Model

In this section, we study random effects one way model

$$Y_{ij} = \mu + T_i + \varepsilon_{ij}, \quad j = 1, \cdots, n_i, \ i = 1, \cdots, I, \tag{5.1}$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$ and $T_i \sim N(0, \sigma_T^2)$. Further more it is assumed that ε_{ij}'s and T_i's are all mutually independent.

From model (5.1), it follows that

$$Y_{ij} \sim N(\mu, \sigma_T^2 + \sigma^2),$$

and

$$\text{cov}(Y_{ij}, Y_{ik}) = \sigma_T^2.$$

Thus the two components σ_T^2 and σ^2 of the variance of Y_{ij} are known as *variance components*, and observations on the same treatment are positive correlated with correlation coefficient $\rho = \sigma_T^2/(\sigma^2 + \sigma_T^2)$. It is obvious that the maximum likelihood estimates of μ is

$$\widehat{\mu} = \overline{Y}_{\cdot\cdot}.$$

But the maximum likelihood estimates for σ_T^2 and σ^2 are complicate.

5.2.1 Estimation of σ^2 and σ_T^2

In order to make statistical inference about model (5.1), we need estimate σ^2 and σ_T^2. There is no an analytic form for maximum likelihood estimates of σ^2 and σ_T^2. For fixed effects one way ANOVA model, σ^2 is estimated by

$$\widehat{\sigma}^2 = \text{SS}_E/(n-I) = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2.$$

The random effects one way model is very similar to the fixed effect one way ANOVA model, so a natural question is whether the estimate $\widehat{\sigma}^2$ is a good estimate in the random effects one way model. Now we show that $\widehat{\sigma}^2$ is an unbiased estimator of σ^2. In fact, since $Y_{ij} \sim N(\mu, \sigma^2 + \sigma_T^2)$, $EY_{ij}^2 = \mu^2 + \sigma^2 + \sigma_T^2$, and as $\overline{Y}_{i\cdot} = \mu + T_i + \bar{\varepsilon}_{i\cdot}$, $E\overline{Y}_{i\cdot} = \mu^2 + \sigma_T^2 + \frac{1}{n_i}\sigma^2$.

$$E[\text{SS}_E] = E\{\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{I} n_i\overline{Y}_{i\cdot}^2\} = (n-I)\sigma^2.$$

Hence $\widehat{\sigma}^2 = \mathrm{SS}_E/(n-I)$ is an unbiased estimate of σ^2. Further, conditioning on T_i, it follows from the fixed effects one way ANOVA model that

$$(\mathrm{SS}_E|T_1,\cdots,T_n) \sim \sigma^2\chi^2_{n-I}.$$

Hence by taking expectation on both sides,

$$\mathrm{SS}_E \sim \sigma^2\chi^2_{n-I}.$$

Thus we obtain the sampling distribution of $\widehat{\sigma}^2$, and we are able to make statistical inference on it.

Now we are constructing an unbiased estimate for σ_T^2 using similar techniques in deriving the estimate of σ^2. Since

$$\overline{Y}_{\cdot\cdot} = \mu + \frac{1}{n}\sum_i n_i T_i + \bar{\varepsilon}_{\cdot\cdot}.$$

Therefore it follows that

$$E[\overline{Y}^2_{\cdot\cdot}] = \mu^2 + \frac{n_i^2}{n^2}\sigma_T^2 + \frac{1}{n}\sigma^2.$$

Thus

$$\begin{aligned} E[\mathrm{SS}_T] &= E\{\sum_{i=1}^{I} n_i\overline{Y}^2_{i\cdot} - n\overline{Y}^2_{\cdot\cdot}\} \\ &= \left(n - \frac{\sum_i n_i^2}{n}\right)\sigma_T^2 + (I-1)\sigma^2. \end{aligned}$$

Denote

$$c = \frac{n^2 - \sum_i n_i^2}{n(I-1)}$$

which equals to J in balanced design in which $n_i = J$, for $i = 1,\cdots,I$. Thus an unbiased estimator for σ_T^2 is

$$\widehat{\sigma}_T^2 = c^{-1}\left\{\frac{\mathrm{SS}_T}{I-1} - \frac{\mathrm{SS}_E}{n-I}\right\}.$$

It is, unfortunately, possible for the observed value of this estimator to be negative even though σ_T^2 cannot be negative. When the resulting estimate considerably less than 0, the model should be questioned as it is unlikely to be a good description of the data.

5.2.2 Testing Equality of Treatment Effects

When the treatment factor is random, we are interested in the variability of the treatment effects in the entire population levels, not just those in the experiment. Since the variance of the effects in the population is σ_T^2, the null hypothesis of interest is of the form.

$$H_0 : \sigma_T^2 = 0 \qquad \text{versus} \qquad H_1 : \sigma_T^2 > 0.$$

In practice, rather than testing whether or not the variance of the population of treatment is zero, it may be of more interest to test whether the variance is less than or equal to some proportion of the error variance, that is

$$H_0 : \sigma_T^2 \leq \gamma\sigma^2 \qquad \text{versus} \qquad H_1 : \sigma_T^2 > \gamma\sigma^2. \tag{5.2}$$

Intuitively, one should reject the null hypothesis H_0 if the ratio of σ_T^2 to σ^2 is too large. Consequently, we should reject H_0 if

$$\frac{\mathrm{SS}_T/(I-1)}{\mathrm{SS}_E/(n-I)} > a^*. \tag{5.3}$$

It can be shown that

$$\mathrm{SS}_T \sim (c\sigma_T^2 + \sigma^2)\chi_{I-1}^2$$

and SS_T and SS_E are independent. Thus it follows that

$$\frac{\mathrm{SS}_T/[(I-1)(c\sigma_T^2 + \sigma^2)]}{\mathrm{SS}_E/[(I-1)\sigma^2]} \sim F_{I-1,n-I}.$$

Thus take $a^* = (c\gamma + 1)F_{I-1,n-I}(1-\alpha)$. The rejection region of H_0 is

$$F = \frac{\mathrm{SS}_T/(I-1)}{\mathrm{SS}_E/(n-I)} > a^*$$

with level α. In particular, for $\gamma = 0$, the test is $F > F_{I-1,n-I}(1-\alpha)$, which is the same as the fixed effect model. From (5.3), one finds

$$P\{F_{I-1,n-I}(\alpha/2) \le F/[c(\sigma_T^2/\sigma^2+1)] \le F_{I-1,n-I}(1-\alpha/2)\} = 1-\alpha.$$

Solving this equation, we obtain that $(1-\alpha)100\%$ confidence interval for σ_T^2/σ^2 is

$$c^{-1}\left\{\frac{F}{F_{I-1,n-I}(1-\alpha/2)} - 1\right\} \le \frac{\sigma_T^2}{\sigma^2} \le c^{-1}\left\{\frac{F}{F_{I-1,n-I}(\alpha/2)} - 1\right\}.$$

5.3 Mixed Effects Models and BLUP

Mixed effects models contain both random effects and fixed effects. The analysis of random effects proceeds in exactly the same way as described in the previous sections. Consider the mixed effects model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ consists of the n responses, $\mathbf{X}$ and $\mathbf{Z}$ are $n \times p$ and $n \times q$ design matrix, $\boldsymbol{\beta}$ is $p \times 1$ fixed effects unknown parameters, $\mathbf{u}$ is $q \times 1$ random effect. Here it is assumed that $E\mathbf{u} = 0$, $\mathrm{var}(\mathbf{u}) = \sigma^2 G$ and $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 R$. Further we assume that $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are uncorrelated. In this model, we allow the random error $\boldsymbol{\varepsilon}$ heteroscadastic. All examples can be expressed in this form. Natural questions arise here are how to estimate $\boldsymbol{\beta}$ and how to predict $\mathbf{u}$.

Hendenson (1950) assumed that $\mathbf{u}$ and $\mathbf{y}$ have a joint normal distribution, and assumed that G and R are known for the moment. Then the joint density is

$$\begin{aligned} & f(\mathbf{y}, \mathbf{u}) \\ = \; & f(\mathbf{y}|\mathbf{u}) f(\mathbf{u}) \\ = \; & (2\pi\sigma^2)^{-n/2}|R|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T R^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})\} \\ & *(2\pi\sigma^2)^{-q/2}|G|^{-1/2} \exp(-\frac{1}{2}\mathbf{u}^T G^{-1}\mathbf{u}). \end{aligned}$$

The maximum likelihood estimator of $\boldsymbol{\beta}$ and a predictor to $\mathbf{u}$ can be derived via maximizing the likelihood function with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$. This is equivalent to minimizing

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T R^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \mathbf{u}^T G\mathbf{u}$$

with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$. Regarding $\mathbf{u}$ as fixed effects, one would have to minimize

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T R^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}).$$

Comparing with the fixed effect model, we shrink $\mathbf{u}$ towards zero and hence change the estimate of $\boldsymbol{\beta}$ as well. The corresponding normal equations are as follows:

$$\begin{aligned} \mathbf{X}^T R^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}) = 0, \\ \mathbf{Z}^T R^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}) - G\hat{\mathbf{u}} = 0. \end{aligned}$$

Using the identity

$$(R + \mathbf{Z}G\mathbf{Z}^T)^{-1} = R^{-1} - R^{-1}\mathbf{Z}(\mathbf{Z}^T R^{-1}\mathbf{Z} + G^{-1})^{-1}\mathbf{Z}^T R^{-1}, \quad (5.4)$$

and eliminating $\widehat{\mathbf{u}}$, we have

$$\mathbf{X}^T(R+\mathbf{Z}G\mathbf{Z}^T)^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T(R+\mathbf{Z}G\mathbf{Z}^T)^{-1}\mathbf{y}.$$

This implies that

$$\widehat{\boldsymbol{\beta}} = [\mathbf{X}^T(R+\mathbf{Z}G\mathbf{Z}^T)^{-1}\mathbf{X}]^{-1}\mathbf{X}^T(R+\mathbf{Z}G\mathbf{Z}^T)^{-1}\mathbf{y}.$$

This estimator coincides with the generalized least squares estimate in the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\text{var}(\boldsymbol{\varepsilon}) = \mathbf{Z}G\mathbf{Z}^T + R$. Consequently, $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE). Using (5.4) again

$$\widehat{\mathbf{u}} = G\mathbf{Z}^T(\mathbf{Z}G\mathbf{Z}^T + R)^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

Note that $E(\mathbf{u}|\mathbf{y}) = G\mathbf{Z}^T(\mathbf{Z}G\mathbf{Z}^T + R)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Hence $\widehat{\mathbf{u}}$ can be regarded as the plug-in of the best linear predictor. As a result, it can be shown that $\widehat{\boldsymbol{\beta}}$ is BLUE and $\widehat{\mathbf{u}}$ is the best linear unbiased predictor (BLUP). In fact, it can be shown that for any linear unbiased estimator $\mathbf{a}^T\mathbf{y}$ of $\mathbf{b}^T\boldsymbol{\beta} + \mathbf{c}^T\mathbf{u}$. That is

$$E(\mathbf{a}^T\mathbf{y}) = E(\mathbf{b}^T\boldsymbol{\beta} + \mathbf{c}^T\mathbf{u}) \qquad \text{for all} \qquad \boldsymbol{\beta}.$$

We have

$$\text{var}(\mathbf{a}^T\mathbf{y}) \geq \text{var}(\mathbf{b}^T\widehat{\boldsymbol{\beta}} + \mathbf{c}^T\widehat{\mathbf{u}}).$$

See Robinson (1991) for a proof.

Example 5.3 *Consider the random effects one way model*

$$Y_{ij} = \mu + T_i + \varepsilon_{ij},$$

where $T_i \sim N(0, \sigma_T^2)$ *and* $\varepsilon_{ij} \sim N(0, \sigma^2)$. *Further* $\{T_i\}$*'s and* ε_{ij}*'s are independent. For this model, we have*

$$\overline{Y}_{i\cdot} = \mu + T_i + \bar{\varepsilon}_{i\cdot}$$

where $\bar{\varepsilon}_{i\cdot} \sim N(0, \sigma^2/J_i)$. *Thus it can be derived by some straightforward algebraic calculation*

$$\widehat{\mu} = \frac{\sum_i \frac{J_i \overline{Y}_{i\cdot}}{\sigma^2 + J_i \sigma_T^2}}{\sum_i \frac{J_i}{\sigma^2 + J_i \sigma_T^2}}$$

and

$$\widehat{T}_i = \frac{J_i \sigma_T^2}{J_i \sigma_T^2 + \sigma^2} (\overline{Y}_{i\cdot} - \widehat{\mu}).$$

Thus for balanced cases, $J_1 = \cdots = J_I = J$, *we have*

$$\widehat{\mu} = \overline{Y}_{\cdot\cdot}$$
$$\widehat{T}_i = \frac{J \sigma_T^2}{J \sigma_T^2 + \sigma^2} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}).$$

5.4 Restricted Maximum Likelihood Estimator (REML)

Restricted maximum likelihood was proposed by Thompson (1961). It is also called residual or modified maximum likelihood, and used for estimating parameters involving covariance matrices. Thus it can be used to estimate parameters in G if G contains unknown parameters.

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\text{var}(\mathbf{e}) = V(\boldsymbol{\theta})$. The parameters $\boldsymbol{\theta}$ has to satisfy certain constraints in order for $V(\boldsymbol{\theta}) > 0$. For a given $\boldsymbol{\theta}$, when $\mathbf{y}$ is normal, the likelihood is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|V(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T V^{-1}(\boldsymbol{\theta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

For a given $\boldsymbol{\theta}$, maximizing the likelihood function with respect to $\boldsymbol{\beta}$, this leads to

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{X}^T V^{-1}(\boldsymbol{\theta})\mathbf{X})^{-1}\mathbf{X}^T V^{-1}(\boldsymbol{\theta})\mathbf{y},$$

and the profile likelihood is

$$\ell_p(\boldsymbol{\theta}) = \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2}\log|V(\boldsymbol{\theta})| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}))^T V^{-1}(\boldsymbol{\theta})(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})).$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ is to minimize

$$-2\ell_p(\boldsymbol{\theta}) = \log|V(\boldsymbol{\theta})| + \mathbf{y}^T V^{-1}(\boldsymbol{\theta})[I - Q(\boldsymbol{\theta})]\mathbf{y},$$

where $Q(\boldsymbol{\theta}) = \mathbf{X}(\mathbf{X}V^{-1}(\boldsymbol{\theta})\mathbf{X})^{-1}\mathbf{X}^T V^{-1}(\boldsymbol{\theta})$. It is well known that the maximum likelihood estimates of parameters in the covariance matrix are biased. To adjust for degrees of freedom lost in estimating $\boldsymbol{\beta}$, one minimize the modified likelihood

$$-2\ell_r(\boldsymbol{\theta}) = \log|V(\boldsymbol{\theta})| + \log|\mathbf{X}^T V^{-1}(\boldsymbol{\theta})\mathbf{X}| + \mathbf{y}^T V^{-1}(\boldsymbol{\theta})[I - Q(\boldsymbol{\theta})]\mathbf{y}$$

subject to constraints on $\boldsymbol{\theta}$. The $\ell_r(\boldsymbol{\theta})$ is called "restricted likelihood".

Example 5.4 *Consider the ordinary linear regression model*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $E\boldsymbol{\varepsilon} = 0$ *and* $var(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. *In this case*

$$-2\ell_p(\sigma^2) = \log\sigma^{2n} + \sigma^{-2}RSS,$$

where $RSS = \mathbf{y}^T(I - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$. *So* $\widehat{\sigma}^2_{MLE} = \frac{1}{n}RSS$, *which is biased estimator. On the other hand,*

$$\begin{aligned} -2\ell_r(\sigma^2) &= \log \sigma^{2n} + \log|\sigma^{-2}\mathbf{X}^T\mathbf{X}| + \sigma^{-2}RSS \\ &= (n-p)\log\sigma^2 + \sigma^{-2}RSS + \log|\mathbf{X}^T\mathbf{X}| \end{aligned}$$

Hence

$$\widehat{\sigma}^2_{RMLE} = \frac{RSS}{n-p}.$$

Example 5.5 *(Neyman-Scott Problem) Consider one way ANOVA model*

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad j = 1, \cdots, J \quad i = 1, \cdots, n,$$

where ε_{ij} *are independent and identically distributed* $N(0, \theta)$. *So the profile likelihood is*

$$-2\ell_p(\theta) = \log\theta^n + \theta^{-1}RSS,$$

where $RSS = \sum_i \sum_i (Y_{ij} - \overline{Y}_{i\cdot})^2$. *Hence* $\widehat{\theta}_{MLE} = \frac{1}{nJ}RSS$, *which is a biased estimator. Indeed,*

$$E\widehat{\theta}_{MLE} = \frac{1}{nJ}n(J-1)\theta = \frac{J-1}{J}\theta.$$

The estimator is not consistent. On the other hand

$$-2\ell_r(\theta) = \log\theta^{nJ} - n\log\theta + \log|\mathbf{X}^T\mathbf{X}| + \theta^{-1}RSS$$

which implies that

$$\widehat{\theta}_{RMLE} = \frac{1}{n(J-1)}RSS.$$

This is an unbiased and consistent estimator.

Example 5.6 *(Random effects one way model) Consider the random effects one way model*

$$Y_{ij} = \mu + T_i + \varepsilon_{ij}, \quad i = 1, \cdots, I, \ \text{and} \ \ j = 1, \cdots, J,$$

where T_i's are independent and identically distributed $N(0, \sigma_T^2)$ and ε_{ij}'s are independent and identically distributed $N(0, \sigma^2)$. Further assume that they are all mutually independent. REML gives the same estimator as that given in Section 5.2. For balanced design,

$$\widehat{\sigma}^2_{RMLE} = MS_E$$

and

$$\widehat{\sigma}^2_T = MS_T - MS_E$$

On the other hand, the maximum likelihood estimate is biased.

5.5 Estimation of Parameters in Mixed Effects Models

Now assume that G and R depend on unknown $\boldsymbol{\theta}$. The estimating procedure is schematically summarized as follows:

Step 1. Use REML to estimate $\widehat{\boldsymbol{\theta}}$.

Step 2. Obtain $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$ and $\widehat{\mathbf{u}}(\widehat{\boldsymbol{\theta}})$ which are empirical the best linear unbiased estimator and the linear unbiased predictor respectively. Further, the covariance matrix can be estimated from the usual maximum likelihood method.

For comparing two nested mixed effect models, one uses the likelihood ratio statistics

$$T = 2\{\ell_r(H_1) - \ell_r(H_0)\}$$

and reject H_0 if

$$T > \chi^2_d(1-\alpha)$$

where $d = dim(H_1) - dim(H_0)$.

5.6 Examples

In this section, we illustrate analysis of mixed effect models by the Female Supply Labor data. The data set was fitted by the model

$$\begin{aligned} \text{earning}_i &= \mu + \beta_0 \text{prestige}_i + \sum_{j=1}^{4} \beta_j I(eduf = j) \\ &\quad + A_i + B_i \text{prestige}_i + \varepsilon_i, \end{aligned} \tag{5.5}$$

where μ and β_is are fixed effects, and A_i and B_i are random effects which are normally distributed with mean 0 and covariance D. From S-plus output, the value of the log-likelihood at the maximum likelihood estimate is

$$\ell_{R1} = -1614.309.$$

Of interest here is to test $H_0 : B_i = 0$. Thus consider the reduced model

$$\text{earning}_i = \mu + \beta_0 \text{prestige}_i + \sum_{j=1}^{4} \beta_j I(eduf = j) + A_i + \varepsilon_i. \tag{5.6}$$

From S-plus output, we have

$$\ell_{R2} = -1671.3.$$

The likelihood ratio test statistic is

$$T = 2\{\ell_{R1} - \ell_{R2}\} = 113.91$$

The likelihood ratio statistic has a χ^2-square distribution with 2 degrees of freedom. Thus the corresponding p-value $\approx$0. So we reject the null hypothesis and model (5.5) should be used.

The estimated fixed effects are displayed as follows

	Value	Approx. Std.Error	z ratio(C)
(Intercept)	6.982	0.8143	8.581
prestige	0.120	0.0140	8.518
edufB−edufA	-1.924	0.569	-3.379
edufC−edufA	-2.421	0.570	-4.250
edufD−edufA	-2.891	0.593	-4.873

Appendix: S-plus codes and outputs

```
>labor <- read.table("d:/rli/book/labor.dat",header=T)
>earning <- labor[,3]
>edu <- labor[,5]
>eduf <- rep("A",length(edu))
>eduf[edu<16]<-"B"
>eduf[edu<13]<-"C"
>eduf[edu<12]<-"D"
>prestige <- labor[,4]
>subj <- seq(1,607,1)
>labor.df <- data.frame(earning, eduf, prestige, subj)
>options(contrasts=c("contr.treatment","contr.treatment"))
>library(nlme2,first=T)
```

```
> fit1 <- lme(fixed= earning ~ prestige+ eduf,
+ random=~1+prestige,cluster=~subj,data=labor.df)

> summary(fit1)
Call:
  Fixed: earning ~ prestige + eduf
 Random:  ~ 1 + prestige
Cluster:  ~ subj
   Data: labor.df

Estimation Method: RML
Convergence at iteration: 22
Restricted Loglikelihood: -1614.309
Restricted AIC: 3246.619
Restricted BIC: 3286.295

Variance/Covariance Components Estimate(s):
  Structure: unstructured
  Parametrization: matrixlog
  Standard Deviation(s) of Random Effect(s)
 (Intercept)  prestige
    3.954297 0.1562036
 Correlation of Random Effects
         (Intercept)
prestige -0.99952

 Cluster Residual Variance: 4.68764

Fixed Effects Estimate(s):
                  Value Approx. Std.Error z ratio(C)
(Intercept)  6.9820033        0.81369533   8.580611
   prestige  0.1196804        0.01404968   8.518374
      edufB -1.9239213        0.56941136  -3.378790
      edufC -2.4209602        0.56969944  -4.249539
      edufD -2.8913486        0.59338502  -4.872635

 Conditional Correlation(s) of Fixed Effects Estimates
         (Intercept)   prestige      edufB      edufC
prestige -0.7808980
   edufB -0.6517700   0.1205947
```

```
   edufC -0.7730582   0.2762747  0.8297279
   edufD -0.7988700   0.3378162  0.8053599  0.8575645

Random Effects (Conditional Modes):
       (Intercept)       prestige
  1  2.008246e+000 -7.943335e-002
  2  6.764489e-001 -2.691101e-002
  3 -2.346625e+000  9.300137e-002
  4 -2.005403e-002  7.935912e-004
  5 -1.834950e+000  7.257029e-002
  .
  .
  .

603  6.991530e-001 -2.773203e-002
604 -1.060450e+000  4.192961e-002
605 -4.738697e+000  1.874799e-001
606  6.946953e+000 -2.748225e-001
607 -7.125518e-001  2.795519e-002


Standardized Population-Average Residuals:
       Min         Q1         Med        Q3      Max
 -2.861871 -0.3486675 -0.01130373 0.2635609 6.324993

Number of Observations: 607
Number of Clusters: 607
>

> fit2 <- lme(fixed= earning ~ prestige+ eduf, random=~1,
+  cluster=~subj,data=labor.df)
> summary(fit2)
Call:
  Fixed: earning ~ prestige + eduf
 Random:  ~ 1
Cluster:  ~ subj
   Data: labor.df

Estimation Method: RML
Convergence at iteration: 1
```

```
Restricted Loglikelihood: -1671.267
Restricted AIC: 3356.533
Restricted BIC: 3387.393

Variance/Covariance Components Estimate(s):
  Structure: unstructured
  Parametrization: matrixlog
  Standard Deviation(s) of Random Effect(s)
 (Intercept)
    3.651806

 Cluster Residual Variance: 1.006645

Fixed Effects Estimate(s):
                 Value Approx. Std.Error z ratio(C)
(Intercept)  7.5305753         0.9148366   8.231607
   prestige  0.1194338         0.0145112   8.230456
      edufB -2.3372648         0.5372856  -4.350135
      edufC -2.9599503         0.5793238  -5.109320
      edufD -3.4787089         0.6360176  -5.469516

 Conditional Correlation(s) of Fixed Effects Estimates
         (Intercept)   prestige      edufB      edufC
prestige -0.8690149
   edufB -0.6115981   0.2241098
   edufC -0.7509996   0.4193302  0.7522324
   edufD -0.7604445   0.4698535  0.7048789  0.7530959

Random Effects (Conditional Modes):
      (Intercept)
  1  -2.669014887
  2  -1.354223805
  3   2.667284763
  4   0.022581953
  5   2.604090005
  .
  .
  .
603  -0.916337727
604   1.704892284
```

```
605   5.322124989
606  -8.803717773
607  -1.561537897


Standardized Population-Average Residuals:
        Min         Q1         Med        Q3     Max
 -0.6722075 -0.1226388 -0.01158055 0.1101916 4.00266

Number of Observations: 607
Number of Clusters: 607

> anova(fit1,fit2)
Response: earning
fit1
 fixed: (Intercept), prestige, edufB, edufC, edufD
 random: (Intercept), prestige
 block: list(1:2)
 covariance structure: unstructured
 serial correlation structure: identity
 variance function: identity
fit2
 fixed: (Intercept), prestige, edufB, edufC, edufD
 random: (Intercept)
 block: list(1:1)
 covariance structure: unstructured
 serial correlation structure: identity
 variance function: identity
     Model Df    AIC    BIC  Loglik    Test Lik.Ratio P value
fit1     1  9 3246.6 3286.3 -1614.3
fit2     2  7 3356.5 3387.4 -1671.3 1 vs. 2    113.91       0
```

6

Introduction to Generalized Linear Models

Generalized linear models provide a unified approach to many of the most common statistical procedures used in applied statistics. They have many applications in various scientific research fields, such as medicine, engineering, psychology, etc. A comprehensive account of generalized linear models can be found in McMullagh and Nelder (1989). Also see Lindsey (1997) for many interesting applications. This chapter will extend techniques of the classical linear models to handle different types of responses. We will focus on three types of responses: count, categorical and continuous responses.

6.1 Introduction

In many applications, response variables can be either discrete or continuous. Let us examine some motivating examples first.

Example 6.1 *These data were collected by the General Hospital Burn Center at the University of Southern California. The data set consists of 981 observations. For each observation, many variables were collected. Among those the binary response variable Y is 1 for those victims who survived their burns and 0 otherwise; and many covariates, such as age, sex, race, weight, ratio of the third degree burned and type of burn, so on. It is of interest to identify the risk factors and their risk contribution. Because the response is binary categorical data, classical linear regression model is not appropriate for this type of data. A natural model for this data set is to regard the data as realizations from the following model:*

$$P\{Y = 1|X_1 = x_1, \cdots, X_p = x_p\} = p(x_1, \cdots, x_p).$$

This is equivalent to assuming that the conditional distribution of Y given $X_1, \cdots, X_p$ is a Bernoulli distribution with success probability $p(X_1, \cdots, X_p)$. To make statistical inferences on this model, we need to know an appropriate forms for $p(x_1, \cdots, x_p)$ and how to fit the model. Furthermore, we are also interested in which variables are significant and what the raw materials for model diagnostic are.

Example 6.2 *Altman (1991, p.199) provides count of T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkins's disease and 20 other patients in remission from disserminated malignancies, as shown in Table 6.1 and taken from Lindsey (1997). Of interest is to examine whether there is a significant difference in cell counts between the two diseases. To this end, one could use the technique in one way analysis of variance model. Since the response are*

counts, a more sophisticated method might be to assume a Poisson distribution of the counts within each group. It is natural to use differences in logarithm of the means, so that we are looking at the difference between the means, themselves, through a ratio instead of by subtraction. However, this model also carries the additional assumption that the variability will be different between the two groups if the mean is, because the variance of a Poisson distribution is equal to its mean.

TABLE 6.1. T_4 cell counts

Hodgkin	Non-Hodgkin
396	375
568	375
1212	752
171	208
554	151
1104	116
257	736
435	192
295	315
397	1252
288	675
1004	700
431	440
795	771
1621	688
1378	426
902	410
958	979
1283	377
2415	503

The above two examples specify the conditional distribution of Y given $X_1, \cdots, X_n$. Further, they admits the conditional variance of

response variable depending on its conditional mean. Just like that one may apply the normal likelihood method to all homoscedastic cases, he may apply the Bernoulli likelihood to all the data with variance $p(1-p)$, and use Poisson likelihood to all the data with variance equaling to its mean, so on. Such as kind of method is called quasi-likelihood method.

6.2 Elements of Generalized Linear Models

Each generalized linear model consists of three components: response distribution, linear predictor and link function. The latter two are used to model the regression function. Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \cdots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$. For this model, we have already looked at two of the three components: the conditional distribution of Y given X and the linear structure. The third component in this simple model in fact is identical link. Let us look at all three more more closely.

6.2.1 Modeling regression functions

The regression function is defined by

$$m(x_1, \cdots, x_p) = E(Y|X_1 = x_1, \cdots, X_p = x_p).$$

It is the best predictor of Y given $X_1 = x_1, \cdots, X_p = x_p$. One can always write a regression model as

$$Y = m(X_1, \cdots, X_p) + \varepsilon \qquad \text{with} \qquad E(\varepsilon|\mathbf{x}) = 0.$$

This model treats the part that cannot be explained by $m(X_1, \cdots, X_p)$ as random error. In classical linear models, it is assumed that

$$m(x_1, \cdots, x_p) = \beta_1 x_1 + \cdots \beta_p x_p.$$

It leads to the liner model

$$Y = \beta_1 X_1 + \cdots \beta_p X_p + \varepsilon.$$

This is the most crucial part of the assumption and needs to verified. When the parameters $\beta_1, \cdots, \beta_p$ varies from $-\infty$ to $+\infty$. So does the linear function $\beta_1 x_1 + \cdots + \beta_p x_p$. Thus the minimum requirement is that $\mu = m(x_1, \cdots, x_p)$ has range $(-\infty, +\infty)$. This condition is not satisfied for the mean functions in Examples 6.1 and 6.2. In this situation, we transform μ by a function g and model this transformed mean by a liner model

$$g(\mu) = \sum_{j=1}^{p} \beta_j x_j \equiv \eta.$$

The minimum requirement is that $g(\mu) \in (-\infty, +\infty)$ and g is strictly monotone. With these assumptions,

$$\mu = m(x_1, \cdots, x_p) = g^{-1}(\eta) = g^{-1}(\sum_{j=1}^{p} \beta_j x_j).$$

This is the most critical assumption in generalized linear model. We call g as a *link function*, linking the regression to the linear predictor.

Different link functions correspond to different models. In practice, link functions can be chosen by users, or chosen to be the corresponding canonical link for mathematical convenience (see Section 6.2.2 for the definition of canonical link).

Example 6.3 *(Bernoulli Model) When the response Y is binary, it is usual to assume that the conditional distribution of Y given $X_1, \cdots, X_p$ is a Bernoulli distribution with success probability $p(x_1, \cdots, x_p)$. Thus the mean regression function is*

$$\mu = E(Y|X_1, \cdots, X_p) = p(X_1, \cdots, X_p) \in [0, 1].$$

Since the range of μ is [0,1], we are looking for a link function defined on [0,1] and its range is $(-\infty, +\infty)$. Intuitively, any continuous strictly increasing distribution can be used to construct a link function for the Bernoulli model. In fact, suppose unobserved variable z follows the linear model

$$z = -\beta_1 x_1 - \cdots - \beta_p x_p + \varepsilon,$$

where $\varepsilon \sim F$. We observed $Y = I(z \leq \tau)$ according to certain thresholding parameter τ. Then

$$(Y|X_1, \cdots, X_p) \sim Bernoulli\{p(x_1, \cdots, x_p)\}$$

with

$$p(x_1, \cdots, x_p) = F(\tau + \beta_1 x_1 + \cdots \beta_p x_p).$$

If there is an intercept term in the linear model, we can assume that $\tau = 0$. Hence

$$p(x_1, \cdots, x_p) = F(\beta_1 x_1 + \cdots \beta_p x_p).$$

(a) If ε is distributed a logistic distribution, then

$$p(x_1, \cdots, x_p) = \frac{\exp(\beta_1 x_1 + \cdots \beta_p x_p)}{1 + \exp(\beta_1 x_1 + \cdots \beta_p x_p)}.$$

This gives us a logit link:

$$g(\mu) = \log \frac{\mu}{1-\mu} = \beta_1 x_1 + \cdots + \beta_p x_p.$$

The logit link is the canonical link for Bernoulli distribution. Therefore it is important in modeling binary response data. In many fields, researchers are interested in examining odds ratio of different treatments. Thus logit link becomes the most popular link function in modeling binary responses as it directly gives us odds, defined as $p/(1-p)$, which is equal to $\exp(\beta_1 x_1 + \cdots + \beta_p x_p)$.

(b) If ε is normal distribution $N(0, \sigma^2)$, then

$$p(x_1, \cdots, x_p) = \Phi\left(\frac{\beta_1 x_1 + \cdots + \beta_p x_p}{\sigma}\right) \widehat{=} \Phi(\alpha_1 x_1 + \cdots + \alpha_p x_p).$$

This leads to probit link: $\Phi^{-1}(u) = \alpha_1 x_1 + \cdots \alpha_p x_p$.

(c) Another commonly used link function is complementary log-log link *:*

$$\eta = \log\{-\log(1-\mu)\} = \beta_1 x_1 + \cdots + \beta_p x_p$$

which is equivalent to

$$p(x_1, \cdots, x_p) = 1 - \exp\{-\exp(\beta_1 x_1 + \cdots \beta_p x_p)\}.$$

Example 6.4 *When responses are counts, we often assume that the conditional distribution of response has a Poisson distribution with*

mean $\lambda(X_1, \cdots, X_p)$. *The mean function of Poisson distribution is equal to the* $\lambda(X_1, \cdots, X_p)$ *whose range is* $[0, \infty)$. *Log-link function is frequently used to model counting data. Thus*

$$\eta = \log(\mu) = \beta_1 x_1 + \cdots + \beta_p x_p.$$

and

$$\lambda(x_1, \cdots, x_p) = \exp(\beta_1 x_1 + \cdots + \beta_p x_p).$$

In fact, the log-link *is the canonical link of Poisson regression. Therefore Poisson regression model is referred to as* log-linear model *in some literature.*

6.2.2 Conditional distributions

In generalized linear models, it is assumed that the conditional density or probability function of Y given $\mathbf{X} = \mathbf{x}$ belongs to a *canonical exponential family*

$$f(y|\mathbf{x}) = \exp([\theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}]/a(\phi) + c(y, \phi)) \tag{6.1}$$

for some known functions $a(\cdot), b(\cdot)$ and $c(\cdot, \cdot)$. The parameter $\theta(\cdot)$ is called a *canonical parameter* and ϕ is called a *dispersion parameter* . The exponential family includes many commonly used distributions, such as normal distributions, binomial distributions, Poisson distributions and gamma distributions. We now illustrate model (6.1) by some useful examples.

Example 6.5 *Suppose that the conditional density of* Y *given* $\mathbf{X} = \mathbf{x}$ *is* $N(\mu(\mathbf{x}), \sigma^2)$, *then by writing the normal density as*

$$f(y; \mu, \sigma) = \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma\right).$$

Here $\theta = \mu,\ a(\phi) = \sigma^2$, $b(\theta) = \theta^2/2$ *and* $c(y,\phi) = -\frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$. *As shown in previous chapters, this model is useful for continuous response with homoscedastic errors. The canonical link function is the identity link* $g(t) = t$.

Example 6.6 *If the conditional distribution of* Y *given* $\mathbf{X} = \mathbf{x}$ *is binomial* $b(m, p(\mathbf{x}))$ $(0 < p(\mathbf{x}) < 1)$, *then its probability function is for* $y = 0, \cdots, m$,

$$\begin{aligned} P(Y = y|\mathbf{X} = \mathbf{x}) &= \binom{m}{y} p^y (1-p)^{m-y} \\ &= \exp\{y \log \frac{p}{1-p} + m \log(1-p) + \log \binom{m}{y}\}. \end{aligned}$$

Here $\theta = \log \frac{p}{1-p}$ *is a canonical parameter,* $b(\theta) = m \log(1-p)$ *and* $c(y,\phi) = \log \binom{m}{y}$ *with* $\phi = 1$. *This model is commonly used for situations with a binary response.*

Example 6.7 *Suppose that the conditional distribution of* Y *given* $\mathbf{X} = \mathbf{x}$ *is a Poisson distribution with mean* $\lambda(\mathbf{x})$, *then the conditional probability function is*

$$\begin{aligned} P(Y = y|\mathbf{X} = \mathbf{x}) &= \frac{\lambda^y \exp(-\lambda)}{y!} \\ &= \exp(y \log \lambda - \lambda - \log y!). \end{aligned}$$

Therefore $\theta = \log \lambda$ *is a canonical parameter,* $b(\theta) = \lambda = \exp(\theta)$ *and* $c(y,\phi) = \log y!$ *with* $\phi = 1$. *This model is useful for situations in which the response variable is a counting variable with mean and variance approximately the same.*

Example 6.8 *Assume that the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ is the gamma distribution with mean $\mu(\mathbf{x})$ and shape parameter α. Then, its conditional density possesses the form*

$$\exp\{\alpha(y/\mu - \log\mu) + (\alpha - 1)\log y + \alpha\log\alpha - \log\Gamma(\alpha)\}.$$

For this example, $\phi = \alpha$, $\theta = -1/\mu$, $a(\phi) = 1/\alpha$, $b(\theta) = -\log(-\theta)$. This model is useful for continuous response with constant coefficient of variation.

For the canonical exponential family, we have the following moment properties.

Theorem 6.1 *If the distribution of Y belongs to the canonical exponential family (6.1), then*

(i) $\mu = E(Y) = b'(\theta)$;

(ii) $var(Y) = a(\phi)b''(\theta)$.

Proof: Note that

$$\int f(y, \theta, \phi)\, d\nu(y) = 1.$$

Taking the derivative with respect to θ, we have

$$\int \frac{\partial f(y,\theta,\phi)}{\partial\theta}\, d\nu(y) = \int \frac{\partial \log f(y,\theta,\phi)}{\partial\theta} f(y,\theta,\phi)\, d\nu(y).$$

Denote by $\ell(\theta)$ the log likelihood, then

$$E_\theta \ell'(\theta) = 0.$$

This is Bartlette's first identity, which gives us the consistency of maximum likelihood estimates of θ. Taking derivative $E_\theta\ell'(\theta)$ with

respect to θ again,

$$\int \frac{\partial^2 \log f(y,\theta,\phi)}{\partial \theta^2} f(y,\theta,\phi)\, d\nu(y) + \int \left(\frac{\partial \log f(y,\theta,\phi)}{\partial \theta}\right)^2 f(y,\theta,\phi)\, d\nu(y) = 0.$$

That is

$$E_\theta \ell''(\theta) + E\{\ell'(\theta)\}^2 = 0.$$

This is Bartlette's second identity. For the canonical exponential family,

$$\ell(\theta) = \{Y\theta - b(\theta)\}/a(\phi) + c(y,\phi).$$

From the Bartlette's first identity

$$E_\theta \ell'(\theta) = E_\theta\{Y - b'(\theta)\}/a(\phi) = 0.$$

which implies that (i) holds.

From the second Bartlette identity

$$E_\theta\{-b''(\theta)/a(\phi)\} + E_\theta\{Y - b'(\theta)\}^2/a^2(\phi) = 0.$$

Hence

$$\mathrm{var}(Y) = a(\phi)b''(\theta).$$

This completes the proof.

It is not difficult to verify Theorem 6.1 for the normal, binomial, Poisson and gamma distribution in Example 6.5 to 6.8.

Definition The function g that links the mean to the canonical parameter is called canonical link

$$g(\mu) = \theta$$

Since $\mu = b'(\theta)$, the canonical link is

$$g(\mu) = (b')^{-1}(\mu).$$

Table 6.2 summarizes canonical link functions for normal, binomial, Poisson and gamma distributions.

TABLE 6.2. Canonical Links

Distribution	$b(\theta)$	Canonical Link	Range
Normal	$\theta^2/2$	$g(\mu) = \mu$	$(-\infty, +\infty)$
Poisson	$\exp(\theta)$	$g(\mu) = \log \mu$	$(-\infty, +\infty)$
Binomial	$-\log(1 - e^{\theta})$	$g(\mu) = \log \frac{\mu}{1-\mu}$	$(-\infty, +\infty)$
Gamma	$-\log(-\theta)$	$g(\mu) = -\frac{1}{\mu}$	$(-\infty, 0)$

Theorem 6.2 *If the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ belongs to the canonical exponential family (6.1), then it holds that*

(i) The canonical link function is strictly increasing;

(ii) The likelihood $\ell(\theta)$ using canonical parameter is strictly concave.

Proof: Note that

$$\frac{dg(\mu)}{d\theta} = \frac{1}{b''(\theta)}.$$

Since $\text{var}(Y) = a(\phi)b''(\theta) > 0$, so $b''(\theta) > 0$. Thus $g(\mu)$ is a strictly increasing.

As to (ii), note that

$$\ell''(\theta) = -\frac{b''(\theta)}{a(\phi)} < 0.$$

Thus $\ell''(\theta)$ is strictly concave. This completes the proof.

Remark: If the canonical link function is used, then the likelihood function $\ell(\theta)$ is strictly concave. Hence maximum likelihood estimate is unique. On the other hand, if other parameterization is used, the likelihood function may not be strictly concave. In this case, there may exist several local maximizers.

6.3 Maximum Likelihood Methods

6.3.1 Maximum Likelihood Estimate and Estimated Standard Errors

Suppose that $(\mathbf{x}_i, y_i), i = 1, \cdots, n$ are independent and identically distributed, and the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ belongs to the canonical exponential family

$$f(y_i|\mathbf{x}_i, \theta_i, \phi) = \exp\{(y_i\theta_i - b(\theta_i))/a_i(\phi) + c(y_i, \phi)\}.$$

Here the mean μ_i is related with θ_i via

$$\theta_i = (b')^{-1}(\mu_i) \equiv \theta(\mu_i)$$

and μ_i is related with $\mathbf{x}_i$ through a link function g

$$g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}.$$

Hence, $(\theta_1, \cdots, \theta_n)$ are only a p-dimensional manifold, called Θ. To accommodate the heteroscedasticity, we will assume that

$$a_i(\phi) = \frac{\phi}{w_i}.$$

For example, if y_i is the average of n_i data points with covariate $\mathbf{x}_i$, the one takes $w_i = n_i$. For the generalized linear models,

$$\theta_i = \theta\{g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})\} \widehat{=} h(\mathbf{x}_i^T\boldsymbol{\beta}).$$

The log-likelihood function is

$$\ell_n(\theta, \phi, ; \mathbf{y}, \mathbf{x}) = \sum w_i\{y_i\theta_i - b(\theta_i)\}/\phi$$

except a constant term. The statistical inference is based on the maximum likelihood estimate. For simplicity of notation, denote

$$\ell(\boldsymbol{\beta}) = \ell_n(\theta, \phi; \mathbf{y}, \mathbf{x})$$

for the model $\theta_i = \theta(g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})) = h(\mathbf{x}_i^T\boldsymbol{\beta})$. Then, the maximum likelihood estimate solves the following likelihood equations

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \ell'(\widehat{\boldsymbol{\beta}}) = 0.$$

By Taylor's expansion around the true parameter $\boldsymbol{\beta}_0$,

$$0 = \ell'(\widehat{\boldsymbol{\beta}}) \approx \ell'(\boldsymbol{\beta}_0) + \ell''(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

Hence

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \approx [\ell''(\boldsymbol{\beta}_0)]^{-1}\ell'(\boldsymbol{\beta}_0) = [\frac{1}{n}\ell''(\boldsymbol{\beta}_0)]^{-1}\frac{1}{n}\ell'(\boldsymbol{\beta}_0). \tag{6.2}$$

Note that for the independent and identically distributed samples (assume $w_i = 1$ for simplicity) of discussions,

$$\frac{1}{n}\ell''(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i\theta_i - b(\theta_i)\}''/\phi = \frac{1}{n}\sum_{i=1}^{n} b''(h(\mathbf{x}_i\boldsymbol{\beta}))\mathbf{x}_i\mathbf{x}_i^T/\phi.$$

This tends to $Eb''(\theta_i)\mathbf{x}\mathbf{x}^T/\phi$ in probability as $n \to \infty$, where $b''(\theta_i) = b''(h(\mathbf{x}_i^T\boldsymbol{\beta})$. So the variance comes from $\ell'(\boldsymbol{\beta}_0)$, which is given by

$$\ell'(\boldsymbol{\beta}_0) = \sum_{i=1}^{n}\{Y_i - b'(\theta_0)\}h'(\mathbf{x}_i^T\boldsymbol{\beta}_0)\mathbf{x}_i.$$

Hence,

$$\text{var}\{\ell'(\boldsymbol{\beta}_0)|\mathbf{x}\} = \sum_{i=1}^{n}\{h'(\mathbf{x}_i^T\boldsymbol{\beta})\}^2 V(\mu_i)\mathbf{x}_i\mathbf{x}_i^T.$$

This can be estimated by the substitution method, denote by the resulting estimator by

$$\widehat{\Sigma} = \text{var}\{\ell'(\boldsymbol{\beta}_0)|\mathbf{x}\}.$$

From (6.2), the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is given by approximated by

$$\text{avar}(\widehat{\boldsymbol{\beta}}) \approx [\ell''(\boldsymbol{\beta}_0)]^{-1}\text{var}(\ell'(\boldsymbol{\beta}_0)|\mathbf{x})[\ell''(\boldsymbol{\beta}_0)]^{-1}.$$

It can be estimated as

$$\widehat{\text{avar}}(\widehat{\boldsymbol{\beta}}) = [\ell''(\widehat{\boldsymbol{\beta}})]^{-1}\widehat{\Sigma}[\ell'(\widehat{\boldsymbol{\beta}})]^{-1}.$$

which is the corresponding sandwich formula, also referred as a robust estimator of variance-covariance matrix. Hence, one obtains estimated covariance matrix, standard errors and correlations.

Example 6.9 *(Kyphosis data) The data frame kyphosis in S-plus consists of measurements on 81 children following corrective spinal surgery (see, for example, Chambers and Hastie, 1993). The binary response variable* Kyphosis *indicates the presence or absence of a postoperative deforming (called Kyphosis). There are three covariates in this data set: Age of the child in month, Number of vertebrae involved in operation, and Start, the beginning of the range of vertbrae involved. Of interest is to examine whether the three covariates relate to the response. The data set was fitted by a simple linear logistic model involving all three covariates as predictors*

$$logit\{P(Y = 1|X_1 = x_1, \cdots, X_p = x_p)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

with $\phi = 1$. The estimated regression coefficients $\widehat{\boldsymbol{\beta}}$ and their estimated variance covariance matrix were obtained. Table 6.3 depicts the estimated coefficients and their standard errors.

TABLE 6.3. Estimated Coefficients and Standard Errors in Example 6.10

	$\widehat{\beta}_i$	S.E.	t-value
Intercept	-2.037	1.449	-1.405
Age	0.0109	0.00644	1.696
Start	-0.207	0.0677	-3.05
Number	0.411	0.225	1.827

From Table 6.3, it seems that only the variable Start appears significant. Note that

$$logit\{p(x_1, x_2, x_3)\} = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

it follows that

$$\frac{p(x_1, x_2 + 1, x_3)/(1 - p(x_1, x_2 + 1, x_3))}{p(x_1, x_2, x_3)/(1 - p(x_1, x_2, x_3))} = \exp\{\beta_2\}.$$

This gives us very good interpretation: given the levels of Age and Start, odds ratio is $\exp(-0.207) = 0.8130$ *as the covariate Number increase 1.*

Example 6.10 *(Wave-soldering data) In this example, we apply Poisson log-linear regression model to wave-soldering data, which have been analyzed in Chambers and Hastie (1993), using several models. In 1988, an experiment was designed and implemented at one of AT&T's factories to investigate alternatives in the "wave-soldering" procedure for mounting electronic components on printed relevant to the engineering of wave-soldering. The response, measured by eye,*

is a count of the number of visible solder skips for a board soldered under a particular choice of levels for the experimental factors. The data consist of 900 observations of the response skips in a balanced subset of all the experimental runs, with the corresponding values of five experimental factors. The data set is available in S-plus package, named solder.balance. The paper by Comizzoli, Landwehr, and Sinclair (1990) gives a readable general discussion. Here is a brief description of the factors:

Opening: amount of clearance around the mounting pad;

Solder: amount of solder;

Mask: type and thickness of the material used for the solder mask;

PadType: the geometry and size of the mounting pad; and

Panel: each board was divided into three panels, with three runs on a board.

A Poisson log-linear regression was fitted to this data set involving all linear terms

$$\log\{\lambda(Opening, Solder, Mask, PadType, Panel)\} \sim Opening + Solder + Mask + PadType + Panel$$

They are many ways to parameterize the problem. Like in the ANOVA, we report estimable contrasts and their estimated standard errors in Table 6.4.

TABLE 6.4. Estimated Contrasts and Standard Errors

	$\widehat{\beta}_i$	SE	t value
(Intercept)	0.7357	0.0295	24.9548
Opening.L	-1.3389	0.0379	-35.3286
Opening.Q	0.5619	0.0420	13.3778
Solder	-0.7776	0.0273	-28.4742
Mask1	0.2141	0.0377	5.6761
Mask2	0.3294	0.0165	19.9285
Mask3	0.3307	0.0089	36.9702
PadType1	0.0550	0.0332	1.6570
PadType2	0.1058	0.0173	6.1034
PadType3	-0.1049	0.0152	-6.9157
PadType4	-0.1229	0.0136	-9.0317
PadType5	0.0131	0.0089	1.4780
PadType6	-0.0466	0.0088	-5.2752
PadType7	-0.0076	0.0070	-1.0871
PadType8	-0.1355	0.0106	-12.7861
PadType9	-0.0283	0.0066	-4.3096
Panel1	0.1668	0.0210	7.9305
Panel2	0.0292	0.0117	2.4876

*6.3.2 Computation**

The computation method for generalized linear model relies on iteratively reweighted least squares. Suppose that we want to solve the likelihood equation

$$\ell'(\widehat{\boldsymbol{\beta}}) = 0.$$

Given an initial value $\boldsymbol{\beta}_0$ of $\widehat{\boldsymbol{\beta}}$, by Taylor's expansion

$$0 = \ell'(\widehat{\boldsymbol{\beta}}) \approx \ell'(\boldsymbol{\beta}_0) + \ell''(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

Hence

$$\widehat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}_0 - [\ell''(\boldsymbol{\beta}_0)]^{-1}\ell'(\boldsymbol{\beta}_0).$$

The Newton-Raphson method is to iteratively use the equation

$$\widehat{\boldsymbol{\beta}}_{\text{new}} = \boldsymbol{\beta}_0 - [\ell''(\boldsymbol{\beta}_0)]^{-1}\ell'(\boldsymbol{\beta}_0)$$

for a given initial value $\boldsymbol{\beta}_0$. When the algorithm converges: $\widehat{\boldsymbol{\beta}}_{\text{new}} = \widehat{\boldsymbol{\beta}}_0 = \widehat{\boldsymbol{\beta}}$, we have

$$\widehat{\boldsymbol{\beta}}_{\text{new}} = \widehat{\boldsymbol{\beta}} - [\ell''(\widehat{\boldsymbol{\beta}})]^{-1}\ell'(\widehat{\boldsymbol{\beta}}),$$

or

$$\ell'(\widehat{\boldsymbol{\beta}}) = 0.$$

Namely, $\widehat{\boldsymbol{\beta}}$ solves the likelihood equation. From the discussion above, we can replace $\ell''(\widehat{\boldsymbol{\beta}}_0)$ by any non-degenerate matrix, resulting in an iterative scheme

$$\widehat{\boldsymbol{\beta}}_{\text{new}} = \widehat{\boldsymbol{\beta}}_0 - [A(\widehat{\boldsymbol{\beta}}_0)]^{-1}\ell'(\widehat{\boldsymbol{\beta}}_0).$$

Different choice of A would have difference speed of convergence.

If we choose $A(\boldsymbol{\beta}_0) = I$, the identity matrix, it leads to a deepest decent method. The advantage of this choice is that we do not need to compute the second derivative matrix and its inverse. However, it converges to the solution very slowly.

As known, $E\ell''(\boldsymbol{\beta}_0) = -I_n(\boldsymbol{\beta}_0)$, the Fisher information matrix. If one uses $A(\boldsymbol{\beta}_0) = -I_n(\boldsymbol{\beta}_0)$, we have

$$\widehat{\boldsymbol{\beta}}_{\text{new}} = \widehat{\boldsymbol{\beta}}_0 + I_n^{-1}(\widehat{\boldsymbol{\beta}}_0)\ell'(\widehat{\boldsymbol{\beta}}_0).$$

This is Fisher score method. Now let us examine the Fisher scoring method closely. For simplicity, take $w_i = 1$. Recall

$$\ell_n(\boldsymbol{\beta}) = \sum_i [\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y, \phi)]$$

and

$$\theta_i = (b')^{-1} \circ g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}).$$

The gradient vector can be found by

$$\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i \frac{y_i - \mu_i}{a(\phi)} h'(\mathbf{x}_i^T \boldsymbol{\beta}) x_{ij}.$$

In vector notation,

$$\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i \frac{y_i - \mu_i}{a(\phi)} h'(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i = \mathbf{X}^T W (\mathbf{y} - \boldsymbol{\mu}),$$

where

$$W = \text{diag}\{h'(\mathbf{x}_i^T \boldsymbol{\beta})/(g'(\mu_i) a(\phi))\}.$$

By some straightforward algebraic calculations,

$$I_n = \mathbf{X}^T W \mathbf{X}.$$

Hence, the Fisher scoring method gives

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_0 + (\mathbf{X}^T W \mathbf{X})^{-1} (\mathbf{X}^T W (\mathbf{y} - \boldsymbol{\mu}) = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W (\mathbf{y} - \boldsymbol{\mu} + \mathbf{X} \boldsymbol{\beta}_0).$$

This is the same as weighted least squares solution with weight W and the working response variable

$$\mathbf{x}_i^T \boldsymbol{\beta}_0 + g'(\mu_i^0)(y_i - \mu_i^0)$$

The values $\{g'(\mu_i)(y_i - \mu_i)\}$ are called working residuals because they are the difference between the working responses and their predictors $\{\mathbf{x}_i^T \boldsymbol{\beta}_0\}$. Note that the conditional variance of the working responses is $a(\phi) b''(\theta_i) g'(\mu_i)^2$. So the weight W provides the right weighting matrix. The algorithm reads as follows.

Step 1 Given initial value $\boldsymbol{\beta}_0$, construct the adjusted (working) variable

$$z_i = \mathbf{x}_i^T \boldsymbol{\beta}_0 + g'(\mu_i)(Y_i - \mu_i).$$

Step 2 Update $\boldsymbol{\beta}_0$ by using the weighted least squares

$$\boldsymbol{\beta}_1 = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W \mathbf{z}.$$

Step 3 Regarding $\boldsymbol{\beta}_1$ as initial value and update it. Iteratively between Step 1 and Step 2 until convergence.

6.4 Deviance and Residuals

6.4.1 Deviance

As discussed in the last section, the maximum likelihood estimate for $\boldsymbol{\beta}$ can be found by maximizing $\ell_n(\boldsymbol{\beta}; \mathbf{y}, \mathbf{x})$. Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood estimate and $\widehat{\theta}$ be its fitted value, namely

$$\widehat{\theta}_i = \theta\{g^{-1}(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}})\}.$$

Without putting restriction on θ_i, we would have to maximize

$$\ell_n(\theta, \phi; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{n} w_i\{y_i\theta_i - b(\theta_i)\}/\phi.$$

This turns out that the unrestricted maximizer $\tilde{\theta}_i = (b')^{-1}(y_i) = \theta(y_i)$. Then the difference is the lack-of-fit due to the model restriction, called *deviance* and denoted by $D(\mathbf{y}; \widehat{\mu})$,

$$\sum_{i=1}^{n} w_i\{y_i(\tilde{\theta}_i - \widehat{\theta}_i) - b(\tilde{\theta}) + b(\widehat{\theta})\}/\phi = \frac{1}{2} D(\mathbf{y}; \widehat{\mu})/\phi.$$

When the conditional distribution of Y given $\mathbf{x}$ is normal distribution, then $\theta(\mu) = \mu$ and $b(\theta) = \frac{\theta^2}{2}$. Then

$$\begin{aligned} D(\mathbf{y}; \widehat{\mu}) &= 2\sum_{i=1}^{n} w_i\{y_i(y_i - \widehat{y}_i) - \frac{y_i^2}{2} + \frac{\widehat{y}_i^2}{2}\} \\ &= \sum_{i=1}^{n} w_i(y_i - \widehat{y}_i)^2. \end{aligned}$$

This is the sum of residual squares under a weighted linear model.

For the useful members, their deviances are listed in Table 6.5.

TABLE 6.5. Deviances

Distribution	deviance ($w_i = 1$)
Normal	$\sum(y_i - \widehat{\mu}_i)^2$
Binomial	$2\sum(y_i \log(y_i/\widehat{\mu}_i) - (n_i - y_i)\log\{(n_i - y_i)/(n_i - \widehat{Y}_i)\})$
Poisson	$2\sum\{y_i \log(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)\}$
Gamma	$2\sum\{-\log(y_i/\widehat{\mu}_i) + (y_i - \widehat{\mu}_i)/\widehat{\mu}_i\}$

6.4.2 Analysis of Deviance

The value of deviance is usually not scaled, however, the reduction of deviance residual provides useful information. Suppose that we are comparing a larger model with a smaller one. Let Θ_1 and Θ_0 be the corresponding parameter space and Θ be the set consisting of all possible (no constraint) model. By the definition,

$$\begin{aligned} & \text{Deviance(smaller model)}-\text{Deviance(larger model)} \\ = \;& 2\phi\{(\max_{\theta\in\Theta} \ell_n(\theta) - \max_{\theta\in\Theta_0} \ell_n(\theta)) - (\max_{\theta\in\Theta} \ell_n(\theta) - \max_{\theta\in\Theta_1} \ell_n(\theta))\} \\ = \;& 2\phi\{\max_{\theta\in\Theta_1} \ell_n(\theta) - \max_{\theta\in\Theta_0} \ell_n(\theta)\} \\ \to \;& \phi\chi_d^2 \end{aligned}$$

in distribution by Wilks' theorem, where $d = \dim(\Theta_1) - \dim(\Theta_0)$. Therefore we have the following proposition

Proposition 6.1 *Under some regularity conditions, it holds that*

$$\textit{Deviance(smaller model)} - \textit{Deviance(larger model)} \to \phi\chi^2_d.$$

For binomial and Poisson models without over-dispersion, we can take $\phi = 1$. Like the ANCOVA table, we can also obtain the Analysis of Deviance Table. Because the models are not orthonormal, like ANCOVA, the conclusion should be drawn with care. Suppose that we have factors A and B and variables $X_1, \cdots, X_p$. The Deviance can be decomposed as

$$\begin{aligned}
& \mathrm{Dev}(\mu) \\
= \; & \{\mathrm{Dev}(\mu) - \mathrm{Dev}(A)\} + \{\mathrm{Dev}(A) - \mathrm{Dev}(A, B)\} \\
& + \{\mathrm{Dev}(A, B) - \mathrm{Dev}(A * B)\} + \{\mathrm{Dev}(A * B) - \mathrm{Dev}(A * B, X_1)\} \\
& + \cdots \\
& + \{\mathrm{Dev}(A * B, X_1, \cdots, X_{p-1}) - \mathrm{Dev}(A * B, X_1, \cdots, X_p)\} \\
& + \mathrm{Dev}(A * B, X_1, \cdots, X_p) \\
\equiv \; & R_A + R_{B|A} + \cdots + R_{X_p|A*B, X_1, \cdots, X_p} + \mathrm{Dev}(A * B, X_1, \cdots, X_p)
\end{aligned}$$

This decomposition depends on the order. For other nested model sequences, we can do similar decomposition as in examples below.

Example 6.9 (Continued) Table 6.6 shows that the deviance decomposition. From this table, one can see that age and number are not significant. For testing significant of number and age simultaneously, we take deviance reduction = 3.5357+3.1565=6.6922 with 2 degrees

of freedom. The corresponding p-value is 0.035, which indicates that there is not very strong evidence against the null hypothesis.

TABLE 6.6. Table of Analysis of Deviance

Resource	DF	Deviance reduction	Residual Deviance	p-value
Null			83.2345	
Start	1	15.1623	68.0722	0.001
Number	1	3.5357	64.5365	0.060
Age	1	3.1565	61.3799	0.076

Example 6.10 (Continued) The deviance can be decomposed in Table 6.7. From Table 6.7, all of the factors are significant. Some possible improvement could be made via changing the order of the fitting or using over-dispersion model to fit these data.

TABLE 6.7. Table of Analysis of Deviance

Resource	DF	Deviance reduction	Residual Deviance	p-value
Null			6855.69	
Opening	2	2524.56	4331.13	0.0000
Solder	1	936.96	3394.17	0.0000
Mask	3	1653.09	1714.08	0.0000
PadType	9	542.46	1198.62	0.0000
Panel	2	68.14	1130.48	0.0000

6.4.3 Deviance residuals

Let $d_i = y_i(\tilde{\theta}_i - \widehat{\theta}_i) - \{b(\tilde{\theta}_i) - b(\widehat{\theta})\}$. Thus d_i in fact is the discrepancy of the i-th unit, and

$$D(\mathbf{y}, \widehat{\mu}) = \sum_{i=1}^{n} w_i d_i^2.$$

Define deviance residual as

$$r_{D,i} = d_i \text{sgn}(y_i - \widehat{y}_i).$$

For Bernoulli model,

$$d_i^2 = -2 \log \widehat{p}_i I(y_i = 1) - 2 \log \widehat{q}_i I(y_i = 0).$$

Thus

$$r_{D,i} = \sqrt{-2 \log \widehat{p}_i} I(y_i = 1) - \sqrt{-2 \log \widehat{q}_i} I(y_i = 0).$$

Deviance residual has the following property.

Proposition 6.2 *For each given y_i, the deviance residual $r_{D,i}$ is an increasing function of $y_i - \widehat{\mu}_i$.*

Proof: We now drop the subscript i and write $z = y - \widehat{\mu}$. Note that $\theta = b'^{-1}(\theta) = \theta(\mu)$. Denote $f(z) = r_D^2$. Thus

$$\begin{aligned} f(z) &= y(\tilde{\theta} - \widehat{\theta}) - \{b(\tilde{\mu}) - b(\widehat{\theta})\} \\ &= y(\tilde{\theta} - \theta(\widehat{\theta})) - [b(\tilde{\theta}) - b\{\theta(\widehat{\mu})\}] \\ &= y\{\tilde{\theta} - \theta(y - z)\} - b(\tilde{\theta}) + b\{\theta(y - z)\}. \end{aligned}$$

We need only to show that $f(z)$ is increasing when $z > 0$ and decreasing when $z < 0$. Note that

$$f'(z) = y\theta'(y - z) - b'\{\theta(y - z)\}\theta'(y - z) = z\theta'(y - z)$$

as $b'\{\theta(y-z)\} = y-z$. Recall that $b'(\theta) = \mu$, which implies $b''(\theta)\, d\theta = d\mu$. This implies that

$$\frac{d\theta}{d\mu} = \frac{1}{b''(\theta)} > 0.$$

Hence the sign of $f'(z)$ is the same as that of z. Thus $f(z)$ is increasing and when $z > 0$ and decreasing when $z < 0$. This completes the proof.

Deviance residuals provide useful raw materials for model diagnostic, in a similar way to the ordinary residuals. The deviance, the sum of squares of deviance residuals, plays a very similar role to the summation of residual squares.

6.4.4 Pearson residuals

Pearson's method was invented before modern computer time. It has simple computation advantage, such as goodness of fit test. Let $E(Y) = \mu$ and $\text{var}(Y) = V(\mu)$, where $V(\mu)$ is known. For generalized linear models, $V(\cdot)$ is known. Let $\widehat{\mu}_i$ be the fitted value. Thus in generalized linear models,

$$\widehat{\mu}_i = g^{-1}(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}).$$

Then, the *Pearson residual* is defined as

$$\widehat{r}_{P,i} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}.$$

The Pearson residual is the same as the ordinary residual for normal distributed error, and is skewed for nonnormal distributed response.

The Pearson goodness-of-fit statistic is

$$X^2 = \sum_{i=1}^{n} \widehat{r}_{p,i}^2.$$

The residual defined above plays an similar role to the usual one. It can be used to construct a residual plot for examining overall pattern, detecting heteoscedasticity and suggesting alternative models.

6.4.5 Anscombe residual

For non-normal data y, a traditional approach is to transform y so that the distribution of $A(y)$ is closed to as normal distribution as possible in some sense. It was shown by Wedderburn (unpublish, but see Barndorff-Nielsen, 1978) that the function is given by

$$A(z) = \int_a^z \frac{d\mu}{V^{1/3}(\mu)},$$

where a is the lower limit of the range of μ. The residual

$$A(y) - A(\widehat{\mu}) \approx A'(\mu)(y - \widehat{\mu})$$

when $\widehat{\mu}$ is close to y. Hence the variance of the "residual" $A(y) - A(\widehat{\mu})$ is approximate

$$A'(\mu)^2 V(\mu) = V^{-2/3}(\mu)V(\mu) = V(\mu)^{1/3}.$$

This leads to define the *Anscombe residual* as

$$r_A = \frac{A(y) - A(\widehat{\mu})}{V(\widehat{\mu})^{1/6}}.$$

For the Poisson distribution,

$$A(z) = \int_0^z \frac{d\mu}{\mu^{1/3}} = \frac{3}{2} z^{2/3}.$$

Hence

$$r_A = \frac{\frac{3}{2}(y^{2/3} - \widehat{\mu}^{2/3})}{\widehat{\mu}^{1/6}}.$$

In contrast, Pearson residual is

$$r_p = \frac{Y - \widehat{\mu}}{\widehat{\mu}^{1/2}}$$

and the deviance residual

$$r_D = 2\{y \log(y/\widehat{\mu}) - (y - \widehat{\mu})\}\text{sgn}(y - \widehat{\mu}).$$

6.5 Comparison with Response Transform Models

Generalized linear models are an alternative to response transformation models of the form

$$\psi(y) = \boldsymbol{\beta}^T \mathbf{x} + \varepsilon$$

which are also used for enhancing linearity and homogeneity of variance. In fact, certain choices of g and V above lead to analyses very similar to the class of response variable reexpression models, but in fact they are more general due to their flexibility in allowing separate functions to specify linearity and variance relationships.

Reexpressions, although very useful at times, suffer from several defects:

1. Familiarity of the measured response variable is sacrificed in the analysis of $\psi(y)$.

2. A single reexpression $\psi(y)$ must simultaneously enhance both linearity and homogeneity of variance.

3. Often the preferred transformations are not defined on the boundaries of the sample space; e.g., the logit transformation is not defined for observed proportions exactly equal to zero or one.

For example, to make the Poisson response look like normal, one uses Anscomb tranform

$$\psi(y_i) = y_i^{2/3}.$$

So the residuals are based on $y_i^{2/3} - \mu_i^{2/3}$, which does not stabilize the variance. For example, for large μ_i

$$\begin{aligned} y_i^{2/3} - \mu_i^{2/3} &= (\mu_i - y_i - \mu_i)^{2/3} - \mu_i^{2/3} \\ &\approx \mu_i^{2/3} + \frac{2}{3}\mu_i^{-1/3}(y_i - \mu_i) - \mu_i^{2/3} \\ &\to N(0, \frac{4}{9}\mu_i^{1/3}) \end{aligned}$$

in distribution. Hence the variance is heteroscedastic.

6.6 S-plus codes

6.6.1 S-plus codes for Example 6.9

```
> kyphosis
   Kyphosis Age Number Start
 1   absent  71      3     5
 2   absent 158      3    14
 3  present 128      4     5
 4   absent   2      5     1
 5   absent   1      4    15
 .
 .
 .
79  present 157      3    13
80   absent  26      7    13
81   absent 120      2    13
82  present  42      7     6
83   absent  36      4    13


> kyph <- glm(Kyphosis ~Age + Start + Number,
+  family=binomial, data=kyphosis)
> summary(kyph)

Call: glm(formula = Kyphosis ~ Age + Start + Number,
```

```
family = binomial, data = kyphosi s)
Deviance Residuals:
       Min         1Q     Median         3Q     Max
 -2.312363 -0.5484308 -0.3631876 -0.1658653 2.16133

Coefficients:
                  Value Std. Error   t value
(Intercept) -2.03693225 1.44918287 -1.405573
        Age  0.01093048 0.00644419  1.696175
      Start -0.20651000 0.06768504 -3.051043
     Number  0.41060098 0.22478659  1.826626

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 83.23447 on 80 degrees of freedom

Residual Deviance: 61.37993 on 77 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:
       (Intercept)        Age      Start
   Age -0.4633715
 Start -0.3784028  -0.2849547
Number -0.8480574   0.2321004  0.1107516

> anova(glm(Kyphosis ~Start  + Number+Age,family=binomial,
+ data=kyphosis),test="Chi")
Analysis of Deviance Table

Binomial model

Response: Kyphosis

Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev    Pr(Chi)
  NULL                    80   83.23447
 Start  1 15.16229        79   68.07218 0.00009865
Number  1  3.53571        78   64.53647 0.06006055
   Age  1  3.15654        77   61.37993 0.07562326
```

6.6.2 S-plus codes for Example 6.10

```
> solder.balance
    Opening Solder Mask PadType Panel skips
  1       L  Thick A1.5      W4     1     0
  2       L  Thick A1.5      W4     2     0
  3       L  Thick A1.5      W4     3     0
  4       L  Thick A1.5      D4     1     0
  5       L  Thick A1.5      D4     2     0
  .
  .
  .
896       S   Thin   B6      W9     2    21
897       S   Thin   B6      W9     3    15
898       S   Thin   B6      L9     1    11
899       S   Thin   B6      L9     2    33
900       S   Thin   B6      L9     3    15

>  solder.glm <- glm(skips ~., family=poisson,
+   data=solder.balance)
> summary(solder.glm)

Call: glm(formula = skips ~ Opening + Solder +
    Mask + PadType + Panel, family = poisson,
    data = solder.balance)
Deviance Residuals:
       Min        1Q     Median        3Q      Max
 -3.661532 -1.086761 -0.4406795 0.6114746 3.942936

Coefficients:
                  Value  Std. Error    t value
(Intercept)  0.735679734 0.029480508  24.954785
  Opening.L -1.338897707 0.037898452 -35.328560
  Opening.Q  0.561940334 0.042005350  13.377828
     Solder -0.777627385 0.027309904 -28.474190
      Mask1  0.214096793 0.037719290   5.676056
      Mask2  0.329383406 0.016528258  19.928501
      Mask3  0.330750657 0.008946418  36.970177
   PadType1  0.055000439 0.033193067   1.656986
   PadType2  0.105788229 0.017332685   6.103395
```

```
   PadType3 -0.104859727 0.015162522  -6.915718
   PadType4 -0.122876529 0.013604968  -9.031740
   PadType5  0.013084728 0.008852899   1.478016
   PadType6 -0.046620368 0.008837701  -5.275169
   PadType7 -0.007583584 0.006976023  -1.087093
   PadType8 -0.135502138 0.010597577 -12.786144
   PadType9 -0.028288193 0.006563991  -4.309603
     Panel1  0.166761164 0.021027817   7.930503
     Panel2  0.029213741 0.011743659   2.487618

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: 6855.69 on 719 degrees of freedom

Residual Deviance: 1130.48 on 702 degrees of freedom

Number of Fisher Scoring Iterations: 4


> anova(solder.glm,test="Chi")
Analysis of Deviance Table

Poisson model

Response: skips

Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev       Pr(Chi)
   NULL                   719   6855.690
Opening  2 2524.563       717   4331.128 0.000000e+000
 Solder  1  936.955       716   3394.173 0.000000e+000
   Mask  3 1653.092       713   1741.080 0.000000e+000
PadType  9  542.463       704   1198.617 0.000000e+000
  Panel  2   68.137       702   1130.480 1.554312e-015
```

References

[1] Altman, D. G. (1991). *Practical Statistics for Medical Research.* Chapman and Hall, London.

[2] Barndorff-Nielsen, O. E. (1978). *Information and Exponenial Families in Statistical Theory*, Chichester, Wiley.

[3] Barnett, V and Lewis, T. (1978). *Outliers in Statistical Data*, Wiley, Chichester.

[4] Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models, *Journal of American Statistical Association*, **95**, 888-902.

[5] Champers, J. M. and Hastie, T. J. (1993). *Statistical Models in S.* Chapman and Hall, New York.

[6] Comizzoli, R. B., Landwehr, J. M. and Sinclair, J. D. (1990). Robust materials and processes: Key to Reliability. *AT&T Technical Journal*, **69**, 113-128.

[7] Davidian, M. and Giltinan, D.M. (1995) *Nonlinear Mixed Effects Models for Repeated Measurement Data*, Chapman and Hall.

[8] Dean, A and Voss, D. (1999). *Design and Analysis of Experiments*, Springer, New York.

[9] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models, *The Annals of Statistics*, **27**, 1491-1518.

[10] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, reprinted in *Contributions to Mathematical Statistics* (by R. A. Fisher) (1950), Wiley & Sons, New York.

[11] Hawkins, D. M. (1980). *Identification of Outliers*, Chapman and Hall, London.

[12] Henderson, C. R. (1950).

[13] Hochberg, Y. and Tamhance, A. C. (1987). *Multiple Comparison Procedures*. Wiley & Sons, New York.

[14] Lindsey, J. K. (1997). *Applying Generalized Linear Models*, Springer, New York.

[15] Longford, N. T. (1993). *Random Coefficient Models*, Oxford University Press, Oxford.

[16] McCullagh, P. and Nedler, J. A. (1989). *Generalized Linear Models*, Second Edition, Chapman and Hall, London.

[17] Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Statistical Models*, Fourth Edition. Irwin, Chicago.

[18] Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects, *Statistical Sciences*, **6**, 15-51.

[19] Smith, R. L. and Young, K. D. S. (2000). *Linear Regression*, Manuscript.

[20] Thompson, (1961).

Author index

Subject index