CHAPTER 4

CIRCULAR REGRESSION MODEL

4.1 Introduction

Regression analysis is a statistical technique for investigating and modelling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, physical sciences, economics, management, life and biological sciences and social sciences. In fact, regression analysis may be the most widely used statistical technique which includes investigating, forecasting and modelling the relationship between variables. There are several types of regression models that have been used widely such as linear regression, logistic regression, multivariate regression, and circular regression models for various purposes. In order to use any of the models, we have to statisfy every single assumption that has been specified. Linear regression is the simplest form of regression models which make several assumptions, including the measurement errors need to be normally distributed. The regression models can be extended to the case of circular variables.

The study of regression models for circular variable started four decades ago. Gould (1969) proposed a regression model to predict a circular response variable θ from a set of linear covariates, where θ has a von Mises distribution with mean μ and concentration parameter κ, $VM(\mu,\kappa)$. The proposed model is given by

$$\mu = \mu_0 + \sum_{j=1}^{k} \beta_j x_j \qquad (4.1)$$

where μ_0 and β_j are unknown parameters, and x_j is a linear covariate, $j = 1,2,\ldots,k$. Assuming that $\theta_1, \theta_2, ..., \theta_n$ is a set of circular independent and identical observations of

von Mises distributions with mean directions $\mu_1, \mu_2, ..., \mu_n$ respectively and unknown concentration parameter κ, Mardia (1972) extended model (4.1) to

$$\mu_i = \mu_0 + \beta\, t_i \tag{4.2}$$

for some known numbers $t_1, t_2, ..., t_n$ and unknown μ_0 and β. Jammalamadaka and Sarma (1993) proposed a regression model for two circular random variables u and v in term of the conditional expectation of the vector e^{iv} given by

$$E(e^{iv_i} \mid u_i) = \rho(u_i)e^{i\mu(u_i)} = g_1(u_i) + ig_2(u_i) \tag{4.3}$$

where $\mu(u_i)$ is the conditional mean direction of v_i given u_i with conditional concentration $0 \le \rho(u_i) \le 1$. Due to the difficulty of estimating $g_1(u_i)$ and $g_2(u_i)$ from the data, Jammalamadaka and Sarma (1993) expressed them in terms of their Fourier series expansions.

Hussin *et al.* (2004) extended model (4.2) for the case when both response and explanatory variables are circular. For any circular observation $(u_1, v_1), (u_2, v_2), ..., (u_n, v_n)$ of circular variables u and v with a linear relationship between them, they proposed a model

$$v_i = \alpha + \beta\, u_i + \varepsilon_i \ (\mathrm{mod} 2\pi) \tag{4.4}$$

where ε_i is circular random errors following a von Mises distribution with mean circular 0 and concentration parameter κ. They imposed a restriction on the model parameters, so that β is an integer and close to one. In this study, the focus will be a circular regression model by Down and Mardia (2002) which will be described in the following section. Herewith, we refer to this model as the DM circular regression model.

4.2 Down and Mardia (DM) Circular RegressionModel

Down and Mardia (2002) proposed the DM circular regression model which maintains a one-to-one correspondence between the independent angle and the mean of the dependent angle. Assume that v is the dependent random angle and u is the fixed independent angle. The parameter α and β are the angular location parameters while ω is a slope parameter in the closed interval [-1, 1]. They proposed the DM model given as

$$\tan\frac{1}{2}(v-\beta)=\omega\tan\frac{1}{2}(u-\alpha). \tag{4.5}$$

The locus of the points (u,v) is a continuous closed curve winding once around a toroidal surface. Eq. (4.5) has a unique solution given by

$$v=\beta+2\tan^{-1}\left\{\omega\tan\frac{1}{2}(u-\alpha)\right\}. \tag{4.6}$$

Eq. (4.6) defines a one to one relationship between u and v where $\omega\neq 0$. Fisher and Lee (1992) suggested the link function $\mu=2\tan^{-1}x$ for linear-circular regression, since it maps the linear variable x to $(-\pi,\pi]$. Suppose that u is the fixed independent angle, v the dependent angle and v in Eq. (4.6) replaced by μ, the mean direction for v given u. The resulting link function, or regression curve, is given by

$$\tan\frac{1}{2}(\mu-\beta)=\omega\tan\frac{1}{2}(u-\alpha) \tag{4.7}$$

which has unique a solution

$$\mu=\beta+2\tan^{-1}\left\{\omega\tan\frac{1}{2}(u-\alpha)\right\}. \tag{4.8}$$

Down and Mardia (2002) classified the regression models into three categories which are class A, class B and class C. In the class A model, the regression model has three

functionally independent parameters α, β and ω. For the class B model, α and β have a relationship such as $\alpha \pm \beta = 0$, or can be written as $(\alpha, \pm\alpha, \omega)$. In the class C model, the slope parameter, ω can take one of the special values of $\{-1,0,1\}$ such as $\omega = \omega_0$. The loglikelihood function for a random sample of *n* pairs (u_j, v_j) from a Class A or B model is given by

$$l(\alpha, \beta, \omega; v_1, \ldots, v_n) = -n \log I_0(\kappa) + \kappa \sum_j \cos(v_j - \beta - v(u_j - \alpha; \omega)) + \text{constant}. \quad (4.9)$$

Differentiating Eq. 4.9 with respect to κ gives

$$\frac{\partial l(\alpha, \beta, \omega; v_1, \ldots, v_n)}{\partial \kappa} = -n \frac{I_1(\kappa)}{I_0(\kappa)} + \sum_j \cos(v_j - \beta - v(u_j - \alpha; \omega)) = -n A_1(\kappa) + \sum_j \cos(v_j - \beta - v(u_j - \alpha; \omega))$$

where $\rho = A_1(\kappa)$ as defined in Section 2.4.4. Recall that for class B models, $\beta = \pm\alpha$. For both classes, the maximum likelihood estimator $\hat{\rho}$ of the precision parameter ρ is defined explicitly by

$$\hat{\rho}(\alpha, \beta, \omega) = \frac{1}{n} \sum_j \cos(v_j - \beta - v(u_j - \alpha; \omega)). \quad (4.10)$$

Class B and C are special cases of Class A; with the parameters of α and β are related for the earlier case, while ω is restricted to take specific values only. Consequently, the estimation of parameters are expected to be simpler for this special cases.

4.3 The Maximum Likelihood Estimation (MLE) of Parameters in DM Model

From the previous section, the loglikelihood function for a random sample of *n* pairs from a Class A or B model is given by Eq. (4.9). In order to maximize the likelihood function, we employ iterative approach which require the determination of initial

values α_0, β_0 and ω_0. These initial values are obtained by calculating the precision parameter ρ in Eq. (4.10) for all posible pairs of α, β and ω in a pre-specified sets. In our case, we consider the following pre-specified sets of parameter values, $\alpha = [-\pi, \pi]$, $\beta = [-\pi, \pi]$and $\omega = [-1,1]$. Thus, we consider a set of initial values; $(\alpha_0, \beta_0, \omega_0)$ correspond to value which maximize the precision parameter ρ. Using these initial values, we obtain the MLE estimates for the three parameters using Eq. (4.9). This can be done by using the MS function available in S-Plus software.

4.4 Covariance Matrix of Circular Regression Model

Down and Mardia (2002) provided the information matrix for DM sircular regression model using the loglikelihood function with known parameters $(\alpha, \beta, \omega, \kappa)$

$$l = const - n \log I_0(\kappa) + \kappa \sum \cos(v_i - \mu_i),$$

where

$$\mu_i = \beta + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(u_i - \alpha)\right\}.$$

Using the fact that

$$E\{\cos(v_i - \mu_i)\} = A(\kappa), \quad E\{\sin(v_i - \mu_i)\} = 0,$$

the Fisher information matrix $\boldsymbol{I} = (I_{ij})$ for $\boldsymbol{\theta}^T = (\beta, \alpha, \omega, \kappa) = (\theta_1, \theta_2, \theta_3, \theta_4)$. Given that, $I_{14} = I_{24} = I_{34} = 0$ so that $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ are independent of $\hat{\theta}_4$ as expected, asymptotically.

$$I = \begin{bmatrix} \boldsymbol{C}_{11} & 0 \\ 0 & \boldsymbol{C}_{22} \end{bmatrix},$$

where $\boldsymbol{0}^T = (0,0,0)$, $\boldsymbol{C}_{11}$ is 3 x 3 and $\boldsymbol{C}_{22}$ is a scalar. Then, $\boldsymbol{C}_{22} = I_{44} = nA'(\kappa)$, $\boldsymbol{C}_{11} = \kappa A(\kappa)\boldsymbol{B}(\theta_1, \theta_2, \theta_3)$ where the elements of the matrix $\boldsymbol{B}(\beta, \alpha, \omega)$ are

$$b_{11} = n,\ b_{12} = \sum(\mu_i)_\alpha,\ b_{13} = \sum(\mu_i)_\omega$$

$$b_{22} = \sum(\mu_i)_\alpha^2,\ b_{23} = \sum(\mu_i)_\alpha(\mu_i)_\omega,\ b_{33} = \sum(\mu_i)_\omega^2,$$

with

$$(\mu_i)_\omega = \frac{2\tan\frac{1}{2}(u_i - \alpha)}{1 + \omega^2 \tan^2\frac{1}{2}(u_i - \alpha)} \text{ and } (\mu_i)_\alpha = -\frac{\omega \sec^2\frac{1}{2}(u_i - \alpha)}{1 + \omega^2 \tan^2\frac{1}{2}(u_i - \alpha)}$$

Thus

$$\text{cov}(\hat{\beta}, \hat{\alpha}, \hat{\omega}) = \{\hat{\kappa}A(\hat{\kappa})\}^{-1}\{\boldsymbol{B}(\hat{\beta}, \hat{\alpha}, \hat{\omega})\}^{-1},$$

$$\text{cov}(\hat{\beta}, \hat{\kappa}) = \text{cov}(\hat{\alpha}, \hat{\kappa}) = \text{cov}(\hat{\omega}, \hat{\kappa}) = 0.$$

We will consider the covariance matrix in Chapter 5 to identify influential observation in bivariate circular data.

4.5 Practical Example

Here we consider real data sets to show the estimation of the DM circular regression model using MLE method; the ocean wind direction data and the circadian biological rhythm data.

4.5.1 Ocean wind direction data

In this section, we introduce briefly the HF (High frequency) radar system and the anchored wave buoy (AB) techniques for measuring the ocean wind direction followed by a description of the data.

1. **The HF radar System**

HF radar is an on-line mapping tool of surface current fields and the spatial distribution of the wave directional spectrum. The HF radar was developed by UK Rutherford and Appleton Laboratories and subsequently by Marex Ltd and The Marconi Radar Company. The system uses pulse radar with high radio frequency (24.4-27 MHz) to map surface current patterns over a large area of ocean.

2. **The Anchored wave buoy**

The anchored wave buoy is often used as the standard tool in evaluating new wind or wave measuring systems. Older models measure the vertical motion of wind and wave at a single point. Typical wave buoy also additionally measure the lope of the sea surface in two directions at the same points. We consider data collected along the Holderness coastline (the Humberside coast of the North Sea, United Kingdom) by using an HF radar system and an anchored wave buoy. The deployment began in October 1994. The following information is assumed:

i. There is temporal stationary over the period of measurements
ii. There is spatial stationary over the area of measurement
iii. The different techniques are measured independently

The wind direction is the direction of the local wind which blows across the sea surface and along the coast where the HF radar system and anchored wave buoy are deployed. The full data set is obtained from Hussin (1997) and is given in Appendix 5. There were a total of 129 measurements recorded by both instruments.

Descriptive Statistics

Several plots can be used to show the distributions of both measurements. In general, from Figures 4.1 and 4.2, both sets of measurement follow the same distribution. It can be seen that there is a high frequency in the second quadrant for both

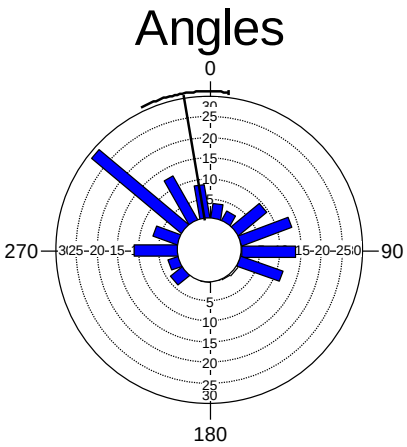

Figure 4.1: Circular Histogram for HF

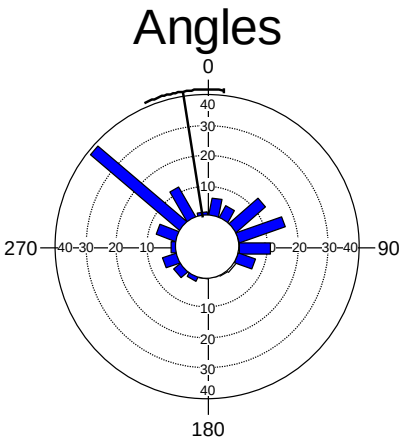

Figure 4.2: Circular Histogram for AB

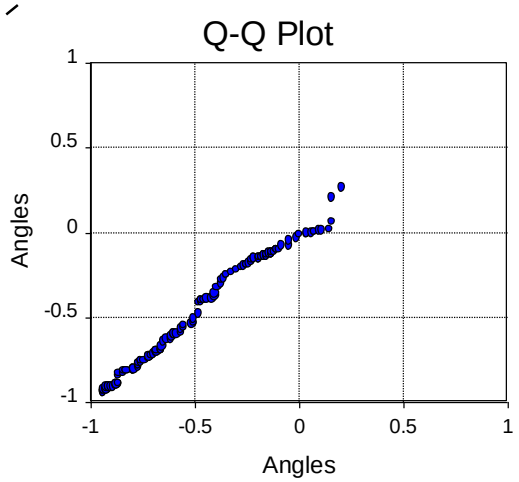

Figure 4.3: Q-Q plot for HF and AB

sets of measurements. From Figure 4.3, there are two points located at the top of the Q-Q plot. These points might correspond to observations which are candidates to be outliers. Some of the descriptive statistics for the ocean wind direction data are given in Table 4.1. The summary statistics of the HF radar and anchored wave buoy are almost similar including the concentration parameter with the value less than one.

Table 4.1: Descriptive statistics for the ocean wind direction data

Variable	HF(v)	AB(u)
Observations	129	129
Mean Direction	350.43°	351.06°
Mean Resultant Length	0.41	0.44
Circular Variance	0.59	0.56
Circular Std Dev	76.374°	73°
Median Direction	334.72°	327.33°
Circular Dispersion	28.3	25.74
Concentration parameter	0.902	0.99

Parameter Estimation

Using the data set, we calculate the precision parameters in the pre-specified sets as described in Section 4.3. The resulting plot of ρ versus index representing different points of (α, β, ω) is given in Figure 4.4. The initial values of each parameter correspond to the highest point observed in the plot giving $\alpha_o = 126°, \beta_o = 126°$ and $\omega_o = 0.9$. Thus, using these initial values, the final estimated parameter values are obtained by maximizing the log likelihood function given by equation (4.9); $\hat{\alpha} = 65.39°$, $\hat{\beta} = 71.82°$ and $\hat{\omega} = 0.91$. Figure 4.5 gives the P-P plot of the residuals from the resulting DM circular regression model. It can be seen that the points are all close to the straight line. Further, the value of the goodness-of-fit measure, $A^*(\hat{\kappa})$, as described in Section 2.5 is 0.90. These suggest that the model fits the data well.

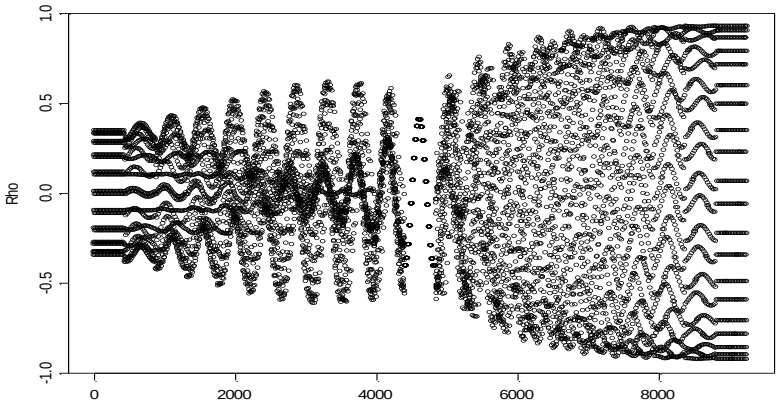

Figure 4.4: Plot of ρ versus index for ocean wind direction data

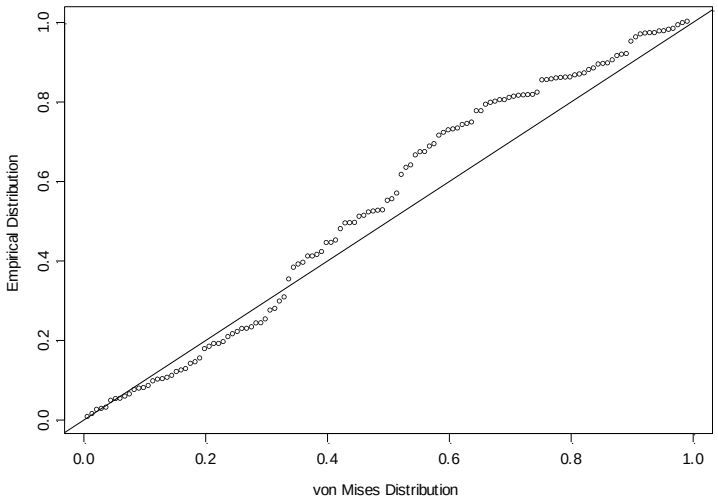

Figure 4.5: P-P plot of the residuals

4.5.2 Circadian data

We consider the circadian data provided by Downs and Mardia (2002). The data were obtained from 10 medical students in Austria. The students are measured for about 20 variables several times daily for a period of few weeks. The study period was split into two prime time periods as part of the study, and the peak time for systolic blood pressure (in degree) was estimated separately for each student for each period, giving values S1 and S2. These data are given in Appendix 6, in degrees, with 15 degrees equal to one hour. The two blood pressure peak times should be equivalent, if circumstances are the same for each of the two periods.

Descriptive Statistics

Several plots are used to study the distribution of S1 and S2. In general, the maximum blood pressures are observed in the upper left quadrant of the circular histogram indicating the same time in both periods as can be seen in Figures 4.6-4.7. From Figure 4.8, the Q-Q plot for testing the von Mises distribution shows that the quantiles are close to the straight line. There is a point observed at the bottom left of the plot and well separated from the others though lies on a straight line.

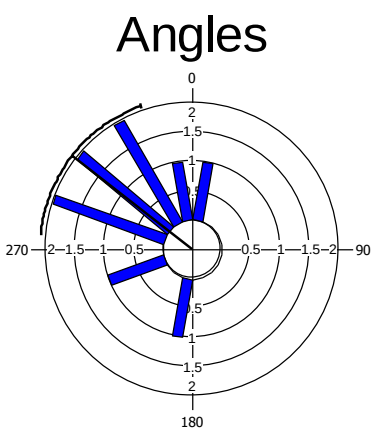

Figure 4.6: Circular Histogram for S1

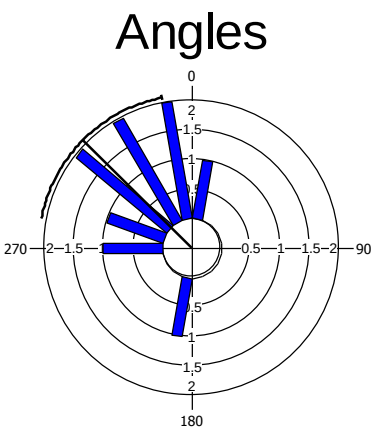

Figure 4.7: Circular Histogram for S2

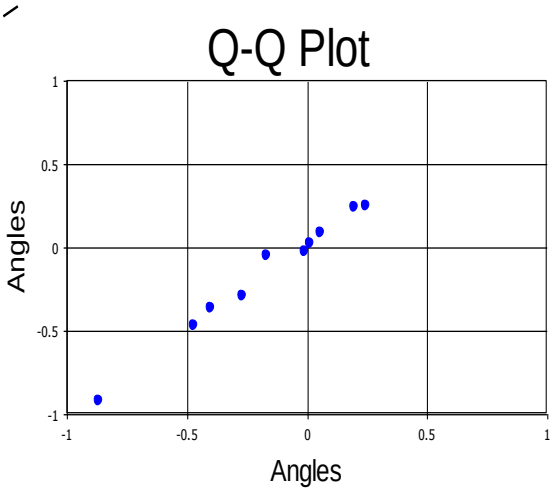

Figure 4.8: Q-Q plot for S1 and S2

The descriptive statistics for the circadian data are given in Table 4.2. Summary statistics of the peak time for systolic blood pressure (in degree) give almost similar values for S1 and S2.

Table 4.2: Descriptive statistics for the circadian data

Variable	S1(u)	S2(v)
Observations	10	10
Mean direction	307.93	314.69
Mean resultant length	0.74	0.72
Circular variance	0.26	0.28
Circular std dev	44.87	46.6
Median direction	314.5	318
Circular dispersion	28.3	25.74
Concentration parameter	2.251	2.125

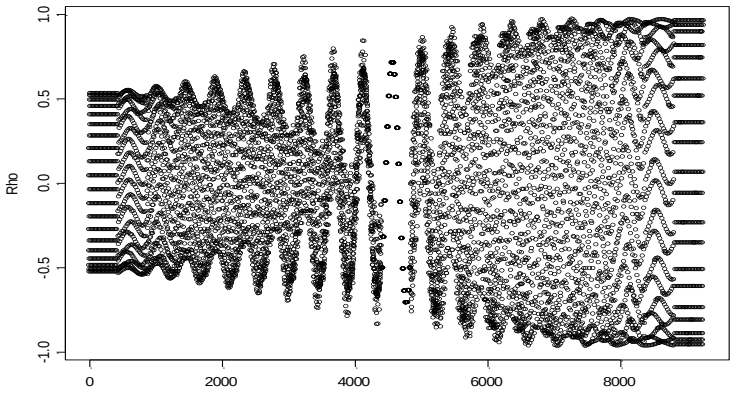

Figure 4.9: Plot of ρ versus index for circadian data

The concentration parameter for both periods are close and greater than 2, suggesting that the observations are concentrated in the directions of 307.93° and 314.69° for S1 and S2 respectively.

Parameter Estimation

Using the data set, we calculate the precision parameters in the pre-specified sets first and obtain the plot of ρ versus index as given in Figure 4.9. The initial values of each parameter correspond to the highest point observed in the plot giving $\alpha_o = 18°$, $\beta_o = 9°$ and $\omega_o = 0.70$. Thus, using these initial values, the final estimated parameter values are obtained by maximizing the log likelihood function from Eq. (4.9) giving $\hat{\alpha} = 16.58°$, $\hat{\beta} = 5.74°$ and $\hat{\omega} = 0.67$. Figure 4.10 gives the P-P plot of the residuals from the resulting DM circular regression model. It can be seen that the points are reasonably close to the straight line. Further, the value of the goodness-of-fit measure, $A^*(\hat{\kappa})$, is 0.945. These suggest that the model fits the data well.

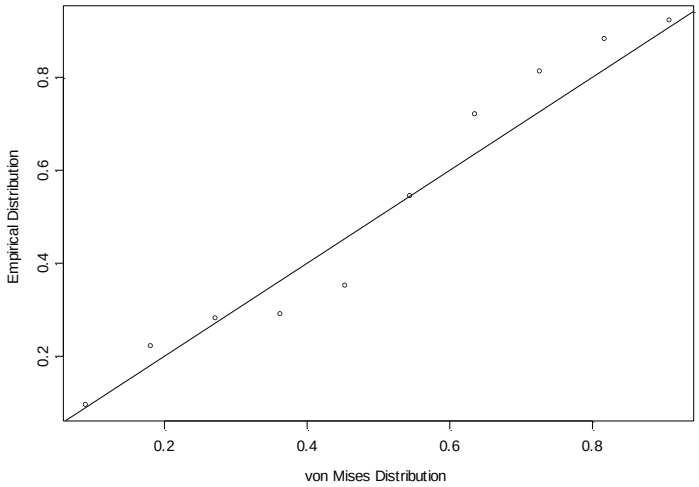

Figure 4.10: P-P plot of the residuals

4.6 Summary

In this chapter, we have described the DM circular regression model and its properties. We used the maximum likelihood estimation (MLE) method to estimate all the parameters in the model. The procedures to find the parameter estimation have been discussed in Section 4.3. Finally, we have illustrated the application of the model to the real data set as described in Section 4.5. We will utilize these estimates in our search of outliers and influential observations in the next two chapters.