# CHAPTER 13

# FURTHER TOPICS

## 13.1 INTRODUCTION

In this chapter we discuss two topics that have come up several times earlier but we did not focus on them. We will be discussing generalized linear models (GLM), and robust regression. These are two vast topics, and would require full-length books. We will give brief descriptions of the topics and provide examples that illustrate the concepts. GLM unifies the concept of linear model building, a primary activity of statistical analysts.

The importance of robust models in any statistical analysis cannot be overemphasized. The earlier chapters have provided us with methods for constructing robust models. In Section 13.5 we discuss methods that exclusively aim at robustness. The discussion on these two topics will not be exhaustive but reflect our personal experience and preferences.

## 13.2 GENERALIZED LINEAR MODEL

As in Chapter 3, given a response variable $Y$ and $p$ predictor variables $X_1$, $X_2$, . . ., $X_p$, the linear regression model can be described as follows: an observation $Y_i$ can

*Regression Analysis by Example, Fourth Edition.* By Samprit Chatterjee and Ali S. Hadi

be written as

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \\ &= \mu_i + \varepsilon_i, \end{aligned} \tag{13.1}$$

where $\mu_i$ is called the linear predictor and $\varepsilon_i$ is a random error assumed to have a Gaussian (normal) distribution.

The GLM extends the linear regression model in two ways. The $\varepsilon_i$ is assumed to have a distribution coming from the exponential family. The exponential family includes several standard distributions, in addition to the Gaussian. For example, it includes the binomial, Poisson, Gamma, and inverse Gaussian distributions.

The second generalization is that the mean function $\mu_i$ is not necessarily the linear predictor, but some monotonic differentiable function of the linear predictor,

$$h(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}, \tag{13.2}$$

where $h(\mu)$ denotes the function that links $\mu$ to the linear predictor. The function relating the mean to the linear predictor is called the link function.

These two generalizations considerably increase the flexibility of linear models. The GLM can be used in situations where a linear regression model would not be appropriate. These models are fitted by the method of maximum likelihood. Most statistical packages have programs that can be used to fit and analyze generalized linear models.

The GLM were first proposed by Nelder and Wedderburn (1972), and extensively developed by McCullagh and Nelder (1989). For computational details the reader should consult the references given above. A very accessible discussion is given in Simonoff (2003).

The logistic regression, which we discussed in the previous chapter, is an example of GLM although we did not describe it in those terms. We now describe logistic regression as a GLM. The probability distribution of the random error was binomial, since there are only two outcomes. Instead of the mean, $\pi_i$, being the linear predictor, we took a function of $\pi_i$, namely, $\ln(\pi_i/(1 - \pi_i))$ as the linear predictor. The logistic regression model can now be described as a GLM from the binomial family with a logit link function. Another example of GLM is the Poisson regression model. This is discussed in the next section.

## 13.3 POISSON REGRESSION MODEL

Poisson regression models are appropriate when the response variable is count data. A researcher in public health area may be interested in studying the number of hospitalizations of a group of people, and the characteristics associated with these patients. Simonoff (2003) studies the number of tornado deaths in relation to the month, year, and the classification of the tornado's severity. In Section 6.4, we have analyzed injury accidents in airlines. These data can be analyzed by Poisson regression, because here we are dealing with count data. We analyzed these data earlier by using the square root transformation, which is an approximation to the exact method that we are now considering. Note that in these data sets the counts are small numbers, and small values are observed more frequently than large values.

The Poisson regression model can be described as follows: the random component has a Poisson distribution (see Section 6.4 for the Poisson distribution), and the mean is linked to the linear predictor by a logarithmic function

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}. \tag{13.3}$$

The test and the inferences on the Poisson model are carried out in the same way as the logit model (logistic regression). In some analysis instead of analyzing the number of cases ($y$) we may be interested in analyzing the rates of occurrence. Let $y_i$ be the number observed out of $a_i$ that are exposed to the risk. To construct a model for the rate we have only to modify the link function. The link function for the rate is

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \ln(a_i). \tag{13.4}$$

The quantity $\ln(a_i)$ is called the off-set, and the $\ln(\mu_i)$ now represents the logarithm of the mean rate of occurrence instead of the logarithm of the mean occurrence.

We now illustrate Poisson regression by an example.

## 13.4 INTRODUCTION OF NEW DRUGS

Number of new drugs ($D$) for 16 diseases brought to the U.S. market between 1992–2005 is given in Table 13.1. Also provided are the prevalence rates ($P$) of these diseases for 100,000 people. The money allocated for research by the National Institute of Health ($M$) during the year 1994 for a specific disease in millions of dollars is also given. This dataset was kindly provided to us by Dr. Salomeh Keyhani of Mount Sinai Medical School. This is a part of a much larger database. We are using these three variables to illustrate the application of Poisson regression.

We are interested in studying the relationship between $D$ (the response variable), with $P$ and $M$. It should be noted that $D$ is an integer variable with small values. A new drug coming to the market is a rare event. Poisson distribution is often used to model rare events. We now will fit a GLM with a Poisson random component and log link function. The result is given in Table 13.2

The large values of the Wald's ($Z$-test) shows that the two variables are strongly related to the response variable $D$. The value of AIC[1] is 82.14. The model fitted values ($\hat{D}$) are given in the last column of Table 13.1. The agreement between observed and fitted values is satisfactory.

We will fit the data by least squares and compare this AIC value with that obtained from the least squares fit. The least squares fit is given in Table 13.3.

[1] See Section 11.5.3.

**Table 13.1** New Drugs Data: Number of New Drugs Introduced ($D$), Prevalent Rate ($P$), Expenditure on Research by the National Institute of Health ($M$), and the Predicted Number of Drugs ($\hat{D}$)

| | | $D$ | $P$ | $M$ | $\hat{D}$ |
|---|---|---|---|---|---|
| 1 | Ischemic Heart Disease | 6 | 8976 | 198.4 | 4.55 |
| 2 | Lung Cancer | 3 | 874 | 80.2 | 2.89 |
| 3 | HIV/AIDS | 21 | 1303 | 1049.6 | 20.29 |
| 4 | Alcohol Use | 2 | 18092 | 222.6 | 6.12 |
| 5 | Cerebrovascular Disease | 2 | 9467 | 108.5 | 3.86 |
| 6 | COPD[a] | 1 | 4271 | 48.9 | 2.98 |
| 7 | Depression | 7 | 12785 | 149.5 | 4.58 |
| 8 | Diabetes | 13 | 37850 | 278.4 | 11.66 |
| 9 | Osteoarthritis | 5 | 12345 | 151.3 | 5.54 |
| 10 | Drug abuse | 1 | 4000 | 442.1 | 6.48 |
| 11 | Dementia | 9 | 8931 | 344.1 | 6.09 |
| 12 | Asthma | 3 | 15919 | 41.8 | 4.02 |
| 13 | Colon Cancer | 2 | 1926 | 70.6 | 2.92 |
| 14 | Prostate Cancer | 4 | 2020 | 40.1 | 2.75 |
| 15 | Breast Cancer | 9 | 2262 | 159.5 | 3.52 |
| 16 | Bipolar Disorder | 2 | 2418 | 35.0 | 2.75 |

[a] Chronic Obstructive Pulmonary Disease

**Table 13.2** Output from the Poisson Regression Using $P$ and $M$

| Variable | Coefficient | s.e. | $z$-test | $p$-value |
|---|---|---|---|---|
| Constant | 0.8778 | 0.2074 | 4.233 | 0.0000 |
| $P$ | $2.700\times10^{-5}$ | $9.508\times10^{-6}$ | 2.840 | 0.0045 |
| $M$ | $1.998\times10^{-3}$ | $3.008\times10^{-4}$ | 6.642 | 0.0000 |
| | Log-Likelihood = –9.721 | $d.f. = 2$ | AIC = 82.14 | |

**Table 13.3** Output from the Linear Regression Using $P$ and $M$

| Variable | Coefficient | s.e. | $t$-test | $p$-value |
|---|---|---|---|---|
| Constant | 0.8362 | 1.379 | 0.607 | 0.5546 |
| $P$ | $1.317\times10^{-4}$ | $8.973\times10^{-5}$ | 1.467 | 0.1661 |
| $M$ | 0.01688 | $3.378\times10^{-3}$ | 4.996 | 0.0002 |
| $R^2 = 0.671$ | $R_a^2 = 0.620$ | $\hat{\sigma} = 3.292$ | $d.f. = 13$ | AIC = 88.207 |

Only the $M$ coefficient is significant in the least squares fit. As we have pointed out earlier, the linear regression model is not appropriate here as the response variable is count data. The value of AIC is 82.1 for the Poisson model compared to 88.2 for the linear regression model. The AIC indicates that the Poisson model fits the data better than does the linear regression model. The pattern of residuals is also more satisfactory for the Poisson model. The Poisson regression model is more appropriate for these data.

## 13.5 ROBUST REGRESSION

The regression model fitted to data should be robust, in the sense that deletion of one or two observations should not cause a drastic change in the model. In Chapter 4 (particularly Sections 4.8–4.9), we have described how to detect these points. Our prescription was to delete these points to get a more stable and realistic model. We will now consider a method in which instead of deleting these points (which may be considered subjective), we reduce the impact of these points. There are several ways of getting a regression model which is robust. We describe a method which is simple and effective. The problem with the least squares is that the procedure gives too much weight to outliers and high leverage points in the fitting. This has been illustrated extensively in Sections 4.8–4.9. The effect of these points can be reduced by down weighting these points in the fitting. We use weighted least squares (WLS), in which low weight is given to points with (i) high leverage and (ii) large residuals. Since the weights are determined by the residuals, and as these change from iteration to iteration, the procedure is an iterative one. The explicit form of the weights and the procedure are given in Algorithm 13.1 below, where we use $Q^j$ to denote the value of $Q$ in the $j$th iteration step.

**Algorithm 13.1.**

**Input**: An $n \times 1$ response vector $\mathbf{Y}$ and the corresponding $n \times p$ predictors matrix $\mathbf{X}$.

**Output**: A weighted least squares robust estimates of the regression coefficients and the corresponding residual vector.

**Step 0:** Compute the weighted least squares estimate of the regression coefficients when using $w_i^0 = 1/max(p_{ii}, p/n)$ as a weight for the $ith$ observation, where $p_{ii}$ is the $ith$ diagonal element of the projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Let this estimate be denoted by $\hat{\beta}^0$.

**Step $j$:** For $j = 1, 2, \ldots$, until convergence, compute

$$\mathbf{e}^{j-1} = \mathbf{Y} - \hat{\mathbf{Y}}^{j-1} = \mathbf{Y} - \mathbf{X}\hat{\beta}^{j-1}, \qquad (13.5)$$

**Table 13.4** Data Illustrating Robust Regression

| $Y$ | $X$ | $Y$ | $X$ |
|---|---|---|---|
| 18.8 | 3.4 | 92.7 | 12.5 |
| 20.5 | 5.7 | 124.2 | 14.2 |
| 29.4 | 7.5 | 142.4 | 15.2 |
| 45.0 | 8.8 | 154.7 | 15.8 |
| 72.4 | 11.1 | 118.4 | 17.9 |

which is the residuals of the fit at Step $j - 1$. Compute the new weights

$$w_i^j = \frac{(1 - p_{ii})^2}{max(\mid e_i^{j-1} \mid, m_e^{j-1})}, \tag{13.6}$$

where $m_e^{j-1}$ is the median of $(\mid e_1^{j-1} \mid, \ldots, \mid e_n^{j-1} \mid)$. Compute the weighted least squares estimate of the regression coefficients when using $w_i^j$ as a weight for the $ith$ observation. Let this estimate be denoted by $\hat{\beta}^j$.

As can be seen from the weighting scheme, those points with high leverage (high $p_{ii}$), or with large residuals ($e_i$) get low weights. The details of this procedure can be found in Chatterjee and Mächler (1997).

We provide two examples to illustrate the procedure.

## 13.6 FITTING A QUADRATIC MODEL

We illustrate the problem with least squares fit by a simple artificial data set given in Table 13.4. There are 10 observations on two variables $Y$ and $X$. The plot of the data in Figure 13.1 shows clearly a quadratic pattern. The least squares fit is given in Table 13.5.

The least squares fit shows both the linear and quadratic term are statistically insignificant. We will now fit the model by the robust regression method that we have outlined. The results are given in Table 13.6.

Figure 13.2 shows the least squares and the robust fits superposed on the scatter plot of $Y$ and $X$. The robust fit tracks the data considerably better than the least squares fit. The least squares fit is pulled away from the main body of the data by the high leverage outlier points in the top right and the bottom left. The robust fit does not suffer from this because such points are down weighted. Here this is visually obvious. In higher dimensions this would not be apparent, but the robust procedure would automatically take this into account. This is a commonly occurring situation.

We now illustrate robust regression by using real-life data.

**Table 13.5** Least Squares Quadratic Fit for the Data Set in Table 13.4

ANOVA Table

| Source | Sum of Squares | $d.f.$ | Mean Square | $F$-test | $p$-value |
|---|---|---|---|---|---|
| Regression | 21206 | 2 | 10603 | 25.14 | 0.001 |
| Residuals | 2952 | 7 | 422 | | |

Coefficients Table

| Variable | Coefficient | s.e. | $t$-test | $p$-value |
|---|---|---|---|---|
| Constant | -28.77 | 37.69 | -0.76 | 0.470 |
| $X$ | 9.329 | 7.788 | 1.20 | 0.270 |
| $X^2$ | 0.041 | 0.359 | 0.12 | 0.911 |
| $n = 10$ | $R^2 = 0.878$ | $R_a^2 = 0.843$ | $\hat{\sigma} = 20.5349$ | $d.f. = 7$ |

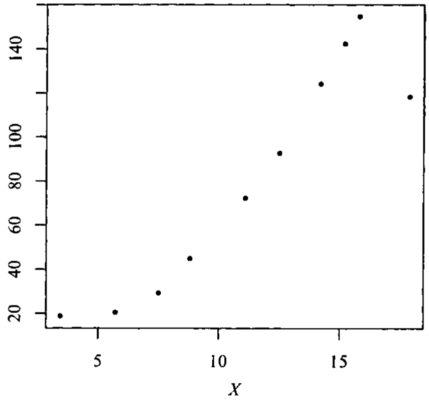

**Figure 13.1** A scatter plot of $Y$ versus $X$ for the Data Set in Table 13.4.

**Table 13.6** Robust Regression Quadratic Fit for the Data Set in Table13.4

| Variable | Coefficient | s.e. | $z$-test | $p$-value |
|---|---|---|---|---|
| Constant | 15.2614 | 0.8003 | 19.07 | 0.000 |
| $X$ | -3.5447 | 0.1829 | -19.38 | 0.000 |
| $X^2$ | 0.783099 | 0.0099 | 79.10 | 0.000 |
| $n = 10$ | $R^2 = 0.837$ | $R_a^2 = 0.780$ | $\hat{\sigma} = 41.425$ | $d.f. = 7$ |

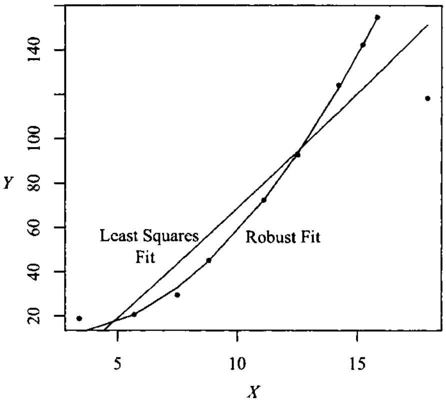

**Figure 13.2** Least Squares and Robust Fits Superposed on the Scatter Plot of $Y$ versus $X$ for the Data Set in Table 13.4.

## 13.7 DISTRIBUTION OF PCB IN U.S. BAYS

Table 13.7 gives the Poly Chlorinated Biphenyl (PCB) concentrations in 1984 and 1985 in 29 U.S. bays and estuaries. PCB is a health hazard found in water from industrial waste and city drainage. The concentration is measured in parts per billion. We do not include any bay or estuary in which no PCB was detected in the two years. We want to study the relationship of PCB level for the two years. The data is taken from *Environmental Quality* 1987–1988, published by the Council on Environmental Quality. An exhaustive description of the data along with a thorough analysis can be found in Chatterjee, Handcock, and Simonoff (1995).

To overcome the skewness of the data we will transform the data. We will work with the logarithm of the PCB concentrations. The result of the least squares fit is given in Table 13.8.

The fit is problematic. Two observations (Boston Harbor and Delaware Bay) require attention. Both are outliers with standardized residuals of –2.12 and 4.11, respectively. Boston Harbor is a high leverage point with a large value of Cook's Distance. Delaware Bay is not a high leverage point, but has a high value of Cook's Distance. These two points have a significant effect on the fit. We will examine the relationship between the PCB levels in two succeeding years when these two aberrant points are removed. The regression result with the two observations deleted is given in Table 13.9.

This fit has no problems, and is an acceptable description of the relationship of the PCB levels between the years 1984 and 1985. It should be pointed out that this relationship does not hold for Boston Harbor and Delaware Bay. These two bays present special conditions and should be investigated. The deletion of the two points from the fit gives us a better picture of the overall relationship of PCB levels for U.S. bays and estuaries.

**Table 13.7** Distribution of PCB in 1984 and 1985 in U.S. Bays

| Bay | PCB84 | PCB85 | ln(PCB84) | ln(PCB85) |
|---|---|---|---|---|
| Casco Bay | 95.3 | 77.55 | 4.55682 | 4.35092 |
| Merrimack River | 53.0 | 29.23 | 3.96973 | 3.37520 |
| Salem Harbor | 533.6 | 403.10 | 6.27961 | 5.99918 |
| Boston Harbor | 17104.9 | 736.00 | 9.74712 | 6.60123 |
| Buzzards' Bay | 308.5 | 192.15 | 5.73159 | 5.25828 |
| Narragansett Bay | 160.0 | 220.60 | 5.07492 | 5.39635 |
| E. Long Island Sound | 10.0 | 8.62 | 2.30259 | 2.15409 |
| W. Long Island Sound | 234.4 | 174.31 | 5.45716 | 5.16084 |
| Raritan Bay | 443.9 | 529.28 | 6.09558 | 6.27152 |
| Delaware Bay | 2.5 | 130.67 | 0.91629 | 4.87268 |
| Lower Chesapeake Bay | 51.0 | 39.74 | 3.93183 | 3.68236 |
| Charleston Harbor | 9.1 | 8.43 | 2.20827 | 2.13180 |
| St. Johns River | 140 | 120.04 | 4.94164 | 4.78783 |
| Apalachicola Bay | 12.0 | 11.93 | 2.48491 | 2.47906 |
| Mississippi R. Delta | 34.0 | 30.14 | 3.52636 | 3.40585 |
| San Diego Harbor | 422.1 | 531.67 | 6.04524 | 6.27602 |
| San Diego Bay | 6.7 | 9.30 | 1.90806 | 2.23001 |
| Dana Point | 7.1 | 5.74 | 1.95445 | 1.74746 |
| Seal Beach | 46.7 | 46.47 | 3.84396 | 3.83881 |
| San Pedro Canyon | 159.6 | 176.90 | 5.07242 | 5.17558 |
| Santa Monica Bay | 14.0 | 13.69 | 2.63906 | 2.61667 |
| Bodega Bay | 4.2 | 4.89 | 1.43031 | 1.58719 |
| Coos Bay | 3.2 | 6.60 | 1.16002 | 1.88707 |
| Columbia River Mouth | 8.8 | 6.73 | 2.17134 | 1.90658 |
| Nisqually Beach | 4.2 | 4.28 | 1.44220 | 1.45395 |
| Commencement Bay | 20.6 | 20.50 | 3.02529 | 3.02042 |
| Elliott Bay | 20.6 | 20.50 | 3.02529 | 3.02042 |
| Lutak Inlet | 5.5 | 5.80 | 1.70475 | 1.75786 |
| Nahku Bay | 6.6 | 5.08 | 1.88707 | 1.62531 |

**Table 13.8** Least Squares Regression of ln(PCB85) on ln(PCB84) for the Data Set in Table 13.7

ANOVA Table

| Source | Sum of Squares | $d.f.$ | Mean Square | $F$-test | $p$-value |
|---|---|---|---|---|---|
| Regression | 59.605 | 1 | 59.605 | 87.96 | 0.000 |
| Residual Error | 18.296 | 27 | 0.678 | | |

Coefficients Table

| Variable | Coefficient | s.e. | $t$-test | $p$-value |
|---|---|---|---|---|
| Constant | 1.001 | 0.315 | 3.17 | 0.004 |
| ln(PCB84) | 0.718 | 0.077 | 9.38 | 0.000 |
| $n = 29$ | $R^2 = 0.765$ | $R_a^2 = 0.756$ | $\hat{\sigma} = 0.823$ | $d.f. = 27$ |

**Table 13.9** Least Squares Regression of ln(PCB85) on ln(PCB84) for the Data Set in Table 13.7, when Boston and Delaware Are Deleted

ANOVA Table

| Source | Sum of Squares | $d.f.$ | Mean Square | $F$-test | $p$-value |
|---|---|---|---|---|---|
| Regression | 64.712 | 1 | 64.712 | 908.24 | 0.000 |
| Residual Error | 1.781 | 25 | 0.071 | | |

Coefficients Table

| Variable | Coefficient | s.e. | $t$-test | $p$-value |
|---|---|---|---|---|
| Constant | 0.093 | 0.122 | 0.76 | 0.456 |
| ln(PCB84) | 0.960 | 0.032 | 30.14 | 0.000 |
| $n = 27$ | $R^2 = 0.973$ | $R_a^2 = 0.972$ | $\hat{\sigma} = 0.267$ | $d.f. = 25$ |

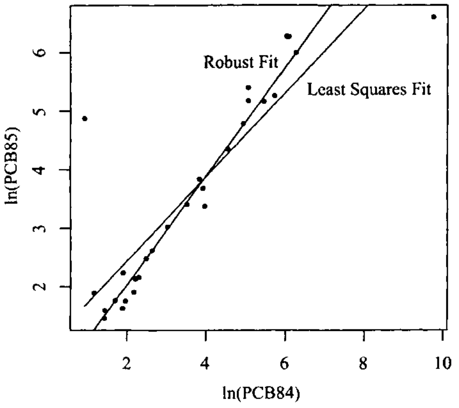

**Figure 13.3** Least squares and robust fits superposed on scatter plot of ln(PCB85) versus ln(PCB84) for the data set in Table 13.7.

Robust regression provides us an alternative approach. Using the robust regression algorithm outlined earlier, we get the following fitted model:

$$\ln(\text{PCB85}) = 0.175 + 0.927 \ln(\text{PCB84}). \tag{13.7}$$

The standard errors of the two coefficients in (13.7) are 0.25 and 0.0056, respectively. The robust regression applied to the complete data gives results similar to those obtained by the diagnostic prescription of deleting the two observations. The weights given to the deleted points are very small compared to the other data points. This is done mechanically by the rules built into the algorithm. The robust procedure does not require a detailed analysis of regression diagnostics. The final weights used in the iteration point out the problematic observations for further investigation.

Figure 13.3 shows the least squares and robust fit superposed on the scatter plot. It is seen that the least squares line is influenced strongly by the outliers and high leverage points. The robust fit tracks the data more accurately without being unduly influenced by leverage points and outliers.

As can be seen from (13.6), in the presence of masking, observations in a high-leverage group tend to have large weights because they tend to have small values of $p_{ii}$. For this reason the Chatterjee-Mächler procedure is not very effective in the presence of masking. The algorithm we have given has been extended to cover problems of masking and swamping, but that is beyond the scope of this book. For details, the reader is referred to Billor, Chatterjee, and Hadi, (2006).

Much work has been done on robust regression but it is not widely used in practice. We hope this brief exposition will bring it to public attention.

## EXERCISES

**13.1** Use the data on injury incidents in airlines given in Table 6.6 to fit a Poisson Regression model. Compare the three fits (least squares, transformed least squares, and Poisson), and decide which procedure provides best description of the data.

**13.2** Using the data on the distribution of PCB in U.S. bays and estuaries given in Table 13.7, do a thorough analysis that relates the 1985 PCB levels to 1984 levels. Compare the results of your analysis to the robust fit given in the text.

**13.3** Use the dataset given in Table 3.3 and regress $Y$ on $X_1$ and $X_3$ by least squares and the robust procedure. Verify that both procedures give similar results.

**13.4** Use the data on Magazine Advertising given in Table 6.15 and regress $\ln R$ on $\ln P$ using least squares. Observations 15, 22, 23, 41 are problematic. Do these points have any special feature? Show that the robust fit for the full data set gives results comparable to the least squares results after deleting the 4 points.

**13.5** Use the data on Field-goal Kicking given in Table 12.16:

(a) Fit a Poison regression model to relate Success with the Distance from which the kick is taken. Use Attempts as offset (See Page 347).

(b) Fit a logistic model relating the probability of a successful kick to the distance from which the kick is taken.

(c) Show that the logistic model gives a better fit than the Poisson regression model.