Chapter 1

Introduction

All models are wrong, but some are useful.
Box (1979)

In *data analysis*, an investigator is presented with a *problem* and *data* from some *population*. The population might be the collection of all possible outcomes from an experiment while the problem might be predicting a future value of the response variable Y or summarizing the relationship between Y and the $p \times 1$ vector of predictor variables $\boldsymbol{x}$. A **statistical model** is used to provide a useful approximation to some of the important underlying characteristics of the population which generated the data. Models for *regression* and *multivariate location and dispersion* are frequently used.

Model building is an *iterative process*. Given the problem and data but no model, the model building process can often be aided by graphs that help visualize the relationships between the different variables in the data. Then a statistical model can be proposed. This model can be fit and inference performed. Then *diagnostics* from the fit can be used to check the assumptions of the model. If the assumptions are not met, then an alternative model can be selected. The fit from the new model is obtained, and the cycle is repeated.

Definition 1.1. *Regression* investigates how the response variable Y changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of predictors. Often this *conditional distribution* $Y|\boldsymbol{x}$ is described by a *1D regression model*, where Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $\boldsymbol{\beta}^T\boldsymbol{x}$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x}. \tag{1.1}$$

The class of 1D models is very rich. Generalized linear models (GLMs) are a special case of 1D regression, and an important class of parametric or semiparametric 1D regression models has the form

$$Y_i = g(\boldsymbol{x}_i^T \boldsymbol{\beta}, e_i) \tag{1.2}$$

for $i = 1, ..., n$ where g is a bivariate function, $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and e_i is a random error. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed) from a distribution that is known except for a scale parameter. For example, the e_i's might be iid from a normal (Gaussian) distribution with *mean* 0 and unknown *standard deviation* σ. For this Gaussian model, estimation of $\boldsymbol{\beta}$ and σ is important for inference and for predicting a future value of the response variable Y_f given a new vector of predictors $\boldsymbol{x}_f$.

Many of the most used statistical models are 1D regression models. An additive error *single index model* uses

$$g(\boldsymbol{x}^T \boldsymbol{\beta}, e) = m(\boldsymbol{x}^T \boldsymbol{\beta}) + e \tag{1.3}$$

and an important special case is *multiple linear regression*

$$Y = \boldsymbol{x}^T \boldsymbol{\beta} + e \tag{1.4}$$

where m is the identity function. The *response transformation model* uses

$$g(\boldsymbol{\beta}^T \boldsymbol{x}, e) = t^{-1}(\boldsymbol{\beta}^T \boldsymbol{x} + e) \tag{1.5}$$

where t^{-1} is a one to one (typically monotone) function. Hence

$$t(Y) = \boldsymbol{\beta}^T \boldsymbol{x} + e. \tag{1.6}$$

Several important *survival models* have this form. In a *1D binary regression model*, the $Y|\boldsymbol{x}$ are independent Bernoulli$[\rho(\boldsymbol{\beta}^T \boldsymbol{x})]$ random variables where

$$P(Y = 1|\boldsymbol{x}) \equiv \rho(\boldsymbol{\beta}^T \boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x}) \tag{1.7}$$

In particular, the *logistic regression model* uses

$$\rho(\boldsymbol{\beta}^T \boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x})}.$$

In the literature, the response variable is sometimes called the dependent variable while the predictor variables are sometimes called carriers, covariates, explanatory variables, or independent variables. The ith *case* $(Y_i, \boldsymbol{x}_i^T)$ consists of the values of the response variable Y_i and the predictor variables $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p})$ where p is the number of predictors and $i = 1, ..., n$. The *sample size* n is the number of cases.

Box (1979) warns that "All models are wrong, but some are useful." For example the function g or the error distribution could be misspecified. *Diagnostics* are used to check whether model assumptions such as the form of g and the proposed error distribution are reasonable. Often diagnostics use *residuals* r_i. If m is known, then the single index model uses

$$r_i = Y_i - m(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$. Sometimes several estimators $\hat{\boldsymbol{\beta}}_j$ could be used. Often $\hat{\boldsymbol{\beta}}_j$ is computed from a subset of the n cases or from different fitting methods. For example, ordinary least squares (OLS) and least absolute deviations (L_1) could be used to compute $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{L_1}$, respectively. Then the corresponding residuals can be plotted.

Exploratory data analysis (EDA) can be used to find useful models when the form of the regression or multivariate model is unknown. For example, suppose g is a monotone function t^{-1} :

$$Y = t^{-1}(\boldsymbol{x}^T \boldsymbol{\beta} + e). \tag{1.8}$$

Then the transformation

$$Z = t(Y) = \boldsymbol{x}^T \boldsymbol{\beta} + e \tag{1.9}$$

follows a multiple linear regression model, and the goal is to find t.

Robust statistics can be tailored to give useful results even when a certain specified model assumption is incorrect. An important class of robust statistics can give useful results when the assumed model error distribution is incorrect. This class of statistics is useful when *outliers,* observations far from the bulk of the data, are present. The class is also useful when the error distribution has heavier tails than the assumed error distribution, eg if the assumed distribution is normal but the actual distribution is Cauchy

or double exponential. This type of robustness is often called *distributional robustness.*

Another class of robust statistics, known as *regression graphics*, gives useful results when the 1D regression model (1.1) is misspecified or unknown. Let the estimated sufficient predictor $ESP = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{OLS}$ where $\hat{\boldsymbol{\beta}}_{OLS}$ is obtained from the OLS multiple linear regression of Y on $\boldsymbol{x}$. Then a **very important** regression graphics result is that the *response plot* of the ESP versus Y can often be used to visualize the conditional distribution of $Y|\boldsymbol{\beta}^T\boldsymbol{x}$.

Distributionally robust statistics and regression graphics have amazing applications for regression, multivariate location and dispersion, diagnostics, and EDA. This book illustrates some of these applications and investigates the interrelationships between these two classes of robust statistics.

1.1 Outlier....s

The main message of this book is that robust regression is extremely useful in identifying outliers ....
Rousseeuw and Leroy (1987, p. vii)

Following Staudte and Sheather (1990, p. 32), we define an *outlier* to be an observation that is far from the bulk of the data. Similarly, Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 21) define outliers to be observations which deviate from the pattern set by the majority of the data. Typing and recording errors may create outliers, and a data set can have a large proportion of outliers if there is an omitted categorical variable (eg gender, species, or geographical location) where the data behaves differently for each category. Outliers should always be examined to see if they follow a pattern, are recording errors, or if they could be explained adequately by an alternative model. Recording errors can sometimes be corrected and omitted variables can be included, but often there is no simple explanation for a group of data which differs from the bulk of the data.

Although outliers are often synonymous with "bad" data, they are *frequently the most important part* of the data. Consider, for example, finding the person whom you want to marry, finding the best investments, finding the locations of mineral deposits, and finding the best students, teachers, doctors, scientists, or other *outliers in ability.* Huber (1981, p. 4) states that outlier resistance and distributional robustness are synonymous while

Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 36) state that the first and most important step in robustification is the rejection of distant outliers.

In the literature there are two important paradigms for *robust procedures*. The *perfect classification paradigm* considers a *fixed* data set of n cases of which $0 \leq d < n/2$ are outliers. The key assumption for this paradigm is that the robust procedure *perfectly classifies* the cases into outlying and non-outlying (or "clean") cases. The outliers should *never* be blindly discarded. Often the clean data and the outliers are analyzed separately.

The *asymptotic paradigm* uses an asymptotic distribution to approximate the distribution of the estimator when the sample size n is large. An important example is the *central limit theorem* (CLT): let $Y_1, ..., Y_n$ be iid with mean μ and standard deviation σ; ie, the Y_i's follow the *location model*

$$Y = \mu + e.$$

Then

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Hence the *sample mean* $\overline{Y}_n$ is asymptotically normal $\mathrm{AN}(\mu, \sigma^2/n)$.

For this paradigm, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large n must be for the approximation to be useful. Moreover, the (asymptotic) standard error (SE), an estimator of the asymptotic standard deviation, must be computable if the estimator is to be useful for inference. Note that the sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the $\mathrm{SE} = S/\sqrt{n}$ where S is the *sample standard deviation.* For many distributions the central limit theorem provides a good approximation if the sample size $n > 30$. Chapter 2 examines the sample mean, standard deviation and robust alternatives.

1.2 Applications

One of the key ideas of this book is that *the data should be examined with several estimators.* Often there are many procedures that will perform well when the model assumptions hold, but no single method can dominate every

other method for every type of model violation. For example, OLS is best for multiple linear regression when the iid errors are normal (Gaussian) while L_1 is best if the errors are double exponential. Resistant estimators may outperform classical estimators when outliers are present but be far worse if no outliers are present.

Different multiple linear regression estimators tend to estimate $\boldsymbol{\beta}$ in the iid constant variance symmetric error model, but otherwise each estimator estimates a different parameter. Hence a plot of the residuals or fits from different estimators should be useful for detecting departures from this very important model. The "RR plot" is a *scatterplot matrix* of the residuals from several regression fits. Tukey (1991) notes that such a plot will be linear with slope one if the model assumptions hold. Let the ith residual from the jth fit $\hat{\boldsymbol{\beta}}_j$ be $r_{i,j} = Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j$ where the superscript T denotes the transpose of the vector and $(Y_i, \boldsymbol{x}_i^T)$ is the ith observation. Then

$$\|r_{i,1} - r_{i,2}\| = \|\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)\|$$

$$\leq \|\boldsymbol{x}_i\| \ (\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\| + \|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}\|).$$

The RR plot is simple to use since if $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ have good convergence rates and if the predictors $\boldsymbol{x}_i$ are bounded, then the residuals will cluster tightly about the *identity line* (the unit slope line through the origin) as n increases to ∞. For example, plot the least squares residuals versus the L_1 residuals. Since OLS and L_1 are consistent, the plot should be linear with slope one when the regression assumptions hold, but the plot should not have slope one if there are Y–outliers since L_1 resists these outliers while OLS does not. Making a scatterplot matrix of the residuals from OLS, L_1, and several other estimators can be very informative.

The FF plot is a scatterplot matrix of fitted values and the response. A plot of fitted values versus the response is called a response plot. For square plots, outliers tend to be $\sqrt{2}$ times further away from the bulk of the data in the OLS response plot than in the OLS residual plot because outliers tend to stick out for both the fitted values and the response.

Example 1.1. Gladstone (1905–1906) attempts to estimate the *weight* of the human brain (measured in grams after the death of the subject) using simple linear regression with a variety of predictors including *age* in years, *height* in inches, *head height* in mm, *head length* in mm, *head breadth* in mm, *head circumference* in mm, and *cephalic index* (divide the breadth of the head

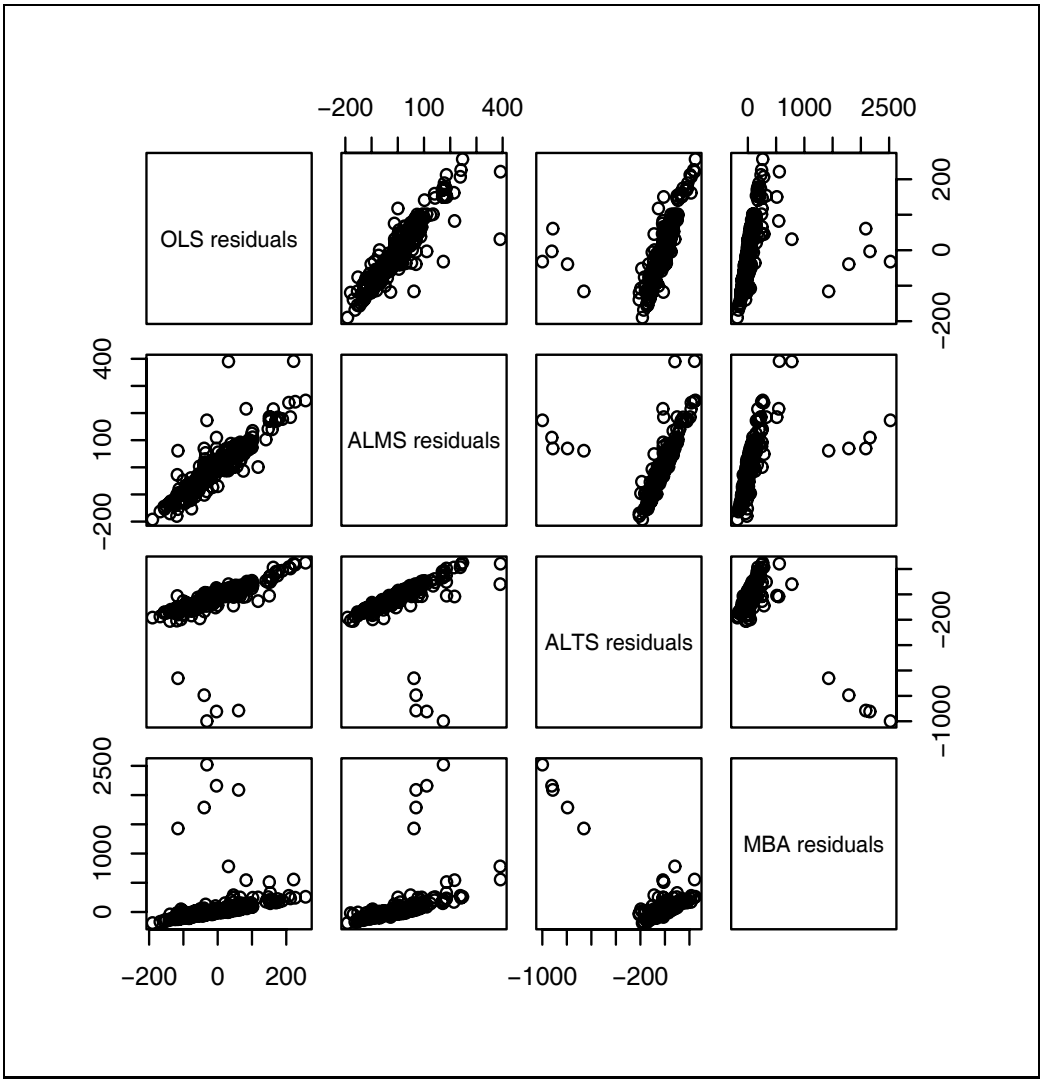

Figure 1.1: RR Plot for Gladstone data

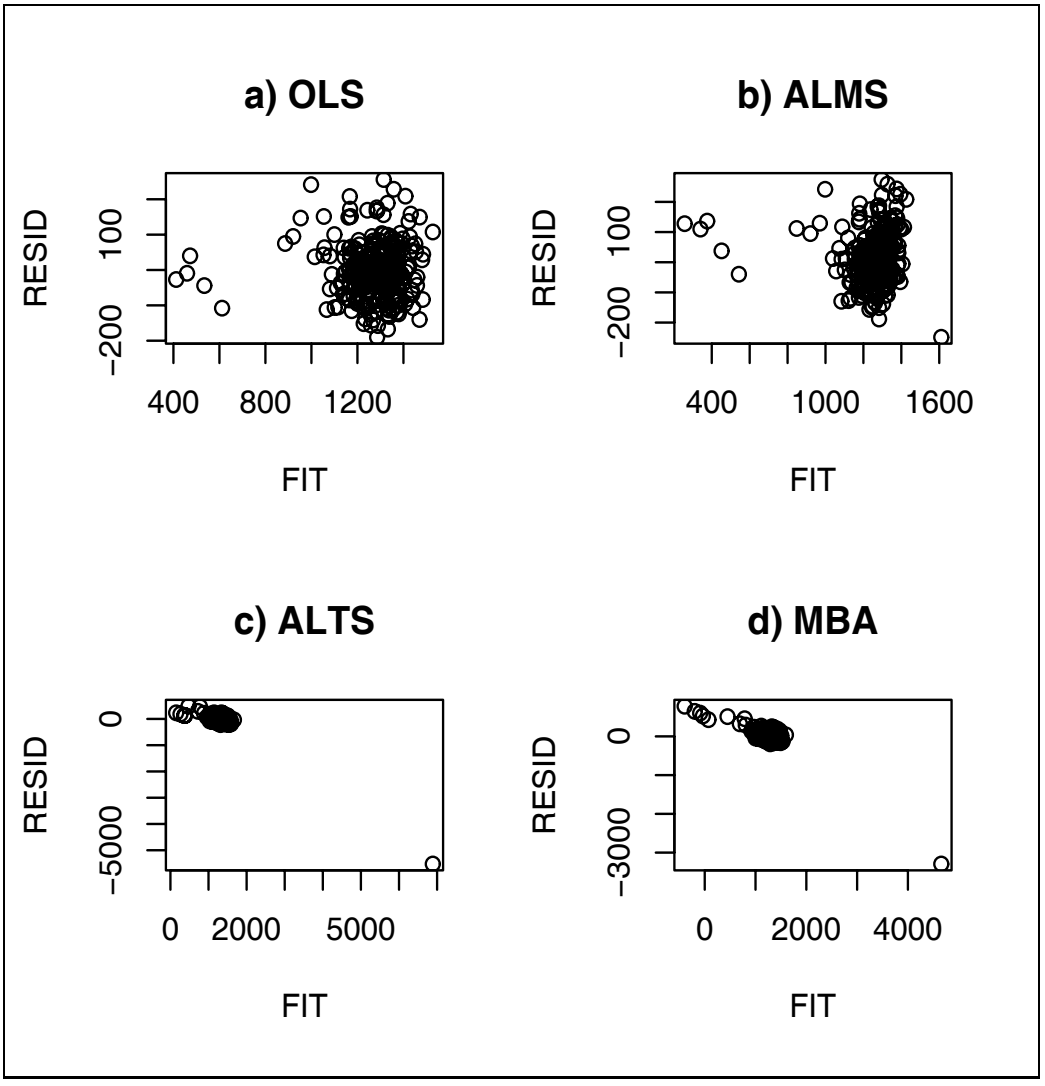

Figure 1.2: Gladstone data where case 119 is a typo

by its length and multiply by 100). The *sex* (coded as 0 for females and 1 for males) of each subject was also included. The variable *cause* was coded as 1 if the cause of death was acute, as 3 if the cause of death was chronic, and coded as 2 otherwise. A variable *ageclass* was coded as 0 if the age was under 20, as 1 if the age was between 20 and 45, and as 3 if the age was over 45. *Head size* is the product of the *head length*, *head breadth*, and *head height*.

The data set contains 276 cases, and we decided to use multiple linear regression to predict brain weight using the six head measurements height, length, breadth, size, cephalic index and circumference as predictors. Cases 188 and 239 were deleted because of missing values. There are five infants (cases 238, 263-266) of age less than 7 months that are $\boldsymbol{x}$-outliers. Nine toddlers were between 7 months and 3.5 years of age, four of whom appear to be $\boldsymbol{x}$-outliers (cases 241, 243, 267, and 269).

Figure 1.1 shows an RR plot comparing the OLS, L_1, ALMS, ALTS and MBA fits. ALMS is the default version of the *R/Splus* function `lmsreg` while ALTS is the default version of `ltsreg`. The three estimators ALMS, ALTS, and MBA are described further in Chapter 7. Figure 1.1 was made with a 2007 version of *R* and the *rpack* function `rrplot2`. ALMS, ALTS and MBA depend on the seed (in *R*) and so the estimators change with each call of `rrplot2`. Nine cases stick out in Figure 1.1, and these points correspond to five infants and four toddlers that are $\boldsymbol{x}$-outliers. The OLS fit may be the best since the OLS fit to the bulk of the data passes through the five infants, suggesting that these cases are "good leverage points."

An obvious application of outlier resistant methods is the detection of outliers. Generally robust and resistant methods can only detect certain configurations of outliers, and the ability to detect outliers rapidly decreases as the sample size n and the number of predictors p increase. When the Gladstone data was first entered into the computer, the variable *head length* was inadvertently entered as 109 instead of 199 for case 119. Residual plots are shown in Figure 1.2. For the three resistant estimators, case 119 is in the lower left corner.

Example 1.2. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. *Height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, numbers 62–66, were reported to be about 0.75

inches tall with head lengths well over five feet! Figure 7.1, made with *Splus* and the *rpack* function `rrplot`, shows that the outliers were accommodated by the all of the estimators, except MBA. Figure 6.2 shows that the outliers are much easier to detect with the OLS response and residual plots.

The Buxton data is also used to illustrate robust multivariate location and dispersion estimators in Example 11.4 and to illustrate a graphical diagnostic for multivariate normality in Example 11.2.

Example 1.3. Now suppose that the only variable of interest in the Buxton data is $Y = height$. How should the five adult heights of 0.75 inches be handled? These observed values are impossible, and could certainly be deleted if it was felt that the recording errors were made at random; however, the outliers occurred on consecutive cases: 62–66. If it is reasonable to assume that the true heights of cases 62–66 are a random sample of five heights from the same population as the remaining heights, then the outlying cases could again be deleted. On the other hand, what would happen if cases 62–66 were the five tallest or five shortest men in the sample? In particular, how are point estimators and confidence intervals affected by the outliers? Chapter 2 will show that classical location procedures based on the sample mean and sample variance are adversely affected by the outliers while procedures based on the sample median or the 25% trimmed mean can frequently handle a small percentage of outliers.

For the next application, assume that the population that generates the data is such that a certain proportion γ of the cases will be easily identified but randomly occurring unexplained outliers where $\gamma < \alpha < 0.2$, and assume that remaining proportion $1 - \gamma$ of the cases will be well approximated by the statistical model.

A common suggestion for examining a data set that has unexplained outliers is to run the analysis on the full data set and to run the analysis on the "cleaned" data set with the outliers deleted. Then the statistician may consult with subject matter experts in order to decide which analysis is "more appropriate." Although the analysis of the cleaned data may be useful for describing the bulk of the data, the analysis may not very useful if prediction or description of the entire population is of interest.

Similarly, the analysis of the full data set will likely be unsatisfactory for prediction since numerical statistical methods tend to be inadequate when outliers are present. Classical estimators will frequently fit neither the bulk of

the data nor the outliers well, while an analysis from a good practical robust estimator (if available) should be similar to the analysis of the cleaned data set.

Hence neither of the two analyses alone is appropriate for prediction or description of the actual population. Instead, information from both analyses should be used. The cleaned data will be used to show that the bulk of the data is well approximated by the statistical model, but the full data set will be used along with the cleaned data for prediction and for description of the entire population.

To illustrate the above discussion, consider the multiple linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1.10}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors. The ith case $(Y_i, \boldsymbol{x}_i^T)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element Y_i of $\boldsymbol{Y}$. Assume that the errors e_i are iid zero mean normal random variables with variance σ^2.

Finding prediction intervals for future observations is a standard problem in regression. Let $\hat{\boldsymbol{\beta}}$ denote the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ and let

$$MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-p}$$

where $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is the ith residual. Following Neter, Wasserman, Nachtsheim and Kutner (1996, p. 235), a $100(1-\alpha)\%$ prediction interval (PI) for a new observation Y_f corresponding to a vector of predictors $\boldsymbol{x}_f$ is given by

$$\hat{Y}_f \pm t_{n-p,1-\alpha/2} se(pred) \tag{1.11}$$

where $\hat{Y}_f = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$, $P(t \leq t_{n-p,1-\alpha/2}) = 1 - \alpha/2$ where t has a t distribution with $n-p$ degrees of freedom, and

$$se(pred) = \sqrt{MSE(1 + \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f)}.$$

For discussion, suppose that $1 - \gamma = 0.92$ so that 8% of the cases are outliers. If interest is in a 95% PI, then using the full data set will fail because outliers are present, and using the cleaned data set with the outliers deleted will fail since only 92% of future observations will behave like the "clean" data.

A simple remedy is to create a nominal $100(1-\alpha)\%$ PI for future cases from this population by making a classical $100(1-\alpha^*)$ PI from the clean cases where

$$1-\alpha^* = (1-\alpha)/(1-\gamma). \tag{1.12}$$

Assume that the data have been perfectly classified into n_c clean cases and n_o outlying cases where $n_c + n_o = n$. Also assume that no outlying cases will fall within the PI. Then the PI is valid if Y_f is clean, and

$$\mathrm{P}(Y_f \text{ is in the PI}) = \mathrm{P}(Y_f \text{ is in the PI and clean}) =$$

$$\mathrm{P}(Y_f \text{ is in the PI} \mid Y_f \text{ is clean})\ \mathrm{P}(Y_f \text{ is clean}) = (1-\alpha^*)(1-\gamma) = (1-\alpha).$$

The formula for this PI is then

$$\hat{Y}_f \pm t_{n_c-p,1-\alpha^*/2} se(pred) \tag{1.13}$$

where $\hat{Y}_f$ and $se(pred)$ are obtained after performing OLS on the n_c clean cases. For example, if $\alpha = 0.1$ and $\gamma = 0.08$, then $1-\alpha^* \approx 0.98$. Since γ will be estimated from the data, the coverage will only be approximately valid. The following example illustrates the procedure.

Example 1.4. STATLIB provides a data set (see Johnson 1996) that is available from the website (http://lib.stat.cmu.edu/datasets/bodyfat). The data set includes 252 cases, 14 predictor variables, and a response variable $Y =$ *bodyfat*. The correlation between Y and the first predictor $x_1 =$ *density* is extremely high, and the plot of x_1 versus Y looks like a straight line except for four points. If simple linear regression is used, the residual plot of the fitted values versus the residuals is curved and five outliers are apparent. The curvature suggests that x_1^2 should be added to the model, but the least squares fit does not resist outliers well. If the five outlying cases are deleted, four more outliers show up in the plot. The residual plot for the quadratic fit looks reasonable after deleting cases 6, 48, 71, 76, 96, 139, 169, 182 and 200. Cases 71 and 139 were much less discrepant than the other seven outliers.

These nine cases appear to be *outlying at random*: if the purpose of the analysis was description, we could say that a quadratic fits 96% of the cases well, but 4% of the cases are not fit especially well. If the purpose of the analysis was prediction, deleting the outliers and then using the clean data to find a 99% prediction interval (PI) would not make sense if 4% of future cases are outliers. To create a nominal 90% PI for future cases from this population,

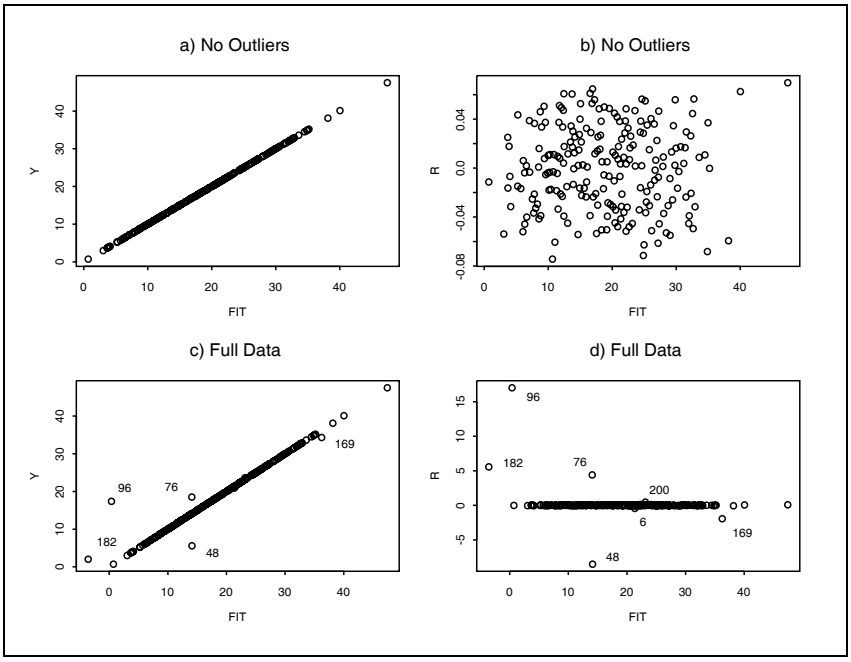

Figure 1.3: Plots for Summarizing the Entire Population

make a classical $100(1-\alpha^*)$ PI from the clean cases where $1-\alpha^* = 0.9/(1-\gamma)$. For the bodyfat data, we can take $1-\gamma \approx 1-9/252 \approx 0.964$ and $1-\alpha^* \approx 0.94$. Notice that $(0.94)(0.96) \approx 0.9$.

Figure 1.3 is useful for presenting the analysis. The top two plots have the nine outliers deleted. Figure 1.4a is a response plot of the fitted values $\hat{Y}_i$ versus the response Y_i while Figure 1.3b is a residual plot of the fitted values $\hat{Y}_i$ versus the residuals r_i. These two plots suggest that the multiple linear regression model fits the bulk of the data well. Next consider using weighted least squares where cases 6, 48, 71, 76, 96, 139, 169, 182 and 200 are given weight zero and the remaining cases weight one. Figure 1.3c and 1.3d give the response plot and residual plot for the entire data set. Notice that seven of the nine outlying cases can be seen in these plots.

The classical 90% PI using $\boldsymbol{x} = (1, 1, 1)^T$ and all 252 cases was $\hat{Y}_h \pm t_{249,0.95} se(pred) = 46.3152 \pm 1.651(1.3295) = (44.12, 48.51)$. When the 9 outliers are deleted, $n_c = 243$ cases remain. Hence the 90% PI using Equation (1.13) with 9 cases deleted was $\hat{Y}_h \pm t_{240,0.97} se(pred) = 44.961 \pm 1.88972(0.0371) = (44.89, 45.03)$. The classical PI is about 31 times longer than the new PI.

For the next application, consider a response transformation model

$$Y = t_{\lambda_o}^{-1}(\boldsymbol{x}^T\boldsymbol{\beta} + e)$$

where $\lambda_o \in \Lambda = \{0, \pm 1/4, \pm 1/3, \pm 1/2, \pm 2/3, \pm 1\}$. Then

$$t_{\lambda_o}(Y) = \boldsymbol{x}^T\boldsymbol{\beta} + e$$

follows a multiple linear regression (MLR) model where the response variable $Y_i > 0$ and the *power transformation family*

$$t_\lambda(Y) \equiv Y^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda} \tag{1.14}$$

for $\lambda \neq 0$ and $Y^{(0)} = \log(Y)$.

The following simple graphical method for selecting response transformations can be used with any good classical, robust or Bayesian MLR estimator. Let $Z_i = t_\lambda(Y_i)$ for $\lambda \neq 1$, and let $Z_i = Y_i$ if $\lambda = 1$. Next, perform the multiple linear regression of Z_i on $\boldsymbol{x}_i$ and make the "response plot" of $\hat{Z}_i$ versus Z_i. If the plotted points follow the identity line, then take $\lambda_o = \lambda$. One plot is made for each of the eleven values of $\lambda \in \Lambda$, and if more than one value of λ works, take the simpler transformation or the transformation that makes the most sense to subject matter experts. (Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of λ_o by adding $\hat{\lambda}$ to Λ.) The following example illustrates the procedure.

Example 1.5. Box and Cox (1964) present a textile data set where samples of worsted yarn with different levels of the three factors were given a cyclic load until the sample failed. The goal was to understand how $Y =$ *the number of cycles to failure* was related to the predictor variables. Figure 1.4 shows the forward response plots for two MLR estimators: OLS and the *R/Splus* function `lmsreg`. Figures 1.4a and 1.4b show that a response transformation is needed while 1.4c and 1.4d both suggest that $\log(Y)$ is the appropriate response transformation. Using OLS and a resistant estimator as in Figure 1.4 may be very useful if outliers are present.

The textile data set is used to illustrate another graphical method for selecting the response transformation t_λ in Section 5.1.

Another important application is *variable selection*: the search for a subset of predictor variables that can be deleted from the model without important loss of information. Section 5.2 gives a graphical method for assessing

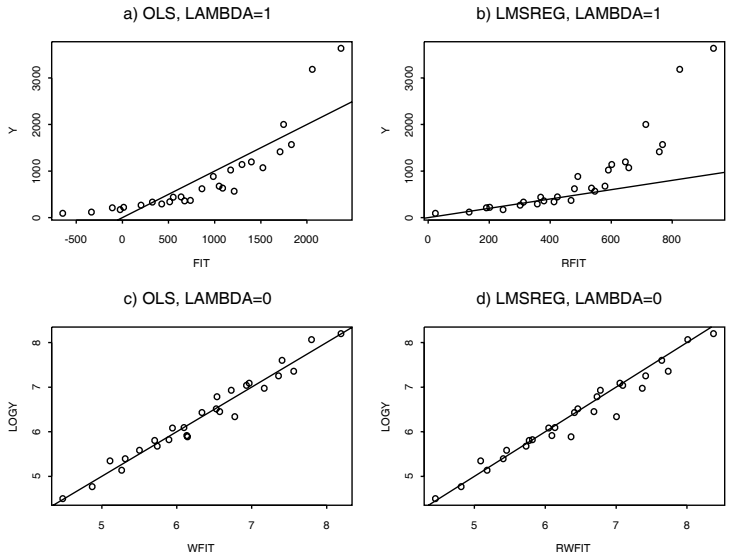

Figure 1.4: OLS and LMSREG Suggest Using log(Y) for the Textile Data

variable selection for multiple linear regression models while Section 12.4 gives a similar method for 1D regression models.

The basic idea is to obtain fitted values from the full model and the candidate submodel. If the candidate model is good, then the plotted points in a plot of the submodel fitted values versus the full model fitted values should follow the identity line. In addition, a similar plot should be made using the residuals.

A problem with this idea is how to select the candidate submodel from the nearly 2^p potential submodels. One possibility would be to try to order the predictors in importance, say $x_1, ..., x_p$. Then let the kth model contain the predictors $x_1, x_2, ..., x_k$ for $k = 1, ..., p$. If the predicted values from the submodel are highly correlated with the predicted values from the full model, then the submodel is "good." This idea is useful even for extremely complicated models: the estimated sufficient predictor of a "good submodel" should be highly correlated with the ESP of the full model. Section 12.4 will show that the all subsets, forward selection and backward elimination techniques of variable selection for multiple linear regression will often work for the 1D regression model provided that the Mallows' C_p criterion is used.

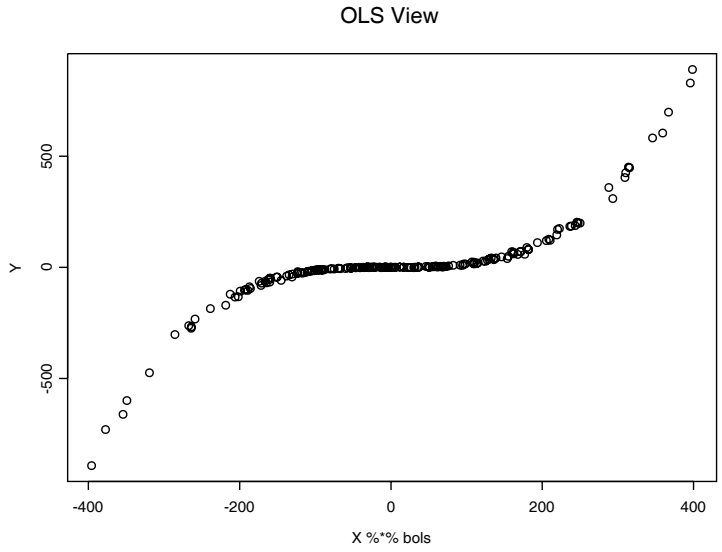

Figure 1.5: Response Plot or OLS View for $m(u) = u^3$

Example 1.6. The Boston housing data of Harrison and Rubinfeld (1978) contains 14 variables and 506 cases. Suppose that the interest is in predicting the *per capita crime rate* from the other variables. Variable selection for this data set is discussed in much more detail in Section 12.4.

Another important topic is fitting 1D regression models given by Equation (1.2) where g and $\boldsymbol{\beta}$ are both unknown. Many types of plots will be used in this text and a plot of x versus y will have x on the horizontal axis and y on the vertical axis. This notation is also used by the software packages *Splus* (MathSoft 1999ab) and *R,* the free version of *Splus* available from (www.r-project.org/). The *R/Splus* commands

```
X <- matrix(rnorm(300),nrow=100,ncol=3)
Y <- (X %*% 1:3)^3 + rnorm(100)
```

were used to generate 100 trivariate Gaussian predictors $\boldsymbol{x}$ and the response $Y = (\boldsymbol{\beta}^T \boldsymbol{x})^3 + e$ where $e \sim N(0, 1)$. This is a model of form (1.3) where m is the cubic function.

An amazing result is that the unknown function m can often be visualized by the response plot or "OLS view," a plot of the OLS fit (possibly ignoring the constant) versus Y generated by the following commands.

```
bols <- lsfit(X,Y)$coef[-1]
plot(X %*% bols, Y)
```

The OLS view, shown in Figure 1.5, can be used to visualize m and for prediction. Note that Y appears to be a cubic function of the OLS fit and that if the OLS fit $= 0$, then the graph suggests using $\hat{Y} = 0$ as the predicted value for Y. This plot and modifications will be discussed in detail in Chapters 12 and 13.

This section has given a brief outlook of the book. Also look at the preface and table of contents, and then thumb through the remaining chapters to examine the procedures and graphs that will be developed.

1.3 Complements

Many texts simply present statistical models without discussing the process of model building. An excellent paper on statistical models is Box (1979).

The concept of outliers is rather vague. See Barnett and Lewis (1994) and Beckman and Cook (1983) for history.

Outlier rejection is a subjective or objective method for deleting or changing observations which lie far away from the bulk of the data. The modified data is often called the "cleaned data." See Rousseeuw and Leroy (1987, p. 106, 161, 254, and 270), Huber (1981, p. 4-5, and 19), and Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 24, 26, and 31). Data editing, screening, truncation, censoring, Winsorizing, and trimming are all methods for data cleaning. David (1981, ch. 8) surveys outlier rules before 1974, and Hampel, Ronchetti, Rousseeuw and Stahel (1986, Section 1.4) surveys some robust outlier rejection rules. Outlier rejection rules are also discussed in Hampel (1985), Simonoff (1987a,b), and Stigler (1973b).

Robust estimators can be obtained by applying classical methods to the cleaned data. Huber (1981, p. 4-5, 19) suggests that the performance of such methods may be more difficult to work out than that of robust estimators such as the M-estimators, but gives a procedure for cleaning regression data. Staudte and Sheather (1990, p. 29, 136) state that rejection rules are the least understood and point out that for subjective rules where the cleaned data is assumed to be iid, one can not find an unconditional standard error estimate. Even if the data consists of observations which are iid plus outliers, some "good" observations will usually be deleted while some "bad" observations

will be kept. In other words, the assumption of perfect classification is often unreasonable.

The graphical method for response transformations illustrated in Example 1.5 was suggested by Olive (2004b).

Seven important papers that influenced this book are Hampel (1975), Siegel (1982), Devlin, Gnanadesikan and Kettenring (1981), Rousseeuw (1984), Li and Duan (1989), Cook and Nachtsheim (1994) and Rousseeuw and Van Driessen (1999). The importance of these papers will become clearer later in the text.

An excellent text on regression (using 1D regression models such as (1.1)) is Cook and Weisberg (1999a). A more advanced text is Cook (1998a). Also see Cook (2003), Horowitz (1998) and Li (2000).

This text will use the software packages *Splus* (MathSoft (now Insightful) 1999ab) and *R,* a free version of *Splus* available from the website (www.r-project.org/), and *Arc* (Cook and Weisberg 1999a), a free package available from the website (www.stat.umn.edu/arc).

Section 14.2 of this text, Becker, Chambers, and Wilks (1988), and Venables and Ripley (1997) are useful for *R/Splus* users. The websites (www.burns-stat.com/), (http://lib.stat.cmu.edu/S/splusnotes) and (www.isds.duke.edu/computing/S/Snotes/Splus.html) also have useful information.

The Gladstone, Buxton, bodyfat and Boston housing data sets are available from the text's website under the file names *gladstone.lsp, buxton.lsp, bodfat.lsp* and *boston2.lsp.*

1.4 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

1.1*. Using the notation on p. 6, let $\hat{Y}_{i,j} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j$ and show that $\|r_{i,1} - r_{i,2}\| = \|\hat{Y}_{i,1} - \hat{Y}_{i,2}\|$.

R/Splus Problems

1.2*. a) Using the *R/Splus* commands on p. 16-17, reproduce a plot like

Figure 1.6. Once you have the plot you can print it out directly, but it will generally save paper by placing the plots in the *Word* editor.

b) Activate *Word* (often by double clicking on a *Word* icon). Click on the screen and type "Problem 1.2." In *R/Splus,* click on the plot and then press the keys *Ctrl* and *c* simultaneously. This procedure makes a temporary copy of the plot. In *Word,* move the pointer to *Edit* and hold down the leftmost mouse button. This will cause a menu to appear. Drag the pointer down to *Paste*. In the future, these menu commands will be denoted by "Edit>Paste." The plot should appear on the screen. To save your output on your diskette, use the *Word* menu commands "File > Save as." In the **Save in** box select "3 1/2 Floppy(A:)" and in the *File name* box enter HW1d2.doc. To exit from *Word*, click on the "X" in the upper right corner of the screen. In *Word* a screen will appear and ask whether you want to save changes made in your document. Click on *No*. To exit from *R/Splus*, type "q()" or click on the "X" in the upper right corner of the screen and then click on *No*.

c) To see the plot of $10\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ versus Y, use the commands

```
plot(10*X %*% bols, Y)
title("Scaled OLS View")
```

d) Include the plot in *Word* using commands similar to those given in b).

e) Do the two plots look similar? Can you see the cubic function?

1.3*. a) Enter the following *R/Splus* function that is used to illustrate the central limit theorem when the data $Y_1, ..., Y_n$ are iid from an exponential distribution. The function generates a data set of size n and computes $\overline{Y}_1$ from the data set. This step is repeated $nruns = 100$ times. The output is a vector $(\overline{Y}_1, \overline{Y}_2, ..., \overline{Y}_{100})$. A histogram of these means should resemble a symmetric normal density once n is large enough.

```
cltsim <- function(n=100, nruns=100){
ybar <- 1:nruns
for(i in 1:nruns){
  ybar[i] <- mean(rexp(n))}
list(ybar=ybar)}
```

b) The following commands will plot 4 histograms with $n = 1, 5, 25$ and 100. Save the plot in *Word* using the procedure described in Problem 1.2b.

```
> z1 <- cltsim(n=1)
> z5 <- cltsim(n=5)
> z25 <- cltsim(n=25)
> z200 <- cltsim(n=200)
> par(mfrow=c(2,2))
> hist(z1$ybar)
> hist(z5$ybar)
> hist(z25$ybar)
> hist(z200$ybar)
```

c) Explain how your plot illustrates the central limit theorem.

d) Repeat parts a), b) and c), but in part a), change *rexp(n)* to *rnorm(n)*. Then $Y_1, ..., Y_n$ are iid N(0,1) and $\overline{Y} \sim N(0, 1/n)$.

Arc Problems

1.4*. a) Activate *Arc* (Cook and Weisberg 1999a). Generally this will be done by finding the icon for *Arc* or the executable file for *Arc*. Using the mouse, move the pointer (cursor) to the icon and press the leftmost mouse button twice, rapidly. This procedure is known as *double clicking* on the icon. A window should appear with a "greater than" > prompt. The menu *File* should be in the upper left corner of the window. Move the pointer to *File* and hold the leftmost mouse button down. Then the menu will appear. Drag the pointer down to the menu command *load.* Then click on *data*, next click on *ARCG* and then click on *wool.lps.* You will need to use the *slider bar* in the middle of the screen to see the file *wool.lsp*: click on the arrow pointing to the right until the file appears. In the future these menu commands will be denoted by "File > Load > Data > ARCG > wool.lsp." These are the commands needed to activate the file *wool.lsp.*

b) To fit a multiple linear regression model, perform the menu commands "Graph&Fit>Fit linear LS." A window will appear. Double click on *Amp*, *Len* and *Load*. This will place the three variables under the *Terms/Predictors* box. Click once on *Cycles,* move the pointer to the *Response* box and click once. Then *cycles* should appear in the *Response* box. Click on *OK*. If a mistake was made, then you can double click on a variable to move it back to the *Candidates* box. You can also click once on the variable, move the pointer to the *Candidates* box and click. Output should appear on the *Listener screen.*

c) To make a residual plot, use the menu commands "Graph&Fit>Plot of." A window will appear. Double click on *L1: Fit–Values* and then double click on *L1:Residuals*. Then *L1: Fit–Values* should appear in the *H* box and *L1:Residuals* should appear in the *V* box. Click on *OK* to obtain the plot.

d) The graph can be printed with the menu commands "File>Print," but it will generally save paper by placing the plots in the *Word* editor. Activate *Word* (often by double clicking on a *Word* icon). Click on the screen and type "Problem 1.4." In *Arc,* use the menu command "Edit>Copy." In *Word,* use the menu commands "Edit>Paste."

e) In your *Word* document, write "1.4e)" and state whether the points cluster about the horizontal axis with no pattern. If curvature is present, then the multiple linear regression model is not appropriate.

f) After editing your *Word* document, get a printout by clicking on the *printer icon* or by using the menu commands "File>Print." To save your output on your diskette, use the *Word* menu commands "File > Save as." In the **Save in** box select "3 1/2 Floppy(A:)" and in the *File name* box enter HW1d4.doc. To exit from *Word* and *Arc*, click on the "X" in the upper right corner of the screen. In *Word* a screen will appear and ask whether you want to save changes made in your document. Click on *No*. In *Arc,* click on *OK*.

Warning: The following problem uses data from the book's webpage. Save the data files on a disk. Next, get in *Arc* and use the menu commands "File > Load" and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:)*. Then click twice on the data set name, eg, bodfat.lsp. These menu commands will be denoted by "File > Load > 3 1/2 Floppy(A:) > bodfat.lsp" where the data file (bodfat.lsp) will depend on the problem.

If the free statistics package *Arc* is on your personal computer (PC), there will be a folder *Arc* with a subfolder *Data* that contains a subfolder *Arcg*. Your instructor may have added a new folder *mdata* in the subfolder *Data* and added *bodfat.lsp* to the folder *mdata*. In this case the *Arc* menu commands "File > Load > Data > mdata > bodfat.lsp" can be used.

1.5*. This text's webpage has several files that can be used by *Arc*. Chapter 14 explains how to create such files.

a) Use the *Arc* menu commands "File > Load > 3 1/2 Floppy(A:) > bodfat.lsp" to activate the file *bodfat.lsp.*

b) Next use the menu commands "Graph&Fit>Fit linear LS" to obtain a window. Double click on *x1* and click once on y. Move the pointer to the *Response* box and click. Then *x1* should be in the *Terms/Predictors* box and y should be in the *Response* box. Click on *OK*. This performs simple linear regression of y on $x1$ and output should appear in the *Listener* box.

c) Next make a residual plot with the menu commands "Graph&Fit>Plot of." A window will appear. Double click on *L1: Fit–Values* and then double click on *L1:Residuals*. Then *L1: Fit–Values* should appear in the *H* box and *L1:Residuals* should appear in the *V* box. Click on *OK* to obtain the plot. There should be a curve in the center of the plot with five points separated from the curve. To delete these five points from the data set, move the pointer to one of the five points and hold the leftmost mouse button down. Move the mouse down and to the right. This will create a box, and after releasing the mouse button, any point that was in the box will be highlighted. To delete the highlighted points, click on the *Case deletions* menu, and move the pointer to *Delete selection from data set.* Repeat this procedure until the five outliers are deleted. Then use the menu commands "Graph&Fit>Fit linear LS" to obtain a window and click on *OK*. This performs simple linear regression of y on $x1$ without the five deleted cases. (*Arc* displays the case numbers of the cases deleted, but the labels are off by one since *Arc* gives the first case the case number zero.) Again make a residual plot and delete any outliers. Use *L2: Fit–Values* and *L2:Residuals* in the plot. The point in the upper right of the plot is not an outlier since it follows the curve.

d) Use the menu commands "Graph&Fit>Fit linear LS" to obtain a window and click on *OK*. This performs simple linear regression of y on $x1$ without the seven to nine deleted cases. Make a residual plot (with L3 fitted values and residuals) and include the plot in *Word.* The plot should be curved and hence the simple linear regression model is not appropriate.

e) Use the menu commands "Graph&Fit>Plot of" and place *L3:Fit-Values* in the *H* box and y in the *V* box. This makes a response plot. Include the plot in *Word*. If the response plot is not linear, then the simple linear regression model is not appropriate.

f) Comment on why both the residual plot and response plot are needed to show that the simple linear regression model is not appropriate.

g) Use the menu commands "Graph&Fit>Fit linear LS" to obtain a win-

dow, and click on the *Full quad.* circle. Then click on *OK*. These commands will fit the quadratic model $y = x1 + x1^2 + e$ without using the deleted cases. Make a residual plot of L4:Fit-Values versus L4:Residuals and a response plot of L4:Fit-Values versus y. For both plots place the fitted values in the H box and the other variable in the V box. Include these two plots in *Word*.

h) If the response plot is linear and if the residual plot is rectangular about the horizontal axis, then the quadratic model may be appropriate. Comment on the two plots.