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 *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. After a reasonable model is found, the model can be used for description or inference.

Response variables are the variables of interest, and are predicted with a $p \times 1$ vector of predictor variables. For regression models, we will often use Y or Z for the response variable and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for predictor variables where $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$. For example, predict Y = *systolic blood pressure* using a constant x_1, x_2 = *age*, x_3 = *weight*, and x_4 = *dosage amount of blood pressure medicine*. The multivariate location and dispersion (MLD) model has no predictor variables, and we will often use $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the p response variables. For regression, the ith case is $(Y_i, x_{i1}, ..., x_{ip})^T = (Y_i, \boldsymbol{x}_i^T)^T$ for $i = 1, ..., n$ where n is the sample size. For MLD, the ith case is $\boldsymbol{x}_i$. To get outlier resistant methods for regression models and MLD models, we will often use a robust MLD estimator on the $\boldsymbol{x}_i$. See Chapter 3.

Definition 1.1. A **case** or **observation** consists of k random variables measured for one person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

Definition 1.2. *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* $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|SP \text{ or } Y \perp\!\!\!\perp \boldsymbol{x}|h(\boldsymbol{x}), \tag{1.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The *estimated sufficient predictor* $\text{ESP} = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ where $\text{ESP} = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. This class of models includes the *generalized linear model* (GLM). Another important special case is a *generalized additive model* (GAM), where Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ given the *additive predictor* $AP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated additive predictor* $\text{EAP} = \text{ESP} = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$.

Notation: In this text, a plot of x versus Y will have x on the horizontal axis, and Y on the vertical axis.

Plots are extremely important for regression. When $p = 1$, x is both a sufficient predictor and an estimated sufficient predictor. So a plot of x versus Y is both a sufficient summary plot and a response plot. Usually the SP is unknown, so only the response plot can be made. The response plot will be extremely useful for checking the goodness of fit of the 1D regression model.

Definition 1.3. A *sufficient summary plot* is a plot of the SP versus Y. An *estimated sufficient summary plot* (ESSP) or **response plot** is a plot of the ESP versus Y.

Notation. Often the index i will be suppressed. If $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, we could redefine $\boldsymbol{x}$ and $\boldsymbol{\beta}$ (or omit α) so that $h(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$. For example, the *multiple linear regression model*

$$Y_i = \boldsymbol{\beta}^T\boldsymbol{x}_i + e_i \tag{1.2}$$

for $i = 1, ..., n$ where $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and e_i is a random error. This model could be written $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$. More accurately, $Y|\boldsymbol{x} = \boldsymbol{\beta}^T\boldsymbol{x} + e$, but the conditioning on $\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed) with *mean* 0 and unknown *standard deviation* σ. For this model, estimation of $\boldsymbol{\beta}$ and σ is important for inference and for predicting a new value of the response variable Y_f given a new vector of predictors $\boldsymbol{x}_f$.

The class of 1D regression models is very rich, and many of the most used statistical models, including GLMs and GAMs, are 1D regression models. Nonlinear regression, nonparametric regression, and linear regression are special cases of the *additive error regression* model

$$Y = h(\boldsymbol{x}) + e = SP + e. \tag{1.3}$$

The *multiple linear regression model* and *experimental design model* or *ANOVA model* are special cases of the linear regression model $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$. Another important class of parametric or semiparametric 1D regression models has the form

$$Y = g(\alpha + \boldsymbol{x}^T\boldsymbol{\beta}, e) \text{ or } \mathrm{Y = g(\boldsymbol{x}^T\boldsymbol{\beta}, e)}. \tag{1.4}$$

Special cases include GLMs and the *response transformation model*

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

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

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

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* $(Y_i, \boldsymbol{x}_i^T)^T$ 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 additive error regression model uses

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

where $\hat{m}(\boldsymbol{x})$ is an estimate of $m(\boldsymbol{x})$. If the sufficient predictor is $\boldsymbol{x}^T\boldsymbol{\beta}$, then 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} :

$$Y = t^{-1}(\boldsymbol{x}^T\boldsymbol{\beta} + e). \tag{1.7}$$

Then the transformation

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

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 *outliers,* observations far from the bulk of the data, are present.

Another class of robust statistics has good large sample theory for a large class of distributions: e.g. $\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{\beta}$ for a large class of error distributions. Examples include OLS and L_1 for multiple linear regression, the sample mean and sample covariance matrix for the multivariate location and dispersion model, least squares and the Yule Walker estimators for AR(p) time series, and least squares for the multivariate linear regression model where there are m response variables.

These two classes of robust statistics 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.

Acronyms are widely used in robust statistics and multivariate analysis, and some of the more important acronyms are in Table 1.1. Also see the text's index. The letter "R" tends to stand for "robust" (RPCA) or "reweighted" (RFCH). The letter "F" before a brand name robust estimator (FMCD) tends to mean a practical estimator that used a fixed number of trial fits, where the criterion of the brand name estimator was used to select the trial fit used in the final estimator. The letter "C" before a brand name estimator (CLTS) tends to mean a concentration algorithm was used for the F–brand name estimator. The letter "A", standing for "algorithm", was also used for concentration algorithms (ALTS). These acronyms (with A, C, F, or R) are often omitted from Table 1.1.

1.1 Outlier....s

An *outlier* is an observation that is far from the bulk 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 (e.g. 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

Table 1.1 Acronyms

Acronym	Description
cdf	cumulative distribution function
cf	characteristic function
CI	confidence interval
CLT	central limit theorem
Det-MCD	practical approximate MCD estimator not backed by theory
DGK	an MLD estimator (DGK are the initials of the paper's authors)
EC	elliptically contoured
ESP	estimated sufficient predictor
Fast-MCD	a slow FMCD estimator
FCH	name of a fast, consistent, highly outlier resistant MLD estimator
FLTS	practical approximate LTS estimator not backed by theory
FMCD	practical approximate MCD estimator not backed by theory
GAM	generalized additive model
GLM	generalized linear model
HB	high breakdown
hbreg	practical high breakdown regression estimator backed by theory
iid	independent and identically distributed
LMS	least median of squares (robust regression)
LR	logistic regression
LTA	least trimmed sum of absolute deviations (robust regression)
LTS	least trimmed sum of squares (robust regression)
MAD	median absolute deviation
MANOVA	multivariate analysis of variance
MB	median ball estimator
MBA	an MLD estimator made obsolete by FCH
MBA	or the median ball algorithm is the mbareg estimator
mbareg	a resistant regression estimator backed by theory
MCD	the impractical minimum covariance determinant estimator
MCLT	multivariate central limit theorem
MED	the median
mgf	moment generating function
MLD	multivariate location and dispersion
MLR	multiple linear regression
MVE	the impractical minimum volume ellipsoid estimator
MVN	multivariate normal
OGK	an MLD estimator not backed by theory
OLS	ordinary least squares
pdf	probability density function
PI	prediction interval
pmf	probability mass function
RFCH	the reweighted FCH estimator
RMVN	a reweighted FCH estimator that works well for MVN data
SE	standard error
SSP	sufficient summary plot
TVREG	a resistant "trimmed views" regression estimator

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 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 and Ronchetti (2009, p. 4) states that outlier resistance and distributional robustness are synonymous while Hampel et al. (1986, p. 36) state that the first and most important step in robustification is the rejection of distant outliers.

Deciding what to do with outliers can be difficult. Sometimes the outliers should be discarded or downweighted. Then inflexible estimators such as resistant multiple linear regression estimators are often useful. The estimator is inflexible since a hyperplane is estimated. Sometimes the oultiers are important and should be fit will by the model. Then flexible estimators, such as the generalized additive model to fit the additive error regression model, are often useful.

Example 1.1. a) The Rousseeuw and Leroy (1987, p. 26) Belgian telephone data has response $Y =$ *number of international phone calls* (in tens of millions) made per year in Belgium. The predictor variable $x =$ year (1950-1973). From 1964 to 1969 total number of minutes of calls was recorded instead, and years 1963 and 1970 were also partially effected. Hence there are 6 large outliers and 2 additional cases that have been corrupted. The 8 cases corresponding to these outliers should be deleted.

b) Wood (2017, pp. 346-348) describes an air pollution data set where the response variable is the daily death rate in Chicago over a number of years. For this data set, there tend to be outliers that occur a few days after days that had both high temperature and high ozone levels. For this data set, the outliers are very important, and should be fit well by the model.

c) While consulting for a chemistry experiment, the data set was fit by a regression method where the expert said some of the Y_i were impossible due to large e_i. The nonparametric bootstrap using all of the data gave results that the expert considered reasonable for inference.

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 nonoutlying (or "clean") cases. The outliers should *never* be blindly discarded. Often the clean data and the outliers are analyzed separately. The clean cases are also called *inliers.*

The *asymptotic paradigm* uses an asymptotic distribution to approximate the distribution of the estimator when the sample size n is large. An impor-

tant example is the *central limit theorem* (CLT): let $Y_1, ..., Y_n$ be iid with mean μ and standard deviation σ; i.e., 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 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 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 $n > 30$, but for any $n > 0$, there are many distributions where the CLT approximation is poor. 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*, and this book provides robust estimators and diagnostics that can be used in tandem with classical 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 1.2. Gladstone (1905) attempts to estimate the *weight* of the human brain using 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 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, ALMS, ALTS and MBA fits. ALMS is the default version of the *R* function `lmsreg` while ALTS is the default version of `ltsreg`. The three estimators ALMS, ALTS, and MBA are described further in Chapters 6, 7, and 8. Figure 1.1 was made with a 2007 version of *R*. ALMS, ALTS and MBA depend on the seed (in *R*) and so the estimators change with each call of `rrplot2`. Also, the ALMS and ALTS estimators change frequently. 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 (with

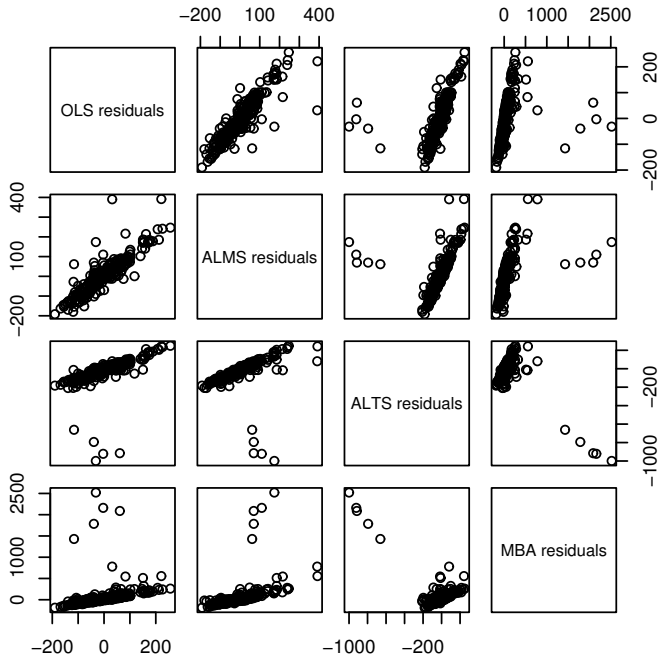

Fig. 1.1 RR Plot for Gladstone data

the nine potential outliers given weight 0) passes through the five infants, suggesting that these cases are "good leverage points."

Assume the book's collection of *R* functions *rpack* and collection of data sets *robdata* are stored on flash drive G. See Section 11.2. RR plots similar to Figure 1.1 can be made in *R* using the following commands.

```
source("G:/rpack.txt")
source("G:/robdata.txt")
library(MASS)
rrplot2(cbrainx,cbrainy)
```

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 right corner.

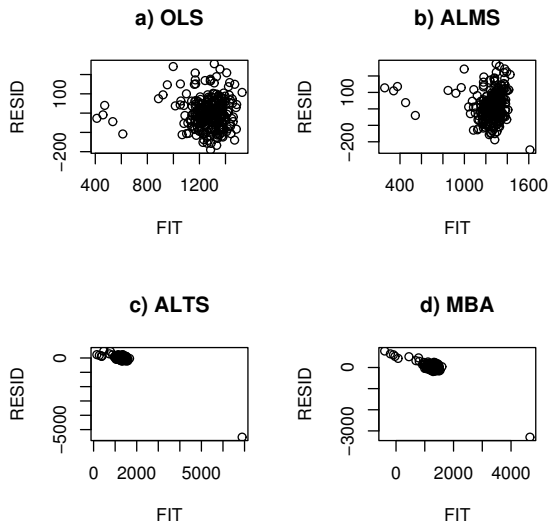

Fig. 1.2 Gladstone data where case 119 is a typo

Example 1.3. 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 around 2000, shows that the outliers were accommodated by OLS, ALMS and ALTS. The outliers had large absolute residuals for the MBA, BB and MBALATA estimators. Figure 5.2 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 3.4 and to illustrate a graphical diagnostic for multivariate normality in Example 3.2.

Example 1.4. 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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1.9}$$

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 $(Y_i, \boldsymbol{x}_i^T)^T$ 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 Olive, (2017a, p. 39), a $100(1-\delta)\%$ prediction interval (PI) for a new observation Y_f corresponding

to a vector of predictors $\boldsymbol{x}_f$ is given by

$$\hat{Y}_f \pm t_{n-p,1-\alpha/2} se(pred) \tag{1.10}$$

where $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}$, $P(t \leq t_{n-p,1-\delta/2}) = 1 - \delta/2$ 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 $100(1 - \delta)\%$ PI for future cases from this population by making a classical $100(1-\delta^*)$ PI from the clean cases where

$$1 - \delta^* = (1 - \delta)/(1 - \gamma). \tag{1.11}$$

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

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

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

The formula for this PI is then

$$\hat{Y}_f \pm t_{n_c-p,1-\delta^*/2} se(pred) \tag{1.12}$$

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

Example 1.5. STATLIB provides the Johnson (1996) data set that is available from the website (http://lib.stat.cmu.edu/datasets/bodyfat) and from the text website file *bodfat.lsp*. 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

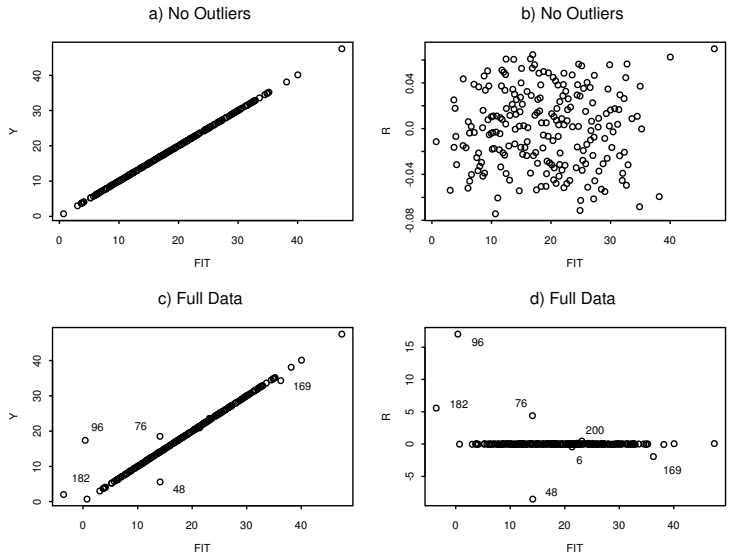

Fig. 1.3 Plots for Summarizing the Entire Population

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, make a classical $100(1-\delta^*)$ PI from the clean cases where $1-\delta^* = 0.9/(1-\gamma)$. For the bodyfat data, we can take $1-\gamma \approx 1-9/252 \approx 0.964$ and $1-\delta^* \approx 0.94$. 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 1.3a 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 $\hat{Y}_f \pm t_{249,0.95} se(pred) = 46.3152 \pm 1.651(1.3295) = [44.12, 48.51]$. When the 9 outliers are deleted, $n_c = 243$ cases remain. Hence the 90% PI using Equa-

tion (1.12) 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*

$$t_\lambda(Y) \equiv Y^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda} \tag{1.13}$$

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 1.6. 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 response plots for two MLR estimators: OLS and the R 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.

Further illustrations of the graphical method for selecting the response transformation t_λ are in Section 4.2.

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 4.3 gives a graphical method for assessing variable

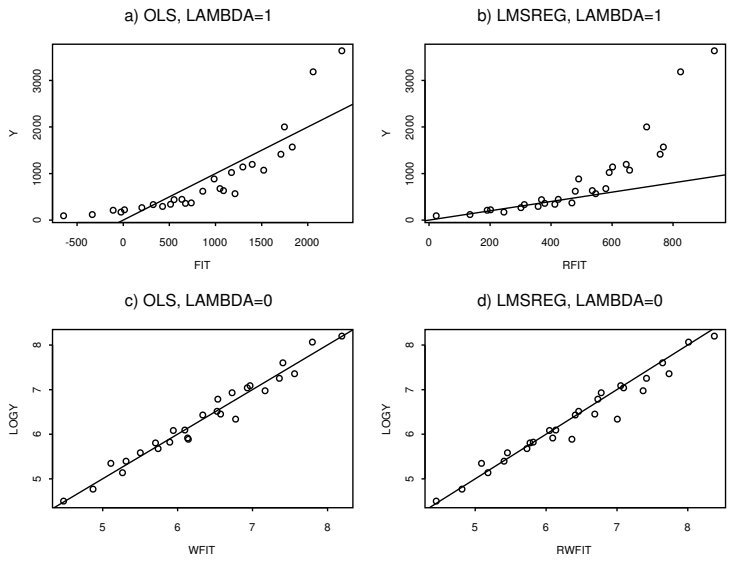

Fig. 1.4 OLS and LMSREG Suggest Using log(Y) for the Textile Data

selection for multiple linear regression models while Section 9.4 gives a similar method for a large class of 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.

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 9.4 will show that the all subsets, forward selection and backward elimination techniques of variable selection for multiple linear regression will often work for a large class of 1D regression models provided that the Mallows' C_p criterion is used.

Example 1.7. 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 9.4.

Another important topic is fitting 1D regression models given by Equation (1.4) 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. The R 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 an *additive error single index model* $Y = m(\boldsymbol{x}^T\boldsymbol{\beta}) + e$ where m is the cubic function.

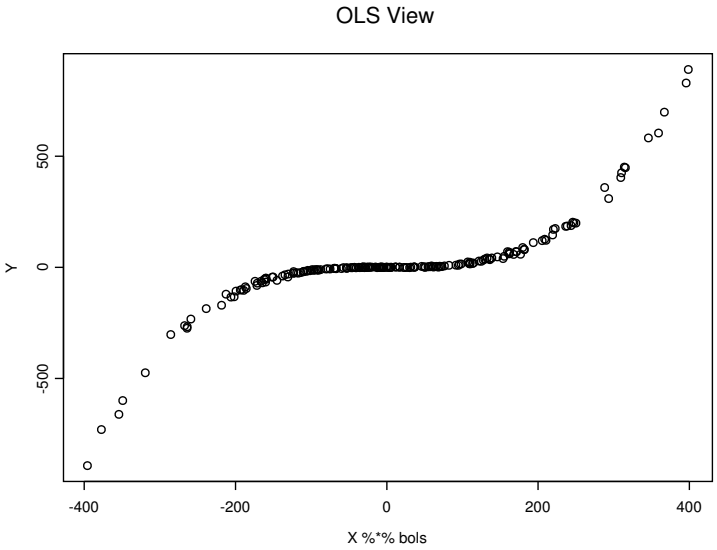

Fig. 1.5 Response Plot or OLS View for $m(u) = u^3$

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 Chapter 9.

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

An excellent paper on statistical models is Box (1979). Several authors consider the model $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$ or $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}_1, ..., \boldsymbol{x}^T\boldsymbol{\beta}_d$ where the structural dimension is d. See Cook and Weisberg (1999a) and Cook (1998a). The 1D regression model, due to Olive (2004b), uses $Y \perp\!\!\!\perp \boldsymbol{x} | h(\boldsymbol{x})$. A dD regression model would use $Y \perp\!\!\!\perp \boldsymbol{x} | h_1(\boldsymbol{x}), ..., h_d(\boldsymbol{x})$. Using $h(\boldsymbol{x})$ is similar to using a minimal sufficient statistic while using $\boldsymbol{x}^T\boldsymbol{\beta}_1, ..., \boldsymbol{x}^T\boldsymbol{\beta}_d$ is similar to using a sufficient statistic, e.g. a 1D regression model could have structural dimension $d > 1$ (this result occurs for the additive error regression model $Y = m(\boldsymbol{x}) + e$ if $m(\boldsymbol{x})$ is a function of $\boldsymbol{x}^T\boldsymbol{\beta}_1, ..., \boldsymbol{x}^T\boldsymbol{\beta}_d$). For more on 1D regression, see Olive (2010, 2017a, 2017b: pp. 427-443, 2020). The graphical method for response transformations illustrated in Example 1.6 was suggested by Olive (2004b).

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." 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 et al. (1986, Section 1.4) surveys some robust outlier rejection rules. Outlier rejection rules are also discussed in Hampel (1985), Simonoff (1987ab), and Stigler (1973b). Aggarwal (2017) covers outliers from a Machine Learning perspective. Olive (2017b) gives many outlier resistant methods.

This text will use the *R* software R Core Team (2016), available from the website (www.r-project.org/). Section 11.2 of this text, Becker, Chambers, and Wilks (1988), Crawley (2013), and Venables and Ripley (2010) are useful for *R* users.

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 in the second paragraph of Section 1.2, 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 Problems Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

1.2*. a) Paste the commands for this problem (from the above link) into *R* to reproduce a plot like Figure 1.5.

b) Activate *Word* (often by double clicking on a *Word* icon, perhaps after typing *word* in the box on the lower left of the computer screen). Click on the screen and type "Problem 1.2." To copy and paste a plot from *R* into *Word*, click on the plot and hit *Ctrl* and *c* at the same time. Then go to *file* in the *Word* menu and select *paste* or hit *Ctrl* and *v* at the same time.

To save your output on your flash drive G, click on the icon in the upper left corner of *Word*. Then drag the pointer to "Save as." A window will appear, click on the *Word Document* icon. A "Save as" screen appears. Click on the right "check" on the top bar, and then click on "Removable Disk (G:)". Change the file name to HW1d2.docx, and then click on "Save."

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*, 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, paste the commands for this problem into *R*.

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) Paste the commands for this problem into *R* 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.

b) Paste the commands for this problem into *R* to plot 4 histograms with $n = 1, 5, 25$ and 200. Save the plot in *Word* and then print the plot using the procedure described in Problem 1.2b.

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 $\overline{Y} \sim N(0, 1/n)$.