1

Introduction

1.1 Classical and robust approaches to statistics

This introductory chapter is an informal overview of the main issues to be treated in detail in the rest of the book. Its main aim is to present a collection of examples that illustrate the following facts:

- Data collected in a broad range of applications frequently contain one or more atypical observations, known as *outliers*; that is, observations that are well-separated from the majority or "bulk" of the data, or in some way deviate from the general pattern of the data.
- Classical estimates, such as the sample mean, the sample variance, sample covariances and correlations, or the least-squares fit of a regression model, can be adversely influenced by outliers, even by a single one, and therefore often fail to provide good fits to the bulk of the data.
- There exist *robust* parameter estimates that provide a good fit to the bulk of the data when the data contains outliers, as well as when the data is free of them. A direct benefit of a good fit to the bulk of data is the reliable detection of outliers, particularly in the case of multivariate data.

In Chapter 3 we shall provide some formal probability-based concepts and definitions of robust statistics. Meanwhile, it is important to be aware of the following performance distinctions between classical and robust statistics at the outset. Classical statistical inference quantities such as confidence intervals, t-statistics and p-values, R^2 values and model selection criteria in regression can be adversely influenced by the presence of even one outlier in the data. In contrast, appropriately constructed

Robust Statistics: Theory and Methods (with R), Second Edition.
Ricardo A. Maronna, R. Douglas Martin, Victor J. Yohai and Matías Salibián-Barrera.

Companion website: www.wiley.com/go/maronna/robust

robust versions of those inference quantities are little influenced by outliers. Point estimate predictions and their confidence intervals based on classical statistics can be spoiled by outliers, while predictive models fitted using robust statistics do not suffer from this disadvantage.

It would, however, be misleading to always think of outliers as "bad" data. They may well contain unexpected, but relevant information. According to Kandel (1991, p. 110):

> The discovery of the ozone hole was announced in 1985 by a British team working on the ground with "conventional" instruments and examining its observations in detail. Only later, after reexamining the data transmitted by the TOMS instrument on NASA's Nimbus 7 satellite, was it found that the hole had been forming for several years. Why had nobody noticed it? The reason was simple: the systems processing the TOMS data, designed in accordance with predictions derived from models, which in turn were established on the basis of what was thought to be "reasonable", had rejected the very ("excessively") low values observed above the Antarctic during the Southern spring. As far as the program was concerned, there must have been an operating defect in the instrument.

In the next sections we present examples of classical and robust estimates of the mean, standard deviation, correlation and linear regression for data containing outliers. Except in Section 1.2, we do not describe the robust estimates in any detail, and return to their definitions in later chapters.

1.2 Mean and standard deviation

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be a set of observed values. The sample mean $\overline{x}$ and sample standard deviation (SD) s are defined by

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{1.1}$$

The sample mean is just the arithmetic average of the data, and as such one might expect that would provide a good estimate of the *center* or *location* of the data. Likewise, one might expect that the sample SD would provide a good estimate of the *dispersion* of the data. Now we shall see how much influence a single outlier can have on these classical estimates.

Example 1.1 *Consider the following 24 determinations of the copper content in wholemeal flour (in parts per million), sorted in ascending order (Analytical Methods Committee, 1989):*

2.20	2.20	2.40	2.40	2.50	2.70	2.80	2.90
3.03	3.03	3.10	3.37	3.40	3.40	3.40	3.50
3.60	3.70	3.70	3.70	3.70	3.77	5.28	28.95

The value 28.95 immediately stands out from the rest of the values and would be considered an outlier by almost anyone. One might conjecture that this

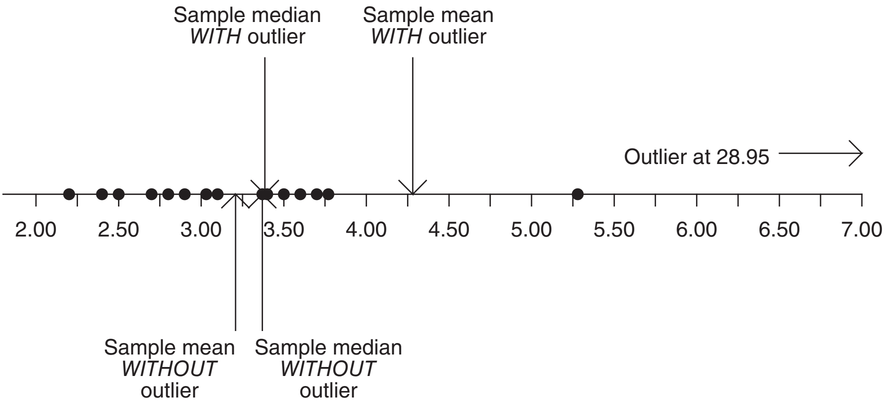

Figure 1.1 Copper content of flour data with sample mean and sample median estimates

inordinately large value was caused by a misplaced decimal point with respect to a “true” value of 2.895. In any event, it is a highly influential outlier, as we now demonstrate.

The values of the sample mean and SD for the above dataset are $\overline{x} = 4.28$ and $s = 5.30$, respectively. Since $\overline{x} = 4.28$ is larger than all but two of the data values, it is not among the bulk of the observations and as such does not represent a good estimate of the center of the data. If one deletes the suspicious value of 28.95, then the values of the sample mean and sample SD are changed to $\overline{x} = 3.21$ and $s = 0.69$. Now the sample mean does provide a good estimate of the center of the data, as is clearly shown in Figure 1.1, and the SD is over seven times smaller than it was with the outlier present. See the leftmost upward pointing arrow and the rightmost downward-pointing arrow in Figure 1.1.

Let us consider how much influence a single outlier can have on the sample mean and sample SD. For example, suppose that the value 28.95 is replaced by an arbitrary value x for the 24th observation, x_{24}. It is clear from the definition of the sample mean that by varying x from $-\infty$ to $+\infty$ the value of the sample mean changes from $-\infty$ to $+\infty$. It is an easy exercise to verify that as x ranges from $-\infty$ to $+\infty$, the sample SD ranges from some positive value smaller than that based on the first 23 observations to $+\infty$. Thus we can say that a single outlier has an *unbounded influence* on these two classical statistics.

An outlier may have a serious adverse influence on confidence intervals. For the flour data, the classical interval based on the t-distribution with confidence level 0.95 is (2.05, 6.51); after removing the outlier, the interval is (2.91, 3.51). The impact of the single outlier has been to considerably lengthen the interval in an asymmetric way.

This example suggests that a simple way to handle outliers is to detect them and remove them from the dataset. There are many methods for detecting outliers (see, for example, Barnett and Lewis, 1998). Deleting an outlier, although better than doing nothing, still poses a number of problems:

- When is deletion justified? Deletion requires a subjective decision. When is an observation "outlying enough" to be deleted?
- The user or the author of the data may think that "an observation is an observation" (i.e., observations should speak for themselves) and hence feel uneasy about deleting them
- Since there is generally some uncertainty as to whether an observation is really atypical, there is a risk of deleting "good" observations, which would result in underestimating data variability
- Since the results depend on the user's subjective decisions, it is difficult to determine the statistical behavior of the complete procedure.

We are thus lead to another approach: why use the sample mean and SD? Maybe are there other better possibilities?

One very old method for estimating the "middle" of the data is to use the sample *median*. Any number t with a value such that the numbers of observations on both sides of it are equal is called a *median* of the dataset: t is a median of the data set $\mathbf{x} = (x_1, \ldots, x_n)$, and will be denoted by

$$t = \mathrm{Med}(\mathbf{x}),\ \mathit{if}\ \#\{x_i > t\} = \#\{x_i < t\},$$

where $\#\{A\}$ denotes the number of elements of the set A. It is convenient to define the sample median in terms of the *order statistics* $(x_{(1)}, x_{(2)}, ..., x_{(n)})$, obtained by sorting the observations $\mathbf{x} = (x_1, ..., x_n)$ in increasing order so that

$$x_{(1)} \leq ... \leq x_{(n)}. \tag{1.2}$$

If n is odd, then $n = 2m - 1$ for some integer m, and in that case $\mathrm{Med}(\mathbf{x}) = x_{(m)}$. If n is even, then $n = 2m$ for some integer m, and then any value between $x_{(m)}$ and $x_{(m+1)}$ satisfies the definition of a sample median, and it is customary to take

$$\mathrm{Med}(\mathbf{x}) = \frac{x_{(m)} + x_{(m+1)}}{2}.$$

However, in some cases (e.g. in Section 4.5.1) it may be more convenient to choose $x_{(m)}$ or $x_{(m+1)}$ ("low" and "high" medians, respectively).

The mean and the median are approximately equal if the sample is symmetrically distributed about its center, but not necessarily otherwise. In our example, the median of the whole sample is 3.38, while the median without the largest value is 3.37, showing that the median is not much affected by the presence of this value. See the locations of the sample median with and without the outlier present in Figure 1.1 above. Notice that for this sample, the value of the sample median with the outlier present is relatively close to the sample mean value of 3.21 with the outlier deleted.

Suppose again that the value 28.95 is replaced by an arbitrary value x for the 24th observation $x_{(24)}$. It is clear from the definition of the sample median that when x ranges from $-\infty$ to $+\infty$ the value of the sample median does not change from $-\infty$ to $+\infty$ as was the case for the sample mean. Instead, when x goes to $-\infty$ the sample median undergoes the small change from 3.38 to 3.23 (the latter being the average of $x_{(11)} = 3.10$ and $x_{(12)} = 3.37$ in the original dataset); when x goes to $+\infty$ the sample median goes to the value 3.38 given above for the original data. Since the sample median fits the bulk of the data well, with or without the outlier, and is not much influenced it, it is a good robust alternative to the sample mean.

Likewise, one robust alternative to the SD is the *median absolute deviation about the median* (MAD), defined as

$$\mathrm{MAD}(\mathbf{x}) = \mathrm{MAD}(x_1, x_2, ..., x_n) = \mathrm{Med}\{|\mathbf{x} - \mathrm{Med}(\mathbf{x})|\}.$$

This estimator uses the sample median twice, first to get an estimate of the center of the data in order to form the set of absolute residuals about the sample median, $\{|\mathbf{x} - \mathrm{Med}(\mathbf{x})|\}$, and then to compute the sample median of these absolute residuals. To make the MAD comparable to the SD, we define the *normalized MAD* (MADN) as

$$\mathrm{MADN}(\mathbf{x}) = \frac{\mathrm{MAD}(\mathbf{x})}{0.6745}.$$

The reason for this definition is that 0.6745 is the MAD of a standard normal random variable, and hence a $\mathrm{N}(\mu, \sigma^2)$ variable has $\mathrm{MADN} = \sigma$.

For the above dataset, one gets $\mathrm{MADN} = 0.53$, as compared with $s = 5.30$. Deleting the large outlier yields $\mathrm{MADN} = 0.50$, as compared to the somewhat higher sample SD value of $s = 0.69$. The MAD is clearly not influenced very much by the presence of a large outlier, and as such provides a good robust alternative to the sample SD.

So why not always use the median and MAD? An informal explanation is that if the data contain no outliers, these estimates have a statistical performance that is poorer than that of the classical estimates $\overline{x}$ and s. The ideal solution would be to have "the best of both worlds": estimates that behave like the classical ones when the data contain no outliers, but are insensitive to outliers otherwise. This is the data-oriented idea of robust estimation. A more formal notion of robust estimation based on statistical models, which will be discussed in the following chapters, is that the statistician always has a statistical model in mind (explicitly or implicitly) when analyzing data, for example a model based on a normal distribution or some other idealized parametric model such as an exponential distribution. The classical estimates are in some sense "optimal" when the data are exactly distributed according to the assumed model, but can be very suboptimal when the distribution of the data differs from the assumed model by a "small" amount. Robust estimates on the other hand maintain approximately optimal performance, not just under the assumed model, but under "small" perturbations of it too.

1.3 The “three sigma edit” rule

A traditional measure of the outlyingness of an observation x_i with respect to a sample, is the ratio between its distance to the sample mean and the sample SD:

$$t_i = \frac{x_i - \overline{x}}{s}, \tag{1.3}$$

Observations with $|t_i| > 3$ are traditionally deemed suspicious (the “three-sigma rule”), based on the fact that they would be “very unlikely” under normality, since $\mathrm{P}(|x| \geq 3) = 0.003$ for a random variable x with a standard normal distribution. The largest observation in the flour data has $t_i = 4.65$, and so is suspicious. Traditional “three-sigma edit” rules result in either discarding observations for which $|t_i| > 3$, or adjusting them to one of the values $\overline{x} \pm 3s$, whichever is nearer. Despite its long tradition, this rule has some drawbacks that deserve to be taken into account:

- In a very large sample of “good” data, some observations will be declared suspicious and be altered. More precisely, in a large normal sample, about three observations out of 1000 will have $|t_i| > 3$. For this reason, normal Q–Q plots are more reliable for detecting outliers (see example below).
- In very small samples the rule is ineffective: it can be shown that

  $$|t_i| < \frac{n-1}{\sqrt{n}}$$

  for all possible data sample values, and hence if $n \leq 10$ then always $|t_i| < 3$. The proof is left to the reader (Problem 1.3).
- When there are several outliers, their effects may interact in such a way that some or all of them remain unnoticed (an effect called *masking*), as the following example shows.

Example 1.2 *The following data (Stigler 1977) are 20 determinations of the time (in microseconds) needed for the light to travel a distance of 7442 m. The actual times are the table values* × 0.001 + 24.8.

28	26	33	24	34	−44	27	16	40	−2
29	22	24	21	25	30	23	29	31	19

The normal Q–Q plot in Figure 1.2 reveals the two lowest observations (−44 and −2) as suspicious. Their respective t_is are −3.73 and −1.35, and so the value of $|t_i|$ for the observation −2 does not indicate that it is an outlier. The reason

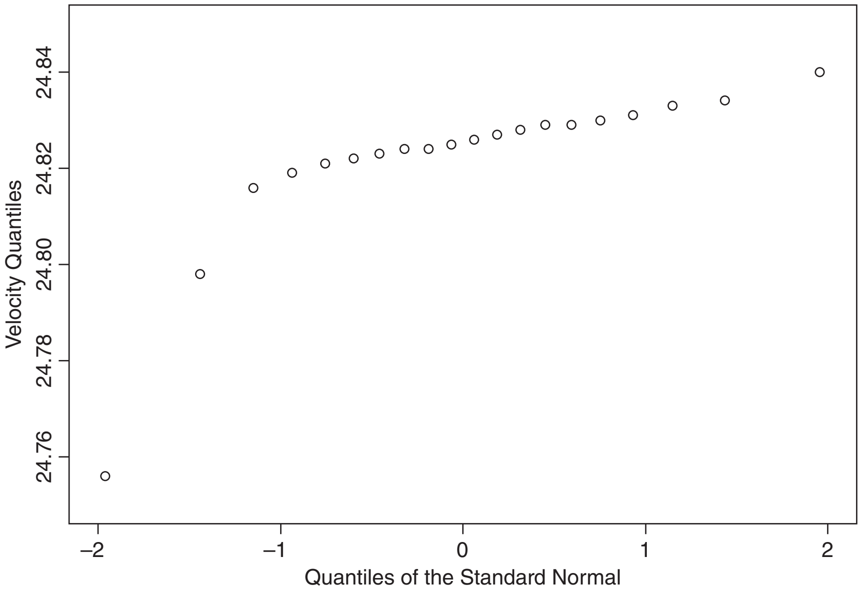

Figure 1.2 Velocity of light: Q–Q plot of observed times

that −2 has such a small $|t_i|$ value is that both observations pull $\overline{x}$ to the left and inflate s; it is said that the value −44 “masks” the value −2.

To avoid this drawback it is better to replace $\overline{x}$ and s in (1.3) by robust location and dispersion measures. A robust version of (1.3) can be defined by replacing the sample mean and SD by the median and MADN, respectively:

$$t'_i = \frac{x_i - \mathrm{Med}(\mathbf{x})}{\mathrm{MADN}(\mathbf{x})}. \tag{1.4}$$

The t_is for the two leftmost observations are now −11.73 and −4.64, and hence the three-sigma edit rule, with t' instead of t, pinpoints both as suspicious. This suggests that even if we only want to detect outliers – rather than to estimate parameters – detection procedures based on robust estimates are more reliable.

A simple robust location estimate could be defined by deleting all observations with $|t'_i|$ larger than a given value, and taking the average of the rest. While this procedure is better than the three-sigma edit rule based on t, it will be seen in Chapter 3 that the estimates proposed in this book handle the data more smoothly, and can be tuned to have certain desirable robustness properties that this procedure lacks.

1.4 Linear regression

1.4.1 Straight-line regression

First consider fitting a straight line regression model to the dataset $\{(x_i, y_i) : i = 1, ., n\}$

$$y_i = \alpha + x_i\beta + u_i, \quad i = 1, \ldots, n$$

where x_i and y_i are the predictor and response variable values, respectively, and u_i are random errors. The time-honored classical way of fitting this model is to estimate the parameters α and β with the least-squares (LS) estimates

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\widehat{\alpha} = \overline{y} - \overline{x}\widehat{\beta}$$

As an example of how influential two outliers can be on these estimates, Figure 1.3 plots the earnings per share (EPS) versus time each year for a company with the stock exchange ticker symbol IVENSYS, along with the straight-line fits of the LS estimate and of a robust regression estimate (called an MM-estimate) that has desirable theoretical properties (to be described in detail in Chapter 5).

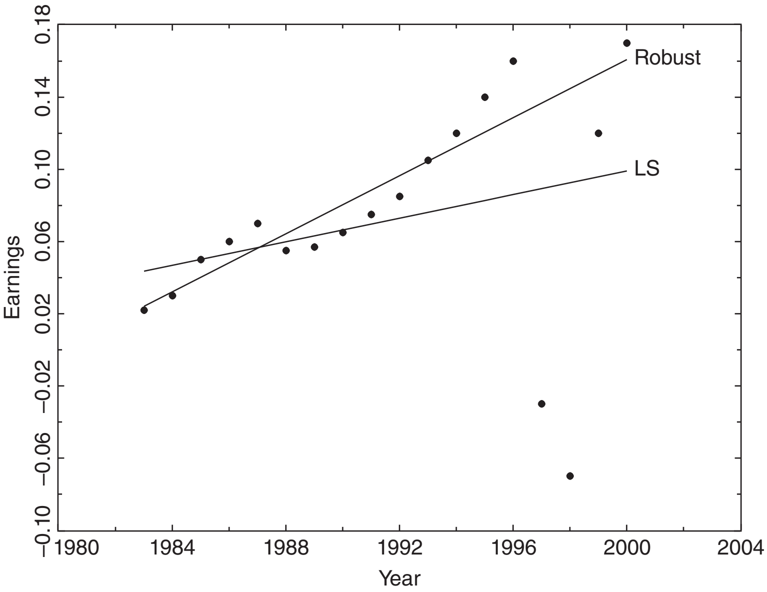

Figure 1.3 EPS data with robust and LS fits

The two unusually low EPS values in 1997 and 1998 cause the LS line to fit the data very poorly, and one would not expect the line to provide a good prediction of EPS in 2001. By way of contrast, the robust line fits the bulk of the data well, and should provide a reasonable prediction of EPS in 2001.

The above EPS example was brought to one of the author's attention by an analyst in the corporate finance department of a well-known large company. The analyst was required to produce a prediction of next year's EPS for several hundred companies, and at first he used the LS fit for this purpose. But then he noticed a number of firms for which the data contained outliers that distorted the LS parameter estimates, resulting in a very poor fit and a poor prediction of next year's EPS. Once he discovered the robust estimate, and found that it gave him essentially the same results as the LS estimate when the data contained no outliers, while at the same time providing a better fit and prediction than LS when outliers were present, he began routinely using the robust estimate for his task.

It is important to note that automatically flagging large differences between a classical estimate (in this case LS) and a robust estimate provides a useful diagnostic alert that outliers may be influencing the LS result.

1.4.2 Multiple linear regression

Now consider fitting a multiple linear regression model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + u_i, \quad i = 1, \ldots, n$$

where the response variable values are y_i, and there are p predictor variables x_{ij}, $j = 1, \ldots, p$, and p regression coefficients β_j. Not surprisingly, outliers can also have an adverse influence on the LS estimate $\widehat{\boldsymbol{\beta}}$ for this general linear model, a fact which is illustrated by the following example that appears in Hubert and Rousseeuw (1997).

Example 1.3 *The response variable values y_i are the rates of unemployment in various geographical regions around Hannover, Germany, and the predictor variables x_{ij}, $j = 1, \ldots, p$ are as follows:*

- *PA: percentage engaged in production activities*
- *GPA: growth in PA*
- *HS: percentage engaged in higher services*
- *GHS: growth in HS*
- *Region: geographical region around Hannover (21 regions)*
- *Period: time period (three periods: 1979–82, 1983–88, 1989–92)*

Note that the categorical variables Region and Period require 20 and 2 parameters respectively, so that, including an intercept, the model has 27 parameters,

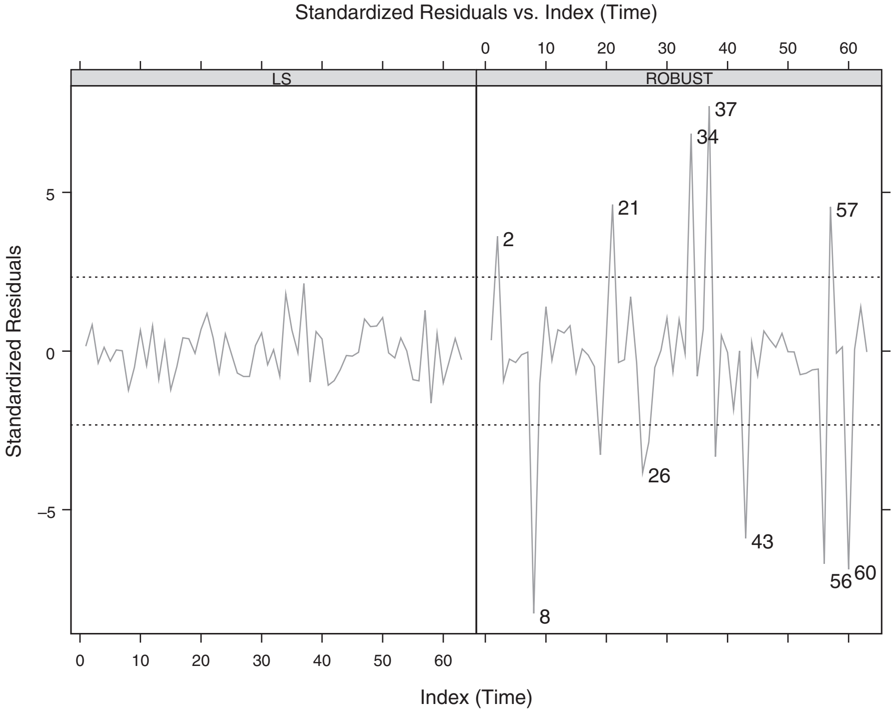

Figure 1.4 Standardized residuals for LS and robust fits

and the number of response observations is 63, one for each region and period. Figures 1.4 and 1.5 show the results of LS and robust fitting in a manner that facilitates easy comparison of the results. The robust fitting is done by a special "M-estimate" that has desirable theoretical properties, and is described in detail in Section 5.7.5.

For a set of estimated parameters $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$, with fitted values $\widehat{y}_i = \sum_{j=1}^{p} x_{ij}\widehat{\beta}_j$, residuals $\widehat{u}_i = y_i - \widehat{y}_i$ and residuals dispersion estimate $\widehat{\sigma}$, Figure 1.4 shows the standardized residuals $\tilde{u}_i = \widehat{u}_i/\tilde{\sigma}$ plotted versus the observations' index values i. Standardized residuals that fall outside the horizontal dashed lines at ±2.33, which occurs with probability 0.02, are declared suspicious. The display for the LS fit does not reveal any outliers, while that for the robust fit clearly reveals 10 to 12 outliers among 63 observations. This is because the robust regression has found a linear relationship that fits the majority of the data points well, and consequently is able to reliably identify the outliers. The LS estimate instead attempts to fit all data points and so is heavily influenced by the outliers. The fact that all of the LS standardized residuals lie inside the horizontal dashed lines is because the outliers have inflated

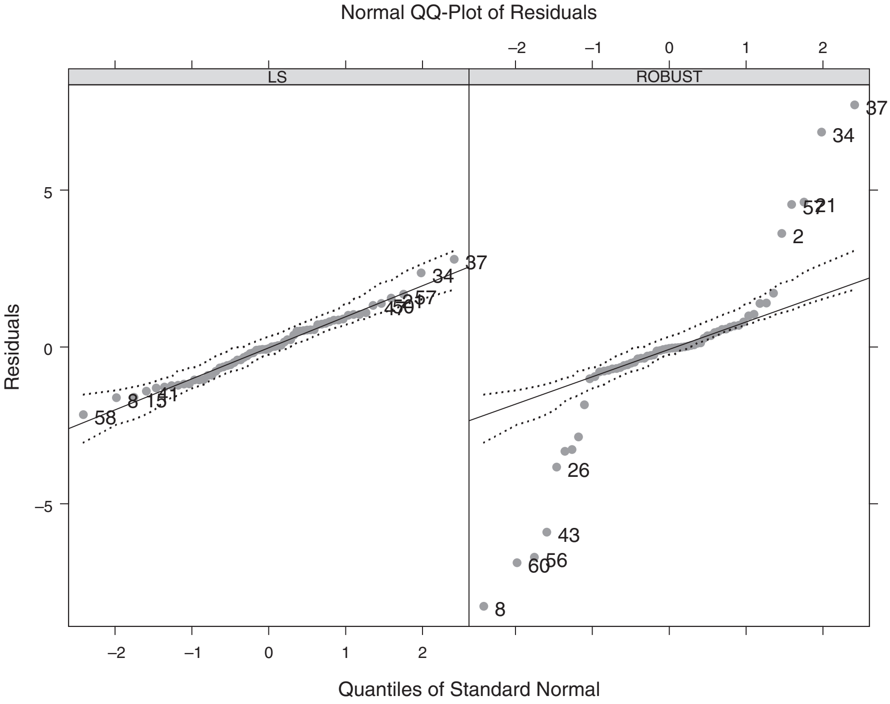

Figure 1.5 Normal Q–Q plots for (left) LS and (right) robust fits

the value of $\tilde{\sigma}$ computed in the classical way based on the sum of squared residuals, while a robust estimate $\tilde{\sigma}$ used for the robust regression is not much influenced by the outliers.

Figure 1.5 shows normal Q–Q plots of the residuals for the LS and robust fits, with light dotted lines showing the 95% simulated pointwise confidence regions to allow an assessment of whether or not there are significant outliers and potential nonnormality. These plots may be interpreted as follows. If the data fall along the straight line (which itself is fitted by a robust method) with no points outside the 95% confidence region, then one is moderately sure that the data are normally distributed.

Performing only the LS fit, and therefore looking only at the normal Q–Q plot in the left-hand plot in Figure 1.5, would lead to the conclusion that the residuals are indeed quite normally distributed, with no outliers. The normal Q–Q plot of residuals for the robust fit in the right-hand panel of Figure 1.5 clearly shows that such a conclusion is wrong. This plot shows that the bulk of the residuals are indeed quite normally distributed, as evidenced by the compact linear behavior in the middle of the plot. At the same time, it clearly reveals the outliers that were evident in the plot of standardized residuals (Figure 1.4).

1.5 Correlation coefficients

Let $\{(x_i, y_i)\}$, $i = 1, \ldots, n$, be a bivariate sample. The most popular measure of association between the x_i and the y_i is the sample correlation coefficient, defined as

$$\widehat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{1/2}\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{1/2}}$$

where $\overline{x}$ and $\overline{y}$ are the sample means of the x_i and y_i.

The sample correlation coefficient is highly sensitive to the presence of outliers. Figure 1.6 shows a scatterplot of the increase (gain) in numbers of telephones versus the annual change in new housing starts, for a period of 15 years in a geographical region within New York City in the 1960s and 1970s, in coded units.

There are two outliers in this bivariate (two-dimensional) dataset that are clearly separated from the rest of the data. It is important to notice that these two outliers are not one-dimensional outliers; they are not even the largest or smallest values in any of the two coordinates. This observation illustrates an extremely important point: two-dimensional outliers cannot be reliably detected by examining the values of bivariate data one-dimensionally; that is, one variable at a time.

The value of the sample correlation coefficient for the complete gain data is $\widehat{\rho} = 0.44$, and deleting the two outliers yields $\widehat{\rho} = 0.91$, which is quite a large difference and in the range of what an experienced user might expect for the dataset with the two outliers removed. The dataset with the two outliers deleted can be seen

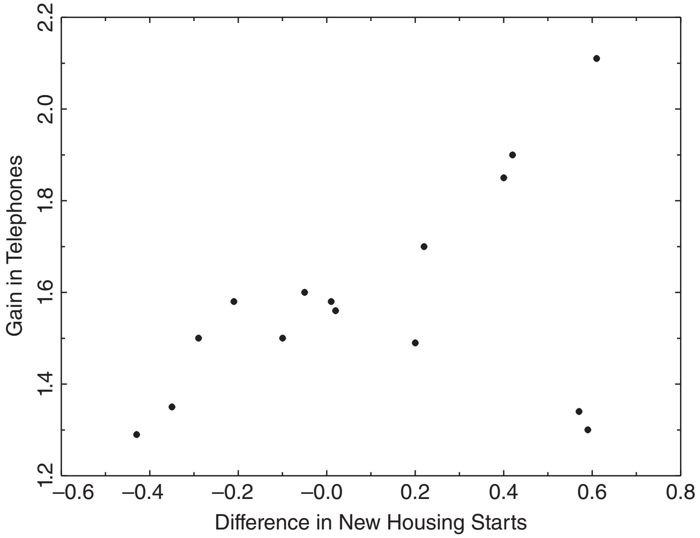

Figure 1.6 Increase in numbers of telephones versus difference in new housing starts

as roughly elliptical, with a major axis sloping up and to the right and the minor axis sloping up and to the left With this picture in mind one can see that the two outliers lie in the minor axis direction, though offset somewhat from the minor axis. The impact of the outliers is to decrease the value of the sample correlation coefficient by the considerable amount of 0.44 from the value of 0.91 it has with the two outliers deleted. This illustrates a general biasing effect of outliers on the sample correlation coefficient: outliers that lie along the minor axis direction of data that is otherwise positively correlated negatively influence the sample correlation coefficient. Similarly, outliers that lie along the minor axis direction of data that is otherwise negatively correlated will increase the sample correlation coefficient. Outliers that lie along a major axis direction of the rest of the data will increase the absolute value of the sample correlation coefficient, making it more positive if the bulk of the data is positively correlated.

If one uses a robust correlation coefficient estimate it will not make much difference whether the outliers in the main-gain data are present or deleted. Using a good robust method $\widehat{\rho}_{Rob}$ for estimating covariances and correlations on the main-gain data yields $\widehat{\rho}_{Rob} = 0.85$ for the entire dataset and $\widehat{\rho}_{Rob} = 0.90$ with the two outliers deleted. For the robust correlation coefficient, the change due to deleting the outlier is only 0.05, compared to 0.47 for the classical estimate. A detailed description of robust correlation and covariance estimates is provided in Chapter 6.

When there are more than two variables, examining all pairwise scatterplots for outliers is hopeless unless the number of variables is relatively small. But even looking at all scatterplots or applying a robust correlation estimate to all pairs does not suffice, for in the same way that there are bivariate outliers that do not stand out in any univariate representation, there may be multivariate outliers that heavily influence the correlations and do not stand out in any bivariate scatterplot. Robust methods deal with this problem by estimating all the correlations simultaneously, in such a manner that points far away from the bulk of the data are automatically downweighted. Chapter 6 considers these methods in detail.

1.6 Other parametric models

We do not want to leave the reader with the impression that robust estimation is only concerned with outliers in the context of an assumed normal distribution model. Outliers can cause problems in fitting other simple parametric distributions such as an exponential, Weibull or gamma distribution, where the classical approach is to use a nonrobust maximum likelihood estimate (MLE) for the assumed model. In these cases one needs robust alternatives to the MLE in order to obtain a good fit to the bulk of the data.

For example, the exponential distribution with density

$$f(x;\lambda) = \frac{1}{\lambda} e^{-x/\lambda}, \quad x \geq 0$$

is widely used to model random inter-arrival and failure times, and it also arises in the context of times-series spectral analysis (see Section 8.14). It is easily shown that the parameter λ is the expected value of the random variable x – in other words, $\lambda = \mathrm{E}(x)$ – and that the sample mean is the MLE. We already know from the previous discussion that the sample mean lacks robustness and can be greatly influenced by outliers. In this case the data are nonnegative so one is only concerned about large positive outliers that cause the value of the sample mean to be inflated in a positive direction. So we need a robust alternative to the sample mean, and one naturally considers use of the sample median $\mathrm{Med}(\mathbf{x})$. It turns out that the sample median is an *inconsistent* estimate of λ: it does not approach λ when the sample size increases, and hence a correction is needed. It is an easy calculation to check that the median of the exponential distribution has value $\lambda \log 2$, where log stands for natural logarithm, and so one can use $\mathrm{Med}(\mathbf{x})/\log 2$ as a simple robust estimate of λ that is consistent with the assumed model. This estimate turns out to have desirable robustness properties, as described in Problem 3.15.

The methods of robustly fitting Weibull and gamma distributions are much more complicated than the above use of the adjusted median for the exponential distribution. We present one important application of robust fitting a gamma distribution due to Marazzi *et al.* (1998). The gamma distribution has density

$$f(x; \alpha, \sigma) = \frac{1}{\Gamma(\alpha)\sigma^{\alpha}} x^{\alpha-1} e^{-x/\sigma}, \quad x \geq 0$$

and the mean of this distribution is known to be $\mathrm{E}(x) = \alpha\sigma$. The problem has to do with estimating the length of stay (LOS) of 315 patients in a hospital. The mean LOS is a quantity of considerable economic importance, and some patients whose hospital stays are much longer than those of the majority of the patients adversely influence the MLE fit of the gamma distribution. The MLE values turn out to be $\widehat{\alpha}_{MLE} = 0.93$ and $\widehat{\sigma}_{MLE} = 8.50$, while the robust estimates are $\widehat{\alpha}_{Rob} = 1.39$ and $\widehat{\sigma}_{Rob} = 3.64$, and the resulting mean LOS estimates are $\widehat{\mu}_{MLE} = 7.87$ and $\widehat{\mu}_{Rob} = 4.97$. Some patients with unusually long LOS values contribute to an inflated estimate of the mean LOS for the majority of the patients. A more complete picture is obtained through the figures below.

Figure 1.7 shows a histogram of the data along with the MLE and robust gamma density fit to the LOS data. The MLE underestimates the density for small values of LOS and overestimates the density for large values of LOS, thereby resulting in a larger MLE estimate of the mean LOS, while the robust estimate provides a better overall fit and a mean LOS that better describes the majority of the patients. Figure 1.8 shows a gamma Q–Q plot based on the robustly fitted gamma distribution. This plot reveals that the bulk of the data is well fitted by the robust method, while approximately 30 of the largest values of LOS appear to come from a sub-population of the patients characterized by longer LOS values. This is best modeled separately using another distribution, possibly another gamma distribution with different values of the parameters α and σ.

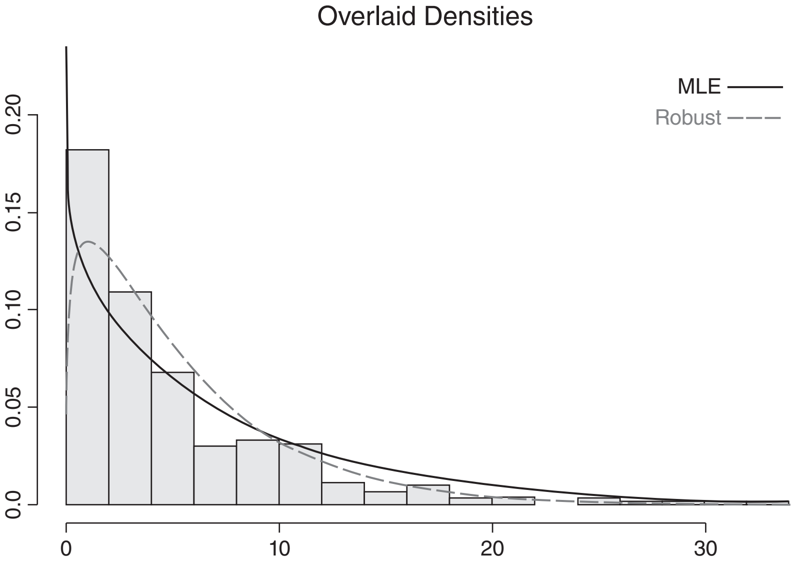

Figure 1.7 MLE and robust fits of a gamma distribution to LOS data

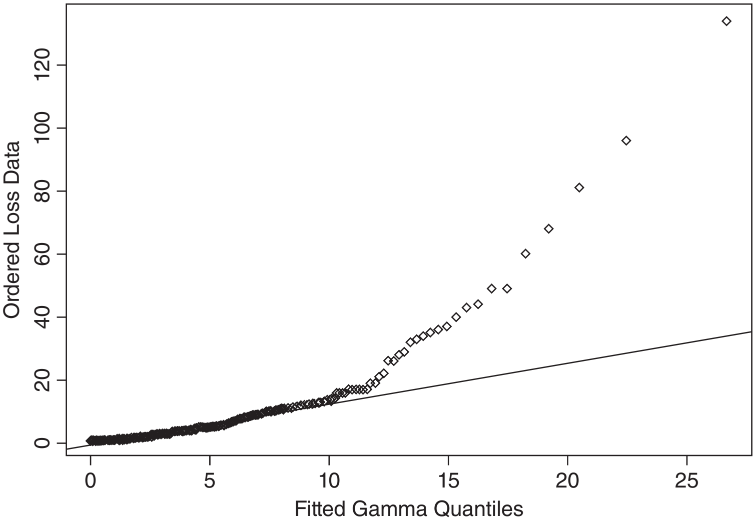

Figure 1.8 Fitted gamma QQ-plot of LOS data

1.7 Problems

1.1. Show that if a value x_0 is added to a sample $\mathbf{x} = \{x_1, ..., x_n\}$, when x_0 ranges from $-\infty$ to $+\infty$, the standard deviation of the enlarged sample ranges between a value smaller than SD($\mathbf{x}$) and infinity.

1.2. Consider the situation of the former problem.

(a) Show that if n is even, the maximum change in the sample median when x_0 ranges from $-\infty$ to $+\infty$ is the distance from Med($\mathbf{x}$) to the next order statistic the farthest from Med($\mathbf{x}$).

(b) What is the maximum change if n is odd?

1.3. Show for t_i defined in (1.3) that $|t_i| < (n-1)/\sqrt{n}$ for all possible datasets of size n, and hence for all datasets $|t_i| < 3$ if $n \leq 10$.

1.4. The interquartile range (IQR) is defined as the difference between the third and the first quartiles.

(a) Calculate the IQR of the $\mathrm{N}(\mu, \sigma^2)$ distribution.

(b) Consider the sample interquartile range

$$\mathrm{IQR}(\mathbf{x}) = \mathrm{IQR}(x_1, x_2, ..., x_n) = x_{(\lfloor 3n/4 \rfloor)} - x_{(\lfloor n/4 \rfloor)}$$

as a measure of dispersion. It is known that sample quantiles tend to the respective distribution quantiles if these are unique. Based on this fact, determine the constant c such that the normalized interquartile range IQRN($\mathbf{x}$) = IQR ($\mathbf{x}$)/c is a consistent estimate of σ when the data has a $\mathrm{N}(\mu, \sigma^2)$ distribution.

(c) Can you think of a reason why you would prefer MADN($\mathbf{x}$) to IQRN($\mathbf{x}$) as a robust estimate of dispersion?

1.5. Show that the median of the exponential distribution is $\lambda \log 2$, and hence Med($\mathbf{x}$)/ $\log 2$ is a consistent estimate of λ.