1

Introduction

1.1 What is Robust Statistics?

The scientific method is a set of principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses (Merriam-Webster online dictionary, http://merriam-webster.com).

Although procedures may be different according to the field of study, scientific researchers agree that hypotheses need to be stated as explanations of phenomena, and experimental studies need to be designed to test these hypotheses. In a more philosophical perspective, the hypothetico-deductive model for scientific methods (Whewell, 1837, 1840) was formulated as the following four steps: (1) characterizations (observations, definitions and measurements of the subject of inquiry); (2) hypotheses (theoretical, hypothetical explanations of observations and measurements of the subject); (3) predictions (possibly through a model, logical deduction from the hypothesis or theory); (4) experiments (test (2) and (3), essentially to disprove them). It is obvious that statistical theory plays an important role in this process. Not only are measurements usually subject to uncertainty, but experiments are also set using the theory of experimental designs and predictions are often made through a statistical model that accounts for the uncertainty or the randomness of the measurements. As statisticians, however, we are aware that models can at best be approximated (at least for the random part), and this introduces another type of uncertainty into the process. G. E. P. Box’s famous citation that ‘all models are wrong, some models are useful’ (Box, 1979) is often cited by the researcher when faced with the data to analyze. Hence, for truly honest scientific
research, statistics should offer methods that not only deal with uncertainty of the collected information (sampling error), but also with the fact that models are at best an approximation of reality. Consequently, statistics should be in ‘some sense’ robust to model misspecifications. This is important since the aim of scientific research is the pursuit of knowledge that is used in fine to improve the wellbeing of people as is obviously the case, for example, in medical research.

Robust methods date back to the prehistory of statistics and they naturally start with outlier detection techniques and the subsequent treatment of the data. Mathematicians of the 18th century such as Bernoulli (1777) were already questioning the appropriateness of rejection rules, a common practice among astronomers of the time. The first formal rejection rules are suggested in the second part of the 19th century; see Hampel et al. (1986, p. 34), for details. Student (1927) proposes repetition (additional observations) in the case of outliers, combined with rejection. Independently, the use of mixture models and simple estimators that can partly downweight observations appears from 1870 onwards; see Stone (1873); Edgeworth (1883); Newcomb (1886) and others. Newcomb even imagines a procedure that can be posthumously described as a sort of one-step Huber estimator (see Stigler, 1973). These attempts to reduce the influence of outliers, to make them harmless instead of discarding them, are in the same spirit as modern robustness theory; see Huber (1972); Harter (1974–1976); Barnett and Lewis (1978) and Stigler (1973). The idea of a ‘supermodel’ is proposed by Pearson (1916) who embedded the normal model that gained a central role at the turn of the 20th century into a system of Pearson curves derived from differential equations. The curves are actually distributions where two additional parameters are added to ‘accommodate’ most deviations from normality. The discovery of the drastic instability of the test for equality of variance by Pearson (1931) sparked the systematic study of the non-robustness of tests. Exact references on these developments can be found in Hampel et al. (1986, pp. 35–36).

The term robust (strong, sturdy, rough) itself appears to have been proposed in the statistical literature by Box (1953). The field of modern robust statistics finally emerged with the pioneering works of Tukey (1960), Huber (1964) and Hampel (1968), and has been intensively developed ever since. Indeed, a rough bibliographic search in the Current Index to Statistics1 revealed that since 1960 the number of articles having the word ‘robust’ in their title and/or in their keywords list has increased dramatically (see Figure 1.1). Compared with other well-established keywords, ‘robust’ appears to be quite popular: roughly half as popular as ‘Bayesian’ and ‘design’, but more popular than ‘survival’, ‘bootstrap’, ‘rank’ and ‘smoothing’. Is robust statistics really as popular as it appears to be, in that it is used fairly routinely in practical data analysis? We do not really believe so. It might be that the word ‘robust’ is associated with other keywords such as ‘rank’, ‘smoothing’ or ‘design’ because of the perceived nature of these methods or procedures. We also performed a rough bibliographic search under the same conditions as before, but with the combination of the words ‘robust’ and each of the other words. The result is presented in Figure 1.2. It appears that although ‘robust’ is relatively more associated

1http://www.statindex.org/
1.2 Against What is Robust Statistics Robust?

Robust statistics aims at producing consistent and reasonably efficient estimators, test statistics with stable level and power, when the model is slightly misspecified.
Model misspecifications encompass a relatively large set of possibilities, and robust statistics cannot deal with all types of model misspecifications. First we characterize the model using a cumulative probability distribution $F_\theta$ that captures the structural part as well as the random part of the model. The parameters needed for the structural part and/or the random part are included in the parameter’s vector $\theta$. For example, in the regression model that is thoroughly studied in Chapter 3, $\theta$ contains the (linear) regression coefficients (structural part) as well as the residual error variance (random part) and $F_\theta$ is the (conditional) normal distribution of the response variable (given the set of explanatory variables). Here $F_\theta$ does not need to be fully parametric, e.g. the Cox model presented in Chapter 7 can also be used. Then, by ‘slight model misspecification’, we assume that the data-generating process lies in a neighborhood of the true (postulated) model $F_\theta$ that is considered as ‘useful’ for the problem under investigation. This notion of a neighborhood, due originally to Huber (1964),
1.2. AGAINST WHAT IS ROBUST STATISTICS ROBUST?

is formalized as

\[ F_\varepsilon = (1 - \varepsilon) F_\theta + \varepsilon G, \]  

(1.1)

where \( F_\theta \) is the postulated model, \( \theta \) is a set of parameters of interest, \( G \) is an arbitrary distribution and \( 0 \leq \varepsilon \leq 1 \). The form of \( G \) is not really important, but there are some interesting special cases. For example, \( G \) can be a gross error-generating process (or point mass distribution), i.e.

\[ G(x) = \Delta_\varepsilon(x) = \begin{cases} 
0 & x < z, \\
1 & x \geq z. 
\end{cases} \]  

(1.2)

In other words, data generated from \( F_\varepsilon \) are from \( F_\theta \) with probability \( 1 - \varepsilon \) and from \( G \) with probability \( \varepsilon \). Since \( G \) is an arbitrary distribution, and \( \varepsilon \in (0, 1) \), the neighborhood \( F_\varepsilon \) is very general. However, the crucial quantity is actually \( \varepsilon \), which in a sense measures the ‘amount’ of model misspecification. When \( \varepsilon = 0 \), then there is no model misspecification in that the data-generating process is exactly the postulated model. This is the fundamental hypothesis in classical estimation based, for example, on the maximum likelihood estimator (MLE) and classical testing based, for example, on the \( F \)-test in analysis of variance (ANOVA). For a data analysis practitioner, experience shows that the chosen ‘useful’ model \( (F_\theta) \) is very rarely equal to the data-generating process \( (F_\varepsilon) \). Assuming that the data analyst does not ignore this fact, he/she is faced with the problem of ‘what to do next’. There exist many ‘practical strategies’ that have been developed over the years to process the data in an \textit{ad-hoc} fashion or tweak the model to ultimately resort to classical inference. Most of the strategies may fail in that what is sought is not necessarily what is found. Indeed, when \( 0 < \varepsilon < 1 \), the situation becomes murkier. If one truly believes that the data-generating process \( F_\varepsilon \) is the true model, then inference should be carried out at \( F_\varepsilon \). A mixture distribution should then be used, assuming that \( G \) can be chosen adequately. For instance, in injury prevention, researchers are interested in modeling the number of crashes involving young drivers or the number of serious injuries. Such outcomes often display a large number of zeros and are typically modeled by a zero-inflated Poisson or negative binomial distribution (see, e.g., Lambert, 1992) or alternatively using hurdle, two-step or conditional models (see, e.g., Welsh \textit{et al.}, 1996). Inference is at the zero-inflated model (or one of the other models), represented as a mixture model with \( \varepsilon \), as the probability of an observation being part of the (excess) spike at zero, commonly described on the logistic scale through a set of covariates. This is a reasonable model if \( \varepsilon \) is relatively large, but there is no guarantee that the resulting mixture (after choosing \( G \)) is the exact data-generating process. Inference is then sought simultaneously on \( \theta \), \( \varepsilon \) and the other parameters (or directly on the quantiles of \( G \) if it is a non-parametric model). The

\footnote{Note that (1.1) is not exactly a neighborhood in the mathematical sense. However, Huber’s idea was to imagine a workable set of distributions that were ‘close enough’ to the assumed model, hence the use of the term neighborhood. He proved that any distribution in (1.1) is within a distance \( \varepsilon \) of \( F_\theta \) for a proper metric on the distribution space such as the Kolmogorov or Prohorov distance; see Huber (1981, Chapter 2).}
procedure can become very cumbersome because the number of parameters can become very large.

If \( \varepsilon \) is relatively small and \( G \) is not obvious to define, then another strategy should be chosen. Indeed, very often \( F_\theta \) is chosen because it makes sense with respect to the problem under investigation, so that another model is less ‘interpretable’. Moreover, when focusing on \( F_\theta \) and \( \varepsilon > 0 \), it is very difficult to define \( G \) in order to use \( F_\varepsilon \) as the ‘true’ model. In practice, discovering the form of \( G \) from the data is often impossible, unless the sample size is very large. Most of the time, the best ‘guess’ is \( F_\theta \) and the data can be assumed to have been generated approximately by \( F_\theta \). As stated previously, even if one has a fairly good idea about \( G \), you still cannot be sure that the mixture \((1 - \varepsilon)F_\theta + \varepsilon G\) is the exact data-generating process. Finally, the mixture can be so complicated that one may wonder whether it is even worth using \( F_\varepsilon \) for small \( \varepsilon \) when one is actually interested in inference about \( F_\theta \).

Another situation (at least in theory) occurs when \( \varepsilon = 1 \). In this case it would make no sense to still seek inference about \( F_\theta \), so the postulated model should be changed to \( G \). However, in generalized linear mixed models, for example, several authors have studied the ‘robustness’ of the Wald test to the assumption of normality of the random effects. Hence, in these cases, \( F_\theta \) is the mixed model with normal random effects, \( G \) is the mixed model with non-normal random effects and \( \varepsilon = 1 \). One then seeks inference about \( \theta \) when \( F_\varepsilon \) with large \( \varepsilon \) is the data-generating process. Some of the proposed procedures have been found to be ‘robust’ in some particular situations and for some specific distributions for the random effects; see e.g. Litière et al. (2007b) for details. Although this type of robustness is also important, it is limited to some particular instances of \( G \) (i.e. for some distributions for the random effects). This actually concerns a type of model misspecification that can be called structural misspecification in that the form of \( G \) is known (and \( \varepsilon \) is large). The robust procedures we propose here are robust in the sense that inference remains correct at \( F_\theta \) even if \( F_\varepsilon \) is the data-generating process and \( \varepsilon \) is unknown but small and \( G \) can be of any form. The type of model misspecification in this case can be called distributional misspecification in that the best that can be said is that the data-generating process is approximately \( F_\theta \) (small \( \varepsilon \)).

Seeking inference about \( F_\theta \) when \( F_\varepsilon \) is the actual data-generating process is not the same as seeking inference about \( F_\varepsilon \) when \( F_\theta \) is fitted. Indeed, sometimes classical procedures (with possible added corrections) are said to be robust to model misspecification in the sense that the estimator of \( \theta \) (when \( F_\theta \) is fitted) still provides consistent estimators for (some of) the parameters of \( F_\varepsilon \). For example, in the important case of the omission of covariates, we would have \( G \) (assuming that it exists) such that \( F_\varepsilon = F_\theta(\theta', \varepsilon) \) where \( \theta' \) is the added parameter corresponding to the missing predictors. This is another case of structural misspecification that is not covered by the robustness theory introduced in this book. In the 1980s there were some important studies of the conditions under which a consistent estimate of \( \theta \), assuming \( F_\theta(\theta', \varepsilon) \) (as the true model) but fitting \( F_\theta \) could still be obtained; see Gail et al. (1984) and Bretagnolle and Huber-Carol (1988) for instance. They essentially

---

3This type of ‘robustness’ is such that the level of the test is preserved under these assumptions.
1.3. AN ALTERNATIVE TO ROBUST STATISTICS?

Table 1.1 Models at which inference can at best be made.

<table>
<thead>
<tr>
<th>Inference</th>
<th>$G$</th>
<th>$\varepsilon = 0$</th>
<th>$0 &lt; \varepsilon \ll 1$</th>
<th>$0 &lt; \varepsilon &lt; 1$</th>
<th>$\varepsilon = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td>Arbitrary</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>$F_\theta$</td>
</tr>
<tr>
<td>$G = \Delta_\varepsilon$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td></td>
</tr>
<tr>
<td>$G$ specified</td>
<td>$G$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td></td>
</tr>
<tr>
<td>Robust</td>
<td>Arbitrary</td>
<td>?</td>
<td>?</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
</tr>
<tr>
<td>$G = \Delta_\varepsilon$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td></td>
</tr>
<tr>
<td>$G$ specified</td>
<td>$G$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td>$F_\theta$</td>
<td></td>
</tr>
</tbody>
</table>

showed that a small residual bias remains although in some simple cases such as the linear model this situation does not occur.

In Table 1.1 we summarize some of the possible situations discussed so far regarding the postulated model, the data-generating process, the value of $\varepsilon$, the form of $G$ and the estimation method. Except in the case when $\varepsilon = 0$, classical inference is not (at least a priori) suitable in most of the situations considered here. Moreover, even if theoretically one could postulate $F_\varepsilon$ instead of $F_\theta$, the former is often difficult to find and/or to estimate. The robust methods we propose in this book provide an alternative (and more effective) approach when $\varepsilon$ is relatively small. We propose a set of statistical tools for correct estimation and inference about $F_\theta$ when the data-generating process is $F_\varepsilon$, not only when $\varepsilon = 0$, as with classical methods, but also for relatively small $\varepsilon$ and any $G$. As a by-product, data not fitting $F_\theta$ exactly can be easily identified, and the model can possibly be changed and refitted.

Hence, one possibility is to manipulate the data so that they ‘fit’ the postulated model, but as argued below, this is not a good method. Another possibility is to change the model, but it is not always clear what a suitable alternative model may be. One could also use a more flexible model (e.g. using non-parametric statistics), but care should be taken as to what the underlying assumptions really are (see the discussion in Section 1.3). The alternative we propose here is to use robust statistics, which allows one to make inferences about $F_\theta$, when the data-generating process is actually $F_\varepsilon$, with small $\varepsilon$ and arbitrary $G$. We spend the remainder of this chapter explaining how robust statistics work in general to achieve these goals.

1.3 Are Diagnostic Methods an Alternative to Robust Statistics?

Since classical methods, when $F_\theta$ is the postulated model, only work when $\varepsilon = 0$, one could be tempted to modify the data by removing ‘dubious’ observations from the sample. By ‘dubious’ observations we mean here that they are in some sense far from the bulk of the data generated by $F_\theta$. The measure of how far an observation is from the bulk of the data is highly dependent on the problem (hence the model). For the problem of estimating the mean of a population, also
called the location problem, measures such as standard deviations are sometimes used to build thresholds (e.g. three standard deviations around the mean) outside which observations are considered as outliers. Hampel (1985) provides an account of several measures for the location problem and compares the properties of the mean estimator computed after the removal of outliers. In more complex situations, the measure can be based on graphical tools such as boxplots and/or scatterplots and constructed before the model is fitted. Alternatively, one could rely on some sort of ‘residual analysis’ (i.e. estimation of the random part of the model, once the estimated fixed part has been removed) for checking the distributional assumptions. More sophisticatedly, in regression models, the so-called ‘diagnostic’ techniques could be used (see, e.g. Atkinson, 1985; Belsley et al., 1980; Cook and Weisberg, 1982). One such well-known tool is the Cook distance. The strategy of removing observations, although apparently simple, can be not only unpractical, but also very misleading. The main arguments are as follows.

- Graphical tools used before the model is fitted are only suitable for simple problems, such as when comparing groups (testing differences in mean responses), without control variables or in correlation settings (such as regression models) when there are only at most three variables (the response and two explanatory variable) if one uses three-dimensional graphs. When the dimension is higher, then combinations of (often pairs of) variables could be used, but multivariate effects could be masked.

- In practice, however, it is not always obvious how to quantify ‘far from’, and some observations might appear to be just at the (imaginary) border. Then the analyst is left with a rather arbitrary decision to make.

- Sometimes raw measures of ‘outlyingness’ are used that are based on standard deviations (e.g. remove observations that are three standard deviations away from the mean). This leaves the question of how the standard deviations (and the mean) are estimated open. The chosen scale estimator could be inflated and the mean itself biased by outlying observations generating a masking effect (see, e.g. Rousseeuw and Leroy, 1987).

- Moreover, when this is done in a univariate fashion, outlying observations are found only with respect to one variable at a time masking the effects of other covariates (see the example below).

- A ‘residual analysis’ can be used to detect ‘outlying’ observations once the model is fitted. This is commonly done in regression models. However, this procedure is not flawless as it does not take into account how the residuals are estimated. Indeed, if classical estimators for the model parameters are computed, they can be seriously biased by model deviations such as outliers. Hence, residuals obtained through these biased estimates will, in turn, be biased. This is another illustration of the masking effect. Removing observations on the basis of potentially biased residual estimates can become a very dangerous strategy.
The same argument applies to other diagnostic tools based on classical estimators of the model parameters. Even if the diagnostic tools are based on the comparison of fitted models with and without one observation at a time such as the Cook distance, the simultaneous effect of multiple outliers could be masked.

The ‘data-cleaning’ process can become very cumbersome in that one or some of the observations are removed on the basis of some criteria, then the model refitted, the criteria calculated again, new data are removed, etc. The process may never end at a satisfactory stage and a large proportion of data are removed before the process is stabilized. It is also unfeasible for large datasets.

However, perhaps the most important argument is inference. A proper inferential procedure should take into account the data manipulation. In other terms, inference (e.g. significance tests) should be conditional on the criteria used for the removal of the observations. The use of classical inference (e.g. $t$-test) after case deletion and refit ignores this problem and is therefore dubious and, in some cases, completely wrong (see also Welsh and Ronchetti, 2002).

In our view the true purpose of diagnostic methods should be to identify genuine structural model misspecifications, e.g. adding a quadratic term, a missing covariate or an interaction in the model, identifying a systematic violation to the proportional hazard assumption in the Cox model or an incorrect formulation of the random component of a mixed linear model. They do not oppose robust methods, they are just complementary.

To illustrate the danger of a relatively naive data-cleaning process used before fitting, we consider the following dataset which will also be reanalyzed in Chapter 3 using robust techniques. The data (kindly provided by Dr Pascal Bovet, IUMSP, Lausanne, Switzerland) come from a study investigating the prevalence of hyperuricemia and the association of uric acid levels with various cardiovascular risk factors in the Seychelles Islands. A total of 998 participants from this population, mainly of African origin, were included in the study; see Conen et al. (2004). The primary outcome, serum uric (uric), is typically analyzed by linear regression with predictors such as the triglycerides level in body fat (trig) and the low-density lipoprotein (ldl) cholesterol; see Section 3.5 for a complete description of all covariates. Before the regression model is fitted, a descriptive data analysis should be performed. Boxplots of each of the variables can be drawn to detect extreme measurements, as well as scatterplots of pairs of variables to study their relationship and possibly detect outlying observations. In Figures 1.3 and 1.4 we present the scatterplots of uric versus trig and ldl versus trig, respectively. The vertical and horizontal lines represent the values of the sample means plus three standard deviations (i.e. the quantile 0.999 on the normal distribution) for each variable. As is routinely performed, a ‘cleaning’ mechanism based on this univariate criterion would remove from the sample all of the data represented by the points lying to the right and above these lines on both graphs. This rather rough mechanism does not take into account the possible correlation between the variables, especially between the
response (\textit{uric}) and the explanatory variables. A more sophisticated but still rough mechanism is to consider as extreme values those with low probability under the bivariate normal model, or in other words, observations lying outside the quantiles with equal density (say with corresponding cumulative probability of 0.999) as illustrated by the ellipses on the scatter plots. To draw these ellipses, one needs to estimate the bivariate center and the covariance matrix between the pairs of variables. The classical estimators are the sample means, variances and Pearson correlation. If there are extreme observations (away from the bulk of data), these estimators can be artificially inflated, and this is the case with both examples in Figures 1.3 and 1.4. Alternatively, one can compute the ellipses based on a robust estimation of the center and covariance matrix between the pairs of variables (see Section 2.3.3), which, in the examples taken here, lead to better ‘centered’ ellipses (with respect to the bulk of data) of smaller volume. Using the ellipse-based criterion for ‘cleaning’ the data does not lead to the same decisions regarding the data to discard. In particular, observations that have not been removed with the three standard deviations would be removed with the classical ellipse, while one outside the three standard deviations on the \textit{ldl} variable in Figure 1.4 would not be discarded with the robust ellipse. One can also notice that more observations would be discarded with the robust ellipses.

The difference in the appreciation of the ‘outlyingness’ of one observation between a univariate and a multivariate approach is due to the fact that the underlying model upon which the decisions are made is not the same. Indeed, with the ellipses, the correlation between the two variables is taken into account, which is not the case with a univariate approach. One could also consider multivariate criteria, i.e. criteria based on the relationship between all of the variables simultaneously. Such a criterion is given by the Mahalanobis distances (Mahalanobis, 1936) based on the multivariate normal assumption of the data; see (2.34). In the bivariate case as in Figures 1.3 and 1.4, the points on the same ellipses have equal Mahalanobis distances. Hence, a limit distance could be chosen and points with corresponding Mahalanobis distance exceeding this limit could be discarded from the sample. This would lead to the rejection of yet different observations. Even if this approach takes into account the relationships between all of the variables simultaneously, and hence is better than a univariate or bivariate approach, it is not satisfactory for the chosen example. Given that the data are actually used to explain the response variable \textit{uric} through a regression model, extreme observations should be chosen with respect to the regression model, and this can only be done through a robust estimation of the latter. The complete robust analysis of the cardiovascular risk factors data by means of a regression model will be presented in Section 3.5. In this analysis, observations are down-weighted according to their degree of ‘outlyingness’. In Figures 1.3 and 1.4 extreme observations (weight less than or equal to 0.3) with respect to the final regression model estimated in Table 3.10 have been drawn using the symbol $\circ$. The striking feature is that, although most of them correspond to observations that would have been discarded using one of the previous \textit{ad-hoc} methods, they do not correspond to all of them (hence, more data are used for the estimators, and consequently inference is more powerful). More dramatically, some observations that would not have been removed with the \textit{ad-hoc} method are strongly downweighted by
means of the robust regression estimator, questioning the validity of the procedure. In other words, the ‘outlyingness’ of an observation is relative to a model, and procedures that do not take this fact into account are not good procedures.

1.4 How do Robust Statistics Compare with Other Statistical Procedures in Practice?

Robust methods have not seen much use in clinical trials or epidemiological studies, except in a few cases such as Conen et al. (2004), Kurtio et al. (2005), Tashkin et al. (2006) and Wen et al. (2004). The only areas were the penetration of such methods is not anecdotal are medical imaging and genetics where robust regression (and smoothing techniques) are commonly used successfully; see, for instance, Ma and Elis (2003), Wager et al. (2003) and Jain et al. (2005). Apart from these particular
cases, there seems to be a general feeling that outliers rarely occur in clinical trials, and if present they can properly be dealt with by means of traditional methods (e.g., rank-based techniques, described below). We believe that there is no reason to think that distributional model misspecifications are less present in clinical research than in any other area. If it is true that in regression settings covariates can be well controlled, extreme responses may nevertheless be present and hence ruin the interpretation of many standard procedures. A sense of false security arises as binary endpoints such as the occurrence of a specific event (e.g., death, disease progression, relapse) or the corresponding times to event are routinely studied. Even if procedures such as the chi-squared test typically used to analyze binary data or the log-rank test for survival times are less sensitive to extreme responses, one often forgets that a treatment estimate generally has to be given and some form of modeling assumed. If, as is done most of time, the Cox proportional hazard model is used, we show in Chapter 7

Figure 1.4 Scatterplot of the variables $\text{trig}$ and $\text{ldl}$ for the cardiovascular risk factors data, together with the three standard deviations univariate limits, the robust and classical 0.999 bivariate normal contours. The symbol $\circ$ denotes observations that have been downweighted by a robust regression analysis.
1.4. HOW DO ROBUST STATISTICS COMPARE?

that the classical estimator for the hazard ratio based on the partial likelihood can be ruined by a small number of abnormal long-term survivors. In addition, if an adjusted analysis is performed as a second step, or if the outcome is simply continuous, the situation is about the same as any other area in biostatistics.

In clinical research, the case deletion and refit is not a satisfactory alternative to robust statistics, not only for the reasons discussed in Section 1.3, but also because such a procedure violates the intention-to-treat (ITT) principle. This principle (Hollis and Campbell, 1999) states that patients must be analyzed as randomized, irrespective of what actually happens and usually assumes that everybody randomized on the trial is included in the analyzed datasets. In contrast, robust techniques have solid theoretical underpinnings, protect against outliers and other model misspecifications and offer an elegant way to preserve the ITT principle by automatically downweighting extreme observations instead of deleting them in an ad-hoc fashion. In addition, there is no reason to think that results obtained from a robust fit (when appropriately used) will favor a particular outcome, say a positive effect of the drug under investigation, which is precisely what the guidelines request as a proper way to deal with outliers.

Rank-based methods (see e.g. Hettmansperger and McKean, 1998) are more sensible and this may explain why they are so popular for some types of analysis, such as survival analysis where the log-rank test is systematically used. However, they are not always available for more complex techniques, e.g. generalized estimating equations, complex models with covariates or mixed linear models. In addition, power issues have to be taken into consideration, an issue that is often overlooked. As described later in this book, it is often possible to calibrate the robust procedures we propose to achieve a pre-specified efficiency in the model (e.g. 90–95%). In other words the price to pay for the use of such techniques with respect to the MLE and related tests is a small loss in efficiency if the model holds.

It is also often believed that resampling methods such as the bootstrap can be used as an alternative to robust methods as ‘one does not need to specify the distribution’. One should first recall that the bootstrap method of Efron (1982) (see also e.g. Davison and Hinkley, 1997) is a technique allowing the computation of standard errors, confidence intervals and $p$-values that are based on given estimators or test statistics. It thus does not provide new estimators or test statistics ‘per se’. This method can be used for parametric, semi-parametric or even non-parametric analyses. What is understood behind this distribution-free assumption is that the sampling error distribution does not usually need to be a given parametric model (such as the normal distribution). One simply assumes instead that the observations are ‘independent and identically distributed’ (i.i.d.). The bootstrap and other non-parametric methods do not become naturally robust to model misspecification just because the model sampling error distribution is not specified.

We do not believe that the applied statistician has no ‘model’ in mind when stating the i.i.d. condition. Indeed, in order to summarize a group response by a ‘mean parameter’, e.g. the mean cholesterol per treatment arm in a statin trial, it is

---

4Statins are drugs that improve blood cholesterol levels.
implicitly assumed that the response distribution is somewhat symmetric around this mean. Otherwise it would make little sense to summarize the outcome in this way but it would be more sensible to compare the (whole) response distribution across the treatment arm. In other words, even if the bootstrap provides good inference techniques without the need to specify a data-generating process (e.g. sampling error distribution), what is tested (choice of the parameters such as the ‘mean’ response) is not necessarily appropriate in all situations (e.g. bimodal distributions instead of symmetric distributions) and the conclusions can be very misleading.

Resampling techniques can be particularly sensitive to some types of model deviations such as outliers. Indeed, some of the bootstrap samples will invariably have a larger proportion of outliers and therefore heavily bias the estimators or test statistics computed on these samples. Confidence intervals or $p$-values derived from these bootstrapped statistics will then represent an ‘average’ between estimates (or tests statistics) computed from samples with different proportions of contaminated data (outliers). One might then wonder whether these confidence intervals (or $p$-values) are really informative, since they are representative of neither the ‘clean’ data nor the outliers. A standard bootstrap procedure may even fail when applied to a robust estimator as it may not necessarily withstand more than a certain proportion of outliers (i.e. the breakdown point of the procedure is reached). A solution to this problem would then be to use a robust bootstrap procedure applied to a robust estimator as originally suggested by Salibian-Barrera and Zamar (2002) in the linear regression model.

Finally, in non-parametric regression, it is often thought that robustness is automatically achieved, given that the approach relies on relaxed hypotheses (no normality assumption for the error term). This feeling is reinforced by the fact that the non-parametric regression estimators (smoothers) are local averages and it is therefore wrongly believed that an outlier occurring in a given subspace of the design only affects the estimation around this region. In fact, quite long ago Huber (1979) had already warned against the non-robustness of non-parametric regression and proposed a robust version of smoothing splines. There are also other alternatives, for example the $M$-kernels of Härdle (1990) or the Locally Weighted Regression and Smoothing Scatterplots (LOWESS) of Cleveland (1979). More recently, Cantoni and Ronchetti (2001a) have shown that the data-driven choice of the smoothing parameter pertaining to smoothers also needs to be made robust. They propose both a cross validation and a $C_p$-like criterion to cope with this issue.

To conclude, the robust methods we propose in this book are based on the specification of a core (parametric) model $F_\theta$ such as, for example, the linear regression model, the mixed linear model, the generalized linear model (GLM), models for longitudinal data or a model which might contain some non-parametric parts such as the Cox model for survival data. We assume, however, that the data are generated by a distribution in a neighborhood (1.1). In order to avoid the potential bias on classical estimators, test statistics and other inferential procedures of this type of model misspecification, we propose instead the use of alternative robust statistics which provide correct inference at the core model $F_\theta$. 