9

Outliers and robustness for ordinal data

Marco Riani, Francesca Torti and Sergio Zani

This chapter tackles the topics of robustness and multivariate outlier detection for ordinal data. We initially review outlier detection methods in regression for continuous data and give an example which shows that graphical tools of data analysis or traditional diagnostic measures based on all the observations are not sufficient to detect multivariate atypical observations. Then we focus on ordinal data and illustrate how to detect atypical measurements in customer satisfaction surveys. Next, we review the generalized linear model of ordinal regression and apply it to the ABC survey. The chapter concludes with an analysis of a set of diagnostics to check the goodness of the suggested model and the presence of anomalous observations.

9.1 An overview of outlier detection methods

There are several definitions of outliers in the statistical literature (see Barnett and Lewis, 1994; Atkinson *et al.*, 2004; Hadi *et al*., 2009). A commonly used definition is that outliers are a minority of observations in a data set that is represented by a common pattern which can be captured by some statistical model. The assumption here is that there is a core of at least 50% of observations that is homogeneous and a set of remaining observations (hopefully few) which has patterns that are inconsistent with this common pattern. Awareness of outliers in some form or another has existed for at least 2000 years. Thucydides, in his third book about the Peloponnesian War (III 20, 3–4), describes how in 428 BC the Plataeans used concepts of robust statistics in order to estimate the height of the ladder which was needed to overcome the fortifications built by the Peloponnesians and the Boeotians who were besieging their city.

> The same winter the Plataeans, who were still being besieged by the Peloponnesians and Boeotians, distressed by the failure of their provisions, and seeing no

Modern Analysis of Customer Surveys: with applications using R, First Edition. Edited by Ron S. Kenett and Silvia Salini.
 Published 2012 by John Wiley & Sons, Ltd.

> hope of relief from Athens, nor any other means of safety, formed a scheme with the Athenians besieged with them for escaping, if possible, by forcing their way over the enemy's walls; the attempt having been suggested by Theaenetus, son of Tolmides, a soothsayer, and Eupompides, son of Daimachus, one of their generals. At first all were to join: afterwards, half hung back, thinking the risk great; about two hundred and twenty, however, voluntarily persevered in the attempt, which was carried out in the following way. Ladders were made to match the height of the enemy's wall, which they measured by the layers of bricks, the side turned towards them not being thoroughly whitewashed. These were counted by many persons at once; and though some might miss the right calculation, most would hit upon it, particularly as they counted over and over again, and were no great way from the wall, but could see it easily enough for their purpose. The length required for the ladders was thus obtained, being calculated from the breadth of the brick.

Rejection of outliers prior to performing classical statistical analysis has thus been regarded as an essential preprocessing step for almost as long as the methods have existed. For example, a century and a half ago Edgeworth stated that: 'The method of Least Squares is seen to be our best course when we have thrown overboard a certain portion of our data – a sort of sacrifice which often has to be made by those who sail upon the stormy seas of Probability'.

For univariate data, Grubbs (1969) defined an outlier as an observation that 'appears to deviate markedly from other members of the sample in which it occurs'. For multivariate data, however, the examination of each dimension by itself or in pairs does not work, because it is possible for some data points, as we will see in the next section, to be outliers in multivariate space, but not in any of the original univariate dimensions. When the variables under study are not measured on a quantitative scale, as often happens in customer satisfaction surveys, the need to jointly consider several variables becomes of paramount importance in detecting atypical observations. For example, in the 11 September attacks on World Trade Center in New York, five out of the eighty passengers on one of the flights displayed unusual characteristics with respect to a set of qualitative variables. These five passengers were not US citizens, but had lived in the USA for some period of time, were citizens of a particular foreign country, had all purchased one-way tickets, had purchased these tickets at the gate with cash rather than credit cards, and did not have any checked luggage. One or two of these characteristics might not be very unusual, but, taken together, they could be seen as markedly different from the majority of airline passengers.

Identification of outlying data points is often by itself the primary goal, without any intention to fit a statistical model. The outliers themselves sometimes can be points of primary interest, drawing attention to unknown aspects of the data, or especially, if unexpected, leading to new discoveries. For example, one of the purposes of the Institute for the Protection and Security of the Citizen, one of the institutes of the European Commission's Joint Research Centre, is to find patterns of outliers in data sets including millions of trade flows grouped in a large number of small to moderate size samples. The statistically relevant cases, which can be due to potential cases of tax evasion or activities linked to money laundering, are presented for evaluation and feedback to subject-matter experts of the anti-fraud office of the European Commission and its partner services in the member states. In the context of international trade, the unexpected presence of a set of atypical transactions from or towards a particular country can be an indication of trade carousels (repetitive trade of the same good among the same persons), potential illegal activities or stockpiling.

The examples above clearly demonstrate the need for outlier identification both on small and very large data sets, whether the available variables are quantitative or qualitative. See Su and Tsai (2011) for a recent overview of the different approaches to outlier detection.

9.2 An example of masking

The purpose of this section is to show that when multivariate (multiple) outliers are present in the data, traditional tools of visual inspection or traditional data analysis based on least squares (LS) in regression may lead us to incorrect conclusions. Atkinson and Riani (2000, pp. 5–9) give an example of a regression data set with 60 observations on three explanatory variables (X1, X2 and X3) where there are six masked outliers that cannot be detected using standard analyses. The scatter plot of the response y against the three explanatory variables (see Figure 9.1) simply shows that y is increasing with each of X1, X2 and X3, but does not reveal particular observations far from the bulk of the data. The traditional plot of residuals against fitted values (see left panel of Figure 9.2) shows no obvious pattern and the largest residual is for a unit (no. 43) which lies well within the envelopes of the quantile–quantile plot of studentized residuals (see the right-hand panel of Figure 9.2).

In the regression context, the best-known robust methods are based on the use of least trimmed squares (LTS), least median of squares (LMS) or on forward search (FS) – see Morgenthaler (2007) and Rousseeuw and Hubert (2011) for recent reviews of robust methods in regression. The LTS regression method searches for an estimate of the vector of regression coefficients which minimizes the sum of the h smallest squared residuals, where h must be at least half the number of observations. On the other hand, the LMS estimator minimizes the median of the squares of the residuals. LTS and LMS are very robust methods in the sense that the estimated regression fit is not unduly influenced by outliers in the data, even if there are

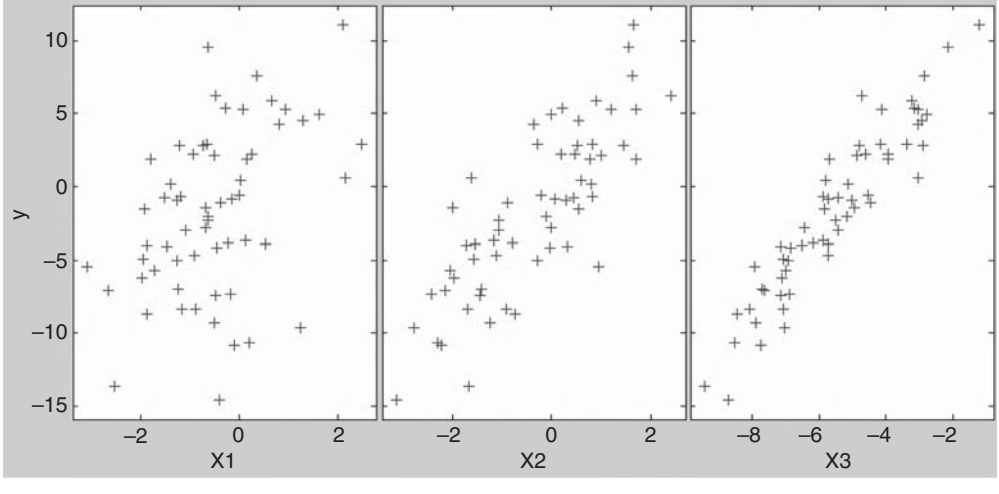

Figure 9.1 Plot of y *against each of the explanatory variables.*

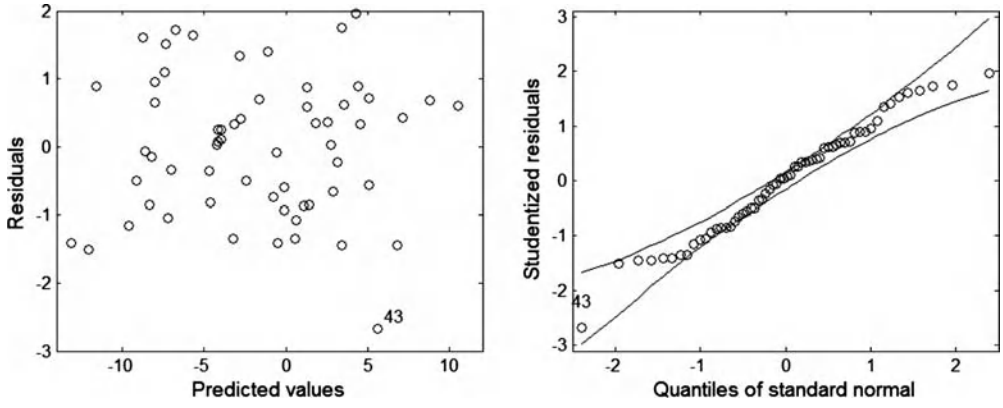

Figure 9.2 Traditional uninformative diagnostic plots. Left: least squares residuals against fitted values. Right: quantile–quantile plot of studentized residuals.

several atypical observations. Due to this robustness, we can detect outliers by their large LTS (LMS) residuals. Given that explicit solutions for LMS and LTS do not exist, approximate solutions are sought using elemental subsets, that is, subsets of p observations, where p is the number of explanatory variables including the intercept (Rousseeuw, 1984; Rousseeuw and Van Driessen, 2006).

Because of the way in which models are fitted, whether with LS, LTS or LMS, we lose information about the effect of individual observations on inferences about the form and parameters of the model. In order to understand the effect that each unit, outlier or not, exerts on the fitted model, it is necessary to start with a subset of data and monitor the required diagnostics. In the example above, if we start with a least squares fit to four observations, robustly chosen, we can calculate the residuals for all 60 observations and next fit to the five observations with smallest squared residuals. In general, given a fit to a subset of size m, we can order the residuals and take, as the next subset, the $m + 1$ cases with smallest squared residuals. This gives a forward search through the data (Atkinson and Riani 2000; Riani *et al.*, 2009), ordered by closeness to the model. We expect that the last observations to enter the search will be those that are furthest from the model and so may cause changes once they are included in the subset used for fitting. Figure 9.3 shows the monitoring of the scaled squared residuals for the 60 units of the data set. In this case we have initialized the search with LTS, investigating all possible $\binom{60}{4}$ subsets and taking the one with the smallest sum of the 50% smallest residuals, although this is not necessary. This fascinating plot reveals not only the presence of six masked outliers but also that:

(1) the six outliers form a cluster, because their trajectories are very similar – in other words, they respond in a similar way to the introduction of units into the subset;

(2) the residuals of the six outliers at the end of the search are completely mixed with those of the other units, therefore traditional methods based on single deletion diagnostics cannot detect them;

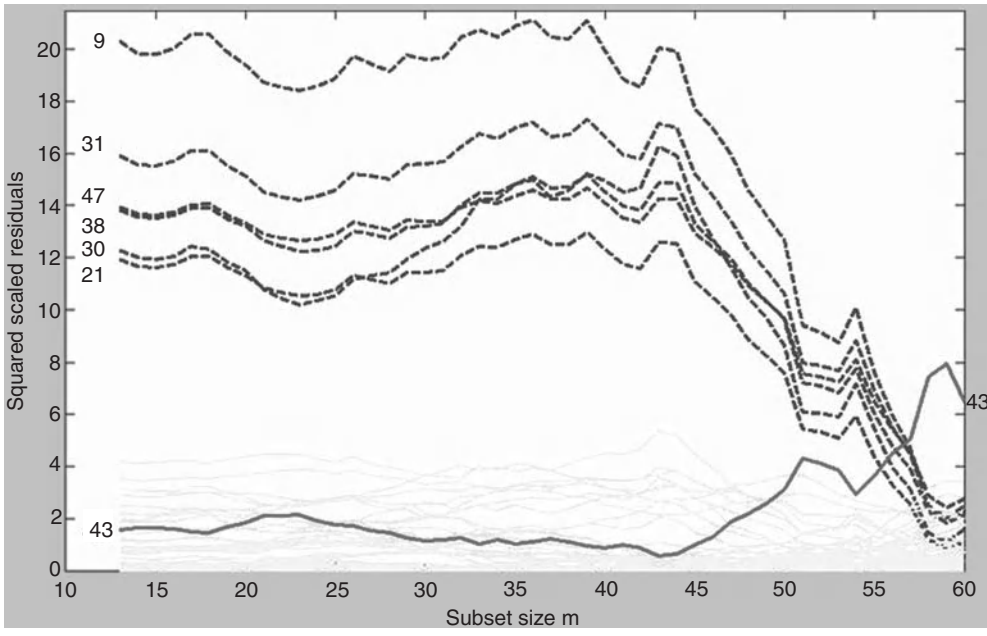

Figure 9.3 Monitoring of squared scaled residuals. The outliers have been drawn with dotted lines, while the trajectory of the case which in the final step shows the largest residual has been drawn with a solid line. All the other unimportant trajectories have been shown in faint grey.

(3) the entry of the six outliers causes a big increase in the trajectory of the residual for unit 43. Indeed this is the unit which in the final step has the largest residual and may be wrongly considered as an outlier from the traditional plot of residuals against fitted values (see Figure 9.2).

9.3 Detection of outliers in ordinal variables

The problem of defining and detecting outliers for ordinal variables has received scant attention in the literature. For example, the recent book of Agresti (2010) on the analysis of ordinal data does not mention this topic. Only a few papers deal with ordinal outliers in multivariate statistical methods (Zijlstra *et al.*, 2007; Pardo, 2010; Dong, 2010; Liu *et al.*, 2010). One of the reasons may be due to the difficulty of defining outliers in ordinal variables. Clearly, if we define as univariate outliers the observations that are different from the majority of the observations in a data set, for an ordinal variable corresponding to a ranking, no unit can be considered as an outlier, because the observations take on values (ranks) from 1 to n. In an ordered categorical variable with k levels, a unit may have each of k, *a priori*, defined categories and therefore no outlier could be detected. However, in a few special cases the frequency distribution of a variable may show univariate outliers. Consider, for example, the fictitious distribution of the responses of 200 customers on overall satisfaction given in Table 9.1. The two 'very unsatisfied' customers may be considered as univariate outliers, because this category of the

Table 9.1 Overall satisfaction of 200 customers.

Levels	Frequencies
Very unsatisfied	2
Unsatisfied	0
Fair	30
Satisfied	120
Very satisfied	48
	200

variable has a very low frequency and is separated from the other categories (the 'unsatisfied' label has zero frequency). If we assume that there is an underlying quantitative variable which has produced the observed ordinal categories (continuum hypothesis), the first category may be considered as coming from a different distribution. However, distributions like the one in Table 9.1 are very unusual in real situations. For example, in the ABC survey, no variable presents such a distribution. Therefore, for ordinal variables we suggest the following definition:

Definition *Bivariate and multivariate outliers for ordinal variables are those units representing an unusual combination of the categories or of the ranks of the variables.*

The previous definition of multivariate outliers can be applied both to rankings and to ordered categorical variables or to a set of nominal and ordinal variables. Notice that, similarly to what we saw in the previous section, for categorical variables, a multivariate outlier may not necessarily be a univariate outlier with respect to single ordinal categorical variables.

9.4 Detection of bivariate ordinal outliers

The joint distribution of two ordinal categorical variables, with k_1 and k_2 categories, can be presented in a $k_1 \times k_2$ contingency table. Consider, for example, the data in Table 9.2, showing the bivariate distribution of the 255 (of 266) not-missing respondents of the ABC survey with respect to the variables 'Overall satisfaction level with ABC' (question 1) and 'Overall

Table 9.2 Contingency table of the 255 non-missing respondents of questions 'Overall satisfaction level with ABC' (q1) and 'Overall satisfaction level with the equipment' (q11) in the ABC survey.

q1\q11	Very unsatisfied	Unsatisfied	Fair	Satisfied	Very satisfied	Total
Very unsatisfied	4	3	3	1	0	11
Unsatisfied	1	8	6	8	⇒2	25
Fair	0	5	40	21	1	67
Satisfied	⇒1	2	20	86	7	116
Very satisfied	0	0	3	26	7	36
Total	6	18	72	142	17	255

satisfaction level with the equipment' (question 11). The marginal distributions of the two variables show no outlier, but the pairs marked with the arrow symbol, (satisfied, very unsatisfied) with frequency 1 and (unsatisfied, very satisfied) with frequency 2, may be considered as unusual (non-coherent) combinations of the categories, that is, bivariate ordinal outliers.

Kendall's tau-b rank correlation index computed on all 255 units is equal to 0.480, with standard error 0.049. If we delete the three potential outliers, the tau-b index becomes 0.518 with a standard error of 0.044. The rank correlation without the outliers is higher and may be more appropriate to describe the association between the two variables for the majority of units.

9.5 Detection of multivariate outliers in ordinal regression

In this section, we apply a generalized linear model for ordinal variables using 'Overall satisfaction level with ABC' (q1) as response variable (see Bradlow and Zaslavsky, 1999, for an approach based on a different model for ordinal data). In Section 9.5.1, we briefly review the building blocks of ordinal regression for better understanding of our results. In Section 9.5.2 we define our model, we analyse the discrepancies between the observed and predicted values for detecting the eventual presence of atypical observations, and finally, we propose a detailed analysis of some anomalous units, or subgroups of units, for better understanding of the key features of the data set.

9.5.1 Theory

In most cases it is implausible to assume normality and homogeneity of variance for an ordered categorical outcome when (as in the case of our application) the ordinal outcome contains only a small number of discrete categories (Chu and Ghahramani, 2005). Thus, the ordinal regression model becomes the preferred modelling tool, because it does not assume normality and constant variance. The model is based on the assumption that there is a latent continuous outcome variable and that the observed ordinal outcome (which in our case is the global satisfaction index) arises from discretizing the underlying continuum into k ordered groups.

To be more precise, the ordinal regression model (McCullagh, 1980), which is nothing more than a generalized linear model (McCullagh and Nelder, 1989), may be written in the following form:

$$\text{link}(\gamma_{ij}) = \alpha_j - [b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}], \quad j = 1, \ldots, k-1 \text{ and } i = 1, \ldots, n. \tag{9.1}$$

The parameter α_j represents the threshold value of the jth category of the underlying continuous variable. The α_j terms often are not of much interest in themselves, because their values do not depend on the values of the independent variables for a particular case. They are like the intercept in a linear regression, except that each level j of the response has its own value. In our application the number of categories $k-1$ is equal to 4 (1 = very unsatisfied, 2 = unsatisfied, 3 = fair, 4 = satisfied). As usual, one of the classes of the response (in this case the category 'very satisfied') must be omitted from the model, because it is redundant.

The parameter γ_{ij} is the cumulative distribution function for the jth category of the ith case. For example, in our case γ_{i2} is the cumulative probability that the ith subject is 'very unsatisfied' or 'unsatisfied'.

In (9.1), link(γ_{ij}) is the so-called 'link function' which is typical of generalized linear models. In the case of ordinal regression, it is a transformation of the cumulative probabilities of the ordered dependent variable that allows for estimation of the model. The most commonly used links are logit, probit, negative log-log, complementary log-log and Cauchit. While the negative log-log link function is recommended when the probability of the lowest category is high, the complementary log-log is particularly suitable when the probability of the highest category is high. Finally, the logit, probit and Cauchit links assume that the underlying dependent variable respectively has a logistic, normal or Cauchy distribution. In general the Cauchit link is mainly used when extreme values are likely to be present in the data.

The number of regression coefficients is p: $b_1, \ldots, b_p$ are a set of regression coefficients and $x_{i1}, x_{i2}, \ldots, x_{ip}$ are a set of p explanatory variables for the ith subject.

With regard to the $-[b_1x_{i1} + b_2x_{i2} + \cdots + b_px_{ip}]$ term in the model, it is interesting to note two aspects. First, the negative sign before the square brackets ensures that larger coefficients indicate an association with larger scores. In our case, for example, a positive coefficient of an explanatory variable means that people who are in that class have a greater probability of showing a global level of satisfaction.

Second, the expression inside the square brackets does not depend on j. In other words, if the link is logit, for example, this implies that the effect of the independent variables is the same for all the logits. Thus the results are a set of parallel lines or hyperplanes, one for each category of the outcome variable. It is possible to test this assumption using the so-called 'test of parallel lines' against the alternative that the relationships between the independent variables and logits are not the same for all logits. If the output of the testing procedure leads us to reject the hypothesis of parallel lines, it is necessary to introduce into the model a scale component and to modify the link function as follows:

$$\text{link}(\gamma_{ij}) = \frac{\alpha_j - [b_1x_{i1} + b_2x_{i2} + \ldots + b_px_{ip}]}{\exp(\tau_1z_{i1} + \ldots + \tau_kz_{ik})}. \tag{9.2}$$

The numerator of equation (9.2) is known in the literature as the *location* component of the model, while the denominator specifies the *scale*. The $\tau_1, \tau_2, \ldots, \tau_k$ are coefficients for the scale component and the $z_{i1}, z_{i2}, \ldots, z_{ik}$ are predictor variables for the ith subject for the scale component (chosen from the same set of variables as the xs). The scale component accounts for differences in variability for different values of the predictor variables. For example, if certain groups have more variability than others in their ratings, using a scale component to account for this may improve the model.

In ordinal regression, in order to evaluate the goodness of fit of the model, the indices which are typically used are Cox and Snell's R^2 (R^2_{CS}) and Nagelkerke's pseudo-R^2(R^2_{N}) (Nagelkerke, 1991). The second index is a modification of the first which adjusts the scale of the statistic to cover the full range from 0 to 1. These two indexes are expressed as follows:

$$R^2_{\text{CS}} = 1 - \left(\frac{l(0)}{l(\hat{\beta})}\right)^{2/n}, \quad R^2_{\text{N}} = \frac{R^2_{\text{CS}}}{\max(R^2_{\text{CS}})}, \tag{9.3}$$

where $l(\hat{\beta})$ is the log-likelihood of the current model, $l(0)$ is the log-likelihood of the initial model which does not contain explanatory variables (null model), and $\max(R^2_{\text{CS}}) = 1 - \{l(0)\}^{2/n}$.

9.5.2 Results from the application

We initially considered all the 133 variables present in the ABC 2010 ACSS data set as explanatory variables. We then repeated the analysis considering only the 21 overall satisfaction level variables (questions 11, 17, 25, 31, 38, 42, 43, 49, 57, 65, 67, 70 and 73–81). Here we present only the results of this last model, since, for the purposes of our exposition, the difference between the two sets of results is not appreciable.

The data set has a large number of missing values. Thus we decided to keep in the analysis only the six variables with less than 10% missing data (questions 11, 17, 25, 38, 57, 65, i.e. overall satisfaction level with the equipment, sales support, technical support, ABC's supplies and orders, purchasing support and contracts and pricing) and 216 observations obtained with the listwise criteria (see Chapter 10 of this volume for a systematic treatment of missing values).

Figure 9.4, which shows the bar-plot of the response variable 'Overall satisfaction level' (q1), highlights that its empirical distribution is slightly negatively skewed. Therefore, as mentioned in Section 9.5.1, suitable link functions, *a priori*, are the logit and the complementary log-log. After fitting the model with these two links, we noticed that all coefficients have the expected sign, with the exception of q57. This variable was always highly non-significant, even after suppressing the observations with the most anomalous combinations with the response. Hence, we decided to remove q57 from all subsequent analyses.

With regard to the goodness of fit, the R^2 indices in equation (9.3) indicated that the model with the complementary log-log link was more satisfactory. Using this link we both tested model (9.1) with just the location component and model (9.2) with the scale component. The introduction of the scale component (in this case only q25 was significant in the scale) made Nagelkerke's pseudo-R^2 much lower than the original one, (0.499 against 0.818). We therefore decided to adopt the specification without the scale. We also investigated all the pairwise interactions among the explanatory variables; we did not find statistical evidence for their inclusion.

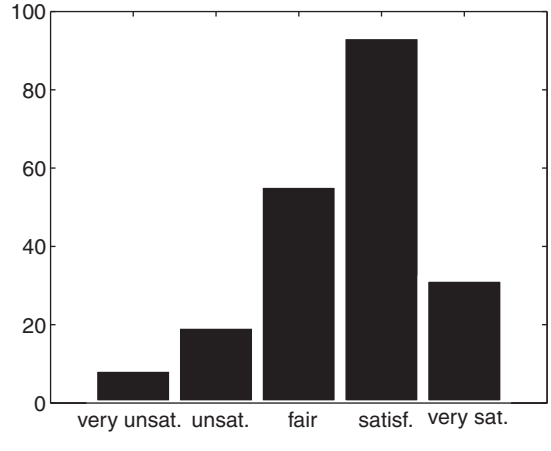

Figure 9.4 Bar-plot of the 'Overall satisfaction level' (q1): the 'satisfied' (4) category is the most frequent one.

Table 9.3 Results of the ordinal generalized linear model: list of thresholds, list of the variables in the location component, estimated coefficients and relative p-values.

	Estimated coefficients	Significance (p-value)
Thresholds		
$y = 1$ (Very unsatisfied)	1.931	0.000
$y = 2$ (Unsatisfied)	3.335	0.000
$y = 3$ (Fair)	4.979	0.000
$y = 4$ (Satisfied)	6.885	0.000
Position		
q11	0.533	0.000
q17	0.058	0.488
q25	0.528	0.000
q38	0.258	0.044
q65	0.266	0.018

The estimated coefficients of our final model together with their significance values are given in Table 9.3. All coefficients, except that for q17, are significant at the 5% level. However, from the contingency table between variables q1 and q17 (Table 9.4), we can see that there are 12 observations, marked with arrows, showing a global level of satisfaction that is different from that collected in q17 by three or more categories. If we suppress these, the p-value of q17 decreases as far as 0.001 and Nagelkerke's pseudo-R^2 becomes equal to 0.971. Therefore, q17 must be kept in the regression model.

The value of the pseudo-R^2 (0.818) in this model is very high even if there is still about 20% of variance that is left unexplained. Now we wish to evaluate the stability of the model in the presence of some units with an unusual combination of the categories, as we did before for q1 and q17, taking into account that, as happens in ordinary regression models, an observation with a large residual does not necessarily mean that it is influential with respect to the fitted equation, and vice versa. To start this check, in Table 9.5 we report the contingency table

Table 9.4 Contingency table of the 216 respondents of questions 1, 'Overall satisfaction level with ABC', and 17, 'Overall satisfaction level with the sales support' in the ABC survey.

		q17					
		Very unsatisfied (1)	Unsatisfied (2)	Fair (3)	Satisfied (4)	Very satisfied (5)	Total
q1	Very unsatisfied (1)	7	2	0	1⇐	0	10
	Unsatisfied (2)	1	9	7	0	4⇐	21
	Fair (3)	6	9	19	19	4	57
	Satisfied (4)	3⇐	13	27	37	15	95
	Very satisfied (5)	1⇐	3⇐	7	12	10	33
	Total	18	36	60	69	33	216

Table 9.5 Contingency table of 'Overall satisfaction level' (q1)' and 'Predicted class of satisfaction level' (PRE).

		Predicted class for q1 (PRE)					
		Very unsatisfied (1)	Unsatisfied (2)	Fair (3)	Satisfied (4)	Very satisfied (5)	Total
q1	Very unsatisfied (1)	2	2	5←	1⇐	0	10
	Unsatisfied (2)	1	4	10	6←	0	21
	Fair (3)	1←	1	21	34	0	57
	Satisfied (4)	0	0	9	77	9	95
	Very satisfied (5)	0	0	1←	15	17	33
	Total	4	7	46	133	26	216

of the 'Overall satisfaction level' (q1)' and the 'Predicted class of satisfaction level' by the model (PRE). This table clearly shows that the predicted values are in general very close to the observed levels of satisfaction. Most of the frequencies lie around the main diagonal of the table.

In Table 9.5, the cells where the absolute difference between predicted and observed class of satisfaction is greater than 1 are indicated with arrows. For example, in the cell with the fat open arrow, the model overestimates q1 by three points: it predicts category 4 instead of category 1. We will label this case with its position in our data subset, 163 (201 in the original data set). The reason why the model largely misclassifies this unit is partially explained in Table 9.6: the ranking assigned by the respondent to the dependent variable q1 is lower than the rankings mainly given by the same respondent to the other questions; q1 is even lower than the minimum category taken by the other variables. If we apply our ordinal regression model to the data set without this unit, Nagelkerke's pseudo-R^2 increases from 0.818 to 0.849 and the p-values of the regression parameters remain approximately invariant.

In Table 9.5, four cells are also indicated by thin arrows, where q1 differs from its predicted values by two categories: there are six cases where the model predicts 4 instead of 2, five cases where the model predicts 3 instead of 1, one case where the model predicts 3 instead of 5, and one case where the model predicts 1 instead of 3. Concerning the first six units for which $q1 = 2$ and $PRE = 4$ (units 33, 96, 101, 128, 186, 207 in the data subset, corresponding to units 39, 121, 127, 157, 229, 256 in the original data set), Table 9.7 shows that also in this case, the respondents assign to q1 a category almost always lower than or equal to the minimum assigned for all the other variables. If we now apply our regression model to the data set without unit 163 and without these six units, Nagelkerke's pseudo-R^2 increases from 0.849 to 0.918.

Table 9.6 Unit 163 ($q1 = 1, PRE = 4$): the 'Overall satisfaction level' (q1) is not in agreement with the other partial satisfaction levels.

	q1	q11	q17	q25	q38	q65
Unit 163	1	2	4	4	3	4

Table 9.7 Units 33, 96, 101, 128, 186, 207 (q1 = 2, PRE = 4). In all cases the 'Overall satisfaction level' (q1) is not in agreement with the other satisfaction levels.

	q1	q11	q17	q25	q38	q65
Unit 33	2	4	5	3	3	3
Unit 96	2	4	5	4	3	3
Unit 101	2	5	2	1	5	4
Unit 128	2	5	5	3	2	5
Unit 186	2	3	3	5	3	2
Unit 207	2	4	2	3	3	3

The same considerations about the inconsistency of the overall level of satisfaction with the partial levels are valid for the other three groups of units for which q1 = 1 and PRE = 3, q1 = 5 and PRE = 3, and q1 = 3 and PRE = 1, as is clear from Tables 9.8–9.10. If we apply our regression model to the data set without unit 163 and without respectively the units shown in Tables 9.8, 9.9 and 9.10, Nagelkerke's pseudo-R^2 increases from 0.849 to 0.883, 0.918 and 0.888. If, instead, we drop all 14 cases mentioned from the data set, Nagelkerke's pseudo-R^2 increases as far as 0.967.

The units discussed so far can be considered influential (anomalous) with respect to our ordinal regression model. For a better interpretation of these results, we can also investigate the observations with anomalous combination of the categories, leaving aside the defined model. Taking all the variables on the same level, we could, for example, compare for each observation i the 'Overall satisfaction level' q1 with the median of the other five variables, as follows:

$$d(i) = \text{q1}(i) - \text{median}[\text{q11}(i), \text{q17}(i), \text{q25}(i), \text{q38}(i), \text{q65}(i)] \tag{9.4}$$

The result of this proposal on the ABC data subset is given in Figure 9.5. The x-axis is the observation number while the y-axis plots quantity (9.4). The observations of interest are labelled with their data subset row number. As expected, for unit 163, which was the most anomalous in Table 9.5, $d(i)$ takes a very low value. Among the six units of Table 9.7, for which q1 = 2 and PRE = 4, units 128, 96 and 101 deserve a mention for the low values of $d(i)$. Of the other five units reported in Table 9.8 for which q1 = 1 and PRE = 3, units 22, 49 and

Table 9.8 Units 14, 22, 49, 70, 77 of our data subset (units 14, 25, 63, 90, 98 in the original data set): q1 = 1 and PRE = 3. In most of the cases the 'Overall satisfaction level' (q1) is not in agreement with the other satisfaction levels.

	q1	q11	q17	q25	q38	q65
Unit 14	1	3	1	2	3	2
Unit 22	1	4	1	2	3	3
Unit 49	1	3	1	4	3	1
Unit 70	1	1	1	5	3	1
Unit 77	1	1	2	4	3	3

Table 9.9 Unit 124 of our data subset (unit 152 in the original data set): q1 = 5 and PRE = 3. The 'Overall satisfaction level' (q1) is higher than the maximum of the partial satisfaction levels.

	q1	q11	q17	q25	q38	q65
Unit 124	5	3	2	4	3	2

Table 9.10 Unit 60 of our data subset (unit 79 in the original data set): q1 = 3 and PRE = 1. The 'Overall satisfaction level' (q1) is higher than the maximum of the partial satisfaction levels.

	q1	q11	q17	q25	q38	q65
Unit 60	3	2	2	1	1	1

77 take low values. Finally, the values of $d(i)$ for observations 124 and 60, for which q1 = 5 and PRE = 3 and q1 = 3 and PRE = 1, are very large.

One of the highest values reported in Figure 9.5 corresponds to unit 135 (unit 164 in the original data set), which was not identified as anomalous in Table 9.5 as it belongs to the combination q1 = 5 and PRE = 4. As clearly shown by Table 9.11, the value assigned by the respondent to the 'Overall satisfaction level' (q1) is greater than the rankings mainly given by the same respondent to the other questions. However, in our regression model, the largest coefficients are those relative to variables q11 and q25. For them unit 135 takes high categories

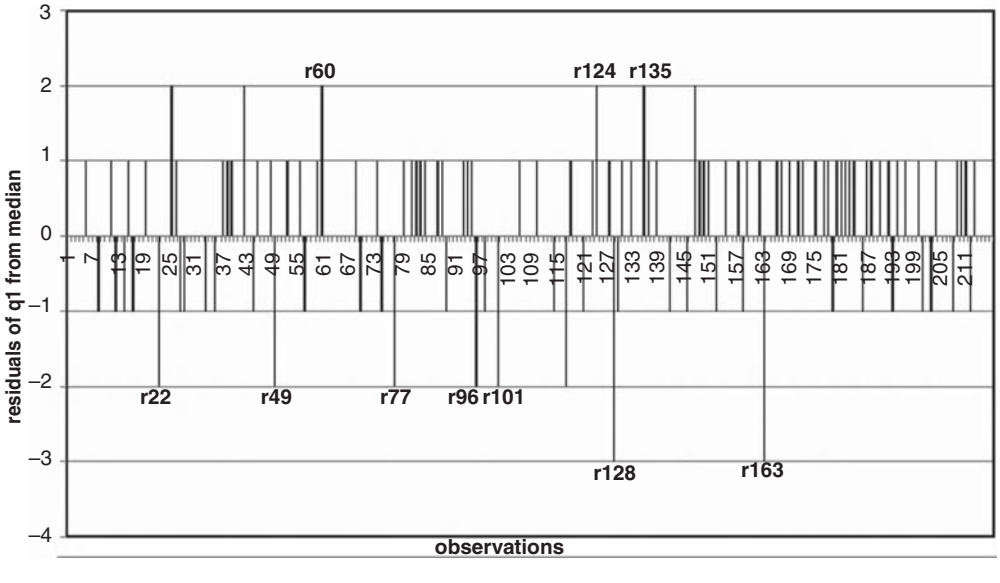

Figure 9.5 Difference between q1 and the median of the other five variables for each observation of the data subset.

Table 9.11 Unit 135 (q1 = 5, PRE = 4). Only variables q11 and q25 are in line with q1.

	q1	q11	q17	q25	q38	q65
Unit 135	5	4	1	5	2	3

(4 and 5, respectively). Therefore, while an exploratory analysis would suggest that unit 135 is anomalous, with our regression analysis such a unit appears to be coherent with the model. The same considerations could be entertained for all the other high values that are plotted, but not labelled.

Remark 1 In Section 9.3, using an exploratory data analysis approach, we defined as outliers for ordinal variables the units which presented unusual combinations of the categories. Here, using a modelling approach, we can consider as outliers the units which deviate markedly from the fitted model, that is, those with the largest residuals.

Remark 2 In this chapter we did not consider the approach of outlier detection based on robust cluster analysis (see Riani *et al*., 2010).

9.6 Summary

In this chapter we have investigated the topics of robustness and multivariate outlier detection for ordinal data. In the first part of the chapter we presented an overview of methods for outlier detection for continuous data in regression. In this context we illustrated an example of a data set with masking effects and showed the difficulties presented by these kinds of data. In the second part of the chapter, we suggested a definition of outliers in ordinal data and proposed methods for identifying them. We also recalled the ordinal regression model and applied it to the ABC 2010 ACSS data, previously cleansed of missing values. We highlighted how the presence of anomalous observations can lead to erroneous conclusions about the choice of the best model. Having chosen the final model, we analysed its stability and highlighted a series of potential anomalous/influential observations.

References

Agresti, A. (2010) *Analysis of Ordinal Categorical Data*, 2nd edition. Hoboken, NJ: John Wiley & Sons, Inc.

Atkinson, A.C. and Riani, M. (2000) *Robust Diagnostic Regression Analysis*. New York: Springer.

Atkinson, A.C., Riani, M. and Cerioli A. (2004) *Exploring Multivariate Data with the Forward Search*. New York: Springer.

Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*. Chichester: John Wiley & Sons, Ltd.

Bradlow, E.T. and Zaslavsky, A.M. (1999) A hierarchical latent variable model for ordinal data from a customer satisfaction survey with 'no answer' responses. *Journal of the American Statistical Association*, 94, 43–52.

Chu, W. and Ghahramani, Z. (2005) Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6, 1019–1041.

Dong, F. (2010) Bayesian method to detect outliers for ordinal data. *Communications in Statistics – Simulation and Computation*, 39, 1470–1484.

Grubbs, F.E. (1969) Procedures for detecting outlying observations in samples. *Technometrics* 11, 1–21.

Hadi, A.S., Imon, A. and Werner, M. (2009) Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1, 37–70.

Liu, Y., Wu, A.D. and Zumbo, B.D. (2010) The impact of outliers on Cronbach's coefficient alpha estimate of reliability: Ordinal/rating scale item responses, *Educational and Psychological Measurement*, 70, 5–21.

McCullagh, P. (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42, 109–142.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman & Hall.

Morgenthaler, S. (2007) A survey of robust statistics. *Statistical Methods and Applications*, 15, 271–293; *Statistical Methods and Applications*, 16, 171–172 (erratum).

Nagelkerke, N. J. D. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.

Pardo, M.C. (2010) High leverage points and outliers in generalized linear models for ordinal data. In C.H. Skiadas (ed.), *Advances in Data Analysis*, pp. 67–80. Boston: Birkhäuser.

Riani, M., Atkinson, A.C. and Cerioli, A. (2009) Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, 71, 447–466.

Riani, M., Cerioli, A. and Rousseeuw, P.J. (eds) (2010) Special issue on robust methods for classification and data analysis. *Advances in Data Analysis and Classification*, 4.

Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P.J. and Hubert, M. (2011) Robust statistics for outlier detection, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 73–79.

Rousseeuw, P.J. and Van Driessen, K. (2006) Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12, 29–45.

Schlatter, R. (ed.) (1975) *Hobbes's Thucydides*, translated by Thomas Hobbes. New Brunswick, NJ: Rutgers University Press.

Su, X. and Tsai, C.-L. (2011) Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1 (3), 261–268.

Zijlstra, W.P., van der Ark, L.A. and Sijtsma, K. (2007) Outlier detection in test and questionnaire data. *Multivariate Behavioral Research*, 42, 531–555.