1

Robust regression

Introduction

This chapter considers the robustness of quantile regression with respect to outliers. A small sample model presented by Anscombe (1973) together with two real data examples are analyzed. The equations are estimated by OLS and by the median regression estimator, in order to compare their behavior in the presence of outliers. The impact of an outlying observation on a selected estimator can be measured by the influence function, and its sample approximation allows to evaluate the robustness of an estimator. The difference between the influence function of the OLS and of the quantile regression estimators is discussed, together with some other diagnostic measures defined to detect outliers.

1.1 The Anscombe data and OLS

In the linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$, the realizations of the variables x_i and y_i, in a sample of size n with independent and identically distributed (i.i.d.) errors, allow to compute the $p = 2$ unknown coefficients β_0 and β_1. The ordinary least squares (OLS) estimator is the vector $\beta^T = (\beta_0, \beta_1)$ that minimizes the sum of squared errors, $\sum_{i=1,n} e_i^2 = \sum_{i=1,n} (y_i - \beta_0 - \beta_1 x_i)^2$. The minimization process yields the OLS estimators $\beta_1 = \frac{cov(x,y)}{var(x)}$ and $\beta_0 = \bar{y} - \beta_1 \bar{x}$, where $\bar{y}$ and $\bar{x}$ are the sample means. These estimators are the best linear unbiased (BLU) estimators, and OLS coincides with maximum likelihood in case of normally distributed errors. However OLS is not the sole criterion to compute the unknown vector of regression coefficients, and normality is not the unique error distribution. Other criteria are available, and they turn out to be very useful in the presence of outliers and when

Quantile Regression: Estimation and Simulation, Volume 2. Marilena Furno and Domenico Vistocco.

Companion website: www.wiley.com/go/furno/quantileregression

Table 1.1 Anscombe data sets and its modifications.

original data set						modified data set				
X_1	Y_1	Y_2	Y_3	Y_4	X_2	Y_1^*	Y_3^*	Y_4^*	X_2^*	Y_1^{**}
10	8.04	9.14	7.46	6.58	8	8.04	7.46	6.58	8	8.04
8	6.95	8.14	6.77	5.76	8	6.95	6.77	5.76	8	6.95
13	7.58	8.74	**12.74**	7.71	8	7.58	8.5	7.71	8	7.58
9	8.81	8.77	7.11	8	8	15	7.11	8	8	15
11	8.33	9.26	7.81	8.47	8	8.33	7.81	8.47	8	8.33
14	9.96	8.10	8.84	7.04	8	9.96	8.84	7.04	8	9.96
6	7.24	6.13	6.08	5.25	8	7.24	6.08	5.25	8	7.24
4	4.26	3.10	5.39	**12.5**	**19**	4.26	5.39	8.5	8	4.26
12	10.84	9.13	8.15	5.56	8	10.84	8.15	5.56	8	14.84
7	4.82	7.26	6.42	7.91	8	4.82	6.42	7.91	8	4.82
5	5.68	4.74	5.73	6.89	8	5.68	5.73	6.89	8	5.68

Note: The first six columns present the Anscombe data sets, where the numbers in bold are anomalous values. The last five columns present the modified data sets, where few observations have been altered. These observations are underlined in the table. In Y_3^*, Y_4^* and X_2^* the outliers are brought closer to the other observations. In Y_1^* and Y_1^{**} outliers are introduced in the original clean data set.

the errors are realization of non-normal distributions. The small data set in Table 1.1 allows to explore some of the drawbacks of OLS that motivate the definition of different objective functions, that is different criteria defining the estimators of β, like in the quantile and the robust regression estimators.

Anscombe (1973) builds an artificial data set comprising $n = 11$ observations of four dependent variables, Y_1, Y_2, Y_3, and Y_4, and two independent variables, X_1 and X_2. This data set is reported in the first six columns of Table 1.1, while the remaining columns modify some observations of the original variables. The variables in the first six columns define four simple linear regression models where the OLS estimate of the intercept is always equal to $\widehat{\beta}_0 = 3$ and the OLS estimated slope is always equal to $\widehat{\beta}_1 = 0.5$. These estimates are significantly different from zero, and the goodness of fit index is equal to $R^2 = 0.66$ in each of the four models. Figure 1.1 presents the plots of these models: the top-left graph shows a regression model where OLS well summarizes the data set $[Y_1\ X_1]$. In the other three models, however, the OLS estimates poorly describe the majority of the data in the sample.

In the top-right graph, the data $[Y_2\ X_1]$ follow a non-linear pattern, which is incorrectly estimated by a linear regression. Here the assumption of linearity is wrong, and the results are totally unreliable since the model is misspecified.

In the two bottom graphs, the OLS line is attracted by one anomalous value, that is by one observation that is far from the majority of the data. In the bottom-left graph, the $[Y_3\ X_1]$ data set is characterized by one observation greater than all the others with respect to the dependent variable Y_3. This is a case of one anomalous observation in the dependent variable, reported in bold in the table, where the third observation

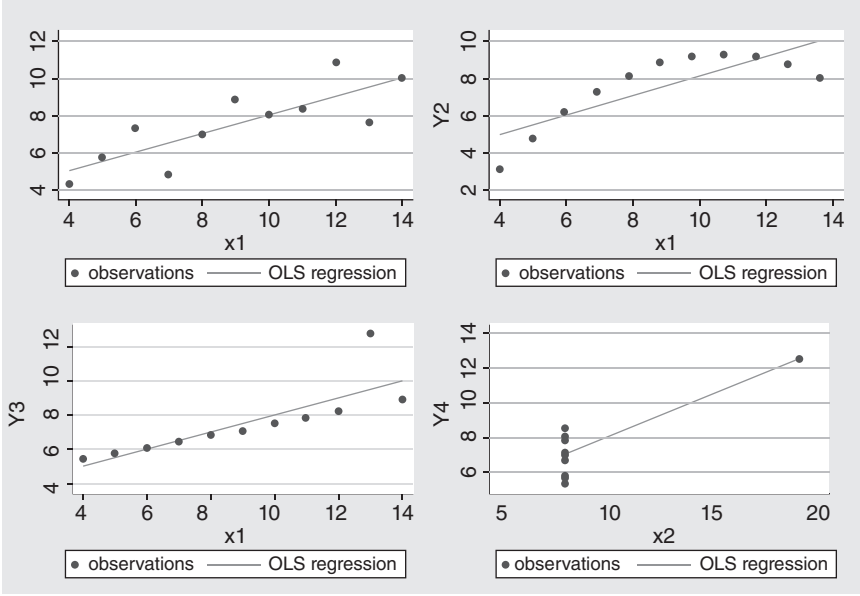

Figure 1.1 OLS estimates in the four Anscombe data sets $[Y_1\ X_1]$, $[Y_2\ X_1]$, $[Y_3\ X_1]$ and $[Y_4\ X_2]$. The four linear regressions yield the same OLS estimated coefficients, $\widehat{\beta}_0 = 3$ and $\widehat{\beta}_1 = 0.5$, and the same goodness of fit value, $R^2 = 0.66$, in all the four regressions. The sample size is $n = 11$.

$(Y_3\ X_1)_3 = (\mathbf{12.74}\ 13)$ presents the largest value of Y_3, the farthest from its median, $\text{Me}(Y_3) = 7.11$, and from its mean $\overline{Y_3} = 7.5$. In this case the outlier attracts the OLS regression, causing a larger OLS estimated slope and a smaller OLS intercept. This example shows how one outlying observation can cause bias in the OLS estimates. If this observation is replaced by an observation closer to the rest of the data, for instance by the point $(Y_3^*\ X_1)_3 = (\underline{8.5}\ 13)$ as reported in the eighth column of the table, the variance of the dependent variable drops from $\widehat{\sigma}^2(Y_3) = 4.12$ of the original series to $\widehat{\sigma}^2(Y_3^*) = 1.31$ of the modified series, all the observations are on the same line, the goodness of fit index attains its maximum value, $R^2 = 1$, and the unbiased OLS estimated coefficients are $\widehat{\beta}_0^* = 4$ and $\widehat{\beta}_1^* = 0.34$. The results for the modified data set $[Y_3^*\ X_1]$ are depicted in the top-right graph of Figure 1.2.

In the bottom-right graph of Figure 1.1, instead, OLS computes a non-existing proportionality between Y_4 and X_2. The proportionality between these two variables is driven by the presence of one outlying observation. In this case one observation, given by $(Y_4\ X_2)_8 = (\mathbf{12.5\ 19})$ and reported in bold in the table, is an outlier in both dimensions, the dependent and the explanatory variable. If the eighth observation of Y_4 is brought in line with all the other values of the dependent variable, that is if $(Y_4\ X_2)_8$ is replaced by $(Y_4^*\ X_2)_8 = (\underline{8.5}\ \mathbf{19})$ as reported in the ninth column of Table 1.1, this observation is anomalous with respect to the independent variable alone. This case, for the data set $[Y_4^*\ X_2]$, is depicted in the bottom-left graph of Figure 1.2. Even

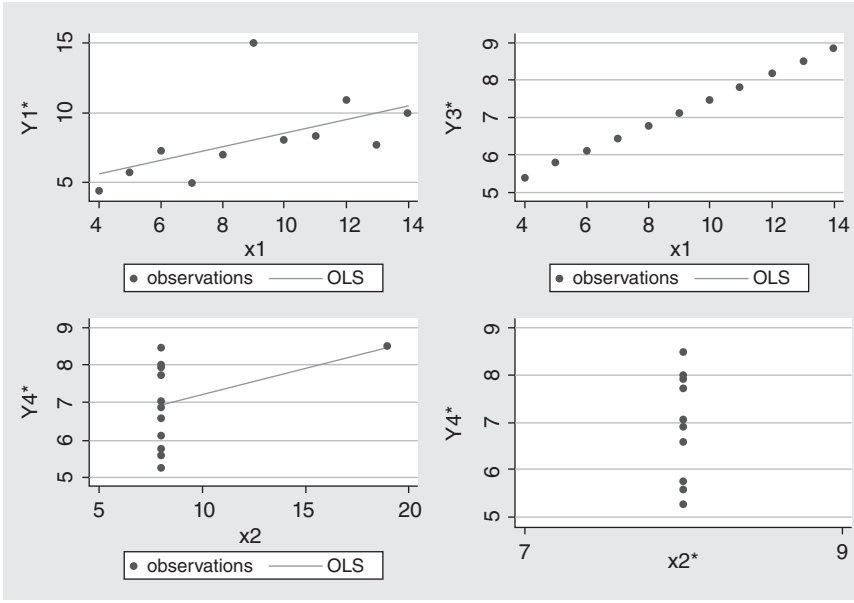

Figure 1.2 OLS estimates of the modified data sets $[Y_1^* \; X_1]$ and $[Y_3^* \; X_1]$ in the top graphs, $[Y_4^* \; X_2]$ and $[Y_4^* \; X_2^*]$ at the bottom section of the figure. The sample size is $n = 11$.

now there is no true link between the two variables; nevertheless OLS computes a non-zero slope driven by the outlying value in $(X_2)_8$. When the eighth observation $(Y_4 \, X_2)_8 = (\mathbf{12.5 \; 19})$ is replaced by $(Y_4^* \; X_2^*)_8 = (\underline{8.5} \; \underline{8})$, so that also the independent variable is brought in line with the rest of the sample, it becomes quite clear that Y_4^* does not depend on X_2^* and that the previously estimated model is meaningless, as can be seen in the bottom-right graph of Figure 1.2 for the data set $[Y_4^* \; X_2^*]$.

These examples show that there are different kinds of outliers: in the dependent variable, in the independent variable, or in both. The OLS estimator is attracted by these observations, and this causes a bias in the OLS estimated coefficients.

The bottom graphs of Figure 1.1 illustrate the impact of the so-called leverage points, which are outliers generally located on one side of the scatterplot of the data. Their sideway position enhances the attraction, and thus the bias, of the OLS estimated line. The bias can be related to the definition of the OLS estimator, which is based on the sample mean of the variables. The mean is not a robust statistic, as it is highly influenced by anomalous values, and its lack of robustness is transmitted to the OLS estimator of the regression coefficients.

There are, however, cases of non-influential outliers, i.e., of anomalous values that do not attract the OLS estimator and do not cause bias. This case is presented in the top-left panel of Figure 1.2. In this graph the data set $[Y_1 \; X_1]$ is modified to include one outlier in Y_1. In particular, by changing the fourth observation $(Y_1 \; X_1)_4 = (8.81 \; 9)$ into $(Y_1^* \; X_1)_4 = (\underline{15} \; 9)$ – as reported in the seventh column of Table 1.1 – the

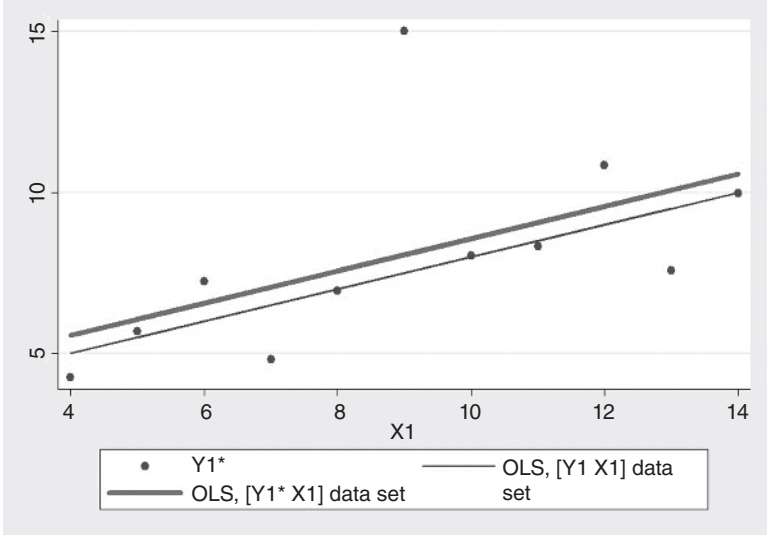

Figure 1.3 OLS estimated regressions in the presence of a non-influential outlier, as in the data set $[Y_1^* \; X_1]$ characterized by one outlier in the dependent variable. The anomalous value in Y_1^* has a central position in the graph and causes a change in the OLS intercept but not in the OLS estimated slope. The lower line is estimated in the $[Y_1 \; X_1]$ data set, without outliers. The upper line is estimated in the $[Y_1^* \; X_1]$ modified data set, with one non-influential outlier in the dependent variable affecting the intercept but not the slope in the OLS estimates. The sample size is $n = 11$.

estimated slope remains the same, $\widehat{\beta}_1^* = 0.5$, while the intercept increases to $\widehat{\beta}_0^* = 3.56$. The comparison of these OLS estimates is depicted in Figure 1.3: the lower line in this graph is estimated in the $[Y_1 \; X_1]$ data set without outlier and yields the values $\widehat{\beta}_0 = 3.0$ and $\widehat{\beta}_1 = 0.5$. The upper line is estimated in the modified data set $[Y_1^* \; X_1]$ with one outlier in the fourth observation of Y_1^*. The stability of the OLS slope in this example is linked to the location of the outlying point, which assumes a central position with respect to the explanatory variable, close to the mean of X_1. Thus a non-influential outlier is generally located at the center of the scatterplot and has an impact only on the intercept, without modifying the OLS slope.

The bias of the OLS estimator in the presence of influential outliers has prompted the definition of a wide class of estimators that, by curbing the impact of outliers, provide more reliable – robust – results. The payout of robust estimators is a reduced efficiency with respect to OLS in data sets without outlying observations. This is particularly true in case of normal error distributions, since under normality, OLS coincides with maximum likelihood and provides BLU estimators. However, in the presence of anomalous data, the idea of normally distributed errors must be discarded. Indeed the presence of outliers in a data set can be modeled by assuming non-normal errors, like Student-t, χ^2, double exponential, contaminated distributions, or any other distribution characterized by greater probability in the tails with respect to the normal

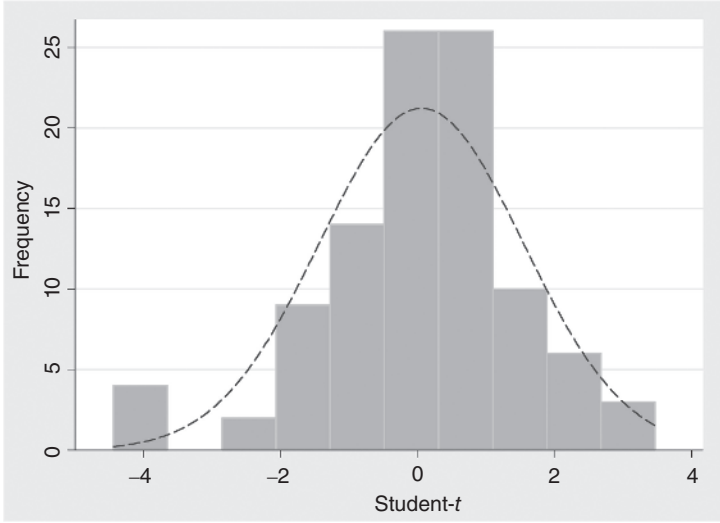

Figure 1.4 Comparison of the histogram of a Student-t with two degrees of freedom and a standard normal density, represented by the dashed line, in a sample of size $n = 50$. With respect to the standard normal, the Student-t histogram presents a small peak in the left tail, thus yielding values far from the rest of the data with frequency higher than in the standard normal case.

case. A greater probability in the tails implies a greater probability of realizations far from the center of the distribution, that is, a greater probability of outliers in the data. Figure 1.4 compares the realizations of a Student-t distribution with 2 degrees of freedom and a standard normal, represented by the dashed line, in a sample of $n = 50$ observations. The realizations of the Student-t distribution present a small peak in the left tail. This peak shows that data far from the center occur with a frequency greater than in the case of a normal density. Analogously, Figure 1.5 presents histogram of the realizations of a contaminated normal distribution f. This distribution is defined as the linear combination of two normal distributions centered on the same mean, in this example centered on zero, but having different variances. The outliers are realizations of the distribution with higher variance. In Figure 1.5 a standard normal density, f_a, generates 95% of the observations while the remaining 5% are realizations of f_b, a contaminating normal distribution having zero mean and a larger standard error, $\sigma_b = 4$. In this example the degree of contamination, i.e., the percentage of observations coming from the contaminating distribution f_b, is 5%. In a sample of size $n = 50$, this amounts to 2 or 3 anomalous data. The figure reports the histogram of $f = (1 - .05) f_a + .05 f_b$, which is more dispersed than the normal distribution, depicted by the solid line, skewed and with a small peak in the left tail. It is of course possible to have all sorts of contaminated distributions, with densities that differ not only in variance but also in shape, like in the case of the linear combination of a normal and a χ^2 or a Student-t distribution. Figure 1.6 reports the linear combination

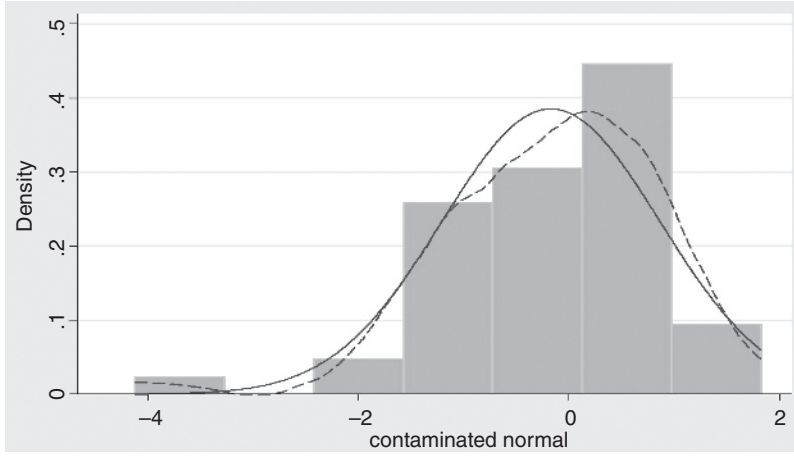

Figure 1.5 Histogram of the realizations of a contaminated normal distribution, where the observations close to the mean are generated by a standard normal while the observations in the tails, the outliers in the left tail, to be precise, are generated by a contaminating normal having zero mean and $\sigma = 4$. The degree of contamination, i.e., the percentage of observations generated by the distribution characterized by a larger variance, is 5% in a sample of size $n = 50$. The dashed line is the Epanechnikov (1969) kernel density, while the solid line is the normal density.

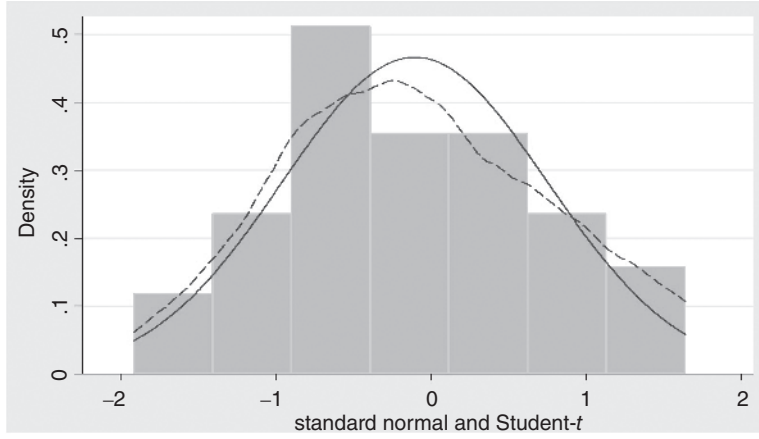

Figure 1.6 Histogram of the realizations of a standard normal distribution contaminated by a Student-t with 4 degrees of freedom. The degree of contamination is 5% in a sample of size $n = 50$. The dashed line is the Epanechnikov (1969) kernel density, which is skewed and thick tailed when compared to the solid line representing the normal density.

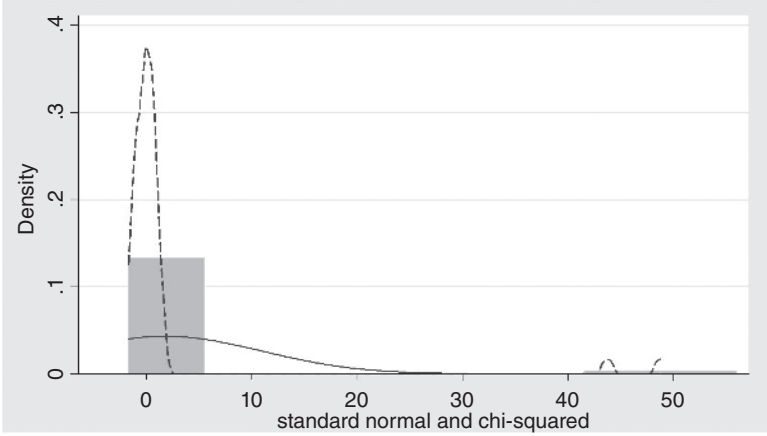

Figure 1.7 Histogram of the realizations of a standard normal distribution contaminated by a χ^2 distribution. The degree of contamination is 5%, in a sample of size $n = 50$, and contamination occurs only in the right tail. The dashed line is the Epanechnikov (1969) kernel density, characterized by two small peaks in the far right, while the solid line is the normal density.

of a standard normal and a Student-t with 4 degrees of freedom. The latter, depicted by the dashed line, is characterized by tails larger than normal, particularly on the right-hand side. Figure 1.7 presents a standard normal contaminated by a χ^2 distribution, which is a case of asymmetric contamination since it generates outliers exclusively in the right tail. The figure shows very large outliers to the right, generated by the contaminating χ^2 distribution.[1] In addition, various degrees of contamination can be considered.

1.2 The Ancombe data and quantile regression

This section considers the behavior of the quantile regression estimator in the presence of outliers, looking in particular at the median regression. The median regression can be directly compared to OLS, which in turn coincides with the conditional mean regression. Table 1.2 reports the OLS and the median regression estimates for the four different Anscombe models. The first two top columns present the model $[Y_1 \; X_1]$ where the median regression estimated coefficients are very close to the OLS values. In this case the median regression coefficients are within the 95% OLS confidence intervals: $0.75 < \widehat{\beta}_0(.5) = 3.24 < 5.25$ and $0.26 < \widehat{\beta}_1(.5) = 0.48 < 0.74$. The weakness of quantile regression is in the standard errors, which at the median are about twice the OLS standard errors, $se(\widehat{\beta}_0(.5)) = 2.043 > se(\widehat{\beta}_0)_{OLS} = 1.125$

[1] An example of asymmetric contamination, and more in general of asymmetric behavior, is provided by variables like market prices, which rarely decrease, or stocks, which rapidly increase during an economic crisis.

Table 1.2 Comparison of OLS and median regression results in the four Anscombe data sets.

	Y_1 median	Y_1 OLS	Y_2 median	Y_2 OLS	Y_3 median	Y_3 OLS
X_1	0.480	0.500	0.500	0.500	0.345	0.500
se	(.199)	(.118)	(.197)	(.118)	(.001)	(.118)
intercept	3.240	3.000	3.130	3.001	4.010	3.002
se	(2.043)	(1.125)	(1.874)	(1.125)	(.007)	(1.124)
	Y_4 median	Y_4 OLS				
X_2	0.510	0.508				
se	(.070)	(.107)				
intercept	2.810	2.857				
se	(1.323)	(1.019)				

Note: Standard errors in parenthesis, sample size $n = 11$.

and $se(\widehat{\beta}_1(.5)) = 0.199 > se(\widehat{\beta}_1)_{OLS} = 0.118$. This occurs since, as mentioned, under normality, OLS is the BLUE and thus more efficient than the quantile regression estimator.

Model $[Y_2\ X_1]$ is incorrect in both OLS and median regression since the true equation is non-linear and any attempt to describe these observations by a linear model causes misspecification. The second pair of columns in the top section of Table 1.2 present these results, with OLS and median regression providing very similar estimates. The top-right graph in Figure 1.1 presents this data set together with the OLS estimated regression, which coincides with the median regression.

The $[Y_3\ X_1]$ data set is characterized by one influential outlier in the dependent variable. Here the median regression improves upon OLS since the median fitted line is not attracted by the anomalous value in the third observation of Y_3. Furthermore, the precision of the median regression estimator improves upon OLS, and the standard errors of the quantile regression coefficients are much smaller than the OLS analogues: $se(\widehat{\beta}_0(.5)) = 0.007 < se(\widehat{\beta}_0)_{OLS} = 1.124$ and $se(\widehat{\beta}_1(.5)) = 0.001 < se(\widehat{\beta}_1)_{OLS} = 0.118$. The right-hand side graph in Figure 1.8 presents the behavior of the median regression in this case, showing that the estimated slope is unaffected by the anomalous value and that the estimated median regression well represents the majority of the data.

The case of outliers in both dependent and independent variables, analyzed in the data set $[Y_4\ X_2]$, yields median regression results comparable to OLS, and quantile regression in this case is just as unreliable as OLS. This can be seen in the two bottom columns of Table 1.2 and in the left plot of Figure 1.9, where the presence of one outlier in $(Y_4\ X_2)_8$ yields very close results in the OLS and the median regression: none of them is robust with respect to outliers in the explanatory variables.

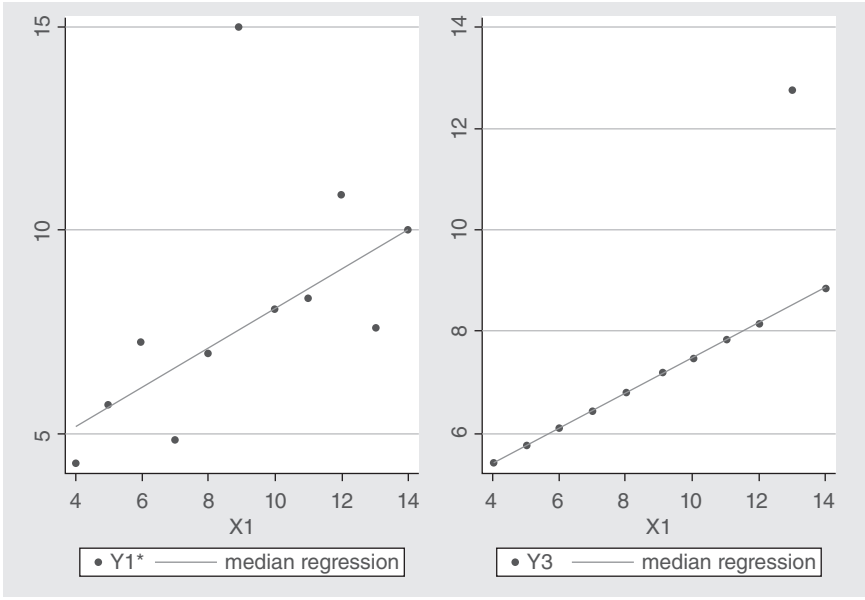

Figure 1.8 Median regression estimates in the presence of outliers. In the left graph is the $[Y_1^* \; X_1]$ modified data set with one outlier in the dependent variable placed in a central position. The right panel depicts the $[Y_3 \; X_1]$ data set with one outlier located on the side. In both cases there is no impact of the outlier on the coefficients as estimated by the median regression. The OLS estimates, instead, present an increase in the intercept in the $[Y_1^* \; X_1]$ data set, as shown in Figure 1.3, and an increased slope in the $[Y_3 \; X_1]$ data set, as can be seen in the bottom left panel of Figure 1.1. Sample size $n = 11$.

Summarizing, the median regression is robust with respect to outliers in the dependent variable, as in $[Y_3 \; X_1]$, but is not robust to outliers in the explanatory variables, as in $[Y_4 \; X_2]$. More insights can be gained by the analysis of the modified data sets.

In $[Y_1^* \; X_1]$ the location of the outlier in $(Y_1^* \; X_1)_4$ is close to the mean of the independent variable X_1, i.e., the outlier is in a central position and is not a leverage point. The first two top columns of Table 1.3 present the estimated coefficients. This is a case of one non-influential outlier, where its central position modifies only the OLS intercept. Indeed in Figure 1.3 the OLS fitted line shifts upward but yields an unchanged slope. This can be compared with the left-hand side graph of Figure 1.8, which shows the same median regression in the two data sets, $[Y_1 \; X_1]$ and $[Y_1^* \; X_1]$. The second column in the top section of Table 1.3 shows that the estimated line does not shift at all, yielding the same intercept and the same slope of Table 1.2, $\widehat{\beta}_0(.5) = \widehat{\beta}_0^*(.5) = 3.24$ and $\widehat{\beta}_1(.5) = \widehat{\beta}_1^*(.5) = 0.48$. In addition, the median regression standard errors are slightly smaller than in the OLS case, $se(\widehat{\beta}_0^*)_{OLS} = 2.441 > se(\widehat{\beta}_0^*(.5)) = 2.043$ and $se(\widehat{\beta}_1^*)_{OLS} = 0.256 > se(\widehat{\beta}_1^*(.5)) = 0.199$, showing that one non-influential outlier worsens the precision of the OLS estimator.

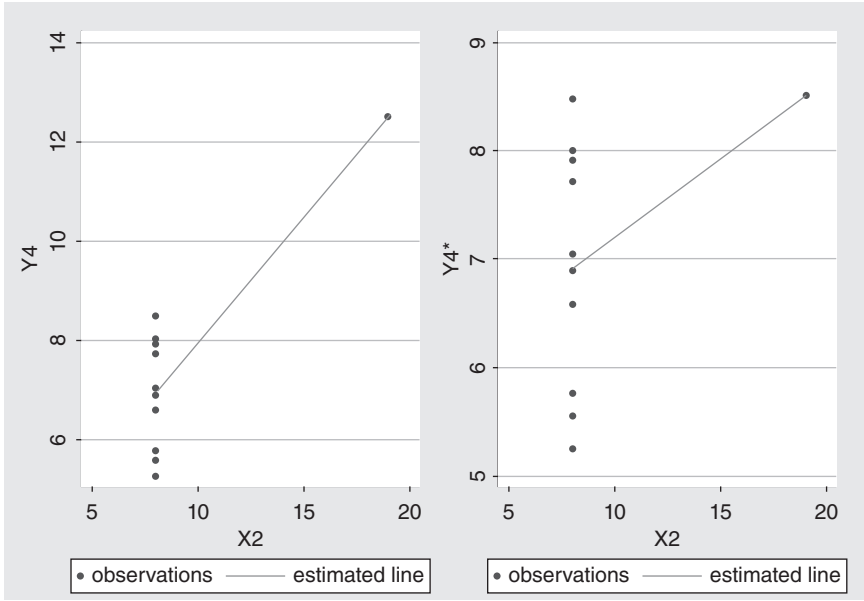

Figure 1.9 OLS and median regression estimates in the $[Y_4\ X_2]$ and $[Y_4^*\ X_2]$ Anscombe data sets overlap. In the left plot there is one outlier in both the dependent and the independent variable, while the right plot presents one outlier only in the explanatory variable X_2. Sample size $n = 11$.

Table 1.3 Comparison of OLS and median regression results in the modified Anscombe data sets.

	Y_1^* OLS	Y_1, Y_1^* median	Y_3^* OLS	Y_3, Y_3^* median	Y_4^* OLS	Y_4^* median
X_1	0.500	0.480	0.346	0.345		
se	(0.256)	(0.199)	(0.0003)	(0.001)		
X_2					0.144	0.146
se					(0.107)	(0.070)
intercept	3.563	3.240	4.005	4.010	5.765	5.719
se	(2.441)	(2.043)	(0.003)	(0.007)	(1.019)	(1.323)

	Y_4^* OLS	Y_4^* median
X_2^*	0 (omitted)	0 (omitted)
se	0	0
intercept	7.061	7.04
se	(.351)	(.628)

Note: Standard errors in parentheses, sample size $n = 11$.

The $[Y_3\ X_1]$ data set is characterized by one influential outlier in the dependent variable that affects the OLS estimates but not the median regression results, as can be seen in the right plot of Figure 1.8. When the outlier is removed, as in the $[Y_3^*\ X_1]$ data set, the OLS results coincide with the median regression estimates, as reported in the second pair of columns in the top section of Table 1.3. Once accounted for the anomalous value in the third observation of Y_3, the standard errors are equal or very close to zero since all the observations are on the same line in the $[Y_3^*\ X_1]$ data set, as can be seen in the top-right graph of Figure 1.2.

The $[Y_4\ X_2]$ data set has been modified in two steps, first bringing the outlier closer to the rest of the data in the dependent variable, so that the $[Y_4^*\ X_2]$ modified data set presents one outlier only in the independent variable. The estimated coefficients are $\widehat{\beta}^*_{0,OLS} = \widehat{\beta}^*_0(.5) = 5.7$ and $\widehat{\beta}^*_{1,OLS} = \widehat{\beta}^*_1(.5) = 0.14$, as reported in the last two top columns of Table 1.3 and in the right graph of Figure 1.9. In both the $[Y_4\ X_2]$ and $[Y_4^*\ X_2]$ data sets the OLS and the median regression estimates are very similar, since the quantile regression estimator is not robust with respect to outliers in the explanatory variables just as OLS. Finally, in the data set $[Y_4^*\ X_2^*]$, where also the outlier in X_2 is brought in line with the other observations, there is no estimated slope at all and only an estimated intercept, as reported in the two columns at the bottom of Table 1.3 and as depicted in the bottom-right graph of Figure 1.2. The intercept computes the selected quantile of the dependent variable, the conditional median and, in the OLS case, the conditional mean.

Summarizing:

- under normality, as in the data set $[Y_1\ X_1]$, the OLS estimator is BLU and provides the smallest variance;
- in case of non-normality with outliers in the dependent variable, as in $[Y_3\ X_1]$ and $[Y_1^*\ X_1]$, the quantile regression estimator is unbiased and has smaller variance;
- in case of incorrect model specification, as in $[Y_2\ X_1]$, and in case of outliers in the explanatory variables, as in $[Y_4\ X_2]$ and $[Y_4^*\ X_2]$, both OLS and quantile regression provide unreliable results.

1.2.1 Real data examples: the French data

This section considers a real data set example provided by the 2004 Survey of Health, Ageing and Retirement in Europe (SHARE).[2] SHARE samples the population aged 50 or above in 11 European countries. The survey involved 19286 households and 32022 individuals, covering a wide range of topics, such as physical health, socioeconomic status, and social interactions. The data set is quite large, but focusing on one country at a time and on a specific issue, the sample size becomes more manageable and apt to compare the behavior of OLS and median regression in the presence of outliers. The focus is on body mass index as predictor of walking speed. Pooling together all the 11 European countries of the SHARE data set, the link between walking speed and body mass index is negative and statistically relevant. When the

[2] The website to download the SHARE data set is http://www.share-project.org/

model is individually estimated in each country, only some of them yield statistically relevant negative slopes.

Consider for instance French data on walking speed, *wspeed*, and body mass index, *BMI*, which provide a subsample of $n = 393$ observations, presented in Figure 1.10. The graph shows the presence of a large walking speed value coinciding with a value of the *BMI* variable that characterizes decidedly overweight individuals. This observation describes an individual with $BMI = 36.62851$ and $wspeed = 3.333333$, where $BMI > 30$ signals obesity and the median walking speed in this sample is $\text{Me}(wspeed) = 0.62$. This observation is highly suspicious and in need of further scrutiny. It provides an example of one anomalous value in the dependent variable, similar to the Anscombe $[Y_3\ X_1]$ example.

Next consider a linear regression model relating walking speed to the *BMI* values, where it is reasonable to expect an inverse proportionality between these two variables. Figure 1.11 presents the OLS and the median regression, showing that the two estimated slopes present a discrepancy: $\widehat{\beta}_{1,OLS} = 0.0033$ and $\widehat{\beta}_1(.5) = -0.00056$. The OLS line is attracted by the large anomalous value of *wspeed* located to the top right of the scatterplot, and the OLS slope becomes positive. This observation is highly influential. It can be brought in line with the rest of the sample by replacing it with the median value of walking speed, given by $\text{Me}(wspeed) = 0.62$. Actually, since the sample size is large and the observation is possibly a coding error, it could also be simply discarded. The OLS regression implemented in the cleaned sample, i.e., computed by replacing the outlier 3.333 with the median value of the variable 0.62,

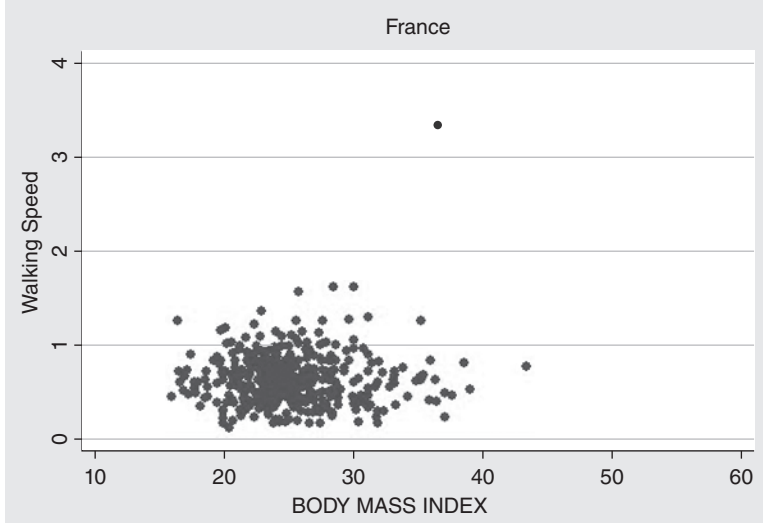

Figure 1.10 French data on walking speed and body mass index in a sample of size $n = 393$. This data set presents one outlier in the dependent variable describing an individual with the fastest walking speed of the sample and with a quite large *BMI* value, well inside the obesity range. This makes the observation quite suspicious, leading to consider it as a possible coding error.

Figure 1.11 OLS and median regression in the French data set, sample size $n = 393$. The two regressions do not have the same slope: OLS yields a positive estimate since the upper line is attracted by the influential outlier in the dependent variable. The median regression, due to its robustness to outliers in the dependent variable, yields a negative slope.

yields the estimated slope $\widehat{\beta}^*_{1,OLS} = -0.0009$, bringing the OLS results in line with the median regression computed in the original sample, as can be seen in Figure 1.12. The estimated slope at the median is $\widehat{\beta}_1(.5) = \widehat{\beta}^*_1(.5) = -0.00056$ in both the original and the cleaned sample. The estimated median lines coincide, and the outlier in *wspeed* does not modify the median regression estimated slope. This is the case since quantile regressions are robust to outliers in the dependent variable.

The final results show that in the French data set there is a small negative link between walking speed and body mass index.

1.2.2 The Netherlands example

Next consider the same regression model, walking speed as a function of *BMI*, computed for the Dutch data, in a subset of size 307. The scatterplot in Figure 1.13 presents at least one outlier in the explanatory variable, with one individual having $BMI = 73.39$ and a reasonably low walking speed. This is not a suspicious observation, but it could be influential and thus dangerous for the reliability of the estimated coefficients. The OLS slope estimated in the original sample, of size $n = 307$, does not really differ from the median regression estimated slope, with $\widehat{\beta}_{1,OLS} = -0.0058$ and $\widehat{\beta}_1(.5) = -0.0065$. The OLS and the median regression are depicted in Figure 1.14.

If the outlying value in *BMI*, equal to 73.39, is replaced with the median value of the variable, $\text{Me}(BMI) = 25.60$, the estimated coefficients do not change much, yielding $\widehat{\beta}^*_{1,OLS} = -0.0054$ and $\widehat{\beta}^*_1(.5) = -0.0063$. Figure 1.15 reports the estimated

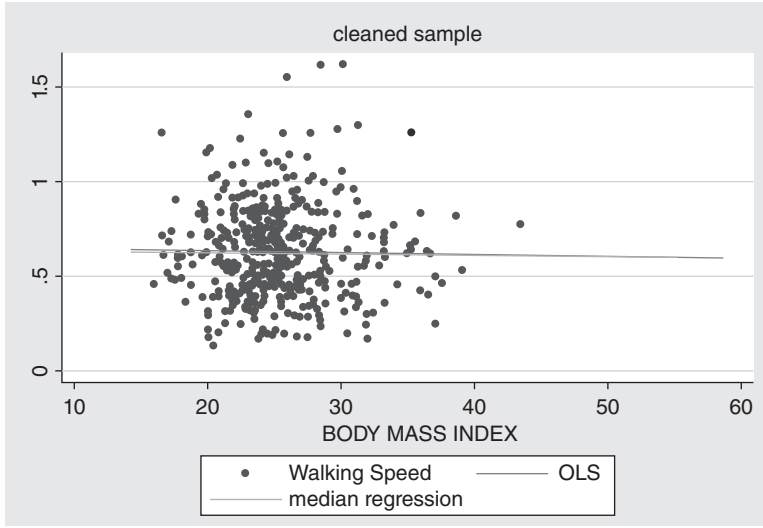

Figure 1.12 OLS and median regression in the cleaned French data set, $n = 393$. The data are cleaned by replacing the outlier with the median value of the *wspeed* variable. The OLS estimated slope becomes small and negative, just as in the median regression estimated in the original, not cleaned sample.

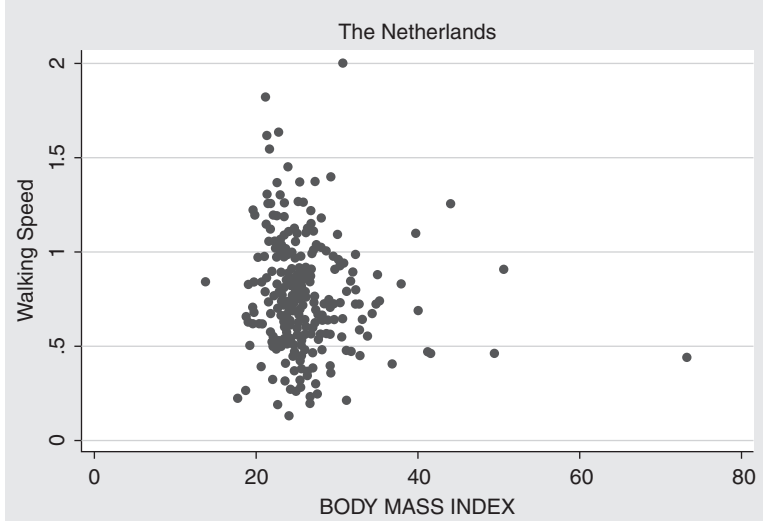

Figure 1.13 The Netherlands data set, sample of size $n = 307$. The data present one very large value of *BMI* and provide an example of at least one very large outlier in the explanatory variable.

Figure 1.14 Dutch data, OLS and median regression, sample size $n = 307$. The OLS and the median regression estimated coefficients do not significantly differ from one another.

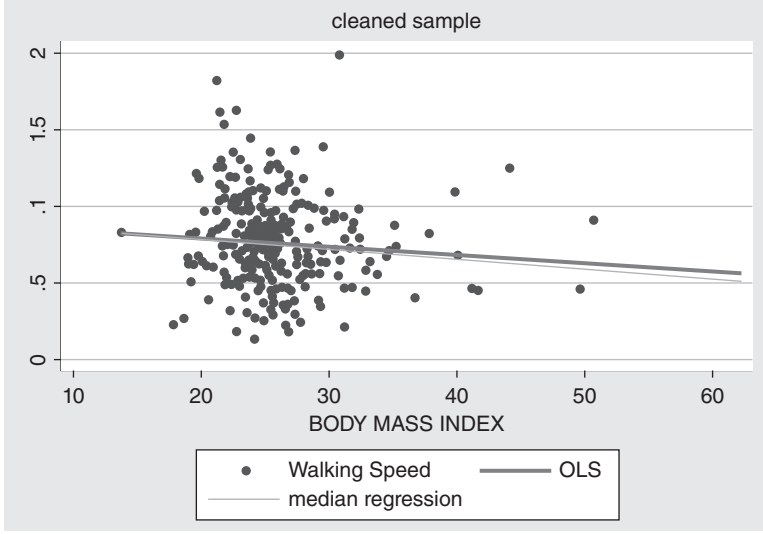

Figure 1.15 Dutch cleaned data set, OLS and median regression. The data are cleaned by replacing the outlier with the median of the explanatory variable *BMI*. The OLS regression estimates in the cleaned data set does not differ much from the median regression estimates computed in the original sample.

lines for the cleaned sample. The results are not very different between the original and the cleaned sample, and the outlier can be considered as non-influential.

In the Netherlands example the link between walking speed and *BMI* is small, negative, and statistically different from zero in the original sample.

1.3 The influence function and the diagnostic tools

In the previous sections, the robustness of quantile regression and of OLS has been analyzed by looking at their behavior in data sets where the position of one outlier and its impact on the estimates could be easily spotted. More generally, the appropriate tool to analyze the robustness of an estimator is provided by the influence function, IF, introduced by Hampel (1974). The IF measures the impact of one observation on the selected estimator, in a sample generated by a given distribution F_a. Consider next Δx_i, which assigns probability mass one to the additional point x_i. The IF measures the impact of x_i on the statistic T. It compares T evaluated at the contaminated distribution $(1-t)F_a + t\Delta x_i$ with the same statistic T evaluated at the uncontaminated F_a. The IF is given by

$$\text{IF}(x_i, T, F_a) = lim_{t\to 0} \ \frac{T((1-t)F_a + t\Delta x_i) - T(F_a)}{t}$$

A large IF signals that the addition of x_i greatly modifies the estimator and T is not robust. Vice versa, a small IF shows a negligible impact of x_i on T thus signaling that T is a robust estimator.

In the case of a linear regression model $y_i = \mathbf{x}_i^T\beta + e_i$, where $\mathbf{x}_i^T$ is the p-row vector comprising the ith observation for all the p explanatory variables, the IF is given by the ratio between the first-order condition defining the estimator and the expectation of the second-order condition. In the simple regression model with $p = 2$ and $\mathbf{x}_i^T = \begin{bmatrix}1 & x_i\end{bmatrix}$, OLS is characterized by the following IF

$$\text{IF}_{OLS} = \frac{e_i\mathbf{x}_i}{E(\mathbf{x}_i\mathbf{x}_i^T)}$$

The IF_{OLS} can be split in two different factors, the scalar term of the Influence of Residuals, $IR_{OLS} = e_i$, and the Influence of Position in the factor space, $IP_{OLS} = \frac{\mathbf{x}_i}{E(\mathbf{x}_i\mathbf{x}_i^T)}$. The former singles out the impact on the estimator of each regression residual, while the latter shows the impact of the position of the ith observation of the explanatory variables:

$$\text{IF}_{OLS} = e_i\frac{\mathbf{x}_i}{E(\mathbf{x}_i\mathbf{x}_i^T)} = IR_{OLS}IP_{OLS}$$

These two components show that a large e_i increases by the same amount IR_{OLS} and that an outlier in the explanatory variables increases the IP_{OLS} term. Thus the OLS estimator is not robust to outliers in the dependent variable, as mirrored by IR_{OLS}, nor to outliers in the explanatory variables as measured by IP_{OLS}.

For the quantile regression, the IF is defined as

$$\text{IF}_\theta = \frac{\psi(e_i)\mathbf{x}_i}{\int f(x_i^T\widehat{\beta}(\theta))\mathbf{x}_i\mathbf{x}_i^T dG(x)} = \psi(e_i)\mathbf{x}_i Q^{-1} = IR_\theta\, IP_\theta$$

where $Q = \int f(x_i^T\widehat{\beta}(\theta))\mathbf{x}_i\mathbf{x}_i^T dG(x)$. In this equation $IR_\theta = \psi(e_i) = (\theta - 1(e_i < 0))$, while $IP_\theta = \mathbf{x}_i Q^{-1}$. The most important difference between IF_{OLS} and IF_θ is in the *IR* term, in the treatment of the residuals. In OLS any large value of $\widehat{e}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ has an equal impact on IF_{OLS}, since the IR_{OLS} term is free to assume any value. In quantile regression, the impact of a residual on the estimator depends upon its sign, positive or negative according to its position above or below the estimated line, and not to its value measuring the distance between the observation and the estimated line. A very large e_i does not influence much the quantile regression estimator since only its position with respect to the estimated equation matters. The different behavior of *IR* is what makes the difference between the OLS and the quantile regression results in the Anscombe example for the $[Y_3\ X_1]$, the $[Y_1^*\ X_1]$, and in the French data sets. For instance in the $[Y_3\ X_1]$ sample, the IF for the outlying observation $(Y_3\ X_1)_3$ is function of $IR_{OLS,3} = \widehat{e}_3 = 3.241$, where $\widehat{e}_i$ is the OLS residual. The latter is larger than $IR_{\theta,3} = \psi(\tilde{e}_3) = \psi(4.245) = \theta$, where $\tilde{e}_i$ is the residual from the median regression and $\theta = .5$. Analogously in the $[Y_1^*\ X_1]$ modified Anscombe data set, characterized by one outlier in the fourth observation of the dependent variable, the $IR_{OLS,4} = \widehat{e}_4 = 6.936 >> IR_{\theta,4} = \psi(\tilde{e}_4) = \psi(7.44) = \theta = .5$. Finally, in the French data, the OLS residual for the outlying observation yields $IR_{OLS,i} = \widehat{e}_i = 2.66$, while for the median regression it is $IR_{\theta,i} = \psi(\tilde{e}_i) = \psi(2.72) = \theta = .5$. At the median IR_θ is always equal to $\pm.5$, while at a selected quantile θ it is $IR_\theta = \theta - 1(e_i < 0)$.[3]

It is worth stressing that the value of the quantile regression residuals for the influential outliers, the third residual in $[Y_3\ X_1]$, the fourth in $[Y_1^*\ X_1]$, and the residual linked to the large walking speed value in the French data are actually larger than their OLS analogues. This occurs because the OLS residuals are not a good diagnostic tool to detect outliers. When the OLS estimated line is attracted by an influential outlier, the corresponding residual is smaller than it should be. Looking at the ith OLS residual and replacing $\widehat{\beta}_0$ and $\widehat{\beta}_1$ by the OLS estimators – both of which depend on the non-robust sample mean – the OLS residual can be written as (Huber, 1981)

$$\begin{aligned}
\widehat{e}_i &= y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i = y_i - \overline{y} + \widehat{\beta}_1\overline{x} - \widehat{\beta}_1 x_i = y_i - \overline{y} - \widehat{\beta}_1(x_i - \overline{x}) \\
&= y_i - \overline{y} - \frac{cov(x,y)}{var(x)}(x_i - \overline{x}) \\
&= y_i - \overline{y} - \frac{(x_i - \overline{x})^2(y_i - \overline{y}) + \sum_{j\neq i}(x_j - \overline{x})(x_i - \overline{x})(y_j - \overline{y})}{\sum (x_i - \overline{x})^2} \\
&= \left[1 - \frac{(x_i - \overline{x})^2}{\sum (x_i - \overline{x})^2}\right](y_i - \overline{y}) - \frac{\sum_{j\neq i}(x_j - \overline{x})(x_i - \overline{x})(y_j - \overline{y})}{\sum (x_i - \overline{x})^2}
\end{aligned}$$

The last equality shows that the ith OLS residual is small when the term $\left[1 - \frac{(x_i-\overline{x})^2}{\sum (x_i-\overline{x})^2}\right] = 0$. This occurs when $\frac{(x_i-\overline{x})^2}{\sum (x_i-\overline{x})^2} \simeq 1$, that is when x_i is very far from

[3] Since $0 < \theta < 1$, its value can slightly curb the quantile regression IP_θ.

its sample mean $\bar{x}$, and its deviation from the mean greatly contributes to the total deviation of X. When this is the case, the numerator, $(x_i - \bar{x})^2$ is approximately equal to the denominator, which represents the dispersion of the entire X. Thus the farther is x_i from its mean, the smaller is the OLS residual, and the OLS fitted line is attracted by the influential outlier. This is the reason why it is not advisable to consider the OLS residuals to detect outliers, but it is preferable to use alternative diagnostic measures.

Various diagnostic tools have been defined in the literature. In order to find outliers in the explanatory variables, the key statistic is $h_{ii} = x_i^T (X^T X)^{-1} x_i$, which is closely related to the influence of position *IP*. To detect outliers in both dependent and independent variables, the appropriate statistic is given by the standardized residuals, defined as $std(e_i) = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$. Alternatively, the studentized residuals can be considered, given by $stu(e_i) = \frac{e_i}{\sigma_{(i)}\sqrt{1-h_{ii}}}$. The difference between $std(e_i)$ and $stu(e_i)$ is in the denominator, where $\sigma_{(i)}$ is computed after the exclusion of the i^{th} observation.[4] These two statistics are distributed respectively as standard normal and Student-t distribution so that a formal test on each observation could be implemented. The test verifies if each data point is or is not a rightful observation, respectively under the null and the alternative hypothesis. However, it suffices to scrutinize only the very large standardized or studentized residual.[5]

Back to the examples, the *IP* term in the influence function of the quantile regression estimators explains why this estimator is not robust to outliers in the explanatory variables: indeed there is no element in $IP_\theta = \mathbf{x}_i Q^{-1}$ to bound outliers in the explanatory variables. This is the reason why OLS and median regression provide similar results in the Anscombe model $[Y_4\ X_2]$ and in its transformed version $[Y_4^*\ X_2]$. Looking at the h_{ii} values in both data sets, all the observations in X_2 yield $h_{11} = 0.1$ but the eighth, which yields $h_{88} = 1$. The large value in the explanatory variable yields similar results in both the OLS and the median regression estimators since *IP* is not bounded.

One simple way to approximate the IF is provided by the change in the estimated coefficients caused by the exclusion of one observation. This involves the comparison of $\widehat{\beta}$ and $\widehat{\beta}_{(i)}$, where the subscript in parenthesis signals that the regression coefficient has been estimated in a sample of size $n-1$, where the ith observation has been excluded. For instance, in the Anscombe data set $[Y_3\ X_1]$, the exclusion of the largest observation, the third one given by $(Y_3\ X_1)_3$=(**12.74** $13)_3$, yields the OLS estimated slope $\widehat{\beta}_{1(3)} = 0.34$ and the difference is $\widehat{\beta}_1 - \widehat{\beta}_{1(3)} = 0.5 - 0.34 = 0.16$. When instead the smallest observation is excluded from the sample, like the data point in position eight, which is equal to $(Y_3\ X_1)_8$=$(5.39\ 4)_8$, the OLS slope is $\widehat{\beta}_{1(8)} = 0.52$ and $\widehat{\beta}_1 - \widehat{\beta}_{1(8)} = 0.5 - 0.52 = -0.02$. This observation has a much smaller impact than the third one on the OLS estimates, i.e., it is less influential.

[4] A subscript within parenthesis implies that the particular observation, or set of observations, in parenthesis has been excluded when computing the statistic under analysis.

[5] For further references on outlier detection, see Chatterjee and Hadi (1988) or Hoaglin, Mosteller, and Tukey (1983).

In the median regression, the drop of the third observation does not cause any change, $\widehat{\beta}_1(.5) - \widehat{\beta}_{1(3)}(.5) = 0$, and the exclusion of the smallest observation yields $\widehat{\beta}_{1(8)}(.5) = .346$ with $\widehat{\beta}_1(.5) - \widehat{\beta}_{1(8)}(.5) = .345 - .346 = -.001$, which is quite small. The change in coefficient, $\widehat{\beta} - \widehat{\beta}_{(i)}$, allows to approximate the IF and to compute the Sample Influence Function, SIF, defined as SIF=$(n-1)(\widehat{\beta} - \widehat{\beta}_{(i)})$. For the third observation of the $[Y_3\ X_1]$ data set, $\text{SIF}(\beta_{1,OLS})_3 = (n-1)\,(\widehat{\beta}_1 - \widehat{\beta}_{1(3)}) = 1.6$ while for the eighth one, $\text{SIF}(\beta_{1,OLS})_8 = (n-1)\,(\widehat{\beta}_1 - \widehat{\beta}_{1(8)}) = -0.2$.

In the Anscombe $[Y_4\ X_2]$ data set, the exclusion of the eighth outlying observation from the sample yields the estimated slope $\widehat{\beta}_{1(8)} = \widehat{\beta}_{1(8)}(.5) = 0$ in both OLS and the median regression. The impact of this observation is quite large, since its presence causes the slope coefficient to move from a positive value to zero, $\widehat{\beta}_1 - \widehat{\beta}_{1(8)} = 0.5 - 0 = 0.5$. For the intercept, the SIF grows even further, since $\widehat{\beta}_0 = \widehat{\beta}_0(.5) = 2.8$ in the full sample turns into $\widehat{\beta}_{0(8)} = \widehat{\beta}_{0(8)}(.5) = 6.9$ in the smaller sample, yielding the difference $\widehat{\beta}_0 - \widehat{\beta}_{0(8)} = -4.1$. The final value of the sample influence function is given by $\text{SIF}(\beta_{1,OLS})_8 = \text{SIF}(\beta_1(.5))_8 = (n-1)\,(\widehat{\beta}_1 - \widehat{\beta}_{1(8)}) = 5$ and $\text{SIF}(\beta_{0,OLS})_8 = \text{SIF}(\beta_0(.5))_8 = (n-1)(\widehat{\beta}_0 - \widehat{\beta}_{0(8)}) = -41$ in both OLS and median regression estimators. This signals the eighth observation as highly influential.

In the $[Y_4^*\ X_2]$ data set, since the outlier is anomalous only in the independent variable, the change in the estimated coefficients due to the exclusion of the eighth observation is smaller than in the $[Y_4\ X_2]$ case. The estimated slope from $\widehat{\beta}_1 = \widehat{\beta}_1(.5) = 0.14$ turns into $\widehat{\beta}_{1(8)} = \widehat{\beta}_{1(8)}(.5) = 0$ in both OLS and median regression. The impact of the anomalous value is $\widehat{\beta}_1 - \widehat{\beta}_{1(8)} = 0.14$ for the slope, and $\widehat{\beta}_0 - \widehat{\beta}_{0(8)} = 5.7 - 6.9 = -1.2$ for the intercept in both OLS and median regression estimators. In this case the sample influence function is $\text{SIF}(\beta_{1,OLS})_8 = \text{SIF}(\beta_1(.5))_8 = (n-1)(\widehat{\beta}_1 - \widehat{\beta}_{1(8)}) = 1.4$ and $\text{SIF}(\beta_{0,OLS})_8 = \text{SIF}(\beta_0(.5))_8 = (n-1)(\widehat{\beta}_0 - \widehat{\beta}_{0(8)}) = -12$.

The Anscombe data sets provide examples for the case of a single outlier, but real data may have more than one anomalous value. In case of multiple outliers, assuming there are m of them in a sample of size n, the sample influence function becomes $\text{SIF}(\beta)_I = \frac{n-m}{m}(\widehat{\beta} - \widehat{\beta}_{(I)})$, where the index (I) refers to the set of m excluded observations, and the coefficient $\widehat{\beta}_{(I)}$is computed by excluding all of them. For instance, in the $[Y_1^*\ X_1]$ modified Anscombe data set, one additional outlier can be introduced. Besides changing the fourth observation, also the ninth observation of the dependent variable can be modified from $(Y_1^*\ X_1)_9 = (10.84\ \ 12)_9$ into $(Y_1^{**}\ X_1)_9 = (\underline{14.84}\ 12)_9$ as reported in the last column of Table 1.1. In the full sample, the estimated regression coefficients are $\widehat{\beta}_{OLS} = \begin{bmatrix} \widehat{\beta}_0 = 2.94 \\ \widehat{\beta}_1 = 0.61 \end{bmatrix}$ and $\widehat{\beta}(.5) = \begin{bmatrix} \widehat{\beta}_0(.5) = 3.24 \\ \widehat{\beta}_1(.5) = 0.48 \end{bmatrix}$. Figure 1.16 reports the OLS and the median regression estimated lines in the sample of size $n = 11$ comprising two outliers. Next both the fourth and the ninth observation are excluded from the sample. The OLS and the median regression estimated coefficients in the subset without outliers are very similar to one another, $\widehat{\beta}_{OLS(4,9)} = \begin{bmatrix} \widehat{\beta}_{0(4,9)} = 3.21 \\ \widehat{\beta}_{1(4,9)} = 0.43 \end{bmatrix}$ and $\widehat{\beta}(.5)_{(4,9)} = \begin{bmatrix} \widehat{\beta}_{0(4,9)}(.5) = 3.27 \\ \widehat{\beta}_{1(4,9)}(.5) = 0.46 \end{bmatrix}$, with OLS getting closer to the median regression

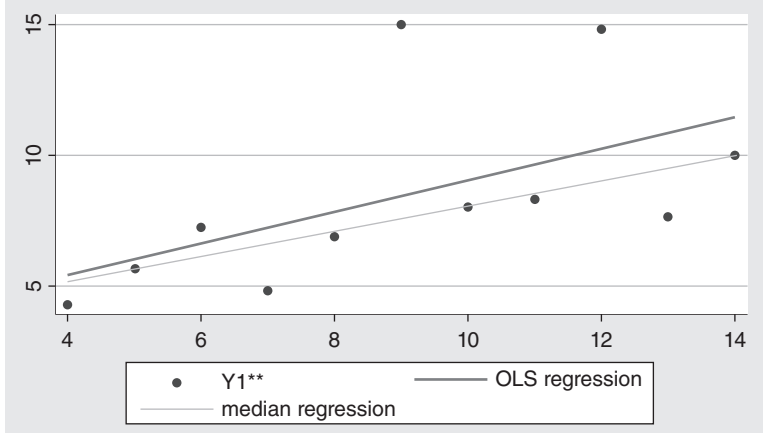

Figure 1.16 OLS and median regression when there are two anomalous values in the dependent variable, the 4^{th} and the 9^{th} observation of the $[Y_1^{**}\ X_1]$ data set, in a sample of size $n = 11$. The scatterplot clearly shows the two outliers, located at the top of the graph. Together they cause an increase in the OLS estimated slope, as depicted by the upper line in the graph. This is not the case for the median regression, which yields the same estimates in the $[Y_1\ X_1]$, $[Y_1^{*}\ X_1]$ and $[Y_1^{**}\ X_1]$ data sets.

estimates. The OLS sample influence function is

$$\begin{aligned}\text{SIF}(\widehat{\beta}_{OLS})_{4,9} &= \frac{11-2}{2}(\widehat{\beta}_{OLS} - \widehat{\beta}_{OLS(4,9)}) \\ &= 4.5\begin{bmatrix} \widehat{\beta}_0 - \widehat{\beta}_{0(4,9)} = 2.94 - 3.21 = -0.27 \\ \widehat{\beta}_1 - \widehat{\beta}_{1(4,9)} = 0.61 - 0.43 = 0.18 \end{bmatrix} = \begin{bmatrix} -1.21 \\ 0.81 \end{bmatrix}\end{aligned}$$

while for the median regression, it is

$$\begin{aligned}\text{SIF}(\widehat{\beta}(.5))_{4,9} &= \frac{11-2}{2}(\widehat{\beta}(.5) - \widehat{\beta}(.5)_{(4,9)}) \\ &= 4.5\begin{bmatrix} \widehat{\beta}_0(.5) - \widehat{\beta}_{0(4,9)}(.5) = 3.24 - 3.27 = -0.03 \\ \widehat{\beta}_1(.5) - \widehat{\beta}_{1(4,9)}(.5) = 0.48 - 0.46 = 0.02 \end{bmatrix} = \begin{bmatrix} -0.13 \\ 0.09 \end{bmatrix}\end{aligned}$$

The two outliers are highly influential in OLS. A more than double value of $\text{SIF}(\widehat{\beta}_{OLS})_{4,9}$ compared to $\text{SIF}(\widehat{\beta}(.5))_{4,9}$ shows once again the robustness of the quantile regression with respect to outliers in the dependent variable.

As a final check, in the $[Y_1^{**}\ X_1]$ data set, the vector of standardized residuals from the median regression can be computed:

$$\begin{aligned}std(\tilde{e}_i) &= \left[\frac{\tilde{e}_i}{\sigma\sqrt{1-h_{ii}}}\right]^T \\ &= \begin{bmatrix}1.13e^{-08} & -.18 & -4.72 & 9.83 & -.26 & 0 & 5.46 & 0 & 9.42 & -3.25 & 0\end{bmatrix}\end{aligned}$$

The largest standardized residuals are located at the fourth and ninth observations, signaling that these two data points, which are the ones intentionally modified, are influential outliers.

1.3.1 Diagnostic in the French and the Dutch data

Starting with the French data, in the original sample, the OLS residual corresponding to the outlying observation with the highest walking speed and large *BMI* is $\widehat{e}_i = 2.66$, while the standardized and studentized residuals for this observation are much larger, respectively $std(\widehat{e}_i) = 9.36$ and $stu(\widehat{e}_i) = 10.61$. There is quite a discrepancy between the simple OLS residual and its standardized and studentized values. This signals that the observation is anomalous and influential. The outlier has attracted the OLS estimated line causing a smaller OLS residual and a positive estimated slope. In the median regression, for this same observation, the residual is $\tilde{e}_i = 2.71$, and the standardized and the studentized residuals are similar, respectively $std(\tilde{e}_i) = 2.72$ and $stu(\tilde{e}_i) = 2.74$. There is not much difference among these three values since the median regression is not attracted by outliers in the dependent variable. For the slope, the SIF_i computing the impact of this observation on the OLS estimator, is $\text{SIF}(\widehat{\beta}_{OLS})_i = (n-1)(\widehat{\beta}_{OLS} - \widehat{\beta}_{OLS(i)}) = 392(.0033 + .0009) = 1.64$, to be compared with its analogue for the median regression, $(n-1)(\widehat{\beta}(.5) - \widehat{\beta}_{(i)}(.5)) = 0$.

Next consider the Dutch data set, which is characterized by one outlier in the explanatory variable. The best tool to identify this kind of outlier is $h_{ii} = x_i^T (X^T X)^{-1} x_i$, which reaches the highest value of $h_{ii} = 0.280$ for $BMI = 73.39$, followed by $h_{ii} = 0.12$ for $BMI = 57.34$ and $h_{ii} = 0.11$ for $BMI = 55.53$. These are the three data points located to the right of Figure 1.13. The largest value in h_{ii} clearly signals the largest outlier in the explanatory variable. However, the corresponding SIF_i in this example is quite small. Indeed, for the OLS slope it is $\text{SIF}(\widehat{\beta}_{OLS})_i = (n-1)(\widehat{\beta}_{OLS} - \widehat{\beta}_{OLS(i)}) = 306(-.0058 + .0054) = -0.122$, and for the median regression slope is $\text{SIF}(\widehat{\beta}(.5))_i = 306(-.0065 + .0069) = 0.122$. Thus the *BMI* outlier is not particularly influential; indeed the corresponding OLS residual, together with the standardized and studentized residual, are respectively $\widehat{e}_i = -0.048$, $std(\widehat{e}_i) = -0.20$, and $stu(\widehat{e}_i) = -0.19$. These values are all quite small and become even smaller in the median regression case: $\tilde{e}_i = -6.07\text{e}^{-9}$, $std(\tilde{e}_i) = -6.13\text{e}^{-9}$ and $stu(\tilde{e}_i) = -6.12\text{e}^{-9}$.

1.3.2 Example with error contamination

This section implements a controlled experiment by analyzing an artificial data set with outliers generated by a contaminated error distribution. In a sample of size $n = 100$, the explanatory variable x_i is given by the realizations of a uniform distribution defined in the [0, 1] interval, and the error term e_i is provided by the realizations of a 10% contaminated normal where a standard normal is contaminated by a zero mean normal with $\sigma = 6$. Figure 1.17 presents the histogram of the realizations of e_i.

Once known x_i and e_i, the dependent variable y_i is defined as $y_i = 3 + 0.9x_i + e_i$. Next the OLS and the median regression estimators are implemented in the artificial

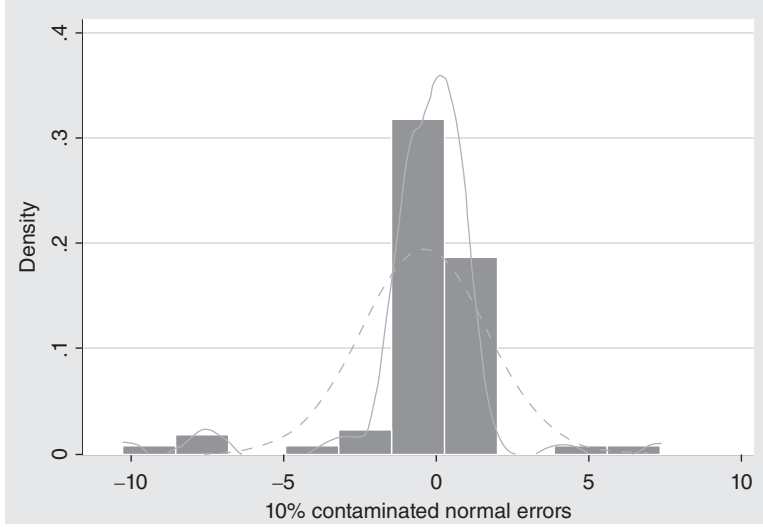

Figure 1.17 Histogram of the realizations of a 10% contaminated normal error distribution. The dashed line is the normal density while the solid line is the Epanechnikov (1969) kernel density of the contaminated normal, sample size $n = 100$. The kernel density shows the presence of values far from the mean, generated by the contaminating distribution.

data set (y_i x_i), and their behavior in the presence of outliers can be fully evaluated since the parameters defining y_i are known in advance. The mean and the median conditional regressions results are reported in the first two columns of Table 1.4, while Figure 1.18 presents the scatterplot of the data. The graph clearly shows the presence of outliers. Although the OLS estimates are close to the true coefficients, $\beta_0 = 3$ and $\beta_1 = 0.9$, the standard error of the slope is so large that $\widehat{\beta}_1$ is not statistically different from zero. The standardized OLS residuals, defined as $std(\widehat{e}_i) = \frac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}}$, and the OLS studentized residuals, $stu(\widehat{e}_i) = \frac{\widehat{e}_i}{\sigma_{(i)}\sqrt{1-h_{ii}}}$, allow to point out the following outliers in the dependent variable: $y_{16} = 7.89$, $y_{36} = 10.79$, $y_5 = -4.28$, $y_{18} = -7.02$,

Table 1.4 OLS and median regression in the case of contaminated error distribution.

	OLS	median	$OLS_{(I)}$	$median_{(I)}$
slope	.940	1.753	1.263	1.696
se	(.71)	(.36)	(.32)	(.41)
constant	2.60	2.372	2.64	2.427
se	(.42)	(.21)	(.19)	(.24)

Note: Standard errors in parenthesis, sample size $n = 100$.

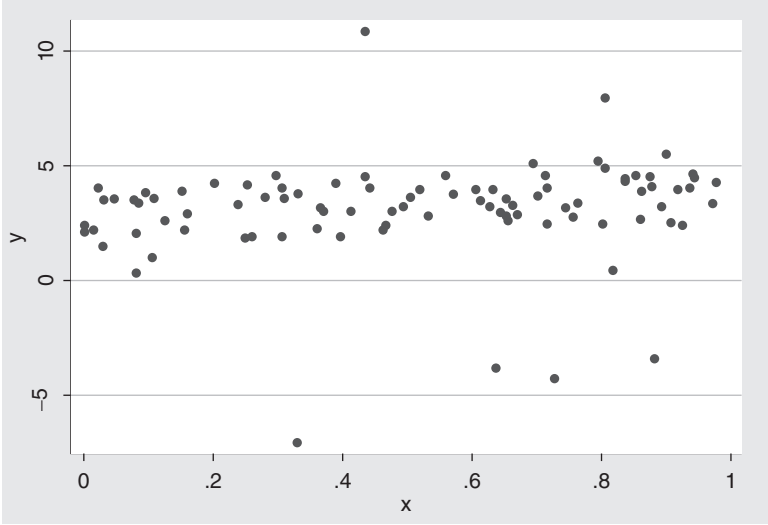

Figure 1.18 Data generated as $y_i = 3 + 0.9x_i + e_i$ where e_i is the contaminated error distribution depicted in Figure 1.17, sample of size $n = 100$. The presence of outliers in the dependent variable, caused by the contaminated error distribution, is clearly visible.

$y_{48} = -3.83$, $y_{52} = -3.46$, as detected by values of $std(\widehat{e}_i)$ outside the ± 1.96 interval. These values are reported in Table 1.5 for both the OLS and the median regression, respectively, $\widehat{e}_i$ and $\tilde{e}_i$ in the table. This table shows that the standardized and studentized residuals of the median regression point out two additional outliers undetected in OLS, $y_{45} = 0.25$ and $y_{97} = 0.38$. Once the anomalous values are detected, it is

Table 1.5 Largest values of the studentized and standardized residuals in both OLS and median regression, $n = 100$.

y_i	OLS $stu(\widehat{e}_i)$	OLS $std(\widehat{e}_i)$	median $stu(\tilde{e}_i)$	median $std(\tilde{e}_i)$
7.89	2.26	2.22	4.54	4.15
10.79	4.08	3.79	12.17	7.69
−4.28	−3.96	−3.69	−13.53	−8.00
−7.02	−5.52	−4.85		−10.06
0.25			−2.36	−2.31
−3.83	−3.63	−3.43	−11.00	−7.37
−3.46	−3.58	−3.38	−11.39	−7.49
0.38			−3.68	−3.47

Note: The empty cells in the table are for the observations that are not detected as outliers.

possible to compute their impact on an estimator by means of the sample influence function, $\text{SIF}(\beta)_I = \frac{n-m}{m}(\widehat{\beta} - \widehat{\beta}_{(I)})$. The last two columns of Table 1.4 report the OLS and median regression estimated coefficients when the outlying values are eliminated, respectively the six observations detected by the standardized and studentized OLS residuals and the eight observations detected by the standardized and studentized residuals of the median regression. For the OLS estimator, the SIF $(\beta)_I$ is

$$\begin{aligned}\text{SIF}(\widehat{\beta}_{OLS})_I &= \frac{(n-m)}{m}(\widehat{\beta}_{OLS} - \widehat{\beta}_{OLS(I)}) \\ &= \frac{100-6}{6}\begin{bmatrix} \widehat{\beta}_0 - \widehat{\beta}_{0(I)} = 2.60 - 2.64 = -0.04 \\ \widehat{\beta}_1 - \widehat{\beta}_{1(I)} = 0.94 - 1.26 = -0.32 \end{bmatrix} = \begin{bmatrix} -0.63 \\ -5.01 \end{bmatrix}\end{aligned}$$

while for the median regression, it is

$$\begin{aligned}\text{SIF}(\widehat{\beta}(.5)_I &= (\widehat{\beta}(.5) - \widehat{\beta}(.5)_{(I)}) \\ &= \frac{100-8}{8}\begin{bmatrix} \widehat{\beta}_0(.5) - \widehat{\beta}_0(.5)_{(I)} = 2.37 - 2.42 = -0.05 \\ \widehat{\beta}_1(.5) - \widehat{\beta}_1(.5)_{(I)} = 1.75 - 1.69 = 0.06 \end{bmatrix} = \begin{bmatrix} -0.57 \\ 0.69 \end{bmatrix}\end{aligned}$$

The comparison of $\text{SIF}(\widehat{\beta}_{OLS})_I$ and $\text{SIF}(\widehat{\beta}(.5))_I$ shows that the impact of the set of observations in I is larger in the OLS model than in the median regression case, and the difference is quite marked for the slope coefficient, although the number of excluded observations in the OLS regression is smaller than in the median regression case. Finally, by computing the quantile regression at $\theta = .10, .25, .50, .75, .90$, all the eight outliers detected by the standardized and the studentized quantile residuals are clearly visible, as shown in Figure 1.19. The estimated coefficients at the θ

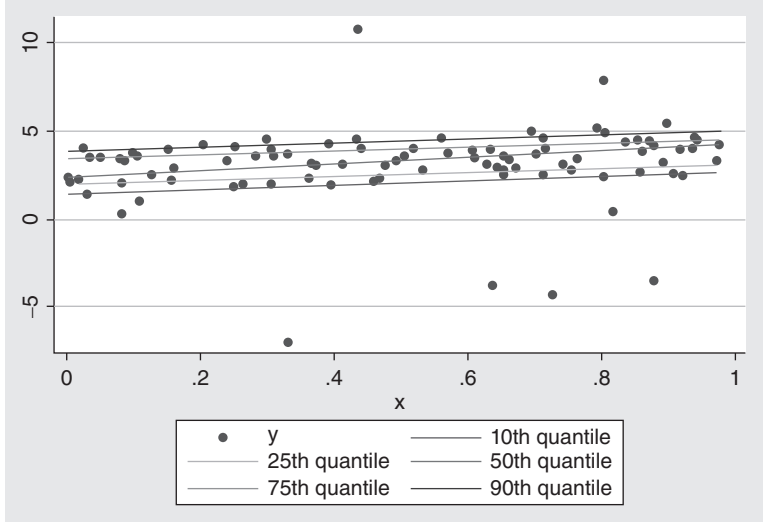

Figure 1.19 Estimated quantile regressions for the model with 10% contamination in the errors, sample of size $n = 100$.

Table 1.6 Quantile regression estimates, contaminated errors.

θ	10^{th}	25^{th}	50^{th}	75^{th}	90^{th}
slope	1.21	1.12	1.75	1.08	1.31
se	(5.8)	(.54)	(.36)	(.39)	(3.7)
constant	1.40	1.99	2.37	3.44	3.81
se	(3.4)	(.32)	(.21)	(.23)	(2.2)

Note: Standard errors in parenthesis, sample size $n = 100$.

quantiles are reported in Table 1.6, where both the lower and the higher quantile regression estimates are not significantly different from zero.

1.4 A summary of key points

Section 1.1 focuses on the shortcomings of the OLS estimator in the linear regression model by analyzing an artificial data set, discussed by Anscombe (1973), where OLS yields the same estimated coefficients in four different models, but the result is correct only in one of them. Section 1.2 considers the behavior of the median regression in the Anscombe data and in two additional models analyzing real data, in order to show the robustness of the quantile regression approach. Section 1.3 considers the influence function, a statistic that measures the impact of outliers on an estimator, together with its empirical approximation, the sample influence function. Examples with real and artificial data sets allow to compare the influence function of the OLS and of the median regression estimators in samples characterized by different kinds of outliers.

References

Anscombe, F. 1973. "Graphs in statistical analysis." The American Statistician 27, 17–21.

Chatterjee, S., and Hadi, A. 1988. *Sensitivity Analysis in Linear Regression*. Wiley.

Epanechnikov, V. 1969. "Nonparametric estimation of a multidimensional probability density." Theory of Probability and its Applications 14 (1), 153–158.

Hampel, F. 1974. "The influence curve and its role in robust estimation." Journal of the American Statistical Association 69, 383–393.

Hoaglin, D., Mosteller, F., and Tukey, J. 1983. *Understanding Robust and Exploratory Data Analysis*. Wiley.

Huber, P. 1981. *Robust Statistics*. Wiley.

Appendix: computer codes in Stata

A) generate contaminated distributions	
`gen norm1=invnorm(uniform())`	generate N(0,1)
`gen norm4 = 4*invnorm(uniform())`	generate N(0,16)
`gen conta = cond(norm1 >-1.96,norm1,norm4)`	left tail contamination
`gen contam = cond(norm1 > 1.96,norm4,conta)`	right tail contamination

B) compute diagnostic measures to detect outliers	
`reg y x (or qreg y x)`	OLS (or quantile regression)
`predict name, resid`	compute regression residuals
`predict name1, hat`	compute h_{ii}
`predict name2, rstand`	compute standardized residuals
`predict name3, rstud`	compute studentized residuals