Diagnostics Plots for Linear Regression Analysis

A set of techniques called *regression diagnostics* provides us with the necessary tools for evaluating the appropriateness of the regression model and can help us to uncover and correct problems. We will examine the standard approach that uses functions that come with R's base installation.

The diagnostic plots show residuals in four different ways. Let's take a look at the first type of plot:

1. Residuals vs Fitted (linearity assumption)

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between explanatory variables and an outcome variable, and the pattern could show up in this plot. If we find equally spread residuals around a horizontal line without distinct patterns, that is a good indication and we have linear associations.
Let's look at residual plots from a 'good' model and a 'bad' model (Figure 9). The good model data are simulated in a way that meets the regression assumptions very well, while the bad model data are not.

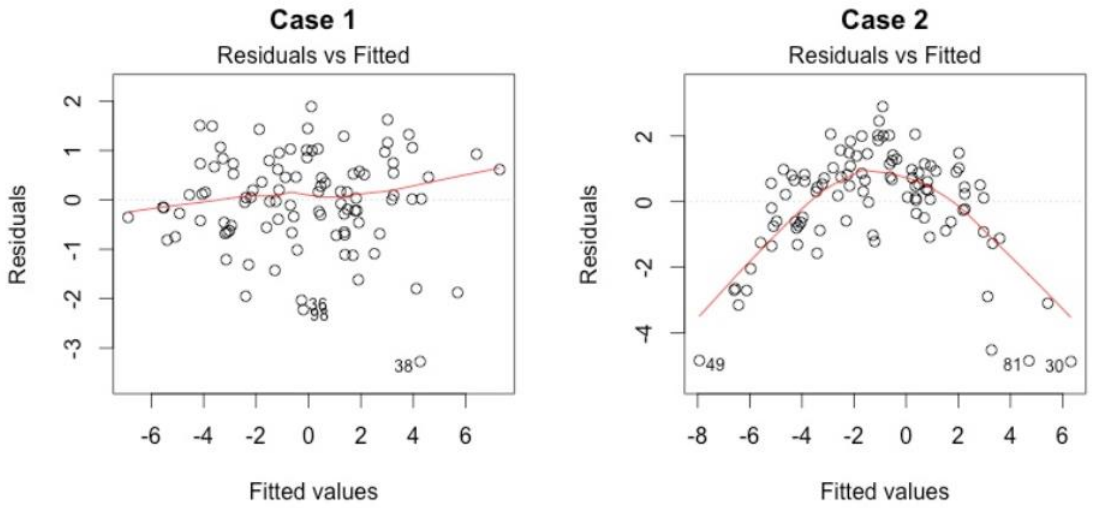

Figure 9: Plots of residuals against predicted (fitted) values in two different cases (case 1: linear pattern, case 2: non-linear pattern)

We can see that there is not any distinctive pattern in Case 1, but probably we can see a parabola in Case 2, where the non-linear association was not explained by the model and was left out in the residuals.

2. Normal Q-Q plot (normality assumption)

The Normal Q-Q plot is a probability plot of the standardized residuals against the values that would be expected under normality. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight 45-degree dashed line.

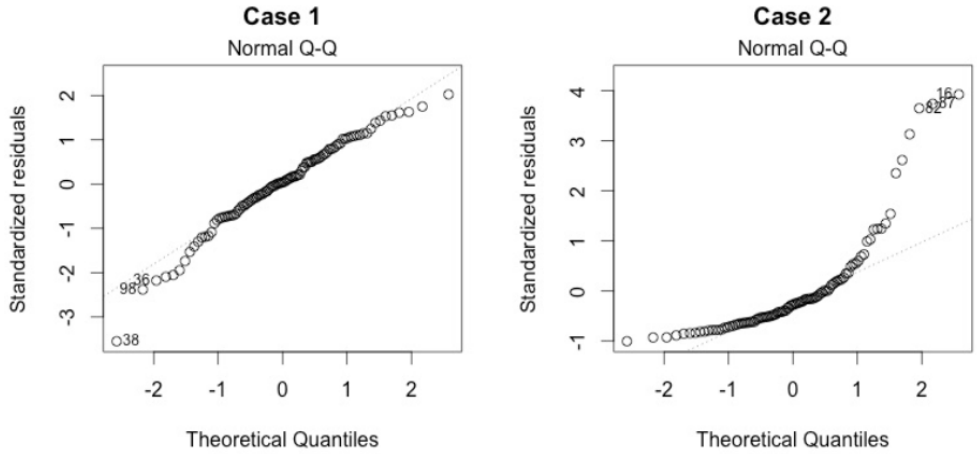

Figure 10: Q-Q plots of approximately normally distributed residuals (left-hand side) and non-normally distributed residuals (right-hand side)

Case 2 definitely should concerns us. We would not be concerned by Case 1 too much, although an observation numbered as 38 looks a bit off. Let's look at the next plot while keeping in mind that observation with number 38 might be a potential problem.

3. Scale-Location (homoscedasticity assumption)

If we've met the constant variance assumption (homoscedasticity), the points in the Scale-Location graph should be a random band around a horizontal line.

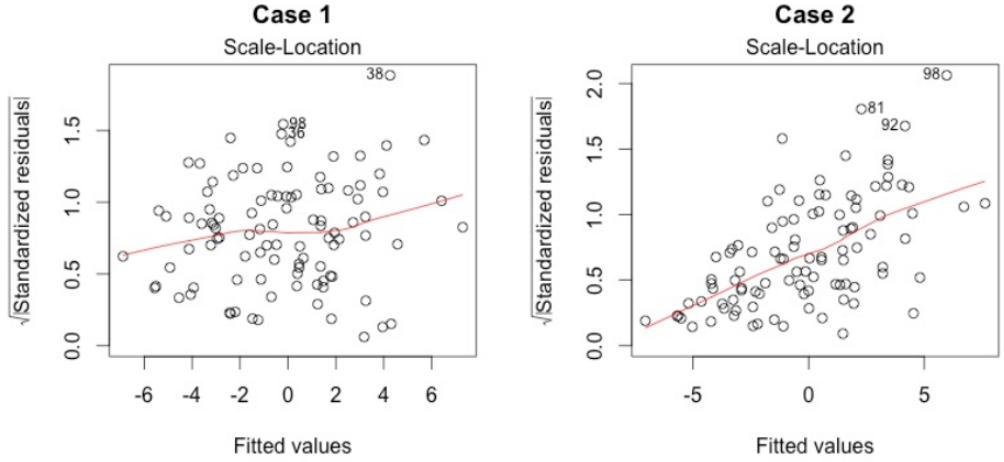

Figure 11: The square root of the standardized residuals versus fitted values plot in two cases (case 1: homoscedasticity, case 2: no homoscedasticity)

In Case 1, the residuals appear randomly spread. Whereas, in Case 2, the residuals begin to spread wider along the x-axis as it passes around 5. Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle in Case 2.

4. Residuals vs Leverage (influential points)

Outliers are points that fall away from the cloud of points. If they fall horizontally (x-axis) away from the center of the cloud are called *leverage points*. High leverage points that actually influence the parameters of the regression model are called *influential points*. A measure of the influence of each point on the fitted model is the *Cook's statistic*. Values ≥ 1 (or even approaching 1) correspond to highly influential observations.

Therefore, not all outliers are influential in linear regression analysis. Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn't be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases, and they don't really matter; they are not influential.

On the other hand, some cases could be very influential even if they look to be within a reasonable range of values. They could be extreme cases against a regression line and can alter the results if we exclude them from the analysis. Another way to put it is that they don't get along with the trend in the majority of the cases.

The **Residual Vs Leverage plot** helps us to find influential cases if any. Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.

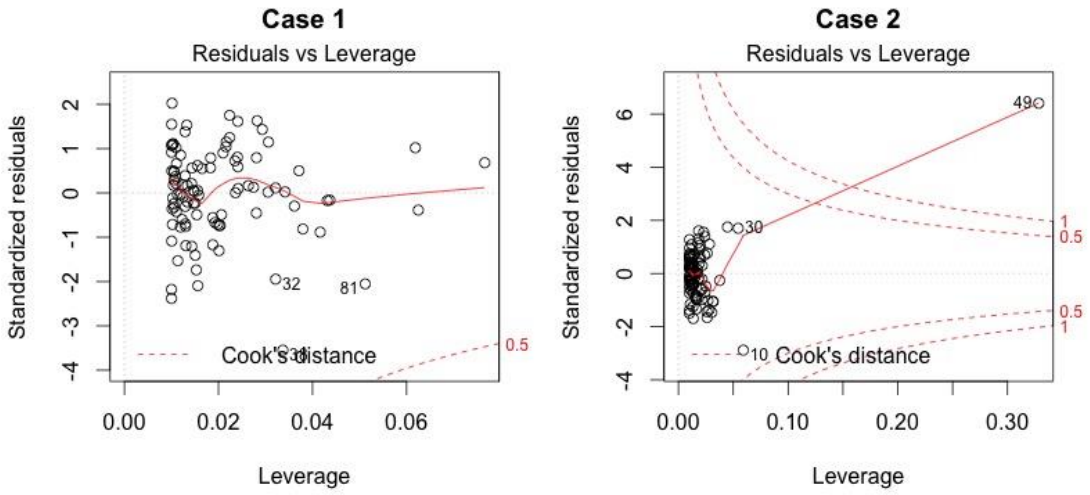

Figure 12: Standardized residuals Vs Leverage plots in two cases (case 1: no influential point, case 2: influential point)

Case 1 is the typical look when there is no influential cases. You can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines. In Case 2, a case is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot). The plot identified the influential observation number 49.

Influential points: three different scenarios

With these facts in mind, consider the plots associated with three different situations (figures 13,14,15) for **the special red point 21**. The plots on the left show the data, the center of the data $(\bar{x}, \bar{y})$ with a blue dot, the underlying data generating process with a dashed gray line, the model fit with a blue line, and the special point with a red dot. On the right are the corresponding residual-leverage plots; the special point is the red `one`. In figure 13 the red point 21 is far away (as far as x- axis) from blue point (the center

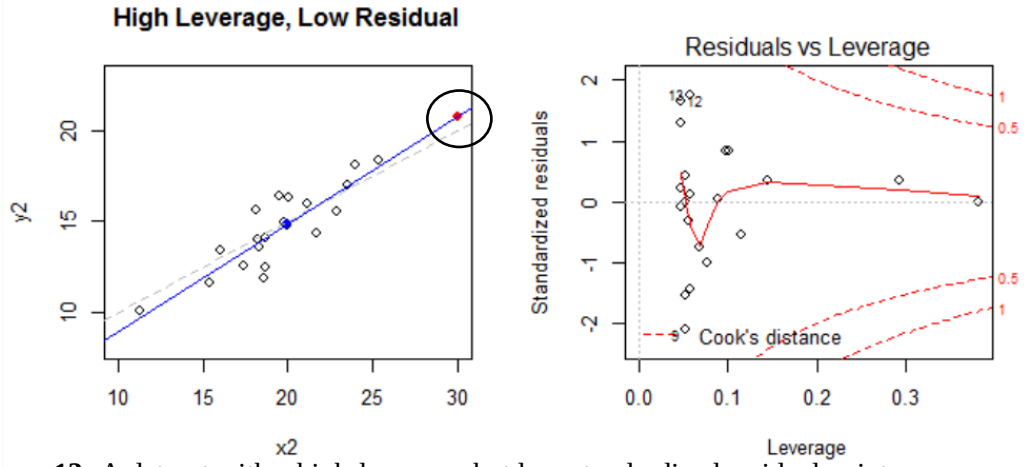

Figure 13: A dataset with a high-leverage, but low-standardized residual point

of the data), so it is a high leverage point (x-axis distance). But the red point is also close to the fitted line, so it has small residual (y-axis). The Cook distance is the result of the combination of these two factors and equals to 0.0000007<1. Therefore, it is not influential point. As we can see from the Residual-Leverage plot there is not any point outside the red dashed lines (Cook's distance).

In figure 14 the red point 21 is close (as far as x- axis) to blue point (the center of the data), so it is a low leverage point (x-axis distance). But the red point is far away to the fitted line, so it has large residual (y-axis). The Cook distance is the result of the combination of these two factors and equals to 0.2968102 <1. Therefore, it is not influential point. As you can see from the Residual-Leverage plot there is not any point outside the red dashed lines (Cook's distance).

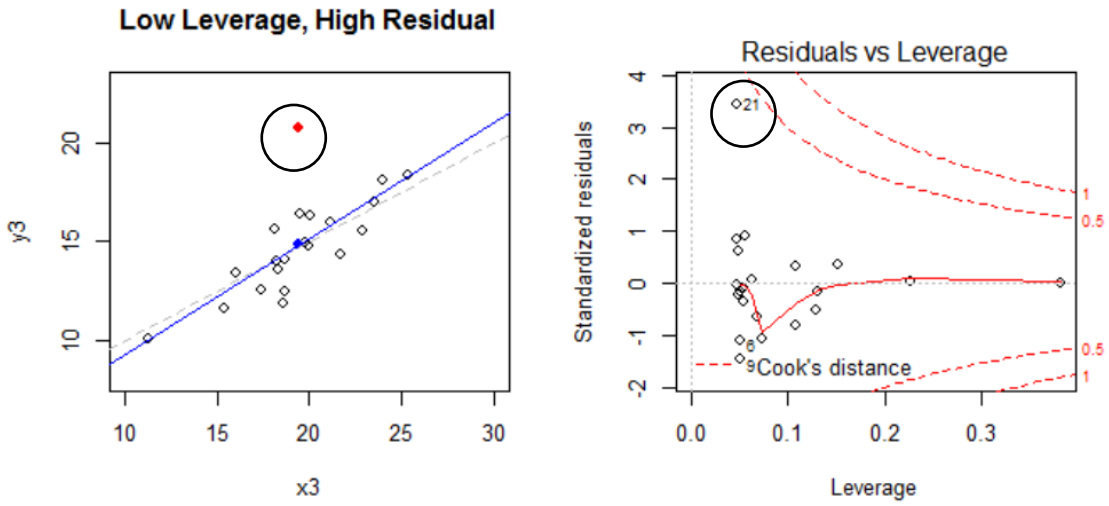

Figure 14: A dataset with a low-leverage, but high-standardized residual point

The model is badly distorted primarily in the figure 15 where there is a point with high leverage and a large (negative) standardized residual.

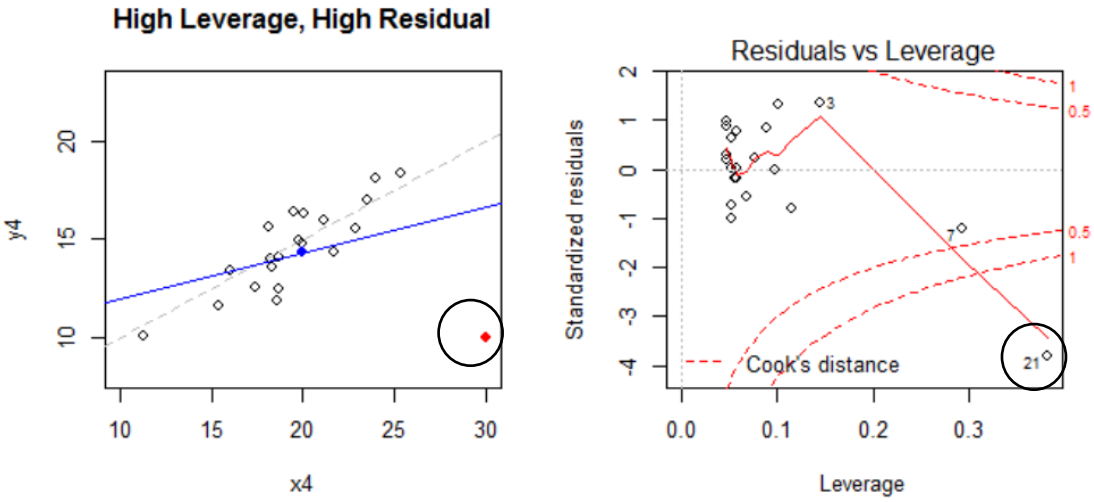

Figure 15: A dataset with a high-leverage, high-standardized residual point (observation 21 is an influential point).

For reference, here are the values associated with the special points:

	leverage	std.residual	cooks.d
high leverage, low residual	0.3814234	0.0014559	0.0000007
low leverage, high residual	0.0476191	3.4456341	0.2968102
high leverage, high residual	0.3814234	-3.8086475	4.4722437