Lecture 5

Outliers and Influential Points in Linear Models

There are a number of methods for dealing with outliers. We are not going to address all of them, but we will get a flavor for the various strategies and tools available in R. We'll also address the reasons for being concerned about these points and how to deal with them.

Outliers and influential points can be the same point or different points.

An outlier is a point that has a particularly large error relative to the regression line. We will look at several ways to identify these points numerically and some hypothesis tests we can apply.

An influential point may be an outlier in the sense above, but it need not be. An influential point is a point such that, if it is removed from the data, has an outsized influence on the regression results: it may, for example, greatly impact the value or even the direction of the slope.

Data sets that are smaller tend to be more effected by either type of point. As the size of the dataset grows, we should expect to have more outliers, but that each will have less impact on the overall trend.

Last semester, when we looked at univariate (single variable) data, we discussed some strategies for identifying outliers. We'll review those here because these methods can also be useful.

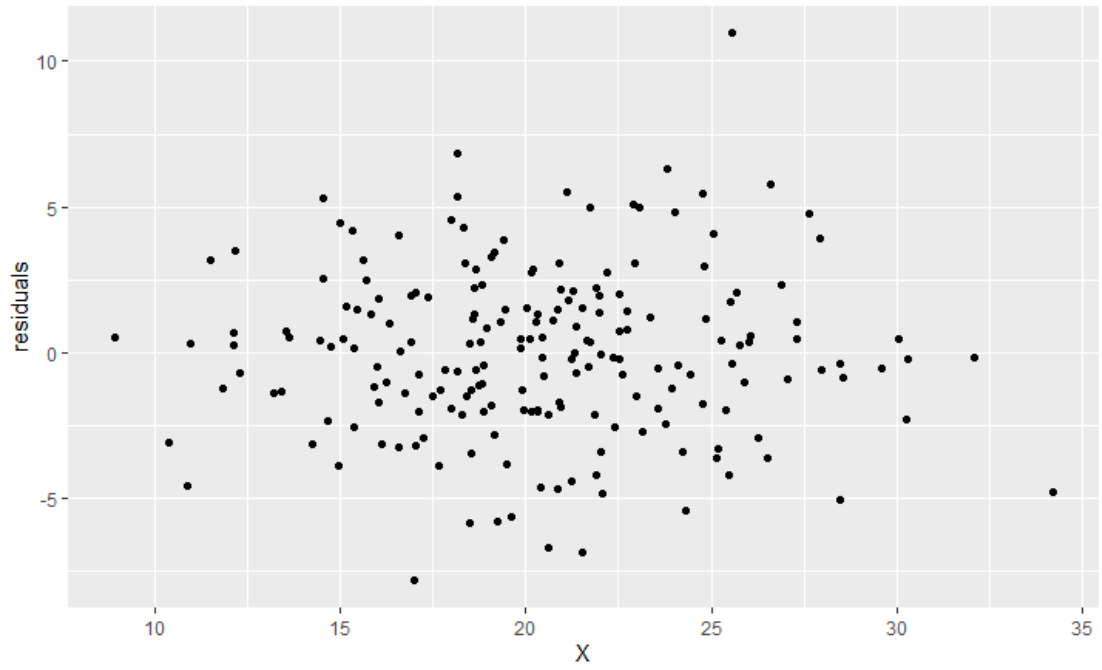
An unusual value is generally considered to be an observation that occurs less than 5% of the time. Recall from the empirical rule (normal distribution), that 95% of normally distributed data falls within 2 standard deviations of the mean. So that is one possible standard we can use to identify potential outliers. Extreme outliers would be more than 3 standard deviations from the mean.

Outliers can also be identified, especially in skewed data, using the interquartile range (IQR). Recall that the upper and lower fences in a box plot are $1.5 \times IQR$ above the third quartile, or below the first quartile. The extreme fences are at $3 \times IQR$. We can apply these standards to also identify potential outliers.

These standards applied to the independent or dependent variables may help you to identify influential points. If a point is a long way from the rest of the data, it can have a big influence on the slope of the line. Or it may not. Applying these outlier standards to the original data can only flag points that are worth inspecting more closely, but the are not a guarantee that these points are problematic.

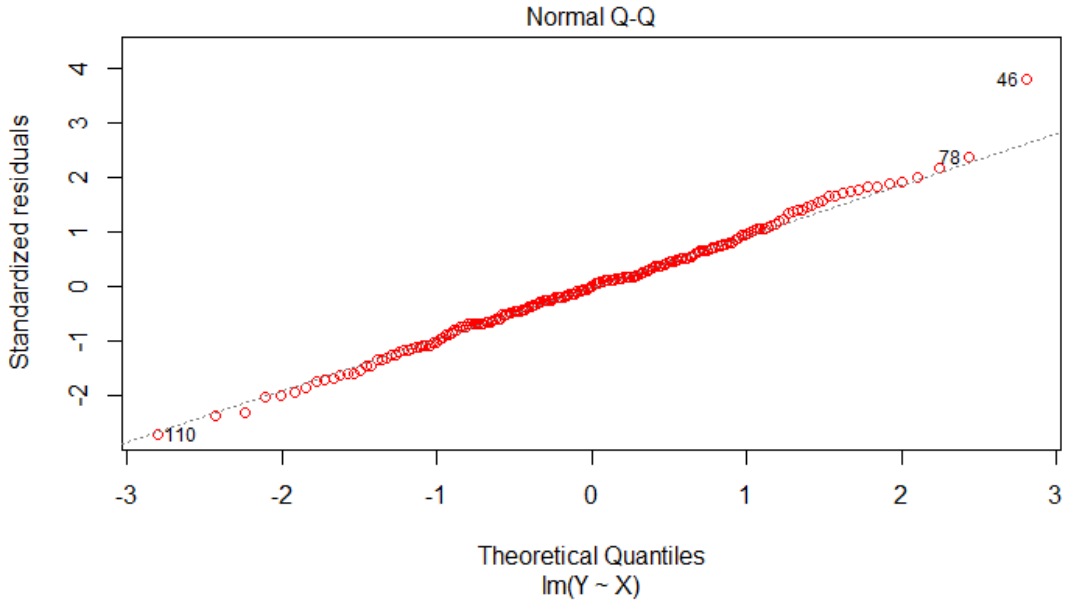
However, when working with regression, the main place we want to apply these standards to is to the residuals. The outliers are those points that are far away from the line of best-fit. Let's look at a specific example. We'll use a simple linear regression model for this discussion, but everything we say about this for a simple linear regression model works the same way with a multivariable regression model.

Let's look at our residual plot for our simulated example that we've discussed previously.



We have one X value we might want to pay attention to out on the far right, but the residual is small. We may return to this later as a potentially influential point. However, the outlier we want to pay attention to the is the residual at the top of the graph, the only one with a value greater than 10. Less concerning, but possibly of interest is the smallest residual. It's the only one smaller than -7.5, so it has the second largest magnitude since the graph has no other residuals bigger than 7.5 on the other end except the one that is bigger than 10.

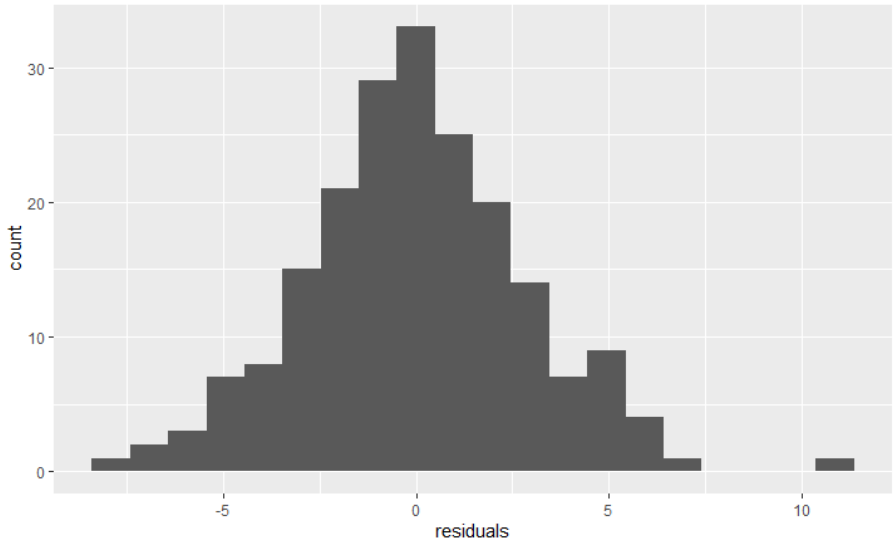
There are many way to identify the observation that these points came from. We can apply a filter to the data. Or the qq-plot we looked at, also flagged values with observation numbers for us.



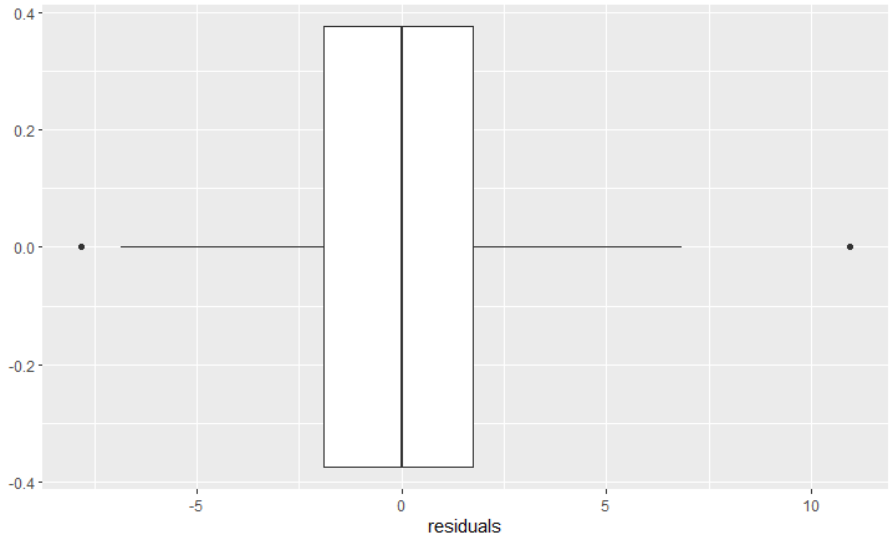
This plot flags our high-value point at observation 46. And the smallest one at observation 110. This dataset has 200 observations, so we would expect $200 \times 0.05 = 10$ potentially unusual values. And from the qq-plot, we can see that there are many values at or beyond two standard deviations from the mean. From the plot, I estimate 5 below the -2 standard deviation mark, and about the same above 2 standard deviation mark. This is what we would expect.

In general, we should be cautious about removing data without good reason. Observation 46 appears to be larger than expected (that's what it means to be above the line), so that one warrants the most attention. This is not a guarantee, however, that we should remove it.

We can also look at the residuals with a histogram or boxplot to see if these types of graphs flag points for us.



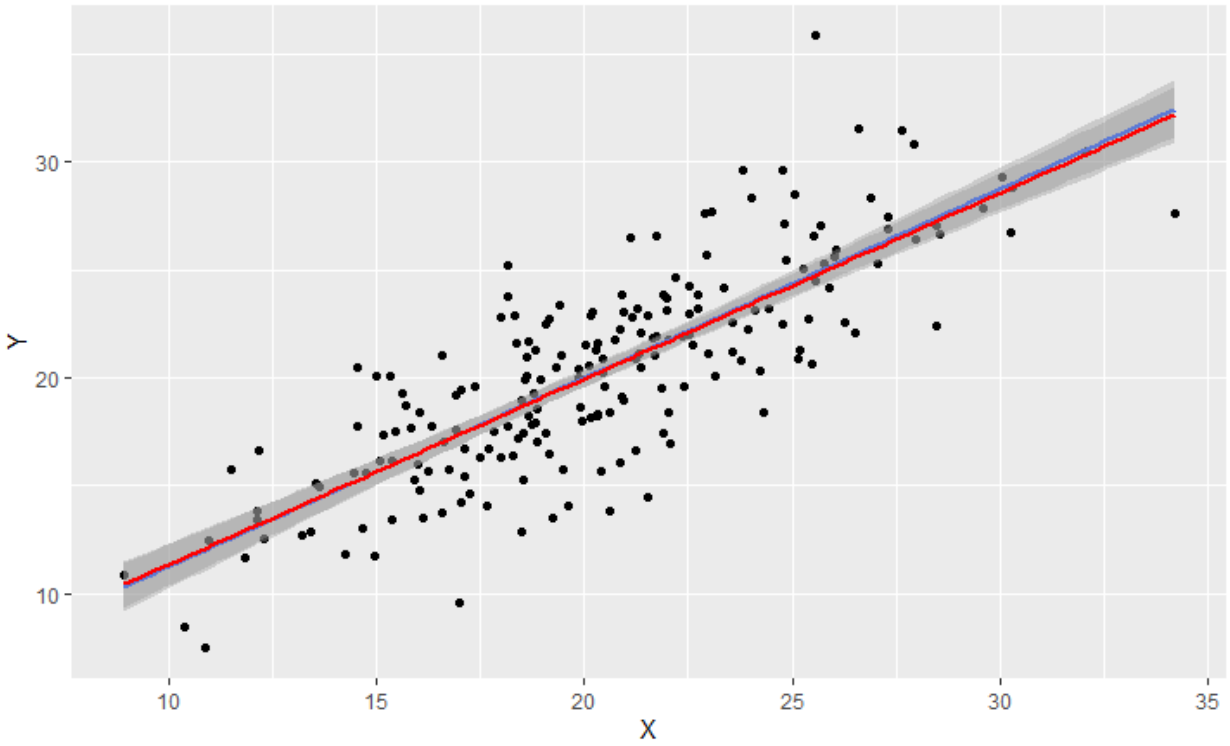
We can see from the histogram that most of the data fits relatively nicely into a "smooth" curve (as smooth as one might expect from data), but there is one value that is way out to the right with a substantial gap.



The boxplot also flags the largest value on the right as a potential extreme value, and the furthest observation on the left as well.

If the outliers don't impact the regression line much, generally, it's better to keep the values, since removing them will reduce the value of the standard error. This could impact your confidence intervals, prediction intervals, and other estimates. It may cause you to underestimate the variability in the data.

Let's see what happens if we remove the most extreme point, observation 46 and then compare the analysis before and after.



This graph shows the blue line with all 200 observations, and the red line with the regression after omitting observation 46. As you can see, you can barely see the blue line at all (a bit on the ends), which means that the regression did not substantially change from omitting the extreme value. We can look at the coefficients of the model and see much the same result.

Model with observation 46:
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.48770	0.95444	2.606	0.00984 **
X	0.87604	0.04581	19.123	< 2e-16 ***

Model without observation 46:
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.72777	0.92318	2.955	0.00351 **
X	0.86152	0.04437	19.418	< 2e-16 ***

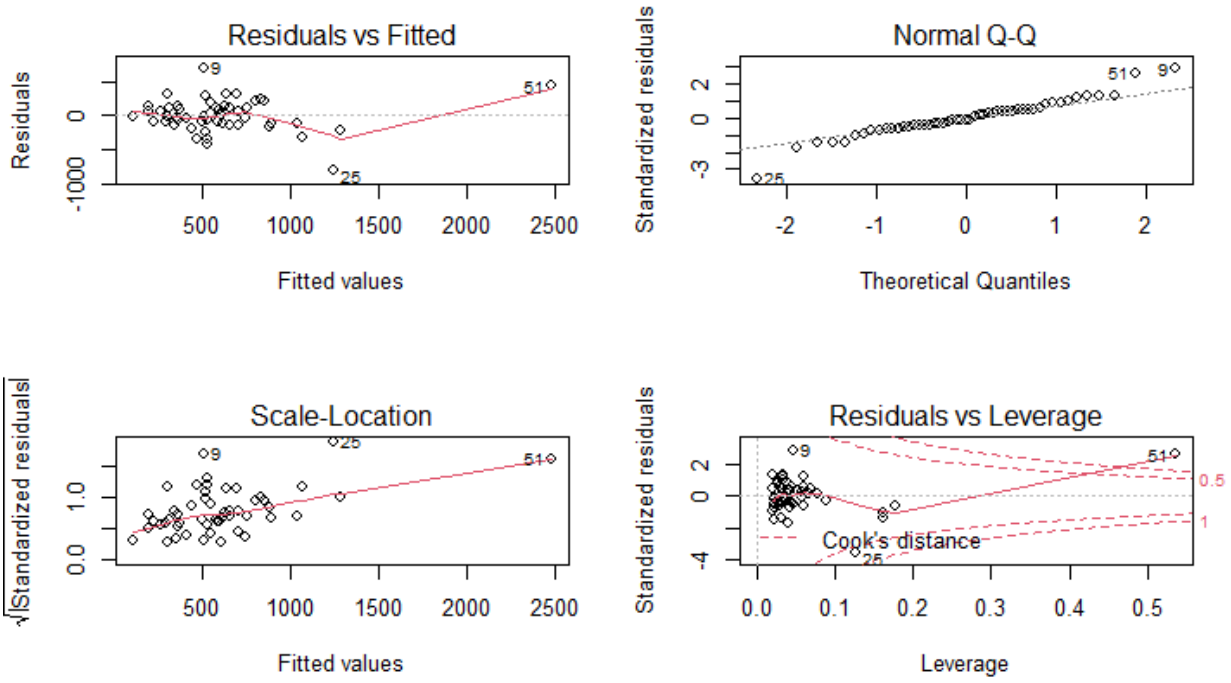
The R^2 value is a bit higher in the second model. I think I would have to conclude that while observation 46 is an outlier, nonetheless, it is not an influential point, and so can be safely retained in the data.

We can conduct a similar test on observation 110 (the most negative residual), but if this one is not influential, given the position of observation 110 is not far to one end of the dataset, the result is likely to be similar.

Let's look at a case where the point is influential.

There are a number of packages in R that can assist you in determining outliers and influential points.

After fitting a model to crime data, we can look at a series of graphs that tell a story about our residuals.



The top left graph is plotting residuals vs. fitted (predicted) values (instead of against the independent values). The qq-plot looks okay for the most part. Scale-location measures the size of the residual relative to their distance from center. Leverages is a measure of how much influence a point has on the regression slope(s) using Cook's distance.

You can use these graphs to identify potentially problematic points. Based on this data 9, 25 and 51 need a further look.

In addition to these graphical approaches combined with a variety of calculations related to extreme values, leverage, percentiles, distances or other measures, there are a number of statistical tests you can apply.

Among them are Grubb's Test which allows you to detect whether the highest or lowest value in a dataset is an outlier. Dixon's test is similar, allowing you to test a specific high or low value. Rosner's Test allows you to test several values at once which can avoid the problem of masking, if two extreme values are close together. You can specify how many values you want to check for, and the test will return a report on all tests with observation numbers that make them easy to track down and further analyze.

There are also additional packages in R with additional tools. We'll look at some of these tools in the lab. It's a good idea to experiment with several of them and see which you like the best. There may be standards in your field to follow.

My biggest piece of advice, though, is to reject data with caution. Use these tools to look for errors. There may be more incentive to reject problematic points in small data sets. But in all cases, preserve the original data and compare, and be prepared to justify your choices with robust reasoning. Don't just reject data because it was flagged in one test. Think of the procedure as a hypothesis test: the null hypothesis should be to preserve the data. The alternative should be supported with strong reasoning before rejecting a point.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. http://r-statistics.co/Outlier-Treatment-With-R.html
3. https://statsandr.com/blog/outliers-detection-in-r/
4. https://rpubs.com/mpfoley73/501093