1 Outliers and Influential Observations

> "Outliers are sample values that cause surprise in relation to the majority of the sample" (W.N. Venables and B.D. Ripley. 2002. Modern applied statistics with S. New York: Springer, p.119).

Crucially, surprise is in the mind of the beholder and is dependent on some tacit or explicit model of the data.

There may be another model under which the outlier is not surprising at all, say if the data really are lognormal or gamma rather than normal.

1.1 Outliers

Data points that diverge in a big way from the overall pattern are called outliers. There are four ways that a data point might be considered an outlier.

- It could have an extreme X value compared to other data points.
- It could have an extreme Y value compared to other data points.
- It could have extreme X and Y values.
- It might be distant from the rest of the data, even without extreme X or Y values.

After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line. If a point lies far from the other data in the horizontal direction, it is known as an influential observation. The reason for this distinction is that these points have may have a significant impact on the slope of the regression line.