**When is linear regression appropriate?**

The sensible use of linear regression on a data set requires that four assumptions about that data set be true:

1. The relationship between the variables is **linear**.
2. The data is **homoskedastic**, meaning the variance in the residuals (the difference in the real and predicted values) is more or less constant.
3. The residuals are **independent**, meaning the residuals are [**distributed randomly**](http://gsb420.blogspot.fi/2008/03/lecture-8-residual-analysis-checking_04.html) and not influenced by the residuals in previous observations. If the residuals are not independent of each other, they’re considered to be *autocorrelated*.
4. The residuals are **normally distributed**. This assumption means the probability density function of the residual values is normally distributed at each x value. I leave this assumption for last because I [**don’t consider it to be a hard requirement**](http://stats.stackexchange.com/questions/29731/regression-when-the-ols-residuals-are-not-normally-distributed) for the use of linear regression, although if this isn’t true, some manipulations must be made to the model.