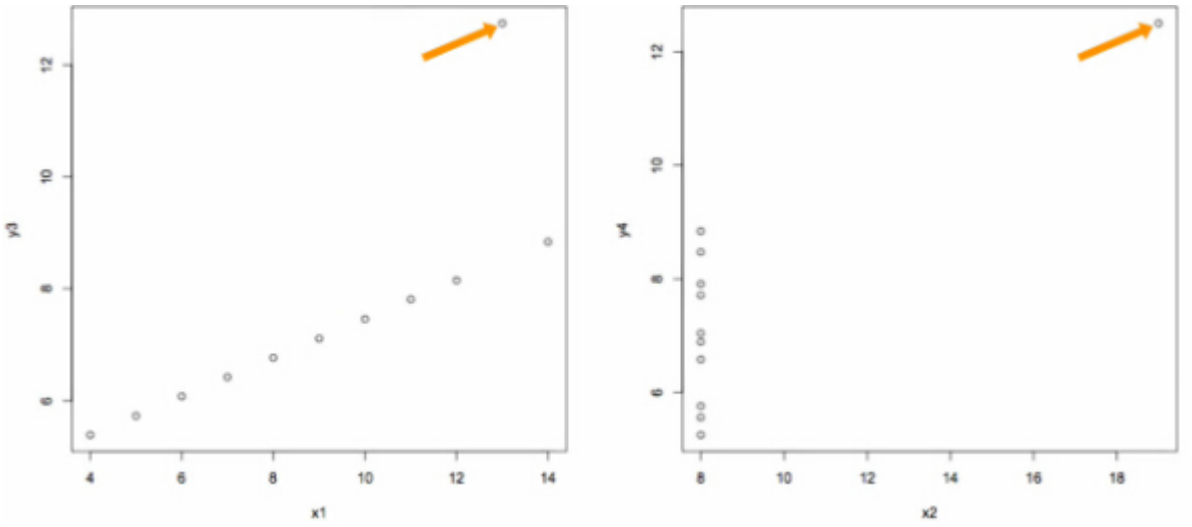
Simple Linear Regression

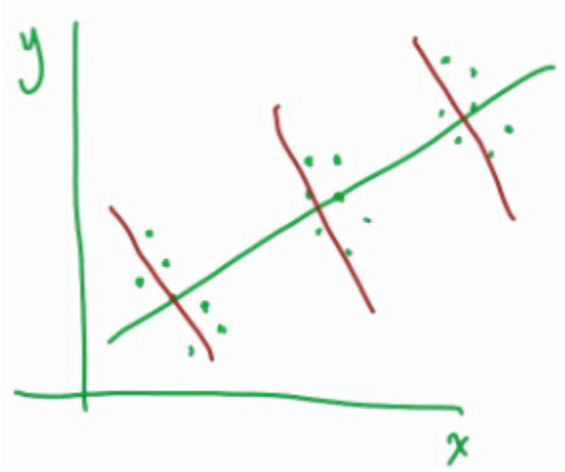
Checking the Conditions for Linear Regression

To make valid conclusions from data, we have to carry out statistical methods appropriately. Here we discuss the conditions required for using linear regression and how we can examine whether they hold for our data.

Condition 1: A linear model is appropriate for the data, no curvature, no influential points, no groups with different patterns.



Influential points



Groups of points for which different models are appropriate

Figure 1: Examples where one linear model is inappropriate for the data.

To check Condition 1, we should first examine a scatterplot of the data. Follow up a linear model with a residual plot to make sure that there aren't patterns we've missed. The residual plot should look like random scatter.

In order to be able to carry out inference on our slope, we need additional conditions.

Condition 2: Observations must be independent.

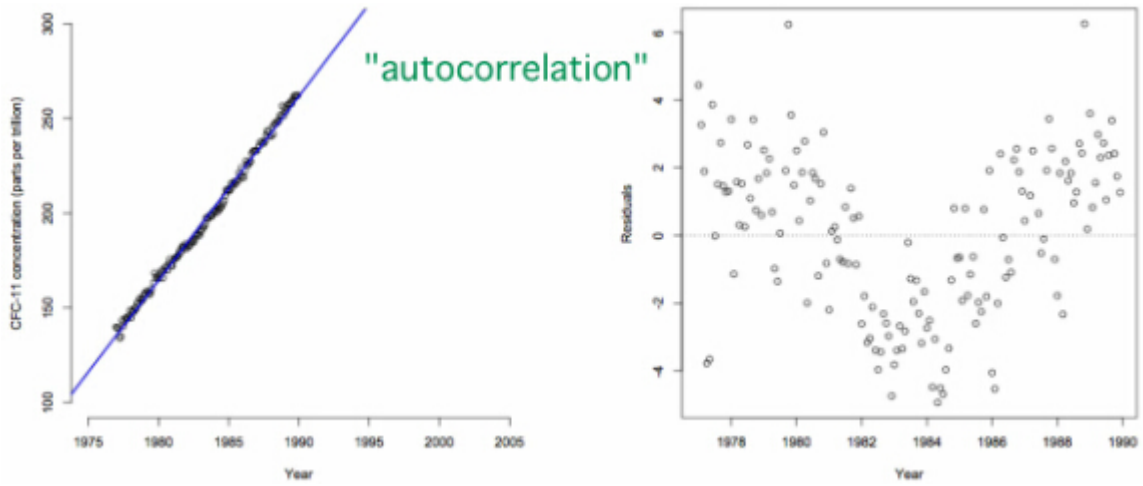


Figure 2: The CFCs example has non-independent observations.

This is the most important assumption, as there is no getting around it without moving to more sophisticated models, but it is also the most difficult to check. For example, if we were measuring the crawling age of babies, we wouldn't want to include twins as we'd expect twin babies to have similar crawling ages, so then they wouldn't be independent. The CFCs example shows another version of lack of independence. The curved pattern in the residual plot in Figure 2 is partly a consequence of the fact that measurements were taken over time, and measurements close together in time tend to be correlated with each other. This is called autocorrelation, and, when it occurs, observations above the line tend to have surrounding observations that are above the line, and observations below the line tend to have surrounding observations that are below the line. Knowing that the data were collected over time would help us catch this lack of independence. Generally speaking, the best defence against working with data that are not independent, is to understand the data and how they were collected.

In order to carry out inference on the slope using the t-test and confidence interval that we've studied, there are two more conditions that must be met.

Condition 3: The variation in the error terms is constant.

A typical violation of this condition looks like variability that increases or decreases with the

explanatory variable.

Figure 3 shows an example of increasing variability. The scatterplot shows the amount of carbohydrates in food items sold at a coffee shop on the vertical axis, and the number of calories in the food items on the horizontal axis. From the scatterplot, we can see a positive relationship, but it also shows increasing variability in the amount of carbohydrates with greater calories. This increasing variability is also seen in the residual plot, and like non-linear patterns, increasing or decreasing variability is often exaggerated and more easily seen in the residual plot. This pattern of increasing variance means that it would not be appropriate to carry out inference on the slope of the line for these data.

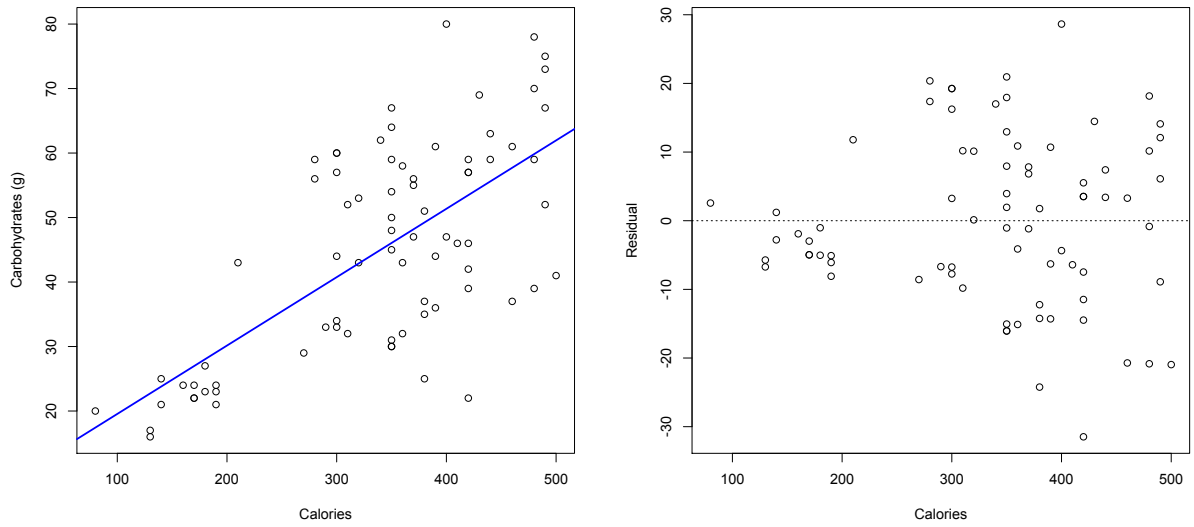


Figure 3: An example of non-constant variance.

Condition 4: The errors are normally distributed.

Typically, Condition 4 is checked by looking at a normal quantile plot of the residuals. In Figure 4, the plot on the left shows a typical example where the residuals appear to have a normal distribution, because the normal quantile plot of the residuals looks approximately like a straight line.. The curvature in the plot on the right is an example in which the residuals appear to have a skewed distribution. The methods we've learned for confidence intervals and statistical tests for the slope of the line would not be appropriate for the data that resulted in the plot on the right.

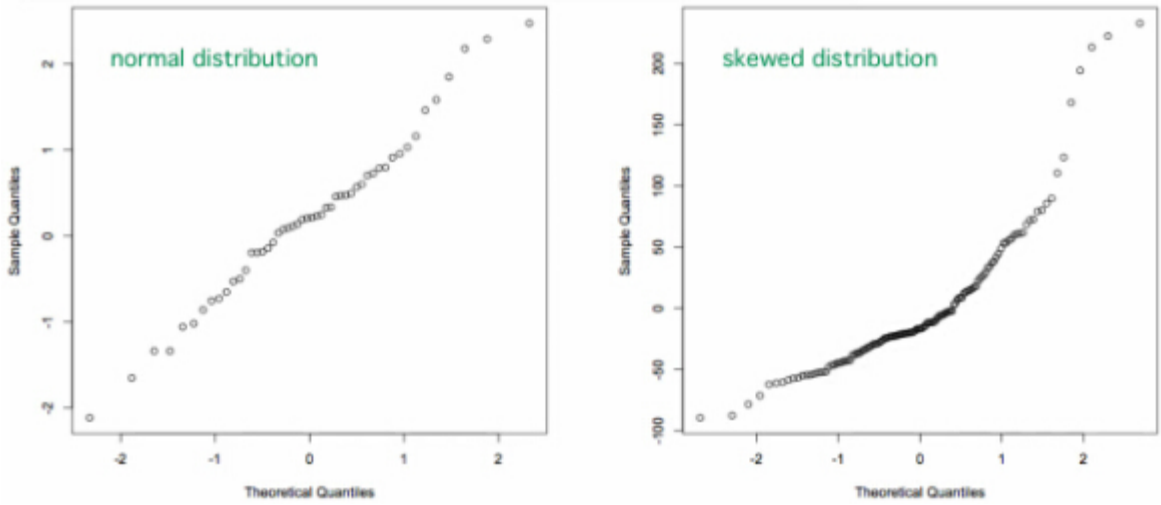


Figure 4: Examples of normal quantile plots.

Summary of Conditions:

To use linear regression:

Condition 1: A linear model is appropriate for the data. (no curvature, no influential points, no groups with different patterns).
To check Condition 1: Look at a scatterplot of the data. Follow-up a linear model with a plot of the residuals versus the predictor variable or the predicted values of the explanatory variable.

To carry out inference on the slope:

Condition 2: Observations must be independent.
To check Condition 2: Understand how the data were collected.

Condition 3: The variation in the error terms is constant.
To check Condition 3: Check the scatterplot and residual plot for signs of non-constant variance.

Condition 4: The errors are normally distributed.
To check Condition 4: Look at a normal quantile plot of the residuals.