Chapter 4

4.1

We will look more critically at regression analysis and consider the times when the results it is giving us may be misleading. We will discuss the different assumptions.

4.2

It is important to have an understanding of the types of scales that are used when making measurements. The type of measuring scale will determine the type of statistical tests that can be carried out on the data. We will discuss four different scales:

Nominal scales

Data measured on a nominal scale are often also known as categorical data. For example the sex of each participant is categorical data. This can be labelled in the data, as ‘1’ for males, and ’2‘ for females, and ’3’ for don’t know. In this case it is not appropriate to calculate the mean.

Ordinal scales

Ordinal data are collected when objects or individuals are put into a rank order. This data is sometimes called ordered categorical data. For example, an experiment that asks you to rank five TV programmes into an order of preference, so 1 signifies your most favoured and 5 signifies your least favoured.

However it could be that the difference in preference between TV programmes is not always equal. There might be little difference in preference between your rank 1 and rank 2, but there can be a big difference between your rank 3 and rank 4. So it can be said that ordinal scales do not necessarily have equal intervals.

Interval scales

Interval data represents a much ‘higher’ level of measurement than ordinal data. An example of an interval scale would be UK and US shoe sizes. The intervals between adjacent values in the shoe size scale are equal. So numbers of this scale can be added and subtracted (we can calculate a mean value). Interval scales do not have a true zero point. In addition, you cannot multiply the values on an interval scale.

Ratio scales

A ratio scale has all the properties of the interval scales. However this scale does have an absolute zero point, and therefore represents the highest level of measurement. An example of ratio scale is the measurement of response times in a psychology experiment.

Short summary of the different levels of measurements:

Scale	Operation	Examples
Nominal	Equal versus not equal	Telephone number
Ordinal	Monotonically increasing	Rank order in class test
Interval	Equality of intervals	Temperature (Celsius)
Ratio	Equality of ratios	Temperature (Kelvin)

In the context of regression, we make the following assumptions about the nature of our data: (a) the dependent variable should be measured on a continuous (interval or ratio) scale, and (b) the independent variable(s) should be measured on a continuous scale or if the independent variables are measured on categorical scales they can be uses, after a little recoding.

These assumptions are not applied very strictly, because if we would do that, we would not use the regression very often. One thing to consider is the number of categories your respondents are able to use, the more categories, the closer to an interval scale you are likely to be. Seven categories is considered to be the optimum number, but people who are used to responding to rating scales can often use more points while people who are not used to them may prefer to use fewer points.

If we conduct a regression analysis, we also make assumptions about the distribution of the dependent variable and residuals. If these assumptions are not satisfied, the conclusions we derive will not have a sound basis and will be incorrect.

The normal distribution is a very important distribution in statistics. For accurate calculations of standard deviations, and therefore of standard errors we must have a normal distribution. If the data are not normally distributes the standard errors (and therefore the significance tests) can be inaccurate.

4.4

In the case of univariate (only one independent variable) we assume that the residuals are normally distributed. For the calculations of the mean and standard error to be any use to us we must be able to assume that the data follow the pattern of a normal distribution. There are two related ways in which the data can fail to follow the pattern of a normal distribution:

The data contain a small number of scores that are extremely large or extremely small compared to the rest; these scores are termed outliers;

If there are outliers in a set of data, the mean may not be an appropriate way of representing these data.

The shape of the whole distribution fails to resemble the bell-shaped curve, which is characteristic of the normal distribution.

If there are no outliers, the distribution may still deviate from normality. The distribution can deviate in two ways:

The distribution can be non-symmetrical. This means that one tail of the distribution is longer than the other tail. We describe a non-symmetrical distribution as skewed.

Skew can occur for a number of reasons but it most commonly occurs when there is some sort of floor effect or ceiling effect. A floor means that the data have a minimum value. This can cause positive skew. A celling effect occurs when it is not possible to score above a certain upper limit. This would cause negative skew.

If the distribution of data is too flat or too peaked, that is the tails are to short or too long, the distribution is described as being kurtosed.

Kurtosis causes fewer problems in the estimation of regression models than skew. If the curve is too flat, the distribution is described as being negatively kurtosed or platykurtosed. If the distribution is too peaked, the distribution is described as being positively kurtosed or leptokurtosed.

There are different methods to detect non-normality. There are graphical and numerical methods. First we will discuss the graphical methods:

- Histogram. This is the easiest way of defining a normal distribution. Detecting skew, kurtosis, and outliers is easy using a histogram. If the data appear to be bell shaped and symmetrical then the distribution is probably approximately normal. You can be easily fooled by histograms, especially when samples are small.
- Boxplot. The first stage in drawing a boxplot is to find the median. This is the middle point when the points are ranked rom the highest to the lowest. Next stage is to mark the quartiles, in the same way as the median represents the halfway point the quartiles represents the one-quarter and three-quarter way points. The whiskers extend from the edge of the box to the highest and lowest points, unless those points are more than 1.5 times further from the central line than the edge of the box. The points that exceed this distance are called outliers.
- An advantage of the boxplot over the histogram is that for a smaller sample size random deviations from normality can make a histogram appear non-normal, but these few deviations do not have the same effect on the boxplot.
- A probability plot (P-P plot) is a more mathematical method. We know the types of scores that we would expect to get if our data where normally distributed. We can use this to compare our dataset with an imaginary dataset that would be found in an ‘ideal’ normal distribution with the same mean and standard deviation. If the distribution we are interested in matches the normal distribution fairly well, we can conclude that our data are normally distributed.

4.5

Besides the graphical methods there are also calculation-based methods:

Skew and Kurtosis. These values will have a value of 0 when there is a normal distribution. If the value of skew or kurtosis is greater than twice the standard error, then the distribution significantly differs from the normal distribution. If the skewness statistic is less than 1.0 there should be little problem. If the skewness is greater than 1.0, but less than 2.0, you should be aware that this might be having an effect on your parameters estimates, but that it is probably OK. If the skewness statistic is greater than 2.0 you should be concerned. This method is useful for detecting general deviations from the normal distribution. However they are not good in detecting outliers.

If we have a normal distribution, we can find out what proportion of data points we expect to find at different distances from the mean. We know from the tables that 68% of scores will lie within one standard deviation from the mean, 95% will lie within two standard deviations of the mean, and 99% will lie within three standard deviations of the mean. We can convert a score into what is called a standardised score, or a z-score, by calculating the number of standard deviations from the mean that any score lies. We do this by this equation:

$$z(x) = \frac{x - \bar{x}}{sd} \qquad (103)$$

It follows that if we have a dataset that contains the z-scores of a normally distributed measure of a sample of 100 people, we would expect to find several scores whose z-values were greater than two and approximately one that had a z-score greater than three standard deviations.

Using the standardised scores to detect outliers is not a wholly satisfactory method. A better method is to use the deleted z-score. This is the z-score of that data, but instead of using the mean and standard deviation of all the data points, we use the mean and standard deviation of all of the data points except the one that we are interested in. To calculate the deleted z-score, the first case is deleted, the mean and standard deviation for the rest of the data are calculated, and this mean and standard deviation are used to calculate z-scores from the deleted case. The process is then repeated.

Influence statistics are rarely used in the univariate case to determine if a data point should be considered an outlier, but because influence statistics are used in regression we will discuss it. The aim of this is to see how much each individual data point influences the model parameters, even if the only parameter of interest is the mean. An influence statistic is calculated by: (1) calculating the parameter estimate for all the variables; (2) recalculating the parameter estimate with the one data point excluded; (3) calculating the difference between results 1 and 2.

For this method, we encounter the problem of the variables having different and arbitrary scales. So we need to use standardised scores. Standardised scores convert the mean to zero and the standard deviation to one, and they make the expected change statistic much easier to interpret.

4.6

If the distribution is not normal, either because of outliers or because of skew, or kurtosis, least squares estimates and their standard errors will be inaccurate.

To decide whether a data point is an outlier and deciding what action to take is much more difficult. The first thing to do is to try to determine why the outlier has occurred. It can be caused by a faulty measurement, or error while entering data. If you cannot find the correct value you may simply want to delete the data point and carry on without it. If the outlier has occurred because of a true, properly measured data point, then you need to look at both your theory and your measuring scales to see if they are appropriate. If the outlier is still there after you have checked your sources of data and reassured yourself that your theory is the appropriate one, you are left with two options. The first is to carry out the analysis with the outlier and be aware that it may be having an undue influence on you parameter estimates. The second option is to delete the outlier.

Points that are further away from the mean will have a greater influence on the results of calculations than nearer ones, therefore the mean will prove to be a biased estimator. In a negatively skewed distribution, the mean will be biased downward; in a positively skewed distribution, it will be biased upward. Positive skew is more common than negative skew.

When the data are kurtosed, the mean value of a variable remains the same, but the standard errors of the mean can become too small or too large. The calculations we used for the standard error only give the correct value when the distribution is normal. The distortion that occurs in a non-normal distribution will have one of the two effects: (1) it may have the effect of making us believe that our estimates are less accurate than they really are, the standard error will be too large. So the type II error rate will be inflated and we will be less likely to find significant effects. If standard errors are too small, type I error rate will become inflated, and we will think our estimates are more accurate than they actually are; (2) more complex case, kurtosis can affect the parameter estimates as well as the standard errors, but the effects are usually small unless the kurtosis is severe.

Transformations are a way of taking data that is not normally distributed to make them fit a normal distribution. A transformation is a calculation that is done to all of the values of a variable together. A commonly used transformation for removing positive skew is the log transform. The ‘unlogging’ or ‘antilogging’ is done using the exponential function in a statistics package.

If your data are negatively skewed, squaring each of the values can make them form a more normal distribution.

4.7

Until now we have looked at the univariate distribution. We will now look at the more complex multivariate distributions. This distribution is a little more complex, however the principles remain the same. A multivariate distribution is a distribution that contains more than one variable. In the previous chapter we have looked at the mean, and slope coefficients, which are both least, squares estimators. The assumptions made of the multivariate distributions are as follows:

At each value of the dependent variable, the distribution of the residuals is normal.

The variance of the residuals at every set of values for the independent variable is equal. This assumption that the variance is equal is called homoscedasticity (and if it should happen that the variance is not equal i.e. our assumption of equality is not satisfied, then that condition is called heteroscedasticity).

At every possible value of the dependent variables, the expected (mean) value of the residuals is equal to zero. In bivariate cases, this assumption means that the relationship between the independent variable and the dependent variable should be linear.

For any two cases, the expected correlation between the residuals should be equal to zero. This is referred to as the independence assumption, or a lack of autocorrelation.

We will now take a closer look at each of the assumptions.

Assumption 1

The first way to check this assumption is to examine whether the distribution of residuals is approximately normal. If the distribution differs from normality, this assumption is violated. Outliers, skew and/or kurtosis can cause this non-normality.

When we discussed the mean, we saw that the residuals can be raw or standardised, and that the residuals can be calculated from a dataset in which a particular individual case was deleted. Residuals can be treated in the same ways, in case of multivariate distribution.

There are different types of outliers there exist:

- Unstandardized residuals (RESID). These are the 'raw' residuals. They are simply the difference between the predicted value and the actual value for the particular cases.
- Standardised residuals (ZRESID). If we do now know the underlying scale of a variable, it is difficult to interpret the residual. The solution using the mean was to standardise the scores to give the mean of 0 and a standard deviation of 1.
- Studentized residuals (SRESID). These residuals make a correction based on the estimated variance of the residual at that value of the predicted value.
- Deleted residuals (DRESID). When calculating residuals we are interested in the difference between the predicted value and the actual value. When we calculate the predicted value, we include the (potential) outlier. This means the outlier has an influence on the predicted value. Because outliers may have undue influence, deleted residuals, and studentized deleted residuals are calculated for each case, based on the predicted value if that case were excluded from analysis.

Raw residuals are hard to interpret, because the measurement scale of the dependent variable affects these.

Outliers do not always have an influence on the results of our regression analysis. Cases that do have an effect on the outcome of the calculation are referred to as influential cases. We will now consider two general types of outlier detection in statistics, these are distance statistics and influence statistics. We will look at three types of distance statistics: leverage; Mahalanobis distance; and Cook's D. And four influence statistics: DfBeta; standardised DfBeta; DfFit, and standardised DfFit.

Distance statistics

Leverage

This is calculated using only the values of the independent variables, so it is possible that an influential case may not be detected solely be examining leverage statistic. The calculating for the leverage statistic (called h). Leverage values have a maximum possible value of 1.

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})}{\sum x^2} \quad (104)$$

There is a rule of thumb whereby a cut-off value for leverage is given by:

$$\frac{2(k+1)}{N} \quad (105)$$

Where k is the number of independent variables and N the number of participants.

Mahalanobis Distance

This distance is calculated from the leverage value. The advantage of this value is that it is possible to use the distances as a value with a known distribution, which can then be tested for significance. It is calculated by:

$$MD_i = (N - 1)\left(h_i - \frac{1}{N}\right) \quad (106)$$

Cook's D

The previous two values only used the values of the independent variables for the calculations. Cook's D uses both the independent and dependent variables. Cook's D uses both the value of the studentized residuals and that of the leverage statistic to calculate a distance. The equation is:

$$D_i = \left(\frac{SRESID_i}{k+1}\right)\left(\frac{h_i}{1-h_i}\right) \quad (107)$$

Influence statistics

DfBeta and standardised DfBeta

DfBeta for a case is the difference between the value of beta when the case is included, and the value of beta when the case is excluded. These values are referred to as dfbetaij. Where the subscript i refer to the case number and j refers to the parameter estimate. A common recommendation is that a cut-off of:

$$\text{DfBeta} > \left|\frac{2}{\sqrt{N}}\right| \quad (108)$$

This can be used to determine whether a case is influential.

DfFit and standardised DfFit

This value is very similar to DfBeta, but rather than looking at the change in the parameters estimates that occur as a result of excluding a case, it examines the change in the predicted value of a case, when that case is excluded. To calculate DfFit the model is estimated using all of the cases, and then estimated again, with the first case excluded. This is repeated for every case. The difference in fit is calculated as the change in the predicted value that occurs when the case is excluded. The rule of thumb that should be used is to examine cases that have a cut-off of:

$$\text{DfBeta} > \left|\frac{2}{\sqrt{N}}\right| \quad (108)$$

Assumption 2

Heteroscedasticity is spotted using the same type of residual scatterplots that we examined when looking for outliers. We are interested in the variance of the residuals at each level of the predicted values. We want to know whether the variances of the residuals, at each predicted value of the dependent variable are equal.

Violation of the heteroscedasticity assumption, from statistical viewpoint, is not as serious as violation of certain other assumptions. The parameter estimates of a heteroscedastic dataset will not be altered, in much the same way that the mean is not altered if a distribution is now skewed, but is positively or negatively kurtosed. The standard errors of the estimates will be inaccurate, and therefore any calculation of significance will be wrong to some extent. The presence of heteroscedasticity means that the model has been miss-specified. We said that from a statistical viewpoint heteroscedasticity is not such a big problem, however from a theoretical standpoint it is. The presence of heteroscedasticity means that there is a more complex relationship between the variables than we have modelled.

Assumption 3

The expected value for a residual is the value that you would expect that residual to have if you have no other information. This assumption can be violated for two reasons:

- There is a ceiling effect. You can reach a ceiling where it is not possible to get more. (Attempting to make predictions beyond a reasonable range is called extrapolation, and can lead to this type of error)
- Violation of this assumption can occur when the effect is fundamentally non-linear. That is, it has a different effect at different values.

Assumption 4

This is the most difficult assumption that we will deal with in regression analysis, and the one, which tends to be ignored more than the others. It is hard to detect whether this assumption is violated. Autocorrelation occurs when a variable correlates with itself: If the cases are auto correlated, then they are related to one another, they are not independent. There are two occasions when this can occur:

In a time-series design. Where multiple measures of the same entity are assessed. This situation is referred to as autocorrelation.

Occurs when units of analysis can be grouped or clustered in some way, often by geographical area; this situation is referred to as non-independence.

This is the more common type of problem. This is for example a big problem when you have data of students, which are clustered in classes. When you fail to consider the structure of the data, we have found that the correlation has been reversed. These types of data are referred to as being hierarchical or multilevel data.

If violation of this assumption is suspected, it can be difficult to determine whether it has actually occurred.