Chapter 2: Looking at Data - Relationships

Introduction

In analysing our data, we often look for relationships between variables. The focus of this chapter is on learning how these relationships can be graphically and numerically described.

2.1: Relationships

We use the term *associated* to describe the relationship between two variables. An example is the relationship between weight and height.

- Two variables are associated if a value of the first variable says something about the value of the other variable.

Describing Data

- A *response variable* is related to the outcome of an investigation. A researcher wants to know whether height affects weight. In this case, weight is the response variable.
- An *explanatory variable* is the variable associated that declares or causes changes in response variables. In our example, the length is the explanatory variable.

A description of the main characteristics of a data set should include the following:

- Cases: Identify the cases an how many are in the dataset.
- Labels: Identify which variable is used as variable label (if any).
- Quantitative or Categorical: Classify each variable as categorical or quantitative.
- Values: Identify the possible values for each variable.
- Explanatory or response: When applicable, classify each variable as explanatory or response variable.

Many researchers are interested in investigating how explanatory variables cause changes in the response variables. Many relationships between these variables, however, will not have a direct form of causality.

Often, explanatory variables are also called *independent* variables. Response variables are called *dependent* variables.

2.2: Scatterplots

Relationships between two quantitative variables are often idisplayed in a scatter plot.

- The two variables need to be measured at the same individuals.
- The values of one variable are put on the X-axis, while the values of the other variable are put on the Y-axis. Each individual in the data is processed as a point in the graph, on the basis of the scores achieved by the person on the X-axis and the Y-axis.
- The explanatory variable corresponds to the X-axis. For this reason, the explanatory variable is also referred to as the X-variable. The response (Y) variable will be put on the Y-axis.
- If there is no distinction between explanatory and response variables, then it does not matter, which variable ends up on which axis.
- Time plots are a special type of scatterplot, that uses time as an explanatory or x-variable.

Interpreting Scatterplots

To get a first impression of a scatter plot, it is useful to:

- Look at the general pattern and deviations.
- Describe the shape, direction, and the strength of the relationship.
- Identify any outliers: These are individual values that fall outside the general pattern.

The relationship between two variables can be positive or negative.

- Two variables are positively associated when high scores on one variable associated with high scores on the other variable. An example is that a high score in height is often associated with high scores in weight.
- Two variables are negatively associated when high scores on one variable are associated with low scores on the other variable. For example, there is a negative correlation between test anxiety and performance on an exam. The more test anxiety, the lower the exam score.

The strength of a relationship is determined by looking at the degree to which points on the graph follow a specific form. Scatter plots can take on many forms and shapes. Many scatter plots show linear relationships; the points lie on a straight line.

Sometimes the points will be on a curve rather than a straight line.

To display a straight line rather than a curve, you can *transform* the data.

The most used transformation is the *log transformation*. Then it is necessary to use only positive values. A logarithm is a mathematical function, it is the exponent that a fixed number need to be raised with to get a certain number as a result. Most statistical software and more extensive calculators have a button that you can easily use to calculate a logarithm.

To add a categorical variable to the scatterplot, it is handy to use different colors or symbols for each category.

Often times, software is used to fine a line of best fit for your data, allowing you to see the overall form of your distribution. To display one curve rather than several smaller curves, you can *smooth* the curve.

Scatterplots show the relationship between two quantitative variables. However, sometimes it can be necessary to analyse the relationship between a categorical variable and a quantitative variable. In that case it is necessary to make a comparison between the distributions for each category.

2.3 Correlation

In short, the scatterplot of a distribution describes the shape, direction, and strength of a relationship between two quantitative variables. It can be misleading to make statements about the strength of this relationship with the naked eye. By changing the numbers on the axes, any distribution can appear to have a strong correlation, while it might not necessarily be the case. The reverse is also possible. For this reason we use the correlation measure.

- The correlation measures the direction and the strength of a linear relationship between two quantitative variables. Often, the letter r is used to describe the correlation.
- Suppose we have collected data for variables X and Y for n number of people. The averages and standard deviation of the two variables are then $\bar{x}$ and s_x for the x-values and y-bar and s_y for the y-values.
- The correlation, r, between X and Y is:

$$r = \frac{1}{n-1} \Sigma \left(\frac{X - \overline{X}}{s_x} \right) \left(\frac{Y - \overline{Y}}{s_y} \right)$$

In this formula the meaning of the symbols is as follows:

- r = correlation between X and Y
- n = the total number of persons
- Σ = sommation/sigma. Add everything that is behind the sigma.

- $\bar{x}$ (X with a bar above) = the average of X
- Y-bar (Y with a bar above) = the average of Y
- s_x = the standard deviation of X
- s_y = the standard deviation of Y

By using this equation, all of the values for the X and Y variables will be standardised.

Characteristics of Correlations

- *r* is negative when there is a negative correlation is and positive when there is a positive correlation.
- Correlation does not distinguish between explanatory and response variables. It doesn't matter which variable you use for X and which one for Y.
- To calculate a correlation both variables must be qualitative.
- Since r makes use of standardized values, the correlation does not change as the units of measurement of X, Y, or both do. For example, measuring height in inches or meters and measuring weight in kilograms or pounds does not change the correlation.
- The correlation r itself does not have a unit of measure; it is only a number.
- The correlation r is always a number between -1 and 1. Values close to 0 indicate that there is a very weak relationship. The strength of a relationship increases as r approaches -1 or 1. A correlation of -1 or 1 is rare and is extreme. In such cases, all of the points lie exactly on a straight line.
- Correlation measures only the strength of the *linear* relationship between two variables. Correlation does not describe the *curve* of the relationship between variables.
- Just as the mean and standard deviation, correlation is not robust: r is strongly influenced by very few scores.
- To adjust the scale of a scatterplot can be misleading, it does not change the standardised values of the variables, and it does not change the correlation.
- Correlation is never a complete description of data with two variables. You also need to look at the average and the standard deviation.

2.4 Regression

Regression Lines

If a scatterplot shows that there is a linear relationship, we often want to find the *line of best fit* to describe this relationship.

- A regression line is a straight line that describes how a response variable Y changes as explanatory variable X changes.
- We often use a regression line is to predict the value of Y is for a given value of X. For regression, in contrast to correlation, however, it is important that we have specific explanatory and response variables.

Finding a Line of Regression

There is often no straight line that passes exactly through all points of the scatter plot. Therefore, we have to depend on a *line of best fit*, or a line that is closest to the most values in your distribution. Suppose that Y is a response variable on the vertical axis and that X is an explanatory variable on the horizontal axis. A straight line that shows the relationship between X and Y has the equation:

- $Y= b_0+ b_1x.$, where b_1 is the regression coefficient (slope) and is b_0 is the Y-intercept.
- The *regression coefficient (slope)* is the value with which Y changes when X increases with 1 unit. The slope gives information about how much the whole is subject to change when the values are changed. This indicates whether a graph is steep or not.
- The *intercept* is the value of Y when X is zero.

Extrapolation

Extrapolation is the use of a regression line in order to make predictions that are far beyond the examined values. For example, you can create a scatter plot based on height and weight scores of a group of people. The tallest person may be, for example, 1.80 m. By using an extrapolation, you could predict how much a person who is 1.95 m tall would weigh.

Least-Squares Regression

We use the least squares regression to find a line that will allow us to predict a value for variable Y, based on variable X. It is important to note that in using the least squares regression, there is always a degree of error:

- Error = observed score - predicted score. Errors are positive if the value of Y is above the regression line and negative if they are below.

- The least squares regression line of Y to X is the line that that is formed when the sum of the squares of the distances between data points and the regression line is as small as possible. To make this regression line, we must first find the values of b_0 and b_1, that give the lowest error: $\Sigma (error)^2 = \Sigma(y_i - b_0 - b_1x_i)^2$
- Often, this line can be found by using computer programs; however, it is also possible to calculate the regression line manually: $\hat{y} = b_0 + b_1x$.
- The value of b_1 is found with the formula $b_1 = (S_y / S_x)$
- So to calculate b_1, first you devide the standard deviation of Y by the standard deviation of X, then you multiple this with r.
- The value of b_0 is found by the formula: $b_0 = \bar{y} - b_1\bar{x}$.

Interpreting the Regression Line

The slope and intercept of the least-squares regression line are very dependent on the sort of measurement units that is used. When only the size of the slope and intercept are known, but the measurement unit is unknown, there nothing can be concluded.

When using software, take a close look at which information you need and which information you do not need. At the point where you understand the statistical method, you can read output from almost every kind of software, so understanding the method is the first step.

Properties of Least-Squares Regression

least-squares regression is the most used method to apply a regression line to data. This method has the following features:

- Correlation and the slope of the least-squares regression line depend on each other. A change in the standard deviation of X results in a change of r standard deviations in Y.
- The least-squares regression line always goes through the point $(\bar{x} , \bar{y})$ on a Y/X graph.
- The distinction between explanatory variables and response variables is important to regression. These two should not be switched in position, because then the regression line will look differently. If you want to measure increased fat gain as a result of activities, then fat is the response variable, this is projected on the Y-axis (vertical axis).

Proportion of Explained Variance (r^2)

The square of the correlation coefficient tells us how close the data are to the fitted regression line. If a correlation is -1 or 1, then r^2 will be exactly 1.

r^2 can also be seen as the variance of the predicted scores ($\hat{Y}$), divided by the variance of the observed values (Y).

Data mining

Explanatory data analysis (EDA) is a concept that implies analysing and interpreting data with scatterplots, regression etc. This can also be used for huge amounts of data and very large databases, then it is called *data mining*. How a database is structured and how data can be processed into it, is called *data warehousing*. For data mining it is important to use efficient algorithms, structure the data in a clear way and use more automated methods for analysis.

2.5 Limitations of Correlation and Regression

Residuals

Even with the best possible regression line, not all of the points lie precisely on the line. Some items might therefore not be well predicted on the basis of the regression line. The points that deviate from the regression line are called residuals.

- A residual is the difference between an observed value of a response variable and the predicted value in accordance with the regression line: $y - \hat{y}$. The average of all residuals is always 0.
- A residual-plot is a dot plot of all regression residuals with respect to the explanatory variable. With such a plot, it can be determined whether a regression line fits well. If the regression line fits the general pattern of the data, no patterns will be present in the residuals. An outlier is an observation that is far from the overall pattern of a residual plot.
- Items that are outliers in the Y direction of a scatter plot have large residuals, but this does not necessarily apply to other residuals.
- A score is *influential* for a statistical calculation if removing it would lead to a major change in the calculation. Outliers in the X-direction often have an impact on the least-square regression line.
- The least-squares regression line is, like the correlation, not robust.

Lurking Variables

The relationship between two variables can often be best understood by also looking at the effect of other variables. Lurking variables can make a correlation or a regression misleading.

- A lurking variable is a variable that is not included in the study as an explanatory or response variable, but may affect the interpretation of the relationship between these variables.

Other Things to Remember about Correlation and Regression

A (strong) relationship between an explanatory variable (X) and a response variable (Y) is not evidence that X *causes* changes in Y. Correlation says nothing about causality. In addition, it is important to be careful when working with regressions of averaged values. Correlations based on averages are often much higher than correlations based on individual scores. In some cases, it is also important to be aware of the *restricted-range problem,* as the range of the explanatory variable can greatly effect its relationship to the dependent variable. In that case, the correlation will be (r), and the proportion of explained variance (r^2) be lower than if all possible scores for the data were taken into account. Investigators often make use of several explanatory variables. A high score on a math test (Y) may, for example be related to natural ability, but also to motivation and education. When an investigator uses multiple explanatory variables, he or she must use multiple regression, where the correlation between all related variable can be calculated.

2.6 Data Analysis for Two-Way Tables

Categorical Data

As previously discussed, scatterplots are used for quantitative variables. Qualitative data, on the other hand, is analysed using two-way tables. Examples of categorical variables are gender and occupation. A two-way table shows how often each combination of category pairs occurs. For example, you might want to find out, out of a population of men and women, how many of each sex are psychologists and how many are doctors. This leaves you with 4 combinations of interest: male doctors, male psychologists, female doctors, and female psychologists. To describe the relationship between two categorical variables, we calculate different rates, for example, the percentage of male doctors or the percentage of female psychologists. Proportions for each categorical pair (female doctors, mal psychologists, etc) can be calculated by simply dividing the value of each categorical pair and dividing by the total number of individuals. The total collection of these proportions is called the *joint distribution* for the two variables.

Marginal and Conditional Distributions

In addition to a joint distribution it is also possible to calculate *marginal distributions.* Marginal distributions are the distributions of a single variable in a two-way table. A *conditional distribution*, however, gives more information than individual marginal distributions. Conditional distributions are when you put a condition on one variable and calculate the distribution of the other variable. An example of a conditional distribution is if you wanted to find out how many doctors are men and how many are women.

Two-way tables are a compact way to show a lot of information, the first step to make a two-way table is deciding which percentages you want to show.

Mosaic plot

A mosaic plot shows a distribution in (usually) four boxes. The way it is given form is similar to a bar graph, but each bar is divided in two parts, the part of the population for which a certain variable gathers positive results and the part for which it gathers negative results. A mosaic plot can be used to display both marginal and conditional distributions in a clear way.

Simpson's Paradox

As with quantitative variables, lurking variables can also affect the relationship between two categorical variables.

- A relationship or equation that applies to all groups can change direction when the data are combined into a single group. This change of direction is called Simpson's Paradox.

To make tree-way tables, it is necessary to combine the results for three variables. This process is called *aggregation*.

2.7: Causation

Correlation

Correlation only says something about the degree to which two variables are related. As discussed above, even a strong correlation between two variables does not prove causality. If we see that anxiety is associated with lower school grades, we can not (yet) conclude that fear of failure is the cause of the low grades.

- If variable X causes Y variable, there is causality ($X \rightarrow Y$). Causality can be discovered through experimentation. In that case, values of variable X may be varied in order to investigate their effect on Y. Other factors are held constant. This is to minimize the influence of lurking variables as much as possible.

- It is also possible for X and Y appear related because they are both influenced by another variable, namely variable Z. This is also called a *common response*. Variable Z, in this case, is a lurking variable. The observed correlation between X and Y is therefore misleading.
- Two variables are "confounded" if their effects on a response variable cannot be distinguished from one another. These variables can be explanatory variables, lurking variables, or both.

Establishing Causation

Sometimes it is not possible to discover causality by means of experiments. For example, it would be unethical to force people to smoke more than they usually would, in order to see if the increased amount of smoking causes cancer. Research shows that smoking is often associated with cancer, but here, we still cannot conclude that smoking causes cancer. This is because we did not run our own experiments. So how can causation be established without doing any experiments?

- There must be a strong correlation between variable X (smoking) and Y (cancer).
- The connection must be consistent. Studies in different countries, for example, shows that smoking and cancer often go together.
- It must also show that high doses go together with stronger reactions. People who smoke more are more likely to get cancer.
- The suspected cause must precede the effect. For example, lung cancer is discovered only after many years of smoking.
- The suspected cause must be plausible. Animal studies, for example, shows that cigarette smoke causes cancer, and it would not be unreasonable to say the same thing about humans.