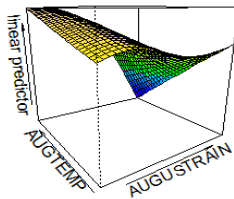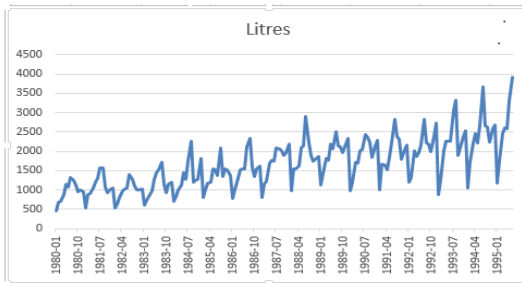# Straightforward Statistics using Excel and Tableau
Stephen Peplow PhD

Straightforward Statistics with Excel and Tableau

A hands-on guide with screencasts and data

Stephen Peplow

Tweet This Book!

Please help Stephen Peplow by spreading the word about this book on Twitter!

The suggested hashtag for this book is #PeplowStephen.

Find out what other people are saying about the book by clicking on this link to search for this hashtag on Twitter:

https://twitter.com/search?q=#PeplowStephen

Contents

1. Why I wrote this book

I began business life as an entrepreneur in business in Hong Kong. I ran trade exhibitions, imported coffee from Kenya, and started and operated two restaurants, one of which (unusually for Hong Kong) served vegetarian food. These businesses were profitable, but I would have saved myself a great deal of stress, and done better if I had used some business intelligence informed by statistics to improve my decision-making. Looking back, I wish I had been able to think through and write analyses on topics such as these among many others:

a. calculation of optimal restaurant staffing levels
b. analysis of sales over time and seasonality in trends
c. receipts per customer by restaurant type and analyzed strength and type of any differences
d. predicted sales data for potential exhibitors with visualizations of various 'what if' scenarios
e. gained deeper insights from visualizing my data
f. won over more clients and investors with better visual and written presentations

In this book we'll explore cases such as these, and also do some more general work so that you can carry the work into other questions.

Most books on stats introduce statistical techniques on a chapter by chapter basis. Instead, I've written this book so that it reflects the way many people learn. The chapters are structured by the type of question you might want to ask. Examples of the type of questions are summarized below, and at the beginning of each chapter. The chapters themselves doesn't include much math and

technical details. Instead these are placed towards the end of the book in a glossary. As a result, this book is useful both for readers wanting to answer specific questions and also for students trying to get through an introductory stats class.

Many of the Excel procedures are linked to screen casts prepared by me to illustrate that particular procedure. The data-sets used in the book are available from my Dropbox public folder: just click on the hyperlink that appears next to each worked example. With the data-sets open in Excel, you can follow along at your own pace. (And then do the same with your own data). I have used Tableau for some visualizations, and you will find links to my workbooks and screen casts.

1.1 The key points in stats

It is easy to get swamped in stats with so many definition and equations. but basically it all comes down to five key areas. If you can get these down, then the rest will follow easily. The key points which you should try to get first off are these:

1. difference between a sample and a population
2. notation (that is the symbols and letters we use to represent operations and unknown quantities)
3. meaning of variance and standard deviation
4. types of variable
5. hypothesis testing

These five key points make up the next five chapters of the book, and we will keep referring back to them. The rest of the book is as follows:

1.2 Chapter descriptions

If you're reading this book, then you are very likely already engaged in business and would like to know how to take that business to new heights. Take a look through the chapter descriptions that follow and then go straight to that chapter.

If you are taking a stats course and think you might need a bit of a statistics refresher, look through the Glossary in Chapter 20 first. A very good free statistics OpenIntro textbook is available here[1]

Chapter 7 introduces Tableau Public[2] which is a free data visualization tool. While Excel does have graphing tools which are easy to use, the results can look a little clunky. Tableau helps us to merge data from different sources and create remarkable visualizations which can then be easily shared. I'll be illustrating results throughout this book with either Tableau, Excel or both. One caution: if you publish your results to the web using Tableau Public, your data is also published. If this is a problem, there is an option to pay for an enhanced version of Tableau. The screenshot below shows work done changes on land use in the Delta region of British Columbia. By clicking on the link, you can open the workbook and alter the settings. You can filter the year (see top right) and also land use type. Thanks to Malcolm Little for his work on this project. Tableau workbook for Delta Land Cover Changes[3]

---

[1] https://www.openintro.org/stat/

[2] http://www.tableausoftware.com/public/

[3] https://public.tableau.com/views/DeltaCropCategories1996-2011/Sheet1?:embed=y&:showTabs=y&:display_count=yes
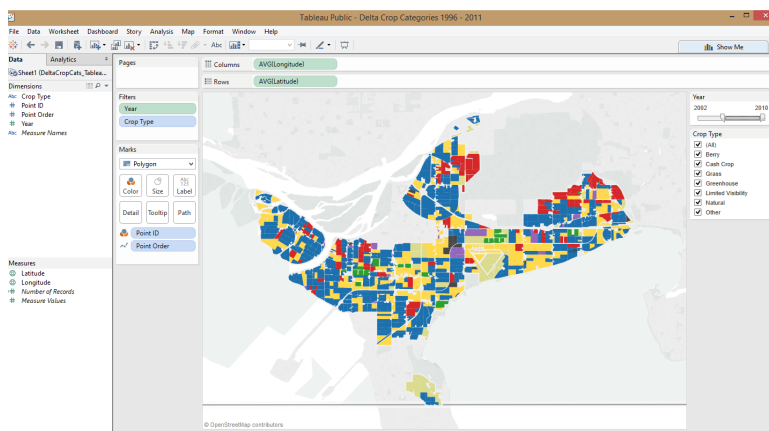
Tableau showing changes in landcover

Tableau has training and demonstration videos available on its website, and there are plenty of examples out there. The screen casts which Tableau provides (available at their homepage) are probably enough. Where I have found some technique (such as boxplots) particularly tricky I have created screen casts for this book. The image below shows results from data we will use in the regression chapter. You run a trucking company and from your logbooks extract the distance and duration and number of deliveries for some delivery jobs. The image shows a trend line, with distance on the hoirizontal axis and time taken on the vertical axis. The points on the scatterplot are sized by number of deliveries. You can see that more deliveries increases the time, as one would expect. Using regression, Chapter 6, we will work out a model which can show how much extra you should charge for each delivery. Here is a Youtube for trucking scatterplot[4]

---

[4]https://youtu.be/oFACLZLJWZI
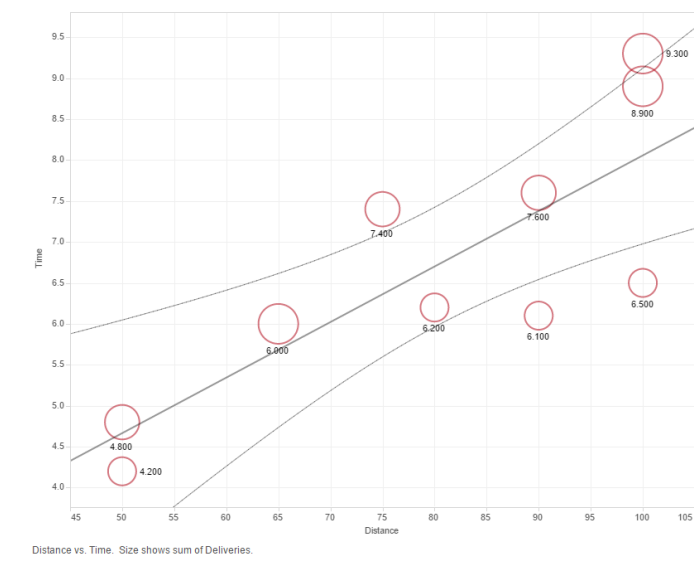
Tableau showing distances, times, and deliveries

Chapter 8: Data and how to get it. If you are considering collecting data yourself, through a survey for example, then you'll find this chapter useful. So-called **big data** is a hot topic, and so I include some discussion. The chapter also includes links to publicly available data sets which might be helpful.

Chapter 9 deals with tests for whether two or more quantities are the same or different. For example, you are a franchisee with two or more coffee-shops. You want to know whether daily sales are the same or different, and perhaps what factors cause any difference that you find. You want to know whether there is a statistically significant common factor or not, to eliminate the possibility that the difference you see occurred purely by chance. Perhaps the average age of the customers makes a difference in the sales? ANOVA uses very similar theory to regression, the subject of the next chapter and perhaps the most important in the book.

Chapter 10: Regression is almost certainly the most important sta-

tistical tool covered in this book. In one form or another, regression is behind a great deal of applied statistics. The example presented in this book imagines you running a trucking business and wanting to be able to provide quotations for jobs more quickly. It turns out that if you have some past log-book data to use, such as the distance of various trips, the time that they took and how many deliveries were involved, an accurate model can be made. Regression can find the average time taken for every extra unit of distance, as well as other variables such as the number of stops. This chapter shows how to create such a model and how to use it for prediction. Using a model such as this, you can make quotations really quickly. This chapter also covers more complex regression and model-building techniques.

Other uses: you want to calculate the beta of a stock, comparing the returns of one particular stock with the S&P 500.

Chapter 11 is about testing whether the regression models we built in Chapters 6 and 7 are in fact any good. Regression appears so easy to do that everybody does it, without checking the validity of the answers. Following the procedures in this chapter will help to ensure that your work is taken seriously. This chapter is more of a guide to thinking critically about the results other people might have. Is there missing variable bias? For example, you notice a strong correlation between sales of ice-creams and deaths by drowning. Can you therefore say that reducing ice-cream sales will make the water safer? Err–no. The missing or lurking variable is temperature. Both drowning deaths and ice cream sales are a function of temperature, not of each other.

Chapter 12 is about time-series and shows how we can detect trends in data over time, and make predictions. The smoothing methods, such as moving averages and exponentially weighted moving averages covered in this chapter are fine for data which lacks seasonality and when only a relatively short-term forecast is needed. When you have longer-run data and for making predictions, the following

chapter has the techniques you need.

Chapter 13 describes the regression-based approach to time series analysis and forecasting. This approach is powerful when there is some seasonality to your data, for example sales of TV sets show a distinct quarterly pattern (as do umbrellas!).
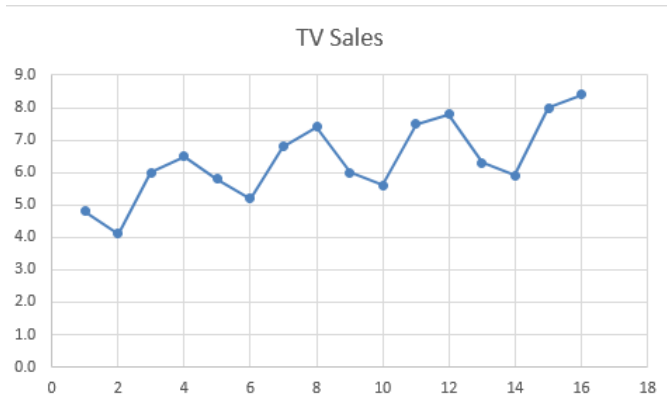
Sales of TV sets showing a quarterly seasonality

Using regression we can detect peaks and troughs, connect them to seasons and calculate their strength. The result is a model which we can use to predict sales into the future.

Chapter 14 is about optimization or making the best use of a limited number of resources. This is highly useful in many business situations. For example, you need to staff a factory with workers of different skills and pay-levels. There is a minimum amount of skills required for each class of worker, and perhaps also a minimum number of hours for each worker. Using optimization, you can calculate the number of hours to allocate to each worker.

Chapter 15 Optimization can also be used for more complex 'blending' problems. Example: You run a paper recycling business. You take in papers and other fibers such as cardboard boxes. How can you mix together the various inputs so that your output meets minimum quality requirements, minimizes wastage, and generates

the most profit?

Another example: what mix or blend of investments would best suit your requirements?

Chapter 16 concerns calculating the probability of items you can count one by one. For example, what is the probability that more than five people will come to the service desk in the next half-hour? What is the probability that all of the next three customers will make a purchase? We can solve these problems using the binomial and the Poisson distributions.

Chapter 17 concerns choice under uncertainty: if you have a choice of different actions, each of which has an uncertain outcome, which action should you choose to maximize the expected monetary value? A farmer knows the payoffs he/she will make from different crops provided the weather is in one state or another (sunny/wet) but at the time when the crop decision has to be made, he obviously doesn't know what the actual state will be. Which crop should he plant? A manufacturer needs to decide whether to invest in constructing a new factory at a time of economic uncertainty: what should he do?

Chapter 18 adds the decision-maker's risk profile to his or her decision process. Most people are risk averse, and are willing to trade off some risk in exchange for certainty. This chapter shows how to construct a utility curve which maps risk attitudes, and then prioritize the decisions in terms of maximum expected utility.

Chapter 19. Writing up your findings. In most cases statistical analysis is done in order to help a decision-maker decide what to do. This chapter suggests ways in which you might write up your report, accompanied by visualizations to get across your message. This is also a great template if you have to hand in some classwork. There are also links to some helpful websites which discuss the preparation of slides and how to make presentations.

Chapter 20 is a Glossary and contains some basic statistics informa-

tion, primarily definitions of terms that the book uses frequently and which have a particular meaning in statistics (for example 'population'). The Glossary also discusses why the inferential techniques used in statistics are so powerful, allowing us to make inferences about a population based on what appears to be a very small sample.

Under E for **Excel** in the Glossary, you'll find some links to screen casts on subjects not directly covered in this book, but which you might find helpful.

2. Samples and populations

The power of statistics lies in the fact that only quite small samples are needed for testing and predicting effects at a much large scale. If we wanted to find out the average weight of a certain species of fish in a river, we do not need to drain the river and weigh each fish. A small sample, properly collected, would be enough to make an estimation of the true mean weight. The result would not be totally accurate but we could calculate the potential inaccuracies and include those with our results. This would save a great deal of time and money, not to mention leaving the fish undisturbed. The same logic applies to collecting opinions on a new product, political polling and any other situation where it is too expensive or time-consuming to test or interview every single person.

Time for some definitions: a **population** (statistically speaking of course) is the collection or set of all elements of interest in the study. This could be the population of a country of state, but it could also be the fish in a particular section of a river. Each person, each fish, is an element in the study and together they form the population of interest.

A **sample** is drawn from the population. The sample is therefore usually smaller than the population. Exceptions are censuses, when the government wants data on every single person. As mentioned above, a small sample can represent the population remarkably well. The sample must have been collected

1. **Randomly**, so that each member of the population from which the sample is drawn has the same probability of being selected as a member of the sample. A small fish has the chance as a large fish. Random selection is best carried out by

computer, not least because humans are terrible at it. Random selection Youtube.

2. **Representatively**, meaning that the sample is true reflection of the population at a smaller scale. Statistics generalizes from the sample to the population, and so a biased or unrepresentative sample will result in a misleading generalization. It is very easy to let biases creep into your sample. Often this is because of a poor definition of the population.

2.1 What can possibly go wrong?

For statistics to work—drawing inferences about a population from a small sample—then the population has to be clearly defined and the sample must be representative. Problems arise when the population is not clearly defined. For example, you want to get the opinions of visitors to a shopping mall on the packaging for your new product. Sounds simple, draw a sample from visitors to the mall and then use the tools provided by statistics to make inferences about the population. But what is the population? Is it all the people in the mall at lunchtime? Does that include the teenagers hanging around near the deliveries entrance?

3. Notation

All those thick black Xs and Greek letters floating about the page are enough to put anyone off statistics. But, as you probably guessed, they do have a purpose. And understanding and using them isn't really that difficult. Let's work through this in a Q&A format. Q. Why are there so many different letters? A. We use a letter to represent a particular variable. Using an example of finding the average weight of fish in a particular stretch of river. First we draw a sample from the population. We weight each element in the sample (each fish) and write it down. The weights we can call x, or any other letter we choose. The letter x is acting as a container for all the observations we have written down about the fish. We can subscript our xs with numbers pinpointing which x exactly we are referring to if necessary. The dataset might look like this:

Observation	Weight
X1	297
X2	303
xn	300

I have called the last observation Xn because that means continue the numbering until you reach the last one. The size of the sample is n, so the numbering would run x1, x2, and so on .... to xn. If we want to draw attention to any one of the n observations, without being specific, then use the expression xi. Q. Why are some letters in what looks like Greek writing? A. in Chapter 1 we talked about samples and populations. Samples are drawn from populations, but they are not the same. We will know the size of the sample, and also we can probably find out important information such as the mean. We need some way of differentiating between the sample and the

population. How do you know whether 'the mean' is referring to the sample or the population? The customary way is to use Greek letters for characteristics of the population and regular Latin letters for the same thing in the sample. Below is a list of some of the quantifiable features and their notation.

Quantifiable feature

Mean or average in the population is μ. In a sample it is

$\bar{x}$, spoken 'x bar'.

Variance in the population is σ^2. In a sample it is just s.

Size of population is N; in a sample it is n.

Q. What other symbols and notation should I know? A. There are hundreds of course, but for our purposes these are the symbols you should probably know:

$$\sum x_i$$

The big symbol is called sigma and means add up eveything. So in this case we should total all the xs in the dataset. Notice the use of the xi here. Anything with an x as part of the dataset gets added up.

4. Variance and standard deviation

It took me a while to get this one down, I'll freely admit. The underlying meaning is the average difference of each observation in a dataset from the mean of the dataset. How spread about are the observations? This is a key characteristic in statistics and worth trying hard to understand properly. An example will probably help. Let's say you run a small store and that you rely on two companies to resupply you. You record the number of days each company took to fulfil your order on five different occasions. The results are below.

Supplier A	Supplier B
1	3
2	3
3	3
4	3
5	3

Which supplier would you choose to work with given only the data above? I think I would pick Supplier B because he or she appears to be much more consistent. You know that your order will most likely be fulfilled in three days. But that isn't the case with Supplier A. You might be waiting five days. But how can we measure the consistency I have just described? If we take the means of each supplier, we find that they are the same at 3. Now, using the notation we just learned, we could write

$$\bar{x}_a = 3$$

and the equivalent for Supplier B. This doesn't help at all.

Perhaps we could subtract each observation from its mean, add up the differences and then divide by 5 (because we have five observations. Let's try that. With Supplier A it is:

Observation	Mean	Difference
1	3	-2
2	3	-1
3	3	0
4	3	1
5	3	2

The total is 10. If the data was a population (all the deliveries) then we would divide by 5, because N = 5 and find a variance of 2. But it is unlikely to be the population because the store will have had many more deliveries than that. To correct for this, we divide by n – 1. n is the sample size, so we divide by 5 – 1 = 4. The variance is 10/4 = 2.5. Let's try the same thing in Excel. The standard deviation is just the square root of the variance. The standard deviation is used more frequently than the variance because of units. We squared the differences to get positive numbers, but in doing so we also squared the units. This makes very little sense. So take the square root of the variance to get the standard deviation which will be in the same units as the original observations. In equation form, the standard deviation is:

$$s = \sqrt{\frac{\sum (\bar{x} - x_i)^2}{n - 1}}$$

It would be worth your while working your way through this equation, matching it to the steps we took above.

5. Types of variable

6. Hypothesis testing

7. Visualization and Tableau: telling (true) stories with data

In this book, we'll work on gaining insights from data by visualization and quantitative analysis, and then presenting the results to others. Recently a powerful new tool called Tableau has become available. Tableau Public[1] is a free version. But be careful with your data because when you publish to the web, as you must do with Tableau Public, then your data also becomes available to anybody. If you require that your data be protected, you can pay for the private version. Tableau 9 has just been made available.

Many of the worked examples in this book are accompanied by a link to a completed Tableau workbook, showing how you could have used Tableau to present your findings. Tableau cannot perform easily the more technical hypothesis-testing aspects of statistics such as regression, but it can help you to get your point across clearly. Tableau has also provided a useful White Paper[2] on visual best practices which is well worth reading

In business intelligence we are generally interesting in detecting patterns and relationships. This might seem obvious, because as humans we are always looking for these phenomena. I'd just like to add that the absence of a pattern or relationship might be just as informative as finding one. An excellent first step is to take a look at your data with an expository graph before moving on to more formal data analysis. Below are examples of the types of charts or graphs most commonly used in either examining data first off,

---

[1]http://www.tableausoftware.com/public/
[2]http://www.tableausoftware.com/whitepapers/visual-analysis-everyone

just to 'see how it looks'. Checking an expository graph shows up problems that might throw you off later: missing data, outliers or, more excitingly, unexpected and intriguing relationships.

Histograms

Histograms break the data into classes and show the distribution of the data. Which classes (sometimes known as bins) are most common. The histogram shows us the 'shape' of the data. Are there many small measurements, or does the data look as though it is normally distributed and consequently mound-shaped. See Distribution in the Glossary for more on this. The plot below is of deaths in car accidents on a weekly basis in the United States over the period 1973-1978. The histogram shows only the distribution and not the sequence.

From the histogram only, we could say that the most common number of deaths in a week is between 8000 and 9000, because the bars of the histogram are highest there. They are highest because they count the number of times in the time-period that there have been (for example 8000 deaths occurred in 15 weeks).

What the histogram doesn't show is that deaths have actually been decreasing since
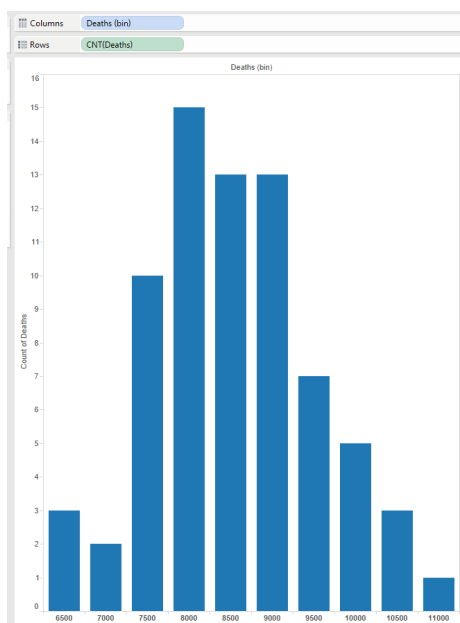
Histogram of car deaths, reported weekly

about 1976. There is probably a simple explanation for this: compulsory seatbelts perhaps? The takehome from this is that expository graphs are really easy to do, and that you should try different methods with the same data to tease out the insights.
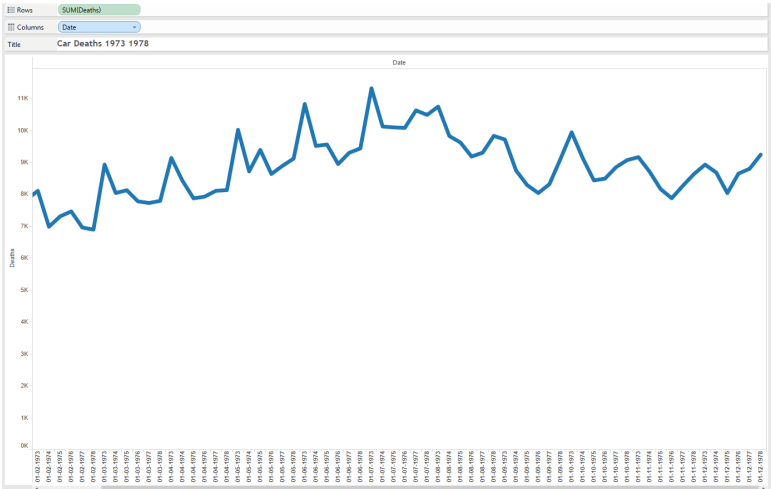
Car deaths over time

Scatterplots

The most common and most useful graph is probably the scatterplot. It is used when we have two continuous variables, such as two quantities (weight of car and mileage) and we want to see the relationship between them. The scatterplot is also useful for plotting time series, where time is one of the variables.

You can use Excel for this, and also Tableau. In Excel, arrange the columns of your data so that the variable which you want on the horizontal x axis is the one on the left. This is the independent variable. Put the other variable, the dependent variable, on the vertical y axis. It is usual practice to put the variable which we think is doing the explaining on the x axis and the response variable on

the y axis. Here is a scatterplot of European Union Expenditure per student as a percentage of GDP:

EU % of GDP spent on Primary Education
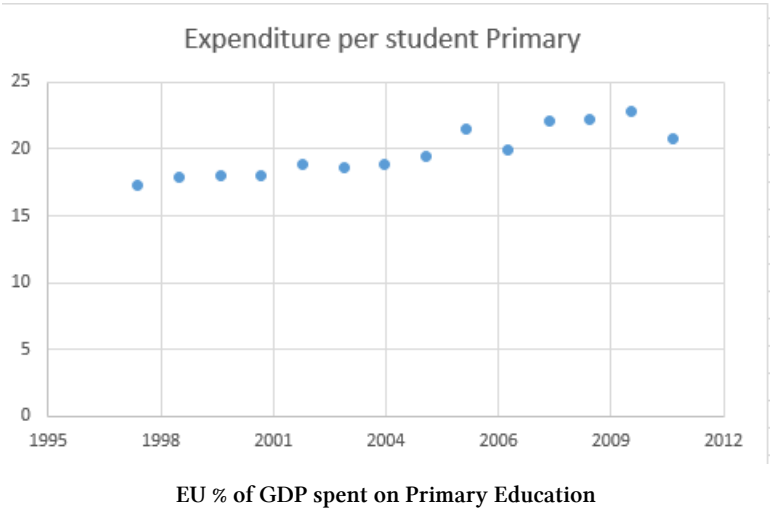
There is a gradual increase over the years 1996 to 2011 as one would expect; educational expenditures tend to remain a relatively fixed share of the budget. However, there was a blip in 2006 which is interesting. Because the data is percentage of GDP, the drop might have been caused by an increase in GDP or alternatively by an actual educational spending decrease. We'd need to look at GDP figures for the period. I got European GDP figures per capita and using Tableau put them on the same plot, with different axes of course. The result is below
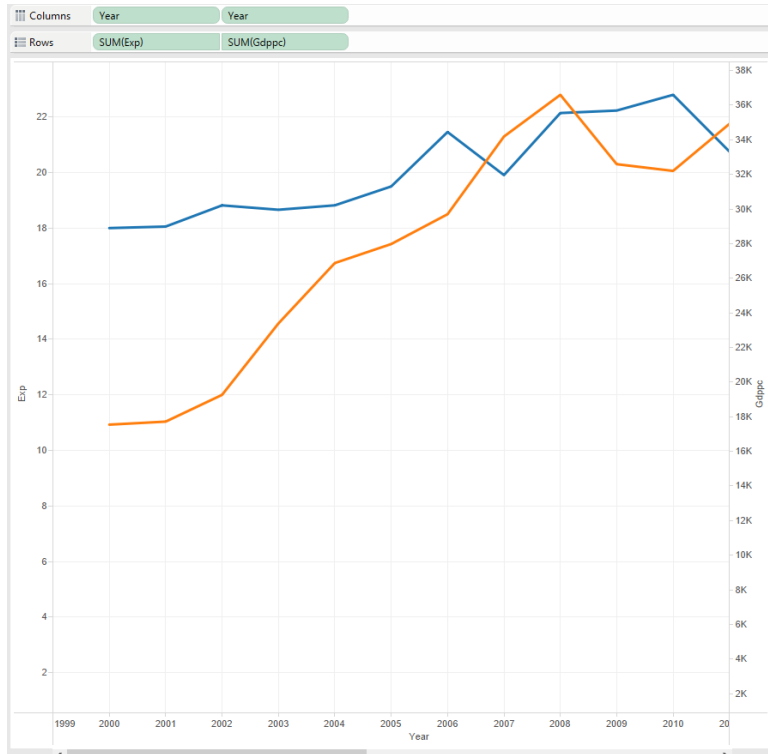
European GDP and Primary Expenditures

GDP dropped around 2008/2009 because of the world financial crisis. The blip in primary expenditures is still unexplained.

Maps

Tableau makes the display of spatial data relatively easy, provided you tell Tableau which of your dimensions contain the geographical variable. The plot below shows changes in greenhouse gas emissions from agriculture in the European Union. I linked world bank data by country. The workbook is here[3].

---

[3]https://public.tableausoftware.com/views/EuropeanGHGfromAgriculture2010-2011/Sheet1?:embed=y&:display_count=no

Unfortunately, not all geographic entities are available for linkage in Tableau. States and counties in the United States are certainly available and provinces in Canada, as well as countries in Europe. There are ways to import ArcGIS shapefiles into Tableau. A shapefile is a list of vertices which delimit the polygons that define the geographical entity.
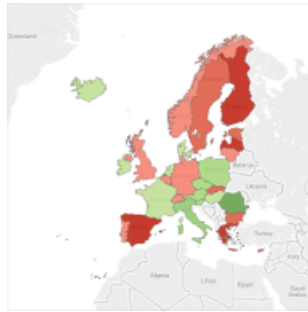
GHG emissions from agriculture in the European Union

8. Data and how to get it

In this chapter, we'll look at the data collection process, the problems that might accompany poorly collected data, discuss so-called **big data** and provide links to some useful public sources of data.

As you might expect, this book works through the analysis of numbers — quantitative data — but it is important to note that the analysis of qualitative data is a rapidly growing area. Unfortunately the analysis of qualitative data is beyond this book and the software available to us.

Quantitative data comes from two main sources. **Primary data** is collected by you or the company you are working for. For example, the market research you do to find out the possible market share for your product provides primary data. Primary data includes both internal company data and data from automated equipment such as website hits. Collecting data is expensive and highly proprietary. It is therefore unlikely to be published and available outside the enterprise.

By contrast, **Secondary data** is plentiful and mostly free. It is collected by governments of all sizes, and also by many non-government organizations as well. There is an increasing trend towards the liberation of data under various open government initiatives. Government data is usually reliable, but be sure to check the accompanying notes which warn of any problems, such as limited sample size or change in classification or collection methods over time. There are some links to secondary data at the end of this chapter.

Experimental design

Experiments don't necessarily have to be conducted in laboratories by people in white coats: a survey in shopping mall is also a form of experiment, as is analyzing the results of your favorite cricket team. There are two different design types: **experimental** and **observational**. The key difference is the amount of randomization that is built into the experiment. In general, more randomization is good, because it helps to remove the influence of fixed effects, such as the quality of the soil in a particular location.

Here's an example to make clear the difference. Imagine a researcher wanting to test the effect of a new drug on mice. In an experimental design, the mice are examined before the drug is administered, and then the drug is applied to a treatment group of mice and a control group. The control group receives no drugs. Its job is to act as a reference group. All the mice are kept in identical conditions apart from the application of the drug. The allocation of mice to the treatment group and to the control group is entirely random.

While we can (and do) carry out drug experiments on mice, it would clearly be very wrong to try to do the same with humans as subjects. Instead we collect data or observations and then analyze those observations looking for differences.

To compensate for the lack of randomization, we control for observed differences by including as many relevant variables as possible. If we knew that person X had received a particular drug and had developed a particular condition, we would want to compare person X with somebody else who had not developed that condition. Relevant variables that we would want to know might be age, gender and possibly pre-existing health conditions. Including these variables reduces fixed effects and allows us to concentrate on the effect of the drug.

Problems with data

It is obvious that to be credible, your analyses must be based on reliable data. Problems with data are usually connected to poor sampling and experimental design techniques, especially:

Sample size too small. The relative size of the sample to the population usually does not matter. It is the absolute size of the sample that counts. You usually want at least several hundred observations.

Population of interest not clearly defined. It is clear that we need to take a sample from a population, but what exactly is the population? Here's an example. You want to survey shoppers in a shopping mall regarding your new product. But of the people inside the mall, who exactly are your population? People just entering, people just leaving, people having their lunch in the food court? Singles, couples, elderly people or the teenagers hanging around outside the door? You can see that picking any one of these groups on its own will lead to a biased sample.

Non-response bias. Many people don't answer those irritating telephone calls which come in the evening because they're busy with dinner. As a result, only answers from those who do choose to answer the survey are counted. Those respondents most likely aren't representative of the population. Perhaps they live alone or do not have too much to do. I'm not saying that they should not be in the sample, just that including only those who do respond may bias your sample.

Voluntary response bias. If you feel strongly about an issue, then you are more likely to respond than if you are indifferent. That's simply human nature. As a result, the survey results will be skewed by the views of those who feel most passionately. This is hardly a representative sample because the strongly-held views drown out the more moderate voices.

8.1 Big data

Primary data frequently comes from automated collection devices, such as scanners, websites, social media, and the like. The volume of such data is enormous, and is aptly called big data. Big data is the term used to describe large datasets generated by traditional business activities and from new sources such as social media. Typical big data includes information from store point-of-sale terminals, bank ATMs, Facebook posts and YouTube videos.

One of the apparently attractive features of big data is simply its size, which supposedly enables deeper insights and reveals connections which would not appear in smaller samples. This argument neglects the power of statistics, and in particular inferential statistics. A small sample, properly collected, can yield superior insights to a very large poorly collected sample. Think of it this way: which is better: a very large sample in which all the respondents are in the same age-group and of the same gender; or a smaller one which more accurately reflects the population?

8.2 Some useful sites

You can of course easily just Google for data, or look at these more focused sites:

Gapminder data: free to use but be sure to attribute[1]

Worker employment and compensation[2]

Interesting and wide-ranging historical data[3]

Google Public Data[4]

---

[1] http://www.gapminder.org/data/
[2] http://www.bls.gov/fls/country/canada.htm
[3] http://www.historicalstatistics.org/
[4] http://www.google.com/publicdata/directory

International Monetary Fund[5]

Food and Agriculture Organisation[6]

United Nations Data[7]

The World Bank[8]

---

[5]http://www.imf.org/external/data.htm
[6]http://www.fao.org/statistics/en/
[7]http://data.un.org/
[8]http://databank.worldbank.org/data/home.aspx

9. Regression: Predicting with continuous variables

This chapter is about discovering the relationships that might or (equally well) might not exist between the variables in the dataset of interest. Here's a typical if rather simplistic example of regression in action: the operator of some delivery trucks wants to predict the average time taken by a delivery truck to complete a given route. The operator needs this information because he charges by the hour and needs to be able to provide quotations rapidly. The customer provides the distance to be driven, and the number of stops en route. The task is to develop a model so that the truck operator can predict the time taken given that information. Regression provides a mathematical 'model' or equation from which we can very quickly predict journey times given relevant information such as the distance, and number of stops. This is extremely useful when quoting for jobs or for audit purposes.

We know the size of the input variables—the distance and the deliveries, but we don't know the rate of change between them and the dependent or response variable: what is the effect on time of increasing distance by a certain amount? Or adding one more stop? Using the technique of **regression**, we make use of a set of historical records, perhaps the truck's log-book, to calculate the average time taken by the truck to cover any given distance. As with any attempt to predict the future based on the past, the predictions from regression depend on unchanged conditions: no new road works (which might speed up or delay the journey), no change in the skills of the driver.

In the language of regression, the time taken in the truck example is the **response** or **dependent** variable, and the distance to be driven is the **explanatory** or **independent** variable. While we will only ever have just one response variable, the number of explanatory variables is unlimited. The number of stops the truck has to make will impact journey time, as will variables such as the age of the truck, weather conditions, and whether the journey is urban or rural. If we know these variables, we also can include them in the model and gain deeper insights into what does (and equally important does not) affect journey time.

Regression models are used primarily for two tasks: to explain events in the past; and to make predictions. Regression provides **coefficients** which provide the size and direction of the change in the response variable for a one unit change in one or more of the explanatory variables.

Regression makes a prediction for how long a truck will take to make a certain number of deliveries and also states how accurate the prediction will be. With a regression model in hand, we can make quotations accurately and quickly. In addition, we can detect anomalies in records and reports because we can calculate how long a journey should have taken under various what-if conditions.

Here I have used a straightforward example of calculating a time problem for a trucking firm, but regression is much more powerful than that. Some form of regression underlies a great deal of applied research. When you read about increased risk due to smoking (or whatever else is the latest danger) that risk was calculated with regression. In this chapter, we'll calculate the difference between used and new marioKart sales on ebay, estimate stroke risk factors, and gender inequalities in pay, all with regression.

Regression is not very difficult to do, but the problem is that everybody does it. Ordinary Least Squares (OLS), linear regression's proper name, rests on some assumptions which should be checked for validity, but often aren't. We cover the assumptions, and what

to do if they're not met, in the following chapter. This chapter is about the hands-on applications of regression.

9.1 Layout of the chapter

The chapter begins with some background on how regression works. We'll illustrate the theory with the trucking example mentioned above, before going on to adding more explanatory variables to improve the accuracy of the prediction.

Particularly useful explanatory variables are dummy variables, which take on a categorical values, typically zero or 1. For example, we could code employees as male and female, and discover from this whether there is a gender difference in pay, and calculate the effect. In a worked example in the text, we analyze the sales of marioKart on ebay, and use dummy variables to find the average difference in cost between a new and a used version.

Sometimes two variables interact: higher or lower levels in one variable change the effect of another variable. In the text, we show that the effect of advertising changes as the list price of the item increases.

Finally, we'll discuss **curvilinearity**. The relationship between salary and experience is non-linear. As people get older and more experienced, their salaries first move quickly upwards, and then flatten out or plateau. We can capture that non-linear function to make prediction more accurate.

9.2 Introducing regression

At school you probably learned how to calculate the slope of a line as 'rise over run'. Let's say you want to go to Paris for a vacation. You have up to a week. There are two main expenses, the airfare and the cost of accommodation per night. The airfare is fixed and

stays the same no matter whether you go for one night or seven. The hotel (plus your meal charges etc) is $200 per night. You could write up a small dataset and graph it like this:

	A	B	C	D
	Airfare	Duration	Hotel bill	Total
	1000	0	0	1000
		1	200	1200
		2	400	1400
		3	600	1600
		4	800	1800
		5	1000	2000
		6	1200	2200
		7	1400	2400

Paris total expenses
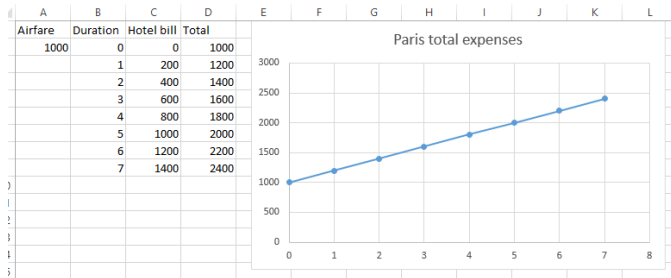
Total Paris trip expenses

The equation for is: Total expenses = 1000 + 200*Nights

Basically, you have just written your first regression model. The model contains two coefficients of interest:

- the intercept which is the value of the response variable (cost) when the explanatory variable is zero. Here the **intercept** is $1000. That is the expense with zero nights. It is the point on the vertical y axis where the trend line cuts through it, where $x = 0$. You still have to pay the airfare regardless of whether you stay zero nights or more.
- the slope of the line which is $200. For every one unit increase in the explanatory variable (nights) the dependent variable (total cost) increases by this amount.

In regression, we usually know the dependent variable and the independent variable, but we don't know the coefficients. We use regression to extract those coefficients from the data so that we can make predictions.

How do we know whether the slope of the regression line reflects the data? Try this thought experiment: if you plotted the data and

the trend line was flat, what would that mean? No relationship. You could stay in Paris forever, free! A flat line has zero slope, and so an extra night would not cost any more.

More formally, the hypothesis testing procedure tests the null hypothesis that the slope is zero. If we can reject the null, then we are required to accept the alternative hypothesis, which is that the slope is not zero. If it isn't zero (the trend line could slope up or down) then we have something to work with. We test the hypothesis using the data found from the sample, and Excel gives us a p value. If the p value is smaller than 0.05, we reject the null hypothesis. If we reject the null we can say that there is a slope and the analysis is worthwhile.

The Paris example had only one independent variable (number of nights) but regression can include many more, as the trucking example below will show. If there is more than one variable, we cannot show the relationship with one trend line in two dimensions. Instead, the relationship is a hyperplane or surface. The plot below shows the effect on taste ratings (the dependent variable) of increasing amounts of lactic acid and H2S in cheese samples.
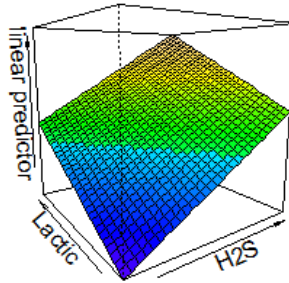
3D hyperplane for responses to cheese

Here is another example. We have the data on the size of the population of some towns and also pizza sales. How can we predict sales given population. Pizza Youtube here[1]

9.3 Trucking example

The records of some journeys undertaken are available in an Excel file named 'trucking'[2].

Take a look at the data. As I mentioned at the beginning of the book, it is always a good idea to begin with at least a simple exploratory plot of the variables of interest in your data. With Excel we can quickly draw a rough plot so that we can see what is going on. What we want to do is predict time when we know the distance. The

---

[1] https://www.youtube.com/watch?v=ib1BfdVxcaQ
[2] https://dl.dropboxusercontent.com/u/23281950/trucking.xlsx

scatter plot below shows time as a function of distance. Youtube: scatterplot of trucking data[3]
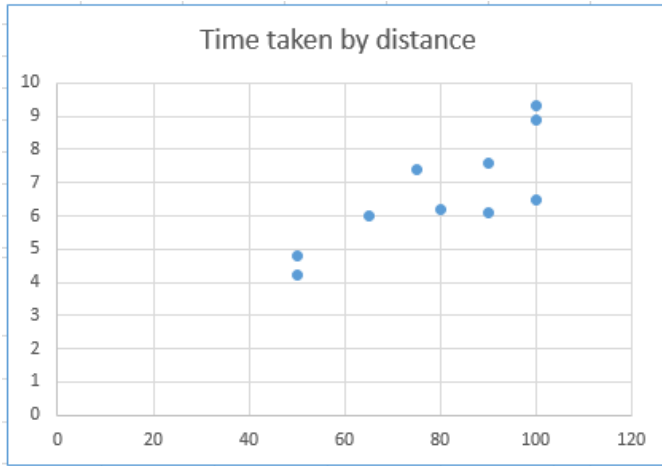
Trucking scatterplot

Time is the dependent variable and distance is the independent variable. The independent variable goes on the horizontal x axis, the dependent variable on the vertical y axis.

From the plot it is clear that there is a positive relationship between the two variables. As the distance increases, then so does the time. This is hardly a surprise. But we want more than this: we want to quantify the relationship between the time taken and the distance traveled. If we can model this relationship, that will be useful when customers ask for quotations.

The data set contains one further variable, which is the number of deliveries on the route. We'll use that later on. For now we'll use just the time and the distance.

To build our model, we want to build an equation which looks like this in symbolic form

---

[3]https://www.youtube.com/watch?v=HSpY12mLXoU

$$\hat{y} = b_o + b_1 x_1 + ... b_n x_n + e$$

yhat (spoken 'y hat') is the dependent variable. It is called 'hat' because it is an estimate. b0 is the intercept, or the point where the trend line (also called the regression line) passes through the vertical y axis. This is the point where the independent variable is zero. You perhaps remember a similar equation from school: y = mx +b.

In the trucking example, the intercept (b0) might be the time taken warming up the truck and checking documentation. Time is running, but the truck is not moving. b1 is the 'coefficient' for the independent variable because it provides the change in the dependent variable for a one-unit change in the independent variable. x1 is the independent variable, in this case distance. We want to be able to plug in some value of the independent variable and get back a predicted time for that distance.

The b and the x both have a subscript of 1 because they are the first (and so far only independent variable). I have put in more variables just to indicate that we could have many.

The coefficient, b1, provides the rate of change of the dependent variable for a one unit change in the independent variable. You can think of this as the slope of the line: a steeper slope means a greater increase in y for a one-unit increase in x.

We can now find the estimated regression equation using the regression application in Excel. First, we use the regress function in Excel's data analysis tool to regress distance on time. Youtube[4]

The regression output is as below:

---

[4] https://www.youtube.com/watch?v=xKsYfa7YGgE

1	SUMMARY OUTPUT							
2								
3	*Regression Statistics*							
4	Multiple R	0.814906						
5	R Square	0.664071						
6	Adjusted R S	0.62208						
7	Standard Errc	1.001792						
8	Observations	10						
9								
10	ANOVA							
11		*df*	*SS*	*MS*	*F*	*Significance F*		
12	Regression	1	15.8713	15.8713	15.8145781	0.004080177		
13	Residual	8	8.028696	1.003587				
14	Total	9	23.9					
15								
16		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0% pper 95.0%*
17	Intercept	1.273913	1.400745	0.909454	0.38968736	-1.956209623	4.504036	-1.95621 4.504036
18	Distance	0.067826	0.017056	3.976755	0.00408018	0.028495716	0.107156	0.028496 0.107156

The regression output

I have marked some of the key results in red, and I explain them below.

9.4 How good is the model? —r-squared

There is a red circle around the adjusted r-squared value, here 0.622. r-squared is a measure of how good the model is. r-squared goes from 0 to 1, with a 1 meaning a perfect model, which is extremely rare in applied work such as this. Zero means there is no relationship at all. The result here of 0.62 means that 62% of the variance in the dependent variable is explained by the model. This isn't bad, but it's not great either. Below, we'll add more independent variables and show how the accuracy of the model improves. The adjusted r-squared takes into account the number of variables in the regression equation and also the sample size, which is why it is slightly smaller than the r-squared value. It doesn't have quite the same interpretation as r-squared, but it is very useful when comparing models. Like r-squared, we want as high a value as possible: higher is better because it means that the model is doing a more accurate job.

Also circled are the words intercept and distance. These give us the

coefficients we need to write the model. Extracting the coefficients from the Excel output, we can write the estimated regression model.

$$\hat{y} = 1.274 + 0.068 * Dist$$

where yhat is the estimated or predicted response time. If you took a very large number of journeys of the same distance, this would be the average of the time taken.

The intercept is a constant, it doesn't change. It is the amount of time taken before a single kilometer has been driven. The distance coefficient of 0.068 is the piece of information that we really want. This is the rate of change of time for a one unit change in the dependent variable, distance. An increase of one kilometer in distance increases the predicted time by 0.0678 hours, and vice-versa of course. Do the math and you'll see that the average speed is 14.7 mph.

9.5 Predicting with the model

A model such as this makes estimating and quoting for jobs much easier. If a manager wanted to know how long it would take for a truck to make a journey of 2.5 kms for example, all he or she would need to do is to plug 2.5 into distance and get:

predicted time = 1.27391 + 0.06783*2.5 = 1.4435 hours.

multiply this result by the cost per hour of the driver, add the miscellaneous charges and you're done.

It is unwise the extrapolate. Only make predictions within the range of the data with which you calculated your model (see also Chapter 4 on this.

9.6 How it works: Least-squares method

The method we just used to find the coefficients is called the **method of least squares**. It works like this: the software tries to find a straight line through the points that minimizes the vertical distance between the predicted y value for each x (the value provided by the trend line) and the actual or observed value of each y for that x. It tries to go as close as it can to all of the points. The vertical distance is squared so that the amounts are always positive.

The plot below is the same as the one above, except that I have added the trend line.
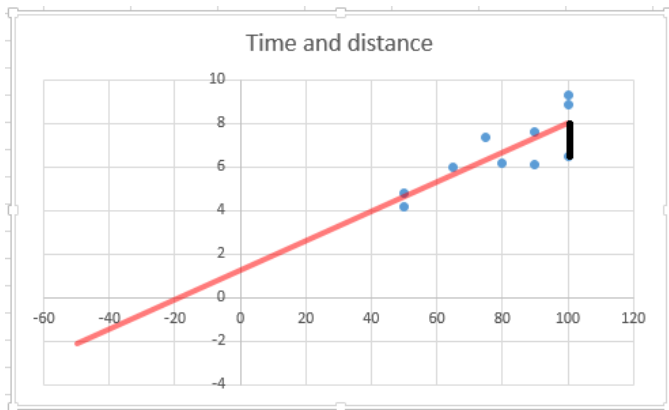
The black line illustrates the error

The slope of the trend line is the coefficient 0.0678. That is the rate of change of the dependent variable for a one-unit change in the independent variable. Think "rise over run". I extended the trend line backwards to illustrate the meaning of intercept. The value of the intercept is 1.27391, which is the value of y when x is zero. This is actually a form of extrapolation, and because we have no observations of zero distance, this is an unreliable estimate.

Now look at the point where x = 100. There are several y values representing time taken for this distance. I have drawn a black line

between one particular y value and the trend line. This vertical distance represents an error in prediction: if the model was perfect, all the observed points would lie along the trend line. The method of least squares works by minimizing the vertical distance. It is possible to do the calculations by hand, but they are tedious and most people use software for practical use. The gap between predicted and observed is known as a residual and it is the 'e' term in the general form equation above.

The error discussed just above is the vertical distance between the predicted and the observed values for every x value. The amount of error is indicated by the **r-squared value**. r-squared runs from 0 (a completely useless model) to 1 (perfect fit).

In the glossary, under Regression, I have written up the math that underpins these results.

9.7 Adding another variable

The r-squared of 0.66 we found with one independent variable is reasonable in such circumstances, indicating that our model explains 66% of the variability in the response variable. But we might do better by adding another variable to explain more of the variability.

The trucking dataset also provides the number of deliveries that the driver has to make. Clearly, these will have an effect on the time taken. Let's add deliveries to the regression model.

Note that you'll need to cut and paste so that the explanatory variables are adjacent to each other. It doesn't matter where the response variable is, but the explanatory variables must be adjacent in one block. The new result is below:

	A	B	C	D	E	F	G	H	I
	SUMMARY OUTPUT								
	Regression Statistics								
	Multiple R	0.950678166							
	R Square	0.903788975							
	Adjusted R Square	0.876300111							
	Standard Error	0.573142152							
	Observations	10							
	ANOVA								
		df	*SS*	*MS*	*F*	*gnificance F*			
	Regression	2	21.60055651	10.80028	32.87837	0.000276			
	Residual	7	2.299443486	0.328492					
	Total	9	23.9						
		Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
	Intercept	-0.86870147	0.951547725	-0.91294	0.391634	-3.11875	1.381351	-3.11875	1.381351
	Deliveries	0.923425367	0.221113461	4.176251	0.004157	0.400575	1.446276	0.400575	1.446276
	Distance	0.061134599	0.009888495	6.182397	0.000453	0.037752	0.084517	0.037752	0.084517

Now with deliveries

Note that the r-squared has increased to 0.903, so the new model explains about 25% more of the variation in time. The improved model is:

$$\hat{y} = -0.869 + 0.06 * Dist + 0.92 * Del$$

A few things to note here: the values of the coefficients have changed. This is because the interpretation of the coefficients in a multiple model like this is based on **only one variable changing**. For example, the coefficient of distance is 0.92, almost one hour for each delivery assuming that the distance doesn't also change. We should interpret the coefficients under the assumption that all the other variables are 'held steady', apart from the coefficient of interest.

9.8 Dummy variables

Above, we saw how adding a further variable has dramatically improve the accuracy of a model. A dummy variable is an additional variable but one that we construct ourselves as a result of dividing data into two classes, for example by gender. Dummy variables are

powerful because they allow us to measure the effect of a binary variable, known as a dummy or sometimes indicator variable. A dummy variable takes on a value of zero or one, and thus partitions the data into two classes. One class is coded with a zero, and is called the **reference group**. The other classes are coded with a one and successive numbers.

There may be more than two groups, but there will always be one reference group. We are generating an extra variable, so the regression equation looks like this:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

In the case of observations which have been coded x1=0 (the reference group), then the b1 term will disappear because it is multiplied by zero. The b2 term remains. For the reference group, the estimated regression equation then simplifies to

$$\hat{y}_{reference} = b_0 + b_2 x_2$$

while for the non-reference group, it is

$$\hat{y}_{nonreference} = b_0 + b_1 x_1 + b_2 x_2$$

The size of b1x1 represents the difference between the average size of the reference level and whatever group is the non-reference group. Let's work through an example. Creating a dummy variable[5]

The dataset ['gender'] (https://dl.dropboxusercontent.com/u/23281950/gender.xlsx) contains records of salaries paid, years of experience and gender. We might want to know whether men and women receive the same salary given the same years of experience. Load the data, then run a linear regression of Salary on Years of Experience, using years of experience as the sole explanatory variable. The result is below:

---

[5]https://www.youtube.com/watch?v=TBJsEb2UCPs

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3		Regression Statistics							
4	Multiple R	0.911394159							
5	R Square	0.830639313							
6	Adjusted R	0.813703245							
7	Standard E	8204.456799							
8	Observatio	12							
9									
10	ANOVA								
11		df	SS	MS	F	ignificance F			
12	Regression	1	3301410596	3.3E+09	49.04558	3.7E-05			
13	Residual	10	673131113.7	67313111					
14	Total	11	3974541710						
15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
17	Intercept	30949.35067	6464.390223	4.787667	0.000737	16545.79	45352.91	16545.79	45352.91
18	Years Exp	4076.498322	582.0862359	7.003255	3.7E-05	2779.529	5373.467	2779.529	5373.467
19									
20									

The gender regression

The interpretation is that the intercept of $30949 is the average starting salary, with years of experience zero. The coefficient $4076 means that every additional year of experience increases the worker's salary by this amount.

The r-squared is 0.83, so the simple one-variable model explains about 83% of the variation in the dependent variable, salary. We have one more variable in the dataset, gender. Add this as a dummy variable and run the regression again. Note that you will have to cut and paste the variables so that the explanatory variables are adjacent.

▲	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	Regression Statistics					
4	Multiple R	0.996685836				
5	R Square	0.993382655				
6	Adjusted R Square	0.991912134				
7	Standard Error	1709.480536				
8	Observations	12				
9						
10	ANOVA					
11		df	SS	MS	F	ignificance F
12	Regression	2	3.95E+09	1.97E+09	675.531	1.56E-10
13	Residual	9	26300913	2922324		
14	Total	11	3.97E+09			
15						
16		Coefficients	tandard Err	t Stat	P-value	Lower 95%
17	Intercept	23607.51734	1434.476	16.45724	5.03E-08	20362.51
18	Gender (X2) 0=Female 1=Male	14683.66667	986.969	14.87754	1.21E-07	12450.99
19	Years Experience (X1)	4076.498322	121.2835	33.61132	9E-11	3802.136
20						

With the gender dummy added

The r-squared has increased to nearly 1, so our new model with the inclusion of gender is very accurate. The important new variable is gender, coded as female = 0 and male = 1. Nothing sexist about this, we could equally well have reversed the coding. If you are female, then the model for your salary is: 23607.5 + 4076*Years

if you are male, then the model for your salary is

23607.5 + 1*14683.67* + *4076*Years

Each extra year of experience provides the same salary increase, but on average males receive $14683.67 more earnings.

Here is another example, using the maintenance dataset. Dummy variable[6]

Another dummy variable example and a cautionary tale

The mariokart dataset came from [OpenIntro Statistics] (www.openintro.org) a wonderful free textbook for entry level students. The dataset

---

[6]https://www.youtube.com/watch?v=Yv681upodDI

contains information on the price of Mario Kart in ebay auctions. First let's test whether condition 'new' or 'used' makes difference. Construct a new column called CONDUMMY, coded new = 1 and used = 0. Now run a regression with your new dummy variable against total price. The output is below

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.127362357			
R Square	0.01622117			
Adjusted R Square	0.009244015			
Standard Error	25.56955174			
Observations	143			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	1520.022598	1520.023	2.324898
Residual	141	92186.07867	653.802	
Total	142	93706.10127		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	47.14809524	2.789866818	16.89977	1.04E-35
condummy	6.622582728	4.34335657	1.524761	0.129558

The condition dummy results

This result is tremendously bad. The p value is 0.129, meaning that condition is not a statistically significant predictor of price. Surely the condition must have some significant effect? Wait—we forgot to do some visualization. To the right is a histogram of the total price variable.

Looks like we might have a problem with outliers...some of the observations are much larger than the others. Take another look with a barchart at the higher prices.
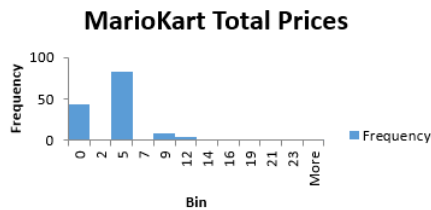
marioKart total prices

We can identify these outliers using z scores (see the Glossary). Or we could just sort them by size and then make a value judgement based on the description. That's what I have done in this Youtube. Outlier removal and regression[7]

The total prices as a bar chart

It seems that two of the items listed were for grouped items which were quite different from the others. There is therefore a legitimate reason for excluding them. Below are the new results:

---

[7]https://www.youtube.com/watch?v=ivLGteHuu3Q

	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.592075				
5	R Square	0.350553				
6	Adjusted R Squar	0.345881				
7	Standard Error	7.370907				
3	Observations	141				
9						
0	ANOVA					
1		*df*	*SS*	*MS*	*F*	*gnificance F*
2	Regression	1	4076.301609	4076.302	75.02818	1.06E-14
3	Residual	139	7551.908374	54.33028		
4	Total	140	11628.20998			
5						
6		*Coefficient*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%Up*
7	Intercept	42.8711	0.813980554	52.66845	1.02E-93	41.26171 4
8	CONDUMMY	10.89958	1.258338778	8.661881	1.06E-14	8.411621 1

Corrected marioKart output

The estimated regression equation is

$$\hat{y} = 42.87 + 10.89 * \text{CONDUMMY}$$

The condition dummy was coded as new = 1, used = 0. If a mariokart is used, then its average price is 42.87, if it new then the average price is 42.87 + 10.89. The average difference in price between old and new is nearly \$11. Makes sense.

Take-home: check your data before doing the regression.

9.9 Several dummy variables

The gender and the marioKart examples above contained just one dummy variable. But it is possible to have more. For example, your sales territory contains four distinct regions. If you make one of the four the reference level, and then divide up the data with dummies

for the remaining three regions, you can compare performance in each of the regions both to each other and to the reference level.

The dataset maintenance contains two variables which you can convert to dummies, following this Youtube.

Maintenance Regression[8]

9.10 Curvilinearity

So far, we have assumed that the relationship between the dependent variable and the independent variable was linear. However, in many situations this assumption does not hold. The plot below shows ethanol production in North America over time.
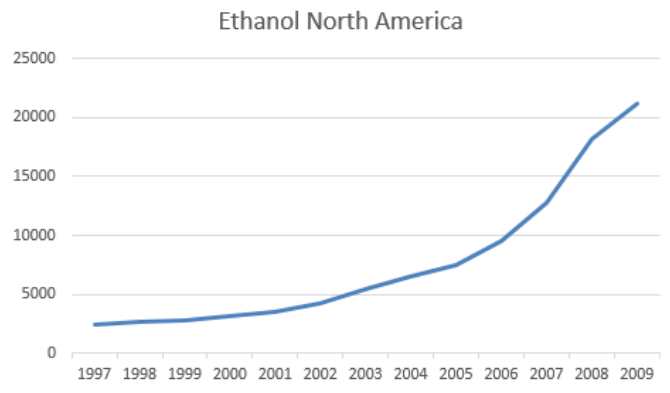
North American ethanol production. Source: BP

The source of the data is BP. It is clear that production of ethanol is increasing yearly but in a non-linear fashion. Just drawing a straight line through the data will miss the increasing rate of production. We can capture the increasing rate with a quadratic term, which is simply the time element squared. I have created a new variable which indexes the years, which I have called t, and a further variable

---

which is t squared.

urvilinear regression Youtube[9]

		Renewable energy – fuel ethanol		
Year		thtonnes	t	t2
	1997	2374	1	1
	1998	2637	2	4
	1999	2831.148	3	9
	2000	3187.299	4	16
	2001	3445.481	5	25
	2002	4172.912	6	36
	2003	5418.881	7	49
	2004	6550.186	8	64
	2005	7507.669	9	81
	2006	9527.401	10	100
	2007	12751.05	11	121
	2008	18154.49	12	144
	2009	21200.5	13	169

The dataset with an index for time. Source: BP

A regression of t against output has an r-squared value of 0.797. Inclusion of the t squared term increases the r-squared markedly to 0.98. The regression output is below.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	4504.0172	920.2783612	4.89419	0.000629	2453.509	6554.525	2453.509	6554.525
t	-1301.06572	302.3291909	-4.30347	0.001553	-1974.7	-627.434	-1974.7	-627.434
t2	194.8751795	21.01333136	9.273883	3.16E-06	148.0546	241.6958	148.0546	241.6958

Regression output with the quadratic term

We would write the estimated regression equation as

$$\hat{y} = 4504 - 1301t + 194.9t^2$$

Notice that for early smaller values of t, the effect of the quadratic

[9]https://www.youtube.com/watch?v=jgjWSpyPBqg

term is negligible. As t gets larger then the quadratic swamps the linear t term.

Making a prediction. Let's predict ethanol production after 5 years. Then

$$\hat{y} = 4505 - 1301 * 5 + 194.9 * 25 = 2872.5$$

The plot below shows predicted against observed values. While the fit is clearly imperfect, it is certainly better than a straight line.

Ethanol: Predicted against Observed
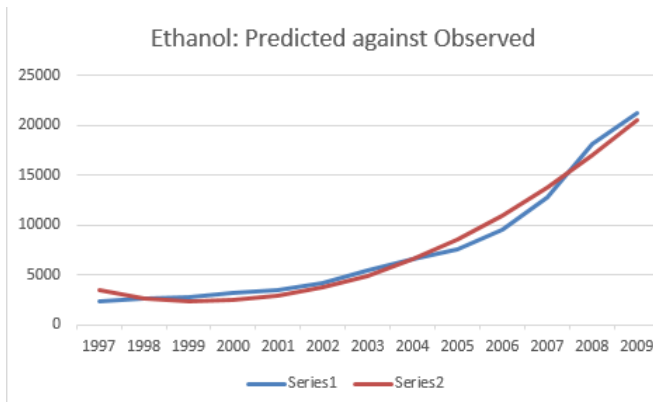
Predicted against observed ethanol production.

9.11 Interactions

Increasing the price of a good usually reduces sales volume (although of course profit might not change if the price increases sufficiently to offset the loss of sales. Advertising also usually increases sales, otherwise why would we do it?

What about the joint effect of the two variables? How about reducing the price and increasing the advertising? The joint effect is called an **interaction** and can easily be included in the explanatory variables as an extra term. The output below is the result of

regressing sales on Price and Ads for a luxury toiletries company. The estimated regression model is

$$\hat{y} = 864 - 281P + 4.48Ads$$

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.922319							
5	R Square	0.850672							
6	Adjusted R	0.83645							
7	Standard E	71.81038							
8	Observation	24							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	2	616900	308450	59.81504	2.13E-09			
13	Residual	21	108291.3	5156.73					
14	Total	23	725191.3						
15									
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	864.1667	101.0249	8.553998	2.77E-08	654.0739	1074.259	654.0739	1074.259
18	Price	-281	35.90519	-7.82617	1.17E-07	-355.669	-206.331	-355.669	-206.331
19	Ads	4.48	0.586329	7.640758	1.71E-07	3.260662	5.699338	3.260662	5.699338
20									

Regression output for just price and ads

So far so good. The signs of the coefficients are as we would expect from economic theory. Sales go down as price rises (negative sign on the coefficient). Sales go up with advertising (positive sign on the coefficient). Caution: do not pay too much attention to the absolute size of the coefficients. The fact that price has a much larger coefficient than advertising is irrelevant. The coefficient is also related to the choice of units used.

Now create another term which is price multiplied by ads. Call this term PAD. The first few lines of the dataset are below.

	A	B	C	D
1	PAD	Price	Ads	Sales
2	100	2	50	478
3	125	2.5	50	373
4	150	3	50	335
5	100	2	50	473
6	125	2.5	50	358
7	150	3	50	329
8	100	2	50	456

First few lines with the new interaction variable

Now do the regression again, including the new term.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.988994							
5	R Square	0.978109							
6	Adjusted R	0.974825							
7	Standard E	28.17386							
8	Observation	24							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	3	709316	236438.7	297.8692	9.26E-17			
13	Residual	20	15875.33	793.7667					
14	Total	23	725191.3						
15									
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-275.833	112.8421	-2.44442	0.023898	-511.218	-40.4488	-511.218	-40.4488
18	PAD	-6.08	0.563477	-10.7901	8.68E-10	-7.25539	-4.90461	-7.25539	-4.90461
19	Price	175	44.54679	3.928453	0.000832	82.07702	267.923	82.07702	267.923
20	Ads	19.68	1.427352	13.78777	1.13E-11	16.7026	22.6574	16.7026	22.6574

Inclusion of the interaction term

The new term is statistically significant and the r-squared has increased to 0.978109. The new model does a better job of explaining the variability. How come price is now positive and the interaction term is negative? We have to look at the results as a whole remembering that the signs work when all the other terms are 'held constant'. The explanation: as the price increases, the effect of advertising on sales is LESS. You might want to lower the price and see if the increased volume compensates.

Another interaction example

Here's another example, this one relating to gender and pay. The dataset is called 'paygender' and contains information on the gender of the employee, his/her review score (performance), years of experience and pay increase. We want to know:

- Is there a gender bias in awarding salary increases?
- In there a gender bias in awarding salary increases based on the interaction between gender and review score?

First, let's regress salary increases on the dummy variable of gender and also Review. My results are below (I have created a new variable which I have called G. It is just the gender variable coded with male = 0 and female = 1.

1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple F	0.913247							
5	R Square	0.834021							
5	Adjusted I	0.810309							
7	Standard I	71.91991							
3	Observati	17							
9									
0	ANOVA								
1		*df*	*SS*	*MS*	*F*	*gnificance F*			
2	Regressio	2	363872.4	181936.2	35.17393	3.47E-06			
3	Residual	14	72414.62	5172.473					
4	Total	16	436287.1						
5									
6		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
7	Intercept	204.5061	43.2204	4.731703	0.000321	111.8075	297.2046	111.8075	297.2046
8	G	-233.286	36.25862	-6.43395	1.56E-05	-311.053	-155.519	-311.053	-155.519
9	Review	2.24689	0.647321	3.471062	0.003745	0.858526	3.635255	0.858526	3.635255
0									

G dummy and Review

Nothing very surprising here: women get paid on average 233.286 less than men. And—holding gender steady–each point increase in Review gives an increase in salary of 2.24689.

How do I know that women are paid less than men? The estimated regression equation is:

$$\hat{y} = 204.5061 - 233.286*(x = 1\text{ifF}) - 233.286*(\text{x} = 0\text{ifM}) + 2.24689*\text{Review}$$

Remember how we coded men and women? The x = 0 if M term disappears, so we are left with this equation for men:

$$\hat{y} = 204.5061 + 2.24689*\text{Review}$$

and this one for women (I've done the subtraction) to give

$$\hat{y} = -28.779 + 2.24689*\text{Review}$$

Conclusion: there is a gender bias against women. For the same Review standard, on average women are paid 233.286 less than men.

How about the second question — possibility that the gender bias increases with Review level? We can test this with an interaction variable. Multiply together your gender dummy and the Review score to create a new variable called Interact. Then do the regression again with salary regressed against G, Review and Interact. My output is below.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple F	0.970032							
5	R Square	0.940963							
6	Adjusted	0.927339							
7	Standard	44.51199							
8	Observati	17							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressio	3	410529.9	136843.3	69.06682	3.05E-08			
13	Residual	13	25757.13	1981.318					
14	Total	16	436287.1						
15									
16		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	59.94472	40.03718	1.497227	0.158216	-26.5503	146.4398	-26.5503	146.4398
18	G	-29.714	47.57536	-0.62457	0.543062	-132.494	73.06636	-132.494	73.06636
19	Review	4.848995	0.669355	7.244276	6.51E-06	3.402941	6.295049	3.402941	6.295049
20	Int	-4.05468	0.83555	-4.8527	0.000316	-5.85977	-2.24958	-5.85977	-2.24958
21									

Interaction output

For men, the equation is Salary = 59.94472 + 4.848995 (because the G and Interact terms go to zero because we coded Male=0. So for every extra increase in Review, a man's salary increases by 4.848995.

For women, the equation is Salary = 59.94472 -29.714 - 1*(4.848995-4.05468) so for every one increase in Review, a women's salary increases by only 4.848995-4.05468=0.79. Notice that the adjusted r-squared for this model is higher (it is 0.927339 compared to 0.810309) than the model with just G and Review. So the interaction is both statistically significant, p value 0.000316) and has the meaning that for women, improving the Review score doesn't increase the salary as much as for men.

Conclusion The analysis showed that women are being treated unfairly. There is a gender bias against them in average salaries for the same performance review; and each increase in review points earn them considerably less in salary than for the equivalent male. Caution: this conclusion is based solely on the limited evidence provided by this small dataset.

9.12 The multicollinearity problem

Multicollinearity refers to way in which two or more variables 'explain' the same aspect of the dependent variable. For example, let's say that we had a regression model which was explaining someone's salary. Employees tend to get paid more as they get older and also as their years of experience increases. So if we had both age and years of experience in the list of independent variables we would almost certainly suffer from multicollinearity.

Multicollinearity can lead to some frustrating and perplexing results. You run a regression. One of the variables has a p value larger than 0.05, so you decide to take it out. You run the regression again and—-the sign and/or significance of another variables changes. This happens because the two variables were explaining the same aspect of the dependent variable jointly.

How to solve this problem: check the correlation of the independent variables first, before putting them into your model. Chapter 14 has a section on correlation. If you find that the correlation of any two variables is higher than 0.7, be suspicious. These two may bring your some grief!

9.13 How to pick the best model

You will be trying out different formulations, adding and removing variables to try to capture as much explanatory power as you can. How you do decide if one model is better than another? There are two approaches:

- Look only at the adjusted r-squared value, even if your model contains variables with a p value larger than 0.05. If the adjusted r-squared value goes up, leave such variables in.
- Prune your model so that it contains only variables with a p value <= 0.05. You still want as high an adjusted r-squared as

you can get, but you also want all your explanatory variables to be statistically significant.

I take the latter approach. You usually have fewer variables but all of them have a reason (statistical significance) for being in your model and you can interpret their meaning intelligently. A parsimonious model is better than a complex model that fits your data very well. Simple and robust is good. The dataset that you used to fit your model is only a sample from a population. An overly complex model may not work well when presented with a different sample from the population.

9.14 The key points

- Think through your model before you start including variables. What variables do you think will have an effect on the dependent variable and in which direction (plus or minus). It is tempting to just put in everything and hope for the best but this rarely works. Some software is able to do stepwise regression, pulling out insignificant variables for you, but Excel is not one of them.
- Keep your model as simple as possible. Complicated models rarely work well.
- Visualise your data first, even with a simple scatter plot as we have done throughout this chapter.
- Check whether you have all the variables that you might need. If you were trying to predict whether a shop selling expensive jewellery would make sufficient sales, you might want to know the average income of residents. If you don't have it—get it. There is a huge amount of data lying about which you can obtain either free or quite cheaply. I've included a very brief list of URLs in Chapter 12.

9.15 Worked examples

1. The estimated regression equation for a model with two independent variables and 10 observations is as follows:

$$\hat{y} = 29 + 0.59x_1 + 4.9x_2$$

What are the interpretations of b1 and b2 in this estimated regression equation?

Answer: the dependent variable changes by on average 0.59 when x1 change by one unit, holding x2 constant. Similarly for b2.

Predict the value of the independent variable when x1 is 175 and x2 is290 :

$$\hat{y} = 29 + 0.59(175) + 4.9(290) = 1553.25$$

10. Checking your regression model

It isn't difficult to build a regression model as the previous chapter has shown. But that is part of the problem: everybody does it. But not everybody takes the trouble to check the results. Checking the results is an important step for two reasons: first, to make sure your calculations and predictions are correct; and secondly to show third parties that your work is solid. In this chapter we'll work on doing just that. To decide whether the models we have been working on are any good we need to look at two areas:

- Is your model statistically significant?
- Are the assumptions behind the least squares method met? In particular, linearity and residual distribution.

The first area is easier to work through than the second, which is probably why one doesn't always see residual analysis discussed when results are presented. This is a shame because there is a great deal to be learned from picking through residuals. Your analysis will be greatly improved by a close attention to this area.

Below we'll work through statistical significance and then test the assumptions.

10.1 Statistical significance

The regression trend line displays a rate of change between two variables. The line slopes upwards when the dependent variable increases for a one-unit increase in the independent variable (

positive relationship) and vice-versa (negative relationship). What we want to know is this: did that slope upwards or downwards occur by chance; if we took another sample from the population would we achieve a similar result? Key point: we are usually working with a sample drawn from a population. The sample in the trucking example was the firm's log book.

The null hypothesis is that there is no slope; with the alternative that there is in fact a slope. The hypotheses (see the Glossary for help on hypotheses) to test whether there is a statistically significant slope are:

The null:

$$H_o : \beta = 0$$

and the alternative:

$$H_a : \beta \neq 0$$

where

$$\beta$$

is the symbol for the coefficients of the population parameter of the independent variable of interest. In words, the hypothesis is testing whether beta is zero or not. If it is zero, then we are 'flatlining' and there is no point in continuing with the analysis. If however we can reject the null, then we accept the alternative which is that beta is not zero. It might be negative or positive, that doesn't matter: what matters is whether it is zero or not. Note that we use a Greek letter when discussing a population parameter which will probably never be accurately known. We use the Latin letter 'b' when we have estimated it using a sample drawn from the population.

Excel does all the testing of statistical significance for us in two ways.

First, it tests the individual variables and provides a p value. Below is the result from the first trucking regression we did. Note that the p value for distance is 0.004. It is customary to use a cut-off of 0.05. The meaning of this result is that, following the **rejection rule**, we reject the null if the p value is smaller than 0.05. Because 0.004 is smaller than 0.05, we reject the null we therefore conclude that there is in fact a statistically significant slope. In summary, we tested the hypothesis that the slope was zero, and rejected it. Therefore we accept the alternative which is that there is a slope of some sort.

1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.814906							
5	R Square	0.664071							
6	Adjusted R Sc	0.62208							
7	Standard Errc	1.001792							
8	Observations	10							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	15.8713	15.8713	15.8145781	0.004080177			
13	Residual	8	8.028696	1.003587					
14	Total	9	23.9						
15									
16		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	1.273913	1.400745	0.909454	0.38968736	-1.956209623	4.504036	-1.95621	4.504036
18	Distance	0.067826	0.017056	3.976755	0.00408018	0.028495716	0.107156	0.028496	0.107156

The trucking regression output

This test tells us that there is a statistically significant slope. It doesn't tell us the sign of the slope (up or down) or its magnitude. But the fact that there is a significant slope is important information. It is worth carrying on with the analysis because the variables actually mean something. By the way, ignore the p value for the intercept. It usually has no substantive meaning.

The second way that Excel tests the validity of the model is to test whether the model as a whole is significant. The test is called the F test after the statistician RA Fisher. For the trucking example, look under *significance F*. The result is a p value, in this case the same

p value for the variable DISTANCE. We have only one explanatory variable and so the p value will be the same. We hope that the p value is smaller than 0.05, which enables us to conclude that at least one of the slopes is non-zero.

The section above examined the first of the tests, which was for the statistical significance of the model. Now we move on to the second area.

10.2 The standard error of the model

The standard error in Excel is found just under the adjusted r-squared output. It is the estimated standard deviation of the amount of the dependent variable which is not explainable by the model. In other words, it is the standard deviation of the residuals.

Under the assumption that the model is correct, it is the lower bound on the standard deviation of any of the model's forecast errors. The figure below shows the output of regression estimating risk of a stroke (multiplied by 100) against blood pressure and age, with a dummy variable for smoking or not.

	A	D
1	SUMMARY OUTPUT	
2		
3	*Regression Statistics*	
4	Multiple R	0.934605168
5	R Square	0.87348682
6	Adjusted R Square	0.849765599
7	Standard Error	5.756574565
8	Observations	20
9		
10	ANOVA	
11		*df*
12	Regression	3
13	Residual	16
14	Total	19
15		
16		*Coefficients*
17	Intercept	-91.75949844
18	Smokedummy	8.739871056
19	Age	1.076741057
20	Pressure	0.251813473

Excel output for stroke likelihood

The standard error is 6 (rounded up). For a normally-distributed variable, 95% of the observations will be within two standard deviations of the mean. For our purposes that means that 2 x 6 = 12 should be added or subtracted to the predicted value to find a confidence interval for predictions. The estimated regression equation from the stroke model above is:

$$\hat{y} = -91 + 1.08 Age + 0.25 \Pr essure + 8.74 Smokedummy$$

An imaginary person who smokes, is aged 68 and has a blood pressure of 175 will have a risk of 34 (all figures rounded). We can be 95% confident that this estimate will fall somewhere between 34 - 12 and 34 + 12 or 22 to 46. These are rather wide confidence intervals and so we might want to work at improving the model by for example increasing the sample size or adding more explanatory variables.

10.3 Testing the least squares assumptions

There are two key assumptions that the least squares method relies on which we need to check. After checking we'll work through ways of retrieving the situation should the assumptions be violated. The assumptions are:

- The relationship between the dependent and the independent variables is linear. Fortunately this is easy to check and also to fix if it isn't satisfied.
- The residuals have a non-constant variance. (You'll sometimes see in other statistics textbooks other stricter requirements, such as the residuals being normally distributed and with a mean of zero. For most purposes just checking for non-constant variance is enough). What does non-constant variance mean? We'll work through this by defining a residual and identifying whether or not the assumption has been met. And finally what to do about it. But first, checking for linearity.

Checking for linearity

The relationship between the dependent variable and the independent variable is assumed to be linear. This is important because the model gives us a rate of change: the coefficient shows the change in the dependent variable for a one-unit change in the independent variable. If the relationship is non-linear, that coefficient will not be valid in some portions of the range of the variables.

We want to see a straight line (either up or down) on a scatter plot. Put the dependent variable on the vertical y axis, and one of the independent variables on the horizontal x axis. If you include the trend line, you can observe how closely the observed values match the predicted.

We can also check for linearity by plotting the predicted values and the observed values. The plot below does this for risk of stroke:

Predicted against observed risk
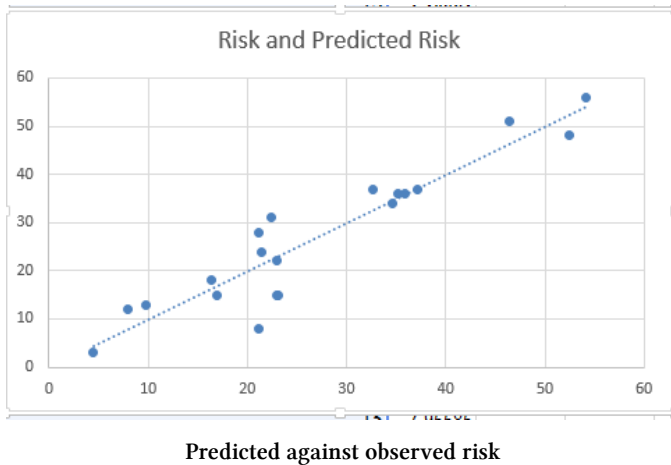
This is a reasonable result, indicating that the linearity condition is met. Most of the points are in a straight line, although I would be concerned about the lower risk levels, especially around risk = 20. There appears to be considerable variance at this point.

10.4 Checking the residuals

A **residual** is the difference between the observed value of y for any given x value, and the predicted value of y for that same x value. It is therefore a prediction error, and is given the notation e for the Greek letter 'epsilon'. It is the vertical distance between the actual and predicted values of the dependent variable for the same value of the independent variable. We discussed this in the previous chapter in connection with the least squares method, where we wrote the estimation equation:

$$\hat{y} = b_0 + b_1 x$$
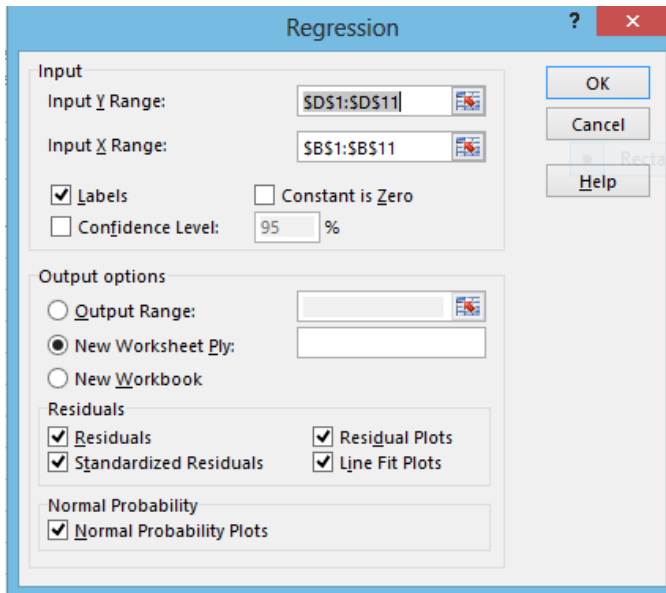
The hat on y, the dependent variable, indicates that it is an estimate of the dependent variable. We know that the estimate cannot be correct unless all the predicted values and the observed values match up exactly. If they don't, then there are residuals. If we call the errors ε (epsilon) which is the Greek letter matching our letter e, then we can rewrite the regression equation as:

$$y = \beta_o + \beta_1 x + \varepsilon$$

The errors have now been absorbed into epsilon and so we can remove the hat from y. It is the behavior of epsilon that is of interest, because the least-squares method rests on the assumption that the errors absorbed into epsilon have a non-constant variance. This means that there no relationship between the size of the error term and the size of the independent variable. Therefore, if we plotted the errors against the independent variables, we should see no clear pattern. How to do this is described below.

10.5 Constructing a standardized residuals plot

Excel provides what is calls standard residuals as part of its regression output, and we will use these. Note however that these are not 'true' standardized residuals, but they are probably close enough. Make sure you check the Standardized Residuals box when setting up your regression.

Check the standardized residuals box

We want to plot the standard residuals against the predicted value yhat. We will create a new column to the left of the column Standard Residuals, and copy the column of predicted times into that new column. Then create a scatter plot of predicted time and standard residuals. You should end up with the image below. Standardized residual plot[1]

---

[1]https://www.youtube.com/watch?v=1wuyZfB39p4

Standard residuals
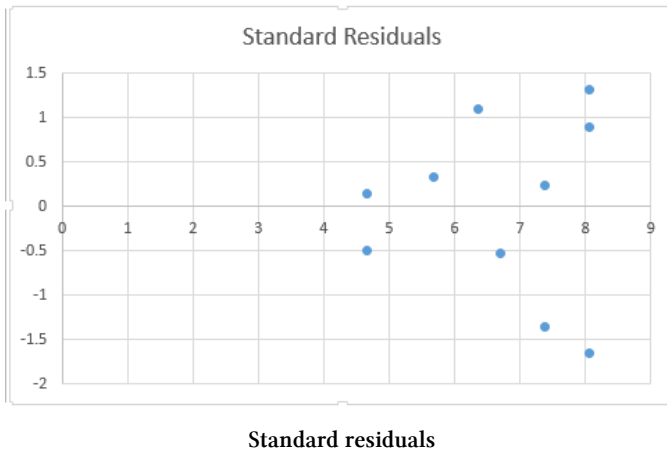
The Y axis indicates the number of standard deviations that each residual is away from the mean, plotted against the predicted values on the horizontal axis. None of the residuals are more that 2 standard deviations away from the mean of zero, so the results are generally satisfactory, although there is one observation in excess of 1.5. We still have a worrying fan shape which would seem to indicate non-constant variance: we can observe a pattern.

Let's run the regression again, but now including the second variable, which is deliveries. The residuals plot is obtained in the same way, and here is the result.

Standardized residuals with two variables
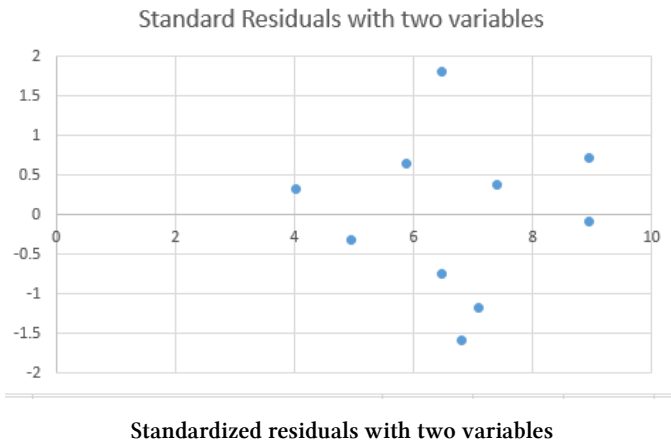
The fan problem seems to have improved a little. The r-squared of the model with two variables was higher, meaning that the residuals were smaller (because there is less unexplained error).

10.6 Correcting when an assumption is violated

Above, we examined possible violations of two of the assumptions underlying linear regression: linearity and non-constant variance. Now let's look at what to do when these assumptions are found to have been violated. First some good news: linear regression is quite robust to such violations, and even so they are quite easy to correct for. We'll deal with the problems in the same order: linearity and non-constant variance.

10.7 Lack of linearity

The first check is to look at a scatter plot of the dependent variable against the independent variable. The plot below shows volume of sales and length of time employed, together with a trend line.
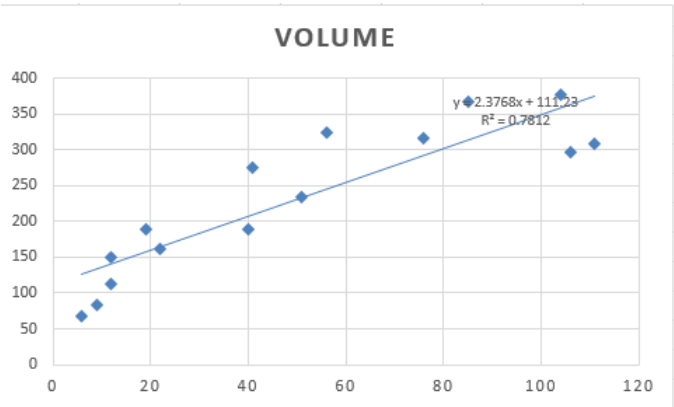
A curvilinear relationship

There are several different ways in which a linear relationship can be achieved so that we can use the least squares method. A common method is to include a quadratic term. This means adding the squared value of the independent variable to the list of independent variables. This is easily done by creating another column, consisting of squared values of the first variable.

	A	B	C
1	**Months**	**MonthSq**	**Sales**
2	41	1681	275
3	106	11236	296
4	76	5776	317
5	104	10816	376
6	22	484	162
7	12	144	150

A new column of the independent variable squared

To do this, create anew column and then label it. Click on the first value in the variable which you want to square, and add ^2. Enter. Then drag downwards. The plot below shows the predicted and observed sales. The model is clearly superior.
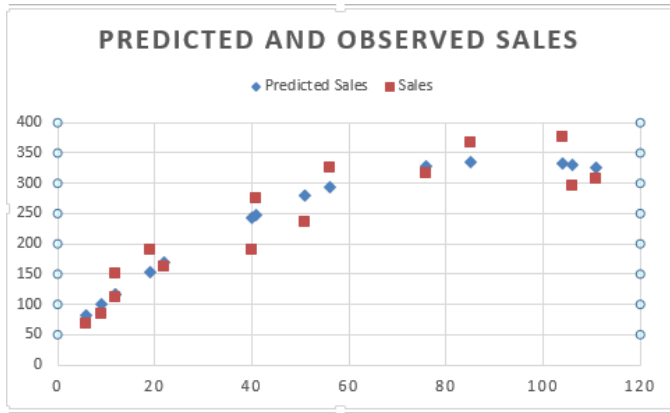
Predicted and observed after inclusion of quadratic term

If the dependent variable has a large number of low values, and is heavily skewed to the right, then a good solution is to transform the dependent variable into its natural logarithm.

10.8 What else could possibly go wrong?

Regression is a very commonly used analytical tool and you are most likely either going to use it yourself or examine the work of others who have used regression. Below is a list of common mistakes to watch for. And if I have made them somewhere in this book, I'm sure you'll be the first to let me know!

10.9 Linearity condition

Regression assumes that the relationship between the variables is linear: the trend line that the software tries to find to minimize the squared difference between the observed and predicted values is straight. So if you try to run a regression on non-linear data, you'll get a result but it will be meaningless.

Action: always visualize the data before doing any analysis. If you see that the data is non-linear, you may be able to transform it using the techniques described in the previous chapter. As an example: the curvilinear example which was transformed by including a quadratic term as an explanatory variable.

10.10 Correlation and causation

Regression is a special case of correlation, and as we all know correlation doesn't mean causation (see the Glossary). In regression, no matter how good the model, all that we have been able to show is that a change in an explanatory variable is associated by a change in the dependent variable. For example, you record the hours you put into studying and your grades. Surprise! More studying = better grades? Or perhaps not....could have been a better instructor. We cannot say that one caused the other. So when writing up results or interpreting those of others, be very careful not to claim more than you are able to.

There is a related problem which is 'reverse causation'. You study more, your grades go up. Tempting to think that one possibly caused the other. But perhaps it is the other way round? Your grades were poor, you studied harder? Or you had good grades and then led you into taking the course seriously.

Action: try not to include explanatory variables which are affected by the dependent variable. You could also try to 'lag' one variable. Cut and paste the hours of studying variable so that it is one time-period behind the grades and then run the regression again.

10.11 Omitted variable bias

Under Correlation in the Glossary I give some examples of lurking variables. Regression, like correlation, is susceptible to the same problem. Example: you notice that in hotter weather there are more

deaths by drowning. Did the hotter weather cause the drownings? Well no, the extra swimming caused by the heat presented more risk scenarios. The lurking variable is hours spent swimming rather than temperature. Try to get to the real variable if you can.

10.12 Multicollinearity

When predicting the price of a house, square footage and number of rooms are likely to be highly correlated because they are both explaining the same thing. This problem results in unusual behaviour in the regression model.

Action: correlate your explanatory variables BEFORE doing any regression. Watch out if you have a pair which has an r value of 0.7 or higher. This doesn't mean that you shouldn't use them, but it's red flag. This is what might happen. You take out on one the variables in a regression, the one that's left reverses its sign or suddenly becomes statistically insignificant.

If you do run the regression and you get an unlikely result, choose the variable with the highest t value (or smallest p value) and ditch the other.

10.13 Don't extrapolate

The coefficients from the regression are calculated based on the data you provided. If you try to predict for a value beyond the range of that data, the results will be unreliable if not totally wrong.

In the years of experience and salary example in Chapter 5, the coefficient for years of experience provided the change in salary that an extra year of experience would give. Would you feel comfortable predicting the salary of someone with 95 years of experience.

Action: before running the numbers, check that the inputs are within the range for which you calculated the model.

11. Time Series Introduction and Smoothing Methods

A time series is a set of observations on the same variable measured at consistent intervals of time. The variable of interest might be monthly sales volume, website hits or any other data of relevance to the business. Using time series analysis we can detect patterns and trends and–just possibly–make forecasts. Forecasting is tricky because (of course!) we only have historical data to go on and there is no assurance that the same pattern will repeat itself. Despite all this, time series analysis has developed into a huge topic, fortunately with a large number of freely available data-sets to use. Links to some of these data-sets are provided at the end of the data chapter (Chapter 3).

There are two basic approaches: the **smoothing** approach and the **regression method**. Both methods attempt to eliminate the background noise. This chapter covers smoothing methods, the following chapter the regression approach.

11.1 Layout of the chapter

Here we will

1. identify the four components that make up a time series
2. introduce the naive, moving average methods and exponentially weighted smoothing methods to make a forecast
3. discuss ways in which the accuracy of the forecasts can be calculated

11.2 Time Series Components

There are four potential components of a time series. One or more of the components may co-exist. The components are: trend component; seasonal component; cyclical component; random component.

- **trend component**: the overall pattern apparent in a time series. The plot below shows French military expenditure as a percentage of GDP. The downward trend is apparent.

French military expenditure as a % of GDP
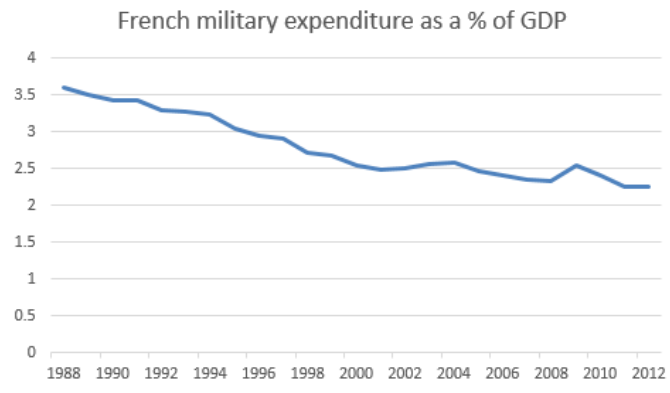
French military expenditure as a percentage of GDP

- **seasonal component**: sales of skis or Christmas decorations are clearly higher in the winter months, while ice cream and suntan cream move more quickly in the summer. If we know the seasonal component, then we can use this information for prediction. The time between the peaks of a seasonal component is known as the **period**. Note here that season doesn't necessarily mean the season of the year.

The plot below shows sales of TV sets by quarter. Here there is a constant upward trend because more people are buying TV sets, but there is also a seasonal effect.

TV Sales by quarter
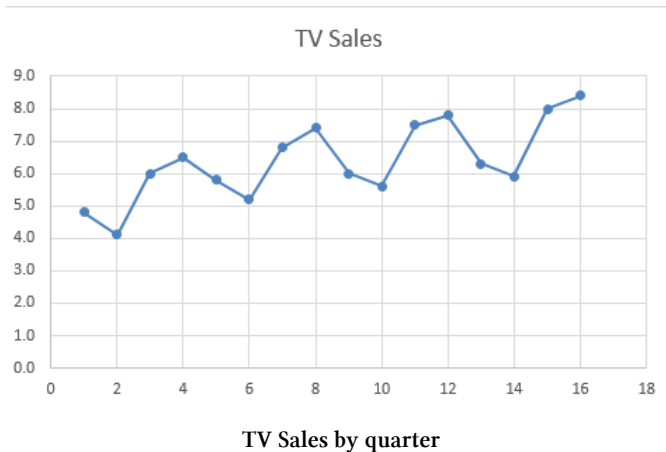
- **cyclical component**: some time series have periods lasting longer than one year. Business cycles such as the 50-year Kondratieff Wave are an example. These are almost impossible to model, but that doesn't stop the hopeful from trying. Kondratieff himself ended up a victim of Joseph Stalin because his suggestion that capitalist economies went in waves undermined Stalin's view that capitalism was doomed, and he was executed in 1938.
- **random component**: the random component is just that: the noise or bits left over after we have accounted for everything else. The tell-tale signs of a random component are: a constant mean; no systematic pattern of observations; constant level of variation.

11.3 Which method to use?

In the next two chapters, we'll work through the two basic approaches: the smoothing approach and the regression approach. How to decide which to use?

First step as usual is to make a simple scatter plot of your data, with time on the horizontal x axis and the variable of interest on the vertical y axis. If the result is something resembling a straight line, for example the French military expenditure, then use a smoothing approach. If there is evidence of some seasonality, such as with the TV sales data, then regression is the way to go. This is especially the case if you want to calculate the size of the seasonal effect.

11.4 Naive forecasting and measuring error

Forecasting is just that: an educated guess about what might happen at some point in the future. Forecasts are based on historical events, and we are hoping that some pattern of behavior will repeat itself which will make the forecast 'true'. The best that we can do is to try out different forecasting methods and work out how well they predicted data which we already knew about. Then we take a leap into the dark and hope that the best of those methods will do a good job on data that we don't know about. This section concerns the measurement of forecasting accuracy.

We're going to construct a **naive** forecast and use that forecast to demonstrate two common methods of accuracy checking: the **Mean Squared Error** (MSE)and the **Mean Absolute Deviation** (MAD) approaches.

A naive forecast assumes that the value of the variable at t+1 will be the same as at t. The values are just carried forward by one time period. Here is an example:

1	Week	Sales	Forecast
2	1	17	
3	2	21	17
4	3	19	21
5	4	23	19
6	5	18	23
7	6	16	18
8	7	20	16
9	8	18	20
10	9	22	18
11	10	20	22
12	11	15	20
13	12	22	15
14			22

Image Gas prices with naive forecast

The naive approach is sometimes surprisingly effective, perhaps because of its simplicity. In general simple models perform well perhaps because of their lack of assumptions about the future. Now we will measure the accuracy of the Naive Forecast with MAD and MSE. In both methods the error is calculated by subtracting the forecast or predicted value from the observed value. The methods differ in what is done with those errors.

MAD measures the absolute size of the errors, sums them and then divides by the number of forecasts. The absolute value of the error is just its size, without the sign. In Excel, you can find an absolute value with =ABS(F1 - F3) where FI is the observed value and FE the predicted value. Use the little corner of the formula box to drag it down. Then find the sum and divide by the n, which is the number of observations. In math this is

$$MAD = \frac{\sum (\text{abserror})}{n}$$

In MSE, the error is squared before being summed and divided.

Squaring the error removes the problem of the negative numbers, but creates another one: large errors are given more weighting because of the squaring. This can distort the accuracy of the results. The maths for the MSE is below

$$MSE = \frac{\sum (\text{error})^2}{n}$$

The plot below shows the errors associated with the naive forecasting, the absolute values of the errors and the MAD. In Excel, use =abs( ) to convert to an absolute value.

	A	B	C	D	E
1	Week	Sales	Forecast	Error	Absolute Value
2	1	17			
3	2	21	17	4	4
4	3	19	21	-2	2
5	4	23	19	4	4
6	5	18	23	-5	5
7	6	16	18	-2	2
8	7	20	16	4	4
9	8	18	20	-2	2
10	9	22	18	4	4
11	10	20	22	-2	2
12	11	15	20	-5	5
13	12	22	15	7	7
14			22		
15				sum	41
16				MAD	3.73
17					

Image Errors and Mean Absolute Deviation

11.5 Moving averages

Slightly more complex than the naive approach is the *moving average* approach, which relies on the fact that the mean has less

variance than individual observations. By focusing on the mean, we can see the general trend, less distracted by noise.

The analyst decides how far back in time the average will go. The longer back in time, then the more values of the observation are averaged, resulting in a flatter more smoothed result. This is good for observing long-term trends, but less satisfactory if you are interested in more recent history.

The number of observations which are to be averaged is known as k, and this number is chosen by the analyst based on experience. In Excel's Data Analysis Toolpak there is a Moving Average tool which can do all the work. Positioning the output is slightly tricky. There are k time periods being used for the calculation of the first forecast; so the first forecast should be at k+1 because k time periods are being used to find the first value. Where exactly to put the first cell of the output is a bit fiddly. If there are k time periods being used, place the first cell at k-1. Excel provides a chart output, example below. There is a Youtube here[1]
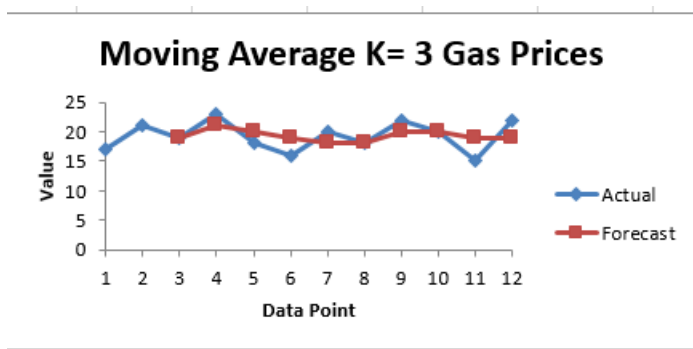
Image Moving average with k=3 for gas prices

The moving average technique is used by investors to detect the 'golden cross' and the 'death cross'. They examine 50 day and 200 day moving averages. When the 50 day moving average is above

[1]http://youtu.be/zrioQOWfxjY

the 200 day moving average this point is the so-called golden cross and a signal to buy. By contrast, when the opposite happens this is the 'death cross'...time to get out!

11.6 Exponentially weighted moving averages

The exponential smoothing method (EWMA) is a further step forward from the naive and the moving average approaches. The moving average forgets data older than the k time periods specified, while the EWMA incorporates both the most recent and more historical observations to construct a forecast. In the Glossary you'll find more detailed math showing how the EWMA brings forward all past history. In general the further back in time you get, the less influence the observations have.

Using EWMA we choose a smoothing constant, alpha, which sets the weight given to the most recent observation against previous forecasts. If we thought that the forecast (t+1) would be similar to the current observation (t=0) and we thought that older forecasts were of little value, then we would choose a high value of alpha. Alpha runs from 0 to 1. The EWMA formula is

$$F_{t+1} = \alpha Y_t + (1 - \alpha)F_t$$

where F is the forecast, and Y is the actual value of the variable. alpha is the smoothing constant, ranging in value between 0 and 1, and chosen by the analyst.

In words, this equation means that the forecast value (at t+1) is the smoothing constant alpha multiplying the actual value at t=0 plus (1-alpha) times the forecast at t=0. So if the previous forecast was totally correct, the error would be zero and the forecast would be exactly what we have today. It turns out that the exponential

smoothing forecast for any period is constructed from a weighted average of all the previous actual values of the time series.

The equation above shows that the size of alpha, the smoothing constant, controls the balance between weighting given to the most previous observation and previous observations. If the alpha is small, then the amount of weight given to Y at t=0 is small and the weight given to previous observations is large and vice-versa. If alpha = 1, then previous observations are given no weight at all, and we assume that the future is the same as the past. This achieves the same result as *naive forecasting* discussed above. Typically quite small value of alpha are used, such as 0.1.

##Application of Excel's exponential smoothing tool

An exponential smoothing tool is available in the Analysis ToolPak. We'll use Canadian Gross Domestic Product per capita as an example. Here is the data plotted on its own. Youtube[2]

Canadian GDP Per Capita
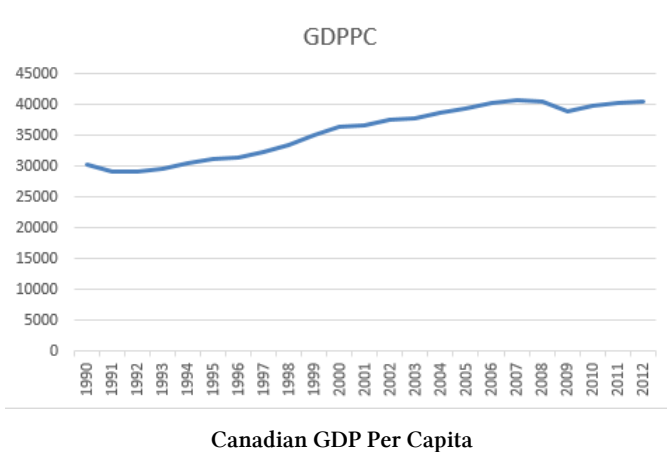
There is a steady upwards trend, but with a dip in 2009 due to the world financial crisis. Excel asks for a damping factor. This is 1 - alpha. I have done the forecasts twice, once with an alpha of 0.1

[2]http://youtu.be/w-GmxjrX1qg

and again with an alpha of 0.9. Therefore the damping factors were 0.9 and 0.1. The result for alpha = 0.9 is shown below.

Forecast with alpha = 0.9
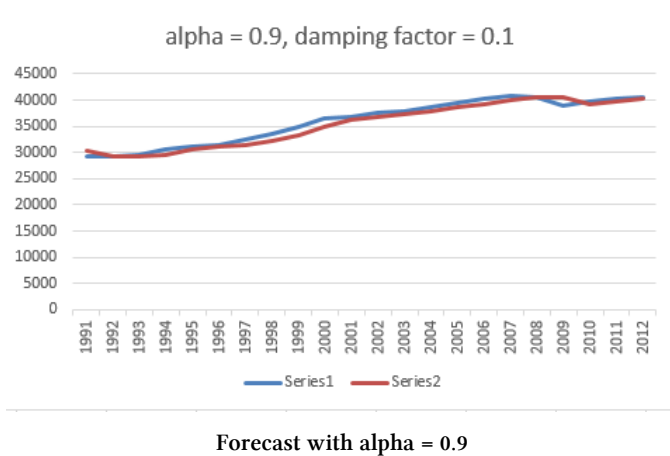
This is pretty good forecast, with the forecast tracking the actual observations closely.

The plot below shows sheep numbers as counts by head in Canada from 1961 to 2006. First, let's plot this using 0.3 as alpha (arbitrarily chosen). The plot is below, with a damping factor of 1-0.3 = 0.7.
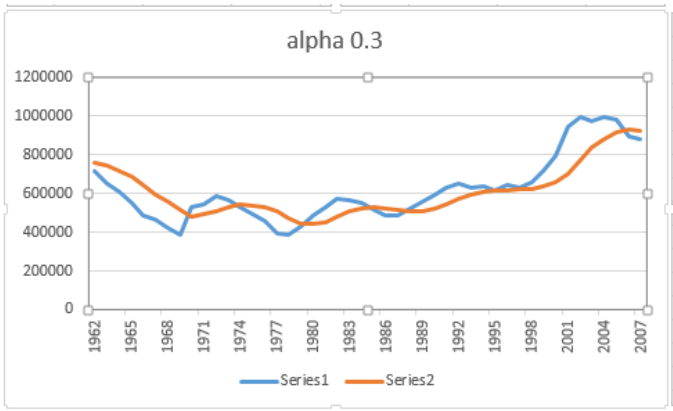
Sheep numbers with alpha = 0.3

12. Time Series Regression Methods

The smoothing methods we worked in the previous chapter are fine when there is no evidence of seasonality, for example with the sheep data. However, smoothing has only limited use for longer term prediction and analysis. Data that is more interesting for us as analysts may be non-linear in trend, and also have seasonal peaks and troughs. Using regression, we can measure the quantitative effect of seasonality and trend, either together or separately. The result is a model which can be used for prediction.

Below we will:

1. use regression to quantify a trend in a time series
2. introduce a quadratic term to account for non-linearity in the time series
3. use dummy variables to measure seasonality

12.1 Quantifying a linear trend in a time series using regression

The plot below shows life expectancy at birth for Canadians.
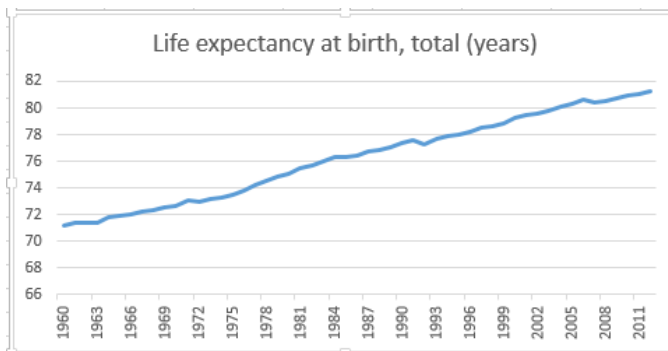
Canadian life expectancy

This is pretty much a straight line as one might expect in a developed country with a large per capita expenditure on health care. Note that this is for both sexes: for women only we might expect to see even better figures. I ran the regression with Year as the independent variable, explaining life expectancy. The result is below.

	A	B	C	D	E
1	SUMMARY OUTPUT				
2					
3	*Regression Statistics*				
4	Multiple R	0.997163796			
5	R Square	0.994335636			
6	Adjusted R Square	0.99422457			
7	Standard Error	0.242976949			
8	Observations	53			
9					
0	ANOVA				
1		*df*	*SS*	*MS*	*F* ig
2	Regression	1	528.5452388	528.5452	8952.658
3	Residual	51	3.010927678	0.059038	
4	Total	52	531.5561665		
5					
6		*Coefficients*	*Standard Error*	*t Stat*	*P-value* L
7	Intercept	58.47245316	0.190590134	306.7969	5.67E-85
8	Year	0.000565205	5.97351E-06	94.61849	5.66E-59
9					

Life expectancy regression results

Notice that the adjusted R-squared value is close to unity, reflecting

the straight line that we see on the graph. The coefficients for the intercept and the independent variable provide this estimated regression equation

$$\hat{y}_t = 58.47 + 0.000565 * \text{Year}$$

The meaning is that for every year after 1960 that a child was born, his/her life expectancy increased by 0.000565 years, or about 5 hours.

12.2 Measuring seasonality

The dataset for TV sales (from Modern Business Statistics by Anderson Sweeney and Williams) contains sales by quarter for four years. There are therefore sixteen observations.

I created a column which I called index so that we can see sales in a consecutive fashion. The first quarter is Spring and the last quarter is winter. It is apparent that quarters which are divisible by four are higher than others, and so forth. Perhaps sales are higher in winter?

We can use the **dummy variable method** of Chapter 5 to determine whether this is true and also the extent of the difference. We will create dummies for Summer, Fall and Winter, leaving Spring as the reference level. Recall that we have four possible states of the season variable, and so we will need k-1, or 4-1 = 3 dummies. It usually doesn't matter which state you choose as your reference level: just don't forget which one you picked. The dataset is below, with a column called Index, which we'll use to measure the time trend.
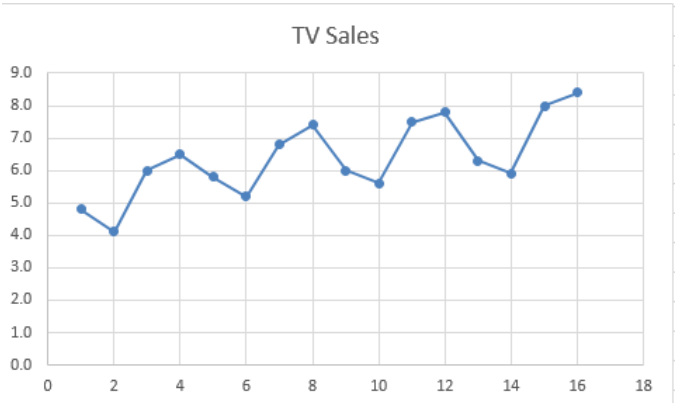
TV Sales with the quarterly dummies

I want to find out two things:

whether sales are increasing over time. I can do this by including the index as an explanatory variable. Because the regression tool requires that all the independent variable be in one block, I have copied and pasted the Index column to the right.

The regression output is

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	4.705625	0.137564422	34.2067	1.6E-12	4.40284775	5.00840225	4.402848	5.008402
Summer	-0.670625	0.153682427	-4.36371	0.001129	-1.00887774	-0.33237226	-1.00888	-0.33237
Fall	1.05875	0.155107642	6.825905	2.85E-05	0.71736038	1.40013962	0.71736	1.40014
Winter	1.363125	0.157454336	8.657272	3.06E-06	1.01657034	1.70967966	1.01657	1.70968
Index	0.145625	0.012111872	12.02333	1.14E-07	0.11896695	0.17228305	0.118967	0.172283

TV Sales regression out

Let's walk through this output line by line. First notice that the p values for all independent variables are smaller than 0.05. Therefore all are significant.

The reference level is Spring, and therefore there is no coefficient for this quarter. Summer has a negative sign in front of its coefficient, meaning that sales for Summer are smaller than those for Spring. Fall is positive, so more TV sales are sold in the Fall than in the Spring. Not unexpectedly, Winter has the largest coefficient of all,

reflecting the plot. The index has a positive sign, meaning that average TV sales are increasing with time.

There are therefore two components in this series: a trend and a seasonal component. Youtube[1]

** Making predictions** from these results. Let's predict the sales for Fall in seven quarters time. This is

$$\hat{y} = 4.7 + 1.05875 + 7 * 0.147 = 6.78$$

The observed value was 6.8, so the prediction was reasonable if slightly pessimistic.

12.3 Curvilinear data

The data above followed a linear trend, which is perfect for ordinary least squares, which assumes that the relationship between the dependent variable and the independent variable is linear. But some interesting data does not follow a neat linear trend. The plot below shows crude oil and gas prices over the period 1949 to 2003.

---

[1]http://youtu.be/wyKIHInMbY8
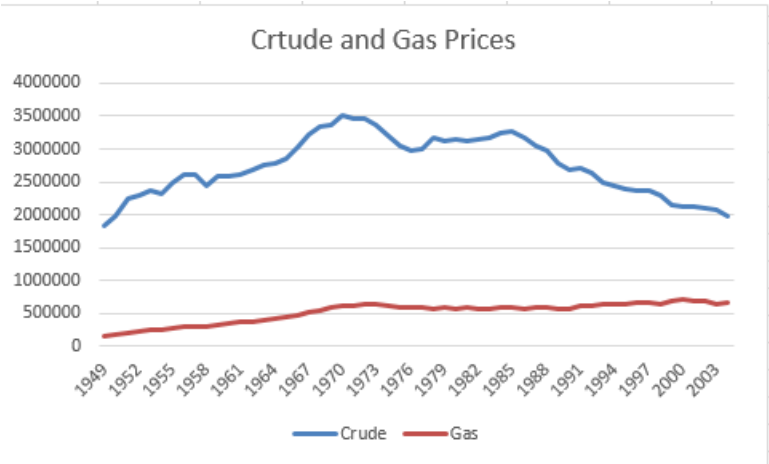
Crude and gas prices showing crude price curvilinearit

However, if we show the two series on the same graph, as on the plot I did in Tableau, with separate y axes for each type of fuel, we can see that the shapes are very close.
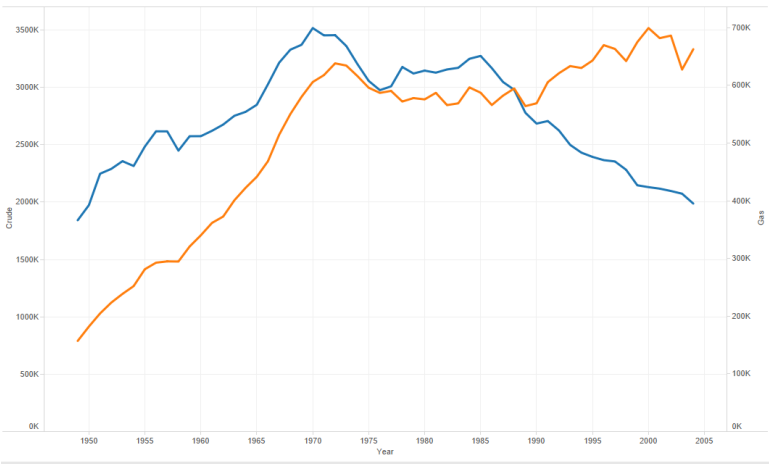
Crude and gas on different axes

The curvilinearity of crude prices is clear. The prices came to a peak

in the 1970s as a result of OPEC's decision to restrict supply. In any event, crude prices cannot be described as linear. The solution is to add an extra term to the regression of price on time, and that is time squared. The first few lines of the dataset are below. I have added a column called to represent the year, and added tsq which is just t squared.

	A	B	C	D	E
1	Year	t	tsq	Crude	Gas
2	1949	1	1	1841940	157086
3	1950	2	4	1973574	181961
4	1951	3	9	2247711	204754
5	1952	4	16	2289836	223515
6	1953	5	25	2357082	238579
7	1954	6	36	2314988	252133
8	1955	7	49	2484428	281371
9	1956	8	64	2617283	292727
10	1957	9	81	2616901	294990
11	1958	10	100	2448987	294749
12	1959	11	121	2574590	320757
13	1960	12	144	2574933	340157
14	1961	13	169	2621758	361689
15	1962	14	196	2676189	372705
16	1963	15	225	2752723	400886
17	1964	16	256	2786822	422471
18	1965	17	289	2848514	441556
19	1966	18	324	3027763	468636

Data with time squared

Now regress crude against time and time squared, and the pleasing result below appears

⊿	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.942064				
5	R Square	0.887484				
6	Adjusted R Square	0.883238				
7	Standard Error	153922.1				
8	Observations	56				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	2	9.90431E+12	4.95E+12	209.0220804	7.19743E-26
13	Residual	53	1.25568E+12	2.37E+10		
14	Total	55	1.116E+13			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%* *Uppe*
17	Intercept	1821447	63977.42731	28.47015	8.56437E-34	1693124.209 19
18	t	99651.67	5178.613405	19.24292	1.42258E-25	89264.68512 110
19	tsq	-1792.25	88.06707176	-20.351	1.0358E-26	-1968.89419 -16
20						
21						
22						

The curvilinear regression for crude

The adjusted r-squared is high at 0.88 and the two time predictors are highly significant statistically. Notice that the two predictors are quite different in size and also have opposite signs. When t is small, then the variable t dominates and the trend is upward. However, as t gets larger, then t-squared gets even larger still. The negative sign on t-squared pulls the regression line down.

13. Optimisation

Linear Programming (LP) is the tool we use to optimize a particular **objective function**. For example, a manufacturer of carpets wants to get the most profit out of his raw materials and labor force. The objective function is the relationship between the inputs and his costs and thus his profits. Optimization helps the manufacturer find the most efficient allocation of resources. The objective function doesn't necessarily have to be in terms of money; it could very well be to allocate working hours efficiently so that everyone has more time off. Frequently we want to maximize the objective function (perhaps make more profit) but we might also want to minimize an objective function, such as costs.

Optimization is not particularly difficult, and we will use the Solver add-in to Excel to do the mathematically tricky parts. What is important is the writing up of a correct model in the first place.

Layout of the chapter

Linear programming is not especially difficult, but the work has to be done in a logical and orderly fashion. In this chapter, we will

*spend some time working through the basic steps in setting up an optimization problem * work through the meaning of the Solver output in terms of what the coefficient values actually mean

13.1 How linear programming works

Linear programming works by solving a set of simultaneous equations. The problem to be solved—to maximize a stock return for

example—is written as a set of simultaneous equations. The equations may be quite simple, but there may be many of them. Linear Programming finds the best solution to the equations. The job is to find the coefficients for the variables in the equation which maximize or minimize the outcome.

We'll work through an example first on paper, and then put it into Excel's Solver add-in to get the solution. The three important steps are:

- Write the simultaneous equations, basically putting into math the wording of the problem. This is probably the toughest part.
- Run the Optimization in Solver, to find the optimal solution.
- Sensitivity analysis to find out how much effect a unit change in a constraint would have.

13.2 Setting up an optimization problem

Optimization problems are usually presented as written questions. The best approach is to determine which are the

- **decision variables** those variables which you can control. For example, in the carpet factory example, the number of hours given to each worker is a decision variable.

It is usually the size of the decision variable that we are trying to optimize. For the workers, we would want them to have exactly the number of hours which produces the most profitable output. Too few hours and the factory is working below capacity. On the other hand, too many and money is being wasted. The decision variables go into what Solver calls the *changing cells*. The convention is to color-code these cells in Excel in red color.

- **constraints** are constants or *givens* which we cannot change. The jurisdiction where the factory is located might have legislation restricting the maximum number of hours that an employee can work per day. We would need to add a constraining equation to take account of this fact, even if the solution was financially sub-optimal.

13.3 Example of model development.

A farmer has 50 acres of land at his/her disposal. He also has up to 150 hours of labor and up to 200 tonnes of fertilizer. He can plant either cotton or corn or a mixture. Cotton produces a profit of $400 per acre, corn $200 per acre. Each acre of cotton requires 5 labor hours and 6 tonnes of fertilizer. For corn the equivalent is 3 and 2. How much land should the farmer allocate to each crop?

Let's call acres of cotton x and acres of corn y. F for fertilizer and L for labor. These four variables are his **decision variables** because he can allocate crops to land and he can also decide how much labor and fertilizer to deploy. He is searching for the combination of the four decision variables which maximizes his profit.

We know he has 50 acres, so the total acreage cannot exceed this amount. This is a constraint, which can be written as an equation, as shown below. Other constraints are the maximum labor and fertilizer. Obviously he cannot use more than he has.

13.4 Writing the constraint equations

The total acreage devoted to each crop cannot exceed 50:

$$x + y \leq 50$$

The labor and fertilizer are also constraints. Refer back to the question to get the amounts of labor and fertilizer per acre per crop.

For labor, the constraint equation is:

$$5x + 3y \leq 150$$

This equation comes about because for each acre of cotton, the farmer needs 5 hours of labor, while for corn it is 3. We do not know the values of x and y, but whatever they are, we know that 5x plus 3 y must be less than or equal to 150, because we only have 150 hours available.

For fertilizer, the constraint equation is:

$$6x + 2y \leq 200$$

13.5 Writing the objective function

What do we want to get out of this: what is the objective? Clearly it is the most efficient mixture of inputs, subject to constraints, producing the largest profit. The objective function is:

$$\text{MaxProfit} = 400x + 200y$$

In words, find the combination of x and y that maximizes the profit, subject to the constraints we have written. The next task is to put all this into Solver.

13.6 Optimisation in Excel (with the Solver add-in)

We will use an Excel spreadsheet and Solver to achieve all three steps.
Optimisation Youtube for the farmer problem[1]

---

[1]https://www.youtube.com/watch?v=WeTgK6wmvSY

Open an Excel spreadsheet so that you can follow along.

Cell coloring conventions

I suggest you follow these conventions for color-coding the cells:

- Input cells. These contain all the numeric data given in the statement of the problem. Color input cells in BLUE.
- Changing cells. The values in these cells change to optimize the objective. Code changing cells in RED.
- Objective cell. One cell contains the value of the objective. Color the objective cell in GREY.

Now I'll work through the farming problem above step by step.

Make sure you have the Solver add-in loaded. To check, open Excel, then click on the Data Tab. If Solver is loaded, you will see the Solver name to the right.
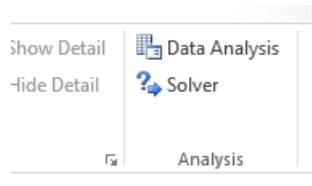
The Solver tab

Below is the Excel spreadsheet with the information that we know already typed in. I have put random values in the red-colored changing cells, just as place holders. These numbers will change when Excel solves for the most profitable allocation.

- Type the address of the objective into the Solver dialogue box, and make sure the Max radio button is selected.
- Type in the range of the changing cells.

- Work through the constraints. There are three: labor; fertil-izer; and land.
- Select Simplex LP as the Solving method.
- Press solve.

A	B	C	D	E	F
		Cotton	Corn		
	labor	5	3		
	fert	6	2		
	profit	400	200		
	Acres	30	0		
	Constraints			Total	
	Labor	150	0	150	150
	Fert	180	0	180	200
	Land	30	0	30	50
	Profit	12000	0	12000	

The farming solution

Solver will change the values in the changing cells to maximize the objective cell. I found that 30 acres of cotton and none of corn provided a profit of $12000.

13.7 Sensitivity analysis

Constraints can be either **binding** or **slack**. If a constraint is binding, that means that all of that particular resource is being used, and more could be employed if it should become available. We can find out whether constraints are binding and also their *shadow*

price by pressing Sensitivity Analysis after running Solver again. The Sensitivity analysis will appear as a tab at the bottom of your worksheet. For the farming problem, it looks like this:

	Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
Microsoft Excel 15.0 Sensitivity Report							
Worksheet: [Book1]Sheet1							
Report Created: 2014-05-15 5:24:33 PM							
Variable Cells							
	C8	Acres Cotton	30	0	400	1E+30	66.66666667
	D8	Acres Corn	0	-40	200	40	1E+30

	Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
Constraints							
	E13	Labor Total	150	80	150	16.66666667	150
	E14	Fert Total	180	0	200	1E+30	20
	E15	Land Total	30	0	50	1E+30	20

Farming sensitivity report

Let's let at the Constraints section. Labor uses 150 hours and has a shadow price of $80. The shadow prices indicates the per-unit value of the constrained commodity if the constraint was increased by one unit. If we could have one more hour of labor, then the profit would increase by $80. Land and fertilizer have a shadow prices of zero because the constraint is slack or non-binding. We are not completely using these resources and so we do not need any extra inputs. There are sacks of fertilizer lying around unused. No need to buy more.

The columns *Allowable Increase* and *Allowable Decrease* are relevant because they tell us how much more of a binding constraint we could use before running the model again. For labor, the amount is 16.67 (rounded). If we eased the constraint by more than this amount then we would need to run the new model in Solver again. The same logic for decrease.

13.8 Infeasibility and Unboundedness

Solver is quite robust, but two problems may occur. A solution to an optimization problem is feasible if it satisfies all the constraints. But it is possible for no feasible solution to exist. This occurs if you make a mistake in writing the model, or the model is too tightly constrained.

Unboundedness

Unboundedness occurs when you have missed out a constraint. There is no maximum (or minimum). Try changing all constraints to >= instead of =< .

13.9 Worked examples

The demanding mother

Most mothers are keen to keep contact with their children. One particular mother is rather demanding. She requires at least 500 minutes per week of contact with you. This can be through telephone, visiting or letter-writing (yes! Some people still do that!). Her weekly minimum for phone is 200 minutes, visiting 40 minutes and letters 200 minutes. You assign a cost to these activities. For phone: $5 per minute; visiting $10 per minute; letter-writing $20 per minute. How do you allocate your time so that your cost is the minimum?

Write the equations first. What is the objective?

Let x stand for minutes of phone; y minutes of visiting; and z minutes of letters. You want to minimize the cost of these activities, so the objective is to minimise 200x + 40y + 200z. The constraints are: x + y + z must be more than or equal to 500.

	Cost				
Phone	5				
Visiting	10				
Letters	20				
	Phone	Visiting	Letters	Actual tot	Total Reqd
Minutes	260	40	200	500	500
Objective	5700				
	Actual		Reqd		
Constraints			500		
Phone	260		200		
Visiting	40		40		
Letters	200		200		

Spreadsheet for demanding mother

Notice that we want to minimize the time, so change the radio button to min rather than max. And when you are typing in the constraints, check that the sign of the inequality is the right way round.

The theater manager

You are the manager of a theater which is in financial trouble. You have to optimize the combination of plays that you will put on to make the most profit. You have five plays in your repertoire, A,B,C,D,E. They have different draw points (appeal to the public) and therefore ticket prices to match. A draw point is how attractive the play is to the general public. The data looks like this:

Play	Draw	Ticket
A	2	20
B	3	20
C	1	20
D	5	35

Play	Draw	Ticket
E	7	40

You have these constraints: you have only 40 possible slots. And no play can be put on less than twice or more than ten times (gives the actors a reasonable turn). Each performance costs $10 per ticket sold, regardless of the ticket price (covers the electricity etc). The total draw points has to exceed 160 to keep the critics happy.

How do you distribute your performances? Which combination of plays produces the highest profit and satisfies the constraints?

We are trying to maximize profit, so look for an objective function that does just that: $20A + 20B + 20C + 35D + 40E$. But wait—-we have the $10 cost. So better take that out first, leaving $10A + 10B + 10C + 25D + 30E$.

With these constraints:

The draw points(the attractiveness of the play to critics): $2A + 3B + 1C + 5D + 7 E >= 160$

and no play can be put on more than twice or more than ten times.

$A <=2$ and $A <= 10$ $B <= 2$ and $B <=10$ $C <= 2$ and $C <= 10$ $D <= 2$ and $D <=10$

and we have only 40 slots, so $A + B + C + D + E <= 40$

Solution:

Play	Draw	Ticket	Performances	Gross Rev	Costs	Net Rev	Draw Points
A	2	20	8	160	80	80	16
B	3	20	10	200	100	100	30
C	1	20	2	40	20	20	2
D	5	35	10	350	100	250	50
E	7	40	10	400	100	300	70

	Num Perf	40		Draw Points	168
				Profit	750

Constraints	Num Perf <=40	40
Each play at least twice		2
And no more than ten times		10
Total Draw Points at least		160

Under Performances, the combination that produces the most profit is shown.

Spreadsheet for theatre problem

More worked examples

19th century farmer

I am a 19th century English farmer. I can grow wheat or barley. Wheat yields 10 bushels an acre, barley 8 bushels an acre. The price of wheat is ten shillings a bushel, barley 5 shillings a bushel. The labour costs for wheat are 3 shillings an acre, for barley 2 shillings an acre. The transportation cost to market for wheat is 1 shilling a bushel, for barley ½ shilling a bushel. I have 100 acres. (a bushel is a measure of volume; an acre is a measure of area; a shilling is a currency unit).

Questions: How do I split up my land?

If I could get one more acre of land, how much would that be worth to me?

Yet another farming question (I use these because most people can imagine fields of land, crops growing and the like. But of course the same techniques are applicable to other business situations).

A farmer has 100 acres of land. He can plant crops or raise sheep. Each hectare of crops provides $100 but requires labor of $50 per acre. He can spend a maximum of $500 on crop labor. Each acre of sheep provides $40 but needs only $10 in labor charges. The labor budget is unlimited.

Worked solution

Call area in crops X and in sheep Y. Then X + Y =< 100 and 50X <= 500 The objective is 100X - 50X + 40Y - 10Y simplifies to 50X + 30Y. My Excel spreadsheet is below.

A	B	C	D	E	F	G	H
			x	y			
	Revenue		100	40			
	Costs		50	10			
	Net		50	30			
						Used	Constraint
	Changing		10	90		100	100
						Used	Constraint
	Labour		500	900		1400	500
	Objective		3200				

the labour budget is only for crops: the question says he can spend only $500 on crop labour. Sheep

Spreadsheet for the above problem

14. More complex optimisation

Optimisation using Solver is a powerful method of solving common business resource-allocation problems. In the previous chapter the problems were relatively simple concerning the allocation of land or time. But linear programming can be used to solve more complex and worthwhile problems as we'll see below. The key requirement is that the analyst is able to define the problem as a set of equations. There is no one method except for thinking carefully and writing out the problem as a set of equations.

To demonstrate, we'll work through three different types of problem:

*problems concerning proportionality: you need to allocate money to different investments while minimizing risk and keeping returns above a certain amount. What proportion do you put in each investment?

*supply chain problems

*blending problems where you need to mix together inputs from different sources

14.1 Proportionality

Investment decisions example

This example shows how we can 'weight' the inputs according to some criteria, in this case their risk. We want to minimise the risk but ensure that the return is above some minimum level. What is the mixture or blend of investments that can do that?

The problem: you have an inheritance of $300,000 from an uncle but there are some restrictions: you must invest all the money in four funds; your annual return has to be at least 5%. And you must minimise your risk. The four funds your uncle has specified are:

Fund	Risk	Return
X1	10.7	4.2
X2	5	4
X3	6	5.6
X4	6.2	4

Let's deal with the objective function first. That is to minimise the risk. We can weight the size of the investment in each fund with its risk index. So, weighting each investment and then dividing by the total investment gives the amount of the risk: which is exactly what we want to minimise. So we want to minimise R (for Risk) like this:

$$R = \frac{10.7X_1 + 5X_2 + 6X_3 + 6.2X_4}{300,000}$$

If you don't get this, think about what would happen if all the funds were as risky as X1: the total risk would increase. Another way to think of this: if we decrease the number of shares in X1 and instead increase the number of X4, what will happen? The risk will decrease.

The constraints: the total investment has to add up to $300,000, so this constraint is

$$X_1 + X_2 + X_3 + X_4 = 300,000$$

We also have to achieve a return of at least 5% (no easy matter these days). This constraint is

$$4.2X_1 + 4X_2 + 5.6X_3 + 4X_4 \geq 5$$

Below is my spreadsheet from this problem:

C	D	E	F	G
	Funds	**Risk**	**Return**	
	X1	10.7	4.2	
	X2	5	4	
	X3	6	5.6	
	X4	6.2	4	
	X1	X2	X3	X4
Changing	0	0	281250	18750
Constraint	300000		300000	
Constraint	5.5		5	
Objective	11.4375			

Investment blending

14.2 Supply chain problems

Working out how much to ship from production centres to demand locations is a common problem in supply chain optimization. As I have been stressing, the key to solving problems of this type is to write out the equations which define the model. Here is an example:

You run a company which has bakeries in location A and B. The bakeries ship cartons of bread to your retail stores at locations X, Y, Z. The bakeries have different capacities and each retail store has different demands. The costs of delivery per carton from each bakery to each store is below

Bakery	X	Y	Z	Capacity
A	12	13	11	150
B	9	17	17	200
Demand	50	100	90	

Notation: a delivery from bakery A to location X is Ax and so forth.

The objective function is to minimize the delivery costs. So we want to minimize: 12Ax + 13Ay + 11Az + 9Bx + 17By + 17Bz

Each bakery has a fixed capacity, so Ax + Ay + Az <=150 and Bx + By + Bz <= 200

Supply has to exactly match demand

Ax + Bx = 50 Ay + By = 100 Az + Bz = 90

Notice the strict equality sign. My results are below:

		X	Y	Z	Used	Total
	A	12.00	13.00	11.00		
	B	9.00	17.00	17.00		
	Demand	50	100	150		
					Used	Total
Changing Cells	A	0	100	50	150	150
	B	50.00	0.00	40.00	90.00	200.00
Constraints	Store X demand	50.00			50	
	Store Y demand	100.00			100	
	Store Z demand	90.00			90	
Objective		1680				

Distribution problem

14.3 Blending problems

Above we discussed linear programming models which were simple but effective. There are other types of Optimisation model which are helpful, especially *Blending Models*, which can also be solved by linear programming.

Blending models are used in situations where we have two or more inputs which have to be mixed to some formula. Through Optimisation we can find the most profitable mixture. Wine, metals, oil, sausages, recycled paper—-this is a powerful technique. You could probably use it for marketing campaigns. We'll work though an example which is for oil.

The oil blending problem

The problem: an oil company has 15000 barrels of Crude oil 1 and 20000 barrels of Crude oil 2 on hand. The company sells gasoline and heating oil. These products are made by blending together Crude oil 1 and Crude oil 2. Each barrel of Crude oil 1 has a quality level of 10, and each barrel of Crude oil 2 has a quality level of 5. The gasoline that we produce must have a quality level of at least 8. The heating oil must have a quality level of at least 6. Gasoline sells for $75 a barrel, heating oil for $60. How can we blend the oils together in such a way that meets minimum quality requirements and maximizes profit?

Oil blending solution

First, let's think through what the decision variables (what goes into the changing cells) might be. You might very well think (as I did first off) that the decision variables would be the amounts of the two oils used and the amounts produced. But this isn't enough: we have to blend together the two types of oil. They have to be mixed before they can be sold, and the mixture has to reach some minimum quality standard. The company needs a blending plan.

The inputs:

selling prices (here gasoline = $75, heating oil = $60) availability of oil from suppliers quality level of crude oils: Crude 1 is 10 and Crude 2 is 5

The constraints

Gasoline quality >=8 Heating oil quality >= 6 Quantity of Crude 1 = 15000 Quantity of Crude 2 = 20000

The blending plan:

Gasoline has to have a minimum quality of at least 8, and heating oil must have a minimum quality of at least 6. The Crude oil 1 we

have on hand has a quality level of 10 while Crude oil 2 has a quality level of 5. We want to blend these two crude oils to both achieve the minimum quality standards and make the greatest profit.

Let's attack the problem by creating total 'quality points' (QPs) which represent the quality of oil in a barrel multiplied by the number of barrels of that oil.

Write equations to calculate the quality points:

Total QPs in the gasoline = 10 * amount of Oil 1 + 5 * amount of Oil 2

If for example, we mixed together 50 barrels of Crude oil 1 and 40 barrels of Crude oil 2, the total QPs would be 50 x 10 + 40 x 5 = 700. The average per barrel would be 700/90 = 7.78. This is too low for gasoline (needs 8) but acceptable for heating oil. Two points:

we could sell the oil as heating oil, but it is exceeding the minimum quality requirement. We could make more profit by reducing the quality to the minimum, or charge a premium. But this is prescriptive work, and we have to work within the inputs given to us).

The only way to get the oil up to gasoline standards is to increase the amount of Crude oil 1 in the mixture.

If you don't get this, try a thought experiment: for gasoline, if there was no oil at all from Oil 2, what would be the QP? It would be 10 * the quantity of oil from Oil 1. Again, how about if we blended together 1000 barrels from each type of oil: how many QPs would be produced? It would be 10 * 1000 + 5 * 1000 = 15000. Read this through again…it is important.

The blending plan provides us with the constraints we need to ensure that the minimum quality levels are achieved. In the example just above, we blended 2000 barrels to provide a QP of 15000. This is an average of 7.5. Good enough for heating oil, not good enough for gasoline.

	Gas	Heating		
Selling Price	75	60		
Min Qual	8	6		
	Crude 1	Crude 2		
Qual	10	5		
Quant	15000	20000		
Blending Plan				
			Total	Constraint
Crude 1	3000	2000	5000	15000
Crude 2	2000	8000	10000	20000
Total	5000	10000		
QPs reqd	40000	60000		
QPs provided	40000	60000		
Profit	975000			

Oil problem solution

Discussion of the solution: note that gasoline sells for more money than heating oil, but the optimal solution suggests that we should sell more heating oil than gasoline. This is because of the constraints on quality.

15. Predicting items you can count one by one

Why you need to know this. Many business decisions involve counts: either binary (yes/no) or within time or space. It would be good to know the probability of a certain number of customers coming up to a service desk in a certain length of time; or the probability of a certain number of car accidents at a given intersection. We are looking for a discrete probability distribution; discrete because the number of occurrences is an integer.

Chapter 4 on regression showed how to predict the size of an outcome which was continuous (money or time perhaps). The dependent variable—what we were trying to predict—could take on almost any value. Now we want to predict probabilities for the occurrence of an independent variable which is an integer. Below we'll work through some hands-on applications, with the theory available in the Glossary.

We'll break this into two parts:

The probability of a **binary outcome**. The probabilities of two faulty items out of the next twenty on the production line. Or, the probability of at least five customers out of the next fifty actually buying something. This is estimated with the **binomial distribution**.

The probability of a particular count of occurrences over time or area. The probability of three or fewer people arriving at your Customer Service Desk within the next half hour. Or more than two mistakes in the next ten lines of code. This is estimated with the **Poisson distribution**.

15.1 Predicting with the binomial distribution

After the normal distribution (see the Glossary for a definition), the binomial is the most important in statistics. The math for the binomial distribution is also defined in the Glossary.

The binomial distribution provides the probability of a 'success' in a certain number of 'trials'. For example, you can calculate the probability of more than seven out of the next twenty people through the door actually buying something. Here a 'success' is somebody buying something; while the number of 'trials' is the number of people coming through the door (here it is twenty).

Some definitions:

What we are looking for is the probability of a pre-defined number of 'successes'. Note that the definition of success is up to the analyst. It could be 'spending more than $20' or wearing a hat.

In the example we'll work through below you run a store. You know that the long-run probability of somebody buying something is 0.6. You obtained this number by counting total numbers of customers and, out of those, the numbers who actually bought something (successes).

Worked example: You want to know the probability of exactly three people through the door out of the next five buying something. Note that 'success' just means that the defined event happens. Whether or not it is a 'good thing' isn't relevant.

For Excel, the arguments required are: the random variable whose probability we want to predict; number of trials, the long-run probability, and a true/false statement. In the example, the number of trials is five. The random variable (X) whose probability we want to predict is 3. We also need the long-run probability of a success. In the example, p = 0.6. We also want the probability of exactly 3, so use the false statement. I'll go into this in some detail shortly.

Open an Excel spreadsheet, and type in =BINOM.DIST(3,5,0.6,false) and you should get this: 0.3456. This is the probability of exactly three people out of the next five buying something (number of successes out of the next five trials). Notice the order of the arguments in the Excel function. And especially the last one, which in the example above is false. (The alternative is true). The difference is important because the results are quite different.

True and false argument. Defining the argument as false provides the probability of exactly the random variable. Defining the argument as true provides a cumulative probability.

Excel adds up probabilities from the left, so changing the arguments to =BINOM.DIST(3,5,0.6,TRUE) = 0.66304 is the sum of the probabilities of X=0 + X=1 and so on up to an including X=3. The probability of 0.66304 is therefore the probability of **three or fewer** customers buying something. The table below shows the probabilities of various values of X, both 'false' and 'cumulative'. As you can see, the cumulative is just the continued addition of each successive probability. There is a further example here[1]

X	P(x)	P(x), cumu
0	0.01024	0.01024
1	0.0768	0.08704
2	0.2304	0.31744
3	0.3456	0.66304
4	0.2592	0.92224
5	0.07776	1

Probabilities calculated both false and true

How about more than three customers buying something? In math notation, we're looking for $P(X >= 3)$. We know that probabilities must sum up to 1. We know that three or fewer is 0.66304. So more than five has to be: 1 - 0.66304 = 0.33696.

---

[1] https://www.youtube.com/watch?v=oZ1DmNQ8wW4

A slightly harder example: the probability that **at least** four customers out of the next ten will buy something? We're looking for P(X >=4). Look carefully at the notation. X >= 4 implies that the distribution of the ten customers is split into two halves: less than four and four or more. It is the latter half whose probability we're looking for.

If we can find the probability of 0 + 1 + 2 + 3, that is the probability of less than four (we only want integers here). So find that probability and then subtract from 1, making use of the fact that probabilities must sum up to 1.

Let's do this step by step.

First, find P(X = < 3), that is the probability that X is three or less: =BINOM.DIST(3,10,0.6,true) = 0.054762. Notice the 'true' which gives us the cumulative probability.

We want four or more, so subtract from 1 like this: 1 - 0.054762 = 0.945238.

Another example. It's winter and you need to wear a sweater every day. You have two blue and three red sweaters. Calculate the probability that during the week you will wear a red sweater:

Exactly twice in the week

Answer: the long-run probability of picking a red sweater is 3/5 = 0.6 because you have five sweaters, and three of them are red. The number of trials is 7 because there are 7 days in a week. The wording of the question contains the word 'exactly' which means that we don't want a cumulative answer, so we'll include the FALSE argument. Therefore the answer is =binom.dist(2, 7, 0.6, FALSE) = 0.077414

More than three times. When you see words such as 'more than' that's a clue that you're looking for a cumulative probability. In math notation, we're looking for P(X>3). So if we find the cumulative probability up to and including two, and then subtract from one, we're done. The cumulative probability of two

or fewer is P(X=0) + P(X=1) + P(X=2). So the answer is = 1 - binom.dist(2,7,0.6,true) = 0.903744

The keys:

Write out what you are trying to predict in math notation. This forces you to be clear. Draw a little sketch (hopefully better than mine!) if you get confused.

If the wording of your problem contains 'exactly' or requires the probability of just one particular outcome (eg P(x=3)) then you want to use false.

15.2 Predicting with the Poisson distribution

The binomial distribution gave us the probability of a binary outcome (yes/no) out of a certain number of trials. The Poisson gives us the probability of a certain number within a specified time-frame or area.

Here's the example. You run the Customer Service Desk. You want to know the probability of five or fewer customers arriving in the next half hour. That would be useful for staffing, wouldn't it? You know that the on average, 20 customers arrive every half hour. You know that because have counted them.

We want: P(X>=5). Like the binomial above, the Poisson has a true/false argument for whether we want cumulative probabilities or 'exact'. The notation include a > sign, which implies cumulative, so we use TRUE. In Excel: =POISSON.DIST(5,20,true). The answer is 7.19088E-05. This looks a bit weird, but it is just math notation. The E-05 means that you should move the decimal point 5 places to the left. So there are four zeroes in front of the leading 7. In other words, an extremely small probability. Perhaps send some staff out for lunch? It is very small because 5 is a long way from your long run average of 20.

16. Choice under uncertainty

Virtually all business decisions involve at least some degree of uncertainty. For example, you don't know (and presumably can never know) some future 'state of nature' which might be an exchange rate or business rental. A farmer does not know the price of wheat at harvest time, but has to decide how much to plant many months ahead. His decision is relatively simple compared to a decision-maker confronted with an opponent who will react to take advantage of whatever decision you do make. The first situation is called 'non-strategic' and is easily handled by the expected monetary value method which we'll work through in the chapter. The second 'strategic' situation is more complicated but can be solved by an application of game theory. We'll cover the EMV method in some detail below. Game theory is an enormous subject and we don't cover it here, but I'll give an example at the end of the chapter along with some recommended reading.

16.1 Influence diagrams

An influence diagram is a sketch of the influences and outcomes of a decision. The influence diagram allows us to think through and depict the forces that influence our choices without having to assign probabilities. It is customary to use oval shapes for variables which are uncertain, a box for a decision node; and lozenges for the outcome.
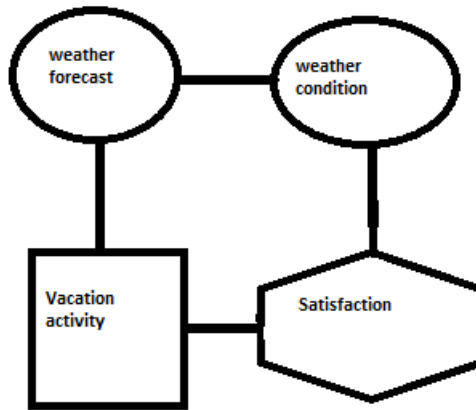
Vacation Influence diagram

You are an outdoors-type and so the choice of activity is affected by the weather. The weather forecast is affected by the weather condition (one hopes!). But you don't know the actual weather condition, just the forecast (which is why it is a forecast). You choose your activity (box) based on the forecast. The amount of satisfaction you get (lozenge) depends on the activity you chose AND the actual weather condition: observe the two lines leading to the lozenge. There is a standard format:

1. Decision Nodes are presented in rectangles. In the example on the slides, the decision is what activity to pursue when on vacation (climb a mountain? read a book?). Which choice we make may to some extent be determined by the weather forecast.

2. Uncertainty Nodes are presented in ovals. The weather forecast is uncertain and so is the actual weather itself.

3. Outcome Nodes are presented in lozenges. The outcome in the vacation example is how much satisfaction you took from the choice you made for vacation activity. This is a utility.

4. Arcs display the flow of the influence. The actual weather condition affects the weather forecast, which affects your planning. Notice that you are taking a decision on activity in advance of knowing what the actual weather will be. That is why there is no arc between weather condition and vacation activity.

You can draw influence diagrams quite easily in Excel using the shapes drop-down tool.

There are different outcomes here, which we can be represented as payoffs. The highest payoff comes from matching your choice of activity to the weather forecast and then finding that the forecast matched reality.

16.2 Expected monetary value

the expected monetary value, or EMV, is just the value of some outcome given a particular state of nature, multiplied by the probability of that state of nature. Here, I am using the customary expression 'state of nature' not necessarily to refer to anything in the natural world, but to a particular set of conditions.

The state of nature has to be mutually exclusive for the probabilities to work. For example, if we have separate probabilities for rain and sunshine, then cannot calculate for rain AND sunshine. The payoffs which result from each state of nature are different, depending on the state of nature.

Let's motivate this with an example using a friendly farmer as the decision-maker. He has the choice of planting wheat, raising cattle, or some mixture of both. Those three potential decisions represent his range of choices. From experience, he knows how much he can expect to receive depending on the weather at the time of harvest. That amount is called the **payoff**.

He faces three states of nature, representing different weather conditions at harvest. These are rain, clear, and sunny. For each choice and for each state of nature he knows the payoffs which are represented in the cells below.

		States of Nature		
		Raining	Clear	Sunny
D1	Mixed	10	10	10
D2	Cattle	-10	20	40
D3	Wheat	-30	30	70

The payoff table for the farmer

There are three decisions and three outcomes, making a total of 9 possible payoffs. Payoffs can be negative, and are not always in terms of money. They could be time, or any other appropriate metric. In the rows we put the choices available to the decision-maker. In the columns we put the 'states of nature' which are the future events not under the control of the decision-maker.

The farmer calculates the probability of each of the three outcomes by working through old weather records. He finds:

probability of raining is 0.4 probability of clear is 0.3 probability of sunny is 0.3 (of course these probabilities must all add up to 1. There could be many more 'states of nature' but I have kept them to three for clarity.

The **Expected Value** is the sum of each probability multiplied by the outcome. In math notation this is:

$$EMV = \sum p_x * x$$

which in words is: the expected monetary value is the sum of all the outcomes multiplied by their probability. In Excel, use the formula =sumproduct(array1, array2). This is a mean, or average, with each outcome weighted by its probability. In decisions involving money,

we usually call the Expected Value the Expected Monetary Value (EMV). For the farmer, the EMVs are as below:

			States of Nature				
						EMV	Re
			Raining	Clear	Sunny		
D1	Mixed		10	10	10	10	
D2	Cattle		-10	20	40	18	
D3	Wheat		-30	30	70	30	
Prob			0.4	0.3	0.3		

The EMVs for the farmer

I multiplied (horizontally by decision) the payoff by the probability of that payoff and then summed. Usually we choose the decision with the highest EMV, so the farmer should choose wheat. Expected value Youtube[1]

Note that a 'good' decision is making the best decision at the time with the information to hand at that time. There may be unlucky consequences but provided the analyst has done a thorough job in selecting the optimal outcome at the time, then she should not be blamed!

It will be highly unusual for an outcome of 30 to actually occur. The EMV is just a weighted average and not a prediction.

16.3 Value of perfect information

The farmer might be tempted to ask for advice from a consultant. What is the maximum he or she should pay for the advice? It is the difference between the best case that the farmer calculates and how much he would receive with perfect information. We calculate the value of perfect information by comparing the EMV of the best choice in each state of nature with the EMV without information. If you knew that the weather would be *poor*, you would select Cattle; *clear* wheat; *sunny* wheat again. Then find the EMV with perfect information by multiplying these best options by their probabilities.

---

[1]https://www.youtube.com/watch?v=fR7cBts1C1w

The working is: $100.4 + 300.3 + 70^*0.3 = 4+9+21 = 34$. The best we could come up without perfect information was 30, so the value of perfect information is 34 - 30 = 4. So if someone offered you perfect information for a price, the highest that you would pay for the perfect information would be not more than 4.

16.4 Risk-return Ratio

There are other ways of choosing the optimal outcome other than the EMV, although the EMV remains the most commonly used. One common measure is the **Return to Risk Ratio**, which provides the dollars returned per dollar put at risk. For each decision, we divide the Expected Value of the decision by the Standard Deviation of the outcome for that decision. We usually choose the decision with the highest RRR, because then the dollar return for each dollar put at risk is highest. The farmer's worksheet is below.

Mixed has a zero because the payoffs are all the same. There is no risk. Cattle has the highest at 0.715. This is higher than the RRR for wheat although wheat has the highest EMV. The farmer might want to think more deeply about the probability of sunny weather at harvest time, because this makes a huge difference.

16.5 Minimax and maximin

A non-probabilistic approach to making choices is the use of maximin and maximax. You can see that your choice depends on whether you are pessimistic (maximin) or optimistic (maximax). If we can somehow obtain probabilities, then a probabilistic approach works best.

16.6 Worked examples

You are planning to market a new coffee-flavored drink. You have a choice between packing the drink in a returnable or non-returnable packaging. Your local government is debating whether non-returnable bottles should be prohibited. The table below shows your profits. If the non-returnable law is passed, you will still get some sales from exports.

Decision	Law passed	Law not passed
Returnable	80	40
Nonreturnable	25	60

Example 1

1. What would be the best decision based on maximin and maximax criteria?
2. A lobbyist tells you that the probability of the government banning non-returnables is 0.7. Assuming the lobbyist is right, what is the best decision based on the EMV?
3. At what level of probability would your decision change?
4. How much would you pay for perfect information?

Answers:

1.Maximin: the worst are 25 and 40. The best is 40, meaning package with returnables. Maximax: the best are 80 and 60. Again returnables.

1. The EMV for returnable is $80 \cdot 0.7 + 40 \cdot 0.3 = 68$. For nonreturnable it is $25 \cdot 0.7 + 60 \cdot 0.3 = 35.5$. Under EMV, you should use returnable bottles.
2. Set this up as equation, with p the unknown which has to balance both sides $80p + 40(1-p) = 25p + 60(1-p)$. Solve for

p and get 0.267. So if you hear rumours that suggest the probability is coming down, think again?

3. If we had perfect information, the EVPI is $80 \times 0.7 + 60 \times 0.3 = 74$. So you should not pay more than $74 - 68 = 6$

Example 2

You run a bank. A customer wants to borrow $15,000 for 1 year at 10% interest. You believe that there is a 5% chance that the customer will default on the loan, in which case you will lose all the money. If you don't lend the money, you will instead place the $15,000 in bonds which return 6% but are risk-free.

1. What are the EMVs of loan and not loan?
2. You have a credit investigation department which can help you with more accuracy in the probability of default. What is the most you should be willing to pay for their advice?
3. Calculate the level of probability of default at which lending the money and investing in bonds have equal EMV.

17. Accounting for risk-preferences

What this chapter is about: the previous chapter showed how we could pick the most attractive choice given payoffs and probabilities. But it might have struck you that there was little discussion of the risks involved and how different people relate to risk. In this chapter we are going to work through picking optimal decisions, taking into account the risk preferences of the decision makers.

Here is a question for you: you are given a lottery ticket which has a 0.5 probability of winning $10,000 and a 0.5 probability of zero. The EMV is therefore $10,000 \times 0.5 + 0 \times 0.5 = \$5000$. Someone comes along and offers you $3000 for the ticket guaranteed. Would take the sure thing? Or would you hope that you win the $10,000? Most people are 'risk averse' and would prefer to give up some of the EMV in exchange for certainty. The $3000 (or however much it is) is called your Certainty Equivalent (CE) in this particular gamble. The difference between the EMV and the CE is called the Risk Premium. So

RP = EMV - CE

You can think of the certainty equivalent as a selling price. It is usually small when the size of the gamble is small, but increases as the gamble gets larger. Here I am using the term 'gamble' because there are probabilities involved. The certainty equivalent is a useful concept which, amongst other qualities, allows us to categorize approaches to risk.

- **Risk-averse**. If your CE is less than the EMV of a gamble, then you are risk-averse (and most people are, so nothing to worry about. You're 'normal')

- **Risk-neutral**. Your CE matches the EMV. You play the averages.
- **Risk-seeking**. You enjoy the gamble inherent in the EMV and need to have a very high CE before you'll give it up. Most people in business who are risk-seeking usually don't last very long. Although we sometimes see risk-seeking decisions when your business is in trouble and a huge gamble is the only possible solution

Business decisions are all about risk, and the probabilities themselves are often unreliable and hard to find. The farmer whose cropping decision we studied in Chapter 2 doesn't know tomorrow's weather, let alone the weather in six months. We generally prefer to give up some of the EMV in return for more certainty because we are risk averse. That's why we buy insurance. I know that I will be covered if I smash my car, and I can even put up with the knowledge that people in Zurich are getting richer because of my aversion to risk.

17.1 Outline of the chapter

Here's what we're going to do:

- Describe utility and calculate utilities
- Show how to use the exponential utility function to calculate utilities based on risk tolerance
- Convert those utilities to certainty equivalents which we can rank and compare with the EMVs of the same decision

To rank choices in terms of their certainty equivalent rather than their EMV, we need the concept of utility, which is a numerical measure of how well a good or service satisfies a need or want. Do you prefer to be reading this book, outside eating an ice-cream, or

talking with a friend? You can rank these choices quickly in your head. You can subconsciously allocate utility numbers to the choices and then pick whichever has the highest utility.

Utility numbers help us to rank our preferences, but there are no units and the ranking is individual-specific. In other words, given a set of alternatives, different people may well have different rankings for them. For the moment, let's assume that it is just you whose utilities we are interested in. If we can somehow figure out your utilities for the different outcomes of a business decision, we can use those utilities to help you make a more psychologically satisfying choice.

Utility

The plot below shows a hypothetical utility curve. The utility value is on the vertical y axis and the outcome of a gamble is on the horizontal x axis. Notice two things:
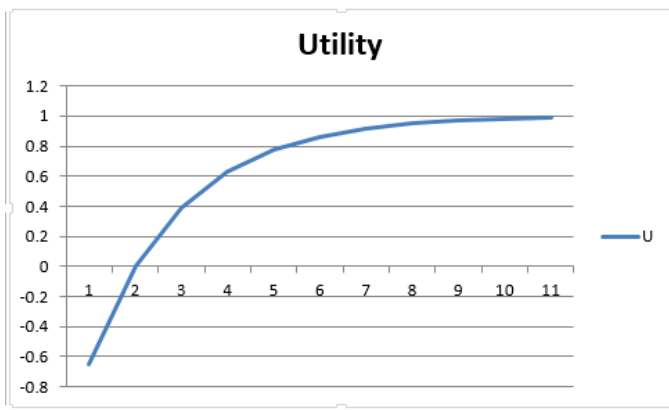
A typical utility curve

*The slope of the line is upwards. This reflects the fact that everyone prefers more money to less

- The rate of increase of the line is decreasing or slowing down. It was steep early on, but towards the end it is almost a plateau. This is because the value of each extra dollar is slightly less than that of the previous dollar. This is the marginal utility of money.

17.2 Where do the utility numbers come from?

There are two approaches.

The first involves asking the decision-maker his/her choice between two gambles.

The second uses a utility function, a mathematical function which converts two input variables (risk tolerance and the size of the outcome) into one utility number.

With a utility number in hand we can begin the process of ranking the decisions in terms of utility rather than EMV.

The intuitive model

We'll work through the intuitive model first and then apply the same thinking to the equation model. Once you get the hang of it, the equation model is quicker and probably more accurate. We'll use the payoff table from the farmer example, repeated below.

			States of Nature					EMV	Re
			Raining		Clear		Sunny		
D1	Mixed		10		10		10		10
D2	Cattle		-10		20		40		18
D3	Wheat		-30		30		70		30
Prob			0.4		0.3		0.3		

The payoff table for the farmer

We'll assign a utility number of 1 to the highest payoff and then zero to the lowest. The beginnings of the utility table looks like this:

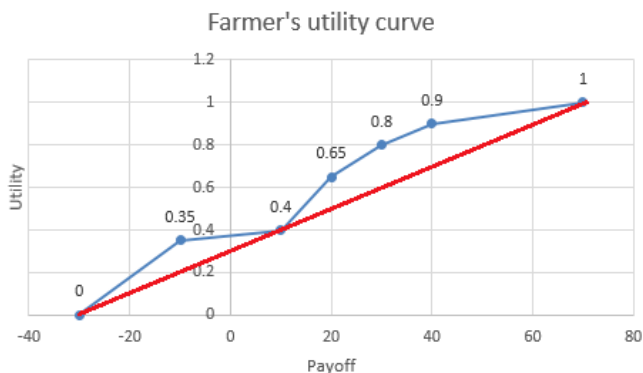| Payoff	Utility
70	1
40	
30	
20	
10	
-10	
-30	0

Now we need to fill in the gaps. What we'll do is to use each payoff as a certainty equivalent, and ask this question:

What value of p would make you indifferent between either receiving a guaranteed payoff of the certainty equivalent (in the first case 40) or accepting a gamble of receiving the highest payoff (70) with probability p or losing 30 (the lowest payoff) with probability 1-p. Let's say that you answer p = 0.9. The expected value of the gamble is 70 * 0.9 + 0.1(-30) = 60.

This is higher than your certainty equivalent of 40, indicating that you are risk averse. The difference between the expected value and the certainty equivalent is the risk premium, and is the amount the individual is willing to give up to forgo risk.

Continue downwards, certainty equivalent by certainty equivalent and complete the table. I've made up the numbers just for illustration. A plot of the utility values is also shown.

Payoff	Utility
70	1
40	0.9
30	0.8
20	0.65
10	0.4
-10	0.35
-30	0

Farmer's utility curve

If you were risk-neutral, you would prefer to take the gamble and so would be an EMV maximizer. The choice of the risk-neutral person is indicated by the straight line.

Now, let's replace the payoff values in the farmer table with utilities. We can calculate the expected utilities in the same way as we found the EMVs. The table shows both for comparison. We multiply the utility of each outcome with its probability.

Decision	Poor	OK	Good	EU
Cattle	0.4	0.4	0.4	0.48
Mixed	0.35	0.65	0.9	0.58
Wheat	0	0.8	1	0.52

Expected utilities

The famer's EMV decision would have been wheat (EMV=14) but taking into account his risk preferences, mixed gives the highest expected utility. It makes senses to spread the risk between two completely different crops.

The exponential utility curve method

It is rather tedious asking so many questions, plus people get tired and confused rather quickly. An attractive alternative is to use a

mathematical function, the exponential utility curve. This requires only one input variable, R, the tolerance for risk. The equation is below:

$$U_x = 1 - e^{-x/R}$$

Here x is the payoff; Ux is the utility of the payoff x; and R is the individuals risk tolerance. A person with a large value of R is more likely to take risks than someone with a smaller R value. As the R value increases, the behavior approaches that of the EMV maximizer. The plot for someone with a risk tolerance of R = 1000 is
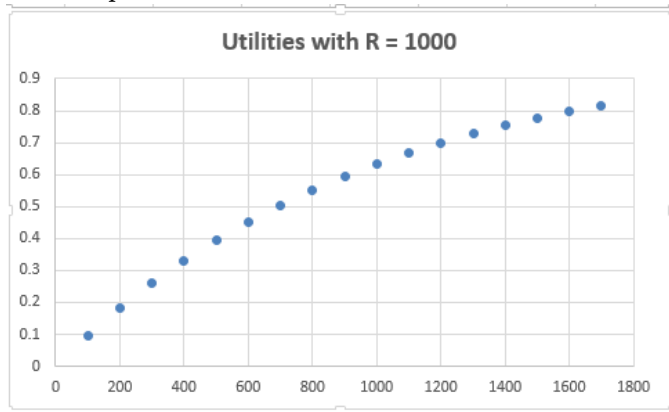

below.

Actually performing the calculations to find utility at each level of payoff is easy using Excel, as we'll see below. But first we have to determine R, the individual's tolerance for risk.

Finding risk tolerance

There are two approaches

By asking this question: consider a gamble which had equal chances of making a **profit** of X or a **loss** of X/2. What is the value of x for which you wouldn't care whether you had the gamble. In other words, what is the value of x for which the certainty equivalent is

zero? The expected value of the alternative is 0.5 x X - 0.5 * (X/2) = 0.25X. As long as X> 0 then the decision-maker is displaying risk-averse behavior, because his/her certainty equivalent is less than the expected value of the gamble. The greater the R, the more tolerant of risk. In his book, Thinking Fast, Thinking Slow, Daniel Kahneman notes that the '2' in the denominator can vary a tad. It isn't a precise formula, but apparently it comes close: In diagram form:

Indifferent between —< Winning X / Losing X/2
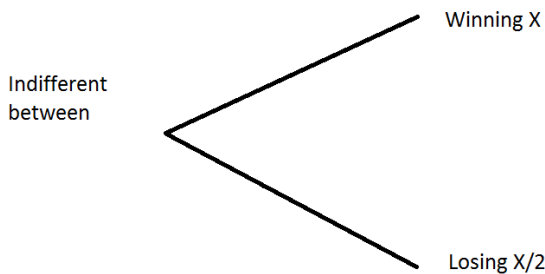
An alternative approach, more used in business and finance, is to employ guideline numbers calculated by Professor Ron Howard[1] , a pioneer of decision-analysis. Based on his years of experience, Howard suggests that the R value for a company is:

- 6.4% of new sales
- 124% of net income
- 15.7% of equity

##Example decision using utility maximisation

What we'll do here is to examine a decision from both the EMV and Expected Utility angles, using the exponential utility function. The calculation of expected values is the same: the difference is that instead of multiplying each monetary payoff by its probability and then adding up, we multiply the utility.

---

[1]https://profiles.stanford.edu/ronald-howard

The decision set-up. All money figures in millions. I run a coffee importing business. I need a special import license to bring in a particular type of coffee. The equity of my company is 12.74. So, my R value is 15.74% of 12.74 = 2.

I have a choice between 'rush' and 'wait' for the permit.

If I rush, I have to pay a special fee of 5, and there is a 50/50 chance of the permit being granted. If it is granted, I will have sales of 8, giving me a net profit of 8-5 = 3. If it isn't granted, I still have to pay the 5, but will have sales of only 6, giving me a net profit of 2. First, find the EMV and EU of rushing. The EMV is 0.5x3 + 0.5(2) = 2.5.

Using the exponential curve formula, the utility of 3 is = 1 - exp(-3/2) = 0.77687. And the utility of 2 is = 1 - exp(-2/2) = 0.632121. The EU for rushing is 0.5(0.77687) + 0.5(0.632121) = 0.704496. All figures are in millions.

Now for the alternative, which is to wait. In this scenario the probabilities are the same at 50/50, but the cost is only 3 if it is issued, nothing if it isn't. The sales are the same at 8 and 4. The net profits are 8 - 3 = 5 and 4. The EMV is $0.55 + 0.54$ = 4.5. The utilities are U(5) = 1 - exp(-5/2) = 0.918 and U(4) = 1 - exp(-4/2) = 0.865. The EU is $0.50.918 + 0.50.865$ = 0.892.

The table below shows the EMVs and EUs.

	A	B	C	D	E	F	G	H
1	Action	Status	Cost	Sales	Net	EMV	Utility	EU
2	Rush	Granted	5	8	3		0.77687	
3		Not Granted	5	4	-1	1	0.632121	0.704496
4	Not rush	Granted	3	8	5		0.918	
5		Not granted	0	4	4	4.5	0.865	0.892
6								

Spreadsheet for the rush decision

But we can go further using certainty equivalents, subject of the next section.

17.3 Converting an expected utility number into a certainty equivalent.

Fortunately there is an easy way to convert expected utilities back to certainty equivalents. It is then straightforward to rank the decisions by certainty equivalents. Remember the concept of the certainty equivalent? The amount of money which it would take for you to sell your gamble? It turns out that we can convert an expected utility number into a certainty equivalent using an equation:

$$CE = -R * \ln(1 - EU)$$

where ln is the natural logarithm. We use this equation where we are talking profits and want more. In the case of costs, take off the negative sign in front of R.

Above we found EU = 0.704496 for the rush decision branch. That is a CE of

= - 2*LN(1-0.704496) = 2.43814

using Excel's ln function. For the wait decision, the EU was 0.892, giving a certainty equivalent of = - 2*LN(1-0.892) = 4.451248

Larger CEs are preferred when the outcomes are profits, and smaller CEs when the outcomes are costs. So I'd wait! This confirms the decision made with the Expected Utilities.

18. Writing up your findings

In chapter 1, I mentioned some of the questions I wished I had been able to answer when I was operating a business. Now I want to suggest ways in which you might structure your response to such questions. The actual answers come when you've done the statistical work (carried out in the following chapters), but it might help to have at least a structure on which you can begin building your work.

Before you do any work on the problem, get straight in your mind exactly what you are being asked to do. You need to be clear and so does the decision-maker to whom you are going to submit your findings. Here is one way to do this:

Copy down the key question that you are being asked in exactly the way that it has been given to you. For example, let us say that you have been asked to answer this admittedly tough question: 'What factors are important for the success of a new supermarket?'

Now keep rewriting it in your own words so that it becomes as clear as possible. Make sure that each and every word is clear. What does 'success' actually mean? —in the business context probably 'most profitable'. What does 'new' mean? Is this a completely new supermarket or a new outlet of an existing brand? You might end up with: 'when we are planning a location for a new outlet for our brand, what factors contribute most to profitability?' Doing all this helps you to identify the statistical method most suitable for the task. You also can include here the hypotheses (if any) that you will be testing.

Here is a suggested section order for your report. However, this probably won't be the order of the work. The order in which you do the statistical work is in Section 3, your plan of attack. So the order

of doing the actual work is that you carry out the plan of attack, and then write up your results following the section by section sequence I am giving you now.

Section 1. State the question to be answered clearly and succinctly, as result of the rewriting you did above. Doing this makes sure that you clearly understand the problem and what is being asked of you. Writing out the question and stating it clearly right at the beginning of your report also ensures that the decision-maker (the person who posed the question in the first place) and you understand each other. You can also write in the hypothesis which you are testing.

Section 2. Provide an executive summary of the results. This should be perhaps six lines long and contain the key findings from your work. Don't put in any technical words, jargon or complicated results. Just enough so that a busy person can skim through it and get the general idea of your findings before going more deeply into your hard work.

Section 3. Outline your **plan of attack**, describing how you have approached the problem. You also can include here the hypotheses (if any) that you will be testing. A typical plan of attack follows later on in this chapter.

Section 4. Data. Provide a brief description of the data that you have used. The source of the data, how you found it, whether you think it is reliable and whether it is sufficient for the task. It is a good idea to include summary statistics at this point. This includes information such as the number of observations, and what the variables are.

Section 5. Statistical methodology. Describe the statistical methodology which you are going to use, for example ANOVA to test whether or not the means are the same.

Section 6. Carry out the tests, making sure to include a description of whether or not the hypotheses can be rejected. This is the central part of the report. Just make sure to focus on answering the question.

Section 7. Write a conclusion which describes the results of the test and how the results answer the question that was asked. You can also include here any shortcomings in the data or in your methodology which might affect the validity of the results. The conclusion is very important because it ties together all the previous sections. Here you could include a link to a Tableau 'story' as an additional or alternative way of presenting results.

Section 8. References/end notes or further information, especially on sources of data, should end the document. If you want to make your document look really good, and you have lots of references, consider using the Zotero plug-in for Firefox. It is an excellent free way to manage your bibliography.

18.1 Plan of attack. Follow these steps.

a. Identification of the statistical method that you will use and why you chose those methods
b. What data is required, and where can it be found?
c. Conduct the statistical tests
d. Discuss whether the results are reliable/answer the question. (If not, start again!)

18.2 Presenting your work

While it is probably best to concentrate on written work because the decision-maker might want to read your work in detail, and discuss it with colleagues, a Tableau presentation is an excellent way of making your point over again. See Chapter 2 for more on visualization and Tableau.

Here is a link to some excellent notes by Professor Andrew Gelman on giving research presentations[1]

---

[1] http://andrewgelman.com/2014/12/01/quick-tips-giving-research-presentations/

19. Glossary

This is a hands-on guide to doing some basic analytic work with data. However, statistics has its own set of vocabulary and techniques. When you are presenting your results to other people, you may wish to follow established techniques and terminology. Also, here you'll find more of the underlying theory if you are interested in how Excel gets the results it does.

Binomial distribution

The binomial distribution is used to find the probability of the occurrence of a certain number of events (called successes, although they could be anything clearly defined) within a number of trials. The binomial distribution will give, for example, the probability of at least ten parts being defective out of the next hundred, provided the long-term defective rate is known. The binomial requires some conditions to work properly. These are:

a sequence of n identical trials there are two outcomes for each trial, denoted success and failure the probability of success doesn't change the trials are independent.

It is quite easy to calculate the probabilities with Excel but for illustration here is the equation which Excel uses:

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{(n-x)}$$

where x is the number of successes; p the probability of success on one trial; n the number of trials, and f(x) the probability of x successes in n trials. This notation points up the fact that the result is a function of x.

Central limit theorem (CLT)

The central limit theorem is fundamental in statistics. What the theorem states is this: let's say you have a large population and you keep taking samples from the population and measuring the mean of each sample. Then you draw a histogram of the means so that each mean is a variable in the distribution. In other words, instead of plotting each observation independently, we plot them by the means of each sample. The histogram which appears will be approximately normally distributed, or in the shape of the well-known bell curve. This is wonderful news because the normal distribution is well behaved, and so we can calculate the area under any part of it.

Correlation

You have a pair of variables: for example: website hits and actual sales. You want to know whether there is a relationship between the variables. And whether it is 'statistically significant', or perhaps what you thought was a relationship occurred just by chance.

Correlation is the study of the linear relationship between two or more random variables. Correlation analysis provides both the direction and the strength of the relationship between the variables. For example, there is a relationship between the height and weight of individuals. There is apparently also a relationship between consumption of sugar and obesity rates at the national level. Correlation analysis can test whether the relationships are spurious or real.

It's true and worth repeating: **correlation does not imply causation**. It does not necessarily follow that one variable causes the other even though they might be highly correlated. There are many examples of such spurious correlations, for example national consumption of chocolate and number of Nobel prizes won. If you do find a correlation, first think about whether you have a `lurking variable` discussed below. The two examples given above for height/weight and sugar consumption might very well include

causal effects, but this is not something that correlation analysis can establish on its own. Indeed, some philosophers (such as David Hume) claim that causation can never be decisively established.

Now we will:

- Visualize a correlation with a scatter plot
- explain why the coefficient of correlation works
- interpret the value of the coefficient of correlation
- test for the statistical significance of the coefficient of correlation

Scatterplots and correlation

A canning factory has data on number of workers employed and output of cans. The dataset is 'canning'. The factory is interested in the relationship between the number of workers and output. Load the data into Excel.

	A	B
1	Workers	Cans
2	2	40
3	2.1	45
4	2.7	52
5	3	57
6	3.1	65
7	3.6	75

First few line of the canning data

The columns of data represents numbers of workers and cans manufactured. For the notation we'll call these X and Y. When I mean the name of the variable, I'll use the upper case. For individual observations, such as x = 23, I'll use lower case. We want to see the relationship between X and Y. To do this we need to have the data laid out in columns so that each observation of X and Y are on the same line. It's always a good thing to visualise the relationship with a scatter plot and the scatter plot is below.
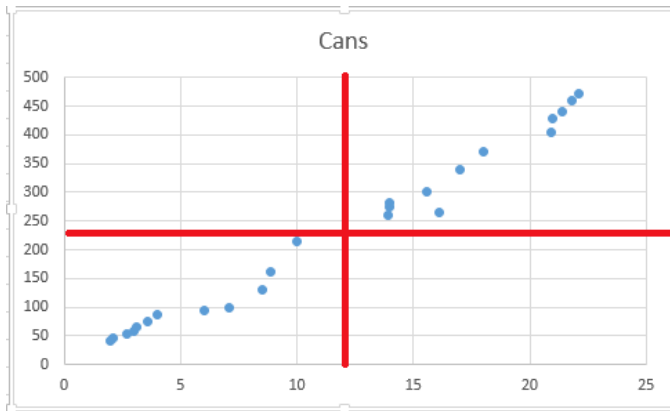
The cans scatter plot

I've drawn a vertical and horizontal line which goes through the means of the two variables. We can divide the observations into quadrants, with quadrant one being top right and so forth. If the relationship is positive then we would expect to see most of the observations in quadrant one and quadrant three. If the relationship was negative then most of the observations would be in quadrants two and four.

The relationship between X and Y is not perfectly linear. If it was all the pairs of points would line up along a straight line. We need some way of measuring how far off being in a straight line these observations are. There are two measures, covariance and coefficient of correlation. There is some maths in what follows but we'll walk through it slowly.

Covariance: Try this thought experiment: if all the Y points were on a vertical line, then the subtraction of the mean from each Y value would give us zero:

$$(y_i - \bar{y}) = 0$$

I've put in a little i for the y values to show that it could be any

of the values in the Y variable. Similarly for the X values, if all the points were on a vertical line then

$$(x_i - \bar{x}) = 0$$

We can see that we don't have vertical lines and there is some difference between each observed value and the mean. If we multiply together the differences and then find the average by dividing by n-1 we can find the covariance. We divide by n-1 because we are in most cases dealing with a sample. Subtracting 1 adjusts for this fact. The sample covariance of X and Y is therefore:

$$S_{XY} = \frac{(x - \bar{x})(y - \bar{y})}{n - 1}$$

The big problem with covariance is that the result is very dependent on the units. We can get round this problem by standardizing the differences between each observation and its mean. We do that by dividing the difference by the standard deviation of the variable. You can think of this technique as providing the difference in units of standard deviations, so

$$\frac{(x - \bar{x}}{s_x}$$

and the same for the Y variable. Now we no longer need to worry about the units. The coefficient of correlation is the covariance now between the standardised differences, not the differences themselves. The equation for the coefficient of correlation of a sample is

$$r_{XY} = \sum \frac{(x - \bar{x})(y - \bar{y})}{(n - 1)S_X S_Y}$$

Performing this calculation is extremely tedious and we usually let the computer do the work. Type in =correl(array1,array2) and hit

enter. For the cans data the result is 0.991083. This is the Pearson Product Moment or 'r' value.

Correlation Youtube[1]

Values of r: values of r are restricted to the range -1 to +1. A negative value means that the slope is negative: as one variable increases then the other decreases. A positive value means that both variables increase together. The larger the absolute value of r, then the closer the correlation. A correlation of zero means that all the points are scattered about with no pattern whatsoever. The r value of 0.991083 for cans means that the points are lined up along a straight line, which is what we would expect from looking at the scatter plot.

Testing r: the observations that we have used to calculate r consist of a sample from a population. Because it is only a sample, we don't know whether the correlation would hold up if we were somehow able to gain access to the population. r is the notation for the coefficient of correlation for a sample. For a population we use the Greek letter ρ "rho". Generally, Greek letters are used for population parameters. This distinguishes them from the Roman letters used for estimates.

The hypothesis we're testing is:

$$H_0 : \rho \leq 0$$

against the null

$$H_a : \rho \neq 0$$

We find the test statistic using this equation

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

---

[1]https://www.youtube.com/watch?v=QDj8aYfvccQ

We'll first work out the value of the test statistic using the r value we found above. Then we'll find the p value from this.

The r value found from the correlation was 0.99 and there were 24 observations (n=24). Plugging in the numbers, we find that t = 32.86. This is a one tailed test, so use = t.dist(your value for t, n-1, and false) to find a p value of 2.67343E-21. The -21 means move the decimal place 21 places to the left, so this number is effectively zero. The r value we found is statistically significant. It is much smaller than 0.05, so we can reject the null hypothesis. There is a correlation.

Descriptive statistics uses graphs and tables to present what we know about the data. This is a powerful method of gaining insights into the data, and also for making presentations. However, descriptive statistics limits itself to the data available does not make any inferences beyond what we have from the data to hand.

Distribution of the observations within a dataset describes the density of the observations: are they spread about evenly, or peaked in the middle and then spread out on either side of their mean?
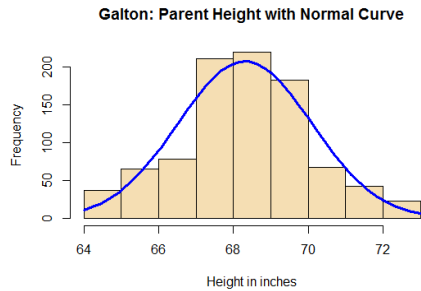
One of the most common and useful distributions is the normal distribution, sometimes known as the bell curve. Many things in the natural world follow this distribution. For example, if you were able to measure the heights of a large number of men or women, you'd find that they followed a normal distribution.

The plot to the right shows the distribution of records of heights of parents collected by Francis Galton in the 19th century. Galton was trying to find out the relationship between the height of parents and their children. Along the way, he discovered the phenomenon known as 'regression to the mean'.

I have plotted the heights as a histogram and added a normal curve with the same mean and standard deviation as the original data. This clearly follows the bell curve.

The normal distribution is very important in statistics for several reasons

-the normal distribution is 'well-behaved' in the sense that it always follows the same mathematical function. The equation is a little complex, but we can draw any normal distribution provided we know its mean and variance. The **mean** is the average of the ob-

Galton: Parent Height with Normal Curve

Galton's height observations

servations, and in a normal distribution occurs right in the middle. The **variance** is a measure of the average distance of each observation from the mean. The larger the variance, the more 'spread out' or dispersed the data.

The graphs here show two datasets, with identical means but different variances. In most of this book we won't use the term variance, but instead **standard deviation**. The standard deviation is just the square root of the variance. We use the standard deviation primarily so that we can use the same units as the mean. The first plot has a standard deviation of 5, and the second one a standard deviation of 2. The plots were created by generating a random set of 1000 numbers, with a mean of ten and the standard deviations just described.

-the second reason is connected with the most important result in statistics, the **Central Limit Theorem**, or CLT. This states that if you keep drawing samples with replacement from a large population, the **means of the samples will follow a normal distribution**. The important thing is that the distribution of the original population doesn't matter as long as we draw at least 30 samples. In fact, it gets better than that: if the original population is not normally distributed, then we don't need as many as 30 in our sample. As few as 10 will do the trick.
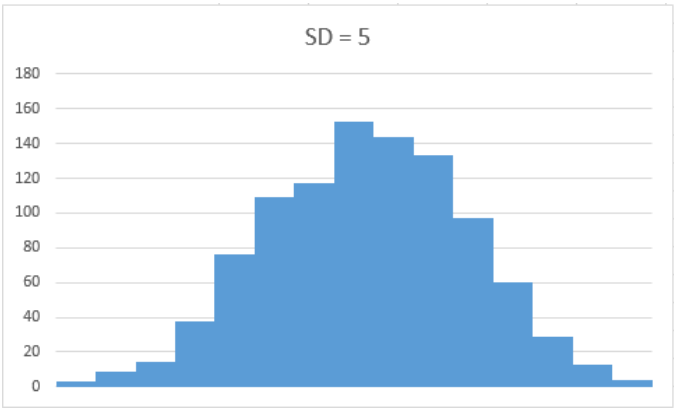
Standard deviation = 5

Because the normal distribution follows an equation so well, we can find the area under any part of it very easily.
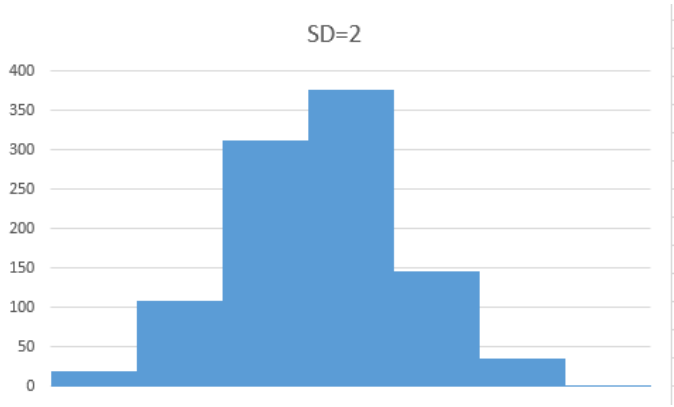
Standard deviation = 2

The mean splits the area under the curve into halves, meaning that 50 % of the observations are smaller than the mean, and 50% larger. A common practice is to mark off the horizontal axis in units of standard deviation. The further from the mean, the greater the standard deviation. This helps both in visualizing the shape of the data, but also in using the **empirical rule**.

Empirical rule: a 'rule of thumb' which states that when data are normally distributed,

-64% of the data are within one standard deviation of the mean -95% of the data are within two standard deviations of the mean -99.7% of the data are within three standard deviations of the mean

This makes intuitive sense: only 100 - 99.7 = 0.3 % of the data are more than three standard deviations from the mean. They would therefore be in either of the extreme ends or tails of the distribution. The probability of their occurrence would be very small. By contrast, observations closer to the mean, those located only a small number of standard deviations from the mean, are much more likely to occur. That is why the histogram is highest on at the mean. Thought experiment: very short or very tall people are unusual (which is why you stare at them). By contrast, people around the mean height are not at all unusual and you just pass them by.

Estimation: The statistical process is circular: we start off by choosing the characteristic of the population that interests us. That characteristic is called a parameter. We draw a random sample from that population. We use the statistic to infer information about the same quantifiable feature in the population. The statistic (for example an average or a proportion) is an **estimate** of the population parameter. The process of finding how close the estimate to the parameter, and building a confidence interval around the estimate, is **inference**. We test claims about the parameter using hypothesis-testing, which is closely related to inference. Hypothesis-testing is the final part of this chapter.

Excel –more screencasts

Inserting the data analysis toolpak add-in[2]

The normal distribution[3]

---

[2]https://www.youtube.com/watch?v=Ods1Z5BLD9g&list=UUVZFhqXDgJgIRq-ljotClgQ

[3]https://i1.ytimg.com/vi/1qn0EIm5-PQ/mqdefault.jpg

Exponential Smoothing

Hypotheses and p values A hypothesis is a claim someone makes about a population. For example, a coffee importer (me!) is told by the supplier that the mean weight of the sacks of coffee beans is 40 kg. That is the claim. I want to test that claim because I am paying for the coffee. All the sacks of coffee in the shipment are the 'population'. I take a random sample of sacks to check. I find that the mean weight of the sample is 38 kgs. Do I reject the whole shipment, or say I must have got a low sample, let's take them? The hypotheses here are: the **null** hypothesis: the population mean weight is 40 kg. Meanwhile the **alternative** o the population mean weight is not 40 kgs. Hypothesis testing provides an answer to the test in the form of a p value, defined below.

A p value is the probability of finding a sample with the measured characteristics IF the null hypothesis was true. In the case of the coffee above, let's say that we calculated the p value (See Chapter 9) and it was 0.1. That means that there is a 10% chance of finding a sample with a mean weight of 38 kgs if the real true mean weight of the shipment was 40 kgs. The most frequently used cut-off point is 0.05 or 5%. If the p value was smaller than 0.05 then we would say that there is a less than 5% chance that this sample came from a population with the claimed weight. Therefore the shipment isn't of the claimed weight and should be rejected.

Hypothesis writing and testing

A hypothesis is a statement or claim. For example, 'the mean weight of the coffee packs is 3 kg' is a hypothesis or claim. The hypothesis needs to be tested so that we know whether the claim stands or is shown to be false.

Hypotheses are written so that the claim, and the counter-claim, are distinct. There are two hypotheses, one for the claim and the other for the counterclaim. Ho, the 'null hypothesis', contains the claim. Ha, the 'alternative hypothesis', contains the counterclaim. We write hypotheses by assuming that the null hypothesis is true.

The null hypotheses for the coffee example above would be

$$Ho : \mu = 3kg$$

This is read as: The null hypothesis is that the population mean weight (mu) of the coffee is equal to 3 kg.

The alternative hypothesis (Ha) is that the population mean weight is not 3 kg, written:

$$Ha : \mu \neq 3kg$$

We test the hypotheses using information (statistics) from the sample (the mean weight, the sample size, and the standard deviation). Note that the counterclaim says nothing about the mean weight in the population being more than or less than 3 kg, it simply says that it not equal to 3 kg. Also, important, the equality sign appears in the null. This is always the case, whether the inequality is 'strict' or 'weak'.

We can also test whether the population mean is smaller than or larger than some constant. The shipper claims that the weight is 'at least' 3 kg. We write this as:

$$H_0 : \mu \geq 3kg$$

Note that the equality remains with the null. The alternative is then the negation of the null, and this must be that the weight is 'less than 3 kg'.

$$H_a : \mu < 3kg$$

Convince yourself that the only possible way to negate Ho is with the alternative hypothesis. If the claim was that that the average weight is less than or equal to 3 kg, we simply switch over the

inequality signs. There are other forms of hypothesis writing which we will work through in this book. However, they all share the rule that the equality sign goes in the null; and the alternative negates the null.

Inference is the key statistical process. It is the way in which information about populations can be gleaned from surprisingly small samples. An example is polling before an election. Provided the sample is drawn randomly and is representative of the population, a sample of perhaps only 1,000 people can provide quite accurate estimates of the proportion of the population predicted to vote for a particular political party. The estimates are used to make **inferences** about the population. We won't know until election day whether or not the inferences were correct or not.

Inferential statistics is a very powerful technique which allows us to make the jump from a sample to a population. We can use a small sample to make inferences about the characteristics of a population about which we perhaps know very little. The key is to obtain a sample which represents the population as accurately as possible. For this, both randomization and good survey design are essential. We will discuss these two important issues later in the chapter. Recently, a third class of statistical methodologies has arisen. This is **machine learning**, which takes advantage of new analytic techniques and greater computing power. We won't be able to go into these techniques in this book, unfortunately.

Lurking variables: It happens that there is close monthly correlation between the consumption of ice cream and deaths by drowning. Is it that people eat an ice cream, go swimming and then drown? The correlation misses the lurking variable which is temperature. In hot weather, people both go swimming more often and buy ice creams. The ice cream and the drowning variables are both correlated with temperature and so are correlated with each other. Another example: it seems that there is a correlation between kids needing reading glasses and sleeping with the light on. So you

should switch the light off? Not so fast. The lurking variable is the short-sightedness of the parents, not the kids. The parents leave the light on so they can see the kid. The kids inherit bad eyesight from their parents.

Parameter: a quantifiable characteristic of interest in the population. The average weight of all of the fish in the Fraser River is a parameter. The same characteristic in the sample is known as a statistic. We can easily find this statistic because the sample is only a subset, and therefore smaller, than the population. We usually never know the actual value of a parameter, but that doesn't matter because we can estimate from the sample.

Point estimate: a statistic, such as the mean or average. It is called a point because it is an exact number such as 4.21. It is not a range, it is a point. It is called an estimate because it is an estimate of the population parameter. Therefore 4.21 is an estimate of what the same quantifiable feature would be in the population. You can think of a point estimate as being the 'best guess' for the population parameter.

If we selected a different sample from the population, then almost certainly the point estimate would be different, giving a new best guess. If we continue to take samples, then we are going to get as many best guesses as we have samples. The range of best guesses is the **sampling variation**.

Poisson distribution

The Poisson probability distribution is, as with the Binomial, a discrete probability tool. We can use it for calculating the probability of events which occur over time or space. For example, number of mistakes in a given number of lines of code. The formula is:

$$f(x) = \frac{\mu^x e^{-x}}{x!}$$

where f(x) is the probability of x occurrences in a given interval. mu

is the expected value or mean number of occurrences in the given interval (or area). e is a constant, value 2.71828.

Population: the group of individuals about whom we want some information. Each individual in the population is called an **element** or sometimes a **case**. Examples of populations are all the fish in the Fraser River, Toyota cars made in 2012, or indeed the population of a country. It is usually not possible to deal with data at the population level because it is either expensive or impossible to obtain. As a result we draw a sample, or subset, from the population.

Randomisation is the technique of selecting elements from the population to build a sample so that each element has the same known probability of being selected. This technique greatly reduces bias, described in more detail in the following chapter.

Regression

Sampling frame: a list of the items to be sampled. Imagine that you wish to survey 100 students from a University. The University provides you with a list of students and their ID numbers. The students form the population, and the list is the sampling frame. You would use a random sampling method, covered below, to randomly select 100 students from the University. The 100 students form the sample.

Standard error: Try this thought experiment. Let's say that there are 1000 jelly babies in a bowl and for some reason you'd like to known the mean weight of a jelly baby without having to weigh all of them. You have a choice of taking a random sample of either n = 100 or n = 500. Which sample would produce the most accurate estimate: obviously the sample size of 500 because it is 'closer' to the population size.

Yet it still won't be completely accurate because it could happen that the sample you select is biased in some way: perhaps you selected all the heavy ones? We can measure the amount of error with the standard error, which helps us to determine how 'close' we are to the

unknown parameter. Below I show how to calculate the standard error (SE). Here is the standard error for a statistic such as a mean:

$$SE = \frac{\sigma}{\sqrt{n}}$$

In words, the standard error is the population standard deviation divided by the square root of the sample size, denoted by n. Note that as the sample size increases, then the standard error decreases.

This is in line with your earlier intuition that the larger the sample size then the more accurate the estimate. Usually we don't know 'sigma', σ, so we replace it with 's' which is the standard deviation of the sample.

Transformations: this is a technique sometimes used in regression. The object is usually to transform the distribution of the dependent variables so that it more closely resembles a normal distribution. We do this because the theoretical underpinning of linear regression is based on an assumption of normality in the dependent variable. The dependent variable is typically transformed by taking its natural logarithm and then using the transformed variable in the regression. This is often used when the dependent variable has only positive values and is highly skewed to the right. Independent variables can also be transformed such as by adding a squared version of the variable in the regression. Independent variables can also be transformed by

z scores: A z score is a measure of how far an observation contained within a variable is from the mean of that variable, in terms of standard deviations. It is quite easy to calculate within Excel, and this Youtube shows you how

z scores Youtube[4].

Why do this? By dividing the difference between the value of the

---

[4]https://www.youtube.com/watch?v=FSZpynSBev8

observation and the mean of the variable by the standard deviation of the variable, like this:

$$z_i = \frac{x_i - \bar{x}}{s}$$

the differences are in units of standard deviations. We can use this for:

- Comparing the distributions of two completely different variables
- Detecting outliers. An outlier is an observation which is very far from the rest of the distribution. Usually, 99.7% of the observations will be within three standard deviations of the mean. So an observation which is more than three standard deviations from the mean is likely to be questionable. This can be good or bad:
- Bad because perhaps someone made a data entry error which we can catch.
- Good because perhaps there is something anomalous and interesting about that 'weird' observation.