MTH5131: Actuarial Statistics

Dr Dr Dudley Stark

Term 2, 2023-2024

Contents

1 Introduction

1.1 Topics to be covered

This module covers material in the Institute and Faculty of Actuaries CS1, Actuarial Statistics, Core Principles Syllabus. It consists of four main topics.

1. **Data analysis** describes aims of data analysis, types of data, and tools for exploratory data analysis. It also covers statistics used to measure correlation as well as the principle components method of reducing the dimensionality of the data.

2. **Point estimation and bootstrapping confidence intervals** is about the properties of point estimators and methods for finding them. As with regards to confidence intervals we will mainly cover the bootstrap method.

3. **Bayesian statistics** is a way of viewing the parameters of a distribution as being random themselves. Bayesian estimators can often be expressed in the form of a Bayesian credibility formula. This leads to the Bayesian empirical approach to credibility theory.

4. **Generalised linear models** is a broad way of generalising linear models. It encompasses logistic and Poisson regression as well as linear regression.

1.2 Software

You should already be familiar with R. If you need refreshing on R, then please see the notes for Probability and Statistics I.

We will use RStudio. RStudio can be downloaded for free from

`https://rstudio.com/products/rstudio/download/#download`

For the most part, we will be coding in a script. A script is just a txt file with the extension .R . To open up a new script in RStudio go to File, New File, R Script. You can then enter a series of commands in the top left panel, line by line. You can put comments in your script. For example

```
# This script is a quick demo
2 + 3 # here we are adding together 2 and 5
```

Thus, anything after a # is not executed. Comments help another user understand your commands, which might otherwise be opaque. If you higlight some lines of the script and press the Run button at the top right of the top left panel with th green arrow, then the highlighted commands will be executed. Otherwise, if no commands are highlighted, then pressing the Run button will execute the script line by line. When you save your script, it will be saved in your working directory. To set your working directory, use the command

```
setwd("<folder address>")
```

or choose Session, Set Working Directory, Choose Directory from the pull down menus at the top. Scripts can be loaded using the menu: File, Open File.

2 Data Analysis

2.1 Aims of a data analysis

Three keys forms of data analysis will be covered in this subsection:

- descriptive
- inferential
- predictive

2.1.1 Descriptive analysis

Data presented in its raw state can be difficult to manage and draw meaningful conclusions from, particularly where there is a large volume of data to work with. A descriptive analysis solves this problem by presenting the data in a simpler format, more easily understood and interpreted by the user.

Simply put, this might involve summarising the data or presenting it in a format which highlights any patterns or trends. A descriptive analysis is not intended to enable the user to draw any specific conclusions. Rather, it describes the data actually presented.

For example, it is likely to be easier to understand the trend and variation in the sterling/euro exchange rate over the past year by looking at a graph of the daily exchange rate rather than a list of values. The graph is likely to make the information easier to absorb.

Two key measures, or parameters, used in a descriptive analysis are the **measure of central tendency** and the **dispersion**. The most common measurements of central tendency are the mean, the median and the mode. Typical measurements of the dispersion are the sample standard deviation and ranges such as the interquartile range. You have seen these in a previous module.

Measures of central tendency tell us about the 'average' value of a data set, whereas measures of dispersion tell us about the 'spread' of the values.

2.1.2 Inferential analysis

Often it is not feasible or practical to collect data in respect of the whole population, particularly when that population is very large. For example, when conducting an opinion poll in a large country, it may not be cost effective to survey every citizen. A practical solution to this problem might be to gather data in respect of a sample, which is used to represent the wider population. The analysis of the data from this sample is called inferential analysis

The sample analysis involves estimating the parameters as described in Section 1.1 above and testing hypotheses. It is generally accepted that if the sample is large and taken at random (selected without prejudice), then it quite accurately represents the statistics of the population, such as distribution, probability, mean, standard deviation, However, this is also contingent upon the user making reasonably correct hypotheses about the population in order to perform the inferential analysis.

Care may need to be taken to ensure that the sample selected is likely to be representative of the whole population. For example, an opinion poll on a national issue conducted in urban locations on weekday afternoons between 2pm and 4pm may not accurately reflect the views of the whole population. This is because those living in rural areas and those who regularly work during that period are unlikely to have been surveyed, and these people might tend to have a different viewpoint to those who have been surveyed.

2.1.3 Predictive analysis

Predictive analysis extends the principles behind inferential analysis in order for the user to analyse past data and make predictions about future events.

It achieves this by using an existing set of data with known attributes (also known as features), known as the training set in order to discover potentially predictive relationships. Those relationships are tested using a different set of data, known as the test set, to assess the strength of those relationships. Training sets are especially used in machine learning.

A typical example of a predictive analysis is regression analysis, which is covered in more detail later. The simplest form of this is linear regression where the relationship between a scalar dependent variable and an explanatory or independent variable is assumed to be linear and the training set is used to determine the slope and intercept of the line. A practical example might be the relationship between a car's braking distance against speed.

In this example, the car's speed is the explanatory (or independent) variable and the braking distance is the dependent variable.

Example 2.1. Based on data gathered at a particular weather station on the monthly rainfall in mm (r) and the average number of hours of sunshine per day (s), a researcher has determined the following explanatory relationship:

$$s = 9 - 0.1r.$$

Using this model

1. Estimate the average number of hours of sunshine per day, if the monthly rainfall is 50mm.

2. State the impact on the average number of hours of sunshine per day of each extra millimetre of rainfall in a month.

Answers:

1. When $r = 50$

$$s = 9 - 0.1 \times 50 = 4.$$

There are 4 hours of sunshine per day on average.

2. For each extra millimetre of rainfall in a month, the average number of hours of sunshine per day falls by 0.1 hours, or 6 minutes.

2.2 The data analysis process

While the process to analyse data does not follow a set pattern of steps, it is helpful to consider the key stages which might be used by actuaries when collecting and analysing data.

The key steps in a data analysis process can be described as follows:

1. Develop a well-defined set of objectives which need to be met by the results of the data analysis.

   The objective may be to summarise the claims from a sickness insurance product by age, gender and cause of claim, or to predict the outcome of the next national parliamentary election.

2. Identify the data items required for the analysis.

3. Collection of the data from appropriate sources.

   The relevant data may be available internally (e.g. from an insurance company's administration department) or may need to be gathered from external sources (e.g. from a local council office or government statistical service).

4. Processing and formatting data for analysis, e.g. inputting into a spreadsheet, database or other model.

5. Cleaning data, e.g. addressing unusual, missing or inconsistent values.

6. Exploratory data analysis, which may include:

   (a) Descriptive analysis; producing summary statistics on central tendency and spread of the data.

   (b) Inferential analysis; estimating summary parameters of the wider population of data, testing hypotheses.

(c) Predictive analysis; analysing data to make predictions about future events or other data sets.

7. Modelling the data.

8. Communicating the results.

   It will be important when communicating the results to make it clear what data was used, what analyses were performed, what assumptions were made, the conclusion of the analysis, and any limitations of the analysis.

9. Monitoring the process; updating the data and repeating the process if required.

   A data analysis is not necessarily just a one-off exercise. An insurance company analysing the claims from its sickness policies may wish to do this every few years to allow for the new data gathered and to look for trends. An opinion poll company attempting to predict an election result is likely to repeat the poll a number of times in the weeks before the election to monitor any changes in views during the campaign period.

Throughout the process, the modelling team needs to ensure that any relevant professional guidance has been complied with. For example, the Financial Reporting Council has issued a Technical Actuarial Standard (TAS) on the principles for Technical Actuarial Work (TAS100) which includes principles for the use of data in technical actuarial work. Knowledge of the detail of this TAS is not required for this module.

Further, the modelling team should also remain aware of any legal requirement to be complied with. Such legal requirement may include aspects around consumer/customer data protection and gender discrimination.

2.3 Data Sources

Primary data can be gathered as the outcome of a designed experiment or from an observational study (which could include a survey of responses to specific questions). In all cases, knowledge of the details of the collection process is important for a complete understanding of the data, including possible sources of bias or inaccuracy. Issues that the analyst should be aware of include:

- whether the process was manual or automated;
- limitations on the precision of the data recorded;
- whether there was any validation at source;
- and if data wasn't collected automatically, how was it converted to an electronic form.

These factors can affect the accuracy and reliability of the data collected. For example:

- in a survey, an individual's salary may be specified as falling into given bands, e.g. $£20,000-£29,999$, $£30,000-£39,999$ etc, rather than the precise value being recorded.
- if responses were collected on handwritten forms, and then manually input into a database, there is greater scope for errors to appear.

Where randomisation has been used to reduce the effect of bias or confounding variables it is important to know the sampling scheme used:

- simple random sampling;
- stratified sampling;
- or another sampling method.

We will illustrate sampling methods through an example.

Example 2.2. A researcher wishes to survey 10% of a company's workforce. The sample could be selected using:

1. simple random sampling

   Using simple random sampling, each subset of employees of employees whose size is 10% of the total number of employees has size employee has an equal chance of being selected. This could be achieved by taking a list of the employees, allocating each a number, and then selecting 10% of the numbers at random (either manually, or using a computer-generated process).

2. stratified sampling

   Using stratified sampling, the workforce would first be split into groups (or strata) defined by specific criteria, e.g. level of seniority. Then 10% of each group would be selected using simple random sampling. In this way, the resulting sample would reflect the structure of the company by seniority.

   This aims to overcome one of the issues with simple random sampling, i.e. that the sample obtained does not fully reflect the characteristics of the population. With a simple random sample, it would be possible for all those selected to be at the same level of seniority, and so be unrepresentative of the workforce as a whole.

Data may have undergone some form of **pre-processing**. A common example is **grouping** (e.g. by geographical area or age band). In the past, this was often done to reduce the amount of storage required and to make the number of calculations manageable. The scale of computing power available now means that this is less often an issue, but data may still be grouped: perhaps to anonymise it, or to remove the possibility of extracting sensitive (or perhaps commercially sensitive) details.

Other aspects of the data which are determined by the collection process, and which affect the way it is analysed include the following:

- Cross-sectional data involves recording values of the variables of interest for each case in the sample at a single moment in time.

  For example, recording the amount spent in a supermarket by each member of a loyalty card scheme this week.

- Longitudinal data involves recording values at intervals over time.

  For example, recording the amount spent in a supermarket by a particular member of a loyalty card scheme each week for a year.

- Censored data occurs when the value of a variable is only partially known, for example, if a subject in a survival study withdraws, or survives beyond the end of the study: here a lower bound for the survival period is known but the exact value isn't.

- Truncated data occurs when measurements on some variables are not recorded so are completely unknown.

  For example, if we were collecting data on the periods of time for which a user's internet connection was disrupted, but only recorded the duration of periods of disruption that lasted 5 minutes or longer, we would have a truncated data set

2.3.1 Big data

The term **big data** is not well defined but has come to be used to describe data with characteristics that make it impossible to apply traditional methods of analysis (for example, those which rely on a single, well-structured data set which can be manipulated and analysed on a single computer). Typically, this means automatically collected data with characteristics that have to be inferred from the data itself rather than known in advance from the design of an experiment. Given the description above, the properties that can lead data to be classified as 'big' include:

- **size**, not only does big data include a very large number of individual cases, but each might include very many variables, a high proportion of which might have empty (or null) values – leading to sparse data;

- **speed**, the data to be analysed might be arriving in real time at a very fast rate – for example, from an array of sensors taking measurements thousands of time every second;

- **variety**, big data is often composed of elements from many different sources which could have very different structures – or is often largely unstructured;

- **reliability**, given the above three characteristics we can see that the reliability of individual data elements might be difficult to ascertain and could vary over time (for example, an internet connected sensor could go offline for a period).

Examples of 'big data' are:

- the information held by large online retailers on items viewed, purchased and recommended by each of its customers

- measurements of atmospheric pressure from sensors monitored by a national meteorological organisation
- the data held by an insurance company received from the personal activity trackers (that monitor daily exercise, food intake and sleep, for example) of its policyholders.

2.3.2 Data security, privacy and regulation

In the design of any investigation, consideration of issues related to data security, privacy and complying with relevant regulations should be paramount. It is especially important to be aware that combining different data from different 'anonymised' sources can mean that individual cases become identifiable.

Another point to be aware of is that just because data has been made available on the internet, doesn't mean that that others are free to use it as they wish. This is a very complex area and laws vary between jurisdictions.

2.4 Reproducible research

2.4.1 The meaning of reproducible research

Reproducibility refers to the idea that when the results of a statistical analysis are reported, sufficient information is provided so that an independent third party can repeat the analysis and arrive at the same results.

In science, reproducibility is linked to the concept of **replication** which refers to someone repeating an experiment and obtaining the same (or at least consistent) results. Replication can be hard, or expensive or impossible, for example if:

- the study is big;
- the study relies on data collected at great expense or over many years; or
- the study is of a unique occurrence (eg the standards of healthcare in the aftermath of a particular event).

Due to the possible difficulties of replication, reproducibility of the statistical analysis is often a reasonably alternative standard.

So, rather than the results of the analysis being validated by an independent third party completely replicating the study from scratch (including gathering a new data set), the validation is achieved by an independent third party reproducing the same results based on the same data set.

2.4.2 Elements required for reproducibility

Typically, reproducibility requires the original data and the computer code to be made available (or fully specified) so that other people can repeat the analysis and verify the results. In all but the most trivial cases, it will be necessary to include full documentation (eg description of each data variable, an audit trail describing the decisions made when cleaning and processing the data, and full documented code).

Doing things 'by hand' is very likely to create problems in reproducing the work. Examples of doing things by hand are:

- editing tables and figures (rather than ensuring that the programming environment creates them exactly as needed);
- downloading data manually from a website (rather than doing it programmatically); and
- pointing and clicking (unless the software used creates an audit trail of what has been clicked).

'Pointing and clicking' relates to choosing a particular operation from an on-screen menu, for example. This action would not ordinarily be recorded electronically.

The main thing to note here is that the more of the analysis that is performed in an automated way, the easier it will be to reproduce by another individual. Manual interventions may be forgotten altogether, and even if they are remembered, can be difficult to document clearly.

2.4.3 The value of reproducibility

Many actuarial analyses are undertaken for commercial, not scientific, reasons and are not published, but reproducibility is still valuable:

- reproducibility is necessary for a complete technical work review (which in many cases will be a professional requirement) to ensure the analysis has been correctly carried out and the conclusions are justified by the data and analysis;
- reproducibility may be required by external regulators and auditors;
- reproducible research is more easily extended to investigate the effect of changes to the analysis, or to incorporate new data;
- it is often desirable to compare the results of an investigation with a similar one carried out in the past; if the earlier investigation was reported reproducibly, an analysis of the differences between the two can be carried out with confidence;
- the discipline of reproducible research, with its emphasis on good documentation of processes and data storage, can lead to fewer errors that need correcting in the original work and, hence, greater efficiency.

There are some issues that reproducibility does not address:

- Reproducibility does not mean that the analysis is correct. For example, if an incorrect distribution is assumed, the results may be wrong – even though they can be reproduced by making the same incorrect assumption about the distribution. However, by making clear how the results are achieved, it does allow transparency so that incorrect analysis can be appropriately challenged.
- If activities involved in reproducibility happen only at the end of an analysis, this may be too late for resulting challenges to be dealt with. For example, resources may have been moved on to other projects.

2.5 Exploratory Data Analysis

Exploratory data analysis (EDA) is the process of analysing data to gain further insight into the nature of the data, its patterns and relationships between the variables, before any formal statistical techniques are applied. That is, we approach the data free of any pre-conceived

assumptions or hypotheses. We first see the patterns in the data before we impose any views on it and fit models.

For numerical data, this process will include the calculation of summary statistics and the use of data visualisations. Transformation of the original data may be necessary as part of this process. For a single variable, EDA will involve calculating summary statistics (such as mean, median, quartiles, standard deviation, IQR) and drawing suitable diagrams (such as histograms. You have learned about most of these methods already.

We will be using a data structure called **data frames**. A data frame is a two-dimensional object (like a matrix). However, whilst each column (ie vector) contains data of the same type, the different columns (ie vectors) can be a different data type. This will be most useful for statistical analysis where each row represents a single observation (eg a single policyholder). Suppose that you have the following information:

Name	Age	Smoker
Alfie	34	TRUE
Belinda	28	FALSE
Charlie	31	FALSE
Delilah	38	TRUE

To create the a data frame, in this case called C, type

```
C <- data.frame(name = c("Alfie", "Belinda", "Charlie", "Delilah"),
                age = c(34, 28, 31, 38),
                smoker = c(TRUE, FALSE, FALSE, TRUE))
```

To view the data frame type its name C. You will see

```
     name age smoker
1   Alfie  34   TRUE
2 Belinda  28  FALSE
3 Charlie  31  FALSE
4 Delilah  38   TRUE
```

Note that the numbers 1, 2, 3, 4 down the lefthand side of the data frame are not index values referring to the first, second, third and fourth rows but the names of the rows and are, therefore, characters '1', '2', '3', '4'. You can look at the structure of C by typing

```
str(C)
```

You will see

```
'data.frame': 4 obs. of 3 variables:
$ name : Factor w/ 4 levels "Alfie","Belinda",..: 1 2 3 4
$ age : num 34 28 31 38
$ smoker: logi TRUE FALSE FALSE TRUE
```

The output here tells us we have a data frame with three columns and describes how each of these columns are treated. Here, the first column, name, is treated as a factor. **Factors** are vectors of characters where the entries are categorical data (eg gender, insurance group, country). Each entry can only take one of a specified number of categories (eg male/female, or groups 1-15, or UK, US, etc). We call these categories the levels of the factor. By default, R will assign the levels alphabetically (so female $= 1$ and male $= 2$).

To access an particular vector, for example age.we would type

```
C$age
```

For example, using plot() on a vector '§Sepal.length' from the built-in data frame called 'iris'

```
plot(iris$Sepal.Length)
```

gives Figure 1.

Using it on 'iris' itself:

```
plot(iris)
```

gives Figure 2. This is a matrix of scatterplots for the different vectors of 'iris'. it gives a quick overview of the structure of the data frame, such as correlations between pairs of vectors.

A file can be loaded by typing in

```
read.table("<filename>.txt")
```

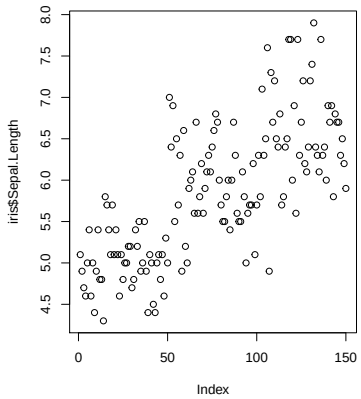


Figure 1: iris$Sepal.Length

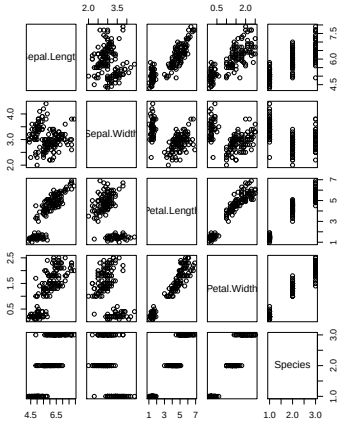


Figure 2: iris

So to load this particular file and store it in the object Data we'll need to type the following:

```
Data <- read.table("Data.txt")
```

If the file has a header, you should use

```
read.table("<filename>.txt",header=TRUE)
```

Recall that you'll need to have set the working directory using the setwd() command or via the menu to ensure R is looking in the correct folder on your computer. You can import csv (column separated value) files with

```
read.csv("<filename>.csv")
```

You can also import datasets in RStudio using the Environment window. Click on Import Dataset and select the first option.

2.6 Measures of correlation

In linear regression we think of there being a **dependent variable** Y and the **independent** or **regressor variable** X. We are often interested to know if there is a linear relationship between X and Y, i.e. one of the form

$$\mathbb{E}(Y|X = x) = a + bx$$

or even

$$Y = a + bX. \tag{1}$$

Recall that the covariance between X and Y is

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

and the correlation is

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

and that if (1) holds, then $\mathrm{Cor}(X, Y)$ is 1 if $b > 0$, 0 if $b = 0$,and -1 if $b < 0$. Data won't follow (1) perfectly and we need methods of estimating the amount of correlation between X and Y. We need statistics r which measure the strength of the dependency of Y upon X. If r is near 1 or -1 the dependency is strong, but if r is close to 0 the dependency is weak or non-existent.

We will use two sample sets. Here is the First Data Set. A sample of ten claims and corresponding payments on settlement for household policies is taken from the business of an insurance company. The amounts, in units of $\pounds 100$, are as follows:

Claim x	2.10	2.40	2.50	3.20	3.60	3.80	4.10	4.20	4.50	5.00
Payment y	2.18	2.06	2.54	2.61	3.67	3.25	4.02	3.71	4.38	4.45

First Data Set

Exercise 2.3. Use the R **plot** command to draw a scatterplot and comment on the relationship between claims and payments. You can make a txt file linear.txt which looks like

```
Claim   Payment
2.10    2.18
2.40    2.06
2.50    2.54
3.20    2.61
3.60    3.67
3.80    3.25
4.10    4.02
4.20    3.71
4.50    4.38
5.00    4.45
```

and use the command

```
DF1<- read.table("linear.txt",header=TRUE)
```

to create a data frame DF1 and plot it using

```
plot(DF1)
```

or type in vectors Claim and Payment and use

```
plot(Claim,Payment)
```

The rate of interest of borrowing, over the next five years, for ten companies is compared to each company's leverage ratio (its debt to equity ratio). The data is as follows:

Leverage ratio x	0.1	0.4	0.5	0.8	1.0	1.8	2.0	2.5	2.8	3.0
Interest rate $(\%), y$	2.8	3.4	3.5	3.6	4.6	6.3	10.2	19.7	31.3	42.9

Second Data Set

Exercise 2.4. Use the R **plot** command to draw a scatterplot and comment on the relationship between company borrowing (leverage) and interest rate. Hence apply a transformation to obtain a linear relationship.

We will be applying our methods to the First Data Set. The Second Data Set is for you to practice.

2.6.1 Pearson's correlation coefficient

Let

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Pearson's correlation coefficient is defined to be

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 /n$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 /n$$

and

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n.$$

Example 2.5. For Data Set 1,

$$\sum_{i=1}^{n} x_i = 35.4,\ \sum_{i=1}^{n} x_i^2 = 133.76,\ \sum_{i=1}^{n} y_i = 32.87,\ \sum_{i=1}^{n} y_i^2 = 115.2025,\ \sum_{i=1}^{n} x_i y_i = 123.81,$$

$$S_{xx} = 8.444,\ S_{yy} = 7.15881,\ S_{xy} = 7.4502 \ \Rightarrow r = \frac{7.4502}{\sqrt{8.444 \times 7.15881}} = 0.95824.$$

As expected, this high (close to 1), and indicates a strong positive linear relationship.

The R code for calculating the Pearson Correlation Coefficient for variables x and y is

```
cor(<Data>, method = "pearson")
```

where $< \mathrm{Data} >$ could either be a data frame or tthe names of two vectors separated by a comma.

Exercise 2.6. Calculate r for Data Set 2.

2.6.2 Spearman's rank correlation coefficient

Next we define Spearman's rank correlation coefficient r_s.

First we must define the ranks of the variables. The rank of a value of x is j if x is the jth smallest of the x's. Similarly, The rank of a value y is j if y is the jth smallest of the y's. For example, for Data Set 1 see the ranks in the third and fourth columns.

Claim x	Payment y	*Rank* x	*Rank* y	d	d^2
2.1	2.18	1	2	-1	1
2.4	2.06	2	1	1	1
2.5	2.54	3	3	0	0
3.2	2.61	4	4	0	0
3.6	3.67	5	6	-1	1
3.8	3.25	6	5	1	1
4.1	4.02	7	8	-1	1
4.2	3.71	8	7	1	1
4.5	4.38	9	9	0	0
5.0	4.45	10	10	0	0

Spearman's rank correlation coefficient is the Pearson correlation coefficient applied to the ranks, $(\mathrm{rank}(x_i), \mathrm{rank}(y_i))$ rather than applied to the raw values (x_i, y_i) of the bivariate data.

If all the x_i's are unique and all the y_i's are separately unique, then

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where

$$d_i = \mathrm{rank}(x_i) - \mathrm{rank}(y_i),$$

Hence, the table above produces

$$r_s = 1 - \frac{6 \times 6}{10 \times (10^2 - 1)} = 0.9636$$

The R code for calculating the Pearson Correlation Coefficient for variables x and y is

```
cor(<Data>, method = "spearman")
```

Exercise 2.7. Calculate r_s for Data Set 2.

2.6.3 The Kendall rank correlation coefficient

Kendall's rank correlation coefficient τ measures the strength of monotonic relationship between two variables. Like the Spearman rank correlation coefficient, the Kendall rank correlation coefficient considers only the relative values of the bivariate data, and not their actual values. It is far more intensive from a calculation viewpoint, however, since it considers the relative values of all possible pairs of bivariate data, not simply the ranks of x_i and y_i.

Despite the more complicated calculation, it is considered to have better statistical properties than Spearman's rank correlation coefficient, particularly for small data sets with large numbers of tied ranks.

If there are n observations (x_i, y_i), $i = 1, 2, \ldots, n$, then there are $\binom{n}{2}$ pairs of observations (x_i, y_i), (x_j, y_j). A pair is said to be **concordant** if $x_i > x_j$ and $y_i > y_j$ or if $x_i < x_j$ and $y_i < y_j$. It is said to be **disconcordant** if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$.

Let n_c be the number ofn concordant pairs and let n_d be the number of disconcordant pairs. The Kendall coefficient τ is defined to be

$$\tau = \frac{n_c - n_d}{n(n-1)/2}.$$

For Data Set 1, marking c for concordant and d for discordant, we get

	$2.1, 2.18$	$2.4, 2.06$	$2.5, 2.54$	$3.2, 2.61$	$3.6, 3.67$	$3.8, 3.25$	$4.1, 4.02$	$4.2, 3.71$	$4.5, 4.38$	$5.0, 4.45$
$2.1, 2.18$		d	c	c	c	c	c	c	c	c
$2.4, 2.06$			c	c	c	c	c	c	c	c
$2.5, 2.54$				c	c	c	c	c	c	c
$3.2, 2.61$					c	c	c	c	c	c
$3.6, 3.67$						d	c	c	c	c
$3.8, 3.25$							c	c	c	c
$4.1, 4.02$								d	c	c
$4.2, 3.71$									c	c
$4.5, 4.38$										c
$5.0, 4.45$										

We see that $n_c = 42$ and $n_d = 3$, so

$$\tau = \frac{42 - 3}{(10)(9)/2} = 0.8667.$$

Again the relatively high value demonstrates the strong correlation between the variables.

It's often easier to determine concordant and discordant pairs by using the ranks instead of the actual numbers. First arrange the values in order of rank for x. Then the number of concordant pairs (C) is the number of observations below which have a higher rank for the y and the number of discordant pairs (D) is the number of observations below which have a

lower rank for the y. For Data Set 1,

	Rank x	Rank y	C	D
$2.1, 2.18$	1	2	8	1
$2.4, 2.06$	2	1	8	0
$2.5, 2.54$	3	3	7	0
$3.2, 2.61$	4	4	6	0
$3.6, 3.67$	5	6	4	1
$3.8, 3.25$	6	5	4	0
$4.1, 4.02$	7	8	2	1
$4.2, 3.71$	8	7	2	0
$4.5, 4.38$	9	9	1	0
$5.0, 4.45$	10	10		

Totalling the columns gives $n_c = 42$ and $n_d = 3$ as before.

The R code for calculating the Pearson Correlation Coefficient for variables x and y is

```
cor(<Data>, method = "kendall")
```

Exercise 2.8. Calculate τ for Data Set 2.

2.6.4 Inference for correlation coefficients

To go further than a mere description/summary of the data, a model is required for the distribution of the underlying variables (X, Y).

Inference under Pearson's correlation

The bivariate normal distribution with parameters μ_X, μ_Y, σ_X, σ_Y and ρ has joint p.d.f.

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right).$$

If $\rho = 0$, the X and Y are independent, while if $\rho = \pm 1$, then X and Y are directly related.

We have

$$\mathbb{E}(Y|X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X).$$

Note that the right hand side is a linear function of x.

Theorem 2.9. *Under $H_0 : \rho = 0$, the statistic $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ is t-distributed with $n - 2$ degrees of freedom.*

Example 2.10. For Data Set 1 we found that $r = 0.958$. The value of the test statistic is

$$\frac{0.95824\sqrt{8}}{\sqrt{1 - (0.95824)^2}} = 9.478$$

Under H_0 the test statistic is t_8 distributed. Let $Y \sim t_8$. The p-value $2P(Y > 9.478)$ is much less than 0.1% and is strong evidence to reject H_0 and conclude that $\rho \neq 0$.

The R code for testing $H : \rho = 0$ vs $H : \rho \neq 0$ and getting a 95% confidence interval for ρ is

```
cor.test(x, y, method = "pearson",alt="two.sided",conf.level=0.95)
```

These are the default values, so you can just type in

```
cor.test(x, y, method = "pearson")
```

You can also set alt="greater" or alt="lower" and conf.level to a different numb er, sat conf.level=0.90 .

There is another result which does not rely on $\rho = 0$.

Theorem 2.11. *If*

$$W = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

then W has approximately a normal distribution with mean

$$\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

and standard deviation

$$\frac{1}{n-3}.$$

From this result on W, tests of $H_0 : \rho = \rho_0$ can be performed. Confidence intervals for μ_W and hence for ρ can also be found.

Example 2.12. For Data Set 1, test $H_0 : \rho = 0.9$ vs $H_1 : \rho > 0.9$.

For $r = 0.95824$, $W = 1.9239$. Under H_0, W is approximately normal distributed with mean 1.4722 and standard deviation $1/\sqrt{10-3} = 0.37796$. Thus,

$$P(W > 1.921) = P\left(Z > \frac{1.9239 - 1.4722}{0.37796}\right) = P(Z > 1.195) \approx 0.12.$$

The p-value of $r = 0.958$ is therefore about 0.12. There is insufficient evidence to justify rejecting H_0.

Inference under Spearman's correlation

Since we are using ranks rather than the actual data, no assumption is needed about the distribution of X, Y or (X, Y), i.e. it is a non-parametric test. However, non-parametric tests are less powerful than parametric tests (ie ones that do assume a distribution) as we have less information. So we would need to obtain a more extreme result before we are able to reject H_0. On the plus side, the test is less affected by outliers.

Under a null hypothesis of no association/no monotonic relationship between X and Y the sampling distribution of r_s can (for small values of n) be determined precisely using permutations. This does not have the form of a common statistical distribution.

For example, if we had a sample size of 4, there would be 4!=24 ways of arranging the ranks of the Y variables, so each arrangement has a probability of $1/24$ of occurring. We then calculate $\sum d^2$ for each arrangement and hence obtain the probabilities of getting each value of $\sum d^2$. We can then carry out a hypothesis test. For example, if we are testing $H_0 : \rho = 0$ vs $H_1 : \rho > 0$ with a 5% significance level, where the data values give $\sum vd^2 = 3$, we can calculate $P(\sum d^2 \leq 3)$ using the probabilities obtained above and if we get less than 5% we would reject H_0. However, for large n this will be time consuming.

For larger values of n ($n > 20$) we can use Theorem 2.11 above. Recall that Spearman's rank correlation coefficient is derived by applying Pearson's correlation coefficient to the ranks

rather than the original data. Under the null hypothesis that the variables are uncorrelated,

$$\frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

is approximately t_{n-2} distributed.

For very large values of n, the sampling distribution of r_s can be approximated by the $N\left(0, \frac{1}{n-1}\right)$ distribution.

The R code for testing $H : \rho = 0$ vs $H : \rho \neq 0$ and getting a 95% confidence interval for ρ_s is

```
cor.test(x, y, method = "pearson", alt="two.sided",conf.level=0.95)
```

Inference under Kendall's rank correlation

Again, since we are using ranks, we have a non-parametric test.

Under the null hypothesis of independence of X and Y, the sampling distribution of τ can be determined precisely using permutations for small values of n.

We can carry out a hypothesis test in the same way as described above but calculating each arrangement. However, again, for large n this will be time consuming.

For larger values of n ($n > 10$), use of the Central Limit Theorem means that an approximate normal distribution can be used, with mean 0 and variance $2(2n + 5)/9n(n - 1)$.

The R code for testing $H : \rho = 0$ vs $H : \rho \neq 0$ and getting a 95% confidence interval for ρ_s is

```
cor.test(x, y, method = "kendall", alt="two.sided",conf.level=0.95)
```

Note that cor.test will determine exact p-values if $n < 50$; for larger samples the test statistic is approximately normally distributed.

2.7 Principle components analysis

First we will give an overview of principal component analysis (PCA) and then explain the theory and go over an example.

PCA, also called factor analysis, provides a method for reducing the dimensionality of the data set. In other words, it seeks to identify the key components necessary to model and understand the data. For many multivariate datasets there is correlation between each of the variables. This means there is some 'overlap' between the information that each of the variables provide. The technical phrase is that there is redundancy in the data. PCA gives us a process to remove this overlap.

The idea is that we create new uncorrelated variables, and we should find that only some of these new variables are needed to explain most of the variability observed in the data. The key thing is that each 'new' variable is a linear combination of the 'old' variables. If we eliminate the new variables with the smallest variation we are still retaining the most important bits of information. These components are chosen to be uncorrelated linear combinations of the variables of the data which maximise the variance.

In order to understand PCA in more depth, we will need to go over some linear algebra.

The **dot product** of two vectors $\mathbf{v} = (v_1, v_2, \ldots, v_n)$ and $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ in $\mathbb{R}^n$ is $\mathbf{v} \cdot \mathbf{w} = \sum_{i=1}^{n} v_i w_i$. Two vectors $\mathbf{v}$, $\mathbf{w}$ are **orthogonal** if $\mathbf{v} \cdot \mathbf{w} = 0$. A **basis** of vectors of $\mathbb{R}^n$ is a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ such that every vector $\mathbf{v} \in \mathbb{R}^n$ can be written uniquely as $\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$ where $\alpha_i \in \mathbb{R}$ for all i. A basis is **orthogonal** if $\mathbf{v}_i$ and $\mathbf{v}_j$ are orthogonal for every pair $i \neq j$.

A column vector $\mathbf{v}$ is an **eigenvector** of A if $A\mathbf{v} = \lambda\mathbf{v}$ for some $\lambda \in \mathbb{R}$. The **eigenvalues** λ are the solutions of the eqation $\det(A - \lambda I) = 0$, where $\det$ stands for determinant and I is the identity matrix. The eigenvectors corresponding to an eigenvalue λ are the vectors satisfying the equation $(A - \lambda I)\mathbf{v} = \mathbf{0}$.

The **transpose** of a matrix A with entries $A_{i,j} = a_{i,j}$ is the matrix A^T with entries $A^T_{i,j} = a_{j,i}$. A square matrix A such that $A^T A = I$ is **orthogonal**. The column vectors of a square $n \times n$ matrix A are the vectors $\mathbf{v}_j = (A_{1,j}, A_{2,j} \ldots, A_{n,j})^T$. A matrix is orthogonal if and only if its column vectors are mutually orthononal.

Suppose we have a $n \times p$ **data matrix** $\hat{X}$, where n is the number of observations and p is the number of variables. We center $\hat{X}$, meaning that from each entry $\hat{x}_{i,j}$ we subtract $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_{i,j}$. This gives a new matrix X which for which all columns sum to 0. The $p \times p$ matrix $X^T X$ divided by $n-1$ is the sample variance/covariance matrix for the variables;

$(X^TX)_{j_1,j_2}$ is an estimator for the covariance between the j_1th and j_2th variable. It is a fact that we are always able to obtain an orthogonal basis in $\mathbb{R}^p$ of eigenvectors of X^TX. These eigenvectors are the **components** of X. It often happens that the largest two or three eigenvalues are significantly larger than the other eigenvalues. The eigenvectors corresponding to the largest eigenvalues are called **principle components**. The observations, i.e. the rows of X, are represented by p values, but the principle components explain most of the variability in the data and we can get a good approximation to the data by taking linear combinations of the principle components. That is the idea behind PCA. The eigenvector of the largest eigenvalue often represents the trend of the data.

Let us see in detail how we can get an approximation to X by using the principle components. Let W be the $p \times p$ orthogonal matrix whose column vectors are the components of X, ordered left to right from largest eigenvalue to smallest eigenvalue. We can represent X by $X = XI = XWW^T = PW^T$ where $P = XW$. Now, P^TP is a diagonal matrix whose diagonal entry $(P^TP)_{i,i}$ is exactly the ith largest eigenvalue of X^TX for each i. (This is because W is what is called a change of basis matrix.) We now zero out the columns of P corresponding to the non-principal components, i.e. the rightmost columns, thereby forming a new matrix $\tilde{P}$. We get a good approximation to X by $\tilde{X} = \tilde{P}W^T$. Because we are using a reduced basis with fewer than p vectors to represent the row vectors of $\tilde{X}$, this is an example of "data compression".

Example 2.13. Suppose we are trying to model the chances of a student passing the MTH5131 exam. We are going to include in our model the average number of days per week each student does some studying (X_1) and the average number of hours each student studies at the weekend (X_2). The data values for one student are $x_1 = 2$, $x_2 = 10$ and for another student we have $x_1 = 4$, $x_2 = 6$. The data matrix is therefore:

$$\hat{X} = \begin{pmatrix} 2 & 10 \\ 4 & 6 \end{pmatrix}.$$

The mean of the entries in the first column is 3 and the mean of the entries in the second

column is 8, so the centred matrix is:

$$X = \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}.$$

We now need to calculate $X^T X$:

$$X^T X = \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} = \begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix}.$$

We can see that this is the covariance matrix for the data in X. The variance of the data set $(-1, 1)$ is 2, the variance of the data set $(2, -2)$ is 8, and the covariance between the data sets is calculated as follows:

$$\frac{1}{n-1}\sum_{j=1}(\hat{x}_{j,1} - \overline{x}_1)(\hat{x}_{j,2} - \overline{x}_2) = \frac{1}{2-1}\sum_{j=1}^{2} x_{j,1}x_{j,2} = (-1)(2) + (1)(-2) = -4.$$

Next we determine the eigenvalues, and from there the eigenvectors, for the matrix $X^T X$:

$$\det(X^T X - \lambda I) = \det \begin{pmatrix} 2-\lambda & -4 \\ -4 & 8-\lambda \end{pmatrix} \Rightarrow (2-\lambda)(8-\lambda) - 16 = 0 \Rightarrow \lambda^2 - 10\lambda = 0 \Rightarrow \lambda = 0, \lambda = 10.$$

For $\lambda = 0$:

$$\begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \left.\begin{array}{l} 2x - 4y = 0 \\ -4x + 8y = 0 \end{array}\right\} x = 2y$$

so one eigenvector $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$. For $\lambda = 10$:

$$\begin{pmatrix} -8 & -4 \\ -4 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \left.\begin{array}{l} -8x - 4y = 0 \\ -4x - 2y = 0 \end{array}\right\} y = -2x$$

so another eigenvector is $\begin{pmatrix} 1 \\ -2 \end{pmatrix}$. The unit eigenvalues (components) are

$$\frac{1}{2^2+1^2}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ and } \frac{1}{1^2+(-2)^2}\begin{pmatrix} 1 \\ -2 \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

By definition,

$$W = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}.$$

(The component corresponding to $\lambda = 10$ goes to the left.) The principle components decomposition matrix P is

$$P = XW = \frac{1}{\sqrt{5}}\begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} -5 & 0 \\ 5 & 0 \end{pmatrix}$$

The second column provides no information as it is all zeroes. Indeed, the rows of the matrix X are simply multiples of the row vector $\begin{pmatrix} 1 \\ -2 \end{pmatrix}^T$. If, say, the second column had been non-zero we might have well chosen to zero it out ourselves. In the general case, we would set the columns in P of the components we are eliminating to zero. Note that

$$P^TP = \begin{pmatrix} 10 & 0 \\ 0 & 0 \end{pmatrix}:$$

a diagonal matrix with the eigenvalues 10 and 0 on the diagonal. If we choose $\tilde{P} = P$, then $X = \tilde{P}W^T$. Generally, $\tilde{P}W^T$ will be different from X but will be a good approximation to it.

Notes:

1. Since the principal components are linear combinations of the variables it is not useful for reducing the dimensionality where there are non-linear relationships. A suitable transformation (such as log) should be applied first.

2. Since the loadings of each variable that make up the components are chosen by max-

imising variance, variables that have the highest variance will be given more weight. It is often good practice (especially if different units of measurement are used for each variables) to scale the data before applying PCA.

3. No explanation has been provided for what these components represent in a practical, real-world sense. Intuitively, the first component is the overall trend of the data. For the second component onwards, intuition for this must be sought elsewhere. This is often done by regressing the components against variables external to the data which the statistical analyst has an a priori cause to believe may have explanatory power.

In choosing which components to keep, we could use the criteria of choosing principle components containing a total of at least 90% of the total variance, say. We could also use a Scree test. This involves the examination of a line chart of the variances of each component (called a Scree diagram). The Scree test retains only those principal components before the variances level off (which is easy to observe from the Scree diagram).

PCA is performed in R through

```
prcomp(<name of data frame>)
```

3 Point Estimation

Poimt estimators are functions of the data which estimate a parameter or some function of the parameters. We will study their properties and also describe two good ways of finding them.

3.1 Definitions

Assume that we have data, each observation of which is considered to be a sample from a population, perhaps all the same, perhaps all different. In other words, we have a probability model for the population from which an observation is assumed to be a random sample.

Example. Suppose that we have data on a single, continuous variable given by $y_1, y_2, \ldots, y_n$. We are interested in the population from which the data are sampled. Suppose that the

probability model is that $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ independently for $i = 1, 2, \ldots, n$. Then the probability model for the population is completely specified, except for the unknown parameters μ and σ^2.

In general, we have the following probability model for the population from which the data $y_1, y_2, \ldots, y_n$ are a sample: $Y_1, Y_2, \ldots, Y_n$ have a joint distribution which is specified, except for the unknown parameters $\theta_1, \theta_2, \ldots, \theta_p$. We are interested in estimating some function of the parameters $\phi = g(\theta_1, \ldots, \theta_p)$. Note that ϕ could be just a single parameter, such as μ in the above example. We will use a **sample statistic**, $T(Y_1, \ldots, Y_n)$, as an **estimator** of ϕ. For example, $\overline{Y} = T(Y_1, \ldots, Y_n)$ is an estimator of μ. The **observed** value of T, $T(y_1, \ldots, y_n) = t$, say, is the **point estimate** of ϕ.

Definition. For discrete data $y_1, y_2, \ldots, y_n$, the **likelihood**, $L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n)$, is the **joint probability mass function** of the data, that is,

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n) = P(Y_1 = y_1, \ldots, Y_n = y_n).$$

For continuous data $y_1, y_2, \ldots, y_n$, the likelihood, $L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n)$, is the **joint probability density function** of the data, that is,

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n) = f_{Y_1, \ldots, Y_n}(y_1, \ldots, y_n).$$

Note. If the data are **independent**, then

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n) = \prod_{i=1}^{n} P(Y_i = y_i)$$

in the discrete case and

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i)$$

in the continuous case.

Example. Suppose that $Y_i \sim \text{Exp}(\lambda)$ independently for $i = 1, 2, \ldots, n$. Then the likelihood is

$$\begin{aligned} L(\lambda; y_1, \ldots, y_n) &= \prod_{i=1}^{n} \lambda e^{-\lambda y_i} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}. \end{aligned}$$

Example. Suppose that $Y_i \sim \text{N}(\mu, \sigma^2)$ independently for $i = 1, 2, \ldots, n$. Then the likelihood is

$$\begin{aligned} L(\mu, \sigma^2; y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}. \end{aligned}$$

For notational convenience, we will sometimes write the data as a vector, $\underline{y} = (y_1, \ldots, y_n)^T$, and write the parameters as a vector, $\underline{\theta} = (\theta_1, \ldots, \theta_p)^T$. So we write the likelihood as $L(\underline{\theta}; \underline{y})$.

3.2 The bias and the mean square error

An estimator is a random variable, a function of the random variables $Y_1, \ldots, Y_n$, and the estimate is a single value taken from the distribution of this random variable. Since we want our estimate, t, to be close to ϕ, the random variable T should be centred close to ϕ and have small variance. Also, if $Y_1, \ldots, Y_n$ are independent and identically distributed, we would want our estimator to be such that, as $n \to \infty$, $T \to \phi$ with probability one.

Definition. The **error** in estimating ϕ by $T(\underline{Y})$ is $T - \phi$.

Definition. The **bias** of the estimator $T(\underline{Y})$ for ϕ is

$$\text{bias}(T) = E(T) - \phi,$$

that is, the expected error.

Definition. The estimator $T(\underline{Y})$ is **unbiased** for ϕ if $E(T) = \phi$, that is, it has zero bias.

Definition. The **mean square error** of $T(\underline{Y})$ as an estimator of ϕ is

$$\mathrm{MSE}(T) = E\left\{(T-\phi)^2\right\}.$$

Note. We may write

$$\begin{aligned}
\mathrm{MSE}(T) &= E\left([\{T-E(T)\}+\{E(T)-\phi\}]^2\right) \\
&= E\left[\{T-E(T)\}^2 + 2\{T-E(T)\}\{E(T)-\phi\} + \{E(T)-\phi\}^2\right] \\
&= E\left[\{T-E(T)\}^2\right] + 2\{E(T)-E(T)\}\{E(T)-\phi\} + \{E(T)-\phi\}^2 \\
&= \mathrm{var}(T) + \{\mathrm{bias}(T)\}^2.
\end{aligned}$$

Definition. The estimator $T(\underline{Y})$ is **consistent** for ϕ if its distribution converges to ϕ as $n \to \infty$.

Notes. 1. A sufficient condition for consistency is that $\mathrm{MSE}(T) \to 0$ as $n \to \infty$, or, equivalently, $\mathrm{var}(T) \to 0$ and $\mathrm{bias}(T) \to 0$ as $n \to \infty$.

2. Consistency means that the probability of our estimator being within some small δ of ϕ can be made as close to one as we like by making the sample size n sufficiently large, and so $P(|T-\phi| < \delta) \to 1$ as $n \to \infty$.

Example. Suppose that $Y_1, \ldots, Y_n$ are independent $\mathrm{Poisson}(\lambda)$ random variables and consider using $\overline{Y}$ to estimate λ. Then we have

$$\begin{aligned}
E(\overline{Y}) &= \frac{1}{n}E(Y_1 + \ldots + Y_n) \\
&= \frac{1}{n}\left\{E(Y_1) + \ldots + E(Y_n)\right\} \\
&= \frac{1}{n}n\lambda = \lambda,
\end{aligned}$$

and so $\overline{Y}$ is an unbiased estimator of λ. Next, by independence, we have

$$\begin{aligned} \mathrm{var}(\overline{Y}) &= \frac{1}{n^2}\left\{\mathrm{var}(Y_1) + \ldots + \mathrm{var}(Y_n)\right\} \\ &= \frac{1}{n^2} n\lambda = \frac{\lambda}{n}, \end{aligned}$$

so that $\mathrm{MSE}(\overline{Y}) = \lambda/n$. Thus, $\mathrm{MSE}(\overline{Y}) \to 0$ as $n \to \infty$, and so $\overline{Y}$ is a consistent estimator of λ.

Note that there are many estimators of λ, such as $n\overline{Y}/(n+1)$ which are biased, since

$$E\left(\frac{n}{n+1}\overline{Y}\right) = \frac{n}{n+1}\lambda,$$

but asymptotically unbiased, since $n/(n+1) \to 1$ as $n \to \infty$. However, because

$$\mathrm{var}\left(\frac{n}{n+1}\overline{Y}\right) = \left(\frac{n}{n+1}\right)^2 \frac{\lambda}{n} = \frac{n\lambda}{(n+1)^2} \to 0$$

as $n \to \infty$, $n\overline{Y}/(n+1)$ is also a consistent estimator of λ.

3.3 The Cramér-Rao Lower Bound

Estimators can be compared through their mean square errors. If they are unbiased, this is equivalent to comparing their variances. In many applications, we try to find an unbiased estimator which has minimum variance, or at least low variance. We can compare the variances of several unbiased estimators, but how do we know if we have found an estimator with the lowest variance among **all** unbiased estimators?

The **Cramér-Rao lower bound** (CRLB) tells us. For a single parameter θ, the CRLB for unbiased estimators of $\phi = g(\theta)$ is

$$\mathrm{CRLB}(\phi) = \frac{\left\{\frac{dg(\theta)}{d\theta}\right\}^2}{E\left\{-\frac{d^2 \log L(\theta;\underline{Y})}{d\theta^2}\right\}},$$

where $E\{-d^2 \log L(\theta;\underline{Y})/d\theta^2\}$ is called the **Fisher information**. For a vector of p parameters

$\underline{\theta} = (\theta_1, \ldots, \theta_p)^T$, the CRLB for unbiased estimators of $\phi = g(\underline{\theta})$ is

$$\mathrm{CRLB}(\phi) = \left(\frac{\partial g}{\partial \theta_1}, \ldots, \frac{\partial g}{\partial \theta_p}\right) V^{-1} \begin{pmatrix} \frac{\partial g}{\partial \theta_1} \\ \vdots \\ \frac{\partial g}{\partial \theta_p} \end{pmatrix},$$

where the $p \times p$ matrix V has (i,j)th element $E\{-\partial^2 \log L(\underline{\theta}; \underline{Y})/\partial\theta_i \partial\theta_j\}$ and is called the **Fisher information matrix**.

For the following result, we need to assume that the sample space does not depend on the parameters, and that suitable differentiability and continuity constraints hold, so that the order of integration and differentiation can be interchanged. Problems which satisfy these conditions are called **regular**.

Theorem 1. Cramér-Rao inequality.
For any unbiased estimator, $T(\underline{Y})$, of ϕ, if the estimation problem is regular,

$$\mathrm{var}\{T(\underline{Y})\} \geq \mathrm{CRLB}(\phi).$$

So, if an unbiased estimator has variance equal to the CRLB, it must have minimum variance amongst **all** unbiased estimators. We call it the **minimum variance unbiased estimator** (MVUE) of ϕ.

Example. Suppose that $Y_1, \ldots, Y_n$ are independent $\mathrm{Poisson}(\lambda)$ random variables. Let $g(\lambda) = \lambda$. Then we have $dg/d\lambda = 1$. The likelihood is

$$\begin{aligned} L(\lambda; \underline{y}) &= \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \\ &= \frac{\lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}}{\prod_{i=1}^n y_i!}, \end{aligned}$$

and so the log-likelihood is

$$\log L = \sum_{i=1}^{n} y_i \log \lambda - n\lambda - \log\left(\prod_{i=1}^{n} y_i!\right).$$

Thus, we have

$$\frac{d \log L}{d\lambda} = \frac{\sum_{i=1}^{n} y_i}{\lambda} - n$$

and

$$\frac{d^2 \log L}{d\lambda^2} = -\frac{\sum_{i=1}^{n} y_i}{\lambda^2}.$$

Hence, the Fisher information is

$$\begin{aligned} E\left(-\frac{d^2 \log L}{d\lambda^2}\right) &= \frac{1}{\lambda^2} E\left(\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{\lambda^2} \sum_{i=1}^{n} E(Y_i) \\ &= \frac{1}{\lambda^2} n\lambda = \frac{n}{\lambda} \end{aligned}$$

and $\mathrm{CRLB}(\lambda) = \lambda/n$. Since $\mathrm{var}(\overline{Y}) = \lambda/n$, it follows that $\overline{Y}$ is the MVUE of λ.

The **efficiency** of an asymptotically unbiased estimator, $T(\underline{Y})$, of ϕ is

$$\mathrm{eff}(T) = \lim_{n\to\infty} \frac{\mathrm{CRLB}(\phi)}{\mathrm{var}\{T(\underline{Y})\}}.$$

An asymptotically unbiased estimator, $T(\underline{Y})$, of ϕ is said to be **efficient** if its efficiency is equal to one. In this case, for large n, $\mathrm{var}\{T(\underline{Y})\}$ is approximately equal to the CRLB.

3.4 Method of moments

Two commonly used methods of finding point estimators ate the method of moments and the method of maximum likelihood. These method will be covered in this and the next section.

The **kth population moment about the origin** of a random variable Y is

$$\mu_k' = E(Y^k), \quad k = 1, 2, \ldots,$$

and the **kth sample moment about the origin** of a random sample $y_1, \ldots, y_n$ is

$$m'_k = \frac{1}{n} \sum_{i=1}^{n} y_i^k, \quad k = 1, 2, \ldots.$$

If $Y_1, \ldots, Y_n$ are assumed to be independent and identically distributed from a distribution with p parameters, the method of moments estimators of $\theta_1, \ldots, \theta_p$ are obtained by solving the equations

$$\mu'_k = m'_k, \quad k = 1, 2, \ldots, p.$$

The method of moments estimator of θ_j is denoted by

$$\tilde{\theta}_j = \tilde{\theta}_j(\underline{y}), \quad j = 1, 2, \ldots, p,$$

and the method of moments estimator of $\phi = g(\underline{\theta})$ is defined to be $\tilde{\phi} = g(\tilde{\underline{\theta}})$, where $\tilde{\underline{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_p)^T$. The following result, stated without proof, summarises the main properties of these estimators.

Theorem 1. Under fairly general conditions, method of moments estimators are asymptotically normal and asymptotically unbiased, but they are not in general efficient.

If there is only one parameter θ (so $p = 1$), then the Method of Moments Estimator is obtained by setting

$$\mathbb{E}(Y) = \overline{y}$$

and solving for θ.

Example. Suppose that $Y_1, \ldots, Y_n$ are independent $\mathrm{Poisson}(\lambda)$ random variables. Then we have $\mu'_1 = E(Y) = \lambda$ and $m'_1 = \overline{y}$. So the method of moments estimator of λ is $\tilde{\lambda} = \overline{y}$.

Example. A random sample from an Exponential(λ) distributed random variable is as follows:

$$14.84,\ 0.19,\ 11.75,\ 1.18,\ 2.44,\ 0.53$$

The expectation of $Y \sim \text{Exponential}(\lambda)$ is $\mathbb{E}(Y) = 1/\lambda$. We set

$$\frac{1}{\lambda} = \overline{y} \Rightarrow \lambda = \frac{1}{\overline{y}}.$$

We find that

$$\overline{y} = \frac{14.84 + 0.19 + 11.75 + 1.18 + 2.44 + 0.53}{6} = 5.155$$

so that

$$\tilde{\lambda} = \frac{1}{5.155} = 0.1940.$$

For some distributions such as Uniform$(-\theta, \theta)$ or N$(0, \sigma^2)$, the mean does not involve the parameter, in which case a hight-order moment must be used.

Example. The Uniform$(-\theta, \theta)$ distribution has expectation $(-\theta+\theta)/2 = 0$ and setting $0 = \overline{Y}$ does not involve θ. We instead note that if $Y \sim \text{Uniform}(-\theta, \theta)$, then $\text{var}(Y) = \theta^2/3$. We equate the sample and population variances. If the random sample is

$$2.6,\ 1.9,\ 3.8,\ -4.1,\ -0.2,\ -0.7,\ 1.1,\ 6.9$$

then $\sum y_i = 11.3$, $\sum y_i^2 = 90.97$,

$$s^2 = \frac{1}{7}\left(90.97 - 8 \times \left(\frac{11.3}{8}\right)^2\right) = 10.7155$$

So,

$$\frac{\tilde{\theta}^2}{3} = 10.7155 \Rightarrow \tilde{\theta} = 5.67$$

If there are two parameters θ_1 and θ_2 (so $p = 2$)). then we solve

$$\mathbb{E}(Y) = \overline{y}, \quad \mathbb{E}(Y^2) = \frac{1}{n}\sum_{i=1}^{n} y_i^2.$$

Equivalently, we may solve

$$\mathbb{E}(Y) = \overline{y}, \quad \text{var}(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2$$

Example. Suppose that $Y_1 \ldots, Y_n$ are independent $\mathrm{N}(\mu, \sigma^2)$ random variables. Then we have $\mu_1' = E(Y) = \mu$, $\mu_2' = E(Y^2) = \sigma^2 + \mu^2$, $m_1' = \overline{y}$ and $m_2' = \sum_{i=1}^n y_i^2/n$. So the method of moments estimators of μ and σ^2 satisfy the equations

$$\tilde{\mu} = \overline{y} \quad \text{and} \quad \tilde{\sigma}^2 + \tilde{\mu}^2 = \frac{1}{n}\sum_{i=1}^n y_i^2.$$

Thus, we obtain $\tilde{\mu} = \overline{y}$ and $\tilde{\sigma}^2 = \sum_{i=1}^n y_i^2/n - \overline{y}^2 = \sum_{i=1}^n (y_i - \overline{y})^2/n$.

Example. A random sample from a Binomial(n, p) distribution yields the following samples:

$$4,\ 2,\ 7,\ 4,\ 1,\ 4,\ 5,\ 4$$

We have $\bar{y} = 31/8 = 3.875$ and $\mathbb{E}(Y) = np$ Equating the last two expressions gives a first equation

$$\tilde{n}\tilde{p} = 3.875$$

We also know that

$$\mathbb{E}(Y^2) = \mathrm{var}(Y) + (\mathbb{E}(Y))^2 = np(1-p) + (np)^2.$$

We calculate that $\frac{1}{n}\sum_{i=1}^n y_i^2 = 143/8 = 17.875$. Equating the last two expressions gives a second equation

$$\tilde{n}\tilde{p}(1-\tilde{p}) + (\tilde{n}\tilde{p})^2 = 17.875$$

Substituting the first equation in the second produces

$$3.875(1-\tilde{p}) + (3.875)^2 = 17.875 \Rightarrow \tilde{p} = 0.26219$$

from which

$$\tilde{n} = 3.875/\tilde{p} = 14.78$$

Since n is the number of trials, the true value cannot be 14.78. Therefore it is likely to be 14 or 15.

Alternatively, we find that

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = \frac{143}{8} - (3.875062)^2 = 2.859375$$

and so we get the third equation

$$\tilde{n}\tilde{p}(1-\tilde{p}) = 2.859375$$

Substituting the first equation in the second produces

$$3.875(1-\tilde{p}) = 2.859375 \Rightarrow \tilde{p} = 0.2621$$

and hence $\tilde{n} = 14.78$ as before.

3.5 Method of maximum likelihood

The maximum likelihood estimates of $\theta_1, \ldots, \theta_p$ are the values of $\theta_1, \ldots, \theta_p$ which maximise the likelihood $L(\underline{\theta}; \underline{y})$. The maximum likelihood estimator of θ_j is denoted by

$$\hat{\theta}_j = \hat{\theta}_j(\underline{Y}), \quad j = 1, 2, \ldots, p,$$

and the maximum likelihood estimator of $\phi = g(\underline{\theta})$ is defined to be $\hat{\phi} = g(\hat{\underline{\theta}})$, where $\hat{\underline{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$. The following result, again stated without proof, gives the **asymptotic distributions** of $\hat{\underline{\theta}}$ and $\hat{\phi}$.

Theorem 2. Under fairly general conditions, maximum likelihood estimators are asymptotically normal, asymptotically unbiased and efficient. Moreover, for large n, $\hat{\underline{\theta}} \sim \mathrm{N}_p(\underline{\theta}, V^{-1})$ and $\hat{\phi} \sim \mathrm{N}\{\phi, \mathrm{CRLB}(\phi)\}$, where V is the Fisher information matrix.

Note. We usually find the maximum likelihood estimates by maximising the **log-likelihood** $\ell(\underline{\theta}; \underline{y}) = \log L(\underline{\theta}; \underline{y})$ and then solving the **likelihood equations**

$$\frac{\partial \ell}{\partial \theta_j} = 0, \quad j = 1, 2, \ldots, p.$$

Example. Suppose that $Y_1, \ldots, Y_n$ are independent $\mathrm{Poisson}(\lambda)$ random variables. Then the likelihood is

$$L(\lambda; \underline{y}) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}}{\prod_{i=1}^n y_i!},$$

and so the log-likelihood is

$$\ell(\lambda; \underline{y}) = \sum_{i=1}^{n} y_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log(y_i!).$$

Thus, solving the equation

$$\frac{d\ell}{d\lambda} = \frac{\sum_{i=1}^n y_i}{\lambda} - n = 0$$

yields the estimator $\hat{\lambda} = \sum_{i=1}^n Y_i/n = \overline{Y}$, which is the same as the method of moments estimator. For large n, $\hat{\lambda} \sim \mathrm{N}(\lambda, \lambda/n)$.

Example. Suppose that $Y_1, \ldots, Y_n$ are independent $\mathrm{N}(\mu, \sigma^2)$ random variables. Then the likelihood is

$$\begin{aligned} L(\mu, \sigma^2; \underline{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}, \end{aligned}$$

and so the log-likelihood is

$$\ell(\mu, \sigma^2; \underline{y}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2.$$

Thus, we have

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)$$

and

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2.$$

Setting these equations to zero, we obtain

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i-\hat{\mu})=0 \Rightarrow \sum_{i=1}^{n} y_i = n\hat{\mu},$$

so that $\hat{\mu}=\overline{Y}$ is the maximum likelihood estimator of μ, and

$$-\frac{n}{2\hat{\sigma}^2}+\frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(y_i-\overline{y})^2=0 \Rightarrow n\hat{\sigma}^2=\sum_{i=1}^{n}(y_i-\overline{y})^2,$$

so that $\hat{\sigma}^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2/n$ is the maximum likelihood estimator of σ^2, which are the same as the method of moments estimators. For large n,

$$\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}^2\end{pmatrix} \sim \mathrm{N}_2\left\{\begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}, \begin{pmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{2\sigma^4}{n}\end{pmatrix}\right\}.$$

In some cases taking the deriviative of the log-likelihood is inappropriate, sa we see in the next example.

Example. Let Y_i be i.i.d. Uniform$[0,\infty]$ distributed:

$$f_{Y_i}(y_i)=\begin{cases}\theta^{-1} & \text{if } 0\le y_i\le\theta;\\ 0 & \text{else.}\end{cases}$$

Then,

$$L(\theta:\underline{y})=\begin{cases}\theta^{-n} & \text{if } 0\le y_i\le\theta \quad \forall 1\le i\le n;\\ 0 & \text{else.}\end{cases}$$

$$\begin{cases}\theta^{-n} & \text{if } \theta\ge\max(y_1,\ldots,y_n);\\ 0 & \text{else.}\end{cases}$$

Taking $\frac{d}{d\theta}$ doesn't work because $L(\theta:\underline{y})$ is discontinuous at $\max(y_1,\ldots,y_n)$ and has no derivative there. However, it is obvious that $L(\theta:\underline{y})$ is maximised at $\max(y_1,\ldots,y_n)$ and so

$$\hat{\theta}=\max(y_1,\ldots,y_n).$$

3.5.1 Incomplete samples

The method of maximum likelihood can be applied in situations where the sample is incomplete. For example, truncated data or censored data in which observations are known to be greater than a certain value, or multiple claims where the number of claims is known to be two or more.

Censored data arise when we have information about the full range of possible values but that information is not complete (e.g. when we only know that there are, say, 6 values greater than 500). Truncated data arise when we have no information about part of the range of possible values (e.g. when we have no information at all about values greater than 500).

In these situations, as long as the likelihood (the probability of observing the given information) can be written as a function of the parameter(s), then the method can be used. Again in such cases the solution may be more complex, perhaps requiring numerical methods.

For example, suppose a sample yields n observations and m observations greater than z. Since the values above z are unknown we cannot use $L(\underline{\theta}) = \left(\prod_{i=1}^{n+m} f(y_i; \underline{\theta})\right)$. We use instad the formula given.

If the information is more detailed than 'greater than z' we can use a more detailed likelihood function. For example, if we have m observed values between x and z, and p observed values above z, in addition to the n known values, then we would use

$$L(\underline{\theta}) = \left(\prod_{i=1}^{n} f(y_i; \underline{\theta})\right) \times P(x < X \leq z)^m \times P(X > z)^p.$$

Example 3.1. Claims in thousands of pounds on a particular policy have a distribution with p.d.f. given by:

$$f(y) = 2cye^{-cy^2}, \quad y > 0.$$

Seven of the last ten claims are

$$1.05, 3.38, 3.26, 3.22, 2.71, 2.37, 1.85$$

The three remaining claims were known to be greater than $£6,000$. Calculate the maximum likelihood estimate of c.

Let $y_1, \ldots, y_7$ be the claims. We have 7 known claims and 3 claims greater than 6. So the likelihood is:

$$L(c) = \prod_{i=1}^{7} f(y_i) \times P(Y > 6)^3$$

Since

$$P(Y > 6) = \int_6^\infty 2cye^{-cy^2}\, du = -e^{-cy^2}\Big|_6^\infty = e^{-36c}$$

and

$$\sum_{i=1}^{7} y_i^2 = 49.91,$$

we have

$$L(c) = \prod_{i=1}^{7}(2cy_i e^{-cy_i^2}) \times (e^{-36c})^3 = C(\underline{y})c^7 e^{-c\sum_{i=1}^7 y_i^2 - 108c} = C(\underline{y})c^7 e^{-157.91c}$$

for some function $C(\underline{y})$. The log-likelihood is

$$\ell(c) = \ln(C(\underline{y})) + 7\ln(c) - 157.91c.$$

Setting $\ell'(c) = 0$ gives

$$\frac{7}{c} - 157,91 = 0 \Rightarrow c = \frac{7}{157.91} = 0.0443.$$

If we have some claims about which nothing is known (i.e. we don't even know whether there are any claims of a particular type), then the data are said to be truncated, rather than censored. We need to take a slightly different approach here.

Example 3.2. The number of claims in a year on a pet insurance policy are distributed as follows

No. of claims	0	1	2	≥ 3
$P(N = n)$	5θ	3θ	θ	$1 - 9\theta$

Information from the claims file for a particular year showed that there were 60 policies with 1 claim, 24 policies with 2 claims and 16 policies with 3 or more claims. There was no information about the number of policies with no claims. Calculate the maximum likelihood

estimate of θ.

Since we have no information at all about zero claims, we need to determine the truncated distribution. All we do is omit the zero claims probability and scale up the remaining probabilities (which only total to $1-5\theta$) so that they now total to 1:

No. of claims	1	2	≥ 3
$P(N=n)$	$\frac{3\theta}{1-5\theta}$	$\frac{\theta}{1-5\theta}$	$\frac{1-9\theta}{1-5\theta}$

These probabilities can also be thought of as conditional probabilities, ie the first probability in the table is actually

$$P(N=1|N>0)=\frac{P(N=1)}{P(N>0)}=\frac{3\theta}{1-5\theta}$$

.

The likelihood is

$$\begin{aligned} L(\theta|N>0) &= \text{constant } \times P(N=1|N>0)^{60}P(N=2|N>0)^{24}P(N\geq 3|N>0)^{16} \\ &= \text{constant } \times \left(\frac{3\theta}{1-5\theta}\right)^{60}\left(\frac{\theta}{1-5\theta}\right)^{24}\left(\frac{1-9\theta}{1-5\theta}\right)^{16} \\ &= \text{constant } \times \frac{\theta^{84}(1-9\theta)^{16}}{(1-5\theta)^{100}}. \end{aligned}$$

The first constant arises from the fact that we don't know which of the policies had 1 claim, etc. and so there is some combinatorial factor to account for this.

The log-likelihhod is

$$\ln L(\theta|N>0)=\text{constant } +84\ln\theta+16\ln(1-9\theta)-100\ln(1-5\theta).$$

Differentiating gives

$$\frac{d}{d\theta}L(\theta|N>0)=\frac{84}{\theta}-\frac{144}{1-9\theta}+\frac{500}{1-5\theta}.$$

Setting this euqal to 0 produces

$$84(1-9\theta)(1-5\theta)-144\theta(1-5\theta)+500\theta(1-9\theta)=0\Rightarrow 84-820\theta=0\Rightarrow\theta=0.102$$

so $\hat{\theta} = 0.102$.

3.5.2 Independent samples

For independent samples from two populations which share a common parameter, the overall likelihood is the product of the two separate likelihoods.

Example 3.3. The number of claims, X, per year arising from a low-risk policy has a Poisson distribution with mean μ. The number of claims, Y, per year arising from a high-risk policy has a Poisson distribution with mean 2μ.

A sample of 15 low-risk policies had a total of 48 claims in a year and a sample of 10 high-risk policies had a total of 59 claims in a year. Determine the maximum likelihood estimate of μ based on this information.

The likelihood for these 15 low-risk and 10 high-risk policies is:

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{15} P(X = x_i) \times \prod_{j=1}^{10} P(Y = y_j) \\ &= \prod_{i=1}^{15} \frac{\mu^{x_i} e^{-\mu}}{x_i!} \times \prod_{j=1}^{10} \frac{(2\mu)^{y_j} e^{-2\mu}}{y_i!} \\ &= \text{ constant } \times \mu^{\sum_{i=1}^{15} x_i} e^{-15\mu} \times \mu^{\sum_{j=1}^{10} y_j} e^{-20\mu} \\ &= \text{ constant } \times \mu^{48} e^{-15\mu} \times \mu^{59} e^{-20\mu} \\ &= \text{ constant } \times \mu^{107} e^{-35\mu} \end{aligned}$$

The log-likelihood is

$$\ln L(\mu) = \text{ constant } + 107 \ln(\mu) - 35\mu$$

Setting the derivative to 0 gives

$$\frac{d}{d\mu} \ln L(\mu) = \frac{107}{\mu} - 35 = 0 \Rightarrow \hat{\mu} = \frac{107}{35} = 3.057.$$

4 Topics for simulating random variables

We will need the following topics for the R demonstrations of generating random variables. The Gamma and Beta distributions are important in the section on Bayesian statistics, as well.

4.1 The Gamma, Beta and LognormalDistributions

The Gamma and Beta distributions are important distributions which you may not have come across before.

4.1.1 The gamma distribution

The gamma family of distributions has two positive parameters and is a versatile family. The PDF can take different shapes depending on the values of the parameters. The range of the variable is $\{x \in \mathbb{R} : x > 0\}$. The parameter α changes the shape of the graph of the PDF, and the parameter λ changes the x scale. The gamma distribution may be written in shorthand as Gamma(α, λ). The gamma function $\Gamma(\alpha)$ is defined for $\alpha > 0$ as follows:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

In particular, $\Gamma(1) = 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for $\alpha > 1$; if α is positive integer, then $\Gamma(\alpha) = (\alpha - 1)!)$; and $\Gamma(1/2) = \sqrt{\pi}$. The PDF of the gamma distribution with shape parameter α and rate parameter λ is defined by

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0.$$

It can alternatively be speicified with shape parameter α and scale parameter $s = 1/\lambda$. It has expectation

$$\mathbb{E}(X) = \frac{\alpha}{\lambda}$$

and variance

$$\text{var}(X) = \frac{\alpha}{\lambda^2}.$$

The Gamma$(1, \lambda)$ distribution is the same as the Exponential(λ) distribution. The Gamma$(\nu/2, 1/2)$ distribution is the same as the chi square χ^2 distribution with ν degrees of freedom.

4.1.2 The beta distribution

Another versatile family of distributions with two parameters $\alpha > 0$, $\beta > 0$, is the Beta distribution. The range of the variable is $\{x \in \mathbb{R} : 0 < x < 1\}$. The beta function $B(\alpha, \beta)$ is defined by:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

The relationship between beta functions and gamma functions is:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The PDF of the beta distribution is defined by:

$$f_X(x) = \frac{1}{B(\alpha, \beta)}x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1.$$

It has expectation

$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$$

and variance

$$\text{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The Beta$(1, 1)$ distribution is the same as the Uniform$(0, 1)$ distribution.

The beta distribution is a useful distribution because it can be rescaled and shifted to create a wide range of shapes – from straight lines to curves, and from symmetrical distributions to skewed distributions. Since the random variable can only take values between 0 and 1, it is often used to model proportions, such as the proportion of a batch that is defective or the percentage of claims that are over £1,000.

4.1.3 The lognormal distribution

The lognormal distributions have range $\{x \in \mathbb{R} : x > 0\}$. They have parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. If $Y \sim \text{Lognormal}(\mu, \sigma^2)$, then Y has the distribution of e^Z, where $Z \sim \text{N}(\mu, \sigma^2)$.

Lemma 4.1. *The p.d.f. of the* $\text{Lognormal}(\mu, \sigma^2)$ *distribution is*

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}y} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right), \quad y > 0. \tag{2}$$

Proof.

$$P(Y \leq y) = P(e^Z \leq y) = P(Z < \ln y)$$
$$= \int_{-\infty}^{\ln y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du$$

Therefore

$$f_Y(y) = \frac{d}{dy} P(Y \leq y)$$

which equals (2). Because $\ln y \to \infty$ a lot slower than $y \to \infty$, the Lognormal distributions have a much longer tail than do the normal distributions.

4.2 Simulating random variables from the uniform distribution

The following lemma is useful for simulating random variables. Note that F^{-1} means functional inverse.

Lemma 4.2. *Let* F *be the c.d.f. of continuous random variable and let* $U \sim \text{Uniform}[0, 1]$. *The random variable* $F^{-1}(U)$ *has c.d.f. identicle to* F.

Proof. We want to show that $P(F^{-1}(U) \leq x) = F(x)$ for $x \in \mathbb{R}$. We have

$$P(F^{-1}(U) \leq x) = P(F(F^{-1}(U)) \leq F(x)) = P(U \leq F(x)).$$

Now, $0 \leq F(x) \leq 1$ and $P(U \leq u)$ if $0 \leq y \leq 1$. Therefore,

$$P(F^{-1}(U) \leq x) = F(x).$$

□

R, like many statistical packages, simulates independent observations of the from the $U \sim$Uniform$[0,1]$ distribution. Suppose we need simulations of a continuous random variable Z with a c.d.f. F (which may not be available in R). We may obtain simulations of uniform random variable U and transform it to $Z = F^{-1}(U)$.

Example 4.3. Let $X \sim$Exponential(λ).

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0$$

and if $u = 1 - e^{-\lambda x}$, then $x = -\frac{1}{\lambda}\ln(1-u)$. Therefore,

$$F^{-1}(U) = -\frac{1}{\lambda}\ln(1-U).$$

is Exponential(λ) distributed. If we can simulate a Uniform$(0,1)$ random variable, call the output u, then $-\frac{1}{\lambda}\ln(1-u)$ simulates a Exponential(λ) random variable.

4.3 For loops in R

Before discussing the R code used for the bootstrap, we must earn about for loops. Loops are a convenient way to repeat the same operation over and over. Here, when we say repeat the same operation, we mean that the structure of the commands executed is the same. So, we may still use different inputs and get different answers for each repetition (or iteration).

The structure of a for loop is as follows:

```
for (<range to iterate over>) {
<commands to perform>
}
```

Inside the round brackets we need to give R information about how we want it to repeat the commands. The format that R expects for this section is:

```
for (<variable> in <vector>)
```

where

- $< \text{vector} >$ is the set of values for which we want the commands to be repeated. The commands inside the curly brackets of the for loop will be executed for each value in $< \text{vector} >$

- $< \text{variable} >$ is the object that will take all the values in the set of values $< \text{vector} >$. It is known as the iterating variable and is used to control the loop. We can also use this object in the commands inside the curly brackets as we'll see shortly.

Inside the curly brackets we need to tell R which operations to perform when executing each loop. So,we can just type any code that we want to be repeated.

A simple example is as follows:

```
for (i in 1:5){
print(1)
}
```

```
[1] 1
[1] 1
[1] 1
[1] 1
[1] 1
```

In this for loop we are asking R to perform an operation for each value from 1 to 5. Here, we just get R to print the number 1 to the console for each iteration. So, it prints it 5 times.

Note that if we didn't include the print() function then the for loop wouldn't output anything. This is just how for loops work in R. If you want to show something in the console for each iteration, remember to print it.

Now we extend our example to use the iterating variable inside the commands that we execute:

```
for (i in 1:5){
print(i)
```

```
}
```

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
```

We are still asking R to perform commands for each number 1 to 5. However, instead of printing 1 to the console each time, we are printing the value of the iterating variable, i.

Although i is a common letter to use as the iterating variable, it can be anything that we like.

Even though the ":" notation is probably what you will use the most, any vector of values will work:

```
for (my.vec in c("a", "b")){
print(my.vec)
}
```

```
[1] "a"
[1] "b"
```

In the above example we are iterating over a character vector instead of a vector of numbers. However, the fundamental execution of the loop is the same.

For very simple for loops with only one command, you can skip the curly brackets and write the command on the same line:

```
for (i in 1:5) print(i)
```

If you want to execute multiple commands inside a loop, then you have two options. Either

you need to separate them on multiple lines:

```
for (i in 1:5) { j=i+1
print(j)
}
```

or you need to separate each command on the same line with ";" , as well as putting everything inside curly brackets:

```
for (i in 1:5) {j = i + 1; print(j)}
```

4.4 The bootstrap method

Suppose that we want to make inferences about parameter θ using observed data $(y_1, y_2, \ldots, y_n)$ which follow a distribution with cumulative distribution function $F(y : \theta)$. Usually inference is based on the **sampling distribution** of an estimator $\hat{\theta}$. The sampling distribution can be obtained by theoretical results, but it can also be based on a large number of samples from $F(y : \theta)$.

For example, suppose we have a sample $(y_1, y_2, \ldots, y_n)$ an exponential distribution with parameter λ and we wish to make inferences about λ. The CLT tells us that asymptotically $\overline{Y} \sim \mathrm{N}(\lambda^{-1}, (n\lambda^2)^{-1})$ and we can use this sampling distribution to estimate quantities of interest (e.g. for confidence intervals or tests about λ). However, there will be cases where assumptions or asymptotic results may not hold (or we may not want to use them – e.g. when samples are small).

Then one alternative option is to use the bootstrap method. Bootstrap allows us to avoid making assumptions about the sampling distribution of a statistic of interest, by instead forming an **empirical sampling distribution** of the statistic. This is generally achieved by resampling based on the available sample.

The procedure for performing the non-parametric bootstrap, when estimating a parameter

θ can be described as follows. The **empirical distribution** $\hat{F}_n(y)$ of the data is defined to be

$$\hat{F}_n(y) = \frac{1}{n} \left| \{1 \leq i \leq n : y_i \leq y\} \right|.$$

Then perform the following steps:

1. Draw a sample of size n from $\hat{F}_n(y)$. This is a a bootstrap sample $(y_1^*, y_2^*, \ldots, y_n^*)$ with each of the y_i^* sampled uniformly at random from $(y_1, y_2, \ldots, y_n)$ with replacement.

2. Obtain an estimate of $\hat{\theta}^*$ from the $(y_1^*, y_2^*, \ldots, y_n^*)$ in the same way that $\hat{\theta}$ is obtained from $(y_1, y_2, \ldots, y_n)$.

Repeat steps 1 and 2 B times, say, thereby obtaining B estimates $(\theta_1^*, \theta_2^*, \ldots, \theta_B^*)$. The set of estimates $(\theta_1^*, \theta_2^*, \ldots, \theta_B^*)$ can be used for any desired inference regarding the estimator $\hat{\theta}$, and particularly to estimate its properties. For example we can:

- estimate the mean of estimator $\hat{\theta}$ by using the sample mean $\frac{1}{B} \sum_{j=1}^{B} \theta_j^*$
- estimate its median, using the 0.5 empirical quantile of the bootstrap estimates θ_j^*.
- estimate the variance of estimator $\hat{\theta}$ by using the sample variance of the bootstrap estimates $\frac{1}{B-1} \left\{ \sum_{j=1}^{B} (\theta_j^*)^2 - \frac{1}{B} \left(\sum_{j=1}^{B} \theta_j^* \right)^2 \right\}$
- estimate a $(1-\alpha)\%$ confidence interval for $\hat{\theta}$ by $(k_{\alpha/2}, k_{1-\alpha/2})$, where k_α denotes the αth empirical quantile of the bootstrap values θ_j^*.

4.4.1 An example

Example 4.4. Suppose we have the following sample of 10 values (to 2 DP) from an exponentially distributed random variable with unknown parameter λ:

```
0.61 6.47 2.56 5.44 2.72 0.87 2.77 6.00 0.14 0.75
```

We can use the following R code to obtain a single resample with replacement from this original sample.

```
sample.data <-c(0.61, 6.47, 2.56, 5.44, 2.72, 0.87, 2.77, 6.00, 0.14, 0.75)
sample(sample.data, replace=TRUE)
```

If we do this, R automatically gives us a sample of the same size as the original data sample, i.e. we obtain a sample of size 10 in this case.

The following R code obtains $B = 1000$ estimates $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_B^*)$ using $\lambda_j^* = 1/\overline{y}_j^*$ and stores them in the vector estimate. The command set.seed() initiates the random selections in sample. Whenever the code is run with the same value in set.seed, 47 for example, exactly the same results are obtained. This allows for reproducability of the results.

```
set.seed(47)
estimate<-rep(0,1000)
for (i in 1:1000)
       {x<-sample(sample.data, replace=TRUE);
        estimate[i]<-1/mean(x)} }
```

An alternative would be to use:

```
set.seed(47)
estimate <-replicate(1000, 1/mean(sample(sample.data, replace=TRUE)))
```

4.4.2 Parametric bootstrap

If we are prepared to assume that the sample is considered to come from a given distribution, we first obtain an estimate of the parameter of interest $\hat{\theta}$ (e.g. using maximum likelihood, or method of moments). Then we use the assumed distribution, with parameter equal to $\hat{\theta}$, to draw the bootstrap samples. Once the bootstrap samples are available, we proceed as with the non-parametric method before.

Example 4.5. Let $Y_i \sim \text{Exp}(\lambda)$, $i = 1, \ldots, n$, where λ is unknown. We know that $\overline{Y}$ is an unbiased estimator of λ. So, we may generate bootstrap samples y^* from the Exponential$(1/\overline{y})$ distribution.

Using our sample of 10 values (to 2 DP) from an Exp(λ) distribution with unknown parameter λ

```
0.61 6.47 2.56 5.44 2.72 0.87 2.77 6.00 0.14 0.75
```

our estimate would for λ would be $\hat{\lambda} = 1/\overline{y} = 1/2.833 = 0.3530$. We now use the Exp$(0.3530)$ distribution to generate the bootstrap samples. Note that this is parametric as we are using the exponential distribution to obtain new samples.

The following R code obtains $B = 1000$ estimates $(\lambda_1^*, \lambda_2^*, \ldots, \lambda_B^*)$ using $\lambda_j^* = 1/\overline{y}_j^*$ and stores them in the vector param.estimate

```
set.seed(47)
param.estimate <-replicate(1000, 1/mean(rexp(10,rate=1/mean(sample.data))))
```

5 Bayesian Statistics

5.1 Introduction to Bayesian Statistics

Bayesian statistics provides us with tools to update our beliefs in light of new data/evidence.

Example 5.1 (F1 race)**.** Consider the following assumptions:

- Suppose, out of four F1 championship races between Lewis Hamilton and Sebastian Vettel, Hamilton won 3 times while Vettel managed only 1.
- If you were to bet on the winner of next race, who would it be ?
- Here's the twist.
- What if you are told that it rained once when Vettel won and once when Hamilton won and it is definite that it will rain on the next date.
- So, who would you bet your money on now ?
- By intuition, it is easy to see that chances of winning for Vettel have increased drastically.
- But the question is: by how much ?

We define conditional probability as the probability of an event A given B equals the probability of B and A happening together, divided by the probability of B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

A **Sample space** is the set of all possible outcomes for the problem. We can subdivide it into different outcomes - called partitioning. Events are subsets of the sample space. If $A_1, A_2, A_3, \ldots, A_n$ are subsets of a sample space S, then they partition S if:

1. $A_1 \cup A_2 \cup \cdots \cup A_n = S$
2. $A_i \cap A_j = \emptyset$ whenever $i \neq j$

Theorem 5.2 (Bayes)**.** *Let $B_1, B_2, \ldots, B_n$ be a patition of sample spavce S such that $P(B_i) > 0$ for all i and suppose A is an event. Then*

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)},$$

where

$$P(A) = \sum_{j=1}^{n} P(A|B_j)P(B_j)$$

for all $i = 1, 2, \ldots, n$.

Returning to our F1 driver's example let's try to answer it by attaching some probabilities to some events. Let A be the event of it raining. Let B be the event of Sebastian Vettel winning. Then: $P(A) = 1/2$ since it rained twice out of four days. $P(B) = 1/4$ since Vettel only won one race out of four. $P(A|B) = 1$, since it rained every time Vettel won. Given that it will rain during the next race, what is the probability of Vettel winning the next race? The answer is

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = (1 \times 1/4)/(1/2) = 1/2.$$

Example 5.3 (Faulty Garment)**.** Consider the following assumptions:

- Three manufacturers supply clothing to a retailer.
- 60% of the stock comes from manufacturer 1, 30% from manufacturer 2 and 10% from manufacturer 3.
- 10% of the clothing from manufacturer 1 is faulty, 5% from manufacturer 2 is faulty and 15% from manufacturer 3 is faulty.

- What is the probability that a faulty garment comes from manufacturer 3?

We first need to define some events. Let A be the event that a garment is faulty. Let B_i be the event that the garment comes from manufacturer i. Then substituting the figures into the formula for Bayes' Theorem we get:

$$\begin{aligned} P(B_3|A) &= P(A|B_3)P(B_3)/P(A) \\ &= P(A|B_3)P(B_3)/[P(A|B_1)P(B_1)/ + P(A|B_2)P(B_2)/ + P(A|B_3)P(B_3)] \\ &= (0.15)(0.1)/[(0.1)(0.6) + (0.05)(0.3) + (0.15)(0.1)] \\ &= 0.167 \end{aligned}$$

So although manufacturer 3 supplies only 10% of the garments to the retailer, nearly 17% of the faulty garments come from that manufacturer.

5.1.1 Prior and posterior distributions

When we apply Bayes' Theorem in statistics, the random variables we are conditioning on will be continuous. In Bayesian statisticxs, there is a probability distribution on θ which represents the belief of the statistician that θ takes different values. This belief is quantified by a probability distribution $f(\theta)$ called the **prior distribution** of θ. This allows the use of any knowledge available about possible values for θ before the collection of any data. Then after collecting appropriate data, the **posterior distribution** of θ is determined and this forms the basis of all inference concerning θ. Suppose $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)$ is a random sample from a population specified by the density or probability function $f(y;\theta)$ and tbat it is required to estimate θ. As θ is a random variable it should really be denoted by the capital Θ and its prior density written as $f_\Theta(\theta)$. However, for simplicity, no distinction will be made between θ and Θ and the density will simply be denoted by $f(\theta)$. Note that referring to a density here implies that θ is continuous. In most applications this will be the case, as even when X is discrete (like in the Binomial or Poisson distributions) the parameter (p or λ) will vary in continuous space $(0,1)$ or $(0,\infty)$. Finally, the population density or probability function will be denoted by $f(x|\theta)$ rather than the earlier $f(x;\theta)$ as it represents the conditional distribution of X given

θ.

We must see how to find the posterior distribution. Suppose that $\underline{X}$ is a random sample from a population specified by $f(x|\theta)$ and that θ has the prior density $f(\theta)$. The posterior density of $\theta|X$ is determined by applying the basic definition of a conditional density:

$$f(\theta|\underline{x}) = \frac{f(\theta, \underline{x})}{f(\underline{x})} = \frac{f(\underline{x}|\theta)f(\theta)}{f(\underline{x})}. \tag{3}$$

Note that $f(\underline{x}) = \int f(\underline{x}|\theta)f(\theta)d\theta$. This result is like a continuous version of Bayes' Theorem.

A useful way of expressing the posterior density is to use proportionality. The denominator $f(\underline{x})$ in (3) does not involve θ and is just the constant needed to make $f(\theta, \underline{x})$ a proper density that integrates to unity so:

$$f(\theta|\underline{x}) \propto f(\underline{x}|\theta)f(\theta). \tag{4}$$

The joint density of the sample values $f(x|\theta)$ is none other than the likelihood and therefore (4) says that the posterior is proportional to the likelihood times the prior.

There is a four-step procedure for determining the posterior.

Step 1 **Select a prior distribution.**

Write down the prior distribution of the unknown parameter.

Step 2 **Determine the likelihood function.**

Write down the (joint) likelihood function for the observations.

Step 3 **Determine the posterior parameter distribution**

Multiply the parameter distribution and the likelihood function to find the form of the posterior parameter distribution.

You can ignore any multipliers that don't contain the unknown parameter. These will be "absorbed" by the proportional sign.

Step 4 **Identify the posterior parameter distribution.**

Either

(a) look for a standard distribution that has a PDF with the same algebraic form and range of values as the posterior distribution you have found (eg by comparing with

the PDFs in your tables).

or

(b) If your posterior distribution does not match any of the standard distributions then "integrate (or sum) out" the unknown parameter to find the constant that makes the integral (or sum) of the PDF/PF of the posterior distribution equal to 1.

Example 5.4 (Number of claims on an insurance policy)**.** A new insurance policy is sold to a limited number of existing policy holders. Suppose the number of claims in the first six months are 18. Assume the number of claims per month to be independent from month to month, and that the number of claims per month has a Poisson distribution with mean θ.

You are to do a), b) and c).

a) Write down the likelihood function.

The prior distribution of θ is given by a gamma distribution. When designing the policy it was thought that based on the claims history of these policy holders that the mean number of claims per month would be 4 claims with variance 1.

b) Show that a Gamma distribution Gamma$(16, 4)$ is a suitable prior.

c) Find the posterior distribution of θ.

Solution:

a) The likelihood is given by

$$f(\underline{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \propto e^{-n\theta}\theta^s,$$

where

$$s = \sum_{i=1}^{n} x_i.$$

Here $n = 6$ and $s = 18$ so the likelihood is proportional to $e^{-6\theta}\theta^{18}$.

b) The mean and variance of a Gamma(a, b) distribution are a/b and a/b^2 (from a table of distributions), so Gamma$(16, 4)$ has mean 4 and variance 1.

c) The prior satisfies $f(\theta) \propto \theta^{15}e^{-4\theta}$ and so then posterior satisfire

$$f(\theta|\underline{x}) \propto \theta^{15}e^{-4\theta}\theta^{18}e^{-6\theta} \propto \theta^{33}e^{-10\theta}.$$

Thus the posterior is Gamma$(34, 10)$ with mean 3.4 and variance 0.34

How do we interpret our answer? We initially think that the parameter θ has mean 4 and variance 1. In light of new data/evidence (number of claims in the first six months), we update our beliefs and say that the parameter θ has mean 3.4 and variance 0.34.

Example 5.5 (Short or long life batteries)**.** -

- You bought some batteries from a shop a short time ago.
- You know that the batteries are all either short-life batteries with a mean life of 100 hours or long-life batteries with a mean life of 1,000 hours.
- As there is no label on either the box or the individual batteries you cannot tell which
- You have no initial opinion as to whether a battery is short or long life as you have not shopped here before.
- After 80 hours the ten batteries you have been using are still going.
- Assuming that the life of an individual battery follows an exponential distribution, what do you believe to be the probability that you bought long-life batteries?

Solution:

Step 1: Select a prior distribution.

We are told that we have no opinion as to whether we have been sold short or long-life batteries. Therefore we think they are equally likely. Therefore the prior distribution for λ, the parameter for the distribution of the lifetimes, is the discrete distribution where $P(\lambda = 0.01) = 0.5$ and $P(\lambda = 0.001) = 0.5$

Step 2: Determine the likelihood.

Let X be the random variable representing the lifetime of a battery. We are told $X \sim$

$\text{Exp}(\lambda)$. Thetrefore

$$P(\text{battery will still be working after 80 hours}) = P(X > 80) = \int_{80}^{\infty} \lambda e^{-\lambda x} dx = e^{-80\lambda}.$$

Now assuming that the lifetime of each battery is independent from the others, which is a reasonable assumption to make, then

$$P(X_1 > 80, X_2 > 80, \ldots, X_{10} > 80) = P(X_1 > 80)P(X_2 > 80) \cdots P(X_{10} > 80)$$

So the likelihood that all ten will still be working at that time is $(e^{-80\lambda})^{10} = e^{-800\lambda}$.

Step 3: Determine the posterior parameter distribution

We can construct the posterior distribution using our proportional approach.

λ	Prior	Likelihood	Prior $\times$ Likelihood	Posterior $\propto$ Prior $\times$ Likelihood
$1/100$	0.5	$e^{-8} = 0.00034$	0.00017	0.000746
$1/100$	0.5	$e^{-0.8} = 0.44933$	0.22466	0.999254
Totals	1		0.22483	1

Step 4: Identify the posterior parameter distribution.

We can see that the posterior distribution for λ given observations from the ten batteries is the discrete distribution where

$$P(\lambda = 1/100 | X_1 > 80, X_2 > 80, \ldots, X_{10} > 80) = 0.000746,$$

$$P(\lambda = 1/1000 | X_1 > 80, X_2 > 80, \ldots, X_{10} > 80) = 0.999254r.$$

Strictly speaking, we do not need Step 4 for this question.

Remember that the question was: (Given observations from the batteries you bought), what do you believe to be the probability that you bought long-life batteries? Our final answer is 0.999254. How do we interpret our answer? We initially think that a battery is equally likely to be short or long life. In light of new data/evidence (observations from ten batteries), we update our beliefs about whether a battery is short or long life.

Gentle reminder: Bayesian statistics provides us with tools to update our beliefs in light of new data/evidence.

5.1.2 Conjugate priors

For a given likelihood function, if the prior distribution leads to a posterior distribution belonging to the same family as the prior distribution, then this prior is called the **conjugate prior** for this likelihood. The likelihood function determines which family of distributions will lead to a conjugate pair, i.e. a prior and posterior distribution that come from the same family. Conjugate distributions can be found by selecting a family of distributions that has the same algebraic form as the likelihood function, treating the unknown parameter as the random variable.

Example 5.6 (IID observations from a geometric distribution)**.** Suppose that $X_1, X_2, \ldots, X_n$ are IID (independent and identically distributed) observations from a geometric distribution with parameter p. i.e. a distribution having probability function:

$$P(X = x) = p(1-p)^{x-1}, x = 1, 2, 3, \ldots$$

where p is unknown. We want to find a family of distributions that would result in conjugate prior and posterior distributions. The likelihood function is:

$$\prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\sum_{i=1}^{n} x_i - n}.$$

We need a family of functions of the form $p^{\text{something}}(1-p)^{\text{something}}$ where $0 < p < 1$. Inspecting our table of distributions we see that this means we need a Beta distribution.

5.1.3 Uninformative prior distributionss

An uninformative prior distribution assumes that an unknown parameter is equally likely to take any value from a given set. In other words, the parameter is modelled using a uniform distribution.

As an example, suppose that we have a random sample $X_1, X_2 \ldots, X_n$ from a normal

population with mean μ but we have no prior information about μ. In this case it would be natural to model μ using a uniform distribution. Since μ can take any value between $-\infty$ and ∞, the appropriate uniform prior is Uniform$(-\infty, \infty)$. This leads to a problem, however, since the PDF of this "distribution" is 0 everywhere.

We can get round this problem by using the distribution Uniform$(-m, m)$ and then letting $m \to \infty$. If $\mu \sim$Uniform$(-m, m)$, then the prior PDF of μ is:

$$f_{\text{prior}} = \begin{cases} \frac{1}{2m} & \text{if } -m < \mu < m; \\ 0 & \text{otherwise.} \end{cases}$$

Also, since the data values come from a normal population, the likelihood function is

$$L(\mu) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right).$$

The likelihood function can alternatively be expressed as:

$$l(\mu) = C \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

where C is a constant that bdoes not depend on μ. Combining the prior PDF with the likelihood function gives:

$$f_{\text{post}} = \begin{cases} K \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right) & \text{if } -m < \mu < m; \\ 0 & \text{otherwise.} \end{cases}$$

where K is a constant that does not depend on μ. This constant is required to ensure that the PDF integrates to 1. Letting $m \to \infty$, we see that the posterior PDF is proportional to

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \quad \text{for } -\infty < \mu < \infty.$$

Not all statisticians agree that this example really shows that it is all right to use uninformative prior distributions, regardless of whether or not they are useful in practise.

Notice that the PDF of this posterior distribution is proportional to the likelihood function.

This should be intuitive as, by definition, a posterior distribution is obtained by combining two pieces of information:

- prior knowledge of the parameter
- the sample data.

§ However, in this case we are using an uninformative prior as we have no prior knowledge of the parameter. The posterior distribution is therefore determined solely by the sample data.

5.1.4 Loss functions

To obtain an estimator of θ, a **loss function** must first be specified. This is a measure of the "loss" incurred when $g(X)$ is used as an estimator of θ. A loss function is sought which is

- zero when the estimation is exactly correct, that is, $g(X) = \theta$
- positive and does not decrease as $g(X)$ gets further away from θ.

The **Bayesian estimator** is the $g(X)$ that minimises the expected loss with respect to the posterior distribution. There is one very commonly used loss function, called quadratic or squared error loss. Two others are also used in practice.

The main loss function is quadratic loss defined by:

$$L(g(\underline{x}), \theta) = [g(x) - \theta]^2$$

It is related to mean square error loss. A second loss function is absolute error loss defined by:

$$L(g(x), \theta) = |g(x) - \theta|$$

A third loss function is "all-or-nothing" loss defined by:

$$L(g(\underline{x}), \theta) = \begin{cases} 0 & \text{if } g(\underline{x}) = \theta \\ 1 & \text{if } g(\underline{x}) \neq \theta \end{cases}$$

The **expected posterior loss** (EPL) is:

$$\text{EPL} = E[L(g(\underline{X}), \theta)] = \int [L(g(\underline{x}), \theta) f(\underline{\theta}|\underline{x}) d\theta.$$

The EPL depends on the data $\underline{x}$ and the estimate $g(\underline{x})$. Remember that the expected value of any function is found by multiplying the function by the appropriate PDF and integrating. We want to find the Bayesian estimator, $g(\underline{x})$, which minimises the EPL.

Under quadratic loss we get:

$$\text{EPL} = \int (g(\underline{x}) - \theta)^2 f(\theta|\underline{x}) d\theta$$

so

$$\frac{d}{dg(\underline{x})}\text{EPL} = 2 \int (g(\underline{x}) - \theta) f(\theta|\underline{x}) d\theta$$

Equating this to 0, we get

$$g(\underline{x}) \int f(\theta|\underline{x}) d\theta = \int \theta f(\theta|\underline{x}) d\theta$$

But since $\int f(\theta|\underline{x}) d\theta = 1$ we get

$$g(\underline{x}) = \int \theta f(\theta|\underline{x}) d\theta = E(\theta|\underline{x}).$$

Therefore, under quadratic loss function, the Bayesian estimator, $g(\underline{x})$, which minimises the EPL is the mean of the posterior distribution.

Under absolute error loss we get:

$$\text{EPL} = \int |g(\underline{x}) - \theta| f(\theta|\underline{x}) d\theta$$

Assuming the range for θ is $(-\infty, \infty)$,

$$\text{EPL} = \int_{-\infty}^{g(\underline{x})} (g(\underline{x}) - \theta) f(\theta|\underline{x}) d\theta + \int_{g(\underline{x})}^{\infty} (\theta - g(\underline{x})) f(\theta|\underline{x}) d\theta$$

and so

$$\frac{d}{dg(\underline{x})}\text{EPL} = \int_{-\infty}^{g(\underline{x})} f(\theta|\underline{x})d\theta - \int_{g(\underline{x})}^{\infty} f(\theta|\underline{x})d\theta.$$

Equating this to zero gives

$$\int_{-\infty}^{g(\underline{x})} f(\theta|\underline{x})d\theta = \int_{g(\underline{x})}^{\infty} f(\theta|\underline{x})d\theta.$$

This means that

$$P(\theta \leq g(\underline{x})) = P(\theta \geq g(\underline{x})) = 1/2.$$

Thus, under absolute error loss function, the Bayesian estimator, g(x), which minimises the EPL is the median of the posterior distribution.

Under all-or-nothing loss the differentiation approach cannot be used so we use a direct approach:

$$L(g(\underline{x}), \theta) = \begin{cases} 0 & \text{if } g(\underline{x}) - \epsilon < \theta < g(\underline{x}) + \epsilon \\ 1 & \text{otherwise} \end{cases}$$

In the limit $\epsilon \to 0$ this tends to0 the required loss function. We have

$$\text{EPL} = 1 - \int_{g(\underline{x})-\epsilon}^{g(\underline{x})+\epsilon} f(\theta|\underline{x})d\theta = 1 - 2\epsilon \times f(g(\underline{x})\underline{x})$$

for small ϵ. The EPL is minimised by taking $g(\underline{x})$ to be the mode of $f\theta|\underline{x})$. Thus, under all-or-nothing loss function, the Bayesian estimator, $g(\underline{x})$, which minimises the EPL is the mode of the posterior distribution.

Example 5.7. We now return to our earlier example on the number of claims on an insurance policy and add part d) and e) to the question.

A new insurance policy is sold to a limited number of existing policy holders. Suppose the number of claims in the first six months are 18. Assume the number of claims per month to be independent from month to month, and that the number of claims per month has a Poisson distribution with mean λ.

a) Write down the likelihood function.

The prior distribution of θ is given by a gamma distribution. When designing the policy it

was thought that based on the claims history of these policy holders that the mean number of claims per month would be 4 claims with variance 1.

b) Show that a Gamma distribution Gamma$(16, 4)$ is a suitable prior.

c) Find the posterior distribution of θ.

d) Calculate the value of the Bayes estimate. Show it can be written as a weighted average of the prior mean and the data mean and interpret the weights.

Solution:

d) The value of the Bayes estimate $=$ the posterior mean $=$ 3.4

The prior mean was 4 and the sample mean $18/6 = 3$.

The weights are the relative information provided by the data (6 observations) and the prior (equivalent to 4 observations).

We can write the posterior mean as a weighted average of the sample mean and prior mean as follows:

$$3.4 = 3 \times \frac{6}{10} + 4 \times \frac{4}{10}$$

Example 5.8 (IID observations from a Gamma distribution)**.** Consider this problem:

- Suppose $X_1, X_2, \ldots, X_n$ are IID observations from a Gamma(α, λ) distribution, where λ is unknown, but α is known.
- The prior distribution of λ is Exp(θ) where θ is a known constant.
- Find the Bayesian estimator of λ under "All-or-nothing" loss.

Answer:

The prior distribution of λ is proportional to $e^{-\theta\lambda}$. The likelihood function is proportional to

$$\prod_{i=1}^{n} \lambda^{\alpha} e^{-\lambda x_i} = \lambda^{n\alpha} e^{-\lambda \sum_{i=1}^{n} x_i}$$

so the posterior distribution is proportional to

$$\lambda^{n\alpha} e^{-\lambda \sum_{i=1}^{n} x_i} \times e^{-\theta\lambda} = \lambda^{n\alpha} e^{-\lambda \sum_{i=1}^{n} x_i + \theta}$$

Looking at our statistical distributions, we can see that this corresponds to a

Gamma$(n\alpha + 1, \sum_{i=1}^{n} x_i + \theta)$ distribution.

Under "All-or-nothing" loss the unknown parameter is estimated by the mode of the posteriorn distribution. By differentiating the $\ln$ of the PDF and setting this equal to zero, we find that the mode of the Gamma(α, λ) distribution is $(\alpha - 1)/\lambda$. The Bayesian estimator of λ is the mode of the Gamma$(n\alpha + 1, \sum_{i=1}^{n} x_i + \theta)$ distribution, i.e.

$$\hat{\lambda} = \frac{n\alpha}{\sum_{i=1}^{n} x_i + \theta}.$$

5.2 Credibility theory

Some results on conditional expectation we will need are, for random variables X and Y:

$$E[X] = E[E[X|Y]],$$

$$E[f(Y)|Y] = f(Y),$$

and

$$E[Xf(Y)] = E[E(Xf(Y)|Y] = E[f(Y)E(X|Y)].$$

Two random variables X_1 and X_2 are **conditionally independent** given a third random variable Y if:

$$E[X_1X_2|Y] = E[X_1|Y]E[X_2|Y]$$

This means that both X_1 and X_2 depend on Y but if we know Y, then they are independent. This does **not** imply that X_1 and X_2 are unconditionally independent so it may be the case that:

$$E[X_1X_2] \neq E[X_1]E[X_2]$$

Example 5.9 (Poisson variables)**.** Suppose that $X_1|Y$ has a $\mathrm{Poisson}(Y)$ distribution and $X_2|Y$ has a $\mathrm{Poisson}(2Y)$ distribution. If we know the value of Y, say $Y = 2$ then $X_1|Y$ is $\mathrm{Poisson}(2)$ and $X_2|Y$ is $\mathrm{Poisson}(4)$ and these two random variables may very well be independent. However, if we don't know the value of Y then knowing something about $X_1|Y$ (for example knowing it takes the value 10, say) tells us something about Y and hence about

$X_2|Y$. So if the value of Y is unknown the two random variables are not independent since knowing something about one of them tells us something about the other. These random variables may be conditionally independent but not unconditionally independent.

The basic idea underlying the **credibility premium formula** is intuitively very simple. Let's consider an example.

Example 5.10 (Buses)**.** Suppose the local authority in a small town has run a fleet of 10 buses for a number of years. They wish to insure this fleet for the coming year against claims from accidents involving these buses. The pure premium for this insurance needs to be calculated. This is the expected cost of claims for the coming year. To calculate this, the following data are available:

- The data for the past five years for this particular fleet of buses show that the average cost of claims per annum (for the ten buses) has been £1,600.
- Suppose that, in addition to this information, there is data relating to a large number of local authority bus fleets from all over the United Kingdom, which show that the average cost of claims per annum per bus is £250.
- While the figure of £2,500 is based on many more fleets of buses than the figure of £1,600, some of the fleets of buses in this large dataset operate under very different conditions from the particular fleet at which we are looking.
- There are two extreme choices for the pure premium in the coming year:
    1. £1,600 could be chosen on the grounds that this estimate is based on the most appropriate data, whereas the estimate of £2,500 is based on less relevant data.
    2. £2,500 could be chosen on the grounds that this is based on more data and so is, in some sense, a more reliable figure.

The credibility approach to this problem is to take a weighted average of these two extreme answers. So we calculate the pure premium as:

$$Z \times 1,600 + (1 - Z) \times 2,500$$

where Z is some number between zero and one. Z is known as the credibility factor. Purely for the sake of illustration, suppose Z is set equal to 0.6 so that the pure premium is calculated to be £1,960. We will return to this example later on, but For now we will express the above ideas more formally.

The problem is to estimate the aggregate claims, or possibly, just the expected number of claims, in the coming year from a risk. By a risk we mean a single policy or a group of policies. These policies are, typically, short term policies and, for convenience, the term of the policies will be taken to be one year, although it could equally well be any other short period. The following information is available:

- $\bar{X}$ is an estimate of the expected aggregate claims / number of claims for the coming year based solely on the data from the risk itself.
- μ is an estimate of the expected aggregate claims / number of claims for the coming year based on **collateral data**, i.e. data for risks similar to, but not necessarily identical to, the particular risk under consideration.

The credibility premium formula (or credibility estimate of the aggregate claims / number of claims) for a risk is:

$$Z\bar{X} + (1 - Z)\mu$$

Where Z is a number between zero and one and is known as the **credibility factor**. The attractive features of the credibility formula are its simplicity and, provided $\bar{X}$ and μ are obviously reasonable alternatives, the ease with which it can be explained to senior management. The credibility factor Z is just a weighting factor. Its value reflects how much "trust" is placed in data from the risk itself $\bar{X}$ compared with the data from the larger group μ as an estimate of next year's expected aggregate claims or number of claims. The higher the value of Z, the more trust is placed in $\bar{X}$ compared with μ and vice versa.

Let's return to our earlier example.

Example 5.11 (Buses)**.** Suppose that data from the particular fleet of buses under consideration had been available for more than just five years. For example, suppose that the estimate of the aggregate claims in the coming year based on data from this fleet itself were £1,600,

as before, but that this is now based on 10 years of data rather than just 5. In this case, the figure of £1,600 is considered more trustworthy than the figure of £2,500 and this means giving the credibility factor a higher value, say 0.75 rather than 0.6. The resulting credibility estimate of the aggregate claims would be £1,825.

Now suppose that the figure of £1,600 is based on just 5 year's data, as before, but the figure of £2,500 is based only on data from bus fleets operating in towns of roughly the same size as the one under consideration. In this case the collateral data would be regarded as more relevant than it was the first time and so the credibility factor would be correspondingly reduced, for example to 0.4 from 0.6, giving a credibility premium of £2,140.

Finally, suppose the situation is exactly as in the original setup but the figure of £2,500 is based only on bus fleets operating in London and Dublin. In this case the collateral data might be regarded as less relevant and so the credibility factor would be correspondingly increased, for example to 0.8 from 0.6 giving a credibility premium of £1,780.

We need to formulate two general conditions that Z should satisfy:

1. The more data there is from the risk itself, the higher should be the value of the credibility factor

2. The more relevant the collateral data, the lower should be the value of the credibility factor.

One final point is that, while the value of the credibility factor should reflect the amount of data available from the risk itself, its value should not depend on the value of $\bar{X}$. If Z were allowed to depend on $\bar{X}$ then any estimate of the aggregate claims / number of claims, say ϕ, taking a value between $\bar{X}$ and μ could be written in the form of $Z\bar{X} + (1 - Z)\mu$ by choosing Z to be equal to

$$\frac{(\phi - \mu)}{(\bar{X} - \mu)}.$$

5.2.1 Bayesian credibility

The Bayesian approach to credibility involves the same steps as Bayesian estimation:

- Start with a **prior distribution** for the unknown parameter under consideration (e.g. : the claim frequency).

- Collect relevant data and use these values to obtain the **likelihood function**.
- We work out the **posterior distribution** using the posterior $\propto$ prior $\times$ likelihood result.
- A loss function is specified to quantify how serious misjudging the parameter value would be.
- The Bayesian estimate of the parameter values is then calculated.

We will consider two models of Bayesian credibility:

1. The Poisson / Gamma model
2. The Normal / Normal model

The Poisson/Gamma model

Suppose the claim frequency for a risk needs to be estimated. The problem can be summarised as follows:

- The number of claims in each year is assumed to have a Poisson distribution with parameter λ.
- The value of λ is not known, but estimates of its value are possible along the lines of, for example, "there is a 25% chance that the value of λ is between 100 and 125".
- More precisely, before having any data available from this risk, we think that λ has a Gamma(α, β) distribution.

Data from this risk is available showing the number of claims arising in each of the past n years.

This problem fits exactly into the framework of Bayesian Statistics and can be summarised as follows:

- The random variable X represents the number of claims in the coming year from a risk
- The distribution of X depends on the fixed, but unknown, value of a parameter, λ

- The conditional distribution of X given λ is Poisson(λ)

- The prior distribution of λ is Gammma(α, β)

- $x_1, x_2, \ldots, x_n$ are past observed values of X (we will denote these as $\underline{x}$)

The problem is to estimate λ given the data $\underline{x}$ and the estimate wanted is the Bayes estimate with respect to quadratic loss, i.e. $E[\lambda|\underline{x}]$. The posterior distribution of λ given $\underline{x}$ is:

$$\text{Gamma}(\alpha + \sum_{i=1}^{n} x_i, \beta + n)$$

and:

$$E[\lambda|\underline{x}] = \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + n}.$$

The observed mean number of claims is

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

and the mean number based on prior beliefs is the mean of the prior gamma distribution $\frac{\alpha}{\beta}$. The expression for the posterior mean can then be written in credibility form using the rearrangement:

$$E[\lambda|\underline{x}] = Z\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] + (1 - Z)\frac{\alpha}{\beta}$$

where:

$$Z = \frac{n}{\beta + n}.$$

If only the data from the risk itself were available to estimate λ, the obvious estimate would be $\frac{\sum_{i=1}^{n} x_i}{n}$ since this is the maximum likelihood estimate of λ. We notice that this estimate, which plays the role of $\bar{X}$ in the credibility premium formula, is a linear function of the observed values $x_1, x_2, \ldots, x_n$. Now suppose that no data was available from the risk itself (so $n = 0$), this means $Z = 0$. The only information available to help estimate λ would be its prior distribution. The best estimate of λ would then be the mean of the prior distribution,

which is $\frac{\alpha}{\beta}$.

The value of Z depends on the amount of data available for the risk, n and the collateral information, through β. As n increases, the sampling error of $\frac{\sum_{i=1}^{n} x_i}{n}$ as an estimate for λ decreases. Similarly β represents the variance of the prior distribution for λ. (The higher β is, the smaller the variance of the prior distribution, and so the more 'reliable' the prior is.) So Z reflects the relative reliability of the two alternative estimates of λ.

Bearing these points in mind, the Bayesian estimate of λ is in the form of a weighted average of the estimate of λ based solely on data from the risk itself and an estimate of λ based on some other information. This is precisely the form of the credibility estimate but with "collateral data" now being given a more precise interpretation as a prior distribution.

The Normal/Normal model

This time we want to estimate the pure premium *ie* the expected aggregate claims, for a risk. Let X be a random variable representing the aggregate claims in the coming year for this risk. We make the following assumptions:

- The distribution of X depends on the fixed, but unknown, value of a parameter, θ
- The conditional distribution of X given θ is $\mathsf{N}(\theta, \sigma_1^2)$
- The uncertainty about the value of θ is modelled in the usual Bayesian way, by regarding it as a random variable.
- The prior distribution of θ is $\mathsf{N}(\mu, \sigma_2^2)$
- The values of μ, σ_1 and σ_2 are known.
- n past values of X have been observed, which will be denoted $x_1, x_2, \ldots, x_n$ (or $\underline{x}$ for short)

If the value of θ was known, the correct pure premium for this risk would be $E[X|\theta]$, which is just θ. The problem, then, is to estimate $E[X|\theta]$ given $\underline{x}$ and, as in the Poisson/Gamma

model, the Bayesian estimate with respect to quadratic loss will be used. This means that the estimate will be

$$E[E(X|\theta)|\underline{x}]$$

which is the same as $E[\theta|\underline{x}]$.

The posterior distribution of θ given $\underline{x}$ is:

$$N\left(\frac{\mu\sigma_1^2 + n\sigma_2^2\bar{x}}{\sigma_1^2 + n\sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + n\sigma_2^2}\right)$$

where:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

So the posterior mean is

$$\begin{aligned} E(\theta|\underline{x}) &= \frac{\mu\sigma_1^2 + n\sigma_2^2\bar{x}}{\sigma_1^2 + n\sigma_2^2} && (5) \\ &= \frac{\sigma_1^2}{\sigma_1^2 + n\sigma_2^2}\mu + \frac{n\sigma_2^2}{\sigma_1^2 + n\sigma_2^2}\bar{x} \\ &= Z\bar{x} + (1 - Z)\mu && (6) \end{aligned}$$

where:

$$Z = \frac{n}{n + \frac{\sigma_1^2}{\sigma_2^2}}$$

This is a credibility estimate of $E[\theta|\underline{x}]$ since it is a weighted average of two estimates: the first, $\bar{x}$, is the maximum likelihood estimate based solely on the data from the risk itself and the second, μ is the best available estimate if no data was available from the risk itself.

We notice that as in the Poisson / Gamma model, the estimate based solely on data from the risk itself is a linear function of the observed data values. There are some further points to be made about the credibility factor, Z in this instance. First of all, it is always between zero and one. Secondly it is an increasing function of n, the amount of data available. Finally, it is an increasing function of σ_2, the standard deviation of the prior distribution. These are exactly the features that we expect from a credibility factor.

The Normal/Normal model can be reformulated as follows:

1. The distribution of X_j depends on the value of a fixed, but unknown, parameter θ.

2. The conditional distribution of X_j given θ is $\mathsf{N}(\theta, \sigma_1^2)$

3. Given θ, the random variables X_j are independent.

4. The prior distribution of θ is $\mathsf{N}(\mu, \sigma_2^2)$.

5. The values of $X_1, X_2, \ldots, X_n$ have already been observed and the expected aggregate claims in the coming i.e. , $(n+1)$th, year, need to be estimated.

It is important to realise that the assumptions and problem outlined here are the same as before – just slightly different notation.

We get some important consequences of the assumptions expressing in this way:

1. Given θ, the random variables X_j are independent and identically distributed (IID).

2. The random variables X_j are (unconditionally) identically distributed.

3. The random variables X_j are not (unconditionally) independent.

Statement 1 is an immediate consequence of assumption 2, which says that given θ, each X_j is $\mathsf{N}(\theta, \sigma_1^2)$. As for statement 2, the unconditional distribution function of each of the X_j can be written as

$$P(X_j \leq y) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma_2^2}\right) \Phi\left(\frac{y-\theta}{\sigma_1}\right) d\theta$$

where Φ is the distribution function or cdf of a standardised normal distribution. This is the same for each j, so statement f is true. The easiest way to see this formula is to use the formula for calculating a marginal probability:

$$P(X_j \leq y) = \int P(X_j \leq y|\theta) f(\theta) d\theta$$

and since $X_j|\theta \sim \mathrm{N}(\theta, \sigma_1^2)$,

$$P(X_j \leq y|\theta) = \Phi\left(\frac{y-\theta}{\sigma_1}\right).$$

and this is the same for each j, so statement B is true.

Statement 3 is not as immediately obvious but can be demonstrated as follows:

$$E(X_1X_2) = E[E(X_1X_2|\theta)] = E[E(X_1|\theta)E(X_2|\theta)] = E(\theta^2) = \mu^2 + \sigma_2^2$$

since $X_j|\theta \sim \mathrm{N}(\theta, \sigma_1^2)$ and $\theta \sim \mathrm{N}(\mu, \sigma_2^2)$. Now if X_1 and X_2 were unconditionally independent:

$$E(X_1X_2) = E(X_1)E(X_2) = E[E(X_1|\theta)]E[E(X_2|\theta)] = E(\theta)E(\theta) = \mu^2.$$

Hence, $E(X_1X_2) \neq E(X_1)E(X_2)$.

The approach used in the Poisson / Gamma and Normal / Normal models is essentially the same. The only important difference is in the distributional assumptions. We can summarise the approach in 4 steps:

1. The problem is stated (for example to determine a claim size or claim number distribution).

   In each case the aim is to estimate some quantity which characterises this distribution, (for example the mean number of claims or the mean claim size).

2. Some model assumptions are then made within a Bayesian framework. For example, it could be assumed that the claim number distribution is Poisson and that the unknown parameter of the Poisson distribution has a gamma distribution with specified parameters.

3. A Bayesian estimate of the particular quantity is derived.

4. This estimate is then shown to be in the form of a credibility estimate. $Z\bar{X} + (1-Z)\mu$.

This approach has been quite successul in the two specific cases shown. What are the drawbacks?

The first difficulty is whether a Bayesian approach is even acceptable. Even if it is, what values should we assign to the prior distribution? Although the Poisson / Gamma model provides a formula for the calculation of the credibility factor it involves the parameter β. How a value for β might be chosen has not been discussed. The Bayesian approach to the choice of

parameter values for a prior distribution is to argue that they summarise the subjective degree of belief about the possible values for the quantity to be estimated. In the case of the Poisson / Gamma model this is the mean claim number, λ.

The second difficulty is that even if the problem fits into a Bayesian framework, the Bayesian approach may not work in the sense that it may not produce an estimate that can be readily rearranged to be in the form of a credibility estimate.

5.2.2 Empirical Bayes credibility theory

Today we will discuss two model of Empirical Bayes Credibility Theory (EBCT). As with the Poisson/Gamma and Normal/Normal models discussed already these two models are used to estimate the "true" claims frequency or risk premium based on the total claim amounts in successive periods. Model 1 gives equal weight to each risk in each year. Model 2 is more sophisticated and takes into account the volume of business written under each risk in each year.

Both approaches use a risk parameter, θ. With the Bayesian approach, the quantity that we wish to estimate **is** θ. For the EBCT models we are looking to estimate a function of θ, usually denoted by $m(\theta)$. Unlike the Bayesian approach, the EBCT models do not assume a specific statistical distribution for θ. Both approaches assume that the conditional variables $X_j|\theta$ are independent and indentically distributed. But, as the EBCT models do not assume a statistical distribution for $X_j|\theta$ we instead develop formulae just using the assumption that the mean $m(\theta)$ and variance $s^2(\theta)$ of $X_j|\theta$ can be expressed as functions of θ. The values of $m(\theta)$ and $s^2(\theta)$ are then estimated by the models. With both approaches, we can express the results for either a claim frequency or risk premium for a given risk using a credibility formula, i.e. in the form of a linear combination of the average derived from past claims and an overall average.

EBCT model 1

EBCT Model 1 can be regarded as a generalisation of the Normal/Normal model.

We want to estimate the pure premium, or possibly the claims frequency for a risk. Let $X_1, X_2, \ldots, X_n$ denote the aggregate claims, or the number of claims, in successive periods

for this risk. A more precise statement is that, having observed the values of $X_1, X_2, \ldots, X_n$, the expected value of X_{n+1} needs to be estimated. From now on, $X_1, X_2, \ldots, X_n$ will be denoted by $\underline{X}$.

Then the X_j for one risk satisfy the **Basic assumption for EBCT Model 1**:

1. The distribution of each X_j depends on a parameter, denoted θ, whose value (which is the same for all the X_j's), which is unknown, is chosen randomly according to some distribution.

2. Given θ, the X_j's are independent and identically distributed.

Consequences of the 2 assumptions above:

1. The random variables X_j are identically distributed.

2. The X_j's are not unconditionally independent.

Next we will introduce some notation: $m(\theta) = E(X_j|\theta)$ and $s^2(\theta) = \mathrm{var}(X_j|\theta)$. The problem will be to estimate $m(\theta)$ given $\underline{X}$. The result for the estimate of $m(\theta)$ given $\underline{x}$ (in other words, the credibility premium formula) under EBCT 1 is given below. The derivation of this result/formula is beyond the scope of this module.

Basic result of EBCT Model 1

The estimate of $m(\theta)$ given $\underline{X}$ given by EBCT Model 1 is:

$$(1 - Z)E[m(\theta)] + Z\bar{X}$$

where:

$$\bar{X} = \sum_{j=1}^{n} \frac{X_j}{n}$$

and:

$$Z = \frac{n}{n + \frac{E[s^2(\theta)]}{var[m(\theta)]}}$$

The result/formula involves $E[m(\theta)]$, var$[m(\theta)]$ and $E[s^2(\theta)]$ so we need to estimate these and to do this we need a further assumption - that data is available from some risks that are

similar, but not identical, to the original risk. This is one of the key differences of EBCT to the pure Bayesian approach. We don't need (collateral) data to estimate parameters in the Bayesian approach but we do in the EBCT approach. We assume that a particular risk is just one of N risks in a collective. A collective just means a collection of different, but related, risks. For simplicity, let's assume that the risk we want is risk number 1 in this collective. We assume that, for each of the past n years, the aggregate claims, or claims frequencies for each of these N risks have been observed.

Let X_{ij} denote the aggregate claims, or number of claims, for Risk Number i, $i = 1, 2, \ldots, N$ in year j, $j = 1, 2, \ldots, n$. The claims can be arranged in an array:

		Year			
		1	2	...	n
Risk	1	X_{11}	X_{12}	...	X_{1n}
Number	2	X_{21}	X_{22}	...	X_{2n}
	...	...	...	...	...
	N	X_{N1}	X_{N2}	...	X_{Nn}

We will now state **Detailed assumptions for EBCT Model 1**:

1. For each i, the distribution of $X_{ij}, j = 1, 2, \ldots, n$ depends on the value of a parameter θ_i which is randomly chosen (but is unknown) and the same for each j.
2. The $X_{ij}|\theta_i$, $j = 1, 2, \ldots, n$ are independent and identically distributed.
3. For $i \neq k$, the pairs (θ_i, X_{ij}) and (θ_k, X_{km}) are independent and identically distributed.

In particular, assumption 3 implies that the θ_i are independent and identically distributed.

We need some more notation so define:

$$\sum_{j=1}^{n} \frac{X_{ij}}{n} = \bar{X}_i$$

and:

$$\sum_{i=1}^{N} \frac{\bar{X}_i}{N} = \left(\sum_{i=1}^{N} \sum_{j=1}^{n} \frac{X_{ij}}{Nn} \right) = \bar{X}.$$

Observe that $\bar{X}_i$ denotes the sample mean of the data from Risk Number i, while

$$\bar{X} = \frac{\bar{X}_1 + \ldots + \bar{X}_N}{N}$$

denotes the sample mean from all of the risks. The credibility estimate of the amount / number of claims for the coming year for Risk Number 1 is:

$$(1 - Z)E[m(\theta)] + Z\bar{X}_1$$

where:

$$\bar{X}_1 = \sum_{j=1}^{n} \frac{X_{1j}}{n}$$

and:

$$Z = \frac{n}{n + \frac{E[s^2(\theta)]}{var[m(\theta)]}}.$$

Each row of our risk table corresponds to a fixed value of θ. Bearing both this and the definitions of $m(\theta_i)$ and $s^2(\theta_i)$ in mind, the obvious estimators for $m(\theta_i)$ and $s^2(\theta_i)$ are:

$$\bar{X}_i$$

and

$$\frac{1}{(n-1)} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2.$$

These are for Risk i in the table. We will need the values for each row "averaged" for our final estimators, which are:

$$\bar{X} \text{ to estimate } E[m(\theta)]$$

$$\frac{1}{N} \sum_{i=1}^{N} \frac{1}{(n-1)} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2 \text{ to estimate } E[s^2(\theta)]$$

and

$$\frac{1}{(N-1)}\sum_{i=1}^{N}(\bar{X}_i-\bar{X})^2-\frac{1}{Nn}\sum_{i=1}^{N}\frac{1}{(n-1)}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2 \text{ to estimate } \text{var}[m(\theta)].$$

Example 5.12. The following data represents the total claim amounts per year, X_{ij}, over a six-year period $(n=6)$ for five fleets of buses $(N=5)$, A to E:

		Year, j					
		2005	**2006**	**2007**	**2008**	**2009**	**2010**
	A	1,250	980	1,800	2,040	1,000	1,180
	B	1,700	3,080	1,700	2,820	5,760	3,480
Fleet,i	**C**	2,050	3,560	2,800	1,600	4,200	2,650
	D	4,690	4,370	4,800	9,070	3,770	5,250
	E	7,150	3,480	5,010	4,810	8,740	7,260

Problem: Calculate the credibility premium (i.e. the credibility estimate of total claim amounts) for each fleet for next year.

Solution:

The following functions can be determined from the data:

		$\sum_{j=1}^{6}\frac{X_{ij}}{n}=\bar{X}_i$	$\sum_{j=1}^{6}(X_{ij}-\bar{X}_i)^2$	$(\bar{X}_i-\bar{X})^2$
	A	1,375	973,150	5,569,600
	B	3,090	11,218,200	416,025
Fleet, i	**C**	2,810	4,562,000	855,625
	D	5,325	18,039,550	2,528,100
	E	6,075	19,130,550	5,475,600
Total		18,675	53,923,450	14,844,950

Now:

$$E[m(\theta)] = \bar{X} = \frac{\sum_{i=1}^{N} \bar{X}_i}{N} = \frac{18,675}{5} = 3,735$$

$$E[s^2(\theta)] = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{(n-1)} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$$

$$= \frac{1}{25} \times 53,923,450 = 2,156,938$$

$$V[m(\theta)] = \frac{1}{(N-1)} \sum_{i=1}^{N} (\bar{X}_i - \bar{X})^2 - \frac{1}{Nn} \sum_{i=1}^{N} \frac{1}{(n-1)} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} (\bar{X}_i - \bar{X})^2 - \frac{1}{n} E[s^2(\theta)]$$

$$= \frac{1}{4} \times 14,844,950 - \frac{1}{6} \times 2,156,938 = 3,351,748$$

Therefore:

$$Z = \frac{n}{n + \frac{E[s^2(\theta)]}{V[m(\theta)]}} = 0.90313$$

For Fleet A, *ie* Risk 1 in the table, the credibility premium would be calculated as:

$$(1 - Z)E[m(\theta)] + Z\bar{X}_1$$

$$= 0.09687 \times 3,735 + 0.90313 \times 1,375 = 1,604$$

and the credibility premiums, for fleets B to E, calculated in a similar manner are:

Fleet	B	C	D	E
Premium	3,152	2,900	5,171	5,848

EBCT model 2

This time the techniques of Empirical Bayes Credibility will be applied to a second, more complicated, model. The problem will be the same as the first model, to estimate either the pure premium or the expected number of claims in the coming year, for a risk. The assumptions

will, however, be slightly different. We will refer to this model as EBCT Model 2.

Let $Y_1, Y_2, \ldots, Y_n$ be random variables representing the aggregate claims or number of claims in successive years for a given risk. It is assumed that the values of $Y_1, Y_2, \ldots, Y_n$ have already been observed and the expected value of Y_{n+1} needs to be estimated. So far the problem looks exactly like EBCT Model 1 but the important difference between EBCT Model 1 and EBCT Model 2 is that Model 2 invoves an extra paramter, known as the risk volume, denoted P_j. Intuitively, the value of P_j measures the "amount of business" in year j. For example P_j might represent the premium income for the risk in year j or the number of separate policies comprising the risk in year j. An important point to note is that the value of P_{n+1} at the start of year $n+1$ is assumed to be known.

Next a sequence of random variables $X_1, X_2, \ldots,$ is defined as follows:

$$X_j = \frac{Y_j}{P_j}, j = 1, 2, \ldots$$

The random variable X_j represents the aggregate claims, or the number of claims, in year j standardised to remove the effect of different levels of business in different years.

The assumptions that specify EBCT Model 2 are as follows.

1. The distribution of each X_j depends on the value of a parameter, θ, whose value is the same for each j but is unknown.

2. Given θ, the X_j's are independent (but not necessarily identically distributed).

3. $E(X_j|\theta)$does not depend on j.

4. $P_j \mathrm{var}(X_j|\theta)$ does not depend on j.

Having made these assumptions we can now define $m(\theta)$ and $s^2(\theta)$ as follows:

$$m(\theta) = E(X_j|\theta)$$

$$s^2(\theta) = P_j \mathrm{var}(X_j|\theta)$$

Note that if all the P_j's are equal to 1, EBCT Model 2 is exactly the same as EBCT Model

1. In EBCT Model 2, the definition of $m(\theta)$ is the exact same as in EBCT Model 1. The definition of $s^2(\theta)$ is, however, slightly different.

To gain a bit more insight consider the following example. Suppose that the risk being considered is made up of a different number of independent policies each year and that the number of policies in year j is P_j. Suppose also that the aggregate claims in a single year from a single policy have mean $m(\theta)$ and variance $s^2(\theta)$ where θ is a fixed, but unknown, risk parameter for all these policies. Finally, let Y_j denote the aggregate claims from all the policies in force in year j. Then we have:

$$E(Y_j) = P_j m(\theta)$$

$$\text{var}(Y_j) = P_j s^2(\theta)$$

$$E(X_j) = m(\theta)$$

$$P_j \text{var}(X_j) = s^2(\theta)$$

And this example satisfies assumptions 3 and 4.

The result for the estimate of $m(\theta)$ given $\underline{X}$ (in other words, the credibility premium formula) under EBCT 2 is given below. The derivation of this result/formula is beyond the scope of this module. The estimate of $m(\theta)$ given $\underline{X}$ given by EBCT Model 2 is:

$$Z\bar{X} + (1 - Z)E[m(\theta)]$$

Where:

$$\bar{X} = \frac{\sum_{j=1}^{n} P_j X_j}{\sum_{j=1}^{n} P_j} = \frac{\sum_{j=1}^{n} Y_j}{\sum_{j=1}^{n} P_j}$$

and:

$$Z = \frac{\sum_{j=1}^{n} P_j}{\sum_{j=1}^{n} P_j + \frac{E[s^2(\theta)]}{\text{var}[m(\theta)]}}$$

The procedure for estimating the parameters $E[m(\theta)], \text{var}[m(\theta)], E[s^2(\theta)]$ for EBCT Model 2 follows exactly the same steps as the procedure for EBCT Model 1. It is now assumed that the risk of interest is one of a collective of N risks and there exists data for each of these risks

for each of the past n years. The data consists of values for the aggregate claims, or number of claims and the corresponding risk volumes. Let Y_{ij} be a random variable denoting the aggregate claims, or the number of claims, for risk number i in year j, $j = 1, 2, \ldots, n, i = 1, 2, \ldots, N$, and let P_{ij} be the corresponding risk volume. For each i and j define:

$$X_{ij} = \frac{Y_{ij}}{P_{ij}}$$

The data are summarised in the table below, which corresponds to the risk table for EBCT Model 1.

		Year			
		1	2	...	n
Risk	1	Y_{11}, P_{11}	Y_{12}, P_{12}	...	Y_{1n}, P_{1n}
Number	2	Y_{21}, P_{21}	Y_{22}, P_{22}	...	Y_{2n}, P_{2n}
	...	...	...	...	...
	N	Y_{N1}, P_{N1}	Y_{N2}, P_{N2}	...	Y_{Nn}, P_{Nn}

Define the notation:

$$\sum_{j=1}^{n} P_{ij} \qquad \text{by } \bar{P}_i$$

$$\sum_{i=1}^{N} \bar{P}_i \qquad \text{by } \bar{P}$$

$$\frac{1}{(Nn-1)} \sum_{i=1}^{N} \bar{P}_i \left(1 - \frac{\bar{P}_i}{\bar{P}}\right) \qquad \text{by } P^*$$

$$\sum_{j=1}^{n} \frac{P_{ij} X_{ij}}{\bar{P}_i} = \frac{\sum_{j=1}^{n} Y_{ij}}{\sum_{j=1}^{n} P_{ij}} \qquad \text{by } \bar{X}_i$$

$$\sum_{i=1}^{N} \frac{\bar{P}_i \bar{X}_i}{\bar{P}} = \sum_{i=1}^{N} \sum_{j=1}^{n} \frac{P_{ij} X_{ij}}{\bar{P}} \qquad \text{by } \bar{X}$$

It is important to notice that $\bar{X}_i$ and $\bar{X}$ are weighted averages of the X_{ij}'s with the weights being the risk volumes P_{ij}.

Results of EBCT Model 2:

The credibility estimate of the pure premium or number of claims, per unit of risk volume for the coming year for Risk number 1 in the collective can be written as:

$$Z_1\bar{X}_1 + (1 - Z_1)E[m(\theta)]$$

where:

$$\bar{X}_i = \frac{\sum_{j=1}^{n} P_{ij}X_{ij}}{\sum_{j=1}^{n} P_{ij}} = \frac{\sum_{j=1}^{n} Y_{ij}}{\sum_{j=1}^{n} P_{ij}}$$

$$Z_1 = \frac{\sum_{j=1}^{n} P_{1j}}{\sum_{j=1}^{n} P_{1j} + \frac{E[s^2(\theta)]}{var[m(\theta)]}}$$

and estimators for $E[m(\theta)], E[s^2(\theta)]$ and $\mathrm{var}[m(\theta)]$ are given by:

$$\bar{X} \text{ to estimate } E[m(\theta)]$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{(n-1)}\sum_{j=1}^{n} P_{ij}(X_{ij} - \bar{X}_i)^2 \text{ to estimate } E[s^2(\theta)]$$

$$\frac{1}{P^*}\left[\frac{1}{Nn-1}\sum_{i=1}^{N}\sum_{j=1}^{n} P_{ij}(X_{ij} - \bar{X})^2 - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(n-1)}\sum_{j=1}^{n} P_{ij}(X_{ij} - \bar{X}_i)^2\right] \text{ to estimate } \mathrm{var}[m(\theta)].$$

Example 5.13. Let's look at how to calculate a credibility premium using Model 2. The table here shows the same claim amounts as in our earlier EBCT Model 1 example. These are denoted Y_{ij}. However, the table also shows the corresponding number of buses, P_{ij} over a six-year period $(n = 6)$ for the five fleets of buses $(N = 5)$, A to E. Unlike EBCT Model 1, we now take into consideration the risk volume, which is represented by the number of buses in each fleet for each year.

		2005	2006	2007	2008	2009	2010
	A	1,250 ; 5	980 ; 5	1,800 ; 4	2,040 ; 6	1,000 ; 5	1,180 ; 5
	B	1,700 ; 11	3,080 ; 13	1,700 ; 10	2,820 ; 12	5,760 ; 15	3,480 ; 14
Fleet	**C**	2,050 ; 3	3,560 ; 4	2,800 ; 4	1,600 ; 3	4,200 ; 3	2,650 ; 2
	D	4,690 ; 9	4,370 ; 9	4,800 ; 8	9,070 ; 8	3,770 ; 9	5,250 ; 10
	E	7,150 ; 7	3,480 ; 7	5,010 ; 8	4,810 ; 8	8,740 ; 9	7,260 ; 10

Problem:

1. Calculate the credibility premiums per unit of risk volume for each fleet for 2011.

2. Suppose the risk volumes for 2011 are

	A	B	C	D	E
Risk Volume	6	15	4	10	10

How would you calculate the credibility premiums for each fleet for 2011?

Answer to 1:

The following functions can be determined from the data:

		$\sum_{j=1}^{6} P_{ij}X_{ij}$	$\sum_{j=1}^{6} P_{ij} = \bar{P}_i$	$\bar{X}_i$	$\sum_{j=1}^{6} P_{ij}(X_{ij} - \bar{X}_i)^2$	$\sum_{j=1}^{6} P_{ij}(X_{ij} - \bar{X})^2$
	A	8,250	30	275.0	217,910	1,680,442
	B	18,540	75	247.2	437,931	5,072,946
Fleet	**C**	16,860	19	887.4	1,812,785	4,726,028
	D	31,950	53	602.8	2,804,038	3,411,218
	E	36,450	49	743.9	1,706,731	4,722,398
	Total	112,050	226	$\bar{X}$ =495	6,979,395	19,613,032

		$\bar{P}_i\left(1-\frac{\bar{P}_i}{\sum_{i=1}^{5}\bar{P}_i}\right)$
	A	26.018
	B	50.111
Fleet	**C**	17.403
	D	40.571
	E	38.376
Total		172.478

Note that $X_{ij} = \frac{Y_{ij}}{P_{ij}}$ We have:

$$E[m(\theta)] = \bar{X} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n} P_{ij}X_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{n} P_{ij}} = \frac{112,050}{226} = 495.796$$

$$E[s^2(\theta)] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(n-1)}\sum_{j=1}^{n} P_{ij}(X_{ij}-\bar{X}_i)^2 = \frac{6,979,395}{25} = 279,175.8$$

$$\bar{P} = \sum_{i=1}^{N}\bar{P}_i = 226, \quad P^* = \frac{1}{(Nn-1)}\sum_{i=1}^{N}\bar{P}_i\left(1-\frac{\bar{P}_i}{\bar{P}}\right) = 5.9475$$

$$\text{var}[m(\theta)] = \frac{1}{P^*}\left[\frac{1}{Nn-1}\sum_{i=1}^{N}\sum_{j=1}^{n} P_{ij}(\bar{X_{ij}}-\bar{X})^2 - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(n-1)}\sum_{j=1}^{n} P_{ij}(X_{ij}-\bar{X}_i)^2\right]$$

$$= \frac{1}{P^*}\left(\frac{1}{Nn-1}\sum_{i=1}^{N}\sum_{j=1}^{n} P_{ij}(\bar{X_{ij}}-\bar{X})^2 - E[s^2(\theta)]\right)$$

$$= \frac{\frac{19,613,032}{29} - 279,175.8}{5.9475}$$

$$= 66,773.4$$

Therefore:

$$Z_i = \frac{\sum_{j=1}^{n} P_{ij}}{\sum_{j=1}^{n} P_{ij} + \frac{E[s^2(\theta)]}{V[m(\theta)]}}$$

will be different for each fleet. For Fleet A (Risk 1) we get:

$$Z_1 = \frac{30}{30 + \frac{279,175.8}{66,774}} = 0.87768$$

We can calculate the credibility factors, Z_i for fleets B to E in a similar manner to get:

Fleet	B	C	D	E
Credibility Factor	0.94720	0.81964	0.92688	0.92138

So for Fleet A (Risk 1), the credibility premium per unit of risk volume would then be determined as:

$$0.12232 \times 495.796 + 0.87768 \times 275 = 302.0$$

And then the credibility premiums per unit of risk volume, for fleets B to E can be calculated in a similar manner:

Fleet	A	B	C	D	E
Credibility Premium per unit risk	302.0	260.3	816.7	595.0	724.4

Answer to 2:

Multiplying these figures by the risk volumes for 2011 would then give the credibility premiums for 2011:

Fleet	A	B	C	D	E
Credibility Premium	1812.04	3904.90	3266.98	5950.04	7243.74

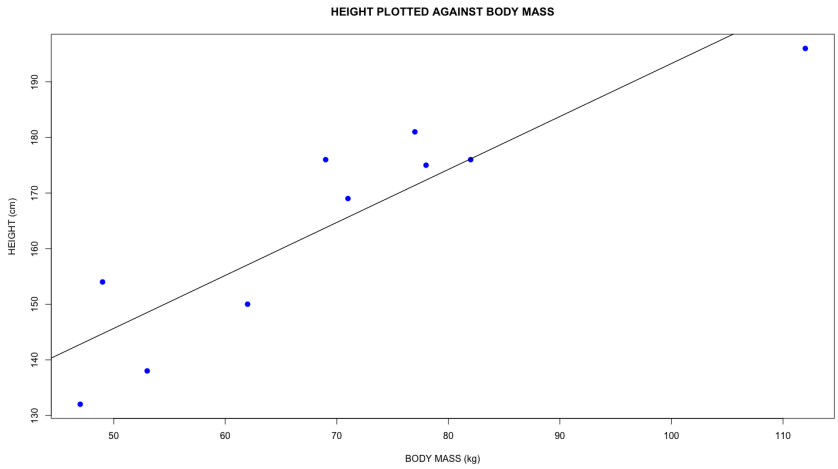
6 Generalised Linear Models

A Generalised Linear Model (GLM) may be regarded as an extension of a linear regression model. In particular, the linear regression model is a simple GLM and many of the ideas of regression modelling are also used in GLMs.

A linear regression takes the form:

$$Y = \beta_0 + \beta_1 x$$

Where β_0 is the intercept on the Y axis and β_1 is the gradient. We choose the line to minimise the sum of the squared error terms where the error terms are the distances from the data points to the straight line. We also assume that the error terms have a $\mathrm{N}(0, \sigma^2)$ distribution.



The essential difference for GLMs is that we now allow the distribution of the data to be non-normal. This is particularly important in actuarial work where the data very often does not have a normal distribution. For example, in mortality, the Poisson distribution is used to model the force of mortality μ_x and the binomial distribution is used for the initial rate of mortality q_x. In general insurance, the Poisson distribution is often used for modelling the claim frequency and the gamma or lognormal distribution for the claim severity.

The aims of a data analysis exercise are usually to decide which variables or factors are important predictors for the risk being considered, and then to quantify the relationship between these predictors and the risk in order to assess appropriate premium levels. For example, in motor insurance, there are a large number of factors that may be used as proxies for the level of risk:

- type of car driven
- age of driver
- number of years of driving experience

We can use a Generalised Linear Model to determine which of these factors are significant to the assessment of the risk and to suggest an appropriate premium to charge for a risk that represents a particular combination of these factors.

GLM techniques are also widely used in life insurance. Some rating factors that an insurance company may consider in the pricing of a single life annuity contract are:

- age
- gender (if permitted by legislation)
- size of single premium
- postcode
- health status

GLMs relate a variable, called the **reponse variable**, which you want to predict, to variables or factors, called **predictors, covariates or independent variables**, about which you have information. In defining a GLM, the first step is to define the distribution of the response variable. The general form of possible response variable distribution used in Generalised Linear Models are called **exponential families** of distributions. We also need to write down an appropriate **linear predictor** which is a function of the covariates and which is linear in unknown constants which are to be fitted. The means of the response variables are made to equal a function of the linear predictor determined by something called the link. The means of the response variables will determine their distributions. A GLM is therefore determined by three things:

1. the distribution of the response variables
2. the linear predictor
3. the link

We will consider these three components in succession.

6.1 Exponential families of distributions

The exponential family of distributions includes the following distributions:

- The Normal distribution

- The Poisson distribution
- The Binomial distribution
- The Gamma distribution
- The Exponential distribution

Definition 6.1. A distribution for a random variable Y belongs to an exponential family if its density has the following form:

$$f_Y(y;\theta;\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right]$$

where a,b and c are functions and θ and ϕ are parameters.

The parameter θ, called the **natural parameter**, is the one which is relevant for relating the response (Y) to the covariates. The parameter ϕ is known as the **scale parameter** or **dispersion parameter**. It can be shown that the mean and variance of Y are:

$$E(Y) = b'(\theta) \tag{7}$$

and

$$\mathrm{var}(Y) = a(\phi)b''(\theta). \tag{8}$$

It follow that θ is a function of $\mu = E[Y]$ only and that in general the mean does not depend on ϕ, so when predicting Y it is θ that is of importance. Also, the variance of the data has two components: one that involves the scale parameter and the other that determines the way the variance depends on the mean. Where a distribution has just one parameter, such as $\mathrm{Poisson}(\lambda)$, we take $\phi = 1$.

Example 6.2 (Normal)**.** We wrote the normal p.d.f. as

$$f_Y(y;\theta;\phi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right]$$

$$= \exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]$$

This is in the form of an exponential family, with $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \frac{\theta^2}{2}$, and $c(y, \phi) = \frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$.

Thus, the natural parameter for the normal distrubtion is μ and the scale parameter is σ^2. Note that we could equally have taken $\phi = \sigma$ and $a(\phi) = \sigma^2$. There is no unique parameterisation.

Applying (7) and (8) produce

$$b(\theta) = \frac{\theta^2}{2} \implies E(Y) = b'(\theta) = \theta = \mu$$

$$a(\phi) = \phi \implies \text{var}(Y) = a(\phi)b''(\theta) = \phi = \sigma^2.$$

These are the results we would expect for the normal distribution.

As was noted above, the natural and dispersion parameters are not uniquely defined. We will show that if we re-parameterise the normal distribution using $\theta = 2\mu$, we still get the same results for the mean and variance of the distribution. If we put $\theta = 2\mu$, we get the following expressions for the various functions: $a(\phi) = 2\phi$, $b(\theta) = \frac{\theta^2}{4}$, and $c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$. Using (7) and (8) results in

$$E(Y) = b'(\theta) = \frac{2\theta}{4} = \frac{2 \times 2\mu}{4} = \mu$$

and

$$\text{var}(Y) = b''(\theta)a(\phi) = \frac{1}{2} \times 2\phi = \phi = \sigma^2$$

So the mean and variance are μ and σ^2, as before.

For the normal distribution, the variance of Y does not depend on the mean (since $b''(\theta) = 1$). However, for other distributions we will see that the variance *does* depend on the mean.

Example 6.3 (Poisson)**.** We write the Poisson p.d.f. as

$$f_Y(y; \theta; \phi) = \frac{\mu^y e^{-\mu}}{y!} = \exp\left(y \log \mu - \mu - \log(y!)\right)$$

This is in the form of an exponential family with: $\theta = \log \mu$, $\phi = 1$, $a(\phi) = 1$, $b(\theta) = e^{\theta}$, and $c(y, \phi) = -\log(y!)$.

Thus, the natural parameter for the Poisson distribution is $log\mu$, the mean is :

$$E(Y) = b'(\theta) = e^{\theta} = \mu$$

and the variance function is:

$$V(\mu) = b''(\theta) = e^{\theta} = \mu$$

The variance function tells us that the variance is proportional to the mean. We can see that the variance is actually *equal* to the mean since $a(\phi) = 1$.

Example 6.4 (Binomial)**.** This is slightly more awkward to deal with, since we have to first divide the binomial random variable by n. Thus, suppose $Z \sim \text{Binomial}(n, \mu)$. Let

$$Y = \frac{Z}{n}$$

so that

$$Z = nY.$$

The distribution of Z is:

$$f_Z(z; \theta, \phi) = \binom{n}{z} \mu^z (1-\mu)^{n-z}$$

and by substituting for z, the distribution of Y is:

$$\begin{aligned} f_Y(y; \theta, \phi) &= \binom{n}{ny} \mu^{ny} (1-\mu)^{n-ny} \\ &= \exp\left[n\left(y \log \mu + (1-y)\log(1-\mu)\right) + \log \binom{n}{ny} \right] \\ &= \exp\left[n\left(y \log(\frac{\mu}{1-\mu}) + \log(1-\mu) \right) + \log \binom{n}{ny} \right]. \end{aligned}$$

This is in the form of an exponential family with: $\theta = \log(\frac{\mu}{1-\mu}) \Rightarrow \mu = \frac{e^{\theta}}{1+e^{\theta}})$, $\phi = n$, $a(\phi) = \frac{1}{\phi}$, $b(\theta) = \log(1 + e^{\theta})$, and $c(y, \phi) = \log \binom{n}{ny}$.

The reason for transforming to Y that in an exponential family θ is a function of μ, the

distribution mean, only. However, the binomial distribution as typically quote it: Bin(n, p) does not have μ as one of its parameters. So we start by considering Bin(n, μ), which does have μ as a parameter, but has mean $n\mu$. We then divide this by n to get a distribution with μ in its probability function and which also has mean μ. Note that $\phi = n$, the "other" parameter in the distribution i.e. the parameter other than the mean).

The natural parameter for the binomial distribution is $\log\left(\frac{\mu}{1-\mu}\right)$, the mean is:

$$E[Y] = b'(\theta) = \frac{e^\theta}{1+e^\theta} = \mu$$

and the variance function is:

$$V(\mu) = b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2} = \mu(1-\mu)$$

Example 6.5 (Gamma)**.** The best way to consider the Gamma distribution is to change the parameters from α and λ to α and $\mu = \frac{\alpha}{\lambda}$, ie $\lambda = \frac{\alpha}{\mu}$. Then

$$\begin{aligned} f_Y(y;\theta,\phi) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} \\ &= \frac{\alpha^\alpha}{\mu^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y\alpha/\mu} \\ &= \exp\left[\left(-\frac{y}{\mu} - \log\mu\right)\alpha + (\alpha-1)\log y + \alpha\log\alpha - \log\Gamma(\alpha)\right]. \end{aligned}$$

This is in exponential family form with: $\theta = -\frac{1}{\mu}$, $\phi = \alpha$, $a(\phi) = \frac{1}{\phi}$, $b(\theta) = -\log(-\theta)$, and $c(y,\phi) = (\phi-1)\log y + \phi\log\phi - \log\Gamma(\phi)$.

Thus, the natural parameter for the gamma distribution is $\frac{1}{\mu}$, ignoring the minus sign. The mean is $E[Y] = b'(\theta) = -\frac{1}{\theta} = \mu$ The variance function is:

$$V(\mu) = b''(\theta) = \frac{1}{\theta^2} = \mu^2$$

and so the variance is:

$$\frac{\mu^2}{\alpha}.$$

6.1.1 Setting the distribution in R

To set the distribution in R, use

```
family=gaussian
family=poisson


and so on ...
```

6.2 Linear predictors

The second component of a GLM is the linear predictor, η, which is a function of the covariates, ie the input variables to the model.

The covariates (also known as explanatory, predictor or independent variables), enter the model through the linear predictor. This is also where the parameters occur which have to be estimated. The requirement is that it is linear in the *parameters* that we are estimating.

There are two kinds of covariates used in GLMs: variables and factors.

1. variables, for which the actual value of the variable enters the linear predictor

2. factors there is a parameter for each value that the factor may take

6.2.1 Variables

In general, variables are covariates where the actual value of a variable enters the linear predictor. The age of the policyholder is an actuarial example of a variable. So far in our linear models we have only met continuous variables. A variable is a type of covariate whose real numerical value enters the linear predictor directly, such as age. Other examples of variables in a car insurance context are annual mileage and number of years for which a driving licence has been held.

The bivariate linear model had a single continuous explanatory variable x with a linear predictor of $\beta_0 + \beta_1 x$. To fit this model it is necessary to estimate the parameters β_0 and β_1. For the multivariate linear regression model with k continuous main effect variables the linear predictor is $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.

We can extend our models to include polynomials, to functions of the variable and to linear predictors including more than one variable. Recall that the linear predictor is linear in the parameters and not necessarily linear in the covariates. For example, $\beta_0 + \beta_1 x^2$ is a linear predictor. You may want to add interactions between covariates, such as $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. When an interaction term is used in a model, both main effects must also be included. Otherwise we are saying that the variables don't contribute anything independently.

Some examples, where age (x_1) and duration (x_2) are treated as variables, are shown in the table below together with the formula used in the glm function in R.

Model	Linear Predictor	R Formula
1(null model)	β_0	$Y \sim 1$
age	$\beta_0 + \beta_1 x_1$	$Y \sim X1$
age^2	$\beta_0 + \beta_1 x_1^2$	$Y \sim I(X1 \wedge 2)$
age + age^2	$\beta_0 + \beta_1 x_1 + \beta_1 x_1^2$	$Y \sim X1 + I(X1 \wedge 2)$
age + duration	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$Y \sim X1 + X2$
log(age)	$\beta_0 + \beta_1 \log(x_1)$	$Y \sim \log(X1)$
age + duration + age.duration	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$Y \sim X12 + X2 + X1 : X2$
age $*$ duration	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$Y \sim X1 * X2$

The null model has no covariates and so there is just the intercept parameter. This is estimated as the sample mean of the response values.

It's fairly easy to see that we start with an intercept parameter and then add a new term with a 'slope' parameter multiplied by the covariate. However, there is actually a little more happening before we get to this simplified linear predictor.

Suppose the linear predictor for age only is $\alpha_1 + \beta_1 x_1$ and the linear predictor for duration only is $\alpha_2 + \beta_2 x_2$. We could then obtain a linear predictor for both of these covariates by summing their individual linear predictors:

$$(\alpha_1 + \beta_1 x_1) + (\alpha_2 + \beta_2 x_2) = (\alpha_1 + \alpha_2) + \beta_1 x_1 + \beta_2 x_2.$$

The final simplified version given in the table above, $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, combines the two constants together, ie $\alpha_1 + \alpha_2$. This simplified formula gives the same final values as the uncombined formula and is more efficient as it requires us to estimate three rather than four parameters. However, it can be shown by example that it is actually impossible to estimate α_1 and α_2 individually from any given data and hence we have to combine them in the linear predictor to overcome this issue.

The two models in the table above are equivalent, and have been shown separately to illustrate the use of the dot and star model notation in R.

An interaction term is denoted using dot notation. In the example above, 'age.duration' denotes the interaction between age and duration (although in R a colon is used to prevent confusion with a decimal point).

The star notation is used to denote the main effects and the interaction term. In the example above, age*duration = age + duration + age.duration.

6.2.2 Factors

The other main type of covariate is a factor, which takes a categorical value. For example, the sex of the policyholder is either male or female, which constitutes a factor with two categories (or levels).

Other examples of factors in a car insurance context are postcode and car type.

This type of covariate can be parameterised so that the linear predictor has a term α_1 for a male, and a term α_2 for a female. In general, there is parameter for each level that the factor may take.

Examples of factors are male/female, smoker/non-smoker etc. For example, the linear predictor may have a term α_1 for a male, and a term α_2 for a female (ie α_i where $i = 1$ for a male and $i = 2$ for a female. In general, there is parameter for each level that the factor may take.

Pairs of these two types enter linear predictors in different ways: If there is more than one factor then we may need to include an interaction term which is another parameter that describes how the effect of one factor depends on the level of the other factor. In that case, If α_i and β_j are factors, the model needs to have additional term $\gamma_{i,j}$.

In the following sex and vrg are factors.

Model	Linear Predictor	R Formula
sex	α_i	$Y \sim \text{sex}$
vehicle rating group	β_j	$Y \sim \text{vrg}$
sex + vehicle rating group	$\alpha_i + \beta_j$	$Y \sim \text{sex} + \text{vrg}$
sex + vehicle rating group +sex.vehicle rating group	$\alpha_i + \beta_j + \gamma_{ij}$	$Y \sim \text{sex} + \text{vrg} + \text{sex} : \text{vrg}$
sex * vehicle rating group	$\alpha_i + \beta_j + \gamma_{ij}$	$Y \sim \text{sex} * \text{vrg}$

The last two models are identical.

It is impossible to estimate all of the α_i and β_j individually from any given data set, and so we have to combine constants together to overcome this issue. This means that one of those constants effectively becomes zero. This is called the base assumption in the model.

We can have models with both variables and factors, as is shown in the following table, where age is a variable and sex is a factor.

Model	Linear predictor	Rformula
age	$\beta_0 + \beta_1 x$	$Y \sim X1$
sex	α_i	$Y \sim \text{sex}$
age + sex	$\alpha_i + \beta x$	$Y \sim X1 + \text{sex}$
age + sex + age.sex	$\alpha_i + \beta_i x$	$Y \sim X1 + \text{sex} + X1 : \text{sex}$
age * sex	$\alpha_i + \beta_i x$	$Y \sim X1 * \text{sex}$

The last two models are identical.

6.3 Links

Let η denote the predictor. The link function connects the mean response to the linear predictor: $g(\mu) = \eta$ where $\mu = E(Y)$. To fit a GL Model, the link function must be differentiable and invertible. We then have $\mu = g^{-1}(\eta)$ and the parameters of η along with the dispersion parameters and covariates determine the distribution of Y.

The canonical link function (sometimes called the natural link function) is such that $g(\mu) = \theta(\mu)$. This is a good place to start in linear modelling and will often suffice for actuarial applications. Before using more complex link functions you would need to consider the implications on possible values of μ and whether these are reasonable given the data.

Here are some canonical link functions:

Distribution	Canonical Link Function	Name
Normal	$g(\mu) = \mu$	identity
Poisson	$g(\mu) = \log(\mu)$	log
Binomial	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	logit
Gamma	$g(\mu) = \frac{1}{\mu}$	inverse

Earlier, we showed that $\theta = -1/\mu$ for the gamma distribution. The minus sign is dropped in the canonical link function. This doesn't affect anything since constants will be absorbed into the parameters in the linear predictor.

The canonical link functions work well for each of the above distributions, but it is not obligatory that they are used in each case. For example, we could use the identity link function in conjunction with the Poisson distribution, we could use the log link function for data which had a gamma distribution, and so on.

The choice of the link function has implications for the possible values of μ. $\mu = g^{-1}(\eta)$. The choice of the link function on the possible values for μ. For example, if the data have a Poisson distribution then μ must be positive. If we use the log link function, then $\eta = \log(\mu)$ and $\mu = e^{\eta}$. Thus, μ is guaranteed to be positive, whatever value (positive or negative) the linear predictor takes. The same is not true if we use the identity link function.

Example 6.6. The inverse of the link function $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ is determined by by setting it equal to η and solving. We find that

$$\mu = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}}.$$

It is an appropriate link function for the binomial distribution since it results in values of μ,

the probability parameter, between 0 and 1.

6.3.1 Setting the link in R

The default in R is to use the canonical link.

If you want to, you can set

```
link=identity

link=log

link=sqrt


and so on ...
```

6.4 Using glm in R

The R code to fit a generalised linear model to a multivariate data frame and assign it to the object model, is:

```
model <-glm(Y ~ ..., family = ... (link = ... ), data=...)
```

The statement Y ... specifies the linear predictor, e.g. Y∼age+sex. The statement family= specifies the distribution, i.e. family="gaussian", family="binomial", family="poisson", family="Gamma"". The canonical link functions are "identity" , "log" , "logit" , "inverse" for normal, Poisson binomial and gamma distribution, respectively. If you don't specify a link it will use the canonical link. After data= put the name of your data frame. The estimates of the parameters and their approximate standard errors can be obtained by:

```
summary(model)
```

6.5 Model fitting and comparison

The process of choosing a model also uses methods which are approximations, based on maximum likelihood theory, and this section outlines this process.

6.5.1 Obtaining the estimates

The log-likelihood function of a GLM, $\ell(y;\theta,\phi) = \log(f(y;\theta,\phi))$, depends on the parameters α_i in the linear predictor. Those parameters are usually estimated using maximum likelihood estimation.

From the estimated parameters $\hat{\alpha}_i$ we have the estimated linear predictors $\hat{\eta}$ and the fitted mean $\hat{\mu} = g^{-1}(\hat{\eta})$ which approximates Y.

Example 6.7. Claim amounts for medical insurance claims for hamsters are believed to have an exponential distribution with mean μ_i:

$$f(y_i) = \frac{1}{\mu_i} e^{-y_i/\mu_i}.$$

We have the following data for hamsters' medical claims, using the model above:

age x_i (months)	4	8	10	11	17
claim amount($\pounds$)	50	52	119	41	163

The insurer believes that a linear function of age affects the claim amount:

$$\eta_i = \alpha + \beta x_i.$$

The log of the likelihood function is

$$\ell(\mu_i) = -\sum_{i=1}^{7} \frac{y_i}{\mu_i} - \sum_{i=1}^{7} \log(\mu_i).$$

The canonical link function for the exponential distribution is $g(\mu_i) = 1/\mu_i$. Hence we have

$$\frac{1}{\mu_i} = \alpha + \beta x_i.$$

Rearranging gives

$$\mu_i = \frac{1}{\alpha + \beta x_i}.$$

This enables us to write the log-likelihood function in terms of α and β:

$$\ell(\alpha, \beta) = -\sum_{i=1}^{7} y_i(\alpha + \beta x_i) + \sum_{i=1}^{7} \log(\alpha + \beta x_i).$$

We can now differentiate this with respect to α and β:

$$\begin{aligned} \frac{\partial}{\partial \alpha}\ell(\alpha, \beta) &= -\sum_{i=1}^{7} y_i + \sum_{i=1}^{7} \frac{1}{\alpha + \beta x_i} \\ \frac{\partial}{\partial \beta}\ell(\alpha, \beta) &= -\sum_{i=1}^{7} x_i y_i + \sum_{i=1}^{7} \frac{x_i}{\alpha + \beta x_i}. \end{aligned}$$

The equations satisfied by the MLE's of α and β are:

$$-\sum_{i=1}^{7} y_i + \sum_{i=1}^{7} \frac{1}{\hat{\alpha} + \hat{\beta} x_i} = 0$$

and

$$-\sum_{i=1}^{7} x_i y_i + \sum_{i=1}^{7} \frac{x_i}{\hat{\alpha} + \hat{\beta} x_i} = 0.$$

Substituting in the given data values gives the following equations:

$$\frac{1}{\hat{\alpha} + 4\hat{\beta}} + \frac{1}{\hat{\alpha} + 8\hat{\beta}} + \frac{1}{\hat{\alpha} + 10\hat{\beta}} + \frac{1}{\hat{\alpha} + 11\hat{\beta}} + \frac{1}{\hat{\alpha} + 17\hat{\beta}} - 425 = 0$$

and

$$\frac{4}{\hat{\alpha} + 4\hat{\beta}} + \frac{8}{\hat{\alpha} + 8\hat{\beta}} + \frac{10}{\hat{\alpha} + 10\hat{\beta}} + \frac{11}{\hat{\alpha} + 11\hat{\beta}} + \frac{17}{\hat{\alpha} + 17\hat{\beta}} - 5028 = 0$$

These are not particularly easy to solve without computer assistance. Solve the equations gives $\hat{\alpha} = 0.160134$ and $\hat{\beta} = -0.000598$. We can then estimate the mean claim amounts for various ages using

$$\hat{\mu}_i = \frac{1}{\hat{\alpha} + \hat{\beta} x_i}.$$

Doing so gives estimates for the claim amounts of 6.34, 6.44, 6.49, 6.51 and 6.67, which are very poor indeed. So the model does not appear to be appropriate at all. (In R we would fit with the Gamma distribution, which is more general than the Exponential distribution.)

6.5.2 Significance of the parameters

We can test whether each of the parameters is significantly different from zero. Generally speaking, it is not useful to include a parameter β for which we cannot reject the hypothesis that $\beta = 0$.

Approximate standard errors of the parameters can be obtained using asymptotic maximum likelihood theory. Recall that estimators are in general asymptotically normal and unbiased with variance equal to the Cramér-Rao lower bound:

$$\hat{\beta} \sim \mathrm{N}(\beta, \mathrm{CRLB}(\beta)).$$

The variance $\mathrm{CRLB}(\beta)$ may involve the parameters and so we estimate it by substituting $\hat{\beta}$ for β, giving $\widehat{\mathrm{CRLB}(\beta)}$. Hence, when testing $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$, we use:

$$\frac{\hat{\beta} - 0}{\sqrt{\widehat{\mathrm{CRLB}(\beta)}}} \sim \mathrm{N}(0, 1).$$

For a two-tailed test, the theory of hypothesis testing shows that we can reject the null hypothesis when $\left| \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{CRLB}(\beta)}}} \right|$ is larger than 1.96. We could approximate 1.96 by 2 for simplicity. As a rough guide, an indication of the significance of the parameters is given by twice the standard error. Thus, if:

$$|\hat{\beta}| > 2 \times \sqrt{\widehat{\mathrm{CRLB}(\beta)}},$$

then the parameter is significant and should be retained in the model. Otherwise, the parameter is a candidate for being discarded.

It should be noted that in some cases, a parameter may appear to be unnecessary using this criterion, but the model without it does not provide a good enough fit to the data.

In R, the statistic and p-value for the tests of $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ for the different parameters β are given in the output of

```
summary(model)
```

6.5.3 The saturated model

To compare models we need a measure of the fit of a model to the data.

To do this, we compare our model to a model that is a perfect fit to the data. The model that fits perfectly is called the saturated model.

A saturated model is defined to be a model in which there are as many parameters as observations, so that the fitted values $\hat{\mu}_i$ are equal to the observed values y_i.

Example 6.8. Claim amounts for medical insurance claims for hamsters are believed to have an exponential distribution with mean μ_i:

$$f_{Y_i}(y_i) = \frac{1}{\mu_i} e^{-y_i/\mu_i}.$$

The μ_i's are parameters, one for each data point y_i. Suppose that there are n data points. The log-likelihood is

$$\ell(\underline{\mu}) = -\sum_{i=1}^{n} \frac{y_i}{\mu_i} - \sum_{i=1}^{n} \log(\mu_i).$$

We use a factor with n categories for our linear predictor, so

$$\eta_i = \alpha_i, \quad \text{for all } 1 \leq i \leq n.$$

The canonical link for the exponential distribution is $g(\mu) = 1/mu$, so setting $g(\mu_i) = \eta_i$ gives $\mu_i = 1/\alpha_i$. Thus we can write the log-likelihhod in terms of the α_i:

$$\ell(\underline{\alpha_i}) = -\sum_{i=1}^{n} y_i \alpha_i + \log(\alpha_i).$$

Diferentiating gives

$$\frac{\partial}{\partial \alpha_i} \ell(\underline{\alpha_i}) = -y_i + \frac{1}{\alpha_i}.$$

All the terms other than those involving the specific α_i we are looking at disappear. Setting the derivative equal to 0 gives

$$-y_i + \frac{1}{\hat{\alpha}_i} = 0 \Rightarrow \hat{\mu}_i = \frac{1}{\hat{\alpha}_i} = y_i.$$

The fitted values $\hat{\mu}_i$ equal the observed values y_i and the log-likelihood of the saturated model is

$$\ell_S = -\sum_{i=1}^{n} \frac{y_i}{\hat{\mu}_i} - \sum_{i=1}^{n} \log(\hat{\mu}_i) = -n - \sum_{i=1}^{n} \log(y_i)$$

The saturated model is uninformative because it does not summarize the data, but merely repeats them in full. However, it does provide an excellent benchmark against which to compare the fit of other models.

6.5.4 The scaled deviance

In order to assess the adequacy of a model for describing a set of data, we can compare the likelihood under this model with the likelihood under the saturated model.

The saturated model uses the same distribution and link function as the current model, but has as many parameters as there are data points. As such it fits the data perfectly. We can then compare our model to the saturated model to see how good a fit it is. Suppose that L_S and L_M denote the likelihood functions of the saturated and current models, evaluated at their respective optimal parameter values. The inequality $L_M \leq L_S$ generally holds, but if the current model is poor, then L_M should be close to L_S. We write

$$\log\left(\frac{L_S}{L_M}\right) = \ell_S - \ell_M,$$

where ℓ_S, ℓ_M are the respective log-likelihoods. The **scaled deviance** SD_M is defined as twice the difference between the log-likelihood of the model under consideration (known as the current model) and the saturated model:

$$SD_M = 2(\ell_S - \ell_M).$$

The **deviance** D_M is defined to be

$$D_M = SD_M \times \phi,$$

where ϕ is the scale parameter. For the Poisson and exponential distributions, $\phi = 1$, so the

scaled deviance and the deviance are identical.

The scaled deviance has approximately a χ^2_{n-p} distribution where p is the number of parameters in the linear estimator in M.

Example 6.9. Suppose the $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$. The log-likelihood is

$$\ell(\underline{y}; \underline{\mu}, \sigma^2) = \sum_{i=1}^{n} \log f_Y(y_i; \mu_i, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma^2}.$$

For the saturated model, the parameter μ_i is estimated by y_i , and so the second term disappears. Thus, the scaled deviance is

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}$$

where $\hat{\mu}_i$ is the fitted value for the current model. (The dispersion parameter σ^2 can be estimated using the method of moments or assumed known; we will not go into this further, but instead treat it as known.) The deviance (remembering that $\phi = \sigma^2$) is the residual sum of squares:

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2.$$

This is why the deviance is defined with a factor of two in it, so that for the normal model the deviance is equal to the residual sum of squares,

If the μ_i equal a single parameter α, then $\hat{\mu}_i = \overline{y}$ for all i and the scaled deviance is χ^2_{n-1} distributed.

The scaled deviance is displayed under "Residual Deviance" as part of the results from

`summary(model)`

The null model is the model with one parameter. The scaled deviance for the null model is displayed under "Null Deviance" in summary(model).

6.5.5 Using scaled deviance and Akaike's Information Criterion to choose between models

From now on we will only be using the scaled deviance, which may sometimes just be called the deviance.

The smaller the scaled deviance, the better the model from the point of view of model fit. However, there will be a trade-off here. A model with many parameters will fit the data well. However a model with too many parameters will be difficult and complex to build, and will not necessarily lead to better prediction in the future. It is possible for models to be 'over-parameterised', i.e. factors are included that lead to a slightly, but not significantly, better fit. When choosing linear models, we will usually need to strike a balance between a model with too few parameters (which will not take account of factors that have a substantial impact on the data, and will therefore not be sensitive enough) and one with too many parameters (which will be too sensitive to factors that really do not have much effect on the results). We use the principle of parsimony here, i.e. we choose the simplest model that does the job.

Adding more covariates will always improve the fit and thus decrease the deviance, however we need to determine whether adding a particular covariate leads to a significant decrease in the deviance.

We make use of the fact that even in the case of data that are not normally distributed, the deviance is still asymptotically a χ^2 distribution with degrees of freedom equal to $n - p$: the number of data points minus the number of parameters.

Suppose that we want to decide if Model 2 (which has $p+q$ parameters and scaled deviance S_2) is a significant improvement over Model 1 (which has p parameters and scaled deviance S_1. Since $\chi^2_{n-p-q} + \chi^2_q \sim \chi^2_{n-p}$ (provided that the random variables are independent) it makes sense to say that the difference in the scaled deviances $S_1 - S_2$ has a χ^2_q distributions. The common procedure is to compare two models by looking at the difference in the scaled deviance and comparing with a χ^2_q distribution. Thus, if we want to decide if Model 2 (which has $p + q$ parameters and scaled deviance S_2) is a significant improvement over Model 1 (which has p parameters and scaled deviance S_1, we see if $S_1 - S_2$ is greater than the 5% upper point of the χ^2_q distribution.

The code for comparing two non-normally distributed) models, model1 and model2, in R is:

```
anova(model1, model2, test="Chi")
```

A very important point is that this method of comparison can only be used for nested models. In other words, Model 1 must be a submodel of Model 2. Thus, we can compare two models for which the distribution of the data and the link function are the same, but the linear predictor has one extra parameter in Model 2. For example $\beta_0 + \beta_1 x$ and $\beta_0 + \beta_1 x + \beta_2 x^2$. But we could not compare in this way if the distribution of the data or the link function are different, or, for example, when the linear predictors are $\beta_0 + \beta_1 x + \beta_2 x^2$ and $\beta_0 + \beta_3 \log x$. We *can* gauge the importance of factors by examining the scaled deviances, but we cannot use the testing procedure outlined above.

An alternative method of comparing models is to use Akaike's Information Criterion (AIC). Since the deviance will always decrease as more covariates are added to the model, there will always be a tendency to add more covariates. However this will increase the complexity of the model which is generally considered to be undesirable. To take account of the undesirability of increased complexity, computer packages will often quote the AIC, which is a penalised log likelihood:

$$\mathrm{AIC} = -2 \times \log L_M + 2 \times \text{number of parameters}$$

where $\log L_M$ is the log likelihood of the model under consideration.

When comparing two models, the smaller the AIC, the better the fit. So if the change in deviance is more than twice the change in the number of parameters then it would give a smaller AIC.

This is approximately equivalent to checking whether the difference in deviance is greater than the 5% upper point of the χ^2 distribution for degrees of freedom between 5 and 15. However, it has the added advantage of being a simple way to compare GLMs without formal testing.

In R the AIC is displayed as part of the results from

```
summary(model)
```

6.5.6 The process of selecting explanatory variables

The process of selecting the optimal set of covariates for a GLM is not always easy. Again, we could use one of the two following approaches:

Forward selection

Add the covariate that reduces the AIC the most or causes a significant decrease in the deviance. Continue in this way until adding any more causes the AIC to rise or does not lead to a significant improvement in the deviance. Note we should start with main effects before interaction terms and linear terms before polynomial.

Suppose we are modelling the number of claims on a motor insurance portfolio and we have data on the driver's age, sex and vehicle group. We would start with the null model (ie a single constant equal to the sample mean). Then we would try each of single covariate models (linear function of age or the factors sex or vehicle group) to see which produces the most significant improvement in a χ^2 test or reduces the AIC the most. Suppose this was sex. Then we would try adding a second covariate (linear function of age or the factor vehicle group). Suppose this was age. Then we would try adding the third covariate (vehicle group). We might then try a quadratic function of the variable age (and maybe higher powers) or each of 2 term interactions (eg sex*age or sex*group or age*group). Finally we would try the 3 term interaction (ie sex*age*group).

Backward selection

Start by adding all available covariates and interactions. Then remove covariates one by one starting with the least significant until the AIC reaches a minimum or there is no significant improvement in the deviance, and all the remaining covariates have a statistically significant impact on the response.

So with the last example we would start with the 3 term interaction sex*age*group and look at which parameter has the largest p-value (in a test of it being zero) and remove that. We should see a significant improvement in a χ^2 test and the AIC should fall. Then we remove the next parameter with the largest p-value and so on.

6.6 Residuals analysis and assessment of mode; fit

Once a possible model has been found it should be checked by looking at the residuals. The residuals are based on the differences

$$y_i - \hat{\mu}_i \tag{9}$$

between the observed responses, y_i, and the fitted responses, $\hat{\mu}_i$. The fitted responses are obtained by applying the inverse of the link function to the linear predictor with the fitted values of the parameters as in the previous subsection.

We looked at how we could obtain predicted responses values in the previous section. The fitted values are the predicted Y values for the observed data set, x. The R code for obtaining the fitted values of a GLM is:

`fitted(model)`.

The raw residuals (9) are a natural extension of the way we calculated residuals for linear regression models. However, because of the different distributions used, we need to transform these 'raw' residuals so we are able to interpret them meaningfully. There are two kinds of transformed residuals: Pearson and deviance

6.6.1 Pearson residuals

The Pearson residuals are defined as

$$\frac{y_i - \hat{\mu}_i}{\sqrt{\mathrm{var}(\hat{\mu}_i)}}.$$

The $\mathrm{var}(\hat{\mu}_i)$ in the denominator refers to the variance of the response distribution, $\mathrm{var}(Y)$ using the fitted values, $\hat{\mu}_i$, in the formula. For example, since the variance of the exponential distribution is μ_i^2 , we have $\mathrm{var}(\hat{\mu}_i) = \hat{\mu}_i^2$ in that case.

The Pearson residual, which is often used for normally distributed data, has the disadvantage that its distribution is often skewed for non-normal data. This makes the interpretation of residual plots difficult.

The R code for obtaining the Pearson residuals is:

```
residuals(model, type= "pearson")
```

If the data come from a normal distribution, then the Pearson residuals will follow the standard normal distribution. By comparing these residuals to a standard normal (eg by using a Q-Q plot), we can determine whether the model is a good fit. However, for non-normal data the Pearson residuals will not follow the standard normal distribution and won't even be symmetrical. This makes it difficult to determine whether the model is a good fit. Hence we will need to use a different type of residual.

6.6.2 Deviance residuals

Deviance residuals are defined as the product of the sign of $y - \hat{\mu}$ and the square root of the contribution of y to the scaled deviance. Thus, the deviance residual is

$$\text{sign}(y_i - \hat{\mu}_i)d_i$$

where the scaled deviance is written $\sum d_i^2$, where d_i depends on y_i and $\hat{\mu}_i$, and

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0; \\ -1 & \text{if } x < 0. \end{cases}$$

Example 6.10. If $Y \sim \text{N}(\mu_i, \sigma^2)$, then the Pearson residuals are:

$$\frac{y_i - \hat{\mu}_i}{\sqrt{\text{var}(\hat{\mu}_i)}} = \frac{y_i - \hat{\mu}_i}{\sigma}.$$

We saw that the scaled deviance is

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2} = \sum_{i=1}^{n} d_i^2,$$

where

$$d_i = \left| \frac{y_i - \hat{\mu}_i}{\sigma} \right|.$$

So the deviance residuals are given by:

$$\text{sign}(y_i - \hat{\mu}_i) \left| \frac{y_i - \hat{\mu}_i}{\sigma} \right| = \frac{y_i - \hat{\mu}_i}{\sigma}.$$

Hence, the Pearson residuals and the deviance residuals are the same.

The R code for obtaining the deviance residuals is:

```
residuals(model)
```

6.6.3 Using residual plots to check the fit

The assumptions of a GLM require that the residuals should show no patterns. The presence of a pattern implies that something has been missed in the relationship between the predictors and the response. If this is the case, other model specifications should be tried.

So, in addition to the residuals being symmetrical, we would expect no connection between the residuals and the explanatory covariates. Rather than plotting the residuals against each of the covariates, we could just see if there is a pattern when plotted against the fitted values.

We could also plot a histogram of the residuals, or another similar diagnostic plot should also be examined in order to assess whether the distributional assumptions are justified.

6.7 Estimating the response variable

Once we have obtained our model and its estimates, we are then able to calculate the value of the linear predictor, η, and by using the inverse of the link function we can calculate our estimate of the response variable $\hat{\mu} = g^{-1}(\hat{\eta})$.

Substituting the estimated parameters into the linear predictor gives the estimated value of the linear predictor. The link function links the linear predictor to the mean of the distribution. In this way we can obtain an estimate for the mean of the distribution of Y for that individual.

In R we can obtain an estimation by

```
predict(model,newdata,type="response")
```

where 'newdata' contains the data for the variables.