Lecture notes for
STAT0028
Statistical Models and Data Analysis

Dr Giampiero Marra
Department of Statistical Science, UCL

2018-2019

Preliminaries

- This book does NOT contain a complete set of notes for this course.
- These notes will be supplemented by handouts of examples of data analysis that will be given out and discussed in the lectures. The computer output in these examples has been obtained using the R software – you will learn how to obtain R output in the Statistical Computing course.
- If you require further explanation of a method or a proof that is not given or wish to study further examples, then do refer to a textbook on the subject. At the end of each chapter, some references to texts are provided for optional background reading – in general you need only do this from one of the texts listed (see booklist below).
- In the early chapters, references are made to the book by Rice used in the Foundation Course to provide a bridge between the Foundation course and this course, and for some proofs of results.
- To summarise: these notes do not stand alone – they need to be studied with reference to the supplementary examples handouts and, for some sections, you may well find it helpful to refer to a textbook, particularly for further examples.
- You are strongly advised to try the exercises set each week – that is the way to learn the material!

References

Main texts

- Dobson A.J. (2002), An Introduction to Generalized Linear Models. 2nd edition, Chapman & Hall.

  Covers most of the course.

- Krzanowski W.J. (1998), An Introduction to Statistical Modelling. Arnold.

  Covers most of the course.

- Hastie T., Tibshirani R., Friedman J. (2002), The Elements of Statistical Learning. Springer.

  Covers what is not covered by the other two books (e.g. Lasso, trees).

Other texts

- The book by Rice used in the Foundation Course will be referenced in the introductory chapters to provide a bridge between the Foundation Course and this course.
- Garthwaite P.H., Jolliffe I.T., Jones B. (2002), Statistical Inference. 2nd edition, Oxford.
- McCullagh P., Nelder J.A. (1989), Generalized Linear Models. 2nd edition,Chapman & Hall.

  Chapters 1 to 6 and 9 of this book cover much of the course and much more, but is more advanced than the above texts.
- Harrell F.E. jr. (2001), Regression Modeling Strategies. Springer.

  This is a recommendation for further reading on a high level; very sophisticated.
- Other texts are mentioned in some chapters – you may find these useful if you find that you need more detailed sources for particular topics (e.g. for proofs, for summer project).

Data sets referred to in the course

Note: this section provides some of the data sets; objectives (if not stated) and statistical analyses will be described at the appropriate times during the lectures.

Example A

From Freund and Wilson example 2.1, table 2.1.

One task assigned to foresters is to estimate the potential lumber harvest of trees. This is typically done by using a prediction formula from non-destructive measures of the trees. A prediction formula is obtained from a study using a sample of trees for which actual lumber yields were obtained by harvesting. The data below show, for a sample of 20 trees, the values of three non-destructive measures:

DBH, the diameter of the trunk at breast height (about 4 feet), in inches

D16, the diameter of the trunk at 16 feet of height, in inches

HT, the height, in feet

and the measure of yield obtained by harvesting the trees:
VOL, the volume of lumber, in cubic feet.

DBH	D16	HT	VOL	DBH	D16	HT	VOL
10.20	9.3	89.00	25.93	13.78	13.6	89.00	56.20
13.72	12.1	90.07	45.87	15.67	14.0	102.00	66.16
15.43	13.3	95.08	56.20	15.67	13.7	99.00	62.18
14.37	13.4	98.03	58.60	15.98	13.9	89.02	57.01
15.00	14.2	99.00	63.36	16.50	14.9	95.09	65.62
15.02	12.8	91.05	46.35	16.87	14.9	95.02	65.03
15.12	14.0	105.60	68.99	17.26	14.3	91.02	66.74
15.24	13.5	100.80	62.91	17.28	14.3	98.06	73.38
15.24	14.0	94.00	58.13	17.87	16.9	96.01	82.87
15.28	13.8	93.09	59.79	19.13	17.3	101.00	95.71

Example B

Krzanowski example 6.3, table 6.4

To study the effect of volume x_1 and rate x_2 of air inspired by human subjects on the occurrence or not ($Y = 1$ or 0, respectively) of transient vasoconstriction response in the skin of the fingers, 39 observations on these variables were obtained and the following shows just 5 of these observations (the complete set of data is in Krzanowski):

x_1	x_2	Y
3.70	0.83	1
0.90	0.75	0
1.70	1.06	0
1.90	0.95	1
⋮		

Example C

Freund and Wilson example 10.2.

A toxicologist is interested in the effect of a toxic substance on tumour incidence in a particular species of laboratory animals. A sample of animals is exposed to various concentrations of the substance and subsequently examined for the presence or absence of tumours. The data obtained are as follows:

Concentration	0.0	2.1	5.4	8.0	15.0	19.5
Number of animals	50	54	46	51	50	52
Number with tumours	2	5	5	10	40	42

Example D

In an experiment designed to simulate a production operation carried out at different speeds, 15 similarly experienced operatives were randomly divided into 5 groups of 3 and each group was randomly allocated to one of the speeds $x = 1, 2, 3, 4, 5$. Each operative was required to perform a routine task repetitively over a given period of time. The total numbers of mistakes over the 3 runs at each speed were as follows:

Speed	1	2	3	4	5
Number of mistakes	2	7	25	47	121

Example E

In a survey on the attidudes of students in New Jersey towards mathematics, school leavers were asked whether they agreed or disagreed with the statement "I'll need mathematics in my future work". The attitude of women towards mathematics was of particular concern. This question refers only to the responses to the above statement of those female students who intended to take a job on leaving school. The data in the table below shows the observed frequencies classified according to three variables:
R = response to statement (agree or disagree),
A = location of school (suburban or urban),
B = course preference (maths/science or liberal arts).

Location	Course	Response: Agree	Disagree
Suburban	Maths/science	18	13
	Liberal arts	17	12
Urban	Maths/science	3	17
	Liberal arts	7	19

Example F

In an experiment to investigate the effect of nitrogen fertilizer on sugar cane, the yields (per plot) from using various nitrogen levels (kg/hectare) were as follows:

Nitrogen level	0	50	100	150	200
Yield	60	125	152	182	198
	73	144	154	167	188
	77	145	160	181	189
	72	116	141	185	182
Mean	70.50	132.50	151.75	178.75	189.25

Example G

The data are from an investigation into the effects of some compounds on the heat evolved during the hardening of Portland cement (Woods, Steinour and Clark, Industrial and Engineering Chemistry, 1932). The amounts of the compounds (the x-variables below) are expressed as percentages of the weight of the clinkers from which the cement was made.

x_1 = amount of calcium aluminate
x_2 = amount of tricalcium silicate
x_3 = amount of tetracalcium alumino ferrite
x_4 = amount of b-dicalcium silicate
y = heat evolved after 180 days in calories per gram of the cement

x_1	x_2	x_3	x_4	y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

Example H

Collett, example 1.2

An experiment was conducted in order to investigate the effect of time of planting and length of cutting on the mortality of plum root-stocks propagated from root cuttings. For each combination of time of planting and length of cutting, 240 cuttings were planted and the following table shows the number that survive.

	Length of cutting	
Time of planting	Long	Short
Autumn	156	107
Spring	84	31

Example I

Collett, example 1.4

In a toxicological study, 32 female rats were randomly assigned to two groups during pregnancy. One group of 16 rats was fed a diet containing a certain chemical; the other group of 16 rats was a control group that received the same diet without the addition of the chemical. After the birth of the litters, the number of pups that survived the twenty-one day lactation period was recorded: these are shown on the table below as a fraction of those alive four days after birth. An objective was to compare the mortality over the lactation period for the two groups.

Treated rats	13/13	12/12	9/9	9/9	8/8	8/8	12/13	11/12
	9/10	9/10	8/9	11/13	4/5	6/7	7/10	7/10
Control rats	12/12	11/11	10/10	9/9	10/11	9/10	9/10	8/9
	8/9	4/5	7/9	4/7	5/10	3/6	3/10	0/7

Additional references

Freund B.J., Wilson W.J. (1998), Regression Analysis: Statistical Modelling of a Response Variable. Academic Press.

Collett D. (1991), Modelling Binary Data. Chapman & Hall.

Chapter 1

Introduction

This course is an introduction to statistical modelling that concentrates mainly on linear and generalised linear statistical modelling. This means that we mainly look at regression models, where we have a response and a number of explanatory variables. One exception are loglinear models for contingency tables, where there is no designated response, instead the (conditional) independence structure among the variables is investigated. In this first chapter some preliminary remarks are made that should be kept in mind throughout.

1.1 Aims of statistical modelling

In general one can say that statistical modelling is about predicting, explaining, investigating structure and causal inference. Depending on the aims of the analysis different methods might be appropriate and different criteria will determine what a good model is.

Prediction: Data set A is an example. Here we want to predict the volume from the two diameter measures and the height. We do not need to analyse data in order to know that the volume of lumber that a tree provides depends on these variables. Also we are not interested in causal statements, i.e. we do not want to produce trees with high volume and also we can't force a tree to have a certain height or diameter. Instead we just want a model that reliably predicts the volume.

Explaining: Data set E is an example. The aim is to explain differences in attitudes towards 'mathematics' by location of school and course preference. Clearly, we are only looking for associations, and if we find some association further investigations would be needed to explain these. E.g. if people in suburbs are less likely to be interested in mathematics it might be because they will typically aspire to different jobs but more data is needed to investigate this.

Causal inference: Data set C is an example. An experiment is conducted using different concentrations of a toxic substance to quantify its causal effect on tumor incidence. It is not of interest whether there are other potential causes, or why this substance causes cancer.

In many situations the ultimate aim is to make causal statements like in the last example and this is reflected in the terminology used, such as 'effect' or 'influence', but the standard statistical literature usually shys away from using the word 'cause' except in the case of experimental data. This is because traditionally statisticians see their role in making statements

about associations and correlations while it is for the subject matter experts to decide which of these can be given a causal explanation. However, in the context of regression models it is very tempting to give a causal interpretation to the relations found to be 'significant' between response and explanatory variables even without subject matter knwoledge. In order to see why one should be cautious about this let's have a closer look at what causality might mean.

Causal interpretation of regression models: Regression models can be regarded as modelling the *conditional distribution* of a response Y on some explanatory variables $X_1, \ldots, X_m$. Remembering the interpretation of conditional probabilities we can say that this conditional distribution describes what we can say about Y when we have *seen* the event $X_1 = x_1, \ldots, X_m = x_m$. A causal interpretation in contrast implies that we want to describe the distribution of Y when we *intervene* in some or all explanatory variables, i.e. when we *set* them to prespecified values. In other words, the target of inference is the intervention distribution of Y given we set $X_1 = x_1, \ldots, X_m = x_m$ but the data might only contain information about the conditional distribution of Y given we see $X_1 = x_1, \ldots, X_m = x_m$ — without further knowledge on the data collection (e.g. experiment) or subject matter background there is no reason why these two distributions should be related. E.g. you might find from data that students who attend all tutorials typically obtain better grades. Does this mean that if you force all students to attend all tutorials they will have better grades? Or: a retailer finds that the sales of stores are strongly associated with the size of stores. Does this mean that if he opens a huge store he will have gigantic sales?

These problems of interpretation arise especially with *observational* data, i.e., data that has been collected under uncontrolled conditions. In contrast, experimental data often result directly from interventions, like in the case of Data set C, and can therefore be used to estimate the intervention distribution. However, experiments can be badly designed so that if an effect is found one cannot always attribute it to the explanatory variable.

1.2 Models and reality

It is important to keep in mind that statistical models, as all mathematical models, are *idealisations*. If we, for example, assume a simple linear regression model with a normally distributed error term, this means that we *think about reality* in terms of a linear relationship between our response and predictor variable, and about the deviations of the data from this relationship as independent random quantities distributed according to a normal distribution. The reason is that such a linear relationship is easily interpretable, and the most important information in the data can easily be summarised by looking at the regression line. The normal distribution enables us to quantify our uncertainty and thus gives us an idea how precise our knowledge is. So in order to be able to give a clear interpretation of the data, we need models that guide us to see some striking clear tendencies even though the reality may be much more complex.

We interpret probability models in this course in a *frequentist* way, which means, for example, that if we assume a normal distribution, *we think about our problem in such a way that we expect* that if we repeat the experiment/situation very often, identically and independently, the distribution of observations would approximate the Gaussian bell curve more and more precisely. (Other interpretations of probability exist, but they are idealisations as well.)

We don't believe that the model assumptions really hold precisely - and there is no means of verifying that they do. Observed data are always compatible with a variety of possible distributions, and it is generally not possible to know whether repetitions of situations are

really identical and independent. Some situations cannot be repeated at all, and it generally depends on interpretation what constitutes a "repetition". (Can a new patient of the same age considered to be a "repetition" of the previous patient the doctor has seen?)

We can only find out whether the model saves its purpose. i.e., whether it leads to reliable predictions or convincing explanations.

Even though we don't believe that our models are precisely true, we always have to be concerned about the model assumptions, because they can be violated in such a way that the resulting model leads to *misleading* conclusions (for example bad predictions in case that a linear regression is assumed but the relationship between predictor and response shows a strongly nonlinear pattern).

On the other hand, some violations of the model assumptions are harmless and the methods based on these assumptions give useful results anyway. For example if, as in Example A, all data have only two digits behind the decimal point, the true underlying distribution cannot be normal, because a normal distribution generates real numbers from a continuum while the observations can only take discret values. However, the effect of the discreteness on a resulting analysis based on the normal linear regression can only be very small, as long as the values of the response are still informative enough.

1.3 Steps of statistical modelling

The following is a general outline of the modelling process, in which the description assumes there is one response variable, although there could be several, and is done in the context of the types of problem that will be covered in this course. The modelling process involves several steps, which will be illustrated in the examples considered in the course.

1. **Exploratory data analysis**

   - Is the response variable quantitative or categorical? If the former, is it continuous or discrete; if the latter, is it nominal or ordinal?
   - Similarly, consider each variable that may affect the response; its type will affect how it is represented in the model – distinguish between quantitative explanatory variables (as in regression) and factors (qualitative explanatory variables, as in factorial experiments). Each type may occur in the same model.
   - As appropriate, obtain frequency tables, dot plots or histograms or stem plots for each variable.
   - Look at appropriate descriptive statistics.
   - As a preliminary to regression modelling, draw scatter plots of pairs of variables; for factorial experiments with replication, tabulate treatment means in particular, and other statistics as appropriate.
   - Consider the implications of the above for model formulation (e.g. what is a reasonable distribution to assume for the response variable, are the points in a scatter plot roughly in a straight line?).

2. **Model formulation**

   With the help of the exploratory data analysis and any subject matter background knowledge, propose the following:

- Distribution of response variable.
- Equation linking the expected response with the explanatory variables and/or factors.

3. **Model fitting**

   The model specified in 2 will contain unknown parameters (e.g. in the straight line regression example, intercept and slope). Estimate the unknown parameters by an appropriate method of estimation which, in the context of this course, is mostly least squares or maximum likelihood.

4. **Model checking**

   This step asks whether the model is an adequate fit to the data and is usually based on analysis of residuals. Raw residuals are the difference of the observed values of the response and their **fitted values** which are estimates of the expected response for each observation. However, these residuals are often standardised. (Again recall the straight line regression example: analysis of residuals included plots of residuals against explanatory variable, and normal probability plot of residuals, and recall the reasons for doing this.)

   There will be further discussion of the model checking process later in the course. You may need to go through the whole process again if your model does not provide an adequate fit to the data

---

1.4 Optional reading for Chapter 1

One of the following:

- Dobson Chapters 1 and 2.
- Krzanowski Chapter 1.
- McCullagh and Nelder Chapter 1.

Chapter 2

Linear Models

In fitting a straight line to data on a response variable Y and explanatory variable x, recall the model

$$Y_i = \beta_0 + \beta_1 x_i + e_i \qquad (i = 1, 2, \ldots, N) \tag{2.1}$$

where Y_i denotes, for the i^{th} observation, the response Y corresponding to the value x_i of the explanatory variable x, e_i is the 'error'[1] associated with the i^{th} observation, and β_0 and β_1 are unknown parameters to be estimated from the data on Y and x.

The above model is a simple example of a **linear model** which is one where the **predictor** of the response (in this example $\beta_0 + \beta_1 x_i$) is linear in unknown parameters (in this example β_0 and β_1).

The model (2.1) is often equivalently written as

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i \qquad (i = 1, 2, \ldots, N)$$

and we define $e_i = Y_i - \mathrm{E}(Y_i)$. When the explanatory variables are not fixed by the design of the study it is more appropriate to write $\mathrm{E}(Y_i | X_i = x_i)$ in order to make clear that the x-values are themselves random variables but what we are interested in is not the distribution of the x-values, only the way how Y depends on them.

Now suppose that there is more than one explanatory variable that could be used in the prediction of a response: see chapter 0, example A.

In general, let there be m explanatory variables labelled $x_1, \ldots, x_m$. Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + e_i \qquad (i = 1, 2, \ldots, N) \tag{2.2}$$

where, for the i^{th} observation, Y_i denotes the response Y corresponding to values $x_{i1}, \ldots, x_{im}$ of the explanatory variables $x_1, \ldots, x_m$. In the following, m will typically denote the number of explanatory variables and p will denote the number of regression parameters in total. In the above we have $p = m + 1$ due to the intercept.

This more general model is also a linear model as its predictor is linear in unknown parameters $\beta_1, \ldots, \beta_m$. If all the explanatory variables $x_1, \ldots, x_m$ are continuous then we call the above a **multiple linear regression**. Note that in this concept of a linear model, some of the x's could be derived from others, e.g. squares of explanatory variables. If the explanatory variables are all discrete with few different categories or levels then we have an **analysis of variance**

[1] The term "error" is used here because it is the usual term, but it should not be read as having the "moral" implication that deviations from the model are "bad" or "erroneous" - just the error term being large does not mean that anything is wrong with the data.

model (the notation has to be modified a bit as described in Section 2.5). If some of the explanatory variables are discrete and some continous we have a **general linear regression** (not to be confused with a generalised linear regression). To cover all generalities we will adopt the following vector notation for a linear model:

$$Y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + e_i \qquad (i = 1, 2, \ldots, N) \tag{2.3}$$

where, for the i^{th} observation, Y_i denotes the response Y and $\mathbf{x}_i$ is a column vector whose elements depend on quantities that may affect the response (factors, explanatory variables), and e_i is the associated 'error'. Finally, $\boldsymbol{\beta}$ is a column vector of unknown parameters to be estimated from the data.

The right hand side of equation (2.3) consists of two components:

- a **systematic** component, $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$: this is a **linear predictor** (of the response) and is linear in the sense that it is linear in the unknown parameters, the elements of $\boldsymbol{\beta}$;
- a **random** component, e_i: the error term; or equivalently $Y_i - \mathrm{E}(Y_i)$.

The errors are often assumed to be independent, normally distributed with zero mean and constant variance, σ^2 say. Then the Y_i's are independent, normally distributed random variables such that

$$\mathrm{E}(Y_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \tag{2.4}$$

and

$$\mathrm{var}(Y_i) = \sigma^2.$$

Some special cases

1. **Simple linear regression.** Fitting a straight line to data – has a response variable Y and just one explanatory variable x.

   Examples and theory: Rice, Sections 14.1 and 14.2 (covered in Foundation Fortnight).

   Model given earlier: see (2.1). To compare (2.1) with the general form (2.3):

   $$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_i \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

   Example: R-output 1.

2. **Multiple linear regression.** Generalisation to cover several explanatory variables $x_1, \ldots, x_m$.

   Examples: Examples A and G; also Rice Section 14.5.

   Model given earlier: see (2.2). To compare with (2.2) with (2.3):

   $$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ . \\ . \\ . \\ x_{im} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_m \end{pmatrix}.$$

   Example: R-output 3, also in many other R-outputs

3. **Quadratic in an explanatory variable x.** Model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$$

As the predictor is linear in unknown parameters β_0, β_1 and β_2, then the above is a linear model. To compare with (2.3):

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_i \\ x_i^2 \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

4. **Linear regression with transformed variables.**

More generally, a model is still linear if the variables are (e.g. nonlinear) transformations of originally observed variables. For example, we may observe responses Z_i and predictors $u_{i1}, u_{i2},\ i = 1, \ldots, n$, but model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

with $Y_i = \log(Z_i),\ x_{i1} = \log(u_{i1}),\ x_{i2} = \sqrt{u_{i2}},\ x_{i3} = \log(u_{i1})\sqrt{u_{i2}}$ (more than one original variable may be involved in the definition of a predictor in the final model), so that

$$Y_i = \log(Z_i),\ \mathbf{x}_i = \begin{pmatrix} 1 \\ \log(u_{i1}) \\ \sqrt{u_{i2}} \\ \log(u_{i1})\sqrt{u_{i2}} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Transformation of the originally observed variables is often helpful if

- model assumptions (linearity, normality) are clearly violated for the original variables (often it has to be tried out experimentally which transformations of which variables improve the situation),
- a multiplicative model $Z_i = u_{i1} u_{i2} \ldots u_{im} v_i$ (v_i being the multiplicative 'error') seems to be appropriate. This becomes linear ($e_i = \log v_i$) by taking logarithms on both sides:

$$\log(Z_i) = \log(u_{i1}) + \log(u_{i2}) + \ldots + \log(u_{im}) + e_i,$$

Example: R-output 1

5. **Interactions.**

Interactions can make a statistical model more flexible and realistic, and are used when considering the impacts that at least two explanatory variables have on a response. If two predictors interact then the relationship between each of the interacting variables and the response depends on the value of the other interacting variable. An interaction variable is typically constructed as the product of some regressors. When there are more than two explanatory variables, several interaction variables can be constructed: pairwise-products representing pairwise-interactions and higher order products representing higher order interactions. Model example:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + e_i.$$

To compare with (2.3):

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i1}x_{i2} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Variables x_{i1} and x_{i2} can of course be categorical or continuous and they are referred to as main terms in this context. Note that β_1 and β_2 do not represent main effects, as in the case of no interaction term, rather they are conditional effects. β_1 is the effect of x_{i1} conditional on $x_{i2} = 0$. For values of x_{i2} other than zero, the effect of x_{i1} is $\beta_1 + \beta_3 x_{i2}$. Remember that if, for instance, β_1 turns out not to be significant, then this does not mean that x_{i1} does not have a significant impact on the response.

6. **Comparison of means of two groups: two-sample problem.**

   Example: Rice Section 11.2 (this is the two-sample problem revised in Foundation Fortnight).

   For the j^{th} observation in the i^{th} group:

$$Y_{ij} = \mu_i + e_{ij} \qquad (i = 1, 2;\ j = 1, \ldots, n_i).$$

   With the assumptions about the errors made earlier (see (2.4) above), then the Y_{ij}'s are independent $\mathcal{N}(\mu_i, \sigma^2)$, as in the standard two-sample problem.

   The above model can be written as

$$Y_{ij} = x_{ij1}\mu_1 + x_{ij2}\mu_2 + e_{ij}$$

   where $x_{ijk} = 1$ for $k = i$ and 0 for $k \neq i$, demonstrating that this model is a linear model.

   To compare with (2.3): the single subscript i is replaced by the double subscript ij,

$$\mathbf{x}_{ij} = \begin{pmatrix} x_{ij1} \\ x_{ij2} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

   In this case the x's are not measurements on some variable; they just indicate whether or not to include the corresponding parameter for each observation and are then called **dummy variables** or **indicator variables**.

7. **Comparison of group means: one-way layout**

   The above modelling for the two-sample problem is easily extended to the comparison of the means of I groups where $I \geq 2$.

   Example: Rice Section 12.2 (called a one-way layout)

   For the j^{th} observation in the i^{th} group:

$$Y_{ij} = \mu_i + e_{ij} \qquad (i = 1, \ldots, I;\ j = 1, \ldots, n_i)$$

   which can be written as

$$Y_{ij} = x_{ij1}\mu_1 + \ldots + x_{ijI}\mu_I + e_{ij}$$

   where $x_{ijk} = 1$ for $k = i$ and 0 for $k \neq i$.

To compare with (2.3): the single subscript i is replaced by the double subscript ij,

$$\mathbf{x}_{ij} = \begin{pmatrix} x_{ij1} \\ . \\ . \\ . \\ x_{ijI} \end{pmatrix} \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ . \\ . \\ . \\ \mu_I \end{pmatrix}.$$

The modelling for a one-way layout will be discussed further in Section 2.5.

Note that in the one-way layout, the response depends on 'levels' (categories) of a qualitative variable called a **factor**. Of the type of examples discussed above, this course will mainly concentrate on the regression models where the x's are quantitative explanatory variables. Designed experiments involving factors are discussed in the STAT0029 course.

2.1 Inference for linear models

In this section we consider methods of inference about the parameters of the model (2.3).

2.1.1 Estimation of regression parameters

(i) **Least squares estimation** is generally used for estimating the unknown parameters in the linear model (2.3). The method is described in Rice Section 14.2. The following generalises the ideas to the linear model (Rice Section 14.3).

The least squares estimators of the elements of $\boldsymbol{\beta}$ minimise the sum of the squares of the errors, ie minimise

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2$$

with respect to the elements of $\boldsymbol{\beta}$.

(ii) The method and results are better described by writing the model (2.3) in **matrix form**:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.5}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_N \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ . \\ . \\ . \\ \mathbf{x}_N^{\mathrm{T}} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ . \\ . \\ . \\ e_N \end{pmatrix}.$$

So with p parameters in $\boldsymbol{\beta}$, $\mathbf{X}$ is a $N \times p$ matrix with i^{th} row $\mathbf{x}_i^{\mathrm{T}}$.

The form of model (2.5) is known as the **general linear model** in that it can include continuous explanatory variables and or factors.

$\mathbf{X}$ is known as the **design matrix** in the context of designed experiments.

(iii) **Normal equations.**

Using the matrix form (2.5),

$$S(\boldsymbol{\beta}) = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

By differentiating $S(\boldsymbol{\beta})$ with respect to the elements of $\boldsymbol{\beta}$, it follows that the least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ satisfies

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{2.6}$$

which are called the **normal equations**. *Proof:* in Rice Section 14.3.

For most of this course it will be assumed that there are fewer parameters than observations, so that $p < N$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ has full rank and is invertible. Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{2.7}$$

Note that the practical calculation of $\hat{\boldsymbol{\beta}}$ is usually not performed by inverting $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ as this is an inefficient computational method. Better methods exist but are not covered in this course. If you wish to pursue this further for your own interest, some details are found in McCullagh and Nelder Section 3.8.

(iv) **Properties of least squares estimators**

- **$\hat{\boldsymbol{\beta}}$ is an unbiased estimator:** Assuming that $\mathrm{E}(e_i) = 0$ we have

  $$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

  Proof: in Rice Section 14.4.

- **Covariance matrix of $\hat{\boldsymbol{\beta}}$:** if in addition $e_i, i = 1, \ldots, N$, are independent with constant variance σ^2, then

  $$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$$

  Proof: in Rice Section 14.4.
  Note: in this course 'variance-covariance matrix' will usually be abbreviated to 'covariance matrix'.

- **Sampling distribution of $\hat{\boldsymbol{\beta}}$:** now assume that the errors are independent normally distributed with mean 0 and constant variance σ^2 (the normal distribution assumption has not been needed so far in this chapter), then

  $$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}) \tag{2.8}$$

  where $\mathcal{N}_p$ denotes multivariate normal distribution with dimension p. This follows from noting that $\hat{\boldsymbol{\beta}}$ is a linear transformation of $\mathbf{Y}$.

(v) **Gauss–Markov Theorem:** if the above assumptions except normality are satisfied then we have the following result. If $\psi = \mathbf{c}^{\mathrm{T}}\boldsymbol{\beta}$ is an estimable function, then the unique linear unbiased estimator of it which has minimum variance is $\hat{\psi} = \mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is any solution of the normal equations.

In the above statement, $\mathbf{c}$ is a known $p \times 1$ vector; a linear estimator of ψ is a linear combination of $Y_1, \ldots, Y_N$; ψ is estimable if it has a linear unbiased estimator.

This theorem has relevance for inferences about linear combinations of the parameters of a linear predictor. A proof can be found in:

Seber G.A.F., Lee A.J. (2003), Linear Regression Analysis. 2nd edition, p. 43.

(vi) **Maximum likelihood estimation**

The least squares estimator is also a maximum likelihood estimator (MLE) when the responses are normally distributed. Assuming the responses are also independent, the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{S(\boldsymbol{\beta})}{2\sigma^2}\right)$$

from which it can be seen that maximisation of $L(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ for given σ^2 is equivalent to minimisation of $S(\boldsymbol{\beta})$.

(vii) **Geometric approach**

Least squares estimation has a geometric interpretation that is useful for proofs of associated results. This approach is beyond the scope of this course, but if you are interested in this more advanced approach to linear model theory, then it is introduced in McCullagh and Nelder Section 3.6, and covered in detail in Chapters 1 and 2 of a classic text by Scheffé H. (1959), *The Analysis of Variance*, Wiley.

2.1.2 Estimation of σ^2

This was also considered in Foundation Fortnight, particulary in the context of simple linear regression models. In assessing the fit of a model we consider the residuals, i.e. the differences of the observed and fitted values.

(i) **Residuals.**

The fitted values are the estimates of the mean response for each observation: for the i^{th} observation, the fitted value is $\hat{\mu}_i = \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. So the residual is

$$\hat{e}_i = Y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$$

In matrix notation, the vector of fitted values is $\mathbf{X}\hat{\boldsymbol{\beta}}$ and the vector of residuals is

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

or

$$\hat{\mathbf{e}} = (\mathbf{I}_N - \mathbf{H})\mathbf{Y}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix and

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

an $N \times N$ matrix.

(ii) **Residual sum of squares.**

The sum of squares of the residuals is called the **residual sum of squares** and is the minimum value of $S(\boldsymbol{\beta})$, and will be simply denoted by RSS, ie

$$\mathrm{RSS} = \hat{\mathbf{e}}^{\mathrm{T}}\hat{\mathbf{e}} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2.$$

By using various matrix properties it can be shown that the expected value of RSS is

$$\mathrm{E}(\mathrm{RSS}) = \mathrm{E}(\hat{\mathbf{e}}^{\mathrm{T}}\hat{\mathbf{e}}) = \mathrm{E}[\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_N - \mathbf{H})\mathbf{Y}] = (N - p)\sigma^2$$

Proof: Rice Section 14.4, who uses the notation $\mathbf{P}$ instead of $\mathbf{H}$.

(iii) Hence an **unbiased estimator** of σ^2 is

$$\hat{\sigma}^2 = \frac{\text{RSS}}{N-p}.$$

(iv) **Sampling distribution of RSS:** under the assumption of independent $e_i \sim \mathcal{N}(0, \sigma^2)$, it can be shown that $\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{N-p}$, independent of $\hat{\boldsymbol{\beta}}$.

A proof can again be found in the book by Seber and Lee.

2.1.3 Weighted least squares estimation

(i) **Independent errors.**

Weighted least squares estimation is used when the error variance is known not to be constant. Weighted least squares estimators of the β_j's minimise

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} w_i (Y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})^2$$

with respect to the elements of $\boldsymbol{\beta}$, where the w_i's are weights. Greater weight is assigned to observations that are more reliable, ie have smaller variance. Ideally we put $w_i = 1/\text{var}(Y_i)$, which is what you would obtain from the usual criterion based on the transformed response $\sqrt{w_i} Y_i$ which has constant variance.

In matrix form the above criterion becomes

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.9}$$

where $\mathbf{V}$ is a $N \times N$ diagonal matrix with the $\text{var}(Y_i)$'s as the diagonal elements (but the exact variance is usually not known and must be estimated or guessed).

(ii) **Correlated errors.**

This method can be generalised to cover the case when the errors are correlated. The criterion is still (2.9) but now $\mathbf{V}$ denotes the covariance matrix of the errors (and also for the responses). The following derivations hold for independent errors as well.

The normal equations can be shown to be

$$\mathbf{X}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{Y}, \tag{2.10}$$

therefore (in case that inversion is possible)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{Y}. \tag{2.11}$$

(Proof omitted.)

Note that if $\mathbf{V}$ is multiplied by a constant, the relevant statistical quantities don't change; for $\hat{\boldsymbol{\beta}}$ this can be directly seen from (2.11).

The problem with the given theory for weighted least squares is that it can only be applied if $\mathbf{V}$ is known (up to a constant factor), which is not the case in practice.

In the situation (i) (independent errors), it is sometimes reasonable to assume that the variances of Y_i are proportional to a function $f(\mathbf{x}_i)$ of one or more of the predictors. It is then possible to choose $w_i = 1/f(\mathbf{x}_i)$ (even if $\text{var}(Y_i) = cf(\mathbf{x}_i)$ for some unknown constant c). In other cases, weighted least squares can be used iteratively, starting from an initial guess of $\mathbf{V}$ and then re-estimating $\mathbf{V}$ based on the weighted least squares estimator from the previous iteration. This requires methods to estimate $\mathbf{V}$; examples occur later in the course.

2.1.4 Confidence intervals (CIs) and tests

CIs and tests for the regression parameters

In the case of a general linear model with $\mathcal{N}(0, \sigma^2)$ errors and σ^2 unknown, we can derive from (2.8) that

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_j),$$

where v_j is the $(j,j)^{th}$ diagonal element of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. From (iv) in Section 2.1.2

$$\frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{N-p}$$

gives

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 v_j}} \sim t_{N-p}. \tag{2.12}$$

where $\hat{\sigma}^2 = \mathrm{RSS}/(N-p)$.

Proof: from definition of t-distribution, Rice Section 6.2.

Hence an exact $100(1-\alpha)\%$ confidence interval for β_j has limits

$$\hat{\beta}_j \pm t_{N-p,\frac{1}{2}\alpha}\mathrm{se}(\hat{\beta}_j)$$

where $\mathrm{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_j}$ and $t_{N-p,\frac{1}{2}\alpha}$ is the upper $100 \times \frac{1}{2}\alpha\%$ point of the t_{N-p} distribution.

The distribution (2.12) can easily be used to test hypotheses of the form H_0: $\beta_j = \beta^*$ for a given j. Particular interest is often paid to H_0: $\beta_j = 0$. In the latter case the test statistic and null distribution are

$$\frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)} \sim t_{N-p} \text{ under } H_0.$$

Hence we can obtain the p-value in the usual way (as described in Foundation Fortnight). In the context of regression, this tests whether the response depends on the associated explanatory variable, *given the inclusion of the other explanatory variables* in the model.

CI and test for a linear combination of parameters

Inferences about $\psi = \mathbf{c}^{\mathrm{T}}\boldsymbol{\beta}$ – an example is estimation of expected response. Let $\hat{\psi} = \mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$.

Use

$$\frac{\hat{\psi} - \psi}{\sqrt{\hat{\sigma}^2 v}} \sim t_{N-p}$$

where $v = \mathbf{c}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{c}$

Proof: result follows from $\hat{\psi} \sim \mathcal{N}(\psi, \sigma^2 v)$, and from the distribution of RSS and the definition of the t-distribution (similar procedure to that used for a single β_j above).

Hence we can easily construct a $100(1-\alpha)\%$ confidence interval for ψ.

The above t–distribution can also be used to derive a test statistic for H_0: $\psi = \psi^*$.

Tests about multiple parameters

Now consider testing the null hypothesis H_0 that a subset of $p-q$ parameters out of the p parameters in the linear predictor are all 0, leaving q non-zero parameters.

Let H_1 denote the alternative hypothesis that all p parameters are not 0.

Let RSS_0 and RSS denote the residual sum of squares under H_0 and H_1, respectively. We know that $\text{RSS}/\sigma^2 \sim \chi^2_{N-p}$ and similarly, under H_0, $\text{RSS}_0/\sigma^2 \sim \chi^2_{N-q}$. In addition it can be shown that the difference $\text{RSS}_0 - \text{RSS}$ is independent of RSS (proof omitted). This yields that

$$\frac{\text{RSS}_0 - \text{RSS}}{\sigma^2} \sim \chi^2_{p-q} \text{ under } H_0$$

and we obtain the F-test:

$$\frac{(\text{RSS}_0 - \text{RSS})/(p-q)}{\text{RSS}/(N-p)} \sim F_{p-q,N-p} \text{ under } H_0.$$

Proof: uses the definition of the F-distribution based on ratios of independent chi-squared random variables, as given in Rice (Section 6.2).

In the case when $p - q = 1$, then $F = t^2$ where t is the t-statistic given above.

Optional reading for Section 2.1

Details of the underlying theory of general linear models are dealt with in hundreds of books. Rice contains details on the properties of least squares without assuming normally distributed errors. The following book has the full details and has been cited earlier.

- Seber G.A.F., Lee A.J. (2003), Linear Regression Analysis. 2nd edition, Wiley, Section 3.

2.2 Multiple linear regression

Example: R-output 3

This chapter concentrates on the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + e_i \qquad (i = 1, \ldots, N)$$

where Y_i is the response Y corresponding to values $x_{i1}, \ldots, x_{im}$ of m explanatory variables $x_1, \ldots, x_m$.

It will be assumed that the errors $e_1, \ldots, e_N$ are independent, normally distributed with zero means and constant variance σ^2, so that the Y_i's are independent $\mathcal{N}(\mu_i, \sigma^2)$ where

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im}.$$

In matrix form the model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ (cf. equation (2.5)) where the i^{th} row of $\mathbf{X}$ is $(1, x_{i1}, \ldots, x_{im})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^{\mathrm{T}}$, so that the number of unknown parameters in the linear predictor is $p = m + 1$.

In this chapter, assume that the $p \times p$ matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is of full rank and thus invertible.

2.2.1 Some basic results

From Section 2.1 we have the following basic results

(i) **Least squares estimates:**

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

(ii) **Sampling distribution of $\hat{\boldsymbol{\beta}}$:**

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}).$$

(iii) **Residual sum of squares:**

$$\mathrm{RSS} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - 2\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} + \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

after using the normal equations $\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

(iv) **Unbiased estimator of σ^2:**

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{N - m - 1}.$$

(v) **t-test for H_0: $\beta_j = 0$.**

$$t = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)} \sim t_{N-m-1} \text{ under } H_0.$$

This is a test for the omission of the j^{th} explanatory variable *given the other explanatory variables* are kept in the model.

(vi) **F-test for H_0: ν of the regression parameters $\beta_1, \ldots, \beta_m$ are 0**
i.e. testing the omission of ν explanatory variables where $\nu \leq m$.

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/\nu}{\mathrm{RSS}/(N - m - 1)} \sim F_{\nu, N-m-1} \text{ under } H_0.$$

where RSS_0 is the residual sum of squares under H_0.

(vii) **Special case** H_0**:** $\beta_1 = \beta_2 = \ldots = \beta_m = 0$, *i.e. none of the explanatory variables affect the response.* In this case, under H_0, the least squares estimator of β_0 is just the sample mean of the responses $\overline{Y}$, and so in (vi),

$$\mathrm{RSS}_0 = \sum_i (Y_i - \overline{Y})^2 = \textbf{corrected total sum of squares}, \text{ denoted by CTSS.}$$

The numerical results for this F-test are often reported in the form of an **analysis of variance table**:

Source of variation	Sum of squares	df	Mean square	F
Regression	SS(reg)	m	SS(reg)$/m$	$\frac{\mathrm{SS(reg)}/m}{\mathrm{RSS}/(N-m-1)}$
Residual	RSS	$N-m-1$	RSS$/(N-m-1)$	
Total	CTSS	$N-1$		

In the above table SS(reg) = CTSS − RSS, the **sum of squares due to the regression**. It can be shown that

$$\mathrm{SS(reg)} = \sum_{i=1}^{N} (\hat{\mu}_i - \overline{Y})^2,$$

where $\hat{\mu}_i = \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}$ are the fitted values.
Proof: see appendix to this section.

Note that Dobson Table 6.1 gives a form of analysis of variance table to cover the more general case first considered in (vi). This form of table is not usually seen in practice because, in general, the model usually includes the constant term, which is then excluded from the list of sources of variation whose degrees of freedom then total to $N-1$ instead of N. This will be more evident when you consider the analysis of data from designed experiments in the STAT0029 course.

The term 'analysis of variance' is often abbreviated to **anova**.

2.2.2 Interpretation of regression parameters

Example: R-output 2

Partial regression coefficient
The coefficient β_j of an explanatory variable x_j in a multiple linear regression model that includes other explanatory variables is called a **partial regression coefficient**. It measures the rate of change of the mean response with x_j while holding constant the values of the other explanatory variables in the model.

Total regression coefficient
The coefficient of x_j in the simple linear regression model of the response variable on x_j on its own is called a **total regression coefficient**. It measures the rate of change of the mean response with x_j ignoring the values of the other explanatory variables.

Thus the partial and total regression coefficients measure different quantities and so, as you would expect, their estimates are different (examples in lecture). However the estimates are the same if the columns in the **X**-matrix (see exercises) are orthogonal, which you do not generally have in observational data.

2.2.3 Checking model adequacy

Examples: R-output 1, 3, 4, 6, 7, 8

(i) **Assessment of the model assumptions.** Here are the crucial model assumptions of the multiple linear regression model:

Linearity of the relationship between predictors and Y.

Normality of errors e_i. Though the normal distribution theoretically can generate arbitrary real numbers (including very large ones), very extreme values occur under the normal distribution with a very small probability. For example, the probability that a point is generated which is further than 5σ away from the mean value (0 for residuals) is about $5 * 10^{-7}$ which means that only one such point can be expected in more than 1.5 millions of points generated from the same normal distribution. Because outliers have a very strong influence on least squares regression, the detection of outliers is the most important task connected to the normality assumption about the ϵ_i. Another possible important deviation from normality could be skewness of the distributional shape (many points one one side but much fewer points, some of them appearing somewhat outlying, on the other side).

Some other deviations from normality are less dangerous. For example, under uniformly distributed random variation with restricted value range (this may hold, for example, if the y-variable consists of percentages), the normal theory still works quite well. Generally, you should have in mind that normality (as well as all the other model assumptions) is an idealization that never holds *precisely* in practice. The important question is not whether a distribution is *really* normal but whether there are deviations from normality that may lead the applied statistical methodology (here linear regression) to misleading conclusions about the data.

Homogeneity of variances ("homoscedasticity", as opposed to "heteroscedasticity") of e_i. This implies particularly that the variances do not depend on any of the predictor variables.

Independence of errors e_i (of each other).

(i-a) **Matrix plot** The so-called matrix plot consists of all scatterplots of any pair of predictor variables and predictors vs. response, arranged in matrix form. The matrix plot can (and should) be plotted without having fitted a linear regression (or any other model) before. The plots of predictor variables vs. the response can be used to assess linearity, outliers and homoscedasticity. Note, however, that there is some danger of over-interpreting impressions from these plots, because to see the response plotted against every single predictor variable does not give a full impression. For example, it is possible that an apparently nonlinear shape of the plot of a single predictor vs. the response is rather caused by values of the other predictors than by a real violation of linearity (though strong nonlinear or heteroscedastic shapes hint at real violations of model assumptions in practice in most cases).

The plots of pairs of predictors can reveal collinearity and leverage points (see below), which are not violations of the model assumptions (there are no model assumptions about the predictor variables in linear regression), but still problematic.

(i-b) **Residuals and standardized residuals.** The residuals are the deviations of the observations of their "ideal" value according to the model. They can be interpreted as

estimators of the errors e_i (and are therefore denoted by $\hat{e}_i$). This suggests that the residuals can be used to assess the model assumptions about the errors. The **standardised residuals** are residuals divided by their estimated standard deviations, so that the standardized residuals should have a standard deviation of 1, which helps assessing their size. Too many standardised residuals with magnitude greater than 2 suggests that the error distribution has heavier tails than a normal distribution, but note that about 5% of standardised residuals are expected to be larger than 2 in magnitude even if the errors are normally distributed (remember that $2 \approx 1.96 = 97.5\%$ quantile of the standard normal distribution).

It can be shown (Rice Section 14.4) that the covariance matrix of the residuals is

$$\mathbf{V}(\hat{\mathbf{e}}) = \sigma^2[\mathbf{I}_N - \mathbf{H}]$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

is called the **hat matrix**, previously met in Section 2.1.2(i).

So $\mathrm{var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$ and $\mathrm{cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2 h_{ij}$ for $i \neq j$,
where h_{ij} is the $(i,j)^{th}$ element of $\mathbf{H}$.

While the errors are assumed to be uncorrelated, the above result shows that the residuals are, in general, correlated. The above result also gives a formula for the standardised residuals. For the i^{th} observation, the standardised residual is

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

(i-c) **Residual plots.** It makes sense to use the standardized residuals are for residual plots (particularly when looking for heteroscedasticity, because they should have constant variance), but the raw residuals may be looked at as well.

There are several possible residual plots.

Predictor vs. residuals. The errors are assumed to be independent from the predictor variables, and therefore the residuals should look randomly scattered when plotted against every single predictor variable. Otherwise, the plot can reveal non-linearity, heteroscedasticity, autocorrelation (i.e. dependence among themselves) of residuals with neighboring values of the predictor and outliers.

Fitted values vs. residuals. If the model is true, the correlation between the residuals and fitted values is zero (appendix, or Rice Section 14.4). The plot should look randomly scattered. The fitted values are, mathematically, a linear combinations of the predictors. Therefore, this plot can reveal the same kind of problems as the predictor vs. residuals plot. (Geometrically, the x-axis of this plot gives information along a further, $m+1$st, direction of the m-dimensional predictor space, complementing the m views along the m predictors.)

Observation order vs. residuals. This makes sense if the observation order is informative (i.e. time). As the errors are assumed to be i.i.d., this plot should look randomly scattered as well. It often reveals autocorrelation among the residuals, but may show heteroscedasticity as well. (As long as the observation order is not a predictor in itself, there is no linearity assumption to be checked here, but strongly nonlinear patterns may still be worth investigation.)

Normal probability plot of residuals. The normal probability plot plots the sorted (standardized) residuals $r_{(i)}$ (denoting the ith smallest residual) against the theoretical quantiles of the normal distribution ($\Phi^{-1}(\frac{i-0.5}{n})$), which are the "ideal" locations of sorted realizations of a standard normal distribution[2]. This should look roughly like a straight line. Otherwise, it indicates deviations from normality, including outliers (which can be seen at the extreme ends of the plot).

(i-d) **Remedies for violated model assumptions**

Non-linearity. Sometimes it helps to transform one or more of the predictors and/or the response. A linear model may hold for some nonlinear functions of the observed variables. First candidates are usually taking logs, square roots, squares, exponentials.

There are also techniques for fitting nonlinear relationships (not treated in this course but see Seber G.A.F., Wild C.J. (2003), Nonlinear regression, Wiley) and techniques for nonparametric regression, i.e., for fitting more general functions that do not have an easy form defined by few regression parameters. More can be found in the book of Hastie, Tibshirani and Friedman. In some cases, a generalized linear model, as treated later in this course, may be appropriate.

Non-normality of the error distribution. Robust linear regression estimators as treated in the Section 2.3 may help, particularly with outliers. In case of skewness, transformations could improve the situation. A possibility is to apply the same transformation to response and predictors.

Heteroscedasticity (as opposed to "homoscedasticity"). Weighted least squares (or iterative reweighting); transformations; sometimes (in the case that heteroscedasticity only accounts for a few more extreme residuals than would be expected under the normal model) robust linear regression.

Dependence of errors. Sometimes this does not affect regression parameter estimators, but it does affect standard deviations and confidence intervals. If assumptions about the nature of dependence can be made, time series models may apply (see "Forecasting"-course).

(ii) **Coefficient of determination**, R^2. This is defined by

$$R^2 = \frac{\sum(\hat{\mu}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{\text{SS(reg)}}{\text{CTSS}} = 1 - \frac{\text{RSS}}{\text{CTSS}}$$

which is the *proportion of the total variation explained by the regression model* (see above anova table).

R^2 is also the square of the correlation between the observed and fitted values (see appendix Section 3). This correlation is known as the **multiple correlation coefficient**.

The coefficient of determination (with a maximum value of 1) measures how well the model accounts for the data. Note, however, that it is *not* directly related to the model assumptions. A small value of R^2 may occur in case that assumptions are violated *or* that some crucial information (further predictors) is missing in the data *or* that the model is fine but the error variance is large, so that the predictors don't have a large explanatory or predictive strength.

[2]Unfortunately, there is more than one possibility to define the "theoretical ideal quantiles"; the given version is used by R for $n > 10$. An alternative is $\Phi^{-1}(\frac{i}{n+1})$.

On the other hand, violations of the model assumptions are still possible if R^2 is relatively high. If the true relationship between predictors and response is strong and monotone but slightly nonlinear, or in case of heteroscedasticity with comparably small error variances, a linear regression may still yield a relatively good fit and a high R^2.

(iii) **Outliers.** Outliers are observations that, in some way, behave differently from the bulk of the data. They can for instance be extreme in the x- or in the y-direction. Hence we distinguish between the following:

Regression outliers: these are observations that have an unusual y-value compared to other observations with similar x-values. If there are only few regression outliers these may show up in residual plots as having large residuals.

Leverage points: these are observations that have unusual x-values compared to the bulk of the data. In a designed experiment they can be prevented but they are very common in observational data. Note that linear regression does not assume normality for the predictors, and therefore **leverage points do not violate the model assumptions** (except if they are "bad", see below). But they cause instability of the regression in the sense that small modifications of the data may lead to large changes in the least squares regression estimator.

Depending on whether the y-value is also unusual we further distinguish:

- *Good leverage points:* if the y-value is 'in line' with other y-values. Typically, for a good leverage point the fitted value $\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ is similar to the observed value y_i and omitting this observation will not change the fitted model dramatically.
- *Bad leverage points:* if the y-value is also unusual. For a bad leverage point the fitted value $\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ can be close to or very different from the observed value y_i — depending on the extend of the leverage effect.

Hence leverage describes the *potential* for affecting the model fit. An intuitive way to measure the leverage of an observation $(\mathbf{x}_i, y_i)$ is to calculate 'how far away is $\mathbf{x}_i$ from the centre of the x-values'. As this is scale dependent we should standardise this distance, which leads to the following notion of **Mahalanobis Distance** MD_i

$$MD_i = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}\hat{\Sigma}_x^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})},$$

where $\hat{\Sigma}_x$ is the empirical covariance matrix of the x-observations. It can be shown that there is a one-to-one relation with the diagonal elements of the Hat matrix:

$$MD_i^2 = (N-1)\left[h_{ii} - \frac{1}{N}\right]. \tag{2.13}$$

The leverage of the i^{th} observation is therefore often just measured by h_{ii}. Under multivariate normality of the predictors (which is usually not assumed in regression), $MD_i^2 \sim \chi^2_{p-1}$, which can be used to assess whether a mahalanobis distance is unusually large.

Note that leverage points can be prevented in designed experiments.

Least squares is heavily affected by outliers and it is theorefore a good idea to check, before even fitting the data, if there are any obvious unusual observations, e.g. by looking at the matrix plot (leverage points can be found in the plots of pairs of the predictor variables). However, in higher dimensions (i.e. more than 2 explanatory variables) we

might not be able to spot such unusual observations by looking at 2-dimensional plots. The residual plots can help, but it has to be kept in mind that the residuals are computed from the fitted model which might itself already be affected and distorted by outliers.

The problem with influential observations such as bad leverage points is that they may have such a large influence on the least squares regression that they actually produce a smaller residual than other points and cannot be revealed by residual plots.

Another possibility of finding out about the influence of an observation is **Cook's statistic.** This is probably the most popular measure of influence among those proposed. Fit the model repeatedly, always omitting one observation, and see how that changes the fitted values: the change in fitted values is $\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$ where $\hat{\boldsymbol{\beta}}_{(i)}$ denotes the least squares estimator of $\boldsymbol{\beta}$ without the i^{th} observation. Cook's statistic is

$$D_i = \frac{1}{p\hat{\sigma}^2}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$

A large value of D_i indicates that the i^{th} observation is influential.

The numerator of D_i is just the sum of the squared differences of the fitted values with and without the i^{th} observation. It can also be shown that

$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2$$

which means that D_i can be computed easily for all i without the need to actually fit the regression model dropping one observation at a time.

Unfortunately, neither the Mahalanobis distance nor Cook's statistic can reliably find all leverage points. A reason for this is the so-called **masking effect**, which means that if there is more than one leverage point (particularly at about the same location), they all together can prevent that every single one of them produces an unusual value on any of these statistics.

The only possibility to cope with this is to use robust estimators which are less sensitive even to groups of leverage points, see Section 2.3. Residual plots with residuals computed from robust estimators are more informative about outliers.

(iv) **Collinearity.** "Strict collinearity" means that there is linear dependence among the predictor variables. In that case, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not invertible and the least squares estimator does not exist. In practice this may happen if the number of predictors is very large compared to the number of observations.

"Approximate collinearity" means that the predictors are nearly linearly dependent. This happens particularly if some of the predictors are strongly correlated. This may be detected from the matrix plot (though sometimes the interplay of more than two variables may produce collinearity, and situations with a high number of variables and a relatively low number of data points are again dangerous). Even though the least squares estimator can be computed in this case and it is not a violation of the model assumptions, approximate collinearity is problematic. The fact that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is close to singularity means that some of the regression parameter estimators (particularly those belonging to strongly correlated predictors) may be very unstable and should be interpreted with care ($\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, the covariance matrix of $\hat{\boldsymbol{\beta}}$, may contain some very large variance entries).

An intuitive reason is that if two predictors measure more or less "the same thing" (i.e. are highly correlated), it cannot be clearly separated what any of them contributes to the explanation of the response.

One method to deal with collinearity is variable selection (Section 2.4). There is also a technique called "Ridge regression", treated for example in Section 3.4.3 of Hastie, Tibshirani and Friedman (2001).

2.2.4 Optional reading for Section 2.2

One of the following:

- Dobson Sections 6.1 to 6.3.
- Krzanowski Sections 3.3 and 3.5.

2.2.5 Appendix: more on multiple linear regression

For this appendix, assume that the linear predictor has the general form in Section 2.1, i.e.

$$\mu_i = \sum_{j=1}^{p} x_{ij}\beta_j = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}.$$

1. **More on least squares estimation.**

   (i) The normal equations (see (2.6)) are

   $$\sum_{i=1}^{N} x_{ij}(Y_i - \hat{\mu}_i) = 0 \text{ for } j = 1, \ldots, p. \tag{2.14}$$

   Multiply the j^{th} equation in (2.14) by β_j and then sum all of the equations over $j = 1, \ldots, p$ gives

   $$\sum_{j=1}^{p} \beta_j \sum_{i=1}^{N} x_{ij}(Y_i - \hat{\mu}_i) = 0,$$

   $$\text{i.e. } \sum_{i=1}^{N} \mu_i(Y_i - \hat{\mu}_i) = 0, \tag{2.15}$$

   $$\text{and in particular } \sum_{i=1}^{N} \hat{\mu}_i(Y_i - \hat{\mu}_i) = 0. \tag{2.16}$$

   (ii) The sum of squares to be minimised

   $$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{N}(Y_i - \mu_i)^2 \\ &= \sum_{i=1}^{N}[(Y_i - \hat{\mu}_i) + (\hat{\mu}_i - \mu_i)]^2 \\ &= \sum_{i=1}^{N}(Y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{N}(\hat{\mu}_i - \mu_i)^2 + 2\sum_{i=1}^{N}(Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) \end{aligned}$$

Use of (2.15) and (2.16) shows that the cross product term in the last expression is zero, so

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N}(Y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{N}(\hat{\mu}_i - \mu_i)^2 \tag{2.17}$$

from which it follows that the least squares estimators do indeed minimise $S(\boldsymbol{\beta})$ because it makes the second term zero.

2. **Partition of CTSS in anova table** (Section 2.2.1 (vii)).

$$\begin{aligned}
\text{CTSS} &= \sum_{i=1}^{N}(Y_i - \overline{Y})^2 \\
&= \sum_{i=1}^{N}[(Y_i - \hat{\mu}_i) + (\hat{\mu}_i - \overline{Y})]^2 \\
&= \sum_{i=1}^{N}(Y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{N}(\hat{\mu}_i - \overline{Y})^2 + 2\sum_{i=1}^{N}(Y_i - \hat{\mu}_i)(\hat{\mu}_i - \overline{Y})
\end{aligned}$$

For the constant term in the multiple linear regression model, the corresponding column in $\mathbf{X}$ consists of all 1's and so the first equation in (2.14) is

$$\sum_{i=1}^{N}(Y_i - \hat{\mu}_i) = 0, \text{ i.e. } \sum_{i=1}^{N}\hat{e}_i = 0 \tag{2.18}$$

Hence

$$\text{CTSS} = \text{RSS} + \sum_{i=1}^{N}(\hat{\mu}_i - \overline{Y})^2 \tag{2.19}$$

as the cross product term is zero by (2.16) and (2.18). Hence the expression for SS(reg) in Section 2.2.1 (vii).

3. **Correlation between residuals and fitted values** (Section 2.2.3 (i-c)).

In vector notation, (2.16), is

$$\hat{\boldsymbol{\mu}}^{\mathrm{T}}(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = 0$$

which implies that the vectors of fitted values and residuals are orthogonal and that the correlation between the residuals and fitted values is 0 assuming the model is true.

4. **Multiple correlation coefficient** (Section 2.2.3 (ii)).

$$R = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})(\hat{\mu}_i - \overline{\hat{\mu}})}{\sqrt{\text{CTSS}\,\sum_{i=1}^{N}(\hat{\mu}_i - \overline{\hat{\mu}})^2}}\,.$$

Equation (2.18) gives $\overline{\hat{\mu}} = \overline{Y}$ and so the numerator of R is

$$\begin{aligned}
\sum_{i=1}^{N}[(Y_i - \hat{\mu}_i) + (\hat{\mu}_i - \overline{Y})](\hat{\mu}_i - \overline{Y}) &= \sum_{i=1}^{N}(\hat{\mu}_i - \overline{Y})^2 \text{ using (2.16) and (2.18)} \\
&= \text{SS(reg)}.
\end{aligned}$$

On the denominator, $\sum_{i=1}^{N}(\hat{\mu}_i - \overline{\hat{\mu}})^2 = \text{SS(reg)}$, hence $R = \sqrt{\text{SS(reg)/CTSS}}$.

2.3 Robust regression

Examples: R-output 4, 6, 7

Outliers are common in many data situations. They can be individual observations that are far away from the main bulk of the data, or they can also be clusters of observations that seem to follow a different pattern than the main set of observations.

Reasons for outliers can be mistakes in the data recording process or missing information which could have been added as categorical variables (e.g. imagine a medical study where a small number of patients is diabetic and this has not been recorded), but there may also be outliers for which there is no clear explanation. Whenever outliers are detected, it makes some sense to check these observations again and to find reasons, if possible. Outliers are bound to occur in observational studies (but not exclusively there), where we can rarely expect that all relevant background variables have been recorded. While they are easy to spot in two dimensions (and the human eye is often better than any algorithm) it becomes really difficult to 'see' them in higher dimensions. Robust methods do not just provide a certain degree of protection against outliers but also against model misspecification in general; for instance, when the error distribution is not normal but follows a distribution with heavier tails than the normal (which basically means that more extreme data points are to be expected even without committing mistakes or missing information).

Least squares estimation has optimality properties (cf. Gauss Markov Theorem) that make it the most popular method for linear regression. However, as mentioned in the previous section, it is also very sensitive with regard to outliers, i.e. even a single outlier can 'divert' the estimate $\hat{\boldsymbol{\beta}}$ arbitrarily far away from the true value.

To understand why least squares is so sensitive recall that the estimator is found by minimising (cf. Section 2.1.1)

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} e_i(\boldsymbol{\beta})^2 = \sum_{i=1}^{N} (Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2. \tag{2.20}$$

Due to squaring the residuals, any residual with a large magnitude will have an extremely large contribution to the sum compared to the others.

There are many different suggestions for robust regression estimators in the literature and there is no unique optimality result that can be used to find the "best one". The reason for this is that robust estimators should achieve various things at the same time. They should be good under the normal model ("efficiency") as well as under linear regression models with other distributions of the error term (particularly those with heavy tails), and they should not be sensitive to outliers ("high breakdown point").

Note that (2.20) is equivalent to minimising the error variance ("scale") estimator $\hat{\sigma}^2$. A general principle for constructing linear regression estimators is to minimise a scale estimator, i.e., a function that is proportional to the variation of the residuals around the regression hyperplane. The general idea is that a regression estimator is good if the residuals are generally small in absolut value, and different approaches to measure the "general size of the residuals" lead to different regression estimators.

Here are two simple ways of 'robustifying' the LS procedure.

(i) Instead of squaring the residuals use another function to reflect the distance between Y_i

and $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, e.g. the absolute distance, i.e. minimise

$$\sum_{i=1}^{N} |e_i(\boldsymbol{\beta})| = \sum_{i=1}^{N} |Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}|.$$

The value of $\boldsymbol{\beta}$ that minimises the above sum is known as the L_1-estimator. The generalisation of this is known as M–estimation and will be introduced in Section 2.3.1. Note that the L_1-estimator can be shown to be the maximum likelihood estimator if the error distribution is double exponential. For $m = 0$, i.e., one-dimensional location, the L_1-estimator of β_0 is the median.

(ii) Instead of minimising the sum in (2.20), which is equivalent to minimising the average squared residuals, we could minimise another measure of "average residual" like the median, i.e. minimise

$$MED\{e_i(\boldsymbol{\beta})^2, i = 1, \ldots, N\} = MED\{(Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2, i = 1, \ldots, N\},$$

where MED means taking the median value of all $e_i(\boldsymbol{\beta})^2$. This is known as the *least median of squares (LMS)* method.

How LMS works is best understood in the case of a simple linear regression. The above minimisation problem then corresponds to finding the narrowest strip covering half the observations (the width of the strip is measured in the vertical direction). The LMS regression line lies in the middle of this band.

Effect of leverage points. Before we discuss robust methods in more detail we first illustrate in more detail the deficiencies of the least squares estimator. The argument given above applies to regression outliers, but least squares is also affected by leverage points. To see this, consider the case of a simple linear regression. Remember that the least squares estimator for the slope β_1 can then be written as

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x}) y_i}{\sum_k (x_k - \overline{x})^2} = \sum_i v_i y_i,$$

with $v_i = (x_i - \overline{x}) / \{\sum_k (x_k - \overline{x})^2\}$. Hence $\hat{\beta}_1$ is a weighted sum of the y_i where the largest weights are given to observations that have large $(x_i - \overline{x})$, which are exactly the leverage points. Similarly, it can be shown that R^2 is affected by leverage points in the sense that good leverage points improve the model fit, i.e. increase R^2, whereas bad leverage points decrease R^2. As we will see later, M–estimators are not much better than least squares in dealing with leverage points but the LMS–estimator is.

Identifying outliers: The methods presented in Section 2.2.3 to identify outliers were based on the leverage h_{ii} and Cook's distances. These work best when there is only one clear outlier. In case of clusters of outliers these methods of outlier identification may themselves be so much distorted that it is not possible to determine the outliers. This is because Cook's statistic is still based on $\hat{\boldsymbol{\beta}}_{(i)}$, the least squares estimator when observation i is deleted, and this is still affected by the other outlieres. Further h_{ii} is based on $\overline{\mathbf{x}}$ and the empiricial covariance matrix $\hat{\Sigma}_x$ as can be seen from its relation the the Mahalanobis distance in (2.13). Both, $\overline{\mathbf{x}}$ and $\hat{\Sigma}_x$ are extremely sensitive to oultiers in the x-values.

Furthermore, a common problem in outlier detection is the **masking effect**, i.e. an outlier is not detected because another outlier is present. As soon as the latter is removed from the data, we detect new outliers and this can go on and on. An alternative approach to outlier

identification is to first calculate a very robust estimator and use the residuals from the fitted model to find outliers.

Efficiency: Robust methods are specifically designed to be only slightly affected by small deviations from the model assumptions (or few outliers) and to not be catastrophically affected by large deviations (or many outliers). The price for this is that if the model assumptions for least squares to be optimal are satisfied then the robust methods based on the above ideas are not as efficient as least squares. This means that their asymptotic variance is larger than the one of the least squares estimator when normality holds, see below.

2.3.1 M–estimation

(i) M–estimation is based on the following idea: instead of minimising the sum of the squared residuals (2.20), another function of the residuals is minimised. This function is denoted by $\rho(\cdot)$ and called the *objective function.* It is assumed to satisfy the following criteria, for $u \in I\!R$,

- positivity, i.e. $\rho(u) \geq 0$;
- it is zero for zero-residual, $\rho(0) = 0$;
- symmetry, i.e. $\rho(u) = \rho(-u)$
- increasing for increasing residuals, i.e. $\rho(u) \geq \rho(u')$ if $|u| \geq |u'|$.

(ii) Assuming that we know σ (which in practice we do not, see below), the *M–estimator* $\hat{\boldsymbol{\beta}}_M$ is then defined as the value $\boldsymbol{\beta}$ that minimises

$$\sum_{i=1}^{N} \rho\left(\frac{e_i(\boldsymbol{\beta})}{\sigma}\right) = \sum_{i=1}^{N} \rho\left(\frac{Y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}{\sigma}\right). \tag{2.21}$$

in $\boldsymbol{\beta}$. Note that the least squares and the L_1-estimator are M-estimators as well, with ρ being the square, absolute value function, respectively. In these cases, the estimator is independent of knowing σ, because σ can be drawn out of the ρ as a constant factor, which does not make a difference for minimisation.

(iii) Differentiating (2.21) w.r.t. $\boldsymbol{\beta}$ and setting it to zero yields that the M–estimator $\hat{\boldsymbol{\beta}}_M$ has to satisfy

$$\sum_{i=1}^{N} \mathbf{x}_i \psi\left(\frac{Y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{\sigma}\right) = 0, \tag{2.22}$$

where $\psi(\cdot)$ is the derivative of $\rho(\cdot)$. Note that (2.22) is a set of equations because $\mathbf{x}_i$ is a vector.

(iv) The equations (2.22) cannot usually be solved analytically. By setting $u_i = (Y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})/\sigma$ we can rewrite (2.22) as

$$\sum_{i=1}^{N} \mathbf{x}_i w_i u_i = 0, \tag{2.23}$$

where $w_i = \psi(u_i)/u_i$. This gets us back to a situation similar to weighted least squares (cf. Section 2.1.3). If the weights w_i were fixed we could calculate $\hat{\boldsymbol{\beta}}_M$ as solution of

$$\hat{\boldsymbol{\beta}}_M = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{Y},$$

where $\mathbf{W} = diag\{w_1, \ldots, w_N\}$. However, the weights w_i depend on the $\hat{e}_i$ and hence on $\hat{\boldsymbol{\beta}}_M$, so that an iterative procedure has to be used to compute $\hat{\boldsymbol{\beta}}_M$: start with an initial estimate $\hat{\boldsymbol{\beta}}_M^{(0)}$, compute $\mathbf{W}^{(0)}$ and then

$$\hat{\boldsymbol{\beta}}_M^{(k+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{W}^{(k)})\mathbf{Y} \tag{2.24}$$

until convergence.

(v) The resulting values $w_1, \ldots, w_N$ can be interpreted as *robustness weights* and can be used for outlier identification. An observation is treated (and can be interpreted) as an outlier by the M-estimator if w_i is very small (zero in case of the Bisquare objective function introduced below).

(vi) From (2.22) it can be seen that the residuals $\hat{e}_i(\boldsymbol{\beta})$ only enter through $\psi(u_i)$. If ψ is bounded, the influence of large residuals on the regression estimator is bounded as well, though the influence of leverage points is not bounded because of the factor $\mathbf{x}_i$, unless $\psi(u) = 0$ or $\psi(u)$ very small for those observations with large $\mathbf{x}_i$.

Distribution of M–estimators and efficiency

The sampling distribution of the least squares estimator under normal errors can be computed precisely. However, finding the distribution of general M-estimators is much more complicated. Generally, only asymptotic results can be given. These also hold for non-normal error distributions (and can be applied to the least squares estimator as well, because it is an M-estimator).

Under fairly general conditions, M–estimators can be shown to be consistent. A necessary (but not sufficient) condition for this is that

$$\mathrm{E}\left[\psi(Z)\right] = 0$$

where $Z \sim F$ and F is the error distribution, which holds for all symmetric error distributions if ψ is bounded.

Further, under certain regularity conditions, one can show that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}) \overset{n\to\infty}{\sim} \mathcal{N}(0, V(\psi, F)L^{-1}),$$

where $V(\psi, F)$ is a (quite complicated) matrix that depends on the influence function ψ and on the true error distribution F, and $L = \lim \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X}$. This can be used to compute (approximated) tests and confidence intervals.

Under normality, remember that the (finite sample and asymptotic) covariance matrix of the least squares estimator is $\mathbf{C} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. For M-estimators with many other ρ-functions, the covariance matrix can be shown to be $V(\psi, F)L^{-1} = b\mathbf{C}$ with some constant b. The factor $\frac{1}{b}$ is then called "efficiency" of the estimator, compared to the least squares estimator under normality. Assuming that the asymptotics hold approximately for finite samples, b can be interpreted as the factor with which the number of observations has to be multiplied to arrive at the same precision as the least squares estimator, i.e. an estimator with efficiency 0.5 needs twice as much observations.

Note, however, that for non-normal error distributions, particularly those with heavy tails, the least squares estimator is often much worse than M-estimators with bounded ψ.

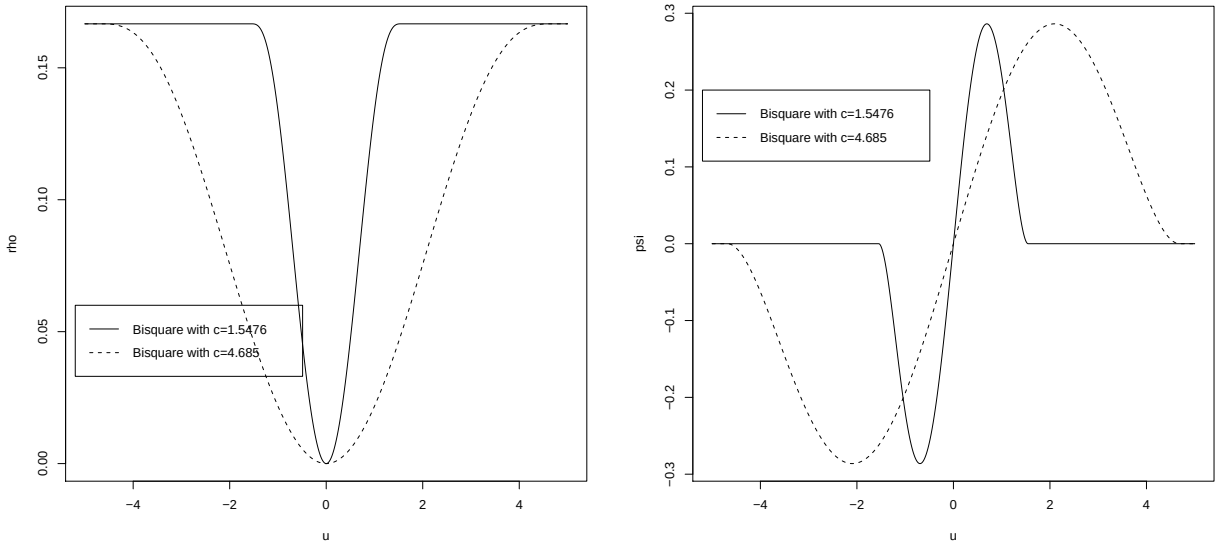

Figure 2.1: Bisquare ρ and ψ-functions for $c = 1.547$ and $c = 4.685$.

Objective functions

$\rho(u) = u^2$ and $\psi(u) = 2u$ yield the LS-estimator. With $\rho(u) = |u|$, we get the L_1-estimator. ψ then is the sign function. According to (vi) above, the L_1-estimator is not robust against leverage points. An alternative is the so-called "**Bisquare**" objective function:

$$\rho_B(u) = \begin{cases} \frac{1}{6}\left\{1 - \left[1 - \left(\frac{u}{c}\right)^2\right]^3\right\} & \text{if } |u| \leq c, \\ \frac{1}{6} & \text{if } |u| > c, \end{cases}$$

where c is a tuning constant, yielding

$$\psi_B(u) = \begin{cases} \frac{u}{c^2}\left[1 - \left(\frac{u}{c}\right)^2\right]^2 & \text{if } |u| \leq c, \\ 0 & \text{if } |u| > c, \end{cases}$$

Note that the Bisquare ψ-function is 0 for large residuals, as required for robustness against leverage points. c can be used to tune robustness and efficiency. For small $|u|$, $\psi_B(u)$ is approximately equal to a line, which is proportional to the ψ-function for the LS-estimator (and ρ is very similar to the simple squared funtion, see Figure 2.3.1). If c is large, most u-values are small and most residuals are treated in a very similar way as by the LS-estimator. For this reason, for $c \to \infty$, the efficiency of the Bisquare-M-estimator converges to 1. On the other hand, c should not be too large in order to correct for extreme outliers. If $\sigma^2 = 1$, $c = 4.658$ is the usually suggested default value.

2.3.2 The S-estimator

A problem with the Bisquare-M-estimator is that it is dependent on σ, which is unknown. Here the aim is to obtain a regression estimator that is robust, efficient, and independent of

knowledge of σ. An **M-estimator of scale** for observations $\mathbf{z} = (z_1, \ldots, z_N)$ (each z_i assumed to have an expected value of 0) can be defined by finding $s_M(\mathbf{z})$ so that

$$M(\mathbf{z}) = \frac{1}{N}\sum_{i=1}^{n} \rho\left(\frac{z_i}{s_M(\mathbf{z})}\right) = E_{\mathcal{N}(0,1)}\rho(Z), \tag{2.25}$$

where ρ fulfills the same conditions as in 2.3.1 (i). Note that if $\rho(u) = u^2$, $E_{\mathcal{N}(0,1)}\rho(Z) = 1$ and $s_M^2(\mathbf{z}) = \frac{1}{N}\sum_{i=1}^{N} z_i^2$, which is the ML estimator of the normal variance σ^2. The motivation is that $s_M(\mathbf{z})$ is chosen in order to scale $z_1, \ldots, z_N$ by dividing by $s_M(\mathbf{z})$ so that they correspond, "on ρ-average", to a standard normal distribution.

From this, a linear regression estimator $\boldsymbol{\beta}_S$, the so-called S-estimator, can be obtained by choosing $\boldsymbol{\beta}$ in order to minimise $s_M\{e_1(\boldsymbol{\beta}), \ldots, e_N(\boldsymbol{\beta})\}$; define

$$S_M = s_M\{e_1(\boldsymbol{\beta}_S), \ldots, e_N(\boldsymbol{\beta}_S)\} = \min s_M\{e_1(\boldsymbol{\beta}), \ldots, e_N(\boldsymbol{\beta})\}.$$

The S-estimator can be computed by an algorithm alternating (2.24) to find $\boldsymbol{\beta}^{(k+1)}$ for given $s_M^{(k)}(\mathbf{z}^{(k)})$ and $\boldsymbol{\beta}^{(k)}$ and solving (2.25) for $s_M^{(k+1)}(\mathbf{z}^{(k+1)})$ for given $\mathbf{z}^{(k+1)} = \left\{e_1(\boldsymbol{\beta}^{(k+1)}), \ldots, e_N(\boldsymbol{\beta}^{(k+1)})\right\}$. Note that this may only lead to a local optimum and depends on the starting parameter $\boldsymbol{\beta}^{(0)}$. The S-estimator may be found by taking the optimum over several runs of this algorithm started by LS-estimators computed from very small random subsets (with $p+1$ points, say) of the dataset.

2.3.3 The MM-estimator

The MM-estimator has been introduced by Yohai (1987, Annals of Statistics) and is still the "state of the art" in robust linear regression. It is based on combining the two principles of robust S-estimation and M-estimation. It is the default estimator computed by the R-function `lmrob` in package `robustbase`.

S_M can be used as an estimator of σ^2. Plugging in the S-scale estimator S_M for σ^2, the MM-estimator is an M-estimator and can be iterated according to (2.24), starting from the S-estimator.

The robustness weights given in 2.3.1 (v) can be used for outlier identification. S_M and the robustness weights are computed by the R-function `lmrob` along with the MM-estimator, regression parameter standard errors, and t- and p-values based on the asymptotic theory for the MM-estimator under the normal model.

Note, however, that the MM-estimator can be more influenced by moderate outliers than the S-estimator and therefore it is, to some extent, less robust.

2.3.4 Optional reading for Section 2.3

- Sections 1 to 3 and 6 of the book: Rousseeuw and Leroy (1987), Robust regression and outlier detection. Wiley.
- Seber G.A.F., Lee A.J. (2003), Linear Regression Analysis. 2nd edition, Wiley, Section 3.13.
- Maronna R.D., Martin R.D., Yohai V.J. (2006), Robust Statistics: Theory and Methods. Wiley.

2.4 Variable selection

Examples: R-output 7, 8

We consider now again specifically multiple regression models of the type

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + e_i, \qquad (i = 1, \ldots, N)$$

where Y_i is the response Y corresponding to values $x_{i1}, \ldots, x_{im}$ of m explanatory variables $x_1, \ldots, x_m$. Variable selection[3] means that we want to reduce the number of explanatory variables to those that are 'really relevant'. Mathematically, the problem is to find out whether one or more of the coefficients $\beta_1, \ldots, \beta_m$ are precisely zero, though in practice, depending on the research aim, it may be desirable to exclude explanatory variables which may have a very small but nonzero contribution as well.

There are various reasons for variable selection:

- If the number of regression parameters p is large compared to the number of observations N, $\mathbf{X}^T\mathbf{X}$ is often close to collinearity and the estimation of the parameters can be very unstable. Getting rid of some variables (and therefore some parameters) can improve the situation. (A rule of thumb is that 10-20 observations are needed per estimated regression parameter in order to estimate them accurately enough.)

- If data analysis is carried out for explanatory reasons, often a clear and simple explanation can be communicated and understood more easily. Therefore it is often desired to base the explanation on only a small number of variables. But of course this is only sensible as long as all really important variables are kept in the model.

- Sometimes, if the aim is prediction, it may be expensive (or, e.g. in some medical applications, risky or painfull) to observe some of the explanatory variables and it would be beneficial if it turned out to be unnecessary to measure some of the variables in future in order to still achieve an acceptable predictive accuracy. (Note, however, that in such a situation often not all variables are equally costly. Individual tests whether the most expensive variables are needed may then be more sensible than the application of general methods of variable selection.)

There are also arguments against variable selection:

- In terms of predictive accuracy, as long as enough observations are available, it is usually much worse to leave out variables that are really important than to keep variables in the model of which the true coefficients are (about) zero. The reason is that with enough observations the true zero coefficients will be estimated to be about zero anyway. Therefore, variable selection should not be done unless there is a real need for it (i.e. one of the reasons given above).

- If some of the variables in the full model are highly correlated (i.e. they measure more or less the same thing), usually not all of these variables are needed, but the decision about precisely which variable should be kept in the model and which should be left out is quite arbitrary. Therefore, variable selection is often unstable and the chosen

[3]Variable selection is one aspect of the more general problem of model selection, and referred to as "model selection" in some texts. Some of the methods introduced here, such as cross-validation, can be used for more general model selection issues such as deciding about the best transformation of some variables.

variables have to be interpreted with care concerning explanation and causal inference. Particularly, it cannot be taken for granted that a variable left out by variable selection is "really unimportant", because it may just be "represented" by another variable still in the model.

- Absolute values of regression parameter estimators become biased high because many methods select them *conditionally on being large* (at least indirectly). For the same reason, p-values become biased low.

2.4.1 t–tests and F–tests for model selection

Before we address some model selection methods we should review how the tests described earlier in Section 2.2.1 have to be interpreted.

- The F-test in the anova table in 2.2.1 (vii) is for testing that *none* of the explanatory variables affect the response.

- The F-test in Section 2.2.1 (vi) is for testing the omission of ν of the explanatory variables *given that the other* $m - \nu$ *explanatory variables are in the model.*

- The t-test in Section 2.2.1 (v) is for testing that the j^{th} explanatory variable x_j does not affect the response *given that the other* $m - 1$ *explanatory variables are in the model.* So this is a test that a partial regression coefficient is 0.

- Computer output usually gives t-statistics and associated p–values against each of the regression coefficients. You should not deduce from these that you can omit several variables: if any of them give non-significant results, you can only select one of these to drop the corresponding explanatory variable from the model, e.g. the 'most non-significant' (highest P-value) – this is the basis of a procedure called 'backward elimination' to be described later.

  The reason why not several non-significant variables should be dropped simultaneously is that if for example two variables are highly correlated, one of them is needed in the model, but every single one can be dropped as long as the other one is still there. In such a situation, the t-tests for both variables may be non-significant, but if both are left out simultaneously, some important information is lost. Analogous situations can be caused by correlations between combinations of variables, so that the problem cannot always be detected by looking at pairwise correlations.

- *You should also note that if you drop a variable, the estimates of the coefficients of the remaining variables DO CHANGE in general.* Similarly if you add a variable. The estimates remain unchanged only if you have orthogonal columns in the $\mathbf{X}$-matrix (as already commented on at the end of Section 2.2.2).

- F-tests and t-tests may be useful for deciding whether some individual variables are needed or not (particularly if they are expensive to measure, see above). However, it should be kept in mind that a non-significant test result does not mean that the true coefficient is zero. Particularly if the sample size is not large, tests may have poor power and even really important variables may not necessarily yield low p-values. Therefore, omitting a non-significant variable always risks to lose useful information for prediction.

2.4.2 Best subset selection

With m explanatory variables, there are 2^m possible regression models. Best subset selection means that all these models are compared and the best one is chosen. Therefore, it is crucial how to decide which model is best.

A starting point is to find the best models for each number of explanatory variables $k = 1, 2, \ldots, m$. For fixed k, the best model can be defined as the model with the smallest RSS, which can easily be shown to be equivalent with choosing the model that minimises $\hat{\sigma}^2$, maximises R^2, and minimises the p-value of the F-test of the null hypothesis that $\beta_1 = \ldots = \beta_m = 0$ against at least one of the parameters of the chosen model to be nonzero.

It is less straightforward to compare the best models for different k, because adding more explanatory variables to a model *always* decreases the RSS and therefore increases R^2 ($\hat{\sigma}^2$ is *not* necessarily decreased, though, because it multiplies the RSS by a factor dependent on the number of explanatory variables). The reason is that for a given set of variables the least squares estimator always minimises the RSS. If a new variable is added to a model, it is always possible to obtain a smaller RSS by minimising it at the new model than if the coefficient for the new variable is fixed at zero (which is equivalent to not having the variable in the model) - even if the true regression parameter for the new variable is zero.

Note that the best model with $l > k$ variables does not necessarily contain all k variables of the best model with k variables. Therefore, t- and F-tests cannot always be used to compare the best subsets of different sizes. The following section introduces methods to compare models with different numbers of explanatory variables.

Note that, because the data is often compatible with more than one model, it may be advisable to look for more than one good model.

2.4.3 Criteria to compare models

Several criteria have been suggested in the literature that take into account the number of explanatory variables in the model. In the present course, only two of them are introduced explicitly. More can be found, e.g., in Hastie, Tibshirani and Friedman (2001).

- **Akaike Information Criterion (AIC):** for quite general models fitted by maximum likelihood,

  $$\text{AIC} = -2\,\hat{\ell}(\text{model}) + 2p$$

  where, for the model under consideration, $\hat{\ell}(\text{model})$ is the maximum of the log-likelihood function and p is the number of regression parameters in the model (including the intercept), i.e., $p = k + 1$ for a model with k explanatory variables.

  AIC is reported in the R-software. Because a good model has a high likelihood and (hopefully) not many parameters, we look for models with the smallest AIC.

  Under the assumptions of this chapter,

  $$\hat{\ell}(\text{model}) = -\frac{N}{2}\log(\sigma^2) - \frac{\text{RSS}}{2\sigma^2} + \text{constant}.$$

  σ^2 is usually unknown and can be estimated by its MLE $\hat{\sigma}^2_{ML} = \text{RSS}/N$. Hence[4]

  $$\text{AIC} = N\log\left(\frac{\text{RSS}}{N}\right) + 2p.$$

[4]Recall $\hat{\sigma}^2_{ML} \neq \hat{\sigma}^2$, the standard estimator of σ^2, see Section 2.1.2. $\hat{\sigma}^2_{ML}$ is used here because the underlying theory of the AIC is based on maximising the likelihood.

The AIC is a so-called "penalised likelihood-method", because the term $2p$ can be interpreted as a penalty for too large models, correcting for the fact that the RSS is always decreased by increasing the model. For fixed k, the AIC just chooses the model with the smallest RSS.

- **Other functions of the RSS** penalising large models have been suggested, such as Mallows C_p and the so-called "adjusted R^2", which delivers a number between 0 and 1 like R^2 but is maximised if $\hat{\sigma}^2$ is minimised. This is, however, a very "soft" criterion which often does not reduce the number of variables enough (or even not at all).

- **Leave-one-out cross validation** (LOO-CV) is a very general method to compare different models. Its aim is to select the model that minimises an estimator of the expected prediction loss.

  The principle is as follows: one point is left out of the dataset, the model is fitted on the remaining $N-1$ points. The left out point is then predicted from the fit and the prediction error is computed (which we can do, because we know the true value). This is done for every single point in the dataset.

  More formally:

  1. Like for Cook's distance, calculate $\hat{\boldsymbol{\beta}}_{(i)}$ based on the data where observation $(y_i, \mathbf{x}_i)$ has been omitted.
  2. 'Predict' the omitted observation by $\hat{y}_{(i)} = \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(i)}$.
  3. Calculate the 'loss', i.e. a suitable measure of distance between the actual value y_i and the predicted value $\hat{y}_{(i)}$. This is denoted by a loss function $L(y_i, \hat{y}_{(i)})$. The most common loss functions are
     - absolute error loss function $L(y_i, \hat{y}_{(i)}) = |y_i - \hat{y}_{(i)}|$;
     - squared error loss function $L(y_i, \hat{y}_{(i)}) = (y_i - \hat{y}_{(i)})^2$.

     Other choices of L are possible.
  4. Do steps 1.–3. for each observation $i = 1, \ldots, N$, and then calculate the average loss as
     $$\frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{y}_{(i)}).$$

  Different models can then be compared with regard to their average loss and you would choose the one with the smallest loss.

  A major advantage of LOO-CV is that it is not based on any model assumption (though data are still required to be i.i.d. conditionally on the explanatory variables). Even if the fitted model is wrong (e.g. if a linear model is fitted to an essentially nonlinear real situation), LOO-CV still yields a reliable estimator of the expected prediction loss.

  A disadvantage of LOO-CV is that in most situations the model has to be fitted N times[5], which is computationally demanding if N is large.

[5] In multiple linear regression and some other situations, there are formulae which can be used to get all N fits easily from the original fit with all points.

2.4.4 Stepwise methods

In the following, 'variable' refers to 'explanatory variable'. Best subset selection has a major drawback if the number of variables is large, because the number of candidate models 2^m quickly becomes computationally untractable.

Stepwise selection is a very general principle that can be applied to a wide range of selection problems in order to reduce the number of models to be fitted, from which then an optimal model is selected.

There are two basic approaches:

- **Backward elimination.** Start by fitting the full model, i.e., with all variables. Call this the backward m-model. Set $k = m$.

  (a) Fit all k models with $k-1$ variables, i.e., the models with all but one of the variables of the backward k-model.

  (b) Choose the best of these models, i.e., the one with minimal RSS, as the backward $k-1$-model.

  (c) Set $k = k-1$. Go back to (a) unless $k = 0$.

  Like best subset selection, this produces a sequence of "best" models for every possible number k of variables. The total number of models to be fitted is much lower, namely $m + (m-1) + \ldots + 1 = m(m+1)/2$.

  After the backward sequence of models has been found, one of the criteria from Section 2.4.3 can be chosen to select the optimal model.

  Because backward elimination produces a "nested" sequence of models, i.e., the smaller models are always submodels of the larger ones, it is possible to use t- or F-tests to compare the models as an alternative. This means that after step (b) the p-value for the test comparing the backward k-model (H_1) with the backward $k-1$-model (H_0) is computed and the algorithm is stopped with the backward k-model as the final chosen model if this model is significantly better than the backward $k-1$-model (because this indicates that no variable should be removed from the backward k-model). This is the way how backward elimination is most often done in practice, though many statisticians hold that this is in most situations worse than optimising LOO-CV or AIC. If tests are used anyway, it has been suggested to use a significance level of 0.1 or even 0.2 instead of the usual 0.05 as stopping criterion in order to prevent too many potentially informative variables from being deleted.

- **Forward selection.** Start with $k = 0$, i.e., without variables, only fitting the mean to the data. Call this the forward 0-model.

  (a) Fit all $m-k$ models with $k+1$ variables, i.e., the models with all variables of the forward k-model plus a single new one.

  (b) Choose the best of these models, i.e., the one with minimal RSS, as the forward $k+1$-model.

  (c) Set $k = k+1$. Go back to (a) unless $k > m$.

  This produces another sequence of "best" models for every possible number k of variables. The total number of models to be fitted is again $m + (m-1) + \ldots + 1 = m(m+1)/2$ and the criteria mentioned above can be used to select the best model.

Again, a nested sequence of models is produced and t- or F-tests can be used (even though this is not necessarily good), this time stopping the algorithm if the larger model does not lead to significant improvement.

Some general remarks about stepwise methods

- Stepwise methods produce heuristically reasonable sequences of models, but they are not guaranteed to find the best subset model, because not all possible subsets are evaluated. Therefore, best subset selection is better as long as it is computationally feasible.
- Backward elimination and forward selection are not guaranteed to arrive at the same sequence of models or at the same "best" model.
- Stepwise methods are often very unstable, i.e., small changes in the data can lead to vastly different selected models. This holds to some smaller extent for best subset selection as well.
- Experience suggests that backward elimination is superior to forward selection in most cases in terms of prediction quality.
- If the number of observations is *very* small compared to the number of variables ($N < 2m$, say), the initial model of backward elimination is very unstable and forward selection is preferable. In some situations (for example in genetics) datasets with $N < m$ occur. In such cases, backward elimination cannot be performed (nor can any regression model with more than $N - 2$ variables be fitted), but forward selection still works for small enough k.

2.4.5 The lasso

The "lasso" is an alternative method for variable selection and generally for dealing with situations close to collinearity such as a too large number of variables. The lasso is more stable than the selection methods mentioned above and leads often to a lower prediction error. However, it introduces some (moderate) bias and can be outperformed in "clear-cut" situations (i.e. if all true regression parameters are either very close to or very far away from zero). For computing the lasso estimator, the original explanatory variables (which are denoted by $z_1, \ldots z_m$ here) are transformed first so that the new transformed explanatory variables have a mean of 0 and a standard deviation of 1, i.e.,

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j},\ i = 1, \ldots, N,\ j = 1, \ldots, m,$$

where

$$\bar{z}_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij},\ s_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(z_{ij} - \bar{z}_j)^2}.$$

The benefit of this transformation is that because the variations of the x-variables are standardised, their regression parameters are of comparable size and quantify the relative contribution of the variables to the linear regression (though this should not be interpreted as "relative importance" of the variables because dependencies between them may exist). Therefore, standardisation of the predictors is sometimes applied not only because the lasso should be applied.

It can be shown that the least squares estimator for β_0 is $\frac{1}{N}\sum_{i=1}^{N} Y_i$ if all explanatory variables are centered (i.e., have average 0).

The lasso estimator $\hat{\boldsymbol{\beta}}_L$ of $\boldsymbol{\beta}$ is defined by minimising the sum of squared errors

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_m x_{im})^2$$

under the constraint that

$$\sum_{i=1}^{m} |\beta_i| \leq t, \tag{2.26}$$

where t is a pre-chosen tuning constant.

- If $\hat{\boldsymbol{\beta}}$ is the usual least squares estimator, and $t_0 = \sum_{i=1}^{m} |\hat{\beta}_i|$, it follows that

  $$t \geq t_0 \Rightarrow \hat{\boldsymbol{\beta}}_L = \hat{\boldsymbol{\beta}},$$

  because $\hat{\boldsymbol{\beta}}$ minimises $S(\boldsymbol{\beta})$ and fulfills (2.26) if $t \geq t_0$.

- The recommended method to choose t is to compute $\hat{\boldsymbol{\beta}}$ first and then to compute the lasso estimator for $t = ct_0$, $c = 0.1, 0.2, \ldots, 1$ (say), and to choose the best value of t from these by LOO-CV. Instead, c could be just fixed at some value, for example 0.5 (the smaller c, the more β_is will be estimated to be zero, and the fewer variables will be selected; therefore c should be smaller if the number of variables, compared to the number of observations, is very large).

- Assume a situation in which some variables in the model are really much less important than some others. It can be shown that under $t < t_0$, minimising $S(\boldsymbol{\beta})$ while fulfilling (2.26) means that the estimated lasso regression parameters $\hat{\beta}_{Li}$ are reduced to zero for the unimportant variables first, before "shrinking" the absolute values of the parameters corresponding to the important variables. Therefore, usually some estimated lasso parameters become *exactly zero*, which means that the lasso really selects variables.

- Note that β_0 is not covered by (2.26), and it can therefore be estimated by least squares (see above) independently of the slope parameters. In most applications, there is either no particular reason why β_0 should be estimated as 0, or for subject matter reasons $\beta_0 = 0$ is known and does not have to be estimated anyway.

- The absolute values of the β_i are required to be of comparable size because they are aggregated by $\sum_{i=1}^{m} |\beta_i|$ in (2.26). Therefore, standardisation is necessary.

- Unless $t \geq t_0$, the computation of the lasso estimator is complicated (though it can be relatively fast) and is not treated in this course. See Section 3.5 of Hastie, Tibshirani and Friedman (2001) for references.

- The lasso is known as a **shrinkage method** because (2.26) effectively forces the absolute values of the $\hat{\beta}_{Li}$ to be smaller than they would be under unrestricted least squares. Apart from forcing some of the $\hat{\beta}_{Li}$ to be exactly zero, there is another benefit of shrinkage. It can be shown that for standardised explanatory variables, stepwise and best subset selection methods tend to choose variables of which the β_i-parameters are estimated to have a large absolute value. This effectively introduces a bias by overestimating the corresponding $|\beta_i|$ (except in "clear-cut" situations, see above), so that a better estimation can be achieved by shrinking them. Unfortunately it is strongly dependent on the true values of $\boldsymbol{\beta}$ and σ^2 how much shrinkage is needed, and therefore there is no clear theoretical guideline how to choose the t.

  There are other methods such as Ridge regression. This will be discussed during the problem class if time permits.

2.4.6 Optional reading for Section 2.4

- Krzanowski Section 3.4
- Hastie, Tibshirani and Friedman, Chapter 3.
- Sections 12.1 – 12.4 of Seber G.A.F., Lee A.J. (2003), Linear Regression Analysis. 2nd edition, Wiley.

Specialist books on regression modelling

If you require much more detail about regression analysis, there are many texts that specialise on this topic. Among them are:

- Chatterjee S., Hadi A.S., Price B. (2000), Regression Analysis by Example. 3rd edition, Wiley.
- Harrell F.E. jr. (2001), Regression Modeling Strategies. Springer.
- Montgomery D.C., Peck E.A., Vining C.G. (2001), Introduction to Linear Regression Analysis. 3rd edition, Wiley.

2.5 Analysis of variance

Example: R-output 9

The theory of linear models in Section 2.1 applies to situations in which the response Y is modelled as dependent on categorical variables or, in other words, group memberships. These models are often called ANOVA-models, though the term ANOVA ("analysis of variance") rather refers to breakdown of the corrected total sum of squares CTSS into a sum of the RSS of a full model plus some sum of squares contributions explained by some (or all) of the regression parameters, see Sections 2.2.1 (vii) and 2.2.5, 2. Sums of squares can be seen as quantifying variation and the usual variance estimators under normality are actually sums of squares divided by some constant. Such decompositions play a stronger role for ANOVA models with categorical predictors.

ANOVA models are treated in more detail with examples in STAT0029. In the present course, only the connection to the general theory of the linear model is presented.

The special case 6 of Chapter 2 **(comparison of several groups/one-way layout)**

$$Y_{ij} = \mu_i + e_{ij} \qquad (i = 1, \ldots, I;\ j = 1, \ldots, n_i)$$

is the easiest ANOVA model. It has already been shown that this can be written down by use of indicator variables as explanatory variables, i.e., the indicator of the first group is 1 for all observations in the first group and 0 for all other observations, the indicator of group i is 1 for all the observations in group i and 0 for all others and so on.

In matrix notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.27}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ . \\ . \\ . \\ Y_{In_I} \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ . \\ . \\ . \\ \mu_I \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ . \\ . \\ . \\ e_N \end{pmatrix},$$

the columns of $\mathbf{X}$ denoting the I indicator variables, the rows the n observations. This notation makes it possible to apply all the theory in Section 2.1. It can actually be shown that the least squares estimator defined by (2.7) amounts to computing all the group means. The F-test in Section 2.1.4 is usually applied to test the $H_0: \ \mu_1 = \mu_2 = \ldots = \mu_I$ (using RSS_0 =CTSS) against a model where at least two of the group means differ.

A more complicated model is the **two-way layout** where the observations are classified according to two factors (such as "urban/suburban", "Maths/Lib. arts course"). In casewise notation:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \qquad (i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, n_{ij}), \tag{2.28}$$

where μ is interpreted as the overall expected value, α_i is the effect of level i of the first factor, β_j is the effect of level j of the second factor and γ_{ij} is an interaction effect (nonzero if the effect of level i of the first factor for particularly those observations having level j of the second factor differs from the average overall effect of the two factors).

(2.27) can be applied again. (2.28) suggests an $\mathbf{X}$-matrix with a first column consisting of ones (corresponding to μ), I columns corresponding to indicator variables for the levels of the first factor (as above), J columns corresponding to levels of the second factor and IJ columns corresponding to the interactions (1 only for observations belonging to level i for the first factor *and* level j for the second factor), i.e., $1 + I + J + IJ$ columns in total.

This could be multiplied by

$$\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_{11}, \ldots, \gamma_{IJ})^T.$$

For the simplest case $I = J = n_{ij} = 2$ for each i and j, here is how it looks like (omitting $\mathbf{e}$):

$$\mathbf{Y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{21} \\ \gamma_{22} \end{pmatrix}.$$

However, it can be shown that $\mathbf{X}^T\mathbf{X}$ is not invertible, because there are too many parameters[6]. The problem here is not a too small number of observations but the fact that, however large the number of observations, only IJ cell means (i.e. means corresponding to all possible combinations of the two factors) can be estimated and therefore more than IJ parameters are not supported.

This is usually resolved by introducing constraints. A possible set of constraints (but not the only one) is

- $\sum_{i=1}^{I} \alpha_i = 0$; one constraint,
- $\sum_{j=1}^{J} \beta_j = 0$; one constraint,
- $\sum_{i=1}^{I} \gamma_{ij} = 0,\ j = 1, \ldots, J; \sum_{j=1}^{J} \gamma_{ij} = 0,\ i = 1, \ldots, I;\ I + J$ constraints of which one is redundant, because if all but one of these sums are 0 it can be shown that the last one has to be zero as well.

In terms of matrix algebra, the first constraint means that $\alpha_1 = -\sum_{i=2}^{I} \alpha_i$. The column belonging to level 1 is omitted from $\mathbf{X}$, α_1 is omitted from $\boldsymbol{\beta}$, and for the observations of level 1 of the first factor, because α_1 is replaced by $-\sum_{i=2}^{I} \alpha_i$, there are matrix entries -1 in the columns corresponding to $\alpha_2, \ldots, \alpha_I$. And so on with the other constraints (the columns of the resulting $\mathbf{X}$-matrix are no longer simple indicator vectors; this could be achieved by constraining some

[6]Using some knowledge from linear algebra, this can for example be seen by verifying that all I indicator variables corresponding to the α_i sum up to a vector of ones, which is already the first column of $\mathbf{X}$ corresponding to μ, proving linear dependence and $\mathbf{X}$ not being of full rank.

parameters to be zero instead). Eventually, a matrix with $1+I+J+IJ-2-(I+J-1)=IJ$ columns results so that $\mathbf{X}^T\mathbf{X}$ is invertible and the least squares estimator can be computed.

Using the situation $I=J=n_{ij}=2$ as an example again, this yields

$$\mathbf{Y}=\begin{pmatrix} Y_{111}\\ Y_{112}\\ Y_{121}\\ Y_{122}\\ Y_{211}\\ Y_{212}\\ Y_{221}\\ Y_{222}\end{pmatrix},\qquad \mathbf{X}=\begin{pmatrix} 1 & -1 & -1 & 1\\ 1 & -1 & -1 & 1\\ 1 & -1 & 1 & -1\\ 1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1\\ 1 & 1 & -1 & -1\\ 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1\end{pmatrix},\qquad \boldsymbol{\beta}=\begin{pmatrix}\mu\\ \alpha_2\\ \beta_2\\ \gamma_{22}\end{pmatrix}.$$

All the other parameters can be obtained from the constraints.

Again, all the theory in Section 2.1 can be applied. F-tests are particularly used in many applications to find out whether

- all interactions could be ignored (with RSS_0 computed from a model where all $\gamma_{ij}=0$),
- "all $\alpha_i=0$" is compatible with the data,
- "all $\beta_j=0$" is compatible with the data.

The ANOVA table decomposes the CTSS (which is computed from a model in which only the overall mean μ is fitted by $\bar{Y}$) into the RSS of the full model plus sums of squares explained by the first factor. the second factor, and the interactions. Unless all n_{ij} are equal (the so-called "balanced" design), this depends on the (somewhat arbitrary) order of the factors.

For variable selection, most methodology in Section 2.4 can be applied, though in most application there is a particular order in which terms are removed in a stepwise backward fashion (or, the other way round, added in a stepwise forward fashion). Usually it is reasonable to have the original factors in the model if there are interactions in the model involving these factors (and lower-order interactions in the presence of higher-order interactions). For example, in a two-way layout with backward selection, usually

- it is checked whether removal of *all* interaction terms improves the model. If not, the model is not reduced. Otherwise, after removal of all interactions,
- it is checked removal of which of the two factors (i.e. *all* parameters belonging to that factor) is better, and whether this improves the model. If so, after removal of this factor,
- it is checked whether removal of the other factor improves the model even further.

These checks can be done using F-tests, AIC or LOO-CV. In more complicated layouts, this may become more complicated, because for example removing all second-order interactions in a three-way layout may be compared to removing one of the factors plus all interactions in which it is involved in the first step, so that quite complicated stepwise strategies are conceivable.

2.5.1 Optional reading for Section 2.5

One of the following:

- Rice Sections 12.2 and 12.3.
- Dobson Section 6.4.
- Krzanowski Sections 4.1 to 4.5.

Chapter 3

Generalised Linear Models

Generalised linear models extend the ideas underlying linear models to situations when the response has binomial, Poisson, gamma and other distributions that belong to the **exponential family of distributions**, which will be discussed in more detail in Section 3.1.

A **g**eneralised **l**inear **m**odel, or **GLM** for short, consists of three parts or components:

- a **random** component: independent observations $Y_1, \ldots, Y_N$;
- a **systematic** component: the linear predictor, denoted by η_i for the i^{th} observation $(i = 1, \ldots, N)$;
- a **link** between the random and systematic components through the use of a link function g, i.e $g(\mu_i) = \eta_i$ where $\mu_i = \mathrm{E}(Y_i)$. (μ will be used as generic symbol for $\mathrm{E}(Y)$ throughout.)

Using the earlier notation, $\eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$.

The function g is called the **link function**: it describes how the **expected** response is linked to (i.e. depends on) the explanatory variables or factors. It is assumed to be a monotonic, differentiable function.

The name GLM is usually used in connection with responses whose distributions are of the exponential family.

Some special cases

1. **Linear model.**
   The linear model discussed in previous sections is the special case $g(\mu) = \mu$, called the **identity link**.

2. **Binary data: linear logistic model.**

   Suppose the i^{th} observation consists of a Bernoulli trial, with outcome $Y_i = 1$ (a "success") or 0 ("a failure").

   If π_i is the associated probability of a success, then under a linear logistic regression model,

   $$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}. \tag{3.1}$$

   In this case $\mu_i = \mathrm{E}(Y_i) = \pi_i$ and so the left hand side of the above equation has (on dropping the subscript i) the link function

   $$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{\mu}{1-\mu}\right) = g(\mu).$$

The quantity $\log\left(\frac{\pi}{1-\pi}\right)$ is called the **logit** of π and the link function in this case is called the **logit** link.

Notice that for $0 < \pi < 1$, then $-\infty < \log\left(\frac{\pi}{1-\pi}\right) < \infty$ with the consequence that there are no constraints on the unknown parameters in $\boldsymbol{\beta}$. This simplifies estimation.

Binomial data. Now suppose that the i^{th} observation corresponds to a specified number $n_i (\geq 1)$ of independent Bernoulli trials with the same $\mathbf{x}_i$. The distribution of the number of successes Y_i is $\text{Bin}(n_i, \pi_i)$ where $\mu = n\pi_i$. Then the linear logistic model given by equation (3.1) is still appropriate.

The link function expressed as a function of μ is now (again dropping the subscript i)

$$g(\mu) = \log\left(\frac{\mu}{n-\mu}\right).$$

3. **Poisson data: log-linear model.**

   Now suppose the i^{th} observation consists of a Poisson count Y_i. If $\text{E}(Y_i) = \mu_i$, a log-linear regression model has

   $$\log(\mu_i) = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}. \tag{3.2}$$

   Here $g(\mu) = \log(\mu)$ is called the **log** link.

   Example: R-output 10

4. **Contingency tables.**
   You will see later in the course (Section 3.4) that log-linear modelling is also relevant for analysis of contingency table data, particularly for contingency tables that summarise observed frequencies from combinations of the levels of more than two variables.

3.1 Exponential families of distributions

We will assume that the response variable is a random variable Y whose pdf (if continuous r.v.) or pmf (if discrete r.v.) depends on parameters θ and ϕ and has the form

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{3.3}$$

where a, b, c are known functions.
In the above, $a(\phi) = \frac{\phi}{w}$ where w is a known weight, and ϕ is the **dispersion parameter** or **scale parameter**, which for some distributions is known and others is unknown. As a dispersion parameter we require $a(\phi) > 0$.

Examples: the common distributions considered in the previous examples (normal, binomial, Poisson) are all special cases: all of these have $w = 1$ and so $a(\phi) = \phi$, and the other functions are displayed in Table 1 below.

Table 1

Distribution	θ	ϕ	$b(\theta)$	$c(y,\phi)$
Poisson(μ)	$\log(\mu)$	1	e^{θ}	$-\log y!$
Bin(n,π)	$\log\left(\frac{\pi}{1-\pi}\right)$	1	$n\log(1+\mathrm{e}^{\theta})$	$\log\begin{pmatrix} n \\ y \end{pmatrix}$
N(μ,σ^2)	μ	σ^2	$\frac{1}{2}\theta^2$	$-\frac{1}{2}[\frac{y^2}{\phi}+\log(2\pi\phi)]$

(Poisson case shown in lecture.)

If ϕ is known, (3.3) defines a one-parameter **exponential family of distributions**.
If ϕ is unknown, (3.3) defines an **exponential dispersion family of distributions**.

Note: The exponential representation given in Dobson's book is

$$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\} \tag{3.4}$$

where a, b, c, d are known functions. The form (3.4) does not explicitly include a scale parameter, as is the case in (3.3). The distributions $\mathcal{N}(\mu,\sigma^2)$ where σ^2 is known, Bin(n,π) and Poisson(μ) all have $a(y) = y$ in (3.4) when (3.4) is said to have the canonical form. Equivalently, if ϕ is known, (3.3) defines a one-parameter canonical exponential family.

The exponential representation given by (3.3) is that used in Krzanowski, McCullagh and Nelder, Garthwaite *et al.*, and Venables and Ripley's S-Plus book.
The representation (3.3) will be used in this course.

3.1.1 Properties of exponential families

Assume that the pdf or pmf of Y is given by (3.3).

(i) **Mean:** $\mathrm{E}(Y) = b'(\theta)$ (Recall that $\mathrm{E}(Y) = \mu$ and that μ is a function of $\boldsymbol{\beta}$.)
where $'$ denotes differentiation with respect to θ.

Proof: In the following $f(y;\theta,\phi)$ is abbreviated to $f(y)$. Also, if Y is a discrete random variable, $f(y)$ is a pmf and integration is replaced by summation.

As $1 = \int_{-\infty}^{\infty} f(y)dy$ then diffentiating both sides of this equation with respect to θ gives

$$0 = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(y)dy = \int_{-\infty}^{\infty}\frac{\partial f(y)}{\partial\theta}dy = \int_{-\infty}^{\infty}\frac{y-b'(\theta)}{a(\phi)}f(y)dy. \tag{3.5}$$

Assuming $a(\phi)\neq 0$, then $0 = \int_{-\infty}^{\infty} yf(y)dy - b'(\theta) = \mathrm{E}(Y) - b'(\theta)$, giving the result.

(ii) **Variance:** $\mathrm{var}(Y) = b''(\theta)a(\phi)$

Proof: Differentiate the last expression in (3.5) again gives

$$0 = \int_{-\infty}^{\infty}\left\{\frac{[y-b'(\theta)]^2}{a(\phi)}f(y) - b''(\theta)f(y)\right\}dy = \frac{\mathrm{var}(Y)}{a(\phi)} - b''(\theta),$$ giving the result.

(iii) **Variance function:** from the previous two results, the variance of Y can be written as

$$V(\mu)a(\phi) \qquad \text{or} \qquad V(\mu)\frac{\phi}{w}$$

where $V(\mu)$ is called the **variance function**. (Note that a can be any function of ϕ, and there would not be any difficulty in dealing with any form of a, when ϕ is known. However, when ϕ is unknown matters are awkward, unless we write $a(\phi) = \phi/w$, where $w = 1$. This form covers all the cases considered in this course.)

Examples:

Distribution	$V(\mu)$
Poisson(μ)	μ
Bin(n, π)	$\mu(n-\mu)/n$
$N(\mu, \sigma^2)$	1

(iv) **Canonical links:** these occur when $g(\mu) = \theta$.
From Table 1, it is immediately obvious that the log link, logit link and identity link are the canonical links for the Poisson, binomial and normal distributions, respectively (as in the examples given at the start).

3.1.2 Optional reading for Section 3.1

One of the following:

- Dobson Chapter 3.
- Krzanowski Section 5.2.
- Garthwaite *et al.* Section 10.2.
- McCullagh and Nelder Section 2.2.

3.2 Some GLM theory

Assume that the responses $Y_1, \ldots, Y_N$ are independent from a distribution with pdf or pmf given by (3.3), i.e.

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}.$$

Note that we are assuming a common parameter ϕ for all observations. Let $\mathrm{E}(Y_i) = \mu_i, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and x_{ij} be the j^{th} element of $\mathbf{x}_i^{\mathrm{T}}$. So then the linear predictor for the i^{th} observation is $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$.

3.2.1 Estimation for GLMs

(i) **Likelihood equations**

If ℓ denotes the log-likelihood function given data $y_1, \ldots, y_N$, then the likelihood equations are

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{N} \frac{y_i - \mu_i}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij} = 0 \tag{3.6}$$

for $j = 1, \ldots, p$. Note the absence of ϕ in the likelihood equations!

Proof: As seen by the examples at the start of this chapter, θ_i is a function of μ_i, which in turn is related to the elements of $\boldsymbol{\beta}$ through the link function. The unknown parameters are $\boldsymbol{\beta}$ and ϕ. Let ℓ denote the resulting log-likelihood function given data $y_1, \ldots, y_N$ on $Y_1, \ldots, Y_N$, i.e.

$$\ell = \sum_{i=1}^{N} \ell_i, \text{ where } \ell_i = \log f(y_i; \theta_i, \phi) = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi). \tag{3.7}$$

The following shows the steps in obtaining the likelihood equations

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \beta_j} = 0 \qquad \text{for } j = 1, \ldots, p.$$

- Firstly, for $i = 1, \ldots, N$,

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)},$$

  as $\mu_i = b'(\theta_i)$ (from Section 3.1.1).

- Then

$$\frac{\partial \ell_i}{\partial \mu_i} = \frac{\partial \ell}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} = \frac{y_i - \mu_i}{a_i(\phi)} \frac{1}{b''(\theta_i)} = \frac{y_i - \mu_i}{\text{var}(Y_i)},$$

  where $\text{var}(Y_i) = \phi b''(\theta_i)$ (from Section 3.1.1). Recall that $d\mu_i/d\theta_i = b''(\theta_i) \Rightarrow d\theta_i/d\mu_i = 1/b''(\theta_i)$.

- Then

$$\frac{\partial \ell_i}{\partial \eta_i} = \frac{\partial \ell}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} = \frac{y_i - \mu_i}{\text{var}(Y_i)} \frac{d\mu_i}{d\eta_i}.$$

- Finally, for $j = 1, \ldots, p$,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial \ell}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{N} \frac{y_i - \mu_i}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij}. \tag{3.8}$$

(ii) **Iterative procedure**

Estimation of $\boldsymbol{\beta}$

The likelihood equations (3.6) cannot be solved algebraically (see Exercise sheet 8). However, these equations are essentially the same as those that would have to be solved to find $\boldsymbol{\beta}$ by non-linear weighted least squares, if the weights $V(\mu_i)$ were known and independent of $\boldsymbol{\beta}$. This correspondence suggests an iterative method for finding the parameter estimates. This is the Fisher scoring method for which some details are included in the appendix to this section (optional reading).

Starting with a guess $\hat{\boldsymbol{\beta}}^{(0)}$ at the MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, successive values $\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \ldots$ are obtained until the iterative procedure converges according to some criterion (e.g. convergence of

these successive values or of the log-likelihood function to a desired accuracy) and the final value is the MLE $\hat{\boldsymbol{\beta}}$ (to the desired accuracy).

The iterative scheme for the estimates can be shown to be

$$(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{(s-1)}\hat{\boldsymbol{\beta}}^{(s)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z})^{(s-1)} \tag{3.9}$$

where the superscript $(s-1)$ in (3.9) indicates evaluation at $\hat{\boldsymbol{\beta}}^{(s-1)}$, $(s = 1, 2, \ldots)$, and, for $i = 1, \ldots, N$, $\mathbf{W}$ is the $N \times N$ diagonal matrix with $(i, i)^{th}$ element

$$w_{ii} = \frac{1}{V(\mu_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2,$$

and $\mathbf{z}$ has i^{th} element

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{d\eta_i}{d\mu_i}\right)$$

Proof: see appendix (optional reading).

Equation (3.9) looks like equation (2.10) for weighted least squares estimation but with weights given by w_{ii} and an 'adjusted response variable' given by z_i for the i^{th} observation. However the procedure is iterative and the weights and adjusted response have to be recalculated at each iteration. Also the iterative procedure can be started by putting $\mu_i^{(0)} = y_i$ instead of specifying a guess $\hat{\boldsymbol{\beta}}^{(0)}$. This method is known as Iteratively Reweighted Least Squares (IRLS). Note that, at convergence, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z}$ and can be seen as the minimizer of $||\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})||^2$.

In the case of normally distributed errors and identity link function,

$$\frac{d\eta_i}{d\mu_i} = 1, z_i = y_i$$

and hence the above procedure reduces to the non-iterative normal equations.

Estimation of ϕ

If a GLM contains a dispersion parameter ϕ whose value is unknown, we need to estimate it (this is not the case for the binomial and Poisson cases, for which we know that $\phi = 1$). Unfortunately, maximum likelihood estimators of dispersion parameters are biased (the most obvious example of this is the error variance in a linear model, for which the ML estimator is $N^{-1}\sum_{i=1}^{N}(Y_i - \mu_i)^2$).

An estimator of ϕ can be obtained from the Pearson statistic

$$X^2 = \sum_{i=1}^{N}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

The quantity X^2/ϕ is the sum of squares of zero mean and unit variance random variables, with $N - p$ degrees of freedom. This suggests that if the model is adequate then approximately $X^2/\phi \sim \chi^2_{N-p}$. So, setting the observed Pearson statistic to its expected value gives

$$\hat{\phi} = \frac{\hat{X}^2}{N - p}.$$

Note that $X^2 = ||\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}})||^2$, at convergence of $\mathbf{W}$ and $\mathbf{z}$.

(iii) **Large sample distribution of $\hat{\boldsymbol{\beta}}$**

Consider the one parameter case. To obtain the large sample distribution of $\hat{\beta}$, we use a Taylor expansion of the derivative of the log likelihood around the true parameter β_0 and evaluate this at $\hat{\beta}$. That is,

$$\left.\frac{\partial \ell}{\partial \beta}\right|_{\hat{\beta}} \approx \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta_0} + (\hat{\beta} - \beta_0) \left.\frac{\partial^2 \ell}{\partial \beta^2}\right|_{\beta_0},$$

which reduces to

$$(\hat{\beta} - \beta_0) \approx -\frac{\left.\frac{\partial \ell}{\partial \beta}\right|_{\beta_0}}{\left.\frac{\partial^2 \ell}{\partial \beta^2}\right|_{\beta_0}},$$

with equality in the large sample limit (by consistency of $\hat{\beta}$). The numerator has expected value equal to zero and variance $\mathcal{I}$ (see next section) and is made up of a sum of *i.i.d.* random variables ℓ_i. Hence, by the central limit theorem, as $n \to \infty$, its distribution will tend to $\mathcal{N}(0, \mathcal{I})$. By the law of large numbers, the denominator tends to $\mathcal{I}$. Therefore, in the large sample limit $(\hat{\beta} - \beta_0)$ follows a $\mathcal{N}(0, \mathcal{I})$ random variable divided by $\mathcal{I}$. This implies that, as $n \to \infty$,

$$(\hat{\beta} - \beta_0) \sim \mathcal{N}(0, \mathcal{I}^{-1}),$$

which also generalises to parameter vectors

$$\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}_p(\boldsymbol{\beta}, \mathcal{I}^{-1}).$$

This result is exact in the case of normally distributed errors. See, e.g., Rice Section 8.5.2 for further details.

(iv) **Covariance matrix of $\hat{\boldsymbol{\beta}}$**

The (Fisher) information matrix is typically used to calculate the covariance matrix associated with maximum likelihood estimates. It is defined as the variance of the score (or alternatively as the expected value of the observed information, i.e. minus the expected value of the Hessian). Specifically, the $(j,k)^{th}$ element of the information matrix can be written as

$$\mathrm{E}\left(\frac{\partial \ell}{\partial \beta_j}\frac{\partial \ell}{\partial \beta_k}\right) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\frac{\mathrm{E}[(Y_i - \mu_i)(Y_{i'} - \mu_{i'})]}{V_i V_{i'}}\right)\frac{d\mu_i}{d\eta_i}\frac{d\mu_{i'}}{d\eta_{i'}}x_{ij}x_{i'k},$$

where $V_i = \mathrm{var}(Y_i)$. Now

$$\mathrm{E}[(Y_i - \mu_i)(Y_{i'} - \mu_{i'})] = \begin{cases} \mathrm{var}(Y_i) & = V_i \quad \text{for } i = i' \\ \mathrm{cov}(Y_i, Y_{i'}) & = 0 \quad \text{for } i \neq i' \end{cases},$$

hence

$$\mathrm{E}\left(\frac{\partial \ell}{\partial \beta_j}\frac{\partial \ell}{\partial \beta_k}\right) = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2$$

from which follows that

$$\mathcal{I} = \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}/\phi.$$

Based on the large sample approximation result discussed in the previous section, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is therefore given by the inverse of $\mathcal{I}^{-1}$, i.e. $(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\phi$.

It is not difficult to prove that $\mathrm{E}\left(\frac{\partial \ell}{\partial \beta_j}\frac{\partial \ell}{\partial \beta_k}\right) = -\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right)$ but this is beyond the scope of this course.

Because $\mathcal{I}$ is also defined as minus the expected value of the Hessian, information may be seen as a measure of the curvature of the log-likelihood near the maximum likelihood estimate of $\boldsymbol{\beta}$. A flat likelihood will have a low negative expected second derivative, and thus low information; while a sharp one would have a high negative expected second derivative and thus high information.

For distributions with known ϕ, the result of the previous section can be used directly to find confidence intervals for the parameters. If ϕ is unknown (e.g. for the gamma distribution), then it must be estimated and intervals have to be based on an appropriate t distribution, as shown for the linear model case.

3.2.2 Confidence intervals for model parameters

Let us assume that we have a GLM with *known* dispersion parameter. The case with unknown ϕ follows the same construction except that ϕ has to be estimated and an appropriate t distribution employed.

(i) **Single parameter**

From Section 3.2.1,

$$\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}_p(\boldsymbol{\beta}, \mathcal{I}^{-1}).$$

An estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is obtained after substituting the MLE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in the matrix $\mathbf{W}$ in $\mathcal{I}^{-1}$. Hence $\mathrm{se}(\hat{\beta}_j)$, the standard error of $\hat{\beta}_j$, is the square root of the estimated $(j,j)^{th}$ element of $\mathcal{I}^{-1}$.

An approximate $100(1-\alpha)\%$ confidence interval for β_j has limits

$$\hat{\beta}_j \pm z_{\alpha/2}\mathrm{se}(\hat{\beta}_j)$$

where $z_{\frac{\alpha}{2}}$ is the upper $100\frac{\alpha}{2}\%$ point of the standard normal distribution.

In the case of a linear model with normally distributed errors, the above result is exact if σ^2 is known.

(ii) **Linear combination of parameters**

Inferences about $\psi = \mathbf{c}^{\mathrm{T}}\boldsymbol{\beta}$ – an example is estimation of expected response (compare: Gauss–Markov theorem in Section 2.1.1 (v)). Let $\hat{\psi} = \mathbf{c}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$.

Use

$$\hat{\psi} \dot{\sim} \mathcal{N}(\psi, \mathbf{c}^{\mathrm{T}}\mathcal{I}^{-1}\mathbf{c}).$$

Proof: follows from results concerning a linear combination of normally distributed random variables (some discussion of this in Rice Section 14.4).

Hence $100(1-\alpha)\%$ confidence interval for ψ.

3.2.3 Hypothesis testing

(i) **Single parameter**

Now consider testing the null hypothesis H_0: $\beta_j = 0$ for some j. As before, remember that in the context of regression, H_0 tests whether the response depends on the associated explanatory variable, given the inclusion of the other explanatory variables in the model. We can use

$$\frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \dot{\sim} \mathcal{N}(0,1) \text{ under } H_0,$$

and hence obtain the P-value in the usual way (as described in the foundation fortnight). The usual changes apply when ϕ is unknown.

(ii) **Model comparison**

Suppose that we want to compare two models M_0 and M, where M_0 is a special case of M, and let $\ell(\hat{\boldsymbol{\beta}}_0)$ and $\ell(\hat{\boldsymbol{\beta}})$ be the maximized log-likelihoods of the two models. The null hypothesis of interest H_0 is that a subset of $p-q$ parameters out of the p parameters in the linear predictor are all 0 (leaving q non-zero parameters), whereas let H_1 denote the alternative hypothesis that all p parameters are not 0.

H_0 can be tested using a likelihood ratio test. That is, if H_0 is true then in the large sample limit,

$$2[\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_0)] \sim \chi^2_{p-q}. \tag{3.10}$$

(The derivation of this result is beyond the scope of this course.) If H_0 is false then M will mostly likely have a higher likelihood than M_0, hence -2 log-likelihood ratio would be too large to be consistent with the chosen χ^2 distribution.

This result applies to Poisson and binomial models, but not to, e.g., gamma where ϕ is not known. Before discussing this case, it is useful to introduce the concept of deviance.

(iii) **Deviance**

When fitting GLMs, it is useful to have a quantity that is similar to the residual sum of squares in a linear model context. This is the deviance and is defined as

$$D = 2[\ell(\hat{\boldsymbol{\beta}}_{sat}) - \ell(\hat{\boldsymbol{\beta}})]\phi,$$

where $\ell(\hat{\boldsymbol{\beta}}_{sat})$ denotes the maximized likelihood of the saturated model, i.e. the model with one parameter per datum. $\ell(\hat{\boldsymbol{\beta}}_{sat})$ is based on setting $\hat{\mu}_i = y_i$ and represents indeed the highest value that the likelihood can possibly have.

The scaled deviance is defined as

$$D^* = D/\phi,$$

which does depend on the dispersion parameter. Obviously, for the binomial and Poisson distributions, the deviance and scaled deviance are the same. Given result (3.10), under the proposed model,

$$D^* \dot{\sim} \chi^2_{N-p}.$$

In the above, the dispersion parameter is assumed to be known so that D^* can be calculated.

Example: normally distributed errors with known constant variance. Here

$$D^* = \frac{\text{RSS}}{\sigma^2}$$

Proof: in lecture. Under H_0: proposed model, $D^* \sim \chi^2_{N-p}$ (exactly: see Section 2.1.2 (iv)).

Examples of deviances (the $\hat{\mu}_i$'s are the fitted values under the proposed model):

Distribution	Deviance
Normal	$\frac{1}{\sigma^2}\sum_{i=1}^{N}(Y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^{N}[Y_i \log(Y_i/\hat{\mu}_i) - (Y_i - \hat{\mu}_i)]$
	$2\sum_{i=1}^{N} Y_i \log(Y_i/\hat{\mu}_i)$ if constant term included in linear predictor
Binomial	$2\sum_{i=1}^{N}[Y_i \log(Y_i/\hat{\mu}_i) + (n_i - Y_i)\log\{(n_i - Y_i)/(n_i - \hat{\mu}_i)\}]$

In the Binomial and Poisson cases, the (scaled) deviance statistics may be used to test for goodness-of-fit,

$$D^* \dot{\sim} \chi^2_{N-p} \text{ under proposed model.}$$

Given the definition of deviance, under H_0, likelihood ratio test (3.10) can be re-expressed as

$$D_0^* - D^* \dot{\sim} \chi^2_{p-q},$$

where, again, the dispersion parameter has to be known so that the deviances can be calculated.

(iv) **Model comparison with unknown ϕ**

Under H_0, we know that $D_0^* - D^* \dot{\sim} \chi^2_{p-q}$ and $D^* \dot{\sim} \chi^2_{N-p}$. If $D_0^* - D^*$ and D^* are treated as asymptotically independent, then under the null and in the large sample limit

$$F = \frac{(D_0^* - D^*)/p - q}{D^*/(N-p)} \sim F_{p-q,N-p}.$$

But this is equivalent to

$$F = \frac{(D_0 - D)/p - q}{D/(N-p)} \dot{\sim} F_{p-q,N-p},$$

hence allowing for model comparison when ϕ is unknown.

Other statistics for inference

Score statistic

The scores $U_1, \ldots, U_p$ are defined by

$$U_j = \frac{\partial \ell}{\partial \beta_j} \text{ for } j = 1, \ldots, p$$

where ℓ is the log-likelihood function.
Let $\mathbf{U} = (U_1, \ldots, U_p)^{\mathrm{T}}$. Properties of score statistics:

- **Expectation**:
$$\mathrm{E}(\mathbf{U}) = \mathbf{0}.$$
- **Covariance matrix**:
$$\mathbf{V}(\mathbf{U}) = \mathcal{I}$$
where $\mathcal{I}$ is the information matrix (see Section 3.2.2(i)).
- **Asymptotic sampling distribution**:
$$\mathbf{U} \sim \mathcal{N}_p(\mathbf{0}, \mathcal{I})$$
$$\mathbf{U}^{\mathrm{T}} \mathcal{I}^{-1} \mathbf{U} \sim \chi^2_p. \tag{3.11}$$

Proofs: see, for example, Dobson Section 5.2 (optional reading).

Wald statistic

If $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ then it can be shown that asymptotically

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \mathcal{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p, \tag{3.12}$$

where $\mathcal{I}(\hat{\boldsymbol{\beta}})$ is the information matrix evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.
Similar results to (3.11) and and (3.12) hold for a subset of the parameters – this involves using the corresponding submatrix of the covariance matrix. Hence the score statistic can be used as alternatives to the likelihood ratio test for multiple parameter hypothesis testing.

3.2.4 Optional reading for Section 3.2

One of the following:

- Dobson Chapters 4 and 5.
- Krzanowski Sections 5.3 and 5.4.
- McCullagh and Nelder Section 2.5.
- Garthwaite *et al.* Sections 10.3 and 10.4.

3.2.5 Appendix: Iterative procedure for solving likelihood equations

The MLE of the β_j's are obtained by equating the p expressions given by (3.8) to zero. It is not difficult to see that these equations cannot be solved algebraically and so a numerical method is used. The procedure used is the **Fisher scoring method** which is a modified version of the Newton-Raphson method.

The numerical method is iterative. For the Newton-Raphson method it is:

$$\hat{\boldsymbol{\beta}}^{(s)} = \hat{\boldsymbol{\beta}}^{(s-1)} - \mathbf{H}(\hat{\boldsymbol{\beta}}^{(s-1)})^{-1}\mathbf{h}(\hat{\boldsymbol{\beta}}^{(s-1)})$$

for $s = 1, 2, \ldots$, starting with a guess $\boldsymbol{\beta}^{(0)}$ at the position of the maximum, where

$\mathbf{h}(\boldsymbol{\beta})$ is a $p \times 1$ vector of the first order partial derivatives of ℓ with respect to the elements of $\boldsymbol{\beta}$ (the gradient vector), and $\mathbf{H}(\boldsymbol{\beta})$ is a $p \times p$ matrix of the second order partial derivatives of ℓ (the Hessian matrix). (Do not confuse this use of the notation $\mathbf{H}$ with that in Section 2.1.2)

For a maximum to be at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ say, the matrix $\mathbf{H}(\boldsymbol{\beta}^*)$ should be negative definite.

More detail on the background to the above procedure will be given in the STAT0030 course.

The Fisher Scoring method replaces the Hessian matrix by its expectation (with respect to the observations) which is the negative of the information matrix $\mathcal{I}$. So the iterative scheme is then

$$\hat{\boldsymbol{\beta}}^{(s)} = \hat{\boldsymbol{\beta}}^{(s-1)} + \mathcal{I}(\hat{\boldsymbol{\beta}}^{(s-1)})^{-1}\mathbf{h}(\hat{\boldsymbol{\beta}}^{(s-1)})$$

or

$$\mathcal{I}(\hat{\boldsymbol{\beta}}^{(s-1)})\hat{\boldsymbol{\beta}}^{(s)} = \mathcal{I}(\hat{\boldsymbol{\beta}}^{(s-1)})\hat{\boldsymbol{\beta}}^{(s-1)} + \mathbf{h}(\hat{\boldsymbol{\beta}}^{(s-1)}). \tag{3.13}$$

Using the result for $\mathcal{I}$, the j^{th} element of the vector on the right hand side of (3.13) is

$$\sum_{i=1}^{N}\sum_{k=1}^{N} w_{ii}x_{ij}x_{ik}\hat{\beta}_k^{(s-1)} + \sum_{i=1}^{N} w_{ii}(y_i - \mu_i)\left(\frac{d\eta_i}{d\mu_i}\right)x_{ij} = \sum_{i=1}^{N} x_{ij}w_{ii}z_i$$

where

$$z_i = \sum_{k=1}^{N} x_{ik}\hat{\beta}_k^{(s-1)} + (y_i - \mu_i)\left(\frac{d\eta_i}{d\mu_i}\right)$$

Hence the right hand side of (3.13) is $\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z}$ where $\mathbf{z} = (z_1, \ldots, z_N)^{\mathrm{T}}$ evaluated at $\hat{\boldsymbol{\beta}}^{(s-1)}$, and the iterative scheme can be written simply as

$$(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{(s-1)}\hat{\boldsymbol{\beta}}^{(s)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z})^{(s-1)} \text{ as stated in Section 3.2.1 equation (3.9).}$$

3.3 Binomial data and logistic regression

Examples: R-output 11, 13, 14

This section concentrates on a special case of generalised linear models as introduced on the previous section.

Suppose that N responses $Y_1, \ldots, Y_N$ are independent such that

$$Y_i \sim \text{Bin}(n_i, \pi_i)$$

where

$$g(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$$

for some link function g and $\mu_i = \mathrm{E}(Y_i) = n_i\pi_i$ $(i = 1, 2, \ldots, N)$.

This includes the binary case in which $n_i = 1$ for all i. Note that the models below are given in terms of π_i. This is equivalent to expression via μ_i as $\pi_i = \frac{\mu_i}{n_i}$.

The canonical link function is the logit link (see Section 3.1.1) which gives the **linear logistic model**

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}.$$

The logit link is usually the default in computer packages. Other models that are used in this context are:

Probit:

$$\Phi^{-1}(\pi_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}.$$

where the link function is the inverse cumulative $\mathcal{N}(0, 1)$ distribution function.

The probit model was one of the original models used for bioassay.

Complementary log-log:

$$\log[-\log(1-\pi_i)] = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}.$$

In fact in all of these cases, the expression for the probability of success, π, obtained by inverting the equation for the model, has been chosen to be a cumulative distribution function:

Logistic:

$$\pi = \frac{1}{1+\mathrm{e}^{-\eta}}, \text{ cdf of a logistic distribution.}$$

Probit:

$$\pi = \Phi(\eta), \text{ cdf of the standard normal distribution.}$$

Complementary log-log:

$$\pi = 1 - \exp[-\exp(\eta)], \text{ cdf of the extreme value distribution.}$$

where, in the current context, $\eta = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$ denotes the linear predictor for any observation. Note that these equations can be used for prediction, after $\boldsymbol{\beta}$ has been estimated by $\hat{\boldsymbol{\beta}}$, by plugging in $\hat{\eta} = \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ for η, giving the estimated probability that a new observation of Y_{N+1} is 1 given $\mathbf{x}_{N+1}$ (this assumes that the new observation can only be 0 or 1, i.e. $n_{N+1} = 1$; if $n_{N+1} = 1$ is larger, the prediction of Y_{N+1} needs to be multiplied by n_{N+1}. Based on these predictions and a suitable loss function, different link functions could be compared using Leave-One-Out Cross-Validation, which, however, needs large datasets because they often lead to quite similar models.

The above choices ensure that $0 \leq \pi \leq 1$, as required. Also the resulting link function has the property that $-\infty < g(\mu) < \infty$ with the consequence that there are no constraints on the unknown parameters in the linear predictor.

3.3.1 Interpretation of logistic regression parameters

If π denotes the probability of a success, then the logit link function $\log\{\pi/(1-\pi)\}$ is the logarithm of the odds on a success, or 'log odds' for short, which we will denote by $\text{logit}(\pi)$. So the coefficient β_j of an explanatory variable x_j in a logistic regression model measures the rate of change of the log odds with x_j, holding constant the values of the other explanatory variables in the model.

In particular, suppose that x_1 is an indicator variable with just two levels, 0 and 1 say, as would be the case if these values represented 'absence' and 'presence' of a factor. Then at $x_1 = 0$,

$$\text{logit}(\pi) = \beta_0 + \beta_2 x_2 + \ldots + \beta_m x_m,$$

and at $x_1 = 1$,

$$\text{logit}(\pi') = \beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_m x_m,$$

where the probability of success in the second case has been denoted by π'.

Subtracting these equations gives

$$\beta_1 = \text{logit}(\pi') - \text{logit}(\pi) = \log\{\pi'/(1-\pi')\} - \log\{\pi/(1-\pi)\} = \log\left\{\frac{\pi'/(1-\pi')}{\pi/(1-\pi)}\right\},$$

which is the logarithm of the ratio of the odds on a success at the two values of x_1, or logarithm of the **odds ratio**, holding constant the values of the other explanatory variables in the model.

So the odds on a success when $x_1 = 1$ is e^{β_1} times the odds on a success when $x_1 = 0$, holding constant the values of the other explanatory variables in the model.

3.3.2 Some basic results

The results described in Section 3.2 of the course notes apply to GLMs in general, so in particular to logistic regression (but also if you choose a probit or complementary log-log link). For logistic regression we have:

(i) **Likelihood equations**: these are given in general by equation (3.6) in Section 3.2.1(i), but for the logistic model they reduce to $\sum_{i=1}^{N}(y_i - \hat{\mu}_i)x_{ij} = 0$ for $j = 1, \ldots, p$.

(ii) **Maximum likelihood estimates**: in general the MLE of $\boldsymbol{\beta}$ is obtained numerically by the iterative procedure (3.9) in Section 3.2.1 (ii): $(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{(s-1)}\hat{\boldsymbol{\beta}}^{(s)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z})^{(s-1)}$ for iterations $s = 1, 2, \ldots$.

For the logistic model, the $(i,i)^{th}$ element in the diagonal matrix $\mathbf{W}$ is $w_{ii} = \frac{1}{V_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2$,

where $V_i = \text{var}(Y_i) = V(\mu_i) = n_i\pi_i(1-\pi_i)$ and $\dfrac{d\mu_i}{d\eta_i} = n_i\pi_i(1-\pi_i)$, hence $w_{ii} = n_i\pi_i(1-\pi_i)$.

Also $\mathbf{z}$ has i^{th} element

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{d\eta_i}{d\mu_i}\right) \qquad \text{where} \qquad \frac{d\eta_i}{d\mu_i} = \frac{1}{n_i\pi_i(1-\pi_i)}.$$

(iii) **Sampling distribution of $\hat{\boldsymbol{\beta}}$:**

$$\hat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}_p(\boldsymbol{\beta}, (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}).$$

(iv) **Deviance** for testing the goodness-of-fit of a particular model is

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right]$$

where the $\hat{\mu}_i$'s are the fitted μ_i's under the model (using the convention that $0 \log 0 = 0$). If the model is true, $D \dot{\sim} \chi^2_{N-p}$, which provides a test statistic for a goodness-of-fit test.

(v) **Test for H_0: $\beta_j = 0$.**

$$\frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \dot{\sim} \mathcal{N}(0,1) \text{ under } H_0.$$

(from Section 3.2.3 (i)). This is a test for the omission of the j^{th} explanatory variable given the other explanatory variables in the model.

(vi) **Test for H_0: ν of the regression parameters $\beta_1, \ldots, \beta_m$ are 0**.

Here we are assuming that the linear predictor consists of a constant term and terms from m explanatory variables (as in Section 2.2 for multiple linear regression) and H_0 tests the omission of ν explanatory variables where $\nu \leq m$.

Let D_0 and D denote the deviances under H_0 and the maximal (full) models, respectively. The likelihood ratio test for H_0 is

$$D_0 - D \dot{\sim} \chi^2_\nu \text{ under } H_0.$$

Special case H_0: $\beta_1 = \beta_2 = \ldots = \beta_m = 0$, *i.e. none of the explanatory variables affect the response*. In this case, under H_0, the π_i's are all equal and the MLE of the common probability of a success is the observed proportion of successes (so no iteration is required). Let the resulting deviance be denoted by C.

The numerical results could be reported in the form of a table, an **analysis of deviance table**:

Source of variation	Deviance	df
Regression	$C - D$	m
Residual	D	$N - m - 1$
Total	C	$N - 1$

(vii) **Pearson chi-squared statistic**: an alternative test for goodness-of-fit.
An alternative goodness-fit-test is to use the Pearson chi-squared statistic, to be denoted by X^2, based on the following table of observed and fitted frequencies (the latter are usually called expected frequencies in introductory texts, but they are really their estimates which are called fitted values or, in this context, fitted frequencies):

Observation	1	2	...	N
Observed # successes	Y_1	Y_2	...	Y_N
Observed # failures	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$

with a similar table of fitted values in which Y_i is replaced by $\hat{\mu}_i = n_i\hat{\pi}_i$.

Aside: using a common notation o for observed frequency and e for fitted (expected) frequency, D and X^2 have the forms:

$$D = 2\sum o \log\left(\frac{o}{e}\right) \text{ and } X^2 = \sum \frac{(o-e)^2}{e}.$$

This form of the deviance D is often denoted by G^2.

The Pearson chi-squared statistic is, after some algebra,

$$X^2 = \sum_{i=1}^{N} \frac{(Y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

which, if the model fitted is true, has the same large sample distribution as the deviance D, i.e. χ^2_{N-p}. By using the Taylor Series expansion of $s\log(s/t)$ about $s = t$ up to the quadratic term, it can be shown that $D \approx X^2$ (Rice Section 9.6).

This large sample distribution for D and X^2 is likely to be poor if any of the fitted values in the $2 \times N$ table are small (recall foundation fortnight, Rice Section 9.6).

(viii) **Binary data.** If each observation has a different pattern of the explanatory variables, then $n_i = 1$ for all i and the responses are all binary (i.e. 0 or 1). The results for estimation still hold. However, the deviance can be shown to depend on the binary observations only through the fitted values and so is no good for assessing goodness-of-fit (proof in Krzanowski Section 6.3.4 – optional reading; note that all observations are 0 or 1 and can be fitted perfectly with zero variance under the saturated model, which is not informative).

Hosmer and Lemeshow proposed a test obtained by grouping observations into about $g = 10$ groups of observations with roughly the same number per group according to their predicted probabilities. This then gives a $2 \times g$ table like that illustrated in (vii) above from which X^2 is calculated. The large sample distribution of the resulting statistic if the model is true is suggested to be approximately χ^2_{g-2}. This test can also be used in the binomial case (some computer packages report it anyway, e.g. Minitab). This test may not be reliable in certain cases. Alternative tests are available; see "A comparison of goodness-of-fit tests for the logistic regression model", Statistics in Medicine, 16, 965–980, 1997.

3.3.3 Checking model adequacy

Besides considering the residual deviance D, model adequacy should also be checked by appropriate plots – cf Section 2.2.3 for linear models.

There are several forms of residuals in the binomial case.

- **Raw residuals**: $\hat{e}_i = Y_i - n_i\hat{\pi}_i$
  Interpretation: these show how well the raw data are fitted.

- **Pearson** or **chi-squared residuals**:

$$X_i = \frac{\hat{e}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

  so that the chi-squared statistic $X^2 = \sum_{i=1}^{N} X_i^2$.
  Interpretation: these standardise the raw residuals by the estimated standard deviation of Y_i, making them comparable in size.

- **Standardised Pearson residuals**:

$$r_{Pi} = \frac{X_i}{\sqrt{1-h_{ii}}}$$

where h_{ii}, the leverage for the i^{th} observation, is the i^{th} diagonal element of the hat matrix which for a GLM is

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{\frac{1}{2}}.$$

Interpretation: these are standardised so that their variance is 1. They are comparable in size as well but adjust for the location in $\mathbf{x}$-space. For mathematical comparison these are better than the X_i, but the X_i are more naturally interpreted in terms of which points are "well fitted".

- **Deviance residual**:

$$d_i = \mathrm{sign}(\hat{e}_i)\left\{2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right]\right\}^{\frac{1}{2}},$$

so that the deviance $D = \sum_{i=1}^{N} d_i^2$. The term $\mathrm{sign}(\hat{e}_i)$ gives d_i the same sign as $\hat{e}_i$.
Interpretation: these formalise how strongly the observation contributes to the deviance, i.e., to the standard way of measuring the quality of the overall fit (or rather misfit; the higher, the worse). They show to what extent the observation indicates that the model is violated and rather a saturated model is needed.

- **Standardised deviance residual**:

$$r_{Di} = \frac{d_i}{\sqrt{1 - h_{ii}}}.$$

Interpretation: this makes the d_i directly mathematically comparable by unifying their variance and adjusting for the location in $\mathbf{x}$-space.

- **Cook's statistic**:

$$D_i = \frac{1}{p}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$

(using the notation of Section 2.2.3 (vi)). The numerator of D_i is just the weighted sum of the squared differences of the fitted logits of the π_i's with and without the i^{th} observation with the w_{ii}'s as weights. However, the D_i's are calculated from

$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1 - h_{ii}}\right) r_{Pi}^2$$

Interpretation: quantifies the effect on $\hat{\boldsymbol{\beta}}$ of omitting observation no. i.

3.3.4 Model selection

Examples: R-output 13, 14

(i) Some computer packages include stepwise methods for the choice of variables in logistic regression.

In looking at best subsets, the number of parameters may be taken into account by use of the Akaike information criterion (AIC) – see Section 2.4.2. For binomial data,

$$\mathrm{AIC} = D + 2p + \mathrm{const}$$

where D is the deviance statistic and p is the number of parameters in the linear predictor of the model under consideration. LOO-CV can be used as well. const is the negative log-likelihood of the saturated model and can be ignored because it is always there regardless of the model for which the AIC is computed (this holds for the GLM in general).

(ii) Two models are compared by the difference of their deviances.
Suppose model M_1 with p_1 regression parameters is a submodel of M_2 with p_2 parameters. Let $\ell(\hat{\boldsymbol{\beta}}_j)$ denote the maximum value of the log-likelihood under model M_j and D_j denote the deviance for model M_j, $j = 1, 2$. We want to test $H_0 : M_1$ vs. $H_1 : M_2$.
Then the likelihood ratio test statistic -2log likelihood ratio is

$$2[\ell(\hat{\boldsymbol{\beta}}_2) - \ell(\hat{\boldsymbol{\beta}}_1)] = D_1 - D_2$$

which, under the null hypothesis, has approximately a $\chi^2_{p_2-p_1}$ distribution.

3.3.5 Analysis of deviance

The regression deviance $C - D$ (Section 3.3.2 (vi)) can be partitioned in the same way as for normal linear models, where RSS is replaced by the deviance D.

The differences in deviances are independent and approximately chi-squared distributed under the respective null hypotheses, with degrees of freedom equal to the respective differences in the number of parameters (see end of Section 3.3.4 (ii)).

3.3.6 Optional reading for Section 3.3

One of the following:

- Dobson Chapter 7.
- Krzanowski Chapter 6.
- McCullagh and Nelder Chapter 4.

Specialist books on modelling binary data

If you require much more detail about analysis of binary data and logistic regression, texts that specialise on this topic include:

- Collett D. (2002), Modelling Binary Data. Chapman & Hall.
- Hosmer D.W., Lemeshow S. (2000), Applied Logistic Regression. 2nd edition, Wiley.

3.4 Contingency tables: log-linear modelling

Example: R-output 15

Two-way contingency tables were revised in the foundation fortnight course (Rice Chapter 13). When introducing this topic, the main emphasis is on testing that the "row" and "column" variables are independent which, for large enough sample sizes, is done using Pearson's chi-squared statistic. We now consider a modelling approach to the problem of analysing contingency table data, including those in more than two dimensions. Note that with contingency tables there typically is no designated 'response' variable, the mutual interrelations among all variables are of interest. Thus we are not considering the usual case of regression models.

3.4.1 Background: two-way contingency tables

Firstly, it is helpful to reconsider two-way tables.

(i) Let the "row" variable be called A with I possible categories (i.e. the outcomes) and the "column" variable be called B with J possible categories.

(ii) Consider an $I \times J$ contingency table that has been obtained by allocating a random sample of N observations on the pair of variables A and B to the IJ possible combinations of the I categories of A and J categories of B.

(iii) For cell (i, j) of the contingency table $(i = 1, \ldots, I; j = 1, \ldots, J)$, let
π_{ij} = probability that an observation belongs to the cell,
μ_{ij} = expected frequency.

(iv) Recall the following from introductory probability:

(a) $\mu_{ij} = N\pi_{ij}$

(b) If the row variable A is independent of the column variable B then

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

where π_{i+} = probability observation belongs to row i and π_{+j} = probability observation belongs to column j. Here the + subscripts indicate summation over the corresponding subscript.

Then from (a) and (b), if A is independent of B,

$$\mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{N} \tag{3.14}$$

where $\mu_{i+} = N\pi_{i+}$ =expected frequency in row i and $\mu_{+j} = N\pi_{+j}$ =expected frequency in row j.

(v) **Model under null hypothesis.**
A typical null hypotheses is H_0 : 'A and B are independent.' We model this by taking logarithms in equation (3.14) to give

$$\log \mu_{ij} = \log \mu_{i+} + \log \mu_{+j} - \log N \tag{3.15}$$

We could rewrite equation (3.15) as any sum of three terms, one depending on i, another on j, and another being a constant. Thus

$$\log \mu_{ij} = \lambda + \alpha_i + \beta_j \tag{3.16}$$

Exercise: verify that (3.16) satisfies (3.14).

(vi) **Model under alternative hypothesis.**
If the variables A and B are not independent (i.e. the alternative hypothesis) then (3.14) does not hold and hence (3.16) does not hold. Let ϕ_{ij} then denote the difference between the left and right hand sides of equation (3.16) then, under the alternative hypothesis,

$$\log \mu_{ij} = \lambda + \alpha_i + \beta_j + \phi_{ij} \tag{3.17}$$

(vii) **Constraints on parameters.**
The linear predictor in (3.17) contains more parameters than cells in the contingency table and so is over-parameterised. As before (e.g. Section 2.5), this is overcome by putting constraints on the parameters to reduce the total number of 'independent' parameters to IJ. Two standard methods are: (i) set the parameters for one specific outcome to zero, e.g. for $\alpha_1 = 0, \beta_1 = 0, \phi_{1j} = 0, \phi_{i1} = 0$, this means that for every variable category '1' is regarded as a reference category; or (ii) make the parameters sum to zero, e.g. $\sum_i \alpha_i = 0$, $\sum_i \phi_{ij} = \sum_j \phi_{ij} = 0$.

(viii) The model (3.17) is a **saturated** log-linear model for the two-way contingency table. This is because there are IJ effective parameters (number of parameters minus number of effective constraints, cp. the discussion of model (2.28) and its constraints in Section 2.5), which is equal to the number of frequencies in the table, so every frequency can be fitted perfectly. The sub-model (3.16) is unsaturated. Estimation in (3.17) gives $\hat{\mu}_{ij} = n_{ij}$, i.e. there is a perfect fit and $G^2 = 0$.

(ix) Compare equation (3.17) with the model underlying the analysis of variance of a two-way table with normally distributed response (2.28), and we see that it is similar in character but with the logarithm of the expected response (frequency) on the left hand side instead of the expected response. So the model defined by (3.17) is a **log-linear model**.

(x) **Parameters.**
Note that the parameters of a loglinear model $(\lambda, \alpha_i, \beta_j, \phi_{ij})$ do not have a straightforward interpretation as regression parameters. It is mainly of interest whether certain parameters vanish or not as this will imply certain independencies (cf. Table 1).

(xi) The extra term introduced in (3.17) is an **interaction** term and, from comparing (3.16) and (3.17), the independence hypothesis is analogous to the hypothesis of no interaction. The usual method of testing a null hypothesis through log-linear modelling is based on the goodness-of-fit statistic

$$G^2 = 2 \sum o \log \left(\frac{o}{e} \right)$$

where o and e denote observed and fitted (expected) frequency, rather than the Pearson chi-squared statistic $X^2 = \sum \frac{(o-e)^2}{e}$ (see Section 3.3.2 (vii) for some discussion of these statistics).

(xii) **Simplified notation for models.**
To simplify the descriptions of the models in general, the notation used for linear predictors in the R computer package will be used from now on: the linear predictor on the right hand side of (3.16) will be written as $A + B$, and in (3.17) as $A + B + A{:}B$ or more simply $A * B$ (so ':' indicates interaction, '$*$' indicates include interaction and all lower order terms involving these factors). In using this notation, it is assumed that a constant term is included in the linear predictor.

3.4.2 Three-way contingency tables

(i) Suppose that the table is now formed by allocating observations from a random sample to the IJK categories formed from the combinations of the I categories of variable A, J categories of variable B and K categories of variable C.

(ii) **Example**: Data example E.

(iii) **Saturated model**
An obvious generalisation of (3.17) to model a three-way contingency table with expected frequency μ_{ijk} in cell (i, j, k) is to equate the log expected cell frequency to a linear predictor with the terms

$$A + B + C + A : B + A : C + B : C + A : B : C \tag{3.18}$$

The final term is an interaction term involving the categories of all three variables, so the predictor (3.18) can be abbreviated to A∗B∗C in the notation used by the R software. Like (3.17), the predictor (3.18) has too many parameters and so constraints are imposed on the terms so that there are only IJK parameters. Then (3.18) defines a saturated log-linear model for a three-way contingency table and we successively test the significance of the interaction terms, starting with A:B:C, i.e. we look for evidence that a submodel of (3.18) may adequately describe the data.

(iv) **Possible sub-models.**
The most common types of log–linear sub–models are summarised and interpreted in Table 1 below.

Table 1

Type	Predictor of log-linear model	Interpretation
1.	$A + B + C$	Complete independence of A, B, C
2.	$A + B + C + A : B$	(A, B) are jointly independent of C
3.	$A + B + C + A : B + A : C$	B and C are independent conditional on the category of A
4.	$A + B + C + A : B + A : C + B : C$	*Has no independence or conditional independence interpretation*

(v) **Interpretation of models.**
The interpretation of the type 1 predictor is an obvious extension of (3.16) to the three-way table.

The inclusion of A:B in the type 2 predictor indicates that A and B are not independent, and the absence of any interaction term with C suggests independence of C with the other two variables jointly.

The interpretation of the type 3 predictor is not so obvious, but the absence of a B:C interaction term suggests some form of independence for B and C. In fact it is independence of B and C conditional on the category of A.

Numerical illustration of the log-linear models

For three-way tables, Table 2 below demonstrates by example what the model types 1, 2 and 3 mean for the underlying probabilities (expressed as percentages for simplicity). In these tables, A_1 and A_2 indicate the categories of variable A; similarly for variables B

and C.

Table 2.

Type 1

		C_1	C_2
A_1	B_1	2	6
	B_2	8	24
A_2	B_1	3	9
	B_2	12	36

Type 2

		C_1	C_2
A_1	B_1	2	6
	B_2	8	24
A_2	B_1	5	15
	B_2	10	30

Type 3

		C_1	C_2
A_1	B_1	2	6
	B_2	8	24
A_2	B_1	30	15
	B_2	10	5

From Table 2 we observe the following:

- In the type 1 illustration (independence case), the proportions for C_2 to C_1 are in the same ratio over all combinations of categories of A and B. The same property applies if we switch the variables round in that statement. This also means that if we add together the percentages over the categories of any one variable, A say, we find that for each category of a second variable, B say, the proportions for C_2 to C_1 are still in the same ratio and so B and C are pairwise independent. Similar calculations show that A and B are pairwise independent, as are A and C.
- In the type 2 illustration (C independent of (A, B)), the proportions for B_2 to B_1 are not now in the same ratio over the categories of A, and so A and B are no longer independent. However, we still see that the proportions for C_2 to C_1 are in the same ratio over all combinations of categories of A and B and so C is still independent of A and B.
- In the type 3 illustration (conditional independence), we see that for each category of A separately, the proportions for C_2 to C_1 are the same for each category of B (3:1 for A_1 and 1:2 for A_2), i.e. for each category of A, we have that B and C are independent

(vi) **Algebraic explanation of the log-linear models**

Mathematical interpretations of the above types of model are given in Table 3 below. These are explained as follows (here the + subscripts indicate summation over the corresponding subscript, as usual, so, for example, π_{i++} is the marginal probability that A has category A_i).

- *Type 1: complete independence of A, B, C.*
  In this case the joint probability of an observation belonging to cell (i, j, k),

$$\pi_{ijk} = \Pr(A_i)\Pr(B_j)\Pr(C_k) = \pi_{i++}\pi_{+j+}\pi_{++k}$$

- *Type 2: joint independence of (A, B) with C.*

$$\pi_{ijk} = \Pr(A_iB_j)\Pr(C_k) = \pi_{ij+}\pi_{++k}$$

- *Type 3: conditional independence of (B, C) with A.*

$$\pi_{ijk} = \Pr(B_jC_k \mid A_i)\Pr(A_i) = \Pr(B_j \mid A_i)\Pr(C_k \mid A_i)\Pr(A_i)$$

  by the conditional indepedence assumption. Hence

$$\pi_{ijk} = \pi_{ij+}\pi_{i+k}/\pi_{i++}.$$

Table 3: summary of relationships.

	Probabilities	Expected values
Type 1:	$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$	$\mu_{ijk} = \mu_{i++}\mu_{+j+}\mu_{++k}/N^2$
Type 2:	$\pi_{ijk} = \pi_{ij+}\pi_{++k}$	$\mu_{ijk} = \mu_{ij+}\mu_{++k}/N$
Type 3:	$\pi_{ijk} = \pi_{ij+}\pi_{i+k}/\pi_{i++}$	$\mu_{ijk} = \mu_{ij+}\mu_{i+k}/\mu_{i++}$

You then follow the steps for getting from (3.14) to (3.16) in Section 3.4.1 to identify the corresponding log-linear model in the notation used in Table 1. For example, for type 1, taking logarithms on both sides in the formula for the expected values, then the right hand side is the sum of a constant term, a term depending on i, a term depending on j and a term depending on k, i.e. A + B + C in the notation used here. In the other models, if we include say A:B then we also include A + B (an example of the **hierarchical principle** whereby for the variables in each term in the linear predictor, you include all "lower order" terms involving those variables).

Note: the figures in Table 2 can also be thought of as expected values from a random sample of $N = 100$ observations (remember that expected value = $N\times$ prob for each cell of the table) and so they satisfy the relationships on the right hand side in Table 3.

(vii) **Higher dimensional contingency tables** The ideas generalise to four-way and higher-way tables, but there are many more models to consider. This will not be discussed in this course. If you wish to pursue this further refer to, for example, Section 6.4 of the text by Agresti (see texts at end of section).

3.4.3 Fitting the models

- **Sampling model.**
  The models are fitted using maximum likelihood estimation which requires the specification of the joint distribution of the observations, in this case of the observed frequencies. This joint distribution is the **multinomial distribution** (if you need to be reminded of this distribution, see Rice Section 3.2).

- It can be shown that the joint distribution of independent Poisson random variables conditional on their sum is a multinomial distribution (exercise). So, in practice, log-linear models for contingency table data are fitted as if the observed frequencies are independent Poisson variables.

- **Fitted values.** For some models, algebraic results for the fitted values exist (use the formulae for the expected values in Table 3 above, substituting observed marginal frequencies on the right hand sides). However, in order to obtain the G^2 values as well, it is simplest to fit all of the models of interest using the command for fitting a GLM in a computer package.

3.4.4 Goodness-of-fit and model checking

Due to the similarity between the multinomial distribution and the Poisson distribution mentioned above, the following methods are almost the same.

- **Goodness-of-fit** is done as for Poisson data using G^2 (cf. e.g. Section 3.4.1).

- **Nested models** are compared by differences of their G^2 values (i.e. differences of deviances).
- **Standardised residuals** are similar to the case of a logistic regression. E.g. for the case of two-way contingency tables we have
  - **Raw residuals**: $\hat{e}_{ij} = n_{ij} - \hat{\mu}_{ij}$
  - **Pearson** or **chi-squared residuals**:

$$X_{ij} = \frac{\hat{e}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

    so that the chi-squared statistic $X^2 = \sum_i \sum_j X_{ij}^2$.
  - **Standardised Pearson residuals**:

$$r_{Pij} = \frac{X_{ij}}{\sqrt{1 - h_{(ij)}}}$$

    where $h_{(ij)}$, the leverage for the observation with combination (i, j) of the two factors (it is a bit fiddly to write this in matrix notation, but as before you can find the leverages as diagonal elements of the Hat matrix.)
  - **Deviance residuals**:

$$d_{ij} = \text{sign}(\hat{e}_{ij}) \left\{ 2 \left[y_{ij} \log \left(\frac{y_{ij}}{\hat{\mu}_{ij}} \right) - (y_{ij} - \hat{\mu}_{ij}) \right] \right\}^{\frac{1}{2}}$$

    Note that the deviance $D = \sum_i \sum_j d_{ij}^2$.
  - **Standardised deviance residual**:

$$r_{Dij} = \frac{d_{ij}}{\sqrt{1 - h_{(ij)}}}.$$

3.4.5 Optional reading for Section 3.4

One of the following:

- Dobson Sections 9.3 to 9.7.
- Krzanowski Chapter 7.

Specialist books on modelling categorical data

If you require much more detail about analysis of contingency table data by log-linear modelling, this is included in the following specialist text:

Agresti A. (1996), An Introduction to Categorical Data Analysis. Wiley.

3.5 Generalized Additive Models (GAMs): An Overview

GAMs allow for flexible functional dependence of a response variable on covariates. Here we provide a brief overview of this flexible class of models by focusing on the penalized likelihood framework with regression splines.

3.5.1 Introduction

GAMs are becoming among the most useful and used of statistical methods in the fields of biology, ecology, economics, environmental science, epidemiology, genetics and medicine, to name a few. This approach extends traditional GLMs by allowing the determination of possible nonlinear effects of covariates on a response variable. GLMs model the effects of predictor variables x_{ij} in terms of a linear predictor of the form $\theta_0 + \sum_j \theta_j x_{ij}$, where the θ_j are regression parameters. GAMs replace $\theta_0 + \sum_j \theta_j x_{ij}$ with, for instance, $\sum_j f_j(x_{ij})$, where the f_j are unknown smooth functions of regressors. The use of smooth terms is crucial since the functional shape of any relationship is rarely known *a priori* and the response of interest may depend on the predictors in a complicated manner.

3.5.2 Model Structure

A GAM can be seen as a GLM with a linear predictor involving smooth functions of covariates

$$g\{\mathrm{E}(Y_i)\} = \eta_i = \mathbf{X}_i^* \boldsymbol{\theta} + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}, x_{i4}) + \ldots, \tag{3.19}$$

where $\mathbf{X}_i^*$ is the i^{th} row of $\mathbf{X}^*$, which is the model matrix for any parametric model components, with corresponding parameter vector $\boldsymbol{\theta}$, and the f_j are smooth functions of the covariates x_{ij}. The f_j are subject to identifiability constraints such as $\sum_i f_j(x_{ij}) = 0$ for each j.

Model (3.19) can flexibly determine the functional shape of the relationship between a response and some explanatory variables, hence avoiding the drawbacks of modelling data using parametric relationships. As an example, let us consider a group of patients from a single hospital who underwent Coronary Artery Bypass Graft surgery. One may wish to identify the risk factors of in-hospital mortality following surgery, where the outcome of interest is *Status* (0=alive, 1=died) and the explanatory variables associated with surgical mortality could be *Age*, *BSA* (Body Surface Area), and *Ejection Fraction* (a measure of heart function summarized in the categories Good, Fair and Poor). In order to explain the in-hospital mortality following surgery from these explanatory variables, several model specifications can be adopted. A possibility would be to fit a GLM with linear predictor

$$\eta_i = \theta_0 + \theta_1 EF_{i,fair} + \theta_2 EF_{i,poor} + \theta_3 Age_i + \theta_4 BSA_i, \tag{3.20}$$

where θ_0 represents the baseline group $Ejection\ Fraction = Good$. But we do not know whether the variables *Age* and *BSA* really enter the model linearly, and (3.20) makes the assumption that such relationships are linear. Instead, one could employ a GAM where the linear predictor is given by

$$\eta_i = \theta_0 + \theta_1 EF_{i,fair} + \theta_2 EF_{i,poor} + f_1(Age_i) + f_2(BSA_i).$$

In this way the relationship between the in-hospital mortality and the continuous variables in the model can be determined flexibly.

The smooth terms can be represented using regression splines. In particular, the regression spline of a predictor is made up of a linear combination of known basis functions, $b_{jk}(x_j)$, and unknown regression parameters, β_{jk},

$$f_j(x_j) = \sum_{k=1}^{q_j} \beta_{jk} b_{jk}(x_j), \tag{3.21}$$

where j indicates the smooth term for the j^{th} explanatory variable, q_j is the number of basis functions, hence regression parameters, used to represent the j^{th} smooth term, and the subscript i is dropped for simplicity. Similarly, the regression spline of two covariates can be written as $f_{jp}(x_j, x_p) = \sum_{k=1}^{q_j} \beta_{jp,k} b_{jp,k}(x_j, x_p)$. As mentioned earlier on, in order to identify (3.19), each smooth component is subject to some identifiability constraint. Basis functions have to be chosen in order to come up with smooth component estimates. For instance, suppose that $f_1(Age)$ is believed to be a 3^{th} order polynomial. A basis for this space is $b_{11}(Age) = 1$, $b_{12}(Age) = Age$, $b_{13}(Age) = Age^2$ and $b_{14}(Age) = Age^3$. Here, expression (3.21) becomes

$$f_1(Age) = \sum_{k=1}^{4} \beta_{1k} b_{1k}(Age) = \beta_{11} + \beta_{12} Age + \beta_{13} Age^2 + \beta_{14} Age^3,$$

which can be easily estimated using standard regression techniques. The number of basis functions, q_j, determines the maximum possible flexibility allowed for a smooth term. For example, a q_j equal to 20 will yield a "wigglier" nonlinear estimate as compared to the estimate that can be obtained when this parameter is set to 10. It is worth observing that, although quite illustrative, polynomial bases are not very useful in practice. As the number of basis functions increases, polynomial bases become increasingly collinear. This yields highly correlated parameter estimators which may lead to high estimator variance and numerical problems. These issues can be overcome by using orthogonal polynomial bases. These tend to be very useful for situations in which interest focuses on properties of f in the vicinity of a single specified point, but when the questions of interest relate to f over its whole domain, polynomial bases have some problems (see R example *Orthogonal polynomial vs spline*). For these reasons, such basis functions should not generally be employed to model nonlinear relationships. As a practical solution, continuous variables can be categorized into groups based on intervals or frequencies. However, categorization has several disadvantages since it introduces the problem of defining cut-points and implies that the relationship between a response variable and a set of covariates is flat within intervals. To overcome all these issues, spline bases are typically used to determine flexibly the relationship between the continuous predictors and the outcome of interest. In fact, they avoid the disadvantages of categorization, are not as correlated as polynomial basis functions, have convenient mathematical properties and good numerical stability. Common choices for representing smooth functions include smoothing splines. These place knots at every data point, and are indeed sometimes referred to as full rank smoothers because the size of the spline basis is equal to the number of observations. However, such smoothers have as many unknown parameters as there are data which results in expensive computations. Regression spline bases are a valid alternative. Figure 3.1 illustrates a thin plate regression spline basis in one dimension.

3.5.3 Parameter estimation

In model (3.19), replacing the smooth functions with their regression spline expressions yields a GLM whose design matrix contains the spline bases representing the smooth components in

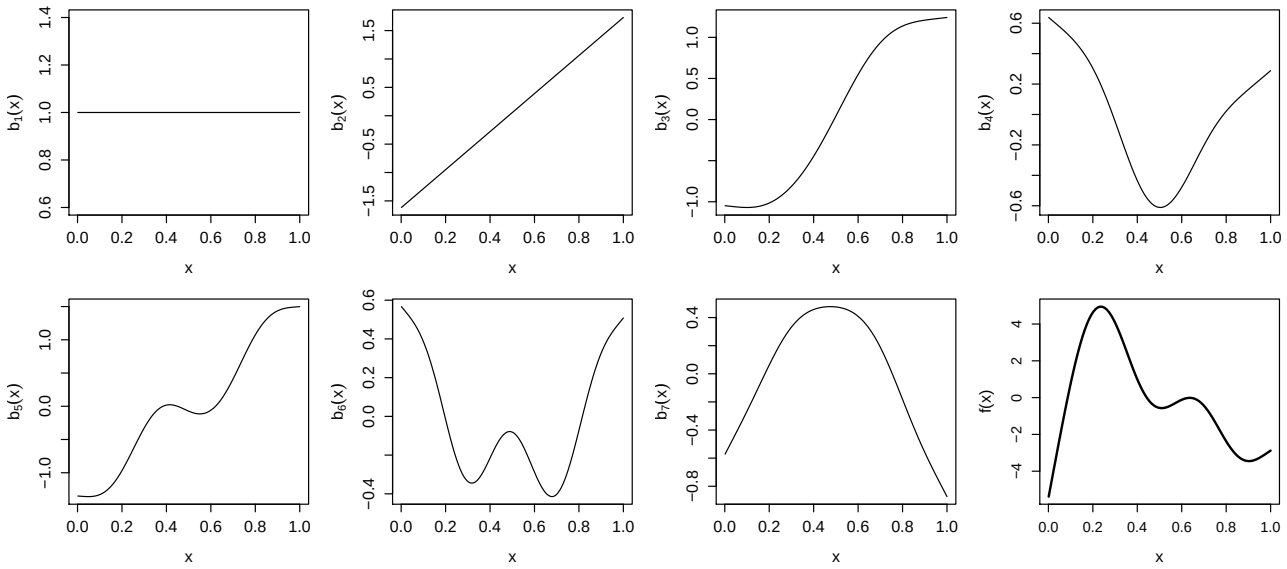

Figure 3.1: This plot illustrates a rank 7 thin plate regression spline basis for representing a smooth function of one variable. The first 7 panels (starting at top left) show the basis functions which multiplied by some coefficients and then summed give the smooth curve in the lower right panel. Based on a second order penalty, the first two bases span the space of functions that are completely smooth. The remaining basis functions represent the wiggly component of the smooth curve.

the model. This means that a GAM can simply be estimated by MLE. However, in a smoothing spline context, unpenalized parameter estimation is likely to result in smooth component estimates that are too 'wiggly', hence undermining the utility of the resulting estimates. This can be overcome by penalized MLE, where the use of penalties allows for the suppression of that part of smooth term complexity which has no support from the data. For simplicity and without loss of generality, let us assume that the linear predictor in model (3.19) is made up of smooth components only. The model can be fitted by maximisation of

$$\ell(\boldsymbol{\beta}) - \frac{1}{2}\sum_j \lambda_j \int \left\{ f_j^{d_j}(x_j) \right\}^2 dx_j. \tag{3.22}$$

The terms in the summation measure the roughness of the smooth functions, d_j (usually set to 2) indicates the order of the derivatives for the j^{th} smooth term to be used in the fitting process, and the λ_j are smoothing parameters that control the trade-off between fit and smoothness. Since regression splines are linear in their model parameters, the penalty $\sum_j \lambda_j \int \left\{ f_j^{d_j}(x_j) \right\}^2 dx_j$ can be written as a quadratic form in $\boldsymbol{\beta}$ with known coefficient matrices $\mathbf{S}_j$. As an example, by setting $d_j = 2$ and for a regression spline basis in one dimension, we have that

$$\begin{aligned}
\int \left\{ f_j^2(x_j) \right\}^2 dx_j &= \int \left\{ \frac{\partial^2 f_j(x_j)}{\partial x_j^2} \right\}^2 dx_j = \int \left\{ \frac{\partial^2 \sum_{k=1}^{q_j} \beta_{jk} b_{jk}(x_j)}{\partial x_j^2} \right\}^2 dx_j \\
&= \int \left\{ \boldsymbol{\beta}^{\mathrm{T}} \mathbf{b}_j''(x_j) \right\}^2 dx_j = \int \boldsymbol{\beta}^{\mathrm{T}} \mathbf{b}_j''(x_j) \mathbf{b}_j''(x_j)^{\mathrm{T}} \boldsymbol{\beta} dx_j \\
&= \boldsymbol{\beta}^{\mathrm{T}} \left\{ \int \mathbf{b}_j''(x_j) \mathbf{b}_j''(x_j)^{\mathrm{T}} dx_j \right\} \boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{S}_j \boldsymbol{\beta},
\end{aligned}$$

where $\mathbf{b}_j''(x_j)$ is a vector containing the second derivatives of the basis functions for the j^{th} smooth term with respect to x_j. It follows that

$$\sum_j \lambda_j \int \left\{ f_j^{d_j}(x_j) \right\}^2 dx_j = \sum_j \lambda_j \boldsymbol{\beta}^{\mathrm{T}} \mathbf{S}_j \boldsymbol{\beta}.$$

The precise mathematical expression of the quantities above depends on the chosen basis and value of d_j. The smoothing parameters play a crucial role in penalized regression spline estimation: very large values for λ_j lead to very smooth estimates and vice versa. Given smoothing parameters, the penalized nonlinear least squares problem can be solved by using the IRLS algorithm. It turns out that the form of the parameter estimators of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{z},$$

where $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$. Note that the estimator for $\boldsymbol{\beta}$ is biased because of penalty-induced bias.

In (3.22), the λ_j are not estimated but fixed to some values. This is because joint estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ via maximization of (3.22) would lead to very wiggly smooth function estimates since the highest value of (3.22) would be obtained when the smoothing parameters are equal to zero. This would result in severe overfitting which is not desirable. Hence the need to select the λ_j using an alternative criterion. Smoothing parameter estimation can be achieved by minimization of a prediction error estimate, such as the generalized cross validation (GCV) score, if a dispersion parameter has to be estimated, or the generalized AIC. Figure 3.2 illustrates how smoothing parameter selection works.

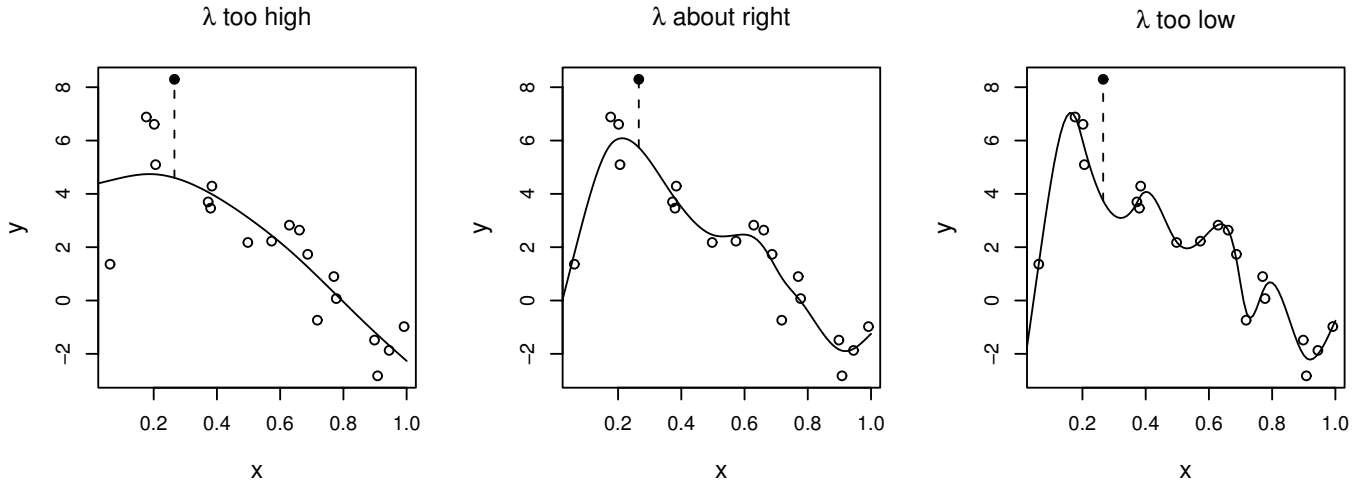

Figure 3.2: Illustration of the principle behind cross validation for smoothing parameter selection. In this case the datum denoted by • has been omitted from fitting and the continuous line shows the curve estimated using the remaining data (∘). When λ is too high the spline fits many of the data points poorly, especially the missing point. When λ is too low the spline over-fits, hence predicting the missing datum rather poorly. For the intermediate λ the spline fits the underlying signal quite well; as a result the missing datum is well predicted.

3.5.4 Inference

The inferential theory for models involving the use of penalized regression splines is not standard. This is because of the presence of smoothing penalties which undermines the use of classic

asymptotic likelihood results for practical modelling. In this section, we briefly describe how confidence intervals for GAM components can be constructed.

Let us consider a generic smooth model component $f(x_i)$. Intervals can be constructed seeking some constants C_i and A, such that

$$\text{ACP} = \frac{1}{n}\text{E}\left\{\sum_i 1(|\hat{f}(x_i) - f(x_i)| \leq q_{\alpha/2}A/\sqrt{C_i})\right\} = 1 - \alpha, \tag{3.23}$$

where 'ACP' denotes 'Average Coverage Probability', $1(\cdot)$ is an indicator function, α is a constant between 0 and 1, and $q_{\alpha/2}$ is the $\alpha/2$ critical point from a standard normal distribution. Defining $b(x) = \text{E}\{\hat{f}(x)\} - f(x)$ and $v(x) = \hat{f}(x) - \text{E}\{\hat{f}(x)\}$, so that $\hat{f} - f = b + v$, and I to be a random variable uniformly distributed on $\{1, 2, \ldots, n\}$, we have that ACP $= \Pr\left(|B + V| \leq q_{\alpha/2}A\right)$, where $B = \sqrt{C_I}b(x_I)$ and $V = \sqrt{C_I}v(x_I)$. At this point, it is necessary to find the distribution of $B+V$ and values for the C_i and A so that requirement (3.23) is met. Such a requirement is approximately met by the posterior distribution

$$\boldsymbol{\beta}|\mathbf{y} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, (\boldsymbol{\mathcal{I}} + \mathbf{S})^{-1}), \tag{3.24}$$

Given result (3.24), confidence intervals for linear and nonlinear functions of the model parameters can be easily obtained. Note that, for any strictly parametric model components, using (3.24) to obtain confidence intervals is equivalent to using classic likelihood results. This is because such model terms are not penalized.

Variable selection and model comparison in the GAM context is not standard and the methods illustrated in the previous sections may not provide reliable results. Model checking is similar to what is done for linear model and GLMs.

3.5.5 Optional reading for Section 3.5

- Marra G., Radice R. (2010), Penalised Regression Splines: Theory and Application to Medical Research. *Statistical Methods in Medical Research*, 19(2), 107–125.
- Ruppert D., Wand M.P., Carroll R.J. (2003), Semiparametric Regression. London, Cambridge University Press.
- Wood S.N. (2006), Generalized Additive Models: An Introduction with R. London, Chapman & Hall.
