Department of Methodology

MY452
Applied Regression Analysis

2022/2023

Course pack written by:

Jouni Kuha and Benjamin Lauderdale

Course information

MY452 — Applied Regression Analysis

(MY452M in Michaelmas Term, MY452L in Lent Term)

Course Description

This course is intended for those with prior training in quantitative methods. It is assumed that participants have already attended course MY451 (Introduction to Quantitative Analysis) or an equivalent course. Topics covered include multiple linear regression and binary, multinomial and ordinal logistic regression, as well as discussions on how inference and estimation should and should not be used in social science research. Students will be introduced to the computer programs Stata and/or SPSS.

Please note that up to 2010-11 the course was called MI452, so older materials will bear that code. For example, you will find past examination papers on the Library website under MI452.

Course Materials

- **Coursepack**: Digital copies of this coursepack are available for download from the MY452 Moodle page.
- **Lecture slides**: Copies of all of the slides displayed during the lectures can be downloaded from the MY452 Moodle page.
- **Required course text**:
  - Alan Agresti and Barbara Finlay (2009), *Statistical Methods for the Social Sciences* (Fourth Ed.). Prentice Hall.

MY452 on Moodle

The course materials are all available on Moodle. Go to `http://moodle.lse.ac.uk/` and login using your *username* and *password* (the same as for your LSE e-mail). Then in the *select courses* dialogue box type in MY452, and in *search results* click on MY452. The site contains the structure of the course week by week, the lecture recordings, the readings, weekly computer class assignments and the associated data sets, and other materials. The coursepack can also be found here, as well as a section on news and announcements. For PhD students and students who are auditing MY452 (rather than taking it for course credit), assignment to computer classes is also by on-line sign up on the Moodle page.

Computing

Software All course exercises and model answers will be practised using the open source statistical software R.

R is completely free and available from `www.r-project.org/`.

We also recommend the use of the dedicated development environment Rstudio, which is also freely available at: https://rstudio.com/products/rstudio/.

Feedback

We would welcome any comments you have on the course. If there are any problems that we can deal with, we will attempt to do so as quickly as possible. Speak to any member of the course team, or to your departmental supervisor if you feel that would be easier for you. Also please let us know if you find any errors or omissions in the coursepack, so that we can correct them for next year.

Acknowledgements

This coursepack still bears traces of previous editions and all of their authors, Colm O'Muircheartaigh, Colin Mills, Matt Mulford, Anders Skrondal and Fiona Steele. We

are grateful to Sally Stares and Piero Stanig for correcting numerous errors in the previous editions.

FAQ: Frequently Asked Questions

Is MY452 a course in statistics? No, it is a course in quantitative data analysis for social scientists. The distinction may appear subtle, but it is nonetheless important. If it were a course in statistics you would have a right to expect a much more theoretical, mathematical and formal treatment of the material. This is not what we offer. The course has a more applied focus than would be appropriate for a course in statistics and much of the explanation appeals to intuition rather than formal proof. Many of the technical details are downplayed, or simply treated as "black boxes" in order to emphasise the logic of data interpretation.

What do I need to know before I take MY452? You should have passed comfortably a first level statistics or quantitative methods course that covered: measures of central tendency and dispersion; contingency tables and chi-square tests; the normal distribution; sampling distributions; hypothesis tests for means and proportions; confidence intervals; correlation and regression. In addition the course should have emphasised the quantitative intuition behind these measures and techniques and required you to make some hand calculations rather than simply teaching you, cook-book style, which buttons of a computer program to click on.

I'm taking MY452 because I want to learn how to use R (or Stata, SPSS, etc.) but we don't seem to learn very much about the program. Why is that? MY452 is a course on statistics, not statistical software. We use statistical software merely to facilitate data analysis and interpretation with the methods covered on the course, and learn enough about the software to be able to do that. If you wish to learn more about particular statistical software, you will have to take a dedicated training course of some sort.

I'm taking MY452 to help me analyse data for my dissertation. Can I discuss my data and my specific problems with the lecturers? Yes, but not during the course. Staff of the Department of Methodology will be happy to talk to you about problems specific to your dissertation during the weekly sessions of the Methodology Surgery (see the website of the department for more).

Does the coursepack contain everything I need to know for the exam? A very large proportion of it, but the course pack is not quite a text-book. You will stand an even better chance of doing well in the exam if you also read the relevant parts of the Agresti & Finlay textbook, and follow the lectures, where the lecturers emphasise and explain the key parts of the material.

The lecturer introduced some material that was not in the coursepack. Do I need to know that material? This is almost certainly an illusion. The lectures will not introduce any genuinely new material not included in the course pack. However, sometimes the lecturer may of course use different words or a different example to further explain some topic. Copies of the most relevant notes displayed at the lectures will be posted in the MY452 Moodle site.

There is one exception to the principle stated in the previous paragraph. Regression models for counts, which are introduced in the lecture and class of week 9 of the course, are not included in the coursepack. For these models, the lecture slides provide the basic information.

I don't like the textbook and I have a different one that I used for another course/I found on a bus/my uncle gave me. Is it OK to use it? There are hundreds, possibly thousands of quantitative methods textbooks, many of them covering almost identical material. If it looks as though it covers the same material at about the same level of sophistication then it is probably OK. But NB *caveat emptor*. If in doubt ask one of the lecturers.

Can I work together on the homework with my friends? Yes, we positively encourage you to discuss the homework assignments with your colleagues.

If I don't agree with the mark, how do I appeal? The cultural norm in the UK is that marks are not arrived at by a process of teacher-student negotiation. You can make a formal appeal through the School's appeal process (see the appropriate section of the Graduate Handbook for details). NB: appeals cannot be made on grounds of academic substance, only on grounds of procedural irregularities. In other words an appeal will not be allowed if the only grounds you have for appealing is that you/your friend/your personal advisor/your spiritual guru think your script was worth more marks than the examiners did.

MY452 is too difficult for me. Is there a more elementary course? Yes, MY451 or MY465. MY451 finishes about where MY452 starts, whereas MY465 covers some of the material in MY451 and some of the material in MY452. MY451 is sufficient preparation for MY452. MY465, like MY452, is sufficient preparation for the more advanced courses offered by the Department of Methodology.

MY452 is too easy for me. Is there a more advanced course? There are several other appropriate courses on quantitative analysis by the the Department of Methodology and the Department of Statistics: MY455, MY456, and MY457. These build on the topics covered on MY452, with about the same style and the same of somewhat higher level of difficulty as MY452. More mathematically oriented courses on statistical methods are offered by the Statistics and Economics departments.

Course Programme

Week 1	
Lecture	Course overview and organisation. Review of statistical inference
Class	Introduction to Stata or SPSS.
Reading	Coursepack: Chapters 1 and 2; Agresti & Finlay (AF): 1–7
Week 2	
Lecture & Class	Linear regression 1: Definition, estimation, inference and interpretation
Reading	Coursepack: Chapter 3 and Sections 4.1–4.5; AF: 9–10, 11.1–11.4
Week 3	
Lecture & Class	Linear regression 2: Types of explanatory variables
Reading	Coursepack: Section 4.6; AF: 11.5, 12.1–12.5, 13.1–13.4, 14.4–14.6
Week 4	
Lecture & Class	Linear regression 3: Model fomulation and selection, diagnostics
Reading	Coursepack: Sections 4.7–4.9; AF: 9.6, 14.1–14.2
Week 5	
Lecture & Class	Binary logistic regression 1: Definition, estimation and interpretation
Reading	Coursepack: Sections 5.2–5.3; AF: 15.1–15.2
Week 6	
Reading Week	No Lecture
Week 7	
Lecture & Class	Binary logistic regression 2: Inference and model selection
Reading	Coursepack: Section 5.4; AF: 15.3
Week 8	
Lecture & Class	Multinomial logistic regression
Reading	Coursepack: Chapter 6; AF: 15.5
Week 9	
Lecture & Class	Ordinal logistic regression
Reading	Coursepack: Chapter 7; AF: 15.4
Week 10	
Lecture & Class	Regression models for counts
Reading	Slides from the Moodle page; AF: 15.6–15.7 (only partly relevant)
Week 11	
Lecture	Review and further topics
Class	Review exercises
Reading	Lecture slides from the Moodle page; Coursepack: Chapter 8; AF: 16

Contents

Chapter 1

Introduction

1.1 What is the purpose of this course?

MY452 is a second-level course on methods of statistical analysis of quantitative data, aimed for graduate students in the social sciences and taught with emphasis on concepts and examples. The main topic of the course is statistical regression modelling, focusing on five important types of such models. Regression models are by far the the most important and widely used class of statistical techniques, and the specific models considered here are the ones most commonly employed in social research.

The material covered on this course relates to those of other courses as follows:

- Because MY452 is an intermediate-level course, it builds on basic statistical tools and ideas introduced on an **introductory statistics** course. We assume that you have taken such a course before, e.g. MY451 at the LSE or a comparable course elsewhere. Some of these basic concepts are briefly reviewed in the first three chapters of this coursepack.

- The course deals with methods of *analysis* of *quantitative data*. Such analysis is only one element of the empirical research process, and learning methods of analysis only makes sense in this broader context. Other parts of the research process, such as the formulation of research questions, data collection, measurement, and reporting of the results, are covered on courses on **research design** such as MY400 (Fundamentals of Social Science Research Design). Methods for collection and analysis of **qualitative data** are also covered on other courses, such as MY421 (Qualitative Research Methods).

- Most importantly, the questions that statistical analysis of quantitative data aims to help to answer are research questions in the social and other sciences. In other words, the claims about regularities in the social world discussed on the **substantive courses** that form most of your degrees are based on analyses of empirical observations, often using methods covered on this course. You may also need to use these methods yourself to produce new findings, for an MSc dissertation or a PhD thesis, or in the world after graduation.

At the end of the course you should be familiar with the most important types of regression modelling. This will enable you to be both a user and a consumer of statistics:

- You will be able to use the methods to analyse your own data and to report the results of the analyses.
- Perhaps even more importantly, you will also be able to understand (and maybe criticize) their use in other people's research. As the methods discussed here are very widely used, this will make accessible to you a large proportion of quantitative empirical research in academic and other publications, far beyond the range of the methods discussed on an introductory statistics course. Furthermore, because even those regression models that are not discussed on this course use the same basic ideas as those that are, and because interpreting results is somewhat easier than carrying out new analyses, you will also have some understanding of many further techniques that are not covered here.

The rest of this chapter gives a brief overview of some of the basic concepts behind the idea of regression models (for all of these topics, more detailed discussions can be found in the text book by Agresti and Finlay and the coursepack for MY451, which you can download from the Moodle site of that course). Different types of variables are compared in Section 1.2. The concept of statistical associations is introduced in Section 1.3, and its relation to the stronger concept of causality is discussed in Section 1.4. Section 1.5 discusses the idea of statistical models and outlines the nature of the specific models considered on this course, while Section 1.6 describes the structure of the rest of the course.

1.2 Data and variables

The information analysed by statistical methods is contained in a set of **data**, consisting of recorded values of several **variables** for a set of **units**. The units may be people, countries, organisations or any other entities relevant for the research question. The number of units for which data are available is known as the **sample size**, and is typically denoted by the symbol n. The data are normally arranged in the form of a rectangular **data matrix** with one row for each unit and one column for each variable. For example, in the computer classes you will see data displayed in this form in Stata or SPSS.

The variables represent *measurements* of the levels of some characteristics of the units. For example, suppose we were collecting a data set on a group of undergraduate students at the LSE. This might include for each student variables on their sex as observed by a survey interviewer, parents' income reported by the student himself/herself in the interview, and the average mark in the student's first-year examinations, obtained from School records. These could then be regarded as measurements of corresponding characteristics, i.e. sex, parental income and examination results (or perhaps academic performance in a more general sense).

In these and all other cases it is clearly crucial that the variables should be good (meaningful, valid and reliable) measures of the relevant concepts. This is not always

easily achieved, and devising adequate ways of measuring variables is often a major challenge in research design. On this course, however, we will set aside questions of quality of measurement, in effect assuming that the variables we are considering have been measured well enough for the analysis to be meaningful.

The data are called quantitative because the values of all of the variables are recorded as numbers. However, there may be important differences between variables in the nature and interpretation of those numbers, which we must take into account in the analysis. The first of these distinctions defines the **measurement level** of a variable:

- A variable is measured on a **nominal scale** (i.e. it is a "nominal-level" variable) if the numbers are simply labels for different possible values (*levels* or *categories*) of the variable. These labels are not in any natural order and have no quantitative meaning; they thus cannot be treated as numbers in the everyday sense, and calculations such as sums or means of them are not meaningful. An example is labour market status, recorded as 1 = Employed, 2 = Unemployed and 3 = Not in the labour force.
- A variable is measured on an **ordinal scale** if its values do have a natural ordering. The numerical codes assigned to the values then need to be in the correct order, but they have no other quantitative meaning and should not, strictly speaking, be used for calculations. An example is the level of agreement with an attitude statement in a survey, recorded as 1 = Completely disagree, 2 = Somewhat disagree, 3 = Neither agree nor disagree, 4 = Somewhat agree, and 5 = Completely agree.
- The values of a variable measured on an **interval scale** are "real" numbers in that standard calculations on them are meaningful. A criterion for this to be true is that *differences* of the values must be comparable: for example, for an interval-level variable the difference between values recorded as 0 and 5 must be of equal magnitude to the difference between 5 and 10, and both must be bigger than the difference between 10 and 12. An example of an interval-level variable is age measured in years.
  - Often a further distinction is made between **ratio-level** variables for which ratios of values are also comparable, and pure interval-level variables for which they are not. This distinction is largely irrelevant for our purposes, so we will ignore it and refer to both as interval-level variables.

The second important distinction is between continuous and discrete variables:

- A variable is **continuous** if it can in principle take infinitely varied fractional values. The idea implies an unbroken scale of possible values. Age is an example of a continuous variable, as we can in principle measure it to any degree of accuracy we like.
- A variable is **discrete** if its basic unit of measurement cannot be subdivided. Thus a discrete variable can only have certain values, and the values between these are logically impossible. An example is the agreement variable mentioned under ordinal scales above, with only possible values 1, 2, 3, 4 and 5.

Table 1.1: Relationships between the types of variables discussed in Section 1.2.

	Measurement level: **Nominal or ordinal**	Measurement level: **Interval**
Discrete	Many • Always **categorical**, i.e. having a fixed set of possible values (categories) • If only two categories, variable is **binary** (**dichotomous**)	*Counts* • If many different observed values, often treated as effectively continuous
Continuous	None	Many

  - An important type of a discrete variable is a **categorical variable**, which has only a finite (in practice usually quite small) number of possible values. For example, labour market status as defined above is categorical. The simplest kind of categorical variable has only *two* possible values, for example sex recorded as "male" or "female" or an opinion recorded just as "agree" or "disagree". Such a variable is said to be **binary** or **dichotomous**.
  - The most common kind of a discrete but non-categorical variable is a **count**, such as the population of a country, which must be an integer but can have many possible values with no obvious upper bound.

These two sets of distinctions are partially related. (Almost) all nominal or ordinal-level variables are discrete, and (almost) all continous variables are interval-level variables. This leaves one further possibility, a discrete interval-level variable, of which counts are the primary instance. These connections are summarized in Table 1.1. Some of the divisions may be further blurred in practical data analysis, for example when we sometimes treat a discrete, ordinal variable effectively as an interval-level variable. This issue is discussed further in Chapter 4.

The types of the variables involved in an analysis largely determine what kinds of statistical methods are appropriate. What this means for the techniques considered on this course is outlined in Section 1.5 below.

1.3 Association and regression

All of the methods considered on this course are used to examine statistical associations between variables. One rather nonrigorous definition of such an association is that

- There is an **association** between two variables if knowing the value of one of the variables will help to predict the value of the other variable.

Other ways of referring to the same concept are that the variables are "related" or "correlated", or that there is a "dependence" between them.

For example, recent UK data on voting behaviour typically show that voters who were born outside Britain are more likely to vote for Labour and less likely to vote for the Conservatives or Liberal Democrats than those born in the country. There is then an association between the two (discrete) variables "country of birth" and "voting intention", and knowing whether a person was born abroad would help us better predict who they intend to vote for. Similarly, a study of education and income might find that people with more years of education completed tend to have higher annual incomes, again suggesting an association between these two (continuous) variables.

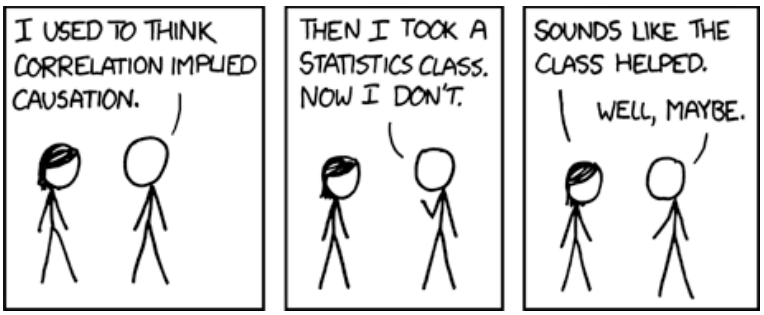
Sometimes the variables in an association are treated on an equal footing. More often, however, they are considered asymmetrically in that it is more natural to think of one of them as being used to predict the other. For example, in the above examples it seems easier to talk about country of birth predicting voting intention than vice versa, and of level of education predicting income than vice versa. This kind of asymmetric formulation of associations is the defining feature of statistical **regression models**. The variable used for prediction is then known as an **explanatory variable** and the variable to be predicted as the **response variable**; an alternative convention is to talk about *independent* rather than explanatory variables and *dependent* instead of response variables, but this terminology will not be used here. In notation, a response variable is typically denoted by Y and an explanatory variable by X, and this convention will also be used throughout this coursepack.

As defined above, an association is a relationship between *two* variables. Often, however, we will carry out *multivariate* analyses involving three or more variables at once, rather than *bivariate* analyses of just two variables at a time. In a multivariate analysis, we will still examine associations between pairs of two variables, but now *controlling for* the other variables. What this means will be discussed in more detail in Chapter 4. The regression models considered on this course have just one response variable, but they may have several explanatory variables. Models involving only one explanatory variable are **simple** regression models, while those with several explanatory variables are **multiple** regression models. In multiple regression models, we can describe associations between the response variable and each explanatory variable in turn, controlling for the other explanatory variables.

1.4 Association vs. causality

As the caption to the cartoon in Figure 1.1 suggests, when we establish an association or correlation it is often tempting to use it also to answer questions at the more demanding level of **causation**. But what exactly does it mean for a relationship between two variables to be causal? There are multiple approaches to defining causality, but the most straightforward is in terms of *counterfactuals*. A relationship between X and Y is causal if changing the value of X would change the value of Y for a given unit. That is, we observe an *factual* value of Y, but we want to know if the *counterfactual* value of Y would have been different if we had somehow intervened to change the value of X to an alternative value.

Figure 1.1: The artist's caption reads: "Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'." Source: `http://xkcd.com/552/`



To return to the examples of Section 1.3, the relationship between country of birth and voting is causal if and only if changing someone's country of birth would make them vote differently. The relationship between education and income is causal if and only if receiving more years of education would increase a given person's income. As these examples illustrate, it can be difficult to determine what the appropriate counterfactual is, or to find one that corresponds to a meaningful intervention. While we can imagine ways to intervene to make someone attend more years of school, it is very difficult to imagine what it means to change someone's country of birth.

Regression methods estimate associations but they do not, in general, estimate causal effects. They can describe the relationships between a response variable Y and one or more explanatory variables X *across* a set of observations, and enable inferences about the population of units from which those observations were drawn. However, they do not enable one to make inferences about the response variable Y would have changed if X had been changed. To see this, consider the example of education and income. We can observe the association between education and income in a sample: For example, let us imagine that in our sample, on average, individuals with 15 years of schooling earn 35% more than those with 12 years of schooling. We can then use this information also to make inferences about the population from which our sample was drawn. However, does this mean that if the individuals with 12 years of schooling were provided with 3 more years of schooling, they would on average earn 35% more than they currently do? No, not necessarily. There may be all sorts of other differences between the two groups that also contribute to their different levels of income. Those differences might have *caused* both the difference in education levels and the difference in income levels. From merely knowing the average income among these two groups, we cannot make any specific claims how changing education levels would change incomes levels. But we do, as the cartoon caption indicates, have reason to investigate further.

While the idea of controlling for other variables through multiple regression seems as though it might be able to address this problem, very often this will not work. Here it would require that we manage to include as explanatory variables in the model *all* the variables that both influence income levels and are associated with years of schooling attained. Clearly, this is not likely to be feasible in this example. In other situations, being able to include all the relevant explanatory variables in a model may be more (or even less) feasible, and whether it is will depend on both the research question and the research design through which data are collected.

In summary, the main job of regression models is to estimate associations between variables. These associations are not causal effects, unless such a claim is justified by further strong assumptions. The feasibility of these assumptions cannot be inferred from the regression analysis itself but must be supported by additional information about the research design and the variables in the model. On this course we will not discuss such questions much further, but will concentrate on the practice and interpretation of regression modelling itself. If you are interested in learning more about causal inference, we strongly advise you to take this course and then take MY457 (Causal Inference for Observational and Experimental Studies). That course provides an in-depth theoretical overview of causality as well as a variety of methods that have been developed for establishing and estimating causal relationships, including the uses of regression modelling as an element in that activity.

1.5 Regression models considered on this course

A **statistical model** is a simplified representation of the process which generated a set of observed data. In particular, a *regression* model describes how values of a response variable depend on the values of one or more explanatory variables. An adequately specified model of this kind can be very useful for two main purposes:

- *Explaining* and *understanding* how the response variable is associated with explanatory variables.
- *Predicting* or *controlling* future values of the response variable through knowledge or control of the values of the explanatory variables.

— which we might, somewhat casually, summarise as the "scientific" and "practical" uses of regression modelling. In practice, of course, our motives for considering such models are usually a mixture of explanation and prediction.

A more careful discussion of the concept of statistical models will be given in Sections 3.2 and 3.3, and details of how specific instances of such models are defined will occupy much of the rest of the course. The detailed specification of the elements of a regression model depends on the number and types of variables involved. The main differences here concern the type of the response variable. The explanatory variables, on the other hand, will be treated essentially similarly in all such models, with nominal-level variables handled rather differently from interval-level ones, and ordinal variables in effect treated as one of these other types.

On this course, we will consider the following five types of regression models for different kinds of response variables:

- **Linear** regression models for interval-level response variables.
- **Binary logistic** regression models for dichotomous response variables.
- **Multinomial logistic** regression models for categorical response variables with more than two categories. The response categories may in reality be unordered

(nominal) or ordered (ordinal), but the model will always treat them as unordered.

- **Ordinal logistic** regression models for categorical response variables with more than two categories, where the categories are ordered and treated as such.
- Regression **models for counts** as the response variable.

These are the most commonly used regression models in most social sciences. Because there are many differences of detail in specification and interpretation of the models, we will discuss each of them separately. However, underlying these surface differences are common basic ideas and strong similarities of broad structures. Discussion of the different models in separate chapters will thus provide much repetition of these commons concepts to aid the learning.

1.6 Structure of the course

The rest of this coursepack divides into three parts:

- The next two chapters continue the introduction by providing a brief review of material assumed known as prerequisites for this course:
  - Chapter 2 is a review of basic concepts of statistical inference.
  - Chapter 3 is a review of simple linear regression models.
- The core chapters cover four of the five models considered on the course:
  - Chapter 4 on multiple linear regression models.
  - Chapter 5 on binary logistic regression models.
  - Chapter 6 on multinomial logistic regression models.
  - Chapter 7 on ordinal logistic regression models.
  - Models for counts are not included in this coursepack. Instead, the lecture notes for week 9 provide an introduction to these models.
- Chapter 8 gives a very brief overview of some other kinds of regression models not covered on this course.
- At the end of the coursepack are instructions for the weekly computer classes.

Chapter 2

Review of statistical inference

An important disctinction is drawn on introductory statistics courses between methods of *descriptive statistics*, used to summarise observed data in convenient and understandable ways, and *statistical inference*, used to generalise conclusions from observed samples to larger populations and to quantify the uncertainty in such conclusions. The regression models considered on this course can be used for both purposes. Although they are quite useful even as purely descriptive summary statistics, questions of statistical inference are also central in many if not most of their applications. In preparation for the inferential elements of regression modelling, this chapter briefly reviews the basic concepts and elements of statistical inference.

The principles are here discussed in the context of a specific type of analysis, namely the comparison of population means of an interval-level variable between two groups. This situation and the significance tests and confidence intervals used for it should be familiar from introductory courses. It is also closely related to linear regression modelling, which will be discussed in Chapters 3 and 4.

Data from the 2002/3 European Social Survey (ESS) will be used for illustration in this chapter[1]. We will compare respondents in Great Britain and France, and consider their values for the following two variables from the survey:

- The number of hours per week the respondent normally works in his or her main job, including overtime (respondents with employment only).

- The length of the survey interview recorded by the interviewer, in minutes.

Summary statistics for these variables are shown in Table 2.1, together with results of the tests and confidence intervals discussed below.

[1] R. Jowell and the Central Co-ordinating Team (2003). *European Social Survey 2002/2003: Technical Report*. Centre for Comparative Social Surveys, City University. The data, which are archived and distributed by Norwegian Social Science Data Services (NSD), were obtained from the web site `www.europeansocialsurvey.org`, which also gives much further information on the survey.

Table 2.1: Summary statistics for the two examples considered in Chapter 2. See the text for definitions of the statistics.

	n	$\bar{Y}$	s	$\hat{\Delta}$	$\hat{\sigma}$	t	P	95% CI
Average weekly working hours								
UK	1956	38.74	15.72					
France	1316	40.35	14.55	1.61	15.26	2.96	0.003	(0.54; 2.68)
Length of the survey interview								
UK	1913	64.74	57.64					
France	1496	66.58	35.78	1.84	49.26	1.08	0.279	$(-1.49; 5.17)$

Source: European Social Survey 2002/3

2.1 Samples and populations

Statistical inference is used when the observations in the available data are regarded as a **sample** from some larger **population**, and our research questions concern the population rather than just the sample. By definition, information on the variables is available only for the sample, and their values for the rest of the population are unknown. It is nevertheless possible to draw conclusions about the whole population, if the sample can be regarded as *representative* of the population. Drawing such conclusions, and quantifying the level of uncertainty involved in them, is the task of statistical inference.

Our two examples illustrate two somewhat different types of samples and populations:

- For the average weekly working hours, the populations are people in the UK and France with a job. These are large but still finite groups of actual people. For each of them the variable has some true value, although this is known to us only for the members of the sample. The sample of respondents in the survey was obtained using methods of *probability sampling*, which allows us to treat the sample as representative of the population.

- For the observations of interview lengths, the situation is less clear. Here the variable has a value for those respondents who were interviewed, but not, it seems, for those were not interviewed. On the other hand, it is clear that it *would* have had a value for any respondent if they *had* been selected to the sample (and agreed to be interviewed), even if they were not included in the sample actually collected. In other words, interview length is a sort of potential variable whose value is realised only when an interview actually takes place.

  Furthermore, it may be argued that the length would have been different even for a respondent in the observed sample, if he or she had been interviewed at a slightly different time or in a different mood or by a different interviewer than in the interview which actually took place. The length is thus arguably best regarded as a characteristic of the interview rather than of the respondent, and the population of interest as one of interviews rather than of people. We may thus treat the interview lengths in the data as observed characteristics of two samples from populations of potential ESS interviews, one in Britain and one

in France. Whether these samples are representative of such populations can only partially (if at all) be justified by appealing to the probability sampling of the respondents. Instead, other, less concrete arguments are needed. Often this might amount to something as vague as arguing that the sample is not obviously *un*representative, here that there is no obvious reason why these interviews would differ from others that might have been or may later be carried out under similar conditions and instructions.

The case of the interview lengths is an illustration of a situation where the population of interest is not a finite group of concrete units such as people. Instead, it is a hypothetical, conceptual population of possible realisations of some events, effectively infinite in size, from which the observed values in the data are regarded as a sample of realisations. This case is more abstract and vague than the finite-population situation which is typically used to introduce ideas of statistical inference on introductory statistics courses (even though instances of actual finite populations are in the minority even there). However, the general motivation of inference is the same in both cases. The idea of such conceptual populations is particularly important on this course and for regression modelling in general, because statistical inference is here virtually always really based on a hypothetical population of some kind. This question will be discussed further in Section 4.9.2.

2.2 Models for the data

Statistical inference of the kind considered here is based on a set of assumptions about the distributions of the variables of interest in the population, and of the nature of the observed samples drawn from them. Such a specification is known as a **statistical model** (the nature of such models will be discussed further in Sections 3.2–3.3). Some model assumptions are needed for virtually any kind of statistical inference to be possible.

All the methods of inference discussed on this course are instances of *parametric* inference, based on a parametric statistical model. What this means is that the model is specified in terms of a small number of population **parameters**, and the inference focuses on questions about the values of those parameters. For example, in the two-sample case discussed in this chapter, we will consider the following model:

1. The n_1 observations of a variable Y [here weekly working hours, or length of interview] in group 1 [here UK, say] are a random sample from a population where Y has a normal distribution with mean μ_1 and variance σ^2.

2. The n_2 observations of Y in group 2 [here France, say] are a random sample from a population where Y has a normal distribution with mean μ_2 and variance σ^2.

3. All observations of Y within and between the samples are *statistically independent*. This is a technical statistical term, which can be for our purposes roughly translated as "unrelated". The condition can usually be regarded as satisfied

when the units of analysis are different entities, such as different people (or different interviews) in the social survey here. All of the methods discussed on this course assume that observations for different units are independent in this sense.

Several aspects of this model are worth noting:

- The model describes the population distributions of Y in each group using only *two* parameters, the *mean* μ and *variance* σ^2 (or, equivalently, the square root of the variance, the *standard deviation* σ).
- The means μ_1 and μ_2 may be different in the two groups. These means and their difference are the quantities of main interest in the analyses discussed here.
- The variance σ^2 is assumed to be the same in both populations. This is a simplifying assumption, which could in fact be easily avoided in two-group analyses. It is used here because an analogous assumption is also a part of the linear regression models discussed later.
- The population distributions are assumed to have the form of a particular *probability distribution*, here the *normal* distribution. This assumption could also be avoided, as discussed in Section 2.4, but is again made here to parallel the standard specification of linear regression models.

2.3 Population parameters and their estimates

A parametric statistical model reduces all questions about the population to questions about a few parameters of the population distributions. Often we are mainly interested in only some of the parameters, while the rest are *nuisance parameters*, necessary but of less interest. In the two-sample case considered here, most research questions are in practice such that the variances will be nuisance parameters and the means the parameters of main interest, so we will present methods of inference for this case.

It will be convenient to express the two population means in the following form:

$$\begin{aligned} &\text{Mean in group 1:} && \mu_1 \\ &\text{Mean in group 2:} && \mu_2 \quad = \mu_1 + (\mu_2 - \mu_1) = \mu_1 + \Delta \end{aligned}$$

where $\Delta = \mu_2 - \mu_1$ is the *difference* of means between groups 2 and 1. It is this difference rather than even the individual means μ_1 and μ_2 which is of most interest here. This is because Δ relates directly to an association between the group variable and Y. There is no association if $\Delta = 0$ (i.e. $\mu_1 = \mu_2$, the mean is the same in both groups) and an association if $\Delta \neq 0$, i.e. if $\Delta > 0$ ($\mu_2 > \mu_1$) or $\Delta < 0$ ($\mu_2 < \mu_1$). In our examples the group variable is country (UK or France), so the questions of interest for the two response variables are whether average weekly working hours or average lenghts of the ESS interview are different in the two countries.

The first step of statistical inference is to obtain sample estimates (**point estimates**) for the parameters. Here an obvious estimate of each population mean is the corre-

sponding sample mean, i.e. for group 1

$$\bar{Y}_1 = \frac{\sum Y_i}{n_1}$$

where the summation is over all the observations of Y from group 1. The estimate of μ_2 is the sample mean $\bar{Y}_2$ in group 2, defined similarly. An obvious estimate of $\Delta = \mu_2 - \mu_1$ is then the difference of sample means, i.e.

$$\hat{\Delta} = \bar{Y}_2 - \bar{Y}_1. \tag{2.1}$$

These estimates in our examples are shown in Table 2.1. For the working hours variable, $\hat{\Delta} = 1.61$, i.e. the sample average of reported weekly working hours is a little over an hour and a half higher among the French than the British respondents[2]. For the interview length, $\hat{\Delta} = 1.84$, i.e. the average recorded ESS interview takes just under two minutes longer in France than in the UK.

An estimate of the population variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{(n_2 - 1)s_2^2 + (n_1 - 1)s_1^2}{n_1 + n_2 - 2} \tag{2.2}$$

where

$$s_1^2 = \frac{\sum (Y_i - \bar{Y}_1)^2}{n_1 - 1}$$

is the sample variance of Y among the observations in group 1, and s_2^2 is defined similarly for group 2. For future comparisons, it is also useful to rewrite the same $\hat{\sigma}^2$ using a slightly different formula. For this, let $\hat{Y}_i$ be a quantity (which will later be called the "fitted value") defined for each unit i so that it is equal to $\bar{Y}_1$ if i is in group 1, and equal to $\bar{Y}_2$ if i is in group 2. In other words, $\hat{Y}_i$ is the estimated population mean in the group to which unit i belongs. We can then write (2.2) as

$$\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2} \tag{2.3}$$

where $n = n_1 + n_2$ denotes the total sample sizes over the two groups together, and the summation in (2.3) is over all of these n observations. The estimate of the parameter corresponding to σ^2 will be of exactly this form also in more general linear models discussed later, of which the two-group model of this section is a special case.

2.4 Significance testing

A **significance test** is a technique of statistical inference designed to examine specific claims about the values of population parameters. Such a test provides an assessment of whether it is plausible, given the evidence provided by the observed data, that a population parameter (or parameters, for some tests) has a particular value specified by the researcher.

[2]This difference, which is statistically significant, is interesting in the context of the 35-hour working week in France and the UK's opt-out of even the 48-hour maximum mandated by the EU working time directive. Note that the variable considered here is *self-reported* weekly working hours, *including overtime*.

In general, the main steps in carrying out any significance test are the following:

1. State the **null hypothesis** H_0 of the test, and the **alternative hypothesis** H_a against which H_0 is to be tested.
2. Calculate the value of a **test statistic**.
3. Identify the **sampling distribution** of the test statistic if H_0 is true.
4. Calculate and report the **P-value**, which is a summary measure of the plausibility of H_0 given the observed data.
5. Report a **conclusion**, rejecting or not rejecting the null hypothesis at a stated significance level.

Below we will discuss each of these steps, in the context of a test of the mean difference Δ defined in Section (2.3); this is known as the "t-test of equality of means for two independent samples" (or words to that effect). The description given here is brief and not meant to be comprehensive, as we assume that you have encountered this test and the general concepts of significance testing before. For a more extensive refresher, please consult the text book by Agresti and Finlay.

2.4.1 Null and alternative hypotheses

The null hypothesis is the specific claim about some population parameters tested by the significance test. For the two-sample t-test, this is

$$H_0 : \Delta = 0 \tag{2.4}$$

where $\Delta = \mu_2 - \mu_1$ is the difference of population means of a variable Y between two groups. The null hypothesis thus claims that there is no mean difference, i.e. that there is no association (at least in terms of means) between the group and Y. In the examples of this chapter, this is thus the claim that there is no difference in average weekly working hours in Britain and France, or no difference in average lenghts of survey interviews in Britain and France. In general, the null hypothesis for most tests which concern associations is that there is no association.

The null hypothesis is tested against an alternative hypothesis, here either the *two-sided* hypothesis

$$H_a : \Delta \neq 0 \tag{2.5}$$

or one of the *one-sided* hypotheses

$$H_a : \quad \Delta < 0 \quad \text{or} \tag{2.6}$$

$$H_a : \quad \Delta > 0. \tag{2.7}$$

Sample evidence which is consistent with the alternative hypothesis will be regarded as evidence against the null hypothesis. This means that for the two-sided alternative (2.5), evidence that Δ differs from zero, whether in the negative or positive direction, will be evidence against H_0. For the one-sided alternatives, on the other hand, only

one side of zero matters. If H_a is (2.6), for example, only evidence of a negative mean difference (i.e. of $\mu_1 < \mu_2$) will ever be regarded as evidence against the null hypothesis.

For the models discussed on this course, two-sided alternative hypotheses are in practice used much more commonly than one-sided ones, so we will also concentrate on them throughout the course. If required, modifications for the one-sided case (which are needed only in calculating the P-value of the test) are easily done as explained below. A decision to consider a one-sided alternative hypothesis should usually be based on subject-matter arguments which suggested that only deviations from H_0 in one direction are important and/or interesting.

2.4.2 Test statistic

Sample evidence on the plausibility of the null hypothesis is distilled into a single number calculated from the data, known as a test statistic. For the case considered here, we will use the two-sample t-test statistic

$$t = \frac{\bar{Y}_2 - \bar{Y}_1}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2}} \tag{2.8}$$

where $\hat{\sigma}$ is given by (2.2). This is a ratio of two quantities:

- The numerator $\bar{Y}_2 - \bar{Y}_1$ is an estimate of Δ (labelled in $\hat{\Delta}$ in equation 2.1).
- The denominator $\hat{\sigma}\sqrt{1/n_1 + 1/n_2}$ is an estimate of the **standard error** of $\hat{\Delta}$. As explained on introductory statistics courses, the *sampling distribution* of a statistic is the distribution of values of the statistic over a hypothetical, infinitely large set of repeated random samples from the same population and of the same size as the sample actually collected. For particular models and statistics, the form of the sampling distribution can often be derived mathematically. The standard deviation of the sampling distribution of a statistic is known as the standard error of that statistic.

  When a statistic is used as an estimate of some population quantity, its standard error is best thought of as a measure of the *precision* of the estimate. An estimate with a low standard error, i.e. one whose values would be expected to vary little from sample to sample, is said to be more precise than an estimate with a higher standard error[3]. Another way of expressing this is to say that the smaller the standard error of an estimate, i.e. the higher its precision, the less *uncertainty* there is in the estimate.

  The estimated standard error of $\hat{\Delta}$ used in (2.8) is

$$\hat{\text{se}}(\hat{\Delta}) = \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{2.9}$$

  where we have introduced the notation $\hat{\text{se}}$, which will also be used later for other estimated standard errors. An important feature of the standard error is that

[3]Note that "precise" is here used in a narrow technical sense which is not quite the same as "good". An estimate with high precision is not very useful if it is a precise estimate of the wrong quantity, i.e. if it is a *biased* estimate of the parameter of interest. This is a question to be discussed separately.

it is a decreasing function of the sample sizes n_1 and n_2. In other words, the standard error is, other things being equal, going to be smaller when the estimate $\hat{\Delta}$ is calculated from large samples than when it is calculated from small samples. This is a general result which holds for virtually all statistical analyses. It also makes intuitive sense, whichever of the equivalent terms we state it in: larger samples imply lower standard errors, more precision and less uncertainty.

In short, the test statistic (2.8) is of the form

$$t = \frac{\hat{\Delta}}{\hat{\text{se}}(\hat{\Delta})}, \tag{2.10}$$

i.e. an estimate divided by its estimated standard error. Many other test statistics are of this same form, including the one- and two-sample t- and z-statistics for means and proportions discussed on introductory courses, as well as tests of individual regression coefficients used repeatedly on this course.

The value of the t-statistic is used to assess the level of evidence against the null hypothesis H_0 that $\Delta = 0$ in the population. Since the numerator $\hat{\Delta} = \bar{Y}_2 - \bar{Y}_1$ is an estimate of Δ, it is clear that values of $\hat{\Delta}$ and thus also of t which are far from zero will be regarded as evidence against H_0, while values close to 0 will not. What counts as "far from 0" depends, first, on the alternative hypothesis: if this is two-sided, both positive and negative values of t may be evidence against against H_0, while for a one-sided alternative only values in its direction (i.e. $t < 0$ for hypothesis 2.6 and $t > 0$ for 2.7) may lead to rejection of H_0. Second, and more importantly, we need to decide which specific values of t are far enough from 0 to be regarded as strong enough evidence against the null hypothesis. This will be discussed in the next section.

It should also be noted that the magnitude of t depends on the standard error $\hat{\text{se}}(\hat{\Delta})$ as well as the sample difference $\hat{\Delta}$. Specifically, t increases as the standard error decreases. This means that if we had two data sets in which the value of $\hat{\Delta}$ was the same, the one with the smaller standard error (i.e., other things being equal, the one based on more observations) would yield a value of t farther from zero. In other words, the same difference of sample means would be regarded as stronger evidence against the null hypothesis of no difference in the case where the sample difference was a more precise (less uncertain) estimate of the population difference. This connection between sample size, level of uncertainty and strength of evidence should be intuitively obvious, in light of the discussion in this section.

To illustrate the calculations, consider the length of the interview in the ESS example. There $\hat{\Delta} = 1.84$,

$$\hat{\sigma} = \sqrt{\frac{1912 \cdot 57.64^2 + 1495 \cdot 35.78^2}{1913 + 1496 - 2}} = 49.26,$$

$$\hat{\text{se}}(\hat{\Delta}) = 49.26 \cdot \sqrt{\frac{1}{1913} + \frac{1}{1496}} = 1.70$$

and

$$t = \frac{1.84}{1.70} = 1.08$$

as shown in Table 2.1. For the data on weekly working hours, $t = 2.96$. In practice, of course, these calculations will usually be done using statistical software such as SPSS.

2.4.3 Sampling distribution of the test statistic

The sampling distribution of a test statistic needs to be known only under the assumption that the null hypothesis is true. For the two-sample t-statistic (2.8) it is a t distribution with $n-2$ degrees of freedom, where $n = n_1 + n_2$ is the total sample size. In our example of interview lengths, this gives $n = 1913 + 1496 = 3409$ and thus 3407 degrees of freedom. The curve (in more technical terms, probability density function) of this distribution is shown in Figure 2.1.

An alternative result about the same distribution is obtained by using the **Central Limit Theorem** (CLT), the most important result of all of mathematical statistics. In its simplest form, the CLT refers to the sampling distribution of a single sample mean, and states that

- If $Y_1, Y_2, \ldots, Y_n$ are independent observations from (almost) any distribution with a population mean μ and variance σ^2, and if n is reasonably large, the sampling distribution of their sample mean $\bar{Y} = \sum_i Y_i/n$ is approximately a normal distribution with mean μ and variance σ^2/n.

More generally, the theorem applies also to many other quantities[4], including the sample difference $\hat{\Delta} = \bar{Y}_2 - \bar{Y}_1$ in the t-test statistic. Combining this with another theorem[5] gives the following result:

- When the null hypothesis $H_0 : \Delta = 0$ is true and the sample sizes n_1 and n_2 are large enough, the sampling distribution of the two-sample t-statistic (2.8) is approximately a standard normal distribution (i.e. a normal distribution with mean 0 and variance 1). This approximation is sufficiently accurate when n_1 and n_2 are both at least 20.

This result implies that when applying the test, we do not in practice need to be unduly concerned about the assumption that the population distributions of Y are normal[6]. Furthermore, since a t distribution with large degrees of freedom is very similar to the standard normal, the difference between the two possible sampling distributions (t_{n-2} or standard normal) is unimportantly small when the total sample size n is even moderately large (bigger than 40, say). It is then a matter of convenience which distribution is used to obtain P-values for the test, as explained in the next section. Finally, note that when the standard normal distribution is used as the sampling distribution, a test statistic of the general form (2.10) is often labelled z and referred to as a z-test rather than a t-test. The label is otherwise unimportant.

[4] Essentially anything that is of the same general form as the sample mean, i.e. a sum of something over the units in the sample, divided by something like the sample size.

[5] Slutsky's theorem, which allows us to replace the $\hat{\sigma}$ in the denominator of (2.8) with σ before applying the CLT.

[6] This is good news in our example on interview lengths, where the sample distributions are actually heavily skewed to the right: although most of the interviews last around an hour, some are recorded as much longer (the longest as 668 minutes). It is not clear whether these represent coding errors, total lengths of interrupted interviews or cases where the interviewer was treated to many cups of tea.

2.4.4 P-values

The logic of drawing conclusions from a significance test is to consider what we would expect to see if the null hypothesis was in fact true in the population, and compare that to what was actually observed in our sample. The null hypothesis is then supported if the observed data, specifically the observed value of the test statistic, is in line with what we would expect under the null hypothesis, while a value which would be surprising (i.e. unlikely) under the null hypothesis will be regarded as evidence against the hypothesis. The level of plausibility of the observed value of the test statistic is quantified by the P-value:

- The **P-value** is the probability, if the null hypothesis was true in the population, of obtaining a value of the test statistic which provides as strong or stronger evidence against the null hypothesis, and in the direction of the alternative hypothesis, as the value of the test statistic in the sample actually observed.

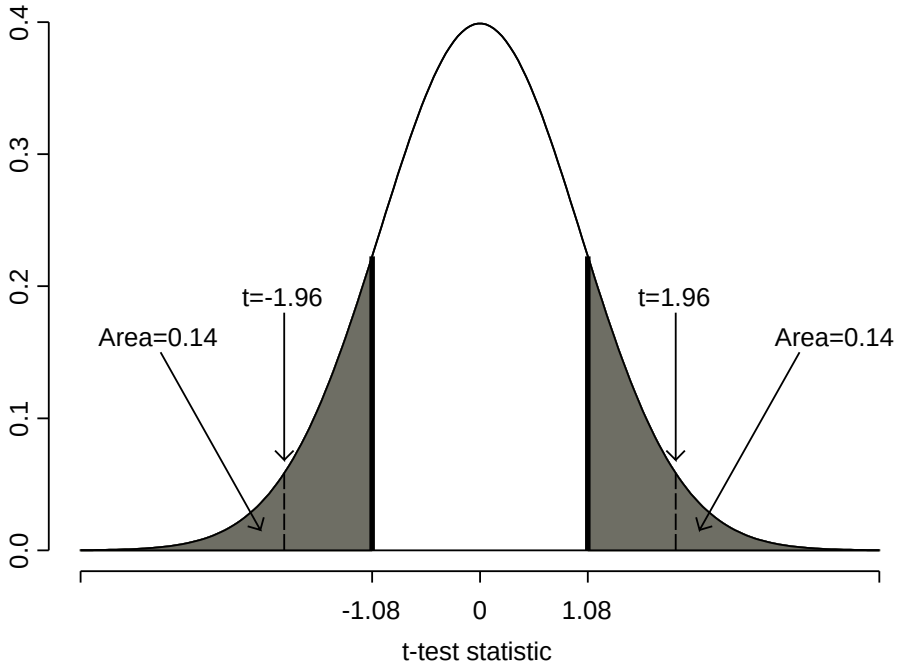
In the case of the two-sample t-test, this means the probability of a value which is as far of further from zero as the observed t, in either direction if the alternative hypothesis is two-sided or in the direction of the alternative if this is one-sided. For example, for the interview length variable we obtained $t = 1.08$. If the two-sided alternative hypothesis $H_a: \Delta \neq 0$ is used, the P-value is the probability of values as far or further from 0 than 1.08, i.e. the values $t \leq -1.08$ and $t \geq 1.08$. The P-value against the one-sided alternative hypothesis $H_a : \Delta > 0$ is then the probability of $t \geq 1.08$, and that against $H_a : \Delta < 0$ is the probability of $t \leq 1.08$.[7]

The P-value is calculated from the sampling distribution of the test statistic under the null hypothesis. This is illustrated for the interview length example by Figure 2.1. The curve in the plot is that of the t distribution with 3407 degrees of freedom (which is indistinguishable from the standard normal distribution), and the two-sided P-value is equal to the combined areas under the curve corresponding to values further than 1.08 from 0, one below and one above zero. These are shown in grey in the plot. Because any t distribution is symmetric, the two areas are of the same size, here both 0.14. The two-sided P-value is thus $0.14 + 0.14 = 0.28$. Similarly, the one-sided P-value against the alternative hypothesis $H_a : \Delta > 0$ is the area of the grey region on the right-hand tail, i.e. 0.14, and the P-value against $H_a : \Delta < 0$ the area of everything but the grey region on the right-hand tail, i.e. $1 - 0.14 = 0.86$. For average weekly working hours, the test statistic is $t = 2.96$ and the two-sided P-value is $P = 0.003$.

For most standard tests, the P-value is produced automatically by statistical software. When it is not available, e.g. if it is required by an exam question but not included in attached output, an approximate P-value can be obtained by comparing the observed value of t to *critical values* for the sampling distribution, as explained on introductory statistics courses. The critical values are obtained from a table of t distributions, which is included in all basic text books (e.g. Table B in the Appendix of Agresti and Finlay). If sample sizes n_1 and n_2 are large enough (e.g. both at least 20), it is sufficient, and more convenient, to use the critical values from the standard normal distribution.

[7] Note that the last of these includes also positive values smaller than 1.08, because they are closer than the observed t to the values claimed by H_a

Figure 2.1: Illustration of the calculation of P-values. Here the value of the t-test statistic is $t = 1.08$, as in the example on lengths of survey interviews in Chapter 2. The areas in grey indicate the two-sided P-values, i.e. the probabilities of values at least as far from 0 as the observed value of t.



These are listed also in Table 2.2, for standard significance levels (discussed below) and for both one- and two-sided tests. We would then report the P-value as being smaller than the significance level corresponding to the largest critical value that the absolute value of t (i.e. dropping the $-$ sign if t is negative) is larger than, or $P > 0.10$ if the absolute value of t is smaller than the critical value for the significance level 0.10[8]. For example, for the interview lengths we have $t = 1.08$, which is smaller than 1.65, so the two-sided P-value is $P > 0.10$. For the working hours variable, $t = 2.96$ is larger than 2.58 but smaller than 3.09, so the two-sided P-value is reported as $P < 0.01$. Both of these agree, of course, with the precise values calculated above, of $P = 0.28$ and $P = 0.003$ respectively.

2.4.5 Conclusions of a significance test

The P-value is the end product of any significance test, and its value should always be reported, either a precise value from computer output or an approximate value obtained using critical values as discussed above. When the P-value is very small, say smaller than 0.001, only an upper bound is needed even when using a computer, so we can report it as $P < 0.001$ (note that SPSS would misleadingly report this as "0.000", even though a P-value is never *exactly* zero).

[8]For one-sided alternative hypotheses, this applies only if t is in the direction of H_a. In the rare cases where it is not, report $P > 0.5$.

Table 2.2: A table of critical values for conventional significance levels for one- and two-sided tests, obtained from the standard normal distribution. The two-sided critical values are also used as multipliers for confidence intervals based on the standard normal distribution, at the confidence levels indicated on the last row.

	Significance level			
Alternative hypothesis	0.10	0.05	0.01	0.001
One-sided	1.28	1.65	2.33	3.09
Two-sided	1.65	1.96	2.58	3.29
(& CI at confidence level of)	(90%)	(95%)	(99%)	(99.9%)

The *smaller* the P-value, the stronger is the evidence *against* the null hypothesis. In our working hours example $P = 0.003$, which would usually be regarded as strong evidence against the null hypothesis (see the dicussion below). There is thus strong evidence that the population means of average weekly working hours are not the same among the British and French working populations. For the interview lengths, on the other hand, $P = 0.28$, which indicates little evidence against the null hypothesis of no difference. There is thus no reason to conclude that ESS interviews carried out in France differ in length from ones done in the UK.

In addition to reporting the P-value, it is quite common to state the conclusion also in the form of a more discrete decision of "rejecting" or "not rejecting" the null hypothesis. This is usually based on conventional reference levels, known as **significance levels** or **α-levels**. The standard significance levels are 0.10, 0.05, 0.01 and 0.001 (also known as the 10%, 5%, 1% and 0.1% significance levels respectively), of which the 5% level is most commonly used; other values than these are rarely considered. The values of the test statistic which correspond exactly to these levels are the critical values discussed in the previous section, for the standard normal distribution the values shown in Table 2.2. When the P-value is *smaller* than a conventional level of significance (i.e. the test statistic is *larger* than the corresponding critical value), it is said that the null hypothesis is *rejected* at that level of significance, or that the results (i.e. evidence against the null hypothesis) are **statistically significant** at that level. When the P-value is larger than a level of significance, the null hypothesis is *not rejected* (but not "accepted", which is neither correct nor the same as "not rejected").

For the interview lengths we obtained $P = 0.28$, so H_0 is not rejected at any conventional level of significance. In contrast, $P < 0.01$ for the working hours, so the null hypothesis is "rejected at the 1 % level of significance", i.e. the evidence that the population mean difference is not zero is "statistically significant at the 1% level" (as well as the 10% and 5% levels of course, but it is enough to state only the strongest level). The conclusion should also be stated with reference to the variables under consideration, for example as

- "The null hypothesis of equal population means is rejected at the 1% significance level ($P = 0.003$). There is strong evidence that the average weekly working hours including overtime of the British and French populations are not the same."; or
- "The average weekly working hours in the British and French samples are statistically significantly different at the 1% significance level ($P = 0.003$)."

2.5 Confidence intervals

A significance test assesses *whether* it is plausible, given the evidence in the observed data, that a population parameter has a particular single value, for example that a population mean difference Δ is equal to 0. A **confidence interval** provides instead a list of all of those values of the parameter which *are* plausible given the data. The interval thus supplements our best estimate (point estimate) of the parameter with a range of values (an **interval estimate**) which are also reasonably well supported by the observed data.

Consider again inference for the difference $\Delta = \mu_2 - \mu_1$ between population means of a variable Y in two groups, for which the point estimate is the sample difference $\hat{\Delta} = \bar{Y}_2 - \bar{Y}_1$. A confidence interval for Δ consists of all the values in the range

$$\text{from} \quad \hat{\Delta} - q_{\alpha/2}\, \hat{\text{se}}(\hat{\Delta}) \quad \text{to} \quad \hat{\Delta} + q_{\alpha/2}\, \hat{\text{se}}(\hat{\Delta})$$

or, in commonly used concise notation,

$$\hat{\Delta} \pm q_{\alpha/2}\, \hat{\text{se}}(\hat{\Delta}) \tag{2.11}$$

where $\hat{\text{se}}(\hat{\Delta})$ is the estimated standard error of $\hat{\Delta}$, and the quantity $q_{\alpha/2}$ is explained below. In fact, all of the confidence intervals used in this coursepack will be of the same general form (2.11), just with different definitions of the parameter Δ and its point estimate and standard error.

The quantity $q_{\alpha/2}$ in (2.11), which multiplies the standard error of $\hat{\Delta}$, is the critical value for the test statistic (2.10) at significance level α. It thus depends on the sampling distribution of the test statistic. As discussed in Section 2.4.3, in the two-sample case we can use either the t-distribution with $n_1 + n_2 - 2$ degrees of freedom or, if both sample sizes are at least 20, the standard normal distribution. The difference between the two is again small and unimportant if the sample sizes are large enough for the normal distribution to be appropriate. Computer software such as SPSS typically produce confidence intervals using multipliers from the t-distribution. If an interval is calculated by hand, it is more convenient to use the normal distribution.

If the critical value $q_{\alpha/2}$ for significance level α is used in (2.11), the confidence interval is said to have the **confidence level** $1 - \alpha$. Conventionally, only the confidence levels 0.90, 0.95 and 0.99 (or the 90%, 95% and 99% levels respectively) are generally used, with the 95% confidence being most common. The multipliers for these levels (as well as the rarer 99.9% level) for the standard normal distribution are shown on the last row of Table 2.2. The symbol $z_{\alpha/2}$ (instead of $q_{\alpha/2}$) is often used for these values. It is worth remembering at least the multiplier $z_{0.025} = 1.96$, which is used to obtain 95% confidence intervals based on the standard normal distribution.

In the two-sample case $\hat{\Delta} = \bar{Y}_2 - \bar{Y}_1$, and its estimated standard error under the model defined in Section 2.2 is again calculated from equation (2.9). If we require the 95% confidence interval (i.e. $\alpha = 0.05$) and use the standard normal distribution as the sampling distribution, the multiplier in (2.11) is $q_{0.05/2} = z_{0.025} = 1.96$, and the 95% confidence interval for Δ is obtained from (2.11) as

$$(\bar{Y}_2 - \bar{Y}_1) \pm 1.96\, \hat{\sigma} \sqrt{\frac{1}{n_2} + \frac{1}{n_1}}. \tag{2.12}$$

In the interview length example we had $\hat{\Delta} = 1.84$ and $\hat{\text{se}}(\hat{\Delta}) = 1.70$, so (2.12) gives

$$1.84 \pm 1.96 \times 1.70 = 1.84 \pm 3.332 = (-1.49; 5.17).$$

The 95% confidence interval for the population difference Δ thus ranges from -1.49 to 5.17, i.e. from the average interview length in the population being between 1.49 minutes longer in the UK to 5.17 minutes longer in France. The result of a confidence interval like this is conventionally stated as something like

- "The 95% confidence interval for the population difference in average interview lengths between France and Britain is -1.49 to 5.17 minutes."; or
- "We are 95% confident that the difference in average interview lengths between France and Britain is between -1.49 and 5.17 minutes."

In our other example, the 95% confidence interval for the difference in average weekly working hours is $(0.54; 2.68)$, as shown in Table 2.1.

We conclude this chapter with two general observations about confidence intervals. The first concerns the width of the interval. We would of course prefer a confidence interval to be narrow, as this would indicate that we had a fairly precise idea of which values the population parameter might have. The width of the interval defined by (2.11) is $2 \times q_{\alpha/2} \times \hat{\text{se}}(\hat{\Delta})$. This depends on two quantities:

- The multiplier $q_{\alpha/2}$, which in turn depends on the confidence level $1 - \alpha$. The higher the confidence level is, the larger the multiplier and the wider the interval. Thus a 99% confidence interval is always wider than a 95% interval for the same data, and wider still than a 90% interval. We thus face an inevitable trade-off between increasing the precision of the interval estimate and decreasing the level of confidence. This explains why we do not consider a 100% confidence interval, as this would contain all possible values of Δ and exclude none. Instead, we aim for a high but not perfect level of confidence, obtaining an interval which contains some but not all possible values, for the price of a small chance of incorrect conclusions.
- The standard error $\hat{\text{se}}(\hat{\Delta})$, which depends, in particular, on the sample size(s). Other things being equal, larger samples yield smaller standard errors and thus narrower intervals. Here the cost of increasing the precision of the confidence interval is thus increasing sample size, which usually also implies increasing cost of data collection.

The second general observation concerns the relationship between confidence intervals and significance tests. Suppose we use the test statistic (2.10) for the null hypothesis $\Delta = 0$ against a *two-sided* alternative hypothesis $\Delta \neq 0$, and the confidence interval (2.11) for Δ, both under the same model assumptions, using the same estimates $\hat{\Delta}$ and $\hat{\text{se}}(\hat{\Delta})$, same sampling distributions and matching significance and confidence levels (e.g. 5% significance level for the test and 95% $(= 100\% - 5\%)$ confidence level for the interval). The following correspondencies then hold in general:

- If the null hypothesis $H_0 : \Delta = 0$ is not rejected by the test, the confidence interval for $\Delta = 0$ contains 0, and vice versa. For instance, this is the case in for the average interview lengths, where $P = 0.28$ and the 95% confidence interval $(-1.49; 5.17)$ does contain $\Delta = 0$.
- If the null hypothesis $H_0 : \Delta = 0$ is rejected by the test, the confidence interval for $\Delta = 0$ does not contain 0, and vice versa. For instance, this is the case for the average weekly working hours, where $P = 0.003$ and the 95% confidence interval $(0.54; 2.68)$ does not contain $\Delta = 0$.

A confidence interval can thus be used also to draw the conclusion of whether the null hypothesis of a two-sided significance test should be rejected. In addition, the interval tells us not only whether a particular single parameter value such as $\Delta = 0$ is consistent with the data, but which values overall are consistent with them. This property makes confidence intervals a very useful tool of statistical inference.

Chapter 3

Review of simple linear regression models

3.1 Introduction

This chapter provides a short review of the basic elements of simple linear regression models, which involve only one explanatory variable X and one interval-level response variable Y. The description is quite concise, for two reasons. First, we assume that you have encountered this material before (or, if not, will supplement this review by reading the corresponding sections of the text book). Second, the results listed here will be repeated and discussed, often in more detail, in the next chapter on the multiple linear regression model, which contains simple linear regression as a special case.

Exceptions to the telegraphic style of this chapter are Sections 3.2 and 3.3, which discuss the nature of statistical models in more general terms. The observations made there apply to all of the regression models considered on this course.

3.2 Nature of a statistical model: an illustration

To introduce the idea of statistical models in general and simple linear regression models in particular, we will first discuss an old application from a quite different field than the ones we usually study. The considerations that arise there will then be translated to a social science context in Section 3.3.

What is the shape of the Earth?[1] The ancient Greeks were the first to get as far as "round", with the Pythagorean school the first to propose the idea of a spherical Earth in the 6th century BCE, and Eratosthenes of Alexandria the first to obtain an estimate of its size based on solid empirical evidence, in the 3rd century BCE.

[1] General information used here comes from the article "geoid" in *Encyclopedia Britannica* (`http://www.britannica.com/eb/article-9036465`; accessed 17 July 2006). Historical information on the statistical analysis of the question comes from Stephen Stigler's magisterial *The History of Statistics*, Cambridge, MA: Belknapp Press, 1986.

In his *Principia* in 1687, Sir Isaac Newton proposed that a Copernican Earth, orbiting the Sun and rotating about its own axis, would be a *spheroid*, essentially a somewhat flattened sphere. Specifically, it should be an *oblate* spheroid, flattened at the poles, rather than a *prolate* one, flattened at the equator.

The geometry of a spheroid suggests a way of examining the shape of the Earth empirically, by using geodetic triangulation to measure the distance covering one degree of latitude along a meridian (line of constant longitude) in different latitudes. If we denote by Y this length centered at θ degrees of latitude, and let $X = \sin^2\theta$, the following equation holds approximately for any spheroid:

$$Y = \alpha + \beta X. \tag{3.1}$$

Here α and β are unknown quantities (*parameters*): α is the length of a degree at the equator, and β is the excess (or deficiency) over α of a degree at the poles. As discussed in more detail later, (3.1) implies that the functional relationship between X and Y is described by a straight line.

Equation (3.1) is a simple mathematical *model* which links the two measurable quantities X and Y. If such a model is correct, it can be used for two important purposes:

1. Estimating the values of the parameters α and β based on a set of measurements of X and Y. In this example, β is of particular interest because it is directly related to the general shape of the Earth: $\beta > 0$ indicates an oblate spheroid, $\beta < 0$ a prolate spheroid and $\beta = 0$ an exact sphere.
2. Once estimates of α and β are available, predicting the value of Y for any value of X. Here this means that we can predict the length of a degree at any latitude from the degree of latitude alone, without further measurements.

In fact, these aims can be achieved even when the model is not exactly correct, as long as it is adequate for the purposes for which it is used. This is a crucial observation, since a simple model like (3.1) is hardly ever exactly correct, and it is not here either:

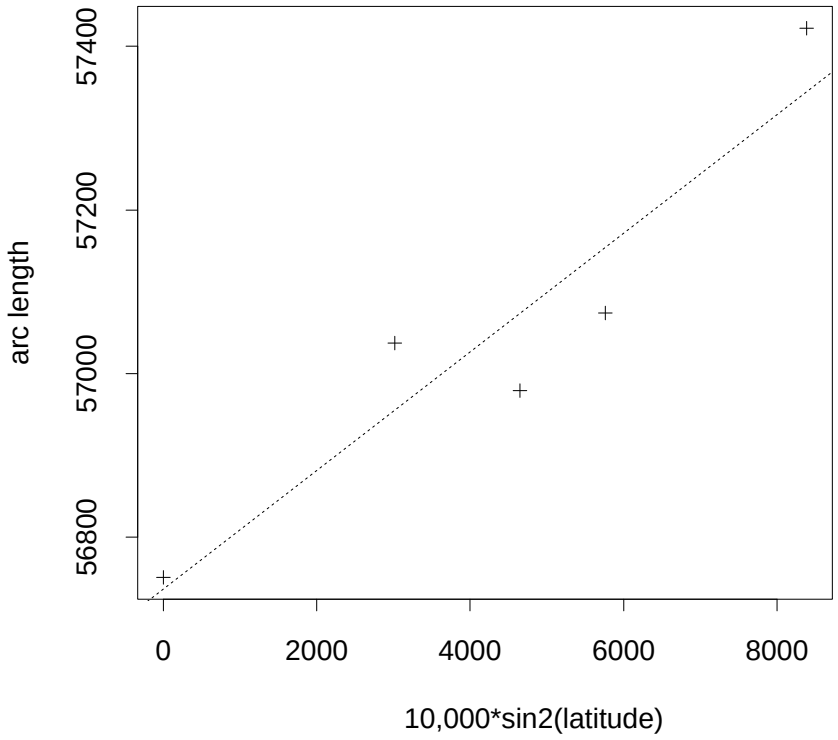
- Both the length of a degree Y and (to a much lesser extent) the degree of latitude (and thus X) may in practice be measured with some error, so empirical measurements of them are typically not exactly on the straight line implied for the corresponding theoretical quantities by equation (3.1). An example of this is given in Figure 3.1, which shows a scatterplot of five measurements of X and Y at a wide range of latitudes, published in 1755 by Roger Boscovich (Rudjer Bošković)[2]. The model for the measurements should thus not be a *deterministic* model where Y can be predicted exactly from X, but a *probabilistic* model which allows for some random error. An extended model of this kind can be written as

  $$Y = \alpha + \beta X + \epsilon \tag{3.2}$$

  where ϵ is a random variable corresponding to measurement error in Y. Models of this kind are discussed in this chapter, and generalised in the next one.

[2] *De litteraria expeditione per pontificiam ditionem ad dimetiendos duos meridiani gradus et corrigendam mappam geographicam, iussu et auspiciis Benedicti XIV Pont. Max. suscepta a Christophoro Maire et Rogerio Josepho Boscovich.* Rome: Palladis, 1755.

Figure 3.1: Five observations of meridian arc length and $\sin^2\theta$, where θ is latitude. The dotted line is the least-squares fitted line for the points.



Source: Boscovich and Maire (1755), reproduced in Stigler (1986). An error noted by Stigler on the recorded latitude of Cape of Good Hope in Boscovich and Mare's data has been corrected here.

- Equation (3.1) is not exactly true even if X and Y were measured perfectly, because the shape of the Earth (the *geoid*) is not in fact an exact spheroid. However, the simple model (3.1) provides a very good approximation of it: the best-fitting spheroid differs from the true geoid by at most ± 100 metres.

Some of the earliest measurements of latitude and degree length used to fit model (3.2) were made by Gian Domenico and Jacques Cassini along a meridian across the whole of France, from Dunkirk to the Spanish border. Their results, published in 1720, suggested a prolate Earth, in a surprising contradiction to Newton's theoretical predictions. Further observations were necessary to resolve the question, preferably on a wider range of latitudes for improved accuracy. The French Academy of Sciences accordingly dispatched two expeditions to the ends of the Earth, one under Bouguer and Condamine to Peru and another under Maupertuis to Lapland. Their results indicated that the Earth was oblate, in agreement with Newton. According to modern measurements, the distance from Earth's centre to the poles is about 21 kilometres shorter than the distance from the centre to the equator.

Measurement of the shape of the Earth is also important in the history of statistics, because it stimulated some of the very first steps in the development of statistical estimation. Consider the observations analysed by Boscovich, shown in Figure 3.1. Suppose first that only two pairs of measurements of X and Y had been available,

Figure 3.2: A portrait of Rudjer Bošković on a Croatian 5-Dinar banknote.

instead of five. It would then be possible to find a straight line (and thus values of α and β) which went through the resulting two points and thus fitted the observed data exactly. This would, of course, not be the "true" line or even a good estimate of it, but it would at least be easy to calculate. When, however, Boscovich (whose portrait can be seen in Figure 3.2) gathered together more than two sets of measurements, which did not fit exactly on a single line, he was faced with a new kind of problem. In effect, he had too much data (not a phrase that comes easily to a modern data analyst) for the simple estimation method of drawing a line through two points. It was then necessary to come up with another approach, in order to identify a single line which fitted the observations as well as possible, even if none of them perfectly. In 1760 Boscovich devised a method, later extended and formalised by Laplace, which was a perfectly sensible solution to the problem, although it was superceded a few decades later by Gauss and Legendre's development of the method of least squares. The least squares estimate of β based on the five observations in Figure 3.1 is $\hat{\beta} = 926$, which correctly indicates an oblate Earth.

3.3 Statistical models in the social sciences

Let us now consider a more familiar-looking example, using some data from the *Global Civil Society 2004/5* yearbook[3]. We will examine the following two variables for $n = 111$ countries:

- Net **primary school enrolment** ratio 2000-01 (%). This will be used as the explanatory variable X.
- **Infant mortality rate** (IMR) 2001 (% of live births). This will be used as the response variable Y.

[3]Anheier, H., Glasius, M. and Kaldor, M. (eds.) (2005). *Global Civil Society 2004/5*. London: Sage. The book gives detailed references to the indices considered here. Many thanks to Sally Stares for providing the data in an electronic form.

Suppose that we would like to model the infant mortality rate (Y) of a country given its school enrolment ratio (X). We might again consider a model of the simple form

$$Y = \alpha + \beta X + \epsilon.$$

This has the same limitations as in the previous example, but to a much greater degree:

- As shown by the scatterplot of the measurements for the 111 countries in Figure 3.3, the observations are not all on a single straight line. For example, even if two countries had the same value of school enrolment, they would not be guaranteed to have exactly the same level of IMR. The error term ϵ in the model aims to capture these deviations. In this example it reflects not only measurement error in infant mortality but also real differences between countries which are not fully described by variation in their school enrolment ratios. We will later have much more to say about such differences, and whether it is appropriate to absorb them all in the error term ϵ.
- The systematic part of the dependence is described by the equation $\alpha + \beta X$. In the previous example this was motivated by a theoretical model for the shape of the Earth, although even there the model was not quite correct. In the case of school enrolment and IMR we do not have an equally detailed theory about the relationship between the variables, and it would clearly be absurd to think that a simple mathematical equation could ever describe the exact "true" form of this relationship. Instead, the model is a simplified, tentative summarization of its basic features. The linear function $\alpha + \beta X$ is commonly used for this purpose, because it is mathematically the simplest specification which allows Y to change (increase or decrease) as X changes.

This spirit of modelling is best summarized by the well-known statement by the statistician George Box[4]:

> *All models are wrong, but some are useful.*

To quote from a previous edition of the MI452 coursepack[5]:

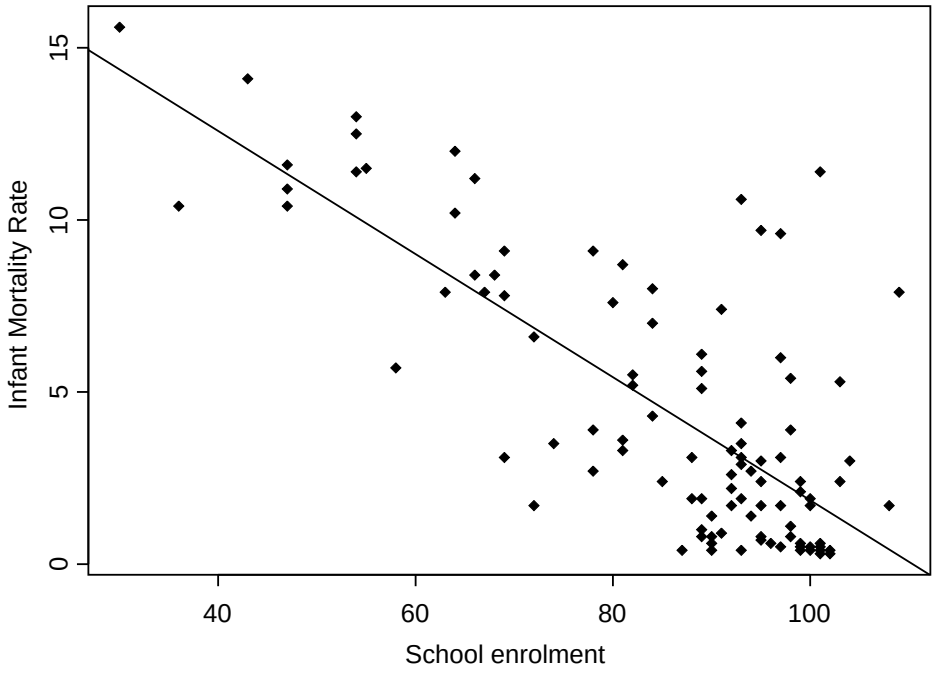
> All models are wrong because the world, especially the social world, is an exceedingly complex place, full of local detail. As social scientists we do not want to reproduce that local detail in our models. What we are trying to do is capture the essentials and leave out the inessentials. A model that was one hundred percent correct would be of no value because it would be as complex as reality itself and if we could understand reality in all its complexity we would have no need for models!

All of the models discussed on this course will be used in this spirit in all of the examples to which they are applied.

[4] This exact phrase apparently appeared first in Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In Launer, R.L. and Wilkinson, G.N., *Robustness in Statistics*, pp. 201–236.

[5] This paragraph was written by Colin Mills.

Figure 3.3: A scatterplot of net primary school enrolment ratio vs. Infant mortality rate for countries in the Global Civil Society data set ($n = 111$). The solid line is the best-fitting (least squares) straight line for the points.



Even (and, in the light of the above, especially) a wrong model can be exceedingly useful for the purposes of explaining and predicting variables (c.f. Section 1.5). Of course, this does not apply to all models, and very many will be both wrong and useless. The results from a model should be seriously presented and interpreted only if the model is deemed to be reasonably adequate for the observed data. For the simple linear regression model, this can be partly assessed by examining whether the scatterplot between X and Y appears to be reasonably consistent with a linear relationship. Further comments on the assessment of model adequacy will be given in various sections throughout the coursepack.

One way of thinking about the models discussed above is as the sum of two general elements:

$$\text{Response variable} \quad = \quad \left[\begin{array}{c}\text{Systematic part} \\ \text{(depends on explanatory variables)}\end{array}\right] \quad + \quad [\text{Random part}]$$

All of the models considered on this course have these elements, even though this is not always obvious from the model formulas. Here the "Systematic part" ($\alpha + \beta X$ for model 3.2) describes those features of the relationship between the response and the explanatory variables which will be used for explanation of and prediction from this relationship. The "Random part" (ϵ in 3.2) captures that part of the variation in the response variable which is not due to its associations with the explanatory variables. Explicitly allowing for such random, unexplained elements in observed data is the defining feature of a *statistical* model.

3.4 Elements of a simple linear regression model

3.4.1 Definition and assumptions

Suppose we have data on n pairs of observations (Y_i, X_i), $i = 1, \ldots, n$, of two variables X and Y. We will consider the following statistical model for them:

1. Observations Y_i are statistically independent of each other.
2. Observations Y_i are a random sample from a population where Y_i has a normal distribution with mean μ_i and variance σ^2.
   - Note that the variance σ^2 is assumed to be the same for all units i, i.e. it is assumed that it does not depend on X_i. This is known as the assumption of *homoscedasticity*.
   - The assumption that the population distribution is normal, although typically included, is not strictly speaking necessary for some purposes.
3. The mean μ_i of Y_i for each unit i depends on the value of the explanatory variable X_i through the linear function

$$\mu_i = \alpha + \beta X_i \tag{3.3}$$

   where α and β are unknown population parameters.

This specification is rather similar to that of the two-sample model considered in Section 2.2. The similarity is not accidental, as the two-sample model is in fact a special case of the model considered here, obtained with a particular definition of X.

The specification of the model is focused on the response variable Y, and says little about the explanatory variable X. In particular, the values X_i are not assumed to be a random sample from any population. While they can be such a sample, they may also be fixed by the researcher, as often happens, in particular, in designed experiments.

A common formulation of the model, equivalent to the one above, is the one used in the previous section:

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{3.4}$$

for all observations $i = 1, \ldots, n$. Here ϵ_i are unobserved random error terms satisfying the following assumptions:

- All ϵ_i are statistically independent of each other.
- The mean (expected value) of ϵ_i is 0 for all i, not depending on X_i. This is often written concisely as $\mathrm{E}(\epsilon_i) = 0$, where E denotes the expected value.
- The variance of ϵ_i is σ^2 for all i, not depending on X_i; in shorter notation, this is written as $\mathrm{var}(\epsilon_i) = \sigma^2$.
- All ϵ_i are normally distributed. Again, this assumption is conventionally included although not absolutely essential.

These specifications define the simple linear regression model:

- **Simple** because it has only one explanatory variable, as opposed to *multiple* linear regression models which will have more than one.
- **Linear** because is a linear function of the *parameters* α and β. The specification (3.3) is also linear in the explanatory variable X, meaning that it defines a linear (straight-line) relationship between X and μ. Later, however, we will encounter more flexible linear models which do not need to be linear in the explanatory variables. This subtlety will be discussed in Section 4.6.3.
- **Regression**: these days, a "regression model" typically refers to any statistical model which describes how response variables depend on explanatory variables. The original reasons for the use of the word in this context are somewhat accidental and of no relevance here[6].
- **Model**, because this is a statistical model in the sense discussed above.

3.4.2 Interpretation of the model parameters

The simple linear regression model has three parameters, α, β and σ^2. The first two of these appear in the model for the conditional mean $\mu = \alpha + \beta X$ of Y given X in the population (dropping the subscript i for now for notational simplicity). The parameters α and β in this formula are known as **regression coefficients**. They are interpreted as follows (a graphical summary of these results is given in Figure 3.4):

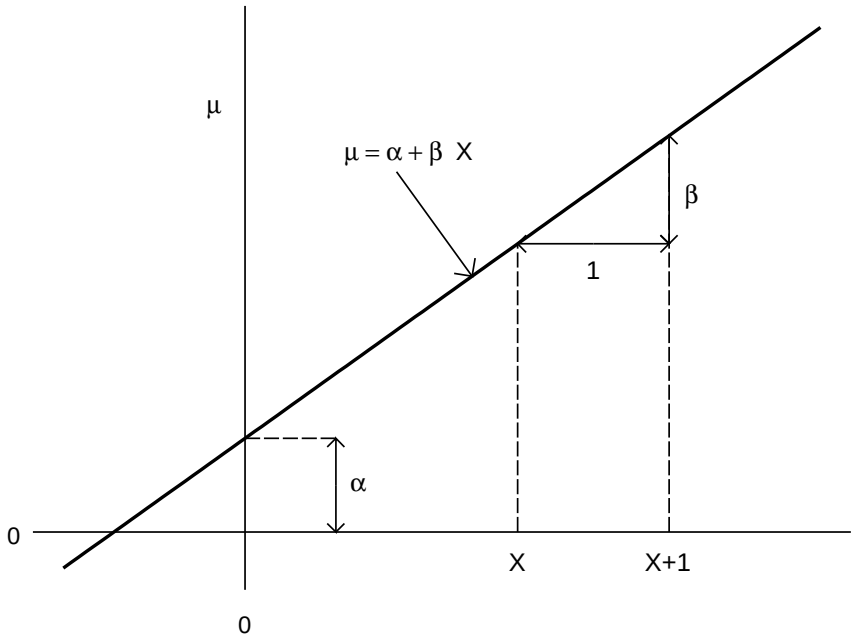
- α is the expected value of Y when X is equal to 0. It is known as the **intercept** or **constant** term of the model. Because 0 is typically not a specifically interesting value of X, α is usually not interpreted in any detail.
- β is the change in the expected value of Y when X increases by 1 unit.
  - To see this, compare two observations, one with value X of the explanatory variable, and the other with one unit more, i.e. $X + 1$. The corresponding means of Y are

with $X+1$:	μ	$= \alpha + \beta \times (X+1)$	$= \alpha + \beta X + \beta$	
with X	: μ		$= \alpha + \beta X$	
Difference	:			β

  The parameter β is known as the **slope** term or the **coefficient of** X. It is typically the only parameter of substantive interest in the model, because it describes the association between X and Y, in the above terms of expected changes in Y corresponding to changes in X. The sign of β indicates the direction of the association. When β is positive, the regression line slopes upwards and increasing X thus also increases the expected value of Y — in other words, the association between X and Y is positive. This is the case illustrated in Figure

[6]The term was first used in Galton, F. (1886). "Regression towards mediocrity in hereditary stature". *Journal of the Anthropological Institute*, **15**, 246–263. The original context is essentially the one discussed on courses on research design as "regression toward the mean".

Figure 3.4: Illustration of the interpretation of the regression coefficients of a simple linear regression model.



3.4. If β is negative, the regression line slopes downwards and the association is also negative. This is the case for the estimated coefficients for IMR given school enrolment in our example, as shown in Figure 3.3. Finally, if β is zero, the line is parallel with the X-axis, so that changing X does not change the expected value of Y. Thus $\beta = 0$ corresponds to no (linear) association between X and Y.

The third parameter of the simple regression model is σ^2. This is the variance of the conditional distribution of Y given X. It is also known as the **conditional variance** of Y, the **error variance** or the **residual variance**. Similarly, its square root σ is known as the conditional, error or **residual standard deviation**; it is in fact more convenient to consider σ rather than σ^2, because σ is on the same scale as the measurements of Y. To understand σ, let us consider a single value of X, say school enrolment of 85 in Figure 3.3. The model specifies a distribution for Y given any such value of X. If we were to (hypothetically) collect a large number of observations, all with this same value of X, the distribution of Y for them would describe the conditional distribution of Y given that value of X. The model states that the average of these values, i.e. the conditional mean of Y, is $\mu = \alpha + \beta X$, which is the point on the regression line corresponding to X. The individual values of Y, however, would of course not all be on the line but somewhere around it, some above and some below.

This distribution of Y given a value of X (which is typically assumed to be a normal distribution) is illustrated by Figure 9.19 (on p. 339) of Agresti and Finlay's text book. The bell shape of the distribution indicates that most of the values of Y for a given X will be close to the regression line, and only small proportions of them far from it. The residual standard deviation σ is the standard deviation of this conditional distribution.

Figure 3.5: SPSS output for a simple linear regression model for Infant mortality rate given School enrolment in the Global Civil Society data.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.753[a]	.567	.563	2.6173

a. Predictors: (Constant), Net primary school enrolment ratio 2000-2001 (%)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	976.960	1	976.960	142.621	.000[a]
	Residual	746.653	109	6.850		
	Total	1723.613	110			

a. Predictors: (Constant), Net primary school enrolment ratio 2000-2001 (%)

b. Dependent Variable: Infant Mortality Rate 2001 (% live births)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	19.736	1.313		15.028	.000	17.133	22.339
	Net primary school enrolment ratio 2000-2001 (%)	-.179	.015	-.753	-11.942	.000	-.209	-.149

a. Dependent Variable: Infant Mortality Rate 2001 (% live births)

In essence, it thus describes how tightly concentrated values of Y tend to be around the regression line. This is usually not of direct interest for interpretation, but it will be a necessary component of some parts of the analyses discussed below.

3.4.3 Estimation of the parameters

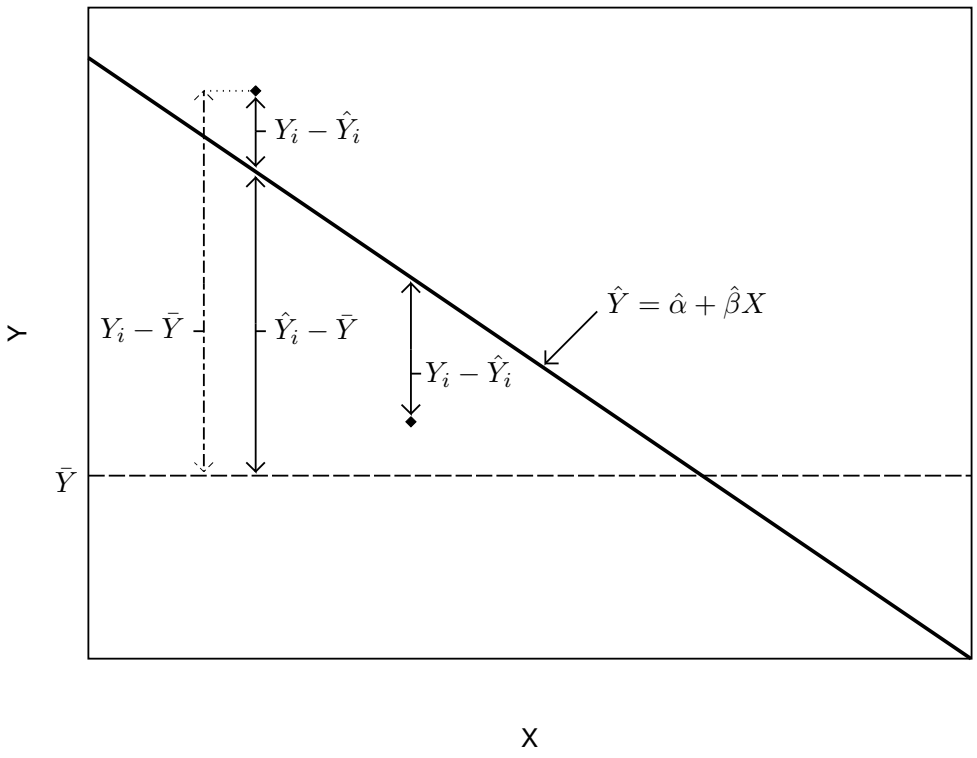
SPSS output for the model for infant mortality given school enrolment ratio fitted to the Global Civil Society data set is shown in Figure 3.5. All of the estimates and other sample statistics for this example quoted in the rest of this chapter appear somewhere in the output (together with several statistics not discussed here). When a statistic is mentioned in the text, you should check that you can find the same number in the output. The contents of the output will be explained in more detail in Section 4.3.1.

In this coursepack, estimates of the regression coefficients α and β will be denoted by $\hat{\alpha}$ and $\hat{\beta}$ ("alpha-hat" and "beta-hat") respectively (other notations are also often used, e.g. a and b). Once they are available, we can also calculate **fitted values**

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

for Y given any value of X. In particular, fitted values $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ can be calculated at the values X_i of the explanatory variable X for each unit i in the observed sample. These can then be compared to the corresponding values Y_i the response variable actually had. Their differences $Y_i - \hat{Y}_i$ are known as the (sample) **residuals**.

Figure 3.6: Illustration of the quantities involved in the definitions of least squares estimates and the coefficient of determination R^2. See the text for explanation.



These quantities are illustrated in Figure 3.6. This shows the fitted regression line for IMR given School enrolment also shown in Figure 3.3. Also shown are two points (X_i, Y_i) from the same figure; the rest have been omitted to simplify the plot. The point further to the left is the one for Mali, which has School enrolment $X_i = 43.0$ and IMR $Y_i = 14.1$. Using the estimated coefficients $\hat{\alpha} = 19.736$ and $\hat{\beta} = -0.179$ (obtained as explained below), the fitted value for Mali is $\hat{Y}_i = 19.736 - 0.179 \times 43.0 = 12.0$. Their difference is the residual $Y_i - \hat{Y}_i = 14.1 - 12.0 = 2.1$. Because the observed value is here larger than the fitted value, the residual is positive and the observed value is above the fitted line. The second point shown in Figure 3.6 corresponds to the observation for Ghana, for which $X_i = 58.0$, $Y_i = 5.7$, and the fitted value is $\hat{Y}_i = 19.736 - 0.179 \times 58.0 = 9.4$. The residual is then $Y_i - \hat{Y}_i = 5.7 - 9.4 = -3.7$, which is negative as the observed value is here smaller than the fitted value.

It remains to define a method for calculating specific values for the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$. In doing so, we are faced with the task of identifying a regression line which provides the best fit to the observed points in a scatterplot like Figure 3.3. Each possible choice of $\hat{\alpha}$ and $\hat{\beta}$ corresponds to a different regression line, and some choices are clearly better than others. To make such considerations explicit, the residuals can be used as a criterion of model fit. The aim will then be to make their total magnitude as small as possible, so that the fitted line is as close as possible to the observed points Y_i in some overall sense. This cannot be done simply by adding up the residuals, because they can have different signs, and positive and negative residuals could thus cancel out each other in the addition. The way around this is to remove the signs by considering the squares of the residuals. Summing these over all units i in the sample

leads to the sum of squared residuals

$$SSE = \sum (Y_i - \hat{Y}_i)^2.$$

Here SSE is short for Sum of Squares of Errors (it is also often called the Residual Sum of Squares or RSS). This is the quantity used as the criterion in estimating regression coefficients for a linear model. Different candidate values for $\hat{\alpha}$ and $\hat{\beta}$ lead to different values of $\hat{Y}_i$ and thus of SSE. The final estimates are the ones which give the smallest value of SSE. Their formulas are

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{s_{xy}}{s_x^2} \tag{3.5}$$

and

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \tag{3.6}$$

where $\bar{Y}$, $\bar{X}$, s_x and s_{xy} are the usual sample means, standard deviations and covariances for Y and X. These values of $\hat{\alpha}$ and $\hat{\beta}$ are known as the **least squares estimates** of the regression coefficients (or as Ordinary Least Squares or OLS estimates), and the reasoning used to obtain them is the **method of least squares**[7]. Least squares estimates for linear regression models are typically obtained using statistical software packages such as Stata and SPSS.

In our example, the estimated coefficients are $\hat{\alpha} = 19.736$ and $\hat{\beta} = -0.179$, and the estimated regression line for expected IMR is thus $19.736 - 0.179X$, where X denotes School enrolment. This is the line shown in Figure 3.3. Because the slope coefficient is negative, the estimated association between the variables is also negative, i.e. higher levels of school enrolment are associated with lower levels of infant mortality. More specifically, every increase of one unit (here one percentage point) in School enrolment is associated with a decrease of 0.179 units (here percentage points) in expected IMR. If we want to express the result in terms of other changes in X than one unit, this is obtained by multiplying $\hat{\beta}$ by the corresponding constant. For instance, the estimated effect of increasing School enrolment by 10 percentage points is $10 \times \hat{\beta} = -1.79$, i.e. an expected decrease in IMR of 1.79 percentage points.

The estimated coefficients can be used to calculate predicted values for Y at any values of X, not just those included in the observed sample. For instance, in the infant mortality example the predicted IMR for a country with School enrolment of 80% would be $\hat{Y} = 19.736 - 0.179 \times 80 = 5.4$. Such predictions should usually be limited to the range of values of X actually observed in the data, and extrapolation beyond these values should be avoided.

The most common estimate of the remaining parameter of the model, the residual standard deviation σ, is

$$\hat{\sigma} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - (k+1)}} = \sqrt{\frac{SSE}{n - (k+1)}} \tag{3.7}$$

where $k = 1$ is the number of explanatory variables in the model, and $k + 1 = 2$ is the number of regression coefficients (α and β) including the constant term α. The

[7] This is another old idea. A description of the method of least squares was first published by Adrien Marie Legendre in 1805 (*Nouvelles méthodes pour la détermination des orbites des comètes*. Paris: Courcier).

quantity $n-(k+1)$, i.e. here $n-2$, is the **degrees of freedom** (df) of the parameter estimates. The formula is given here in this general form involving the symbol k, because we will later be able to use it unchanged also for models with more explanatory variables and thus k greater than 1. In the infant mortality example $n = 111$, $k = 1$, $df = 111 - 2 = 109$, and $\hat{\sigma} = 2.6173$.

3.4.4 Statistical inference for the regression coefficients

The only parameter of the simple linear regression model for which we will describe methods of statistical inference is the slope coefficient β. Tests and confidence intervals for population values of the intercept α are rarely and ones about the residual standard deviation σ almost never substantively interesting, so they will not be considered.

As discussed above, the coefficient β describes the direction and strength of the association between X and Y. It is thus comparable to the mean difference Δ, which was the focus of attention in the two-group comparisons considered in the previous chapter. Indeed, inference for β proceeds almost exactly as in Chapter 2. As there, the only null hypothesis on β discussed here is that its value is zero, i.e.

$$H_0 : \ \beta = 0. \tag{3.8}$$

Expressed in words, this is the hypothesis

$$H_0 : \ \textit{There is no linear association between } \mathrm{X} \textit{ and } \mathrm{Y} \textit{ in the population.} \tag{3.9}$$

Tests of this are usually carried out against a two-sided alternative hypothesis H_a : $\beta \neq 0$, and we will also concentrate on this case.

The tests and confidence intervals involve both the estimate $\hat{\beta}$ and its estimated standard error, which is here denoted $\hat{\mathrm{se}}(\hat{\beta})$. It is calculated as

$$\hat{\mathrm{se}}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum \left(X_i - \bar{X}\right)^2}} = \frac{\hat{\sigma}}{s_x\sqrt{n-1}} \tag{3.10}$$

where $\hat{\sigma}$ is the estimated residual standard deviation given by (3.7), and s_x is the sample standard deviation of X. The last expression in (3.10) shows that the sample size n appears in the denominator of the standard error formula, which implies that the standard error becomes smaller as the sample size increases. As always, the precision of the estimate $\hat{\beta}$ thus increases with increasing sample size.

The test statistic for the null hypothesis (3.8) is again of the general form (2.10) discussed on page 16, i.e. a point estimate divided by its standard error. Here this gives

$$t = \frac{\hat{\beta}}{\hat{\mathrm{se}}(\hat{\beta})}. \tag{3.11}$$

The logic of this is the same as in previous applications of the same idea. Since the null hypothesis (3.8) claims that the population β is zero, values of its estimate $\hat{\beta}$ far from zero will be treated as evidence against the null hypothesis. What counts as "far from zero" depends on how precisely β is estimated from the observed data by $\hat{\beta}$ (i.e.

how much uncertainty there is in $\hat{\beta}$), so the test statistic is obtained by standardising $\hat{\beta}$ by dividing by its standard error.

When the model defined in Section 3.4.1 (page 30) holds and null hypothesis (3.8) is true, the sampling distribution of the test statistic (3.11) is a t distribution with $n-2$ degrees of freedom (i.e. $n-(k+1)$ where $k=1$ is the number of explanatory variables in the model). The P-value for the test against a two-sided alternative hypothesis $\beta \neq 0$ is then the probability that a value from a t_{n-2} distribution is at least as far from zero as the value of the observed test statistic.

As for the test of two means discussed in Chapter 2, it would again be possible to consider a large-sample version of the test which relaxes the assumption that Y_i given X_i are normally distributed, and uses (thanks to the Central Limit Theorem again) the standard normal distribution to obtain the P-value. This would in practice be relevant mostly in the rare cases where the P-value for t is for some reason not available in computer output, so that the conclusion has to be drawn by comparing t to critical values for its sampling distribution. It would then be most convenient to use the critical values for the standard normal distribution, given in Table 2.2. Usually, however, P-values from the t distribution are reported in standard computer output for linear regression, so only they will be discussed here. The difference between P-values from the t_{n-2} and standard normal distributions is in any case minimal when the sample size is reasonably large (at least 30, say).

In the infant mortality example, the estimated coefficient of School enrolment is $\hat{\beta} = -0.179$, and its estimated standard error is $\hat{\text{se}}(\hat{\beta}) = 0.015$, so the test statistic is

$$t = \frac{-0.179}{0.015} = -11.94$$

(up to some rounding error). The P-value, obtained from the t distribution with $n-2 = 109$ degrees of freedom, is very small, and can be reported as $P < 0.001$. The null hypothesis of no association is thus clearly rejected. The data provide very strong evidence that primary school enrolment is associated with infant mortality rate in the population.

In many analyses, rejecting the null hypothesis of no association will be entirely unsurprising. The question of interest is then not *whether* there is an association in the population, but *how strong* it is. This question is addressed with the point estimate $\hat{\beta}$, combined with a confidence interval which reflects the level of uncertainty in $\hat{\beta}$ as an estimate of the population parameter β. A confidence interval for β with the confidence level $1-\alpha$ is given by

$$\hat{\beta} \pm t_{\alpha/2}^{(n-2)} \hat{\text{se}}(\hat{\beta}) \tag{3.12}$$

where the multiplier $t_{\alpha/2}^{(n-2)}$ is the critical value for a t_{n-2} distribution for a two-sided test at the significance level α. For a 95% confidence interval in the infant mortality example, this is $t_{0.025}^{(109)} = 1.98$, and the endpoints of the interval are

$$-0.179 - 1.98 \times 0.015 = -0.209 \quad \text{and} \quad -0.179 + 1.98 \times 0.015 = -0.149.$$

In this example we are thus 95% confident that the expected change in IMR associated with an increase of one percentage point in School enrolment is a decrease of between

0.149 and 0.209 percentage points. If you are calculating this confidence interval by hand, it is (if the sample size is at least 30) again acceptable to use the multiplier 1.96 from the standard normal distribution instead of the t-based multiplier. Here this would give the confidence interval $(-0.208; -0.150)$.

It is often more convenient to interpret the slope coefficient in terms of larger or smaller increments in X than one unit. As noted earlier, a point estimate for the effect of this is obtained by multiplying $\hat{\beta}$ by the appropriate constant. A confidence interval for it is calculated by multiplying the end points of an interval for $\hat{\beta}$ by the same constant. For example, the estimated effect of a 10-unit increase in School enrolment is $10 \times \hat{\beta} = -1.79$, and a 95% confidence interval for this is $10 \times (-0.209; -0.149) = (-2.09; -1.49)$. In other words, we are 95% confident that the effect is a decrease of between 2.09 and 1.49 percentage points.

3.4.5 Sums of squares and R^2

A number of useful concepts and results in linear regression models involve sample statistics known as *sums of squares*. Consider first two possible models for Y. The first of these is the very simple one where the explanatory variable X is not included at all. The estimate of the expected value of Y is then the sample mean $\bar{Y}$. This is the best prediction of Y we can make, if the same predicted value is to be used for all observations. The error in the prediction of each value Y_i in the observed data is then $Y_i - \bar{Y}$ (c.f. Figure 3.6 for an illustration of this for one observation). The sum of squares of these errors is $TSS = \sum(Y_i - \bar{Y})^2$, where TSS is short for "Total Sum of Squares". This can also be regarded as a measure of the **total variation** in Y_i in the sample (note that $TSS/(n-1)$ is the usual sample variance s_y^2).

When an explanatory variable X is included in the model, the predicted value for each Y_i is $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, the error in this prediction is $Y_i - \hat{Y}_i$, and the error sum of squares is $SSE = \sum(Y_i - \hat{Y}_i)^2$. The two sums of squares are related by

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2. \tag{3.13}$$

Here $SSM = \sum(\hat{Y}_i - \bar{Y})^2 = TSS - SSE$ is the "Model sum of squares". It is the reduction in squared prediction errors achieved when we make use of X_i to predict values of Y_i with the regression model, instead of predicting $\bar{Y}$ for all observations. In slightly informal language, SSM is the part of the total variation TSS "explained" by the fitted regression model, and SSE is the part left "unexplained" by the model. In this language, (3.13) can be stated as

Total variation of Y	=	Variation explained by regression	+	Unexplained variation
TSS	=	SSM	+	SSE

The most common summary of these sums of squares is the **coefficient of determination**, more commonly known as $\mathbf{R^2}$ ("R-squared"). It is a measure of association very often used to describe the results of linear regression models. The R^2 statistic is

defined as

$$R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}. \quad (3.14)$$

This is the *proportion* of the total variation of Y in the sample explained by the fitted regression model. Its smallest possible value is 0, which is obtained when $\hat{\beta} = 0$, so that X and Y are completely unassociated, X provides no help for predicting Y, and thus $SSE = TSS$. The largest possible value of R^2 is 1, obtained when $\hat{\sigma} = 0$, so that the observed Y can be predicted perfectly from the corresponding X and thus $SSE = 0$. More generally, R^2 is somewhere between 0 and 1, with large values indicating strong linear association between X and Y. For our model for IMR given School enrolment, the coefficient of determination is $R^2 = 0.567$.

Because the sums of squares TSS and SSE can be thought of as measures of errors in predicting the sample values of Y_i, the difference $TSS - SSE$ shows the *reduction* in the level of prediction errors when we allow the predicted values $\hat{Y}_i$ to depend on the values X_i of the explanatory variable for each unit, compared to using the overall sample mean $\bar{Y}$ as the prediction for every unit. R^2 then expresses this reduction as a proportion of TSS. In our example $R^2 = 0.567$, so using each country's level of School enrolment to predict its IMR reduces the prediction errors by 56.7% compared to the situation where the predicted IMR is the overall sample mean (here 4.34) for every country. Another conventional way of describing this R^2 result is to say that the variation in rates of School enrolment explains 56.7% of the observed variation in Infant mortality rates.

R^2 is also related to the *correlation* coefficient, a simple measure of linear association between two interval-level variables. In simple linear regression, R^2 is the square of the correlation r between X_i and Y_i. Furthermore, the square root of R^2 is the correlation between Y_i and the fitted values $\hat{Y}_i$. This quantity, known as the **multiple correlation coefficient** and typically denoted R, is always between 0 and 1. It is equal to the correlation r between X_i and Y_i when r is positive, and the absolute value (removing the $-$ sign) of r when r is negative. For example, for our infant mortality model $r = -0.753$, $R^2 = r^2 = 0.567$ and $R = \sqrt{R^2} = 0.753$.

Chapter 4

Multiple linear regression models

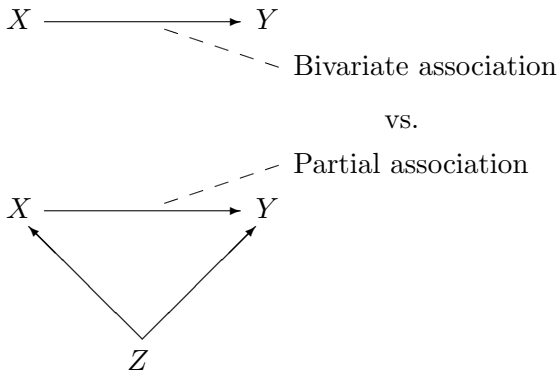
Multiple linear regression modelling is the most widely used of all techniques of statistical modelling. It is employed in situations where the response variable Y is a continuous, interval-level variable, and the explanatory variables can be of any type.

The requirement of a "continuous, interval-level" response variable is in fact often interpreted in a fairly relaxed manner. First, linear models are sometimes fitted for responses which are undeniably discrete, in particular counts of some events; situations in which it is or is not sensible to do this will be discussed further in Section 4.7. Second, the models are frequently used for response variables whose measurement level might strictly speaking be better regarded as ordinal rather than interval. A very common instance of this in the social sciences is modelling of *scales* for attitudes, opinions, beliefs and other such characteristics, obtained by aggregating responses to a set of survey questions on the same topic. Such scales are arguably measured only at an ordinal level, yet they are routinely modelled using linear regression. This is partly a practical matter, as regression models designed specifically for ordinal-level responses are impracticable when the variable has many possible values. For example, in Section 4.3.1 below we fit linear models to an index of general health, aggregated from responses to 22 questions on self-assessed health. This has several hundred distinct values in the data set we analyse.

It can thus be more broadly stated that linear regression models may be considered for modelling a response variable which has a reasonably large number of distinct observed values, and whose measurement level is clearly at least ordinal. This clearly contrasts them with categorical responses with a small number of possible values, for which the kinds of models discussed in Chapters 5–7 should be used.

An example introduced in Section 4.2 will be used in this chapter to illustrate elements of linear models. Section 4.1 discusses some of the reasons why multiple regression models with several explanatory variables are fundamentally more powerful than simple models. The basic elements of the model are introduced in Section 4.3, parameter estimation in Section 4.4, and statistical inference in Section 4.5. Section 4.6 takes a closer look at different ways in which explanatory variables may enter into the models. The rest of the chapter considers other important questions, such as the specification of models and the assessment of their adequacy.

Figure 4.1: A graphical representation of a bivariate association between X and Y (upper graph) and a partial association controlling for a third variable Z (lower graph).



4.1 Why multiple regression?

Multiple regression modelling is like simple regression, except with more than one explanatory variable used at once. Why is this necessary? In particular, why should we use multiple regression — as we routinely will — even when we are only really interested in how the response variable depends on one particular explanatory variable?

The upper part of Figure 4.1 depicts graphically an analysis of two variables. Here the symbols X and Y and the arrow from X to Y indicate that the variables are considered asymmetrically, so that X is regarded as explanatory for Y. The arrow from X and Y is here also used to indicate that the association has been found to be statistically significant in a bivariate analysis of these variables. For example, we might have fitted a simple linear regression model for Y given X, and found that the estimated regression coefficient $\hat{\beta}$ is statistically significant.

The lower part of Figure 4.1 introduces a third variable, a **control variable** denoted by Z (in practice there will usually be several such variables). For sensible interpretation, a potential control variable should usually be logically prior to Y, and prior to or on an equal footing with X. In other words, it should not be explanatory for Y but a response to X (a so-called intervening variable) and it should not be a response to Y.

The arrows from Z to X and Y in Figure 4.1 indicate possible associations between Z and the other two variables. The dashed arrow between X and Y denotes the effect of X on Y, *controlling for* Z, also known as the *partial effect*. This may well be different from the bivariate effect of X on Y denoted in the upper part of the figure, because of the associations Z may have with X and Y. An association like this, controlling for relevant control variables, is in most contexts regarded as the more fundamental association between X and Y, and will be reported instead of the bivariate association.

What, then, does it mean to "control for" variables in the analysis of associations? Informally, this involves "holding constant" the values of control variable(s) Z while examining the association between X and Y. Multiple regression modelling provides a general way of implementing this, in a way explained in Section 4.3.3. Here we will

first illustrate the ideas in a number of simple cases where Z has only three possible values. These examples are shown in Figures 4.2–4.4.

Consider first Figures 4.2 and 4.3. These show sets of artificial data, generated for this illustration. In Figure 4.2 we see a standard scatterplot of two variables X and Y, together with an estimated least squares regression line. The bivariate association between the variables is clearly quite strong (and statistically significant), and the value of the estimated slope coefficient $\hat{\beta}$ is around 0.8. This is what we would conclude from an analysis using only the simple regression tools of the previous chapter.

The four plots of Figure 4.3 show similar scatter plots. These are all comparable to Figure 4.2 in that the bivariate association is positive, with a slope of around 0.8. Conclusions about the effect of X on Y from simple linear regression would thus be the same in every case. Controlling for a third variable Z, however, changes the picture in different ways in the four examples. Here Z is a variable with three possible values, and units with different values of it are plotted with different symbols in Figure 4.3.

Recall that statistical control essentially means holding the control variable constant when assessing the association between X and Y. Here we can do this literally by considering the regression lines for Y given X separately for each subset of observations with the same value of Z. There are three such lines, one for each value of Z; these are shown as solid lines in the plots.

Plots (a) and (b) of Figure 4.3 illustrate situations where controlling for Z does not change the association between X and Y. In general, this will be the case if the control variable is *not* associated with X and/or Y (graphically, if at least one of the arrows from Z in Figure 4.1 is missing). In plot (a) of Figure 4.3 there is a strong association between Z and X (note how the observations with one value of Z all have small values of X, and so on) but no association between Z and Y. Because of the latter property, the regression line for Y given X is the same at all values of Z, and it is also the same as the bivariate regression line for Y given X alone. This line is shown in the plot.

In plot (b), there is an association between Z and Y (note how some values of Z tend to go together with higher values of Y at all levels of X) but no association between Z and X (note how observations with different values of Z seem to be similarly evenly distributed across values of X). Because Z has an effect on Y, the regression lines for Y given X at different levels of Z are now not the same (we will return to issues like this, and the interpretation of multiple linear regression in general, in later sections). However, each of these lines has the same slope (because there is no interaction between X and Z, a concept explained in Section 4.6.2), which is also the same as the slope of bivariate regression for Y given X because Z is not associated with X. The conclusions about the effect of X on Y would thus be unchanged by controlling for Z.

In plot (c), the slope of the bivariate regression line (the dashed line) is again around 0.8. Here, however, the third variable Z is associated with both X and Y, so the association will be different when we control for Z. The regression lines at all levels of Z are in fact parallel to the horizontal axis, i.e. each has a zero slope. There is thus no association between X and Y at any level of Z. Controlling for Z has here completely explained away the original bivariate association, which is revealed to have been a *spurious correlation*, entirely due to the fact that both X and Y were associated with

the third variable Z. Specifically, both of the associations with Z are here positive. This means that units with high values of X tend also to have high values of Z — and units with high values of Z tend also to have high values of Y. If we then ignore Z, we only observe that high values of X tend to go together with high values of Y, i.e. that the bivariate association between X and Y is strongly positive. However, basing interpretation on this association would be misleading for most purposes.

Plot (d) shows a less dramatic case than plot (c), but still one where controlling for Z changes the conclusions about the association between X and Y. Here that association is not completely eliminated, but it does become weaker when controlling for Z.

To round off this discussion, let us consider a real example with a similar structure. Here the units of analysis are the 50 states of the United States[1]. The explanatory variable X is each state's school expenditure per student in public elementary and secondary schools, and the response variable Y is the average total score on the SAT academic performance test for all the students taking the test in the state. The upper plot of Figure 4.4 shows the scatterplot and fitted regression line for these. There is clearly a negative (and statistically significant) association, suggesting that states which spend more per student on public schools tend to have lower average SAT scores.

The lower plot introduces a third variable (in the role of Z above), the percentage of all eligible students in the state who actually took the SAT. This is strongly negatively correlated with average SAT scores, so states with smaller proportions of students taking the SAT tend to have higher average scores (i.e. it is the brighter students who tend to take the test in such states). It is also positively correlated with school expenditure, i.e. states with high expenditures tend to enter large proportions of their students to the SAT. The effect of these associations can be seen in the lower plot of Figure 4.4. Here the control variable is considered in three groups, corresponding to states in the lowest, middle and highest third of the U.S. distribution of the percentage of students taking the SAT. The plot also shows the regression lines for SAT score given school expenditure, separately for these groups. Compared to Figure 4.3, we observe here some new possibilities of what may emerge from multiple regression models:

- The solid lines are not parallel. In other words, the strength of the association between X and Y is different at different levels of the control variable Z. This is an example of a statistical *interaction*, a subtle but important concept to which we will return in Section 4.6.2[2].
- Two of the solid lines slope upwards. In other words, these partial associations between spending and SAT performance are positive, the opposite of the negative bivariate association. This reversal is known as *Simpson's paradox*. It is a consequence of a particular pattern of positive and negative correlations between the control variable and the other two variables (which is not really paradoxical, so Simpson's "paradox" is a misnomer).

[1]The data are for 1994–95. They were originally collated by D. L. Guber from 1997 *Digest of Education Statistics* published by the U.S. Department of Education. The file used here was obtained from Guber, D. L. (1999). Getting what you pay for: The debate over equity in public school expenditures. *Journal of Statistics Education* [Online], **7(2)**. `www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm`.

[2]The interaction is not actually statistically significant here. A better model for these data is thus obtained by fitting a line with the same slope in all of the groups. This line has a positive slope.

Figure 4.2: A scatter plot of an artificial data set of variables X and Y, with the fitted least-squares regression line. The slope $\hat{\beta}$ of the line is around 0.8.

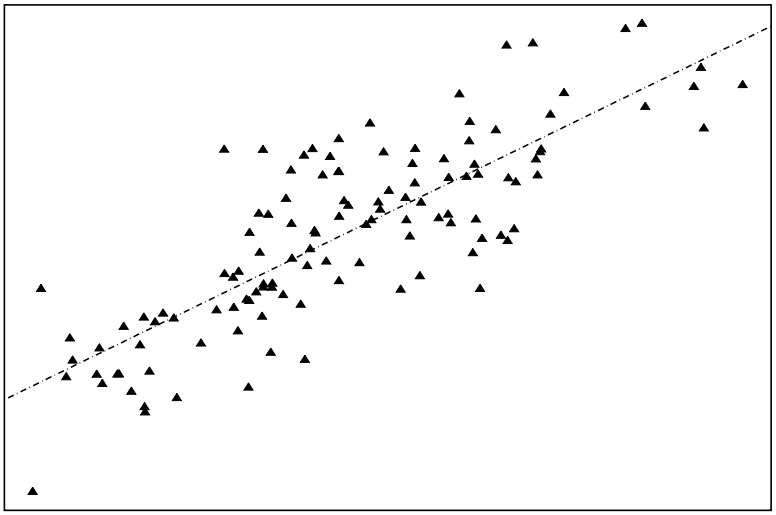


Figure 4.3: Scatter plots of artificial data sets of variables X and Y. The dashed line (where visible) is the least-squares regression line fitted to X and Y alone; its slope is in each plot around 0.8. The plotting symbols indicate observations with three different values of a control variable Z, and the solid lines show the regression lines for Y given X separately at different values of Z.

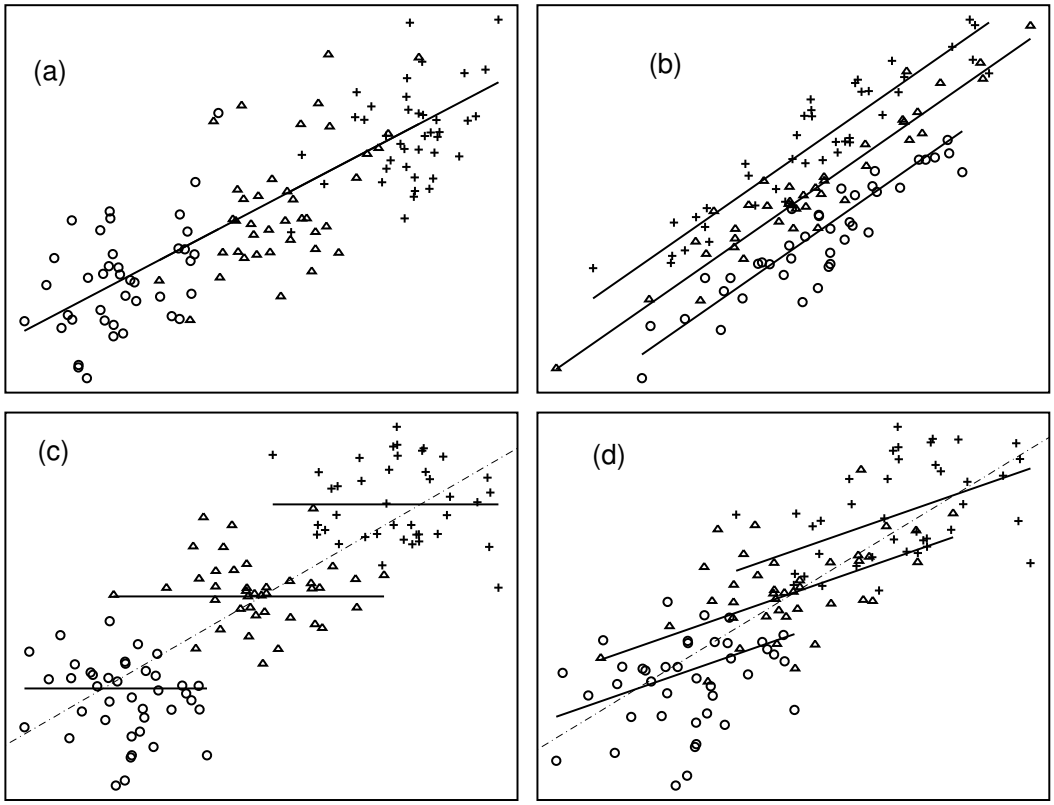
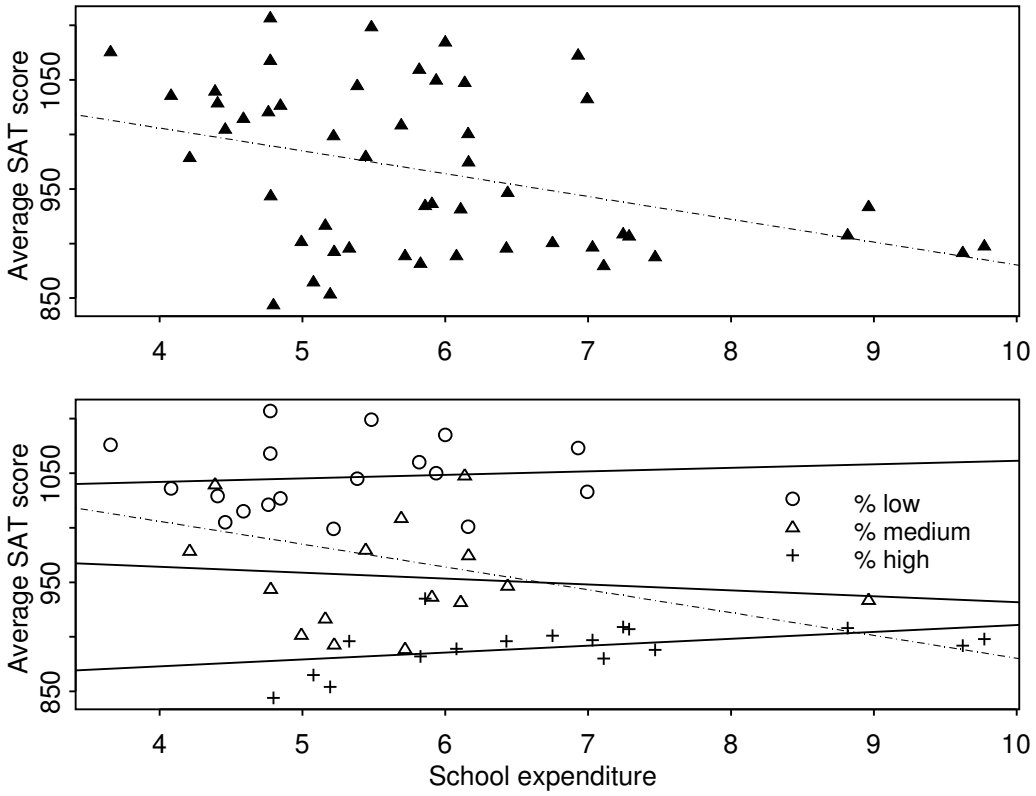


Figure 4.4: Scatterplots of per-student expenditure (in $1000) on public schools vs. average total SAT score for U.S. states in 1994-95. Both plots show the same points, and the dashed line is the same bivariate fitted regression line in both. In the lower plot, states in different tertiles of the percentage of eligible students taking the SAT test are shown with different plotting symbols, and the solid lines show fitted regression lines for SAT score given school expenditure separately in these three groups.



Still other patterns of effects of statistical control are possible, especially in more complex situations with many control variables. In such cases we would not in practice investigate all the effects graphically as above. Instead, we would simply implement the control by using multiple regression models, and see what happens. The general idea is always the same: when controlling for a variable changes some of the estimated associations, it does so because of correlations between the control variables on the one hand, and the other explanatory variables and the response variable on the other.

As noted above, the one situation where controlling for further explanatory variables will *not* change the conclusions about the association between X and Y is when the control variables are not correlated with X or Y or both. Unfortunately, however, this will not usually be the case when we just measure the values of the variables of interest for the units in our data (e.g. respondents in a survey) but have no way of influencing those values: the associations between variables are then whatever they are in the population. Such research designs are known as **observational studies**.

A crucially different research design is possible if the values of X are in the control of the researchers. The best thing to do is then to assign them to the participants at random. If we do this, we have a **randomized experiment** (also known as a randomized

controlled trial or RCT) of the effect of X on Y. The profound difference between a randomized experiment and an observational study is that the random allocation of the values of X to the units of study makes X uncorrelated (on average) with all possible confounding variables Z, known and unknown. In other words, randomization removes the arrow from any Z to X in Figure 4.1. Valid conclusions about the effect of X on Y can then be drawn from bivariate analyses of these variables alone, and statistical control for other variables is unnecessary.

The randomized experiment is the gold standard research design for examining causal questions, and it is widely and routinely used for this purpose in, for example, natural sciences, engineering, agriculture and medicine. In these fields, the complete control of the assignment of the values of the variable (treatment) of interest required by the experiment is both practically and ethically feasible. In the social sciences, on the other hand, this is much less often the case. While randomized *laboratory* experiments are quite often employed in, say, social psychology, *field* experiments on people in their real-life settings are less common, although by no means unknown. For example, in Section 4.2 we describe a large-scale field experiment in health care policy.

Sometimes we use data from randomized experiments also to examine the effects on outcomes of other explanatory variables than the experimentally manipulated treatments. We then need to treat the data as observational, and to use multiple regression as required. In fact, we often include some control variables even when we *do* examine the effects of the experimental treatments, even though the randomization makes this strictly speaking unnnecessary. This may be done for two main reasons. First, it allows us to estimate also the effects of the control variables on the responses, and to examine possible statistical interactions between the controls and the experimental treatments. Second, multiple regression allows us to further control for any imbalances in the distributions of the control variables across the treatment groups which may remain after the randomization (although the effect of this should usually be minor).

So far we have discussed reasons for multiple regression relevant for valid *explanation* of the values of the response variable (c.f. the discussion on page 7). For purposes of *prediction*, the logic is even simpler: it is easier to predict a variable well using information on multiple explanatory variables than a single one only. This, however, should not usually be taken to imply that a prediction model should always include *all* the available variables; this question will be discussed further in Section 4.4.3.

4.2 Example: The Rand Health Insurance Experiment

The RAND Health Insurance Experiment (hereafter abbreviated HIE) was a large, federally funded study conducted in the United States between 1971 and 1982. Its aim was to examine how the cost of health services for individuals affected their use of the services, their satisfaction with health care, the quality of care, and the state of the individuals' health. In brief, the design of the study was as follows:[3]

[3] A detailed description of the study and its main findings can be found in Newhouse, Joseph P. (1993). *Free for All? Lessons from the RAND Health Insurance Experiment.* Cambridge, MA: Harvard University Press. The data from the study used in this coursepack are from Newhouse, Joseph P. *RAND Health Insurance Experiment [In Metropolitan and Non-metropolitan Areas of the United States], 1974–*

- A total of around 2000 nonelderly families were recruited for the study, from four urban and two rural locations across the United States.

- All eligible members of each selected family were included in the study. The most important exclusion criterion was age, with people aged 62 or older not included because they were eligible for the Medicare programme. In the end, 3,565 adults and 1,720 children stayed with the experiment throughout the intended period.

- Each family was allocated into one of fourteen treatment groups, with different types of health insurance.[4] The plans differed mainly along two dimensions:

  - The *coinsurance rate*, which is the fraction of medical charges paid by the individual. This had values of 0, 25, 50 and 95 per cent.
  - The *maximum dollar expenditure* (MDE), which is the maximum amount a family may pay in a 12-month period. This had values of 5, 10, or 15 per cent of family income, up to a maximum of $1000 per year. In one plan (known as the individual-deductible plan), MDE was set at $150 per person, although still subject to a maximum of $450 per family.

  Various modifications and combinations of these were used, resulting in the 14 experimental plans. For example, the most generous plan from a participant's point of view was one with a 0% coinsurance rate (i.e. completely free medical care), and the most stringent a plan with 95% coinsurance and a 15% MDE.

- For the duration of the experiment, the participants transferred their medical insurance to a company set up for the experiment, which paid for their medical care according to the conditions of the experimental plan allocated to them. A system of incentive payments (independent of health care use) was used to ensure that no participant was worse off than they would have been under their previous health insurance.

- Participating families were randomly allocated to one of the 14 insurance plans. HIE was thus a large-scale randomized field experiment.

- Participating families were enrolled in HIE for 3 or 5 years. At the beginning, during and at the end of that period, various types of data were collected from them using interviews, medical tests and insurance claim forms.

The data set used in this chapter contains, for a subset of the participants of the experiment, information on

- the insurance plan to which the person was assigned;

- age, sex, education, income and other background variables, obtained from an interview at the time of enrolment to the study;

1982 [Computer File]. ICPSR Version. Santa Monica, CA: The RAND Corporation [producer], 1986. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1989. Available from `http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/06439.xml`. The data set used here combines variables from HIE data files 91, 160, 163, and 190.

[4]There was also a second comparison involving two groups who received their health care through a Health Maintenance Organization (HMO). These are not included in the analyses of this chapter.

- self-reported health and satisfaction with health care services from interviews at enrolment and at the end of the study;
- physiological measurements (e.g. of blood pressure, cholesterol and vision) at enrolment and at the end of the study; and
- use of health care services and their costs.

Many analyses of the full HIE data set require statistical methods which go somewhat beyond those covered on this course[5]. The main reason for this is that some of the observations are inevitably correlated, in a way which violates the assumption of statistical independence required by the models considered here. First, HIE data often include several members of the same family, which induces dependencies between observations within a family. Second, some of the variables, especially those on health care use and costs, are recorded annually, and observations for different years for the same person are likely to be correlated. To avoid complications arising from this, the data set used here has been simplified as follows:

- From each family, only the person recorded as head of household is included. This leaves a total of 1932 people, all of them between the ages of 16 and 62. It was HIE practice to record a woman as head of household whenever possible, so 87% of the people in the data set are women, and all the men resided in families with no adult female head of household.
- For the expenditure variable considered in Section 4.7, data are used only for one year, each respondent's second year of enrolment in the study.

Elsewhere too the analyses described here have often been somewhat simplified, where the interests of clarity for teaching and the best possible analysis for substantive purposes did not coincide. The results should thus not be treated as definitive answers to the research questions of the experiment. Anyone interested in such conclusions should consult Newhouse (1993) and the publications cited therein.

Summary statistics for all the HIE variables considered somewhere in this chapter are shown in Tables 4.1 and 4.2. Note that the sample sizes are not the same for all the variables, because some had missing values for some respondents.

To conclude this section, let us briefly relate HIE to some of the concepts discussed in Section 4.1. Suppose first that we decided to carry out an observational study to examine the research question, i.e. the effect of health insurance on expenditure and health outcomes. This would be relatively straightforward (if still expensive) to do, e.g. by carrying out and analysing a population survey of the relevant characteristics. As noted before, we would then have no control over the correlations between level of health insurance the respondents had and other characteristics of them, such as age, education, occupation and income, that are also associated with health and the other outcomes of interest. Such correlations would be very likely to be present and even quite strong. To have any chance of giving useful answers, the analysis would then have to control for these background variables.

[5]In particular, models which allow observations to be correlated between units or over time are discussed on the courses ST416 (Multilevel modelling) and ST442 (Longitudinal data analysis).

Table 4.1: Summary statistics for the continuous variables from the HIE data used in Chapter 4. Unless stated otherwise, the variables are from the time of enrolment to the study.

Variable	n	Min	Max	Mean	st.dev.
General Health Index	1878	5.7	100	69.4	15.3
General Health Index (at exit)	1804	8	100	67.4	15.8
Diastolic blood pressure (mmHg)	1151	36	134	75.1	11.1
Diastolic blood pressure (mmHg, at exit)	1801	40	130	78.3	12.1
Outpatient medical expenses in year 2 ($)†	1532	2	3159	185	260.6
Age (years)	1932	16	62	35.9	11.9
Weight (kg)	1851	37	159	65.8	15.0
Family size	1932	1	9	2.2	1.3
Education (years)	1916	0	25	12.1	2.7
Family income (year before enrolment, 1000$)	1837	0	32.3	10.2	6.0
Adjusted family income (1000$)‡	1932	0	49.8	12.0	7.1
Work experience (months)	1821	0	600	124	109

† Including only respondents for whom these expenses were not 0.

‡ Average of two years before enrolment, adjusted for family size and cost of living.

Table 4.2: Percentages of respondents in the levels of the categorical variables from the HIE data used in Chapter 4. All the variables are from the time of enrolment to the study.

Variable	Category					n
	Free care	25% CI	50% CI	95% CI	Individual deductible	
Insurance plan (by coinsurance rate)	32.3	19.9	6.5	18.7	22.5	(1932)
	Female	Male				
Sex	87.3	12.7				(1932)
	Never smoked	Former smoker	Current smoker			
Cigarette smoking status	46.8	13.7	39.5			(1876)
	Excellent	Good	Fair	Poor		
Self-assessed current health	40.1	45.4	11.6	3.0		(1887)

The HIE, however, was not an observational study but a randomized experiment. Clearly it had been decided that the research question was of such public importance, and the difficulties with even the best observational studies so formidable, that a randomized experiment was warranted. At the same time, it is also obvious that carrying out the study will have involved immense financial, administrative, legal and ethical challenges, on a scale which would have made them insurmountable in many other contexts. The HIE thus also serves as an illustration of why the randomized field experiment can never be the research design for all studies in the social sciences.

4.3 Basic elements of the model

4.3.1 Example and computer output

In this section, we will use for illustration data on five variables from the Health Insurance Experiment. As the response variable, we will consider

- Y: General Health Index (GHI). This is a summary measure of a person's self-assessed general health, based on a set of 22 survey questions. Each of these items was answered on a five-point scale, and the results were then summated over the 22 items and transformed to a 0–100 scale, where higher scores represent better health. This index is here treated as a continuous, interval-level variable. Newhouse (1993) characterises the magnitude of its values with the example that "a five-point reduction in the GHI is equivalent to being diagnosed as hypertensive".

Four potential explanatory variables for each respondent will initially be considered:

- X_1: age in years
- X_2: education, in years of school completed
- X_3: family income in the year preceding enrolment, in thousands of 1973 dollars
- X_4: work experience, measured by the number of months the person has been employed since leaving full-time education

The analyses use data for $n = 1699$ respondents for whom the values of all the variables were observed. Summary statistics for the variables are shown in Table 4.1. All of these variables were measured at the time of enrolment to the study. We are thus not yet considering the experimental conditions and subsequent outcomes.

Figure 4.5 contains basic SPSS output for a multiple linear regression model for Y given X_1, X_2 and X_3 for these data. You can refer to this to see where elements of the estimated model discussed in this chapter appear in such output. In summary, the following entries are included:

1. In the **Model Summary** table:
   (a) "R" is the multiple correlation coefficient $R = \sqrt{R^2}$. (p. 60)
   (b) "R Square" is the coefficient of determination R^2 (equation 4.9 on p. 60).
   (c) "Adjusted R Square" is the adjusted R^2 (equation 4.10 on p. 61).
   (d) "Std. Error of the Estimate" (which is in fact a thoroughly misleading lable) is the estimated residual standard error $\hat{\sigma}$ (equation 4.6 on p. 57).
2. In the **ANOVA** [Analysis of Variance] table:
   (a) "Sum of Squares" column shows the Model sum of squares SSM (row labelled "Regression"), Error sum of squares SSE ("Residual") and Total sum of squares TSS ("Total") (decomposition 4.8 on p. 60).
   (b) "df" shows the degrees of freedom of SSM (k), SSE ($n-(k+1)$) and TSS ($n-1$).
   (c) "Mean Square" shows the model mean square (SSM/k) and error mean square ($SSE/[n-(k+1)]$).
   (d) "F" is the overall F-test statistic for testing the null hypothesis that none of the explanatory variables are associated with Y (equation 4.17 on p. 67). This is the ratio of the model mean square and the error mean square.
   (e) "Sig." is the P-value of the overall F-test.
3. In the **Coefficients** table, the row labelled "(Constant)" shows results for the estimated intercept (constant) term $\hat{\alpha}$, and each of the other rows show results for the coefficient of the explanatory variable named in the first column:
   (a) "B" shows least squares estimates $\hat{\alpha}$ and $\hat{\beta}_j$ of the coefficients (Section 4.4.1).
   (b) "Std. Error" shows the estimated standard errors $\hat{\text{se}}(\hat{\alpha})$ and $\hat{\text{se}}(\hat{\beta}_j)$ of the estimated coefficients (Section 4.4.1).
   (c) "Standardized Coefficients: Beta" shows the standardised regression coefficients $\hat{\beta}_j^{\text{std}}$ (equation 4.11 on p. 62)
   (d) "t" shows the t-test statistic for testing the null hypothesis that β_j (or α) is 0 in the population (equation 4.13 on p. 64)
   (e) "Sig." shows the two-sided P-value for t, obtained from the t distribution with $n-(k+1)$ degrees of freedom (Section 4.5.1)
   (f) "95% Confidence Interval for B" shows the 95% confidence interval for β_j (or α) obtained from equation (4.14) on p. 65. This is included in the output only if specifically requested.

The same quantities, possibly with some added or omitted, appear in the linear regression output from any statistical software. Stata output for the same model as the one in Figure 4.5 is shown in Section 4.9.5 on page 107, together with output from three other common packages. In a research report it is usually not advisable to include such output from any package directly. Instead, the relevant model estimates should be transferred to a neatly formatted table, such as the one shown in Table 4.3 and other examples in this coursepack.

Figure 4.5: SPSS output for a linear model for General Health Index given age, education and income in the HIE data.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.247[a]	.061	.059	14.60145

a. Predictors: (Constant), income, education, age

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23525.132	3	7841.711	36.781	.000[a]
	Residual	361377.77	1695	213.202		
	Total	384902.91	1698			

a. Predictors: (Constant), income, education, age

b. Dependent Variable: ghi

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	59.417	2.221		26.747	.000	55.060	63.775
	age	-.128	.032	-.101	-4.029	.000	-.190	-.066
	education	.990	.143	.172	6.906	.000	.709	1.272
	income	.275	.063	.109	4.345	.000	.151	.398

a. Dependent Variable: ghi

Table 4.3: Results for a linear regression model for General Health Index given age, education and income in the HIE data.

Response variable: General Health Index					
Explanatory variable	Coefficient	Standard error	t	P-value	95 % Confidence interval
Constant	59.417				
Age (years)	−0.128	0.032	−4.029	< 0.001	(−0.190; −0.066)
Education (years)	0.990	0.143	6.906	< 0.001	(0.709; 1.272)
Family Income ($1000)	0.275	0.063	4.345	< 0.001	(0.151; 0.398)
$\hat{\sigma} = 14.6$; $R^2 = 0.061$; $n = 1699$; $df = 1695$					

4.3.2 Definition of the model

The data considered here consist of n sets of observations $(Y_i, X_{1i}, X_{2i}, \ldots, X_{ki})$, one for each unit $i = 1, \ldots, n$, where Y is a response variable and $X_1, X_2, \ldots, X_k$ are k different explanatory variables. The multiple linear regression model for such data is an obvious generalisation of the simple linear model defined in Section 3.4.1, extended to include several explanatory variables. The model assumptions are as follows:

1. Observations Y_i are statistically independent of each other.
2. Observations Y_i are a random sample from a population where Y_i has a normal distribution with mean μ_i and variance σ^2.
   - Note that the variance σ^2 is assumed to be the same for all units i, i.e. it is assumed that it does not depend on X_i. This is known as the assumption of *homoscedasticity*.
   - The assumption that the population distribution is normal, although typically included, is not strictly speaking necessary for some purposes.
3. The mean μ_i of Y_i for each unit i depends on the value of the explanatory variables $X_{1i}, X_{2i}, \ldots, X_{ki}$ through the linear function

$$\mu_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \tag{4.1}$$

   where α and $\beta_1, \beta_2, \ldots, \beta_k$ are unknown population parameters.

The only difference to the simple linear model is thus that the mean μ_i now depends on several explanatory variables, each contributing to the linear function (4.1) with its own regression coefficient.

The model is again often expressed in an equivalent alternative form

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i \tag{4.2}$$

for observations $i = 1, \ldots, n$. Here ϵ_i are unobserved random error terms (often called model **residuals**) satisfying the following assumptions:

- All ϵ_i are statistically independent of each other.
- The mean (expected value) of ϵ_i is 0 for all i, not depending on $X_{1i}, X_{2i}, \ldots, X_{ki}$. This is often written concisely as $\mathrm{E}(\epsilon_i) = 0$, where E denotes the expected value.
- The variance of ϵ_i is σ^2 for all i, not depending on $X_{1i}, X_{2i}, \ldots, X_{ki}$; in shorter notation, this is written as $\mathrm{var}(\epsilon_i) = \sigma^2$.
- All ϵ_i are normally distributed. Again, this assumption is conventionally included although not absolutely essential.

4.3.3 Interpretation of the parameters

For the simple linear regression model, the conditional mean μ of Y is related to X by the straight-line relationship $\mu = \alpha + \beta X$. For the multiple linear model, μ is given by equation (4.1), which is a higher-dimensional linear generalisation of a straight line. When there are two explanatory variables ($k = 2$), μ is described by a flat *plane* as X_1 and X_2 take different values (think of a sheet of paper, at an angle and extended indefinitely in all directions, cutting across a room in the air; this is illustrated by Figure 11.1 of Agresti and Finlay, p. 384). When k is larger than 2, the regression surface is a higher-dimensional linear object known as a hyperplane. This is impossible to visualise in our three-dimensional world, but the essence of the model remains mathematically and conceptually unchanged.

Values of Y which may be observed exist in a yet higher dimension, so they cannot in general be predicted exactly even with multiple explanatory variables. For a model with two explanatory variables, you might try to visualise the values of Y as a swarm of bees in the air in three-dimensional space, some perhaps sitting on the sheet of paper corresponding to the regression plane but most hovering above or below it. The model states that, at any set of values for the explanatory variables $X_1, X_2, \ldots, X_k$, the population distribution of Y is a normal distribution with mean μ given by (4.1) and standard deviation given by the **residual standard deviation** σ. As for the simple linear model, values of Y are thus distributed around the regression surface, and σ quantifies how widely they tend to vary around the surface.

The **constant term** (intercept) α is interpreted as the expected value of Y when all of the explanatory variables have the value 0. This can be seen by setting $X_{1i}, X_{2i}, \ldots, X_{ki}$ all to 0 in (4.1). As before, α is rarely of any substantive interest. In the example in Table 4.5, the estimated value of α is 59.4. This represents the predicted score of the General Health Index for a person who is 0 years old and has no education and no family income. Here this is not only not specifically interesting but also out of the range of the data (where the lowest age was 16), so basing any serious interpretation on this estimate would be nonsensical.

The parameters of main interest in a multiple linear model are the regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$ of the explanatory variables. To explain their interpretation, consider a model with three explanatory variables X_1, X_2 and X_3. This specifies the expected value of Y as

$$\mu = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{4.3}$$

for any values of X_1, X_2 and X_3. Suppose now that we consider two observations, where the second one has the same values of X_1 and X_2 as the first one, but the value of X_3 larger by one unit, i.e. $X_3 + 1$. The expected value of Y is given by (4.3) for the first observation, and by

$$\mu = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3(X_3 + 1) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_3. \quad (4.4)$$

for the second. Subtracting (4.3) from (4.4) leaves us with β_3. In other words, β_3 is the change in expected value of Y when X_3 is increased by one unit, while keeping the values of X_1 and X_2 unchanged. A similar result would obviously be obtained for X_1 and X_2, and for models with any number of explanatory variables. In general, the interpretation can thus be stated as follows:

- The regression coefficient β_j of any explanatory variable X_j in a multiple linear regression model shows the change in expected value of the response variable Y when X_j is increased by one unit (from any value), while holding all other explanatory variables constant (at any values).

The association between Y and X_j described in this way by the coefficient β_j is known as a **partial association** between X_j and Y, controlling for the other explanatory variables in the model. The *sign* of the coefficient indicates the *direction* of the partial association in the same way as the coefficient in a simple linear model does for the bivariate (marginal) association: if $\beta_j > 0$, the partial association between X_j and Y is positive, and if $\beta_j < 0$, the partial association is negative. If $\beta_j = 0$, there is no partial association; in other words, X_j is then not associated with Y after we control for the other explanatory variables.

In the example output in Figure 4.5 we have three explanatory variables for GHI, age X_1, education X_2 and income X_3, for which the estimated coefficients are $\hat{\beta}_1 = -0.128$, $\hat{\beta}_2 = 0.990$ and $\hat{\beta}_3 = 0.275$. These are interpreted as follows:

- Holding levels of education and family income constant, increasing age by one year *decreases* expected GHI by 0.128 points.
- Holding levels of age and family income constant, increasing education by one year of school completed increases expected GHI by 0.990 points.
- Holding levels of age and education constant, increasing family income by 1000 dollars increases expected GHI by 0.275 points.

As before, it may be more convenient to express the interpretations in terms of other increments than one unit, by multiplying the coefficient by the correponding value. For example, the estimated effect of a 10-year increase in age would here be an expected decrease of $10 \times 0.128 = 1.28$ GHI points, holding education and income constant.

The intepretation of regression coefficients is obviously connected to the idea of statistical control discussed in Section 4.1. In essence, this was to examine the association between a response variable and a particular explanatory variable, while holding all other explanatory variables at constant values. In Section 4.1 we examined this informally in examples where a single control variable had only three possible values,

so it was to possible carry out separate analyses where it was literally held constant. This is not practicable when some of the control variables are continuous, because they then have too many distinct values for it to be feasible to consider each one separately. Instead, statistical control is implemented with the help of a multiple regression model, and interpreted in terms of the regression coefficients in the way stated above. When discussing the coefficients, we will thus often talk of "controlling for" other variables instead of "holding constant" their values.

This interpretation of the coefficients is obtained by "increasing by one unit" and "holding constant" values of explanatory variables by mathematical manipulations alone. It is thus true within the model even when the values of the explanatory variables are not and cannot actually be controlled and set at different values by the researcher. This, however, also implies that this appealing interpretation is a mathematical construction which does not automatically correspond to reality. The interpretation of the regression coefficients is always mathematically true, but whether it is also an approximately correct description of an association in the real world depends on the appropriateness of the model for the data and study at hand. In some studies it is indeed possible to manipulate at least some explanatory variables, and corresponding regression models can then help to draw reasonably strong conclusions about associations between variables. Useful results can also be obtained in studies where no variables are in our control, as long as the model is selected carefully. This requires, in particular, that a linear model specification is adequate for the data, and that no important explanatory variables have been omitted from the model.

4.4 Estimation

4.4.1 Estimates of the model parameters

Estimates of the regression coefficients are here denoted with hats as $\hat{\alpha}$ and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ Fitted (predicted) values for Y_i are given by

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} \tag{4.5}$$

for every unit $i = 1, \ldots, n$ in the data. The estimated regression coefficients are again obtained with the method of least squares, by finding the values for $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ which make the error sum of squares

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

as small as possible. These estimates are known as the least squares (or Ordinary least squares, or OLS) estimates of the parameters. Finding them is both mathematically and intuitively the same exercise as least squares estimation for a simple linear model (c.f. Section 3.4.3). The only difference is that we are now operating in more dimensions, trying to find the best-fitting hyperplane through a high-dimensional cloud of points rather than the best-fitting straight line through a two-dimensional scatterplot.

With more than one explanatory variable, the computational formulas for the OLS estimates become difficult to write down, and the estimates themselves would be practically impossible to calculate by hand. This is not a problem in practice, as the

calculations are easily done by statistical software such as SPSS. In Table 4.3 on page 53, the least squares estimates of the regression coefficients are shown in the "Coefficient" column. Each row of the table gives the coefficient for one explanatory variable, identified in the first column. For example, the estimated coefficient of Age (X_1) is here $\hat{\beta}_1 = -0.128$. A similar format is adopted in the output from SPSS (see the "Coefficients" table of Figure 4.5) and other software packages (see Section 4.9.5).

Estimation of the model provides also estimates of the standard errors of each estimated coefficient $\hat{\beta}_j$ (where $j = 1, \ldots, k$). Here they will be denoted $\hat{\text{se}}(\hat{\beta}_j)$. For example, the standard error of the estimated coefficient of Age in our example is $\hat{\text{se}}(\hat{\beta}_1) = 0.032$ (c.f. the "Standard error" column of Table 4.3). The formulas of the standard errors are complicated and uninteresting in themselves. As with standard errors of most estimates in statistics, $\hat{\text{se}}(\hat{\beta}_j)$ depends on the sample size, with large samples giving smaller standard errors than small ones.

Besides being intuitively easy to explain and understand, least squares estimates of the regression coefficients also have several attractive theoretical properties. First, the OLS estimate $\hat{\beta}_j$ of any coefficient β_j is *unbiased*, meaning that the mean of its sampling distribution is equal to β_j (in short, $\text{E}(\hat{\beta}_j) = \beta_j$). In other words, the estimates are, on average, estimating the correct parameters, and have no systematic bias. Second, among all the unbiased estimates of β_j which are linear functions of the observations $Y_1, \ldots, Y_n$, the OLS estimate has the smallest standard error, for which it is known as the Best Linear Unbiased Estimate or BLUE (the fact that $\hat{\beta}_j$ is a linear function of the observations is neither obvious nor particularly important here, but it is). This means that for any given set of data, the least squares method gives estimates which are more precise than any other estimates, and thus make the most efficient use of the available data. The BLUE property holds even without the assumption of normality for the error terms ϵ_i of the model (c.f. Section 4.3.2). If this *is* assumed, the condition on linearity in BLUE can be removed. The OLS estimates are then also *Maximum likelihood estimates* of the coefficients, which implies that they possess further useful properties. The upshot of these results (which you do not need to know to use the models) is that least squares is the standard method of estimation for linear models, and other kinds of estimates are very rarely considered.

The most common estimate of the remaining parameter of the model, the residual standard deviation σ, is the same as before, i.e.

$$\hat{\sigma} = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n - (k+1)}} = \sqrt{\frac{SSE}{n - (k+1)}} \tag{4.6}$$

where k is the number of explanatory variables and $df = n - (k+1)$ is the degrees of freedom associated with the error sum of squares SSE. In the model shown in Table 4.3, we have $n = 1699$, $k = 3$, $df = 1699 - (1+3) = 1695$, and $\hat{\sigma} = 14.6$.

In mathematical statistics, formulas for multiple linear regression models are usually expressed in the notation of matrix algebra. This makes it possible to manipulate with ease expressions which would in standard notation be impenetrable thickets of products and summations. For example, the OLS estimate of the regression coefficients (now including $\hat{\alpha}$) is in matrix notation

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are matrices of explanatory and response variables respectively, and the estimated residual standard error is written as

$$\hat{\sigma} = \sqrt{\frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - (k+1)}}.$$

It is not necessary to know these formulas, or even to be aware that they exist, to use multiple linear models. They are mentioned here only so that you have an idea of what they are for, if you ever come across them elsewhere.

4.4.2 Fitted values

The most effective way to describe and illustrate the implications of a fitted regression model is typically to do so in two complementary ways. One of these is the interpretation of the regression coefficients, as explained in Section 4.3.3. The second is the presentation of a selection of *fitted values* for the response variable Y, calculated from the estimated model. Such fitted (or *predicted*) values are also the key output of the model when it is used mainly to predict future values of Y at some values of the Xs. For both purposes, the fitted values are calculated from the formula

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k \tag{4.7}$$

by substituting the estimated coefficients and required values of $X_1, X_2, \ldots, X_k$ on the right-hand side. Note that (4.7) is the same as (4.5) in the previous section, except that the subscript i is omitted. This emphasises the fact that fitted values for prediction or illustration can be calculated at *any* values of the explanatory variables, not just those that actually appear in the original data.

Figure 4.6 shows some fitted values for GHI obtained from the model shown in Table 4.3. Each of the lines shows the fitted values given family incomes between 0 and 32 (thousands of dollars), roughly the range of values in the observed data (recall that these are 1973 dollars). The upward slope of the lines obviously shows that increasing income is associated with better levels of predicted general health. Whenever we draw lines like this for one explanatory variable in a multiple linear model, it is necessary to fix all the other explanatory variables at some specific values. Here education is fixed at 12 years, roughly its mean in the sample. Age is fixed at two values, 30 and 60 years, resulting in the two lines in the plot. The difference between these (of about 3.8 GHI points) is the predicted decrease in GHI corresponding to a 30-year increase in age, other things being equal.

A tabular presentation of fitted values is often more economical than a graphical one. It can, for example, be used to display the fitted values at all combinations of selected values of the explanatory variables (or, if there are too many variables, combinations of some of them, holding the rest constant at some values). An illustration of this for the GHI model is shown in Table 4.4. It contains fitted values of GHI at combinations of age of 16, 35, and 62 years, income of 0, 10 and 32 thousands of dollars, and education of 0, 12 and 18 years; the extreme values are in each case approximately the minimum and maximum of the variable in the HIE sample, and the middle value is approximately the sample mean. The table shows how different combinations of the explanatory

Figure 4.6: Fitted values of General Health Index as function of family income, at two values of age, based on the fitted model in Table 4.3. Education is fixed at 12 years.

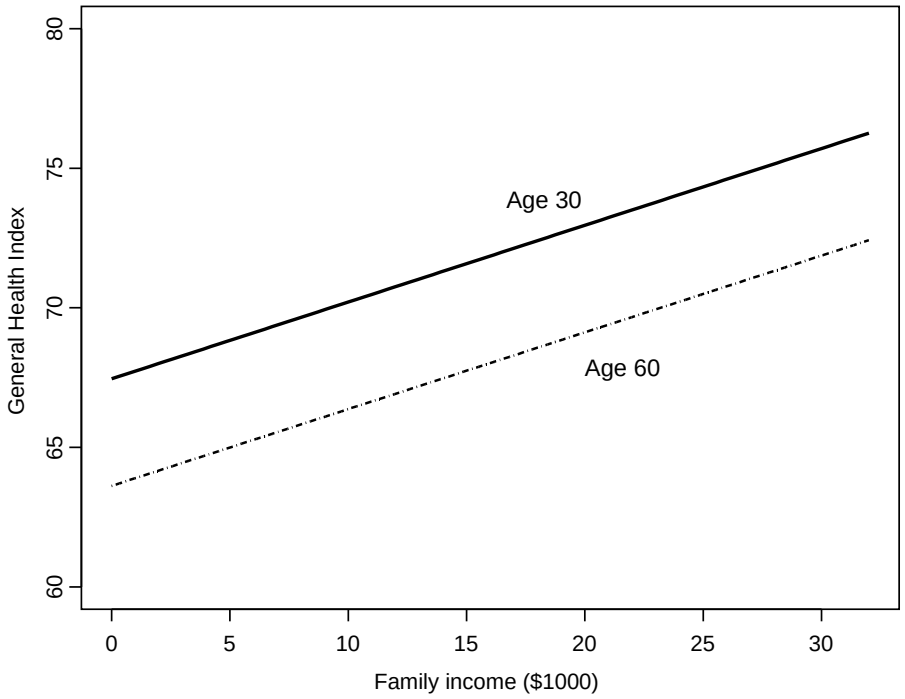


Table 4.4: Fitted values of General Health Index at selected values of age, family income and education, based on the fitted model in Table 4.3.

	Education								
	0			12			18		
Income	0	10	32	0	10	32	0	10	32
Age 16	57.4	60.1	66.2	69.2	72.0	78.0	—	—	—
35	54.9	57.7	63.7	66.8	69.6	75.6	72.8	75.5	81.6
62	51.5	54.2	60.3	63.4	66.1	72.2	69.3	72.1	78.1

variables result in larger or smaller fitted values, and illustrates the magnitudes of these differences. For example, the strongest gradient in predicted general health is observed when moving from the least favourable combination (old, poor and uneducated) in the bottom-left corner of the table towards the top-right corner of young, rich and well-educated. Even within, say, age 35, the difference between the extremes shown here is 26.7 points, or over five times the five-point difference which Newhouse et al. describe as roughly equivalent to being diagnosed with high blood pressure.

Calculation of fitted values, and interpretation of the model in general, should be limited to the range of values of the explanatory variables about which the data provide information, and *extrapolation* beyond this should be avoided. This applies, first, to the values of individual variables. Here, for example, we could calculate predicted GHI scores for someone who is 95 years old or has a family income of a million dollars. We could not, however, place any confidence on such predictions, because these values are

well beyond the ranges of ages and incomes in the actual data.

More complex kind of extrapolation occurs when we consider values of the explanatory variables which are individually plausible but never observed in combination. For example, there are many respondents in the HIE data who are aged 16, and many who have completed 18 years of education, but obviously none with both these characteristics. Here the combination is logically impossible and easy to spot (and, as such, omitted from Table 4.4), but other rare combinations may be less obvious. This should be investigated with some care before placing strong interpretation on fitted values which might correspond to such extremes.

4.4.3 R^2 and sums of squares

The discussion of sums of squares and the coefficient of determination R^2 given in Section 3.4.5 for simple linear models is essentially unchanged for multiple linear models. The variation of the observed values Y_i of the response variable, summarised by the total sum of squares $\sum(Y_i - \bar{Y})^2$, can again be decomposed as

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 \tag{4.8}$$

where the fitted values $\hat{Y}_i$ are now calculated from formula (4.5). In words and symbols, this can again be summarised as

$$\begin{array}{ccccc} \text{Total variation of } Y & = & \text{Variation explained by regression} & + & \text{Unexplained variation} \\ TSS & = & SSM & + & SSE. \end{array}$$

The R^2 statistic is still defined as

$$R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}. \tag{4.9}$$

It shows the proportion of the total variation of Y in the sample explained by the fitted regression model, i.e. by the variation of the values of the explanatory variables. An alternative expression for this is that R^2 is the proportional reduction in prediction errors when we use fitted values from the model to predict the observed values Y_i, compared to ignoring the explanatory variables and using the overall mean $\bar{Y}$ as the predicted value for every observation.

The square root of R^2 is the multiple correlation coefficient R, which is the correlation between Y_i and the fitted values $\hat{Y}_i$. Unlike for a simple linear model, R does not now correspond to the correlation between Y_i and any individual explanatory variable. The possible values of both R^2 and R range from 0 (none of explanatory variables are associated with Y, so the model is of no help in predicting Y) to 1 (values of Y can be predicted perfectly using the explanatory variables). In our example, $R^2 = 0.061$ and $R = 0.247$. Thus 6.1% of the observed variation in the values of General Health Index is explained by variation in the respondents' age, education and family income.

R^2 is a useful statistic with a convenient interpretation. However, its importance should not be exaggerated. R^2 is rarely the only or the most important part of the model results. This may be the case if the regression model is fitted solely for the purpose of *predicting* future observations of the response variable. More often, however, we are at least or more interested in examining the nature and strength of the associations between the response and explanatory variables, in which case the regression coefficients are the main parameters of interest. This point is worth emphasising because in our experience many users of linear regression models tend to place far too much importance on R^2, often hoping to treat it as the ultimate measure of the goodness of the model. We are frequently asked questions along the lines of "My model has R^2 of 0.42 — is that good?". The answer tends to be "I have no idea" or, at best, "It depends". This not a sign of ignorance, because it really does depend:

- Which values of R^2 are large or small or "good" is not a statistical question but a substantive one, to which the answer depends on the nature of the variables under consideration. For example, most associations between variables in the social sciences involve much unexplained variation, so their R^2 values tend to be smaller than for quantities in, say, physics. Similarly, even in social sciences models for aggregates such as countries often have higher values of R^2 than ones for characteristics of individual people. For example, in Chapter 3 (c.f. output in Figure 3.5 on page 33) we obtained $R^2 = 0.567$ for a simple linear model for infant mortality rate given school enrolment for a sample of countries, while here $R^2 = 0.061$ for the model for GHI values of individuals.
- In any case, achieving large R^2 is usually not the ultimate criterion for selecting a model, and a model can be very useful without having a large R^2. The R^2 statistic reflects the magnitude of the variation around the fitted regression line, corresponding to the residual standard deviation $\hat{\sigma}$. Because this is an accepted part of the model, R^2 is not a measure of how well the model fits: we can have a model which is essentially true (in that the Xs are linearly associated with Y) but has large residual standard error and thus small R^2. This may be the case for our model for GHI: while an R^2 of 6.1% shows that a person's self-assessed general health cannot be very precisely predicted from his or her age, education and income alone, it is still useful to know that all of these appear to be associated with it.

4.4.4 Other quantities in computer output

Two quantities in standard SPSS output for linear regression (c.f. Figure 4.5 and the explanation on p. 51) are not discussed above or in the inference section below. They are of lesser interest in most analyses, so they will only be explained briefly here.

The "Adjusted R^2" statistic is defined as

$$R^2_{adj} = \frac{(n-1)R^2 - k}{n - (k+1)}. \tag{4.10}$$

This is most relevant in situations where the main purpose of the model is prediction of future observations of Y. The population value of the R^2 statistic is then a key criterion

of model selection. R^2_{adj} is a better estimate of it than standard R^2. Unlike R^2, R^2_{adj} does not always increase when new explanatory variables are added to the model. As a sample statistic, R^2_{adj} does not have the same interpretation as the proportion of variation of Y explained as standard R^2.

The "standardized" regression coefficient of an explanatory variable X_j is defined as

$$\hat{\beta}_j^{\text{std}} = \left(\frac{s_{xj}}{s_y}\right) \hat{\beta}_j, \tag{4.11}$$

where $\hat{\beta}_j$ is the usual (OLS) estimated coefficient of X_j, and s_{xj} and s_y are sample standard deviations of X_j and Y respectively. This is equal to the correlation of Y and X_j when X_j is the only explanatory variable in the model, but not otherwise.

The standardized coefficient describes the expected change in Y in units of its sample standard error s_y, when X_j is increased by one of its standard errors s_{xj}, holding other explanatory variables constant. Such a quantity may be used for two largely related purposes. The first is rescaling, which here reexpresses the model results as the effect of X_j on Y when both are measured in standard deviation units. This may be useful when the original scales of the variables are unfamiliar to us, as may be the case, say, with a new attitude scale. When, however, the variables are measured in familiar units, such as years or dollars in our example, such rescaling is unhelpful.

The second purpose of standardisation is to facilitate direct comparisons between the effects of different explanatory variables. This ultimately refers to the question of their *relative importance*, i.e. "which of the explanatory variables in the model is the most important?" Standardisation aims to address this by comparing effects of similar changes in the explanatory variables, where "similar" is defined as "change of one standard deviation". For example, the standardised coefficient of Age in our example (c.f. Figure 4.5 on p. 52) is -0.101. This means that the estimated effect of increasing age by one sample standard deviation (which is about 12 years, see Table 4.1 on p. 49) is to decrease expected GHI by 0.101 standard deviation units (each of which is about 15 GHI points), controlling for education and income. The standardised coefficients of education and income are 0.172 and 0.109 respectively. This suggests that education is the "most important" of these variables, in the narrow sense that an increase of one standard deviation (around 2.7 years) in it has the largest estimated effect on GHI.

Despite their appealing title, however, standardized coefficients do not really provide a simple and widely accepted tool for judging relative importance. First, the standardisation with respect to Y is unnnecessary for comparison purposes, because Y is the same for all of the explanatory variables[6]. This suggests that we might consider *semi-standardised* coefficients $s_{xj}\hat{\beta}_j$, which indicate the estimated effect on the expected Y (in its natural units) of a one standard deviation increase in X_j. This, however, still leaves the question of whether a standard deviation is really the increase which is most meaningfully comparable as being of "similar magnitude" across variables, or whether some other increments might be considered instead. For example, the difference between the largest and smallest values of each variable might be another such reference increment. Using this (e.g. 46 years for age) to multiply the coefficients in our example gives new "standardised" coefficients of -5.9, 24.8 and 8.9 for age, education and

[6] Unless we contemplate comparisons across models for different responses or for different data sets, but that is an exercise rarely to be attempted.

income respectively. While these have many problems of their own, they show that the usual standardised coefficients are not the only possibility.

A more fundamental problem with assessment of relative importance using *any* version of the parameter estimates is that it often involves conceptually troublesome comparisons of variables of very different natures and practical implications. For instance, suppose that we have fitted a model for the General Health Index of a person, given the amount of physical exercise the person takes (which can be changed by him/herself), investment in preventive healthcare in the area where the person lives (which can be changed, but with more effort and not by the individual) and the person's age (which cannot be manipulated at all). The values of the unstandardized or standardized coefficients of these explanatory variables can certainly be compared, but it is not clear what statements about the relative sizes of the effects of "increasing" them would really mean. All of this makes the question of relative importance of explanatory variables, while understandably one of great interest in many cases, also one of the hardest questions in modelling, and one without a purely statistical solution.

4.5 Inference for the regression coefficients

Virtually all interesting questions about the population values of the parameters of a multiple linear model concern the regression coefficients $\beta_1, \ldots, \beta_k$ rather than α or σ. This is because the coefficients are the parameters corresponding to (partial) associations between the explanatory variables and the response. We will thus discuss only inference for the regression coefficients. First, t-tests and confidence intervals for single coefficients are described in Sections 4.5.1 and 4.5.2 respectively. These procedures are essentially identical to the ones for simple linear models discussed in Section 3.4.4. A new tool is the F-test for hypotheses involving several coefficients at once, which is introduced in Section 4.5.3.

4.5.1 Tests of single coefficients

Significance testing of individual regression coefficients in a multiple linear model differs from that for the simple model in interpretation but not in execution. Let β_j denote the population coefficient of an explanatory variable X_j (where j may be any of $1, 2, \ldots, k$). The null hypothesis considered here is

$$H_0 : \ \beta_j = 0, \quad \text{other regression coefficients are unrestricted} \tag{4.12}$$

against the two-sided alternative hypothesis

$$H_a : \ \beta_j \neq 0, \quad \text{other regression coefficients are unrestricted.}$$

(one-sided alternative hypotheses are again possible but rarely considered; if needed, a one-sided P-value is obtained as explained in Section 2.4.4). These are subtly but crucially different from the hypotheses in the case of a simple linear model. Here the statement about "unrestricted" other parameters implies that neither hypothesis makes any claims about the values of other coefficients than β_j, and these are allowed

to have any values. The null hypothesis is a claim about the association between X_j and Y when the other explanatory variables are already included in the model. In other words, (4.12) can also be stated as

H_0 : There is no partial association between X_j and Y, controlling for the other explanatory variables.

A t-test statistic for the null hypothesis (4.12) is given by

$$t = \frac{\hat{\beta}_j}{\hat{\text{se}}(\hat{\beta}_j)} \tag{4.13}$$

where $\hat{\beta}_j$ is the least squares estimate of β_j, and $\hat{\text{se}}(\hat{\beta}_j)$ is its estimated standard error. This is identical in form to statistic (3.11) for the simple regression model (see p. 36), and also yet another instance of the general test statistic (2.10) on p. 16, with β_j now in the role of Δ.

The sampling distribution of (4.13) when the null hypothesis (4.12) holds is a t distribution with $n - (k + 1)$ degrees of freedom, where k is the number of explanatory variables in the model. The (two-sided) P-value, evaluated from this distribution, is the probability of values which are further from 0 than the observed value of t. Small P-values indicate, as always, evidence against the claim that β_j is 0 in the population. The test statistic and P-value are calculated by statistical software and typically included in output for linear regression (see Table 4.3 on page 53).

It is important to note two things about the results of such tests for multiple regression models. First, (4.12) implies that if the null hypothesis is not rejected, X_j is not associated with Y, *if* the other explanatory variables are already included in the model. We would typically react to this by removing X_j from the model altogether, while keeping the other variables in it. Second, the k tests and P-values in the output actually refer to k *different* hypotheses of the form (4.12), one for each explanatory variable. This raises the question of what to do if, say, tests for two variables have large P-values, suggesting that either of them could be removed from the model. The appropriate reaction is to remove one of the variables (perhaps the one with the larger P-value) rather than both at once, and then see whether the other still remains nonsignificant (if so, it can then also be removed). This is part of the general area of **model selection**, which will be discussed further in Section 4.8.

To illustrate the use of the t-test, consider the models for GHI shown in Table 4.14, starting with model (5). This has four explanatory variables. Consider first the variable on work experience, for which the estimated coefficient is (to five decimal places) $\hat{\beta} = 0.00240$ and its estimated standard error (not shown in the table) $\hat{\text{se}}(\hat{\beta}) = 0.00415$, so $t = 0.00240/0.00415 = 0.58$. The P-value for this is obtained from the t distribution with $n - (k + 1) = 1699 - 5 = 1694$ degrees of freedom, and is $P = 0.563$. This means that we clearly do not reject the null hypothesis that there is no association between experience and GHI, once we control for age, education and income. The P-values for the other three explanatory variables are very small, indicating that they *are* associated with GHI even after controlling for the others. These results suggest that we can drop experience from the model, retaining the other three variables. This gives model (4) in Table 4.14, for which all the details are shown in Table 4.3. All of the three explanatory

variables in this model are clearly significant (with P-values less than 0.001), so we would keep all of them and settle for model (4).

We might also have proceeded in a different direction, starting with a simple model and adding explanatory variables to it as needed. For example, model (1) contains only age, which is clearly significant. Adding either education (model 2) or income (model 3) produces a model where two explanatory variables are both significant, and adding the remaining one leads us back to model (4) where all three explanatory variables are significant. If we then added also work experience, we would obtain model (5), where we would again discover that experience is not significant once the other three are included.

As the last observation on this example, consider model (6) which includes education, income and experience, but not age. The coefficient of experience is now significant at the 5% significance level (with $P = 0.045$), so experience does appear to have an effect on GHI, controlling for income and experience. What is happening here is that experience is picking up some of the effect of the missing age variable, with which it is (for obvious logical reasons) quite strongly correlated (with correlation 0.618). It is only when age itself is included in the model (giving us model 5) that this effect is attributed back to age, and experience is left nonsignificant. Results of the tests for a sequence of models thus depend on the *order* in which explanatory variables are added to or removed from the models. Similar behaviour was observed for model sums of squares and R^2 in Section 4.9.3, and the two findings are obviously closely related. This order dependence is one reason why model selection through significance tests like this is typically not a a straightforward exercise, and requires careful consideration.

4.5.2 Confidence intervals for single coefficients

A confidence interval with confidence level $1 - \alpha$ for any β_j is given by

$$\hat{\beta}_j \pm t_{\alpha/2}^{(n-(k+1))} \, \hat{\text{se}}(\hat{\beta}_j) \tag{4.14}$$

where $t_{\alpha/2}^{(n-(k+1))}$ denotes the two-sided critical value at significance level α for the t distribution with $n - (k + 1)$ degrees of freedom. This is identical in form and interpretation to the interval (3.12) for simple regression (except for a change in the degrees of freedom), so no new issues arise. The interval is routinely produced for linear regression procedures of statistical software. If it is for some reason calculated by hand, it is again acceptable to substitute multipliers for the standard normal distribution (e.g. 1.96 for 95% intervals) instead of $t_{\alpha/2}^{(n-(k+1))}$.

Table 4.3 shows 95% confidence intervals for the coefficients in that model (where $df = n - 4 = 1695$ and $t_{0.025}^{(1695)} = 1.96$). For example, we are 95% confident that the population effect of a one-year increase in level of education is an expected increase of between 0.709 and 1.272 GHI points, controlling for age and income. Similarly, the 95% confidence interval for the partial effect of, say, a five-year increase in education is $5 \times (0.709; 1.272) = (3.545; 6.360)$.

4.5.3 F-tests of several coefficients

The t-test statistic (4.13) is used to test the null hypothesis (4.12) that the population coefficient of a single explanatory variable is 0. Sometimes we may instead want to consider a hypothesis that the coefficients of several variables are all 0, i.e.

$$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0, \tag{4.15}$$

against the alternative hypothesis

$$H_a : \text{ at least one of } \beta_{g+1}, \beta_{g+2}, \ldots, \beta_k \text{ is not } 0$$

where the coefficients $\beta_1, \ldots, \beta_g$ are unrestricted under both hypotheses. Note also that the numbering of the variables is arbitrary here, so the null hypothesis (4.15) does not need to refer to the "last" explanatory variables in any sense but can be about any subset of the k explanatory variables.

Another view of this is that the hypotheses specify two linear models for μ:

$$\textbf{Restricted Model } M_0 : \ \mu = \alpha + \beta_1 X_1 + \cdots + \beta_g X_g$$

under H_0, and

$$\textbf{Full model } M_a : \ \mu = \alpha + \beta_1 X_1 + \cdots + \beta_g X_g + \beta_{g+1} X_{g+1} + \cdots + \beta_k X_k$$

under H_a. In short, the null hypothesis specifies a model M_0 which includes only the explanatory variables $X_1, \ldots, X_g$, and the alternative hypothesis a model M_a which includes both $X_1, \ldots, X_g$ and $X_{g+1}, \ldots, X_k$. The model M_a thus contains all of the explanatory variables in M_0, as well as some addtional ones. The smaller restricted model M_0 is then said to be **nested** within the larger full model M_a. For example, below we consider an illustration where $k = 3$ and $g = 1$, so M_0 includes X_1 while M_a includes X_1, X_2 and X_3.

The general idea of testing a null hypothesis of type (4.15) is to fit both the full and the restricted model, and to compare the results. Let SSE_a and R_a^2 denote the error sum of squares and R^2 for the full model, and define SSE_0 and R_0^2 for the restricted model similarly. For symmetry of notation, denote the numbers of explanatory variables by $k_a = k$ and $k_0 = g$ for M_a and M_0 respectively, and let $df_a = n - (k_a + 1)$ and $df_0 = n - (k_0 + 1)$ denote their degrees of freedom. The test statistic for the null hypothesis (4.15) is the general **F test statistic**

$$\begin{aligned} F &= \frac{(SSE_0 - SSE_a)/(k_a - k_0)}{SSE_a/[n - (k_a + 1)]} = \frac{(SSE_0 - SSE_a)/(df_0 - df_a)}{SSE_a/df_a} \\ &= \frac{(R_a^2 - R_0^2)/(k_a - k_0)}{(1 - R_a^2)/[n - (k_a + 1)]} \\ &= \frac{R^2_{\text{change}}/df_{\text{change}}}{(1 - R_a^2)/[n - (k_a + 1)]} \end{aligned} \tag{4.16}$$

where $df_{\text{change}} = df_0 - df_a = [n - (k_0 + 1)] - [n - (k_a + 1)] = k_a - k_0$. Note that $k_a - k_0$ is the number of explanatory variables which are included in the full model but not in the restricted model, i.e. the number of regression coefficients which are set to 0 by the null hypothesis (4.15).

The test statistic is here expressed in a number of different but equivalent forms, because different ones may be most convenient in different contexts (and for different readers). Here we can concentrate on the last version, expression (4.16). The core of this is $R^2_{\text{change}} = R^2_a - R^2_0$, i.e. the *increase in* R^2 when we switch from the restricted model M_0 to the full model M_a (the other elements of the formula (4.16) are there for scaling purposes, to give the F statistic a mathematically convenient sampling distribution). In other words, R^2_{change} is the amount by which R^2 increases when we add the explanatory variables $X_{g+1}, \ldots, X_k$ to the restricted model.

A large value of R^2_{change} indicates that the additional explanatory variables included in the full model increase the predictive power of the model by a large amount compared to the restricted model. Intuitively it seems obvious that this would indicate evidence against the null hypothesis (4.15), i.e. evidence that at least one of the additional explanatory variables has an effect on Y, even after controlling for the variables already in the restricted model. So it is large values of R^2_{change} and thus also of F which are evidence against H_0, while small values of F indicate that H_0 should not be rejected.

The sampling distribution of the F statistic (4.16) under the null hypothesis (4.15) is the F distribution with $k_a - k_0$ and $n-(k_a+1)$ degrees of freedom (the F distributions are a family of probability distributions specified by two degrees of freedom values). The P-value of the F-test is the probability, calculated for this distribution, of values at least as large as the observed value of F; this calculation and the curve of one F distribution are illustrated by Figure 11.9 on p. 400 of Agresti and Finlay. The calculation of the test statistic and P-value are normally done using statistical software such as SPSS. They could also be calculated by hand, obtaining approximate P-values using tables of critical values for F distributions (see Table D in the Appendix of Agresti and Finlay); this, however, is not practiced here nor required in the examination.

F-tests for two extreme versions of the null hypothesis (4.15) are worth discussing separately. The first is one where only one coefficient is set to zero, i.e. a case where (4.15) reduces to hypothesis (4.12) on page 63. This can then be tested using either the t-statistic (4.13) or the F-statistic (4.16). Both will in fact give the same result, as then $F = t^2$ and the P-values (from $t_{n-(k+1)}$ and $F_{1,n-(k+1)}$ distributions) will be the same. The t-test is usually used to test this hypothesis, as its results for every coefficient are routinely available in computer output.

If (4.15) is stated simultaneously for *all* of the coefficients $\beta_1 = \cdots = \beta_k$, we get the null hypothesis that *none* of the explanatory variables $X_1, \ldots, X_k$ are associated with the response. The restricted model M_0 is then the constant-only model where $\mu = \alpha$. Since for this model $k_0 = 0$, $R^2_0 = 0$ and $SSE_0 = TSS$, the F statistic (4.16) becomes

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{SSM/k}{SSE/[n-(k+1)]} \tag{4.17}$$

where all the quantities are for the model with all the explanatory variables included. This test statistic and its P-value (from the F distribution with k and $n-(k+1)$ degrees of freedom) are included in standard computer output. A large P-value would indicate that none of the explanatory variables had an effect on the response, in which case there is little point in proceeding with the model. In practice, however, this is not often the case, as it is more common that we have enough understanding of the variables to select for consideration at least some which are associated with the

response. Thus it is more common to obtain a very small P-value for the overall F-test, indicating that at least one of the explanatory variables is significant. The test is then of little further use, because it gives us no information on *which* of the explanatory variables should be retained.

As an illustration, consider a comparison between two models for GHI in Table 4.14. Suppose the restricted model M_0 is model (1) which has age (X_1) as its only explanatory variable, and that the full model M_a is model (4) which includes education (X_2) and income (X_3) in addition to age. The null hypothesis (4.15) is thus that

$$H_0 : \beta_2 = \beta_3 = 0,$$

i.e. that neither education nor income is associated with GHI, controlling for age.

In the notation of (4.16), here $R_0^2 = 0.0119$, $R_a^2 = 0.0611$, $k_0 = 1$, $k_a = 3$ and $n = 1699$, so

$$F = \frac{(0.0611 - 0.0119)/(3 - 1)}{(1 - 0.0611)/[1699 - (3 + 1)]} = 44.4,$$

for which the P-value from the $F_{2,1695}$ distribution is $P < 0.001$[7]. Thus there is very strong evidence that education or income or both have an effect on GHI even after controlling for age. This comes as no surprise, since we know from the t-tests above that both of them individually are strongly statistically significant.

This illustration of the F-test is rather artificial. In practice we would not normally try to test for the effects of two separate explanatory variables at once, but would use the t-test for them one at a time. There are, however, also situations where testing the effects of several variables together is natural and useful. The most common of these situations will be discussed in Section 4.6.1.

The terms "restricted model" and "full model" used here are always relative to a specific comparison. Often we may consider a set of F-tests for a sequence of nested models. It is then common that a model which was the full (larger) model in one comparison takes the role of the restricted (smaller) one in the next. For example, if we here decided next to carry out an F-test comparing models (4) and (5), model (4) would become the restricted model and model (5) the new full model. Similarly, the restricted model from one comparison would be the full model in another where it was compared to an even smaller model which omits some of its explanatory variables.

Finally, note that the quantities in the F-test statistic (4.16) should be based on the restricted and full models fitted to the same set of observations, i.e. n should be the same for both. This is an issue when values of some of the explanatory variables are missing for some observations. Statistical software then usually use so-called complete-case analysis (listwise deletion) where all such incomplete observations are omitted. If an observation has missing values only for explanatory variables included only in the full model, it would normally be included if we were fitting just the restricted model. For the F-test, however, the requirement of the common set of observations implies that any such observations should be omitted when fitting both of the models. This is done automatically when the test is carried out by SPSS or other software.

[7] If the F-statistic is calculated by hand, it is actually better to use the formulas in (4.16) expressed in terms of the error sums of squares, as these are typically reported in the output with more significant digits than R^2. Here SPSS calculates $F = 44.464$.

4.6 Specification of explanatory variables

Various types of explanatory variables may be included in multiple linear regression models. Some of these are described here, again illustrated using data from the Health Insurance Experiment. Models for two different response variables will be considered:

- General Health Index (GHI), defined as explained on page 50, measured at the *end* of a person's involvement in the HIE.
- Diastolic blood pressure (in mmHg), also measured at exit from the study.

The following explanatory variables, all measured at the time of enrolment to the study, will be used somewhere in this section:

- Value of GHI or diastolic blood pressure at the time of enrolment to the HIE. This is typically a very strong predictor of the same quantity at exit, and will be included in all the models. For other explanatory variables, we will thus examine whether they are significant predictors of GHI or blood pressure at exit even after the initial value of the same variable is controlled for.
- Sex of the respondent.
- Age in years.
- Family income, in thousands of 1973 dollars. This is a slightly different measure from the one used in Section 4.3, which was simply the family income in the year preceding enrolment. The variable used for the rest of this chapter (and also in Newhouse et al. 1993) is an average of reported incomes in the two years preceding enrolment, adjusted for size of the family and cost of living.
- Cigarette smoking, recorded as Current smoker, Ex-smoker, or Never smoked.
- Weight in kilograms.
- Respondent's own evaluation of his or her health, recorded as a response to a single question with four categories (Excellent, Good, Fair, Poor).

All the models for a given response variable are fitted to the set of observations for which the response and all the explanatory variables are observed. This sample size is $n = 1716$ for GHI and $n = 1045$ for blood pressure (mainly due to many missing values of initial blood pressure). The models shown here are purely illustrations of different types of explanatory variables in the context of these real data. In particular, no effort has been made to select the other explanatory variables in each model with any care.

In short, the following cases are discussed below:

- Categorical explanatory variables, included as dummy variables (Section 4.6.1).
- Interactions of two explanatory variables, included as products of those variables (Section 4.6.2).

- Nonlinear transformations (e.g. X^2 or $\log X$) of continuous explanatory variables (Section 4.6.3).
- Treatment of ordinal-level categorical explanatory variables (Section 4.6.4).

4.6.1 Categorical explanatory variables: Dummy variables

Consider first the variables on sex and smoking status defined above, and suppose that their categories are recorded as 1=Man, 2=Women; and 1=Never smoked, 2=Ex-smoker, 3=Current smoker. The social sciences are full of categorical variables like this (on marital status, race, employment status, party affiliation etc. etc.), which might obviously be of interest as explanatory variables in many analyses. Yet we clearly cannot simply use the numerical codes of the categories in the analysis as if they were values of a continuous variable. Doing so might complicate interpretation if the variable was dichotomous, and would be at best a partial solution if it was ordinal (c.f. Section 4.6.4), and entirely meaningless if it was nominal.

Instead, categorical explanatory variables are included with the help of dummy variables. A **dummy variable** (or **indicator variable**) has only two values, 0 and 1. Its value is 1 if a unit is in a particular category of a categorical variable, and 0 if it is not in that category. For example, we can define for each respondent the variable

$$S_m = \begin{cases} 1 & \text{if the person is male} \\ 0 & \text{otherwise.} \end{cases}$$

This could be referred to as "dummy for men". Note that the label S_m (for "sex male") used here has no special significance; dummy variables will be treated as regular explanatory variables, and we could denote them as Xs just like all the others.

A dummy variable S_w for women would be defined similarly, with values $S_w = 1$ for women and $S_w = 0$ for men. For cigarette smoking, we could define

$$C_c = \begin{cases} 1 & \text{if the person is a current smoker} \\ 0 & \text{otherwise,} \end{cases}$$

and C_e for ex-smokers and C_n for never-smokers similarly.

The dummies for the categories of a variable are not all needed to identify the category for every respondent. For sex, one of them is enough: using S_m alone, we would know that $S_m = 1$ indicates a man and $S_m = 0$ a woman. For smoking status, two of the three are sufficient, for example C_c and C_e: $(C_c = 1, C_e = 0)$ identifies a current smoker, $(C_c = 0, C_e = 1)$ an ex-smoker and $(C_c = 0, C_e = 0)$ a non-smoker; note that $(C_c = 1, C_e = 1)$ is not possible as no-one can be in more than one category at once.

In general, if a categorical variable has K categories, only $K - 1$ dummy variables are needed to identify the category of every unit. Dichotomous variables with only two categories ($K = 2$) are thus fully identified by just one dummy variable. The category which is not given a dummy of its own is known as the **reference category** or **baseline category**. Any category can be the baseline, and the results of the model will be the same for any choice of it.

Table 4.5: Estimated coefficients for linear regression models for Diastolic blood pressure at exit from HIE, given blood pressure at enrolment, sex and smoking status. Estimates listed as "0" denote reference categories of categorical explanatory variables.

	Model		
Variable	(1)	(2)	(3)
Past blood pressure	0.573	0.573	0.573
Sex			
Female	0	-2.033	0
Male	2.033	0	2.033
Smoking status			
Never smoked	0	1.239	1.382
Ex-smoker	−1.239	0	0.143
Current smoker	−1.382	−0.143	0
(Constant)	35.383	36.177	34.001

Categorical variables are used as explanatory variables in regression models by including the dummy variables for them in the model. This requires no changes in the definition or estimation of the model, and the parameter estimates, standard errors and quantities for statistical inference are obtained exactly as before. The only aspect which requires some further explanation is the interpretation of the coefficients of the dummy variables. This is illustrated in Table 4.5 with estimated coefficients for a model for blood pressure at exit given blood pressure an enrolment, sex and smoking status. Columns (1)–(3) of the table refer to the same model, with different choices of reference levels for sex and smoking status. We will first consider the estimates in column (1), where the reference levels are Woman and Never-smoker.

Recall that the regression coefficient of a continuous explanatory variable X is the expected change in the response variable when X is increased by one unit, holding all other explanatory variables constant. Exactly the same interpretation works for dummy variables, except that it is limited to the only one-unit increase possible for them, i.e. from 0 to 1. For example, column (1) of Table 4.5 shows that the estimated coefficient of the dummy variable S_m for men is 2.033. This is the expected change in the blood pressure when smoking status and blood pressure at enrolment are held constant but S_m increases from 0 (woman) to 1 (man). In other words, the expected blood pressure is 2.033 units higher for men than for women, controlling for the other two explanatory variables.

For smoking status, consider a comparison between a never-smoker and an ex-smoker, both with the same sex and initial blood pressure. The value of the ex-smoker dummy is $C_e = 0$ for the former and $C_e = 1$ for the latter. The value of the current smoker dummy must be $C_c = 0$ for both, so it too is being held constant in this comparison. The estimated effect of moving from $C_e = 0$ to $C_e = 1$ is the coefficient of C_e in model (1), which is -1.239. In other words, the expected blood pressure is 1.239 units lower for ex-smokers than for never-smokers, controlling for the other two explanatory variables. A similar argument for the coefficient of C_c shows that the expected blood

pressure is 1.382 units lower for current smokers than for never-smokers, other things being equal.[8] From these, it follows logically that it is also $1.382 - 1.239 = 0.143$ units lower for current smokers than for ex-smokers. The difference of the two non-baseline categories is thus obtained as the difference of their coefficients.

In general, we have the following interpretation:

- The coefficient of a dummy variable for a particular level of a categorical explanatory variable is the *difference* in the expected value of the response variable Y between a unit with that level of the categorical variable and a unit in the baseline category, holding all other explanatory variables constant.

The choice of the reference category does not affect the fitted model, and exactly the same results are obtained with any choice. For example, column (2) in Table 4.5 shows estimates for the same model, but now with Men and Ex-smokers as the reference levels. Expected differences in blood pressure are now -2.033 between women and men, 1.239 between never-smokers and ex-smokers, -0.143 between current and ex-smokers, and $-0.143 - 1.239 = -1.382$ between current and never-smokers. These are the same estimates as in column (1). You can confirm similarly that the same results are obtained also from column (3), where the reference level for smoking status is Current smoker. The models are also identical in fit (e.g. R^2) and significance of the effects, although these are not shown here. Note also that changing the baseline of one variable does not change the coefficients of others; here, for example, the coefficient of initial blood pressure is 0.573 throughout. The estimate $\hat{\alpha}$ of the constant term *is*, however, changed, because its interpretation depends on the choice of the baselines. For example, the constant 35.383 in column (1) is the expected blood pressure at exit for a woman ($S_m = 0$) who has never smoked ($C_e = C_c = 0$), while the 36.177 in column (2) refers to a male ex-smoker, both with initial blood pressure of 0[9].

Because the choice is arbitrary, the baseline level should be selected in a way which is convenient for stating the interpretation. If the categorical variable is ordinal, the baseline should usually be the first or last category. For example, smoking status is here arguably ordinal, and using its middle category (ex-smoker) as the reference level would seem rather awkward. For many variables, it is fairly easy to identify a neutral or "default" or most common category which serves well as a the reference level. Often this is simply a matter of trying out the interpretations of the cofficients as above, and picking the one which slips off the tongue most easily.

Predicted values are again obtained by substituting values for the explanatory variables into the estimated regression equation, now including appropriate zeros and ones for the dummy variables. For example, the predicted exit blood pressure for a man with initial blood pressure of 75 is

[8] These results seem to be largely due to differences in weight of people in these categories. The differences between all smoking categories become nonsignificant if we include weight as an additional control variable.

[9] This is another illustration of how the constant is often not substantively meaningful.

$$\begin{aligned}
\hat{Y} &= 35.383 + 0.573 \times 75 + 2.033 \times 1 - 1.239 \times 0 - 1.382 \times 0 \\
&= 80.4 \qquad \text{for a never-smoker, and} \\
\hat{Y} &= 35.383 + 0.573 \times 75 + 2.033 \times 1 - 1.239 \times 0 - 1.382 \times 1 \\
&= 80.4 - 1.382 = 79.0 \qquad \text{for a current smoker,}
\end{aligned}$$

using estimates from column (1) of Table 4.5. Repeating the calculation using the estimates in columns (2) and (3) to obtain the same values is another good way of convincing you that changing the reference categories does not affect the fitted model.

Significance tests and confidence intervals for coefficients of dummy variables are obtained exactly as for any regression coefficients. Since the coefficient is now interpreted as an expected difference between two levels of a categorical variable, the null hypothesis of a zero coefficient is that there is no such difference. The standard t-test of an individual coefficient is thus a test of the hypothesis of no difference between a particular category and the reference category. In cases where a variable has more than two categories, this is not the same as the hypothesis that the variable has no effect at all on the response. Such a claim corresponds to the hypothesis that the coefficients of the dummies for *all* of the categories are zero. This can be tested using the F-test.

To illustrate the differences between such hypotheses and their implications, consider the models in Table 4.6. Here the response variable is again diastolic blood pressure at exit, and we include initial blood pressure, age and sex as control variables. In addition, the model includes the main focus of HIE, the insurance plan to which the participant was assigned by the experiment (as explained in Section 4.2). Following Newhouse et al. (1993), the 14 plans are initially grouped into 5 groups according to the coinsurance rate: plans with rates of 25%, 50% and 95%, the individual deductible plan and the free-care plan. These are included in the model as four dummy variables, with 95% coinsurance rate as the reference level. Parameter estimates for this are shown as the first model in Table 4.6. A number of different tests can be carried out on these estimates, with different implications:

- The t-test of the coefficient of the dummy for each individual insurance plan tests the hypothesis that the expected exit blood pressure is the same for participants on that plan as on the 95% coinsurance plan, controlling for initial blood pressure, age and sex. The P-value for three of the coefficients is large, indicating that the null hypothesis is not rejected. There is thus no evidence of a differential effect on blood pressure between any of these plans and the 95% plan. The estimated difference is statistically significant (at the 5% significance level) only for the free-care plan, for which $P = 0.020$. The estimated coefficient for it is -2.009. The expected blood pressure is thus estimated to be around 2 mmHg lower for participants who received completely free medical care than for those on the 95% coinsurance plan, even after accounting for the three control variables.
- The finding that three of the coefficients are not significantly different from zero suggests that we might drop the corresponding dummy variables from the model. This is done in the second model of Table 4.6. Doing so has the effect of *combining* three of the levels of insurance coverage with the reference level of 95% coinsurance (and thus also with each other), leaving only the free-care plan separate

Table 4.6: Estimated coefficients for linear regression models for Diastolic blood pressure at exit from HIE, given insurance plan and control variables.

	Insurance plan in			
	5 categories		2 categories	
Variable	Coefficient (s.e.)	t (P)	Coefficient (s.e.)	t (P)
(Constant)	32.42		31.98	
Initial blood pressure	0.493	17.44	0.493	17.47
	(0.029)	(< 0.001)	(0.028)	(< 0.001)
Age	0.249	9.10	0.249	9.10
	(0.027)	(< 0.001)	(0.027)	(< 0.001)
Sex: male	3.991	4.08	3.938	4.03
	(0.979)	(< 0.001)	(0.977)	(< 0.001)
Insurance plan				
95% coinsurance	0	—	0	—
	—	—	—	—
50% coinsurance	0.091	0.07	0	—
	(1.382)	(0.947)	—	—
25% coinsurance	−0.772	−0.79	0	—
	(0.979)	(0.430)	—	—
Individual deductible	−0.727	−0.77	0	—
	(0.947)	(0.442)	—	—
Free care	−2.009	−2.32	−1.544	−2.49
	(0.866)	(0.020)	(0.621)	(0.013)
R^2	0.3536		0.3530	

from the rest. In the second model, insurance plan is thus used as a dichotomous explanatory variable, with the values free care vs. something else (non-free care). The coefficient of the free-care dummy is −1.544, and it is statistically significant, with $P = 0.013$. A 95% confidence interval for it is given by

$$-1.544 \pm 1.96 \times 0.621 = (-2.76; -0.33).$$

We are thus 95% confident that the expected diastolic blood pressure is between 0.33 and 2.76 mmHg lower for participants on the free-care plan than for those on any of the other plans, controlling for age, sex and initial blood pressure.

- Another inferencial strategy would be to start by testing the hypothesis that there are no differences between *any* of the insurance plans. This is the hypothesis that the coefficients of all of the plan dummies in the first model of Table 4.6 are 0. It can be tested using the general F-test of Section 4.5.3. In fact, testing for no effect of a categorical explanatory variable with more than two categories is

one of the most important uses of the F-test, because it inevitably implies a null hypothesis (4.15) that several coefficients are simultaneously zero.

Here the F-test statistic (4.16), calculated by statistical software, is $F = 1.80$, with $P = 0.127$. We are thus faced with a contradiction: the F-test indicates that none of the insurance plans differ from each other in their effect on blood pressure, whereas a t-test for the coefficient of the free-care dummy suggests that this plan does differ from the others. Apparent inconsistencies like this, while reasonably uncommon, are perfectly possible when testing related hypotheses in different ways. Here the finding is mostly due to the fact that the two tests have different *powers*, i.e. different probabilities of detecting true non-zero effects if they exist. The power of a significance test depends partly on the number of parameters involved in its null hypothesis, and it tends to be highest for tests of fewest parameters. Here the t-test is for a hypothesis that a single parameter is zero, so it has higher power than an F-test of four parameters. In essence, the t-test has a better chance of detecting a non-zero effect because it is focused on looking for a single specific difference, while the F-test tries to examine several possible effects at once. Because of this, the same data result in a smaller P-value for the t-test than for the F-test.

This, however, does not automatically imply that the result of the t-test is correct here. What we did in the first stage of the analysis was to look at four separate differences at once, and then decided to retain the largest and most significant of them. Such *multiple comparisons* are problematic, because the significance level of the combined process is different from the levels (of, say, 5%) of the individual tests. This is easily seen in the case of even more tests: if, for example, we carried out (independent) tests of 100 null hypotheses, around 5 of them would be expected to be significant at the 5% level purely by chance, even if all of the null hypotheses were actually true. For such reasons, extensive multiple comparisons should usually be avoided (extreme versions of them are sometimes described with such deservedly disparaging terms as "data dredging" or "data mining") or carried out with appropriate adjustments. A useful practical guideline is that effects should not be selected for further interpretation purely because they were the most significant in a multiple comparison, but only if they also make substantive sense. Ideally, such an effect would be one which was listed as a possible research hypothesis even before the model was fitted.

Here the free-care insurance plan is clearly different from the other plans in a substantially meaningful way, so it does not seem unreasonable to retain the distinction and interpret its effects (this is also what was done by Newhouse et al.). Nevertheless, the contradictory test results indicate that the evidence of its effect on blood pressure is not very strong.

Similar results for models for the General Health Index are shown in Table 4.7. Here the findings are clear, in that no differences between the insurance plans are found after we control for initial GHI, age and sex. This is in fact typical of the results of more careful analyses of the HIE data, which found very few effects of different insurance plans on health outcomes (the exceptions included blood pressure, correctable vision and periodontal health). Possible reasons for these results are discussed in Chapter 11 of Newhouse et al. (1993).

Table 4.7: Estimated coefficients for linear regression models for General Health Index at exit from HIE, given insurance plan and control variables.

	Insurance plan in			
	5 categories		2 categories	
Variable	Coefficient (s.e.)	t (P)	Coefficient (s.e.)	t (P)
(Constant)	28.97		29.71	
Initial GHI	0.635	32.38	0.634	32.39
	(0.020)	(< 0.001)	(0.020)	(< 0.001)
Age	-0.186	-7.24	-0.186	-7.25
	(0.026)	(< 0.001)	(0.026)	(< 0.001)
Sex: male	0.463	0.50	0.526	0.57
	(0.918)	(0.614)	(0.917)	(0.57)
Insurance plan				
95% coinsurance	0	—	0	—
	—	—	—	—
50% coinsurance	-0.339	-0.25	0	—
	(1.350)	(0.802)	—	—
25% coinsurance	0.816	0.85	0	—
	(0.954)	(0.393)	—	—
Individual deductible	1.355	1.46	0	—
	(0.927)	(0.144)	—	—
Free care	0.741	0.86	0.074	0.12
	(0.857)	(0.387)	(0.624)	(0.906)
R^2	0.4082		0.4072	

4.6.2 Interactions

Suppose we have a linear model with two explanatory variables X_1 and X_2, and the mean model $\mu = \alpha + \beta_1 X_1 + \beta_2 X_2$. Here the coefficient β_2 of X_2 describes the effect of X_2 on Y, controlling for X_1. Importantly, this effect is the same irrespective of the value at which we fix X_1 while considering the partial effect of X_2.

The situation changes if we create a new variable X_3 which is the product of X_1 and X_2, i.e. $X_3 = X_1 X_2$, and add this to the model. The expression for μ is now

$$\mu = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) = \alpha + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2. \tag{4.18}$$

The last expression shows that the coefficient of X_2 in (4.18) is now $\beta_2 + \beta_3 X_1$, which depends also on X_1. In other words, the partial effect of X_2 on Y depends on the value at which X_1 is controlled. In this case there is said to be a (two-way) statistical *interaction* between X_1 and X_2. In general, such an interaction is defined as follows:

- There is an **interaction** between two explanatory variables in a regression model if the magnitude of the partial effect of one of them on the response variable depends on the value at which the other explanatory variable is fixed.

The definition is symmetric in the two variables. For example, we could rearrange (4.18) as $\mu = \alpha + \beta_2 X_2 + (\beta_1 + \beta_3 X_2)X_1$. This emphasises how the coefficient of X_1 depends on X_2 rather than vice versa, but the model is unchanged. In practice the roles of the two variables are discussed in whichever order is more natural for interpretation.

An interaction is included in a model by using the product of the two variables as an additional explanatory variable. A model which includes such an interaction term — e.g. $\beta_3(X_1X_2)$ above — should also always include the corresponding terms involving the two variables alone, e.g. $\beta_1 X_1$ and $\beta_2 X_2$ above; these terms are known as the **main effects** of X_1 and X_2. A model which satisfies this condition is known as a **hierarchical model**. It is technically possible to fit non-hierarchical models which include an interaction but omit some of the corresponding main effects, but they are inconvenient for interpretation and have other technical flaws. Non-hierarchical models should usually be avoided in both linear models and other kinds of regression modelling.

Each of the variables in an interaction (X_1 and X_2 in 4.18) can be a continuous or a dummy variable. Although the general principle is the same in every case, the detailed interpretation of an interaction depends slightly on the types of the variables. To illustrate the three possible cases, we use models for diastolic blood pressure at exit, contolling for initial blood pressure, age, sex, family income and insurance plan. Three models are considered. Each of them includes an interaction between income and insurance plan, but the versions of these variables used in the models are dfferent. Estimated coefficients are shown in Table 4.8.

Model (1) includes income as a continuous variable (as explained on p. 69) and insurance plan as a dummy variable for the free-care plan. The coefficients of these are -0.096 and -0.874 respectively, and the coefficient of their interaction (product) term is -0.064. These can be interpreted in two equivalent ways (both controlling for initial blood pressure, age and sex):

- The estimated coefficient of family income is -0.096 for respondents who were not on the free-care insurance plan, and $-0.096 - 0.064 = -0.160$ for those on the free-care plan. In other words, for each additional 1000 dollars of income, expected blood pressure at exit decreases by 0.096 units for those on non-free care, and by 0.160 units for those on the free-care plan.
- The coefficient of the dummy variable for free care is $-0.874 - 0.064 \times$ Income. In other words, expected blood pressure at exit is 0.874 units lower for respondents with free care than for others among those with family income of 0, and this difference increases by 0.064 units for each additional 1000 dollars of income.

These results are summarized in the upper plot of Figure 4.7, which shows how fitted values for blood pressure at exit depend on income, separately for free care and the other plans (fixing the control variables at average values). The interaction is indicated by the fact that the fitted lines are not parallel (c.f. Figure 4.6 for a model without an

Table 4.8: Estimated coefficients for linear regression models for Diastolic blood pressure at exit from HIE, given main effects and an interaction of family income and insurance status, and three control variables. Three different versions of the income and insurance variables are considered; see the text for explanation.

	Model		
Variable	(1)	(2)	(3)
Initial blood pressure	0.487	0.483	0.471
Age	0.270	0.260	0.273
Sex: Male	4.093	3.981	3.643
Income:			
in $1000	−0.096	—	−0.143
low income (lowest 20%)	—	2.662	—
Insurance plan:			
free care	−0.874	−1.299	—
coinsurance rate (%)	—	—	0.013
Income×Insurance plan	−0.064	−1.262	0.00072
(Constant)	32.88	31.83	33.23

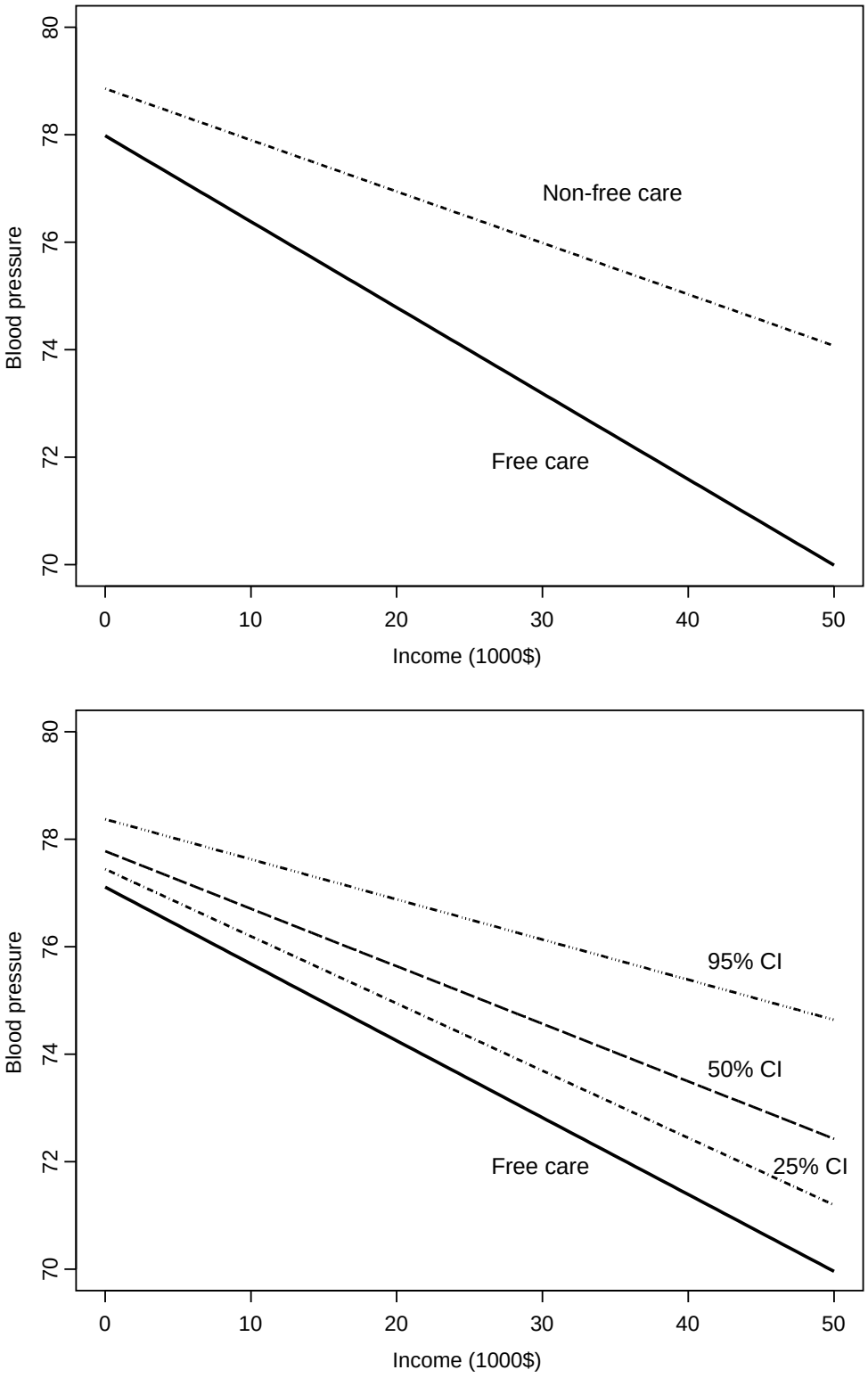
interaction). In short, the effect of income on blood pressure seems to be stronger for those on free care than for others.

In Model (2), dichotomous variables are used for both income and insurance plan. Income is here classified into two groups, one for the respondents whose family income is in the lowest 20% of the incomes in the data set, and the other for the remaining 80%. Dummy variables are included for free care and for the low-income group. Their product, which corresponds to the interaction, has the value 1 for a respondent who is on the free-care plan *and* has low income, and 0 for everyone else. The estimated coefficients of these variables can be interpreted in two equivalent ways (both controlling for initial blood pressure, age and sex):

- For a respondent who is not in the low-income group, the coefficient of the free-care dummy is -1.299, i.e. expected blood pressure at exit is 1.299 units lower on the free-care plan than on other insurance plans. For a respondent in the low-income group, this coefficient is $-1.299 - 1.262 = -2.561$, i.e. expected blood pressure is 2.561 units lower on the free-care plan.
- For respondents who do not have a free-care insurance plan, expected blood pressure is 2.662 units higher for people with low income than for others. Among respondents with free health care, this difference is $2.662 - 1.262 = 1.4$ units.

In short, differences between income groups are smallest among people receiving free care, or, free care has the biggest effect on blood pressure among those with low income.

Figure 4.7: Fitted values of blood pressure at exit as functions family income in thousand of dollars for models (1) [upper plot] and (3) [lower plot] in Table 4.8.



Control variables are fixed at initial blood pressure 75, age 35 and sex: female.

In model (3), both insurance plan and income (in \$1000, as in model 1) are used as continuous variables. For the insurance plan, respondents on the individual deductible plan are now omitted, and the coinsurance rate is used as a measure of insurance coverage for the rest. This has only the values of 0, 25, 50 and 95 per cent in these data, but obviously could in principle take other values between 0 and 100.

One way of illustrating an interaction between two continuous variables is to plot the dependence of the response on one of the variables at selected values of the other. This is done in the lower plot of Figure 4.7, which shows fitted blood pressures given income, at the values 0, 25, 50, and 95 for the coinsurance rate. Clearly the effect of income decreases with increasing rate of coinsurance. For example, the coefficient of income for respondents on free care is -0.143, while for those with 95% coinsurance it is $-0.143 + 0.00072 \times 95 = -0.075$.

A case not covered by these examples is one where one or both of the variables in the interaction are categorical and have more than two categories. The main effect of such a variable is modelled using several dummy variables, one for all but one of the categories. To include an interaction term, we then need to add to the model the products of *all* the pairs of variables corresponding to the two main effects. For example, if one variable (X_1) is continuous and the other has three categories captured by two dummy variables X_2 and X_3, the product variables X_1X_2 and X_1X_3 are needed to describe the interaction. Similarly, for an interaction between the categorical variable with dummy variables X_2 and X_3 and another with dummy variables X_4 and X_5, the products X_2X_4, X_2X_5, X_3X_4, and X_3X_5 are required. This can clearly become somewhat complicated when one or both of the variables have many categories, and the interaction then requires particularly careful interpretation and presentation.

Standard tests are used to test whether an interaction is needed in a model. For the models in Table 4.8, the interaction between income and insurance plan is captured by the coefficient of a single product variable (Income×Plan). A t-test of the null hypothesis that this coefficient is 0 is also a test of the hypothesis of no interaction. In the cases discussed in the previous paragraph, where interaction effects are described by the coefficients of several product variables, the hypothesis of no interaction is tested with an F-test of the hypothesis that all of those coefficients are 0.

For all of the models in Table 4.8, the interactions are actually nonsignificant, with P-values of 0.48, 0.42, and 0.57 for models (1), (2) and (3) respectively. These models were used here purely to illustrate interactions of different types of variables with actual estimated coefficients, and should not be interpreted seriously. It should be noted, however, that Newhouse et al. (1993), using a model for blood pressure with a different set of control variables, did find and discuss a moderately significant interaction comparable to that in model (2), i.e. between low income and free health care.

You may have noticed that the qualitative interpretation of the interaction in these examples was different for model (2) than for (1) and (3): for the former, free health care has the largest effect when income is low, while the opposite is the case for the latter. This is not an error, but reflects somewhat different observed effects at different ends of the income distribution. It would be possible to elaborate the model in ways which combined the two findings. This, however, is not really sensible here, as the interaction effects are not actually significant.

Statistical interactions are a fairly complex and subtle element of regression models, and understanding them may seem difficult at first. Learning about interaction effect is, however, worth the effort, as some of the most powerful and interesting findings of modelling correspond to interactions. Further examples of interaction effects will be provided later in this coursepack, in the computer classes for linear regression models, and in both classes and the main text for logistic models.

4.6.3 Nonlinear functions of explanatory variables

The modelling of interaction effects involves the use of explanatory variables which are not individual statistical variables in the usual sense of the word (c.f. Section 1.2), but products of two such variables. This may seem like a change to the definition of the linear model stated in Section 4.3.2, but this is not really the case. The symbols $X_1, X_2, \ldots, X_k$ in the model equation (4.1) may in fact represent either individual original variables or derived variables obtained through calculations on the values of one or more of them. This is why we were able to add an interaction simply by calculating, say, $X_3 = X_1X_2$ and including it in the model just like any other explanatory variable.

The same idea works also for calculations (transformations) of individual variables, with one restriction: the model may not include a set of variables which are *linear* combinations of each other in a way which gives rise to exact multicollinearity of the kind discussed in Section 4.9.1. For example, we may use height in inches X_1 or height in centimetres $X_2 = 2.54X_1$ but not both, or the model may not simultaneously include two variables X_1 and X_2 and their mean $X_3 = (X_1 + X_2)/2$.

Nonlinear transformations of continuous explanatory variables are allowed, and provide very flexible ways of extending the model. One common example of this is the use of *polynomial* regression models, which include both a variable X and higher powers $X^2, X^3, \ldots$ of it. For example, using X and its square X^2 gives the *quadratic model*

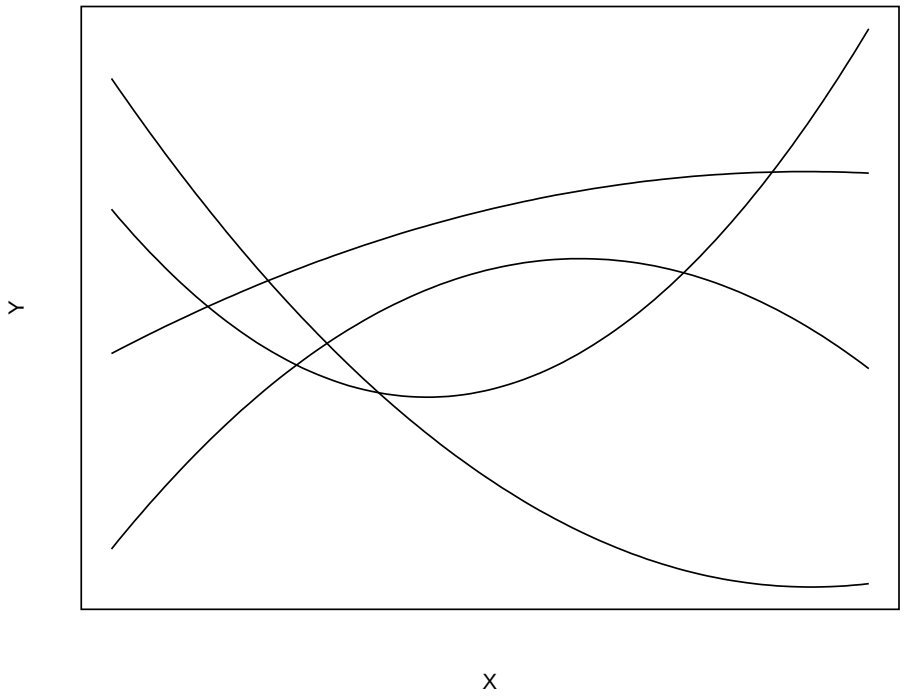
$$\mu = \alpha + \beta_1 X + \beta_2 X^2 \tag{4.19}$$

where $Y = \mu + \epsilon$, as usual. Adding to this higher powers X^3, X^4, and so on, would produce higher-order polynomial models. Here we will concentrate on the quadratic model, which is in practice the most common. This would typically also include other explanatory variables, although these have been omitted from (4.19) for simplicity.

Before discussing the quadratic model further, a few incidental points should be noted:

- Including both X and X^2 does not lead to exact multicollinearity, because they are not *linearly* related. Their sample correlation is thus not exactly 1.
- A polynomial model should usually contain all powers of an explanatory variable up to the highest one included. For the quadratic model, this means that X should always be included when X^2 is. The reason for this is the same as the reason we do not include interactions without corresponding main effects, and do not omit the constant term from any linear model.
- The possibility of nonlinear transformations of explanatory variables explains why a careful definition of the linear model (see the discussion on p. 31) states

Figure 4.8: Examples of nonlinear dependencies produced by different values of the parameters of the quadratic model (4.19).



that the mean model is linear in the *parameters* rather than the Xs. Thus (4.19) is a linear model, but $\mu = \alpha X^{\beta}$ is not.

Quadratic models provide a simple way of representing some nonlinear patterns for how the mean μ of Y depends on X. Equation (4.19) defines a *parabola*, the shape of which is determined by the values of the parameters α, β_1 and β_2. With different choices for these, and by limiting consideration to different ranges of X, a wide variety of shapes of dependence are obtained. Some examples of these are shown in Figure 4.8. Note, in particular, that even though a full parabola both increases and decreases, this does not need to be the case within the observed range of values of X. Thus the dependence implied by the model can also be one where the rate of increase (or decrease) just decreases (or increases) over the relevant range of X.

As an example of this kind of dependence, consider the models displayed in Table 4.9 and Figure 4.9. These are for blood pressure at exit, given initial blood pressure and the person's weight in kilograms. The first model includes only the linear effect of weight. The second also contains the quadratic effect of it, which has been included simply by creating a new variable for weight^2, and adding it to the model. The coefficient of this squared weight variable is statistically significant, so there is evidence that the quadratic model is an improvement over the linear one.

When a model includes a quadratic of effect of a variable X, the coefficients of X and X^2 cannot be interpreted in the usual simple way for coefficients in linear models. The reason is that it is not possible to change one of X and X^2 without changing the other. Thus, β_X no longer corresponds to the increase in the expected value of Y when we

Table 4.9: Estimated coefficients (with P-values in parentheses) for linear models for Diastolic blood pressure at exit from HIE, given initial blood pressure and linear and quadratic effects of initial weight.

	Effect of weight			
Variable	Linear		Quadratic	
(Constant)	27.36		18.06	
Initial blood pressure	0.520	(< 0.001)	0.518	(< 0.001)
Weight	0.174	(< 0.001)	0.435	(< 0.001)
Weight^2	—		-0.0017	(0.023)
R^2	0.343		0.346	

increase X by one, holding other variables constant because when we change X, we are also changing X^2. Of course this is the whole point of including the X^2 term: a nonlinear model is one where increasing X by one unit does not correspond to the same increase in the expected value of Y regardless of the value of X.

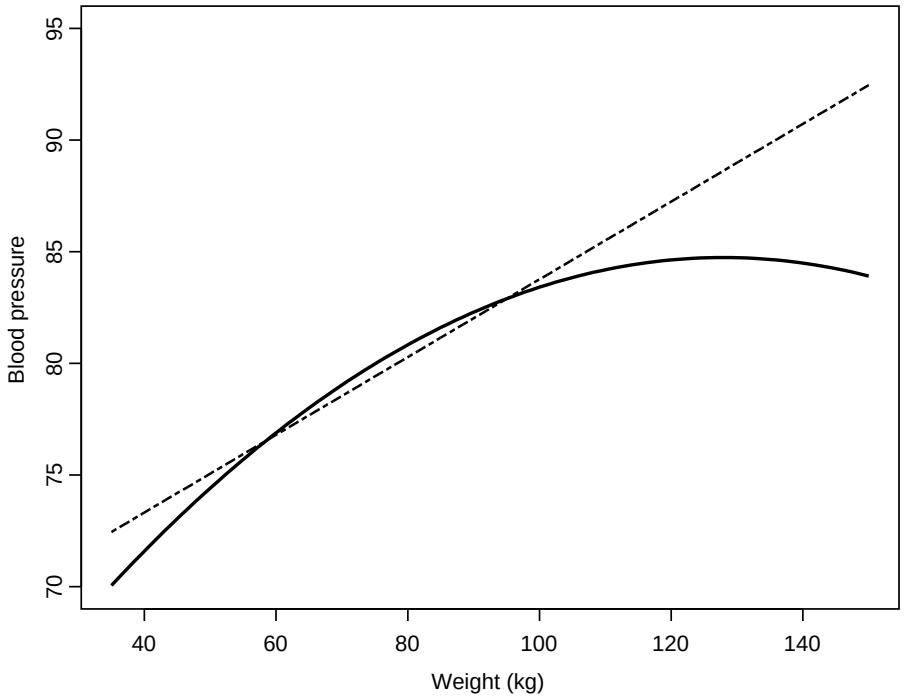
As discussed above, an interaction between two explanatory variables X_1 and X_2 is entered in a model by including their product X_1X_2 as an explanatory variable. The square of an individual variable, $X^2 = X \cdot X$, is also a product, that of X with itself. Thus we might think of a quadratic effect as an interaction of a variable with itself. This is not wrong, and indeed provides one intuitive way of explaining a quadratic model: the association between X and Y (the steepness of the curve of X against the expected value of Y) is not constant but depends on the specific value of X where we want to state the association.

Because the coefficients cannot be directly interpreted, it is easiest to illustrate the model by plotting or tabulating fitted values. Figure 4.9 shows an example for the models in Table 4.9, including the one with linear weight effect for comparison. It can be seen that the difference between the models emerges mainly for people with the highest weights. Beyond about 100kg, the quadratic model suggests that expected blood pressure will not continue to increase steadily (as implied by the linear model), but will stabilise at around 83. The curve also appears to start to decrease beyond about 140kg, but this is an artifactual result due to sampling variation and should not be seriously interpreted, as there were very few people with such weights in the sample (recall that the data are from the 1970s). Note also that the two curves are actually quite similar up to about 100kg, a range which contains around 96.5% of the respondents in the sample.

The second nonlinear transformation discussed here is the logarithm of an explanatory variable X, here denoted $\log(X)$.[10] Unlike X^2 and other powers of X, this is included

[10]Here and elsewhere in this coursepack log denotes the natural or base-e logarithm (where $e \approx 2.71828$), also commonly written (e.g. in calculators) as $\log_e$ or ln. Some basic properties of logarithms are summarised in the Appendix, on p. **??**. These are used to derive some of the results here and in Chapter 5. However, they are not essential for understanding the basic ideas.

Figure 4.9: Fitted values of exit blood pressure given weight for the linear (dashed line) and quadratic (solid line) models in Table 4.9.



Initial blood pressure fixed at 75.

in the model instead of, rather than in addition to, X. The model is thus of the form

$$\mu = \alpha + \beta \log(X), \tag{4.20}$$

which may again include other explanatory variables as well. The reason for considering $\log(X)$ is typically less to improve the fit of the model (although this may also happen) than to obtain a convenient interpretation. The estimated coefficient of $\log(X)$ can be interpreted in terms of effects on expected value of Y of *proportional* (rather than additive) changes in X itself. For example, suppose that X increases by 10%. Because $\log(1.1 \times X) = \log(X) + \log(1.1) = \log(X) + 0.095$, this proportional increase corresponds to an additive increase of about 0.095 in $\log(X)$ in model (4.20). Applying the usual interpretation of the coefficients of linear models, this means that 0.095β for model (4.20) is the expected change in Y when X increases by 10% (controlling for other explanatory variables, if included). Similarly (because $\log(2.72) = 1.0$), β itself indicates the expected change corresponding to multiplying X by 2.72, i.e. a 172%-increase in X.

As an illustration, consider the models for blood pressure at exit given family income and other variables shown in Table 4.10 and Figure 4.10. The first model includes income itself, while the second uses log-income instead. For the log model, we first need to discuss a common problem with models which include log-transformations of any variables. This is that the logarithm cannot be calculated when the value of a variable is 0 (or negative). One relatively common ad-hoc solution is then to add a small constant to every observation of the variable before taking the logarithm (or sometimes just to every observation which is 0). Here we have done this, by defining the log-income variable as the log of the recorded family income plus 1; in effect, we

pretend that everyone's family income was \$1000 higher than it actually was. We note, however, that this is in general a rather unsatisfactory solution, with a degree of arbitrariness and various conceptual problems, so it cannot be recommended as a general strategy. A different, also imperfect remedy to the same problem is to exclude observations with values of 0. This is discussed in the context of a different example on page 93.

In Table 4.10, the estimated coefficients of the other explanatory variables as well R^2 are essentially the same for the two models. For income, its coefficient of -0.115 in the first model indicates that expected blood pressure decreases by 0.115 for every \$1000-dollar increase in income. The coefficient of log-income in the second model is -1.298, indicating that expected change in blood pressure when income increases by 10% is a decrease of $0.095 \times 1.298 = 0.123$. The fitted values in Figure 4.10 suggest that these effects of income on blood pressure implied by the two models are actually rather similar, especially for incomes below \$30,000, which account for 98% of the respondents in the sample. There is, however, some difference at the lowest end of the income scale, where the logarithmic model indicates a more dramatic increase in expected blood pressure.

How should you choose whether to use $\log(X)$, or include X^2, or X^3, etc? There are a few ways to make this decision. First, you should think about whether a log transformation makes sense in the particular case you are considering. Would it make sense for a multiplicative change in X to correspond to an additive change in Y? This might make sense for an X variable like income and a Y variable like life expectancy. Having a higher income might be associated with longer life expectancy, but we would expect the effects to be diminishing as income increases. Thus, it might make sense to assume that the difference between the life expectancy of someone with an income of £10000 and someone with an income of £20000 would be the same as the difference between £20000 and £40000, rather than assuming a linear relationship. Once you have made this decision, you should examine the residual plots as described in Section 4.7 to see if your model fits well.

While log transformations of explanatory variables often can be justified in terms of the application, as above, quadratic, cubic, and higher polynomial transformations seldom reflect any substantive theory about the relationship between an explanatory variable and the outcome. They are useful when one is really not sure of the relationship between X and Y, and one wants to allow for the possibility that the relationship is not linear. The cost is that the coefficients are not directly interpretable in most cases, so one typically must plot fitted values to make sense of the estimates.

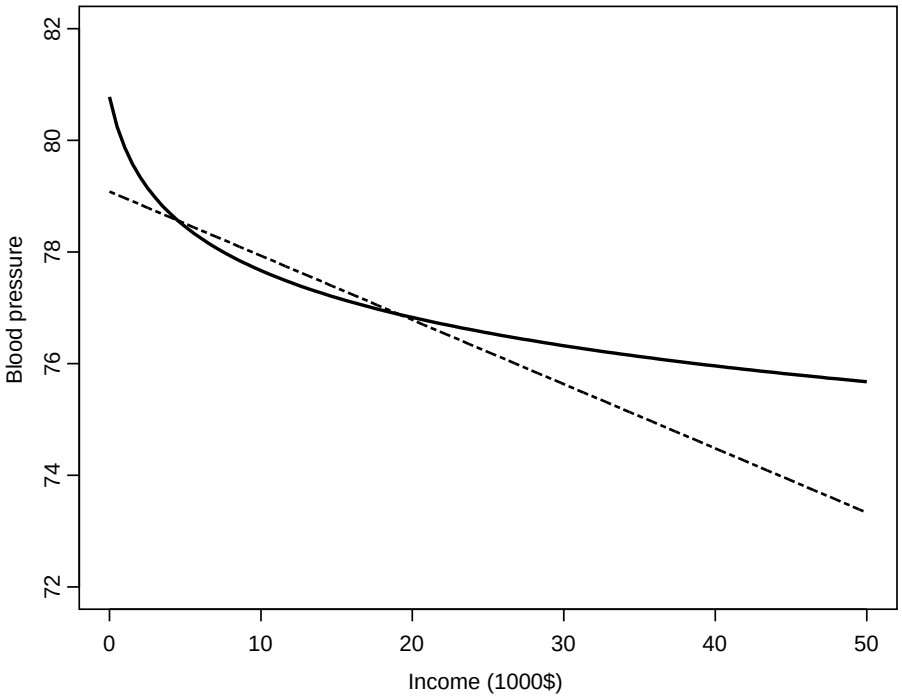
4.6.4 Ordinal-level explanatory variables

The values of an interval-level explanatory variable are included in a regression model directly, as with age and income in the examples of this chapter. Inclusion of nominal categorical variables is also straightforward, as they must be used in the form of dummy variables. This leaves the case of an ordinal-level explanatory variable. As an example, consider the responses to the following HIE survey question, measured at enrolment: *"Would you say that your health, in general, is excellent, good, fair, or, poor?"* This

Table 4.10: Estimated coefficients (with P-values in parentheses) for linear models for diastolic blood pressure at exit from HIE, given family income or its logarithm, and other control variables.

	Income included as			
Variable	Linear		Log income	
(Constant)	33.14		43.99	
Initial blood pressure	0.487	(< 0.001)	0.485	(< 0.001)
Age	0.269	(< 0.001)	0.268	(< 0.001)
Sex: male	4.088	(< 0.001)	4.097	(< 0.001)
Free health care	-1.633	(0.009)	-1.610	(0.010)
Family income	-0.115	(0.008)	—	
Log of family income	—		-1.298	(0.007)
R^2	0.357		0.358	

Figure 4.10: Fitted values of exit blood pressure given income models in Table 4.10, when income is included as linear (dashed line) or logarithmic (solid line).



Control variables are fixed at initial blood pressure 75, age 35, female, and non-free care.

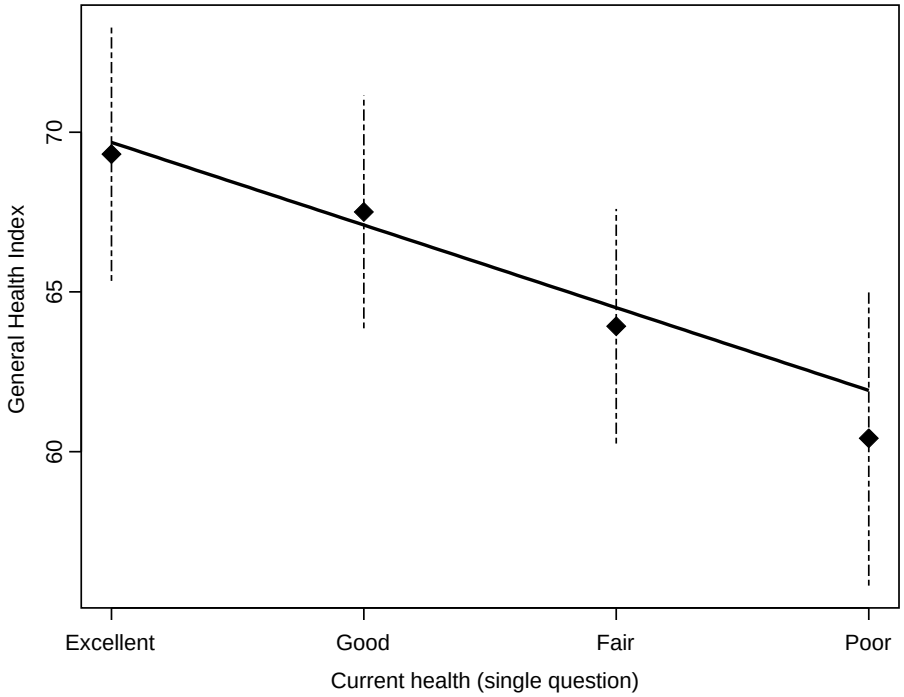
Table 4.11: Estimated coefficients for linear models for General Health Index at exit from HIE, given GHI at enrolment, age and self-reported current health (from a single survey question) at enrolment (with standard errors in parentheses). Current health is included either as continuous (with values 0, 1, 2, 3) or categorical.

	Current health as		
Variable	Categorical		Continuous
Initial GHI	0.567	0.567	0.567
	(0.023)	(0.023)	(0.023)
Age	−0.159	−0.159	−0.160
	(0.026)	(0.026)	(0.026)
Current health (continuous)	—	—	−2.588
			(0.474)
Current health:			
Excellent	0	35.187	—
	—	(2.023)	
Good	−1.809	33.378	—
	(0.677)	(1.857)	
Fair	−5.389	29.798	—
	(1.154)	(1.871)	
Poor	−8.895	26.292	—
	(1.995)	(2.352)	
(Constant)	35.187	0	35.445
	(2.023)	—	(2.017)
R^2	0.421	0.421	0.420

is here referred to as "Current health". We will consider it as an explanatory variable for General Health Index at exit, controlling for initial GHI and age. Estimates for two models like this are shown in Table 4.11.

Ordinal explanatory variables are in practice either demoted or promoted, i.e. treated as if their measurement level was actually nominal or interval. The nominal approach involves simply defining dummy variables for the categories. Doing this is always appropriate, but it has the disadvantage that it ignores the ordering of the categories and requires many parameters (one for each of the dummy variables). A model like this is shown in the first column of Table 4.11. Here "Excellent" is the reference category for current health, and dummy variables for the other three are included. Their coefficients are all negative, so the expected GHI at exit is highest when current health is excellent, even controlling for age and initial GHI itself (the differences are statistically significant). The coefficients are most negative for categories corresponding to poorest current health, indicating consistently increasing difference from being in excellent health. The estimates are thus in the order suggested by the categories of current health, even though this ordering was not imposed by the dummy variables.

Figure 4.11: Fitted values of GHI at exit given current health at enrolment, calculated from the models in Table 4.11. The solid line shows fitted values for the model where current-health score is treated as an interval-level variable. The diamonds show fitted values for the model where current health is treated as a nominal variable, with approximate 95% confidence intervals around them (these intervals ignore the variability in the coefficients of the two control variables).



Control variables are fixed at initial GHI 70, age 35.

The second way to treat an ordinal explanatory variable is to assign values (*scores*) to its categories and to include these scores as if they were values of an interval-level variable. This has the advantage that the effect of the variable is then described by a single parameter, but the disadvantage that doing so might not actually be appropriate. The numerical values of the scores must be in the correct order, but otherwise they are arbitrary. The scores do not even have to have equal intervals: something like 1, 5, 7, 15 would not necessarily be any less valid than, say, 1, 2, 3, 4. Normally, however, such exotics would be used only if there was a substantive reason to do so. The most important example of this is when an ordinal variable represents grouped values of a continuous variable. Suppose, for example, that age is recorded only in four groups, 20–30, 31–50, 51–70, and over 70. It is then common practice to use midpoints of the intervals as scores for the categories, e.g. here 25, 40, 60 and, say, 80 (because the last category is open-ended, the choice of its score is somewhat arbitrary).

A model like this is shown in the third column of Table 4.11. Here the values 0, 1, 2, 3 were assigned to levels of current health from Excellent to Poor. The estimated coefficient of this score is -2.588. In other words, moving from any one category to the next is associated with an expected decrease of 2.588 GHI points, controlling for age and GHI at enrolment. Figure 4.11 show fitted value for GHI given these scores, fixing age at 35 and initial GHI at 75.

Also shown in Figure 4.11 are comparable fitted values for the model where current health is treated as nominal. Here the contribution of age and initial GHI to the fitted values is always $0.567 \times 75 - 0.159 \times 35 = 36.960$. From the first column of the table, where Excellent is the reference level for current health, we obtain the fitted values of, say, $36.960 + 35.187 = 72.147$ when current health is Excellent, and $36.960 + 35.187 - 1.809 = 36.960 + 33.378 = 70.338$ when it is Good. Another way of calculating these values is illustrated by the second column of Table 4.11. This shows estimates for the same model, now fitted without the constant term but including dummy variables for all four categories of current health. The coefficients of the dummy variables now show directly the contributions of the categories to fitted values: for example, this is 35.187 for Excellent and 33.378 for Good current health. These are, of course, the same as for the previous version of the same model. This second, less common way of fitting the model is shown here partly to provide further illustration of the interpretation of dummy variables. It also allows us to calculate confidence intervals for the contributions of the categories to the fitted values simply as standard intervals for the coefficients of the four dummy variables. These are also shown in Figure 4.11[11].

The two models in Table 4.11 and the fitted values in Figure 4.11 are actually quite similar. Estimates for the model where current health is treated as an interval-level variable with equal-interval scores imply that each one-category increase in the score (corresponding to worsening health) is associated with an expected decrease of 2.588 GHI points. From the model where current health is treated as a categorical variable, the estimated decreases between successive categories are 1.809, 3.580 (= 5.389−1.809), and 3.506. Allowing for their standard errors, these are actually all fairly consistent with 2.588 (see also the confidence intervals in Figure 4.11). It thus appears that it might not be inapproriate to include current health in the form of equal-interval scores.

We can actually make a formal decision between the two models, by calculating the F-test statistic (4.16) between them. This is possible because the model with an ordinal explanatory variable included as category scores is nested in the model where it is included as dummy variables, in a way which makes the test appropriate (the null hypothesis is then a generalisation of 4.15). The degrees of freedom for this test are $K - 2$ and $n - (k + 1)$, where K is the number categories of the ordinal variable, and k the number of all the explanatory variables (including the dummies) in the model where it is treated as nominal. Here the test statistic is $F = 1.31$, and $P = 0.27$. There is thus no reason to reject the simpler model in favour of the more complex one. Including current health in the form of scores with equal intervals thus seems appropriate for these data.

In general, when an ordinal explanatory variable is treated as nominal, the coefficients of the dummy variables show the estimated effects of moving from one category to the next when no constraints are imposed on them. These may or may not be consistent with a simple linear trend implied by using the ordinal variable in the form of category scores. Informal comparisons and significance tests like the ones above can help us to choose between the two approaches.

[11] Note that these are not the confidence intervals for the fitted values themselves, which would also incorporate uncertainty in the estimated coefficients of the two control variables, and would be somewhat wider than the intervals shown here. The calculation of confidence intervals for fitted values is not covered on this course.

4.7 Model diagnostics using residual plots

The second tool of model assessment discussed here is the use of **sample residuals**

$$e_i = Y_i - \hat{Y}_i,$$

i.e. the differences between the observed (Y_i) and fitted ($\hat{Y}_i$) values of the response variable, calculated for each observation $i = 1, \ldots, n$ in the data set. These can be regarded as "estimates" (more accurately, predictions) of the model residuals ϵ_i in equation (4.2).

The magnitude of the residuals depends partly on the units in which Y is measured. For model assessment, it is often somewhat more convenient to use standardised residuals which do not have this property. Statistical software packages sometimes provide separately "standardised" residuals $e_i/\hat{\sigma}$ and "studentised" residuals, which are divided by a further quantity so that they will be approximately normally distributed with a standard deviation of 1 if the model is correct. We will use studentised residuals here and standardised ones in the computer classes. It is important to note, however, that raw, standardised and studentised residuals are not qualitiatively different: all of them will always give residual plots of a similar shape.

Scatter plots of the residuals against selected other quantities are used for model assessment. This can help to detect signs of possible failures of some of the model assumptions, an activity known as model **diagnostics**. Depending on what the residuals are plotted against, diagnostics of different aspects of the model are obtained.

Plots of residuals (usually placed on the Y-axis) against individual explanatory variables X can suggest changes to the model. If the X-variable in this plot is already included in the model, the plot cannot have a linear trend (which is already included) but it can show a nonlinear one. An example of this is Figure 14.3(c) of Agresti & Finlay (p. 451 in the 4th Edition). This suggests that a nonlinear transformation of X, here obviously its square X^2, should be added to the model. If, on the other hand, X is not included in the model from which the residuals were calculated, this plot can also show a linear trend. The obvious remedy is then to add X to the model.

A plot of this kind which most closely matches the ideas of multiple regression is actually one that plots the residuals against not X but the residual from a fitted model for X given all the other explanatory variables in the model. This is known as the *partial regression plot* or *added-variable plot*. It is discussed in Section 11.2 of Agresti & Finlay. It is interpreted in the same way as a plot of residuals against X.

Plots of residuals against explanatory variables provide suggestions about which variables might be added to a model. These variables would then usually be examined further with a significance test, and included if they are significant. Very often this is done directly as discussed above, without considering the plots at all. It is for another purpose that residual plots are particularly useful. This is the examination of the assumption of constant variance (homoscedasticity) of the error terms of the model. If this assumption is violated, the error terms are said to be *heteroscedastic*. If there is serious heteroscedasticity, usual estimates of the standard errors of the regression coefficients will be incorrect, and so will test statistics and confidence intervals. Results

of statistical inference will then also be invalid, in that confidence intervals will be too wide or too narrow, and tests may lead us to conclude incorrectly that partial effects are zero when they are not or vice versa.

To illustrate the use of residual plots to examine the variability of the error terms, we will use another example from HIE. Here the response variable is

- "Outpatient expenses": a participant's total annual covered expenses for outpatient medical services during their second year of participation (in dollars).

The explanatory variables are age, General Health Index and log of family income (the variable used also in Section 4.6.3) at enrolment, and a dummy variable for being on the free-care insurance plan. The models reported here include only those $n = 1496$ respondents whose outpatient expenses were not zero, for reasons discussed below.

A plot of (studentised) residuals for this model is shown in the top-left corner of Figure 4.12. Here, and in general for this purpose, the residuals are plotted against the fitted values $\hat{Y}_i$ from the model. There is then only one plot even for multiple linear models, rather than separate ones against each explanatory variable.

Model diagnostics from a residual plot like this are based on the general shape of the cloud of points. What we would like to see is a plot where the points form a band of roughly even width, as in the lower-left corner of Figure 4.12 (explained below) or in Figure 14.3(a) of Agresti & Finlay. Such a plot indicates that the variability of the residuals is of similar magnitude at different fitted values (and, by implication, at the corresponding values of the explanatory variables), i.e. it is evidence that the assumption of homoscedasticity is not violated.

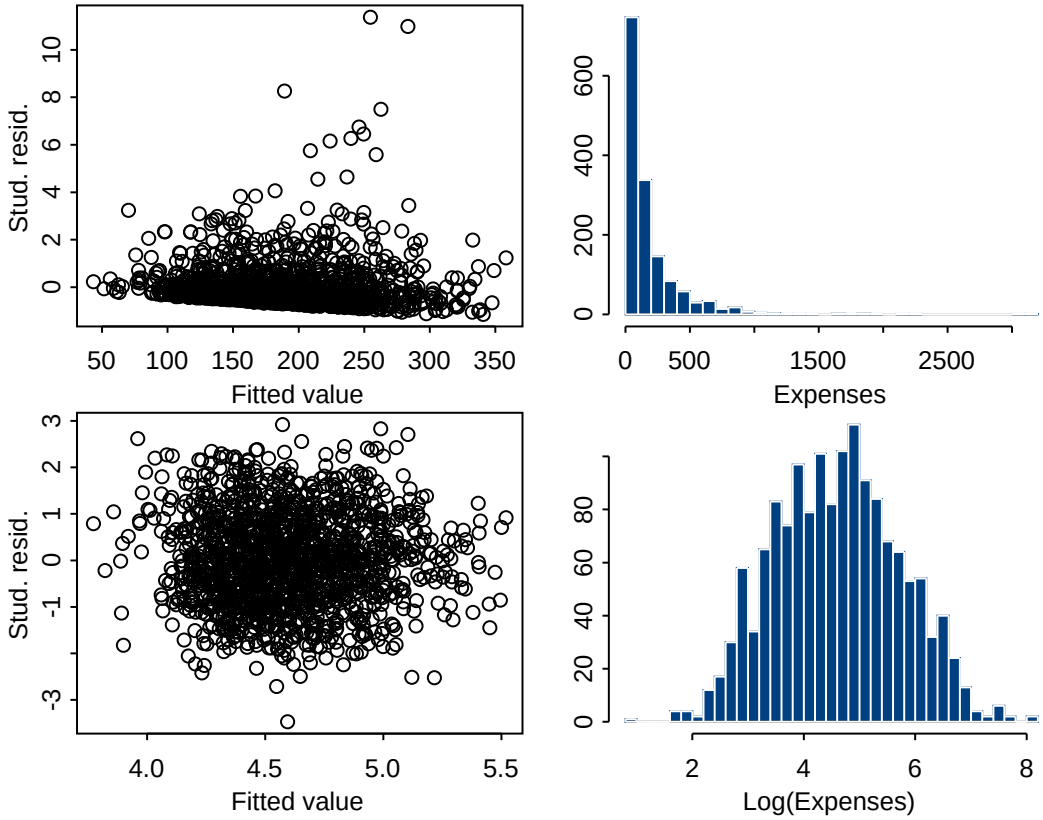
The first plot in Figure 4.12 is not of this kind. Instead, it has a "funnel" shape where the variability of the residuals increases when the size of the fitted values decreases or (as here) increases (see also Fig. 14.3(b) of Agresti & Finlay). This is evidence that the error terms are heteroscedastic, i.e. that their variance is not constant.

There are a number of possible remedies to heteroscedasticity of error terms, but only one of them is discussed here. This is to change the model by transforming the *response* variable, specifically by using $\log(Y)$ instead of Y as the response. Residual plot for such a model in this example, with the logarithm of outpatient expenses as the response, is shown in the lower-left corner of Figure 4.12. Clearly the funnel shape of the first plot has disappeared, and there is now no obvious evidence of heteroscedasticity.

Obvious heteroscedasticity which is removed by a log-transformation of the response variable is often found when the original response has a very skewed sample distribution. This is very clearly the case for outpatient expenses, as shown by the histogram in the top-right corner of Figure 4.12. The median of this distribution is around 100, while 20 of the 1496 respondents had expenses of over 1000 dollars, largest of them 3159. The histogram of the log-expenses shows that the log-transformation removes the skewness, turning the distribution into a roughly symmetric one.

As discussed in Section 4.6.3, the interpretation of the coefficient of a log-transformed explanatory variable is based on the general result that additive changes of logarithms

Figure 4.12: On the left, plots of studentised residuals against fitted values for linear models for outpatient expenses in year 2 (top row) or its logarithm (bottom row) given four explanatory variables. On the right, histograms of the response variables in these models. Only respondents with non-zero outpatient expenses are included.



correspond to multiplicative changes on the original scale. The same idea is used when the response is log-transformed. Suppose that β is the coefficient of an explanatory variable X in a model for $\log(Y)$. Thus a one-unit increase in X, controlling for other explanatory variables, is expected to lead to a β-unit change in $\log(Y)$. Since

$$\log(Y) + \beta = \log(Y) + \log(e^{\beta}) = \log(e^{\beta}Y)$$

this corresponds to a multiplicative change by a factor of e^{β} in Y itself[12]. More precisely, a result like this applies, as usual, to changes in the expected value μ of Y, although showing this involves some further subtleties not discussed here (see also S. 14.6 of Agresti & Finlay for some further discussion).

To illustrate the interpretation, consider the estimated model for the log of outpatient expenses shown in Table 4.12. For example, the coefficient of the dummy variable for the free-care insurance plan is here 0.309. Using the interpretation above, this means that the expected effect of being on free care rather than on one of the other insurance plan is to multiply expected outpatient expenses by $\exp(0.309) = 1.36$, i.e. to increase them by 36%, controlling for the other explanatory variables.

[12]Here e^{β}, often denoted $\exp(\beta)$ instead, is the exponential or antilog of β.

Table 4.12: Results for a linear regression model for the log of outpatient medical expenses in year 2 of HIE, given four explanatory variables measured at enrolment. Only respondents with non-zero expenses are included.

Response variable: Logarithm of outpatient medical expenses (year 2)					
Explanatory variable	Coefficient	Standard error	t	P-value	95 % Confidence interval
Constant	3.785				
Age (years)	0.006	0.002	2.368	0.018	(0.001; 0.011)
General Health Index	−0.014	0.002	−7.338	< 0.001	(−0.017; −0.010)
Log of Family Income	0.152	0.047	3.240	0.001	(0.060; 0.245)
Free health care (dummy variable)	0.309	0.059	5.243	< 0.001	(0.193; 0.424)
$\hat{\sigma} = 1.09$; $R^2 = 0.066$; $n = 1496$; $df = 1491$					

The model also includes the log of family income, with estimated coefficient 0.152. Its interpretation needs to combine the interpretation of coefficients of log-transformed explanatory variables (p. 84) with the one for log-transformed responses. Since here $\exp(0.095 \times 0.152) = 1.015$, the estimated effect of a 10%-increase in family income is a 1.5%-increase in outpatient expenses, controlling for the other explanatory variables.

On page 84 we discussed the issue that the logarithm of a variable, here $\log(Y)$, cannot be calculated when Y is 0 (or negative). In the current example we have avoided this by fitting the models only to data for respondents who had some outpatient expenses, and omitting everyone with zero expenses (about 20% of the sample). This must be made clear in the interpretation of the results, here by stating that the estimated coefficients in Table 4.12 describe the partial effects of explanatory variables on outpatient expenses only for people whose expenses are not zero. To complete the picture, we could also fit a model (of the kind described in the next chapter) for the dichotomous outcome of whether or not a person had some expenses at all. This is in fact essentially what was done by Newhouse et al. (1993) for this part of their analysis.

One type of response variable which often has a skewed sample distribution is a *count*: for example, the HIE data contain information on the annual number of visits to a doctor by each participant. Linear regression is sometimes used to model such counts, typically log-transformed and adding a small constant to deal with zero counts. This, however, is not a very good way to model count outcomes. In particular, if a large proportion of the counts are zero or very small, a linear model may seriously distort the results. In the HIE data, for example, 26% of the respondents did not visit a doctor in year 2, and 75% made at most 4 visits. A much better way to analyse such data is by using models which are specially designed for count outcomes. They are introduced in Week 9 of this course.

Finally, we will briefly mention another use of sample residuals, the detection of **outliers**. An outlier in a regression model is an individual observation which differs markedly from the rest of the data, in that its value of Y is much less well predicted by

the model than those of other observations. This would be indicated by an unusually large residual for such an observation. To decide what may be "unusually large", we can make use of the fact that, for a correctly specified model, the studentised residuals can be treated as approximately statistically independent and each having a standard normal distribution. For such variables, we would expect approximately 3 per 1000 to be greater than 3 or less than -3 purely by chance, and 6 per 100,000 to be further than 4 from 0. Thus observations with studentised residuals with absolute values of over 3 are commonly regarded as potential outliers, and ones over 4 as clear outliers.

An outlier may be a sign that the value of the response variable has been recorded inaccurately, or that it really is different from the other observations in ways not captured by the model. The treatment of a suspected outlier partly depends on which of these is more likely to be the case. Outliers which are possibly erroneous or otherwise of no major interest are often omitted from the fitting of the model (or given a dummy variable each, which amounts to the same thing), in order to avoid them possibly distorting the estimated coefficients. In other cases, outliers may be investigated further, either through further model selection (in case they are a sign that the model is misspecified) or by discussing them separately (in case it is thought they really are idiosyncratically different from the rest of the units).

4.8 Model specification and selection

The purpose of regression models is to describe the relationship between a response variable Y and one or more explanatory variables X, in a sample or population. There are multiple reasons that we might want to do this, but three common ones are:

- There are observations for which we observe X, but not Y, and we would like to produce a **prediction** of what the value of Y will be.
- We want to provide a **description** of how observed variation in Y is related to several variables X.
- We want to provide a **counter-argument** to a claim that the partial association between a particular variable X_1 and Y is actually due to another variable X_2.

It is important to know which of these is your goal, because it has consequences for how you should make decisions about which variables to include and which to exclude from your model.

4.8.1 Model specification for prediction

For prediction, one should always include as many predictive variables as possible, and as few non-predictive variables as possible. Here, the question is not *why* there is a partial association of Y with a given variable X_1, but only what the best guess of Y is given a set of X variables. In this situation, one does not want to include variables that are not significantly associated with the outcome. This is not because doing so will

lead to *biased* predictions (too high or too low), but because it will lead to *imprecise* predictions (too variable).

In the context of the regression models we consider in this course, the most common tool used for selection of explanatory variables is significance testing. In all the most common instances of selection, the null hypothesis of a test is that the coefficients of one or more explanatory variables are zero. This is tested using the standard tests described in Section 4.5. If the null hypothesis is not rejected, this is taken as justification for omitting the explanatory variables corresponding to it from the model. If the hypothesis is rejected, the variables are retained.

Implementing this idea usually leads to a sequence of tests, after each of which the current model is either retained, or variables are added to or removed from it. This process continues until we find a model in which all its explanatory variables are significant (i.e. should not be removed), while all of the variables omitted from the model would not be significant if they were included (and hence can be left out). There are two general approaches to such a sequential search of models: *forward* model selection which starts from a small model and adds variables to it one at a time until the remaining ones are no longer significant, and *backward* selection which starts from a large model and removes nonsignificant variables until all the remaining ones are significant. In practice, however, careful model selection usually requires a combination of the two (so-called *stepwise* selection), alternating between steps forwards and backwards as required.

The result of a sequential model selection process is not unique: there can be (and usually are) several non-nested models such that all variables in them are significant and none of the omitted variables would be if included. In other words, there need not be a single "final" model which the process will always identify, but the selected model may depend on the order in which variables are added and/or removed.

Various apparent inconsistencies are perfectly possible when related hypotheses are tested with slightly different tests. One example of this was seen on page 75, where a joint test of three dummy variables gave a different conclusion than a test of one of them alone. In other cases, for example, a variable which is nonsignificant in one model may become significant when further variables are added to the model, or an interaction term may be significant even though neither of the main effects are significant in the model without the interaction.

Sequential model selection is an obvious example of multiple comparisons of the kind discussed on page 75. This means that the overall significance level of the process is not the same as that (say 5%) used for each of the individual tests.

For consistency of the tests, all the models should be fitted using the same set of observations. When there are missing observations and complete-case analysis is used to deal with them, this means that the model selection should be carried out on the set of observations with no missing values for any of the potential explanatory variables (see also the discussion on p. 68). If this causes a large number of observations to be omitted because of missing values in a clearly nonsignificant variable, it may be necessary to repeat the model selection with that variable omitted and those observations included. At the end, the final model should be refitted to all available observations

for its variables.

Many statistical software packages have "automatic model selection" procedures, which carry out an entire sequence of tests and select a model with one command. For reasons like those listed above, such procedures cannot be relied on to produce a good selection. There are more advanced regression procedures that do yield a unique answer, such as the "lasso", which are covered in more advanced courses such as MY459.

4.8.2 Model specification for description

Some regression models are meant to be "models" in the literal sense: a simple story of the approximate relationships between a set of variables. If this is the aim of a regression model, it becomes important to select variables carefully so that the model is interpretable. This means limiting the number of variables in the model, so that it is as *parsimonious* as possible[13]. Models with no insignificant variables are simpler and easier to explain, and so typically this kind of regression will have few if any insignificant variables.

At the same time, when thinking about a descriptive model, it is important to think carefully about what it will mean to "hold all other variables constant". In this type of model, we typically want to be able to interpret all of the coefficients, and so it is important to think about all the issues discussed in Section 4.6, such as the choice of the set of dummy variables for categorical variables, inclusion of interaction terms, and possible nonlinear transformations of explanatory variables. "Correct selection" of explanatory variables here implies that the forms in which the variables are used in the model should be sensible given the particular quantities that are being modeled.

4.8.3 Model specification for counter-argument

Research articles sometimes report fitted models in which a large number of control variables are included, without model selection, (i.e. even when some of them are not statistically significant). In particular, this is often done with such standard controls as age, sex, race, education etc. Often, in these cases, the regression analysis is intended as a form of counter-argument. Here, the researcher is particularly interested in the partial effect of a variable X_1. Usually this variable is associated with Y in a simple linear regression model. Adding additional explanatory variables can then be seen as a way of convincing a skeptical reader that the partial association with X_1 is not better understood as a partial association with some other variables X_2, X_3, etc. In this context, these other explanatory variables are usually called "control variables", and we are interested in whether their inclusion reduces the magnitude and/or the significance of the relationship between X_1 and Y.

In this context, it makes sense to keep these control variables in the model, whether they are significant or not. It is important to emphasize that there is never anything

[13] It is conventional to mention at this point Ockham's [or Occam's] razor, usually attributed to William of Ockham (1285–1347/49): *Pluralitas non est ponenda sine necessitate* — roughly, "Entities are not to be multiplied beyond necessity". It may be an old idea, but it is still a good one.

wrong with including an insignificant explanatory variable X_2 in the model.[14] Even if control variables have no partial effect, they will only introduce a small cost in terms of the imprecision (variability) of the predictions. But for making a counter-argument agains the hypothesis that X_2 better explains the association of X_1 with Y, you want to keep the insignificant variable in the model because that variable is part of the argument that you are making.

4.8.4 Summary

Even when done very carefully, selection of regression models is not easy. Indeed, the whole exercise may seem absurd, if we argue that there is a infinite number of possible models, essentially all of which are guaranteed to be wrong. This, however, is not quite the right way to view the task. Even if all models are wrong, some are more fit to purpose than others. Careful model selection involves thinking about the role of the regression model in the social scientific argument that is being made.

4.9 Other topics on multiple regression

4.9.1 Multicollinearity of explanatory variables

As discussed above, key differences between simple and multiple linear regression models emerge when the multiple explanatory variables are correlated. These correlations are not a problem, because the model is designed to allow and adjust for them (indeed, that is what it is for: there would be no need for multiple regression if all Xs were uncorrelated). They do, however, become a complication of a kind when they are very high. This case, where two or more of the explanatory variables are very highly (linearly) associated, is known as (approximate) **multicollinearity**.

Consider the situation depicted in plot (3) of Figure 4.13. Here two explanatory variables are strongly correlated. As such, they are in effect two very nearly identical copies of the same information and thus have nearly the same association with Y. This is indicated by the large area of overlap between them. If we now fit a simple model with either X_1 or X_2 as the only explanatory variable, both the overlap and the small unique contribution of the variable are wholly assigned to that variable. The complications due to multicollinearity emerge when we try to include *both* variables in the model. Loosely speaking, the problem is how to divide the shared contribution between them. The estimated model often fails to achieve this with any confidence, resulting in large standard errors and low precision for the coefficients of both variables.

As a rather silly example of approximate multicollinearity, consider the models for GHI shown in Table 4.13. Here two measurements of annual family income are included, one for the year before enrolment and one for the year before that. These are unsurprisingly strongly related, with a correlation $r = 0.887$. When either one is included in the model

[14] You may have heard otherwise in previous courses, but this is just a widespread and mistaken idea about model specification.

Table 4.13: Estimated coefficients for linear regression models for General Health Index at enrolment in the HIE data, given family income (in $1000) in the two years before enrolment. The numbers in square brackets are estimated standard errors and those in parentheses P-values for t-tests of the coefficients.

	Model				
Variable	(1)	(2)	(3)	(4)	(5)
Income 1 year before	0.274	—	0.170	—	—
	[0.067]		[0.146]		
	(< 0.001)		(0.245)		
Income 2 years before	—	0.254	0.111	—	—
		[0.064]	[0.138]		
		(< 0.001)	(0.421)		
Average of incomes	—	—	—	0.281	0.279
1 and 2 years before				[0.068]	[0.068]
				(< 0.001)	(< 0.001)
Difference of incomes	—	—	—	0.029	—
1 and 2 years before				[0.138]	
				(0.832)	
R^2	0.013	0.012	0.013	0.013	0.013

on its own (models 1 and 2), its effect on GHI is significant and positive. When both are included in model (3), their standard errors become much bigger and neither is now significant (the test considered here is discussed in the next section; in short, the large P-values for model (3) indicate that neither variable has an effect on GHI when controlling for the other).

One obvious way to deal with multicollinearity is to omit one of the variables causing it. Here we might simply include only one of the income variables (probably the more recent one, as this seems to make more substantive sense). Model (4) represents an alternative approach. It still includes two variables corresponding to income, but in a different way: one variable is now the mean of the incomes in the previous two years, and the other their difference. The two very similar income variables in model (3) have thus been transformed into two pieces of information which have distinct interpretations (average income and the *change* in income between the two years) and a low sample correlation ($r = -0.119$). It can be seen that only the average is significant, so we can actually remove the difference variable to end up with model (5). Crucially, however, making this decision is now easy because the two variables are not multicollinear and, in particular, the standard error and significance of average income are unaffected by the presence of the difference variable.

When approximate multicollinearity is suspected, its extent can be easily assessed. For pairs of explanatory variables, this means examining sample correlations between them; a common rule of thumb is that a correlation of 0.8 or above indicates serious multicollinearity. Groups of more than two explanatory variables may also be jointly

multicollinear. This is the case if R^2 for a model for one explanatory variable given the others is very high.

Explanatory variables are *perfectly multicollinear* if they are exactly linearly dependent. As an example, suppose that the true model for the mean of some response variable is $\mu = 10X_1$, where X_1 denotes height measured in inches. Suppose now that we absent-mindedly fit a model which includes both X_1 and the same height (denoted X_2) expressed in centimetres. Here X_1 and X_2 are exactly related by the formula $X_2 = 2.54X_1$. Trying to include both of them in the model for μ, we observe that the same model can be expressed in many different ways, e.g.

$$\begin{aligned} \mu &= 10X_1 = 10X_1 + 0X_2 \\ &= 5X_1 + 1.97X_2 = 0X_1 + 3.94X_2 = -10X_1 + 7.87X_2 \end{aligned}$$

and so on. There are in fact infinitely many equivalent ways of expressing μ as a linear combination of X_1 and X_2, i.e. of dividing the effect of height on μ between height in inches and height in centimetres. This is obviously an extreme version of approximate multicollinearity: instead of empirical difficulty with assigning separate estimated contributions to variables with any precision, here the question is one of complete impossibility of finding a unique way of doing so. In statistical terms, separate coefficients for X_1 and X_2 in the above example are formally *unidentified*.

As in this example, perfect multicollinearity is usually the result of accidentally including variables which are deterministically linearly related. Once spotted, it can be removed simply by omitting some of these variables from the model[15]. If model with perfectly multicollinear explanatory variables is fitted with statistical software, this either issues an error message or drops (with or without warning) enough variables to remove the problem.

4.9.2 What is the population?

In Section 4.5 we described statistical inference for the regression coefficients, i.e. methods for drawing conclusions about population coefficients β_j based only on their sample estimates $\hat{\beta}_j$. What is meant by "population" and "sample" in this context is a question which often causes some confusion, so it is is worth reviewing briefly here. This discussion builds on the general one in Section 2.1.

As a basis for the discussion, suppose that, for a particular group of units, a linear model of the kind defined in Section 4.3.2 holds for a response variable Y given a set of explanatory variables. In the context of the HIE data, for example, this might be a model for some health outcome, in the group of all U.S. adults below the age of 62. Such a model is here referred to as the "true model", even though it can (bearing in mind the discussion in Sections 3.2-3.3) of course ever be only approximately correct. The group of units for which the model applies is a population in the intuitive sense of the word, although for statistical inference for the model the "population" will in fact

[15] One manifestation of perfect multicollinearity where the effects of all of the contributing variables are of separate interest is estimation of so-called age-period-cohort models in demography and elsewhere. In this context, other approaches than simply dropping one of the three have also been considered, but with limited success.

be defined slightly differently below. The parameters of interest for which inference is required are the values of the regression coefficients in this true model.

The most straightforward case is one where the units in the data are a probability sample from the group for which the true model holds. In statistical terms, we then have a representative sample from the *joint distribution* of the response variable and the explanatory variables in that group. Inference for linear regression is then clearly valid and conceptually unproblematic (bearing in mind, however, the comments in the last paragraph of this section).

More often than not, however, we do not have a sample of this kind. For example, the HIE data can be regarded as reasonably representative of the populations (below age 62) of the six locations where the study was carried out, but not of the general U.S. population. Furthermore, inference for estimated linear models is very often used with data which are quite obviously not probability samples from the joint distribution of the variables in *any* well-defined finite population. For example, the following types of analyses are common:

- The values of some or all of the explanatory variables may be controlled and assigned by the researchers, as happens in randomized experiments.
- The units of analysis may form the entire group of interest, as when we analyse the aggregate school performance in every U.S. state (c.f. the example in Section 4.1) or the voting behaviour of every member of a parliament.

Fortunately, valid and meaningful inference is, at least in many cases and to the philosophical satisfaction of most people, possible also in such situations. To see why, recall the definition of the linear model on page 53. First, note that the only distributions that this refers to are *conditional* distributions of Y, conditional on values of $X_1, X_2, \ldots, X_k$. There is no reference to any distribution of the explanatory variables themselves, which thus need not be a sample from any meaningful population. In particular, their distributions in the sample do not need to match the distributions in the group of interest: for example, inference for models for the HIE data could be perfectly valid even if it was known that some demographic groups were overrepresented in it. In fact, the "distribution of the Xs in the group of interest" does not even need to be a meaningful concept, as happens when the values of these variables are not inherent characteristics of the units but are assigned by the researchers.

These ideas might suggest that the sampling should also be conditional, i.e. that we should collect separate probability samples of Y at selected values of the Xs. Although sometimes done, this again is not and cannot be the practice in all studies. Instead, we need to broaden the view further to allow for a *conceptual* population of potential realisations, of the kind discussed in Section 2.1. In this spirit, the population which inference for linear regression actually refers to can be regarded as a set of conceptual populations of possible realisations of Y given the explanatory variables, one at each possible combination of the values of $X_1, X_2, \ldots, X_k$. In the example considered earlier in this chapter this would mean the populations of possible realisations of the General Health Index among people with each combination of age, education and income.

Far from a single finite population, we have thus ended up with a possibly infinite

number of infinite conceptual populations. While this makes statistical inference for regression modelling an arguably meaningful exercise for many types of studies, it is obviously a somewhat less than straightforward construction to grasp. For example, what does it then mean for the data to be a representative sample from these populations, and how can we tell whether this is the case? To obtain some insight into this, we can consider the second formulation of the model in Section 4.3.2, i.e. equation (4.2) and the list of assumptions below it. Here all the key assumptions of the model are stated as ones about the random error terms (residuals) ϵ_i, which are assumed to have a zero mean and constant variance, and to be uncorrelated with the explanatory variables. In short, the residuals are required to have the characteristics of pure random *noise*, and our observations of (estimates of) them should be samples of such noise.

One (partial and informal) way of describing "sampling noise" is that the errors ϵ around the regression surface (4.1) at any value of the Xs should also be unrelated to all other possible explanatory variables which are not included in the model. If this is not the case, some such additional variables should have been included in the first place. The model for μ is then *misspecified*, i.e. we are fitting a model which differs from the true model in important ways. Conversely, for the true model the assumptions are satisfied by definition. This suggests that questions of "representativeness" and "generalisability" in the context of regression models can to a large extent be reformulated as questions of model adequacy: the more correctly specified the model is, especially in including an appropriate set of explanatory variables, the more likely it is that both the estimates of the regression coefficients and inference for them are valid and meaningful[16].

Even though the definition of samples and populations thus ended up with little reference to values of the explanatory variables or to concrete units of analysis, these are still central to a broader notion of the generalisability of the results. Most models are adequate only in a limited set of conditions, and should not be generalised beyond their reach. This means, first, that predictions and conclusions should not be *extrapolated* to beyond the range of the explanatory variables in the observed data, in the sense discussed in Section 4.4.2. Second, the group of units for which the true model may be thought to hold is usually limited. For example, the HIE researchers can claim with some confidence that the models and parameter estimates they present are reasonably valid for people in a particular social setting of late 20th-century United States, but they might not want to argue that the same is true for other places, times and situations. How far the generalisability of a model extends is usually a matter for a careful and largely non-statistical discussion. Ultimately it can only be really resolved empirically, through replication of studies with data from a range of populations.

Finally, let us return to the first case discussed in this section, the one where the data *are* a probability sample from a meaningful finite population. For example, this would be the case with data from one of the social surveys of the general UK population. Data from such surveys usually includes *survey weights*, which reflect the dependence of sampling and nonresponse probabilies on some background characteristics of the respondents. These weights need to be included in the analysis to get valid estimates of population characteristics such as means and proportions of individual variables.

[16]Here "adequacy" encompasses also the assumptions of constant σ^2 and normally distributed residuals, which should and can be checked too. However, the adequacy of the specification of μ is the most important prerequisite for meaningful conclusions about the regression coefficients.

For example, suppose that a national survey has deliberately oversampled people in inner London, in order to get a particularly clear picture of their conditions. If we then want estimates of average levels in the whole population for variables which are correlated with living in inner London (such as income, political attitude or ethnic group), it will be essential to incorporate the survey weights in the analysis, in order to avoid biasing estimates towards the oversampled groups.

How about regression in such cases — if survey weights are available, should they be used in the fitting of regression models? The very short answer is "it depends", and the short answer is "typically not, assuming the model is reasonably adequate". As defined above, the population distribution which regression inference refers to is not really the joint distribution of the explanatory and response variables in a finite population, so using weights which relate specifically to that population is not essential. Survey weights are sometimes used in regression analysis, as a kind of partial insurance against model misspecification, but this practice is perhaps not to be recommended for routine use. If the model is indeed seriously misspecified, even the weighted estimates cannot really be interpreted as properties of some general (perhaps causal) relationships between the variables. Instead, they are best regarded as essentially descriptive statistics for certain partial associations in the finite population.

4.9.3 Decompositions of sums of squares

The basic decomposition discussed on page 60 shows how the total sum of squares TSS, which measures the overall sample variation in Y, can be decomposed into a part (SSM) explained by a regression model and a part (SSE) left unexplained by it. Applying this idea repeatedly to a sequence of models where explanatory variables are added one at a time, we can further decompose the explained part (and thus also R^2) into contributions from different explanatory variables. Ambiguities arising in this exercise also highlight an important feature of multiple regression models where the explanatory variables are correlated with each other.

Table 4.14 shows estimates for some models for GHI. Consider first model (1), which has age as its only explanatory variable. The sums of squares for this are $SSM = 4566$ and $SSE = 380337$ (and thus $R^2 = 0.012$), which add up to $TSS = 384903$. Suppose now that we add education to this model, i.e. move to model (2). For this, $SSM = 19500$ and $SSE = 365403$. These sum to the same TSS as before, but in a different way. What has happened is that 14934 of the total sum of squares has moved from SSE to SSM, from the unexplained to the explained. In short, adding education to the model explains that much of the sum of squares left unexplained by the model which includes age alone. The R^2 statistic increases correspondingly, from 0.012 to 0.051.

We could continue by including further explanatory variables. For example, adding income to model (2) to give model (4) explains a further 4026 of the total sum of squares and raises SSM to 23525 and R^2 to 0.061. Furthermore, the model sums of squares for this sequence of models [(1)–(2)–(4)] provide a breakdown of the contributions of the three explanatory variables into the variation explained by model (4), as shown in Table 4.15. Such decompositions are easily produced by statistical software packages.

Table 4.14: Estimated coefficients for some linear regression models for General Health Index at enrolment in the HIE data ($n = 1699$). The numbers in parentheses are P-values for t-tests of the coefficients.

	Model					
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Age	−0.138	−0.089	−0.184	−0.128	−0.142	—
	(< 0.001)	(0.004)	(< 0.001)	(< 0.001)	(< 0.001)	
Education	—	1.157	—	0.990	0.981	1.117
		(< 0.001)		(< 0.001)	(< 0.001)	(< 0.001)
Income	—	—	0.391	0.275	0.277	0.219
			(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
Experience	—	—	—	—	0.002	-0.007
					(0.563)	(0.045)
(Constant)	74.777	58.801	72.383	59.417	59.723	54.666
R^2	0.012	0.051	0.035	0.061	0.061	0.054

Table 4.15: Sequential decomposition of the total sum of squares for model (4) in Table 4.14, with model sum of squares decomposed into contributions of the explanatory variables (when included in the order shown in the table).

Age	4566
Education	14934
Income	4026
Model sum of squares	23525
Error sum of squares	361378
Total sum of squares	384903

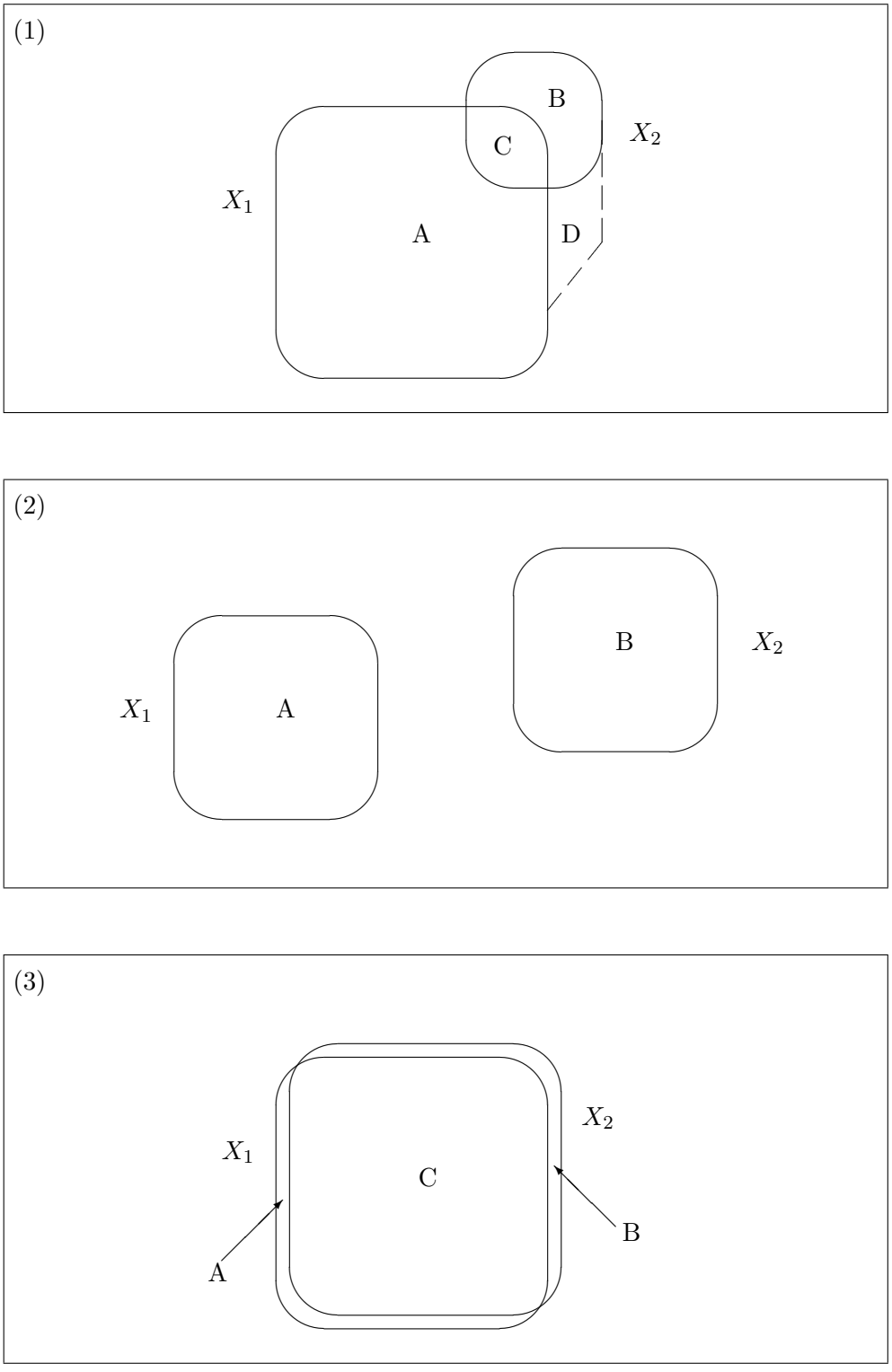
A decomposition like this is, however, not quite as useful as it might at first appear. The main reason is that it is not unique: the model sum of squares can usually be divided between the explanatory variables in several different ways, depending on the *order* in which the variables were included. This idea is illustrated graphically in Figure 4.13. Consider first plot (1). Here the rectangle represents the total sum of squares, i.e. the total sample variation in Y which the models aims to explain. The two ovals represent the portions of this explained by two explanatory variables X_1 and X_2 individually. Thus if X_1 was included in the model alone, it would explain the proportion corresponding to the combined areas of A and C, while the sum of squares explained by X_2 alone would be the sum of B and C.

The fact that the two ovals in plot (1) overlap is used to indicate that there is an association (correlation) between X_1 and X_2. In terms of sums of squares, this correlation implies that the variation explained when X_1 and X_2 are both in the model is not simply the sum of the amounts each of them explains when included alone. Somewhat loosely speaking, some of what either of them explains alone is common to both variables (area C), while some variation (are D) is explained *only* if both variables are in the model. In sequential sums of squares, the variable entered first is credited for its own unique contribution (e.g. A for X_1) and C, and the one entered second for its unique contribution (e.g. B for X_2) and D. The total explained by X_1 and X_2 together ($A + B + C + D$) may be smaller than the sum of the single contributions ($A + C$ plus $B + C$) or larger than it (i.e. D may be smaller or larger than C, in the informal notation of this graph). For example, for the models in Table 4.14 the former is the case for age and education and the latter for age and income.

The details of the sequential sums of squares in particular cases, and the net effect of including a set of the explanatory variables, depend on the intricacies of the associations between the explanatory variables. The reasons for the observed patterns are usually not easy (or necessary) to examine in detail. What is more important is the general principle: when explanatory variables are correlated with each other, results for a multiple regression model cannot be predicted from those of simple models given each of the explanatory variables alone. Instead, the multiple model has to be actually fitted to see what happens. The same, of course, applies also to the values of the regression coefficients (c.f. the discussion in Section 4.1). In Table 4.14, for example, we observe that the magnitude of the coefficient of age in the simple model (1) decreases when we control for education (model 2) but increases when controlling for income (model 3).

The only case where simple and multiple regression models give identical results is when the explanatory variables are uncorrelated. This is represented by plot (2) of Figure 4.13, where the lack of orrelation is indicated by no overlap area (C) between X_1 and X_2 (in this case there will be no joint contribution D either). The parts of the total sum of squares explained by each variable (A and B in the plot) are then uniquely defined and the same whether or not the other variable is included in the model; in a sequential decomposition of the sum squares these contributions are also the same irrespective of the order in which the variables are included. Similarly, the estimated regression coefficient of each variable in a simple linear model does not then change when we control for the other variable. These results are not limited to the case of two variables, but hold for any explanatory variable which is uncorrelated with *all* other explanatory variables in a model (even if these other variables are correlated with each other, so that similar results do not hold for them).

Figure 4.13: Symbolic representations of three examples of the decomposition of R^2 between two explanatory variables. See the text for explanation.



Because of these convenient features, randomized experiments involving two or more experimental treatments are usually designed in such a way that the treatment variables are uncorrelated. In observational studies, however, this is not possible, because we cannot control the values of the explanatory variables. It is then inevitable that many of them will be correlated, so fitting multiple regression models is essential for these correlations to be properly controlled for.

As a final observation on this theme, note that adding new explanatory variables to the model will always increase (or at least not decrease) the model sum of squares and R^2. It is indeed often possible (by including a dummy variable for every observation, using terminology introduced in Section 4.6.1 below) to obtain a model with $R^2 = 1$, i.e. one which exactly reproduces the observed data. This model, however, *is* the observed data, so it is of no use for parsimonious explanation of Y or for prediction of its values in future data. Nor do we normally pursue ever higher values of R^2 in less extreme ways, by including as many explanatory variables as are available. Instead, the aim is to include all of the explanatory variables which are necessary but no others. How this difficult goal might be achieved has been discussed in Sections 4.5 and 4.8.

4.9.4 ANOVA and ANCOVA

The terms *Analysis of Variance* (ANOVA) and *Analysis of Covariance* (ANCOVA) are among the most venerable and widely used in statistics, especially in the context of randomized experiments. In essence, they refer to methods of analysing the effects on a continuous response variable of categorical explanatory variables or *factors* (ANOVA), or of both categorical and continuous variables (ANCOVA). As such, basic ANOVA and ANCOVA models can be fitted as linear regression models, with the use of dummy variables and interactions. We will thus not discuss them in detail separately. The connection is briefly pointed out here mainly because it is not always obvious, as some of the terminology and presentation of ANOVA and ANCOVA models are rather different from those of linear models, largely for historical reasons. As a small example, a test of the effect of a single dichotomous factor is in ANOVA output typically given as an F-test rather than the equivalent t-test (c.f. p. 67). Further discussion of ANOVA and ANCOVA can be found in Chapters 12 and 13 of Agresti & Finlay.

The simplest Analysis of Variance model is "one-way" ANOVA with a single dichotomous factor. This is identical to a simple linear regression model with a dummy variable for one of the two categories of the factor as the only explanatory variable. The estimates and inference from both approaches are also identical to the methods of two-group comparison which were used as a simple example of inference in Chapter 2.

4.9.5 Output from other software

All of these are for the model for which SPSS output is shown in Figure 4.5 on page 52.

Stata:

```
      Source |       SS       df       MS              Number of obs =    1699
-------------+------------------------------           F(  3,  1695) =   36.78
       Model |  23525.1326     3  7841.71088           Prob > F      =  0.0000
    Residual |  361377.772  1695  213.202226           R-squared     =  0.0611
-------------+------------------------------           Adj R-squared =  0.0595
       Total |  384902.905  1698  226.680156           Root MSE      =  14.601

------------------------------------------------------------------------------
         ghi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1280583    .031783    -4.03   0.000    -.1903963   -.0657203
    educatio |    .990496    .143432     6.91   0.000     .7091735    1.271818
      income |   .2745179   .0631757     4.35   0.000     .1506074    .3984284
       _cons |   59.41749   2.221486    26.75   0.000     55.06034    63.77463
------------------------------------------------------------------------------
```

R:

```
Call:
lm(formula = ghi ~ age + educatio + income, data = hiedata.ch4m1)

Residuals:
    Min      1Q  Median      3Q     Max
-66.075  -8.616   2.188  10.270  33.174

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.41749    2.22149  26.747  < 2e-16 ***
age         -0.12806    0.03178  -4.029 5.85e-05 ***
educatio     0.99050    0.14343   6.906 7.04e-12 ***
income       0.27452    0.06318   4.345 1.47e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.6 on 1695 degrees of freedom
Multiple R-squared: 0.06112,    Adjusted R-squared: 0.05946
F-statistic: 36.78 on 3 and 1695 DF,  p-value: < 2.2e-16
```

Minitab:

```
The regression equation is
GHI = 59.4 - 0.128 AGE + 0.990 EDUCATIO + 0.275 INCOME


Predictor      Coef  SE Coef      T      P
Constant     59.417    2.221  26.75  0.000
AGE        -0.12806  0.03178  -4.03  0.000
EDUCATIO     0.9905   0.1434   6.91  0.000
INCOME      0.27452  0.06318   4.35  0.000


S = 14.6014   R-Sq = 6.1%   R-Sq(adj) = 5.9%


Analysis of Variance

Source            DF        SS      MS      F      P
Regression         3   23525.1  7841.7  36.78  0.000
Residual Error  1695  361377.8   213.2
Total           1698  384902.9
```

SAS:

```
Dependent Variable: GHI
Number of Observations Used        1699

Analysis of Variance
                            Sum of        Mean
Source                DF    Squares      Square   F Value    Pr > F

Model                  3      23525  7841.71065     36.78    <.0001
Error               1695     361378   213.20223
Corrected Total     1698     384903

Root MSE            14.60145    R-Square     0.0611
Dependent Mean      69.82383    Adj R-Sq     0.0595
Coeff Var           20.91184

                Parameter Estimates
                      Parameter       Standard
Variable     DF       Estimate        Error      t Value    Pr > |t|

Intercept     1       59.41749        2.22149      26.75      <.0001
AGE           1       -0.12806        0.03178      -4.03      <.0001
EDUCATIO      1        0.99050        0.14343       6.91      <.0001
INCOME        1        0.27452        0.06318       4.35      <.0001
```

Chapter 5

Binary logistic regression

The main difference between the different regression models considered on this course is the type of response variable for which each of the models is most appropriate. Linear models are used when we treat the response as a continuous, interval-level variable. In this chapter, we consider *binary logistic* (or "binary logit") models. These are designed for the case of a **binary** (**dichotomous**) response variable, i.e. one which has only *two* possible values.

One example with a binary response variable is introduced in Section 5.1. Many others are encountered in all social and other sciences. For example, any outcome which can be thought of as an answer to a Yes/No question (or True/False, Change/No change, etc.) produces a binary variable. Did a person vote in the last election? Or default on a mortgage in the past year? Does she support an increase of a carbon tax? Did a candidate win an election? Has a country ratified the Ottawa Treaty on landmines? And so on — there are clearly many interesting binary variables which we might want to consider as response variables in regression models.

As is the style of this course, the discussion in this section is fairly nontechnical and focuses on those aspects of logistic models which are most important for their use and interpretation. This leaves out a number of interesting topics which might be covered on a longer or more technically oriented course. Some of these are discussed briefly in Section 5.5. That section is included only as additional material for the interested reader, and none of it is examinable.

5.1 Example: A question on biotechnology

Data for our main example come from the Eurobarometer 58.0 survey, carried out in 2002[1]. The survey was mainly concerned with public attitudes to and engagement with modern biotechnology, including specific applications such as genetically modified food and therapeutic cloning. It also included a set of ten questions designed to assess

[1] Gaskell, G., Allum, N. and Stares, S. (2003). *Europeans and Biotechnology in 2002 (Eurobarometer 58.0)*. Report to the EC Directorate General for Research from the project "Life Sciences in European Society".

each respondent's level of knowledge of biology and genetics. These would typically be used to derive a measure of knowledge based on all ten items, for example the number of correct answers. Here, however, we will focus on the single question

"By eating a genetically modified fruit, a person's genes could also become modified".

This was answered with "True", "False" or "Don't know", of which False is the factually correct answer. The responses were coded into a binary variable with the two values "Correct" and "Incorrect or Don't know". The question was answered correctly by 46.4% of the respondents in our sample.

The Eurobarometer is a European cross-national survey. The survey considered here was carried out in 15 countries, and the most recent Eurobarometers already in 30. Here we will only use data for three countries: United Kingdom (30% of the sample members), Germany (47%) and Greece (23%). These countries were chosen because previous analyses suggested that they tend to display fairly different levels and patterns of attitudes to biotechnology. The data set includes 2185 respondents.

The response variable in this example is thus

- Y: Answer to the knowledge question, with values Correct (coded as 1) and Incorrect/Don't know (0)

A number of possible explanatory variables will be considered:

- *Country* of the respondent (UK, Germany or Greece)
- *Gender* (coded as 0 for women and 1 for men, with 52% and 48% of the respondents respectively)
- *Age* in years (mean 45, standard deviation 18, range 15–99)
- *Education*, as the age at which full-time education was completed. This is recorded in three categories: up to 19 years (73%), 20 or more (22%), or still studying and aged 15–19 (5%). Table 5.1 shows the crosstabulation of education and answer to the knowledge question. This will be used as an example for some initial illustrations.
- A measure of *technological optimism*. This is the total number of eight listed technologies (e.g. the internet and nuclear energy) which the respondent thought would "improve our way of life in the next 20 years" (mean 4.0, s.d. 2.3, range 0–8).
- *Attitude towards genetically modified (GM) food*, in three classes labelled Opposition (45%), Support (34%) and Don't know (21%). This is a derived variable based on four survey questions.

For the last two of these variables, the direction of any effects between them and an answer to the knowledge question is not obvious. We will model knowledge as a

Table 5.1: Table of education level vs. answer to the knowledge question on biotechnology in the example of Section 5.1.

Education (age at which completed)	Answer Incorrect	Correct	% Correct	Total
Up to 19 [1]	890	696	43.9	1586
20 or more [2]	223	257	53.5	480
Still studying (aged 15–19) [3]	59	60	50.4	119
Total	1172	1013	46.4	2185

response to attitudes, but the reverse might be just as plausible. The results of the models are thus best considered in a descriptive or predictive spirit: if we are told a person's attitude and other characteristics, what would we predict about that person's chances of correctly answering the knowledge question?

5.2 Definition of the model

5.2.1 The response variable

The two values of a binary response variable Y are conventionally coded as 0 and 1. In our biotechnology example,

$$Y = \begin{cases} 1 & \text{if respondent answered the question correctly} \\ 0 & \text{otherwise.} \end{cases}$$

The probability distribution appropriate for a binary variable is the **binomial distribution**[2]. The population parameter we are interested in explaining and predicting is

$$\pi = \text{Probability that } Y = 1.$$

In textbooks "probability that $Y = 1$" is usually abbreviated as $P(Y = 1)$ or $\Pr(Y = 1)$. In the biotechnology example, this is the probability that a respondent answers the question correctly. Note also that $P(Y = 0) = 1 - \pi$; in the example, this is the probability of an incorrect or "Don't know" response. Considering for the moment all respondents together (c.f. the last row of Table 5.1), the value of π in our sample is $1013/2185 = 0.464$ and that of $1 - \pi$ is $1172/2185 = 0.536 = 1 - 0.464$.

The probability π is also the expected value of Y, denoted by $\pi = \mathrm{E}(Y)$. For a binary response variable, π thus plays the same role as the expected value μ for continuous variables, and it is the parameter of interest in modelling. It is in fact the only parameter we need to consider. For a binary variable the population variance is $\sigma^2 = \pi(1 - \pi)$, which is a function of π rather than a separate parameter as before. This is the reason why the variance σ^2 is never mentioned below, and why a logistic regression model does not require any separate assumptions about it (e.g. of homoscedasticity).

[2]Specifically, the Binomial(1,π) distribution, which is also known as the *Bernoulli* distribution.

5.2.2 Explanatory variables

In the same spirit as in multiple linear regression, logistic regression involves modelling variation in π between individuals using multiple explanatory variables. These explanatory variables can again be of any type and level of measurement. For example, in the biotechnology data set two explanatory variables (age and optimism) will be treated as interval-level and continuous, and four as nominal or ordinal categorical variables, three of the latter with three categories and one (gender) with two.

Explanatory variables are used in exactly the same way in logistic as in linear regression. In particular, categorical explanatory variables will be included as sets of dummy variables, and interaction terms are specified just as before. Other issues that involve only explanatory variables, such as the need to avoid extreme multicollinearity, are also unchanged.

5.2.3 Why not use a linear model?

If we didn't know about better alternatives, we might consider using a multiple linear regression model also for a binary Y. This would mean fitting the model

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon \tag{5.1}$$

which, since for a binary variable $\pi = \mathrm{E}(Y) = \mu$, also implies

$$\pi = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k. \tag{5.2}$$

This is called the **linear probability model**. It is perfectly possible to fit such models, in that a statistical software package will not complain, and will produce parameter estimates in the usual way, if you try to do so. So why do we not do just that, and save the time and effort of learning about logistic regression?

There are two main problems with the linear probability model:

1. The error terms $\epsilon = Y - \pi$ are no longer even approximately normally distributed, and other standard assumptions about them cannot possibly be satisfied. For example, the variance of ϵ (i.e. the conditional variance of Y given the explanatory variables) is $\sigma^2 = \pi(1 - \pi)$, which depends on the explanatory variables, because π depends on them. The homoscedasticity assumption of a standard linear model thus cannot be satisfied. As a consequence, standard tests and confidence intervals from fitted linear regression will not be valid, and conclusions from them will be misleading.

2. Some expected values may be smaller than 0 or greater than 1 — i.e. poetical but meaningless values of "less than impossible" or "more than certain". For example, suppose we fit a linear probability model for the knowledge question in the biotechnology example (the results are not shown here). Consider then a Greek woman who is 90 years old, finished education before age 19, recorded a 0 for the technological optimism scale, and belongs to the attitude class of "Don't know". The fitted probability that she answers the question correctly is

$$\hat{\pi} = 0.506 - 0.192 - 0.0037 \times 90 = -0.019.$$

This is admittely not a very convincing example, because an extreme configuration of the variables is needed to produce a negative fitted probability. This is because average fitted probabilities are here close to 0.5, and ones which are very far from it are very unusual. In general, however, the problem cannot be ignored. In examples where typical values of π are closer to 0 or 1, a linear probability model can easily produce many inappropiate fitted probabilities.

5.2.4 The logistic transformation

The first of the above problems with a linear probability model could be remedied by maintaining (5.2) but changing the assumptions about ϵ in (5.1)[3]. The problem of impossible fitted values, however, is unavoidable in general. A logistic model avoids such values by modelling not π directly, but a transformation of it. We start by considering the **odds**:

$$\text{Odds} = \frac{\pi}{1-\pi},$$

i.e. the ratio of the probability that an event will occur to the probability that it will not. In the knowledge example, this is the ratio of the probability of a respondent answering vs. not answering the question correctly:

- If Odds = 1, the probability of a correct answer is *equal* to the probability of an incorrect answer.
- If Odds > 1, the probability of a correct answer is *greater* than the probability of an incorrect answer.
- If Odds < 1, the probability of a correct answer is *smaller* than the probability of an incorrect answer.

In our sample $\pi = 0.464$ and $\text{Odds} = 0.464/(1-0.464) = 0.866$. Thus the probability of a correct answer is smaller than the probability of an incorrect (or "Don't know") answer, at least for an average respondent. Later we will consider how these odds vary by the characteristics of the respondents.

Logistic regression models the logarithm of the odds, known as the **logistic transformation of π** or simply the **logit** of π. It is defined as

$$\text{logit} = \log(\text{Odds}) = \log\left(\frac{\pi}{1-\pi}\right).$$

Logits can be transformed back into probabilities using the formula

$$\pi = \frac{\exp(\text{logit})}{1+\exp(\text{logit})} = \frac{\text{Odds}}{1+\text{Odds}} \tag{5.3}$$

where $\exp(x)$ denotes the exponential of x (or "e to the power of x"). You can calculate $\exp(x)$ for any number x using statistical software such as Stata or SPSS (as explained

[3]In the terminology of Section 5.5.1, this is a generalised linear model for a Binomial Y and with an identity link.

in the computer class) or with a pocket calculator (where the button for $\exp(x)$ is often labelled e^x). Please refer to page **??** of the Appendix for a summary of the main properties of logarithms and exponentials.

The upper plot of Figure 5.1 shows the values of the probability π, obtained from the transformation (5.3), as the logit varies between -6 and $+6$. This illustrates various characteristics of the logits and probabilities:

1. If $\pi = 0.5$, then Odds $= 0.5/(1 - 0.5) = 1$ and logit $= \log(1) = 0$. Thus when the probability of an event happening (π) is equal to the probability of the event not happening ($1 - \pi$), the odds are one and the logit is zero.

2. If $\pi > 0.5$, then Odds > 1 and logit > 0. Thus the logit is positive when the probability of an event happening is greater than the probability of the event not happening.

3. If $\pi < 0.5$, then Odds < 1 and logit < 0. Thus the logit is negative when the probability of an event happening is smaller than the probability of the event not happening.

4. The logit increases as π increases, and vice versa; the logit decreases as π decreases, and vice versa.

5. The logit can take any values, but π is constrained to be between 0 and 1. As the logit takes on increasingly large values, π becomes closer and closer to 1; as the values of the logit become smaller and smaller, the values of π approach 0. This gives the probability curve based on the transformation (5.3) its characteristic S-shape, as the curve bends to avoid going below 0 or above 1.

6. Although π can be arbitrarily close to 0 and 1, it can never be exactly 0 or 1.[4]

In our example, Odds $= 0.866$ and logit $= \log(0.866) = -0.144$. Transforming back gives $\pi = \exp(-0.144)/[1 + \exp(-0.144)] = 0.866/(1 + 0.866) = 0.464$, as it should.

5.2.5 The logistic regression model

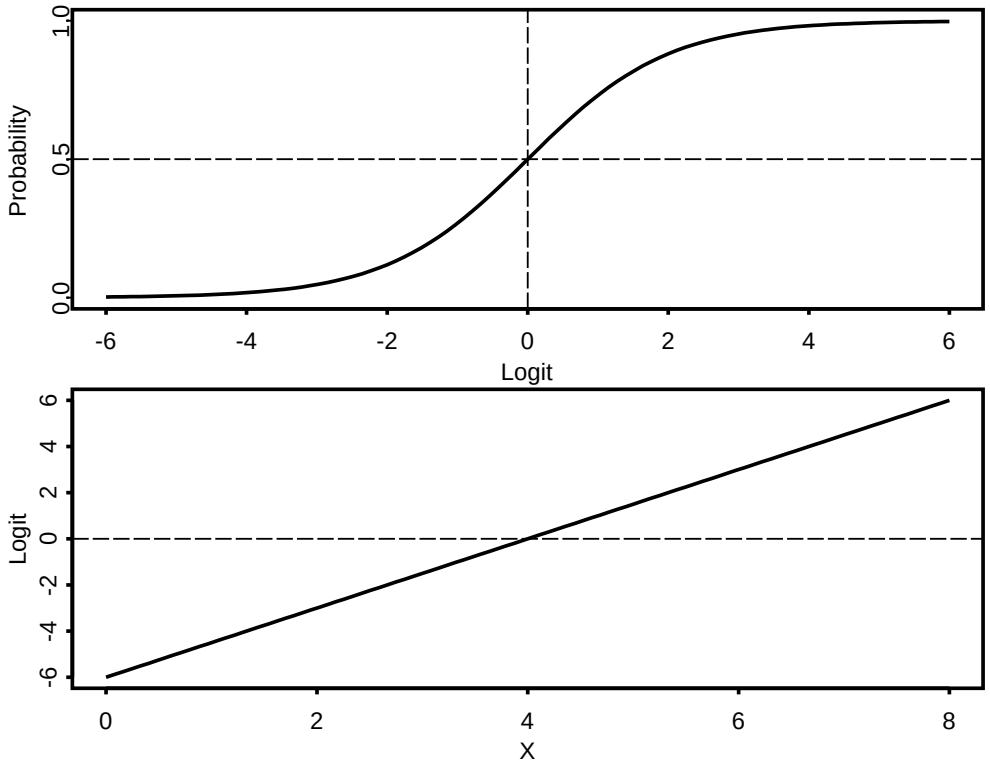
The basic idea of a logistic regression model is to model the logit of the probability π — instead of π directly — as a linear additive function of the explanatory variables. Suppose we have data on n sets of observations $(Y_i, X_{1i}, \ldots, X_{ki})$, $i = 1, \ldots, n$, where Y is a dichotomous response variable with values 0 and 1, and $X_1, \ldots, X_k$ are explanatory variables. The binary logistic regression model is defined by the following assumptions:

1. Observations Y_i are statistically independent of each other.

2. Observations Y_i are a random sample from a population where Y_i has a binomial distribution with probability parameter $\pi_i = P(Y_i = 1)$.

[4] Mathematically, 0 and 1 are the limits of π from (5.3) as the logit tends to negative and positive infinity respectively.

Figure 5.1: Plots of a probability against its logit transformation (upper plot) and the logit against an explanatory variable X (lower plot, here with logit $= -6 + 1.5X$).



3. The logit of π_i for each unit i depends on the values of the explanatory variables through the linear function

$$\text{logit}_i = \log(\text{Odds}_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \tag{5.4}$$

where α and $\beta_1, \ldots, \beta_k$ are unknown population parameters.

Note that because Y_i are dichotomous and assumed to have a binomial distribution, no further assumptions about any error terms $\epsilon_i = Y_i - \pi_i$ or their variances are required. The error terms are usually not explicitly included in formulas for logistic regression.

It is thus the logit transformation of π which is modelled as a linear expression of the explanatory variables. This is illustrated by the lower plot of Figure 5.1, which shows the logit against a single explanatory variable X. Here $\alpha = -6$ and $\beta = 1.5$, so that logit $= -6 + 1.5X$. The logit can then be transformed back into the probability of an occurrence using the formula

$$\pi_i = \frac{\exp(\text{logit}_i)}{1+\exp(\text{logit}_i)} = \frac{\exp(\alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}{1+\exp(\alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}, \tag{5.5}$$

e.g. in the example of Figure 5.1

$$\pi = \frac{\exp(-6+1.5X)}{1+\exp(-6+1.5X)}.$$

These probabilities are guaranteed to be between 0 and 1, as illustrated by the upper plot of Figure 5.1. An alternative formula, which is equivalent to (5.5) but slightly more convenient for calculations with a computer or a calculator, is

$$\pi_i = \frac{1}{1 + \exp[-(\alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})]}. \tag{5.6}$$

5.2.6 Estimating the logistic model

By fitting model (5.4) to a set of data, we obtain estimates $\hat{\alpha}$, $\hat{\beta}_1, \ldots, \hat{\beta}_k$ of the unknown parameters $\alpha, \beta_1, \ldots, \beta_k$, and the fitted (estimated) regression equation

$$\widehat{\text{logit}}_i = \log\left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i}\right) = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}. \tag{5.7}$$

Here the "ˆ" (hat) above logit (and elsewhere where it appears) again denotes the estimated (fitted, predicted) value. The fitted logits can then be transformed into fitted probabilities by

$$\hat{\pi}_i = \frac{1}{1 + \exp[-(\hat{\alpha} + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki})]}. \tag{5.8}$$

Unlike for linear models, the parameters are not estimated using the method of least squares, but something called **Maximum Likelihood Estimation** (MLE). A discussion of this method is beyond the scope of this class. A brief explanation of it is given in (unexaminable) Section 5.5.4, and more information in many textbooks.

The calculations of ML estimation are handled by computer programs like Stata and SPSS. They produce both estimates of the parameters ($\hat{\alpha}$ and $\hat{\beta}_j$) and their estimated standard errors $\hat{\text{se}}(\hat{\alpha})$ and $\hat{\text{se}}(\hat{\beta}_j)$. For our purposes it is enough to know that these maximum likelihood estimates have, at least for moderately large samples, good statistical properties, in that they are approximately unbiased, efficient, and approximately normally distributed. This allows statistical inference using statistics which are very similar to the ones we already know about.

Figure 5.2 shows parts of SPSS output for a fitted logit model in the knowledge example. The explanatory variables included here are age, sex and education. The column "B" of the table "Variables in the Equation" includes the parameter estimates $\hat{\alpha}$ (on the row labelled "Constant") and $\hat{\beta}_j$ for different explanatory variables X_j. The estimated standard errors of the coefficients are shown in the column "S.E." of the same table. Other parts of the output will be explained later.

5.3 Interpretation of the model

5.3.1 Effects of the parameters: general

The parameters α and $\beta_1, \ldots, \beta_k$ of the logistic model (5.4) are broadly analogous to the "intercept" (constant) and "slope" parameters of a linear model respectively (c.f.

Figure 5.2: SPSS output for a logistic regression model for the knowledge example

Dependent Variable Encoding

Original Value	Internal Value
Incorrect (incl. DK)	0
Correct	1

Categorical Variables Codings

		Frequency	Parameter coding	
			(1)	(2)
Education in 3 groups	up to 19 years old	1586	.000	.000
	20+ years old	480	1.000	.000
	still studying (15-19)	119	.000	1.000

Model Summary

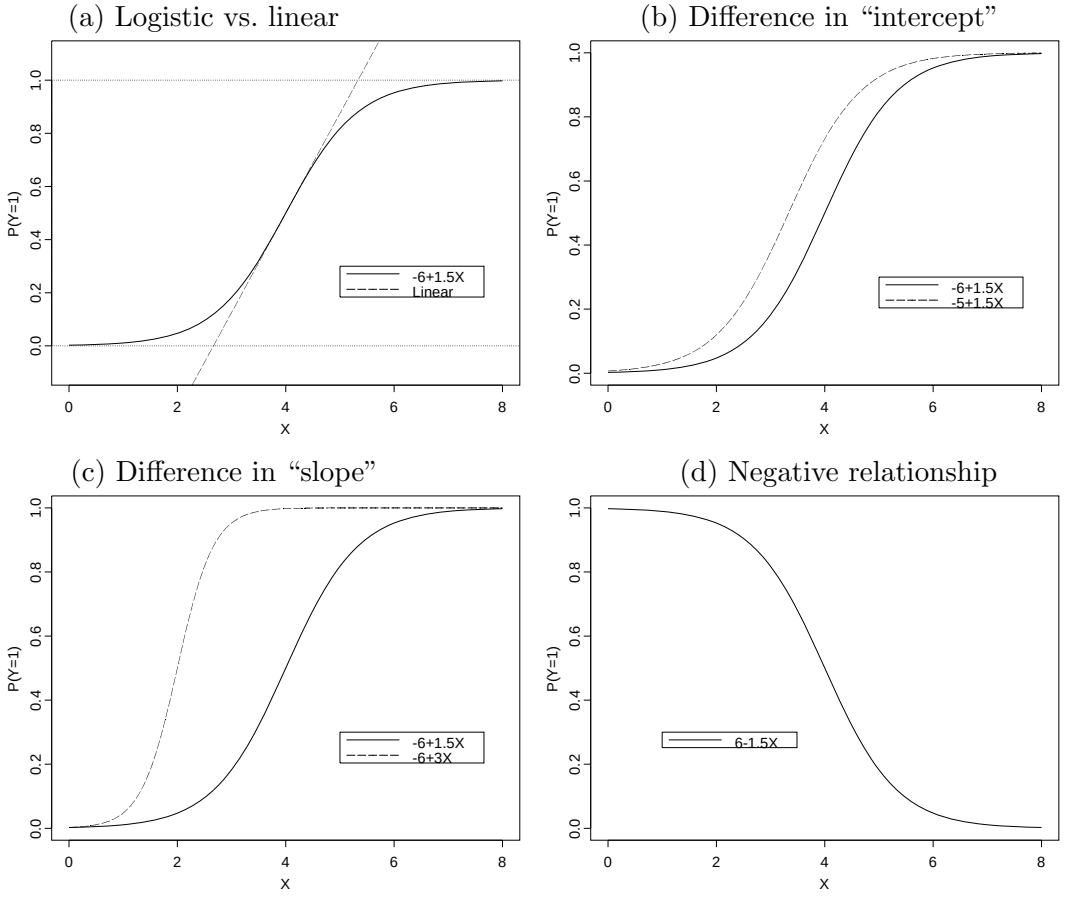
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	2947.681[a]	.031	.042

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	age	-.020	.003	53.079	1	.000	.980
	male	.048	.088	.305	1	.581	1.050
	educ			8.836	2	.012	
	educ(1)	.223	.108	4.268	1	.039	1.250
	educ(2)	-.353	.208	2.882	1	.090	.703
	Constant	.686	.146	22.037	1	.000	1.986

Figure 5.3: Illustrations of the effects of the parameters of the logistic model on the probability $\pi = P(Y = 1)$. See the text for more details.



Section 4.3.3), although their detailed interpretation is somewhat different. The roles of the parameters are illustrated by Figure 5.3, in the context of a simple logistic model with one continuous explanatory variable X.[5] First, part (a) shows how $\pi = P(Y = 1)$ depends on X in the model where the intercept is $\alpha = -6$ and the coefficient (slope) of X is $\beta = 1.5$, i.e. where $\text{logit}(\pi) = -6 + 1.5X$. This is the model which was also illustrated in Figure 5.1 through the plots of $\text{logit}(\pi)$ vs. X and π against the logit. In Figure 5.3 we see how these combine to translate the linear dependence of the logit on X into a nonlinear dependence of π on X. This also guarantees that π is always between 0 and 1, whatever the value of X. This is not the case for a linear probability model, an example of which is also shown in the plot. It can be seen that a linear model can match the probabilities from a logit model quite well for values of π around 0.5, but for some values of (continuous) X it will inevitably give values of π below 0 or over 1.

Plot (b) of the figure compares the curves of π against X for two models with the same value of β but different values of the intercept term α. It can be seen that the curves have the same shape, but that the larger value of α (here $\alpha = -5$) gives the larger

[5] I am grateful to Anders Skrondal for the idea of this plot.

value of π at any value of X. Another way of saying this is that increasing α shifts the curve to the left, so that a given value of π is achieved with a smaller value of X. The intercept is thus related to the overall level of the probabilities implied by the model. In particular, when $X = 0$, we have $\pi = \exp(\alpha)/[1 + \exp(\alpha)]$.

The regression coefficient ("slope" parameter) β determines the strength of the association between X and Y. This is illustrated by plot (c) of Figure 5.3, which compares fitted probabilities for two models with the same value of α but different values of β. Here the curve with the larger value of β is steeper, so that for the same change in X, the resulting change in π is larger than for a curve with a smaller β. Thus β describes the strength of the dependence of π on X. This dependence is positive when β is positive as in plot (c), and negative when β is negative as in plot (d) of Figure 5.3. In both cases the association is strong and the curve steep when the value of β is far from 0. The value $\beta = 0$ would correspond to no association and a flat curve where π does not change at all as a function of X.

In addition to such generalities about the roles of the parameters, we also need a quantitative interpretation of their values — especially that of the coefficients β_j. Recalling the story for linear models, the most obvious and most intuitive interpretation for the coefficients of logit models would be in terms of changes in the probability π in response to changes in an explanatory variable. Unfortunately, however, this is not equally convenient for logit as for linear models, for reasons which are illustrated by Figure 5.3. The problem is that since the dependence of π on X is nonlinear, the same change in X results in different changes in π depending on which values of X we consider. For example, in plot (a) a change of X from 0 to 2 clearly increases π much less than a change from 2 to 4, even though both are 2-unit increases. Furthermore, in a multiple logistic model the change in π in response to even exactly the same change in one X_j depends on the values at which the other explanatory variables are fixed. Thus what was for linear models the universal interpretation of "the effect of a change of 1 unit in X_j (*from any starting value*), holding other explanatory variables constant (*at any values*), is a β_j-unit change in μ" becomes for logit models a particular interpretation where the change in π is different for different starting values of X_j and choices of the fixed values of the other variables. We will consider such interpretations in Section 5.3.5, in the form of comparisons of fitted values of π. For the regression coefficients, however, we will consider a different direct interpretation. This takes the form of consideration of *odds ratios*.

5.3.2 Odds ratios

Before discussing the coefficients of the logit model, we will introduce the idea of odds ratios in the simple example of the two-way table of measures of education and knowledge shown in Table 5.1. Consider the three levels of education separately, denoting them by 1, 2 and 3 as shown in the table. The sample probabilities of a correct answer for these three groups are $\pi_1 = 0.439$, $\pi_2 = 0.535$ and $\pi_3 = 0.504$, and the corresponding odds are $\text{Odds}_1 = 0.439/(1 - 0.439) = 0.783$, $\text{Odds}_2 = 1.151$ and $\text{Odds}_3 = 1.016$.

An *odds ratio* (OR) is simply the ratio of two odds corresponding to two groups or, more generally, two values of an explanatory variable. For example, the odds ratio

between groups 2 and 1 is

$$\text{OR}_{21} = \frac{\text{Odds}_2}{\text{Odds}_1}.$$

This can also be treated as a comparison of odds given values 1 and 0 of an explanatory variable X which is a dummy variable for group 2. The odds ratio describes the association between the group and the binary response variable Y as follows:

- If $\text{OR}_{21} = 1$, the odds of $Y = 1$ are the same in both groups. This also implies that the *probability* $\pi = P(Y = 1)$ is the same in both groups. There is *no association* between X (being in group 2 rather than 1) and Y.
- If $\text{OR}_{21} > 1$, the odds (and π) are larger for group 2 than for group 1. There is a *positive association* between X and Y.
- If $\text{OR}_{21} < 1$, the odds (and π) are smaller for group 2 than for group 1. There is a *negative association* between X and Y.

In the example we have

$$\text{OR}_{21} = \frac{\text{Odds}_2}{\text{Odds}_1} = \frac{1.151}{0.783} = 1.47$$

for education levels 2 vs. 1. This is greater than 1, so the odds of a correct answer for group 2 are higher than for group 1. More specifically,

$$\text{Odds}_2 = \text{OR}_{21} \cdot \text{Odds}_1 = 1.47 \cdot \text{Odds}_1.$$

In other words, the odds of a correct answer for a member of group 2 (those who finished their education at age 20 or later) are 1.47 times the odds of a correct answer for a member of group 1 (those who finished their education at age 19 or earlier). Another way of stating this is to say that the odds are 47% higher for group 2 than for group 1 (since multiplying a number by 1.47 is the same as increasing it by $(1.47 - 1) \times 100$ per cent).

Odds ratios between other groups can be calculated similarly, as

$$\begin{aligned}
\text{OR}_{31} &= \frac{\text{Odds}_3}{\text{Odds}_1} = \frac{1.016}{0.783} = 1.30 \\
\text{OR}_{32} &= \frac{\text{Odds}_3}{\text{Odds}_2} = \frac{1.016}{1.151} = 0.88 = \frac{\text{Odds}_3}{\text{Odds}_1} \cdot \frac{\text{Odds}_1}{\text{Odds}_2} = \frac{\text{OR}_{31}}{\text{OR}_{21}} \\
\text{OR}_{12} &= \frac{\text{Odds}_1}{\text{Odds}_2} = \frac{0.783}{1.151} = 0.68 = \frac{1}{\text{OR}_{21}}
\end{aligned}$$

as well as $\text{OR}_{13} = 0.77 = 1/\text{OR}_{31}$ and $\text{OR}_{23} = 1.13 = 1/\text{OR}_{32}$. Here there are thus six possible odds ratios of comparisons of two out of the three groups. However, only two of the six are really needed to describe the associations between education and knowledge, as the other four can be calculated from them if needed. For example, the odds ratio comparing group 1 to group 2 (OR_{12}) is simply the inverse of that comparing 2 to 1 (i.e. $1/\text{OR}_{21}$). Here $\text{OR}_{12} = 0.68$, so the odds of a correct answer for a member of group 1 are 0.68 times the odds of a correct answer for a member of group 2; in other words, the odds are 32% lower for group 1 (since $(0.68 - 1) \times 100 = -32$, multiplying a number by 0.68 is the same as decreasing it by 32 per cent).

The same results are obtained when we fit a logistic model for knowledge with education level as the only explanatory variable.[6] The estimated coefficients of dummy variables of education levels 2 and 3 are $\hat{\beta}_2 = 0.388$ and $\hat{\beta}_3 = 0.263$ respectively. From these we obtain $\text{OR}_{21} = \exp(\hat{\beta}_2) = 1.47$, $\text{OR}_{31} = \exp(\hat{\beta}_3) = 1.30$, $\text{OR}_{32} = \exp(\hat{\beta}_3)/\exp(\hat{\beta}_2) = 0.88$, $\text{OR}_{12} = 1/\exp(\hat{\beta}_2) = 0.68$, $\text{OR}_{13} = 1/\exp(\hat{\beta}_3) = 0.77$, and $\text{OR}_{23} = \exp(\hat{\beta}_2)/\exp(\hat{\beta}_3) = 1.13$, the same values as above.

5.3.3 Regression coefficients as log odds ratios

The regression coefficients of a logistic model are interpreted in terms of odds ratios, controlling for other explanatory variables. To see this, consider first the definition of the model in (5.4). This states that the logit (the log of the Odds) is a linear function of the X-variables. On the scale of the logits, we can thus use exactly the same argument as for linear regression models in Section 4.3.3. Suppose that we consider two observations, labelled 1 and 2, such that both observations have the same values of explanatory variables $X_1, \ldots, X_{k-1}$ but that the value of X_k is 1 unit higher for observation 2 than for observation 1. The logits for the two observations are then

$$\begin{aligned}
\log(\text{Odds}_2) &= \alpha + \beta_1 X_1 + \cdots + \beta_{k-1} X_k + \beta_k (X_k + 1) \\
&= \alpha + \beta_1 X_1 + \cdots + \beta_{k-1} X_k + \beta_k X_k + \beta_k \\
\log(\text{Odds}_1) &= \alpha + \beta_1 X_1 + \cdots + \beta_{k-1} X_k + \beta_k X_k.
\end{aligned}$$

Subtracting the latter from the former gives us

$$\log(\text{Odds}_2) - \log(\text{Odds}_1) = \beta_k. \tag{5.9}$$

The coefficient β_k can thus be interpreted as the change in the logit as X_k increases by 1 unit, while the other explanatory variables are held unchanged. Note that this interpretation holds for any initial value of X_k and for any values of the other explanatory variables.

Differences in log odds are not easily understandable, so the interpretation is usually stated in terms of ratios of odds instead. Using the properties of exponentials and logarithms (c.f. page **??**), we know that $\log(\text{Odds}_2) - \log(\text{Odds}_1) = \log(\text{Odds}_2/\text{Odds}_1)$, and we can then restate (5.9) as

$$\exp\left[\log\left(\frac{\text{Odds}_2}{\text{Odds}_1}\right)\right] = \frac{\text{Odds}_2}{\text{Odds}_1} = \text{OR}_{21} = \exp(\beta_k).$$

which also means that

$$\text{Odds}_2 = \exp(\beta_k) \times \text{Odds}_1.$$

The exponential function thus transforms a difference of log odds to a ratio of odds (i.e. the **odds ratio**), and this is given by the exponential of the regression coefficient. In general, the coefficients of a logistic model are interpreted is as follows:

- Suppose that β_j is the regression coefficient of an explanatory variable X_j in a multiple logistic regression model. When X_j is increased by one unit (from any

[6] This uses concepts from the next section, so it is best to read this paragraph again after reading Section 5.3.3.

value), while holding all other explanatory variables constant (at any values), the odds of the outcome $Y = 1$ are multiplied by $\exp(\beta_j)$.

To illustrate what this means with a few concrete examples, consider the estimated logistic model shown in Figure 5.2. The exponentials $\exp(\hat{\beta}_j)$ of the estimated coefficients $\hat{\beta}$ are also included in the SPSS output, in the "Exp(B)" column, so we do not need to calculate them separately. The model in this output illustrates the interpretation for different types of explanatory variables. Consider first the age of the respondent, for which the estimated coefficient is $\hat{\beta}_{\text{age}} = -0.020$. This is a continuous variable measured in years. Its coefficient is interpreted as follows:

- Increasing age by 1 year, controlling for gender and education level, multiplies the odds of a correct answer by $\exp(\hat{\beta}_{\text{age}}) = \exp(-0.020) = 0.980$, i.e. decreases them by 2%.

The *sign* of the regression coefficient also has a clear interpretation. If, as here, $\hat{\beta}_j$ is negative, $\exp(\hat{\beta}_j)$ is smaller than 1, and the effect of increasing X_j is to decrease the odds (and thus also the probability) of $Y = 1$. Here this means that older respondents are, other things being equal, less likely than younger ones to answer the question correctly. If $\hat{\beta}_j$ is positive, $\exp(\hat{\beta}_j)$ is greater than 1, and the effect of increasing X_j is to increase the odds of $Y = 1$. For example, if the coefficient of age had been $\hat{\beta}_{\text{age}} = 0.02$, the effect of a one-year increase in age would have been to multiply the odds of a correct answer by $\exp(0.020) = 1.020$, i.e. to increase the odds by 2%. Finally, since $\exp(0) = 1$ and since an odds ratio of 1 means no association, $\hat{\beta}_j = 0$ implies that X_j has no effect on the odds of the response.

An interpretation which refers to a one-unit change is not the most appropriate for many continuous explanatory variables. For example, one year is a rather small increase in age. It would be more convenient to describe its effects in terms of larger increments, say five years. This can be done using the same logic as above. The simple result is that the effect of an increase of c units in X_j, while holding other explanatory variables constant, is to multiply the odds of $Y = 1$ by $\exp(c\beta_j)$. For example, the effect of a five-year increase in age in the example is to multiply the odds of a correct answer by $\exp(5 \times -0.020) = \exp(-0.100) = 0.905$, i.e. to decrease them by 9.5%.

The general interpretation of the logistic regression coefficients applies equally well when an explanatory variable is not continuous but a dummy variable. The only difference is again that the "1-unit increase" we consider is from 0 to 1, i.e. a comparison of the group for whom the dummy is 1 to the group for whom it is 0. The first illustration of this in Figure 5.2 is the coefficient of *male*, which is a dummy variable for men. Its estimated coefficient $\hat{\beta}_{\text{male}} = 0.048$ is interpreted as follows:

- Controlling for age and education level, the odds of a correct answer for a man are $\exp(\hat{\beta}_{\text{male}}) = \exp(0.048) = 1.050$ times the odds for a woman (i.e. the man's odds are 5.0% higher).

If needed, the odds in the other direction — women vs. men — are obtained simply as an inverse of this, i.e. $1/\exp(\hat{\beta}_{\text{male}}) = \exp(-\hat{\beta}_{\text{male}}) = 0.952$. Exactly the same result

can thus be stated by saying that the odds are 4.8% lower for a woman than for a man, or 5.0% higher for a man than for a woman.

As before, a categorical explanatory variable with three or more categories is included in the model as several dummy variables, one for all but one of its categories. The coefficients of each of these dummy variables are interpreted as odds ratios relative to the reference category. To illustrate this, consider the education variable *educ* in Figure 5.2. This has three categories, coded in the data set as 1, 2, and 3 as shown in Table 5.1. When fitting the model in SPSS, the programme creates appropriate dummy variables automatically once the variable is declared to be categorical. To interpret the output correctly, we first need to understand how SPSS defines these dummies and displays the results. In the output table "Variables in the equation" there are two estimated coefficients for two dummy variables corresponding to *educ*, labelled *educ(1)* and *educ(2)*. The table "Categorical Variables Codings" then explains the definition of these dummies. The column "(1)" shows the values of the variable "Educ(1)". It shows a "1.000" in the row "20+ years old", indicating that "Educ(1)" is a dummy variable for this category (i.e. category 2) of *educ*. Similarly, "Educ(2)" is shown to be a dummy variable for category 3 (aged 15-19 and still studying). This leaves category 1 (finished education at age 19 or earlier) as the reference category. The two coefficients for *educ* are thus interpreted as odds ratios between one of the other categories and the reference category:

- Controlling for age and gender, the odds of a correct answer for a person who finished education at age 20 or later are $\exp(\hat{\beta}_{\text{educ}(1)}) = \exp(0.223) = 1.250$ times the odds for a person who finished education at age 19 or earlier (i.e. they are 25.0% higher).
- Controlling for age and gender, the odds of a correct answer for a person who is aged 15–19 and still studying are $\exp(\hat{\beta}_{\text{educ}(2)}) = \exp(-0.353) = 0.703$ times the odds for a person who finished education at age 19 or earlier (i.e. they are 29.7% lower).

All other odds ratios between the education levels can be obtained from these. Odds ratios which compare level 1 to one of the other levels are simply the inverses of the ratios in the other direction (c.f. the example of the gender variable above). Odds ratios between two non-reference categories are obtained as ratios of their odds ratios with the reference level. For example, the odds ratio of level 2 vs. level 3 is here $\exp(\hat{\beta}_2)/\exp(\hat{\beta}_3) = 1.250/0.703 = 1.778$. The odds of a correct answer are thus 77.8% higher for someone who was in education to age 20 or beyond than for someone aged 15–19 who is still studying[7].

In summary, we conclude that, controlling for the other explanatory variables, an older person is less likely to answer the knowledge question correctly than a younger one, and a woman less likely to answer correctly than a man (although this association turns out not to be statistically significant). Respondents with high levels of education are most likely to answer correctly, and even those who finished their education by age 19 more likely than young people who are still studying. We will return to the interpretation

[7]Note, however, that this comparison is logically impossible while holding age constant. The calculation is included here purely as an illustration of the general idea.

of the results for these data in Section 5.4.4, after the explanatory variables have been selected more carefully.

The interpretation of the estimates in terms of proportional changes in odds may seem somewhat unintuitive, at least until we get used to it. It would be more convenient to give such interpretations in terms of changes, either proportional or absolute, in the probabilities themselves. Unfortunately, however, there is no equally simple way of describing these changes in a logistic model, for the reasons outlined in Section 5.3.1. The probabilities implied by the model are best illustrated through examples, as discussed in Section 5.3.5 below.

5.3.4 Interpretation of interaction effects

In linear regression, a (two-way) interaction between two explanatory variables means that the effect on the response variable of a variable involved in the interaction depends on the value at which the other variable in the interaction is considered fixed (c.f. Section 4.6.2). The same is true of interactions in logistic regression models, but now with the effects defined as odds ratios. In other words, an interaction in a logit model implies that the effect of an explanatory variable on the odds of a response depends on the value at which the other variable in the interaction is fixed. For instance, below we will consider two interactions in the biotechnology example:

- An interaction between age and education level. This means that the effect of education on knowledge depends on age. In other words, the odds ratios of correct answer between different education levels are higher at some ages than at others; specifically, it turns out that the difference in chances of a correct answer between more and less educated respondents is largest for young people and smallest for old ones. Alternatively, the same interaction can be presented by focusing on the effect of age on knowledge, by stating that it is different at different education levels. Specifically, it turns out that the probability of a correct answer declines with age for less educated respondents, but is virtually unrelated to age for more educated ones.

- An interaction between gender and education. This means that the effect of education on knowledge depends on gender (i.e. that the ratios of odds of a correct answer between different education levels are different for men and women) or, expressed the other way round, that the effect of gender on knowledge depends on education (i.e. that differences in chances of a correct answer between men and women vary by education level).

Before the example, let us consider a model with an interaction in general terms. For simplicity, suppose the model includes only two explanatory variables X_1 and X_2, and their interaction. As in linear regression, the interaction is included by including the product variable X_1X_2 as another explanatory variable in the model. The principle of hierarchical models dictates that the individual variables X_1 and X_2 (the main effects)

must then also be included. The logit model with an interaction is thus

$$\begin{aligned} \text{logit} &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) && (5.10)\\ &= \alpha + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 && (5.11)\\ &= \alpha + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) X_1. && (5.12) \end{aligned}$$

where (5.11) and (5.12) are the same as (5.10), but with their terms rearranged in different ways. In (5.11), the focus is on the effect of X_2 on Y. Here $\beta_2 + \beta_3 X_1$ plays the role of the coefficient of X_2. Thus a 1-unit increase in X_2 multiplies the odds of $Y = 1$ by $\exp(\beta_2 + \beta_3 X_1)$. The value of this odds ratio depends on the value at which X_1 is fixed, so there is indeed an interaction between X_1 and X_2 in this model. Alternatively, we can focus on the effect of X_1, and read from (5.12) that the coefficient of X_1 is $\beta_1 + \beta_3 X_2$, which depends on X_2. This highlights the fact that an interaction between two explanatory variables can always be interpreted in two ways, focusing on the effect of one or the other of the variables. Often one of these is substantively more natural than the other, so the results are reported in one way only.

Interaction effects are inevitably more complicated than results for models with no interactions. It is then particularly important that the results are explained carefully using a combination of odds ratios and fitted probabilities.

Table 5.2 shows estimated coefficients for two logistic models for the answer to the knowledge question. Both models include the main effects of gender, age and education, plus one interaction. Consider first model (1), which includes the interaction between age and education level. Here the third level of education (still studying and aged 15–19) is defined in such a way that it also constrains the value of the age variable. This peculiarity renders the interaction between age and this level of education nearly meaningless. We will ignore it and focus on the other two education levels. Below, all terms involving the dummy variable for the third level are thus 0.

This is an example of an interaction between a categorical variable (education) and a continuous one (age). Let "Educ2" denote the dummy variable for education level 2. The effects of the variables are captured by three coefficients, the main effect coefficients of age ($\hat{\beta}_{\text{age}} = -0.025$) and Educ2 ($\hat{\beta}_{\text{educ2}} = -0.758$), and the coefficient of their product ($\hat{\beta}_{\text{age}\times\text{educ2}} = 0.024$). These play the roles of β_1, β_2 and β_3 in (5.10)–(5.12) respectively. Focusing on them, the fitted model can be written as

$$\begin{aligned} \text{logit} &= (A) + \hat{\beta}_{\text{age}}\text{Age} + \hat{\beta}_{\text{educ2}}\text{Educ2} + \hat{\beta}_{\text{age}\times\text{educ2}}\,(\text{Age} \times \text{Educ2})\\ &= (A) - 0.025\,\text{Age} - 0.758\,\text{Educ2} + 0.024\,(\text{Age} \times \text{Educ2}) && (5.13) \end{aligned}$$

where (A) denotes the sum of the constant term $\hat{\alpha}$ and the effects of the explanatory variables — here gender — which are not involved in the interaction. The values of the latter are held fixed, and they do not affect the interpretation of the interaction.

Consider first the effect of age on the odds of a correct answer. From (5.13), the coefficient of Age is $-0.025 + 0.024\,\text{Educ2}$. This gives two different estimated effects depending on the value of the dummy variable Educ2, one at Educ2 = 0 (i.e. education level 1) and one at Educ2 = 1 (i.e. education level 2):

$$\begin{aligned} &\text{At education level 1:} && -0.025 && \text{and } \exp(-0.025) = 0.975\\ &\text{At education level 2:} && -0.025 + 0.024 && \\ & && = -0.001 && \text{and } \exp(-0.001) = 0.999. \end{aligned}$$

Table 5.2: Estimated coefficients for logistic regression models for correctly answering a knowledge question on genetics, with interactions between explanatory variables.

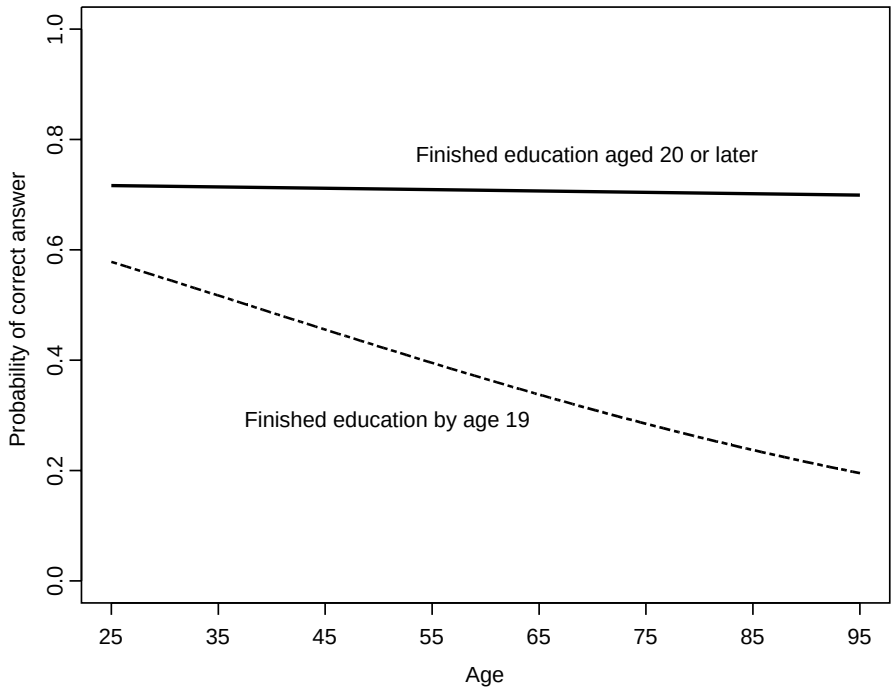
	Model	
Variable	(1)	(2)
Age	−0.025	−0.020
Sex: Male	0.027	0.083
Education		
Finished at age 20 or later [2]	−0.758	0.312
Aged 15–19 and still studying [3]	−3.842	−0.360
Age×Education		
Age×Education level 2	0.024	—
Age×Education level 3	0.195	—
Sex×Education		
Male×Education level 2	—	−0.164
Male×Education level 3	—	0.025
(Constant)	0.933	0.664

For respondents at education level 1, every one-year increase in age, controlling for gender, is thus estimated to decrease the odds of a correct answer by 2.5%. At education level 2, the effect is a decrease of 0.1%. In other words, the nature of the interaction between age and education is here such that age has a clear negative effect on knowledge for respondents with the lower level of education, but virtually no effect for those with the higher level of education. This difference is clearly illustrated by fitted probabilities obtained from the model, such as those shown in Figure 5.4 (the use of fitted probabilities is dicussed further in Section 5.3.5).

The same interaction can also be interpreted in terms of the effect of education on knowledge, and how this is modified by age. In (5.13), the coefficient of Educ2 is $-0.758 + 0.024\,\text{Age}$. This has a different value at each value of age. For example, for Age = 75 we get $-0.758 + 0.024 \cdot 75 = 1.042$ and $\exp(1.042) = 2.83$. For a respondent who is 75 years old and has a high level of education, the odds of a correct answer are 2.83 times (i.e. 183% higher than) the odds for a respondent of same age and gender but with a low level of education. At age 50, on the other hand, the coefficient is $-0.758 + 0.024 \cdot 50 = 0.442$ and the odds ratio $\exp(0.442) = 1.56$ is smaller, while at age 25 we get $-0.758 + 0.024 \cdot 25 = -0.158$ and $\exp(-0.158) = 0.85$. The fitted model thus suggests that at age 25, respondents with a higher level of education actually have a *lower* chance of answering the question correctly; more realistically, this indicates that there is little difference in knowledge between young respondents with different education levels.

Model (2) of Table 5.2 includes an interaction between gender and education. These are

Figure 5.4: Fitted probabilities of a correct answer for model (1) of Table 5.2, for women with different levels of age and education.



both categorical variables. In this case we can report the effect of each of the variables on the reponse variable at every level of the other explanatory variable. Considering the effect of gender first, the odds ratio of a correct response for men vs. women is

$\exp(0.083) = 1.087$	for respondents with education level 1
$\exp(0.083 - 0.164) = \exp(-0.081) = 0.922$	for respondents with education level 2
$\exp(0.083 + 0.025) = \exp(0.108) = 1.114$	for respondents with education level 3.

Conversely, the odds ratio for education level 2 vs. level 1 is

$\exp(0.312) = 1.366$	for women
$\exp(0.312 - 0.164) = \exp(0.148) = 1.160$	for men

and the odds ratio for education level 3 vs. level 1 is

$\exp(-0.360) = 0.698$	for women
$\exp(-0.360 + 0.025) = \exp(-0.335) = 0.715$	for men.

We can aid the interpretation of the interaction by creating a table of fitted probabilities for combinations of the two variables, as shown in Table 5.3. Here the remaining explanatory variable age is fixed at 40. Typically such a table shows all combinations of the two explanatory variables in the interaction, but here we have again omitted education level 3 because of its peculiar age-related definition.

In short, the odds ratios from model (2) and the corresponding fitted probabilities

Table 5.3: Fitted probabilities of a correct answer for model (2) of Table 5.2, separately for men and women with different levels of education. Here age is fixed at 40 years.

	Age at which education finished	
	19 or lower	20 or higher
Women	0.47	0.54
Men	0.49	0.52

indicate that, given age, the probability of answering the knowledge question correctly is less strongly dependent on education level among men than among women. Also, women are more likely to answer correctly than men when we consider people with a high level of education, but men are more likely to answer correctly among those with lower level of education and among young people who are still studying. Here the interaction thus implies that even the *direction* of the association between gender and knowledge depends on the level at which we control education. It should be noted, however, that the interaction between gender and education is not actually statistically significant here (the results of the significance test of this are not shown), so they can be entirely due to sampling variation rather than a real interaction. They were considered here purely as an illustration of how such interactions are explained and interpreted.

5.3.5 Presentation of fitted probabilities

Logistic regression coefficients give concise information about the direction and strength of the effects of the explanatory variables on the response variable. In addition, it is helpful to calculate and display some fitted probabilities from the model, in order to illustrate the magnitude of the regression effects in more familiar units than the odds. Figure 5.4 and Table 5.3 in the previous section are examples of this.

First we need to decide which probabilities to calculate. Unless the model contains only a small number of categorical explanatory variables, it is impractical or impossible to report the results at all possible values of the explanatory variables. Instead, we can present fitted probabilities which cover a representative range of the values of the explanatory variables. Usually this is done for one or two variables at a time, fixing the other explanatory variables at some appropriate values (see Section 4.4.2 for some discussion of how to choose these values).

Fitted probabilities are calculated from formula (5.8). For example, consider the entry in the lower left corner of Table 5.3. Here *Age*=40, *Male*=1 and *Educ2*=*Educ3*=0. Using the estimated coefficients of model (2) in Table 5.2, the predicted log odds of a correct answer are

$$\begin{aligned}\hat{\text{logit}} &= 0.664 - 0.020 \times 40 + 0.083 \times 1 \\ &\quad +0.312 \times 0 - 0.360 \times 0 - 0.164 \times (1 \times 0) + 0.025 \times (1 \times 0) = -0.053\end{aligned}$$

and the fitted the probability is

$$\hat{\pi} = \frac{1}{1 + \exp[-(-0.053)]} = \frac{1}{1 + \exp(0.053)} = 0.49.$$

Calculating a larger number of such probabilities is clearly easiest with the aid of a computer. The way to do this in Stata or SPSS is explained in the computer classes (some which concern linear and multinomial or ordinal logistic rather than binary logistic models, but the general procedure is the same for all of them).

5.4 Statistical inference

Recall from Section 5.3.3 that a value of 0 for a regression coefficient in a logistic model means that the corresponding explanatory variable has no effect on the odds of the response, controlling for the other explanatory variables. Statistical inference is used to examine whether or not this is the case in the population. For a single regression coefficient β_j of an explanatory variable X_j, the null hypothesis considered in significance testing is thus usually

$$H_0 : \ \beta_j = 0, \quad \text{other regression coefficients are unrestricted} \tag{5.14}$$

against the two-sided alternative hypothesis

$$H_a : \ \beta_j \neq 0, \quad \text{other regression coefficients are unrestricted.} \tag{5.15}$$

For a single coefficient, we may also want to calculate confidence intervals for β_j or the odds ratio $\exp(\beta_j)$. If we are interested in the coefficients of two or more explanatory variables at once, we can consider the null hypothesis

$$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0, \tag{5.16}$$

against the alternative hypothesis

$$H_a : \ \text{at least one of } \beta_{g+1}, \beta_{g+2}, \ldots, \beta_k \text{ is not } 0 \tag{5.17}$$

where the coefficients $\beta_1, \ldots, \beta_g$ of all the other explanatory variables are unrestricted under both hypotheses.

These are the same hypotheses, and considered for the same reasons, that we discussed for linear regression models in Section 4.5. The methods of inference are also similar in spirit to those used for linear regression, albeit different in some detail. For logistic models we will consider

- *Wald test* for single coefficients (Section 5.4.1). This is comparable to the t-test of coefficients in linear regression (c.f. Section 4.5.1).
- *Confidence intervals* for single coefficients and odds ratios (Section 5.4.2). These are very similar to the ones for linear regression (c.f. Section 4.5.2).
- *Likelihood ratio test* for one or more regression coefficients (Section 5.4.3). This is analogous to the F-test in linear regression (c.f. Section 4.5.3).

If the null hypothesis (5.14) or (5.16) is not rejected, this implies that the corresponding explanatory variable or variables have no partial effect on the response, so they may be omitted from the model. The significance tests can thus be used for model selection, in a way similar to that discussed for linear models in Section 4.8 — and with the same considerations and reservations. Examples for logistic models are given in Section 5.4.4.

5.4.1 Wald test for single regression coefficients

To test the single-coefficient null hypothesis (5.14), we can use as a test statistic the ratio of the estimated coefficient $\hat{\beta}_j$ to its standard error $\hat{\text{se}}(\hat{\beta}_j)$, i.e.

$$z = \frac{\hat{\beta}_j}{\hat{\text{se}}(\hat{\beta}_j)}. \tag{5.18}$$

This is similar to the t-test statistic (4.13) in linear regression. Here, however, z is not compared to the t-distribution but to a standard normal distribution. As a rule of thumb, if this ratio is greater than 2, we can reject H_0 at the 5% level of significance.

In SPSS output (c.f. Figure 5.1), the test statistic is in fact shown in the form

$$z^2 = \left(\frac{\hat{\beta}_k}{\hat{\text{se}}(\hat{\beta}_k)}\right)^2 = \frac{\hat{\beta}_j^2}{\hat{\text{var}}(\hat{\beta}_j)} \tag{5.19}$$

where $\hat{\text{var}}(\hat{\beta}_j) = [\hat{\text{se}}(\hat{\beta}_j)]^2$ is the estimated *variance* of $\hat{\beta}_j$. The statistic z^2 is simply the square of z. It is known as the Wald statistic, and the test based on it is a *Wald test*. The sampling distribution of z^2 under the null hypothesis (5.14) is a χ^2 distribution on 1 degree of freedom, so (2-sided) P-values for the test are obtained from this distribution. The Wald test statistic for each coefficient in the fitted model is shown in the column labelled "Wald" in the SPSS output (c.f. Figure 5.2), and the P-value in the "Sig." column.

The statistics (5.18) and (5.19) give us two versions of the same test, not two different tests. In other words, a P-value obtained from the standard normal distribution for z is identical to the P-value from the χ^2_1 distribution for z^2. More specifically, this is the case for a test with a 2-sided alternative hypothesis; a 1-sided hypothesis can be tested only with z. But since we consider only 2-sided tests here, the choice between the two test statistics is purely a matter of convenience. We report results in terms of z^2, because it is the one displayed by SPSS.

As an example, consider the model output in Figure 5.2. The estimated coefficient of the explanatory variable *Male* is $\hat{\beta} = 0.048$ and its standard error $\hat{\text{se}}(\hat{\beta}) = 0.088$, and the test statistic is

$$z^2 = \left(\frac{0.048}{0.088}\right)^2 = 0.55^2 = 0.303.$$

SPSS reports that the P-value for this, obtained from the χ^2_1 distribution, is $P = 0.581$ (the same would be obtained for $z = 0.55$ from the standard normal distribution). Since $P > 0.05$, the null hypothesis is not rejected at the 5% level of significance.

Controlling for age and education, gender thus has no effect on the odds of a correct answer to the knowledge question. In contrast, for the explanatory variable *Age* we obtain $z^2 = 53.079$ and $P < 0.001$, so the effect of age on knowledge is statistically significant, even controlling for gender and education.

An extension of the single-parameter Wald test can also be used to test the multivariate null hypothesis (5.16). This is essentially a multivariate version of z^2 above (its formula requires some matrix notation, so it is not given here). Its sampling distribution under the null hypothesis is χ^2 with the degrees of freedom equal to the number of regression coefficients set to 0 by the null hypothesis. SPSS calculates the multivariate Wald test for any categorical explanatory variables with more than two categories (and hence two or more dummy variables). An example of this is the test statistic for the effect of education in Figure 5.2. The row labelled "educ" in the table of estimated coefficients shows the Wald test for the 2-parameter hypothesis that the coefficients of *both* dummy variables for education are zero. For hypotheses like this, we usually prefer the likelihood ratio test described in Section 5.4.3; some further discussion of the multivariate Wald test is given there.

5.4.2 Confidence intervals for coefficients and odds ratios

Confidence intervals for regression coefficients in logistic models are obtained in the same way as for linear models (c.f. Section 4.5.2). The main difference is that the calculation is now always based on a standard normal distribution, rather than a t-distribution. Suppose that β_j is the coefficient of an explanatory variable X_j, and that $\hat{\beta}_j$ is its estimate and $\widehat{\text{se}}(\hat{\beta}_j)$ the estimated standard error of $\hat{\beta}_j$. A 95% confidence interval for β_j is calculated as

$$\hat{\beta}_j \pm 1.96\,\widehat{\text{se}}(\hat{\beta}_j). \tag{5.20}$$

Since the most convenient interpretation of the coefficient is based on the odds ratio $\exp(\beta_j)$, a confidence interval for this is also useful. Such an interval is obtained simply by applying the exponential transformation to the end points of the interval (5.20). In other words, a 95% confidence interval for $\exp(\beta_j)$ is the interval

$$\exp[\hat{\beta}_j - 1.96\,\widehat{\text{se}}(\hat{\beta}_j)] \quad \text{to} \quad \exp[\hat{\beta}_j + 1.96\,\widehat{\text{se}}(\hat{\beta}_j)]. \tag{5.21}$$

Note that this is asymmetric in the sense that the lower limit of (5.21) is always closer to the estimated odds ratio $\exp(\hat{\beta}_j)$ than the upper limit. The interval (5.21) can be included in Stata and SPSS output for fitted logistic models, as explained in the computer class.

The confidence intervals are related to the Wald test (5.19) of β_j in a familiar way. Thus if the null hypothesis that $\beta_j = 0$ is not rejected at the 5% level of significance, $\beta_j = 0$ will be included in the 95% confidence interval (5.20) and, equivalently, $\exp(\beta_j) = 1$ will be included in interval (5.21). Conversely, if the null hypothesis is rejected, 0 will not be included in (5.20) and 1 not in (5.21).

As an example, consider the coefficient of Age in the model output in Figure 5.2. The estimated coefficient is $\hat{\beta}_{\text{age}} = -0.020$ and its standard error $\widehat{\text{se}}(\hat{\beta}_{age}) = 0.003$. The 95% confidence interval for β_{age} is thus

$$(-0.020 - 1.96 \cdot 0.003;\ -0.020 + 1.96 \cdot 0.003) = (-0.026;\ -0.014)$$

and the 95% confidence interval for the odds ratio $\exp(\beta_{\text{age}})$ is

$$[\exp(-0.026);\ \exp(-0.014)] = (0.974;\ 0.986).$$

We are thus 95% confident that the effect of a 1-year increase in age, controlling for gender and education, is to decrease the odds of a correct answer by between 2.6% and 1.4%. The interval for the odds ratio does not include the null value of 1 (i.e. a 0% change), so it agrees with the conclusion from the Wald test of the coefficient of the age variable.

5.4.3 Likelihood ratio tests

The two general hypotheses (5.16) and (5.17) — of which the single-parameter hypotheses (5.14) and (5.15) are a special case — introduced on page 129 correspond to two logistic models: H_0 to a model which includes only explanatory variables $X_1, X_2, \ldots, X_g$, and H_a to a model which also includes $X_{g+1}, X_{g+2}, \ldots, X_k$ (as always, the subscripts are here arbitrary, so the set of variables included only in H_a can be any subset of the explanatory variables). This is an example of a pair of nested models. Two models are said to be nested if one model may be constructed from the other by either removal of variables or addition of variables, but not both, so that one model is a simpler (restricted) version of the other (full) model. This is thus exactly the same idea of nesting as the one discussed for linear models in Section 4.5.3.

A very general statistical test that can be used to compare any pair of nested models is the *likelihood ratio test* (LR test). It is a generalisation of the F-test for linear models discussed in Section 4.5.3. The LR test compares two nested models by comparing the values of the *likelihoods* of the models. The likelihood is, loosely speaking, a measure of how well a model fits the data. Some more information about it is given in Section 5.5.4.

Suppose we have fitted two nested models. Let Model 1 be the simpler model and Model 2 a more complex version of Model 1 with some extra variables added. We wish to know whether the extra variables included in Model 2 are necessary, i.e., whether the extra parameters in Model 2 are equal to zero. Model 1 thus corresponds to the null hypothesis (5.16) and Model 2 to the alternative hypothesis (5.17). We begin by fitting both models and recording the likelihood values for each of them. Let L_1 be the likelihood for Model 1 and L_2 the likelihood for Model 2. If the null hypothesis is correct — i.e. the larger Model 2 is no better than the simpler Model 1 — we would expect L_1 and L_2 to be similar, while a large difference between them — indicating that the larger model fits much better than the simpler one — would suggest that H_0 should be rejected. The likelihood ratio test is based on this simple idea. For mathematical reasons the test statistic is not defined as the difference between L_1 and L_2 themselves, but as the *likelihood ratio statistic*

$$L^2 = -2\log L_1 - (-2\log L_2). \tag{5.22}$$

This is compared with a χ^2 distribution with degrees of freedom equal to the number of extra parameters (regression coefficients) included in Model 2 but not in Model 1. In practice, an easy way to check that the difference (5.22) has been calculated the correct way round is that its value must always be non-negative.

As an example, consider the model shown in Figure 5.2 on page 117. The value of $-2 \log L$ for it is 2947.68, as shown under "-2 Log likelihood" in the "Model Summary" table of the output. This model includes age, education and gender as explanatory variables. Consider now the null hypothesis that gender has no effect on knowledge, once we control for age and education. We fit a model with age and education included as explanatory variables, but gender (i.e. the dummy variable for men) excluded. The value of $-2 \log L$ for this model is 2947.99. This model has the role of the smaller Model 1 in (5.22), and the model with gender included is Model 2. Thus the LR test statistic is

$$L^2 = -2 \log L_1 - (-2 \log L_2) = 2947.99 - 2947.68 = 0.31.$$

This is compared to a χ^2 distribution with 1 degree of freedom. The degrees of freedom are 1 because the larger model includes 1 parameter — the coefficient of the dummy variable for men — which is not used in the smaller model.

The P-value for 0.31 from the χ^2_1 distribution is 0.58. The null hypothesis is thus *not* rejected. There is no evidence that the larger model fits better than the smaller one, i.e. no evidence that gender has an effect on the probability of a correct answer, once we control for age and education. We can drop gender from the model.

Consider now the model with only age and education included. Suppose we try the even simpler model which further excludes age. The value of $-2 \log L$ for this is 3002.84, and the LR test statistic is $L^2 = 3002.84 - 2947.99 = 54.84$. The P-value for this, again from the χ^2_1 distribution, is $P < 0.001$. The null hypothesis is clearly rejected — age is associated with knowledge, and we cannot drop age from the model. Similarly, comparing the model with age and education to a model with education omitted gives $L^2 = 9.12$. The degrees of freedom for this are now 2, because leaving out education reduces the number of parameters by 2 — the coefficients of the two dummy variables for two of the three levels of education. The P-value for $L^2 = 9.12$ from the χ^2_2 distribution is $P = 0.010$. This null hypothesis too is rejected at the 5% level of significance, and we keep education in the model.

The LR test can be used for any kinds of nested models. Another example is combining categories of a categorical explanatory variable. This is achieved by omitting dummy variables for some but not all of the categories of the variable (see also the discussion of this in Section 4.6.1). As an example, consider education in Figure 5.2. Here the Wald test of the coefficient of "educ(2)", which denotes the dummy variable for education level 3 (still studying and aged 15–19), has a P-value of 0.090. This suggests that this dummy variable might be omitted from the model, while keeping the dummy for level 2. Doing so has the effect of combining level 3 with the reference level 1 (finished education before age 20). We are then comparing only two levels of education, "High" (finished education at age 20 or later) against "Low" (finished education before age 20 or aged less than 20 and still studying). Fitting models in this way with education as a two-level vs. three-level variable gives another nested pair of models (with age included in both). The LR test statistic is $L^2 = 2.90$, with 1 degree of freedom (because the larger model has 1 extra parameter, the coefficient of the dummy for education level 3), and $P = 0.089 > 0.05$. We can thus simplify the model into one which includes age and the two-level education variable.

In the last example, the Wald test (with $P = 0.090$) and the LR test ($P = 0.089$) gave

very similar results for the same hypothesis. This is typically the case, for hypotheses of both single and multiple parameters. Either test can be used in practice. The Wald test is more convenient for tests of regression coefficients of logit models, because it is automatically included in Stata and SPSS output. It is, however, very useful to know also how to use the likelihood ratio test. Its applicability is not limited to logistic models or tests of regression coefficients, but extends much wider to other models and hypotheses. The general idea of the test and the form of the test statistic (5.22) are the same in all situations.

Note that the LR test cannot be used for pairs of *non-nested* models. For example, it cannot be used to compare the model with only age included to a model with only education included. This comparison does not correspond to hypotheses of the form (5.16)–(5.17).

5.4.4 Model selection for logit models

Likelihood ratio tests (or Wald tests) can be used for model selection for logistic regression models. The general issues and caveats of this are the same as for linear models, as discussed on page 75 and in Section 4.8. We will not repeat them here. Instead, we consider an example of sequential model selection for our knowledge example, starting with all the potential explanatory variables listed in Section 5.1. For simplicity, we will not consider interactions, except for those involving country as discussed below.

This analysis involves data from three countries, UK, Germany and Greece. It is thus an example of the analysis *cross-national* surveys. Country can be included in the models as a categorical explanatory variable, here as dummy variables for Germany and Greece (thus leaving UK as the reference country). The coefficients of these describe differences between countries (as odds ratios) in chances of a correct answer, when controlling for differences in the values of the other explanatory variables in the country samples. However, the *interactions* between country and the other explanatory variables are also of particular interest here. Such an interaction indicates that the effect of an explanatory variable on the response varies between countries. For example, an interaction between country and education means that the effect of education on knowledge is different in different countries. In the extreme case, country has an interaction with every other explanatory variable, which would mean that the whole model was different (i.e. had different regression coefficients) in different countries. Here we start with this extreme model and see whether it can be simplified, using likelihood ratio tests to carry out sequential tests for model selection in the backwards (simplifying) direction. The results of these tests are summarised in Table 5.4.

We began by considering the Wald test statistics in the output for the full model (this is not shown here), and considered first for removal the interaction for which these P-values were the largest. The order of subsequent tests was also decided in a similar ad hoc manner. The first null hypothesis to be considered was that the interaction between country and gender could be omitted once everything else was included in the model. This hypothesis was not rejected, so the interaction was removed. The interaction of country and technological optimism was then considered and also removed, and so on. The end result was that *all* the interactions involving country could be deleted

Table 5.4: Summary of likelihood ratio tests (test statistic L^2) for backward selection of explanatory variables for logit models for answer to a knowledge question on biotechnology

Model	L^2	df	P-value
Model (A): all explanatory variables, all country interactions			
(A1): (A)−Country×Gender	0.56	2	0.76
(A2): (A1)−Country×Optimism	1.94	2	0.38
(A3): (A2)−Country×Attitude	2.98	4	0.56
(A4): (A3)−Country×Education	5.74	4	0.22
(A5): (A4)−Country×Age	4.50	2	0.11
Model (B)=(A5): all explanatory variables, no interactions			
(B1): (B)−Gender	0.07	1	0.80
(B2): (B1) with "high" education only	2.71	1	0.10
Model (C)=(B2): all main effects but gender, education with 2 levels			
(C1): (C)−Education	4.43	1	0.035
(C2): (C)−Country	49.63	2	< 0.001
(C3): (C)−Age	31.01	1	< 0.001
(C4): (C)−Optimism	14.41	1	< 0.001
(C5): (C)−Attitude	39.74	2	< 0.001
Final model (C): See Table 5.5			

Table 5.5: Results for the selected logit model for answer to a knowledge question on biotechnology.

Variable	Coeff. ($\hat{\beta}$)	s.e.	z^2	P-value	$\exp(\hat{\beta})$	95% CI for $\exp(\beta)$
(Constant)	−0.078	0.201				
Country (vs. UK)						
Germany	−0.150	0.105	2.035	0.154	0.861	(0.701; 1.058)
Greece	−0.868	0.132	43.51	< 0.001	0.412	(0.324; 0.543)
Age	−0.015	0.003	30.61	< 0.001	0.985	(0.980; 0.991)
Education: High	0.234	0.111	4.421	0.036	1.264	(1.016; 1.571)
Technological optimism	0.081	0.021	14.34	< 0.001	1.084	(1.040; 1.131)
Attitude to GM food (Ref: "Don't know")						
Opposition	0.426	0.123	11.99	0.001	1.530	(1.203; 1.947)
Support	0.811	0.131	38.17	< 0.001	2.250	(1.739; 2.910)

from the model. All the explanatory variables thus appear to have the same effect on knowledge in each of these countries, which greatly simplifies interpretation of the models.

We next considered whether the model could be further simplified. Wald tests at this stage again suggested the same simplifications which we applied to the smaller model considered previously, i.e. removing gender and combining categories 1 and 3 of the original education variable. LR tests confirmed that this could be done here also. After that, however, all of the remaining explanatory variables were statistically significant and could not be removed. The explanatory variables in our selected model are thus country, age, education (as "high" vs. "low") , technological optimism and attitude to genetically modified (GM) food. The estimated coefficients of this model are shown in Table 5.5. This still includes one parameter which is not significant, according to both Wald and LR tests. This is the coefficient of the dummy variable for Germany, indicating that the levels of knowledge are similar in Germany and UK. However, since there is no strong theoretical justification for combining these, and since we may well be interested in each country separately, we retain country as a variable with three categories.

The effects of age and education are broadly the same as in the smaller model considered previously. Perhaps of more interest here are the results for two of the new variables, technological optimism and attitude to GM food. For them, controlling for the other explanatory variables,

- A one-point increase in the measure of technological optimism increases the odds of a correct answer to the knowledge question by 8.4 %.
- Compared to respondents with an uncertain ("Don't know") attitude to GM food, the odds of a correct answer are 53% higher for those with a definite attitude of opposition, and 125% higher for those with a supportive attitude. This also means (since $2.25/1.53 = 1.47$) that the odds are 47% higher for the supporters than for the opponents.

In short, our prediction of the probability of a correct answer would be highest, other things being equal, for a person who professes an optimistic and supportive attitude toward science and technology.

5.5 Other topics on logistic models*

5.5.1 Logit model as a generalised linear model

We have so far not explained the subtitle of this course. *Generalised Linear Models* are a broad family of statistical regression models which includes, as its special cases, all of the models considered on this course. This framework is convenient for statisticians, because it allows general results to be derived for all members of this family at once,

*The material in this section is not required in the examination.

rather than separately for each model. Knowing about this generality is not, however, crucial when first learning about the models, so we do not make much use of it on the course.

We will first introduce the main components of a generalised linear model (GLM) by specifying a normal linear model as a GLM. This states that, for independent observations Y_i of a response variable Y,

(1) Y_i follows a normal distribution with mean μ_i and variance σ^2

(2) We consider the transformation $g(\mu_i) = \mu_i$ of the mean μ_i

(3) The model for $g(\mu_i)$ is

$$g(\mu_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \tag{5.23}$$

where $X_{1i}, \ldots, X_{ki}$ are the observed values of explanatory variables $X_1, \ldots, X_k$ for unit i, and $\alpha, \beta_1, \ldots, \beta_k$ are unknown parameters

Here (1) specifies the distribution of Y. We are mainly interested in modelling its mean (expected value) parameter μ. The variance parameter σ^2 is a *nuisance parameter* of lesser interest; it will be modelled and estimated somewhat separately from μ. Statement (2) specifies a particular transformation $g(\mu)$ of μ which will be modelled as a linear function of the explanatory variables. This transformation is known as the *link function* of the model. In the normal linear model, μ itself is modelled, and the link function $g(\mu) = \mu$ is known as the *identity link*. Finally, (3) specifies that $g(\mu)$ depends on a specific set of explanatory variables through a linear expression, with parameters (regression coefficients) which remain to be estimated from observed data.

Making changes to (3) means selecting different sets of explanatory variables (including dummy variables, nonlinear functions of explanatory variables, and interactions) to model $g(\mu)$. This is thus a matter of model selection within a particular member of the GLM family, as in sections 4.8 and 5.4.4. Changes to (1) and/or (2) lead to different models in a broader sense, i.e. to different members of the GLM class — (1) by changing the assumed distribution of Y, and (2) by changing the link function. The binary logistic model is obtained by specifying the *binomial distribution* for (1) and the *logit link* for (2):

(1) Y_i follows a Binomial (Bernoulli) distribution with mean μ_i $(= \pi_i)$

(2) We consider the transformation $g(\mu_i) = \log[\mu_i/(1 - \mu_i)] = \text{logit}(\mu_i)$

(3) The model for $g(\mu_i)$ is

$$g(\mu_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Note that here $\mu = \pi$, i.e. the expected value of Y is the probability parameter π introduced in Section 5.2.1. Because π is the only parameter of the binomial distibution, this model has no nuisance parameters. Note also that using the identity link function

would not be ideal here because it would lead to the problem of inappropriate fitted values for μ (c.f. Section 5.2.3); using the logit link prevents such values.

The multinomial logistic model considered in Chapter 6 is an example of a *multivariate* generalised linear model, where we need to specify model for a set of several parameters $\pi^{(1)}, \ldots, \pi^{(C-1)}$ at once.

5.5.2 Probit and other binary regression models

The logit link of the logistic model was motivated above mainly as a way of avoiding fitted values of the probability π below 0 or above 1. In principle, any link function which achieves this could also be used, and each one would define a different generalised linear model for a binary Y. Some such models are used fairly frequently as alternatives to logit models. The most common is obtained by choosing the link function as $g(\mu_i) = \Phi^{-1}(\mu_i)$, where Φ^{-1} denotes the inverse function of the function

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\, du. \tag{5.24}$$

The function Φ is the "cumulative distribution function" of the standard normal distribution. This means that if Z has a standard normal distibution, $P(Z \leq z) = \Phi(z)$ for any number z. The function can be used for calculations of probabilities from any normal distributions, as explained on introductory statistics courses; tables of Φ are included in most text books (e.g. appendix Table A in Agresti and Finlay 2009).

This link function is known as the *probit* link, and the resulting binary GLM as the **probit model**. It is in practice very similar to the logistic model in terms of model fit. Detecting any differences in fit between the logit and probit models is virtually impossible except in very large data sets, and substantially pointless even then. The question of whether the logit or the probit is the better or "more correct" model for a particular problem is thus essentially meaningless. The choice between the two models is more of a matter of convenience and convention. On this course we concentrate on the logit model because it is somewhat simpler in certain theoretical and computational respects (which are not at all apparent here) and because its coefficients can be relatively conveniently interpreted as log odds ratios[8]. The coefficients of the probit model can only really be interpreted in a meaningful way with the help of the alternative motivation of binary regression models which is described in Section 5.5.3.

The logit and probit models are the most commonly used binary logistic models. In some contexts we may also need the *complementary log-log model*, obtained with the link function $g(\mu) = \log[-\log(1-\mu)]$. This differs from the other models in that it is not symmetric with respect to the values of Y: if we reverse the coding of Y so that 0 becomes 1 and vice versa, the coefficients of logit and probit models simply change sign, but those of a complemetary log-log model change in a nontrivial way. The fitted probabilities from a complementary log-log model are very similar to logistic ones when π is close to 0 but rather different as it approaches 1. The reverse is true for another asymmetric binary model, the very rare *log-log* model which has $g(\mu) = -\log[-\log(\mu)]$.

[8]There is one context where this interpretation is absolutely decisive. The odds ratio is the only interesting measure of association, and logit models thus the only sensible choice, for data arising from so-called *case–control studies*. This research design is, however, not common in the social sciences.

5.5.3 Latent-variable motivation of binary models

It is sometimes helpful to think of categorical variables as coarse measurements of unobservable, *latent* continuous variables. When applied to a binary response variable Y, this idea provides an alternative motivation for binary regression models.

Suppose that Y^* is a continuous response variable which depends on a set of explanatory variables through a linear model

$$Y^* = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon. \tag{5.25}$$

Suppose further that Y^* itself cannot be directly observed. All we can detect is whether or not Y^* is positive. This is recorded in a binary variable Y such that $Y = 1$ when $Y^* > 0$ and $Y = 0$ when $Y^* \leq 0$. All we can then do is to derive a model for Y and hope that we can use it also to learn about the parameters of the model for Y^*.

It follows from (5.25) that the model for Y is

$$\begin{aligned} P(Y=1) &= P(Y^* > 0) = 1 - P(Y^* \leq 0) \\ &= 1 - P(\epsilon \leq -\alpha - \beta_1 X_1 - \cdots - \beta_k X_k) \\ &= 1 - F(-\alpha - \beta_1 X_1 - \cdots - \beta_k X_k) \\ &= F(\alpha + \beta_1 X_1 + \cdots + \beta_k X_k) \end{aligned} \tag{5.26}$$

where F denotes the cumulative distribution function of the distribution of ϵ, and the last step in (5.26) only holds when this distribution is symmetric around 0. This then implies that

$$F^{-1}[P(Y=1)] = F^{-1}(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k, \tag{5.27}$$

i.e. the model for Y is a GLM with the link function F^{-1}, and with the same coefficients as the linear model (5.25) for Y^*. In other words, even though Y^* is unobserved, we can estimate the model for it using a model for the observed binary Y.

The most obvious distribution for ϵ which is symmetric around 0 is the standard normal distribution, i.e. a normal distribution with mean 0 and variance 1 (the variance of the distribution is not separately identifiable in this context, so it is set to 1). Then F is the function Φ defined in (5.24), and (5.27) becomes $\Phi^{-1}(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$ — i.e. the probit model. A normal linear model for the latent Y^* thus implies the probit model for the observed binary Y.

The logit model is obtained if instead of the standard normal, we assume the *standard logistic distribution* for ϵ in (5.25). This is another distribution which is symmetric around 0. Its cumulative distribution function is $F(z) = \exp(z)/[1 + \exp(z)]$, and the inverse of that is $F^{-1}(z) = \log[z/(1-z)]$. When ϵ follows this distribution, (5.27) yields the logit model $\log[\pi/(1-\pi)] = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$.

The standard logistic distribution is very similar to a normal distribution with variance $\pi^2/3 \approx 3.29$ (where π now denotes the mathematical constant, not a probability). It is because of this that the logit and probit models are in practice virtually indistinguishable empirically, and lead to similar conclusions. If we fit both models to the same data, the estimated coefficients of the probit model will be roughly $\sqrt{3}/\pi \approx 0.55$ times the coefficients of the logit models.

Any binary logistic model can be, and none need be, formally represented in terms of the latent variable formulation (5.25)-(5.26). Statisticians often use it as a convenient mathematical device through which some properties of the logistic model and extensions of it can be most easily examined. Whether the latent variable Y^* is also regarded as a real, substantively meaningful concept depends on the nature of the variables. In some applications this makes a great deal of sense, in others none at all. When Y^* is given a substantive interpretation, the estimated coefficients of a binary model for Y are also interpreted as estimates of the coefficients of model (5.25) for Y^*. To do that, however, we also need to assume that the measurement model for Y is also correct, i.e. that Y really is 1 if and only if the substantively defined Y^* is greater than 0.[9] This assumption is not trivial, and not always obviously satisfied.

When substantive theory leads us to consider (5.25), it is natural to take ϵ to have a normal distribution. This leads to a probit model for Y. Indeed, the only convenient interpretation of the coefficients of the probit model is as coefficients of the linear model (5.25) for the latent Y^*. Logit models, on the other hand, have a clear interpretation on the scale of the observed binary Y, but involve a less familiar distribution on the latent scale. As a result, the use of probit models for binary response variables is most common in fields where the latent-variable motivation has a natural link to substantive theory. A prominent example is economics, where many concepts such as utility or demand can play the role of Y^*.

5.5.4 Maximum likelihood estimation

When fitting a logistic model, we are considering observations $Y_1, \ldots, Y_n$ of a binary response variable Y. The *likelihood* is then the probability, calculated under some model, of observing the values of Y that were actually observed.[10] For a given set of data, the value of the likelihood depends on the values of the parameters of the model. The broad idea of maximum likelihood estimation is to find those values of the parameters under which the likelihood of the observed data is highest.

To demonstrate what this means, let $\pi_i = P(Y_i = 1)$ denote the probability parameter of the distribution of Y_i, and let y_i denote the value of Y_i which is actually observed in our data. The two values that y_i may have are 0 and 1. The contribution of observation i to the likelihood is

$$L_i = P(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}; \tag{5.28}$$

to see why this is so, note that $x^1 = x$ and $x^0 = 1$ for any number x, and try the two cases $y_i = 0$ and $y_i = 1$ in (5.28).

When $Y_1, \ldots, Y_n$ are assumed to be statistically independent, the likelihood L for the whole data set is a product of the contributions of individual observations, i.e.

$$L = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}. \tag{5.29}$$

[9] Note that the specific value 0 for the threshold is selected for mathematical convenience, and is not itself substantively meaningful. The scale of the latent variable Y^* cannot be identified from the observed variables.

[10] More generally, when Y is a continuous variable, the likelihood is not actually a probability. But the general idea is the same.

In practice we usually work with the *log likelihood*

$$\log L = \sum_{i=1}^{n} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]. \tag{5.30}$$

Both lead to the same estimation results, but the computations are easier with $\log L$.

Suppose now that we specify a model for Y, so that the probabilities $\pi_1, \ldots, \pi_n$ are not all separate but are determined by a smaller set of common parameters. The simplest model is that the π_i are all the same, i.e. $\pi_i = \pi$. Then (5.29) becomes

$$L = \pi^m (1 - \pi)^{n-m} \tag{5.31}$$

where $m = \sum y_i$ is the number of observations for which the observed value y_i is 1.

The *maximum likelihood estimate* (MLE) $\hat{\pi}$ of π is the value of it which maximizes the value of (5.31) given the set of values $y_1, \ldots, y_n$ in our data set. Suppose, for example, that almost all of the observed y_i are 1. It then seems intuitively obvious that the MLE of π should be large (close to 1), as such a value of π would make it most likely that we observe mostly ones. Conversely, most of the observations are 0, we would expect $\hat{\pi}$ to be close to 0. The value of π which actually maximizes (5.31) is

$$\hat{\pi} = m/n$$

which is simply the proportion of observations of 1 in the sample. This is certainly an eminently sensible estimate of the population probability.

In a logistic model we do not model the probabilities π_i as all equal, but let them depend on a set of explanatory variables through the model (5.5) on page 115. The log likelihood (5.30) is then

$$\log L = \sum_{i=1}^{n} \left\{ y_i(\alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}) - \log[1 + \exp(\alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})] \right\}. \tag{5.32}$$

Here the values of y_i and all of the Xs are known numbers. The unknown parameters are α and $\beta_1, \ldots, \beta_k$, and their ML estimates are again found by finding those values which maximize the likelihood. This is done by solving a set of $k + 1$ equations, obtained by setting the partial derivatives of $\log L$ with respect to each of the parameters to zero. These equations do not have closed-form solutions, so they have to be solved numerically, using an iterative computational method. That is what SPSS and other software packages do in the moment between you pressing OK and the output appearing on the screen. If done by hand, that fraction of a second would turn into some months. The estimated standard errors of the ML estimates $\hat{\alpha}$ and $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are obtained as a byproduct of the estimating procedure. The log likelihood value for the fitted model which is reported in the output, and which was used for likelihood ratio tests in Sections 5.4.3 and 5.4.4, is the number obtained when we calculate (5.32) using the observed data *and* the ML estimates of the parameters.

The method of ML estimation is not limited to logistic regression, but can be applied very generally to all kinds of statistical models. For example, when the linear regression model is treated as a GLM for a normally distributed Y, the ML estimates of its

coefficients are the familiar least squares estimates. In (almost) all of these many contexts, ML estimates have a number of desirable statistical properties. The results which establish these properties are *asymptotic*, which means that they hold reasonably well when the sample size n is sufficiently large (and exactly when it is infinite). Under such conditions, ML estimates have the following characteristics:

- As the sample size increases, the estimates approach the true values of the parameters (in technical parlance, they are "asymptotically unbiased").
- It is not possible to find other estimates that have smaller standard errors (ML estimates are "efficient").
- The sampling distributions of the estimates are approximately ("asymptotically") normal distributions.

Given these nice things, it is not suprising that the method of maximum likelihood plays a central role in the theory and practice of estimation of statistical models.

Chapter 6

Multinomial logistic models

One of the examples of Chapter 5 concerned the predictors of how a person voted in an election. There the choices were grouped in two (Democrat or other), so binary logistic models could be used. In many applications, however, there may instead be three or more outcomes of interest. This is the case in the following example, which is also about voting:

Example: voting in a UK General Election

The data for this example come from the UK component of the 2002/3 European Social Survey[1]. We consider the $n = 1228$ respondents who voted for one of the three main parties in the 2001 General Election. The response variable Y will be the party a person voted for, recorded as 0=Labour (often abbreviated "Lab" below), 1=Conservative ("Cons") or 2=Liberal Democrat ("Lib"). The following (somewhat haphazard) set of potential explanatory variables will be considered:

- Age in years
- Education level (recorded as lower secondary or less, upper secondary, or post-secondary)
- Sex
- Legal marital status (married; separated; divorced; widowed; never married)
- Employment status (employed; self-employed; not in paid work)
- The respondent lives in a big city (yes; no)
- Total time the respondent spends watching TV on an average weekday, in hours
- Whether the respondent was born in the UK (yes; no)

[1] R. Jowell and the Central Co-ordinating Team (2003). *European Social Survey 2002/2003: Technical Report*. Centre for Comparative Social Surveys, City University. The data, which are archived and distributed by Norwegian Social Science Data Services (NSD), were obtained from the web site `www.europeansocialsurvey.org`, which also gives much further information on the survey.

Here the response variable has 3 possible values, so binary logistic models are not applicable. It is not difficult to think of many other examples in the social sciences where we might want to model outcomes with three or more categories. This chapter introduces the **multinomial logistic model**, which can be used for this purpose. The model turns out to be a straightforward extension of the binary logistic model.

The multinomial logistic model is most appropriate when the levels of a categorical response variables are regarded as *unordered*. This is the case in the voting example, where it would not be entirely natural to consider the three parties as being ordered along some single political dimension.

We may use the multinomial logistic model even when the response categories are naturally ordered. This, however, means that the ordering is simply ignored, which is often something we would not want to do. A better approach may then be to use a model which takes the ordering of the response categories explicitly into account. One model of this kind, the *ordinal logistic model*, will be described in Chapter 7.

6.1 Definition of the model

6.1.1 Motivating example

The multinomial logistic model is an obvious generalisation of the binary logistic model, so there are few conceptually new issues. The interpretation of the multinomial model is, however, somewhat more labourious, simply because the larger number of categories of the response variable leads to a corresponding increase in comparisons which need to be considered for a full interpretation. For this reason, it is useful to introduce the model first in the context of a simple, specific example.

Let us consider the association between voting and just education, using the data introduced above. For further simplicity, the two higher categories of the education variable are combined, creating a dichotomous variable called *HIGHED*. Its two values are coded as 0='Low' (lower secondary education or less) and 1='High' (upper secondary or post-secondary education).

In this simple case, the data can be presented in a two-way contingency table of vote choice by education level, as shown in Table 6.1. The upper part of the table gives the sample frequencies of different combinations of the variables. The lower part shows the proportions of vote conditional on education level, obtained by dividing the frequencies in each row by the corresponding row total. For example, the proportion of respondents with low level of education who voted for the Conservatives is $196/646 = 0.30$. This is denoted below by $\hat{P}(\text{Cons}|\text{Highed} = 0) = 0.30$, and the other row proportions are defined similarly. These proportions can be regarded as estimates of corresponding population probabilities of vote given education level. A multinomial logistic model for these data is a model for these probabilities.

In a multinomial logistic model, one of the categories of the response variable is chosen as the *reference category*, and the model is interpreted in terms of comparisons between

Table 6.1: Table of vote choice by education level in the example considered in Section 6.1.1. The upper part of the table shows the frequencies in this two-way contingency table, and the lower part shows conditional proportions of vote choice given education level.

Frequencies:

	Party (Y)			
Education (X)	Lab	Cons	Lib	Total
Low (Highed=0)	369	196	81	646
High (Highed=1)	255	190	137	582
Total	624	386	218	1228

Proportions given education level:

Education (X)	Lab	Cons	Lib	Total
Low (Highed=0)	0.57	0.30	0.13	1
High (Highed=1)	0.44	0.33	0.24	1

the other categories and the reference category. The reference level is typically chosen so as to make this interpretation as convenient as possible. Any category can be used as the reference, and all choices will give the same estimated model (this will be illustrated later). Here we will first consider Labour as the reference category.

A multinomial logistic model is basically a model, given the explanatory variables, for the log of the odds that the value of the response variable is in a particular category rather than in the reference category. For the data in Table 6.1, these odds and log odds are

- For voting Conservative rather than Labour:
  - For respondents with Low education:

$$\begin{aligned} \widehat{\text{Odds}}_L(\text{Cons}|\text{Highed}=0) &= \frac{\hat{P}(\text{Cons}|\text{Highed}=0)}{\hat{P}(\text{Lab}|\text{Highed}=0)} = \frac{0.30}{0.57} = 0.53 \\ \log[\widehat{\text{Odds}}_L(\text{Cons}|\text{Highed}=0)] &= \log(0.53) = -0.63 \end{aligned}$$

  - For respondents with High education:

$$\begin{aligned} \widehat{\text{Odds}}_L(\text{Cons}|\text{Highed}=1) &= \frac{\hat{P}(\text{Cons}|\text{Highed}=1)}{\hat{P}(\text{Lab}|\text{Highed}=1)} = \frac{0.33}{0.44} = 0.75 \\ \log[\widehat{\text{Odds}}_L(\text{Cons}|\text{Highed}=1)] &= \log(0.75) = -0.29 \end{aligned}$$

- For voting Liberal Democrat rather than Labour:
  - For respondents with Low education:

$$\begin{aligned} \widehat{\text{Odds}}_L(\text{Lib}|\text{Highed}=0) &= \frac{\hat{P}(\text{Lib}|\text{Highed}=0)}{\hat{P}(\text{Lab}|\text{Highed}=0)} = \frac{0.13}{0.57} = 0.23 \\ \log[\widehat{\text{Odds}}_L(\text{Lib}|\text{Highed}=0)] &= \log(0.23) = -1.47 \end{aligned}$$

– For respondents with High education:

$$\begin{aligned} \widehat{\text{Odds}}_L(\text{Lib}|\text{Highed}=1) &= \frac{\hat{P}(\text{Lib}|\text{Highed}=1)}{\hat{P}(\text{Lab}|\text{Highed}=1)} = \frac{0.24}{0.44} = 0.55 \\ \log[\widehat{\text{Odds}}_L(\text{Lib}|\text{Highed}=1)] &= \log(0.55) = -0.60 \end{aligned}$$

Here the subscript L indicates that Labour is the reference category, and the hats on the various quantities show that these are regarded as sample estimates of corresponding population quantities.

Associations between an explanatory variable and the response are described by ratios of odds given different values of the explanatory variable, and logs of these ratios. Here the comparison is between respondents with High (Highed=1) and Low (Highed=0) education. The sample estimates of these odds ratios (OR) and log odds ratios are

- For voting Conservative rather than Labour:

$$\begin{aligned} \widehat{\text{OR}}_L(\text{Cons}|\text{Highed}) &= \frac{\widehat{\text{Odds}}_L(\text{Cons}|\text{Highed}=1)}{\widehat{\text{Odds}}_L(\text{Cons}|\text{Highed}=0)} = \frac{0.75}{0.53} = 1.40 \\ \log[\widehat{\text{OR}}_L(\text{Cons}|\text{Highed})] &= \log(1.40) = 0.34 \end{aligned}$$

- For voting Liberal Democrat rather than Labour:

$$\begin{aligned} \widehat{\text{OR}}_L(\text{Lib}|\text{Highed}) &= \frac{\widehat{\text{Odds}}_L(\text{Lib}|\text{Highed}=1)}{\widehat{\text{Odds}}_L(\text{Lib}|\text{Highed}=0)} = \frac{0.55}{0.23} = 2.45 \\ \log[\widehat{\text{OR}}_L(\text{Lib}|\text{Highed})] &= \log(2.45) = 0.90 \end{aligned}$$

In other words, the sample odds of voting Cons rather than the reference Lab are 1.40 times higher for those with high education than for those with low education. Similarly, the sample odds of voting Lib rather than the reference Lab are 2.45 times higher for those with high education than for those with low education.

We now start to see how the interpretation of multinomial logistic models will involve a profusion of comparisons. On the one hand, we compare different values of *explanatory* variables (e.g. High vs. Low education) and, on the other, see how these affect the relative odds of one category of the *response* variable rather than another (e.g. voting Cons rather than Lab). Getting used to all these comparisons takes some practice and concentration. It is, however, unavoidable if we want to obtain a full picture of the associations between the variables. The complexity arises because the response categories are regarded as *unordered*, so we have to consider them one at a time and cannot simplify the interpretation by making use of any ordering of the categories. Because of the large number of the odds ratios, it is also desirable to be able to supplement them with other ways of interpreting the model. Using the presentation of fitted probabilities for this purpose will be discussed in Section 6.3.

A multinomial logistic model for the example is defined as

$$\begin{aligned}
\log[\text{Odds}_L(\text{Cons}|\text{Highed})] &= \log\left[\frac{P(\text{Cons}|\text{Highed})}{P(\text{Lab}|\text{Highed})}\right] \\
&= \alpha_L^{(\text{Cons})} + \beta_L^{(\text{Cons})}\text{Highed} \qquad (6.1) \\
\log[\text{Odds}_L(\text{Lib}|\text{Highed})] &= \log\left[\frac{P(\text{Lib}|\text{Highed})}{P(\text{Lab}|\text{Highed})}\right] \\
&= \alpha_L^{(\text{Lib})} + \beta_L^{(\text{Lib})}\text{Highed} \qquad (6.2)
\end{aligned}$$

where 'Highed' is a dummy variable for High education as defined above, and the αs and βs are population parameters. The superscript (Cons) indicates that a parameter refers to the model for voting Cons rather than Lab and superscript (Lib) to a model for Lib vs. Lab, and the subscript L indicates that Labour is the reference level for the response variable. In short, (6.1) and (6.2) define separate models for the log odds of each non-reference category of the response variable against the reference category.

Substituting the values 0 or 1 for 'Highed' in (6.1) and (6.2) we get the log odds

$$\begin{aligned}
\log[\text{Odds}_L(\text{Cons}|\text{Highed}=0)] &= \alpha_L^{(\text{Cons})} && (= \text{-0.63}) \\
\log[\text{Odds}_L(\text{Cons}|\text{Highed}=1)] &= \alpha_L^{(\text{Cons})} + \beta_L^{(\text{Cons})} && (= -0.29) \\
\log[\text{Odds}_L(\text{Lib}|\text{Highed}=0)] &= \alpha_L^{(\text{Lib})} && (= \text{-1.47}) \\
\log[\text{Odds}_L(\text{Lib}|\text{Highed}=1)] &= \alpha_L^{(\text{Lib})} + \beta_L^{(\text{Lib})} && (= -0.60)
\end{aligned}$$

and thus the log odds ratios and odds ratios

$$\begin{aligned}
\log[\text{OR}_L(\text{Cons}|\text{Highed})] &= \log[\text{Odds}_L(\text{Cons}|\text{Highed}=1)] - \log[\text{Odds}_L(\text{Cons}|\text{Highed}=0)] \\
&= \beta_L^{(\text{Cons})} && (= \text{0.34}) \\
\text{OR}_L(\text{Cons}|\text{Highed}) &= \exp(\beta_L^{(\text{Cons})}) && (= \text{1.40}) \\
\log[\text{OR}_L(\text{Lib}|\text{Highed})] &= \beta_L^{(\text{Lib})} && (= \text{0.90}) \\
\text{OR}_L(\text{Lib}|\text{Highed}) &= \exp(\beta_L^{(\text{Lib})}) && (= \text{2.45}).
\end{aligned}$$

Here the numbers in parentheses are the sample estimates of these quantities, obtained directly from Table 6.1 as shown above. Another way of estimating them is to fit the multinomial logistic model (6.1)–(6.2) to the data. Figure 6.1 shows SPSS output for this model. Some estimates are highlighted in the output as well as above. Comparing the estimates of matching quantities obtained in these two ways shows that they are indeed the same, apart from rounding error. In more complex cases, when there are several explanatory variables, or even just one continuous explanatory variable, there is usually no simple alternative to estimating the model using computer algorithms. Even then, however, the parameters can be interpreted as log odds ratios, as explained in general terms in the next section.

A key feature of the quantities considered above was that each of them compared probabilities of just two categories of the response variable at a time, for one non-reference category against the reference category. Each of these comparisons is similar to ones that would be considered for a binary logistic model for a dichotomous response

Figure 6.1: SPSS output for a multinomial logistic model for voting given education, fitted to the data in Table 6.1.

Parameter Estimates

Vote in 2001 General Election[a]		B	Std. Error	Wald	df	Sig.	Exp(B)
Conservative	Intercept	-.633	.088	51.240	1	.000	
	highed	.338	.130	6.739	1	.009	1.403
Liberal Democrat	Intercept	-1.516	.123	152.720	1	.000	
	highed	.895	.162	30.489	1	.000	2.447

a. The reference category is: Labour.

variable. Indeed, an easy way to understand the idea of a multinomial logistic model is to think of it roughly as a set of parallel binary logistic models, one for each comparison of a non-reference level to the reference level. This also means that if the response variable actually has only two categories, the multinomial logistic model is identical to the binary logistic model.

The choice of Labour as the reference category in our example was arbitrary, and the same quantities could have been defined with any other reference category. Furthermore, it is straightforward to convert results for one reference level into those for another. To illustrate this, suppose Conservative is used as the reference level (indicated by the subscript C in the quantities below). We can then define, for example,

$$\widehat{\text{Odds}}_C(\text{Lab}|\text{Highed}=0) = \frac{\hat{\text{P}}(\text{Lab}|\text{Highed}=0)}{\hat{\text{P}}(\text{Cons}|\text{Highed}=0)} = \frac{0.57}{0.30} = 1.90.$$

This can also be derived from the results for the reference level Lab, by observing that

$$\begin{aligned}\widehat{\text{Odds}}_C(\text{Lab}|\text{Highed}=0) &= \frac{1}{\hat{P}(\text{Cons}|\text{Highed}=0)/\hat{P}(\text{Lab}|\text{Highed}=0)} \\ &= \frac{1}{\widehat{\text{Odds}}_L(\text{Cons}|\text{Highed}=0)} = \frac{1}{0.53} = 1.90.\end{aligned}$$

Similarly, it can shown that, for example,

$$\begin{aligned}\widehat{\text{Odds}}_C(\text{Lib}|\text{Highed}=0) &= \frac{0.13}{0.30} = 0.43 \\ &= \frac{\widehat{\text{Odds}}_L(\text{Lib}|\text{Highed}=0)}{\widehat{\text{Odds}}_L(\text{Cons}|\text{Highed}=0)} = \frac{0.23}{0.53}.\end{aligned}$$

From results like these it also follows that

$$\begin{aligned}\widehat{\text{OR}}_C(\text{Lab}|\text{Highed}) &= 1/\widehat{\text{OR}}_L(\text{Cons}|\text{Highed}) = 1/1.40 = 0.71 \\ \log[\widehat{\text{OR}}_C(\text{Lab}|\text{Highed})] &= -\log[\widehat{\text{OR}}_L(\text{Cons}|\text{Highed})] = -0.34 \\ \widehat{\text{OR}}_C(\text{Lib}|\text{Highed}) &= \frac{\widehat{\text{OR}}_L(\text{Lib}|\text{Highed})}{\widehat{\text{OR}}_L(\text{Cons}|\text{Highed})} = \frac{2.45}{1.40} = 1.75 \\ \log[\widehat{\text{OR}}_C(\text{Lib}|\text{Highed})] &= \log[\widehat{\text{OR}}_L(\text{Lib}|\text{Highed})] - \log[\widehat{\text{OR}}_L(\text{Cons}|\text{Highed})] \\ &= 0.90 - 0.34 = 0.56.\end{aligned}$$

Analogous relations hold also for the parameters of a multinomial logistic model. Suppose we use Conservative as the reference level and define the model as

$$\begin{aligned} \log[\text{Odds}_C(\text{Lab}|\text{Highed})] &= \alpha_C^{(\text{Lab})} + \beta_C^{(\text{Lab})}\text{Highed} \\ \log[\text{Odds}_C(\text{Lib}|\text{Highed})] &= \alpha_C^{(\text{Lib})} + \beta_C^{(\text{Lib})}\text{Highed}. \end{aligned}$$

The coefficients of this model are related to those of (6.1)-(6.2), where Labour was the reference level, by

$$\alpha_C^{(\text{Lab})} = -\alpha_L^{(\text{Cons})} \tag{6.3}$$

$$\beta_C^{(\text{Lab})} = -\beta_L^{(\text{Cons})} \tag{6.4}$$

$$\alpha_C^{(\text{Lib})} = \alpha_L^{(\text{Lib})} - \alpha_L^{(\text{Cons})} \quad \text{and} \tag{6.5}$$

$$\beta_C^{(\text{Lib})} = \beta_L^{(\text{Lib})} - \beta_L^{(\text{Cons})}. \tag{6.6}$$

These relationships hold also for estimates of the parameters. From estimates of the model fitted with one reference level, such as those in Figure 6.1, it is thus possible to derive parameter estimates for the same model with any reference level.

6.1.2 General definition

To define the multinomial logistic model in general terms, let C denote the number of categories of the response variable Y. In the voting example there are three possible choices (Lab, Cons and Lib), so $C = 3$. If $C = 2$, the multinomial logistic model reduces to the binary logistic model. The categories of Y are here numbered $0, 1, \ldots, C-1$, where 0 denotes the category which will be treated as the reference category. Both the assignment of numbers to categories (including the choice of the reference category) and the set of numbers used are arbitrary, and changes in them will not change the model itself. For example, the levels of Y may often be coded in a data set as $1, 2, \ldots, C$, instead of starting from zero.

An appropriate probability distribution for such categorical variables is the **multinomial distribution**. Its parameters are the category probabilities

$$\pi^{(j)} = P(Y = j) \quad \text{for } j = 0, 1, \ldots, C-1.$$

In the voting example, these are thus the probabilities of voting Lab, Cons and Lib. Because the sum of the probabilities over all the possibilities must be 1, only $C-1$ of the probabilities need to be known to determine them all. For example, if $C = 3$, we can calculate $\pi^{(0)} = 1 - \pi^{(1)} - \pi^{(2)}$ if $\pi^{(1)}$ and $\pi^{(2)}$ are known. If $C = 2$, the multinomial distribution becomes the binomial distribution discussed in Section 5.2.1. There we used simpler notation where $\pi^{(1)} = \pi$ and $\pi^{(0)} = 1 - \pi$. With more than two categories, however, it is useful to indicate the categories j explicitly in the notation.

Suppose we have data on n sets of observations $(Y_i, X_{1i}, \ldots, X_{ki})$, $i = 1, \ldots, n$, where Y is a response variable with C categories, and $X_1, \ldots, X_k$ are explanatory variables. The multinomial logistic model is defined by the following assumptions:

1. Observations Y_i are statistically independent of each other.

2. Observations Y_i are a random sample from a population where Y_i has a multinomial distribution with probability parameters $\pi_i^{(0)}, \pi_i^{(1)}, \ldots, \pi_i^{(C-1)}$.

3. The log odds of each non-reference category $j = 1, \ldots, C-1$ against the reference category 0 depend on the explanatory variables through

$$\log\left(\frac{\pi_i^{(j)}}{\pi_i^{(0)}}\right) = \alpha^{(j)} + \beta_1^{(j)} X_{1i} + \cdots + \beta_k^{(j)} X_{ki} \quad \text{for each } j = 1, \ldots, C-1 \tag{6.7}$$

where $\alpha^{(j)}$ and $\beta_1^{(j)}, \ldots, \beta_k^{(j)}$ are unknown population parameters.

Comparing (6.7) to (5.4) on page 115, we can see that the multinomial logistic model is essentially a set of $C-1$ binary logistic models, one for each non-reference category of the response variable against the reference category. For example, when $C = 3$, (6.7) defines the two submodels

$$\log\left(\frac{\pi_i^{(1)}}{\pi_i^{(0)}}\right) = \alpha^{(1)} + \beta_1^{(1)} X_{1i} + \cdots + \beta_k^{(1)} X_{ki} \quad \text{and} \tag{6.8}$$

$$\log\left(\frac{\pi_i^{(2)}}{\pi_i^{(0)}}\right) = \alpha^{(2)} + \beta_1^{(2)} X_{1i} + \cdots + \beta_k^{(2)} X_{ki}. \tag{6.9}$$

Model (6.1)–(6.2) for the example of the previous section was clearly of this form (although using slightly different notation, for the didactic purposes of that section), with 0=Lab, 1=Cons and 2=Lib, and X_1 ='Highed' as the only explanatory variable.

The specification (6.7) assumes that the same explanatory variables $X_1, \ldots, X_k$ are included in each of the $C-1$ submodels. We will consider only this case, as is standard practice in multinomial logistic modelling. In many software packages (including SPSS, but not Stata) it is not even possible to fit models with different sets of explanatory variables for different log odds.

6.1.3 Interpretation of the coefficients

From the discussion above, it is clear that the interpretation of the coefficients of a multinomial logistic model is very similar to that of a binary logistic model. Specifically, each coefficient in (6.7) is interpreted as follows:

- If the value of an explanatory variable X_l increases by 1 unit while all the other variables are held fixed, the log odds of the response variable Y being in category j rather than the reference category 0 increase by $\beta_l^{(j)}$. Similarly, the odds of being in category j than 0 are then multiplied by $\exp(\beta_l^{(l)})$.

This implies, first, that the *signs* of the coefficients are interpreted in expected ways:

- if $\beta_l^{(j)} > 0$, $\exp(\beta_l^{(j)}) > 1$: increasing X_l **increases** the odds of being in category j rather than the reference category 0

Figure 6.2: Edited SPSS output for a multinomial logistic model for voting in the 2001 UK General Election, given age and education.

Parameter Estimates

Vote in 2001 General Election[a]		B	Std. Error	Wald	df	Sig.	Exp(B)
Conservative	Intercept	-1.861	.265	49.147	1	.000	
	age	.021	.004	24.804	1	.000	1.021
	educ=3	.638	.151	17.807	1	.000	1.892
	educ=2	.474	.225	4.422	1	.035	1.606
	educ=1	0[b]	.	.	0	.	.
Liberal Democrat	Intercept	-1.809	.316	32.796	1	.000	
	age	.005	.005	1.044	1	.307	1.005
	educ=3	1.026	.181	32.031	1	.000	2.791
	educ=2	.746	.263	8.068	1	.005	2.108
	educ=1	0[b]	.	.	0	.	.

a. The reference category is: Labour.

b. This parameter is set to zero because it is redundant.

- if $\beta_l^{(j)} = 0$, $\exp(\beta_l^{(j)}) = 1$: increasing X_l has **no effect** on the odds of being in category j rather than 0
- if $\beta_l^{(j)} < 0$, $\exp(\beta_l^{(j)}) < 1$: increasing X_l **decreases** the odds of being in category j rather than 0

The quantitative interpretation of the coefficients as odds ratios applies both when an explanatory variable is continuous and when it is a dummy variable, although it is described in somewhat different ways in these two cases, in a by now familiar way. To illustrate this with the voting data, consider two explanatory variables, age in years as a continuous variable and education in three categories as defined on page 143. Two dummy variables are defined for education, here for its two highest categories (2=upper secondary and 3=post-secondary), leaving the lowest (1=lower secondary or less) as the reference level.

SPSS output (slightly edited) for a fitted model for these variables is shown in Figure 6.2. We will use these estimates to illustrate the interpretation of the coefficients. Subscripts are used to refer to the explanatory variables as 1=Age, 2=Dummy variable for education level 2 and 3=Dummy variable for education level 3, and superscripts to refer to categories of the response variable as (1)=Cons and (2)=Lib. Using this notation, the estimated coefficients in Figure 6.2 are interpreted as:

- Coefficients of age:
  - $\hat{\beta}_1^{(1)} = 0.021$ and $\exp(\hat{\beta}_1^{(1)}) = 1.021$: a one-year increase in age, holding education level constant, multiplies the odds of voting Cons rather than Lab by 1.021, i.e. increases them by 2.1%.
  - $\hat{\beta}_1^{(2)} = 0.005$ and $\exp(\hat{\beta}_1^{(1)}) = 1.005$: a one-year increase in age, holding education level constant, multiplies the odds of voting Lib rather than Lab by 1.005, i.e. increases them by 0.5%.

    * Odds ratios corresponding to other changes in age are obtained as exponentials of multiples of the coefficients, as for binary logistic models. For example, a 10-year increase in age, holding education constant, multiplies the odds of voting Lib instead of Lab by $\exp(10 \times \hat{\beta}_1^{(2)}) = \exp(0.05) = 1.05$.

- Coefficients of the dummy variable for upper secondary education:

  - $\hat{\beta}_2^{(1)} = 0.474$ and $\exp(\hat{\beta}_2^{(1)}) = 1.606$: holding age constant, the odds for someone with upper secondary education of voting Cons rather Lab are 1.606 times (i.e. 60.6% higher than) the odds for someone with lower secondary or less education.
  - $\hat{\beta}_2^{(2)} = 0.746$ and $\exp(\hat{\beta}_2^{(2)}) = 2.108$: holding age constant, the odds for someone with upper secondary education of voting Lib rather Lab are 2.108 times the odds for someone with lower sedondary or less education.

- Coefficients of the dummy variable for post-secondary education:

  - $\hat{\beta}_3^{(1)} = 0.638$ and $\exp(\hat{\beta}_3^{(1)}) = 1.892$: holding age constant, the odds for someone with post-secondary education of voting Cons rather Lab are 1.892 times the odds for someone with lower secondary or less education.
  - $\hat{\beta}_3^{(2)} = 1.026$ and $\exp(\hat{\beta}_3^{(2)}) = 2.791$: holding age constant, the odds for someone with post-secondary education voting Lib rather Lab are 2.791 times the odds for someone with lower sedondary or less education.

Qualitatively, increasing age thus increases the odds (and thus also the probability) that a person will vote for Cons rather than Lab, but has little effect on the choice between Lib and Lab. Having higher levels of education also increases the probability of voting Cons rather than Lab, and even more strongly the probabililty of voting Lib rather than Lab.

Comparisons against non-reference categories of *explanatory* variables are obtained in the same way as for binary logistic models. For example,

- $\exp(-\hat{\beta}_2^{(1)}) = \exp(-0.474) = 0.623 = 1/1.606$. Thus, holding age constant, the odds for someone with lower secondary or less education of voting Cons rather than Lab are 0.61 times the odds for someone with upper secondary education.

- $\exp(\hat{\beta}_3^{(1)} - \hat{\beta}_2^{(1)}) = \exp(0.638 - 0.474) = \exp(0.164) = 1.178 = 1.892/1.606$. Thus, holding age constant, the odds for someone with post-secondary education of voting Cons rather than Lab are 1.18 times the odds for someone with upper secondary education.

In a very similar way, comparisons of odds against non-reference categories of the *response* variable can be expressed in terms of differences of the coefficients. Suppose, for example, that we want to use level 1 instead of 0 as the reference level. This can

be done by re-expressing model (6.7) as

$$\begin{aligned}
\log\left(\frac{\pi_i^{(0)}}{\pi_i^{(1)}}\right) &= -\alpha^{(1)} - \beta_1^{(1)}X_{1i} - \cdots - \beta_k^{(1)}X_{ki} \quad \text{and} \\
\log\left(\frac{\pi_i^{(j)}}{\pi_i^{(1)}}\right) &= (\alpha^{(j)} - \alpha^{(1)}) + (\beta_1^{(j)} - \beta_1^{(1)})X_{1i} + \cdots + (\beta_k^{(j)} - \beta_k^{(1)})X_{ki} \\
&\qquad \text{for each } j = 2, \ldots, C-1
\end{aligned}$$

where the αs and βs are the coefficients in the model defined by (6.7), where 0 was the reference level. In short, the model for the log odds of 0 vs. 1 is obtained by changing the signs of the coefficients in the model for 1 vs. 0, and the model for any other category against 1 by using differences of the coefficients in model (6.7). The expressions (6.3)–(6.6) in the previous section are an example of these transformations. For the estimated model in Figure 6.2, we can for example derive

- $\exp(-\hat{\beta}_1^{(1)}) = \exp(-0.021) = 0.979 = 1/1.021$. Thus, holding education level constant, a one-year increase in age multiplies the odds of voting Lab rather than Cons by 0.979, i.e. decreases them by 2.1%.
- $\exp(\hat{\beta}_1^{(2)} - \hat{\beta}_1^{(1)}) = \exp(0.005 - 0.021) = \exp(-0.016) = 0.984 = 1.005/1.021$. Thus, holding education level constant, a one-year increase in age multiplies the odds of voting Lib rather than Cons by 0.984, i.e. decreases them by 1.6%.

These estimates can of course be obtained also by simply refitting the model with Cons specified as the reference category.

Like any other regression models, multinomial logistic models may also include interaction terms and nonlinear transformations of the explanatory variables. Interactions are interpreted as for binary logistic models (see e.g. Section 5.3.4 for examples), except now with reference to odds of one response category against the reference category. For example, suppose that a model for voting included an interaction between age and education. This would imply, for example, that the effect of age on the odds of voting Cons rather than Lab and Lib rather than Lab would depend on a person's level of education.

6.1.4 Probabilities of the response categories

The multinomial logistic model (6.7) is defined as $\log(\pi^{(j)}/\pi^{(0)}) = L^{(j)}$, where

$$L^{(j)} = \alpha^{(j)} + \beta_1^{(j)}X_1 + \cdots + \beta_k^{(j)}X_k \tag{6.10}$$

(omitting the subject subscript i here for simplicity). From this, we can calculate the probabilities of the categories $j = 1, \ldots, C-1$ as

$$\pi^{(j)} = P(Y = j) = \frac{\exp(L^{(j)})}{1 + \exp(L^{(1)}) + \cdots + \exp(L^{(C-1)})} \tag{6.11}$$

and the probability of the reference category 0 as

$$\pi^{(0)} = P(Y = 0) = \frac{1}{1 + \exp(L^{(1)}) + \cdots + \exp(L^{(C-1)})}. \tag{6.12}$$

These are of the same general form as the probability formula (5.5) for the binary logistic model. If there are only two response categories, (6.11) becomes $\pi^{(1)} = \exp(L^{(1)})/[1 + \exp(L^{(1)})]$ and (6.12) becomes $\pi^{(0)} = 1/[1 + \exp(L^{(1)})]$, exactly the same as from the binary logistic formula (apart from slightly different notations here and in Chapter 5).

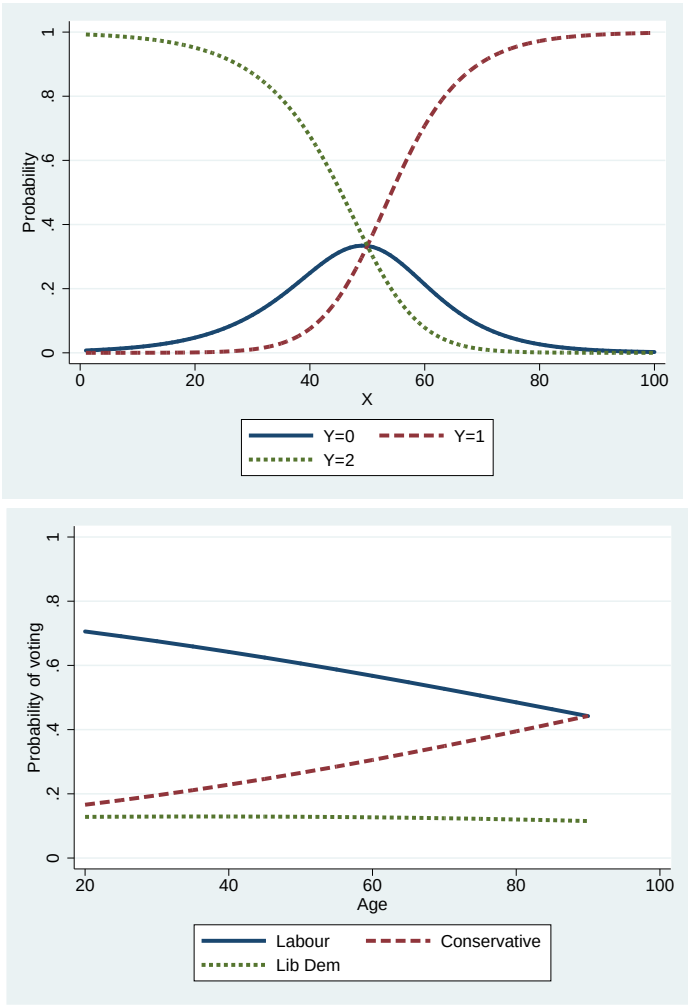
To understand the probabilities obtained from the formulas (6.11) and (6.12), recall first the plot in the lower part of Figure 5.1 on page 115. This showed the probabilities $\pi = P(Y = 1)$ from the formula (5.5) for the *binary* logistic model, as a single continuous explanatory variable X varies from large negative to large positive values. The probabilities trace an S-shaped curve which either increases from 0 to 1 or decreases from 1 to 0 across the whole range of X.[2] The curve is increasing if the regression coefficient of X is positive and decreasing if it is negative. The curve for $P(Y = 0) = 1 - \pi$ is a mirror image of the one for π, so only one of them is typically plotted.

The upper plot of Figure 6.3 shows a comparable example of response probabilities as functions of a continuous X from a multinomial logistic model for $C = 3$ response categories. The probabilities of all three categories are now shown. Only two of the curves are S-shaped, one increasing monotonically from 0 to 1, and one decreasing from 1 to 0. The third curve, and in general all the remaining $C - 2$ probabilities for a C-category response variable, are increasing for some values of X and decreasing for others. The shapes of the curves are related to the relative sizes of the coefficients of X for different response categories. Suppose we associate the notional coefficient 0 with the reference category 0 and the coefficient of X in the model for $\log[\pi^{(j)}/\pi^{(0)}]$ with each other category j. The largest of these numbers identifies the category for which the probability curve is monotonically increasing, and the smallest the category for which it is decreasing. For example, the upper plot of Figure 6.3 is based on a model where the coefficient of X is largest for category 1 and smallest for category 2.

The lower plot of Figure 6.3 shows fitted probabilities of vote given age for the estimated model shown in Figure 6.2, with education fixed at level 1 (lower secondary or less). Here the coefficients of age associated with the response categories are 0 for Lab, 0.021 for Cons and 0.005 for Lib. The predicted probabilities for voting Lab should thus be consistently decreasing and those of voting Cons consistently increasing with age, while the probabilities for Lib increase for some and decrease for some values of age. The plot shows that this is indeed the case. In other ways, however, it looks rather different from the upper plot of Figure 6.3. This is because it is limited to values for age (20 to 90 years) which are sensible and actually observed, We are, in effect, looking at just a slice of the full potential pattern of probabilities displayed by the upper plot. Over this range of ages, the probabilities change fairly gradually and do not vary all the way from 0 to 1. The changes in the probabilities for Lab and Cons are nearly linear over this age range, while the probabilities of Lib hardly change at all.

[2] The probabilities may not be monotonic, i.e. they may both increase and decrease at different values of X, if the model includes, for example, quadratic effects X^2. Examples of such cases are not considered here.

Figure 6.3: Examples of fitted probabilities from a multinomial logistic model with three response categories, as functions of one continuous explanatory variable.



6.2 Estimation and inference for the parameters

6.2.1 Estimation of coefficients and their standard errors

As for binary logistic models (c.f. Section 5.2.6), the parameters of a multinomial logistic model are estimated using the method of Maximum Likelihood Estimation. The calculations are done with computer algorithms. This gives us maximum likelihood estimates $\hat{\beta}_l^{(j)}$ of the regression coefficients, and estimated standard errors $\hat{\text{se}}(\hat{\beta}_l^{(j)})$ of them. Figures 6.1 and 6.2 are examples of SPSS output for estimated multinomial logistic models.

6.2.2 Inference for individual coefficients

Confidence intervals for individual coefficients and odds ratios are also calculated the same way as for binary logistic models (c.f. Section 5.4.2). For example, a 95% confidence interval for the population coefficient $\beta_l^{(j)}$ is given by

$$\hat{\beta}_l^{(j)} \pm 1.96\,\hat{\text{se}}(\hat{\beta}_l^{(j)}) \tag{6.13}$$

and a 95% confidence interval for the corresponding odds ratio by exponentiating the end points of (6.13), i.e. as

$$\left(\exp[\hat{\beta}_l^{(j)} - 1.96\,\hat{\text{se}}(\hat{\beta}_l^{(j)})];\ \exp[\hat{\beta}_l^{(j)} + 1.96\,\hat{\text{se}}(\hat{\beta}_l^{(j)})]\right). \tag{6.14}$$

For example, consider the coefficient of the dummy variable for the middle level of education in a model for the log odds of voting Cons rather than Lab given education and age, a cofficient which was previously labelled $\beta_2^{(1)}$. The exponential $\exp(\beta_2^{(1)})$ of this is thus the ratio of the odds of voting Lab rather than Cons between a respondent with upper secondary education and one with lower seondary or no education, controlling for age. From Figure 6.2, the estimate of this coefficient is $\hat{\beta}_2^{(1)} = 0.474$, with the standard error $\hat{\text{se}}(\hat{\beta}_2^{(1)}) = 0.225$. From (6.14) we obtain the 95% confidence interval

$$\begin{aligned}(\exp[0.474 - 1.96 \times 0.225];\ &\exp[0.474 + 1.96 \times 0.225]) \\ &= \quad (\exp[0.033];\ \exp[0.915]) = (1.03,\ 2.50).\end{aligned}$$

for $\exp(\beta_2^{(1)})$. Standard Stata and SPSS output shows confidence intervals for these odds ratios for every coefficient in the model (although here they had been omitted from the output shown in Figures 6.1 and 6.2, to save space).

Parallelling the familiar elements of other regression models, we could also define for each individual coefficient the Wald test statistic

$$z = \frac{\hat{\beta}_l^{(j)}}{\hat{\text{se}}(\hat{\beta}_l^{(j)})}. \tag{6.15}$$

SPSS output in fact shows the square of this (z^2), labelled as "Wald" (c.f. the example outputs above). The difference is unimportant, as both versions of the statistic lead to the same (two-sided) P-value.

Test statistic (6.15) can be used to test the null hypothesis $H_0 :\ \beta_l^{(j)} = 0$. This is the hypothesis that, controlling for the other explanatory variables, X_l does not affect the population odds of the response category being j rather than 0. The sampling distribution of z under H_0 is approximately standard normal (and that of z^2 is χ^2 with 1 degree of freedom), as for similar test statistics for binary logistic models.

For example, consider the coefficients of the explanatory variable Age in the model shown in Figure 6.2. These were previously denoted $\beta_1^{(1)}$ and $\beta_1^{(2)}$. To ease the discussion of them here and below, let us relabel them $\beta_{\text{Age}}^{(\text{Cons})}$ and $\beta_{\text{Age}}^{(\text{Lib})}$ respectively, using text sub- and superscripts in an obvious way. The model output shows that the P-value for the test statistic (6.15) is $P < 0.001$ for $\beta_{\text{Age}}^{(\text{Cons})}$ and $P = 0.307$ for $\beta_{\text{Age}}^{(\text{Lib})}$.

In other words, age has a significant effect on the odds of voting Cons rather than Lab, but no significant effect on the odds of voting Lib rather than Lab. This seems interesting from a substantive point of view. For the use of the multinomial logistic model, however, it is in practice relatively unimportant or at least inconsequential, for reasons discussed below.

6.2.3 Model selection

The reason why we do not in practice pay much attention to significance tests of individual coefficients in multinomial logistic models is that these tests cannot easily be used to aid the selection of explanatory variables. In the example above we concluded that $\hat{\beta}_{\text{Age}}^{(\text{Cons})}$ was significantly different from 0, while $\hat{\beta}_{\text{Age}}^{(\text{Lib})}$ was not. This suggests that the explanatory variable age might be omitted from the model for the log odds of Lib vs. Lab, but not from the model for Cons vs. Lab. In practice, however, this is rarely done: as noted in Section 6.1.2, we usually consider only multinomial logistic models where all the models for the different log odds include the same explanatory variables.

The hypothesis regarding age which we are really interested in is whether it has any effect on voting at all, after controlling for other explanatory variables. That hypothesis is not the same as either $H_0: \beta_{\text{Age}}^{(\text{Cons})} = 0$ or $H_0: \beta_{\text{Age}}^{(\text{Lib})} = 0$ individually, but

$$H_0: \beta_{\text{Age}}^{(\text{Cons})} = \beta_{\text{Age}}^{(\text{Lib})} = 0. \tag{6.16}$$

When (6.16) is true, age affects neither the odds of voting Cons rather than Lab, nor the odds of Lib vs. Lab; furthermore, these together also imply that age has no effect on the odds of Cons vs. Lib either. The hypothesis that age has no effect on voting is thus the hypothesis that both coefficients of age in the model are 0 at once. This can also be seen by considering the formulas (6.11) and (6.12) for the response probabilities. When (6.16) is true, none of the linear predictors $L^{(j)}$ contain anything involving age, so the probabilities $\pi^{(j)}$ do not depend on age either.

A similar hypothesis for a categorical explanatory variable with more than two categories involves coefficients of several dummy variables. In the voting model in Figure 6.2, the hypothesis that, controlling for age, education has no effect on voting is thus

$$H_0: \beta_{\text{Educ2}}^{(\text{Cons})} = \beta_{\text{Educ2}}^{(\text{Lib})} = \beta_{\text{Educ3}}^{(\text{Cons})} = \beta_{\text{Educ3}}^{(\text{Lib})} = 0, \tag{6.17}$$

again identifying coefficients by obvious sub- and superscripts.

Hypotheses of main interest in multinomial logistic models are thus inevitably of the form of (6.16) and (6.17), where several regression coefficients are set to zero at once. In Section 5.4.3, we described how the likelihood ratio (LR) can be used to test such hypotheses for binary logistic model. Exactly the same test is used again here. Suppose that the null hypothesis H_0 claims that all the coefficients corresponding to a particular explanatory variable X_l (or possibly several of them, as in the case of several related dummy variables) are 0. The test is carried out as follows:

1. Fit the multinomial logistic model under the null hypothesis, i.e. a model which omits the explanatory variable X_l. Record the log likelihood value $\log L_1$ for this model (the restricted model).

2. Fit the multinomial logistic model which includes X_l, and record the log likelihood value $\log L_2$ for this model (the full model).

3. The LR test statistic, here denoted L^2, for testing H_0 is

$$L^2 = -2 \log L_1 - (-2 \log L_2).$$

4. The P-value of the test is obtained by referring L^2 to its approximate sampling distribution, which is a χ^2 distribution with degrees of freedom *df* equal to the number of coefficients set to 0 by the null hypothesis. For example, if H_0 concerns a single continuous explanatory variable, $df = C-1$. If it refers to the coefficients of a set of K related dummy variables, $df = K(C-1)$.

As an example, consider a model for voting where we include all the eight explanatory variables listed on page 143. Suppose we want to test whether age has an effect on voting, after we control for the other seven variables. This is done as follows:

- Fit a model without age, but including the other explanatory variables. For this, -2 times the log likelihood is $-2 \log L_1 = 2250.782$.
- Fit a model with age and the other explanatory variables included. For this, $-2 \log L_2 = 2228.298$.
- The LR test statistic is thus $L^2 = 2250.782 - 2228.298 = 22.484$.
- The degrees of freedom of the test are $df = 2$. This is because there are 2 coefficients of age, one for the model for the log odds of voting Cons vs. Lab, and one for Lib vs. Lab. The P-value for $L^2 = 22.484$ from a χ^2_2 distribution is $P < 0.001$. The null hypothesis is thus clearly rejected. Age has an effect on vote choice, even after controlling for the other explanatory variables.

When a multinomial logistic model is fitted in SPSS, the output includes the results of such LR tests for each explanatory variable. In other words, in addition to the model requested, SPSS fits all the models with one of the explanatory variables excluded, and uses the LR test to compare each of them in turn to the full model with all the variables included. Figure 6.4 shows examples of this output. The upper table in it contains results for voting models for the variables considered here. The row labelled "age" shows the results obtained above, for the test of removing age from the model.

Likelihood ratio tests like these are used to select which explanatory variables should be included in the model. The principles and procedures of this are the same as those discussed in Sections 4.8 and 5.4. For illustration, suppose that we start with the set of potential explanatory variables for voting in the upper part of Figure 6.4. Suppose further that we have a substantive research hypothesis about the explanatory variable (labelled *brn_abr* in the table) which indicates whether the respondent was born outside the UK. Because this variable is the focus of our research question, it is kept in the model throughout. The other explanatory variables are regarded as control variables. We can then examine whether any of them may be omitted as not statistically significant, using the kind of a stepwise testing procedure illustrated for binary logistic models in Section 5.4.4 and computer class 7.

Figure 6.4: Examples of SPSS output of Likelihood ratio tests of individual explanatory variables in multinomial logistic models, applied to models for voting in the 2001 UK General Election. See the text for more details.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	2228.298[a]	.000	0	.
age	2250.782	22.484	2	.000
city	2245.789	17.492	2	.000
brn_abr	2236.165	7.867	2	.020
female	2228.956	.659	2	.719
tvtotx	2251.178	22.881	2	.000
marital	2231.024	2.726	8	.950
educ	2258.094	29.796	4	.000
empl	2240.025	11.727	4	.019

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	1994.426[a]	.000	0	.
age	2018.559	24.132	2	.000
city	2012.440	18.013	2	.000
brn_abr	2002.320	7.893	2	.019
tvtotx	2017.465	23.039	2	.000
educ	2024.641	30.215	4	.000
empl	2007.456	13.030	4	.011

The upper table of Figure 6.4 shows that either sex (the dummy variable *female*) or marital status may be removed from the model if the other seven explanatory variables are included. Further tests (not shown here) show that the other one of these two is not significant even after the other one has been omitted, so both sex and marital status can be removed from the model. After that, all of the remaining six explanatory variables are significant at the 5% level of significance, as shown by the output in the lower part of Figure 6.4. In particular, the dummy variable for having been born outside the UK is statistically significant ($P = 0.019$), indicating that this variable is associated with vote choice even after controlling for age, domicile, education, employment status and TV viewing habits.

Standard Stata or SPSS output does not contain test results for some common types of hypotheses about the explanatory variables. One such case is the hypothesis that some but not all categories of a categorical explanatory variable have the same effect on the response variable, i.e. that those categories can be combined in the model. This can be formulated as a hypothesis about the coefficients of individual dummy variables (c.f. the discussion in Section 4.6), which can also be tested using likelihood ratio tests. Two examples from the voting model illustrate these possibilities:

- In the model reached above, the P-values of the Wald test statistic (6.15) for the coefficients of the dummy variable for the third category of employment status (not in paid employment) are 0.87 in the model for voting Cons rather than Lab, and 0.35 in the model for Lib vs. Lab. On the other hand, at least one of the coefficients of the second category (self-employed) is clearly significant according to this test. This suggests that there are no differences in probabilities of voting between the employed and those in paid work, but both may differ from the self-employed. The null hypothesis corresponding to this is that both coefficients of the dummy variable of the third employment category are zero. We can test it with a LR test which compares two models, both with the other five explanatory variables included: a model with dummy variables for both the self-employed and for those not in work included (the full model), and a model with only the dummy variable for the self-employed included (the reduced model). The resulting test statistic is $L^2 = 1.07$ with 2 degrees of freedom, and $P = 0.59$. We can thus combine the categories "employed" and "not in paid work", and proceed with a model which only distinguishes these from being self-employed.

- In the model obtained after the previous step, the coefficients (not shown here) of the two dummy variables for the two higher education levels are clearly significantly different from 0, but quite similar to each other (relative to their standard errors) in the models for both Cons vs. Lab and Lib vs. Lab. This suggests the null hypothesis that these coefficients are equal in both submodels separately. If this is not rejected, we can combine the two higher categories of education (upper secondary and post-secondary) and contrast them together with the lowest level (lower secondary or less). The hypothesis is tested with a LR test which compares two models, both with the other five explanatory variables included: a model with separate dummy variables for the two highest categories included (the full model), and a model with only the dummy variable for the combined category included (the reduced model). The resulting test statistic is $L^2 = 0.97$ with 2 degrees of freedom, and $P = 0.62$. We can thus combine the two higher categories, and proceed with a model with a dichotomous education variable.

Table 6.2: Estimates for the selected multinomial logistic model for voting in the 2001 UK General Election.

Variable	Coeff.	(s.e.)	z	P-value	Odds ratio [exp(β)] Est.	95% CI
Voting Conservative vs. Labour:						
(Constant)	-1.451	(0.285)				
Age (years)	0.024	(0.004)	5.57	< 0.001	1.02	(1.02; 1.03)
High education	0.479	(0.148)	3.24	0.001	1.62	(1.21; 2.16)
Self-employed	0.763	(0.247)	3.09	0.002	2.14	(1.32; 3.48)
Living in a city	-1.753	(0.492)	-3.56	< 0.001	0.17	(0.07; 0.45)
TV watching (hours/day)	-0.154	(0.034)	-4.58	< 0.001	0.86	(0.80; 0.92)
Born abroad	-0.622	(0.265)	-2.35	0.019	0.54	(0.32; 0.90)
Voting Lib Dem vs. Labour:						
(Constant)	-1.436	(0.335)				
Age (years)	0.007	(0.005)	1.39	0.164	1.01	(0.997; 1.02)
High education	0.896	(0.179)	5.00	< 0.001	2.45	(1.72; 3.48)
Self-employed	-0.110	(0.340)	-0.32	0.746	0.90	(0.46; 1.74)
Living in a city	-0.345	(0.354)	-0.98	0.330	0.71	(0.35; 1.42)
TV watching (hours/day)	-0.114	(0.040)	-2.81	0.005	0.89	(0.82; 0.97)
Born abroad	-0.642	(0.322)	-1.99	0.046	0.53	(0.28; 0.99)

The selected model has two continuous explanatory variables, age and weekly hours of TV watching, and four dichotomous explanatory variables, education (low or high), employment status (self-empoyed or other), domicile (city or other) and having been born outside the UK (yes or no). All of these are significant at the 5% level according to LR tests. Parameter estimates for the model are shown in Table 6.2. In broad terms, the model suggests that, in each case controlling for the other explanatory variables,

- Increasing age increases the probability of voting Cons rather than Lab or Lib.
- Having higher levels of education is associated with increased probabilities of voting Cons and, in particular, Lib.
- Being self-employed increases the chances of voting Cons rather than Lab or Lib.
- Living in a big city is associated with decreased probability of voting Cons rather than Lab or Lib.
- Respondents who spend most time watching television are most likely to vote Lab rather than Cons or Lib.
- Having been born outside the UK is associated with an increased probability of voting Lab rather than Cons or Lib.

The precise quantitative interpretation of the results is in terms of odds ratios, as outlined in Section 6.1.3. You can practise this with the estimates in Table 6.2.

6.3 Interpreting the model with fitted probabilities

The most concise way of describing a fitted multinomial logistic model is in terms of its coefficients, interpreted as odds ratios as explained in Section 6.1.3. Even this, however, is not entirely straightforward nor as intuitive as for many other regression models, largely because of the inevitably large number of coefficients involved in a multinomial model. To ease the interpretation further, it is desirable to supplement the discussion of the coefficients with other ways of presentation. Fitted response probabilities are the main method of doing this.

Fitted probabilities for different values of the response variable can be calculated from the formulas (6.11) and (6.12) given any values of the explanatory variables. We can thus use such probabilities at selected values of the explanatory variables to illustrate the implications of the model. As an example, consider the estimated model in Table 6.2. Suppose we first consider a hypothetical respondent who has lower secondary or less education (i.e. the "High education" dummy variable in the model is 0), is self-employed, lives in a city and was born in the UK; furthermore, his or her age and daily hours of watching TV are equal to the sample averages of these variables in the data, at 52.7 years and 3.25 hours respectively. Substituting these values of the explanatory variables and the estimated coefficients from Table 6.2 into equation (6.10), we obtain the linear predictors

$$\begin{aligned} L^{(\text{Cons})} &= -1.451 + 0.024 \times 52.7 + 0.479 \times 0 + 0.763 \times 1 \\ &\quad -1.753 \times 1 - 0.154 \times 3.25 - 0.622 \times 0 = -1.677 \\ L^{(\text{Lib})} &= -1.436 + 0.007 \times 52.7 + 0.896 \times 0 - 0.110 \times 1 \\ &\quad -0.345 \times 1 - 0.114 \times 3.25 - 0.642 \times 0 = -1.893 \end{aligned}$$

and thus from (6.12) and (6.11)

$$\begin{aligned} P(\text{Lab}) &= \frac{1}{1 + \exp(L^{(\text{Cons})}) + \exp(L^{(\text{Lib})})} = \frac{1}{1 + \exp(-1.677) + \exp(-1.893)} = 0.75, \\ P(\text{Cons}) &= \frac{\exp(L^{(\text{Cons})})}{1 + \exp(L^{(\text{Cons})}) + \exp(L^{(\text{Lib})})} = \frac{\exp(-1.677)}{1 + \exp(-1.677) + \exp(-1.893)} = 0.14, \\ P(\text{Lib}) &= \frac{\exp(L^{(\text{Lib})})}{1 + \exp(L^{(\text{Cons})}) + \exp(L^{(\text{Lib})})} = \frac{\exp(-1.893)}{1 + \exp(-1.677) + \exp(-1.893)} = 0.11. \end{aligned}$$

This is clearly not something we would want to calculate by hand very often, so computers are used to carry out the computations. An example in computer class 8 shows how this can be done in Stata and SPSS.

Fitted probabilities may be displayed in tabular or graphical form. Table 6.3 shows an example of a table of such probabilities, again for the voting model in Table 6.2. The fitted probabilities of all three parties are shown, given all 16 combinations of the four dichotomous explanatory variables in the model, in each case fixing the two continuous ones at their sample mean values (52.7 years for age, 3.25 hours for TV watching)[3]. For example, the first row of the table gives the probabilities for the case (low education, self-employed, living in a city and born in the UK) considered above.

[3]Note that some of the 16 combinations may be fairly rare in the actual sample. Predicted values for such cases should not interpreted too seriously.

Table 6.3: Predicted probabilities of voting for the three parties, for the estimated model shown in Table 6.2, given all combinations of the dichotomous explanatory variables. Age (at 52.7) and TV watching (at 3.25) are fixed at their sample mean values. In each row, the highest probability is shown highlighted and the second highest framed.

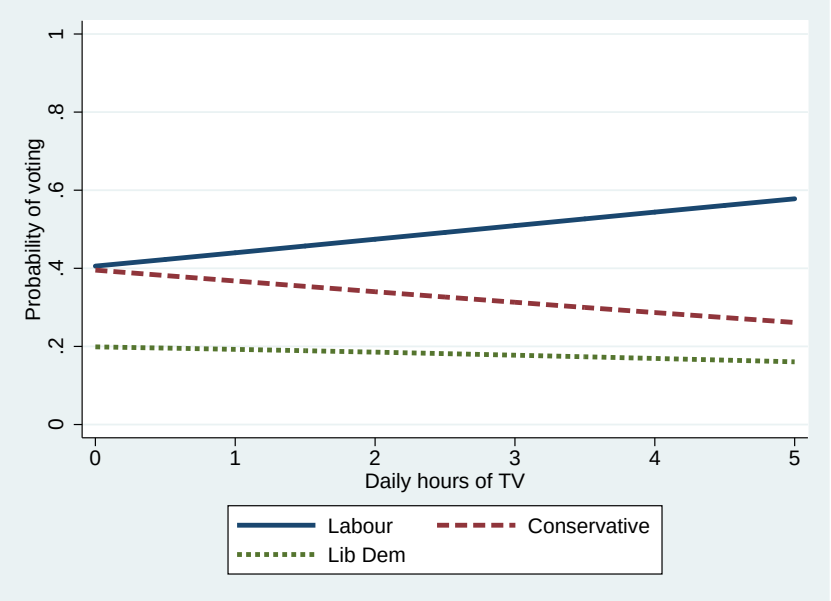
Education	Employment	Domicile	Born	Lab	Cons	Lib
Low	Self	City	UK	.75	.14	.11
Low	Self	City	Abroad	.85	.09	.07
Low	Self	Other	UK	.30	.47	.09
Low	Self	Other	Abroad	.45	.34	.06
Low	Other	City	UK	.80	.07	.14
Low	Other	City	Abroad	.88	.04	.09
Low	Other	Other	UK	.57	.29	.13
Low	Other	Other	Abroad	.72	.19	.08
High	Self	City	UK	.60	.18	.22
High	Self	City	Abroad	.74	.12	.14
High	Self	Other	UK	.30	.54	.16
High	Self	Other	Abroad	.45	.42	.12
High	Other	City	UK	.64	.09	.27
High	Other	City	Abroad	.77	.06	.17
High	Other	Other	UK	.42	.34	.24
High	Other	Other	Abroad	.57	.25	.17

The table shows that the predicted probability is highest for Labour in all but two cases, and the second highest for the Conservative party in most cases. The cases where Liberal Democrats have the second highest probability are all ones where the respondent lives in a city.

For continuous explanatory variables, we may display a table of fitted probabilities calculated at selected values, or a plot of the probabilities over an appropriate range of values of the explanatory variable. Figure 6.5 shows fitted probabilities of voting as the number of daily hours spent watching TV varies between 0 and 5, holding all the other explanatory variables at their sample mean values[4]. The plot shows, in particular, the extent to which higher amounts of TV watching are associated with increased probabilities of voting Lab and decreased probabilities of voting Cons.

[4]For the dummy explanatory variables, this sample mean is the *proportion* of people in the sample who have the value 1 for the dummy variable. For example, for the dummy variable for having been born abroad it is 0.0811, the proportion of the respondents who were born outside the UK. A dummy variable fixed at such a value clearly does not represent a meaningful single individual. This, however, does not affect the calculations, and such sample proportions are often used to represent "average" values of the dummy variables in examples of fitted probabilities.

Figure 6.5: Predicted probabilities of voting for the three parties, obtained from the estimated model shown in Table 6.2, given the self-reported number of hours spent watching TV on an average weekday. All the other explanatory variables are fixed at their sample mean values.



Chapter 7

Ordinal logistic models

7.1 Introduction

The regression models described in this chapter are comparable to multinomial logistic models in that the response variable is again a categorical variable with more than two levels. Now, however, it will be assumed that this variable is ordinal, i.e. that its levels are regarded as having a known and fixed ordering. If this is the case, we might choose one of three broad approaches to modelling the ordinal response:

- Use the multinomial logistic model nevertheless, effectively ignoring the ordering of the categories. This approach is always possible, but it has the disadvantage that the interpretation of the results is not as convenient as for models which do take the ordering into account.

- Assign numerical scores (e.g. $1, 2, 3, 4$) to the levels of the variable and use a linear model, treating the scores as values of a continuous, interval-level response variable. This is a common approach, especially when the number of response categories is large. It is, however, obviously not entirely satisfactory, as it implies specifying not just an ordering but also specific intervals between the categories of the response. Such intervals have no real substantive justification if the variable is measured only at an ordinal level.

- Use models which treat the measurement level of the response variable specifically as ordinal, rather than too weakly as nominal or too strongly as interval. In this chapter we focus on one type of such *ordinal regression models*, the *ordinal logistic* (or *proportional odds*) *model*. Other types of ordinal regression models are discussed briefly in Section 7.7.

Ordinal regression models should not be used if the response levels cannot be treated as ordered, because then the interpretation of the model will be nonsensical. If the response is binary, ordinal logistic models are equivalent to multinomial logistic models, and both are equivalent to binary logistic models.

7.2 Example

We will use for illustration models where the response variable is a survey measure on **self-reported health**. The question wording was *"How is your health in general? Would you say it is..."* and the response options were *Very bad* (coded as 1), *Bad* (2), *Fair* (3), *Good* (4) and *Very good* (5). Such self-reports have often been found to be useful measures of general health status, especially in broad population studies. The five-level measure considered here can clearly be treated as ordinal at face value, with higher-numbered levels corresponding to better self-reported health.

The data used here come from Round 5 of the European Social Survey (ESS), carried out in 2010[1], and includes the 48,979 respondents from 27 countries for whom complete data were available on self-reported health and all the explanatory variables listed below. In this sample, the proportions of levels 1–5 of self-reported health were 1.8%, 8.1%, 27.7%, 39.3% and 23.1% respectively.

Our examples are loosely motivated by the questions addressed in an article by von dem Knesebeck and Geyer (2007)[2], but we do not aim to replicate their analyses exactly (and we use data from a more recent round of the ESS). The primary research questions of von dem Knesebeck and Geyer concerned the associations that a person's level of *education* and his or her access to *emotional support* may have with self-reported health. The measure of emotional support that they used was the following:

- Whether the person has anyone with whom they can discuss intimate and personal matters: This is recorded as Yes or No, and labelled *intimate* below.

Alongside this variable, we also consider three other measures of somewhat related constructs:

- Partnership status, with four levels: Living currently with a partner; Not currently living with a partner and has been widowed in the past; Not currently living with a partner or ever widowed but has lived with a partner in the past; Never lived with a partner (*partnership*).
- How often the person meets socially with friends, relatives or work colleagues: Never; Less than once a month; Once a month; Several times a month; Once a week; Several times a week; Every day (*meeting friends*).
- Compared to other people of his or her age, how often the person takes part in social activities: Much less than most; Less than most; About the same; More than most; Much more than most (*social activities*).

For education, an obvious challenge in international comparative research is measuring educational attainment in a cross-nationally comparable way. Here we use the classifi-

[1] ESS Round 5: European Social Survey Round 5 Data (2010). Data file edition 3.0. Norwegian Social Science Data Services, Norway Data Archive and distributor of ESS data.

[2] von dem Knesebeck, O. and Geyer, S. (2007). Emotional support, education and self-rated health in 22 European countries. *BMC Public Health*, 7:272.

cation employed by the ESS, which is a modified version of the International Standard Classification of Education[3]:

- Person's highest level of education: Less than lower secondary; Lower secondary; Upper secondary without university access; Upper secondary with university access; Post-secondary below bachelor's level; Bachelor's level (lower tertiary); Master's level or above (upper tertiary) (*education*).

In addition, we will include the respondent's *age* (in years) and *sex* (Male or Female) as control variables in the models.

7.3 Motivation of the models: Cumulative probabilities

As before, we denote a categorical response variable by Y and the number of its categories (levels) by C. In this chapter, it will be convenient to label the categories as $1, 2, \ldots, C$, rather than $0, 1, \ldots, C-1$ as we did in the multinomial logistic Chapter 6. Since Y is now regarded as ordinal, the categories must be numbered in the right order. Two such numberings are always possible, depending on which end of the ordered list of categories is labelled as 1 and which one as C. We can choose either one of them, as the substantive interpretation of the fitted model will be the same for both choices (this will be justified later). When we describe the ordinal logistic model below, we will refer to lower-numbered categories as "low" values of the variable and higher-numbered ones as "high" ones, irrespective of the substantive interpretation of the levels.

In the example of self-rated health, there are $C = 5$ categories. We will number them from 1 for "Very bad" to 5 for "Very good" as listed in Section 7.2, so that in this example "higher" values in the generic language above correspond to better levels of self-reported health. The other possible numbering would be in reverse, from 1 for "Very good" to 5 for "Very bad". Both of these would be equally valid and would result in the same fitted models. No other numberings would be appropriate, as they would imply a different (and substantively unmotivated) ordering of the categories. For example, if we used 1 = Good, 2 = Very bad, 3 = Bad, 4 = Fair, 5 = Very good, the conclusions from an ordinal regression model would be nonsensical.

As a simple example, consider a model for self-reported health given sex. The data for it is summarised by the contingency table in part (a) of Table 7.1. In the previous notation and terminology, the proportions in this table are estimates of the probabilities

$$\pi^{(j)} = P(Y = j), \quad \text{for } j = 1, \ldots, C,$$

of each of the levels j of the response variable Y, here the levels of self-reported health. The table shows these proportions conditional on sex, i.e. separately for men and women. For example, the proportion of respondents who report Fair general health (corresponding to $\pi^{(3)}$) is 0.254 among men and 0.297 among women.

[3]UNESCO (2006 [1997]). *International Standard Classification of Education: ISCED 1997* (re-edition). Montreal: UNESCO-UIS; modified according to suggestions in Schneider, S. L. (2010). Nominal comparability is not enough: (In-)Equivalence of construct validity of cross-national measures of educational attainment in the European Social Survey. *Research in Social Stratification and Mobility*, 28, 343–357.

Table 7.1: Table of sex by self-reported health in data from Round 5 of the European Social Survey. Part (a) of the table shows conditional proportions of the levels of self-reported health, and part (b) shows their cumulative proportions, counting from the level "Very bad" upwards.

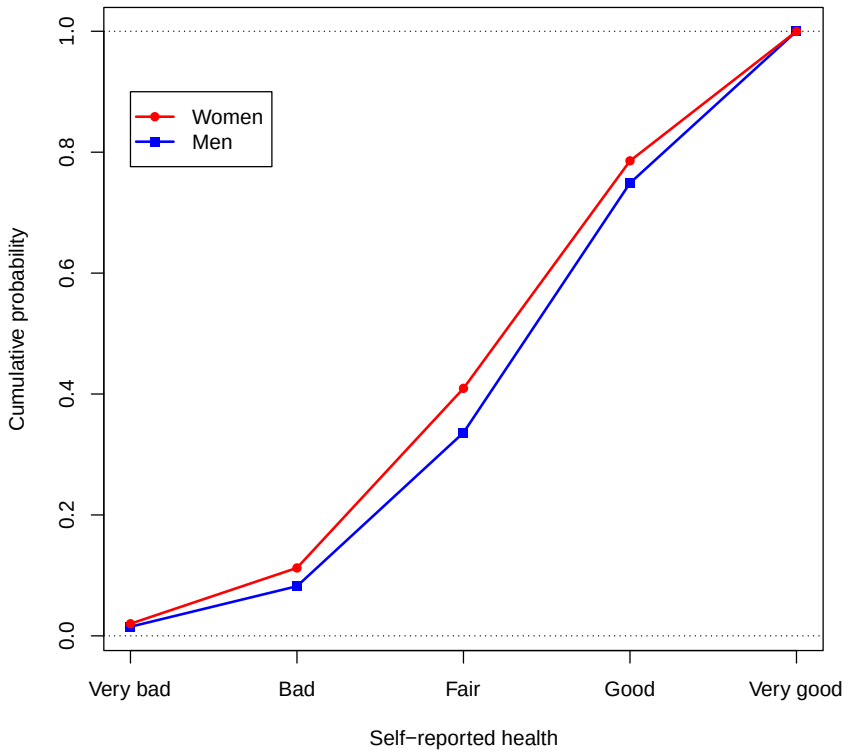
(a) Proportions of levels of Self-reported health (given Sex)

	Self-reported health						
Sex	Very bad	Bad	Fair	Good	Very good	Total	n
Men	.015	.067	.254	.413	.251	1.000	22240
Women	.020	.092	.297	.376	.215	1.000	26739

(b) Cumulative proportions of levels of Self-reported health (given Sex)

	Self-reported health						
Sex	Very bad	Bad	Fair	Good	Very good	Total	n
Men	.015	.082	.336	.749	1.000	—	22240
Women	.020	.112	.409	.785	1.000	—	26739

Figure 7.1: Cumulative proportions of self-reported health conditional on sex, in data from Round 5 of the European Social Survey. This figure shows the same information as part (b) of Table 7.1.



In multinomial logistic models for unordered response variables we modelled these category probabilities $\pi^{(j)}$ directly. For ordinal response variables, we will consider instead the **cumulative probabilities**

$$\gamma^{(j)} = P(Y \leq j) = P(Y = 1) + \cdots + P(Y = j) = \pi^{(1)} + \cdots + \pi^{(j)} \quad \text{for } j = 1, \ldots, C. \tag{7.1}$$

The cumulative probability $\gamma^{(j)}$ of level j is the sum of the individual category probabilities summed ("cumulated") from level 1 to level j. In other words, it is the probability that the value of Y is *at most* j — i.e. j *or* one of the lower-numbered categories. In our example, $\gamma^{(1)}$ is the probability that self-reported health is Very bad, $\gamma^{(2)}$ the probability that it is Very bad or Bad, and so on. The corresponding proportions in the example, conditional on sex, are shown in part (b) of Table 7.1. This shows, for example, that in this sample the proportion of respondents who report Very bad, Bad or Fair general health (corresponding to the cumulative probability $\gamma^{(3)}$) is 0.336 among men and 0.409 among women. The same information is also displayed graphically in Figure 7.1, which shows the cumulative proportions for each of the five categories against the numbers of the categories.

Considering the definition in (7.1) and the example in Table 7.1 and Figure 7.1, we can observe the following general facts about cumulative probabilities $\gamma^{(j)}$:

- $\gamma^{(j-1)} \leq \gamma^{(j)}$: The cumulative probability for a category is never smaller than the one for the previous category.
- $\gamma^{(1)} = \pi^{(1)}$: For the lowest category, the cumulative probability is the same as the probability of the category itself.
- $\gamma^{(C)} = 1$: The cumulative probability of the highest category is 1. This is simply a way of stating the fact that the value of Y must be one of its possible levels.
- Putting these results together, we get formulas for the category probabilities expressed in terms of the cumulative probabilities:

$$\pi^{(1)} = \gamma^{(1)}, \tag{7.2}$$

$$\pi^{(j)} = \gamma^{(j)} - \gamma^{(j-1)} \quad \text{for } j = 2, \ldots, C-1, \tag{7.3}$$

$$1 - \sum_{j=1}^{C-1} \pi^{(j)} = \pi^{(C)} = 1 - \gamma^{(C-1)}. \tag{7.4}$$

  These show that all the information about the probabilities of Y is contained in the $C-1$ category probabilities $\pi^{(1)}, \ldots, \pi^{(C-1)}$ or, equivalently, the $C-1$ cumulative probabilities $\gamma^{(1)}, \ldots, \gamma^{(C-1)}$.

- $P(Y < j) = \gamma^{(j-1)}$, $P(Y > j) = 1 - \gamma^{(j)}$ and $P(Y \geq j) = 1 - \gamma^{(j-1)}$: These show how the cumulative probabilities, i.e. the probabilities that Y is *less than or equal to* a particular level, can be used to calculate also the probabilities of *less than*, *greater than*, and *greater than or equal to* any level.
- If we decided to order the categories in reverse order, cumulative probabilities would also be calculated in that order. In the health example this would mean numbering the levels from Very good = 1 to Very bad = 5. The cumulative probabilities would then be obtained by summing probabilities from right to left

(instead of left to right) in part (a) of Table 7.1; for men, for example, they would thus be 0.251, 0.664, 0.918, 0.985, and 1.000. Since cumulative probabilities in one direction can always be transformed into cumulative probabilities in the other direction (as shown in the previous bullet point), both sets carry exactly the same information about the distribution of Y.

As explained in the next section, an ordinal logistic model is essentially a set of binary logistic models for cumulative probabilities. It can be used to examine associations between explanatory variables and the ordinal response variable. For example, in Figure 7.1 we can see that at all levels of self-reported health the cumulative probability of that level and all worse levels of health is higher for women than for men. It thus seems appropriate to say that women's self-reported health tends to be worse than men's in this sense. We will see later how the ordinal logistic model summarises this association.

7.4 Basic elements of the model

7.4.1 Definition of the model

Suppose that we have data on n sets of observations $(Y_i, X_{1i}, \ldots, X_{ki})$, $i = 1, \ldots, n$, where Y is an ordinal response variable with C categories labelled from lowest to the highest as $j = 1, \ldots, C$, and $X_1, \ldots, X_k$ are explanatory variables. The *ordinal logistic model* — which is also known as the *cumulative logistic model* or the *proportional odds model* — is defined by the following assumptions:

1. Observations Y_i are statistically independent of each other.

2. Observations Y_i are a random sample from a population where Y_i has a multinomial distribution with probability parameters $\pi_i^{(1)}, \pi_i^{(2)}, \ldots, \pi_i^{(C)}$, and thus with the cumulative probabilities

$$\gamma_i^{(j)} = P(Y_i \leq j) = \pi_i^{(1)} + \cdots + \pi_i^{(j)}$$

   for $j = 1, \ldots, C$, where $\gamma_i^{(C)}$ is equal to 1 for all i and thus need not be modelled separately.

3. The log odds based on the cumulative probabilities depend on the explanatory variables through

$$\log\left(\frac{\gamma_i^{(j)}}{1-\gamma_i^{(j)}}\right) = \log\left[\frac{P(Y_i \leq j)}{P(Y_i > j)}\right] = \alpha^{(j)} - (\beta_1 X_{1i} + \cdots + \beta_k X_{ki}) \quad (7.5)$$

$$\text{for each } j = 1, \ldots, C-1$$

   where $\alpha^{(1)}, \ldots, \alpha^{(C-1)}$ and $\beta_1, \ldots, \beta_k$ are unknown population parameters.

We note immediately three points about the model parameters:

- The model has $C - 1$ distinct intercept terms $\alpha^{(1)}, \ldots, \alpha^{(C-1)}$. These must be (and their estimates will always be) ordered in size, so that $\alpha^{(1)} \leq \alpha^{(2)} \leq \cdots \leq \alpha^{(C-1)}$. This ensures that the necessary conditions $\gamma_i^{(j-1)} \leq \gamma_i^{(j)}$ for cumulative probabilities always hold.
- Unlike for the multinomial logistic model (6.7), the regression coefficients $\beta_1, \ldots, \beta_k$ in (7.5) do not have a superscript j. In other words, in an ordinal logistic model each explanatory variable has only one regression coefficient, which applies to the models for each of the cumulative probabilities $\gamma_i^{(1)}, \ldots, \gamma_i^{(C-1)}$.
- The negative sign $(-)$ in front of the $(\beta_1 X_{1i} + \cdots + \beta_k X_{ki})$ in (7.5) will be explained in Section 7.4.3 below.

In Chapter 6 we observed that a multinomial logistic model could be thought of as a set of binary logistic models, each of them for just two of the C categories of Y at a time. The ordinal logistic model can also be thought of as a set of binary models, but now for sets of dichotomous responses obtained by *combining* adjacent categories of Y. This motivation is as follows:

- For each level $j = 1, \ldots, C-1$ of Y, define a new binary variable with two levels: 1 if Y is at most j, and 0 if Y is greater than j.
  - For instance, if $C = 5$, the first of these dichotomies (corresponding to $j = 1$) constrasts level 1 of Y with levels 2–5 combined, the second 1–2 vs. 3–5, the third 1–3 vs. 4–5 and the fourth 1–4 vs. 5.
  - For self-reported health, these dichotomies are thus Very bad vs. all other levels, Very bad and Bad vs. the other three levels, and so on.
- Define a binary logistic model for each of these new dichotomous variables, in such a way that all of these models have the same values of the regression coefficients (the βs) but not of the intercept terms (the $\alpha^{(j)}$s).

This is not a literal description of how an ordinal logistic model is actually estimated, but it is what the model means conceptually.

7.4.2 Fitted probabilities

From (7.5), the formula for cumulative probabilities from the ordinal logistic model is

$$\gamma^{(j)} = P(Y \leq j) = \frac{\exp[\alpha^{(j)} - (\beta_1 X_1 + \cdots + \beta_k X_k)]}{1 + \exp[\alpha^{(j)} - (\beta_1 X_1 + \cdots + \beta_k X_k)]} \tag{7.6}$$

for each $j = 1, \ldots, C-1$. The probabilities $\pi^{(j)}$ of the individual categories can then be calculated as differences of these, using the formulas (7.2)–(7.4). To examine the forms of these various probabilities, consider first a hypothetical example where Y has $C = 4$ levels and there is one continuous explanatory variable X. Suppose that an

ordinal logistic model (7.5) holds for Y given X, withe intercept terms $\alpha^{(1)} = -2$, $\alpha^{(2)} = 1$ and $\alpha^{(3)} = 2$, and the coefficient of X being $\beta_X = 1$. The model is thus given by $\text{logit}[\gamma^{(j)}] = \alpha^{(j)} - \beta_X X$ for $j = 1, 2, 3$. The topmost plot of Figure 7.2 shows the cumulative probabilities (7.6) in this case. From this plot, we note the following general points:

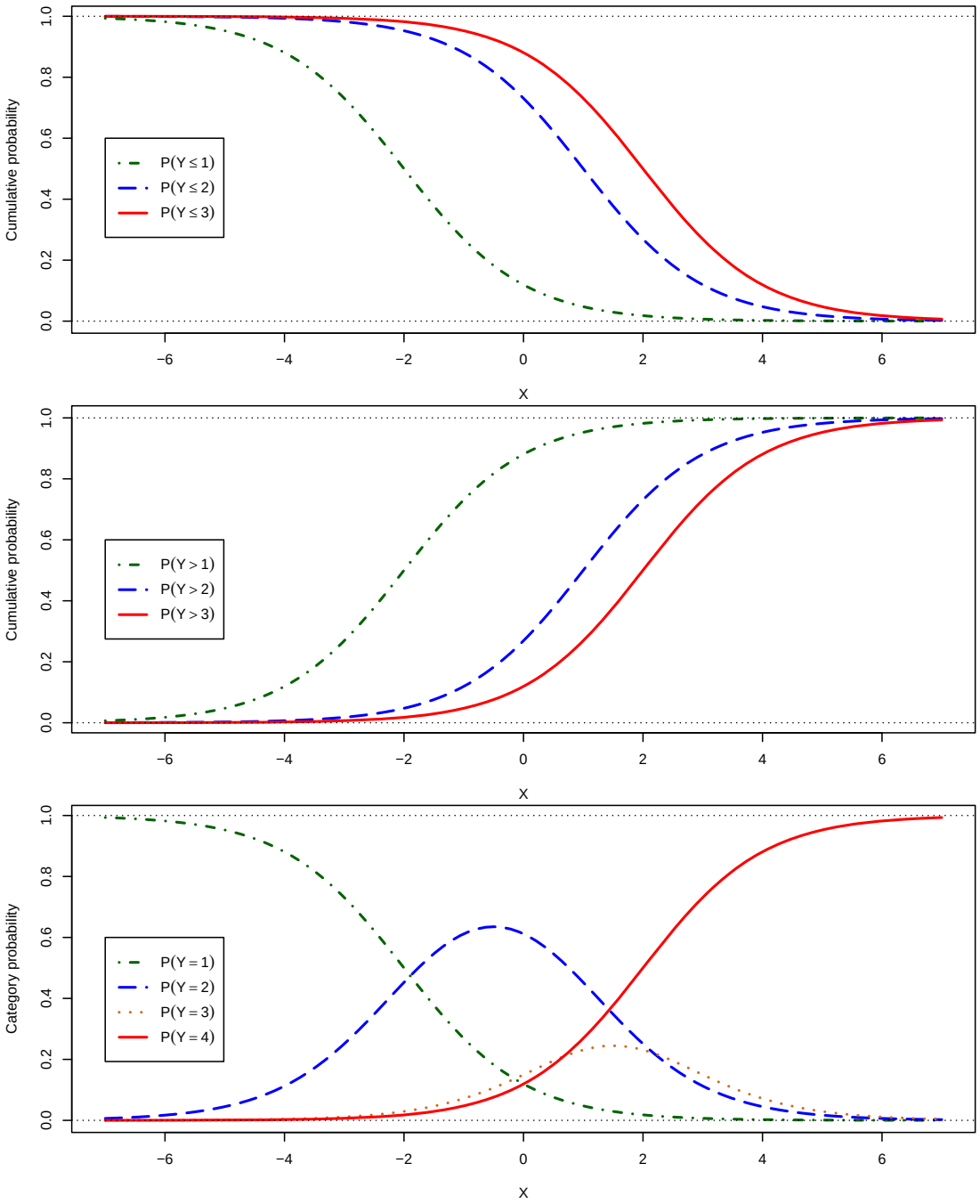
- The magnitude of the coefficient β_X determines the steepness of the curves, that is how strongly $\gamma^{(j)}$ depend on X. The value $\beta_X = 0$ would correspond to no association, i.e. flat curves which would not depend on X at all.
- The curves for $\gamma^{(j)}$ have the same steepness for each category j. This is because they all have the same coefficient β_X for X. This is the *proportional odds* assumption of the ordinal logistic model, which we will discuss further in other sections below.
- Here the *positive* value of $\beta_X = 1$ translates into curves where the probabilities $\gamma^{(j)}$ *decrease* as X *increases* (and, correspondingly, a negative β_X would imply that the curves would increase wth X). This is a consequency of the $-$ sign in front of $\beta_X X$ in the model specification, and in general in (7.5). The motivation of this choice is discussed below.
- The intercept terms $\alpha^{(j)}$ determine the heights of the curves. The *differences* between $\alpha^{(j)}$ for different j determine how far the corresponding curves are from each other. For example, here $\alpha^{(2)} = 1$ is closer to $\alpha^{(3)} = 2$ than to $\alpha^{(1)} = -2$, so the curve of $\gamma^{(2)}$ is closer to that of $\gamma^{(3)}$ than that of $\gamma^{(1)}$. These differences are related to the probabilities $\pi^{(j)}$ of the individual categories, as discussed below.

The middle plot of Figure 7.2 shows the complementary cumulative probabilities $P(Y > j) = 1 - \gamma^{(j)}$, that is, the probabilities that Y has a value *higher than* a given category j. These curves are mirror images of those in the top plot. The important point to note from the middle plot is the following:

- When β_X is positive, the curves for $P(Y > j)$ increase when X increases. This is again a consequence of the $-$ sign in the model equation, and the reason for why that sign is there. With it included, we obtain a model specification where a *positive* regression coefficient of an explanatory variable X implies that *increasing* the value of X *increases* the cumulative probability of *high* values of Y (and, similarly, a negative coefficient means that increasing X decreases the probability of high values of X). It is in order to obtain this convenient (and easiest-to-remember) interpretation for what 'positive' and 'negative' associations mean in an ordinal logistic model that the negative sign is included in front of the regression coefficients in (7.5). This convention is always used on this course, and it is also used in computer output from both SPSS and Stata[4]

[4]It is also used by the `polr` function in R which is employed in the example code for the MY452 computer classes. But other R functions, and other software such as SAS, may define an ordinal logistic model without the $-$, in which case the interpretation of the coefficients is reversed. It is important to check this if you use any unfamiliar software for ordinal logistic modelling.

Figure 7.2: Fitted probabilities from an ordinal logistic model for a response Y with $C = 4$ levels, given a single continuous explanatory variable X. The model has intercept terms $\alpha^{(1)} = -2$ $\alpha^{(2)} = 1$ and $\alpha^{(3)} = 2$, and the regression coefficient $\beta_X = 1$. The plot at the top shows the cumulative probabilities $P(Y \leq j)$, the middle plot the cumulative probabilities $P(Y > j)$ and the bottom plot the category probabilities $P(Y = j)$.



The bottom plot of Figure 7.2 shows the probabilities $\pi^{(j)} = P(Y = j)$ of the individual categories, which are obtained as differences of the cumulative probabilities as shown in equations (7.2)–(7.4). From these, we note the following:

- In the same way as for multinomial logistic models (see Figure 6.3 on page 155), the probabilities of two categories tend to 0 and 1 when a continuous variable like X here takes on very large or very small values. For an ordinal logistic model, these two are always the highest and lowest category of Y. The intermediate categories obtain their highest probabilities at some intermediate values of X.

- If the coefficient of X is positive, the probability of the lowest category of Y tends to 0 and that of the highest category to 1 as X increases, and the values of X at which the intermediate categories of Y obtain their maximum probabilities are in the same order as the numbers of the categories. All of these results are reversed if the coefficient of X is negative.

- The maximum values of the probabilities of the intermediate categories depend on the differences between successive values of the intercepts $\alpha^{(j)}$. For example, here $\alpha^{(1)}$ and $\alpha^{(2)}$ are further apart than are $\alpha^{(2)}$ and $\alpha^{(3)}$. A consequence of this is that the largest value that $P(Y = 2)$ obtains as X varies is larger than the largest value of $P(Y = 3)$.

Next, we illustrate fitted probabilities with an example with the real data, considering a model for self-reported health given sex and age. Estimates for this model are shown in Figure 7.3, in the SPSS and Stata output formats. The estimated values of the parameters are here

- intercepts: $\hat{\alpha}^{(1)} = -6.903$, $\hat{\alpha}^{(2)} = -5.026$, $\hat{\alpha}^{(3)} = -3.086$, and $\hat{\alpha}^{(4)} = -1.077$; the (unimportant) reason for why these are labelled "cuts" (in Stata) or "thresholds" (in SPSS) will be explained in Section 7.7;

- regression coefficients: $\hat{\beta}_{\text{female}} = -0.247$ and $\hat{\beta}_{\text{age}} = -0.049$.

As always, we can choose to calculate fitted probabilities at whatever values of the explanatory variables we want for sensible and helpful interpretation of the model. For example, suppose we decide to calculate them for both men and women, and for the two widely separated ages of 25 and 75. For ordinal logistic models there is also the further choice of several different probabilities that we might calculate, including both cumulative probabilities and probabilities of individual categories. For illustration, suppose that we want to calculate estimated values of the individual probability of Fair health (i.e. $\pi^{(3)}$) and the cumulative probability of Fair or worse health (i.e. $\gamma^{(3)}$). The cumulative probability is obtained directly from (7.6). For a woman who is 25 years old, it is

$$\begin{aligned}\hat{\gamma}^{(3)} &= \frac{\exp(\hat{\alpha}^{(3)} - [\beta_{\text{female}} \times 1 + \beta_{\text{age}} \times 25])}{1 + \exp(\hat{\alpha}^{(3)} - [\beta_{\text{female}} \times 1 + \beta_{\text{age}} \times 25])} \\ &= \frac{\exp(-3.086 - [-0.247 \times 1 - 0.049 \times 25])}{1 + \exp(-3.086 - [-0.247 \times 1 - 0.049 \times 25])} = 0.166.\end{aligned}$$

Figure 7.3: Estimated parameters for an ordinal logistic model in Stata (upper figure) and SPSS (lower figure) output. The fitted model is here for self-reported health given sex and age, in data from Round 5 of the European Social Survey.

health	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
2.female	-.2471966	.0169076	-14.62	0.000	-.280335	-.2140583
age	-.0488667	.0005004	-97.65	0.000	-.0498475	-.0478859
/cut1	-6.902782	.0466941			-6.994301	-6.811264
/cut2	-5.026161	.0342307			-5.093252	-4.95907
/cut3	-3.086451	.0288904			-3.143075	-3.029827
/cut4	-1.076891	.0254569			-1.126785	-1.026996

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[health = 1]	-6.903	.047	21945.896	1	.000	-6.994	-6.811
	[health = 2]	-5.026	.034	21627.769	1	.000	-5.093	-4.959
	[health = 3]	-3.086	.029	11408.346	1	.000	-3.143	-3.030
	[health = 4]	-1.077	.025	1793.770	1	.000	-1.127	-1.027
Location	female	-.247	.017	213.976	1	.000	-.280	-.214
	age	-.049	.000	9615.306	1	.000	-.050	-.048

To calculate $\hat{\pi}^{(3)}$, we need also the cumulative probability $\hat{\gamma}^{(2)}$. In the same way as above — but now using the intercept $\hat{\alpha}^{(2)} = -5.026$ — we get $\hat{\gamma}^{(2)} = 0.028$. The required probability is the difference of these two:

$$\hat{\pi}^{(3)} = \hat{\gamma}^{(3)} - \hat{\gamma}^{(2)} = 0.166 - 0.028 = 0.138.$$

Please take a moment to convince yourself that this makes sense: Since $\hat{\gamma}^{(3)}$ is the predicted probability of Very Bad, Bad, or Fair self-reported health, and $\hat{\gamma}^{(2)}$ is the probability of Very Bad or Bad health, their difference is the probability of Fair alone.

Similar calculations given the other required combinations of sex and age give the fitted probabilities shown in Table 7.2. Considering the cumulative probabilities, it is clear that age has a very substantial effect on these, in the direction that older people tend to report worse levels of health: over 60% of both men and women aged 75, but less than 20% of those aged 25, are predicted to report Fair or worse health. There is also a difference, but a much smaller one, between men and women given age, with women predicted to report somewhat worse levels of health. Considering then the probabilities of the individual category of Fair health, we may note that these appear here to be less helpful for interpretation than the cumulative probabilities, for the reason that it is difficult to judge or compare their values if we do not know how much of the rest of the probability is in higher and lower categories than Fair.

Table 7.2: Some fitted probabilities for self-reported health given sex and age, calculated from the estimated ordinal logistic model in Figure 7.3.

	Age=25		Age=75	
	Men	Women	Men	Women
$\hat{P}$(At most Fair)	0.134	0.166	0.641	0.695
$\hat{P}$(Fair)	0.112	0.138	0.437	0.448

7.4.3 Interpretation of the coefficients

The regression coefficients of ordinal logistic models are interpreted as log odds ratios. This is done in essentially the same way as for binary logistic models, with only the following modifications: (i) the odds ratios that we consider for an ordinal model are for dichotomies defined by grouping the response levels into two groups, "high" and "low"; (ii) the odds ratio is for the high group against the low, because of the change of sign of the coefficients in the model formula; and (iii) the same odds ratio applies to each grouping into high vs. low, because of the proportional odds assumption.

To explain these in more detail, consider the example of self-reported health, in five ordered categories. As noted in Section 7.4.1, an ordinal logistic model can be thought of as a set of binary logistic models for the binary outcomes obtained by grouping adjacent values of the ordinal response into two groups, "high" (i.e. higher-numbered) and "low", divided at each level $j = 1, \ldots, C-1$ in turn. For self-reported health there are four such groupings:

- Very Good [high] vs. (Good, Fair, Bad, Very Bad) [low]
- (Very Good, Good) [high] vs. (Fair, Bad, Very Bad) [low]
- (Very Good, Good, Fair) [high] vs. (Bad, Very Bad) [low]
- (Very Good, Good, Fair, Bad) [high] vs. Very Bad [low]

The regression coefficient of an explanatory variable X is then interpreted as the log odds ratio (and its exponential as the odds ratio), associated with a one-unit increase in X, for the choice of the high rather than low value in these dichotomies. Although there are $C-1$ ways of grouping the response levels into high and low, the same odds ratio applies to each one of them, because the model (7.5) specifies the same regression coefficient for the log odds of the cumulative probability $\gamma^{(j)}$ at each level j. Because of the negative sign inserted in front of the terms which depend on the βs and Xs in (7.5), the odds ratio is specifically for the comparison of high against low, in this direction. This implies the interpretation for positive and negative signs of the regression coefficients which has already been discussed in Section 7.4.2.

For example, consider the estimated model in Figure 7.3. Here the coefficient of age is $\hat{\beta}_{\text{age}} = -0.049$. Because the coefficient is negative, we can immediately conclude that increasing age is associated with a decreasing probability of higher-numbered levels of self-reported health — i.e. levels which correspond to better levels of health. A

quantitative summary of this negative association is provided by the odds ratio, given by $\exp(\hat{\beta}_{\text{age}}) = 0.952$:

- Controlling for a person's sex, a one-year increase in age multiplies the odds of high rather than low levels of self-reported health by 0.952, i.e. decreases them by 4.8%.

The estimated coefficient for the dummy variable for women is $\hat{\beta}_{\text{female}} = -0.247$ and $\exp(\hat{\beta}_{\text{female}}) = 0.781$. The negative sign indicates that women tend to report worse levels of health than men. The odds ratio interpretation is as follows:

- Controlling for a person's age, the odds for a woman of high rather than low levels of self-reported health are 0.781 times the odds for a man, i.e. 21.9% lower.

For the last comment in this section, recall that we noted earlier that an ordinal logistic model will be substantively unchanged if we reverse the numbering of the response levels, e.g. if for self-reported health we switch to 1–5 from Very good to Very bad rather than vice versa. If this is done, all the estimated regression coefficients will have the same absolute values as before, but with opposite signs. This will mean that the interpretation of the odds ratios will be exactly the same as before, since the reversal of the codes will also reverse the meaning of "high" and "low" levels correspondingly.

7.4.4 Estimation and inference

All the methods of estimation and inference are defined and used in the same way for ordinal logistic models as they were for other logistic models, so nothing new needs to be introduced here.

- The parameters of the model are typically estimated using the method of maximum likelihood estimation (see Section 5.5.4).
- Significance tests of the regression coefficients are carried out using the types of tests we are already familiar with. Wald tests, most often applied to single coefficients (see Section 5.4.1), and likelihood ratio tests for any number of parameters (see Section 5.4.3) are both defined and conducted in exactly the same way as for binary logistic models. The hypothesis of no association between an explanatory variable and the response variable is again that the regression coefficient of the explanatory variable is 0, so this is again the null hypothesis that we most often test in practice.
- Confidence intervals for regression coefficients and exponentials of them (i.e. odds ratios) are again calculated as described in Section 5.4.2.

In the model output in Figure 7.3, for example, we can see that the P-values of the coefficients of sex and age are both very small, so the null hypothesis of no association is clearly rejected for each of them. Both age and sex thus have a statistically significant partial association with self-reported health in these data.

7.5 Models for the example

In this section we provide further examples of ordinal logistic models, in the context of the self-rated health example which was introduced in Section 7.2. The models now include also the explanatory variables related to emotional support and education, which were the main motivation of the example. The results below are still mainly illustrations of how models like these may be used and presented, rather than attempted answers to substantive research questions on these topics.

Table 7.3 shows estimates for models which included as explanatory variables the respondent's age, sex, level of education and different measures of emotional or social support. The table shows only the estimated regression coefficients (log odds ratios), omitting the intercept terms and estimated standard errors. All of the explanatory variables included in these models have statistically significant associations with the response, at least at the 5% level of significance (but for the categorical explanatory variables, the coefficients of some of their individual categories are not significantly different from each other).

In each of the models in Table 7.3, the coefficients of age and sex are similar to the ones in models where only these variables were included, and which were interpreted in previous sections. The coefficients of levels of education are also comparable across all of these models. They indicate that, controlling for age, sex and emotional support, higher levels of education are associated with better levels of self-reported health. The largest expected difference is between individuals with less than lower secondary education and ones with a bachelor's degree. The partial log odds ratio between these levels is around 0.8–0.9, which corresponds to an odds ratio of 2.2–2.4 for better levels of self-reported health, comparing the higher education level to the lower one. For comparison with the other explanatory variables, we may note that this odds ratio is about twice the size of the one for men vs. women, and comparable to the association corresponding to an age difference of around 18 years.

Of the measures of emotional support, Model (1) in Table 7.3 uses the survey question of whether or not the respondent has anyone with whom they can discuss intimate and personal matters. Recall that this was also the measure used in the article by von dem Knesebeck and Geyer (2007) which served as a partial motivation for this example. The coefficient of this explanatory variable (as a dummy variable for No) is here strongly significant, with a P-value less than 0.001 according to both the Wald test and the likelihood ratio test (which are not shown here). Its estimated coefficient in Model (1) is $\hat{\beta}_{\text{intimate}} = -0.582$, and $\exp(\hat{\beta}_{\text{intimate}}) = 0.558$. The odds ratio interpretation of the coefficient is thus that the odds of higher (i.e. better) levels of self-reported health are 44.2% lower for someone without emotional support (in the sense of not having someone to discuss intimate matters with) than with it, controlling for sex, age and educational attainment. Compared to the other partial associations in the fitted model, this odds ratio is about 35% higher than the one for the comparison between men and women, and comparable to the ones associated with a 12-year difference in age or between the lowest and highest levels of education in this model. The magnitudes of such associations can be further illustrated by fitted probabilities such as the ones shown in Table 7.4.

Table 7.3: Estimated regression coefficients (log odds ratios) for ordinal logistic models for self-rated health, estimated using data from Round 5 of the European Social Survey ($n = 48979$). All the coefficients except for the ones shown in parentheses are significantly different from 0 at the 5% level of significance.

	Model				
Variable	(1)	(2)	(3)	(4)	(5)
Age (years)	−.046	−.044	−.045	−.046	−.042
Sex: Female	−.279	−.210	−.252	−.244	−.205
Education (vs. Less than lower secondary)					
Lower secondary	(.032)	(.016)	(.030)	(−.016)	(−.015)
Lower tier upper secondary	.137	.139	.136	.112	.089
Upper tier upper secondary	.285	.283	.305	.222	.229
Advanced vocational, sub-degree	.257	.267	.272	.186	.178
Lower tertiary (Bachelor's)	.895	.905	.889	.808	.779
Upper tertiary (Master's or higher)	.492	.502	.494	.394	.378
Measures of emotional or social support:					
Intimate: No	−.582				−.296
Partnership (vs. With partner now)					
Widowed		−.506			−.430
Previously with partner		−.345			−.326
Never with partner		.065			(.040)
Meeting friends (vs. Every day)					
Several times a week			(−.004)		(.015)
Once a week			(−.055)		(.031)
Several times a month			−.192		−.106
Once a month			−.403		−.211
Less than once a month			−.836		−.456
Never			−1.321		−.731
Social activities (vs. Much more than most)					
More than most				(−.043)	(−.050)
About the same				−.252	−.238
Less than most				−.705	−.601
Much less than most				−1.271	−.998

Table 7.4: Fitted probabilities of Very Good or Good self-rated health, based on Model (1) in Table 7.3. Here age is fixed at 45. Education level 'I' corresponds to less than lower secondary education, and 'V2' to at least a Master's degree.

		Women		Men	
Has intimate friend?	Education:	I	V2	I	V2
No		0.46	0.58	0.53	0.65
Yes		0.60	0.71	0.67	0.77

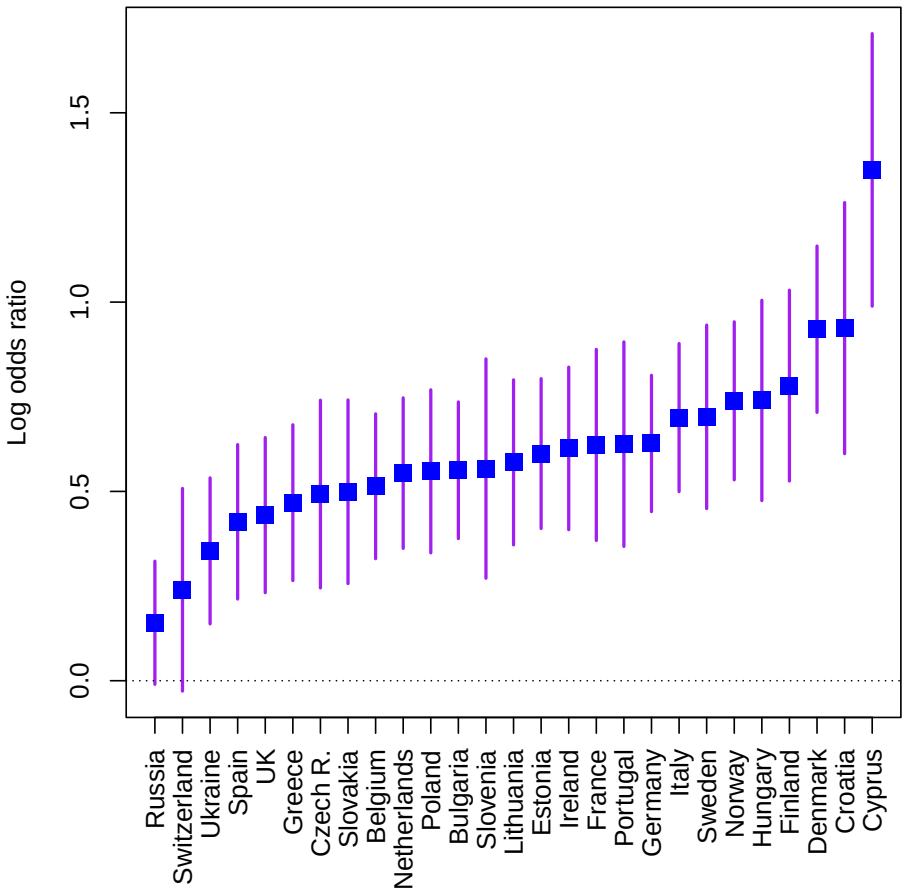
In the other models of Table 7.3 we use the other variables introduced in Section 7.2 which focus on different aspects of emotional or social support. In Models (2)–(4) each of them in turn is included on its own. In each case the variable is strongly statistically significant, and its association with the response variable is broadly in the direction that values which correspond to higher levels of emotional support are associated with better levels of self-reported health, even after controlling for age, sex and level of education. For partnership status, individuals who have previously lived with a partner but now no longer do (due to separation or widowhood) tend to report worse levels of health than those who currently live with a partner, while those who have never lived with a partner report very slightly (and only marginally significantly) better health than those who live with a partner currently. For the variables on frequency of meeting friends and of social activities, perhaps the most noticeable feature is that the levels which indicate the very lowest frequencies (e.g. never meeting friends) are associated with particularly high odds ratios toward poor levels of self-reported health.

All of these possible measures of emotional support are fairly strongly and positively associated with each other, so it is not surprising that each of them individually has a significant association with self-reported health. In Model (5) of Table 7.3 we go beyond this by including all of them in the same model. Here each of them is still strongly significant, even after controlling for the other three (this is inferred from likelihood ratio tests which are not shown here). It thus appears that these variables are not simply substitutes for each other, but that each of them carries some distinct information which is associated with self-reported health. Compared to Models (1)–(4), the estimates change most for having an intimate friend and the frequency of meeting friends, for both of which the coefficients become generally smaller in absolute value. These two are most strongly associated with the other measures of emotional support, particularly with each other. Partnership status and frequency of social acitivities are more distinct, and their coefficients change less when we control for the other measures.

In the second example of this section we consider also the country in which each respondent lives, to illustrate the types of comparative analyses which might often be conducted using cross-national survey data such as the European Social Survey. Here we focus on the respondent's education and its association with self-reported health. Education may be seen as one element of a person's socio-economic status, and any associations between education and health thus as instances socio-economic *inequalities* in health. Furthermore, the extent to which these associations vary between different countries can suggest cross-national variation in the magnitudes of such inequalities. This leads us to consider models for health which include an *interaction* between an individual's education and the country in which he or she lives.

For simplicity of presentation, we consider a dichotomous education variable which contrasts university education (the two highest levels in Table 7.3) with all the other levels. The models also include age, sex and emotional support (whether the person has an intimate friend), plus the country as a categorical explanatory variable, i.e. in the form of 26 dummy variables (there are 27 countries in the data). If the model has no interaction between country and education, it implies that the association between education and health is the same in every country. If we then include such an interaction, we conclude that it is clearly significant ($P < 0.001$ for a likelihood ratio test with 26 degrees of freedom). There is thus strong evidence that the association between an individual's education and health varies in strength across the countries.

Figure 7.4: Estimated regression coefficients and their 95% confidence intervals for a dummy variable of university education (vs. any other level of education), separately for each of 27 countries, in an ordinal logistic model for self-rated health, estimated using data from Round 5 of the European Social Survey ($n = 48979$). The model also includes as explanatory variables sex, age, dummy variable for whether the person has an intimate friend, and the main effect of country.



From this model with the interaction, Figure 7.4 shows the estimated coefficients of the education variable in each country, together with their 95% confidence intervals (the rest of the estimates are not shown here; the coefficients of age, sex and emotional support are broadly unchanged from Table 7.3). We can see that the coefficient is between 0.4 and 0.8 for most countries, again indicating that individuals with university education tend to report better levels of health than ones with other levels of education. The countries are here ordered by size of the coefficient. This, however, should not be treated as a firm ordering of the countries. The confidence intervals remind us that for most pairs of countries the coefficients would not be regarded as significantly different from each other. Formal methods of carrying out such pairwise comparisons are somewhat complicated, because they need to make allowance for the large number of possible pairs which may compared. These methods are not discussed here.

The most prominent individual result in Figure 7.4 is that for Cyprus, where the estimated association between education and self-reported health appears to be sub-

stantially stronger than in any other country in these data. Since we did not designate this one country as being of specific interest before we carried out the analysis, we should not overinterpret this finding after it emerges from the model, but it might well lead us to explore the case of Cyprus in more detail in subsequent research. More generally, in some (but by no means all) cross-national research we might aim to go beyond the essentially decriptive summary in Figure 7.4 of how the individual-level association between education and health varies between countries, by considering theories or hypotheses which aimed to *explain* this variation in terms of characteristics of the countries. Here, for example, someone might formulate such hypotheses based on the different types of wellfare state represented by the countries in these data. We will not pursue this possibility here.

7.6 Assessing the adequacy of the model

A defining feature of the ordinal logistic model is that in its equation (7.5) the regression coefficients (the βs) are the same for all $j = 1, \ldots, C-1$. In other words, the model states that the odds ratios for the cumulative probabilities are the same irrespective of the point at which the response categories are divided into a Low and a High set. A visual expression of this assumption is that in a figure like the top graph of Figure 7.2 the curves of $P(Y \leq j)$ for each j have the same shape and do not cross each other. This condition is known as the *proportional odds assumption* of the ordinal logistic model, and the model itself is also known accordingly as the *proportional odds model*.

The proportional odds assumption is not necessarily satisfied. In other words, even when the response variable is clearly ordinal, its cumulative probabilities do not need to depend on explanatory variables in the particular way specified by the ordinal logistic model. If they do not, conclusions from the model can in principle be misleading, at least if the deviation from the proportional odds assumption is severe. So we have some reason to examine the adequacy of the assumption when we use ordinal logistic models. Unfortunately, however, the tools for doing so are not entirely satisfactory or convenient for routine use. We discuss them only briefly in this section, finishing with some tentative practical suggestions.

Addressing the proportional odds assumption involves two steps: (i) assessing whether the assumption holds, and (ii) if it does not, deciding how to modify the model to relax the assumption. Although (i) clearly comes first in order of action, it is convenient here to begin by discussing (ii). There are very broadly speaking three general approaches that we might take if the proportional odds assumption of an ordinal logistic model does not hold:

- Extending the right-hand side of the model, for example by adding interactions of the explanatory variables. This is of course easy to do, and may (or may not) yield a model for which the proportional odds assumption is more appropriate.
- Modifying the model in some other way, while still maintaining its basic form as a model for cumulative probabilities. The various possible ways of doing this will not be discussed here in more detail, with one exception. This is the extension which might seem the most obvious, that of abandoning the proportional odds

assumption and allowing the regression coefficients in (7.5) to vary with j. If this is done for every explanatory variable X_k, we end up with a model which has the same number of parameters as the multinomial logistic model but which is less appealing than it in a number of ways. An intermediate possibility is to relax the proportional odds assumption in this way only for some (but not all) of the explanatory variables, resulting in a 'partial proportional odds model'. This, however, has one unappealing implication for these variables. Recall that the proportional odds assumption implies that fitted probability curves like the ones in the top plot of Figure 7.2 have similar shapes, so that they are essentially shifted versions of each other. This is no longer true if the proportional odds assumption is relaxed, which in turn has the implication that the curves will cross at some value of the explanatory variable. But this means that for some values of the explanatory variable we have $P(Y \leq j) > P(Y \leq j+1)$ for some j, i.e. that the cumulative probabilities will be in a logically impossible order. This crossing may not actually happen within the observed range of the explanatory variables, but even the possibility of it is a weakness of the partial proportional odds model.

- Fit the multinomial logistic model and use the ordering of the response categories only in an informal way as part of the interpretation of the estimated model. This is the approach which we will focus on in this section.

The situation where we would contemplate these actions is the one where we had first concluded that the proportional odds assumption is inadequate for the data at hand. This returns us to the question of how this conclusion would have been reached. There are several possible methods of assessing the adequacy of the assumption, but none of them fully satisfactory. An obvious possibility would be a significance test of some kind where the null hypothesis corresponds to the proportional odds model and rejecting that hypothesis implies a failure of the proportional odds assumption. Such tests exists, but they are not always fully satisfactory (and not implemented in standard software) and will not be described here. A test which would seem possible given the last bullet point above would be a likelihood ratio test of the ordinal logistic against the multinomial logistic model. This, however, is not appropriate because the assumptions of the test are not satisfied: although the ordinal logistic model is smaller (has fewer parameters) than the multinomial logistic model, it is not formally nested within the multinomial logistic in the way required by the likelihood ratio test.

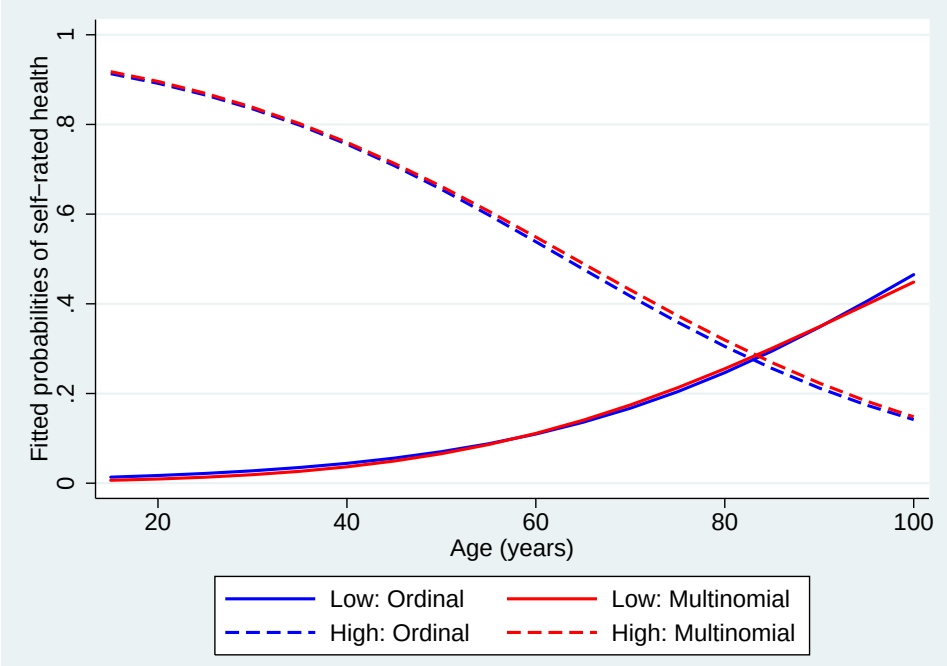
Instead, the most we will suggest on this course is to compare the ordinal and multinomial logistic models in less formal ways. If they appear to be similar, the simpler ordinal logistic model can be used; if not, it is prudent to use the multinomial model even when the response variable is ordinal. To illustrate such comparisons, we return to the model for self-reported health given only age and sex, for which the estimated ordinal model is shown in Figure 7.3 on page 175, and (Stata output of) the estimated multinomial model in Figure 7.5.

We will mention two partial ways of using the multinomial logistic model to assess the appropriateness of the ordinal model. The first of them considers the minimal requirement that the multinomial should at least agree with the ordering of the response categories. This can be examined, still somewhat informally, by checking whether the coefficients of each explanatory variable are in the right order of magnitude. Consider,

Figure 7.5: Estimated parameters for a multinomial logistic model for self-reported health given sex and age, in data from Round 5 of the European Social Survey. Estimates for an ordinal logistic model for the same variables is shown in Figure 7.3.

health	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
very_bad	(base outcome)					
bad						
2.female	.03863	.0770103	0.50	0.616	-.1123075	.1895674
age	-.0119411	.0025282	-4.72	0.000	-.0168963	-.0069859
_cons	2.258891	.1752473	12.89	0.000	1.915413	2.60237
fair						
2.female	-.116378	.0721178	-1.61	0.107	-.2577262	.0249703
age	-.0400007	.0023795	-16.81	0.000	-.0446644	-.035337
_cons	5.255027	.1647191	31.90	0.000	4.932184	5.57787
good						
2.female	-.3599883	.0720972	-4.99	0.000	-.5012963	-.2186804
age	-.0720576	.002386	-30.20	0.000	-.0767341	-.0673811
_cons	7.364301	.1644434	44.78	0.000	7.041998	7.686604
very_good						
2.female	-.419159	.073837	-5.68	0.000	-.5638768	-.2744411
age	-.1030816	.0024499	-42.08	0.000	-.1078834	-.0982798
_cons	8.150099	.1660007	49.10	0.000	7.824744	8.475455

Figure 7.6: Fitted cumulative probabilities for Very Bad or Bad ('Low') and for Very Good or God ('High') self-reported health given age, for men, based on the ordinal logistic model in Figure 7.3 and the multinomial logistic model in Figure 7.5.



for example, the coefficients of age in the multinomial model in Figure 7.5. Taking the coefficient of the reference category Very Bad to be 0, these coefficients, from Very Bad to Very Good, are $0, -0.011, -0.048, -0.072$, and -0.103. The thing to note here that the coefficients are in order of size, here in the sense that they get smaller for the submodels for the categories which are further from Very Bad according to the proposed ordering of the response categories. If this is the case, the multinomial logistic model implies the same ordering of the categories as the one assumed a priori in the ordinal model. For the dummy variable for women we have in Figure 7.5 the coefficients $0, 0.039, -0.116, -0.360$ and -0.419. These are in order, *except* for the 0.039 in the model for Bad vs. Very Bad, which is not between 0 and -0.116. However, this coefficient is not significantly different from 0, so the evidence is also consistent with the possibility that it could be slightly less than 0 and the estimated coefficients thus all in order. In short, the multinomial logistic model does not give any strong reason here to question the proposed ordering of the response categories, as reflected in their associations with age and sex.

Apparently correct ordering is, however, just the minimum requirement for appropriateness of an ordinal model, and one which we would expect to be satisfied in most sensible applications of the model. Checking it is not the same as checking the more stringent requirement of the proportional odds assumption. To assess the latter, the approach which we suggest here is to compare fitted probabilities from ordinal and multinomial logistic models for the same variables. An example of this, for the model introduced above, is given in Figure 7.6. This shows two curves of cumulative probabilities as functions of age (for men), one for Very Bad or Bad self-reported health, and one for Very Good or Good health (for the multinomial logistic model, each of these is obtained by calculating the probabilities from equations (6.10)–(6.12) for the two response categories separately, and summing them). It is clear that for at least these probabilites there is essentially no difference between the ordinal and multinomial models, so we have no reason to abandon the simpler ordinal model.

Comparing fitted probabilities is not a comprehensive check of the proportional odds assumption. To use if for a full evaluation we would in principle need to apply it given all the different explanatory variables in the model, or even for any one of them given different values of the other explanatory variables, which is practically infeasible in general. However, even spot checks like the one in Figure 7.6 are useful for revealing whether there appear to be major differences between the multinomial and ordinal models. This is particularly convenient if the focus of interest is on one explanatory variable in particular. Fitted probabilities given that variable are then the ones we should compare. If they are very similar for the two types of model — e.g. as they are given age in Figure 7.6 — we can be reasonably confident that conclusions concerning that explanatory variable are not substantially affected by the specific form of the ordinal regression model.

7.7 Other topics[1]

In this section we mention briefly some additional topics related to ordinal regression models. First, we note that some of the further considerations discussed in Section 5.5 for binary regression models are also relevant for ordinal ones. Instead of a logistic model, it is possible to specify other model formulations for cumulative probabilities, by replacing the logit function on the left-hand side of (7.5) with another appropriate link function (c.f. Section 5.5.2). In particular, by choosing the inverse cumulative distribution function $\Phi^{-1}(\cdot)$ of the standard normal distribution gives the *ordinal probit model*. This is an extension of the binary probit model in the same way that the ordinal logit model is of the binary logit model.

The motivation of binary regression models in terms of an underlying latent continuous variable, which was discussed in Section 5.5.3 on page 139, can also be extended to ordinal models. This is done by starting again from equation (5.25), and then supposing that the ordinal outcome Y is obtained by dividing all the values of the latent variable Y^* into three or more ranges of adjacent values at two or more cut-off points (rather than into just two ranges at the cut-off point of 0, as in the binary case). This yields an ordinal model for cumulative probabilities, the exact form of which depends on the distribution of ϵ in (5.25); in particular, an ordinal logistic model is obtained if ϵ follows the standard logistic distribution, and an ordinal probit model if ϵ has the standard normal distribution. The values of the cut-off points translate into the intercept terms $\alpha^{(j)}$ of the ordinal model, and it is this reasoning which explains the use of terms like "cuts" or "thresholds" for the intercept terms in computer output for ordinal logistic models. As in the binary case, this latent variable motivation of the model is a mathematically convenient device which is always possible but never necessary for interpretation of ordinal regression models for cumulative probabilities.

As the second topic of this section, we mention briefly two other types of ordinal regression models. These are genuinely different from the ordinal logistic model in that they are not based on modelling cumulative probabilities but use the ordering of the response categories in other ways. The goodness of fit of the different ordinal models tends to be fairly similar in most practical applications, so none of the models is consistently 'better' than the others in this sense. Their differences are mostly in the interpretation of the parameters, which may in any particular application be more convenient for one of the models than for the others. The ordinal logistic model is very often preferable in this sense and in other ways (such as easy availability in standard software), so it is the one we use on this course.

For simplicity of notation, consider models with a single explanatory variable X (models with more explanatory variables are obvious extensions of these). The *adjacent-categories logit model* is defined by

$$\log \frac{\pi^{(j+1)}}{\pi^{(j)}} = \alpha^{(j)} + \beta X \quad \text{for } j = 2, \ldots, C. \tag{7.7}$$

In other words, it is based on modelling each of the comparisons between the probabilities of a response category j and the next category $j + 1$. This too is an ordinal model, because the concept of the 'next' category is only meaningful if the categories

[1] The material in this section is not required in the examination.

are treated as ordered. Model (7.7) assumes that the same regression coefficient β applies to the comparison of all pairs of adjacent categories, which is analogous to the proportional odds assumption of the ordinal logistic model.

An interpretation of a model in terms of adjacent categories is not very often more convenient than the one in terms of cumulative probabilities which applies to the ordinal logistic model. The adjacent-categories model does, however, have one specific advantage, which is that it is formally nested within the multinomial logistic model. This is because (7.7) implies that

$$\log \frac{\pi^{(j)}}{\pi^{(1)}} = \alpha_*^{(j)} + \beta(j-1)X \tag{7.8}$$

where $\alpha_*^{(j)} = \alpha_2 + \cdots + \alpha_j$. Model (7.8) is equivalent to (6.7), with the regression coefficients constrained to obey the expression $\beta^{(j)} = (j-1)\beta$ for $j = 2, \ldots, C$. Because of this, we can use a likelihood ratio test to compare (7.7) to the standard multinomial logistic model, in order to assess the appropriateness of the assumption that β in (7.7) is the same across all j.

The *continuation-ratio logit model* is specified by

$$\log \frac{\pi^{(j)}}{\pi^{(j+1)} + \cdots + \pi^{(C)}} = \alpha^{(j)} + \beta X \quad \text{for } j = 1, \ldots, C-1. \tag{7.9}$$

What this amounts to is a set of models for a sequence of binary choices between response categories:

- category 1 rather than one of categories $(2, 3, \ldots, C)$
- 2 rather than one of $(3, 4, \ldots, C)$, given that it is not 1
- ... and so on, up to ...
- $C-1$ rather than C, given that it is not any of $1, 2, \ldots, C-2$.

For example, suppose that the response variable is the rank of an academic, classified in order of increasing rank as Assistant Professor, Associate Professor or Professor. The continuation-ratio model would then model two dichotomies, (i) Assistant Professor vs. higher, and (ii) Associate Professor vs. Professor, for individuals who are not Assistant Professors. In examples like this such a formulation as a sequence of binary choices may make substantive sense and be very useful for interpretation; in others, for instance the self-reported health example we have considered in this chapter, it is not very natural, in which case the continuation-ratio logit model itself is not particularly useful.

The description above also provides an explicit recipe for fitting the continuation-ratio logit model: It can be estimated as a set of distinct binary logistic models corresponding to this sequence of binary choices, each restricted to the appropriate subset of observations. We also note that for this model it is not particularly necessary to constrain the regression coefficients (e.g. β in (7.9)) to be the same for each $j = 1, \ldots, C-1$. If they are not, the model has the same number of parameters as a multinomial logistic model. Even then, however, (7.9) retains a strongly ordinal flavour, because its interpretation is so firmly based on the idea of an ordered sequence of choices between the categories.

Chapter 8

Further topics

Chapter 8 of the MY451 coursepack states that "the topics covered on this course have not quite exhausted the list of available statistical methods". You may or may not be thrilled to learn that this is still true after MY452. There are many statistical methods that are not regression models, and even the broad family of regression models contains many members which have not been considered on this course.

This final chapter briefly describes quantitative methods courses at the LSE that naturally follow MY452, and that are offered in LT. Each listing begins with the Calendar entry for the course in question, and then describes how the topics covered in the course build on the topics covered in MY452. Some of these courses are targeted for a statistical audience, so their mathematical level may be higher than in this course. This, however, should not deter you from exploring them. Much of statistical modelling makes use of the basic ideas discussed on this course in the context of linear and logistic models. You should thus be able to recognise the same aims and elements also in applications of other types of models.

- **MY455: Multivariate Analysis and Measurement**

  An introduction to the application of modern multivariate methods used in the social sciences, with particular focus on latent variable models for continuous observed variables, and their application to questions of measurement in the social sciences. At least the following topics will be covered: principal components analysis, exploratory factor analysis, confirmatory factor analysis and structural equation models. In addition, a selection from the following topics will be covered: cluster analysis, correspondence analysis, multidimensional scaling, latent class models, latent trait models.

  In MY452, all the regression models we consider have a single (scalar) response variable. MY455 considers models for *multivariate* response variables, and with the further twist that the explanatory variables in those models may be unobserved ("latent") variables. In a typical application the response variables are individual survey questions and the latent variables represent underlying constructs which the questions are designed to measure. These models are particularly common in psychological and behavioural research, where the aim is to examine the properties such measurement and models for the latent constructs.

- **MY456: Survey Methodology**

  This course provides an introduction to the methodology of the design and analysis of social surveys. It is intended both for students who plan to design and collect their own surveys, and for those who need to understand and use data from existing large-scale surveys. Topics covered include basic ideas of target populations, survey estimation and inference, sampling error and nonsampling error; sample design and sampling theory; methods of data collection; survey interviewing; cognitive processes in answering survey questions; design and evaluation of survey questions; nonresponse error and imputation for item nonresponse; survey weights; analysis of data from complex surveys; accessing, preparing and working with secondary data from existing social surveys. The course includes computer classes, using the statistical computer package Stata; no previous knowledge of Stata is required.

  In MY452, we treated all data as a simple random sample from the population we were interested in, and the data magically appeared on Moodle ready for our analysis. In practice, survey data are seldom simple random samples, and one often needs to design a new survey in order to have the data needed to answer a research question of interest, or to use and understand data from a survey collected by someone else. MY456 covers both the practice of survey design and implementation, and the additional statistical issues that arise in the analysis of survey data.

- **MY457: Causal Inference for Observational and Experimental Studies**

  This course provides an introduction to statistical methods used for causal inference in the social sciences. Using the potential outcomes framework of causality, topics covered include research designs such as randomized experiments and observational studies. We explore the impact of noncompliance in randomized experiments, as well as nonignorable treatment assignment in observational studies. To analyze these research designs, the methods covered include matching, instrumental variables, difference-in-difference, and regression discontinuity. Examples are drawn from different social sciences. The course includes computer classes, where standard statistical computer packages (Stata or R) are used for computation.

  In MY452, we very carefully avoided making *causal* claims from regression models. MY457 carefully considers the assumptions and data necessary to make such claims on the basis of quantitative data. The course considers a range of research designs that scholars have used to make causal claims more credible given various kinds of data.

- **ST416: Multilevel Modelling**

  A practical introduction to multilevel modelling with applications in social research. This course deals with the analysis of data from hierarchically structured populations (e.g. individuals nested within households or geographical areas) and longitudinal data (e.g. repeated measurements of individuals in a panel survey). Multilevel (random-effects) extensions of standard statistical techniques, including multiple linear regression and logistic regression, will be considered. The course will have an applied emphasis with computer sessions using appropriate software (e.g. Stata).

In MY452, the observations in our samples were treated as independent. ST416 considers models for data with multilevel or hierarchical structure: certain groups of units share some common attributes or exposures. As this is a Statistics Department course, it will use statistical notation and theory at somewhat higher level than MY452. However, every year students who have previously taken MY452 successfully complete ST416.

- **ST442: Longitudinal Data Analysis**

  A practical introduction to methods for the analysis of repeated measures data, including continuous and binary outcomes. Topics include: longitudinal study designs, models for two measures, (random effects) growth curve models, marginal models, dynamic (autoregressive) models, latent class models, and models for multivariate outcomes. The course will have an applied emphasis with fortnightly computer classes using the Stata software.

  In MY452, the observations in our samples were treated as independent. Like ST416, ST442 considers models where this assumption is abandoned. In the case of ST442, the lack of independence arises because the outcome data has a longitudinal (i.e. repeated-measures or time-series) structure: each unit has the outcome measured at multiple points in time. As this is a Statistics Department course, it will use statistical notation and theory at somewhat higher level than MY452. However, like ST416 it is also intended to be accessible for students who are comfortable with the material on MY452.