NIELS RICHARD HANSEN

REGRESSION

WITH R

UNIVERSITY OF COPENHAGEN

Preface

Disclaimer: This is a second draft. Chapters 1, 2 and 4 are relatively complete. Chapter 5 lacks some practical examples and a treatment of the practical computation of nonparametric likelihood intervals. Chapter 6 on survival analysis is a little raw. This draft was printed for the course Regression, University of Copenhagen, 2015.

This book was written as the textbook material for a graduate statistics course in regression analysis. The prerequisites include univariate linear regression, t-tests and constructions of standard confidence intervals, and knowledge of standard distributions such as the normal, the binomial and the Poisson distribution. The reader will also benefit from an introductory statistics courses covering likelihood methods, one- and two-sided analysis of variance, and aspects of asymptotic theory. In addition, a solid knowledge of linear algebra is assumed.

The exposition is mathematical, but the emphasis is on data modeling as well as theory. Mathematics is used to make model descriptions and assumptions precise, and to give precise results relevant for the practical analysis of data and the computations required for carrying out data analysis.

The book attempts to be complete and thorough on the topics covered, yet to be practical and relevant for the applied statistician. The means for achieving the latter is by larger case studies using R. The R code included is complete and covers most aspects of the data analysis from reading data into R, cleaning and plotting data to data analysis, data modeling and model diagnostics.

Splendid books such as *Regression Modeling Strategies*[1] and *The Elements of Statistical Learning*[2], which offers a plethora of predictive models and methods, were important inspirations. The present book is less ambitious in terms of breadth and more ambitious in terms of depth.

[1] Frank Harrell. *Regression Modeling Strategies*, Springer-Verlag New York, Inc., 2010

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, Springer, New York, 2009

Contents

1 Introduction

Regression modeling in this book

This book is on *descriptive and predictive regression modeling.* A regression model relates the distribution of a *response* variable to one or more *predictor variables.* All models considered are direct models of the conditional distribution of the response given the predictors, and a main purpose of such models is to be predictive of the response conditionally on observations of the predictors. There is no claim that this is the only purpose of modeling in general, but it is arguably important. The book treats linear models, generalized linear models and survival regression models. They are widely used in practice and obligatory parts of an education in statistics. At the same time they serve as good examples when developing general statistical principles and methodologies such as likelihood based methods. The intention of this book is to bridge the gap between a mathematical treatment of the model classes considered and their practical use.

In predictive modeling it is fairly clear what makes a good model. Given a method for quantification of predictive accuracy, the best model is the most accurate model. The accuracy of a prediction is typically quantified by a *loss* function, which actually quantifies how *inaccurate* a prediction is. The smaller the loss is the more accurate is the prediction. The expected loss quantifies how accurate the predictive model is on average. The specification of a relevant loss function is, perhaps, not always easy. A good loss function should ideally reflect the consequences of wrong predictions. On the other hand there exists a selection of useful, reasonable and convenient standard loss functions that can cope with many situations of practical interest. Examples include (weighted) squared error loss, the 0-1-loss and the negative log-likelihood loss. The book will not plunge into any elaborate discussions of the choice of loss but focus on how to fit models to data with standard choices.

Generalization is good, overfitting is bad.

One of the biggest challenges in practice is to select and fit a model to a data set in such a way that it will preserve its predictive accuracy when applied to new independent data. The model should be able to *generalize* well to cases not used for the model fitting and selection process. Otherwise the model has been overfitted to the data, and this is to be avoided. The book will develop a range of methods for fitting models to data, for assessing the model fit and for making appropriate model selections while avoiding overfitting.

It is important to understand that descriptive or predictive regression models do not in themselves provide information about *causal* relations between the response and the predictors. Such information can only be obtained if either the data comes from a randomized study, or if we are willing to make untestable assumptions. In the latter case, this will involve a thorough understanding of the subject matter field (which assumptions can we make?). Causal modeling is not just a technical question about distributional assumptions, but a structural modeling question that the data cannot reveal by themselves. It goes deeper than predictive modeling, and the techniques needed for causal modeling go beyond what is treated in this book. Typically they will involve some modeling of the distribution of the predictor variables as well.

With such words of warning about not making causal interpretations of predictive models, we should remind ourselves of the usefulness of predictive models. They are invaluable in automatized processes like spam filters or image and voice recognition. They make substantial contributions to medical diagnosis and prognosis, to business intelligence, to prediction of customer behavior, to risk prediction in insurance companies, pension funds and banks, to weather forecasts and to many other areas where it is of interest to know what we cannot (yet) observe.

Organization

The book consists of model chapters interspersed by methodology chapters, with the model chapters consisting of theory sections interspersed by real data modeling and data analysis. A decision was made that instead of providing a lot of small, simple and somewhat artificial data examples to illustrate a point, the relevant points are illustrated by larger real case studies. The hope is that this will ease the transition from theory to practice. The price to pay is that there are some distractions in forms of real data problems. Data rarely behaves well. There are missing observations and outliers, the models do not fit the data perfectly, the data comes with a strange encoding of variables and many other issues. Issues that require decisions to be made and issues on which many textbooks on statistical theory are silent.

By working through the case studies in detail it is the hope that many relevant practical problems are illustrated and appropriate solutions are given in such a way that the reader is better prepared to turn the theory into applications on her own.

The theory sections focus, on the other hand, on presenting the mathematical framework for regression modeling together with the relevant mathematical theory. In each model chapter the focus is on one particular model class, and the mathematical theory deals with derivations of likelihood functions, likelihood estimators, estimation algorithms and properties of likelihood estimators.

The methodology chapters treat general questions on regression methodology. This includes developing methods for the complex decision process that practical data analysis is. They also include a general treatment of parametric and nonparametric likelihood based statistical methods and methods for multimodel inference[1]. Finally, computational methodology, such as bootstrapping, is treated in the methodology chapters.

[1] How to sensibly and without overfitting select or average multiple model fits

The use of R

We use the programming language R[2] throughout to illustrate how good modeling strategies can be carried out in practice on real data. The book will not provide an introduction to the language though. Consult the R manuals[3] or the many introductory texts on R.

The case studies in this book are complete with R code that covers all aspects of the analysis. They represent an integration of data analysis in R with documentation in LaTeX. This is an adaptation to data analysis of what is known as *literate programming.* The main idea is that the writing of a report that documents the data analysis and the actual data analysis are merged into one document. This supports the creation of *reproducible analysis*, which is a prerequisite for reproducible research. To achieve this integration the book was written using the R package `knitr`[4]. The package supports alternative documentation formats, such as HTML or the simpler Markdown format, and they may be more convenient than LaTeX for day-to-day data analysis. An alternative to `knitr` is the `Sweave` function in the `utils` package (comes with the R distribution). The functionality of `knitr` is far greater than `Sweave` or any other attempt to improve upon `Sweave`, and `knitr` is thus recommended. The use of the RStudio[5] integrated development environment (IDE) is also recommended. The RStudio IDE is developed and distributed separately from R by RStudio Inc., but it is still open source and available for free.

Most figures in the book were produced using the `ggplot2` package[6] originally developed by Hadley Wickham and based on The Grammar of Graphics[7]. It is an extensive plotting and visualization system written in R. It is essentially a language within the language that allows you to specify how figures are plotted in a logical and expressive way. The package is well documented, see the web page or consult the book[8]. Occasionally, an alternative package, `lattice`, was used. There exists a nice series of blogs[9] recreating plots from the book *Lattice: Multivariate Data Visualization with R* using `ggplot2`.

Another classical resource worth mentioning is the influential book *Modern Applied Statistics with S*[10] and the corresponding `MASS` package (comes with the R distribution). Many classical sta-

[2] `www.r-project.org`

[3] R CORE TEAM. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013a

[4] `yihui.name/knitr/`

[5] `www.rstudio.com`

[6] `ggplot2.org`

[7] LELAND WILKINSON. *The grammar of graphics*, Springer, New York, 2005

[8] HADLEY WICKHAM. *ggplot2: elegant graphics for data analysis*, Springer New York, 2009

[9] `learnr.wordpress.com/2009/06/28/`

[10] W. N. VENABLES and B. D. RIPLEY. *Modern Applied Statistics with S*, Springer, New York, 2002

tistical models and methods are supported by this package and documented in the book. The MASS package is, furthermore, and by a wide margin the single package that most other packages depend upon (at the time of writing).

Once you have become a experienced user of R for data analysis (or perhaps earlier if you are a programmer) you will want to learn more about programming in R. Perhaps you want to develop your own functions, data structures or entire R packages. For package development the official manual[11] is an important resource. Another splendid resource is the book *Advanced R*[12]. It also freely available online[13]. This is a very well written and pedagogical treatment of R programming and software development.

To conclude this section we list (and load) all the R packages that are explicitly used in this book.

```r
library(ggplot2)   ## Grammar of graphics
library(reshape2)  ## Reshaping data frames
library(lattice)   ## More graphics
library(hexbin)    ## and more graphics
library(gridExtra) ## ... and more graphics
library(xtable)    ## LaTeX formatting of tables
library(splines)   ## Splines -- surprise :-)
library(survival)  ## Survival analysis
require(grid)      ## for 'unit'
```

[11] R Core Team. *Writing R Extensions*. R Foundation for Statistical Computing, Vienna, Austria, 2013b

[12] Hadley Wickham. *Advanced R*, Chapman Hall/CRC, 2014

[13] `adv-r.had.co.nz`

2

The linear model

The linear model is the classical workhorse of regression modeling. It is a tractable model with many desirable practical and theoretical properties. In combination with nonlinear variable transformations and basis expansions it can also be surprisingly flexible. And even if it is not the optimal model choice in a given situation, it can work well as a benchmark or fallback model whenever more sophisticated models are considered. This chapter introduces the linear model assumptions as well as the general regression modeling framework treated throughout. A case study on the modeling of birth weight is introduced early on. It illustrates many of the practical aspects of regression modeling, and covers initial exploratory data analysis, regression modeling and model diagnostics. The theory is kept to a minimum. The section *Estimation theory* covers how the linear model is fitted to data, and the section *Sampling theory* gives results on the sampling distributions of parameter estimates and test statistics that are used in relation to the linear model. The chapter concludes by revisiting the case study to show how basis expansions can be used to capture nonlinear relations within the framework of the linear model.

Model assumptions

This section briefly introduces the linear model and the typical assumptions made. This settles notation for the case study in the following section. In a first reading it can be read quickly and returned to later to better digest the model assumptions and their implications.

The linear model relates a continuous *response* variable Y to a p-dimensional vector X of *predictors*[1] via the identity

$$E(Y \mid X) = X^T\beta. \tag{2.1}$$

Here

$$X^T\beta = X_1\beta_1 + \ldots + X_p\beta_p$$

is a linear combination of the predictors weighted by the β-parameters. Thus the linear[2] model is a model of the conditional expectation of the response variable given the predictors.

An *intercept* parameter, β_0, is often added,

$$E(Y \mid X) = \beta_0 + X^T\beta.$$

It is notationally convenient to assume that the intercept parameter is included among the other parameters. This can be achieved by joining the predictor $X_0 = 1$ to X, thereby increasing the dimension to $p + 1$. In the general presentation we will not pay particular attention to the intercept. We will assume that if an intercept is needed, it is appropriately included among the other parameters, and we will index the predictors from 1 to p. Other choices of index set, e.g. from 0 to p, may be convenient in specific cases.

In addition to the fundamentel assumption (2.1) we will need two other model assumptions. For later reference we collect all three model assumptions here.

The assumption A2 is known as *homoskedasticity*, which is derived from the Greek words "homo" (same) and "skedastios" (dispersion). The opposite is *heteroskedasticity*.

A1 The conditional expectation of Y given X is

$$E(Y \mid X) = X^T\beta.$$

A2 The conditional variance of Y given X does not depend upon X,

$$V(Y \mid X) = \sigma^2.$$

[1] The X goes by many other names; explanatory variables, covariates, independent variables, regressors, inputs or features.

[2] The linearity that matters for statistics is the linearity in the unknown parameter vector β.

A3 The conditional distribution of Y given X is a normal distribution,

$$Y \mid X \sim \mathcal{N}(X^T\beta, \sigma^2).$$

It is, of course, implicit in A1 and A2 that Y has finite expectation and variance, respectively. It may not be obvious why Assumption A2 is needed, but it at least conceivable that A2 makes it easier to estimate the variance, since it doesn't depend upon X. The assumption has, furthermore, several consequences for the more technical side of the statistical analysis as well as the interpretation of the resulting model and the assessment of the accuracy of model predictions.

Assumption A3 is for many purposes unnecessarily restrictive. However, it is only under this assumption that a complete statistical theory can be developed. Some results used in practice are formally derived under this assumption, and they must thus be regarded as approximations when A3 is violated.

There exists a bewildering amount of terminology related to the linear model in the literature. Notation and terminology has been developed differently for different submodels of the linear model. If the X-vector only represents continuous variables, the model is often referred to as the linear *regression* model. Since any categorical variable on k levels can be encoded in X as k binary dummy variables, the linear model includes all ANalysis Of VAriance (ANOVA) models. Combinations, which are known in parts of the literature as ANalysis of COVAriance (ANCOVA), are of course also possible. These special cases and the special terminology, which is unnecessary, are most likely a consequence of historically different needs in different areas of applications. We give a unified treatment of the linear model. It is a fairly simple model with a rather complete theoretical basis. That said, many modeling questions still have to be settled in a practical data analysis, which makes applications of even the simple linear model non-trivial business.

The j'th dummy variable being 1 if the value of the categorical variable is the j'th level and 0 otherwise.

We need to introduce a couple of additional distributional assumptions. These are assumptions on the joint distribution of multiple observations. We assume that we have n observations, $Y_1, \ldots, Y_n$, of the response with corresponding predictors $X_1, \ldots, X_n$. We collect the responses into a column vector $\mathbf{Y}$, and we collect the predictors into an $n \times p$ matrix $\mathbf{X}$ called the *model matrix*. The *i*'th row of $\mathbf{X}$ is X_i^T. The additional assumptions are:

A4 The conditional distribution of Y_i given $\mathbf{X}$ depends upon X_i only, and Y_i and Y_j are conditionally *uncorrelated* given $\mathbf{X}$,

$$\mathrm{cov}(Y_i, Y_j \mid \mathbf{X}) = 0.$$

A5 The conditional distribution of Y_i given $\mathbf{X}$ depends upon X_i only, and $Y_1, \ldots, Y_n$ are conditionally *independent* given $\mathbf{X}$.

Assumption A4 implies together with A1 and A2 that

$$E(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}\beta, \tag{2.2}$$

and that

$$V(\mathbf{Y} \mid \mathbf{X}) = \sigma^2 \mathbf{I} \tag{2.3}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

Assumption A5 implies A4, and A5 and A3 imply that

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}). \tag{2.4}$$

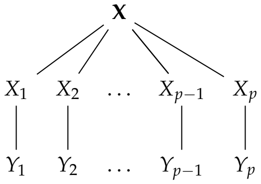


Figure 2.1: Graphical illustration of the assumptions on the joint distribution.

In summary, there are two sets of distributional assumptions. The weak set A1, A2 and A4, which imply the moment identities (2.2) and (2.3), and the strong set A3 and A5, which, in addition, imply the distributional identity (2.4).

The equation

$$Y_i = X_i^T \beta + \varepsilon_i.$$

defines ε_i – known as the *error* or *noise* term. For the linear model, the model assumptions can be formulated equally well in terms of ε_i; A1 is equivalent to $E(\varepsilon_i \mid X_i) = 0$, A2 to $V(\varepsilon_i \mid X) = \sigma^2$ and A3 is equivalent to $\varepsilon_i \mid X_i \sim \mathcal{N}(0, \sigma^2)$. As this conditional distribution does not depend upon X_i, assumption A3 implies that ε_i and X_i are independent.

It is quite important to realize that the model assumptions cannot easily be justified prior to the data analysis. There are no magic arguments or simple statistical summaries that imply that the assumptions are fulfilled. A histogram of the marginal distribution of the response Y can, for instance, not be used as an argument for or against Assumption A3 on the conditional normal distribution of Y. Justifications and investigations of model assumptions are done *after* a model has been fitted to data. This is called *model diagnostics*.

Birth weight – a case study

The question that we address in this case study is how birth weight of children is associated with a number of other observable variables. The data set comes from a sub-study of The Danish National Birth Cohort Study. The Danish National Birth Cohort Study was a nationwide study of pregnant women and their offspring[3]. Pregnant women completed a computer assisted telephone interview, scheduled to take place in pregnancy weeks 12-16 (for some, the interview took place later). We consider data on women interviewed before pregnancy week 17, who were still pregnant in week 17. One of the original purposes of the sub-study was to investigate if fever episodes during pregnancy were associated with fetal death.

We focus on birth weight as the response variable of interest. If Y denotes the birth weight of a child, the objective is to find a good predictive model of Y given a relevant set of predictor variables X. What we believe to be relevant can depend upon many things, for instance, that the variables used as predictors should be observable when we want to make a prediction. Causal mechanisms (known or unknown) may also be taken into account. If coffee happened to be a known course of preterm birth, and if we are interested in estimating a total causal effect of drinking coffee on the birth weight, we should not include the gestational age (age of fetus) at birth as a predictor variable. If, on the other hand, there are unobserved variables associated with coffee drinking as well as preterm birth, the inclusion of gestational age might give a more appropriate estimate of the causal effect of coffee. We will return to this discussion in subsequent sections – the important message being that the relevant set of predictors may very well be a subset of the variables in the data set.

First, we obtain the data set by reading it directly from the internet source.

[3] Jørn Olsen, Mads Melbye, Sjurdur F. Olsen, et al. The Danish National Birth Cohort - its background, structure and aim. *Scandinavian Journal of Public Health*, 29(4):300–307, 2001

```
pregnant <- read.table(
  "http://www.math.ku.dk/~richard/regression/data/pregnant.txt",
  header = TRUE,
  colClasses = c("factor", "factor", "numeric", "factor", "factor",
                 "integer", "factor", "numeric", "factor", "numeric",
                 "numeric", "integer")
)
```

Mistakes are easily made if the classes of the columns in the data frame are not appropriate.

The standard default for **read.table** is that columns containing characters are converted to factors. This is often desirable. Use the **stringsAsFactors** argument to **read.table** or set the global option **stringsAsFactors** to control the conversion of characters. Categorical variables encoded as integers or other numeric values, as in the present data set, are, however, turned into **numeric** columns, which is most likely not what is desired. This is the reason for the explicit specification of the column classes above.

It is always a good idea to check that the data was read correctly, that the columns of the resulting data frame have the correct names and are of the correct class, and to check the dimensions of the resulting data frame. This data set has 12 variables and 11817 cases.

```
head(pregnant, 4)

##   interviewWeek fetalDeath   age abortions children gestationalAge
## 1            14          0 36.73         0        1             40
## 2            12          0 34.99         0        1             41
## 3            13          1 33.70         0        0             35
## 4            16          0 33.06         0        1             38
##   smoking alcohol coffee length weight feverEpisodes
## 1       1       0      1     NA     NA             0
## 2       3       2      2     53   3900             2
## 3       1       0      1     NA     NA             0
## 4       1       4      2     48   2800             0
```

Note that there are missing observations represented as NA. One explanation of missing length and weight observations is fetal death.

Descriptive summaries

The first step is to summarize the variables in the data set using simple descriptive statistics. This is to get an idea about the data and the variable ranges, but also to discover potential issues that we need to take into account in the further analysis. The list of issues we should be aware of includes, but is not limited to,

- extreme observations and potential outliers,

- missing values
- and skewness or asymmetry of marginal distributions.

Anything worth noticing should be noticed. It should not necessarily be written down in a final report, but figures and tables should be prepared to reveal and not conceal.

A quick summary of the variables in a data frame can be obtained with the `summary` function. It prints quantile information for numeric variables and frequencies for factor variables. This is the first example where the class of the columns matter for the result that R produces. Information on the number of missing observations for each variable is also given.

```
summary(pregnant)

##  interviewWeek  fetalDeath        age       abortions children
##  14     :2379   0    :11659   Min.   :16.3   0:9598    0:5304
##  15     :2285   1    :  119   1st Qu.:26.6   1:1709    1:6513
##  16     :2202   NA's:   39   Median :29.5   2: 395
##  13     :2091                 Mean   :29.6   3: 115
##  12     :1622                 3rd Qu.:32.5
##  11     :1089                 Max.   :44.9
##  (Other): 149
##  gestationalAge smoking     alcohol        coffee         length
##  Min.   :17.0   1:8673   Min.   : 0.000   1    :7821   Min.   : 0.0
##  1st Qu.:39.0   2:1767   1st Qu.: 0.000   2    :3624   1st Qu.:51.0
##  Median :40.0   3:1377   Median : 0.000   3    : 368   Median :52.0
##  Mean   :39.4            Mean   : 0.512   NA's:    4   Mean   :51.8
##  3rd Qu.:41.0            3rd Qu.: 1.000                3rd Qu.:54.0
##  Max.   :47.0            Max.   :15.000                Max.   :99.0
##                          NA's   :1                     NA's   :538
##      weight     feverEpisodes
##  Min.   :   0   Min.   : 0.0
##  1st Qu.:3250   1st Qu.: 0.0
##  Median :3600   Median : 0.0
##  Mean   :3572   Mean   : 0.2
##  3rd Qu.:3950   3rd Qu.: 0.0
##  Max.   :6140   Max.   :10.0
##  NA's   :538
```

Further investigations of the marginal distributions of the variables in the data set can be obtained by using histograms, density estimates, tabulations and barplots. Barplots are preferable over histograms for numeric variables that take only a small number of different values, e.g. counts. This is the case for the `feverEpisodes` variable. Before such figures and tables are produced – or perhaps

`interviewWeek`: Pregnancy week at interview.

`fetalDeath`: Indicator of fetal death (1 = death).

`age`: Mother's age at conception in years.

`abortions`: Number of previous spontaneous abortions (0, 1, 2, 3+).

`children`: Indicator of previous children (1 = previous children).

`gestationalAge`: Gestational age in weeks at end of pregnancy.

`smoking`: Smoking status; 0, 1–10 or 11+ cigs/day encoded as 1, 2 or 3.

`alcohol`: Number of weekly drinks during pregnancy.

`coffee`: Coffee consumption; 0, 1–7 or 8+ cups/day encoded as 1, 2 or 3.

`length`: Birth length in cm.

`weight`: Birth weight in gram.

`feverEpisodes`: Number of mother's fever episodes before interview.

Table 2.1: The 12 variables and their encoding in the pregnant data set.

after they have been produced once, but before they enter a final report – we may prefer to clean the data a little. We can observe from the summary information above that for some cases weight or length is registered as 0 – and in some other cases weight or length is found to be unrealistically small – which are most likely registration mistakes. Likewise, some lengths are registered as 99, and further scrutiny reveals an observation with `weight` 3550 gram with `gestationalAge` registered as 18. We exclude those cases from the subsequent analysis.

For convenience, `disVar` and `contVar` are the variables that will be summarized as discrete or as continuous variables, respectively.

```
pregnant <- subset(pregnant,
                   weight > 32 & length > 10 & length < 99 &
                     gestationalAge > 18,
                   select = -c(interviewWeek, fetalDeath))
disVar <- sapply(pregnant, class) == "factor"
contVar <- names(which(!disVar))[-6]  ## Excluding 'feverEpisodes'
disVar <- c(names(which(disVar)), "feverEpisodes")
```

We present density estimates of the 5 continuous variables, see Figure 2.2. The density estimates, as the majority of the figures presented in this book, were produced using the `ggplot2` package. Readers familiar with ordinary R graphics can easily produce histograms with the `hist` function or density estimates with the `density` function. For this simple task, the `qplot` (for quick plot) and the general `ggplot` functions do not offer much of an advantage – besides the fact that figures have the same style as other `ggplot2` figures. However, the well-thought-out design and entire functionality of `ggplot2` has resulted in plotting methods that are powerful and expressive. The benefit is that with `ggplot2` it is possible to produce quite complicated figures with clear and logical R expressions – and without the need to mess with a lot of low-level technical plotting details.

What is most noteworthy in Figure 2.2 is that the distribution of `alcohol` is extremely skewed, with more than half of the cases not drinking alcohol at all. This is noteworthy since little variation in a predictor makes it more difficult to detect whether it is associated with the response.

See `?melt.data.frame` on the `melt` method for data frames from the `reshape2` package.

```
mPregnant <- melt(pregnant[, contVar])
qplot(value, data = mPregnant, geom = "density", adjust = 2,
      fill = I(gray(0.5)), xlab = "", ylab = "") +
  facet_wrap(~ variable, scales = "free", ncol = 6)
```

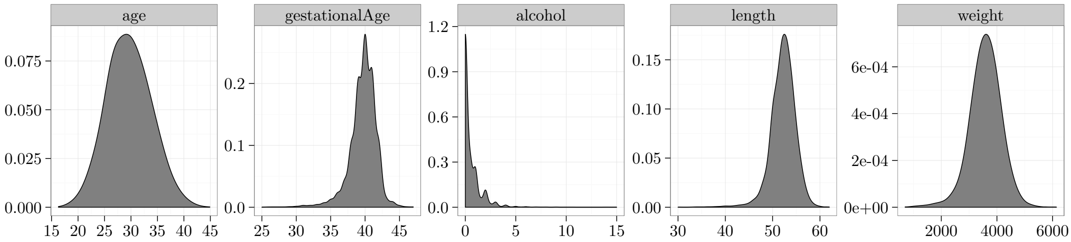


Figure 2.2: Density esti-

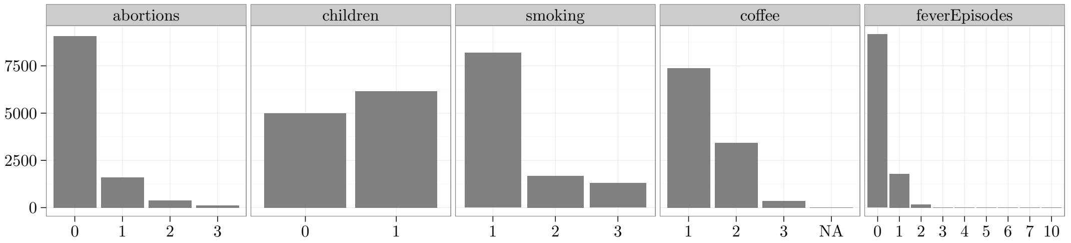


Figure 2.3: Barplots of discrete variables.

For the discrete variables – the categorical or count variables – we produce barplots instead of density estimates. Figure 2.3 shows that all discrete variables except `children` have quite skewed distributions.

In summary, the typical pregnant woman does not smoke or drink alcohol or coffee, nor has she had any previous spontaneous abortions or any fever episodes. About one-third drinks coffee or alcohol or smokes. These observations may not be surprising – they reflect what is to be expected for a random sample of cases. Little variation of a predictor can, however, make estimation and detection of associations between the response and the predictors more difficult. In this case the data set is quite large, and cases with the least frequently occurring predictor values are present in resonable numbers.

```
mPregnant <- melt(pregnant[, disVar], id.var = c())
qplot(factor(value, levels = 0:10), data = mPregnant, geom = "bar",
      fill = I(gray(0.5)), xlab = "", ylab = "") +
  facet_wrap(~ variable, scales = "free_x", ncol = 5)
```

The coercion of `value` to `factor` is needed to get the order of the levels correct.

Pairwise associations

The next step is to investigate associations of the variables. We are still not attempting to build a predictive model and the response does not yet play a special role. One purpose is again to get better acquainted with the data – this time by focusing on covariation – but there is also one particular issue that we should be aware of.

- Collinearity of predictors.

Add this bullet point to the previous list of issues. If two or more predictors in the data set are strongly correlated, they contain, from a predictive point of view, more or less the same information, but perhaps encoded in slightly different ways. Strongly correlated predictors result in the same problem as predictors with little variation. It can become difficult to estimate and detect joint association of the predictors with the response. A technical consequence is that statistical tests of whether one of the predictors could be excluded become non-significant if the other is included, whereas a test of joint exclusion of the predictors can be highly significant. Thus it will become difficult to determine on statistical grounds if one predictor should be included over the other. It is best to know about such potential issues upfront. Perhaps it is, by subject matter arguments, possible to choose one of the predictors over the other as the most relevant to include in the model.

A scatter plot matrix is a useful graphical summary of the pairwise association of continuous variables. It can be supplemented with computations of Pearson correlations.

Function **cor.print** formats the correlations for printing. The **na.omit** function removes cases containing missing observations. This is needed here to get the estimated correlations computed.

```
cor.print <- function(x, y) {
  panel.text(mean(range(x)), mean(range(y)),
             paste('$', round(cor(x, y), digits = 2), '$', sep = '')
             )
}

splom(na.omit(pregnant)[, contVar], xlab = "",
      upper.panel = panel.hexbinplot,
      pscales = 0, xbins = 20,
      varnames = c("age", "gest. age", contVar[3:5]),
      lower.panel = cor.print
)
```

The scatter plots, Figure 2.4, show that `length` and `weight` are (not surprisingly) very correlated, and that both of these variables

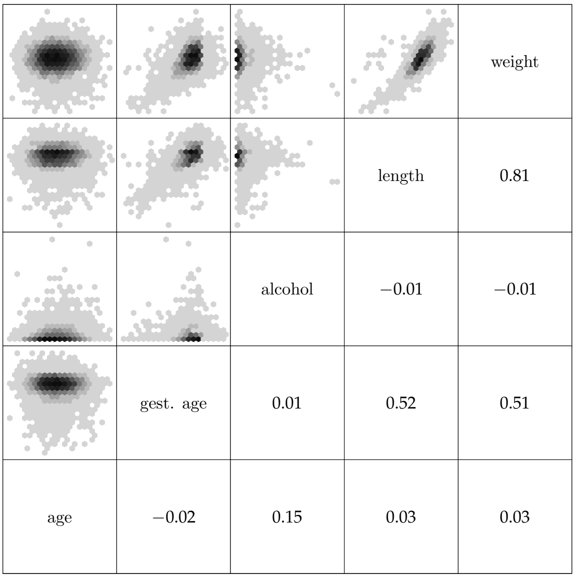


Figure 2.4: Scatter plot matrix of the continuous variables and corresponding Pearson correlations.

are also highly correlated with `gestationalAge`. The `alcohol` and `age` variables are mildly correlated, but they are virtually uncorrelated with the other three variables.

The scatter plot matrix was produced using the `splom` function from the `lattice` package. The data set is quite large and a naive plot of a scatter plot matrix results in a lot of overplotting and huge pdf graphics files. Figures can be saved as high-resolution png files instead of pdf files to remedy problems with file size. The actual plotting may, however, still be slow, and the information content in the plot may be limited due to the overplotting. A good way to deal with overplotting is to use hexagonal binning of data points. This was done using the `panel.hexbinplot` function from the `hexbin` package together with the `splom` function.

If two categorical variables are strongly dependent the corresponding vectors of dummy variable encoding of the categorical levels will be collinear. In extreme cases where only certain pairwise combinations of the categorical variables are observed, the resulting dummy variable vectors will be perfectly collinear. Dependence between categorical variables may be investigated by cross-tabulation and formal tests of independence can, for instance, be computed. However, neither test statistics of independence nor corresponding p-values are measures of the degree of dependence – they scale with the size of the data set and become more and more extreme for larger data sets. There is no single suitable substitute for the Pearson correlation that applies to categorical variables in general. In this particular example all the categorical variables are, in fact, ordinal. In this case we can use the Spearman correlation. The Spearman correlation is simply the Pearson correlation between the ranks of the observations. Since we only need to be able to sort observations to compute ranks, the Spearman correlation is well defined for ordinal as well as continuous variables.

```r
cp <- cor(data.matrix(na.omit(pregnant)), method = "spearman")
ord <- rev(hclust(as.dist(1-abs(cp)))$order)
colPal <- colorRampPalette(c("blue", "yellow"), space = "rgb")(100)
levelplot(cp[ord, ord],  xlab = "", ylab = "",
          col.regions = colPal, at = seq(-1, 1, length.out = 100),
          colorkey = list(space = "top", labels = list(cex = 1.5)),
          scales = list(x = list(rot = 45),
                        y = list(draw = FALSE),
                        cex = 1.2))
```

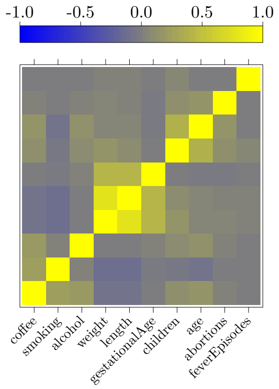


Figure 2.5: Spearman correlation matrix. Variables are ordered according to a hierarchical clustering.

Figure 2.5 shows Spearman correlations of all variables – categorical as well as continuous. For continuous variables the Spearman correlation is, furthermore, invariant to monotone transformations and less sensitive to outliers than the Pearson correlation. These properties make the Spearman correlation more attractive as a means for exploratory investigations of pairwise association.

For the production of the plot of the correlation matrix, Figure 2.5, we used a hierarchical clustering of the variables. The purpose was to sort the variables so that the large correlations are concentrated around the diagonal. Since there is no natural order of the variables, the correlation matrix could be plotted using any order. We want to choose an order that brings highly correlated variables close together to make the figure easier to read. Hierarchical clustering can be useful for this purpose. For the clustering, a dissimilarity measure between variables is needed. We used 1 minus the absolute value of the correlation. It resulted in a useful ordering in this case.

What we see most clearly from Figure 2.5 are three groupings of positively correlated variables. The `weight`, `length` and `gesta-`

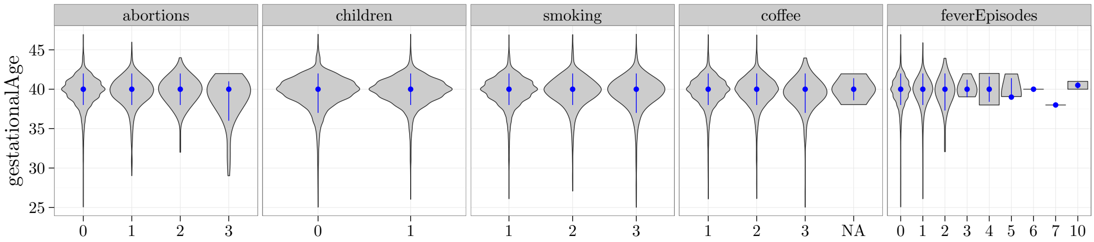


Figure 2.6: Violin plots, medians and interdecile ranges for the distribution of `gestationalAge`. Note that there are very few observations with many fever episodes.

tionalAge group, a group consisting of **age**, **children** and **abortions** (not surprising), and a grouping of **alcohol**, **smoking** and **coffee** with mainly coffee being correlated with the two others.

An alternative way to study the relation between a continuous and a categorical variable is to look at the distribution of the continuous variable stratified according to the values of the categorical variable. This can be done using violin plots.

```
mPregnant <- melt(pregnant[, c("gestationalAge", disVar)],
                       id = "gestationalAge")
deciles <- function(x) {
  quan <- quantile(x, c(0.1, 0.5, 0.9))
  data.frame(ymin = quan[1], y = quan[2], ymax = quan[3])
}
ggplot(mPregnant,
       aes(x = factor(value, levels = 0:10), y = gestationalAge)) +
  geom_violin(scale = 'width', adjust = 2, fill = I(gray(0.8))) +
  stat_summary(fun.data = deciles, color = "blue") + xlab("") +
  facet_wrap(~ variable, scale = "free_x", ncol = 5)
```

The `deciles` function is used to add median and decile information to the violin plots.

A violin plot can be seen as an alternative to a boxplot, and it is easy to produce with `ggplot2`. It is just a rotated kernel density estimate.

Figure 2.6 shows violin plots of `gestationalAge` stratified according to the discrete variables. The violin plots have been supplemented with median and interdecile range information. The figure shows that there is no clear relation between `gestationalAge` and the other variables. This concurs with the information in Figure 2.5. Figure 2.7 shows a similar violin plot but this time with the continuous variable being the response variable `weight`. From this figure we observe that `weight` seems to be larger if the mother has had children before and to be negatively related to coffee drinking and smoking.

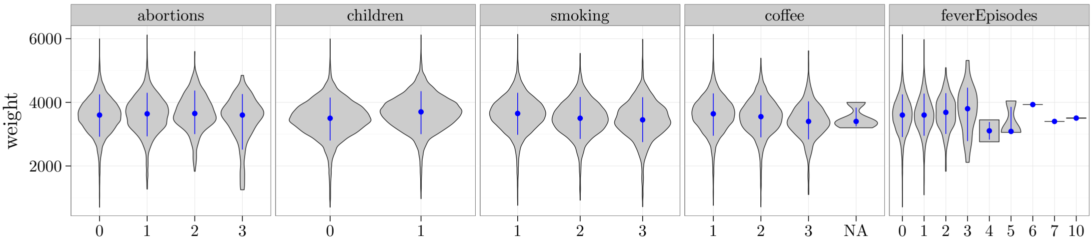


Figure 2.7: Violin plots, medians and interdecile ranges of `weight`.

A linear regression model

To build a linear regression model of the response variable `weight`, we need to decide which of the predictors we want to include. We also need to decide if we want to include the predictor variables as is, or if we want to transform them. Before we make any of these decisions we explore linear regression models where we just include one of the predictors at a time. This analysis is not to be misused for variable selection, but to supplement the explorative studies from the previous sections. In contrast to correlation considerations this procedure for studying single predictor association with the response can be generalized to models where the response is discrete.

```
form <- weight ~ gestationalAge + length + age + children +
  coffee + alcohol + smoking + abortions + feverEpisodes
pregnant <- na.omit(pregnant)
nulModel <- lm(weight ~ 1, data = pregnant)
oneTermModels <- add1(nulModel, form, test = "F")
```

Table 2.2: Marginal association tests sorted according to the p-value.

	Df	Sum of Sq	RSS	F value	Pr(>F)
length	1	2.352e+09	1.262e+09	20774.87	0
gestationalAge	1	9.323e+08	2.682e+09	3876.54	0
children	1	9.762e+07	3.516e+09	309.55	2.29e−68
smoking	2	5.332e+07	3.561e+09	83.48	1.03e−36
coffee	2	2.199e+07	3.592e+09	34.13	1.67e−15
abortions	3	6.273e+06	3.608e+09	6.46	0.000229
age	1	3.954e+06	3.610e+09	12.21	0.000476
feverEpisodes	1	1.086e+06	3.613e+09	3.35	0.0672
alcohol	1	1.700e+05	3.614e+09	0.52	0.469

Table 2.2 shows the result of testing if inclusion of each of the predictors by themselves is significant. That is, we test the model with only an intercept against the alternative where a single predic-

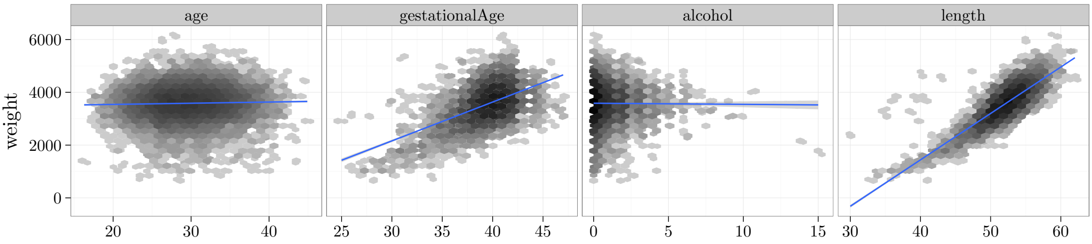


Figure 2.8: Scatter plots including linear regression line.

tor is included. The test used is the F-test – see the next section, page 36, for details on the theory. For each of the categorical predictor variables the encoding requires Df (degrees of freedom) dummy variables in addition to the intercept to encode the inclusion of a variable with Df + 1 levels.

Figure 2.8 shows the scatter plots of `weight` against the 4 continuous predictors. This is just the first row in the scatter plot matrix in Figure 2.4, but this time we have added the linear regression line. For the continuous variables the tests reported in Table 2.2 are tests of whether the regression line has slope 0.

```
mPregnant <- melt(pregnant[, contVar],
                  id.vars = "weight")
binScale <- scale_fill_continuous(breaks = c(1, 10, 100, 1000),
                                  low = "gray80", high = "black",
                                  trans = "log", guide = "none")
qplot(value, weight, data = mPregnant, xlab = "", geom = "hex") +
  stat_binhex(bins = 25) + binScale +
  facet_wrap(~ variable, scales = "free_x", ncol =  4) +
  geom_smooth(size = 1, method = "lm")
```

To decide upon the variables to include in the first multivariate linear model, we summarize some of the findings of the initial analyses. The `length` variable is obviously a very good predictor of `weight`, but it is also close to being an equivalent “body size” measurement, and it will be affected in similar ways as `weight` by variables that affect fetus growth. From a predictive modeling point of view it is in most cases useless, as it is will not be observable unless `weight` is also observable. The `gestationalAge` variable is likewise of little interest if we want to predict `weight` early in pregnancy. The variable is, however, virtually unrelated to the other predictors, and age of the fetus at birth is a logic cause of the weight of the child. It could also be a relevant predictor late in pregnancy for predicting the weight if the woman were to give birth at a given time. Thus we keep `gestationalAge` as a predictor. The remaining predictors are not strongly correlated, and we have not found reasons to exclude any of them. We will thus fit a main effects linear model with 8 predictors. We include all the predictors as they are.

The main effects model.

```
form <- update(form, . ~ . - length)
pregnantLm <- lm(form, data = pregnant)
summary(pregnantLm)
```

Table 2.3: Summary table of parameter estimates, standard errors and t-tests for the linear model of weight fitted with 8 predictors.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−2169.44	98.60	−22.00	4.6e−105
gestationalAge	145.16	2.30	63.01	0
age	−2.00	1.20	−1.66	0.097
children1	185.95	9.90	18.79	1.5e−77
coffee2	−65.54	10.39	−6.31	2.9e−10
coffee3	−141.78	27.24	−5.20	2e−07
alcohol	−2.75	5.09	−0.54	0.59
smoking2	−101.95	13.05	−7.81	6.1e−15
smoking3	−131.19	14.91	−8.80	1.6e−18
abortions1	27.84	13.09	2.13	0.033
abortions2	48.76	25.45	1.92	0.055
abortions3	−50.03	45.80	−1.09	0.27
feverEpisodes	6.36	9.39	0.68	0.5

Table 2.3 shows the estimated β-parameters among other things. Note that all categorical variables (specifically, those that are encode as factors in the data frame) are included via a dummy variable representation. The precise encoding is determined by a linear constraint, known as a *contrast*. By default, the first factor level is constrained to have parameter 0, in which case the remaining parameters represent differences to this base level. In this case it is only occasionally of interest to look at the t-tests for testing if a single parameter is 0. Table 2.4 shows instead F-tests of excluding any one of the predictors. It shows that the predictors basically fall into two groups; the strong predictors `gestationalAge`, `children`, `smoking` and `coffee`, and the weak predictors `abortions`, `age`, `feverEpisodes` and `alcohol`. The table was obtained using the `drop1` function. We should at this stage resist the temptation to use the tests for a model reduction or model selection.

```
drop1(pregnantLm, test = "F")
```

	Df	Pr(>F)
gest. Age	1	0
children	1	1.5e−77
smoking	2	5.6e−26
coffee	2	5.2e−13
abortions	3	0.028
age	1	0.097
feverEpisodes	1	0.5
alcohol	1	0.59

Table 2.4: Tests of excluding each term from the full model.

Model diagnostics are then to be considered to justify the model assumptions. Several aspects of the statistical analysis presented so far rely on these assumptions, though the theory is postponed to the subsequent sections. Most notably, the distribution of the test statistics, and thus the p-values, depend on the strong set of assumptions, A3 + A5. We cannot hope to prove that the assumptions are fulfilled, but we can check – mostly using graphical methods – that they are either not obviously wrong, or if they appear to be wrong, the methods should reveal how we can improve the model.

Model diagnostics for the linear model are mostly based on the residuals, which are estimates of the unobserved errors ε_i, or the *standardized* residuals, which are estimates of ε_i/σ. Plots of the standardized residuals against the fitted values, or against any one of the predictors, are useful to detect deviations from A1 or A2. For A3 we consider a qq-plot against the standard normal distribution. The assumptions A4 or A5 are more difficult to investigate. If we don't have a specific idea about how the errors, and thus the observations, might be correlated, it is very difficult to do anything.

```
pregnantDiag <- fortify(pregnantLm)

p1 <- qplot(.fitted, .stdresid, data = pregnantDiag, geom = "hex") +
  binScale + geom_smooth(size = 1) +
  xlab("fitted values") + ylab("standardized residuals")
p2 <- qplot(gestationalAge, .stdresid, data = pregnantDiag,
            geom = "hex") + binScale +
  stat_binhex(bins = 25) + geom_smooth(size = 1) +
  xlab("gestationalAge") + ylab("")
p3 <- qplot(sample = .stdresid, data = pregnantDiag, stat = "qq") +
  geom_abline(intercept = 0, slope = 1, color = "blue", size = 1) +
  xlab("theoretical quantiles") + ylab("")
grid.arrange(p1, p2, p3, ncol = 3)
```

Why not use the plot method for `lm`-objects? That may be useful for quick interactive usage, but the resulting plot is difficult to customize for publication quality.

The residual plot in Figure 2.9 shows that the model is not spot on. The plot of the residuals against `gestationalAge` shows that there is a non-linear effect that the linear model does not catch. Thus A1 is not fulfilled. We address this specific issue in a later section, where we solve the problem using splines. The qq-plot

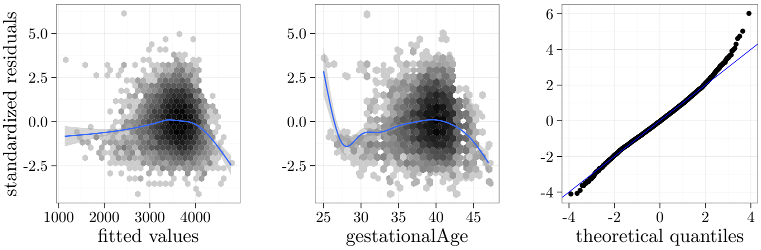


Figure 2.9: Diagnostic plots. Standardized residuals plotted against fitted values, the predictor `gestationalAge`, and a qq-plot against the normal distribution.

shows that the tails of the residuals are heavier than the normal distribution an right skewed. However, given the problems with A1, this issue is of secondary interest.

The diagnostics considered above address if the data set as a whole does not comply to the model assumptions. Single observations can also be extreme and, for instance, have a large influence on how the model is fitted. For this reason we should also be aware of single extreme observations in the residual plots and the qq-plot.

INTERACTIONS between the different predictors can then be considered. The inclusion of interactions results in a substantial increase in the complexity of the models, even if we have only a few predictors. Moreover, it becomes possible to construct an overwhelming number of comparisons of models. Searching haphazardly through thousands of models with various combinations of interactions is not recommended. It will result in spurious discoveries that will be difficult to reproduce in other studies. Instead, we suggest to focus on the strongest predictors from the main effects model. It is more likely that we are able to detect interactions between strong predictors than between weak predictors. To comprehend an interaction model it is advisable to visualize the model to the extend it is possible. This is a point where the `ggplot2` package is really strong. It supports a number of ways to stratify a plot according to different variables.

```
qplot(gestationalAge, weight, data = pregnant, geom = "hex") +
  facet_grid(coffee ~ children + smoking, label = label_both) +
  binScale + stat_binhex(bins = 25) +
  geom_smooth(method = "lm", size = 1, se = FALSE)
```

Figure 2.10 shows a total of 18 scatter plots where the stratifica-

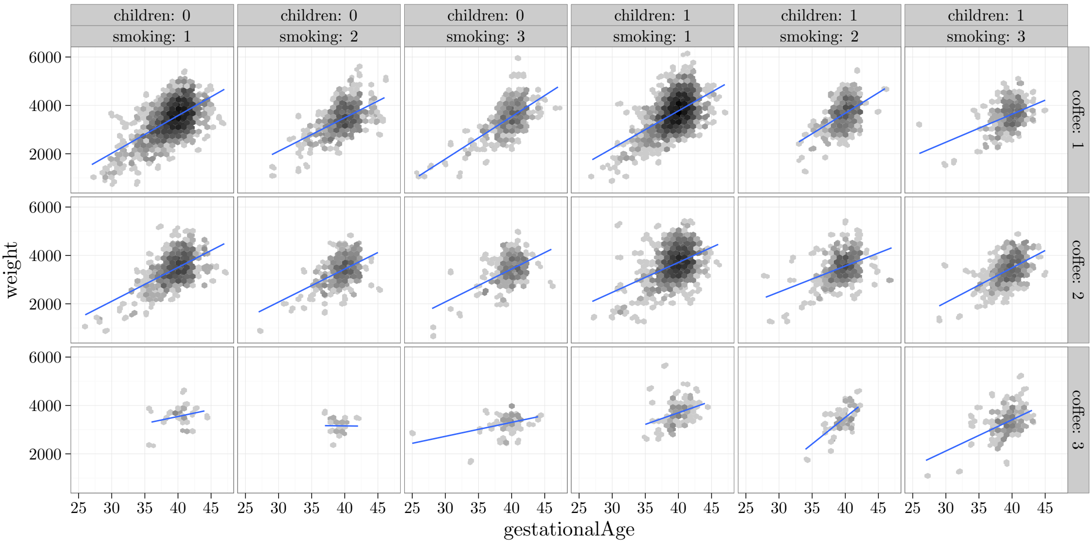


Figure 2.10: Scatter plots of **weight** against **gestationalAge** stratified according to the values of **smoking**, **children** and **coffee**

tion is according to **children**, **smoking** and **coffee**. A regression line was fitted separately for each plot. This corresponds to a model with a third order interaction between the 4 strong predictors (and with the weak predictors left out). Variations between the regression lines are seen across the different plots, which is an indication of interaction effects. For better comparison of the regression lines it can be beneficial to plot them differently. Figure 2.11 shows an example where the stratification according to **coffee** is visualized by color coding the levels of **coffee**. We can test the model with a third order interaction between the strong predictors against the main effects model. In doing so we keep the weak predictors in the model.

```r
form <- weight ~ smoking * coffee * children * gestationalAge +
  age + alcohol + abortions + feverEpisodes
pregnantLm2 <- lm(form, data = pregnant)
anova(pregnantLm, pregnantLm2)
```

```r
ggplot(pregnant, aes(gestationalAge, weight, color = coffee)) +
  facet_grid(. ~ children + smoking, label = label_both) +
  geom_smooth(method = "lm", size = 1, se = FALSE)
```

Table 2.5 shows that the *F*-test of the full third order interaction model against the main effects model is clearly significant. Since

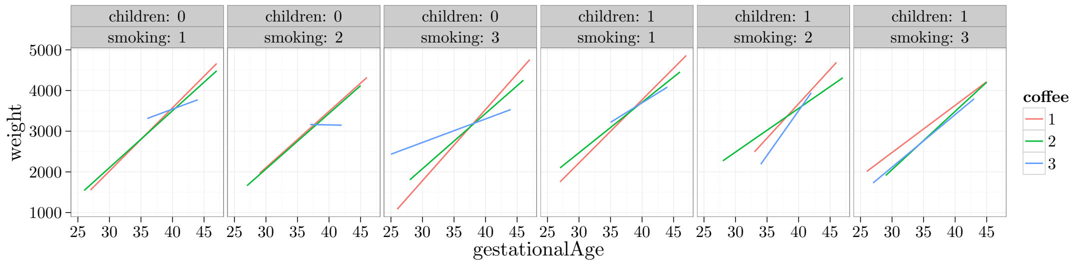


Figure 2.11: Comparison of estimated regression lines for **gestationalAge** stratified according to the values of **smoking**, **coffee** and **children**

there is some lack of model fit, we should be skeptical about the conclusions from formal hypothesis tests. However, deviations from A1 result in an increased residual variance, which will generally result in more conservative tests. That is, it will become harder to reject a null hypothesis, and thus, in this case, conclude that inclusion of the interactions is significant. The third order interaction model contains 42 parameters, so a full table of all the parameters is not very comprehensible, and it will thus not be reported.

Table 2.5: Test of the model including a third order interaction against the main effects model.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11139	2.5376e+09				
2	11110	2.5143e+09	29	2.3341e+07	3.56	3.49e−10

We reconsider model diagnostics for the extended model, where we have included the interactions. Figure 2.12 shows the residual plot. The inclusion of the interactions did not solve the earlier observed problems with the model fit. This is hardly surprising as the problem with the model appears to be related to a non-linear relation between `weight` and `gestationalAge`. Such an apparent nonlinearity could be explained by interaction effects, but this would require a strong correlation between the predictors, e.g. that heavy coffee drinkers (`coffee = 3`) have large values of `gestationalAge`. We already established that this was not the case.

```
pregnantDiag2 <- fortify(pregnantLm2)
qplot(.fitted, .stdresid, data = pregnantDiag2, geom = "hex") +
  binScale + geom_smooth(size = 1) +
  xlab("fitted values") + ylab("standardized residuals")
```

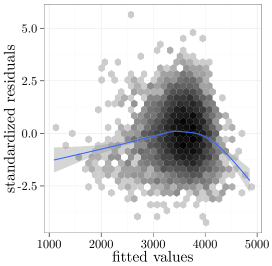


Figure 2.12: Residual plot for the third order interaction model.

Before we conclude the analysis, we test if the inclusion of the 4 weak predictors together is necessary. Table 2.6 shows that the test results in a borderline p-value of around 5%. On the basis of this we choose to exclude the 4 weak predictors even though Table 2.4 suggested that the number of abortions is related to `weight`. The highly skewed distribution of `abortions` resulted in large standard errors, and low power despite the size of the data set. In combination with the different signs on the estimated parameters in Table 2.3, depending upon whether the woman had had 1, 2 or 3+ spontaneous abortions, the study is inconclusive on how `abortions` is related to `weight`.

```
form <- weight ~ smoking * coffee * children * gestationalAge
pregnantLm3 <- lm(form, data = pregnant)
anova(pregnantLm3, pregnantLm2)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11116	2.5172e+09				
2	11110	2.5143e+09	6	2.8727e+06	2.12	0.04825

Table 2.6: Test of the full third order interaction model against the model excluding the 4 weak predictors.

In conclusion, we have arrived at a predictive model of `weight` given in terms of a third order interaction of the 4 predictors `gestationalAge`, `smoking`, `coffee` and `children`. The model is not a perfect fit, as it doesn't catch a nonlinear relation between `weight` and `gestationalAge`. The fitted model can be visualized as in the Figures 2.10 or 2.11. We note that the formal F-test of the interaction model against the main effects model justifies the need for the increased model complexity. It is, however, clear from the figures that the actual differences in slopes are small, and the significance of the test reflects that we have a large data set. There is no clear-cut interpretation of the interactions either. The regression lines in the figures should, preferably, be equipped with confidence bands. This can be achieved by removing the `se = FALSE` argument to the `geom_smooth` function. However, this will result in a separate variance estimate for each combination of `smoking`, `coffee` and `children`. If we want to use the pooled variance estimate obtained by our model, we have to do something else. How this is achieved is shown in a later section, where we also consider how to deal with the nonlinearity using spline basis expansions.

Estimation theory

The theory that we will cover in this section is on the estimation of the unknown parameter β using least squares methods. We give theoretical results on the existence and uniqueness, and we provide characterizations of the least squares estimator. We also discuss how the estimator is computed in practice.

Weighted linear least squares estimation

We will consider the generalization of linear least squares that among other things allows for weights on the individual cases. Allowing for weights can be of interest in itself, but serves, in particular, as a preparation for the methods we will consider in Chapter 4.

We introduce the *weighted squared error loss*[4] as

$$\ell(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta) \tag{2.5}$$

where $\mathbf{W}$ is a positive definite matrix. An $n \times n$ matrix is positive definite if it is symmetric and

$$\mathbf{y}^T \mathbf{W} \mathbf{y} > 0$$

for all $\mathbf{y} \in \mathbb{R}^n$ with $\mathbf{y} \neq 0$. A special type of positive definite weight matrix is a diagonal matrix with positive entries in the diagonal. With

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & w_n \end{pmatrix}$$

we find that the weighted squared error loss becomes

$$\ell(\beta) = \sum_i w_i (Y_i - X_i^T \beta)^2.$$

That is, the i'th case receives the weight w_i.

The parameter β is estimated by minimization of ℓ.

Theorem 2.1. *If* $\mathbf{X}$ *has full column rank* p*, the unique solution of the normal equation*

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \beta = \mathbf{X}^T \mathbf{W} \mathbf{Y} \tag{2.6}$$

is the unique minimizer of ℓ.

[4] With $\mathbf{W} = \mathbf{I}$ this loss is proportional to the negative log-likelihood loss under assumptions A3 and A5 as derived in Chapter 4.

Proof. The derivative of ℓ is

$$D_\beta \ell(\beta) = -2(\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W}\mathbf{X}.$$

For the differentiation it may be useful to think of $\ell(\beta)$ as a composition. The function $a(\beta) = (\mathbf{Y} - \mathbf{X}\beta)$ from $\mathbb{R}^p$ to $\mathbb{R}^n$ has derivative $D_\beta a(\beta) = -\mathbf{X}$, and ℓ is a composition of a with the function $b(z) = z^T \mathbf{W} z$ from $\mathbb{R}^n$ to $\mathbb{R}$ with derivative $D_z b(z) = 2z^T \mathbf{W}$. By the chain rule

$$D_\beta \ell(\beta) = D_z b(a(\beta))) D_\beta a(\beta) = -2(\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W}\mathbf{X}.$$

Note that the derivative is a row vector[5]. The second derivative is

$$D_\beta^2 \ell(\beta) = 2\mathbf{X}^T \mathbf{W}\mathbf{X}.$$

If $\mathbf{X}$ has rank p, $D_\beta^2 \ell(\beta)$ is (globally) positive definite, and there is a unique minimizer found by solving $D_\beta \ell(\beta) = 0$, which amounts to a transposition of the normal equation. □

Under the rank-p assumption on $\mathbf{X}$, the solution to the normal equation can, of course, be written as

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

As we discuss below, the practical computation of $\hat{\beta}$ does not rely on explicit matrix inversion.

A geometric interpretation of the weighted least squares estimator provides additional insights. The inner product induced by $\mathbf{W}$ on $\mathbb{R}^n$ is given by $\mathbf{y}^T \mathbf{W}\mathbf{x}$, and the corresponding norm is denoted $||\cdot||_{\mathbf{W}}$. With this notation we see that

$$\ell(\beta) = ||\mathbf{Y} - \mathbf{X}\beta||_{\mathbf{W}}^2.$$

If $L = \{\mathbf{X}\beta \mid \beta \in \mathbb{R}^p\}$ denotes the column space of $\mathbf{X}$, ℓ is minimized whenever $\mathbf{X}\beta$ is the orthogonal projection of $\mathbf{Y}$ onto L in the inner product given by $\mathbf{W}$.

$||\mathbf{y}||_{\mathbf{W}}^2 = \mathbf{y}^T \mathbf{W}\mathbf{y}$ specifies a norm if and only if $\mathbf{W}$ is positive definite.

Lemma 2.2. *The orthogonal projection onto* L *is*

$$P = \mathbf{X}(\mathbf{X}^T \mathbf{W}\mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}$$

provided that $\mathbf{X}$ *has full column rank* p.

[5] The gradient,

$$\nabla_\beta \ell(\beta) = D_\beta \ell(\beta)^T,$$

is a column vector.

Proof. We verify that P is the orthogonal projection onto L by verifying three characterizing properties:

$$\begin{aligned} P\mathbf{X}\beta &= \mathbf{X}\beta \quad (P \text{ is the identity on } L) \\ P^2 &= \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W} = P \\ P^T\mathbf{W} &= (\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W})^T\mathbf{W} \\ &= \mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W} = \mathbf{W}P. \end{aligned}$$

The last property is self-adjointness w.r.t. the inner product given by $\mathbf{W}$. □

Note that since $P\mathbf{Y} = \mathbf{X}\hat{\beta}$, Theorem 2.1 follows directly from Lemma 2.2 – using the fact that when the columns of $\mathbf{X}$ are linearly independent, the equation $P\mathbf{Y} = \mathbf{X}\beta$ has a unique solution.

If $\mathbf{X}$ does not have rank p the projection is still well defined, and it can be written as

$$P = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-}\mathbf{X}^T\mathbf{W}$$

where $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-}$ denotes a generalized inverse[6]. This is seen by verifying the same three conditions as in the proof above. The solution to $P\mathbf{Y} = \mathbf{X}\beta$ is, however, no longer unique, and the solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$

is just one possible solution.

Algorithms

The actual computation of the solution to the normal equation is typically based on a QR-decomposition instead of a direct matrix inversion. The R function `lm` – or rather the underlying R functions `lm.fit` and `lm.wfit` – are based on the QR-decomposition. If we write[7] $\mathbf{W} = \mathbf{L}\mathbf{L}^T$ and introduce $\tilde{\mathbf{X}} = \mathbf{L}^T\mathbf{X}$ and $\tilde{\mathbf{Y}} = \mathbf{L}^T\mathbf{Y}$, the normal equation can be rewritten as

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\beta = \tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}.$$

Then we compute the QR-decomposition of $\tilde{\mathbf{X}}$, that is,

$$\tilde{\mathbf{X}} = \mathbf{Q}\mathbf{R}$$

[6] A generalized inverse of a matrix A is any matrix A^- with the property that $AA^-A = A$

[7] This could be the Cholesky decomposition. For a diagonal $\mathbf{W}$, $\mathbf{L}$ is diagonal and trivial to compute by taking square roots. For unstructured $\mathbf{W}$ the computation of the Cholesky decomposition scales as n^3.

where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{R}$ is an upper triangular matrix. Since

$$\mathbf{X}^T\mathbf{W}\mathbf{X} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathbf{R}^T \underbrace{\mathbf{Q}^T\mathbf{Q}}_{\mathbf{I}} \mathbf{R} = \mathbf{R}^T\mathbf{R}, \tag{2.7}$$

the normal equation becomes

$$\mathbf{R}^T\mathbf{R}\beta = \mathbf{R}^T\mathbf{Q}^T\tilde{\mathbf{Y}}.$$

This equation can be solved efficiently and in a numerically stable way in a two-step pass by exploiting first that $\mathbf{R}^T$ is lower triangular and then that $\mathbf{R}$ is upper triangular. Note that the computations based on the QR-decomposition don't involve the computation of $\mathbf{X}^T\mathbf{W}\mathbf{X}$. The factorization (2.7) of the positive definite matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ as a lower and upper triangular matrix is called the Cholesky decomposition.

An alternative to the QR-decomposition is to compute $\mathbf{X}^T\mathbf{W}\mathbf{X}$ and then compute its Cholesky decomposition directly. The QR-decomposition is usually preferred for numerical stability. Computing $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is essentially a squaring operation, and precision can be lost.

Sampling distributions

In this section we give results on the distribution of the estimator, $\hat{\beta}$, and test statistics of linear hypotheses under the weak assumptions A1, A2 and A4 and under the strong assumptions A3 and A5. Under the weak assumptions we can only obtain results on moments, while the strong distributional assumptions give exact sampling distributions. Throughout we restrict attention to the case where $\mathbf{W} = \mathbf{I}$.

Moments

Some results involve the unknown variance parameter σ^2 (see Assumption A2) and some involve a specific estimator $\hat{\sigma}^2$. This estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta})^2 = \frac{1}{n-p}||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2 \tag{2.8}$$

provided that $\mathbf{X}$ has full rank p. With the i'th *residual* defined as

$$\hat{\varepsilon}_i = Y_i - X_i^T \hat{\beta},$$

the variance estimator is – up to division by $n - p$ and not n – the empirical variance of the residuals. Since the residual is a natural estimator of the unobserved error ε_i, the variance estimator $\hat{\sigma}^2$ is a natural estimator of the error variance σ^2. The explanation of the denominator $n - p$ is related to the fact that $\hat{\varepsilon}_i$ is an estimator of ε_i. A partial justification, as shown in the following theorem, is that division by $n - p$ makes $\hat{\sigma}^2$ unbiased.

Theorem 2.3. *Under the weak assumptions A1, A2 and A4, and assuming that* $\mathbf{X}$ *has full rank* p*,*

$$\begin{aligned} E(\hat{\beta} \mid \mathbf{X}) &= \beta, \\ V(\hat{\beta} \mid \mathbf{X}) &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}, \\ E(\hat{\sigma}^2 \mid \mathbf{X}) &= \sigma^2. \end{aligned}$$

Proof. Using assumptions A1 and A4 we find that

$$\begin{aligned} E(\hat{\beta} \mid \mathbf{X}) &= E((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \mid \mathbf{X}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{Y} \mid \mathbf{X}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= \beta. \end{aligned}$$

Using assumptions A2 and A4 it follows that

$$\begin{aligned} V(\hat{\beta} \mid \mathbf{X}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T V(\mathbf{Y} \mid \mathbf{X})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned}$$

For the computation of the expectation of $\hat{\sigma}^2$, the geometric interpretation of $\hat{\beta}$ is useful. Since $\mathbf{X}\hat{\beta} = P\mathbf{Y}$ with P the orthogonal projection onto the column space L of $\mathbf{X}$, we find that

$$\mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbf{I} - P)\mathbf{Y}.$$

Because $E(\mathbf{Y} - \mathbf{X}\hat{\beta} \mid \mathbf{X}) = 0$

$$E(||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2 \mid \mathbf{X}) = \sum_{i=1}^{n} V(\mathbf{Y} - \mathbf{X}\hat{\beta} \mid \mathbf{X})_{ii}$$

and

$$\begin{aligned} V(\mathbf{Y} - \mathbf{X}\hat{\beta} \mid \mathbf{X}) &= V((\mathbf{I} - P)\mathbf{Y} \mid \mathbf{X}) \\ &= (\mathbf{I} - P)V(\mathbf{Y} \mid \mathbf{X})(\mathbf{I} - P)^T \\ &= (\mathbf{I} - P)\sigma^2\mathbf{I}(\mathbf{I} - P) \\ &= \sigma^2(\mathbf{I} - P). \end{aligned}$$

The sum of the diagonal elements in $(\mathbf{I} - P)$ is the trace of this orthogonal projection onto $L^{\perp}$ – the orthogonal complement of L – and is thus equal to the dimension of $L^{\perp}$, which is $n - p$. □

Tests and confidence intervals

For the computation of exact distributional properties of test statistics and confidence intervals we need the strong distributional assumptions A3 and A5.

Theorem 2.4. *Under the strong assumptions A3 and A5 it holds, conditionally on* $\mathbf{X}$, *that*

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

and that

$$(n - p)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}.$$

Moreover, for the standardized Z*-score*

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-p},$$

or more generally for any $a \in \mathbb{R}^p$

$$Z_a = \frac{a^T\hat{\beta} - a^T\beta}{\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a}} \sim t_{n-p}.$$

Proof. See EH, Chapter 10. □

The standardized Z-scores are used to test hypotheses about a single parameter or a single linear combination of the parameters. The Z-score is computed under the hypothesis (with the hypothesized value of β_j or $a^T\beta$ plugged in), and compared to the t_{n-p} distribution. The test is two-sided. The Z-scores are also used to construct confidence intervals for linear combinations of the parameters. A 95% confidence interval for $a^T\beta$ is computed as

$$a^T\hat{\beta} \pm z_{n-p}\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a} \tag{2.9}$$

where $\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a}$ is the estimated standard error of $a^T\hat{\beta}$ and z_{n-p} is the 97.5% quantile in the t_{n-p}-distribution.

For the computation of $a^T(\mathbf{X}^T\mathbf{X})^{-1}a$ it is noteworthy that $(\mathbf{X}^T\mathbf{X})^{-1}$ is not needed, if we have computed the QR-decomposition of $\mathbf{X}$ or the Cholesky decomposition of $\mathbf{X}^T\mathbf{X}$ already. With $\mathbf{X}^T\mathbf{X} = \mathbf{L}\mathbf{L}^T$ for a lower triangular[8] $p \times p$ matrix $\mathbf{L}$ we find that

$$\begin{aligned} a^T(\mathbf{X}^T\mathbf{X})^{-1}a &= a^T(\mathbf{L}\mathbf{L}^T)^{-1}a \\ &= (\mathbf{L}^{-1}a)^T\mathbf{L}^{-1}a \\ &= b^Tb \end{aligned}$$

where b solves $\mathbf{L}b = a$. The solution of this lower triangular system of equations is *faster* to compute than the matrix-vector product $(\mathbf{X}^T\mathbf{X})^{-1}a$, even if the inverse matrix is already computed and stored. This implies that the computation of $(\mathbf{X}^T\mathbf{X})^{-1}$ is never computationally beneficial. Not even if we need to compute estimated standard errors for many different choices of a.

To test hypotheses involving more than a one-dimensional linear combination, we need the F-tests. Let $p_0 < p$ and assume that $\mathbf{X}'$ is an $n \times p_0$-matrix whose p_0 columns span a p_0-dimensional subspace of the column space of $\mathbf{X}$. With $\hat{\beta}'$ the least squares estimator corresponding to $\mathbf{X}'$ the F-test statistic is defined as

$$F = \frac{||\mathbf{X}\hat{\beta} - \mathbf{X}'\hat{\beta}'||^2/(p-p_0)}{||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2/(n-p)}. \tag{2.10}$$

Note that the denominator is just $\hat{\sigma}^2$. The F-test statistic is one-sided with large values critical.

Theorem 2.5. *Under the strong assumptions A3 and A5 and the hypothesis that*

$$E(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}'\beta_0'$$

the F-test statistic follows an F-distribution with $(p-p_0, n-p)$ degrees of freedom.

[8] If we have computed the QR-decomposition, $\mathbf{L} = \mathbf{R}^T$.

Proof. See EH, Chapter 10. □

The terminology associated with the F-test and reported by the R function `anova` is as follows (R abbreviations in parentheses). The norm $||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2$ is called the residual sum of squares (RSS) under the model, and $n - p$ is the residual degrees of freedom (Res. Df). The norm $||\mathbf{X}\hat{\beta} - \mathbf{X}'\hat{\beta}'||^2$ is the sum of squares (Sum of Sq.), and $p - p_0$ is the degrees of freedom (Df). The norm $||\mathbf{Y} - \mathbf{X}'\hat{\beta}'||^2$ is the residual sum of squares under the hypothesis, and it follows from Pythagoras that

$$||\mathbf{X}\hat{\beta} - \mathbf{X}'\hat{\beta}'||^2 = ||\mathbf{Y} - \mathbf{X}'\hat{\beta}'||^2 - ||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2.$$

Thus the sum of squares is the difference between the residual sum of squares under the hypothesis and under the model. The R function `anova` computes and reports these numbers together with a p-value for the null hypothesis derived from the appropriate F-distribution.

It is important for the validity of the F-test that the column space, L', of $\mathbf{X}'$ is a subspace of the column space, L, of $\mathbf{X}$. Otherwise the models are not nested and a formal hypothesis test is meaningless even if $p_0 < p$. It may not be obvious from the matrices if the models are nested. By definition, $L' \subseteq L$ if and only if

$$\mathbf{X}' = \mathbf{X}C \tag{2.11}$$

for a $p \times p_0$ matrix C (of rank p_0), and we can verify that $L' \subseteq L$ if there is such a C matrix. The situation where the columns of X' is a subset of the columns of X, which is a hypothesis on the complementary set of parameters being 0, corresponds to C being diagonal with 0's or 1's appropriately placed in the diagonal. In the ANOVA literature, where categorical predictors and their interactions are considered, a considerable amount of work has been invested on choosing the $\mathbf{X}$ matrix, and thus the parametrization, so that relevant scientific hypotheses can be formulated in terms of certain parameters being 0. Though this can ease communication in certain cases, it can also be quite confusing. Linear hypotheses considered within the linear model are not specific to a given choice of parametrization, and there is no requirement that a linear hypothesis is linked to certain parameters being 0. This will be illustrated in the next section.

The actual C matrix is not needed if theoretical considerations show that it exists.

Birth weight – nonlinear expansions

We found in the previous analysis of the birth weight data a lack of model fit that appeared to be related to a nonlinear relation between `weight` and `gestationalAge`. To handle nonlinear relations between the response and one or more predictors, the predictors (as well as the response) can be nonlinearly transformed before they enter into the linear model. Such pre-modeling transformations extend the scope of the linear model considerably. It allows us, for instance, to model power-law relations. We just have to remember that the errors are then additive on the transformed scale, and that the model assumptions must hold for the transformed data. In general, it may be difficult to come up with just the right transformation, though. An alternative is to use a small but flexible class of *basis functions*, which can capture the nonlinearity. One possibility is to use low degree polynomials. This can be done by simply including powers of a predictor as additional predictors. Depending on the scale and range of the predictor, this may work just fine. However, raw powers can result in numerical difficulties. An alternative is to use orthogonal polynomials, which are numerically more well behaved.

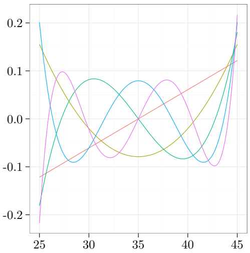


Figure 2.13: Orthogonal polynomials.

Figure 2.13 shows an example of an orthogonal basis of degree 5 polynomials on the range 25–45. This corresponds approximately to the range of `gestationalAge`, which is a target for basis expansion in our case. What should be noted in Figure 2.13 is that the behavior near the boundary is quite erratic. This is characteristic for polynomial bases. To achieve flexibility in the central part of the range, the polynomials become erratic close to the boundaries. Anything can happen beyond the boundaries.

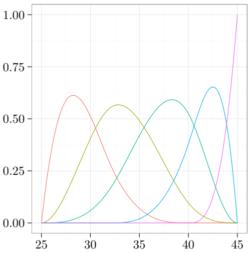


Figure 2.14: *B*-splines computed using the `bs` function.

An alternative to polynomials is splines. A spline[9] is piecewisely a polynomial, and the pieces are joined together in a sufficiently smooth way. The points where the polynomials are joined together are called *knots*. The flexibility of a spline is determined by the number and placement of the knots and the degree of the polynomials. A degree k spline is required to be $k-1$ times continuously differentiable. A degree 3 spline, also known as a *cubic spline*, is a popular choice, which thus has a continuous (piece-wise linear) second derivative.

Figure 2.14 shows a basis of 5 cubic spline functions. They are so-called B-splines (basis splines). Note that it is impossible to visually detect the knots where the second derivative is non-differentiable. The degree k B-spline basis with r knots has $k + r$ basis functions[10]. As seen in Figure 2.14, the B-spline basis is also somewhat erratic close to the boundary. For a cubic spline, the behavior close to the boundary can be controlled by requiring that the second and third derivatives are 0 at the boundary knots. The result is known as a *natural cubic spline*. The extrapolation (as a spline) of a natural cubic spline beyond the boundary knots is linear.

Due to the restriction on the derivatives of a natural cubic spline, the basis with r knots has $r + 1$ basis functions. Thus the basis for the natural cubic splines with $r + 2$ knots has the same number of basis functions as the raw cubic B-spline basis with r knots. This means in practice that, compared to using raw B-splines with r knots, we can add two (internal) knots, and thus increase the central flexibility of the model, while retaining its complexity in terms of $r + 3$ parameters.

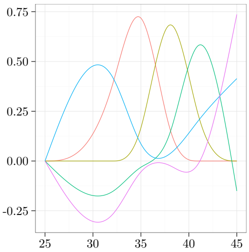


Figure 2.15: B-spline basis for natural cubic splines computed using the `ns` function.

Expansions should only be tried if we expect to get something out of it, that if, if we expect that there are some nonlinear relations in the data. On the other hand, we should be careful not to construct models tailor-made to capture nonlinear relations we have spotted by eye-balling residual plots. In particular, if it is followed up by a formal justification using a statistical test. It would be a model selection procedure with statistical implications, which it is difficult to account for. It is always extremely difficult to take informal model selection into account in such a the statistical test, which will result in a test with too large a level.

We decided to expand `gestationalAge` using natural cubic splines with three knots in 38, 40, and 42 weeks. The boundary knots were determined by the range of the data set, and were thus 25 and 47. We also expanded `age`, but we let the `ns` function determine the knots automatically for those two predictors. The last continuous predictor, `alcohol` has a very skewed marginal distribution, and it was not judged to be suitable for a standard basis expansion. We also present a test of the nonlinear effect.

[10] This is when the constant function is excluded from the basis, which is the way to go when the basis expansion is used in a regression model including an intercept

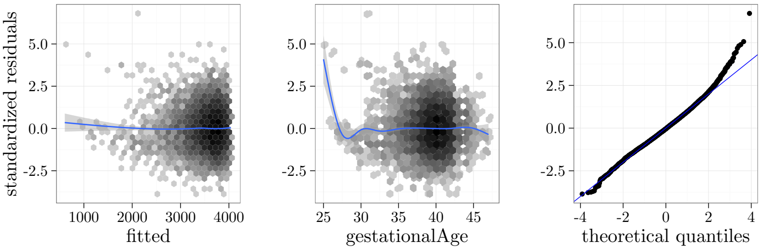


Figure 2.16: Diagnostic plots for the model with `gestationalAge` expanded using splines.

Nonlinear main effects model.

```
nsg <- function(x)
  ns(x, knots = c(38, 40, 42), Boundary.knots = c(25, 47))
form <- weight ~ nsg(gestationalAge) + ns(age, df = 3) + children +
  coffee + alcohol + smoking + abortions + feverEpisodes
pregnantLm3 <- lm(form, data = pregnant)
anova(pregnantLm, pregnantLm3)
```

Table 2.7: Test of the model including a spline expansion of `gestationalAge` against the main effects model.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11139	2.5376e+09				
2	11134	2.4664e+09	5	7.1213e+07	64.3	2.254e−66

The main effects model is a submodel of the nonlinear main effects model. This is not obvious. The nonlinear expansion will result in 4 columns in the model matrix $\mathbf{X}$, none of which being `gestationalAge`. However, the linear function is definitely a natural cubic spline, and it is thus in the span of the 4 basis functions. With $\mathbf{X}'$ the model matrix for the main effects model, it follows that there is a C such that (2.11) holds. This justifies the use of the F-test. The conclusion from Table 2.7 is that the nonlinear model is highly significant.

The positioning of the knots for spline expansion is selected based on the marginal distribution of the predictor[11]. The knots should be placed reasonably relative to the distribution of the predictor, so that we learn about nonlinearities where there is data to learn from. In this case we placed knots at the median and the 10% and 90% quantiles of the distribution of `gestationalAge`. An eye-ball decision based on 2.9 is that a single knot around 41 would have done the job, but we refrained from making such a decision. A subsequent test of the nonlinear effect with 1 degrees of freedom would not appropriately take into account how the placement of the knot was made. The `ns` function makes a similar automatic

[11] Letting the knots be parameters to be estimated is not a statistically viable idea.

selection of knots based on the marginal distribution of the predictor variables when it is applied in the formula. One just have to specify the number of basis functions using the `df` argument.

Figure 2.16 shows diagnostic plots for the nonlinear main effects model. They show that the inclusion of the nonlinear effect removed the previously observed problem with assumption A1. The error distribution is still not normal, but right skewed with a fatter right tail than the normal distribution. There is a group of extreme residuals for preterm born children, which should be given more attention than they will be given here.

Reduced nonlinear main effects model.

```
form <- weight ~ nsg(gestationalAge) + children + coffee + smoking
pregnantLm4 <- lm(form, data = pregnant)
anova(pregnantLm4, pregnantLm3)
```

Table 2.8: Test of weak predictors in the nonlinear main effects model.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11142	2.4694e+09				
2	11134	2.4664e+09	8	2.9381e+06	1.66	0.1032

Table 2.8 shows that dropping the 4 weak predictors from the nonlinear main effects model is again borderline and not really significant. They will be excluded from the remaining analysis.

Figure 2.17 shows examples of the fit for the reduced nonlinear main effects model. The figure illustrates the general nonlinear relation between `weight` and `gestationalAge`. Differences due to other variables are purely additive in this model, which amounts to translations up or down of the curve. The figure shows a couple of extreme cases; the majority group who have had children before and who don't smoke or drink coffee, and a minority group who have had children before and smoke and drink coffee the most. What we should notice is the wider confidence band on the latter (`smoke = 3`, `coffee = 3`) compared to the former, which is explained by the skewness of the predictor distributions. Table 2.9 gives confidence intervals for the remaining parameters based on the nonlinear main effects model.

```
predFrame <- expand.grid(children = factor(1),
                         smoking = factor(c(1, 3)),
                         coffee = factor(c(1, 3)),
                         gestationalAge = seq(25, 47, 0.1),
                         alcohol = 0,
```

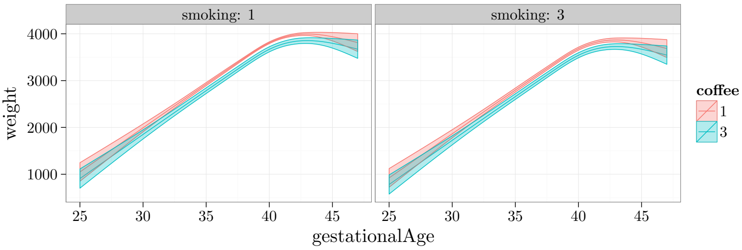


Figure 2.17: Main effects model with basis expansion of `gestationalAge`. Here illustrations of the fitted mean and 95% confidence bands for `children=1`.

```
                                   age = median(pregnant$age),
                                   feverEpisodes = 0,
                                   abortions = factor(0)
                                   )
predGest <- predict(pregnantLm3, newdata = predFrame,
                    interval = "confidence")
predFrame <- cbind(predFrame, predGest)
qplot(gestationalAge, fit, data = predFrame, geom = "line",
      color = coffee) + ylab("weight") +
  geom_ribbon(aes(ymin = lwr, ymax = upr, fill = coffee),
              alpha = 0.3) +
  facet_grid(. ~ smoking, label = label_both)
```

Table 2.9: Confidence intervals.

	2.5 %	97.5 %
(Intercept)	703.58	1141.28
children1	155.47	194.09
coffee2	−82.61	−42.43
coffee3	−193.02	−87.63
smoking2	−125.76	−75.26
smoking3	−155.76	−97.98

To conclude the analysis, we will again include interactions. However, the variable `gestationalAge` is in fact only taking the integer values 25 to 47, and the result of a nonlinear effect coupled with a third order interaction, say, results in an almost saturated model. That is, the interaction model has more or less a separate mean for all observed combinations of the predictors. We choose to consider a less complex model with all second order interactions between `gestationalAge` and the three factors that we have judged to be strong predictors.

```
form <- weight ~ (smoking + coffee + children) * nsg(gestationalAge) +
   ns(age, df = 3) + alcohol + abortions + feverEpisodes
pregnantLm5 <- lm(form, data = pregnant)
```

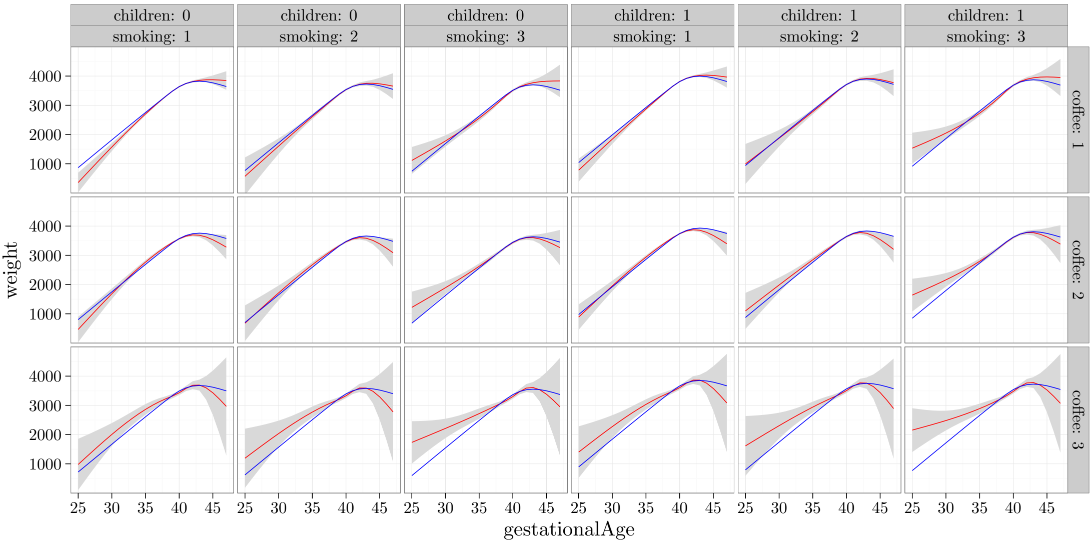


Figure 2.18: Comparison of the interaction model (red, 95% gray confidence bands) with the reduced nonlinear main effects model (blue).

```
anova(pregnantLm4, pregnantLm5)
```

Nonlinear interaction model.

```
predFrame <- expand.grid(children = factor(c(0, 1)),
                         smoking = factor(c(1, 2, 3)),
                         coffee = factor(c(1, 2, 3)),
                         gestationalAge = 25:47,
                         alcohol = 0,
                         age = median(pregnant$age),
                         feverEpisodes = 0,
                         abortions = factor(0)
                         )
predFrame <- cbind(predFrame,
  predict(pregnantLm5, newdata = predFrame, interval = "confidence")
)
predFrame$fit4 <- predict(pregnantLm4, newdata = predFrame)
ggplot(predFrame, aes(gestationalAge, fit)) +
  facet_grid(coffee ~ children + smoking, label = label_both) +
  geom_ribbon(aes(ymin = lwr, ymax = upr), fill = gray(0.85)) +
  geom_line(color = "red") + coord_cartesian(ylim = c(0, 5000)) +
  geom_line(aes(y = fit4), color = "blue") + ylab("weight") +
  scale_y_continuous(breaks = c(1000, 2000, 3000, 4000))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11134	2.4664e+09				
2	11114	2.4512e+09	20	1.5261e+07	3.46	2.612e−07

Table 2.10: Test of interactions.

Table 2.10 shows that inclusion of interaction terms is still significant. We should, however, try to visualize how the nonlinear interaction model differs from the model with only main effects. Figure 2.18 shows the predicted values for the reduced nonlinear main effects model for all combinations of `children`, `smoking` and `coffee` (blue curves). These curves are all just translations of each other. In addition, the figure shows the predicted values and a 95% confidence band as estimated with the nonlinear interaction model. We observe minor deviations for the reduced main effects model, which can explain the significance of the test, but the deviations appear unsystematic and mostly related to extreme values of `gestationalAge`. The conclusion is that even though the inclusion of interaction effects is significant, there is little to gain over the reduced nonlinear main effects model.

Exercises

Exercise 2.1. Show that if (2.11) holds then

$$C = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}'.$$

Explain how this can be used to check if the column space of an arbitrary $\mathbf{X}'$ is contained in the column space of $\mathbf{X}$. ○

Exercise 2.2. Consider the following matrices

```
X <- bs(seq(0, 1, 0.01), knots = c(0.3, 0.5, 0.7))
Xprime <- seq(0, 1, 0.01)
```

of B-spline basis functions evaluated on a grid in the interval $[0,1]$. Use the previous exercise to verify that `Xprime` is in the column space of `X`. You can use the formula from the exercise to compute C. Can you also use the `lm.fit` function? ○

Exercise 2.3. Consider a *linear estimator*

$$\tilde{\beta} = C^T\mathbf{Y}$$

for some $N \times p$ matrix C. Show that it is unbiased if A1 holds and if

$$\beta = C^T\mathbf{X}\beta.$$

Show that if it is unbiased and A1, A2 and A4 hold then

$$V(\tilde{\beta}|\mathbf{X}) = \sigma^2 C^T C$$

and that

$$V(\hat{\beta} - \tilde{\beta}|\mathbf{X}) = \sigma^2(C^T C - (\mathbf{X}^T\mathbf{X})^{-1}).$$

Explain that this shows the Gauss-Markov theorem: Under the assumptions A1 + A2 + A4 the least squares estimator of β has minimal variance among all linear, unbiased estimators of β. ○

The following three exercises relate the empirical correlations between the predictor variables to the variances of the parameters in the linear model. The common setup is as follows. The matrix $\mathbf{X}$ denotes an $n \times p$ matrix of predictor variables, $\mathbf{1}$ denotes an n dimensional column vector of ones and $(\mathbf{1}\ \ \mathbf{X})$ denotes the $n \times (p+1)$ matrix with the first column being $\mathbf{1}$ and the remaining $n \times p$ block being $\mathbf{X}$. We let

$$\bar{X} = \frac{1}{n}\mathbf{X}^T\mathbf{1}$$

denote the p dimensional vector of mean values for the predictor, and we let

$$\hat{\Sigma} = \frac{1}{n}(\mathbf{X} - \mathbf{1}\bar{X}^T)^T(\mathbf{X} - \mathbf{1}\bar{X}^T)$$

denote the empirical covariance matrix. It is assumed that $\hat{\Sigma}$ is invertible.

Exercise 2.4. With the setup above, show that

$$n\hat{\Sigma} = \mathbf{X}^T\mathbf{X} - n\bar{X}\bar{X}^T.$$

Use block matrix inversion to show that

$$\left((\mathbf{1}\ \ \mathbf{X})^T(\mathbf{1}\ \ \mathbf{X})\right)^{-1} = \begin{pmatrix} * & * \\ * & \frac{1}{n}\hat{\Sigma}^{-1} \end{pmatrix}.$$

Here the $*$'s mean some values that we are not interested in computing in this exercise. ○

Exercise 2.5. Use the result from the previous exercise to show that if we fit a linear model with an intercept (which is denoted β_0), then the variance of $\hat{\beta}_i$, conditionally on $\mathbf{X}$, for $i = 1, \ldots, p$ is

$$\frac{\sigma^2}{n}\left(\hat{\Sigma}^{-1}\right)_{ii}.$$

○

Exercise 2.6. With the setup as in the previous two exercises, write the empirical covariance matrix in the following block form

$$\hat{\Sigma} = \left(\begin{array}{cc} \hat{\sigma}_1^2 & \hat{\gamma}^T \\ \hat{\gamma} & \hat{\Sigma}_{-1} \end{array} \right),$$

and use block inversion (again) to show that

$$\left(\hat{\Sigma}^{-1}\right)_{11} = \frac{1}{\hat{\sigma}_1^2 - \hat{\gamma}^T \left(\hat{\Sigma}_{-1}\right)^{-1} \hat{\gamma}}.$$

Show that in the case $p = 2$ the formula above reduces to

$$\left(\hat{\Sigma}^{-1}\right)_{11} = \frac{1}{\hat{\sigma}_1^2 (1 - \widehat{\mathrm{corr}}(X_1, X_2)^2)}.$$

Show finally that if $\mathbf{X}_{\cdot 1} - \mathbf{1}\bar{X}_1$ is perpendicular to the remaining columns in $\mathbf{X} - \mathbf{1}\bar{X}^T$, then

$$\left(\hat{\Sigma}^{-1}\right)_{11} = \frac{1}{\hat{\sigma}_1^2}.$$

Interpret the results in terms of how the empirical variance and empirical correlation between predictor variables affect the variances of the estimators $\hat{\beta}_i$ for $i = 1, \ldots, p$. ○

3

Data analysis with regression models

This chapter deals with the practical process of data analysis. Most of the general considerations apply to data analysis in a broad sense, but to keep focus, practical data analysis is treated in the framework of regression models. The process of practical data analysis involves a lot of decision making. For an inexperienced data analyst the decision making can be a daunting task, first of all because there is never a set of clearly correct decisions. Instead it appears as if many decisions have to be made on the basis of little knowledge, and whenever a decision is made, it may be open to criticism. This can have a paralyzing effect, and worse yet, it can result in that the data analyst attempts to hide the decision process. This chapter develops methods for understanding and controlling the decision making process involved in practical data analysis. This includes some treatment of how predictive regression models can be interpreted and applied, since this is important knowledge when strategic decisions about the data analysis are considered.

Practical data analysis

Practical data analysis ranges from tedious data manipulations[1],

[1] Anybody that has worked with real data knows that data manipulations are a time consuming but essential part of data analysis.

such as quality control, data cleaning and tidying, over initial and descriptive data explorations to model fitting, interpretations and applications. Most data require, due to size and nature, a computer for storage and manipulation, and data analysis belongs just as much to computer science as it does to statistics. Indeed, the famous Danish computer scientist Peter Naur preferred *datalogy*[2] over computer science to emphasize that the science is about data and data treatment and not computers. When we speak of data analysis in a statistical context we usually leave data storage and database management to computer science, but most other aspects of data analysis are statistical research topics. It is, moreover, a salient fact that among data analysts, statisticians form a minority. Yet statistical models and statistical methodologies are in widespread use. Data analysis is therefore not a subject that can be isolated to statistics and the professional statistical community.

One of the big challenges is to know which models and which methods that should be applied to which data. And even if this is known in principle, it is still a challenge to put the different computational techniques and tools together to form an actual analysis of the data. A data analyst need to know not just the individual techniques but also how they are appropriately put together.

The case on birth weight modeling in the previous chapter is an example of a practical data analysis using the linear regression model. It illustrates some aspects of data analysis, which we will build upon and use to introduce general considerations, reflections and principles in relation to the process of data analysis. Other cases follow in the subsequent chapters, and they will illustrate other aspects.

Data analysis may be considered a craft – or even an art – that is ideally learned in a master-apprentice relationship. It certainly contains a strong component of practical skills that require practical training, and which cannot be mastered by theoretical studies alone. The theory of data analysis is a rather undeveloped subject anyway, and most writings of the topic have more of an anecdotal than a theoretical nature. Data analysis is usually taught, learned and discussed more by examples than by theory.

This chapter will not propose a theory of data analysis, but it will attempt to raise the readers consciousness of the steps and choices

[2] Computer Science is translated into *datalogi* in Danish, and Datalogisk Institut is the Danish name of the Department of Computer Science at the University of Copenhagen

that are required in practical data analysis. There is ample but scattered advice in the literature[3], and we will attempt to put this advice into a common framework. The aim is to provide a framework that is specific and concrete enough that it will support and help the reader with practical problems. We will, however, refrain from bullet point lists of best practices. Such lists may be praised by students and readers who want to checkmark that they followed the correct procedure, but they encourage automatic compliance to rules. Rather than promoting a rule based approach to data analysis we want to promote an approach where the data analyst is adaptive and responsive to the outcomes during the analysis, and is consciously aware of the decision making process during analysis and how these decisions may affect the analysis and the conclusions.

Techniques, tactics and strategies

To get a better mental representation of the process of data analysis it is useful to distinguish between three different levels of operation. At the bottom level there are the techniques for data analysis. They are the tools in our toolbox, and they are well defined units in a mathematical and computational sense. The middle level consists of the tactics, which are the ways to arrange[4] the techniques in combination to analyze the data. The top level consists of the strategies, which concern themselves with the overall picture.

Techniques are ubiquitous in books dealing with mathematical and methodological aspects of regression. The least squares estimator as introduced in the previous chapter is an example of a technique. Given a n-dimensional vector $\mathbf{Y}$ and an $n \times p$ matrix $\mathbf{X}$ of rank p it is a well defined solution of an optimization problem, it is characterized as the solution to the normal equation, and it can be computed safely using the QR-decomposition.

The construction of the standard confidence interval (2.9), within the framework of the linear model and under the strong distributional assumptions A3 and A5, is another example of a technique, and so is the construction of the residual plot. In isolation, techniques produce unambiguous and interpretable numbers or figures. The main focus of mathematical or methodological textbooks and

[3] This includes the electronic literature consisting of wikis, blogs, and QA-sites

[4] Tactic is derived from the Greek word taktike, which means *art of arrangement.*

scientific papers on regression is techniques. Techniques are important. They are, in fact, essential. The better we understand the techniques, and the more techniques we master, the less time we need to ponder about them, and the more time can be spend on intricate tactical and strategic decisions.

Tactics for stringing together techniques are illustrated in the case study on birth weight. The initial exploration of data prepared the data for fitting a regression model, it provided us with clues on what to expect, and it helped us recognizing that we did not want to include `length` as a predictor. The residual plots of the first model suggested the need for nonlinear effects. It was a tactical choice to introduce a spline basis expansion for dealing with the nonlinearity. Other possible tactical choices, that were not pursued, are low order polynomial expansions or transformations of the predictor variable. The fit and visualization of the nonlinear model with and without interactions showed us what we can sensibly squeeze out of the data[5]. Along the way, considerations of p-values and confidence bands assisted us in our interpretations of what the model showed about the data. In the process of making tactical decisions, we will never be able to guarantee that the assumptions of a given technique are fulfilled, whence we should always be cautious with our interpretations. On the other hand, the core of the matter is that the tactical choices we make should support us towards a trustworthy analysis. A central tactical question in regression is the choice of which variables that should be included as predictors.

The strategy behind the case on birth weights was not explicit though several of the comments made in the case pertain to the chosen strategy. The analysis was exploratory and unfolded a strategy where the focus was on communication of models and their interpretations. Model fit was discussed in some detail and the aim was to find a model that fitted the data well and clearly showed how the different predictors contributed to the prediction of birth weight. The strategic decisions lay out a framework for tactical decisions. If the purpose had been more specific – to establish if birth weight is related to smoking when correcting for gestational age, say – a different analysis would have been carried out. Then we would have

[5] At least what this author could squeeze out.

focused on a model of birth weight regressed on smoking and gestational age only. Tactical considerations about the nonlinear effect of genstational age would still be be needed, but more attention would have been paid to estimating and testing the smoking effect. When we think about the strategy of the data analysis, we think about what purpose the analysis is supposed to serve, and what the outcome is going to be used for.

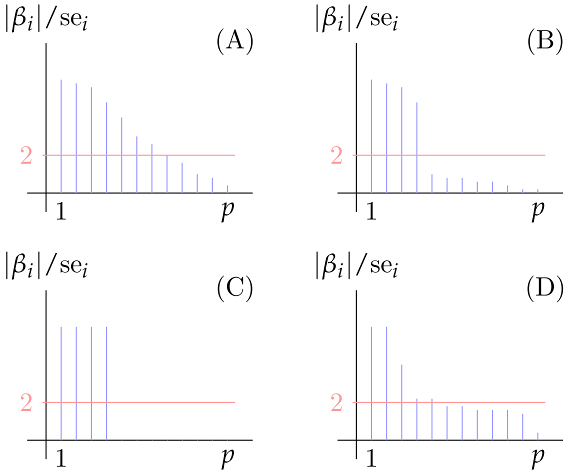
Iterations

Data analysis is an iterative process. Do not expect to get everything right the first time, or to think up the entire data analysis pipeline without looking at the data. Decisions made early on in a data analysis are constantly up for reconsideration in the light of what you may discover. A good example is distributional assumptions. If you have fitted a regression model to data and discover that the residuals are not normal (assumption A3 is not fulfilled) or that the variance is not constant (Assumption A2 is not fulfilled), then you want to reconsider the decision of fitting the linear model. Perhaps the data is better modeled using a generalized linear model from the following chapter?

The data analyst should generally feel encouraged whenever a decision has to be reconsidered. It means that there is progress in the data analysis. Deciding to fit just a single regression model and report the estimated parameters does not constitute a data analysis. The data analyst must be critical about the model, explore alternatives as well as the model fit, and iteratively attempt to improve the model used.

There is one caveat with the iterative approach, and that is the risk of overanalyzing the data. This problem is closely related to the problem with overfitting. If the data analyst begins to adapt the model to the random variation in the data, then there is a risk of "discovering" spurious effects. In the subsequent section on the tactical decision process advice is given on how we can try to avoid overanalyzing the data while still being adaptive and responsive.

Figure 3.1: Four different scenarios of true parameter values relative to the standard error of the estimators. Recall that the t-test statistic for the null hypothesis $H_0 : \beta_i = 0$ is $\hat{\beta}_i/\hat{\text{se}}_i$. In which of these scenarios will variable selection work well? Which of the scenarios are realistic?



Tactical considerations

This section treats the tactical decision process and highlights some of the possible and common pitfalls that might jeopardize the validity of an otherwise technically sound data analysis. Some recommendations are also given, but it is important to emphasize once again that there is no checklist to follow that will guarantee success.

The problem with variable selection

The main tactical decisions in regression analysis concern how predictor variables should be selected, and how the variables should enter into the model. In the case study on birth weight we excluded length but otherwise avoided any variable selection, and we did, in particular, not attempt to remove "insignificant variables" or remove variables because they were weak predictors. This is the most important overall advice: Avoid data driven variable selection if possible. And if you do variable selection anyway, acknowledge the consequences and the difficulties that will follow for the downstream data analysis. The main reason to avoid variable selection is that most statistical methods do not have the desired properties when applied conditionally on the result of a variable selection procedure. At least not if the procedure depends on the response variable. To appreciate the problem, it is useful to study Figure 3.1. It illustrates four different scenarios of true regression parameters β_i, and the figure shows the ratios $|\beta_i|/\text{se}_i$ where se_i is the standard error of the estimator of β_i. Before reading any further, try to answer the questions phrased in the figure legend.

There are many methods for variable selection, and we cannot possibly cover them all in any detail, but a common idea is that for a subset of the variables the true regression coefficients are 0 or small, and that we can determine (perhaps iteratively or in groups) if some estimates could be put equal to 0 based on the estimates $\hat{\beta}_i$. A reestimate of the remaining parameters is then in most cases necessary after we put some equal to 0. The value of the t-test statistic

$$t_i = |\hat{\beta}_i|/\hat{\text{se}}_i$$

can serve as a first screening for variables with small (or 0) coefficients. Variable selection is easy if t_i is large for the nonzero coefficients and small for the zero coefficients.

Among variable selection procedures are the infamous forward variable selection and backward variable elimination.

Figure 3.1 (C) represents a scenario where variable selection is easy, while (D) represents a difficult scenario. Figure 3.1 (B) might be similar to the birth weight case with four variables having relatively large coefficients while the rest have relatively small or almost zero coefficients.

The scenarios (A), (B) and (D) are realistic, while scenario (C) is not, and Scenario (C) is quite irrelevant for a practical discussion of methods for variable selection. It can at best be seen as a surrogate for Scenario (B). In the following, we focus the discussion on variables with relatively small coefficients. If we use variable selection in Scenario (B) we will make two mistakes for those variables. Either they are not selected for inclusion in the model, and the corresponding estimate is 0. This introduces a small bias that is rarely problematic as it also reduces the variance. The other mistake occurs when the $\hat{\beta}_i$ estimate has a (by chance) large enough value that the variable is selected. This is problematic because conditionally on selecting the variable we will report a biased estimate and a too narrow confidence interval. In Scenario (D) both mistakes will be pronounced, and the subset of variables that is actually included will be quite arbitrary. If we avoid variable selection based on data adaptive procedures the resulting model fit and methods for uncertainty assessment, e.g. confidence intervals, will work as expected, but conditionally on data adative variable selection such methods will not work as expected. Theoretical or simulation based reassurance under Scenario (C) that the methods will almost work even conditionally on variable selection are mostly irrelevant as Scenario (C) is not realistic.

Whether interactions should be included into the model and whether variables should be transformed or expanded in a basis to account for nonlinear effects is closely related to variable selection. Indeed, if we focus on interactions and expansions the resulting problem is a question of which derived variables we should include. The exact same arguments as above applies. Remember that visual inspections of plots are just informal data adaptive test procedures that should be treated with the same care as formal test based methods.

The following list of recommendations should be considered in questions related to variable selection.

- Consult subject matter knowledge and the literature.
- Base decisions on marginal distributional considerations. If needed, eliminate predictor variables with limited variability or many missing values.
- Base decisions on pairwise distributional considerations. If needed, eliminate the least relevant variables among collinear predictor variables. Or form an "index" from multiple collinear predictor variables, e.g. a weighted average.
- Data adaptive variable selection based on formal tests and stepwise procedures are discouraged.

As a rule of thumb, data based methods for variable selection or aggregation that only rely on the distribution of the predictor variables can be used without problems.

Nonlinear expansions

A pertinent problem in data analysis is whether one or more variables should enter nonlinearly into the model. Variable transformations, such as taking the log or the square root of a predictor

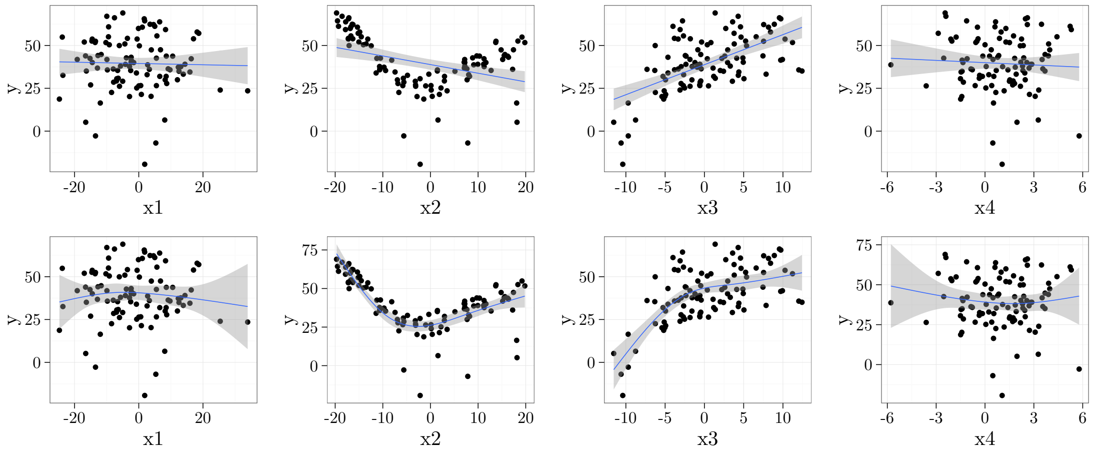


Figure 3.2: Marginal association of the response and the four predictors. A linear regreesion (top) and a general smoother (bottom) help capture the trend or possible nonlinear marginal relations.

variable, can sometimes be useful, but if we start searching trough several possible variable transformations and selecting the one that gives the best fit to data, we introduce the same problem as with variable selection. It is hard to account for the search as part of the estimation proceduce by standard statistical methods. Basis expansions are, on the contrary, quite easy to put into the framework of the linear model and allow for the use of standard statistical methods. This is true as long as the basis functions, e.g. the knots in a spline basis expansions, are not selected in a data adaptive way. Or rather, in a response adaptive way. It is perfectly OK to base the choice of knots on the marginal distribution of the predictor variable.

When we use basis expansions in practice we use combinations of tests and plots to judge the importance of including nonlinear effects. To illustrate appropriate usage we consider in this section a small simulated example with $n = 100$ and $p = 4$. Figure 3.2 shows plots of the response against the four predictor variables.

```r
lmA <- lm(y ~ x2)
lmB <- lm(y ~ x3)
lmC <- lm(y ~ x1 + x2 + x3 + x4)
lmD <- lm(y ~ ns(x1, df = 3) + ns(x2, df = 3) +
            ns(x3, df = 3) + ns(x4, df = 3))
pA <- qplot(.fitted, .resid, data = fortify(lmA)) + geom_smooth() +
  annotate("text", x = 48, y = -50, label= "(A)")
pB <- qplot(.fitted, .resid, data = fortify(lmB)) + geom_smooth() +
```

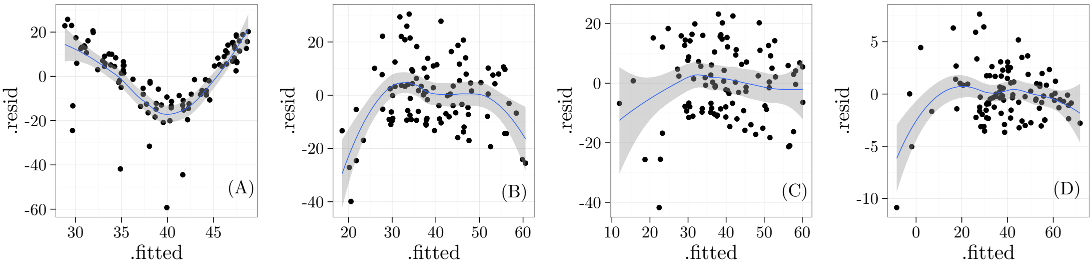


Figure 3.3: Residual plots. The models are: (A) only X_2 is included linearly; (B) only X_3 is included linearly; (C) all four variables are included linearly; (D) all four variables are included nonlinearly.

```
    annotate("text", x = 58, y = -36, label= "(B)")
pC <- qplot(.fitted, .resid, data = fortify(lmC)) + geom_smooth() +
    annotate("text", x = 58, y = -36, label= "(C)")
pD <- qplot(.fitted, .resid, data = fortify(lmD)) + geom_smooth() +
    annotate("text", x = 66, y = -9, label= "(D)")
grid.arrange(pA, pB, pC, pD, ncol = 4)
```

Figure 3.2 suggests that some variables should be included nonlinearly into the model. There is in general no guarantee that nonlinear relations will show themselves in the marginal association or vice versa, but the circumstances should be quite strange if we see a figure like Figure 3.2 where we don't need nonlinear effects in the full model. In Figure 3.3 we see residual plots after having fitted four different models. It is evident from the residual plots in Figure 3.3 (A) and (B) that if we include just X_2 or X_3, the model does not fit, and that there is some nonlinear deviation. What is interesting is that the residual plot in Figure 3.3 (C), where all four variables X_1, X_2, X_3 and X_3 are included, does not reveal a clear lack of model fit. There are a few large negative residuals for the small fitted values, but otherwise the plot suggests that the model fits. The residual plot in Figure 3.3 (D), where all four variables are expanded in a spline basis, does not suggest a better fit per see. But if we notice the scales on the y-axes of the residual plots, we see that the residuals for the nonlinear model are much smaller. This does not necessarily imply a better model fit, but it does imply that the model is a more accurate prediction model. Table 3.1 also shows that the F-test of the linear model against the nonlinear model is highly significant.

Note how we did not pursue a nonlinear expansion of only those variables that appeared to be nonlinearly associated with the response. Instead we expanded all variables and made a joint assessment (using the F-test) of the significance of the nonlinearity. Using this tactic for dealing with nonlinear expansions we avoid to some extent the common pitfall of cherry picking random wrinkles in the data as nonlinear relations.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
98	16954.86				
87	676.45	11	16278.41	190.33	0.0000

Table 3.1: F-test for the linear model against the nonlinear model.

Missing values

Missing values for some variables is an ubiquitous problem for real data sets, and the data analyst must make a conscious decision about how missing values should be treated. In some cases, throwing away observations with missing values is a viable solution, which in the worst case reduces the power of the analysis, but doesn't do any serious harm. In other cases, throwing away observations – even if it is only a small fraction – can seriously bias the conclusions. To get an idea about whether missing values form a serious problem or not, think about the three scenarios below in relation to the birth weight case. Try to explain what the problem is, and how serious it is in each case.

- 10% of the women did not answer on their alcohol consumption because it was the last question, and the time to complete the interview was to short.
- 10% of the women did not answer on their alcohol consumption because they were smokers and did not want to give information on their alcohol consumption too.
- 10% of the women did not answer on their alcohol consumption because they had been drinking more than 5 drinks per week.

To deal with missing values we need to understand mechanisms of missingness. That is, why the values are missing. We focus the discussion here on missing values for the predictor variables only. There are three overall mechanisms.

Missing completely at random (MCAR) means in words that the reason that X_j is missing is unrelated to the actual value of the entire vector X. A mechanism for generating MCAR missingness is that a (loaded) coin flip independent of the data and with success probability p_j determines if the value of the j'th variable is missing. MCAR can also be expressed mathematically as follows. If Z_j denotes the indicator variable of whether X_j is missing then

$$Z_j \perp\!\!\!\perp X_1, \ldots, X_m.$$

If a woman did not answer on her alcohol consumption because time ran out, then alcohol consumption is MCAR – provided that time usage is independent of the responses to the other questions.

Missing at random (MAR) means in words that the reason that X_j is missing is unrelated to the actual values of the variables with missing values in X. We can construct a mechanism for MAR missingness as follows. We denote the observed variables (those with non-missing values) X_z for an index set $z \subseteq \{1, \ldots, m\}$, where z is a random set with probability $p(z|X) = p(z \mid X_z)$ depending only on the observed variables. MAR can be expressed mathematically as

$$(Z = z) \perp\!\!\!\perp X \mid X_z$$

for all z, where Z is the random index set of the observed variables in X. The terminology – missing at *random* – is misleading because it covers mechnisms that allows for quite systematic (that is, non-random) forms of missingness. If a woman did not answer on her alcohol consumption because she was a smoker and did not want to give information on alcohol consumption in this case, then alcohol consumption is MAR.

Missing not at random (MNAR) or informative missingness is everything that is not MCAR. If a woman did not answer on her alcohol consumption because she had been drinking more than 5 drinks per week, then alchohol consumption is MNAR. This is problematic, and a subsequent data analysis can only really be carried out correctly if we know the actual missingness mechnism. Something we never do.

Removing observations with missing values under MCAR will not bias the results of the data analysis, but it can be an inefficient usage of the data available. Especially if the fraction of observations with missing values is large but the missingness pattern is sparse.

If we just remove observations with missing values under a MAR mechanism, we can get biased results from the data analysis. However, MAR allows us to predict, or impute, the missing values using the conditional distribution $X_{z^c} \mid X_z$ where we condition on observed values. The imputation can be done for continuous variables by, for instance, imputing the conditional mean

$$\hat{X}_{z^c} = E(X_{z^c} \mid X_z)$$

into the data set. For discrete variables the most probable value may be imputed. Since MCAR implies MAR, imputation can also be used under MCAR.

The assumption MCAR allows for imputation without specifying the precise mechanism of missingness but only that it fulfills certain distributional properties.

In practice, we don't know the conditional distribution $X_{z^c} \mid X_z$, and it will have to be estimated from the data. This can in itself be done using regression models as developed in this book. For this to be reliable, we have to have sufficient amounts of observations where we actually observe X_{z^c} and X_z. If alcohol consumption is systematically missing for smokers, we cannot build a model to impute alcohol consumption when it is missing. If, however, it is only half the smokers that don't report alcohol consumption, then we can build a model of alcohol consumption for smokers under the MCAR assumption.

If the missingness pattern is not sparse, the imputation may artificially inflate the information in the data. That is, the imputed data set will not reflect the random variation correctly, which may lead to overly optimistic confidence intervals and too small p-values. To remedy this problem, the imputation can be done by randomly sampling values to impute from the conditional distribution $X_{z^c} \mid X_z$. If we want to get rid of the arbitrariness introduced by the random sampling, we can do multiple imputations and average the resulting parameter estimates.

Robustness

All statistical methods are based on some assumptions about the distribution of data. In this book, these assumptions are spelled out in detail and their individual consequences are studied. A method does not necessarily break down just because the assumptions are not fulfilled, but it is at risk of doing so, if the deviations become too large. It is therefore necessary to have ways of discovering and dealing with such deviations. Through initial exploratory data analysis and model diagnostics it is possible to discover, for instance, suspicious extreme values or deviations from distributional assumptions.

There is a whole research field on robust statistical methods that grew out of the concern that misspecified distributional assumptions and extreme observations such as outliers can affect the results of a data analysis severely. Indeed, a single outlier can have a heavy influence on the fitted model if we use least squares estimation. The quadratic loss function is sensitive to large deviations from the regression model, and the consequence is that an outlier influences the least squares estimator a lot. For the linear model this influence is measured by the leverage.

Whether extreme observations are outliers or consequences of a heavy tailed error distribution instead of a normal error distribution, say, is not easy to decide. In either case, the substitution of the quadratic loss with a loss function, such as the the absolute deviation $|Y_i - X_i^T \beta|$, that is less sensitive to large deviations, can make the resulting estimator more resilient to extreme observations.

If the model fitting is going to be implemented in an automatic system where model diagnostics cannot be performed, then robust methods should be considered. The same holds true if thousands or millions of models are fitted in parallel[6], in which case it is impossible for the data analyst to study diagnostics for each model. If the data analysis is carried out iteratively on a single data set, then the sensitivity of the model to the individual data points can be assessed by diagnostic tools, and the data analyst can take appropriate actions, e.g. disregarding or down weighting extreme observations, or modifying the distributional assumptions to fit the distribution of the data better.

Strategic considerations

This section treats the strategic decision process. The strategy defines the overall approach to the data collection process, the data analysis and the regression modeling. The focus is on context, pur-

[6] This happens on an everyday basis in genomics today

pose, interpretation and deployment. How does the modeling fit into the greater picture? There are several questions that should be asked before we launch into the actual data analysis. Is it, for instance, important that the model gives an interpretable relation between the response and the predictors, or are interpretations irrelevant? Is there a specific scientific hypothesis that should be addressed, or is the modeling of a more exploratory nature? Should predictive performance be optimized at all costs or are there constraints – financial, practical or idiosyncratic[7] – that limit the usefulness of some models over others? The data analyst should also think about the form of the outcome from the analysis and modeling. Should the results be presented in a written report, or is the model going to be implemented in an automatic prediction system? Should actionable conclusions be emphasized (are you supporting business decisions), or should descriptive facts dominate (are you part of a scientific project)? A predictive regression model can be a useful model in many circumstances, but it also has its limitations, and the analyst should recognize those.

Statistical inference or predictive modeling

A distinction is made in data analysis between statistical inference and predictive modeling. Explicit interpretations of fitted models and parameters, and conclusions drawn about populations, say, on the basis of models, are traditionally known as statistical inference. The use of data analysis for the specific purpose of prediction of (future) unobserved variables is known as predictive modeling. If the data analysis is supposed to answer inferential questions, a black box prediction model will most likely not be satisfactory. Similarly, if the purpose is to deploy a model in an industrial or business process for automatic predictions, an analysis of inferential questions only will probably not do the job. The overall strategy to be used for the data analysis is tightly intertwined with the purpose of the data analysis. But rather than regarding inference and prediction as two disjoint objectives for regression modeling, inference and prediction should be seen as dual viewpoints on data analysis that can mutually benefit from each other.

The classical example is a randomized trial with a treatment and

[7] In an application field it may be quite a lot easier to get results communicated and accepted if a standard model class is used, even if that model class appears inferior from a predictive viewpoint.

a control group and a well specified response variable. The data analysis should document a difference in the distribution of the response between the two groups – if there is one – and provide an effect size estimate on a population level. The average difference between a treated and an untreated population could, for instance, be estimated. Statistical inference procedures must provide us with information on how strong the statistical evidence is for the effect, and how uncertain the estimated effect size on population level is. We could instead ask for a prediction of the response variable depending upon whether an individual is treated or not. This would be a predictive modeling formulation. The inferential approach focuses on whether the treatment has a documentable effect and what the effect size is. The actual distribution of the respose is less important. To judge how large the effect size is from a predictive viewpoint, we need to contrast it with the response distribution. The predictive modeling approach focuses on the response distribution and to what extent it can be explained by the treatment.

Causality and interpretations of predictive models

As mentioned in the introduction, a regression model does not in itself imply any particular causal relations. A regression coefficient estimated in a linear regression model does not represent the effect on the response of changing the value of the corresponding predictor by an intervention. Suppose that we measure the two variables Y and X for a number of individuals in a population, and that the linear model

$$Y = \beta_0 + \beta X + \varepsilon$$

actually represents a causal relation between the variables. To make the example concrete, X could be the average calorie intake per day of an individual and Y his or hers body fat percentage. The causal interpretation means that an increase of the individuals average calorie intake will increase the body fat percentage. There is, however, a difference between the two genders. Women have naturally a higher body fat percentage than men but generally a lower calorie intake. The following simulation illustrates what can happen if we have samples from both genders and fit a regression model without taking the gender difference into account.

```
n <- 50
x1 <- rnorm(n, 2200, 200)
x2 <- rnorm(n, 1800, 200)
simpson <- data.frame(
  x <- c(x1, x2),
  y <- c(0.01 * (x1 + 100) + rnorm(n, sd = 4),
         0.01 * (x2 + 1200) + rnorm(n, sd = 4)),
  gender = rep(c("male", "female"), each = n)
)
qplot(x, y, data = simpson) +
  geom_smooth(method = "lm")
```

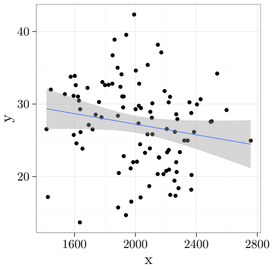


Figure 3.4: Scatter plot and fitted regression model for a data set with a hidden gender grouping.

The linear regression model fitted without taking gender into account has a negative slope as Figure 3.4 shows. The "effect" of calorie intake in the model has the opposite sign of the causal effect, and the model seems to suggest that an increase of average calorie intake will decrease the body fat percentage. If we adjust for gender, the conclusion is reversed to what is expected as Figure 3.5 shows.

```
qplot(x, y, data = simpson, color = gender) +
  geom_smooth(method = "lm") +
  scale_color_discrete(guide = "none")  ## Removes legend
```

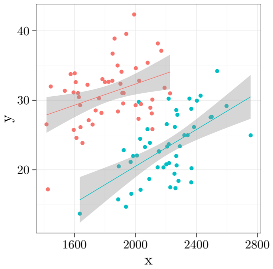


Figure 3.5: The same scatter plot as above but with the gender grouping exposed. The regression models are fitted within each gender group.

The phenomenon that the simulation illustrates is known as the *Yule-Simpson effect*. This is the phenomenon that the direction of a relation or trend between two variables within subgroups disappears or is reversed when the groups are combined. It is often referred to as Simpson's paradox.

When we think about the Yule-Simpson effect from a predictive viewpoint, it is not paradoxical at all. The model above that predicts body fat percentage from average calorie intake is not wrong. It is as good as it gets when we only observe average calorie intake. Because of the association between calorie intake and gender, the calorie intake variable encodes some information about gender, which in this case translates into a prediction of body fat percentage. A larger calorie intake indicates a male, who has a lower body fat percentage. When we include gender in the model, we can make more accurate predictions. The association between calorie intake and gender results in the sign-change of the calorie regression coefficient when gender in included. This does not make the model more correct as a predictive model, but it does make it more accurate. It is possible to have recursive[8] alternating Yule-Simpson

[8] In theory, this can be an infinite recursion.

effects, where subgroups can be further subdivided to reverse the sign again. We can in practice never be sure we have not missed a relevant variable in our data collection that could expose a Yule-Simpson effect.

The existence of the Yule-Simpson effect warn us about causal interpretations of regression models. We can never from the regression model itself know if there are unknown subgroups within which a regression coefficient has a different sign compared to the overall sign. We can therefore not tell from the data analysis alone if we have found the correct causal relation between the response and the predictors.

From the discussion above we conclude that a predictive regression model does not generally allow for a causal interpretation, that is, it does not necessarily explain the underlying mechanisms of how the observed variables are related. Sometimes the only purpose of the model is to be predictive, in which case we are perfectly happy with the model being a black prediction box without any causal interpretations. Few people really care about the mechanisms of spam email generation, but we care a lot about the performance of our spam email filter. Spam filters are good examples of models that first and foremost must be good predictors. In other cases it is quite important to understand the mechanisms behind certain observations. This is in particular the case if we want to *intervene* on one or more variables. It is, for instance, well known that education level is a strong predictor of income, but are we ever interested in predicting income based on education level? Are we not more likely to be interested in how education affects income – for an individual as well as for a population? Even if we have an accurate prediction model of income given education level, an increase of the general education level in the population may not result in a corresponding increase of income – as the model would otherwise predict. For a predictive model to be accurate certain distributional assumptions must be fulfilled. If we, for instance, sample a random individual from the population and predict her income based on her education level we get an accurate prediction. If we *intervene* and change the education level in the population we might not observe a corresponding effect on the income level. In this particular example income as well as education may be partly determined by unobserved factors such as social background and native intelligence, which will remain unaffected by changes in education level. Whether a model explains mechanisms, and allows for computations of intervention effects or not, cannot be turned into a purely mathematical or statistical question. It is a problem that is deeply entangled with the subject matter field to which the model is applied.

Causal modeling is an important and active research field, and many of the ideas have been around for decades. The Rubin causal model is a framework for thinking about causality in terms of what is called potential outcomes or counterfactual conditions. That is, what would we have observed had the conditions be different. Other ideas include path analysis developed by geneticist Sewall Wright and structural equation models (SEMs). More recently, Judea Pearl's book *Causality*[9] has been very influential on the development of the field. One main difference from predictive modeling is that causal modeling is concerned with predictions of intervention effects. Predictions are thus important in causal modeling as well, but in a setup that may differ from the setup we have data from. An intervention is not the same as a conditioning, and to formulate an intervention calculus it is necessary to formulate a general causal structure between all the variables considered and possibly also unobserved variables. We will not have the space to develop the causal modeling ideas. We discuss the randomized experiment below as one example where causal conclusions can be drawn from a regression model, but it is not the only case.

It is necessary once again to to warn against causal (mis)interpretations of predictive models. A phrase like[10] "the *effect* of the mother drinking more than 8 cups of coffee per day during her pregnancy is a reduction of the birth weight by 142 gram *all other things being equal*" is problematic. At least if it is, without further considerations, taken to imply that a mother's choice of whether or not to drink coffee can affect the birth weight by 142 gram. The *all other things being equal* condition does not save the day. In a technical model sense it makes the claim correct, but it may be impossible to keep all other (observed) variables fixed when intervening on one variable. More seriously[11], a variable may be affected by, or may affect when intervened upon, an unobserved variable related to the response. A generally valid interpretation of a regression coefficient

[9] Judea Pearl. *Causality*, Cambridge University Press, Cambridge, 2009

[10] See the birth weight case study, p. 19.

[11] Since issues related to variables we don't have data on are difficult to address.

is that it quantifies a difference between subpopulations – and not the effect of moving individuals from one subpopulation to another. The documentation that such differences exist, and the estimation of their magnitude, are important contributions to the understanding of causal relations, but it is a discussion we have to take within the subject matter field, and a discussion related to the variables we observe, their known or expected causal relations, and how the data was obtained.

Another example of the difficulties of drawing correct causal conclusions from a regression model is closely related to the birth weight case study in the previous chapter, where we found an association between smoking and birth weight. In the marginal analysis as well as in the full model, smoking was predictive of a lower birth weight[12]. In studies of infant mortality, babies with a low birth weight are found to have a higher mortality rate than other babies. In fact, low birth weight is a very strong predictor of infant mortality. Somewhat surprisingly, low birth weight babies of smoking mothers have a *lower* infant mortality than low birth weight babies of non-smoking mothers. This is known as the *low birth weight paradox*[13]. Does smoking then actually benefit the health of small babies? Not likely! Low birth weight is not the causal explanation of infant mortality, but rather an indicator of other unobserved causal risk factors. Hence even if smoking causes lower birth weight there is not necessarily any effect on the causal risk factors. The shift of the weight distribution due to smoking results in a *percieved* decrease of infant mortality for low birth weight babies of smoking mothers.

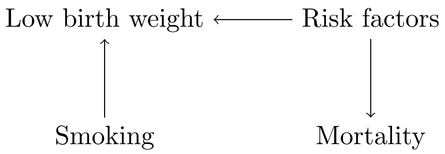


Figure 3.6: Causal diagram capable of explaining the low birth weight paradox. If low birth weight is caused by unobserved risk factors that also cause increased infant mortality, and if smoking causes a decrease of birth weight while having no effect on mortality or the risk factors, then the mortality risk for low birth weight babies of smoking mothers will be lower than for non-smoking mothers.

The difficulties with causality arise when data is from an *observational study*. That is, when the data collected are samples of the pairs (X_i, Y_i). This is sometimes called a "natural experiment", because it is the natural circumstances that dictate the "design", that

[12] With the knowledge of the Yule-Simpson effect in mind we understand that neither of these regression model "effects" in themselves imply that smoking causes the lower birth weight, though this may be a plausible hypothesis.

[13] Allen J Wilcox. On the importance—and the unimportance— of birthweight. *International Journal of Epidemiology*, 30(6): 1233–1241, 2001

is, the distribution of the values of the predictors X_i. The main alternative is a randomized designed experiment where the values of the predictors are fixed by a design. In both cases we use the same regression models of the conditional distribution of $Y_i \mid X_i$. In a randomized design, the values of the predictors are determined and fixed by the design, and these values are then assigned randomly to the experimental units[14]. The randomization is crucial, and its purpose is to break all relations between the response from the unit and unobserved variables, so that observed differences can be ascribed to the variation of the predictors.

It is important that the randomization is carried out correctly. It should be done by a randomization process that is completely independent of all aspects of the experimental units from the recruitment to all downstream experimental procedures applied. This is particularly important for biological samples that are processed in a laboratory. The randomization should be respected at this stage as well, as there are differences between laboratories as well as differences over time within laboratories. We must ensure that all potential laboratory effects are independent of the predictors. When the units are human subjects the experiment should also be blinded, which means that the subjects are not informed about the value of the predictors (whether they are, in fact, treated, what the dosage is etc.). This is to prevent that knowledge about the predictors induces behavioral differences or placebo effects, which will then compete with the actual values of the predictors in explaining differences in the response. Preferably the experiment is double blinded, which means that the experimenter – the medical doctor, say – does not know the values of the predictors either, so that his interaction with the subject cannot induce problems. It may even be argued that it should be triple blinded, which imply that the data analyst should not know the predictor values either[15]. Blinding the data analyst is sensible if it is important to avoid any potentially biased analysis decision, for instance, in the comparison of the effect of two drugs. It may, on the other hand, also prevent the data analyst from making informed decisions based on experience.

If we lump together samples for laboratory analysis depending on predictor values, we may introduce what is known as *batch effects*.

A regression model that fits data obtained from a correctly randomized experiment allows for a causal interpretation. Differences in the response predicted by the regression model for different val-

[14] An individual, a plot in an agricultural experiment, a test component in an industrial process etc.

[15] This can be achieved by scrambling the values of the predictors by a coordinatewise linear transformation, and not informing the data analyst about variable names

ues of the predictors are due to the differences in the predictors. A famous quote of G. E. Box[16] reads: *To find out what happens when you change something, it is necessary to change it.* So according to Box we need experiments where we change a variable to find out what happens when we change it. The randomized experiment is the gold standard for causal inference, and for finding out what happens when we change something. We should think this into the data collection process and see the decisions made about data collection as part of the data analysis strategy. If we need to draw causal conclusions, we should aim for a randomized experiment, and if this is not possible, we might need to consider alternatives to regression modeling where the causal questions can be answered properly.

Reports and deployment

When we analyze data and build regression models it should be done using reproducible methods. That is, we should document the analysis so that others or yourself can reproduce it later. This is well supported in R by using, for instance, the R package `knitr` as mentioned in the introduction. The resulting integration of code, analysis, results and text is, however, only suitable for like-minded people who need the details.

In many cases the results of a data analysis are to be summarized and written up in a report. Either to be published, to serve as documentation for officials, to serve decision makers, to form the basis for writing a scientific just to name some common purposes. In the writing of such a report, it is important to have the intended reader in mind. Many technical details that are treated in this book are irrelevant to the reader of the report, even if they are necessary for you to understand the techniques and tactics used in the data analysis. In the strategic decision process one needs to factor in the subsequent use of the analysis. You should, for instance, be able to communicate the results to the target reader. The data analysis is also, as described earlier in this chapter, an iterative process where decisions often have to be revised during the analysis. It is not very likely that the reader wants to read about the entire process, but rather the results of the process.

When you write a report you want to communicate the results

[16] George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for experimenters*, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2005

and some information about the models and methods used. The level of detail will depend on the target reader. You want to think very carefully about how the results are presented. It is easy to print out a table of parameter estimates – using the `summary` function, say – from a model fitted in R, but this is not the optimal way to communicate the model. To the very least, the output should be better formated, which can be achieved quite easily using the `xtable` package, and you also want to think about the order of the variables. You also need to be selective and not mindlessly present a table for every single model considered. If you include results and output from the analysis in the report make sure the you comment on it and that it is important for the reader to interpret the results of the analysis correctly. If it is not, then it is just clutter that removes the readers attention from the relevant results. If you include interactions, expansions or categorical predictors with many levels you definitely don't want a long list of parameter estimates that are difficult to interpret. It is much better if you can present the fitted model graphically. Figures and graphs summarizing the models, analysis and results are generally recommended over tables, as it is quite difficult for the human brain to see the overall picture from tables of numeric data except for very simple models. There is a tradeoff though. Some precision is lost in figures, so for numeric accuracy tables are preferable. In can be quite difficult and require a lot of work to find a good way to present a model, if it is not extremely simple. It is worth the effort, and with practical training you will become better. And use your critical common sense to evaluate your presentation.

In some cases, the data analysis will not need to be written up in a report. Instead, the resulting model will be used for automatic prediction. It might need to be integrated into already existing software. In this case, the ultimate deployment of the model can put restrictions on what modeling approaches are suitable. It may, for instance, be that R cannot be used for deployment of the model. The effort required to recode a prediction algorithm for a complicated model may easily outweigh the benefits it has over simpler models.

4 Generalized linear models

This chapter treats a generalization of the linear model that preserves much of its tractability, but can be adapted to cases where the linear model is inappropriate. A case study on the modeling of the sale of a particular item in a supermarket chain illustrates the practical use of generalized linear models. In this case study the response variable – the number of sold items – has a highly non-constant conditional variance. This can be captured within the framework of generalized linear models, which allows for a more flexible specification of the mean and variance relations between the response and the predictors than the linear model. This generalization of the linear model is tightly linked to non-normal choices of response distributions. Among the possible alternatives are discrete distributions like the binomial and the Poisson distribution. The theory in this chapter falls into two parts. The section *Exponential dispersion models* covers the class of response distributions that define generalized linear models. The two sections *Estimation theory* and *Sampling distributions* cover how a model is fitted to data and distributional properties of parameter estimates and test statistics, respectively. The content of these two last sections is an extension of similar results obtained in Chapter 2.

Model assumptions

The linear model assumptions A1–A3 can be generalized by allowing for nonlinear relations between a linear combination of the predictors and the mean and variance. This also allows for a non-constant variance. To motivate the generalization, we consider an example.

Example 4.1. For a binary response $Y \in \{0,1\}$ we have

$$E(Y \mid X) = P(Y = 1 \mid X) \in [0,1]$$

and

$$V(Y \mid X) = P(Y = 1 \mid X)(1 - P(Y = 1 \mid X)).$$

A linear model of the conditional expectation is problematic. The constraint that a probability must be in $[0,1]$ can only be enforced by restricting the parameter space. The restriction will depend on the values of the observed predictors, and it is generally impossible to ensure that all future predictions will be in $[0,1]$. Moreover, the variance is not constant but depends upon the expectation.

One solution is to consider a model of the form

$$P(Y = 1 \mid X) = \mu(X^T\beta)$$

with $\mu : \mathbb{R} \to [0,1]$ a given function[1]. Thus the model of the expectation is still given in terms of the linear combination $X^T\beta$, but we (nonlinearly) transform it using μ. By assumption, μ takes values in $[0,1]$.

The variance is

$$V(Y \mid X) = \mu(X^T\beta)(1 - \mu(X^T\beta))$$

and is given completely in terms of the mean. By introducing the variance function $\mathcal{V}(\mu) = \mu(1 - \mu)$ we get that

$$V(Y \mid X) = \mathcal{V}(\mu(X^T\beta)).$$

○

THE ABOVE EXAMPLE IS A PROTOTYPICAL generalized linear model. The expectation is a nonlinear transformation of $X^T\beta$, and the relation between the expectation and the variance is given by

[1] Typically a continuous and monotone function – a distribution function for instance. A possible choice is the logistic function.

mathematical necessity. If we model a binary variable, its variance is unavoidably linked to its expectation.

The assumptions GA1–GA3 below, that replace A1–A3 for generalized linear models, follow as natural abstractions of the example. To ease notation we introduce the *linear predictor* as

$$\eta = X^T \beta$$

for a vector of predictors X and a vector of parameters β.

GA1 The conditional expectation of Y given X is

$$E(Y \mid X) = \mu(\eta),$$

with $\mu : \mathbb{R} \to \mathbb{R}$ the *mean value function.*

GA2 The conditional variance of Y given X is

$$V(Y \mid X) = \psi \mathcal{V}(\mu(\eta)),$$

with $\mathcal{V} : \mathbb{R} \to (0, \infty)$ the *variance function* and $\psi > 0$ the *dispersion parameter.*

GA3 The conditional distribution of Y given X is the $(\theta(\eta), \nu_\psi)$-exponential dispersion distribution,

$$Y \mid X \sim \mathcal{E}(\theta(\eta), \nu_\psi).$$

The exponential dispersion distributions referred to in GA3 are introduced in the later section on exponential dispersion models, see page 87. They replace the normal distribution in Assumption A3, and form a broad class of possible response distributions. The normal distribution is a special case. As for the linear model, the assumption GA3 is a strong distributional assumption, which imply GA1 and GA2 for *specific* choices of μ and $\mathcal{V}$ that depend on how the map $\eta \mapsto \theta(\eta)$ is chosen and which the exponential dispersion model we choose. The situation is, however, a little more complicated than for the linear model. With arbitrary choices of μ, $\mathcal{V}$ and ψ we cannot always expect to be able to find a corresponding exponential dispersion model. Before we introduce the main case study in this chapter, we introduce another of the prototypical examples of a generalized linear model.

Example 4.2. If $Y \in \mathbb{N}_0$ is a counting variable then certainly

$$E(Y \mid X) \geq 0.$$

With a linear model of the conditional expectation we cannot easily enforce the positivity constraint. A possible solution is the model

$$E(Y \mid X) = e^{X^T \beta}$$

corresponding to $\mu = \exp$ being the exponential function. If $Y \mid X$ is Poisson distributed the variance function becomes

$$\mathcal{V}(\mu) = \mu,$$

because for the Poisson distribution the variance equals the mean. ○

If the mean value map is exp, the model is called a log-linear model because the log of the mean is linear in the parameters.

Advertising – a case study

In this case we consider data extracted from a large database from a major supermarket chain. The overall goal is to predict the number of items sold in those weeks where the chain runs an advertising campaign, which consists of a combination of advertisements and discounts. The item considered here is frozen vegetables, but in reality the objective would be to build a model for all the items in each store. For this example case the data available are from a few weeks and a selection of stores in the supermarket chain in Sweden.

The supermarket chain has an interest in a predictive model on several levels. We will focus on predictions for the individual store of how many items the store can expect to sell in a week. The predictive model will be based on the stores normal sale in the week in combination with campaign variables. All observations are from weeks with a discount on the item, but the discount differs between weeks.

Descriptive summaries

```
vegetables <- read.table(
  "http://www.math.ku.dk/~richard/regression/data/vegetablesSale.txt",
  header = TRUE,
  colClasses = c("numeric", "numeric", "factor", "factor",
                 "numeric", "numeric", "numeric")
)
summary(vegetables)

##       sale          normalSale         store        ad     
##  Min.   :  1.0   Min.   :  0.20   1      :   4   0:687  
##  1st Qu.: 12.0   1st Qu.:  4.20   102    :   4   1:379  
##  Median : 21.0   Median :  7.25   106    :   4          
##  Mean   : 40.3   Mean   : 11.72   107    :   4          
##  3rd Qu.: 40.0   3rd Qu.: 12.25   11     :   4          
##  Max.   :571.0   Max.   :102.00   110    :   4          
##                                   (Other):1042          
##     discount      discountSEK         week     
##  Min.   :12.1   Min.   : 2.40   Min.   :2.00  
##  1st Qu.:29.9   1st Qu.: 7.50   1st Qu.:7.00  
##  Median :37.0   Median :10.30   Median :7.00  
##  Mean   :37.5   Mean   : 9.95   Mean   :6.95  
##  3rd Qu.:46.0   3rd Qu.:12.80   3rd Qu.:8.00  
##  Max.   :46.0   Max.   :12.80   Max.   :9.00  
##  NA's   :26     NA's   :26                    
```

sale: Total number of sold items.

normalsale: A proxy of the normal sale in the same week.

store: The id-number of the store.

ad: Advertising (0 = no advertising, 1 = advertising).

discount: Discount in percent.

discountSEK: Discount in Swedish kroner.

week: Week number (2, 4, 5, 7, 8, 9).

Table 4.1: The 7 variables and their encoding in the vegetables data set.

From the summary we note that there are 26 missing observations of `discount` and `discountSEK`. A further investigation shows that it is the same 26 cases for which observations are missing for both variables. Moreover, 25 of these cases are the 25 cases from week 2. The last case is from week 4. This is the only case from week 4 registered as no advertising. We believe this to be an error in the database. The cross tabulation of `week` and `ad` in Table 4.2 shows that for all other cases, advertising is either 1 or 0 within each week.

	0	1
2	25	0
4	1	164
5	0	44
7	344	0
8	317	0
9	0	171

Table 4.2: Cross tabulation of week number and advertising indicator.

We remove the 25 cases from week 2 – we have no data to support an imputation in this case. We impute the discount (and correct what we believe to be an error) in the last case. Another modification is to reorder the levels for the factor `store`. The default ordering is the lexicographical order of the factor levels, which in this case amounts to a meaningless and arbitrary ordering. Instead, we order the levels according to the mean normal sale of the stores over the observed weeks.

```
vegetables <- subset(vegetables, week != 2)
naid <- is.na(vegetables$discount)
impute <- with(subset(vegetables, !is.na(discount) & week == 4),
  c(1, median(discount), median(discountSEK))
               )
vegetables[naid, c("ad", "discount", "discountSEK")] <- impute
```

```
vegetables <- within(vegetables, {
    meanNormSale <- sort(tapply(normalSale, store, mean))
    store <- factor(store, levels = names(meanNormSale))
    meanNormSale <- meanNormSale[store]
}
)
```

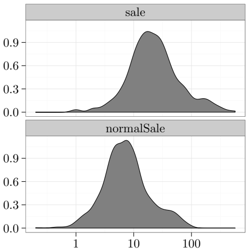


Figure 4.1: Density estimates of `sale` and `normalSale`. Note the log-scale on the x-axis.

The categorical variable `store` represents the total of 352 stores. Not all stores are represented each week. The stores are represented between 1 and 4 of the total of 5 weeks on which we have data.

```
mVegetables <- melt(vegetables[, c("sale", "normalSale")])
qplot(value, data = mVegetables, geom = "density",
      fill = I(gray(0.5)), xlab = "", ylab = "") +
  scale_x_log10() + facet_wrap(˜ variable, ncol = 1)
```

The marginal distributions of `sale` and `normalSale` are seen in Figure 4.1. It shows that the number of items sold is larger than the normal sale in a distributional sense, which is not surprising given that we are considering the sale in weeks with a discount. Note that we have log-transformed the variables because their distributions are quite right skewed. The distribution of the two categorical variables `ad` and `week` is given in the summary above, and the distribution of the remaining two variables `discount` and `discountSEK` is given as barplots in Figure 4.2. We note a quite skewed distribution with around one-third of the observations corresponding to the largest discount.

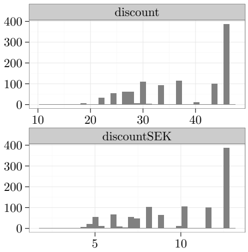


Figure 4.2: Barplots of discount variables.

```
mVegetables <- melt(vegetables[, c("discount", "discountSEK")])
qplot(value, data = mVegetables, geom = "bar",
      fill = I(gray(0.5)), xlab = "", ylab = "") +
  facet_wrap(˜ variable, scales = "free", ncol = 1)
```

Pairwise associations

We consider scatter plots and Pearson correlations between the 4 continuous variables. As for the histograms we log-transform `sale` and `normalSale`.

```
contVar <- c("sale", "normalSale", "discount", "discountSEK")
vegLog <- vegetables[, contVar]
vegLog <- transform(vegLog,
                    sale = log10(sale),
```

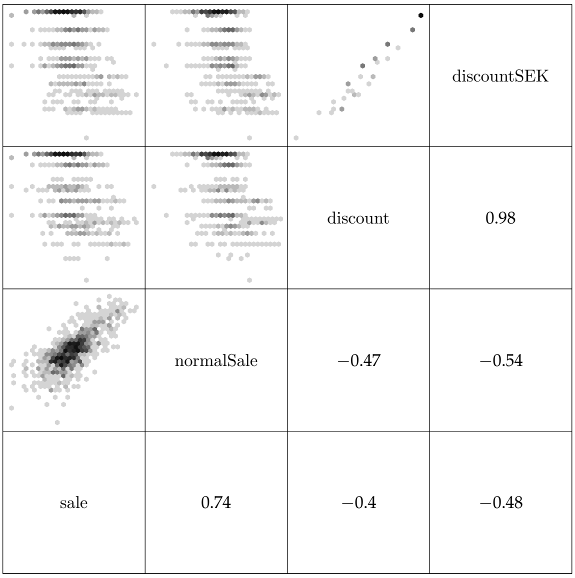


Figure 4.3: Scatter plot matrix of the 4 continuous variables and the corresponding Pearson correlations.

```
                    normalSale = log10(normalSale))
splom(vegLog,
      xlab = "",
      upper.panel = panel.hexbinplot,
      pscales = 0, xbins = 30,
      lower.panel = cor.print
)
```

The scatter plot matrix, Figure 4.3, shows that `sale` and `normal sale` are strongly correlated. As one should expect, `discount` and `discountSEK` are also extremely correlated and close to being collinear. What is, one the other, notable is, that the discount variables and the total sale are strongly negatively correlated. However, as the figure also shows, the `discount` variables are negatively correlated with the normal sale as well. Thus the negative marginal correlation may simply reflect that stores with a larger sale is only present with a smaller discount. Figure 4.4 further shows how the `discount` and `normalSale` distributions change depending upon the week number and the advertising variable.

The data consist of observations from different stores (of different size) in different weeks, and as mentioned above we don't have

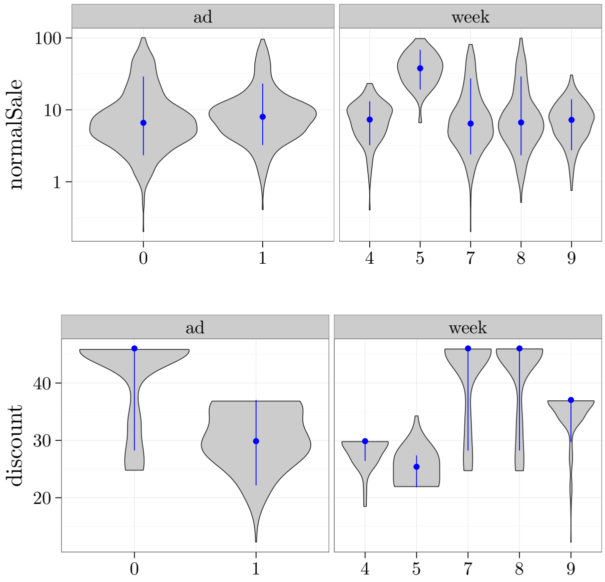


Figure 4.4: Violin plots, medians and interdecile ranges for the distribution of `normalSale` and `discount` stratified according to `ad` and `week`.

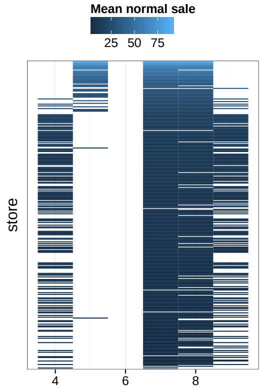


Figure 4.5: Illustration of which stores we have observations from for each of the weeks. The stores are ordered according to the mean normal sale.

observations from the same stores for all weeks. Since the advertising campaigns run for a given week, we don't have observations from all stores for all combinations of campaigns. In fact, Figure 4.5 shows that stores with a large normal sale are only included in weeks 5, 7 and 8, and that a considerable number of stores are only included in weeks 7 and 8.

```
ggplot(vegetables, aes(x = store, ymin = week - 0.5,
                       ymax = week + 0.5,
      group = store, color = meanNormSale)) + geom_linerange() +
  coord_flip() + scale_x_discrete(breaks = c()) +
  theme(legend.position = "top") +
  scale_color_continuous("Mean normal sale",
    guide = guide_colorbar(title.position = "top"))
```

To further illustrate this point we consider the relation between the stores and the discount. Figure 4.6 shows that for stores with a large normal sale we only have observations with moderate discounts. This can explain why we observe a negative marginal correlation between the discount and the sale as well as the normal sale.

```
qplot(discount, store, data = vegetables, geom = "point",
      group = store, color = meanNormSale) +
  scale_y_discrete(breaks = c()) + theme(legend.position = "top") +
  scale_color_continuous(guide = "none")
```

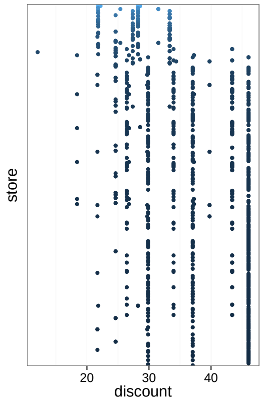


Figure 4.6: Illustration of the relation between the stores and the discount.

A Poisson regression model

The response variable Y is the number of sold items in a given week and in a given store. The proxy variable `normalSale` is based on historical data and provides, for the given store and week, an estimate of the expected number of sold items without taking advertising campaigns into account. We let N denote the normal sale. The normal sale is clearly predictive for the number of sold items, and we will attempt to build of model of how to adjust the normal sale with the information on the advertising campaign. We will consider models of the form

$$E(Y \mid N, X) = Ne^{X^T\beta} = e^{\log(N)+X^T\beta} \tag{4.1}$$

where X is the vector of additional predictors besides the normal sale. This is a log-linear model, but with one of the predictors having a fixed coefficient. Fixing the coefficient of a term in the linear predictor is achieved using the `offset` function in the formula specification of the model.

The log-linear model, that is, the model with an exponential mean value map, is the default choice for the Poisson response distribution when it is fitted using the `glm` function in R. As the very first thing we investigate the marginal association of each of the predictors `ad`, `discount`, `discountSEK` and `store` using the Poisson regression model.

```
form <- sale ~ ad + discount + discountSEK + store
nulModel <- glm(sale ~ offset(log(normalSale)),
                family = poisson,
                data = vegetables)
oneTermModels <- add1(nulModel, form, test = "LRT")
```

Table 4.3 shows the results of testing each of the predictors individually. Note that all models contain the fixed control for the normal sale. As we see from the table, the predictors are marginally significantly related to the number of sold items, though `discount`

Table 4.3: Marginal association tests sorted according to the p-value.

	Df	Deviance	LRT	Pr(>Chi)
store	351	8.824e+03	8132.31	0
discountSEK	1	1.695e+04	8.89	0.00287
ad	1	1.695e+04	6.94	0.00841
discount	1	1.695e+04	3.19	0.0743

is borderline. This conclusion rests on some model assumptions, in particular the relation between mean and variance that is determined by the Poisson model. Below, we find evidence against this relation in a more general model.

Note that `week` was not included as a predictor. First of all, it would be of limited use in a predictive model, if we can only predict for the weeks included in the data set. Second, the value of `ad` is completely determined by `week`, and including both would result in perfect collinearity of the `ad` column in the design matrix with the columns representing `week`. The `week` variable is, however, useful for subsequent model diagnostics.

The next thing we consider is a main effects model including the four predictors linearly and additively. We should remember that `discount` and `discountSEK` are highly correlated, and we should expect the collinearity phenomenon that neither of them are significantly related to the response when the other is included.

A main effects model.

```
form <- sale ~ offset(log(normalSale)) + store + ad +
  discount + discountSEK - 1
vegetablesGlm <- glm(form,
                     family = poisson,
                     data = vegetables)
```

Table 4.4: Summary table of parameter estimates, standard errors and t-tests for the poisson model of sale with 4 predictors.

	Estimate	Std. Error	z value	Pr(>|z|)
ad1	0.32	0.03	9.90	4.3e−23
discount	−0.08	0.02	−4.43	9.4e−06
discountSEK	0.42	0.06	7.01	2.4e−12

We do not want to consider all the estimated parameters for the individual stores – only the parameters for the other variables. Table 4.4 shows that in this model all three variables `ad`, `discount` and `discountSEK` are significant, but somewhat surprisingly, the estimated coefficient of `discount` is negative. This counter intuitive result makes us suspecious about the model. The collinearity might explain a negative estimate, but it should not be significant then. We return to this point later in the chapter where we have developed enough theory to do model diagnostics.

Exponential dispersion models

This section treats the theory of exponential dispersion models – a class of distributions on the real line. Each such model is a two-parameter family of distributions determined by an exponential family and a dispersion parameter.

We let ν be a σ-finite measure on $\mathbb{R}$ and define $\varphi : \mathbb{R} \to [0, \infty]$ by

$$\varphi(\theta) = \int e^{\theta y} \, \nu(\mathrm{d}y). \tag{4.2}$$

Define also

$$I = \{\theta \in \mathbb{R} \mid \varphi(\theta) < \infty\}^\circ$$

as the interior of the set of θ's for which $\varphi(\theta) < \infty$.

Note that it is possible that $I = \emptyset$ – take, for instance, ν to be the Lebesgue measure. If ν is a finite measure then $\varphi(0) < \infty$, but it is still possible that $\varphi(\theta) = \infty$ for all $\theta \neq 0$, which results in $I = \emptyset$. The case where I is empty is not of any relevance. There are two other special situations that are not relevant either. If ν is the zero measure, $\varphi(\theta) = 0$, and if ν is a one-point measure, that is, $\nu = c\delta_y$ for $c \in (0, \infty)$ and δ_y the Dirac measure in y, then $\varphi(\theta) = ce^{\theta y}$. Neither of these two cases will be of any interest, and they result in pathological problems that we want to avoid. We will therefore make the following regularity assumptions about ν throughout.

- The measure ν is not the zero meaure, nor is it a one-point measure.

- The open set I is non-empty.

By the assumption that ν is not the zero measure, it follows that $\varphi(\theta) > 0$ for $\theta \in I$. This allows us to make the following definition.

Definition 4.3. *The exponential family with structure measure ν is the one-parameter family of probability measures, ρ_θ for $\theta \in I$, defined by*

$$\frac{\mathrm{d}\rho_\theta}{\mathrm{d}\nu} = \frac{1}{\varphi(\theta)} e^{\theta y}.$$

The parameter $\theta \in I$ is called the canonical parameter.

Note that the exponential family is determined completely by the choice of the structure measure ν. Introducing $\kappa(\theta) = \log \varphi(\theta)$ we have that

$$\frac{\mathrm{d}\rho_\theta}{\mathrm{d}\nu} = e^{\theta y - \kappa(\theta)}$$

for $\theta \in I$. The function κ is called the *unit cumulant function* for the exponential family. It is closely related to the cumulant generating functions for the probability measures in the exponential family, see Exercise 4.4.

Under the regularity assumptions we have made on ν we can obtain a couple of very useful results about the exponential family.

Lemma 4.4. *The set I is an open interval, the parametrization $\theta \mapsto \rho_\theta$ is one-to-one, and the function $\kappa : I \mapsto \mathbb{R}$ is strictly convex.*

Proof. We first prove that the parametrization is one-to-one. If $\rho_{\theta_1} = \rho_{\theta_2}$ their densities w.r.t. ν must agree ν-almost everywhere, which implies that

$$(\theta_1 - \theta_2)y = \kappa(\theta_1) - \kappa(\theta_2)$$

for ν-almost all y. Since ν is assumed not to be the zero measure or a one-point measure, this can only hold if $\theta_1 = \theta_2$.

To prove that I is an interval let $\theta_1, \theta_2 \in I$, and let $\alpha \in (0,1)$. Then by Hölders inequality

$$\begin{aligned} \varphi(\alpha\theta_1 + (1-\alpha)\theta_2) &= \int e^{\alpha\theta_1 y} e^{(1-\alpha)\theta_2 y}\, \nu(\mathrm{d}y) \\ &\leq \left(\int e^{\theta_1 y}\, \nu(\mathrm{d}y)\right)^{\alpha} \left(\int e^{\theta_2 y}\, \nu(\mathrm{d}y)\right)^{1-\alpha} \\ &= \varphi(\theta_1)^{\alpha} \varphi(\theta_2)^{1-\alpha} < \infty. \end{aligned}$$

Hölders inequality says that

$$\int |fg| \leq \left(\int |f|^p\right)^{\frac{1}{p}} \left(\int |g|^q\right)^{\frac{1}{q}}$$

for $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. It is used with

$$\begin{aligned} f(y) &= e^{\alpha\theta_1 y}, \\ g(y) &= e^{(1-\alpha)\theta_2 y}, \end{aligned}$$

$$p = \frac{1}{\alpha} \quad \text{and} \quad q = \frac{1}{1-\alpha}.$$

This proves that I is an interval. It is by definition open.

Finally, it follows directly from the inequality above that

$$\kappa(\alpha\theta_1 + (1-\alpha)\theta_2) \leq \alpha\kappa(\theta_1) + (1-\alpha)\kappa(\theta_2),$$

which shows that κ is convex. If we have equality in this inequality, we have equality in Hölders inequality. This happens only if $e^{\theta_1 y}/\varphi(\theta_1) = e^{\theta_2 y}/\varphi(\theta_2)$ for ν-almost all y, and just as above we conclude that this implies $\theta_1 = \theta_2$. The conclusion is that κ is strictly convex. □

The structure measure ν determines the unit cumulant function by the formula

$$\kappa(\theta) = \log \int e^{\theta y}\, \nu(\mathrm{d}y).$$

The assumption that I is open implies that κ determines ν uniquely, see Exercise 4.5. In many cases the structure measure belongs to a family of σ-finite measures ν_ψ parametrized by $\psi > 0$, and whose cumulant functions are $\theta \mapsto \kappa(\psi\theta)/\psi$. That is, $\nu_1 = \nu$ and

$$\frac{\kappa(\theta)}{\psi} = \log \int e^{\frac{\theta y}{\psi}}\, \nu_\psi(\mathrm{d}y).$$

Since this cumulant function is defined on the same open interval I, it uniquely determines ν_ψ – if there exists such a ν_ψ. It is, however, not at all obvious whether there exists a σ-finite measure ν_ψ with cumulant function $\kappa(\psi\theta)/\psi$ for a given unit cumulant function κ. We will find this to be the case for all $\psi > 0$ for a number of concrete examples, but we will not pursue a systematic study. What we can say is that if there is such a family ν_ψ for $\psi > 0$, it is uniquely determined by ν, and that all such measures will satisfy the same regularity conditions we have required of ν. In this case, we introduce the *exponential dispersion model* determined by ν – a two-parameter family of probability measures – by

$$\frac{\mathrm{d}\rho_{\theta,\psi}}{\mathrm{d}\nu_\psi} = e^{\frac{\theta y - \kappa(\theta)}{\psi}}. \tag{4.3}$$

The parameter ψ is called the dispersion parameter, and we call $\nu = \nu_1$ the unit structure measure for the exponential dispersion model. For fixed ψ the exponential dispersion model is an exponential family with structure measure ν_ψ and canonical parameter θ/ψ. We abuse the terminology slightly and call θ the canonical parameter for the exponential dispersion model. Thus, whether θ/ψ or θ is canonical depends upon whether we regard the model as an exponential family with structure measure ν_ψ or an exponential dispersion model with unit structure measure ν. Whenever we consider an exponential dispersion model, the measure ν will always denote the unit structure measure ν_1.

The parameters ψ and θ are only uniquely determined up to a scale transformation, which corresponds to specifying which of the structure measures ν_ψ is the unit structure measure.

Definition 4.5. *The probability distribution $\rho_{\theta,\psi}$ is called the (θ, ν_ψ)-exponential dispersion distribution and is denoted $\mathcal{E}(\theta, \nu_\psi)$.*

In practice we check that a given parametrized family of distributions is, in fact, an exponential dispersion model by checking that it can brought on the form (4.3).

Example 4.6. The normal distribution $\mathcal{N}(\mu, 1)$ has density

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}} = e^{y\mu - \frac{\mu^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

w.r.t. the Lebesgue measure m. We identify this as an exponential family with[2]

$$\theta = \mu, \quad \kappa(\theta) = \frac{\theta^2}{2} \quad \text{and} \quad \frac{\mathrm{d}\nu}{\mathrm{d}m} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$$

The $\mathcal{N}(\mu, \sigma^2)$ distribution is an exponential dispersion model with dispersion parameter $\psi = \sigma^2$ and

$$\frac{\mathrm{d}\nu_\psi}{\mathrm{d}m} = \frac{1}{\sqrt{2\pi\psi}} e^{-\frac{y^2}{2\psi}}.$$

Thus with the notation introduced in Definition 4.5,

$$\mathcal{E}\left(\mu, \frac{1}{\sigma} \cdot m\right) = \mathcal{N}(\mu, \sigma^2)$$

○

Example 4.7. The Poisson distribution with mean $\mu > 0$ has density (point probabilities)

$$e^{-\mu} \frac{\mu^n}{n!}$$

w.r.t. the counting measure τ on $\mathbb{N}_0$. With $\theta = \log \mu$ we can rewrite this density as

$$e^{n \log \mu - \mu} \frac{1}{n!} = e^{n\theta - e^\theta} \frac{1}{n!},$$

and we identify the distribution as an exponential family with

$$\theta = \log \mu, \quad \kappa(\theta) = e^\theta \quad \text{and} \quad \frac{\mathrm{d}\nu}{\mathrm{d}\tau} = \frac{1}{n!}.$$

We conclude that

$$\mathcal{E}\left(\log \mu, \frac{1}{n!} \cdot \tau\right) = \mathrm{Poi}(\mu).$$

○

[2] The κ and ν are not unique. The constant

$$\frac{1}{\sqrt{2\pi}}$$

in ν can be moved to a $-\log \sqrt{2\pi}$ term in κ.

The choice of parametrization most convenient for an exponential family or an exponential dispersion model depends upon what we want to do. The canonical parametrization is convenient for theoretical considerations. It is, however, not always directly interpretable. For the normal distribution the canonical parameter happens to be equal to the mean. For the Poisson distribution the canonical parameter is the log-mean. Sometimes a third parametrization is convenient, such that the canonical parameter is given as a function, $\theta(\eta)$, of an *arbitrary* parameter η. We call η arbitrary because we don't make any assumptions about what this parameter is[3], just that it is used to parametrize the exponential dispersion model. The density for the exponential dispersion model w.r.t. ν_ψ in the arbitrary parametrization is

$$e^{\frac{\theta(\eta)y-c(\eta)}{\psi}}$$

where $c(\eta) = \kappa(\theta(\eta))$. The arbitrary parameter is allowed to take values in H, and the parameter function $\theta : H \to I$ is allowed to depend upon additional (nuisance) parameters. We suppress such dependences in the abstract notation. It is, however, not allowed to depend upon the dispersion parameter[4].

Example 4.8. The Γ-distribution with shape parameter $\lambda > 0$ and scale parameter $\alpha > 0$ has density

$$\frac{1}{\alpha^\lambda \Gamma(\lambda)} y^{\lambda-1} e^{-y/\alpha} = e^{\frac{-y/(\lambda\alpha)-\log(\lambda\alpha)}{1/\lambda}} \frac{\lambda^\lambda}{\Gamma(\lambda)} y^{\lambda-1}$$

w.r.t. the Lebesgue measure m on $(0,\infty)$. We identify this family of distributions as an exponential dispersion model with dispersion parameter $\psi = 1/\lambda$, canonical parameter

$$\theta = -1/(\lambda\alpha) < 0,$$

$$\kappa(\theta) = -\log(-\theta),$$

and structure measure given by

$$\frac{\mathrm{d}\nu_\psi}{\mathrm{d}m} = \frac{1}{\psi^{1/\psi}\Gamma(1/\psi)} y^{\frac{1}{\psi}-1}$$

on $(0,\infty)$. We have $I = (-\infty, 0)$. The mean value of the Γ-distribution is $\mu := \alpha\lambda = -\frac{1}{\theta}$, and the variance is $\alpha^2\lambda = \psi\mu^2 = \psi/\theta^2$.

[3] In most cases it will be the linear predictor, though.

[4] Otherwise the dispersion parameter cannot be eliminated from the estimation equations.

The Γ-family of distributions is an example where we have one parametrization in terms of shape and scale, a different parametrization in terms of the canonical parameter and the dispersion parameter, and yet another parametrization in terms of mean and variance. $\circ$

The different parametrizations can be related through differentiation.

For the Γ-distribution:

$$\kappa'(\theta) = -\frac{1}{\theta}$$

and

$$\kappa''(\theta) = \frac{1}{\theta^2}.$$

Theorem 4.9. *The function κ is infinitely often differentiable on I. If $Y \sim \mathcal{E}(\theta, \nu_\psi)$ for $\theta \in I$ then*

$$EY = \kappa'(\theta) \tag{4.4}$$

and

$$VY = \psi\kappa''(\theta). \tag{4.5}$$

Proof. It follows by a suitable domination argument that, for $\theta \in I$,

$$\frac{\mathrm{d}^n}{\mathrm{d}\theta^n}\varphi(\theta) = \int y^n e^{\theta y}\,\nu(\mathrm{d}y) = \varphi(\theta)EY^n.$$

Since $\kappa(\theta) = \log\varphi(\theta)$, it follows that also κ is infinitely often differentiable. In particular,

$$EY = \frac{\varphi'(\theta)}{\varphi(\theta)} = (\log\varphi)'(\theta) = \kappa'(\theta),$$

$$\begin{tikzcd}
J & H \arrow[l, "\mu"'] \arrow[dl, "\theta"] \\
I \arrow[u, "\kappa'"] &
\end{tikzcd}$$

Figure 4.7: Relations between the three parametrizations.

and

$$EY^2 = \frac{\varphi''(\theta)}{\varphi(\theta)},$$

and we find that

$$VY = EY^2 - (EY)^2 = \frac{\varphi''(\theta)\varphi(\theta) - \varphi'(\theta)^2}{\varphi(\theta)^2} = \kappa''(\theta).$$

This proves the theorem for $Y \sim \mathcal{E}(\theta, \nu_1)$. The general case can be proved by similar arguments. Alternatively, observe that for an arbitrary dispersion parameter $\psi > 0$, the distribution of Y is an exponential family with canonical parameter $\theta_0 = \theta/\psi$ and with corresponding cumulant function

$$\kappa_0(\theta_0) = \frac{\kappa(\psi\theta_0)}{\psi}.$$

The conclusion follows by differentiating κ_0 twice w.r.t. θ_0. □

We already know from Lemma 4.4 that κ is strictly convex, which actually implies, since κ was found to be differentiable on I, that κ' is a strictly increasing function. This can also be seen directly from the previous theorem. Indeed, by the regularity assumptions on ν_ψ, $\rho_{\theta,\psi}$ is not a Dirac measure, which implies that $VY > 0$ in (4.5), see Exercise 4.3. Hence κ'' is strictly positive, and it follows that κ' is strictly increasing, and that κ is strictly convex. That κ' is continuous and strictly increasing imply that the range of κ', $J := \kappa'(I)$, is an open interval. This range is the range of possible mean values for the exponential dispersion model.

Since κ' bijectively maps I onto J we can always express the variance in terms of the mean.

Definition 4.10. *The variance function $\mathcal{V} : J \to (0, \infty)$ is defined as*

$$\mathcal{V}(\mu) = \kappa''((\kappa')^{-1}(\mu)).$$

Recall that in terms of an arbitrary parametrization the exponential dispersion model has the density

$$e^{\frac{\theta(\eta)y - c(\eta)}{\psi}},$$

where $\theta : H \to I$. The *mean value function*, in the arbitrary parametrization, is given as

$$\mu(\eta) = \kappa'(\theta(\eta)).$$

We can then express the mean and the variance functions in terms of the function $\eta \mapsto c(\eta)$.

Corollary 4.11. *If θ is twice differentiable as a function of η, and if $\theta'(\eta) \neq 0$, we have that*

$$\mu(\eta) = \frac{c'(\eta)}{\theta'(\eta)}$$

and

$$\mathcal{V}(\mu(\eta)) = \frac{c''(\eta)\theta'(\eta) - c'(\eta)\theta''(\eta)}{\theta'(\eta)^3} = \frac{\mu'(\eta)}{\theta'(\eta)}.$$

Proof. From the definition $c(\eta) = \kappa(\theta(\eta))$ we get by differentiation and (4.4) that

$$c'(\eta) = \kappa'(\theta(\eta))\theta'(\eta) = \mu(\eta)\theta'(\eta),$$

and the first identity follows. An additional differentiation yields

$$\begin{aligned} c''(\eta) &= \kappa''(\theta(\eta))\theta'(\eta)^2 + \kappa'(\theta(\eta))\theta''(\eta) \\ &= \mathcal{V}(\mu(\eta))\theta'(\eta)^2 + \frac{c'(\eta)\theta''(\eta)}{\theta'(\eta)}, \end{aligned}$$

and the second identity follows by isolating $\mathcal{V}(\mu(\eta))$. □

When the mean value function is bijective, its inverse, g, is called *the link function*. It maps the mean value μ to the arbitrary parameter η, that is, $\eta = g(\mu)$. The choice of the link function (equivalently, the mean value function) completely determines the θ-map and vice versa. The choice that makes $\theta = \eta$ plays a particularly central role. If $\theta = \eta$ then $\mu = \kappa'$, and the corresponding link function is thus $g = (\kappa')^{-1}$.

Definition 4.12. *The canonical link function is the link function*

$$g = (\kappa')^{-1}.$$

Example 4.13. For the Poisson distribution we have $\kappa(\theta) = e^{\theta}$, see Example 4.7. This implies that $\kappa'(\theta) = e^{\theta}$ and the canonical link function is, for the Poisson distribution, $g(\mu) = \log \mu$. ○

Example 4.14. For the Γ-distribution we have from Example 4.8 that $\kappa(\theta) = -\log(-\theta)$ and

$$\kappa'(\theta) = -\frac{1}{\theta}.$$

This gives that the canonical link function for the Γ-distribution is

$$g(\mu) = -\frac{1}{\mu}.$$

○

$$\begin{tikzcd} \mu \arrow[r, bend right, "g"'] & \eta \arrow[l, bend right, "\mu"'] \arrow[dl, bend left, "\theta"] \\ \theta \arrow[u, "\kappa'"] & \beta \arrow[u, bend right, "X^T"'] \end{tikzcd}$$

Figure 4.8: The linear predictor, $\eta = X^T\beta$, enters the exponential dispersion model through the mean value function.

Returning to regression, the response distribution is specified in terms of the linear predictor $\eta = X^T\beta$, that determines the mean, and an exponential dispersion model, that determines the remaining parts of the distribution. The specification of the mean is given in terms of the mean value map, or equivalently in terms of the link function. That is, with link function g,

$$g(E(Y \mid X)) = \eta = X^T\beta$$

or

$$E(Y \mid X) = \mu(\eta).$$

This specifies implicitly the canonical parameter in the exponential dispersion model, and results in a model with

$$V(Y \mid X) = \psi \mathcal{V}(\mu(\eta))$$

where $\mathcal{V}$ is the variance function.

Deviance

In this section we introduce the (unit) *deviance* for exponential dispersion families. It can be understood as a generalization of the squared error

$$(Y - \mu)^2$$

for the $\mathcal{N}(\mu, 1)$ distribution. For an exponential dispersion model, Lemma 4.22 shows that the log-likelihood for $\psi = 1$ is

$$\ell_Y(\theta) = \theta Y - \kappa(\theta).$$

The deviance is defined as twice the negative log-likelihood up to an additive constant. In terms of the canonical parameter we introduce the *unit deviance*[5] as

$$d(Y, \theta) = 2\left(\sup_{\theta' \in I} \ell_Y(\theta') - \ell_Y(\theta)\right)$$

for $Y \in \overline{J}$ and $\theta \in I$. The supremum in the definition is attained in $g(Y)$ for $Y \in J$ where $g = (\kappa')^{-1}$ is the canonical link. For practical computations it is more convenient to express the unit deviance in terms of the mean value parameter.

Definition 4.15. *The unit deviance is*

$$d(Y, \mu) = 2\left(\sup_{\mu' \in J}\left\{g(\mu')Y - \kappa(g(\mu'))\right\} - g(\mu)Y + \kappa(g(\mu))\right)$$

for $Y \in \overline{J}$ and $\mu \in J$.

For $Y \in J$, the supremum is attained in $\mu' = Y$ and the unit deviance can be expressed as

$$d(Y, \mu) = 2\left(Y(g(Y) - g(\mu)) - \kappa(g(Y)) + \kappa(g(\mu))\right).$$

[5] Here, *unit* refers to the dispersion parameter being 1.

Note that by the definition of the unit deviance we have that $d(Y,Y) = 0$ and

$$d(Y,\mu) > 0$$

for $\mu \neq Y$.

The deviance has simple analytic expressions for many concrete examples, which are interpretable as measures of how the observation Y deviates from the expectation μ.

Example 4.16. For the normal distribution the canonical link is the identity and $\kappa(\theta) = \theta^2/2$, hence

$$d(Y,\mu) = 2(Y(Y-\mu) - Y^2/2 + \mu^2/2) = (Y-\mu)^2.$$

○

Example 4.17. For the Poisson distribution the canonical link is the logarithm, and $\kappa(\theta) = e^\theta$, hence

$$d(Y,\mu) = 2(Y(\log Y - \log \mu) - Y + \mu) = 2(Y \log(Y/\mu) - Y + \mu)$$

for $Y,\mu > 0$. It is clear from the definition that $d(0,\mu) = 2\mu$, and the identity above can be maintained even for $Y = 0$ by the convention $0 \log 0 = 0$. ○

Example 4.18. For the binomial case it follows directly from the definition that for $Y = 1,\ldots,m$ and $\mu \in (0,m)$

$$\begin{aligned} d(Y,\mu) &= 2\Big(Y \log Y/m + (m-Y)\log(1-Y/m) \\ &\qquad -Y\log\mu/m - (m-Y)\log(1-\mu/m)\Big) \\ &= 2\Big(Y\log(Y/\mu) - (m-Y)\log\left((m-Y)/(m-\mu)\right)\Big). \end{aligned}$$

Again, by the convention $0 \log 0 = 0$ the identity is seen to extend also to the extreme cases $Y = 0$ or $Y = m$. ○

The unit deviance is approximately a quadratic form for $Y \simeq \mu$.

Theorem 4.19. *For the unit deviance it holds that*

$$d(Y,\mu) = \frac{(Y-\mu)^2}{\mathcal{V}(\mu)} + o((Y-\mu)^2)$$

for $Y,\mu \in J$.

Proof. We consider the function

$$Y \mapsto d(Y, \mu)$$

around μ. Since we know that $d(\mu, \mu) = 0$, that $Y = \mu$ is a local minimum, and that d is twice continuous differentiable in $J \times J$, we get by Taylors formula that

$$d(Y, \mu) = \frac{1}{2}\partial_Y^2 d(\mu, \mu)(Y - \mu)^2 + o((Y - \mu)^2).$$

Using that $\partial_Y \kappa(g(Y)) = Yg(Y)$ we find that

$$\begin{aligned} \frac{1}{2}\partial_Y^2 d(Y, \mu) &= \partial_Y^2 \Big\{ Yg(Y) - \kappa(g(Y)) - Yg(\mu) \Big\} \\ &= \partial_Y \Big\{ g(Y) + Yg'(Y) - Yg'(Y) - g(\mu) \Big\} \\ &= g'(Y) = \frac{1}{\kappa''(Y)} = \frac{1}{\mathcal{V}(Y)}. \end{aligned}$$

Plugging in $Y = \mu$ completes the proof. □

Model diagnostics

As for the linear model we base model checks and model diagnostics on residuals, but for generalized linear models there are several possible choices. We assume in this section that we have observations $Y_1, \ldots, Y_n$ that are used to fit a generalized linear model, which gives the fitted mean values $\hat{\mu}_i, \ldots, \hat{\mu}_n$. For the i'th observation Y_i we could first of all consider the raw residual

$$Y_i - \hat{\mu}_i.$$

In the R terminology the raw residual is called the *response residual.* Whenever the variance function is not a constant, the raw residual is not particularly useful. To take a non-constant variance function into account, a natural choice of residual is the *Pearson residual* defined as

$$\frac{Y_i - \hat{\mu}_i}{\sqrt{\mathcal{V}(\mu_i)}}.$$

The sum of the squared Pearson residuals is known as the Pearson χ^2-statistic. Pearson residuals are used in the same way as the raw or standardized residuals are used for the linear model. We plot the

Pearson residuals againts the fitted values or a predictor variable to visually check the model assumptions – the residuals should show no distributional dependence upon what we plot it against. Systematic trends show that GA1 is not fulfilled, and variance inhomogeneity shows that the variance function in GA2 is not correct. We cannot expect, however, that the Pearson residuals appear to have a normal distribution, nor is it clear what the distribution of the residuals is even if the model assumptions are fulfilled.

The Pearson residual is based on the mean and variance assumptions GA1 and GA2 only. Based on the distributional assumption GA3 we can also introduce the *deviance* residual for the i'th observation as

$$\operatorname{sign}(Y_i - \hat{\mu}_i)\sqrt{d(Y_i, \hat{\mu}_i)}.$$

We will later introduce the deviance statistic, which is the sum of the squared deviance residuals – just as the χ^2-statistic is the sum of the squared Pearson residuals. The deviance residuals can be used like the Pearson residuals, and Theorem 4.19 gives an approximate relation between them.

Finally, the *working* residual should be mentioned. It is defined for the i'th observation as

$$\frac{Y_i - \hat{\mu}_i}{\mu'(\hat{\eta}_i)},$$

and is the raw residual from the weighted least squares problem (4.11).

Example 4.20. For the binomial case we find that the raw residual is

$$Y_i - m_i \hat{p}_i$$

where $\hat{p}_i$ is the estimate of the success probability for the i'th observation. The Pearson residual is

$$\frac{Y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i)}},$$

and the deviance residual is

$$\operatorname{sign}(Y_i - m_i \hat{p}_i)\sqrt{2Y_i \log \frac{Y_i}{m_i \hat{p}_i} + 2(m_i - Y_i) \log \frac{m_i - Y_i}{m_i(1 - \hat{p}_i)}}.$$

○

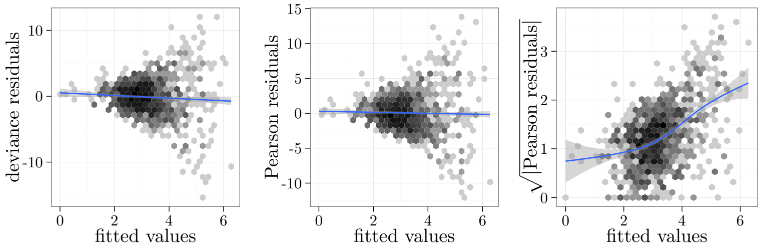


Figure 4.9: Diagnostic plots. Deviance residuals, Pearson residuals and the square root of the absolute value of the Pearson residuals plotted against fitted values.

Advertising revisited

The first analysis of the advertising data ended with a fitted log-linear Poisson regression model. We were somewhat suspicious of this model. Before we consider any nonlinear and interaction effects we will investigate the model fit. For this purpose we consider residual plots using the deviance residuals and the Pearson residuals.

```
vegetablesDiag <- transform(vegetables,
                            .fitted = predict(vegetablesGlm),
                            .deviance = residuals(vegetablesGlm),
                            .pearson = residuals(vegetablesGlm,
                                                 type = "pearson")
)
p1 <- qplot(.fitted, .deviance, data = vegetablesDiag,
          geom = "hex") + binScale + geom_smooth(size = 1) +
  xlab("fitted values") + ylab("deviance residuals")
p2 <- qplot(.fitted, .pearson, data = vegetablesDiag,
          geom = "hex") + binScale + geom_smooth(size = 1) +
  xlab("fitted values") + ylab("Pearson residuals")
p3 <- qplot(.fitted, sqrt(abs(.pearson)), data = vegetablesDiag,
          geom = "hex") + binScale + geom_smooth(size = 1) +
  xlab("fitted values") +
  ylab("$\\sqrt{|\\text{Pearson residuals}|}$")
grid.arrange(p1, p2, p3, ncol = 3)
```

Figure 4.9 shows two things. First, there is is a clear overdispersion in the model corresponding to the Poisson distribution (the residuals are too large). Second, the linear relation between mean and variance that the Poisson model dictates does not seem to be appropriate.

The overdispersion can be handled by using the quasi Poisson family. The quasi Poisson family does not correspond to a real dispersion model – there is no model on the integers where the variance equals the mean times a constant besides the Poisson model, where the constant is 1. The quasi family is just a way to specify a mean-variance relation that allows for a dispersion parameter. We illustrate the effect of this by reconsidering the table of estimated parameters but using the quasi family.

```
vegetablesGlm2 <- glm(form,
                      family = quasipoisson,
                      data = vegetables)
```

Table 4.5: Summary table of parameter estimates, standard errors and t-tests for the quasi poisson model of sale with 4 predictors.

	Estimate	Std. Error	t value	Pr(>|t|)
ad1	0.32	0.11	2.89	0.004
discount	−0.08	0.06	−1.29	0.2
discountSEK	0.42	0.21	2.04	0.041

As Table 4.5 shows, the estimates stay the same, but by the introduction of a dispersion parameter in the quasi Poisson model the z-scores and the corresponding p-values change. The `ad` variable remains significant, but the two discount variables are no longer both significant.

For the second issue we can try to use a mean-variance relation more appropriate for the data. From the previous diagnostic plot the variance increases more rapidly than linearly with the mean. We can thus try the Γ-model where the variance is the square of the mean.

At the same time we decide to include the `discount` variable only, and we choose to consider a basis expansion of `discount` using natural cubic splines with 3 internal knots.

```
form <- sale ~ offset(log(normalSale)) + store  + ad +
  ns(discount, knots = c(20, 30, 40), Boundary.knots = c(0, 50)) - 1
vegetablesGlm3 <- glm(form,
                    family = Gamma("log"),
                    data = vegetables)
```

Figure 4.10 shows the diagnostic plots. It is clear from this plot that the Γ-model is a much better fit.

Now that we have developed a model that fits the data reasonably well, it makes more sense to consider formal tests. The decision to include the nonlinear expansion of `discount` can, for instance, be justified by a likelihood ratio test. In the glm jargon the likelihood ratio test statistic is known as the deviance, and it is computed as the difference in deviance between the two (nested) models.

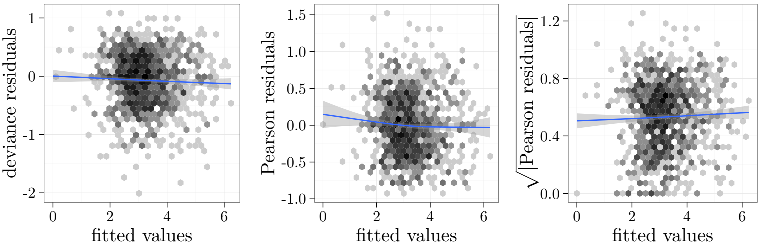


Figure 4.10: Diagnostic plots for the Γ-model. Deviance residuals, Pearson residuals and the square root of the absolute value of the Pearson residuals plotted against fitted values.

```
form <- sale ~ offset(log(normalSale)) + store + ad + discount - 1
vegetablesGlm4 <- glm(form,
                      family = Gamma("log"),
                      data = vegetables)
anova(vegetablesGlm4, vegetablesGlm3, test = "LRT")
```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	687	231.24			
2	684	228.77	3	2.4733	0.02923

Table 4.6: Test of the Γ-model including a non-linear expansion of `discount` against a linear effects model.

To conclude the analysis we will illustrate predictions from the model. Since there are many stores we need to summarize the model somehow. We do this by choosing a few stores representing different coefficients. One of these, store 206, has regression coefficient −0.06, which is the median coefficient, and we refer to this store as the *typical store*. For this store we present predictions and confidence bands corresponding to a normal sale of 1 item.

```
predFrame <- expand.grid(
  normalSale = 1,
  store = factor(c(91, 84, 66, 206, 342, 256, 357)),
  ad = factor(c(0, 1)),
  discount = seq(10, 50, 1)
)
predSale <- predict(vegetablesGlm3,
                    newdata = predFrame,
                    se.fit = TRUE)
predFrame <- cbind(predFrame, as.data.frame(predSale))
p1 <- qplot(discount, exp(fit),
            data = subset(predFrame, store == 206), geom = "line") +
  ylab("sale") +   geom_ribbon(aes(ymin = exp(fit - 2 * se.fit),
                                   ymax = exp(fit + 2 * se.fit)),
              alpha = 0.3) + facet_grid(. ~ ad, label = label_both) +
  coord_cartesian(ylim = c(0, 10)) +
  scale_y_continuous(breaks = c(1, 3, 5, 7, 9))
p2 <- qplot(discount, fit, data = predFrame,
```

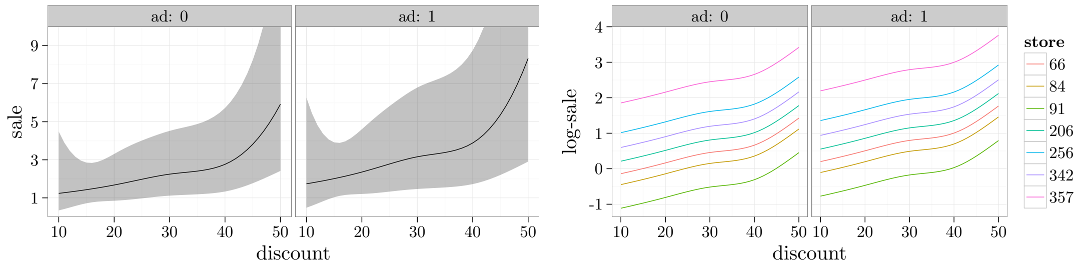


Figure 4.11: Predictions and confidence bands for the sale per normal sale for one typical store (206), and predictions for 7 stores spanning the range of the different stores.

```
             geom = "line", color = store) + ylab("sale") +
  facet_grid(. ~ ad, label = label_both) +
  scale_y_continuous("log-sale")
grid.arrange(p1, p2, ncol = 2)
```

As Figure 4.11 shows, the model predicts that the sale will increase by a factor larger than 1 (larger than 0 on a log-scale) for many stores. The factor is increased if the campaign includes advertising. The model predicts a non-linear relation between the discount (in percent) and the increase in sale. However, the confidence bands (for the median store) show that the predictions are estimated with a large uncertainty.

Estimation theory

In this section we cover the theory behind estimation in generalized linear models. We introduce maximum likelihood estimation for generalized linear models based on the exponential dispersion distributions. This includes the derivation of the nonlinear estimating equation, known as the score equation, and the iterative weighted least squares (IWLS) algorithm that is used in practice to fit the models to data. Then we give a result on the existence and uniqueness of the solution to the score equation for the canonical link.

Maximum likelihood estimation

We first consider the simple case where $Y \sim \mathcal{E}(\theta(\eta), \nu_\psi)$, that is, the distribution of Y is given by the exponential dispersion model with canonical parameter $\theta(\eta)$ and structure measure ν_ψ. Derivations of the score equation and Fisher information in this case can then be used to derive the score equation and Fisher information in the general case when we have observations $Y_1, \ldots, Y_n$ that are conditionally independent given the predictors with $Y_i \mid \mathbf{X} \sim \mathcal{E}(\theta(\eta_i), \nu_\psi)$ and $\eta_i = X_i^T \beta$.

Definition 4.21. *The score statistic is the gradient of the log-likelihood function,*

$$U(\eta) := \nabla_\eta \ell(\eta).$$

The Fisher information,

$$\mathcal{J}(\eta) = -E_\eta D_\eta U(\eta),$$

is minus the expectation of the derivative of the score statistic, or, equivalently, the expectation of the second derivative of the negative log-likelihood.

The score equation is obtained by equating the score statistic equal to 0. In all that follows, the dispersion parameter ψ is regarded as fixed.

Lemma 4.22. *If $Y \sim \mathcal{E}(\theta(\eta), \nu_\psi)$ then the log-likelihood function is*

$$\ell_Y(\eta) = \frac{\theta(\eta)Y - c(\eta)}{\psi},$$

the score function is $U(\eta) = \theta'(\eta)(Y - \mu(\eta))/\psi$, and the Fisher information is

$$\mathcal{J}(\eta) = \frac{\theta'(\eta)\mu'(\eta)}{\psi}.$$

Proof. The density for the distribution of Y w.r.t. ν_ψ is by definition

$$e^{\frac{\theta(\eta)y - c(\eta)}{\psi}},$$

and it follows that $\ell_Y(\eta)$ has the stated form. Differentiation of $\psi \ell_Y(\eta)$ yields

$$\psi U(\eta) = \psi \ell'(\eta) = \theta'(\eta)Y - c'(\eta) = \theta'(\eta)\Big(Y - \underbrace{\frac{c'(\eta)}{\theta'(\eta)}}_{\mu(\eta)}\Big),$$

where we have used Corollary 4.11. Furthermore, we find that

$$\psi U'(\eta) = \theta''(\eta)(Y - \mu(\eta)) - \theta'(\eta)\mu'(\eta),$$

and since $E_\eta Y = \mu(\eta)$ it follows that

$$\mathcal{J}(\eta) = -\frac{E_\eta U'(\eta)}{\psi} = \frac{\theta'(\eta)\mu'(\eta)}{\psi}.$$

□

With only a single observation, the score equation is equivalent to $\mu(\eta) = Y$, and it follows that there is a solution to the score equation if $Y \in J = \mu(I)$. However, the situation with a single observation is not relevant for practical purposes. The result is only given as an intermediate step towards the next result.

The score function U above is a function of the univariate parameter η. We adapt in the following the convention that for a vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$

$$U(\boldsymbol{\eta}) = (U(\eta_1), \ldots, U(\eta_n))^T.$$

That is, the score function is applied coordinatewisely to the vector $\boldsymbol{\eta}$. Note that the derivative (the Jacobian) of $\boldsymbol{\eta} \mapsto U(\boldsymbol{\eta})$ is an $n \times n$ diagonal matrix.

$$\partial_{\eta_i} U(\boldsymbol{\eta})_j = \begin{cases} U'(\eta_j) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Theorem 4.23. *Assume that $Y_1, \ldots, Y_n$ are conditionally independent given $\mathbf{X}$ and that $Y_i \mid \mathbf{X} \sim \mathcal{E}(\theta(\eta_i), \nu_\psi)$ where $\eta_i = X_i^T\beta$. Then with $\boldsymbol{\eta} = \mathbf{X}\beta$ the score function expressed in the β-parameter is*

$$\mathcal{U}(\beta) = \mathbf{X}^T U(\boldsymbol{\eta}).$$

The Fisher information is

$$\mathcal{J}(\beta) = \mathbf{X}^T\mathbf{W}\mathbf{X},$$

with the entries in the diagonal weight matrix $\mathbf{W}$ being

$$w_{ii} = \frac{(\mu_i')^2}{\psi\mathcal{V}(\mu_i)} = \frac{\theta'(\eta_i)\mu_i'}{\psi}$$

where $\mu_i = \mu(\eta_i)$ and $\mu_i' = \mu'(\eta_i)$.

The diagonal weight matrix $\mathbf{W}$ is

$$\frac{1}{\psi}\begin{pmatrix} \frac{(\mu_1')^2}{\mathcal{V}(\mu_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{(\mu_n')^2}{\mathcal{V}(\mu_n)} \end{pmatrix}$$

Proof. By the independence assumption the log-likelihood is

$$\ell_{\mathbf{Y}}(\beta) = \sum_{i=1}^{n} \ell_{Y_i}(\eta_i)$$

where $\eta_i = X_i^T \beta$. By the chain rule,

$$\mathcal{U}(\beta) = \nabla_\beta \ell_{\mathbf{Y}}(\beta) = \sum_{i=1}^{n} X_i U(\eta_i) = \mathbf{X}^T U(\boldsymbol{\eta}).$$

As argued above, $-D_{\boldsymbol{\eta}} U(\boldsymbol{\eta})$ is diagonal, and the expectation of the diagonal entries are according to Lemma 4.22

$$w_{ii} = \frac{\theta'(\eta_i)\mu'(\eta_i)}{\psi}.$$

The alternative formula for the weights follow from Corollary 4.11. Thus

$$\mathcal{J}(\beta) = -E_\beta D_\beta \mathcal{U}(\beta) = -E_\beta \mathbf{X}^T D_{\boldsymbol{\eta}} U(\boldsymbol{\eta}) \mathbf{X} = \mathbf{X}^T \mathbf{W} \mathbf{X}.$$

□

By observing that

$$U(\boldsymbol{\eta})_i = \frac{\theta'(\eta_i)(Y_i - \mu_i)}{\psi},$$

it follows that the score equation $\mathcal{U}(\beta) = 0$ is equivalent to the system of equations

$$\sum_{i=1}^{n} \theta'(\eta_i)(Y_i - \mu_i) X_{ij} = 0$$

for $j = 1, \ldots, p$. Note that the score equation and thus its solution does not depend upon the dispersion parameter. Note also, that for the canonical link function the equations simplify because $\theta'(\eta_i) = 1$, and the weights also simplify to

$$w_{ii} = \frac{\mu_i'}{\psi} = \frac{\mathcal{V}(\mu_i)}{\psi}.$$

Whether there is a solution to the score equation, and whether it is unique, has a complete solution for the canonical link. For arbitrary link functions the situation is less clear, and we must be prepared for the existence of multiple solutions or no solutions in practice.

Example 4.24. For the normal distribution $\mathcal{N}(\mu, \sigma^2)$ and with the canonical link function the log-likelihood function becomes

$$\begin{aligned} \ell(\beta) &= \frac{1}{\psi} \sum_{i=1}^{n} Y_i X_i^T \beta - \frac{(X_i^T \beta)^2}{2} \\ &= \frac{1}{2\psi} \left(2\mathbf{Y}^T \mathbf{X} \beta - \beta^T \mathbf{X}^T \mathbf{X} \beta \right) \\ &= \frac{1}{2\psi} \left(||\mathbf{Y}||^2 - ||\mathbf{Y} - \mathbf{X}\beta||^2 \right). \end{aligned}$$

Up to the term $||\mathbf{Y}||^2$ – that doesn't depend upon the unknown β-vector – the log-likelihood function is proportional to the squared error loss with proportionality constant $-1/(2\psi)$. The maximum likelihood estimator is thus equal to the least squares estimator. ○

Algorithms

The general, nonlinear score equation does not have a closed form solution and must be solved by iterative methods. Newton's algorithm is based on a first order Taylor approximation of the score function. The resulting approximation of the score equation is a linear equation. Newton's algorithm consists of iteratively computing the first order Taylor approximation and solving the resulting linear approximation. The preferred algorithm for estimation of generalized linear models is a slight modification where the derivative of the score is replaced by its expectation, that is, by the Fisher information. To present the idea we consider a simple example of estimation in the exponential distribution with i.i.d. observations.

If Z is Weibull distributed with shape parameter k and scale parameter η then $Y = Z^k$ is exponentially distributed with scale parameter η^k. This explains the interest in this particular parametrization as it allows us to fit models with Weibull distributed responses.

Example 4.25. Consider the parametrization $\theta(\eta) = -\eta^{-k}$ for $\eta > 0$ (and a fixed $k > 0$) of the canonical parameter in the exponential distribution. That is, the density is

$$e^{\theta(\eta)y - k\log\eta}$$

w.r.t. the Lebesgue measure on $(0, \infty)$. The mean value function is

$$\mu(\eta) = -\frac{1}{\theta(\eta)} = \eta^k.$$

With $Y_1, \ldots, Y_n$ i.i.d. observations from this distribution and with

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

the score function amounts to

$$U(\eta) = \sum_{i=1}^{n} \theta'(\eta)(Y_i - \mu(\eta)) = nk\eta^{-1}(\eta^{-k}\bar{Y} - 1)$$

where we have used that $\theta'(\eta) = k\eta^{-k-1}$. The score equation is thus equivalent to

$$\eta^{-k}\bar{Y} - 1 = 0.$$

It is straight forward to solve this equation analytically[6], and the solution is

$$\eta = \bar{Y}^{1/k}.$$

However, to illustrate the general techniques we Taylor expand the left hand side of the equation around η_m to first order and obtain the linear equation

$$\eta_m^{-k}\bar{Y} - 1 - k\eta_m^{-k-1}\bar{Y}(\eta - \eta_m) = 0.$$

The solution of this linear equation is

$$\eta_{m+1} = \eta_m - \frac{\eta_m^{-k}\bar{Y} - 1}{-k\eta_m^{-k-1}\bar{Y}} = \eta_m - \frac{U(\eta_m)}{U'(\eta_m)}$$

provided that $U'(\eta_m) \neq 0$. This is Newton's algorithm. With a suitable choice of starting value η_1 we iteratively update η_m until convergence.

If we replace the derivative of the score function in the approximating linear equation with its expectation we arrive at the linear equation

$$\eta_m^{-k}\bar{Y} - 1 - k\eta_m^{-1}(\eta - \eta_m) = 0,$$

whose solution is

$$\eta_{m+1} = \eta_m - \frac{\eta_m^{-k}\bar{Y} - 1}{-k\eta_m^{-1}} = \eta_m + \frac{U(\eta_m)}{\mathcal{J}(\eta_m)}.$$

The general technique of replacing the derivative of the score function with its expectation in Newton's algorithm is known as Fisher scoring. ○

The iterative weighted least squares algorithm in the general case is no more difficult to formulate than for the one-dimensional example above. First note that the dispersion parameter enters as a multiplicative constant in the log-likelihood, and its

[6] Which shows that if $Z_1, \ldots, Z_n$ are i.i.d. Weibull distributed with known shape parameter k and scale parameter η the MLE of η is

$$\hat{\eta} = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i^k\right)^{1/k}.$$

value does not affect the maximum likelihood estimate of β. We take it to be equal to 1 for the subsequent computations. The derivative of minus the score function[7] is found as in the proof of Theorem 4.23 to be

$$-D_\beta \mathcal{U}(\beta) = \mathbf{X}^T \mathbf{W}^{\text{obs}} \mathbf{X}$$

where

$$U'(\eta_i) = \theta''(\eta_i)(Y_i - \mu(\eta_i)) \\ -\theta'(\eta_i)\mu'(\eta_i).$$

$$\mathbf{W}^{\text{obs}} = -\begin{pmatrix} U_1'(\eta_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & U_n'(\eta_n) \end{pmatrix}.$$

A first order Taylor expansion of the score function around β_m results in the following linearization of the score equation

$$\mathbf{X}^T U(\boldsymbol{\eta}_m) - \mathbf{X}^T \mathbf{W}_m^{\text{obs}} \mathbf{X}(\beta - \beta_m) = 0.$$

Note that $\mathbf{W}_m^{\text{obs}}$ as well as $\mathbf{W}_m$ depend upon the current β_m through $\boldsymbol{\eta}_m = \mathbf{X}\beta_m$, hence the subscript m.

Its solution is

$$\beta_{m+1} = \beta_m + (\mathbf{X}^T \mathbf{W}_m^{\text{obs}} \mathbf{X})^{-1} \mathbf{X}^T U(\boldsymbol{\eta}_m),$$

provided that $\mathbf{X}^T \mathbf{W}_m^{\text{obs}} \mathbf{X}$ has full rank p. Replacing $\mathbf{W}_m^{\text{obs}}$ by $\mathbf{W}_m$ from Theorem 4.23 we get the Fisher scoring algorithm. We may note that the diagonal entries in $\mathbf{W}_m$ are always strictly positive if the mean value map is strictly monotone, which implies that $\mathbf{X}^T \mathbf{W}_m \mathbf{X}$ is positive definite and has rank p if and only if $\mathbf{X}$ has rank p. By contrast, the diagonal weights in $\mathbf{W}_m^{\text{obs}}$ may be negative.

We can rewrite the update formula for the Fisher scoring algorithm as follows

$$\begin{aligned} \beta_{m+1} &= \beta_m + (\mathbf{X}^T \mathbf{W}_m \mathbf{X})^{-1} \mathbf{X}^T U(\boldsymbol{\eta}_m)^T \\ &= (\mathbf{X}^T \mathbf{W}_m \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_m \Big(\underbrace{\mathbf{X}\beta_m + \mathbf{W}_m^{-1} U(\boldsymbol{\eta}_m)^T}_{\mathbf{Z}_m} \Big). \end{aligned}$$

The vector $\mathbf{Z}_m$ is known as the *working response*, and its coordinates can be written out as

$$Z_{m,i} = X_i^T \beta_m + \frac{Y_i - \mu_{m,i}}{\mu'_{m,i}}. \tag{4.6}$$

In terms of the working response, the vector β_{m+1} is the minimizer of the weighted squared error loss

$$(\mathbf{Z}_m - \mathbf{X}\beta)^T \mathbf{W}_m (\mathbf{Z}_m - \mathbf{X}\beta), \tag{4.7}$$

[7] Often called the observed Fisher information.

see Theorem 2.1. The Fisher scoring algorithm for generalized linear models is known as iterative weighted least squares (IWLS), since it can be understood as iteratively solving a weighted least squares problem. To implement the algorithm we can also rely on general solvers of weighted least squares problems. This results in the following version of IWLS. Given β_1 we iterate over the steps 1–3 until convergence:

1. Compute the working response vector $\mathbf{Z}_m$ based on β_m using (4.6).

2. Compute the weights

$$w_{m,ii} = \frac{(\mu'_{m,i})^2}{\mathcal{V}(\mu_{m,i})}.$$

3. Compute β_{m+1} by minimizing the weighted sum of squares (4.7).

The dispersion parameter is eliminated (by taking $\psi = 1$) in the IWLS algorithm. It doesn't mean that the dispersion parameter is irrelevant. It matters for the subsequent statistical analysis, but not for the estimation.

It is noteworthy that the computations only rely on the mean value map μ, its derivative μ' and the variance function $\mathcal{V}$. Thus the IWLS algorithm depends on the mean and variance structure, as specified in the assumptions GA1 and GA2, and not on any other aspects of the exponential dispersion model.

Existence and uniqueness

This section deals with investigating existence and uniqueness of the solution to the score equation when the link function is the canonical link. This rests on some special mathematical structure for the canonical link that we don't have in the general case. It is of practical relevance to understand if the estimator we try to compute using the IWLS algorithm exists and is unique. Convergence of the algorithm is then assured to be towards the unique solution.

We have previously seen that for the canonical link the weights and thus the IWLS algorithm simplify. We may further observe that for the canonical link function $\theta''(\eta) = 0$, and the observed Fisher information coincides with the Fisher information. That is, for the canonical link function

$$\mathbf{W}^{\text{obs}} = \mathbf{W},$$

and Newton's algorithm coincides with the IWLS algorithm. Moreover, if we introduce

$$\tau(\beta) = \sum_{i=1}^{n} \mu(X_i^T \beta) X_i \quad \text{and} \quad t = \sum_{i=1}^{n} Y_i X_i,$$

then the score equation is equivalent to the equation

$$\tau(\beta) = t. \tag{4.8}$$

The main result in what follows is a complete characterization of existence and uniqueness of the solution to the score equation.

Define the convex, open set

$$D = \{\beta \in \mathbb{R}^p \mid \mathbf{X}\beta \in I^n\}$$

of parameters for which the linear predictors are in the open interval I. The set depends upon $\mathbf{X}$ and the map τ is defined on D. We only search for solutions in D. Observe that by the general regularity assumptions $\mu'(\eta) = \mathcal{V}(\eta) > 0$ for $\eta \in I$. This implies that the diagonal entries in $\mathbf{W}$ are strictly positive for $\beta \in D$, and thus that the Fisher information

$$\mathcal{J}(\beta) = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

is positive definite for $\beta \in D$ if and only if $\mathbf{X}$ has rank p.

Theorem 4.26. *If $\mathbf{X}$ has full rank p the map $\tau : D \to \mathbb{R}^p$ is one-to-one. In this case there is a unique solution to (4.8) if and only if $t \in C := \tau(D)$.*

Proof. The result is proved if we establish that τ is one-to-one. To reach a contradiction, assume that $\tau(\beta) = \tau(\beta')$ with $r = \beta' - \beta \neq 0$. Then consider the function

$$h(\alpha) = r^T \tau(\beta + \alpha r)$$

with the property that $h(0) = h(1)$. We find that h is continuously differentiable with

$$h'(\alpha) = r^T \mathcal{J}(\beta + \alpha r) r.$$

Since $h(0) = h(1)$ there is an $\alpha' \in (0, 1)$ where h attains a local optimum and thus $h'(\alpha') = 0$. This implies that $\mathcal{J}(\beta + \alpha' r)$ is not positive definite (only positive semidefinite), which contradicts the full rank p assumption on $\mathbf{X}$. □

Recall that $J = \mu(I)$. *Disclaimer: In the following proof we assume that $D = \mathbb{R}^p$, which is not needed for the result to hold*

Lemma 4.27. *If $t_0 = \sum_{i=1}^n \mu_i X_i$ with $\mu_i \in J$ then $t_0 \in C$.*

Proof. Take $\nu = \nu_1$, that is, ν is the structure measure corresponding to the dispersion parameter $\psi = 1$. If $\mu \in J$ there is a θ such that

$$\begin{aligned} 0 &= \int (y-\mu)e^{\theta y}\nu(\mathrm{d}y) \\ &= \int_{\{y\leq\mu\}} (y-\mu)e^{\theta y}\nu(\mathrm{d}y) + \int_{\{y>\mu\}} (y-\mu)e^{\theta y}\nu(\mathrm{d}y). \end{aligned}$$

If the latter integral is 0 the former integral must be 0 too, which implies that ν is degenerate at μ. This contradicts the assumption that the mean value map μ is strictly increasing (that is, that $\mathcal{V}(\eta) = \mu'(\eta) > 0$). Thus the latter integral is non-zero, and the important conclusion is that $\nu(\{y \mid y-\mu > 0\}) > 0$. Likewise, $\nu(\{y \mid \mu - y > 0\}) > 0$.

With

$$L(\beta) = e^{\sum_{i=1}^n \mu_i X_i^T\beta - \kappa(X_i^T\beta)} = \prod_{i=1}^n e^{\mu_i X_i^T\beta - \kappa(X_i^T\beta)}$$

we see that $D_\beta \log L(\beta) = 0$ is equivalent to the equation $\tau(\beta) = t_0$. Thus if we can show that the function L attains a maximum we are done. To this end fix a unit vector $e \in \mathbb{R}^p$. By definition

$$e^{\kappa(\lambda X_i^T e)} = \int e^{\lambda y_i X_i^T e}\nu(\mathrm{d}y_i),$$

and if we plug this into the definition of L we get that

$$\begin{aligned} L(\lambda e)^{-1} &= \prod_{i=1}^n e^{-\lambda\mu_i X_i^T e}\int e^{\lambda y_i X_i^T e}\nu(\mathrm{d}y_i) \\ &= \int e^{\lambda\left(\sum_{i=1}^n (y_i-\mu_i)X_i^T e\right)}\nu^{\otimes n}(\mathrm{d}y) \end{aligned}$$

for $\lambda > 0$. With $A_+ = \{(y_1,\ldots,y_n) \mid (y_i-\mu_i)\mathrm{sign}(X_i^T e) > 0\}$ it follows from the previous considerations that $\nu^{\otimes n}(A_+) > 0$ and by monotone convergence that

$$L(\lambda e)^{-1} \geq \int_{A_+} e^{\lambda\left(\sum_{i=1}^n (y_i-\mu_i)X_i^T e\right)}\nu^{\otimes n}(\mathrm{d}y) \to \infty \cdot \nu^{\otimes n}(A_+) = \infty$$

for $\lambda \to \infty$.

If 0 is a maximizer we are done so we assume it is not. Then there is a sequence β_n such that

$$L(0) \leq L(\beta_n) \nearrow \sup_{\beta \in D} L(\beta)$$

and such that $\lambda_n := ||\beta_n|| > 0$ for all n. Define the unit vectors $e_n = \beta_n / \lambda_n$, then since the unit sphere is compact this sequence has a convergent subsequence. By taking a subsequence we can thus assume that $e_n \to e$ for $n \to \infty$. By taking a further subsequence we can assume that either λ_n is convergent or $\lambda_n \to \infty$. In the former case we conclude that $\beta_n = \lambda_n e_n$ is convergent, and by continuity of L the limit is a maximizer.

To reach a contradiction, assume therefore that $\lambda_n \to \infty$. Choose λ_e according to previous derivations such that $L(\lambda_e e) < L(0)$ then $\lambda_n > \lambda_e$ from a certain point, and by log-concavity of L it holds that

$$L(\lambda_e e_n) \geq L(0)$$

from this point onward. Since the left hand side converges to $L(\lambda_e e)$ we reach a contradiction. We conclude that L always attains a maximum, that this maximum is a solution to the equation $\tau(\beta) = t_0$, and thus that $t_0 \in \mathsf{C}$. □

Corollary 4.28. *The set $\mathsf{C} = \tau(D)$ has the representation*

$$\mathsf{C} = \left\{ \sum_{i=1}^{n} \mu_i X_i \mid \mu_i \in J \right\} \tag{4.9}$$

and is convex. If $\mathbf{X}$ has full rank p then C is open.

Proof. Lemma 4.27 shows that C has the claimed representation. Recall also that $J = \mu(I)$ is an open interval. This is because the function μ is continuous, which implies that J is an interval, and strictly increasing, which implies that J is also open. Since J is an interval, C is convex. The full rank condition ensures that $\mathbf{X}^T$ maps $\mathbb{R}^n$ onto $\mathbb{R}^p$ as a linear map, which implies that $\mathbf{X}^T$ is an open map (it maps open sets onto open sets). In particular, we have that

$$\mathsf{C} = \mathbf{X}^T(J \times \ldots \times J)$$

is open. □

To prove that there is a unique solution to the score equation amounts to proving that $t \in C$. This is, by Corollary 4.28, clearly the case if

$$P_\theta(Y \in J) = 1,$$

but less trivial to check if $P_\theta(Y \in \partial J) > 0$.

Note that the solution, if it exists, is unique if $\mathbf{X}$ has full rank p. Note also that $Y \in \overline{J}$ – the observations are in the closure of J. The following is a useful observation. Suppose that $\mathbf{X}$ has full rank p such that C is open and assume that $t \in C$. Consider one additional observation (Y_{n+1}, X_{n+1}) and let C' denote the C-set corresponding to the enlarged data set. Then $t + Y_{n+1}X_{n+1} \in C'$. This is obvious if $Y_{n+1} \in J$. Assume that Y_{n+1} is the left end point of J, then for sufficiently small $\delta > 0$

$$t + Y_{n+1}X_{n+1} = t - \delta X_{n+1} + \underbrace{(Y_{n+1} + \delta)}_{\in J} X_{n+1},$$

and $t - \delta X_{n+1} \to t \in C$ for $\delta \to 0$. Since C is open, $t - \delta X_{n+1} \in C$ for sufficiently small δ. A similar argument applies if Y_{n+1} is the right end point. In conclusion, if $\mathbf{X}$ has full rank p and $t \in C$ such that the score equation has a unique solution, there will still be a unique solution if we add more observations. Figuratively speaking, we cannot loose the existence and uniqueness of the solution to the score equation once it is obtained.

Example 4.29. This example illustrates the general results for the Poisson model where $J = (0, \infty)$. The existence of the MLE for the Poisson likelihood is determined by how and if the data set contains zero counts ($Y_i = 0$). If there are no zero counts, the MLE always exists.

If there are zero counts, the existence of the MLE is characterized as follows. The n predictors $\tilde{X}_1, \ldots, \tilde{X}_n$ span a convex cone in $\mathbb{R}^p$ with vertex in 0. The set C is the interior of this cone. The MLE exists if and only if

$$\tilde{t} = \sum_{i=1}^n Y_i \tilde{X}_i$$

is in the interior of the cone. If there is an intercept in the model (which there usually is) and $\tilde{X}_n = (1, X_n^T)^T \in \mathbb{R}^{p+1}$ for $X_n \in \mathbb{R}^p$

then $\tilde{t}$ is in the interior of the cone if and only if

$$t = \frac{1}{\sum_{i=1}^n Y_i} \sum_{i=1}^n Y_i X_i$$

is in the interior of the convex hull of $X_1, \ldots, X_n$ (the smallest convex set spanned by the predictors). This condition is easy to visualize for $p = 1, 2$.

We generate a small toy example with $p = 2$ and $n = 4$. We construct the predictors and the figure illustrating their convex hull.

```
X <- data.frame(x1 = c(-2, -1, 2, 0), x2 = c(1, -1, 0, 2))
p <- qplot(x1, x2, data = X, geom = "polygon", alpha = I(0.5)) +
  geom_point(size = 5, alpha = 1) + xlab("") + ylab("")
```

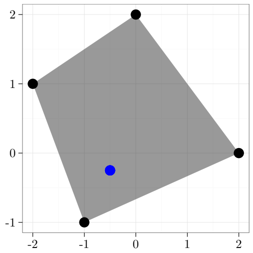


Figure 4.12: The convex hull of the predictors with the sufficient statistic in the interior.

We consider first an example where the average t ends up in the interior of the convex hull, and we fit the Poisson model using `glm`.

```
y <- c(1, 2, 1, 0)
Xy <- cbind(y, X)
t <- c(y %*% X$x1, y %*% X$x2)/sum(y)
p + geom_point(aes(t[1], t[2]), size = 5, color = "blue")
```

```
summary(glm(y ~ x1 + x2, family = poisson, data = Xy))

##
## Call:
## glm(formula = y ~ x1 + x2, family = poisson, data = Xy)
##
## Deviance Residuals:
##       1        2        3        4
##  0.3582  -0.1618   0.2128  -0.7445
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.02674    0.54607   0.049    0.961
## x1          -0.12367    0.37418  -0.331    0.741
## x2          -0.65497    0.51569  -1.270    0.204
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.77259  on 3  degrees of freedom
## Residual deviance: 0.75402  on 1  degrees of freedom
## AIC: 13.368
##
## Number of Fisher Scoring iterations: 5
```

Next we consider an example where the average t ends up on the boundary of the convex hull. We can fit a Poisson model using glm, and we will not get an explicit warning that the MLE does not exist.

```
y <- c(1, 2, 0, 0)
Xy <- cbind(y, X)
t <- c(y %*% X$x1, y %*% X$x2)/sum(y)
summary(glm(y ~ x1 + x2, family = poisson, data = Xy))

##
## Call:
## glm(formula = y ~ x1 + x2, family = poisson, data = Xy)
##
## Deviance Residuals:
##          1           2           3           4
##  0.000e+00   0.000e+00  -3.519e-07  -2.015e-05
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -12.836  25534.580  -0.001        1
## x1             -8.789  17023.053  -0.001        1
## x2             -4.741   8511.527  -0.001        1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4.4987e+00  on 3  degrees of freedom
## Residual deviance: 4.0610e-10  on 1  degrees of freedom
## AIC: 10.614
##
## Number of Fisher Scoring iterations: 21

p + geom_point(aes(t[1], t[2]), size = 5, color = "blue")
```

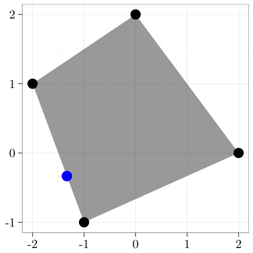


Figure 4.13: The convex hull of the predictors with the sufficient statistic on the boundary.

There are several signs in the summary output that the MLE does not exist. First, the residual deviance is almost driven to 0, which is a warning sign. Second, the estimated standard errors are huge. Moreover, the 21 iterations is actually a lot. When the MLE exists the IWLS algorithm will converge in very few iterations. ○

Sampling distributions

It is not possible to derive exact distributional results about the estimators or the various test statistics that are used for generalized linear models – except for the special case of the linear model with normal errors. We will need to rely on approximations. By introducing an idealized least squares estimator and reinterpreting the IWLS algorithm in this context, it is possible to introduce sensible *approximate* moment results under the weak GA1 and GA2 assumptions. Approximate confidence intervals are subsequently introduced based on Z-scores, and tests of linear hypotheses on β under the GA3 assumption are discussed using likelihood ratio (deviance) tests. The operational procedures for using the tests is the same as for the linear model – the only difference being that the distributions used to compute p-values are approximations. The formal asymptotic justifications will only be treated briefly, see also page 135.

An idealized least squares estimator

We introduce a weighted least squares estimator with weights and responses depending on the unknown parameters for which we can derive exact moment results. We can then interpret the IWLS algorithm as computing an approximate solution to this idealized least squares problem. The derivation emphasizes the fact that the algorithm only depends on the assumptions GA1 and GA2.

Supposing that we know the true parameter, β, we can form the weighted quadratic loss

$$\ell(\tilde{\beta}) = \sum_{i=1}^{n} \mathcal{V}(\mu_i)^{-1}(Y_i - \mu(X_i^T\tilde{\beta}))^2$$

where $\mu_i = \mu(X_i^T\beta)$. Besides the fact that the weights depend upon the unknown true parameter, the $\tilde{\beta}$ enters nonlinearly through $\mu(X_i^T\tilde{\beta})$. We can Taylor expand the mean once[8] around β to obtain

$$\mu(X_i\tilde{\beta}) \simeq \mu_i + \mu_i' X_i^T(\tilde{\beta} - \beta)$$

and plug this approximation back into the loss. This yields the approximating quadratic loss

$$\begin{aligned} \tilde{\ell}(\tilde{\beta}) &= \sum_{i=1}^{n} \mathcal{V}(\mu_i)^{-1}\Big(Y_i - \mu_i - \mu_i' X_i^T(\tilde{\beta} - \beta)\Big)^2 \\ &= \sum_{i=1}^{n} \frac{(\mu_i')^2}{\mathcal{V}(\mu_i)} \left(X_i^T\beta + \frac{Y_i - \mu_i}{\mu_i'} - X_i^T\tilde{\beta}\right)^2. \end{aligned}$$

By introducing the *idealized* responses

$$Z_i = X_i^T\beta + \frac{Y_i - \mu_i}{\mu_i'} \tag{4.10}$$

[8] This is the key step in the Gauss-Newton algorithm for nonlinear least squares estimation.

and the idealized weights

$$w_{ii} = \frac{(\mu_i')^2}{\mathcal{V}(\mu_i)},$$

the approximating quadratic loss can be written as

$$\tilde{\ell}(\tilde{\beta}) = (\mathbf{Z} - \mathbf{X}\tilde{\beta})^T \mathbf{W}(\mathbf{Z} - \mathbf{X}\tilde{\beta}) = ||\mathbf{Z} - \mathbf{X}\tilde{\beta}||^2_{\mathbf{W}},$$

with **W** the diagonal weight matrix with the w_{ii}'s in the diagonal. Under the assumptions GA1 and GA2 we can observe that

$$E(Z_i \mid X_i) = X_i^T \beta$$

and

$$V(Z_i \mid X_i) = \psi w_{ii}^{-1},$$

but since the weights as well as $\mathbf{Z}$ depend upon the unknown β, we cannot compute $\tilde{\ell}(\tilde{\beta})$. The IWLS algorithm can be understood as iteratively approximating the idealized response by the working response – by plugging in the current estimate of β. In this way we get approximations of the idealized loss, and the idealized weighted least squares estimator

$$\hat{\beta}^{\text{ideal}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Z}.$$

It is worth observing that under GA1, GA2 and A4 we have the following exact distributional results on $\hat{\beta}^{\text{ideal}}$

$$\begin{aligned} E(\hat{\beta}^{\text{ideal}} \mid \mathbf{X}) &= \beta, \\ V(\hat{\beta}^{\text{ideal}} \mid \mathbf{X}) &= \psi(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}, \\ E(||\mathbf{Z} - \mathbf{X}\hat{\beta}^{\text{ideal}}||^2_{\mathbf{W}} \mid \mathbf{X}) &= (n-p)\psi. \end{aligned}$$

If we plug the maximum likelihood estimator, $\hat{\beta}$, in for β in the definition of the idealized quadratic loss, and thus replacing the idealized responses by

$$\hat{Z}_i = X_i^T \hat{\beta} + \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i'}$$

we obtain a computable loss[9]

$$||\hat{\mathbf{Z}} - \mathbf{X}\beta||^2_{\hat{\mathbf{W}}}. \tag{4.11}$$

[9] In reality, $\hat{\beta}$ is β_m, the value of β in the m'th iteration of IWLS, when the algorithm is judged to have converged.

The fact that $\hat{\beta}$ is a fixed point for the IWLS algorithm implies that $\hat{\beta}$ is also the minimizer of (4.11). Provided that (4.11) is a good approximation of the idealized quadratic loss, we can expect that the distributional results on the idealized estimator will be good approximations for $\hat{\beta}$ as well. Observe that

$$\mathcal{X}^2 := ||\hat{\mathbf{Z}} - \mathbf{X}\hat{\beta}||^2_{\hat{\mathbf{W}}} = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\mathcal{V}(\hat{\mu}_i)}$$

where $\mathcal{X}^2$ is known as the Pearson or χ^2-statistic. The results above suggest, in particular, the estimator

$$\hat{\psi} = \frac{1}{n-p}\mathcal{X}^2 \tag{4.12}$$

of the dispersion parameter.

Tests and confidence intervals

The distributional results for the idealized weighted least squares estimator suggest the approximation

$$\sqrt{\psi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}_{jj}}$$

of the standard error of $\hat{\beta}_j$. By plugging in the estimate of the dispersion parameter and the estimate of the weights we define the j'th Z-score as

$$Z_j = \frac{\hat{\beta}_j - \beta}{\sqrt{\hat{\psi}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}_{jj}}}.$$

More generally, we can introduce

$$Z_a = \frac{a^T\hat{\beta} - a^T\beta}{\sqrt{\hat{\psi}a^T(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}a}}$$

for $a \in \mathbb{R}^p$.

The Z-score is used exactly as it is used for linear models. First, it is used to test a one-dimensional restriction on the parameter vector – typically a hypothesis of the form $H_0 : \beta_j = 0$ for some index j. Second, it is used to compute confidence intervals for $a^T\beta$ of the form

$$a^T\hat{\beta} \pm z\sqrt{\hat{\psi}a^T(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}a} \tag{4.13}$$

where z is chosen suitably. Often, a confidence interval with 95% nominal coverage is sought, and $z = 1.96$ is chosen as the 97.5% quantile for the normal distribution. The actual coverage depends upon how well the distribution of Z_a is approximated by $\mathcal{N}(0, 1)$. Occasionally, z is chosen as the 97.5% quantile in the t_{n-p}-distribution instead – still aiming for a 95% coverage. The theoretical support for this practice is not strong, but the practical consequence is clear. Since the tail quantiles for the t-distribution are larger than the tail quantiles for the normal distribution, the use of the t-distribution results in wider confidence intervals and more conservative conclusions. The difference is, however, negligible unless $n - p$ is small – less than 20, say.

An alternative to using theoretical approximations is to use bootstrapping to compute approximating quantiles, see page 136. The square of the Z-statistic, Z_a^2, is known as the Wald statistic. It is possible to construct multivariate Wald statistics, but they are used less frequently.

The likelihood ratio tests are alternatives to z-tests and multivariate Wald tests. Their derivation is based on the stronger distributional assumption GA3 that the responses conditionally on the predictors an exponential dispersion distribution. In the framework of generalized linear models the likelihood ratio tests are usually formulated in terms of *deviances* as introduced in Section 4.

For a generalized linear model we have response observations $Y_1, \ldots, Y_n$, and we let $\hat{\mu}_i$ denote the maximum likelihood estimate of the mean of Y_i. That is,

$$\hat{\mu}_i = \mu(X_i^T \hat{\beta})$$

with $\hat{\beta}$ the maximum likelihood estimate of $\beta \in \mathbb{R}^p$. Likewise, if $\hat{\beta}^0$ denotes the maximum likelihood estimate of β under the null hypothesis

$$H_0 : \beta \in L$$

where $L \subseteq \mathbb{R}^p$ is a p_0 dimensional subspace of $\mathbb{R}^p$, we let $\hat{\mu}_i^0$ denote the corresponding maximum likelihood estimate of the mean for the i'th observation under the hypothesis. Recall that all such hypotheses can be rephrased in terms of a $p \times p_0$ matrix C of rank

p_0 such that

$$H_0 : E(Y_i \mid X_i) = \mu(X_i^T C\beta^0).$$

Definition 4.30. *The deviances for the model and the null hypothesis are*

$$D = \sum_{i=1}^n d(Y_i, \hat{\mu}_i) \quad \text{and} \quad D_0 = \sum_{i=1}^n d(Y_i, \hat{\mu}_i^0),$$

respectively. The deviance test statistic is

$$D_0 - D$$

and the F-test statistic is

$$F = \frac{(D_0 - D)/(p - p_0)}{D/(n-p)}.$$

The deviance test statistic is simply the log-likelihood ratio test statistic for the null hypothesis. The F-test statistic is inspired by the F-test for linear models. In both cases, large values of the test statistic are critical, and thus evidence against the null hypothesis. In contrast to the linear model, it is not possible to derive the exact distributions of the deviance test statistic or the F-test statistic under the null hypothesis. To compute p-values we need to rely on approximations or simulations. The general theoretical approximation of the deviance test statistic is

$$D_0 - D \overset{\text{approx}}{\sim} \psi\chi^2_{p-p_0},$$

which is exact for the linear model with normal errors (under the assumptions A3 + A5). A purpose of the F-test is, as for linear models, to (approximately) remove the dependence upon the unknown dispersion parameter. The approximation of the F-test statistic is

$$F \overset{\text{approx}}{\sim} F_{p-p_0, n-p},$$

which is exact for the linear model with normal errors. The approximation is generally good when ψ is small.

A formal justification of the approximations is quite involved, and the typical strategy is to consider asymptotic scenarios with $n \to \infty$ combined with suitable conditions on the predictors X_i. We will explain the basis for these approximations later, see page 127, but we will not consider any formal asymptotic arguments.

Exercises

Exercise 4.1. The point probabilities for any probability measure on $\{0,1\}$ can be written as

$$p^y(1-p)^{1-y}$$

for $y \in \{0,1\}$. Show that this family of probability measures for $p \in (0,1)$ form an exponential family. Identify I and how the canonical parameter depends on p. ○

Exercise 4.2. Show that the binomial distributions on $\{0,1,\ldots,n\}$ with success probabilities $p \in (0,1)$ form an exponential family. ○

Exercise 4.3. Show that if the structure measure $\nu = \delta_y$ is the Dirac-measure in y then for the corresponding exponential family we have $\rho_\theta = \delta_y$ for all $\theta \in \mathbb{R}$. Show next, that if ν is not a one-point measure, then ρ_θ is not a Dirac-measure for $\theta \in I^\circ$ and conclude that its variance is strictly positive. ○

Exercise 4.4. If ρ is a probability measure on $\mathbb{R}$, its cumulant generating function is defined as the function

$$s \mapsto \log \int e^{sy}\, \rho(\mathrm{d}y)$$

for s such that the integral is finite. Let ρ_θ be an exponential family with unit cumulant function κ. Show that

$$s \mapsto \kappa(\theta + s) - \kappa(\theta)$$

is the cumulant generating function for ρ_θ. ○

Exercise 4.5. Let ρ_θ be the exponential family given by the structure measure ν. Show that

$$\frac{\mathrm{d}\nu}{\mathrm{d}\rho_{\theta_0}} = e^{\kappa(\theta_0) - \theta y}$$

for any $\theta_0 \in I$. Then show that ν is uniquely specified by the unit cumulant function κ.

Hint: Use that the cumulant generating function for ρ_{θ_0} is determined by κ, and that a probability measure is uniquely determined by its cumulant generating function if it is defined in an open interval around 0. ○

5 Statistical methodology

This chapter deals more extensively with statistical methodology for quantifying model uncertainty and for model assessment. General methods based on exact and approximate, parametric and nonparametric likelihoods for *interval estimates* are treated in detail. These methods define nested families of intervals with a clear interpretation. Sampling properties, such as coverage probabilities, of likelihood based intervals can be obtained by distributional approximations or bootstrap methods. The chapter is concluded with a treatment of model assesment, model comparison and model averaging techniques for multimodel inference and model aggregation.

Interval estimates

Interval estimates of parameters of interest are important supplements to point estimates as they add information about the uncertainty of the point estimate. How an interval estimate is to be constructed – not to mention the interpretation of the interval – is, however, not completely obvious.

Definition 5.1. *A confidence interval of the parameter of interest γ with coverage $1 - \alpha$ is a data dependent interval $I(\mathbf{Y})$ with the property*

$$P(I(\mathbf{Y}) \ni \gamma) \geq 1 - \alpha.$$

The determining property of a confidence interval – or rather of the data dependent procedure used to construct the confidence interval – is the lower bound on the probability that it contains γ. This is a frequentistic sampling property. The abstract definition of a confidence interval does, for instance, not grade the different points in the interval – the center points are neither more nor less "certain" than close-to-boundary points. It is no problem to construct interval estimates with the right coverage that systematically misses the MLE, say. Several other peculiar and counter reasonable constructions are possible. Say we have a procedure for constructing bounded confidence intervals that miss γ to the left with probability 5% and to the right with probability 5%. Perfectly valid confidence intervals with 95% coverage can be obtained by extending the intervals to be unbounded to the left, say. We might feel that a symmetric extension sounds more reasonable, but there is nothing in the definition of a confidence interval that enforces this. Most useful confidence intervals, like the standard confidence intervals introduced in the previous chapters, do actually inherit a grading of the different points from the fact that they are either likelihood or approximate likelihood intervals. These intervals are part of naturally defined families of nested intervals. Being (approximate) likelihood intervals they have intrinsic value and can be interpreted without reference to sampling properties, though they can also be calibrated to have the sampling property of a 95% confidence interval, say. The concept of likelihood intervals is to us more fundamental and important than that of confidence intervals.

In the next section we treat likelihood intervals at length, but before doing so we will discuss some general confidence interval constructions.

Definition 5.2. *A combinant is function $R(\mathbf{Y}, \gamma)$ of the data as well as the parameter of interest γ. A combinant is a pivot for a class of distributions of* $\mathbf{Y}$ *if the distribution of $R(\mathbf{Y}, \gamma)$ does not depend on the distribution of* $\mathbf{Y}$.

Example 5.3. The canonical example of a pivot is the Z-score in the linear model under assumptions A3 and A5. The parameter of interest is $\gamma = a^T\beta$ and the combinant is

$$Z_a = \frac{\hat{\gamma} - \gamma}{\hat{\sigma}\sqrt{a^T(\mathbf{X}^T\mathbf{X})^{-1}a}}.$$

From Theorem 2.4 its distribution is a t_{n-p}-distribution, which does not depend on the distribution of $\mathbf{Y}$ – as long as A3 and A5 are fulfilled. It is therefore a pivot. Its square,

$$Z_a^2 = \frac{(\hat{\gamma} - \gamma)^2}{\hat{\sigma}^2 a^T(\mathbf{X}^T\mathbf{X})^{-1}a}$$

is likewise a pivot, now with an $F_{1,n-p}$-distribution. ○

It is important to realize that pivotality is a property determined by the combinant R *as well as* the class of allowed distributions of $\mathbf{Y}$. In the example above $\mathbf{Y}$ must have a normal distribution albeit we do not need to know the parameters. If other distributions of $\mathbf{Y}$ are allowed, the Z-score is no longer guarantied to be an exact pivot.

The point about pivots is that they allow us to derive tests and interval estimates with sampling distributions that are independent of the unknown distribution of $\mathbf{Y}$. A prototypical interval estimate constructed from a pivot such as Z_a is the interval[1]

$$\{\gamma \mid R(\mathbf{Y}, \gamma) \in (w, z)\}.$$

When $R(\mathbf{Y}, \gamma)$ is a pivot we can in principle choose $w = w_\alpha$ and $z = z_\alpha$ to be the α- and $(1-\alpha)$-quantiles of its distribution, in which case the interval is a confidence interval with coverage $1 - 2\alpha$. For combinants like Z_a^2 we would rather compute the interval

$$\{\gamma \mid R(\mathbf{Y}, \gamma) < c\},$$

this time choosing $c = c_\alpha$ as the $(1-\alpha)$-quantile of the pivot's distribution, which gives a confidence interval with coverage $1 - \alpha$.

For the Z-score the constructions above lead to the standard confidence intervals introduced in (2.9). Since the t_{n-p} distribution is symmetric it does not matter if we use the construction based on Z_a or Z_a^2 – they give the same result in this case.

Exact pivotality is a rare and fragile property. It is difficult to come up with other examples than those from the linear model with

[1] We use the term “interval” in a liberal sense. The construction will not result in an interval for all combinants.

the strong A3 normal distributional assumption. And if the distributional assumption does not hold, exact pivotality does usually not hold either. The practical use of combinants for confidence interval construction is therefore based on assuming that the combinant is *approximately* a pivot. If γ denotes the univariate parameter of interest, and $\widehat{\mathrm{se}}$ denotes an estimate of its standard deviation[2], the Z-score

$$\frac{\hat{\gamma} - \gamma}{\widehat{\mathrm{se}}}$$

is widely used as an approximate pivot. Its distribution is approximated by either a $\mathcal{N}(0,1)$-distribution, an appropriate t-distribution, a bootstrap distribution as considered in a subsequent section or by other means. For generalized linear models the Z-score based confidence intervals were introduced in (4.13).

The justification of approximate pivotality – not to mention the analytic formulas needed for the computation of $\widehat{\mathrm{se}}$ – are usually based on asymptotic theory. That is, scenarios where $n \to \infty$, and where the dimension p of the parameter space is fixed. In the next section we will focus on likelihood based combinants, in particular parametric and nonparametric profile log-likelihoods. These have intrinsic non-asymptotic interpretations. Quadratic approximations of these combinants, whose quality can be discussed without reference to asymptotic theory, are then introduced.

Bayesian credibility intervals should be mentioned as an alternative way to produce interval estimates. It is quite important to understand that a credibility interval is neither a likelihood interval nor a confidence interval. We can try to measure the quality of a credibility interval by the yardstick of confidence intervals and vice versa, but this is pointless. They are answers to different questions and fundamentally incomparable. That said, a common credibility construction is quite similar to the likelihood intervals we discuss below, and likelihood intervals can be interpreted as credibility intervals with a noninformative prior. The Bayesian calibration consists of choosing the interval so large that it has 95% posterior probability, say, whereas the frequentist calibration consists of ensuring that the sampling probability that the interval covers the parameter of interest is greater than or equal to 95%.

[2] Also known as the standard error of the estimator.

Likelihood methods

Disregarding philosophical differences, the likelihood ratio quantifies the relative evidence in the data for one probability distribution in our statistical model over another. In this section we treat parametric as well as nonparametric methods for likelihood interval constructions. Sampling properties are discussed in the subsequent section on calibration.

Parametric likelihood methods

If ℓ denotes a generic log-likelihood function and β the generic parameter then

$$\ell(\beta_1) - \ell(\beta_2)$$

quantifies the evidence in the data of[3] β_1 over β_2. If the difference of the log-likelihoods is positive, there is more evidence for β_1, and if it is negative, there is more evidence for β_2.

The maximum likelihood estimator (MLE) is the parameter – or corresponding probability distribution among those in our model – for which there is the largest evidence in the data. The maximum likelihood estimator was considered in the context of generalized linear models as a method for point estimation of parameters in the model. The likelihood function does, however, contain more information than what is extracted by the MLE alone. A likelihood purist might claim that the only thing we need is the ability to compute the likelihood function. In practice, it is quite difficult to comprehend and communicate the information contained in a likelihood function of a multivariate[4] parameter. Thus we need ways to summarize the information. The MLE is a partial summary, but it does not contain any information about how peaked the likelihood is around it. We also want to learn about the other distributions in the statistical model, for which the data provides only a little less evidence than for the MLE. For this purpose we introduce general *likelihood regions*.

Definition 5.4. *Assume that $\ell_{max} := \sup_\beta \ell(\beta) < \infty$. The family $\mathcal{L}(c)$ for $c \geq 0$ of likelihood regions for β is defined*[5] *as*

$$\mathcal{L}(c) = \{\beta \mid 2(\ell_{max} - \ell(\beta)) < c\}.$$

[3] To ease notation we do not distinguish between the distribution and the parameter here.

[4] Possibly high-dimensional or even infinite dimensional.

[5] The "2" in the definition implies that the regions are defined in terms of "$-2\log Q$", which is an arbitrary but convenient convention.

The interpretation of $\mathcal{L}(c)$ is that there is more evidence in the data for any model in $\mathcal{L}(c)$ than for any model in its complement. Also, the likelihood regions increase as a function of c: If $c_1 < c_2$ then $\mathcal{L}(c_1) \subseteq \mathcal{L}(c_2)$, and there is more evidence for any model in $\mathcal{L}(c_1)$ than in $\mathcal{L}(c_2) \backslash \mathcal{L}(c_1)$. Finally, if the MLE, $\hat{\beta}$, exists then

$$\hat{\beta} \in \bigcap_{c>0} \mathcal{L}(c),$$

and this likelihood region is a singleton if and only if the MLE is unique.

The general likelihood regions are not of much more use than the likelihood itself, but we will now introduce a version of likelihood regions for univariate parameters of interest, which has more practical value. Assume in the following that $\gamma(\beta) \in \mathbb{R}$ is a univariate parameter of interest, which is defined as a function of the distributions (or parameters) in our statistical model.

Definition 5.5. *The profile log-likelihood for γ is defined as*

$$\ell(\gamma) = \sup_{\beta : \gamma(\beta) = \gamma} \ell(\beta).$$

The family of likelihood intervals[6] *for γ is defined as*

$$\mathcal{I}(c) = \{\gamma \mid 2(\ell_{max} - \ell(\gamma)) < c\}.$$

The family of likelihood intervals is, just as the family of abstract likelihood regions, a nested family of sets. Being subsets of $\mathbb{R}$ they are somewhat more comprehensible and easier to communicate. The obvious question left is how the cutoff c should be chosen? In the subsequent section on calibration we discuss different ways to chose c and their justifications. It is beneficial to separate the discussion of the type of interval to compute from the calibration of the interval to have a particular property[7]. In this way the (approximate) computation and justification of the form of the interval is completely separated from its distributional properties. That said, c is often chosen to be $c = 3.841$, which hinges on $2(\ell_{\max} - \ell(\gamma))$ being asymptotically χ^2-distributed with 1 degrees of freedom under some regularity conditions. An alternative way to choose a sensible cutoff is based on an absolute interpretation of log-likelihood ratios.

The 95% quantile in the χ^2-distribution with 1 degrees of freedom is 3.841.

[6] The sets defined need not be intervals, but the terminology becomes less heavy if we accept this slight abuse of the word “interval”.

[7] Properties as in sampling properties, e.g. to have a certain coverage as a confidence interval.

Suppose we have two (infinite) urns, one with only white balls and one with half black and half white balls. If we draw b white balls without drawing a black the log-likelihood ratio of the "all white" urn over the alternative is $b \log(2) \simeq 0.69b$. If we choose the cutoff $c = 3.841$ this corresponds to 5.5 white balls. Is this strong evidence for the "all white" urn over the alternative? The simple urn example can be used to translate the abstract log-likelihood ratio to an interpretable quantity – the number of white balls drawn. For a thorough treatment of the likelihood principle see the book Statistical Evidence[8] by Richard Royall. The urn example is covered in his Section 1.6.

Likelihood intervals can be computed using numerical methods, and are implemented in R for generalized linear models via the `confint` function. Approximate likelihood intervals can be computed via quadratic approximations of the full log-likelihood. We first give a profiling result for quadratic functions and then discuss its applications.

Lemma 5.6. *Let* $\beta = (\beta_1, \gamma)$ *with* $\beta_1 \in \mathbb{R}^{p-1}$ *and* $\gamma \in \mathbb{R}$. *Let* $\ell(\beta) = (\hat{\beta} - \beta)^T \mathcal{J} (\hat{\beta} - \beta)$ *for a positive definite matrix* $\mathcal{J}$. *Then*

$$\ell(\gamma) = \inf_{\beta_1} \ell((\beta_1, \gamma)) = \frac{(\hat{\gamma} - \gamma)^2}{(\mathcal{J}^{-1})_{pp}}.$$

Proof. Write the matrix $\mathcal{J}$ as a block matrix

$$\mathcal{J} = \left(\begin{array}{cc} \mathcal{J}_{11} & \mathcal{J}_{1p} \\ \mathcal{J}_{1p}^T & \mathcal{J}_{pp} \end{array} \right).$$

Expanding $\ell(\beta)$ in terms of these blocks gives

$$\ell(\beta) = (\hat{\beta}_1 - \beta_1)^T \mathcal{J}_{11} (\hat{\beta}_1 - \beta_1) + (\hat{\gamma} - \gamma)^2 \mathcal{J}_{pp} + 2(\hat{\beta}_1 - \beta_1)^T \mathcal{J}_{1p} (\hat{\gamma} - \gamma).$$

Minimization over β_1 for fixed γ gives

$$\hat{\beta}_1(\gamma) = \hat{\beta}_1 + (\hat{\gamma} - \gamma) \mathcal{J}_{11}^{-1} \mathcal{J}_{1p}.$$

Plugging this into $\ell(\beta)$ gives the profile log-likelihood

$$\ell(\gamma) = (\hat{\gamma} - \gamma)^2 (\mathcal{J}_{pp} - \mathcal{J}_{1p}^T \mathcal{J}_{11}^{-1} \mathcal{J}_{1p}).$$

By the general formula for inversion of block matrices the lower left corner of $\mathcal{J}^{-1}$ is

$$(\mathcal{J}^{-1})_{pp} = (\mathcal{J}_{pp} - \mathcal{J}_{1p}^T \mathcal{J}_{11}^{-1} \mathcal{J}_{1p})^{-1},$$

[8] Richard M. Royall. *Statistical evidence*, Chapman & Hall, London, 1997

and the result follows. □

For a twice continuously differentiable parametric log-likelihood ℓ and an MLE $\hat{\beta}$ in the interior of the parameter space a second order Taylor expansion of ℓ around $\hat{\beta}$ gives

$$\ell(\beta) \simeq \ell_{\max} - \frac{1}{2}(\hat{\beta} - \beta)^T \mathcal{J}^{\text{obs}}(\hat{\beta} - \beta).$$

Here $\mathcal{J}^{\text{obs}} = -D^2\ell(\hat{\beta})$ is the positive (semi)definite observed Fisher information. From this

$$2(\ell_{\max} - \ell(\beta)) \simeq (\hat{\beta} - \beta)^T \mathcal{J}^{\text{obs}}(\hat{\beta} - \beta)$$

and Lemma 5.6 can be used to compute the quadratic approximation to the profile log-likelihood as well as approximate likelihood intervals. As a combinant it is of the same form as a squared Z-score with $\hat{\text{se}} = \sqrt{((\mathcal{J}^{\text{obs}})^{-1})_{pp}}$.

Nonparametric likelihood methods

In this section we consider likelihood methods in a nonparametric context. This serves two purposes. It is first of all of intrinsic interest to be able to use likelihood methods – and likelihood intervals in particular – for statistical inference without the need to make (restrictive) parametric model assumptions. It does, in addition, serve as a useful preparation for the methods we will consider in the following chapter on survival analysis. The approach taken is quite general and provides, among other thing, a definition of a maximum likelihood estimator in cases where the statistical model does not have a dominating measure.

Consider a probability measure ρ and a σ-finite measure ν such that $\rho \ll \nu$, and let $\frac{\mathrm{d}\rho}{\mathrm{d}\nu}$ denote the Radon-Nikodym derivative of ρ w.r.t. ν. With i.i.d. observations $Y_1, \ldots, Y_n$ having distribution ρ the likelihood w.r.t. ν is defined as

$$L(\rho; \nu) = \prod_i \frac{\mathrm{d}\rho}{\mathrm{d}\nu}(Y_i).$$

For a parametrized family ρ_β dominated by ν, that is, $\rho_\beta \ll \nu$, the likelihood as a function of the parameter β is $L(\beta) = L(\rho_\beta; \nu)$.

In the following, $\mathcal{M}$ denotes the set of all probability measures.

Definition 5.7. *Given the data set we define a total order on the set of all probability measures by $\rho \succ \rho'$ for $\rho, \rho' \in \mathcal{M}$ if and only if*

$$L(\rho; \rho + \rho') \geq L(\rho'; \rho + \rho').$$

Note that Radon-Nikodym derivatives are only determined uniquely up to null-sets for $\rho + \rho'$. For the definition to make sense it is in principle important that the likelihood is not changed according to the given data set. That is, for all pairs of measures we should upfront decide on the version of the Radon-Nikodym derivative that we want to use in the definition of the total order. Different choices can result in different orders. For the subsequent applications this turns out to be immaterial – all orders give rise to the same nonparametric maximum likelihood estimator. A maximum likelihood estimator is now defined as a maximal element in a statistical model $\mathcal{M}_0 \subseteq \mathcal{M}$.

Definition 5.8. *The measure $\rho \in \mathcal{M}_0 \subseteq \mathcal{M}$ is a nonparametric maximum likelihood estimator (NPMLE) in $\mathcal{M}_0$ if $\rho \succ \rho'$ for all $\rho' \in \mathcal{M}_0$.*

Theorem 5.9. *The NPMLE in $\mathcal{M}$ is the empirical measure,*

$$\hat{\rho} = \frac{1}{n} \sum_i \delta_{Y_i}.$$

Proof. To simplify the proof we assume that $Y_1, \ldots, Y_n$ are all distinct. If they are not, the proof has to be modified slightly so that the multiplicity of each observed value is taken into account. We also assume that $n \geq 2$ leaving the case $n = 1$ to the reader.

First assume that ρ has strictly positive point probabilities in $Y_1, \ldots, Y_n$ and that ρ' doesn't. Then $\nu = \rho + \rho'$ has point masses in $Y_1, \ldots, Y_n$, and it follows that $\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(Y_i) > 0$ for all i whereas $\frac{\mathrm{d}\rho'}{\mathrm{d}\nu}(Y_i) = 0$ for at least one i. Thus $L(\rho; \nu) > 0 = L(\rho'; \nu)$ and $\rho \succ \rho'$.

We conclude that the maximum must be found among those measures with strictly positive point probabilities in $Y_1, \ldots, Y_n$. If ρ' has point probabilities in $Y_1, \ldots, Y_n$ as well, we have that

$$\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(Y_i) = \frac{\rho(Y_i)}{\rho(Y_i) + \rho'(Y_i)} = \frac{p_i}{p_i + q_i}$$

and

$$\frac{\mathrm{d}\rho'}{\mathrm{d}\nu}(Y_i) = \frac{\rho'(Y_i)}{\rho(Y_i) + \rho'(Y_i)} = \frac{q_i}{p_i + q_i},$$

where $p_i = \rho(Y_i) \in (0,1)$ and $q_i = \rho'(Y_i) \in (0,1)$. It follows that

$$Q := \frac{L(\rho;\nu)}{L(\rho';\nu)} = \prod_i \frac{p_i}{q_i}.$$

Since $\log(x) \leq x - 1$ with equality if and only if $x = 1$ we find, with $p_i = \frac{1}{n}$, that

$$\log Q = -\sum_i \log(nq_i) \geq \sum_i (1 - nq_i) = n - n\sum_i q_i \geq 0,$$

with equality if and only if $q_i = \frac{1}{n}$. In conclusion, the empirical distribution is the unique nonparametric MLE. □

If the observations are real valued it follows that the NPMLE of the distribution function is the empirical distribution function

$$\hat{F}(x) = \frac{1}{n}\sum_i 1(Y_i \leq x).$$

The definition of the NPMLE relies entirely on pairwise comparisons. It is natural to ask if it can be understood as the maximizer of a likelihood function. This would, among other things, allow us to construct likelihood intervals. Clearly, we cannot hope for all probability measures to be dominated by a single measure, and thus for a likelihood function in a classical sense. Fortunately, an alternative and useful definition is possible.

Definition 5.10. *The nonparametric likelihood is defined as*

$$L(\rho) = \prod_{i=1}^{n} \rho(\{Y_i\}).$$

This definition is a very literal definition of a likelihood – the probability of the data given the model. Moreover, for probability measures ρ and ρ' with point masses in $Y_1, \ldots, Y_n$ the nonparametric likelihood ratio is given as

$$Q = \frac{L(\rho;\rho+\rho')}{L(\rho';\rho+\rho')} = \frac{L(\rho)}{L(\rho')}.$$

This implies that L gives the same ordering of all probability measures with point probabilities in the observations. In fact, if ρ is a probability measure with point masses in the observations and ρ' is any probability measure, then $L(\rho) \geq L(\rho')$ if and only if $\rho \succ \rho'$. We observe, in particular, that the NPMLE maximizes L over all probability measures.

With

$$\ell(\rho) = \sum_{i=1}^{n} \log \rho(Y_i).$$

the nonparametric log-likelihood it is tempting to define the profile log-likelihood for a univariate parameter of interest $\gamma = \gamma(\rho) \in \mathbb{R}$ as

$$\ell(\gamma) = \sup_{\rho : \gamma(\rho) = \gamma} \ell(\rho).$$

This does not lead to a useful theory. With this unrestricted form of the profile log-likelihood it is easy to construct examples where

$$\ell(\gamma) = \ell_{\max}$$

for all $\gamma \in \mathbb{R}$, see Exercise 5.1. Such examples are constructed by taking y different from all observations and considering

$$\rho_y = \varepsilon \delta_y + (1 - \varepsilon)\hat{\rho}$$

where $\hat{\rho}$ is the NPMLE and $\varepsilon \in (0, 1)$, in which case

$$\ell(\rho) = \ell_{\max} + n \log(1 - \varepsilon).$$

The last term can be made small by choosing ε small. If we can solve the equation $\gamma(\rho_y) = \gamma$ in y for a given γ and all small ε it follows that $\ell(\gamma) = \ell_{\max}$.

If the observations are all distinct

$$\ell_{\max} = -n \log n.$$

Otherwise ties have to be taken into account. For practically relevant computations ties play no role.

A useful *restricted* profile log-likelihood is obtained by restricting the probability measures to those whose support is the observations. Denote by $\mathcal{S} \subseteq \mathcal{M}$ the set of probability measures of the form

$$\rho = \sum_{i=1}^{n} \rho_i \delta_{Y_i}.$$

We identify $\rho \in \mathcal{S}$ with the n-dimensional vector $(\rho_1, \ldots, \rho_n)$ of weights[9] satisfying $\rho_i \geq 0$ and $\sum_{i=1}^{n} \rho_i = 1$.

Definition 5.11. *The nonparametric profile log-likelihood for γ is defined as*

$$\ell(\gamma) = \sup_{\rho \in \mathcal{S} : \gamma(\rho) = \gamma} \ell(\rho).$$

[9] The ρ_i's are not uniquely defined if there are ties among the observations.

The nonparametric family of likelihood intervals is defined as

$$\mathcal{I}(c) = \{\gamma \mid 2(\ell_{\text{max}} - \ell(\gamma)) < c\}.$$

Lemma 5.12. *The log-likelihood ratio identity*

$$\ell(\gamma) - \ell_{\text{max}} = \sup_{\rho:\gamma(\rho)=\gamma} \sum_{i=1}^{n} \log(n\rho_i)$$

holds.

Proof. If there are no ties, this follows directly. If the first $k \geq 2$ observations are equal the first k terms in $\ell(\gamma) - \ell_{\text{max}}$ sum to

$$k \log\left(\sum_{i=1}^{k} \rho_i\right) - k \log\left(\frac{k}{n}\right) \quad = \quad k \log\left(\frac{\sum_{i=1}^{k} \rho_i}{k}\right) + k \log(n).$$

All probability measures with the same sum $\sum_{i=1}^{k} \rho_i$ are identical (because of the ties). Due to the concavity of $\log$ and Jensen's inequality the term above is greater than or equal to

$$\sum_{i=1}^{k} \log(\rho_i) + k \log(n) = \sum_{i=1}^{k} \log(n\rho_i)$$

with equality if and only if the ρ_i's are identical for $i = 1, \ldots, k$. This shows the result. □

Observe how the likelihood intervals can be computed by profiling the "$-2\log(Q)$" function as follows

$$\mathcal{I}(c) = \{\gamma \mid \inf_{\rho:\gamma(\rho)=\gamma} -2 \sum_{i=1}^{n} \log(n\rho_i) < c\},$$

for which ties play no role.

The nonparametric likelihood theory works well with general estimating equations. If β is defined by the equation

$$Em(Y, \beta) = 0 \tag{5.1}$$

as a function of the distribution of Y, a nonparametric MLE of β is a solution to

$$\frac{1}{n} \sum_{i=1}^{n} m(Y_i, \beta) = 0.$$

If the parameter of interest, $\gamma(\beta)$, is defined as a function of the β that solves (5.1), we get that

$$\{\rho \mid \gamma(\rho) = \gamma\} = \left\{ \rho \;\middle|\; \sum_{i=1}^{n} \rho_i m(Y_i, \beta) = 0,\ \gamma(\beta) = \gamma \right\}.$$

Calibration

To calibrate the cutoff for a likelihood interval[10] to achieve a desired sampling property such as 95% coverage, knowledge of the sampling distribution of the combinant is required.

Asymptotic results

We will present two results from asymptotic theory that are useful for calibration of likelihood intervals.

Theorem 5.13 (Parametric Wilks). *Suppose that*

$$(Y_1, X_1), \ldots, (Y_n, X_n)$$

are i.i.d. such that the distribution of $Y_i \mid X_i$ *is a generalized linear model with parameter* $\beta^0 \in \mathbb{R}^p$*. Let* $\gamma^0 = \gamma(\beta^0)$ *and let* $\hat{\gamma} = \gamma(\hat{\beta})$ *be the MLE. With* ℓ *the parametric log-likelihood it holds, under suitable regularity conditions, that*

$$2(\ell(\hat{\gamma}) - \ell(\gamma^0)) \xrightarrow{\mathcal{D}} \chi^2_1$$

for $n \to \infty$*.*

We refrain from discussing the technical regularity conditions needed to make Wilks theorem a precise mathematical statement. What we emphasize is that for the parametric Wilks theorem to be true, the model must be correct. That is, the conditional distribution of the observations Y_i given the predictors X_i must be the generalized linear model given by β^0. A conditional version of the parametric Wilks theorem is possible by requirering certain asymptotic conditions on the predictors.

Theorem 5.14 (Nonparametric Wilks). *Suppose that*

$$(Y_1, X_1), \ldots, (Y_n, X_n)$$

[10] Or some other interval estimate based on a combinant for that matter.

are i.i.d. such that $E(m(Y_i, X_i, \beta^0)) = 0$*. Let* $\gamma^0 = \gamma(\beta^0)$*. With* ℓ *the nonparametric log-likelihood it holds, under suitable regularity conditions, that*

$$2(\ell_{max} - \ell(\gamma^0)) \xrightarrow{\mathcal{D}} \chi_1^2$$

for $n \to \infty$*.*

A precise version of this theorem can be found as Theorem 3.4 in the book Empirical Likelihood[11]. To paraphrase Art Owen:

> The interesting thing about Theorem 5.14 is what is not there. It includes no conditions to make $\hat{\beta}$ a good estimate of β^0, nor even conditions to ensure a unique value for β^0, nor even that any solution $\hat{\beta}$ exists.

Bootstrapping

Bootstrapping is an alternative to asymptotic theory for approximation of sampling distributions of statistics or combinants of interest. Bootstrapping is usually based on simulations rather than analytic methods. Specifically, by *resampling* new data sets. This can be used for calibration of likelihood intervals as well as other forms of intervals based on approximate pivots. The general outline of the application of bootstrap methods is the following two-step procedure.

- Sample B independent new data sets $\mathbf{Y}_1^*, \ldots, \mathbf{Y}_B^*$.
- Compute a combinant of interest from each resampled data set.

The combinant is typically either a statistic, e.g. a parameter estimator, or an approximate pivot. From the B samples it is possible to compute estimates of sampling quantities such as the standard error or the bias of an estimator, or quantiles of a combinant. As we discuss further below, the principle behind bootstrapping does not require that we use simulations. The idea is to approximate sampling distributions in a data driven way that avoids model dependent analytic approximations. The data dependent sampling distribution can, however, typically not be treated by analytic computations, and practical applications therefore require simulations. The finite number B of resampled data sets introduces a sampling error, and B should be chosen as large as possible to avoid this

[11] ART OWEN. *Empirical likelihood*, Chapman & Hall/CRC, 2001

sampling error to influence on the reported results. An appropriate choice is dictated by the trade-off between computational costs and the accuracy required by the application. We will give more advice on the choice of B below when we consider different concrete applications of bootstrapping.

We only consider how bootstrapping can be used for the construction of confidence intervals for parameters of interest. In this case there are two independent decisions to make. First, which kind of interval do we want to compute[12]? Second, how should we resample the data sets? We first treat the interval constructions, and then we treat different ways, known as parametric and nonparametric bootstrapping, for resampling data.

Bootstrap estimation of standard errors is useful for computing standard confidence intervals in cases where we don't have an analytic formula for the standard error. The parameter of interest will be called γ.

From the resampled data $\mathbf{Y}_1^*, \ldots, \mathbf{Y}_B^*$ we compute reestimates

$$\hat{\gamma}_b = \hat{\gamma}(\mathbf{Y}_b^*).$$

Then we estimate the standard error of $\hat{\gamma}$ as

$$\hat{\text{se}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\gamma}_b - \overline{\gamma})^2}$$

with $\overline{\gamma} = \frac{1}{B}\sum_{b=1}^{B}\hat{\gamma}_b$.

We can use $\hat{\text{se}}$ to construct the standard confidence intervals

$$\hat{\gamma} \pm 1.96\hat{\text{se}}.$$

For estimation of the standard error, $B = 200$ is usually enough for acceptable precision.

Note that the bootstrap procedure for estimating the standard error does not in itself have an edge over analytic methods, and the confidence interval construction still rely on an approximate $\mathcal{N}(0,1)$-distribution of

$$\frac{\hat{\gamma} - \gamma}{\hat{\text{se}}}.$$

[12] A standard interval, a likelihood interval or an interval based on some other combinant.

For combinant based confidence intervals the simplest combinant to consider is arguably

$$R(\mathbf{Y}, \gamma) = \hat{\gamma} - \gamma.$$

With w_α and z_α the α and $1 - \alpha$ quantiles for R, the nominal level $(1 - 2\alpha)$-confidence interval is

$$\{\gamma \in \mathbb{R} \mid \hat{\gamma} - \gamma \in [w_\alpha, z_\alpha]\} = [\hat{\gamma} - z_\alpha, \hat{\gamma} - w_\alpha].$$

Estimates, $\hat{w}_\alpha$ and $\hat{z}_\alpha$, of the quantiles are obtained as the empirical quantiles for the bootstrapped sample of the combinant

$$\hat{\gamma}_1^* - \hat{\gamma}, \ldots, \hat{\gamma}_B^* - \hat{\gamma}.$$

Here $\hat{\gamma}_b^* = \hat{\gamma}(\mathbf{Y}_b^*)$ is the estimate based on the b'th resampled data set $\mathbf{Y}_b^*$. This is a simple procedure, but approximate pivotality of R is often questionable.

Note that if $\hat{q}_\alpha$ and $\hat{r}_\alpha$ are the α and $(1 - \alpha)$-quantiles for the empirical distribution of the reestimates

$$\hat{\gamma}_1^*, \ldots, \hat{\gamma}_B^*,$$

then

$$\hat{w}_\alpha = \hat{q}_\alpha - \hat{\gamma} \quad \text{and} \quad \hat{z}_\alpha = \hat{r}_\alpha - \hat{\gamma}$$

and the confidence interval is

$$[2\hat{\gamma} - \hat{r}_\alpha, 2\hat{\gamma} - \hat{q}_\alpha].$$

Why is the confidence interval not just $[\hat{q}_\alpha, \hat{r}_\alpha]$? Indeed, the latter is known as the *percentile* confidence interval[13]. If the distribution of $\hat{\gamma}$ is close to being symmetric around γ, the two intervals are almost identical.

A couple of alternative combinants to consider are

$$R(\mathbf{Y}, \gamma) = \frac{(\hat{\gamma} - \gamma)^2}{\widehat{\mathrm{se}}^2}$$

or the (profile) negative log-likelihood

$$R(\mathbf{Y}, \gamma) = 2(\ell_{\max} - \ell(\gamma)).$$

[13] The justification of the percentile interval is based on pivotality and symmetry of a monotone transformation. It has the nice property of being invariant to monotone parameter transformations

For either of these two combinants we compute the estimate $\hat{c}_\alpha$, from the bootstrapped sample, as the empirical $(1-\alpha)$-quantile, and the corresponding confidence interval for γ is

$$\{\gamma \mid R(\mathbf{Y}, \gamma) < \hat{c}_\alpha\}.$$

Whether the actual coverage is close to the nominal level $1-\alpha$ still depends upon approximate pivotality of the combinant, but not on the distribution being approximately a χ^2-distribution. With R the log-likelihood combinant we get likelihood intervals calibrated to have approximate $1-\alpha$ coverage.

When we use bootstrapping to compute quantile based confidence intervals as above, the rule of thumb is to take $B = 999$. The odd number $B = 999$ is suggested because the empirical 95% quantile is then uniquely defined. If we aim for larger coverage[14] we need to take B even larger to get an acceptable relative accuracy.

Parametric bootstrapping amounts to resampling data from the parametrically specified full model of the data. That is, the distribution of $\mathbf{Y}$ is assumed to be in a parametrized family ρ_β of probability measures, and $\hat{\beta}$ is the estimate of β based on $\mathbf{Y}$. Then we resample $\mathbf{Y}^*$ from $\rho_{\hat{\beta}}$. The parameter of interest, γ, may be one coordinate of a multivariate β-parameter, but it may also be any (nonlinear) function of β.

In the regression setup we need the strong distributional assumptions GA3 and A5 to be able to implement parametric bootstrapping. We rarely want to make parametric assumptions about the distribution of the predictors, and we therefore condition on the predictors. In this case $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is an n-dimensional vector, whose coordinates are conditionally independent given the predictors. Thus we have to resample the coordinates $Y_i^* \mid X_i$ independently from the fitted exponential dispersion model.

Nonparametric bootstrapping does not require a fully specified parametric model, but it does require distributional assumptions that allow us to sample new data sets from an empirical distribution. The standard assumption is that $\mathbf{Y} = (Y_1, \ldots, Y_n)$ with $Y_1, \ldots, Y_n$ i.i.d. Then each resampled data set is obtained by sampling n observations from $Y_1, \ldots, Y_n$ with replacement. This is in

[14] From a testing point of view this means computing a more accurate p-value for hypotheses on γ.

practice carried out by sampling n independent indices uniformly from $\{1, \ldots, n\}$. A resampled data set $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*)$ thus consists of n i.i.d. samples from the empirical distribution

$$\sum_{i=1}^{n} \delta_{Y_i}.$$

In a regression context the response variables $Y_1, \ldots, Y_n$ are **not** identically distributed conditionally on the predictors. Thus we cannot just resample the responses with replacement. In observational studies it may be a reasonable assumption that the pairs (Y_i, X_i) are i.i.d., while in designed experiments no sampling assumptions are made about the predictors. In either case, we can implement nonparametric bootstrapping by sampling n pairs from $(Y_1, X_1), \ldots, (Y_n, X_n)$ with replacement. The resampled data set thus consists of n i.i.d. samples from the the empirical distribution

$$\sum_{i=1}^{n} \delta_{(Y_i, X_i)},$$

and the actual sampling is still done in practice by sampling n independent indices uniformly from $\{1, \ldots, n\}$. This version of nonparametric bootstrapping adapted to the regression setup is known as *pair sampling*.

RESIDUAL SAMPLING is a version of nonparametric bootstrapping for regression models where residuals and not responses are resampled. For designed experiments in particular, the sampling distribution sought is the conditional distribution given the predictors, which is not what pair sampling gives. In this case pair sampling appears inappropriate[15]. We will only consider the case where

$$Y_i = \mu(X_i^T \beta) + \varepsilon_i$$

with $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. In this case the raw residuals are

$$e_i = y_i - \mu(X_i^T \hat{\beta}) = y_i - \hat{\mu}_i.$$

We resample a new data set by resampling raw residuals with replacement. That is, a resampled data set $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*)$ is constructed as follows:

[15] Appearance is not everything. It usually gives appropriate results in practice.

- Sample n residuals $e_1^*, \ldots, e_n^*$ from $\{e_1, \ldots, e_n\}$ with replacement.
- Set $Y_i^* = \hat{\mu}_i + e_i^*$.

For residual sampling the explanatory variables are held fixed as for parametric bootstrapping. The resampling of residuals is in practice carried out as for the other nonparametric bootstrapping procedures by uniformly sampling indices.

The distributional assumptions made for residual sampling are GA1 together with the i.i.d. assumption on the raw residuals. The i.i.d. assumption is, in fact, a relatively strong assumption that does not allow for a non-constant variance function among other things. It is possible to introduce generalizations of residual sampling by sampling standardized errors or by including weights to attempt to handle non-constant variance, but we will not pursue this any further.

The bootstrap method was introduced in a seminal paper[16] by Bradley Efron. It was inspired, in particular, by jackknife methods, which, among other things, had been introduced earlier for nonparametric bias and variance estimation of statistics. Like jackknife methods, bootstrapping was developed in a nonparametric framework, and the original resampling method was what is now known as nonparametric bootstrapping. The goal of bootstrapping is to compute an approximation of the sampling distribution of a combinant $R(\mathbf{Y}, \gamma)$ when $\mathbf{Y} \sim \rho$ for an unknown ρ. The parameter of interest, $\gamma = \gamma(\rho)$, is a functional of the unknown sampling distribution ρ of $\mathbf{Y}$. The principle behind bootstrapping is the distributional approximation

$$R(\mathbf{Y}, \gamma) \overset{\mathcal{D}}{\simeq} R(\mathbf{Y}^*, \hat{\gamma}) \tag{5.2}$$

with $\mathbf{Y}^* \sim \hat{\rho}$. Here $\hat{\rho}$ is an estimate based on $\mathbf{Y}$ of the distribution of $\mathbf{Y}$.

Suppose that we have a parametrized model ρ_β, that $\rho = \rho_{\beta_0}$ and that $\gamma = \gamma(\beta_0)$. With $\hat{\beta}$ the estimate of β we can use the plugin estimates $\hat{\gamma} = \gamma(\hat{\beta})$ and $\hat{\rho} = \rho_{\hat{\beta}}$ of γ and ρ, respectively. The combinant R is a pivot if the sampling distribution of

$$R(\mathbf{Y}, \gamma(\beta_0)), \qquad \mathbf{Y} \sim \rho_{\beta_0}$$

[16] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979

does not depend upon β_0. If R is a pivot, the approximation (5.2) is a distributional identity. That is, there is no approximation error, and the sampling distribution can be computed exactly as the distribution of $R(\mathbf{Y}^*, \gamma(\hat{\beta}))$ for $\mathbf{Y}^* \sim \rho_{\hat{\beta}}$. The standardized Z-score in Theorem 2.4 is an example of a pivot, whose distribution can be found analytically. The analytic and the parametric bootstrap solution coincide in this case. For nonparametric bootstrapping we cannot expect R to be a pivot, and a systematic study of the degree of pivotality of different combinants is beyond the scope of this book. The take-home message is simply that the distribution of the combinant R should be as insensitive as possible to the joint approximation error $(\hat{\rho}, \hat{\gamma}) \simeq (\rho, \gamma)$.

Note that it is quite possible that $\hat{\gamma} \neq \gamma(\hat{\rho})$. Indeed, γ is in a regression context typically a plugin estimate $\gamma(\hat{\beta})$ based on the parametric estimate of β, while $\hat{\rho}$ is (derived from) an empirical distribution.

Advertising – a bootstrap example

We use the vegetables data set from the advertising case to illustrate computation of confidence and likelihood intervals and the use of bootstrapping.

Likelihood and approximate likelihood intervals

We fit a Poisson model with the same formula as the Γ-model used earlier (with a natural cubic spline expansion of `discount`) and we compute confidence intervals for the regression parameter associated with the variable `ad`. We first compute standard confidence intervals based on estimates of the standard errors. This is achieved using the `confint.default` function. Then we compute likelihood intervals based on the profile log-likelihood function. This is achieved by calling the generic `confint` function, which for objects of class `glm` will invoke a method that computes likelihood intervals.

```r
vegetablesGlm <- glm(form,
                     family = poisson,
                     data = vegetables)
confint.default(vegetablesGlm, "ad1")
```

```
##      2.5 % 97.5 %
## ad1 0.4223 0.5162

confint(vegetablesGlm, "ad1")

## Waiting for profiling to be done...

##  2.5 % 97.5 %
## 0.4223 0.5162
```

The standard intervals and the likelihood intervals are in this case identical up to the fourth decimal.

Bootstrap calibration

We then turn to nonparametric bootstrapping using pair sampling. The implementation is based on `sample` for sampling indices from 1 to `n` (the number of cases) with replacement. We store only the parameter of interest for each fitted model. An alternative is to store the entire object returned by `glm` and then subsequently extract what is needed.

```
B <- 999
n <- nrow(vegetables)
beta <- numeric(B)
for(b in 1:B) {
  i <- sample(n, n, replace = TRUE)
  bootGlm <- glm(form,
                 family = poisson,
                 data = vegetables[i, ])
  beta[b] <- coefficients(bootGlm)["ad1"]
}
```

An alternative to the nonparametric bootstrapping is to sample conditionally on the predictors from the fitted Poisson model. This is almost as easy as nonparametric bootstrapping. It can be done by extracting the fitted means from the model and using the `rpois` function for the simulation based on the fitted means. The generic `simulate` function

```
parbeta <- numeric(B)
vegetablesSamp <- vegetables
for(b in 1:B) {
  vegetablesSamp$sale <- simulate(vegetablesGlm)[, 1]
  bootGlm <- glm(form,
                 family = poisson,
                 data = vegetablesSamp)
  parbeta[b] <- coefficients(bootGlm)["ad1"]
}
```

We will compute two bootstrap confidence intervals below, but first we take a look at the standard deviations estimated from the bootstrap samples. These are estimates of the standard error, and can be used for computing standard confidence intervals.

```
sebeta <- sd(beta)
separbeta <- sd(parbeta)
## Standard error based on nonparametric bootstrapping
sebeta

## [1] 0.1553868

## Standard error based on parametric bootstrapping
separbeta

## [1] 0.02321514

## Standard error based on ordinary analytic approximations
coef(summary(vegetablesGlm))["ad1", 2]

## [1] 0.023956
```

We observe that the standard error based on parametric bootstrapping is almost identical to that obtained from the analytic approximation. The standard error based on nonparametric bootstrapping is, however, around 6 times larger. The corresponding confidence interval becomes much wider than the one based on the Poisson model.

```
betahat <- coefficients(vegetablesGlm)["ad1"]
betahat + 1.96 * sebeta * c(-1, 1)

## [1] 0.1646853 0.7738016
```

This wider (and more realistic) confidence interval is explainable by the overdispersion in the data and the poor fit of the variance dictated by the Poisson model, as has already been observed in the model diagnostics.

An alternative to the confidence interval based on the standard error is to use the simple combinant

$$\hat{\beta} - \beta.$$

It results in the following confidence interval.

```
betastar <- beta - betahat
qbeta <- quantile(betastar, probs = c(0.975, 0.025), type = 1)
betahat - qbeta

##     97.5%      2.5%
## 0.1809314 0.7899651
```

We observe that this confidence interval is moved slightly to the right. If we consider the density estimate of the bootstrapped distribution we see a left skewness in the distribution relative to $\hat{\beta}$, which is reflected as a translation to the right of the confidence interval.

```
qplot(beta, fill = I('gray'), geom = "density") +
  geom_vline(aes(xintercept=betahat), color = "red")
```

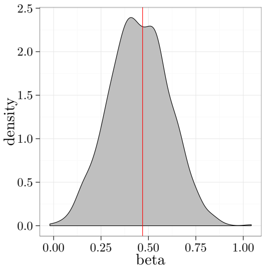


Figure 5.1: Density estimate of the bootstrapped distribution of $\hat{\beta}$.

The choice of $B = 999$ above is recommended for computation of quantile based confidence intervals. A choice of $B = 200$ is sufficient if we just want to estimate standard errors.

The quasi Poisson and the Γ*-model*

For comparison we compute likelihood intervals based on the quasi Poisson model and the Γ-model.

```
vegetablesGlm2 <- glm(form,
                      family = quasipoisson,
                      data = vegetables)
confint(vegetablesGlm2, "ad1")

## Waiting for profiling to be done...

##  2.5 % 97.5 %
## 0.3081 0.6306
```

The use of the quasi Poisson model clearly improves the confidence intervals, but they are still not as wide as those obtained from nonparametric bootstrapping. The quasi Poisson model is only capable of correcting for overdispersion, and does not correct for a lack of fit of the variance function.

```
vegetablesGlm3 <- glm(form,
                      family = Gamma("log"),
                      data = vegetables)
confint(vegetablesGlm3, "ad1")
```

```
## Waiting for profiling to be done...

##     2.5 %    97.5 %
## -0.001019  0.369257
```

Though the Γ-model with the log-link has the same mean value structure as the Poisson model, the fit and the confidence intervals are quite different. The fit will be different because the weights obtained from the variance function are different. Cases with a large expected value will be weighted further down when we use the Γ-model, and will thus contribute less to the fit. If the variance given by the Γ-model is more appropriate, this can improve on the efficiency of the estimator.

Model assessment

When we fit a model to data we make assumptions, and given the assumptions we find an approximating model. Our hope is that it can give us insights that can be generalized beyond the given data. To investigate model assumptions we can consider model diagnostics, whose purpose it is to validate if the model actually fits the data. If there are glaring deviations from the model assumptions we must be suspicious about statistical conclusions[17] based on the model assumptions.

For purely predictive purposes a good model fit is neither necessary nor sufficient[18]. It appears plausible that if a model does not fit, we can improve the predictive power of the model by improving its fit. In the framework of the linear model, for instance, the inclusion of a nonlinear expansion of a predictor variable might improve the fit as well as the predictive power of the model. What lurks in the background is the risk of *overfitting*. Overfitting means that we model properties of the data that are specific to the given data set, and which do not generalize to other data. For a given data set there is always a limit to how complex a model we can fit without overfitting. This section deals with methods for assessing the model as a predictive model.

The aim of predictive model assessment is to quantify how well a model works as a predictive model. That is, how does the model predict new data that is independent of the data set used for fitting

[17] Such as p-values of statistical tests or confidence intervals of parameters of interest

[18] Is this surprising?

the model? This is of interest in itself and is also essential for *model selection*, that is, for selecting one model among several possible models.

For the linear model the residual sum of squares

$$\text{RSS} = ||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2$$

quantifies how well $X_i^T \hat{\beta}$ predicts Y_i for $i = 1, \ldots, n$. The absolute value of RSS is difficult to interpret. The (in)famous R^2 is a first attempt to introduce an interpretable index of predictive power[19].

Definition 5.15. *Let RSS_0 be the residual sum of squares for the intercept only model and RSS the residual sum of squares for the model of interest (of dimension p). The coefficient of determination is*

$$R^2 = \frac{RSS_0 - RSS}{RSS_0} = 1 - \frac{RSS}{RSS_0}.$$

Observe that the F-test statistic for the intercept only model is

$$F = \frac{n-p}{p-1}\frac{R^2}{1-R^2}$$

Thus the coefficient of determination and the F-test are in one-to-one correspondence via a monotonely increasing function. The closer R^2 is to 1, the larger is the F-test. With a single continuous predictor R^2 is the *square of the Pearson correlation* between the response and the predictor. For this reason the square root of R^2 is called the *multiple correlation coefficient.*

For two non-nested models with the same complexity (same dimension p in this case), the one with the largest R^2 (smallest `RSS`) is the best predictor on the data set. But R^2 is monotonely increasing in p for nested models, which implies that among nested models the more complex model will always predict better on the data. This does not reflect what will happen if we want to make predictions on an independent observation $(Y^{\text{new}}, X^{\text{new}})$. The average residual sum of squares, `RSS`$/n$, will generally underestimate the *prediction error*

$$E\left((Y^{\text{new}} - (X^{\text{new}})^T\hat{\beta})^2 \mid \hat{\beta}\right),$$

and the more complex the model is (the larger p is), the more serious is the underestimation. It is quite possible that `RSS` $\simeq 0$ ($R^2 \simeq 1$)

[19] The R^2 quantity has surely been widely overused and overinterpreted, but it is a sensible quantity to consider, in particular for observational data. Note also the definition of the adjusted R^2 below.

even though the prediction error is still considerable. To (partly) remedy the optimism of R^2 we introduce the *adjusted* R^2. We can interpret $1 - R^2$ as a ratio of variance estimates,

$$1 - R^2 = \frac{\mathrm{RSS}/n}{\mathrm{RSS}_0/n} = \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}$$

using biased estimates of the variance. The adjusted R^2 is defined as

$$\overline{R}^2 = 1 - \frac{\mathrm{RSS}/(n-p)}{\mathrm{RSS}_0/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-p},$$

where the variance ratio is based on unbiased estimates of the variance.

The definition of the adjusted R^2 can easily be extended to generalized linear models by replacing $\mathrm{RSS}/(n-p)$ with $\mathcal{X}^2/(n-p)$, where

$$\mathcal{X}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

is the Pearson χ^2-statistic, or by $D/(n-p)$, where

$$D = \sum_{i=1}^{n} d(Y_i, \hat{\mu}_i)$$

is the deviance.

General methods for assessment are based on the introduction of a *loss* function. The loss function is a function, $\ell(Y, \rho)$, that quantifies how badly the probability measure ρ predicts Y. If ρ has a density h w.r.t. the Lebesgue measure, say, the negative log-likelihood loss

$$\ell(Y, h \cdot m) = -\log h(Y)$$

can be used. In some cases the loss function only depends on certain aspects of ρ. For a real valued Y the squared error loss is, for instance, given as

$$\ell(Y, \rho) = (Y - \int x \, \mathrm{d}\rho(x))^2,$$

which is the squared distance from Y to the expected value under ρ.

In a regression context ρ is typically the conditional distribution of Y given the predictors X. If we use the squared error loss, the loss function is really only a function of the conditional expectation of Y given X. With $f : \mathbb{R}^p \to \mathbb{R}$ a candidate model of the function $x \mapsto E(Y \mid X = x)$ the squared error loss for Y given $X = x$ can be expressed as

$$\ell(Y, f(x)) = (Y - f(x))^2$$

as a function of $f(x)$ only.

The *prediction error*[20] of ρ is defined as

$$\mathrm{PE} = E\ell(Y, \rho).$$

If $\hat{\rho}$ is based on the data $Y_1, \ldots, Y_n$ we would like to compute PE as a measure of how well the fitted model predicts. The *empirical* prediction error

$$\mathrm{PE}_{\mathrm{emp}} = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \hat{\rho}_i)$$

is, just as R^2, an optimistic estimate of the actual prediction error. In fact, it will generally underestimate the prediction error, and the more complex a model we use for $\hat{\rho}$ the more severe is the underestimation.

The only way to estimate the prediction error of $\hat{\rho}$ is to have an independent data set, called a test or validation data set, available. With m observations in the validation data set $\mathbf{Y}^{\mathrm{new}} = (Y_1^{\mathrm{new}}, \ldots, Y_m^{\mathrm{new}})$ and $\hat{\rho}_i^{\mathrm{new}}$ computed based on $\mathbf{Y} = (Y_1, \ldots, Y_n)$,

$$\widehat{\mathrm{PE}} = \frac{1}{m} \sum_{i=1}^{m} \ell(Y_i^{\mathrm{new}}, \hat{\rho}_i^{\mathrm{New}})$$

is an unbiased estimate of the prediction error if $\mathbf{Y}^{\mathrm{new}}$ and $\mathbf{Y}$ are independent. In practice, the validation data set is obtained by setting aside a randomly selected subset containing 10-25% of the data available. The remaining 75-90% of the data is used for model fitting. Setting aside a part of the data to be used for model assessment only is in many cases not a viable option.

Without independent validation data the best we can hope for is to estimate the *expected* prediction error

$$\mathrm{EPE} = E\ell(Y, \hat{\rho})$$

[20] The expectation defining PE is also over the predictors. See below for a discussion about conditioning on the predictors.

where the expectation is over Y as well as the data $Y_1, \ldots, Y_n$ (and the predictors) used to compute $\hat{\rho}$.

DATA SPLITTING TECHNIQUES comprise a class of similar methods for estimation of the expected prediction error by mimicking the estimation-validation data splitting idea. The model is fitted using the entire data set, but EPE is estimated by the data splitting method.

The methods we will mention work by the same general principle. The data set is split into two disjoint subsets with indices I_1 and I_2. Then

- the estimates $\hat{\rho}_i$ for $i \in I_2$ are computed based on Y_j for $j \in I_1$,
- and the losses $\ell(Y_i, \hat{\rho}_i)$ are computed for $i \in I_2$.

From a single round of data splitting the obvious estimate

$$\widehat{\mathrm{EPE}} = \frac{1}{|I_2|} \sum_{i \in I_2} \ell(Y_i, \hat{\rho}_i)$$

of EPE can be computed. In practice, several rounds of data splitting will be used and all the losses will be averaged at the end.

The three major splitting schemes used are:

- Subsampling. Subset I_2 of size $m < n$ is sampled *without replacement* from $\{1, \ldots, n\}$. The procedure is replicated independently.
- Cross-validation (K-fold). The index set $\{1, \ldots, n\}$ is randomly divided into k disjoint sets $J_1, \ldots, J_K$ of roughly equal size $\simeq n/K$. We then use K rounds of data splitting with $I_1 = \{1, \ldots, n\} \backslash J_k$ and $I_2 = J_k$ for $k = 1, \ldots, K$. The entire procedure can be replicated with new random K-folde division.
- Bootstrapping. Subset I_1 of size n (replicates counted) is sampled *with replacement* from $\{1, \ldots, n\}$. The procedure is replicated independently.

Of these three, cross-validation with $K = 5$ or $K = 10$ is the most popular, and it is not too computationally demanding. Leave-one-out CV ($K = n$) requires in principle n refits of models with $n - 1$ observations, but can in some special cases be computed analytically from the model fitted to the complete data set.

Theoretical alternatives to data splitting techniques are abundant. A concern with data splitting techniques is that they are computationally demanding just as the bootstrap is for the calibration of confidence intervals. Analytic alternatives can typically be computed efficiently. The price to pay is added model assumptions.

Definition 5.16. *Assume that $\rho_i = \rho_i(\beta)$ for $\beta \in \mathbb{R}^p$ a p-dimensional parameter β. With ℓ the negative log-likelihood loss we define Akaike's information criteria as*

$$AIC = 2\sum_{i=1}^{n} \ell(\hat{\rho}_i) + 2p.$$

History has determined the arbitrary constant "2" in the definition. In the context of generalized linear models[21] we observe that

$$\text{AIC} = \frac{1}{\psi}D + 2p$$

where $D = \sum_{i=1}^{n} d(Y_i, \hat{\rho}_i)$ is the deviance in Definition 4.30. The log-likelihood and hence AIC is only determined up to an additive constant. The estimator of EPE derived from AIC is

$$\frac{1}{2n}\text{AIC}.$$

The justification of AIC hinges on a number of assumptions, and its justification is asymptotic for $n \to \infty$. We will not pursue the arguments in any detail, but mention that we need $\hat{\rho}_i = \rho_i(\hat{\beta})$ with $\hat{\beta}$ being the maximum likelihood estimator, and that the data generating model must be of the form $\rho(\beta)$ for some β. For comparability of AIC between two models it is also essential that the same reference measure is used for the definition of the likelihood. One has to pay attention to this for generalized linear models. The structure measures for the Poisson and the binomial distributions, for example, as exponential families differ[22]. If we do not correct for this the models become incomparable in terms of their AIC.

Finite sample results conditionally can $\mathbf{X}$ are possible in special cases. If we consider the linear model with the weak assumptions A1, A2 and A4, and if $\mathbf{Y}^{\text{new}} \mid \mathbf{X} \overset{\mathcal{D}}{=} \mathbf{Y} \mid \mathbf{X}$ is a new data set of

[21] Either assuming that we know the dispersion parameter or that we just plainly ignore that it is estimated too. If it is estimated we should use a single value for all models, see the discussion below.

[22] One has the factor $\frac{1}{Y!}$ and the other $\binom{m}{Y}$ on the counting measure.

responses conditionally independent of $\mathbf{Y}$ it holds that

$$C(p_0) := E(||\mathbf{Y}^{\text{new}} - \hat{\boldsymbol{\mu}}||^2 \mid \mathbf{X}) = E(||\mathbf{Y} - \hat{\boldsymbol{\mu}}||^2 \mid \mathbf{X}) + 2p_0\sigma^2 \quad (5.3)$$

where $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\beta}_0$ is the vector of predicted values from a model of dimension p_0.

An estimator of $C(p_0)$ is

$$\hat{C}(p_0) = ||\mathbf{Y} - \hat{\boldsymbol{\mu}}||^2 + 2p_0\hat{\sigma}^2 = \text{RSS}(p_0) + 2p_0\hat{\sigma}^2$$

where $\hat{\sigma}^2 = \frac{1}{n-p}\text{RSS}(p)$. Here $\text{RSS}(p)$ denotes the residual sum of squares for a model with p predictors. We should base the estimation of σ^2 on a single large and flexible model. The corresponding estimator of the expected prediction error with the squared error loss function is $\hat{C}(p_0)/n$. From a predictive point of view it is sensible to choose among two models, nested or not, by choosing the one with the smallest value of $\hat{C}(p_0)$. The quantity

$$\frac{C(p_0)}{\hat{\sigma}^2} - n = \frac{\text{RSS}(p_0)}{\hat{\sigma}^2} - n + 2p_0$$

is known as Mallows's C_p.

For generalized linear models we can likewise estimate the dispersion parameter as

$$\hat{\psi} = \frac{1}{n-p}\mathcal{X}^2(p),$$

and a natural generalization of the $C(p_0)$ quantity above is

$$\hat{C}(p_0) = \mathcal{X}^2(p_0) + 2p_0\hat{\psi}.$$

Here $\mathcal{X}^2(p_0)$ is the Pearson χ^2-statistic. If we replace the χ^2-statistic with the deviance we get AIC back with a plug-in estimate of the dispersion parameter. The corresponding generalization of Mallows's C_p is

$$\frac{\mathcal{X}^2(p_0)}{\hat{\psi}} - n + 2p_0.$$

Exercises

Exercise 5.1. Assume that the sample space is $\mathbb{R}^k$ and the parameter of interest is the mean

$$\gamma(\rho) = \int y \, \mathrm{d}\rho(y).$$

Assume that we have observations $Y_1, \ldots, Y_n$ and let K denote the convex hull spanned by these observations (the smallest convex subset containing the observations). Show that if the support of ρ' is contained in K then $\gamma(\rho') \in K$. Show that in this case there is a

$$\rho = \sum_{i=1}^{n} \rho_i \varepsilon_{Y_i}.$$

such that $L(\rho) \geq L(\rho')$. Show that if the support of ρ' is not restricted then

$$\sup_{\rho' : \gamma(\rho') = \gamma} L(\rho') = L(\hat{\rho})$$

for all $\gamma \in \mathbb{R}^k$ where $\hat{\rho}$ is the NPMLE. ○

Exercise 5.2. Prove formula (5.3). ○

6

Survival analysis

This chapter deals with regression models for time-to-event data with possible right censoring, which is common for survival data but also in other applications of statistics such as reliability analysis[1]. Throughout this chapter the response variable will be denoted T instead of Y, which is the convention in the survival literature, and T is also a natural notation for a time-to-event observation.

Survival distributions and hazards

Regression models for survival data differ in a couple of fundamental ways from the linear and generalized linear regression models considered in previous chapters. This is primarily due to the following two issues:

- Survival distributions are skewed distributions on the positive half line. It is the *shape* of the distribution rather than the location of the distribution that is of interest.
- There is almost always a *censoring mechanism*, and certain aspects of the data are consequently missing. We need to deal with this in the modeling.

That said, the idea of modeling the conditional distribution of survival times given predictor variables through a linear predictor still works. The linear predictor just needs to enter into the survival

[1] Time to failure of an electronic component, say.

distribution in different ways than through the mean.

A typical real application of survival analysis is to the study of survival of patients after the initial diagnosis. In such a study patients are enrolled whenever they are diagnosed with a given (serious, life threatening) disease. Data on the subjects are collected, and at a planned calendar time the statistical analysis is done. All patients still alive at the time of the analysis have right censored survival times. The overall question is then how the different predictors are associated with the survival time after diagnosis for this particular disease. One objective could be to compare survival distributions for two or more treatments. Another, to give a survival prognosis – if not for the individual patient then at least for (sub)populations.

The Kaplan-Meier estimator

Before we develop the general theory of regression models we deal with the fundamental problem of just estimating a distribution when we have right censored observations. We consider n individuals and $T_1^*, \ldots, T_n^*$ identically distributed positive random variables (the survival times). We don't observe the T_i^*'s but rather the censored variables

$$T_i = \min\{T_i^*, C_i\}$$

with *censoring times* $C_1, \ldots, C_n$. We also assume that the censoring times are identically ditributed, and we assume that

$$T_1^*, \ldots, T_n^*, C_1, \ldots, C_n$$

are independent. In addition to observing the censored variable T_i we observe if the variable is censored, that is, we observe

$$e_i = 1(T_i^* \leq C_i).$$

We call e_i the *status indicator.* It is 1 if the observation is not censored and 0 otherwise. In conclusion we assume that we observe the i.i.d. pairs

$$(T_1, e_1), \ldots, (T_n, e_n).$$

The distribution function is $F(t) = P(T_1^* \leq t)$ and the *survival function* is defined as

$$S(t) = 1 - F(t) = P(T_1^* > t).$$

Without censoring, and with

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} 1(T_i^* \leq t)$$

the empirical distribution function, the natural nonparametric estimator of the survival function is the empirical survival function

$$\hat{S}(t) = 1 - \hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} 1(T_i^* > t).$$

To handle the censoring we introduce the *individuals at risk*

$$Y(t) = \sum_{i=1}^{n} 1(t \leq T_i)$$

at time t. It can be computed based on the censored data.

Definition 6.1. *With $t_1 < \ldots < t_k \leq t$ the ordered, uncensored survival times before time t the Kaplan-Meier estimator of the survival function is*

$$\begin{aligned} \hat{S}(t) &= \left(1 - \frac{1}{Y(t_1)}\right)\left(1 - \frac{1}{Y(t_2)}\right)\cdots\left(1 - \frac{1}{Y(t_k)}\right) \\ &= \prod_{i:t_i \leq t}\left(1 - \frac{1}{Y(t_i)}\right). \end{aligned}$$

The intuition behind the estimator is that each factor is an estimator of the conditional probability of surviving the time interval $(t_i, t_{i+1}]$ given survival beyond time t_i. Precisely,

$$\begin{aligned} S(t) &= P(T_1^* > t) \\ &= P(T_1^* > t \mid T_1^* > t_k) \times P(T_1^* > t_k \mid T_1^* > t_{k-1}) \times \\ &\qquad \ldots \times P(T_1^* > t_2 \mid T_1^* > t_1) \times P(T_1^* > t_1). \end{aligned}$$

There is one individual out of $Y(t_i)$ that dies in the interval $(t_i, t_{i+1}]$, whence conditionally on having survived beyond t_i the probability of dying is estimated as $1/Y(t_i)$ and the probability of surviving beyond t_{i+1} is thus $1 - 1/Y(t_i)$. This argument is based on the assumption that multiple deaths do not occur at the same time. A version that allows for multiple deaths at the same time is given as follows. With

$$N(t) = \sum_{i=1}^{n} 1(T_i \leq t, e_i = 1)$$

the *counting process* of deaths (non-censored events), and the jumps for the counting process given as

$$\Delta N(t) = N(t) - N(t-) = N(t) - \lim_{\varepsilon \to 0+} N(t - \varepsilon),$$

the Kaplan-Meier estimator can be reformulated as

$$\hat{S}(t) = \prod_{s \leq t} \left(1 - \frac{\Delta N(s)}{Y(s)}\right).$$

The factors are equal to 1 for all s where $\Delta N(s) = 0$. This version allows for $\Delta N(s) > 1$ to accommodate multiple deaths at the same time.

Hazards

If F is continuously differentiable with derivative f (the density for the survival distribution), we introduce the *hazard function*

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

Observe that

$$\begin{aligned} \lambda(t) &= \lim_{\varepsilon \to 0+} \frac{1}{\varepsilon} \frac{F(t+\varepsilon) - F(t)}{S(t)} \\ &= \lim_{\varepsilon \to 0+} \frac{1}{\varepsilon} P(T_i^* \in (t, t+\varepsilon] \mid T_i^* > t). \end{aligned}$$

Thus $\lambda(t)$ is the instantaneous rate of death at time t.

Example 6.2. The *exponential distribution* has hazard function

$$\lambda(t) = \lambda,$$

which is constant.

The *Weibull distribution* has hazard function

$$\lambda(t) = \alpha \gamma t^{\gamma - 1}$$

for $\alpha, \gamma > 0$.

The Weibull distribution can, in particular, capture increasing hazard functions over time by taking $\gamma > 1$. ○

For continuously differentiable distribution functions

$$\lambda(t) = -(\log S(t))',$$

whence

$$\Lambda(t) := \int_0^t \lambda(s)\mathrm{d}s = -\log S(t),$$

which is called the *cumulative hazard function.* Note that

$$S(t) = \exp\left(-\Lambda(t)\right).$$

With $\hat{S}$ the Kaplan-Meier estimator of the survival function the cumulative hazard can be estimated as $-\log \hat{S}$. This is a step function. An alternative, and direct, estimator of the cumulative hazard is the *Nelson-Aalen* estimator given as

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{1}{Y(t_i)}.$$

The intuition behind the Nelson-Aalen estimator is as follows. The probability that an individual that survived beyond time t_i dies in (a small) interval $(t_i, t_{i+1}]$ is approximately $\lambda(t_i)(t_{i+1} - t_i)$ by the definition of the hazard rate. The cumulative hazard function $\Lambda(t)$ is approximately the sum of these quantities for $t_i < t$. By estimating the probability of a death in $(t_i, t_{i+1}]$ as $1/Y(t_i)$ and aggregating the probabilities by summation we get the Nelson-Aalen estimator of Λ. We could plug the Nelson-Aalen estimator into the formula above for the survival function, which gives the estimator

$$\exp(-\hat{\Lambda}(t)) = \prod_{i:t_i \leq t} \exp\left(-\frac{1}{Y(t_i)}\right).$$

We may observe that

$$\exp\left(-\frac{1}{Y(t_i)}\right) \simeq \left(1 - \frac{1}{Y(t_i)}\right)$$

for large $Y(t_i)$ using the Taylor expansion $\exp(x) = 1 + x + o(x^2)$.

In a subsequent section on nonparametric survival analysis we will show how the Nelson-Aalen and Kaplan-Meier estimators can be understood as nonparametric MLEs.

Parametric survival analysis

In this section we turn to parametric modeling of survival distributions. The first issue that we address is how to get the likelihood when we have right censoring.

Assume that T^* is a positive random variable with density f and survival function S, C is a positive random variable with density g and survival function H.

We define

$$T = \min\{T^*, C\} \quad \text{and} \quad e = 1(T^* \leq C).$$

Theorem 6.3. *If T^* and C are independent the joint distribution of (T, e) has density*

$$f(t)^e S(t)^{1-e} g(t)^{1-e} H(t)^e$$

w.r.t. the product measure $m \otimes \tau$ (the Lebesgue measure times the counting measure).

Proof.

$$\begin{aligned} P(T \leq t, e = 1) &= P(T^* \leq t, e = 1) \\ &= P(T^* \leq t, C \geq T^*) \\ &= \int_0^t f(s) \int_s^\infty g(u)\mathrm{d}u\mathrm{d}s \\ &= \int_0^t f(s)H(s)\mathrm{d}s. \end{aligned}$$

Likewise,

$$P(T \leq t, e = 0) = \int_0^t g(s)S(s)\mathrm{d}s,$$

and we conclude that the density is

$$h(t, e) = \begin{cases} f(t)H(t) & \text{if } e = 1 \\ g(t)S(t) & \text{if } e = 0. \end{cases}$$

□

Based on the density result above we can introduce the full likelihood for right censored observations. With $(T_1, e_1), \ldots, (T_n, e_n)$

independent the full likelihood is

$$L = \prod_{i=1}^{n} f_i(T_i)^{e_i} S_i(T_i)^{1-e_i} g_i(T_i)^{1-e_i} H_i(T_i)^{e_i}.$$

We assume that $f_i = f_{i,\beta}$ is parametrized by β, and that the distributions, given by the g_i's, of the censoring mechanism hold no information about β. This implies that

$$L(\beta) = \prod_{i=1}^{n} f_{i,\beta}(T_i)^{e_i} S_{i,\beta}(T_i)^{1-e_i} K_i$$

with K_i depending on the observations but not the parameter β.

Note that the literature is full of heuristic "derivations" of the likelihood, which leaves some room for wondering exactly what the dominating measure is. This is made explicit above. The derivation also makes it clear how the distribution of the censoring mechanism enters, and why it can be ignored if it does not depend on the unknown parameter β. This has nothing to do with independence[2]. The ignorability is due to the assumption that g does not depend upon β.

It is possible to take a slightly different point of view and condition on the censoring variables instead. One arrives at the same likelihood, but this time the dominating measure for the i'th observation becomes $m + \delta_{C_i}$ (δ_{C_i} is the Dirac measure in C_i).

From hereon the likelihood we consider is

$$L(\beta) = \prod_{i=1}^{n} f_{i,\beta}(T_i)^{e_i} S_{i,\beta}(T_i)^{1-e_i}. \tag{6.1}$$

If we recall the definition of the hazard function

$$\lambda_\beta(t) = \frac{f_{i,\beta}(t)}{S_\beta(t)},$$

the likelihood can be written as

$$L(\beta) = \prod_{i=1}^{n} \lambda_{i,\beta}(T_i)^{e_i} e^{-\Lambda_{i,\beta}(T_i)} \tag{6.2}$$

where

$$\Lambda_{i,\beta}(t) = -\log S_{i,\beta}(t).$$

For continuously differentiable survival functions this is the cumulative hazard function.

[2] Which does play a role, though, for this particular derivation of the full likelihood.

Example 6.4. If T_i is exponentially distributed with rate λ, the MLE is

$$\hat{\lambda} = \frac{n_u}{\sum_{i=1}^n T_i}$$

with n_u the number of deaths (the number of uncensored observations). If we ignore censoring the MLE is

$$\frac{n}{\sum_{i=1}^n T_i} > \hat{\lambda},$$

which will overestimate the rate. If we discard censored observations the MLE is

$$\frac{n_u}{\sum_{i=1}^n e_i T_i} > \hat{\lambda},$$

which will overestimate the rate. ○

Regression models

We introduce two ways that predictor variables can affect the survival distribution. The accelerated failure time models exemplified by the log-logistic model, and the proportional hazards model exemplified by the Weibull model. In the latter case it turns out that the likelihood can be computed and optimized using methods for Poisson generalized linear models.

Example 6.5. If Z has a logistic distribution the distribution function is

$$G(z) = \frac{e^{\lambda z}}{1 + e^{\lambda z}}$$

and $T = e^Z$ – the *log-logistic* distribution – has distribution function

$$F(t) = G(\log t) = \frac{t^\lambda}{1 + t^\lambda}$$

and density

$$f(t) = F'(t) = \frac{\lambda t^{\lambda - 1}}{(1 + t^\lambda)^2}$$

for $t > 0$. We introduce a *scale* parameter as follows

$$f_\eta(t) = \frac{\lambda e^{-\lambda\eta} t^{\lambda-1}}{(1 + e^{-\lambda\eta} t^\lambda)^2}.$$

If $\eta = X^T\beta$ is a linear predictor we can model its effect on the survival distribution as a scale transformation

$$e^{\eta}T$$

of the baseline distribution of T. If T has log-logistic distribution with density f_0 then $e^{\eta}T$ is log-logistic distributed with density f_{η}, cf. Exercise 6.1. The λ parameter is a nuisance parameter that determines the shape of the baseline distribution. The survival function is

$$S_{\eta}(t) = \frac{1}{1 + e^{\lambda\eta}t^{\lambda}}.$$

○

Definition 6.6. *An accelerated failure time (AFT) model has survival function given as*

$$S_{\eta}(t) = 1 - G((\log t - \eta)/\sigma)$$

with η the linear predictor, G a distribution function (on $\mathbb{R}$), and $\sigma > 0$ called the scale parameter.

The log-logistic model introduced in the example above is an AFT model with scale parameter $\sigma = \lambda^{-1}$, and G is the distribution function for the logistic distribution.

The interpretation of the AFT model is that a unit change of the j'th predictor increases – or *accelerates* – the failure time by a factor e^{β_j}. For maximum likelihood estimation we need to compute and optimize the likelihood (6.1). Efficient computation requires access to the density as well as the survival function, which are both explicitly available for the log-logistic distribution.

The proportional hazards models are among the most widely used survival models. The linear predictor enters here as a multiplicative effect on the hazard function.

Definition 6.7. *With λ_0 a baseline hazard function, the proportional hazards model is given by the hazard function*

$$\lambda(t) = \lambda_0(t)e^{\eta}$$

with η the linear predictor.

It follows that for the cumulative hazard function,

$$\Lambda(t) = \Lambda_0(t)e^{\eta},$$

the proportionality holds too.

The factor e^{β_j} is the *hazard ratio* between two models corresponding to a unit change of the j'th predictor.

Example 6.8. The Weibull baseline hazard function and cumulative hazard function are

$$\lambda_0(t) = \gamma t^{\gamma-1} \quad \text{and} \quad \Lambda_0(t) = t^{\gamma}.$$

The log-likelihood is

$$\begin{aligned} \ell &= \sum_{i=1}^{n} e_i \log(\gamma T_i^{\gamma-1} e^{\eta_i}) - T_i^{\gamma} e^{\eta_i} \\ &= \underbrace{\sum_{i=1}^{n} e_i \log(T_i^{\gamma} e^{\eta_i}) - T_i^{\gamma} e^{\eta_i}}_{\text{Poisson log-likelihood}} + \sum_{i=1}^{n} e_i \log(\gamma T_i^{-1}). \end{aligned}$$

This is (surprisingly) up to a constant the log-likelihood for a Poisson model of the e_i's with log link and mean value $T_i^{\gamma} e^{\eta_i}$ for fixed γ. Implementations for computations with generalized linear models – and Poisson regression models, in particular – can thus be used to fit the Weibull proportional hazards model (for fixed γ) with the survival times entering as an offset term. Note that the α parameter has been dropped in the Weibull distribution above as it is captured by an intercept in the linear predictor.

To estimate γ we can use an iterative procedure or compute the profile likelihood using a glm implementation for the optimization for a range of γ-parameters. For a fixed value of the linear predictor we find that

$$\partial_{\gamma} \ell = \sum_{i=1}^{n} (e_i - T_i^{\gamma} e^{\eta_i}) \log T_i + \frac{e_i}{\gamma}.$$

Thus γ solves the equation

$$\gamma = \frac{n_u}{\sum_{i=1}^{n} (T_i^{\gamma} e^{\eta_i} - e_i) \log T_i}$$

with n_u the number of uncensored observations.

There is no closed form solution to the equation. One idea is to use an iterative procedure and approximate the solution by

$$\gamma^{(k)} = \frac{n_u}{\sum_{i=1}^{n}(T_i^{\gamma^{(k-1)}} e^{\eta_i} - e_i)\log T_i}$$

and then reestimate the linear predictors before the next iteration.

∘

Nonparametric survival analysis

If ρ and ρ' are survival distributions on $[0,\infty)$ with corresponding survival functions S and S' we can compute the nonparametric likelihood ratio

$$Q = \frac{L(\rho;\rho+\rho')}{L(\rho';\rho+\rho')} = \prod_i \frac{\rho(T_i)^{e_i} S(T_i)^{1-e_i}}{\rho'(T_i) S'(T_i)^{1-e_i}}$$

for two measures with point probabilities in $T_1,\ldots,T_n$ in the presence of censoring.

Definition 6.9. *The nonparametric likelihood with right censoring is*

$$L(\rho) = \prod_i \rho(T_i)^{e_i} S(T_i)^{1-e_i}.$$

In the previous section we only introduced hazards for continuous distributions. In the discrete case the *hazard function* is defined as

$$\lambda_i = \frac{\rho(T_i)}{S(T_i-)} = P(T^* = T_i \mid T^* \geq T_i).$$

Lemma 6.10. *For a discrete probability measure with point masses in $T_1,\ldots,T_n$ and corresponding hazards λ_i the survival function is*

$$S(t) = \prod_{i:T_i \leq t} (1-\lambda_i).$$

Proof. Note that $S(T_j-) = S(T_{j-1})$ and that it is sufficient to show the identity for $t = T_j$. The proof is by induction. For T_1 we have $S(T_1-) = 1$ and

$$S(T_1) = 1 - \rho(T_1) = 1 - \lambda_1,$$

which gives the induction start. For the induction step, observe that the definition of the hazard implies that

$$S(T_j) = S(T_j-) - \rho(T_j) = S(T_j-) - S(T_j-)\lambda_j = S(T_{j-1})(1-\lambda_j),$$

which gives the induction step. □

The lemma above shows that the discrete hazard function still determines the survival function and thus the distribution, but in a slightly different way than for continuous distributions. We should note that the discrete hazards are restricted to be in the interval $[0,1]$.

Lemma 6.11. *With $y \geq 1$ the function*

$$\lambda \mapsto \lambda(1-\lambda)^{y-1}$$

attains a unique maximum over $[0,1]$ in $1/y$.

Proof. With $f(\lambda) = \lambda(1-\lambda)^{y-1}$ we have

$$\begin{aligned} f'(\lambda) &= (1-\lambda)^{y-1} - \lambda(y-1)(1-\lambda)^{y-2} \\ &= (1-\lambda)^{y-2}(1-\lambda-\lambda(y-1)) \\ &= (1-\lambda)^{y-2}(1-\lambda y). \end{aligned}$$

For $y = 1$, the function is the identity, which attains its maximum in $1 = 1/y$. For $y < 1$ we have $f(\lambda) \geq 0$ for $\lambda \in [0,1]$, $f(0) = f(1) = 0$ and the function has a unique stationary point in $(0,1)$ in $1/y$ with $f(1/y) > 0$. Thus $1/y$ is the unique maximizer in $[0,1]$ for $y \geq 1$. □

Theorem 6.12. *The NPMLE of the hazards – the hazards for the probability measure that maximizes L – is given by*

$$\hat{\lambda}_i = \frac{e_i}{Y_i}$$

where $Y_i = Y(T_i) = \sum_j 1(T_i \leq T_j)$ is the number of individuals at risk at time T_i.

Proof. The nonparametric likelihood function becomes, by an interchange of the product order,

$$\begin{aligned} L &= \prod_i \lambda_i^{e_i} \prod_{j:T_j<T_i} (1-\lambda_j)^{e_i} \prod_{j:T_j\leq T_i} (1-\lambda_j)^{1-e_i} \\ &= \prod_i \lambda_i^{e_i}(1-\lambda_i)^{Y_i-e_i}. \end{aligned}$$

Each factor equals

$$\lambda_i^{e_i}(1-\lambda_i)^{Y_i-e_i} = \begin{cases} (1-\lambda_i)^{Y_i} & \text{if } e_i = 0 \\ \lambda_i(1-\lambda_i)^{Y_i-1} & \text{if } e_i = 1 \end{cases}$$

In the first case, when $e_i = 0$, the factor is maximized in 0, and in the latter case, when $e_i = 1$, the factor is maximized in $1/Y_i$. □

The convention is that the cumulative hazard function in the discrete case is the jump function

$$\Lambda(t) = \sum_{i:T_i \leq t} \lambda_i.$$

We then recognize the NPMLE of the cumulative hazard function as the Nelson-Aalen estimator. It is the jump function with jumps e_i/Y_i in T_i. By plugging in the NPMLE in the formula in Lemma 6.10 for the survival function we find that the NPMLE of the survival function is

$$\hat{S}(t) = \prod_{i:T_i \leq t} \left(1 - \frac{e_i}{Y_i}\right).$$

We recognize this as the *Kaplan-Meier* estimator.

The fact that the hazard function determines the survival function in slightly different ways in the continuous and the discrete case is a nuisance. It leads to the earlier apparent difference between the Nelson-Aalen estimator and the Kaplan-Meier estimator of the survival distribution. We have resolved this issue by showing that the Nelson-Aalen estimator is the NPMLE of the *discrete* hazard function.

Our earlier definition of the nonparametric likelihood was based on (6.1). It is also possible to introduce an alternative nonparametric likelihood based on (6.2).

Definition 6.13. *The Poisson nonparametric likelihood is*

$$L^*(\Lambda) = \prod_i \lambda_i^{e_i} e^{-\Lambda_i}$$

where $\Lambda_i = \Lambda(T_i)$

In the definition we have emphasized that L^* is to be understood as a function of the (jump) cumulative hazard function. In this way,

L^* is a direct extension the likelihood (6.2) defined initially only on the set of continuous cumulative hazard functions. How we should relate the survival function to the cumulative hazard in general is not completely obvious. We can use the correct relation for discrete distributions from Lemma 6.10 even though L^* is not the correct likelihood, or we can use the relation $S(t) = e^{-\Lambda(t)}$ for continuous distributions, for which L^* is the correct likelihood. Subsequent arguments will show that the NPMLE of the cumulative hazard function using L^* is still the Nelson-Aalen estimator

$$\hat{\Lambda}(t) = \sum_{i:T_i \leq t} \frac{e_i}{Y_i}.$$

In a more general framework of counting processes the likelihood L^* can be justified as a real nonparametric likelihood[3].

Cox's proportional hazards model

Returning to the continuous proportional hazards model, the i'th individual has hazard function

$$\lambda_i(t) = w_i \lambda_0(t),$$

for $w_i > 0$ a weight, and consequently cumulative hazard function $\Lambda_i(t) = w_i \Lambda_0(t)$.

Theorem 6.14. *With $\Lambda_i(t) = w_i \Lambda_0(t)$, $W(t) = \sum_{j:t \leq T_j} w_j$ and $W_i = W(T_i)$ it holds that*

$$L = \prod_{i:e_i=1} \frac{w_i}{W_i} \left(\prod_i (W_i \lambda_0(T_i))^{e_i} \right) e^{-\int_0^\infty W(t)\lambda_0(t)\mathrm{d}t}.$$

Proof. First note that with $W(t) = \sum_{j:t \leq T_j} w_j$,

$$\sum_i \Lambda_i(T_i) = \int_0^\infty W(t)\lambda_0(t)\mathrm{d}t.$$

[3] Søren Johansen. An extension of Cox's regression model. *International Statistical Review*, 51(2):165–174, 1983

The likelihood can therefore be factorized as

$$\begin{aligned} L &= \prod_i \lambda(T_i)^{e_i} e^{-\Lambda_i(T_i)} \\ &= \left(\prod_{i:e_i=1} (w_i \lambda_0(T_i))^{e_i} \right) e^{-\int_0^\infty W(t)\lambda_0(t)\mathrm{d}t} \\ &= \prod_{i:e_i=1} \frac{w_i}{W_i} \left(\prod_i (W_i \lambda_0(T_i))^{e_i} \right) e^{-\int_0^\infty W(t)\lambda_0(t)\mathrm{d}t}. \end{aligned}$$

□

Definition 6.15. *Cox's partial likelihood is*

$$L_{par} = \prod_{i:e_i=1} \frac{w_i}{W_i}.$$

It is possible, on the basis of heuristic arguments, to justify that all information about parameters that enter in the weights is contained in L_{par}, and that L_{par} can sensibly be regarded as a profile likelihood. We will instead turn to the Poisson nonparametric likelihood and present a rigorous profiling argument. What is particularly nice about this derivation of the partial likelihood as a nonparametric profile likelihood is that we directly obtain an estimator of the cumulative baseline hazard in terms of the estimates of the weights. The weights are in turn estimated by optimizing Cox's partial likelihood numerically over a suitable parameter space.

The Poisson log-likelihood for the proportional hazards model is

$$\ell^* = \sum_i e_i \log(\lambda_0(T_i)) + e_i \log(w_i) - \Lambda_0(T_i) w_i$$

in terms of the linear predictor and the (nonparametric) baseline hazard λ_0.

We consider jump cumulative hazard functions of the form

$$\Lambda_0(t) = \sum_{j:T_j \leq t} e^{h_j},$$

that is, $\log(\lambda_0(T_j)) = h_j$. The resulting log-likelihood is

$$\ell^* = \sum_i e_i h_i + e_i \log(w_i) - w_i \sum_{j:T_j \leq T_i} e^{h_j}.$$

Interchanging the sums in the log-likelihood we get

$$\ell^* = \sum_i e_i h_i + e_i \log(w_i) - e^{h_i} W_i.$$

Lemma 6.16. *The maximizer of ℓ^* over $h_1, \ldots, h_n$ is given by*

$$e^{h_i} = \frac{e_i}{W_i}$$

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

Proof. The parameters h_i are variation independent, and we can thus maximize the function by maximizing each term. Differentiating we find that

$$\partial_{h_i} \ell^* = e_i - e^{h_i} W_i.$$

If $e_i = 0$ the function is monotonely decreasing and maximized in $-\infty$ corresponding to $e^{h_i} = 0$. If $e_i = 1$ we equate the derivative equal to 0 and find the claimed solution to be the unique stationary point. An additional differentiation shows that the second derivative is

$$-e^{h_i} W_i < 0,$$

and this implies that the solution is a global maximum. □

With all weights being 1 in the lemma above, the estimator is again recognized as the Nelson-Aalen estimator. This shows that the Poisson likelihood L^* gives the same NPMLE as L. Plugging the maximizer into the log-likelihood gives

$$\ell^* = \underbrace{\sum_{i:e_i=1} \log\left(\frac{w_i}{W_i}\right)}_{\text{Cox's partial log-likelihood}} - N$$

where $N = \sum_i e_i$ is the total number of deaths.

One has to pay attention to ties in the data. The above derivations are made under the assumption that all the observed survival times are different.

The discrete proportional hazards model

It is legitimate to consider the proportional hazards model for discrete probability distributions, and use the ordinary nonparametric likelihood function L for profiling. Contrary to the derivations above using the Poisson likelihood, the maximizer is not explicit. Moreover, the proportionality assumption for the discrete model does not really provide an approximation of the proportionality assumption for the continuous model.

The appropriate generalization of the proportional hazards model to the discrete case is actually the model

$$\lambda_i = 1 - (1 - \lambda_{0,i})^{w_i}. \tag{6.3}$$

The nonparametric likelihood becomes

$$L = \prod_i (1 - (1 - \lambda_{0,i})^{w_i})^{e_i} (1 - \lambda_{0,i})^{W_i - w_i e_i},$$

and similar arguments as in Theorem 6.12 will show that the NPMLE of $\lambda_{0,i}$ for given weights is

$$\hat{\lambda}_{0,i} = e_i \left(1 - \left(1 - \frac{w_i}{W_i}\right)^{1/w_i}\right).$$

The corresponding profile likelihood is

$$\prod_{i:e_i=1} \frac{w_i}{W_i} \left(1 - \frac{w_i}{W_i}\right)^{W_i/w_i - e_i}. \tag{6.4}$$

The discrete proportional hazards model (6.3) is arguably the "correct" generalization of the continuous proportional hazards model[4], and using the ordinary nonparametric likelihood yields the profile likelihood in (6.4). The resulting profile likelihood is not Cox's partial likelihood, but it is for most practical purposes very close. In the literature, Cox's partial likelihood has caught on as the default choice for semiparametric estimation of parameters that enter the weights.

Survival Residuals

Model fit can be assessed for survival regression models just as for generalized linear models by means of residuals. There are, however, several different competing choices of residuals.

[4] Martin Jacobsen. Maximum likelihood estimation in the multiplicative intensity model: A survey. *International Statistical Review*, 52(2):193–207, 1984

Cox-Snell residuals are not really residuals in a classical sense, but a transformation of the observed survival times to a set of random variables whose distribution should be exponential if the model is correct. They are based on the following result.

Lemma 6.17. *If T has continuous cumulative hazards function Λ then $\Lambda(T)$ is exponentially distributed with rate 1.*

Proof. The survival function is $S(t) = \exp(-\Lambda(t))$ is continuous, whence

$$S(T) \sim \text{unif}([0,1]).$$

From this it follows that

$$P(\Lambda(T) \leq t) = P(S(t) > \exp(-t)) = 1 - \ exp(-t)$$

for $t \geq 0$, or $\Lambda(T) \sim \text{Exp}(1)$. □

From the lemma above we conclude that if $(T_1, e_1), \ldots, (T_n, e_n)$ are right censored (independent censoring) survival times with corresponding survival functions $S_1, \ldots, S_n$, then

$$(\Lambda_1(T_1), e_1), \ldots, (\Lambda_n(T_n), e_n)$$

are (independently) right censored exponentially distributed variables. We use this in practice as follows. If $\hat{\Lambda}_i$ is the fitted model of the cumulative hazards function for the i'th survival time we fit a nonparametric cumulative hazard function to the *Cox-Snell residuals*

$$\hat{\Lambda}_i(T_i).$$

This can be done using the Nelson-Aalen estimator. The resulting fitted cumulative hazards function is plotted against time. This should be a straight line with slope 1 (the cumulative hazard function for the exponential distribution).

The martingale residuals are defined as

$$e_i - \hat{\Lambda}_i(T_i),$$

and are thus the status indicator minus the Cox-Snell residual. Their usefulness is partly justified by the following result.

Lemma 6.18. *If $T = \min\{T^*, C\}$ with T^* and C having continuously differentiable survival functions it holds that*

$$E\Lambda(T) = P(T^* \leq C).$$

Proof. Let S and H denote the distribution functions of T^* and C, respectively. By partial integration

$$\begin{aligned} E\Lambda(T) &= -\int_0^\infty \Lambda(t)(SH)'(t)\mathrm{d}t = \int_0^\infty \Lambda'(t)S(t)H(t)\mathrm{d}t \\ &= \int_0^\infty f(t)H(t)\mathrm{d}t = P(T^* \leq C). \end{aligned}$$

□

The survival function for the minimum of the two independent random variables T^* and C is the product SH, whence the density of their distribution is the negative derivative $-(SH)'$.

It follows from the lemma above that with $e = 1(T^* \leq C)$ then

$$E(e - \Lambda(T)) = 0.$$

Thus if $\hat{\Lambda}_i$ were the true cumulative hazards function, the martingale residuals would have mean zero. Now $\hat{\Lambda}_i$ is only an estimate, but if the model fits the data, the martingale residuals should have approximately mean 0. In the more advanced theory of survival analysis they also enjoy a martingale property, which explains their name. Here we will not pursue this, but just observe that we can investigate the model fit by investigating if the martingale residuals approximately have mean 0. The residuals can be used much as regression residuals to judge overall fit as well as the fit of how the individual predictor variables enter into the model. It should be noted, though, that they by definition have a left skewed distribution concentrated on $(-\infty, 1)$.

The martingale residuals do *not* approximately follow a normal distribution!

Alternative residuals include the deviance residuals and the Schoenfeld residuals. We will not treat the latter, but the former are based on the following observation. The Poisson nonparametric log-likelihood as well as the parametric log-likelihood with a continuous baseline can be written as

$$\ell^* = \sum_{i=1}^n \underbrace{e_i \log(w_i \Lambda_{0,i}) - w_i \Lambda_{0,i}}_{\text{Poisson log-likelihood term}} + \sum_{i=1}^n e_i \log\left(\frac{\lambda_{0,i}}{\Lambda_{0,i}}\right)$$

The first term is identical to the Poisson log-likelihood, and the second term does not depend on parameters that enter into the weights. Based on this we define the deviance residuals in terms of the Poisson deviances

$$d(e_i, \hat{\Lambda}_i) = 2\left(e_i \log(e_i/\hat{\Lambda}_i) - e_i + \hat{\Lambda}_i\right),$$

where $\hat{\Lambda}_i = \hat{w}_i \hat{\Lambda}_{0,i}$. The result is the *deviance residuals*

$$\operatorname{sign}(e_i - \hat{\Lambda}_i)\sqrt{d(e_i, \hat{\Lambda}_i)} = \begin{cases} -\sqrt{2\hat{\Lambda}_i} & \text{if } e_i = 0 \\ \operatorname{sign}(1 - \hat{\Lambda}_i)\sqrt{2(\hat{\Lambda}_i - 1 - \log \hat{\Lambda}_i)} & \text{if } e_i = 1 \end{cases}$$

Observe that the deviance residuals can be expressed as

$$\operatorname{sign}(r_i)\sqrt{-2(\hat{r}_i + e_i \log(e_i - \hat{r}_i))}$$

with $\hat{r}_i = e_i - \hat{\Lambda}_i$ the martingale residuals. Compared to the martingale residuals, the distribution of the deviance residuals is less skewed.

Prostate cancer survival

We reconsider a data set from Chapter 20 in Frank Harrell's book *Regression modeling strategies*[5].

```
prostate <- read.table("prostate.txt",
                       header = TRUE,
                       colClasses =
                         c("factor", "factor", "factor", "numeric",
                           "factor", "numeric", "numeric", "factor",
                           "factor", "numeric", "numeric", "factor",
                           "numeric", "numeric", "numeric", "numeric",
                           "factor", "character")
)
```

The data comes from a randomized clinical trial on prostate cancer patients with 4 different treatments (one placebo and three different dosages of estrogen). To document differences between treatment groups in the survival distribution we do not need to take other variables into account. As discussed elsewhere, the purpose of the randomization is to make treatment independent of other predictor variables (observed and unobserved), thus allowing us to ascribe differences between the treatment groups to the treatment alone. If we want to predict the survival distribution of individual patients (or subpopulations) we can take other variables into account. That is, for prognostic purposes other predictor variables can make the prediction more accurate.

[5] Frank Harrell. *Regression Modeling Strategies*, Springer-Verlag New York, Inc., 2010

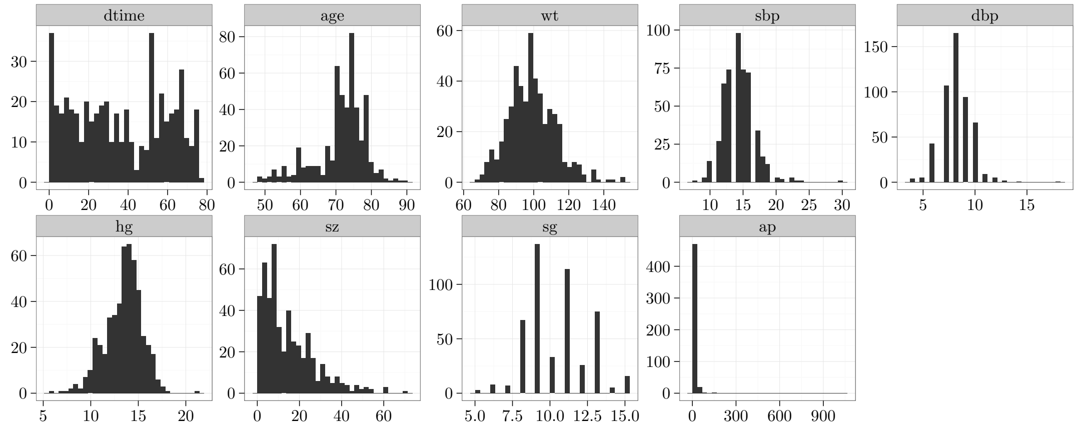


Figure 6.1: Histograms for continuous predictor variables.

Descriptive statistics

The data set contains the following variables:

```
names(prostate)

##  [1] "patno"  "stage"  "rx"     "dtime"  "status" "age"    "wt"
##  [8] "pf"     "hx"     "sbp"    "dbp"    "ekg"    "hg"     "sz"
## [15] "sg"     "ap"     "bm"     "sdate"

## For later usage
conVar <- c("dtime", "age", "wt", "sbp", "dbp","hg", "sz", "sg", "ap")
disVar <- c( "stage", "rx", "status", "pf", "hx", "ekg", "bm")
```

Then we look at the marginal distributions of the variables in terms of histograms and barplots.

```
meltedProstate <- melt(prostate[, conVar])
qplot(value, data = meltedProstate, geom = "histogram",
      xlab = "", ylab = "") +
  facet_wrap(˜ variable, scales = "free", ncol = 5)
```

```
meltedProstate <- melt(prostate[, disVar], id.vars = c())
qplot(value, data = meltedProstate, geom = "bar",
      xlab = "", ylab = "") +
  facet_wrap(˜ variable, scales = "free", ncol = 4) +
  theme(axis.text.x = element_text(angle = -30,
         size = 8, hjust = 0, vjust = 1))
```

Some noteworthy observations are that `ap` has a very skewed and nonuniform distribution, which suggest that a log-transform is

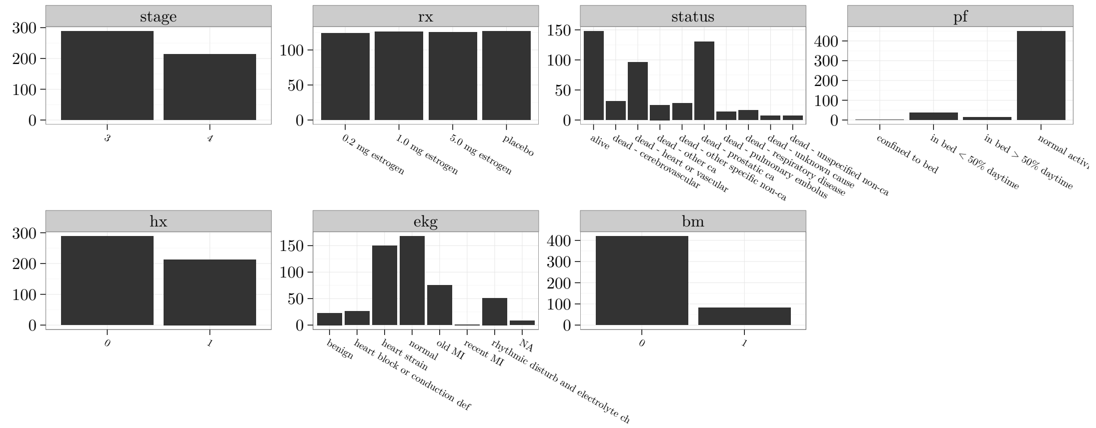


Figure 6.2: Barplots for categorical predictor variables.

beneficial. We note that `pf` has a very asymmetric distribution with only a small number of patients having a nonnormal activity. We should also observe that 148 out of the total of 502 individuals, that is, approximately one third, have censored survival times (are still alive).

We also consider the scatter plot matrix of the continuous variables.

```
splom(na.omit(prostate[, conVar]),
      upper.panel = panel.hexbinplot,
      pscale = 0,
      varname.cex = 0.7,
      nbins = 15,
      lower.panel = function(x, y) {
        panel.text(mean(range(x)), mean(range(y)),
                   round(cor(x, y), digits = 2),
                   cex = 0.7
                   )
      }
      )
```

What is most notable is that the two measures of blood pressure are (not surprisingly) very correlated. We let their mean enter into the model below. They are both also correlated with weight, which in turn is also correlated with `hg`. The indicator `sg` shows some correlation with the size variable `sz` as well. Besides the correlation between systolic and diastolic blood pressures, there is no evidence

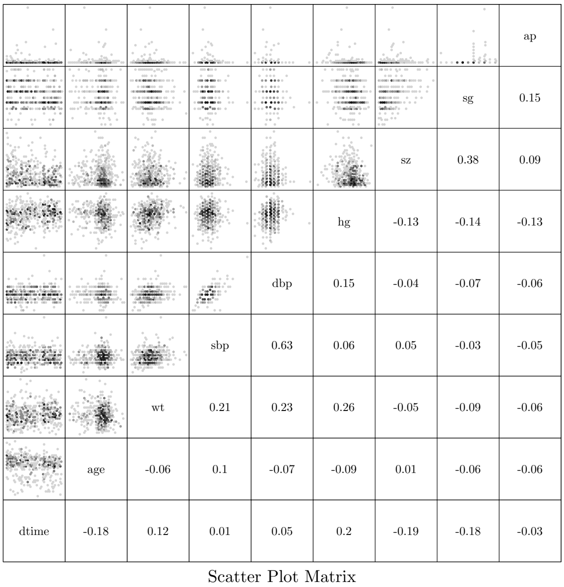


Figure 6.3: Scatter plot matrix and Pearson correlations for the continuous variables.

of strong collinearity.

We will for now just drop the missing observations (Harrell gives a treatment of imputation for this particular data set in his Chapter 8).

```
### The number of missing values for each variable.

sapply(prostate, function(x) sum(is.na(x)))

##   patno  stage     rx  dtime status    age     wt     pf     hx    sbp
##       0      0      0      0      0      1      2      0      0      0
##     dbp    ekg     hg     sz     sg     ap     bm  sdate
##       0      8      0      5     11      0      0      0

### Dropping observations with missing variables,
### and other data cleaning / manipulations.

subProstate <-
  transform(na.omit(prostate),
            bp = (sbp + dbp) / 2,      ## Average blood presure.
            logap = log(ap)
  )
conVar <- c(conVar, "logap", "bp")
```

To focus on the treatment we turn to a Kaplan-Meier estimator for each of the four treatment groups.

```
prostateSurv <- survfit(Surv(dtime, status != "alive") ~ rx,
                        data = subProstate)
plot(prostateSurv, mark.time = FALSE, conf.int = FALSE,
     col = c("red", "blue", "purple", "cyan"))
legend(30, 1.08, levels(subProstate$rx),
       col = c("red", "blue", "purple", "cyan"),
       lty = 1, bty = "n")
```

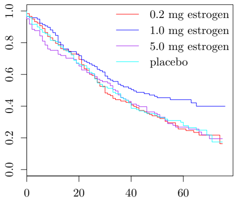


Figure 6.4: Kaplan-Meier estimators of the survival functions within each of the four treatment groups.

The plot suggests that individuals treated with 1.0 mg estrogen have a longer survival time than the other three groups. Adding confidence bands (which will mesh up the reading of the figure in general) shows that the difference is borderline significant.

Cox model

We fit Cox's proportional hazards model using the `coxph` function from the Survival package. Then we construct tests for each of the variables in the full model.

```
form <- Surv(dtime, status != "alive") ~ rx + age + wt + pf +
  hx + ekg + hg + sz + sg + logap + bp + bm
prostateCox <- coxph(form, data = subProstate)
testtab <- drop1(prostateCox, test = "Chisq")
ord <- order(testtab[, 4][-1]) + 1
testtab[ord, ]
```

Table 6.1: Likelihood ratio tests of excluding one predictor variable in the proportional hazards model with the variables entering linearly. The variables are ordered according to the *p*-value of the test.

	Df	LRT	Pr(>Chi)
hx	1	16.52	0.0000
sz	1	11.37	0.0007
age	1	7.21	0.0073
hg	1	4.16	0.0414
sg	1	3.96	0.0465
wt	1	3.83	0.0502
ekg	6	11.70	0.0690
rx	3	5.77	0.1235
bm	1	1.58	0.2091
pf	3	3.50	0.3210
bp	1	0.08	0.7829
logap	1	0.02	0.8744

The tests suggest that the difference between the four treatment groups can be explained by other variables. In particular `hx` (history of cardiovascular disease), tumor size and age. In fact, these findings suggest that a more appropriate model is one with competing risks. That is, a model where we model the risk of dying from the cancer as well as dying from a heart attack or a similar non-cancer related cause.

Baseline estimation is achieved from the profile likelihood argument, which in the derivation of the partial likelihood gave a nonparametric estimator the cumulative baseline hazard function as well.

```
w <- predict(prostateCox, type = "risk")  ## Individual weights
orddtime <- order(subProstate$dtime)
stat <- (subProstate$status != "alive")[orddtime]
W <- rev(cumsum(w[rev(orddtime)]))
Lambda <- cumsum(stat / W)
qplot(subProstate$dtime[orddtime], exp(-Lambda), geom = "step") +
  ylab("Survival function") + xlab("time (months)") + ylim(c(0,1))
```

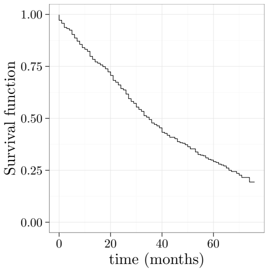


Figure 6.5: The nonparametric estimate of the baseline survival function.

The computations can be carried out using the `survfit` function on the `coxph` object as well, but it is useful to understand how the estimator arise from the profile method. The explicit R computations above show that.

```
plot(survfit(prostateCox), mark.time = FALSE)
```

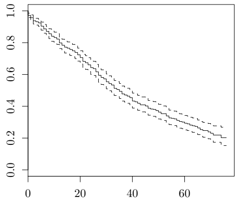


Figure 6.6: The `plot` method for a `survfit` object plots the nonparametric estimate of the baseline survival function including a pointwise 95% confidence band.

Residuals such as the Cox-Snell residuals can be used to reveal overall lack of fit. The Cox-Snell residuals are based on the result that the transform of the survival times by the cumulative hazard function is exponentially distributed for continuous survival distributions and with independent censoring.

The Cox-Snell residuals are just the individual estimated cumulative hazards in the corresponding censored survival times. The Nelson-Aalen estimate of the cumulative hazards function for the Cox-Snell residuals plotted against the residuals should be on a straight line.

```
CSres <- Lambda * w[orddtime] ## Cox-Snell residuals
ordres <- order(CSres)
CSres <- CSres[ordres]
statres <- stat[ordres] ## Indicator of non-censoring
## Computation of cumulative hazards using 'survfit'.
```

```
tmp <- survfit(Surv(CSres, statres) ~ 1, type = "fleming-harrington")
CumHaz <- -log(tmp$surv)[statres]
## Alternative direct computation of cumulative hazards
## Y <- seq(length(Lambda), 1)[statres]
## CumHaz <- cumsum(1/Y)
## Nelson-Aalen estimator of exponential cumulative hazard
qplot(CSres[statres], CumHaz) +
  geom_abline(intercept = 0, slope = 1) +
  geom_point(color = "red") +
  ylab("Cum. Hazards") + xlab("Cox-Snell residuals")
```

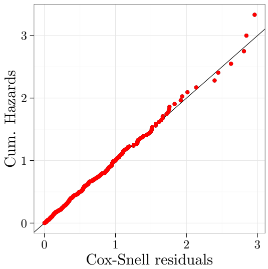


Figure 6.7: Cox-Snell residual plot.

The Nelson-Aalen estimate was computed using the **survfit** function. It can also be computed "by hand" as illustrated in the comments to the R code above. Note the specification of the **type** argument in the call of **survfit**. The **"fleming-harrington"** value gives an estimate of the survival function based on the Nelson-Aalen estimator, so that $-\log \hat{S}$ is identical to the Nelson-Aalen estimator. This is the way to compute the Nelson-Aalen estimator using the survival package. The default **type** argument is **"kaplan-meier"**, which of course gives the Kaplan-Meier estimator of the survival function. In this case, $-\log \hat{S}$ will differ slightly from the Nelson-Aalen estimator.

The Cox-Snell residuals are mostly useful for overall assessment of model fit, and not so useful for detecting how the model might fail to fit. In this respect it resembles the qq-plot. For linear models, a heterogeneous variance or a lack of fit in the mean value specification can result in a a deviation from Gaussianity in a qq-plot, but the qq-plot will not show clearly where the lack of model fit comes from.

Martingale or deviance residuals for survial models are more similar to regression residuals from linear and generalized linear models, and they are useful for investigations of the model fit, for instance if the predictors enter correctly into the model.

```
prostateData <- subProstate[, c("patno", all.vars(form)[-c(1, 2)])]
## Change the 'type' argument below to 'deviance' for deviance residu-
als
predProstate <-
  cbind(prostateData,
        data.frame(mgres = residuals(prostateCox,
                                     type = 'martingale'))
  )

meltedPredProstate <-
  melt(predProstate,
```

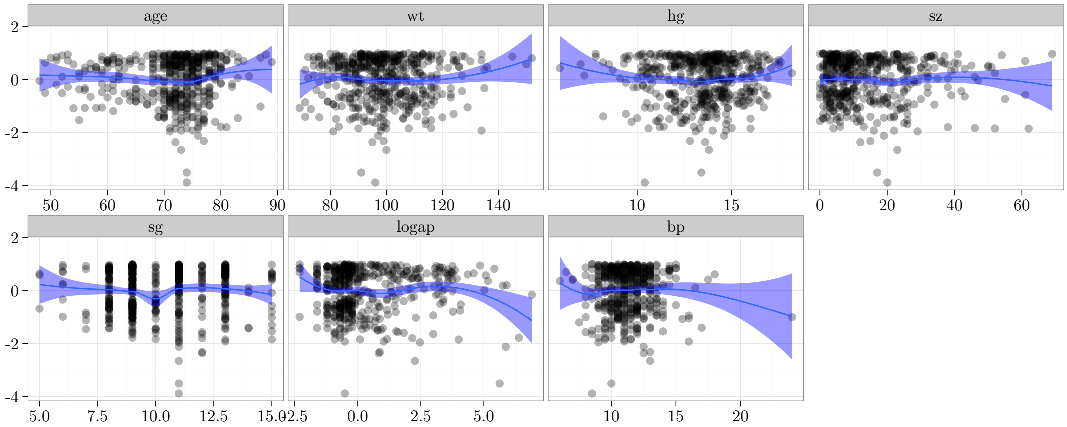


Figure 6.8: Martingale residuals plotted against predictor variables. The residuals have a skewed distribution but zero mean if the model is correct. The smoothed line can assist in the detection of variables with nonlinear effects.

```
        id.vars = c("patno", "mgres"),
        measure.vars = conVar[conVar %in% all.vars(form)[-c(1, 2)]])

qplot(value, mgres, data = meltedPredProstate,
       xlab = "", ylab = "", geom = "point",
       alpha = I(0.3), size = I(3)) +
  facet_wrap(~ variable, scales = "free_x", ncol = 4) +
  geom_smooth(size = 1, fill = "blue")
```

These plot do not suggest obvious misfits, but we attempt a model with nonlinear effects of all continuous predictors anyway.

```
form <- Surv(dtime, status != "alive") ~ rx + ns(age, 4) + ns(wt, 4) +
  + pf + hx + ekg +  ns(hg, 4) + ns(sz, 4) +
  ns(sg, 4) + ns(logap, 4) + ns(bp, 4) + bm
prostateCox2 <- coxph(form, data = subProstate)
drop1(prostateCox2, test = "Chisq")[ord, ]
anova(prostateCox2, prostateCox)
```

loglik	Chisq	Df	P(>|Chi|)
−1845.17	36.15	21	0.0210

Table 6.2: Likelihood ratio test of the linear model against the nonlinear model.

The `anova` call above for the `coxph` objects gives in this case a test of the model with only linear effects against this larger model including nonlinear effects, see Table 6.2. It is borderline significant showing that there is some, but not a lot, evidence in the data for nonlinear relations.

Table 6.3: Likelihood ratio tests of the individual variables in the nonlinear model.

	Df	LRT	Pr(>Chi)
hx	1	16.71	0.0000
ns(sz, 4)	4	14.00	0.0073
ns(age, 4)	4	14.88	0.0050
ns(hg, 4)	4	9.54	0.0489
ns(sg, 4)	4	4.08	0.3958
ns(wt, 4)	4	8.87	0.0644
ekg	6	14.27	0.0268
rx	3	4.31	0.2295
bm	1	0.30	0.5857
pf	3	6.68	0.0828
ns(bp, 4)	4	2.91	0.5731
ns(logap, 4)	4	7.68	0.1039

Table 6.3 shows the likelihood ratio tests for excluding each of the variables from the full nonlinear model. Compare with Table 6.1 where the variables enter linearly in the model.

THE PROPORTIONAL HAZARDS ASSUMPTION can be investigated by excluding a variable and visually inspect different baseline estimates for the variable. This works well for factor variables with few levels.

To illustrate the method, we first return to the model where we only include treatment. But now we fit the proportional hazards model and consider the different estimated parameters for the treatments.

```
prostateCox0 <- coxph(Surv(dtime, status != "alive") ~ rx,
                      data = subProstate)
summary(prostateCox0)
```

Table 6.4: Estimated coefficients for the proportional hazards model including only treatment as a predictor.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
rx1.0 mg estrogen	−0.40	0.67	0.16	−2.50	0.01
rx5.0 mg estrogen	−0.01	0.99	0.15	−0.03	0.97
rxplacebo	−0.03	0.97	0.15	−0.21	0.83

This shows, in concordance with the findings above, that the group given estrogen treatment with a dosage of 1.0 mg have a decreased risk of dying (the hazard rate is lowered), and that this effect is borderline significant. However, what about the proportional hazards assumption? We check that by checking if the Nelson-Aalen estimates of cumulative hazard functions within each of the four treatment groups are, in fact, proportional.

```
## The cloglog transform gives the log cumulative hazards plotted
## against the log times
plot(prostateSurv, mark.time = FALSE, conf.int = FALSE,
     col = c("red", "blue", "purple", "cyan"),
     fun = "cloglog")
```

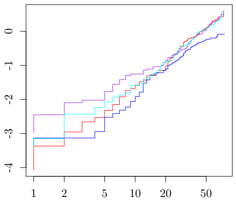


Figure 6.9: Diagnostic plot for checking the proportional hazards assumption for the treatment without other predictors.

The proportional hazards assumption is not perfect. If the proportional hazards assumption were true, the four curves should have been roughly parallel. If we have a model with more predictors, we can investigate a single one of them using a similar technique. We fit the model using all other predictors and then nonparametrically estimate the baseline for each value of the factor. This is done below for the treatment variable.

```
prostateCox2a <- update(prostateCox2, . ~ . - rx)
w <- predict(prostateCox2a, type = "risk")  ## Individual weights
orddtime <- order(subProstate$dtime)
stat <- (subProstate$status != "alive")[orddtime]
rxord <- subProstate$rx[orddtime]
W <- tapply(w[orddtime], rxord, function(ww) rev(cumsum(rev(ww))))
statstrat <- tapply(stat, rxord, function(x) x)
tmp <- list()
for(i in 1:4) {
  tmp[[i]] <- data.frame(
    logLambda = log(cumsum(statstrat[[i]] / W[[i]])),
    time = subProstate$dtime[orddtime][as.numeric(rxord) == i],
    rx = names(statstrat)[i]
  )
}

tmp <- do.call(rbind, tmp)
ggplot(tmp, aes(time, logLambda, color = rx)) +
  geom_line() +
  scale_color_manual("",
                     values = c("red", "blue", "purple", "cyan")) +
  scale_x_log10(breaks = c(1, 2, 5, 10, 20, 50)) +
  xlab("log time") + ylab("log cum. hazards") +
  theme(legend.position = c(1, 0),
        legend.justification = c(1, 0))
```

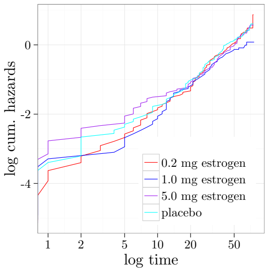


Figure 6.10: Diagnostic plot for checking the proportional hazards assumption for the treatment with the other predictors included.

The conclusion is the same – not surprisingly, because treatment is by randomization independent of the remaining predictors.

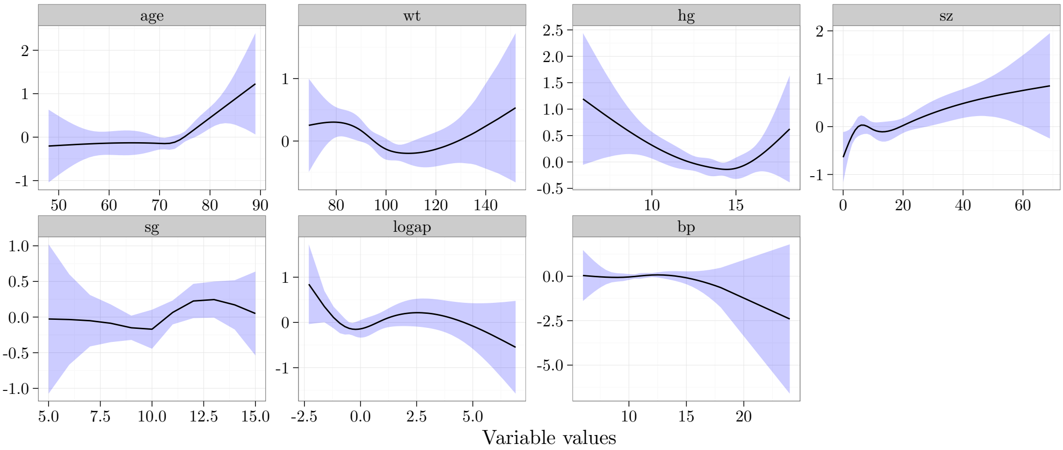


Figure 6.11: Estimates and 95% confidence bands for the nonlinear effects of the continuous variables.

Reporting the model

As always, we should try to investigate and present the fitted model. Here we give curves with confidence bands of the fitted terms for each continuous variable and fitted parameters with confidence intervals for each parameter associated with a factor variable.

```r
predictProstate <- predict(prostateCox2, type = "terms", se.fit = TRUE)
predictData <- melt(subProstate[, c("patno", conVar[-c(1, 4, 5, 9)])],
                    id.vars = "patno")

selectedTerms <- which(all.vars(form)[- c(1, 2)] %in% conVar)

plotData <-
  cbind(predictData,
        se = melt(predictProstate$se.fit[, selectedTerms])$value,
        y = melt(predictProstate$fit[, selectedTerms])$value
  )

ggplot(plotData, aes(x = value), xlab = "", ylab = "") +
  geom_ribbon(aes(ymin = y - 2*se, ymax = y + 2*se),
              alpha = I(0.2),
              fill = I("blue")) +
  geom_line(aes(y = y), size = 1) +
  facet_wrap(~ variable, ncol = 4, scale = "free") +
  ylab("") + xlab("Variable values")
```

We observe that the effect of the variables `age` and `sz` (and perhaps also `hg`) show some nonlinearity.

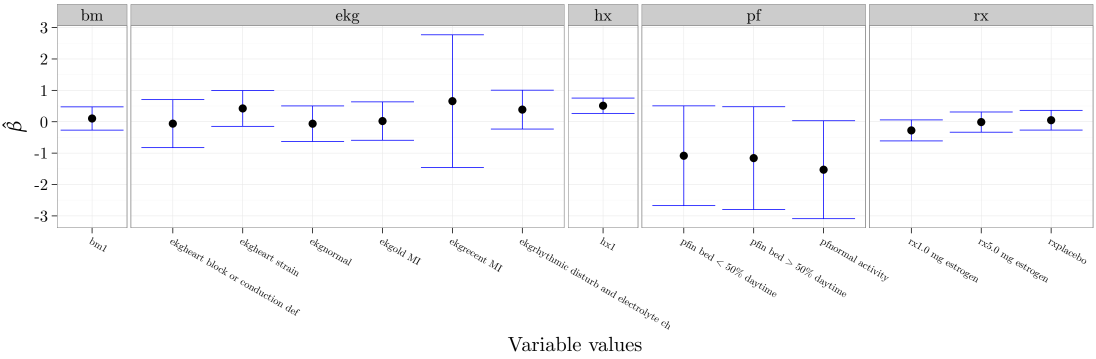


Figure 6.12: Estimates and 95% confidence intervals of the parameters related to the factor variables.

```
plotData <- cbind(
  as.data.frame(confint(prostateCox2, c(1:3, 12:21, 42))),
  coef(prostateCox2)[ c(1:3, 12:21, 42)])
names(plotData) <- c("low", "up", "hat")
plotData$variable <-
  c(rep("rx", 3), rep("pf", 3), "hx", rep("ekg", 6), "bm")
plotData$value <- rownames(plotData)

ggplot(plotData, aes(x = value, xlab = "", ylab = "")) +
  geom_errorbar(aes(ymin = low, ymax = up),
                colour = I("blue")) +
  geom_point(aes(y = hat), size = 3) +
  facet_grid(.~ variable, scales = "free_x", space = "free_x") +
  theme(axis.text.x = element_text(angle = -30,
        size = 8, hjust = 0, vjust = 1)) +
  ylab("$\\hat{\\beta}$") + xlab("Variable values")
```

The conclusion from the analysis is that `age`, the size of the primary tumor (`sz`) and the indicator of a history of cardiovascular disease (`hx`) are the three variables that most predictive of the survival time. The former two enter nonlinearly in the proportional hazards model.

Exercises

Exercise 6.1. Show that if T has a log-logistic distribution with density $f = f_0$ then

$$e^{\eta} T$$

has density f_η. Show that if T has a log-logistic distribution with density f_η then

$$\log T - \eta$$

has a logistic distribution. ○

Bibliography

George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for experimenters*, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2005.

B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

Frank Harrell. *Regression Modeling Strategies*, Springer-Verlag New York, Inc., 2010.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, Springer, New York, 2009.

Martin Jacobsen. Maximum likelihood estimation in the multiplicative intensity model: A survey. *International Statistical Review*, 52(2):193–207, 1984.

Søren Johansen. An extension of Cox's regression model. *International Statistical Review*, 51(2):165–174, 1983.

Jørn Olsen, Mads Melbye, Sjurdur F. Olsen, et al. The Danish National Birth Cohort - its background, structure and aim. *Scandinavian Journal of Public Health*, 29(4):300–307, 2001.

Art Owen. *Empirical likelihood*, Chapman & Hall/CRC, 2001.

Judea Pearl. *Causality*, Cambridge University Press, Cambridge, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013a.

R Core Team. *Writing R Extensions.* R Foundation for Statistical Computing, Vienna, Austria, 2013b.

Richard M. Royall. *Statistical evidence*, Chapman & Hall, London, 1997.

W. N. VENABLES and B. D. RIPLEY. *Modern Applied Statistics with S*, Springer, New York, 2002.

HADLEY WICKHAM. *ggplot2: elegant graphics for data analysis*, Springer New York, 2009.

HADLEY WICKHAM. *Advanced R*, Chapman Hall/CRC, 2014.

ALLEN J WILCOX. On the importance—and the unimportance—of birthweight. *International Journal of Epidemiology*, 30(6):1233–1241, 2001.

LELAND WILKINSON. *The grammar of graphics*, Springer, New York, 2005.