NIELS RICHARD HANSEN

REGRESSION

WITH R

UNIVERSITY OF COPENHAGEN

Preface

This book was written as the textbook material for a graduate statistics course in regression analysis. The prerequisites include acquaintance with standard statistical methodology, such as ordinary least squares linear regression methods, including t-tests and constructions of standard confidence intervals, and standard distributions such as the univariate and multivariate normal distributions. The reader will also benefit from 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 rather than formal theoretical arguments. That is, mathematics is used to make model descriptions and assumptions precise, and to analyze practical estimation problems and computations required for carrying out data analysis. Less attention is paid to mathematical justifications of methods, e.g. bounds on the estimation error, theoretical optimality results or formal asymptotic arguments.

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 all aspects of the data analysis from reading data into R, cleaning and plotting data to data analysis and model diagnostics.

Contents

1 Introduction

The purpose of statistical modeling

This book is primarily on *predictive regression modeling*. That is, the viewpoint is that the main purpose of a model is to be predictive. There is no claim that this is the only purpose of modeling in general, but it is arguably important. The topics chosen and the treatment given owe a lot to two other books in particular. The book *Regression Modeling Strategies*[1] was an important inspiration, and is an excellent supplement – this book being more mathematical. The other book is *The Elements of Statistical Learning*[2], which offers a plethora of predictive models and methods. The present book is far less ambitious with a narrower focus on fundamental regression models and modeling strategies – the aim is to be more detailed. Indeed, the book can be seen as providing the statistical foundation for *The Elements of Statistical Learning* as well as the literature on predictive regression modeling and machine learning in general.

In predictive modeling it is fairly clear how to compare models. The only thing required is a quantification of predictive accuracy, and the best model is then the most accurate model. The accuracy of a prediction is typically quantified by a *loss* function, which actu-

[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

ally quantifies how *inaccurate* the prediction is. Thus, the smaller the loss is the more accurate is the model. The specification of a relevant loss function is, perhaps, not always easy. A good loss function should ideally reflect the consequences of wrong predictions. There is, on the other hand, 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 biggest challenge 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 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, which is a situation we want to avoid.

Generalization is good, overfitting is bad.

A prediction model does not need to explain the underlying mechanisms of how observed variables are related. This may be a point of criticism. What good is the model if we can't interpret it – if it is just a black box producing predictions? Sometimes a black box is completely adequate. Nobody really cares about the mechanisms behind spam emails[3], but we care a lot about the performance of our spam email filter. On the other hand, it is well known that education level is a strong predictor of income, but are we ever interested in predicting income based on education level? We are 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 we require that if we sample a random individual and predict her income based on her education level we get an accurate prediction. However, 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 case both variables may be partly determined by 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

[3] Perhaps except those that design spam filters.

the subject matter field to which the model is applied.

Causal modeling is an important, interesting and active research field. Judea Pearl's *Causality*[4] book has been exceptionally influential on the development of the field. One main difference from predictive modeling is that causal modeling is concerned with prediction of intervention effects. Predictions are thus important in causal modeling as well, but the predictions are to be made in a setup that differs from the setup we have data from. We will not pursue causal modeling in any systematic way, but we will bring up causal interpretations and predictions when relevant, and we will discuss which assumptions we are making to allow for a causal interpretation of a predictive model. It is necessary to warn against causal misinterpretations of predictive models. A regression coefficient does not generally represent an effect. A phrase like[5] "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[6], 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 is that it quantifies a difference between subpopulations – and not the effect of moving individuals from one subpopulation to another. However, the documentation that such differences exist, and the estimation of their magnitude, are important contributions to the understanding of causal relations. It is, however, 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. In particular, if the data was obtained from an observational study. By contrast, in a randomized trial the purpose of the randomization is to break all relations between the response and unobserved variables, so that observed differences can be ascribed to the variation of (controlled) predictors, e.g. a treatment, and thus be given a

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

[5] See the birth weight case study, p. 16.

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

causal interpretation.

With these words of warning and reminders of carefulness in making causal interpretations of predictive models, we should again 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.

Case studies

The book consists of theory sections interspersed by real data modeling and data analysis. A decision was made that instead of providing small simple data examples to illustrate a point, the relevant points are illustrated by real case studies. The hope is that this will ease the transition from theory to practice. The price to pay is that there are constant distractions in forms of real data problems. Data never behaves well. There are missing observations and outliers, the model does 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.

R

We use the programming language R[7] 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[8] 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 adap-

[7] www.r-project.org

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

tation 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`[9]. 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[10] 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[11] developed by Hadley Wickham. 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[12]. Occasionally, an alternative package, `lattice`, was used. There exists a nice series of blogs[13] 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*[14] and the corresponding `MASS` package (comes with the R distribution). Many classical statistical 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[15] is an important resource. Another splendid resource is the book *Advanced R development: mak-*

[9] `yihui.name/knitr/`

[10] `www.rstudio.com`

[11] `ggplot2.org`

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

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

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

[15] R CORE TEAM. *Writing R Extensions*. R Foundation for Statistical Computing, Vienna, Austria, 2013b

ing reusable code by Hadley Wickham. It is at the time of writing a book in progress, but it is fortunately available as a wiki[16] – and will continue to be so after publication. 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.

```
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 :-)
```

[16] github.com/hadley/devtools/wiki

2 The linear model

The fundamental linear model

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 equality

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

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

[1] The X goes by many 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.

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.

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,

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

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

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

Assumption A2 is often made and also often needed[3], but it is perhaps not obvious why. It is first of all 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 precision 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[4],

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

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

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. The fractionation of the linear model in the literature into different submodels has resulted in special terminology for special cases, which is unnecessary, and most likely a consequence of historically different needs in different areas of applications. A unified treatment is preferable to understand that, in reality, the linear model 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.

We need to introduce a couple of additional distributional assumptions. These are assumptions on the joint distribution of multiple observations. If we have n observations, $Y_1, \ldots, Y_n$, of the response with corresponding predictors $X_1, \ldots, X_n$. 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 $\mathbf{X}$.

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. Assumptions A4 imply 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

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

In summary, there are two sets of distributional assumptions. The weak set A1, A2 and A4, which imply the moment identities (2.2)

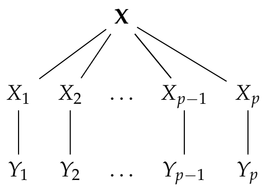

Figure 2.1: Graphical illustration of the assumptions on the joint distribution.

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[5]. 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

[5] 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

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.

```
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
```

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

```
## 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
```

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
```

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.

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 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.

```
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")
```

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

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

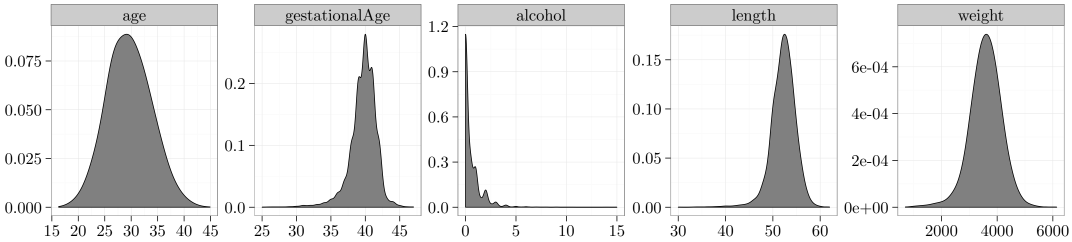

Figure 2.2: Density estimates of continuous variables.

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)
```

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 dif-

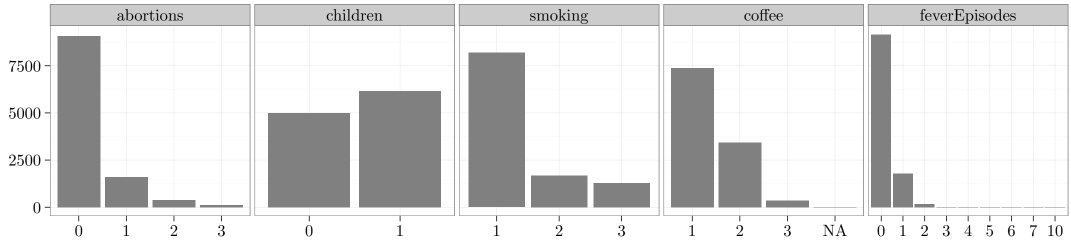

Figure 2.3: Barplots of discrete variables.

ficult. 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

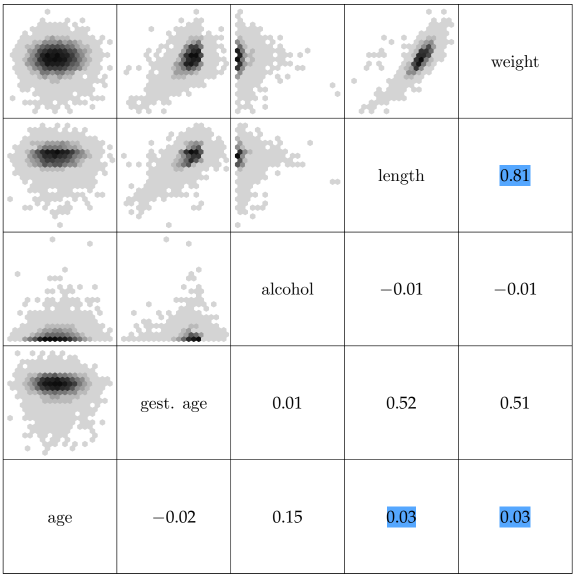

Figure 2.4: Scatter plot matrix of the continuous variables and corresponding Pearson correlations.

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 – in this case to get the 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 are also highly correlated with `gestationalAge`. The `alcohol` and `age` variables are mildly correlated, but they are virtually uncorre-

	abortions					children			coffee			
		0	1	2	3		0	1		1	2	3
smoking	1	6669	1172	270	78	1	3577	4612	1	5939	2140	109
	2	1367	222	66	18	2	848	825	2	890	717	64
	3	1043	201	36	15	3	574	721	3	552	565	177

Table 2.2: Cross-tabulation of `smoking` with `abortions`, `children` and `coffee`.

lated 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 just blindly producing a scatter plot matrix results in a lot of overplotting and huge 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.

Just as the scatter plot is useful for continuous variables, cross-tabulation is useful for categorical variables. 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.

```
crossTabA <- with(pregnant, table(smoking, abortions))
crossTabB <- with(pregnant, table(smoking, children))
crossTabC <- with(pregnant, table(smoking, coffee))
```

Table 2.2 shows the cross-tabulation of `smoking` with the variables `abortions`, `children` and `coffee`. The table shows, for instance, a clear association between coffee drinking and smoking. A χ^2-test (on 4 degrees of freedom) of independence yields a test statistic of 953.7 with a corresponding p-value of 3.8×10^{-205}. To summarize, all the cross-tabulations for the 4 categorical variables and corresponding χ^2-tests of independence are computed.

```
vars <- c("smoking", "coffee", "children", "abortions")
tests <- outer(1:4, 1:4,
           Vectorize(function(i, j) {
             tmp <- summary(table(pregnant[, c(vars[i], vars[j])]))
```

```
                ifelse(i <= j, tmp$p.value, tmp$statistic)
              }
              )
)
colnames(tests) <- rownames(tests) <- vars
```

Table 2.3: Test statistics (below diagonal) and p-values (above diagonal) for testing independence between the different variables.

	smoking	coffee	children	abortions
smoking		3.79e−205	9.58e−07	0.376
coffee	954		1.19e−40	0.00155
children	27.7	184		1.49e−41
abortions	6.44	21.4	193	

Table 2.3 shows that all variables are significantly dependent except `abortions` and `smoking`. However, neither the p-value nor the test statistic 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.

diskretne

```
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))
```

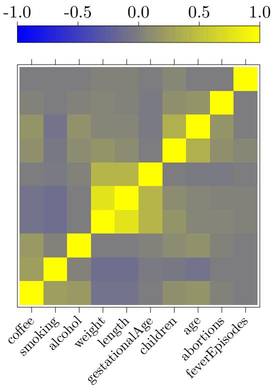

Figure 2.5: Spearman correlation matrix. Variables are ordered according to a hierarchical clustering.

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

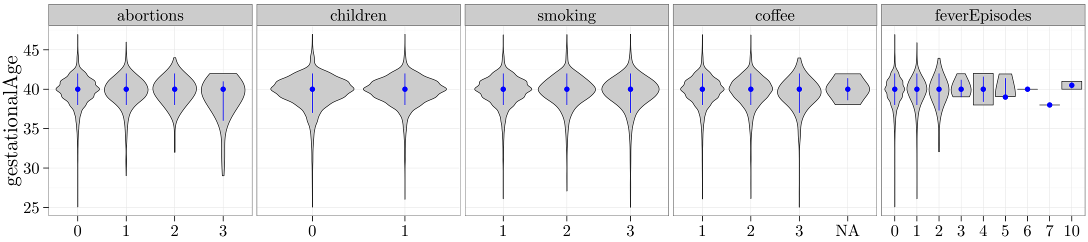

Figure 2.6: Violin plots, medians and interdecile ranges for the distribution of `gestationalAge`. Note that there are very few observations with many fever episodes.

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.

was to sort the variables so that th

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

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 2.4 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 predictor is included. The test used is the F-test – see the next section, page 34, 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

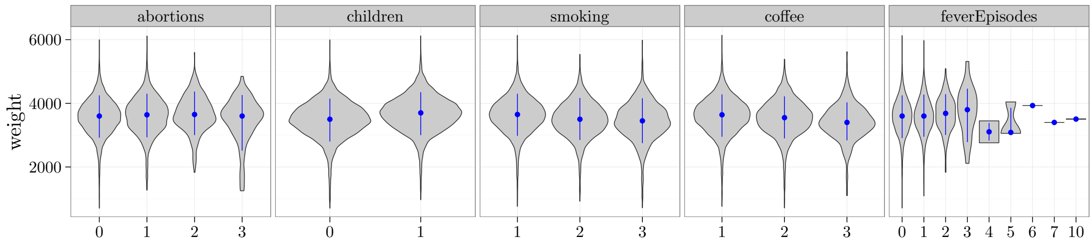

Figure 2.7: Violin plots, medians and interdecile ranges of `weight`.

Table 2.4: 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

in Figure 2.4, but this time we have added the linear regression line. For the continuous variables the tests reported in Table 2.4 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.

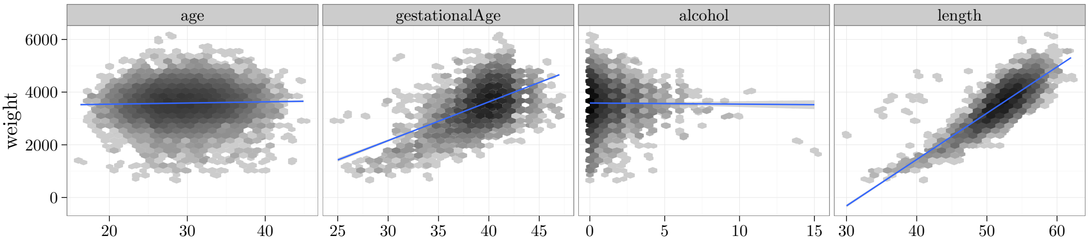

Figure 2.8: Scatter plots including linear regression line.

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 2.5 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 2.6 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.

Table 2.5: 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

```
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 2.6: 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, what we can do about it.

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 qq-plots 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 +
```

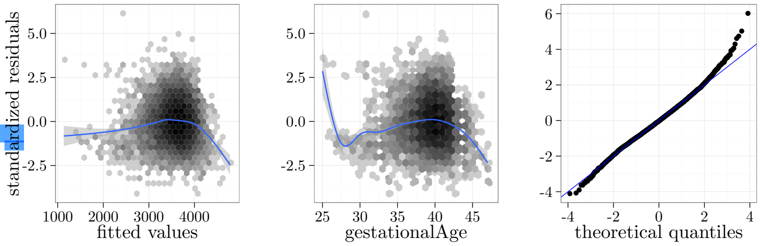

Figure 2.9: Diagnostic plots. Standardized residuals plotted against fitted values, the predictor `gestationalAge`, and a qq-plot against the normal distribution.

Why not use the plot method for `lm`-objects? That's OK for interactive usage, but difficult to customize for publication quality.

```
  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)
```

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 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

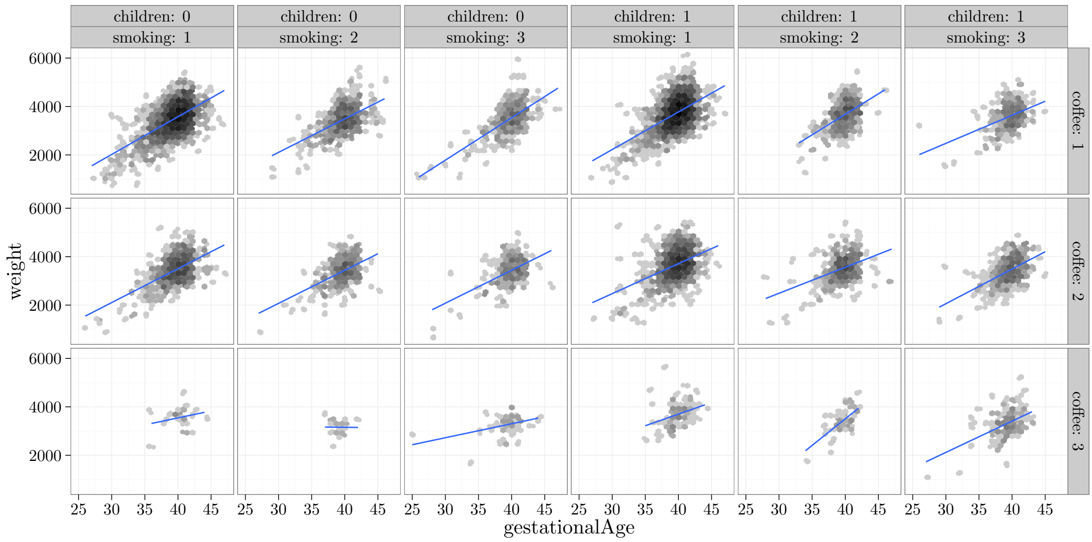

Figure 2.10: Scatter plots of `weight` against `gestationalAge` stratified according to the values of `smoking`, `children` and `coffee`

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 stratification 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

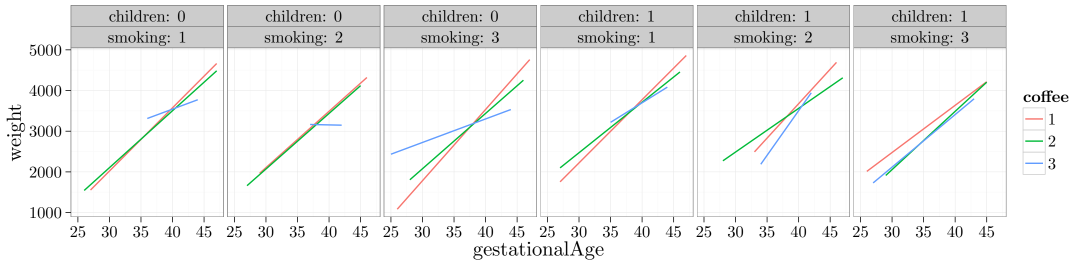

Figure 2.11: Comparison of estimated regression lines for `gestationalAge` stratified according to the values of `smoking`, `coffee` and `children`

model.

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

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

Table 2.7 shows that the F-test of the full third order interaction model against the main effects model is clearly significant. Since 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 2.7: 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 non-linearity 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")
```

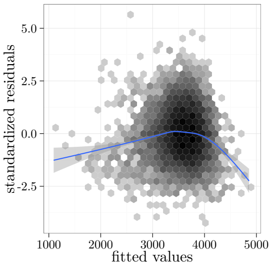

Figure 2.12: Residual plot for the third order interaction model.

Before we conclude the analysis, we test if the inclusion of the 4 weak predictors together is necessary. Table 2.8 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 2.6 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 2.5, depending upon whether the woman had had 1, 2 or 3+ spontaneous abortions, the study is inconclusive on how `abortions` is related `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 2.8: 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 non-linear 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 slope 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 non-linearity using spline basis expansions.

The theory of the linear model

The theory that we will cover in this section is divided into two parts. First, we will consider how the unknown β-parameters are estimated in theory and in practice using the least squares estimator. Second, we consider results on the distribution of the estimators and tests under the weak assumptions A1, A2 and A4 and under the strong assumptions A3 and A5. Needless to say, the conclusions obtained under A3 and A5 are stronger.

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 3.

We introduce the *weighted squared error loss*[6] 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.

[6] With $\mathbf{W} = \mathbf{I}$ this loss is proportional to the negative log-likelihood loss under assumptions A3 and A5 as derived in Chapter 3.

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 β-parameters are 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*.*

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[7]. The second derivative is

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

If $\mathbf{X}$ has rank p, $D^2_\beta \ell(\beta)$ 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

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

[7] The gradient,

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

is a column vector.

As we discuss below, the practical computation of the solution does not rely on explicit matrix inversion.

THE GEOMETRIC interpretation of the solution provides additional insight into the weighted least squares estimator. 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

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

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

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 $\mathbf{W}$.

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.

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[8]. This is seen by

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

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.

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[9] $\mathbf{W} = \mathbf{L}\mathbf{L}^T$ 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,

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

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.

[9] 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.

Distributional results

The results above are all on the estimation of β. The results below are on the distribution of $\hat{\beta}$. They are based on different combinations of assumptions A1–A5. Throughout we restrict attention to the case where $\mathbf{W} = \mathbf{I}$.

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

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

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 A4a 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, in addition, assumptions A2 and A4b 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 $n - p$. □

Just as assumptions A1, A2 and A4 are distributional assumptions on the first and second moments, the distributional results are, under these assumptions, results on the first and second moments. If we want precise results on the distribution of $\hat{\beta}$ and $\hat{\sigma}^2$ 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 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[10] $p \times p$ 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.

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

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.*

Proof. See EH, Chapter 10. □

The terminology associated with the F-test is as follows. 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. These numbers are computed and reported by the R function `anova` in addition to the actual F-test statistic and corresponding p-value 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[11]. 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. The literature shows that a considerable amount of work has been put into choosing the $\mathbf{X}$ matrix, and thus the parametrization, so that scientific hypotheses can be formulated in terms of parameters being 0. This is particularly so in the ANOVA literature when categorical predictors and their interactions are considered.

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

Birth weight – non-linear expansions

We found in the previous analysis of the birth weight data a lack of model fit that appeared to be related to a non-linear relation between `weight` and `gestationalAge`. To handle non-linear relations between the response and one or more predictors, the predictors (as well as the response) can be non-linear 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 non-linearity. 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.

```
x <- seq(25, 45, 0.1)
pol <- as.data.frame(cbind(x, poly(x, degree = 5)))
mPol <- melt(pol, id.vars = "x")
p <- ggplot(NULL, aes(x, value, colour = variable), size = 1) +
  scale_color_discrete(guide = "none") + xlab("") + ylab("")
p + geom_line(data = mPol)
```

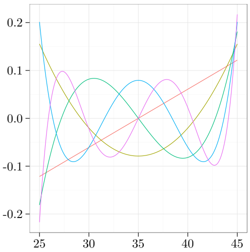

Figure 2.13: Orthogonal polynomials.

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.

An alternative to polynomials is splines. A spline[12] 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

[12] It is also a thin and flexible wood or metal strip used for smooth curve drawing.

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.

```
bSpline <- as.data.frame(cbind(x, bs(x, knots = c(32, 40))))
mBSpline <- melt(bSpline, id.vars = "x")
p + geom_line(data = mBSpline)
```

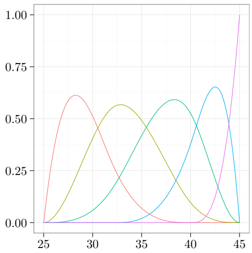

Figure 2.14: *B*-splines computed using the `bs` function.

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[13]. 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.

```
nSpline <- as.data.frame(cbind(x, ns(x, knots = c(32, 35, 38, 41))))
mNSpline <- melt(nSpline, id.vars = "x")
p + geom_line(data = mNSpline)
```

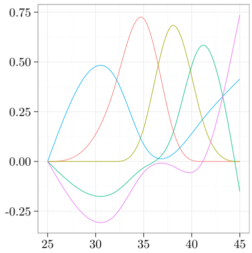

Figure 2.15: *B*-spline basis for natural cubic splines computed using the extttns function.

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.

Expansions should only be tried if we can expect to get something out of it. If a predictor is only weakly marginally associated with the response, it will be unlikely that we get a significant non-linear effect out of expanding it. On the other hand, we should be careful not to construct models tailor-made to capture non-linear 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

[13] 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

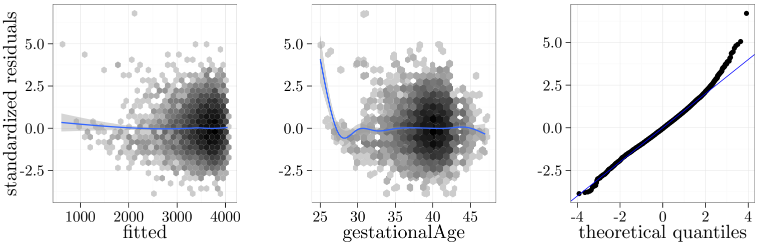

Figure 2.16: Diagnostic plots for the model with `gestationalAge` expanded using splines.

be a model selection procedure with statistical implications, which it is very difficult to account for. It is always extremely difficult to take informal model selection into account in the statistical test, which will result tests 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 present a test of the non-linear effect.

Non-linear main effects model.

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

Table 2.9: 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	11136	2.4667e+09	3	7.088e+07	107	4.424e−68

The main effects model is a submodel of the non-linear main effects model. This is not obvious. The non-linear 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 2.9 is that the non-linear model is highly significant.

The argument for expanding `gestationalAge`, and not `age` or `alcohol`, based on the initial correlation analyses and Figure 2.4, is that only `gestationalAge` is sufficiently correlated with `weight` to justify an expansion. We look for strong association rather than non-linear association as an argument for increased flexibility. The same line of thought applies to the selection of knots[14] for the splines. The knots should be placed reasonably relative to the distribution of the predictor, so that we learn about non-linearities 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 non-linear effect with 1 degrees of freedom would not appropriately take into account how the placement of the knot was made.

Figure 2.16 shows diagnostic plots for the non-linear main effects model. They show that the inclusion of the non-linear 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.

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

Reduced non-linear main effects model.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11142	2.4694e+09				
2	11136	2.4667e+09	6	2.605e+06	1.96	0.06764

Table 2.10: Test of weak predictors in the non-linear main effects model.

Table 2.10 shows that dropping the 4 weak predictors from the non-linear 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 non-linear main effects model. The figure illustrates the general non-linear 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 =`

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

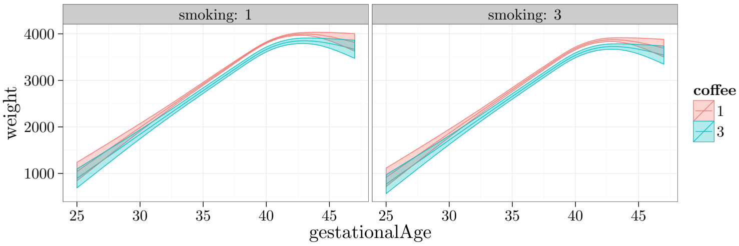

Figure 2.17: Main effects model with basis expansion of `gestationalAge`. Here illustrations of the fitted mean and 95% confidence bands for `children=1`.

`3`, `coffee = 3`) compared to the former, which is explained by the skewness of the predictor distributions. Table 2.11 gives confidence intervals for the remaining parameters based on the non-linear 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)
                         )
predGest <- predict(pregnantLm4, 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 2.11: Confidence intervals.

	2.5 %	97.5 %
(Intercept)	669.15	1065.22
children1	154.39	190.02
coffee2	−85.84	−46.67
coffee3	−199.25	−94.69
smoking2	−123.66	−73.42
smoking3	−153.04	−95.64

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 non-linear effect coupled with a third order interaction, say, results in an almost saturated model.

Non-linear interaction model.

```
form <- weight ~ (smoking + coffee + children) * nsg(gestationalAge)
pregnantLm5 <- lm(form, data = pregnant)
anova(pregnantLm4, pregnantLm5)
```

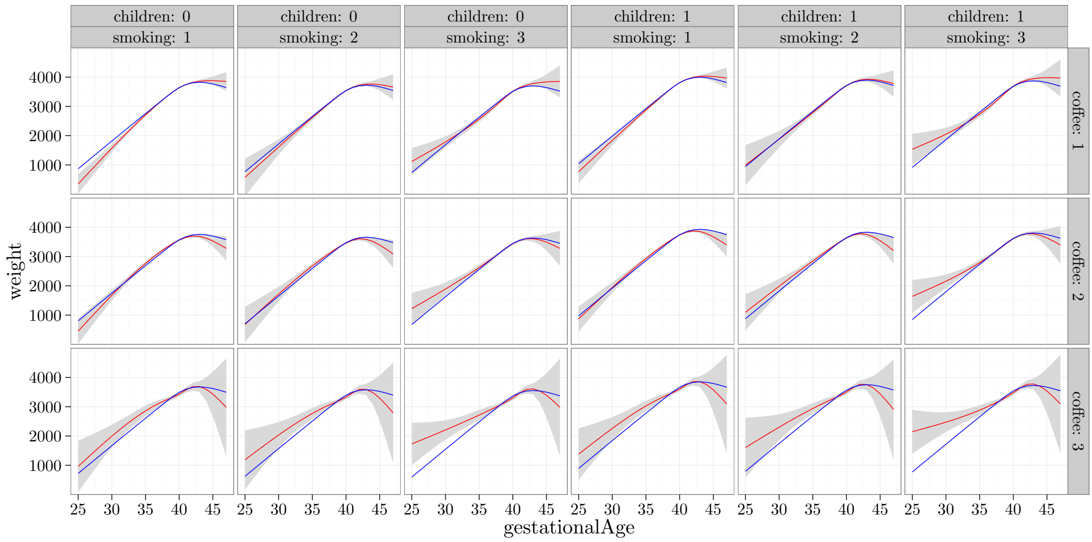

Figure 2.18: Comparison of the interaction model (red, 95% gray confidence bands) with the reduced non-linear main effects model (blue).

```
predFrame <- expand.grid(children = factor(c(0, 1)),
                         smoking = factor(c(1, 2, 3)),
                         coffee = factor(c(1, 2, 3)),
                         gestationalAge = 25:47
                         )
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))
```

That is, the interaction model has more or less a separate mean for all observed combinations of the predictors. Table 2.12 shows that inclusion of interaction terms is still significant.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11142	2.4694e+09				
2	11122	2.4542e+09	20	1.5198e+07	3.44	2.944e−07

Table 2.12: Test of interactions.

We should, however, try to visualize how the non-linear interaction model differs from the model with only main effects. Figure 2.18 shows the predicted values for the reduced non-linear 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 non-linear 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 non-linear 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

```r
X <- bs(seq(0, 1, 0.01), knots = c(3, 5, 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?

3

Generalized linear models

This chapter treats a generalization of the linear model that preserves much of its tractability, but can be adapted to a number of cases where the linear model is inappropriate. The generalization is tightly linked to alternative, non-normal, choices of response distributions. Among the possible alternatives are discrete distributions like the binomial and the Poisson distribution. The generalization can also be understood purely in terms of mean and variance relations between the response and the predictors. This extends, as for the linear model, the scope of generalized linear models beyond the specific response distributions.

The fundamental model assumptions

The linear model assumptions A1–A3 can be generalized by allowing for non-linear 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 3.1. 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 (non-linearly) 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 EXAMPLE IS A PROTOTYPICAL generalized linear model. The expectation is a non-linear transformation of $X^T\beta$, and the relation between the expectation and the variance is given by 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*.

[1] Typically a continuous and monotone function – a distribution function for instance. The most frequent choice is the logistic 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 models referred to in GA3 are introduced in the next section. They replace the normal distribution, 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 of the exponential dispersion models we use. 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.

Exponential families and dispersion models

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

$$\varphi(\theta) = \int e^{\theta y} \, \nu(\mathrm{d}y). \tag{3.1}$$

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 result 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.

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

2. 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 3.2. *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 3.4.

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

Lemma 3.3. *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

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

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

$$\begin{aligned}
f(y) &= e^{\alpha\theta_1 y}, \\
g(y) &= e^{(1-\alpha)\theta_2 y}, \\
p = \frac{1}{\alpha} &\text{ and } q = \frac{1}{1-\alpha}.
\end{aligned}$$

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 3.5. 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

$$\frac{d\rho_{\theta,\psi}}{d\nu_\psi} = e^{\frac{\theta y - \kappa(\theta)}{\psi}}. \tag{3.2}$$

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.

Definition 3.4. *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 (3.2).

Example 3.5. 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]

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

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

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

[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.

Example 3.6. The Poisson distribution with mean λ has density (point probabilities)

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

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

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

and we identify the distribution as an exponential family with

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

Which particular parametrization that is 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 3.7. The Γ-distribution with shape parameter $\lambda > 0$ and scale parameter $\alpha > 0$ has density

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

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

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

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 $\alpha^2\lambda = \psi\mu^2 = \psi/\theta^2$.

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.

THE DIFFERENT PARAMETRIZATIONS can be related through differentiation.

Theorem 3.8. *The function κ is infinitely often differentiable on I. If $Y \sim \mathcal{E}(\theta, \nu_\psi)$ for $\theta \in I$ then*

$$EY = \kappa'(\theta) \tag{3.3}$$

and

$$VY = \psi\kappa''(\theta). \tag{3.4}$$

For the Γ-distribution:

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

and

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

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

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

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

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

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. □

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

Figure 3.1: Relations between the three parametrizations.

We already know from Lemma 3.3 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 (3.4), see Exercise 3.3. 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 3.9. *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 $c(\eta)$ function.

Corollary 3.10. *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 (3.3) 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 3.11. *The canonical link function is the link function*

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

Example 3.12. For the Poisson distribution we have $\kappa(\theta) = e^\theta$, see Example 3.6. This implies that $\kappa'(\theta) = e^\theta$ and the canonical link function is, for the Poisson distribution, $g(\mu) = \log \mu$.

Example 3.13. For the Γ-distribution we have from Example 3.7 that $\kappa(\theta) = -\log(-\theta)$ and

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

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

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

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.

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

Figure 3.2: The linear predictor, $\eta = X^T\beta$, enters the exponential dispersion model through the mean value function.

Advertising – a case study

In this case we consider data from a large database from a major supermarket chain. The overall goal is to predict the number of items sold weeks where the chain runs an advertising campaign, that is, advertising combined with a discount. The item considered here is frozen vegetables, but in reality the goal is 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 3.1: 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 data base. The cross tabulation of `week` and `ad` in Table 3.2 shows that for all other cases, advertising is either 1 or 0 within each week.

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 `stores` factor. 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.

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

Table 3.2: Cross tabulation of week number and advertising indicator.

```
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]
}
)
```

The categorical store variable 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)
```

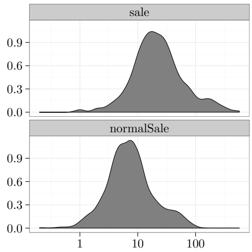

Figure 3.3: Density estimates of `sale` and `normalSale`. Note the log-scale on the x-axis.

The marginal distributions of `sale` and `normalSale` are seen in Figure 3.3. 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 3.4. We note a quite skewed distribution with around one-third of the observations corresponding to the largest discount.

```
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)
```

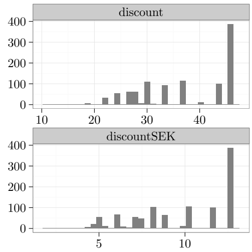

Figure 3.4: Barplots of discount variables.

Pairwise associations

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

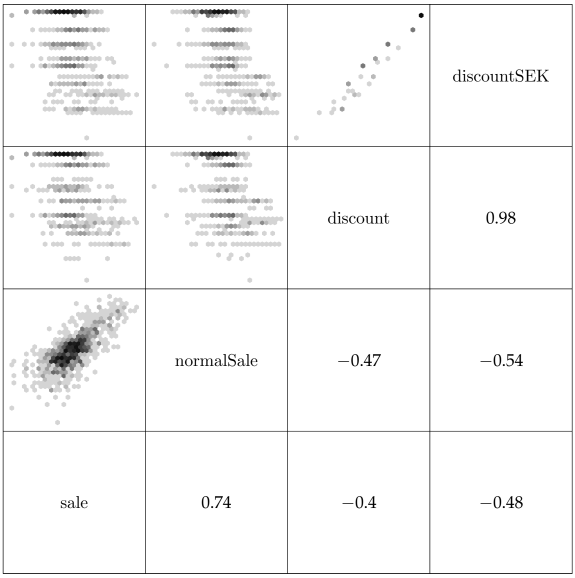

Figure 3.5: Scatter plot matrix of the 4 continuous variables and the corresponding Pearson correlations.

```
contVar <- c("sale", "normalSale", "discount", "discountSEK")
vegLog <- vegetables[, contVar]
vegLog <- transform(vegLog,
                    sale = log10(sale),
                    normalSale = log10(normalSale))
splom(vegLog,
      xlab = "",
      upper.panel = panel.hexbinplot,
      pscales = 0, xbins = 30,
      lower.panel = cor.print
)
```

The scatter plot matrix, Figure 3.5, 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 3.6 further shows how the `discount` and `normalSale` distributions change depending upon the

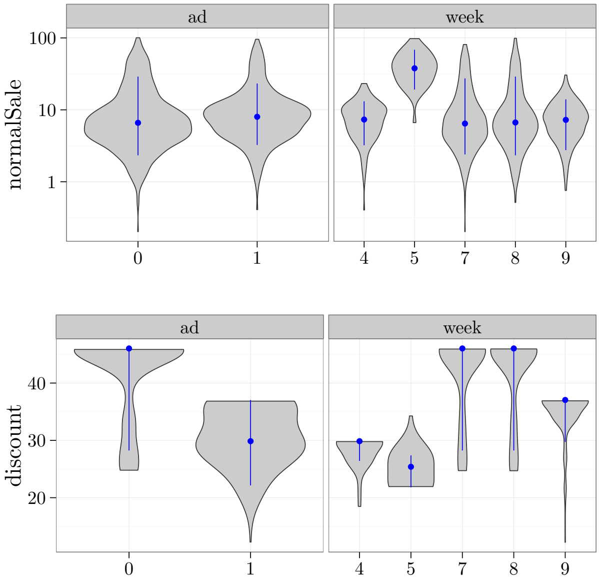

Figure 3.6: Violin plots, medians and interdecile ranges for the distribution of `normalSale` and `discount` stratified according to `ad` and `week`.

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 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 3.7 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"))
```

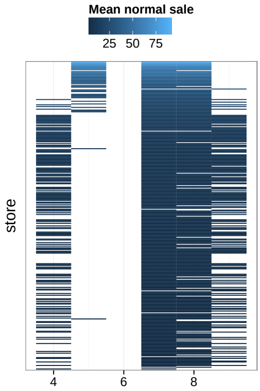

Figure 3.7: Illustration of which stores we have observations from for each of the weeks. The stores are ordered according to the mean normal sale.

To further illustrate this point we consider the relation between the stores and the discount. Figure 3.8 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")
```

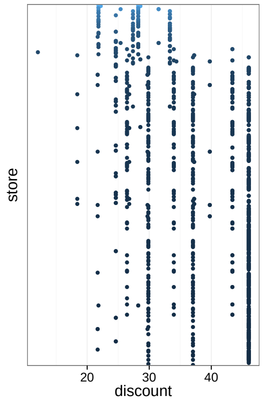

Figure 3.8: Illustration of the relation between the stores and the discount.

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

$$E(Y \mid N, X) = Ne^{X^T\beta} = e^{\log(N)+X^T\beta} \tag{3.5}$$

where X is the vector of additional predictors besides the normal sale. This is a log-linear model – a model with a logarithmic link function – 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 logarithmic link is the canonical link for the Poisson distribution, and we will first consider a Poisson model. As the very first thing we test 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 3.3 shows the results of testing each of the predictors individually. Note that all models contain the fixed control for the

	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

Table 3.3: Marginal association tests sorted according to the p-value.

normal sale. As we see from the table, the predictors are marginally significantly related to the number of sold items, though `discount` 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.

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

A main effects model.

	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

Table 3.4: Summary table of parameter estimates, standard errors and t-tests for the poisson model of sale with 4 predictors.

We do not want to consider all the estimated parameters for the individual stores – only the parameters for the other variables. Table 3.4 shows that in this model all three variables `ad`, `discount` and `discountSEK` are significant, but somewhat surprisingly, the estimated coefficient of `discount` is negative. A possible explanation is the collinearity combined with overdispersion, which will make the tests too optimistic.

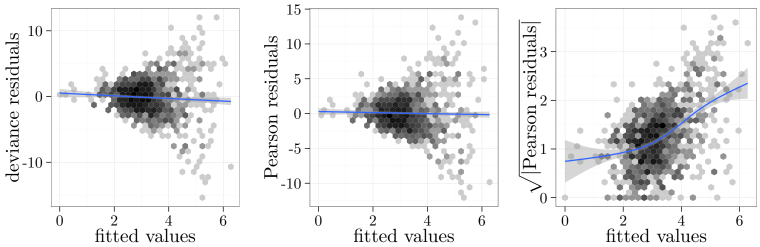

Figure 3.9: Diagnostic plots. Deviance residuals, Pearson residuals and the square root of the absolute value of the Pearson residuals plotted against fitted values.

However, before we consider non-linear 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 3.9 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.

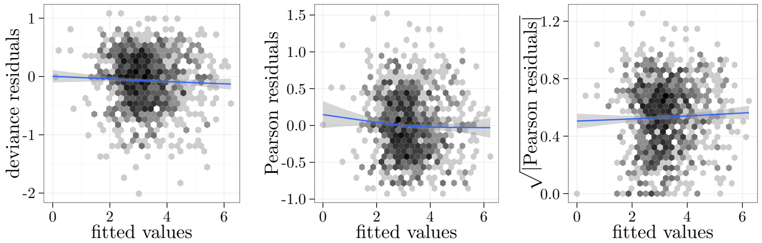

Figure 3.10: Diagnostic plots for the Γ-model. Deviance residuals, Pearson residuals and the square root of the absolute value of the Pearson residuals plotted against fitted values.

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

	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

Table 3.5: Summary table of parameter estimates, standard errors and t-tests for the quasi poisson model of sale with 4 predictors.

As Table 3.5 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 3.10 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 non-linear 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.

```
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 3.6: 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,
            geom = "line", color = store) + ylab("sale") +
  facet_grid(. ~ ad, label = label_both) +
  scale_y_continuous("log-sale")
grid.arrange(p1, p2, ncol = 2)
```

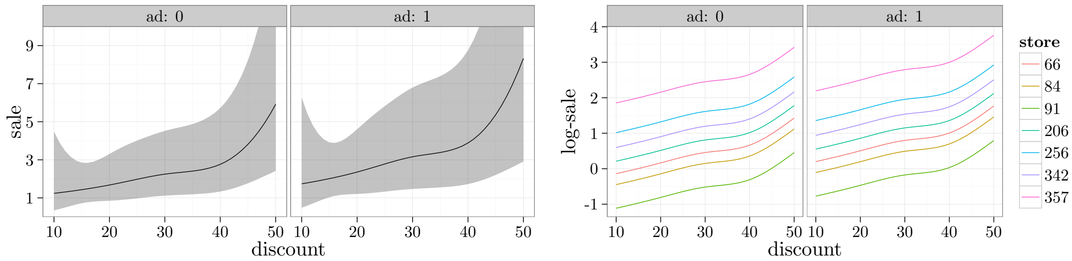

Figure 3.11: Predictions and confidence bands for the sale per normal sale for one typical store (206), and predictions for 7 stores spanning the range of the different stores.

As Figure 3.11 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.

The theory of the generalized linear models

We cover in this section three aspects of the statistical theory for generalized linear models. We treat maximum likelihood estimation based on the exponential dispersion models. This includes the derivation of the non-linear estimation equation, known as the score equation, and the iterative weighted least squares algorithm that is used in practice to fit the models to data. We introduce the likelihood ratio tests (deviance tests) and z-tests. These tests replace the F- and t-tests used for the linear model. In contrast to the linear model it is not possible to derive exact distributional results in the general case. However, the operational procedures for using the tests are the same – the only difference being that the distributions used to compute p-values are approximations. Formal justifications can be based on asymptotic arguments, which are discussed again in the next chapter, page 92, but asymptotic arguments will only be treated briefly. We also discuss model diagnostics and, among other things, the different possible generalizations of residuals. Finally, we show 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 independent observations $Y_1, \ldots, Y_n$ with $Y_i \mid X_i \sim \mathcal{E}(\theta(\eta_i), \nu_\psi)$ and $\eta_i = X_i^T \beta$.

Definition 3.14. *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 3.15. *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 3.10. 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 $\boldsymbol{\eta}$-vector. 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 3.16. *Assume that $Y_1, \ldots, Y_n$ are independent and that $Y_i \mid X_i \sim \mathcal{E}(\theta(\eta_i), \nu_\psi)$ where $\eta_i = X_i^T\beta$. Then with $\boldsymbol{\eta} = \mathbf{X}\beta$ the score statistic is*

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

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*

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

The diagonal weight matrix **W** is

$$\frac{1}{\psi}\begin{pmatrix} \frac{\mu'(\eta_1)^2}{\mathcal{V}(\mu(\eta_1))} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\mu'(\eta_n)^2}{\mathcal{V}(\mu(\eta_n))} \end{pmatrix}$$

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,

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

As argued above, $-D_{\boldsymbol{\eta}} U(\boldsymbol{\eta})$ is diagonal, and the expectation of the diagonal entries are according to Lemma 3.15

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

The alternative formula for the weights follow from Corollary 3.10. Thus

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

□

By observing that

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

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

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

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

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

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 3.17. 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.

THE GENERAL, NON-LINEAR 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.

Example 3.18. 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

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

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.

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[5], 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

[5] 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[6] is found as in the proof of Theorem 3.16 to be

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

where

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

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

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

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

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 3.16 we get the Fisher scoring algorithm. We may note that the diagonal entries in $\mathbf{W}$ are always strictly positive if the mean value map is stricly 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.

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.

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

$$Z_{m,i} = X_i^T \beta_m + \frac{Y_i - \mu(\eta_{m,i})}{\mu'(\eta_{m,i})}. \tag{3.6}$$

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

$$(\mathbf{Z}_m - \mathbf{X}\beta)^T \mathbf{W}_m (\mathbf{Z}_m - \mathbf{X}\beta), \tag{3.7}$$

[6] 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:

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.

1. Compute the working response vector $\mathbf{Z}_m$ based on β_m using (3.6).

2. Compute the weights
$$w_{ii} = \frac{\mu'(\eta_{m,i})^2}{\mathcal{V}(\mu(\eta_{m,i}))}.$$

3. Compute β_{m+1} by minimizing the weighted sum of squares (3.7).

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.

AN ALTERNATIVE understanding of IWLS is provided by relating it to an idealized least squares problem. This derivation emphasizes the fact that the algorithm depends on the assumptions GA1 and GA2 only.

Supposing that we know the true parameter, β, we can form the weighted quadratic loss

$$\ell(\beta') = \sum_{i=1}^{n} \mathcal{V}(\mu_i)^{-1}(Y_i - \mu(X_i^T\beta'))^2$$

where $\mu_i = \mu(X_i^T\beta)$. Besides the fact that the weights depend upon the unknown true parameter, the β' enters non-linearly through $\mu(X_i^T\beta')$. We can Taylor expand the mean once[7] around β to obtain

$$\mu(X_i\beta') \simeq \mu(\eta_i) + \mu'(\eta_i)X_i^T(\beta' - \beta)$$

and plug this approximation back into the loss. This yields the

[7] This is the key step in the Gauss-Newton algorithm for non-linear least squares estimation.

approximating quadratic loss

$$\begin{aligned}\tilde{\ell}(\beta') &= \sum_{i=1}^{n} \mathcal{V}(\mu_i)^{-1}\Big(Y_i - \mu(\eta_i) - \mu'(\eta_i)X_i^T(\beta' - \beta)\Big)^2 \\ &= \sum_{i=1}^{n} \frac{\mu'(\eta_i)^2}{\mathcal{V}(\mu_i)}\left(X_i^T\beta + \frac{Y_i - \mu(\eta_i)}{\mu'(\eta_i)} - X_i^T\beta'\right)^2.\end{aligned}$$

By introducing the *idealized* responses

$$Z_i = X_i^T\beta + \frac{Y_i - \mu(\eta_i)}{\mu'(\eta_i)} \tag{3.8}$$

where $\eta_i = X_i^T\beta$ and the idealized weights

$$w_{ii} = \frac{\mu'(\eta_i)^2}{\mathcal{V}(\mu(\eta_i))},$$

the approximating quadratic loss can we written as

$$\tilde{\ell}(\beta') = (\mathbf{Z} - \mathbf{X}\beta')^T\mathbf{W}(\mathbf{Z} - \mathbf{X}\beta') = ||\mathbf{Z} - \mathbf{X}\beta'||^2_{\mathbf{W}},$$

with $\mathbf{W}$ 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 $\tilde{\ell}(\beta')$. The IWLS algorithm can be understood as iteratively approximating the idealized responses by the working response – by plugging in the current estimate of the β-vector. 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 $\hat{\beta}^{\text{ideal}}$

$$\begin{aligned}E(\hat{\beta}^{\text{ideal}} \mid \mathbf{X}) &= \mathbf{X}\beta, \\ V(\hat{\beta}^{\text{ideal}} \mid \mathbf{X}) &= \psi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}, \\ E(||\mathbf{Z} - \mathbf{X}\hat{\beta}^{\text{ideal}}||^2_{\mathbf{W}} \mid \mathbf{X}) &= (n-p)\psi.\end{aligned}$$

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

$$\hat{Z}_i = X_i^T \hat{\beta} + \frac{Y_i - \mu(\hat{\eta}_i)}{\mu'(\hat{\eta}_i)}$$

we obtain a computable loss[8]

$$||\hat{\mathbf{Z}} - \mathbf{X}\beta||^2_{\hat{\mathbf{W}}}. \tag{3.9}$$

The fact that $\hat{\beta}$ is a fixed point for IWLS implies that $\hat{\beta}$ is also the minimizer of this loss function. Provided that this loss 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, that we can estimate the dispersion parameter as

$$\hat{\psi} = \frac{1}{n-p}\mathcal{X}^2. \tag{3.10}$$

Deviance and z-tests

The distributional results for the idealized weighted least squares estimator suggest the approximation

$$\sqrt{\psi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}_{ii}}$$

of the standard error of $\hat{\beta}_i$. 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

$$Z_a = \frac{a^T\hat{\beta} - a^T\beta}{\sqrt{\hat{\psi}a^T\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}a}}.$$

[8] 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.

for $a \in \mathbb{R}^p$.

The Z-score is used exactly as it is be 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

$$a^T\hat{\beta} \pm z\sqrt{\hat{\psi}a^T(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}a} \tag{3.11}$$

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. An alternative to using theoretical approximations is to use bootstrapping to compute approximating quantiles. The square of the Z-statistic, Z_a^2, is known as the Wald statistic. Multivariate Wald statistics are possible, but they are used less frequently.

The likelihood ratio tests are alternatives to z-tests or the multivariate Wald tests. Their derivation are based on the stronger distributional assumption that the the responses have a distribution from an exponential dispersion model. In the framework of generalized linear models the likelihood ratio tests are usually formulated in terms of *deviances*. For an exponential dispersion model, Lemma 3.15 shows that the log-likelihood for $\psi = 1$ is

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

We introduce the deviance as twice the negative log-likelihood up to an additive constant. In the canonical parameter we introduce the *unit deviance*[9] as

$$d_0(Y,\theta) = 2\left(\sup_{\theta' \in I} \ell_Y(\theta') - \ell_Y(\theta)\right)$$

for $Y \in \bar{J}$ 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.

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

Definition 3.19. *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 $Y \in \bar{J}$ *and* $\mu \in J$.

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

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

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 a simple analytical expression for many concrete examples, which is interpretable as a measure of how the observation Y deviates from the expectation μ.

Example 3.20. 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 3.21. 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 3.22. 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 3.23. *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. □

For a generalized linear model in general 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 $\beta \in \mathbb{R}^p$ – computed by the IWLS algorithm. Likewise, if $\hat{\beta}^0$ is the maximum-likelihood estimate of β under the null hypothesis

$$H_0 : \beta \in L$$

where $L \subset \mathbb{R}^p$ 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 3.24. *The deviances for the model and the null hypothesis are*

$$D = \sum_{i=1}^{n} d(Y_i, \hat{\mu}_i) \quad \textit{and} \quad D_0 = \sum_{i=1}^{n} d(Y_i, \hat{\mu}_i^0),$$

respectively. The deviance test statistic is

$$D - D_0$$

and the F-test statistic is

$$\frac{(D - D_0)/(p - p_0)}{D/(n-p)}.$$

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

$$D - D_0 \overset{\text{approx}}{\sim} \psi \chi^2_{p-p_0},$$

which is exact for the linear model with normal errors (under the assumptions A3 and 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.

The 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 92, but we will not consider any formal asymptotic arguments.

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 could first of all consider the raw residuals

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

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

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

Note that the sum of the squared Pearson residuals is the Pearson χ^2-statistic. The Pearson residuals are used much 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, however, expect the Pearson residuals to appear normal, not is it clear what the distribution of the residuals is even if the model assumptions are fulfilled.

The deviance of a model is a sum of deviance contributions – one for each observation. – and we define the *deviance* residuals as

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

We observe that the deviance is then the sum of the squared deviance residuals. They can be used like the Pearson residuals.

Finally, the *working* residuals should also be mentioned. They are

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

and are the raw residuals from the weighted least squares problem (3.9).

Example 3.25. For the binomial case we find that the raw residuals are

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

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

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

and the deviance residuals are

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

Existence and uniqueness

The canonical link function offers some simplifications compared to the general case. If the link function is the canonical link, we have already mentioned the simplifications in the IWLS algorithm. 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

$$\mathbf{W}^{\mathrm{obs}} = \mathbf{W},$$

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

$$\tau(\beta) = t. \tag{3.12}$$

The main result in what follows is a quite satisfactory characterization of existence and uniqueness of the solution to the score equation.

Define the convex, open set

$$D = \{\beta \in \mathbb{R}^p \mid \mathbf{X}\beta \in (I^\circ)^n\}$$

of parameters for which the linear predictor is an interior point of I. The set depends upon $\mathbf{X}$ and the map τ is defined on D. We only search for solutions in D. We will also assume in the following that $\mu'(\eta) = \mathcal{V}(\eta) > 0$ for $\eta \in I^\circ$. 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 3.26. *If $\mathbf{X}$ has full rank p the map $\tau : D \to \mathbb{R}^p$ is one-to-one. With $C := \tau(D)$ there is a unique solution to (3.12) if and only if $t \in C$.*

Proof. All we need is to 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 $J = \mu(I^\circ)$. *Disclaimer: In the following proof we assume that $D = \mathbb{R}^p$, which is not needed for the result to hold*

Lemma 3.27. *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 $t_0 \in C$. □

Corollary 3.28. *The set $C = \tau(D)$ has the representation*

$$C = \left\{ \sum_{i=1}^{n} \mu_i X_i \mid \mu_i \in J \right\} \tag{3.13}$$

and is convex. If $\mathbf{X}$ has full rank p then C is open.

Proof. Lemma 3.27 shows that C has the claimed representation. The function μ is continuous and increasing, and, by assumption, it is strictly increasing. Since $J := \mu(I^\circ)$ is the image of the open interval I° the continuity of μ assures that J is an interval and strict monotonicity assures that J is 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

$$C = \mathbf{X}^T (J \times \ldots \times J)$$

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 3.28, 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 $Y \in \bar{J}$ – 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.

Something on using the `biglm` package and the `bigmemory` project.

Exercises

Exercise 3.1. 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 3.2. Show that the binomial distributions on $\{0, 1, \ldots, n\}$ with success probabilities $p \in (0, 1)$ form an exponential family.

Exercise 3.3. 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 3.4. 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 3.5. 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.

4

Statistical methodology

In the previous chapters we have considered least squares and maximum likelihood estimation as well as F- and deviance-tests for hypothesis testing. Confidence intervals based on standard errors were also considered. Tests and confidence intervals quantify in different ways the uncertainty associated with the estimators. In this chapter we will deal more systematically and more extensively with statistical methodology for quantifying uncerntainty. We will focus on general methods for interval estimation as well as methods for model validation and model validation.

INTERVAL ESTIMATES add uncertainty considerations to point estimates. Three types of intervals: likelihood intervals, confidence intervals, and credibility intervals.

It is useful to separate the construction of interval estimators from the procedures used for calibrating the estimators to have desirable distributional properties. Likelihood intervals and the Bayesian credibility intervals are, for instance, meaningful intervals in their own right whether or not they have been calibrated to be confidence intervals. In Section 4 we deal with several procedures for exact as well as approximate likelihood based interval constructions. Any of these procedures can be calibrated to give approximate confidence intervals. This requires distributional knowledge, which can be obtained from asymptotic theory or by simulations as treated in Section 4.

model validation

Model validation is within the context of predictive modeling a question about how well the model works as a predictive model. This is quantified as the expected loss of predictions, know as the risk of the model, for a given loss function. We will discuss methods for risk estimation. Validation is often used for comparative purposes and for model selection.

Likelihood and quasi-likelihood methods

Likelihood intervals, profiling, profiling approximate log-likelihoods using quadratic approximations.

Uniqueness of the variance function. Representation of the deviance in terms of the variance function.

Calibration

Decision theory

General loss functions, risk

The linear model

The Gauss-Markov Theorem

Does there exist a minimal variance, unbiased estimator of β? We consider linear estimators only

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

for some $N \times p$ matrix C requiring that

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

for all β. That is, $C^T \mathbf{X} = I_{p+1} = \mathbf{X}^T C$. Under Assumption 1

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

and we have

$$\begin{aligned} V(\hat{\beta}-\tilde{\beta}|\mathbf{X}) &= V(((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - C^T)Y|\mathbf{X}) \\ &= \sigma^2((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - C^T)((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - C^T)^T \\ &= \sigma^2(C^TC - (\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

The matrix $C^TC - (\mathbf{X}^T\mathbf{X})^{-1}$ is positive semidefinite, i.e. for any $a \in \mathbb{R}^{p+1}$

$$a^T(\mathbf{X}^T\mathbf{X})^{-1}a \leq a^TC^TCa$$

Theorem 4.1. *Under assumptions A1 + A2 + A4 the least squares estimator of β has minimal variance among all linear, unbiased estimators of β.*

This means that for any $a \in \mathbb{R}^p$, $a^T\hat{\beta}$ has minimal variance among all estimators of $a^T\beta$ of the form $a^T\tilde{\beta}$ where $\tilde{\beta}$ is a linear, unbiased estimator.

It also means that $V(\tilde{\beta}) - (\mathbf{X}^T\mathbf{X})^{-1}$ is positive semidefinite – or in the partial ordering on positive semidefinite matrices

$$V(\tilde{\beta}) \succeq (\mathbf{X}^T\mathbf{X})^{-1}.$$

Model validation

Model comparison

5

Survival analysis

6

Penalized regression

A

Standard distributions

B

Linear algebra

C

R

Bibliography

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.

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.

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.

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.