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 checking.

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 block 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 mothers 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.

[9] yihui.name/knitr/

[10] www.rstudio.com

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`.

[11] ggplot2.org

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

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

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).

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

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-*

[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.

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

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 :-)
```

2 The linear model

The fundamental linear model

This section briefly reviews 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 discussed and their implications.

THE LINEAR MODEL relates a continuous *response* variable Y to a p-dimensional vector X of *predictors*[1] via the relation

$$Y = X^T\beta + \varepsilon. \tag{2.1}$$

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

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

$$Y = \beta_0 + X^T\beta + \varepsilon.$$

It is notationally convenient to assume that the intercept parameter is included among the other parameters. This can be achieved by joining the predictor $X_0 = 1$ to X, thereby increasing the dimension to $p + 1$. In the general presentation we will not pay particular

attention to the intercept. We will assume that if an intercept is needed, it is appropriately included among the other parameters, and we will index the predictors from 1 to p. Other choices of index set, e.g. from 0 to p, may be convenient in specific cases.

The ε is called the *error* or *noise* term. The model is not much of a model if the error can be arbitrary. The typical model assumptions are distributional and specified in terms of the *conditional* distribution of ε given X. We list them in decreasing order of importance.

A1 The conditional expectation of ε given X is 0,

$$E(\varepsilon \mid X) = 0.$$

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

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

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

$$\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2).$$

Assumption A1 is the key assumption, which implies that

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

Thus the linear[2] model is a model of the conditional expectation of the response variable given the predictors. This assumption is crucial if we want $X^T\beta$ to be interpretable and useful.

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

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.

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

Assumption A3 implies that ε and X are independent, as the conditional distribution of ε given X does not depend upon X in this case. 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], 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.

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

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

A4a The conditional distribution of Y_i given $\mathbf{X}$ depends upon X_i only.

A4b The variables Y_i and Y_j are conditionally uncorrelated given $\mathbf{X}$,

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

A5 The variables $Y_1, \ldots, Y_n$ are conditionally independent given $\mathbf{X}$.

Assumptions A4a and A4b 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. We refer to A4a and A4b collectively as assumption A4.

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) and (2.3), and the strong set A3 and A5, which, in addition, imply the distributional identity (2.4).

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 normal distribution of the errors. 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.

[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

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

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

```
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.

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

```
head(pregnant, 4)

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

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

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

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

```
summary(pregnant)

##  interviewWeek  fetalDeath        age       abortions children
##  14     :2379   0   :11659   Min.   :16.3   0:9598    0:5304
##  15     :2285   1   :  119   1st Qu.:26.6   1:1709    1:6513
##  16     :2202   NA's:   39   Median :29.5   2: 395
##  13     :2091                Mean   :29.6   3: 115
##  12     :1622                3rd Qu.:32.5
```

```
##  11     :1089                Max.   :44.9
##  (Other): 149
##  gestationalAge smoking     alcohol        coffee         length
##  Min.   :17.0   1:8673   Min.   : 0.000   1   :7821   Min.   : 0.0
##  1st Qu.:39.0   2:1767   1st Qu.: 0.000   2   :3624   1st Qu.:51.0
##  Median :40.0   3:1377   Median : 0.000   3   : 368   Median :52.0
##  Mean   :39.4            Mean   : 0.512   NA's:   4   Mean   :51.8
##  3rd Qu.:41.0            3rd Qu.: 1.000               3rd Qu.:54.0
##  Max.   :47.0            Max.   :15.000               Max.   :99.0
##                          NA's   :1                    NA's   :538
##      weight     feverEpisodes
##  Min.   :   0   Min.   : 0.0
##  1st Qu.:3250   1st Qu.: 0.0
##  Median :3600   Median : 0.0
##  Mean   :3572   Mean   : 0.2
##  3rd Qu.:3950   3rd Qu.: 0.0
##  Max.   :6140   Max.   :10.0
##  NA's   :538
```

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

We present density estimates of the 5 continuous variables, see Figure 2.1. The density estimates, as the majority of the figures presented in this book, were produced using the `ggplot2` package.

`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 data set.

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

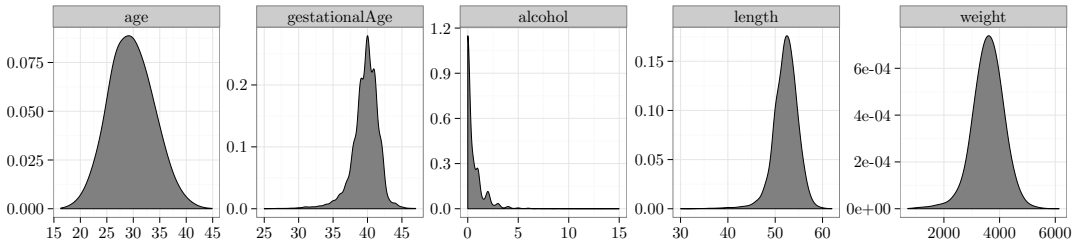

Figure 2.1: Density estimates of continuous variables.

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

What is most noteworthy in Figure 2.1 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 2.2 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

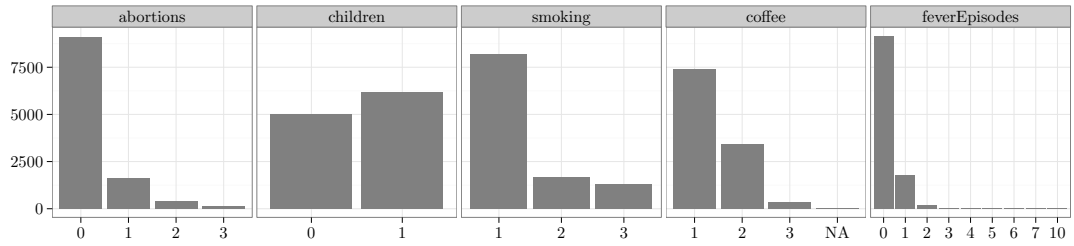

Figure 2.2: Barplots of discrete variables.

what is to be expected for a random sample of cases. A small variation of a predictor can result in difficulties with estimation and detection of associations between the response and the predictors. However, the data set is quite large, which can potentially compensate for this fact.

```
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

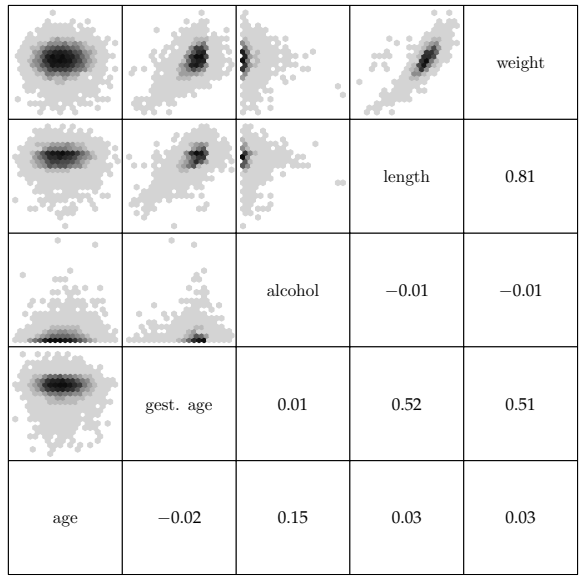

Figure 2.3: Scatter plot matrix of the continuous variables and corresponding Pearson correlations.

will become difficult to determine on statistical grounds if one predictor should be included over the other. It is best to know about such potential issues upfront. Perhaps it is, by subject matter arguments, possible to choose one of the predictors over the other as the most relevant to include in the model.

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

Function `cor.print` formats the correlations for printing. The `na.omit` function removes cases containing missing observations – 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 2.3, show that `length` and `weight` are (not surprisingly) very correlated, and that both of these variables

	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
smoking 2	1367	222	66	18	2	848	825	2	890	717	64
smoking 3	1043	201	36	15	3	574	721	3	552	565	177

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

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

The scatter plot matrix was produced using the `splom` function from the `lattice` package. The data set is quite large and 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.

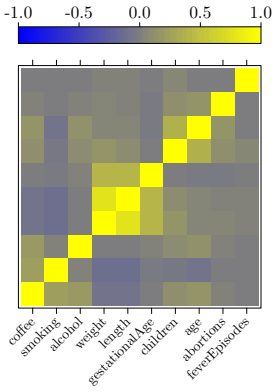

Figure 2.4: Spearman correlation matrix. Variables are ordered according to a hierarchical clustering.

```
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 2.4 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 2.4, we used a hierarchical clustering of the variables. The purpose was to sort the variables so that the large correlations are concentrated around the diagonal. Since there is no natural order of the variables, the correlation matrix could be plotted using any order. We want to choose an order that brings highly correlated variables close together to make the figure easier to read. Hierarchical clustering can be useful for this purpose. For the clustering, a dissimilarity measure between variables is needed. We used 1 minus the absolute value of the correlation. It resulted in a useful ordering in this case.

What we see most clearly from Figure 2.4 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 2.5 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 2.4. Figure 2.6 shows a similar violin plot but this time with the continuous variable being the response variable `weight`. From this

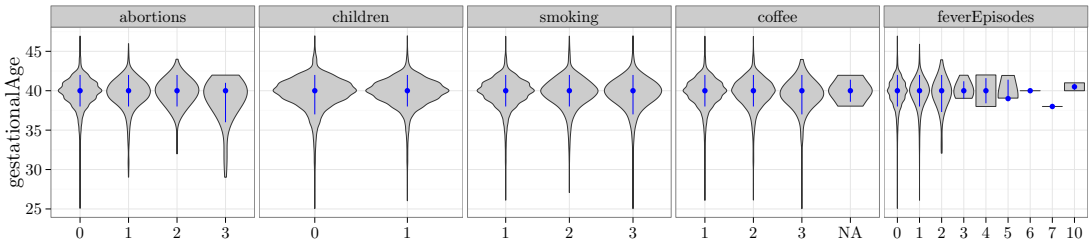

Figure 2.5: Violin plots, medians and interdecile ranges for the distribution of `gestationalAge`. Note that there are very few observations with many fever episodes.

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

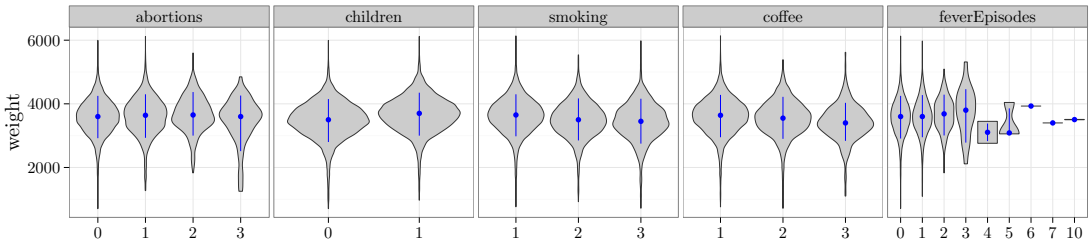

Figure 2.6: Violin plots, medians and interdecile ranges of `weight`.

variable with Df + 1 levels.

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

Table 2.4: Marginal association tests sorted according to the p-value.

Figure 2.7 shows the scatter plots of `weight` against the 4 continuous predictors. This is just the first row in the scatter plot matrix in Figure 2.3, 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

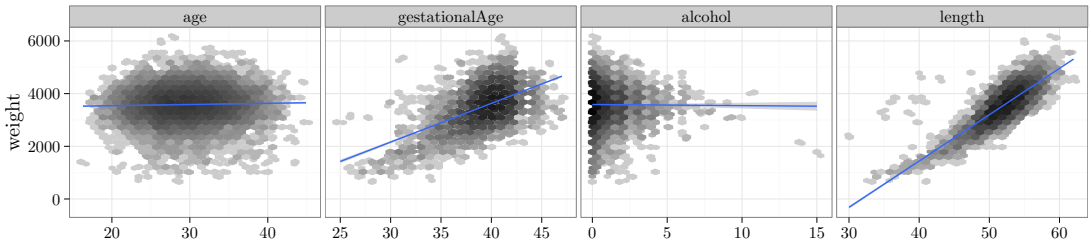

Figure 2.7: Scatter plots including linear regression line.

`weight` is also observable. The `gestationalAge` variable is likewise of little interest if we want to predict `weight` early in pregnancy. The variable is, however, virtually unrelated to the other predictors, and age of the fetus at birth is a logic cause of the weight of the child. It could also be a relevant predictor late in pregnancy for predicting the weight if the woman were to give birth at a given time. Thus we keep `gestationalAge` as a predictor. The remaining predictors are not strongly correlated, and we have not found reasons to exclude any of them. We will thus fit a main effects linear model with 8 predictors. We include all the predictors as they are.

The main effects model.

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

Table 2.5 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

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

Table 2.5: Summary table of parameter estimates, standard errors and t-tests for the linear model of weight fitted with 8 predictors.

to use the tests for a model reduction or model selection.

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

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

Table 2.6: 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,
```

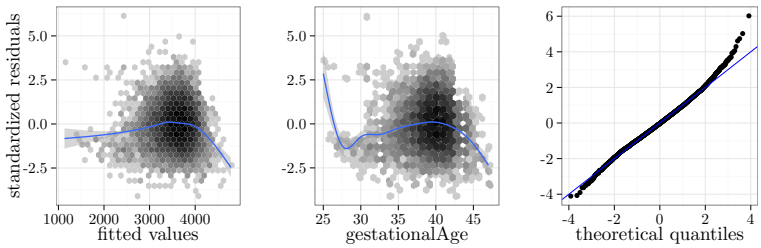

Figure 2.8: Diagnostic plots. Standardized residuals plotted against fitted values, the predictor `gestationalAge`, and a qq-plot against the normal distribution.

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

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

The residual plot in Figure 2.8 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 an 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

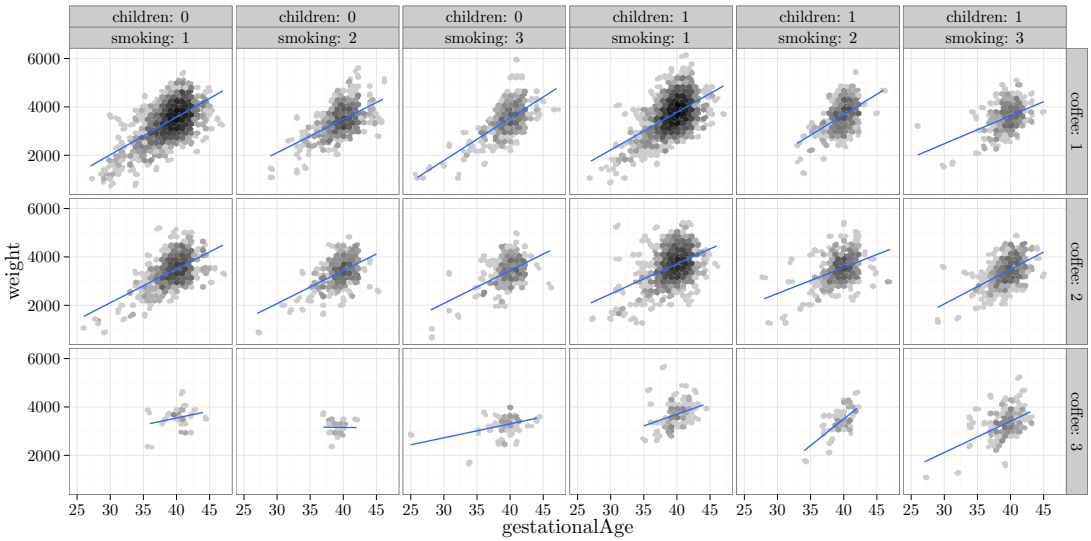

Figure 2.9: Scatter plots of `weight` against `gestationalAge` stratified according to the values of `smoking`, `children` and `coffee`

is more likely that we are able to detect interactions between strong predictors than between weak predictors. To comprehend an interaction model it is advisable to visualize the model to the extend it is possible. This is a point were 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 2.9 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 2.10 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

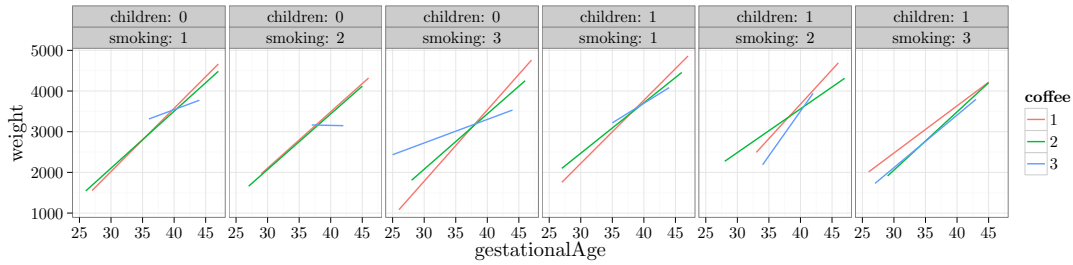

Figure 2.10: Comparison of estimated regression lines for `gestationalAge` stratified according to the values of `smoking`, `coffee` and `children`

main effects model. In doing so we keep the weak predictors in the 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 additive 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 2.11 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")
```

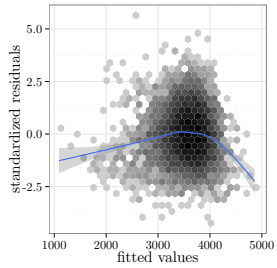

Figure 2.11: Residual plot for the third order interaction model.

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 2.9 or 2.10. 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] That this loss with $\mathbf{W} = \mathbf{I}$. is proportional to the negative log-likelihood loss under assumptions A3 and A5 is derived in Chapter 3

With

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

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

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

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

[7] The gradient, $\nabla_\beta \ell(\beta) = D_\beta \ell(\beta)^T$, is a column vector.

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

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

As we discuss below, the practical computation of 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

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

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

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

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

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

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

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

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

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

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.