

# Topics in Applied Regression

J.H. Maindonald\*

June 4, 2007

## Abstract

The following topics will be treated at various levels of detail, depending on time available.

- Review the theory of linear models, noting the utility of Householder reflections and the QR decomposition both for computation and for theory.  
[c.f., Chapter 1 of [Wood \(2006\)](#)]
- Discuss regression splines, noting in passing the further extension into generalized additive models.  
[Maindonald & Braun \(2007, Section 7.1, pp. 234–238\)](#).
- Introduce the theory of Generalized Linear Models, with logistic regression and Poisson regression as special cases.  
[[Maindonald & Braun \(2007, Sections 8.1 – 8.5\)](#); [Wood \(2006, Sections 2.1 – 2.3\)](#)]
- Brief discussion of supervised learning (classification), using tree-based methods, Breiman’s random forests, and linear discriminant analysis as implemented in R’s function `lda()`. The interest here in the use of these methods to provide constructed variables (in this context, known as “propensities”) for use as explanatory variables in regression.  
[[Maindonald & Braun \(2007, Sections 12.2 and 11.7, with a brief overview of Sections 11.1–11.6\)](#)]
- Note issues in the use of linear and generalized linear models, and of other regression models. This will be, largely, an exercise in awareness raising.
  - Preliminary data exploration; [[Maindonald & Braun \(2007, Chapters 2 & 6\)](#)]
  - Transformation of explanatory variables, where possible, to ensure linearly related predictors;  
[[Maindonald & Braun \(2007, Chapter 6\)](#); [Cook and Weisberg \(1999\)](#)]
  - Diagnostic plots; uses of simulation;  
[[Maindonald & Braun \(2007, Chapter 6\)](#)]
  - Missing variables; noting very striking examples that arise in multi-way tables, perhaps modeled using logistic regression;  
[[Maindonald & Braun \(2007, Subsections 2.2.1, 3.4.5, 6.8.3 & Section 8.3\)](#)]

---

\*Centre for Mathematics & Its Applications, Australian National University, Canberra ACT 0200, AUSTRALIA. <mailto:john.maindonald@anu.edu.au>

- Observational versus experimental data – implications for interpretation and inference;  
[Mairdonald & Braun (2007, Chapter 6); Rosenbaum (2002)]
- Variable selection, noting the use of resampling methods to obtain realistic “error” estimates;  
[Mairdonald & Braun (2007, Chapter 6)]
- Errors in explanatory variables; implications of classical measurement error for inference;  
[Mairdonald & Braun (2007, Chapter 6); Carroll (2006, Chapter 1)]
- Regression on constructed variables – principal components, partial least squares, and propensity scores;  
[Mairdonald & Braun (2007, Chapter 13)]
- Worked examples, some using data that have been a basis for published research, will be used as a basis for a more extended discussion of selected issues from the list given above;
- Laboratory exercises, 2 hours per week, using the R system for the computations. These will adapt and extend relevant material from the set of laboratory exercises used for the 2006 Mathematics Department “Data Mining” course at Australian National University. [See <http://www.maths.anu.edu.au/johnm/courses/dm/statminers/labs> For the full set of course materials, go to the parent directory.]

A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions. M.J.Moroney

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Examples	5
1.1.1	The Age of the Universe	5
1.1.2	Tomato plant growth – three different nutrients	5
1.1.3	Electrical resistance of fruit, vs apparent juice content	6
1.1.4	Record times for Scottish hillraces	6
1.2	Terminology	7
1.3	Least squares and maximum likelihood	7
1.4	Justification of causative interpretations	8
1.5	Ordinal and categorical outcomes	8
<b>2</b>	<b>Linear Models – Basis Functions</b>	<b>8</b>
2.1	Terms – groups of basis functions	9
2.2	Factor basis functions	9
2.3	Spline basis functions	10
<b>3</b>	<b>Solving Least Squares Systems</b>	<b>10</b>
3.1	Householder Reflections, and QR	11
3.1.1	Properties of Householder reflections	11
3.2	Least Squares	14
3.2.1	Normal equations	14
3.2.2	Computational issues	15
3.3	Demonstrations of the computational steps	17
3.4	The Analysis of Variance Table	20
3.5	Weighted Least Squares	21
<b>4</b>	<b>Model Assumptions and Model Choice</b>	<b>22</b>
4.1	Distributional Theory	22
4.1.1	Strict Requirements	22
4.1.2	Distributional Results	23
4.2	Model choice	23
4.3	The interpretation of regression parameters	24
<b>5</b>	<b>Factor and Spline Terms</b>	<b>24</b>
5.1	Factor terms	24
5.1.1	Ordered factors	24
5.2	Regression splines	25
5.2.1	How many degrees of freedom?	26
<b>6</b>	<b>Generalized Linear Models (GLMs)</b>	<b>28</b>
6.1	Brief summary of theory	29
6.2	GLMs – commentary on the theory	29
<b>7</b>	<b>Discriminant Methods</b>	<b>30</b>

<b>8</b>	<b>Validity Issues</b>	<b>30</b>
8.1	Errors in $x$	31
8.1.1	Regression with a single covariate	31
8.1.2	One covariate measured with error; others without error	31
8.1.3	Implications for variable selection	32
8.2	Missing variables and/or mis-specification of the model	32
8.2.1	Do airbags reduce risk of death in an accident	32
8.3	Model and variable selection	33
<b>9</b>	<b>Resampling Methods</b>	<b>34</b>
<b>10</b>	<b>Examples and Issues</b>	<b>35</b>
10.1	A severely constrained sample of books	35
10.2	Hill race data	35
10.2.1	Spline Terms	36
10.3	Diet-disease studies	37
10.4	Lalonde's data – effectiveness of a labour training program	38
<b>11</b>	<b>References and Further Reading</b>	<b>38</b>

# 1 Introduction

Applications of linear models are the focus of this course. Much of the discussion will apply also to non-linear models. In the sense used here, linear models are linear in the parameters, not necessarily in the variables. Initial discussion will focus on  $E[y]$ , often called the fixed part of the model.

Models in which  $E[y]$  is a linear combination of the explanatory variables are obviously linear models. These are linear both in the parameters and in the variables. This is unnecessarily restrictive. The classical theory required for such models applies providing only that models are linear in the parameters.

## 1.1 Examples

We will delay attention to the magic of how models are fitted. We'll first examine several examples of output from the fitting process.

### 1.1.1 The Age of the Universe

Here, there is a single explanatory variable:

```
library(gamair)
data(hubble)
names(hubble) <- c("Velocity", "Distance")
plot(Velocity ~ Distance, data=hubble)
hubble.lm <- lm(Velocity ~ -1 + Distance, data=hubble)
hubble.rlm <- rlm(Velocity ~ -1 + Distance, data=hubble)
```

Note the recourse to the function `rlm()`. This gives a robust fit, i.e., the effect of points that are identified as outliers is downweighted.

Note also `lqs()`, which gives a resistant fit. Points with large residuals are ignored. If the number of data points  $n$  is large relative to the total number of model parameters  $p$ , the default settings have the effect of ignoring slightly less than half of the points.

The plot gives a suggestion of curvature. This might be accommodated by including the square of the distance as a further explanatory variable, thus:

```
hubble.lm2 <- rlm(Velocity ~ -1 + Distance + I(Distance^2), data=hubble)
```

### 1.1.2 Tomato plant growth – three different nutrients

```
tomato <-
  data.frame(weight=
    c(1.5, 1.9, 1.3, 1.5, 2.4, 1.5, # water
      1.5, 1.2, 1.2, 2.1, 2.9, 1.6, # Nutrient
      1.9, 1.6, 0.8, 1.15, 0.9, 1.6), # Nutrient+24D
    trt = factor(rep(c("water", "Nutrient", "Nutrient+24D"),
      c(6, 6, 6))))
## Now make water the first level of trt. It will then appear as
## the initial level in the graphs. In aov or lm calculations, it
## will appear as the baseline or reference level.
tomato$trt <- relevel(tomato$trt, ref="water")
```

Now fit a one-way analysis of variance model:

```
tomato.lm <- aov(weight ~ 0 + trt, data=tomato)
summary(tomato.lm)
termplot(tomato.lm, partial=T, col.res="gray30")
```

The 0 is a device that determines how R chooses the parameters that describe the model. Examine `model.matrix(tomato.lm)` to see how R has set up the model.

### 1.1.3 Electrical resistance of fruit, vs apparent juice content

The following plots shows clear evidence of curvature.

```
library(DAAG)
plot(ohms ~ juice, data=fruitohms)
```

With the `hubble` data set, we could attempt to model the hint of curvature by using the square of `Distance`, as well as `Distance`, as an explanatory variable. Here, a linear combination of several curves is needed.

The following works quite well.

```
library(splines)
juice.ns3 <- lm(ohms~ns(juice, 3), data=fruitohms)
plot(ohms ~ juice, data=fruitohms)
ord <- with(fruitohms, order(juice))
lines(fitted(juice.ns3)[ord] ~ juice[ord], data=fruitohms, col=2)
coef(juice.ns3)
```

We have used a natural spline basis of degree 3. Here are the “basis” curves that were used:

```
library(lattice)
ns3 <- as.data.frame(with(fruitohms, ns(juice, 3))[, 1:3])
names(ns3) <- c("Curve_1", "Curve_2", "Curve_3")
ns3$juice <- fruitohms$juice
xyplot(Curve_1 + Curve_2 + Curve_3 ~ juice, type="l", data= ns3,
       auto.key = list(columns=3, points=FALSE, lines=TRUE))
```

Above, we saw that the coefficients for the three curves were, respectively, -4534.6, -6329.0 and -2569.0.

### 1.1.4 Record times for Scottish hillraces

We will work with logarithms of all variables. The rationale will be discussed when we later examine these data in more detail.

```
hills.loglm <- lm(log(time) ~ log(dist) + log(climb), data=hills2000)
par(mfrow=c(1,2), pty="s")
termplot(hills.loglm, partial=TRUE, smooth=panel.smooth)
```

## 1.2 Terminology

Note the distinction between fixed and random effects. In

$$y = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \dots, n$$

$\alpha$  and  $\beta$  are fixed effects, while the  $\epsilon_i$  are random effects. A common assumption is that the  $\epsilon_i$  are distributed as  $N(0, \sigma^2)$ , independently between observations. This is commonly known as the iid normal assumption.

Sequential dependence models may be used for data in which observations are sequential in time or space. The dependence between any two observations is a function of their distance apart.

The general linear model will be written, using matrix notation, as

$$\mathbf{y}_i = (\mathbf{X}\boldsymbol{\beta})_i + \epsilon_i, \quad i = 1, 2, \dots, n$$

where as a minimum it is required that  $E[\boldsymbol{\epsilon}] = \mathbf{0}$ . This is commonly strengthened to  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ , where  $\mathbf{I}$  is  $n$  by  $n$  with ones on the diagonal and zeros elsewhere.

More succinctly

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

## 1.3 Least squares and maximum likelihood

Unless justified by more fundamental considerations, least squares appears ad hoc. The more fundamental justification, from maximum likelihood, is available for data that are independently and identically distributed (iid) as normal.

Both least squares and maximum likelihood ensure that values that are predicted by the model (fitted or predicted values) are, according to one or other criterion, close to observed values. This is so even if the model is wrong. If the model is wrong because the assumed error structure is incorrect (e.g., correlated observations), the closeness measure that is used may however be inappropriate.

### Correct and incorrect models

If the model is correct, parameter estimates have certain optimality properties. These derive from the Gauss-Markov theorem, and happen because the parameter estimates are linear functions of the fitted values.

If the model is incorrect, there is no longer a guarantee that these virtues will be realized, not even approximately. Here, note malign consequences that may result from mis-specification of the fixed effects part of the model. One or more parameters may be reversed in sign, while remaining statistically significant. (If the true perpetrator is not on the list of suspects, there is a risk that others who could be found at the same places at similar times will be incriminated. The presence of one or more of these individuals may turn out to have explanatory power.)

There are interesting recent examples in the epidemiological literature that illustrate this point. Given certain common types of model failure, large biases are almost inevitable. The models that thus mislead may have substantial predictive power, at least for the population from which the data have been sampled.

Errors in explanatory variables, if they are large enough, have a similar potential to lead to biased parameter estimates. Biases may be generated in parameters other than those for the variables in which the errors appear.

## 1.4 Justification of causative interpretations

Occam’s razor may not be used in this context. On the contrary, the proper advice is to “make your hypotheses complex” (R. A. Fisher, quoted in [Cochran, 1965](#), , Section 6). Estimation of a parameter that codes for a treatment effect is an important special case, discussed in detail in [Rosenbaum \(2002\)](#).

A further step is to give a causative interpretation to one or more parameters. [Rosenbaum \(2002\)](#) notes two types of objection to causative interpretation of parameters derived from observational data – the dismissive and the tangible. One important type of tangible objection involves drawing attention to variables that were not included as explanatory variables in the regression equation. These are the sorts of “complex hypotheses” that Fisher had in mind.

## 1.5 Ordinal and categorical outcomes

Models with a 0/1 outcome are a particular case of models with a categorical outcome, otherwise known as classification or discriminant models. This special case will get some limited attention in this course. More general classification models are outside of the scope of this course, aside from some cursory discussion.

Ordinal outcomes, except to the extent that they can be handled using the same approaches as for continuous variables, are outside of the scope of this course.

# 2 Linear Models – Basis Functions

We begin by defining what we mean by a “linear model”. Define basis functions

$$\phi_1(x_1, x_2, \dots, x_k), \phi_2(x_1, x_2, \dots, x_k), \dots, \phi_p(x_1, x_2, \dots, x_k)$$

In the simplest case  $p = k$  and  $\phi_1(x_1, x_2, \dots, x_p) = x_1$ ,  $\phi_2(x_1, x_2, \dots, x_p) = x_2$ ,  $\dots$ ,  $\phi_p(x_1, x_2, \dots, x_p) = x_p$ .

Then any function with values on the real line such that

$$f(x_1, x_2, \dots, x_k) = \beta_1\phi_1(x_1, x_2, \dots, x_k) + \beta_2\phi_2(x_1, x_2, \dots, x_k) + \dots + \beta_p\phi_p(x_1, x_2, \dots, x_k)$$

where the elements of  $\beta = (\beta_1, \beta_2, \dots, \beta_p)$  are the only unknowns, specifies a linear model. In words, any model in which  $E[y]$  is a linear combination of the  $\phi_i$  is a linear model.

Note three important non-trivial special cases:

- Several  $\phi_i$ ’s are defined that together account for a factor term. See below for the definition of “term”.
- Several  $\phi_i$ ’s are defined that together account for a spline term.
- Interaction terms may be created by multiplying together columns that relate to other terms in the model.

The model is linear in the values that the  $\phi$ ’s take on the sample data. It is not, in general, linear in the  $x_i$ ’s.

## 2.1 Terms – groups of basis functions

A factor with  $k$  levels requires, assuming that the model already has a constant term,  $k - 1$  basis elements to represent it. These basis elements together constitute a “term”, in this case a factor. The Wilkinson & Rogers notation makes it possible to specify the factor as a term in the model, leaving code that is called by R’s `lm()` function to determine the basis functions that are needed.

Similarly a 4 degree of freedom spline term requires four basis elements. Again, code that is called by R’s `bs()` (B-splines) or `ns()` (natural splines) function will determine the basis functions.

Quite generally, the basis functions  $\phi_1, \phi_2, \dots, \phi_p$  can be categorized into groups, with one group for each term the model, thus:

$$\underbrace{\phi_1, \dots, \phi_{m_1}}_{\text{Term1}}, \underbrace{\phi_{m_1+1}, \dots, \phi_{m_2}}_{\text{Term2}}, \dots$$

## 2.2 Factor basis functions

A simple example will do for now. Consider the `sugar` data frame in the R package. There are three levels of `trt` – `Control`, `A`, `B` and `C`. Type into R:

```
> contrasts(sugar$trt)
      A B C
Control 0 0 0
A       1 0 0
B       0 1 0
C       0 0 1
```

The first factor level, ie `Control`, is taken as the baseline. The parameter associated with `A` is then its difference from the baseline, and similarly for `B` and `C`.

Now see how this works in practice:

```
> sugar.lm <- lm(weight ~ trt, data=sugar)
> model.matrix(sugar.lm)
  (Intercept) trtA trtB trtC
1           1    0    0    0
2           1    0    0    0
3           1    0    0    0
4           1    1    0    0
5           1    1    0    0
6           1    1    0    0
7           1    0    1    0
8           1    0    1    0
9           1    0    1    0
10          1    0    0    1
11          1    0    0    1
12          1    0    0    1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$trt
```

```
[1] "contr.treatment"
```

In the above, (Intercept) is really Control. The columns trtA, trtB and trtC measure differences from Control

Another possibility is

```
> sugar.lm0 <- lm(weight ~ -1 + trt, data=sugar)
> model.matrix(sugar.lm0)
  trtControl trtA trtB trtC
1          1  0  0  0
2          1  0  0  0
3          1  0  0  0
4          0  1  0  0
5          0  1  0  0
6          0  1  0  0
7          0  0  1  0
8          0  0  1  0
9          0  0  1  0
10         0  0  0  1
11         0  0  0  1
12         0  0  0  1
attr(,"assign")
[1] 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$trt
[1] "contr.treatment"
```

Omission of the “constant term” from the formula forces the use of one parameter for each different factor level. With this parameterization no restriction on the parameters is needed.

### 2.3 Spline basis functions

We will come to these later. Spline basis functions are, in essence, a construction kit for constructing curves that have some predefined flexibility.

## 3 Solving Least Squares Systems

This section will review important computational and theoretical aspects of least squares. Least squares may be used because it seems intuitively sensible. Or it may be justified by maximum likelihood, assuming independently and identically distributed normal errors.

The QR decomposition, usually achieved by a sequence of Householder reflections, is widely used in the practical computer solution of linear least squares systems. It is also a good basis for the derivation of important results in least squares theory. An understanding of the computational details can be important when efficient computation is important, e.g., a computation is repeated a large number of times.

Other methods are available that can be more efficient in particular types of problem, e.g., for sparse systems. For very large linear least squares systems,

it may be important to use one of these alternative methods. See for example [Bates \(2006\)](#); [Koenker and Ng \(2003\)](#).

### 3.1 Householder Reflections, and QR

Given an  $n \times p$  matrix  $\mathbf{X}$ , the QR decomposition derives an orthogonal matrix  $\mathbf{Q}$  and an upper triangular matrix  $\mathbf{R}$  such that

$$\mathbf{Q}'\mathbf{X} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

Then

$$\mathbf{X} = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

In the sequel, it will be shown that the decomposition can be achieved by constructing a sequence of reflections that successively reduce to zero below diagonal elements in columns 1, 2,  $\dots$   $p$  of  $\mathbf{X}$ .

A Householder reflection of a vector is achieved by pre-multiplication by a Householder matrix, ie a matrix of the form:

$$\mathbf{H} = \mathbf{I} - \gamma \mathbf{u}\mathbf{u}', \quad \text{where} \quad \gamma = \frac{2}{\|\mathbf{u}\|^2}$$

In least squares applications, with a model matrix  $\mathbf{X}$  and vector of observations  $\mathbf{y}$ , a sequence of Householder reflections is used to reduce  $\mathbf{X}$  to upper triangular form. The same sequence of reflections is applied also to  $\mathbf{y}$ . The first reflection yields  $\mathbf{H}_1\mathbf{X}$ . The second reflection, which operates only on rows of  $\mathbf{H}_1\mathbf{X}$  subsequent to the first, replaces below diagonal elements in the second column with zeros, yielding  $\mathbf{H}_2\mathbf{H}_1\mathbf{X}$ .

Figure 1 shows diagrammatically the sequence of Householder reflections, applied to a  $9 \times 4$  model matrix  $\mathbf{X}$ , for reducing a  $9 \times 4$  model matrix  $\mathbf{X}$  to upper triangular form.

Before proceeding, note important properties of Householder reflections.

#### 3.1.1 Properties of Householder reflections

1. A Householder matrix is orthogonal, ie,  $\mathbf{H}'\mathbf{H} = \mathbf{I}$ .
2. The product of two Householder matrices is an orthogonal matrix.
3. For a suitable choice of  $\mathbf{H}$

$$\mathbf{H}\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ x_{k-1} \\ \eta_k \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

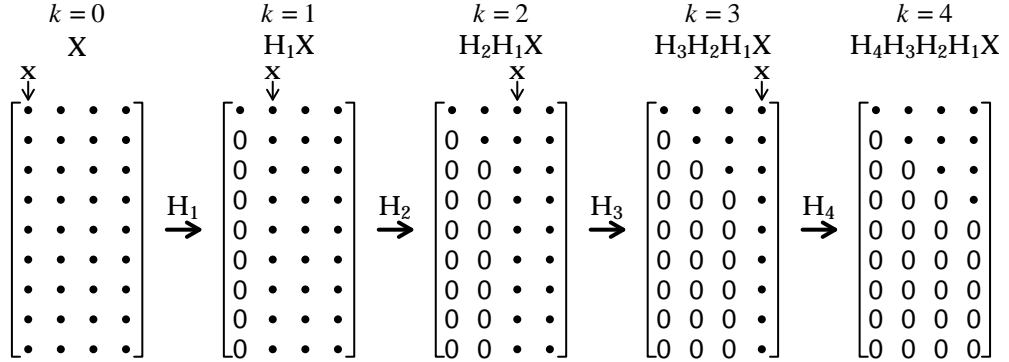


Figure 1: Diagrammatic representation of the sequence of steps that are applied to a  $9 \times 4$  model matrix  $\mathbf{X}$  when Householder reflections are used in least squares calculations. The vector  $\mathbf{x}$  whose final  $n - k$  elements are used in forming  $\mathbf{u}_k$  and hence  $\mathbf{H}_k$  is identified, for each of the successive steps, with an arrow ( $\downarrow$ ). Thus  $\mathbf{H}_1$  is formed using the first column of  $\mathbf{X}$  ( $k = 0$ ),  $\mathbf{H}_2$  is formed using the second and later elements in the second column of  $\mathbf{H}_1\mathbf{X}$  ( $k = 1$ ),  $\mathbf{H}_3$  is formed using the third and later elements in the third column of  $\mathbf{H}_2\mathbf{H}_1\mathbf{X}$  ( $k = 2$ ), and so on. The same sequence of Householder reflections is applied to  $\mathbf{y}$ .

The proof of (1) is a straightforward use of matrix algebra.

For (2), observe that

$$\begin{aligned}
 (\mathbf{H}_2\mathbf{H}_1)'(\mathbf{H}_2\mathbf{H}_1) &= \mathbf{H}_1'\mathbf{H}_2'\mathbf{H}_2\mathbf{H}_1 \\
 &= \mathbf{H}_1'(\mathbf{H}_2'\mathbf{H}_2)\mathbf{H}_1 \\
 &= \mathbf{H}_1'\mathbf{I}\mathbf{H}_1 \\
 &= \mathbf{H}_1'\mathbf{H}_1
 \end{aligned}$$

For (3), observe that

$$(\mathbf{I} - \gamma\mathbf{u}\mathbf{u}')\mathbf{x} = \mathbf{x} - \mathbf{u}\gamma(\mathbf{u}'\mathbf{x})$$

Consider the case  $k = 1$ . We set  $u_i = x_i$  ( $i = 2, \dots, n$ ), and require that  $\gamma(\mathbf{u}'\mathbf{x}) = 1$ . The elements  $x_i, i = 2, \dots, n$  will then be replaced by zeros.

We ensure that  $\gamma(\mathbf{u}'\mathbf{x}) = 1$  by a suitable choice of  $u_1$ ; it is convenient to write  $u_1 = x_1 + \alpha$ .

Thus we have

$$\mathbf{u} = \begin{bmatrix} x_1 + \alpha \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

Then

$$\gamma\mathbf{u}'\mathbf{x} = \frac{2}{\|\mathbf{u}\|^2}(\|\mathbf{x}\|^2 + \alpha x_1)$$

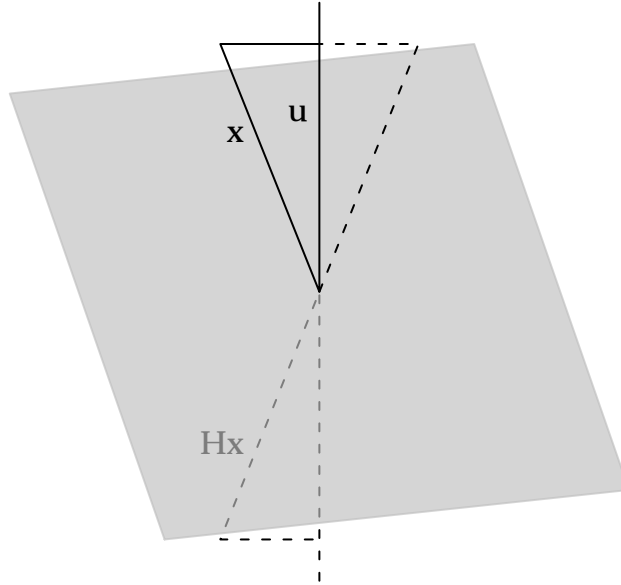


Figure 2: Geometrical representation of the reflection of  $\mathbf{x}$  in the plane that is orthogonal to  $\mathbf{u}$ . Here  $\mathbf{H} = I - \gamma \mathbf{u} \mathbf{u}'$ , with  $\mathbf{u} = x_1 + \text{sgn}(x_1) \|\mathbf{x}\|$ .

For this to equal 1

$$\begin{aligned} 2(\|\mathbf{x}\|^2 + \alpha x_1) &= \|\mathbf{u}\|^2 \\ &= \|\mathbf{x}\|^2 + 2\alpha x_1 + \alpha^2 \end{aligned}$$

Thus

$$\alpha^2 = \|\mathbf{x}\|^2 \quad \alpha = \pm \|\mathbf{x}\|$$

Figure 2 gives a geometrical representation of the reflection of  $\mathbf{x}$  in the plane that is orthogonal to  $\mathbf{u}$ . Note that  $\mathbf{H}\mathbf{x}$  is a positive or negative multiple  $(1, 0, \dots, 0)$ .

For numerical stability, we require  $\alpha = \text{sgn}(x_1) \|\mathbf{x}\|$ , thus ensuring that  $x_1$  and  $\alpha$  have the same sign. It follows that

$$\|\mathbf{u}\|^2 = 2\alpha(x_1 + \alpha)$$

$$\mathbf{u}'\mathbf{x} = \alpha(x_1 + \alpha)$$

$$\begin{aligned} \gamma &= \frac{2}{\|\mathbf{u}\|^2} \\ &= \frac{1}{\alpha(x_1 + \alpha)} \end{aligned}$$

More generally, choose

$$\mathbf{u}_{(k)} = \begin{bmatrix} 0 \\ \cdot \\ 0 \\ x_k + \alpha_k \\ x_{k+1} \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

where

$$\alpha_k = \operatorname{sgn}(x_k) \sqrt{\sum_{i=k}^n x_i^2}$$

Then

$$\mathbf{H}_k = \mathbf{I} - \gamma \mathbf{u}_{(k)} \mathbf{u}'_{(k)}, \quad \text{where} \quad \gamma = \frac{2}{\|\mathbf{u}_{(k)}\|^2} = \frac{1}{\alpha(x_k + \alpha)}$$

The sequence of reflections  $\mathbf{H}_1, \mathbf{H}_2, \dots$ , with  $\mathbf{x}$  chosen as indicated in Figure 1, then achieves the result that is required for the QR reduction.

## 3.2 Least Squares

Form

$$\mathbf{Q}'\mathbf{X} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{Q}'\mathbf{y} = \begin{bmatrix} \mathbf{f} \\ \mathbf{r} \end{bmatrix}$$

where  $\mathbf{R}$  ( $p \times p$ ) is upper triangular,  $\mathbf{0}$  is an  $(n-p) \times p$  array of zeros,  $\mathbf{f}$  is  $p \times 1$  and  $\mathbf{r}$  is  $(n-p) \times 1$ . Then

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 &= \|\mathbf{Q}'\mathbf{y} - \mathbf{Q}'\mathbf{X}\mathbf{b}\|^2 \\ &= \left\| \begin{bmatrix} \mathbf{f} - \mathbf{R}\mathbf{b} \\ \mathbf{r} \end{bmatrix} \right\|^2 \\ &= \|\mathbf{f} - \mathbf{R}\mathbf{b}\|^2 + \|\mathbf{r}\|^2 \end{aligned}$$

Hence  $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$  is minimized by choosing  $\mathbf{b}$  such that

$$\mathbf{R}\mathbf{b} = \mathbf{f}$$

Solve first for  $b_p$ , then for  $b_{p-1}, \dots, b_1$ , in a procedure called *backward elimination*. The residual sum of squares is  $\|\mathbf{r}\|^2$ .

### 3.2.1 Normal equations

We have

$$\mathbf{R}\mathbf{b} = \mathbf{f} = \mathbf{f}$$

Then

$$\mathbf{R}'\mathbf{R}\mathbf{b} = \mathbf{R}'\mathbf{f}$$

i.e.

$$\mathbf{R}'\mathbf{R}\mathbf{b} = \mathbf{R}'\mathbf{f}$$

Observe that

$$\mathbf{R}'\mathbf{f} = \begin{bmatrix} \mathbf{R}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{r} \end{bmatrix} = \mathbf{X}'\mathbf{Q}\mathbf{Q}'\mathbf{y} = \mathbf{X}'\mathbf{y}$$

and that

$$\mathbf{R}'\mathbf{R}\mathbf{b} = \mathbf{X}'\mathbf{Q}\mathbf{Q}'\mathbf{X} = \mathbf{X}'\mathbf{X}$$

Hence the so-called normal equations

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

Writing  $\mathbf{X}\mathbf{b} = \boldsymbol{\mu}$  (the fitted values), the normal equations can then be written

$$\mathbf{X}'\boldsymbol{\mu} = \mathbf{X}'\mathbf{y}$$

This form of the least squares equations carries over to generalized linear models, in the special case where the canonical link is specified.

### 3.2.2 Computational issues

1. Algebraically, one can write

$$\mathbf{Q}'\mathbf{X} = \mathbf{H}_p\mathbf{H}_{p-1}\dots\mathbf{H}_1\mathbf{X}$$

Neither  $\mathbf{Q}'$  nor any of the matrices  $\mathbf{H}_i$  is ever formed explicitly. For large matrices, this would be an impossibly computationally expensive way to do the computations. In

$$\mathbf{H}\mathbf{X} = (\mathbf{I} - \gamma\mathbf{u}\mathbf{u}')\mathbf{X}$$

it is important to do the calculation as

$$\mathbf{X} - \gamma\mathbf{u}(\mathbf{u}'\mathbf{X})$$

Thus for each column  $\mathbf{x}_j$  of  $\mathbf{X}$ , first form  $\gamma\mathbf{u}'\mathbf{x}_j$ , then subtract this multiple of  $\mathbf{u}$  from  $\mathbf{x}_j$ .

2. For calculation of row  $k$  of the array resulting from pre-multiplication by  $\mathbf{H}_k$ , formulae are available that are simplified and numerically more accurate than from direct use of equation 1. A further simplification is to form the diagonal element as  $|\alpha_k|$ , multiplying remaining elements in the  $i$ th row by  $\text{sgn}(\alpha_k)$ . In the examples shown below, sequences of such modified Householder reflections have been used. We no longer have Householder reflections, but the  $\mathbf{H}_i$  are, just as before, orthogonal matrices. The difference is that when  $\alpha_k$  is negative, all elements in row  $k$  change sign relative to the Householder matrix  $\mathbf{H}_k$ .
3. If the calculated diagonal element is zero to within machine precision, the easiest recourse is to set all elements in the row to zero. The vector  $\mathbf{b}$  is no longer uniquely defined. The residual sum of squares is unique, independent of the action that may be taken to resolve the indeterminacy in the parameters.

4. Let  $\mathbf{X}_k$  be the matrix that consists of the first  $k$  columns of  $\mathbf{X}$ . Then the information needed to minimize

$$\|\mathbf{y} - \mathbf{X}_k \mathbf{b}\|^2$$

is available once the first  $k$  Householder steps are complete. It is found in the  $k \times k$  submatrix of  $\mathbf{R}$  and in the first  $k$  elements of  $\mathbf{Q}'\mathbf{y}$ .

### 3.3 Demonstrations of the computational steps

#### Example 1

We will start with

$$\mathbf{X} = \begin{bmatrix} 1 & -6 & 0 \\ 1 & -3 & 6 \\ 1 & 6 & 15 \\ 1 & 21 & 9 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -3 \\ 1 \\ 2 \\ 6 \end{bmatrix}$$

The aim is to minimize

$$[-3-(b_0-6b_1)]^2 + [1-(b_0-3b_1+6b_2)]^2 + [2-(b_0+6b_1+15b_2)]^2 + [6-(b_0+21b_1+9b_2)]^2$$

The sequence of sweeps is

$$\begin{bmatrix} 1 & -6 & 0 & -3 \\ 1 & -3 & 6 & 1 \\ 1 & 6 & 15 & 2 \\ 1 & 21 & 9 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{2} & \mathbf{9} & \mathbf{15} & \mathbf{3} \\ 1 & -4 & 1 & 1 \\ 1 & 5 & 10 & 2 \\ 1 & 20 & 4 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{2} & \mathbf{9} & \mathbf{15} & \mathbf{3} \\ 1 & \mathbf{21} & \mathbf{6} & \mathbf{6} \\ 1 & 5 & 9 & 1 \\ 1 & 20 & 0 & 2 \end{bmatrix}$$

The first sweep is equivalent to pre-multiplication by  $\mathbf{H}_1$ , and the second to pre-multiplication by  $\mathbf{H}_2$ . At this point, the element in the (4,3) position is zero, and use of  $\mathbf{H}_3$  is not needed. ( $\mathbf{H}_3$ , if it were formed, would be the identity matrix.)

Notice that we have not bothered to set below diagonal elements to zero, once calculations for column  $i$  are complete. Instead, we have printed them in small type. In practice, it is useful to leave them in place. They can be used to reconstruct the sequence of Householder steps.

#### Use of R

```
X1df <- data.frame(X1=c(-6,-3,6,21), X2=c(0,6,15,9), y=c(-3,1,2,6))
lm(y ~ X1+X2, data=X1df)
summary(lm(y ~ X1+X2, data=X1df)) # Gives more information
anova(lm(y ~ X1+X2, data=X1df)) # Gives different information
lm(y ~ X1+X2, data=X1df)$qr # qr decomposition
```

Observe that R stores somewhat different information in below diagonal positions.

Alternatively, R has a suite of functions that can be used for the underlying computations of least squares calculations. These include `qr()`, `qr.coef()` and `qr.solve()`. Because they work with matrices rather than data frames, they may be much faster than `lm()` for the handling of large least squares problems. Approaches to the calculations that use these will be demonstrated for the second slightly more substantial example that will now be presented.

## Example 2

Table 1 shows the application of a sequence of 4 modified Householder reflections to a  $9 \times 5$  array

$$(\mathbf{X}, y) = \begin{bmatrix} \text{X0} & \text{X1} & \text{X2} & \text{X3} & y \\ 1 & 1 & -1 & -14 & -7.7 \\ 1 & -1 & 6 & -2 & 8.5 \\ 1 & 1 & 9 & 8 & 19.6 \\ 1 & -2 & 8 & -3 & 11.4 \\ 1 & 0 & 5 & 1 & 11.4 \\ 1 & 0 & 3 & -1 & 6.8 \\ 1 & 4 & 2 & 9 & 2.5 \\ 1 & 7 & 0 & 8 & -1.6 \\ 1 & 8 & -5 & 3 & -13 \end{bmatrix}$$

The numbers in the X-matrix have been chosen so that the values in the R-matrix are integers. The least squares system is for a model in which there is a constant term X0 and explanatory variables X1, X2 and X3.

The least squares estimates are obtained by solving

$$\begin{bmatrix} \mathbf{3} & \mathbf{6} & \mathbf{9} & \mathbf{3} \\ \mathbf{0} & \mathbf{10} & \mathbf{-10} & \mathbf{10} \\ \mathbf{0} & \mathbf{0} & \mathbf{8} & \mathbf{16} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{8} \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{12.63} \\ \mathbf{-20.04} \\ \mathbf{19.95} \\ \mathbf{4} \end{bmatrix}$$

Observe the subsystems for which information is available

$$\begin{bmatrix} \mathbf{3} \end{bmatrix} \mathbf{12.63} \leftarrow \begin{array}{l} \text{Solve } 3b_0 = 12.63 \Rightarrow b_0 = 4.21. \\ \text{SS reduces by } 12.63^2 \text{ (to CSS}^a \text{)} \end{array}$$

<sup>a</sup>SS = sum of squares; RSS = residual SS; CSS = corrected SS, i.e., RSS about mean

$$\begin{bmatrix} \mathbf{3} & \mathbf{6} \\ \mathbf{0} & \mathbf{10} \end{bmatrix} \begin{array}{l} \mathbf{12.63} \\ \mathbf{-20.04} \end{array} \leftarrow \begin{array}{l} 10b_1 = -20.04 \Rightarrow b_1 = -2.004; 3b_0 + \\ 6b_1 = 12.63 \Rightarrow b_0 = 8.22. \\ \text{Additional reduction in RSS} = 20.04^2. \end{array}$$

$$\begin{bmatrix} \mathbf{3} & \mathbf{6} & \mathbf{9} \\ \mathbf{0} & \mathbf{10} & \mathbf{-10} \\ \mathbf{0} & \mathbf{0} & \mathbf{8} \end{bmatrix} \begin{array}{l} \mathbf{12.63} \\ \mathbf{-20.04} \\ \mathbf{19.95} \end{array} \leftarrow \begin{array}{l} 8b_2 = 19.95; 10b_1 - 10b_2 = -20.04; 3b_0 + 6b_1 + \\ 9b_2 = 12.63. \\ \text{Additional reduction in RSS} = 19.95^2. \end{array}$$

In the full model, the RSS reduces by a further  $4^2$ . The residual sum of squares from the full model is

$$1.46^2 + 0.68^2 + 4.26^2 + 1.11^2 + 1.58^2 = 24.46$$

(Note that 24.46 is the result of rounding the RSS when it is calculated to full machine accuracy.)

$$\begin{array}{c}
\begin{bmatrix} 1 & 1 & -1 & -14 \\ 1 & -1 & 6 & -2 \\ 1 & 1 & 9 & 8 \\ 1 & -2 & 8 & -3 \\ 1 & 0 & 5 & 1 \\ 1 & 0 & 3 & -1 \\ 1 & 4 & 2 & 9 \\ 1 & 7 & 0 & 8 \\ 1 & 8 & -5 & 3 \end{bmatrix} \begin{array}{l} 7.7 \\ 8.5 \\ 19.6 \\ 11.4 \\ 11.4 \\ 6.8 \\ 2.5 \\ -1.6 \\ -13.0 \end{array} \rightarrow \begin{bmatrix} \mathbf{3} & \mathbf{6} & \mathbf{9} & \mathbf{3} \\ 1 & -2.75 & 4 & 0.75 \\ 1 & -0.75 & 7 & 10.75 \\ 1 & -3.75 & 6 & -0.25 \\ 1 & -1.75 & 3 & 3.75 \\ 1 & -1.75 & 1 & 1.75 \\ 1 & 2.25 & 0 & 11.75 \\ 1 & 5.25 & -2 & 10.75 \\ 1 & 6.25 & -7 & 5.75 \end{bmatrix} \begin{array}{l} \mathbf{12.63} \\ 7.27 \\ 18.37 \\ 10.17 \\ 10.17 \\ 5.57 \\ 1.27 \\ -2.83 \\ -14.23 \end{array} \\
\\
\begin{array}{c} \rightarrow \begin{bmatrix} \mathbf{3} & \mathbf{6} & \mathbf{9} & \mathbf{3} \\ 1 & \mathbf{10} & \mathbf{-10} & \mathbf{10} \\ 1 & -0.75 & 6.18 & 11.29 \\ 1 & -3.75 & 1.88 & 2.47 \\ 1 & -1.75 & 1.08 & 5.02 \\ 1 & -1.75 & -0.92 & 3.02 \\ 1 & 2.25 & 2.47 & 10.12 \\ 1 & 5.25 & 3.76 & 6.94 \\ 1 & 6.25 & -0.14 & 1.22 \end{bmatrix} \begin{array}{l} \mathbf{12.63} \\ \mathbf{-20.04} \\ 16.76 \\ 2.14 \\ 6.42 \\ 1.82 \\ 6.09 \\ 8.41 \\ -0.85 \end{array} \rightarrow \begin{bmatrix} \mathbf{3} & \mathbf{6} & \mathbf{9} & \mathbf{3} \\ 1 & \mathbf{10} & \mathbf{-10} & \mathbf{10} \\ 1 & -0.75 & \mathbf{8} & \mathbf{16} \\ 1 & -3.75 & 1.88 & -1.15 \\ 1 & -1.75 & 1.08 & 2.94 \\ 1 & -1.75 & -0.92 & 4.79 \\ 1 & 2.25 & 2.47 & 5.36 \\ 1 & 5.25 & 3.76 & -0.31 \\ 1 & 6.25 & -0.14 & 1.48 \end{bmatrix} \begin{array}{l} \mathbf{12.63} \\ \mathbf{-20.04} \\ \mathbf{19.95} \\ -2.74 \\ 3.63 \\ 4.21 \\ -0.31 \\ -1.34 \\ -0.49 \end{array} \\
\\
\begin{array}{c} \rightarrow \begin{bmatrix} \mathbf{3} & \mathbf{6} & \mathbf{9} & \mathbf{3} \\ 1 & \mathbf{10} & \mathbf{-10} & \mathbf{10} \\ 1 & -0.75 & \mathbf{8} & \mathbf{16} \\ 1 & -3.75 & 1.88 & \mathbf{8} \\ 1 & -1.75 & 1.08 & 2.94 \\ 1 & -1.75 & -0.92 & 4.79 \\ 1 & 2.25 & 2.47 & 5.36 \\ 1 & 5.25 & 3.76 & -0.31 \\ 1 & 6.25 & -0.14 & 1.48 \end{bmatrix} \begin{array}{l} \mathbf{12.63} \\ \mathbf{-20.04} \\ \mathbf{19.95} \\ \mathbf{4} \\ 1.46 \\ 0.68 \\ -4.26 \\ -1.11 \\ -1.58 \end{array} \end{array}
\end{array}$$

Table 1: Reduction of a  $9 \times 4$  array to upper triangular form, with the same operations applied to the fifth column  $\mathbf{y}$ . Four modified Householder reflections are applied to the  $9 \times 5$  array  $(\mathbf{X}, \mathbf{y})$ .

## Use of R

```
X2df <- data.frame(X1=c(1, -1, 1, -2, 0, 0, 4, 7, 8),
                  X2=c(-1, 6, 9, 8, 5, 3, 2, 0, -5),
                  X3=c(-14, -2, 8, -3, 1, -1, 9, 8, 3),
                  y=c(-7.7, 8.5, 19.6, 11.4, 11.4, 6.8, 2.5, -1.6, -13))
lm(y ~ X1+X2+X3, data=X2df)
summary(lm(y ~ X1+X2+X3, data=X2df)) # Gives more information
anova(lm(y ~ X1+X2+X3, data=X2df))   # Gives different information
lm(y ~ X1+X2+X3, data=X2df)$qr      # qr decomposition
```

## Alternatives to lm()

For computationally intensive least squares calculations, consider the use of `qr()` and associated functions. For use of this approach, the user must first extract or construct the design matrix. The following demonstrates the use of `qr()` and allied functions:

```
X <- with(X2df, model.matrix(~ X1+X2+X3))
## X <- cbind(rep(1,9), X2df[, 1:3]) # More efficient alternative
QR <- qr(X)
qr.coef(QR, X2df$y)
qr.resid(QR, X2df$y)
## etc, etc
```

Calculations may be computationally intensive because a major component of the calculation is repeated a large number of times, and/or because they involve one or more large design matrices.

## 3.4 The Analysis of Variance Table

We show the sequential analysis of variance table that is relevant to Example 2 in Subsection 3.3. From Table 1, we see that

$$\mathbf{Q}'\mathbf{y} = \begin{bmatrix} 12.63 \\ -20.04 \\ 19.95 \\ 4 \\ 1.46 \\ 0.68 \\ -4.26 \\ -1.11 \\ -1.58 \end{bmatrix}$$

The usual form of sequential analysis of variance table, which partitions the sum of squares about the mean, is obtained by picking off the elements of  $\mathbf{Q}'\mathbf{y}$

in turn:

Term	Sum of squares	DF
X1	20.04 <sup>2</sup>	1
X2	19.95 <sup>2</sup>	1
X3	4 <sup>2</sup>	1
Residual	24.46	$n - p$

Compare this with

```
X2df.lm <- lm(y ~ X1+X2+X3, data=X2df)
anova(X2df.lm)
```

Now observe how this table can be obtained from the output of `qr()` and friends.

```
## Now extract the table from the successive components of Qy
QR <- qr(cbind(rep(1,9), as.matrix(X2df[,1:3])))
Qty <- qr.qty(QR, X2df$y)
duetoSS <- Qty[2:4]^2 # Why is Qty[1] omitted?
rss <- sum(Qty[-(1:4)]^2)
aovtab <- data.frame(row.names=c("X1", "X2", "X3", "Residual"),
  Df=c(rep(1,3), length(Qty)-4), SS=c(duetoSS, rss))
aovtab
```

It is important to note that this is a sequential analysis of variance table. In general the contribution of each term depends on which (if any) term(s) precede it in the model.

### 3.5 Weighted Least Squares

If  $\text{var}[\mathbf{y}] = \mathbf{V}$  is known to within the scale factor  $\sigma^2$ , then a linear transformation of the vector  $\mathbf{y}$  can be determined such that elements of the linearly transformed vector are iid.

A suitable linear transformation is conveniently obtained from the Cholesky decomposition of  $\mathbf{V}$ , by which

$$\mathbf{V} = \mathbf{L}\mathbf{L}'$$

Then

$$\begin{aligned}\text{var}[\mathbf{L}^{-1}\mathbf{y}] &= \mathbf{L}^{-1}\text{var}[\mathbf{y}]\mathbf{L}'^{-1} \\ &= \mathbf{L}^{-1}\mathbf{V}\mathbf{L}'^{-1} \\ &= \mathbf{I}_n\sigma^2\end{aligned}$$

Then we minimize

$$\|\mathbf{L}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{b})\|^2$$

For this, replace  $\mathbf{y}$  by  $\mathbf{y}^* = \mathbf{L}^{-1}\mathbf{y}$ , and  $\mathbf{X}$  by  $\mathbf{X}^* = \mathbf{L}^{-1}\mathbf{X}$ , and proceed as before.

The normal equations become

$$\begin{aligned}\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{W}\mathbf{y} \\ \text{i.e., } \mathbf{X}^*\mathbf{X}^*\mathbf{b} &= \mathbf{X}^*\mathbf{y}^*\end{aligned}$$

where  $\mathbf{W} = \mathbf{V}^{-1}$

Then setting  $\mathbf{X}\mathbf{b} = \boldsymbol{\mu}$  (the fitted values), this becomes

$$\mathbf{X}'\mathbf{W}\boldsymbol{\mu} = \mathbf{X}'\mathbf{W}\mathbf{y}$$

This is form of the equations that carries across to Generalized Linear Models.

## 4 Model Assumptions and Model Choice

This section will discuss the the classical theory, and note implications for model choice. Finally, recourses will be noted that can be used when the classical theory fails or is in doubt.

### 4.1 Distributional Theory

The classical theory that will now be described assumes that conditional on  $\mathbf{X}$

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n\sigma^2) \quad (1)$$

#### 4.1.1 Strict Requirements

According to the strict requirements of the theory, the model must be known in advance. Neither data based choice of transformation(s) nor variable selection are permitted. It is however permissible to divide the data set into two parts – a training set that is used to develop the model, and a test set that is used to derive the eventual model fit. Or three parts may be required – a training set, a holdout set on which which performance is optimized in order to tune the model, and a test set that is used for the eventual model fit. An objection is that this makes poor use of the data, which may be a serious issue for smallish data sets. Matters can be improved somewhat by repeating the process, with the roles of training, test and any holdout set interchanged.

With respect to Equation 1, note that:

- Depending on how the fitted model will be used, the normality assumption is commonly not of critical importance; approximate normality is enough. Independence is more crucial, and less open to checking. If the form of departure from independence is known or can be guessed, the dependence can be modelled. This leads into areas (time series modeling, multi-level models, repeated measures, etc.) whose details are beyond the scope of the present course.
- In practice limited use of plots or other mechanisms to determine appropriate transformations may be essential, to get a model fit that does not do violence to the data. Depending on the details of the specific model and associated data, this may not not seriously invalidate assumptions. Limited variable selection, e.g., choose two explanatory variables out of four, may be similarly acceptable. Again, much will depend on the details of the model and associated data. Selection effects are in general a more serious problem than effects from limited use of data-based transformations.
- Where one model emerges as very substantially superior to other models considered, selection effects are not an issue. Why?

When the model is fitted to the data used to derive the model, the effect is anti-conservative. Thus, standard errors will be smaller than indicated by the theory, and coefficients and  $t$ -statistics larger. Such anti-conservative estimates of standard errors and other statistics may, unless the bias is huge, nevertheless provide the useful guidance

Recourses that are available when the classical theory fails will be discussed below.

#### 4.1.2 Distributional Results

Observe that

$$\text{var}[\mathbf{Q}\mathbf{y}] = \mathbf{Q}'\mathbf{I}_n\mathbf{Q}\sigma^2 = \mathbf{I}_n\sigma^2$$

Then

$$\mathbb{E} \begin{bmatrix} \mathbf{f} \\ \mathbf{r} \end{bmatrix} = \mathbb{E}[\mathbf{Q}'\mathbf{y}] = \mathbf{Q}'\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} \boldsymbol{\beta}$$

Thus,

$$\mathbf{f} \sim N(\mathbf{R}\boldsymbol{\beta}, \mathbf{I}_p\sigma^2), \quad \mathbf{r} \sim N(\mathbf{0}, \mathbf{I}_{n-p}\sigma^2)$$

Moreover  $\|\mathbf{f}\|^2$  and  $\|\mathbf{r}\|^2$  are each sums of independent  $\sigma^2\chi_1^2$  terms and therefore independent.

Finally  $\mathbf{b} = \mathbf{R}^{-1}\mathbf{f}$  is distributed as

$$N(\boldsymbol{\beta}, \mathbf{R}^{-1}\mathbf{R}'^{-1}\sigma^2), \quad \text{i.e., as } N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$$

As we will use an estimate of  $\sigma^2$  that has  $n - p$  degrees of freedom, individual elements of  $\mathbf{b}$  are distributed as  $t_{n-p}$ . Elements  $b_i$  of  $\mathbf{b}$  are independent if and only if  $\mathbf{X}'\mathbf{X}$  is diagonal.

## 4.2 Model choice

Note first that it is important whether the aim is the derivation of a predictive model, or whether the hope is to obtain interpretable regression coefficients. The derivation of interpretable regression coefficients may be difficult or impossible. More is required than to obtain a model that is a good fit to the data.

If scientific understanding suggests a suitable model, at least to within use of one or other transformation(s), this model should be investigated as a starting point. As noted above, limited data snooping to determine suitable transformation(s), and/or possible modification by deletion or addition of a limited number of variables, may be acceptable.

Recourses when the number of potential explanatory variables is large and there is recourse to variable selection are:

- The training/test set approach;
- Cross-validation or the bootstrap, with the selection process repeated at each cross-validation fold, or for each new bootstrap sample;
- Use some form of variable selection, recognizing that the classical theory will then give lower bounds for SEs of parameter estimates.

Another possibility is to fit a penalized version of the “full” model; e.g., use ridge regression;

### 4.3 The interpretation of regression parameters

Do not lightly assume that the regression answers the question(s) of interest. Issues here include:

1. All relevant explanatory variables must be included. They must appear in the “correct” form, e.g., use a logarithmic transformation or spline term if this is needed to give the correct model.
2. Results are conditional on the observed  $x$ . If any of the  $x$ -variables in a regression are observed with error (“measurement error”), the regression coefficients for all variables, both those measured with error and those measured without error, may be misleading as estimates of the regression coefficients that are of interest.

These issues will be further discussed in Section 8.

## 5 Factor and Spline Terms

### 5.1 Factor terms

See the R help pages for `contr.treatment()`, `contr.sum()`, `contr.SAS()` and `contr.poly()`.

Try the following

```
## The following is the default
with(sugar, C(trt, contr.treatment))
sugar.lm <- lm(weight ~ C(trt, contr.treatment), data=sugar)
sugar.lm
model.matrix(sugar.lm)
```

1. Repeat, replacing `contr.treatment` with `contr.SAS`. and reconcile the two sets of parameter estimates.
2. Replace with `contr.sum` and reconcile with the parameter estimates from `contr.treatment` and `contr.SAS`.

See further [Maindonald & Braun \(2007, Section 7.1, pp. 219-223\)](#)

#### 5.1.1 Ordered factors

Enter

```
tinting$tint      # tinting is in DAAG
```

The values in the column are followed with

```
Levels: no < lo < hi
```

The default contrasts are polynomial contrasts (`contr.poly`). Observe `with(tinting, C(tinting$tint, contr.poly))`

The "contrasts" attribute is

```
attr("contrasts")
      .L      .Q
no -7.071068e-01  0.4082483
lo -7.850462e-17 -0.8164966
hi  7.071068e-01  0.4082483
Levels: no < lo < hi
```

This equals

```
      .L      .Q
no      -1      1
lo       0     -2
hi       1      1
  Xply by 0.707    0.408
```

Taking levels as equally spaced, the first column accounts for a linear change as one moves from "no" to "lo" to "hi", while the second column allows for a quadratic form of change.

**Exercise** Show that the above polynomial contrasts are equivalent to the fitting of a quadratic polynomial, i.e., they give the same set of fitted values.

## 5.2 Regression splines

See [Maindonald & Braun \(2007, Section 7.5, pp. 234–238\)](#). See [Ruppert et al \(2003\)](#) for a more complete account of regression splines.

We will use cubic splines, i.e., cubic polynomial curves are joined at internal *knots*, with the requirement that the slopes and the second derivative must be continuous over the whole ranges of values. Specifically, the slope and the second derivative must agree at the knots. Cubic splines are the most commonly used form of spline, and have “nice” theoretical properties. Natural splines have the (often sensible) further constraint that the second derivative is zero at the boundary points, or *boundary knots*. Boundary knots are commonly taken to coincide with the extremes of the data, but this is not necessary. It turns out that such curves can be expressed as a linear combination of basis functions.

For our purposes, spline basis terms are a kitset for constructing a flexible variety of curves. As the number of degrees of freedom increases, it becomes possible to accommodate an ever increasing variety of curves. We will use natural splines, generated by the function `ns()` in R’s `spline()` package.

A good way to get a feel for spline functions is to plot the basis functions. These depend on the  $x$ -values that are chosen. We will examine (Figure 3) the spline basis functions when  $x$ -values range from 1 to 50 in increments of one.

```
## Generate the matrix of values of the basis functions
x <- 1:50
Xns <- ns(x, df=5)
```

Observe the attributes `knots` and `Boundary.knots`

```
> attributes(Xns)$knots
 20% 40% 60% 80%
10.8 20.6 30.4 40.2
```

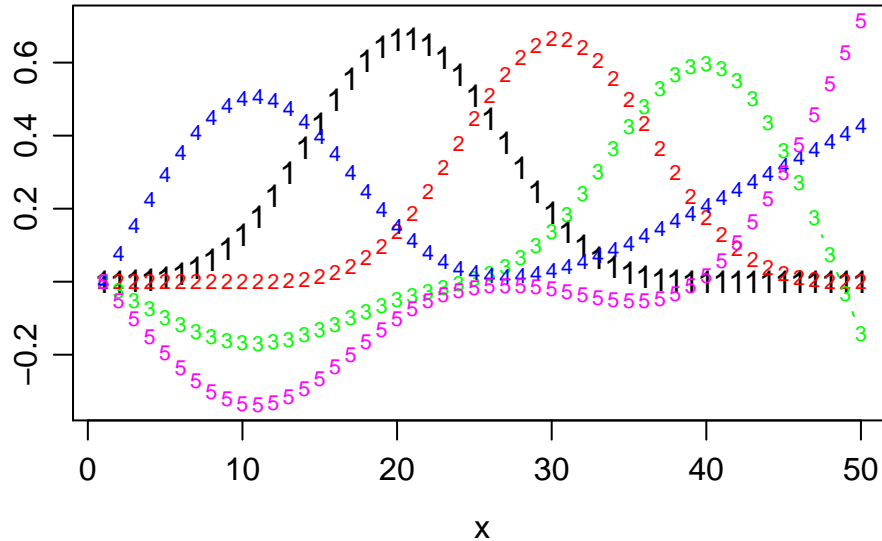


Figure 3: Spline basis functions, with  $x = 1:50$ . Basis function 1 is plotted with 1's, and so on.

```
> attributes(Xns)$Boundary.knots
[1] 1 50
```

Code that will create the plots is

```
## Create the plots
plot(1:50, Xns[,1], ylim=range(Xns), pch="1")
points(1:50, Xns[,2], pch="2", col="red")
points(1:50, Xns[,3], pch="3", col="green")
points(1:50, Xns[,4], pch="4", col="blue")
points(1:50, Xns[,5], pch="5", col="magenta")
```

The figure shows the first five curves in the kitset (all that are available with  $df=5$ ), when  $x$  ranges from 1 to 50 in increments of 1.

### 5.2.1 How many degrees of freedom?

A quadratic (hill or valley-shaped curve) has one degree of freedom for the intercept, plus one degree of freedom for the slope, plus one degree of freedom for the curvature. Every additional point where the second derivative is zero (the slope instantaneously ceases changing) accounts for one additional degree of freedom.

Consider the following sine curve

```
x <- seq(from=0, to = 6*pi, length=50)
y <- sin(x)
plot(x,y)
```

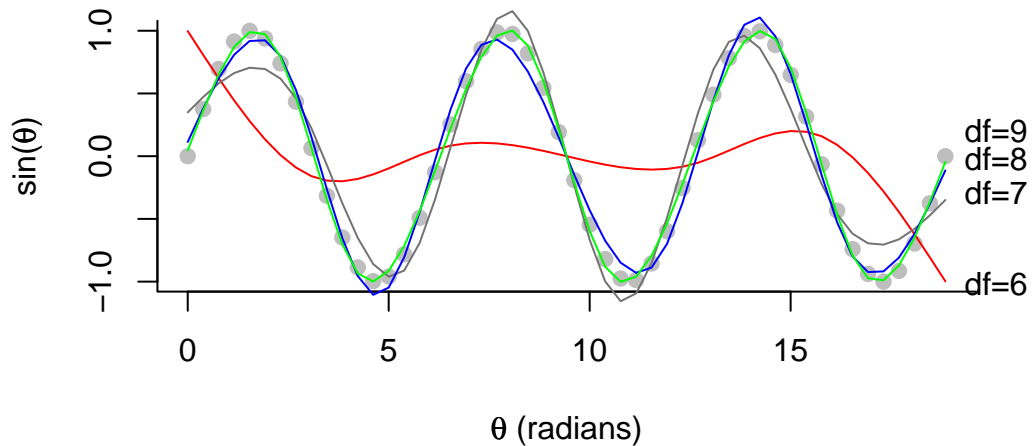


Figure 4: Regression spline fit to points that are determined by a sine curve, over the range 0 to  $6\pi$ . A 6 d.f. natural spline curve is clearly grossly inadequate. A 7 d.f. curve gets the broad shape more or less correct. An 8 d.f. curve is clearly preferable. The improvement with a 9 d.f. curve is more marginal.

The first "hill" accounts for two degrees of freedom. There are five additional valleys/hills, accounting in total for seven degrees of freedom, additional to the intercept. Hence we try

```
lines(x, fitted(lm(y~ns(x,7))), col="red")
```

Any fewer degrees of freedom is grossly inadequate. The choice `df=6` shows how bad the spline fit can be when there are not enough degrees of freedom to capture the major features of the data. The choice `df=8` is an obviously worthwhile improvement on `df=7`. Figure 4 provides a comparison.

**Note** If B-splines are used (function `bs()`), two extra degrees of freedom might seem required to capture the broad shape of the curve. There is however a choice in whether the available degrees of freedom are used to give freedom to change shape near the boundary knots, or whether to fit the necessary number of hills and valleys. The least squares algorithm uses that discretion to good effect.

### Regression splines – an example

The `covsample` dataset that is in the *DAAGxtras* package gives forest cover type (one of eight types) as a function of various environmental attributes. There is no information on geographical coordinates. However, it might be expected that there would be an ordering in the data, reflecting some kind of geographical ordering. If so, this is likely to be reflected in a systematic variation in

the environmental attributes. The first 10 of these are continuous variables. These are: Elevation, Aspect, Slope, Horizontal\_Distance\_To\_Hydrology, Vertical\_Distance\_To\_Hydrology, Horizontal\_Distance\_To\_Roadways, Hillshade\_9am, Hillshade\_Noon, Hillshade\_3pm, Horizontal\_Distance\_To\_Fire\_Points.

We will investigate whether the first of these (Elevation) seems to vary systematically through the data. We create a variable `dist` that measures distance through the data.

```
covdf <- cbind(covsample[, 1:10], dist=(1:11318-0.5)/11318)
## First, investigate use of lowess()
with(covdf, plot(lowess(dist, V1))) ## too smooth
with(covdf, plot(lowess(dist, V1, f=0.1))) ## less smooth
with(covdf, plot(lowess(dist, V1, f=0.01))) ## reasonable?
```

Now, use regression splines. The final plot from the `lowess` has 11 or 12 features that might be characterised as hills or valleys or plateaus. We will try 15 d.f.

```
hat <- fitted(lm(V1 ~ ns(dist,15), data=covdf))
## Overplot on the earlier lowess smooth
lines(hat ~ dist, data=covdf, col="red")
```

This seems not quite adequate. It does rather poorly at approximating one of the major features, as identified by the `lowess` curve. Use of a much higher number of degrees of freedom makes little difference, however. The difference is probably that `lowess()` gives a robust smooth. This can be checked by using `rlm()` (a robust version of `lm()`) for the fit. We will be generous in the number of degrees of freedom that are allowed.

```
with(covdf, plot(lowess(dist, V1, f=0.01)))
hat <- fitted(rlm(V1 ~ bs(dist,40), data=covdf))
lines(hat ~ dist, data=covdf, col="red")
## Compare with use of lm() with 40 d.f.
hat <- fitted(lm(V1 ~ bs(dist,40), data=covdf))
lines(hat ~ dist, data=covdf, col="blue")
```

## 6 Generalized Linear Models (GLMs)

The main case of interest will be logistic regression models, where the outcome is binary, i.e. 0 or 1. There will be some discussion of the general theory, but mainly for its relevance to this special case.

For examples of logistic regression models, see [Maindonald & Braun \(2007, Sections 8.1 & 8.2\)](#). Suitable texts for further reading and reference, in increasing order of technical demands, are [Faraway \(2006\)](#); [Wood \(2006\)](#); [McCullagh and Nelder \(1989\)](#).

In principle, models with a 0/1 outcome can be fitted using least squares. If unweighted least squares is used, and the variance really is that for a Binomial( $n$ ,  $p$ ) distribution with  $n = 1$ , then estimates will be inefficient, though the effect will be of little consequence if fitted proportions lie between perhaps 0.25 and 0.75. Some fitted values may be less than 0 or greater than 1. Where a continuous explanatory variable has a non-zero coefficient, extrapolation to suitably

small or suitably large values will always yield predicted proportions that are outside the range (0,1).

The efficiency of least squares estimates can be improved by taking the fitted proportions from an initial least squares fit, and using these to determine weights for a second fit. This is a step on the way to using a maximum likelihood fit for a generalized linear model. The remedy for preventing silly predicted values is to fit a linear model on the scale of a suitably transformed predicted value, which is exactly what GLMs are designed to do.

It will be interesting to compare logistic regression fits with least squares fits, for roughly equivalent models, to see what difference it makes.

## 6.1 Brief summary of theory

- As before, we have  $\boldsymbol{\mu}$  ( $n$  by 1),  $\mathbf{X}$  ( $n$  by  $p$ ),  $\boldsymbol{\beta}$  ( $p$  by 1), &  $\boldsymbol{\epsilon}$  ( $n$  by 1), with  $\boldsymbol{\mu} = E[\mathbf{y}]$
- The model is now

$$f(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{where } E(\mathbf{y}) = \boldsymbol{\mu}$$

Here,  $f()$  has the name *link* function. For example,

$$f(\mu_i) = \log\left(\frac{\mu_i}{N_i - \mu_i}\right)$$

- The theory requires independently distributed  $\mathbf{y}_i$ , from the exponential family of distributions. Common distributions that are allowed are the normal, binomial and Poisson.
- Output is in much the same form as for the `lm` models. There are additional subtleties of interpretation – a `z` value is not a `t`-statistic, though for some GLMs that yield `z` values there are specific circumstances where it is reasonable to treat them `z` values as `t`-statistics. [More technically, they are Wald statistics.]
- GLMs with binomial errors are formally equivalent to discriminant models where there are two categories. The GLM framework has advantages for some problems.

## 6.2 GLMs – commentary on the theory

### Solution by maximum likelihood

- Recall that the equation is

$$f(\boldsymbol{\mu}) = E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

where  $\boldsymbol{\mu} = E[\mathbf{y}]$

- Assuming a distribution from the exponential family, the maximum likelihood estimates of the parameters are given by

$$\mathbf{X}'\mathbf{W}\boldsymbol{\mu} = \mathbf{X}'\mathbf{W}\mathbf{y}$$

where  $f(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$ , and  $\mathbf{W}$  is a diagonal matrix.

- Note that the (diagonal) elements of  $\mathbf{W}$  are functions both of  $\text{var}[\mathbf{y}_i]$  and of  $f(\boldsymbol{\mu}_i)$
- The ML equations must in general be solved by iteration ( $\boldsymbol{\beta}$  appears on both sides of the equation.) Iteratively reweighted least squares is used, i.e. Newton-Raphson.

### Asymptotic theory

- Except in special cases, the statistical properties of parameters rely on asymptotic results.
- For binary logistic regression, the theory should be used with extreme care. Assessments of predictive accuracy can be derived using cross-validation.

## 7 Discriminant Methods

This will follow [Maindonald & Braun \(2007, Sections 12.2 and 11.7, , with a brief overview of Sections 11.1–11.6\)](#).

In this course, the main use of discriminant methods will be for the derivation of propensity scores that may serve to characterize the differences between two or more sets of observational data. If one or more propensity scores can replace many covariates that together account for much of the difference between different groups of data, this can greatly simplify the analysis, allow more effective use of standard diagnostic tools, and give results that are more readily interpretable.

In linear discriminant analysis, discriminant scores in as many dimensions as seem necessary are used to classify the points, and thus emerge directly from the analysis. However the linearity assumptions are restrictive, even allowing for the use of regression spline terms to model non-linear effects. It is not obvious how to choose the appropriate degree for each of a number of terms. The attempt to investigate and allow for interaction effects adds further complications. In order to make progress with the analysis, it may be expedient to rule out any but the most obvious interaction effects. The problem affects regression methods (including GLMs) as well as discriminant methods.

Where there are two groups, logistic or other binary regression models can be used to determine a discriminant. From that perspective, linear discriminant analysis, as implemented e.g. by R's function `lda()`, can be viewed as a variation on and extension of the use of a GLM for binary data.

Random forests offer a nonparametric approach. The proportion of trees in which any pair of points appear together at the same node may be used as a measure of the “proximity” between that pair of points. Then, using 1-proximity as a measure of distance, an ordination method can be used to find a representation of those points in a low-dimensional space.

## 8 Validity Issues

Here will be discussed:

1. Errors in  $x$ .

2. Effects that arise from model and variable selection. The chief interest will be in effects that are, or can be, large.
3. Missing variables and/or mis-specification of the model.
4. Failure of the independence assumption.

## 8.1 Errors in $x$

Here will be discussed just one of a variety of possible “errors in  $x$ ” models, described in [Carroll \(2006\)](#) as the “classical” model. See [Carroll \(2006, pp. 49-52\)](#) for a summary of different types of models that have been proposed. In the following, we discuss implications for the interpretation of the regression coefficients.

### 8.1.1 Regression with a single covariate

Consider first regression with a single covariate. Under the classical model errors in explanatory variables, if they are sufficiently extreme, have two effects:

1. Estimates of the coefficient will be reduced, relative to the coefficient for the variable that is measured without error.
2. Very large samples may be required to show a statistically detectable coefficient.

### 8.1.2 One covariate measured with error; others without error

The coefficient of the variable that is measured with error is attenuated, as in the single variable case. The coefficients of other variables may be reversed in sign, or show an effect when there is none. See [Carroll \(2006, pp. 52-55\)](#) for summary comment.

Suppose that

$$y = \beta_x \mathbf{x} + \beta_z z + \epsilon$$

If  $w$  is unbiased for  $x$  and the measurement error  $u$  is independent of  $x$  and  $z$ , then least squares regression yields a consistent estimate of  $\lambda\beta_x$

$$\lambda = \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}$$

The  $\sigma_x^2$  that appears in the single covariate case is replaced by  $\sigma_{x|z}^2$ .

A new feature is the bias in the least squares estimate of  $\beta_z$ . The naive least squares estimator estimates

$$\beta_z + \beta_x(1 - \lambda)\gamma_{x|z} \tag{2}$$

where  $\gamma_{x|z}$  is the coefficient of  $z$  in the least squares regression of  $x$  on  $z$ . The least squares estimate may be non-zero value even though  $\beta_z = 0$ . Where  $\beta_z \neq 0$ , the least squares estimate may, depending on the relative values of  $\beta_z$ ,  $\beta_x$  and  $\lambda$  be reversed in sign from  $\beta_z$ .

Where there are multiple explanatory variables that are measured without error, equation 2 can be applied to each of them in turn.

### 8.1.3 Implications for variable selection

The implications are clearly damaging. If one covariate only is measured with substantial error has a non-zero effect, then any variable that

- has a non-zero correlation with that covariate, and
- has no effect on the response

will have a non-zero expected least squares coefficient.

Where two or more variables are measured with substantial error, effects on other least squares coefficients may in fortuitous circumstances cancel. More important, in most practical circumstances, is a widening of the range of possibilities for obtaining least squares coefficients that are spuriously non-zero.

## 8.2 Missing variables and/or mis-specification of the model

The issue will be illustrated with examples. Some striking examples come from the analysis of multi-way tables. For a formal analysis, logistic regression can be used. However, the effects that are of interest are perhaps better demonstrated from examination of relevant tables.

An important reference is [Rosenbaum \(2002\)](#).

### 8.2.1 Do airbags reduce risk of death in an accident

Each year the National Highway Traffic Safety Administration in the USA collects, using a random sampling method, data from all police-reported crashes in which there is a harmful event (people or property), and from which at least one vehicle is towed. The data in [Table 2](#) summarize data in the data frame `nassCDS` (*DAAGxtras*).<sup>1</sup>

The data are a sample. The use of a complex sampling scheme has the consequence that the sampling fraction differs between observations. Each point has to be multiplied by the relevant sampling fraction, in order to get a proper estimate of its contribution to the total number of accidents. The column `national` (*national inflation factor*) gives the relevant multiplier.

Other variables than those included in `nassCDS` might be investigated – those extracted into `nassCDS` are enough for present purposes.

seatbelt	airbag	dead	total	Prop_dead
none	none	24067	1366089	0.01762
belted	none	15609	4118833	0.00379
none	airbag	13760	885635	0.01554
belted	airbag	12159	5762975	0.00211

Table 2: Number of fatalities, by use of seatbelt and presence of airbag. Data are for front-seat occupants.

<sup>1</sup>They hold a subset of the columns from a corrected version of the data analyzed in [Meyer and Finney \(2005\)](#). More complete data are available from one of the web pages <http://www.stat.uga.edu/~mmeyer/airbags.htm> (SAS transport file) or <http://www.maths.anu.edu.au/~johnm/datasets/airbags/> (R image file).

Meyer and Finney (2005) and Meyer (2006) conclude that on balance (over the period when their data were collected) airbags gave no statistically detectable benefit. There is a suggestion that airbags may have cost lives. Their study seems a large improvement over an official National Highway Traffic Safety Administration assessment of the evidence that was based on accidents where there was at least one death. Farmer (2006) offers an alternative form of analysis that does suggest a benefit. A definitive conclusion is impossible; see the further discussion below.

In order to obtain a fair comparison, it is necessary to adjust, not only for the effects of seatbelt use, but also for speed of impact. When this is done, airbags appear on balance to be dangerous, with the most serious effects in high impact accidents. Strictly, the conclusion of the two Meyer papers is that, conditional on involvement in an accident that was sufficiently serious to be included in the database (at least one vehicle towed away from the scene), airbags are harmful.

Both sets of data are from accidents, and there is no way to know how many cases there were with airbags where accidents (serious enough to find their way into the database) were avoided, as opposed to the cases without airbags where accidents were avoided. Tests with dummies do not clinch the issue; they cannot indicate how often it will happen that an airbag disables a driver to an extent that they are unable to recover from an accident situation enough to avoid death or serious injury.

Before installation of airbags was made mandatory, should there have been a large controlled trial in which one out of every two cars off the production line was fitted with an airbag? Would it have worked? Or would there be too much potential for driver behaviour to be influenced by whether or not there was an airbag in the car? Would it have been possible to sell the idea of such a trial to the public?

### 8.3 Model and variable selection

Approaches that may be used include:

- Stepwise regression; either forward (starting with a simple model) or backward (starting with a maximal model). All such methods suffer from the difficulties that:
  - Optimal decisions at each local step do not ensure a globally optimal model;
  - Decisions on whether to include or drop variables at each local step have a large element of arbitrariness. Should the  $F$ -statistic or  $p$ -value be used, or should an information-based measure (AIC, BIC, ...) be used? In either case, what are the appropriate thresholds for adding or dropping variables? Should the threshold be held constant (on one or other scale) throughout the process?
  - Separate data must in general be set aside for use in deriving the statistical properties of estimates. Other selection effects will bias statistics that for the model that is, finally, selected.
- Best subsets regression. This is, except in relatively simple situations, computationally expensive. Cross-validation or bootstrap approaches to

deciding model size make it even more expensive. The training/test set approach may offer a way out, at the cost of making poor use of what may be limited data.

Among recent papers on variable selection, note Luo et al (2006) and Zhu and Chipman (2006). The exposition in Luo et al (2006) is less than satisfactory, and the examples that they give are unconvincing. Zhu and Chipman (2006) is interesting. The main usefulness of the genetic algorithm may be in the insertion of randomness into the selection process. This could be achieved in other ways, e.g., by taking bootstrap samples. Model selection remains, except in the simplest cases, a difficult and challenging problem.

## 9 Resampling Methods

Such methods can be a useful recourse when theoretical results are unavailable, or when asymptotic results seem too inaccurate. They allow a removal or weakening of normality assumptions. While a recourse that is often useful, they should not be regarded as the answer to every problem of failure of assumptions.

Three methods will be noted:

1. Permutation methods locate a fitted model statistic (e.g., the residual mean square) within the distribution that is obtained when observed values are randomly permuted. If the residual mean square for the fitted model is in the upper tail of this distribution, this is taken as evidence that the model has some predictive power. Another possibility is to work with permuted residuals, thus estimating a permutation distribution for model coefficients.
2. In cross-validation, observations are split into  $k$  parts. For each of  $i = 1, 2, \dots, k$ , the  $i$ th part is left aside for use in testing, the model is fitted to the remaining  $k - 1$  parts, and predictions made for the  $i$ th part. This yields predictions for all observations that have been derived independently of those observations. An accuracy measure can then be computed.
3. The idea of the bootstrap is to regard the sample as a microcosm of the population. Repeated with replacement resamples are taken, and the model fitted to each resample, yielding a sampling distribution of parameter estimates. There are many variations on this simple scheme.

For examples of the use of resampling methods, see Maindonald & Braun (2007) thus:

**p.90:** Brief general comments;

**pp.129–134 (Section 4.7):** Permutation methods & the bootstrap;

**pp.159–164 (Section 5.5):** Cross-validation & the bootstrap, in straight line regression;

**pp.257–258 (Subsection 8.2.3):** Cross-validation, in logistic regression;

**pp.381–383 (Subsections 12.1.2):** Use of the bootstrap to check the stability of a principal components plot;

**pp.386–388 (Subsections 12.2.2 & 12.2.3):** Cross-validation, in discriminant analysis.

**pp.400–403 (Subsection 12.3.3):** Use of cross-validation in variable selection for discriminant analysis.

## 10 Examples and Issues

### 10.1 A severely constrained sample of books

Data, on book dimensions and book weight, are from the `oddbooks` dataset in R. The discussion will follow (Maindonald & Braun, 2007, Section 6.5, pp. 196–199).

### 10.2 Hill race data

The data are from the `hills2000` data set from the DAAG package. To make the data available, do the following:

```
> library(DAAG)
> names(hills2000)
[1] "h"      "m"      "s"      "h0"     "m0"     "s0"     "dist"   "climb"  "time"
[10] "timef"
```

The row names store the names of the hillraces. I have recently discovered that for the `Caerketton` race, where the time seems anomalously small, the value of `dist` seems in doubt. Possibly it should be 1.5mi not 2.5mi. The safest option may be to omit this point. For later reference, note the row number:

```
> match("Caerketton", rownames(hill2k))
[1] 42
> hill2k[42, "dist"]
[1] 2.5
```

The interest is in prediction of `time` as a function of `dist` and `climb`. First examine the scatterplot matrices, for the untransformed variables, and for the log transformed variables. The pattern of relationship between the two explanatory variables – `dist` and `climb` – is much closer to linear for the log transformed data, i.e., the log transformed data are consistent with a form of parsimony that is advantageous if we hope to find a relatively simple form of model. Note also that the graphs of `log(dist)` against `log(time)` and of `log(climb)` against `log(time)` are consistent with approximately linear relationships. Thus, we will work with the logged data:

```
loghill2k <- log(hill2k[-42, ])
names(loghill2k) <- c("ldist", "lclimb", "ltime", "ltimef")
loghill2k.lm <- lm(ltime ~ ldist + lclimb, data = loghill2k)
par(mfrow = c(2, 2))
plot(loghill2k.lm)
par(mfrow = c(1, 1))
```

We pause at this point and look more closely at the model that has been fitted. Does `log(time)` really depend linearly on the terms `ldist` and `log(lclimb)`?

The function `termplot()` gives a graphical summary that can be highly useful. The graph is called a termplot because it shows the contributions of the different terms in the model. We use the function `mfrow()` to place the graphs side by side in a panel of one row by two columns:

```
par(mfrow = c(1, 2))
termplot(loghill2k.lm, col.term = "gray", partial = TRUE,
         col.res = "black", smooth = panel.smooth)
par(mfrow = c(1, 1))
```

The plot shows the “partial residuals” for `log(time)` against `log(dist)` (left panel), and for `log(time)` against `log(climb)` (right panel). They are partial residuals because, for each point, the means of contributions of other terms in the model are subtracted off. The vertical scales show changes in `ltime`, about the mean of `ltime`.

The lines, which are the contributions of the individual linear terms (“effects”) in this model, are shown in gray so that they do not obtrude unduly. For the lines as well as the points, the contributions of each term are shown after averaging over the contributions of all other terms. The dashed curves, which are smooth curves that are passed through the partial residuals, are the primary feature of interest in these plots. In both panels, they show clear indications of curvature.

### 10.2.1 Spline Terms

```
loghill2k.lm <- lm(ltime ~ ldist + lclimb, data = loghill2k)
par(mfrow = c(1, 2))
termplot(loghill2k.lm, col.term = "gray", partial = TRUE,
         col.res = "black", smooth = panel.smooth)
par(mfrow = c(1, 1))
```

A spline of degree 3 (by default a cubic polynomial) seemed adequate for capturing the curvature in the partial residuals for `ldist`, while a spline of degree 4 seemed adequate for capturing the slightly more complicated pattern of curvature in the partial residuals for `lclimb`:

```
library(splines)
loghill2ks.lm <- lm(ltime ~ ns(ldist, 3) + ns(lclimb, 4), data = loghill2k)
```

Notice that the first plot brings together the information associated with the basis functions that are generated by `bs(ldist,3)`, while the second plot brings together the information associated with the basis functions that are generated by `bs(lclimb,4)`

**Diagnostic plots:** The following is a series of diagnostic plots, designed to highlight issues that it may be important to consider:

```
if (dev.cur() == 2) invisible(dev.set(3))
par(mfrow = c(2, 2))
plot(loghill2ks.lm)
par(mfrow = c(1, 1))
```

The diagnostic plots cannot possibly identify all possible problems with the fit of the models to the data. It is possible to have models where the diagnostic plots look fine, but the model is lousy. They can however be very useful in picking up some issues that commonly merit attention – outliers, non-normality in the residuals, heterogeneity of variance, and points that individually have a large effect on the fitted model.

Notice that, in the diagnostic plot, one point (row 19: 12 Trig Trog) has a huge Cook's distance. With a time of 8.3h, it is the longest of any of the races.

The following plots the contributions of the individual spline curves (“the effects”), shows the partial residuals, and passes a smooth curve (red dashes) through the partial residuals:

```
if (dev.cur() == 3) invisible(dev.set(2))
par(mfrow = c(1, 2))
termplot(loghill2ks.lm, col.term = "gray", partial = TRUE,
         col.res = "black", smooth = panel.smooth)
par(mfrow = c(1, 1))
```

Also the fitted curve for `lclimb` is not monotonic for small values of `lclimb`. It would be desirable to constrain it to be monotonic.

### \*The basis functions

Use the following to inspect and plot the basis functions:

```
bases <- model.matrix(loghill2ks.lm)
colnames(bases)
options(digits = 3)
bases[1:5, ]
par(mfrow = c(2, 2))
for (i in 0:3) plot(loghill2k$lclimb, bases[, 5 + i])
par(mfrow = c(1, 1))
```

## 10.3 Diet-disease studies

The attempt to use food frequency questionnaires (FFQs) or food diaries, in studies that are designed to detect diet-disease associations, provides a telling and interesting case study. A recent major study with biomarkers has demonstrated large person-specific biases in standard dietary intake measurement “instruments” (diaries or questionnaires). These biases severely complicate the finding of a relationship between such measures and health outcomes. Not only is there an error that varies from recording time to recording time, for an individual. There is also a person-specific bias that can be substantially larger than the random occasion to occasion error. See [Schatzkin et al \(2003\)](#) and the power point presentation [Carroll \(2006\)](#).

This is a multi-million dollar issue. The following prospective studies that use such instruments are complete or nearly complete:

NHANES:	n = 3,145 women aged 25-50 (National Health and Nutrition Examination Survey)
Nurses Health Study:	n = 60,000+
Pooled Project:	n = 300,000+
Norfolk (UK) study:	n = 15,000+
AARP:	n = 250,000+

Only 1 prospective study has found firm evidence suggesting a fat and breast cancer link, and 1 has found a negative link. The lack of consistent (even positive) findings led to the Women’s Health Initiative Dietary Modification Study in which 60,000 women have been randomized to two groups: healthy eating and typical eating. Objections to this study are:

- Cost (\$100,000,000+)
- Can Americans can really lower % fat calories from to 20%, from the current 35%
- Even if the study is successful, difficulties in measuring diet mean that we will not know what components led to the decrease in risk.

#### 10.4 Lalonde’s data – effectiveness of a labour training program

This will review the discussion in [Maindonald & Braun \(2007, Section 13.2\)](#), though working with the `nswdemo` dataset in the `DAAGxtras` package rather than with the `nsw74psid1` dataset from `DAAG`. Proximities from software that uses bootstrap aggregation offer an alternative and it will be argued, preferable approach to the determination of distances and hence ordination scores that can be used in a regression. I will explore this approach as a preferred alternative.

There has been a long-standing debate in the econometric literature over whether social programs can be reliably evaluated without a randomized experiment. The Lalonde data have received wide attention, from several different authors, in the course of this debate. A recent contribution to the debate, which gives a good summary of the controversy, is [Smith & Todd \(2005\)](#).

## 11 References and Further Reading

### References

- Bates, D, 2007. Comparing least squares calculations. *Vignette “Comparisons” accompanying the package “Matrix” for R.*
- Bishop, C. M, 2006. *Pattern Recognition and Machine Learning*. Springer.  
[Chapters 3 and 4 offer an interesting and somewhat novel perspective on regression and discriminant methods. The theoretical framework is that of Bayesian Decision theory. This is a demanding text. There is helpful comparative commentary on the methods that are described.]
- Carroll, R., Ruppert, D. and Stefanski, L. A. 2006. *Measurement Error in Nonlinear Models*. 2<sup>nd</sup> edition, Chapman and Hall.  
[This is a definitive text on errors in linear and nonlinear models.]

- Cochran, W. G. 1965. The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*, **128**:134-155
- Cook, R D and Weisberg, S., 1999. *Applied Regression Including Computing and Graphics*. Wiley.  
 [This emphasizes geometric insights, linear predictors (transformation of predictors, if possible, so that pairwise regression relationships are linear), and dimension reduction.]
- Faraway, J. J. 2006. *Extending the Linear Model with R*. Chapman & Hall/CRC.
- Faraway, J. J. 2005. *Linear Models with R*. Chapman & Hall/CRC.
- Farmer, C.H. 2006. Another look at Meyer and Finney's 'Who wants airbags?' *Chance* 19:15-22.
- Koenker, R and Ng, P, 2003. SparseM: A sparse matrix package for R. *Journal of Statistical Software* 8(6).
- Luo, X. H., Stefanski, L. A., and Boos, D. D. (2006). Tuning variable selection procedures by adding noise. *Technometrics* 48, 165-175.
- Maindonald, J. H. and Braun, W.J. 2007. *Data Analysis and Graphics Using R – An Example-Based Approach*. 2<sup>nd</sup> edition, Cambridge University Press.  
 URL:<http://www.maths.anu.edu.au/~johnm/r-book.html>  
 [As the title says, this is example-based, drawing attention to theoretical issues as they arise in the context of specific examples. There is extensive use of graphs that may provide insight on data and on fitted models. It has extended critiques of alternative approaches, and gives detailed advice on a wide range of practical data analysis issues.]
- McCullagh, P. and Nelder, J. A., 1989. *Generalized Linear Models*. Chapman and Hall, 2<sup>nd</sup> edition.
- Meyer, M C and Finney, T. 2005. Who wants airbags?. *Chance* 18:3-16.
- Meyer, M C, 2006. Commentary on Another look at Meyer and Finney's 'Who wants airbags?'. *Chance* 19:23-24.
- Rosenbaum, P. R., 2002. *Observational Studies*. 2<sup>nd</sup> edition, Springer-Verlag.  
 [This is required reading for anyone who works with observational data.]
- Ruppert, D., Wand, M., and Carroll, R. 2003. *Semiparametric Regression*. Cambridge University Press.
- SCHATZKIN, A., KIPNIS, V., CARROLL R.J., MIDTHUNE, D., SUBAR, A.F., BINGHAM, S., SCHOELLER D.A., TROIANO, R.P. AND FREEDMAN, L.S. 2003. A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. *International Journal of Epidemiology* 32: 1054-1062.
- Smith, J. A. and Todd, P.E. 2005. Does Matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125: 305-353.

Wood, S. N., 2006. *Generalized Additive Models*. An Introduction with R. Chapman & Hall/CRC.

[This has an elegant treatment of linear models and generalized linear models, as a lead-in to generalized additive models.]

Zhu, M. and Chipman, H.A. 2006 *Darwinian Evolution in Parallel Universes: A Parallel Genetic Algorithm for Variable Selection*. *Technometrics* 48, 491–502.