# Practical Aspects of the Use of Linear & Generalized Linear Models

J.H. Maindonald*

August 7, 2007

**Abstract**

Primarily, the purpose of this document is to note issues for the use of linear and generalized linear models, and of other regression models. Worked examples, as far as possible using data that have been a basis for published research, are used as a basis for discussion of the following issues:

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)]

Observational versus experimental data – implications for interpretation and inference;
Maindonald & Braun (2007, Chapter 6); Rosenbaum (2002)]

Variable selection, noting the use of resampling methods to obtain realistic "error" estimates;
[Maindonald & Braun (2007, Chapter 6)]

Errors in explanatory variables; implications of classical measurement error for inference;
[Maindonald & Braun (2007, Chapter 6); Carroll (2006, Chapter 1)]

Regression on constructed variables – propensity scores, with brief mention of principal components and partial least squares;
[Maindonald & Braun (2007, Chapter 13)]

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

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

# Contents

# 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.

Models in which E[$y$] (the "fixed part" of the model) 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 for linear models requires only that models are linear in the parameters.

## 1.1 Examples

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

### 1.1.1 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).

### 1.1.2 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)
```

### 1.1.3 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.

### 1.1.4 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.

### 1.1.5 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

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

$\alpha$ 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 asssumption.

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

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

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

where as a minimum it is required that $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$. This is commonly strengthened to $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, 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 paremeter 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 deerived 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.

### 1.6 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 reccourse 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;

### 1.7 Model diagnostics

Most model diagnostics rely on various forms of scrutiny of the model residuals. R's plot method for `lm` objects gives, by default, four standard types of diagnostic plot.

Where there is more than one outlier, these can so distort the fitted model the plots of residuals from an `lm` analysis are misleading. It is better to work with the residuals from the robust linear model function `rlm()`.

If data are noisy, residuals from `lqs()`, which does resistant regression, should be examined. Currently, there is no usable plot method for `lqs` objects.

Output from `termplot()` is a useful supplement to output from the plot method for `lm` or `rlm` objects.

### 1.8 The interpretation of regression parameters

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

1. All relevant explantory 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.

## 2 Factor and Spline Terms

### 2.1 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)

### 2.1.1 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 columns 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 as quadratic polynomial, i.e., they give the same set of fitted values.

## 2.2 Balanced vs Unbalanced Data

This section illustrates, in a very simple case, confounding effects that may arise when data are unbalanced. It demonstrates, also, how the estimate for a term may change when a further term is included in the model.

### Post-operative pain profiles

The table shows, separately for males and females, the effect of pentazocine on post-operative pain profiles (average VAS scores), with (mbac and fbac) and without (mpl and fpl) preoperatively administered baclofen. Pain scores were recorded every 20 minutes, from 10 minutes to 170 minutes. Results are shown for 50 minutes only. The complete data may be found in the `gaba` dataset in the *DAAGxtras* package.

3 males were given baclofen, as against 15 females. 9 males received the placebo, as against 7 females.

| | male | female | Average over all subjects | Average of averages |
|---|---|---|---|---|
| placebo | 0.67 (9) | 3.66 (7) | (9*0.67+7*3.67)/16 = 1.98 | (0.67+3.67)/2 = 2.17 |
| baclofen | 0.05 (3) | 3.13 (15) | (3*.05+15*3.13)/18 = 2.62 | (0.05+3.13)/2 = 1.59 |

Notice that, both for males and for females, the scores are lower when baclofen is administered. Reliance on the overall average would suggest that the scores are higher when baclofen is administered.

Instead of taking the average over all subjects, we might take the overall average of the baclofen scores and the overall average of the placebo scores, ignoring the different numbers of subjects contributing to the separate male and female scores. This is shown in the "Average of averages" column, and gives a result that is defendable.

A competent analyst will, with such data, look for effects that may be due to factors other than the treatment. If information on a relevant factor is not included in the data, it is obviously not possible to allow for it at the time of analysis. Thus, if details of gender were not available for the subjects who contributed to the present data, the only averages that could be calculated would be the misleading overall averages given in the "Average over all subjects" column.

### 2.2.1 Analysis Using `lm()`

```
> pain <- data.frame(vasScore=c(0.67, 0.05, 3.67, 3.13),
+                    trt=factor(rep(c("placebo", "baclofen"), 2),
                               levels=c("placebo", "baclofen")),
+                    gender=factor(rep(c("male", "female"), c(2,2)),
                                  levels=c("male", "female")),
+                    number=c(3, 9, 15, 7))
> pain
   vasScore      trt gender number
1      0.67  placebo   male      9
2      0.05 baclofen   male      3
3      3.67  placebo female      7
4      3.13 baclofen female     15

> pain.lm <- lm(vasScore ~ gender + trt, data=pain)
> round(coef(pain.lm), 2)
 (Intercept) genderfemale  trtbaclofen
        0.65         3.04        -0.58
```

Now omit consideration of the Gender effect:

```
> pain1.lm <- lm(vasScore ~ trt, data=pain)
> round(coef(pain1.lm), 2)
(Intercept) trtbaclofen
       2.17       -0.58
```

Observe that the estimate of the treatment effect is unchanged.

#### Weighted analysis

```
> pain.wlm <- lm(vasScore ~ gender + trt, weight=number, data=pain)
> round(coef(pain.wlm), 2)
 (Intercept) genderfemale  trtbaclofen
        0.66         3.03        -0.57
```

Observe that the estimate of the treatment effect has hardly changed. In general, the change may not be so small, but will not reverse an effect that goes in the same direction for the genders separarately.

Now omit consideration of the gender effect:

```
> pain.wlm1 <- lm(vasScore ~ trt, weight=number, data=pain)
> round(coef(pain.wlm1), 2)
(Intercept) trtbaclofen
       1.98        0.63
```

Observe that the treatment effect now goes in the other direction. The combination of unequal weights and omission of a relevant factor generates this misleading result.

# 3 Applications of Logistic Regression Models

## 3.1 Logistic Regression vs Multi-way Tabulation

Models for multi-way tables that allow or all interactions, at all levels, are said to be "saturated". The probabilities may alternatively be derived from an equivalent multi-way table. The probabalities equal the fitted values from a logistic regression with a model that is thus "saturated" with respect to the multi-way table.

Code will be given that can be used to verify the equivalence of the multi-way table to fitted values from a logistic model. (Actually, because all interactions are included in the model, fitted values are the same irrespective of the link that is used with the binomial or quasibinomial model.)

### 3.1.1 US Accident Mortality Data

Data, in the data frame `nassCDS` in the `DAAGxtras` package, were collected according to a sampling design where different cells had different weights. Whether using the logistic regression or tabulating the proportions, it is necessary to take account of the weights.

The data have had a central role in a controversy, debated in three articles in the journal *Chance*, on the effectiveness of airbags. References, both to these articles and to relevant web pages, are given on the help page for `nassCDS`.

Various biases may affect the result. There are alternatives to the style of analysis discussed here. Farmer (2006) discusses an approach that is preferred by the National Highway and Traffic Safety Administraton (NHTSA), and which gives an answer that is favourable to the use of airbags.

### 3.1.2 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 1 summarize data in the data frame `nassCDS` (*DAAGxtras*).[1]

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 `weight` (or textttnif = textitnational 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.

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

[1] 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).

| 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 1: Number of fatalities, by use of seatbelt and presence of airbag. Data are for front-seat occupants.

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 aribags 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?

Notwithstanding care to consider all relevant effects, it remains possible that there will be relevant factors of interactions that have not been considered. The relevant information may not be included in the data. A useful strategy, with data such as these, may be:

- Account first for those effects that on prior grounds (relevant science, previous experience with related data), seem certain to have a role. Such arguments justify use, for the present data, of the factors `seatbelt`, `airbag` and `dvcat`.
- Investigate addition of other possible effects one at a time.

### 3.1.3 Factors that affect mortality

The analysis here will be limited to the factors `seatbelt` and `airbag`, leaving as an exercise extension to account for the force of impact measure (`dvcat`). Such a more extended analysis makes it clear that any defendable analysis in the style of the analysis discussed here must, as a minimum, include these three factors and their interactions.

Almost certainly, there are other factors, not considered in any of the analyses presented in the *Chance* articles, that have affected results. A more complete analysis will require consideration of further possible effects for which data are available. If more than one or two of those factors are included there is a risk, even with this relatively large data set, that it will become impossible to distinguish the likely effects of airbags from those of other factors.

Here is the code for the limited (and misleading!) tabulations presented here:

```
tot <- xtabs(weight ~ seatbelt+airbag, data=nassCDS)
dead <- xtabs(I(weight*(unclass(dead)-1)) ~ seatbelt+airbag, data=nassCDS)
```

Here is code for the `glm` analysis:

```
nass.glm <- glm(dead ~ seatbelt*airbag, weight=weight, family=binomial,
                data=nassCDS)
```

Note that `seatbelt*airbag` expands to `seatbelt + airbag + seatbelt:airbag`, i.e., all possible main effects and interactions.

Now reconcile this with the tabulated result.

```
## Reconcile with tabulated result
df <- with(nassCDS, expand.grid(seatbelt=factor(levels(seatbelt),
                                  levels=levels(seatbelt)),
                                airbag=factor(levels(airbag),
                                  levels=levels(airbag))))
df$tot <- as.vector(xtabs(weight ~ seatbelt+airbag, data=nassCDS))
nasshat <- predict(nass.glm, newdata=df, type="response", se=TRUE)
df$estdead <- nasshat$fit*df$tot
xtabs(tot ~ seatbelt+airbag, data=df)      # Table of totals
xtabs(estdead ~ seatbelt+airbag, data=df)  # Table of dead
```

Bootstrap esimates of the excess risk from airbags can be obtained thus:

```
xtra <- matrix(0, nrow=2, ncol=1000)
nass <- nassCDS[nassCDS$weight>0,]
prob=with(nass, weight/sum(weight))
for(i in 1:1000){
  nrows <- sample(1:dim(nass)[1], prob=prob, replace=TRUE)
  xtra[,i] <- excessRisk(form = weight ~ seatbelt + airbag,
                         data=nass[nrows, ])[, 8]
}
```

Calculations can take a long time. Run this to calculate 10 or 100 bootstrap samples before running it for the full 1000 samples. Percentile estimates of the confidence limits may in this instance be satisfactory.

# 4 Validity Issues – 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.

## 4.1 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.

The model is

$$y = \alpha + \beta x + \epsilon$$

We measure, not $x$, but $w = x + u$, where $u$ is "measurement error".

Now assume that $w$ is unbiased for $x$ and that $u$ is independent of $x$ and $\epsilon$.

Then, conditional on $x$, instead of

$$\hat{\beta} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

we have

$$\beta^* = \frac{\sum (w - \bar{w})(y - \bar{y})}{\sum (w - \bar{w})^2}$$

Then

$$\mathrm{E}[\sum (\mathrm{w} - \bar{\mathrm{w}})(\mathrm{y} - \bar{\mathrm{y}})] = \sum (\mathrm{x} - \bar{\mathrm{x}})(\mathrm{y} - \bar{\mathrm{y}})$$

This happens because $y$ is independent of $u$.

$$\begin{aligned} \mathrm{E}[\sum (\mathrm{w} - \bar{\mathrm{w}})^2] &= \mathrm{E}[\sum (\mathrm{x} - \bar{\mathrm{x}} + \mathrm{u} - \bar{\mathrm{u}})^2] \\ &= \sum (x - \bar{x})^2 + \sum (u - \bar{u})^2 \\ &= (n-1)\sigma_x^2 + (n-1)\sigma_u^2 \end{aligned}$$

Then $\beta^*$ is a consistent estimate of $\lambda\beta$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

### Simulation

Use the function `g6.17`, included in the image file **figs6.RData** that is available from:
http://www.maths.anu.edu.au/~johnm/r-book/2edn/figures/

## 4.2 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

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

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 1 can be applied to each of them in turn.

#### One covariate measured without error – a simulation

The function `errorsINxGP()`, available from http://www.maths.anu.edu.au/~johnm/r/functions/, simulates the effect when the variables that are measured without error code for a categorical effect.

### 4.3 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.

### 4.4 Example – diet-disease association studies

The attempt to use food frequency questionnnaires (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.

## 5 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).

## 5.1 Model and variable selection

### 5.1.1 Strategy issues

Stepwise and best subsets automatic variable selection procedures have a much more limited usefulness than older textbooks on regression may have suggested. Cross-validation or bootstrap approaches should be used to check out the stability of the selection with respect to statistical variation. Note that the use of automatic selection priocedures invalidates, in general, estimates that classical linear model theory gives for the standard errors of parameters. Bootstrap estimates of the standard errors, perhaps obtained along with the procedure used to check out the stability of the selection, may provide a workaround.

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 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.

Note further:

- Consider data where one variable, or a small number of variables jointly, have effects that, in the preferred model, are large relative to statistical error, while other variables have effects that, at best, are marginally detectable. Then classical selection techniques (stepwise regression, etc.) are likely to find those variables that have large effects, and their coefficients will be estimated without selection bias.
- In contexts where automatic selection techniques are tested more severely, they may not do much better than chance.
- To see the potential, with automatic selection algorithms, to get highly significant effects from random data, run the function `bestset.noise()`, from the *DAAG* package.
- There may be no unique "best" set of explanatory variables.
- The paper by Zhu and Chipman (2006) is interesting. The key here seems to be the incorporation of a random element. I suspect that a bootstrap approach, used in a similar way as in the random forests algorithm, would do as well or better.
- The selection problem is fraught with further hazards when one or more of the variables is measured with substantial error.

Attempts to interpret regression coefficients raise further hazards. Conditions that may make coefficients interpretable include: a) It is possible to identify a few variables that have large effects; b) the data allow their contribution to the regression to be estimated accurately; c) there is good reason to believe that no variables or interactions with substantial effects have been left out; d) there is a context of scientific understanding that supports proposed interpretations. Note further the Bradford Hill criteria, in Hill (1965).

# 6 Propensities

Propensities are one of a number of devices that may be used in the attempt to reduce the number of explanatory variables that need to be considered in a regression. The method is intended for a very specific, but important, context where there are two (or, potentially, more) levels of a treatment factor. The aim is to investigate treatment effects, after adjustment for covariate effects.

The attempt to adjust for multiple potential covariate effects has a variety of complications. The correct functional form must be used – it may not be adequate to assume additive linear effects, even after transformation of covariates in cases where this seems desirable. Diagnostic checking may be difficult; failure to account adequately for the effect of one or more variable may lead to misleading diagnostics for other variables.

The derivation and use of propensity scores can simplify the model fitting process. The complications that arise from the attempt to adjust for multiple covariates are limited to the modeling used to predict the propensity scores. Having derived a vector of propensity scores, the regression model that incorporates the treatment effect has two terms – a treatment effect, and a single covariate adjustment term. This can greatly simplify the analysis, allow more effective use of standard diagnostic tools, and give results that are more readily interpretable.

With more than two classes, a further set or sets of scores will in general be rewquired.

## 6.1 Discriminant Methods

There is a wide choice of classification methods that can be used to derive scores. Here, note two very different methods – linear discriminant analysis and random forests.

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. The linearity assumptions are of course restrictive, even allowing for the use of regression spline terms to model non-linear effects. With two classes, there is a linear relationship between the scores and the posterior log odds for the two classes.

Random forests are a highly nonparametric approach. The estimated log odds for the two classes can be used be used as propensity scores. A similar approach can be used with any other classifaction method that yields probabilities or posterior probabilities.

For random forests, note that the proportion of trees in which any pair of points appear together at the same terminal 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.

## 6.2 What is a propensity?

A propensity is the conditional probability $\lambda(\mathbf{x})$ of assignment to a particular treatment given a vector of observed covariates $\mathbf{x}$. The methodology requires that treatment assignment should be ignorable given the propensity, i.e., treatment assignment should be unrelated to potential outcomes within strata defined by $\lambda(\mathbf{x})$. Conditional on the propensity score, the distributions of the observed covariates are independent of the binary treatment assignment.[2] This allows use of the propensity score as a balancing score. See Rosenbaum (2002, pp.296-297), perhaps supplemented by Rosenbaum (1999) and Rosenbaum and Rubin (1983), for a discussion.

[2]The ignorability assumption seems to me implausible for the present data.

The propensity score, or a monotone function of the score, can be estimated using discriminant analysis methodology, independently of the outcome $y_i$. The regression equation becomes

$$y_i = t_i + \beta\phi(\lambda_i) + \epsilon_i$$

where the functional form of $\phi()$ has to be estimated or guessed. Scores from use of the logit transformation are often used as a starting point.

Compare this with the use of regression adjustments of the form

$$y_i = t_i + f(x_1, x_2, \ldots, x_k) \tag{2}$$

where in the simplest situation it might be hoped that

$$f(x_1, x_2, \ldots, x_k) = a_1x_1 + +a_2x_2 + \ldots + +a_kx_k$$

This requires the stronger condition that treatment assignment should be ignorable given the observed covariates **x**. i.e., treatment assignment should be unrelated to potential outcomes within strata defined by **x**.

The propensity score approach reduces the regression equation that is of primary interest to a simple form. Decisions on which variables and interactions to include, and on transformation and/or modeling using spline terms where this seems required, is relegated to the earlier discriminant function calculations. Diagnostics for the model for $y_i$ need be studied for one covariate only.

## 6.3 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).

## 6.4 Scores calculated using a linear discriminant

In Maindonald & Braun (2007, pp.412-419), we discuss the use, as propensity scores, of functions $f(x_1, x_2, \ldots, x_k)$ that are linear in covariates $x_1, x_2, \ldots, x_k$. This may be too restrictive. Even after use of spline terms (how many d.f.? what interactions, if any, should be included?) the model may be unable to capture well the nuances of the regression dependence in cases where there are more than one or two explanatory variables.

Here is code for the calculations:

```
common <- multilap(maxf=30)  # Mild preliminary filtering on
                             # variables educ, re74 and re75
## Calculate propensity scores: data frame nsw74psidA (DAAG)
disc.glm <- glm(formula = trt ~ age + educ + black + hisp + marr +
                re74 + re75, family = binomial,
                data = nsw74psid1, subset=common)
Pscores <- predict(disc.glm)
## Now filter further, based on values of Pscores
xchop <- with(subset(nsw74psid1, common),
              overlapDensity(Pscores[trt==0], Pscores[trt==1],
                             compare.numbers=FALSE,
```

```
                                ratio=c(1/30, 30)))
overlap <- common
overlap[common] <- Pscores > xchop[1] & Pscores < xchop[2]
nsw74psidC <- subset(nsw74psid1, overlap)
Pscores <- Pscores[Pscores > xchop[1] & Pscores < xchop[2]]
```

The hope is that, conditional on values of Pscores, controls and treated are now relatively similar with respect to the various covariates. This can be checked directly, by splitting the data set up in to, e.g., 5 parts, based on values of `Pscores`.

```
cut5 <- cut(Pscores, breaks=5)
for (cutlev in levels(cut5)){
print(cutlev)
nsw74 <- subset(nsw74psidC, cutlev==cut5)
print(
sapply(nsw74[, c("black","hisp","marr","nodeg")],
       function(x){tab <- table(nsw74[,"trt"], x)
                   tab[1,]/apply(tab,2,sum)})
)
}
```

The balancing is, in most cases, reasonable. There is however a big disparity in the numbers of hispanics in one of the categories, and none of the treated group in the final category were married. Similar comparisons can be done for the continuous variables.

Now try fitting the models:

```
nsw.lm <- lm(log(re78+25) ~ trt + propensity, data=nsw74psidC)
nsw.glm <- glm(I(re78>0) ~ trt + propensity,
               family=binomial, data=nsw74psidC)
```

## 6.5 Scores Calculated from `randomForest` Analysis

This is at the other extreme, relative to linear discriminant methods. Random forest models builds in as little structure as possible. There is no insistence on continuous forms of dependence on continuous variables. Remarkably as it seems to me, these models can, in some classification problems, do very well. It may be that the loss from ignoring of obvious structure is more than compensated by the ability to handle complex interactions and non-linear responses.

The following is exploratory. As a technique for revealing structure in data, it can work well. Here, we use the larger data set in which not all observations have information on `re75`. We therefore omit this variable from consideration.

```
nsw <- rbind(psid1, nswdemo[nswdemo$trt==1,])
nsw$trt <- factor(nsw$trt)
nswx <- nsw[, c(1:7,9)]
nsw.rf <- randomForest(trt ~ ., data=nswx, proximity=TRUE)
distmat <- 1-nsw.rf$proximity
distmat[distmat==0] <- 0.001  # Half minimum of non-zero distances
## Apply arcsine transformation to stretch the scale out at both ends
distmat <- asin(distmat)
## Start with classical multi-dimensional scaling (Euclidean distances)
nsw.cmd <- cmdscale(distmat)
plot(nsw.cmd, col=unclass(nsw$trt))
## Apply Sammon (semi-metric) scaling
nsw.sam <- sammon(distmat, nsw.cmd)
plot(nsw.sam$points, col=unclass(nsw$trt))
```

The plot suggests that the two groups are not well matched. There may however be subgroups that overlap substantially. One might try to separate out those regions of the plot where both

controls and (especially, as there are many fewer of these) treatment observations are reasonably well represented. A more direct approach is to use the predict method for `randomForest` objects to return a matrix of class probabilities, with one row for each data point.

```
## Take the second column; the first would do equally well.
prob <- predict(nsw.rf, type="prob")[,2]
prob[prob==0] <- 0.5*min(prob[prob>0])
prob[prob==1] <- 1-0.5*min(1-prob[prob<1])
scores <- log(prob/(1-prob))
```

Now find the range of values of scores where the ratios of the densities are in the range (0.025, 40), and try the regressions with attention limited to data where the scores are in that range:

```
z <- with(nsw, overlapDensity(ratio=c(.025,40),
          scores[trt==0], scores[trt==1]))
retain <- scores>z[1] & scores<z[2]
nsw.lm <- lm(log(re78+25) ~ trt + scores, data=nsw, subset=retain)
termplot(nsw.lm, par=T, smooth=panel.smooth)
nsw.glm <- glm(I(re78>0) ~ trt + scores,
                family=binomial, data=nsw, subset=retain)
```

The results depend on the chosen range of values of the scores. The estimates from `lm()` do not reproduce the results from the experimental comparison; in fact they suggest a negative effect from training. The estimates from `glm()` do favour the treatment group, providing the range of ratios is set small enough. The difference does not, however, reach the 5% level of statistical significance.

# 7 Further Example – the 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.

## 7.1 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))
```

# 8 References and Further Reading

## References

Carroll, R., Ruppert, D. and Stefanski, L. A. 2006. *Measurement Error in Nonlinear Models.* $2^{nd}$ 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

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.

Gordon, N. C. et al. 1995. Enhancement of Morphine Analgesia by the GABA_B against Baclofen'. *Neuroscience* 69:345-349.

Hill, A.B. 1965. The environment and disease. Association or causation? *Proceedings of the Royal Society of Medicine* 58: 295-300.

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.* $2^{nd}$ edition, Cambridge University Press.
URL:http://wwwmaths.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.]

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., 1999. Choice as an alternative to control in observational studies. *Statistical Science*, 14:259–278. With following discussion, pp.279–304.

Rosenbaum, P. R., 2002. *Observational Studies*. $2^{nd}$ edition, Springer-Verlag.
[This is required reading for anyone who works with observational data.]

Rosenbaum, P. and Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

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.