Chapter 6
Epilog

Now that you have nearly made it through the whole book, let me give you a little food for further thought and some additional ideas on the way. Ironically, some of these will probably shake up a bit what you have learnt so far, but I hope they will also stimulate some curiosity for what else is out there to discover and explore.

Let me first mention a few areas that you should begin to explore as you become more familiar with regression modeling. One issue I have only alluded to in passing in the code file is that of *(cross) validation*. Regressions often run the risk of what is called *overfitting*: they fit a particular data set rather well, but generalize badly to others, which of course jeopardizes the generalizability of the findings to the population as a whole. Very often, results can be validated by splitting up the existing sample into, often, 10 parts and then do 10 analyses, in each of which you obtain a regression equation from 90% of the data and apply it to the unseen 10%. Such methods can reveal a lot about the internal structure of a data set and there are several functions available in R for these methods. A related point is that, given the ever increasing power of computers, resampling and permutation approaches become more and more popular; examples include the *bootstrap*, the *jackknife* procedure, or *exhaustive permutation procedures*. These procedures are non-parametric methods you can use to estimate means, variances, but also correlations or regression parameters without major distributional assumptions. Such methods are not the solution to all statistical problems, but can still be interesting and powerful tools (cf. the libraries `boot` as well as `bootstrap`).

Recommendation(s) for further study
Good (2005), Rizzo (2008: Ch. 7, 8)

Also, the analysis of special data points in your sample(s) is very important, given the impact that *outliers* and *points with high leverage* can have on the data. In addition, learning more about what to do with missing data should be high on your list of things. On the one hand, it may be useful, for instance, to run a regression on missing data to see whether there is something in the data that allows you to predict well when, say, subjects do

respond to a stimulus. On the other hand, small proportions of missing data may be *imputed*, that is predicted from other data points (see Torgo: Section 2.5).

Then, there is a range of additional techniques you may wish to explore. This book focused on hypothesis-testing approaches, in particular regressions, but there are many interesting exploratory tools that, for reasons of space, I could not discuss: *principal components analysis* and *correspondence analysis* are two well-known cases in point, *association rules* or *naïve Bayes classifiers* are others.

It is also worth pointing out that R has many many more possibilities of graphical representation than I could mention here. I only used the traditional graphics system, but there are other more powerful tools, which are available from the libraries `lattice` and `ggplot2` (you should explore <http://www.yeroon.net/ggplot2/>). The website <http://gallery.r-enthusiasts.com/> provides many very interesting and impressive examples for R plots, and several good books illustrate many of the exciting possibilities for exploration (cf. Unwin, Theus, and Hofmann 2006, Cook and Swayne 2007, Sarkar 2008, Keen 2010, and of course Murrell 2011).

Finally, note that the *null hypothesis significance testing (NHST) paradigm* that is underlying most of the methods discussed here is not as uncontroversial as this textbook (and most others) may make you believe. While the computation of *p*-values is certainly still the standard approach, there are researchers who argue for a different perspective. Some of these argue that *p*-values are problematic because they do in fact not represent the conditional probability that one is really interested in. Recall, the above *p*-values answer the question "How likely is it to get the observed data when H_0 is true?" but what one actually wants to know "How likely is H_1 given the data I have?" Suggestions for improvement include:

- one should focus not on *p*-values but on effect sizes and/or confidence intervals (which is why I mentioned these above again and again);
- one should report so-called p_{rep}-values, which according to Killeen (2005) provide the probability to replicate an observed effect (but are not uncontroversial themselves);
- one should test reasonable H_0s rather than hypotheses that could never be true in the first place (there will always be some effect or difference).

Another interesting approach is the so-called *Bayesian approach* to statistics, which allows to include subjective prior knowledge or previous results with one's own data. All of these things are worth exploring.

Recommendation(s) for further study
- Cohen (1994), Loftus (1996), Denis (2003) for discussion of the NHST
- Killeen (2005) on p_{rep}-values
- Iversen (1984) on Bayes statistics

I hope you can use the techniques covered in this book for many different questions, and when this little epilog also makes you try and extend your knowledge and familiarize yourself with additional tools and methods – for example, there are many great web resources, <http://www.statmethods.net/index.html> and <http://www.r-bloggers.com/> are among my favorites – then this book has achieved one of his main objectives.