R - Statistical Modelling

Some slides that appear in this presentation were obtained from Thomas Lumley of the R-core group (http://www.r-project.org/) and the Bioconductor webpage (http://www.bioconductor.org/).

Statistical Modelling in S

The systematic part of a model is specified a model formula with basic structure

\[
\text{outcome} \sim \text{exposure} \ast \text{modifier} + \text{confounder}
\]

- The left-hand side is the outcome (response, independent) variable, the right-hand side describes the predictors.
- The \( \ast \) specifies an interaction and the corresponding main effects (\( a:b \) specifies just the interaction term).
- Factors (eg race, subtype of disease) are coded by default with indicator variables for all except the first category.
- Terms can be variables, simple expressions, or composite objects.
Statistical Modelling in S

- \( \text{depress} \sim \text{rural}^* \text{agegp} + \text{partner} + \text{parity} + \text{income} \)
  Does the risk of postnatal depression vary between urban and rural areas, separately for each age group, adjusted for having a domestic partner, previous number of pregnancies, income?

- \( \text{asthma} \sim \text{pm25} + \text{temp} + I(\text{temp}^2) + \text{month} \)
  How does the number of hospital admissions for asthma vary with fine particulate air pollution, adjusted for temperature and month of the year?

- \( \log(\text{pm25}) \sim \text{temp} + \text{stag} + \text{month} + \text{lag(tem}p,1) \)
  Predict (log-transformed) fine particulate air pollution from temperature, air stagnation, month, and yesterday’s temperature.

- \( \text{Surv(ttoMI,MI)} \sim \text{LDL} + \text{age} + \text{sex} + \text{hibp} + \text{diabetes} \)
  How does LDL cholesterol predict (time to) myocardial infarction after adjusting for age, sex, hypertension, and diabetes?

**Generalized linear models**

Generalized linear models (linear regression, logistic regression, poisson regression) are handled by the \( \text{glm}() \) function. This requires:

- A model formula
- A dataframe containing the variables [optional]
- A model family:
  - \( \text{binomial()} \) logistic regression,
  - \( \text{gaussian()} \) linear regression,
  - \( \text{poisson()} \) Poisson regression,
    and others less commonly used.

\( \text{glm(} \text{asthma} \sim \text{pm25} + \text{temp} + I(\text{temp}^2) + \text{month}, \text{data=pmdat, family=poisson()} \)
Model objects

Typical statistics packages fit a model and output the results. In S a model object is created that stores all the information about the fitted model. Coefficients, diagnostics, and other model summaries are produced by methods for this object.

- `coef(model)` returns the coefficients.
- `summary(model)` gives a table with coefficients, standard errors, perhaps other information.
- `resid(model)` returns (various flavours of) residuals.
- `anova(model)` gives an ANOVA table showing likelihood ratio tests for adding each term sequentially. Also, the function `anova(model1, model2)` compares the two models directly.
- `plot(model)` may give some useful diagnostic plots.
- many more!

Classes of model

R has most of the commonly used regression models:

- `lm()` Linear regression.
- `glm()` generalized linear models.
- `coxph()` Cox model (in “survival” package).
- `survreg()` Parametric survival models (in “survival” package).
- `clogit()` Conditional logistic regression (in “survival” package).
- `lme()` Linear mixed models (in “nlme” package).
- simple meta-analyses (in package “rmeta”).

Note: Currently, R does not have generalized linear mixed models (SAS PROC NLMIXED, BUGS, HLM, MLwin).