Linear and Generalized Linear Models

John Maindonald

August 7, 2007

1 Linear Models

- $\boldsymbol{y}$ (n by 1) is a vector of observed values, $\boldsymbol{X}$ (n by p) is model matrix, and $\boldsymbol{\beta}$ (p by 1) is a vector of coefficients.
- The model is $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, i.e. $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ where the vector $\boldsymbol{\epsilon}$ of residuals is n by 1
- Least squares *normal* equations are

  $$\boldsymbol{X}'\boldsymbol{X\beta} = \boldsymbol{X}'\boldsymbol{y}$$

  (assuming $\boldsymbol{\epsilon}_i$ are iid normal, these are the maximum likelihood estimates)
- If variances are unequal, modify *normal* equations to

  $$\boldsymbol{X}'\boldsymbol{WX\beta} = \boldsymbol{X}'\boldsymbol{Wy}$$

  where $\boldsymbol{W}$ is a diagonal matrix with elements equal to the inverses of the variances (justification is from maximum likelihood, or argue that leverage should be independent of variance)
- Assume $E[\boldsymbol{y}] = \mu = \boldsymbol{X\beta}$, i.e. $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.

1.1 Linear Models – general variance-covariance matrix

More generally, if $\boldsymbol{\epsilon}$ is multivariate normal with known variance-covariance matrix $\boldsymbol{\Sigma}$, then ML theory gives the equation as above with $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$.

Two values with a high positive correlation contain, jointly, less information than two independent values. Consider an extreme case; if the correlation is 1, they duplicate the same information.

If the variance-covariance matrix $\boldsymbol{\Sigma}$ is not known, many different special methods are available for various special cases that occur in pracitce. Models that may be relevant include time-series models, spatial analysis and multi-level models.

1.2 Least squares computations

A separate set of notes describes the approach, based on the QR matrix decomposition, that is used in R and in most of the R packages. Where methods that are directly based on QR are too slow, there may be a specialized method that takes advantage of structure in $\boldsymbol{X}$ to greatly speed up computation. Sparse least squares is an important special case, See Bates (2006); Koenker and & Ng (2003).

2 Generalized Linear Models

- As before, we have $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{y}]$ (n by 1), $\boldsymbol{X}$ (n by p), $\boldsymbol{\beta}$ (p by 1), and $\boldsymbol{\epsilon}$ (n by 1).
- The model is now
  $$f(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}, \quad \text{where } \mathrm{E}[\boldsymbol{y}] = \boldsymbol{\mu}$$
  Here, $f()$, which must be monotonic, has the name *link* function. For example,
  $$f(\boldsymbol{\mu}_i) = \log(\frac{\boldsymbol{\mu}_i}{N_i - \boldsymbol{\mu}_i})$$
- The distribution of y_i is a function of the predicted value μ_i, independently for different observations. The different y_i are from the same parametric family, but the distributions are not identical.
- An extension is to the quasi-exponential family, where the variance is a constant multiple of an exponential family variance. The multiplying constant is estimated as part of the analysis.
- Commonly used distributions are the normal, binomial and Poisson. Applications for models with quasibinomial and quasipoisson errors may if anything be more extensive than for their exponential family counterparts.
- GLMs with binomial errors are formally equivalant to discriminant models where there are two categories. The GLM framework has advantages for some problems.
- 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 value`s there are specific circumstances where it is reasonable to treat them `z value`s as t-statistics. [More technically, they are Wald statistics.]

2.1 Maximum likelihood parameter estimates

- Recall that the equation is
  $$f(\boldsymbol{\mu}) = E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$$
  where $\boldsymbol{\mu} = E[\boldsymbol{y}]$
- Assuming a distribution from the exponential family, the maximum likelihood estimates of the parameters are given by
  $$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{\mu} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$$
  where $f(\boldsymbol{\mu}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$
- Note that the (diagonal) element $\boldsymbol{W}_{ii}$ of $\boldsymbol{W}$ are functions both of var$[\boldsymbol{y}_i]$ 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.
- Just as for linear models, spline terms can be fitted.

2.2 Theoretical approximations

- Except in special cases, the statistical properties of parameters rely on asymptotic results. Standards errors and t-statistics rely on first-order Taylor series approximations that, in the worst case, can fail badly. This applies, especially, to binary logistic regression.
- For logistic regression models, and Poisson models with small expected values, assessments of predictive accuracy should be derived using a resampling approach, perhaps cross-validation.

3 Selection of Regression Models

When the model is fitted to the data used to select the model from a set of possible models, 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. Use of test data that are separate from data used to develop the model deals with this issue.

There is a further important issue, that use of separate test data does not address. Almost inevitably, none of the models on offer will be strictly correct. Mis-specification of the fixed effects, and to a lesser extent of the random effects, is likely to bias model estimates, at the same time inflating the error variance or variances, i.e., it may to some extent work in the opposite direction to selection effects.

References

Bates, D, 2007. Comparing least squares calculations. *Vignette "Comparisons" accompanying the package "Matrix" for R.*

Koenker, R and Ng, P, 2003. SparseM: A sparse matrix package for R. *Journal of Statistical Software* 8(6).

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.]