Supplement A to Oron et al.: Details and Formulae for Linear-Model Diagnostics

July 29, 2008

Residuals

As written in the article, the ordinary or "response" residuals $e_{gi} \equiv y_{gi} - \hat{y}_{gi}$ are inadequate for aggregation over many genes, and they need to be normalized. Following is a description of standard approaches to such normalization, based mostly on Cook and Weisberg (1982). All residual types discussed here are available from the `getResidPerGene` function in the GSEAlm package.

The naive approach to normalization would be to simply divide the e_{gi} by the estimate of residual standard deviation:

$$r_{gi}^{(naive)} \equiv \frac{e_{gi}}{S_g}, \tag{A-1}$$

where

$$S_g \equiv \sqrt{\frac{\sum_l e_{gl}^2}{n-p-1}}, \tag{A-2}$$

with p being the number of covariates in the model. These residuals are obtained from the `getResidPerGene` function by setting `type='normalized'` (ordinary residuals are available via `type='response'`).

Equation (A-1) may seem analogous to simple normalization via the sample s.d., and therefore accurate. However, careful inspection of the regression process reveals that this is not the case. To see why, we need to introduce matrix notation. Let $\mathbf{X}$ be the model matrix – i.e., a matrix with all covariate values arranged by sample – having dimensions n by $p+1$ (the first column of $\mathbf{X}$ represents the intercept and is all 1's). The regression model for a single gene (temporarily omitting the g subscript) can be written in matrix form:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon. \tag{A-3}$$

The fitted values in matrix notation are

$$\hat{\mathbf{Y}} = \mathbf{HY} \quad where \quad \mathbf{H} \equiv \mathbf{X}\left(\mathbf{X^TX}\right)^{-1}\mathbf{X^T}. \tag{A-4}$$

$\mathbf{H}$ is known as "the hat matrix". The ordinary residuals in matrix form are $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, with $\mathbf{I}$ being the identity n by n matrix. The values on $\mathbf{H}$'s diagonal are positive and therefore, the variance of the e_{gi} is smaller than $Var(Y)$, for which S^2 is an unbiased estimate. Using "the hat matrix", though, we can correct the bias, yielding what are known as **internally-Studentized residuals:**

$$r_{gi}^{(int.)} \equiv \frac{e_{gi}}{S_g\sqrt{1-h_{ii}}}, \tag{A-5}$$

where h_{ii} is the i-th diagonal element of $\mathbf{H}$. These residuals are obtained from `getResidPerGene` via `type='intStudent'`, and their value is bounded. If one calculates S_g in the denominator of (A-5) without including e_{gi}, the resulting residuals are known as “leave-one-out” or “externally-Studentized” residuals. They have several good properties:

- They are unbounded, unlike the naively-normalized and internally-Studentized residuals;
- The numerator and denominator are independent;
- If model assumptions for i.i.d. normal errors hold, they are t distributed with $n-p-2$ d.f., and function as immediate tests for “shift” outliers - i.e., for observations whose true mean systematically deviates from the model-predicted mean.

For these reasons, the externally-Studentized residuals

$$r_{gi}^{(ext.)} \equiv \frac{e_{gi}}{S_{g(i)}\sqrt{1-h_{ii}}}, \quad with \quad S_{g(i)} \equiv \sqrt{\frac{\sum_{l\neq i} e_{gl}^2}{n-p-2}}, \tag{A-6}$$

are the default option for `getResidPerGene`.

Cook’s D and Related Influence Measures

Residuals inform us whether or not certain samples deviate from the model. A related but different question is whether individual samples have disproportionate effect upon model fit. The simplest and most commonly-used measure is Cook’s D. In matrix notation and suppressing the gene index, an observation’s D value is defined as

$$D_i \equiv \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T (\mathbf{X}^\mathbf{T}\mathbf{X}) \left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{(p+1)S^2}, \tag{A-7}$$

where $\hat{\beta}$ is the vector of covariate estimates, and the subscript (i) indicates, as above, calculation with the i-th sample omitted. D is a square norm in p-dimensional space, and its “leave-one-out” type calculation may seem a formidable task. However, the calculation is simplified greatly via the shortcut identity

$$D_i = \frac{\left(r_i^{(int.)}\right)^2 h_{ii}}{(p+1)(1-h_{ii})}. \tag{A-8}$$

This formula exposes influence as a direct function of normalized-residual magnitude and hat-matrix magnitude (a.k.a. ”leverage”).

Two related influence measures are $DFFITS$ and $DFBETAS$, quantifying the amount by which an individual observation affects fitted values and individual parameter estimates, respectively.[1] All these functions are implemented for matrix input in the GSEAlm package.

References

R.D. Cook and S. Weisberg. *Residuals and Influence in Regression.* Monographs on Statistics and Applied Probability. Chapman and Hall, 1982.

[1] $DFBETAS$ is essentially a decomposition of Cook’s D into single-parameter components. However, due to historical conventions the decomposition is not direct, but multiplied by a function of $\mathbf{X}$ and h_{ii}.