

Homework #7 – Simulation Study Problems

ST790R

02 November 2009

1 Logistic Regression via Iterative GLS

The logistic regression model from Section 9.x has y_i independent Binomial(m_i, π_i) random variables, where $\pi_i = \exp\{\beta^T \mathbf{x}_i\} / (1 + \exp\{\beta^T \mathbf{x}_i\})$ for $i = 1, \dots, n$. The usual (Newton, scoring) method for computing the MLE is Iteratively Reweighted Least Squares or IRWLS. Notice that this is not GLS, as GLS would be minimizing

$$S(\beta) = \sum_i^n \frac{(y_i - m_i \pi_i)^2}{\text{Var}(y_i)},$$

where $\text{Var}(y_i) = m_i \pi_i (1 - \pi_i)$. An iterative GLS estimator that minimizes $S(\beta)$ can be computed using a general optimizer (like *nlm*) or a nonlinear least squares algorithm (like *nls*). (Don't use algorithm as a factor.) Compare this estimator with the MLE. (Note that these estimators will be the same in a saturated model.)

2 Asymptotic normality of L1 Regression

For the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

$i = 1, \dots, n$, an alternative to the usual least squares estimates are L_1 regression estimates, the values $\hat{\beta}$ (not the usual) that minimize the L_1 norm between \mathbf{y} and its mean vector $\mathbf{X}\beta$:

$$S(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_1,$$

which for the simple linear regression problem simplifies to $S(\beta_0, \beta_1) = \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|$.

The asymptotic theory says that $\hat{\beta}$ is asymptotically normal with mean vector β and covariance matrix $(2f(0))^{-2}(\mathbf{X}^T \mathbf{X})^{-1}$, where $f(0)$ is the density of the *iid* errors at 0. Some of the issues which could be explored are the sample size needed to achieve normality, whether the asymptotic standard deviations are valid, whether the design has much of an effect, or the effect of the error density.

(See the code *slr1.r* for a demonstration of using *optimize* to minimize $S(\hat{\beta}_0(\beta_1), \beta_1)$ to compute the minimum of $S(\beta_0, \beta_1)$ for the case of simple linear regression.) (Multiple regression problems can be written as a linear programming problem.)

3 Power Loss when Overfitting

In linear models, the main consequence of overfitting (including covariates whose coefficients are zero) is the increased variances of least squares estimators due to multicollinearity. How does overfitting affect the power of hypothesis tests? Some suggested situations include analysis of covariance, interaction in two-way ANOVA models, or multiple regression with many covariates.

4 Measurement Error

In simple linear regression, $y_i \sim Normal(\beta_0 + \beta_1 x_i, \sigma^2)$, if the covariate x_i is measured with error, say, we observe only $w_i = x_i + z_i$, where z_i are iid $N(0, \sigma_z^2)$, then the usual least squares estimate of the slope is biased downward. (See Example 5.4 in the Linear Models book.) Some suggestions for analysis are: the effect of measurement error on the bias of estimates, on the coverage of confidence intervals, or on the level and power of tests, in the simple linear regression case, or in multiple regression where one variable is measured with error, or in analysis of covariance.

5 Wald vs LRT in Nonlinear Regression

One of the examples we will discuss in simulation is a comparison of different methods for constructing standard errors in nonlinear regression. Choose one of these methods and compare Wald tests based on these standard errors to Likelihood Ratio Tests. Some suggestions are from the nonlinear models you've seen: 1) one of the explanatory variables in *nllsqt2* is a dummy/group variable, 2) in the orange tree data(*oranget3*), one issue is whether the growth rate parameters were different across five trees, 3) in the Yafune, et al., data, again the rate parameters might be the same across subjects with the same dosage.