so when you first `g.data.attach` the directory, the object loaded into memory as `x1` is a "promise object" (which is very small). When you actually use `x1` (e.g., to query its dimensions), the promise is fulfilled, and `g.data.load` does two things:

1. It loads the actual large object, and
2. It returns that object for the query.

Henceforth, the object in memory as `x1` is the real (large) object.

David E. Brahm
Geode Capital Management
brahm@alum.mit.edu

geepack: Yet Another Package for Generalized Estimating Equations

Modeling Both Mean and Association of Multivariate Responses

by Jun Yan

Introduction

geepack is designed to provide an inferential basis for both the association structure and the mean structure in multivariate analysis, using the Generalized Estimating Equations (GEE) approach.

Consider a sample of K independent clusters $y_i^T = (y_{i1}, \cdots, y_{in_i})$, $i = 1, \cdots, K$, of n_i-variate responses. In a generalized linear model setup, the variance of y_{it}, V_{it}, can be factored as

$$\mathrm{var}(y_{it}) = \phi_{it} v(\mu_{it}),$$

where ϕ_{it} is the scale parameter, v is the variance function $v(\mu_{it})$, where $\mu_{it} = E(y_{it})$. To model the association, we decompose $\mathrm{cov}(y_i)$ into two parts, the variance and the correlation,

$$\mathrm{cov}(y_i) = V^{1/2} R V^{1/2},$$

where V is the diagonal matrix of V_{it}, and R is the correlation matrix of y_i.

Let X_{1i}, X_{2i} and X_{3i} be the covariate matrices for the mean, the scale, and the correlation of the response y_i, with dimensions $n_i \times p$, $n_i \times r$, and $n_i(n_i - 1)/2 \times q$, respectively. The models are

$$\begin{aligned} g_1(\mu_i) &= X_{1i}\beta, & (1)\\ g_2(\phi_i) &= X_{2i}\gamma, & (2)\\ g_3(\rho_i) &= X_{3i}\alpha, & (3) \end{aligned}$$

where $g_i, i = 1, 2, 3$, are known link functions, μ_i is a $n_i \times 1$ vector containing $E(y_i|X_{1i})$, ϕ_i is a $n_i \times 1$ vector containing $\mathrm{var}(y_i|X_{2i})/v_{it}$, where $v_{it} = v(\mu_{it})$ is the variance function, and ρ_i is a $n_i(n_i - 1)/2 \times 1$ vector containing $\mathrm{cor}(y_{is}, y_{it}|X_{3i})$. β, γ, and α are the mean, the scale, and the correlation parameters of dimension $p \times 1$, $r \times 1$, and $q \times 1$, respectively.

The mean link has been well studied. The scale link is often taken to be log, while it is natural to let the correlation link be "logistic" (i.e., Fisher's z transformation), in which case the inverse link function is the hyperbolic tangent, that is,

$$\rho_{its} = \mathrm{cor}(y_{is}, y_{it}|X_{3i}) = \frac{\exp(X_{3i(s,t)}\alpha) - 1}{\exp(X_{3i(s,t)}\alpha) + 1}, \qquad (4)$$

where $X_{3i(s,t)}$ is the row in matrix X_{3i} corresponding to the correlation of y_{is} and y_{it}. These links ensure that the scale is positive and that the correlation is in $(-1, 1)$. The scale model is useful in situations where parameters are needed for covariate effects either on over- or under-dispersion or on heteroscedasticity.

A convenient set of estimating equations for the three-link model is

$$U_1(\beta, \gamma, \alpha) = \sum_{i=1}^{K} D_{1i}^T V_{1i}^{-1}(y_i - \mu_i) = 0 \qquad (5)$$

$$U_2(\beta, \gamma, \alpha) = \sum_{i=1}^{K} D_{2i}^T V_{2i}^{-1}(s_i - \phi_i) = 0 \qquad (6)$$

$$U_3(\beta, \gamma, \alpha) = \sum_{i=1}^{K} D_{3i}^T V_{3i}^{-1}(z_i - \rho_i) = 0 \qquad (7)$$

where s_i is the $n_i \times 1$ vector of $s_{it} = (y_{it} - \mu_{it})^2/v_{it}$, z_i is the $n_i(n_i - 1)/2 \times 1$ vector of $z_{its} = (y_{it} - \mu_{it})(y_{is} - \mu_{is})/\sqrt{\phi_{it} v_{it} \phi_{is} v_{is}}$, $D_{1i} = \partial\mu_i/\partial\beta^T$, $D_{2i} = \partial\phi_i/\partial\gamma^T$, $D_{3i} = \partial\rho_i/\partial\alpha^T$, and V_{1i}, V_{2i} and V_{3i} are the conditional working covariance matrices of y_i, s_i, and z_i.

The matrix V_{1i} generally contains scale parameters γ and correlation parameters α. The matrices V_{2i} and V_{3i} may contain other estimated quantities which characterize the third and fourth order moments. In practice, in order to avoid specification of higher order moments, estimation of higher order nuisance parameters, and convergence problems, V_{2i} may be chosen to be a diagonal matrix whose diagonal elements are $2\phi_{it}$, following the independence Gaussian working matrix in Prentice and Zhao (1991), and V_{3i} may be an identity matrix (Ziegler et al., 1998, p.129), at the cost of potential efficiency loss. These simplifications are implemented in **geepack**.

Features

- Allows different covariates in separate models for the mean, scale, and correlation via various link functions.
- Provides "sandwich" and jackknife variance estimators for all the parameter estimates, extending Ziegler et al. (2000).
- Handles clustered ordinal data, allowing covariates in the odds ratio model, using the method in Heagerty and Zeger (1996).

An example: Epileptic seizures

As an illustration, the epileptic seizure data (Thall and Vail, 1990) is analyzed. The dataset arose from a clinical trial of 59 epileptic patients, randomized to receive either the anti-epileptic drug pragabide or a placebo, as an adjuvant to standard chemotherapy. There are four 2-week interval seizure counts for each patient. The covariates are treatment, age, and baseline counts on a 8-week interval before the trial. We first reshape the data into a "long" format for longitudinal data, and create new covariates identical to those used in Thall and Vail (1990),

```
> data(seizure)
> seiz.l <-
+   reshape(seizure,
+           varying = list(c("y1", "y2",
+                            "y3", "y4")),
+           v.names = "y", direction = "long")
> seiz.l <-
+   seiz.l[order(seiz.l$id, seiz.l$time),]
> seiz.l$lbase <- log(seiz.l$base / 4)
> seiz.l$lage <- log(seiz.l$age)
> seiz.l$v4 <- ifelse(seiz.l$time == 4, 1, 0)
```

Next we use the function geese to fit a GEE model for the seizure counts, with the same mean model as that in Thall and Vail (1990). We treat time as a factor and include it in the scale model using a `log` link. To illustrate the usage of the correlation model, we use an `ar1` correlation structure (together with the scale model, this specifies a heterogeneous AR(1) covariance structure), and `fisherz` link, and include patient age in the correlation model. The models for the mean, scale, and correlation are fit using GEE and the results are summarized in the following.

```
> z <- model.matrix(~ age, data = seizure)
> m1 <- geese(y ~ lbase*trt + lage + v4,
+             sformula = ~ as.factor(time) - 1,
+             id = id, data = seiz.l,
+             corstr = "ar1", family = poisson,
+             zcor = z, cor.link = "fisherz",
+             sca.link = "log")
> summary(m1)

Mean Model:
 Mean Link:                 log
 Variance to Mean Relation: poisson

 Coefficients:
            estimate san.se   wald        p
(Intercept)   -2.544 0.8291   9.41 0.002154
lbase          0.964 0.0898 115.22 0.000000
trt           -1.491 0.4557  10.71 0.001068
lage           0.826 0.2387  11.98 0.000539
v4            -0.143 0.0721   3.95 0.046850
lbase:trt      0.601 0.1864  10.38 0.001272

Scale Model:
 Scale Link:                log

 Estimated Scale Parameters:
                 estimate san.se wald        p
as.factor(time)1    1.240  0.255 23.6 1.18e-06
as.factor(time)2    1.544  0.366 17.8 2.49e-05
as.factor(time)3    2.019  0.498 16.5 4.98e-05
as.factor(time)4    0.864  0.204 18.0 2.24e-05

Correlation Model:
 Correlation Structure:     ar1
 Correlation Link:          fisherz

 Estimated Correlation Parameters:
            estimate san.se  wald        p
(Intercept)    2.558 0.7064 13.11 0.000294
age           -0.047 0.0229  4.21 0.040192

Returned Error Value:     0
Number of clusters: 59  Maximum cluster size: 4
```

Since there is one possible outlier in the dataset (Diggle et al., 1994, pp.166–168), it might be interesting to compare the "sandwich" variance estimate with the jackknife variance estimate. Jackknife variance estimate may be obtained by setting `jack`, `j1s`, or `fij` to `TRUE`, requesting approximated, one-step, and fully iterated jackknife variance estimate, respectively; see Ziegler et al. (2000).

```
> m2 <- geese(y ~ lbase*trt + lage + v4,
+             sformula = ~ as.factor(time) - 1,
+             id = id, data = seiz.l,
+             corstr = "ar1", family = poisson,
+             zcor = z, cor.link = "fisherz",
+             sca.link = "log", jack = TRUE,
+             j1s = TRUE, fij = TRUE)
```

Summarizing the fitted object (not shown here) suggests that there is noticeable difference between the sandwich and the jackknife variance estimator for the covariate effect of `trt` and `lbase:trt`. If the jackknife variance estimators were used, these two effects would become insignificant at level 0.05.

Future developments

Different components within a cluster may have different link functions. For example, the data analyzed by Prentice and Zhao (1991) have two responses for each patient. One is continuous and its mean is modeled with the identity link, and the other is binary and its mean is modeled with the logit link. The C++ code for **geepack** was designed to permit this situation. An R interface will be developed for this extension.

Acknowledgments

The C++ code in **geepack** is based on the Template Numerical Toolkit (TNT) version 0.94, developed at the National Institute of Standards and Technology (NIST). At the time of writing, version 1.1 is available at http://math.nist.gov/tnt/.

I am grateful to Douglas Bates and Jason Fine for encouragement, discussions, and comments.

Bibliography

Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger. *Analysis of longitudinal data (ISBN 0198522843)*. Clarendon Press [Oxford University Press], 1994. ISBN 0198522843. 13

Patrick J. Heagerty and Scott L. Zeger. Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association*, 91:1024–1036, 1996. 13

Ross L. Prentice and Lue Ping Zhao. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47:825–839, 1991. 12, 14

Peter F. Thall and Stephen C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46:657–671, 1990. 13

Andreas Ziegler, Christian Kastner, and Maria Blettner. The generalised estimating equations: An annotated bibliography. *Biometrical Journal*, 40:115–139, 1998. 12

Andreas Ziegler, Christian Kastner, Daniel Brunner, and Maria Blettner. Familial associations of lipid profiles: A generalized estimating equations approach. *Statistics in Medicine*, 19(24):3345–3357, 2000. 13

Jun Yan
University of Wisconsin–Madison, U.S.A.
jyan@stat.wisc.edu

On Multiple Comparisons in R

by Frank Bretz, Torsten Hothorn and Peter Westfall

Description

The multiplicity problem arises when several inferences are considered simultaneously as a group. If each inference has a 5% error rate, then the error rate over the entire group can be much higher than 5%. This article shows practical examples of multiple comparisons procedures that control the error of making any incorrect inference.

The **multcomp** package for the R statistical environment allows for multiple comparisons of parameters whose estimates are generally correlated, including comparisons of *k* groups in general linear models. The package has many common multiple comparison procedures “hard-coded”, including Dunnett, Tukey, sequential pairwise contrasts, comparisons with the average, changepoint analysis, Williams’, Marcus’, McDermott’s, and tetrad contrasts. In addition, a free input interface for the contrast matrix allows for more general comparisons.

The comparisons themselves are not restricted to balanced or simple designs. Instead, the package is designed to provide general multiple comparisons, thus allowing for covariates, nested effects, correlated means, likelihood-based estimates, and missing values. For the homoscedastic normal linear models, the functions in the package account for the correlations between test statistics by using the exact multivariate *t*-distribution. The resulting procedures are therefore more powerful than the Bonferroni and Holm methods; adjusted p-values for these methods are reported for reference. For more general models, the program accounts for correlations using the asymptotic multivariate normal distribution; examples include multiple comparisons based on rank transformations, logistic regression, GEEs, and proportional hazards models. In the asymptotic case, the user must supply the estimates, the asymptotic covariance matrix, and the contrast matrix.

Basically, the package provides two functions. The first, `simint`, computes confidence intervals for the common single-step procedures. This approach is uniformly improved by the second function (`simtest`), which utilizes logical constraints and is closely related to closed testing. However, no con-