0.1 `logit.gee`: Generalized Estimating Equation for Logistic Regression

The GEE logit estimates the same model as the standard logistic regression (appropriate when you have a dichotomous dependent variable and a set of explanatory variables). Unlike in logistic regression, GEE logit allows for dependence within clusters, such as in longitudinal data, although its use is not limited to just panel data. The user must first specify a "working" correlation matrix for the clusters, which models the dependence of each observation with other observations in the same cluster. The "working" correlation matrix is a $T \times T$ matrix of correlations, where T is the size of the largest cluster and the elements of the matrix are correlations between within-cluster observations. The appeal of GEE models is that it gives consistent estimates of the parameters and consistent estimates of the standard errors can be obtained using a robust "sandwich" estimator even if the "working" correlation matrix is incorrectly specified. If the "working" correlation matrix is correctly specified, GEE models will give more efficient estimates of the parameters.

Syntax

```
> z.out <- zelig(Y ~ X1 + X2, model = "logit.gee",
                 id = "X3", data = mydata)
> x.out <- setx(z.out)
> s.out <- sim(z.out, x = x.out)
```

where `id` is a variable which identifies the clusters. The data should be sorted by `id` and should be ordered within each cluster when appropriate.

Additional Inputs

- `robust`: defaults to `TRUE`. If `TRUE`, consistent standard errors are estimated using a "sandwich" estimator.

Use the following arguments to specify the structure of the "working" correlations within clusters:

- `corstr`: defaults to `"independence"`. It can take on the following arguments:
  - Independence (`corstr = "independence"`): $\text{cor}(y_{it}, y_{it'}) = 0$, $\forall t, t'$ with $t \neq t'$. It assumes that there is no correlation within the clusters and the model becomes equivalent to standard logistic regression. The "working" correlation matrix is the identity matrix.
  - Fixed (`corstr = "fixed"`): If selected, the user must define the "working" correlation matrix with the `R` argument rather than estimating it from the model.

– Stationary m dependent (`corstr = "stat_M_dep"`):

$$\text{cor}(y_{it}, y_{it'}) = \begin{cases} \alpha_{|t-t'|} & \text{if} \quad |t-t'| \leq m \\ 0 & \text{if} \quad |t-t'| > m \end{cases}$$

If (`corstr = "stat_M_dep"`), you must also specify `Mv` $= m$, where m is the number of periods t of dependence. Choose this option when the correlations are assumed to be the same for observations of the same $|t-t'|$ periods apart for $|t-t'| \leq m$.

Sample "working" correlation for Stationary 2 dependence (`Mv`=2)

$$\begin{pmatrix} 1 & \alpha_1 & \alpha_2 & 0 & 0 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 & 0 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ 0 & \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ 0 & 0 & \alpha_2 & \alpha_1 & 1 \end{pmatrix}$$

– Non-stationary m dependent (`corstr = "non_stat_M_dep"`):

$$\text{cor}(y_{it}, y_{it'}) = \begin{cases} \alpha_{tt'} & \text{if} \quad |t-t'| \leq m \\ 0 & \text{if} \quad |t-t'| > m \end{cases}$$

If (`corstr = "non_stat_M_dep"`), you must also specify `Mv` $= m$, where m is the number of periods t of dependence. This option relaxes the assumption that the correlations are the same for all observations of the same $|t-t'|$ periods apart.

Sample "working" correlation for Non-stationary 2 dependence (`Mv`=2)

$$\begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & 0 & 0 \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} & 0 \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} & \alpha_{35} \\ 0 & \alpha_{24} & \alpha_{34} & 1 & \alpha_{45} \\ 0 & 0 & \alpha_{35} & \alpha_{45} & 1 \end{pmatrix}$$

– Exchangeable (`corstr = "exchangeable"`): $\text{cor}(y_{it}, y_{it'}) = \alpha$, $\forall t, t'$ with $t \neq t'$. Choose this option if the correlations are assumed to be the same for all observations within the cluster.

Sample "working" correlation for Exchangeable

$$\begin{pmatrix} 1 & \alpha & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & \alpha & 1 \end{pmatrix}$$

– Stationary mth order autoregressive (`corstr = "AR-M"`): If (`corstr = "AR-M"`), you must also specify `Mv` $= m$, where m is the number of periods t of dependence. For example, the first order autoregressive model (AR-1) implies $\text{cor}(y_{it}, y_{it'}) = \alpha^{|t-t'|}, \forall t, t'$ with $t \neq t'$. In AR-1, observation 1 and observation 2 have a correlation of α. Observation 2 and observation 3 also have a correlation of α. Observation 1 and observation 3 have a correlation of α^2, which is a function of how 1 and 2 are correlated (α) multiplied by how 2 and 3 are correlated (α). Observation 1 and 4 have a correlation that is a function of the correlation between 1 and 2, 2 and 3, and 3 and 4, and so forth.

Sample "working" correlation for Stationary AR-1 (`Mv`=1)

$$\begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 \\ \alpha & 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha^2 & \alpha & 1 & \alpha & \alpha^2 \\ \alpha^3 & \alpha^2 & \alpha & 1 & \alpha \\ \alpha^4 & \alpha^3 & \alpha^2 & \alpha & 1 \end{pmatrix}$$

– Unstructured (`corstr = "unstructured"`): $\text{cor}(y_{it}, y_{it'}) = \alpha_{tt'}, \forall t, t'$ with $t \neq t'$. No constraints are placed on the correlations, which are then estimated from the data.

- `Mv`: defaults to 1. It specifies the number of periods of correlation and only needs to be specified when `corstr` is `"stat_M_dep"`, `"non_stat_M_dep"`, or `"AR-M"`.
- `R`: defaults to `NULL`. It specifies a user-defined correlation matrix rather than estimating it from the data. The argument is used only when `corstr` is `"fixed"`. The input is a $T \times T$ matrix of correlations, where T is the size of the largest cluster.

Examples

1. Example with Stationary 3 Dependence

   Attaching the sample turnout dataset:

   ```
   > data(turnout)
   ```

   Variable identifying clusters

   ```
   > turnout$cluster <- rep(c(1:200), 10)
   ```

   Sorting by cluster

   ```
   > sorted.turnout <- turnout[order(turnout$cluster), ]
   ```

   Estimating parameter values for the logistic regression:

```
> z.out1 <- zelig(vote ~ race + educate, model = "logit.gee", id = "cluster",
+     data = sorted.turnout, robust = TRUE, corstr = "stat_M_dep",
+     Mv = 3)
```

Setting values for the explanatory variables to their default values:

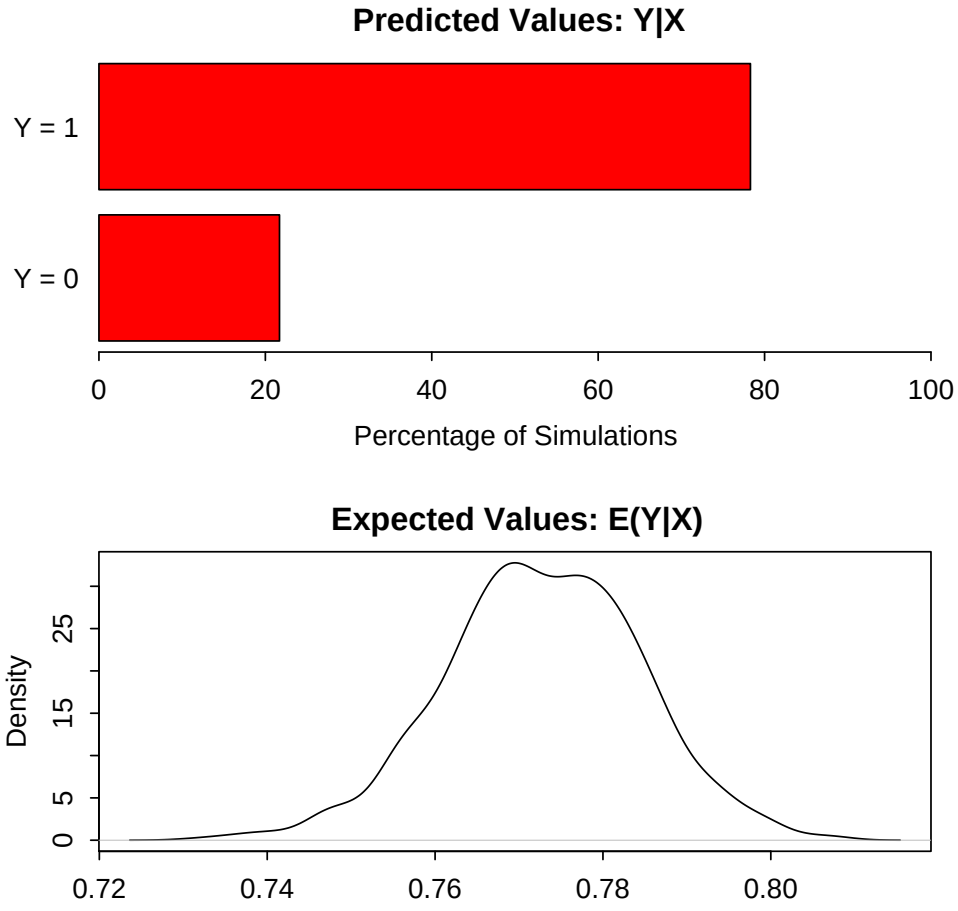
```
> x.out1 <- setx(z.out1)
```

Simulating quantities of interest from the posterior distribution.

```
> s.out1 <- sim(z.out1, x = x.out1)
```

```
> summary(s.out1)
```

```
> plot(s.out1)
```



2. Simulating First Differences

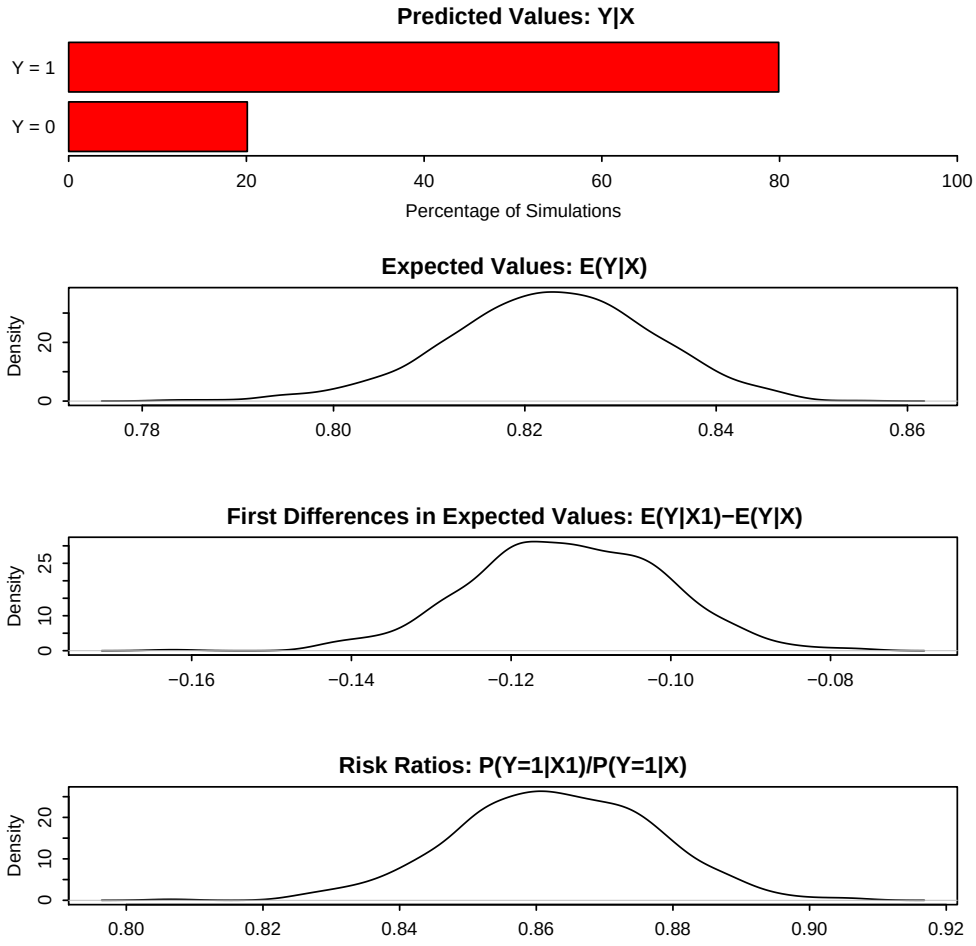
   Estimating the risk difference (and risk ratio) between low education (25th percentile) and high education (75th percentile) while all the other variables held at their default values.

```
> x.high <- setx(z.out1, educate = quantile(turnout$educate, prob = 0.75))
> x.low <- setx(z.out1, educate = quantile(turnout$educate, prob = 0.25))

> s.out2 <- sim(z.out1, x = x.high, x1 = x.low)

> summary(s.out2)

> plot(s.out2)
```



3. Example with Fixed Correlation Structure

   User-defined correlation structure

```
> corr.mat <- matrix(rep(0.5, 100), nrow = 10, ncol = 10)
> diag(corr.mat) <- 1
```

Generating empirical estimates:

```
> z.out2 <- zelig(vote ~ race + educate, model = "logit.gee", id = "cluster",
+     data = sorted.turnout, robust = TRUE, corstr = "fixed", R = corr.mat)
```

Viewing the regression output:

```
> summary(z.out2)
```

The Model

Suppose we have a panel dataset, with Y_{it} denoting the binary dependent variable for unit i at time t which takes the value of either 0 or 1. Y_i is a vector or cluster of correlated data where y_{it} is correlated with $y_{it'}$ for some or all t, t'. In the GEE model, we must specify a mean function, a variance function, and the structure of the correlation matrix of the within-cluster observations. Observations are assumed to be correlated within each cluster but not across clusters.

- The *stochastic component* is given by the joint distribution

$$Y_i \sim f(y_i \mid \pi_i)$$

where f is an unspecified multivariate distribution. The marginal distributions of f, which characterize each nonindependent unit-time observation Y_{it}, is given by

$$\begin{aligned} Y_{it} &\sim \text{Bernoulli}(y_{it} \mid \pi_{it}) \\ &= \pi_{it}^{y_{it}}(1-\pi_{it})^{1-y_{it}} \end{aligned}$$

where $\pi_{it} = \Pr(Y_{it} = 1)$ for $t = 1, ..., T$. The correlations within each unit i are modeled by defining the structure of a $T \times T$ "working" correlation matrix, which is specified by the user *a priori*. Note that the model assumes correlations within i but independence across i. The "working" correlation matrix then enters the variance term for each i, given by:

$$V_i = \phi\, A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$$

where A_i is a $T \times T$ diagonal matrix with $var(\pi_{it}) = \pi_{it}(1-\pi_{it})$ as the tth diagonal element, $R_i(\alpha)$ is the "working" correlation matrix, and ϕ is a scale parameter. The parameters are then estimated via a quasi-likelihood approach.

- The *systematic component* is given by:

$$\pi_{it} \;=\; \frac{1}{1+\exp(-x_{it}\beta)}.$$

  where x_{it} is the vector of k explanatory variables for unit i at time t and β is the vector of coefficients.

- GEE models require three specifications: a mean function (the systematic component above), a variance function (given by the variance of the Bernoulli stochastic component: $var(\pi_{it}) = \pi_{it}(1-\pi_{it})$), and a correlation structure (the "working" correlation matrix above). If the mean is correctly specified, but the variance and correlation structure are incorrectly specified, then GEE models provide consistent estimates of the parameters and thus the mean function as well, while consistent estimates of the standard errors can be obtained via a robust "sandwich" estimator. Similarly, if the mean and variance are correctly specified but the correlation structure is incorrectly specified, the parameters can be estimated consistently and the standard errors can be estimated consistently with the sandwich estimator. If all three are specified correctly, then the estimates of the parameters are more efficient.

- The robust "sandwich" estimator gives consistent estimates of the standard errors when the correlations are specified incorrectly only if the number of units i is relatively large and the number of repeated periods t is relatively small. Otherwise, one should use the "naïve" model-based standard errors, which assume that the specified correlations are close approximations to the true underlying correlations. See **?**) for more details.

Quantities of Interest

- All quantities of interest are for marginal means rather than joint means.

- The expected values (`qi$ev`) for the logit model are simulations of the predicted probability of a success:

$$E(Y) = \pi_c = \frac{1}{1+\exp(-x_c\beta)},$$

  given draws of β from its sampling distribution, where x_c is a vector of values, one for each independent variable, chosen by the user.

- The predicted values (`qi$pr`) are draws from the Binomial distribution with mean equal to the simulated expected value π_c.

- The first difference (`qi$fd`) for the logit model is defined as

$$\text{FD} = \Pr(Y=1 \mid x_1) - \Pr(Y=1 \mid x).$$

- The risk ratio (`qi$rr`) is defined as

$$\text{RR} = \Pr(Y=1 \mid x_1) \;/\; \Pr(Y=1 \mid x).$$

- In conditional prediction models, the average expected treatment effect (`att.ev`) for the treatment group is

$$\frac{1}{\sum_{i=1}^{n}\sum_{t=1}^{T} tr_{it}} \sum_{i:tr_{it}=1}^{n} \sum_{t:tr_{it}=1}^{T} \left\{ Y_{it}(tr_{it}=1) - E[Y_{it}(tr_{it}=0)] \right\},$$

  where tr_{it} is a binary explanatory variable defining the treatment ($tr_{it}=1$) and control ($tr_{it}=0$) groups. Variation in the simulations are due to uncertainty in simulating $E[Y_{it}(tr_{it}=0)]$, the counterfactual expected value of Y_{it} for observations in the treatment group, under the assumption that everything stays the same except that the treatment indicator is switched to $tr_{it}=0$.

- In conditional prediction models, the average predicted treatment effect (`att.pr`) for the treatment group is

$$\frac{1}{\sum_{i=1}^{n}\sum_{t=1}^{T} tr_{it}} \sum_{i:tr_{it}=1}^{n} \sum_{t:tr_{it}=1}^{T} \left\{ Y_{it}(tr_{it}=1) - Y_{it}\widehat{(tr_{it}=0)} \right\},$$

  where tr_{it} is a binary explanatory variable defining the treatment ($tr_{it}=1$) and control ($tr_{it}=0$) groups. Variation in the simulations are due to uncertainty in simulating $Y_{it}\widehat{(tr_{it}=0)}$, the counterfactual predicted value of Y_{it} for observations in the treatment group, under the assumption that everything stays the same except that the treatment indicator is switched to $tr_{it}=0$.

Output Values

The output of each Zelig command contains useful information which you may view. For example, if you run `z.out <- zelig(y ~ x, model = "logit.gee", id, data, corstr)`, then you may examine the available information in `z.out` by using `names(z.out)`, see the `coefficients` by using `z.out$coefficients`, and a default summary of information through `summary(z.out)`. Other elements available through the `$` operator are listed below.

- From the `zelig()` output object `z.out`, you may extract:
  - `coefficients`: parameter estimates for the explanatory variables.
  - `residuals`: the working residuals in the final iteration of the fit.
  - `fitted.values`: the vector of fitted values for the systemic component, π_{it}.
  - `linear.predictors`: the vector of $x_{it}\beta$
  - `max.id`: the size of the largest cluster.
- From `summary(z.out)`, you may extract:
  - `coefficients`: the parameter estimates with their associated standard errors, p-values, and z-statistics.

  - `working.correlation`: the "working" correlation matrix

- From the `sim()` output object `s.out`, you may extract quantities of interest arranged as matrices indexed by simulation × `x`-observation (for more than one `x`-observation). Available quantities are:

  - `qi$ev`: the simulated expected probabilities for the specified values of `x`.
  - `qi$pr`: the simulated predicted values for the specified values of `x`.
  - `qi$fd`: the simulated first difference in the expected probabilities for the values specified in `x` and `x1`.
  - `qi$rr`: the simulated risk ratio for the expected probabilities simulated from `x` and `x1`.
  - `qi$att.ev`: the simulated average expected treatment effect for the treated from conditional prediction models.
  - `qi$att.pr`: the simulated average predicted treatment effect for the treated from conditional prediction models.

How to Cite

To cite the *logit.gee* Zelig model:

> Patrick Lam. 2007. "logit.gee: Generalized Estimating Equation for Logistic Regression," in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software," `http://gking.harvard.edu/zelig`.

To cite Zelig as a whole, please reference these two sources:

> Kosuke Imai, Gary King, and Olivia Lau. 2007. "Zelig: Everyone's Statistical Software," `http://GKing.harvard.edu/zelig`.

> Kosuke Imai, Gary King, and Olivia Lau. 2007. "Toward A Common Framework for Statistical Analysis and Development," `http://gking.harvard.edu/files/abs/z-abs.shtml`.

See also

The logit function is part of the base packages by William N. Venable (Venables and Ripley 2002). Advanced users may wish to refer to `help(glm)` and `help(family)`, as well as McCullagh and Nelder (1989). Robust standard errors are implemented via sandwich package by Achim Zeileis (Zeileis 2004).Sample data are a selection of $2,000$ observations from King et al. (2000)

Bibliography

King, G., Tomz, M., and Wittenberg, J. (2000), "Making the Most of Statistical Analyses: Improving Interpretation and Presentation," *American Journal of Political Science*, 44, 341–355, http://gking.harvard.edu/files/abs/making-abs.shtml.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, no. 37 in Monograph on Statistics and Applied Probability, Chapman & Hall, 2nd ed.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, Springer-Verlag, 4th ed.

Zeileis, A. (2004), "Econometric Computing with HC and HAC Covariance Matrix Estimators," *Journal of Statistical Software*, 11, 1–17.