CSGLM

Introduction

CSGLM is a procedure for regression analysis as well as analysis of variance and covariance based on complex samples.

Complex sample data must contain both the values of the variables to be analyzed and the information on the current sampling design. Sampling design includes the sampling method, strata and clustering information, inclusion probabilities and the overall sampling weights.

Sampling design specification for CSGLM may include up to three stages of sampling. Any of the following general sampling methods may be assumed in the first stage: random sampling with replacement, random sampling without replacement and equal probabilities and random sampling without replacement and unequal probabilities. The first two sampling methods can also be specified for the second and the third sampling stage.

Notations

n	Total number of elements in the sample.
p	Number of regression parameters in the model.
$\mathbf{Y}$	Dependent variable vector containing values $y_i, i = 1, \ldots, n$.
$\mathbf{X}$	n x p design matrix. The rows correspond to the observations and the columns to the model parameters. The i^{th} row is $\mathbf{x}_i', i = 1, \ldots, n$.
$\mathbf{W}$	Diagonal matrix with sampling weights $w_i, i = 1, \ldots, n$ on the diagonal.
$\mathbf{B}$	Vector of p unknown population parameters.
N	Total number of elements in the population.

Weights

Overall weights specified for each ultimate element are processed as given. See "Complex Samples: Covariance Matrix of Total" (cs_covariance.pdf) for more information on weights and variance estimation methods.

Model Specification

Let the linear model be specified by the equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$ where $\mathbf{Y}$ is a vector of observed dependent variable values, $\mathbf{X}$ is the linear model design matrix, $\boldsymbol{\beta}$ is a vector of model parameters and $\mathbf{E}$ is a vector of random errors with zero mean. Each column of the design matrix corresponds to a parameter in the model equation. Each parameter corresponds to one of the intercept, factor main effects, factor interaction effects, factor nested effects, covariate effects and factors by covariates interaction effects. For every factor effect level occurring in data there is a separate parameter. This results in an over-parametrized model.

Estimation method

Assuming that the entire finite population has been observed, we can obtain the least square parameter estimates for the linear model by solving the following normal equations

$$\mathbf{X}'_N \mathbf{X}_N \boldsymbol{\beta} = \mathbf{X}'_N \mathbf{Y}_N$$

where $\mathbf{X}_N$ and $\mathbf{Y}_N$ denote design matrix and dependent variable for all elements in the given population. A solution vector for this system, estimating the model parameters $\boldsymbol{\beta}$, is denoted by $\mathbf{B}$. In our analyses we take the established design-based approach concerned with estimating the finite population parameters $\mathbf{B}$ developed by Kish and Frankel (1974), Fuller (1975), Shah, Holt and Folsom (1977) and others. See Särndal et al. (1992) for an overview.

Estimates for the population matrices $\mathbf{X}'_N \mathbf{X}_N$ and $\mathbf{X}'_N \mathbf{Y}_N$ are given by $\mathbf{X'WX}$ and $\mathbf{X'WY}$ respectively. We solve the following set of weighted normal equations

$$\mathbf{X'WXB} = \mathbf{X'WY}$$

where $\mathbf{W}$ is a diagonal matrix with sampling weights $w_i, i = 1 \ldots n$ on the diagonal. A solution for $\mathbf{B}$ is then given by the equation

$$\hat{\mathbf{B}} = (\mathbf{X'WX})^{-} \mathbf{X'WY}$$

where $(\mathbf{X'WX})^{-}$ is a generalized g2 inverse of $\mathbf{X'WX}$.

Predicted values and residuals

Predicted values for each observation are given by $\hat{y}_i = \mathbf{x}'_i \hat{\mathbf{B}}$, where $\mathbf{x}'_i$ is the i^{th} row of the design matrix $\mathbf{X}$. Vector of residual $\mathbf{r}$ is defined with $r_i = y_i - \hat{y}_i, i = 1, \ldots n$.

The residual sum of squares $\mathbf{r'Wr}$ is computed directly by the following:

$$\mathbf{r}'\mathbf{W}\mathbf{r} = \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i'\hat{\mathbf{B}})^2 .$$

Estimation algorithm

Estimation begins by construction of the weighted sum-of-squares and crossed products (SSCP) matrix. Let $\mathbf{z}_i' = (\mathbf{x}_i', y_i)$ be the i^{th} row of matrix $\mathbf{Z}$, where $\mathbf{x}_i'$ is the i^{th} row of design matrix $\mathbf{X}$, and y_i is the corresponding dependent variable value. Then the SSCP matrix is computed by

$$\mathbf{Z}'\mathbf{W}\mathbf{Z} = \sum_{i=1}^{n} w_i \mathbf{z}_i \mathbf{z}_i'$$

where $\mathbf{z}_i \mathbf{z}_i'$ is the outer product for the vector $\mathbf{z}_i$.

This matrix can be partitioned as follows

$$\mathbf{Z}'\mathbf{W}\mathbf{Z} = \begin{pmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Y} \\ \mathbf{Y}'\mathbf{W}\mathbf{X} & \mathbf{Y}'\mathbf{W}\mathbf{Y} \end{pmatrix}.$$

After applying sweep operator to the first p rows and columns of the matrix above, we obtain the following solution matrix

$$\begin{pmatrix} -(\mathbf{X}'\mathbf{W}\mathbf{X})^- & \hat{\mathbf{B}} \\ \hat{\mathbf{B}}' & \mathbf{r}'\mathbf{W}\mathbf{r} \end{pmatrix}.$$

$(\mathbf{X}'\mathbf{W}\mathbf{X})^-$ is a generalized g2 inverse of $\mathbf{X}'\mathbf{W}\mathbf{X}$, $\hat{\mathbf{B}}$ is a parameter solution, and $\mathbf{r}'\mathbf{W}\mathbf{r}$ is the residual sum of squares.

When a column of $\mathbf{X}'\mathbf{W}\mathbf{X}$ is found to be dependent on previous columns, the corresponding parameter is treated as redundant. Solution for redundant parameters is set to 0 as well as corresponding rows and columns in $(\mathbf{X}'\mathbf{W}\mathbf{X})^-$.

Variance estimates

Variances of parameter estimates are computed according to the Taylor linearization method as presented by Binder (1983).

Define vector $\mathbf{d}_i = \mathbf{x}_i (y_i - \mathbf{x}_i'\hat{\mathbf{B}})$ for $i = 1, \ldots, n$ and its total population estimate by

$$\hat{\mathbf{d}}_T = \sum_{i=1}^{n} w_i \mathbf{x}_i (y_i - \mathbf{x}_i'\hat{\mathbf{B}}).$$

Let $\hat{\mathbf{V}}(\hat{\mathbf{d}}_T)$ be its sample design-based covariance matrix computed by the methods described in "Complex Samples: Covariance Matrix of Total" (cs_covariance.pdf). Then the covariance matrix of $\hat{\mathbf{B}}$ is estimated by

$$\hat{\mathbf{V}}(\hat{\mathbf{B}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\hat{\mathbf{V}}(\hat{\mathbf{d}}_T)(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}.$$

Note: If any diagonal element of $\hat{\mathbf{V}}(\hat{\mathbf{d}}_T)$ happens to be non-positive due to the use of Yates-Grundy-Sen estimator, all elements in the corresponding row and column are set to zero.

Subpopulation estimates

When analyses are requested for a given subpopulation S, we redefine $(\mathbf{x}_i', y_i)'$ as follows:

$$(\mathbf{x}_i', y_i) = \begin{cases} (\mathbf{x}_i', y_i) & \text{if the } i^{th} \text{ element is in } S \\ (0,\ldots,0) & \text{otherwise} \end{cases}$$

When computing point estimates, this substitution is equivalent to including only the subpopulation elements in the calculations. This is in contrast to computing the variance estimates where all elements in the sample need to be included.

Standard Errors

Let $\hat{B}_i$ denote a non-redundant parameter estimate. Its standard error is the square root of its estimated variance:

$$SE(\hat{B}_i) = \sqrt{\hat{V}(\hat{B}_i)}.$$

Standard error is undefined for redundant parameters.

Degrees of freedom

Number of the degrees of freedom ν used for computing confidence intervals and test statistics below is calculated as the difference between the number of primary sampling units and the number of strata in the first stage of sampling. We shall also refer to this quantity as the sample design degrees of freedom. Alternatively, ν may be specified by the user.

Confidence Intervals

A level $1-\alpha$ confidence interval is constructed for a given $0 \le \alpha \le 1$ for each non-redundant model parameter $\hat{B}_i$. Confidence bounds are given by

$$\hat{B}_i \pm SE(\hat{B}_i)t_\nu(1-\alpha/2)$$

where $SE(\hat{B}_i)$ is the estimated standard error of $\hat{B}_i$, and $t_\nu(1-\alpha/2)$ is the $100(1-\alpha/2)$ percentile of t distribution with ν degrees of freedom.

t Tests

Testing hypothesis $H_{0i}: \hat{B}_i = 0$ for each non-redundant model parameter $\hat{B}_i$ is performed using the t test statistic:

$$t(\hat{B}_i) = \frac{\hat{B}_i}{SE(\hat{B}_i)}.$$

The *p*-value for the two-sided test is given by the probability $P(|T| > |t(\hat{B}_i)|)$, where T is a random variable from the t distribution with ν degrees of freedom.

Design Effects

Design effect $Deff(\hat{B}_i)$ for non-redundant parameter estimate $\hat{B}_i$ is given by

$$Deff(\hat{B}_i) = \frac{\hat{V}(\hat{B}_i)}{\hat{V}_{srs}(\hat{B}_i)}$$

Design effect is undefined for redundant parameters.

$\hat{V}(\hat{B}_i)$ is the estimate of variance of $\hat{B}_i$ under the appropriate sampling design, while $\hat{V}_{srs}(\hat{B}_i)$ is the estimate of variance of $\hat{B}_i$ under the simple random sampling assumption. The latter is computed as the i^{th} diagonal element of the following matrix:

$$\hat{V}_{srs}(\hat{B}_i) = [(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}\hat{\mathbf{V}}_{srs}(\hat{\mathbf{d}}_T)(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}]_{ii}$$

where

$$\hat{\mathbf{V}}_{srs}(\hat{\mathbf{d}}_T) = (1-\frac{n}{\hat{N}})\frac{\hat{N}}{n-1}\sum_{i=1}^{n} w_i \mathbf{d}_i \mathbf{d}_i'$$

with $\mathbf{d}_i$ as specified earlier.

Design effects are undefined when $\frac{n}{\hat{N}} \geq 1$.

For subpopulation analysis we have that $\mathbf{d}_i = \mathbf{0}$ whenever observation i does not belong to a given subpopulation.

We also provide the square root of design effect $Defft$ by computing

$$Defft = \sqrt{Deff} \ .$$

Design effects and their application have been discussed by Kish (1965) and Kish (1995).

Multiple R-square

Multiple R-square is computed by the following formula

$$R^2 = 1 - \frac{\mathbf{r}'\mathbf{W}\mathbf{r}}{(\mathbf{Y} - \hat{\bar{Y}}_S\mathbf{1})'\mathbf{W}(\mathbf{Y} - \hat{\bar{Y}}_S\mathbf{1})}$$

where $\hat{\bar{Y}}_S = \hat{Y}_S / \hat{N}_S$ is the estimated subpopulation mean for variable Y.

If the specified model contains no intercept we use the following expression

$$R^2 = 1 - \frac{\mathbf{r}'\mathbf{W}\mathbf{r}}{\mathbf{Y}'\mathbf{W}\mathbf{Y}}.$$

Hypothesis Testing

Given matrix $\mathbf{L}$ with r rows and p columns, and vector $\mathbf{K}$ with r elements, CSGLM performs testing of linear hypothesis $H_0 : \mathbf{LB} = \mathbf{K}$. It is necessary that $\mathbf{LB}$ is estimable. Wald X^2 statistic is given by

$$\mathrm{X}^2 = (\mathbf{L}\hat{\mathbf{B}} - \mathbf{K})'(\mathbf{L}\hat{\mathbf{V}}(\hat{\mathbf{B}})\mathbf{L}')^{-}(\mathbf{L}\hat{\mathbf{B}} - \mathbf{K}).$$

Asymptotic distribution of the X^2 test statistic is chi-square with r_I degrees of freedom, where $r_I = rank(\mathbf{L}\hat{\mathbf{V}}(\hat{\mathbf{B}})\mathbf{L}')$. If $r_I < r$, $(\mathbf{L}\hat{\mathbf{V}}(\hat{\mathbf{B}})\mathbf{L}')^{-}$ is a generalized inverse such that Wald tests are effective for restricted set of hypothesis $\mathbf{L}_I\mathbf{B} = \mathbf{K}_I$ containing a particular subset I of independent rows from H_0.

Each row l_i' of matrix $\mathbf{L}$ is also tested separately. Estimate for the i^{th} row is given by $l_i'\hat{\mathbf{B}}$ and its standard error by $\sqrt{l_i'\hat{\mathbf{V}}(\hat{\mathbf{B}})l_i}$.

See "Complex Samples: Model Testing" (cs_modeltesting.pdf) for additional tests and p-value adjustments.

Custom tests

Custom hypothesis tests are conducted only when $\mathbf{L}$ is such that $\mathbf{LB}$ is estimable. This condition is verified using the following equality:

$$\mathbf{L} = \mathbf{L}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-}(\mathbf{X}'\mathbf{W}\mathbf{X}).$$

Default tests of model effects

For each effect specified in the model, Type III test $\mathbf{L}$ matrix is constructed such that $\mathbf{LB}$ is estimable. It involves parameters only for the given effect and the containing effects and it does not depend on the order of effects specified in the model. If such a matrix cannot be constructed, the effect is not testable. Matrix $\mathbf{K}$ is always set to $\mathbf{0}$ when computing the test statistics for model effects.

Hypothesis for the corrected model is that all the parameters except for the intercept are zero.

Estimated marginal means

Estimated marginal means (EMMEANS) are based on the estimated cell means. For a given fixed set of factors, or their interactions, we estimate marginal means as the mean value averaged over all cells generated by the rest of the factors in the model. Covariates may be fixed at any specified value. If not specified, the value for each covariate is set to its overall mean estimate.

When missing cells are present in the data, EMMEANS may not be estimable. In such circumstance, we provide a modified estimate proposed by Searle, Speed and Milliken (1980) that ignores the non-estimable cells.

Each marginal estimate is finally constructed in the form $l'\hat{\mathbf{B}}$ such that $l'\mathbf{B}$ is estimable.

Comparing EMMEANS

For a given factor in the model, a vector of EMMEANS is created for all levels of the factor. This vector can be expressed in the form $\hat{\boldsymbol{\mu}} = \mathbf{L}\hat{\mathbf{B}}$ where each row of $\mathbf{L}$ matrix is generated as described above. Variance is then computed by the following formula:

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\mu}}) = \mathbf{L}\hat{\mathbf{V}}(\hat{\mathbf{B}})\mathbf{L}'.$$

A set of contrasts for the factor is created according to the selected contrast type. Let this set of contrasts define the matrix $\mathbf{C}$ used for testing the following hypothesis $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$. The Wald X^2 statistic is used for testing given set of contrasts for the factor as follows:

$$\mathrm{X}^2 = (\mathbf{C}\hat{\boldsymbol{\mu}})'(\mathbf{C}\hat{\mathbf{V}}(\hat{\boldsymbol{\mu}})\mathbf{C}')^{-}(\mathbf{C}\hat{\boldsymbol{\mu}})$$

The asymptotic distribution of the X^2 test statistic is chi-square with r_I degrees of freedom, where $r_I = rank(\mathbf{C}\hat{\mathbf{V}}(\hat{\boldsymbol{\mu}})\mathbf{C}')$.

Each row c_i' of matrix $\mathbf{C}$ is also tested separately. The estimate for the i^{th} row is given by $c_i'\hat{\boldsymbol{\mu}}$ and its standard error by $\sqrt{c_i'\hat{\mathbf{V}}(\hat{\boldsymbol{\mu}})c_i}$.

See "Complex Samples: Model Testing" (cs_modeltesting.pdf) for additional tests and p-value adjustments. Substitute the following formula for the simple random sampling covariance: $\hat{\mathbf{V}}_{srs}(\hat{\boldsymbol{\mu}}) = \mathbf{L}\hat{\mathbf{V}}_{srs}(\hat{\mathbf{B}})\mathbf{L}'$.

References

Binder, D. A. (1983), "On the variances of asymptotically normal estimators from complex surveys", International Statistical Review, 51, 279-292.

Fuller, W. A. (1975), "Regression analysis for sample survey", Sankhya, Series C 37, 117-132.

Kish, L. (1965), Survey Sampling, New York: John Wiley & Sons.

Kish, L. (1995), "Methods for Design Effects", Journal of Official Statistics, volume 11, pages 119 - 127.

Kish, L. and Frankel, M. R. (1974), "Inference from complex samples", Journal of the Royal Statistical Society B, 36, 1-37.

Särndal, C. E., Swenson, B., and Wretman, J. H. (1992), Model Assisted Survey Sampling, New York: Springer-Verlag.

Searle, S. R., Speed, F. M., and Milliken, G. A. (1980), "Population marginal means in the linear model: an alternative to least square means", The American Statistician, volume 34, pages 216 - 221.

Shah, B. V., Holt, M. M., and Folsom, R. E. (1977), "Inference about regression models from sample survey data", Bulletin of the International Statistical Institute XLVII, 3, 43-57.