GLM
Univariate and Multivariate

GLM (general linear model) is a general procedure for analysis of variance and covariance, as well as regression. It can be used for both univariate and multivariate designs. Repeated measures analysis is also available. Algorithms that apply only to repeated measures are in the chapter *GLM Repeated Measures*.

For information on post hoc tests, see Appendix 10. For sums of squares, see Appendix 11. For distribution functions, see Appendix 12. For Box's *M* test, see Appendix 14.

Notation

The following notation is used throughout this chapter. Unless otherwise stated, all vectors are column vectors and all quantities are known.

n	Number of cases.
N	Effective sample size.
p	Number of parameters (including the constant, if it exists) in the model.
r	Number of dependent variables in the model.
$\mathbf{Y}$	$n \times r$ matrix of dependent variables. The rows are the cases and the columns are the dependent variables. The *i*th row is $\mathbf{y}_i'$, $i = 1, \ldots, n$.
$\mathbf{X}$	$n \times p$ design matrix. The rows are the cases and the columns are the parameters. The *i*th row is $\mathbf{x}_i'$, $i = 1, \ldots, n$.
r_X	Number of nonredundant columns in the design matrix. Also the rank of the design matrix.
w_i	Regression weight of the *i*th case.
f_i	Frequency weight of the *i*th case.
$\mathbf{B}$	$p \times r$ *unknown* parameter matrix. The columns are the dependent variables. The *j*th column is $\mathbf{b}_j$, $j = 1, \ldots, r$.
Σ	$r \times r$ *unknown* common multiplier of the covariance matrix of any row of $\mathbf{Y}$. The *(i, j)*th element is σ_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, r$.

Model

The model is $\mathbf{Y} = \mathbf{XB}$ and $\mathbf{y}_i'$ is independently distributed as a *p*-dimensional normal distribution with mean $\mathbf{x}_i'\mathbf{B}$ and covariance matrix $w_i^{-1}\mathbf{\Sigma}$. The *i*th case is ignored if $w_i \le 0$.

Frequency Weight and Total Sample Size

The frequency weight f_i is the number of replications represented by an SPSS case; therefore, the weight must be a non-negative integer. It is computed by rounding the value in the SPSS weight variable to the nearest integer. The total sample size is $N = \sum_{i=1}^{n} f_i \, \mathrm{I}(w_i > 0)$, where $\mathrm{I}(w_i > 0) = 1$ if $w_i > 0$ and is equal to 0 otherwise.

The Cross-Product and Sums-of-Squares Matrices

To prepare for the SWEEP operation, an augmented row vector of length $(p+r)$ is formed:

$$\mathbf{z}_i' = (\mathbf{x}_i', \mathbf{y}_i')$$

Then the $(p+r)\times(p+r)$ matrix is computed:

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

This matrix is partitioned as

$$\mathbf{Z'WZ} = \begin{pmatrix} \mathbf{X'WX} & \mathbf{X'WY} \\ \mathbf{Y'WX} & \mathbf{Y'WY} \end{pmatrix}$$

The upper left $p \times p$ submatrix is $\mathbf{X'WX}$ and the lower right $r \times r$ submatrix is $\mathbf{Y'WY}$.

Sweep Operation

Three important matrices, **G**, $\hat{\mathbf{B}}$, and **S**, are obtained by sweeping the **Z′WZ** matrix as follows:

1. Sweep sequentially the first *p* rows and the first *p* columns of **Z′WZ**, starting from the first row and the first column.
2. After the *p*th row and the *p*th column are swept, the resulting matrix is

$$\begin{pmatrix} -\mathbf{G} & \hat{\mathbf{B}} \\ \hat{\mathbf{B}}' & \mathbf{S} \end{pmatrix}$$

where **G** is a $p \times p$ symmetric g_2 generalized inverse of **X′WX**, $\hat{\mathbf{B}}$ is the $p \times r$ matrix of parameter estimates and **S** is the $r \times r$ symmetric matrix of sums of squares and cross products of residuals.

The SWEEP routine is adapted from Algorithm AS 178 by Clarke (1982) and Remarks R78 by Ridout and Cobby (1989).

Residual Covariance Matrix

The estimated $r \times r$ covariance matrix is $\hat{\Sigma} = \mathbf{S} / (N - r_{\mathbf{X}})$ provided $r_X < N$. If $r_X = N$, then $\hat{\Sigma} = 0$. If $r_X > N$, then all elements of $\hat{\Sigma}$ are system missing.

The residual degrees of freedom is $N - r_X$. If $r_X > N$, then the degrees of freedom is system missing.

Parameter Estimates

Let the elements of $\hat{\Sigma}$ be $\hat{\sigma}_{ij}$, the elements of **G**, g_{ij}, and the elements of $\hat{\mathbf{B}}$, $\hat{b}_{ij}$. Then $\text{var}(\hat{b}_{ij})$ is estimated by $\hat{\sigma}_{jj} g_{ii}$ for $i = 1, \ldots, p;\ j = 1, \ldots, r$ and $\text{cov}(\hat{b}_{ij}, \hat{b}_{rs})$ is estimated by $\hat{\sigma}_{js} g_{ir}$ for $i, r = 1, \ldots, p;\ j, s = 1, \ldots, r$.

Standard Error of $\hat{b}_{ij}$

$$\text{se}\left(\hat{b}_{ij}\right) = \sqrt{\hat{\sigma}_{jj} g_{ii}}$$

When the *i*th parameter is redundant, the standard error is system missing.

The *t* Statistic

For testing $H_0: b_{ij} = 0$ versus $H_1: b_{ij} \neq 0$, the *t* statistic is

$$t = \begin{cases} \hat{b}_{ij} / \text{se}(\hat{b}_{ij}) & \text{if the standard error is positive} \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

The significance value for this statistic is $2\left(1 - \text{CDF.T}\left(|t|, N - r_x\right)\right)$ where CDF.T is the SPSS function for the cumulative *t* distribution.

Partial Eta Squared Statistic

$$\eta^2 = \begin{cases} \hat{b}_{ij}^2 \Big/ (\hat{b}_{ij}^2 + (N - r_X)\,\text{var}(\hat{b}_{ij})) & \text{if } r_X < N \text{ and the denominator is positive} \\ 1 & \text{if } r_X = N \text{ but } b_{ij} \neq 0 \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

The value should be within $0 \leq \eta^2 \leq 1$.

Noncentrality Parameter

$$c = |t|$$

Observed Power

$$p = \begin{cases} 1 - \text{NCDF.T}(t_c, N - r_X, c) + \text{NCDF.T}(-t_c, N - r_X, c) & r_x < N \\ \text{SYSMIS} & r_x \geq N, \\ & \text{or any arguments to NCDF.T} \\ & \text{or IDF.T are SYSMIS} \end{cases}$$

where $t_c = \text{IDF.T}\left(1 - \alpha/2, N - r_X\right)$ and α is the user-specified chance of Type I error $(0 < \alpha < 1)$. NCDF.T and IDF.T are the SPSS functions for the cumulative noncentral *t* distribution and for the inverse cumulative *t* distribution, respectively.

The default value is $\alpha = 0.05$. The observed power should be within $0 \leq p \leq 1$.

Confidence Interval

For the *p*% level, the individual univariate confidence interval for the parameter is

$$\hat{b}_{ij} \pm t_{\alpha} \operatorname{se}\left(\hat{b}_{ij}\right)$$

where $t_{\alpha} = \text{IDF.T}\left(0.5(1+p/100), N - r_X\right)$ for $i = 1,\ldots,n; j = 1,\ldots,r$. The default value of *p* is 95 ($0 < p < 100$).

Correlation

$$\text{corr}(\hat{b}_{ij}, \hat{b}_{rs}) = \begin{cases} \hat{\sigma}_{js}\, g_{ir} \Big/ \left(\operatorname{se}\left(\hat{b}_{ij}\right) \times \operatorname{se}\left(\hat{b}_{rs}\right)\right) & \text{if the standard errors are positive} \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

for $i, r = 1,\ldots, p;\ j, s = 1,\ldots, r$.

Estimated Marginal Means

Estimated marginal means (EMMEANS) are computed as the generic $\mathbf{l}'\hat{\mathbf{B}}\mathbf{m}$ expression with appropriate **l** and **m** vectors. **l** is a column vector of length *p* and **m** is a column vector of length *r*. Since the **l** vector is chosen to be always estimable, the quantity $\mathbf{l}'\hat{\mathbf{B}}\mathbf{m}$ is in fact the estimated modified marginal means (Searle, Speed, and Milliken, 1980). When covariates (or products of covariates) are present in the effects, the overall means of the covariates (or products of covariates) are used in the **l** matrix. Suppose X and Y are covariates and they appear as X*Y in an effect; then the mean of X*Y is used instead of the product of the mean of X and the mean of Y.

L Matrix

For each level combination of the between subjects factors in TABLES, identify the nonmissing cases with positive caseweights and positive regression weights which are associated with the current level combination. Suppose the cases are classified by three between-subjects factors: A, B and C. Now A and B are specified in TABLES and the current level combination is A=1 and B=2. A case in the cell A=1, B=2, and C=3 is associated with the current level combination,

whereas a case in the cell A=1, B=3 and C=3 is not. Compute the average of the design matrix rows corresponding to these cases.

If an effect contains a covariate, then its parameters which belong to the current level combination are equal to the mean of the covariate, and are equal to 0 otherwise. Using the above example, for effect A*X where X is a covariate, the parameter [A=1]*X belongs to the current level combination where the parameter [A=2]*X does not. If the effect contains a product of covariates, then the mean of the product is applied.

The result is the **l** vector for the current between-subjects factor level combination. When none of the between-subjects effects contain covariates, the vector always forms an estimable function. Otherwise, a non-estimable function may occur, depending on the data.

M Matrix

The **M** matrix is formed as a series of Kronecker products

$$\mathbf{M} = \mathbf{I}_c \otimes \mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_t$$

where

$$\mathbf{A}_k = \begin{cases} \mathbf{I}_{r_k} & \text{if the } k\text{th within subjects factor is specified in TABLES} \\ (1/r_k)\mathbf{1}_{r_k} & \text{otherwise} \end{cases}$$

with $\mathbf{1}_{r_k}$ a column vector of length r_k and all of its elements equal to 1.

If OVERALL or only between-subjects factors are specified in TABLES, then $\mathbf{A}_k = (1/r_k)\mathbf{1}_{r_k}$ for $k = 1, \ldots, t$.

The column for a particular within-subjects factor level combination, denoted by **m**, is extracted accordingly from this **M** matrix.

Standard Error

$$\text{se}\left(\mathbf{l}'\hat{\mathbf{B}}\mathbf{m}\right) = \begin{cases} \sqrt{(\mathbf{l}'\mathbf{G}\mathbf{l})\left(\mathbf{m}'\hat{\boldsymbol{\Sigma}}\mathbf{m}\right)} & \text{if } N - r_{\mathbf{X}} > 0 \\ \text{SYSMIS} & \text{otherwise} \end{cases} \qquad \textbf{(1)}$$

Since **l** are coefficients of an estimable function, the standard error is the same for any generalized inverse **G**.

Significance

The *t* statistic is

$$t = \begin{cases} \mathbf{l}'\hat{\mathbf{B}}\mathbf{m} \Big/ \text{se}\left(\mathbf{l}'\hat{\mathbf{B}}\mathbf{m}\right) & \text{if } \text{se}\left(\mathbf{l}'\hat{\mathbf{B}}\mathbf{m}\right) > 0 \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

If the *t* statistic is not system missing, then the significance is computed based on a *t* distribution with $N - r_{\mathbf{X}}$ degrees of freedom.

Pairwise Comparison

Between-Subjects Factor

Suppose the **l** vectors are indexed by the level of the between-subjects factor as $\mathbf{l}_{i_1,\ldots,i_b}$, $i_s = 1,\ldots,n_s$ and $s = 1,\ldots,b$ where n_s is the number of levels of between-subjects factor *s* and *b* is the number of between-subjects factors specified inside TABLES. The difference in estimated marginal means of level i_s and level i_s' of between-subjects factor *s* at fixed levels of other between-subjects factors is

$$\left(\mathbf{l}_{i_1,\ldots,i_{s-1},i_s,i_{s+1},\ldots,i_b} - \mathbf{l}_{i_1,\ldots,i_{s-1},i_s',i_{s+1},\ldots,i_b}\right)' \hat{\mathbf{B}}\mathbf{m} \text{ for } i_s, i_s' = 1,\ldots,n_s; i_s \neq i_s'.$$

The standard error of the difference is computed by substituting for **l** in **(1)**: $\mathbf{l}_{i_1,\ldots,i_{s-1},i_s,i_{s+1},\ldots,i_b} - \mathbf{l}_{i_1,\ldots,i_{s-1},i_s',i_{s+1},\ldots,i_b}$.

Within-Subjects Factor

Suppose the **m** vectors are indexed by level of the within-subjects factor as $\mathbf{m}_{j_1,\ldots,j_w}$, $j_s = 1,\ldots,n_s$ and $s = 1,\ldots,w$, where n_s is the number of levels of within-subjects factor *s* and *w* is the number of within-subjects factors specified inside TABLES. The difference in estimated marginal means of level j_s and level j_s' of within-subjects factor *s* at fixed levels of other within-subjects factors is

$$\mathbf{l}'\mathbf{B}\left(\mathbf{m}_{j_1,\ldots,j_{s-1},j_s,j_{s+1},\ldots,j_b} - \mathbf{m}_{j_1,\ldots,j_{s-1},j_s',j_{s+1},\ldots,j_b}\right) \text{ for } j_s, j_s' = 1,\ldots,n_s; j_s \neq j_s'.$$

The standard error of the difference is computed by substituting for **m** in **(1)** $\mathbf{m}_{i_1,\ldots,i_{s-1},i_s,i_{s+1},\ldots,i_b} - \mathbf{m}_{i_1,\ldots,i_{s-1},i_s',i_{s+1},\ldots,i_b}$.

Confidence Interval

The $(1-\alpha)\times 100\%$ confidence interval is

$$\mathbf{l}'\hat{\mathbf{B}}\mathbf{m} \pm t_{1-\alpha/2;N-r_{\mathbf{X}}} \times \text{se}\left(\mathbf{l}'\hat{\mathbf{B}}\mathbf{m}\right)$$

and $t_{1-\alpha/2;N-r_{\mathbf{X}}}$ is the $(1-\alpha/2)\times 100\%$ percentile of a t distribution with $N-r_{\mathbf{X}}$ degrees of freedom. No confidence interval is computed if $N-r_{\mathbf{X}} \le 0$.

Saved Values

Temporary variables can be added to the working data file. These include predicted values, residuals, and diagnostics.

Predicted Values

The $n\times r$ matrix of predicted values is $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$. The ith row of $\hat{\mathbf{Y}}$ is $\hat{\mathbf{y}}_i' = \mathbf{x}_i'\hat{\mathbf{B}}$, $i = 1,\ldots,n$. Let the elements of $\hat{\mathbf{Y}}$ be $\hat{y}_{ij}$ and the elements of $\mathbf{XGX}'$ be π_{ij}.

The standard error of $\hat{y}_{ij}$ is

$$\text{se}\left(\hat{y}_{ij}\right) = \sqrt{\hat{\sigma}_{jj}\,\pi_{ii}} \qquad \textbf{for } i = 1,\ldots,n;\ j = 1,\ldots,r$$

The weighted predicted value of the ith case is $\sqrt{w_i}\,\hat{\mathbf{y}}_i'$.

Residuals

The $n\times r$ matrix of residuals is $\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}$.

The ith row of $\hat{\mathbf{E}}$ is $\hat{\mathbf{e}}_i' = \mathbf{y}_i' - \hat{\mathbf{y}}_i'$, $i = 1,\ldots,n$.

Let the elements of $\hat{\mathbf{E}}$ be $\hat{e}_{ij}$; then

$$\hat{e}_{ij} = y_{ij} - \hat{y}_{ij}\textbf{, for } i = 1,\ldots,n;\ j = 1,\ldots,r$$

The weighted residual is $\sqrt{w_i}\,\hat{\mathbf{e}}_i'$.

Deleted Residuals (PRESS Residuals)

The deleted residual is the predicted residual for the *i*th case that results from omitting the *i*th case from estimation. It is:

$$\text{DRESID}_{ij} = \begin{cases} \hat{e}_{ij} \big/ (1/w_i - \pi_{ii}) & \text{if } w_i > 0 \text{ and } w_i \pi_{ii} < 1; \\ \text{SYSMIS} & \text{otherwise.} \end{cases}$$

for $i = 1, \ldots, n; j = 1, \ldots, r$.

Standardized Residuals

The standardized residual is the residual divided by the standard deviation of data:

$$\text{ZRESID}_{\text{ij}} = \begin{cases} (y_{ij} - \hat{y}_{ij}) \Big/ \left(\sqrt{\hat{\sigma}_{jj} / w_i}\right) & \text{if } w_i > 0 \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

Studentized Residuals

The standard error for $\hat{e}_{ij}$ is

$$\text{se}(\hat{e}_{ij}) = \begin{cases} \sqrt{\hat{\sigma}_{jj}(1/w_i - \pi_{ii})} & \text{if } w_i > 0 \text{ and } w_i \pi_{ii} < 1; \\ \text{SYSMIS} & \text{otherwise.} \end{cases}$$

for $i = 1, \ldots, n; j = 1, \ldots, r$. The Studentized residual is the residual divided by the standard error of the residual.

$$\text{SRESID}_{\text{ij}} = \begin{cases} \hat{e}_{ij} / \text{se}(\hat{e}_{ij}) & \text{if } w_i > 0 \text{ and } se(\hat{e}_{ij}) > 0 \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

Diagnostics

Cook's Distance

Cook's Distance *D* measures the change to the solution that results from omitting each observation. The formula is

$$D_{ij} = \left(\frac{\hat{e}_{ij}}{\sqrt{\hat{\sigma}_{jj}(1/w_i - \pi_{ii})}} \right)^2 \left(\frac{\pi_{ii}}{(1/w_i - \pi_{ii})} \right) \frac{1}{r_X}$$

for $i = 1,\ldots,n; j = 1,\ldots,r$. This formula is equivalent to

$$D_{ij} = \left(\hat{e}_{ij} / \text{se}(\hat{e}_{ij})\right)^2 \left(\text{se}\left(\hat{y}_{ij}\right) / \text{se}\left(\hat{e}_{ij}\right)\right) / r_X \quad \textbf{provided } w_i > 0 \text{ and } \text{se}\left(\hat{e}_{ij}\right) > 0.$$

When $w_i \le 0$ or $\text{se}\left(\hat{e}_{ij}\right) = 0$, D_{ij} is system missing.

Leverage (Uncentered)

The leverage for the *i*th case ($i = 1,\ldots,n$) for all dependent variables is

$$\text{LEVER}_i = \begin{cases} w_i \pi_{ii} & \text{if } w_i > 0 \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

Hypothesis Testing

Let **L** be an $l \times p$ known matrix, **M** be an $r \times m$ known matrix and **K** be an $l \times m$ known matrix. The test hypotheses $\text{H}_0: \mathbf{LBM} = \mathbf{K}$ versus $\text{H}_1: \mathbf{LBM} \neq \mathbf{K}$ are *testable* if and only if **LB** is estimable.

The following results apply to testable hypotheses only. Nontestable hypotheses are excluded.

The hypothesis SSCP matrix is $\mathbf{S}_H = (\mathbf{L}\hat{\mathbf{B}}\mathbf{M} - \mathbf{K})'(\mathbf{LGL}')^{-1}(\mathbf{L}\hat{\mathbf{B}}\mathbf{M} - \mathbf{K})$ and the error SSCP matrix is $\mathbf{S}_E = \mathbf{M}'\mathbf{SM}$.

Four test statistics, based on the eigenvalues of $\mathbf{S}_E^{-1}\mathbf{S}_H$, are available: Wilks' lambda, Hotelling-Lawley trace, Pillai's trace, and Roy's largest root.

Let the eigenvalues of $\mathbf{S}_E^{-1}\mathbf{S}_H$ be $\lambda_1 \geq \ldots \geq \lambda_{r_E} \geq 0$ and $\lambda_{r_E+1},\ldots,\lambda_m = 0$, and let $r_E = \text{rank}(\mathbf{S}_E)$; $s = \min(l, r_E)$; $n_e = n - r_{\mathbf{x}}$; $m^* = \frac{1}{2}(|r_E - l| - 1)$; $n^* = \frac{1}{2}(n_e - r_E - 1)$.

Wilks' Lambda

$$\Lambda = \frac{\det(\mathbf{S}_E)}{\det(\mathbf{S}_H + \mathbf{S}_E)} = \prod_{k=1}^{m} \frac{1}{(1+\lambda_k)}.$$

When H_o is true, the F statistic

$$F = \frac{(\varsigma\tau - 2\upsilon)}{lr_E}\frac{\left(1-\Lambda^{1/\tau}\right)}{\Lambda^{1/\tau}}$$

follows asymptotically an F distribution, where

$$\varsigma = n_e - \tfrac{1}{2}(r_E - l + 1)$$
$$\upsilon = \tfrac{1}{4}(lr_E - 2)$$
$$\tau = \begin{cases} \sqrt{\left(l^2 r_E^2 - 4\right)/\left(l^2 + r_E^2 - 5\right)} & \text{if } \left(l^2 + r_E^2 - 5\right) > 0 \\ 1 & \text{otherwise} \end{cases}$$

The degrees of freedom are $(lr_E, \varsigma\tau - 2\upsilon)$. The F statistic is exact if $s = 1, 2$. See Rao (1951) and Section 8c.5 of Rao (1973) for details.

The eta-squared statistic is $\eta^2 = 1 - \Lambda^{1/s}$.

The noncentrality parameter is $\lambda = (\xi\tau - 2\upsilon)\eta^2 / \left(1 - \eta^2\right)$.

The power is $1 - \text{NCDF.F}\left(\text{F}_\alpha, lr_E, (\xi\tau - 2\upsilon), \lambda\right)$ where F_α is the upper 100α percentage point of the central F distribution, and α is user-specified on the ALPHA keyword on the CRITERIA subcommand.

Hotelling-Lawley Trace

In the SPSS software, the name *Hotelling-Lawley trace* is shortened to *Hotelling's trace*

$$T = \text{trace}\left(\mathbf{S}_E^{-1}\mathbf{S}_H\right) = \sum_{k=1}^{m} \lambda_k$$

When H_o is true, the F statistic

$$F = \frac{2\left(sn^* + 1\right)}{s\left(2m^* + s + 1\right)} \frac{T}{s}$$

follows asymptotically an F distribution with degrees of freedom $\left(s\left(2m^* + s + 1\right), 2\left(sn^* + 1\right)\right)$. The F statistic is exact if $s = 1$.

The eta-squared statistic is $\eta^2 = (T / s) / (T / s + 1)$.

The noncentrality parameter is $\lambda = 2\left(sn^* + 1\right)\eta^2 / \left(1 - \eta^2\right)$.

The power is $1 - \text{NCDF.F}\left(\text{F}_\alpha, s\left(2m^* + s + 1\right), 2\left(sn^* + 1\right), \lambda\right)$ where F_α is the upper 100α percentage point of the central F distribution, and α is user-specified on the `ALPHA` keyword on the `CRITERIA` subcommand.

Pillai's Trace

$$V = \text{trace}\left(\mathbf{S}_H\left(\mathbf{S}_H + \mathbf{S}_E\right)^{-1}\right) = \sum_{k=1}^{m} \lambda_k / \left(1 + \lambda_k\right)$$

When H_o is true, the F statistic

$$F = \frac{\left(2n^* + s + 1\right)}{\left(2m^* + s + 1\right)} \frac{V}{(s - V)}$$

follows asymptotically an F distribution with degrees of freedom $\left(s\left(2m^* + s + 1\right), s\left(2n^* + s + 1\right)\right)$. The F statistic is exact if $s = 1$.

The eta-squared statistic is $\eta^2 = V / s$.

The noncentrality parameter is $\lambda = s\left(2n^* + s + 1\right)\eta^2 / \left(1 - \eta^2\right)$.

The power is $1 - \text{NCDF.F}\left(\text{F}_\alpha, s\left(2m^* + s + 1\right), s\left(2n^* + s + 1\right), \lambda\right)$ where F_α is the upper 100α percentage point of the central F distribution and α is user-specified on the ALPHA keyword on the CRITERIA subcommand.

Roy's Largest Root

$$\Theta = \lambda_1$$

which is the largest eigenvalue of $\mathbf{S}_E^{-1}\mathbf{S}_H$. When $H_\circ$ is true, the F statistic is

$$F = \Theta\left(n_e - \omega + r_H\right)/\omega$$

where $\omega = \max\left(l, r_E\right)$ is an upper bound of F that yields a lower bound on the significance level. The degrees of freedom are $\left(\omega, n_e - \omega + r_H\right)$. The F statistic is exact if $s = 1$.

The eta-squared statistic is $\eta^2 = \Theta / (1 + \Theta)$.

The noncentrality parameter is $\lambda = \left(n_e - \omega + r_H\right)\eta^2 / \left(1 - \eta^2\right)$.

The power is $1 - \text{NCDF.F}\left(\text{F}_\alpha, \omega, n_e - \omega + l, \lambda\right)$, where F_α is the upper 100α percentage point of the central F distribution and α is user-specified on the ALPHA keyword on the CRITERIA subcommand.

Individual Univariate Test

$$F = \frac{\mathbf{S}_{H;i}/l}{\mathbf{S}_{E;i}/(n-r_{\mathbf{X}})},\ i = 1,\ldots,m$$

where $\mathbf{S}_{H;i}$ and $\mathbf{S}_{E;i}$ are the *i*th diagonal elements of the matrices $\mathbf{S}_H$ and $\mathbf{S}_E$ respectively. Under the null hypothesis, the *F* statistic has an *F* distribution with degrees of freedom $(l, n-r_X)$.

The partial eta-squared statistic is $\eta^2 = \mathbf{S}_{H;i} / (\mathbf{S}_{H;i} + \mathbf{S}_{E;i})$.

The noncentrality parameter is $\lambda = (n-r_{\mathbf{X}})\mathbf{S}_{H;i}/\mathbf{S}_{E;i}$.

The power is $1 - \text{NCDF.F}(\text{F}_{\alpha}, 1, n-r_X, \lambda)$ where F_{α} is the upper 100α percentage point of the central *F* distribution and α is user-specified on the ALPHA keyword on the CRITERIA subcommand.

Bartlett's Test of Sphericity

Bartlett's test of sphericity is printed when the Residual SSCP matrix is requested.

Hypotheses

In Bartlett's test of sphericity the null hypothesis is $H_o: \Sigma = \sigma^2 \mathbf{I}_r$ versus the alternative hypothesis $H_1: \Sigma \neq \sigma^2 \mathbf{I}_r$ where $\sigma^2 > 0$ is unspecified and $\mathbf{I}_r$ is an $r \times r$ identity matrix.

Likelihood Ratio Test Statistic

$$\lambda = \begin{cases} \dfrac{|\mathbf{A}|^{n/2}}{(\text{trace}(\mathbf{A})/r)^{nr/2}} & \text{if trace}(\mathbf{A}) > 0 \\ \text{SYSMIS} & \text{if trace}(\mathbf{A}) \leq 0 \end{cases}$$

where $\mathbf{A} = \left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)' \mathbf{W}\left(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}\right)$ is the $r \times r$ matrix of residual sums of squares and cross products.

Chi-Square Approximation

Define $W = \lambda^{2/n}$. When n is large and under the null hypothesis that for $n - r_X \geq 1$ and $r \geq 2$,

$$\Pr\left(-\rho(n - r_X)\log W \leq c\right) = \Pr\left(\chi_f^2 \leq c\right) + \omega_2\left(\Pr\left(\chi_{f+4}^2 \leq c\right) - \Pr\left(\chi_f^2 \leq c\right)\right) + O\left(n^{-3}\right)$$

where

$$f = r(r+1)/2 - 1$$

$$\rho = 1 - \left(2r^2 + r + 2\right)\Big/\left(6r(n - r_X)\right)$$

$$\omega_2 = \frac{(r+2)(r-1)(r-2)\left(2r^3 + 6r^2 + 3r + 2\right)}{288r^2\left(n - r_X\right)^2\rho^2}$$

Chi-Square Statistic

$$c = \begin{cases} -\rho(n - r_X)\log W & \text{if } W > 0 \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

Degrees of Freedom

$$f = r(r+1)/2 - 1$$

Significance

$$1 - \text{CDF.CHISQ}(c, f) - \omega_2\left(\text{CDF.CHISQ}(c, f+4) - \text{CDF.CHISQ}(c, f)\right)$$

where CDF.CHISQ is the SPSS function for the cumulative chi-square distribution. The significance is reset to zero whenever the computed value is less than zero due to floating point imprecision.

Custom Hypothesis Tests

The TEST subcommand offers custom hypothesis tests. The hypothesis term is any effect specified (either explicitly or implicitly) in the DESIGN subcommand. The error term can be a linear combination of effects that are specified in the DESIGN subcommand or a sum of squares with specified degrees of freedom. The TEST subcommand is available only for univariate analysis; therefore, an F statistic is computed. When the error term is a linear combination of effects and no value for degrees of freedom is specified, the error degrees of freedom is approximated by the Satterthwaite (1946) method.

Notation

The following notation is used in this section:

S	Number of effects in the linear combination
q_s	Coefficient of the sth effect in the linear combination, $s = 1, \ldots, S$
l_s	Degrees of freedom of the sth effect in the linear combination, $s = 1, \ldots, S$
MS_s	Mean square of the sth effect in the linear combination, $s = 1, \ldots, S$
Q	Linear combination of effects
l_Q	Degrees of freedom of the linear combination
MS_Q	Mean square of the linear combination

Error Term

Mean Squares

If the error term is a linear combination of effects, the error mean square is

$$\mathrm{MS}_Q = \sum_{s=1}^{S} q_s \times \mathrm{MS}_s$$

If the user supplied the mean squares, MS_Q is equal to the number specified after the keyword VS. If $\mathrm{MS}_Q < 0$, the custom error term is invalid, and MS_Q is equal to the system-missing value and an error message is issued.

Degrees of Freedom

If $\mathrm{MS}_Q \geq 0$ and the user did not supply the error degrees of freedom, then the error degrees of freedom is approximated using the Satterthwaite (1946) method. Define

$$d_s = \begin{cases} (q_s \mathrm{MS}_s)^2 / l_s & \text{if } l_s > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then $D = \sum_{s=1}^{S} d_s$. The approximate error degrees of freedom is

$$l_Q = \begin{cases} (\mathrm{MS}_Q)^2 / D & \text{if } D > 0 \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

If $\mathrm{MS}_Q \geq 0$ and the user supplied the error degrees of freedom, l_Q is equal to the number following the keyword DF. If $l_Q < 0$, the custom degrees of freedom is invalid. In this case, l_Q is equal to the system-missing value and an error message is issued.

Hypothesis Test

F Statistic

The null hypothesis is that all parameters of the hypothesis effect are zero. The *F* statistic is used for testing this null hypothesis. Suppose the mean square and the degrees of freedom of the hypothesis effect are MS_H and l_H; then the *F* statistic is

$$F = \begin{cases} \dfrac{\mathrm{MS}_H}{\mathrm{MS}_Q} & \text{if } \mathrm{MS}_Q > 0 \,\&\, \mathrm{MS}_H \geq 0 \\ \mathrm{SYSMIS} & \text{otherwise} \end{cases}$$

Significance Level

The significance level is

$$\text{significance} = \begin{cases} 1 - \mathrm{CDF.F}(F, l_H, l_Q) & \text{if } l_H > 0, l_Q > 0 \,\&\, F \neq \mathrm{SYSMIS} \\ \mathrm{SYSMIS} & \text{otherwise} \end{cases}$$

where CDF.F is the SPSS function for the *F* cumulative distribution function.

Univariate Mixed Model

This section describes the algorithms pertaining to a random effects model. GLM offers mixed model analysis only for univariate models—that is, for $r = 1$.

Notation

The following notation is used throughout this section. Unless otherwise stated, all vectors are column vectors and all quantities are known.

k	Number of random effects, $k \geq 0$.
p_0	Number of parameters in the fixed effects, $p_0 \geq 0$.
p_i	Number of parameters in the *i*th random effect, $p_i \geq 0$, $i = 1, \ldots, k$.
σ_i^2	*Unknown* variance of the *i*th random effect, $\sigma_i^2 \geq 0$, $i = 1, \ldots, k$.

σ_e^2	*Unknown* variance of the residual term, $\sigma_e^2 > 0$.
$\mathbf{X}_i$	The $n \times p_i$ design matrix, $i = 0,1,\ldots,k$.
β_0	The length p_0 vector of parameters of the fixed effects.
β_i	The length p_i vector of parameters of the *i*th random effect, $i = 1,\ldots,k$.
$\mathbf{L}$	The $s \times p$ full row rank matrix. The rows are estimable functions. $s \geq 1$.

Relationships between these symbols and those defined at the beginning of the chapter are:

- $p = p_0 + p_1 + \cdots + p_k$
- $\mathbf{X} = \left[\mathbf{X}_0 | \mathbf{X}_1 | \ldots | \mathbf{X}_k\right]$
- $\mathbf{B} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$

Model

The mixed model is represented, following Rao (1973), as

$$\mathbf{Y} = \mathbf{X}_0\beta_0 + \sum_{i=1}^{k} \mathbf{X}_i\beta_i + \mathbf{e}$$

The random vectors $\beta_1,\ldots,\beta_k$ and $\mathbf{e}$ are assumed to be jointly independent. Moreover, the random vector β_i is distributed as $N_{p_i}\left(\mathbf{0}, \sigma_i^2 \mathbf{I}_{p_i}\right)$ for $i = 1,\ldots,k$ and the residual vector $\mathbf{e}$ is distributed as $N_n\left(\mathbf{0}, \sigma_e^2 \mathbf{W}^{-1}\right)$. Thus,

$$E(\mathbf{Y}) = \mathbf{X}_0\beta_0$$
$$\text{cov}(\mathbf{Y}) = \sum_{i=1}^{k} \sigma_i^2 \mathbf{X}_i \mathbf{X}_i' + \sigma_e^2 \mathbf{W}^{-1}$$

Expected Mean Squares

For the estimable function **L**, the expected hypothesis sum of squares is

$$E(SS_L) = E\left(\mathbf{Y}'\mathbf{W}^{\frac{1}{2}}\mathbf{A}_L\mathbf{W}^{\frac{1}{2}}\mathbf{Y}\right)$$
$$= \beta_0'\mathbf{X}_0'\mathbf{W}^{\frac{1}{2}}\mathbf{A}_L\mathbf{W}^{\frac{1}{2}}\mathbf{X}_0\beta_0 + \sum_{i=1}^{k}\sigma_i^2 \text{ trace}\left(\mathbf{X}_k'\mathbf{W}^{\frac{1}{2}}\mathbf{A}_L\mathbf{W}^{\frac{1}{2}}\mathbf{X}_k\right) + \sigma_e^2 \text{ trace}(\mathbf{A}_L)$$

where

$$\mathbf{A}_L = \mathbf{W}^{\frac{1}{2}}\mathbf{X}\mathbf{G}\mathbf{L}'(\mathbf{L}\mathbf{G}\mathbf{L}')^{-1}\mathbf{L}\mathbf{G}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$$

Since $\mathbf{L} = \mathbf{L}\mathbf{G}\mathbf{X}'\mathbf{W}\mathbf{X}$, $\text{trace}(\mathbf{A}_L) = s$ and $\mathbf{X}'\mathbf{W}^{\frac{1}{2}}\mathbf{A}_L\mathbf{W}^{\frac{1}{2}}\mathbf{X} = \mathbf{L}'(\mathbf{L}\mathbf{G}\mathbf{L}')^{-1}\mathbf{L}$. The matrix $\mathbf{X}'\mathbf{W}^{\frac{1}{2}}\mathbf{A}_L\mathbf{W}^{\frac{1}{2}}\mathbf{X}$ can therefore be computed in the following way:

1. Compute an $s \times s$ upper triangular matrix **U** such that $\mathbf{U}'\mathbf{U} = \mathbf{L}\mathbf{G}\mathbf{L}'$ by the Cholesky decomposition.
2. Invert the matrix **U** to give $\mathbf{U}^{-1}$.
3. Compute $\mathbf{C} = \mathbf{L}'\mathbf{U}^{-1}$.

Now we have $\mathbf{X}'\mathbf{W}^{\frac{1}{2}}\mathbf{A}_L\mathbf{W}^{\frac{1}{2}}\mathbf{X} = \mathbf{C}\mathbf{C}'$. If the rows of **C** are partitioned into the same-size submatrices as those contained in **X**—that is,

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_0 \\ \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_k \end{bmatrix}$$

where $\mathbf{C}_i$ is a $p_i \times s$ submatrix—then $\mathbf{X}_k'\mathbf{W}^{\frac{1}{2}}\mathbf{A}_L\mathbf{W}^{\frac{1}{2}}\mathbf{X}_k = \mathbf{C}_i\mathbf{C}_i'$, $i = 0,1,\ldots,k$.

Since $\text{trace}(\mathbf{C}_i\mathbf{C}_i')$ is equal to the sum of squares of the elements in $\mathbf{C}_i$, denoted by $\text{SSQ}(\mathbf{C}_i)$, the matrices $\mathbf{C}_i\mathbf{C}_i'$ need not be formed. The preferred computational formula for the expected sum of squares is

$$E(SS_L) = \beta_0'\mathbf{C}_0\mathbf{C}_0'\beta_0 + \sum_{i=1}^{k}\sigma_i^2\,\text{SSQ}(\mathbf{C}_i) + s\sigma_e^2$$

Finally the expected mean square is

$$E(MS_L) = \frac{1}{s}E(SS_L) = \frac{1}{s}\beta_0'\mathbf{C}_0\mathbf{C}_0'\beta_0 + \frac{1}{s}\sum_{i=1}^{k}\sigma_i^2\,\text{SSQ}(\mathbf{C}_i) + \sigma_e^2$$

For the residual term, the expected residual mean square is: $E(MSE) = \sigma_e^2$.

Note: GLM does not compute the term $\frac{1}{s}\beta_0'\mathbf{C}_0\mathbf{C}_0'\beta_0$ but reports the fixed effects whose corresponding row block in $\mathbf{C}_0$ contains nonzero elements.

Hypothesis Test in Mixed Models

Suppose MS_L is the mean square for the effect whose estimable function is **L**, and s_L is the associated degrees of freedom. The *F* statistic for testing this effect is

$$F = \frac{MS_L}{MS_{E(L)}}$$

where $MS_{E(L)}$ is the mean square of the error term with $s_{E(L)}$ degrees of freedom.

Null Hypothesis Expected Mean Squares

If the effect being tested is a fixed effect, its expected mean square is

$$E(MS_L) = \sigma_e^2 + c_1\sigma_1^2 + \cdots + c_k\sigma_k^2 + Q(L)$$

where $c_1,\ldots,c_k$ are coefficients and $Q(L)$ is a quadratic term involving the fixed effects. Under the null hypothesis, it is assumed that $Q(L)=0$. Although the quadratic term may involve effects that are unrelated to the effect being tested, such effects are assumed to be zero in order to draw a correct inference for the effect being tested. Therefore, under the null hypothesis, the expected mean square is

$$E(MS_L)=\sigma_e^2+c_1\sigma_1^2+\cdots+c_k\sigma_k^2$$

If the effect being tested is a random effect, say the *j*th $(1\le j\le k)$ random effect, its expected mean square is

$$E(MS_L)=\sigma_e^2+c_1\sigma_1^2+\cdots+c_k\sigma_k^2$$

Under the null hypothesis $\sigma_j^2=0$; hence, the expected mean square is

$$E(MS_L)=\sigma_e^2+\sum_{1\le i\le k,i\ne j}c_i\sigma_i^2$$

Error Mean Squares

Let MS_i be the mean square of the *i*th $(i=1,\ldots,k)$ random effect. Let s_i be the corresponding degrees of freedom. The error term is then found as a linear combination of the expected mean squares of the random effects:

$$MS_{E(L)}=q_1MS_1+\cdots+q_kMS_k+q_{k+1}MSE$$

such that

$$E\left(MS_{E(L)}\right)=q_1E(MS_1)+\cdots+q_kE(MS_k)+q_{k+1}E(MSE)=\sigma_e^2+c_1\sigma_1^2+\cdots+c_k\sigma_k^2$$

If $s_i=0$ $(1\le i\le k)$ then $q_i=0$.

The error degrees of freedom is computed using the Satterthwaite (1946) method:

$$s_{E(L)} = \frac{\left(MS_{E(L)}\right)^2}{\sum_{1 \le i \le k; s_i > 0} \left(q_i MS_i\right)^2 / s_i}$$

If the design is balanced, the above F statistic is approximately distributed as an F distribution with degrees of freedom $\left(s_L, s_{E(L)}\right)$ under the null hypothesis. The statistic is exact when only one random effect is used as the error term—that is, $q_{i_0} = 1$ and $q_i = 0$ for $i \neq i_0$. If the design is not balanced, the above approximation may not be valid (even when only one random effect is used as the error term) because the hypothesis term and the error term may not be independent.

References

Belsley, D. A., Kuh, E., and Welsch R. E. 1980. *Regression diagnostics*. New York: John Wiley & Sons, Inc.

Clarke, M. R. B. 1982. Algorithm AS 178: The Gauss-Jordan sweep operator with detection of collinearity. *Applied Statistics*, Vol. 31, No. 2: 166–168.

Goodnight, J. H. 1979. A tutorial on the SWEEP operator. *The American Statistician*, Vol. 33, No. 3: 149–158.

Rao, C. R. 1951. An asymptotic expansion of the distribution of Wilks' criterion. *Bulletin of the International Statistical Institute*, 33, Part 2: 177–180.

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley & Sons, Inc.

Ridout, M. S., and Cobby, J. M. 1989. A remark on algorithm AS 178. *Applied Statistics*, 38: 420–422.

Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2: 110–114.

Searle, S. R., Speed, F. M., and Milliken, G. A. 1980. Population marginal means in the linear model: an alternative to least squares means. *The American Statistician*, 34: 216–221.