The Pearson product-moment method is used to estimate the amount of linear correlation between paired columns, say X and Y, of a n by m data matrix where it is assumed that the values are of the continuous type from a normal bivariate distribution, and not integers such as frequencies or categorical variables. The null hypothesis is that X and Y are independent, i.e. have zero covariance, that is

$$H_0 : X \text{ and } Y \text{ are from a bivariate normal distribution with } \rho = 0.$$

Example 1

From the SIMFIT main menu choose [Statistics], [Multivariate], [Correlation], then analyze `g02baf.tf1`, the test file provided, using the Pearson product-moment technique. This file contains the following 5 by 3 data matrix

2.0	3.0	3.0
4.0	6.0	4.0
9.0	9.0	0.0
0.0	12.0	2.0
12.0	-1.0	5.0

and analysis leads first to the correlation coefficients and corresponding p values

Matrix A, Pearson correlation results		
Upper triangle: r		
Lower triangle: corresponding two-tail p values		
.....	-0.5704	0.1670
0.3153	.....	-0.7486
0.7883	0.1455	.....

which is in the following simplified but comprehensive format

$$A = \begin{bmatrix} \cdots & r_{12} & r_{13} \\ p_{12} & \cdots & r_{23} \\ p_{13} & p_{23} & \cdots \end{bmatrix}$$

where the values a_{ij} for matrix A in the table are interpreted as now described. For $j > i$ in the strict upper triangle, then $a_{ij} = r_{ij} = r_{ji}$ are the correlation coefficients, while for $i > j$ in the strict lower triangle $a_{ij} = p_{ij} = p_{ji}$ are the corresponding two-tail probabilities. In other words, since $r_{ij} = r_{ji}$, $p_{ij} = p_{ji}$, while $r_{ii} = 1$, there will only be $m(m-1)/2$ independent correlations coefficients, and so the diagonal $r_{ii} = 1$ are shown as dots. For instance $r_{12} = -0.5704$ is the correlation coefficient for columns 1 and 2, while $p_{12} = 0.3153$ is the two-tail p value for this correlation coefficient. The table indicates that none of the correlations are significant in this case, that is, the probability of obtaining such pairwise linearity in a random swarm of points from a multivariate normal distribution is not low.

This is then followed by a likelihood ratio test that the full correlation matrix $R = r_{ij}$ for the data matrix is the identity matrix with the following results.

Test for absence of any significant correlations	
H_0: correlation matrix is the identity matrix	
Determinant	0.2290
Test statistic (TS)	3.194
Degrees of freedom	3
$P(\chi^2 \geq TS)$	0.3627

To test the hypothesis of no significant correlations, i.e.
H_0: the covariance matrix is diagonal, or equivalently
H_0: the correlation matrix R is the identity matrix, the likelihood ratio test statistic TS, i.e.

$$-2\log\lambda = -(n-(2m+11)/6)\log|R|$$

is used, where $|R|$ is the determinant of the full correlation matrix (not the previous A matrix) which has the asymptotic chi-square distribution with $m(m-1)/2$ degrees of freedom.

Example 2

This example illustrates the analysis of SIMFIT test file `cluster.tf1` which contains the following data set

1.0	4.0	2.0	11.0	6.0	4.0	3.0	9.0
8.0	5.0	1.0	14.0	19.0	7.0	13.0	21.0
3.0	1.0	3.0	1.0	3.0	6.0	23.0	37.0
9.0	0.0	7.0	7.0	1.0	2.0	21.0	2.0
7.0	12.0	9.0	5.0	14.0	9.0	12.0	14.0
2.0	13.0	15.0	2.0	23.0	6.0	34.0	8.0
11.0	7.0	2.0	1.0	4.0	17.0	11.0	4.0
6.0	3.0	7.0	12.0	11.0	8.0	8.0	0.0
8.0	21.0	1.0	10.0	31.0	9.0	3.0	18.0
19.0	14.0	12.0	9.0	16.0	10.0	0.0	27.0
17.0	18.0	10.0	6.0	19.0	14.0	1.0	24.0
15.0	21.0	8.0	7.0	17.0	12.0	4.0	22.0

leading to this correlation and probability matrix

Upper triangle = r, Lower = corresponding two-tail p values

.....	0.5295	0.2874	0.0662	0.1941	0.6255	-0.5876	0.3010
0.0766	.....	0.3285	-0.0219	0.7930	0.5338	-0.4230	0.3006
0.3650	0.2971	.....	-0.2833	0.2165	0.0264	0.2314	-0.0304
0.8381	0.9460	0.3723	.....	0.2787	-0.2837	-0.5238	-0.1166
0.5455	0.0021	0.4992	0.3804	.....	0.2029	-0.1949	0.2144
0.0296	0.0738	0.9351	0.3715	0.5271	.....	-0.4532	0.1360
0.0445	0.1706	0.4694	0.0805	0.5439	0.1390	.....	-0.1696
0.3418	0.3424	0.9253	0.7181	0.5035	0.6735	0.5983	.....

followed by the results displayed next for a likelihood ratio test.

Test for absence of any significant correlations

H_0: correlation matrix is the identity matrix		
Determinant	0.002476	
Test statistic (TS)	45.01	
Degrees of freedom	28	
$P(\chi^2 \geq TS)$	0.0220	*Reject H_0 at 5% significance level*

From the r values in the strict upper triangle, the p values in the strict lower triangle, and the chi-square test there are linear correlations, and in such cases it would be usual to select pairs of columns for closer analysis.

Analyzing selected pairs of columns

For example, the results for analyzing columns 1 and 2 will be considered.

For the next analysis: X is column 1, Y is column 2

Linear regression: $y(x) = A + B * x$, $x(y) = C + D * y$			
Sample	size = 12		
For X	mean = 8.8333	std. dev. = 5.7814	var. = 33.424
For Y	mean = 9.9167	std. dev. = 7.5973	var. = 57.720

First the parameter estimates for linear regression are calculated, where Estimate/Standard Error are t values to test for parameters significantly different from zero, Ppmcc is the Pearson product-moment correlation coefficient, and the Fisher z value is used to estimate a 95% confidence region for ρ. In this type of table $p \leq 0.05$ would be required to suggest a nonzero parameter at the 5% significance level.

Parameter	Estimate	Standard Error	Estimate/Standard Error	p
B (slope)	0.69583	0.35252	1.9739	0.0766
A (const)	3.7702	3.6748	1.0260	0.3291
r (Ppmcc)	0.52951	0.26826	1.9739	0.0766
r^2	0.28038			
y-variation due to x = 28.04%				
z(Fisher)	0.58946			

Note: $z = (1/2)\log[(1+r)/(1-r)]$

$r^2 = B * D$, and $t = r * \sqrt{[(n-2)/(1-r^2)]}$ = Estimate/Standard Error for B and D

The Pearson product-moment correlation coefficient r estimates ρ and

95% confidence limits using z are $-0.0771 \leq \rho \leq 0.8500$

Then this analysis of variance (ANOVA) table is displayed, where the F value is used to test for a significant regression slope. In this type of table $p \leq 0.05$ would be required to suggest a nonzero regression slope at the 5% significance level.

Source	Sum of squares	$ndof$	Mean square	F-value	p
due to regression	178.02	1	178.02	3.8962	0.0766
about regression	456.90	10	45.690		
total	634.92	11			

Conclusions:

B is not significantly different from zero ($p > 0.05$)

A is not significantly different from zero ($p > 0.05$)

The two best-fit unweighted regression lines are:

$y(x) = 3.7702 + 0.69583x$, and $x(y) = 4.8375 + 0.40294y$

Various options for plotting follow, and the theory necessary to interpret such correlation tests and visual displays will be presented next.

Theory

Given any set of n nonsingular (x_i, y_i) pairs, a correlation coefficient r can be calculated as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $-1 \leq r \leq 1$ and, using b_{xy} for the slope of the regression of X on Y, and b_{yx} for the slope of the regression of Y on X

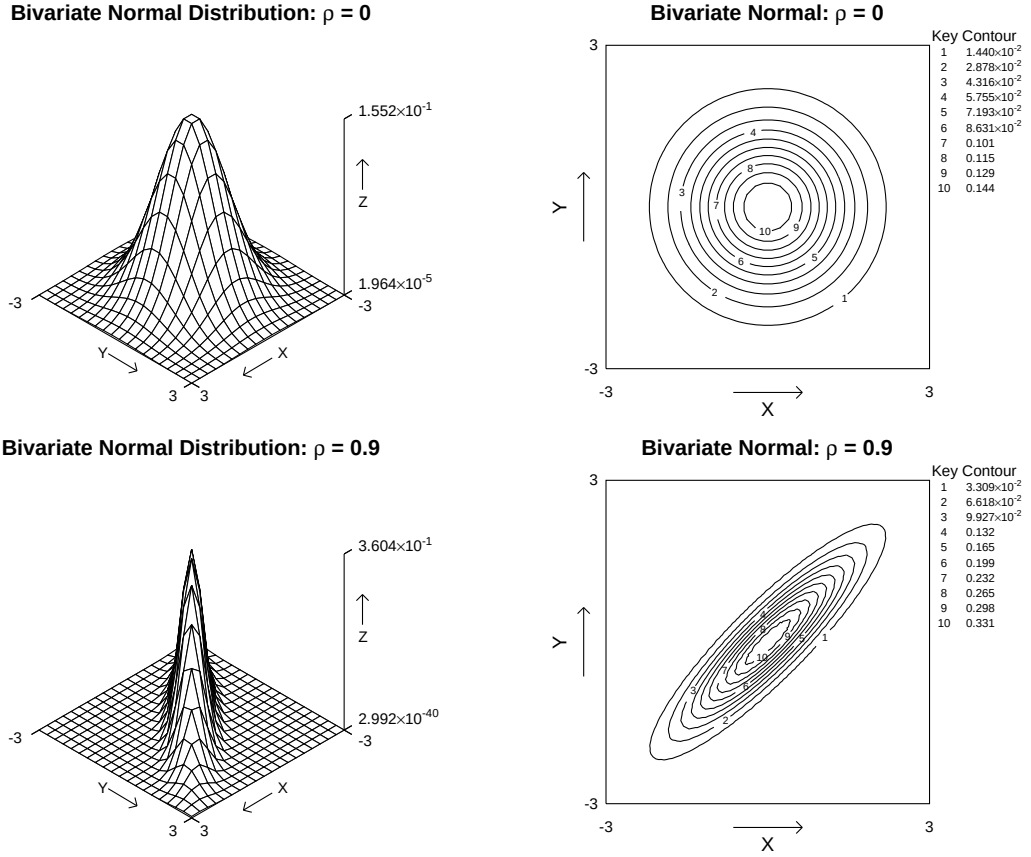
$$r^2 = b_{yx}b_{xy}.$$

However, only when X is normally distributed given Y, and Y is normally distributed given X can simple statistical tests be used for significant linear correlation. For instance, when the (x_i, y_i) pairs are from such a bivariate normal distribution, the statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

has a Student's t-distribution with $n-2$ degrees of freedom. It is also the t value required to test for nonzero slope in the regression of Y on X, and X on Y, for which a p value can be calculated.

The next figure illustrates how the elliptical contours of constant probability for a bivariate normal distribution are aligned with the X and Y axes when X and Y are uncorrelated, i.e., $\rho = 0$ but are inclined otherwise. In this example $\mu_X = \mu_Y = 0$ and $\sigma_x = \sigma_Y = 1$, but in the upper figure $\rho = 0$, while in the lower figure $\rho = 0.9$. The Pearson product-moment correlation coefficient r is an estimator of ρ, and it can can be used to test for independence of X and Y.



The SIMFIT product-moment correlation procedure can be used when you have a data matrix X consisting of $m > 1$ columns of $n > 1$ measurements (not counts or categorical data) and wish to test for pairwise linear correlations, i.e., where pairs of columns can be regarded as consistent with a bivariate normal distribution. In matrix notation, the relationships between such a n by m data matrix X, the same matrix Y after centering by subtracting each column mean from the corresponding column, the sum of squares and products matrix C, the covariance matrix S, the correlation matrix R, and the diagonal matrix D of standard deviations are

$$\begin{aligned} C &= Y^T Y \\ S &= \frac{1}{n-1} C \\ D &= \mathrm{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \ldots, \sqrt{s_{mm}}) \\ R &= D^{-1} S D^{-1} \\ S &= DRD. \end{aligned}$$

So, for all pairs of columns, the sample correlation coefficients r_{jk} are given by

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}},$$

$$\text{where } s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k),$$

and the corresponding t_{jk} values and significance levels p_{jk} are calculated then output in matrix format with the correlations as a strict upper triangular matrix, and the significance levels as a strict lower triangular matrix.

Plotting lines on correlation diagrams

You can plot either both unweighted regression lines, the unweighted reduced major axis line, or the unweighted major axis line on such scattergrams and the difference between these types will now be outlined.

For n pairs (x_i, y_i) with mean $x = \bar{x}$ and mean $y = \bar{y}$, the variances and covariance required are

$$S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$S_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Also, for an arbitrary point (x_i, y_i) and a straight line defined by $y = a + bx$ the squares of the vertical, horizontal, and orthogonal (i.e. perpendicular) distances, v_i^2, h_i^2, and o_i^2 between the point and the line are

$$v_i^2 = [y_i - (a + bx_i)]^2$$
$$h_i^2 = v_i^2/b^2$$
$$o_i^2 = v_i^2/(1 + b^2).$$

Ordinary least squares

If x is regarded as an exact variable free from random variation or measurement error while y has random variation, then the best fit line from minimizing the sum of v_i^2 is

$$y_1(x) = \hat{\beta}_1 x + [\bar{y} - \hat{\beta}_1 \bar{x}]$$

where $\hat{\beta}_1 = S_{xy}/S_{xx}$. However, if y is regarded as an exact variable while x has random variation, then the best fit line for x as a function of y from minimizing the sum of h_i^2 would be

$$x_2(y) = (1/\hat{\beta}_2)y + [\bar{x} - (1/\hat{\beta}_2)\bar{y}]$$

where $\hat{\beta}_2 = S_{yy}/S_{xy}$ or, rearranging to express the line as $y_2(x)$,

$$y_2(x) = \hat{\beta}_2 x + [\bar{y} - \hat{\beta}_2 \bar{x}],$$

emphasizing that the slope of the regression line for $y_2(x)$ is the reciprocal of the slope for $x_2(y)$. Since neither of these two best fit lines can be regarded as satisfactory, SimFIT plots both lines such that $y_1(x)$ covers the range of x values while $x_2(y)$ covers the range of y values. However these two lines intersect at $(\bar{x}, \bar{y})$ and, from the fact that the ratio of slopes equals the square of the correlation coefficient, that is,

$$r^2 = \hat{\beta}_1/\hat{\beta}_2,$$

then two best fit lines with similar slopes suggests strong linear correlation, whereas one line almost parallel to the x axis and the other almost parallel to the y axis would indicate negligible linear correlation. For instance,

if there is no linear correlation between x and y, then the slope of the regression line for $y(x)$ i.e. $\hat{\beta}_1$ would be zero, as would be the slope of the regression line for $x(y)$ i.e. $1/\hat{\beta}_2$ leading to $r^2 = 0$. Conversely strong linear correlation would lead to $\hat{\beta}_1 = \hat{\beta}_2$ and $r^2 = 1$.

The major axis and reduced major axis lines to be discussed next are attempts to get round the necessity to plot two lines and just have one best fit line intermediate between these two lines to represent the correlation.

The major axis line

Here it is the sum of o_i^2, the squares of the orthogonal distances between the points and the best fit line, that is minimized to yield the slope as

$$\hat{\beta}_3 = \frac{1}{2}\left(\hat{\beta}_2 - (1/\hat{\beta}_1) + \gamma\sqrt{4 + (\hat{\beta}_2 - (1/\hat{\beta}_1))^2}\right)$$

where $\gamma = 1$ if $S_{xy} > 0$, $\gamma = 0$ if $S_{xy} = 0$, and $\gamma = -1$ if $S_{xy} < 0$, so that the major axis line is

$$y_3(x) = \hat{\beta}_3 x + [\bar{y} - \hat{\beta}_3 \bar{x}].$$

Actually $\hat{\beta}_3$ is the slope of the first principal component axis and so it points in the direction of maximum variability.

The reduced major axis line

Instead of minimizing the sum of squares of the vertical distances v_i^2, or horizontal distances h_i^2, it is possible to minimize the sum of the areas of the triangles formed by the v_i, h_i with the best fit line as hypotenuse, i.e. $v_i h_i/2$, to obtain the reduced major axis line as

$$y_4(x) = \hat{\beta}_4 x + [\bar{y} - \hat{\beta}_4 \bar{x}].$$

Here

$$\begin{aligned}\hat{\beta}_4 &= \gamma\sqrt{S_{yy}/S_{xx}} \\ &= \gamma\sqrt{\hat{\beta}_1\hat{\beta}_2}\end{aligned}$$

so that the slope of the reduced major axis line is the geometric mean of the slopes of the regression of y on x and x on y.

Recommendations for plotting lines on scattergrams

1. Plotting both both simple regression lines is the most useful and least controversial. Such lines tending to coincidence indicate strong linear correlation, while lines approaching perpendicularity indicate absence of significant linear correlation.
2. If a single line must be plotted to summarize the overall correlation it should be the reduced major axis line, as this allows for uncertainty in both variables and is not so controversial as the major axis line, which requires both axes to have similar units, as in allometry.
3. It should not be just one of the simple regression lines, since the line plotted must be independent of which variable is regarded as x and which is regarded as y.

Plotting bivariate confidence ellipses: basic theory

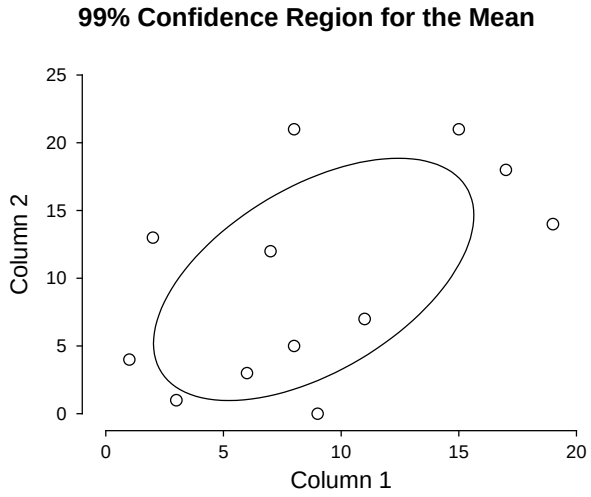
For a p-variate normal sample of size n with mean $\bar{x}$ and variance matrix estimate S, the region

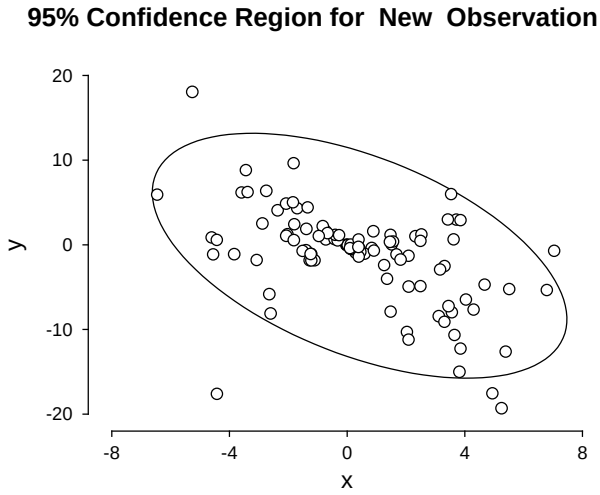
$$P\left\{(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) \le \frac{p(n-1)}{n(n-p)} F^{\alpha}_{p,n-p}\right\} \le 1-\alpha$$

can be regarded as a $100(1-\alpha)\%$ confidence region for μ. The next figure illustrates this for columns 1 and 2 of `cluster.tf1` discussed previously. Alternatively, the region satisfying

$$P\left\{(x-\bar{x})^T S^{-1}(x-\bar{x}) \le \frac{p(n^2-1)}{n(n-p)} F^{\alpha}_{p,n-p}\right\} \le 1-\alpha$$

can be interpreted as a region that with probability $1-\alpha$ would contain another independent observation x, as shown for the swarm of points in the next figure.

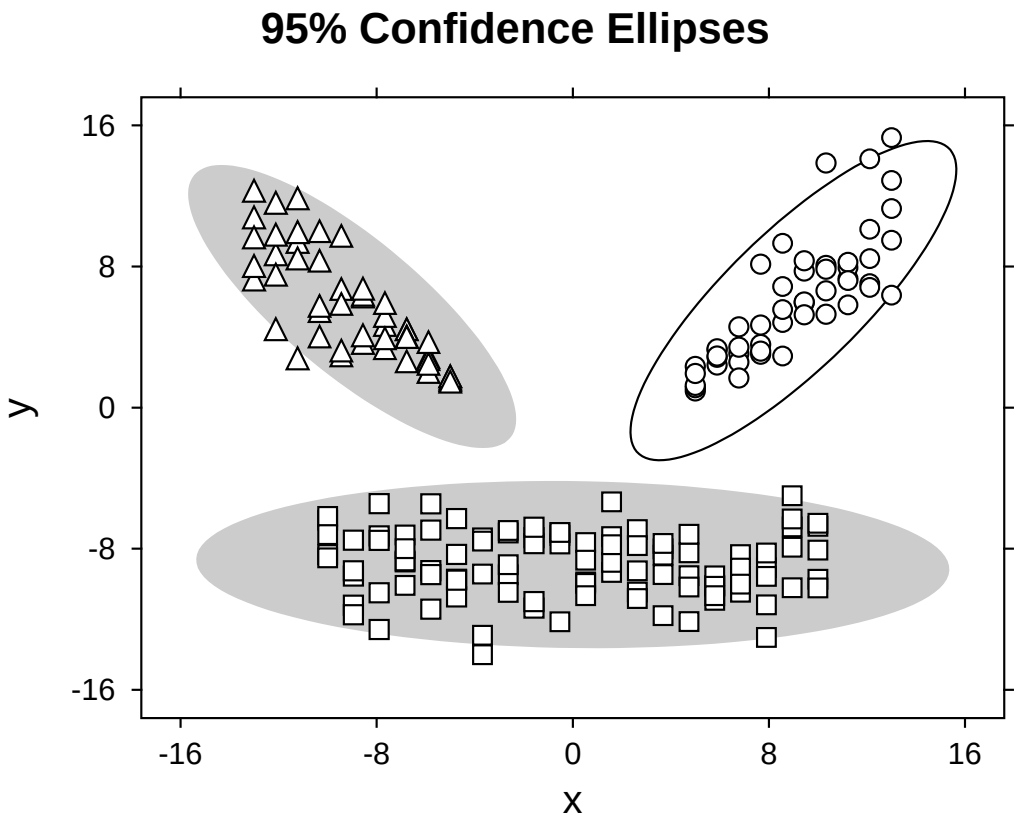




The μ confidence region contracts with increasing n, limiting application to small samples, but the new observation ellipse does not, making it useful for visualizing if data do represent a bivariate normal distribution, while inclination of the principal axes away from parallel with the plot axes demonstrates linear correlation. This technique is only justified if the data are from a bivariate normal distribution and are independent of the variables in the other columns, as indicated by the correlation matrix.

Plotting bivariate confidence ellipses: regions

Often a two dimensional swarm of points results from projecting data that have been partitioned into groups into a subspace of lower dimension in order to visualize the distances between putative groups, e.g., after principal components analysis or similar. If the projections are approximately bivariate normal then confidence ellipses can be added, as in the figure below.



The following steps were used to create this figure and can be easily adapted for any number of sets of two dimensional group coordinates.

1. For each group a file of values for x and y coordinates in the projected space was saved.
2. Each file was analyzed for correlation using the SimFIT correlation analysis procedure.
3. After each correlation analysis, the option to create a 95% confidence ellipse for the data was selected, and the ellipse coordinates were saved to file.
4. A library file was created with the ellipse coordinates as the first three files, and the groups data files as the next three files.
5. The library file was read into **simplot**, then colors and symbols were chosen.

Note that, because the ellipse coordinates are read in as the first coordinates to be plotted, the option to plot lines as closed polygons can be used to represent the confidence ellipses as colored background regions.