Ridge Regression

Summary

The **Ridge Regression** procedure is designed to fit a multiple regression model when the independent variables exhibit *multicollinearity*. Multicollinearity refers to the situation in which the X variables are correlated amongst themselves, which often leads to imprecise estimates of the regression model coefficients using ordinary least squares. By allowing a small amount of bias in the estimates, ridge regression can often reduce the variability of the estimated coefficients and give a more stable and interpretable model.

A ridge trace and plot of the variance inflation factors (VIF) is provided to help select the value of the ridge parameter.

Sample StatFolio: *ridge reg.sgp*

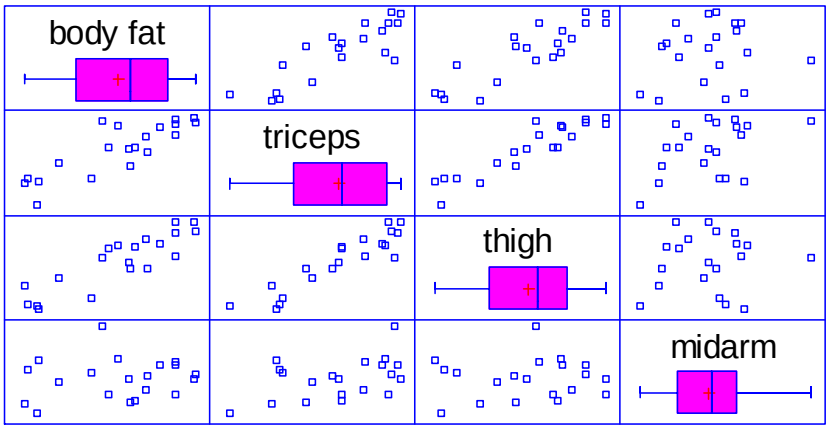
Sample Data:

The file *bodyfat.sgd* contains a set of data from Neter et al. (1998) describing $n = 20$ females between 25 and 35 years of age. The data include measurements of their body fat, triceps skinfold thickness, thigh circumference, and midarm circumference.

subject	*bodyfat*	*triceps*	*thigh*	*midarm*
1	11.9	19.5	43.1	29.1
2	22.8	24.7	49.8	28.2
3	18.7	30.7	51.9	37
4	20.1	29.8	54.3	31.1
5	12.9	19.1	42.2	30.9
6	21.7	25.6	53.9	23.7
7	27.1	31.4	58.5	27.6
8	25.4	27.9	52.1	30.6
9	21.3	22.1	49.9	23.2
10	19.3	25.5	53.5	24.8
11	25.4	31.1	56.6	30
12	27.2	30.4	56.7	28.3
13	11.7	18.7	46.5	23
14	17.8	19.7	44.2	28.6
15	12.8	14.6	42.7	21.3
16	23.9	29.5	54.4	30.1
17	22.6	27.7	55.3	25.7
18	25.4	30.2	58.6	24.6
19	14.8	22.7	48.2	27.1
20	21.1	25.2	51	27.5

It is desired to construct a model relating *body fat* (Y) to the other 3 variables. However, as might be expected, the 3 predictor variables are themselves highly correlated.

To illustrate the relationships between all of the variables, consider the following matrix plot:



Both *triceps* and *thigh* show a strong positive correlation with *bodyfat*. However, they are also strongly correlated with each other.

The results of using the *Multiple Regression* on this data are shown below:

Multiple Regression - body fat

Dependent variable: body fat

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	117.085	99.7824	1.1734	0.2578
triceps	4.33409	3.01551	1.43727	0.1699
thigh	-2.85685	2.58202	-1.10644	0.2849
midarm	-2.18606	1.5955	-1.37014	0.1896

Analysis of Variance

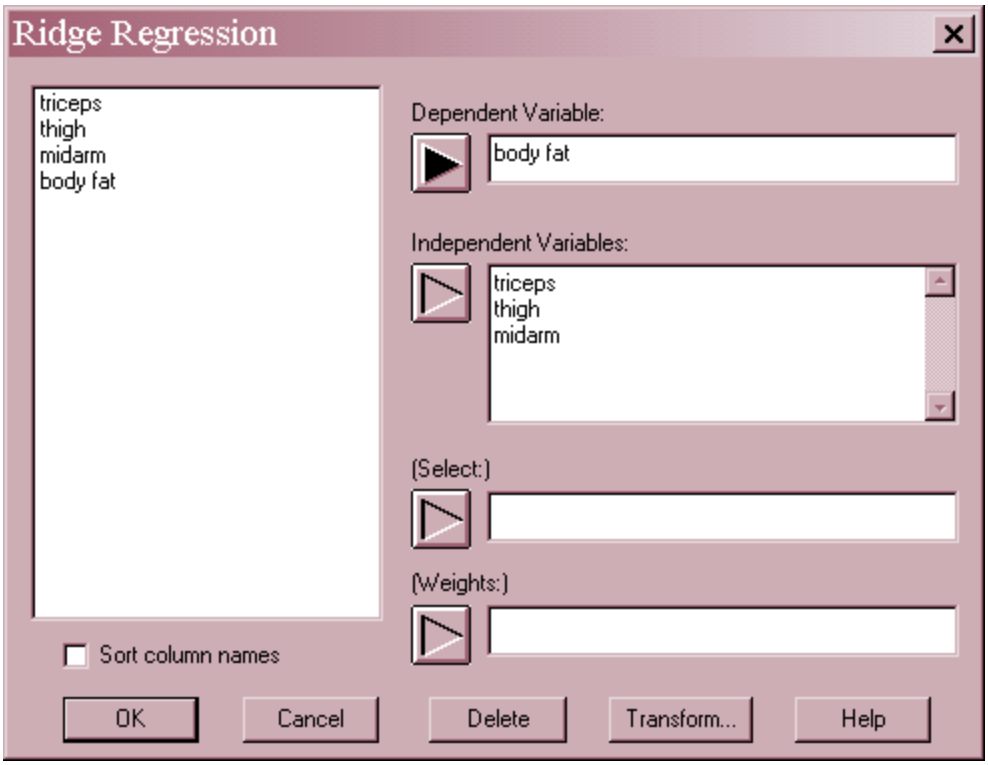
Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	396.985	3	132.328	21.52	0.0000
Residual	98.4049	16	6.15031		
Total (Corr.)	495.389	19			

R-squared = 80.1359 percent
R-squared (adjusted for d.f.) = 76.4113 percent
Standard Error of Est. = 2.47998
Mean absolute error = 1.88563
Durbin-Watson statistic = 2.24291 (P=0.2420)
Lag 1 residual autocorrelation = -0.167729

Notice that none of the coefficients has a significant P-value, although the model itself is highly significant. Note also that the coefficient on *thigh* is negative, despite its strong positive correlation with *bodyfat*. These imprecisions, caused by the multicollinearity amongst the predictor variables, is what ridge regression is designed to overcome.

Data Input

The data input dialog box requests the names of the columns containing the dependent variable Y and the independent variables X:



- **Y:** numeric column containing the *n* observations for the dependent variable Y.
- **X:** numeric columns containing the *n* values for the independent variables X.
- **Select:** subset selection.
- **Weight:** an optional numeric column containing weights to be applied to the squared residuals when performing a weighted least squares fit.

Analysis Summary

The *Analysis Summary* displays information about the fitted model at a selected value of the ridge parameter λ. The ridge parameter controls the amount of bias allowed in the coefficient estimates, with λ = 0 corresponding to the unbiased, ordinary least squares estimates. As λ increases, the bias increases, but the variance of the coefficients decreases.

Ridge Regression - body fat
Dependent variable: body fat
Number of complete cases: 20

Model Results for Ridge Parameter = 0.02

		Variance
		Inflation
Parameter	*Estimate*	*Factor*
CONSTANT	-7.40343	
triceps	0.555353	1.10255
thigh	0.368144	1.08054
midarm	-0.191627	1.01051

R-Squared = 77.2602 percent
R-Squared (adjusted for d.f.) = 72.9965 percent
Standard Error of Est. = 2.59924
Mean absolute error = 1.92607
Durbin-Watson statistic = 2.38078
Lag 1 residual autocorrelation = -0.210305

Residual Analysis

	Estimation	*Validation*
n	20	
MSE	6.75602	
MAE	1.92607	
MAPE	10.4569	
ME	-3.90799E-15	
MPE	-1.71498	

- **Coefficients:** the estimated coefficients and variance inflation factors. The estimates of the model coefficients can be used to write the fitted equation, which in the example is

$$body\ fat = -7.40343 + 0.555353\ triceps + 0.368144\ thigh - 0.191627\ midarm \qquad (1)$$

  The variance inflation factors measure how large the variance of the coefficients is compared to what it would be if the independent variables were uncorrelated. Note that the VIF's are quite close to 1. (Try setting the ridge parameter to 0, which corresponds to ordinary least squares. The variance inflation factors are 709, 564, and 105, respectively!)

- **Analysis of Variance:** decomposition of the variability of the dependent variable Y into a model sum of squares and a residual or error sum of squares. Of particular interest is the F-test and its associated P-value, which tests the statistical significance of the fitted model. A small P-Value (less than 0.05 if operating at the 5% significance level) indicates that a significant relationship of the form specified exists between Y and the independent variables. In the sample data, the model is highly significant.

- **Statistics:** summary statistics for the fitted model, including:

R-squared - represents the percentage of the variability in Y which has been explained by the fitted regression model, ranging from 0% to 100%. For the sample data, the regression has accounted for about 77.3% of the variability in the miles per gallon. The remaining 22.7% is attributable to deviations from the model, which may be due to other factors, to measurement error, or to a failure of the current model to fit the data adequately.

Adjusted R-Squared – the R-squared statistic, adjusted for the number of coefficients in the model. This value is often used to compare models with different numbers of coefficients.

Mean Absolute Error – the average absolute value of the residuals.

Durbin-Watson Statistic – a measure of serial correlation in the residuals. If the residuals vary randomly, this value should be close to 2. A small P-value indicates a non-random pattern in the residuals. For data recorded over time, a small P-value could indicate that some trend over time has not been accounted for. In the current example, the P-value is greater than 0.05, so there is not a significant correlation at the 5% significance level.

Lag 1 Residual Autocorrelation – the estimated correlation between consecutive residuals, on a scale of –1 to 1. Values far from 0 indicate that significant structure remains unaccounted for by the model.

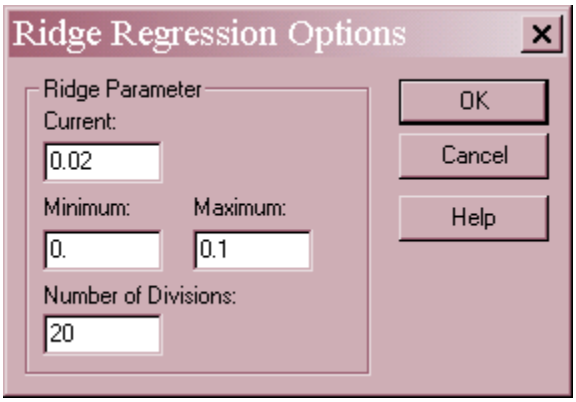
Residual Analysis – if a subset of the rows in the datasheet have been excluded from the analysis using the *Select* field on the data input dialog box, the fitted model is used to make predictions of the Y values for those rows. This table shows statistics on the prediction errors, defined by

$$e_i = y_i - \hat{y}_i \qquad (2)$$

Included are the mean squared error (MSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the mean error (ME), and the mean percentage error (MPE). The validation statistics can be compared to the statistics for the fitted model to determine how well that model predicts observations outside of the data used to fit it.

The fitted model now makes much more intuitive sense, with the coefficients on *triceps* and *thigh* both positive, as expected given their positive correlation with *body fat*.

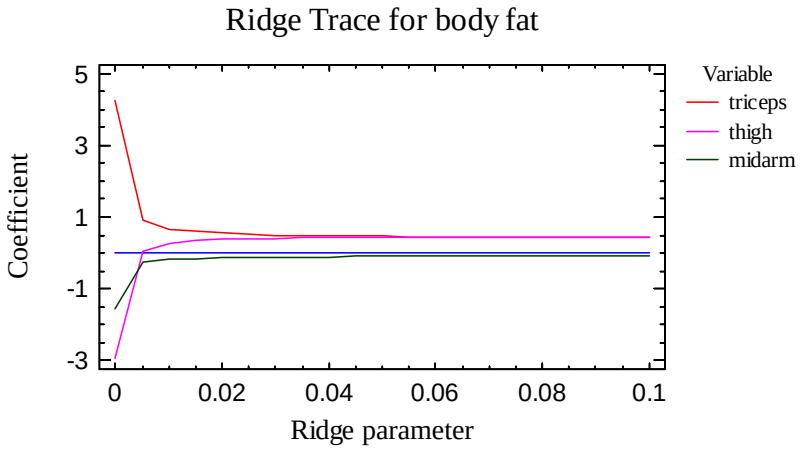
Analysis Options



- **Current**: value of the ridge parameter λ used to fit the model.
- **Minimum and maximum**: the range of values used in the comparative tables and plots.
- **Number of divisions**: the number of divisions into which the range of values is divided when creating the tables and plots.

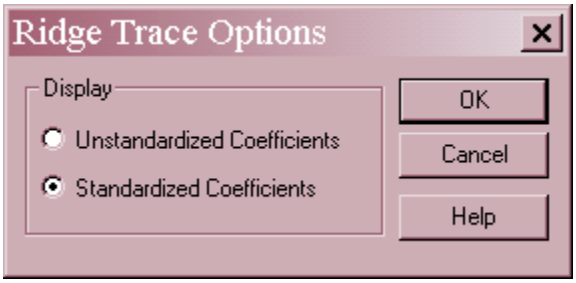
Ridge Trace

The *Ridge Trace* displays the coefficient estimates using various values of the ridge parameter λ:



In many cases, the coefficient estimates will change dramatically as λ is moved away from 0. Eventually, the coefficients will stabilize, as they do in the vicinity of λ = 0.02 in the above plot. Note the change in sign for the coefficient on *thigh*, as well as the approach to 0 of the coefficient on *midarm*. Unfortunately, standard statistical significance tests no longer apply, so it is difficult to determine which coefficients are statistically significant. None the less, the resulting model is likely to be much more stable than the model fit by ordinary least squares, particularly when extrapolating outside of the experimental region.

Pane Options



- **Display**: select standardized or unstandardized coefficients. The standardized coefficients are the coefficients of a model in which all variables have been standardized by subtracting their sample means and dividing by their sample standard deviations and then dividing by the square root of $n - 1$.

Regression Coefficients

This table shows the regression coefficients at selected values of the ridge parameter:

Regression Coefficients

Ridge Parameter	triceps	thigh	midarm
0.0	4.33409	-2.85685	-2.18606
0.005	0.917717	0.0653496	-0.385209
0.01	0.685303	0.261826	-0.261854
0.015	0.600002	0.332377	-0.216024
0.02	0.555353	0.368144	-0.191627
0.025	0.527656	0.389415	-0.176177
0.03	0.508638	0.403272	-0.165315
0.035	0.494658	0.412832	-0.157122
0.04	0.483862	0.419681	-0.150622
0.045	0.475207	0.424713	-0.145264
0.05	0.468061	0.428468	-0.140717
0.055	0.462021	0.431293	-0.136765
0.06	0.456813	0.433419	-0.133266
0.065	0.45225	0.435009	-0.13012
0.07	0.448195	0.436178	-0.127255
0.075	0.444548	0.437012	-0.124619
0.08	0.441236	0.437574	-0.122172
0.085	0.438199	0.437914	-0.119882
0.09	0.435394	0.43807	-0.117727
0.095	0.432785	0.438072	-0.115687
0.1	0.430343	0.437945	-0.113748

The coefficients apply to the model in the unstandardized form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad (3)$$

Standardized Regression Coefficients

This table shows the standardized regression coefficients at selected values of the ridge parameter:

Standardized Regression Coefficients

Ridge Parameter	triceps	thigh	midarm
0.0	4.2637	-2.9287	-1.56142
0.005	0.902813	0.0669932	-0.27514
0.01	0.674173	0.268411	-0.187032
0.015	0.590258	0.340737	-0.154298
0.02	0.546334	0.377404	-0.136872
0.025	0.519087	0.399209	-0.125836
0.03	0.500378	0.413414	-0.118078
0.035	0.486624	0.423215	-0.112226
0.04	0.476004	0.430237	-0.107583
0.045	0.467489	0.435395	-0.103757
0.05	0.46046	0.439245	-0.100508
0.055	0.454517	0.44214	-0.0976858
0.06	0.449394	0.44432	-0.0951867
0.065	0.444905	0.44595	-0.0929396
0.07	0.440916	0.447148	-0.0908935
0.075	0.437329	0.448003	-0.0890105
0.08	0.43407	0.448579	-0.0872624
0.085	0.431083	0.448928	-0.0856271
0.09	0.428323	0.449088	-0.0840878
0.095	0.425756	0.44909	-0.0826309
0.1	0.423354	0.44896	-0.0812455

The standardized coefficients apply to the model in the standardized form:

$$Y' = \beta'_1 X'_1 + \beta'_2 X'_2 + \beta'_3 X'_3 \tag{4}$$

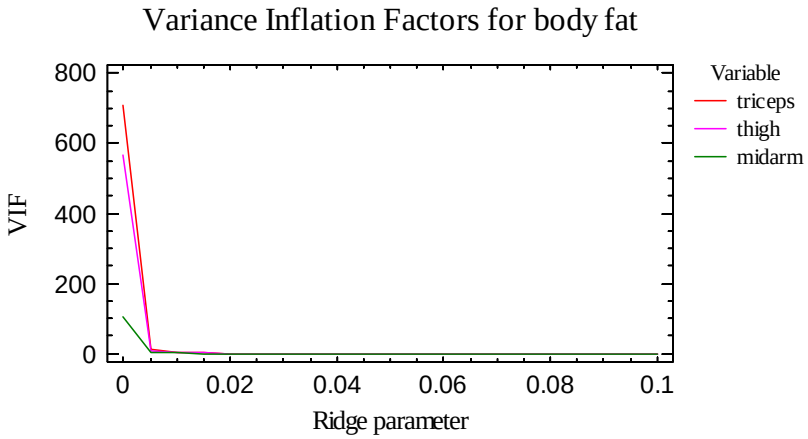
where

$$Y' = \frac{1}{\sqrt{n-1}}\left(\frac{Y - \overline{Y}}{s_Y}\right) \tag{5}$$

$$X'_j = \frac{1}{\sqrt{n-1}}\left(\frac{X_j - \overline{X}_j}{s_{X_j}}\right) \tag{6}$$

Variance Inflation Factors

The Variance Inflation Factors (VIF) measure how large the variance of the coefficients is compared to what it would be if the independent variables were uncorrelated. They are both plotted and tabled:



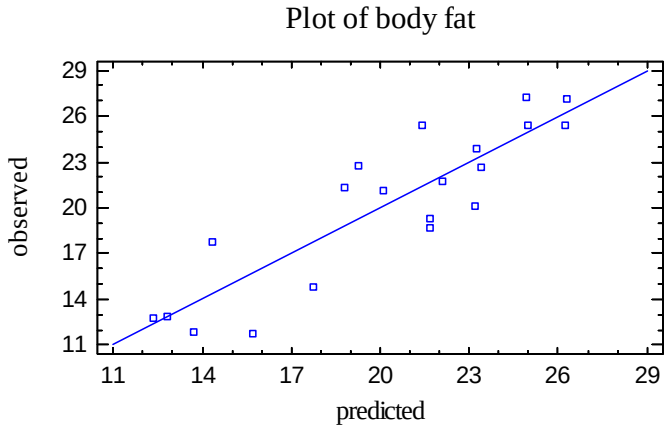
Variance Inflation Factors

Ridge Parameter	triceps	thigh	midarm	R-Squared
0.0	708.843	564.343	104.606	80.14
0.005	11.6434	9.47592	2.57985	78.09
0.01	3.4855	2.98127	1.37703	77.76
0.015	1.74548	1.59433	1.11343	77.50
0.02	1.10255	1.08054	1.01051	77.26
0.025	0.795809	0.834338	0.956922	77.03
0.03	0.625698	0.696905	0.923458	76.81
0.035	0.521438	0.611912	0.899761	76.60
0.04	0.452789	0.555289	0.881403	76.39
0.045	0.405069	0.515354	0.866234	76.18
0.05	0.370454	0.485877	0.853107	75.97
0.055	0.344461	0.463292	0.841362	75.76
0.06	0.324374	0.445435	0.8306	75.56
0.065	0.308468	0.430934	0.820567	75.35
0.07	0.295607	0.418882	0.811093	75.15
0.075	0.285014	0.408663	0.802063	74.95
0.08	0.276148	0.399844	0.793395	74.75
0.085	0.268619	0.392116	0.78503	74.55
0.09	0.262143	0.38525	0.776925	74.36
0.095	0.256506	0.379077	0.769046	74.16
0.1	0.251548	0.373468	0.761368	73.97

Note that the VIF's are also close to 1.0 in the vicinity of $\lambda = 0.02$, which prompted its selection.

Observed versus Predicted

The *Observed versus Predicted* plot shows the observed values of Y on the vertical axis and the predicted values $\hat{Y}$ on the horizontal axis.



If the model fits well, the points should be randomly scattered around the diagonal line. It is sometimes possible to see curvature in this plot, which would indicate the need for a curvilinear model rather than a linear model. Any change in variability from low values of Y to high values of Y might also indicate the need to transform the dependent variable before fitting a model to the data.

Residual Plots

As with all statistical models, it is good practice to examine the residuals. In a regression, the residuals are defined by

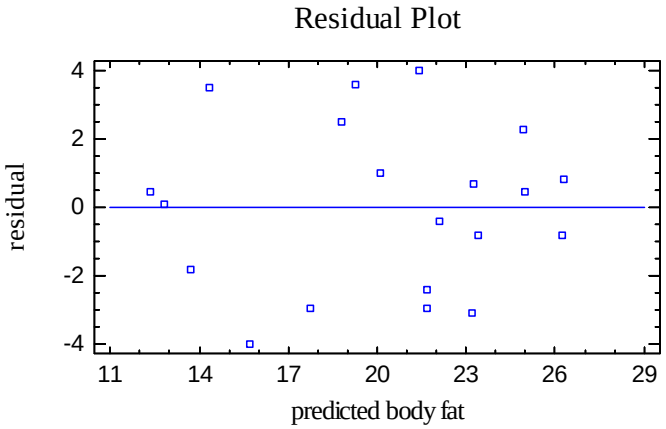
$$e_i = y_i - \hat{y}_i \tag{7}$$

i.e., the residuals are the differences between the observed data values and the fitted model.

The *Ridge Regression* procedure various type of residual plots, depending on *Pane Options*.

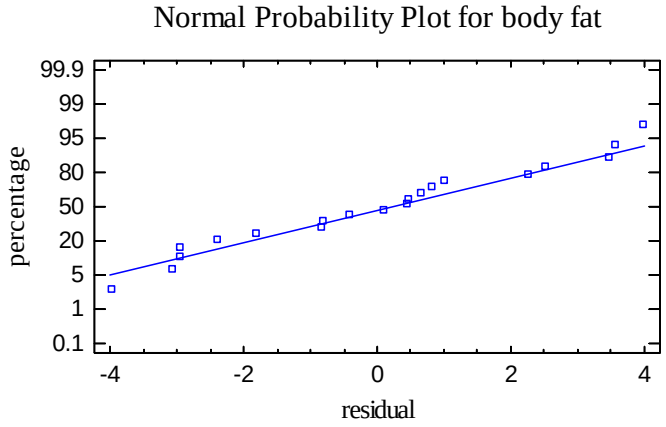
Scatterplot versus Predicted Value

This plot is helpful in visualizing whether the variability is constant or varies according to the magnitude of Y.



Normal Probability Plot

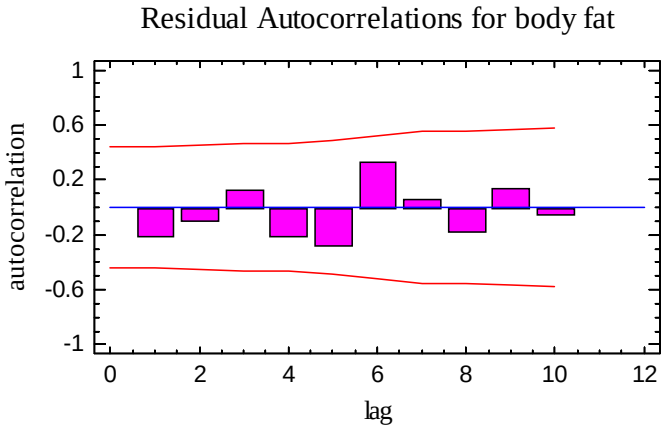
This plot can be used to determine whether or not the deviations around the line follow a normal distribution.



If the deviations follow a normal distribution, they should fall approximately along a straight line.

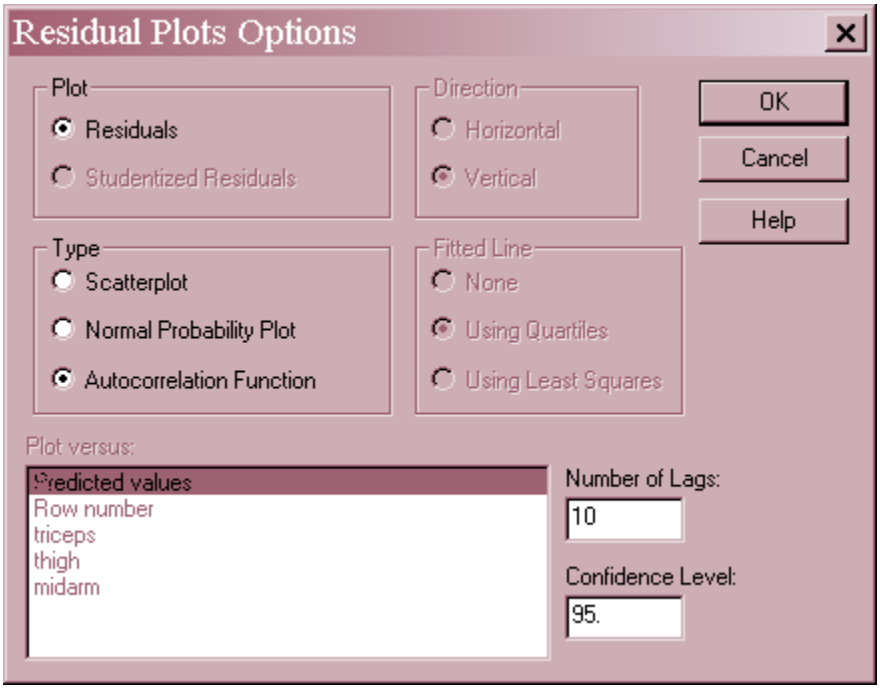
Residual Autocorrelations

This plot calculates the autocorrelation between residuals as a function of the number of rows between them in the datasheet.



It is only relevant if the data have been collected sequentially. Any bars extending beyond the probability limits would indicate significant dependence between residuals separated by the indicated "lag".

Pane Options



- **Plot:** the type of residuals to plot:

    1. *Residuals* – the residuals from the least squares fit.
    2. *Studentized residuals* – the difference between the observed values y_i and the predicted values $\hat{y}_i$ when the model is fit using all observations except the *i-th*, divided by the estimated standard error. These residuals are sometimes called *externally deleted residuals*, since they measure how far each value is from the fitted

model when that model is fit using all of the data except the point being considered. This is important, since a large outlier might otherwise affect the model so much that it would not appear to be unusually far away from the line.

- **Type:** the type of plot to be created. A *Scatterplot* is used to test for curvature. A *Normal Probability Plot* is used to determine whether the model residuals come from a normal distribution. An *Autocorrelation Function* is used to test for dependence between consecutive residuals.

- **Plot Versus**: for a *Scatterplot*, the quantity to plot on the horizontal axis.

- **Number of Lags**: for an *Autocorrelation Function*, the maximum number of lags. For small data sets, the number of lags plotted may be less than this value.

- **Confidence Level:** for an *Autocorrelation Function*, the level used to create the probability limits.

Reports

The *Reports* pane creates predictions using the fitted model. By default, the table includes a line for each row in the datasheet that has complete information on the X variables and a missing value for the Y variable. This allows you to add columns to the bottom of the datasheet corresponding to levels at which you want predictions without affecting the fitted model.

For example, suppose a prediction is desired for a new subject with *triceps* = 30, *thigh* = 55, and *midarm* = 26. The values for each predictor variable would be added to row #21 of the datasheet, leaving the cell for *Body fat* empty. The resulting table is shown below:

Regression Results for body fat

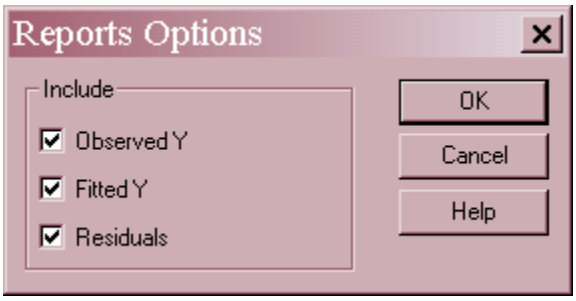
	Observed	*Fitted*	
Row	*Value*	*Value*	*Residual*
1	11.9	13.7166	-1.81664
2	22.8	19.2435	3.55649
3	18.7	21.6624	-2.96241
4	20.1	23.1767	-3.07674
5	12.9	12.8182	0.0817575
6	21.7	22.115	-0.415041
7	27.1	26.2822	0.817792
8	25.4	21.4075	3.99253
9	21.3	18.7945	2.50546
10	19.3	21.7015	-2.40146
11	25.4	24.9562	0.443777
12	27.2	24.9301	2.26994
13	11.7	15.693	-3.99297
14	17.8	14.3285	3.47151
15	12.8	12.3428	0.457156
16	23.9	23.2386	0.661422
17	22.6	23.4134	-0.813431
18	25.4	26.2275	-0.82748
19	14.8	17.7546	-2.95456
20	21.1	20.0971	1.0029
21		24.5228	

Included in the table are:

- **Row** - the row number in the datasheet.
- **Observed Value** – the observed value of Y, if any.
- **Fitted Value** - the predicted value of the dependent variable using the fitted model.
- **Residual** – the observed value minus the predicted value.

For row #21, the predicted *Body fat* equals 24.5.

Pane Options



- **Include**: the columns to be included in the table. If only *Fitted Y* is selected, the table will include only rows with values for all the predictor variables and a missing value for Y.

Save Results

The following may be saved to the datasheet:

1. *Predicted Values* – predicted values from the fitted model for each row of the datasheet.
2. *Residuals* – the *n* residuals.
3. *Coefficients* – the estimated model coefficients.

Calculations

Standardized Model Coefficients

$$\hat{\beta} = \left(X'X + \lambda I\right)^{-1} X'Y \qquad (8)$$

using the transformed variables defined earlier, where *I* is an identity matrix.

