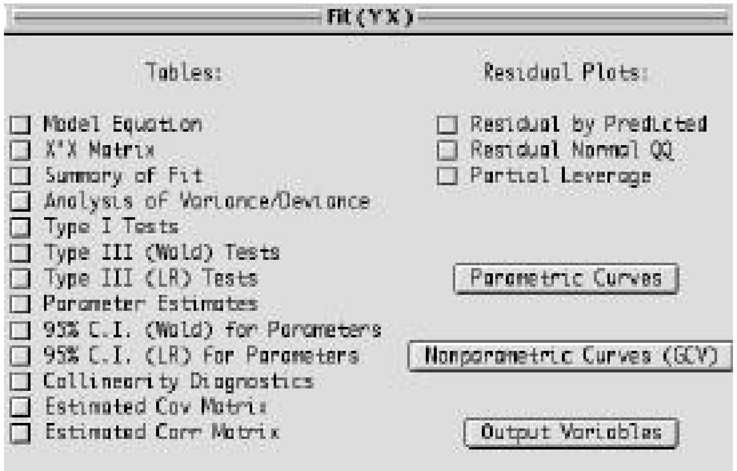


X'X matrix

If model has p variables, or p+1 terms including the interecept term, the X'X (pronounced "X transpose X") matrix is a (p+1) × (p+1) matrix; the entry in a particular row/column is the sum (over all n observations) of the products of the two variables in question. It is not of any great help in and of itself...however, its inverse is central to inferences: the entries of this inverse matrix, multiplied by the MeanSquare Error (Mean Square Residual) provides the estmated variances and covariances of the p+1 parameter estimates.

Type I (Wald) Tests

Tests of sequential incremental improvement as each effect is added to the model (Variables Added in Order"). Order is order in which variables are "clicked" or listed in model.

Type III (Wald) Tests

Tests of partial effects, after the inclusion of the other effects in the model. ("Variables Added Last"). The order in which variables are "clicked" or listed in model does not matter, since the computation is as though the variable in question were the last in the list.

The "Wald" refers to tests based on the Mean Squares. They are the same as the Likelihood Ratio (LR) tests in the case of (measured) Y's analyzed using Gaussian Errors.

Collinearity Diagnostics

(from SAS wording) When an explanatory variable is nearly a linear combination of the other explanatory variables in the model, the affected estimates are unstable and have high standard errors. This problem is called collinearity or multi-collinearity.

Draper and Smith (Applied Regression Analysis, 3rd Edition, page 369) complain that this use of the term collinearity is too loose. To them, there is collinearity when at least one of the X's is linearly depndent on (a linearcombination of) the other X's. They make a distinction between this situation of "exact" collinearity and the "near dependency" in the usage by many modern authors. Unless calculations are programmed very carefully, near dependency (or other ill-conditioned data --- such as having a variable in the model, all of whose values are very close to zero) can create accuracy problems because of the accumulation of rounding errors. Moreover, and more serious statistically, (I'm quoting loosely from

Graybill and Iyer) the presence of multicollinearity, has the following implications: results are highly sensitive to errors in the sample data; resulting parameter estimates cannot be taken seriously, even though overall predictions may be more accurate; and it is not possible with the data at hand to separate the influences of the predictors of the response. this will be reflected in large standarderrors for the individual estimated parameter values Whereas we may be able to find good prediction functions, we have to choose arbitrarily from amomg several sets of nearly equally good prediction functions. Knowledge related to the field of application can often guide us in making a rational selection..

Collinearity Diagnostics...

The **Tolerance** and **Variance Inflation Factor** (VIF) are printed by INSIGHT on same line as each parameter estimate

Tolerance = 1 - the R-square that results from the regression of the the variable in question on the other X variables in the model. If all X variables are orthogonal to each other (ie have zero correlation with each other) then their tolerances are 1. At the other extreme, if a variable is a perfect linear combination of the others, its tolerance is zero.

The **Variance Inflation Factor (VIF)** associated with a particular X variable in a model is the **reciprocal of the Tolerance**. One can think of it as the extra sample size needed to estimate --- to the same precision -- the beta in question, relative to that needed if the X in question were uncorrelated with all linear combinations of the other X's in the model.

VIF = 3 => tol. = 0.33 => variable in question has a multiple R^2 of 0.67 with the other X's in the model.

VIF = 4 => tol. = 0.25 => multiple R^2 of 0.75;

VIF = 5 => tol. = 0.20 => multiple R^2 of 0.80;

VIF=10 => tol. = 0.90 => multiple R^2 of 0.90 etc

But with which other X's ???

Entries in rows with high **Condition Indices** give some clues..

Condition Index (printed if **Collinearity Diagnostics** requested)

Look for rows with condition indices above 100 (Graybill and Iyer say an index above 30 is taken to indicate strong collinearity).

In such rows (each row is a "principal axis"), examine the "Variance Proportion" for each variable: the proportion is the proportion of the variance in the variable that is "explained by" the principal axis. Variables with "Variance Proportions" of say > 0.70 in a row (principal axis) with a high condition index are taken as highly collinear.

Estimated Cov(ariance) Matrix

The estimated covariance matrix of the p+1 parameters; the standard errors (SE's) of the p+1 parameters are the square roots of the p+1 variances given by the diagonal entries; the off diagonal entries are the product of the SE's of the two parameter estimates in question and the correlation of the two estimates. These are useful if one wished to compute the SE of a linear combination of two or more parameter estimates, e.g.

$$SE(b_1 - b_2) \quad = \quad sqrt\{ \ Var[b_1] + Var[b_2] - 2 \ CoVar[b_1 \ , \ b_2] \ \}$$

Estimated Corr(elation) Matrix

The estimated correlation matrix of the p+1 parameters. The p+1 diagonal entries are all 1, reflecting the perfect correlation of an estimate with itself!.; each off diagonal entry is the correlation of the two parameter estimates in question.estimates. These are quite helpful in that they indicate how "separable" are the two parameter estimates, given the the sample size and the degree of collinearity in the "X" data.

Residual plots

Residual by Predicted

Self Explanatory

Residual Normal QQ Plot

Residuals from Gaussian distribution==> plot close to straight line

Plot of Partial Leverage Plots

(Called "Partial Regression Residual plots" if run as an option in PROC REG in SAS Editor Window)

One for each X variable in the model...

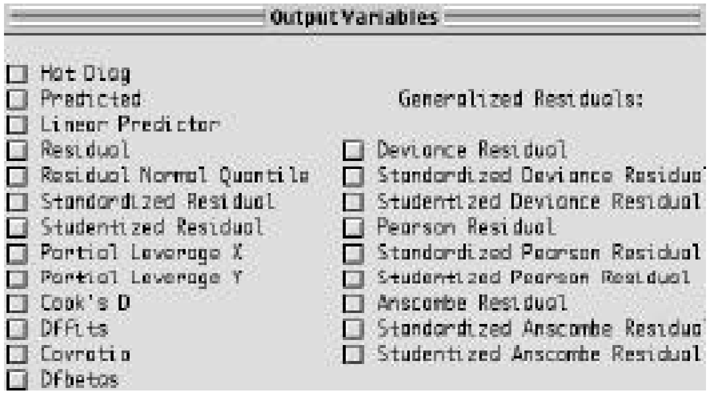
Vertical Axis: Y residual after adjusting Y for other X's in model
i.e. Y - Yhat based on *other* X's

Horizontal Axis: X residual after adjusting X for the other X's in model
i.e. X - Xhat based on *other* X's

Used to assess strength and form of relationship between Y and X after adjusting for other X's [see earlier "Multiple Regression as a series of simple regressions"]

For a particular X, the slope of the regression line of the Partial Y residuals on the partial X residuals is none other than the beta_hat of that X in the multiple regression with this and the other X's.. Also, this is a better way to decide what to do next than to plot the regular Y residuals (from the full model) against each X.

Note that all of the following quantities are at the observation level ... there are as many of them as there are observations.



Output Variables

□ Hat Diag
□ Predicted
□ Linear Predictor
□ Residual
□ Residual Normal Quantile
□ Standardized Residual
□ Studentized Residual
□ Partial Leverage X
□ Partial Leverage Y
□ Cook's D
□ Dffits
□ Covratio
□ Dfbetas

Generalized Residuals:

□ Deviance Residual
□ Standardized Deviance Residua
□ Studentized Deviance Residual
□ Pearson Residual
□ Standardized Pearson Residual
□ Studentized Pearson Residual
□ Anscombe Residual
□ Standardized Anscombe Residua
□ Studentized Anscombe Residual

Hat diagonal

The hat (or leverage) value of an observation measures how *typical* or *atypical* the X values of the observation are. Typical (central) observations have less potential influence on the fitted value; conversely atypical (extreme) observations have more influence. The hat value is a measure of the distance -- in “X space” - between the datapoint and the centroid (centre of gravity) of the X-space.

The hat value (leverage) *is determined solely by the predictor variables (the X’s)* and is not affected by the response variable Y. Thus it is a measure of the potential influence of an observation, and can be calculated as soon as one knows the X values of all of the n observations, and before one knows their associated Y values.

Should worry about/investigate observations with hat values greater than 2 times # parameters / n.

Residual

Observed Y minus Predicted Y

Residual Normal Quantile

The residuals are ranked from smallest (1) to largest (n). if they were from a single Gaussian distribution, one would expect them to be at approximately the (1/)n-th, (2/n)-th, ... (n/n)-th percentiles of the corresponding Gaussian distribution. The Normal Quantile of the i-th ordered residual is computed as the Z value corresponding to the fraction (i-3/8)/(n+1/4) of the distribution. The plot of the observed versus the theoretical (expected) Normal Quantiles should be roughly a straight line.

Standardized Residual

Residual scaled by a measure of its sampling variability. Residuals associated with more extreme X data points have somewhat less sampling variability (since the fitted line or plane is determined more by -- and is thus closer to the Y values of -- the datapoints with extreme X values). This standardization puts all residuals on the same scale. The scaling is achieved by dividing the residual by { RMSE times sqrt[1 - its hat or “leverage” value]}.

Studentized Residual

Since the RMSE in the denominator of the standardized residual was obtained by summinmg all n residuals, including the one in question, the standardized residual has a complicated distribution (for a t-ratio, the numerator and denominator have to be independent). Moreover, if the observation in question is an "outlier" with respect to its Y value, its residual will be large, and will make the RMSE in the denominator of the standardized residual too large, and may make the scaled residual too small to be noticed! To avoid these problems, one can estimate the residual (and the RMSE) from the model that does not include the observation in question. The scaled residual so obtained is called a "studentized" or "studentized deleted" or "jackknife" residual. It is sometimes referred to as an "externally" studentized residual, since the scaling uses a RMSE that is independent of the actual residual in the numerator. In contrast, the standardized residual is sometimes referred to as "internally" studentized, since the RMSE includes the contribution from the residual in the numerator.

Cook's D(istance)

A (standardized) measure of the amount by which the estimates of the beta parameters [or the fitted values] change if the observation in question is deleted froim the analysis. It is an amalgam of the hat value or leverage (potential to influence) and the actual value of the Y residual.

Some authors recommend examining those observations for which Cook's distance is greater than the median (50%) value of an F variable with p and n-p degrees of freedom. See table A-10 in KKMN.

DFFitsS (DiFference in the FITted value -- Standardized)

The influence of an individual observation can also be assessed by examining the amount by which the predicted Y value for the observation in question changes when the observation itself is excluded from the analysis. Again, it is a standardized measure, with authors suggesting that absolute values greater than 2sqrt[p/n] merit attention.

Dfbeta's

(Standardized) measures (1 per term in the model) of the effect of the observation on the estimated regression coefficients. Values above 2 indicate influential observations.