REGRESSION

This procedure performs multiple linear regression with five methods for entry and removal of variables. It also provides extensive analysis of residual and influential cases. Caseweight (CASEWEIGHT) and regression weight (REGWGT) can be specified in the model fitting.

Notation

The following notation is used throughout this chapter unless otherwise stated:

y_i	Dependent variable for case i with variance σ^2 / g_i
c_i	Caseweight for case i; $c_i = 1$ if CASEWEIGHT is not specified
g_i	Regression weight for case i; $g_i = 1$ if REGWGT is not specified
l	Number of distinct cases
w_i	$c_i g_i$
W	$\sum_{i=1}^{l} w_i$
p	Number of independent variables
C	Sum of caseweights: $\sum_{i=1}^{l} c_i$
x_{ki}	The kth independent variable for case i
$\overline{X}_k$	Sample mean for the kth independent variable: $\overline{X}_k = \left(\sum_{i=1}^{l} w_i x_{ki} \right) \Big/ W$
$\overline{Y}$	Sample mean for the dependent variable: $\overline{Y} = \left(\sum_{i=1}^{l} w_i y_i \right) \Big/ W$
h_i	Leverage for case i

$\tilde{h}_i$	$\frac{g_i}{W} + h_i$
S_{kj}	Sample covariance for X_k and X_j
S_{yy}	Sample variance for Y
S_{ky}	Sample covariance for X_k and Y
p^*	Number of coefficients in the model. $p^* = p$ if the intercept is not included; otherwise $p^* = p+1$
R	The sample correlation matrix for $X_1, \ldots, X_p$ and Y

Descriptive Statistics

$$\mathbf{R} = \begin{bmatrix} r_{11} & \cdots & r_{1p} & r_{1y} \\ r_{21} & \cdots & r_{2p} & r_{2y} \\ \cdot & \cdots & \cdot & \cdot \\ r_{y1} & \cdots & r_{yp} & r_{yy} \end{bmatrix}$$

where

$$r_{kj} = \frac{S_{kj}}{\sqrt{S_{kk} S_{jj}}}$$

and

$$r_{yk} = r_{ky} = \frac{S_{ky}}{\sqrt{S_{kk} S_{yy}}}$$

The sample mean $\overline{X}_i$ and covariance S_{ij} are computed by a provisional means algorithm. Define

$$W_k = \sum_{i=1}^{k} w_i = \text{ cumulative weight up to case } k$$

then

$$\overline{X}_{i(k)} = \overline{X}_{i(k-1)} + \left(x_{ik} - \overline{X}_{i(k-1)}\right)\frac{w_k}{W_k}$$

and, if the intercept is included,

$$C_{ij(k)} = C_{ij(k-1)} + \left(x_{ik} - \overline{X}_{i(k-1)}\right)\left(x_{jk} - \overline{X}_{j(k-1)}\right)\left(w_k - \frac{w_k^2}{W_k}\right)$$

Otherwise,

$$C_{ij(k)} = C_{ij(k-1)} + w_k x_{ik} x_{jk}$$

where

$$\overline{X}_{i(1)} = x_{i1}$$

and

$$C_{ij(1)} = 0$$

The sample covariance S_{ij} is computed as the final C_{ij} divided by $C-1$.

Sweep Operations (Dempster, 1969)

For a regression model of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + e_i$$

sweep operations are used to compute the least squares estimates **b** of β and the associated regression statistics. The sweeping starts with the correlation matrix **R**.

Let $\tilde{\mathbf{R}}$ be the new matrix produced by sweeping on the kth row and column of $\mathbf{R}$. The elements of $\tilde{\mathbf{R}}$ are

$$\tilde{r}_{kk} = \frac{1}{r_{kk}}$$

$$\tilde{r}_{ik} = \frac{r_{ik}}{r_{kk}}, \quad i \neq k$$

$$\tilde{r}_{kj} = \frac{r_{kj}}{r_{kk}}, \quad j \neq k$$

and

$$\tilde{r}_{ij} = \frac{r_{ij}r_{kk} - r_{ik}r_{kj}}{r_{kk}}, \quad i \neq k,\ j \neq k$$

If the above sweep operations are repeatedly applied to each row of $\mathbf{R}_{11}$ in

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix}$$

where $\mathbf{R}_{11}$ contains independent variables in the equation at the current step, the result is

$$\tilde{\mathbf{R}} = \begin{pmatrix} \mathbf{R}_{11}^{-1} & -\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \\ \mathbf{R}_{21}\mathbf{R}_{11}^{-1} & \mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \end{pmatrix}$$

The last row of

$$\mathbf{R}_{21}\mathbf{R}_{11}^{-1}$$

contains the standardized coefficients (also called BETA), and

$$\mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$$

can be used to obtain the partial correlations for the variables not in the equation, controlling for the variables already in the equation. Note that this routine is its own inverse; that is, exactly the same operations are performed to remove a variable as to enter a variable.

Variable Selection Criteria

Let r_{ij} be the element in the current swept matrix associated with X_i and X_j. Variables are entered or removed one at a time. X_k is eligible for entry if it is an independent variable not currently in the model with

$r_{kk} \geq t$ (tolerance with a default of 0.0001)

and also, for each variable X_j that is currently in the model,

$$\left(r_{jj} - \frac{r_{jk} r_{kj}}{r_{kk}}\right) t \leq 1$$

The above condition is imposed so that entry of the variable does not reduce the tolerance of variables already in the model to unacceptable levels.

The *F*-to-enter value for X_k is computed as

$$F - to - enter_k = \frac{\left(C - p^* - 1\right) V_k}{r_{yy} - V_k}$$

with 1 and $C - p^* - 1$ degrees of freedom, where p^* is the number of coefficients currently in the model and

$$V_k = \frac{r_{yk} r_{ky}}{r_{kk}}$$

The *F*-to-remove value for X_k is computed as

$$F - to - remove_k = \frac{\left(C - p^*\right) \left|V_k\right|}{r_{yy}}$$

with 1 and $C - p^*$ degrees of freedom.

Methods for Variable Entry and Removal

Five methods for entry and removal of variables are available. The selection process is repeated until the maximum number of steps (MAXSTEP) is reached or no more independent variables qualify for entry or removal. The algorithms for these five methods are described below.

Stepwise

If there are independent variables currently entered in the model, choose X_k such that $F-to-remove_k$ is minimum. X_k is removed if $F-to-remove_k < F_{out}$ (default = 2.71) or, if probability criteria are used, $P(F-to-remove_k) > P_{out}$ (default = 0.1). If the inequality does not hold, no variable is removed from the model.

If there are no independent variables currently entered in the model or if no entered variable is to be removed, choose X_k such that $F-to-enter_k$ is maximum. X_k is entered if $F-to-enter_k > F_{in}$ (default = 3.84) or, $P(F-to-enter_k) < P_{in}$ (default = 0.05). If the inequality does not hold, no variable is entered.

At each step, all eligible variables are considered for removal and entry.

Forward

This procedure is the entry phase of the stepwise procedure.

Backward

This procedure is the removal phase of the stepwise procedure and can be used only after at least one independent variable has been entered in the model.

Enter (Forced Entry)

Choose X_k such that r_{kk} is maximum and enter X_k. Repeat for all variables to be entered.

Remove (Forced Removal)

Choose X_k such that r_{kk} is minimum and remove X_k. Repeat for all variables to be removed.

Statistics

Summary

For the summary statistics, assume *p* independent variables are currently entered in the equation, of which a block of *q* variables have been entered or removed in the current step.

Multiple *R*

$$R = \sqrt{1 - r_{yy}}$$

R Square

$$R^2 = 1 - r_{yy}$$

Adjusted *R* Square

$$R^2_{adj} = R^2 - \frac{\left(1 - R^2\right)p}{C - p^*}$$

R Square Change (when a block of *q* independent variables was added or removed)

$$\Delta R^2 = R^2_{current} - R^2_{previous}$$

F Change and Significance of *F* Change

$$\Delta F = \begin{cases} \dfrac{\Delta R^2 \left(C - p^*\right)}{q\left(1 - R^2_{current}\right)} & \text{for the removal of } q \text{ independent variables} \\ \dfrac{\Delta R^2 \left(C - p^* - q\right)}{q\left(R^2_{previous} - 1\right)} & \text{for the addition of } q \text{ independent variables} \end{cases}$$

the degrees of freedom for the addition are q and $C - p^*$, while the degrees of freedom for the removal are q and $C - p^* - q$.

Residual Sum of Squares

$$SS_e = r_{yy}(C-1)S_{yy}$$

with degrees of freedom $C - p^*$.

Sum of Squares Due to Regression

$$SS_R = R^2(C-1)S_{yy}$$

with degrees of freedom p.

ANOVA Table

Analysis of Variance	*df*	*Sum of Squares*	*Mean Square*
Regression	p	SS_R	$(SS_R)/p$
w	$C-p^*$	SS_e	$(SS_e)/(C-p^*)$

Variance-Covariance Matrix for Unstandardized Regression Coefficient Estimates

A square matrix of size *p* with diagonal elements equal to the variance, the below diagonal elements equal to the covariance, and the above diagonal elements equal to the correlations:

$$\mathrm{var}(b_k) = \frac{r_{kk} r_{yy} S_{yy}}{S_{kk}(C-p^*)}$$

$$\mathrm{cov}(b_k, b_j) = \frac{r_{kj} r_{yy} S_{yy}}{\sqrt{S_{kk} S_{jj}}(C-p^*)}$$

$$\mathrm{cor}(b_k, b_j) = \frac{r_{kj}}{\sqrt{r_{kk} r_{jj}}}$$

Selection Criteria

Akaike Information Criterion (AIC)

$$AIC = C \ln\left(\frac{SS_e}{C}\right) + 2p^*$$

Amemiya's Prediction Criterion (PC)

$$PC = \frac{\left(1 - R^2\right)\left(C + p^*\right)}{C - p^*}$$

Mallow's C_p *(CP)*

$$CP = \frac{SS_e}{\hat{\sigma}^2} + 2p^* - C$$

where $\hat{\sigma}^2$ is the mean square error from fitting the model that includes all the variables in the variable list.

Schwarz Bayesian Criterion (SBC)

$$SBC = C\ln\left(\frac{SS_e}{C}\right) + p^* \ln(C)$$

Collinearity

Variance Inflation Factors

$$VIF_i = \frac{1}{r_{ii}}$$

Tolerance

$$Tolerance_i = r_{ii}$$

Eigenvalues, λ_k

The eigenvalues of scaled and uncentered cross-product matrix for the independent variables in the equation are computed by the QL method (Wilkinson and Reinsch, 1971).

Condition Indices

$$\eta_k = \frac{\max \lambda_j}{\lambda_k}$$

Variance-Decomposition Proportions

Let

$$\mathbf{v}_i = \left(v_{i1}, \ldots, v_{ip}\right)$$

be the eigenvector associated with eigenvalue λ_i. Also, let

$$\Phi_{ij} = v_{ij}^2 \big/ \lambda_i \quad \text{and} \quad \Phi_j = \sum_{i=1}^{p} \Phi_{ij}$$

The variance-decomposition proportion for the *j*th regression coefficient associated with the *i*th component is defined as

$$\pi_{ij} = \Phi_{ij} \big/ \Phi_j$$

Statistics for Variables in the Equation

Regression Coefficient b_k

$$b_k = \frac{r_{yk}\sqrt{S_{yy}}}{\sqrt{S_{kk}}} \quad \textbf{for} \quad k = 1, \ldots, p$$

The standard error of b_k is computed as

$$\hat{\sigma}_{b_k} = \sqrt{\frac{r_{kk} r_{yy} S_{yy}}{S_{kk}\left(C - p^*\right)}}$$

A 95% confidence interval for β_k is constructed from

$$b_k \pm \hat{\sigma}_{b_k} t_{0.025,\, C-p^*}$$

If the model includes the intercept, the intercept is estimated as

$$b_0 = \bar{y} - \sum_{k=1}^{p} b_k \bar{X}_k$$

The variance of b_0 is estimated by

$$\hat{\sigma}^2_{b_0} = \frac{(C-1) r_{yy} S_{yy}}{C\left(C - p^*\right)} + \sum_{k=1}^{p} \bar{X}_k^2 \hat{\sigma}^2_{b_k} + 2 \sum_{k=j+1}^{p} \sum_{j=1}^{p-1} \bar{X}_k \bar{X}_j est.\mathrm{cov}\left(b_k, b_j\right)$$

Beta Coefficients

$$Beta_k = r_{yk}$$

The standard error of $Beta_k$ is estimated by

$$\hat{\sigma}_{Beta_k} = \sqrt{\frac{r_{yy} r_{kk}}{C - p^*}}$$

F-test for $Beta_k$

$$F = \left(\frac{Beta_k}{\hat{\sigma}_{Beta_k}} \right)^2$$

with 1 and $C - p^*$ degrees of freedom.

Part Correlation of X_k with Y

$$Part - Corr(X_k) = \frac{r_{yk}}{\sqrt{r_{kk}}}$$

Partial Correlation of X_k with Y

$$Partial - Corr(X_k) = \frac{r_{yk}}{\sqrt{r_{kk} r_{yy} - r_{yk} r_{ky}}}$$

Statistics for Variables Not in the Equation

Standardized regression coefficient $Beta_k^*$ if X_k enters the equation at the next step

$$Beta_k^* = \frac{r_{yk}}{r_{kk}}$$

The F-test for $Beta_k^*$

$$F = \frac{\left(C - p^* - 1\right) r_{yk}^2}{r_{kk} r_{yy} - r_{yk}^2}$$

with 1 and $C - p^*$ degrees of freedom

Partial Correlation of X_k with Y

$$Partial(X_k) = \frac{r_{yk}}{\sqrt{r_{yy} r_{kk}}}$$

Tolerance of X_k

$$Tolerance_k = r_{kk}$$

Minimum tolerance among variables already in the equation if X_k enters at the next step is

$$\min_{1 \le j \le p} \left(\frac{1}{r_{jj} - \left(r_{kj} r_{jk}\right) / r_{kk}}, r_{kk} \right)$$

Residuals and Associated Statistics

There are 19 temporary variables that can be added to the active system file. These variables can be requested with the RESIDUAL subcommand.

Centered Leverage Values

For all cases, compute

$$h_i = \begin{cases} \dfrac{g_i}{(C-1)} \displaystyle\sum_{j=1}^{p} \sum_{k=1}^{p} \frac{\left(X_{ji} - \overline{X}_j\right)\left(X_{ki} - \overline{X}_k\right) r_{jk}}{\sqrt{S_{jj} S_{kk}}} & \text{if intercept is included} \\ \dfrac{g_i}{(C-1)} \displaystyle\sum_{j=1}^{p} \sum_{k=1}^{p} \frac{X_{ji} X_{ki} r_{jk}}{\sqrt{S_{jj} S_{kk}}} & \text{otherwise} \end{cases}$$

For selected cases, leverage is h_i; for unselected case i with positive caseweight, leverage is

$$h_i' = \begin{cases} g_i\left[\left(\frac{1}{W}+h_i\right)\Big/\left(1+\frac{1}{W}+h_i\right)-\frac{1}{W+1}\right] & \text{if intercept is included} \\ h_i\big/\left(1+h_i/g_i\right) & \text{otherwise} \end{cases}$$

Unstandardized Predicted Values

$$\hat{Y}_i = \begin{cases} \sum_{k=1}^{p} b_k X_{ki} & \text{if no intercept} \\ b_0 + \sum_{k=1}^{p} b_k X_{ki} & \text{otherwise} \end{cases}$$

Unstandardized Residuals

$$e_i = Y_i - \hat{Y}_i$$

Standardized Residuals

$$ZRESID_i = \begin{cases} \frac{e_i}{s} & \text{if no regression weight is specified} \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

where s is the square root of the residual mean square.

Standardized Predicted Values

$$ZPRED_i = \begin{cases} \dfrac{\hat{Y}_i - \overline{Y}}{sd} & \text{if no regression weight is specified} \\ \text{SYSMIS} & \text{otherwise} \end{cases}$$

where *sd* is computed as

$$sd = \sqrt{\sum_{i=1}^{l} \frac{c_i\left(\hat{Y}_i - \overline{Y}\right)^2}{C-1}}$$

Studentized Residuals

$$SRES_i = \begin{cases} \dfrac{e_i / s}{\sqrt{\left(1-\tilde{h}_i\right)/g_i}} & \text{for selected cases with } c_i > 0 \\ \dfrac{e_i / s}{\sqrt{\left(1+\tilde{h}_i\right)/g_i}} & \text{otherwise} \end{cases}$$

Deleted Residuals

$$DRESID_i = \begin{cases} e_i \Big/ \left(1-\tilde{h}_i\right) & \text{for selected cases with } c_i > 0 \\ e_i & \text{otherwise} \end{cases}$$

Studentized Deleted Residuals

$$SDRESID_i = \begin{cases} \dfrac{DRESID_i}{s_{(i)}} & \text{for selected cases with } c_i > 0 \\ \dfrac{e_i}{s\sqrt{\left(1+\tilde{h}_i\right)/g_i}} & \text{otherwise} \end{cases}$$

where $s_{(i)}$ is computed as

$$s_{(i)} = \frac{1}{\sqrt{C-p^*-1}}\sqrt{\frac{\left(C-p^*\right)s^2}{1-\tilde{h}_i} - DRESID_i^2}$$

Adjusted Predicted Values

$$ADJPRED_i = Y_i - DRESID_i$$

DfBeta

$$DFBETA_i = b - b(i) = \frac{g_i e_i \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}_i^t}{1-\tilde{h}_i}$$

where

$$\mathbf{X}_i^t = \begin{cases} \left(1, X_{1i}, \ldots, X_{pi}\right) & \text{if intercept is included} \\ \left(X_{1i}, \ldots, X_{pi}\right) & \text{ottherwise} \end{cases}$$

and $\mathbf{W} = \text{diag}\left(w_1, \ldots, w_l\right)$.

Standardized DfBeta

$$SDBETA_{ij} = \frac{b_j - b_j(i)}{s_{(i)}\sqrt{\left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}_{jj}}}$$

where $b_j - b_j(i)$ is the *j*th component of $\mathbf{b} - \mathbf{b}(i)$.

DfFit

$$DFFIT_i = \mathbf{X}_i\left[\mathbf{b} - \mathbf{b}(i)\right] = \frac{\tilde{h}_i e_i}{1 - \tilde{h}_i}$$

Standardized DfFit

$$SDFIT_i = \frac{DFFIT_i}{s_{(i)}\sqrt{\tilde{h}_i}}$$

Covratio

$$COVRATIO_i = \left(\frac{s_{(i)}}{s}\right)^{2p^*} \times \frac{1}{1 - \tilde{h}_i}$$

Mahalanobis Distance

For selected cases with $c_i > 0$,

$$MAHAL_i = \begin{cases} (C-1)h_i & \text{if intercept is included} \\ C\,h_i & \text{otherwise} \end{cases}$$

For unselected cases with $c_i > 0$

$$MAHAL_i = \begin{cases} C\,h_i' & \text{if intercept is included} \\ (C+1)h_i' & \text{otherwise} \end{cases}$$

Cook's Distance (Cook, 1977)

For selected cases with $c_i > 0$

$$COOK_i = \begin{cases} \left(DRESID_i^2 \tilde{h}_i g_i\right) \Big/ \left[s^2(p+1)\right] & \text{if intercept is included} \\ \left(DRESID_i^2 h_i g_i\right) \Big/ \left(s^2 p\right) & \text{otherwise} \end{cases}$$

For unselected cases with $c_i > 0$

$$COOK_i = \begin{cases} \left(DRESID_i^2 \left(h_i' + \frac{1}{W}\right)\right) \Big/ \left[\tilde{s}^2(p+1)\right] & \text{if intercept is included} \\ \left(DRESID_i^2 h_i'\right) \Big/ \left(\tilde{s}^2 p\right) & \text{otherwise} \end{cases}$$

where h_i' is the leverage for unselected case *i*, and $\tilde{s}^2$ is computed as

$$\tilde{s}^2 = \begin{cases} \frac{1}{C-p}\left[SS_e + e_i^2\left(1 - h_i' - \frac{1}{1+W}\right)\right] & \text{if intercept is included} \\ \frac{1}{C-p+1}\left[SS_e + e_i^2(1-h_i')\right] & \text{otherwise} \end{cases}$$

Standard Errors of the Mean Predicted Values

For all the cases with positive caseweight,

$$SEPRED_i = \begin{cases} s\sqrt{\tilde{h}_i / g_i} & \text{if intercept is included} \\ s\sqrt{h_i / g_i} & \text{otherwise} \end{cases}$$

95% Confidence Interval for Mean Predicted Response

$$LMCIN_i = \hat{Y}_i - t_{0.025,\, C-p^*} SEPRED_i$$

$$UMCIN_i = \hat{Y}_i + t_{0.025,\, C-p^*} SEPRED_i$$

95% Confidence Interval for a Single Observation

$$LICIN_i = \begin{cases} \hat{Y}_i - t_{0.025,\, C-p^*} s\sqrt{\left(\tilde{h}_i + 1\right)/g_i} & \text{if intercept is included} \\ \hat{Y}_i - t_{0.025,\, C-p} s\sqrt{\left(h_i + 1\right)/g_i} & \text{otherwise} \end{cases}$$

$$UICIN_i = \begin{cases} \hat{Y}_i + t_{0.025,\, C-p^*} s\sqrt{\left(\tilde{h}_i + 1\right)/g_i} & \text{if intercept is included} \\ \hat{Y}_i + t_{0.025,\, C-p} s\sqrt{\left(h_i + 1\right)/g_i} & \text{otherwise} \end{cases}$$

Durbin-Watson Statistic

$$DW = \frac{\sum_{i=2}^{l} \left(\tilde{e}_i - \tilde{e}_{i-1}\right)^2}{\sum_{i=1}^{l} c_i \tilde{e}_i^2}$$

where $\tilde{e}_i = e_i \sqrt{g_i}$.

Partial Residual Plots

The scatterplots of the residuals of the dependent variable and an independent variable when both of these variables are regressed on the rest of the independent variables can be requested in the RESIDUAL branch. The algorithm for these residuals is described in Velleman and Welsch (1981).

Missing Values

By default, a case that has a missing value for any variable is deleted from the computation of the correlation matrix on which all consequent computations are based. Users are allowed to change the treatment of cases with missing values.

References

Cook, R. D. 1977. Detection of influential observations in linear regression, *Technometrics*, 19: 15–18.

Dempster, A. P. 1969. *Elements of Continuous Multivariate Analysis.* Reading, Mass.: Addison-Wesley.

Velleman, P. F., and Welsch, R. E. 1981. Efficient computing of regression diagnostics. *The American Statistician*, 35: 234–242.

Wilkinson, J. H., and Reinsch, C. 1971. Linear algebra. In: *Handbook for Automatic Computation*, Volume II, J. H. Wilkinson and C. Reinsch, eds. New York: Springer-Verlag.