Multiple Regression

Influential points

Outliers are data points which lie outside the general linear pattern of which the midline is the regression line. A rule of thumb is that outliers are points whose studentized residual is greater than 2.0. The removal of outliers from the data set under analysis can at times dramatically affect the performance of a regression model. Outliers should be removed if there is reason to believe that other variables not in the model explain why the outlier cases are unusual -- that is, these cases need a separate model. Alternatively, outliers may suggest that additional explanatory variables need to be brought into the model (that is, the model needs respecification).

The following statistics are computed using the option “influence” in the model statement.

```
PROC REG;
MODEL yvar = xvar_1 xvar_2 …xvar_k / influence;
RUN;
```

The **leverage statistic, $\mathbf{h_{ii}}$**, also called the *hat-value*, is available to identify cases which influence the regression model more than others. The $\mathbf{h_{ii}}$ values are computed as the diagonal entries of the matrix $X(X^TX)^{-1}X^T$. The leverage statistic varies from 0 (no influence on the model) to 1 (completely determines the model). A rule of thumb is that cases with leverage under .2 are not a problem, but if a case has leverage over .5, the case has undue leverage and should be examined for the possibility of measurement error or the need to model such cases separately.

Cook's distance, D, is another measure of the influence of a case. Cases with larger D values than the rest of the data are those which have unusual leverage. A cut-off for detecting influential cases is set for values of D greater than 1.

$$D_i = \frac{r_i^2}{p}\frac{V(\hat{y})}{V(e_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$$

where r_i is the studentized residual for i-th observation, h_{ii} is the leverage statistic, p is the number of parameters beta. The value of D_i is a function on how well the model fits the data and how far the point is from the rest of the data.

DFBETAS distance is another measure for influential points. It measures the influence of the i-th observation on the parameters estimates. This measure is obtained by deleting the i-th observation and by refitting the model without this observation. If the parameter estimates vary significantly, then the i-th obs can be considered influential.

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$$

where C_{jj} is the diagonal element of matrix $(X^TX)^{-1}$, $S^2_{(i)}$ is the mean square error of the model fitted without the i-th observation. A cutoff value for detecting influential cases with DFBETAS is $| DFBETAS_{ij}|>2/sqrt(n)$, for sample size n.

DFFITS distance measures the changes in the predictions of y if the i-th observation is eliminated.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{S^2_{(i)} h_{ii}}$$

A cutoff value for detecting influential cases with DFFITS is $| DFFITS_i|>2*sqrt(p/n)$, where n is the sample size and p is the number of parameters.

Studentized residuals are also used to detect outliers with high leverage. The studentized residual is also called the *deleted studentized residual* because its calculation involves leaving out one case in turn for each of the cases. Other terms include *externally studentized residual* or, misleadingly, *standardized residual*. In a plot of studentized residuals, one may draw lines at plus and minus two standard units to highlight cases outside the range where 95% of the cases normally lie.

Remark on cutoff points All cutoff points are only guidelines. Influential points detected by the influential statistics listed above need to be examined individually, and only after a full analysis of the problem we can draw conclusions about outliers and influential points.

Multicollinearity

Multicollinearity is the intercorrelation of independent variables. While simple correlations tell something about multicollinearity, the preferred method of assessing multicollinearity is to regress each independent on all the other independent variables in the equation. Inspection of the correlation matrix reveals only bivariate multicollinearity, for bivariate correlations > .90. To assess multivariate multicollinearity, one uses tolerance or VIF, which build in the regressing of each independent on all the others. Even when multicollinearity is present, note that estimates for other variables in the equation (variables which are not collinear with others) are not affected.

The following multi-collinearity statistics are computed using the option “vif” or “tol” in the model statement.

```
PROC REG;
MODEL yvar = xvar_1 xvar_2 …xvar_k / vif tol;
RUN;
```

Note that a corollary is that very high standard errors of beta coefficients is an indicator of multi-collinearity in the data

- **Tolerance** is $(1 - R_j^2)$ for the regression of X_j on all the other X-variables, ignoring the y-variable. There will be as many tolerance coefficients as there are X-variables. The higher the collinearity of the X-variables, the more the tolerance will approach zero. As a rule of thumb, if tolerance is less than .20, a problem with multicollinearity is indicated.

- **Variance-inflation factor, VIF** *VIF* is the variance inflation factor, which is simply the reciprocal of tolerance. So VIF=1/(1-R_j^2). Therefore, when VIF is high there is high multicollinearity and instability of the beta coefficients. The table below shows the inflationary impact on the standard error of the regression coefficient (beta) of the jth independent variable for various levels of multiple correlation (R_j), tolerance, and VIF (adapted from Fox, 1991: 12). Note that in the "Impact on SE" column, 1.0 corresponds to no impact, 2.0 to doubling the standard error, etc.:

R_j	**Tolerance**	**VIF**	**Impact on SE_b**
0	1	1	1.0
.4	.84	1.19	1.09
.6	.64	1.56	1.25
.75	.44	2.25	1.5
.8	.36	2.78	1.67
.87	.25	4.0	2.0
.9	.19	5.26	2.29

- Standard error is doubled when VIF is 4.0 and tolerance is .25, corresponding to R_j = .87. Therefore VIF >= 4 (or 5) is an arbitrary but common cut-off criterion for deciding when a given independent variable displays "too much" multicollinearity: values above 4 or 5 suggest a multicollinearity problem.