DIAGNOSTIC STATISTICS GENERATED
FOR THE SEGMENT TOTALS TABLE

by
RICK KESTLE

1984

The Segment Totals program in EDITOR was altered recently by Rick Kestle and Martin Ozga. These alterations make use of diagnostic procedures developed by Belsley, Kuh, and Welsch (1980) for the identification of influential points in regression analysis. The result of these alterations is the generation of four diagnostic statistics whose values are printed out in the segment totals table after the regression of classified pixels on reported acres (segment data) takes place. Critical values of these statistics, based on sample size and the nature of the statistics themselves, are also internally generated. Any segment values which result in diagnostic statistics exceeding the critical values are marked with an asterisk (*) in the table.

Care should be taken in drawing conclusions from the results of these diagnostic statistics. Each statistic should be understood in its relationship to the regression model. Diagnostic values exceeding the critical limit for one or more statistics do <u>not</u> necessarily point to outlier data points, but rather to points which are influential in determining the estimated regression coefficients, residuals, etc. Many outlier points, however, will be found, including those which may not be apparent in a plot of the observations or residuals.

A key to understanding these statistics is to remember that they all analyze in some manner the <u>effect</u> on the resulting regression model when the i^{th} data point (segment) is not included. This gives the analyst an idea of the <u>influence</u> exerted by that segment's values.

NOTATION

The DCLC regression model is as follows:

$$\begin{array}{cccccccc} (nx1) & & (nx2) & & (2xl) & & (nxl) & \\ \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} & = & \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} & \cdot & \begin{bmatrix} B_o \\ B_1 \end{bmatrix} & + & \begin{bmatrix} E_1 \\ E_2 \\ \cdot \\ \cdot \\ \cdot \\ E_n \end{bmatrix} & \text{, where} \\ \underline{Y} & & \underline{X} & & \underline{B} & & \underline{E} & \end{array}$$

x_i is the i^{th} row of the $\underline{X}$ matrix - for Application Section's DCLC work

x_i will be a scalar value representing the number of classified pixels in the i^{th}segment,

y_i is the number of reported acres in the i^{th} segment,

$i = 1, 2 \ldots n$, where n is number of segments in the analysis district,

B_o, B_1 are the regression coefficients estimated by b_o and b_1, respectively,

E_i are the errors estimated by the residuals e_i, and

$\underline{Y}$, $\underline{X}$, $\underline{B}$, and $\underline{E}$ are the corresponding matrices.

Also of interest will be s^2, the estimated error variance (MSE) of the regression model population variance δ^2.

In the following sections the four diagnostic statistics are described and defined mathematically with the corresponding critical value provided. Whenever a term is used in the mathematical definition with an (i) designation, it is understood that the i^{th} row of $\underline{X}$ and/or $\underline{Y}$ has been deleted. Thus, $b_{o(i)}$ and $b_{1(i)}$are estimates of the regression coefficients derived when the i^{th} row (segment) of $\underline{X}$ and $\underline{Y}$ have been deleted.

HAT

The hat matrix or least square projection matrix is defined as follows:

$H = \underline{X}\ (X'\ X)^{-1}\ X'$, where $\hat{y} = \underline{X}\ \underline{b} = \underline{H}\ \underline{Y}$.

The diagonal elements of $\underline{H}$ (hat values) are h_i, defined as $x_i\ (\underline{X}'\ \underline{X})^{-1} x_i'$ where $o \leq h_i \leq 1$. Since it is the Hat matrix which determines the predicted values $\hat{\underline{Y}}$ from the response values $\underline{Y}$, the influence of any one response value y_i is most directly reflected in the corresponding predicted value y_i and the influence information is contained in h_i. The h_i also possess a distance interpretation. Large values of h_i reflect outliers -- x_i which are far from $\bar{x}$. The average size of the h_i is 2/n. Critical values (called leverage points) are those which exceed 4/n (n should be > 5 however). The h_i values are printed out in the Segment ~~Table~~ Totals ~~total~~ table under the heading "HAT".

RSTUD

Because var $(e_i) = \delta^2(1-h_i)$, residuals are often "standardized" by dividing e_i by $s\sqrt{1-h_i}$. "Studentized" residuals are similar, with ~~a~~ s replaced by $s_{(i)}$. Thus,

$$\text{"RSTUD"}_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_i}}.$$

Since RSTUD is distributed closely to the t distribution critical values of $|\text{RSTUD}|$ are > 2.0. Influential data points may have small e_i or RSTUD_i , however.

DFFITS

Of interest here is the change in fit which results when the i^{th} segment is deleted. The value DFFITS_i is defined as the studentized difference in fit of the predicted value $\hat{y}_i$ (full model) from the predicted value $\hat{y}_i(i)$ model developed without the i^{th} segment data).

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_i(i)}{\delta\sqrt{h_i}} = \left[\frac{h_i}{1-h_i}\right]^{\frac{1}{2}} \cdot \frac{e_i}{s_{(i)}\sqrt{1-h_i}}, \text{ where } s_{(i)} \text{ estimates } \delta.$$

Since DFFITS does depend on sample size (like h_i but unlike RSTUD), a scaled critical value is $|\text{DFFITS}| \geq 2\sqrt{2/n}$. DFFITS will help detect those points which greatly affect the fit of the model, and thus the form of the model coefficients, but these points should not be deleted solely in order to obtain a more desirable set of estimated coefficients.

COVRAT

This statistic looks at the sensitivity of change in the covariance matrix of the estimated coefficients by looking at the ratio of the determinants of the covariance matrix with all the data to the covariance matrix without the i^th segment. COVRAT is defined as:

$$\text{COVRAT} = \frac{\det\{s^2(i)\,[\underline{X}'(i)\underline{X}(i)]^{-1}\}}{\det\left[s^2\,(\underline{X}'\underline{X})^{-1}\right]} = \frac{1}{\left(\frac{n-3}{n-2} + \frac{\text{RSTUD}_i^2}{n-2}\right)^2 (1-h_i)}.$$

Since these matrices differ only by the inclusion or deletion of the i^th segment, values of COVRAT far from 1.0 indicate the presence of influential points. Critical values are:

$$|\text{COVRAT} - 1.0| > 6/n.$$

FINAL NOTE

All of the definitions and critical values in this document are based on the simple linear regression model used in DCLC projects. In this model $\underline{X}$ is an (nx2) matrix; only 2 coefficients are being estimated. In general, though, $\underline{X}$ will be an nxp matrix. The number of estimated coefficients, p, does appear in nearly every diagnostic statistic definition and critical value, but was replaced by 2.0 in the values reported in this document. Also, the software coded into the EDITOR system assumes the value p=2, and would need to be reprogrammed for other regression models.

REFERENCES

Belsley, D.A., E. Kuh, and R.E. Welsch. Regression Diagnostics. John Wiley and Sons, 1980.

SAS 79.5 Changes and Enhancements. SAS Institute Inc. SAS Technical Report P-115, February 1, 1981.