Additional Regression Diagnostics

Model adequacy : evaluation of independent variables

It is possible to graphically examine “Y on X" after adjusting for another variable. This is called a **partial regression plot**.

The idea is that we are going to examine the residuals of Y, after adjusting for other variable(s). However, in the sweepout EVERYTHING gets adjusted, including the X variable.

Therefore, we plot e_Y in e_X, both adjusted for other variable(s) in the model.

eg. Given the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

now we want to add X_{2i}, and we would like to examine its relation graphically to Y after adjusting for X_1.

Then we get: $e(Y|X_1)$ and $e(X_2|X_1)$

and plot $e(Y|X_1)$ on $e(X_2|X_1)$

SEE HANDOUT

Partial regression plots are not used like residual plots (to examine a fit)

even though I put in a VREF line for comparison

but are used more like raw data plots to examine

a) the potential for a fit, or the need for fitting a variable

b) examine for curvature

More on Identifying outliers : beyond the residual plot

Identifying unusual X values : leverage

We usually think of outliers as being values of Y which are out of line.

However, it may be values of X which are off. In order to determine this we can calculate **leverage**.

$H = X(X'X)^{-1}X'$, recall the HAT MATRIX

uses

$\hat{Y} = HY$

$e = Y - \hat{Y} = (I - H)Y$

$\sigma_e^2 = \sigma^2(I - H)$ gives an individual variance of each e_i since the variance changes along the reg line

in the hat matrix, the vector X is some observed set of X values

eg. $X' = \{1 \quad X_1 \quad X_2 \quad X_3]$

If we take a particular OBSERVED set of X's, call it X_i, and calculate

$H = X_i(X'X)^{-1}X_i'$ this gives a value we will call h_i

likewise, we take the whole n*p X matrix, and get a H matrix (the hat matrix)

$H = X(X'X)^{-1}X'$

then the diagonal elements are those “h" elements, called h_{ii} values

These h_{ii} values have certain properties,

1) the values are between 0 and 1

2) the values sum to p (the number of regressors plus the intercept)

these values are called leverage of the i^{th} case

they provide a relative measure of the distance between the i^{th} case and the mean of all cases

therefore, a large leverage value indicates an “outlier" in that the value is far removed from the mean

furthermore, since $\hat{Y} = HY$, the values of the hat matrix represent weights applied to the Y vector in calculating the predicted values of Y.

How are the h_{ii} values used?

The mean of the h_{ii} values (since they sum to p) is

$$\overline{h}_{ii} = \frac{p}{n}$$ (note that this is < 1)

a leverage value is considered “large" if it is more than twice the value $\overline{h}_{ii}$

ie. $\overline{h}_{ii} > \frac{2p}{n}$ is considered a possible outlier

also, as a general rule, h_{ii} values greater than 0.5 are “large"

while those between 0.2 and 0.5 are moderately large

also look for a leverage value which is noticeably larger than the next largest

SEE HANDOUT

Identifying unusual Y values : outliers

First look at the residuals, this is always a valuable technique.

Studentized residuals

Since we cannot look a an individual residual and determine if it is “too large", we could standardize the residuals to a mean of zero and a variance of 1.

Since the residuals already have a mean of zero, we need only calculate

Studentized residual = $\frac{e_i}{\sqrt{MSE}}$

If normally distributed (which we have already assumed) then

about 65% are between -1 and +1
about 95% are between -2 and +2
about 99% are between -2.5 and +2.5

These are avail in SAS as student

Internally adjusted studentized residuals : instead of adjusting all of the residuals to MSE, we could adjust each residual to its own variance

where $s^2_{e_i} = MSE(1-h_{ii})$ - this is not the same as $s^2_{\hat{Y}}$

$e^*_i = \frac{e_i}{s_{e_i}}$

These values are not provided in SAS, but the individual components e_i and s_{e_i} are provided, so the values could be calculated readily.

!!! SAS gives these (not divided by MSE)

Deleted residuals - Sometimes a value is such a great outlier that it pulls the whole regression line towards itself. It would then be useful to know how far that point lies from a regression line which is fitted with that point excluded. This requires that we fit the regression line WITHOUT each point, and then calculate the deviation of each point.

This sounds like a lot of work, but is readily calculated with the h_{ii} values previously mentioned.

$$d_i = \frac{e_i}{1-h_{ii}} = Y_i - Y_{i(i)}$$

This will identify residuals which pull the line to themselves, and may not be otherwise easily detected.

Further, we could calculate the variance (MSE) adjusted values of the deleted residuals

either values adjusted for the residuals individual variance or

studentized values of all deleted residuals

These studentized deleted residuals values are available in SAS as RStudent

Internally adjusted studentized residuals are not available in SAS, but the components are available for each observation.

Identifying influential cases – an outlier may not have much influence on the regression line, or its influence may be very great.

Once an outlier is found, we ask; “How much influence does the outlier have on the regression line?" If the point were omitted, would the results change?

There are three statistics which measure “influence", all calculated for each observation in the regression line.

How much does the predicted value change?

DFFITS (difference in fits as judged by the predicted values)

$$\text{DFFITS} = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

since the value is standardized by MSE, it represents roughly the number of standard deviation units that the predicted value changes when the particular point is omitted.

How much does a particular regression coefficient change?

DFBETAS (difference in fits as judged by the regression coefficients)

$$\text{DFBETAS} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$ not that this is also a standardized value

interpretation is similar to DFFITS

!!! this value is for standardized b_i in SAS ??? check this

How much do the regression coefficients change overall?

Cook's D (D is for distance)

The boundary of a simultaneous regional confidence region for all regression coefficients can be calculated as;

$$D = \frac{(b-\beta)'X'X(b-\beta)}{pMSE} = F(1-\alpha; p, n-p)$$

This can be modified to determine the effect of removing a point on all of the regression coefficients simultaneously by

$$D_i = \frac{(b-b_{(i)})'X'X(b-b_{(i)})}{pMSE} = F(1-\alpha; p, n-p)$$

this does not follow an F distribution, but it is useful to compare it to the percentiles of the F distribution.

if < 10 or 20 percentile, little effect

if $\geq$ 50 percentile, this is considered large

in practice, we can calculate

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

Suppose you find an influential point which appears to be an outlier.
What do you do?

1) It is an obvious error? If so, correct or delete.

2) What if it is a correct point? It shouldn't be automatically deleted.

Perhaps a transformation is indicated (non-homogeneous variance).

Perhaps the model is wrong (should be curved).