LINEAR REGRESSION ANALYSIS

MODULE – IV

Lecture - 16

Model Adequacy Checking

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Regression variable hull (RVH)

It is the smallest convex set containing all the original data $x_i = (x_{i1}, x_{i2}, ..., x_{ik}),\ i = 1, 2, ..., n.$

The h_{ii} depend on the Euclidian distance of x_i from the centroid and on the density of the points in RVH.

In general, if a point has largest value of h_{ii}, say h_{max}, then it will lie on the boundary of the RVH in a region of the x-space. In such region, where the density of the observations is relatively low. The set of points x (not necessarily the data points used to fit the model) that satisfy $x'(X'X)^{-1}x \le h_{\max}$ is an ellipsoid enclosing all points inside the RVH. So the location of a point, say, $x_0 = (x_{01}, x_{02}, ..., x_{0k}),$ relative to RVH is rejected by $h_{00} = x_0'(X'X)^{-1}x_0.$

Points for which $h_{00} > h_{max}$ are outside the ellipsoid containing RVH. If $h_{00} < h_{max}$ then the point is inside the RVH. Generally, a smaller the value of h_{00} indicates that the point x_0 lies closer to the centroid of the x-space.

Since h_{ii} is a measure of location of the i^{th} point in x-space, the variance of e_i depends on where the point x_i lies. If h_{ii} is small, then $Var(e_i)$ is larger which indicates a poorer fit. So the points near the center of the x-space have poorer least squares fit than the residuals at more remote locations. Violation of model assumptions are more likely at remote points and these violations may be hard to detect from the inspection of ordinary residuals e_i (or the standardized residuals d_i) because their residuals will usually be smaller.

So a logical procedure is to examine the studentized residuals of the form $r_i = \dfrac{e_i}{\sqrt{MS_{res}(1-h_{ii})}}$ in place of e_i (or d_i).

For r_i,

$E(r_i) = 0$

$Var(r_i) = 1$

regardless of the location of x_i when the form of the model is correct.

In many situations, the variance of residuals stabilizes (particularly in large data sets) and there may be little difference between d_i and r_i. In such cases d_i and r_i often convey equivalent information.

However, since any point with a

- large residual and
- large h_{ii}

is potentially highly influential on the least-squares fit, so examination of r_i is generally recommended.

If there is only one explanatory variable then

$$r_i = \frac{e_i}{\sqrt{MS_{res}\left[1-\left(\frac{1}{n}+\frac{(x_i-\overline{x})^2}{s_{xx}}\right)\right]}},\ i=1,2,...,n.$$

- When x_i is close to the midpoint of x-data, i.e., $x_i-\overline{x}$ is small then estimated standard deviation of e_i is large.

- Conversely, when x_i is near the extreme ends of the range of x-data, then $x_i-\overline{x}$ is large and estimated standard deviation of e_i is small.

- When n is really large, the effect of $(x_i-\overline{x})^2$ is relatively small. So in big data sets, r_i may not differ dramatically from d_i.

PRESS residuals

The PRESS residuals are defined as $(y_i - \hat{y}_{(i)})$ where $\hat{y}_{(i)}$ is the fitted value of the i^{th} response based of all the observation except the i^{th} one.

Reason: If y_i is really unusual, then the regression model based on all the observations may be overly influenced by this observations. This could produce a $\hat{y}_i$ that is very similar to y_i and consequently e_i will be small. So it will be difficult to detect any outlier.

If y_i is deleted, then $\hat{y}_{(i)}$ cannot be influenced by that observation, so the resulting residual should be likely to indicate the presence of the outlier.

Procedure

- Delete the i^{th} observation.
- Fit the regression model to remaining (n -1) observations.
- Calculate the predicted value of y_i corresponding to the deleted observation.
- The corresponding prediction error $e_{(i)} = y_i - \hat{y}_{(i)}$.
- Calculate $e_{(i)}$ for each $i = 1,2,\ldots,n$.

These prediction errors are called **PRESS residuals** because they are used in computing the prediction error sum of squares. They are also called as **deleted residuals**.

Now we establish a relationship between e_i and $e_{(i)}$.

Relation between e_i and $e_{(i)}$

Let $b_{(i)}$ be the vector of regression coefficients estimated by with holding the i^{th} observations. Then $b_{(i)} = \left(X_{(i)}^{'} X_{(i)}\right)^{-1} X_{(i)}^{'} y_{(i)}$

where $X_{(i)}$ is the X-matrix without the vector of i^{th} observation and $y_{(i)}$ is the y-vector without the i^{th} observation. Then

$$\begin{aligned} e_{(i)} &= y_i - \hat{y}_{(i)} \\ &= y_i - x_i \hat{b}_{(i)} \\ &= y_i - x_i (X_{(i)}^{'} X_{(i)})^{-1} X_{(i)}^{'} y_{(i)}. \end{aligned}$$

We use the following result in further analysis.

Result: If $X'X$ is a k x k matrix and x be its i^{th} row vector then $(X'X - x'x)$ denotes the $X'X$ - matrix with the i^{th} row withheld. Then

$$\left[X'X - x'x\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1} x'x (X'X)^{-1}}{1 - x(X'X)^{-1} x'}$$

Using this result, we can write

$$\left[X_{(i)}^{'} X_{(i)}\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1} x_i^{'} x_i (X'X)^{-1}}{1 - h_{ii}}$$

where $h_{ii} = x_i (X'X)^{-1} x_i^{'}$.

Then $e_{(i)} = y_i - x_i \left(X'_{(i)} X_{(i)} \right)^{-1} X'_{(i)} y_{(i)}$

$$= y_i - x_i \left[(X'X)^{-1} + \frac{(X'X)^{-1} x_i' x_i (X'X)^{-1}}{1 - h_{ii}} \right] X'_{(i)} y_{(i)}$$

$$= y_i - x_i (X'X)^{-1} X'_{(i)} y_{(i)} - \frac{x_i (X'X)^{-1} x_i' x_i (X'X)^{-1} X'_{(i)} y_{(i)}}{1 - h_{ii}}$$

$$= y_i - x_i (X'X)^{-1} X'_{(i)} y_{(i)} - \frac{h_{ii} x_i (X'X)^{-1} X'_{(i)} y_{(i)}}{1 - h_{ii}}$$

$$= \frac{(1 - h_{ii}) y_i - (1 - h_{ii}) x_i (X'X)^{-1} X'_{(i)} y_{(i)} - h_{ii} x_i (X'X)^{-1} X'_{(i)} y_{(i)}}{1 - h_{ii}}$$

$$= \frac{(1 - h_{ii}) y_i - x_i (X'X)^{-1} X'_{(i)} y_{(i)}}{1 - h_{ii}}.$$

Using $X'y = X'_{(i)} y_{(i)} + x_i' y_i$ (as x_i is $1 \times k$ vector) we can write

$$e_{(i)} = \frac{(1 - h_{ii}) y_i - x_i (X'X)^{-1} (X'y - x_i' y_i)}{1 - h_{ii}}$$

$$= \frac{(1 - h_{ii}) y_i - x_i (X'X)^{-1} X'y + x_i (X'X)^{-1} x_i' y_i}{1 - h_{ii}}$$

$$= \frac{(1 - h_{ii}) y_i - x_i b + h_{ii} y_i}{1 - h_{ii}}$$

$$= \frac{y_i - x_i b}{1 - h_{ii}}$$

$$= \frac{e_i}{1 - h_{ii}}.$$

Looking at the relationship between e_i and $e_{(i)}$, it is clear that calculating the PRESS residuals does not require fitting in different regressions. The $e_{(i)}$'s are just the ordinary residuals weighted according to the diagonal elements h_{ii} of H.

It is possible to calculate the PRESS residuals from the residuals of a single least squares fit to all n observations.

Residuals associated with points for which h_{ii} is large will have large PRESS residuals. Such points will generally be **high influence** points.

Large difference between ordinary residual and PRESS residual indicates a point where the model fits to the data well and a model without that point **predicts** poorly.

Now
$$Var(e_{(i)}) = Var\left(\frac{e_i}{1-h_{ii}}\right)$$
$$= \frac{1}{(1-h_{ii})^2} Var(e_i)$$
$$= \frac{1}{(1-h_{ii})^2} (1-h_{ii})\sigma^2$$
$$= \frac{\sigma^2}{1-h_{ii}}.$$

The **standardized PRESS residual** is $\frac{e_{(i)}}{\sqrt{Var(e_{(i)})}} = \frac{\left(\frac{e_i}{1-h_{ii}}\right)}{\sqrt{\frac{\sigma^2}{(1-h_{ii})}}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$

which is same as the Studentized residuals.

4. *R*-student

The studentized residual r_i is often considered as an outlier diagnostic and MS_{res} is used as an estimate of σ^2 in computing r_i. This is referred to as **internal scaling** of the residuals because MS_{res} is an internally generated estimate of σ^2 obtained from the fitting the model to all n observation .

Another approach is to use an estimate of σ^2 based on a data set with i^{th} observation removed, say $s^2_{(i)}$.

First we derive an expression for $s^2_{(i)}$. Using the identity

$$\left[X'_{(i)}X_{(i)}\right]^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i'x_i(X'X)^{-1}}{1-h_{ii}}.$$

Post multiply both sides by $(X'y - x_i'y_i)$, we get

$$b_{(i)} = b - (X'X)^{-1}x_i'y_i + \frac{(X'X)^{-1}x_i'x_i(X'X)^{-1}(X'y - x_i'y_i)}{1-h_{ii}}$$

$$b - b_{(i)} = (X'X)^{-1}x_i'y_i - \frac{(X'X)^{-1}x_i'x_i\left[b-(X'X)^{-1}x_i'y_i\right]}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})(X'X)^{-1}x_i'y_i - (X'X)^{-1}x_i'x_ib + (X'X)^{-1}x_i'h_{ii}y_i}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i'\left[y_i - x_ib\right]}{1-h_{ii}}$$

$$= \frac{(X'X)^{-1}x_i'e}{1-h_{ii}}$$

$$b_{(i)} = b - \frac{(X'X)^{-1}x_i'e}{1-h_{ii}}.$$

Now consider

$$(n-k-1)s^2_{(i)} = \sum_{j\neq i=1}^{n}(y_j - x_j b_{(i)})^2$$

$$= \sum_{j=1}^{n}\left[y_j - x_j b + \frac{x_j (X'X) x_i' e_i}{1-h_{ii}} \right]^2 - \left(y_i - x_i b + \frac{h_{ii} e_i}{1-h_{ii}} \right)^2$$

$$= \sum_{j=1}^{n}\left[e_j + \frac{h_{ij} e_i}{1-h_{ii}} \right]^2 - \frac{e_i^2}{(1-h_{ii})^2}$$

$$= \sum_{j=1}^{n} e_j^2 + \frac{2e_i}{1-h_{ii}} \sum_{j=1}^{n} e_j h_{ij} + \frac{e_i^2}{(1-h_{ii})^2} \sum_{j=1}^{n} h_{ij}^2 - \frac{e_i}{(1-h_{ii})}$$

$$= \sum_{j=1}^{n} e_j^2 + \frac{h_{ii} e_i^2}{(1-h_{ii})^2} - \frac{e_i^2}{(1-h_{ii})^2}$$

$$= \sum_{j=1}^{n} e_j^2 - \frac{e_i^2}{1-h_{ii}} \quad \text{(using } Hy = H\hat{y},\ \sum_{j=1}^{n} e_j h_{ij} = 0, \sum_{j=1}^{n} h_{ij}^2 = h_{ij} \text{ as } H \text{ is idempotent)}$$

$$= (n-k) MS_{res} - \frac{e_i^2}{1-h_{ii}}.$$

Thus

$$s^2_{(i)} = \frac{1}{n-k-1}\left[(n-k) MS_{res} - \frac{e_i^2}{1-h_{ii}} \right].$$

This estimate of σ^2 is used instead of MS_{res} to produce an **externally studentized residual**, usually called ***R*-student** given by

$$t_i = \frac{e_i}{\sqrt{s^2_{(i)}(1-h_{ii})}}, \; i = 1, 2, ..., n.$$

In many situations, t_i will differ little with r_i. However, if i^{th} observation is influential , then $S^2_{(i)}$ can differ significantly from MS_{res} and the *R* - student statistic will be more sensitive to this point.

Under usual regression assumption, *t* follows a *t*-distribution with (*n – k – 1*) degrees of freedom.