LINEAR REGRESSION ANALYSIS

MODULE – IV

Lecture - 19

Model Adequacy Checking

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

An outlier test based on *R*-student

A common way to model an outlier is the mean shift outlier model.

Suppose we fit a model $y = X\beta + \varepsilon$

when the true model is $y = X\beta + \delta + \varepsilon$

where δ is a $n \times 1$ vector of zeros except for the u^{th} observation which has a value δ_u. Thus

$$\delta = (0, 0, ..., 0, \delta_u, 0, ..., 0).$$

Assume $\varepsilon \sim N(0, \sigma^2 I)$ for both the models we fit. Our objective is to find an appropriate statistic for testing $H_0 : \delta_u = 0$ verus $H_0 : \delta_u \neq 0$ This procedure assumes that we are specifically interested is u^{th} observation, i.e., that we have a priori information that the u^{th} observation may be an outlier.

First we find an appropriate estimate of δ_u. Consider u^{th} residual as its estimate. The $n \times 1$ residual vector is

$$e = [I - H]y = [I - X(X'X)^{-1}X']y.$$

Then

$$\begin{aligned} E(e) &= \bar{H}y \\ &= \bar{H}E(y) \\ &= \bar{H}(X\beta + \delta) \\ &= \bar{H}X\beta + \bar{H}\delta \\ &= 0 + [I - H]\delta \\ &= [I - X(X'X)^{-1}X]\delta. \end{aligned}$$

Thus $E(e_u) = (1 - h_{uu})\delta_u$

$$\Rightarrow \hat{\delta}_u = \frac{e_u}{1 - h_{uu}}$$

is an unbiased estimator of δ_u where h_{uu} is the u^{th} diagonal element of *H.*

It may be observed that $\hat{\delta}_u$ is simply the u^{th} PRESS residual. Further, the covariance matrix of *e* is

$$V(e) = V\left[(I - H)y\right]$$
$$= (I - H)V(y)(I - H)$$
$$= \sigma^2(I - H)$$
$$Var(e_u) = (1 - h_{uu})\sigma^2.$$

So

$$Var(\hat{\delta}_u) = Var\left(\frac{e_u}{1 - h_{uu}}\right)$$
$$= \frac{(1 - h_{uu})\sigma^2}{(1 - h_{uu})^2}$$
$$= \frac{\sigma^2}{1 - h_{uu}}.$$

Also *e* is a linear combination of normally distributed *y*. So *e* is also normally distributed. Thus $\hat{\delta}_u$ is also normally distributed.

Consequently, under $H_0 : \delta_u = 0$,

$$\frac{\left(\frac{e_u}{1-h_{uu}}\right)}{\left(\frac{\sigma}{\sqrt{1-h_{uu}}}\right)} = \frac{e_u}{\sigma\sqrt{1-h_{uu}}} \sim N(0,1).$$

The quantity $\frac{e_u}{\sigma\sqrt{1-h_{uu}}}$ is simply an example of studentized residual. Since σ^2 is unknown and $\frac{MS_{res}}{\sigma^2}$ is a Chi-square random variable, so a candidate test statistic is

$$\frac{e_u}{\sqrt{MS_{res}(1-h_{uu})}}$$

which follows a *t*-distribution if $e = [I - H]y$ and $SS_{res} = y'(I - H)y$ are independent. Since

$$[I - H]\sigma^2 I[I - H] = \sigma^2(I - H) \neq 0$$

so *e* and SS_{res} are not actually independent.

We already have developed $S^2_{(i)}$ which is related to residual mean square in a regression model with i^{th} observation withheld given by

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

This estimate of σ^2 is independent of e_u by the basic independence assumption on random errors. So σ^2 can be replaced by $s^2_{(u)}$ and an appropriate test statistic for the mean shift outlier model is

$$\frac{e_u}{s_{(u)}\sqrt{1-h_{uu}}}$$

which is the externally studentized residual or *R* - student.

Under $H_0 : \delta_u = 0$,

$$\frac{e_u}{s_{(u)}\sqrt{1-h_{uu}}} \sim t(n-k-1)$$

and under $H_0 : \delta_u \neq 0$,

$$\frac{e_u}{s_{(u)}\sqrt{1-h_{uu}}} \sim \text{noncentral } t\left[(n-k-1, \gamma\right],$$

i.e., *t*-distribution with (*n* - k) degrees of freedom and noncentrality parameter

$$\gamma = \frac{\delta_u}{\sigma / (\sqrt{1-h_{uu}})} = \frac{\delta_i \sqrt{1-h_{uu}}}{\sigma}.$$

Note that the power of this test depends on h_{uu}. If we fit an intercept to our model, then $\frac{1}{n} \leq h_{uu} \leq 1$. So maximum power occurs when $h_{uu} = \frac{1}{n}$, i.e., at the center of the data cloud is terms of the $X's$. As $h_{uu} \to 1$, the power goes to 0.

In other words, this test has less ability to detect outliers at the high leverage data points (Note that the concept of leverage point is discussed in later sections).

Test for lack of fit of a regression model

This test for lack of fit of a regression model is based on the assumptions of normality, independence and constant variance which are satisfied. Only the first order or straight line character of the relationship is in doubt . For example, the data in the following scatter plot where the indication is there that straight line fit is not very satisfactory.

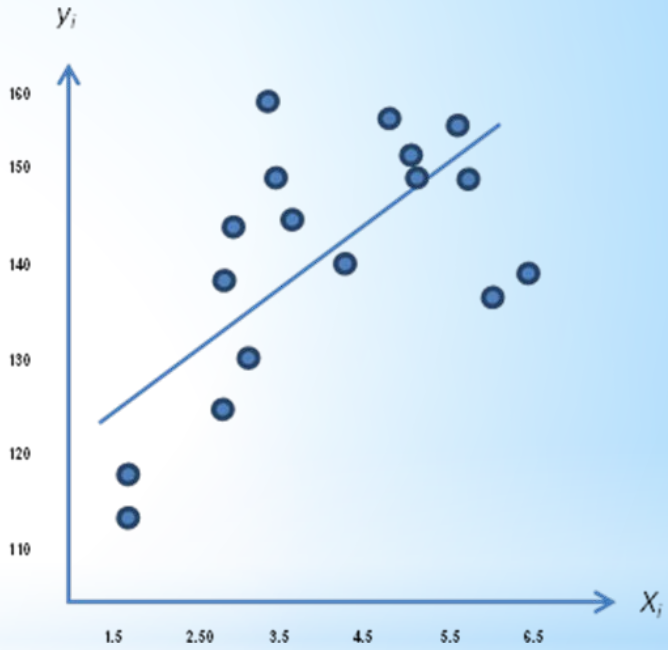
The test procedure determines if systematic curvature is present. The test requires replicate observations on y for at least one level of x and they should be true replications and not just the duplicate readings or measurement of y.

The true replications consists of running n_i separate experiments at $x = x_i$ and observe y. It is not just running a single experiment at $x = x_i$ and measuring y n_i times in which the information only on the variability of the method of measuring y is obtained. These replicated observations are used to obtain a model-independent estimate of σ^2 .

Suppose we have n_i observations on y at the i^{th} level of x_i, $i = 1, 2, ..., m.$

Let y_{ij} be the j^{th} observation on y at x_i, $i = 1, 2, ..., m$, $j = 1, 2, ..., n_i$; $n = \sum_{i=1}^{m} n_i$ is the total number of observations.

Consider the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



Let $\overline{y}_i$ be the mean of n_i observations on x_i. Then the $(i, j)^{th}$ residual is

$$(y_{ij} - \hat{y}_i) = (y_{ij} - \overline{y}_i) + (\overline{y}_i - \hat{y}_i)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 + \sum_{i=1}^{m} n_i(\overline{y}_i - \hat{y}_i)^2$$ (obtained by squaring and summing over i and j)

$$SS_{res} = SS_{PE} + SS_{LOF}$$

SS_{res}	SS_{PE}	SS_{LOF}
Residual sum of squares	Sum of squares due to pure error	Sum of squares due to lack of fit
	Measures pure error	Measures lack of fit

If assumption of constant variance is satisfied, then SS_{PE} is a **model independent measure of pure error** because only the variability of *y's* at each x level is used to compute SS_{PE}.

Since there are $(n_i - 1)$ degrees of freedom for pure error at each level of x_i, the number of degrees of freedom associated with SS_{PE} is $\sum_{i=1}^{m}(n_i - 1) = n - m$. SS_{LOF} is a weighted sum of squared deviations between $\overline{y}_i$ at each level of x and corresponding fitted value.

If $\hat{y}_i$ are close to $\overline{y}_i$, then there is a strong indication that the regression function is linear.

If $\hat{y}_i$ deviate greatly from $\overline{y}_i$ then it is likely that the regression function is not linear. The degrees of freedom associated with SS_{LOF} is $m-2$ because there are m levels of x and two degrees of freedom are lost because two parameters must be estimated to obtain $\overline{y}_i$.

Computationally,

$$SS_{LOF} = SS_{res} - SS_{PE}$$

The test statistic for lack of fit is

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)}$$

$$= \frac{MS_{LOF}}{MS_{PE}}$$

$$E(MS_{LOF}) = \sigma^2 + \frac{\sum_{i=1}^{n} n_i \left[E(y_i) - \beta_0 - \beta_1 x_i\right]^2}{(m-2)}.$$

If true regression is linear, then $E(y_i) = \beta_0 + \beta_1 x_i$ and $E(MS_{LOF}) = \sigma^2$.

If true regression is nonlinear, then $E(y_i) \neq \beta_0 + \beta_1 x_i$ and $E(MS_{LOF}) > \sigma^2$.

If true regression function is linear, then

$$F_0 \sim F(m-2, n-m).$$

So to test for lack of fit, compute F_0 and conclude that regression function is not linear if $F_0 > F_\alpha(m-2, n-m)$ at α level of significance.

If we conclude that regression function is not linear then the tentative model must be abandoned and we attempt to find a more appropriate model. If $F_0 < F_\alpha(m-2, n-m)$ then there is no strong evidence of lack of fit.

The mean sum of squares MS_{PE} and MS_{LOF} are often combined to estimate σ^2.

If *F* ratio for lack of fit is not significant and $H_0 : \beta_1 = 0$ is rejected, then this does not guarantee that model will be satisfactory for prediction. It is suggested that the *F* - ratio must be at least four or five times of $F_\alpha(m-2, n-m)$ if the regression model is to be useful for prediction.

A simple measure of potential prediction performance is found by comparing the range of fitted values, i.e., $(\hat{y}_{\max} - \hat{y}_{\min})$ to their average standard error. Regardless of the term of the model, the average variance of the fitted values is

$$\overline{Var(\hat{y})} = \frac{1}{n}\sum_{i=1}^{n} Var(\hat{y}_i) = \frac{k\sigma^2}{n}$$

where *k* is the number of parameters in the model.

In general, the model is not likely to be satisfactory predictor unless the range of $\hat{y}_i$ is large relative to estimated standard error $\sqrt{\frac{k\hat{\sigma}^2}{n}}$ where $\hat{\sigma}^2$ is a model-independent estimate of error variance.

Estimation of pure error from near-neighbours

In test of lack of fit

$$SS_{res} = SS_{PE} + SS_{LOF}$$

SS_{PE} is computed using responses at repeat observations at some level of *x*. This is model independent estimate of σ^2.

This general principle can be applied to any regression model.

Calculation of SS_{PE} requires repeat observations on the response *y* at the same set of levels on the explanatory variables $x_1, x_2 ..., x_{k,}$, i.e., some of the rows of X - matrix must be same.

In practice, repeat observations do not often occur in multiple regression and the procedure of lack of fit is not often useful.

A method to obtain a model independent estimate of error when there are no exact repeat points are the procedures which search for those points is *x* - space that are near-neighbours.

This is the sets of observations that have been taken with near identical levels of $x_1, x_2, ..., x_k$. The response from such near-neighbours can be considered as repeat points and used to obtain an estimate of pure error.

As a measure of the distance between any two points, $x_1, x_2, ..., x_k$ and $x_{i'1}, x_{i'2}, ..., x_{i'k}$, use weighted sum of squared distance (WSSD)

$$D_{ii'}^2 = \sum_{j=1}^{k} \left[\frac{\hat{\beta}_j (x_{ij} - x_{i'j})}{\sqrt{MS_{res}}} \right]^2 .$$

The pairs of points with small values of $D^2_{ii'}$ are “near neighbours” , i.e., they are relatively close together in *x*-space.

Pairs of points for which $D^2_{ii'}$ is large (e.g., $D^2_{ii'} >> 1$) are widely separated is *x*-space. The residuals at two points with a small value of $D^2_{ii'}$ can be used to obtain an estimate of pure error.

The estimate is obtained from the range of residuals at the points *i* and *i'*, say

$$E_i = \left| e_i - e_{i'} \right| .$$

There is a relationship between the range of a sample from a normal population and the population standard deviation.

For example, for sample size = 2, this relationship is

$$\sigma^2 \equiv \frac{E}{1.128} = 0.886E.$$

The quantity σ^2 so obtained is an estimate of standard deviation of pure error.

An efficient algorithm may be used to compute this estimate like as follows:

- First arrange the data points $x_{ii}, ..., x_{ik}$ in order of increasing $\hat{y}_i$.
- Note that points with different values of $\hat{y}_i$ cannot be near neighbour but those with similar values of $\hat{y}_i$ could be neighbours (or they could be near the same contour of constant $\hat{y}$ but for apart in some *x* -coordinates).

Then

1. Compute the values of $D^2_{ii'}$ for all $(n-1)$ pairs of points with adjacent values of $\hat{y}$. Repeat this calculation for the pairs of points separated by one, two and three intermediate $\hat{y}$ values. This will produce $(4n-10)$ values of $D^2_{ii'}$.
2. Arrange the $(4n-10)$ values of $D^2_{ii'}$ found is step 1. Let $E_u, u=1,2,...,(4n-10)$ be the range of the residuals at these points.
3. For the first m values of E_u, calculate an estimate of the standard deviation of pure error as

$$\hat{\sigma} = \frac{0.886}{m}\sum_{u=1}^{m} E_u.$$

Note that $\hat{\sigma}$ is based on the average range of the residuals associated with the m smallest values of $D^2_{ii'}, m$ must be chosen after inspecting the values of $D^2_{ii'}$ One should not include values of E_u is the calculation for which the weighted sum of squared distance is too large.