Spring 2015 Statistics 151 (Linear Models) : Lecture Twelve

Aditya Guntuboyina

26 February 2015

1 Regression Diagnostics

Regression diagnostics indicate if there is anything wrong with the linear model that has been fitted to the data. They are a must while performing regression analysis. For regression diagnostics, we need to know about the following quantities:

1. Leverage
2. Standardized or Studentized Residuals
3. Predicted Residuals
4. Cook's Distance
5. Externally Standardized or Studentized Residuals

2 Leverage

Let H denote the hat matrix $H = X(X^TX)^{-1}X^T$. The diagonal entries of H, denoted by h_{ii}, are called leverages. In particular, h_{ii} is called the leverage of the ith subject. Leverage has two interpretations:

1. h_{ii} measures the power of the ith subject to influence the regression plane.
2. h_{ii} measures how far the explanatory variable values corresponding to the ith subject are compared to the explanatory variable values of the remaining subjects.

These interpretations can be understood from the discussion below. The leverages h_{ii} have the following properties:

1. $h_{ii} = x_i^T(X^TX)^{-1}x_i$.
2. They lie between 0 and 1 i.e., $0 \leq h_{ii} \leq 1$. This is because $var(\hat{y}_i) = \sigma^2 h_{ii} \geq 0$ and $var(\hat{e}_i) = \sigma^2(1 - h_{ii}) \geq 0$.
3. Their average value is $(p+1)/n$: To see this, note that

$$\sum_{i=1}^{n} h_{ii} = tr(H) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_{p+1}) = p + 1.$$

4. If h_{ii} is large, then $\hat{y}_i$ is close to y_i. To see this, observe that by the fact that H is idempotent, we get
$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$
which implies that $\sum_{j \neq i} h_{ij}^2 = h_{ii}(1 - h_{ii})$. When $h_{ii} \approx 1$, we have $h_{ii}(1 - h_{ii}) \approx 0$ which implies that $h_{ij} \approx 0$ for every $j \neq i$. As a result, we get
$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \approx y_i.$$

The fourth property above means that points with high leverage have the potential to influence the estimated regression equation a great deal. This is the reason h_{ii} is called leverage.

It is important to note that the leverage for the ith subject does not depend on y_i. It only depends on the explanatory variable values.

h_{ii} also measures how far the explanatory variable values corresponding to the ith subject are compared to the explanatory variable values of the remaining objects. To see this, we need to know how to define "far compared to the values of the remaining subjects". This is usually done via the notion of *Mahalanobis Distance.*

Given a set of points (also called a point cloud) $v_1, \dots, v_n$ on the real line and a test point v, can we quantify how far v is from the set? A natural idea is to look at $|v - \bar{v}|^2$ where $\bar{v} = (v_1 + \cdots + v_n)/n$. This makes sense but it does not give a sense of scale. For example if I get $|v - \bar{v}|^2 = 1$, then does this mean that v is far from the point cloud or near it. A better idea is to look at
$$\frac{(v - \bar{v})^2}{s^2} \qquad \text{where } s^2 := \frac{1}{n-1} \sum_{j=1}^{n} (v_j - \bar{v})^2.$$

This easily generalizes to the multivariate case. Suppose $v_1, \dots, v_n$ are all in $\mathbb{R}^p$. Then for a test point $v \in \mathbb{R}^p$, its distance to the point cloud $\{v_1, \dots, v_n\}$ is measured by the *Mahalanobis distance* defined by
$$(v - \bar{v})^T S^{-1} (v - \bar{v}) \qquad \text{where } S := \frac{1}{n-1} \sum_{j=1}^{n} (v_j - \bar{v})(v_j - \bar{v})^T.$$

Let us now see how this applies to the regression model $Y = X\beta + e$. Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one and the rest of the entries carry the values of the explanatory variables for the ith subject. We can write $x_i^T = [1, z_i^T]$ where z_i^T just contains the values of the explanatory variables (without 1) for the ith subject. The Mahalanobis distance for the ith case is defined as
$$\Gamma_i := (z_i - \bar{z})^T S^{-1} (z_i - \bar{z}) \qquad \text{where } S := \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})(z_j - \bar{z})^T \text{ and } \bar{z} = \sum_{j=1}^{n} z_j / n.$$

Γ_i clearly measures how far the ith subject is from the rest of the subjects in terms of the values of the explanatory variables. It turns out that the leverage for the ith subject, h_{ii}, is related to Γ_i by the following simple expression:
$$\Gamma_i = (n-1) h_{ii} - \frac{n-1}{n}.$$
This explains the second interpretation of leverage. For the above formula to hold, it is necessary that there be an intercept term in the model. In other words, this formula does not hold without an intercept term. We will not go over the proof of this formula.