Applied Regression

Multiple Linear Regression Model (MLR Model)

Adequacy

Module 5 Lecture - 5-1

MLR Model

While building up the MLR model to predict, we have made some assumptions and the model is expected to perform well if all the assumptions are correct. Else its performance can be extremely bad. So we should check the following assumptions:

1. The approximate relationship between the response y and the regressors is linear.

2. The errors are normally distributed with mean 0 and constant variance.

3. The errors are uncorrelated.

All of the above assumptions should be tested and effectively corrected whenever necessary. The types of model inadequacies discussed here can have serious consequences. We usually cannot detect departures from the underlying assumptions by examination of the standard summary statistics, such as the t or F statistics, or R^2. These are global model properties, and as such they do not ensure model adequacy.

So the goal here is to understand how to diagnose the violations of the basic regression assumptions. Most of the diagnostic methods are primarily

based on estimated errors.

1. Residual (Error) Analysis:

Residual (error) is estimated as $\hat{e}_i = y_i - \hat{y}_i$.

Almost any deviation from the assumptions on the errors shows up in residuals in different ways. Analysis of the residuals is an effective way to discover several types of model inadequacies. As we will see, plotting residuals is a very effective way to visually investigate how well the regression model fits the data.The residuals have several important properties.

Here we will introduce four popular methods for scaling residuals. These scaled residuals are helpful in finding observations that are outliers, or extreme values.

And Those are

(a) Standardized Residuals

(b) Studentized Residuals

(c) PRESS Residual

(d) R-Student

(a) Standardized Residuals:

As $Var(e) = \sigma^2$ it implies $V\left(\frac{e}{\sigma}\right) = 1$, and sigma is estimated by $\sqrt{MSE}$. Hence a logical scaling for the residuals would be the standardized residuals

$$d_i = \frac{\hat{e}_i}{\sqrt{MSE}}$$

The standardized residuals have mean zero and approximately unit variance. So if a standardized residual is above 3, it is a potential outlier.

Notice that σ^2 is overall assumed variance but it can computed more precisely for every point.

(b) Studentized Residuals:

MSE ia an overall average variation of the errors, but not exact for e_i. We can improve the residual scaling by dividing e_i by the exact standard deviation of the ith residual comes from the fact that

$e = (I - H)y$ where $H = X(X'X)^{-1}X'$ so called the hat matrix.

After simple derivation it is easy to see that $Var(e_i) = \sigma^2(1 - h_{ii})$ where h_{ii} is the ith diagonal element of H. Now note that

(1) Because $0 \leq h_{ii} \leq 1$ $\sqrt{MSE}$ overestimates the variance of e_i.

(2) Since h_{ii} is a measure of the location of the ith point in x space, the variance of e_i depends on where the point x_i lies.

Generally points near the center of the x's is expected to have larger variance as h_{ii} is smaller than the residuals at more remote locations (as h_{ii} is comparatively larger). Violations of model assumptions are more likely at remote points, and these violations may be hard to detect from inspection of the ordinary residuals e_i (or the standardized residuals d_i) because their residuals will usually be smaller.

LEVERAGE POINT and OUTLIER

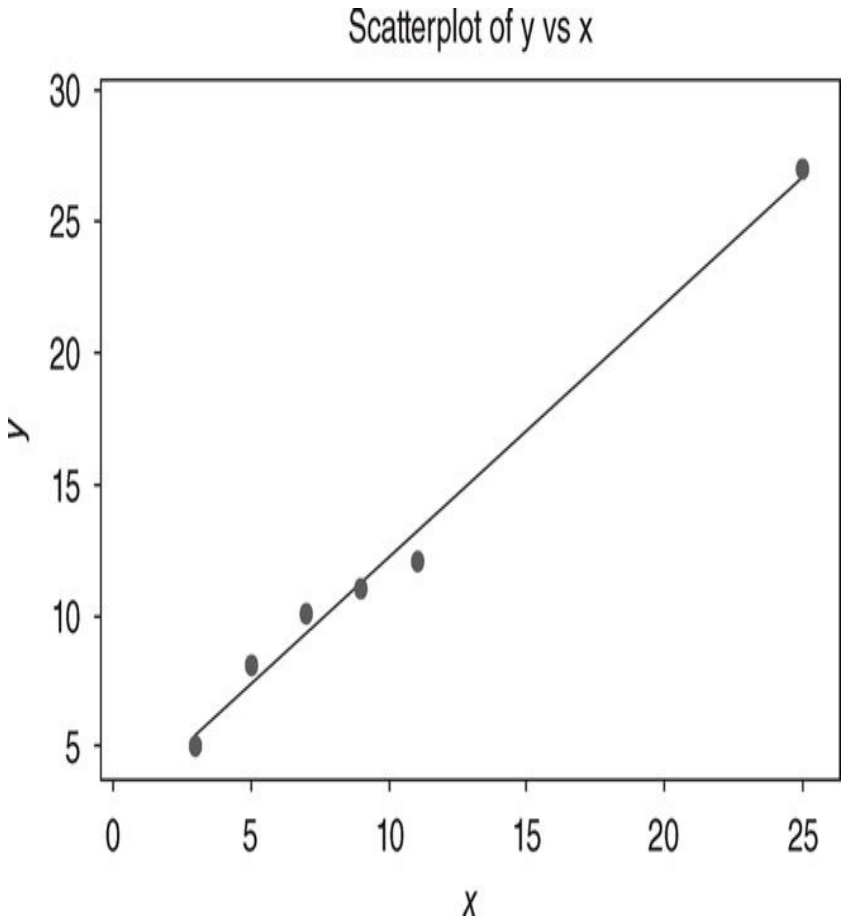

Figure 1: Leverage Point

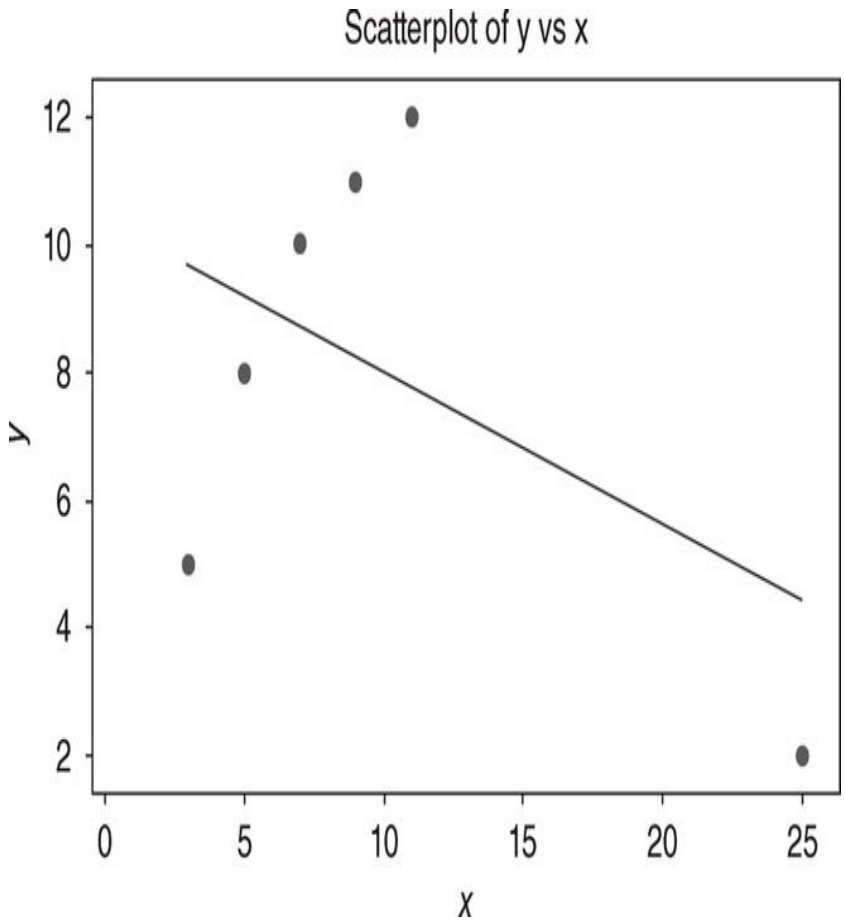

Figure 2: Outlier

Both the figures are drawn with 5 points and their LS-line. Among the five points 4 are common except one at x=25. In Figure-1, the value of the response is 25, and in Figure-2, the value is 2. In both figures x=25 is far

from the center of the other x-values but in Figure-1 observed value for the response is consistent with the prediction based on the other data values. The data point with x = 25 is an example of a pure leverage point. Figure-2 the observed response is not consistent with the values that would be predicted based on only the other data points. Hence it drags the line towards it to reduce the large error and that's why it is called an outlier. Hence the studentized residual is defined as :

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} \quad \text{for} \quad i = 1, 2 \ldots n.$$

$Var(r_i) = 1$ regardless of location of x_i when model is correct. So for high values of r_i, it is suspect for the outliers.

Some of these points are very easy to see by examining the studentized residuals for a simple linear regression model. For SLR model, it is easy to show that the studentized residuals are

$$r_i = \frac{e_i}{\sqrt{MSE\Big[1 - \big(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\big)\Big]}} \quad \text{for} \quad i = 1, 2 \ldots n.$$

(c) PRESS RESIDUAL:

As it is evident that one point with high h_{ii} can swing the estimated regression plane and it may show smaller error. That's why the idea of Press residual is to look at ith point with respect to the MLR model developed by other points except the ith one. However, if the ith observation is deleted, then estimated model cannot be influenced by that observation, so the resulting residual should be likely to indicate the presence of the outlier.

If we delete the ith observation, fit the regression model to the remaining (n - 1) observations, and let $\hat{y}_{(i)}$ is the predicted value corresponding to the deleted observation y_i. Then the corresponding prediction error is

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

This prediction error calculation is repeated for each observation i = 1, 2, , n. These prediction errors are usually called **PRESS** residuals.

Even though it seems that for every point you need to fit a new model with the rest (n-1), but it is not necessary because mathematically it can shown that

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \quad \text{for} \;\; 1 = 1, 2, \ldots, n.$$

Hence it is easy to see that press residuals are high with higher values of h_{ii}.

the variance of the ith PRESS residual is $\quad Var(e_{(i)}) = \dfrac{\sigma^2}{1 - h_{ii}}$

So if we standardize the PRESS residual it becomes

$$\frac{e_{(i)}}{\sqrt{Var(e_{(i)})}} = \frac{e_i/(1 - h_{ii})}{\sqrt{\sigma^2/(1 - h_{ii})}} = \frac{e_i}{\sqrt{\sigma^2(1 - h_{ii})}}$$

Now if we use $\sqrt{MSE}$ to estimate σ^2 and plug it in the above formula it is exactly same as the studentized residual.

(d) R-Student:

It is customary to use MSE as an estimate of σ^2 in computing r_i. But as the concept of PRESS residual is to drop the ith observation it is natural to estimate σ^2 using the remaining (n-1) observation after ith observation removed. It is easy to show that estimate of σ^2 so obtained is

$$\hat{\sigma}^2 = S^2_{(i)} = \frac{(n-p)MSE - e_i^2/(1-h_{ii})}{n-p-1}$$

If we use the above estimate σ^2 instead of MSE it is called R-student, given by

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}, \quad \text{for} \quad i = 1, 2, \ldots, n.$$

It turns out that under the usual regression assumptions, t_i will follow the t_{np1} distribution. Hence if $|t_i| > t_{\alpha/2,n-p-1}$ for certain chosen α then it is called outlier. One could use a Bonferroni-type approach and compare all n values of $|t_i|$ to $t_{\alpha/2n,n-p-1}$ to provide guidance regarding outliers.

Thus, large h_{ii} hat diagonals reveal observations that are potentially influential because they are remote in x space from the rest of the sample. It turns out that the average size of a hat diagonal is $\bar{h} = p/n$ as their sum is is p = Rank of the H. It is customary to identify a point as "remote" or so called leverage point if $h_{ii} > 2 \times \bar{h}$. But all leverage points are not going to be an outlier or influential on the regression coefficients as we have seen in the side-by-side graph before. Because the hat diagonals examine only the location of the observation in x space, it is better to examine the studentized residuals or R-student in conjunction with the h_{ii}. Observations with large hat diagonals and large residuals are likely to be influential.

2. Measure of Influence:

(a) Cook's D

Cook has suggested a way to measure the influence of a point on the estimated values of $\beta's$. To do this, he used the measure of the squared distance between the least-squares estimate based on all n points $\hat{\beta}$ (a p by 1 vector) and the estimate obtained by deleting the ith point $\hat{\beta_{(i)}}$.

This square distance is usually measured in general form as

$$D_i = \frac{(\hat{\beta_{(i)}} - \hat{\beta})' M (\hat{\beta}_{(i)} - \hat{\beta})}{c} \quad \text{for} \quad i = 1, 2, \ldots, n.$$

The usual choices of M and c are $M = X'X$ and c = p MSE. Hence the D_i becomes

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X'X (\hat{\beta}_{(i)} - \hat{\beta})}{pMSE} \quad \text{for} \quad i = 1, 2, \ldots, n.$$

Points with large values of D_i have considerable influence on the least-squares estimates.

Another way to express D_i is

$$D_i = \frac{r_i^2}{p} \frac{Var(\hat{y}_i)}{Var(e_i)} = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}$$

Thus, D_i is made up of two components (1) One that reflects how well the model fits the ith observation y_i and (2) Other that measures how far that point is from the rest of the data.

It also can be shown that $D_i = \frac{(\hat{y}_{(i)} - \hat{y})'(\hat{y}_{(i)} - \hat{y})}{pMSE}$

Therefore, another way to interpret Cook's distance is that it is the squared Euclidean distance (apart from p MSE) that the vector of fitted values moves when the ith observation is deleted.

(b) DFBETAS

Cook's D measure the influence of any point on the estimated β values as a whole. But it does not tell you much about which component got influenced and by how much. Hence a new measure has been proposed to measure the influence of each point on each β_j.

The first of these is a statistic that indicates how much the regression coefficient $\hat{\beta}_j$ changes, in standard deviation units, if the ith observation were deleted. This statistic is

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S^2_{(i)} C_{jj}}}$$

where C_{jj} is the jth diagonal element of $(X'X)^{-1}$ and $\hat{\beta}_{j(i)}$ is the jth regression coefficient computed without use of the ith observation. A large (in magnitude) value of $DFBETAS_{j,i}$ indicates that ith observation has considerable influence on the jth regression coefficient. Notice that $DFBETAS_{j,i}$ is an n by p matrix that conveys similar information to the composite influence information in Cook's distance measure.

In practice if $|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$ then ith observation requires examination.

(c) <u>DFFITS</u>

It is also important to investigate the deletion influence of the ith observation on the predicted or fitted value. It is defined as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}$$

where $\hat{y}_{(i)}$ is the predicted value of y_i without the use of ith observation.

It also can be shown that

$$DFFITS_i = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} \frac{e_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}} = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} t_i$$

Similarly, if $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ then ith observation requires examination.

Example 4.1 (Revisit). Consider the multiple regression model fit to the National Football League team performance data in Example 4.1.

Given $\sqrt{MSE} = 1.7062,\ y_1 = 10,\ \hat{y}_1 = 6.2951,\ h_{11} = 0.0534$

Compute the following quantities and interpret it.

(1) Standardized Residual (2) Studentized Residual (3) PRESS residual
(4) R-Student (5) Cook's D (6) DFFITS

Solution:

Residual $e_1 = y_1 - \hat{y}_1 = 10 - 6.2951 = 3.7049$

(1) Standardized Residual $= d_1 = \frac{\hat{e_1}}{\sqrt{MSE}} = \frac{3.7049}{1.7062} = 2.1714$

Remark: Standardized Residuals have expected value 0 and variance approximately 1. Hence it is a suspect if it is close or above 3. But here it is not.

(2) Studentized Residual $= r_1 = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} = \frac{3.7049}{\sqrt{2.9113(1-0.0534)}} = 2.2318$

Remark: Studentized Residual provides similar information as the previous ones but if leverage is also high along with the residual then it needs more attention. But that's not the case for this point

(3) PRESS Residual $= e_{(i)} = \frac{e_i}{(1-h_{ii})} = \frac{3.7049}{(1-0.0534)} = 3.9141$

Remark: PRESS residual is a leave out type number which is very informative and higher in absolute value is a concern. But sometime it can be higher due to higher value of leverage. Here it has a low value of leverage and yet the PRESS residual is high and hence it needs to closely looked at.

(4) R-Student: To compute R-Student we first need to compute the estimate of the variance without ith observation, Hence

$$\hat{\sigma}^2 = S^2_{(i)} = \frac{(n-p)MSE - e_i^2/(1-h_{ii})}{n-p-1} = \frac{(28-4)\times 2.9113 - 3.7049^2/(1-0.0534)}{28-4-1} = 2.4073$$

Now, $\quad t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}} = \frac{3.7049}{\sqrt{2.4073\times(1-0.0534)}} = 2.4544$

R-Student follow t-distribution with (n - p - 1) degrees of freedom.
Hence at 5% level if it is above the table value the point is called outlier.
As it is 2.4544, it is an outlier.

(5) Cook's D: $D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1-h_{ii}} = \frac{2.2318^2}{4} \times \frac{0.0534}{1-0.0534} = 0.0703$

This measures influence on estimation of the coefficients and the
point is a suspect if it is above 2 which is not the case.

(5) DFFITS: $DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i = \left(\frac{0.0534}{1-0.0534}\right)^{1/2} 2.4544 = 0.5831$

It measure the influence as well and it is only a suspect if it is
higher than $2\sqrt{p/n} = 2 \times \sqrt{4/28} = 0.7559$.
Though it is high but not higher than cutoff.

Residual Analysis Example 4.1

The SAS System

Obs	Y	YHAT	STDERROR_RES	RESIDUAL	LEVERAGE	STUD_RES	PRESS	R_STUDENT	COOKD	DFFITS
1	10	6.2951	1.66003	3.70494	0.05343	2.23185	3.91407	2.45435	0.070291	0.58311
2	11	9.0387	1.60029	1.96135	0.12033	1.22562	2.22964	1.23922	0.051369	0.45833
3	11	8.2710	1.60279	2.72895	0.11758	1.70263	3.09259	1.77759	0.096571	0.64888
4	13	11.3893	1.56415	1.61071	0.15962	1.02977	1.91665	1.03112	0.050354	0.44939
5	10	9.9906	1.53351	0.00939	0.19222	0.00612	0.01163	0.00600	0.000002	0.00292
6	11	11.6557	1.56543	-0.65572	0.15825	-0.41888	-0.77899	-0.41156	0.008246	-0.17845
7	10	11.9040	1.57772	-1.90405	0.14497	-1.20684	-2.22689	-1.21899	0.061738	-0.50195
8	11	10.5202	1.60284	0.47978	0.11753	0.29933	0.54367	0.29357	0.002983	0.10714
9	4	1.9255	1.55041	2.07450	0.17432	1.33803	2.51246	1.36163	0.094494	0.62564
10	2	4.3060	1.59942	-2.30597	0.12130	-1.44176	-2.62429	-1.47681	0.071734	-0.54869
11	7	7.0551	1.51216	-0.05515	0.21456	-0.03647	-0.07021	-0.03570	0.000091	-0.01866
12	10	7.9382	1.64799	2.06178	0.06712	1.25109	2.21011	1.26675	0.028153	0.33978
13	9	9.1365	1.62790	-0.13650	0.08972	-0.08385	-0.14996	-0.08210	0.000173	-0.02577
14	9	9.2582	1.60681	-0.25816	0.11315	-0.16067	-0.29110	-0.15737	0.000823	-0.05621
15	6	8.2197	1.66223	-2.21969	0.05092	-1.33537	-2.33878	-1.35870	0.023918	-0.31471
16	5	3.9499	1.62813	1.05013	0.08946	0.64499	1.15331	0.63695	0.010219	0.19966
17	5	5.2896	1.47028	-0.28955	0.25746	-0.19694	-0.38995	-0.19295	0.003362	-0.11362
18	5	5.4853	1.32951	-0.48529	0.39284	-0.36501	-0.79927	-0.35832	0.021551	-0.28822
19	6	6.1274	1.61218	-0.12736	0.10721	-0.07900	-0.14265	-0.07735	0.000187	-0.02680
20	4	4.3317	1.60647	-0.33168	0.11353	-0.20646	-0.37416	-0.20230	0.001365	-0.07240
21	3	6.0370	1.62410	-3.03697	0.09396	-1.86994	-3.35192	-1.98052	0.090655	-0.63779
22	3	1.7101	1.57833	1.28993	0.14431	0.81727	1.50747	0.81144	0.028162	0.33323
23	4	4.8846	1.60535	-0.88464	0.11476	-0.55106	-0.99932	-0.54290	0.009842	-0.19547
24	10	10.4417	1.59728	-0.44172	0.12364	-0.27654	-0.50404	-0.27115	0.002697	-0.10185
25	6	7.6708	1.64036	-1.67085	0.07573	-1.01859	-1.80775	-1.01942	0.021252	-0.29180
26	8	8.1504	1.59901	-0.15040	0.12174	-0.09406	-0.17124	-0.09209	0.000307	-0.03429
27	2	2.3690	1.40774	-0.36901	0.31928	-0.26213	-0.54209	-0.25698	0.008057	-0.17599
28	0	1.6487	1.57210	-1.64873	0.15105	-1.04875	-1.94209	-1.05103	0.048925	-0.44334