

More on Outlier Diagnostics

Supplement to Section 8.9

Brian Habing – University of South Carolina

Last Updated: July 15, 2004

There are a variety of statistics for detecting outliers available, perhaps the first thought of is the residual ($e_i = y_i - \hat{\mu}_{y|x_i}$). But, as section 8.9 discusses, just because a point has a large residual doesn't mean it is the point we should be most worried about. Further, it is difficult to tell how far off a point is simply looking at the residual because we also need to know the size of the \sqrt{MSE} so that we can compare it appropriately. That is, a residual of 3 is large when the \sqrt{MSE} is 1, but is very small when the \sqrt{MSE} is 100. In this section we examine the hat-diagonal h_i (a measure of leverage, how extreme the x -value is), the externally studentized residual t_i (the residual adjusted for the \sqrt{MSE} , how extreme the y -value is), and the $DFFITs_i$ (a measure of influence, how much the observation actually affects the regression).

Extreme x values – Leverage/Potential and the Hat Diagonal

One way in which a point can be an outlier is for its x -variables to be extreme relative to those of the other observations. This could be troublesome because the point would have the potential (or leverage) to single-handedly cause a large change in the estimated regression line (see the discussion on pages 390-391). Furthermore, an extreme x -value would mean that there is no way to check whether the assumptions for the regression line seem reasonable for the intervening x -values. (Consider a height and weight chart calibrated from twenty individuals between 5'2" and 6'2" with one extra person who is 7'2".)

In simple linear regression it is fairly easy to notice an extreme x -value just by looking at a plot of the values. Numerically we could use the z -score $= (x_i - \bar{x})/s_x$ and a standard rule of thumb about values >2 or >3 being potential outliers. Unfortunately neither of these methods transfers directly for the case with several x -variables. This is because it is very difficult to use plots of more than two-dimensions, and because a point that had a $z=1.5$ in all of the variables would look fairly extreme when all the variables are simultaneously considered.

The quantity that is commonly used to examine x -values in multiple regression can be written as:

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$$

where \mathbf{X} is the matrix seen in Section 8.2, and \mathbf{x}_i^T is the row-vector made up of the i^{th} observations x -values. The h_{ii} then is the i^{th} diagonal element of the "hat matrix" $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. The name "hat matrix" comes from the matrix equation for getting the estimated Y values (the Y -hats) from the observed Y values.

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{H}\mathbf{Y}$$

The reason that these “hat-diagonals” are useful for detecting extreme x -values can be better seen by considering the case of simple linear regression:

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Here we see that the value is closely related to the idea of using the z -score (simply remove the $1/n$ and put a $1/(n-1)$ in the denominator). Some algebra would also show that $0 \leq h_{ii} \leq 1$ and the sum of the $h_{ii} = m+1$.

The observation with the largest h_{ii} can be said to have the most extreme predictor variables, while the observation with the smallest h_{ii} values might be said to be the most typical. As a rule of thumb we say that any observation with an h_{ii} twice the average of all of the h_{ii} is a leverage point, and has the potential to change the model:

$$h_{ii} > 2 \frac{(m+1)}{n}$$

Examining the data in table 8.17 on page shows that observation 8 has the most extreme x -values ($H_{ii} = h_{ii} = 0.7976$) which is greater than $2(3+1)/20 = 0.4$ and would be considered large. Observation 12 on the other hand has the smallest h_{ii} and has the least extreme set of x -values.

	9	Int	Int	Int	Int	Int	Int	Int	Int	Int		
20		Y	X1	X2	X3	R_Y	P_Y	H_Y	RT_Y	F_Y		
■	1	763	19.8	128	86	-17.1641	780.1641	0.2479	-1.0376	-0.5957		
■	2	650	20.9	110	72	-15.5857	665.5857	0.0862	-0.8451	-0.2595		
■	3	554	15.1	95	62	-24.1866	578.1866	0.3462	-1.6460	-1.1979		
■	4	742	19.8	123	82	-6.4881	748.4881	0.1863	-0.3659	-0.1751		
■	5	470	21.4	77	52	6.5247	463.4753	0.3037	0.3980	0.2629		
■	6	651	19.5	107	72	0.3587	650.6413	0.1065	0.0192	0.0066		
■	7	756	25.2	123	84	10.1436	745.8564	0.1328	0.5573	0.2181		
■	8	563	26.2	95	83	-29.3300	592.3300	0.7976	-6.3146	-12.5352		
■	9	681	26.8	116	76	-16.2656	697.2656	0.1204	-0.9018	-0.3337		
■	10	579	28.8	100	64	-15.9962	594.9962	0.2685	-0.9768	-0.5918		
■	11	716	22.0	110	80	42.1602	673.8398	0.1370	2.8554	1.1377		
■	12	650	24.2	107	71	4.8223	645.1777	0.0603	0.2525	0.0639		
■	13	761	24.9	125	81	7.5239	753.4761	0.1398	0.4131	0.1665		
■	14	549	25.6	89	61	14.0446	534.9554	0.1831	0.8037	0.3805		
■	15	641	24.7	103	71	17.9164	623.0836	0.0678	0.9687	0.2612		
■	16	606	26.2	103	67	-11.0784	617.0784	0.1254	-0.6072	-0.2300		
■	17	696	21.0	110	77	24.7172	671.2828	0.0938	1.3979	0.4496		
■	18	795	29.4	133	83	0.0588	794.9412	0.3666	0.0037	0.0028		
■	19	582	21.6	96	65	0.8860	581.1140	0.0883	0.0470	0.0146		
■	20	559	20.0	91	62	6.9385	552.0615	0.1419	0.3811	0.1550		

Extreme y values – Externally Studentized Residuals

An observation can also be extreme because it has extreme y-values. Unfortunately the residual itself is not ideal for examining this. As mentioned above, the major reason for is that the magnitude of the residual depends on the scale of the data (e.g. the \sqrt{MSE}). In particular it would be best to look for outliers by looking for points that were outside of the prediction intervals for that particular x-value (prediction intervals are discussed on pages 304-309). If we calculate the prediction interval using the data point we are worried about we would risk missing some outliers. This is because a point with an extreme y-value would cause the estimated \sqrt{MSE} to be larger, and make it harder to notice the point was an outlier! Therefore we should leave out the observation we are looking at when trying to determine if it is an outlier. The quantity that adjusts the residual in this way is called the externally studentized residual:

$$t_i = \frac{e_i}{\sqrt{MSE_i(1-h_{ii})}}$$

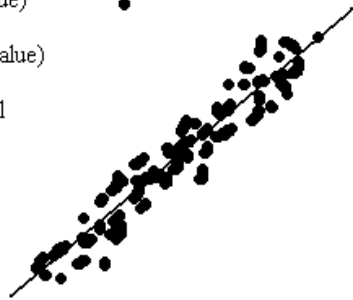
where e_i is the residual for the i^{th} observation and MSE_i is the mean squared error for the regression model if the i^{th} observation is left out.

If the assumptions of the regression model are met and y_i is not an outlier then it will follow a t-distribution with $n-m-2$ degrees of freedom. We could therefore use a t-table to get the exact cut-off values for being extreme, or since the degrees of freedom will usually be quite large we could use the rule of thumb that externally studentized residuals > 2 should be uncommon and those > 3 should be rare. In the example above, observation 11 would be extremely suspicious with $RT_{t_i} = -6.3146$. Observation 11 has a value of 2.8554 and would also be worth a second look. It is also worth noting in this example that what appears most extreme using the externally studentized residuals is different than simply using the raw residuals itself.

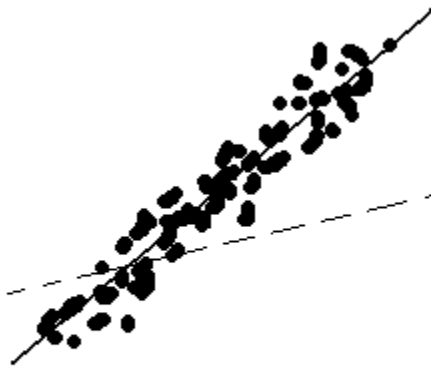
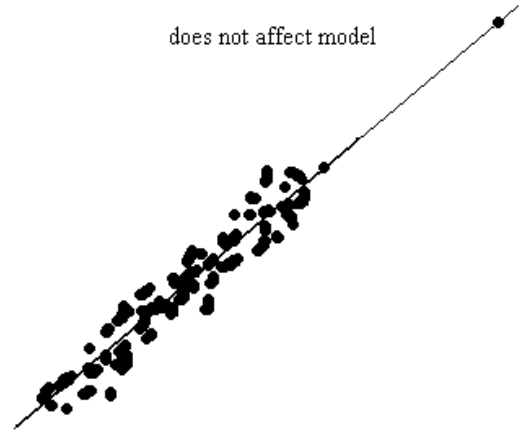
Outliers that Affect the Regression – Influence and DFFITS

The hat-diagonal (h_{ii}) and externally studentized residual (t_i) are useful for detecting potential outliers that probably deserve extra attention. Having an extreme value in one of these two does not necessarily guarantee that the observation actually has any undue influence on the regression. In order to actually affect the estimation the point needs to both have the leverage to influence it and the weight (extreme y-value) to take advantage of that leverage. This is illustrated in the three plots below.

low h_{ii} (middle x value)
 high t_i (extreme y value)
 does not affect model



high h_{ii} (extreme x value)
 low t_i (expected y value)
 does not affect model



high h_{ii} (extreme x value)
 high t_i (extreme y value)
 does affect model



The *DFFITS* statistic combines both the hat-diagonal and externally studentized residual to give a statistic for detecting observations that actually influence the estimated parameters:

$$DFFITS_i = t_i \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2}$$

One rule of thumb is that an observation is influential if the *DFFITS* statistic exceeds $2\sqrt{(m+1)/n}$. In the examples above the *DFFITS* for observations 3, 8, and 11 exceed the cut-off of $2\sqrt{(3+1)/20} \approx 0.894$. If we were to remove the most influential observation (8) from the regression, the parameter estimates would change from

Model Equation											
Y	=	6.3838	-	0.9161	X1	+	5.4090	X2	+	1.1577	X3

Summary of Fit			
Mean of Response	648.2000	R-Square	0.9617
Root MSE	19.1198	Adj R-Sq	0.9545

to

Model Equation												
Y	=	-	42.2676	+	0.9825	X1	+	1.7382	X2	+	6.7386	X3

Summary of Fit			
Mean of Response	652.6842	R-Square	0.9890
Root MSE	10.3243	Adj R-Sq	0.9868

Removing observation 1 (the highest DFFITS not over the cut-off) would only have changed the estimates to:

Model Equation											
Y	=	1.2881	-	1.3434	X1	+	5.4933	X2	+	1.2557	X3

Summary of Fit			
Mean of Response	642.1579	R-Square	0.9606
Root MSE	19.0742	Adj R-Sq	0.9528

One weakness in the DFFITS statistic is that it will not always detect cases where there are two similar outliers. In the third figure above, imagine that there were two points in the bottom right of the graph. Simply removing one would still leave the other there to affect the regression, and so neither point would count as influential by itself. In the numerical example if we were to add an observation 21 that was identical to observation 8 there DFFITS values drop from -12.5252 to -1.0190 .

Removing Outliers

It is important to remember that an outlier cannot simply be erased just because it makes the result of our regression model less clean... in fact many of the great discoveries in science were made by paying attention to the outliers! Unless the observation is clearly in error (a negative distance, a person who was 10' tall, someone spilled soda on the instrument) then most that can be done is to report the results both with and without the outlier (maybe with one of the results in an appendix). The exception to this is the case of extreme x -values. It is possible to reduce the range over which your predictions will be valid. That is, it is ok to say your height and weight relationship is only usable for those between 5'5" and 6'5" for example.