STAT 635 Report:
Identification of multiple high leverage points in logistic regression

Dec.13 2019

Group 7

Dinghao Wang, Penkun Liang

Contents:

Abstract:

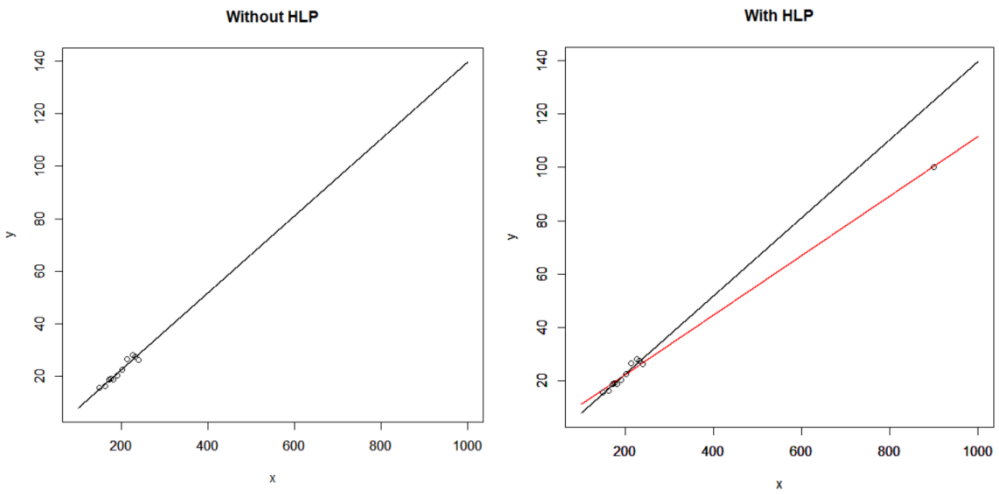
In this paper the authors mainly discuss some commonly used methods and introduce a new modified method to find multiple high leverage points in logistic regression. In their previous research they had developed a method called MDM to identify single leverage point which had been proved to be efficient especially in extreme cases but MDM is not good enough to identify multiple high leverage points, which occurs in most data analysis cases. So they combined the well-developed methods like MCD and MVE to develop their MDM method into new one called MDDM.The usefulness of the proposed method is then investigated through several well-known examples.

Introduction:

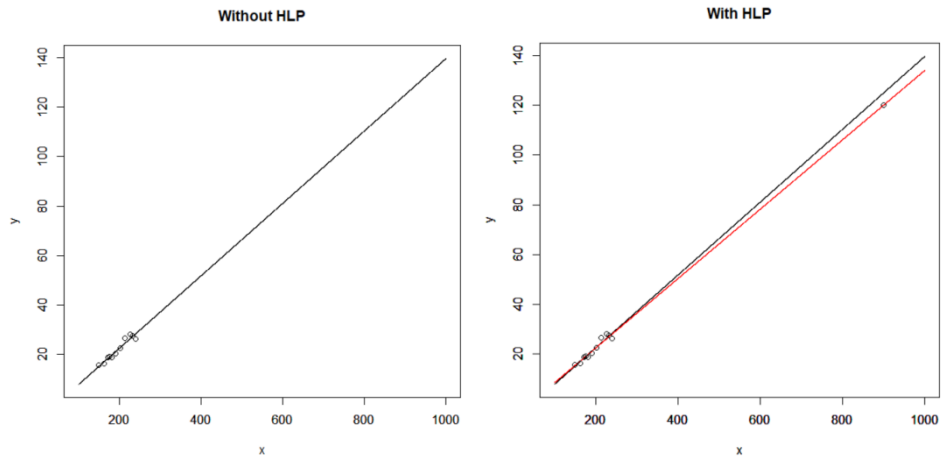
Background:

Regression diagnostics methods have become essential parts of logistic regression in recent years. Outliers, influential observation and high leverage points are discussed together. Generally, a high leverage point is an unusual observation in the X-space.

We can see that in the following graphs:



In this graph, a HLP is added and it has a big influence on the original model. We can see the regression line is greatly changed by one point, which means it’s an influential observation.



In this graph, we still add a HLP in the original model, but it doesn’t have a big influence on the regression line, which means it’s not an influential observation(but still HLP).

There are two definitions:

Masking:

After the deletion of one high leverage point another observation may emerge as a point of extremely high leverage that was not visible at first. This effect is generally known as masking, for which high leverage cases appear as points previously of low leverage.

Swamping:

The opposite effect of masking is known as swamping for which for which low leverage cases are classified as HLP.

We also have the result of logistic regression:

We assume:

$$Y \sim Bernoulli\left(\pi\left(X\right)\right)$$

And do the logistic regression follow the same steps we learned in 635:

$$\pi\left(x_i\right)=\frac{\exp\left(\beta_0+\beta_1X_1+...+\beta_kX_k\right)}{1+\exp\left(\beta_0+\beta_1X_1+...+\beta_kX_k\right)},i=1,2,...n$$

Motivation:

Why We need to find HLP?

We often observe that HLP greatly affect the fitted values and consequently its presence might cause all kind of interpretative problems such as erroneous goodness-of-fit statistics, wrong odds ratios, wrong Wald statistics, etc. So it is important to identify these toxic data point prior to the modeling and eliminate them to ensure better-fitted modeling .

Why we need a new method?

The commonly used leverage measures may fail to identify the HLP in logistic regression.

In linear regression, we can find HLP by the diagonal elements of the weight matrix:

$$W = X\left(X^T X\right)^{-1} X^T$$

And w_i indicate how much the corresponding vector deviates from the bulk of the explanatory variable. The larger value of w_i the more extreme is the corresponding vector of observations of the explanatory variable.

We can check it by the expansion of ith diagonal element:

$$w_i = \frac{1}{n} + \frac{\left(x_{i1} - \bar{x}_{\bullet 1}\right)^2}{\sum\left(x_{i1} - \bar{x}_{\bullet 1}\right)^2} + \frac{\left(x_{i2} - \bar{x}_{\bullet 2}\right)^2}{\sum\left(x_{i2} - \bar{x}_{\bullet 2}\right)^2} + \ldots + \frac{\left(x_{ip} - \bar{x}_{\bullet p}\right)^2}{\sum\left(x_{ip} - \bar{x}_{\bullet p}\right)^2}$$

Similarly, we define a hat matrix for logistic regression:

$$H = V^{1/2} X\left(X^T V X\right)^{-1} X^T V^{1/2}$$

Where V is the diagonal matrix with general elements v_i defined by :

$$v_i = \pi_i\left(1 - \pi_i\right)$$

Then the i-th diagonal element of the matrix H is:

$$h_i = \pi_i\left(1 - \pi_i\right) x_i^T\left(X^T V X\right)^{-1} x_i$$

We also have this property:

$$\sum_{i=1}^{n} w_i = Trace\left(W\right) = k = Trace\left(H\right) = \sum_{i=1}^{n} h_i$$

Proof: We have $tr\left(AB\right) = tr\left(BA\right)$, then

$$tr\left(X\left(X^T X\right)^{-1} X^T\right) = tr\left(X^T X\left(X^T X\right)^{-1}\right)$$

$$= tr\left(I\right) = k$$

Similarly for matrix H.

Methods & Analysis

What are the old methods?

There are not many methods we can use to identify High leverage points since the definition is somehow ambiguous. One generally used method is to calculate the leverage value from hat matrix. Hoaglin and Welsch consider obervations unusual when h_i exceeds $2k/n$ which is also known as **twice-the-mean(2M)** , Vellman and Welsch suggest considering the **thrice-the-mean(3M).** which is $3k/n$ .However in cases with extreme π_i value, this method might fail to identify some HLPs thus lead to wrong Wald Statistics and model fitting.

Since:

$$h_i = \pi_i\left(1-\pi_i\right)x_i^T\left(X^TVX\right)^{-1}x_i$$

$$b_i = x_i^T\left(X^TVX\right)^{-1}x_i$$

If we look at the leverage value as defined above, we will observe that a quantity that does increase with the distance from the mean(DM) is b_i. However, to compute the leverage value, this quantity is multiplied by $\pi_i\left(1-\pi_i\right)$. For an extreme data point in X-space, it is expected that the quantity π_i is close to 0, which means even if b_i is large, h_i could be small. That's why Hosmer and Lemeshow suggest focusing on b_i if we are only interested in measuring the distance.

Imon considered b_i is large if :

$$b_i > Median\left(b_i\right)+3MAD\left(b_i\right)$$

Where

$$MAD\left(b_i\right) = median\left(abs\left(b_i - median\left(b_i\right)\right)\right)/0.6745$$

MAD is a robust statistic, being more resilient to outliers in a data set than the standard deviation .

As far as we are concerned, this is like a 3σ event, since it can be shown that:

$$\lim_{n\to\infty} E\left(MAD\right) = \sigma\Phi^{-1}\left(0.75\right) \approx 0.6745\sigma$$

That's why we have this scale parameter. And this rule is based on the **median of distances from the mean and for this reason call it MDM.**

Up to now, we have 2M, 3M and MDM to detect HLP.

How can we modify the old methods?(new method)

Since we already know there might be Masking and Swamping after we delete single HLP. Mainly because of there two effects, a group deletion version of leverage measure is required for logistic regression.

Let us partition the entire data set into two groups. We assume that d observations among a set of n observations are omitted before the fitting of the model. Here, we denote a set of (n-d) cases 'remaining' in the analysis by R and a set of d cases 'deleted' by D

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \qquad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}$$

And base on X_R and Y_R, we refit the model and have:

$$h_i^{(-D)} = \pi_i^{(-D)}\left(1-\pi_i^{(-D)}\right) x_i^T \left(X_R{}^T V_R X_R\right)^{-1} x_i$$

$$b_i^{(-D)} = x_i^T \left(X_R{}^T V_R X_R\right)^{-1} x_i$$

We also could define $b_i^{(-D)}$ is the deletion distance from the mean(DDM) and :

$$b_i^{(-D)} > Median\left(b_i^{(-D)}\right) + 3MAD\left(b_i^{(-D)}\right)$$

We call it median of deletion distance from the mean.

Now, the author suggest their two-steps method:

Step1:

They employing any suitable robust multivariate technique such as the robust Mahalanobis distance based on minimum volume ellipsoid(MVE) or minimum covariance determinants(MCD) or other methods to identify suspect HLP.

Step2:

After the omission of suspect cases, they refit the model with the rest of the data and the DDM values are computed for the entire dataset. Observations corresponding to DDM values satisfying $b_i^{(-D)} > Median\left(b_i^{(-D)}\right) + 3MAD\left(b_i^{(-D)}\right)$ are declared as HLP.

Why the new method is better?

However, because when we set these rules, we don't know how many HLPs are in the dataset and these rules can be affected by HLP. That's why it's not an easy task to deal with the masking and swamping cases.

1.Cook and Hawkins use MCD and MVE in the detection, but their method identified 6 out of 20 observations as outliers. It's like "outliers everywhere"

2.On the other hand, if we have a group of suspect cases before applying diagnostics.

If we consider the full sample results then the presence of HLP contaminate the entire leverage, and we may not identify the genuine cases at the first place or by successive diagnostics.

The author's new method is a combination of diagnostic and robust approaches.

The advantage is :

By using MCD or MVE, we have a set D contains all potential HLP but it's not the final choice.

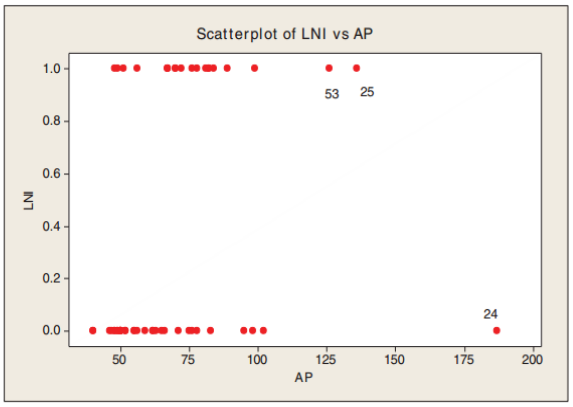
We put back all non-high leverage cases into the estimation so that it might be the entire data set if there is no high leverage cases.

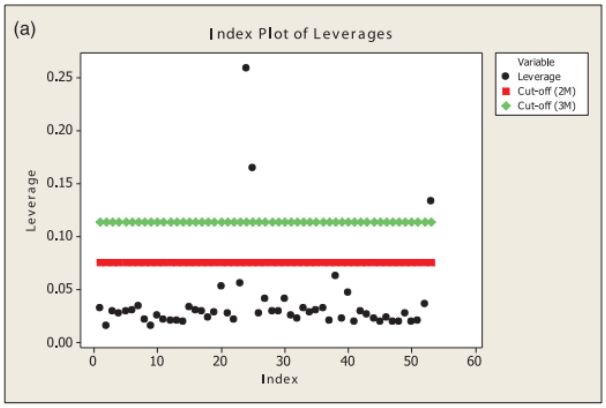
Thus they expect their conclusion should be more accurate than the robust method because they are allowing much more information.

Experiments & Conclusions

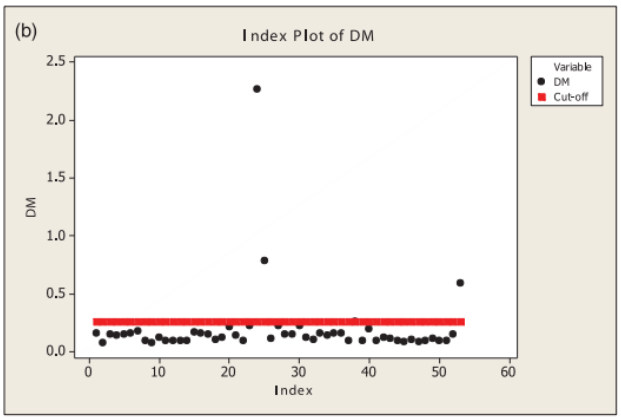
We compared these methods using Brown data set and modified Brown data set.

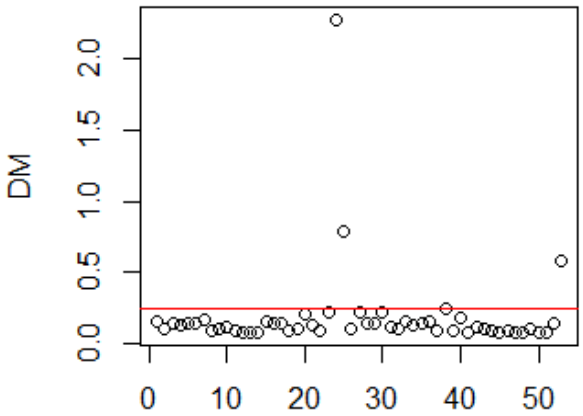
① *Experiment on original Brown dataset*





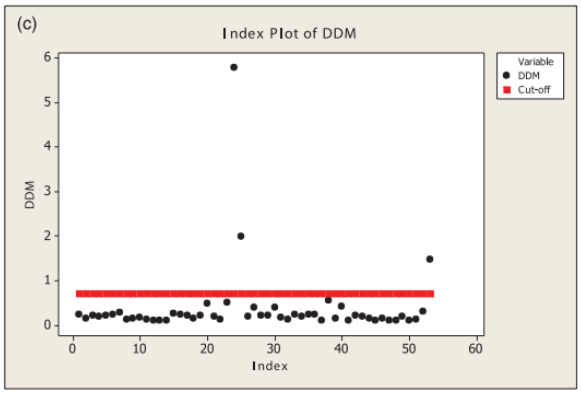
From the scatter plot and 2M&3M method we can easily detect 24,25,53 to be HLPs.

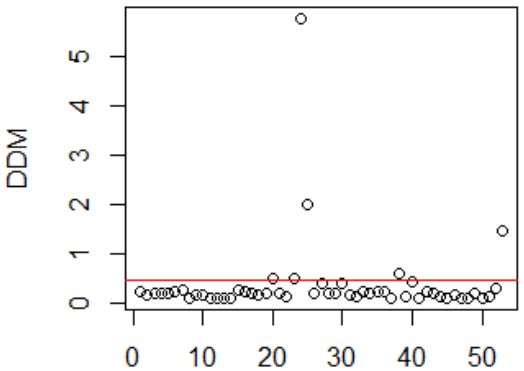




The graph on the left hand side is from the paper that the DM method detect 24,25,53.

The graph on the right hand side is the outcome of our program that we detect 24,25,38,53.



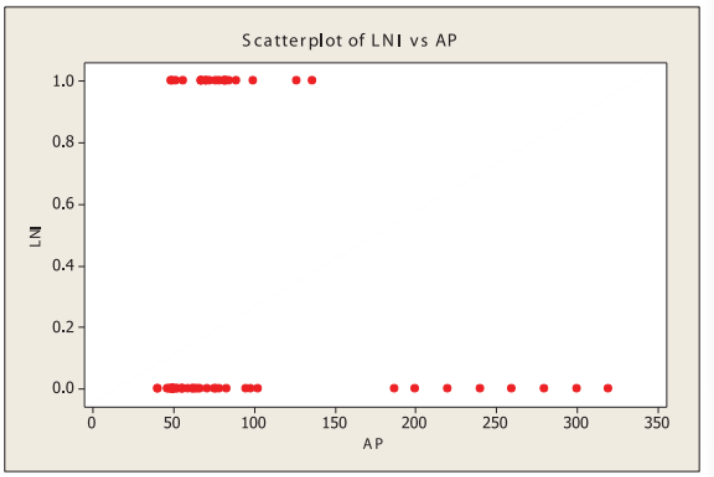


Similarly, we have difference in the outcome of MDDM.

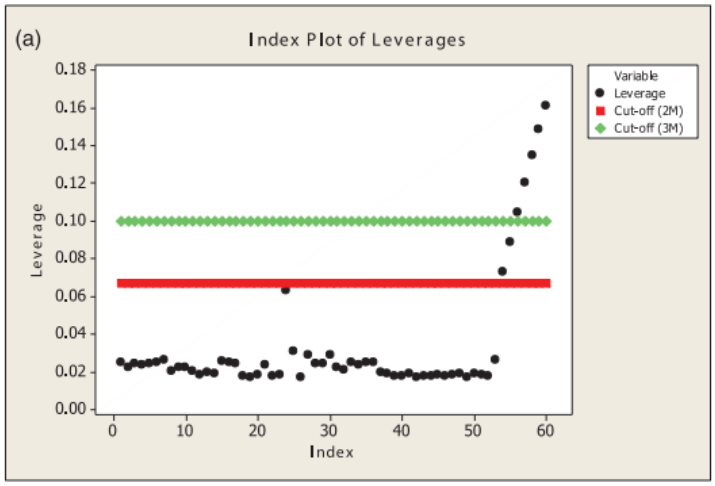
The paper claims 24,25,53 are HLP but we have more suspects 20,23,24,25,38,53 after deletion in MDDM.

② *Experiment on modified Brown dataset*

In this experiment we insert seven more data points with index 54-60 and let them be HLP by setting the AP value high as 200,220,240,260,280,300,320 and LNI all 0.



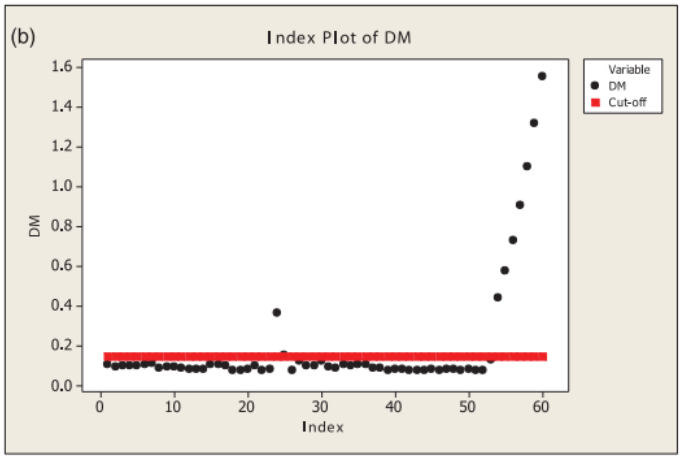
In this case we can hardly tell which points are HLP by scatter plot so we need more accurate methods.

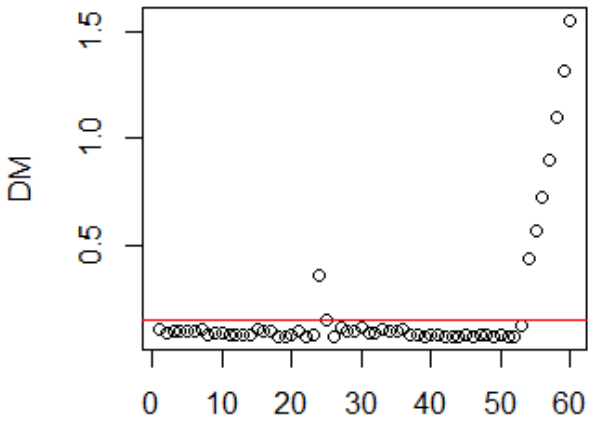


2M method detect: 54 ,55, 56 ,57, 58 ,59 ,60, 3M method detect: 56, 57, 58, 59, 60.

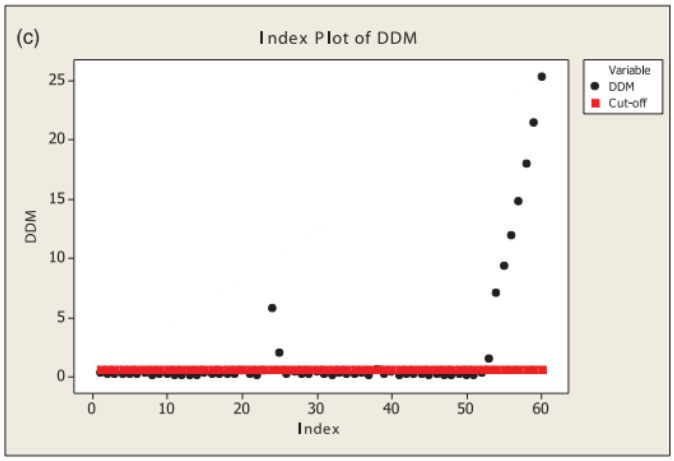
Now we can observe the restriction of 2M&3M as mentioned that they might fail to identify HLP

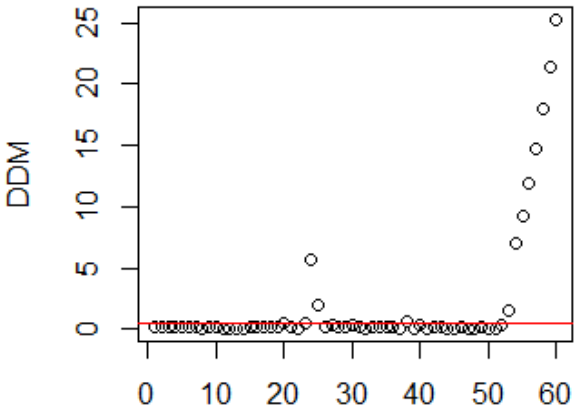
with extreme π_i value such as 54,55 even though they are not 'extreme' enough.





As for MDM method, we have the exactly same outcome as the paper that we identify 24,25,54,55,56,57,58,59,60 including all inserted HLP and previous 24 and 25.





As for MDDM method, we have the almost the same outcome as the paper that it identifies all inserted HLP and previous 24,25 and 53 but we have got one more which is 38.

Conclusions

In the experiment on modified Brown data set, we can easily observe the influence of extreme π_i value that the leverage values of NO.54 and NO.55 could be not significantly large enough to be identified as HLP since their π_i values are relatively small, which is the main reason why we need to develop new method other than 2M&3M base on leverage value. The outcomes also

suggests that the MDM and MDDM perform better while 2M and 3M fail to identify some manually added HLP and previous HLP.

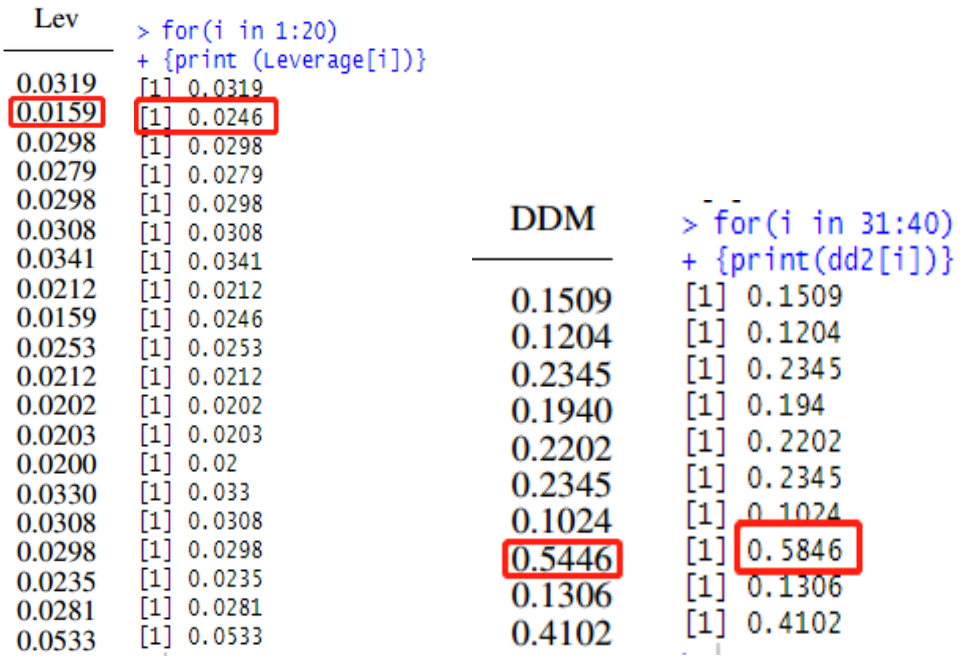
Index	$\hat{\pi}$	Lev	DM	DDM	Index	$\hat{\pi}$	Lev	DM	DDM
1	0.3836	0.0250	0.1057	0.2345	31	0.3743	0.02220	0.0949	0.1509
2	0.3730	0.0219	0.0935	0.1421	32	0.3690	0.0209	0.0898	0.1204
3	0.3809	0.0241	0.1024	0.2067	33	0.3836	0.0250	0.1058	0.2345
4	0.37830	0.0233	0.0992	0.1820	34	0.3796	0.0237	0.1008	0.1940
5	0.3809	0.0241	0.1024	0.2067	35	0.3823	0.0246	0.1040	0.2202
6	0.3823	0.0246	0.1040	0.2202	36	0.3836	0.0250	0.1058	0.2345
7	0.3863	0.0259	0.1093	0.2654	37	0.3638	0.0198	0.0856	0.1024
8	0.3651	0.0201	0.0865	0.1057	38	0.3144	0.0189	0.0876	0.5446
9	0.3730	0.0219	0.0935	0.1421	39	0.3469	0.0176	0.0775	0.1306
10	0.3743	0.0222	0.0949	0.1509	40	0.3230	0.0178	0.0814	0.4102
11	0.3651	0.0201	0.0865	0.1057	41	0.3598	0.0191	0.0829	0.0971
12	0.3534	0.0182	0.0796	0.1041	42	0.3367	0.0172	0.0769	0.2139
13	0.3611	0.0193	0.0837	0.0981	43	0.3405	0.0172	0.0767	0.1768
14	0.3585	0.0189	0.0821	0.0970	44	0.3469	0.0176	0.0775	0.1306
15	0.3850	0.0254	0.1076	0.2496	45	0.3546	0.0183	0.0801	0.1011
16	0.3823	0.0246	0.1040	0.2202	46	0.3444	0.0174	0.0770	0.1467
17	0.3809	0.0241	0.1024	0.2067	47	0.3546	0.0183	0.0801	0.1011
18	0.3444	0.0174	0.0770	0.1467	48	0.3585	0.0189	0.0821	0.0970
19	0.3380	0.0172	0.0768	0.2007	49	0.3393	0.0172	0.0767	0.1883
20	0.3193	0.0182	0.0838	0.4803	50	0.3585	0.0189	0.0821	0.0970
21	0.3783	0.0233	0.0992	0.1820	51	0.3521	0.0180	0.0790	0.1078
22	0.34826	0.0177	0.07780	0.1237	52	0.33043	0.0173	0.0782	0.2914
23	0.3180	0.0184	0.0847	0.5052	53	0.2859	0.0261	0.1280	**1.4742**
24	0.2209	0.0628	**0.3651**	**5.7680**	54	0.20851	0.07274	**0.4405**	**7.0602**
25	0.2745	0.0306	**0.1537**	**1.9781**	55	0.1904	0.0884	**0.5737**	**9.3070**
26	0.3393	0.0172	0.0767	0.1883	56	0.1736	**0.1044**	**0.7276**	**11.8675**
27	0.3944	0.0290	0.1213	0.3771	57	0.1580	**0.1200**	**0.9022**	**14.7416**
28	0.3809	0.0241	0.1024	0.2067	58	0.1435	**0.1349**	**1.0975**	**17.9293**
29	0.38095	0.0241	0.1024	0.2067	59	0.1301	**0.1487**	**1.3135**	**21.4307**
30	0.3944	0.0290	0.1213	0.3771	60	0.1179	**0.1612**	**1.5503**	**25.2458**

the leverage values of NO.54-60 could be not significantly large enough to be identified

On the other hand, the newest method MDDM could lead to other problem like masking and swamping. In the experiment on original Brown data set, MDDM identify NO.38 as HLP but other methods don’t which can be a good example of swamping. As is said, the definition of HLP is somehow ambiguous so criterion for identifying them varies. We can’t say the outcome is ‘wrong’ since we can’t agree on an unanimous criterion if one single data point is HLP due to lax definition. In our opinion, further tests are needed to tell if the model fits better if we omit data point NO.38 which is result from swamping.

There are also some differences between our outcomes from that in the paper. We have checked

all the calculation and programming and find out that there are some minor errors in this paper which might be the mean reason of different outcomes.



As is shown in the graphs, some of the calculations are different but we checked the properties the data should suffice(eg. the sum of Lev should equals to 2) and prove that our data are correct so there might be some errors in their calculation.

Despite some minor errors, the greatest innovation of the author’ s work is that we can focus on bi rather than leverage value and if we find a new criterion we can diminishes the influence of extreme pi value. Another innovation is that since the HLP is not mathematically-strictly defined one can define other reasonable criteria to identify HLP.

References

[1] A.H.M. Rahmatullah Imon & Ali S. Hadi (2013) Identification of multiple

high leverage points in logistic regression, Journal of Applied Statistics, 40:12, 2601-2616

[2] B.W. Brown, Jr, *Prediction analysis for binary data*, in *Biostatistics Casebook*, R.G. Miller, Jr, B. Efron, B. W.Brown, Jr, and L.E. Moses, eds., Wiley, New York, 1980, pp. 3–18.

[3] A.H.M.R. Imon, *Identification of high leverage points in logistic regression*, Pak. J. Stat. 22 (2006), pp. 147–156

[4] D. Pregibon, *Logistic regression diagnostics*, Ann. Stat. 9 (1981), pp. 705–724