FINDING HIGH RISK FACTORS RELATED WITH EARLY STAGE DIABETES

Bohao Zou[1]

[1]Department of statistics of UC Davis,

Background

Diabetes is threatening our human's health and the number of people who infect this disease are increasing significantly. I found a dataset which comes from UCI Machine Learning Repository and related with diabetes.
In this project, I will use this dataset to figure out which factors have a high relationship with the form of diabetes and give conclusion that if this factor would increase or decrease the risk to get diabetes.
In this dataset, there are 520 instances and 17 attributes. The response variable Y for this analysis is a binary variable. "Positive" means this instance has diabetes. In the remaining 16 attributes, only one of them is a continuous variable, it is "Age". The others attributes are all binary category variables. In this project, the statistically significant level is 0.05.

Method

1 Initial Model Build

In this dataset, the response variable Y, "class" is a binary variable. We can use ***logistic regression*** for solving this problem. The initial regression model only contains 16 main effects and we call this model as first logistic model.

2 Model Selection

At the beginning of model selection, we will use BIC as the criteria and set the direction of stepwise procedure as "both".
After model diagnosis, if we found the model is lack-of-fit by using Runs test. This may indicate we need to add more variables into model. This can be treated as model selection again. In this procedure, we will use AIC as the criteria and set the direction of stepwise process as "forward".

3 Model Diagnostic

3.1 Check lack-of-fit

1. Boxplots for pearson residuals and deviance residuals.
2. Residual plots for pearson residuals and deviance residuals against fitted values.
3. Runs test.

If one of them do not pass, we will treat this model lack-of-fit.

3.2 Multicollinearity detection

For each model, we will use Generalized variance-inflation factors (GCIF) to check the multicollinearity between variables. If there are more than one GCIF values of some variables higher than 10, we will remove the variable which has the highest GCIF value first and then fitted the regression model without removed variable. Repeat this process until all GCIF are all lower than 10.

Method

3.3 Outlier detection

We will identify outlying points by leverage value and Cook's distance.If the i-th case's leverage value bigger than $\frac{2p}{n}$ and Cook's distance bigger than $\frac{4}{n-p}$, it will be treated as an outlier.

Result

After model selection by BIC and "both" stepwise, there are 10 predictors in our model, 9 of them are the main effects and 1 of them are interaction effects. We call this model as BIC logistic model.

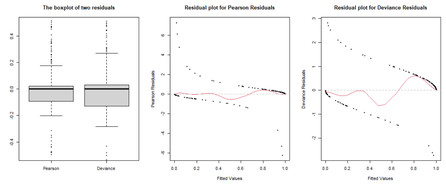

Fig. 1: The diagnostic of BIC logistic model.

The P-Values of Runs test for pearson residuals and deviance residuals are all smaller than $2.2e^{-16}$. This means this model lack-of-fit.
After model selection by AIC and "forward" stepwise, there are 14 predictors in our regression model. 9 of them are main effects and 5 of them are the interaction items. We call this model as AIC logistic model.

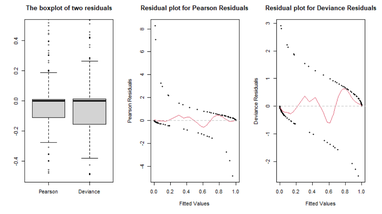

Fig. 2: The diagnostic of AIC logistic model.

The P-Values of Runs test for pearson residuals and deviance residuals are all smaller than $2.2e^{-16}$. This means this model lack-of-fit.
Why this complex model still lack-of-fit? We think this is because the particularity of our dataset. Most($\frac{15}{16}$) of variables are binary category variables. This means the flexibility of regression model that established by those binary category variables is very poor. Therefore, it is very hard for us to find a perfect regression model to have no lack-of-fit in it. In the next analysis, we still check if the model has lack-of-fit but we may not solve it.
By using GCIF to remove some variable from AIC logistic model and using leverage and Cook's distance to remove the ourliers. We can build the Ultimate model.

Result

The estimated coefficients and corresponding P-Values of ultimate model are showed below:

Coefficients	Estimated Value	Std Error	Z-Value	P-Value
Age	-0.02446	0.02594	-0.943	0.345698
Gender,Male	-4.19523	0.61559	-6.815	9.43e-12
Polyuria,Yes	4.18321	0.70964	5.895	3.75e-09
Polydipsia,Yes	6.06275	1.00879	6.01	1.86e-09
Polyphagia,Yes	1.00506	0.62183	1.616	0.106031
Itching,Yes	-2.84043	0.78739	-3.607	0.000309
Irritability,Yes	3.51199	0.86747	4.049	5.15e-05
Delayed.healing,Yes	-0.45029	0.73223	-0.615	0.538585
Partial.paresis,Yes	1.84429	0.65415	2.819	0.004811

Fig. 3: The table of coefficients of Ultimate model

The model diagnostic plots are showed below:

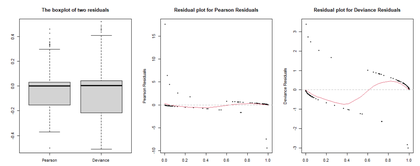

Fig. 4: The diagnostic of Ultimate model.

The result of Runs test still shows that there exist lack-of-fit in our ultimate model. But the residual plots shows that this ultimate model is much better than any previous models.

Conclusions

Those comparisons or conclusions are all in the situation that the others variables are all the same except the compared variable.

1. Compare with women, male will have lower probability to get diabetes.
2. Compare with the instances who do not have Polyuria, the instances who have Polyuria will increase the risk to get diabetes.
3. Compare with instances who do not have Polydipsia, those instances who have Polydipsia will increase the risk to get diabetes.
4. Compare with instances who do not have itching, the instances who have itching will decrease the risk to get diabetes.
5. Compare with instances who do not have irritability, the instances who have irritability will increase the risk to get diabetes.
6. Compare with the instance who do not have partial paresis, the instance who have partial paresis will increase the risk to get diabetes.

Most of our conclusions are all very closed with our common sense. This means we have a great assessment of this dataset.