STA 223 Project 1, Finding High Risk Factors Related with Early Stage Diabetes

Bohao Zou[1]
[1]Department of Statistics of UC Davis

Student ID:917796070
Keywords: Logistic Regression, Model Selection and Diagnostic
8th February, 2021

1 Introduction

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 it gives us some factors which may relate with diabetes. In this project, I will use this dataset to figure out which factors have a high relationship with the form of diabetes. This result may give us some inspiration for preventing the form of diabetes.

In this dataset, there are 520 instances and 17 attributes. The response variable Y for this analysis is "class" and it 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 category variables ("Sex", "Polyuria", "Polydipsia", "Sudden weight loss", "Weakness", "Polyphagia", "Genital thrush", "Visual blurring", "Itching", "Irritability", "Delayed healing", "Partial paresis", "Muscle stiness", "Alopecia", "Obesity") and they are all binary variables.

In this project, the statistically significant level is 0.05.

2 Methods

2.1 Initial Model Build

In this dataset, the response variable Y, "class" is a binary variable. By this property, we know we can use logistic regression in this project, it is a very common case in generalized linear model. The basic form of logistic regression is:

$$Linear \quad Predictor : \eta = \beta_0 + \sum_{j=1}^{p-1} x_j \beta_j \tag{1}$$

$$Link \quad Function : \eta = g(\mu) = log(\frac{\mu}{1-\mu}), E[Y] = \mu = \frac{e^\eta}{1+e^\eta}, 0 < \mu < 1 \tag{2}$$

The initial regression model only contains 16 main effects and **we call this model as first logistic model**.

2.2 Model Selection

There are 16 attributes in our dataset as the candidates of predictors. If we consider the interaction between those attributes, it is a huge number for us to pick and analyze them one by one. In this step, we will use some criteria to help us for model selection like AIC or BIC. For a wider scope of this model, we will set the upper bound of predictors as the 16 main effects plus interaction of each item with "Age" variable. This is because it is very hard for interpretation if we add the interaction item between two binary category variables. The lower bound is only containing the interception of regression model. Compare the formula of AIC

and BIC,

$$AIC: \quad AIC_p = nlog(\frac{SSE_p}{n}) + 2p, \quad k = 2 \tag{3}$$

$$BIC: \quad BIC_p = nlog(\frac{SSE_p}{n}) + log(n)p, \quad k = log(n) \tag{4}$$

we can know BIC penalties more on the model complexity (number of predictors in model, p). Because the number of predictor candidates in upper bound are too much. For tending to find a small and suitable model, we will use BIC as the criteria. The direction in this part is "both", which means variable can be added or removed from model by the BIC criteria.

After model diagnosis, if we found the model is lack-of-fit by using Runs Test. This means we need to add some higher order variable or interaction items into the model and it can be treated as model selection again. In this selection, the upper bound and the lower bound are as same as previous descriptions. This is because in this dataset, most of x variables are binary category variables,we can't add higher order terms. We used AIC criteria and set the direction as "forward". This is because we were more focused on the good fitness of this model. We decrease the penalties of the model complexity and we only want to add variable into model.

2.3 Model Diagnostic

2.3.1 Check lack-of-fit

We will use deviance residuals and pearson residuals to check if lack-of-fit exist in our model. If there is no lack of fit, those two different types of residuals will have similar distributions. We can use the boxplot of those two residuals to check if those two types of residuals have similar distribution.

We can also draw the different types of residuals against fitted values plots to check if our model contains lack-of-fit. Then we add a overlaying smoothing splines to fit those points in those residual plots. If the spline fluctuate around 0 slightly, it indicate that there is no lack-of-fit in our model. Otherwise, we need to use Runs test to give the final result.

If the spline fluctuate around 0 violent, we can run a Runs test to confirm our suspicions. The hull hypothesis of this test is there are no systematic patterns in those residuals. If one of them do not pass, we will treat this model lack-of-fit.

2.3.2 Multicollinearity detection

Because of the speciality of our dataset, most of x variables are binary category variable. We found that in some situation, there is no predictor is significant and their P-Values were all approach 1. But by using log likelihood test, we found there is a significant regression relationship between Y and entire set of x variables. This is the feature of multicollinearity. With the multicollinearity, the estimated regression coefficients tend to have a large standard

errors. It is possible that non of regression coefficients is statistically significantly but there is a significant regression relation between response variable Y and the entire set of x variables.

For generalized linear model, we can use Generalized variance-inflation factors (GCIF) to detect the multicollinearity in those x variables. In practice, if $GCIF > 10$, this may indicate that multicollinearity of that variable is high. We may remove that variable from our model. 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.

2.3.3 Outlier detection

We will identify outlying points by leverage value and Cook's distance. A high leverage value or high Cook's distance are considered this point as a potential outlier point. In practice, if leverage value bigger than $\frac{2p}{n}$, then the i-th case is identified as influential points. If Cook's distance bigger than $\frac{4}{n-p}$, the i-th case will be considered as potential influential case. In this project, if the leverage value and Cook's distance of i-th case are all bigger than each threshold, it will be treated as an outlier.

3 Results

3.1 Model Selection

The initial regression model for this model selection procedure is first logistic model. After model selection by BIC, there are 10 predictors in our model, 9 of them are the main effects and 1 of them are interaction effects. The 9 main effects are "Age", "Gender", "Polyuria", "Polydipsia", "Itching", "Irritability", "Delayed healing" and "Partial paresis". The one interaction effects is the interaction between "Age" and "Delayed healing". **We call this model as BIC logistic model.** The boxplot of two types of residuals shows below:

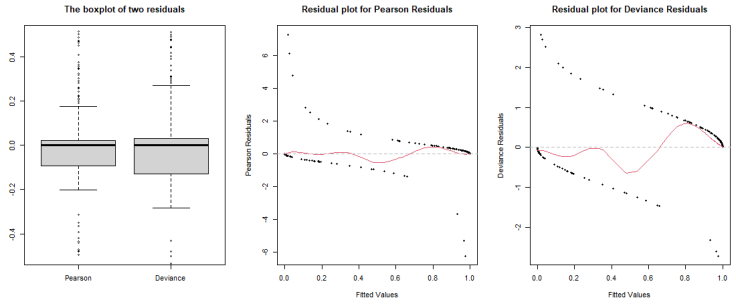

Fig. 1. Left Plot: The left plot shows the result of compare the distribution of two types of residuals. **Medium Plot:** The medium plot is the residual plot of pearson residuals. **Right Plot:** The right plot is the residual plot of deviance residuals.

From this plot we can know the pearson residuals and deviance residuals have similar distribution. However, from figure one we can know the spline fluctuates significantly around

0, This may indicate there exist systematic pattern in those residuals. We used Runs test to check the result. The results of Run test are showed below:

Runs Test	Standardized Runs Statistic	P-value
Pearson Residuals	-13.432	<2.2e-16
Deviance Residuals	-13.432	<2.2e-16

From the result of this table,the tests are all statistically significant under level $\alpha = 0.05$. Therefore, the tests indicate lack-of-fit in the model.

For solving this problem, we need to add higher order terms or interaction terms to see if the pattern persists. By the description of the second part of **Model Selection** of **Method**, we did model selection again and used the AIC as criteria for decreasing the penalties of model complexity. We set the direction as "forward", the variable can only be added into model but can't be removed. The initial regression model for this model selection procedure is BIC logistic model. After this model selection, there are 14 predictors in our regression model. 9 of them are main effects and 5 of them are the interaction items. The 9 main effects are "Age", "Gender", "Polyuria", "Polydipsia", "Itching", "Irritability", "Delayed healing" and "Partial paresis". The 5 interaction effects are the interaction between "Age" with "Delayed healing", "Partial.paresis", "Gender", "Irritability" and "Polydipsia". **We call this model as AIC logistic model**. We also found that there exist lack-of-fit in this AIC logistic model and by comparing the residual plots and the statistic of Runs test, we found the AIC logistic model is worse than BIC logistic model. (Please see the 5-appendix).

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.

3.1.1 Multicollinearity remove

We computed the GCIF value for AIC logistic model. We found most of variables have high GCIF value and most of variables are not significant. This may conduct by multicollinearity. By using the method that showed in the part of **Multicollinearity detection** of **Model Diagnostic**, we can get a model and **call it final model**. In the final model, there are 9 main effects but without interaction item. The predictors in the final model are "Age", "Gender", "Polyuria", "Polydipsia", "Itching", "Irritability", "Delayed healing" and "Partial paresis". The GCIF value of those variables are all lower than 2.1. This means we have a great regression model at present.

3.1.2 Outlier remove

We used leverage value and Cook's distance for finding the outliers based on the final model. If the leverage value and Cook's distance of i-th case are all bigger than corresponding threshold, we treated this case as outlier and then removed from our dataset. Finally, there

are 32 data points are removed from our dataset. The leverage points and Cook's distance points scatter plots are showed below:

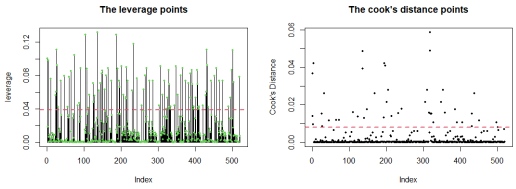

Fig. 2. Left Plot: The left plot is the scatter plot of leverage values. **Right Plot:** The right plot is the scatter plot of Cook's distance values.

3.1.3 Ultimate model build

After define the form of final model and removed the outliers, we can build the ultimate model by using the form of final model and trained with a clean dataset that removed outliers. The estimated coefficients and corresponding P-Values 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

We also did model diagnostic for ultimate model. The plots and tables are showed in 6-appendix . The boxplot of two types of residuals are still similar with each other and the residual plots are much better than BIC or AIC logistic model. However, The result of Runs test still shows that there exist systematic patterns in our ultimate model. Fortunately, the residual plots shows that this ultimate model is much better than any previous models.

4 Discussion

From the coefficient table, we can get some 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. This may against our common sense. We think itching is the feature of other disease and this disease would not occur with diabetes at same time. 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.

5 Appendix, AIC logistic model diagnostic

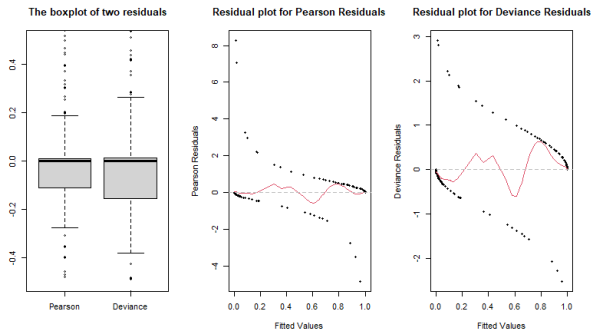

Fig. 3. Left Plot: The left plot shows the result of compare the distribution of two types of residuals. **Medium Plot:** The medium plot is the residual plot of pearson residuals. **Right Plot:** The right plot is the residual plot of deviance residuals.

Runs Test	Standardized Runs Statistic	P-value
Pearson Residuals	-14.658	<2.2e-16
Deviance Residuals	-14.658	<2.2e-16

6 Appendix, Ultimate model diagnostic

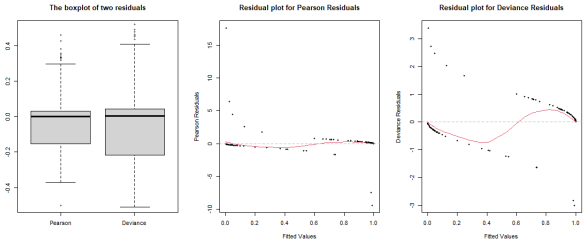

Fig. 4. Left Plot: The left plot shows the result of compare the distribution of two types of residuals. **Medium Plot:** The medium plot is the residual plot of pearson residuals. **Right Plot:** The right plot is the residual plot of deviance residuals.

Runs Test	Standardized Runs Statistic	P-value
Pearson Residuals	-13.683	<2.2e-16
Deviance Residuals	-13.683	<2.2e-16

7 Appendix, R Codes

```r
library("MASS")
library(lawstat)
library(car)
oriData = read.csv("C:\\Users\\Admin\\Desktop\\BST 223\\Project1\\
    diabetes_data_upload.csv")
classNum = rep(0, dim(oriData)[1])
classNum[which(oriData$class == "Positive")] = 1
classNumData = oriData
classNumData$class = classNum

### Model Selection

fit1 = glm(class ~., data = classNumData, family = binomial())
summary(fit1)
Scope = list(upper = ~. + Age*(.), lower = ~1)

BICFit= stepAIC(fit1, trace = FALSE, scope = Scope,k = log(520))
summary(BICFit)


### Model Diagnostic

# We expect these two types of residuals have similar distributions.
# no lack-of-fit => similar boxplots
# similar boxplots -> next step: Residual Plots
par()
res.P = residuals(BICFit, type="pearson")
res.D = residuals(BICFit, type="deviance") #or residuals(fit), by
    default

par(mfrow=c(1,3))
boxplot(cbind(res.P, res.D), names = c("Pearson", "Deviance"), ylim=c
    (-0.5,0.5),
        main="The boxplot of two residuals")

# * Residual Plots ----------------

# no lack-of-fit => no systematic pattern
# next step: Runs Test


plot(BICFit$fitted.values, res.P, pch=16, cex=0.6, ylab='Pearson
    Residuals', xlab='Fitted Values',
     main="Residual plot for Pearson Residuals")
lines(smooth.spline(BICFit$fitted.values, res.P, spar=2.1), col=2)
abline(h=0, lty=2, col='grey')
```

```
plot(BICFit$fitted.values, res.D, pch=16, cex=0.6, ylab='Deviance Residuals', xlab='Fitted Values',
     main="Residual plot for Deviance Residuals")
lines(smooth.spline(BICFit$fitted.values, res.D, spar=2.1), col=2)
abline(h=0, lty=2, col='grey')



runs.test(y = res.P, plot.it = TRUE)
title(main='Pearson Residual Runs Test')
runs.test(y = res.D, plot.it = TRUE)
title(main='Deviance Residual Runs Test')

### Add variable
AICFit= stepAIC(BICFit, trace = FALSE, scope = Scope,k = 2,direction = "both")
summary(AICFit)

par(mfrow=c(1,3))
res.P = residuals(AICFit, type="pearson")
res.D = residuals(AICFit, type="deviance") #or residuals(fit), by default
boxplot(cbind(res.P, res.D), names = c("Pearson", "Deviance"), ylim=c(-0.5,0.5), main = "The boxplot of two residuals")


plot(AICFit$fitted.values, res.P, pch=16, cex=0.6, ylab='Pearson Residuals', xlab='Fitted Values',
     main = "Residual plot for Pearson Residuals")
lines(smooth.spline(AICFit$fitted.values, res.P, spar=2), col=2)
abline(h=0, lty=2, col='grey')
plot(AICFit$fitted.values, res.D, pch=16, cex=0.6, ylab='Deviance Residuals', xlab='Fitted Values',
     main = "Residual plot for Deviance Residuals")
lines(smooth.spline(AICFit$fitted.values, res.D, spar=2), col=2)
abline(h=0, lty=2, col='grey')

runs.test(y = res.P, plot.it = TRUE)
title(main='Pearson Residual Runs Test')
runs.test(y = res.D, plot.it = TRUE)
title(main='Deviance Residual Runs Test')

### vif aicFit
sort(vif(AICFit))
AICFit = glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
               Polyphagia + Itching + Irritability + delayed.healing +
    partial.paresis + Age:partial.paresis + Age:Gender +
```

```
                Age:Irritability + Age:Polydipsia, family = binomial(),
    data = classNumData)
sort(vif(AICFit))
AICFit = glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
                Polyphagia + Itching + Irritability + delayed.healing +
    partial.paresis + Age:partial.paresis  +
                Age:Irritability + Age:Polydipsia, family = binomial(),
    data = classNumData)
sort(vif(AICFit))
AICFit = glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
                Polyphagia + Itching + Irritability + delayed.healing +
    partial.paresis   +
                Age:Irritability + Age:Polydipsia, family = binomial(),
    data = classNumData)
sort(vif(AICFit))
AICFit = glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
                Polyphagia + Itching + Irritability + delayed.healing +
    partial.paresis +
                Age:Polydipsia, family = binomial(), data = classNumData
    )
sort(vif(AICFit))
AICFit = glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
                Polyphagia + Itching + Irritability + delayed.healing +
    partial.paresis, family = binomial(), data = classNumData)
sort(vif(AICFit))
summary(AICFit)


### leverage
leverage = hatvalues(AICFit)
par(mfrow=c(1,2))
plot(names(leverage), leverage, xlab="Index", type="h", main="The
    leverage points")
points(names(leverage), leverage, pch=16, cex=0.6, col = 3)
p <- length(coef(AICFit))
n <- nrow(classNumData)
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPts <- which(leverage>2*p/n)

### Cooks
cooks = cooks.distance(AICFit)
cooksTh = 4/(n-p)
plot(cooks, ylab="Cook's Distance", pch=16, cex=0.6, main = "The cook's
     distance points")
abline(h=4/(n-p),col=2,lwd=2,lty=2)
infPtsCook = which(cooks>cooksTh)
interOutlier = intersect(infPts, infPtsCook)
```

```
120
121 removedData = classNumData[-interOutlier,]
122 finalFittedModel = glm(formula = class ~ Age + Gender + Polyuria +
    Polydipsia +
123                 Polyphagia + Itching + Irritability + delayed.healing +
    partial.paresis, family = binomial(), data = removedData)
124
125 summary(finalFittedModel)
126
127 par(mfrow=c(1,3))
128 res.P = residuals(finalFittedModel, type="pearson")
129 res.D = residuals(finalFittedModel, type="deviance") #or residuals(fit)
    , by default
130 boxplot(cbind(res.P, res.D), names = c("Pearson", "Deviance"), ylim=c
    (-0.5,0.5), main = "The boxplot of two residuals")
131
132 plot(finalFittedModel$fitted.values, res.P, pch=16, cex=0.6, ylab='
    Pearson Residuals', xlab='Fitted Values',
133      main = "Residual plot for Pearson Residuals")
134 lines(smooth.spline(finalFittedModel$fitted.values, res.P, spar=2.1),
    col=2)
135 abline(h=0, lty=2, col='grey')
136 plot(finalFittedModel$fitted.values, res.D, pch=16, cex=0.6, ylab='
    Deviance Residuals', xlab='Fitted Values',
137      main = "Residual plot for Deviance Residuals")
138 lines(smooth.spline(finalFittedModel$fitted.values, res.D, spar=2.1),
    col=2)
139 abline(h=0, lty=2, col='grey')
140
141 runs.test(y = res.P, plot.it = TRUE)
142 title(main='Pearson Residual Runs Test')
143 runs.test(y = res.D, plot.it = TRUE)
144 title(main='Deviance Residual Runs Test')
```