STA 138 FINAL PROJECT

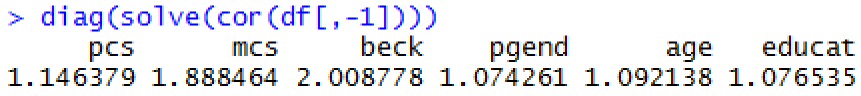
Yibing Luo 912491862

Introduction

In this project, we try to build a logistic model to analyze and predict the diagnosis of depression in primary care with 400 patients and related six variables. We expect the logistic regression can **perform a good prediction about diagnosis result**. Meanwhile, the logistic model should contain interpretable items, which allow us to **infer the relationship between response and explanatory variables**.

Material and Methods

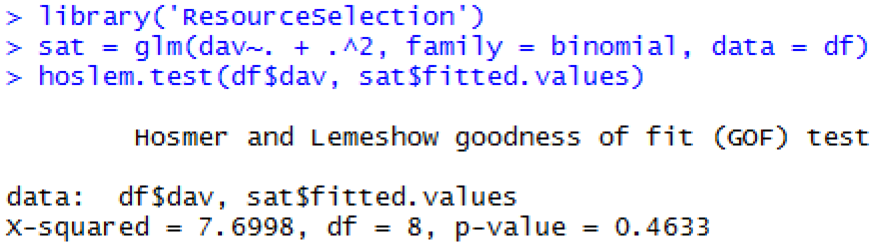
First of all, let's look into the depression data set. There are 400 observations. Among them, 64 observations are diagnosed as depression, which owns 16% of the whole sample. The six related variable are: PCS, MCS, BECK, PGEND, AGE, EDUCAT. Noticed that PCS, MCS, BECK are about patient mental of physical status, thus they may be correlated, which could lead to multicollinearity. So **VIF is calculated to decide whether multicollinearity exist:**



```
> diag(solve(cor(df[,-1])))
     pcs      mcs     beck    pgend      age   educat 
1.146379 1.888464 2.008778 1.074261 1.092138 1.076535 
```

From the table above, no VIF is larger than 10 which indicate the **multicollinearity is not severe.**

Further, a model with all second-order terms and second interaction terms is fitted and its goodness of fit is checked by doing **Hosmoer-Lemeshow goodness of fit test**:

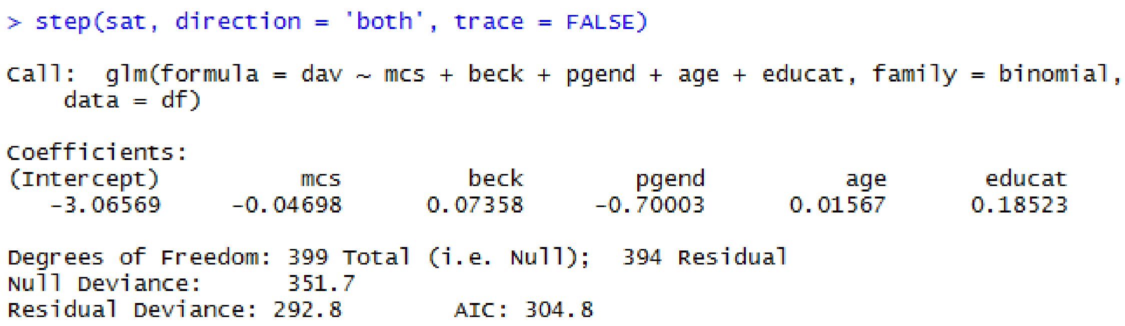


```
> library('ResourceSelection')
> sat = glm(dav~. + .^2, family = binomial, data = df)
> hoslem.test(df$dav, sat$fitted.values)

	Hosmer and Lemeshow goodness of fit (GOF) test

data:  df$dav, sat$fitted.values
X-squared = 7.6998, df = 8, p-value = 0.4633
```

Noticed that the X^2 equals to 7.7 and p-value equals to 0.4633. Thus we can conclude this **model fits the data well**. Percent concordance is also calculated: 76.6%. Therefore, we would like to choose this model as a 'full model' and a **backward stepwise procedure with AIC criterion** is applied:



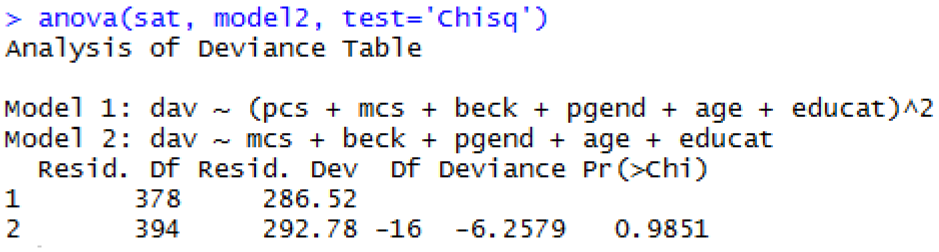
```
> step(sat, direction = 'both', trace = FALSE)

Call:  glm(formula = dav ~ mcs + beck + pgend + age + educat, family = binomial, 
    data = df)

Coefficients:
(Intercept)          mcs         beck        pgend          age       educat  
   -3.06569     -0.04698      0.07358     -0.70003      0.01567      0.18523  

Degrees of Freedom: 399 Total (i.e. Null);  394 Residual
Null Deviance:      351.7 
Residual Deviance: 292.8        AIC: 304.8
```

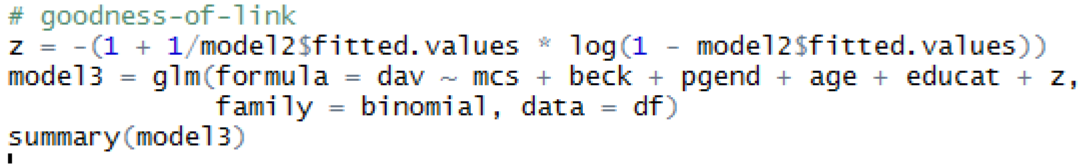
From the result below, we can obtain the final model selected by backward stepwise procedure. Notice that, in final model, **all the second-order terms are eliminated and PCS is also excluded**. Still, we want to compare these two nested model:



```
> anova(sat, model2, test='Chisq')
Analysis of Deviance Table

Model 1: dav ~ (pcs + mcs + beck + pgend + age + educat)^2
Model 2: dav ~ mcs + beck + pgend + age + educat
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       378     286.52                      
2       394     292.78 -16  -6.2579   0.9851
```

The p-value is 0.9851, which means there is **no statistically significant different between two models**. So the selected model is also good. To be robust, goodness-of-link test is also conducted:

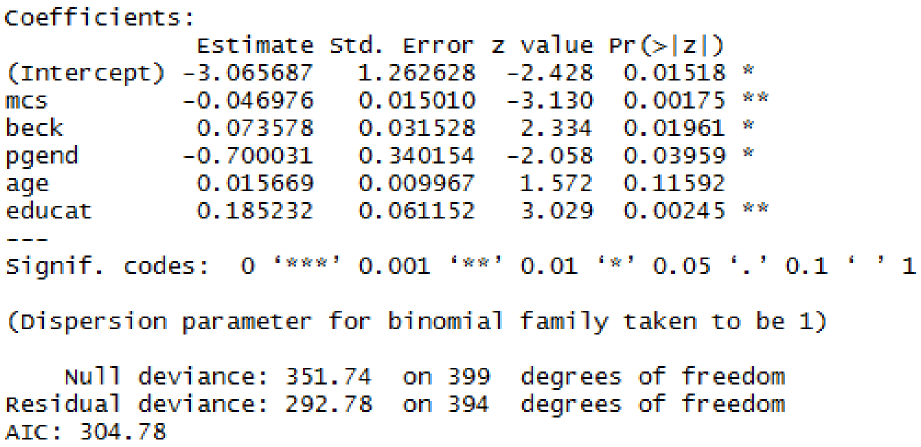


```
# goodness-of-link
z = -(1 + 1/model2$fitted.values * log(1 - model2$fitted.values))
model3 = glm(formula = dav ~ mcs + beck + pgend + age + educat + z,
             family = binomial, data = df)
summary(model3)
```

Notice the p-value of z is 0.82, which imply the link is appropriate. As a result, we decide to use this model to predict and analysis.

Result

Here is the brief summary of the model:



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)   
(Intercept) -3.065687   1.262628  -2.428  0.01518 * 
mcs         -0.046976   0.015010  -3.130  0.00175 **
beck         0.073578   0.031528   2.334  0.01961 * 
pgend       -0.700031   0.340154  -2.058  0.03959 * 
age          0.015669   0.009967   1.572  0.11592   
educat       0.185232   0.061152   3.029  0.00245 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 351.74  on 399  degrees of freedom
Residual deviance: 292.78  on 394  degrees of freedom
AIC: 304.78
```

So the model is as following:

$$\boldsymbol{\hat{\pi}(depression)} = \frac{\mathbf{exp(-3.07 - 0.05}\boldsymbol{mcs}\mathbf{ + 0.07}\boldsymbol{beck}\mathbf{ - 0.7}\boldsymbol{pgend}\mathbf{ + 0.02}\boldsymbol{age}\mathbf{ + 0.19}\boldsymbol{educat}\mathbf{)}}{\mathbf{1 + exp(-3.07 - 0.05}\boldsymbol{mcs}\mathbf{ + 0.07}\boldsymbol{beck}\mathbf{ - 0.7}\boldsymbol{pgend}\mathbf{ + 0.02}\boldsymbol{age}\mathbf{ + 0.19}\boldsymbol{educat}\mathbf{)}}$$

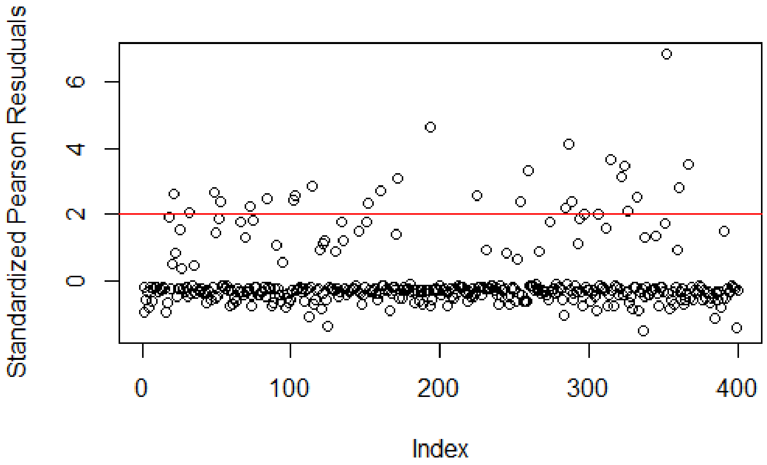
As shown, two of the predictor variables are significant under level 0.01, and four of them under 0.05. There are both quantitative variables and category variables.

For **PGEND (category variable)**, male is denoted by 1 and female is 0. The parameter is -0.7, which indicates that man is less likely to be diagnosed as depression. The **odds ratio for PGEND is 0.497**, which implies that the odds for male being diagnosed as depression is 0.497 times odds of women. **The 95% Wald CI is (0.255, 0.97)**

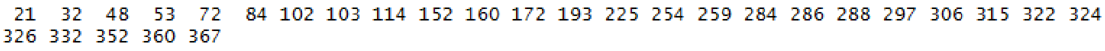
For **MCS (continuous variable)**, the estimated parameter is -0.047, which suggests that the higher one person's MCA, the less likely he or she to be diagnosed as depression. The 95% Wald CI is (0.93, 0.98). This implies that, under 95% confidence, one unit increase in MCS will lead to at least 2% and up to 7% decrease in odd ratio of being diagnosed as depression. Similarity, we can obtain the odds ratio and construct the 95% Wald CI for BECK, AGE, EDUCAT:

	Odds ratio	Lower	Upper
BECK	1.0764	1.01185	1.1450
AGE	1.0158	0.996144	1.0358
EDUCAT	1.2035	1.06756	1.3567

Next, we will do residual diagnose to find out if there is any outlier, or any influential point. First, we compute **standardized Pearson residual**:

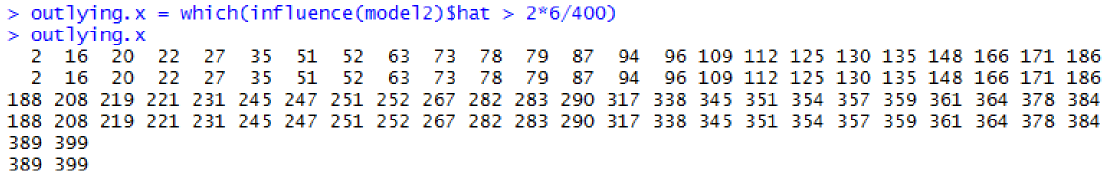


From the above plot, we can see that the residuals of some points is greater than 2, which may be considered as outliers and potential influential points. Their indexes are:



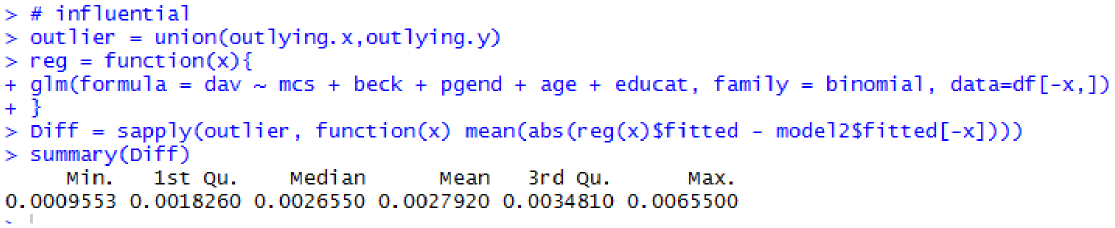
```
 21  32  48  53  72  84 102 103 114 152 160 172 193 225 254 259 284 286 288 297 306 315 322 324
326 332 352 360 367
```

Then we will consider leverage to find out the outlier in explanatory variables. The following output shows cases with leverage larger than 2p/n:



```
> outlying.x = which(influence(model2)$hat > 2*6/400)
> outlying.x
  2  16  20  22  27  35  51  52  63  73  78  79  87  94  96 109 112 125 130 135 148 166 171 186
  2  16  20  22  27  35  51  52  63  73  78  79  87  94  96 109 112 125 130 135 148 166 171 186
188 208 219 221 231 245 247 251 252 267 282 283 290 317 338 345 351 354 357 359 361 364 378 384
188 208 219 221 231 245 247 251 252 267 282 283 290 317 338 345 351 354 357 359 361 364 378 384
389 399
389 399
```

To see where any of the outliers is influential one, we can take each of outlier out and then calculate the average change of estimated probability:



```
> # influential
> outlier = union(outlying.x,outlying.y)
> reg = function(x){
+ glm(formula = dav ~ mcs + beck + pgend + age + educat, family = binomial, data=df[-x,])
+ }
> Diff = sapply(outlier, function(x) mean(abs(reg(x)$fitted - model2$fitted[-x])))
> summary(Diff)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0009553 0.0018260 0.0026550 0.0027920 0.0034810 0.0065500 
```

All the percentage is small, which shows that **none of these cases affect the prediction severely.**

Conclusion and Discussion

Above all, a logistic model might fit the data well. Five of the six potential predictors are selected without any interaction and higher order items included in the model.

The sign of MCS and BECK are reasonable, since MCS measure the mental health of a person while BECK shows the depression degree. And surprisingly, we can find that people who have higher education background are more likely to be depressed. Further, gender could be a significant factor. To be specifically, woman is more likely to be diagnosed as depression compared with man.

Generally speaking, the logistic model is well used in this project. But we can use cross-validation to test the model for a further step.