STA 138 Final Project - Multiple Logistic Regression

Ming Yang (998942775)

December 7th, 2016

Introduction

Depression(major depressive disorder or clinical depression) is a common but serious mood disorder. It causes severe symptoms that affect how people feel, think, and handle daily activities, such as sleeping, eating, or working. Research shows that depression is one of the most common mental disorder in the U.S. Current research suggests that depression is caused by a combination of genetic, biological, environmental, and psychological factors. Although depression, even the most severe cases, can be treated, the earlier that treatment can begin, the earlier that treatment can begin, the more effective it is. Therefore, finding the potential reason is critical in treating depression. In this project, we are going to discover the association between the explanatory factors including PCS(physical component of SF-36 measuring health status of the patient), MCS(Mental component of SF-36 measuring health status of the patient), BECK(The Beck depression score), PGEND(patient gender), AGE, and EDUCAT(Number of years of formal schooling) to the response factor DAV (diagonosis of depression in any visit during one year of care).

Material and Methods

The data used in this report is provided by Professor Azari. In the data, it contains 400 randomly selected patients. The explanatory variables are PCS(physical component of SF-36 measuring health status of the patient), MCS(Mental component of SF-36 measuring health status of the patient), BECK(The Beck depression score), PGEND(patient gender), AGE, and EDUCAT(Number of years of formal schooling) and the response variable is a binary variable specifying wether the patient is diagnosed with depression. Since the outcome variable for estimation is binary, logistic regression will be used in this project The DAV reflects the actual result of diagnosis of depression; whether the patient not diagnosed (0) or diagnosed (1). I will be modeling the probability of diagnosing depression with the multiple logistic regression model seen below:

$$log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = \underline{\hat{\beta}\prime} * \underline{x_i}$$

By applying the multiple logistic regression, I can quantify the risk of diagnosing depression with different factors. The estimated parameters will be analyzed by Wald hypothesis tests and Wald confidence intervals at 0.05 significance level. Association of predicted probabilities and observed reponses will be summarized. The criterion for goodness of fit will be based on both analysis of AIC and the Hosmer and Lemeshow goodness of fit test at the 0.05 significance level. Model diagnostics will be analyzed in a discussion of the standardized Pearson residuals. R will be the parimary tools used to conduct the computations and plots. The final model will be used to predict the probability of diagnosing depression when other explanatory variables are given. The validity of this prediction will be discussed.

Procedure

To investigate the association between explanatory variables and presence of depression, we fit a multiple logistic regression on the data. The following of code are information of the estimated parameters.

```
##                 Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -2.45884620 1.48497146 -1.6558205 0.097758172
## pcs         -0.01078479 0.01390324 -0.7757033 0.437924188
## mcs         -0.04923048 0.01533954 -3.2093856 0.001330190
## beck         0.06657292 0.03284446  2.0269146 0.042671146
## pgend       -0.67024460 0.34223571 -1.9584298 0.050179606
## age          0.01365697 0.01033161  1.3218623 0.186214022
## educat       0.18818150 0.06190234  3.0399738 0.002365987
```

The R reported model with Maximum Likelihood Estimates for the model is:

$$log(\frac{\hat{\pi_i}}{1-\hat{\pi_i}}) = \underline{\hat{\beta}'} * \underline{x_i} = 2.459+0.011*pcs-0.049*mcs+0.067*beck-0.670*pgend+0.014*age+0.188*educat$$

Based on the above formula, it is clear the predicted probability of diagnosing depression would change with change of explanatory variables.

Then we want to perform the hypothesis testing: For the test of $H_0 : \beta_i = 0$ and $H_1 : \beta_i \neq 0$, according to the R output, we know the values of Wal Chi-Square test statistic and p-values for 7 parameters (including the intercept and 6 explanatory variables). Therefore we can say that when $\alpha = 0.05$, we fail to separately reject the null hypothese in favor of H_0 for parameters includes intercept, pcs, and age, and separately reject the null in favor of H_1 for parameters includes mcs, beck, pgend, and educat. When $\alpha = 0.1$, we fail to separately reject the null in favor of H_0 for parameters includes pcs and age, and separately reject the null in favor of H_1 for rest of the parameters includes intercept, mcs, beck, pgend, and educat.

action	Df	Deviance	AIC
- pcs	1	292.78	304.78
- age	1	293.89	305.89
		292.18	306.18
- pgend	1	296.23	308.23
- beck	1	296.29	308.29
- educat	1	302.51	314.51
- mcs	1	302.87	314.87

Then use stepwise method to eliminate insignificant parameters: As we can see from the R output, we can say that the model fits the best when the parameter 'pcs' is eliminated from the original model. Therefore, let's have a new model which does not contain the factor 'pcs', and name it as 'model2'. After perform the stepwise method on model2, we know that the model2 has the smallest AIC value, then let's make model2 as our best model without intersections.

Test the intersections: In order to choose the best model, first we need to analyze the possible intersections. Again, use stepwise method to identify which model fits our data the best. Let's start with model2, and following is the R output we received after perform stepwise method:

actions	Df	Deviance	AIC
		292.78	304.78
+ mcs:educat	1	291.80	305.80
+ mcs:age	1	291.83	305.83
+ pcs	1	292.18	306.18

actions	Df	Deviance	AIC
+ age:educat	1	292.40	306.40
+ pgend:educat	1	292.45	306.45
+ mcs:pgend	1	292.68	306.68
+ pgend:age	1	292.77	306.77

According to the output, we have the 8 models with smallest AIC values, and according to the chart, we can see that the original model ,which does not have any interestion, has the smallest AIC value. Therefore, we decide model2 as our fianl model:

$$log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = \underline{\hat{\beta}'} * \underline{x_i} = 2.459 - 0.049 * mcs + 0.067 * beck - 0.670 * pgend + 0.014 * age + 0.188 * educat$$

Now we want to use our chosen model to find the odds ratios: Here are the R output of the 95% Wald Confidence Intervals, and we expect that all parameters are significant and does not contain the value 0.

```
##                    2.5 %      97.5 %
## (Intercept) -5.597638104 -0.63172439
## beck         0.011909498  0.13591837
## mcs         -0.076963698 -0.01792054
## pgend       -1.392997462 -0.05162981
## age         -0.004122931  0.03512301
## educat       0.069348270  0.30957561
```

We can see that besides the parameter 'age', all other parameters are significant. So we conclude that the variable 'age' has little or no effect on whether if a patient has depression; But from the stepwise method, we know that when the parameter 'age' is eliminated from the model, the model has a higher AIC value comparing to model2. So I decide to keep parameter 'age' in our final model.

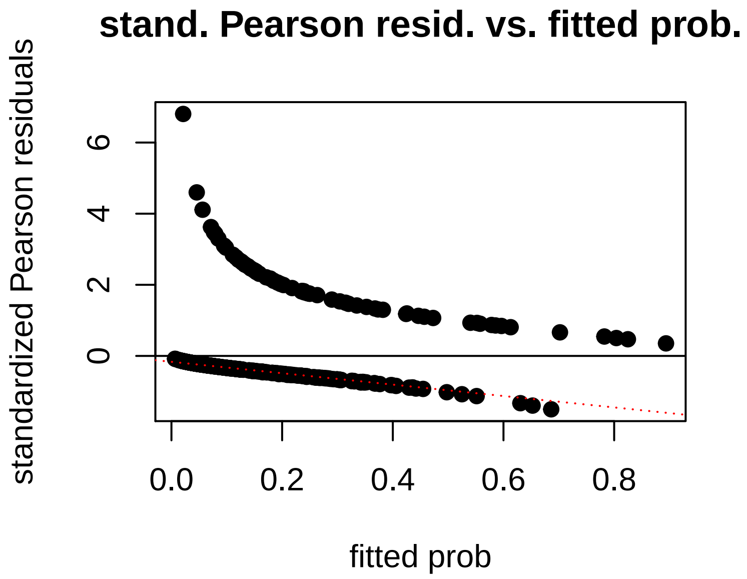
Following are the R output for the 95% odds ratio Confidence intervals:

```
##                   2.5 %    97.5 %
## (Intercept) 0.003706608 0.5316742
## beck        1.011980699 1.1455884
## mcs         0.925923466 0.9822391
## pgend       0.248329829 0.9496804
## age         0.995885556 1.0357471
## educat      1.071809423 1.3628466
```

This means when the explanatory variables are given and fixed, a patient who has one point increase in 'beck', we are 95% confident that the odds of the diagnosis of depression will increase by as small as 1.1% to as much as 14.6%. When a patient who has one point increase in 'mcs'(mental component), we are 95% confident that the odds of the diagnosis of depression will decrease by as small as 1.8% to as much as 7.4%. When a patient who is a male, then we are 95% confident that the odds of the diagnosis of depression will decrease by as small as 5% to as much as 7.52%. When a patient who is one year older, then we are 95% confident that the odds of the diagnosis of depression will increase by as small as -0.4% to as much as 3.6%. When a patient who has one more year of education, then we are 95% confident that the odds of the diagnosis of depression will increase by as small as 7.18% to as much as 36.28%.

Results

Now we want to check the residuals and see if it supports our model: By looking at the standardized Pearson residual plot, we noticed that there are a few observations with standardized Pearson residuals higher than 2, meaning that they are outliers. Yet, by fitting a lowess smoothing line in between, we found that interception of the line equals to zero and the slope of the line is negative. Thus, we conclude that the outliers do have some impact on the model, and our model might underestimate the data.



Perform Hosmer and Lemeshow goodness of fit test

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data$dav, fitted(model2)
## X-squared = 6.8257, df = 8, p-value = 0.5556
```

By applying Hosmer and Lemeshow test, we can directly measure the goodness of fit of the chosen model. The test has an null hypothesis that the model is a good fit of our data. We performed the test on R and received a Chi-Square statistic of 6.8257 with 8 degrees of freedom and p-value of 0.5556. The high p-value ($>\alpha$) indicates that we will not reject the null hypothesis, and thus the model we chose fits well with the given observed data on depression.

Conclusion and Discussion

A multiple logistic regression model was fit to the data. The effect of variables including mcs, beck, pgend, and educat are found to be significant, while the variables including pcs and age are not significantly different from zero. The predicted value for diagnosis of depression will change while our significant explanatory variables change, as mentioned when examine the fitted model. The model's fit was analyzed by examining the standardized Pearson residual plot and by using the Hosmer and Lemeshow goodness of fit test. Based on the results, I can not reject the hypothesis that the model is fitting the data; yet this might not be the best model for fitting our data. Although we perform a few tests on the effectiveness and the goodness of fit of the model, there are several problems in the model. The prediction power of the model is relatively weaker at two ends of the model. Moreover, there are several outliers in this model, which has some but not significant impact on the estimation of parameters. We can only conclude that there are associations between the reponse variable and our explanatory variables, yet there is no evidence of causation in this model. Thus, more studies need to be done to attain a more comprehensive view of depression.

Appendix

```r
data = read.table('final.dat', header = TRUE)
model = glm(dav ~ pcs + mcs + beck + pgend + age + educat,data = data, family = binomial)
step(model,scope=dav~pcs + mcs + beck + pgend + age + educat, direction="both")
model2  = glm(dav~beck + mcs + pgend + age + educat, data=data, family = binomial)
step(model2, scope=dav~ beck + mcs + pgend + age + educat +
       mcs * pgend * age * educat + age*pcs, direction="forward")
confint(model2)
exp(confint(model2))
pear.stdresid=resid(model2,type="pearson")/sqrt(1-lm.influence(model2)$hat)
plot(model2$fitted,pear.stdresid,pch=19,xlab="fitted prob",
     ylab="standardized Pearson residuals",
     main = "stand. Pearson resid. vs. fitted prob.")
abline(h=0)
abline(line(lowess(pear.stdresid~model2$fitted, f = 0.8)), col = "red",lty = "dotted")
library(ResourceSelection)
hoslem.test(data$dav, fitted(model2))
```