Spencer Phillips  
998887568  
12/7/16

**STA 138 Final Project**

**Introduction**

Depression is a new medical issue that has hit many individuals and has damaged and ended many lives. Depression if caught early by doctors or other individuals could help individuals who are fighting it a chance at help for the condition be it medication, therapy, or other treatments. The data that I am working with in this project is the diagnosis of depression by Primary Care doctors, the individuals that most people see for checkups. This is important to find what ways can be used to lead to people getting the right diagnosis in order to prevent unnecessary deaths or anguish that comes with depression and get those individuals the help that they need. In this project I will see which of the given factors, including multiple scoring styles and components of the SF-36 along with the age and gender of the patient and the years of formal education the patient had, to see which are significant and if there are any possible interactions between them that could lead to a more proper model.

**Material and Methods**

In this project, I am using data that was provided by Professor Azari on his course website about 400 randomly selected patients of primary care doctors which included one response variable which was categorical about if the patient was diagnosed with depression in any visit over an entire year. This was put against six different explanatory variables which included: Mental and Physical components of the SF-36 which measures the health status of individuals, the patient’s Beck depression score, the patient’s age, gender of patient, and the number of years of formal schooling that the patient had. This data analysis will be using R as the main way to perform any computations, plots, or charts.

To find the appropriate model for my data, I am going to use the step() function in R and perform backward and forward stepwise model selection on an initial model to be able to find the best possible model for the data. My initial data will be a simple model at first that just includes the explanatory variables and no interactions while going forward into a more saturated model that has up to three levels of interactions between the variables. I will also be going backwards from the saturated model towards the simple model to see if there is any difference in which model is best. I am going to be using AIC as the primary decision for the logistic regression model.

Once the optimal logistic model is found, I will perform multiple logistic regression analysis on the data through a Wald hypothesis test and Wald Confidence Interval. I will also be testing Goodness of Fit through the Hosmer and Lemeshow GOF test. I will then analyze the residuals that are available to analyze the test’s significance and any possible changes for the future.

**Results**

When attempting to find the optimal model for the data, I used the step() function in R and set the direction to be “both” meaning that values would be added or subtracted depending on what was best for the model. When running both the saturated model and the simple model through the step() function I found that running stepwise through R I found that the simple model gave me a less complicated model and lower AIC of 304.78 vs the saturated model’s 310.3. The Residual deviances were also approximately the same as well. Since these two were very close I decided to run an ANOVA Chi-squared test to see if they were at all significantly different to one another (Fig 1).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Analysis of Deviance Table (FIG 1)** | | | | |
| Model 1: dav ~ mcs + beck + pgend + age + educat | | | | |
| Model 2: dav ~ pcs + mcs + beck + pgend + educat + pcs:mcs + pcs:pgend + mcs:pgend + pcs:mcs:pgend | | | | |
| Resid. DF | Resid. Dev | DF | Deviance | Pr(>Chi) |
| 394 | 292.78 |  |  |  |
| 390 | 290.3 | 4 | 2.4787 | .6485 |

In this, Model 1 is our stepwise model from the simple model while Model 2 is the stepwise from the saturated model. We see in this ANOVA table that there is not much of a difference when testing at the .05 significance level. Therefore, we will use the first model in the ANOVA since it is a simpler model, making analysis easier. This gives us the following model of Depression Diagnosis vs Mental component of SF-36, Beck Depression score, Patient gender, Patient age, and the number of years of formal schooling of the patient. This shows that the physical status of the individual is less relevant in diagnosing depression in this model.

We will now look at the coefficients of our chosen model after performing the glm() and summary() functions on it in R (Thus follows is **Fig 2**)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficients** | Estimate | Std. Error | Z-Value | Pr(>|Z|) | Lower | Upper |
| (Intercept) | -3.065687 | 1.262628 | -2.428 | 0.01518 | -5.540437 | -0.59094 |
| MCS | -0.046976 | 0.015010 | -3.130 | 0.00175 | -0.076396 | -0.017556 |
| Beck | 0.073578 | 0.031528 | 2.334 | 0.01961 | 0.011783 | 0.13537 |
| Pgend | -0.700031 | 0.34015 | -2.058 | 0.03959 | -1.366733 | -0.03333 |
| Age | 0.015669 | 0.009967 | 1.572 | 0.11592 | -0.003867 | 0.0352 |
| Educat | 0.185232 | 0.061152 | 3.029 | 0.00245 | 0.065374354 | 0.30509 |

From this we can see that all of these apart from age are significant with the Wald test of H0: βi=0 rejecting that null hypothesis in favor of H1≠0 at the .05 significance level, with the βi values being equal to the values in the estimate column in Fig 2. We can see that the 95% Wald Confidence Interval shown in Fig 2 also corroborate this to be true for all except the Age variable.

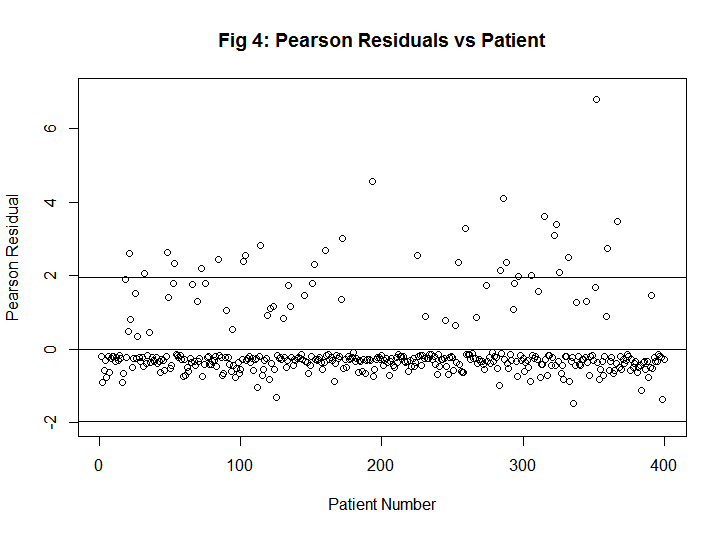
To test the Goodness of Fit of this model, I am going to be performing the Hosmer and Lemeshow GOF test. We get from this a Chi-Squared test statistic of 6.8257 (df=8) giving me a p-value of 0.5556. This means that testing at the .05 significance level that we cannot reject the null hypothesis that the model is fitting the data. This shows that there are no major problems in our data regarding fit.

These beta values do not mean much alone but by taking the exponential function of it we will be able to get the odds ratio and we can also look at the 95% Wald Confidence Interval of the Odds Ratio as well to see if they are significant at this level that there is a difference.

|  |  |  |  |
| --- | --- | --- | --- |
| **Fig 3 Odds Ratio** | Estimate | Lower bound | Upper bound |
| (Intercept) | 0.04662 | 0.003925 | 0.5538085 |
| MCS | 0.95411 | 0.9264487 | 0.9825971 |
| Beck | 1.07635301 | 1.011852759 | 1.1449648 |
| Pgend | 0.49656975 | 0.254938537 | 0.9672195 |
| Age | 1.01579245 | 0.996141768 | 1.0358308 |
| Educat | 1.20349755 | 1.067558594 | 1.3567465 |

This shows us a lot of interesting information with the model showing that in regards to gender that if a patient is male, their odds of being diagnosed with depression is 0.497 times that of female patients, showing that female patients have higher odds to be diagnosed with depression than men in primary care. Another is the influence on education where one additional year of formal education lead to approximately 1.2 times increase in their odds to be diagnosed with depression than someone who has one less year of education. The MCS Odds Ratio being less than one makes sense due to the fact that a lower score on the SF-36 means something is wrong, while a high score on the Beck Depression Inventory indicates depression which makes sense that it is larger than 1. Again, Age is the one that is ambiguous due to the fact that it straddles 0 and 1 in the 95% confidence interval and we cannot make a decision regarding it.

Now to measure the influence of data points in the model I looked at the Pearson residuals. If the absolute value of any of the residuals is greater than the absolute value of 1.96, we would find it to be an outlier in the model. To check for any errors in the fit of the model I looked at Pearson, Likelihood, and standard deviation residuals and found that for all of them there were approximately 5% of the values that exceeded the absolute value of 1.96 which is not majorly significant when working with 400 values, but are concerning due to some being above 4 as shown in Fig 4 for the Pearson residuals (lines are ±1.96 and 0).



**Conclusion and Discussion**

These results are somewhat fascinating in regards to this model while some reinforce what may already be true or are evidence in support of things. In particular, with each additional year of education leading to an increase of odds by 1.07 to 1.36 times is very interesting to see that if this is due to knowledge of depression or if income has another effect on this too. Another interesting finding is with women being having from 1.03 to 3.9 times the odds of men being diagnosed with depression is interesting if that is due to there being higher rates of depression in women or if men feel pressured by society to hide depression symptoms or even that doctors are less willing to diagnose men with depression in general. This lone variable should be looked into much more such as seeing the difference in scores of the SF-36 or Beck tests in men and women for example. The SF-36 and Beck tests had data reinforcing their validity with this model showing that an increase in the Beck score by one lead to an increase of the odds of depression diagnosis by 1.01 to 1.14 times and the SF-36 saw a change in the odds by 0.93 to 0.98 times for each one-point increase which corroborates what the scores on the tests mean. This shows that doctors take both tests seriously in their diagnosis.

Our model was unfortunately lacking in any interactions due to this model being the most optimal one to work with. It would be interesting however to see the data from the other model that we had found in order to see what effect the interactions had if any on diagnosis. As depression and the discussion on mental health care in the United States grows, it will be necessary to see what influences the diagnoses of the front line doctors of most patients, their primary care doctor. Finding what influences these diagnoses could lead to more education for doctors to be able to understand any biases they may have in their diagnoses and show what are important factors to look at. With this it could be done to be able to help our health care providers.

**Code Appendix**

care <- read.table("C:/Users/Spencer/OneDrive/STA138/care.txt", header=TRUE, quote="\"")

View(care)

#make models

simple= glm(dav~.,family=binomial,data=care)

saturated =glm(dav~.+.^2+.^3 ,family=binomial, data=care)

#use step to see which is better

stepper=step(saturated,scope = simple,direction = "both")

smallstep = step(simple,scope=saturated,direction="both")

#use ANOVA to check differences

anova(smallstep,stepper, test = "Chisq")

#summarize best model

summary(smallstep)

#Hosmer Lemeshow test

library(ResourceSelection)

hoslem.test(care$dav, smallstep$fitted.values)

#Odds Ratio and Wald CI of OR

OR=cbind(exp(coef(smallstep)), exp(summary(smallstep)$coefficients[,1] - 1.96\*summary(smallstep)$coefficients[,2]), exp(summary(smallstep)$coefficients[,1] + 1.96\*summary(smallstep)$coefficients[,2]))

OR

#Wald CI Coefficients

coef(smallstep)

summary(smallstep)$coefficients[,1] - 1.96\*summary(smallstep)$coefficients[,2]

summary(smallstep)$coefficients[,1] + 1.96\*summary(smallstep)$coefficients[,2]

#pearson resid

pear =resid(smallstep, type='pearson')

sum(abs(resid(smallstep, type='pearson'))>1.96)

#stdev resid

rstandard(smallstep)

sum(abs(rstandard(smallstep)>1.96))

#likelyhood resid

rstudent(smallstep)

sum(abs(rstudent(smallstep))>1.96)

#pearson chisquared

sum(residuals(smallstep, type = "pearson")^2)

deviance(smallstep)

1 - pchisq(deviance(smallstep), df.residual(smallstep))

plot(pear, main = "Fig 4: Pearson Residuals vs Patient", ylab="Pearson Residual", xlab = "Patient Number", ylim=c(-2,7))

abline(h=1.96)

abline(h=-1.96)

abline(h=0)