STA 138 Final Project

Which Variables Most Accurately Predict a Diagnosis of Depression?

Chirag Kashyap

998388067

Introduction

Dealing with depression on a daily basis can be debilitating for many people. Because depression is not well defined, there is no way of accurately diagnosing it using physical measures (like a blood test). Many depressed people think they are just going through tough times, and seek out the advice of therapists, psychiatrists, psychologists, and counselors. Even these people have a tough time deciding whether a patient is depressed or going through a tough time. The objective of this report is to find a series of variables that can accurately predict if a patient is depressed. In this project, I will fit a multiple logistic regression model that accurately predicts a diagnosis of depression, according to our dataset.

Material and Methods

The data used in the project consists of 400 patients who were either diagnosed with depression or not. The following variables that will help us predict a diagnosis of depression: PCS, MCS, Beck, PGend, Age, and Education. Professor Azari gave us this dataset, and the whole thing can be viewed at http://www.stat.ucdavis.edu/~azari/sta138/final.dat. He does not tell us where the data set is from, or where it was sourced from. PCS is the physical component of the SF-36, which measure the health status of the patient. MCS is the mental component of the SF-36, which measure the health status of the patient. Beck is the beck depression score. PGend is the patient gender. Age is self-explanatory. Education is the number of years of formal scoring. All of the variables are continuous, beside PGend, which is binary. Our binary response is DAV, which indicates the diagnosis of depression. Using multiple logistic regression I can model the probability of a diagnosis of depression with the following formula: $\text{logit}[\pi(x)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$, for $\beta_j = 1, 2, \ldots, m$.

I will first model univariate logistic regressions for each of the six possible explanatory variables in order to get a better view of the data. Then, I will estimate the parameters β_j's by using a stepwise procedure with AIC as my model selection criterion. This will not include interactions, so this model will assume independence between the explanatory variables. Next, I will examine the interactions by again using a stepwise procedure with AIC as the selection criterion. After finding the best possible logistic model for the data, I will use Wald hypothesis test and Wald confidence intervals to make sure the estimated parameters are statistically significant. Confidence intervals will also be provided to explain the various odds ratios within the model. A Hosmer and Lemeshow test goodness of fit test will be done to make sure the model fits properly. Percent concordant and percent discordant will be used to analyze the association between observed and predicted probabilities. Finally, I will also analyze the standardized Pearson residuals to make sure none of them deviate. The final model will be used to predict the probability of a diagnosis of depression and I will comment on the validity of this prediction.

Results

The AIC values from modeling a univariate logistic model for each of the size possible explanatory variables are below, along with the z-value and p-value from the Wald hypothesis test.

	PCS	MCS	Beck	PGend	Age	Education
AIC	352.68	314.37	324.69	347.51	355.04	352.2
z-Value	-1.746	-6.041	5.474	-2.728	0.842	1.819
p-value	0.0808	$1.53*10^{-9}$	$4.44*10^{-8}$	0.00637	0.4	0.0689

From the AIC and p-values, it is clear that MCS, Beck, and PGend should be in our final independent model. For PCS, Age, and Education, it is unclear if those variables will add to the final logistic model, based solely on their univariate logistic models.

The probability of a diagnosis of depression can be modeled with the maximum likelihood estimates for all of the β_js. Using multiple logistic regression and a stepwise procedure with AIC as my model selection criterion, I found $\beta_0 = -3.066$, $\beta_1 = -0.046$, $\beta_2 = 0.183$, $\beta_3 = 0.074$, $\beta_4 = -0.700$, $\beta_5 = 0.016$. PCS did not make it into the final model.

```
Start:  AIC=353.74
dav ~ 1

          Df Deviance    AIC
+ mcs      1   310.37 314.37
+ beck     1   320.69 324.69
+ pgend    1   343.51 347.51
+ educat   1   348.20 352.20
+ pcs      1   348.68 352.68
<none>         351.74 353.74
+ age      1   351.04 355.04
```

```
Step:  AIC=304.78
dav ~ mcs + educat + beck + pgend + age

          Df Deviance    AIC
<none>         292.78 304.78
- age      1   295.20 305.20
+ pcs      1   292.18 306.18
- pgend    1   297.27 307.27
- beck     1   298.24 308.24
- mcs      1   302.90 312.90
- educat   1   302.97 312.97
```

```
Call:
glm(formula = dav ~ mcs + educat + beck + pgend + age, family = binomial(link = "logit"),
    data = final.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5226  -0.5861  -0.3899  -0.2609   2.7757

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 **
educat       0.185232   0.061152   3.029  0.00245 **
beck         0.073578   0.031528   2.334  0.01961 *
pgend1      -0.700031   0.340154  -2.058  0.03959 *
age          0.015669   0.009967   1.572  0.11592
```

By the same process above, I tested to see if any interactions between variables should be included. It turned out that none of the interactions would help the model at all, so I ended up not including any of them. Therefore, our final model is still the model described above.

```
Start:  AIC=304.78
dav ~ mcs + educat + beck + pgend + age

              Df Deviance    AIC
<none>            292.78 304.78
- age          1   295.20 305.20
+ mcs:educat   1   291.80 305.80
+ beck:age     1   291.80 305.80
+ mcs:age      1   291.83 305.83
+ educat:beck  1   291.96 305.96
+ educat:age   1   292.40 306.40
+ educat:pgend 1   292.45 306.45
+ beck:pgend   1   292.58 306.58
+ mcs:pgend    1   292.68 306.68
+ mcs:beck     1   292.74 306.74
+ pgend:age    1   292.77 306.77
- pgend        1   297.27 307.27
- beck         1   298.24 308.24
- mcs          1   302.90 312.90
- educat       1   302.97 312.97
```

Now we test all of the β_js using the Wald Hypothesis test, with $H_o : \beta = 0$ and $H_a : \beta \neq 0$. For MCS (β_1), we have a z-value of -3.130 and a p-value of 0.00174, so we reject so we reject H_o and conclude H_a at an $\alpha = 0.05$. For Education (β_2), we have a z-value of 3.029 and a p-value of 0.00245, so we reject H_o and conclude H_a at an $\alpha = 0.05$. For Beck (β_3), we have a z-value of 2.334 and a p-value of 0.01961, so we reject H_o and conclude H_a at an $\alpha = 0.05$. For PGend (β_4), we have a z-value of -2.058 and a p-value of 0.03959, so we reject H_o and conclude H_a at an $\alpha = 0.05$. For Age (β_5), we have a z-value of 1.572 and a p-value of 0.11592, so we fail to reject H_o. Although, the hypothesis test does not think Age is significant, AIC still thinks that it marginally helps the model, so I decided to keep the variable. Likewise, the 95% Wald confidence interval for β_1 is $-0.046 \pm (1.96)0.015 = (-0.0754, -0.0166)$, for β_2 is $0.183 \pm (1.96)0.061 = (0.0634, 0.3025)$, for β_3 is $0.074 \pm (1.96)0.032 = (0.0113, 0.1367)$, for β_4 is $-0.700 \pm (1.96)0.340 = (-1.3664, -0.0366)$, for β_5 is $0.016 \pm (1.96)0.010 = (-0.0036, 0.356)$. All of my confidence intervals agree with my conclusion to reject the null hypothesis for β_1, β_2, β_3, and β_4, and fail to reject the null for β_5.

The 95% confidence interval of the odds ratios can be determined by taking the exponential of β_js. As seen in the table below.

β_j	CI for β_j	CI for Odds Ratios
β_1 (MCS)	$(e^{-0.0754}, e^{-0.0166})$	(0.9274, 0.9835)
β_2 (Education)	$(e^{0.0634}, e^{0.3025})$	(1.0655, 1.3532)
B_3 (Beck)	$(e^{0.0113}, e^{0.1367})$	(1.0114, 1.1465)
β_4 (PGend)	$(e^{-1.3664}, e^{-0.0366})$	(0.2550, 0.9641)
B_5 (Age)	$(e^{-0.0036}, e^{0.356})$	(0.9964, 1.4276)

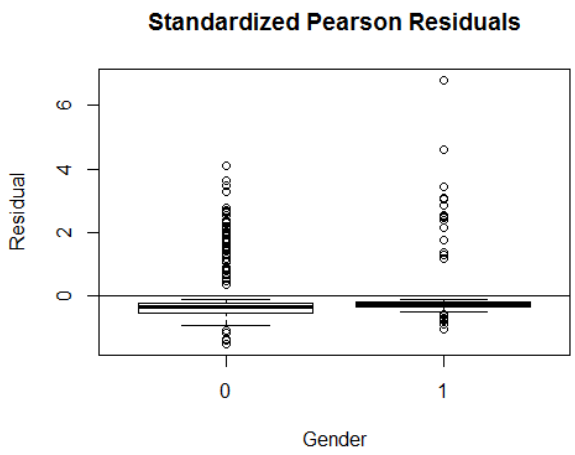
When MCS component increases by 1, we are 95% confident that the odds of diagnosing depression decreases by 1.65% to 7.26%. When Education level increases by 1, we are 95% confident that the odds of diagnosing depression increases by 6.55% to 35.32%. When the Beck score increases by 1, we are 95% confident that the odds of diagnosing depression increases by 1.14% to 14.65%. The conditional odds ratio of diagnosing depression between males and females is between (0.2550, 0.9641), so a diagnosis depression in a male is 3.59% to 74.50% less likely to happen than in a female. Since Age has 1 in the confidence interval for odds ratios, we cannot say anything about what impact it has on the diagnosis of depression.

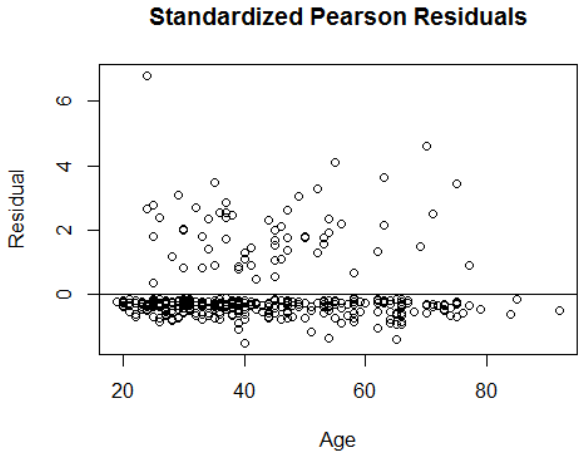
A concordance percentage of 77.96% indicates a moderate association between the predicted and observed probabilities of success. The data has 64 diagnosis of depression out of 400 patients. Therefore the model should be satisfactory in predicting a diagnosis of depression.

Percent Concordant	Percent Discordant	Percent Tied	Number of Pairs
77.96%	22.04%	0%	21504

The plots of standardized Pearson residuals are by gender and age, in order to see if there was a difference in the residuals over time and by gender. There does not seem to be a trend over time, but there are many residual over 2. All of these residuals over 2 are patients who were diagnosed with depression, but the model would not have predicted this. It appears

that while the model does a good job of predicting the diagnosis of no depression, it lacks power in predicting a diagnosis of depression.





The Hosmer and Lemeshow Chi-Square test statistic was 6.8257 with 8 degrees of freedom and a p-value of 0.5556, so we fail to reject the null hypothesis that the model is fitting the data. Although the fit is not perfect, it could work in the right setting.

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  depression.best$y, fitted(depression.best)
X-squared = 6.8257, df = 8, p-value = 0.5556
```

Conclusion and Discussion

Using depression diagnosis data from 400 patients, I was able to fit a multiple logistic regression model. β_1, β_2, β_3, and β_4 were found to be significant by both the Wald hypothesis test and Wald confidence interval, while β_5 was not. The 95% confidence interval of the odds ratios showed the changes in diagnosis percentage for each explanatory variable. The concordance percentage indicates a moderate association between the predicted and observed probabilities of success. However, the standardized Pearson residuals some deviation because of the diagnoses of depression that the model did not think was depression. This means that this model is not good at predicting a diagnosis of depression, but better at predicting a diagnosis of no depression. A multiple linear regression with data from diagnoses of depression could help with this problem, since there were only 64 cases were the patient was diagnosed with depression. Nonetheless, I was surprised to see so many of the residuals above 2. The Hosmer-Lemeshow chi-square test showed that although the model was not perfect, it would fit the data adequately. This model could be used by therapists, counselors, and psychiatrists to see if their patient was at risk of a diagnosis of depression and then plan accordingly.

Code Appendix

```
#STA 138 Final Project
#Chirag Kashyap 998388067
library(MASS)
library(boot)
library(ResourceSelection)

final.dat <- read.table("C:/Users/ckashyap/Desktop/Current Courses/STA
138/FinalProject/final.dat.txt", header=TRUE, quote="\"")
final.dat$dav = factor(final.dat$dav)
final.dat$pgend = factor(final.dat$pgend)
depression.pcs = glm(dav ~ pcs,data=final.dat, family = binomial(link="logit"))
depression.mcs = glm(dav ~ mcs,data=final.dat, family = binomial(link="logit"))
depression.beck = glm(dav ~ beck,data=final.dat, family = binomial(link="logit"))
depression.pgend = glm(dav ~ pgend,data=final.dat, family = binomial(link="logit"))
depression.age = glm(dav ~ age,data=final.dat, family = binomial(link="logit"))
depression.educat = glm(dav ~ educat,data=final.dat, family = binomial(link="logit"))
summary(depression.educat)

depression.null = glm(dav ~ 1,data=final.dat, family = binomial(link="logit"))
depression.full = glm(dav ~ pcs + mcs + beck + pgend + age + educat ,data=final.dat, family =
binomial(link="logit"))
depression.best = stepAIC(depression.null, scope = list(upper=depression.full),
direction="both", data=final.dat)
summary(depression.best)
depression.fullsq = update(depression.best, .~.^2)
depression.best = stepAIC(depression.best, scope = list(upper=depression.fullsq),
direction="both", data=final.dat)
summary(depression.best)

plot(final.dat$pgend, glm.diag(depression.best)$rp, ylab = "Residual", xlab = "Gender", main =
"Standardized Pearson Residuals" )
abline(0,0)
plot(final.dat$age, glm.diag(depression.best)$rp, ylab = "Residual", xlab = "Age", main =
"Standardized Pearson Residuals" )
abline(0,0)

concordance(depression.best)
hoslem.test(depression.best$y, fitted(depression.best), g = 10)

#from https://discuss.analyticsvidhya.com/t/how-to-get-the-percentage-concordant-and-
discordant-values-for-a-logistic-regression-model-in-r/1458/2
```

```
concordance<-function(model){
  # Get all actual observations and their fitted values into a frame
  fitted<-data.frame(cbind(model$y,model$fitted.values))
  colnames(fitted)<-c('respvar','score')
  # Subset only ones
  ones<-fitted[fitted[,1]==1,]
  # Subset only zeros
  zeros<-fitted[fitted[,1]==0,]
  # Initialise all the values
  pairs_tested<-0
  conc<-0
  disc<-0
  ties<-0
  # Get the values in a for-loop
  for(i in 1:nrow(ones)){
    for(j in 1:nrow(zeros)) {
      pairs_tested<-pairs_tested+1
      if(ones[i,2]>zeros[j,2]) {conc<-conc+1}
      else if(ones[i,2]==zeros[j,2]){ties<-ties+1}
      else {disc<-disc+1} }}
  # Calculate concordance, discordance and ties
  concordance<-conc/pairs_tested
  discordance<-disc/pairs_tested
  ties_perc<-ties/pairs_tested
  return(list("Concordance"=concordance,
              "Discordance"=discordance,
              "Tied"=ties_perc,
              "Pairs"=pairs_tested))}
```