**STA 138**

**Fall 2016**

**Final Project:**

**Diagnosis of Depression in Primary Care**

**Madeline Ye**

**ID# 998108849**

**Introduction:**

There are a number of common symptoms that help diagnose those with depression. As this disorder becomes more prevalent within today’s society, our desire to understand what correlations it may have with age and other common factors become of interest. In the following regression analysis, we wish to study factors related to the diagnosis of depression in primary care. With a better understanding, better treatments can be developed to help those who suffer from depression.

**Materials and Methods:**

A sample size of 400 randomly selected patients is used. There are seven variables of interest: Depression Diagnosis (DAV, 0=Not Diagnosed, 1 = Diagnosed), Physical Component (PCS), Mental Component (MCS), The Beck Depression Score (BECK), patient gender (PGEND, 0 = Female, 1= Male), age, and the number of years of formal schooling (EDU). All variables are continuous except for gender and our response variable of depression diagnosis.

The seven variables are being considered as potential predictors for the diagnosis of depression and will be included in our full initial model. The method that will be used is a multiple logistic regression analysis in R. The dependent variable of interest, diagnosis of depression, is binary thus a binomial distribution is suitable. The estimated coefficient parameters will be assessed using a Wald test and the Wald confidence interval at a significance level of 0.05. Step-wise model selection will be conducted in order to choose an appropriate model to represent the data. Once a model is selected, a goodness of fit test using Hosmer and Lemeshow will be conducted to confirm our choice. Finally, an analysis of the standardized Pearson residuals will be done to further assess the fit of the model.

**Results:**

The following output is for the final model. As we can see, the best model with AIC of 304.78 contains the variables mental component, Beck depression score, gender, age, and education. Of the five variables, the physical component was the only variable with a rather large p-value, suggesting that it lacks significance.

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 \*\*

edu 0.185232 0.061152 3.029 0.00245 \*\*

beck 0.073578 0.031528 2.334 0.01961 \*

gend -0.700031 0.340154 -2.058 0.03959 \*

age 0.015669 0.009967 1.572 0.11592

The result of the Hosmer and Lemeshow test at a 0.05 significance level has a p-value of 0.5556. This means we fail to reject the null hypothesis and can say the model fits the data well with 95% confidence.

The next test for the true values of coefficients is the Wald test. For each of the coefficients except for age, we reject the null hypothesis because the p-values are <0.05. The Wald Confidence intervals further support the test conclusion and are as follows:

2.5 % 97.5 %

(Intercept) -5.597638104 , -0.63172439

mcs -0.076963698 , -0.01792054

edu 0.069348270 , 0.30957561

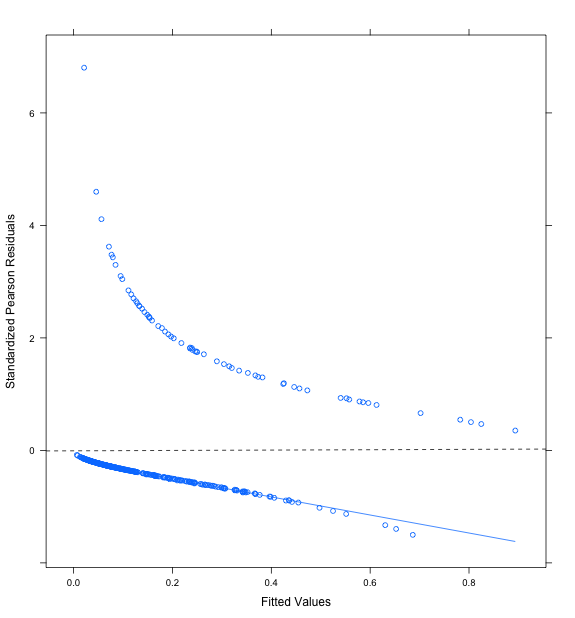
beck 0.011909498 , 0.13591837

gend -1.392997462 , -0.05162981

age -0.004122931 , 0.03512301

In order to further examine our previous results, a 95% Wald confidence interval on the odds ratio of the patient’s years of formal schooling and its effects on the diagnosis of depression is conducted. The 95% odds ratio confidence interval for education was [1.071809423, 1.3628466]. This is interpreted as for every additional year of formal schooling, the patient is about 7%-36% more likely to be diagnosed with depression.

Lastly, the Standard Pearson residuals showed us that our model is effective in representing the data. It shows that the data meets the assumptions of the model.



From the plot of residuals against fitted we see that the points are split into two lines. Since the two lines have a strong positive trend, small width values for an observation of a patient being diagnosed with depression will have a relatively large positive residual. Equally, for large width values of an observation, when the person is diagnosed with no depression, it will have a relatively large negative residual. As the two lines of Pearson residuals plotted against fitted, we see they are parallel towards the end. The plot shows that there are many outliers but suggests there is no significant error in model selection based on best fit to our data.

**Conclusion:**

From the data provided from the study on the diagnosis of depression in primary care, an estimated model was developed to estimate the probability of a patient being diagnosed with depression while considering several factors. A multiple logistic regression model is used with the following variables: Depression Diagnosis (DAV), Physical Component (PCS), Mental Component (MCS), The Beck Depression Score (BECK), patient gender (PGEND), age, and the number of years of formal schooling (EDU). After the step-wise model selection, we decide to drop the physical component variable as it tests to be not as significant as compared to the other variables. Our model selection is further confirmed by the Hosmer and Lemeshow goodness of fit test as well as the Wald test and 95% confidence intervals. An interesting find was that years of formal education had a significant effect on whether a patient was diagnosed or not. Further research into that given variable may contribute to further understanding variables of depression diagnoses.

Appendix

##

## Call:

## glm(formula = y ~ pcs + mcs + beck + gend + age + edu, family = binomial(link = logit),

## data = data)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -1.5372 -0.5778 -0.3912 -0.2609 2.8214

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -2.45885 1.48497 -1.656 0.09776 .

## pcs -0.01078 0.01390 -0.776 0.43792

## mcs -0.04923 0.01534 -3.209 0.00133 \*\*

## beck 0.06657 0.03284 2.027 0.04267 \*

## gend -0.67024 0.34224 -1.958 0.05018 .

## age 0.01366 0.01033 1.322 0.18621

## edu 0.18818 0.06190 3.040 0.00237 \*\*

## ---

## 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: 292.18 on 393 degrees of freedom

## AIC: 306.18

##

## Number of Fisher Scoring iterations: 5

##

## Call:

## glm(formula = y ~ 1, family = binomial(link = logit), data = data)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.5905 -0.5905 -0.5905 -0.5905 1.9145

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.6582 0.1364 -12.16 <2e-16 \*\*\*

## ---

## 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: 351.74 on 399 degrees of freedom

## AIC: 353.74

##

## Number of Fisher Scoring iterations: 3

## Start: AIC=306.18

## y ~ pcs + mcs + beck + gend + age + edu

##

## Df Deviance AIC

## - pcs 1 292.78 304.78

## - age 1 293.89 305.89

## <none> 292.18 306.18

## + mcs:age 1 291.01 307.01

## + beck:age 1 291.11 307.11

## + mcs:edu 1 291.17 307.17

## + beck:edu 1 291.51 307.51

## + pcs:age 1 291.52 307.52

## + pcs:mcs 1 291.56 307.56

## + pcs:gend 1 291.65 307.65

## + pcs:beck 1 291.68 307.68

## + age:edu 1 291.73 307.73

## + gend:edu 1 291.85 307.85

## + beck:gend 1 291.95 307.95

## + mcs:gend 1 292.09 308.09

## + gend:age 1 292.17 308.17

## + mcs:beck 1 292.17 308.17

## + pcs:edu 1 292.18 308.18

## - gend 1 296.23 308.23

## - beck 1 296.29 308.29

## - edu 1 302.51 314.51

## - mcs 1 302.87 314.87

##

## Step: AIC=304.78

## y ~ mcs + beck + gend + age + edu

##

## Df Deviance AIC

## <none> 292.78 304.78

## - age 1 295.20 305.20

## + mcs:edu 1 291.80 305.80

## + beck:age 1 291.80 305.80

## + mcs:age 1 291.83 305.83

## + beck:edu 1 291.96 305.96

## + pcs 1 292.18 306.18

## + age:edu 1 292.40 306.40

## + gend:edu 1 292.45 306.45

## + beck:gend 1 292.58 306.58

## + mcs:gend 1 292.68 306.68

## + mcs:beck 1 292.74 306.74

## + gend:age 1 292.77 306.77

## - gend 1 297.27 307.27

## - beck 1 298.24 308.24

## - mcs 1 302.90 312.90

## - edu 1 302.97 312.97

##

## Call:

## glm(formula = y ~ mcs + beck + gend + age + edu, family = binomial(link = logit),

## data = data)

##

## 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 \*\*

## beck 0.073578 0.031528 2.334 0.01961 \*

## gend -0.700031 0.340154 -2.058 0.03959 \*

## age 0.015669 0.009967 1.572 0.11592

## edu 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

##

## Number of Fisher Scoring iterations: 5

## y ~ mcs + beck + gend + age + edu

##

## Call:

## glm(formula = y ~ mcs + edu + beck + gend + age, family = binomial(link = logit),

## data = data)

##

## 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 \*\*

## edu 0.185232 0.061152 3.029 0.00245 \*\*

## beck 0.073578 0.031528 2.334 0.01961 \*

## gend -0.700031 0.340154 -2.058 0.03959 \*

## age 0.015669 0.009967 1.572 0.11592

## ---

## 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

##

## Number of Fisher Scoring iterations: 5

## Warning: package 'ResourceSelection' was built under R version 3.1.3

## ResourceSelection 0.2-6 2016-02-15

##

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

##

## data: y, fitted(model)

## X-squared = 6.8257, df = 8, p-value = 0.5556

## [[1]]

## Wald test:

## ----------

##

## Chi-squared test:

## X2 = 5.9, df = 1, P(> X2) = 0.015

##

## [[2]]

## Wald test:

## ----------

##

## Chi-squared test:

## X2 = 9.8, df = 1, P(> X2) = 0.0018

##

## [[3]]

## Wald test:

## ----------

##

## Chi-squared test:

## X2 = 9.2, df = 1, P(> X2) = 0.0025

##

## [[4]]

## Wald test:

## ----------

##

## Chi-squared test:

## X2 = 5.4, df = 1, P(> X2) = 0.02

##

## [[5]]

## Wald test:

## ----------

##

## Chi-squared test:

## X2 = 4.2, df = 1, P(> X2) = 0.04

##

## [[6]]

## Wald test:

## ----------

##

## Chi-squared test:

## X2 = 2.5, df = 1, P(> X2) = 0.12

## Waiting for profiling to be done...

## 2.5 % 97.5 %

## (Intercept) -5.597638104 -0.63172439

## mcs -0.076963698 -0.01792054

## edu 0.069348270 0.30957561

## beck 0.011909498 0.13591837

## gend -1.392997462 -0.05162981

## age -0.004122931 0.03512301

## Waiting for profiling to be done...

## OR 2.5 % 97.5 %

## (Intercept) 0.04662182 0.003706608 0.5316742

## mcs 0.95410996 0.925923466 0.9822391

## edu 1.20349755 1.071809423 1.3628466

## beck 1.07635301 1.011980699 1.1455884

## gend 0.49656975 0.248329829 0.9496804

## age 1.01579245 0.995885556 1.0357471