STA 138 Final Project
Diagnosis of Depression in Primary Care

Christopher Wong (999234204)

Introduction

Depression is a brain disorder characterized by persistently depressed mood or loss of interest in activities, causing significant impairement in daily life. It affects how an individual feels, thinks, and behaves and can lead to emotional and physical problems. Depression is not something that can be "snapped out" of. Instead, it may be treated with medication and counseling.

Material and Methods

The data can be found on the course website: http://www.stat.ucdavis.edu/~azari/sta138/final.dat

Attribute Information:

DAV:	Diagnosis of depression in any visit during one year of care. 0 = Not diagnosed 1 = Diagnosed
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 0 = Female 1 = Male
AGE:	Patient's age in years.
EDUCAT:	Number of years of formal schooling.

The data consists of 400 observations of these variables. **DAV** is the response variable and the remaining variables are explanatory variables.

To analyze the data we use multiple logistic regression which is appropriate because DAV is a binary response and PCS, MCS, BECK, AGE, and EDUCAT are continuous explanatory variables with PGEND being a binary response variable.

Results

To start the analysis, we fit a multiple logistic regression model to our data using all variables:

$$\begin{aligned} logit(\pi) = log(odds(\pi)) = log(\frac{\pi}{1-\pi}) = -2.4588462 &- 0.0107848(PCS) \\ &- 0.0492305(MCS) \\ &+ 0.0665729(BECK) \\ &- 0.6702446(PGEND) \\ &+ 0.013657(AGE) \\ &+ 0.1881815(EDUCAT) \end{aligned}$$

where π = probability that the patient has been diagnosed with depression in any visit during one year of care.

Checking the coefficients, we see that PCS and AGE may be insignificant based on their p-values from the Wald test. Recall that the Wald test tests hypotheses $H_0 : \beta_k = 0$ vs. $H_A : \beta_k \neq 0$ using the Wald statistic $z_W = \frac{\hat{\beta}_k - 0}{SE(\beta_k)}$ which is distributed standard normal.

```
##                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
```

So, our naively fitted full model may not be the best model of this data. We use the **step** function from the **stats** package to perform forward selection and backward elimination stepwise procedures. These procedures minimize AIC in each step. Both procedures fit the same model which is the reduced model with MCS, EDUCAT, BECK, PGEND, and AGE as explanatory variables.

Thus, our actual model is:

$$\begin{aligned} logit(\pi) = log(odds(\pi)) = log(\frac{\pi}{1-\pi}) = -3.0656867 &- 0.0469763(MCS) \\ &+ 0.1852319(EDUCAT) \\ &+ 0.0735785(BECK) \\ &- 0.7000313(PGEND) \\ &+ 0.015669(AGE) \end{aligned}$$

where π = probability that the patient has been diagnosed with depression in any visit during one year of care.

The estimated conditional odds ratios can be interpreted by fixing all other variables and looking at e^{β_k}. The corresponding confidence intervals for the conditional odds ratios are given by $\left[e^{\beta_k - 1.96SE(\beta_k)}, e^{\beta_k + 1.96SE(\beta_k)}\right]$.

The conditional odds ratios and their confidence intervals are given below:

```
##       mcs    educat      beck     pgend       age
## 0.9541100 1.2034975 1.0763530 0.4965698 1.0157924
```

```
##            lower     upper
## mcs    0.9264487 0.9825971
## educat 1.0675586 1.3567465
## beck   1.0118528 1.1449648
## pgend  0.2549385 0.9672195
## age    0.9961418 1.0358308
```

For example, fixing the other explanatory variables, the conditional odds ratio for MCS can be interpreted as for every unit increase in a patient's mental component of SF-36, we expect to see a 4.5890035% decrease in the odds of them being diagnosed with depression. Going further, we are 95% confident that the true conditional odds ratio for MCS lies in the confidence interval [0.9264487, 0.9825971]. We are 95% confident

that there is a 1.7402884% to 7.3551296% decrease in the odds of a patient being diagnosed with depression for every unit increase in their MCS.

For this data, we choose not to use a likelihood ratio test for goodness of fit because we have n = 400 independent observations each with different values for the explanatory variables. This means that $n_i = 1$ for all i leading to the deviance only depending on the reduced model and thus is not a good measure of fit.

Instead, we use a Hosmer-Lemeshow test for goodness of fit and find that

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

Since our p-value = 0.5556, we conclude that the fit of our model is good.

Next, we perform residual diagnostics for our model using studentized Pearson residuals $\hat{r}_{p,i} = \frac{r_i}{\sqrt{1-h_i}}$ which are close to standard normal. We notice that there are some studentized Pearson residuals not in the range of [-2, 2] which correspond to outliers in our data.

```
##     dav     pcs     mcs beck pgend age educat Pearson stdresid outlier.hat
## 21    1 32.2315 58.0860    5     0  47     15         2.617510 0.013608050
## 32    1 23.5279 35.9691   11     0  30     11         2.062820 0.018388189
## 48    1 53.7016 32.7951    1     0  24     12         2.652614 0.017221466
## 53    1 21.2015 35.9125    4     0  34     12         2.353422 0.024008013
## 72    1 29.6330 48.4587    4     0  56     14         2.210641 0.011440129
## 84    1 33.8180 35.1747    6     1  38     14         2.459376 0.015496374
## 102   1 40.7527 55.0323   10     0  37     14         2.409841 0.017009308
## 103   1 44.6555 35.4889    6     0  36     10         2.565829 0.009738885
## 114   1 55.6982 41.8934    4     1  37     15         2.844677 0.011074245
## 152   1 50.3570 45.5971    6     0  44     13         2.309539 0.008594207
## 160   1 30.1590 44.3445    1     0  32     14         2.702588 0.013091063
## 172   1 57.6159 50.3594    0     1  49     17         3.044938 0.014007182
## 193   1 31.9046 42.5512    2     1  70      8         4.598019 0.013552424
## 225   1 42.2539 40.6580    6     1  37     15         2.568892 0.011100960
## 254   1 29.0489 39.6889    2     0  54     12         2.368172 0.013981177
## 259   1 24.6544 58.6858    3     0  52     13         3.298494 0.010720210
## 284   1 38.5470 27.6100   12     1  63      9         2.174931 0.028097948
## 286   1 41.4101 63.5514    2     0  55     12         4.110705 0.007898855
## 288   1 33.9159 33.5005   11     1  26     13         2.376481 0.019928349
## 306   1 22.2715 37.9221    7     0  45     12         2.023715 0.015056911
## 315   1 42.9568 53.7292    0     0  63     11         3.622068 0.009247746
## 322   1 26.7603 39.7015    7     1  29     13         3.100038 0.016646029
## 324   1 45.1633 52.2031    5     1  75     12         3.432579 0.014903299
## 326   1 37.4024 32.0619    7     0  46     10         2.112154 0.011943277
## 332   1 46.2415 40.7488    2     1  71     14         2.518144 0.020068843
## 352   1 58.0773 56.7069    0     1  24     12         6.803639 0.004250068
## 360   1 31.3771 53.2316    0     0  25     17         2.773556 0.020409407
## 367   1 55.9457 54.5684    0     0  35     14         3.479163 0.007884265
```

Before we consider removing these outliers in our data, we check to see if they are influential observations. Recall that an observation is influential if $h_i > \frac{2p}{n}$. The h_i values for these observations are given above.

Notice that none of these values are greater than $\frac{2p}{n} = \frac{12}{400} = 0.03$. Thus, none of these outliers are influential observations and we do not remove them.

Conclusion and Discussion

In summary, our model is

$$\begin{aligned} logit(\pi) = log(odds(\pi)) = log(\frac{\pi}{1-\pi}) = -3.0656867 &- 0.0469763(MCS) \\ &+ 0.1852319(EDUCAT) \\ &+ 0.0735785(BECK) \\ &- 0.7000313(PGEND) \\ &+ 0.015669(AGE) \end{aligned}$$

and we have found that MCS, years of education, Beck depression score, gender, and age is significant in determining whether a patient is diagnosed with depression or not.

For our conditional odds ratios, we find that as MCS increases, the odds of a patient being diagnosed with depression decreases. As years of education increases, the odds of a patient being diagnosed with depression increases. Similarly, as a patient's Beck score increases, the odds of them being diagnosed with depression increases. Gender also plays a role in whether a patient is diagnosed with depression with men having lower odds than women. Age is a tossup in that the odds of a patient being diagnosed with depression can either increase or decrease as age increases. This is indicative that while age plays a small role in the odds of a patient being diagnosed with depression (we estimate a 1.5792446% increase in odds for every additional year), it may not be significant at a 95% confidence level.

Of course, this project does not give a definitive answer of whether these are the only factors in determining if a patient has depression, but it does show that they are significant. Depression is very complex and it is unknown exactly what causes depression. There are many other variables that can be considered when diagnosing a patient with depression or not.

Code Appendix

```
setwd("C:/Users/Christopher/Desktop/STA 138/Final Project")

# DAV:
#   Diagnosis of depression in any visit during one year of care.
#   0 =  Not diagnosed
#   1 = Diagnosed
# 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
#   0 = Female
#   1 = Male
# AGE:
#   Patient's age in years.
# EDUCAT:
#   Number of years of formal schooling.

depression <- read.table("final.dat", header = TRUE)
attach(depression)

model <- glm(dav~., data = depression, family = binomial(link = logit))
summary(model)$coefficients

nothing <- glm(dav~1, data = depression, family = binomial(link = logit))
full <- glm(dav~., data = depression, family = binomial(link = logit))

forwards <- step(nothing, scope = list(lower = formula(nothing), upper = formula(full)), direction = "f

backwards <- step(full, scope = list(lower = formula(nothing), upper = formula(full)), direction = "bac

# Forwards and backwards stepwise lead to the same model, minimizing AIC
final.model <- forwards

# Odds ratios and CIs for odds ratios
est.coef <- coef(final.model)
est.coef.se <- summary(final.model)$coefficients[, "Std. Error"]

odds <- exp(est.coef[-which(names(est.coef) == "(Intercept)")])

odds.CI <- data.frame(lower = exp(est.coef[-which(names(est.coef) == "(Intercept)")]-1.96*est.coef.se[-
      upper = exp(est.coef[-which(names(est.coef) == "(Intercept)")]+1.96*est.coef.se[-which(names(est.

# Hosmer-Lemeshow test for goodness of fit
library(ResourceSelection)
hoslem.test(dav, fitted(final.model))

# standardized Pearson residuals #
```

```r
pear.stdresid=resid(final.model,type="pearson")/sqrt(1-lm.influence(final.model)$hat)

# Outliers
outlier.pear.stdresid <- pear.stdresid[which(abs(pear.stdresid) > 2)]
h <- lm.influence(model)$hat
outlier.hat <- h[which(abs(pear.stdresid) > 2)]
outliers.df <- data.frame(cbind(depression[which(abs(pear.stdresid) > 2), ],
                                outlier.pear.stdresid,
                                outlier.hat))
names(outliers.df)[8] <- "Pearson stdresid"
outliers.df

# Influential observations
n <- nrow(depression)
p <- length(final.model$coefficients)
influential.bound <- 2*p/n

# None of the outliers are influential observations
# outlier.hat > influential.bound
```