Diagnosis of Depression in Primary Care

Analysis of Categorical Data

Statistics 138

Audrey Chu
999138148

December 2016

Abstract

To study factors related to the diagnosis of depression in primary care through the categorical data analysis of 400 randomly selected patients. Data is gathered from the *Journal of Women's Health and Gender-Based Medicine, Vol. 10, Number 7*. With the application of multiple logistic regression modeling through SAS and R and evaluation of odds ratios, residual analysis, estimation, and hypothesis testing, a final predictive and generalized linear model (GLM) will be selected by goodness of fit to determine the identification of depression.
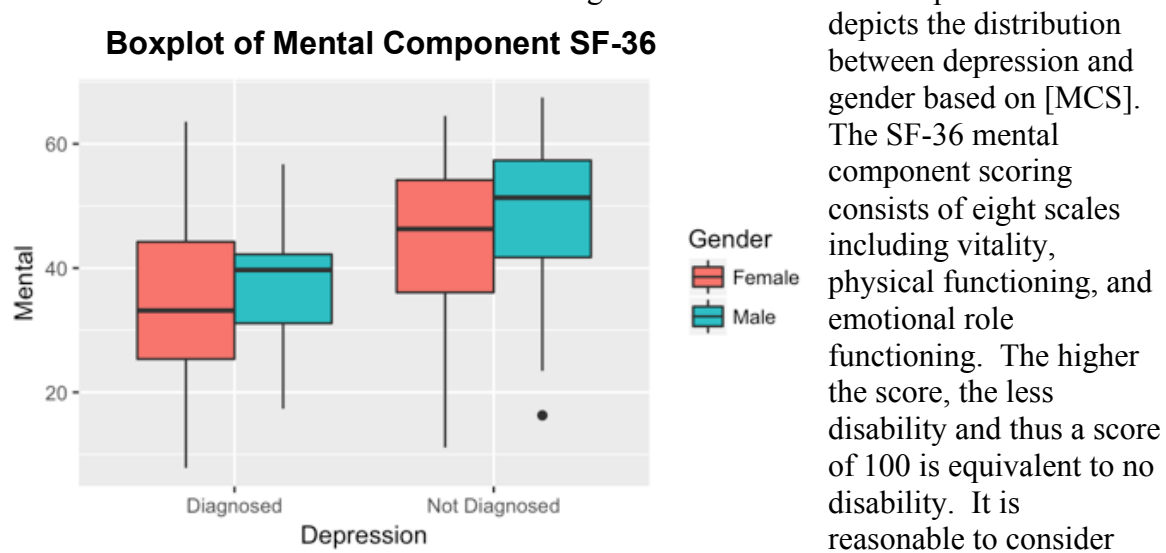
I. Introduction

This report analyzes logical and continuous variables to best determine the response variable of diagnosis of depression in any visit during one year of primary care. The diagnosis will be denoted as [DAV] with understanding that 0 equals not diagnosed and 1 equals diagnosed. Other variables in consideration include continuous variables: physical component of SF-36 measuring health status of patient [PCS], mental component of SF-36 measuring health status of the patient [MCS], the Beck depression score [BECK], patient's age in years [AGE], and number of years of formal schooling [EDUCAT]. The dataset also includes a binary variable of patient gender [PGEND] where 0 indicates female and 1 indicates male.

II. Material and Methodology

A multiple logistic regression is appropriate since these data have one nominal and two or more measurement variables. The correlogram in Figure 1.0 shows a slight relationship between mental component and the Beck depression score with a correlation of -0.67032. The data also show that more females are diagnosed than men. The boxplot below depicts the distribution between depression and gender based on [MCS]. The SF-36 mental component scoring consists of eight scales including vitality, physical functioning, and emotional role functioning. The higher the score, the less disability and thus a score of 100 is equivalent to no disability. It is reasonable to consider gender in fitting a model. Boxplot diagnostic demonstrates that for those who are diagnosed with depression, females have a lower SF-36 score. Sample mean scoring for females is approximately 7 values lower. In both diagnosed and not diagnosed, females

Boxplot of Mental Component SF-36

have a lower SF-36 score. Appendix A contains a similar visualization with the Beck score as a measurement and further corroborates this claim.

Fitting a model to measure the presence or absence of depression [DAV] starts with a logit function that contains all the variables,

$$logit(\pi(x)) = \log\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \alpha + \beta_1 PCS + \beta_2 MCS + \cdots + \beta_6 EDUCAT$$

In Appendix C, the SAS output for shows the results for all these coefficients. The likelihood-ratio testing $H_0 = \beta_1 = \cdots = \beta_6 = 0$ has a test statistic of 59.5605 with df = 6 (P < 0.0001) and thus suggests extremely strong evidence that at least one predictor has an effect.

Stepwise Model Selection
A model is then fitted using stepwise selection in SAS for both forward and backwards procedures. The output illustrates the forward selection procedure in which significant variables are added to the model starting with an intercept only model depending on a significance level. For the backwards procedure, insignificant variables are removed at each step dependent on selected significance level. The PROC LOGISTIC model statements for both forward and backwards procedures consider all possible interaction terms if they fit the criteria specified in the stepwise algorithm. For simplicity, interaction terms up to two-factor levels are only tested.

III. Findings
The forward stepwise selection results in a model containing variables [MCS], [BECK], [PGEND], and [EDUCAT]. The regression's multiple explanatory variables are a mixture of quantitative and qualitative and the model for $\pi(x) = \text{P}(DAV = 1)$ is $\pi(x)$

$$= -\frac{\exp(-2.6410 - .0470MCS + 0.0721BECK + 0.1785EDUCAT + 0.3316PGEND}{1 + \exp(-2.6410 - .0470MCS + 0.0721BECK + 0.1785EDUCAT + 0.3316PGEND)}$$

Similarly, the backwards stepwise procedure gives a model with the same variables excluding [PGEND]. The backwards model is $\pi(x)$

$$= -\frac{\exp(-2.2498 - .0496MCS + 0.0724BECK + 0.1655EDUCAT)}{1 + \exp(-2.2498 - .0496MCS + 0.0724BECK + 0.1655EDUCAT)}$$

For the forward and backwards model, it is important to note that the maximum likelihood estimates are very similar in value. In consideration of the variables and their significance, the forward model will best fit the data. Although the backwards model does not consider [PGEND], gender does have an association in that the 95% Wald confidence interval contains 1 in [0.999, 3.772].

Odds Ratios

The odds ratios estimates produced for the forward and final model are as shown below,

Odds Ratio Estimates and Wald Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
MCS	1.0000	0.954	0.926	0.983
BECK	1.0000	1.075	1.010	1.143
EDUCAT	1.0000	1.195	1.059	1.349
PGEND Female vs Male	1.0000	1.941	0.999	3.772

The estimate of [MCS], calculated from exp $(-0.0470) = 0.954$, says that the chance an individual will be diagnosed depression decreases 4.6% for every increase in mental component of SF-36. In terms of variables increasing odds, [BECK]'s estimate is 1.075. Calculated from exp $(0.0721) = 1.075$, the ratio states that the estimated odds of being diagnosed with depression increases by 7.5% for every increase in Beck depression scoring. There is also a 19.5% higher chance of being diagnosed with depression in primary care for every increase in number of years of formal schooling.

The estimate for gender is 1.941, exp (-0.0496), which states that the odds of being diagnosed with depression for males increases is 1.941 times that of female. However, [PGEND] has a confidence interval of [0.999, 3.772]. Since the value of 1 is within the interval and the estimate is largely greatly than one, we cannot reject the null hypothesis to conlude that gendera and depression diagnosis have an association between each other. However, when looking at the small p-value says that this variable is significant.

Residual Analysis

As in classical linear models, model checking is also important in logistic regression and probit analysis. Pearson residuals and deviance residuals are calculated to show the difference between observed and fitted values. Note that since this is not a linear model, pearson and deviance residuals take into account the fact that each observation has a different variances.

The residual distribution diagnostics are as shown on the next page. The disperson of data points show an overestimation for individuals that have not been diagnosed with depression. In contrast, the disperson of data points for diagnoses individuals show not only a problem of underestimation, but also higher spread between 0 and 6 as well as more outliers. Thus, the residual analysis suggests that this model may not be the best in predicting depression, especially in the case of those who are diagnosed.

Standardized Pearson and Deviance Residuals
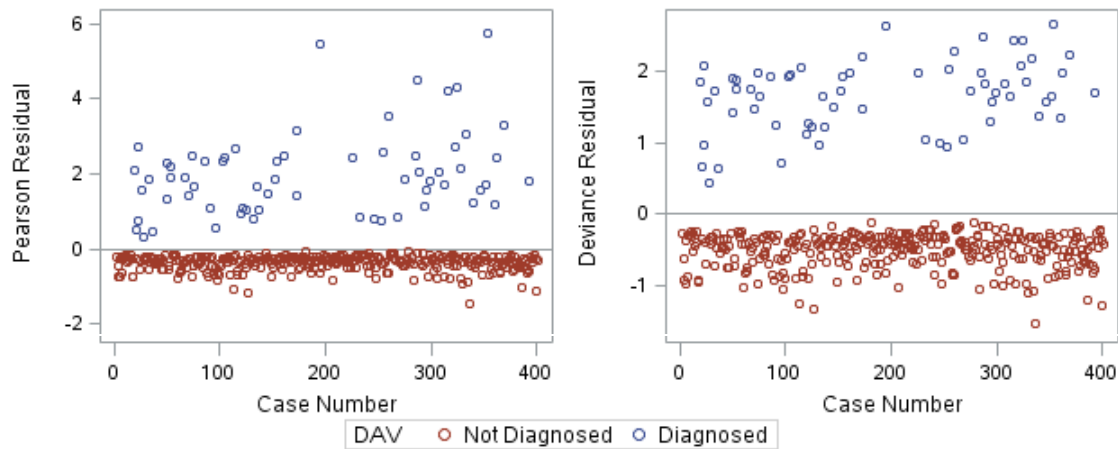
When further analyzing the residuals by looking at each variable, education seems to be the best predictor with a residual bound between -0.4 and 0.2 (shown on next page). This shows a strong relationship and supports the odds ratio stating that the chance of being diagnosed with depression for an individual with more schooling is 1.195 that of an individual with one less year of formal schooling.

Standardized Pearson Residuals for Education
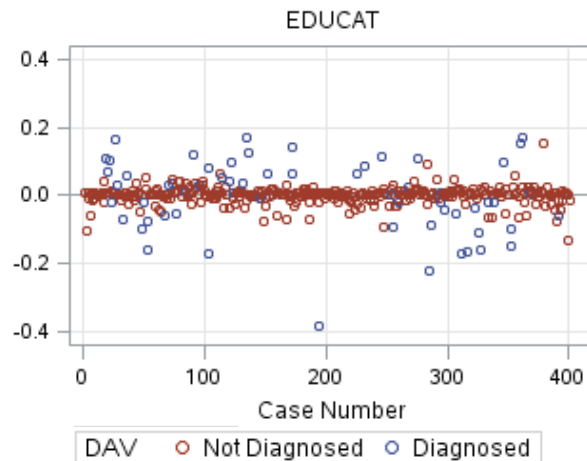
Goodness of Fit

The criterion for goodness of fit for the final multiple logistic regression is based on both the analysis of the Hosmer and Lemeshow Test and the Akaike Information Criterion (AIC) at a $\alpha = 0.05$ significance level. AIC, which is founded on information theory, offers a relative estimate of information lost when a give model is used to represent the data and also deals with balancing between goodness of fit and complexity of models and variables. The AIC of the intercept only model is 353.736 and the AIC of the final model containing on [MCS], [BECK], [PGEND], and [EDUCAT] is 305.201. The decrease in

AIC after eliminating age and the physical component [PCS] has a better effect on the model.

The Hosmer and Lemeshow practice uses predicted probability to create groups. In the case of using SAS, the procedure automatically produces a group of 10, and so in large samples, the usual chi-squared distribution has degrees of freedom equal to $10 - 2 = 8$. The Hosmer and Lemeshow Chi-Squared test results in 7.4172 (df $= 8$) and p-value of 0.4924. Therefore at a $\alpha = 0.05$ significance level, the null hypothesis that the model is fitting the data cannot be rejected.

IV. Conclusion and Discussion

In analyzing this categorical dataset and fitting a predictive model through stepwise procedures, the variables of mental component, Beck depression score, patient gender, and number of years of formal school best predict the diagnosis of depression in primary care. The multiple logistic regression, also a predictive model or generalized linear model, has binary and continuous variables. Calculated odds ratios show that as a patient receives more years of formal schooling, there is a higher chance of 19.5% that he or she will receive a diagnosis of depression. This is a logical conclusion in that over education is commonly linked with poor mental health. Education can have detrimental effects for innumerable reason, one of which is that overeducated people might not be challenged by their jobs and cannot use all of their skills they acquired during their education (Bracke).

Hence, in determining the diagnosis of depression, a generalized linear model using multiple logistic regression suggests that the Beck depression score and years of education are worthy indicators. The main findings of this predictive model also show that although there is a clear difference in distribution between gender and diagnosis, gender is not the best indicator when compared to the other logical or continuous variables.

Next Steps

Improvements on this final model would include looking more closely between gender and it's effect on diagnosis of depression. Given the time constraint of this analysis, it was not feasible. It would also be interesting to research what other variables would contribute to a higher Beck depression score and lower mental component of SF-36 in measuring the mental health status of the patient.

Appendix A: R Code and Output

---

install.packages("corrgram") install.packages("ggplot2")

```
library(corrgram)
library(ggplot2)
library(plyr)
```

```
final = read.table("~/Desktop/final.txt", header=TRUE, quote="\"")    # Read in data

corrgram(final, order=TRUE,
         lower.panel = panel.ellipse,
         upper.panel = panel.pts,
         text.panels = panel.txt,
         diag.panel = panel.minmax,
         main = "Figure 1.0: Correlogram of Depression Data")         # Note MCS v Beck
```

Figure 1.0: Correlogram of Depression Data
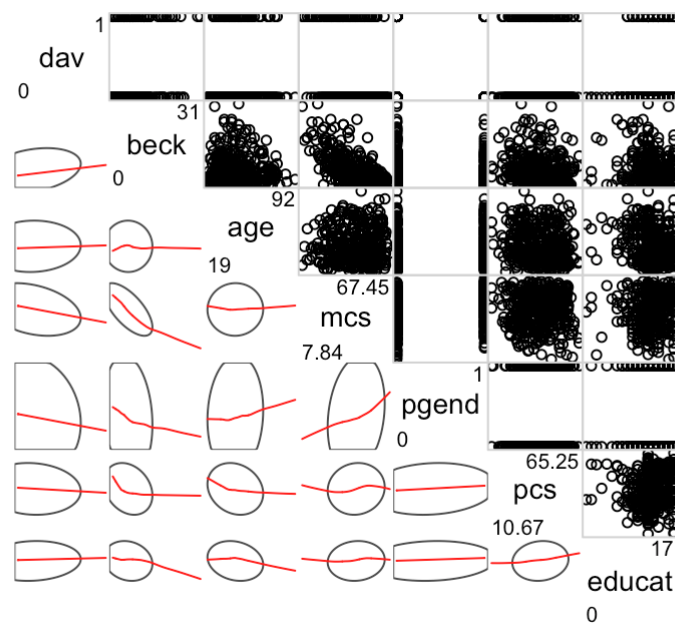

```
cor(final$beck, final$mcs)
```

```
## [1] -0.670318
```

```
cor(final$beck, final$pcs)
```

```
## [1] -0.2525155
```

```
Beck = final$beck                                              # Rename
Mental = final$mcs
Education = final$educat
Gender = revalue(as.character(final$pgend), c("1" = "Male", "0" = "Female"))
Depression = revalue(as.character(final$dav), c("1" = "Diagnosed", "0" = "Not Diagnosed"))


final = cbind(final, Gender, Depression)                       # New data frame

ggplot(final, aes(x=Depression, y=Beck, fill=Gender)) + geom_boxplot() + labs(title="Fig
ure 2.0: Boxplot of Beck Score")
```

Figure 2.0: Boxplot of Beck Score
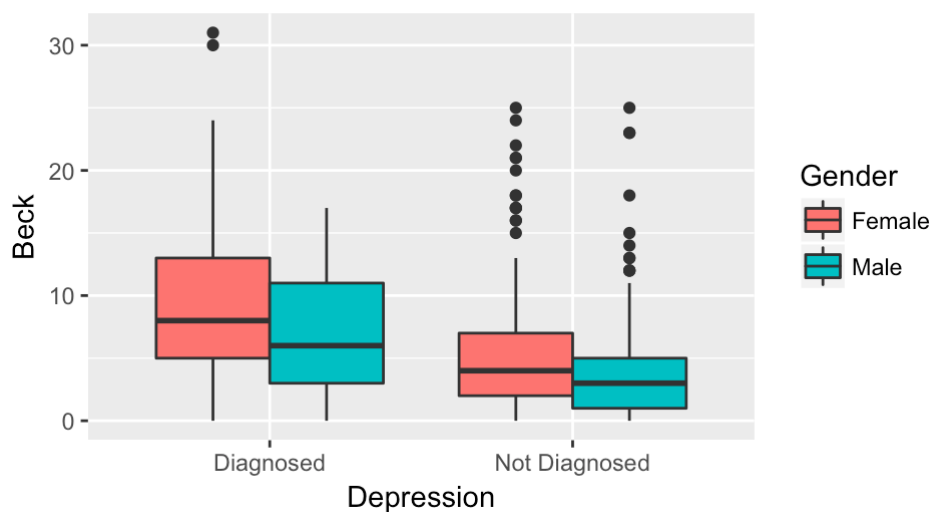

```
ggplot(final, aes(x=Depression, y=Mental, fill=Gender)) + geom_boxplot() + labs(title="F
igure 2.1: Boxplot of Mental Component")
```

Figure 2.1: Boxplot of Mental Component
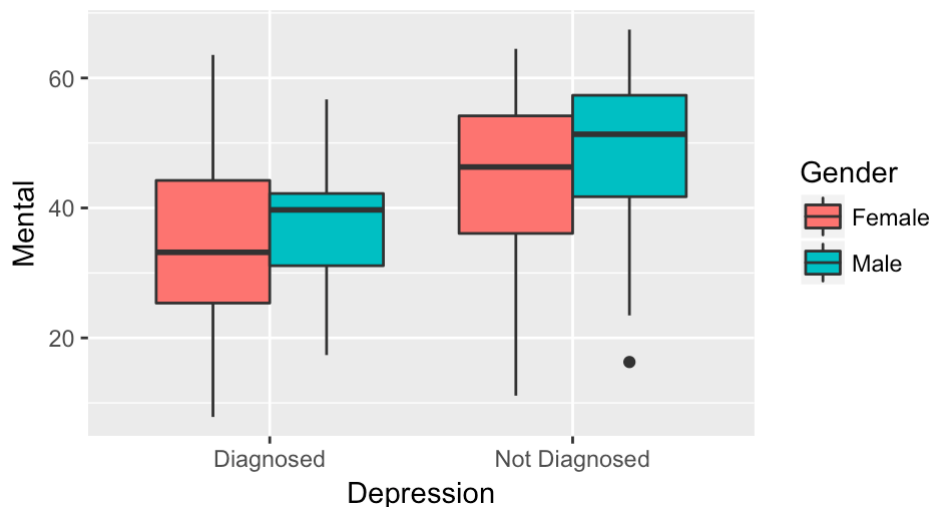
Appendix B: SAS Code

```
data final;
      infile '/folders/myfolders/final.txt';
      input DAV PCS MCS BECK PGEND AGE EDUCAT;
run;

proc format;
      value dav 0 = 'Not Diagnosed' 1 = 'Diagnosed';
      value pgend 0 = 'Female' 1 = 'Male';
run;

proc logistic data = final;
      class PGEND;
      model DAV = PCS MCS BECK PGEND AGE EDUCAT;
      format DAV dav.;
      format PGEND pgend.;
title `'All Variable Logistic Regression';
run;

proc logistic data = final;
      class PGEND;
      model DAV = PCS MCS BECK PGEND AGE EDUCAT/
                     selection = backward details;
      format DAV dav.;
      format PGEND pgend.;
title `'Stepwise Logistic Regression of Depression Data
(Backward)';
run;

proc logistic data = final;
      class PGEND;
      model DAV = PCS|MCS|BECK|PGEND|AGE|EDUCAT @2/
                     selection = forward details;
      format DAV dav.;
      format PGEND pgend.;
title `'Stepwise Logistic Regression of Depression Data
(Forward)';
run;

proc logistic data = final;
      class PGEND;
      model DAV = MCS BECK EDUCAT PGEND/ rsq lackfit plcl
plrl risklimits influence iplots;
      output out = finalo p = predprob;
      format DAV dav.;
      format PGEND pgend.;
title 'Final Logistic Regression Model for the Depression
Data';
run;
```

Appendix C: SAS Partial Output

`All Variable Logistic Regression

The LOGISTIC Procedure

Model Information	
Data Set	WORK.FINAL
Response Variable	DAV
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	401
Number of Observations Used	400

Response Profile		
Ordered Value	DAV	Total Frequency
1	Diagnosed	64
2	Not Diagnosed	336

Probability modeled is DAV='Diagnosed'.

Note: 1 observation was deleted due to missing values for the response or explanatory variables.

Class Level Information		
Class	Value	Design Variables
PGEND	Female	1
	Male	-1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	353.736	306.175
SC	357.727	334.116
-2 Log L	351.736	292.175

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.7940	1.5100	3.4238	0.0643
PCS		1	-0.0108	0.0139	0.6017	0.4379
MCS		1	-0.0492	0.0153	10.3000	0.0013
BECK		1	0.0666	0.0328	4.1083	0.0427
PGEND	Female	1	0.3351	0.1711	3.8354	0.0502
AGE		1	0.0137	0.0103	1.7473	0.1862
EDUCAT		1	0.1882	0.0619	9.2413	0.0024

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
PCS	0.989	0.963	1.017
MCS	0.952	0.924	0.981
BECK	1.069	1.002	1.140
PGEND Female vs Male	1.955	0.999	3.823
AGE	1.014	0.993	1.034
EDUCAT	1.207	1.069	1.363

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.2	Somers' D	0.563
Percent Discordant	21.8	Gamma	0.563
Percent Tied	0.0	Tau-a	0.152
Pairs	21504	c	0.782

`Stepwise Logistic Regression of Depression Data (Backward)

Backward Elimination Procedure

Class Level Information		
Class	Value	Design Variables
PGEND	Female	1
	Male	-1

Step 0. The following effects were entered:

Intercept PCS MCS BECK PGEND AGE EDUCAT

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
PCS	1	0.6017	0.4379
MCS	1	10.3000	0.0013
BECK	1	4.1083	0.0427
PGEND	1	3.8354	0.0502
AGE	1	1.7473	0.1862
EDUCAT	1	9.2413	0.0024

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.7940	1.5100	3.4238	0.0643
PCS		1	-0.0108	0.0139	0.6017	0.4379
MCS		1	-0.0492	0.0153	10.3000	0.0013
BECK		1	0.0666	0.0328	4.1083	0.0427
PGEND	Female	1	0.3351	0.1711	3.8354	0.0502
AGE		1	0.0137	0.0103	1.7473	0.1862
EDUCAT		1	0.1882	0.0619	9.2413	0.0024

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
PCS	0.989	0.963	1.017
MCS	0.952	0.924	0.981
BECK	1.069	1.002	1.140
PGEND Female vs Male	1.955	0.999	3.823
AGE	1.014	0.993	1.034
EDUCAT	1.207	1.069	1.363

Analysis of Effects Eligible for Removal

Effect	DF	Wald Chi-Square	Pr > ChiSq
PCS	1	0.6017	0.4379
MCS	1	10.3000	0.0013
BECK	1	4.1083	0.0427
PGEND	1	3.8354	0.0502
AGE	1	1.7473	0.1862
EDUCAT	1	9.2413	0.0024

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.4157	1.2814	7.1054	0.0077
MCS		1	-0.0470	0.0150	9.7943	0.0018
BECK		1	0.0736	0.0315	5.4463	0.0196
PGEND	Female	1	0.3500	0.1701	4.2352	0.0396
AGE		1	0.0157	0.00997	2.4716	0.1159
EDUCAT		1	0.1852	0.0612	9.1750	0.0025

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
MCS	0.954	0.926	0.983
BECK	1.076	1.012	1.145
PGEND Female vs Male	2.014	1.034	3.922
AGE	1.016	0.996	1.036
EDUCAT	1.203	1.068	1.357

Analysis of Effects Eligible for Removal			
Effect	DF	Wald Chi-Square	Pr > ChiSq
MCS	1	9.7943	0.0018
BECK	1	5.4463	0.0196
PGEND	1	4.2352	0.0396
AGE	1	2.4716	0.1159
EDUCAT	1	9.1750	0.0025

Step 2. Effect AGE is removed:

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.6410	1.1878	4.9439	0.0262
MCS		1	-0.0470	0.0150	9.7730	0.0018
BECK		1	0.0721	0.0315	5.2214	0.0223
PGEND	Female	1	0.3316	0.1695	3.8280	0.0504
EDUCAT		1	0.1785	0.0617	8.3609	0.0038

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
MCS	0.954	0.926	0.983
BECK	1.075	1.010	1.143
PGEND Female vs Male	1.941	0.999	3.772
EDUCAT	1.195	1.059	1.349

Note: No (additional) effects met the 0.05 significance level for removal from the model.

	Summary of Backward Elimination				
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	PCS	1	5	0.6017	0.4379
2	AGE	1	4	2.4716	0.1159
3	PGEND	1	3	3.8280	0.0504

`Stepwise Logistic Regression of Depression Data (Forward)

	Analysis of Maximum Likelihood Estimates				
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6582	0.1364	147.8250	<.0001

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
71.1523	21	<.0001

Analysis of Effects Eligible for Entry			
Effect	DF	Score Chi-Square	Pr > ChiSq
PCS	1	3.0780	0.0794
MCS	1	42.5047	<.0001
BECK	1	36.7771	<.0001
PGEND	1	7.7565	0.0054
AGE	1	0.7112	0.3990
EDUCAT	1	3.3201	0.0684

Step 1. Effect MCS entered:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	353.736	314.373
SC	357.727	322.356
-2 Log L	351.736	310.373

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
MCS	0.933	0.912	0.954

Step 2. Effect EDUCAT entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	353.736	310.594
SC	357.727	322.568
-2 Log L	351.736	304.594

Step 3. Effect BECK entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	353.736	307.253
SC	357.727	323.219
-2 Log L	351.736	299.253

Step 4. Effect PGEND entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

| Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	353.736	305.201
SC	357.727	325.159
-2 Log L	351.736	295.201

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.6410	1.1878	4.9439	0.0262
MCS		1	-0.0470	0.0150	9.7730	0.0018
BECK		1	0.0721	0.0315	5.2214	0.0223
PGEND	Female	1	0.3316	0.1695	3.8280	0.0504
EDUCAT		1	0.1785	0.0617	8.3609	0.0038

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
MCS	0.954	0.926	0.983
BECK	1.075	1.010	1.143
PGEND Female vs Male	1.941	0.999	3.772
EDUCAT	1.195	1.059	1.349

Final Logistic Regression Model for the Depression Data

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	353.736	305.201
SC	357.727	325.159
-2 Log L	351.736	295.201

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
MCS	1.0000	0.954	0.926	0.982
BECK	1.0000	1.075	1.010	1.144
EDUCAT	1.0000	1.195	1.064	1.355
PGEND Female vs Male	1.0000	1.941	1.017	3.873

Odds Ratios with 95% Profile-Likelihood Confidence Limits
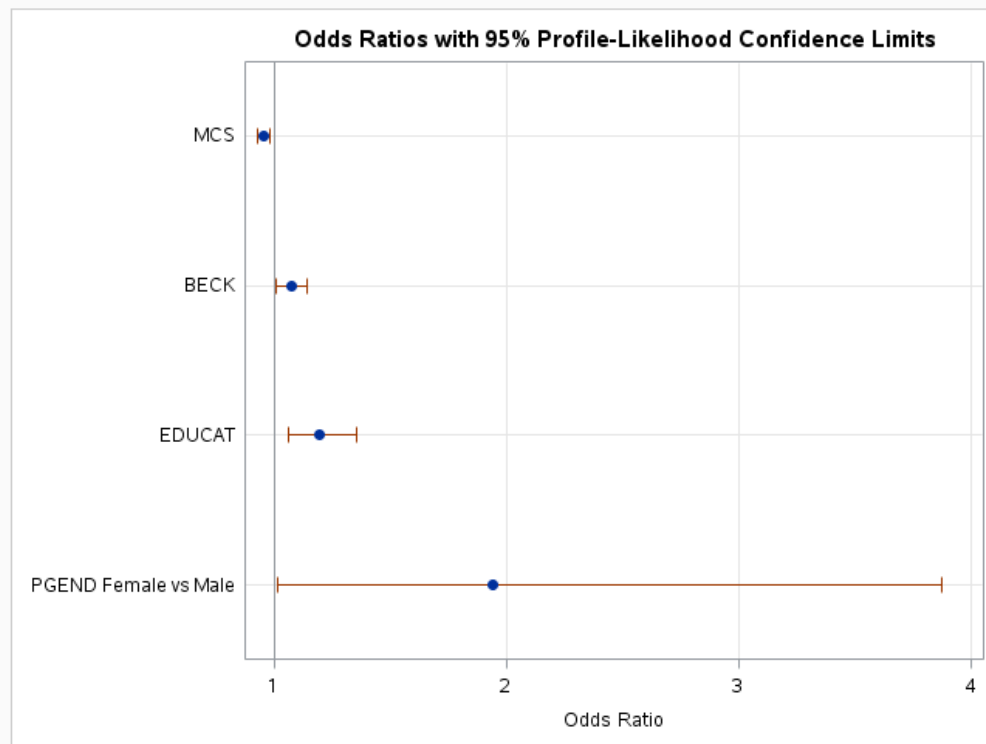
Partition for the Hosmer and Lemeshow Test					
		DAV = Diagnosed		DAV = Not Diagnosed	
Group	Total	Observed	Expected	Observed	Expected
1	40	2	1.02	38	38.98
2	40	1	1.76	39	38.24
3	40	2	2.26	38	37.74
4	40	2	3.02	38	36.98
5	40	2	3.83	38	36.17
6	40	6	4.94	34	35.06
7	40	11	6.45	29	33.55
8	40	9	8.58	31	31.42
9	40	10	12.04	30	27.96
10	40	19	20.11	21	19.89

Influence Diagnostics