Final Project

Lijing Xu

2018/3/13

### 1. Introduction.

###### Depression is a disease that affects millions of people around the world. It is the disease of the 21st century. But this was not always the case. Depression as an endemic problem only started from the 21st century.So that we are is going to analyze the relationship between depression and some other variables in primary care.To study factors related to the diagnosis of depression in primary care, 400 patients were randomly selected and the following variables were recorded: DAV, PCS, MCS, BECK, PGEND, AGE, and EDUCAT.

### 2. Material and Method

###### The dataset is included 400 patients with one response variable(diagnosed "1" and not diagnosed "0") and six explanatory variables.

###### 1. pcs ??? physical component of SF-36 measuring health care of the patient (a continuous variable)

###### 2. mcs ??? mental component of SF-36 measuring health status of the patient (a continuous variable)

###### 3. beck ??? the beck depression score (a continuous variable)

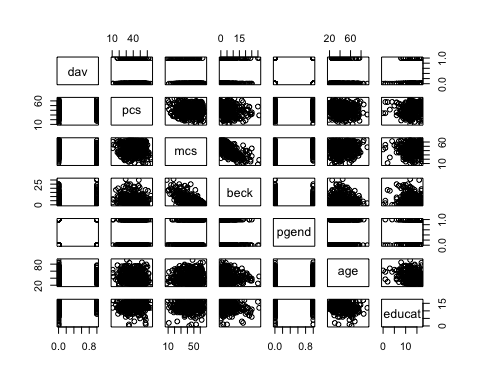
###### 4. pgend ??? patient???s gender (a categorical variable)

###### 5. age ??? patient???s age in years (a continuous variable)

###### 6. educat ??? total years of formal schooling

###### Because the response variable is categorical, in this report we will perform a multiple logistic regression analysis of this data using R. This includes estimation, hypothesis testing, model selection, odds ratios, residual analysis and diagnostics.

### 3. Results

 ###### I plotted this graph to look for which two variables that are almost linear to each other. I could not be 100 percent sure which two variables that are likely linear to each other by observing the graph; however, the "MCS" and "BECK" seems like have the highest collinearity.

##### Forward Selection

###### In the next step, I build the model with the forward selection method. The "step" function is the essential way to do it, which used AIC by default as the criterion. At each step, the function will compute AIC for each model with one additional explanatory variable and choose the one that lowers AIC the most. If none of the variables improves AIC, the current model will be 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=314.37  
## dav ~ mcs  
##   
## Df Deviance AIC  
## + educat 1 304.59 310.59  
## + pgend 1 307.23 313.23  
## + beck 1 307.44 313.44  
## + pcs 1 308.13 314.13  
## <none> 310.37 314.37  
## + age 1 309.42 315.42  
##   
## Step: AIC=310.59  
## dav ~ mcs + educat  
##   
## Df Deviance AIC  
## + beck 1 299.25 307.25  
## + pgend 1 300.44 308.44  
## + pcs 1 301.10 309.10  
## <none> 304.59 310.59  
## + age 1 302.92 310.92  
##   
## Step: AIC=307.25  
## dav ~ mcs + educat + beck  
##   
## Df Deviance AIC  
## + pgend 1 295.20 305.20  
## <none> 299.25 307.25  
## + age 1 297.27 307.27  
## + pcs 1 297.51 307.51  
##   
## Step: AIC=305.2  
## dav ~ mcs + educat + beck + pgend  
##   
## Df Deviance AIC  
## + age 1 292.78 304.78  
## <none> 295.20 305.20  
## + pcs 1 293.89 305.89  
##   
## Step: AIC=304.78  
## dav ~ mcs + educat + beck + pgend + age  
##   
## Df Deviance AIC  
## <none> 292.78 304.78  
## + pcs 1 292.18 306.18

##   
## Call:  
## glm(formula = dav ~ mcs + educat + beck + pgend + age, family = binomial,   
## data = patient)  
##   
## 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 \*   
## pgend -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

##### Backward Selection

###### Here our final model is the one with 'MCS', 'EDUCAT', 'BECK', 'PGEND' and 'AGE'. In the next step, I will begin with the full model and use the backward elimination method. The criterion is still AIC here and we stop when eliminating any variable in the model would increase AIC.

## Start: AIC=306.18  
## dav ~ pcs + mcs + beck + pgend + age + educat  
##   
## Df Deviance AIC  
## - pcs 1 292.78 304.78  
## - age 1 293.89 305.89  
## <none> 292.18 306.18  
## - pgend 1 296.23 308.23  
## - beck 1 296.29 308.29  
## - educat 1 302.51 314.51  
## - mcs 1 302.87 314.87  
##   
## Step: AIC=304.78  
## dav ~ mcs + beck + pgend + age + educat  
##   
## Df Deviance AIC  
## <none> 292.78 304.78  
## - age 1 295.20 305.20  
## - 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 + beck + pgend + age + educat, family = binomial,   
## data = patient)  
##   
## 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 \*   
## pgend -0.700031 0.340154 -2.058 0.03959 \*   
## age 0.015669 0.009967 1.572 0.11592   
## educat 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

##### Both Selection

###### The more ideal way is to check both adding and eliminating a variable at each step. We can do that similarly to the code for forward selection except the ???direction??? argument will be ???both???.

## 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=314.37  
## dav ~ mcs  
##   
## Df Deviance AIC  
## + educat 1 304.59 310.59  
## + pgend 1 307.23 313.23  
## + beck 1 307.45 313.45  
## + pcs 1 308.13 314.13  
## <none> 310.37 314.37  
## + age 1 309.42 315.42  
## - mcs 1 351.74 353.74  
##   
## Step: AIC=310.59  
## dav ~ mcs + educat  
##   
## Df Deviance AIC  
## + beck 1 299.25 307.25  
## + pgend 1 300.44 308.44  
## + pcs 1 301.10 309.10  
## <none> 304.59 310.59  
## + age 1 302.92 310.92  
## - educat 1 310.37 314.37  
## - mcs 1 348.20 352.20  
##   
## Step: AIC=307.25  
## dav ~ mcs + educat + beck  
##   
## Df Deviance AIC  
## + pgend 1 295.20 305.20  
## <none> 299.25 307.25  
## + age 1 297.27 307.27  
## + pcs 1 297.51 307.51  
## - beck 1 304.59 310.59  
## - educat 1 307.44 313.45  
## - mcs 1 310.58 316.58  
##   
## Step: AIC=305.2  
## dav ~ mcs + educat + beck + pgend  
##   
## Df Deviance AIC  
## + age 1 292.78 304.78  
## <none> 295.20 305.20  
## + pcs 1 293.89 305.89  
## - pgend 1 299.25 307.25  
## - beck 1 300.44 308.44  
## - educat 1 304.47 312.47  
## - mcs 1 305.30 313.30  
##   
## 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,   
## data = patient)  
##   
## 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 \*   
## pgend -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

###### The three model selections indicated that five variables are significant, include MCS EDUCAT BECK PEGND AGE which give us same model.

##   
## Call:  
## glm(formula = dav ~ mcs + beck + pgend + age + educat, family = binomial,   
## data = patient)  
##   
## 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 \*   
## pgend -0.700031 0.340154 -2.058 0.03959 \*   
## age 0.015669 0.009967 1.572 0.11592   
## educat 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

##### H0: the models does fitted

##### Ha: the models does not fitted

## ResourceSelection 0.3-2 2017-02-28

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: fit.both$y, fitted(fit.both)  
## X-squared = 6.8257, df = 8, p-value = 0.5556

###### As we can see the p-value = 0.3316 which is larger than the alpha(0.05), then we reject alternattive hypothesis and the model fits well. Next I'm going to use the wald test to get the confident interval:

##### Wald-type Confident Interval

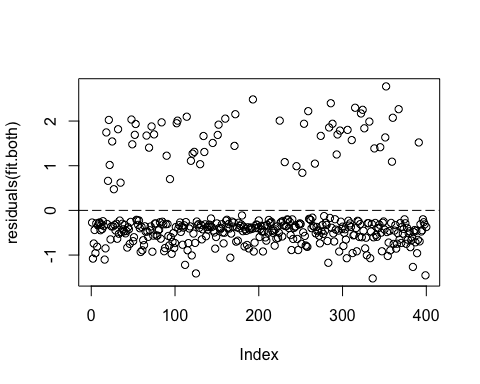
## [,1] [,2]  
## (Intercept) -5.540391541 -0.59098184  
## mcs -0.076396060 -0.01755664  
## educat 0.065376556 0.30508732  
## beck 0.011784200 0.13537278  
## pgend -1.366720544 -0.03334209  
## age -0.003865335 0.03520342

##### The Odd Ratio

## (Intercept) mcs educat beck pgend age   
## 0.04662182 0.95410996 1.20349755 1.07635301 0.49656975 1.01579245

###### I am 95% confident that increased one unit in mcs will make change in the odds of having diagnosed with depression around 0.9264496 to 0.9825968 times. The odd of having diagnosed with depression when X=1 is 0.9541 times as the odd when X=0. The other factor are same as MCS.

##### The Residual Plot

 ######As we can see, the plot show that there are few outlier because we just want residuals to be randomly scattered around 0 and between 2. There also is lots of patients that is confrom to the fitted modle, but only few is contracdict agains the fitted model.

### 4. Conclusion

###### In this report, I used the selecting model method to chose the fitted model. It's was analysis by choose the lowest AIC and the Hosmer-Lemeshow goodness(GOF) of fit test at 0.05 significance level. However, the Hosmer and Lemeshow goodness of fit test and Wald CI also support the null hypothesis test at 0.05 significant level. When I calculated the conditional odd ratios, the factor of the MCS equal to 0.9541, it means the odd of diagnosis of depression would decrease 0.9541 when the MCS increase by one unit.

###### In addition, according to the odds ratio, we get the effect of mental component score of depression and number of years of formal schooling are found to be significant. The predicted value ??(x) would decreases with increasing number of MCS and decreasing Beck depression score, age and number of years of formal schooling.