Calvin Makelky

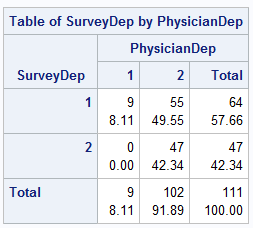
A01872013

HW3

**1.**

a)

i.



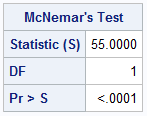
ii. McNemar’s Test

iii.

Ho: PhysicianDep and SurveyDep agree when success (depression) occurs

Ha: PhysicianDep and SurveyDep do not agree when success (depression) occurs

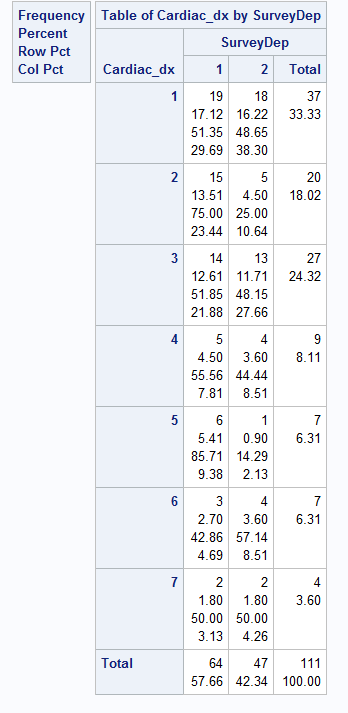
iv. p-value = < .0001



v. With the highly significant p-value, we conclude there is strong evidence that the Physician and survey patients filled out do not agree on whether patients have depression. There is a signficant difference between the two.

b)

i.

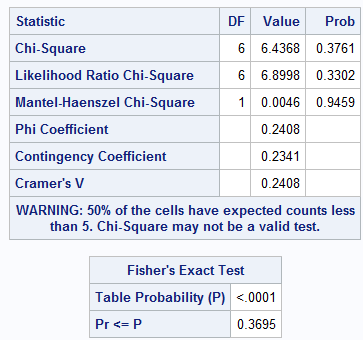


ii. Fisher’s Exact Test (expected cell frequency was not >5 for each cell)

iii. Ho: no association between depression and different diagnostic groups

Ha: association between depression and different diagnostic groups

iv. Fisher’s p-value = .1649

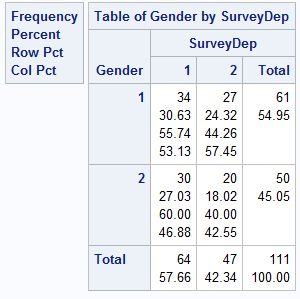


v.

With the insignificant p-value (0.367), we accept the null and say there is insufficient evidence to say there is any difference in depression rates between the seven different diagnostic groups in a general association.

c)

i.

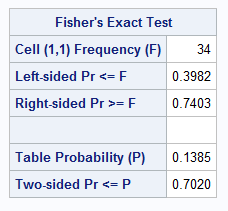


ii. Fisher’s Exact Test

iii. Ho: no general association between gender and depression

Ha: general association between gender and depression

iv. p-value = .0702



v.

Because the p-value is not below the alpha = .05 threshold, we fail to reject the null and conclude there is insufficient evidence to conclude there is a general association between gender and depression rates. No significant difference in depression rates between genders.

**3.**

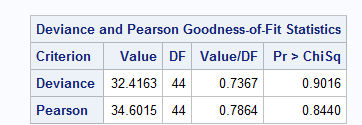
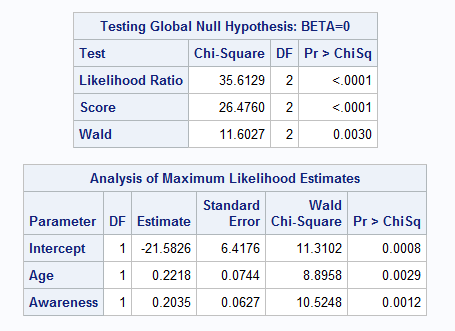
a)

logistic regression: Binary logit model using Fisher’s scoring

Age p –value = .0029

Awareness p-value = .0012

The highly significant p-values of the tests on null hypothesis that beta=0 indicates that we should reject that null, meaning that the probability of getting a flu shot is significantly affected by age and awareness. The odd ratio confidence intervals for age and awareness both contain one, so they are not statistically different from each other in their effect on getting the flu shot. The Deviance and Pearson Goodness of Fit table shows with insignificant p-values, that there is no evidence against the null that sigma equals 1. So there is no evidence of overdispersion.

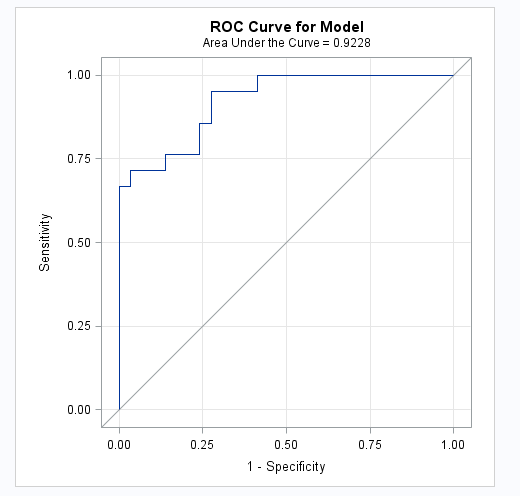
 

b)

With the significant p-value we can conclude that age and awareness are significant predictors of whether people get flu shots after getting a reminder in the mail.

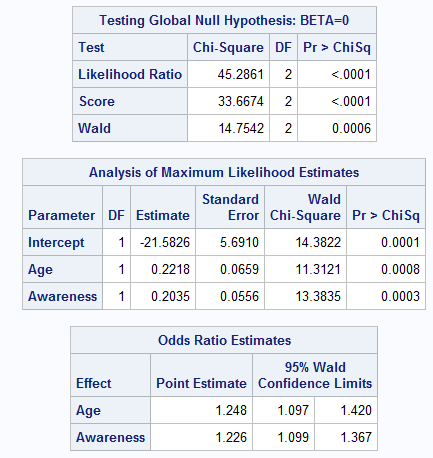
c)

The logistic model using age and awareness as predictors fits the data extremely well, evidenced by c and the area under the ROC curve equals .9228. This indicates very high concordance and predictive ability of the model.



d)

The odds ratio for a one-unit increase in the variable Age is 1.226 (exp(.02218) from the regression coefficients above the odds ratios for getting the flu vs not getting the flu. In other words, the odds of getting a flu shot are about 25% higher for every unit of increase in age, and the odds of getting a flu shot are about 23% higher for every unit increase in awareness. Neither confidence limits contain one, so this is another indicator both are significant factors in whether one gets the flu shot.



5.

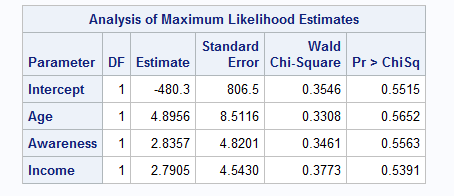
a)

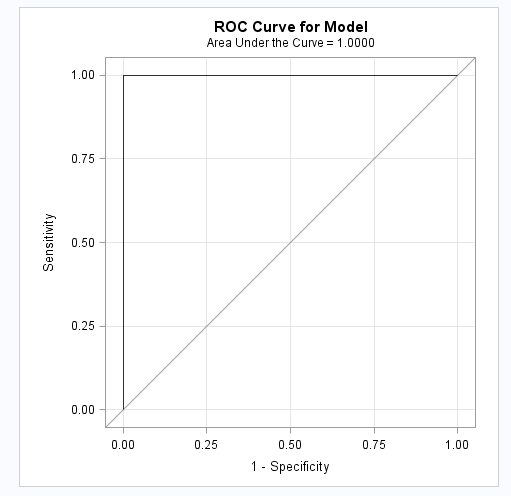
Age p-value: .5652

Awareness p-value: .5563

Income p-value: .5391

The c and area under the ROC curve equals 1.



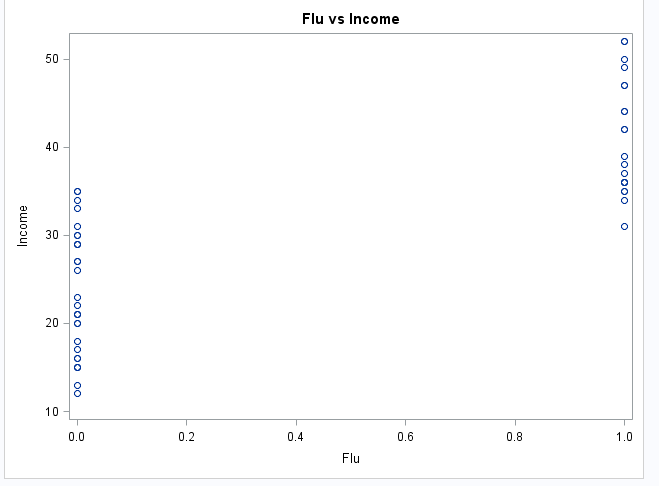
 

b)

This result differs from 3b in that age and awareness are not significant predictors for whether someone got a flu shot. The ROC curve in 3c is similar to this ROC curve in problem 5, but this one covers all the area under the curve that is possible whereas the one problem 3c does not. However, the maximum likelihood may not exist, so whether we can accept the result is questionable. There appears to be separation of points, so we need to fix this in order to really interpret the results.

c)

The plot below shows that income seems to be a very good predictor of whether someone would get a flu shot, with most of the people who got the flu shot having higher incomes than the highest income of all the people that didn’t get the flu shot. Also, there is almost a complete separation of points between those who got the shot and those who didn't in terms of income. This is what caused the separation of points in our model.



d)

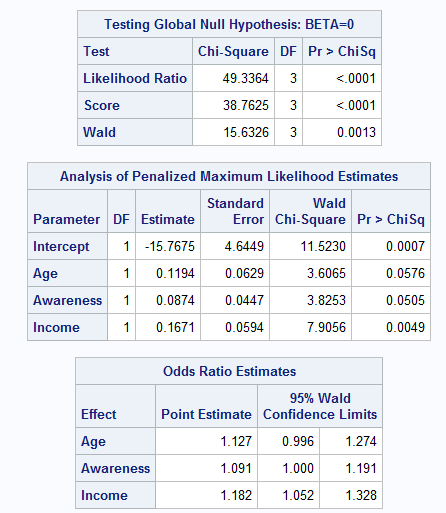
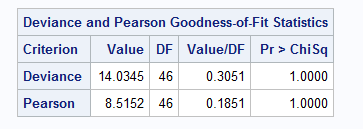
Logistic regression using FIRTH:

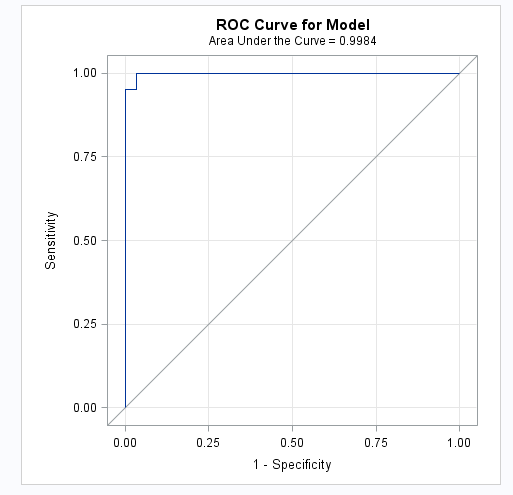
Age p-value: .0576

Awareness p-value: .0505

Income p-value: .0049

Only the p-value for income was significant, and only the odds ratio confidence interval did not contain one. Using this model, the odds of getting a flu increase by about 18% for every unit increase in income. The c and area under the ROC curve equals .9984. This is extremely high, meaning the model has amazing predictive ability and great concordance.



**Problem 1 code**

**proc** **freq** data=depression;

tables SurveyDep\*PhysicianDep / agree norow nocol;

title1 'McNeamars test';

**run**;

**proc** **freq** data=depression;

tables Cardiac\_dx\*SurveyDep / fisher;

**run**;

**proc** **freq** data=depression;

tables Gender\*SurveyDep / fisher;

**run**;

**Problem 3 code**

**proc** **import**

datafile="C:\Users\itcl\Desktop\flushot.csv"

dbms=csv out=flu replace;

getnames=yes;

datarow=**2**;

**run**;

**proc** **logistic** data=flu plots(only)=roc;

model Flu(event='1') = Age|Awareness / aggregate = (Age Awareness) scale=none ;

title1 'Logistic regression';

title2 '(accounting for covariate)';

**run**;

**Problem 5 code**

**proc** **logistic** data=flu plots(only)=roc;

model Flu(event='1') = Income / scale=none ;

title1 'Logistic regression';

title2 '(accounting for covariate)';

**run**;

**proc** **sgplot** data=flu;

scatter x=Flu y=Income ;

title1 ‘Flu vs Income’;

**run**;

**proc** **logistic** data=flu plots(only)=roc;

model Flu(event='1') = Age Awareness Income Age\*Awareness\*Income / firth aggregate = (Age Awareness Income) scale=Pearson ;

title1 'Logistic regression';

title2 '(accounting for covariate)';

**run**;