GLM HW1

Charlotte Li

8/18/2020

Part I Binary Outcome

```
setwd("~/Documents/nyu/1stGradSpring/Generalized Linear Model/dataset")
dat<-read.csv("RELIGION.csv")
dat$relschol <- as.numeric(dat$relschol=="yes")
dat$white <- as.numeric(dat$race=="white")
```

1

```
## probability of attending religious school
sum(dat$relschol==1)/sum(nrow(dat))
```

```
## [1] 0.1277955
```

```
## compute the corresponding odds
sum(dat$relschol==1)/sum(dat$relschol==0)
```

```
## [1] 0.1465201
```

2

```
table(dat$white,dat$relschol)
```

```
##
##       0   1
##   0  76  26
##   1 470  54
```

```
## probability of non-white attend religious school
26/102
```

```
## [1] 0.254902
```

```
## probability of white attend religious school
54/524
```

```
## [1] 0.1030534
```

3

```
# odds that non-white students attend religious school
26/76
```

```
## [1] 0.3421053
```

```
# odds that white students attend religious school
54/470
```

```
## [1] 0.1148936
```

4

the odds ratio that compares the odds of white versus non-white attend religious school

```
(54/470)/(26/76)
```

```
## [1] 0.3358429
```

5

Build a logistic regression to predict “relschol”using variables “white”, “attend” and “age”, treating the latter two as continuous variables. Report the odds ratio for variable “white”

```
summary(glm(formula=relschol~as.factor(white)+attend+age,data=dat,family="binomial"))
```

```
##
## Call:
## glm(formula = relschol ~ as.factor(white) + attend + age, family = "binomial",
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3670  -0.5638  -0.3961  -0.2686   2.5524
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.874156   0.684945  -1.276 0.201870
## as.factor(white)1 -0.941297   0.280942  -3.351 0.000807 ***
## attend             0.356197   0.128341   2.775 0.005513 **
## age               -0.045979   0.009039  -5.086 3.65e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 477.66  on 622  degrees of freedom
## Residual deviance: 426.85  on 619  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 434.85
##
## Number of Fisher Scoring iterations: 5
```

```
exp(-0.94)
```

```
## [1] 0.3906278
```

6

In one short sentence, explain the meaning of the odds ratio for “white” you reported in question 5: –The odds ratio for white students of attending religious school is 0.39 times the odds of non white students attending religious school, holding other variables constant.

7

Report the adjusted odds ratio comparing Non-white students versus White students based one the results from the model in question 5.

```
1/exp(-0.94)
```

```
## [1] 2.559981
```

8

Further extend the model in previous question by including a quadratic term of age, “agesq”. Run a logistic regression now with predictors: white, attend, age and agesq. Choose the answer that is best informed by the results of this model (as compared with the previous one)

```
dat$agesq<-dat$age*dat$age
summary(model8<-glm(formula=relschol~as.factor(white)+attend+age+agesq,data=dat,family="binomial"))
```

```
##
## Call:
## glm(formula = relschol ~ as.factor(white) + attend + age + agesq,
##     family = "binomial", data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.35846  -0.61149  -0.30745  -0.01421   3.07916
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -12.246178   2.374920  -5.156 2.52e-07 ***
## as.factor(white)1  -0.964465   0.293193  -3.290 0.001004 **
## attend              0.455887   0.135428   3.366 0.000762 ***
## age                 0.533997   0.117833   4.532 5.85e-06 ***
## agesq              -0.007178   0.001519  -4.726 2.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 477.66  on 622  degrees of freedom
## Residual deviance: 379.40  on 618  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 389.4
##
## Number of Fisher Scoring iterations: 8
```

The age of respondent has a quadratic (curvilinear) relationship with the log odds of attending religious school.

9

```

res <- glm(relschol~as.factor(white+attend) + age +agesq, data=dat,family=binomial(link="logit"))
newdat <- matrix(0, 2,4)
newdat[1,]<-c(0, 5, 45, 45^2)
newdat[2,]<-c(1,5,45,45^2)
newdat<-as.data.frame(newdat)
```

```
names(newdat)<-c("white","attend","age", "agesq")
pp <- predict(model8, newdata=newdat, type="response", se.fit=TRUE)
pp
```

```
## $fit
##         1         2 
## 0.3841089 0.1920710 
## 
## $se.fit
##          1          2 
## 0.06857122 0.03046408 
## 
## $residual.scale
## [1] 1
```

For those who attend religious services five days per month (attend=5) and age at 45, what is the predicted probability of having attended a religious school for non-white students: 0.384 with a standard error of 0.069 ; and the predicted probability for white students: 0.192 with a standard error of 0.030

10

ABCDEF

11

Run another model, include white, attend and age and agesq, treating "attend" as categorical variable. Based on AIC and BIC, which model fits the data better? Note a smaller AIC indicates better model fit considering the number of predictors used (model degrees of freedom) [hint: in R use AIC(res) to pull out the AIC value of a model. In Stata, use glm version AIC and BIC will be shown in the output]

```
summary(model11<-glm(relschol~white+factor(attend)+age+agesq,data=dat,family=binomial(link="logit")))
```

```
## 
## Call:
## glm(formula = relschol ~ white + factor(attend) + age + agesq, 
##     family = binomial(link = "logit"), data = dat)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -1.26835  -0.55089  -0.30578  -0.01285   2.99884  
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)    
## (Intercept)     -24.721734 917.294606  -0.027  0.97850    
## white            -0.986433   0.300398  -3.284  0.00102 ** 
## factor(attend)2  14.070280 917.291841   0.015  0.98776    
## factor(attend)3  13.804263 917.291834   0.015  0.98799    
## factor(attend)4  13.920903 917.291748   0.015  0.98789    
## factor(attend)5  15.176137 917.291698   0.017  0.98680    
## factor(attend)6  14.997963 917.291762   0.016  0.98695    
## age               0.526181   0.118325   4.447 8.71e-06 ***
## agesq            -0.007090   0.001525  -4.648 3.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 477.66  on 622  degrees of freedom
## Residual deviance: 372.01  on 614  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 390.01
##
## Number of Fisher Scoring iterations: 15
```

A. attend as continuous variable

12

Run a probit regression predicting the probability attending religious school using white, attend (as continuous variable), age and agesq. In one sentence, explain the meaning of the coefficient for "attend" in this probit model.

```
summary(model12<-glm(relschol~white+attend+age+agesq,data=dat,family=binomial(link="probit")))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = relschol ~ white + attend + age + agesq, family = binomial(link = "probit"),
##     data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.29456  -0.61975  -0.31630  -0.00186   3.14693
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.2797816  1.1736654  -5.351 8.77e-08 ***
## white       -0.5659651  0.1700767  -3.328 0.000876 ***
## attend       0.2499339  0.0719364   3.474 0.000512 ***
## age          0.2647927  0.0572947   4.622 3.81e-06 ***
## agesq       -0.0035403  0.0007291  -4.856 1.20e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 477.66  on 622  degrees of freedom
## Residual deviance: 380.24  on 618  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 390.24
##
## Number of Fisher Scoring iterations: 8
```

The latent variable of attending religious school is on average 0.25 standard deviation higher for those attending religious services 5 days per month than for those who do not attend

Part II Multinomial Regression

```
setwd("~/Documents/nyu/1stGradSpring/Generalized Linear Model/dataset")
drug<-read.csv("DRUGTEST.csv")
head(drug)
```

```
##                                  drugtest age educate fulltime gender      race
## 1          preemployment testing program  30      12      yes female non-white
## 2                     no testing program  24      14       no female non-white
## 3          preemployment testing program  32       9      yes   male non-white
## 4                     no testing program  37       7      yes   male non-white
## 5                     no testing program  21       6      yes   male non-white
## 6 preemployment & random testing program  28      12      yes female non-white
##       married         income south sales construc               othwork mjuser
## 1 not married            low    no   yes       no construction or sales     no
## 2 not married medium or high    no    no       no      other occupation    yes
## 3     married medium or high   yes    no       no      other occupation     no
## 4     married medium or high    no    no       no      other occupation     no
## 5 not married medium or high    no    no      yes construction or sales     no
## 6 not married medium or high   yes    no       no      other occupation     no
```

1

Reference Group: No Testing

2

BFCDAE

3

Based on the table above, we know that relative to no testing option, marijuana users are less likely to take the drug test (any group) than the non-users.-T

4

```
library(nnet)
```

```
## Warning: package 'nnet' was built under R version 4.0.2
```

```
drug$mjuser2<-relevel(factor(drug$mjuser),ref = "no")
drug$gender2<-relevel(factor(drug$gender),ref = "female")
drug$income2<-relevel(factor(drug$income),ref="low")
drug$south2<-relevel(factor(drug$south),ref = "no")
drug$construc2<-relevel(factor(drug$construc),ref="no")
drug$drugtest2<-relevel(factor(drug$drugtest),ref = "no testing program")
multi.fit<-multinom(drugtest2~mjuser2+age+educate+gender2+income2+south2+construc2,data=drug)
```

```
## # weights:  36 (24 variable)
## initial  value 12611.119803
## iter  10 value 9598.637922
## iter  20 value 9070.466499
## iter  30 value 8550.317852
## final  value 8535.237213
## converged
```

```
summary(multi.fit)
```

```
## Call:
## multinom(formula = drugtest2 ~ mjuser2 + age + educate + gender2 +
##     income2 + south2 + construc2, data = drug)
##
## Coefficients:
##                                        (Intercept) mjuser2yes           age
## preemployment & random testing program   -3.103282 -0.3524729  0.0109052791
## preemployment testing program            -2.143932 -0.2256318 -0.0045284552
## random testing program                   -4.213029 -0.1284840 -0.0007987289
##                                             educate gender2male
## preemployment & random testing program -0.009181182   0.6006046
## preemployment testing program           0.014431326   0.1826743
## random testing program                  0.045643149   0.1677110
##                                        income2medium or high south2yes
## preemployment & random testing program             0.8123096 0.7465224
## preemployment testing program                      0.6163895 0.2673895
## random testing program                             0.3766506 0.6104563
##                                        construc2yes
## preemployment & random testing program   -0.2490078
## preemployment testing program            -0.7994182
## random testing program                   -0.9412505
##
## Std. Errors:
##                                        (Intercept) mjuser2yes         age
## preemployment & random testing program   0.2313508 0.08937287 0.003337318
## preemployment testing program            0.2117882 0.08086701 0.003277593
## random testing program                   0.4336617 0.16379155 0.006707063
##                                           educate gender2male
## preemployment & random testing program 0.01156211  0.06478729
## preemployment testing program          0.01133410  0.05975907
## random testing program                 0.02412900  0.12242806
##                                        income2medium or high  south2yes
## preemployment & random testing program             0.1556821 0.06246605
## preemployment testing program                      0.1343519 0.05998320
## random testing program                             0.2578418 0.12025146
##                                        construc2yes
## preemployment & random testing program    0.1449877
## preemployment testing program             0.1848357
## random testing program                    0.4231678
##
## Residual Deviance: 17070.47
## AIC: 17118.47
```

5

Based on the model from previous question, predict the probabilities of having each of the four drug testing outcomes for a male who is 30 years old, has 12 year of education, low income, comes from south, and works in the construction who has used marijuana in part year.

```
newdrug <- matrix(c("yes","male","low","yes","yes",30,12),nrow = 1,ncol = 7,byrow = T)
newdrug<-as.data.frame(newdrug)
names(newdrug)<-c("mjuser2","gender2","income2","south2","construc2","age","educate")
```

```r
newdrug$age<-as.numeric(as.character(newdrug$age))
newdrug$educate<-as.numeric(as.character(newdrug$educate))
predict(multi.fit,newdrug[1,],"probs")
```

```
##                     no testing program preemployment & random testing program
##                             0.83008083                             0.09759940
##          preemployment testing program                 random testing program
##                             0.05682221                             0.01549756
```

```r
sum(predict(multi.fit,newdrug[1,],"probs"))
```

```
## [1] 1
```

```r
predict(multi.fit,newdrug[1,],"probs")
```

```
##                     no testing program preemployment & random testing program
##                             0.83008083                             0.09759940
##          preemployment testing program                 random testing program
##                             0.05682221                             0.01549756
```

For this individual,

the probability of no testing is 0.830

the probability of pre-employment testing & random testing is 0.098 the probability of random testing is 0.0155

the probability of pre-employment testing is 0.057