**Biostatistics 140.654: Applied Generalized Linear Models**

**Fourth Term, 2018**

Midterm

**Instructions**: This **is a closed book midterm. Do not consult with any other person or any materials not on the attached pages in answering the questions below.**

**Do any three of the four problems offered. Be sure to indicate which ones you have chosen below in the space below. If you do all 4 problems, we will only grade the first 3.**

**Problems chosen for evaluation; \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**.**

**Good luck.**

**Name (print)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

By signing my name, I agree to abide by the Johns Hopkins University School of Public Health Academic Code.

Signature\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**1.** Below find a data set of 125 infants showing whether an infant began breast-feeding within the first 2 weeks (1-yes, 0-no; jittered) against the child’s birth weight (kg). Vertical lines are at the quintiles of birth weight. 

Use these data to estimate the coefficients in a linear logistic regression model. Report:

i. logistic regression equation

ii. your approximate estimates of the coefficients in your model

iii. approximate predicted value at birth weight of 2 kg

iv. your findings in a sentence or two for a public health journal. Be numerate, eliminate jargon to the extent possible

2. Below find two 2x2 tables showing whether or not a person spent more than $1,000 on medical

|  |  |
| --- | --- |
| Age < 65 | Age >= 65 |
| mscd  y 0 1  0 5436 119  1 2028 323 | mscd  y 0 1  0 2647 280  1 1878 881 |

The scientific goals are: (1) to quantify the extent to which the chance of spending more than $1,000 is greater for persons with an mscd; and (2) to determine whether this mscd effect is the same for persons younger than 65 vs older.

i. Conduct a simple analysis to answer each of the two questions.

ii. Write a sentence or two to report your findings to a public health audience. Be numerate. Avoid jargon.

In the exam appendices A and B, find results of two logistic regression models fit to a subset of the National Medical Expenditures Survey data. For both models, presence (1) or absence (0) of a lung cancer type disease (LC) is the outcome. There were 223 LC’s out of 11,684 people. The predictors of interest are indicator variables of smoking history: never smoked, former smoker, or current smoker. These results were obtained using programs provided in Appendix B.

3. Using the figures and regression results, address the following two questions. Be numerate. Don’t use unnecessary jargon.

***How do the rates of LC differ across the three smoking groups among persons of a similar age; does the evidence support the hypothesis that the effects of smoking do not vary by age.***

4. The two models in Question 3 were used to predict whether a person would have a LC group disease or not. The predictions were made using the same data as were used to fit the models and then again using 10-fold cross validation. The four (2 models x with and without CV) Receiver Operator Characteristic (ROC) Curves are shown in Appendix B. The AUC’s are: Model 1 – no CV =0.796; Model 1 – with CV =0.793; Model 2 – no CV =0.802; Model 2 – with CV =0.788.

In a brief paragraph, explain for a public health person what these curves show and why they are so similar to one another.

**Bio654-2018 Midterm Exam Appendix B of Tables, R Code**

**Problem 3 Table**

**glm(formula = lc5 ~ factor(smoke.status) + age, family = binomial(),**

**data = d)**

Estimate SE Z value Pr(>|z|)

Intercept -3.37148 0.11568 -29.144 <2e-16

former -0.01536 0.15261 -0.101 0.92

never -2.19833 0.24516 -8.967 <2e-16

age 0.06604 0.00616 10.720 <2e-16

Null deviance:2207.3 on 11683df; Residual deviance: 1953.7 on 11680 df

**lrtest(reg1,reg2)**

Likelihood ratio test

Model 1: lc5 ~ factor(smoke.status) + age

Model 2: lc5 ~ factor(smoke.status) \* ns(age, 3)

#Df LogLik Df Chisq Pr(>Chisq)

1 4 -976.86

2 12 -965.60 8 22.508 0.004057 \*\*

**Problems 3 and 4 R Code**

# wrangle NMES data set

nmes.use=data.frame(nmes[,c("lastage","male","RACE3","totalexp","lc5","chd5","beltuse","educate","marital","povstalb","eversmk","current")])

nmes.use$eversmk = ifelse(nmes.use$eversmk==".",NA,nmes.use$eversmk)

nmes.use$smoke.status=ifelse(nmes.use$eversmk=="1",ifelse(nmes.use$current=="0","former","current"),"never")

d=nmes.use[complete.cases(nmes.use),]

#

d$age=d$lastage-65

d$race=d$RACE3

#

# fit two logistic regressions with lung cancer disease group presence/absence as the outcome and smoking status (never, former, current) and age as the main predictors.

# reg1 does not take into account SES variables; reg2 does;

# reg1 has a simpler dependence on smoking status and age than reg2

#

reg1=glm(data=d, lc5~ factor(smoke.status)+age,family=binomial())

fit1=predict.glm(reg1, type="response")

o1=order(fit1)

#

reg2=glm(data=d, lc5~ factor(smoke.status)\*ns(age,3),

family=binomial())

fit2=predict.glm(reg2,type="response")

o2=order(fit2)

#

# make plots of predicted values against age for both models and weatherman plots to compare models 1 and 2 against empirical frequencies

# because there are so many points, we take a random subset of 1000 people for plotting in model checking; this makes printing results faster

set.seed(1234)

select=sample(1:nrow(d),1000)

par(mfrow=c(1,2))

#

plot(d$lastage[select],fit1[select],xlim=c(30,100),xlab="Age (Years)",ylim=c(0,.2),ylab="Probability",main="LC Predictions", pch="o",col="red")

points(d$lastage[select],fit2[select],pch="\*",col="blue")

text(x=50,y=.18,"ooo - Model 1",col="red"); text(x=50,y=.16,"\*\*\* - Model 2",col="blue")

#

plot(fit1,jitter(d$lc5),pch=".",xlab="Predicted Probabilities",ylab="Observed Frequencies",main="LC Weatherman Plots")

points(fit1[o1][select],predict(lm(d$lc5~ns(fit1,3)))[o1][select],col="red",pch="o")

points(fit2[o2][select],predict(lm(d$lc5~ns(fit2,3)))[o2][select],col="blue",pch="\*")

abline(0,1); text(x=.15,y=.60,"ooo - Model 1",col="red")

text(x=.15,y=.50,"\*\*\* - Model 2",col="blue")

#

# summarize both regressions; compare them with a likelihood ratio test

#

summary.glm(reg1);summary.glm(reg2);lrtest(reg1,reg2)

#

# this function does k-fold cross-validation for logistic regression

#

CVlogistic=function(

data, # dataframe with complete cases only

formula, # regression equation

k # number of subsets of approximately equal size

)

{

#

# split the data into k clusters

#

group.ind=floor(runif(nrow(d))\*k)

sets=split(1:nrow(d),group.ind)

cvfit=NA

for (j in 1:k) {

set.ind=sets[[j]]

cvfit[set.ind]=predict.glm(glm(data=d[-set.ind,],

formula=formula,family=binomial()),newdata=d[set.ind,],type="response")

}

return(fit.cv=cvfit)

}

#

#

# calculate the roc curves and auc values without and with cross-validation for the two models

#

roc1=roc(d$lc5,fit1)

roc2=roc(d$lc5,fit2)

roc1.cv.pred=CVlogistic(d,formula = "lc5~ factor(smoke.status)+age",10)

roc1.cv=roc(d$lc5,roc1.cv.pred)

roc2.cv.pred=CVlogistic(d,"lc5~ factor(smoke.status)\*ns(age,3)",10)

roc2.cv=roc(d$lc5,roc2.cv.pred)

#

# plot the 4 ROC curves and print out their AUC values

#

plot.roc(roc1,main="ROC Curve for Logistic Regression to Predict Lung Cancer Group",lty=1,col="blue")

#

lines(roc2,col="red",lty=1)

lines(roc1.cv,col="blue",lty=2)

lines(roc2.cv,col="red",lty=2)

#

text(x=.4,y=.2,"Model 1 No CV",col="red")

text(x=.4,y=.16,"Model 1 --- CV",col="red")

text(x=.4,y=.12,"Model 2 No CV",col="blue")

text(x=.4,y=.08,"Model 2 --- CV",col="blue")

print(paste("auc.reg1=",round(roc1$auc,3),sep=""))

print(paste("auc.reg2=",round(roc2$auc,3),sep=""))

print(paste("auc.cv.reg1=",round(roc1.cv$auc,3),sep=""))

print(paste("auc.cv.reg2=",round(roc2.cv$auc,3),sep=""))