Categorical Analysis on Coronavirus-2019 data in Florida

Abstract: In this step of term project, both Poisson and negative binomial models are fitted based on the dataset of Coronavirus-2019 in Florida, containing about 22,000 observations. The response is a binary categorical variable presented by death status, other explanatory variables include age, gender, days after the first infected case. From the criteria of AIC, Deviance and AUC, logit, probit, Poisson and negative binomial models fit the data well. According to the AIC, the probit model is the best; according to the AUC, the negative model is slightly better than other ones.
Keywords: logistic regression; probit; categorical analysis; Poisson regression; negative binomial regression.

1. Introduction

A dataset of Coronavirus-2019 updated by April, 15th, 2020 is used in the data analysis, containing 21890 observations of infected patients in Florida. The response variable is a categorical variable, represented by death status. It equals to 1 if death status is 'Yes' after getting infected with Coronavirus-2019. Otherwise, it equals to 0 with only being infected but still alive. The explanatory variables consist of age, gender, days after the first case on March 2nd, 2020. Age and days after the first case (briefly named as days) are numerical variables while gender is a categorical variable. The goals of this step are to fit the Poisson and negative binomial regression model, and compare four models (logit, probit, Poisson and negative binomial) through AIC, Deviance, AUC, etc.

2. Poisson Regression Model

Through the backward elimination, the significant explanatory variables containing age, days, gender and the quadratic term of days are selected among age, days, gender, the quadratic terms of age and days, two-way and three-way interaction terms, with AIC= 4104.1 and Deviance=2936.1. However, the anova test shows the quadratic term of days is not significant through LR test chi-square=-0.03117, df=1, p-value=1 that fail to reject the null hypothesis that the model only contains age, days and gender. Finally, the fitted model equation is shown as below.

log(death)= -6.971681 + 0.084142 * Age+ 0.513622 * Gender(Male) - 0.074541*DAYS.

4. Negative Binomial Regression Model

In the same fashion of variable selection above, at first, age, days, gender and the quadratic term of days are considered as significant variables. Furthermore, the anova test with LR test chi-square=27.815, p-value=2.34e-06 << 0.05, indicates the null hypothesis that the model doesn't contain the quadratic term of days is rejected. Then, the fitted model equation is shown as below with AIC=4080.6, Deviance=2905.7.

log(death)= -9.6504962 + 0.0846064 * Age + 0.5132698 * Gender(Male) + 0.1204673 * DAYS - 0.0033714 * DAYS*DAYS.

5. Analysis and Conclusions

In this section, comparing the performance of the logit, probit, Poisson and negative binomial regression models is illustrated. Their AIC, Deviance, AUC are displayed as below, respectively.

logit model: AIC=4006.7, Deviance=3996.7, AUC=0.8833.

probit model: AIC=4002.6, Deviance=3990.6, AUC=0.8836.

poisson model: AIC=4104.1, Deviance=2936.1, AUC=0.8827.

negative binomial model: AIC= 4080.6, Deviance= 2905.7, AUC=0.8850.

Therefore, we can conclude that four models fit the data very well. From the aspect of AIC, the probit model is the best. According to the AUC, the negative binomial regression model is slightly better than others.

Appendix

```
##STA5505_ term project

#################poisson model##########
poisson.reg=glm(Death~Age+Gender+Days_after_first_case

+Age:Gender+Gender:Days_after_first_case+Age:Days_after_first_case+Age:Days_after
_first_case:Gender+I(Age^2)+I(Days_after_first_case^2),
family=poisson(link="log"),data=cov2019)
summary(poisson.reg)
step(poisson.reg,test="Chisq")
drop1(poisson.reg,test="Chisq")
#age,gender,days,days^2 are kept, AIC=4105.7,Deviance=2925.7

poisson.reg1=glm(Death~Age+Gender+Days_after_first_case,
        family=poisson(link="log"),data=cov2019)
summary(poisson.reg1) #AIC: 4104.1, Deviance=2936.1

poisson.reg2=glm(Death~Age+Gender+Days_after_first_case+I(Days_after_first_
case^2),family=poisson(link="log"),data=cov2019)
summary(poisson.reg2) #AIC: 4106.2,deviance=2936.2
anova(poisson.reg1,poisson.reg2)
1-pchisq(-0.031174,1) # fails to reject H0,that's to say, poisson.reg1 is better.
#########negative binomial########
library(MASS)
neg.bio=glm.nb(Death~Age+Gender+Days_after_first_case,data=cov2019)
summary(neg.bio)

neg.bio1=glm.nb(Death~Age+Gender+Days_after_first_case
+Age:Gender+Gender:Days_after_first_case+Age:Days_after_first_case+Age:Days_after
_first_case:Gender+I(Age^2)+I(Days_after_first_case^2),data=cov2019)
summary(neg.bio1)
#AIC: 4082.1,deviance: 2895.8  #I(Days_after_first_case^2) is significant
step(neg.bio1,test="Chisq")   #Deviance: 2896         AIC: 4082
drop1(neg.bio1,test="Chisq")

neg.bio2=glm.nb(Death~Age+Gender+Days_after_first_case+I(Days_after_first_c
ase^2),data=cov2019)
summary(neg.bio2) #AIC: 4080.6,deviance: 2905.7
anova(neg.bio,neg.bio2)
```

#LR test:chi-square=27.81515,p-value=2.343688e-06, reject H0,so neg,bio2 is better

```
library(pROC)
#AUC of poisson model
rocplot.poisson=roc(Death~fitted(poisson.reg1),data=cov2019)
plot.roc(rocplot.poisson,legacy.axes=TRUE)
auc(rocplot.poisson)
#Area under the curve: 0.8827
#log(death)=-6.971681+0.084142*Age+0.513622*Gender(Male)-0.074541*DAYS

#AUC of negative binomial model
rocplot.neg.bio2=roc(Death~fitted(neg.bio2),data=cov2019)
plot.roc(rocplot.neg.bio2,legacy.axes=TRUE)
auc(rocplot.neg.bio2)
#Area under the curve: 0.885
#log(death)=-
9.6504962+0.0846064*Age+0.5132698*Gender(Male)+0.1204673*DAYS-
0.0033714*DAYS*DAYS

#conclusion: logit, probit, Poisson, negative binomial models
# logit model:   AIC=4006.7, Deviance=3996.7, AUC=0.8833
# probit model:  AIC=4002.6, Deviance=3990.6, AUC=0.8836
# poisson model: AIC=4104.1, Deviance=2936.1, AUC=0.8827
# negative binomial model: AIC= 4080.6, Deviance= 2905.7, AUC=0.8850
```