**OM386 Advanced Data Analytics in Marketing**

**Assignment 3**

**Due: March 23rd, 11:59pm**

**Name: Mahika Bansal (mb62835)**

**Count Data Analysis for Shopping Mall Visits**

In this exercise, we will apply regression models for count data, including a Poisson log-linear model and a negative binomial model to analyze a data set on the shopping mall visitation frequencies. The goal is to evaluate whether target marketing is effective in attracting consumers to visit the shopping mall.

Please download the data file "Mall\_visit.csv" from Canvas. In this data set, "customerID" is for 500 customers who have downloaded and used a mobile app through which the shops in the mall sends target marketing messages. The data track each customer for 50 weeks, so there are 50 observations for each ID. "Visit" is the number of visits to the mall in a week; "Discount" is an index of various discounts offered by the mall; "Target" is a dummy variable which indicates whether a customer receives a targeting message; "Distant" is the distance from the customer's residence to the mall; "Income" is the customer's estimated income and "Gender" is the customer's gender (1 for female).

1). Use the function glm( ) to run the Poisson log linear model regression

*log(λit )= β0 + β1×Discount +* *β2×Target + β3×Income+ β4×Distant + β5×Gender*

Copy and paste the results here. Check the estimates of *β*1, *β*2*, β*3, *β*4, *β*5. Are they statistically significant? Please interpret this regression coefficients and calculate the AIC of this model.

Ans.

df <- fread('Mall\_visit.csv',stringsAsFactors = TRUE)

df$customerID <- as.factor(df$customerID)

glm1 = glm(Visit~Discount+Target+Income+Distant+Gender,data=df, family=poisson)

summary(glm1)

Text

Description automatically generated with medium confidence

All coefficients except Target are statistically significant and signify change in visits per unit the associated variable i.e. Discount, Target, Income, Distant, Gender.

The AIC of the model is :

AIC(glm1)

[1] 45274.32

2).Next, we will allow each individual customer to have a different intercept

*Log(λit )= β0 +ζi + β1×Discount + β2×Target + β3×Income+ β4×Distant + β5×Gender,*

where *ζi*  is be a random effect (500 of them) grouped by customerID. Run this regression using the glmer( ) function in the package "lme4" Copy and paste the results here.

Check the estimates of *β*1, *β*2*, β*3, *β*4, *β*5. Are they statistically significant? Please also calculate the AIC of this regression model.

Ans.

glmer2=glmer(Visit~(1|customerID)+Discount+Target+Income+Distant+Gender,data=df, family=poisson)

summary(glmer2)

Text

Description automatically generated

All the coefficients except Target and Gender, i.e. Discount, Income & Distant are statistically significant.

AIC(glmer2)

[1] 44695.9

3). We will also fit the negative binomial model for the count data. Let the mean of the negative binomial distribution be

*log(λit )= β0 + β1×Discount + β2×Target + β3×Income+ β4×Distant + β5×Gender,*

You can run this regression using the glm.nb( ) function in the package "MASS". Copy and paste the results here

Check the estimates of *β*1, *β*2*, β*3, *β*4, *β*5. Are they statistically significant? Please also calculate the AIC.

Based on the AIC's of the models in (1), (2) and (3), which is the best model for the data?

Ans.

glmnb3 = glm.nb(Visit~Discount+Target+Income+Distant+Gender,data=df)

summary(glmnb3)

Text

Description automatically generated

All coefficients except Target are statistically significant.

AIC(glmnb3)

[1] 45248.16

Comparing AIC, the 2nd model performs better than the other 2.

4). For the model in (2), use the MCMCpack function MCMChpoisson() to estimate the same parameters with Bayesian estimation. The model only has a random intercept, so you can specify random=~1 and r=2, R=1. Set burnin=10000, mcmc=20000 and thin=20. Copy and paste the Bayesian estimation results of the fixed effects in the model using summary("*yourBayesianModelName"*$mcmc[,1:6]). From the Bayesian posterior intervals, are the fixed effects significant at the 5% level?

Ans.

mcmc4 = MCMChpoisson(fixed = Visit~Discount+Target+Income+Distant+Gender,

data=df, random = ~1, group = "customerID",burnin = 10000,

r = 2, R=1, thin=20, mcmc=20000)

summary(mcmc4$mcmc[,1:6])

A screenshot of a computer

Description automatically generated with medium confidence

All fixed effects are statistically significant.

**Logistic and C-log-log Regressions for Discrete Hazard Models**

In this exercise, we will use the logit and cloglog links in the glm( ) function for binary data to estimate discrete hazard models. The data file is “Papertowel\_repurchase.csv” on Canvas. For 500 consumer households, we track their paper towel purchase incidences in 52 weeks. The dataset has the following variables:

|  |  |
| --- | --- |
| consumerID | The ID of the customer |
| papertowel | Whether the household buys paper towel in that week{1 = Yes, 0 = No} |
| week | A weekly time period indicator |
| price | The price of paper towel in that week |
| feature | Whether paper towel is a featured product of the supermarket {1 = Yes, 0 = No} |
| famsize | The size of the household |

The exercise is to study the effects of time, price, feature and household size on the hazard of buying paper towel. On the time dimension, the hazard of buying paper towel is considered to be “renewed” after a purchase and is seasonal based on the calendar time. Therefore the hazard function is a function of both the time intervals between purchases and the calendar time.

4). Use read.csv( ) to read the data into R as a data frame. Create a seasonal indicator variable “season” too as follows:

papertowel.data = read.csv("Papertowel\_repurchase.csv", header=T)

papertowel.data$season = as.factor(ceiling(papertowel.data $week/13))

Create a new variable in the data frame called "interval", which registers the number of weeks since the previous order as we discussed in the class. You can modify the R code on Page 4 of Lecture 7 to calculate the “interval” variable. Please paste your code here.

Ans.

papertowel.data = read.csv("Papertowel\_repurchase.csv", header=T)

papertowel.data$season = as.factor(ceiling(papertowel.data $week/13))

interval = c()

for (i in 1:500){

ppr.twl.i = papertowel.data[papertowel.data$consumerID == i,]

interval.i = rep(0,52)

sB = 0

for(t in 1:52){

sB = sB+1

interval.i[t] = sB

if(ppr.twl.i$papertowel[t] == 1) sB =0

}

interval = c(interval,interval.i)

}

papertowel.data$Interval = interval

5). Estimate the following logistic regression model for the hazard function using the R function glm( )

*log*(*λi*(*t*)*/*(*1- λi*(*t*)) = **0 + **1×*Intervalit* + **2×*Seasonit +*3×*Priceit +*4×*Featureit*

*+*5×*Famsizei*

And paste results here. How do you interpret **1, **2*, *3, **4, **5? Are they statistically significant? Please calculate the AIC of this model.

Ans.

glm\_surv = glm(papertowel~Interval+season+price+feature+famsize,data=papertowel.data, family=binomial(link = "logit"))

summary(glm\_surv)

Table

Description automatically generated

Interval, season3, season4 and price are statistically significant. The coefficients signify the change in paper towels wrt the coefficients.

AIC(glm\_surv)

[1] 20254.74

6). Estimate the following cloglog regression model for the hazard function using the R function glm( )

*log*(*-log*(*1- λi*(*t*)) = **0 + **1×*Intervalit* + **2×*Seasonit +*3×*Priceit +*4×*Featureit*

*+*5×*Famsizei*

And paste results here. How do you interpret **1, **2*, *3, **4, **5? Are they statistically significant? Please calculate the AIC of this model.

Ans.

glm\_surv2= glm(papertowel~Interval+season+price+feature+famsize,data=papertowel.data, family=binomial(link = "cloglog"))

summary(glm\_surv2)

Table

Description automatically generated

Interval, season3, season4 and price are statistically significant; interpretation of the coefficients remains the same.

AIC(glm\_surv2)

[1] 20240.32

7). Estimate the following cloglog regression model with a random effect for the intercept in the hazard function using the R function glmer( )

*log*(*-log*(*1- λi*(*t*)) = **0i + **1×*Intervalit* + **2×*Seasonit +*3×*Priceit +*4×*Featureit*

*+*5×*Famsizei*

And paste results here. Please calculate the AIC of this model. Based on the AIC's of the models in (5), (6) and (7), which is the best model for the data?

Ans.

glm\_surv3 = glmer(papertowel~(1|consumerID) + Interval+season+price+feature+famsize,data=papertowel.data, family=binomial(link = "cloglog"))

summary(glm\_surv3)

Text

Description automatically generated

Interval, season3, season4 and price are statistically significant; interpretation of the coefficients remains the same.

AIC(glm\_surv3)

[1] 20240.64

Based on AIC, model in 6) performs best, being slightly better than model in 7).