**OM386 Marketing Analytics II**

**Assignment 5**

**Due: April 8th, 11:59pm**

**Khyathi Balusu**

**Kb42582**

**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 to estimate discrete Hazard models. The data file is “HHonors\_booking.csv” on Canvas. For 400 Hilton HHonors members, we have the following variables:

|  |  |
| --- | --- |
| customer ID | The ID of the customer |
| Booking | Whether the customer books a Hilton hotel room in that week{1 = Yes, 0 = No} |
| Week | A weekly time period indicator |
| Price | The average price of hotel rooms in that week |
| Promotion | Whether a promotion email is send to the customer in that week {1 = Yes, 0 = No} |
| Income | The income level of the customer |
| Gender | Gender indicator {1 = Male, 0 = Female} |

The exercise it to study the effects of time, price and promotion on the hazard of booking a hotel room for each customer. The model also control for the customer's demographics including income and gender. The hazard of booking a hotel is considered to be "renewed" after a customer books a hotel; i.e., the baseline hazard *λ*0(*t*) is reset the *λ*0(*t*+1)= λ0(1) if the customer books a hotel in period (week) *t*.

1). Use read.csv( ) to read the data into R as a data frame. Create a new variable in the data frame called "Interval", which records the number of weeks since the previous hotel booking as we discussed in the class, using the following R code.

**hotel = read.csv("HHonors\_booking.csv", header=T)**

**interval = c( )**

**for(i in 1:400) {**

**hotel.i = hotel[hotel$customerID==i,]**

**interval.i = rep(0, 50)**

**sinceBooking = 0**

**for(t in 1:50) {**

**sinceBooking = sinceBooking + 1**

**interval.i[t] = sinceBooking**

**if (hotel.i$Booking[t] == 1) sinceBooking = 0**

**}**

**interval = c(interval, interval.i)**

**}**

**hotel$Interval = interva**l

**Code:**

**hotel =** read.csv**("HHonors\_booking.csv", header=T)**str**(hotel)**

**## 'data.frame': 20000 obs. of 7 variables:  
## $ customerID: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Booking : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Week : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Price : int 224 230 148 208 195 169 228 113 232 275 ...  
## $ Promotion : int 0 1 1 1 0 0 1 1 1 1 ...  
## $ Income : num 83.6 83.6 83.6 83.6 83.6 83.6 83.6 83.6 83.6 83.6 ...  
## $ Gender : int 0 0 0 0 0 0 0 0 0 0 ...**

**hotel**$**customerID =** as.factor**(hotel**$**customerID)  
  
  
interval =** c**( )**for**(i** in **1**:**400) {  
 hotel.i = hotel[hotel**$**customerID**==**i,]  
 interval.i =** rep**(0, 50)  
 sinceBooking = 0** for**(t** in **1**:**50) {  
 sinceBooking = sinceBooking** + **1  
 interval.i[t] = sinceBooking** if **(hotel.i**$**Booking[t]** == **1) sinceBooking = 0  
 }  
 interval =** c**(interval, interval.i)  
}  
hotel**$**Interval = interval**head**(hotel)**

## customerID Booking Week Price Promotion Income Gender Interval  
## 1 1 0 1 224 0 83.6 0 1  
## 2 1 0 2 230 1 83.6 0 2  
## 3 1 0 3 148 1 83.6 0 3  
## 4 1 0 4 208 1 83.6 0 4  
## 5 1 0 5 195 0 83.6 0 5  
## 6 1 0 6 169 0 83.6 0 6

2). Estimate the following logistic regression model using the R function glm( )

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

+ **4×*Incomei +*5×*Genderi*

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

Next, we will estimate the model:

*log*(*λi*(*t*)*/*(*1- λi*(*t*)) = **0 + **1×*Intervalit* + **2×*Intervalit2* + **3×*Priceit +*4×*Promotionit*

+ **5×*Incomei +*6×*Genderi*

Use poly(Interval, 2) in the glm() function to represent **1×*Intervalit*+**2×*Intervalit2* in this model. Are **1, **, **6 still statistically significant? Please calculate the AIC and BIC of this model.

**Code & Summary:**

**# Logistic Regression with linear coefficients for Interval  
hotel.logit1 =** glm**(Booking**~**Interval**+**Price**+**Promotion**+**Income**+**Gender, data=hotel, family=**binomial**(link="logit"))**print**("The Summary of Logit")**

**## [1] "The Summary of Logit"**

summary**(hotel.logit1)**

##   
## Call:  
## glm(formula = Booking ~ Interval + Price + Promotion + Income +   
## Gender, family = binomial(link = "logit"), data = hotel)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9661 -0.4311 -0.3377 -0.2444 3.1175   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.9024419 0.1310427 -6.887 5.71e-12 \*\*\*  
## Interval 0.0126689 0.0031368 4.039 5.37e-05 \*\*\*  
## Price -0.0132550 0.0006317 -20.984 < 2e-16 \*\*\*  
## Promotion -0.0243461 0.0561804 -0.433 0.665   
## Income 0.0056736 0.0005596 10.139 < 2e-16 \*\*\*  
## Gender 0.0101275 0.0561707 0.180 0.857   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10207.5 on 19999 degrees of freedom  
## Residual deviance: 9579.1 on 19994 degrees of freedom  
## AIC: 9591.1  
##   
## Number of Fisher Scoring iterations: 6

**hotel.logit2 =** glm**(Booking**~poly**(Interval,2)**+**Price**+**Promotion**+**Income**+**Gender, data=hotel, family=**binomial**(link="logit"))**print**("The Summary of Logit with Square of Interval")**

**## [1] "The Summary of Logit with Square of Interval"**

summary**(hotel.logit2)**

##   
## Call:  
## glm(formula = Booking ~ poly(Interval, 2) + Price + Promotion +   
## Income + Gender, family = binomial(link = "logit"), data = hotel)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9636 -0.4306 -0.3376 -0.2444 3.1236   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.7707118 0.1247842 -6.176 6.56e-10 \*\*\*  
## poly(Interval, 2)1 15.5913337 3.8916288 4.006 6.17e-05 \*\*\*  
## poly(Interval, 2)2 3.3175535 3.7485973 0.885 0.376   
## Price -0.0132556 0.0006316 -20.986 < 2e-16 \*\*\*  
## Promotion -0.0244618 0.0561822 -0.435 0.663   
## Income 0.0056686 0.0005595 10.131 < 2e-16 \*\*\*  
## Gender 0.0098917 0.0561731 0.176 0.860   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10207.5 on 19999 degrees of freedom  
## Residual deviance: 9578.3 on 19993 degrees of freedom  
## AIC: 9592.3  
##   
## Number of Fisher Scoring iterations: 6

AIC**(hotel.logit1)**

## [1] 9591.052

AIC**(hotel.logit2)**

## [1] 9592.28

BIC**(hotel.logit1)**

## [1] 9638.473

BIC**(hotel.logit2)**

## [1] 9647.604

**Model 1: Interpretation of **1, **2*, *3, **4, **5:**

p-values show that **1, **2*, and* **4 are statistically significant:

* Interval
* Price
* Income

Will have statistically significant effect on booking

* Promotion
* Gender

No statistically significant effect on booking.

* +ve values show +ve effect on booking when other variables are held constant and -ve values show -ve effect on booking
* Magnitude of values shows the extent of effect.

Conclusion:

* Interval from previous booking increases causes higher probability of booking
* Income of the customer increases probability of room booking, with directly proportional relations
* Price of the room increases probability of room booking

**Model 2: Interpretation of **1, **2*, *3, **4, **5, **6:**

p-values show that **1, **3*, and* **5 are statistically significant:

* Income
* Interval
* Price

Will have statistically significant effect on booking

* Interval^2 Promotion
* Gender

No statistically significant effect on booking.

* +ve values show +ve effect on booking when other variables are held constant and -ve values show -ve effect on booking
* Magnitude of values shows the extent of effect

Conclusion:

* Interval from previous booking increases causes higher probability of booking
* Income of the customer increases probability of room booking, with directly proportional relations
* Price of the room increases probability of room booking

**AIC and BIC:**

* The AIC of the model1 and model2 are 9591.052 and 9592.28
* The BIC of the model1 and model2 are 9638.473 and 9647.604

Model1 performs marginally better than the model2 with square of interval.

3). Estimate the following cloglog regression model using the R function glm( )

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

+ **4×*Incomei +*5×*Genderi*

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

**Code and Summary:**

**hotel.cloglog1=**glm**(Booking**~**Interval**+**Price**+**Promotion**+**Income**+**Gender, data=hotel, family=**binomial**(link="cloglog"))**print**("The Summary of C-loglog Model")**

**## [1] "The Summary of C-loglog Model"**

summary**(hotel.cloglog1)**

##   
## Call:  
## glm(formula = Booking ~ Interval + Price + Promotion + Income +   
## Gender, family = binomial(link = "cloglog"), data = hotel)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.0158 -0.4294 -0.3370 -0.2458 3.0979   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.0117417 0.1241479 -8.149 3.65e-16 \*\*\*  
## Interval 0.0120030 0.0029704 4.041 5.32e-05 \*\*\*  
## Price -0.0126884 0.0005983 -21.207 < 2e-16 \*\*\*  
## Promotion -0.0226868 0.0532868 -0.426 0.670   
## Income 0.0053640 0.0005196 10.324 < 2e-16 \*\*\*  
## Gender 0.0105020 0.0532741 0.197 0.844   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10207.5 on 19999 degrees of freedom  
## Residual deviance: 9578.8 on 19994 degrees of freedom  
## AIC: 9590.8  
##   
## Number of Fisher Scoring iterations: 6

AIC**(hotel.cloglog1)**

## [1] 9590.795

BIC**(hotel.cloglog1)**

## [1] 9638.216

**Interpretation of **1, **2*, *3, **4, **5:**

p-values show that **1, **2*, and* **4 are statistically significant:

* Interval
* Price
* Income

Will have statistically significant effect on booking

* Promotion
* Gender

No statistically significant effect on booking.

* +ve values show +ve effect on booking when other variables are held constant and -ve values show -ve effect on booking
* Magnitude of values shows the extent of effect.

Conclusion:

* Interval from previous booking increases causes higher probability of booking
* Income of the customer increases probability of room booking, with directly proportional relations
* Price of the room increases probability of room booking

**AIC and BIC:**

**The AIC and BIC of the model are 9590.795and 9638.216**

4) Next, we will let the intercept be a random effect **0*i* in both the logistic and cloglog models

*log*(*λi*(*t*)*/*(*1- λi*(*t*)) = **0i + **1×*Intervalit* + **2×*Priceit +*3×*Promotionit*

+ **4×*Incomei +*5×*Genderi*

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

+ **4×*Incomei +*5×*Genderi*

Using the R function glmer() with link="logit" and link="cloglog" to estimate these two model and paste results here. Please also calculate the AIC and BIC of these two models.

Based on the AIC's and BIC's of the five models in (2), (3) and (4), which is the best model for the data?

**Code and Summary:**

library**(lme4)**

**## Warning: package 'lme4' was built under R version 3.6.2**

**## Loading required package: Matrix**

**hotel.logit\_re=**glmer**(Booking**~**Interval**+**Price**+**Promotion**+**Income**+**Gender**+**(1**|**customerID), data=hotel, family=**binomial**(link="logit"))**

print**("The Summary of logit RE Model")**

**## [1] "The Summary of logit RE Model"**

summary**(hotel.logit\_re)**

## Generalized linear mixed model fit by maximum likelihood (Laplace  
## Approximation) [glmerMod]  
## Family: binomial ( logit )  
## Formula: Booking ~ Interval + Price + Promotion + Income + Gender + (1 |   
## customerID)  
## Data: hotel  
##   
## AIC BIC logLik deviance df.resid   
## 9588.2 9643.6 -4787.1 9574.2 19993   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -0.8006 -0.3081 -0.2389 -0.1715 11.4066   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## customerID (Intercept) 0.06312 0.2512   
## Number of obs: 20000, groups: customerID, 400  
##   
## Fixed effects:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.9744082 0.1393634 -6.992 2.71e-12 \*\*\*  
## Interval 0.0170843 0.0037663 4.536 5.73e-06 \*\*\*  
## Price -0.0133162 0.0006342 -20.996 < 2e-16 \*\*\*  
## Promotion -0.0261630 0.0563782 -0.464 0.643   
## Income 0.0058230 0.0006330 9.199 < 2e-16 \*\*\*  
## Gender 0.0093249 0.0618858 0.151 0.880   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation of Fixed Effects:  
## (Intr) Intrvl Price Promtn Income  
## Interval -0.383   
## Price -0.714 -0.044   
## Promotion -0.205 -0.025 0.020   
## Income -0.479 0.212 -0.025 -0.006   
## Gender -0.230 -0.007 0.005 0.005 0.019  
## convergence code: 0  
## Model failed to converge with max|grad| = 0.0054431 (tol = 0.001, component 1)  
## Model is nearly unidentifiable: very large eigenvalue  
## - Rescale variables?

AIC**(hotel.logit\_re)**

## [1] 9588.241

BIC(**hotel.logit\_re)**

## [1] 9643.565

#Part-B Cloglog with RE for CustomerID

library**(lme4)  
hotel.cloglog\_re=**glmer**(Booking**~**Interval**+**Price**+**Promotion**+**Income**+**Gender**+**(1**|**customerID), data=hotel, family=**binomial**(link="cloglog"))**

print**("The Summary of cloglog RE Model")**

## [1] "The Summary of cloglog RE Model"

summary**(hotel.cloglog\_re)**

## Generalized linear mixed model fit by maximum likelihood (Laplace  
## Approximation) [glmerMod]  
## Family: binomial ( cloglog )  
## Formula: Booking ~ Interval + Price + Promotion + Income + Gender + (1 |   
## customerID)  
## Data: hotel  
##   
## AIC BIC logLik deviance df.resid   
## 9587.8 9643.1 -4786.9 9573.8 19993   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -0.8551 -0.3069 -0.2384 -0.1725 11.0681   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## customerID (Intercept) 0.05842 0.2417   
## Number of obs: 20000, groups: customerID, 400  
##   
## Fixed effects:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.0818920 0.1323090 -8.177 2.91e-16 \*\*\*  
## Interval 0.0162822 0.0035743 4.555 5.23e-06 \*\*\*  
## Price -0.0127425 0.0006006 -21.217 < 2e-16 \*\*\*  
## Promotion -0.0241189 0.0533963 -0.452 0.651   
## Income 0.0055030 0.0005936 9.271 < 2e-16 \*\*\*  
## Gender 0.0093618 0.0588422 0.159 0.874   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation of Fixed Effects:  
## (Intr) Intrvl Price Promtn Income  
## Interval -0.393   
## Price -0.714 -0.033   
## Promotion -0.205 -0.023 0.020   
## Income -0.491 0.212 -0.007 -0.006   
## Gender -0.229 -0.009 0.005 0.002 0.018  
## convergence code: 0  
## Model failed to converge with max|grad| = 0.00562954 (tol = 0.001, component 1)  
## Model is nearly unidentifiable: very large eigenvalue  
## - Rescale variables?

AIC**(hotel.cloglog\_re)**

## [1] 9587.758

BIC**(hotel.cloglog\_re)**

## [1] 9643.083

Logit link: The AIC and BIC of RE Model for Intercept term with are 9588.241 and 9643.565

Cloglog link: The AIC and BIC of RE Model for Intercept term are 9587.758 and 9643.083 respectively

This shows that Cloglog model with RE’s has the lower AIC and BIC values, and hence is a better model for the given data