**OM386 Advanced Data Analytics in Marketing**

**Assignment 3**

**Due: March 22nd, 11:59pm**

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

In this exercise, we will estimate discrete hazard models for repeated purchase event data. 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 studies 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 (“season” as explained below).

1). 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 counts 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 show your code here.

library(Matrix)

papertowel.data = read.csv("C:/Users/SAMSUNG/Downloads/Papertowel\_repurchase.csv", header=TRUE)

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

interval = c()

for (i in 1:500) {

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

interval.i = rep(0, 52)

sincePurchase = 0

for (t in 1:52) {

sincePurchase = sincePurchase + 1

interval.i[t] = sincePurchase

if (papertowel.i$papertowel[t] == 1) sincePurchase = 0

}

interval = c(interval, interval.i)

}

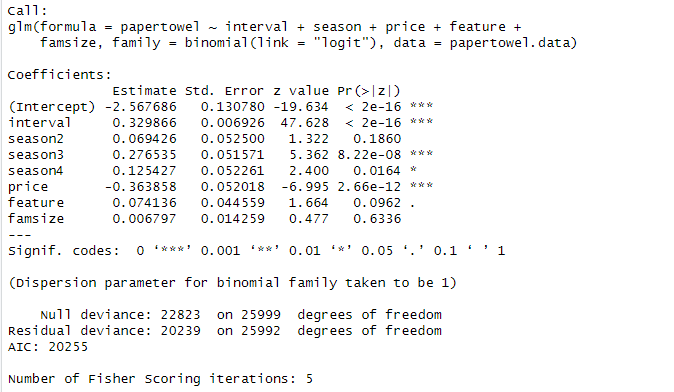
papertowel.data$interval = interval

2). 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. Are they statistically significant? Please calculate the AIC of this model.



* Intercept: This coefficient is statistically significant as the p-value is less than 2e-16 (which is effectively zero).
* interval: The coefficient for interval is statistically significant with a p-value much less than 0.05.
* season2: The coefficient for season2 is not statistically significant (p-value > 0.05).
* season3: The coefficient for season3 is statistically significant (p-value < 0.05).
* season4: The coefficient for season4 is statistically significant (p-value < 0.05).
* price: The coefficient for price is highly statistically significant (p-value < 2e-16).
* feature: The coefficient for feature is not statistically significant at the 0.05 level (p-value is slightly above 0.05, but might be considered significant at the 0.1 level).
* famsize: The coefficient for famsize is not statistically significant (p-value > 0.05).

Three asterisks (), as seen with the intercept, interval, season3, season4, and price, indicate a very high level of significance.

> print(hazard1\_aic)

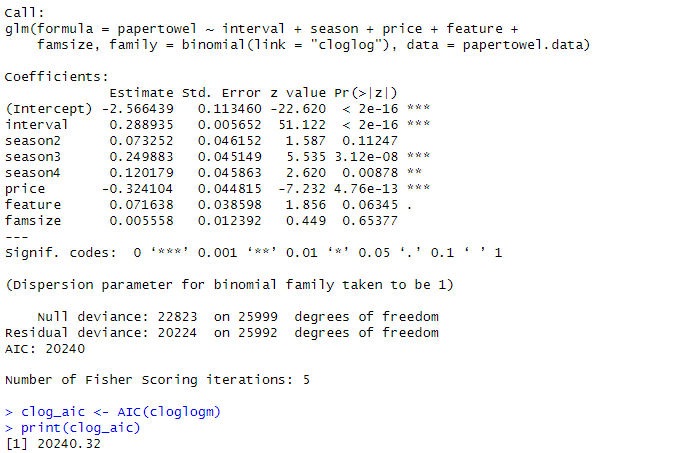
[1] 20254.74

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



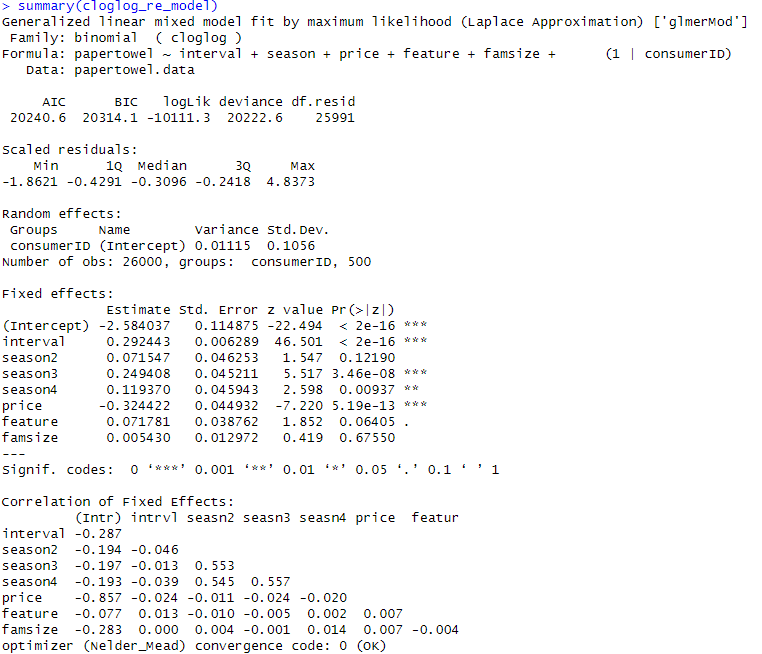
* *b*1​ (Intervalit): Represents the change in the log-odds of making a purchase for each additional week since the last purchase. Coefficient is highly significant (p < 2e-16). A positive coefficient suggests that the longer the time since the last purchase, the greater the odds of making a purchase.
* *b*2​ (Seasonit): Represents the change in the log-odds of making a purchase associated with each season. Since the model output does not differentiate between seasons (all seasons are referred to as season in the coefficients), it's likely that the seasons were entered into the model as a factor and season2, season3, and season4 are being compared to a reference season (probably season1). The coefficients for season3 and season4 are significant, suggesting that these seasons have a different impact on purchase probability compared to the reference season.
* *b*3​ (Priceit): Represents the change in the log-odds of making a purchase for each unit change in price. Coefficient is highly significant (p < 2e-16). The negative coefficient suggests that higher prices are associated with lower odds of making a purchase.
* *b*4​ (Featureit): Represents the change in the log-odds of making a purchase when paper towels are featured in the store. The positive coefficient suggests that when paper towels are featured, the odds of making a purchase increase, though this effect is not statistically significant at the 0.05 level.
* *b*5​ (Famsizei): Represents the change in the log-odds of making a purchase for each additional member in the household. The coefficient is not statistically significant, suggesting that family size does not have a discernible effect on the odds of making a purchase.

4). 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 (2), (3) and (4), which is the best model for the data?



> cloglog\_re\_model\_aic <- AIC(cloglog\_re\_model)

> print(cloglog\_re\_model\_aic)

[1] 20240.64

Based on the AIC's of the models in (2), (3) and (4), the cloglog regression model is the best fit since it has the lowest AIC, 20240.32, but the cloglog model with random effect also presented a similar score with slightly higher score of 20240.64. The cloglog link without random effects assumes that the observations are independent of each other, which might be suitable if there is no reason to believe that there is unobserved heterogeneity among subjects or clusters in the data. if the AIC is similar or the complexity added by the random effects does not lead to a better understanding or prediction, the simpler model without random effects may suffice.

**Censored Regression and Its Bayesian Estimation**

In this exercise, we will apply censored regression to the dataset "CreditCard\_SOW\_Data2.csv". The dataset has the following variables.

|  |  |
| --- | --- |
| ConsumerID | ID's of the sampled consumers |
| SOW | The card's share of wallet in the consumer's total monthly spending |
| Promotion | Index of monthly promotion activity –higher index indicates more promotions |
| Balance | The customer's unpaid balance at the beginning of the month |

1). In this data set, the share of wallet (SOW) can be 0% or 100%. Therefore, the share of wallet is considered a truncated-normal (censored) variable at 0% (censored at 0) or 100% (censored at 1). We would like fit the following regression model

*SOWij\** *= β0 + β1×Balanceij + β2×Promotionij +* *εij*

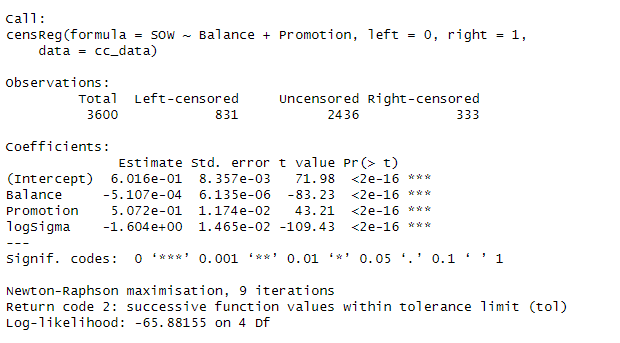
*SOWij = SOWij\** if 0 < *SOWij\** < 1

*SOWij =*0 if *SOWij\** ≤ 0

*SOWij =*1 if *SOWij\** ≥ 1

*εij ~N(0,* *σ2)* (We can reparametrize *τ = 1/σ2,* which is often named the precision)

Please use the R function censReg() in library(censReg) to fit this model by MLE. Copy and paste the summary of the results here. Interpret the parameters *β1 β2* in the model.



* *β*1​ (Balance): The coefficient for Balance is -5.107e-04, which is negative and statistically significant (p < 2e-16). This implies that for each unit increase in Balance, there is a decrease in the log-odds of having a higher share of wallet, on average. Because SOW is bounded between 0 and 1, a negative relationship indicates that customers with higher balances are less likely to use this particular card for a larger share of their total spending.
* *β*2​ (Promotion): The coefficient for Promotion is 5.072e-01, which is positive and statistically significant (p < 2e-16). This suggests that for each unit increase in the Promotion index, the log-odds of having a higher share of wallet increase, on average. This indicates that more promotions are associated with a higher probability that the card will be used for a larger share of the consumer's total spending.

Both parameters are statistically significant, which means that we have strong evidence to believe that these relationships are not due to random chance in the sample of data collected. The negative coefficient for Balance could be interpreted in the context of risk management, where consumers with higher outstanding balances might be using their credit more cautiously, while the positive coefficient for Promotion aligns with marketing efforts encouraging more usage of the card.

2).Next, we will fit the model above using Bayesian estimation, which involves sampling the latent *SOWij\** when the observed *SOWij =*0 or *SOWij =*1. The R code using MCMC (Gibbs sampling) for this inference problem is in "Assignment-3\_blankcode.r". Please read the code carefully and fill in the code for sampling the latent *SOWij\**, the regression coefficients and the precision (inverse of the variance) of the error. You may use the rtruncnorm( ) function in the library(truncnorm) to sample from truncated normal distributions. This method is called data augmentation, which is widely applied in Bayesian statistics for missing data.

Please run the completed code. Use the plot() function to plot the posterior sampling chains and hist() to plot posterior histograms for *β0, β1, β2* and *τ .* Copy and paste the results here. Please also calculate the 95% posterior intervals for *β0, β1, β2* and *τ .* Copy and paste the results here.

