MKT386
Assignment 3

By: Callie Gilmore (cgg756)

Due: March 24th, 11:59pm

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 by which the shopping 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

$$log(\lambda_{it}) = \beta_0 + \beta_1 \times Discount + \beta_2 \times Target + \beta_3 \times Income + \beta_4 \times Distant + \beta_5 \times Gender$$

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 also calculate the AIC and BIC of this regression model.

```
Call:
glm(formula = Visit ~ Discount + Target + Income + Distant +
    Gender, family = poisson(link = "log"), data = mallVisitData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5136  -0.9896  -0.7675   0.5602   3.8721

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.4143685  0.0343574 -41.166  < 2e-16 ***
Discount     0.0007975  0.0002693   2.962  0.00306 **
Target      -0.0275468  0.0179415  -1.535  0.12469
Income       0.0049092  0.0001252  39.216  < 2e-16 ***
Distant     -0.0469158  0.0036635 -12.806  < 2e-16 ***
Gender       0.0448540  0.0179698   2.496  0.01256 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 26722  on 24999  degrees of freedom
Residual deviance: 24853  on 24994  degrees of freedom
AIC: 45274

Number of Fisher Scoring iterations: 6
```

From the results above, we see that Discount, Income, Distant and Gender are statistically significant. In other words, β_1, β_3, β_4, β_5 are statistically significant as their p-values are less than 0.05 but β_2 (Target) is not as the p-value is greater than 0.05.

The AIC and BIC of the model are shown below:

```
> cat("AIC: ")
AIC: > AIC(visit.poisson)
[1] 45274.32
> cat("BIC: ")
BIC: > BIC(visit.poisson)
[1] 45323.08
```

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

$$Log(\lambda_{it}) = \beta_{0i} + \beta_1 \times Discount + \beta_2 \times Target + \beta_3 \times Income + \beta_4 \times Distant + \beta_5 \times Gender,$$

where the individual intercept β_{0i} will 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 and BIC of this regression model.

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: poisson  ( log )
Formula:
Visit ~ (1 | customerID) + Discount + Target + Income + Distant +
    Gender
   Data: mallVisitData

     AIC      BIC   logLik deviance df.resid
 44695.9  44752.8 -22340.9  44681.9    24993

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.2248 -0.6673 -0.5086  0.6231  5.9048

Random effects:
 Groups     Name        Variance Std.Dev.
 customerID (Intercept) 0.09248  0.3041
Number of obs: 25000, groups:  customerID, 500

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.4612174  0.0570385 -25.618  < 2e-16 ***
Discount     0.0008834  0.0002700   3.272  0.00107 **
Target      -0.0269285  0.0179825  -1.497  0.13427
Income       0.0049041  0.0002231  21.981  < 2e-16 ***
Distant     -0.0473141  0.0067030  -7.059 1.68e-12 ***
Gender       0.0458259  0.0333838   1.373  0.16985
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
         (Intr) Discnt Target Income Distnt
Discount -0.125
Target   -0.155 -0.012
Income   -0.707 -0.003 -0.002
Distant  -0.560 -0.004  0.004  0.039
Gender   -0.255  0.000  0.001 -0.038 -0.030
optimizer (Nelder_Mead) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.00207379 (tol = 0.002, comp
onent 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
```

From the results above, we see that Discount, Income and Distant are statistically significantly. In other words, β_1, β_3, and β_5 are statistically significant as their p-values are less than 0.05 and β_2 and β_4 (Target and Gender) are not as their p-values are greater than 0.05.

The AIC and BIC of the model are reported below:

```
> cat("AIC: ")
AIC:
> AIC(visit.poisson.random)
[1] 44695.9
> cat("BIC: ")
BIC:
> BIC(visit.poisson.random)
[1] 44752.78
```

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

$$log(\lambda_{it}) = \beta_0 + \beta_1 \times Discount + \beta_2 \times Target + \beta_3 \times Income + \beta_4 \times Distant + \beta_5 \times Gender,$$

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 and BIC of the model.

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

```
Call:
glm.nb(formula = Visit ~ Discount + Target + Income + Distant +
    Gender, data = mallVisitData, init.theta = 10.82765227, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4760  -0.9787  -0.7623   0.5435   3.6554

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.4140463  0.0350814 -40.308   <2e-16 ***
Discount     0.0007957  0.0002764   2.878   0.0040 **
Target      -0.0275706  0.0184075  -1.498   0.1342
Income       0.0049096  0.0001281  38.328   <2e-16 ***
Distant     -0.0470039  0.0037561 -12.514   <2e-16 ***
Gender       0.0449952  0.0184358   2.441   0.0147 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(10.8277) family taken to be
 1)

    Null deviance: 25524  on 24999  degrees of freedom
Residual deviance: 23738  on 24994  degrees of freedom
AIC: 45248

Number of Fisher Scoring iterations: 1


              Theta:  10.83
          Std. Err.:  2.18

 2 x log-likelihood:  -45234.17
```

From the results above, we see that Discount, Income, Distant and Gender are statistically significant. In other words, β_1, β_3, β_4, β_5 are statistically significant as their p-values are less than 0.05, β_2 (Target) is not statistically significant as the p-value is greater than 0.05.

The AIC and BIC of the model are reported below:

```
> cat("AIC: ")
AIC: > AIC(visit.nb)
[1] 45248.16
> cat("BIC: ")
BIC: > BIC(visit.nb)
[1] 45305.05
```

Based on all the model results, the best model is the 2nd model (random effects) because its AIC is the lowest at 44695.9.

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?

```
Iterations = 10001:29981
Thinning interval = 20
Number of chains = 1
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                      Mean        SD  Naive SE Time-series SE
beta.(Intercept) -1.458994 0.0275274 8.705e-04      6.657e-03
beta.Discount     0.000866 0.0001912 6.046e-06      3.836e-05
beta.Target      -0.026017 0.0111462 3.525e-04      2.306e-03
beta.Distant     -0.049543 0.0040123 1.269e-04      1.353e-03
beta.Income       0.004965 0.0001243 3.932e-06      3.683e-05
beta.Gender       0.033019 0.0158983 5.027e-04      3.999e-03

2. Quantiles for each variable:

                       2.5%        25%        50%       75%     97.5%
beta.(Intercept) -1.5190465 -1.4773325 -1.4596487 -1.439661 -1.409694
beta.Discount     0.0005456  0.0007166  0.0008468  0.001001  0.001246
beta.Target      -0.0484926 -0.0330770 -0.0247604 -0.018701 -0.004972
beta.Distant     -0.0567958 -0.0524544 -0.0497428 -0.046609 -0.041958
beta.Income       0.0046794  0.0048813  0.0049757  0.005046  0.005170
beta.Gender       0.0069093  0.0223233  0.0309312  0.041938  0.071768
```

The fixed effects are significant at the 5% level as 0 is not included in the range of 2.5% to 25% so since it is not 0, they are 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 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 $\lambda_0(t)$ is reset the $\lambda_0(t+1)= \lambda_0(1)$ if the customer books a hotel in period (week) *t*.

5). 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 = interval
```

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

$$log(\lambda_i(t)/(1-\lambda_i(t)) = \beta_0 + \beta_1 \times Interval_{it} + \beta_2 \times Price_{it} + \beta_3 \times Promotion_{it} + \beta_4 \times Income_i + \beta_5 \times Gender_i$$

And paste results here. How do you interpret β_1, β_2, β_3, β_4, β_5? Are they statistically significant? Please calculate the AIC and BIC of this model.

```
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
```

β_1, β_3, β_4 (Interval, Price and Income) are statistically significant as their p-values are less than 0.05 and will have a statistically significant effect on Booking.
β_3 and β_5 (Promotion and Gender) are not statistically significant as their p-values are greater than 0.05 and will not have a statically significant impact on Booking.

AIC and BIC of the model are reported below:

```
> cat("AIC: ")
AIC: > AIC(hotel.logit1)
[1] 9591.052
> cat("BIC: ")
BIC: > BIC(hotel.logit1)
[1] 9638.473
```

Next, we will estimate the model:

$$log(\lambda_i(t)/(1-\lambda_i(t)) = \beta_0 + \beta_1 \times Interval_{it} + \beta_2 \times Interval_{it}^2 + \beta_3 \times Price_{it} + \beta_4 \times Promotion_{it} + \beta_5 \times Income_i + \beta_6 \times Gender_i$$

Use poly(Interval, 2) in the glm() function to represent $\beta_1 \times Interval_{it} + \beta_2 \times Interval_{it}^2$ in this model. Are $\beta_1, \ldots, \beta_6$ still statistically significant? Please calculate the AIC and BIC of this model.

```
Call:
glm(formula = Booking ~ poly(Interval, 2) + Price + Promotion +
    Income + Gender, family = binomial(link = "logit"), data = hote
l)

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
```

$\beta_1, \beta_3, \beta_5$ (poly(Interval, 2) 1, Price and Income) are statistically significant as their p-values are less than 0.05 and will have a statistically significant effect on Booking. β_2, β_4 and β_6 (poly(Interval, 2) 2, Promotion and Gender) are not statistically significant as their p-values are greater than 0.05 and will not have a statically significant impact on Booking.

AIC and BIC of the model are reported below:

```
> cat("AIC: ")
AIC: > AIC(hotel.logit2)
[1] 9592.28
> cat("BIC: ")
BIC: > BIC(hotel.logit2)
[1] 9647.604
```

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

$$log(-log(1-\lambda_i(t)) = \beta_0 + \beta_1 \times Interval_{it} + \beta_2 \times Price_{it} + \beta_3 \times Promotion_{it}$$

$+ \beta_4 \times Income_i + \beta_5 \times Gender_i$

Paste results here. Are they statistically significant? How do you interpret β_1, β_2, β_3, β_4, β_5? Please calculate the AIC and BIC of this model.

```
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
```

β_1, β_2, β_4 (Interval, Price and Income) are statistically significant as their p-values are less than 0.05 and will have a statistically significant effect on Booking.
β_3 and β_5 (Promotion and Gender) are not statistically significant as their p-values are greater than 0.05 and will not have a statically significant impact on Booking.

AIC and BIC of the model are reported below:

```
> cat("AIC: ")
AIC: > AIC(hotel.cloglog)
[1] 9590.795
> cat("BIC: ")
BIC: > BIC(hotel.cloglog)
[1] 9638.216
```

8) Next, we will let the intercept be a random effect β_{0i} in both the logistic and cloglog models

$$log(\lambda_i(t)/(1-\lambda_i(t)) = \beta_{0i} + \beta_1 \times Interval_{it} + \beta_2 \times Price_{it} + \beta_3 \times Promotion_{it} + \beta_4 \times Income_i + \beta_5 \times Gender_i$$

$$log(-log(1-\lambda_i(t)) = \beta_{0i} + \beta_1 \times Interval_{it} + \beta_2 \times Price_{it} + \beta_3 \times Promotion_{it} + \beta_4 \times Income_i + \beta_5 \times Gender_i$$

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.

Model 1:

```
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) -0.9743867  0.1393726  -6.991 2.72e-12 ***
Interval     0.0170834  0.0037664   4.536 5.74e-06 ***
Price       -0.0133163  0.0006342 -20.996  < 2e-16 ***
Promotion   -0.0261650  0.0563795  -0.464    0.643
Income       0.0058229  0.0006330   9.198  < 2e-16 ***
Gender       0.0093213  0.0618861   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 -0.024 -0.006
Gender    -0.230 -0.007  0.005  0.005  0.019
optimizer (Nelder_Mead) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.00920581 (tol = 0.002, c
omponent 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
```

Model 2

```
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
-0.8550 -0.3069 -0.2384 -0.1725 11.0679

Random effects:
 Groups     Name        Variance Std.Dev.
 customerID (Intercept) 0.05836  0.2416
Number of obs: 20000, groups:  customerID, 400

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.0815193  0.1322923  -8.175 2.95e-16 ***
Interval     0.0162718  0.0035743   4.552 5.30e-06 ***
Price       -0.0127428  0.0006006 -21.218  < 2e-16 ***
Promotion   -0.0241017  0.0533963  -0.451    0.652
Income       0.0055012  0.0005935   9.269  < 2e-16 ***
Gender       0.0093199  0.0588361   0.158    0.874
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
          (Intr) Intrvl Price  Promtn Income
Interval  -0.392
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
optimizer (Nelder_Mead) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.029181 (tol = 0.002, com
ponent 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
```

Based on the AIC's and BIC's of the five models in (6), (7) and (8), which is the best model for the data?

The AIC and BIC for Model 1 are reported below:

```
> cat("AIC: ")
AIC: > AIC(hotel.logit.re)
[1] 9588.241
> cat("BIC: ")
BIC: > BIC(hotel.logit.re)
[1] 9643.565
```

The AIC and BIC are reported below:

```
> cat("AIC: ")
AIC: > AIC(hotel.cloglog.re)
[1] 9587.758
> cat("BIC: ")
BIC: > BIC(hotel.cloglog.re)
[1] 9643.083
```

Based on all of these results, the CLogLog model with random effects (model 2 question 8) is the best model for the data as it has the lowest AIC and BIC.