**MKT386**

**Assignment 3**

**By: Callie Gilmore (cgg756)**

**Due: March 24th, 11:59pm**

**Count Data Analysis for Shopping Mall Visits**

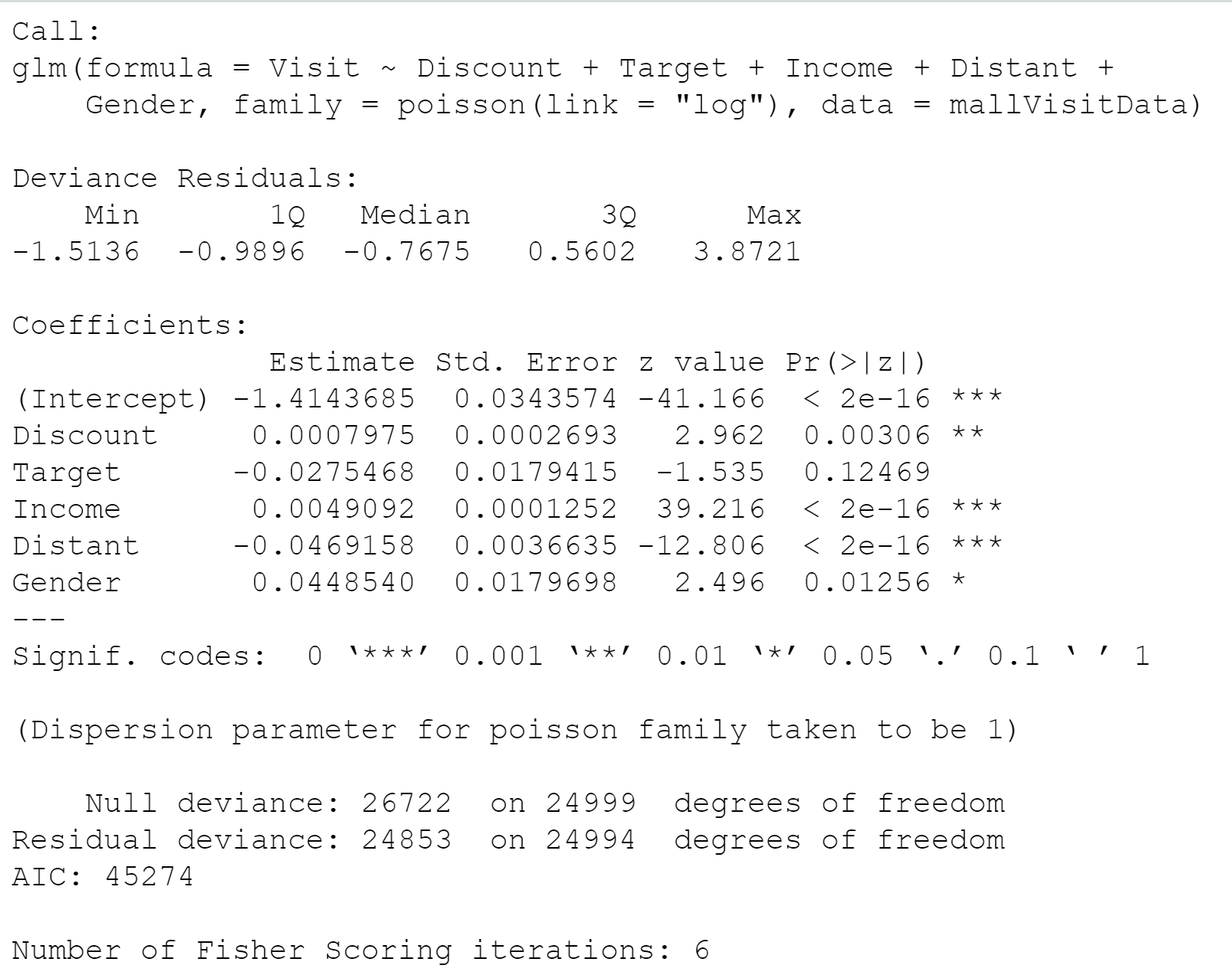
In this exercise, we will apply regression models for count data, including a Poisson log-linear model and a negative binomial model to analyze a data set on the shopping mall visitation frequencies. The goal is to evaluate whether target marketing is effective in attracting consumers to visit the shopping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

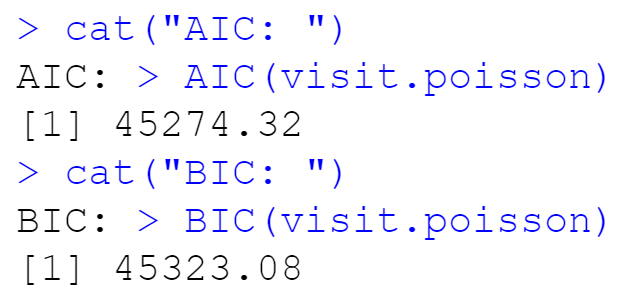
*log(λit )= β0 + β1×Discount +* *β2×Target + β3×Income+ β4×Distant + β5×Gender*

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 also calculate the AIC and BIC of this regression model.

****

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:

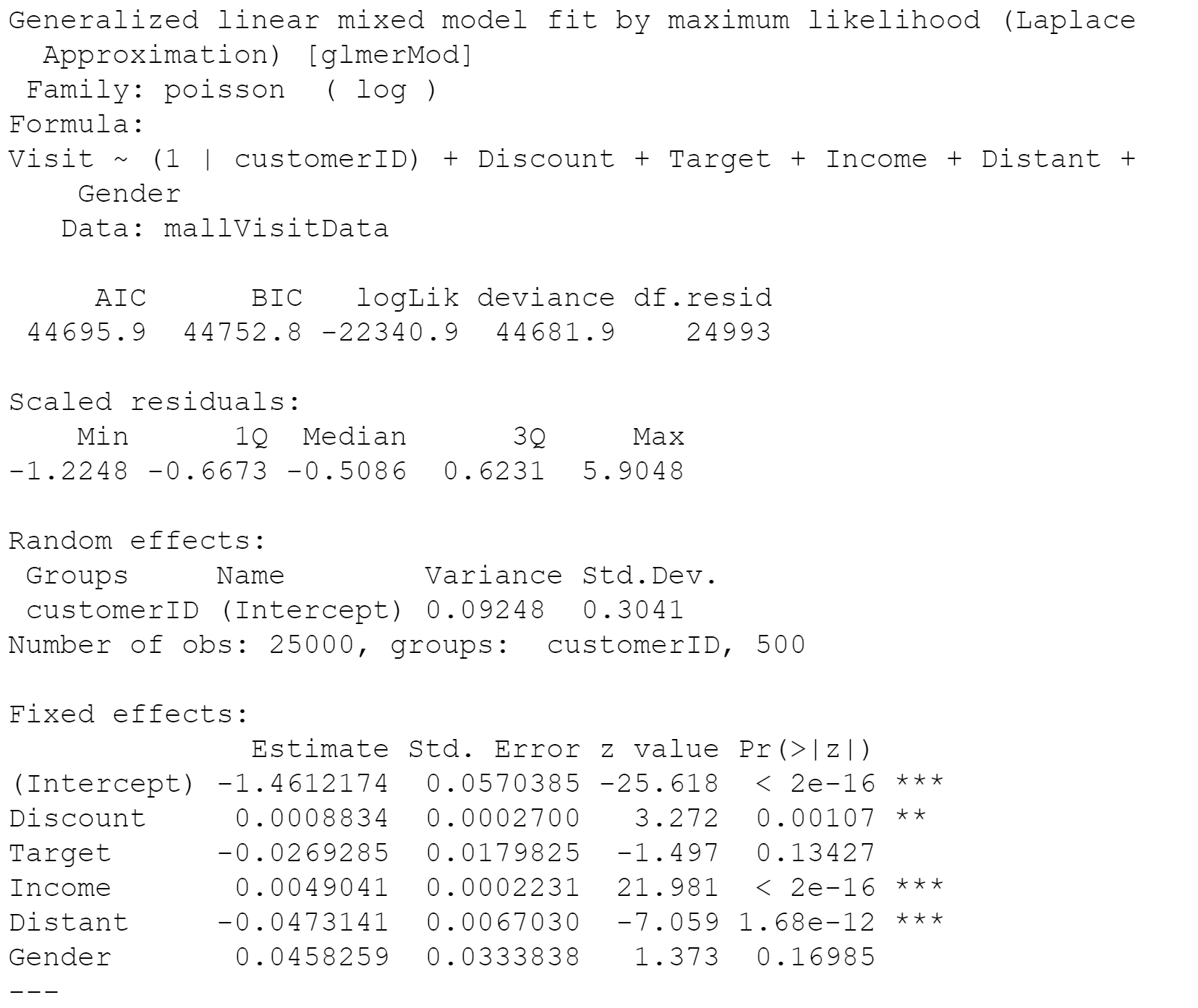


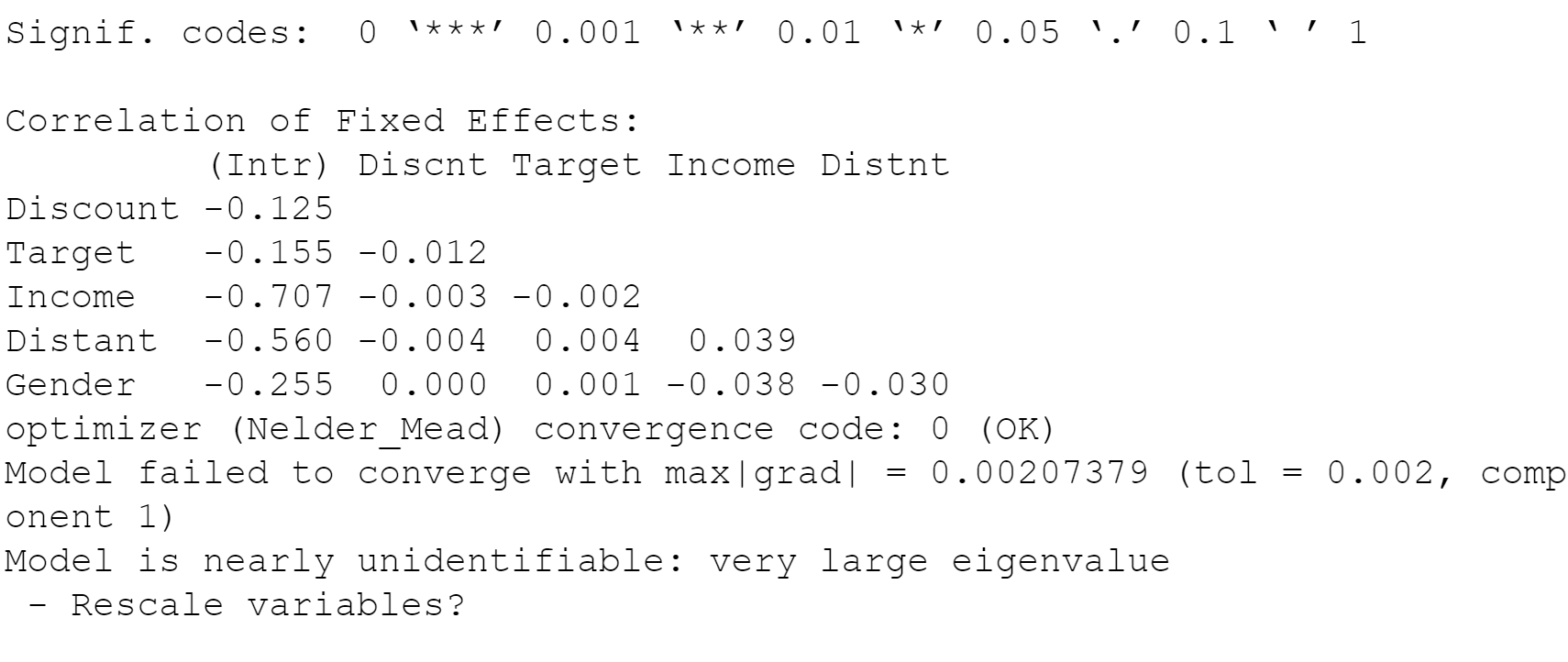
2).Next, we will allow each individual customer to have a different intercept

*Log(λit )= β0i + β1×Discount + β2×Target + β3×Income+ β4×Distant + β5×Gender,*

where the individual intercept *β0i* 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.

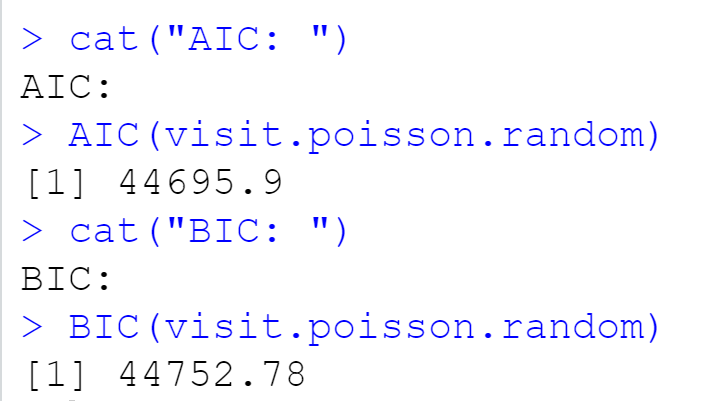
Check the estimates of *β*1, *β*2*, β*3, *β*4, *β*5. Are they statistically significant? Please also calculate the AIC and BIC of this regression model.





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*2and *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:



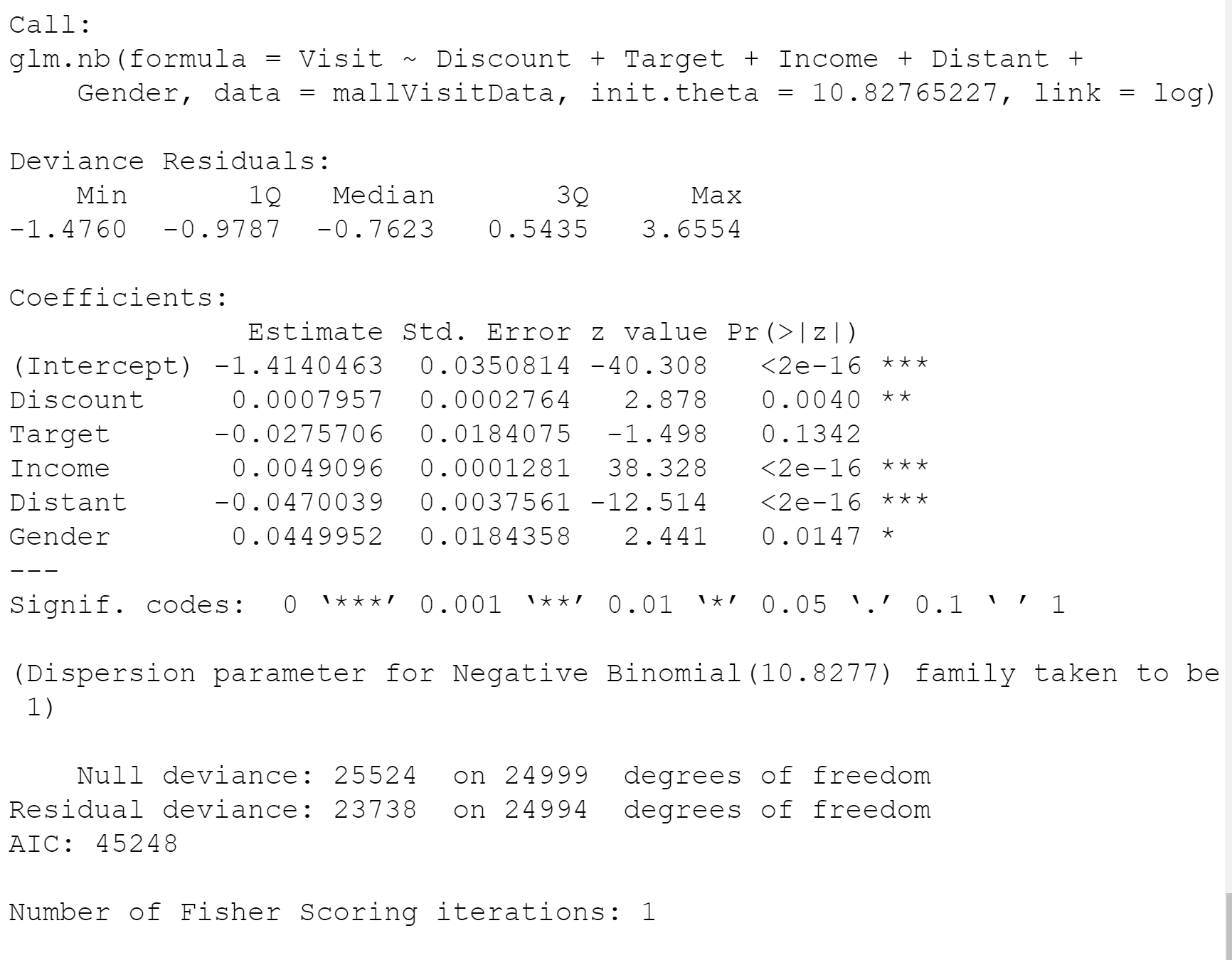
3). We will also fit the negative binomial model for the count data. Let the mean of the negative binomial distribution be

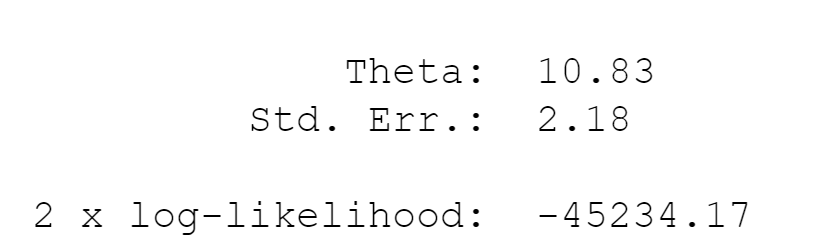
*log(λit )= β0 + β1×Discount + β2×Target + β3×Income+ β4×Distant + β5×Gender,*

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 and BIC of the model.

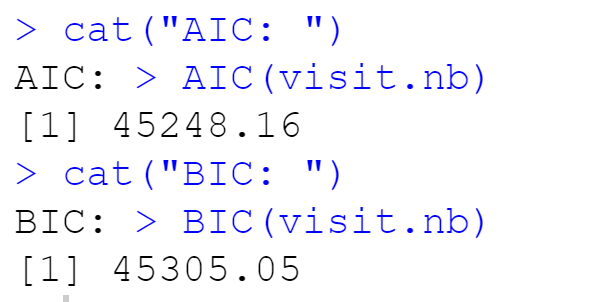
Based on the AIC's and BIC's of the four models in (1), (2) and (3), which is the best model for the data?





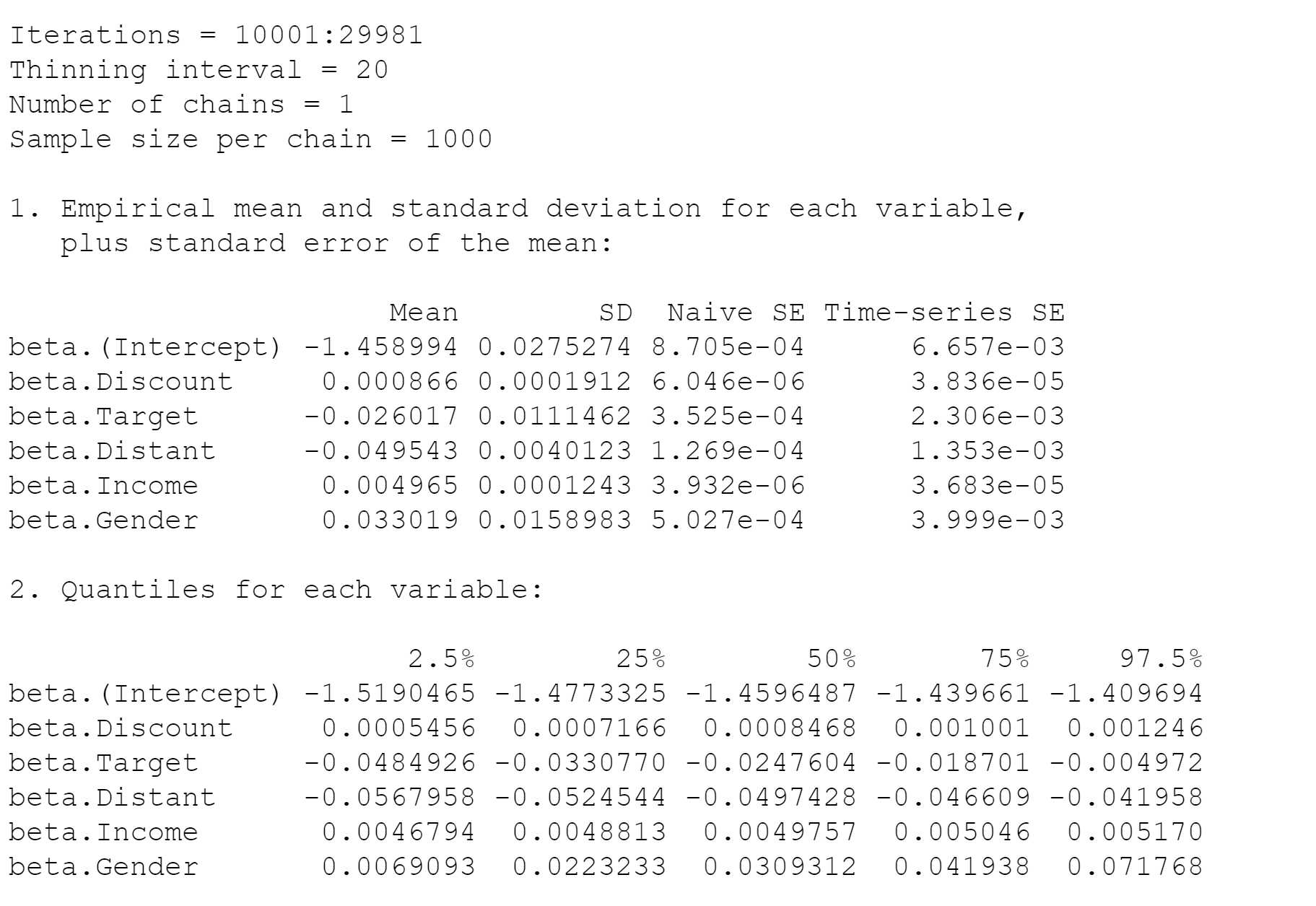
From the results above, we see that Discount, Income, Distant and Gender are statistically significant. In other words, *β*1, *β*3, *β*4, *β*5 are statistically significant as their p-values are less than 0.05, *β*2(Target) is not statistically significant as the p-value is greater than 0.05.

The AIC and BIC of the model are reported below:



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?



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 *λ*0(*t*) is reset the *λ*0(*t*+1)= λ0(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)

}

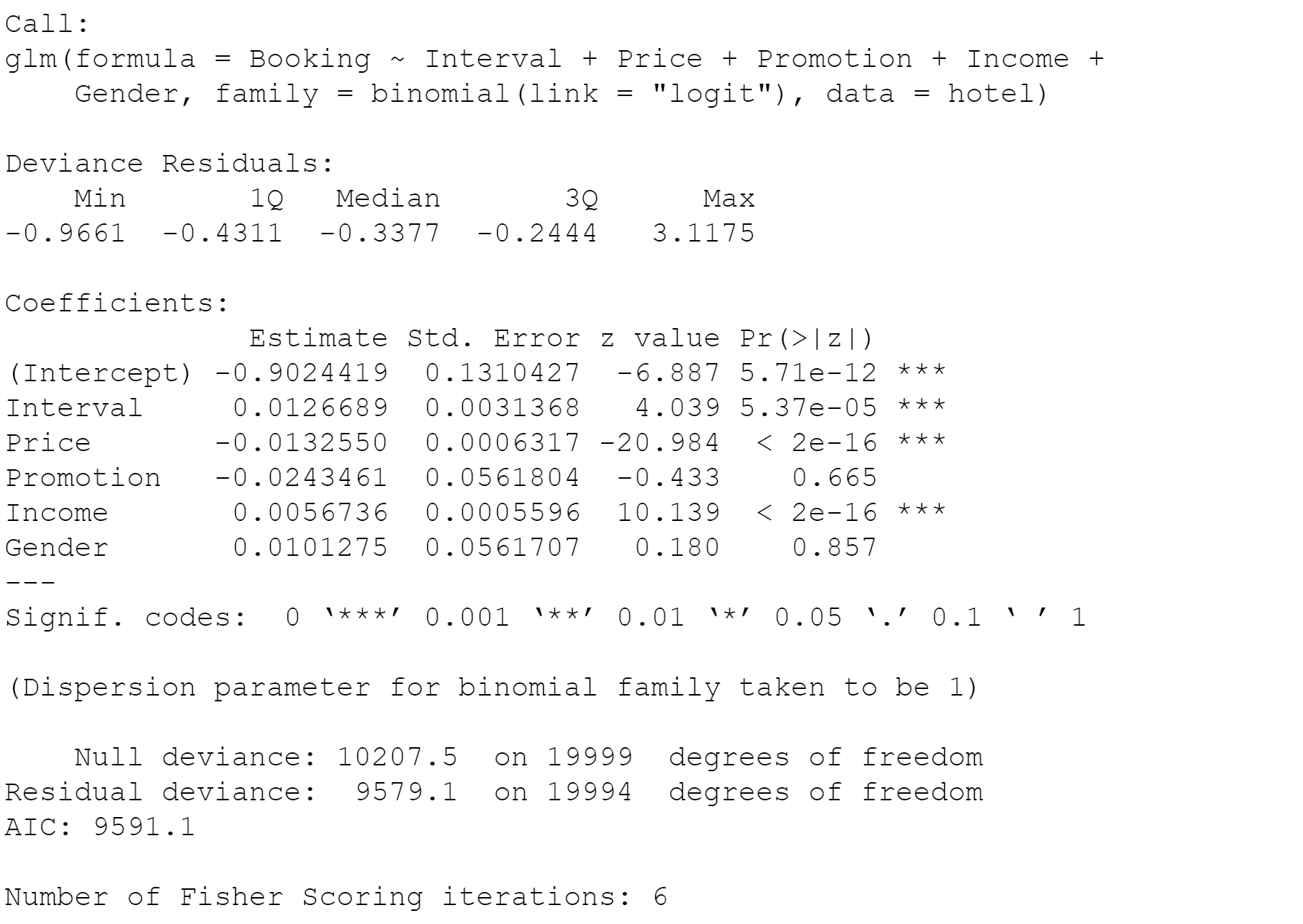
hotel$Interval = interval

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

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

+ **4×*Incomei +*5×*Genderi*

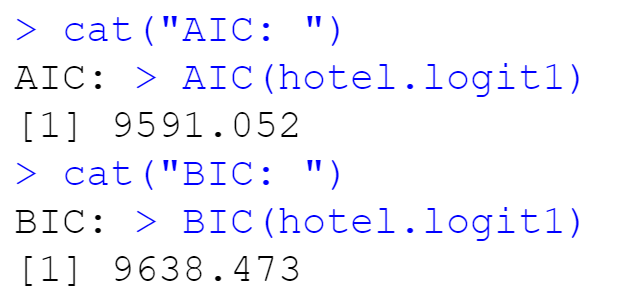
And paste results here. How do you interpret **1, **2*, *3, **4, **5? Are they statistically significant? Please calculate the AIC and BIC of this model.



**1, **3, **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.

**3 and**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:

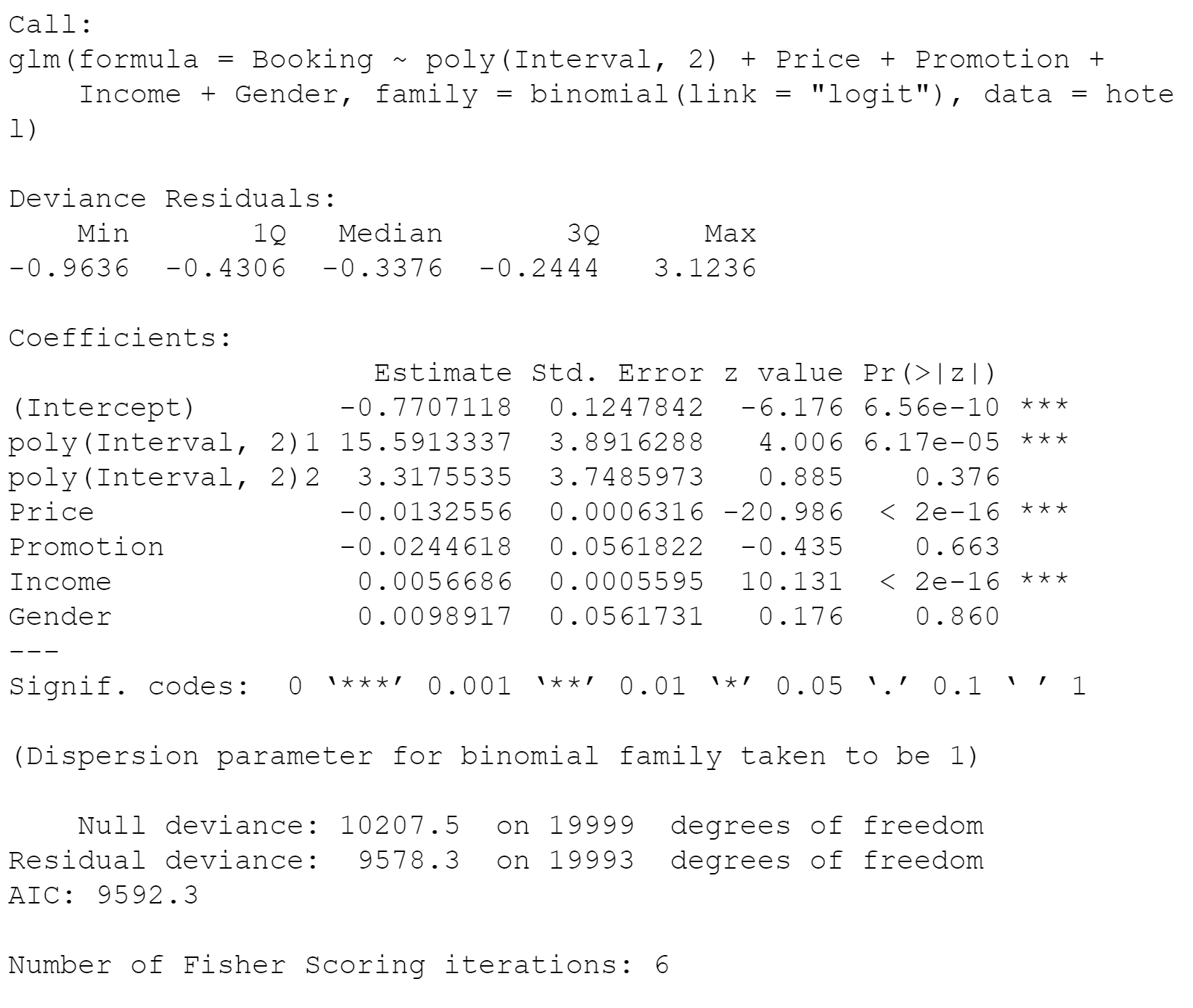


Next, we will estimate the model:

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

+ **5×*Incomei +*6×*Genderi*

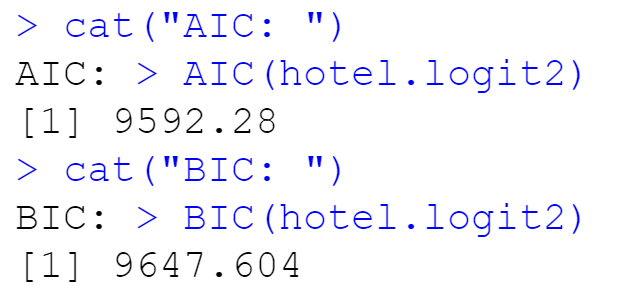
Use poly(Interval, 2) in the glm() function to represent **1×*Intervalit*+**2×*Intervalit2* in this model. Are **1, **, **6 still statistically significant? Please calculate the AIC and BIC of this model.



**1, **3, **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.

**2, **4 and **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:

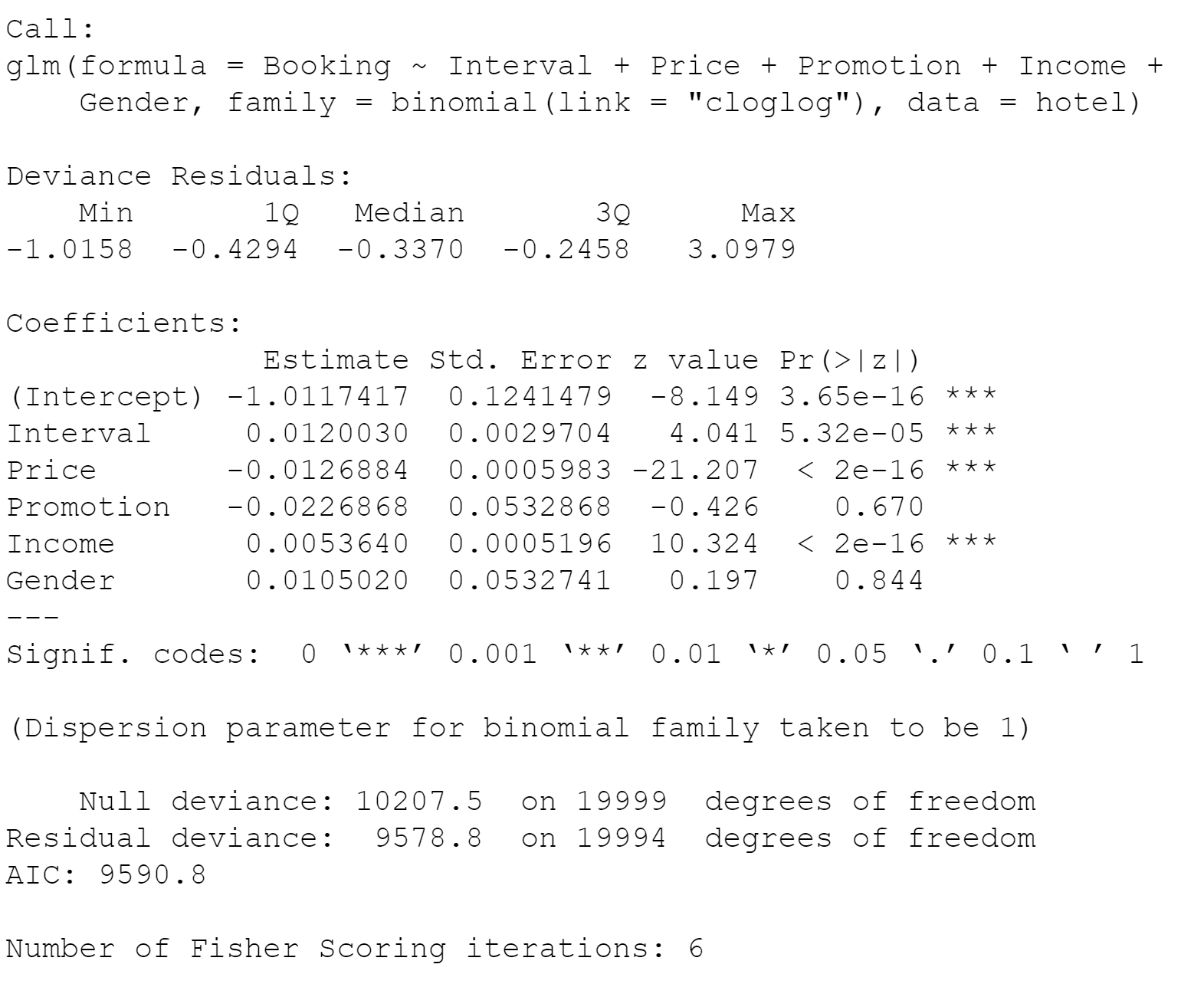


7). Estimate the following cloglog regression model using the R function glm( )

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

+ **4×*Incomei +*5×*Genderi*

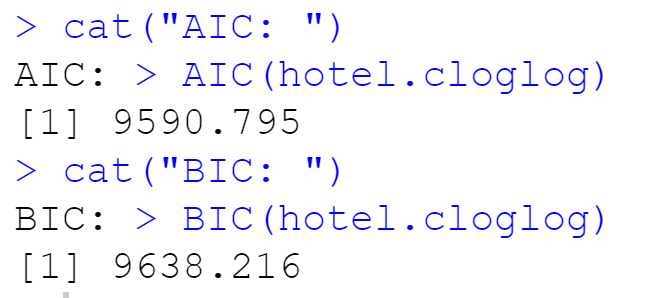
Paste results here. Are they statistically significant? How do you interpret **1, **2*, *3, **4, **5? Please calculate the AIC and BIC of this model.



**1, **2, **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.

**3 and**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:



8) 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*

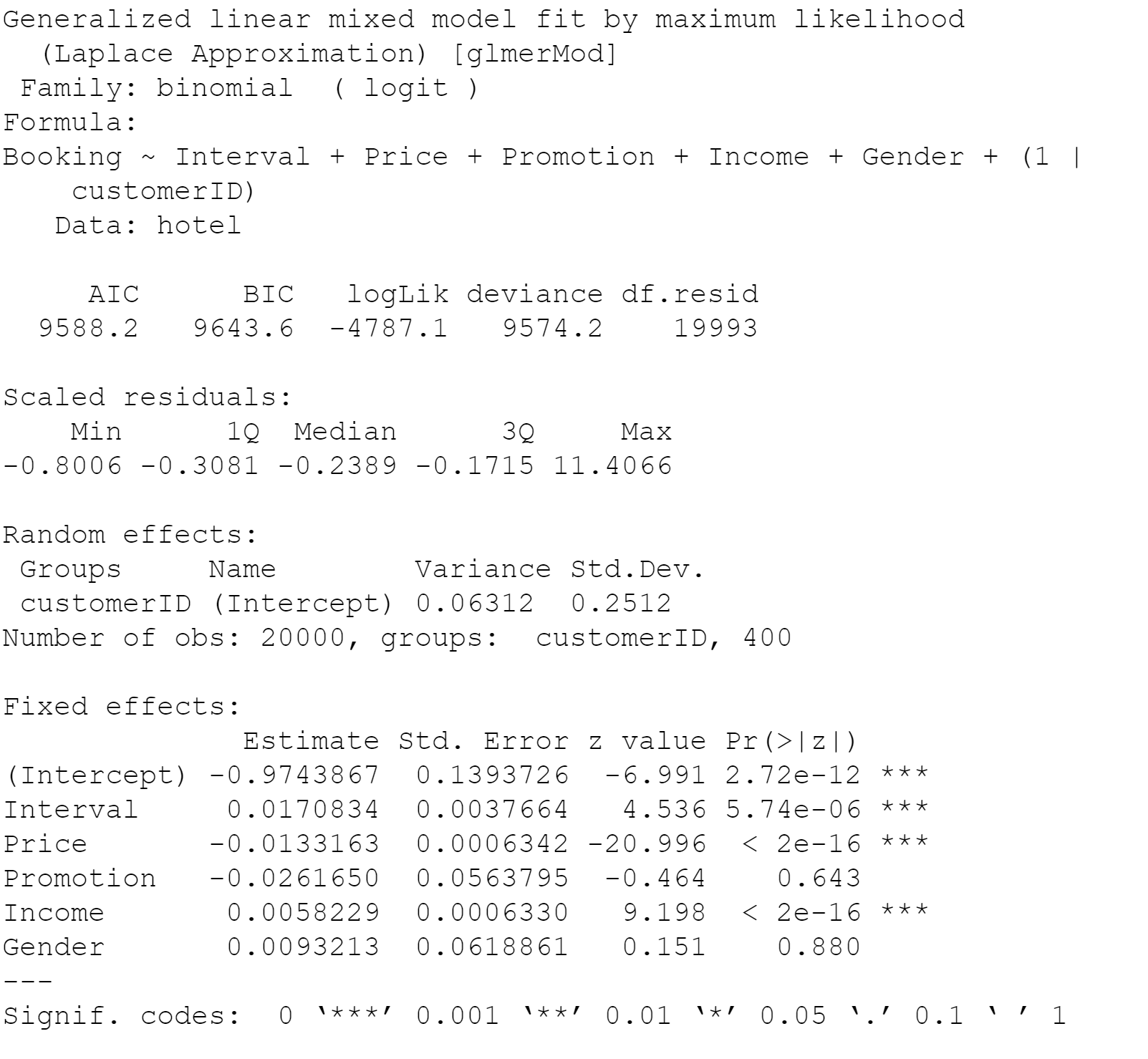
+ **4×*Incomei +*5×*Genderi*

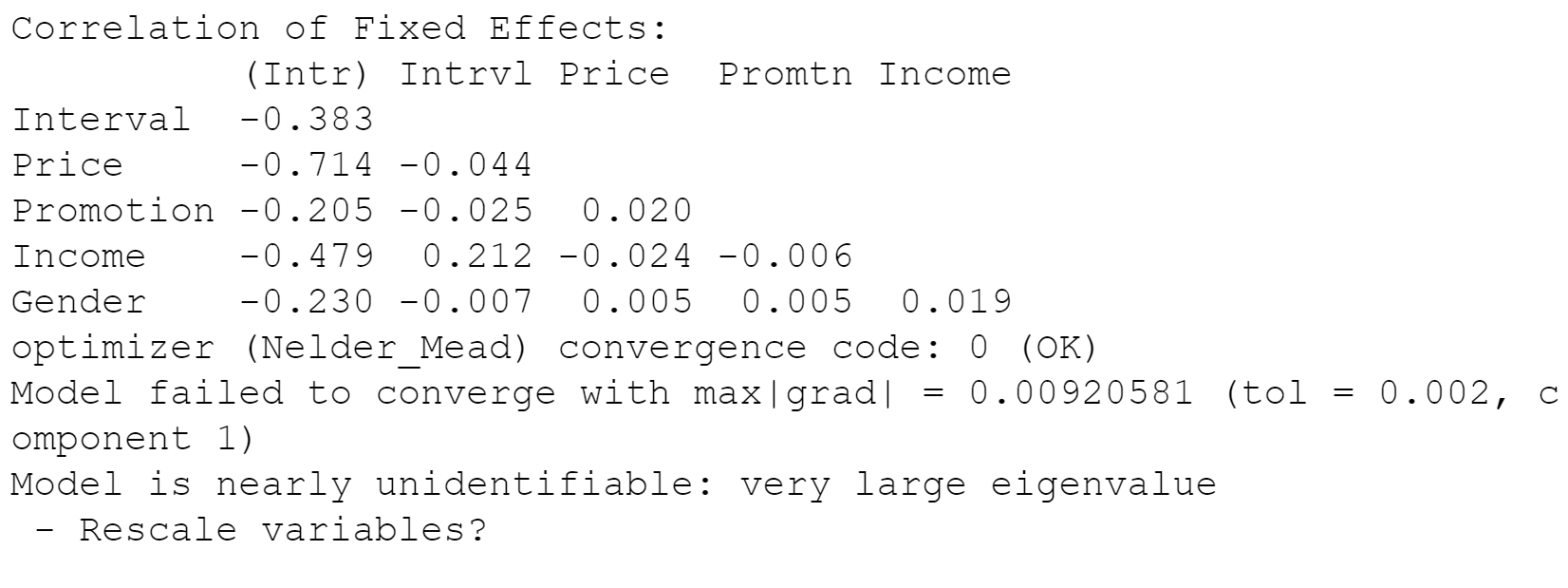
*log*(*-log*(*1- λi*(*t*)) = **0i + **1×*Intervalit* + **2×*Priceit +*3×*Promotionit*

+ **4×*Incomei +*5×*Genderi*

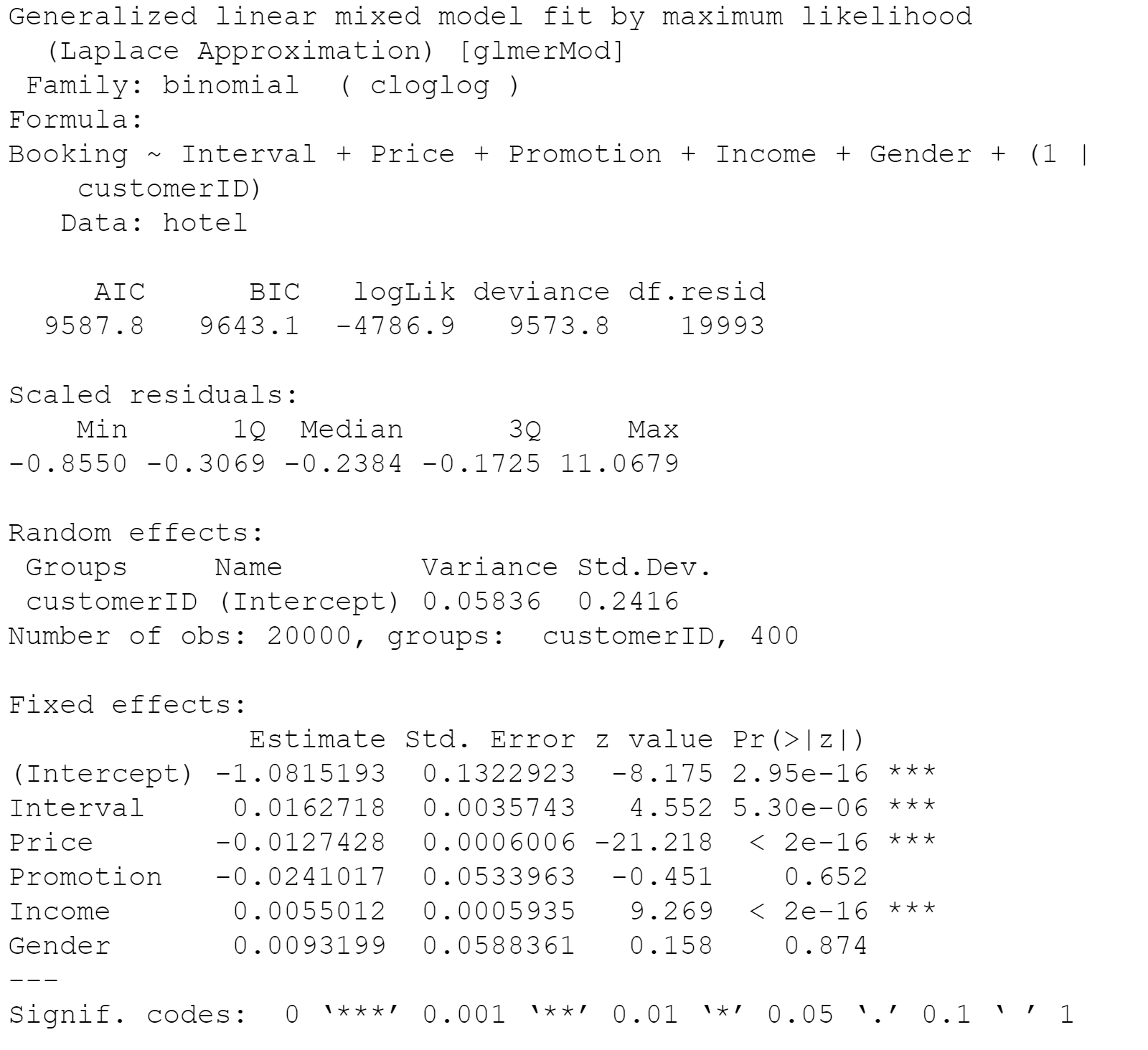
Using the R function glmer() with link="logit" and link="cloglog" to estimate these two model and paste results here. Please also calculate the AIC and BIC of these two models.

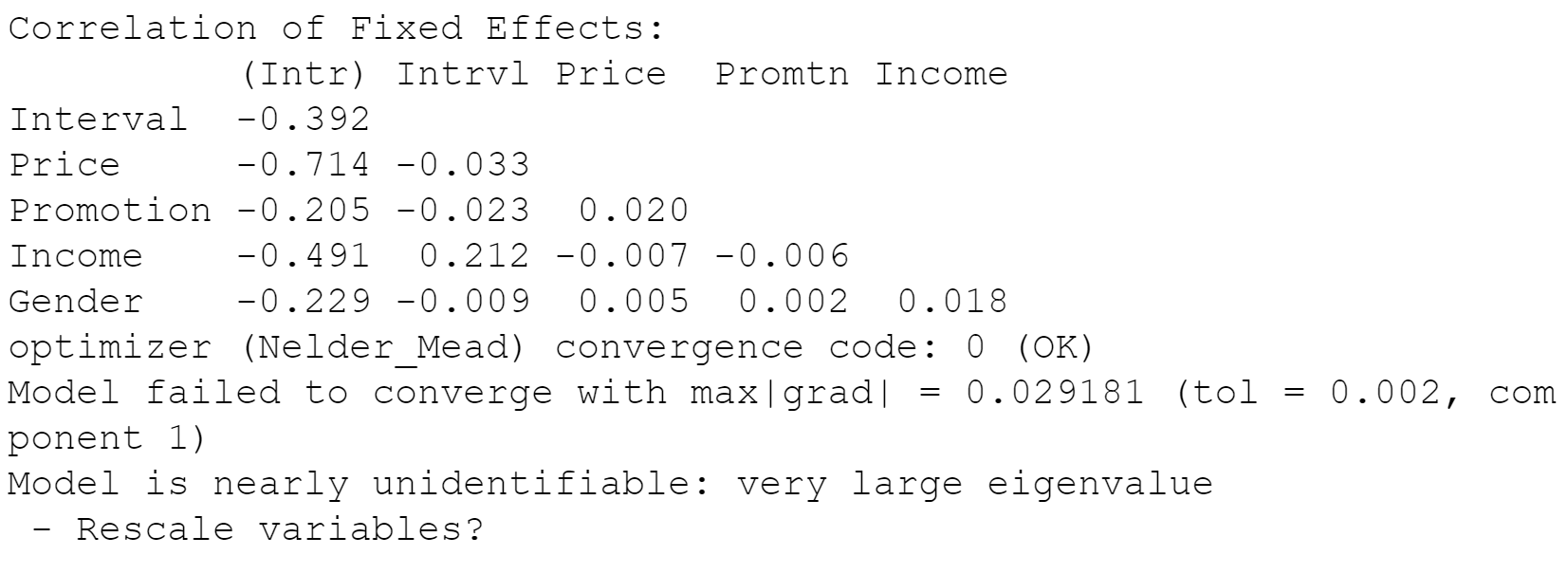
Model 1:





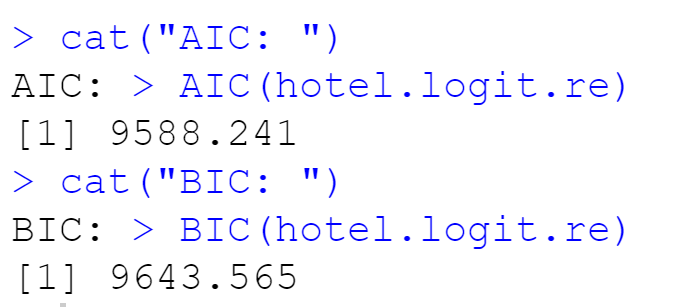
Model 2



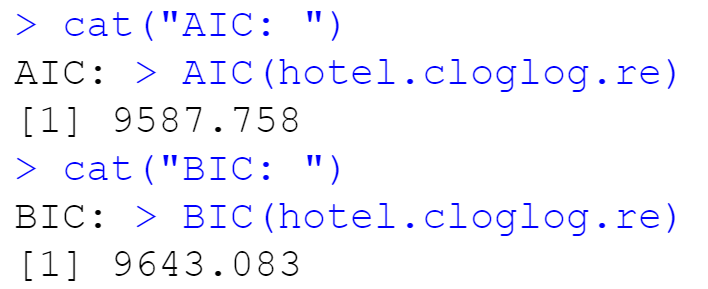


Based on the AIC's and BIC's of the five models in (6), (7) and (8), which is the best model for the data?

The AIC and BIC for Model 1 are reported below:



The AIC and BIC are reported below:



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.