Afnan Ur Rehman

Student ID: 21324930

ECOM6004 ASSESSMENT 2

1. EMPIRICAL COUNT MODELS

i) Training data:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   0.000   3.457   2.000  21.000 

   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21 
3204   23   25   35   27   74   27   20   35    5  210    3   36   10    5  176    2    9    8    1  256  133 
```

Summary Statistics of Training Data:

```
      cigs            con         age              AGE2             male            stress           actany      
 Min.   : 0.000   Min.   :1   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.: 0.000   1st Qu.:1   1st Qu.:0.2250   1st Qu.:0.1004   1st Qu.:0.0000   1st Qu.:0.3333   1st Qu.:0.0000  
 Median : 0.000   Median :1   Median :0.4000   Median :0.2286   Median :0.0000   Median :0.6667   Median :1.0000  
 Mean   : 3.457   Mean   :1   Mean   :0.4184   Mean   :0.2848   Mean   :0.4415   Mean   :0.6159   Mean   :0.5518  
 3rd Qu.: 2.000   3rd Qu.:1   3rd Qu.:0.6125   3rd Qu.:0.4430   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :21.000   Max.   :1   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
    heavdrnk         actdrink          sad               happy             illness          nowhite       
 Min.   :0.0000   Min.   :0.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000  
 Median :0.0000   Median :0.000   Median :0.00000   Median :0.00000   Median :0.0000   Median :0.00000  
 Mean   :0.2227   Mean   :0.145   Mean   :0.09598   Mean   :0.09297   Mean   :0.4126   Mean   :0.04741  
 3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :1.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000  
     single            INC2             INC3             INC4             INC5             INC6       
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   :0.4311   Mean   :0.2021   Mean   :0.1284   Mean   :0.1138   Mean   :0.1582   Mean   :0.1987  
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
     employ           olevel           alevel           males             dep            illness0          lnage      
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :2.773  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.1705   1st Qu.:0.0000   1st Qu.:3.526  
 Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.2543   Median :0.0000   Median :3.871  
 Mean   :0.5821   Mean   :0.3735   Mean   :0.3277   Mean   :0.4415   Mean   :0.3094   Mean   :0.4126   Mean   :3.819  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.4160   3rd Qu.:1.0000   3rd Qu.:4.174  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :4.564  
    LA_RENT           child           nowhite0           single0           olevels           alevel0      
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   :0.3011   Mean   :0.2914   Mean   :0.04741   Mean   :0.4311   Mean   :0.3735   Mean   :0.3277  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
    northmid          south       
 Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1.0000   Median :0.0000  
 Mean   :0.5069   Mean   :0.4931  
 3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000  
```

The mean cigarette consumption (cigs) i.e., the average amount of cigarettes per day is around 3-4 which is consistent across various summaries and suggests dataset of moderate average daily consumption. The age has a median of 0.23 while the Square of age (AGE2) is 0.40 for age. Both variables have been scaled to normalize distribution. The variable Gender (male) has the binary value of 0 or 1. The Income Categories (INC2, INC3, INC4, INC5, INC6) are binary and can be critical in finding socio and economic effect on habits of smoking. The variable Employment Status (employ) shows the employed individuals with mean of 0.58. The variable Health (illness) indicates the status of health, with a mean of 0.41 which indicates that the dataset has significant proportion having health issues. The variables Region (northmid, south) indicates geographical segmentation, useful in analysis of regions in smoking habits. The variable stress has a mean from 0.615 shows a moderate stress level across dataset. The Behavioral Indicators such as such as heavdrnk (heavy drink), actdrink (active drinking) and actany (active in any form) are significant behavioral factors which corelate with habits of smoking.

Variance:

````
```{r}
var(train_data$cigs) # variance of daily cigarettes
```

[1] 44.31438
````

The daily consumption of cigarettes in the train dataset indicates a variance of 44.31438. To understand the dispersion and distribution of consumption of cigarettes, this statistic is critical. The high variance shows the number of smoked cigarettes among individuals in datasets.

EDA

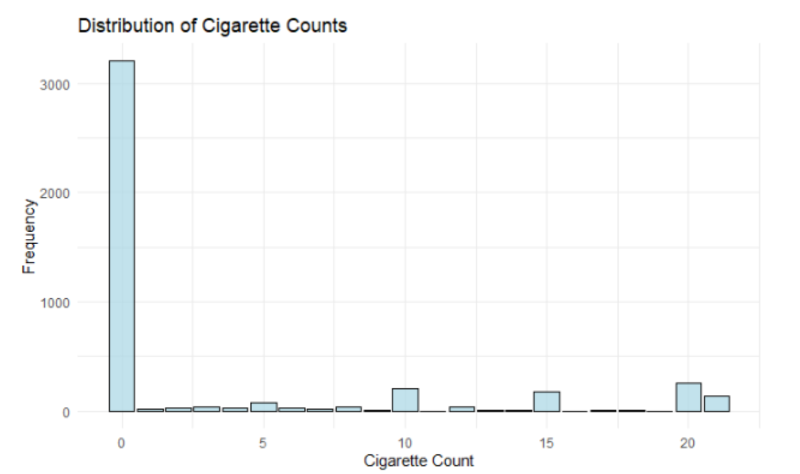
````
```{r}
count_table = table(train_data$cigs) # recommended way of summarising a count variable
count_table
```

   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21
3204   23   25   35   27   74   27   20   35    5  210    3   36   10    5  176    2    9    8    1  256  133
````

This exploratory data analysis indicates the amount of cigarette consumption frequency in training dataset. The dataset shows that the significant number of nonsmokers is 3204. This may include people who have quit smoking or are nonsmokers. The dataset indicates that few people consume 1 to 5 cigarettes daily. As the count increases, there is a noticeable drop in frequency, however the results show that 74 people smoke 5 cigarettes while 210 people smoke 10 cigarettes, who fall under this category. As the frequency decreases further, the result shows heavy smokers. There is a huge spike of 20 cigarettes which is smoked byb256 individuals and shows a common habit of consuming a pack of cigarette a day. Around 133 individuals, a smaller group consume 21 cigarettes daily, indicates a very heavy smoker.

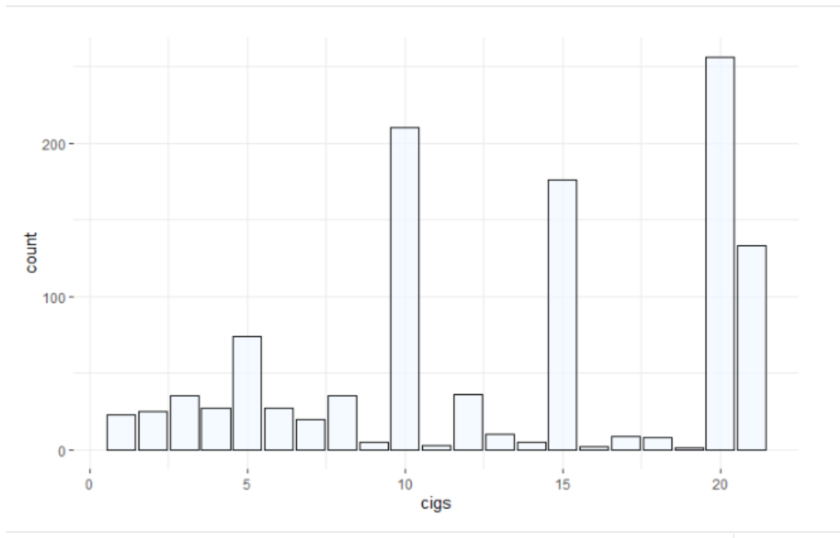
Percentage Table:

```
           0            1            2            3            4            5            6            7            8 
0.7409805735 0.0053191489 0.0057816836 0.0080943571 0.0062442183 0.0171137835 0.0062442183 0.0046253469 0.0080943571 
           9           10           11           12           13           14           15           16           17 
0.0011563367 0.0485661425 0.0006938020 0.0083256244 0.0023126735 0.0011563367 0.0407030527 0.0004625347 0.0020814061 
          18           19           20           21 
0.0018501388 0.0002312673 0.0592044403 0.0307585569 
```

These results indicate the proportions of smokers and their implications**.** The majority of people I.e., around 74.09%, do not smoke, confirms 0 inflation in data. The results above suggest that a significant number of people have quit smoking or are nonsmokers. A low percentage of people consume 1 to 9 cigarettes per day which shows a very small amount of dataset, less than 1%. This is a very low percentage, and it decreases with an increase in number of cigarettes. A noticeable number of individuals, i.e., 4.85% are the people who smoke nearly 10 cigarettes per day. This may suggest a cultural o1r social smoking pattern or smoking habits that favors a half cigarette pack. There's another significant spike of 5.92% at exactly 20 cigarettes, which likely corresponds to the common purchasing size of full pack of cigarette, suggests a pattern between the regular smokers. 3.08% of individuals consume above one pack a day.



In the sample size of 75%, the largest bar confirms a significant number of individuals do not smoke, which indicates the zero-inflation presence in data. The cigarette consumption distribution is rightly skewed, indicating the limited number of individuals who smoke more cigarettes. In count data, the skewness is typical and is related to health behavior like smoking. At 10 and 20 cigarettes per day, there is a sudden spike which indicates the behavior, potentially tied to a pack of cigarettes, as numerous smokers may prefer to consume half or whole pack



The zero inflation is displayed by given histogram which shows a large number of people with consumption of zero cigarettes. Furthermore, the distributions seen shows spike near 10 to 20 cigarettes, that aligns with the estimation of common behavior of smoking a whole pack of cigarettes.

ii) Experiment with the variables available to you and present your preferred Poisson model and explain your choice and how you arrived at it.

```
Call:
glm(formula = cigs ~ age + AGE2 + male + olevel + alevel + single +
    nowhite + INC2 + INC3 + INC4 + INC5 + INC6 + employ + sad +
    happy + actany + heavdrnk + actdrink + stress + illness +
    child + northmid + south + dep, family = poisson, data = train_data)

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.036652   0.052356  19.800  < 2e-16 ***
age          5.126347   0.243034  21.093  < 2e-16 ***
AGE2        -7.995352   0.295667 -27.042  < 2e-16 ***
male         0.215727   0.017249  12.506  < 2e-16 ***
olevel      -0.233734   0.020822 -11.226  < 2e-16 ***
alevel      -0.476276   0.025084 -18.987  < 2e-16 ***
single       0.241228   0.019526  12.354  < 2e-16 ***
nowhite     -0.778728   0.048851 -15.941  < 2e-16 ***
INC2        -0.016177   0.024447  -0.662  0.50814    
INC3        -0.246288   0.030602  -8.048 8.42e-16 ***
INC4        -0.057823   0.030602  -1.890  0.05882 .  
INC5        -0.345028   0.032085 -10.753  < 2e-16 ***
INC6        -0.584177   0.034513 -16.926  < 2e-16 ***
employ      -0.019087   0.021415  -0.891  0.37277    
sad         -0.129847   0.047008  -2.762  0.00574 ** 
happy       -0.139682   0.048567  -2.876  0.00403 ** 
actany      -0.364355   0.021175 -17.207  < 2e-16 ***
heavdrnk     0.421896   0.025671  16.435  < 2e-16 ***
actdrink    -0.008716   0.035554  -0.245  0.80633    
stress       0.408164   0.028694  14.225  < 2e-16 ***
illness      0.106854   0.018073   5.912 3.37e-09 ***
child        0.050103   0.020484   2.446  0.01445 *  
northmid    -0.011150   0.017331  -0.643  0.51999    
south              NA         NA      NA       NA    
dep          0.739616   0.041341  17.891  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 44173  on 4323  degrees of freedom
Residual deviance: 37321  on 4300  degrees of freedom
AIC: 42159

Number of Fisher Scoring iterations: 7
```

Model 1 is the model with all the relevant variables that can directly influence cigarette consumption. The residual deviance has come out to be 37,631while the AIC is 42,466. Many coefficients are significant, suggesting various predictors have substantial effects on cigarette counts except for INC2, employ, actdrink, northmid and south which are non-significant variables in the full model.

```
Call:
glm(formula = cigs ~ age + AGE2 + male + olevel + alevel + single +
    nowhite + INC3 + INC5 + INC6 + sad + happy + actany + heavdrnk +
    stress + illness + child + dep, family = poisson, data = train_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99537    0.04862  20.473  < 2e-16 ***
age          5.05303    0.23801  21.230  < 2e-16 ***
AGE2        -7.88567    0.28681 -27.495  < 2e-16 ***
male         0.21475    0.01717  12.505  < 2e-16 ***
olevel      -0.24037    0.02058 -11.679  < 2e-16 ***
alevel      -0.48645    0.02462 -19.759  < 2e-16 ***
single       0.25304    0.01880  13.463  < 2e-16 ***
nowhite     -0.76976    0.04841 -15.901  < 2e-16 ***
INC3        -0.22530    0.02633  -8.557  < 2e-16 ***
INC5        -0.32024    0.02632 -12.167  < 2e-16 ***
INC6        -0.55627    0.02869 -19.391  < 2e-16 ***
sad         -0.12955    0.04697  -2.758  0.00581 **
happy       -0.13577    0.04848  -2.800  0.00510 **
actany      -0.37004    0.01818 -20.349  < 2e-16 ***
heavdrnk     0.41629    0.01823  22.837  < 2e-16 ***
stress       0.41046    0.02867  14.316  < 2e-16 ***
illness      0.11297    0.01783   6.337 2.35e-10 ***
child        0.05265    0.02039   2.582  0.00981 **
dep          0.74135    0.04010  18.488  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 44173  on 4323  degrees of freedom
Residual deviance: 37327  on 4305  degrees of freedom
AIC: 42155

Number of Fisher Scoring iterations: 7
```

Model 2 contains only the significant variables from model 1 and all the insignificant variables have been removed from the model as they don't seem to have direct effect on cigarette consumption. The residual deviance has come out to be 37,327 while the AIC 42,155 which is lower than the model 1 with all significant and insignificant variables. Almost all the variables in model 2 are significant and have a direct relationship with cigarette consumption.

Preferred Model

Model 2 has a low residual deviance of 37 and 327 while Model 1 has residual of of 37 and 631. The low residual in Model 2 indicates it better fits the data. Furthermore, Model 2 AIC value of 42,155, is lower than the AIC value of Model 1 which is 42,466 which suggests that model 2, which has excluded the insignificant variables, is a better fit and has a better explanatory power.

Thus, the preferred model is Model 2 as the model gives better fit, lower deviance and has a lower AIC, with fewer variable as compared to model 1.

We have further applied the quasi-Poisson Model and Negative Binomial Model.

Quasi-Poisson Model

```
Call:
glm(formula = cigs ~ age + AGE2 + male + olevel + alevel + single +
    nowhite + INC3 + INC5 + INC6 + sad + happy + actany + heavdrnk +
    stress + illness + child + dep, family = quasipoisson, data = train_data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.99537    0.16821   5.918 3.52e-09 ***
age          5.05303    0.82346   6.136 9.20e-10 ***
AGE2        -7.88567    0.99227  -7.947 2.42e-15 ***
male         0.21475    0.05941   3.614 0.000304 ***
olevel      -0.24037    0.07121  -3.376 0.000743 ***
alevel      -0.48645    0.08518  -5.711 1.20e-08 ***
single       0.25304    0.06503   3.891 0.000101 ***
nowhite     -0.76976    0.16749  -4.596 4.43e-06 ***
INC3        -0.22530    0.09110  -2.473 0.013429 *
INC5        -0.32024    0.09106  -3.517 0.000441 ***
INC6        -0.55627    0.09925  -5.605 2.22e-08 ***
sad         -0.12955    0.16250  -0.797 0.425383
happy       -0.13577    0.16773  -0.809 0.418312
actany      -0.37004    0.06291  -5.882 4.37e-09 ***
heavdrnk     0.41629    0.06307   6.601 4.59e-11 ***
stress       0.41046    0.09919   4.138 3.57e-05 ***
illness      0.11297    0.06168   1.832 0.067094 .
child        0.05265    0.07053   0.746 0.455444
dep          0.74135    0.13873   5.344 9.57e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 11.96982)

    Null deviance: 44173  on 4323  degrees of freedom
Residual deviance: 37327  on 4305  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

The Quasi-Poisson Model results shows that male, Age, AGE2, single, olevel, alevel, no white, INC3, INC5, INC6, heavdrnk, stress, actany and dep are highly significant variables. Furthermore, the Quasi-Poisson model has demonstrated substantial overdispersion with having a dispersion estimate of 11.97. To account for overdispersion, quasi-Poisson model adjusts the variance without modifying the likelihood, the standard errors while depending on Poisson distribution structure. It is useful when mean and variance are expected to be different that is case in various datasets of real world.

Negative Binomial Model

```
Call:
glm.nb(formula = cigs ~ age + AGE2 + male + olevel + alevel +
    single + nowhite + INC3 + INC5 + INC6 + sad + happy + actany +
    heavdrnk + stress + illness + child + dep, data = train_data,
    init.theta = 0.09929590897, link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.988541   0.300831   3.286 0.001016 **
age          6.349858   1.293647   4.908 9.18e-07 ***
AGE2        -9.804409   1.442357  -6.797 1.06e-11 ***
male         0.379126   0.103844   3.651 0.000261 ***
olevel      -0.207177   0.132676  -1.562 0.118401
alevel      -0.461799   0.149274  -3.094 0.001977 **
single       0.375735   0.116332   3.230 0.001239 **
nowhite     -0.801177   0.240266  -3.335 0.000854 ***
INC3        -0.115484   0.157666  -0.732 0.463889
INC5        -0.420040   0.157165  -2.673 0.007526 **
INC6        -0.591185   0.157972  -3.742 0.000182 ***
sad         -0.252952   0.202209  -1.251 0.210956
happy       -0.077881   0.203974  -0.382 0.702595
actany      -0.494441   0.110910  -4.458 8.27e-06 ***
heavdrnk     0.363045   0.122380   2.967 0.003012 **
stress       0.253959   0.169530   1.498 0.134128
illness      0.108624   0.108670   1.000 0.317513
child       -0.003389   0.128397  -0.026 0.978942
dep          0.892943   0.266410   3.352 0.000803 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.0993) family taken to be 1)

    Null deviance: 2644.2  on 4323  degrees of freedom
Residual deviance: 2421.2  on 4305  degrees of freedom
AIC: 13690

Number of Fisher Scoring iterations: 1


              Theta:  0.09930
          Std. Err.:  0.00368

 2 x log-likelihood:  -13650.01600
```

Similar to Quasi-Poisson model, significant predictors in this case were found to be male, age, AGE2, nowhite, alevel, single, INC5, INC6, heavdrnk, actany, and dep. The theta (estimated dispersion parameter) has come out to be 0.0993, with AIC has come out to be 13690, suggesting that Negative Binomial model is suitable for over dispersed data modeling. In explaining cigarette consumption on the basis of statistical significance, similar sets of variables are utilized by Both models. Variables such as economic status, age and gender were retained due to the potential effect on the dependent variables.

Both models address the issue of overdispersion. The Negative Binomial model for count data is more appropriate as it deals with overdispersion and provides AIC for model comparison unlike quasi-Poisson model, with a high variance compared to the mean, as seen in analysis.

iii) Pick one variable of your choice, and comment on the effect of this variable based on their partial/marginal effect (and its standard error and/or z-score and/or p-value) as well as the value of exp(βvar i able).

The variable of choice is Age from model 1. The Coefficient (β_age) from Model 1 has come out to be 5.126347. The Standard Error for the age has come out to be is 0.243034. The z-value is 21.093 while the p-value is < 2e-16. The age coefficient of 5.126347 shows that for every additional year of age, the expected count log of smoked cigarettes increases around 5.13 units keeping all other variables in constant state. The suggestion shows that old individuals are likely to smoke more cigarettes as compared to younger individuals and the difference is very significant. The coefficient exponent is 168.42 shows that for every additional age year, the expected number of cigarettes is multiplied by 168.42 times. That means that if the age of individual increase by one years, then the expected number of smoked cigarettes count increases by 168.42 times factor, suggesting a strong positive relationship among cigarette consumption and age. The extremely low p-value of < 2e-16 and z-value of 21.093 suggests that the impact of cigarette count on age is highly significant, which means that there is a significant impact on age with the number of smoked cigarettes. There is a substantial positive impact of consumption of cigarettes on variable age, as old individuals smoke more cigarettes in comparison with young individuals. The robustness of the relationship is confirmed by statistical significance of this effect, as shown by low p-value and high z-score.

iv) Based on the estimation results of a more general count model, test for overdispersion in the Poisson model and comment on your findings and the implications for your Poisson model

For the Psn_model, the z value has come out to be 23.34 while the p-value is < 2.2e-16. The Estimated Dispersion for the first model is 11.93411. For Psn_model2 the z value has come out to be 23.318. The p-value is < 2.2e-16 while the estimated dispersion in this model is 11.93678. As both models get a p-value less than 0.001, null hypothesis is rejected, which proposes that significance evidence of overdispersion is seen in both Poisson models.

The estimation of dispersion is approximately 11.93, which is greater than 1. This suggests that response variable variance is much greater than its mean value. The assumption of Poisson distribution model is violated where the variance is equal to mean. The coefficients standard errors in Poisson model are likely underestimated, due to overdispersion, which leads to optimistic significance tests for predictors, creates incorrect inferences related to their impacts.

The overdispersion presence in Poisson model shows that for analyzing the count data of consumption of cigarettes, they might not be the best choice. Negative Binomial model leads to more reliable results about relationships among the dependent variable and independent variables.

v) Estimate two different models that explicitly allow for a preponderance of zeros in the data, and carefully explain the difference between them; make sure to carefully explain your choice of feature/explanatory variables here. Briefly comment on your model findings (note you are not expected to report any partial effects etc. here).

Two different models are estimated to clearly account for this phenomenon: the **Zero-Inflated Poisson (ZIP) Model** and the **Hurdle Model**.

1. Hurdle Model

It is comprised of two parts. The first part is a binary model that shows the count is non-zero or zero by using logistic regression. The second part is a truncated count model such as Negative Binomial or Poisson model, which predicts the counts between the positive counts like non-zero.

```
Call:
hurdle(formula = cigs ~ age + AGE2 + male + olevel + alevel + single + nowhite + INC2 + INC3 + INC4 +
    INC5 + INC6 + employ + sad + happy + actany + heavdrnk + actdrink + stress + illness | male + lnage +
    single + olevel + alevel + nowhite + child + illness + dep + LA_RENT + south, data = train_data, dist = "negbin",
    zero.dist = "binomial", link = "probit")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.3231 -0.5148 -0.3912 -0.2218  5.9902 

Count model coefficients (truncated negbin with log link):
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.255121   0.081111  27.803  < 2e-16 ***
age          3.185887   0.455237   6.998 2.59e-12 ***
AGE2        -3.709902   0.554629  -6.689 2.25e-11 ***
male         0.100424   0.031348   3.204  0.00136 ** 
olevel       0.016131   0.037165   0.434  0.66426    
alevel      -0.009980   0.044985  -0.222  0.82443    
single       0.042933   0.034723   1.236  0.21629    
nowhite     -0.236134   0.082746  -2.854  0.00432 ** 
INC2         0.035279   0.045489   0.776  0.43801    
INC3        -0.046196   0.055920  -0.826  0.40874    
INC4        -0.048746   0.056079  -0.869  0.38472    
INC5        -0.065599   0.057688  -1.137  0.25548    
INC6        -0.092200   0.062028  -1.486  0.13717    
employ      -0.017943   0.038642  -0.464  0.64240    
sad          0.069211   0.083587   0.828  0.40766    
happy       -0.106820   0.083997  -1.272  0.20348    
actany      -0.152197   0.038162  -3.988 6.66e-05 ***
heavdrnk     0.079888   0.048892   1.634  0.10226    
actdrink     0.055960   0.066139   0.846  0.39750    
stress       0.070282   0.051152   1.374  0.16945    
illness     -0.005756   0.033145  -0.174  0.86212    
Log(theta)   1.768454   0.067465  26.213  < 2e-16 ***
Zero hurdle model coefficients (binomial with probit link):
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.6261885  0.2620528   6.206 5.45e-10 ***
male         0.1388002  0.0440423   3.152  0.00162 ** 
lnage       -0.6827492  0.0612506 -11.147  < 2e-16 ***
single       0.0900848  0.0470869   1.913  0.05573 .  
olevel      -0.1808506  0.0566034  -3.195  0.00140 ** 
alevel      -0.3470722  0.0613501  -5.657 1.54e-08 ***
nowhite     -0.4904194  0.1081646  -4.534 5.79e-06 ***
child        0.1019061  0.0534221   1.908  0.05645 .  
illness      0.1295643  0.0460063   2.816  0.00486 ** 
dep          0.6074149  0.1147895   5.292 1.21e-07 ***
LA_RENT      0.3846127  0.0494034   7.785 6.96e-15 ***
south        0.0005469  0.0445938   0.012  0.99021    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Theta: count = 5.8618
Number of iterations in BFGS optimization: 49 
Log-likelihood: -5916 on 34 Df
```

```
Call:
zeroinfl(formula = cigs ~ age + AGE2 + male + olevel + alevel + single + nowhite + INC2 + INC3 + INC4 +
    INC5 + INC6 + employ + sad + happy + actany + heavdrnk + actdrink + stress + illness | male + lnage +
    single + olevel + alevel + nowhite + child + illness + dep + LA_RENT + south, data = train_data, dist = "poisson",
    link = "probit")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.6551 -0.5652 -0.4256 -0.2414  6.4382 

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.259529   0.045150  50.045  < 2e-16 ***
age          3.128764   0.257694  12.141  < 2e-16 ***
AGE2        -3.649921   0.314285 -11.613  < 2e-16 ***
male         0.096534   0.017118   5.639 1.71e-08 ***
olevel       0.018955   0.020511   0.924  0.35541    
alevel      -0.004550   0.024628  -0.185  0.85343    
single       0.041467   0.018935   2.190  0.02853 *  
nowhite     -0.249540   0.048684  -5.126 2.96e-07 ***
INC2         0.038975   0.024703   1.578  0.11462    
INC3        -0.042609   0.030833  -1.382  0.16700    
INC4        -0.043974   0.030634  -1.435  0.15116    
INC5        -0.059741   0.031775  -1.880  0.06009 .  
INC6        -0.091389   0.034418  -2.655  0.00793 ** 
employ      -0.018806   0.021161  -0.889  0.37416    
sad          0.064040   0.046356   1.381  0.16713    
happy       -0.104705   0.048317  -2.167  0.03023 *  
actany      -0.154544   0.021249  -7.273 3.52e-13 ***
heavdrnk     0.081929   0.025692   3.189  0.00143 ** 
actdrink     0.059301   0.035815   1.656  0.09777 .  
stress       0.072756   0.028851   2.522  0.01168 *  
illness     -0.005104   0.018330  -0.278  0.78065    

Zero-inflation model coefficients (binomial with probit link):
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.6263224  0.2620706  -6.206 5.45e-10 ***
male        -0.1387838  0.0440431  -3.151  0.00163 ** 
lnage        0.6827738  0.0612552  11.146  < 2e-16 ***
single      -0.0901153  0.0470878  -1.914  0.05565 .  
olevel       0.1808732  0.0566049   3.195  0.00140 ** 
alevel       0.3471012  0.0613518   5.658 1.54e-08 ***
nowhite      0.4903074  0.1081725   4.533 5.82e-06 ***
child       -0.1018941  0.0534233  -1.907  0.05648 .  
illness     -0.1295619  0.0460074  -2.816  0.00486 ** 
dep         -0.6074074  0.1147920  -5.291 1.21e-07 ***
LA_RENT     -0.3846124  0.0494045  -7.785 6.97e-15 ***
south       -0.0005504  0.0445947  -0.012  0.99015    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Number of iterations in BFGS optimization: 43 
Log-likelihood: -6354 on 33 Df
```

For count model, the hurdle uses Negative Binomial distribution, that has an advantage in handling overdispersion in counting the data. That means that it can capture variability in data which Poisson distribution may not account for effectively. The count coefficient in Hurdle Model shows a significant impact for variables such as AGE2 (-3.683) and age (3.136) which indicates that there is a relationship which is nonlinear, where smoking increases with age and decreases with high age.

In order to predict the zeros occurrence by utilizing binomial distribution with link of probit, the zero-hurdle part is separately modeled. This helps models to distinguish between those individuals who have never smoked. This zero-hurdle part is modeled by means of a binomial distribution with a probit link. This helps the model to distinguish among individuals, those who smoke a certain amount and who do not smoke at all. The age coefficients in zero hurdle model are significant, which is -2.280. This indicates that younger people are very less likely to report zero consumption of cigarettes.

ZIP Model:

```
Call:
zeroinfl(formula = cigs ~ age + AGE2 + male + olevel + alevel + single + nowhite + INC2 + INC3 + INC4 +
    INC5 + INC6 + employ + sad + happy + actany + heavdrnk + actdrink + stress + illness | male + lnage +
    single + olevel + alevel + nowhite + child + illness + dep + LA_RENT + south, data = train_data, dist = "poisson",
    link = "probit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.6551 -0.5652 -0.4256 -0.2414  6.4382

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.259529   0.045150  50.045  < 2e-16 ***
age          3.128764   0.257694  12.141  < 2e-16 ***
AGE2        -3.649921   0.314285 -11.613  < 2e-16 ***
male         0.096534   0.017118   5.639 1.71e-08 ***
olevel       0.018955   0.020511   0.924  0.35541
alevel      -0.004550   0.024628  -0.185  0.85343
single       0.041467   0.018935   2.190  0.02853 *
nowhite     -0.249540   0.048684  -5.126 2.96e-07 ***
INC2         0.038975   0.024703   1.578  0.11462
INC3        -0.042609   0.030833  -1.382  0.16700
INC4        -0.043974   0.030634  -1.435  0.15116
INC5        -0.059741   0.031775  -1.880  0.06009 .
INC6        -0.091389   0.034418  -2.655  0.00793 **
employ      -0.018806   0.021161  -0.889  0.37416
sad          0.064040   0.046356   1.381  0.16713
happy       -0.104705   0.048317  -2.167  0.03023 *
actany      -0.154544   0.021249  -7.273 3.52e-13 ***
heavdrnk     0.081929   0.025692   3.189  0.00143 **
actdrink     0.059301   0.035815   1.656  0.09777 .
stress       0.072756   0.028851   2.522  0.01168 *
illness     -0.005104   0.018330  -0.278  0.78065

Zero-inflation model coefficients (binomial with probit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.6263224  0.2620706  -6.206 5.45e-10 ***
male        -0.1387838  0.0440431  -3.151  0.00163 **
lnage        0.6827738  0.0612552  11.146  < 2e-16 ***
single      -0.0901153  0.0470878  -1.914  0.05565 .
olevel       0.1808732  0.0566049   3.195  0.00140 **
alevel       0.3471012  0.0613518   5.658 1.54e-08 ***
nowhite      0.4903074  0.1081725   4.533 5.82e-06 ***
child       -0.1018941  0.0534233  -1.907  0.05648 .
illness     -0.1295619  0.0460074  -2.816  0.00486 **
dep         -0.6074074  0.1147920  -5.291 1.21e-07 ***
LA_RENT     -0.3846124  0.0494045  -7.785 6.97e-15 ***
south       -0.0005504  0.0445947  -0.012  0.99015
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 43
Log-likelihood: -6354 on 33 Df
```

Zip Model utilizes Poisson distribution. As this model can handle count data, it also struggles with overdispersion by assuming that the variance and mean of count distribution are equal. For AGE2 (-3.650) and age (3.129), the count model of ZIP model shows significant estimates, showing similar trends to hurdle model but might not be able address overdispersion shown by the data. In order to account for the excess zeros in data, the model has binomial components, but it cannot separate the counts of zeros from one or more cigarettes, the same way as the hurdle model. The zero-inflation part is included in ZIP model which is also modeled as binomial distribution. The significant predictors in ZIP model, such as lnage (0.682) and male (-0.138) suggest that old and male individuals are linked with high probability of being in category of zero consumption. Both of the models i.e., hurdle model and zip model, have utilized similar variables to maintain consistency for model selection. In both models, gender (male), race (nowhite) and Age are included. Both variables help to survey the demographic impact on consumption of cigarettes and the chance of being nonsmoker. Age, no white, single, male inclusion shows demographic impact on behavior of smoking. The age coefficients are positive in both models which indicates that as age increases, the number of cigarettes smoking also increases till a certain point. Income categories such as INC2, INC3, etc. and Education levels such as olevel, alevel, are important in understanding the social and economic status impact on behavior of smoking. Variables such as alevel and olevel included to find the educational attainment impact on behavior of smoking. Similar impacts are seen in both models, where the alevel and olevel coefficients are not statistically significant and suggest that levels of education may strongly correlate with smoking frequency. Behavioral and health variables such as heavdrnk (heavy drinking) and actany (individual engages in activities) shows lifestyle choice which might relate with smoking. Also to evaluate smoking behavior, smoking the illness and stress are also included.

Model Results:

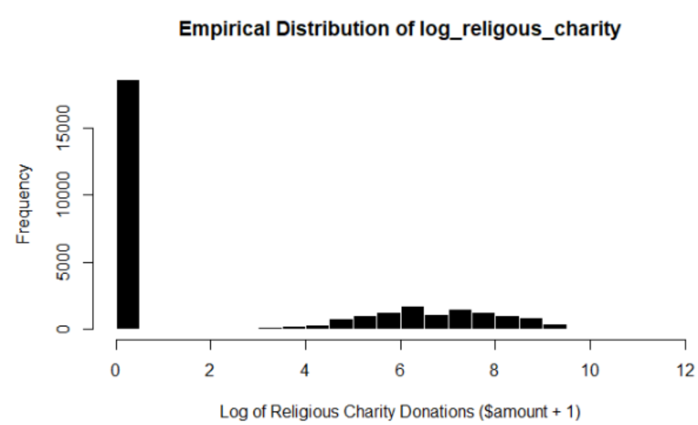
The Hurdle model count part shows significant negative effects of AGE^2 and positive effects of age, suggesting a non-linear relationship, which indicates that increase in age leads to higher consumption of cigarettes, but it is only till a certain point after which the likelihood starts decreasing. The findings of the zero hurdle also suggests that being single, experiencing stress and male increases the likelihood that these characteristics leads to lower smoking rates. The zip model suggests the negative influence of AGE^2 and positive influence of age on cigarette consumption. For zero inflation model, The ZIP model also indicates a significant impact, with social status, males and younger age also affect the zero-count reporting likelihood. In both models, the findings are consistent showing that social characteristics and stress significantly impact behavior of smoking. Both models show the preponderance of zeros effectively in data but the approach for both is different. The ZIP model gives a simple structure based on Poisson distribution while it does not fully capture overdispersion whereas the Hurdle Model is more suited to deal with data having overdispersion.

In both models, the choice of explanatory variables shows the multifaceted nature of behavior of smoking influenced by health-related variables, demographics and socioeconomic factors. As both models get significant insights, the findings show that the Hurdle model might give a nuanced understanding for factors which drive consumption of cigarettes and differentiates among processes which lead to positive counts and zero.

Due to its use of Negative Binomial distribution, the hurdle model is better suited for handling overdispersion in the count process. Although the ZIP model offers a simpler structure that might be less robust in handling variability. The explanatory variables choice shows the complex interplay of socioeconomic, health factors and demographic impacting smoking behavior.

2. <u>CENSORED MODELS</u>

i) Firstly, plot the empirical distribution of l og _r el i g ous and comment on this.



The histogram shows a positively skewed distribution which means that although most donations are low, there are some donations of high value, which increases the Mean more than the Median, which results in long right tail. The extended right tail suggests that small number of individuals make large significant donations. However, a significant number of donations are near zero, which indicates that a huge number of entities and individuals make very small donations or might not make significant contributions to religious causes. The low value donations predominance might reflect a broader trend in giving charity, there small donations are common, which are supported by high value donors of small groups. The dataset shows left censoring at zero which suggests that for some respondents the donation is observed which could be because they didn't donate but it is recorded as zero.

ii) Justify your preferred model and in terms of the summary model output only, explain your findings.

```
Call:
tobit(formula = log_religous_charity ~ num_adults + num_kids +
    male + years_of_schooling + employed + unemployed + log_labour_income +
    log_wealth + marrried + own_home + health + aged_lt_20 +
    aged_20_30 + aged_30_40 + aged_40_50 + aged_50_60 + Catholic +
    Jewish + Protestant + other_relig + black + white, left = 0,
    data = charity_data.df)

Observations:
         Total  Left-censored     Uncensored Right-censored
         30779          18653          12126              0

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.595441   0.302925 -31.676  < 2e-16 ***
num_adults         -0.340683   0.071355  -4.774 1.80e-06 ***
num_kids            0.165726   0.043580   3.803 0.000143 ***
male               -0.723181   0.136585  -5.295 1.19e-07 ***
years_of_schooling  0.374788   0.014813  25.301  < 2e-16 ***
employed           -0.790366   0.115335  -6.853 7.24e-12 ***
unemployed         -2.768042   0.270750 -10.224  < 2e-16 ***
log_labour_income   0.146690   0.012989  11.293  < 2e-16 ***
log_wealth          0.171754   0.013645  12.587  < 2e-16 ***
marrried            3.310954   0.139938  23.660  < 2e-16 ***
own_home            2.303122   0.108096  21.306  < 2e-16 ***
health              0.541871   0.044479  12.183  < 2e-16 ***
aged_lt_20         -4.634521   0.736370  -6.294 3.10e-10 ***
aged_20_30         -4.622461   0.188881 -24.473  < 2e-16 ***
aged_30_40         -3.401039   0.177611 -19.149  < 2e-16 ***
aged_40_50         -2.142747   0.161389 -13.277  < 2e-16 ***
aged_50_60         -1.263707   0.156732  -8.063 7.45e-16 ***
Catholic            1.554633   0.146515  10.611  < 2e-16 ***
Jewish             -0.050144   0.361105  -0.139 0.889559
Protestant          2.288184   0.116926  19.569  < 2e-16 ***
other_relig         1.489361   0.257459   5.785 7.26e-09 ***
black               0.807793   0.196416   4.113 3.91e-05 ***
white              -0.001360   0.184628  -0.007 0.994121
Log(scale)          1.848419   0.007491 246.758  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Scale: 6.35

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -4.992e+04 on 24 Df
Wald-statistic:  5550 on 22 Df, p-value: < 2.22e-16
```

The Tobit model has been utilized with comprehensive set of variables in the initial model which is appropriate for datasets where dependent variables such as logarithm of charity of religious donations, are censored at specific value. The donations are left censored at zero in this dataset, which indicates that various observations may be zero. log_religous_charity, in the dependent variable is the logarithm of donated amount to religious purposes plus one, while the control variables cover various aspects such as demographic, economic, employment status, housing and health, age categories, religious affiliations and ethnicity. Gender may impact on the likelihood or amount of donations, as studies suggest differing donation patterns between genders. num_adults and num_kids could influence donation behavior based on household composition. log_labour_income, log_wealth and years_of_schooling relate with the capacity of donations. The higher income and education might lead to higher charity contributions. Unemployed and employed show how status of employment impacts charitable contributions.

Health and home ownership correlate with altruistic behavior and disposable income. The variables of aged_lt_20 and aged_20_30 etc., capture the impact of age in behavior of donation, as there is a change in priorities and financial capabilities with age. Jewish, Catholic, Protestant and other_relig in religious charity context, helps us to understand the impact of various religions on behavior of donation is critical. Black and white both variables capture how identity of racial impacts the behavior of charity.

Findings:

All socioeconomic and demographic variables indicate the strong significance of p<0.001 which indicates vigorous relationship with dependent variables. The positive coefficiecnt of years_of_schooling suggests that higher education also relates with increase in donations. IN the same way, the individuals own a house and are married also impact positively in the likelihood of higher donations. The negative coefficients aged_lt_20 and num_adults, implies that as there is an increase in the number of these variables, there is a decrease in likelihood of higher donations. The number of iterations and value of log likelihood indicates the model fit. The high number of iterations shows that estimation is converged well.

The intercept has come out to be negative and is significant, which indicates the donations at base level when all other variables are kept constant. The coefficient for the number of adults has come out to be negative which suggests that as the number of adults increases in a household, the donation amount decreases. Unemployed and employed both has negative coefficients while both of the variables are significant which suggests that the employment status has a negative impact on the donated amount. The coefficient for number of kids has come out to be positive which suggests that increase in children in household is also linked with increase in donations. The coefficient of married people suggests that there is higher donations seen in married individuals. The transformation log of labor wealth and income shows a positive donation relationship which shows that wealthy people try to donate more. Protestants, Catholic and those belongs to other religions tries to donate more as compared to non-religious individuals. Around 61% of the observations are left censored.

Tobit model with interation terms:

```
Call:
tobit(formula = log_religous_charity ~ num_adults + num_kids +
    male + years_of_schooling + employed + unemployed + log_labour_income +
    log_wealth + marrried + own_home + health + aged_lt_20 +
    aged_20_30 + aged_30_40 + aged_40_50 + aged_50_60 + Catholic +
    Protestant + other_relig + black + log_labour_income:marrried,
    left = 0, data = charity_data.df)

Observations:
         Total  Left-censored     Uncensored Right-censored
         30779          18653          12126              0

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -9.759296   0.274263 -35.584  < 2e-16 ***
num_adults                  -0.341313   0.071193  -4.794 1.63e-06 ***
num_kids                     0.173863   0.043635   3.984 6.76e-05 ***
male                        -0.746298   0.136946  -5.450 5.05e-08 ***
years_of_schooling           0.373486   0.014658  25.480  < 2e-16 ***
employed                    -0.808917   0.114271  -7.079 1.45e-12 ***
unemployed                  -2.768496   0.270417 -10.238  < 2e-16 ***
log_labour_income            0.173342   0.017294  10.023  < 2e-16 ***
log_wealth                   0.172176   0.013620  12.641  < 2e-16 ***
marrried                     3.684259   0.212746  17.318  < 2e-16 ***
own_home                     2.305670   0.107872  21.374  < 2e-16 ***
health                       0.540925   0.044451  12.169  < 2e-16 ***
aged_lt_20                  -4.626895   0.736452  -6.283 3.33e-10 ***
aged_20_30                  -4.641888   0.188918 -24.571  < 2e-16 ***
aged_30_40                  -3.412839   0.177441 -19.234  < 2e-16 ***
aged_40_50                  -2.148058   0.161175 -13.327  < 2e-16 ***
aged_50_60                  -1.262636   0.156639  -8.061 7.58e-16 ***
Catholic                     1.565071   0.142715  10.966  < 2e-16 ***
Protestant                   2.300497   0.113854  20.206  < 2e-16 ***
other_relig                  1.500617   0.254914   5.887 3.94e-09 ***
black                        0.800619   0.107209   7.468 8.15e-14 ***
log_labour_income:marrried -0.046167   0.019773  -2.335   0.0196 *
Log(scale)                   1.848332   0.007491 246.752  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 6.349

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -4.992e+04 on 23 Df
Wald-statistic:  5550 on 21 Df, p-value: < 2.22e-16
```

Non-Linear Model:

```
Call:
tobit(formula = log_religous_charity ~ num_adults + num_kids +
    male + years_of_schooling + employed + unemployed + log_labour_income +
    I(log_labour_income^2) + log_wealth + marrried + own_home +
    health + aged_lt_20 + aged_20_30 + aged_30_40 + aged_40_50 +
    aged_50_60 + Catholic + Protestant + other_relig + black,
    left = 0, data = charity_data.df)

Observations:
         Total  Left-censored     Uncensored Right-censored
         30779          18653          12126              0

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.997883   0.270535 -33.260  < 2e-16 ***
num_adults              -0.334179   0.071049  -4.704 2.56e-06 ***
num_kids                 0.162837   0.043444   3.748 0.000178 ***
male                    -0.808450   0.136668  -5.915 3.31e-09 ***
years_of_schooling       0.357529   0.014676  24.361  < 2e-16 ***
employed                -0.846277   0.114091  -7.418 1.19e-13 ***
unemployed              -2.633586   0.270151  -9.749  < 2e-16 ***
log_labour_income       -0.423353   0.064050  -6.610 3.85e-11 ***
I(log_labour_income^2)   0.053439   0.005885   9.081  < 2e-16 ***
log_wealth               0.157844   0.013666  11.550  < 2e-16 ***
marrried                 3.243151   0.139754  23.206  < 2e-16 ***
own_home                 2.176202   0.108419  20.072  < 2e-16 ***
health                   0.514674   0.044437  11.582  < 2e-16 ***
aged_lt_20              -4.216269   0.732807  -5.754 8.74e-09 ***
aged_20_30              -4.496086   0.188512 -23.850  < 2e-16 ***
aged_30_40              -3.437230   0.176932 -19.427  < 2e-16 ***
aged_40_50              -2.217135   0.160905 -13.779  < 2e-16 ***
aged_50_60              -1.342244   0.156427  -8.581  < 2e-16 ***
Catholic                 1.553827   0.142399  10.912  < 2e-16 ***
Protestant               2.301952   0.113536  20.275  < 2e-16 ***
other_relig              1.512110   0.254398   5.944 2.78e-09 ***
black                    0.863534   0.107123   8.061 7.56e-16 ***
Log(scale)               1.846019   0.007490 246.478  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Scale: 6.335

Gaussian distribution
Number of Newton-Raphson Iterations: 4
Log-likelihood: -4.988e+04 on 23 Df
Wald-statistic:  5628 on 21 Df, p-value: < 2.22e-16
```

Preferred Model:

In **simplified model,** fewer variables and no interaction terms have been used which makes it easier to interpret and has the capacity to capture the main effects efficiently. Some of the interactions have been added in the **interaction model** which adds complexity to the model, but it helps in understanding the Impact of one variable on another variable such as the impact of income differs by marital status. The nonlinear model on the other hand provides an in-depth insight of how income impacts the amounts of donations but has used an added complexity.

Model	Log-Likelihood	Wald Statistic	Total Observations	Left-Censored	Uncensored	Key Significant Coefficients
Simplified Model	-4.992×10^4	5550 (22 Df)	30,779	18,653	12,126	num_adults (-0.3407), years_of_schooling (0.3747), married (3.3110)
Interaction Model	-4.992×10^4	5550 (21 Df)	30,779	18,653	12,126	log_labour_income (-0.0462)
Nonlinear Model	-4.988×10^4	5628 (21 Df)	30,779	18,653	12,126	log_labour_income (-0.4234), I(log_labour_income^2) (0.0534)

The log-likelihood is higher in case on nonlinear model, which indicates that non- linear model is a better fit as compared to simplified model and interaction model, which indicates that the underlying data structure has been captured more effectively by the addition of income complexity in the model. Wald statistics are high in case of all models which suggests that the statistical significance of overall model fit however, the high value of Wald statistics in case of nonlinear model indicates a better explanatory power of nonlinear model keeping the same amount of censored and uncensored observations in case of all models.

iii) Conduct the ‘rule-of-thumb’ specification test here and comment on your findings.

```
Call:
lm(formula = log_religous_charity ~ num_adults + num_kids + male +
    years_of_schooling + employed + unemployed + log_labour_income +
    I(log_labour_income^2) + log_wealth + marrried + own_home +
    health + aged_lt_20 + aged_20_30 + aged_30_40 + aged_40_50 +
    aged_50_60 + Catholic + Protestant + other_relig + black,
    data = uncensored_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6474 -0.8911  0.0573  0.9631  4.9312

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.183290   0.078699  65.862  < 2e-16 ***
num_adults              0.008646   0.020543   0.421   0.6738
num_kids                0.012782   0.012867   0.993   0.3206
male                   -0.080480   0.043239  -1.861   0.0627 .
years_of_schooling      0.041619   0.004357   9.553  < 2e-16 ***
employed                0.023211   0.033934   0.684   0.4940
unemployed             -0.094925   0.092284  -1.029   0.3037
log_labour_income      -0.211454   0.018056 -11.711  < 2e-16 ***
I(log_labour_income^2)  0.021614   0.001639  13.187  < 2e-16 ***
log_wealth              0.031889   0.003674   8.680  < 2e-16 ***
marrried                0.574054   0.043621  13.160  < 2e-16 ***
own_home                0.353578   0.032530  10.869  < 2e-16 ***
health                  0.080569   0.012775   6.307 2.95e-10 ***
aged_lt_20             -1.194223   0.274020  -4.358 1.32e-05 ***
aged_20_30             -0.890295   0.055660 -15.995  < 2e-16 ***
aged_30_40             -0.702761   0.050303 -13.971  < 2e-16 ***
aged_40_50             -0.389782   0.045070  -8.648  < 2e-16 ***
aged_50_60             -0.217960   0.042815  -5.091 3.62e-07 ***
Catholic               -0.393618   0.041479  -9.490  < 2e-16 ***
Protestant              0.389333   0.034432  11.307  < 2e-16 ***
other_relig            -0.014533   0.076066  -0.191   0.8485
black                   0.249580   0.030845   8.092 6.45e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.32 on 12104 degrees of freedom
Multiple R-squared:  0.1955,    Adjusted R-squared:  0.1941
F-statistic: 140.1 on 21 and 12104 DF,  p-value: < 2.2e-16
```

The Ols model has fit been fit using the uncensored observations. The results of OLS model suggest that the multiple R-squared has come out to be 0.1955 which suggests that the 19.55% of the variability in the log of religious charity donations can be explained by the model. The adjusted R-square has come out to be 0.1941. The F-statistics in case of OLS model is 140.1 on 21 and 12104 DF while the p-value has come out to be < 2.2e-16. F-statistics is very significant which indicates that the model has some explanatory variables that has meaning contribution in the model. Most of the variables in the model are highly significant and has a p value of <0.001 such as years_of_schooling showing a Positive relationship and log_labour_income which shows a negative relationship, married which indicates a positive relationship and aged_20_30, aged_30_40, etc indicates a negative relationships. While some of the non-significant variables are num_adults, num_kid, employed, unemployed and other_religion all has a p-values greater than 0.05, which suggest that they don’t have any significance in the predicting power of the model.

The OLS model findings suggest that the adjusted r square value is 19.55% indicating that 19.55% of the variance in log religious charity donations can be explained by the model. However, even though the OLS model is easier to interpret in terms of R square value and coefficients, it fails to account for zeros in the dataset, which could lead to misleading results. Considering the presence of left censoring in the dataset, Tobit model is more preferred as compared to OLS model as the Tobit model captures the relationship between the variables that are affecting the religious charity donation and provide a more accurate representation of relationship.

iv) Discard all the zero observations on log _religious and estimate an appropriate model on this new sample:

Another Ols model has been fitted after discarding the zero observations on log_religious charity. The same variables as what were used in the non-linear Tobit model have been used. Some of the significant predictors came out to be log_labour_income, years_of_schooling, log_wealth, married, own_home, health, catholic, protestant, square of log_labour_income and black. Log_labou_income has shown a negative relationship while its squared value shows a positive relationship indicates that with rising income there is decline in donations at initial levels however the donations start to increase as a higher income level. The variables married and own home have a positive impact of donations.

The tobit model all had the same significant variables but with different coefficient values. Furthermore, the tobit model indicats the importance of certain predictors such as employed and unemployed have an impact on whether if someone donates or not. The preferred model Is the tobit model as it includes both positive and zero observations which are essential in capturing the factors that drive the donation decision. It provides an insight into variables that affect the likelihood of donation. While the OLS model without zero observation is based purely on the donation size which misses a lot of behavioral variables that affects the donations behavior.

Aspect	Tobit Model (With Zeros)	OLS Model (On Non-Zero Data)
Purpose	Models both the decision to donate and donation size, accounting for zero observations (censoring).	Models ‘donation amount among donors only, without censoring.
Significant Predictors	years_of_schooling, log_labour_income, log_wealth, marrried, own_home, employed, Catholic, Protestant, black.	years_of_schooling, log_labour_income, I(log_labour_income^2), log_wealth, marrried, own_home, Catholic, Protestant, black.
Effect of Income	log_labour_income is significant, reflecting impact on both likelihood of donation and donation size.	log_labour_income shows a non-linear effect on donation amount, with the squared term also significant.
Impact of Employment Variables	Significant for employed and unemployed, indicating these factors influence donation likelihood.	Employed and unemployed are not significant, suggesting no strong effect on donation size alone.
Interpretation of Model Fit	Captures both zero and positive values, providing a more comprehensive view of donation behavior.	Focuses only on donation size among those who donate, without modeling the decision to donate.