Group Project 2

Group 1 QIJIA HE, Jianan Liu, Subasish Behera,Shreyansi Jain, Fanting Kai

2023-03-20

Contents

1 Introduction **1**

2 Exploratory Data Analysis **1**

3 Generalised Linear Model **4**

3.1 Model 1: Poisson Model 4

1 Introduction

The total number of people living in a household can be influenced by multiple variables. In certain cases, a specific type of variable may have a significant impact, which can be determined by different actual situations. By studying the correlation between variables, we can examine the living characteristics of local people.

In this dataset, we notice that six numeric variables are displayed, so we initially consider all of them. We plan to apply a generalized linear model to fit the data and discover the relevance of the variables.

At the same time, we should take into account the interaction between variables. The interaction between variables can sometimes provide insights that may not be evident when considering each variable independently. By including interaction terms in our generalized linear model, we can better understand the combined effects of these variables on the total number of people living in a household. This can help us develop a more accurate and comprehensive understanding of the factors that influence household size in the area under study.

This project begins with an explanatory analysis into the data including data graphics and summaries in 2. The formal analysis has been procured in a GLM model at which will be seen in sections 3 and **??**.

2 Exploratory Data Analysis

The exploratory analysis is started with analyzing the summaries for the numerical variables

```
 Total.Household.Income Total.Food.Expenditure Household.Head.Age
 Min.   :  11988        Min.   :  6781         Min.   :17.00
 1st Qu.: 118565        1st Qu.: 51922         1st Qu.:41.00
 Median : 188580        Median : 73578         Median :52.00
 Mean   : 269540        Mean   : 80353         Mean   :52.23
 3rd Qu.: 328335        3rd Qu.: 98493         3rd Qu.:63.00
```

```
Max.   :6042860        Max.   :327724         Max.   :99.00
Total.Number.of.Family.members House.Floor.Area   House.Age
Min.   : 1.000                 Min.   :  5.00   Min.   :  0.00
1st Qu.: 3.000                 1st Qu.: 32.00   1st Qu.: 12.00
Median : 4.000                 Median : 54.00   Median : 20.00
Mean   : 4.669                 Mean   : 90.92   Mean   : 22.98
3rd Qu.: 6.000                 3rd Qu.:102.00   3rd Qu.: 31.00
Max.   :15.000                 Max.   :900.00   Max.   :100.00
Number.of.bedrooms
Min.   :0.000
1st Qu.:1.000
Median :2.000
Mean   :2.259
3rd Qu.:3.000
Max.   :9.000
```

Table 1: Summary statistics for numerical variables.

Variable	n	Mean	SD	Min	Median	Max	IQR
Total.Household.Income	1725	269540.48	274564.17	11988	188580	6042860	139755
Total.Food.Expenditure	1725	80352.78	41194.36	6781	73578	327724	24915
Household.Head.Age	1725	52.23	14.52	17	52	99	11
Total.Number.of.Family.members	1725	4.67	2.33	1	4	15	2
House.Floor.Area	1725	90.92	99.20	5	54	900	48
House.Age	1725	22.98	15.32	0	20	100	11
Number.of.bedrooms	1725	2.26	1.44	0	2	9	1

The variables used in the dataset tend to vary together in real life scenarios, e.g. total income and total food expenditure. Analyzing this first hand has the advantage of reducing repeated examination of same relationships.

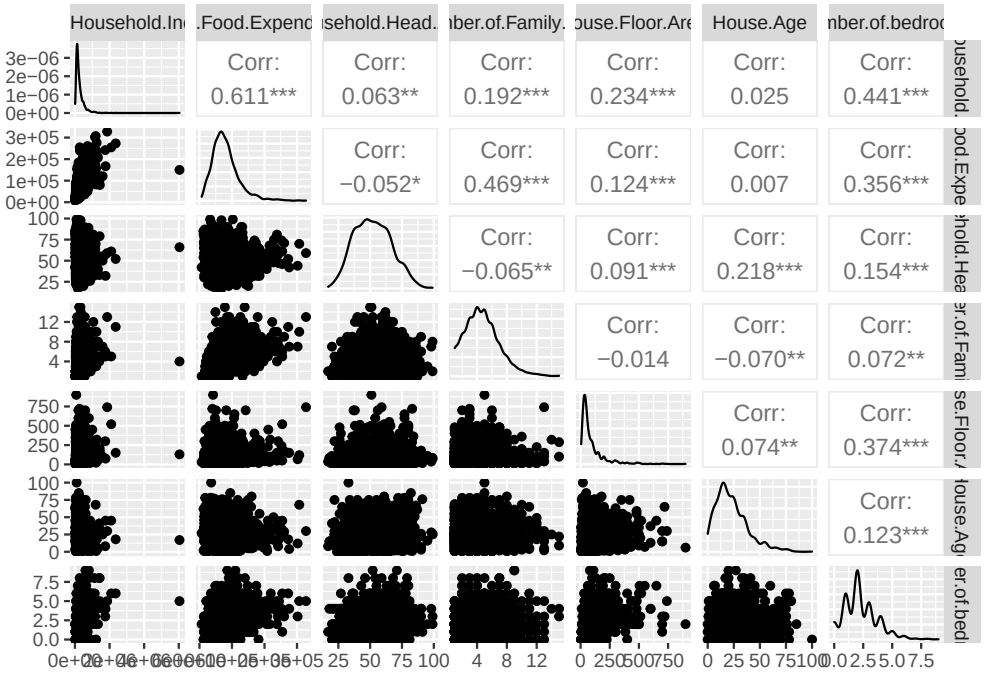

Figure 1: Correlation plots for numerical variables

It is evident from the pair plot that multi-collinearity exists. Total income is highly correlated to food expenditure. Also, the income, house area and food expenditure variables are heavily skewed to the right. This suggests using log scale for these variables.

The graphical analysis of categorical variables can be handled with bar charts and summaries.

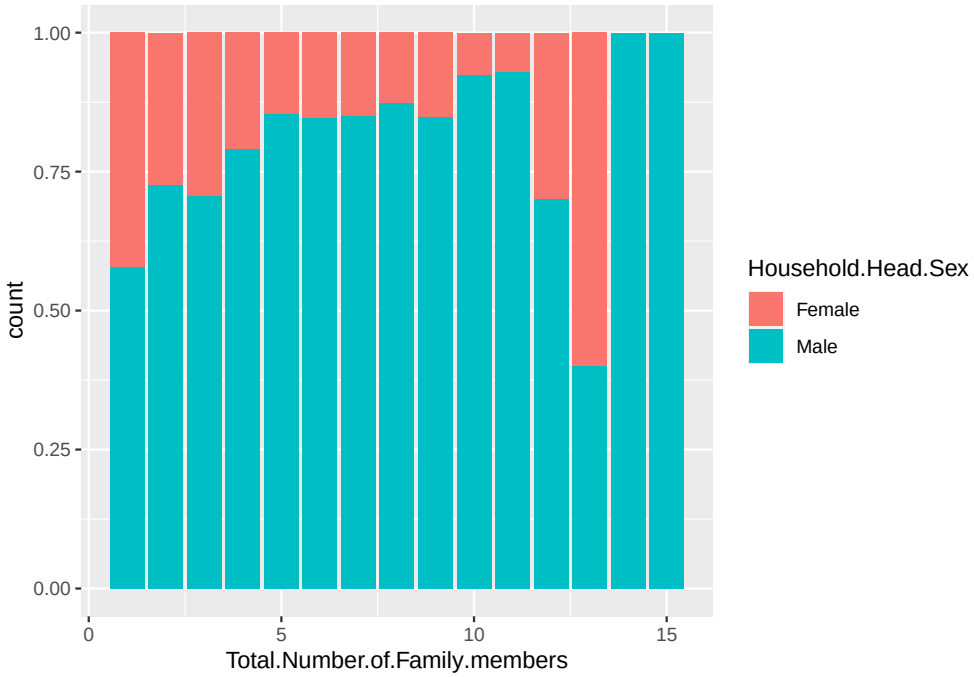

Figure 2: Distribution of number of family members across gender of household head

With an increase in total members, there is as associated increase in proportion of a male being the household head. This trend is seen up until the value of 10, after which the sample size plummets. Thus, any inference after that value is to be drawn cautiously.

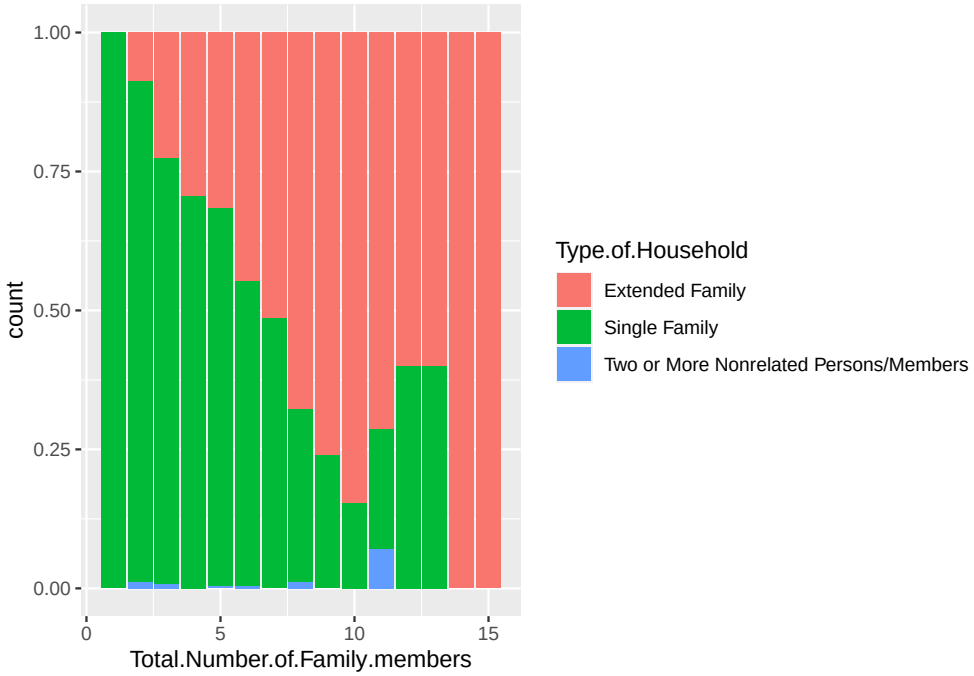

Figure 3: Distribution of number of family members across gender of household head

For 'type of household' variable, a similar trend is seen as before. With higher family members, there is higher proportion of extended family as the type.

The variables 'number of bedrooms' and 'total floor area' tend to vary together. Households having high floor area generally have a higher number of bedrooms. This suggests that once we control for the floor area, number of bedrooms will not be significant. This can tested empirically during the model fit.

Using the trends seen in the data exploration, the analysis can be extended further to examine the relationships using generalized linear models and we bring out 2 different ways to solve the research problem.

3 Generalised Linear Model

3.1 Model 1: Poisson Model

The first model that is used for this data is the possion model. This assumes that the response, given the covariates, follows poisson distribution and the mean and the variance are equal. The check for assumptions will be done after the model fitting.

We start with the full model, using all the variables, except food expenditure because of the high correlation with the income variable.

```
Call:
glm(formula = Total.Number.of.Family.members ~ log_income + Household.Head.Sex +
```

```
    Household.Head.Age + Type.of.Household + log_floorarea +
    House.Age + Number.of.bedrooms + Electricity, family = poisson(),
    data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3692  -0.7292  -0.1354   0.5286   3.6089

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                             -0.0322993  0.2258494
log_income                                               0.1182066  0.0127336
Household.Head.SexMale                                   0.2335947  0.0296685
Household.Head.Age                                      -0.0038907  0.0008554
Type.of.HouseholdSingle Family                          -0.4147942  0.0241591
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.1246197  0.1598410
log_floorarea                                           -0.0342849  0.0105677
House.Age                                               -0.0020011  0.0007731
Number.of.bedrooms                                      -0.0141210  0.0099169
Electricity1                                             0.0376406  0.0478872
                                                        z value Pr(>|z|)
(Intercept)                                              -0.143  0.88628
log_income                                                9.283  < 2e-16 ***
Household.Head.SexMale                                    7.873 3.45e-15 ***
Household.Head.Age                                       -4.548 5.41e-06 ***
Type.of.HouseholdSingle Family                          -17.169  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.780  0.43560
log_floorarea                                            -3.244  0.00118 **
House.Age                                                -2.589  0.00964 **
Number.of.bedrooms                                       -1.424  0.15447
Electricity1                                              0.786  0.43185
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1483.6  on 1715  degrees of freedom
AIC: 7166.1

Number of Fisher Scoring iterations: 4
```

Using the p-values as a metric to simplify the model, it can be seen that there are variables that need to be taken out, as they are not significant. As was suspected, the variable, number of bedrooms did not turn to be significant.Along with that, whether or not the household has electricity also did not affect the number of members. During the exploratory analysis, it was seen that the category 'Two or more non related persons' had low sample size. The corresponding high standard error could be the effect of that. This also suggests that the 'Type of household' variable could be transformed into a binary variable. A new model is then fit below:

```
Call:
glm(formula = Total.Number.of.Family.members ~ log_income + Household.Head.Sex +
    Household.Head.Age + householdtype_binary + log_floorarea +
```

```
    House.Age, family = poisson(), data = dataset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3503  -0.7354  -0.1360   0.5214   3.6785

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             0.1094018  0.2022987   0.541  0.58865
log_income                              0.1129868  0.0114992   9.826  < 2e-16
Household.Head.SexMale                  0.2294369  0.0295836   7.756 8.80e-15
Household.Head.Age                     -0.0040981  0.0008486  -4.830 1.37e-06
householdtype_binaryNot Extended Family -0.4145507  0.0240528 -17.235  < 2e-16
log_floorarea                          -0.0396252  0.0097853  -4.049 5.13e-05
House.Age                              -0.0019835  0.0007682  -2.582  0.00982

(Intercept)
log_income                             ***
Household.Head.SexMale                 ***
Household.Head.Age                     ***
householdtype_binaryNot Extended Family ***
log_floorarea                          ***
House.Age                              **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1489.3  on 1718  degrees of freedom
AIC: 7165.8

Number of Fisher Scoring iterations: 4
```

Using p-values as a metric, it is seen that all of the variables seem to be significant. The variable of 'Type of Household' was transformed to a binary variable,with its levels being 'Extended family' and 'not extended family'; the later includes both 'single family' and 'Two or more no related persons'.

The interpretation of coefficients in this model is different from that of coefficients in OLS. The model itself is multiplicative. So, for example, one unit increase in log_income means the number of members increase by 11.96%

The deviance for the model, as read from the output, is 1489.3 at 1718 degrees of freedom. This value can be compared with the chi-square quantile for assessing lack of fit. The null hypothesis. The chi-square quantile is 1815.54. The deviance is less, which suggests that fit is better than the saturated model(at 5% significance level).

After the appropriate model is fitted, the assumptions are needed to be checked. Diagnostic plots can be used which involves plot of fitted values and deviance or pearson residuals.

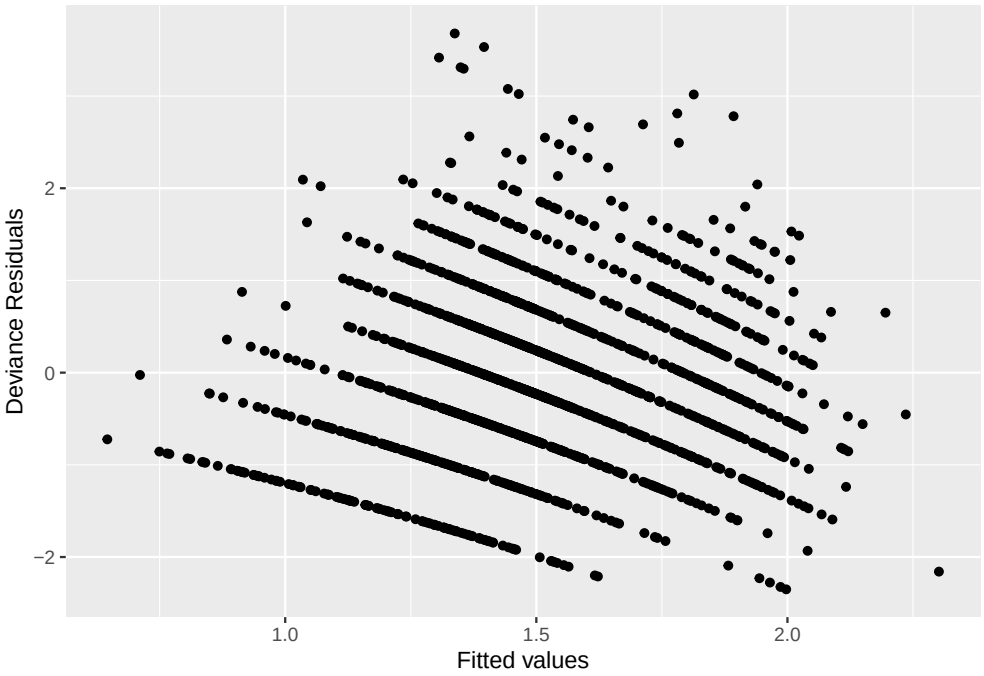

Figure 4: Residual plot for Poisson model.

The point of focus is the range of y-axis. Ideally, if the points are between +-2, it suggests a good fitting model. It can be seen that most of the points are within this range.The plot exhibits curvature, suggesting there might be some non-linearity in the relationship between the fitted values and residuals. The model may be improved with inclusion of non-linear terms in the model.

The dispersion parameter is then calculated to check for overdispersion. The estimate of the dispersion parameter for the model is 0.89. As the estimated parameter is < 1, overdispersion might not be an issue for this model. This gives assurance to the standard errors calculated for the parameters. A formal test could be done to check for the opposite case i.e., underdispersion, but this situation is unlikely in practice.

The variable 'Total food expenditure' was found to be correlated with 'Total income' and was arbitrarily excluded. The poisson model could be refitted using that variable and keeping out the income variable. The model improvement or deterioration can then judged from metrics like AIC and deviance values. Note that, the variable is used in a log base-2 scale.

```
Call:
glm(formula = Total.Number.of.Family.members ~ log_food_exp +
    Household.Head.Sex + Household.Head.Age + householdtype_binary +
    log_floorarea + House.Age, family = poisson(), data = dataset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1173  -0.6524  -0.1191   0.4602   3.7207

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -3.3491844  0.2888214 -11.596  < 2e-16
log_food_exp                       0.3265437  0.0171625  19.027  < 2e-16
Household.Head.SexMale             0.2025974  0.0295793   6.849 7.42e-12
```

```
Household.Head.Age                       -0.0017581  0.0008733  -2.013   0.0441
householdtype_binaryNot Extended Family -0.3081896  0.0248629 -12.396  < 2e-16
log_floorarea                            -0.0410523  0.0094133  -4.361 1.29e-05
House.Age                                -0.0017739  0.0007704  -2.303   0.0213

(Intercept)                              ***
log_food_exp                             ***
Household.Head.SexMale                   ***
Household.Head.Age                       *
householdtype_binaryNot Extended Family ***
log_floorarea                            ***
House.Age                                *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1214.7  on 1718  degrees of freedom
AIC: 6891.3

Number of Fisher Scoring iterations: 4
```

There seems to good improvement in the model fitting when AIC and deviance values are considered. The variables again seem to be significant at 5% significance level. The diagnostic plot, similar to the previous model, could be graphed to assess the assumptions.

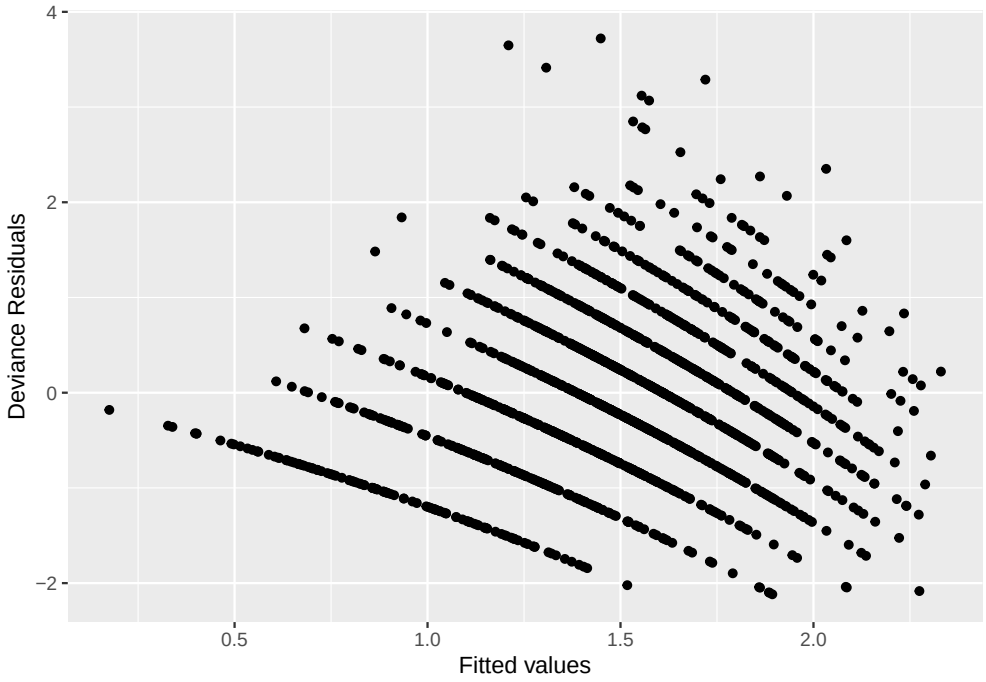

Figure 5: Residual plot for Poisson model with Food Expenditure variable.

The scale of y-axis has increased, going upto a value of 4, but the quantity of points lying outside the preferred range of +-2 has decreased. The model is relatively superior to the one previously considered that

used income instead of food expenditure.

Next we consider one last model that uses the Negative binomial distribution.

##Model 2: Negative Binomial model

In this model, the response is assumed to be distributed according to negative binomial distribution. This has an added benefit in the sense that it does not restrict the mean to be equal to the variance and thus could fit better to the data. The link function in this case is also log link and hence the interpretation of regression coefficients is similar to that of poisson model.

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ log_income +
    Household.Head.Sex + Household.Head.Age + householdtype_binary +
    log_floorarea + House.Age, data = dataset2, init.theta = 59513.02653,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3502  -0.7354  -0.1360   0.5214   3.6783

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          0.1093850  0.2023086   0.541  0.58873
log_income                           0.1129883  0.0114997   9.825  < 2e-16
Household.Head.SexMale               0.2294380  0.0295850   7.755 8.82e-15
Household.Head.Age                  -0.0040983  0.0008486  -4.829 1.37e-06
householdtype_binaryNot Extended Family -0.4145526  0.0240540 -17.234  < 2e-16
log_floorarea                       -0.0396255  0.0097858  -4.049 5.14e-05
House.Age                           -0.0019835  0.0007682  -2.582  0.00982

(Intercept)
log_income                          ***
Household.Head.SexMale              ***
Household.Head.Age                  ***
householdtype_binaryNot Extended Family ***
log_floorarea                       ***
House.Age                           **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(59513.03) family taken to be 1)

    Null deviance: 2024.2  on 1724  degrees of freedom
Residual deviance: 1489.2  on 1718  degrees of freedom
AIC: 7167.8

Number of Fisher Scoring iterations: 1

              Theta:  59513
          Std. Err.:  247760
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -7151.818
```

At 5% significance, the coefficients are again significant as was previously seen with poisson model. Not much difference in AIC and deviance is seen across the two models.

The diagnostic plot can also be graphed for the negative binomial model.

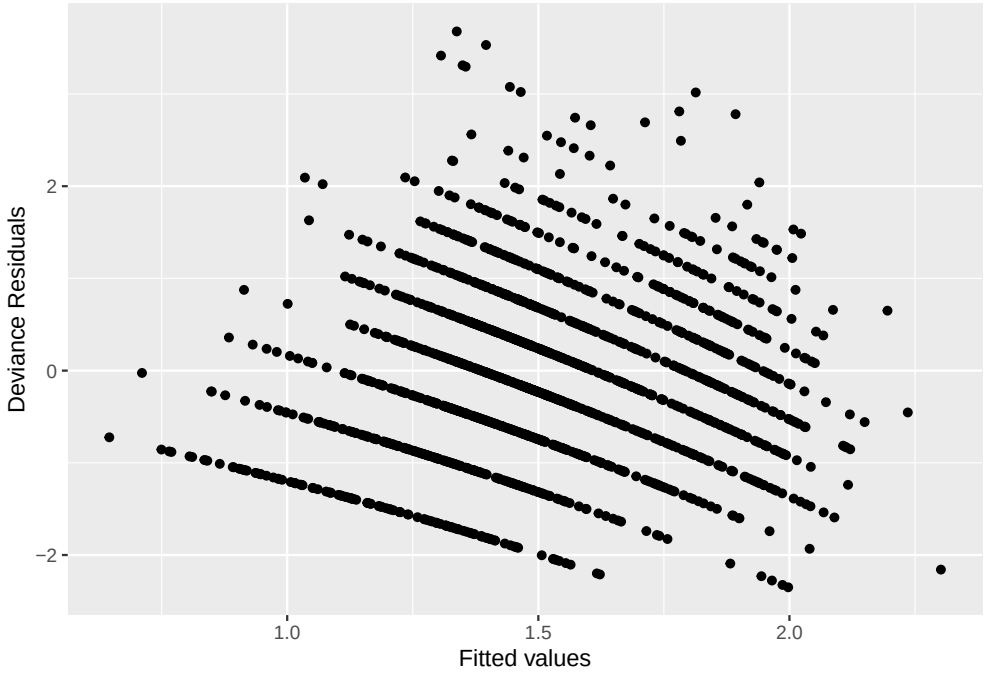

Figure 6: Residual plot for Negative binomial model.

The plot is very similar to the residual plot seen with the poisson model.

#Conclusion {#sec:conc}

Pois.fit3 superior. . . .comment on significant variables..significance level. . .