Final Project of MA5771 – Applied Generalized Linear Models
Cody Rorick
Impacts of Education, Monthly income, and Occupation on Family Size

Section 1 Introduction:

The objectives of this project are to find out if there is a relationship between Education, Occupation, and Salary with the size of families. There are studies in popular culture that have shown that people are having fewer kids and family sizes are shrinking in nations that have become wealthier and more educated. I would like to investigate if this trend persists while controlling for people within the same area. The data was obtained from the popular website kaggle.com. On kaggle, it is stated that the data was collected by customers on an application that was used for various online food orders in the City of Bangalore. The response variable is the self reported family size of the customer, including the customers themselves. This data was manipulated to place family sizes of 2 or less as a factor level, and family members of 3 or greater as another factor level. The explanatory variables will all be factor variables with levels indicating education level, occupation type, and salary range. The goal is to see if any of these variables are significant predictors for the size of family that was indicated and if there is any significant interaction between these predictors when it comes to predicting family size.

Section 2 Statistical Methods

The software being used in this analysis is the open source programming language R. The level of significance that will be used throughout is 0.05, or 95% confidence levels.

Section 2.1 Exploratory Data Analysis

Boxplots were created to compare the size of families to see if there were any obvious patterns between the explanatory variables and the response variables. The raw family size instead of the binary ‘small’ and ‘large’ family size variables was used as the y variable for these plots for better interpretability. These graphs will give a better sense of what sort of relationships to expect from the data.

Section 2.1 Generalized Linear Model

The binomial GLM will be used due to the binary nature of the response variable. The canonical logit link function will be used due to the ease of interpretability. The dispersion parameter will be examined to ensure that neither underdispersion nor overdispersion are areas of concern. If overdispersion is an issue, other models such as the quasi-binomial will be investigated. The uniform association model will be fit first. The analysis of deviance table and the AIC will then be used to decide upon a final model. The AIC will be heavily weighted in this step because a simpler model that does not contain overfitting is preferred for interpretation reasons. Once a final model is settled upon, diagnostic plots such as the qq plot, working residuals vs linear predictors, fitted values vs residuals plot, and a cook's distance plot will all be analyzed to verify that the results from the model can be trusted.

Section 3 Results and Conclusions

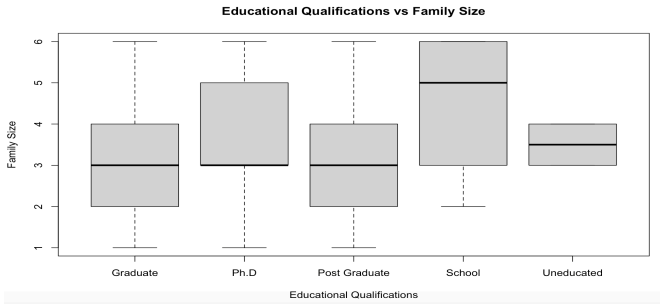

Figure 1 Education and Family Size Relationship

Figure 1 Shows that family sizes are similar amongst most education levels other than school, which appears like it may be slightly higher than the others.

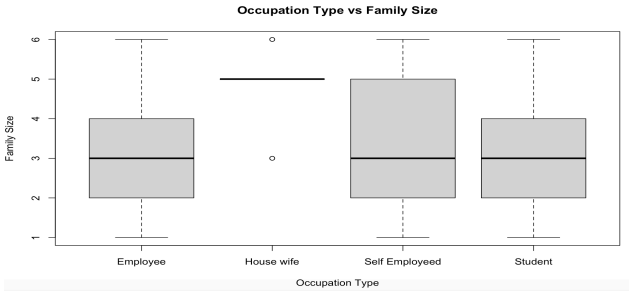

Figure 2 Occupation and Family Size Relationship

Figure 2 shows that the housewife occupation could have larger family size on average than the other occupation types. The rest of the occupation types appear to have similar family sizes.

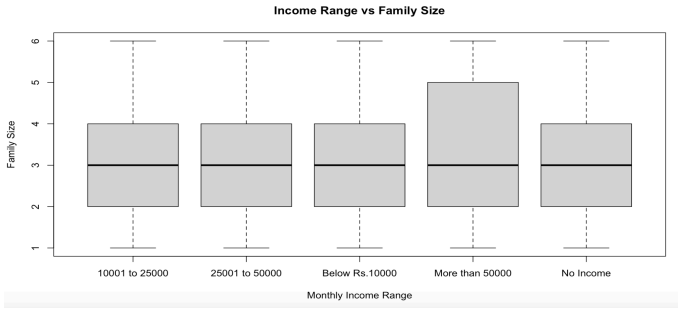

Figure 3 Monthly Income and Family Size Relationship

The boxplot in Figure 3 does not appear to show any relationship between income range and family size, family size has comparable distributions for all income ranges.

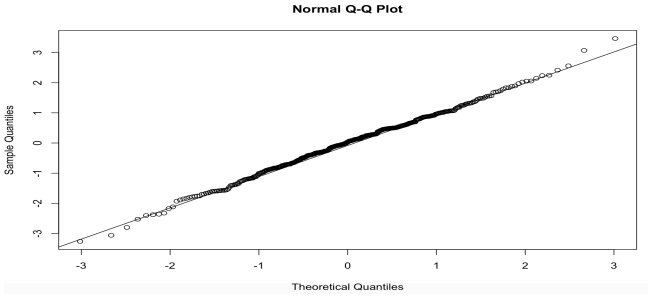

Figure 4 QQ Plot of Quantile Residuals

The QQplot in Figure 4 shows normality in the data. There are a couple points that appear like they could be outliers, as they have quantile residuals with absolute values greater than 3.

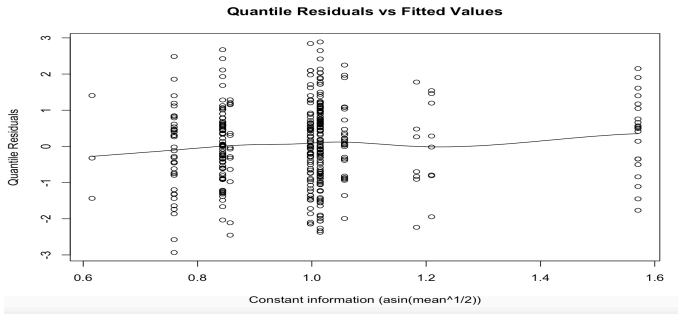

Figure 5 Quantile Residuals against Constant Information Fitted Values

The residuals in Figure 5 are shown to be scattered relatively evenly across the horizontal axis. This is shown by a nearly horizontal trend line at 0. There is no obvious curvature or trends. The variance appears to be approximately constant throughout the predicted values.

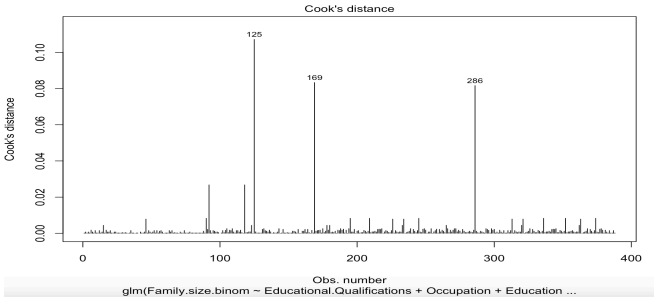

Figure 6 Cook's Distance Plot

This plot of Cook's Distance vs Observation Number in Figure 6 shows that there are not any highly influential observations. Observations 125, 169, and 286 have the largest cook's distance out of any of the data points. The assumption of no influential points in the data is being met here as these are all below the rule of thumb of 1.

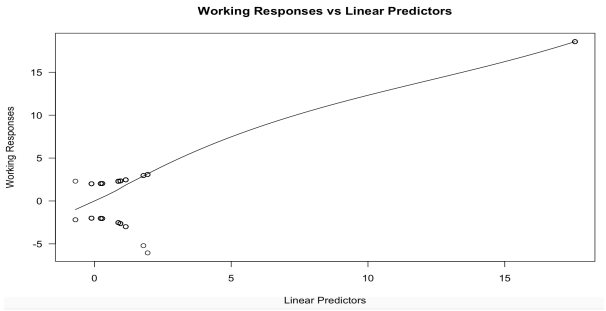

Figure 7 Working Responses Against Linear Predictors

The data in Figure 7 is somewhat difficult to interpret. This diagnostic tool is used to indicate if the correct link function is being used. The trend line drawn through the residuals is linear as would be desired out of such a plot, however there are nonlinear patterns..

Multiple models were tested based on their AIC, and the most desirable model according to this metric was the simplest, Education + Occupation + Education:Occupation. This model was chosen to use as the final model.

Table 1 Type I Analysis of Deviance of Model

Variable	Degrees of Freedom	Test Statistic	P-value

Education	4	10.84	0.0284
Occupation	3	3.84	0.2759
Education:Occupation	6	21.24	0.0017

A type I analysis of deviance table was created in Table 1 to show the significance of each variable. The only variable in the final model that was shown not to be significant at the 0.05 significance level is monthly Occupation. There are significant interaction terms involving this variable so it is still included in the final model. The chi square test was used to test the deviance and come up with the P values.

```
Coefficients: (6 not defined because of singularities)
                                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                0.8755     0.2661   3.289  0.00100 **
Educational.QualificationsPh.D                            16.6906  1142.0509   0.015  0.98834
Educational.QualificationsPost Graduate                   -0.9808     0.4200  -2.335  0.01952 *
Educational.QualificationsSchool                           0.6466     1.1640   0.556  0.57855
Educational.QualificationsUneducated                      16.4209  3956.1804   0.004  0.99669
OccupationHouse wife                                      16.6906  2284.1018   0.007  0.99417
OccupationSelf Employed                                    0.2697     0.5091   0.530  0.59630
OccupationStudent                                         -0.6406     0.3514  -1.823  0.06831 .
Educational.QualificationsPh.D:OccupationHouse wife           NA         NA      NA       NA
Educational.QualificationsPost Graduate:OccupationHouse wife  NA         NA      NA       NA
Educational.QualificationsSchool:OccupationHouse wife     -0.6466  2889.1858   0.000  0.99982
Educational.QualificationsUneducated:OccupationHouse wife -16.4209  6043.1653  -0.003  0.99783
Educational.QualificationsPh.D:OccupationSelf Employed    -18.5289  1142.0517  -0.016  0.98706
Educational.QualificationsPost Graduate:OccupationSelf Employeed   0.1234     0.8102   0.152  0.87896
Educational.QualificationsSchool:OccupationSelf Employed      NA         NA      NA       NA
Educational.QualificationsUneducated:OccupationSelf Employed  NA         NA      NA       NA
Educational.QualificationsPh.D:OccupationStudent          -14.9795  1142.0514  -0.013  0.98953
Educational.QualificationsPost Graduate:OccupationStudent   1.6970     0.5195   3.267  0.00109 **
Educational.QualificationsSchool:OccupationStudent            NA         NA      NA       NA
Educational.QualificationsUneducated:OccupationStudent        NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Figure 8 Wald Test on Parameters

The Wald test on each coefficient in Figure 8 shows that an education level of postgraduate, and the combination of being a postgraduate and a student are both significant at the 0.05 significance level. There is heightened standard error for multiple variables, indicating that collinearity could be an issue in the model. Everything else being held constant, someone with a postgraduate education is a factor of 0.375 times as likely than someone with a graduate (reference) education to have a large family. Someone with a postgraduate education and a student occupation is 2.05 times as likely to have a large family than someone with a graduate level education and an employee (reference) occupation type.

Section 4 Discussion

Education level was shown to be significant in predicting family size at the 0.05 significance level. Occupation alone was not shown to be significant in predicting family size, however the interaction between education and occupation were shown to be significant, so it was included in the final model. An interesting relationship was found with post graduate education and occupation type. Post graduate educated individuals who were currently students were found to be more likely to have larger family sizes than just people that were educated at the postgraduate level. This could have a lot of interpretations. Potentially people that answered at a student level were younger, and were considering their siblings and parents as their family instead of their own children. In order to better understand the relationship between family size and education level, income, and occupation, it might be beneficial to collect data with more specific constraints on family size. A question such as, ‘how many children do you have’ might do a better job at finding the relationships that I was desiring to uncover. There was also a significant amount of collinearity muddying the results.