Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1.Rawat,Nikhil 2.Wen,Hongbo 3. Bui, Dan	4.Verma, Shubham 5. 6. 7. **GROUP NUMBER:** **3**
MSc in: Business Analytic	
Module Code: SMM634	
Module Title: Analytic Methods for Business	
Lecturer: Rosalba Radice	**Submission Date: 8 December 2023**

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: %

Question 1

A) The Gamma Generalised Linear Model (GLM) is chosen to analyse the "HousePrices" dataset for two main reasons:

Positive and Skewed Distribution of Response Variable: The Gamma GLM is ideally suited for data where the response variable (house prices in this case) is positive and skewed. Given that house prices cannot be negative and typically exhibit a right-skewed distribution, the Gamma GLM's inherent ability to model such data is particularly advantageous.

Variance-Mean Relationship: A key feature of the Gamma distribution is that the variance of the response variable is proportional to the mean squared. This characteristic is relevant for housing data, where the variability in prices often increases with higher property values.

Some assumptions that the Gamma GLM must hold include: 1) Error terms associated with the dataset adhere to a gamma distribution. 2) The model presupposes homoscedasticity, implying that the residuals exhibit constant variance across the spectrum of predictions. 3) Residuals are independent of one another. 4) Residuals should be normally distributed. 5) Assume that no single observation disproportionately influences the overall model estimates.

These assumptions are integral to ensuring the structural soundness of the Gamma GLM and the reliability of its interpretations regarding house prices.

B) Assessing the gamma regression output, the intercept serves as the estimated baseline (β0) for housing prices, starting at $22,895.09 when all other predictors are 'zero' or 'no' for binary variables (e.g., prefer variable). Looking at individual estimates, some interesting output, including bathrooms, stands out, revealing that each additional bathroom is associated with a substantial 1.176619 times increase to the intercept value, and the presence of gas heating contributes significantly, with a rate of 1.203573. This implies that homes with gas heating systems have an estimated price around 1.2 times higher than the base price coefficient for air conditioning, which indicates a rate of 1.188344. Furthermore, "lot size" has a slight positive effect, raising the estimated price by around 1.000051 times for each additional unit; intuitively, this makes a lot of sense as the value of the "lotsize" variable is usually high; therefore, the coefficient must be small to not inflate the response variable. On the other hand, the presence of extra bedrooms contributes positively, with each additional bedroom correlating to an estimated response about 1.039453 times larger.

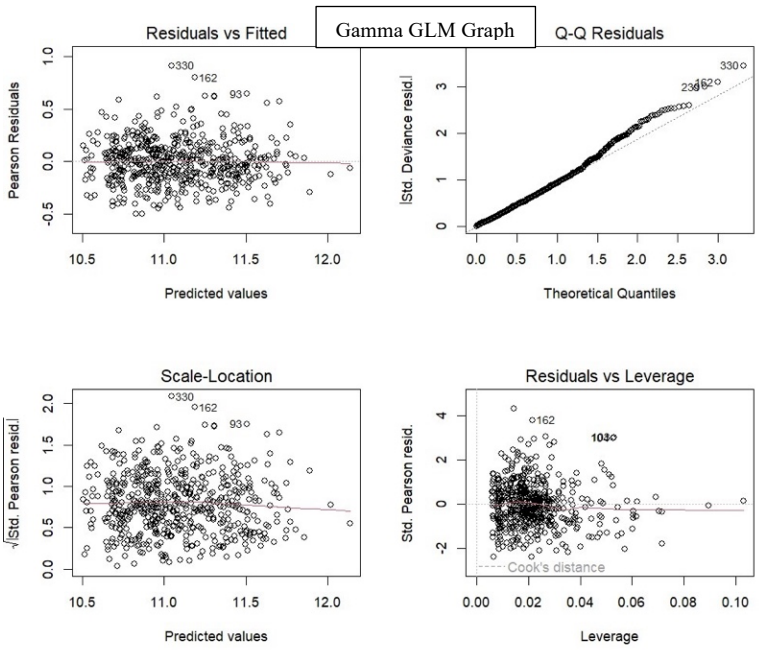
	Gamma GLM					
Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	Exponential Estimates	Monetary value($)
(Intercept)	10.04000	0.0469	213.994	< 2e-16 *	22895.090	22895.09
lotsize	0.00005	0.0000	10.481	< 2e-16 *	1.000	1.17
bedrooms	0.03869	0.0144	2.686	0.00745 **	1.039	903.28
bathrooms	0.16260	0.0205	7.934	1.25e-14 *	1.177	4043.71
stories	0.09199	0.0127	7.226	1.73e-12 *	1.096	2206.10
driveway	0.13320	0.0281	4.734	2.83e-06 *	1.142	3262.44
recreation	0.07264	0.0261	2.779	0.00565 **	1.075	1725.01
fullbase	0.10180	0.0219	4.660	3.99e-06 *	1.107	2454.06
gasheat	0.18530	0.0443	4.185	3.33e-05 *	1.204	4660.82
aircon	0.17260	0.0214	8.065	4.84e-15 *	1.188	4312.15
garage	0.05525	0.0116	4.777	2.30e-06 *	1.057	1300.58
prefer	0.11790	0.0230	5.132	4.02e-07 *	1.125	2863.74

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for Gamma family taken to be 0.04503212)
Null deviance: 76.537 on 545 degrees of freedom
Residual deviance: 24.064 on 534 degrees of freedom
AIC: 11951
Number of Fisher Scoring iterations: 4
RSME: 15744.56

$$\log(\mu) = \beta_0 + \beta_{lotsize}x_{i1} + \beta_{bedrooms}x_{i2} + \beta_{bathrooms}x_{i3} + \beta_{stories}x_{i4} + \beta_{driveway}x_{i5} + \beta_{recreation}x_{i6} + \beta_{fullbase}x_{i7} + \beta_{gasheat}x_{i8} + \beta_{aircon}x_{i9} + \beta_{garage}x_{i10} + \beta_{prefer}x_{i11}$$

When we attempt to omit variables, the AIC increases, so to maintain the AIC level, we concluded from the above explanation that all the variables were significant. The Akaike Information Criterion (AIC) is calculated at 11951, signifying the effectiveness of our model in capturing essential patterns in the data. The dispersion parameter we have is (0.04503212). We use it to measure how much the observed values vary from the predicted values. Generally, the lower this value, the better our model is at simulating housing prices, as it suggests less variability in the data. The output shows a residual deviance of 24.064 on 534 degrees of freedom. Considering how the response variable “price” typically has a mean of 68122, this amount of deviance can be deemed acceptable.

Graph Analysis



Residuals vs. Fitted: This plot is used to check the homoscedasticity assumption (constant variance of residuals). Ideally, we expect to see a random scatter of points with no discernible pattern. The red line should be approximately horizontal at zero, which seems to be the case here. There doesn't appear to be any systematic pattern, suggesting that the model's variance of residuals is fairly constant.

Normal Q-Q Plot: This plot compares the distribution of residuals (errors) to a normal distribution. The points should fall along the 45-degree line if the residuals are normally distributed. The plot shows slight deviations from normality at the tails, indicating that extreme values may not be perfectly modelled. However, the deviation is not extreme, which is often acceptable in large samples.

Scale-Location Plot: This plot, also known as the Spread-Location plot, is another way to check homoscedasticity. The y-axis shows the square root of the standardised residuals, which should be spread randomly with no particular pattern for well-fitted models. The fairly random spread without a clear pattern suggests that the homoscedasticity assumption is reasonable.

Residuals vs. Leverage Plot: This plot helps identify influential cases (outliers that have an undue influence on the model fit). Points that stand out far to the right on the x-axis or far away from zero on the y-axis are of particular interest. The Cook's distance lines (dashed lines) help identify influential points. Points outside Cook's distance lines might be influential. There are a few points, like point 330, that stand out, indicating they may be influential. It might be worth investigating these further.

In summary, the diagnostic plots indicate that the model fits the data reasonably well. There is no evidence of problematic patterns in the residuals, suggesting that the assumptions of constant variance and (to a lesser extent) normality of residuals are met. A few potential outliers and influential observations are identified, which is common in real-world data and does not necessarily undermine the model.

C) Comparative Analysis

Before, we used a linear model for normal distribution for the house price model, which is hard to compare with the GLM gamma output. In fact, we know that the LM for a normal distribution is a subset of the GLM, specifically the same as the GLM Gaussian, where the link is equal to 1. Evidently, the coefficient output from LM and GLM (gaussian) is the same.

Significance: Some coefficients have different levels of significance in the two models. For instance, the p value for 'bedrooms' is not significant in the Gaussian GLM (8%) but is significant in the Gamma GLM (0.7%).

Intercept: In the Gaussian model, when all the independence variables are zero, the intercept (β0) shows a negative house price, which can't be true. On the other hand, Gamma GLM has given a positive house price, which makes more sense in the real world.

Deviance: The Gaussian GLM null deviance (3.8860e+11) is greater than the Gamma GLM (76.537), which shows how well the Gamma GLM predicts the response variable with only the intercept. The Gamma GLM (24.064) has a lower residual deviance, indicating that it explains the variance in house prices more effectively than the Gaussian GLM (1.2703e+11).

AIC (Akaike Information Criterion): The AIC of the Gamma GLM (11951) is lower than the Gaussian GLM (12094), proving it is providing a better balance between model fit and complexity. Lower AIC values indicate a better-fitting model.

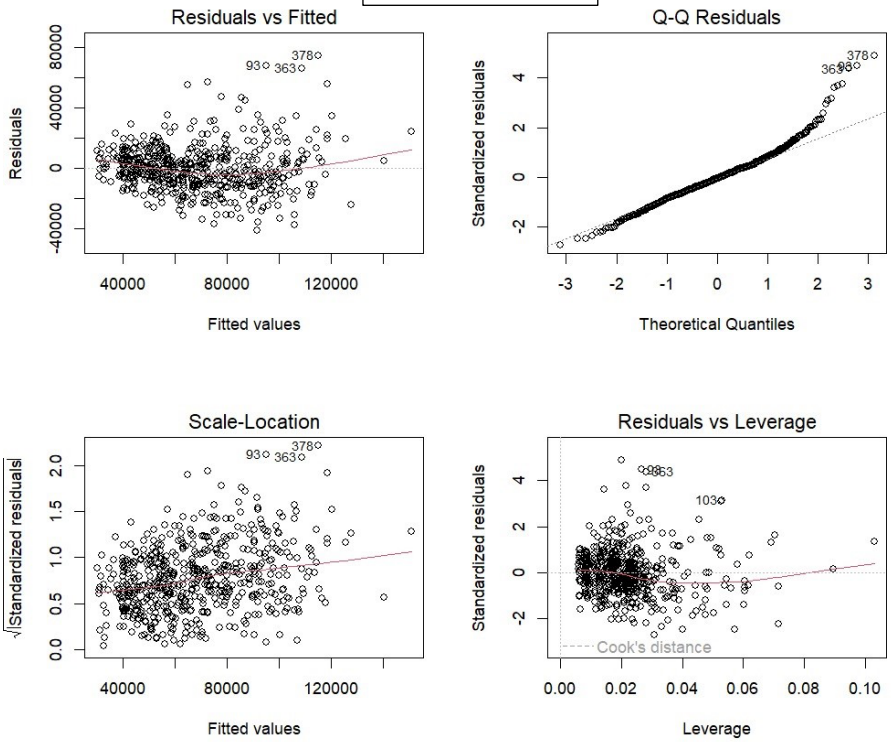
Root Mean Squared Error (RMSE): To make it more comparative, we also check the root mean squared error (RMSE) for each model. The Gamma GLM (15744.56) has a lower RMSE in comparison to the Gaussian GLM (15808.47). Which means Gamma GLM predictions are closer to the actual observed values on average than Gaussian GLM.

Plot Analysis: Graphical analyses further support the superiority of the Gamma GLM. In QQ plots and residual vs. fitted graphs, the Gamma GLM consistently displays patterns closer to normality, confirming its robustness in capturing the underlying structure of the data.

In conclusion, based on the comparison of the specific results from both models, the Gamma GLM demonstrates a better fit for the "HousePrices" dataset. It more effectively captures the data's variability and skewness, provides a potentially more intuitive understanding of the effects of predictors, and, overall, represents a more suitable model for this analysis than the Gaussian GLM used in Assignment 1.

Appendix (Question 1)

Gaussian GLM Graph



	Gaussian GLM			
Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4038.35	3409.47	-1.18	0.236762
lotsize	3.55	0.35	10.12	< 2e-16 *
bedrooms	1832.00	1047.00	1.75	0.080733 .
bathrooms	14335.56	1489.92	9.62	< 2e-16 *
stories	6556.95	925.29	7.09	4.37e-12 *
driveway	6687.78	2045.25	3.27	0.001145 **
recreation	4511.28	1899.96	2.37	0.017929 *
fullbase	5452.39	1588.02	3.43	0.000642 *
gasheat	12831.41	3217.60	3.99	7.60e-05 *
aircon	12632.89	1555.02	8.12	3.15e-15 *
garage	4244.83	840.54	5.05	6.07e-07 *
prefer	9369.51	1669.09	5.61	3.19e-08 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 237874666)
Null deviance: 3.8860e+11 on 545 degrees of freedom
Residual deviance: 1.2703e+11 on 534 degrees of freedom
AIC: 12094
Number of Fisher Scoring iterations: 2
RSME: 15808.47

Question 2

A) Upon inspecting the response variable ("visits") data using the summary code in R, we can categorise the "visit" variable as count data since it is positive discrete data that ranges from 0 to 9. In addition, since the goal is to examine the "number of doctor visits in the last two weeks" and attempt to model it, it is assumed that the data will follow a Poisson distribution since it can be used for the probability of an event happening a certain number of times within a given interval of time or space. Therefore, we are determined to use generalized linear models with the Poisson family.

The generalised linear model with the Poisson family has the following assumptions: **1)** The response variable is Poisson-distributed. **2)** Observations are independent. **3)** It has a link function of "log." **4)** The mean and variance are equal. **5)** $g(\mu_i)$, $\log(\mu_i)$ and predictors have a linear relationship. In addition, if the models have been correctly specified, then **6)** the residuals are normally distributed around 0 with constant variance. **7)** Also, the dataset should not be affected by influential points.

$$\log(\mu) = \beta_0 + \beta_{gender}x_{i1} + \beta_{age}x_{i2} + \beta_{income}x_{i3} + \beta_{illness}x_{i4} + \beta_{reduced}x_{i5} + \beta_{health}x_{i6} + \beta_{private}x_{i7} + \beta_{freepoor}x_{i8} + \beta_{freerepat}x_{i9} + \beta_{nchronic}x_{i10} + \beta_{lchronic}x_{i11}$$

In total, 11 possible variables were given to estimate the response variable "visit". Therefore, we implement a top-down approach and start analysing by fitting all 11 variables and omitting problematic predictors afterward. However, after trial and error, the full model that included all 11 variables was chosen as it provided the best regression output.

B) To assess whether a predictor should be omitted from the model or not, each predictor variable was plotted against the response variable "visit." Although it should be noted that the relationship between response and explanatory variables can differ when considered individually versus when it is viewed in a multivariate setting, it is still deemed useful to gain insights when considering omitting variables. When examining the relationship between each individual variable and "visit," it is found that "gender," "income," "private," "freepoor," and "nchronic" variables seem to have little to no visible relationship with the number of visits to the doctor (see Question 2 Appendix). As a result, these factors were omitted to create a reduced model.

The output is as follows:

MODEL FIT:
$\chi^2(11) = 1254.69$, $p = 0.00$
Pseudo-R^2 (Cragg-Uhler) = 0.27
Pseudo-R^2 (McFadden) = 0.16
AIC = 6735.70, *BIC* = 6814.35

Standard errors: MLE

	exp(Est.)	2.5%	97.5%	z val.	p
(Intercept)	0.14	0.12	0.17	-19.22	0.00
gendermale	0.86	0.77	0.95	-2.79	0.01
age	1.32	0.95	1.83	1.68	0.09
income	0.83	0.70	0.98	-2.19	0.03
illness	1.20	1.16	1.25	10.19	0.00
reduced	1.14	1.12	1.15	25.18	0.00
health	1.03	1.01	1.05	3.05	0.00
privateyes	1.13	0.99	1.31	1.77	0.08
freepooryes	0.65	0.45	0.92	-2.44	0.01
freerepatyes	1.09	0.91	1.30	0.91	0.36
nchronicyes	1.12	0.99	1.28	1.76	0.08
lchronicyes	1.16	0.99	1.37	1.83	0.07

Table 1: Full Model Output

Table 2: Reduced Model Output

MODEL FIT:
$\chi^2(6) = 1222.07$, $p = 0.00$
Pseudo-R^2 (Cragg-Uhler) = 0.27
Pseudo-R^2 (McFadden) = 0.15
AIC = 6758.32, *BIC* = 6804.20

Standard errors: MLE

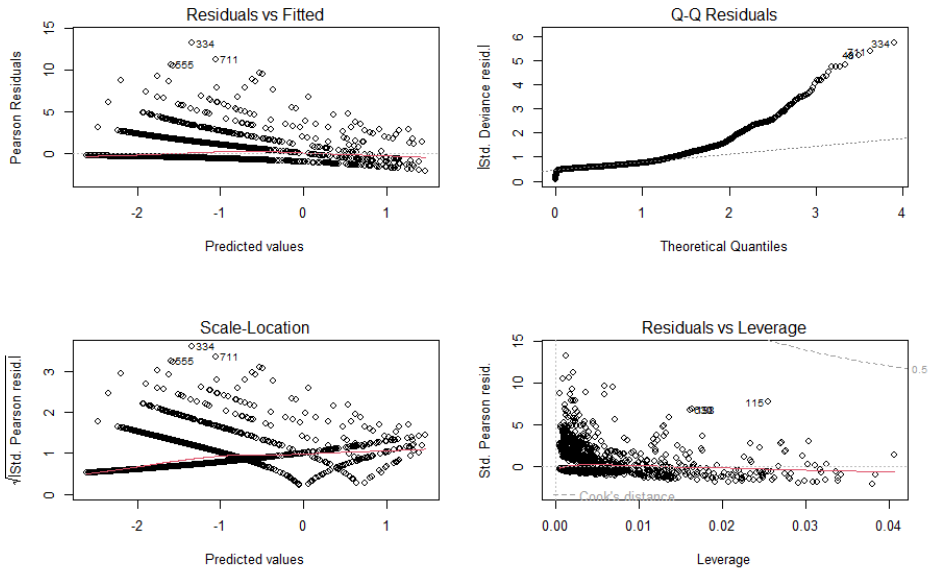
	exp(Est.)	2.5%	97.5%	z val.	p
(Intercept)	0.11	0.10	0.13	-31.43	0.00
age	1.86	1.38	2.50	4.09	0.00
illness	1.22	1.18	1.26	11.36	0.00
reduced	1.13	1.12	1.15	25.14	0.00
health	1.03	1.01	1.05	3.08	0.00
freerepatyes	1.05	0.92	1.20	0.77	0.44
lchronicyes	1.07	0.94	1.21	0.99	0.32

Since it is not possible to directly compare residual deviant when using Poisson generalised multiple linear, we chose to calculate the chi-squared p value using the pchisq function, which examines the goodness-of-fit. For the full model, assuming the residual deviant is 4380.1 with 5178 degrees of freedom, the calculated p value is 1. Interestingly, the reduced model also got an output of 1,

assuming a residual deviant value of 4412.7 with 5183 degrees of freedom. Judging p values at the 5% significant level, we reject the null hypothesis that the current model and the saturated model are not statistically different. This means that both the full model and reduced model manage to simulate the response variable "visit" with 100% goodness-of-fit. Although this indicates that the model is a good fit, it is alarming to see a p value of 1, especially considering how the p value does not change even when predictors are omitted in the reduced model.

There are two reasons for the selection of the full model. Firstly, a common problem observed while omitting variables in this exercise is that the p value of predictors tends to worsen. Evidently, when comparing the p values from Tables 1 and 2, specifically the p values of the "lchronic" and "freerepat" variables, they significantly increased from 0.07 and 0.36 in the full model to 0.32 and 0.44, respectively, in the reduced model. This was also observed in other variable combinations as well. Secondly, the AIC value, which estimates the prediction error of the model, demonstrates an increasing trend whenever variables are omitted. The AIC value of 6735.7 from the full model was the lowest value observed from all the variable combinations tested.

The coefficients have been exponentiated to reverse the "log" effect. The intercept as observed in Table 1 represents a woman with age, income, number of illnesses within the last two weeks, number of days. of reduced activity in the past two weeks due to illness or injury and a health questionnaire score using Goldberg all of 0. Also, she does not have private health insurance or free government insurance due to low income, old age, disability, or veteran status; nor does she have any chronic condition that limits or does not limit activities. Then the exponentiated explanatory variable coefficients in Table 1 represent the rate that each unit increase in that variable will increase the base estimate value. For example, the exponentiated estimate for age is 1.32, which represents, that for each additional unit in age, the base intercept value (0.14) will increase at a rate of 1.32. The Same principle can be applied to interpret the rest of the coefficient outputs. As expected, the coefficient for "age" has the highest weight in predicting the number of visits to the doctor; intuitively, an old person is more prone to health issues and requires a more frequent check-up to monitor their current condition. However, it should be noted that the age variable is recorded in fractions instead of whole discrete numbers; for example, a 36-year-old person will be recorded as 0.36 instead of 36. Furthermore, the p value of age is observed to be 0.09, which can be concluded to be insignificant at the 5% level; however, it can be interpreted as an error in the regression when all variables are being used as the p value for "age" becomes significant at the 0.1% level, as observed in Table 2 with the reduced model. Illness is the variable with the second highest coefficient of 1.2. For each number of illnesses that a person has, the person will be more likely to visit at a rate of 1.2 compared to a person who does not have any illness. It is understandable that this variable has a high contribution rate, as intuitively, a person who experiences more symptoms of illness will be more likely to visit the doctor for medical assistance. The "freerepat" variable is an interesting variable as, while it has a relatively high coefficient, its p-value is always high (0.36 in the full model and 0.44 in the reduced model) rendering it highly insignificant. However, it is believed that it remains an important indicator due to its intuition about the variable. This variable is for individuals who are either old, disabled, or have veteran status. One potential rationale for its high p value might be due to its collinearity characteristic with "age," "lchronic" or "nchronic" variables.



Graph 1. Residual Analysis

Residual analysis is important to examine whether the stated assumptions in part a) hold true when the Poisson generalised linear model is being used to model this problem set.

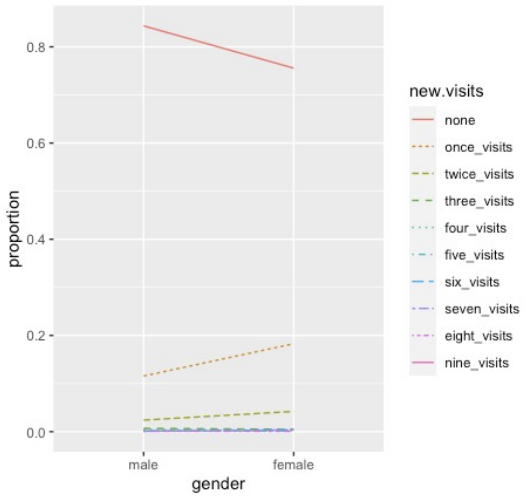
Residual vs. Fitted plot do not show a clear non-linear pattern, indicating that the model form might be appropriate; the fan shape and outliers suggest that the model's assumptions (particularly constant variance and no outliers) may not be fully met. This could affect the reliability of the model's predictions, especially at the higher end of the predicted values.

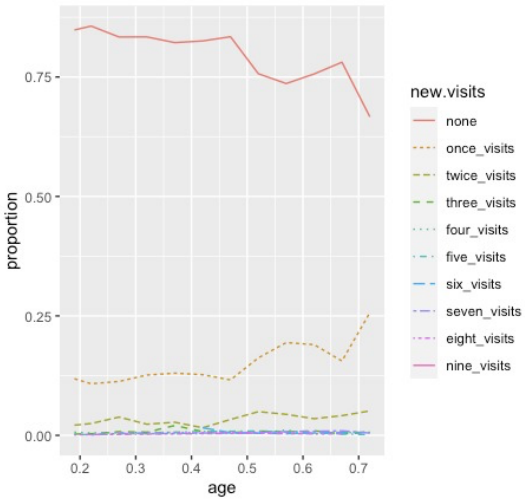
The Q-Q Residuals plot shows that the model violates the linearity assumption, as the data points can be seen heavily deviating from the linear line as the theoretical quantiles increase.

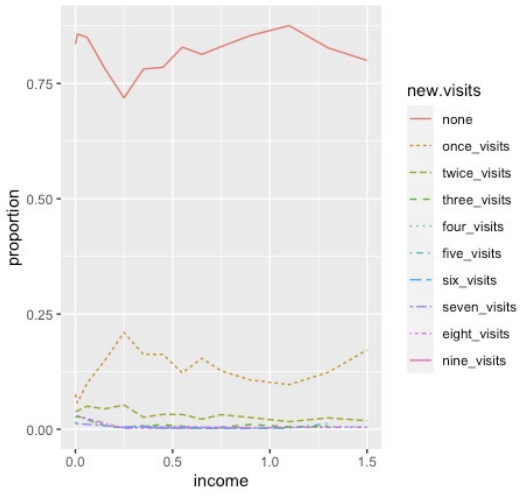
The Residuals vs Leverage plot shows that the model does not violate the assumption that the dataset suffers from influential points, as no point can be seen beyond the dotted barrier and no visible curvature can be seen on the red line.

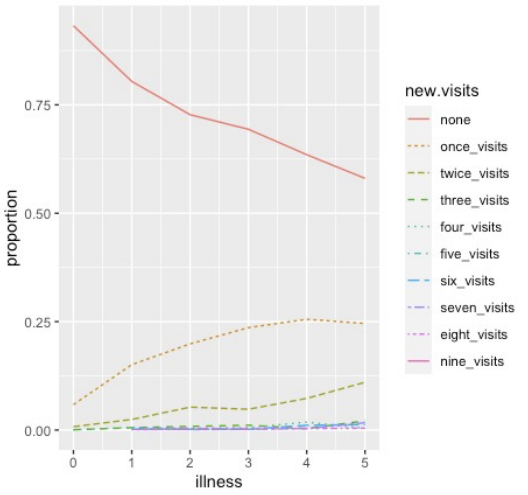
C) The pro of this analysis is that using Poisson to model the response variable “visit” is a good alternative compared to the commonly used gaussian/normally distributed multiple linear regression. This is because Poisson assumed that the response variable is positive and discrete, which is fitting for count data such as the number of visits to the hospital. However, as the regression output and residual analysis graph showed, there are critical problems such as the failure of the linearity assumption as evident from the Q-Q Residual plot in Graph 1 or the alarmingly high Chi-squared p value of 1 calculated from the full model and reduced model. Perhaps the reason for this con can stem from the restrictive assumption of mean equal to variance that the Poisson regression assumed; an alternative to relaxing this strict restriction can be to assume the quasi-Poisson or negative binominal distribution model instead. Besides that, the coefficient of predictors can be easily understood in terms of probability, therefore making the interpretation relevant in the context of estimating the number of visits to the doctor by an individual. Since the generalised linear model is a model that can regress multiple variables at once, it provides a fair evaluation of the impact of each predictor while taking into consideration the impact of other predictors at the same time. An inherent con in this analysis is that the dataset was taken from 1977-1978, which might not be useful if the purpose is to make a prediction on the number of visits to the doctor in current time due to healthcare systems, demographics, and societal factors that may have changed since.

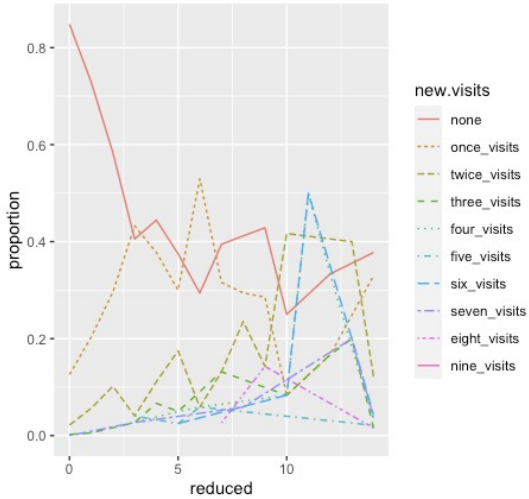
Question 2 Appendix

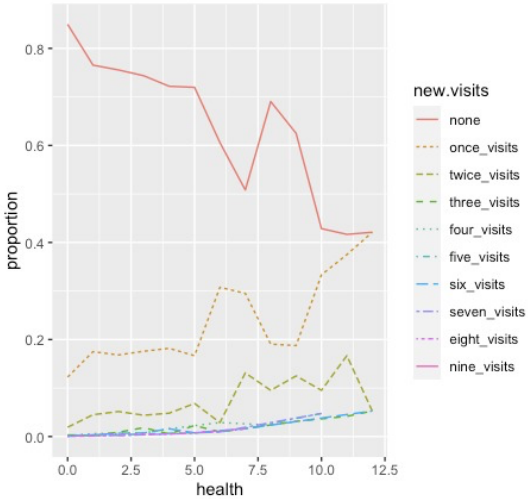


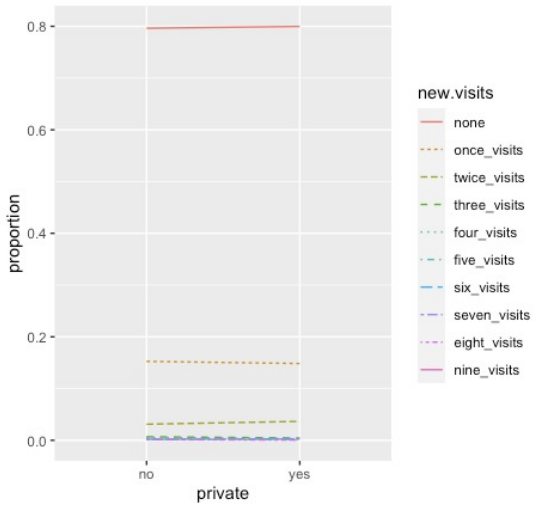


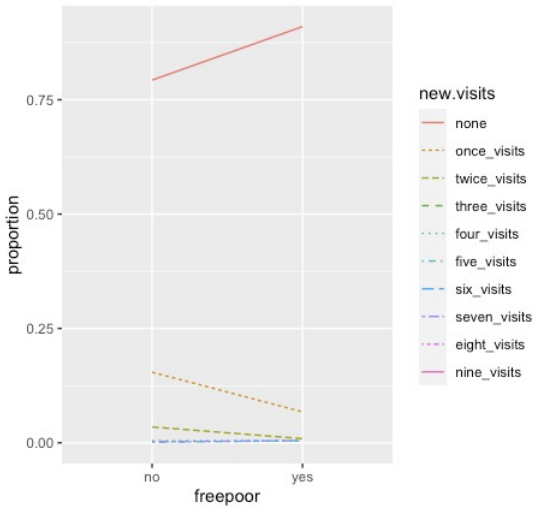


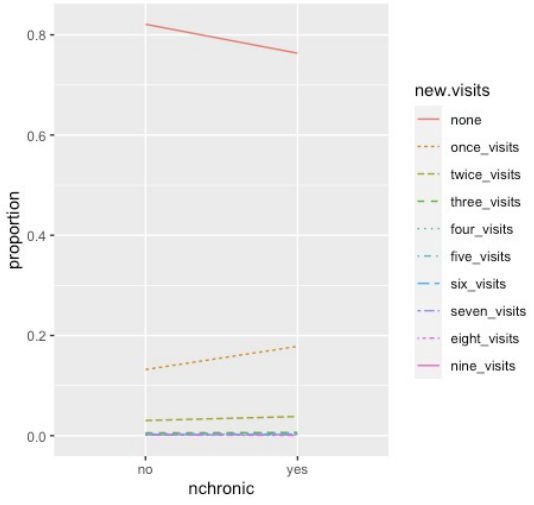


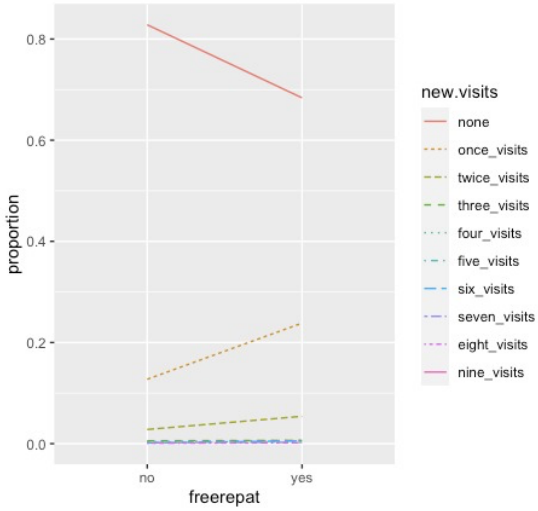


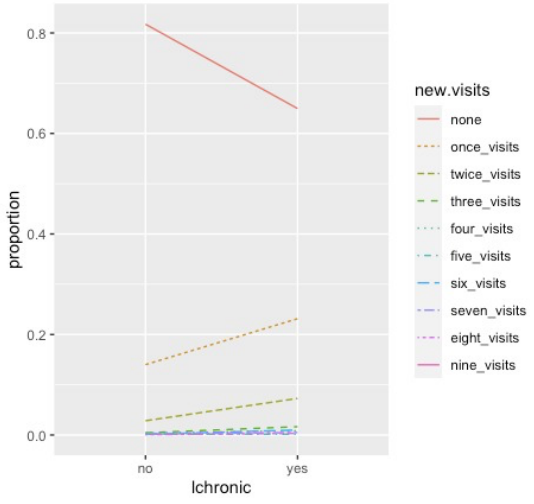



proportion
0.8
0.6
0.4
0.2
0.0
no
yes
private
new.visits
none
once_visits
twice_visits
three_visits
four_visits
five_visits
six_visits
seven_visits
eight_visits
nine_visits


proportion
0.75
0.50
0.25
0.00
no
yes
freepoor
new.visits
none
once_visits
twice_visits
three_visits
four_visits
five_visits
six_visits
seven_visits
eight_visits
nine_visits


proportion
0.8
0.6
0.4
0.2
0.0
no
yes
nchronic
new.visits
none
once_visits
twice_visits
three_visits
four_visits
five_visits
six_visits
seven_visits
eight_visits
nine_visits


proportion
0.8
0.6
0.4
0.2
0.0
no
yes
freerepat
new.visits
none
once_visits
twice_visits
three_visits
four_visits
five_visits
six_visits
seven_visits
eight_visits
nine_visits


proportion
0.8
0.6
0.4
0.2
0.0
no
yes
lchronic
new.visits
none
once_visits
twice_visits
three_visits
four_visits
five_visits
six_visits
seven_visits
eight_visits
nine_visits