JUSTIFICATION OF THE CHOSEN REGRESSION MODEL

The aim of this analysis is to construct a statistical model that explores the effect that several variables may have on the number of doctor visits in past 2 weeks. According to our analysis, the subsequent model presents the estimated association between our response variable and the explanatory variables: $\text{visits}_i \sim \text{Poisson}(\mu)$ where $\log(\mu_i) = -2.004 - 0.200 \times \text{gendermale} + 0.5168 \times \text{age} + 0.1988 \times \text{illness} + 0.1277 \times \text{reduced} + 0.0334 \times \text{health} - 0.4375 \times \text{freepooryes}$ We chose a Poisson GLM with a canonical log link function due to the discrete count type of data of our response variable, namely 'visits'. Consequently, this model assumes that the conditional mean of 'visits', given the set of predictors listed above follows a Poisson distribution, with its variance equal to the mean. The choice was further supported by a goodness-of-fit test p-value of 1, which is based on the chi-square distribution. This result suggests that the model fits the data well, leading us to fail to reject the null hypothesis that there is no significant difference between the saturated model and our current model. However, this high p-value warrants careful interpretation, as it may overlook subtle discrepancies. The limitations of this model will be further discussed in section C2.

SUMMARY OF THE RESULTS AND INTERPRETATION OF THE EMPRICAL FINDINGS

The first step of our analysis involved examining the correlation matrix among all the predictor variables. The purpose of this step was to identify any pair of variables that might be highly correlated, indicating potential multicollinearity. According to our results, only age and freerepat exhibited a moderate correlation, with a correlation coefficient of approximately 0.605. Besides these, the correlation coefficients between all other variables fell below the commonly accepted threshold of 0.8, which would indicate potential issues of multicollinearity. The second step involved fitting the Poisson GLM. We fitted a model with six predictors, namely: gender male, age, illness, reduced, health and freepoor yes. The third step was to calculate Variance Inflation Factors (VIF) to have a more comprehensive understanding of multicollinearity. As shown in Table 1, all VIF values corresponding to the predictors in our analysis were below the typical thresholds of 5 or 10, suggesting that multicollinearity is not a concern in our model.

gender	age	illness	reduced	health	freepoor
1.078	1.203	1.297	1.329	1.367	1.025

Table 1: VIF values

Subsequently, we performed a goodness-of-fit test. As previously mentioned, the p-value was

1, indicating a good fit for the model. This led us to not reject the null hypothesis, suggesting no significant difference between the saturated model and our model. Next, we analysed the diagnostic plots of the residuals (Figure 1) in which under a correctly specified Poisson GLM the Pearson residuals should asymptotically follow a normal distribution with a constant variance. However, as depicted in Figure 1, the normal assumption and constant variance of the deviance residuals do not seem to hold. In fact, in the Residuals vs Fitted plot the Pearson residuals should be randomly dispersed around the 0 and 1, with no discernible pattern. In our plot, it looks like the Pearson residuals are following a pattern, as they don't appear to be completely random, indicating that the model may not be capturing some of the non-linear relationships. Additionally, the Q-Q plot reveals some deviation at the end tail from the line which indicate a departure from the expected normal distribution. Overall, these problems indicate that the Poisson GLM does not fit the data well, which might be due to overdispersion. Additionally, the Residuals vs Leverage plot shows that there are a few points with high leverage, but none of them appear to have a high Cook's distance that would indicate they are overly influential.

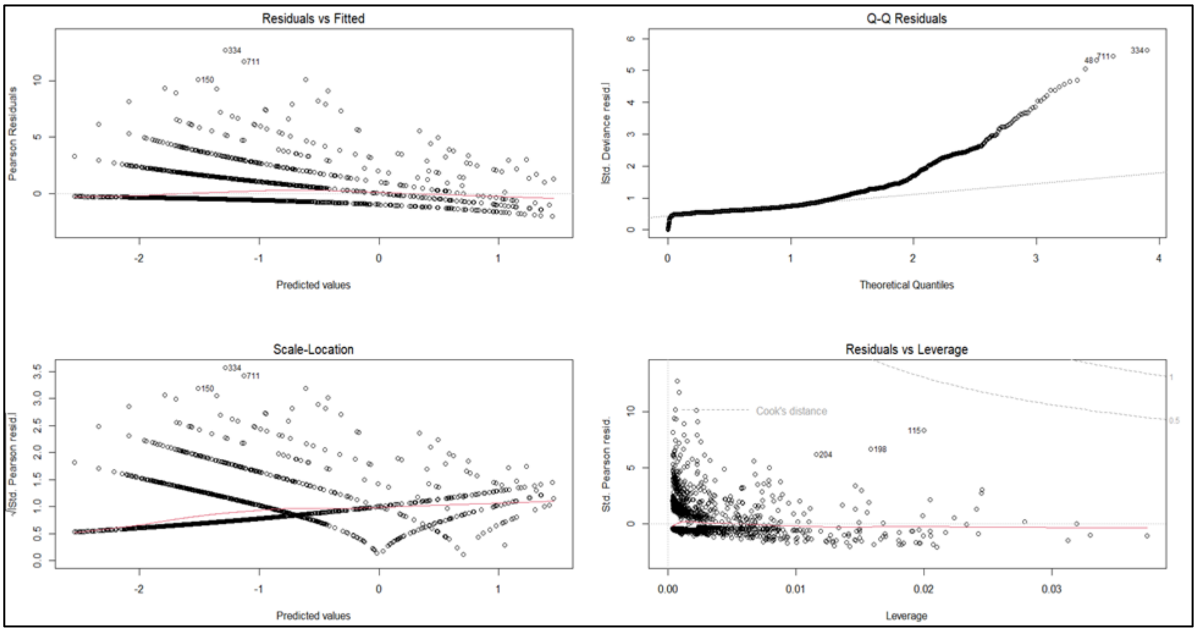

Figure 1: Diagnostic plots panel

We proceed with the interpretation of the parameters although this should be done with caution given the lack of fit. To interpret the result of the change in terms of percentage our model, we need to apply the following formula: $(e^{\beta i} - 1) * 100$. This formula gives you the percent change directly, where βi is the coefficient for the i-th predictor variable. When we input the estimated coefficient into this equation, $e^{\beta i}$ gives the rate ratio, and subtracting 1 and multiplying by 100 converts this ratio into a percentage change. Therefore, being male reduces the expected number of doctor visits by 18%. Essentially, men are expected to visit the doctor 18% less often than women in the past two weeks. Each additional year of age is associated with an increase of 68% in the expected number of doctor visits. In regard to reduced, as days of reduced activity increase, so does the expected number of doctor visits by 14%. The same logic applies to the interpretation of the other predictors.

	Estimate	**Std. Error**	**Pr(>|t|)**
(Intercept)	-2.004	0.076	< 2e-16
gendermale	-0.201	0.054	0.00002
age	0.517	0.132	8.88E-05
illness	0.199	0.018	< 2e-16
reduced	0.128	0.005	< 2e-16
health	0.033	0.01	0.0008
freepooryes	-0.438	0.173	0.0115

Table 2: Regression model summary output

(2C) DISCUSSION OF THE PROS AND CONS OF YOUR ANALYSIS

As mentioned earlier, our analysis cannot be reliable due to lack of fit of the residuals. This is despite we implemented several selection variables techniques with the aim to improve the model. Firstly, we employed Lasso regression for variable selection which identified the variable 'nchronic' as non-significant with a coefficient of 0. We fitted another Poisson GLM but this did not imply any signs of improvements. We also performed stepwise variable selection using stepwise regression, which led us to exclude the variables nchronic, freerepat and lchronic. Despite these efforts, our analysis revealed persistent lack of fit, mainly due to overdispersion. To mitigate this issue, we attempted to use more flexible distributions that do not impose equality of mean and variance, like Quasipoisson and Negative Binomial. However, this brought minimal impact on the Pearson residuals. We also introduced interaction terms, including 'illness' and 'reduced activity’, to explore potential differential impacts regarding the number of illnesses on the amount of reduced activity. Yet, these additions did not enhance our model. Ultimately, we chose a a simpler and more interpretable model, while ensuring statistical significance of the predictors. However, given the higher-than-expected proportion of zero counts compared to the mean of non-zero counts in the visit’s variable, we suggest considering alternative regression techniques, like Zero-Inflated Poisson Regression for count data with numerous zero counts. This model addresses data from two scenarios: one with consistent zero counts, and another following a Poisson distribution. It incorporates covariates through the Poisson mean, differentiating between always-zero outcomes and Poisson-distributed counts.