An Analysis of the Factors Influencing Monthly Doctor Visits in Germany

Giulio Caputi

University of Oxford, MSc in Statistical Science, 11 December 2024

Contents

1 Introduction

In this study, we investigate factors influencing monthly doctor visits in Germany using demographic and socioeconomic variables. We begin with **exploratory data analysis** and **feature engineering** to extract as much meaning as possible from our data. Subsequently, multiple **Poisson GLMs** are developed, starting with a full model, refined through both manual and automatic variable selection processes, with **AIC minimisation** guiding the final model choice. Model evaluation techniques, including **goodness-of-fit tests**, **residual analyses**, and **out-of-sample error estimations**, are employed to assess performance and validate assumptions. Additionally, we present a throughout **interpretation** of results. We conclude by adjusting previous variance estimates to account for **overdispersion** and ensure more accurate inference.

2 Exploratory Data Analysis

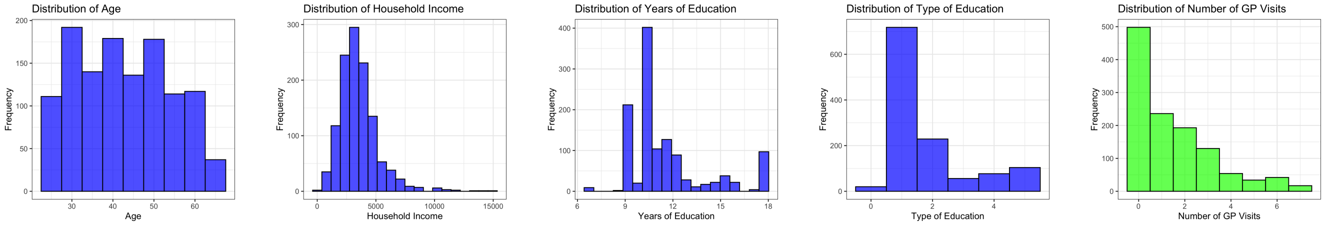
2.1 Univariate Analysis

Variable	**Percentage (%)**
Women	47.59
With kids under 16	39.37
Married	74.5
Employed	70.27
With private insurance	12.38
With additional insurance	2.492

Table 1: Distributions of binary variables

Variable	**Min**	**1st Qu**	**Median**	**Mean**	**3rd Qu**	**Max**	**St dev**
Age	25.0	32.0	42.0	42.6	51.0	64.0	11.21
Household income	50	2400	3300	3531	4200	15000	1671
Years of education	7.0	10.5	10.5	11.5	12.0	18.0	2.507
Type of education	0.0	1.0	1.0	1.8	2.0	5.0	1.306
Number of GP visits	0.0	0.0	1.0	1.47	2.0	7.0	1.739

Table 2: Summary statistics of non-binary variables



(a) *age* (b) *hhninc**1000 (c) *educyrs* (d) *eductype* (e) *docvis*

Figure 1: Distributions of non-binary variables

We first ensure no missing or duplicate values are present, then compute the **frequency of binary variable categories** (Table 1), and analyse **histograms** (Figure 1) **and summary statistics** (Table 2) **for other variables**. The distribution of *docvis* is skewed (Figure 1, panel (e)), indicating most individuals did not visit the GP in the month preceding the interview. The dataset appears **representative of the overall population**, as age and gender are quite balanced, and the distributions of many variables align with larger German samples (e.g., income is skewed (Clementi and Gallegati, 2005), approximately 70% of people between 25 and 64 years old are married (Statistisches Bundesamt [Destatis], 2021), the percentage of employed people in the same age range is more than 70% (Statistisches Bundesamt [Destatis], 2022), and slightly more than 10% of individuals have private medical insurance (Simple Germany, 2024)). However, the distribution of *eductype* differs from population-wide data, showing significantly fewer individuals with university degrees (Statistisches Bundesamt [Destatis], 2020).

2.2 Bivariate Analysis

To evaluate linear relationships and potential **multicollinearity**, we plot a **correlation matrix** of our variables (Figure 2). The correlation between *docvis* and other variables is weak, ranging from -0.12 to 0.16. The strongest positive correlations are with *female* (0.16) and *age* (0.12), suggesting **being female or older may slightly increase doctor visits**. Additionally, given the 0.94 correlation between *eductype* and *educyrs*, we remove the latter to avoid multicollinearity. We choose this variable both because spending a certain number of years in school does not imply earning a certain degree (while the opposite is true), and because *educyrs* contains some weird values (10.80549, 11.44142, 11.81824), likely to be entry errors. Lastly, *female* and *employed* have a strong negative correlation of -0.37, and indeed 53% of women in our data are employed, versus 86% of men.

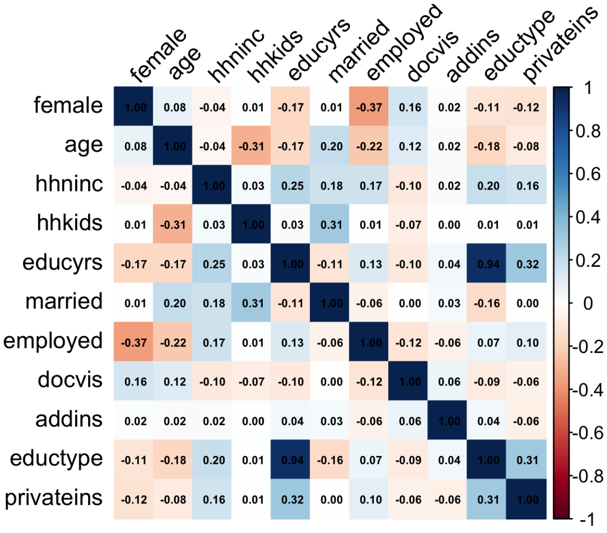


Figure 2: Correlation matrix

To analyse the relationship between explanatory variables and *docvis*, we plot histograms of the latter based on values of the formers (Figure 3), creating **grouping variables** for *age* and *hhninc*. A notable pattern is that the median number of GP visits is always 1, except for individuals with monthly household earnings exceeding 10,000 German Marks, who have a median *docvis* of 0, likely due to better access to preventive healthcare. Overall, **lower-income individuals and women visit doctors more often**, which is consistent with prior studies (e.g., (Hoebel et al., 2016) and (Ladwig et al., 2000)). Older people, as expected, have a higher third quartile of *docvis* compared to younger ones, while individuals with children under 16 visit less, possibly due to **limited time** or unobserved factors. Similar reasons likely cause **employed individuals to visit doctors less than the unemployed** (Kraut et al., 2002). Lastly, individuals with additional insurance show a significantly higher 3rd quartile of GP visits, reflecting proactive healthcare behaviour, though this is based on only 2.5% of the sample.

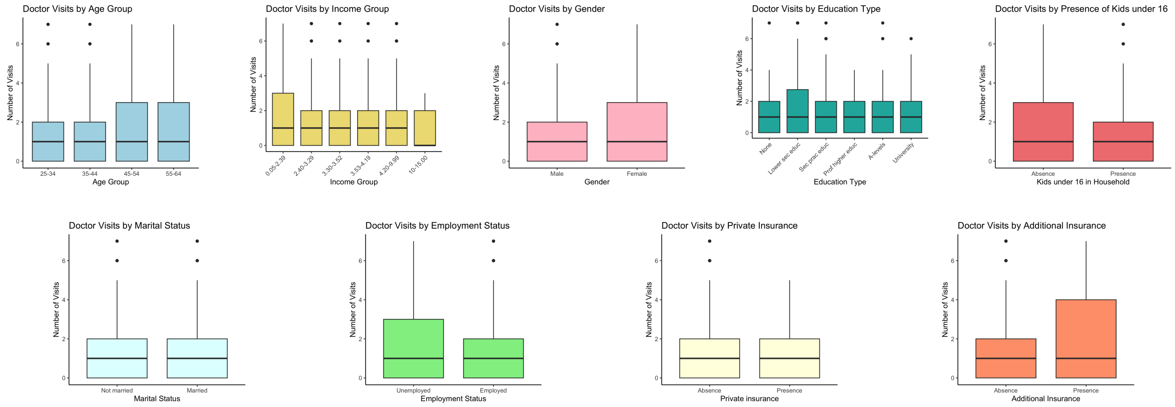


Figure 3: Histograms of number of GP visits grouped by the values of each predictor

2.3 Feature Engineering

In an attempt to **extract more meaning from available variables**, we create:

- *hhninc_per_capita* → We estimate the number of people in household *i* by 1 if individual *i* is not married and does not have kids under 16, by 2 if only one of the previous two conditions is verified, and by 3 otherwise.

For each individual, we then divide *hhninc* by the estimate size of the household to get *hhninc_per_capita*. Subsequently, we eliminate *hhninc*.

- *single_parent* → This is a binary variable taking the value of 1 for unmarried individuals living with children under 16.

3 Modelling

3.1 First Model: Full Poisson GLM

Assuming $docvis_i \sim Pois(\lambda_i)$, with $\mathbb{E}[docvis_i] = Var(docvis_i) = \lambda_i$, we start by fitting a Poisson GLM with canonical link function using all available variables, except for *hhninc* and *educyrs*, as discussed above. We also include the newly created *hhninc_per_capita* and *single_parent* and all possible interactions with $female$. For improved granularity, we use *age* and *hhninc_per_capita* instead of corresponding grouping variables. Table 3 summarises this full model, and highlights **low interpretability and significance of coefficients**. Therefore, **we need to eliminate some variables**, and we do so by performing both a manual and an automatic selection.

	Estimate	**Std. Error**	**z value**	**P(> |z|)**	**Sig. code**
(Intercept)	0.25510	0.21966	1.16	0.2455	
female	0.71201	0.31625	2.25	0.0244	*
age	0.01155	0.00373	3.09	0.0020	**
hhninc_per_capita	-0.10107	0.03860	-2.62	0.0088	**
hhkids	-0.15955	0.09648	-1.65	0.0982	.
married	-0.16382	0.11350	-1.44	0.1489	
employed	-0.14945	0.10357	-1.44	0.1490	
addins	0.30755	0.21002	1.46	0.1431	
eductype	-0.03440	0.02977	-1.16	0.2479	
privateins	0.03087	0.10963	0.28	0.7783	
single_parent	-0.43192	0.32558	-1.33	0.1846	
female:age	-0.00999	0.00510	-1.96	0.0498	*
female:hhninc_per_capita	0.01088	0.05554	0.20	0.8447	
female:hhkids	-0.06490	0.13253	-0.49	0.6243	
female:married	0.05526	0.14535	0.38	0.7038	
female:employed	0.07972	0.12543	0.64	0.5250	
female:addins	0.12817	0.26528	0.48	0.6290	
female:eductype	-0.04429	0.04367	-1.01	0.3105	
female:privateins	-0.10415	0.16975	-0.61	0.5395	
female:single_parent	0.51503	0.37039	1.39	0.1644	
Null deviance:	2543.0 on 1203 degrees of freedom				
Residual deviance:	2405.3 on 1184 degrees of freedom				
AIC:	4324				
log-likelihood:	-2142				

Table 3: Summary of full model

3.2 Second Model: Manual Variable Selection

During this phase, we adopt a **lenient strategy** since further selection will follow. Specifically, we eliminate any predictor X_j such that (1) $\hat{\beta_j}$ is not different from 0 at the 0.1 level, and (2) X_j does not significantly reduce deviance when added to the intercept-only model. For this second check, we fit a null model and compare its deviance to that of a model containing only X_j, using a χ^2_1 test. As Table 3 shows, the variables meeting the first criterion are *married*, *employed*, *addins*, *eductype*, *privateins*, and *single_parent*. Table 4, **unlike a typical analysis-of-deviance table**, shows the impact of adding each predictor to the null model (not sequentially). From this, we see that *married* and *single_parent* do not significantly reduce deviance even when added to the null model, so we exclude them from our analysis.

Variable	Df	Deviance	Resid Df	Resid Dev	Sig code
NULL	–	–	1203	2543	
female	1	61.8	1202	2481	***
age	1	33.4	1202	2510	***
hhninc_per_capita	1	11.8	1202	2531	***
hhkids	1	13.7	1202	2529	***
married	1	0.01	1202	2543	
employed	1	33.9	1202	2509	***
addins	1	8.9	1202	2534	**
eductype	1	20.7	1202	2522	***
privateins	1	8.3	1202	2535	**
single_parent	1	0.03	1202	2543	

Table 4: Analysis of deviance between null model and one-variable models (NOT typical analysis-of-deviance table)

3.3 Third Model: Automatic Variable Selection with AIC

With $p-1=15$ variables (8 single-variable terms and 7 interaction terms), we implement a **computationally intensive automatic variable selection**. Specifically, we fit a model for every possible subset of variables (temporarily ignoring the hierarchical principle) and compute their Akaike Information Criterion (AIC), selecting the model with the lowest value, thus **balancing complexity and predictive power**. This involves evaluating $2^{p-1}-1=32,767$ models. Unlike classic stepwise AIC minimisation, this exhaustive search **does not risk missing the global minimum**. Table 5 presents the two models with minimal AIC, and the *delta AIC* row displays the difference between the AIC of the model and the lowest AIC. Between models 1 and 2, we choose the latter for having one fewer predictor but an essentially equivalent AIC. All interaction terms except $age:female$ are excluded, preserving the hierarchical principle.

Parameter	Model 1	Model 2
(Intercept)	0.1824805	0.0895933
addins	0.3765	0.3915
age	0.009513	0.010113
eductype	-0.05156	-0.05013
employed	-0.09304	NA
female	0.6649	0.6684
hhkids	-0.2290	-0.2277
hhninc_per_capita	-0.07918	-0.08718
privateins	NA	NA
addins:female	NA	NA
age:female	-0.008151	-0.007552
eductype:female	NA	NA
employed:female	NA	NA
female:hhkids	NA	NA
female:hhninc_per_capita	NA	NA
female:privateins	NA	NA
Overall model information		
df:	9	8
log-likelihood:	-2146	-2148
AIC:	4311	4311
delta AIC:	0.0000	0.6937

Table 5: AIC-minimising models

4 Model Evaluation

4.1 Leverage and Influence

We classify a point as “noteworthy” if it has **both high leverage** ($> \frac{2p}{n}$) **and** high Cook’s distance ($> \frac{8}{n-2p}$), i.e., **high influence**. Our dataset contains 17 noteworthy individuals (1.4% of the total), 16 of whom have purchased additional insurance. Since 30 individuals have $addins = 1$, more than half are classified as noteworthy, and they are the only ones to be, except for 1 other observation. It follows that **our model should consider interactions with** ***addins***. We therefore perform another AIC-minimising variable selection as in Section 3.3 among models with these interactions. As shown in Table 6, adding *addins* : *age* and *addins* : *hhninc_per_capita* reduces the AIC from 4311 to 4309 and increases the log-likelihood from -2148 to -2144. This becomes our final model.

Variable	**Estimates**	**Std. Error**	**z value**	**P(> |z|)**	**Sig. code**	**95% Conf. interval**
(Intercept)	0.087778	0.17688	0.50	0.61972		[-0.25891, 0.43447]
addins	1.41876	0.60833	2.33	0.01969	*	[0.22627, 2.61125]
age	0.010495	0.00333	3.16	0.00160	**	[0.00397, 0.01703]
eductype	-0.04804	0.02077	-2.31	0.02072	*	[-0.08877, -0.00731]
female	0.6433	0.19428	3.31	0.00093	***	[0.26254, 1.02412]
hhkids	-0.2314	0.06023	-3.84	0.00012	***	[-0.34940, -0.11330]
hhninc_per_capita	-0.09422	0.02653	-3.55	0.00038	***	[-0.14622, -0.04222]
addins:age	-0.03192	0.01418	-2.25	0.02438	*	[-0.05973, -0.00411]
addins:hhninc_per_capita	0.1836	0.10015	1.83	0.06675	.	[-0.01288, 0.38012]
age:female	-0.007177	0.00427	-1.68	0.09283	.	[-0.01566, 0.00130]
Overall model information						
df:	10					
log-likelihood:	-2144					
AIC:	4309					
delta AIC:	0.0000					
Residual Deviance:	2410					

Table 6: AIC-minimising model, now considering interactions with *addins*. This is our final model

4.2 Comparison with Full Model

We use a **likelihood ratio test** (LRT) to compare our final model with the full model. Under the null, according to which the additional variables of the full model are irrelevant, the test statistic $\Lambda(y) = D(y)^{(final)} - D(y)^{(full)}$ (the deviance difference) approximates a χ^2 distribution with degrees of freedom equal to the difference in number of parameters between the models. Since the test excludes the saturated log-likelihood, it is independent of n, making the approximation valid. The p-value of 0.8956 suggests **the excluded variables in the final model are likely irrelevant**.

4.3 Deviance Goodness-of-Fit Test

Let $l(\theta^{(s)}; y)$ be the saturated log-likelihood, and $l(\hat{\beta}; y)$ be the model log-likelihood. Under H_0, which assumes the model is well-specified, the deviance $D(y) = 2[l(\theta^{(s)}; y) - l(\hat{\beta}; y)]$ approximates a $\chi^2(n-p)$ distribution. The p-value of 0 (10^{-84}) raises doubts about model validity, suggesting issues like **missing covariates or overdispersion** (i.e., variances exceed means, violating Poisson assumptions), which is expanded on in Section 6. It is worth mentioning that this test risks not being particularly reliable as, contrary to the case above, **the dimension of the alternative parameter space is** n. This, together with the fact that observed counts are very small (never greater than 7), makes the MLEs not converge to their limiting distributions.

4.4 Deviance-Based R-squared

The deviance-based R-squared $R^2_D = 1 - \frac{D(y)}{D(y)^{(null)}}$ (Cameron and Windmeijer, 1997) is only 5.2%, further indicating that **our approach may be misspecified**, with important explanatory variables missing and/or a wrong type of model.

4.5 Analysis of Deviance Residuals

Deviance residuals, $r_{D_i} = \text{sign}(y_i - \hat{y}_i)\sqrt{d_i}$, are standardised as $r'_{D_i} = \frac{r_{D_i}}{\sqrt{1-h_{ii}}}$, where the h_{ii}'s are leverage components. We have two positive signs: the r'_{D_i}'s are roughly **symmetric around 0**, and **their correlation with fitted values $\hat{y}$ is only 0.0034** (Table 7). As Figure 4 shows, residuals stratify by *docvis*, something common for Poisson data and which confirms **the model performs well for low** (and thus common) **values**.

Statistic	Value
Min	-5.006
1st Qu	-1.569
Median	-0.454
Mean	-0.310
3rd Qu	0.708
Max	3.797
Standard dev	1.392
Cor with fitted values	0.003439

Table 7: Summary statistics of standardised deviance residuals

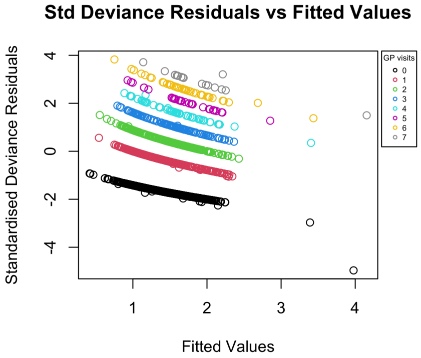


Figure 4: Plot of standardised deviance residuals against fitted values

Figure 5 shows a Standard Gaussian Q-Q plot of r'_D, which approximates a $N(0,1)$ distribution for most middle quantiles. However, deviations at the tails suggest overdispersion, where *docvis* variability exceeds Poisson assumptions. Additionally, 9.8% of the r'_{D_i}'s lie outside the $[-2,2]$ range, further indicating that **the Poisson GLM may not be ideal**.

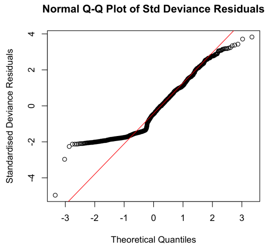


Figure 5: $N(0,1)$ Q-Q plot of standardised deviance residuals, with reference line

4.6 Estimation of Out-of-Sample Error

To estimate the out-of-sample error, we split the data 80-20 into train and test sets, fitting a Poisson GLM on the train set. The absolute error on the test set (Table 8, Figure 6) has a median of 1.163, with 75% of errors below 1.641, indicating that **the model performs well in most cases**. However, a standard deviation of 0.9255 and some large errors suggest **model limitations**.

Statistic	Value
Min	0.018
1st Qu	0.670
Median	1.163
Mean	1.282
3rd Qu	1.641
Max	5.585
St dev	0.9255

Table 8: Summary statistics of out-of-sample absolute errors

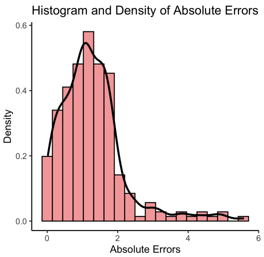


Figure 6: Histogram and density plot of out-of-sample absolute errors

5 Model Interpretation

We have modelled the mean of each $Y_i \sim Pois(\lambda_i)$ with $\mu_i = \lambda_i = g^{-1}(\eta_i) = e^{\eta_i}$, where $\eta_i = x_i^T\beta$ is the linear predictor, and $\eta_i = g(\mu_i) = log(\mu_i)$ is the canonical link function. Our final model, a summary of which is shown in Table 6 in Section 4.1, is the following:

$$docvis \sim female + age + hhninc\_per\_capita + hhkids + addins + eductype + female : age + age : addins + hhninc\_per\_capita : addins$$

In our GLM, each estimate $\hat{\beta}_j$ represents the **expected log change of** *docvis* **for a one-unit increase in predictor** j, with other variables at reference level. So, $e^{\hat{\beta}_j}$ is the **multiplicative factor** by which the expected number of doctor visits changes for a one-unit increase in predictor j. Now on to the actual interpretation.

- Intercept → When all variables are at 0, the expected *docvis* is $e^{\hat{\beta}_0} = e^{0.08778} \approx 1.092$ (95% CI: $[0.7719, 1.544]$). However, this interpretation involves a strong extrapolation, since we never observe values close to 0 for *age* and *hhninc_per_capita*. Indeed the estimate is not statistically significant, and its 95% confidence interval includes 1.

- $female$ → Suppose $age = a$. Our model predicts the factor by which the expected *docvis* changes for a female compared to a male to be $e^{\hat{\beta}_{female} + a*\hat{\beta}_{age:female}} = e^{0.64333-0.00718*a}$ (95% CI: $[e^{0.26254-0.01566*a}, e^{1.02412+0.0013*a}]$). For instance, the average 30-year-old woman is expected to have a *docvis* value $e^{0.64333-0.00718*30} \approx 1.53$ times that of the average 30-year-old man. The fact that $\hat{\beta}_{age:female}$ is negative means that the age effect on *docvis* is smaller for females than for males.

- *age* → For men with no additional insurance ($female = addins = 0$) we expect every one-year increase in age to multiply their month GP visits by $e^{\hat{\beta}_{age}} = e^{0.01050} \approx 1.011$ (95% CI: $[1.004, 1.017]$). Instead, for women with no additional insurance ($female = 1$ and $addins = 0$), our model predicts a one-year age increase to multiply *docvis* by $e^{\hat{\beta}_{age} + \hat{\beta}_{age:female}} = e^{0.01050-0.00718} \approx 1.003$ (95% CI: $[0.988, 1.019]$). For men with additional insurance ($female = 0$ and $addins = 1$), we expect every one-year age increase to multiply *docvis* by $e^{\hat{\beta}_{age} + \hat{\beta}_{addins:age}} = e^{0.01050-0.03192} \approx 0.9788$ (95% CI: $[0.946, 1.013]$). Lastly, women with additional insurance ($female = addins = 1$) are expected to multiply their monthly GP visits by $e^{\hat{\beta}_{age} + \hat{\beta}_{age:female} + \hat{\beta}_{addins:age}} = e^{0.01050-0.00718-0.03192} \approx 0.972$ (CI: $[0.931, 1.014]$) every year. It is worth remembering that, since only 2.5% of our observations have additional insurance, we have to be cautious when making claims about this category. Indeed, in the last 3 cases the 95% confidence intervals contain 1, implying uncertainty about the direction of these effects.

- *hhninc_per_capita* → As the net monthly household income per capita increases by 1000 German Marks, we expect *docvis* to decrease by 9% for individuals with no additional insurance (as $e^{\hat{\beta}_{hhninc\_per\_capita}} = e^{-0.09422} \approx 0.91$ (CI: $[0.864, 0.959]$)), and to increase by 9.4% for individuals with additional medical insurance (as $e^{\hat{\beta}_{hhninc\_per\_capita} + \hat{\beta}_{addins:hhninc\_per\_capita}} = e^{-0.09422+0.18362} \approx 1.094$ (CI: $[0.853, 1.402]$)). This interval contains 1, implying uncertainty about the true direction of this effect.

- *hhkids* → *Ceteris paribus*, individuals living with children under 16 have an expected *docvis* value $e^{\hat{\beta}_{hhkids}} = e^{-0.23135} \approx 0.793$ (CI: $[0.705, 0.893]$) times that of people with $hhkids = 0$. In our dataset, the average docvis value for people with $hhkids = 1$ is 0.8331 the average *docvis* value for individuals with $hhkids = 0$, well inside the model's confidence interval.

- *addins* → Our model predicts individuals who purchased additional health insurance to have a *docvis* value $e^{\hat{\beta}_{addins} + a*\hat{\beta}_{addins:age} + i*\hat{\beta}_{addins:hhninc\_per\_capita}} = e^{1.41876-0.03192*a+0.18362*i}$ (95% CI: $[e^{0.22627-a*0.05973-i*0.01288}, e^{2.61125-a*0.00411+i*0.38012}]$) times that of people without additional insurance, for $age = a$, $hhninc\_per\_capita = i$, and other variables at their reference levels. Since $\hat{\beta}_{addins:age}$ is negative, our model predicts the effects of *age* to be attenuated for people with additional insurance. Conversely, given that $\hat{\beta}_{addins:hhninc\_per\_capita}$ is positive, as income rises, we expect the presence of additional insurance to further increase the number of GP visits.

- *eductype* → For an additional level of education, we expect the number of *docvis* to decrease by roughly 4.7%, as $e^{\hat{\beta}_{eductype}} = e^{-0.04804} \approx 0.953$ (95% CI: $[0.915, 0.993]$).

6 Estimating the Dispersion Parameter

Throughout this work, we have assumed the dispersion parameter ϕ to be 1. Calculating $\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$, with $V(\hat{\mu}_i) = \hat{\mu}_i$, gives $\hat{\phi} = 2.073$, indicating **overdispersion** (variances exceed means). This violates a Poisson assumption, resulting in **too conservative variance estimates** $Var(\hat{\beta}_j)$'s, making our **confidence intervals too narrow**, and our **p-values for the significance of the** $\hat{\beta}_j$ **too low**. To correct this, we scale variances by $\hat{\phi}$. As shown in Table 9, standard errors and p-values increase and z-values decrease, reducing the significance of many coefficients, with some losing significance entirely. With these updated and more realistic assessments of uncertainty in the $\hat{\beta}_j$'s, alternative approaches like quasi-Poisson or Negative Binomial models would be the ideal next steps, but trying them is outside the scope of this work.

Variable	Estimate	Std. Error	z value	P(> |z|)	Sig. code	95% Conf. interval
(Intercept)	0.08778	0.25467	0.34	0.7339		[-0.41137, 0.58693]
female	0.64333	0.27972	2.30	0.0215	*	[0.09507, 1.19159]
age	0.01050	0.00480	2.19	0.0285	*	[0.00110, 0.01990]
hhinc_per_capita	-0.09422	0.03820	-2.47	0.0135	*	[-0.16909, -0.01935]
hhkids	-0.23135	0.08672	-2.67	0.0076	**	[-0.40132, -0.06138]
addins	1.41876	0.87587	1.62	0.1052		[-0.2979, 3.1355]
eductype	-0.04804	0.02991	-1.61	0.1074		[-0.10665, 0.01057]
female:age	-0.00718	0.00614	-1.17	0.2417		[-0.01922, 0.00486]
age:addins	-0.03192	0.02040	-1.56	0.1184		[-0.07190, 0.00806]
hhinc_per_capita:addins	0.18362	0.14411	1.27	0.2042		[-0.09884, 0.46608]

Table 9: Recomputed results for our final model, with dispersion parameter of 2.073

7 Conclusion

Our analysis identified significant factors influencing monthly doctor visits, including gender, age, household income per capita, and additional insurance. While the final Poisson GLM provided insights, its **low deviance-based R-squared** and **evidence of overdispersion** show limitations in explanatory power. Adjusted variance estimates reduced some coefficients' significance. Future work could deal with overdispersion using **quasi-Poisson or Negative Binomial models**. Additionally, the fact that the distribution of some variables (e.g., type of education) does not resemble what we would observe in the overall population, coupled with the small sample size for certain subgroups (e.g., those with additional insurance), both highlight the **need for caution in generalising results**. Lastly, we are probably **missing some important covariates**, related, for instance, to health status, lifestyle, environment, and past medical history. Despite these limitations, our study offers valuable insights into healthcare utilisation in Germany.

8 References

Cameron, A. and Windmeijer, F. (1997). "An R-squared Measure of Goodness of Fit for some Common Nonlinear Regression Models". *Journal of econometrics*, vo. 77(2), pp. 329-342

Clementi, F. and Gallegati, M. (2005). "Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States". *Econophysics of wealth distributions: Econophys-Kolkata I*, pp. 3-14

Hoebel, J., Rattay, P., Prütz, F., Rommel, A. and Lampert, T. (2016). "Socioeconomic Status and Use of Outpatient Medical Care: The Case of Germany". *PloS one*, vo. 11(5)

Kraut, A., Mustard, C., Walld, R. and Tate, R. (2002). "Unemployment and Health Care Utilization". *Health Effects of the New Labour Market*, pp. 25-42

Ladwig, K., Marten-Mittag, B., Formanek, B. and Dammann, G. (2000). "Gender Differences of Symptom Reporting and Medical Health Care Utilization in the German Population". *European Journal of Epidemiology*, vo. 16, pp. 511-518

Simple Germany (2024). “Private vs Public Health Insurance in Germany”. Available at https://www.simplegermany.com/private-vs-public-health-insurance-germany/

Statistisches Bundesamt [Destatis] (2020). “Educational Attainment of the Population in Germany”. Available at https://www.destatis.de/EN/Themes/Society-Environment/Education-Research-Culture/Educational-Level/Tables/educational-attainment-population-germany.html

Statistisches Bundesamt [Destatis] (2021). “Population by Marital Status”. Available at https://www.destatis.de/EN/Themes/Society-Environment/Population/_Graphic/_Interactive/population-marital-status-group.html

Statistisches Bundesamt [Destatis] (2022). “Employment from 1991 to 2021”. Available at https://www.destatis.de/EN/Themes/Labour/Labour-Market/Employment/Tables/etq-1991-2021.html

9 R Code

```r
# We import necessary libraries
library(ggplot2)
library(dplyr)
library(corrplot)
library(MuMIn)


# We load the data
dvis = read.csv("/Users/Giulio/Library/Mobile Documents/com~apple~CloudDocs/Oxford/
                Courses/Michaelmas Term/Applied Statistics/Practicals/Practical Week 8/
                dvis.csv")
options(digits = 4)




##############################################
#### SECTION 2: EXPLORATORY DATA ANALYSIS ####
##############################################

### Section 2.1: Univariate Analysis ###
# This is the data in Table 1: Distributions of binary variables
dvis %>%
  summarise(
    female_percentage = mean(female, na.rm = TRUE) * 100,
    individuals_with_kids_under_16_percentage = mean(hhkids, na.rm = TRUE) * 100,
    married_percentage = mean(married, na.rm = TRUE) * 100,
    employed_percentage = mean(employed, na.rm = TRUE) * 100,
    private_insurance_percentage = mean(privateins, na.rm = TRUE) * 100,
    additional_health_insurance_percentage = mean(addins, na.rm = TRUE) * 100
  ) %>% print()


# This is the data in Table 2: Summary statistics of non-binary variables
print(summary(dvis$age))
sd(dvis$age)
print(summary(dvis$hhninc*1000))
sd(dvis$hhninc*1000)
print(summary(dvis$educyrs))
sd(dvis$educyrs)
```

```r
print(summary(dvis$eductype))
sd(dvis$eductype)
print(summary(dvis$docvis))
sd(dvis$docvis)


# This creates Figure 1: Distributions of non-binary variables
# Age, panel (a)
ggplot(dvis, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Age", x = "Age", y = "Frequency") + theme_bw()
# Household Income * 1000, panel (b)
ggplot(dvis, aes(x = hhninc*1000)) +
  geom_histogram(bins = 20, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Household Income", x = "Household Income",
       y = "Frequency") + theme_bw()
# Years of Education, panel (c)
ggplot(dvis, aes(x = educyrs)) +
  geom_histogram(bins = 19, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Years of Education", x = "Years of Education",
       y = "Frequency") + theme_bw()
# Type of Education, panel (d)
ggplot(dvis, aes(x = eductype)) +
  geom_histogram(bins = 6, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Type of Education", x = "Type of Education",
       y = "Frequency") + theme_bw()
# GP visits, panel (e)
ggplot(dvis, aes(x = docvis)) +
  geom_histogram(binwidth = 1, fill = "green", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Number of GP Visits", x = "Number of GP Visits",
       y = "Frequency") + theme_bw()



### Section 2.2: Univariate Analysis ###

# This creates Figure 2: Correlation matrix
cor_matrix = cor(dvis, use = "complete.obs", method = "pearson")
corrplot(cor_matrix, method = "color", addCoef.col = "black",
         number.cex = 0.5, tl.col = "black", tl.srt = 45)


# Here we create the "age_group" variable
dvis$age_group = cut(
  dvis$age,
  breaks = c(25, 34, 44, 54, 64), # Define the age range intervals
  labels = c("25-34", "35-44", "45-54", "55-64"), # Assign labels to the intervals
  include.lowest = TRUE)


# Here we create the "income_group" variable
dvis$income_group = cut(
  dvis$hhninc,
  breaks = c(0.05, 2.399, 3.299, 3.529, 4.199, 9.99, 15.00),
  labels = c("0.05-2.39", "2.40-3.29", "3.30-3.52", "3.53-4.19", "4.20-9.99",
             "10-15.00"), include.lowest = TRUE)


# This creates Figure 3: Histograms of number of GP visits grouped by the values
# of each predictor
# Boxplots of GP visits by age group
```

```r
ggplot(dvis, aes(x = age_group, y = docvis)) +
  geom_boxplot(fill = "lightblue") + labs(
    title = "Doctor Visits by Age Group",
    x = "Age Group",
    y = "Number of Visits") + theme_classic()
# Boxplots of GP visits by income group
ggplot(dvis, aes(x = income_group, y = docvis)) +
  geom_boxplot(fill = "lightgoldenrod") +
  labs(
    title = "Doctor Visits by Income Group",
    x = "Income Group",
    y = "Number of Visits") + theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
# Boxplots of GP visits by gender
ggplot(dvis, aes(x = factor(female), y = docvis)) +
  geom_boxplot(fill = c("pink")) +
  labs(
    title = "Doctor Visits by Gender",
    x = "Gender",
    y = "Number of Visits") + theme_classic() +
  scale_x_discrete(labels = c("Male", "Female"))
# Boxplots of GP visits by education type
ggplot(dvis, aes(x = factor(eductype), y = docvis)) +
  geom_boxplot(fill = c("lightseagreen")) +
  labs(
    title = "Doctor Visits by Education Type",
    x = "Education Type",
    y = "Number of Visits") + theme_classic() +
  scale_x_discrete(labels = c("None", "Lower sec educ", "Sec prac educ",
                              "Prof higher educ", "A-levels", "University"))  +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
# Boxplots of GP visits by presence of kids under 16
ggplot(dvis, aes(x = factor(hhkids), y = docvis)) +
  geom_boxplot(fill = c("lightcoral")) +
  labs(
    title = "Doctor Visits by Presence of Kids under 16",
    x = "Kids under 16 in Household",
    y = "Number of Visits") + theme_classic() +
  scale_x_discrete(labels = c("Absence", "Presence"))
# Boxplots of GP visits by marital status
ggplot(dvis, aes(x = factor(married), y = docvis)) +
  geom_boxplot(fill = c("lightcyan")) +
  labs(
    title = "Doctor Visits by Marital Status",
    x = "Marital Status",
    y = "Number of Visits") + theme_classic() +
  scale_x_discrete(labels = c("Not married", "Married"))
# Boxplots of GP visits by employment status
ggplot(dvis, aes(x = factor(employed), y = docvis)) +
  geom_boxplot(fill = c("lightgreen")) +
  labs(
    title = "Doctor Visits by Employment Status",
    x = "Employment Status",
    y = "Number of Visits") + theme_classic() +
  scale_x_discrete(labels = c("Unemployed", "Employed"))
# Boxplots of GP visits by presence of private insurance
ggplot(dvis, aes(x = factor(privateins), y = docvis)) +
  geom_boxplot(fill = c("lightyellow")) +
  labs(
    title = "Doctor Visits by Private Insurance",
    x = "Private insurance",
    y = "Number of Visits") + theme_classic() +
  scale_x_discrete(labels = c("Absence", "Presence"))
# Boxplot of GP visits by presence of additional insurance
ggplot(dvis, aes(x = factor(addins), y = docvis)) +
  geom_boxplot(fill = c("lightsalmon")) +
  labs(
    title = "Doctor Visits by Additional Insurance",
    x = "Additional Insurance",
    y = "Number of Visits") + theme_classic() +
  scale_x_discrete(labels = c("Absence", "Presence"))



### Section 2.3: Feature Engineering ###
# Here we create the varaibles "hhninc_per_capita" and "single_parent", and we add
# them to our dataset
dvis$hhninc_per_capita=dvis$hhninc/(1+dvis$married+dvis$hhkids)
dvis$single_parent=dvis$single_parent=ifelse(dvis$married==0&dvis$hhkids==1,1,0)




##############################
#### SECTION 3: MODELLING ####
##############################

### Section 3.1: First Model: Full Poisson GLM ###
# Here we fit the full Poisson GLM
# We exclude the columns hhninc, educyrs, age_group, and income_group, as stated
# in the report
full_model = glm(docvis ~ female + age + hhninc_per_capita + hhkids + married +
               employed + addins + eductype + privateins + single_parent +
               female * (age + hhninc_per_capita + hhkids + married + employed +
               addins + eductype + privateins + single_parent),
             family = poisson, data = dvis)


# This is the data used in Table 3: Summary of full model
summary(full_model)



### Section 3.2: Second Model: Manual Variable Selection ###
# We test the difference in deviance between an intercept-only model and a
# one-variable model.
# In turn, we try every predictor as the one in the one-variable model
# This creates the data in Table 4: Analysis of deviance between null model and
# one-variable models (NOT typical analysis-of-deviance table)
variables = c("female", "age", "hhninc_per_capita", "hhkids", "married", "employed",
              "addins", "eductype", "privateins", "single_parent")
for (variable in variables) {
  formula = as.formula(paste("docvis ~", variable))
  model = glm(formula, family = poisson, data = dvis)
  print(anova(model, test = "Chisq"))
  print("----------------------")
}

# This is our second model
second_model = glm(docvis ~ female + age + hhninc_per_capita + hhkids + employed +
                   addins + eductype + privateins + female * (age + hhninc_per_capita +
                   hhkids + employed + addins + eductype + privateins),
                 family = poisson, data = dvis)
summary(second_model)



### Section 3.3: Third Model: Automatic Variable Selection with AIC

# Here we compute the AIC for all possible subsets of variables (2^15 - 1 models)
options(na.action = "na.fail")

# This is used to create Table 5: AIC-minimising models
# (in Table 5, rows and columns are swapped)
all_models = as.data.frame(dredge(second_model, trace = TRUE))
# These are the two models with minimum AICc
all_models[all_models$delta==min(all_models$delta), ]


# We therefore arrive at this model
third_model = glm(docvis ~ female + age + hhninc_per_capita + hhkids + addins +
                  eductype + female*age,
                family = poisson, data = dvis)
summary(third_model)




#####################################
#### SECTION 4: MODEL EVALUATION ####
#####################################

### Section 4.1: Leverage and Influence
# This part of the analysis is necessary to arrive at our final model
# We define the "noteworthy" variable, thus identifying the "noteworthy" observations
leverage = hatvalues(third_model)
cooks_dist = cooks.distance(third_model)
p = length(coef(third_model))
n = nrow(dvis)
leverage_threshold = (2 * p) / n
cooks_dist_threshold = 8 / (n - 2 * p)
dvis$noteworthy=as.integer(leverage>leverage_threshold&cooks_dist>cooks_dist_threshold)
sum(dvis$noteworthy) / n # 17/1204 = 1.412% of observations are noteworthy
16 / sum(dvis$addins) # 53% of people with additional insurance are "noteworthy"


# We perform another global search to minimise AIC, this time considering interactions
# with addins
# The code below is used to get the data in Table 6: AIC-minimising model, now
# considering interactions with addins
global_model = glm(docvis ~ female + age + hhninc_per_capita + hhkids + addins +
                   eductype + female:age + addins * (female + age + hhninc_per_capita +
                   hhkids + eductype),
  family = poisson, data = dvis)

options(na.action = "na.fail")
all_models2 = dredge(global_model, trace = TRUE)

# We arrive at our final model
final_model = glm(docvis ~ female + age + hhninc_per_capita + hhkids + addins +
                  eductype + female:age + addins * (age + hhninc_per_capita),
  family = poisson, data = dvis)
summary(final_model)
```

```r



### Section 4.2: Comparison with Full Model
# Here we perform the LRT of full_model VS final_model
lrt_stat = 2*(logLik(full_model) - logLik(final_model))
lrt_df = length(coef(full_model)) - length(coef(final_model))
# p-value of the test
pchisq(lrt_stat, df = lrt_df, lower.tail = FALSE)
# This is equivalent to:
anova(final_model, full_model, test = "Chisq")



### Section 4.3: Deviance Goodness-of-Fit Test
# p-value of the test
pchisq(final_model$deviance, df = final_model$df.residual, lower.tail = FALSE)



### Section 4.4: Deviance-Based R-squared
1 - final_model$deviance/final_model$null.deviance



### Section 4.5: Analysis of Deviance Residuals

deviance_residuals = residuals(final_model, type = "deviance")
h_values = hatvalues(final_model)
standardised_deviance_residuals = deviance_residuals / sqrt(1 - h_values)
different_colours = as.factor(dvis$docvis)


# This creates Figure 4: Plot of standardised deviance residuals against fitted values
plot(fitted(final_model), standardised_deviance_residuals,
     col = different_colours,
     xlab = "Fitted Values",
     ylab = "Standardised Deviance Residuals",
     main = "Std Deviance Residuals vs Fitted Values")
legend("topright", legend = levels(different_colours),
       col = seq_along(levels(different_colours)),
       pch = 1, cex = 0.5, title = "GP visits", xpd = TRUE, inset = c(-0.13, 0))


# This is the data for Table 7: Summary statistics of standardised deviance residuals
summary(standardised_deviance_residuals)
sd(standardised_deviance_residuals)
cor(standardised_deviance_residuals, fitted(final_model))


# This creates Figure 5: N(0,1) Q-Q plot of standardised deviance residuals,
# with reference line
qqnorm(standardised_deviance_residuals,
       main = "Normal Q-Q Plot of Std Deviance Residuals",
       xlab = "Theoretical Quantiles",
       ylab = "Standardised Deviance Residuals")
qqline(standardised_deviance_residuals, col = "red")

# Almost 10% of the standardised deviance residuals are outside the [-2,2] range
sum(standardised_deviance_residuals > 2 | standardised_deviance_residuals < -2) / n

```

```r


# Section 4.6: Estimation of Out-of-Sample Error
set.seed(1)
# We split the available data between a train and a test set
train_index = sample(seq_len(nrow(dvis)), size = 0.8 * nrow(dvis))
train_set = dvis[train_index, ]
test_set = dvis[-train_index, ]
# We fit a model on the train set
fitted_model = glm(docvis ~ female + age + hhninc_per_capita + hhkids + addins +
                   eductype + female*age + addins * (age + hhninc_per_capita),
    family = poisson, data = train_set)
# We use it to predict docvis for the test set
predictions = predict(fitted_model, newdata = test_set, type = "response")

# This data is needed for Table 8: Summary statistics of out-of-sample absolute errors
absolute_errors = abs(test_set$docvis - predictions)
summary(absolute_errors)
sd(absolute_errors)

# This creates Figure 6: Histogram and density plot of out-of-sample absolute errors
ggplot(data.frame(values = absolute_errors), aes(x = values)) +
  geom_histogram(aes(y = ..density..), bins = 20, fill = "lightcoral", color = "black",
                 alpha = 0.7) +
  geom_density(color = "black", size = 1) +
  labs(
    title = "Histogram and Density of Absolute Errors",
    x = "Absolute Errors",
    y = "Density") + theme_classic()




########################################################
#### SECTION 6: ESTIMATING THE DISPERSION PARAMETER ####
########################################################
p = length(coef(final_model))
y = dvis$docvis
mu_hat = fitted(model)
phi = 1/(n-p) * sum((y - mu_hat)^2 / mu_hat)
# This is out estimate for the dispersion parameter phi
# This is used to compute the data in Table 9: Recomputed results for our final model,
# with dispersion parameter of 2.073
phi
```