SB1 Assessed Practical 2

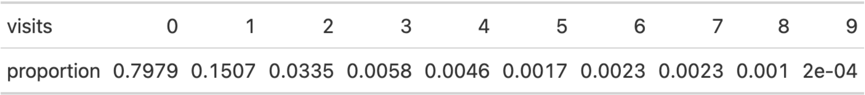
P882

December 7, 2022

Contents

1 Exploratory Data Analysis

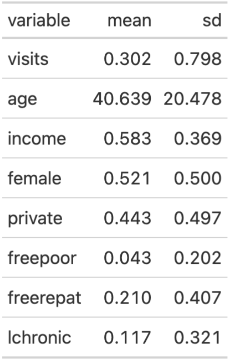
I begin by looking at some summary statistics and exploratory plots of the data. Immediately in Figure 1 we can see that there are very few large counts, with 98% of visits being 0, 1 or 2. This already implies that our intended Poisson model may not be the best choice, due to the high number of zeroes and generally low counts. Also, although the data contains only discrete values for age and income, I treat them as continuous in all model fitting.



visits	0	1	2	3	4	5	6	7	8	9
proportion	0.7979	0.1507	0.0335	0.0058	0.0046	0.0017	0.0023	0.0023	0.001	2e-04

Figure 1:

I also calculate the mean and standard deviation of each of the 8 variables in Figure 2. The important thing to notice here, is that the mean number of visits does not equal its standard deviation. This is a sign of overdispersion, further implying possible problems with Poisson.



variable	mean	sd
visits	0.302	0.798
age	40.639	20.478
income	0.583	0.369
female	0.521	0.500
private	0.443	0.497
freepoor	0.043	0.202
freerepat	0.210	0.407
lchronic	0.117	0.321

Figure 2:

Figure 3 shows histograms for the other two numeric variables: age and income. The data appears to contain a spread of ages, from 19 up to 72 with higher frequency at these extremes. Similarly for income, we have an equal spread in the range. Considering now at the categorical variables and numeric variables, Figure 4 shows the distribution of values for each variable. In each case the highest frequency is on the left. Noticeable features here are the overwhelming majority of people not having free health insurance due to income/disability/age etc. The most common income is $2,500 however the income is spread slightly more evenly. Overall the data seems to be reasonably complete, especially in key variables like gender, age, income and private.

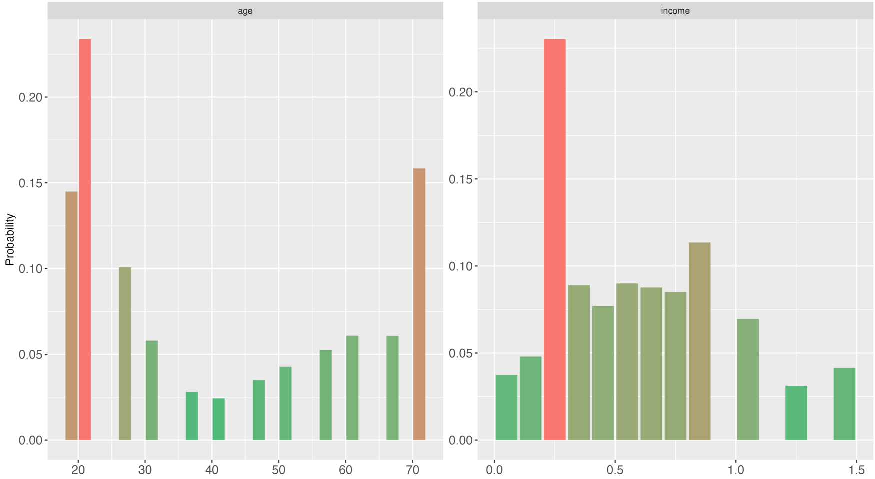


Figure 3:

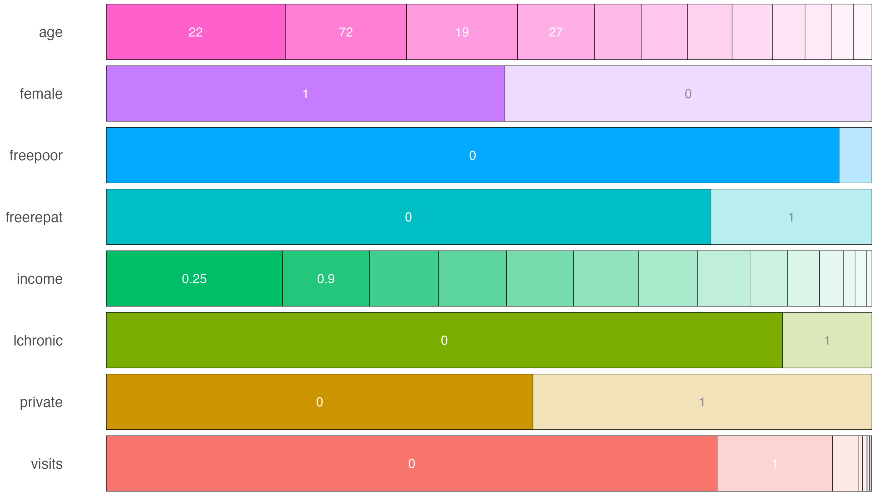


Figure 4:

Now I consider a mosaic plot of age conditional on the number of the visits. The aim here is to see any link between the two variables, and from the plot we can see that higher visit counts (6+) become dominated by older people. However, this plot must be taken with a pinch of salt, as we already know the data is inflated by 0, 1 and 2 visit counts.

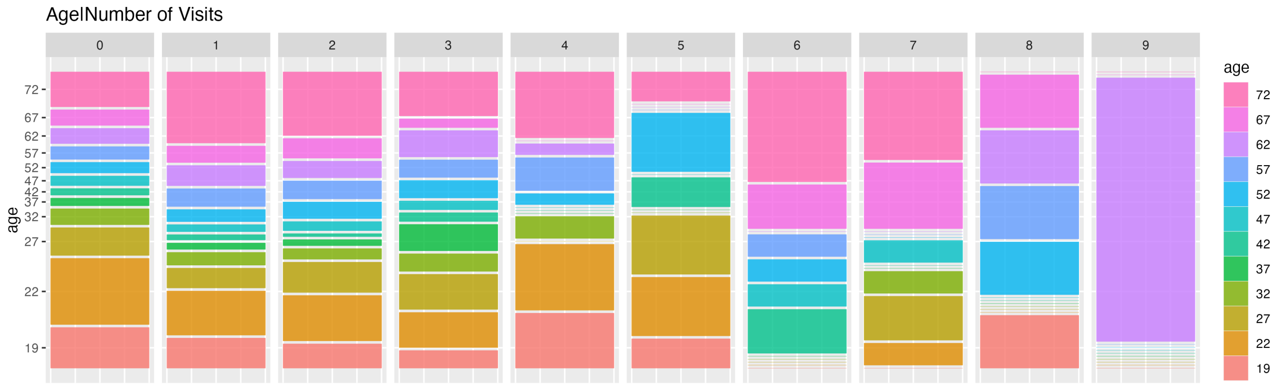


Figure 5:

Figure 6 investigates how gender and private healthcare may affect the number of visits. Overall being female appears to result in increased visits to the doctor, however the presence or lack of private healthcare seems to not have much effect at all.

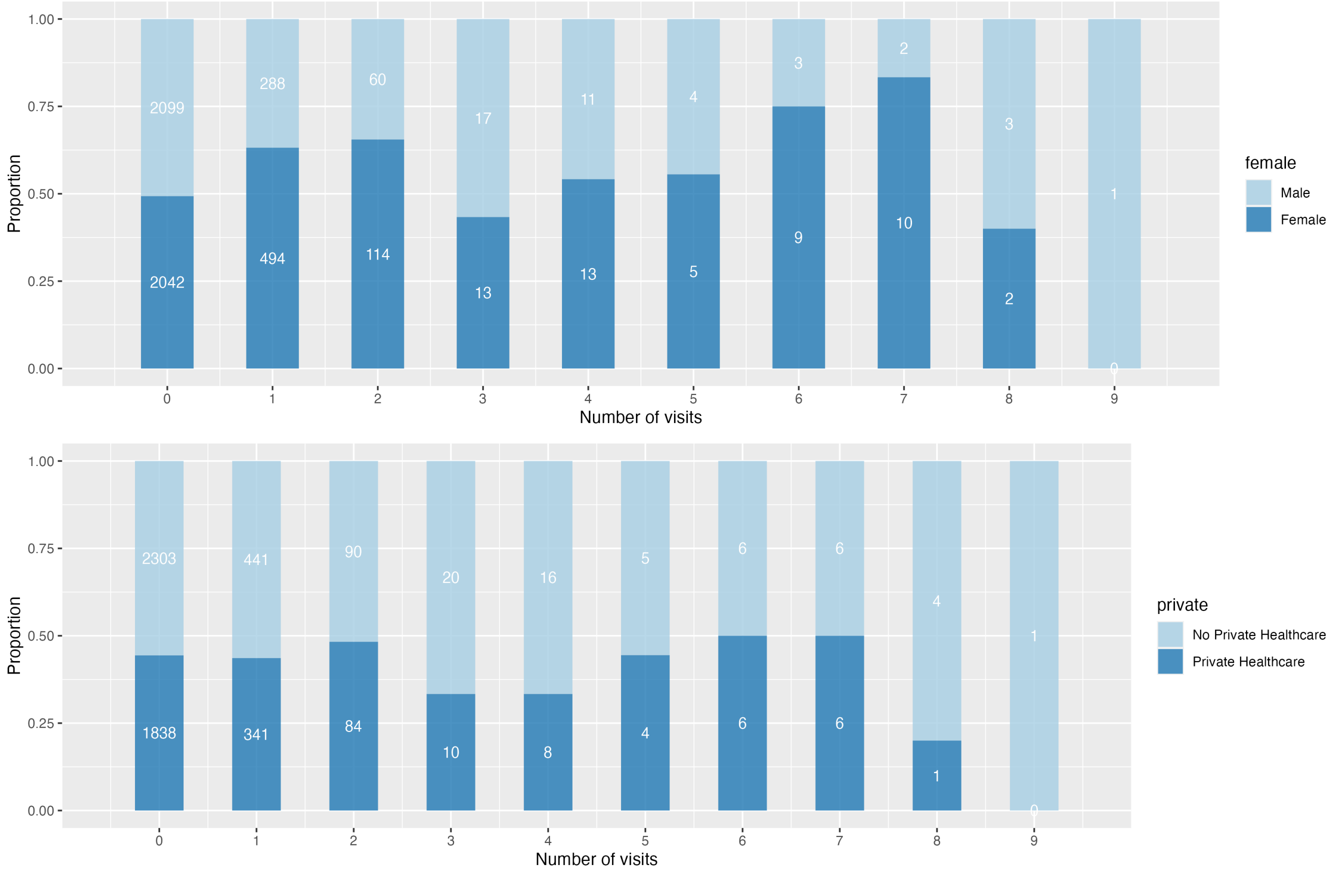


Figure 6:

2 Model Fitting

I begin by fitting a Poisson GLM modelling the expected number of doctor visits, using all the available predictor variables (including interactions of the female variable with each of the rest.):

$$visits \sim age + income + private + freepoor + freerepat + lchronic + female + female : * \tag{1}$$

Figure 7 shows model (1)'s summary. Of the 14 initial predictors, 7 of them are deemed to be (highly) significant according to a Wald test with the "z_value" column. Before cutting down any of these variables, it's worth first checking if this full model is any good in itself. Firstly, it has a null deviance of 5634.8 and a residual deviance of 5271.9. Using these two values we can calculate an LRT statistic of 357.35, for the null hypothesis of $\beta_2 = \cdots = \beta_{14} = 0$ (assuming β_1 is the intercept). Under this null, the LRT has a limiting χ^2_{13} distribution, and the value of $\Lambda = 357.35$ is huge with respect to this, hence we easily reject the null. As a result I am confident this "full" (not saturated!) model is at least better than the null model.

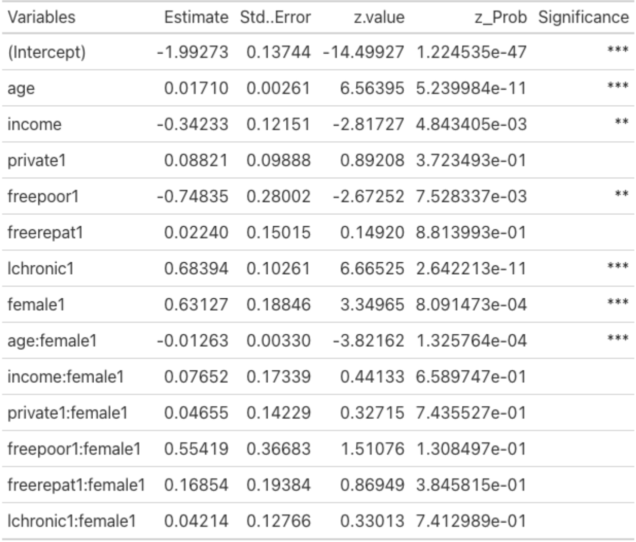
However, 7 of the Wald tests in Figure 7 show that we could simplify our model quite a bit. As a result I drop all 7 insignificant variables and fit the following model:

$$visits \sim age + income + freepoor + lchronic + female + female : age \tag{2}$$

Firstly this is nice because it is has half the number of predictors, increasing interpretability. Secondly, looking at the summary table in Figure 8 shows that all the variables are still highly significant. Overall this model is much more parsimonious, since we appear to have not lost much explaining power at all. To further confirm this we can consider the deviance-based R^2_{kl} of both models, which can be interpreted as the fraction of uncertainty explained by the fitted model (Colin Cameron and Windmeijer 1997, p. 334). Model (1) (full model) has an $R^2_{kl} = 0.0644$ and model (2) (reduced model) has an $R^2_{kl} = 0.0634$. Also stated in the paper is that R^2_{kl} is nondecreasing with more predictors, so it makes sense that my full model has a higher R^2_{kl} value. However, the two values are very similar, hence we clearly have not lost much explainability by reducing the predictors. Also important to note is that the values we get for R^2_{kl}, are still quite low, implying there is room for improvement by finding better predictor variables/a different GLM.

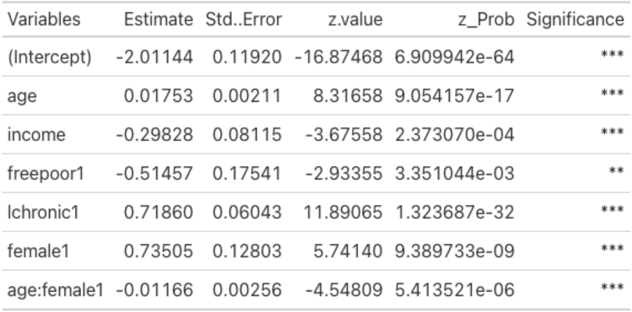
An even further confirmation that model (2) is the best one to choose, is that backward stepwise selection selects the same model based on AIC (ending with an AIC = 7623).

I chose not to use the scaled deviance, D(y) as a goodness-of-fit test itself, due to the data containing only low counts. As a result, D(y) no longer has an asymptotic $\chi^2_{(n-p)}$ distribution and any tests based on it make no sense.



Variables	Estimate	Std..Error	z.value	z_Prob	Significance
(Intercept)	-1.99273	0.13744	-14.49927	1.224535e-47	***
age	0.01710	0.00261	6.56395	5.239984e-11	***
income	-0.34233	0.12151	-2.81727	4.843405e-03	**
private1	0.08821	0.09888	0.89208	3.723493e-01	
freepoor1	-0.74835	0.28002	-2.67252	7.528337e-03	**
freerepat1	0.02240	0.15015	0.14920	8.813993e-01	
lchronic1	0.68394	0.10261	6.66525	2.642213e-11	***
female1	0.63127	0.18846	3.34965	8.091473e-04	***
age:female1	-0.01263	0.00330	-3.82162	1.325764e-04	***
income:female1	0.07652	0.17339	0.44133	6.589747e-01	
private1:female1	0.04655	0.14229	0.32715	7.435527e-01	
freepoor1:female1	0.55419	0.36683	1.51076	1.308497e-01	
freerepat1:female1	0.16854	0.19384	0.86949	3.845815e-01	
lchronic1:female1	0.04214	0.12766	0.33013	7.412989e-01	

Figure 7:



Variables	Estimate	Std..Error	z.value	z_Prob	Significance
(Intercept)	-2.01144	0.11920	-16.87468	6.909942e-64	***
age	0.01753	0.00211	8.31658	9.054157e-17	***
income	-0.29828	0.08115	-3.67558	2.373070e-04	***
freepoor1	-0.51457	0.17541	-2.93355	3.351044e-03	**
lchronic1	0.71860	0.06043	11.89065	1.323687e-32	***
female1	0.73505	0.12803	5.74140	9.389733e-09	***
age:female1	-0.01166	0.00256	-4.54809	5.413521e-06	***

Figure 8:

3 GLM Diagnostics

I begin by checking a plot of the deviance residuals against $\hat{\eta}$ as per Faraway's (2016) recommendation (instead of using $\hat{\mu}$). Figure 9 shows this plot, however we immediately notice it's not great. This due to the fact we are considering a Poisson model with small counts, which tends to break these plots due to the small number of possible values of the visits variable.

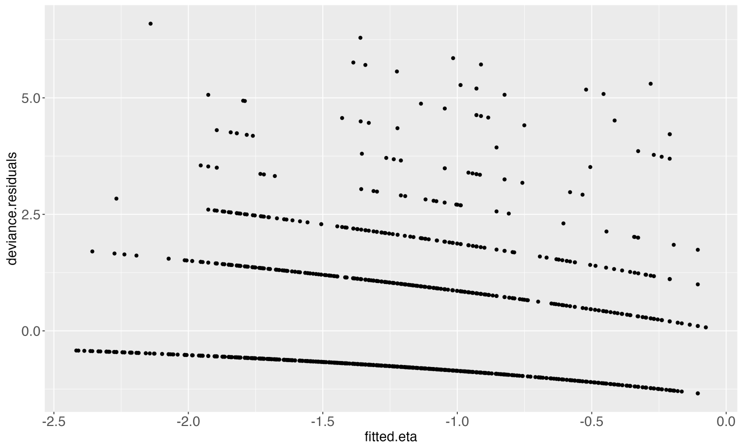


Figure 9:

The next thing we can check is the distribution of the standardised deviance residuals. Under Poisson (with large count) models, we can expect these values to be approximately standard normal. However, with our low count model we can no longer expect this fact completely. They should have roughly unit variance, but no longer normality. Figure 10 shows a Q-Q plot for the standardised deviance residuals, of my chosen reduced model. Clearly the normality is violated, on the right hand tail. The residuals do have a variance of 0.920, close to the desired unit variance, meaning it's not a complete disaster at least.

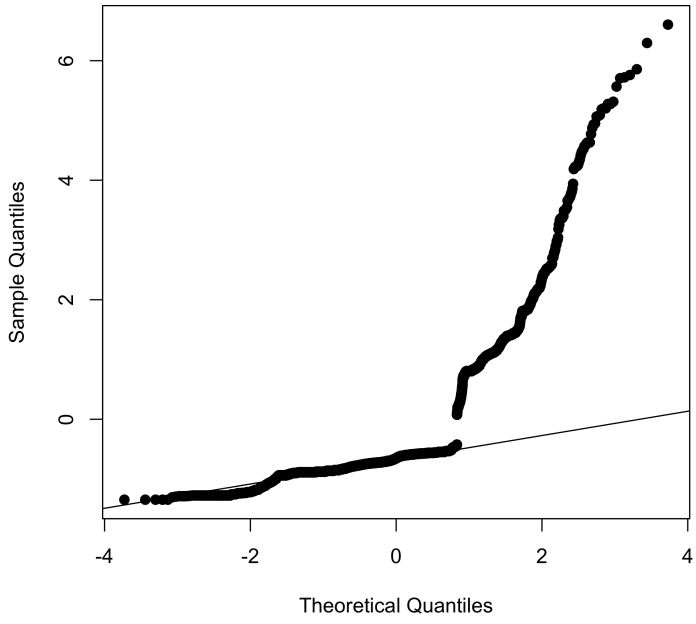


Figure 10:

Finally I considered leverage/influence values for the model. Figure 11 is a (quite hectic) plot which shows the Cook's distance, residual and leverage value for each data point at the same time. We can see that many of the points have residuals that are greater than 2, however this is no longer a truly valid outlier check since we have lost the normality of the deviance residuals. Also from the plot, the three most influential points are highlighted as observations 115, 198 and 630. Looking at the data for these points in more detail doesn't ring any alarm bells for me and I choose to remove not outliers. In particular, the three points seem to just have high visit counts (7, 5 and 8 respectively) and no obvious errors. Due to the lack of high counts in the data, it makes sense these points ended up being influential.

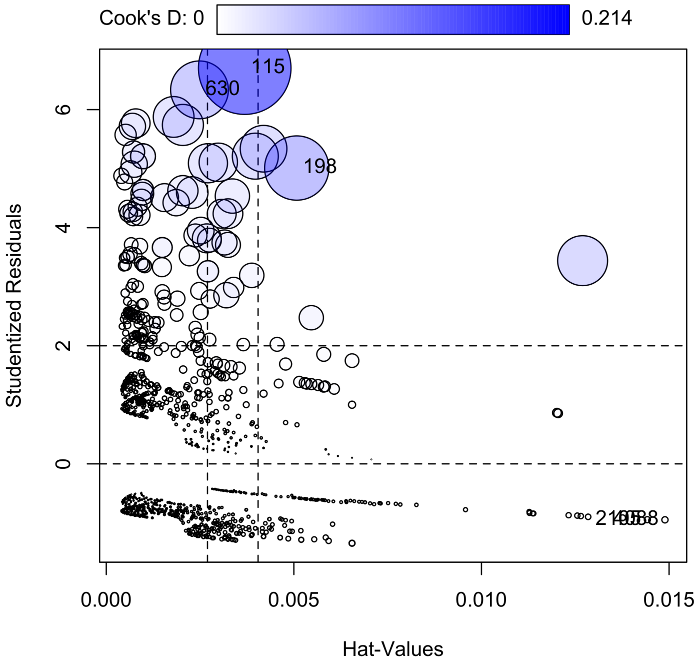


Figure 11:

4 Interpretation of Variable Effects

Having fit model (2) we end up with the following explicit model:

$$\lambda_i = \exp(-2.011 + 0.018 * \text{age} - 0.298 * \text{income} - 0.515 * \text{freepoor} + 0.719 * \text{lchronic} + 0.735 * \text{female} - 0.012 * \text{age:female})$$

Using these coefficients I produced a table (Figure 12) of the effects for each of the variables. The first three variables, 1 continuous and 2 categorical, represent the effect on the mean number of visits when you change each variable by one unit. In the case of the age and female variables, we must also consider the effect of the interaction term. As a result the effect value isn't the same as the coefficient from the model summary in Figure 8, but is a sum of the main effect and interaction effect when the interaction is present. For example, the female effect is calculated as $0.735 - 0.012 = 0.723$. In the case of the age effect, it changes depending on whether the person is male or female, hence why there are two age effect rows.

I now consider the interpretation of these effects, and how they affect the mean number of doctor visits. Firstly, the exponent of an effect relates to a multiplicative change in the mean response count. For example, the income effect of -0.29 corresponds to a $\exp(-0.29) = 0.75$ multiplicative factor change in doctor visits, after a unit change in income. Since this value is less than 1, our fitted model is saying that expected doctor visits decreases as income increases.

Now consider the lchronic variable, which is instead a treatment variable. Its coefficient of 0.719 corresponds to the presence of a chronic condition and since $\exp(0.719) = 2.052$, the model suggests the presence of a chronic condition increases the mean number of doctor visits by a multiplicative factor of ~ 2.

Finally, let's consider one of the variables which had an interaction. Take the age effect, which incorporates the age:female interaction. If the person is female then the coefficient of the age term is now $0.018 - 0.012 = 0.006$, which is lower than if the person is male. This shows that being female, means being older has less of an effect on doctor vists, but in both cases being older results in more doctor visits on average.

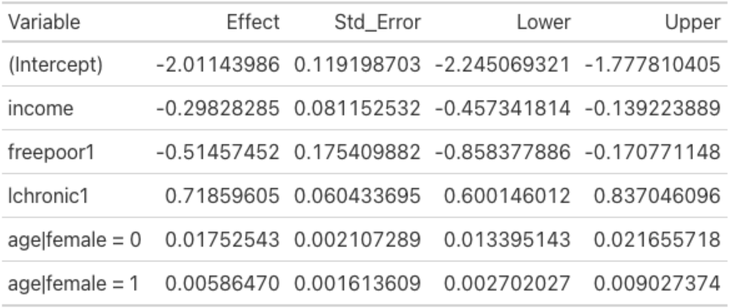
Finally, not shown in the table is the effect of the female variable. Consider one data point $\boldsymbol{x} = (1, \ldots, \text{female} = 0)$ and $\boldsymbol{x}^* = (1, \ldots, \text{female} = 1)$. In each case the model will be $\lambda = \exp(\beta_0 + \ldots)$ and $\lambda^* = \exp(\beta_0 + \cdots + \beta_5 * \text{female} +$

$\beta_6 * \text{age:female}$) respectively. Now consider $\frac{\lambda^*}{\lambda}$:

$$\begin{aligned} \frac{\lambda^*}{\lambda} &= \frac{\exp(\beta_0 + \cdots + \beta_5 + \beta_6 * \text{age})}{\exp(\beta_0 + \dots)} && (3) \\ &= \exp(\beta_5 + \beta_6 * \text{age}) && (4) \\ &= \exp(0.735 - 0.012 * \text{age}) && (5) \end{aligned}$$

From the above, we can conclude that the effect of being female on doctor visits, depends on the value of age. In particular, being female changes your average number of doctor visits by a multiplicative factor of $\exp(0.735 - 0.012 * \text{age})$.

Also shown in Figure 12, are 95% confidence intervals for each of the estimated effects. Since the coefficients are asympotically normal we can simply construct normal confidence intervals using the standard errors. In the case of the age|female=1 effect, it was necessary to include the covariance of the interaction and age coefficients according to $\text{Var}(\beta_4 + \beta_6) = \text{Var}(\beta_4) + \text{Var}(\beta_6) + 2\text{Cov}(\beta_4, \beta_6)$, where β_6 is the coefficient of the age:female interaction. In terms of a confidence interval for the female effect, the same covariance calculation, will include the age term as a linear factor.



Variable	Effect	Std_Error	Lower	Upper
(Intercept)	-2.01143986	0.119198703	-2.245069321	-1.777810405
income	-0.29828285	0.081152532	-0.457341814	-0.139223889
freepoor1	-0.51457452	0.175409882	-0.858377886	-0.170771148
lchronic1	0.71859605	0.060433695	0.600146012	0.837046096
age|female = 0	0.01752543	0.002107289	0.013395143	0.021655718
age|female = 1	0.00586470	0.001613609	0.002702027	0.009027374

Figure 12:

References

Colin Cameron, A. and Frank A.G. Windmeijer (1997). “An R-squared measure of goodness of fit for some common nonlinear regression models”. In: *Journal of Econometrics* 77.2, pp. 329–342. ISSN: 0304-4076. DOI: https://doi.org/10.1016/S0304-4076(96)01818-0. URL: https://www.sciencedirect.com/science/article/pii/S0304407696018180.