KU LEUVEN

Generalized Linear Models

Exam Project

Tutorial Group 5 : G0A18B-GB

ALIU Andi - 0765695

CUYPERS Tina - 0159173

DEQUIER Roman - 0959361

GJELAJ Jozefina - 1007504

POKHODILO Inga - 0893873

TUMULURUR Karthik - 0687716

KU Leuven, Master of Statistics and Data Science

2023-2024

*The students in in red did not contribute to the project

Introduction

In early 1972, the Danish town of Fredericia captured heated national debates when newspapers started publishing headlines about the devastating impact of cancer in town. These headlines were sparked by physicians at Fredericia's hospital when they observed a significant rise in lung cancer admissions in their city. Consequently, studies were conducted to determine if Fredericia indeed had a higher rate of cases than other cities. Notably, one such study conducted by J. Clemmesen analyzed lung cancer cases in Fredericia and three neighboring cities from the period 1968-1971, with his latest findings pinpointing age and residence as significant factors [1].

Building upon this context and Clemmesen's provided data, this report will focus on the effect of age and city on lung cancer cases, keeping in mind previous discussions around Fredericia's high lung cancer cases. To achieve this, we fit a generalized linear Poisson model to analyze the lung cancer case data.

Data

The analyzed dataset consists of 24 rows categorized by age. The variable of interest is cases, with three covariates - population size (pop), four cities (city) and age (age) categorized into six groups.

Table 1 – Lung Cancer Cases and Population by Age and City

Age	**Fredericia**		**Horsens**		**Kolding**		**Vejle**	
	Cases	Population	Cases	Population	Cases	Population	Cases	Population
40–54	11	3059	13	2879	4	3142	5	2520
55–59	11	800	6	1083	8	1050	7	878
60–64	11	710	15	923	7	895	10	839
65–69	10	581	10	834	11	702	14	631
70–74	11	509	12	634	9	535	8	539
Over 74	10	605	2	782	12	659	7	619
Total	64	6264	58	7135	51	6983	51	6026

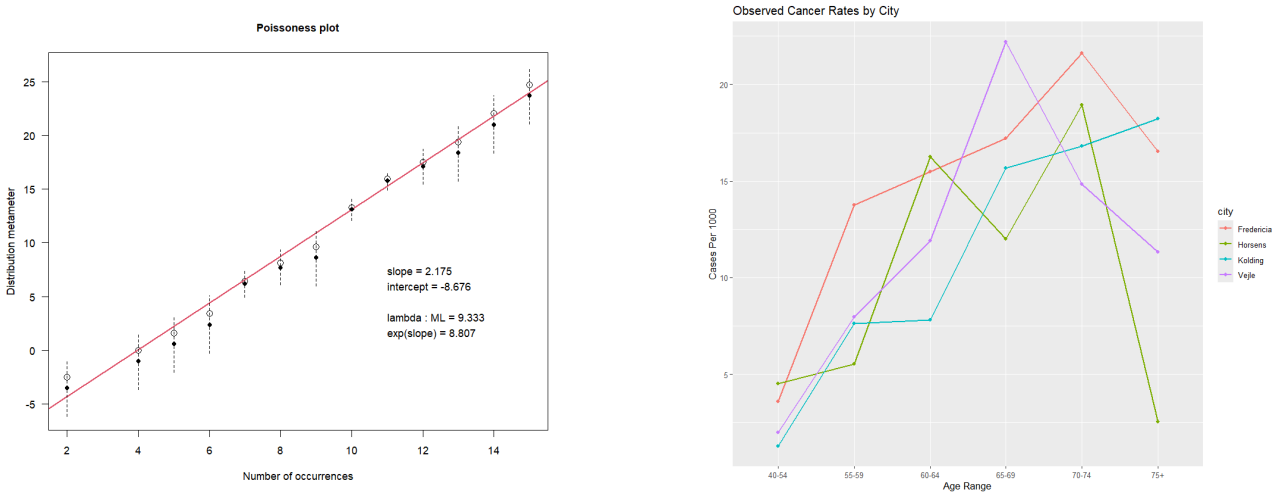
Table 1 shows the lung cancer cases by city, the number of lung cancer cases, and the total population of each age group by city. We see some variability among the cases, but more among population of each city. The sample standard deviation of the variable pop is $s_{pop} = 842$, indicating significant variability between groups. Therefore, we examine the case rates per thousand inhabitants by city. Figure 1(b) shows these rates, indicating that cancer rates increase in the older age groups, but then decrease in the oldest group of people aged 75 and older. The similar trend lines indicate no presence of interaction between age and city. Table 2 shows descriptive statistics for these rates by age and city. We must note that Fredericia has a much higher rate of lung cancer, while the rates of Horsens, Kolding and Vejle are all quite similar. We will try to determine if Fredericia's rate of cancer is higher than the other three cities.

TABLE 2 – Statistics by City and Age

	Statistics			
	Mean	Median	S.D.	Variance
(a) By City				
Fredericia	14.70	16.01	6.04	36.48
Horsens	9.96	8.77	6.75	45.56
Kolding	11.24	11.75	6.69	44.72
Vejle	11.70	11.61	6.76	45.58
(b) By Age				
40-54	2.84	2.79	1.48	2.19
55-59	8.72	7.80	3.52	12.39
60-64	12.87	13.71	3.86	14.90
65-69	16.76	16.44	4.23	17.90
70-74	18.05	17.87	2.90	8.42
75+	12.15	13.92	7.0	49.50

To assess whether lung cancer cases follow a Poisson distribution, Friendly[2] recommends using a "poissoness" plot, which plots the observed counts against a "count metameter," which should result in a linear trend if the counts follow a Poisson distribution. Figure 1(a) shows this plot and it indicates that the a Poisson model is reasonable in this setting.

An important assumption in Poisson regression is that the mean is equal to the variance. Violating it often results in over or under dispersion. The overall mean (9.33) and variance (9.97) of the lung cancer cases are quite close to one another. However, when grouped by city, Horsens seems to have a much higher variance than the mean, while Fredericia has a much lower variance than the mean. Despite this, we proceed with the assumption that the assumption is satisfied as the overall mean and variance are nearly equal. However, we do not ignore this finding, instead we will correct the model for dispersion *pos hoc*, if we see that the assumption is violated.



(a) Poissoness Plot

(b) Cases Per 1000 People

FIGURE 1

Another important assumption that must be satisfied for to fit a generalized linear Poisson model is independence of the counts. This is true if one person having lung cancer in a city has no impact on the risk of another person having lung cancer in the same city. Since lung cancer is not contagious, we assume independence in the counts. We note that this assumption may be violated, but there is no way to check it with the current data.

Model Analysis and Interpretation

As we have seen that the overall lung cancer counts can be adequately described by a Poisson distribution. We therefore fit a generalized linear Poisson model, meaning the parameters will be estimated using maximum likelihood. To identify the model we must specify both the distribution and the systematic parts of the model. The model we work with in this study is :

$$\text{Systematic}: \ \ln(\text{cases}) = \ln(\text{pop}) + \beta_1 \text{city} + \beta_2 \text{age} \quad \text{Distribution}: \text{cases} \sim \text{Poisson}(\lambda)$$

Age is a categorical variable with six levels and is modeled using five dummy variables. We choose to encode city as a binary variable indicating whether the city is Fredericia or not. This choice is made for two reasons. The first is because of small difference in observed cancer rates in the three other cities compared to Fredericia. The second is that in Andersen's study of this dataset, he mentions that original purpose of the data was to determine whether Fredericia had a higher rate of lung cancer compared to the other three cities.[3]. Finally, because we are modelling rates with unequal population sizes we include an offset for the variable pop.

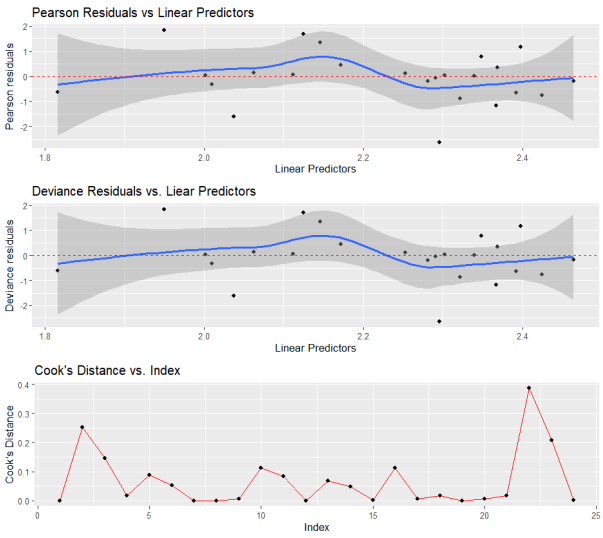


FIGURE 2 – Model Residuals

Fitting this model results in a residual deviance of 23.7 on 17 degrees of freedom, which are close enough for a good fit. Testing for goodness-of-fit using Pearson residuals shows an acceptable fit with $\chi^2 = 22.6, \quad p = 0.16$. The results are similar using deviance residuals with $\chi^2 = 23.7, \quad df = 17, \quad p = 0.13$. Both Pearson and deviance residuals plotted against the linear predictors should result in a horizontal line at 0. The plots in Figure 2 do not display any significant deviation from the expected trend. There is also no evidence of influential observations when plotting Cook's distance vs. the index of the residuals.

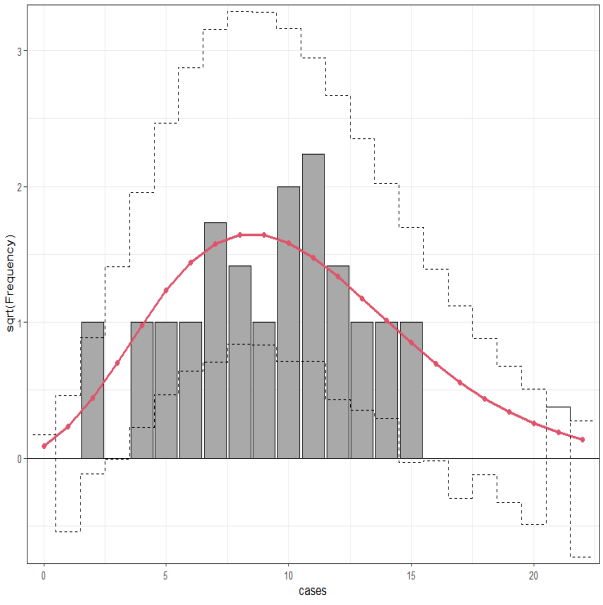


FIGURE 3 – Observed vs. Expected Counts

Another common way to assess the quality of fit of discrete data is by using a rootogram, shown in Figure 3. A rootogram compares predicted counts (in red) vs. the observed counts. We see deviation, suggesting that the model may not fully capture the underlying relationship between city, age and the number of lung cancer cases. However, other than the number of 2 counts, the predictions are well within the dashed confidence intervals.

Failing to detect overdispersion leads to an inflation of standard error and hence, Type-I error. One way to assess over dispersion is by using DHARMa residuals, which are obtained by simulating new response data.

The simulated data is used to create an empirical distribution function and then residuals. These residuals can be used to test whether the simulated dispersion is equal to the observed dispersion. With our model we find a p-value of 0.856, which fails to reject the null hypothesis that there is no overdispersion. Figure 4 plots the predicted mean vs. the variance of the DHARMa residuals, which should follow a linear trend through the origin. The model acceptably models the mean and variance relationship, but does show deviations from the 95% confidence interval.

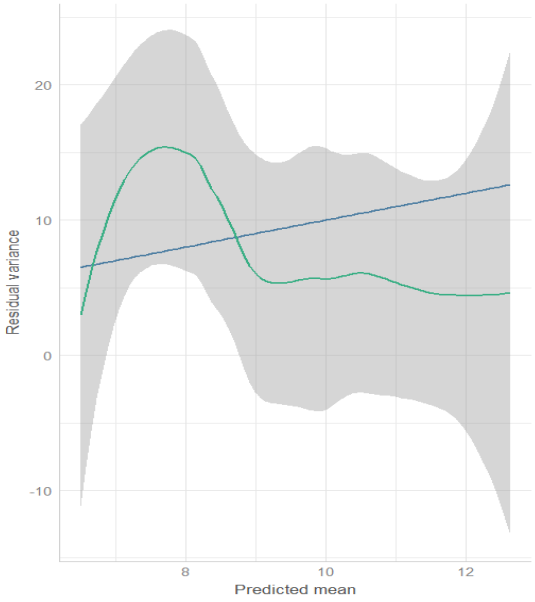


FIGURE 4 – Mean vs. Variance

Pearson residuals can help quantify overdispersion. The dispersion parameter in a generalized linear model is estimated by $\hat{\phi} = \frac{\chi^2}{n-p}$, where χ^2 is the sum of the squared Pearson residuals, and $n-p$ is the residual degrees of freedom in the model. Our model has a dispersion paramater of $\hat{\phi} = 1.33$. For moderate sized datasets, Joseph Hilbe recommends[4] correcting any model with $\hat{\phi} > 1.25$. We see that we have a small to moderate level of overdispersion, for which Hilbe recommends a quasipoisson model. A quasipoisson has the same same point estimates as a Poisson model, but corrects for overdispersion by scaling the standard errors by a factor of $\sqrt{\hat{\phi}}$. This results in wider confidence intervals and reduces the risk of Type I error.

We therefore continue our analysis with the quasipoisson counterpart of the model above. Table 3 shows that under Type III Wald and likelihood ratio tests, the effect of age is highly significant, while the effect of city is borderline significant. The effect of city is borderline significant, underscoring the importance of using a quasi model. Otherwise we would have concluded there is a significant effect of city.

TABLE 3 – Type III Wald and Likelihood Ratio Tests

(a) Likelihood Ratio Test

	χ^2	Df	Pr($>\chi^2$)
city	3.468	1	0.063
age	77.86	5	<0.001

(b) Wald Test

	χ^2	Df	Pr($>\chi^2$)
Intercept	817.3	1	<0.001
city	3.643	1	0.056
age	61.455	5	<0.001

To interpret the coefficients of the model we consider e^{β_i}, which is the rate ratio (RR) between group i and the reference group. Table 4 contains a summary of the results. The reference group for age is 40-54 years old, while for city is the three cities Horsens, Kolding, and Vejle grouped together. The intercept is the rate of lung cancer among 40-54 year-old's.

Our dataset covers the time period between 1968-1971, so this is a lung cancer rate of 0.3% per four person-years. For the cities, Fredericia has a 38.5% higher rate of lung cancer than the combined cities. Since 1 is barely contained within the confidence interval, the Wald and Likelihood ratio tests are borderline significant. We also see that lung cancer rates increase as age increases, to a maximum of 6.4 times as high for those ages 65-69 compared to those who are in the reference group of 40-54. The oldest age group, has a lower rate ratio of 4.14 which may be explained by the high mortality of lung cancer. If people happen to develop lung cancer they are unlikely to live as long.

Conclusion

First, when accounting for city, the data provides strong evidence that the number of lung cancer depend on age (p-value = <0.001 , from $\chi^2 = 61.45$ with 5 d.f.). The cases seem to increase with age, peaking in the age group of 65-69 years old. However, there is also some evidence that city of residence has an effect on the number of lung cancer cases (p-value = 0.056, from $\chi^2 = 3.643$, with 1 d.f.). Fredericia has a higher rate of lung cancer cases compared to the combined rates of Horsens, Kolding and Vejle, altough at a lower significance level.

TABLE 4 – Model Coefficients

	RR	2.5 %	97.5 %
(Intercept)	0.003	0.002	0.004
Fredericia	1.385	0.983	1.921
age55-59	3.008	1.709	5.286
age60-64	4.574	2.719	7.787
age65-69	5.863	3.507	9.941
age70-74	6.419	3.778	11.004
age75+	4.142	2.341	7.304

However, some cautious notes should be stated. We do not have information on the measurement strategy of the lung cancer cases. Nor can we randomly assign lung cancer cases to citizens or randomly select hospitals in the four cities treating these cases. So causality has to be ruled out. Moreover, the effect of age, as well as the effect of city on the lung cancer cases could be overestimated, as other factors that are associated with both, might explain the found association. For instance, factors like smoking may be more common at a certain age or in a certain socioeconomic group. While the population of one city has a different composition when it comes to their socioeconomic groups or fashion in smoking. These, and other confounding variables, such a genetic predisposition to lung cancer, air pollution, should be then considered in future studies.

References

[1] Jens Steensberg. *Environmental Health Decision Making : The Politics of Disease Prevention.* 1989.

[2] M. Friendly. *Visualizing Categorical Data.* SAS Institute, 2000.

[3] Erling B. Andersen. Multiplicative poisson models with unequal cell rates. *Scandinavian Journal of Statistics*, 4(4) :153–158, 1977.

[4] Joseph M. Hilbe. *Statistical Modeling with Count Data.* Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA, first edition, 2014.