Advanced Statistics: Final Project

Dataset: German health registry for the year 1984

A Statistical Analysis of Factors Impacting the Utilization of HealthCare in Germany, 1984

Alli Ajagbe (U20210004)
Aditya Tyagi (U20210003)
Niranjani A. (U20210044)

December 17, 2023

Introduction

The dataset used for our project is *rwm1984,* data collected by the German Health Registry for the year 1984 containing 3,874 observations on 17 variables. Key variables are divided into numerical and binary, covering a range of attributes regarding the usage of medical facilities (*docvis, hospvis*, etc) to the lifestyle of the patient (married, kids, educ).

Taking a closer look at the target population, we can note that all patients are between the ages of 25 to 64, with 47.9% of those being female, 78.9% married, and 44.9% having children. The focus on individuals aged 25 to 64 is strategic, capturing a significant portion of the working-age population. This demographic is critical in understanding the impacts of healthcare policy, as they are often the most affected by changes in healthcare systems due to their economic and social roles.

We leverage methodologies and models featured in Joseph M. Hilbe's work on "Modelling Count Data" to substantiate our findings within the framework of existing literature.

Why Bother

The timing of significant healthcare reforms in Germany underscores the relevance of this study. These reforms, aimed at cost containment and improving healthcare efficiency, present an opportunity to assess their impact. The Cost Containment Act(1977), Hospital Reform Act (1982-1986), and Healthcare Reform Act (1989) were all policies aimed at providing sickness funds and healthcare planning to the working class of Germany. Analyzing this data could provide valuable insights into the effectiveness of these reforms and inform future policy decisions.

Methodology

Data Preprocessing:

To prepare our dataset for potential model applications, we inspected specific columns. The age column was normalized to centre its values around zero, while the income column was subject to logarithmic transformation. Concurrently, we checked for the presence of missing values (NaN) to assess the necessity of handling procedures.

Hypotheses:

In our dataset, we selected two variables, the count of doctor visits and the number of hospital visits, as our outcome variables. We aimed to investigate whether these variables were influenced by the information provided by other variables. Hence, we started our project with the null hypothesis that doctor visits were not dependent on the other variables. Correspondingly, our alternative hypothesis asserted that at least one of the variables significantly influenced doctor visits. We posited the same hypotheses for hospital visits.

Model Selection:

docvis

Based on the data type of both of our outcome variables, we considered starting with a simple Poisson regression model. Our visual plot from trying to fit a Poisson distribution parameterized by the calculated mean, 3.16, of the *docvis* column shows that Poisson could not accurately fit the underlying distribution. This was particularly concerning as the fitted Poisson showed that the expected proportion of zero visits was about 4.23% when there was actually 41% in the data. We went further to plot the residuals of the Poisson regression fit and ascertained it was not centred at zero. The mean, 3.16, and the variances, 39.39, were not equal. The dispersion statistic, 11.34, being greater than 1, showed overdispersion and helped us verify that Poisson was not the best model for the data. Proceeding with it could lead to inaccurate findings where we take some predictors as significant when they are not, as the standard errors of the estimates might be underestimated.

Hence, we decided to look at another count-based regression model, but this time, one that is adjusted for overdispersion - Negative Binomial (NB), as the variance for the distribution, $\boldsymbol{\sigma}^2$, is increased by $\boldsymbol{\alpha\sigma}^2$, where $\boldsymbol{\alpha}$ is the dispersion parameter. The assumptions for NB are such that the variance has to be greater than the mean, and there is a linear relationship between the logarithm of the expected count and the predictor variables. While NB provided a better fit compared to Poisson, the plot of its residuals was also not centred at zero, and the dispersion statistic was 3.13.

Finally, further scrutiny of our data motivated the need to explore models that considered data with excess zeros. We realized that the *docvis* had quite several zeros. These zero counts could be attributed to the individual belonging to a particular subgroup along with being part of some standard distribution. Therefore, we decided to experiment with a Zero-Inflated Poisson Regression model. Apart from modelling the count data as a Poisson distribution, this model uses logistic regression to determine whether an observation is more likely to be a true zero or a count from the Poisson distribution. Zero-inflated Poisson Regression was then fitted to the data and the model converged successfully. We used this for subsequent probings in our data.

hospvis

We fitted a Poisson regression to the *hospvis* data. As seen in the plot, the fitted line aligns well with that of the data, with both peaking at zero. Despite the dispersion statistic being higher than 1, we decided to proceed with Poisson regression as the fitted regression line is still able to yield approximately the same output as seen in the data. For example, using the calculated mean for *hospvis*, 0.12, we expect that 88.6% of the observations in the data have a zero count, compared to the 92.2% in the actual data.

Subsequent Hypotheses

Provided that we reject the hypothesis that none of the explanatory variables explain the variability in our outcome variable, we can only say that at least one of the variables impacts the number of visits. Hence, following the happenings in Germany in 1984, we wanted to check whether age, income, or years of education affected the outcome.

Variable of Interest	**Null Hypothesis**	**Alternate Hypothesis**
Age	There is no significant association between age and the frequency of doctor visits.	Age is associated with variations in the frequency of doctor visits.
Household Income	Household income does not significantly impact the frequency of doctor visits.	There is a significant association between household income and the frequency of doctor visits.
Years of Education	The number of years of education does not influence the frequency of doctor visits.	The number of years of education is associated with variations in the frequency of doctor visits.

We posited the same hypotheses for hospital days.

For each of the hypotheses, the zero-inflated model was fit once with all predictor variables and again with all predictor variables except the one that is tested in the hypothesis, for doctor visits, while we used the Poisson regression for hospital visits. The Likelihood Ratio Test (LRT) is performed to compare the full model to the reduced model to determine if the full model significantly improves the fit. LRT is chosen as it is specifically designed for comparing nested models and generalized linear models. The p-value of the test is also considered during hypothesis testing.

Results

We conducted the LRT for the full model and the reduced model for both of the outcomes. The full model had all the variables and the reduced model had none of the variables.

For *docvis*, the LRT statistic was 792.8 with a p-value of 0. Hence, we reject the null hypothesis that none of the explanatory variables had an effect on the number of doctor visits. This leads to the conclusion that at least one of the explanatory variables affects the outcome variable. Then, we formulate hypotheses and test which of these variables influence the outcome of the response variable.

For *hospvis*, the LRT statistic was 44.4 with a corresponding p-value of 4.78e-07. Hence, we rejected the null hypothesis that none of the explanatory variables influenced hospital visits. Based on this, we can only conclude that at least one of the explanatory variables is affecting the outcome variable. We then proceeded to look at the impact of particular variables that might be influencing the outcome.

From the results of the LRT, we conclude the following for the three hypotheses:

Outcome Variable	**Hypothesis variable**	**Statistic**	**Conclusion**	**Effect on response variable**
docvis	**Age**	LRT Statistic: 171.36 P-value: 0.0	**Reject** the null hypothesis	13.4% increase for a unit change in years holding other variables constant.
	Household income (logarithm)	LRT Statistic: 58.15 P-value: 2.42e-14	**Reject** the null hypothesis	14.3% decrease for 1% increase in income holding other variables constant.
	Years of education	LRT Statistic: 4.74 P-value: 0.0293	**Reject** the null hypothesis	5.66% increase for a unit change in years of education holding other variables constant.

hospvis	**Age**	LRT Statistic: 2.73 P-value: 0.098	**Fail to reject** the null hypothesis	In the presence of other variables, age does not have a significant effect on hospital visits.
	Household Income (logarithm)	LRT Statistic: 7.32 P-value: 0.0068	**Reject** the null hypothesis	24.2% decrease with a 1% increase in income holding other variables constant.
	Years of Education	LRT Statistic: 0.0094 P-value: 0.92	**Fail to reject** the null hypothesis	In the presence of other variables, years of education does not have a significant effect on hospital visits.

Interaction:

We consider the combined effect of two predictor variables on the response variable. Hypotheses testing is done to infer the significance of the combination term. The following hypotheses were formulated:

1. H0: Interaction between log of household income and years of education is not significant.
   H1: Interaction between log of household income and years of education is significant.
2. H0: Interaction between age and years of education is not significant.
   H1: Interaction between age and years of education is significant.

Interaction Terms	**Statistic**	**Conclusion**	**Effect on response variable (*docvis*) for a unit change in predictor variable, holding other variables constant**
Household income (logarithm) and years of education	P-value: 0.015	**Reject** the null hypothesis	2.57% decrease

Age and years of education	P-value: 0.0	**Reject** the null hypothesis	2.04% increase

Conclusion

An increment of one unit in years, while holding other variables constant, corresponds to a 13.4% increase in the number of visits to the doctor. Conversely, a 1% increase in income, with other variables held constant, is associated with a 14.3% decrease in the number of doctor visits. Furthermore, a one-unit change in years of education, under the condition of other variables remaining constant, is linked to a 5.66% increase in the number of visits to the doctor. Interacting some of the variables also yielded interesting insights regarding the additional effect a variable might have with another. For instance, the interaction of household income and years of education showed a 2.57% decrease in the number of visits to the doctor, and an interaction of age and years of education presented a 2.04% increase in the number of visits to the doctor.

However, for days spent in the hospital, when considering the influence of other variables, it was determined that age does not exhibit a significant effect. Also, in the presence of other variables, a 1% increase in income is associated with a 24.2% decrease in the number of hospital days. Similarly, in the presence of other variables, years of education do not exert a significant effect on the number of days spent in the hospital.

These findings are not random and may have their roots in the state of the German Healthcare System at the time of the study. There were significant differences in the delivery of ambulatory and hospital care, especially for the proportion of the population unable to avail of privately funded healthcare, leading to issues such as lengthy referral chains, duplication of equipment, and repetition of diagnostic tests by different doctors in the public sector. These inefficiencies were particularly pronounced for sickness funds, which covered the aforementioned 60% of the population, especially those on the lower income side of the spectrum. Therefore, the findings make sense given the context of existing disparities in healthcare access and quality in Germany for lower-income groups and older populations.

Incorporating feedback:

Initially, we fit the regression model to the data with all predictor variables and looked at the coefficients and the p-value for each of the predictor variables. This was used to tell if a predictor variable (used in the hypothesis testing) had a significant effect on the response variable, after checking for p-value < 0.05. This method was incorrect as this accounts for all predictor variables in the dataset and is not a true representation of the effect of one variable alone. Therefore, the conclusions regarding our hypothesis tests were not true.

The right way to conduct the test would be to fit the regression model to the data with all predictor variables (full model) and then fit the model to the data without the variable of interest for a particular hypothesis (reduced model) when all other predictor variables are held constant. The LRT statistic and p-value for this are calculated using the results of the full model and the reduced model. If the statistic is found to be significant, the LRT is used to conclude if the removed predictor variable is statistically significant or not.

Another feedback was to incorporate interaction terms and see the combined effect of two variables on the response variable. This was to account for the probability that one variable might have a higher impact on the response variable, depending on the presence or absence of a second predictor variable. The interaction terms were taken to be a product of two predictor variables.

Therefore, we fit the zero-inflated Poisson regression model to the *docvis* data and Poisson for the *hospvis* data, and then for each of the hypotheses we removed the predictor variables age, household income, years of education and their interaction terms. The LRT statistic and p-value were computed for each of these and the conclusions of the hypothesis test were re-done. The corresponding change in response variable for significant predictor variables was calculated. This concludes our analysis.

Work Cited:

1. Hilbe, J. M. (2014). Modelling Count Data. Cambridge University Press.