Project 1

Modeling the Number of Physician Visits and it's Potential Correlation with Insurance Status in Elderly Individual's

February 28, 2024

Abstract

In this report, I'll be building various negative binomial models to potentially portray the underlying propensity of elderly individuals visit their physician. Furthermore, I will be reviewing the differences in this propensity between groups of elderly individuals that are insured and not insured. Additionally, looking at the parameters, there is additional insight on the heterogeneity or homogeneity of physician visits between the groups. For example, depending on healthcare status, whether or not the decision of elderly individuals vary widely or if the population truly generally share a similar tendency to visit their physician.

1 Introduction

The dataset used originated from the US National Medical Expenditure Survey (NMES) collected between 1987 and 1988 of elderly individuals above the age of 66. It contains various count dataset of elderly individuals ranging from the count physician visits, chronic, etc., to more identifying features such as race, location, and employment, among a few others (refer to table 1 for detailed description of features). For the purposes of the project, I'll be focusing on the count of physician office visits for each individually. More specifically, I'll be looking at the underlying differences in propensity to visit the physicians office between those that are insured and uninsured. In other words, do elderly individuals that visit their physicians more often perhaps have a higher inclination to do so due to the freedom of being insured?

Variable	Definition
ID	Patient ID
visits	Number of physician office visits
nvisits	Number of non-physician office visits
ovisits	Number of physician hospital outpatient visits
novisits	Number of non-physician hospital outpatient visits
emergency	Emergency room visits
hospital	Number of hospital stays
exclhlth	Factor indicating self-perceived health (Excellent Health)
poorhlth	Factor indicating self-perceived health (Poor Health)
chronic	Number of chronic conditions
adl	Factor indicating if individual has condition that limits living condition
noreast	Factor indicating region (Northeast)
midwest	Factor indicating region (Midwest)
west	Factor indicating region (West)
age	Age in years (divided by 10)
black	Factor indicating if the individual is African-American
male	Factor indicating if the individual is a male
married	Factor indicating if the individual is married
school	Number of years of education
faminc	Family income in USD by 10,000
employed	Factor if the individual is employed
insurance	Factor if the individual is covered by private insurance
medicaid	Factor if the individual is covered by Medicaid

Table 1: Variable name and definitions

2 Features of the Dataset

A notable characteristic of the data is that there appears to be a high proportion of zeroes in both populations. This may be a result of many elderly individuals that absolutely refuse to regularly or even rarely visit the physicians office, regardless of coverage. This is an issue that I'll cover in the modeling section. Another characteristic of the data is that there seems to be some over-dispersion specifically in the group of elderly individuals that are uninsured. This can be observed by the excess amount of zeros prevalent in the data in figure 1. Although there is also an excess amount of zeros in those that are insured in figure 2, there doesn't appear to be as much over-dispersion as there seems to be much more homogeneity in that population.

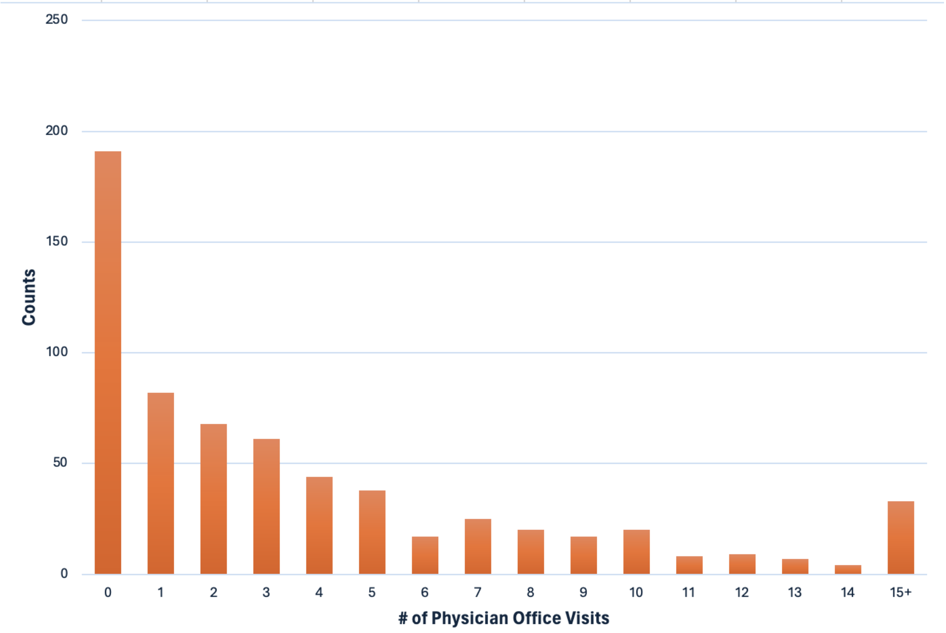

Figure 1: Distribution of the number of physician office visits with no Insurance.

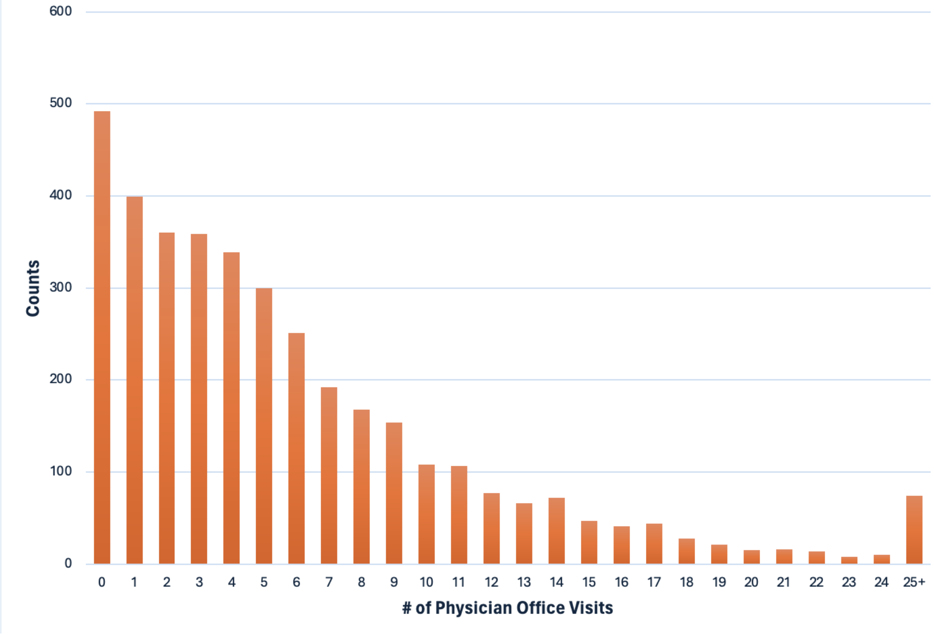

Figure 2: Distribution of the number of physician office visits with Insurance.

Below at table 1, there is a summary statistics of the columns of interest with respect to the pooled and sub- populations. Notice that although I right-censor for graphing in figure 1 and figure 2, the mean, variance, and median found below are found within the original data set containing all the individual level values prior to aggregation. Notice that despite the larger zero-inflation in the uninsured population, the mean still appears to be higher. This can potentially indicate signs of heterogeneity where there seems to exist a higher density of individual on the extreme ends of those that will never visit or visit extremely often.

Model	Parameter	Value
# of Physician Office Visits	Mean	5.7744
(Pooled)	Variance	6.7592
	Median	4
# of Physician Office Visits	Mean	5.6800
(Insured)	Variance	6.7400
	Median	4
# of Physician Office Visits	Mean	6.7139
(Uninsured)	Variance	6.8863
	Median	5

Table 2: Summary Statistics of Column of Interest for Pooled and Subgroups

3 Negative Binomial Modeling on the Pooled Population

Before modeling, notice that I have decided to right-censor the dataset beyond 25+ visits as the bins beyond to get sparse beyond this point and the range even extends to 83 visits. To determine whether or not there might be a potential difference between the pooled population or the individual subgroups, I first built a plain and spiked NBD model to gauge the performance of the spike and also provide a baseline Log-Likelihood of the pooled data when later comparing models using the Chi-square Likelihood Ratio Test. Below at table 1, there is some valuable metric values regarding the regular and spiked NBD model to tell us about model fitment and the heterogeneity of the pooled population.

Model	Parameter	Value
Plain NBD	r	1.0484
	α	0.1859
	Chi-sq GoF P-value	0.0072
NBD w/ Spike @ 0	r	1.3028
	α	1.2387
	Spike Value	0.2098
	Chi-sq GoF P-value	0.0461
	Chi-sq LRT P-value	<0.001

Table 3: Parameter and Test Metric Values for NBD Model(s) over pooled population

Notice that the chi-square goodness of fit test for the plain NBD from table 1 reveals that the base model isn't actually a good fit as it yields a p-value below 0.01. Therefore, as I mentioned earlier, I'll be progressing with the future models using a spike at 0 as there seems to be excess 0's in both subgroups and even the overall population. After the spike at zero, notice that there is a significant improvement in model fitment yielding a goodness of fit p-value greater than 0.25 and, from visual inspection from 3 and the Chi-square Likelihood Ratio Test, the fitment is quite exceptional. One caveat to this analysis was the right censoring of nearly 20+ additional values that had very few observations so potential issues that may arise from this loss of information. From inspection though, it's clear that the model does a moderately acceptable job at estimating the value of the right censored cell, although future alterations to the right censored cell could potentially be added to add more accuracy to this right-tail end.

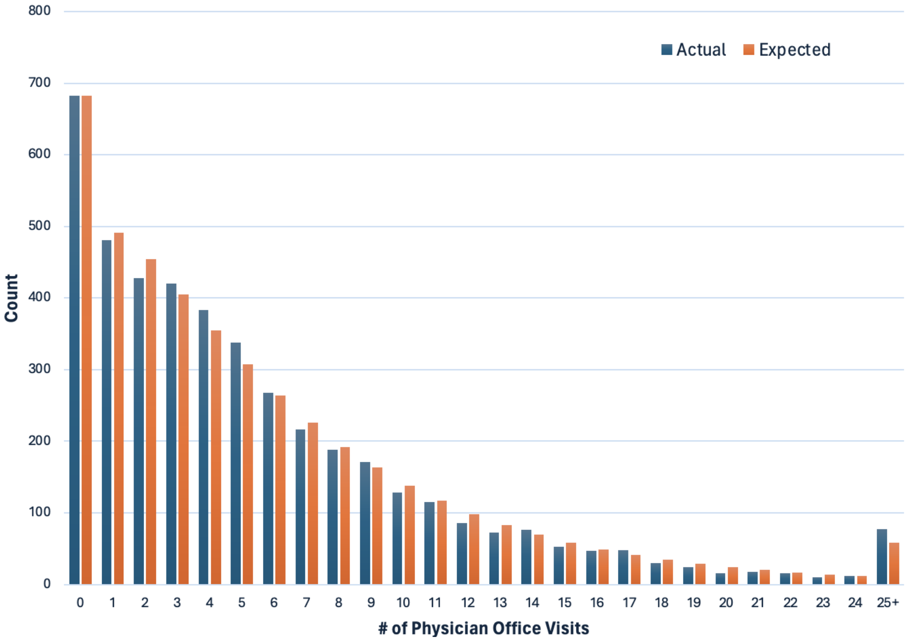

Figure 3: Actual vs. expected distribution of pooled population count of physician office visits

Looking back at table 1, I want to put emphasis primarily on the parameter value of r and the spike at zero. The r value of ∼1.2387 tells us that the pooled population is generally more homogeneous that heterogeneous revealing that the pooled population generally has a similar propensity to visit their physician office. It's important to note that the r value isn't significantly high, indicating that although the pooled population is more homogeneous than heterogeneous, there is still a good portion of the population on the extreme that chooses to either never visit the physicians office or excessively visit the physicians office. This is further supported by the presence of the spike at zeros, where the spike value of 0.0459 or 4.59% of zeros in the pooled population are potentially classified as individuals that under no circumstance will they choose to visit the physicians office. Although, I am skeptical of r value results because there is such a size imbalance between the two subgroups so the larger group will have a tendency to pull the overall distribution shape towards its shape.

4 Negative Binomial on the Subgroups

Before modeling, notice that I have decided to right-censor the dataset beyond 25+ visits for the insured group and 15+ for the uninsured group, which were the respective points in the subgroups before the bin counts began to fall below 5. I suspect that since the uninsured population is more heterogeneous and contain far less data points overall, this led to the counts in the right-tail to eventually reduce to below 5 far earlier. As for the analysis, I'll now be exploring the Zero-Inflated Negative Binomial model on both the subgroups of those which are insured and those that are not insured. Before I begin my analysis, from initial inspection from figure 1, I suspect that for the elderly individuals without insurance, there is likely much more heterogeneity among this group. On the other hand, the elderly individuals with insurance will be more homogeneous since the ability to visit the physicians office may exists in their insurance plan. Therefore, for the figure 1, I'd expect a lower r value and, for figure 2, I'd expected a higher r value.

Model	Parameter	Value
Plain NBD (No Insurance)	r	0.6262
	α	0.1560
	Chi-sq GoF P-value	0.2581
Spike NBD (No Insurance) @ zero	r	0.8727
	α	0.1943
	Spike Value	0.1152
	Chi-sq GoF P-value	0.4837
	Chi-sq LRT P-value	0.0514
Plain NBD (Insurance)	r	1.1615
	α	0.1960
	Chi-sq GoF P-value	0.0168
Spike NBD (Insurance) @ zero	r	1.2974
	α	0.2127
	Spike Value	0.0293
	Chi-sq GoF P-value	0.0730
	Chi-sq LRT P-value	0.0168

Table 4: Parameter and Test Metric Values for NBD Model(s) over subgroups

As seen previously, within the pooled population, results of our spiked model revealed that there potentially exists individuals in our population that might never visit the physician regardless of insurance condition. Therefore, in table 4 the parameter values and goodness of fit results below for reference to provide insight on the improvement through a spike on our subgroups. Note that the plain model for the uninsured group already performed quite well and the spike isn't entirely necessary as shown through the chi-squared LRT test. Regardless, for consistency and a moderate increase in p-value, I will be using the zero-inflated models as my primary model for analyzing r and α.

Notice in our no insurance group, the group appears to be much more heterogeneous with r<1. On the other hand, the group with some form of insurance boasts much more homogeneity with r>1. Additionally, the difference in spike values indicate there are clearly more "Hardcore Never Buyers" in the no insurance group. The results from these models reinforce my previous intuition from figure 1 and figure 2, where I had already suspected a difference in heterogeneity and homogeneity in the two subgroups. Regarding the difference in spike, this could be potentially reasoned that more uninsured elderly individual's will refuse to visit the physician's office since it's out of their convenient bandwidth. Therefore, I would expect the spike percentage for the uninsured population to be much higher than the latter. Ultimately, the model fit is quite satisfactory both visually, found in figure 5 and figure 5, and through the Chi-square Goodness of Fit test in table 4.

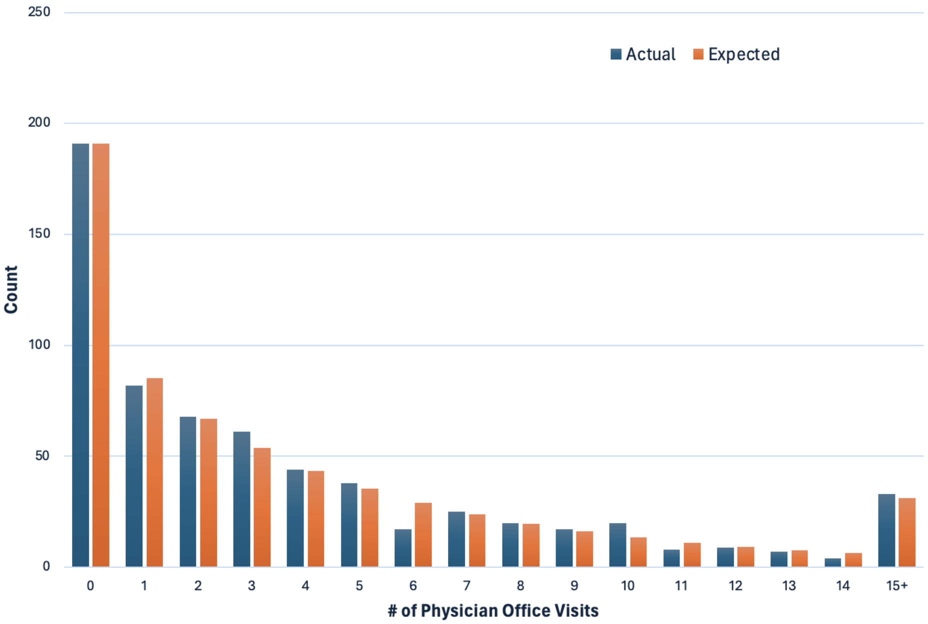

Figure 4: Actual vs. Expected Zero-Inflated NBD of elderly individuals with no insurance.

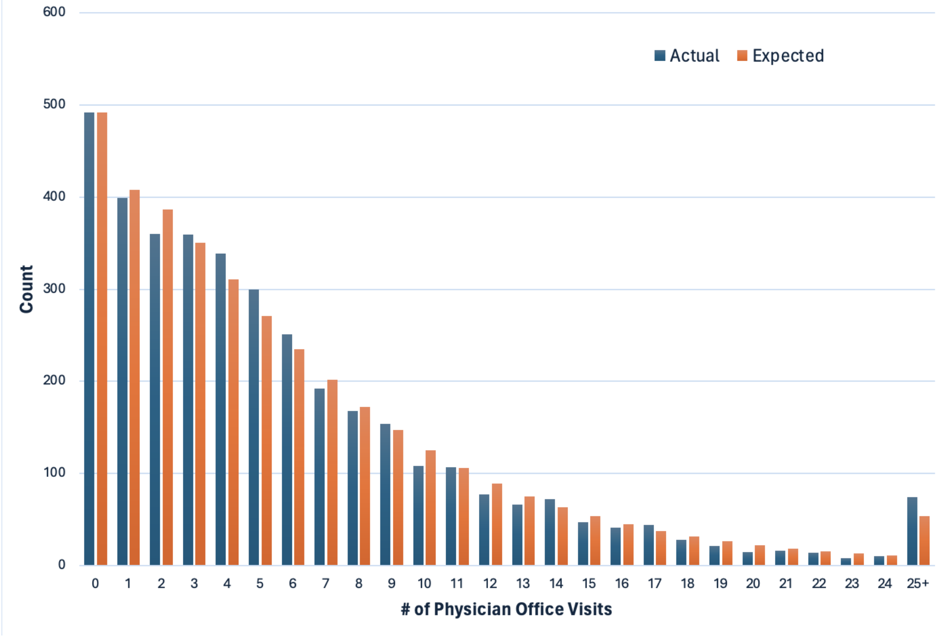

Figure 5: Actual vs. Expected Zero-Inflated NBD of elderly individuals with insurance.

5 Evaluating a Method of Moments Negative Binomial

Another method I explored was an NBD model using method of moments. Before any modeling, I had already suspected that the model would produce poor performances overall since the mean and variance were calculated through the observed data rather than knowing the true population mean and variance.

Below in figure 5 and table 4, the distribution and parameter metrics of the Method of Moments NBD on the pooled data can be found:

Model	Parameter	Value
MoM NBD	r	0.8354
	α	0.1446
	Chi-sq GoF P-value	<0.001

Table 5: Important Metrics for MoM NBD on pooled data

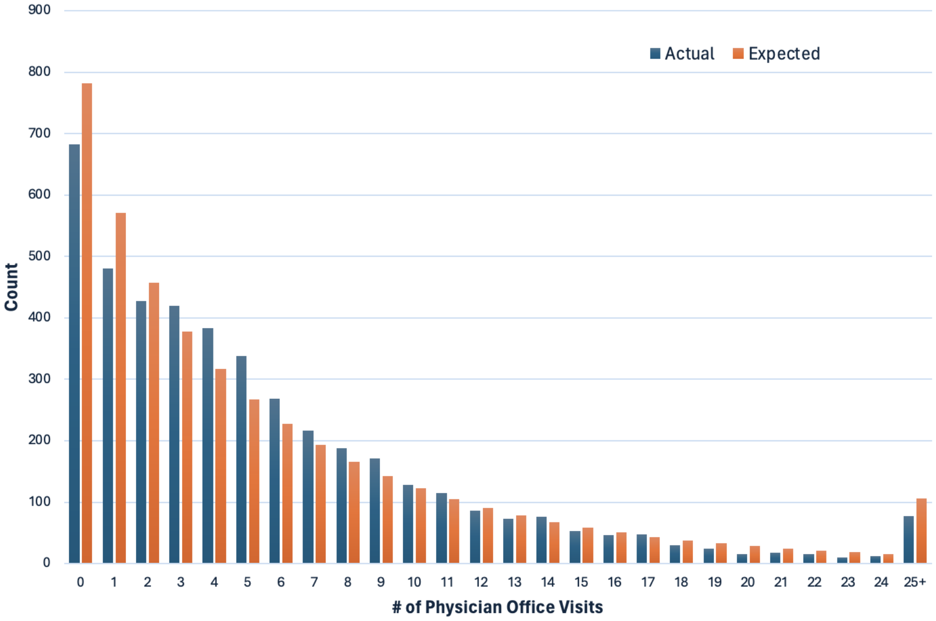

Figure 6: Actual vs. expected distribution of pooled population using MoM

As expected, from visual inspection of figure 6 and table 5, there is a notable performance decrease from even the regular NBD without zero-inflation. The method of moments approach assumes that that the sample moments are an unbiased estimator of the true population moments. Therefore, there might be a potential violation of this measurement that's inhibiting the method of moments method from producing sufficient results. Notice that another feature of the estimation is that the model overshot the values of 0 and 1 by a large margin. This may indicate that the method of moments approach doesn't provide a sufficient answer to the excess zero values since the r and α were built through the means and variance of the sample data. As such, the sample mean and variance might be heavily swayed by the surplus zero values from elderly individuals that might never visit the physician's office regardless of circumstance and this may not be representative of the true population propensity.

6 Subgroup vs. Pooled Model Evaluation

To explore whether the subgroups provide a substantial improvement rather than pooling the whole data through a Chi-square LRT. Based on the results from table 6, this indicates that there appears to significant difference in the count of physician office visits among the two subgroups. Through the visual analysis of the previous figures and tables, it's visible that the two population do show two inherently different patterns. Those without insurance seem to be much more heterogeneous than those with insurance. Therefore, prior to the Chi-Square LRT results, I expected to receive results that would indicate a significant difference between the two models as there shouldn't be a "one-size fit all" r-value from the pooled population that could individually describe the distribution from both subgroups comparably.

Metric	Value
Chi-Square LRT P-value	<0.001

Table 6: Chi-Square LRT between Pooled Data and Subgroups (Insurance and No Insurance)

7 Conclusion and Future Implications

To conclude, the results reaffirm that there is evidence towards the fact that elderly individuals without insurance behave different that those when it comes to visiting their physician's. This is important to understand because visiting the physician's is something that becomes increasingly more important as one gets older. As humans, our bodies slowly deteriorate as time progresses; therefore, as we become more and more susceptible to diseases and ailments, it's important that elderly individuals have access to the preventative care that they need despite insurance status. That involves being able to go to the physician's office regularly and detecting any early onset conditions. As the model's above suggest, elderly individual's without insurance contain a higher proportion of individual's that will absolutely refuse to visit physician's under any circumstance. Additionally, the more heterogeneous behavior indicates that a large margin of them is more likely to never visit or very frequently visit, but that also suggest that not many of them are going regular visits.

A limitation to my analysis is that the temporal scope of my data is only from 1987 to 1988. Therefore, I will not extrapolate definitive conclusions to the present as many things may have changed since then.

Although, I'd like to emphasize that though practices may have changed and healthcare may be more accessible, I believe the underlying propensity of elderly individuals will not change much even as time progresses. If healthcare for elderly individuals remain difficult and costly, then I believe the same heterogeneous behavior will occur regardless of time period. Ultimately, this reveals the importance of affordable and accessible healthcare in elderly population.

A Appendix

The website *NMES 1988* can be found at:

- https://search.r-project.org/CRAN/refmans/AER/html/NMES1988.html