Count Data Analysis of Physician Visits

Ahmed M. Abdelmaksoud

April 02, 2021

Abstract

The following document uses a combination of R Markdown and Knitr to analyze the open source data-set provided under the Ecdat (Econometrics Data) package on US Physician visits. We begin with exploratory data visualization, follow with an analysis and implementation of Ordinary Least Squares (OLS), examining whether the Gauss-Markov assumptions are satisfied, and then proceed with a modern Count Regression approach, as introduced by J.W. Tukey, to better understand the predictors and key determinants of hospitalizations. We find that a linear model built on the two-parameter negative binomial distribution, provides superior performance to either the single-parameter Poisson distribution or the classical OLS model.

1. Introduction

Recent advances in machine learning algorithms have challenged the notions of explainability and hypotheses that once drove science. Hypothesis testing, model-building and the statistical fields that undergird them have become ever-less attractive to the modern-day researcher, whose work is driven, instead, by forecasting, classification accuracy, and other now-popular metrics. One field which is particularly likely to be revolutionized by this zeitgeist is Medical Statistics. For the purposes of providing a case-study for the effectiveness of these novel methods, we employ both classical statistical algorithms and modern Machine Learning ones to study the predictors and correlates of hospitalizations.

2. Data-set and Methods

The OFP open source data-set on physician office visits, provided by R. Below is a table detailing the variable names and descriptions of the data-set. The data consist of 4406 observations and 19 variables. To access the data-set in-full outside of the R environment please refer to the University of Penn R Ecdat Library

We begin with an exploratory data analysis approach, proceed to apply OLS estimation to model hospitalizations, manually selecting features through iterative changes to the number of model parameters. Following this, we test whether each of the Gauss-Markov Assumptions hold. We find that the Gauss-Markov assumptions fail, entailing that the OLS estimates are not the Best Unbiased Estimators. By examining the ways in which the OLS assumptions are unsatisfied, we proceed to implement count data regression techniques.

Variable	Description
faminc:	family income in 10000$
hosp:	number of hospitalizations
sex:	whether the person is male
age:	age in years (divided by 10)
numchron:	number of chronic conditions
school:	number of years of education
married:	whether the person is married
employed:	whether the person is employed
emr:	number of emergency room visits
ofp:	number of physician office visits
region:	the region (noreast, midwest, west)
ofnp:	number of non-physician office visits
opp:	number of physician outpatient visits
black:	whether the person is African-American
opnp:	number of non-physician outpatient visits
medicaid:	whether the person is covered by medicaid
hlth:	self-perceived health (excellent, poor, other)
privins:	whether the person is covered by private health insurance
adldiff:	whether the person has a condition limiting daily activities

2. Exploratory Data Analysis

To probe the data-set for possible relationships, we begin by constructing and visualizing a correlation matrix of the numerical variables available. The correlation coefficient, formulated as:

$$\rho_{x,y} = \frac{Cov(x,y)}{\sqrt{Var(x)} * \sqrt{Var(y)}}$$

Is used to provide a measure, with both magnitude and direction, of the linear associations between two variables. While it may fail to capture causality in the way the conditional expectation paradigm of linear regression manages, it is often used as an initial metric for exploratory data analysis.

The resulting plot (see: **1**) indicates that, of 10 numerical variables, only moderate to low correlation coefficient values exist. The most notable being a 0.5 correlation between *emr* and *hosp* – that is, between emergency room visits and hospitalizations. Then there exits significant correlations of .3 between *adldiff* and *age*, which is to be expected. Likewise, the correlation between *school* and *faminc* is unsurprising. However, insofar as we are concerned with predictors of hospitalization, all but two numerical variables, those measuring family income and schooling, offer promise for significant relationships.

If we further examine the dat-aset by visualizing relationships between numerical and non-numerical data, as in figure **2**, we see that the number of hospitalizations a patient incurs varies by both his/her self-percieved health level and sex. Precisely, figure **2** shows the proportion of people with zero hospitalizations is lower in those who report poor health than in either those reporting either excellent or other health. Furthermore, it appears as though male's are more likely to

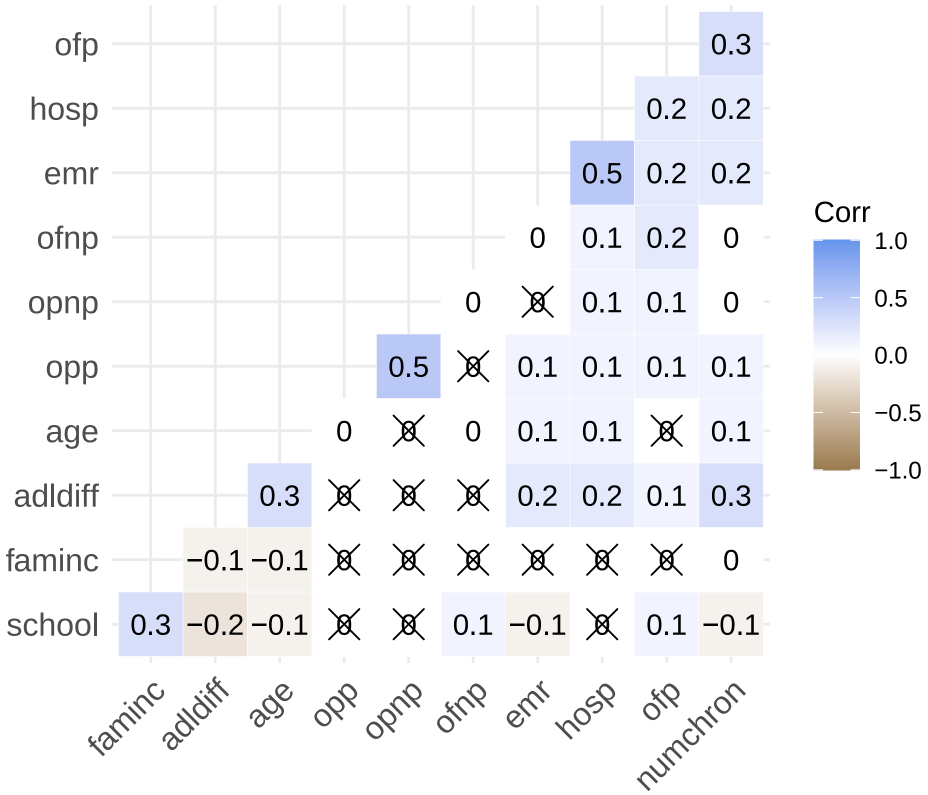

Figure 1: Correlation Matrix for OFP Data-Set

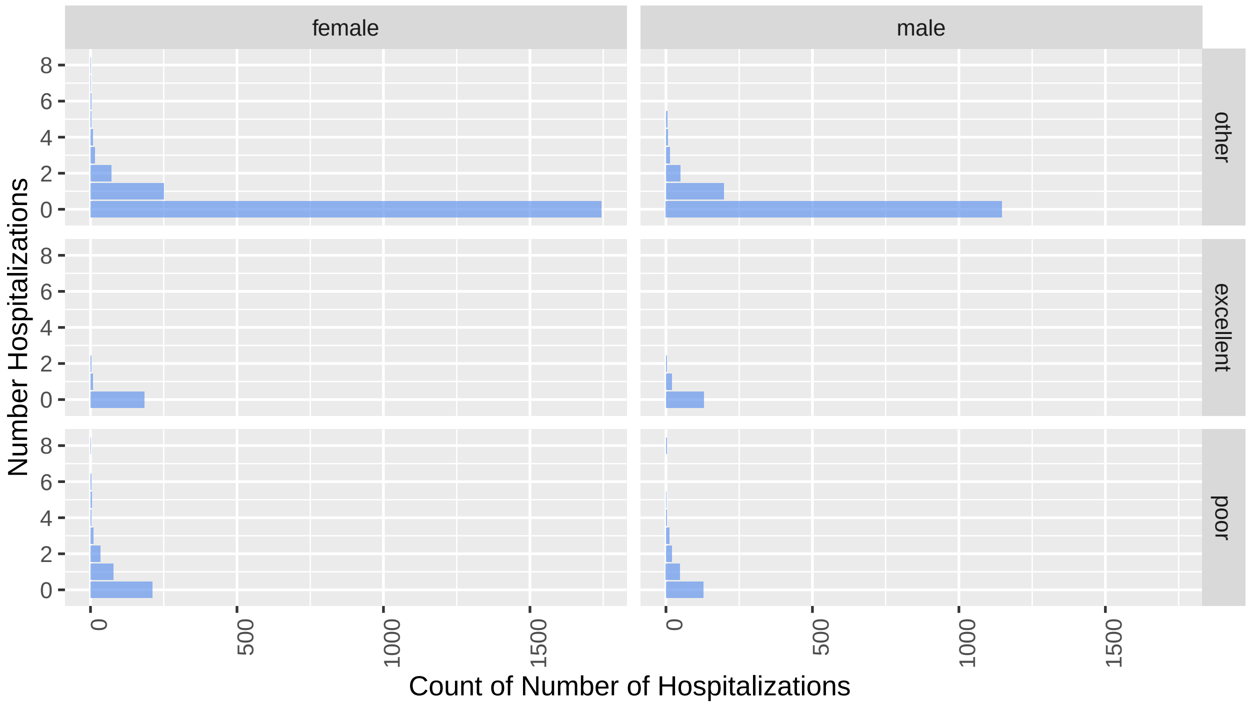

Figure 2: Number of Hospitalizations by Self-Perceived Health and Sex

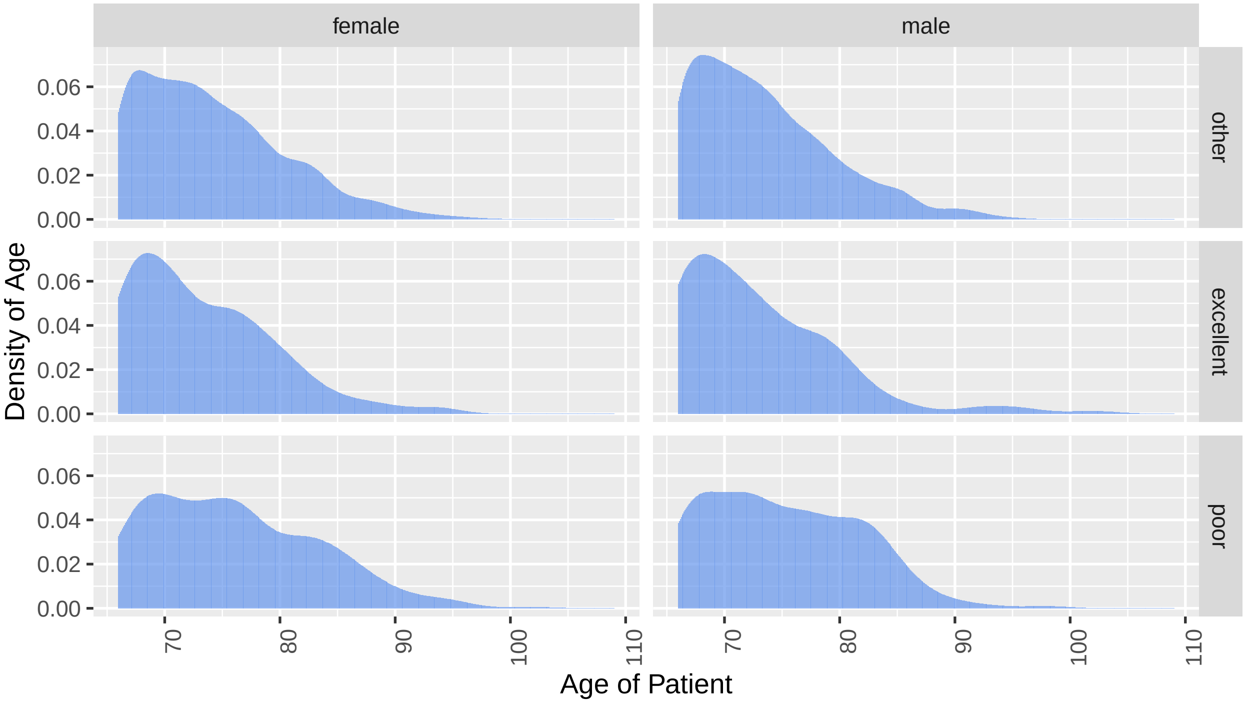

Figure 3: Density of Age by Self-Percieved Health and Sex

report excellent health and less likely to report poor health while experiencing the same number of hospitalizations as females. This, however, may not be a significant difference.

Moving forward to examine another relevant variable, if merely as a control, we observe a clear relationship in figure **3** between age, self perceived health, and sex. Once more, we parse the relationship between self-percieved health and age first. We can see, most pronouncedly that the tail of the distributions of those perceiving themselves to be poor are much longer than that of those perceiving themselves to have excellent health and, thus, as expected, those who are older tend to perceive themselves as having poorer health. Furthermore, we observe that a greater proportion of females aged 80 and above percieve themselves as having poor health, as compared to males of the same age. Conversely, a greater proportion of males 80 and above percieve themselves to have excellent health, as compared to females. This may signal a difference in temperament across the sexes that ought to be included as an interaction.

Likewise, if we examine the relationship between family income, self-percevied health, and sex we observe some evidence of differences between the sexes and, particularly saliently, an over-representation of high incomes in the group perceiving themselves to possess excellent health an a sever under representation of those with higher incomes in those perceiving themselves to have poor health. This might be an indicator that there is information to be garnered from an interaction between self-perceived health and family income.

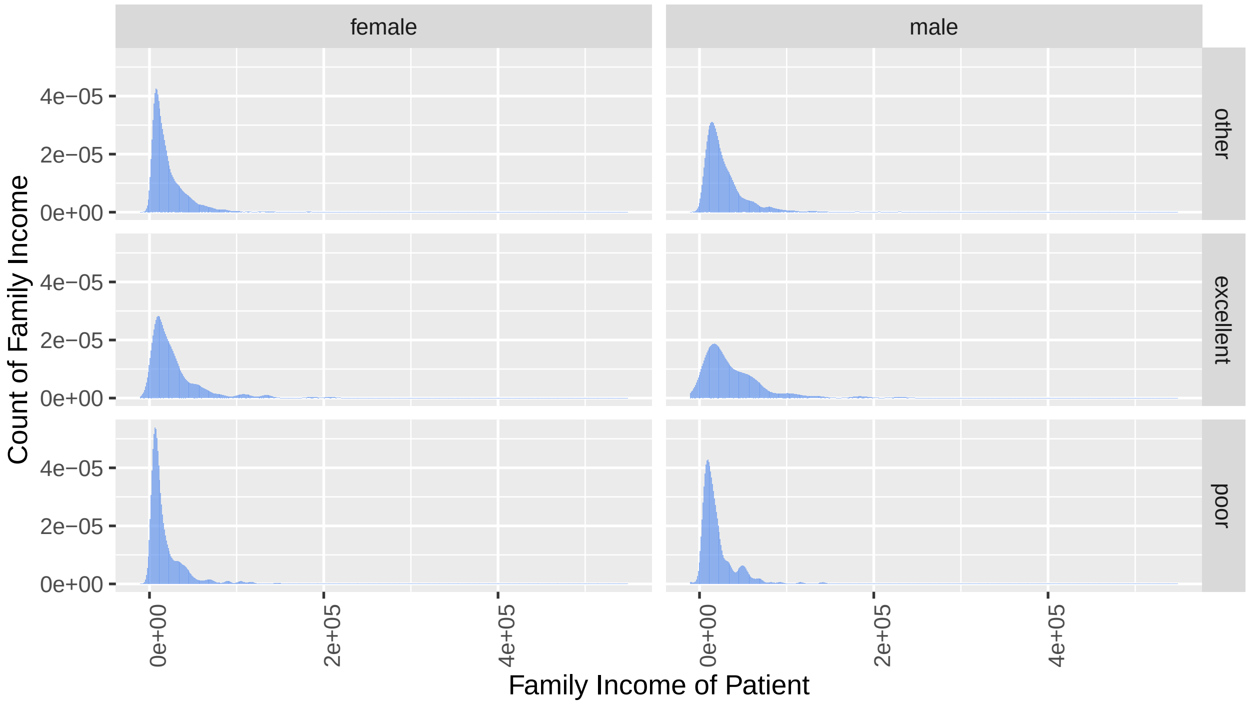

Figure 4:   count of Family Income by Self-Perceived Health and Sex

3. OLS Estimation and Checking the Gauss-Markov Assumptions

OLS as the Best Linear Unbiased Estimator (BLUE)

Ordinary Least Squares (OLS) estimation, a method for fitting a linear model to data by minimizing the average squared distance between the true value of an observation and its fitted value, has become popular on account of its ubiquity, canned implementations available in almost every computational software, and its interpretive ease. These facts notwithstanding, the effectiveness of OLS is governed by the Gauss-Markov Assumptions and the adherence of each use-case to them. The assumptions are as follows.

I. Linearity in Parameters:

That the model's parameters can be modeled as a linear combination of the following form:

$$\hat{y}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + +\beta_k x_k + \hat{u}_i$$

II. Random Sampling:

That each observation is drawn independently and from an identically distributed population, such that:

$$P(X_i = x_i | X_{i-1} = x_{i-1}) = P(X_i = x_i), \forall\ i \in\ \ n$$

III. Non-Collinearity:

That the regressors are not perfectly correlated such that:

$$|corr(x_j, x_l)| \neq 1, \forall\ j \neq l \in k$$

IV. Exogeneity:

Also known as the conditional mean assumption, the exogeneity assumption states that the expected value of the error term should be indendent of the regressors and equal to zero:

$$E(\hat{u_i}|x_1, ..., x_k) = E(\hat{u_i}) = 0$$

V. Homoskedasticity

This is a condition on the variance of the error term. Namely, it is the condition that the variance of the error term be independent of the regressors:

$$Var(\hat{u_i}|x_1, ..., x_k) = \hat{\sigma}^2$$

OLS Model Implementation and Selection

Having derived some ideas of which interactions we might wish to include in our on the basis of the above visualizations, it is vital to assess which predictors are to be included as regressors in our OLS model. For this, we employ a process of iterative feature selection, whereby, in each cycle, a previously absent feature is added to the model, the model is implemented, analyzed, and then, finally, the variable is discarded or kept depending on whether it improves our model.

To determine whether the model has been improved or not, we use the adjusted R^2, a measure of the proportion of variation in our dependent variable (hosp) explained by variation in our independent variables. Moreover, we compare each independent variable's p-value against a significance level of .05 to determine whether its influence on the conditional expectation of hospitalizations is significant.

After 15 rounds we find that the most effective model, accounting for approximately 27% of the variation in hosptializations is one including the following variables: *ofp*, *opp*, *emr*, *numchron*, *adldiff*, *age*, *sex*, *faminc*, *hlth.* Furthermore, we find that running interactions between hlth and age, as well as hlth and faminc produces significant results and augment's the model's explanatory power. The results of this final model, alongside the raw model, are shown below.

Model Results and Interpretation

The resulting output of the two models, one including and the other excluding interaction variables is as follows. Most notably, we observe that the Adj. R^2 value of the model including the interactions between *hlth* and *faminc* and between *hlth* and *age* are is .28, as compared with the .27 of the model without the interaction. Furthermore, the former interaction term is significant; It appears that a 10000$ increase in family income entails a 0.06 increase in the number of hospitalizations an individual incurs if she perceives her health as poor. This might imply that people become more sensitive to disease and ailment when they are wealthier. Aside from this difference, much of what the two models, OLS Model 0 and OLS Model 1 is the same; thus, for brevity, I shall proceed only with an interpretation of OLS Model 1.

	OLS Model 0	OLS Model 1
(Intercept)	-0.30^{*}	-0.38^{**}
	(0.12)	(0.13)
Office Physician Visits	0.02^{***}	0.02^{***}
	(0.00)	(0.00)
Outpatient Physician Visit	0.01^{***}	0.01^{***}
	(0.00)	(0.00)
Emergency Visits	0.45^{***}	0.44^{***}
	(0.01)	(0.01)
Chronic Illnesses	0.05^{***}	0.04^{***}
	(0.01)	(0.01)
Debilitating Disase?	0.08^{**}	0.05^{*}
	(0.03)	(0.03)
Age	0.04^{*}	0.05^{**}
	(0.02)	(0.02)
Male?	0.05^{*}	0.05^{*}
	(0.02)	(0.02)
Family Income	0.00	-0.00
	(0.00)	(0.00)
S-P Hlth: Excellent		0.37
		(0.43)
S-P Health: Poor		0.50
		(0.33)
Age*Excellent S-P Hlth		-0.05
		(0.06)
Age*Poor S-P Health		-0.06
		(0.04)
Family Income*Excellent S-P Hlth		0.00
		(0.01)
Family Income*Poor S-P Hlth		0.06^{***}
		(0.01)
R^2	0.27	0.28
Adj. R^2	0.27	0.28
Num. obs.	4406	4406

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: OLS Statistical models

Somewhat unsurprisingly, both office physician visits and outpatient physician visits appear to be significant predictors of nuumber of hospitalizations; a unit increase in the number of office visits to physician and number of outpatient visits to physicians increases the number of hospitalizations by 0.02 and 0.01, respectively. A mild increase but, across years this may very well compound. The effect of Emergency room visits are both significant and greater in magnitude than that of any other regressor. An additional visit to the emergency room increases the number of hospitalizations for an individual by 0.44. this is an order of magnitude greater than the effect of number of chronic illnesses on hospitalizations. The effect of an additional chronic illness is to increase the number of hospitalizations an individual experiences by .04.

The aforementioned variables have been significant at 0.001 level, whereas the effects of age and being male is are only significant at the .01 and .05 levels, respectively. Each of an increase in a patient's age by one year and being male yields a .05 increase in the number of hospitalizations. The increase in hospitalizations due to maleness might be a function of males being over-represented in risky jobs and risk-loving temperaments.

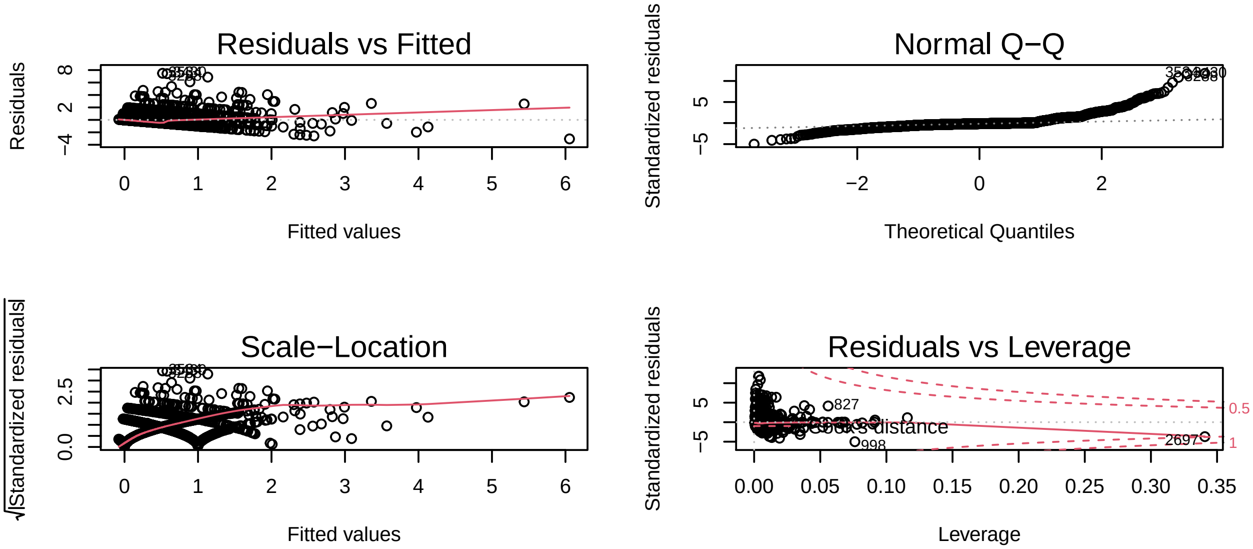

Figure 5: OLS Model Analysis Plots

Gauss-Markov Assumption Check

We check Assumption I: Linearity in Parameters by examining the first plot in figure **5**, which is of the model residuals against the fitted values of the regression. The appropriateness of linearity is attested to by the degree to which the data can be approximately modeled by a straight line. As we can see, in this case, linearity seems appropriate but for an initial curve which complicates the trajectories of the data. Since the graph's departure from linearity is mild, we can proceed to examine the rest of the assumptions.

The second plot in our figure, the Normal Q-Q Plot (see: **5**), suggests that there is severe departure, on the part of our data from normality. The fact that our residuals do not, following normality, cluster in the middle and distribute flatly at the tails, means that our employed Gaussian model is inappropriate and, what is more, OLS is not Unbiased, let alone the Best Unbiased Linear Estimator (BLUE).

Assumption V of the Gauss-Markov Assumpitons, that of homoskedastic errors, is not to be merely examined visually. Nevertheless, the Scale-Location plot in figure **5** suggests far from uniformly distributed residuals across observations. In fact, it is clear that the degree to which observations vary around the model's fitted values does itself vary. This will be examined further below.

Another concerning facet of our model, albeit not strictly related to the Gauss-Markov Assumptions, of which endogeneity is yet to be explored, are the pronounced effects of outliers that seem to ail the OLS model estimated above. In figure **5**, our fourth plot uses Cook's Distances, formulated as:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y_j} - y_{\widehat{j(i)}})^2}{(p+1)\hat{\sigma}^2}$$

to visualize the degree to which our model would change were we to remove the outliers. The fact that there are several outliers crossing the red lines, which constitute a threshold for determining the magnitude of outlier effects, indicates that our model is not robust.

Testing for Heteroskedasticity

Homoskedasticity Tests	**Breusch-Pagan Test**	**White's Test**
p-value:	2.2e-16	2.013312e-30

We further investigate assumption V of the Gauss-Markov Assumptions, to examine the suitability and whether Ordinary Least Squares method is BLUE for the purposes of our analysis. To do this, we employ both a classical Breush-Pagan Test of the following form and then a more advanced White-Test for Homoskedasticity. The latter is used, in particular, to detect non-linear correlations between the variance of the error term and our model independent variables.

In both cases, the hypothses being tested are the following:

$$H_0 : Var(\hat{u_i}|x_1, ..., x_k) = \sigma^2, \ \forall \ x \in \ X$$

where X is the vector of independent variables in our model.

The result of the two tests converge; that is, on accounts of p-values lower than 0.01, we can reject the null hypothesis that our errors are not homoskedastic.

$$H_1 : Var(\hat{u_i}|x_1, ..., x_k) = \sigma_i^2, \ for \ at \ least \ one \ x \in \ X$$

4. Count Data Regressions

The inapporpriatenesss of OLS on several accounts as has been demonstrated above demands that we find an alternative method for estimating parameters of our models. If it were the case that OLS was merely inapporpriate due to heteroskedasticity or endogeneity, we might try a gaussian generalized linear model, estimated by Maximum Likelihood, and perhaps augmented by robust standard errors. However, we know the problems with our data are deeper than mere heteroskedasticity. Namely, that our data significantly depart from normality, as attested to by the QQ plot of our residuals.

To this, the field of count data regression provides an answer which might be insightful. In fact, multiple such answers are available. Our data, which is positively skewed, as count data is, is likely better fitted by a count model. One such model is the Poisson model of the following form:

$$\hat{y} = e^{\hat{\beta_0} + \hat{\beta_1} x_1 + \hat{\beta_2} x_2 + ... + \hat{\beta_k} x_k + \hat{u_i}}$$

or, alternatively:

$$log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k + \hat{u}_i$$

This model is estimated via Maximum Likelihood, which involves solving the following optimization problem and finding the stationary points of the likelihood function.

$$\max_{\theta} \; l(\theta \mid X, Y) = \sum_{i=1}^{m} (y_i \theta' x_i - e^{\theta' xi})$$

$$\frac{\partial \; l(\theta \mid X, Y)}{\partial \; \theta} = 0$$

where, θ' is the vector of parameters whose values are to be solved for, X is the matrix of independent variables and Y is the vector of the dependent variable.

Another model, which is often used to accommodate problems of over-dispersion that are often encountered due to the single-parameter nature of the Poisson distribution, is the negative binomial model. The model is similar in form but has the unique advantage of having an additional free parameter, in the form of the variance, over the Poisson model.

The Poisson and Negative Binomial models are clearly superior fits to the data, as indicated by the models' lower Akaike Information Criterion (AIC) and Bayesian Information Criterion, which provide measures of goodness-of-fit and are formulated as:

$$AIC = 2k - 2log(\hat{L})$$

Where k is the number of estimated parameters in the model and $\hat{L}$ is the maximum value of the likelihood function for the model.

$$BIC = -2 * log(likelihood) + log(N) * k$$

Where N is the number of observations and k is as above.

Whereas the AIC of the Gaussian linear model, estimated by Maximum Likelihood, is 8517, the Poisson model's AIC is 5574, clearly a significant improvement. Nevertheless, the Negative Binomial model appears to provide a slightly better fit, having an AIC of 5236 and a BIC of 5338 compared to the 5670 of the Poisson model. A closer inspection of the models, specifically through rootograms (see **6**), indicates there is an apparent difference in the models' fit. Overdispersion, a problem arising when a model is unable to accomodate the amount of variability within a dataset, is a common problem when using single-parameter distributions, such as the Poisson distribution, to fit data. Thus, it is likely the Poisson model is suffering from over-dispersion and a significant benefit is garnered from using a two-parameter model, such as the negative binomial model,instead. To quantify this difference, we conduct a likelihood ratio test to determine whether the Poisson Model is suffering from a significant amount of over-dispersion and under-predicting tail values while over-predicting medium-range ones, as a result.

To run a log-likelihood test we compute the following test statistic with degrees of freedom equal to the difference between those of our complex model, the Negative Binomial, and of our nested model, the Poisson.

$$LR = -2 * [log(likelihood_{nested}) - log(likelihood_{complex})]$$

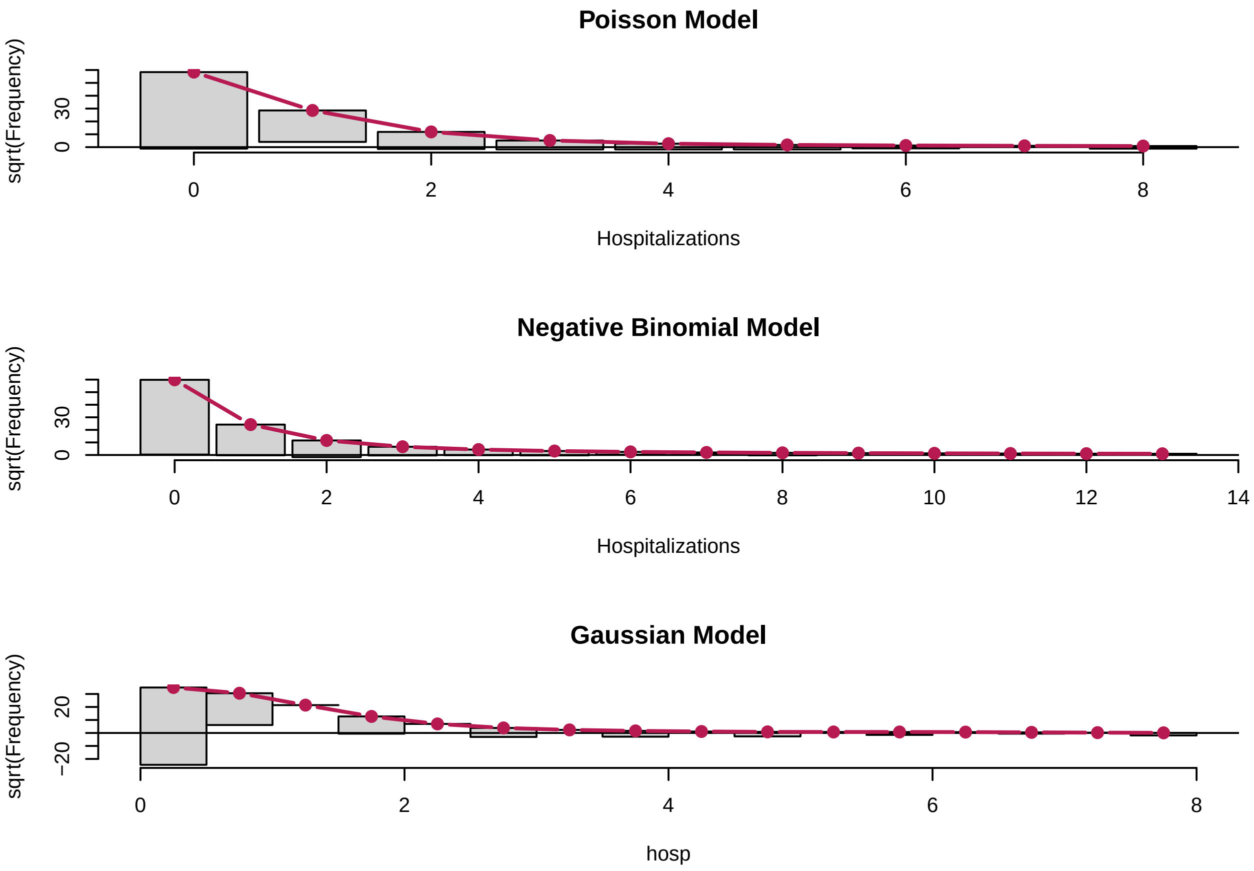

Figure 6: Rootograms of Count Regressions

	OLS	Poisson	Negative	Gaussian
(Intercept)	-0.38^{**}	-3.92^{***}	-4.15^{***}	-0.38^{**}
	(0.13)	(0.39)	(0.48)	(0.13)
Office Physician Visits	0.02^{***}	0.03^{***}	0.04^{***}	0.02^{***}
	(0.00)	(0.00)	(0.00)	(0.00)
Outpatient Physician Visit	0.01^{***}	0.02^{***}	0.03^{***}	0.01^{***}
	(0.00)	(0.00)	(0.01)	(0.00)
Emergency Visits	0.44^{***}	0.30^{***}	0.70^{***}	0.44^{***}
	(0.01)	(0.01)	(0.03)	(0.01)
Chronic Illnesses	0.04^{***}	0.17^{***}	0.15^{***}	0.04^{***}
	(0.01)	(0.02)	(0.03)	(0.01)
Debilitating Disase?	0.05^{*}	0.20^{**}	0.15	0.05^{*}
	(0.03)	(0.07)	(0.09)	(0.03)
Age	0.05^{**}	0.25^{***}	0.24^{***}	0.05^{**}
	(0.02)	(0.05)	(0.06)	(0.02)
Male?	0.05^{*}	0.14^{*}	0.21^{**}	0.05^{*}
	(0.02)	(0.06)	(0.07)	(0.02)
Family Income	0.37	1.60	1.23	0.37
	(0.43)	(2.08)	(2.26)	(0.43)
S-P Hlth: Excellent	0.50	1.64^{*}	0.40	0.50
	(0.33)	(0.70)	(0.96)	(0.33)
S-P Health: Poor	-0.00	-0.01	-0.02	-0.00
	(0.00)	(0.01)	(0.02)	(0.00)
Age|Excellent S-P Hlth	-0.05	-0.30	-0.24	-0.05
	(0.06)	(0.28)	(0.30)	(0.06)
Age|Poor S-P Health	-0.06	-0.21^{*}	-0.06	-0.06
	(0.04)	(0.09)	(0.13)	(0.04)
Family Income|Excellent S-P Hlth	0.00	-0.00	0.01	0.00
	(0.01)	(0.05)	(0.06)	(0.01)
Family Income|Poor S-P Hlth	0.06^{***}	0.08^{**}	0.12^{***}	0.06^{***}
	(0.01)	(0.03)	(0.04)	(0.01)
R^2	0.28			
Adj. R^2	0.28			
Num. obs.	4406	4406	4406	4406
AIC		5574.16	5236.15	8517.30
BIC		5670.02	5338.41	8619.55
Log Likelihood		-2772.08	-2602.08	-4242.65
Deviance		3597.61	2414.34	1769.93

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: Final Statistical Models

We then use the degrees of freedom and test statistic to find the corresponding p-value on a chi-squared distribution. The result is a p-value of 3.172041e-46, allowing us to reject the Null Hypothesis that there is no difference between the two models, in favor of the conclusion that the

Negative Binomial model is a superior fit. It is likely that the relaxation of the assumption inherent in the Poisson model – namely, that the variance and mean are equivalent – that results from a shift to the Negative Binomial model, is what ameliorates the over-dispersion problem.

For our final negative binomial model, in order to interpret the magnitude of each regressor's effect on the number of hospitalizations, we must first exponentiate the coefficients and take that value away from 1. This, ultimately produces far better measures of effect sizes, as it provides contextual information about the proportion of the effects rather than merely the raw sizes.

The first thing we observe, most notably is that, in our final model, contrary to in our OLS or Poisson model, having a debilitating disease does not seem to be a driving factor in the number of hospitalizations, when other variables are controlled for. Furthermore, being male is now singificant at .01% level and a unit increase in age is significant .001% level.

This means that a unit change in the number of physician office visits increases the number of hospitalizations by approximately 4% upward. Furthermore, an increase by one unit in the number of outpatient physician visits increases it hospitalizations by 3%. Most notably, however an additional visit to the emergency room seems to entail a 101% increase in the number of hospitalizations. Thus, as was the case with the OLS mode, this is an especially indicative predictor. We can also observe that each of a unit increase in number of chronic illnesses and an additional year of age increases the number of hospitalizations by 16%. Being male seems to also be associated with a 21% increase in the number of hospitalizations. In the final instance, we see that an increase in family income by 1000$ induces an approximately 13% increase in hospitalizations if you perceive yourself to have poor health.

5. Conclusion

To conclude, it is clear from the extensive analysis above that the count regression models provide a significantly better fit to the data, with its non-normal distribution, than either the OLS model or the Generalized Linear Model with Gaussian parameterization. Furthermore, among count regression models, it is the negative binomial model, with the versatility afforded by its two-parameter distribution, that can best fit the data and avoid over-dispersion. Thus, from our resulting final model we can conclude that the best predictors of hospitalizations, which follow a negative binomial distribution are emergency room visits, poor self-perceived health interacted with family income, and debilitating diseases which limit daily movement. We hope these factors can provide a basis for predicting and anticipating individuals most at risk for hospitalization and that a broader study of these trends, likely through time series data, is conducted.