220B Final Project Report –

Application of General Linear Model on Consultation Data

Xiaoran Ma

9841479

03/20/2020

Summary

This report summarizes the result of analyzing the data regarding individual consultations with pharmacists in San Francisco. Possible relations among the number of consultations and other variables are found. Diagnostics, inference and predictions are conducted through Generalized Linear Models (GLM). Under some assumptions, explanatory variables: sex, income, number of illness, self-reported days of reduced activity, health score and their interactions have certain relations with the dependent variable, number of consultations with pharmacists. And Poisson Regression turns out to fit the data well. Due to the special structure of the data, simulation and bootstrapping methods are also implemented in supporting the analysis and diagnostics.

The rest of this report is organized as follows: Part I discusses data collection, data integrity and assumptions on our data. Part II uses exploratory data analysis (EDA) to find out the pattern and relations among different variables. Part III focuses on building Generalized Linear Models (including assumptions on models and diagonalizations) to explain those relations quantitatively and makes inference and predictions based on the model. The last part is the conclusion and contains some possible improvements to this study.

Key words: Generalized Linear Models, Poisson Regression, Bootstrapping

Part I Data Collection, Integrity and Assumptions

1.1 Variables used in this study

Table 1.1 : Variables Used and their meanings

pc	Number of consultations with a pharmacist in the past 4 weeks
sex	1 if female, 0 if male
age	Age in years divided by 100
income	Annual income in US dollars divided by 100000
lp	1 if covered by private health insurance with pharmacy coverage, 0 otherwise
fp	1 if covered by private health insurance without pharmacy coverage, 0 otherwise
fr	1 if not covered by private health insurance, 0 otherwise
ill	Number of illnesses in the past 4 weeks (5 or more illnesses are coded as 5 in this dataset)
ad	Number of self-reported days of reduced activity in the past 4 weeks due to illness or injury
hs	General health questionnaire score calculated by the investigators. A high score indicates poor health.
ch1	1 if individual has chronic medical condition(s) but is not limited in activity, 0 otherwise
ch2	1 if individual has chronic medical condition(s) that limit individual's activity, 0 otherwise

1.2 Data Collection and Integrity

Data Integrity refers to the maintenance of, and the assurance of the accuracy and consistency of data over its entire lifecycle. Possible ways that may compromise the data accuracy and consistency during data collection are as follows:

a) Collect data over a long period of time. If the data is collected over a long term, we may need to include time as another independent variable. For example, some seasonal disease like flu may occurs in one period but not the other which increases the needs of consultation.

b) Collect data in some specific locations. Data is better collected across the entire city. Collecting data in some

predetermined location may result in inaccuracy. One possible example is that people in the same area may consult at the same pharmacy and these pharmacists may suggest a more frequent visit than other pharmacist at other locations.

c) The quality of the data may not be guaranteed since the survey may be conducted by many different staffs. Different attitudes could result in an inconsistent quality of the data.

d) Dependent data. Later in our GLM analysis, we need our observations independent of each other. But it is possible that our data is collected in a way such that they are not independent. For example, each pharmacy has different consultation regulation. If our responders happen to go to the same pharmacy, the number of their visits may be related. Also, families may go to the pharmacy together. If the data is collected on people in the same family, those data may be highly correlated.

1.3 Assumptions on Data

This report makes the following assumptions according to the possible issues described above:

a) Data was collected over a short period of time and across the entire city. Time periods in each day and locations are randomly assigned so that the data could represent the entire population well.

b) Our staffs are well trained and the data they collected has high quality. Later in part II, we will see that there is an issue related to this part. And possible remedies are implemented there.

c) Data is independent of each other. This is an important assumption for later data analysis.

Part II Exploratory Data Analysis (EDA)

2.1 Data Reformulation

In this section, we convert some of the variables to a new form. They are described in Table 2.2. The followings are reasons to do this.

Table 2.1 Variables after reformulation

pc	Number of consultations with a pharmacist in the past 4 weeks
sex	1 if female, 0 if male
age	Age in years divided by 100
income	Annual income in US dollars divided by 100000
insurance	1 if covered by private health insurance with pharmacy coverage
	2 if covered by private health insurance without pharmacy coverage
	3 if not covered by private health insurance
	4 if recorded 0 in all three variables / unspecified insurance type
ill	Number of illnesses in the past 4 weeks (5 or more illnesses are coded as 5 in this dataset)
ad	0 if 0 days of reduced activity in the past 4 weeks due to illness or injury
	1 if more than 0 days of reduced activity in the past 4 weeks due to illness or injury
hs	General health questionnaire score calculated by the investigators. A high score indicates poor health.
ch	1 if individual has chronic medical condition(s) but is not limited in activity
	2 if individual has chronic medical condition(s) that limit individual's activity
	3 if individual does not have chronic medical condition(s) (0 in other two variables)

In section 1.3 part (b), we mentioned an issue in collecting data. All the variables introduced in section 1.1 are easy to understand but two sets of them requires further justification.

The first set of variables are lp, fp and fr[1]. The second set of variables are ch1 and ch2. They deserve careful justification because they partition some sample spaces. The first set partition the whole sample space and the second set partition the sample space conditioning on having chronic medical condition(s). The illustrations are in graph 2.1 and 2.2.

The first set of three variables forms a partition of the entire sample space since if a respondent is recorded 1 in any one of these three variables, the other two variables for this respondent must be 0. Also, each respondent has to be in at least one of these three parts. The second set of variables forms a partition on sample space condition on having chronic medical conditions since if a person has chronic medical conditions, he/she must be in either one of these two parts.

lp = 1: covered by private health insurance with pharmacy coverage	fr = 1: not covered by private health insurance,
fp = 1: covered by private health insurance without pharmacy coverage	

ch1 = 1: individual has chronic medical condition(s) but is not limited in activity	ch2 = 1: individual has chronic medical condition(s) that limit individual's activity

Graph 2.2 Demonstration on Variables: the square represents the sample space

However, the problem occurs when we have respondent who is recorded 0 in all three variables in the first set. Our data contains 156 observations who reply 0 to all lp, fp and fr variables. This type of response should not occur. One possible explanation is that we did not have data on these three variables at all because our respondent refused to provide this information.

Since we have a large proportion of data having the same problem, we do not want to simply remove all of them, which leads to a loss of information. Instead, one possible remedy is to combine the variable lp, fp and fr into a categorical variable (in table 2.2) and create a new category for those having 0 in all three of them. We could interpret the new category as unspecified insurance type.

Having 0 in both ch1 and ch2 variable is possible in reality. A person without chronic medical condition(s) has this response. For exploratory analysis purpose, I also covert ch1 and ch2 variables to one categorical variable called ch (table 2.2).

For variable ad, I convert it to a variable with only 0-1 values. 0 means 0 days of self-reported reduced activity and 1 means more than 0 days of self-reported reduced activity. This is because only 15.4% of the value in this variable is non-zero in a large range (from 1 to 14). This skewness may lead to outlier points in our later work[2] and converting such variable to 0-1 is frequently used.

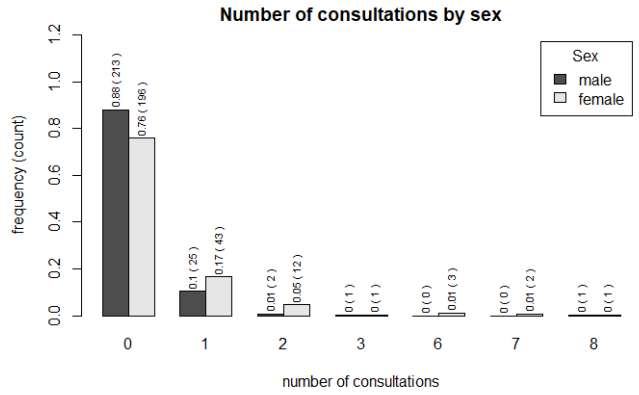
We have checked all other variables and did not find any other problems with our variables[3].

[1] Please refer to section 1.1 for variable meanings.
[2] This will be explained more in Part III.
[3] We also checked whether there exist responses that record multiple 1's in the set lp, fp and fr as well as the set ch1 and ch2. And the answer is no which is good.

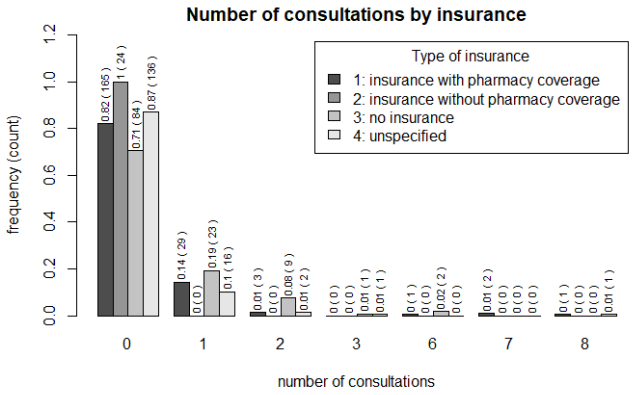
2.2 Visualization -- dependent variable versus explanatory variable

(Reason we need visualization) In practical data analysis, it is preferred to visualize the relationship among explanatory variables and dependent variable before building our models. The reason is that plots and graphs make it easier to get intuition behind the data. Also, it could help us reveal the possible relations among them.

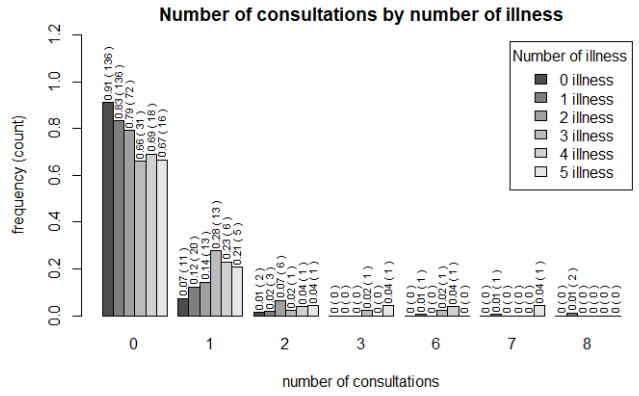
(Use bar plot in our dataset) In our dataset, since we have different data types, we need to consider a good way of visualization, which could capture the pattern of data and will not loss much information. Our dependent variable is count data. It makes visualization clearer to use a bar plot for count data versus count data or count data versus categorical data. For count data versus continuous data, we could use regular dot plot. However, regular dot plot suffers from the weakness that duplication points are overlaid on each other which makes it hard to find the general pattern or tell how many points in total[4]. The way we did is to split our continuous variable into different segments and regard it as a count data. Take income variable for example, we split it evenly into three parts and count how many observations in each part. This adjustment make income variable performs like a count data. Then we use bar plot to find out intuitions between it and our dependent variable. All the plots are shown in Graph 2.3 (a) ~ (h).



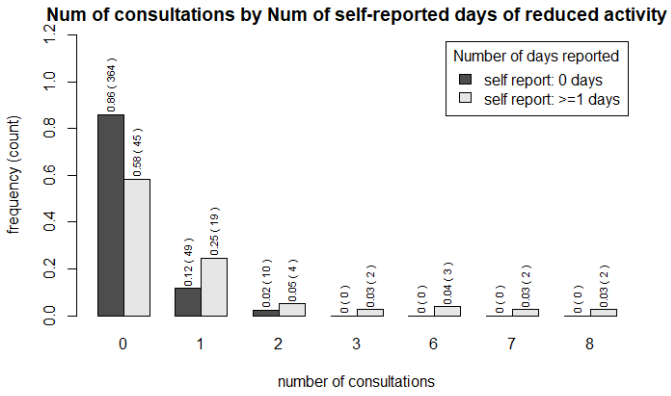
Graph 2.3(a)



Graph 2.3(b)

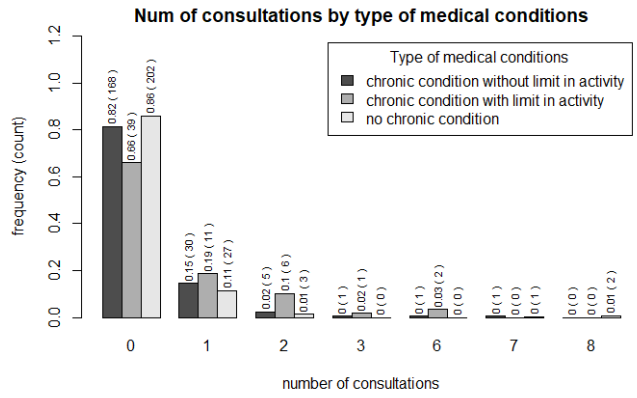


Graph 2.3(c)

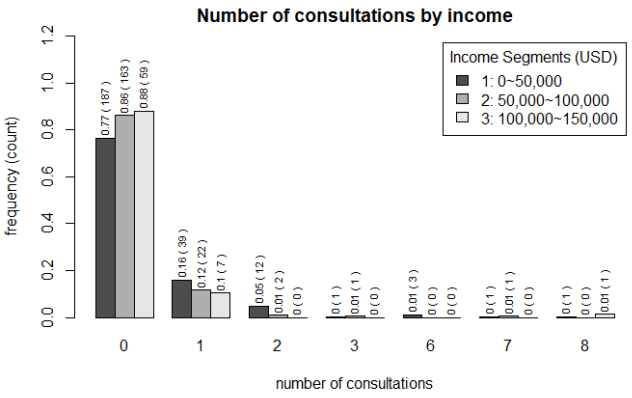


Graph 2.3(d)

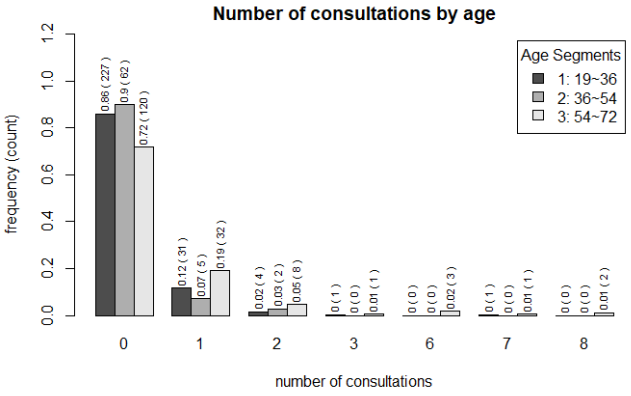
[4] There are ways to separate duplications a little bit on the plot to differentiate from each other. But when we have too many observations taking the same value, this plot could get really messy.



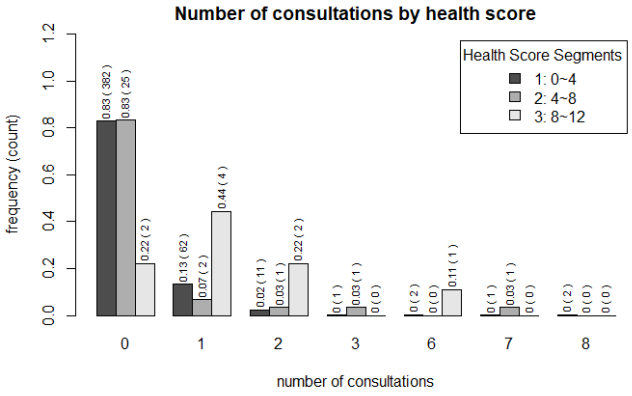
Graph 2.3(e)



Graph 2.3(f)



Graph 2.3(g)



Graph 2.3(h)

Graph 2.3 Visualization between dependent variable and each explanatory variable (for the legend of segments, inclusive on the left and exclusive on the right for the first 2 segments. For the last segment, inclusive on both sides)

(Understanding the bar plots) All above plots are formulated in the following way. x-axis represents different values of number of consultations (dependent variable). y-axis contains both frequency and counts (in the parenthesis). The height of each bar represents the proportion (frequency) of observations having that number of consultations in that explanatory variable segment (or category). Using frequency instead of count as the height of each bar makes more sense because the segment (or category) with a larger population will boost the height of its bars which make the comparison among segments (or categories) meaningless. Take Graph 2.3 (h) for example, the darkest bar standing at 0 has value 0.83 (382). This means that 83% observations in the segment 1 (health score between 0 and 4) have 0 consultations in the past 4 weeks. The corresponding number of people is 382.

(Information provided by bar plots) From these bar plots, we could get information about whether the variation in an explanatory variable will influence the value of dependent variable[5]. In general, if the bars are at the same (or similar) height in all segments (or categories), this means that the variation in this explanatory variable does not change the value of dependent variable much (or the value of this explanatory variable does not change at the same time as the dependent variable). In contrast, fluctuate height in different segments (categories) at the same value of dependent variable indicates that the change of value in this explanatory variable may result in a change in dependent variable.

With the above reasoning, we can see that the variation between segments at the same number of consultations are not very evident for variable income and age. They themselves alone[6] may not explain our dependent variable much. All other variables seem to have some influence on the dependent variable.

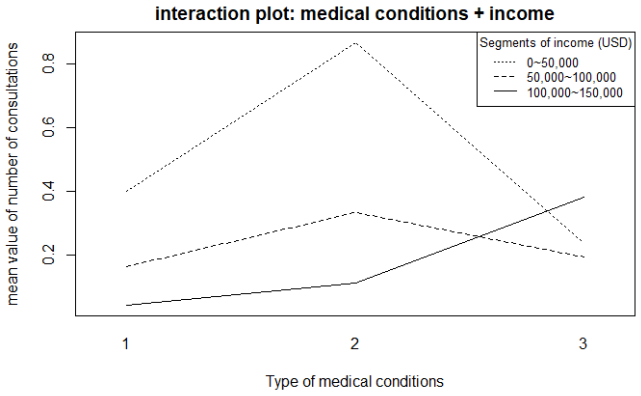
[5] Here we need to be careful about "correlation does not mean causation". But throughout this report, we use this relation loosely.

[6] They may also contribute to the variation of dependent variable through interactions with other variables.

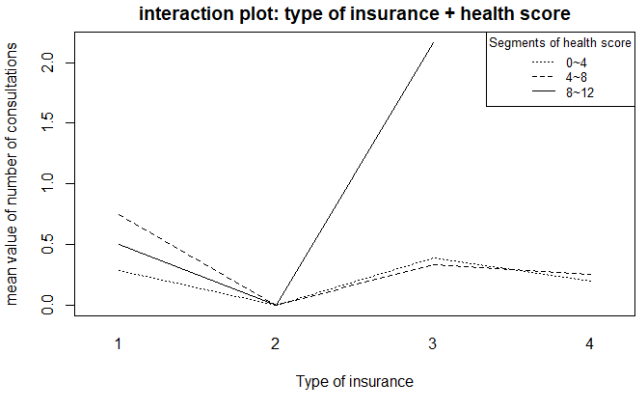
2.3 Visualization -- dependent variable versus pair of explanatory variables

(Reason we need interaction plots) As mentioned in the above section, some explanatory variables may not explain much on the dependent variables if we only consider their main effect. In other words, these variables could influence our dependent variable through interaction with other explanatory variables.

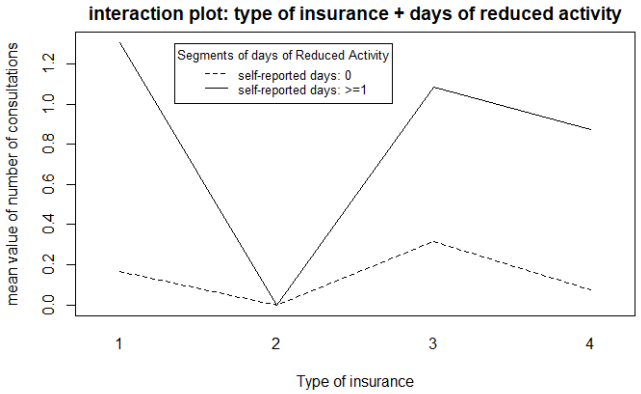
(Interaction plots for our dataset) For interaction plots, we are considering effect on dependent variable through more than one explanatory variable. In practice, we usually consider no more than two variables at the same time. For higher order interactions, it could be hard to interpret and may not be meaningful. The following are interaction plots for our dataset (Graph 2.4(a)~(d)). Here I only list 4 of them. A complete interaction result from visualization is available in the code, but we prefer not listing all of them.



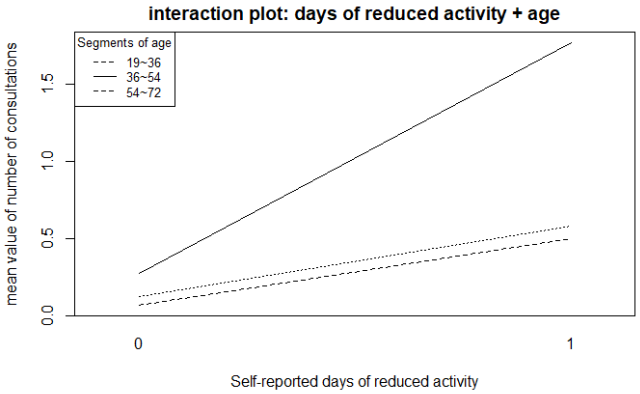
Graph 2.4(a)



Graph 2.4(b)



Graph 2.4(c)



Graph 2.4(d)

Graph 2.4 Some Interaction plots (for the legend of segments, inclusive on the left and exclusive on the right for the first 2 segments. For the last segment, inclusive on both sides)

(Understanding the interaction plots) Here we only list 4 interaction plots as an illustration. Each line in the plot represents the "trend" of the segment specified in the legend according to another categorical (or count) variable with segments specified on x-axis. Take Graph 2.4(d) for example, the solid line is drawn by linking two end points on it. The lower end point represents the average number of consultations for those people who have 0 self-reported days of reduced activity and age between 36~54. The upper end point represents the average number of consultations for those people who have at least 1 self-reported days of reduced activity and age between 36~54.

In general, if two different lines have an intersection, it may indicate that there is an interaction effect between these two explanatory variables. The reason is that, if there is no interaction between two explanatory variables, then fix the value of one explanatory variable, the value of dependent variable is supposed to be monotonically increasing or decreasing. Therefore, no intersection should occur among lines. From the four plots we have, they indicate that there may be interactions between type of medical conditions and income as well as type of insurance and health score. And the interaction between type of insurance and days of reduced activity or age and days of reduced activity are not significant.

Part III Generalized Linear Model – Poisson Regression

With the exploratory analysis in the above section, we have some possible relations among variables in mind. In this section, we will find out a generalized linear model to fit our data and use it to do inference and prediction. The relation between Part II and Part III can be regarded as qualitative and quantitative analysis. Generalized linear models are like linear models but with an extension to allow more flexibility in our dependent variable.

3.1 Model Assumptions and structure

(Reason why we prefer GLM to LM) Regression models aim at finding out the relations among dependent variable and explanatory variables. Although linear regression has many desirable properties, it assumes our dependent variable taking continuous values from negative infinity to positive infinity, which is not the case in our dataset. We have count data which only takes discrete values from zero to infinity.

(Poisson Regression) One model that meets the assumption about our dependent variable is Poisson regression model. It also has the following assumptions and structure.

Let Y_i be the Number of consultations with a pharmacist in the past 4 weeks for the i-th people, for $i \in \{1, \dots, 500\}$.
Let $\mu_i = E(Y_i)$.

We have three components:
Random Component:

$$Y_i \overset{indep}{\sim} Poisson(\mu_i) \quad \text{for} \quad i \in 1, \dots, 500 \quad \text{with} \quad \mu_i = E(Y_i)$$

Systematic Component:

$$\eta_i = X_i{}^T \beta = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{i1}X_{i2}\beta_{12} + \dots \tag{1}$$

Link Function:

$$g(\mu_i) = ln(\mu_i) = \eta_i$$

In this setting, random component specifies that our dependent variable can only take discrete value from 0 to infinity by the property of Poisson distribution. Also, we assume each of them are independent and has a different mean value μ_i. Systematic and link function specify the form of regression. Similar to linear regression, here, we regress the natural logarithm of the mean of our dependent variable on linear combination of explanatory variables. The reason we regress the natural logarithm of the mean instead of the mean itself is because the linear combination of explanatory variables can take values from negative infinity to positive infinity, but the mean of Poisson distribution only allows positive values. And natural logarithm maps the entire real line into the positive numbers[7]. Note that in the systematic component, we first use vector form and then write out the equivalent expansion. It should be clear that the X vector also contains terms of interactions. The other assumption on this model is that the systematic component is linear. In other words, linear in terms of β's.

[7] Other possible link functions are available for Poisson regression. But the natural logarithm is usually preferred because of its ease on interpretation and computation.

3.2 Poisson Regression Model for our dataset

In this section, we propose a Poisson Regression Model based on Exploratory Data Analysis (EDA). We also conducted diagnostic and variable selections based on this model.

(General procedures in finding a model) In practice, we build a model with following steps:

Step 1: postulate general class of models

Step 2: identify model to be tentatively entertained

Step 3: fit this model

step 4: tune this model

Step 5: diagnostics on the model. If the model is adequate, do step 6. If not, go back to step 2.

Step 6: Inference and Predictions

(***Step 1, 2 and 3***) We have finished step 1. For step 2, using the result from EDA, since there are possible interactions from the interaction plots, we want a model with interaction terms. Now, we come to step 3. There are many possible models containing interaction terms and we need a criterion to compare them.

The criterion we used is called AIC (Akaike information criterion):

$$AIC = D_M + 2p\phi$$

where D_M is deviance of our current model M. D_M is a goodness-of-fit measure, the smaller it is, the better our model fit the data. p is the number of explanatory variables in the model. ϕ is the dispersion parameter. In Poisson regression, $\phi = 1$.

We want to choose the model with the smallest AIC because we want a model that both fits the data well (smaller D_M) and contains as fewer explanatory variables as possible. The reason is that we want to balance between Bias and Variance (in the sense of data science). In other words, if the model has too many explanatory variables, it tends to overfitting, capturing the noise along with the underlying pattern in data (low bias and high variance). If the model has too few explanatory variables, it may be underfitting, unable to capture the underlying pattern of the data (high bias and low variance). Since our dataset is relatively small (500 observations), too many covariates may lead to overfitting.

(step 4) After fitting a prototype[8] (use only AIC criterion to choose among the model with interactions), we need to do some tuning on the model. The main purpose in this step is trying to further reduce the number of explanatory variables while remain AIC almost unchanged. The reason is that AIC alone does not always provide a satisfying model. Sometimes, AIC leads to a model with too many explanatory variables and it is reasonable to sacrifice the AIC (a little bit) to reduce the number of predictors. The way we tune this model is to remove those variables with an insignificant estimation.

There are two different tests help us find out insignificant estimations. Like the hypothesis testing in linear regressions, in Poisson regression, we also conduct hypothesis testing on the coefficients of explanatory variables. Since the estimations in GLM is calculated using maximum likelihood estimation (MLE), we could use Wald test statistic. This statistic approximately follows a standard normal distribution under the assumption of MLE and large sample size. In testing our estimated coefficients, we test the null hypothesis: the coefficient of a specific variable is 0 versus the alternative: the coefficient of a specific variable is not 0. We remove variables with large p-values. The output for our prototype model is in appendix 1.

Another method of testing whether the estimation of coefficients is significant is called embedded model test. The idea is to fit two models M0 and M1. M1 contains the variable we want to test while M0 contains all the other variables M1 contains except this target variable. Since we can add a constraint on M1 to get M0 (simply constrain the coefficient of our target variable to be 0), we say M0 is embedded in M1. Again, since we use MLE for Poisson Regression, we can find out

[8] The exact form of this prototype is in appendix 1.

the likelihood function for M1 and M0. Under large sample assumption and asymptotic property of MLE, it is possible use likelihood ratio test to see whether the estimation of target variable's coefficient is significant or not.

All the above tuning process on prototype model can be found in appendix 1. After tuning the model, we have the following model:

(final model)[9]

After estimation, the (estimated) systematic component in formula (1) has the following form (linear combination of explanatory variables and their interactions):

$$\widehat{\eta_i} = \widehat{ln(\mu_i)} = X_i{}^T\hat{\beta} = X_{i1}\widehat{\beta_1} + X_{i2}\widehat{\beta_2} + \cdots + X_{i1}X_{i2}\widehat{\beta_{12}} + \cdots \quad (2)$$

$$\widehat{\mu_i} = exp\left(X_{i1}\widehat{\beta_1} + X_{i2}\widehat{\beta_2} + \cdots + X_{i1}X_{i2}\widehat{\beta_{12}} + \cdots\right) \quad (3)$$

where X_i's and $\hat{\beta}$'s are specified in the following table.

Table 2.5 Final Model Result

X_i	$\hat{\beta}$	X_i	$\hat{\beta}$
intercept	-2.746	$I_{ch=3}(i)$	-0.774
$X_{i,sex}$	$+0.980$	$X_{i,income} * I_{ch=2}(i)$	$+1.122$
$X_{i,age}$	-0.738	$X_{i,income} * I_{ch=2}(i)$	$+2.540$
$X_{i,income}$	-1.655	$I_{hs=1}(i) * I_{insurance=2}(i)$	-0.117
$X_{i,ill}$	$+0.246$	$I_{hs=1}(i) * I_{insurance=3}(i)$	-0.006
$I_{ad=1}(i)$	$+1.876$	$I_{hs=1}(i) * I_{insurance=4}(i)$	$+0.563$
$I_{hs=1}(i)$	-0.146	$I_{ad=1}(i) * I_{insurance=2}(i)$	-1.987
$I_{insurance=2}(i)$	-14.90	$I_{ad=1}(i) * I_{insurance=3}(i)$	-1.031
$I_{insurance=3}(i)$	-0.008	$I_{ad=1}(i) * I_{insurance=4}(i)$	$+0.600$
$I_{insurance=4}(i)$	-0.979	$X_{i,age} * I_{hs=1}(i)$	$+0.524$
$I_{ch=2}(i)$	-0.002	$X_{i,ill} * I_{hs=1}(i)$	-0.040

where

$X_{i,j}$ denotes the value of the sj-th explanatory variable for the i-th person.

$I_{j=k}(i)$ denotes the indicator function. It evaluates to 1 if the j-th explanatory variable equals to level k for the i-th person

j is the abbreviation of explanatory variables we specified in section 2.1. k is the index of each level in a categorical variable.

i is the index for i-th person.

All the other components and assumptions in Poisson Regression remains the same as in section 3.1.

3.3 Model Diagnostics

(step 5) Now let's look at the diagnostic on our model.

(residual plots) As in the linear regression model, we first look at residual plots. Different from linear regression models, in generalized linear regression models, we have different residuals because our model has more complicated structures. Deviance residuals is the one we usually use in residual plot. Graph 2.6(a) is deviance residual versus estimated systematic component. Since the deviance residual[10] is scaled by a "variance" parameter[11], the plot is supposed to have constant variance. Graph 2.6(b) is squared raw residual[12] versus estimated mean. If the response is Poisson, then the squared raw

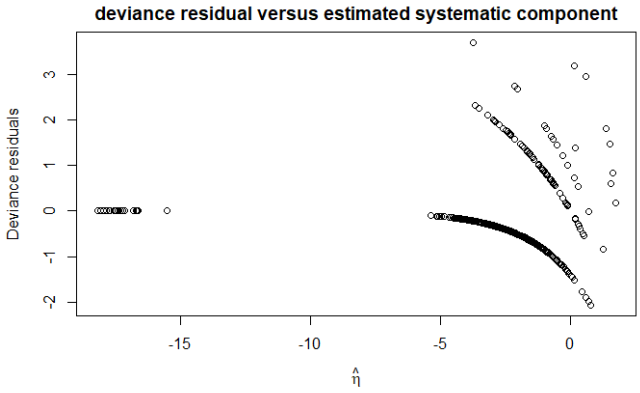
[9] We did not know this is our final model before conducting any diagnostics. Fortunately, this is the one we finally adopted.

[10] Definition of Deviance residual is defined Appendix 2

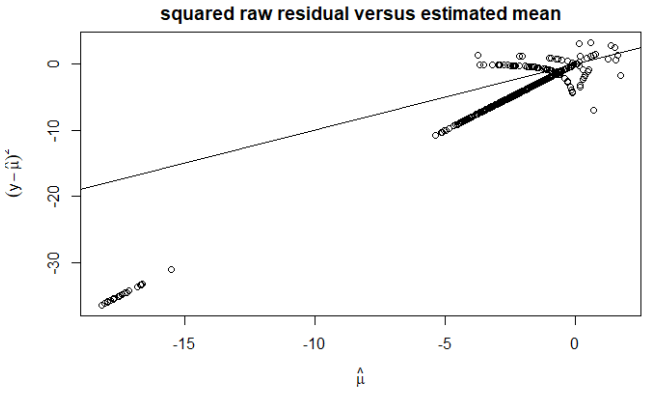
[11] The parameter is called variance function. It has a certain relationship with the variance. Luckily, in Poisson Regression, variance function equals to the variance.

[12] raw residual is defined as $y_i - \widehat{\mu_i}$

residual should be on average equal to the mean because its an estimate for the variance and variance equals mean for Poisson distribution.



Graph 2.6(a)



Graph 2.6(b)

Graph 2.6 residual plots

Unfortunately, residual plots do not really help in our case. The reason is that our dataset contains too many zeros (about 83%). We have a theoretical explanation on why excess zeros result in a residual plot like this (see Appendix 2). Simulations are also done in simulating Poisson dependent variables with very small mean and regular-sized mean. We use simulated dependent variables to do Poisson regression. The result is similar to what we have here. Residual plots are difficult to interpret when the mean is very small. But residual plots are helpful for regular-sized means (see Appendix 3 for more on this simulation).

(bootstrap)[13] Although regular residual plots are hard to interpret, we have other methods to see whether the model fits our data well. One method is to use parametric bootstrap. In our case, parametric bootstrap is done by generating y_i's (dependent variables) from our estimated mean. Then we refit a model with the same set of explanatory variables and these new generated dependent variables. Record the coefficients estimated for this model. Repeat this many times and record all the coefficients estimated each time. Then, for each coefficient (β_i), use its recorded values to construct a 95% quantile confidence interval. The idea is that if our final model fits the data (mean of the data) well, our coefficients are supposed to land in these confidence intervals. Otherwise, the estimation of mean from our model deviates from the true mean of the data. The result of bootstrap is shown in Table 2.7.

Table 2.7 Bootstrap Confidence Intervals and Final Model Estimation

Variable	95% lower bound	95% upper bound	estimation	in	Variable	95% lower bound	95% upper bound	estimation	in
intercept	-3.823	-1.743	-2.746	yes	$I_{ch=3}(i)$	-1.664	-0.031	-0.774	yes
$X_{i,sex}$	+0.569	+1.453	+0.980	yes	$X_{i,income} * I_{ch=2}(i)$	-1.467	+3.037	+1.122	yes
$X_{i,age}$	-0.780	+2.017	+0.738	yes	$X_{i,income} * I_{ch=2}(i)$	+1.341	+4.232	+2.540	yes
$X_{i,income}$	-3.167	-0.618	-1.655	yes	$I_{hs=1}(i) * I_{insurance=2}(i)$	-0.042	+0.351	+0.117	yes
$X_{i,ill}$	+0.077	+0.428	+0.246	yes	$I_{hs=1}(i) * I_{insurance=3}(i)$	-0.175	+0.185	-0.006	yes
$I_{ad=1}(i)$	+1.353	+2.424	+1.876	yes	$I_{hs=1}(i) * I_{insurance=4}(i)$	+0.323	+0.853	+0.563	yes

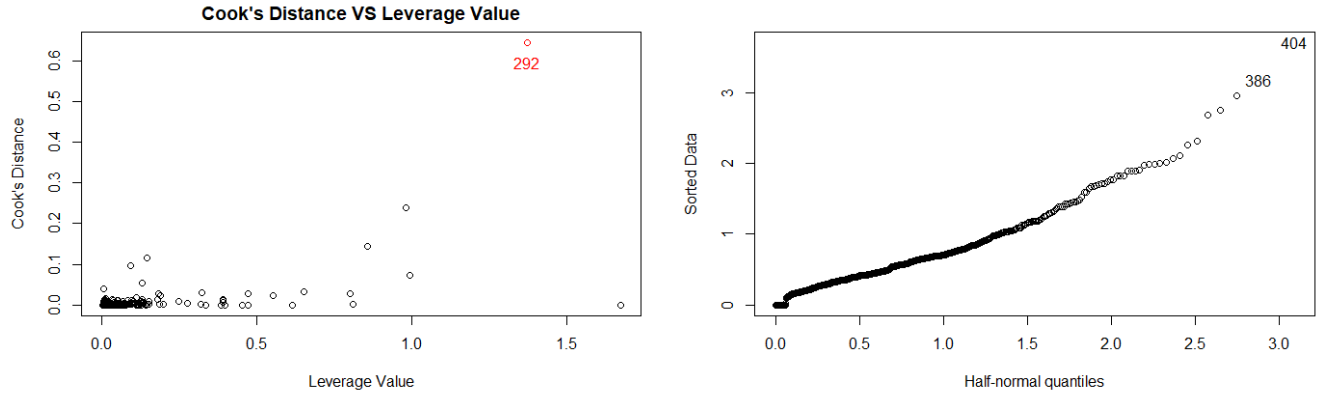
[13] Bootstrapping Regression Models Appendix to An R and S-PLUS Companion to Applied Regression (2002)

$I_{hs=1}(i)$	-0.504	+0.088	-0.146	yes	$I_{ad=1}(i) * I_{insurance=2}(i)$	-2.681	-1.389	-1.987	yes
$I_{insurance=2}(i)$	-16.288	-14.817	-14.895	yes	$I_{ad=1}(i) * I_{insurance=3}(i)$	-1.847	-0.308	-1.031	yes
$I_{insurance=3}(i)$	-0.677	+0.762	-0.008	yes	$I_{ad=1}(i) * I_{insurance=4}(i)$	-0.323	+1.759	+0.600	yes
$I_{insurance=4}(i)$	-2.104	-0.130	-0.979	yes	$X_{i,age} * I_{hs=1}(i)$	0.125	+1.067	+0.524	yes
$I_{ch=2}(i)$	-0.854	+1.012	-0.002	yes	$X_{i,ill} * I_{hs=1}(i)$	-0.090	+0.003	-0.040	yes

We see from the result that all of our estimation lands in the 95% quantile confidence interval. Therefore, we do not have warning that our model deviates from the real data structure.

(overdispersion) Overdispersion means the variance of our observation is larger than the nominal (theoretical) variance determined by Poisson distribution. It usually arises when we sampled from clustered population or correlated with each other through time. If the overdispersion occurs, the model assumption of Poisson Regression is violated, and we need to find another model to solve this problem. One parameter that closely related to overdispersion is the dispersion parameter. It is a parameter obtained in the canonical form of the exponential family. In Poisson Regression, dispersion parameter is assumed to be 1. We can also estimate this parameter in our regression using Pearson chi-squared statistic or deviance. Usually, if the estimated dispersion parameter is greater than 1 (to some extent), we think our model suffers from overdispersion. We can either compare estimated dispersion parameter with 1 or construct a test for over dispersion through some statistics[14]. In our model, the estimation based on Pearson Chi-squared statistic is 1.49 while the estimation based on deviance is 0.71. So, there is not enough evidence indicates that our model suffers from overdispersion. For the overdispersion test, the p-value is 0.108 and we cannot reject the null hypothesis that there is no overdispersion in the model under 10% significance level.

(outliers) Finding out possible outliers is part of diagnostics. We usually concern about outliers that have high leverage values and inconsistent with data. If an observation has both high leverage and inconsistent with data, it will influence the regression a lot (the regression result could differ a lot with and without it). Since this kind of data is usually resulted by experimental error and contribute less to our underlying data structure, we tend to remove them. 3 types of detecting potential outliers are used in this study: plot cook's distance versus leverage (Graph 2.8), half normal plot (Graph 2.8) and outlier test.



Graph 2.8 cook statistic versus leverage & Half-normal plot

Similar to linear regression model, in GLM, we have cook's distance and leverage value. Leverage can be used to identify points which might have large influence on the overall fit based on covariates alone. Notice that, high leverage points may

[14] See Appendix 4 for theoretical explanation on estimation of dispersion parameter and overdispersion tests.

not be outliers if they are consistent with the model. However, if they are inconsistent with the model, they usually result in a huge change in fitted values[15]. Since data points with large residuals or high leverage may distort the outcome and accuracy of a regression, Cook's distance measures the effect of deleting a given observation. Points with a large Cook's distance are considered to merit closer examination in the analysis. In GLM, leverage depends on covariates and a weight matrix while cook's distance depends also on responses in a more direct approach. In our case, the 292th observation may be outlier. To see whether it has a huge influence on the model, we refit the model without this observation[16]. All coefficients do not change a lot. And in our report, we include the 292th observation.

We can use half-normal plot to detect outliers by checking whether all the residuals approximately lie on the straight line. Half-normal plot uses simulated confidence envelopes as a reference for model's residuals. For our plot, no evident residual is spotted.

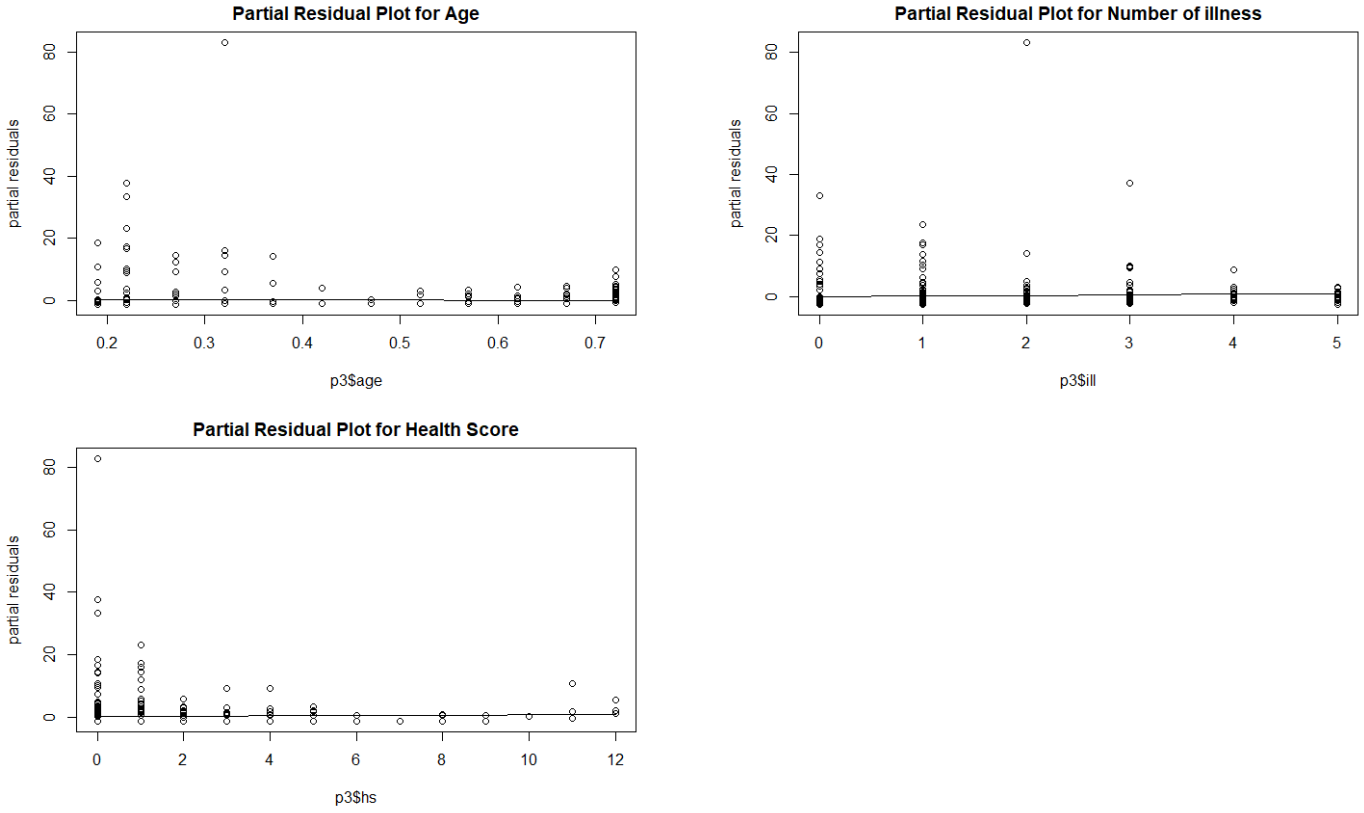
(partial residual plot) The partial residual plots are used for finding higher order terms. If the plot indicates a linear pattern, it indicates that no higher terms are needed. Partial residual is defined as

$$r_{W,i} + x_{ij}\widehat{\beta}_J$$

where

$$r_{W,i} = (y_i - \widehat{y}_i)\frac{d\widehat{\eta}_i}{d\widehat{\mu}_i}$$

For (discrete and continuous) numerical variables in the final model, we have the following plots:



Graph 2.9 Partial Residual Plots for (discrete and continuous) numerical variables

The result indicates that there is no evidence of higher terms. This result is consistent with the AIC result since we do not have any second order effects (except interactions) when we use AIC to choose among models. Note that higher terms of categorical variables are hard to interpret and may not be meaningful. Therefore, we do not include partial residuals for categorical variables.

---

[15] For simple linear regression, points at the end of the regression line have high leverage values.

[16] Comparison of models with and without 292th observation is in Appendix 5.

3.4 Model Interpretation and prediction

(interpretation)

Variables without an interaction term. Take sex variable for example, given all other conditions the same, on average, the natural logarithm of mean of number of consultations (in the past 4 weeks) for female is 0.98 units higher than males. The natural logarithm is specified through our systematic component.

Variables with an interaction term. When a variable has both main effect and interaction effect, interpreting the main effect does not really make sense. Therefore, we will only interpret the interaction effect. Take income and type of medical conditions for example. Given a fixed type of medical conditions (for example at level 2), one unit increase in income leads to 1.12 increase in the logarithm of mean number of consultations. When interpret interaction terms, we need to fix one of the variable values and interpret the other one. To save space, we will not interpret all the coefficients in Table 2.5.

(prediction) The following demonstrates how to use our model to do prediction.

Using our systematic component and estimated coefficients in table 2.5, we could estimate the expected number of pharmacist consultations by a new customer within a 4-week period, if the pharmacy was provided with values of the other variables for that new customer. We plug in the value of other variables into formula (3) and calculated the estimated mean.

Take the first person in our data for example, we estimate the expected number of his/her consultations by plug in the values of all the other variables into formula (3). This gives us the estimated value 0.135. In other words, our model stands for the idea that the number of consultations for the first person in our data follows a Poisson distribution with parameter 0.135. We could use this to estimate the probability that his/her number of pharmacist consultations equals to set of different numbers. The result is summarized in the following table.

Table 2.10 Estimated Probability for Different Number of Consultations

Num Of Consultation	Prob	Num Of Consultation	Prob
0	0.87359062	5	0.00000033
1	0.11806002	6	0.00000001
2	0.00797752	7	0.00000000
3	0.00035937	8	0.00000000

Since we have very small parameter 0.135 for our Poisson distribution, it is expected to see that the probability of having a large number of consultations is very small. Note that the probability greater than 6 is not identically 0. They are very small and rounded to be 0.

Part IV Conclusion

In this study, we explore the relationship among number of consultations with a pharmacist in the past 4 weeks and other variables. The quantitative relation is summarized in Table 2.5. Diagnostics are conducted regarding residuals, overdispersion, outliers and higher terms. Since residual plots are hard to interpret in our case, we use bootstrap method to see whether our model fits the data well. After passing those diagnostics, we illustrate how to use our model to estimate the expected number of pharmacist consultations by a new customer within a 4-week period, provided with values of the other variables for that new customer.

Some highlights are:

a) We discuss the possible problems regarding data collection and integrity and make assumptions related to these problems and GLM.

b) The set of variables lp, fp, fr is supposed to form a partition of the sample space. But we have response with all zeros in all three variables. In order to solve this problem and prevent from losing information, we combine these three variables into a new categorical variable and add a new level called unspecified.
c) Besides part b), we reformulate other variables for better interpretation and modeling.
d) We provide a theoretical explanation on why the residual plots are hard to interpret when we have many zero observations (Appendix II).

Possible improvements are:

a) We need to figure out why the problem described above happens. We can see from Table 2.5 that level 2 in insurance variable has an abnormally large influence on dependent variable. If the person belongs to this category, then the estimated mean would be really small. This may be true since this category correspond to those who do not have pharmacy coverage. However, the large standard error of this estimate may indicate a collinearity problem among variables. Figuring out why we have observations like this may help us solve this problem.
b) This study uses bootstrap method in place of regular residual plots. We could also implement other possible models to deal with excess zeros. One possible model is called zero-inflated Poisson Regression. It models the values of zero and values greater than zero separately. It allows us to use different variables in fitting the two parts while maintaining a relationship between two parts. It may be reasonable to use this model for our dataset because whether a person needs consultations or not may be modeled differently from the number of consultations given that the person needs consultations.

This is the end of this report. Please feel free to contact me if any additional information is needed.

Reference

[1] Diane Lambert, Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, Technometrics(1992)

[2] John Fox, Bootstrapping Regression Models Appendix to An R and S-PLUS Companion to Applied Regression, 2002

[3] Tsung-han Tsai and Jeff Gill, Interactions in Generalized Linear Models: Theoretical Issues and an Application to Personal Vote-Earning Attributes, social sciences(2013)

Appendix 1 Prototype Model selected using AIC criterion

$$\widehat{\eta_i} = \widehat{ln(\mu_i)} = X_i{}^T\hat{\beta} = X_{i1}\widehat{\beta_1} + X_{i2}\widehat{\beta_2} + \cdots + X_{i1}X_{i2}\widehat{\beta_{12}} + \cdots$$

Appendix Table 1 Prototype Model Result

X_i	$\hat{\beta}$	P-value	X_i	$\hat{\beta}$	P-value
intercept	-2.746	0.0008	$I_{hs=1}(i) * I_{insurance=2}(i)$	-0.117	0.9999
$X_{i,sex}$	$+0.980$	0.0000	$I_{hs=1}(i) * I_{insurance=3}(i)$	-0.006	0.9783
$X_{i,age}$	-0.738	0.3621	$I_{hs=1}(i) * I_{insurance=4}(i)$	$+0.563$	0.0001
$X_{i,income}$	-1.655	0.0055	$I_{ad=1}(i) * I_{insurance=2}(i)$	-1.987	0.9994
$X_{i,ill}$	$+0.246$	0.0009	$I_{ad=1}(i) * I_{insurance=3}(i)$	-1.031	0.0222
$I_{ad=1}(i)$	$+1.876$	0.3909	$I_{ad=1}(i) * I_{insurance=4}(i)$	$+0.600$	0.1765
$I_{hs=1}(i)$	-0.146	0.6987	$X_{i,age} * I_{hs=1}(i)$	$+0.524$	0.0580
$I_{insurance=2}(i)$	-14.90	0.9855	$X_{i,ill} * I_{hs=1}(i)$	-0.040	0.0154
$I_{insurance=3}(i)$	-0.008	0.8777	$X_{i,age} * I_{ad=1}(i)$	$+2.322$	0.0399
$I_{insurance=4}(i)$	-0.979	0.0253	$I_{ad=1} * I_{ch=2}(i)$	-0.509	0.2644
$I_{ch=2}(i)$	-0.002	0.6702	$I_{ad=1} * I_{ch=3}(i)$	$+0.816$	0.0709
$I_{ch=3}(i)$	-0.774	0.1467	$X_{i,ill} * I_{ch=2}(i)$	-0.005	0.9747
$X_{i,income} * I_{ch=2}(i)$	$+1.122$	0.2839	$X_{i,ill} * I_{ch=3}(i)$	-0.365	0.0624
$X_{i,income} * I_{ch=3}(i)$	$+2.540$	0.0003	AIC: 583.74		

P-values are calculated using Wald-test. We remove the last 5 variables on the right column and fit an embedded model. The reason is that from the interaction plot, these interaction terms are not significant. Embedded model test has p-value is 0.02. It seems that we removed too many of them. But we cannot reject it with significant level 1%. Also, the AIC for embedded model is 586.99 which is not a large increase. We prefer using the model with less predictors (our final model).

Appendix 2

Theoretical Explanation on Why the Residual Plot Is Hard To Interpret

The deviance residuals are defined as:

$$r_i = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$$
$$d_i = 2w_i\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - \mathrm{b}(\tilde{\theta}_i) + \mathrm{b}(\hat{\theta}_i)\right]$$

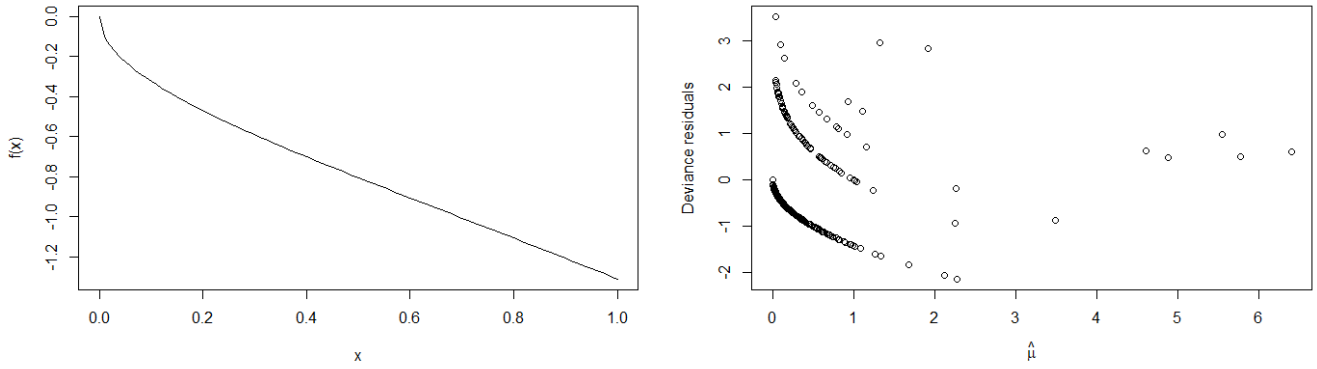
where $w_i = 1$ for Poisson distribution. $\tilde{\theta}_i$ and $\hat{\theta}_i$ are MLE for Poisson parameter in saturated model and current model.

For saturated model, $\tilde{\theta}_i = y_i$. $b(\theta) = e^{\theta}$ in Poisson distribution.

For every zero observation,

$$r_i = sign(y_i - \hat{\mu}_i)\sqrt{2w_i\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - e^{\tilde{\theta}_i} + e^{\hat{\theta}_i}\right]}$$
$$= sign(0 - \hat{\mu}_i)\sqrt{2\left[0(0 - \hat{\theta}_i) - e^{0} + e^{\hat{\theta}_i}\right]}$$
$$= -\sqrt{2e^{\hat{\theta}_i} - 2}$$

Recall the graph for $f(x) = -\sqrt{e^x - 1}$



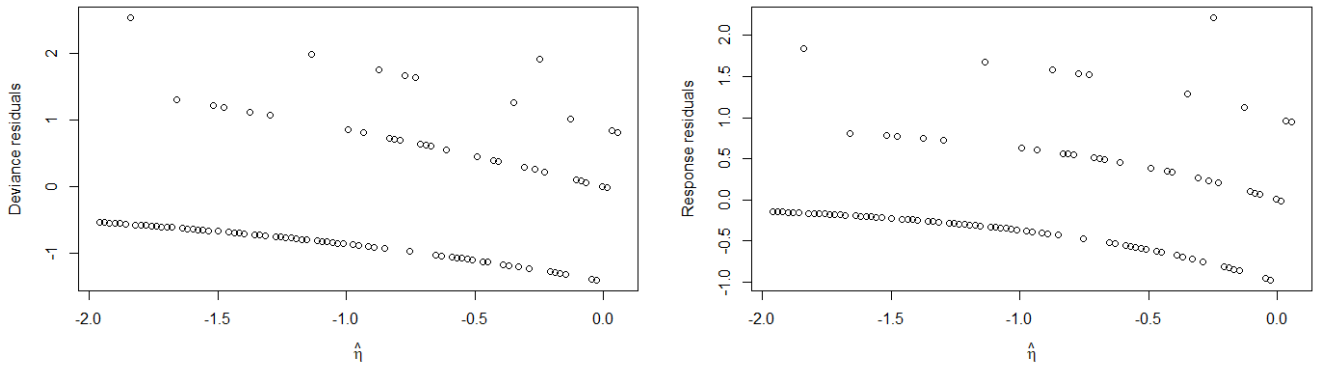
Appendix Graph 2 Comparing theoretical result with our data

We can see that our plot of deviance residuals versus estimated mean is indeed in this form. Other residual plots have similar shape because we have this square root of exponential term.

Appendix 3

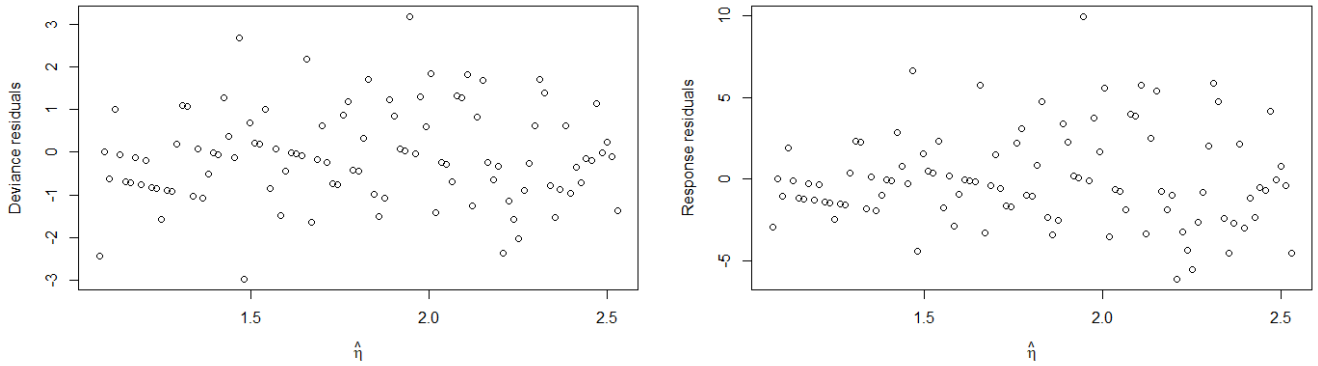
Simulate Data from Poisson Distribution with Small and Regular Parameter Value

1. simulated 100 observations with means less than 1 (0 to 1.01)



Appendix Graph 3 Residual plots with simulated data (small parameter)

2. simulated 100 observations with means greater than 1 (2 to 12)



Appendix Graph 4 Residual plots with simulated data (regular-sized parameter)

Appendix 4 Estimate Dispersion Parameter and Overdispersion Test

Dispersion parameter can be estimated using the following two methods:

$$\hat{\phi} = \frac{D_M}{n-p}$$

$$\tilde{\phi} = \frac{X^2}{n-p}$$

where D_M is deviance of our model M and X^2 is the Pearson chi-squared statistic.

Reason is that

$$\frac{X^2}{\phi} \sim \chi^2_{n-p} \text{ (approximately)}$$

$$\frac{D_M}{\phi} \sim \chi^2_{n-p} \text{ (approximately)}$$

and we use the mean of chi-squared distribution to estimate the ϕ.

Using these two approximate distributions, the rule of thumb is therefore:
Concern about potential overdispersion if

$$X^2 > \chi^2_{n-p,\alpha}$$

or

$$D_M > \chi^2_{n-p,\alpha}$$

Appendix 5 Estimation with and without potential outlier

Appendix table 5 Final model with and without potential outlier (292-th).

X_i	$\hat{\beta}$ (with)	$\hat{\beta}$ (without)	X_i	$\hat{\beta}$ (with)	$\hat{\beta}$ (without)
intercept	-2.746	-2.652	$I_{ch=3}(i)$	-0.774	-1.295
$X_{i,sex}$	$+0.980$	+1.101	$X_{i,income} * I_{ch=2}(i)$	$+1.122$	+1.000
$X_{i,age}$	-0.738	+0.286	$X_{i,income} * I_{ch=2}(i)$	$+2.540$	+2.995
$X_{i,income}$	-1.655	-1.609	$I_{hs=1}(i) * I_{insurance=2}(i)$	-0.117	+0.027
$X_{i,ill}$	$+0.246$	+0.270	$I_{hs=1}(i) * I_{insurance=3}(i)$	-0.006	+0.000
$I_{ad=1}(i)$	$+1.876$	+1.852	$I_{hs=1}(i) * I_{insurance=4}(i)$	$+0.563$	+0.327
$I_{hs=1}(i)$	-0.146	-0.057	$I_{ad=1}(i) * I_{insurance=2}(i)$	-1.987	-1.787
$I_{insurance=2}(i)$	-14.90	-14.620	$I_{ad=1}(i) * I_{insurance=3}(i)$	-1.031	-1.008
$I_{insurance=3}(i)$	-0.008	+0.053	$I_{ad=1}(i) * I_{insurance=4}(i)$	$+0.600$	+0.108
$I_{insurance=4}(i)$	-0.979	-0.523	$X_{i,age} * I_{hs=1}(i)$	$+0.524$	+0.373
$I_{ch=2}(i)$	-0.002	+0.052	$X_{i,ill} * I_{hs=1}(i)$	-0.040	-0.039