Poisson Regression Modelling for Analysis of Health Survey Data

Smita Sukhadeve

November 10, 2016

1 Introduction

Purpose of this paper is to demonstrate of use of Poisson Regression Modelling for the analysis of health survey data. We intend to identify factors that contribute towards the increasing number of consultations with the non-doctor health professional. Sample data used during this study is derived from the dataset used in the research study[1]. This dataset is not the representative of true population since it only considers 5120 individuals over 18 years old. Original source of this data traced back to the Australian Health Survey conducted in the year 1977-78. This survey data comprised of 40,650 individuals with 20 variables and represents the Australian population during the period 1977-78. For the purpose this study, we randomly selected 400 individuals among 5120. For the sake of simplicity, we limit our attention to only 13 variables including response. Detailed description of the variables included in the study is as follows:

Dataset Summary

Variable	Description
NONDOCCO	Number of consultations with non-doctor health professionals (Dependent variable)
SEX	1 or 0 for representing female and male respectively
INCOME	Annual income in tens of thousands of dollars
LEVYPLUS	Equals 1 if it's private insurance
FREEPOOR	Equals 1 if free government insurance due to low income
FREEREPA	Equals 1 if free government insurance due to old-age, disability or veteran status
ILLNESS	Number of illnesses in past 2 weeks
ACTDAYS	Number of days of reduced activity in past two weeks due to illness or injury
HSCORE	General health questionnaire score using Goldberg's method (High score bad hlth)
CHCOND1	Equals 1 if chronic condition(s) but not limited in activity, 0 other
CHCOND2	Equals 1 if chronic condition(s) and limited in activity, 0 otherwise
AGE	Participants age in year divided by 100
AGESQ	Age Squared

Average age of individuals in the sample is 49. Response, number of consultations takes values from 0 to 12. Average of number of consultations is 2.15 and sample variance is 5.3 which shows data is slightly over-dispersed. It appears that number of consultations are correlated with variables like FREEREPA, ACTDAYS, ILLNESS and HSCORE[Appendix-I]. Distribution chart for the number of consultation shows that

we have around 54% one's in our sample indicating around half of the participants consulted non-doctor health professional only once during the study period. Following charts explains the distribution of participants w.r.t. different parameters.

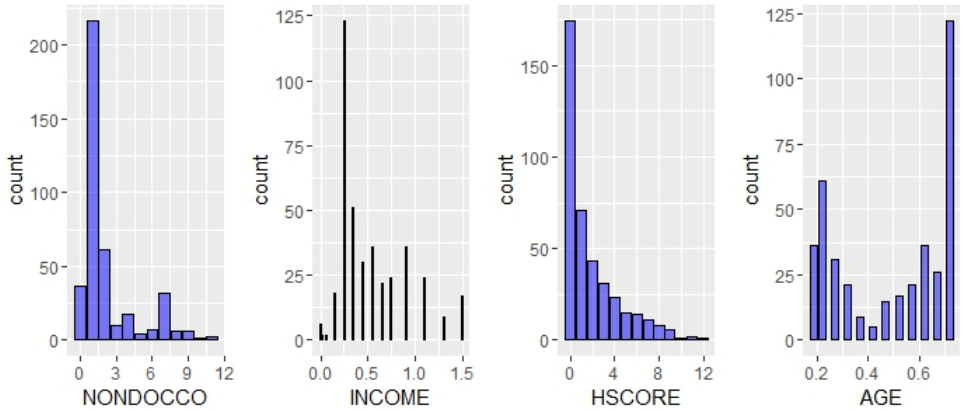


Figure 1: Participants Distribution by consultation count, Income, health score, Age receptively

The distribution of participants w.r.t. categorical variables used in the study can be explained by following charts:

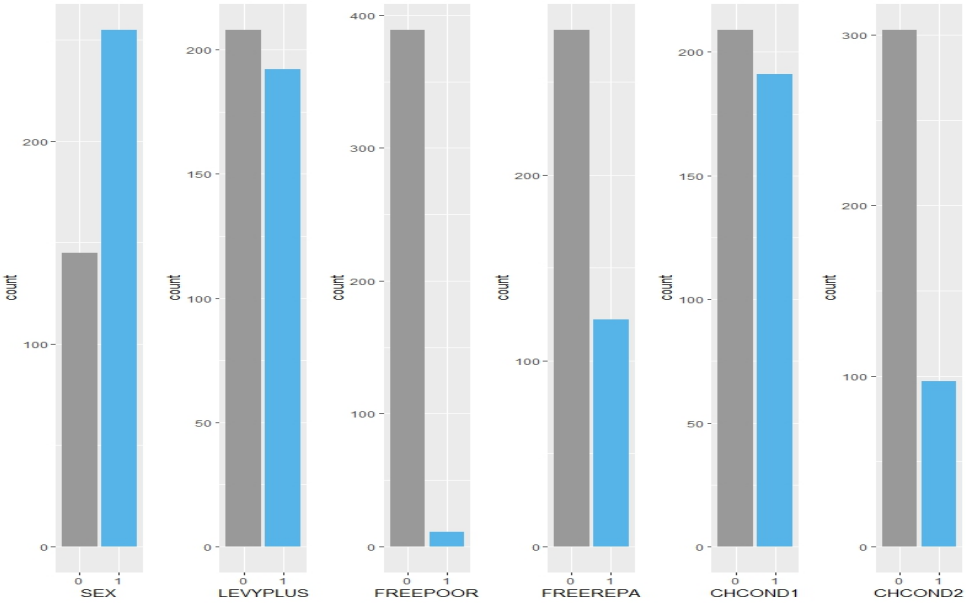


Figure 2: Distribution of participants by gender, insurance type, and chronic conditions

2 Methods

The histogram of the response(Figure-1) is skewed, non-normal and takes only positive values. However, Simple Linear Regression is suitable approach only when the response is normally distributed. In Poisson Regression setting, we assume response is discrete and takes only non-negative values. Therefore, we assume Poisson Distribution of the our dependent variable and decide to use Poisson Regression to analyse the effect of explanatory variables on the consultations count over specific period of time i.e. 1977-1978.

We start with analysing the data to find out presence of any missing values and outliers. Luckily, we don't have any missing values in our sample dataset. We then convert selected categorical variables into factors before fitting any model. This tells a statistical tool to consider these variables as categorical. After initial Data exploration and cleaning, we first fit Poisson Regression Model using all the explanatory variables to get a glimpse of how Poisson Regression works on overall data and to find out the significant explanatory variables. We use R-glm() function to fit the model. Pearson "goodness-of-fit" test is used to assess how well this model fits the observed data. glm() function uses Wald Statistic test to decide significance of variable.

Next, we use stepwise backward variable selection method to find the best subset of predictors to discover the optimal model. This variable selection process uses Akaike Information Criteria (AIC) information criteria to find out the optimal model using best subset of predictors. Process starts with fitting model using all the predictors. We follow repetitive approach where we drop variable at each iteration, refit the model and estimate 'Akaike Information Criteria'. At each iteration, we check the AIC of model before and after removal of predictor. We decide to keep removed predictor if the AIC of model increases after removal of current predictor. Finally, we choose the model with least value of AIC. Akaike Information Criteria for a given model is calculated as below:

$$AIC = \frac{(RSS + 2d\hat{\sigma}^2)}{n\hat{\sigma}^2}$$

where,
n = number of observations
$\hat{\sigma}^2$ = estimated variance of error associated with each response measurement
RSS = residual sum of squares
d = number of explanatory variables

Next, we analyse both models and evaluate their performances using following approach:

- Observe Residual Deviance of both the models
- Global fit statistics using Pearson goodness-of-fit test
- Estimation of Pearson Residual
- Conduct Analysis of variance (ANOVA) test to compare both the models

In final step, we select best model between 1 and 2 after analysing result of anova test. we calculate ratio of residual deviance to residual degrees of freedom to decide presence of over-dispersion in the selected model. If the estimated ratio > 1, we confirm presence of over-dispersion in the model. In case dispersion is present, we attempt to correct the model by adding dispersion parameter(ϕ) and evaluate the performance of the modified model. We plan to use F-statistic to find the best model among a modified Poisson Regression or regular Poisson Regression when the over-dispersion is presents. F- statistic is simply ratio of two variances

taking degree of freedom into account. Larger value of F-test represents greater dispersion. In technical section, we have explained in detail theory behind the process of model building and parameters used for the model analysis.

3 Result

3.1 Model-1 using all the explanatory variables

We used all the explanatory variables to build the first model. We modeled dependent variable as log of the mean of the number of consultations. The estimated model using the complete predictors set is as follows:

log(NONDOCCO) = 0.57 - 0.02SEX1 - 1.23AGE + 1.027AGESQ - 0.29INCOME + 0.26LEVYPLUS - 0.20FREEPOOR1 + 0.68FREEREPA1+ 0.02ILLNESS+ 0.04ACTDAYS + 0.019HSCORE+ 0.04CHCOND11 + 0.19CHCOND21

where,
FREEPOOR1 = participants with free government insurance due do poor condition
FREEREPA1 = participants with free government insurance due to old age/disability/veteran status
CHCOND11 = participants having chronic condition(s) but not limited to activity
CHCOND21 = participants having chronic condition(s) but limited to activity

For this model, we have a null deviance of 793.21 on 399 degrees of freedom. Residual deviance decreased to 630.33 after including the explanatory variables with the loss of 12 degrees of freedom. We performed Pearson goodness-of-fit test to measure the model performance on the overall data. P-value using Chi-square test statistics for this model is $6.40e^{-14}$ which is significant at 0.05 significance level. It means estimated model appears to perform well on the data. Ratio of residual deviance to the residual degrees of freedom is (1.63 = 630.33/387) which is slightly greater than 1 indicate slight presence of over-dispersion. Result of the anova test shows how deviance reduces from 8.9 to 3.01 after addition of predictor at each step. We observed FREEREPA produces significant reduction in the residual deviance from 737.38 to 703.46. Following table presents the result of anova test and variables marked with '***' are significant at 0.05 significance level as (Chisq test statistics < 0.05).

Table-1 ANOVA result for model-1

Variable	Resid.Dev	Pr(>Chi)
NULL	399.00	NA
SEX	784.24	$2.758e^{-03}$ ***
AGE	764.22	$7.66e^{-06}$ ***
AGESQ	763.92	0.58
INCOME	745.73	$1.99e^{-05}$ ***
LEVYPLUS	740.84	0.03
FREEPOOR	737.38	0.06
FREEREPA	703.46	$5.74e^{-09}$ ***
ILLNESS	694.18	$2.31e^{-03}$ ***
ACTDAYS	637.31	$4.66e^{-14}$ ***
HSCORE	634.40	0.08
CHCOND1	633.33	0.30
CHCOND2	630.32	0.082

Although, result of global fit test favours the model-1, we can not ignore the significant value of residual deviance(630.33). Moreover, we have used all the predictors and out of which only few variables appears to be significant at 0.05 significance level. Thus, next we attempted to find the model-2 using variable selection approach.

3.2 Model-2 using Stepwise Backward Selection

We used stepwise backward selection process to find the optimal model using best subset of predictors. We used the R function step() to perform variable selection. We started with full model(model-1) fitted using all the explanatory variables. AIC column in following table represent model AIC resulting from the deletion of variable at each step. In AIC approach of model selection, we favor model with smaller residual but we also penalise the model for including further predictors. This helps to avoid over-fitting of the model.

Table-2 AIC estimates for Model-2

Step	Resid.Dev	AIC
	630.33	1558.4
- SEX	630.37	1556.49
- CHCOND1	630.57	1554.69
- AGESQ	630.99	1553.11
- FREEPOOR	631.48	1551.59
- AGE	632.52	1550.63
- ILLNESS	634.01	1550.13

Estimated optimal model using variable selection approach is as follows:

log(NONDOCCO) = 0.33 - 0.25INCOME + 0.23LEVYPLUS1 + 0.63FREEREPA1 + 0.04ACTDAYS+ 0.026HSCORE + 0.16CHCOND21

Residual deviance of model-2 is 634.01 which is slightly greater than model-1. However, we have 393 degrees of freedom for the model-2 which is better than model-1. We performed the Pearson goodness-of-fit test to measure the model performance on the overall data and found P-value using Chi-square test statistics is $1.36e^{-13}$ which is still significant at 0.05 significance level. Thus, we can still say that model performed well on the data. Ratio of residual deviance to the residual degrees of freedom is (1.61) which is slightly less than model-1. Model-2 is nested within model-1. Therefore, we can use anova() function to compare the performance of the model-2. p-value is 0.7 which is insignificant at 0.05 significance level. This indicates that additional parameters in model-1 doesn't add much to the performance of the model and we can safely drop them. Finally, we conclude that model-2 is better than model-1 in term of simplicity and overall performance.

3.3 Model-3 Modified Poisson Model

For Poisson distribution, we assume mean = variance. Model violates this assumption, when over-dispersion is present. For Model-2 dispersion is 1.61 which is slightly greater than 1. We used R function dispersiontest() to decide whether this value is significant at 0.05 level. We performed two-sided test which

tests under and over-dispersion for the model. We found p-value is $3.6e^{-8}$ which is significant at 0.05 significance level. Hence, we reject null hypothesis and confirm the absence of over-dispersion. Slight greater value of variance could be due to presence of outlier or wrong systematic component. However, we attempted to fit the model-3 by adding dispersion parameter(1.82) in model-2 and test the overall performance of two model using F-statistics.

For the modified Poisson model, value of regression co-efficient remains same. However, we observed larger standard error for the regression coefficient. p-value of variable slightly increased due to increase in S.E. We can no longer use regular deviance Chi-Square test statistics to compare modified Poisson and regular Poisson models. Instead, we performed F-test to compare between two models and choose best among them. We used R's drop1() function to conduct this test. Result of the F-statistics is as follows:

Table 3.1 F-Test for model-2 and model-3 comparison		
Variable	F-value	Pr(>F)
NONE	NA	NA
- SEX	2.96	$8.6e^{-2}$
- LEVYPLUS	3.17	$7.5e^{-2}$
- FREEREPA	22.3	$3.23e^{-6}$
- ACTDAYS	22.8	$62.55e^{-6}$
- HSCORE	2.35	0.12
- CHCOND2	2.7	0.1

3.4 Selection and interpretation of Final Model

For F-test, we assume Null hypothesis H_0 : modified null hypothesis fits the data. Table 3.1 shows estimation of F-value based on modified model and regular Poisson model. We can see F-value for the model with all selected predictors is 2.7 and value is not significant (p-value > 0.05) at 0.05 significance level. Thus, we fail to reject null hypothesis and conclude modified Poisson i.e. model-3 is better than the regular Poisson model.

Table 3.2 Regression Coefficient estimates for Model 2 and Model-3				
Variable	Estm-Coeff	SE-Model2	SE-Model3	p-value based on model-2
Intercept	0.33	0.11	0.15	0.003
INCOME	-0.25	0.12	0.16	0.03
LEVYPLUS1	0.23	0.10	0.14	0.03
FREEREPA1	0.63	0.11	0.14	$8.31e^{-09}$
ACTDAYS	0.04	0.006	0.008	$3.21e^{-10}$
HSCORE	0.03	0.013	0.018	0.048
CHCOND21	0.16	0.07	0.106	0.03

Our Final Model is as follows:

log(NONDOCCO) = 0.33 - 0.25INCOME + 0.23LEVYPLUS1 + 0.63FREEREPA1 + 0.04ACTDAYS+ 0.026HSCORE + 0.16CHCOND21

Intercept is the log of mean of number of consultations when all the predictors are zero. Thus, we can say that when all the predictors equal to zero, log of mean of number of consultations is 0.57. The regression parameter -0.25 of the income indicates that one unit increase in annual income is associated with decrease in the log scale of number of consultation visits by 0.25 holding rest of the parameters constant. In other word, 1 unit increase in annual income increases the average of number of consultations by exp(-0.25)= 0.77 times. FREEREPA1 and ACTDAYS are significant predictors. Average count of consultations to non-doctor health professional among the individuals with free government insurance due to old age or disability is around 1.9 times more than the individuals without free government insurance. 1 unit increase in number of days with reduced activity increases the average of number of consultations by 1.04. Similarly, we see average of number of consultations increase by 1.02 with the 1 unit change of health score. Average number of consultations to non-doctor health professional is 1.18 times more among individuals with chronic condition and limited activity than the individuals without these condition.

Table 3.2 Regression Coefficients of Selected Model

Variable	Estm-Coeff	exp(estm-coeff)	S.E.
Intercept	0.33	1.39	0.15 **
INCOME	-0.25	0.77	0.16 *
LEVYPLUS1	0.23	1.26	0.14 *
FREEREPA1	0.63	1.87	0.14 ***
ACTDAYS	0.04	1.04	0.008 ***
HSCORE	0.03	1.02	0.018 *
CHCOND21	0.16	1.18	0.106 *

Note: Explanatory variables indicated with *** found significant at 95% confidence level.

Discussion

To summarize, our response, number of consultations with non-doctor health professional is count and non-normal. Therefore, we assumed Poisson distribution $Y \sim P(\mu)$ for the response and used log link function to model it. We fitted 3 models and compared their performance using global fitness test - Pearson test statistics, compared their residual deviance. We used variable selection method two find the optimal model. However, this approach doesn't always guarantee that the model, we selected is the best model. For Poisson distribution, we always assume mean = variance. However, in practice we often observe variance is much larger than mean. In such cases, model violets property of Poisson distribution. We calculated dispersion parameter and performed dispersion test. Using F-test, we concluded modified Poisson Model is better than regular Poisson Model. In this study, model-2 (Regular Poisson Model) did not have very large value of over-dispersion. Therefore, we could have chosen model-2 instead of fitting modified Poisson model. In practice, we use modified Poisson Model when the dispersion is significantly high. One can use negative binomial regression which don't have a strict restriction for the values of variance and mean as an alternative to modified Poisson Regression. I think, we can get a better model than selected if we discretize the variables such age or income. Because, Pearson statistic better approximated by Chi-square distributions after collapsing explanatory variables.

Technical Description

I. Poisson Regression Model

For Poisson Distribution, we focus on the average number of rare events occurred during specific interval of the time.
$E(Y) = \mu = \alpha + \beta_1 X$.....(1.2)

We take log as link function which guarantees that non-negative value of response. Thus, GLM for the response having Poisson distribution is represented as follows:

$$log(\mu) = \alpha + \sum_{i=1}^{p} \beta_i X_i \qquad ....1.2$$

We take exponential to get the value of μ

$$\mu = e^{\alpha + \sum_{i=1}^{p} \beta_i X_i}$$

Thus, GLM components are as follows:
1. Random Component $= E(\mu(x)) = Y$
2. Systematic component $= \alpha + \sum_{i=1}^{p} \beta_i X_i$
3. Link function : log

Next we use Maximum likelihood approach to find the unknown parameter β
p.d.f. for the Poisson distribution is given as follows:

$$P(Y = y_i) = \frac{\mu^{y_i} e^{-\mu}}{y_i!} .....y_i = 0, 1, 2...n$$

where $\mu = \alpha + \sum_{k=1}^{p} \beta_k X_k)$

$$L(\mu) = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$$

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{y_i!} . \prod_{i=1}^{n} \mu^{y_i} . e^{-\sum_{i=1}^{n} \mu}$$

For sake of simplicity, we assume, $\mu = \alpha + \beta_1 X_1$

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{y_i!} . \prod_{i=1}^{n} (\alpha + \beta_1 x_1)^{y_i} . e^{-\sum_{i=1}^{n} \alpha + \beta_1 x_1}$$

To find the unknown parameter β we will use MLE. Taking log both side, we get

$$l(\alpha, \beta_1) \propto \sum_{i=1}^{n} y_i log(\alpha + \sum_{k=1}^{p} \beta_k X_k) - \sum_{i=1}^{n} (\alpha + \sum_{k=1}^{p} \beta_k X_k) .....(1)$$

$$\frac{\partial [l(\alpha, \beta_1)]}{\partial \alpha} = 0$$

and

$$\frac{\partial[l(\alpha, \beta_1)]}{\partial \beta_1} = 0$$

Thus, using MLE approach, we can find $\hat{\beta}$ after taking derivative of equation 1 w.r.t. β and equating it to zero.

II. Interpretation of Poisson Regression Model

Consider a very simple Poisson regression model as follows:

$$log(\mu^{(x)}) = \alpha + \beta_1 X - - - 2$$

After one unit increment of x,

$$log(\mu^{(x+1)}) = \alpha + \beta_1 (X + 1)$$

$$log(\mu^{(x+1)}) = \alpha + \beta_1 X + \beta_1 - -3$$

From equation 2 and 3 we get ,

$$\beta_1 = log(\frac{\mu^{(x+1)}}{\mu^{(x)}})$$

$$e^{\beta_1} = \frac{\mu^{(x+1)}}{\mu^{(x)}}$$

Thus, coefficient has a multiplicative effect on the mean by a factor of e^{β_1}

III. Residual Deviance

Residual Deviance of model calculated as below:
Residual deviance = (deviance of estimated model) - (deviance of ideal model where predicted values equals to the observed model)

IV. Global Fit Statistics

We use Global-Fit Statistics to assess how well model fits the observed data.
Null hypothesis for these statistic is $H_0 =$ The model fits the data and lack of fit is not statistically large
Pearson "goodness-of-fit" is:

$$\chi^2 = \sum_{i=1}^{N} \frac{y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

where,
$y_i =$ observed count
$\mu_i =$expected count
χ^2 is approximately Chi-square distributed with degrees of freedom(residual df)=number of counts - number of model parameter

IV. Pearson Residual

Pearson residual= $\frac{r_i}{\sqrt{\hat{\mu}_i}}$

To control the over-dispersion in the model, we use dispersion parameter ϕ and calculate Pearson residual as below: $p_i^c = \frac{r_i}{\sqrt{\phi \hat{\mu}_i}}$
Where,
$\phi = \frac{1}{n-k} \sum_{i=1}^{n} \frac{r_i^2}{\mu_i}$, r_i = residual = $y_i - \hat{\mu_i}$, $\mu_i = e^{\hat{\beta_0} + \beta_1 \hat{X}_1 i + .. + \beta_k \hat{X}_{ki}}$

V. Wald Statistic to test Significance of coefficients

For Wald Statistics, we consider $H_0 : \beta = 0$ Vs $H_a = \beta \neq 0$

$$z = \frac{(\hat{\beta} - \beta_0)}{ASE} = \frac{(\hat{\beta})}{ASE}$$

Wald Statistic $= z^2 \approx \chi^2$
Where, $\hat{\beta}$ = estimate value of β , β_0 = observed value of β, ASE = Asymptotic Standard Error
We can use both z and z^2 for the Hypothesis testing.
We reject null hypothesis when $p(z) < 0.05$ and conclude variable is significant at 0.05 significance level.

VI. Detection of over-dispersion

(Residual Deviance) / (Residual degrees of freedom) > 1 indicates over-dispersion

VII. F-Statistics for goodness-of-fit of Regular Poisson Model and Modified Poisson Model

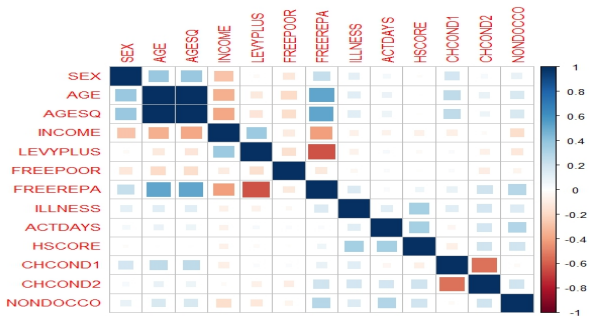
F-Statistics based on the Likelihood ratio Test.

$$F_{df_S - df_L, k-p} = \frac{\frac{Dev_S - Dev_L}{df_S - df_L}}{\hat{\sigma}^2}$$

where,
Dev_S = Deviance of Saturated model (Model S)
Dev_L = Residual Deviance of Least saturated Model (Model L)
$df_S - df_L$ = number of restriction, df_S and df_L are residual degrees of freedom for model S and L respt.
$\hat{\sigma}$ = estimated variance of Models

4 Appendix

I. Correlation Plot



II. Summary Statistics

```
summary(myData[, c(2,4,8,9,13)])
```

```
      AGE              INCOME          ILLNESS         ACTDAYS
 Min.   :0.1900   Min.   :0.0000   Min.   :0.000   Min.   : 0.000
 1st Qu.:0.2700   1st Qu.:0.2500   1st Qu.:1.000   1st Qu.: 0.000
 Median :0.5700   Median :0.3500   Median :2.000   Median : 0.000
 Mean   :0.4905   Mean   :0.5326   Mean   :1.927   Mean   : 2.123
 3rd Qu.:0.7200   3rd Qu.:0.7500   3rd Qu.:3.000   3rd Qu.: 1.000
 Max.   :0.7200   Max.   :1.5000   Max.   :5.000   Max.   :14.000
    NONDOCCO
 Min.   : 0.00
 1st Qu.: 1.00
 Median : 1.00
 Mean   : 2.15
 3rd Qu.: 2.00
 Max.   :11.00
```

III. Data Summary

```
str(myData)
```

```
'data.frame':   400 obs. of  13 variables:
 $ SEX      : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 1 ...
 $ AGE      : num  0.72 0.62 0.57 0.72 0.32 ...
 $ AGESQ    : num  0.518 0.384 0.325 0.518 0.102 ...
 $ INCOME   : num  0.25 0.25 0.15 0.45 0.75 ...
 $ LEVYPLUS: Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 1 1 1 ...
 $ FREEPOOR: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ FREEREPA: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 2 ...
 $ ILLNESS : int  1 1 5 5 0 0 5 3 2 2 ...
 $ ACTDAYS : int  0 0 2 0 0 0 0 0 0 0 ...
 $ HSCORE  : int  1 0 0 2 0 2 2 0 2 0 ...
 $ CHCOND1 : Factor w/ 2 levels "0","1": 2 2 1 2 1 1 1 1 1 2 ...
 $ CHCOND2 : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 2 2 2 1 ...
 $ NONDOCCO: int  1 1 1 1 1 2 1 1 7 2 ...
```

5 References

1. A.C. Cameron and P.K. Trivedi (1986) 'Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests'
2. Dr. Cheng Peng. (Fall 2016 - STA 588) Note 9 Introduction to Generalized Linear Models
3. Dr. Cheng Peng. (Fall 2016 - STA 588) Note 11 Inference on Poisson Regression
4. Dr. Cheng Peng. (Fall 2016 - STA 588) Note 12 Dispersion in Poisson Regression
5. James, G. et al. (2013). An introduction to statistical learning (Vol. 112). New York: Springer
6. https://onlinecourses.science.psu.edu/stat504/node/169