Regression Techniques Homework

Subhrajyoty Roy (MB1911)

27 September 2019

Generalized Linear Model

A generalized linear model (GLM) is comprised of three components.

- The response variable is distributed according to a general exponential family of distribution F with mean $\mu = E(Y|X)$.
- A linear function of the predictor variables, namely $\eta = X\beta$, where β is the unknown vector of coefficients to be estimated.
- A link function $g(\cdot)$ that links component 1 and component 2. Namely, $g(\mu) = \eta$.

To estimate β, we can write down the joint likelihood of them given the data X and Y, and then use the method of maximum likelihood.

Datasets

In this assignment, we shall demonstrate the idea of Logistic, Probit and Poisson Regression. For the Logistic and Probit regression model, we require the response variable to be binary in nature, while for the Poisson regression model, we generally need response data as result of some counting processes.

- For the binary response data, we shall consider **NBA Rookie 5 Year Career Longevity Data**, which is available in the link https://data.world/exercises/logistic-regression-exercise-1. The dataset contains the first year player profile of the rookie NBA (Basketball) players, and the goal is to predict whether they would sustain a career longer than 5 years. This study is extremely useful to the owner of the teams in NBA leagues which helps them to extend the contract with basketball players based on their future career aspects.

- The dataset for Poisson regression is from Cameron and Johansson (1997) data used in Count Data Models. The dataset file is named **health.dta**, available in the link http://www.econ.uiuc.edu/~econ508/data.html and http://www.econ.uiuc.edu/~econ508/Stata/e-ta16_Stata.html. This datasets tries to model the number of consultations in the past four week with non-doctor health professionals (chemist, optician, physiotherapist, etc.) based on the patient's age, sex, gender, income level and chronic disease status etc.

Logistic Regression

In the logistic regression model, we have the following:

- The response variable Y conditional on the predictors X follows a binomial distribution, with mean μ.
- The link function is logit function, i.e. $\log\left(\frac{\mu}{1-\mu}\right) = X\beta$.

```
NBAdata <- read.csv('./nba_logreg.csv')
head(NBAdata)
```

```
               Name GP  MIN PTS FGM FGA  FG. X3P.Made X3PA X3P. FTM FTA  FT.
1  Brandon Ingram 36 27.4 7.4 2.6 7.6 34.7      0.5  2.1 25.0 1.6 2.3 69.9

2 Andrew Harrison 35 26.9 7.2 2.0 6.7 29.6      0.7  2.8 23.5 2.6 3.4 76.5
3  JaKarr Sampson 74 15.3 5.2 2.0 4.7 42.2      0.4  1.7 24.4 0.9 1.3 67.0
4     Malik Sealy 58 11.6 5.7 2.3 5.5 42.6      0.1  0.5 22.6 0.9 1.3 68.9
5     Matt Geiger 48 11.5 4.5 1.6 3.0 52.4      0.0  0.1  0.0 1.3 1.9 67.4
6    Tony Bennett 75 11.4 3.7 1.5 3.5 42.3      0.3  1.1 32.5 0.4 0.5 73.2
  OREB DREB REB AST STL BLK TOV TARGET_5Yrs
1  0.7  3.4 4.1 1.9 0.4 0.4 1.3           0
2  0.5  2.0 2.4 3.7 1.1 0.5 1.6           0
3  0.5  1.7 2.2 1.0 0.5 0.3 1.0           0
4  1.0  0.9 1.9 0.8 0.6 0.1 1.0           1
5  1.0  1.5 2.5 0.3 0.3 0.4 0.8           1
6  0.2  0.7 0.8 1.8 0.4 0.0 0.7           0
```

We remove the name column which is not a potential predictor of the response variable *TARGE_5Yrs*.

```
NBAdata <- NBAdata[, -1]
NBAdata <- NBAdata[complete.cases(NBAdata), ]
NBAdata$TARGET_5Yrs <- factor(NBAdata$TARGET_5Yrs)
dim(NBAdata)
```

```
[1] 1329   20
```

We note that the data now contains 1329 many observations on 20 variables. Now, we fit a logistic regression model to the above data.

```
fit <- glm(TARGET_5Yrs ~ ., data = NBAdata,
           family = binomial(link = "logit"))
summary(fit)
```

```


Call:
glm(formula = TARGET_5Yrs ~ ., family = binomial(link = "logit"),
    data = NBAdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9787  -0.9907   0.5050   0.8673   2.2837

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.616876   1.221426  -3.780 0.000157 ***
GP           0.036016   0.004720   7.630 2.34e-14 ***
MIN         -0.061869   0.033207  -1.863 0.062441 .
PTS         -0.264321   0.884469  -0.299 0.765057
FGM         -0.025728   1.747283  -0.015 0.988252
FGA          0.346054   0.231207   1.497 0.134464
FG.          0.040506   0.021616   1.874 0.060943 .
X3P.Made     3.532284   1.330084   2.656 0.007915 **
X3PA        -1.170455   0.409834  -2.856 0.004291 **
X3P.         0.003916   0.005266   0.744 0.457058
FTM          0.770755   1.021638   0.754 0.450591
FTA         -0.231268   0.469066  -0.493 0.621984
FT.          0.008795   0.009912   0.887 0.374950
```

```
OREB         0.332455   1.286364   0.258 0.796063    
DREB        -0.662017   1.283314  -0.516 0.605949    
REB          0.546356   1.276438   0.428 0.668628    
AST          0.309696   0.112028   2.764 0.005702 ** 
STL          0.001577   0.318116   0.005 0.996045    
BLK          0.571704   0.270996   2.110 0.034889 *  
TOV         -0.304079   0.271926  -1.118 0.263463    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1763.1  on 1328  degrees of freedom
Residual deviance: 1461.5  on 1309  degrees of freedom
AIC: 1501.5

Number of Fisher Scoring iterations: 5
```

We find that, only a few variables are actually significant, like the number of games played, the minute of play, fields goal success rate, 3 pointers made, number of assists and blocks. Therefore, it is reasonable to refit a logistic regression model only with those varibles which are actually significant.

```r
fit <- glm(TARGET_5Yrs ~ GP + MIN + FG. + X3P.Made + X3PA + AST + BLK,
           data = NBAdata, family = binomial(link = "logit"))
summary(fit)
```

```


Call:
glm(formula = TARGET_5Yrs ~ GP + MIN + FG. + X3P.Made + X3PA + 
    AST + BLK, family = binomial(link = "logit"), data = NBAdata)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4468  -1.0135   0.5451   0.8702   2.2811  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.536558   0.567461  -6.232 4.60e-10 ***
GP           0.035885   0.004578   7.839 4.54e-15 ***
MIN          0.029057   0.015718   1.849  0.06451 .  
FG.          0.028578   0.012936   2.209  0.02716 *  
```

```
X3P.Made     2.732580   0.984090   2.777  0.00549 **
X3PA        -1.060804   0.368109  -2.882  0.00395 **
AST          0.101874   0.072409   1.407  0.15945
BLK          0.545544   0.246653   2.212  0.02698 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1763.1  on 1328  degrees of freedom
Residual deviance: 1491.8  on 1321  degrees of freedom
AIC: 1507.8

Number of Fisher Scoring iterations: 4
```

We see that the residual deviance does not increase much. However, we also note that *AST* is now insignificant. So, we again refit the model without this *AST* variable.

```
fit <- glm(TARGET_5Yrs ~ GP + MIN + FG. + X3P.Made + X3PA + BLK,
           data = NBAdata, family = binomial(link = "logit"))
summary(fit)
```

```


Call:
glm(formula = TARGET_5Yrs ~ GP + MIN + FG. + X3P.Made + X3PA +
    BLK, family = binomial(link = "logit"), data = NBAdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4207  -1.0169   0.5506   0.8702   2.2562

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.477379   0.565429  -6.150 7.75e-10 ***
GP           0.036427   0.004559   7.990 1.35e-15 ***
MIN          0.042507   0.012614   3.370 0.000752 ***
FG.          0.025788   0.012780   2.018 0.043602 *
X3P.Made     2.590193   0.973596   2.660 0.007804 **
X3PA        -1.013645   0.364880  -2.778 0.005469 **
BLK          0.403778   0.221879   1.820 0.068787 .
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1763.1  on 1328  degrees of freedom
Residual deviance: 1493.9  on 1322  degrees of freedom
AIC: 1507.9

Number of Fisher Scoring iterations: 4
```

We again see that residual deviance increase very small. Since, all variables are now significant, we stick with this current model. We see that the AIC of the final logistic model is 1507.8829977.

Probit Regression

In the probit regression model, we have the following:

- The response variable Y conditional on the predictors X follows a binomial distribution, with mean μ.
- The link function is probit function, i.e. $\Phi^{-1}(\mu) = X\beta$, where $\Phi(\cdot)$ is the cdf of standard normal distribution.

We fit the probit regression model with all predictors included first.

```
fit <- glm(TARGET_5Yrs ~ ., data = NBAdata,
           family = binomial(link = "probit"))
summary(fit)
```

```


Call:
glm(formula = TARGET_5Yrs ~ ., family = binomial(link = "probit"),
    data = NBAdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1212  -1.0026   0.5030   0.8779   2.3694

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -2.929745   0.713537  -4.106 4.03e-05 ***
GP           0.022005   0.002801   7.856 3.96e-15 ***
MIN         -0.031666   0.019270  -1.643  0.10033    
PTS         -0.169199   0.526061  -0.322  0.74773    
FGM         -0.054016   1.038489  -0.052  0.95852    
FGA          0.227945   0.132330   1.723  0.08497 .  
FG.          0.027238   0.012668   2.150  0.03155 *  
X3P.Made     2.101847   0.780658   2.692  0.00709 ** 
X3PA        -0.694110   0.239674  -2.896  0.00378 ** 
X3P.         0.002383   0.003129   0.762  0.44631    
FTM          0.402823   0.599955   0.671  0.50195    
FTA         -0.100234   0.267972  -0.374  0.70837    
FT.          0.005488   0.005853   0.938  0.34842    
OREB         0.201632   0.768178   0.262  0.79295    
DREB        -0.403765   0.765982  -0.527  0.59811    
REB          0.309885   0.762166   0.407  0.68431    
AST          0.165515   0.063824   2.593  0.00951 ** 
STL          0.016527   0.184777   0.089  0.92873    
BLK          0.347701   0.153649   2.263  0.02364 *  
TOV         -0.153605   0.157431  -0.976  0.32922    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1763.1  on 1328  degrees of freedom
Residual deviance: 1462.2  on 1309  degrees of freedom
AIC: 1502.2

Number of Fisher Scoring iterations: 5
```

We find that, only a few variables are actually significant, like the number of games played, fields goal success rate, 3 pointers made, number of assists and blocks. Therefore, it is reasonable to refit a probit regression model only with those varibles which are actually significant. Note that, logistic and probit regression although chooses different sets of predictors, the most significant predictors remain same in both cases.

```r
fit <- glm(TARGET_5Yrs ~ GP + FG. + X3P.Made + X3PA + AST + BLK,
           data = NBAdata, family = binomial(link = "probit"))
summary(fit)
```

```
Call:
glm(formula = TARGET_5Yrs ~ GP + FG. + X3P.Made + X3PA + AST +
    BLK, family = binomial(link = "probit"), data = NBAdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5670  -1.0194   0.5558   0.8688   2.3369

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.264443   0.332952  -6.801 1.04e-11 ***
GP           0.023803   0.002566   9.278  < 2e-16 ***
FG.          0.020876   0.007545   2.767 0.005658 **
X3P.Made     1.543764   0.577524   2.673 0.007516 **
X3PA        -0.554036   0.214050  -2.588 0.009644 **
AST          0.102945   0.033744   3.051 0.002282 **
BLK          0.442848   0.121217   3.653 0.000259 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1763.1  on 1328  degrees of freedom
Residual deviance: 1494.8  on 1322  degrees of freedom
AIC: 1508.8

Number of Fisher Scoring iterations: 4
```

We again see that residual deviance does not increase a lot. Since, all variables are now significant, we stick with this current model. Note that, the residual deviance here is slightly larger than the residual deviance for logistic model.We see that the AIC of the final logistic model is 1508.7820057.

Therefore, in terms of AIC, logistic regression performs slightly better than the probit model.

Poisson Regression

In the Poisson regression model, we have the following:

- The response variable Y conditional on the predictors X is assumed to follow a Poisson distribution, with mean λ.

- The link function is log, i.e. $\log(\lambda) = X\beta$, where $\log$ is the natural logarithm.

First, we load the data into *R*.

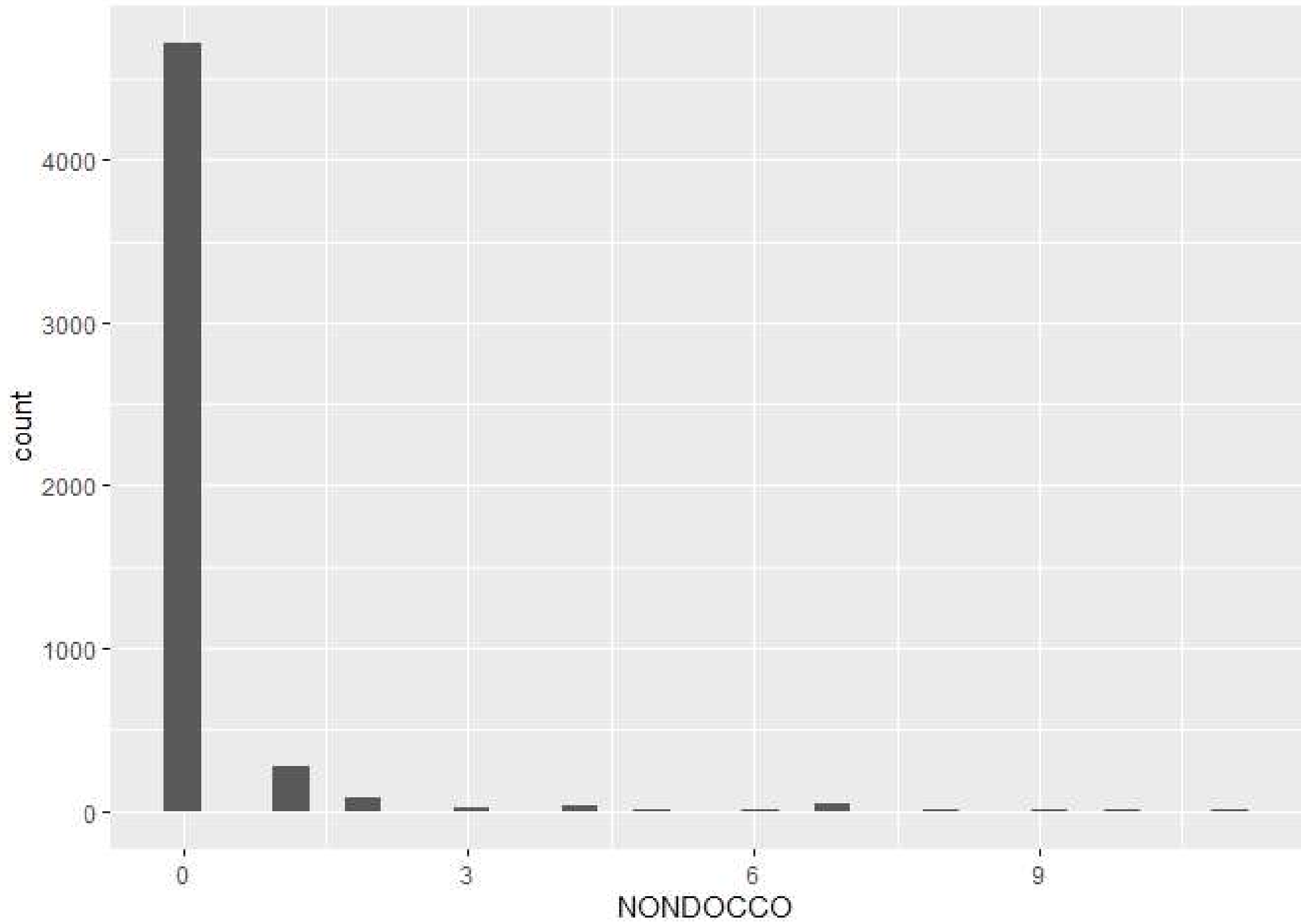
```
library(foreign)
library(ggplot2)
healthdata <- read.dta('./health.dta')
```

Before proceeding with the regression, let us first remove any *NA* values from the data and try plotting a histogram for the response variable.

```
healthdata <- healthdata[complete.cases(healthdata), ]
summary(healthdata)
```

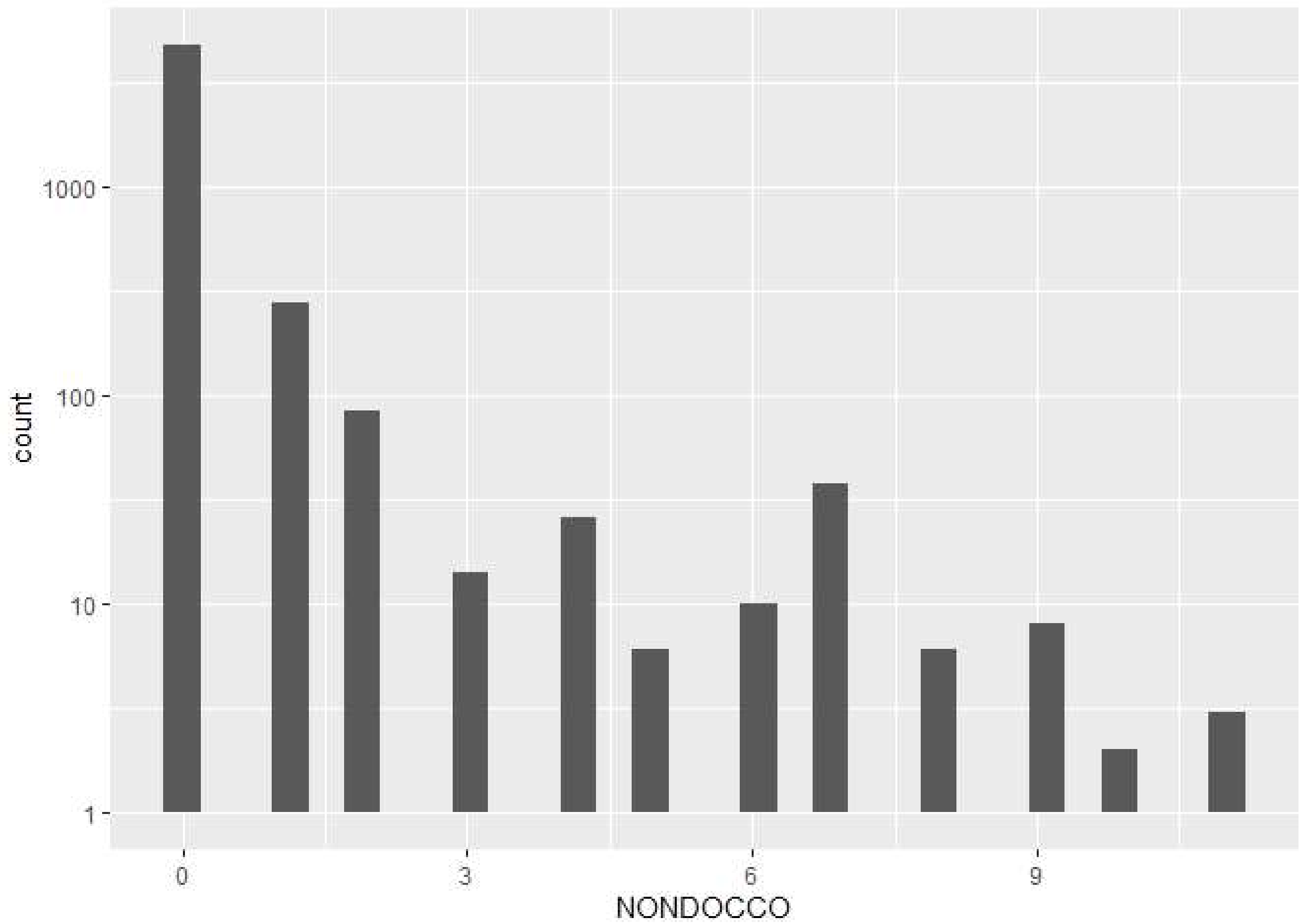
```
    NONDOCCO            SEX               AGE              INCOME
 Min.   : 0.0000   Min.   :0.0000   Min.   :0.1900   Min.   :0.0000
 1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.2200   1st Qu.:0.2500
 Median : 0.0000   Median :1.0000   Median :0.3200   Median :0.5500
 Mean   : 0.2146   Mean   :0.5206   Mean   :0.4064   Mean   :0.5832
 3rd Qu.: 0.0000   3rd Qu.:1.0000   3rd Qu.:0.6200   3rd Qu.:0.9000
 Max.   :11.0000   Max.   :1.0000   Max.   :0.7200   Max.   :1.5000
    LEVYPLUS          FREEPOOR          FREEREPA          ILLNESS
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000
 Median :0.0000   Median :0.00000   Median :0.0000   Median :1.000
 Mean   :0.4428   Mean   :0.04277   Mean   :0.2102   Mean   :1.432
 3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:2.000
 Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :5.000
    ACTDAYS            HSCORE          CHCOND1          CHCOND2
 Min.   : 0.0000   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 0.0000   Median : 0.000   Median :0.0000   Median :0.0000
 Mean   : 0.8619   Mean   : 1.218   Mean   :0.4031   Mean   :0.1166
 3rd Qu.: 0.0000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :14.0000   Max.   :12.000   Max.   :1.0000   Max.   :1.0000
```

```
ggplot(healthdata, aes(NONDOCCO)) + geom_histogram()
```



The diagram says that most of the counts are actually 0. Let us look at the counts in a logarithmic scale to see a better visualization

```
ggplot(healthdata, aes(NONDOCCO)) + geom_histogram() + scale_y_log10()
```



We note that, the above setup clearly seems like a zero inflated situation. Nevertheless, we still fit a Poisson regression just to see how it performs.

```
healthdata$SEX <- factor(healthdata$SEX)
healthdata$LEVYPLUS <- factor(healthdata$LEVYPLUS)
healthdata$FREEPOOR <- factor(healthdata$FREEPOOR)
healthdata$FREEREPA <- factor(healthdata$FREEREPA)
healthdata$CHCOND1 <- factor(healthdata$CHCOND1)
healthdata$CHCOND2 <- factor(healthdata$CHCOND2)

fit <- glm(NONDOCCO ~ ., healthdata, family = "poisson")
summary(fit)
```

```


Call:
glm(formula = NONDOCCO ~ ., family = "poisson", data = healthdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4975  -0.6500  -0.4728  -0.3683   7.7588
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.116128   0.137763 -22.620  < 2e-16 ***
SEX1         0.336123   0.069605   4.829 1.37e-06 ***
AGE          0.782335   0.200369   3.904 9.44e-05 ***
INCOME      -0.123275   0.107720  -1.144 0.252459    
LEVYPLUS1    0.302185   0.097209   3.109 0.001880 ** 
FREEPOOR1    0.009547   0.210991   0.045 0.963910    
FREEREPA1    0.446621   0.114681   3.894 9.84e-05 ***
ILLNESS      0.058322   0.021474   2.716 0.006610 ** 
ACTDAYS      0.098894   0.006095  16.226  < 2e-16 ***
HSCORE       0.041925   0.011613   3.610 0.000306 ***
CHCOND11     0.496751   0.086645   5.733 9.86e-09 ***
CHCOND21     1.029310   0.097262  10.583  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6127.9  on 5189  degrees of freedom
Residual deviance: 5052.5  on 5178  degrees of freedom
AIC: 6254.3

Number of Fisher Scoring iterations: 7
```

Note that, most of the variables turned out to be significant other than Income and the presence of Government Insurance Coverage. However, the residual deviance has not been decreasing much compared to the null deviance due to the fitting of this Poisson model.

On the other hand, we could have used a Zero Inflated Poisson model (as clear from the above plot of histogram).

```
library(pscl)

fit <- zeroinfl(NONDOCCO ~ ., healthdata)
summary(fit)
```

```


Call:
zeroinfl(formula = NONDOCCO ~ ., data = healthdata)
```

```
Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.0360 -0.2997 -0.2189 -0.1681 19.6996 

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.398990   0.167872   2.377  0.01747 *  
SEX1         0.064186   0.089403   0.718  0.47279    
AGE         -0.736371   0.243868  -3.020  0.00253 ** 
INCOME      -0.315433   0.137162  -2.300  0.02146 *  
LEVYPLUS1    0.258797   0.129494   1.999  0.04566 *  
FREEPOOR1    0.202052   0.266160   0.759  0.44777    
FREEREPA1    0.704008   0.146837   4.794 1.63e-06 ***
ILLNESS      0.013602   0.026104   0.521  0.60233    
ACTDAYS      0.052131   0.006472   8.054 7.98e-16 ***
HSCORE       0.025924   0.013694   1.893  0.05835 .  
CHCOND11     0.037017   0.116202   0.319  0.75007    
CHCOND21     0.337704   0.117784   2.867  0.00414 ** 

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.71482    0.23268  15.965  < 2e-16 ***
SEX1        -0.28544    0.12291  -2.322  0.02021 *  
AGE         -1.93159    0.35896  -5.381 7.40e-08 ***
INCOME      -0.19574    0.18681  -1.048  0.29474    
LEVYPLUS1   -0.06998    0.16512  -0.424  0.67171    
FREEPOOR1    0.26769    0.36231   0.739  0.46000    
FREEREPA1    0.29425    0.20541   1.432  0.15200    
ILLNESS     -0.09130    0.04028  -2.267  0.02341 *  
ACTDAYS     -0.06773    0.01300  -5.211 1.88e-07 ***
HSCORE      -0.03157    0.02328  -1.356  0.17504    
CHCOND11    -0.44480    0.14531  -3.061  0.00221 ** 
CHCOND21    -0.78581    0.17084  -4.600 4.23e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Number of iterations in BFGS optimization: 35 
Log-likelihood: -2282 on 24 Df
```

```r
AIC(fit)
```

```
[1] 4611.18
```

We also see that AIC is much smaller compared to the usual Poisson Regression model. Therefore, a Zero Inflated Poisson model is better to model this data. Also, the variable *Income* now is significant once we remove the zero inflation from the data using logistic modelling.

THANK YOU