MH8321-STATISTICAL MODELLING & DATA NALYSIS

Group Team Member:

HUANG RUNCHENG, WANG JIAHENG, ZHOU DUO, ZHANG YIQUN

1. Introduction

DebTrivedi data will be used for modelling and data analysis in this assignment, the project is to model the demand for hospital ward as captured by the number of hospital stays. In the dataset, **hosp** (the number of hospital stays) is adopted as the depend variable. And health status variables and socioeconomic variables as regressors.

2. Analysis of data

For this analysis, we select the variables used from the full data set:

dt <- DebTrivedi[, c(6, 7:19)]
head(dt)

	hosp	health	numchron	adldiff	region	age	black	gender	married	school	faminc	employed	privins	medicaid
1	1	average	2	no	other	6.9	yes	male	yes	6	2.8810	yes	yes	no
2	0	average	2	no	other	7.4	no	female	yes	10	2.7478	no	yes	no
3	3	poor	4	yes	other	6.6	yes	female	no	10	0.6532	no	no	yes
4	1	poor	2	yes	other	7.6	no	male	yes	3	0.6588	no	yes	no
5	0	average	2	yes	other	7.9	no	female	yes	6	0.6588	no	yes	no
6	0	poor	5	yes	other	6.6	no	female	no	7	0.3301	no	no	yes

To obtain a first overview of the dependent variable, we employ a histogram of the observed count frequencies. Histogram plot shown in Figure1. It gives high count of zeros in the dependent variables.

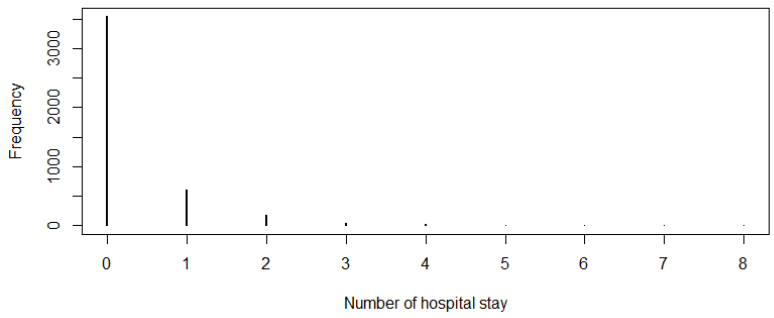

Figure 1

A Poisson GLM is used as first attempt to identify the relationship between the number of hospital stays and regressors. And we have the coefficient estimates along with corresponding Wald tests which is shown in Figure 2

m_pois <- glm(hosp ~ ., data = dt, family = poisson)
summary(m_pois)

```
Call:
glm(formula = hosp ~ ., family = poisson, data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9491  -0.7369  -0.6090  -0.4639   5.7675

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.968393   0.370788  -8.006 1.19e-15 ***
healthpoor       0.534764   0.070373   7.599 2.99e-14 ***
healthexcellent -0.709976   0.176284  -4.027 5.64e-05 ***
numchron         0.251134   0.018566  13.527  < 2e-16 ***
adldiffyes       0.344568   0.067759   5.085 3.67e-07 ***
regionnoreast   -0.120814   0.085900  -1.406  0.15959
regionother     -0.111628   0.072106  -1.548  0.12160
regionwest      -0.012271   0.085020  -0.144  0.88524
age              0.117454   0.045152   2.601  0.00929 **
blackyes         0.095894   0.091634   1.046  0.29534
gendermale       0.154379   0.062975   2.451  0.01423 *
marriedyes      -0.027354   0.065631  -0.417  0.67684
school           0.002369   0.008318   0.285  0.77581
faminc           0.006760   0.009847   0.686  0.49244
employedyes      0.038944   0.106914   0.364  0.71567
privinsyes       0.214679   0.080413   2.670  0.00759 **
medicaidyes      0.179618   0.101851   1.764  0.07781 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4662.5  on 4405  degrees of freedom
Residual deviance: 4109.0  on 4389  degrees of freedom
AIC: 6089.5

Number of Fisher Scoring iterations: 6
```

Figure 2

From the coefficients, health status variable such as self-perceived health status, number of chronic conditions and socioeconomic variable such as age, gendermale, adldiff (indicator that person has a condition that limits activities of daily living), privinsyes (private insurance indicator) gives highly significance impact on the number of hospital stays.

However, count data often exhibit overdispersion meaning that the variance exceeds the mean. For current case, variance is 0.55711 is slight greater than mean 0.29596. To

accommodate such overdispersion, quasi-Poisson regression is used as second attempt shown in Figure 3.

```
m_qp <- glm(hosp ~ ., data = dt, family =quasipoisson)
summary(m_qp)
```

```
Call:
glm(formula = hosp ~ ., family = quasipoisson, data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9491  -0.7369  -0.6090  -0.4639   5.7675

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.968393   0.467480  -6.350 2.37e-10 ***
healthpoor       0.534764   0.088725   6.027 1.80e-09 ***
healthexcellent -0.709976   0.222255  -3.194  0.00141 **
numchron         0.251134   0.023408  10.729  < 2e-16 ***
adldiffyes       0.344568   0.085429   4.033 5.59e-05 ***
regionnoreast   -0.120814   0.108301  -1.116  0.26468
regionother     -0.111628   0.090909  -1.228  0.21955
regionwest      -0.012271   0.107191  -0.114  0.90886
age              0.117454   0.056926   2.063  0.03914 *
blackyes         0.095894   0.115530   0.830  0.40656
gendermale       0.154379   0.079398   1.944  0.05191 .
marriedyes      -0.027354   0.082746  -0.331  0.74098
school           0.002369   0.010487   0.226  0.82130
faminc           0.006760   0.012415   0.544  0.58616
employedyes      0.038944   0.134795   0.289  0.77266
privinsyes       0.214679   0.101383   2.118  0.03427 *
medicaidyes      0.179618   0.128411   1.399  0.16195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.589551)

    Null deviance: 4662.5  on 4405  degrees of freedom
Residual deviance: 4109.0  on 4389  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Figure 3

A more formal way to accommodate over-dispersion in count data regression model is to use a negative binomial model, results are illustrated in Figure 4. Furthermore, Figure 5 presents another approach that zero-inflated Poisson regression is used to model **hosp** which includes excess zero counts. And a different way of augmenting the negative binomial count model with additional probability weight for zero counts is a zero-inflated negative binomial regression. The default model is fitted shown in Figure 6.

```
m_nb <- MASS::glm.nb(hosp ~ ., data = dt)
summary(m_nb)
summary(hosp <- zeroinfl(hosp ~ ., data = dt))
summary(hosp <- zeroinfl(hosp ~ ., data = dt, dist="negbin"))
```

```
Call:
MASS::glm.nb(formula = hosp ~ ., data = dt, init.theta = 0.5840497975,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3483  -0.6676  -0.5587  -0.4380   3.6735

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.3771329  0.4812990  -7.017 2.27e-12 ***
healthpoor       0.5112550  0.0984701   5.192 2.08e-07 ***
healthexcellent -0.6894085  0.1937542  -3.558 0.000373 ***
numchron         0.2764486  0.0256397  10.782  < 2e-16 ***
adldiffyes       0.3377104  0.0905693   3.729 0.000192 ***
regionnoreast   -0.1346648  0.1092496  -1.233 0.217712
regionother     -0.1344601  0.0929022  -1.447 0.147804
regionwest       0.0004014  0.1091862   0.004 0.997067
age              0.1720692  0.0588332   2.925 0.003448 **
blackyes         0.1028964  0.1182114   0.870 0.384058
gendermale       0.2131734  0.0805704   2.646 0.008150 **
marriedyes      -0.0342766  0.0844652  -0.406 0.684884
school           0.0011724  0.0107181   0.109 0.912897
faminc           0.0007634  0.0131543   0.058 0.953722
employedyes      0.0435324  0.1309582   0.332 0.739576
privinsyes       0.1842365  0.1037327   1.776 0.075721 .
medicaidyes      0.1411514  0.1369332   1.031 0.302632
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.584) family taken to be 1)

    Null deviance: 2907.8  on 4405  degrees of freedom
Residual deviance: 2552.1  on 4389  degrees of freedom
AIC: 5728.8

Number of Fisher Scoring iterations: 1


              Theta:  0.5840
          Std. Err.:  0.0536

 2 x log-likelihood:  -5692.8270
```

Figure 4

```
Call:
zeroinfl(formula = hosp ~ ., data = dt)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.0911 -0.4380 -0.3489 -0.2693 11.0116

Count model coefficients (poisson with log link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.172048   0.640980  -0.268 0.788381
healthpoor       0.250247   0.111635   2.242 0.024983 *
healthexcellent -0.953203   0.464364  -2.053 0.040101 *
numchron         0.119496   0.035777   3.340 0.000838 ***
adldiffyes       0.171518   0.106458   1.611 0.107151
regionnoreast   -0.149438   0.148997  -1.003 0.315880
regionother     -0.039937   0.116956  -0.341 0.732746
regionwest      -0.017970   0.137821  -0.130 0.896260
age             -0.073779   0.076503  -0.964 0.334851
blackyes         0.012907   0.148738   0.087 0.930847
gendermale      -0.015244   0.118992  -0.128 0.898062
marriedyes      -0.012964   0.113043  -0.115 0.908694
school          -0.010002   0.013848  -0.722 0.470137
faminc           0.015414   0.015119   1.019 0.307976
employedyes      0.007584   0.180680   0.042 0.966518
privinsyes       0.339808   0.143874   2.362 0.018184 *
medicaidyes      0.241304   0.169413   1.424 0.154345

Zero-inflation model coefficients (binomial with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      4.071640   1.113868   3.655 0.000257 ***
healthpoor      -0.556252   0.223347  -2.491 0.012755 *
healthexcellent -0.450032   0.746322  -0.603 0.546509
numchron        -0.276915   0.068444  -4.046 5.21e-05 ***
adldiffyes      -0.334389   0.194355  -1.721 0.085340 .
regionnoreast   -0.058310   0.240409  -0.243 0.808357
regionother      0.124997   0.189690   0.659 0.509927
regionwest      -0.015821   0.226471  -0.070 0.944305
age             -0.386115   0.136500  -2.829 0.004674 **
blackyes        -0.133972   0.250755  -0.534 0.593151
gendermale      -0.349744   0.192470  -1.817 0.069196 .
marriedyes       0.008068   0.181383   0.044 0.964523
school          -0.020887   0.023239  -0.899 0.368755
faminc           0.016264   0.021648   0.751 0.452468
employedyes     -0.060981   0.264236  -0.231 0.817485
privinsyes       0.274595   0.260793   1.053 0.292377
medicaidyes      0.176986   0.310194   0.571 0.568294
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 42
Log-likelihood: -2865 on 34 Df
```

Figure 5

```
Call:
zeroinfl(formula = hosp ~ ., data = dt, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.7110 -0.4537 -0.3431 -0.2237 12.3578

Count model coefficients (negbin with log link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.290512   0.607919  -2.123 0.033768 *
healthpoor       0.412943   0.110971   3.721 0.000198 ***
healthexcellent -1.081763   0.282195  -3.833 0.000126 ***
numchron         0.150399   0.033685   4.465 8.01e-06 ***
adldiffyes       0.223955   0.107751   2.078 0.037667 *
regionnoreast   -0.035042   0.142368  -0.246 0.805578
regionother     -0.042422   0.112899  -0.376 0.707103
regionwest       0.022786   0.132188   0.172 0.863140
age             -0.046152   0.071083  -0.649 0.516168
blackyes        -0.058633   0.137228  -0.427 0.669185
gendermale       0.043213   0.114645   0.377 0.706229
marriedyes      -0.070099   0.109810  -0.638 0.523233
school           0.003947   0.012741   0.310 0.756710
faminc           0.026353   0.017031   1.547 0.121781
employedyes      0.069575   0.187806   0.370 0.711038
privinsyes       0.296416   0.122493   2.420 0.015527 *
medicaidyes      0.322474   0.158286   2.037 0.041622 *
Log(theta)      -0.176909   0.136406  -1.297 0.194654

Zero-inflation model coefficients (binomial with logit link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      8.95917    2.69833   3.320 0.000899 ***
healthpoor      -1.29191    1.09201  -1.183 0.236789
healthexcellent -2.26535    2.01520  -1.124 0.260958
numchron        -0.85863    0.19954  -4.303 1.68e-05 ***
adldiffyes      -1.06443    0.66786  -1.594 0.110982
regionnoreast    0.48336    0.50586   0.956 0.339313
regionother      0.41538    0.41023   1.013 0.311277
regionwest       0.11551    0.47996   0.241 0.809815
age             -1.30075    0.39138  -3.324 0.000889 ***
blackyes        -0.92410    0.68456  -1.350 0.177042
gendermale      -0.87047    0.41482  -2.098 0.035870 *
marriedyes      -0.24405    0.39699  -0.615 0.538713
school           0.02246    0.05150   0.436 0.662701
faminc           0.07601    0.03717   2.045 0.040884 *
employedyes      0.05383    0.54727   0.098 0.921642
privinsyes       0.86118    0.60937   1.413 0.157590
medicaidyes      1.40996    0.71990   1.959 0.050165 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 0.8379
Number of iterations in BFGS optimization: 54
Log-likelihood: -2816 on 35 Df
```

Figure 6

3. Conclusion

DebTrivedi **hosp** model is fitted via several approaches which includes Poisson regression, Quasi-Poisson regression, Negative Binomial model, zero-inflated Poisson regression and zero-inflated Poisson. Poisson GLM is not fitted data appropriately because of slight overdispersion issue in count data. Therefore, quasi-Poisson model leads to an dispersion parameter 1.59 which is slightly larger than 1. And Negative Binomial, zero-inflated regression and zero-inflated negative binomial are fitted data well with improved AIC value than Poisson regression. there are no much differences in AIC for these three approaches. Nevertheless, Zero-inflated negative binomial model is selected to fit **hosp** data with the best AIC performance.

The predictors self-perceived health status (**healthpoor, healthexcellent**)**, numchrom, adldiffyes, privinsyes, medicaidyes** in the part of the count model predicting **hosp** are significant predictors.The predictors **numchron, age, gendermale, faminc** in the part of the logit model predicting excessive zeros is statistically significant.

For these data, the expected change in log(**hosp**) for a one unit increase in **healthpoor** is 0.412**.** With one unit increase the **healthexcellent** will lead to 1.08 decrease in log(**hosp**). Similarly, one unit increase in **numchron** result 0.15 increase in log(**hosp**). Furthermore, **adldiffyes** has an expected log(**hosp**) of 0.224 higher than **adldiffno**. The log odds of being an excessive zero would decrease by 1.3 for every unit increasing for **age.** In other words, the older people in the group, the less likely that zero would be due to not visit doctor.

In summary, the model is used to predict number of hospitals stays based on health status and socioeconomic status of patient. It will be helpful for hospital on ward arrangement and maximize the ward utilizations.