811_GLM

Sukhdeep Singh

September 29, 2018

Preliminary analysis

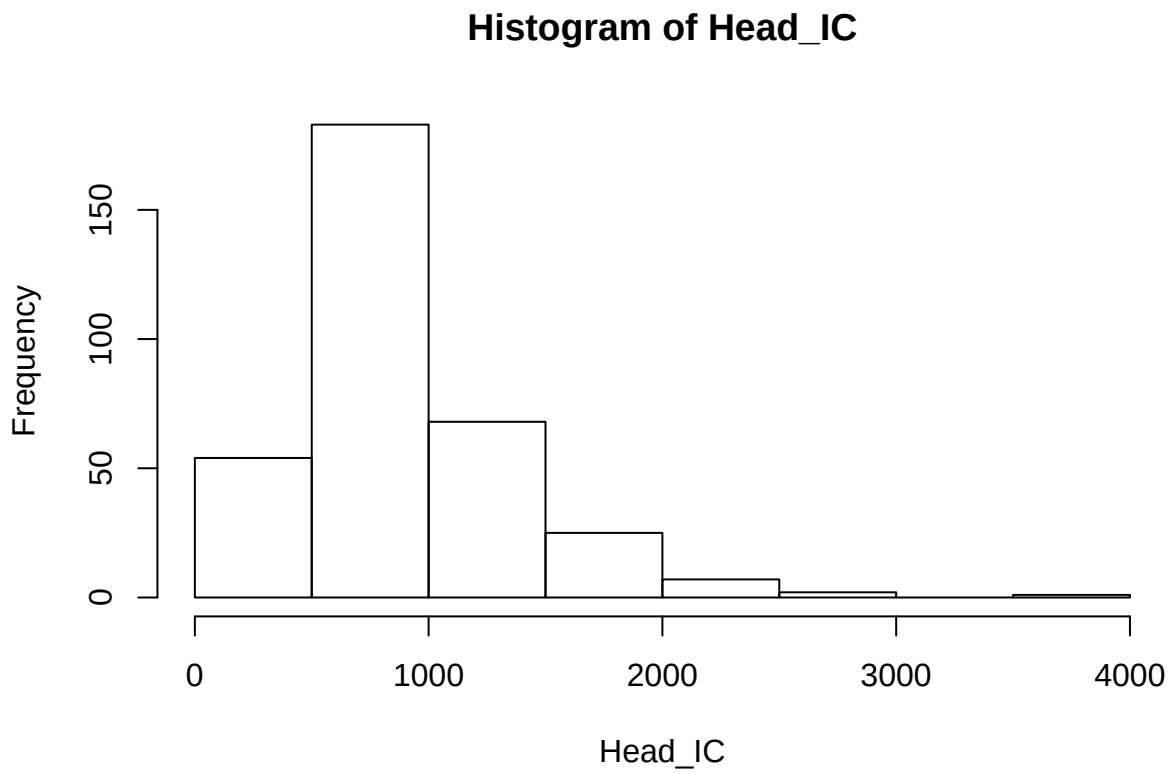
```
data = read.csv('crash_dat.csv',header = TRUE)

attach(data)

library('plyr')
```

```
## Warning: package 'plyr' was built under R version 3.4.4
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.4
## Loading required package: carData
## Warning: package 'carData' was built under R version 3.4.4
```

```
hist(Head_IC)
```



Checking the parameters for Initial screening:

```
count(make)
```

```
##             x freq
## 1       Acura    6
## 2        Audi    4
## 3         BMW    2
## 4       Buick    8
## 5    Cadillac    2
## 6   Chevrolet   42
## 7    Chrysler   12
## 8    Daihatsu    2
## 9       Dodge   16
## 10      Eagle    4
## 11       Ford   38
## 12        Geo    8
## 13      Honda   14
## 14    Hyundai   10
## 15   Infiniti    2
## 16      Isuzu   16
## 17       Jeep    8
## 18      Lexus    2
## 19    Lincoln    4
## 20      Mazda   12
## 21   Mercedes    2
## 22    Mercury    6
## 23 Mitsubishi   14
## 24     Nissan   32
## 25 Oldsmobile    4
## 26    Peugeot    6
## 27   Plymouth    8
## 28    Pontiac    8
## 29    Renault    2
## 30       Saab    4
## 31     Saturn    2
## 32     Subaru    6
## 33     Suzuki    4
## 34     Toyota   28
## 35 Volkswagen   10
## 36      Volvo    2
## 37       Yugo    2
```

```
model = glm(Head_IC~factor(make),family = Gamma(link="log"))
Anova(model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Head_IC
##              LR Chisq Df Pr(>Chisq)
## factor(make)   107.22 36  5.299e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Make is a factor which has 37 levels, not all of them are equally balanced, we might need to combine some of them based a criteria like country of origin instead.

P-value is significant indicating that this is an important variable for the model.

```
count(DP)
```

```
##        x freq
## 1 Driver  176
## 2 Passen  176
```

```
model = glm(Head_IC~factor(DP),family = Gamma(link="log"))
Anova(model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Head_IC
##            LR Chisq Df Pr(>Chisq)
## factor(DP)   8.9085  1   0.002838 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

DP is factor with two levels(binary), p-value suggests that this is an important variable for the model

```
count(Protection)
```

```
##                 x freq
## 1        d airbag   60
## 2     d&p airbags    4
## 3    manual belts  196
## 4 Motorized belts   44
## 5   passive belts   48
```

```
model = glm(Head_IC~factor(Protection),family = Gamma(link="log"))
Anova(model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Head_IC
##                     LR Chisq Df Pr(>Chisq)
## factor(Protection)   48.992  4   5.86e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Protection is factor with 5 levels, the categories are not equally balanced we can combine the 2nd level with the first one as it has only 4 observations. This variable passes the screening as well.

```
count(as.factor(Doors))
```

```
##      x freq
## 1    2  118
## 2    4  168
## 3 <NA>   66
```

```
model = glm(Head_IC~factor(Doors),family = Gamma(link="log"))
Anova(model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Head_IC
##               LR Chisq Df Pr(>Chisq)
## factor(Doors)   1.3837  1     0.2395
```

Doors variable is coerced as factor with 2 levels, initial screening reveals that we can reject this predictor

from the model.

```
count(as.factor(Year))
```

```
##    x freq
## 1 87   76
## 2 88   74
## 3 89   70
## 4 90   68
## 5 91   64
```

```
model = glm(Head_IC~factor(Year),family = Gamma(link="log"))
Anova(model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Head_IC
##              LR Chisq Df Pr(>Chisq)
## factor(Year)   6.1776  4     0.1863
```

Year, is coerced as factor which has 5 levels, they look fairly balanced and initial screening reveals that it barely makes it through and can be considered for the model.

```
model = glm(Head_IC~Wt,family = Gamma(link="log"))
Anova(model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Head_IC
##    LR Chisq Df Pr(>Chisq)
## Wt   14.466  1  0.0001427 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

WT can be included in the analysis as per the initial screening.

```
count(Size)
```

```
##      x freq
## 1 comp   86
## 2  hev   16
## 3   lt   74
## 4  med   62
## 5 mini   14
## 6  mpv   34
## 7   pu   36
## 8  van   30
```

```
model = glm(Head_IC~factor(Size),family = Gamma(link="log"))
Anova(model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Head_IC
##              LR Chisq Df Pr(>Chisq)
## factor(Size)    54.32  7  2.033e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Size is a factor predictor, which passes our screening test.

Data Manipulation:

1. We combine the 'make' of the car company based on country of origin. (Done manually in excel)
2. Protection airbags levels are combined. (Done manually in excel)
3. All the NA's are removed from the Data frame (R command used below)

```
data = read.csv('crash.csv',header = T)
data=data[complete.cases(data),]
attach(data)
```

```
## The following objects are masked from data (pos = 6):
##
##     Doors, DP, Head_IC, make, Protection, Size, Wt, Year
```

```
summary(data)
```

```
##       make          DP                Protection      Doors
##  America:128   Driver:142   d airbag       : 58   Min.   :2.000
##  Japan  :111   Passen:132   manual belts   :129   1st Qu.:2.000
##  Other  : 35                Motorized belts: 43   Median :4.000
##                             passive belts  : 44   Mean   :3.168
##                                                   3rd Qu.:4.000
##                                                   Max.   :4.000
##
##       Year             Wt            Size       Head_IC
##  Min.   :87.00   Min.   :1590   comp   :83   Min.   : 157.0
##  1st Qu.:88.00   1st Qu.:2370   lt     :73   1st Qu.: 546.0
##  Median :89.00   Median :2790   med    :58   Median : 768.5
##  Mean   :88.96   Mean   :2811   mpv    :32   Mean   : 846.4
##  3rd Qu.:90.00   3rd Qu.:3182   hev    :14   3rd Qu.:1001.8
##  Max.   :91.00   Max.   :5619   mini   :14   Max.   :2482.0
##                                 (Other): 0
```

Fitting the Final model:

```
library(boot)
```

```
## Warning: package 'boot' was built under R version 3.4.4
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
model = glm(Head_IC~factor(DP)+factor(Protection)+factor(Size),family = Gamma(link="log"),data = data)
anova(model)
```

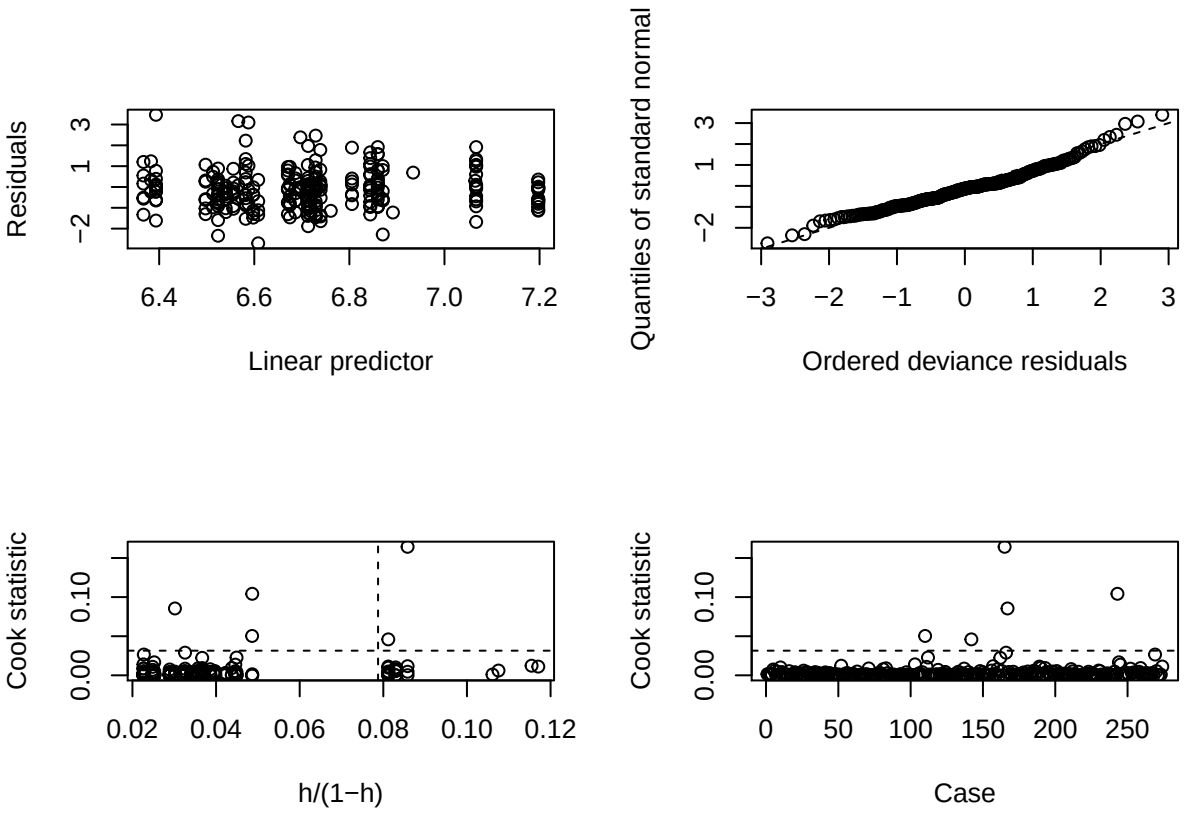
```
## Analysis of Deviance Table
##
## Model: Gamma, link: log
##
## Response: Head_IC
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev
```

```
## NULL                                  273     59.469
## factor(DP)           1   0.9842       272     58.485
## factor(Protection)   3   7.0991       269     51.386
## factor(Size)         5   3.5612       264     47.825
```

```
summary(model)
```

```
##
## Call:
## glm(formula = Head_IC ~ factor(DP) + factor(Protection) + factor(Size),
##     family = Gamma(link = "log"), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.23598  -0.33036  -0.05933   0.14095   1.54192
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        6.49754    0.08692  74.756  < 2e-16 ***
## factor(DP)Passen                  -0.13063    0.05579  -2.342 0.019949 *
## factor(Protection)manual belts     0.34640    0.08695   3.984 8.77e-05 ***
## factor(Protection)Motorized belts  0.17318    0.09914   1.747 0.081849 .
## factor(Protection)passive belts    0.21505    0.09835   2.187 0.029654 *
## factor(Size)hev                    0.22083    0.14050   1.572 0.117190
## factor(Size)lt                     0.01581    0.07496   0.211 0.833096
## factor(Size)med                    0.02628    0.08302   0.317 0.751806
## factor(Size)mini                  -0.03850    0.13997  -0.275 0.783499
## factor(Size)mpv                    0.35337    0.10513   3.361 0.000891 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.2125205)
##
##     Null deviance: 59.469  on 273  degrees of freedom
## Residual deviance: 47.825  on 264  degrees of freedom
## AIC: 3963.9
##
## Number of Fisher Scoring iterations: 6
```

```
glm.diag.plots(model)
```



Looking at the diagnostic plots we can see that, they conform to our model assumptions

1. Residual show an even spread about the x=0 line
2. Deviance residuals follow the straights line thereby confirming the assumptions
3. Cooks distance, does'nt show any cause of concern.

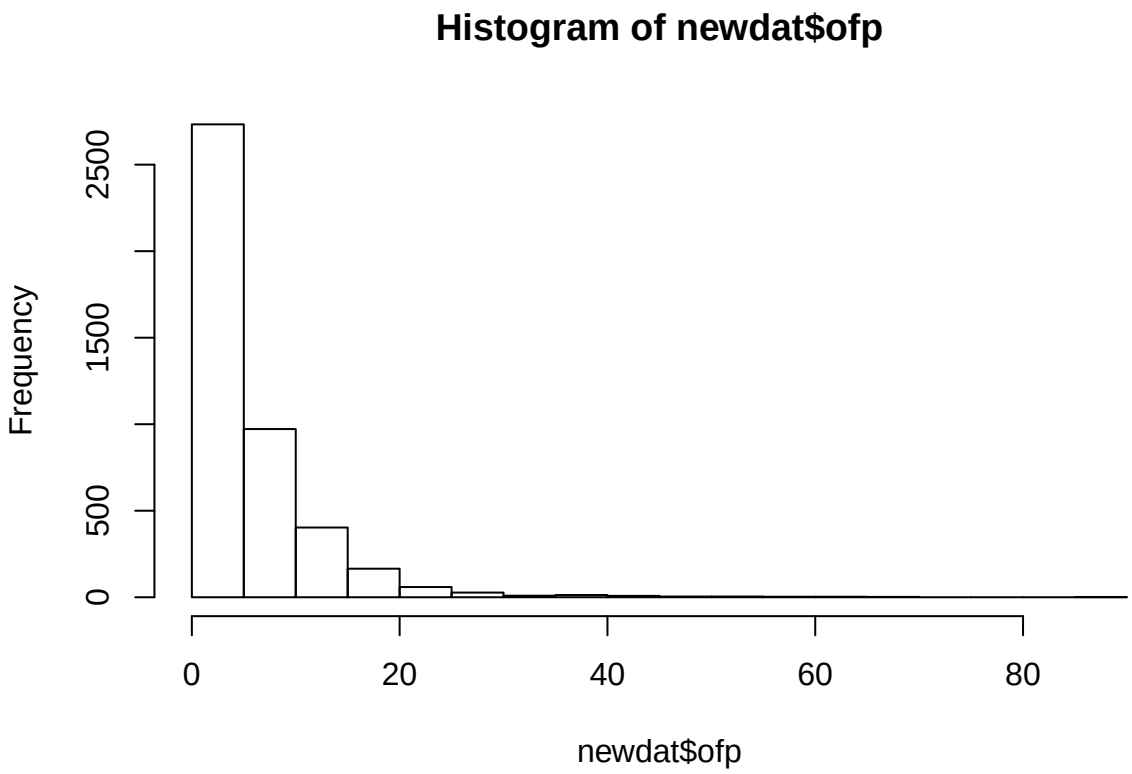
Model equation

$log(HIC) = 6.49754 - 0.13063factor(DP)Passenger + 0.34640factor(Protection)manualbelts + 0.17318factor(Protection)Motorizedbelts + 0.21505factor(Passivebelts) + 0.22083factor(Hev) + 0.01581factor(lt) + 0.02628factor(med) - 0.03850factor(mini) + 0.35337factor(mpv)$

Model Interpration: log(HIC) decreases by a factor of 0.013063 when person is sitting as a passenger which is expected, we can also see that the lighter car reduce the chances of HIC by a factor of 0.03850

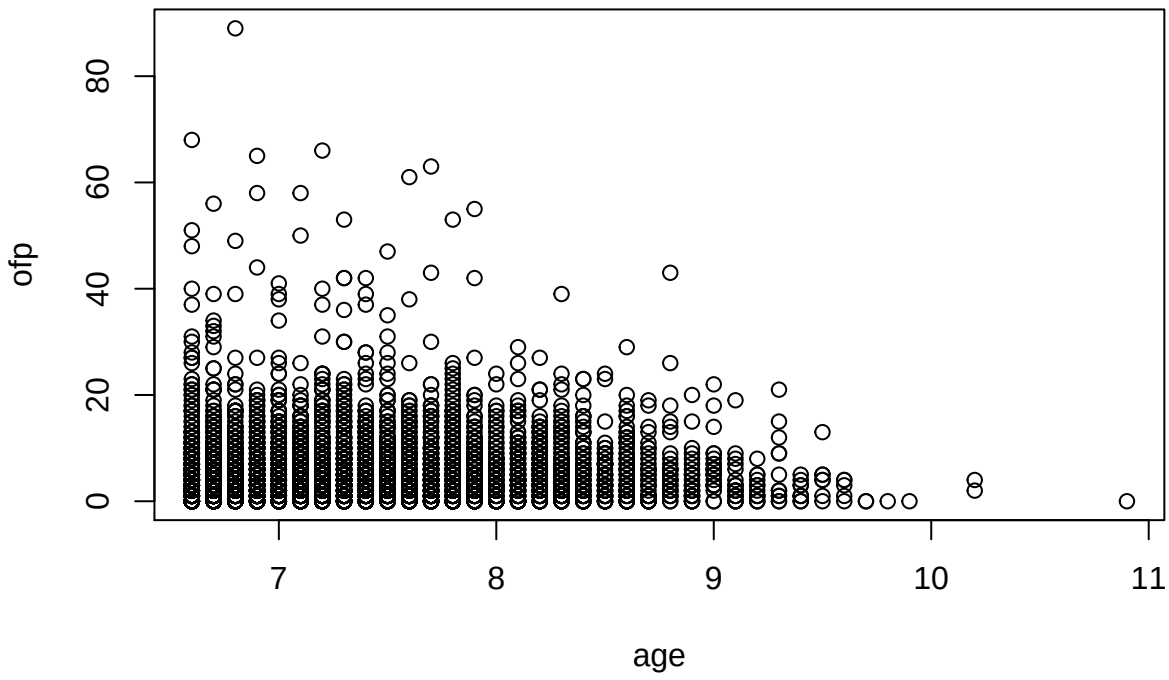
Question 2:

```
newdat = read.csv('ofp.csv', header = T)

hist(newdat$ofp)
```



Checking the relationship of reponse variable "ofp with the various predictors"

```
plot(ofp ~age,data=newdat)
```



We can see a negative trend here, mainly because of the human life expectancy.

Let's try to fit a poisson model to model the counts of the clinic visits explained by age.

```
model_2 = glm(ofp ~ age, family = poisson(link = "log"),data = newdat)
summary(model_2)
```

```
##
## Call:
## glm(formula = ofp ~ age, family = poisson(link = "log"), data = newdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4359  -2.4495  -0.7806   0.8747  17.9178
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.706946   0.073439  23.243   <2e-16 ***
## age         0.006279   0.009881   0.635    0.525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 26943  on 4405  degrees of freedom
## Residual deviance: 26943  on 4404  degrees of freedom
## AIC: 39722
##
```

```
## Number of Fisher Scoring iterations: 5
```

Looking at the residual deviance which is 26943 on 4404 degrees of freedom, we can confirm that this is a case of overdisperion. Therefore let's try to fit a negative binomial model instead.

```
library(MASS)
model_2= glm.nb(ofp~age,data = newdat)
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ age, data = newdat, init.theta = 0.9949490591,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9636  -1.1667  -0.3150   0.3199   4.6119
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.703414   0.191790   8.882   <2e-16 ***
## age         0.006756   0.025813   0.262    0.794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9949) family taken to be 1)
##
##     Null deviance: 5036.5  on 4405  degrees of freedom
## Residual deviance: 5036.4  on 4404  degrees of freedom
## AIC: 24992
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.9949
##           Std. Err.:  0.0260
##
##  2 x log-likelihood:  -24985.5950
```

```
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(0.9949), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                  4405     5036.5
## age   1 0.063718      4404     5036.4   0.8007
```

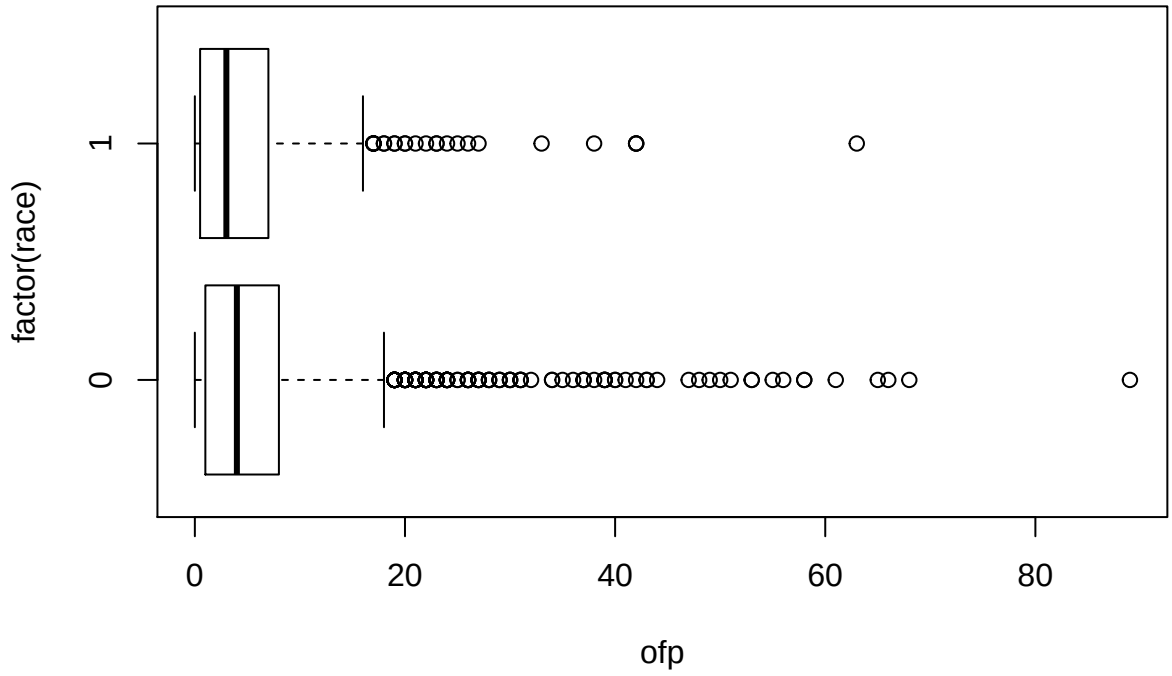
Looking at the above results we can see that now the residual deviance is now 5036 which is much more

comparable to the degrees of freedom i.e 4404.

Looking at the anova table we can see that this predictor does not pass initial screening for the parameter selection.

Let's look at the rest of parameters

```
plot(ofp ~ factor(race),data = newdat,horizontal=T)
```



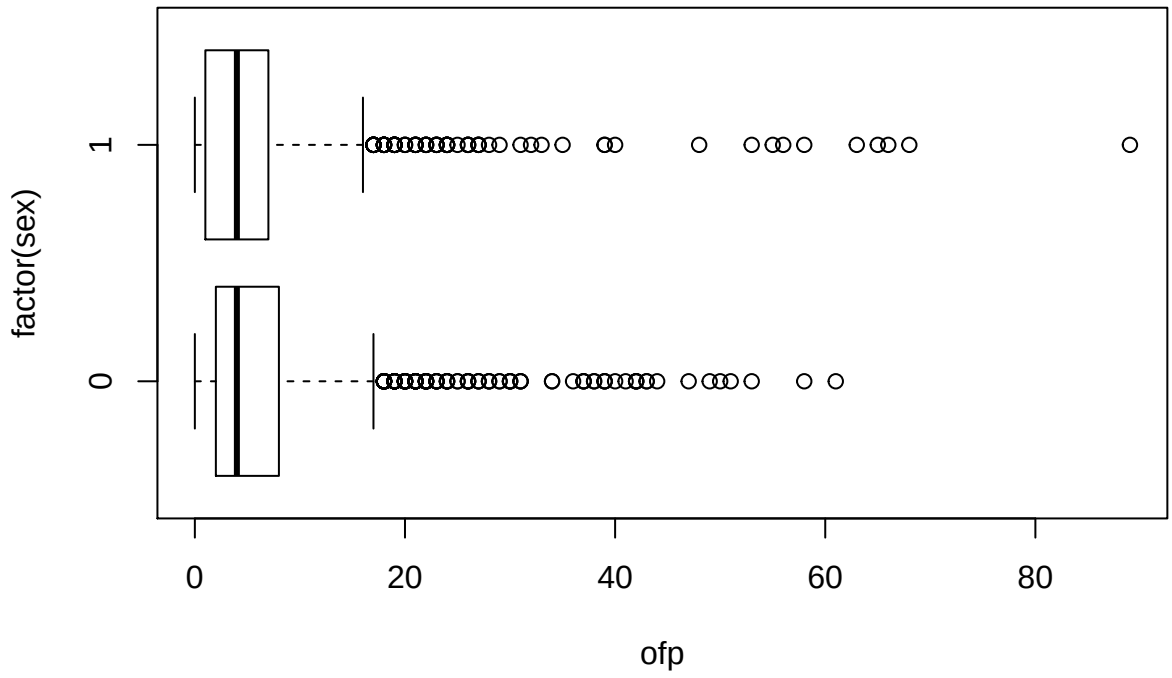

```
model_2 = glm.nb(ofp~factor(race),data = newdat)

anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(0.9979), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                          4405     5046.9
## factor(race)  1   10.372      4404     5036.5 0.001279 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There's isn't much disparity in number of visits based on race, we can that race passes the prelimnary screening for paramter selection.

```
plot(ofp ~ factor(sex),data =newdat,horizontal=T)
```




```
model_2= glm.nb(ofp~factor(sex),data = newdat)
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ factor(sex), data = newdat, init.theta = 0.9977345548,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9725  -1.1326  -0.3494   0.2789   4.7772
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.79410    0.02109  85.082  < 2e-16 ***
## factor(sex)1 -0.10398    0.03335  -3.118  0.00182 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9977) family taken to be 1)
##
##     Null deviance: 5046.2  on 4405  degrees of freedom
```

```
## Residual deviance: 5036.5  on 4404  degrees of freedom
## AIC: 24982
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.9977
##           Std. Err.:  0.0261
##
##  2 x log-likelihood:  -24975.9960
```

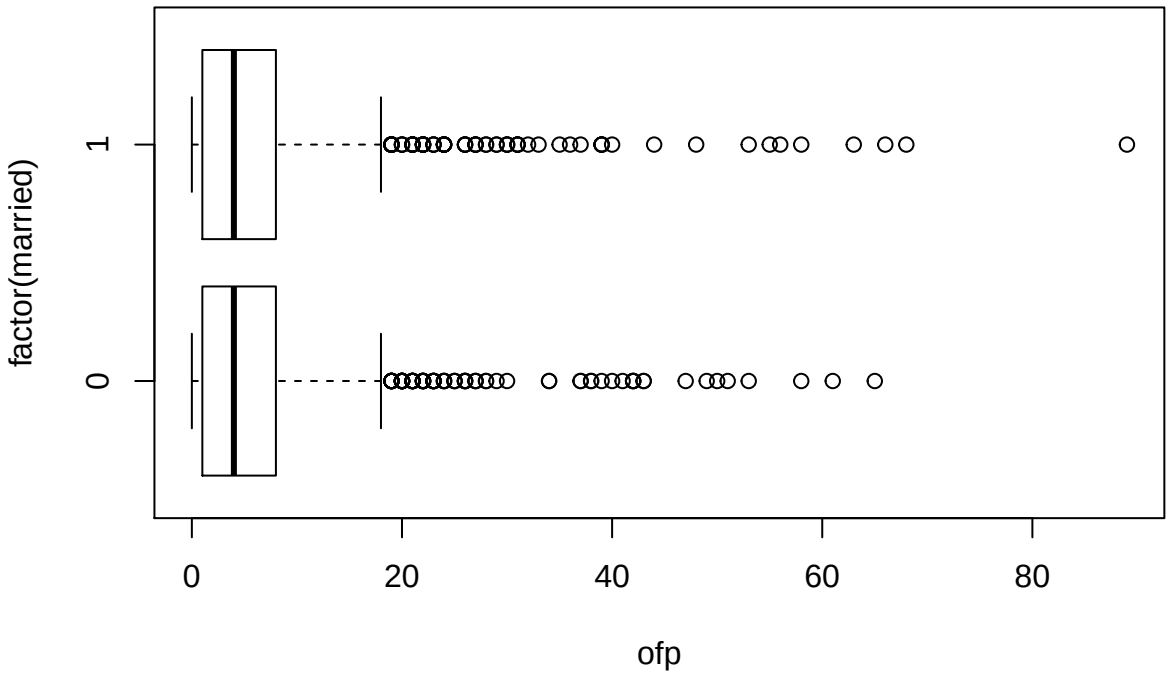
```
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(0.9977), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                         4405     5046.2
## factor(sex)  1   9.6746      4404     5036.5 0.001868 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similar trend is observer in this case as well, predictor passes the initial screening as p-value on 0.001868 indicates we can include it in the final model for further analysis.

```
plot(ofp ~ factor(married),data =newdat,horizontal=T)
```




```
model_2= glm.nb(ofp~factor(married),data = newdat)
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ factor(married), data = newdat, init.theta = 0.9955967073,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9654  -1.1561  -0.2967   0.3435   4.6635
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.78028    0.02422  73.511   <2e-16 ***
## factor(married)1 -0.04972    0.03283  -1.515     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9956) family taken to be 1)
##
##     Null deviance: 5038.7  on 4405  degrees of freedom
## Residual deviance: 5036.4  on 4404  degrees of freedom
## AIC: 24989
##
## Number of Fisher Scoring iterations: 1
```

```
##
##
##               Theta:  0.9956
##           Std. Err.:  0.0260
##
##  2 x log-likelihood:  -24983.3630
```

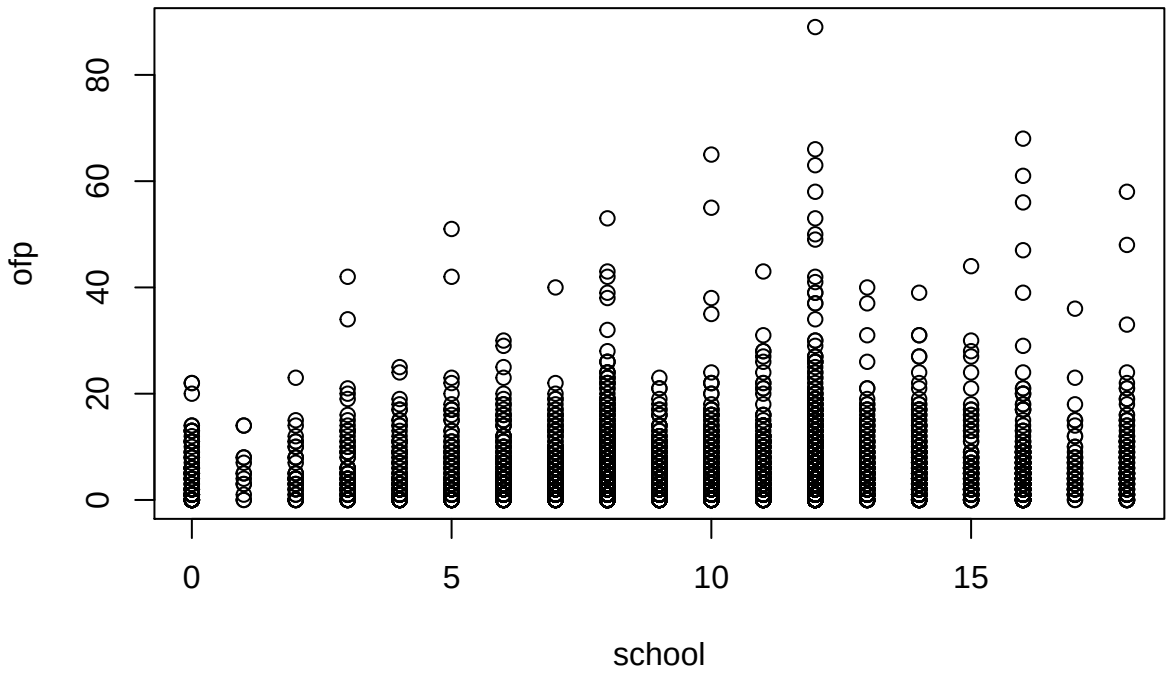
```
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(0.9956), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                             4405     5038.7
## factor(married)  1   2.2962      4404     5036.4   0.1297
```

Variable, passes the screening. We may not end up keeping this in the final model

```
plot(ofp ~school,data =newdat)
```



```r
model_2= glm.nb(ofp~school,data = newdat)
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ school, data = newdat, init.theta = 1.001162113,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0229  -1.1182  -0.3425   0.2934   4.5303
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.543139   0.048079  32.096  < 2e-16 ***
## school      0.020159   0.004377   4.606 4.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0012) family taken to be 1)
##
##     Null deviance: 5058.2  on 4405  degrees of freedom
## Residual deviance: 5036.8  on 4404  degrees of freedom
## AIC: 24970
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0012
##           Std. Err.:  0.0262
##
##  2 x log-likelihood:  -24964.2820
```

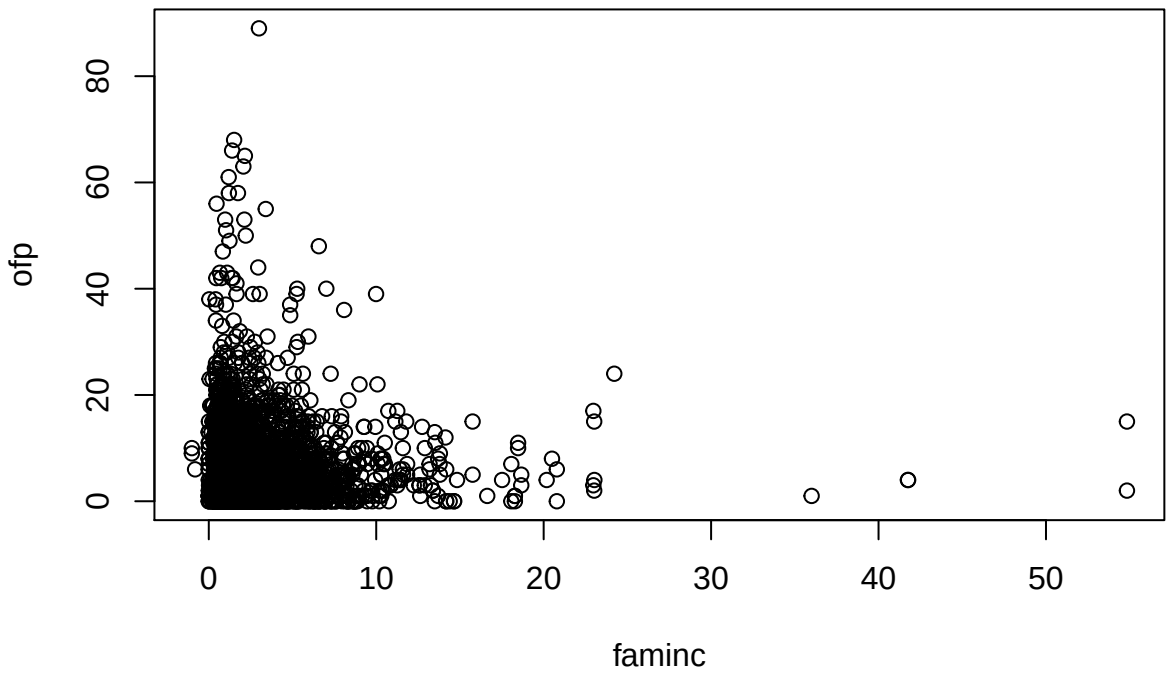
```r
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(1.0012), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    4405     5058.2
## school  1   21.434      4404     5036.8 3.663e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variable passes the screening.

```r
plot(ofp ~ faminc,data =newdat)
```




```
model_2= glm.nb(ofp~faminc,data = newdat)
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ faminc, data = newdat, init.theta = 0.9949658958,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9554  -1.1670  -0.3166   0.3181   4.6034
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.758174   0.021627  81.294   <2e-16 ***
## faminc      -0.001882   0.005606  -0.336    0.737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.995) family taken to be 1)
##
##     Null deviance: 5036.5  on 4405  degrees of freedom
## Residual deviance: 5036.4  on 4404  degrees of freedom
## AIC: 24992
##
## Number of Fisher Scoring iterations: 1
```

```
##
##
##               Theta:  0.9950
##           Std. Err.:  0.0260
##
##  2 x log-likelihood:  -24985.5390
```

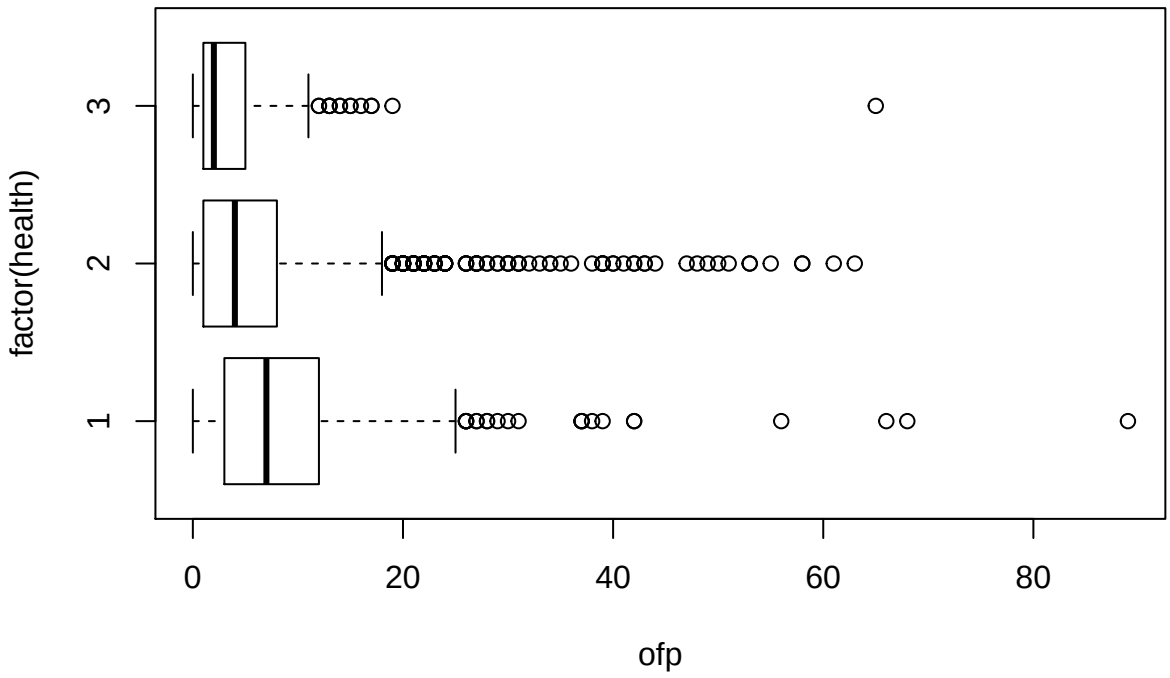
```
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(0.995), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    4405     5036.5
## faminc  1  0.11976      4404     5036.4   0.7293
```

Family income does not pass the initial screening test.

```
plot(ofp ~ factor(health),data =newdat,horizontal=T)
```



```
model_2= glm.nb(ofp~factor(health),data = newdat)
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ factor(health), data = newdat, init.theta = 1.049074119,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1724  -1.1641  -0.2829   0.3657   5.1938
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.18573    0.04386   49.84   <2e-16 ***
## factor(health)2 -0.47904    0.04740  -10.11   <2e-16 ***
## factor(health)3 -0.95358    0.07452  -12.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0491) family taken to be 1)
##
##     Null deviance: 5223.7  on 4405  degrees of freedom
## Residual deviance: 5040.3  on 4403  degrees of freedom
## AIC: 24814
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0491
##           Std. Err.:  0.0279
##
##  2 x log-likelihood:  -24806.3220
```

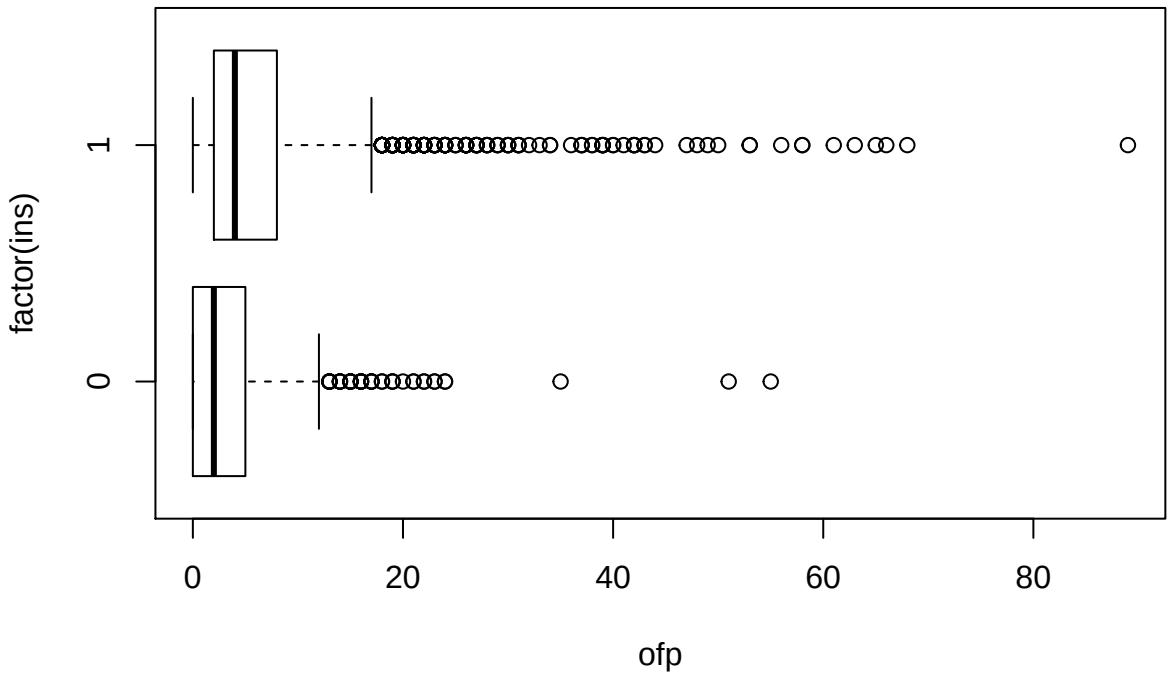
```
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(1.0491), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             4405     5223.7
## factor(health)  2   183.45      4403     5040.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This passes our initial screening.

```r
plot(ofp ~ factor(ins),data =newdat,horizontal=T)
```




```r
model_2= glm.nb(ofp~factor(ins),data = newdat)
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ factor(ins), data = newdat, init.theta = 1.015655441,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9868  -0.9529  -0.3604   0.2708   4.5096
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.39094    0.04376  31.783   <2e-16 ***
## factor(ins)1  0.41323    0.04712   8.769   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0157) family taken to be 1)
##
##     Null deviance: 5108.6  on 4405  degrees of freedom
## Residual deviance: 5036.4  on 4404  degrees of freedom
## AIC: 24920
```

```
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0157
##           Std. Err.:  0.0267
##
##  2 x log-likelihood:  -24914.0350
```

```
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(1.0157), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         4405     5108.6
## factor(ins)  1   72.246      4404     5036.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Insurance status also passes our screening test.

Fitting the full model with the predictors which passed the test.

```
model_2= glm.nb(ofp ~  factor(race)+  factor(sex)+ factor(married)+school+factor(health)+factor(ins),da
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ factor(race) + factor(sex) + factor(married) +
##     school + factor(health) + factor(ins), data = newdat, init.theta = 1.09287414,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3404  -0.9822  -0.3112   0.3020   5.3348
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.692867   0.071033  23.832  < 2e-16 ***
## factor(race)1    -0.104917   0.052331  -2.005   0.0450 *
## factor(sex)1     -0.066389   0.035147  -1.889   0.0589 .
## factor(married)1 -0.055224   0.035145  -1.571   0.1161
## school            0.027287   0.004517   6.040 1.54e-09 ***
## factor(health)2  -0.545149   0.047226 -11.543  < 2e-16 ***
## factor(health)3  -1.064858   0.074534 -14.287  < 2e-16 ***
## factor(ins)1      0.380292   0.047473   8.011 1.14e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for Negative Binomial(1.0929) family taken to be 1)
##
##     Null deviance: 5371.9  on 4405  degrees of freedom
## Residual deviance: 5040.1  on 4398  degrees of freedom
## AIC: 24685
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0929
##           Std. Err.:  0.0294
##
##  2 x log-likelihood:  -24666.8020
```

```
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(1.0929), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                             4405     5371.9
## factor(race)     1   11.196      4404     5360.7 0.0008197 ***
## factor(sex)      1   11.796      4403     5348.9 0.0005936 ***
## factor(married)  1    0.440      4402     5348.5 0.5072579
## school           1   19.156      4401     5329.3 1.204e-05 ***
## factor(health)   2  227.977      4399     5101.3 < 2.2e-16 ***
## factor(ins)      1   61.217      4398     5040.1 5.112e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected looks like we can drop the predictor married for our predictive model.We can also drop Race & Sex for the sake of parimony of the model.

Final model:

```
model_2= glm.nb(ofp ~ school+factor(health)+factor(ins),data=newdat)
summary(model_2)
```

```
##
## Call:
## glm.nb(formula = ofp ~ school + factor(health) + factor(ins),
##     data = newdat, init.theta = 1.088554318, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3344  -0.9789  -0.3121   0.3012   5.3250
##
## Coefficients:
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.60851    0.06580  24.447  < 2e-16 ***
## school           0.02766    0.00439   6.301 2.96e-10 ***
## factor(health)2 -0.54568    0.04729 -11.540  < 2e-16 ***
## factor(health)3 -1.06611    0.07460 -14.290  < 2e-16 ***
## factor(ins)1     0.39601    0.04687   8.449  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0886) family taken to be 1)
##
##     Null deviance: 5357.4  on 4405  degrees of freedom
## Residual deviance: 5039.7  on 4401  degrees of freedom
## AIC: 24692
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0886
##           Std. Err.:  0.0293
##
##  2 x log-likelihood:  -24679.8350
```

```
anova(model_2)
```

```
## Warning in anova.negbin(model_2): tests made without re-estimating 'theta'
```

```
## Analysis of Deviance Table
##
## Model: Negative Binomial(1.0886), link: log
##
## Response: ofp
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           4405     5357.4
## school          1   23.009      4404     5334.4 1.612e-06 ***
## factor(health)  2  226.902      4402     5107.5 < 2.2e-16 ***
## factor(ins)     1   67.770      4401     5039.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final model: $log(ofp) = 1.60851 + 0.02766School - 0.54568factor(health)2 - 1.06611factor(health)3 + 0.39601factor(ins)1$

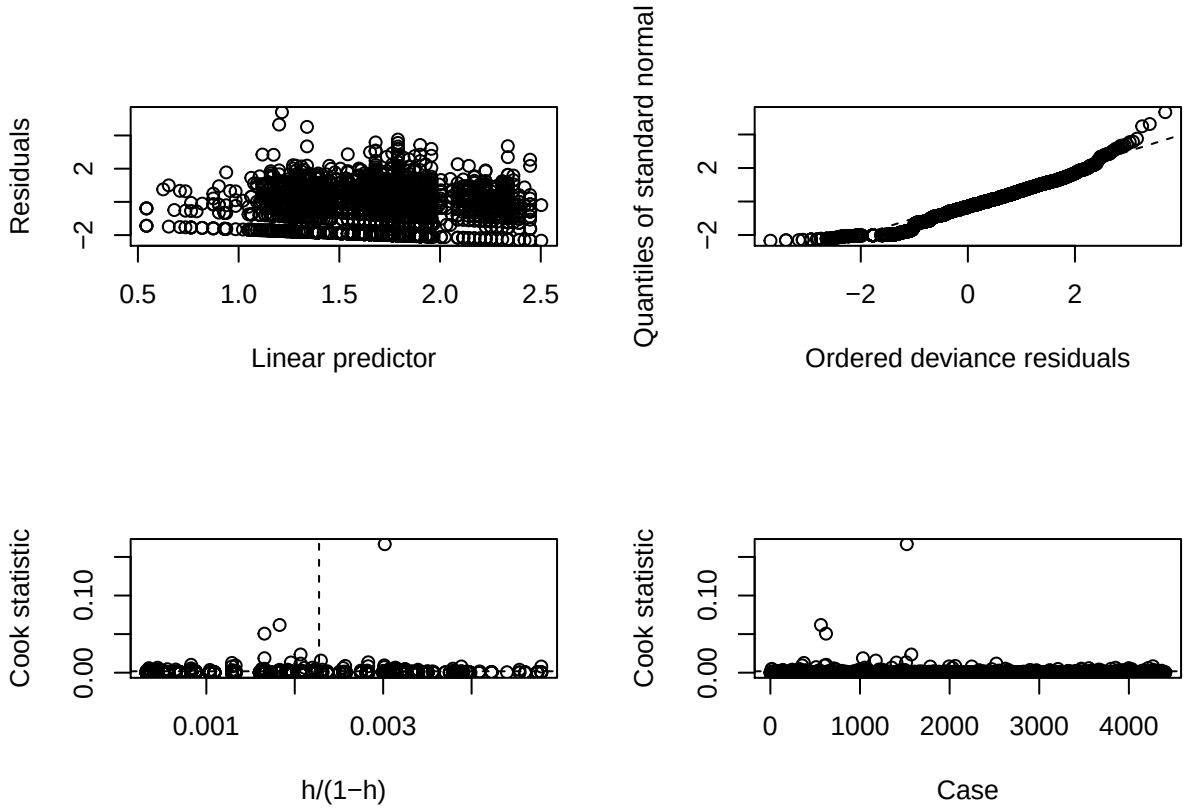
Model interpertation: Log of Visits to physician's clinic increases by a factor of 0.02766 with each year spent at school, thereby educated people are more aware are likely to visit to the doctor as expected.

If the self perceived state of health is moderate or good, it reduces the factor of visit by 0.54568 and 1.066 as expected. People will not go to they the doctor if they think they are healthy. This comparison is drawn with respect to Health factor level 1 i.e poor health.

If people are covered by medical insuarance, they are more likely to visit the physician's clinic as it increass the log(ofp) by 0.39601

Diagnostic plots

```
glm.diag.plots(model_2)
```



Looking at the above diagnostic plots we can see that ordered devaince residuals follow a linear trend and cook's distance/leverage plots show we have nothing much to worry about. Residual spread is fairly random and spread evenly.

part b)

Introducing a new variable in the data frame.

```
newdat$visit <- ifelse(newdat$ofp >0, 1, 0)
```

Using a logistic regression model to predict the using using same predictors as part 1

```
model_3 = glm( visit ~ factor(sex)+school+factor(health)+factor(ins), family = binomial,data = newdat)
summary(model_3)
```

```
##
## Call:
## glm(formula = visit ~ factor(sex) + school + factor(health) +
##     factor(ins), family = binomial, data = newdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4714   0.4287   0.4994   0.5825   1.4003
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)      1.04467    0.17808   5.866 4.46e-09 ***
## factor(sex)1    -0.34620    0.08552  -4.048 5.16e-05 ***
## school           0.06434    0.01133   5.679 1.36e-08 ***
## factor(health)2 -0.60066    0.14934  -4.022 5.77e-05 ***
## factor(health)3 -1.20865    0.19395  -6.232 4.61e-10 ***
## factor(ins)1     0.93157    0.10255   9.084  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3800.7  on 4405  degrees of freedom
## Residual deviance: 3621.2  on 4400  degrees of freedom
## AIC: 3633.2
##
## Number of Fisher Scoring iterations: 4
```

```
anova(model_3)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: visit
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev
## NULL                           4405     3800.7
## factor(sex)     1   20.244      4404     3780.5
## school          1   43.909      4403     3736.6
## factor(health)  2   38.502      4401     3698.1
## factor(ins)     1   76.817      4400     3621.2
```

Model equation: log {p(visit)/1-p(visit)}= 1.04467 -0.34620factor(sex)1 + 0.06434School -0.60066factor(health)2 - 1.20865factor(health)3 + 0.93157factor(ins)1

Model interperation: Insurance coverage enhances the log ratio on the lhs by a factor of 0.93157, as expected, if people are covered by the medical insurance they are more likely to visit the physician's clinic.

Similarly we can see the effect of self perceived health state and other predictors from the above table.

Let's check the goodness of the fit of the model:

```
rs= fitted(model_3)
table(newdat$visit,rs>=0.5)
```

```
##
##     FALSE TRUE
##   0     4  679
##   1     7 3716
```

As we can see our model correctly predicted whether the person is going to visit the clinic for majority of the dataset. We can also see a lot of false postives.

Let's look at the ROC curve for our model:

```
library(pROC)
```

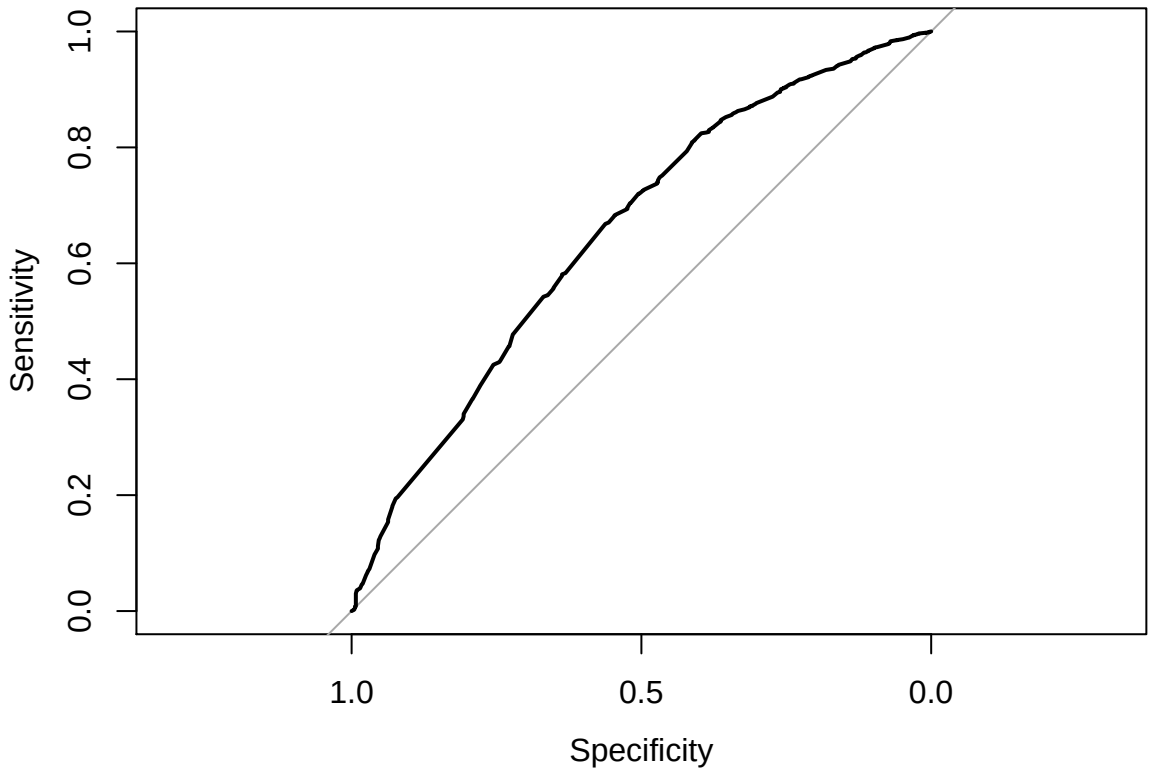
```
## Warning: package 'pROC' was built under R version 3.4.4
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc(newdat$visit,rs,plot=TRUE)
```




```
##
## Call:
## roc.default(response = newdat$visit, predictor = rs, plot = TRUE)
##
## Data: rs in 683 controls (newdat$visit 0) < 3723 cases (newdat$visit 1).
## Area under the curve: 0.6524
```

This model gives us an AUC of 0.6524, which is moderately than random guess or a coin toss.