Homework 3: Binary and Count Regressions

Solution Set

This homework set will cover problems concerning binary and count regression models. Point totals for specific problems are given, 10 points will be reserved for correct submission of the homework assignment.

Problem 1 [10 points]: Manually write your own Fisher scoring algorithm which maximizes the logistic regression log likelihood for the CCSO example in the notes. Report $\hat{\beta}$ and reproduce the summary table (up to convergence tolerance differences) without using the `glm` or `summary` commands. You can ignore deviance residuals.

Solution 1:

The log-likelihood is

$$l(\beta) = \sum_{i=1}^{n} y_i x_i^T \beta - \log(1 + exp(x_i^T \beta))$$

$$l'(\beta) = \sum_{i=1}^{n} \left(y_i x_i - \frac{x_i}{1 + exp(x_i^T \beta)} exp(x_i^T \beta) \right)$$

$$\implies l'(\beta) = X^T (Y - \pi)$$

where $\pi_i = \frac{exp(x_i^T \beta)}{1+exp(x_i^T \beta)}$

$$l''(\beta) = - \sum_{i=1}^{n} \left(\frac{x_i^2}{(1 + exp(x_i^T \beta))^2} exp(x_i^T \beta) \right)$$

$$\implies l''(\beta) = -X^T W X$$

where $W = diag(\pi_i(1 - \pi_i))$

Thus the Fisher scoring algorithm:

$$\beta^{(t+1)} = \beta^{(t)} + (X^T W^{(t)} X)^{-1} X^T (Y - \pi)$$

```
# Reading in the data
library(data.table)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.4.1     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
CCSO = fread("https://uofi.box.com/shared/static/9elozjsg99bgcb7gb546wlfr3r2gc9b7.csv")
```

```r
# Pre-Processing the data
CCSO_small <- CCSO %>% rename(Days = "Days in Jail", Age = "Age at Arrest",
Date = "BOOKING DATE", Sex = "SEX", Race = "RACE",
Crime = "CRIME CODE") %>%
mutate(atleastone = ifelse(Days > 0,1,0)) %>%
filter(Crime == "OTHER TRAFFIC OFFENSES") %>%
filter(Race %in% c("Asian/Pacific Islander","Black","White","Hispanic")) %>%
filter(Sex %in% c("Female","Male")) %>%
dplyr::select(atleastone, Age, Sex, Date, Race) %>%
mutate(Race = fct_drop(Race), Sex = fct_drop(Sex))
CCSO_small <- CCSO_small[complete.cases(CCSO_small), ]
head(CCSO_small)
```

```
##    atleastone Age    Sex     Date     Race
## 1:          0  22   Male 1/1/2011    White
## 2:          0  26   Male 1/1/2011    White
## 3:          0  32 Female 1/1/2011    White
## 4:          0  22   Male 1/2/2011    White
## 5:          0  35   Male 1/2/2011 Hispanic
## 6:          0  35   Male 1/2/2011 Hispanic
```

```r
#Creating the model matrix
X = model.matrix(atleastone ~ -1 + Race + Sex + Age,data = CCSO_small)
n = nrow(X)
p = ncol(X)
Y = CCSO_small$atleastone
```

```r
# Initializing the beta
beta = matrix(rep(0,6))

#Running the Fisher scoring iterations
for(t in 1:10)
{
  pi =  exp(X%*%beta)/(1+exp(X%*%beta))
  W = diag(c(pi*(1-pi)))
  beta = beta + solve(t(X) %*% W %*% X)%*%t(X)%*%(Y - pi)
}
```

```r
# The final values
pi_CCSO =  exp(X%*%beta)/(1+exp(X%*%beta))
W_CCSO = diag(c(pi_CCSO*(1-pi_CCSO)))
var_matrix_CCSO = solve(t(X) %*% W_CCSO %*% X)
sd_beta = sqrt(diag(var_matrix_CCSO))
z_Val = beta/sd_beta
pvalue = 2*(1 - pnorm(abs(z_Val)))
```

```r
#Deviance results
deviance_res = -2*(t(Y)%*%X%*%beta - sum(log(1+exp(X%*%beta))))
beta_0 = 0
deviance_null =  -2*(beta_0*sum(Y) - n*log(1+exp(beta_0)))
AIC = deviance_res + 2*p
```

The summary table:

```r
tab = data.frame("Estimate" = beta,"Std.Error" = sd_beta,"z value" = z_Val, "Pvalue" = pvalue)
list("Coefficients" = tab, "Null deviation" = deviance_null, "Residual Deviance" = deviance_res, "Null d
```

```
## $Coefficients
##                                Estimate   Std.Error    z.value       Pvalue
## RaceAsian/Pacific Islander -4.38986549 0.523613419  -8.383791 0.000000e+00
## RaceBlack                  -1.87654964 0.144600944 -12.977437 0.000000e+00
## RaceHispanic               -2.80454861 0.173349092 -16.178617 0.000000e+00
## RaceWhite                  -3.04322627 0.147160127 -20.679693 0.000000e+00
## SexMale                     0.73983403 0.105379772   7.020646 2.208456e-12
## Age                         0.00770504 0.003186262   2.418207 1.559721e-02
##
## $`Null deviation`
## [1] 8201.317
##
## $`Residual Deviance`
##          [,1]
## [1,] 4668.728
##
## $`Null df`
## [1] 5916
##
## $`Residual df`
## [1] 5910
##
```

```
## $AIC
##          [,1]
## [1,] 4680.728
```

Problem 2 [10 points]: Manually write your own Fisher scoring algorithm which maximizes the Poisson regression log likelihood for the Galapagos example in the notes. Report $\hat{\beta}$ and reproduce the summary table (up to convergence tolerance differences) without using the `glm` or `summary` commands. You can ignore deviance residuals.

Solution 2:

The log-likelihood is

$$l(\beta) = \sum_{i=1}^{n} \left(y_i x_i^T \beta - exp(x_i^T \beta) \right)$$

$$l'(\beta) = \sum_{i=1}^{n} \left(y_i x_i - x_i exp(x_i^T \beta) \right)$$

$$\implies l'(\beta) = X^T(Y - \pi)$$

where $\pi_i = exp(x_i^T \beta)$

$$l''(\beta) = -\sum_{i=1}^{n} x_i^2 exp(x_i^T \beta)$$

$$\implies l''(\beta) = -X^T W X$$

where $W = diag(\pi_i)$

Thus the Fisher scoring algorithm:

$$\beta^{(t+1)} = \beta^{(t)} + (X^T W^{(t)} X)^{-1} X^T (Y - \pi)$$

```
library(faraway)
```

```
#Pre-processing the data
data(gala)
gala <- gala %>%
mutate(Size = as.factor(1 + ifelse(Area > 1,1,0) + ifelse(Area > 25,1,0)))
head(gala)
```

```
##              Species Endemics  Area Elevation Nearest Scruz Adjacent Size
## Baltra            58       23 25.09       346     0.6   0.6     1.84    3
## Bartolome         31       21  1.24       109     0.6  26.3   572.33    2
## Caldwell           3        3  0.21       114     2.8  58.7     0.78    1
## Champion          25        9  0.10        46     1.9  47.4     0.18    1
## Coamano            2        1  0.05        77     1.9   1.9   903.82    1
## Daphne.Major      18       11  0.34       119     8.0   8.0     1.84    1
```

```r
# Creating the model matrix
X = model.matrix(Species ~ Elevation + Nearest + Scruz + Adjacent + Size,data = gala)
n = nrow(X)
p = ncol(X)
Y = gala$Species

# Initialising the beta
beta = c(0,apply(X,2,mean)[-1]/apply(X,2,var)[-1])

#Running the fisher scoring iterations
for(t in 1:10)
{
  pi =  exp(X%*%beta)
  W = diag(c(pi))
  beta = beta + solve(t(X) %*% W %*% X)%*%t(X)%*%(Y - pi)
}
```

```r
# The final values
pi =  exp(X%*%beta)
W = diag(c(pi))

var_matrix = solve(t(X) %*% W %*% X)
sd_beta = sqrt(diag(var_matrix))
z_Val = beta/sd_beta
pvalue = 2*(1 - pnorm(abs(z_Val)))
```

```r
#Deviance results

deviance_res = -2*(t(Y)%*%X%*%beta - sum(pi)+ sum(Y - Y*log(Y)))
beta_0 = log(mean(Y))
deviance_null =  -2*(beta_0*sum(Y) - n*exp(beta_0) + sum(Y - Y*log(Y)))
AIC = -2*sum(dpois(Y,pi,log = TRUE)) + 2*(p)
```

The summary table:

```r
tab = data.frame("Estimate" = beta,"Std.Error" = sd_beta,"z value" = z_Val, "Pvalue" = pvalue)
list("Coefficients" = tab, "Null deviation" = deviance_null, "Residual Deviance" = deviance_res, "Null d
```

```
## $Coefficients
##                  Estimate    Std.Error   z.value       Pvalue
## (Intercept)  2.7897964692 8.107802e-02 34.408787 0.0000000000
## Elevation    0.0009360990 5.402069e-05 17.328527 0.0000000000
## Nearest      0.0064693041 1.747557e-03  3.701912 0.0002139805
## Scruz       -0.0062664946 6.268336e-04 -9.997063 0.0000000000
## Adjacent    -0.0002857805 2.960795e-05 -9.652152 0.0000000000
## Size2        1.1276155415 9.535272e-02 11.825730 0.0000000000
## Size3        2.0586771315 9.419392e-02 21.855732 0.0000000000
##
## $`Null deviation`
## [1] 3510.729
##
## $`Residual Deviance`
```

```
##          [,1]
## [1,] 594.1753
##
## $`Null df`
## [1] 29
##
## $`Residual df`
## [1] 23
##
## $AIC
## [1] 769.0063
```

Problem 3 [5 points]: Derive the log-odds ratio of $x+1$ to x when $Y = 1$, and observe that the log-odds ratio does not depend on x. Comment on this finding.

Solution 3:

We know that in the logistic regression model we have

$$logit(\pi(x)) = x\beta \quad \text{where } \pi(x) = E(Y|X = x) = P(Y = 1|X = x)$$

Thus the log of their ratio

$$\begin{aligned}
\log\left(\frac{odds(Y = 1|X = x + 1)}{odds(Y = 1|X = x)}\right) &= log\left(\frac{P(Y = 1|X = x + 1)/P(Y = 0|X = x + 1)}{P(Y = 1|X = x + 1)/P(Y = 0|X = x + 1))}\right) \\
&= log\left(\frac{\pi(x + 1)/(1 - \pi(x + 1))}{\pi(x)/(1 - \pi(x))}\right) \\
&= log\left(\pi(x + 1)/(1 - \pi(x + 1))\right) - log\left(\pi(x)/(1 - \pi(x))\right) \\
&= logit(\pi(x + 1)) - logit(\pi(x)) \\
&= (x + 1)\beta - x\beta \\
&= \beta
\end{aligned}$$

Thus it does not depend on X. This shows that a unit increase in X leads to an increase of beta in the log-odds ratio i.e a unit increase in x leads to the ratio of their odds increasing by e^{β}

Problem 4 [10 points]: Complete the following parts:

(a) Explain important findings and model information from the summary table produced by a call to summary(m1) in the CCSO example in the logistic regression notes. Keep in mind that we restricted attention to "other traffic offenses" in the CCSO example, and that this data is observational.

(b) Explain important findings and model information from the summary table produced by a call to summary(m1) in the Galapagos islands example in the count regression notes.

Solution 4:

```r
# Creating the logistic regression model for the CCSO model
m1 <- glm(atleastone ~ -1 + Race + Sex + Age, data = CCSO_small,
family = "binomial", x = "TRUE")
summary(m1)
```

```
##
## Call:
## glm(formula = atleastone ~ -1 + Race + Sex + Age, family = "binomial",
##     data = CCSO_small, x = "TRUE")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9393  -0.5485  -0.4817  -0.3391   2.6527
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## RaceAsian/Pacific Islander -4.389865   0.523612  -8.384  < 2e-16 ***
## RaceBlack                  -1.876550   0.144601 -12.977  < 2e-16 ***
## RaceHispanic               -2.804549   0.173349 -16.179  < 2e-16 ***
## RaceWhite                  -3.043226   0.147160 -20.680  < 2e-16 ***
## SexMale                     0.739834   0.105380   7.021 2.21e-12 ***
## Age                         0.007705   0.003186   2.418   0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8201.3  on 5916  degrees of freedom
## Residual deviance: 4668.7  on 5910  degrees of freedom
## AIC: 4680.7
##
## Number of Fisher Scoring iterations: 6
```

The estimate column gives the estimate for β for the logistic model

$$logit(E(Y|X)) = X\beta$$

A unit increase in the predictor variable X_j corresponds to an increase of β_j (estimated by $\hat{\beta}_j$) in the log-odds ratio with everything else being held fixed. A simpler interpretation is that $\hat{\beta}_j > 0$ can be interpreted as: An increase in X_j implies that $P(Y = 1|X = x)$ increases.

Thus

- Race being Asian/Pacific Islander decreases the propensity of incarcerations lasting longer than one day for "other traffic offenses"
- Race being Black also decreases the propensity of incarcerations lasting longer than one day for "other traffic offenses"
- Race being Hispanic also decreases the propensity of incarcerations lasting longer than one day for "other traffic offenses"

- Race being White also decreases the propensity of incarcerations lasting longer than one day for "other traffic offenses"

This shows that race does not really increase the propensity of incarceration.

- Sex being Male increases the propensity of incarcerations lasting longer than one day for "other traffic offenses"
- Age increasing also increases the propensity of incarcerations lasting longer than one day for "other traffic offenses"

The standard error column gives the standard error of the estimate of the β coefficients. The Z-value and P-value help in detecting the significance of the covariates. At a level of $\alpha = 0.05$ we can see that all the covariates are significant.

The null deviance and residual deviance give information about the goodness of fit of the null model (with no covariates) and the submodel we consider respectively. To check if the sub-model is better than the saturated model we can do a χ^2 test because under H_0 (The submodel is a better fit)

$$D(y; \hat{\mu}) \sim \chi^2_{n-p},$$

where D is the deviance.

```
pchisq(m1$deviance, df = m1$df.residual, lower = FALSE)
```

```
## [1] 1
```

Since the p-value is 1 this shows that the submodel is indeed a good fit to the data. We can also check if the null model (M_0) is better than the submodel (M_1) we choose.

$$H_0 \ : \ M_0 \text{ true} \quad H_a \ : \ M_1 \text{ true , but not } M_0$$

Then

$$D(y; \hat{\mu}_0) - D(y; \hat{\mu}_1) \sim \chi^2_{p_0 - p_1}$$

where $D(y; \hat{\mu}_0)$ is the null deviance and $D(y; \hat{\mu}_1)$ is the residual deviance.

```
pchisq(m1$null.deviance - m1$deviance, df = m1$df.null - m1$df.residual,
lower = FALSE)
```

```
## [1] 0
```

Since the pvalue is 0 this means that the submodel we choose is a better fit than the null model.

(b) Summarize the summary tables produced by a call to summary(m1) in the Galapagos islands example in the count regression notes.

```
m2 <- glm(Species ~ Elevation + Nearest + Scruz + Adjacent + Size,
family = "poisson", data = gala, x = TRUE)
summary(m2)
```

```
##
## Call:
## glm(formula = Species ~ Elevation + Nearest + Scruz + Adjacent +
##     Size, family = "poisson", data = gala, x = TRUE)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10.3723   -3.5214   -0.9947    1.7193   10.6627
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.790e+00  8.108e-02  34.410  < 2e-16 ***
## Elevation    9.361e-04  5.402e-05  17.329  < 2e-16 ***
## Nearest      6.469e-03  1.748e-03   3.702 0.000214 ***
## Scruz       -6.266e-03  6.268e-04  -9.997  < 2e-16 ***
## Adjacent    -2.858e-04  2.961e-05  -9.652  < 2e-16 ***
## Size2        1.128e+00  9.535e-02  11.826  < 2e-16 ***
## Size3        2.059e+00  9.419e-02  21.856  < 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: 3510.73  on 29  degrees of freedom
## Residual deviance:  594.18  on 23  degrees of freedom
## AIC: 769.01
##
## Number of Fisher Scoring iterations: 5
```

The estimate column gives the estimate for β for the logistic model

$$log(E(Y|X)) = X\beta$$

. A unit increase in the predictor variable X_j corresponds to an increase of β_j (estimated by $\hat{\beta}_j$) in the log of the mean response with everything else being held fixed. A simpler interpretation is that $\hat{\beta}_j > 0$ can be interpreted as: An increase in X_j implies that the mean response increases.

Thus

- Elevation increases the number of plant species found on each island
- Nearest increases the number of plant species found on each island
- Scruz decreases the number of plant species found on each island
- Adjacent also decreases the number of plant species found on each island
- Size of the island being in category 2 and 3 also increases the number of plant species found on each island

The standard error column gives the standard error of the estimate of the β coefficients. The Z-value and P-value help in detecting the significance of the covariates. At a level of $\alpha = 0.05$ we can see that all the covariates are significant.

The null deviance and residual deviance give information about the goodness of fit of the null model (w ith no covariates) and the submodel we consider respectively. To check if the sub-model is better than the saturated model we can do a χ^2 test because under H_0 (The submodel is a better fit)

$$D(y;\hat{\mu}) \sim \chi^2_{n-p},$$

where D is the deviance.

```
pchisq(m2$deviance, df = m2$df.residual, lower = FALSE)
```

```
## [1] 7.617409e-111
```

Since the p-value is very small this shows that the submodel not really a good fit to the data. We prefer the saturated model over it. We can also check if the null model (M_0) is better than the submodel (M_1) we choose.

$$H_0 \ : \ M_0 \text{ true} \quad H_a \ : \ M_1 \text{ true , but not } M_0$$

Then

$$D(y;\hat{\mu}_0) - D(y;\hat{\mu}_1) \sim \chi^2_{p_0-p_1}$$

where $D(y;\hat{\mu}_0)$ is the null deviance and $D(y;\hat{\mu}_1)$ is the residual deviance.

```
pchisq(m2$null.deviance - m2$deviance, df = m2$df.null - m2$df.residual,
lower = FALSE)
```

```
## [1] 0
```

Since the pvalue is 0 this means that the submodel we choose is a better fit than the null model.

Problem 5 [10 points]: Derive expressions and compute standard errors $\text{se}(\hat{\mu})$ in the logistic regression CCSO example without using `predict.glm`. Then construct Wald based confidence intervals for the estimated mean value parameters. Also construct confidence intervals $(g(\hat{\beta} - z_{\alpha/2}\text{se}(\hat{\beta})), g(\hat{\beta} + z_{\alpha/2}\text{se}(\hat{\beta})))$. Comment on the differences between these two confidence intervals for $\hat{\mu}$.

###solution 5

We know that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma^{-1})$$

where $\Sigma = X^T W X$, $W = diag(\mu_i(1-\mu_i))$ and $\mu = e^{X\beta}/(1+e^{X\beta})$

Now $\mu_i = g(\beta) = e^{M_i^T\beta}/(1+e^{M_i^T\beta}) \implies g'(\beta) = M_i * \mu_i(1-\mu_i)$

Thus by the delta method

$$\sqrt{n}(\hat{\mu}_i - \mu_i) \xrightarrow{d} N(0, \hat{\mu}_i^2(1-\hat{\mu}_i)^2 M_i^T \Sigma^{-1} M_i)$$

Thus our estimation of the SE is:

```r
#Creating the model matrix
X = model.matrix(atleastone ~ -1 + Race + Sex + Age,data = CCSO_small)
#Calculating the se
se_pihat = sqrt(apply(X,1,function(j) t(j)%*%var_matrix_CCSO%*%j))*pi_CCSO*(1-pi_CCSO)
se_pihat[1:20]
```

```
##  [1] 0.006773575 0.006392931 0.005695285 0.006773575 0.013586275 0.013586275
##  [7] 0.011956810 0.013052798 0.005695285 0.006269441 0.010557231 0.010557231
## [13] 0.009911238 0.009911238 0.012797440 0.014138891 0.006552717 0.013037138
## [19] 0.006472744 0.006370288
```

The Wald based confidence intervals for the estimated mean value parameters are

$$\hat{\mu}_i(x) \pm z_{1-\alpha/2}\sigma_i \quad \text{where } \sigma_i = \hat{\mu}_i(1-\hat{\mu}_i)\sqrt{M_i^T\Sigma^{-1}M_i}$$

```r
#Creating the conf intervals
conf_lower = pi_CCSO - qnorm(0.975)*se_pihat
conf_upper = pi_CCSO + qnorm(0.975)*se_pihat
waldci = cbind(conf_upper, conf_lower)
waldci = waldci %>% as.data.frame()  %>%
  mutate(length = conf_upper -  conf_lower)
head(waldci)
```

```
##           V1         V2     length
## 1 0.11912305 0.09257112 0.02655193
## 2 0.12132956 0.09626973 0.02505983
## 3 0.06866721 0.04634211 0.02232511
## 4 0.11912305 0.09257112 0.02655193
## 5 0.16908475 0.11582753 0.05325722
## 6 0.16908475 0.11582753 0.05325722
```

And, for the other confidence interval type $g(\hat{\beta} + z_{\alpha/2}\text{se}(\hat{\beta})))$:

```r
m1 <- glm(atleastone ~ -1 + Race + Sex + Age, data = CCSO_small,
          family = "binomial", x = "TRUE")

betahat = m1$coefficients
M = m1$x
alpha = 0.025
z = qnorm(p = 1-alpha)
n = nrow(M)
pici_upper = 1/(1 + exp( - M %*% (betahat + z*diag(vcov(m1))) ))
pici_lower = 1/(1 + exp( - M %*% (betahat - z*diag(vcov(m1))) ))
pici = cbind(pici_upper, pici_lower)
pici = pici %>% as.data.frame()  %>%
  mutate(length = pici_upper -  pici_lower)
head(pici)
```

```
##           V1         V2      length
## 1 0.11212335 0.09988262 0.012240731
## 2 0.11523654 0.10268058 0.012555961
```

```
## 3 0.05988461 0.05521374 0.004670872
## 4 0.11212335 0.09988262 0.012240731
## 5 0.15268704 0.13280332 0.019883714
## 6 0.15268704 0.13280332 0.019883714
```

Then, the average length of the Wald and plug-in approaches are given as:

```
avg_length_wald = round(mean(waldci$length), digits=4)
avg_length_wald
```

```
## [1] 0.0392
```

```
avg_length_pi = round(mean(pici$length), digits=4)
avg_length_pi
```

```
## [1] 0.0157
```

Then, the average Wald CI is somewhat larger than the average plug-in CI.

Problem 6 [10 points]: Construct a nonparametric bootstrap procedure that estimates the uncertainty associated with both estimates of the average treatment effect (ATE) of online learning in the logistic regression notes. Do the conclusions change when we factor in the uncertainty obtained from the nonparametric bootstrap procedure? Explain.

Solution 6:

Here we are estimating the ATE for the scores earned by students in online learning as opposed to in-person learning.

```
#Reading in the data

dat = read.csv("/Users/diptarka/Documents/GitHub/stat528resources/notes/3_binary_response/online.csv")
dat_small <- dat %>% dplyr::select(Online, ACTMath, ACTMajor, ACT, Gender,
International, F17, S18, S19, Fa19, FR, SO, JR)

ATE_alt = NULL

#Taking 1000 bootstrap samples
for(i in 1:1000)
{
  rand = sample(nrow(dat_small),replace = T)
  m <- glm(Online ~., data = dat_small[rand,], family = "binomial")
  trt <- dat_small[rand,]$Online
  preds <- predict(m, type = "response")
  weights_alt_trt <- 1 / sum(trt / preds) * trt /preds
  weights_alt_notrt <- 1 / sum((1 - trt)/(1 - preds)) * (1-trt)/(1-preds)
  dat_new <- data.frame(dat[rand,], weights = weights_alt_trt - weights_alt_notrt)
  ATE_alt <- c(ATE_alt,sum(weights_alt_trt * dat_new$ObjExam) -
sum(weights_alt_notrt * dat_new$ObjExam))
```

```
}

mean(ATE_alt)
```

```
## [1] 0.5911978
```

```
var(ATE_alt)
```

```
## [1] 0.47968
```

```
quantile(ATE_alt,prob = c(0.025,0.975))
```

```
##       2.5%      97.5%
## -0.7862695  1.9076603
```

Since the mean and variance are small and the confidence intreval contains 0 we can still conclude that there is no difference between the two types of learning.

```
ATE_DR = NULL
for(i in 1:1000)
{
  rand = sample(nrow(dat),replace = T)
  trt = dat_small[rand,]$Online
  m <- glm(Online ~., data = dat_small[rand,], family = "binomial")
  preds <- predict(m, type = "response")
  dat_boot = dat[rand,]
  m_trt <- lm(ObjExam ~ ACTMath + ACTMajor + ACT + International + Gender +
FR + SO + JR + F17 + S18 + S19,
data = dat_boot[trt == 1, ])
  Y_trt <- predict(m_trt, newdata = dat_boot)
  m_notrt <- lm(ObjExam ~ ACTMath + ACTMajor + ACT + International + Gender +
FR + SO + JR + F17 + S18 + S19,
data = dat_boot[trt == 0, ])
  Y_notrt <- predict(m_notrt, newdata = dat_boot)
  ATE_DR <- c(ATE_DR,mean( (dat_boot$ObjExam * trt - (trt - preds) * Y_trt) / preds -
(dat_boot$ObjExam * (1 - trt) + (trt - preds)*Y_notrt) / (1 - preds)))
}

mean(ATE_DR)
```

```
## [1] 0.4409087
```

```
var(ATE_DR)
```

```
## [1] 0.3987808
```

```
quantile(ATE_DR,prob = c(0.025,0.975))
```

```
##       2.5%      97.5%
## -0.8100616  1.6835348
```

Since the mean and variance are small even for the robust estimate and the confidence intreval contains 0 we can still conclude that there is no difference between the two types of learning.

Solution 7

Problem 7 [15 points]: Use the `dvisits` data in the `faraway` package to answer the follow parts:

(a) Make plots which show the relationship between the response variable, doctorco, and the potential predictors, age and illness.

(b) Combine the predictors chcond1 and chcond2 into a single three-level factor. Make an appropriate plot showing the relationship between this factor and the response. Comment.

(c) Build a Poisson regression model with doctorco as the response and sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore and the three-level condition factor as possible predictor variables. Considering the deviance of this model, does this model fit the data?

(d) Plot the residuals and the fitted values — why are there lines of observations on the plot? Make a QQ plot of the residuals and comment.

(e) Use a stepwise AIC-based model selection method. What sort of person would be predicted to visit the doctor the most under your selected model?

(f) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0, 1, 2, etc. times.

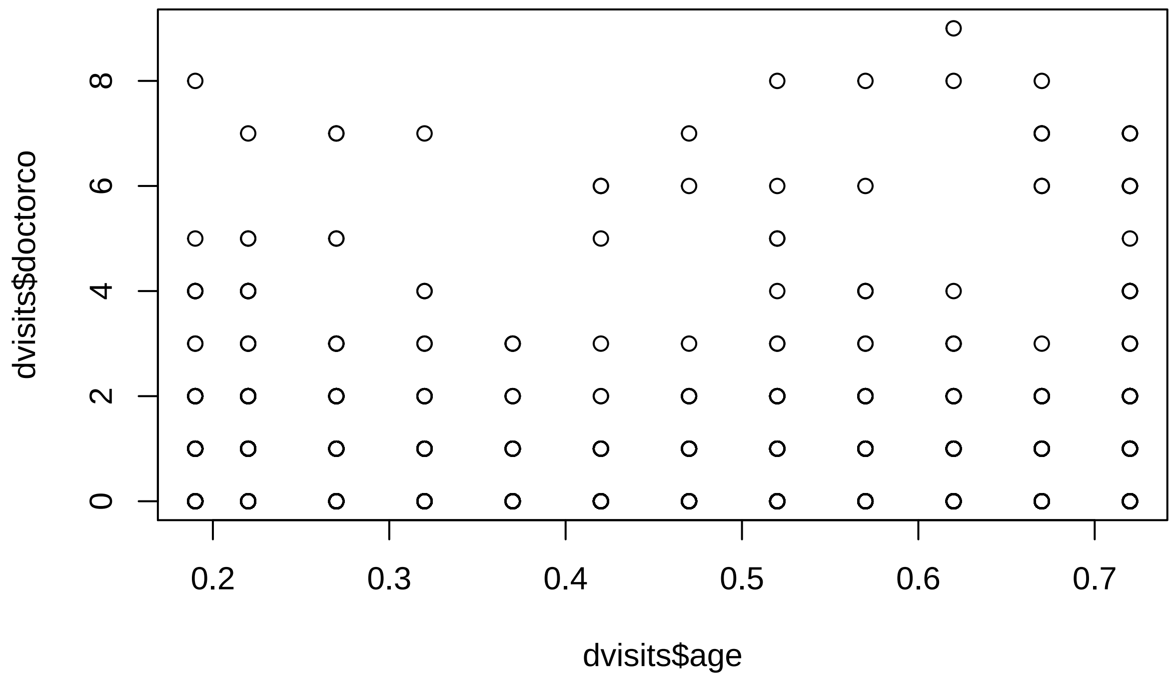
(g) Tabulate the frequencies of the number of doctor visits. Compute the expected frequencies of doctor visits under your most recent model. Compare the observed with the expected frequencies and comment on whether it is worth fitting a zero-inflated count model.

(h) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.

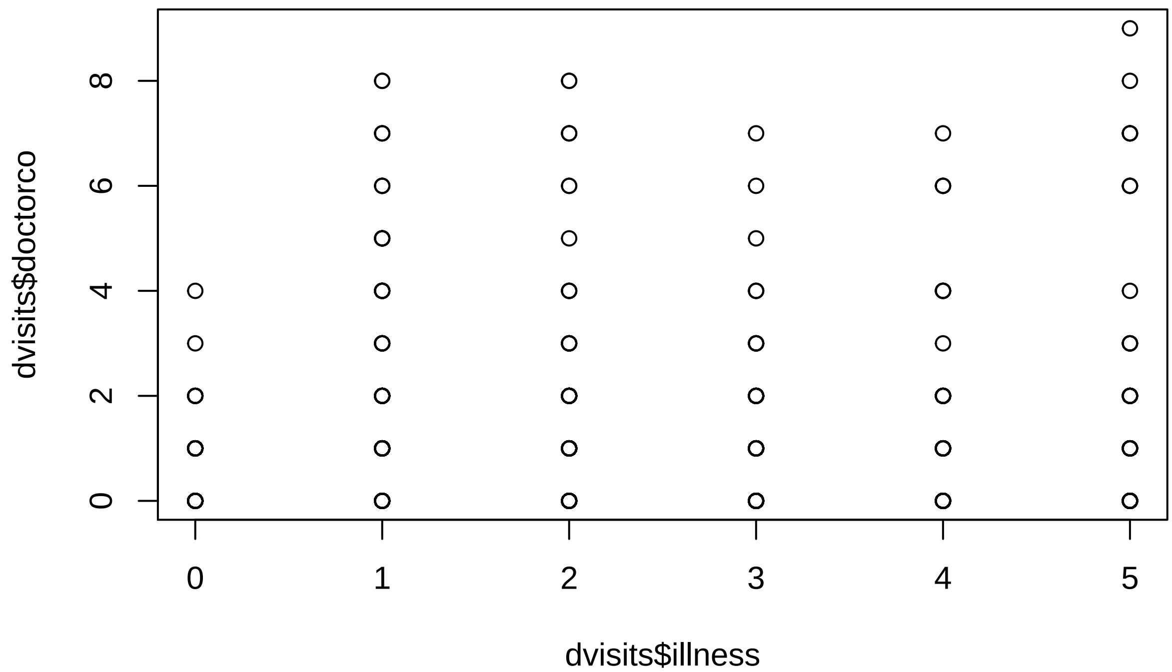
Solution 7:

(a) Make plots which show the relationship between the response variable, doctorco, and the potential predictors, age and illness.

```
data(dvisits)
plot(dvisits$age,dvisits$doctorco)
```




```
plot(dvisits$illness,dvisits$doctorco)
```

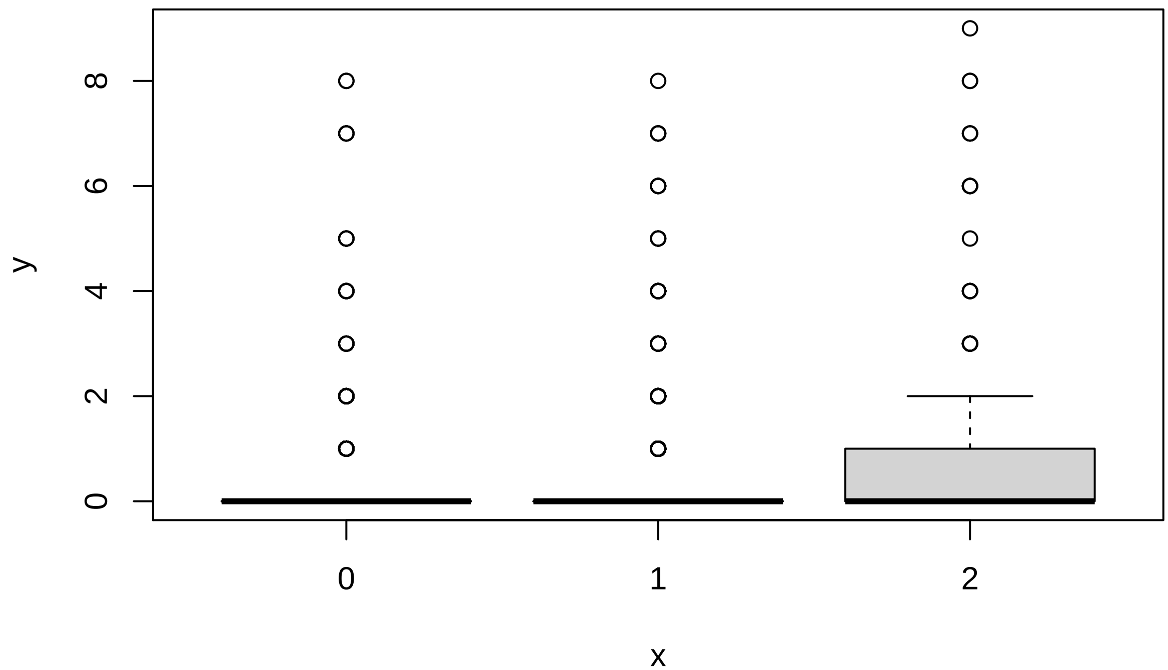


(b) Combine the predictors chcond1 and chcond2 into a single three-level factor. Make an appropriate plot showing the relationship between this factor and the response. Comment.

We create a new variable `chcond` which takes 3 factor values

- 1 if the patient has a chronic condition(s) but is not limited in activity
- 2 if the patient has a chronic condition(s) but is limited in activity
- 0 Otherwise

```
chcond = as.factor(dvisits$chcond1+2*dvisits$chcond2)
plot(chcond,dvisits$doctorco)
```



From the plot we can see that patients which chronic conditions which limit activity have more visits to the doctor. This possible is a factor which influences the response variable.

(c) Build a Poisson regression model with doctorco as the response and sex, age, agesq, income, levyplus, freepoor, freerepa, illness, actdays, hscore and the three-level condition factor as possible predictor variables. Considering the deviance of this model, does this model fit the data?

```
dat = dvisits %>% dplyr::select(doctorco, sex, age, agesq, income,
levyplus, freepoor,freerepa, illness, actdays, hscore)
dat = cbind(dat,"chcond" = chcond)
head(dat)
```

```
##   doctorco sex  age  agesq income levyplus freepoor freerepa illness actdays
## 1        1   1 0.19 0.0361   0.55        1        0        0       1       4
## 2        1   1 0.19 0.0361   0.45        1        0        0       1       2
## 3        1   0 0.19 0.0361   0.90        0        0        0       3       0
## 4        1   0 0.19 0.0361   0.15        0        0        0       1       0
## 5        1   0 0.19 0.0361   0.45        0        0        0       2       5
## 6        1   1 0.19 0.0361   0.35        0        0        0       5       1
##   hscore chcond
## 1      1      0
## 2      1      0
## 3      0      0
## 4      0      0
## 5      1      1
## 6      9      1
```

```r
#Creating the model
mod = glm(doctorco~.,data = dat,family = "poisson",x  =TRUE)
summary(mod)
```

```
##
## Call:
## glm(formula = doctorco ~ ., family = "poisson", data = dat, x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9170  -0.6862  -0.5743  -0.4839   5.7005
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.223848   0.189816 -11.716   <2e-16 ***
## sex          0.156882   0.056137   2.795   0.0052 **
## age          1.056299   1.000780   1.055   0.2912
## agesq       -0.848704   1.077784  -0.787   0.4310
## income      -0.205321   0.088379  -2.323   0.0202 *
## levyplus     0.123185   0.071640   1.720   0.0855 .
## freepoor    -0.440061   0.179811  -2.447   0.0144 *
## freerepa     0.079798   0.092060   0.867   0.3860
## illness      0.186948   0.018281  10.227   <2e-16 ***
## actdays      0.126846   0.005034  25.198   <2e-16 ***
## hscore       0.030081   0.010099   2.979   0.0029 **
## chcond1      0.114085   0.066640   1.712   0.0869 .
## chcond2      0.141158   0.083145   1.698   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4379.5  on 5177  degrees of freedom
## AIC: 6737.1
##
## Number of Fisher Scoring iterations: 6
```

```r
#Testing for the goodness of fit of the model against the saturated moddel
pchisq(mod$deviance, df =  mod$df.residual, lower = FALSE)
```

```
## [1] 1
```

Since the pvalue is 1 we can conclude that the model is indeed a better fit than the saturated model.

```r
pchisq(mod$null.deviance - mod$deviance, df = mod$df.null - mod$df.residual,
lower = FALSE)  %>%  round(4)
```

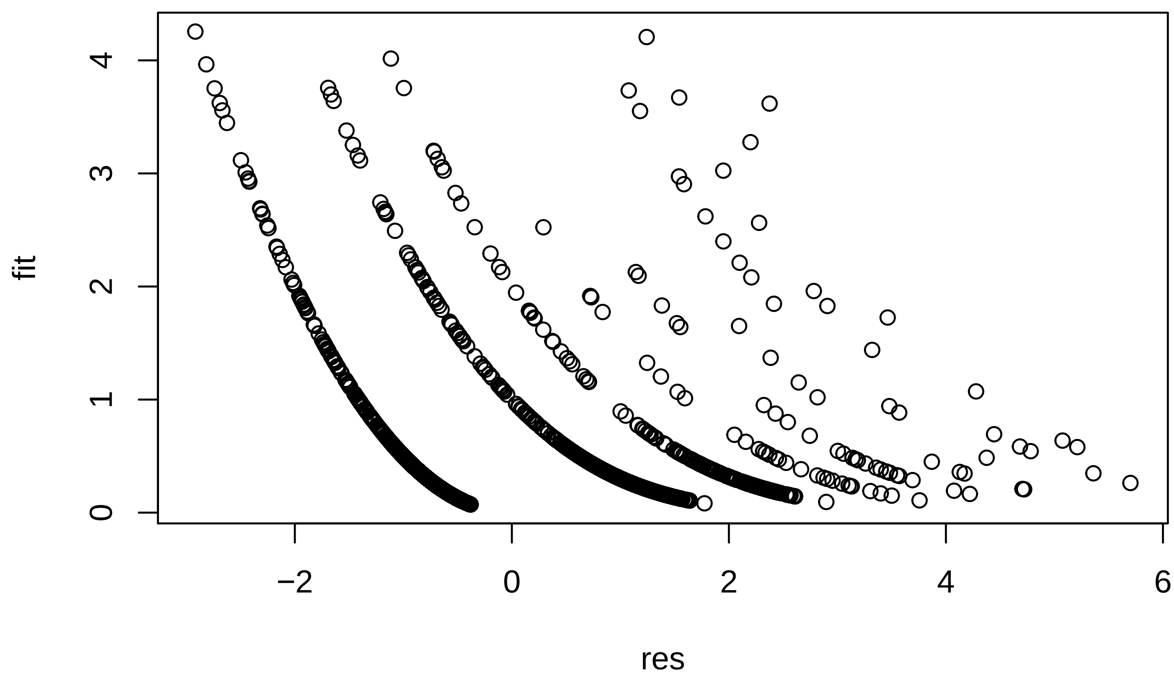
```
## [1] 0
```

Since the pvalue is nearly 0 we can conclude that the model is also better than the null model. Thus it is a appropriate fit to the data.

(d) Plot the residuals and the fitted values — why are there lines of observations on the plot? Make a QQ plot of the residuals and comment.

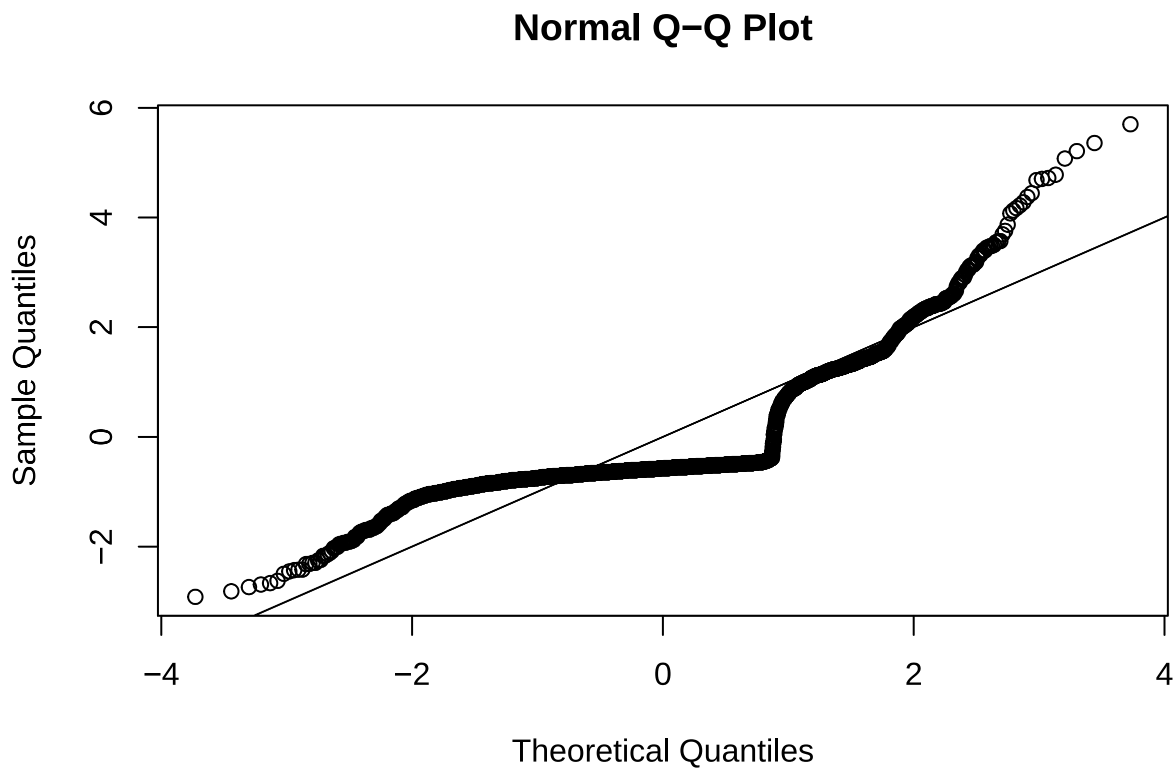
```
res = residuals(mod)
fit = fitted(mod)

plot(res,fit)
```



We observe lines of observations because most of the variables are factor variables with a small number of levels.

```
qqnorm(res)
abline(0, 1)
```



The QQ-plot shows that the residuals do not follow a normal distribution very well indicating the normality of residuals assumption is not reasonable.

(e) Use a stepwise AIC-based model selection method. What sort of person would be predicted to visit the doctor the most under your selected model?

```
#Selecting the model
library(MASS)
mod_Select = stepAIC(mod)
```

```
## Start:  AIC=6737.08
## doctorco ~ sex + age + agesq + income + levyplus + freepoor +
##     freerepa + illness + actdays + hscore + chcond
##
##            Df Deviance    AIC
## - agesq     1   4380.1 6735.7
## - freerepa  1   4380.3 6735.8
## - age       1   4380.6 6736.2
## - chcond    2   4383.2 6736.7
## <none>          4379.5 6737.1
## - levyplus  1   4382.5 6738.1
## - income    1   4385.0 6740.5
## - freepoor  1   4386.2 6741.8
## - sex       1   4387.4 6743.0
## - hscore    1   4388.1 6743.7
```

```
## - illness   1   4481.8 6837.4
## - actdays   1   4917.1 7272.7
##
## Step:  AIC=6735.7
## doctorco ~ sex + age + income + levyplus + freepoor + freerepa +
##     illness + actdays + hscore + chcond
##
##            Df Deviance    AIC
## - freerepa  1   4381.0 6734.5
## <none>          4380.1 6735.7
## - chcond    2   4384.2 6735.8
## - age       1   4383.0 6736.5
## - levyplus  1   4383.3 6736.9
## - income    1   4385.0 6738.6
## - freepoor  1   4386.8 6740.4
## - sex       1   4388.0 6741.5
## - hscore    1   4389.1 6742.7
## - illness   1   4481.9 6835.4
## - actdays   1   4917.1 7270.7
##
## Step:  AIC=6734.53
## doctorco ~ sex + age + income + levyplus + freepoor + illness +
##     actdays + hscore + chcond
##
##            Df Deviance    AIC
## <none>          4381.0 6734.5
## - levyplus  1   4383.4 6735.0
## - chcond    2   4385.5 6735.0
## - income    1   4386.7 6738.2
## - age       1   4387.1 6738.7
## - freepoor  1   4389.1 6740.6
## - sex       1   4389.5 6741.0
## - hscore    1   4390.2 6741.8
## - illness   1   4482.7 6834.2
## - actdays   1   4917.6 7269.2
```

```
#Outputting the best models
mod_Select$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## doctorco ~ sex + age + agesq + income + levyplus + freepoor +
##     freerepa + illness + actdays + hscore + chcond
##
## Final Model:
## doctorco ~ sex + age + income + levyplus + freepoor + illness +
##     actdays + hscore + chcond
##
##
##         Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                              5177   4379.515 6737.083
## 2    - agesq  1 0.6180113      5178   4380.133 6735.701
```

```
## 3 - freerepa  1 0.8279216      5179   4380.961 6734.529
```

```r
# The beta coefficients of our best moddel
beta = mod_Select$coefficients
beta
```

```
## (Intercept)         sex         age      income    levyplus    freepoor
## -2.08906349  0.16199995  0.35513074 -0.19980641  0.08368852 -0.46959634
##     illness     actdays      hscore     chcond1     chcond2
##  0.18610078  0.12661065  0.03111559  0.12110045  0.15889355
```

We can see that the number of doctor consultations increases when the patient is female, with increasing age, with low income, if covered by private health insurance, not covered by the government insurance for low income, high number of illness, high number of days of reduced activity, bad health score and with presence of chronic conditions. This indicates a poorer older woman with private insurance and higher number of illness is predicted to visit doctor more often.

(f) For the last person in the dataset, compute the predicted probability distribution for their visits to the doctor, i.e., give the probability they visit 0, 1, 2, etc. times.

```r
options (scipen = 99)
X = mod_Select$x
Y = mod_Select$y
hat_lambda_last = exp(t(X[nrow(dat),])%*%beta)
data.frame("Value" = 0:9, "Prob" = dpois(0:9,hat_lambda_last) %>%  round(6) )
```

```
##    Value     Prob
## 1      0 0.858916
## 2      1 0.130628
## 3      2 0.009933
## 4      3 0.000504
## 5      4 0.000019
## 6      5 0.000001
## 7      6 0.000000
## 8      7 0.000000
## 9      8 0.000000
## 10     9 0.000000
```

(g) Tabulate the frequencies of the number of doctor visits. Compute the expected frequencies of doctor visits under your most recent model. Compare the observed with the expected frequencies and comment on whether it is worth fitting a zero-inflated count model.

```r
observed_freq <- with(dvisits, table(doctorco))
est <- matrix(nrow=dim(dvisits)[1], ncol=10)
for(i in 1:dim(dvisits)[1]){
est[i,] <- dpois(0:9, fitted.values(mod_Select)[i])
}
expected_freq <- colMeans(est)*dim(dvisits)[1]
cbind.data.frame(observed_freq, expected_freq)
```

```
##    doctorco Freq expected_freq
## 1         0 4141  4013.6020569
## 2         1  782   928.3492327
## 3         2  174   168.0095991
## 4         3   30    45.4859546
## 5         4   24    18.9118479
## 6         5    9     8.8170066
## 7         6   12     4.0118473
## 8         7   12     1.7230627
## 9         8    5     0.6931622
## 10        9    1     0.2608052
```

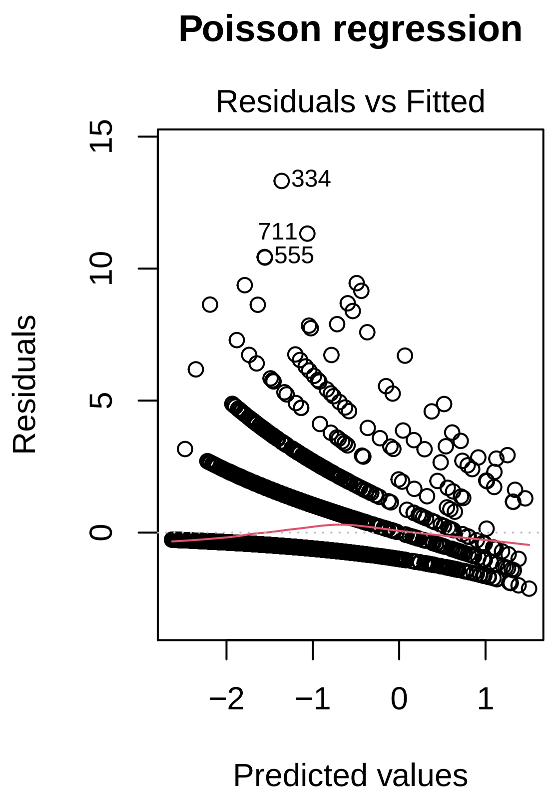
From the two tables, the observed and expected frequencies are close enough and thus it does not seem worth fitting a zero inflated model.

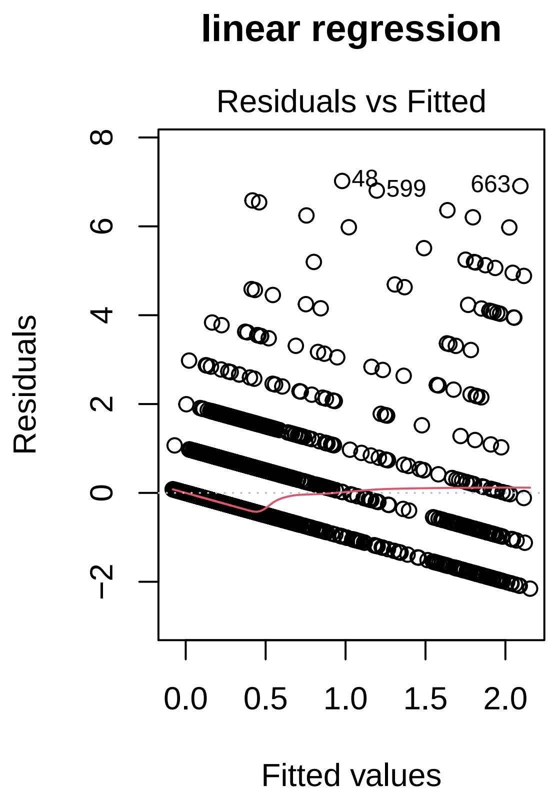
(h) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.

```r
dvisitsmod_lm <- lm(Y ~  X)
summary(dvisitsmod_lm)
```

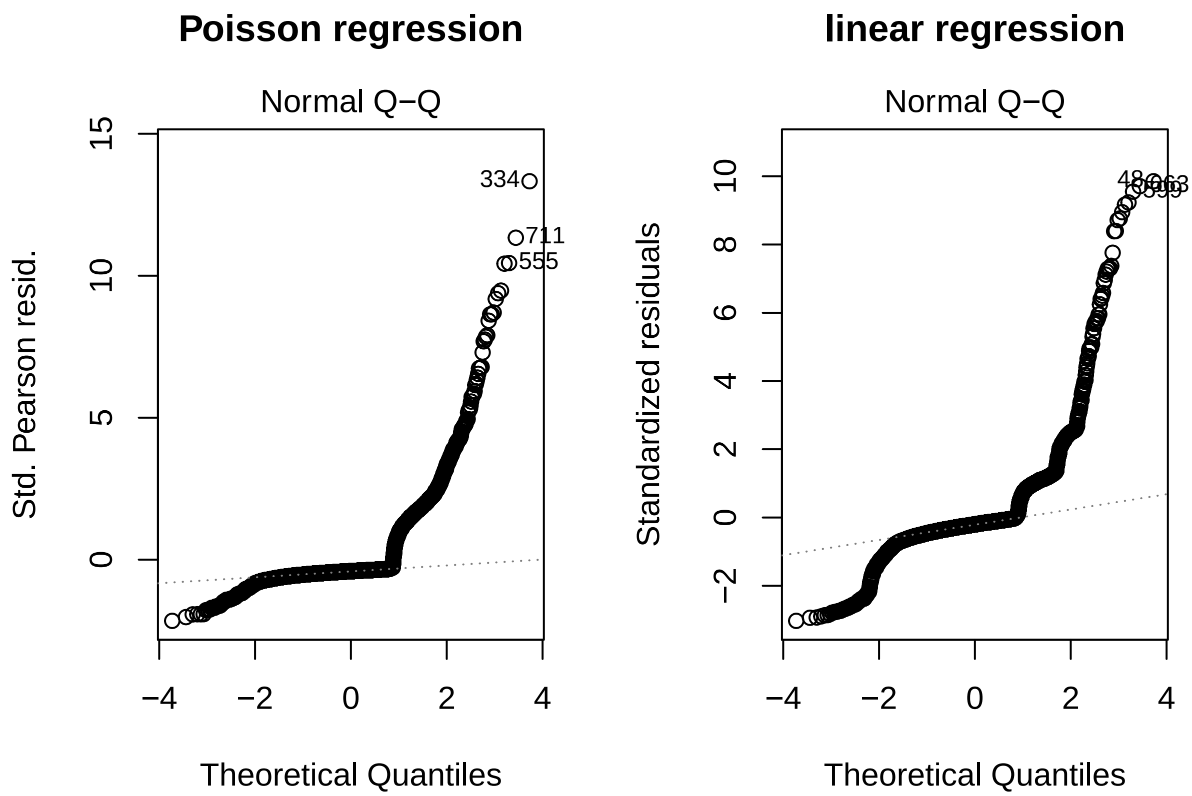
```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1543 -0.2584 -0.1440 -0.0434  7.0211
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  0.036781   0.035794   1.028             0.304201
## X(Intercept)       NA         NA      NA                   NA
## Xsex         0.035574   0.021505   1.654             0.098137 .
## Xage         0.180024   0.055912   3.220             0.001291 **
## Xincome     -0.061208   0.030522  -2.005             0.044975 *
## Xlevyplus    0.024041   0.021235   1.132             0.257626
## Xfreepoor   -0.111650   0.051532  -2.167             0.030308 *
## Xillness     0.060148   0.008332   7.219    0.000000000000602 ***
## Xactdays     0.103140   0.003656  28.213 < 0.0000000000000002 ***
## Xhscore      0.017064   0.005180   3.294             0.000994 ***
## Xchcond1     0.005006   0.023709   0.211             0.832776
## Xchcond2     0.045319   0.035352   1.282             0.199921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7138 on 5179 degrees of freedom
## Multiple R-squared:  0.2017, Adjusted R-squared:  0.2002
## F-statistic: 130.9 on 10 and 5179 DF,  p-value: < 0.00000000000000022
```

```r
par(mfrow=c(1,2))
plot(mod_Select, which=1)
title(main = "Poisson regression\n")
plot(dvisitsmod_lm, which=1)
title(main = "linear regression\n")
```






```
par(mfrow=c(1,2))
plot(mod_Select, which=2)
title(main = "Poisson regression\n")
plot(dvisitsmod_lm, which=2)
title(main = "linear regression\n")
```



This seems to indicate how the Poisson regression is a little better because the line for the residuals plot is much more linear around zero, along with the Q-Q plot. In conclusion, we can see that even though we need to account for overdispersion in the Poisson model it is still a better fit than the linear model.

Problem 8 [20 points]: Analyze the CCSO data set with Days in Jail as the response variable. You are allowed to dichotomize the response into a binary variable. Restrict attention to other traffic offenses as done in class. Your analysis needs to consider the variables considered in class as well as repeat offenders, multiple offenses, released reason, and agency. The determination of repeat offenders and multiple offenses can be done via the jacket number variable. Report interesting significant and null findings and determine that your final model is appropriate. You are allowed to use other inferential techniques than GLM. If you do so, then you need to justify your choices.

Solution 8

Data Wrangling: We first select the necessary data, here we focus on `OTHER TRAFFIC OFFENSES` type crimes. Our response variable is the binary variable denoting whether the offender stayed in jail for at least one day or not. We also restrict ourselves with some specific Race, Sex, and Arrest Agencies.

```
CCSO <- fread("https://uofi.box.com/shared/static/9elozjsg99bgcb7gb546wlfr3r2gc9b7.csv")
```

The full data has 67764 observations and 35 features.

```
## data wrangling
CCSO_imp <- CCSO %>% rename(Days = "Days in Jail", Age = "Age at Arrest",
                              Date = "BOOKING DATE", Sex = "SEX", Race = "RACE",
                              Crime = "CRIME CODE", Agency = "ARREST AGENCY", JNum = "JACKET NUMBER") %>
  mutate(atleastone = ifelse(Days > 0,1,0)) %>%
  filter(Crime == "OTHER TRAFFIC OFFENSES") %>%
  filter(Race %in% c("Asian/Pacific Islander","Black","White","Hispanic")) %>%
  filter(Sex %in% c("Female","Male")) %>%
  ## Agency (and possible interactions with Agency) is interesting if you want to go deepe filter(Agency
  filter(Agency %in%   c( "Illinois State Police", "Rantoul Police Department",
         "University of Illinois Police Department", "Urbana Police Department",
         "Champaign County Sherriff's Office","Champaign Police Department") ) %>%
  dplyr::select(atleastone, Age, Sex, Date, Race, Agency, JNum, `RELEASED REASON`, `EMPLOYMENT STATUS`,
  mutate(Race = fct_drop(Race), Sex = fct_drop(Sex), Agency = fct_drop(Agency))

CCSO_imp <- CCSO_imp[complete.cases(CCSO_imp), ]
```

Now, we will combine categories for the features RELEASED REASON, EMPLOYMENT STATUS, INCARCERATION REASON in order to reduce number of factors in these features.

```
CCSO_imp$`RELEASED REASON`[CCSO_imp$`RELEASED REASON` %in% c("Cash Bond Posted","Paid Fine, Court Costs
CCSO_imp$`RELEASED REASON`[CCSO_imp$`RELEASED REASON` %in% c("Credit Card Bond Posted")] = "CardBonds"
CCSO_imp$`RELEASED REASON`[CCSO_imp$`RELEASED REASON` %in% c("Placed on Probation

CCSO_imp$`RELEASED REASON`[CCSO_imp$`RELEASED REASON` %in% c("Release on Personal Recognizance
CCSO_imp$`RELEASED REASON`[CCSO_imp$`RELEASED REASON` %in% c("Served Sentence of Incarceration
CCSO_imp$`RELEASED REASON`[CCSO_imp$`RELEASED REASON` %in% c("Transfer to other county/state authorities
CCSO_imp$`RELEASED REASON`[!(CCSO_imp$`RELEASED REASON` %in% c("Bonds paid","Conditional Release","Relea

CCSO_imp$`EMPLOYMENT STATUS`[CCSO_imp$`EMPLOYMENT STATUS` %in% c("Employed - Full Time","Employed - Part
CCSO_imp$`EMPLOYMENT STATUS`[CCSO_imp$`EMPLOYMENT STATUS` %in%  c("Student","Retired")] = "Student/Reti
CCSO_imp$`EMPLOYMENT STATUS`[CCSO_imp$`EMPLOYMENT STATUS` %in% c("Unemployed","Laid Off")] ="Unemployed


CCSO_imp$`INCARCERATION REASON`[CCSO_imp$`INCARCERATION REASON` %in% c("Arrest - Without Warrant","Arres
CCSO_imp$`INCARCERATION REASON`[CCSO_imp$`INCARCERATION REASON` %in% c("FTA - CITY WARRANT (OV)","FTA -
CCSO_imp$`INCARCERATION REASON`[CCSO_imp$`INCARCERATION REASON` %in% c("Sentenced - EHD","Sentenced")] =
CCSO_imp$`INCARCERATION REASON`[!(CCSO_imp$`INCARCERATION REASON` %in% c("Arrest","FTA","Sentenced"))] =

CCSO_small <- CCSO_imp %>% rename(Inc_Reason = "INCARCERATION REASON",
                            Emp_Status = "EMPLOYMENT STATUS",
                            Rls_Reason = "RELEASED REASON")
```

Next, we add two derived features as directed in the problem: One denoting whethere the offender did multiple offenses in the same day (called multiple), and the other denoting whether they are repeat offenders (called repeatOffender)

```
date_and_jnum <- paste(CCSO_small$Date, CCSO_small$JNum)
dup_date_and_jnum <- date_and_jnum[duplicated(date_and_jnum)]
dup_date_and_jnum <- unique(dup_date_and_jnum)
CCSO_small <- cbind(CCSO_small, as.numeric(is.element(date_and_jnum, dup_date_and_jnum)))
CCSO_small <- CCSO_small %>% rename(multiple = 'V2')
```

```r
CCSO_small <- cbind(CCSO_small, as.numeric(duplicated(CCSO_small$JNum)))
CCSO_small <- CCSO_small %>% rename(repeatOffender = 'V2')

CCSO_small <- CCSO_small %>% dplyr::select(-c('JNum', 'Date'))
```

We show three rows from our final dataset. It has 5682 observations and 10 features

```r
head(CCSO_small, 3)
```

```
##    atleastone Age    Sex  Race                              Agency Rls_Reason
## 1:          0  22   Male White        Champaign Police Department      Other
## 2:          0  26   Male White Champaign County Sherriff's Office      Other
## 3:          0  32 Female White        Champaign Police Department   Released
##         Emp_Status Inc_Reason multiple repeatOffender
## 1: Student/Retired     Arrest        0              0
## 2:        Employed     Arrest        0              0
## 3:        Employed     Arrest        0              0
```

Model Fitting

Let us fit our full model here

```r
mymod <- glm(atleastone ~ -1 + ., data = CCSO_small,
             family = "binomial", x = "TRUE")

summary(mymod)
```

```
##
## Call:
## glm(formula = atleastone ~ -1 + ., family = "binomial", data = CCSO_small,
##     x = "TRUE")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2979  -0.4575  -0.3068  -0.2451   2.7894
##
## Coefficients:
##                                             Estimate Std. Error z value
## Age                                         0.003014   0.004225   0.713
## SexFemale                                   1.848287   1.008944   1.832
## SexMale                                     2.347515   1.001090   2.345
## RaceBlack                                   1.842612   0.560254   3.289
## RaceHispanic                                0.790797   0.576802   1.371
## RaceWhite                                   0.927702   0.561009   1.654
## AgencyChampaign Police Department           0.164304   0.149798   1.097
## AgencyIllinois State Police                 0.271086   0.147166   1.842
## AgencyRantoul Police Department             1.194293   0.186301   6.411
## AgencyUniversity of Illinois Police Department  0.114082   0.198807   0.574
## AgencyUrbana Police Department              0.512306   0.164309   3.118
## Rls_ReasonOther                            -5.740143   0.725863  -7.908
## Rls_ReasonReleased                         -5.600271   0.741579  -7.552
```

```
## Rls_ReasonServed/Acquitted                      -0.643636   1.286021  -0.500
## Rls_ReasonTransfered                            -1.135399   0.827509  -1.372
## Emp_StatusEmployed                              -0.670441   0.389714  -1.720
## Emp_StatusStudent/Retired                       -0.693590   0.425651  -1.629
## Emp_StatusUnemployed                             0.376140   0.391178   0.962
## Inc_ReasonFTA                                    1.193609   0.254473   4.691
## Inc_ReasonOther                                  0.818230   0.379699   2.155
## Inc_ReasonSentenced                              3.858879   1.088797   3.544
## multiple                                        -0.295395   0.122215  -2.417
## repeatOffender                                   0.197053   0.118757   1.659
##                                                            Pr(>|z|)
## Age                                                        0.475730
## SexFemale                                                  0.066966 .
## SexMale                                                    0.019029 *
## RaceBlack                                                  0.001006 **
## RaceHispanic                                               0.170375
## RaceWhite                                                  0.098202 .
## AgencyChampaign Police Department                          0.272714
## AgencyIllinois State Police                                0.065468 .
## AgencyRantoul Police Department                  0.00000000014497449 ***
## AgencyUniversity of Illinois Police Department             0.566080
## AgencyUrbana Police Department                             0.001821 **
## Rls_ReasonOther                                  0.00000000000000262 ***
## Rls_ReasonReleased                               0.00000000000004292 ***
## Rls_ReasonServed/Acquitted                                 0.616733
## Rls_ReasonTransfered                                       0.170042
## Emp_StatusEmployed                                         0.085370 .
## Emp_StatusStudent/Retired                                  0.103211
## Emp_StatusUnemployed                                       0.336273
## Inc_ReasonFTA                                    0.00000272532190133 ***
## Inc_ReasonOther                                            0.031166 *
## Inc_ReasonSentenced                                        0.000394 ***
## multiple                                                   0.015649 *
## repeatOffender                                             0.097056 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7876.9  on 5682  degrees of freedom
## Residual deviance: 3033.4  on 5659  degrees of freedom
## AIC: 3079.4
##
## Number of Fisher Scoring iterations: 7
```

Let us first test the goodness of the model against the null model. We can see the model is a better fit than the null model

```
#Testing against the null model
pchisq(mymod$null.deviance - mymod$deviance, df = mymod$df.null - mymod$df.residual,
lower = FALSE)
```

```
## [1] 0
```

Model Selection We can see not all the features considered show a significant effect on the response. Thus, we can do feature selection via stepwise method

```
mod_Select = stepAIC(mymod)
```

```
## Start:  AIC=3079.41
## atleastone ~ -1 + (Age + Sex + Race + Agency + Rls_Reason + Emp_Status +
##     Inc_Reason + multiple + repeatOffender)
##
##                  Df Deviance    AIC
## - Age             1   3033.9 3077.9
## <none>                3033.4 3079.4
## - repeatOffender  1   3036.2 3080.2
## - multiple        1   3039.3 3083.3
## - Sex             2   3055.7 3097.7
## - Agency          5   3077.1 3113.1
## - Inc_Reason      3   3080.6 3120.6
## - Race            3   3130.2 3170.2
## - Emp_Status      3   3137.6 3177.6
## - Rls_Reason      4   3639.4 3677.4
##
## Step:  AIC=3077.92
## atleastone ~ Sex + Race + Agency + Rls_Reason + Emp_Status +
##     Inc_Reason + multiple + repeatOffender - 1
##
##                  Df Deviance    AIC
## <none>                3033.9 3077.9
## - repeatOffender  1   3036.7 3078.7
## - multiple        1   3039.7 3081.7
## - Sex             2   3056.8 3096.8
## - Agency          5   3078.4 3112.4
## - Inc_Reason      3   3081.2 3119.2
## - Race            3   3131.1 3169.1
## - Emp_Status      3   3138.7 3176.7
## - Rls_Reason      4   3639.5 3675.5
```

```
summary(mod_Select)
```

```
##
## Call:
## glm(formula = atleastone ~ Sex + Race + Agency + Rls_Reason +
##     Emp_Status + Inc_Reason + multiple + repeatOffender - 1,
##     family = "binomial", data = CCSO_small, x = "TRUE")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2933  -0.4536  -0.3058  -0.2461   2.7775
##
## Coefficients:
##                                                 Estimate Std. Error z value
## SexFemale                                        1.90635    1.00581   1.895
## SexMale                                          2.40881    0.99755   2.415
## RaceBlack                                        1.85213    0.56024   3.306
```

```
## RaceHispanic                                     0.79093    0.57699   1.371
## RaceWhite                                        0.93927    0.56090   1.675
## AgencyChampaign Police Department                0.16927    0.14963   1.131
## AgencyIllinois State Police                      0.27921    0.14671   1.903
## AgencyRantoul Police Department                  1.20142    0.18608   6.456
## AgencyUniversity of Illinois Police Department  0.09876    0.19765   0.500
## AgencyUrbana Police Department                   0.51466    0.16428   3.133
## Rls_ReasonOther                                 -5.72966    0.72553  -7.897
## Rls_ReasonReleased                              -5.59484    0.74136  -7.547
## Rls_ReasonServed/Acquitted                      -0.63323    1.28523  -0.493
## Rls_ReasonTransfered                            -1.12623    0.82720  -1.361
## Emp_StatusEmployed                              -0.65760    0.38932  -1.689
## Emp_StatusStudent/Retired                       -0.68934    0.42565  -1.620
## Emp_StatusUnemployed                             0.39054    0.39067   1.000
## Inc_ReasonFTA                                    1.19580    0.25442   4.700
## Inc_ReasonOther                                  0.82280    0.37985   2.166
## Inc_ReasonSentenced                              3.85425    1.08849   3.541
## multiple                                        -0.29416    0.12222  -2.407
## repeatOffender                                   0.19810    0.11875   1.668
##                                                            Pr(>|z|)
## SexFemale                                                  0.058049 .
## SexMale                                                    0.015747 *
## RaceBlack                                                  0.000946 ***
## RaceHispanic                                               0.170440
## RaceWhite                                                  0.094017 .
## AgencyChampaign Police Department                          0.257937
## AgencyIllinois State Police                                0.057024 .
## AgencyRantoul Police Department                 0.00000000010724171 ***
## AgencyUniversity of Illinois Police Department             0.617327
## AgencyUrbana Police Department                             0.001731 **
## Rls_ReasonOther                                 0.00000000000000285 ***
## Rls_ReasonReleased                              0.00000000000004464 ***
## Rls_ReasonServed/Acquitted                                 0.622224
## Rls_ReasonTransfered                                       0.173358
## Emp_StatusEmployed                                         0.091198 .
## Emp_StatusStudent/Retired                                  0.105334
## Emp_StatusUnemployed                                       0.317472
## Inc_ReasonFTA                                   0.00000260034427526 ***
## Inc_ReasonOther                                            0.030302 *
## Inc_ReasonSentenced                                        0.000399 ***
## multiple                                                   0.016093 *
## repeatOffender                                             0.095270 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7876.9  on 5682  degrees of freedom
## Residual deviance: 3033.9  on 5660  degrees of freedom
## AIC: 3077.9
##
## Number of Fisher Scoring iterations: 7
```

The `Age` feature gets completely removed. Let us examine if this is a good fit. We can notice this model is significantly better than the null model and the full model is not significantly better than this model. We are satisfied with this selected model

```
#Testing against the saturated model
pchisq(mod_Select$deviance - mymod$deviance, df = mod_Select$df.residual - mymod$df.residual, lower = FA
```

```
## [1] 0.4773979
```

```
#Testing against the null model
pchisq(mymod$null.deviance - mod_Select$deviance, df = mymod$df.null - mod_Select$df.residual,
lower = FALSE)
```

```
## [1] 0
```

Discussion Now, finally - if we look at the summary table - the features that seem most useful for predicting the response are Some interesting findings are:

- – Race again has significant effect on response; Especially being black seems to be associated with longer incarceration
- – One agency 'Rantoul Police Department' stands out among all agencies. It appears people arrested by this agency tends to have longer incarcerations.
- – Among incarceration reason, 'FTA' reasons tend to make incarceration longer. This maybe because since they already have failed to appear.
- – Multiple offenses is related to longer incarcerations, per expectations. However, repeat offenses do not seem to have such strong associations. s
- – Being Male also seems to be more problematic than being female.

We can thus understand, black male are more likely to spend at least one day in jail than others. Specially if they are under Rantoul PD and/or have FTA against them. However, we should remember this is an observational data and we shouldn't make causal connections with any confidence.