Homework 3: Binary and Count Regressions

Solution Set

Problem 1[15 points]

This problem concerns manual creation of summary tables from nothing more than the observed data and the assumed model.

- **part a** [5 points]: Manually write your own iteratively reweighted least squares algorithm which maximizes the logistic regression log likelihood for the CCSO example in the notes. Report the estimated submodel canonical parameter vector $\hat{\beta}$ and reproduce the summary table (up to convergence tolerance differences) without using the `glm` or `summary` commands. You can ignore deviance residuals.
- **part b** [5 points]: Manually write your own iteratively reweighted least squares 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.
- **part c** [5 points]: Manually write your own Fisher scoring algorithm for one of the parts above, and compare estimates of β from the Fisher scoring algorithm and the iteratively rewighted least squares algorithm.

Solution 1

- **part a**

Now we will define the IRLS function for this problem.

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

```
CCSO = fread("https://uofi.box.com/shared/static/9elozjsg99bgcb7gb546wlfr3r2gc9b7.csv")
CCSO <- CCSO %>% rename(daysInJail = "Days in Jail", arrestAge = "Age at Arrest",
bookingDate = "BOOKING DATE", sex = "SEX", race = "RACE",
crimeCode = "CRIME CODE", jacketNumber = 'JACKET NUMBER',
releasedReason = 'RELEASED REASON', arrestAgency = 'ARREST AGENCY',
employmentStatus = 'EMPLOYMENT STATUS', city = 'CITY'
)
```

```r
CCSO_small = CCSO %>%
mutate(atleastone = ifelse(daysInJail > 0,1,0)) %>%
filter(crimeCode == "OTHER TRAFFIC OFFENSES") %>%
filter(race %in% c("Asian/Pacific Islander","Black","White","Hispanic")) %>%
filter(sex %in% c("Female","Male")) %>%
dplyr::select(atleastone, arrestAge, sex, race, bookingDate) %>%
mutate(race = fct_drop(race), sex = fct_drop(sex))
CCSO_small = CCSO_small[complete.cases(CCSO_small), ]
head(CCSO_small)
```

```
##    atleastone arrestAge    sex     race bookingDate
## 1:          0        22   Male    White    1/1/2011
## 2:          0        26   Male    White    1/1/2011
## 3:          0        32 Female    White    1/1/2011
## 4:          0        22   Male    White    1/2/2011
## 5:          0        35   Male Hispanic    1/2/2011
## 6:          0        35   Male Hispanic    1/2/2011
```

```r
M1 <- model.matrix(~ -1 + race + sex + arrestAge, data = CCSO_small)
Y1 = CCSO_small$atleastone
```

```r
irls_logistic_regression <- function(M, Y, max.iter = 100, tolerance = 1e-8) {
  # Number of parameters
  p <- dim(M)[2]

  # Initialize beta
  beta <- matrix(0, nrow = p, ncol = 1)

  # Loop until convergence or max iterations
  for (iter in 1:max.iter) {
    # Calculate probabilities
    eta <- M %*% beta
    p <- 1 / (1 + exp(-eta))

    # Calculate weights
    W <- diag(as.vector(p * (1 - p)))

    # Update beta using the Newton-Raphson method (IRLS)
    z <- eta + (Y - p) / (p * (1 - p))
    beta.new <- solve(t(M) %*% W %*% M) %*% t(M) %*% W %*% z

    # Check for convergence
    if (sqrt(sum((beta.new - beta)^2)) < tolerance) {
      cat("Converged in", iter, "iterations\n")
      break
    }

    beta <- beta.new
  }

  return(beta)
}
beta_1 = irls_logistic_regression(M1, Y1)
```

```
## Converged in 8 iterations
```

```
eta <- M1 %*% beta_1
p <- 1 / (1 + exp(-eta))
W <- diag(as.vector(p * (1 - p)))
H <- t(M1) %*% W %*% M1
var_cov_matrix <- solve(H)
sd_err <- sqrt(diag(var_cov_matrix))

z_val = beta_1/sd_err

p_val = 2 * (1 - pnorm(abs(beta_1/sd_err)))

sum_table_1 = data.frame(beta_1, sd_err, z_val, p_val)
sum_table_1
```

```
##                                beta_1      sd_err      z_val        p_val
## 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
## arrestAge                   0.00770504 0.003186262   2.418207 1.559721e-02
```

- **part b**

```
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
```

```
gala = gala %>%
mutate(Size = as.factor(1 + ifelse(Area > 1,1,0) + ifelse(Area > 25,1,0)))

#m1 = glm(Species ~ Elevation + Nearest + Scruz + Adjacent + Size,
#family = "poisson", data = gala, x = TRUE)
#M2 = m1$x
#Y2 = m1$y
#head(M2)
```

```r
M2 <- model.matrix(~ Elevation + Nearest + Scruz + Adjacent + Size, data = gala)
Y2 = gala$Species
```

```r
irls_poisson_regression <- function(M, Y, max.iter = 100, tolerance = 1e-8) {
  # Initialize beta
  p <- dim(M)[2] # Number of parameters
  set.seed(96)
  #beta <- matrix(0, nrow = p, ncol = 1)

  beta <- matrix(rnorm(p)/100, nrow = p, ncol = 1)

  # Loop until convergence or max iterations
  for (iter in 1:max.iter) {
    # Calculate lambda (mean) for Poisson distribution
    eta <- M %*% beta
    lambda <- exp(eta)

    # Calculate weights
    W <- diag(as.vector(lambda),nrow=nrow(M),ncol=nrow(M))

    # Calculate the working response
    z <- eta + solve(W) %*% (Y - lambda)

    # Update beta
    beta_new <- solve(t(M) %*% W %*% M) %*% t(M) %*% W %*% z

    # Check for convergence
    if (sqrt(sum((beta_new - beta)^2)) < tolerance) {
      cat("Converged in", iter, "iterations\n")
      break
    }

    beta <- beta_new
  }

  return(beta)
}

beta_2 = irls_poisson_regression(M2, Y2)
```

```
## Converged in 22 iterations
```

```r
eta <- M2 %*% beta_2
lambda <- exp(eta)
W <- diag(as.vector(lambda),nrow=nrow(M2),ncol=nrow(M2))
H <- t(M2) %*% W %*% M2
var_cov_matrix <- solve(H)
sd_err <- sqrt(diag(var_cov_matrix))

z_val = beta_2/sd_err

p_val = 2 * (1 - pnorm(abs(beta_2/sd_err)))
```

```
sum_table_2 = data.frame(beta_2, sd_err, z_val, p_val)
sum_table_2
```

```
##                   beta_2       sd_err     z_val        p_val
## (Intercept)  2.7897964688 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        1.1276155410 9.535272e-02 11.825730 0.0000000000
## Size3        2.0586771297 9.419392e-02 21.855732 0.0000000000
```

- **part c**

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:

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

```
#Creating the model matrix
X = model.matrix(atleastone ~ -1 + race + sex + arrestAge, data = CCSO_small)
n = nrow(X)
p = ncol(X)
Y = CCSO_small$atleastone

# 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)
}
```

```
# 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)))
```

```
#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:

```
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 df" = n,
     "Residual df" = n-p, "AIC" = AIC)
```

```
## $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
## arrestAge                   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
```

Both algorithms yield the same estimate.

Problem 2[10 points]

Complete the following parts: - **part a** [5 points]: 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. - **part b** [5 points]: 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 2

- **part a**

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

```
##
## Call:
## glm(formula = atleastone ~ -1 + race + sex + arrestAge, 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 ***
## arrestAge                   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 is significant when testing at reasonable significance levels. We observe that Black individuals are estimated to have comparatively larger propensity of incarcerations lasting one day or longer for "other traffic offenses". We would need to look into other factors such as socio-economic status, repeat offenders, and multiple offenses before we could conclude that race is the driver of longer incarcerations.
- Sex being Male is estimated to increase the propensity of incarcerations lasting longer than one day for "other traffic offenses".
- Age increasing also is estimated to increase 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.

- **part 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 is estimated to increase the expected number of plant species found on each island
- Distance to the nearest island is estimated to increases the expected number of plant species found on each island
- Distance to Scruz is estimated to decreases the number of expected plant species found on each island

- Larger adjacent islands are also estimated to decrease the number of expected plant species found on each island
- Medium and large islands are estimated to have more expected 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 3[10 points]

This problem concerns MLEs and inferences of modeling parameters using the CCSO example in class. Do the following:

- **part a** [5 points]: Compute MLEs and estimated standard errors for the saturated model parameter μ from the logistic regression with race, sex, and arrestAge as predictors for atleastone fit to the CCSO data restricted to "other traffic offenses". Compare with `predict.glm`.
- **part b** [5 points]: 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 any noticeable differences between these two confidence intervals for $\hat{\mu}$.

Solution 3

- **part a**

The logistic regression with race, sex, and arrestAge as predictors for atleastone fit to the CCSO data restricted to "other traffic offenses".

```
f1 <- function(x){1 /(1 + exp(-x))}
f2=function(x) {exp(x)/(1+exp(x))^2}
X = model.matrix(atleastone ~ -1 + race + sex + arrestAge,data = CCSO_small)
m1 = glm(atleastone ~ -1 + race + sex + arrestAge, data = CCSO_small,
family = "binomial", x = "TRUE")

beta = m1$coefficient
M = m1$x
mu_mle = as.numeric(f1(M %*% beta))

p1 = predict(m1, type = "response", se.fit = TRUE)
p1.fit = as.numeric(p1$fit)

(cbind.data.frame('Calculated MLEs'=  mu_mle[1:10],'Obtained MLEs' =  p1.fit[1:10]))
```

```
   Calculated MLEs Obtained MLEs
1       0.10584708    0.10584708
2       0.10879965    0.10879965
3       0.05750466    0.05750466
4       0.10584708    0.10584708
5       0.14245614    0.14245614
6       0.14245614    0.14245614
7       0.29426808    0.29426808
8       0.13781432    0.13781432
9       0.05750466    0.05750466
10      0.11030287    0.11030287
```

```
(sqrt( sum(mu_mle - p1.fit)^2 ))
```

```
[1] 2.406755e-14
```

Comparison: From the output above, we can see that the MLEs we computed and the values obtained by using the "predict" function are all same.

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
#Calculating the se
se_pihat = sqrt(apply(X,1,function(j) t(j)%*%var_matrix_CCSO%*%j))*pi_CCSO*(1-pi_CCSO)
```

```r
p1.se = as.numeric(p1$se.fit)
(cbind.data.frame('Calculated SEs'=  se_pihat[1:10],'Obtained SEs' =  p1.se[1:10]))
```

```
   Calculated SEs Obtained SEs
1     0.006773575  0.006773575
2     0.006392931  0.006392931
3     0.005695285  0.005695285
4     0.006773575  0.006773575
5     0.013586275  0.013586275
6     0.013586275  0.013586275
7     0.011956810  0.011956810
8     0.013052798  0.013052798
9     0.005695285  0.005695285
10    0.006269441  0.006269441
```

```r
(sqrt( sum(se_pihat - p1.se)^2 ))
```

```
[1] 7.355313e-06
```

Comparison: From the output above, we can see that the estimated standard errors we computed and the values obtained by using the "predict" function are all same.

- **part b**

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.data.frame('lower' =  conf_lower, 'upper' = conf_upper)
waldci = waldci %>% as.data.frame()  %>%
  mutate(length = conf_upper -  conf_lower) %>%  round(4)
head(waldci)
```

```
##    lower  upper length
## 1 0.0926 0.1191 0.0266
## 2 0.0963 0.1213 0.0251
## 3 0.0463 0.0687 0.0223
## 4 0.0926 0.1191 0.0266
## 5 0.1158 0.1691 0.0533
## 6 0.1158 0.1691 0.0533
```

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

```r
m1 <- glm(atleastone ~ -1 + race + sex + arrestAge, 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.data.frame('lower' =  pici_upper, 'upper' = pici_upper)
pici = pici %>% as.data.frame()  %>%
  mutate(length = pici_upper -  pici_lower) %>%  round(4)
head(pici)
```

```
##    lower  upper length
## 1 0.1121 0.1121 0.0122
## 2 0.1152 0.1152 0.0126
## 3 0.0599 0.0599 0.0047
## 4 0.1121 0.1121 0.0122
## 5 0.1527 0.1527 0.0199
## 6 0.1527 0.1527 0.0199
```

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

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

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

```r
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 4[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 4 Here we are estimating the ATE for the scores earned by students in online learning as opposed to in-person learning.

```r
#Reading in the data
dat = read.csv("/Users/diptarka/Documents/GitHub/1_sp24_STAT528_DS/Solutions/HW3/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] 0.5965967
```

```r
var(ATE_alt)
```

```
## [1] 0.4604335
```

```r
quantile(ATE_alt,prob = c(0.025,0.975))%>%  round(4)
```

```
##    2.5%   97.5%
## -0.6667  2.0191
```

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.

```r
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] 0.4653583
```

```
var(ATE_DR)
```

```
## [1] 0.4021998
```

```
quantile(ATE_DR,prob = c(0.025,0.975))%>%  round(4)
```

```
##    2.5%   97.5%
## -0.7520  1.7417
```

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.

Problem 5[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.

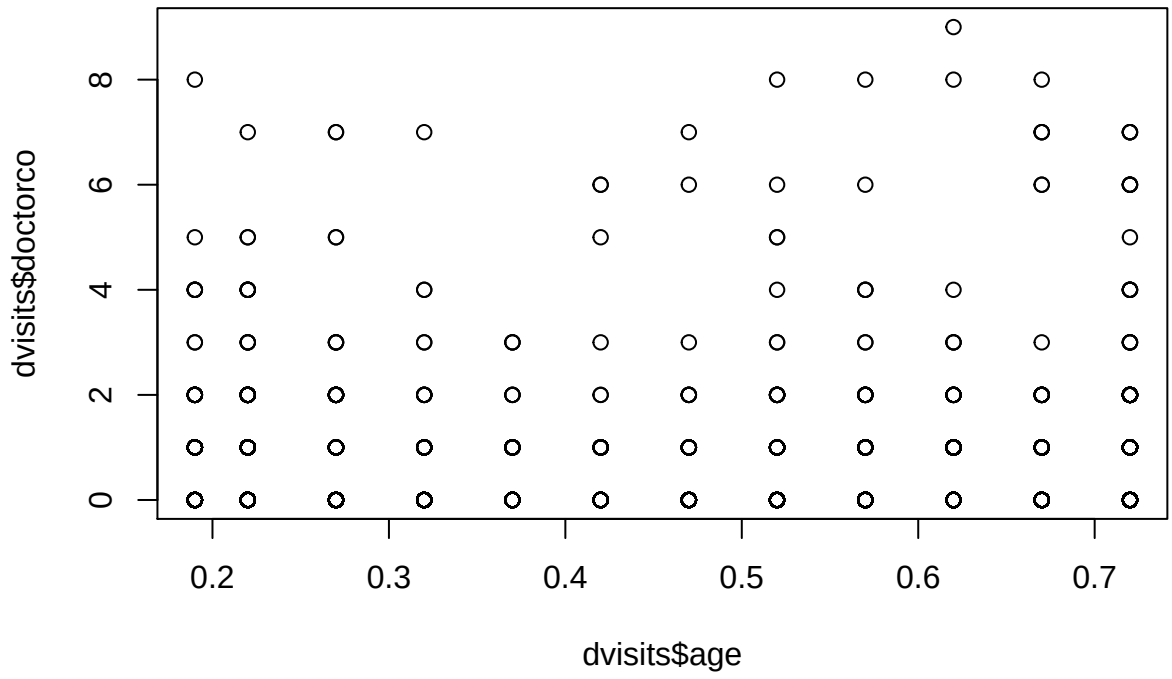
(g) Tabulate the frequencies of the number of doctor visits. Compute the expected frequencies of doctor visits under your most recent model. Compare the observed with the expected frequencies and comment on whether it is worth fitting a zero-inflated count model.

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

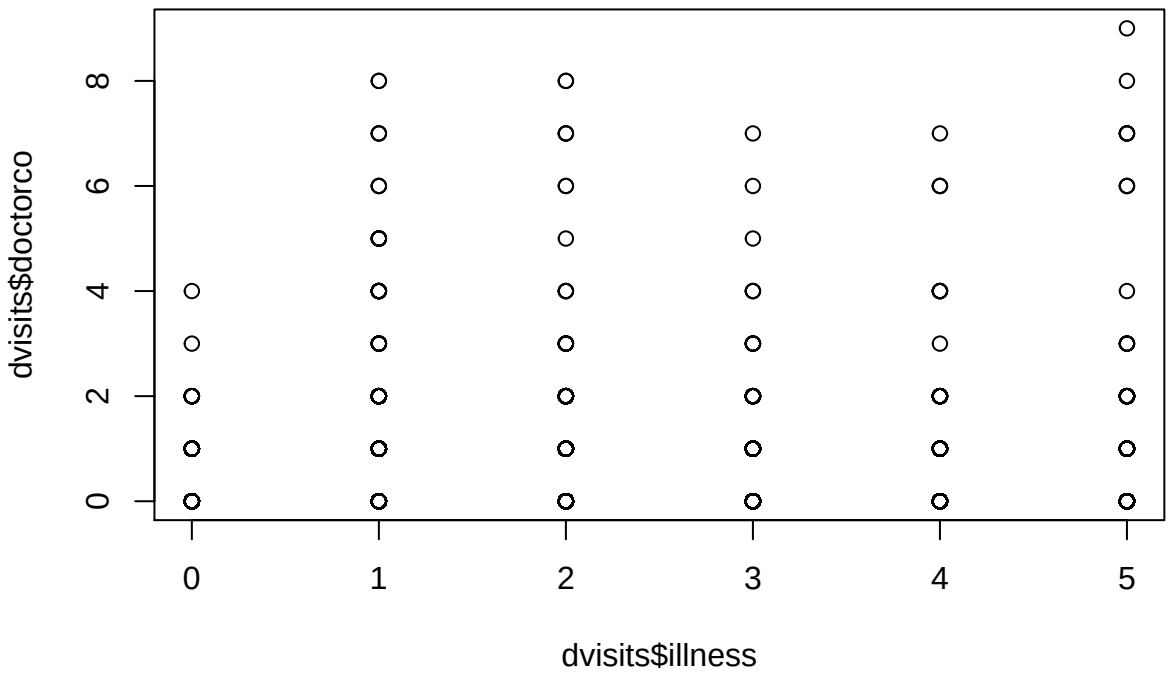
Solution 5

(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)
```

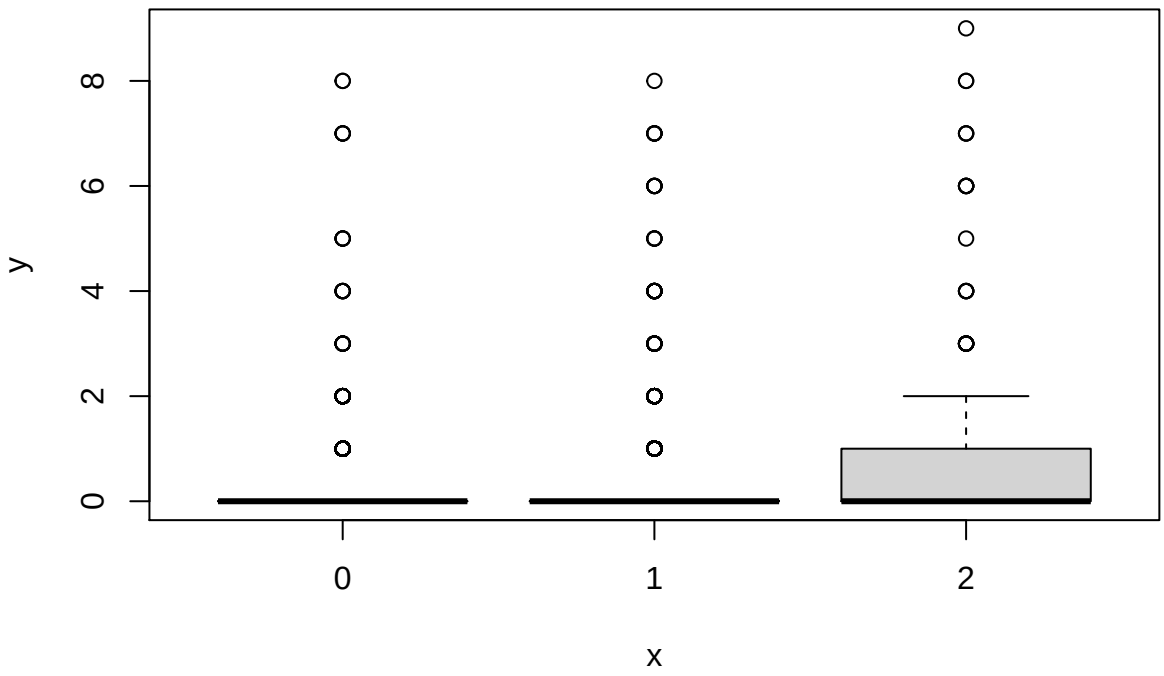


(b) Combine the predictors chcond1 and chcond2 into a single three-level factor. Make an appropriate plot showing the relationship between this factor and the response. Comment.

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)
```

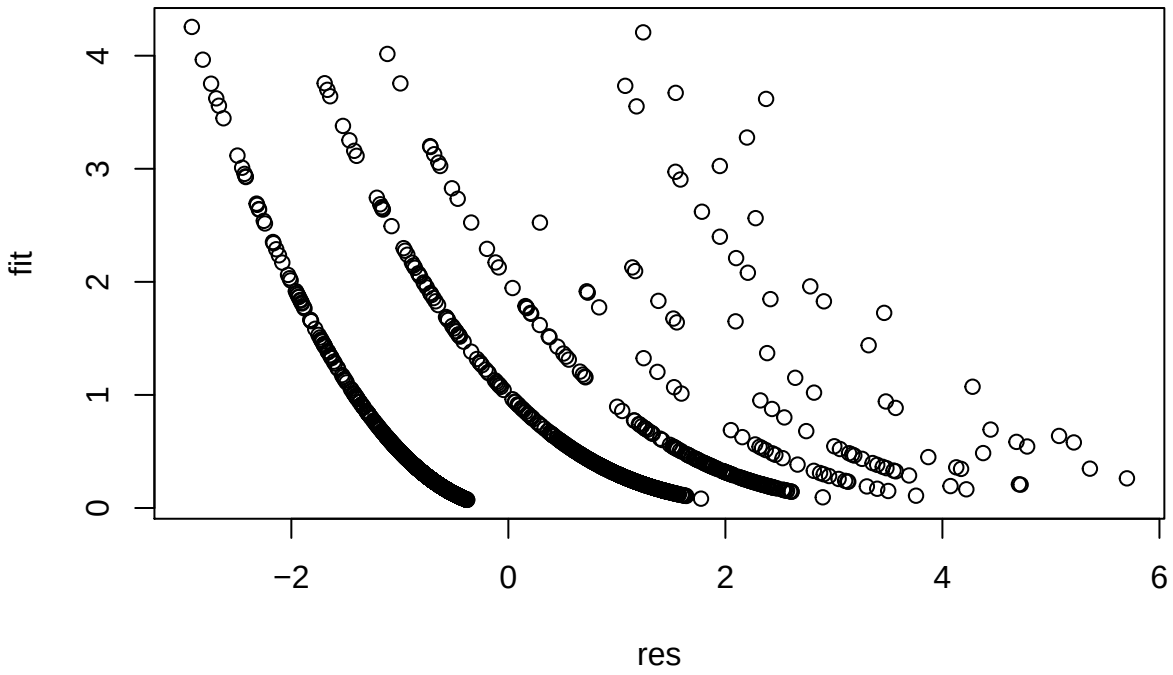
```
## [1] 0
```

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.

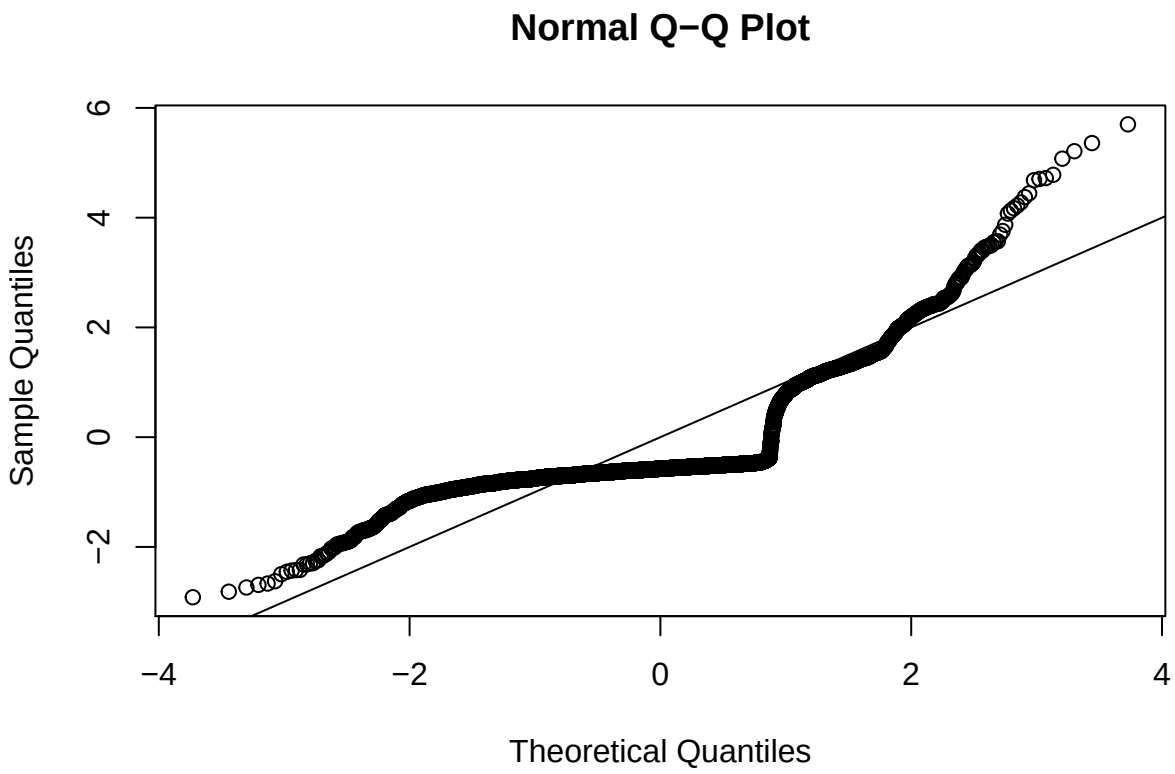
```
res = residuals(mod)
fit = fitted(mod)

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
```

```
# 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.

```
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.

```
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
```

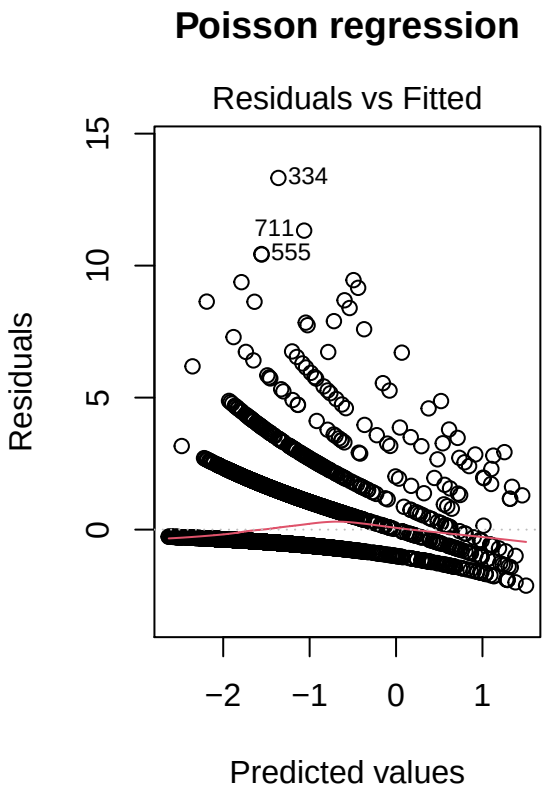
From the two tables, the observed and expected frequencies are close enough and thus it does not seem worth fitting a zero inflated model.

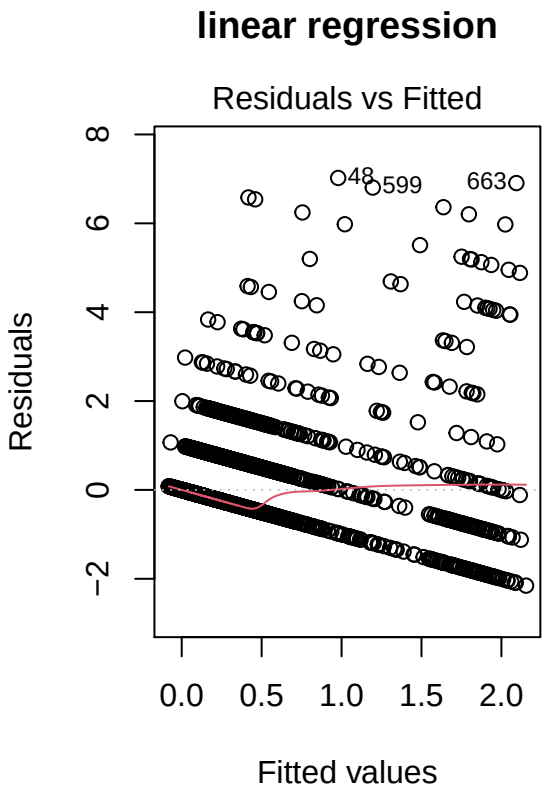
(h) Fit a comparable (Gaussian) linear model and graphically compare the fits. Describe how they differ.

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

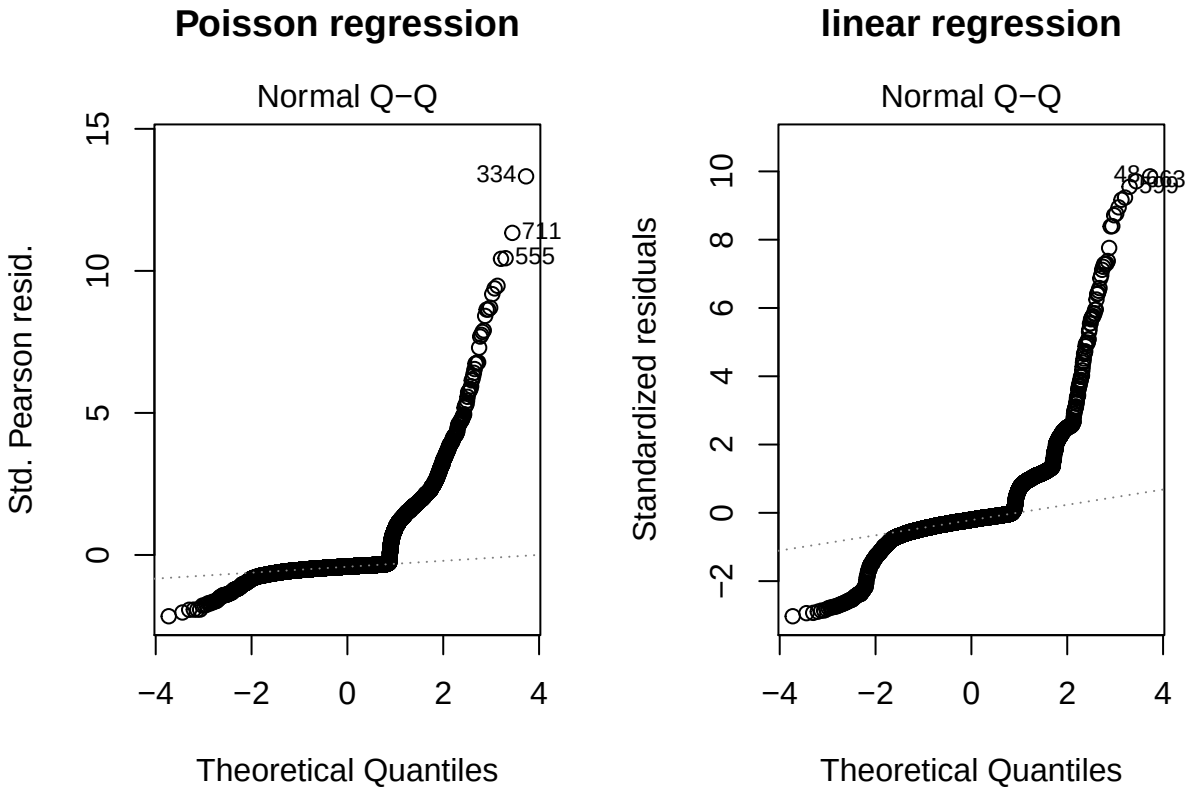
```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1543 -0.2584 -0.1440 -0.0434  7.0211
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   0.036781   0.035794   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
```

```
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 6[30 points]

This problem will expand on analyses of the CCSO data done in the course notes. The goal is to investigate racial biases in the propensity of people who spend at least one day in jail. We will restrict attention to other traffic offenses as done in class. Do the following:

This problem will expand on analyses of the CCSO data done in the course notes. The goal is to investigate racial biases in the propensity of people who spend at least one day in jail. We will restrict attention to other traffic offenses as done in class. Do the following:

- **part a** [15 points]: Analyze the CCSO data restricted to other traffic offenses using a binary response regression models. Your analysis must consider the following variable:
  - race
  - sex
  - arrestAge
  - employmentStatus
  - releaseReason
  - repeatOffenders: a variable that you will have to create to indicate whether or not an arrested individual was previously arrested.

  - multipleOffenses: a variable that you will have to create to indicate whether the arrested individual has committed multiple offenses upon a single arrest.

Note that "consider" does not mean that a variable has to be included in a regression model in this context; it means that your analysis needs to be informed by the above variables. You may want to combine factor levels in these variables, or you may want to throw out individuals belonging to a factor level that may be hard to interpret or is sparse. You are encouraged, but not required, to consider other variables. Report your final regression model, and justify your choice for your final model. Report interesting findings.

- **part b** [5 points]: Report observed propensities of spending at least one day in jail broken up by race and employment status after restricting attention to people who were released because of a bond payment. Comment on racial discrepancies.
- **part c** [5 points]: Pretend you are an expert witness in a court case where the Champaign County Sheriff's Office is being sued for racial bias in sentencing. Suppose you are hired by the side prosecuting the CCSO. Present an argument for racial bias in sentencing based on your analysis above. You can add further analyses if you think they are needed.
- **part d** [5 points]: Pretend you are an expert witness in a court case where the CCSO is being sued for racial bias in sentencing. Suppose you are hired by the CCSO. Present an argument that there is no racial bias in sentencing based on your analysis above and any additional analyses if you think they are needed. You are allowed to criticize the model you presented in part a.

Solution 6 part a:

We perform the relevant data wrangling to construct multipleOffenses and repeatOffenders. We ignore individuals that have more than one offense. This is to investiage the propensity of spending at least one day in jail for people who were booked with just OTHER TRAFFIC OFFESNES, and nothing else. Levels of releaseReason are combined.

```
library(tidyverse)
#library(stat528materials)
library(heatmapFit)
#data("CCSO")

## data wrangling
CCSO_small = CCSO %>%
  mutate(atleastone = ifelse(daysInJail > 0,1,0)) %>%
  filter(crimeCode == "OTHER TRAFFIC OFFENSES") %>%
  filter(race %in% c("Asian/Pacific Islander","Black","White","Hispanic")) %>%
  filter(sex %in% c("Female","Male")) %>%
  mutate(race = fct_drop(race), sex = fct_drop(sex)) %>%
  group_by(jacketNumber, bookingDate) %>%
  reframe(n = n(),
          releasedReason = releasedReason,
          daysInJail = daysInJail,
          race = race,
          sex = sex,
          city = city,
          arrestAgency = arrestAgency,
```

```
          employmentStatus = employmentStatus,
          arrestAge = arrestAge) %>%
  filter(n == 1) %>%
  ungroup() %>%
  group_by(jacketNumber) %>%
    reframe(repeatOffender = 1:n() - 1,
            releasedReason = releasedReason,
            daysInJail = daysInJail,
            race = race,
            sex = sex,
            city = city,
            arrestAgency = arrestAgency,
            employmentStatus = employmentStatus,
            arrestAge = arrestAge) %>%
  mutate(releasedReason = as.factor(releasedReason)) %>%
  mutate(atleastoneDay = ifelse(daysInJail > 0, 1, 0))

levels(CCSO_small$releasedReason) = c("Other","Bond Posted","Other","Other","Other","Bond Posted",
  "Other","Other","Other","Other","Other","Other","Other","Probation","Other","Personal Recognizance",
  "Transfer","Served","Transfer","Transfer","Transfer","Transfer","Transfer","Transfer")

CCSO_small = CCSO_small[complete.cases(CCSO_small), ]
dim(CCSO_small)
```

```
## [1] 2247   11
```

A main-effects only model revelas that black individuals are expected to have a higher propensity for spending at least one day in jail.

```
m1 = glm(atleastoneDay ~ race + sex + arrestAge +
           employmentStatus + releasedReason + repeatOffender,
         data = CCSO_small, family = "binomial")
summary(m1)
```

```
##
## Call:
## glm(formula = atleastoneDay ~ race + sex + arrestAge + employmentStatus +
##     releasedReason + repeatOffender, family = "binomial", data = CCSO_small)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7107  -0.5266  -0.3968  -0.2995   2.5823
##
## Coefficients:
##                                    Estimate Std. Error z value      Pr(>|z|)
## (Intercept)                       -1.645784   1.108357  -1.485      0.137574
## raceBlack                          1.739917   0.890785   1.953      0.050791
## raceHispanic                       0.913401   0.910594   1.003      0.315821
## raceWhite                          1.163289   0.891228   1.305      0.191802
## sexMale                            0.567644   0.176215   3.221      0.001276
```

```
## arrestAge                                0.002092   0.006213   0.337      0.736353
## employmentStatusEmployed - Full Time -0.977426   0.552015  -1.771      0.076619
## employmentStatusEmployed - Part Time -0.751722   0.570871  -1.317      0.187906
## employmentStatusLaid Off             -0.470092   1.185679  -0.396      0.691754
## employmentStatusRetired              -1.092182   0.939686  -1.162      0.245120
## employmentStatusSelf Employed        -0.340738   0.652055  -0.523      0.601280
## employmentStatusStudent              -1.178451   0.628180  -1.876      0.060659
## employmentStatusUnemployed            0.053640   0.550022   0.098      0.922310
## releasedReasonBond Posted            -1.678707   0.286228  -5.865 0.00000000449
## releasedReasonProbation               3.886521   1.057116   3.677      0.000236
## releasedReasonPersonal Recognizance  -1.570575   0.341792  -4.595 0.00000432507
## releasedReasonTransfer                2.799245   0.462400   6.054 0.00000000142
## releasedReasonServed                  3.597690   0.786763   4.573 0.00000481313
## repeatOffender                        0.215751   0.179750   1.200      0.230030
##
## (Intercept)
## raceBlack                            .
## raceHispanic
## raceWhite
## sexMale                              **
## arrestAge
## employmentStatusEmployed - Full Time .
## employmentStatusEmployed - Part Time
## employmentStatusLaid Off
## employmentStatusRetired
## employmentStatusSelf Employed
## employmentStatusStudent              .
## employmentStatusUnemployed
## releasedReasonBond Posted            ***
## releasedReasonProbation              ***
## releasedReasonPersonal Recognizance  ***
## releasedReasonTransfer               ***
## releasedReasonServed                 ***
## repeatOffender
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2328.5  on 2246  degrees of freedom
## Residual deviance: 1531.7  on 2228  degrees of freedom
## AIC: 1569.7
##
## Number of Fisher Scoring iterations: 6
```

Moreover, a likelihood ratio test suggests that race is an important variable at any reasonable testing level.

```
m1_small = glm(atleastoneDay ~ sex + arrestAge +
           employmentStatus + releasedReason + repeatOffender,
         data = CCSO_small, family = "binomial")
anova(m1_small, m1, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: atleastoneDay ~ sex + arrestAge + employmentStatus + releasedReason +
##     repeatOffender
## Model 2: atleastoneDay ~ race + sex + arrestAge + employmentStatus + releasedReason +
##     repeatOffender
##   Resid. Df Resid. Dev Df Deviance    Pr(>Chi)
## 1      2231     1558.0
## 2      2228     1531.7  3   26.231 0.000008531 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

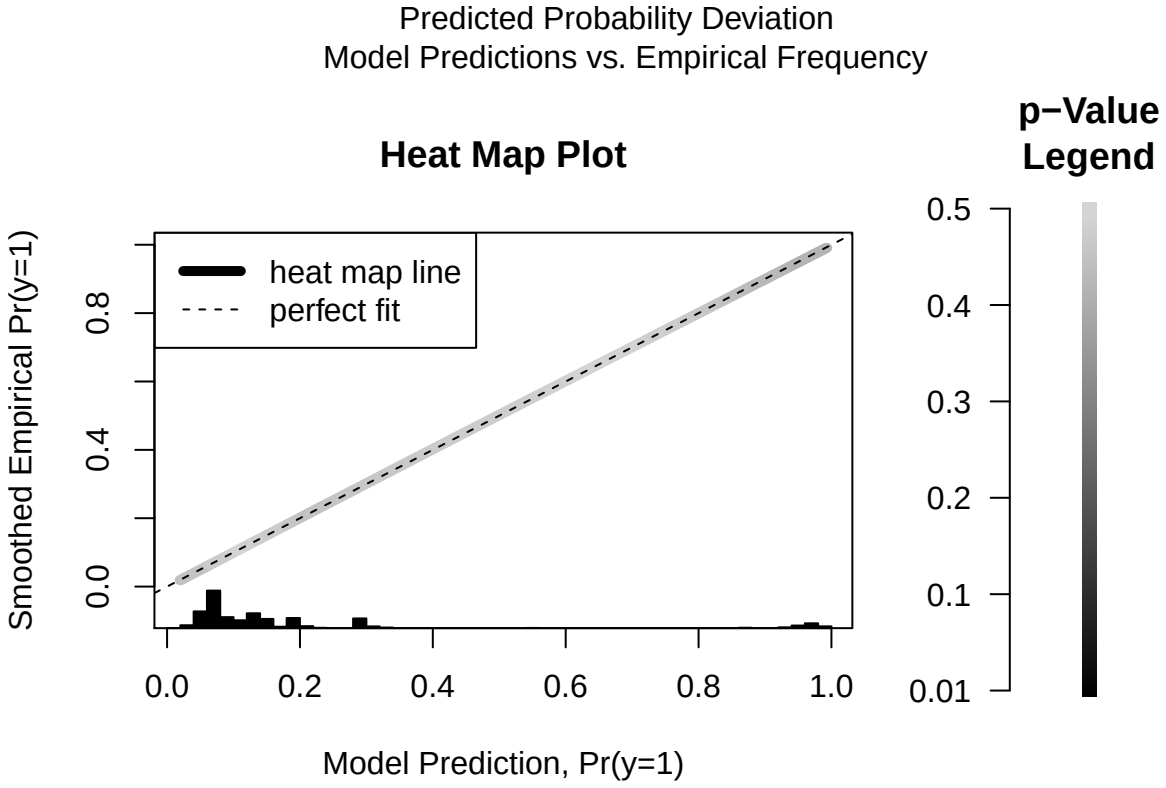
Our main-effect only model with race included fits better than a saturated model, and conditional success probability estimates conform with the observed data. This model seems to fit the data well.

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

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

```
p1 = predict(m1, type = "response")
y = CCSO_small$atleastoneDay
heatmap.fit(y = y, pred = p1)
```

```
##
## Calculating optimal loess bandwith...
## aicc Chosen Span =  0.9899359
##
## Generating Bootstrap Predictions...
##   |                                                                      |
```




```
##
##
## ******************************************
## 0% of Observations have one-tailed p-value <= 0.1
## Expected Maximum = 20%
## ******************************************
```

part b:

There seem to be racial disparities in the propensity for spending at least one day in jail across employment status. A careful look also reveals that there are racial disparities in employment rates. For example, 450 and 159 white people are, respectively, fully employed and unemployed. However, 226 and 220 black people are, respectively, fully employed and unemployed.

```
CCSO_small %>% filter(releasedReason == "Bond Posted") %>%
  dplyr::select(race, employmentStatus,
         atleastoneDay) %>%
  group_by(employmentStatus, race) %>%
  summarise(n = n(), prop = mean(atleastoneDay)) %>%
  as.data.frame()
```

```
##        employmentStatus                race   n       prop
## 1                                     Black   5 0.20000000
```

```
## 2                                    Hispanic   9 0.11111111
## 3                                       White   9 0.11111111
## 4  Employed - Full Time Asian/Pacific Islander  13 0.00000000
## 5  Employed - Full Time                   Black 230 0.12173913
## 6  Employed - Full Time                Hispanic 121 0.04132231
## 7  Employed - Full Time                   White 455 0.06373626
## 8  Employed - Part Time Asian/Pacific Islander   1 0.00000000
## 9  Employed - Part Time                   Black 103 0.19417476
## 10 Employed - Part Time                Hispanic  27 0.03703704
## 11 Employed - Part Time                   White 134 0.06716418
## 12             Laid Off                   Black   3 0.00000000
## 13             Laid Off                Hispanic   1 0.00000000
## 14             Laid Off                   White   2 0.50000000
## 15              Retired                   Black   8 0.12500000
## 16              Retired                   White  16 0.06250000
## 17        Self Employed Asian/Pacific Islander   1 0.00000000
## 18        Self Employed                   Black  14 0.21428571
## 19        Self Employed                Hispanic   4 0.00000000
## 20        Self Employed                   White  37 0.10810811
## 21              Student Asian/Pacific Islander  12 0.00000000
## 22              Student                   Black  51 0.13725490
## 23              Student                Hispanic  13 0.00000000
## 24              Student                   White  51 0.05882353
## 25           Unemployed Asian/Pacific Islander   6 0.33333333
## 26           Unemployed                   Black 222 0.26576577
## 27           Unemployed                Hispanic  64 0.14062500
## 28           Unemployed                   White 161 0.16770186
```

part c:

Our modeling of the propensity for spending at least one day in jail suggests that there are racial biases in sentencing. Moreover, this model offers a convincing description of the data as judged by statistical tests and model diagnostics for binary response regressions.

part d: The model may suggest racial disparities in sentencing. However, it appears that these disparities are more due to an individual's ability to pay a bond. Ability to pay a bond is associated with race, but not necessarily sentencing. More is needed to conclude that the CCSO exhibits racial bias in sentencing.