Problem 1

$Y \sim NB(k, p)$

The density of Y is given by:

$$f_Y(y) = \binom{y+k-1}{k-1} p^k (1-p)^y \; ; \; y = 0, 1, 2, \ldots$$

(a) 'k' is fixed and thus $f_Y(y)$ is a member of exponential family.

Now, $f_Y(y) = \exp\left\{ \log \binom{y+k-1}{k-1} + k \log p + y \log(1-p) \right\}$

$$= \exp\left\{ \frac{y \log(1-p) - (-k \log p)}{1} + \log \binom{y+k-1}{k-1} \right\} \quad -(*)$$

From the canonical form,

$\theta = \log(1-p)$ and $b(\theta) = -k \log p$

$\Rightarrow e^\theta = 1-p$ $\Rightarrow b(\theta) = \boxed{-k \log(1-e^\theta)}$

$\Rightarrow p = 1-e^\theta$

Again, $\mu = E(Y) = \frac{d}{d\theta} b(\theta) = \frac{d}{d\theta} \left\{ -k \log(1-e^\theta) \right\}$

$$= -k \frac{1}{1-e^\theta} (-e^\theta) = \frac{k e^\theta}{1-e^\theta} = \boxed{\frac{k(1-p)}{p}}$$

Now, from $(*)$, $a(\phi) = 1$

So, $V(Y) = a(\phi) \frac{d^2}{d\theta^2} b(\theta) = \frac{d}{d\theta} \left[\frac{d}{d\theta} b(\theta) \right] = \frac{d}{d\theta} \left[\frac{k e^\theta}{1-e^\theta} \right]$

$$= k \cdot \frac{e^\theta(1-e^\theta) - e^\theta(-e^\theta)}{(1-e^\theta)^2} = \frac{k \cdot e^\theta}{1-e^\theta} \cdot \left[\frac{1-e^\theta+e^\theta}{1-e^\theta} \right]$$

$$= \mu \left[\frac{1}{1-e^{\alpha}} \right] = \mu/p = \frac{k(1-p)}{p^2}$$

Again,

$$\mu + \frac{\mu^2}{k} = \frac{k(1-p)}{p} + \frac{k^2(1-p)^2}{p^2 k}$$

$$= \frac{pk^2(1-p) + k^2(1-2p+p^2)}{p^2 k}$$

$$= \frac{pk^2 - p^2k^2 + k^2 - 2pk^2 + p^2k^2}{p^2 k}$$

$$= \frac{k^2 - pk^2}{p^2 k} = \frac{k^2(1-p)}{p^2 k} = \frac{k(1-p)}{p^2}$$

Thus, $\text{Var}(Y) = \mu + \frac{\mu^2}{k}$ □.

(b) For the negative binomial distribution, the canonical link is given by $\log\left(\frac{\mu}{k+\mu}\right)$.

To show this, let $g(\cdot)$ be the link function such that,

$g(\mu) = \theta$

$\Rightarrow g(\mu) = \log(1-p)$

$\Rightarrow g(\mu) = \log\left[1 - \frac{k}{\mu+k}\right]$

$\Rightarrow g(\mu) = \log\left(\frac{\mu}{k+\mu}\right)$ □

[since, $\mu = \frac{k(1-p)}{p} \Rightarrow \mu p = k - kp$

$\Rightarrow p(\mu+k) = k$

$\Rightarrow p = \frac{k}{\mu+k}$]

(e) We know that,

$$\text{Deviance} = \frac{2}{a(\phi)} \sum_i \{ y_i(\tilde{\theta}_i^* - \hat{\theta}_i^*) - b(\tilde{\theta}_i^*) + b(\hat{\theta}_i^*) \}$$

Here, $a(\phi) = 1$, $\tilde{\theta}_i^* = \theta(y_i)$ and $\hat{\theta}_i^* = \theta(\hat{\mu}_i)$

So, $\tilde{\theta}_i^* - \hat{\theta}_i^* = \log(k y_i + y_i \hat{\mu}_i) - \log(k \tilde{\mu}_i + y_i \hat{\mu}_i)$

Also, $b(\tilde{\theta}_i^*) - b(\hat{\theta}_i^*) = -k \log(k + \hat{\mu}_i) + k \log(k + y_i)$

As from the previous parts,

$$b(\theta) = -k \log(1 - e^{\theta}) \quad [\text{where } \theta = \log\left(\frac{\mu}{\mu + k}\right)] \quad -(*)$$

$$= -k \log\left(1 - \frac{\mu}{\mu + k}\right) = -k \log(\mu + k)$$

Thus, $b(\tilde{\theta}_i^*) - b(\hat{\theta}_i^*) = k \log\left(\frac{k + y_i}{k + \hat{\mu}_i}\right)$

$$\Rightarrow -b(\tilde{\theta}_i^*) + b(\hat{\theta}_i^*) = k \log\left(\frac{k + \hat{\mu}_i}{k + y_i}\right)$$

From (*),

$\tilde{\theta}_i^* = \log\left(\frac{y_i}{y_i + k}\right)$ and $\hat{\theta}_i^* = \log\left(\frac{\hat{\mu}_i}{\hat{\mu}_i + k}\right)$

So, $\tilde{\theta}_i^* - \hat{\theta}_i^* = \log\left(\frac{y_i}{y_i + k}\right) - \log\left(\frac{\hat{\mu}_i}{\hat{\mu}_i + k}\right) = \log\left[\frac{y_i(\hat{\mu}_i + k)}{\hat{\mu}_i(y_i + k)}\right]$

So, $$\text{Deviance} = 2 \sum_{i=1}^{n} \left[y_i \log\left\{\frac{y_i \hat{\mu}_i + y_i k}{\hat{\mu}_i y_i + \hat{\mu}_i k}\right\} + k \log\left(\frac{k + \hat{\mu}_i}{k + y_i}\right)\right]$$

(d) Assuming a natural log link, i.e., $\eta_i = \log \mu_i$

observations ranging from $i = 1(1)n$ parameters ranging from $j = 1(1)J$

in $\eta_i = x_i'\beta$.

Now, we know that:

$$\frac{\partial L(\beta, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{a(\phi)} \times \frac{1}{V(\mu_i)} \times \frac{\partial \mu_i}{\partial \eta_i} \times x_{ij} \right] \quad \text{— } *$$

From previous parts, $1/V(\mu_i) = 1/\left(\mu_i + \frac{\mu_i^2}{k}\right) = \frac{k}{\mu_i(k+\mu_i)}$

Also, $\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\frac{\partial \eta_i}{\partial \mu_i}} = \frac{1}{\frac{\partial \log \mu_i}{\partial \mu_i}} = \mu_i$

So, (*) becomes

$$\frac{\partial L(\beta, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \left[(y_i - \mu_i)\mu_i \times \frac{k}{\mu_i(k+\mu_i)} \times x_{ij} \right]$$

$$\Rightarrow \frac{\partial L(\beta, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{k(y_i - \mu_i))\, x_{ij}}{(k+\mu_i)} \right]$$

□

(e) We know that the $(j,k)^{th}$ element of the Fisher Information matrix $= \sum_{i=1}^{n} \frac{w_i}{a(\phi)} x_{ij} x_{ik}$ $[= I_{jk}]$

where, $w_i = \left\{ \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V(\mu_i) \right\}^{-1} = \left\{ \left(\frac{\partial \log \mu_i}{\partial \mu_i} \right)^2 \times \left(\mu_i + \frac{\mu_i^2}{k} \right) \right\}^{-1}$

$$= \left\{ \frac{1}{\mu_i^2} \times \mu_i \left(1 + \frac{\mu_i}{k}\right) \right\} = \left\{ \frac{1}{\mu_i} + \frac{1}{k} \right\}^{-1} = \frac{k\mu_i}{k+\mu_i}$$

Thus,

$$I_{jk} = \sum_{i=1}^{n} \frac{k\mu_i}{k+\mu_i} x_{ij} x_{ik}$$

□

(f) The mathematical expression of $\hat{\beta}^{(1)}$ is :

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}^{(1)} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}^{(0)} + I^{-1}(\beta) \begin{pmatrix} \frac{\partial}{\partial \beta_1} L(\beta, \phi) \\ \frac{\partial}{\partial \beta_2} L(\beta, \phi) \\ \vdots \\ \frac{\partial}{\partial \beta_p} L(\beta, \phi) \end{pmatrix} \Bigg|_{\beta = \beta^{(0)}}$$

where, $I_{jk}(\beta) = \sum_{i=1}^{n} \frac{k\mu_i}{k+\mu_i} \cdot x_{ij} x_{ik}$

where, $\frac{\partial L}{\partial \beta} = \frac{1}{a(\phi)} F^T V^{-1} (y - \mu)$, here, $a(\phi) = 1$.

where, $F^{n \times p = 3} = \frac{\partial \underset{\sim}{\mu}}{\partial \underset{\sim}{\beta}} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \frac{\partial \mu_1}{\partial \beta_2} & \cdots & \frac{\partial \mu_1}{\partial \beta_p} \\ \vdots & & & \\ \frac{\partial \mu_n}{\partial \beta_1} & \frac{\partial \mu_n}{\partial \beta_2} & \cdots & \frac{\partial \mu_n}{\partial \beta_p} \end{pmatrix}$, Here, $p = 3$, $n = 7$

Here, the link function is, $\eta_i = \log \mu_i$

$$\Rightarrow \underset{\sim}{x_i}' \underset{\sim}{\beta} = \log \mu_i$$

$$\Rightarrow \boxed{\mu_i = \exp\{ \underset{\sim}{x_i}' \underset{\sim}{\beta} \}}$$

Thus, $F = \begin{pmatrix} x_{11} e^{\underset{\sim}{x_1}' \underset{\sim}{\beta}} & x_{12} e^{\underset{\sim}{x_1}' \underset{\sim}{\beta}} & x_{13} e^{\underset{\sim}{x_1}' \underset{\sim}{\beta}} \\ \vdots & & \\ x_{71} e^{\underset{\sim}{x_7}' \underset{\sim}{\beta}} & x_{72} e^{\underset{\sim}{x_7}' \underset{\sim}{\beta}} & x_{73} e^{\underset{\sim}{x_7}' \underset{\sim}{\beta}} \end{pmatrix}$

and $V^{n=7 \times p=3} = \begin{pmatrix} V(\mu_1) & 0 & \cdots & 0 \\ 0 & V(\mu_2) & 0 & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & V(\mu_7) \end{pmatrix} = \begin{pmatrix} \mu_1 + \mu_1^2/k & 0 & \cdots & 0 \\ 0 & \mu_2 + \mu_2^2/k & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & \mu_7 + \mu_7^2/k \end{pmatrix}$

□

STAT_8320_HW2_JayadityaNath

JAYADITYA NATH

2024-03-04

Part (f)

```
k = 15
data_1 <- data.frame(y = c(3, 17, 12, 1, 19, 9, 8),x_1 = c(0, 1, 1, 0, 1, 0, 1),x_2 = c(35, 60, 25, 20,
beta_0_hat =  c(2.7, -0.4, -0.01)
design_mat = cbind(1, data_1$x_1, data_1$x_2)

mu = exp(design_mat %*% beta_0_hat )
V = diag(c(mu + (mu^2/k)))

F = matrix(data = NA, nrow = 7, ncol = 3)

for(i in 1:nrow(F)){
  for(j in 1:ncol(F)){
    F[i,j] = design_mat[i,j]*mu[i]
  }
}

library(matlib)
beta_1_hat <- beta_0_hat + inv(t(F) %*% inv(V) %*% F) %*% (t(F) %*% inv(V) %*% (data_1$y - mu))
print("The value of F is : ")
```

```
## [1] "The value of F is : "
```

```
F
```

```
##           [,1]     [,2]     [,3]
## [1,] 10.485570 0.000000 366.9949
## [2,]  5.473947 5.473947 328.4368
## [3,]  7.767901 7.767901 194.1975
## [4,] 12.182494 0.000000 243.6499
## [5,]  6.049647 6.049647 302.4824
## [6,]  8.584858 0.000000 472.1672
## [7,]  7.389056 7.389056 221.6717
```

```
print("The value of V is : ")
```

```
## [1] "The value of V is : "
```

```
V
```

```
##          [,1]     [,2]     [,3]    [,4]    [,5]     [,6]     [,7]
## [1,] 17.81538 0.000000  0.00000  0.0000 0.00000  0.00000  0.00000
## [2,]  0.00000 7.471554  0.00000  0.0000 0.00000  0.00000  0.00000
## [3,]  0.00000 0.000000 11.79059  0.0000 0.00000  0.00000  0.00000
## [4,]  0.00000 0.000000  0.00000 22.0767 0.00000  0.00000  0.00000
## [5,]  0.00000 0.000000  0.00000  0.0000 8.48953  0.00000  0.00000
## [6,]  0.00000 0.000000  0.00000  0.0000 0.00000 13.49818  0.00000
## [7,]  0.00000 0.000000  0.00000  0.0000 0.00000  0.00000 11.02893
```

```
print("The value of mu is : ")
```

```
## [1] "The value of mu is : "
```

```
mu
```

```
##           [,1]
## [1,] 10.485570
## [2,]  5.473947
## [3,]  7.767901
## [4,] 12.182494
## [5,]  6.049647
## [6,]  8.584858
## [7,]  7.389056
```

The value of $\hat{\beta}^{(1)}$ is (0.5940271, 1.1058426, 0.0335505)'.

Part (g)

```
L = matrix(c(0,1,0,0,1,1), nrow = 2, byrow = TRUE)
diff_vec = c(-0.1, -0.5)
W = t(L %*% beta_0_hat - diff_vec) %*% inv(L %*% (inv(t(F) %*% inv(V) %*% F)) %*% t(L)) %*% (L %*% beta
p_val = 1 - pchisq(W, df = qr(L)$rank)

print("The value of L is : ")
```

```
## [1] "The value of L is : "
```

```
L
```

```
##      [,1] [,2] [,3]
## [1,]    0    1    0
## [2,]    0    1    1
```

The value of the Wald test statistic is 1139.8553869.

The p-value for the given test is 0.

As he p-value is less than $\alpha = 0.05$, we reject the null hypothesis.

Part (h)

```r
data_1_h <- read.csv("C:/Users/Jayaditya Nath/Documents/nbreg.csv")

library(MASS)
mod_h <- glm.nb(daysabs ~ male + math + langarts, data = data_1_h)
summary(mod_h)
```

```
##
## Call:
## glm.nb(formula = daysabs ~ male + math + langarts, data = data_1_h,
##     init.theta = 0.7761669366, link = log)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.716069   0.234174  11.598  < 2e-16 ***
## male        -0.431185   0.139516  -3.091  0.00200 **
## math        -0.001601   0.005300  -0.302  0.76259
## langarts    -0.014348   0.005372  -2.671  0.00756 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.7762) family taken to be 1)
##
##     Null deviance: 378.43  on 315  degrees of freedom
## Residual deviance: 356.93  on 312  degrees of freedom
## AIC: 1771.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.7762
##           Std. Err.:  0.0742
##
##  2 x log-likelihood:  -1761.7460
```

```r
mod_h_null = model <- glm.nb(daysabs ~ 1, data = data_1_h)
summary(mod_h_null)
```

```
##
## Call:
## glm.nb(formula = daysabs ~ 1, data = data_1_h, init.theta = 0.7156822877,
##     link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.75960    0.07047   24.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.7157) family taken to be 1)
##
##     Null deviance: 356.99  on 315  degrees of freedom
## Residual deviance: 356.99  on 315  degrees of freedom
```

```
## AIC: 1786.5
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.7157
##           Std. Err.:  0.0667
##
##  2 x log-likelihood:  -1782.4850
```

```
dev_test = mod_h$null.deviance - mod_h$deviance
p_val_dev_test = 1-pchisq(q = dev_test, df = 1)

over_disp_0_1 = mod_h$deviance/mod_h$df.residual
```

Primarily, the outputs of the negative-binomial model show that 'male' and 'langarts' seem to be significant predictors for predicting the attendance behaviour of high school juniors. This means that for one unit increase in male and langarts, we would get to see a 0.43 and 0.01 decrease in the expected log-count of the number of days absent.

The p-value $= 3.5379304 \times 10^{-6}$ of the deviance test being less than 0.05, we have strong evidence against the null hypothesis mentioning that the null model(intercept only) is better.

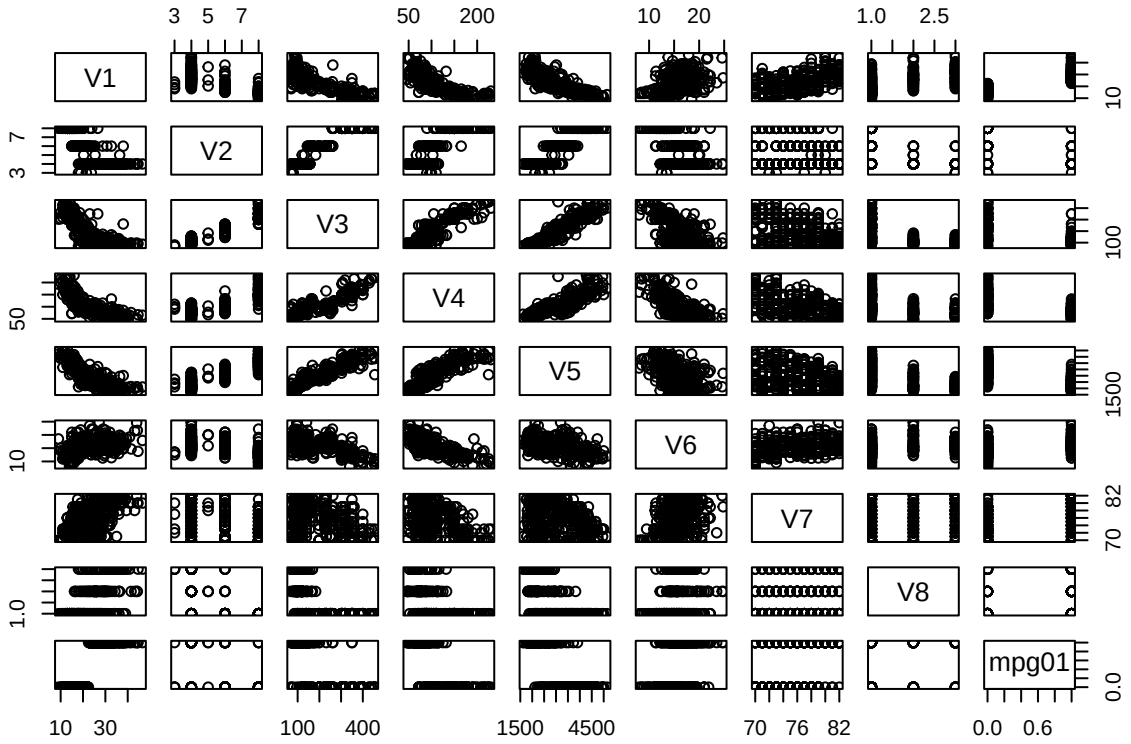
The value of the dispersion parameter $= 1.1440217$ being more than 1 depicts that there exists a problem of over-dispersion.

PROBLEM 2.

```
data_2_trai = read.table("C:/Users/Jayaditya Nath/Documents/auto_mpg_data2_train.dat")
data_2_test = read.table("C:/Users/Jayaditya Nath/Documents/auto_mpg_data2_test.dat")
mpg01 <- ifelse(data_2_trai$V1 > median(data_2_trai$V1), 1, 0)
data_2_train = cbind.data.frame(data_2_trai,mpg01)
```

Part (a)

```
pairs(data_2_train)
```



From the pairs plot, it seems that cylinders, horsepower, weight and acceleration seems to play a significant role in the prediction of the response variable. Although, there seems to some problem of multicollinearity existing between some of the predictors.

Part (b)

```
library(bestglm)
```

```
## Warning: package 'bestglm' was built under R version 4.3.3
```

```
## Loading required package: leaps
```

```
dat_2 = cbind(data_2_trai[,2:8],mpg01=data_2_train$mpg01)

mod_2 = bestglm(dat_2,IC="AIC",family = binomial,method = "exhaustive")
```

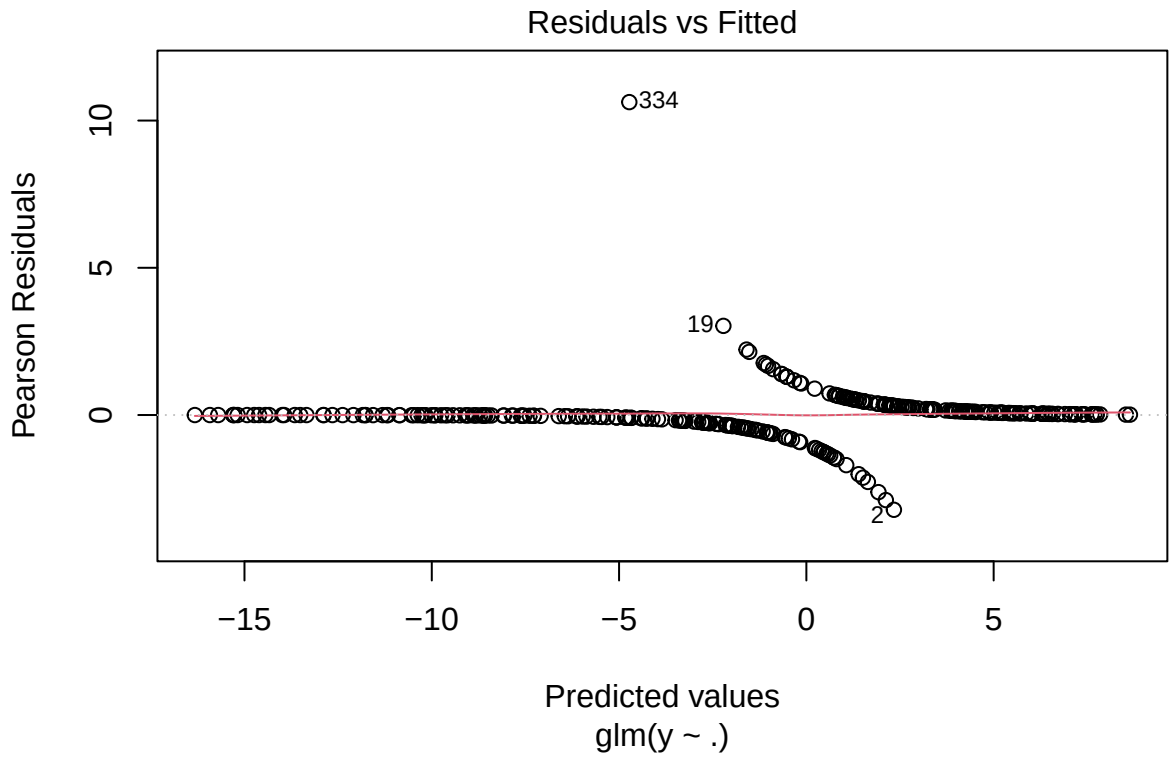
```
## Morgan-Tatar search since family is non-gaussian.
```

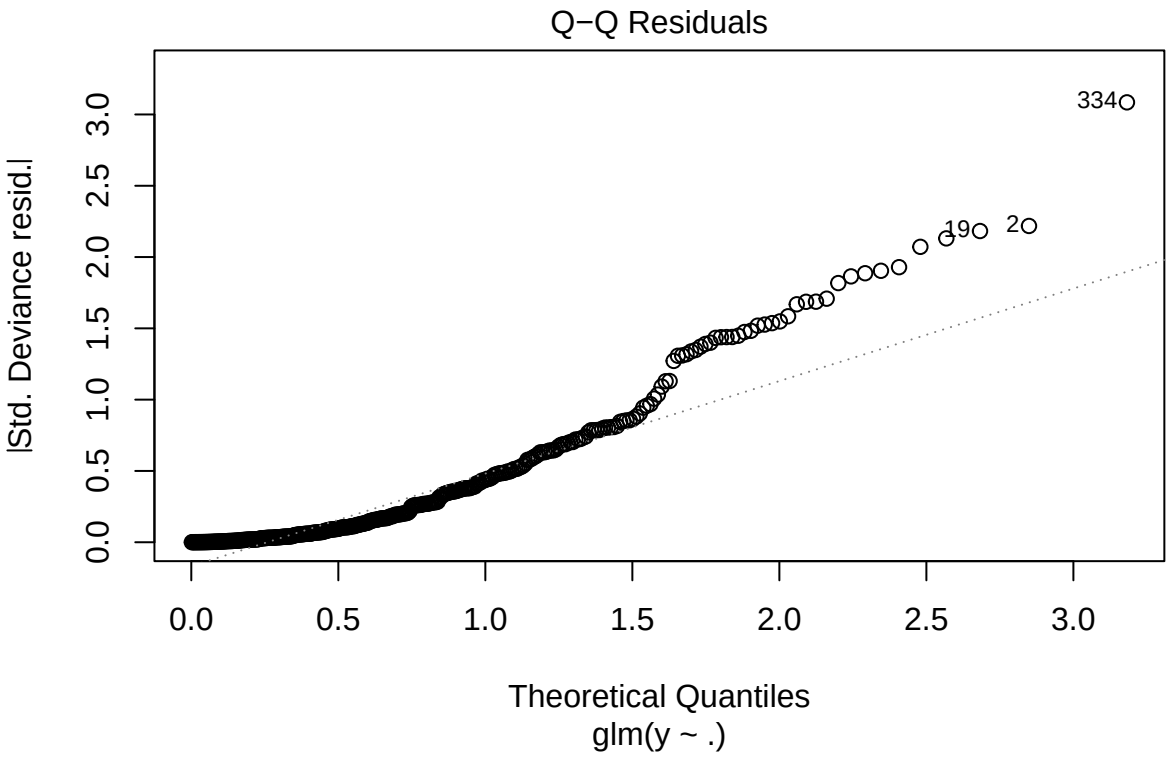
```
summary(mod_2$BestModel)
```

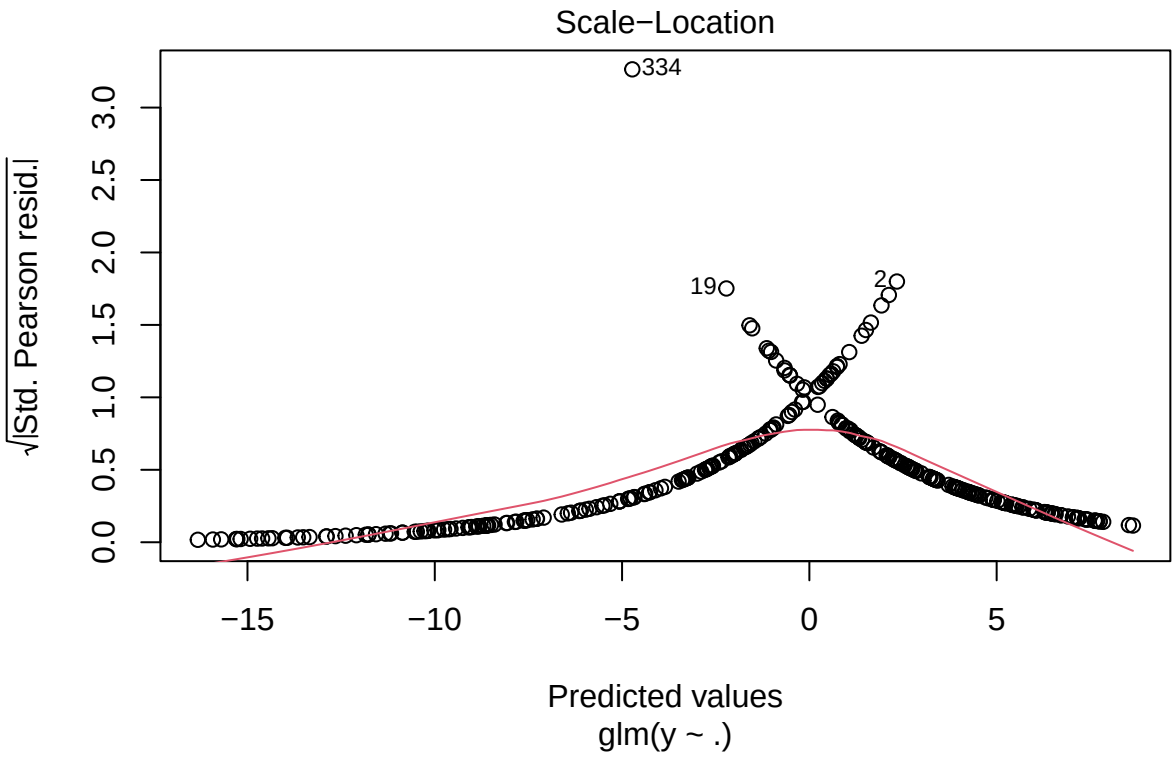
```
##
## Call:
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
```

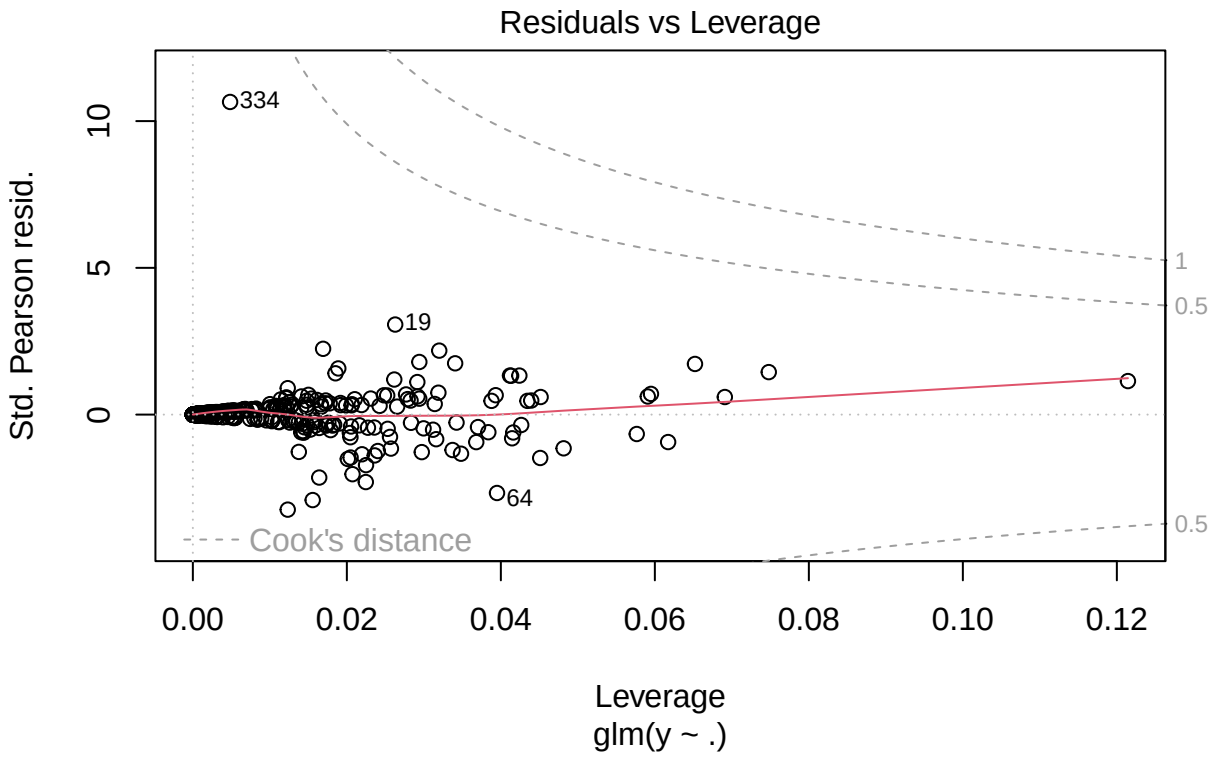
```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -16.419180   5.156240  -3.184  0.00145 **
## V4           -0.042507   0.015950  -2.665  0.00770 **
## V5           -0.004495   0.000675  -6.660 2.75e-11 ***
## V7            0.437192   0.079324   5.511 3.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 474.11  on 341  degrees of freedom
## Residual deviance: 140.21  on 338  degrees of freedom
## AIC: 148.21
##
## Number of Fisher Scoring iterations: 8
```

```r
par(mfrow=c(1,1))
plot(mod_2$BestModel)
```




Q−Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
334
19
2
Theoretical Quantiles
glm(y ~ .)


Scale−Location
334
19
2
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−15
−10
−5
0
5
Predicted values
glm(y ~ .)



From the model summary, we can see that horsepower, weight and model year seem to be significant covariates in predicting the response mpg01.

Also, from the plots, we can see that the residuals are non-normally distributed from the QQ-plot. The residual vs fitted shows that the residuals have heterogeneous variance and also seem to be independently distributed.

Finding the confusion matrix for the train set,

```
predict_train <- ifelse(predict(mod_2$BestModel, newdata = data_2_train, type = "response") > 0.5, 1, 0)

conf_mat_train <- table(mpg01, predict_train)
conf_mat_train
```

```
##       predict_train
## mpg01   0   1
##     0 152  19
##     1  16 155
```

Part (c)

```
mpg01_test = ifelse(data_2_test$V1 > median(data_2_train$V1), 1, 0)

predict_test = ifelse(predict(mod_2$BestModel, newdata = data_2_test, type = "response") > 0.5, 1, 0)

conf_mat_test = table(mpg01_test, predict_test)
conf_mat_test
```

```
##           predict_test
## mpg01_test  0  1
##          0 19  5
##          1  1 25
```

```
sensitivity_test = conf_mat_test["0","0"]/(conf_mat_test["0","0"]+conf_mat_test["0","1"])
```

The sensitivity value of 0.7916667 is quite decent for prediction purpose.

```
library(pROC)
```

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

##
## Attaching package: 'pROC'

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

```
library(ROCR)
auc_val = auc(roc(mpg01_test,predict_test))
```

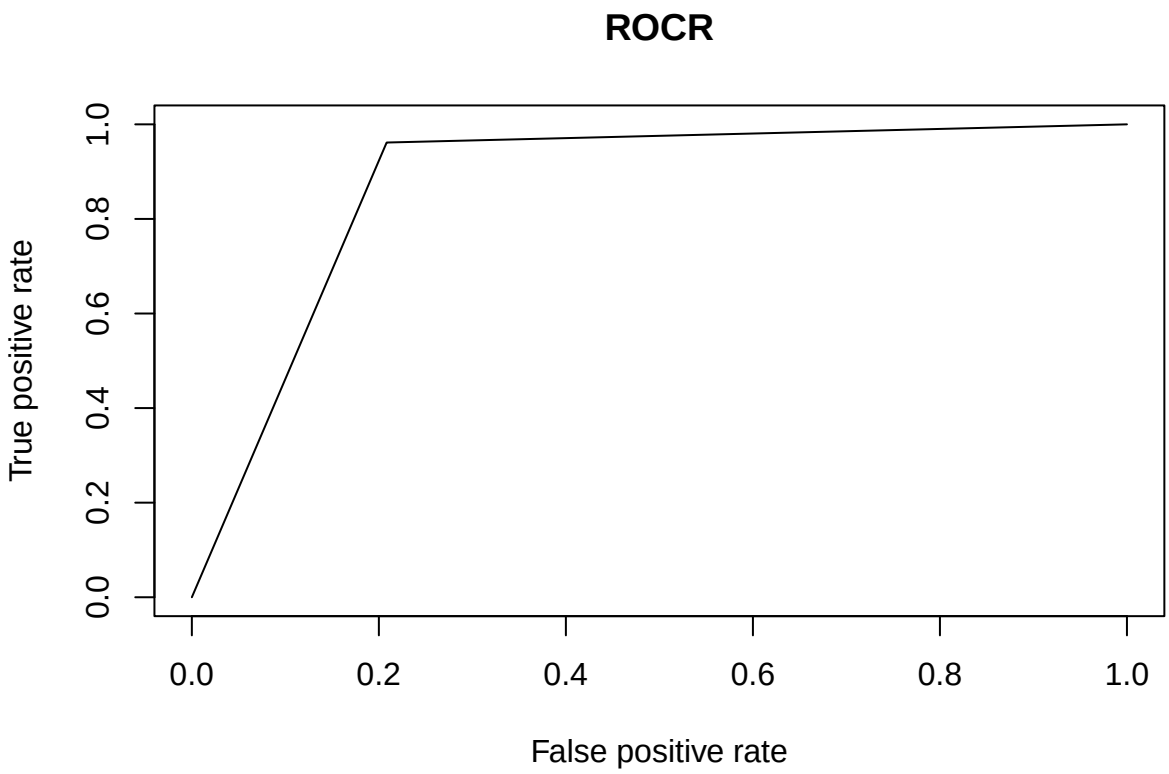
```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
auc_val
```

```
## Area under the curve: 0.8766
```

```
par(mfrow=c(1,1))
plot(performance(prediction(predict_test,mpg01_test),"tpr","fpr"),main="ROCR")
```



The high AUC value and the Receiver Operating Characteristics curve shows that the selected predictor variables would quite accurately predict the value of mpg01 using a generalized regression model.

PROBLEM 3.

```
data_3 = read.table("C:/Users/Jayaditya Nath/Documents/benthicfish.dat")

X1 = ifelse(data_3$V2==1,1,0)
X2 = ifelse(data_3$V2==2,1,0)
X3 = ifelse(data_3$V2==3,1,0)

data_3 = cbind(data_3,X1,X2,X3)
```

Part (a)

```
mod_3 = glm(data_3$V1 ~ matrix(c(X1,X2,X3),ncol = 3),data = data_3, family = poisson)
summary(mod_3)
```

```
##
## Call:
## glm(formula = data_3$V1 ~ matrix(c(X1, X2, X3), ncol = 3), family = poisson,
##     data = data_3)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       0.7492     0.1348   5.556 2.75e-08 ***
```

```
## matrix(c(X1, X2, X3), ncol = 3)1   1.0086     0.1443   6.988 2.79e-12 ***
## matrix(c(X1, X2, X3), ncol = 3)2   0.5124     0.1504   3.408 0.000654 ***
## matrix(c(X1, X2, X3), ncol = 3)3   0.2183     0.1954   1.117 0.263811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1760.3  on 173  degrees of freedom
## Residual deviance: 1672.6  on 170  degrees of freedom
## AIC: 1992.8
##
## Number of Fisher Scoring iterations: 6
```

```
anova(mod_3,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: data_3$V1
##
## Terms added sequentially (first to last)
##
##
##                                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                              173     1760.3
## matrix(c(X1, X2, X3), ncol = 3)  3   87.667       170     1672.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that macrohabitat type is statistically significant as the p-value from the ANOVA test is less than 0.05, thus having strong evidence against the null hypothesis. Also, the macrohabitat types 1,2 and 4 are statistically significant on the basis of the p-value obtained.

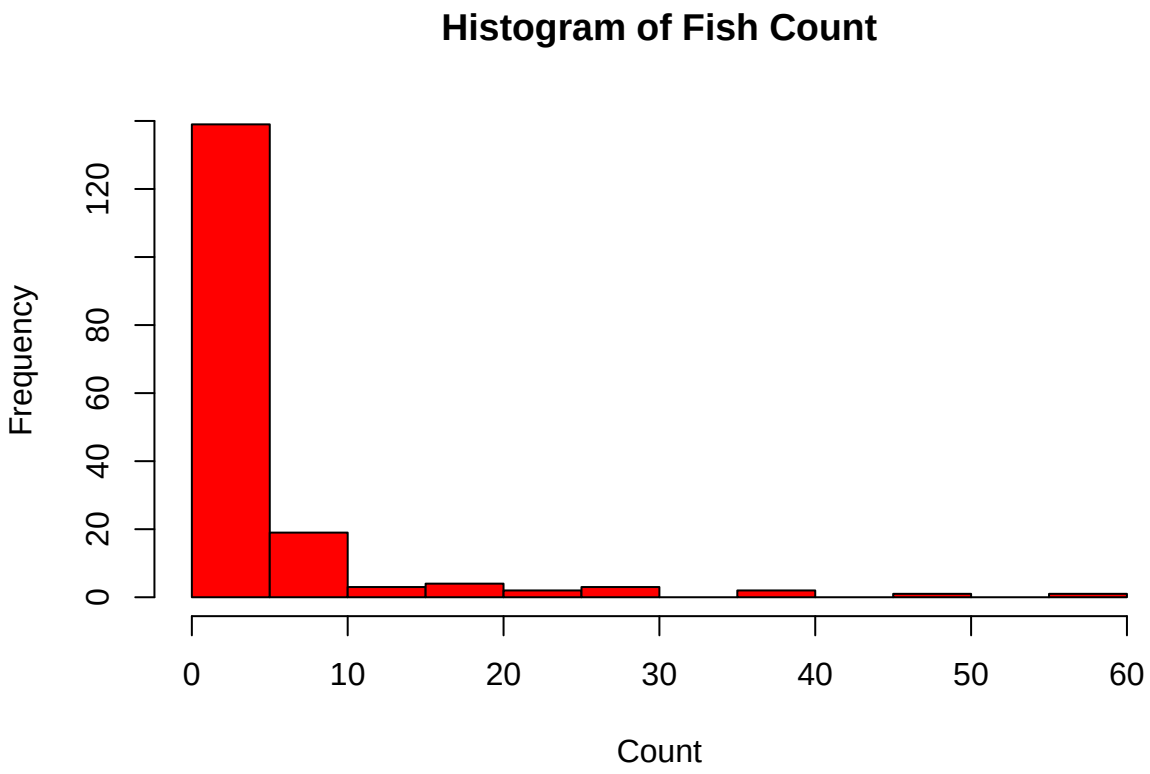
```
over_disp_0_1 = mod_3$deviance/mod_3$df.residual
over_disp_0_1
```

```
## [1] 9.838986
```

The value of the dispersion parameter = 9.8389864 is more than 1, thus indicating a problem of over-dispersion in the fit of the model to the given data.

Part (b)

```
hist(data_3$V1, main = "Histogram of Fish Count", xlab = "Count", ylab = "Frequency",col="red")
```



I do not find anything astounding as it is clear from the histogram that the data is immensely over-dispersed, which completely is similar to the results obtained in part (a).

Part (c)

```
library(pscl)
```

```
## Warning: package 'pscl' was built under R version 4.3.3
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```
# Fit zero-inflated Poisson model
zero_inf_mod <- zeroinfl(V1 ~ as.factor(V2) | V3, data = data_3, dist = "poisson", link = "logit")
summary(zero_inf_mod)
```

```
##
## Call:
## zeroinfl(formula = V1 ~ as.factor(V2) | V3, data = data_3, dist = "poisson",
##     link = "logit")
##
```

```
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.05051 -0.98940 -0.73977  0.04976 11.00179
##
## Count model coefficients (poisson with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.37687    0.05151  46.145  < 2e-16 ***
## as.factor(V2)2 -0.45415    0.08437  -5.383 7.32e-08 ***
## as.factor(V2)3 -1.18475    0.15306  -7.741 9.90e-15 ***
## as.factor(V2)4 -1.03031    0.14984  -6.876 6.15e-12 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.32126    0.36447  -0.881    0.378
## V3           0.02582    0.10974   0.235    0.814
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 12
## Log-likelihood: -682.4 on 6 Df
```

I get similar results for the zero inflated Poisson model as the macrohabitat types 1, 2 and 4 are still statistically significant, but, the predictors in the inflation part are not significant and thus we can conclude that the zero-inflated Poisson regression is not a good fit to predict the number of excess zeroes for the given count data.