Parin Patel
IST 772
HW9

## Attribution Statement:

Homework 9 by Parin Patel: I did this homework by myself, with help from the book and the professor.

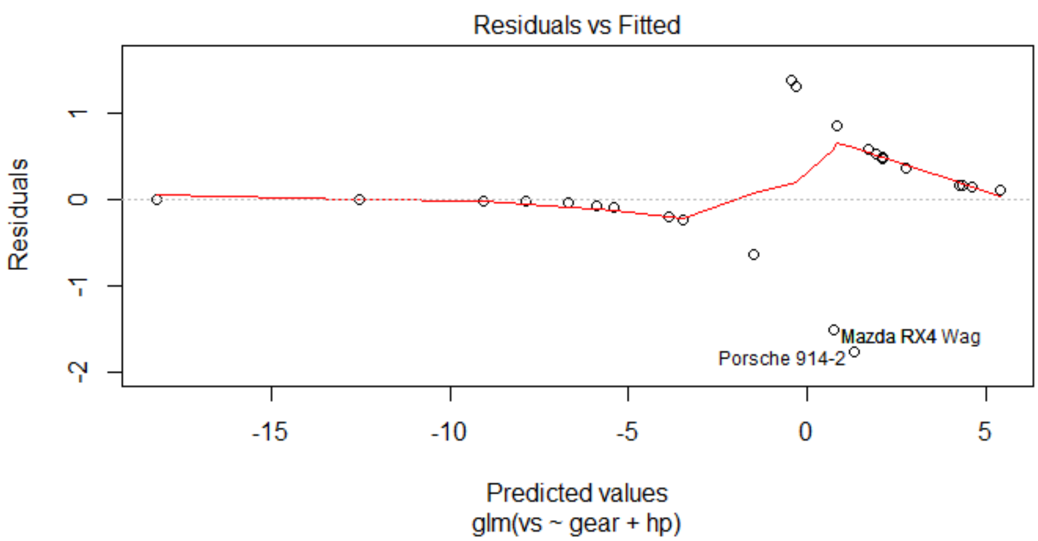
## Exercises:

1. **The built-in data sets of R include one called "mtcars," which stands for Motor Trend cars. Motor Trend was the name of an automotive magazine and this data set contains information on cars from the 1970s. Use "?mtcars" to display help about the data set. The data set includes a dichotomous variable called vs, which is coded as 0 for an engine with cylinders in a v-shape and 1 for so called "straight" engines. Use logistic regression to predict vs, using two metric variables in the data set, gear (number of forward gears) and hp (horsepower). Interpret the resulting null hypothesis significance tests.**

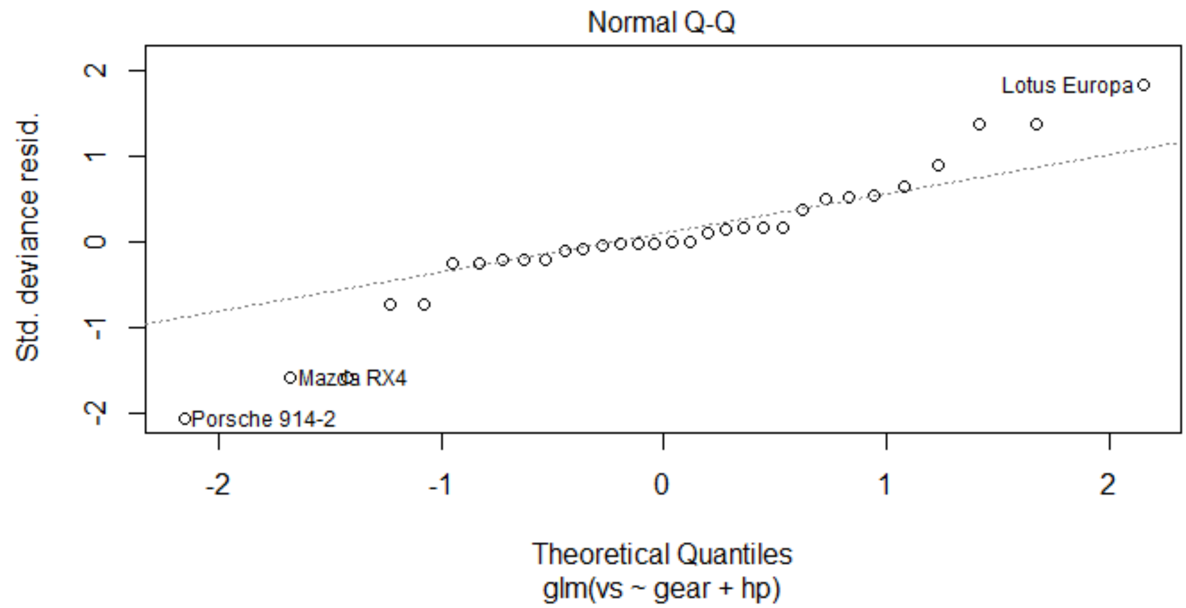
Using the results of our logistic regression to predict ,vs (engine shape), using the horsepower and gears as independent variables, showed that you cannot determine the engine shape based on horsepower and gears. The null hypothesis significance test failed to reject the null hypothesis since p was not less than the alpha for the intercept and gears. It's important to note that the intercept and gears both have  variables are near 0, and therefore, will not have a meaningful effect in the model.  However, looking at the hp, we can determine it will be a more significant predictor since its p value is less than 0.05, at 0.014.

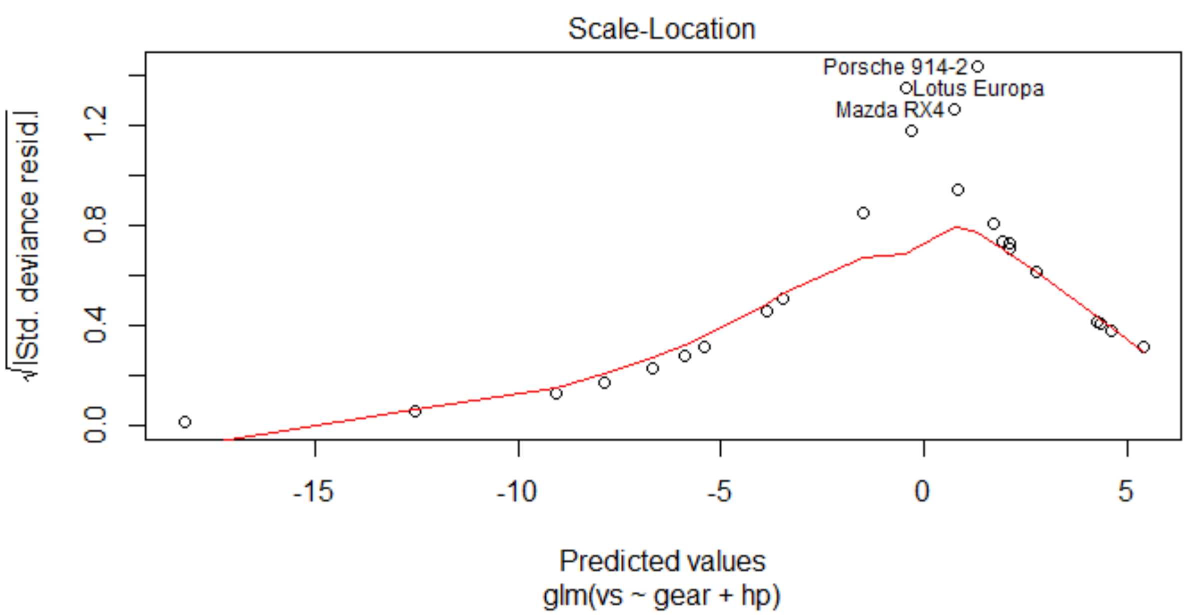
Parin Patel
IST 772
HW9

```
> #Q1
> mtcars <- data.frame(mtcars)
> summary(mtcars)
      mpg             cyl             disp             hp             drat      
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930  
       wt             qsec             vs               am        
 Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :3.325   Median :17.71   Median :0.0000   Median :0.0000  
 Mean   :3.217   Mean   :17.85   Mean   :0.4375   Mean   :0.4062  
 3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :1.0000  
      gear            carb      
 Min.   :3.000   Min.   :1.000  
 1st Qu.:3.000   1st Qu.:2.000  
 Median :4.000   Median :2.000  
 Mean   :3.688   Mean   :2.812  
 3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.000   Max.   :8.000  
```

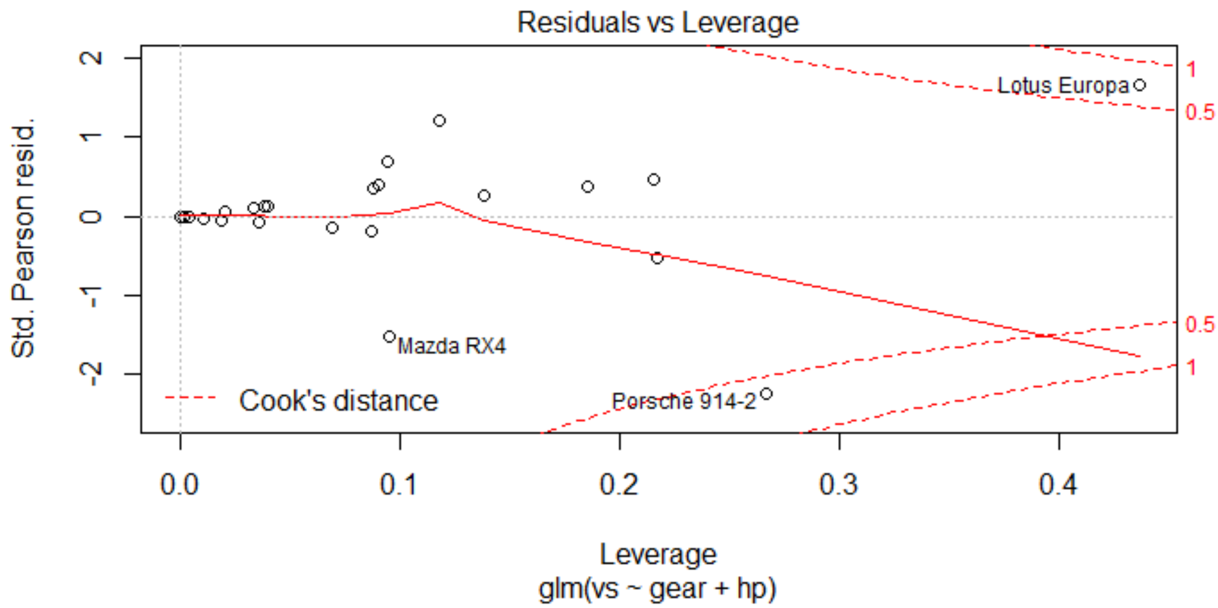


Parin Patel
IST 772
HW9



Parin Patel
IST 772
HW9




```
> summary(glmOut)

Call:
glm(formula = vs ~ gear + hp, family = binomial(link = "logit"),
    data = mtcars)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.76095  -0.20263  -0.00889   0.38030   1.37305

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 13.43752    7.18161   1.871   0.0613 .
gear        -0.96825    1.12809  -0.858   0.3907
hp          -0.08005    0.03261  -2.455   0.0141 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.860  on 31  degrees of freedom
Residual deviance: 16.013  on 29  degrees of freedom
AIC: 22.013

Number of Fisher Scoring iterations: 7
```

```
> exp(coef(glmOut))
 (Intercept)         gear           hp
6.852403e+05 3.797461e-01 9.230734e-01
>
```

**5 As noted in the chapter, the BaylorEdPsych add-in package contains a procedure for generating pseudo-R-squared values from the output of the glm() procedure. Use the results of Exercise 1 to generate, report, and interpret a Nagelkerke pseudo-R-squared value.**

```
> #Question 5
> library("BaylorEdPsych")
Warning message:
package ‘BaylorEdPsych’ was built under R version 3.5.2
> PseudoR2(glmOut)
        McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
       0.6349042        0.4525061        0.5811397        0.7789526
McKelvey.Zavoina           Effron            Count        Adj.Count
       0.8972195        0.6445327        0.8125000        0.5714286
             AIC    Corrected.AIC
      22.0131402       22.8702830
>
```

The results of the Nagelkerke pseudo-R-squared values how an approximate value of the r-squared for linear models in categorical models. Tyically, the Nagelkerke pseudo-R-squared value is higher than the R-squared value. In our model, we got a Nagelkerke of 0.78. This shows a strong model, but based on our results from question 1, this result is either because horsepower is a strong predictor for vs, or because of the smaller sample size of the dataset.

**6. Continue the analysis of the Chile data set described in this chapter. The data set is in the “car” package, so you will have to install.packages() and library() that package first, and then use the data(Chile) command to get access to the data set. Pay close attention to the transformations needed to isolate cases with the Yes and No votes as shown in this chapter. Add a new predictor, statusquo, into the model and remove the income variable. Your new model specification should be vote ~ age + statusquo. The statusquo variable is a rating that each respondent gave indicating whether they preferred change or maintaining the status quo. Conduct general linear model and Bayesian analysis on this model and report and interpret all relevant results. Compare the AIC from this model to the AIC from the model that was developed in the chapter (using income and age as predictors).**

The results of both our logistic model and Bayesian model showed that status quo is actually the strongest predictor within our model for voting. First, when we ran our logistic regression model, we got a p-value of 2e-16, which was very low. Since this p-value was less than our alpha, it was

Parin Patel
IST 772
HW9

highly significant. The results of our next Bayesian model's HDI of statusquo further supported the findings of our logistic model. The Bayesian result showed that the lowest quantile of the statuquo is 2.91 while the upper quantile is at 3.49. Finally, since the statusquo overalps with 0, we can infer that statusquo is in fact, a strong predictor of in determining how an individual will vote.

```
> #Question 6
> library("car")
> data(Chile)
> Chile <- data.frame(Chile)
> ChileNo <- Chile[Chile$vote =='N',]
> ChileYs <- Chile[Chile$vote =='Y',]
> ChileYN <- rbind(ChileYs, ChileNo)
> ChileYN <- ChileYN[complete.cases(ChileYN),]
>
> #General linear model:
> ChileYN$vote <- factor(ChileYN$vote, levels = c('N','Y'))
> Chile_linear <- glm(vote ~ age + statusquo, family = binomial(), data = ChileYN)
> summary(Chile_linear)

Call:
glm(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2095  -0.2830  -0.1840   0.1889   2.8789

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.193759   0.270708  -0.716   0.4741
age          0.011322   0.006826   1.659   0.0972 .
statusquo    3.174487   0.143921  22.057   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2360.29  on 1702  degrees of freedom
Residual deviance:  734.52  on 1700  degrees of freedom
AIC: 740.52

Number of Fisher Scoring iterations: 6
```

Parin Patel
IST 772
HW9

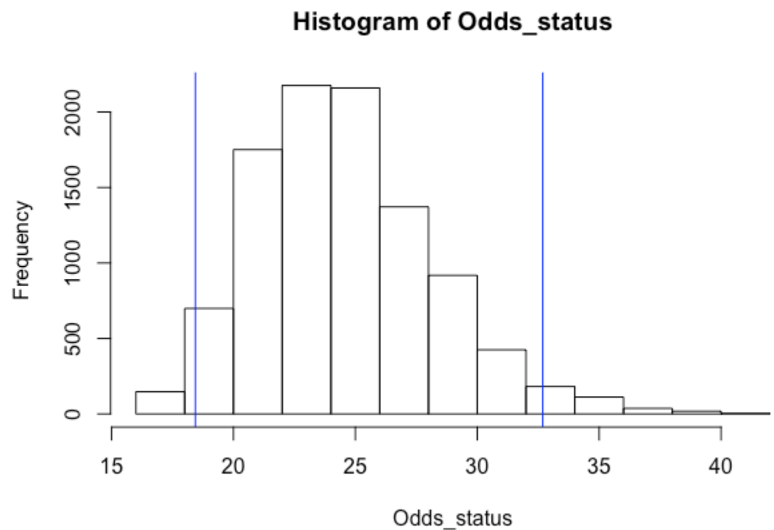
```
> ##Bayes analysis
> library("MCMCpack")
>
> ChileYN$vote <- as.numeric(ChileYN$vote) - 1
> Chile_Bayes <- MCMClogit(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN)
> summary(Chile_Bayes)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                Mean       SD  Naive SE Time-series SE
(Intercept) -0.18272 0.272640 2.726e-03       0.008938
age          0.01123 0.006817 6.817e-05       0.000223
statusquo    3.19061 0.145853 1.459e-03       0.004993

2. Quantiles for each variable:

                 2.5%       25%      50%        75%   97.5%
(Intercept) -0.742761 -0.365241 -0.17552 -0.0003872 0.34439
age         -0.002005  0.006733  0.01121  0.0157683 0.02499
statusquo    2.914442  3.087259  3.18546  3.2847388 3.48698
```

**7. Bonus R code question: Develop your own custom function that will take the posterior distribution of a coefficient from the output object from an MCMClogit() analysis and automatically create a histogram of the posterior distributions of the coefficient in terms of regular odds (instead of log-odds). Make sure to mark vertical lines on the histogram indicating the boundaries of the 95% HDI.**

Histogram below.

```
> #Question 7
>
> LogOddsPost_status <- as.matrix(Chile_Bayes[,"statusquo"])
> Odds_status <- apply(LogOddsPost_status,1,exp)
> hist(Odds_status)
> abline(v=quantile(Odds_status,c(.025)),col='blue')
> abline(v=quantile(Odds_status,c(.975)),col='blue')
>
```

Parin Patel
IST 772
HW9




## Appendix 1: Source Code

```
#Question 1

mtcars <- data.frame(mtcars)

summary(mtcars)


glmOut <- glm(formula = vs ~ gear + hp, family = binomial(link="logit"), data =  mtcars)

plot(glmOut)



summary(glmOut)


exp(coef(glmOut))



#Question 5

library("BaylorEdPsych")
```

Parin Patel
IST 772
HW9
PseudoR2(glmOut)

```
#Question 6

library("car")

data(Chile)

Chile <- data.frame(Chile)

ChileNo <- Chile[Chile$vote =='N',]

ChileYs <- Chile[Chile$vote =='Y',]

ChileYN <- rbind(ChileYs, ChileNo)

ChileYN <- ChileYN[complete.cases(ChileYN),]


#General linear model:

ChileYN$vote <- factor(ChileYN$vote, levels = c('N','Y'))

Chile_linear <- glm(vote ~ age + statusquo, family = binomial(), data = ChileYN)

summary(Chile_linear)


##Bayes analysis

library("MCMCpack")


ChileYN$vote <- as.numeric(ChileYN$vote) - 1

Chile_Bayes <- MCMClogit(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN)

summary(Chile_Bayes)




#Question 7


LogOddsPost_status <- as.matrix(Chile_Bayes[,"statusquo"])
```

Parin Patel
IST 772
HW9

```
Odds_status <- apply(LogOddsPost_status,1,exp)

hist(Odds_status)

abline(v=quantile(Odds_status,c(.025)),col='blue')

abline(v=quantile(Odds_status,c(.975)),col='blue')
```

## Appendix 2: Console Output:

```
> #Question 1

> mtcars <- data.frame(mtcars)

> summary(mtcars)

      mpg             cyl             disp

 Min.   :10.40   Min.   :4.000   Min.   : 71.1

 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8

 Median :19.20   Median :6.000   Median :196.3

 Mean   :20.09   Mean   :6.188   Mean   :230.7

 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0

 Max.   :33.90   Max.   :8.000   Max.   :472.0

       hp             drat             wt

 Min.   : 52.0   Min.   :2.760   Min.   :1.513

 1st Qu.: 96.5   1st Qu.:3.080   1st Qu.:2.581

 Median :123.0   Median :3.695   Median :3.325

 Mean   :146.7   Mean   :3.597   Mean   :3.217

 3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.610

 Max.   :335.0   Max.   :4.930   Max.   :5.424

      qsec             vs               am

 Min.   :14.50   Min.   :0.0000   Min.   :0.0000
```

Parin Patel
IST 772
HW9

```
 1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000
 Median :17.71   Median :0.0000   Median :0.0000
 Mean   :17.85   Mean   :0.4375   Mean   :0.4062
 3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :22.90   Max.   :1.0000   Max.   :1.0000
      gear            carb
 Min.   :3.000   Min.   :1.000
 1st Qu.:3.000   1st Qu.:2.000
 Median :4.000   Median :2.000
 Mean   :3.688   Mean   :2.812
 3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :5.000   Max.   :8.000
>
> glmOut <- glm(formula = vs ~ gear + hp, family = binomial(link="logit"), data =  mtcars)
> plot(glmOut)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot: summary(glmOut)
Hit <Return> to see next plot:
> exp(coef(glmOut))
 (Intercept)         gear           hp
6.852403e+05 3.797461e-01 9.230734e-01
>
>
> #Question 5
> library("BaylorEdPsych")
> PseudoR2(glmOut)
       McFadden     Adj.McFadden       Cox.Snell
      0.6349042       0.4525061       0.5811397
```

Parin Patel
IST 772
HW9

```
   Nagelkerke McKelvey.Zavoina          Effron
    0.7789526       0.8972195       0.6445327
        Count       Adj.Count           AIC
    0.8125000       0.5714286      22.0131402
 Corrected.AIC
   22.8702830
>
>
> #Question 6
> library("car")
> data(Chile)
> Chile <- data.frame(Chile)
> ChileNo <- Chile[Chile$vote =='N',]
> ChileYs <- Chile[Chile$vote =='Y',]
> ChileYN <- rbind(ChileYs, ChileNo)
> ChileYN <- ChileYN[complete.cases(ChileYN),]
>
> #General linear model:
> ChileYN$vote <- factor(ChileYN$vote, levels = c('N','Y'))
> Chile_linear <- glm(vote ~ age + statusquo, family = binomial(), data = ChileYN)
> summary(Chile_linear)


Call:
glm(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN)


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2095  -0.2830  -0.1840   0.1889   2.8789
```

Parin Patel

IST 772

HW9

Coefficients:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.193759   0.270708  -0.716   0.4741
age          0.011322   0.006826   1.659   0.0972 .
statusquo    3.174487   0.143921  22.057   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2360.29  on 1702  degrees of freedom
Residual deviance:  734.52  on 1700  degrees of freedom
AIC: 740.52

Number of Fisher Scoring iterations: 6

>
> ##Bayes analysis
> library("MCMCpack")
>
> ChileYN$vote <- as.numeric(ChileYN$vote) - 1
> Chile_Bayes <- MCMClogit(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN)
> summary(Chile_Bayes)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000
```

Parin Patel
IST 772
HW9

1. Empirical mean and standard deviation for each variable,

   plus standard error of the mean:

```
              Mean       SD  Naive SE Time-series SE
(Intercept) -0.18272 0.272640 2.726e-03       0.008938
age          0.01123 0.006817 6.817e-05       0.000223
statusquo    3.19061 0.145853 1.459e-03       0.004993
```

2. Quantiles for each variable:

```
                2.5%       25%      50%        75%   97.5%
(Intercept) -0.742761 -0.365241 -0.17552 -0.0003872 0.34439
age         -0.002005  0.006733  0.01121  0.0157683 0.02499
statusquo    2.914442  3.087259  3.18546  3.2847388 3.48698
```

```
>
>
>
> #Question 7
>
> LogOddsPost_status <- as.matrix(Chile_Bayes[,"statusquo"])
> Odds_status <- apply(LogOddsPost_status,1,exp)
> hist(Odds_status)
> abline(v=quantile(Odds_status,c(.025)),col='blue')
> abline(v=quantile(Odds_status,c(.975)),col='blue')
```