IST772 Problem Set 10

Abhijith Anil Vamadev

The homework for week 11 is based on exercises 1, 5, 6, and 7 on page 234 but with changes as noted in this notebook (i.e., follow the problems as given in this document and not the textbook).

Attribution statement: (choose only one) 1. I did this homework by myself, with help from the book and the professor

Chapter 10, Exercise 1

The carData package in R contains a small data set called Prestige that contains n = 102 observations of different occupations in Canada in 1971. Load the carData package and use "?Prestige" to display help about the data set. The data in this data set are mostly metric with one factor variable for job type, but the women variable (% women in the occupation) is highly skewed. We can dichotomize it to create binary variable as follows:

```
library(carData)
data <- Prestige
data$male.dominated <- as.integer(data$women < 10)
#if less than 10, then it is a 1, if greater 0.
?Prestige
```

```
## starting httpd help server ... done
```

```
Prestige$income <- Prestige$income/1000 #reducting the variable by 1000
```

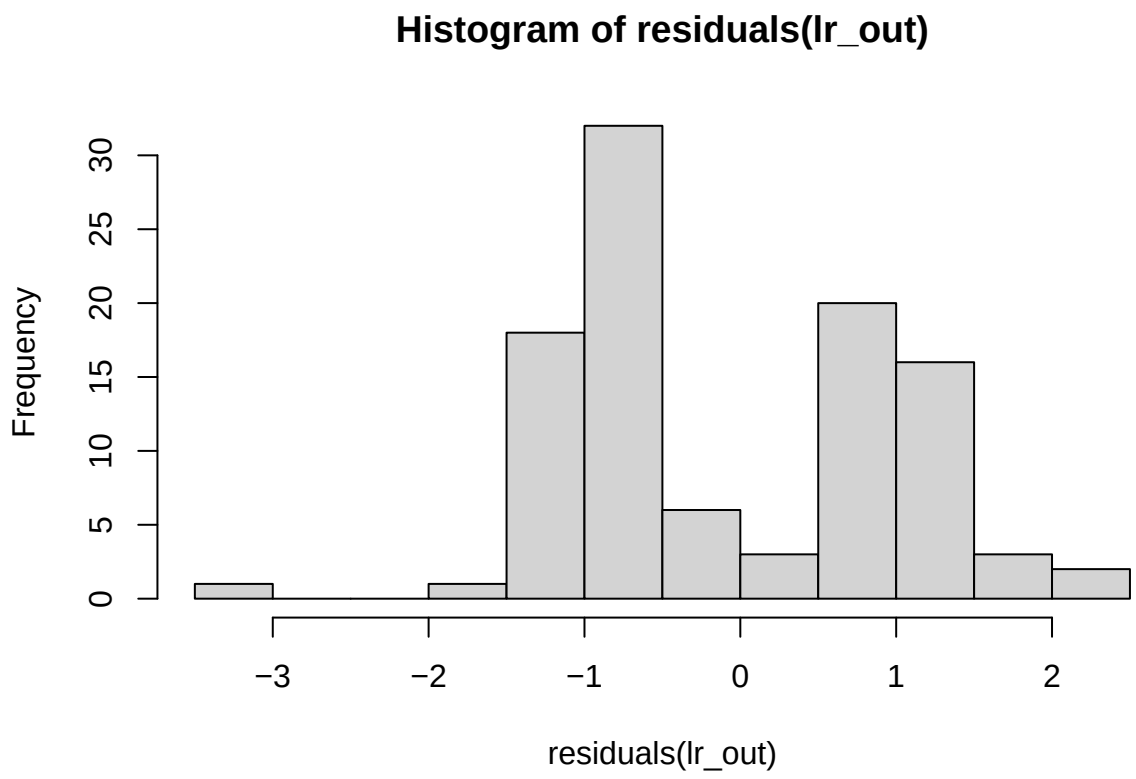
Use logistic regression to predict which occupations are male dominated (i.e., using the male.dominated variable), using two metric variables in the data set, income and education (1 pt). The results will be more interpretable if you transform income to income in $1000. Run all necessary diagnostics. (1 pt) Interpret the resulting null hypothesis significance tests. (1 pt)

```
lr_out <- glm(male.dominated~income +education, data = data, family = binomial())
#model
summary(lr_out)#summary
```

```
##
## Call:
## glm(formula = male.dominated ~ income + education, family = binomial(),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1328  -0.8656  -0.5040   0.9637   2.3882
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.6849568  0.9244550   1.823 0.068357 .
## income       0.0004041  0.0001051   3.845 0.000120 ***
## education   -0.4393786  0.1187607  -3.700 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 139.47  on 101  degrees of freedom
## Residual deviance: 113.92  on  99  degrees of freedom
## AIC: 119.92
##
## Number of Fisher Scoring iterations: 5
```

```
hist(residuals(lr_out))#histogram of residuals
```




```
mean(residuals(lr_out)) #mean of the residuals
```

```
## [1] -0.04595578
```

```
exp(coef(lr_out))#odds
```

```
## (Intercept)      income   education
##   5.3922179   1.0004042   0.6444368
```

- The mean of the residuals is near 0. The median residuals is negative which indicates, that it is slightly positivly skewed. Which is confirmed by the histogram of hte residuals
- The Y-intercept is not significantly different from 0, the estaimte of the intercept is 1.684, with a z-value of 1.823 and an associated p-value of 0.068 which is greater than the, associated p-value of 0.05 so we fail to reject the null hypothesis.
- We observe that the coefficient for income is significantly different from 0, 0.000404 as this is supported by the "Wald" z-test 3.845 and
  the associated p-value (0.000120). So we reject the null hypothesis that the coefficient on income is 0 in the population.
- We observe that the coefficient for education is significantly different from 0, -0.4393, as this is supported by the "Wald" z-test -3.7 and
  the associated p-value (0.000216). So we reject the null hypothesis that the coefficient on income is 0 in the population.
- The null model shows 101 degrees of freedom because we started with 102 cases and we lose 1 degree of freedom simply for calculating the proportion of truth to lies in our Y variable
- One unit change in income gives us a 1:5.39, change in odds of male-domination.
- One unit change in education gives us a 0.65:5.39 change in odds of male-domination. # Chapter 10, Exercise 5

As noted in class, the performance package contains a procedure for generating various measures of model performance, including pseudo-R-squared values from the output of the glm() procedure. Use the results of Exercise 1 to generate, report, and interpret a pseudo-R-squared value. You might also examine the confusion matrix. (1 pt)

```
library(performance)
library(BaylorEdPsych)
exp(confint(lr_out)) # Look at confidence intervals
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 %     97.5 %
## (Intercept) 0.9092588 34.9516781
## income      1.0002172  1.0006310
## education   0.5024012  0.8033069
```

```
anova(lr_out, test="Chisq") #25.55
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: male.dominated
##
## Terms added sequentially (first to last)
##
##
```

```
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        101     139.47
## income     1   8.8587       100     130.62  0.002917 **
## education  1  16.6953        99     113.92 4.389e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PseudoR2(lr_out)#pseudo -r
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##        0.1832168        0.1258585        0.2216117        0.2973734
## McKelvey.Zavoina           Effron            Count        Adj.Count
##        0.3763733        0.2872997        0.7549020        0.4318182
##              AIC    Corrected.AIC
##      119.9203345      120.1652325
```

```
table(round(predict(lr_out,type="response")),data$male.dominated)#confusin matrix
```

```
##
##      0  1
##   0 47 14
##   1 11 30
```

```
(47+30)/(47+30+11+14)#accuracy
```

```
## [1] 0.754902
```

- Looking at the condidence intervals for the intercept it is 0.90 and 34.951, which contradicts the previous test, which showed that the y-intercept wasn't significant.
- The income has a confidence interval that runs from a low of 1:1 and high of 1:1.With a 95% confidence interval that if you constructed a very large number of similar experiments based on new samples, that 95% of the confidence intervals you would calculate would contain the population value.
- The confidence interval for education is a low of 0.5:1 to a high of 0.803:1.
- The anova, is an omnibus test for this analysis. You can see that the probability of observing the income deviation of 8.85 and education 16.69 - 25.55 (Null deviation - Residual deviation), on 1 DF, is extremly low for both income and variable that has an associcated p-value of 0.0029 and 4.38e-05 respectively. Since both the values are less than alpha, we can reject the null hypothesis that introducing income and education into the model caused zero reduction of model error.
- We can consider this rejection of the null hypothesis as evidence that the two-predictor model is preferred over the null model.
- For any of the given measures, you can loosely interpret it as the proportion of variance in the outcome variable (male_dominated) accounted for by the predictor variables (income and education). In that light, these are all very small effect size values. Given our finding that both income and education as significant, these results suggest that the variables has only a rather small role in accounting for the male_dominated variable.
- There was 47 cases where the observed male_dominated was 0. Likewise there was 30 cases, where the observed value was 1. The overall accuracy of the model is (47+30)/(47+30+11+14) is 0.75. 75% accuracy is a good iscore indicating a good model. This shows that a predictive model with income and experience, is a good in predicting model_domination, although there may be better variables to predict male_domination.

Chapter 10, Exercise 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 and "? Chile" to see the documentation.

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 (1 pt + 1 pt) and Bayesian analysis on this model (1 pt) and report and interpret all relevant results (1 pt). Compare the AIC from this model to the AIC from the model that was developed in the chapter (using income and age as predictors).

```
library(car)
library(MCMCpack)
```

```
## Loading required package: coda
```

```
## Loading required package: MASS
```

```
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
```

```
## ## Copyright (C) 2003-2021 Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park
```

```
## ##
## ## Support provided by the U.S. National Science Foundation
```

```
## ## (Grants SES-0350646 and SES-0350613)
## ##
```

```
new_data <- Chile
y <-new_data[new_data$vote == "Y" | new_data$vote == "N",] #only Y or No
ChileYN <- y[complete.cases(y),] # Get rid of missing data
ChileYN$vote <- factor(ChileYN$vote,levels=c("N","Y"))  #converint as factor
chOut <- glm(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN)
#first model
chOut1 <- glm(formula = vote ~ age + income, family = binomial(), data = ChileYN)
 #AIC model - 2332 with age and income
ChileYN$vote <- as.numeric(ChileYN$vote) - 1 #convering to numeric
bayesLogitOut <- MCMClogit(formula = vote ~ age + statusquo, data = ChileYN)
#bayes
```

```
summary(chOut1) #AIC is 2332
```

```
##
## Call:
## glm(formula = vote ~ age + income, family = binomial(), data = ChileYN)
##
## Deviance Residuals:
```

```
##    Min      1Q  Median      3Q     Max
## -1.435  -1.126  -1.004   1.191   1.373
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.581e-01  1.418e-01  -5.346 9.01e-08 ***
## age          1.924e-02  3.324e-03   5.788 7.11e-09 ***
## income      -2.846e-07  1.142e-06  -0.249    0.803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2360.3  on 1702  degrees of freedom
## Residual deviance: 2326.0  on 1700  degrees of freedom
## AIC: 2332
##
## Number of Fisher Scoring iterations: 4
```

```
summary(chOut)#AAC is 740.52
```

```
##
## 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
```

```
exp(coef(chOut)) #odds
```

```
## (Intercept)         age   statusquo
##   0.8238564   1.0113863  23.9145451
```

```
exp(confint(chOut)) #confidence interval
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %    97.5 %
## (Intercept)  0.4847068  1.402937
## age          0.9979335  1.025033
## statusquo   18.2483505 32.107663
```

```
anova(chOut, test = "Chisq") #anova
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: vote
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       1702    2360.29
## age        1     34.2      1701    2326.09 4.964e-09 ***
## statusquo  1   1591.6      1700     734.52 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

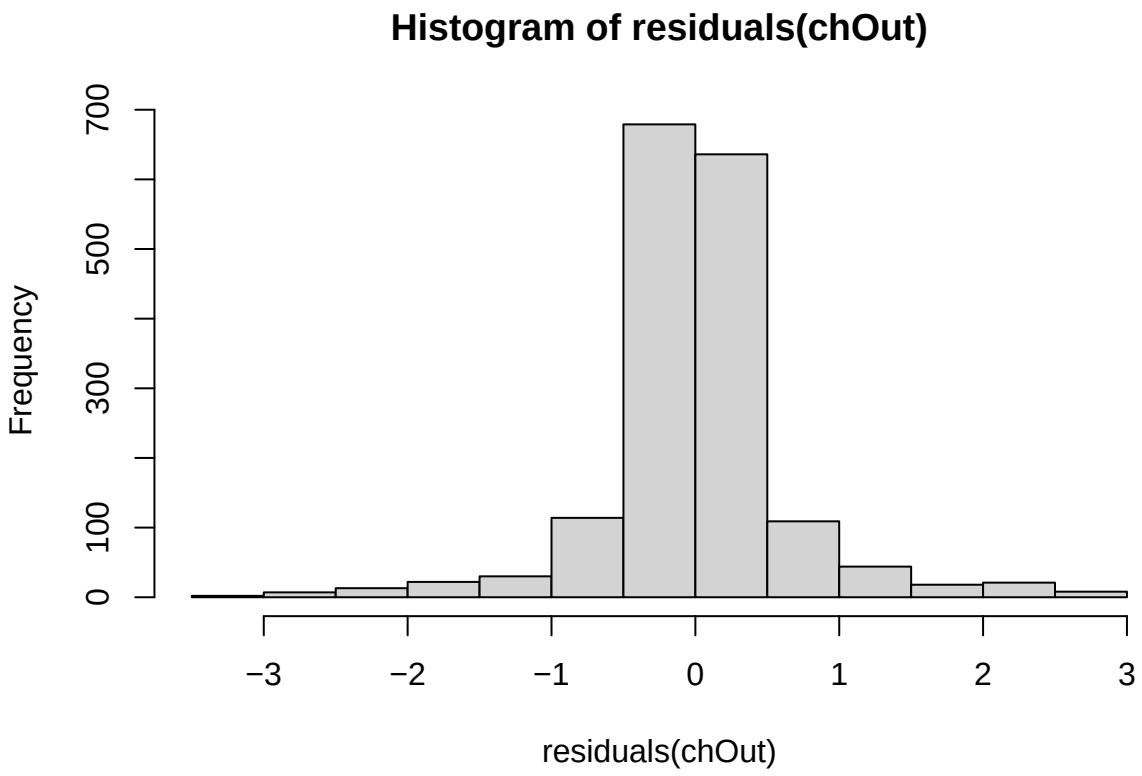
```
PseudoR2(chOut) #pseudo r-square
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##        0.6888013        0.6854119        0.6150544        0.8201631
## McKelvey.Zavoina           Effron            Count        Adj.Count
##        0.7855565        0.7553412        0.9230769        0.8433014
##              AIC    Corrected.AIC
##      740.5206862      740.5348122
```

```
table(round(predict(chOut, type="response")), ChileYN$vote)#confusion matrix
```

```
##
##       0   1
##   0 810  74
##   1  57 762
```

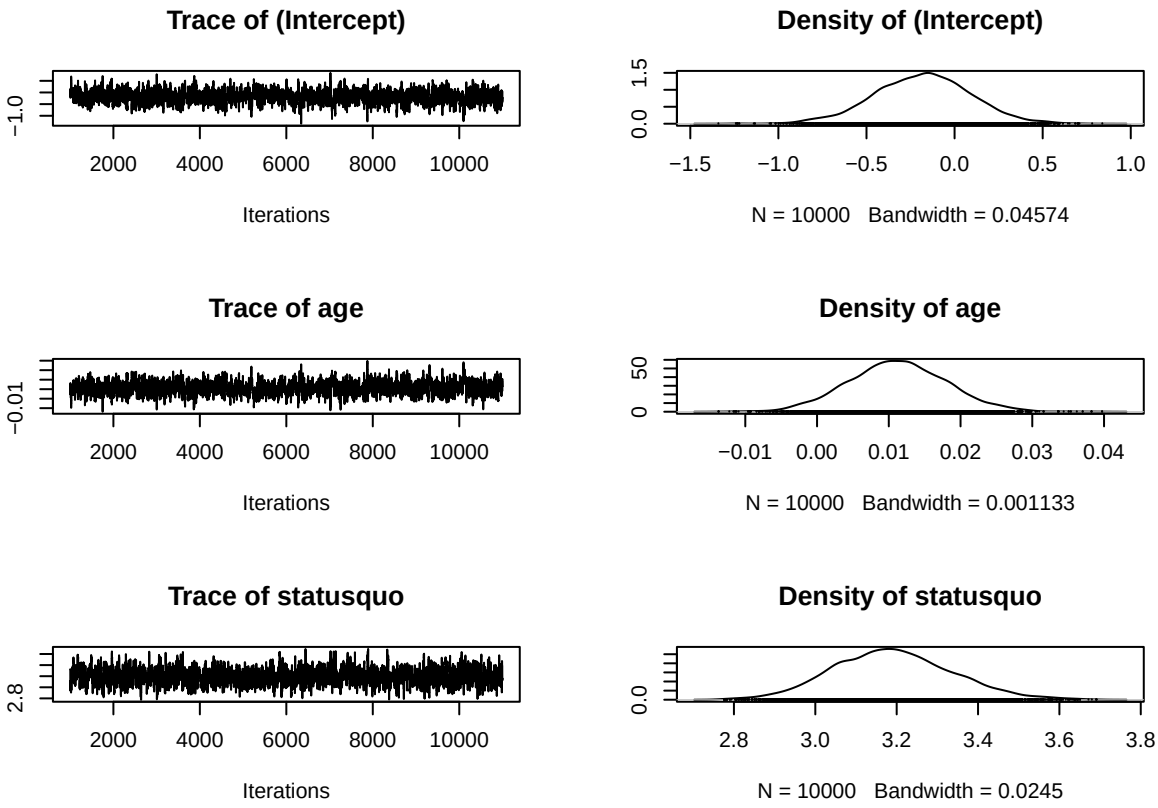
```
hist(residuals(chOut)) #histogram of residuals
```




```
summary(bayesLogitOut) #summary of the 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
```

```
plot(bayesLogitOut) #plot o fhte bayes
```




```
(810+762)/(810+74+57+762) #accuracy of the model
```

```
## [1] 0.9230769
```

- The residuals look normally distributed
- The Y-intercept is not significantly different from 0, the estaimte of the intercept is -0.193, with a z-value of 0.716 and an associated p-value of 0.4741 which is greater than the, associated p-value of 0.05 so we fail to reject the null hypothesis.
- The coefficient on the age predictor is 0.01132 and isn't statistically significant, based on the Wald's z-test value of 1.659 and the associated p-value of 0.09. Because the scientific notation p = 0.0972 means that p > 0.06, we fail to reject the null hypothesis,that the log-odds of income is equal to 0 in the population which makes age statistically insignifcant predictor in this model.
- The coefficient on the statusquo predictor is 3.174 and is statistically significant, based on the Wald's z-test value of 22.057 and the associated p-value of 2e-16. Because the scientific notation p = 2e-16 means that p < 0.01, we reject the null hypothesis, which makes statusquo statistically significant in this model.
- After running the anova command, the age and statusquo with DF of 1, has deviance of 34.2 and 1591.6 , with p-values 4.964e-09 and 2.2e-16. The first chi-squared test compares the null model, to a model with just age to a model that has both age and statusquo as predictors. Both the p-values, are less than alpha, which makes them both statistically signifcant.
- For any of the given measures, you can loosely interpret it as the proportion of variance in the outcome variable (vote) accounted for by the predictor variables (age and statusquo). In that light, these are all pretty good effect size values. Nagelkreke has a r-squqred value of 0.82 indicating a good fit model.
- Looking at the confusion matrix, there were 810 cases where the observed vote in the data set was No and the dichotomized predicted value was 0. Those are the correct classifications of No. Likewise,

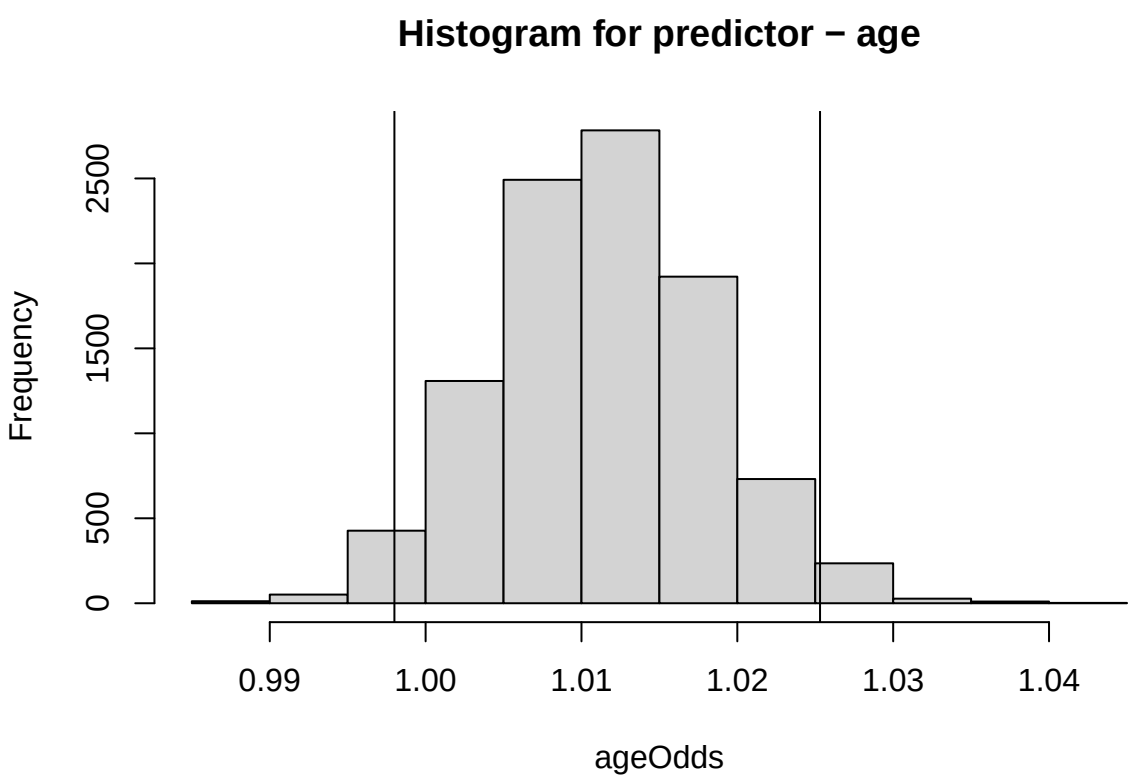
there were 762 cases where the observed vote in the data set was Yes and the dichotomized predicted value was 1. The accuracy of the model is 92.3% which indicates a really good model.

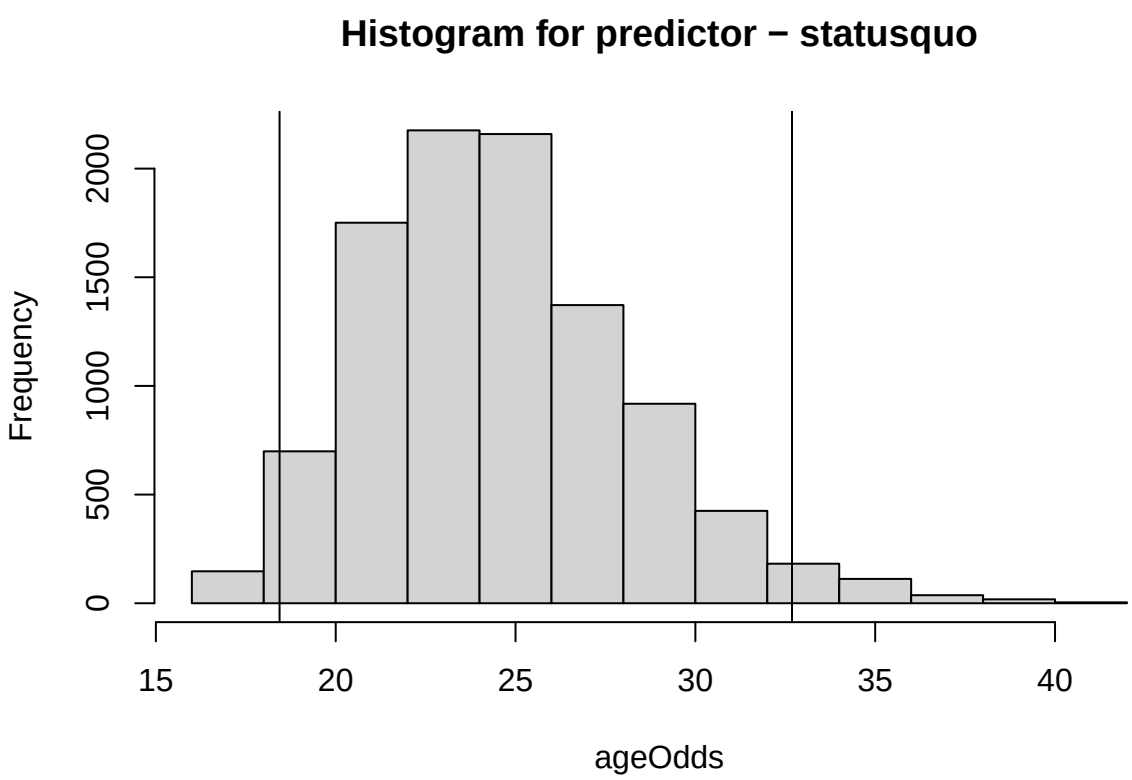
- The intercept represents odds of 0.82:1 for a Yes vote by a penniless new-born. The odds of 1.011:1 for age shows that for every additional year of age, a person is above 1.01% likely to vote Yes. For Statusquo the odds are 23.91:1 that for statusquo a person is 23.91% likely to vote yes.
- Looking at the confidence intervals, the confidence interval for intercept is 0.48:1 low to 1:40:1 high. These results jibe with the hypothesis tests: the confidence interval for age straddles 0.99:1.02, close to 1:1, confirming the nonsignificant result for that coefficient.
- The second output focuses on describing the posteriror distributions of parameters representing both the intercept and the coefficents - age and statusquo, calibrated as log-odds. The point-estimate of age is 0.011 and status quo 3.19 with y-intercept being -0.182. The second part of the output displays the 2.5% and 97.5% quantiles. The HDI for age is between -0.002 and 0.024, which dosen't include 0 and status quo is between 2.9144 and 3.48, which also dosen't include 0.
- Looking at the second set of plots, there seems to be no variation in the trace plots as, all the plots looks similar with similar variations, as there are no obvious drop or rise in the plots.
- Looking at the plots on the right, the true population value of each coefficent is supposed to be somwhere near the middle of each distribution.
- Comparing the two AIC values of the two models - looks like the two models have the different AIC values, 2332 for the first model and 740.5 for the second model, indicating that the second model is more valuable.

Chapter 10, Exercise 7

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. (1 pt) Run the function on your regression results. (1 pt)

```
new_function <- function(arg1){
  length <- length(colnames(bayesLogitOut)) #lenght of the colnames
  names <- colnames(bayesLogitOut) #names of the colnames
  for (i in 2:length ){
    ageLogOdds <- as.matrix(arg1[,i]) # Create a matrix for apply()
    ageOdds <- apply(ageLogOdds,1,exp) # apply() runs exp() for each one
    title <- sprintf("Histogram for predictor - %s", names[i]) #title
    hist(ageOdds, main = title) #histogram
    abline(v=quantile(ageOdds,c(0.025)),col="black")#marking the 2.5% percentile
    abline(v=quantile(ageOdds,c(0.975)),col="black")#marking the 97.5% percentie
    }
}
new_function(bayesLogitOut)
```


Histogram for predictor − age
Frequency
0
500
1500
2500
0.99
1.00
1.01
1.02
1.03
1.04
ageOdds


Histogram for predictor − statusquo
Frequency
0
500
1000
1500
2000
15
20
25
30
35
40
ageOdds