HW9_Videtti

```
##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.
?mtcars

## starting httpd help server ... done

Exercise1 <- glm(vs~gear+hp, data = mtcars, family = binomial)
summary(Exercise1)

##
## Call:
## glm(formula = vs ~ gear + hp, family = binomial, 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

#For the NHSTs in these results, we see that the intercept is NOT
statistically significant from 0, nor is the coefficient for the gear
variable. However, the coefficient for the hp variable is indeed
statistically significant from 0.
```

```r
##5. As noted in the chapter, the install 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.
library(BaylorEdPsych)
PseudoR2(Exercise1)["Nagelkerke"]

## Nagelkerke
##  0.7789526

#The Nagelkerke Pseudo R-squared is 0.7789526. This can loosely be interpreted by saying that about 77.9% of the variance in vs is caused by gear(number of forward gears) and hp(horsepower).



##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).
library(car)

## Loading required package: carData

#GENERAL LINEAR MODEL
data(Chile)
ChileY <- Chile[Chile$vote == 'Y',]
ChileN <- Chile[Chile$vote == 'N',]
ChileYN <-rbind(ChileY,ChileN)
ChileYN <- ChileYN[complete.cases(ChileYN),]
ChileYN$vote <- factor(ChileYN$vote,levels=c('N','Y'))
summary(chOut <- glm(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN))

##
## 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))
```

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

```
exp(confint(chOut))
```

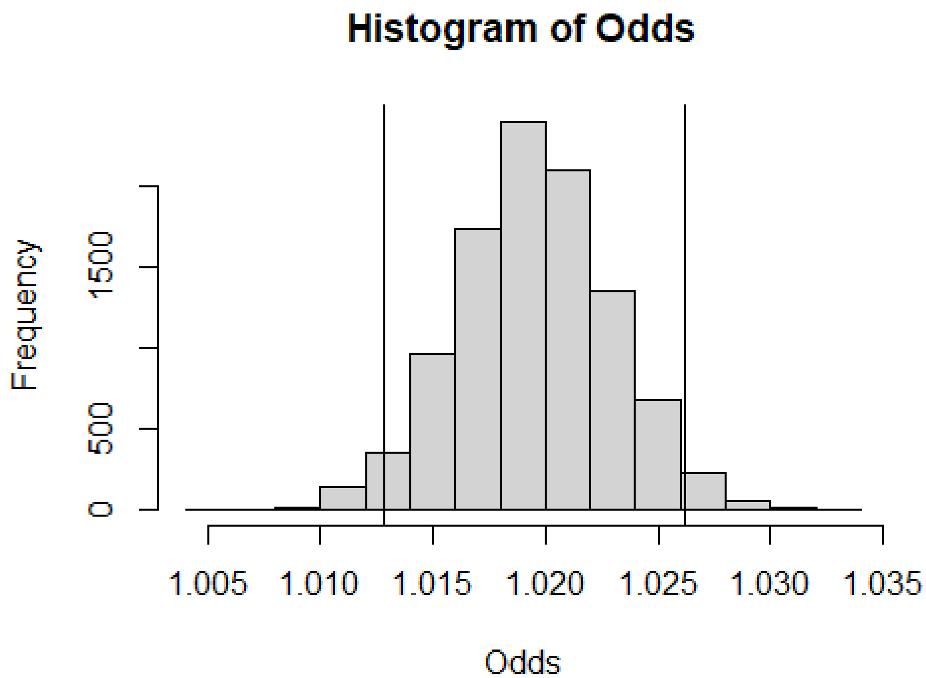
```
## Waiting for profiling to be done...
```

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

```
#Our output from the summary of the model shows us that only the statusquo
variable is significantly different from zero. That is, the log Odds of a
"Yes" vote are not statistically significantly affected at the intercept
level, or by the age variable. Also, the straight odds for the intercept,
age, and statusquo look to be approximately 0.82:1, 1.01:1, and 23.91:1,
respectively, in favor of a "Yes" vote. However, only statusquo was
statistically significant, so we can really only interpret that. The
interpretation is that for each 1 point increase in statusquo, the chances of
a yes vote increase by almost 2300% (2291.45% to be more specific). Our
confidence intervals for the straight odds are seen as the last output. Note
that the confidence interval straddles 1:1 odds for Intercept and age,
confirming our findings from earlier that they are not statistically
significant. The 95% confidence interval for statusquo, however, is
approximately 18.3:1 to 32.1:1 odds in favor of a "Yes" vote. One last thing
to note is the AIC for this model, which is 740.52, is much smaller than the
AIC for the model in the chapter, which was 2332. This means that the model
with statusquo is a much better model than the one from the chapter.
```

```r
#BAYESIAN ANALYSIS
library(MCMCpack)

## Loading required package: coda
## Loading required package: MASS
## ##
## ## Markov Chain Monte Carlo Package (MCMCpack)
## ## Copyright (C) 2003-2022 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)
## ##

ChileYN$vote <- as.numeric(ChileYN$vote) - 1
set.seed(1)
bayesLogitOut <- MCMClogit(formula = vote ~ age + statusquo, data = ChileYN)
summary(bayesLogitOut)

##
## 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

#The output gives us point estimates for each coefficient in the first
section under the "Mean" column. The second section gives us our quantiles
for each variable, which can be used to create HDI's. We do need to keep in
mind that these are in terms of log odds, so to interpret, we need to convert
to straight odds. For the HDI's we see that for straight odds, they range
over the following values.

#Intercept
c(exp(-0.742761),exp(0.34439))
```

```r
## [1] 0.4757984 1.4111289

#age
c(exp(-0.002005),exp(0.02499))

## [1] 0.997997 1.025305

#statusquo
c(exp(2.914442),exp(3.48698))

## [1] 18.43852 32.68708

#Once again, we find that the intercept and the age variable are not
significant, but the statusquo variable is, since the 95% HDI's for the
intercept and for age overlap 1:1 odds, and the status quo 95% HDI does not.
We can interpret the point estimate for statusquo after we turn it from log
odds to straight odds.

exp(3.19061)

## [1] 24.30325

#we get roughly 24.3:1 odds increase in favor of a "Yes" vote for each 1
point increase in statusquo. This is very similar to the 23.9:1 odds we found
for the same variable in the earlier part of this question where we performed
an analysis on the generalized linear model.




##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.
Exercise7 <- function(formula,data,coef_name){
  LogitOut <- MCMClogit(formula = formula, data = data)
  LogOdds <- as.matrix(LogitOut[,coef_name])
  Odds <- apply(LogOdds,1,exp)
  hist(Odds)
  abline(v=quantile(Odds,c(0.025)),col='black')
  abline(v=quantile(Odds,c(0.975)),col='black')}

#TESTING (using example from book):
Exercise7(formula = vote ~ age + income, data = ChileYN, coef_name = "age")
```


Histogram of Odds
Frequency
0
500
1500
1.005
1.010
1.015
1.020
1.025
1.030
1.035
Odds

```
#Looks good!
```