IST772 Problem Set 10

Shivani Sanjay Mahaddalkar

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 2. I did this homework with help from the book and the professor and these Internet sources: 3. I did this homework with help from but did not cut and paste any code

Chapter 10, Exercise 1

The data sets package in R contains a small data set called swiss that contains n = 47 observations of socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. Use "?swiss" to display help about the data set. All the data in this data set are metric, but one, Catholic, shows a very bimodal distribution. We can dichotomize this variable to create binary variable as follows:

```
swiss$Catholic.b <- as.integer(swiss$Catholic > 60) # 60 looks like a gap in
the histogram
summary(swiss) #There are no missing values
```

```
##    Fertility      Agriculture     Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##     Catholic       Infant.Mortality   Catholic.b
##  Min.   :  2.150   Min.   :10.80    Min.   :0.0000
##  1st Qu.:  5.195   1st Qu.:18.15    1st Qu.:0.0000
##  Median : 15.140   Median :20.00    Median :0.0000
##  Mean   : 41.144   Mean   :19.94    Mean   :0.3404
##  3rd Qu.: 93.125   3rd Qu.:21.70    3rd Qu.:1.0000
##  Max.   :100.000   Max.   :26.60    Max.   :1.0000
```

Use logistic regression to predict Catholic.b, using two metric variables in the data set, Fertility and Agriculture (1 pt). Run any necessary diagnostics. (1 pt) Interpret the resulting null hypothesis significance tests. (1 pt)

```
glmCatholic <- glm(Catholic.b~Fertility+Agriculture, data = swiss, family =
binomial())
summary(glmCatholic)
```

```
##
## Call:
## glm(formula = Catholic.b ~ Fertility + Agriculture, family = binomial(),
##     data = swiss)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.50316  -0.24820  -0.01404   0.14290   2.39654
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -34.07275   11.31906  -3.010  0.00261 **
## Fertility     0.35010    0.11768   2.975  0.00293 **
## Agriculture   0.13078    0.05314   2.461  0.01384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 60.284  on 46  degrees of freedom
## Residual deviance: 21.470  on 44  degrees of freedom
## AIC: 27.47
##
## Number of Fisher Scoring iterations: 8
```

The median being slightly negative indicates that the distribution of residuals is slightly positively skewed. Null hypothesis: The variables Fertility and Agriculture have coefficients of 0. The intercept indicates that the if the fertility and agriculture is zero, it is likely that the population is less than 60% catholic. In the analysis, the p-values for the coefficients Fertility and Agriculture are less than 0.05, therefore they are statistically significant.

```
exp(coef(glmCatholic))
```

```
##  (Intercept)    Fertility  Agriculture
## 1.593646e-15 1.419211e+00 1.139720e+00
```

The odds show that the intercept is close to zero. The 1.4219:1 odds for fertility indicate that for every unit change in fertility there is a 42.19% more likely to have more than 60% catholic population. The 1.1397:1 odds for fertility indicate that for every unit change in fertility there is a 13.97% more likely to have more than 60% catholic population.

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

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
```

```
## Response: Catholic.b
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           46     60.284
## Fertility    1   24.032        45     36.252 9.474e-07 ***
## Agriculture  1   14.781        44     21.470 0.0001207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p<0.001, we reject the null hypothesis and say that the two-predictor model is preferred over the null model.

Chapter 10, Exercise 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. You might also examine the confusion matrix. (1 pt)

```
library(performance)

## Warning: package 'performance' was built under R version 4.0.5

r2_nagelkerke(glmCatholic)

## Nagelkerke's R2
##       0.7778181
```

The Nagelkerke's R2 indicates the proportion of variance in the outcome variable(Catholic) accounted for the independent variables Fertility and Agriculture.

```
table(round(predict(glmCatholic, type="response")), swiss$Catholic.b)

##
##      0  1
##   0 29  2
##   1  2 14
```

The model shows that it predicted 4 instances wrong out of the 47 instances. # 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)
```

```
## Loading required package: carData
```

```
ChileDF <- data.frame(Chile)
summary(ChileDF)
```

```
##  region     population     sex           age        education
##  C :600   Min.   :  3750   F:1379   Min.   :18.00   P   :1107
##  M :100   1st Qu.: 25000   M:1321   1st Qu.:26.00   PS  : 462
##  N :322   Median :175000            Median :36.00   S   :1120
##  S :718   Mean   :152222            Mean   :38.55   NA's:  11
##  SA:960   3rd Qu.:250000            3rd Qu.:49.00
##           Max.   :250000            Max.   :70.00
##                                     NA's   :1
##      income         statusquo         vote
##  Min.   :  2500   Min.   :-1.80301   A   :187
##  1st Qu.:  7500   1st Qu.:-1.00223   N   :889
##  Median : 15000   Median :-0.04558   U   :588
##  Mean   : 33876   Mean   : 0.00000   Y   :868
##  3rd Qu.: 35000   3rd Qu.: 0.96857   NA's:168
##  Max.   :200000   Max.   : 2.04859
##  NA's   :98       NA's   :17
```

```
ChileY <- ChileDF[ChileDF$vote =='Y', ] #Getting the Yes cases
ChileN <- ChileDF[ChileDF$vote =='N', ] #Getting the No cases
ChileDF <- rbind(ChileY, ChileN)        #Combining both the Yes and No cases
ChileDF <- ChileDF[complete.cases(ChileDF), ] #Getting rid of the missing
data
ChileDF$vote <- factor(ChileDF$vote, levels= c("N", "Y")) #Simplifying the
factor
```

```
ChOut <- glm(formula = vote ~ age + statusquo, family = binomial(), data =
ChileDF)
summary(ChOut)
```

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

The intercept is not statistically significant. Null Hypothesis: Age and statusquo do not have any impact on whether the vote will be Yes or No. p-value of Age is greater than 0.0972 which means that the coefficient is not statistically significant. p-value of statusquo is less than 0.05 and the coefficient of statusquo is statistically significant. The log odds are not zero for statusquo.

```
exp(coef(ChOut))
```

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

The odds for age are almost 1:1, which means that change in age do not result in change of No to Yes. The odds for statusquo are 23.9145 which means that for every unit change in status quo is 2391% likely to vote Yes.

```
library(MCMCpack)
```

```
## Warning: package 'MCMCpack' was built under R version 4.0.5
```

```
## 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)
## ##
```

```
ChileDF$vote <- as.numeric(ChileDF$vote) -1
bayesLogitOut <- MCMClogit(formula = vote~age+statusquo, data = ChileDF)
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
```

This output focuses on describing the posterior distributions of parameters representing both the intercept and the coefficients on age and status quo, calibrated as log-odds. The point estimates for the intercept and the coefficients are quite similar to the output from the traditional logistic regression. The second part of the output displays quantiles for each coefficient, including the 2.5% and 97.5% quantiles. The region in between the 2.5% and the 97.5% quantiles for each coefficient is the highest density interval (HDI) for the given coefficient.

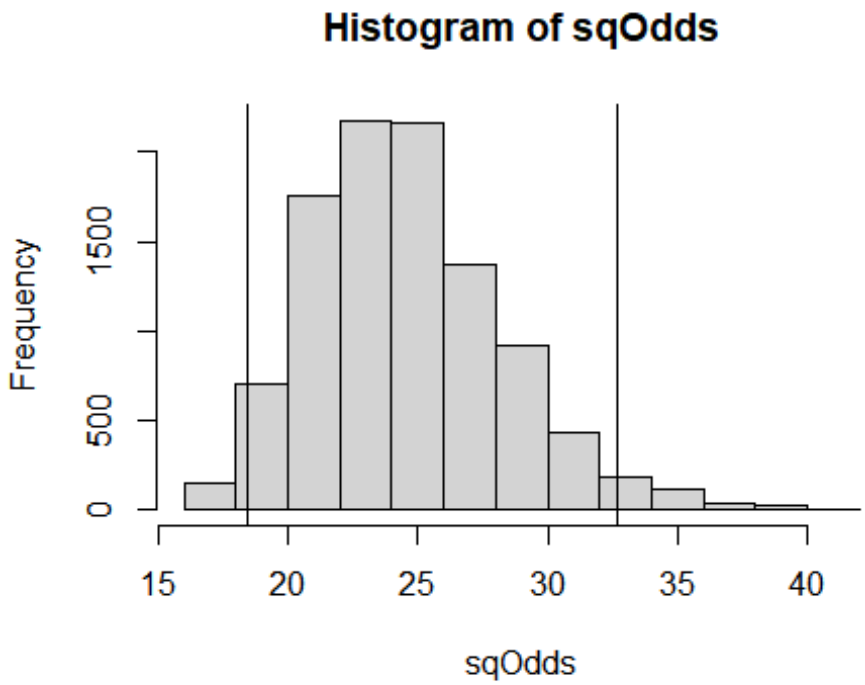
Since the quantiles for intercept and age straddle 0, we look into the status quo and convert log-odds into regular odds.

Chapter 10, Exercise 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. (1 pt) Run the function on your regression results. (1 pt)

```
sqLogOdds <- as.matrix(bayesLogitOut[,"statusquo"]) # Create a matrix for
apply()
sqOdds <- apply(sqLogOdds,1,exp) # apply() runs exp() for each one
```

```r
hist(sqOdds) # Show a histogram
abline(v=quantile(sqOdds,c(0.025)),col="black") # Left edge of 95% HDI
abline(v=quantile(sqOdds,c(0.975)),col="black") #
```



The odds 23.91 from the regular logistic regression is present in the HDI. Which strengthens our evidence.