Categorical Analysis Homework

Tim Hulak

I produced the material below with no assistance.

The answer to Ex. 7 was aided by the HW09 R script.

Exercises 1, 5, 6 and 7 on page 234 of *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R*

```
# install.packages("BaylorEdPsych_0.5.tar.gz", repos = NULL, type = "source")
library(BaylorEdPsych)
```

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.

```
data("mtcars")
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

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

```
##
## 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
```

```
exp(coef(logistic_model))
```

```
##  (Intercept)         gear           hp
## 6.852403e+05 3.797461e-01 9.230734e-01
```

Answer: The null-hypothesis is that the log-odds is 0. The p-value of the *gear* variable is **0.3907**, which is higher than the traditional **0.05** threshold. The z-test value of -0.858 and the associated p value of **0.3907** means that it is not significant. The p-value of the *hp* variable is **0.0141**, which is lower than the traditional **0.05** thresholds. The z-test value of **-2.455** and the associated p value of **0.0141** means that it is significant. As for the intercept, the z-test value of **1.871** and the associated p value of **0.0613** means that we fail to reject the null hypothesis as the p value is higher than the traditional **0.05** threshold.

---

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.

```
PseudoR2(logistic_model)
```

```
##         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
```

Answer: The Nagelkerke R-squared value of **0.7790** can be interpreted as the amount of variance in the dependent variable which depends on the independent variables, hp and gear.

---

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
```

```
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)
## ##
```

```
data("Chile")
head(Chile)
```

```
##   region population sex age education income statusquo vote
## 1      N     175000   M  65         P  35000   1.00820    Y
## 2      N     175000   M  29        PS   7500  -1.29617    N
## 3      N     175000   F  38         P  15000   1.23072    Y
## 4      N     175000   F  49         P  35000  -1.03163    N
## 5      N     175000   F  23         S  35000  -1.10496    N
## 6      N     175000   F  28         P   7500  -1.04685    N
```

```
# remove income
data <-  subset(Chile, select = -c(income) )

YES <- data[data$vote=='Y',]
NO <- data[data$vote=='N',]

dataYN=rbind(NO, YES)
dataYN=dataYN[complete.cases(dataYN),]
dataYN$vote=factor(dataYN$vote, levels=c("N", "Y"))

dataGLM=glm(vote ~ age + statusquo, family=binomial(), data=dataYN)
summary(dataGLM)
```

```
##
## Call:
## glm(formula = vote ~ age + statusquo, family = binomial(), data = dataYN)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2125  -0.2795  -0.1813   0.1876   2.8883
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.196940   0.265723  -0.741    0.459
## age          0.011167   0.006729   1.659    0.097 .
## statusquo    3.191015   0.143314  22.266   <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: 2421.50  on 1746  degrees of freedom
## Residual deviance:  749.28  on 1744  degrees of freedom
## AIC: 755.28
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(dataGLM))
```

```
## (Intercept)         age   statusquo
##   0.8212403   1.0112292  24.3131012
```

```
dataYN$vote=as.numeric(dataYN$vote)-1 # adjust outcome variable
dataBayes=MCMClogit(formula=vote~ age + statusquo, data=dataYN)
summary(dataBayes)
```

```
##
## 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) -0.18783 0.270710 2.707e-03      0.0089037
## age          0.01112 0.006788 6.788e-05      0.0002227
## statusquo    3.20648 0.142919 1.429e-03      0.0047798
##
## 2. Quantiles for each variable:
##
##                  2.5%       25%      50%       75%   97.5%
## (Intercept) -0.732332 -0.369748 -0.19219 -0.001519 0.33434
## age         -0.001716  0.006493  0.01125  0.015782 0.02454
## statusquo    2.948473  3.105790  3.20149  3.296716 3.49959
```

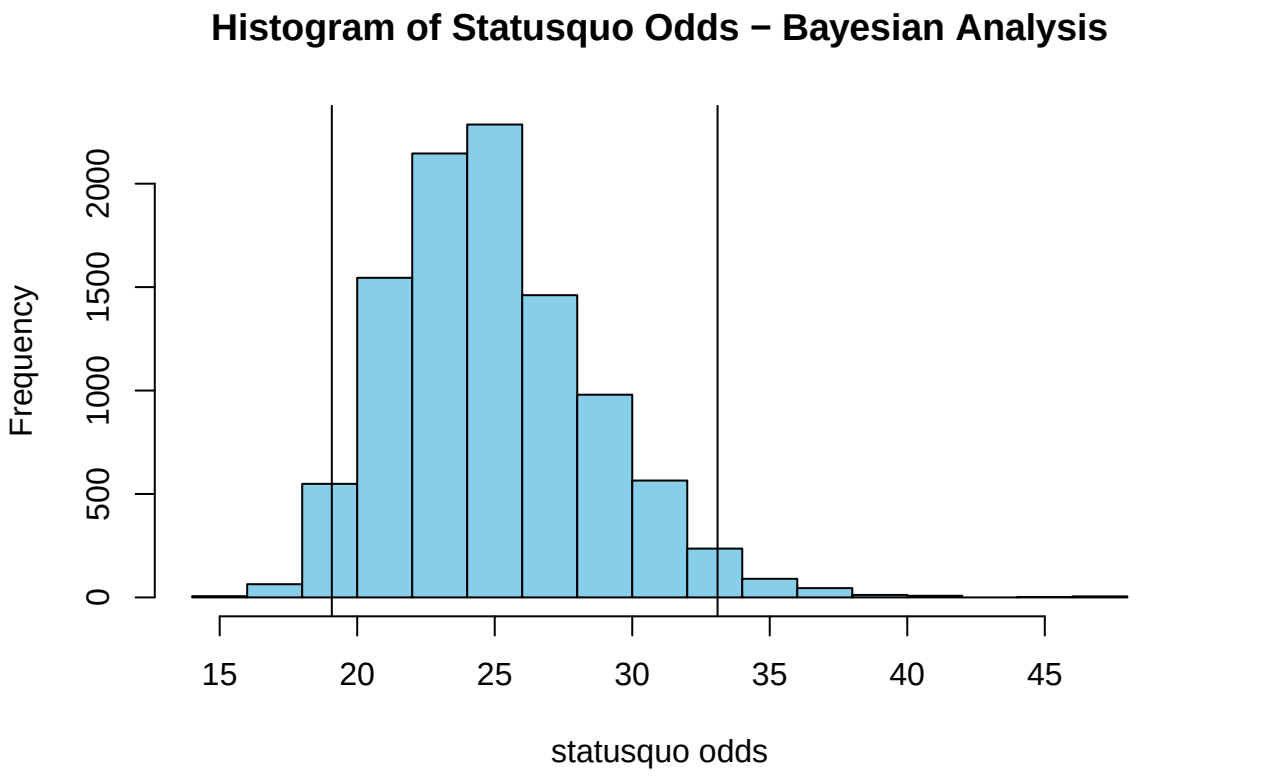
Answer: The *intercept* p-value was **0.459** and *age* p-value was **0.097**, meaning that they were not significant We would fail to reject the null hypothesis that *intercept* is 0 and *age* is 0. The *statusquo* variable has a very small p-value of **0.00000000000000022**, which is well below the traditional alpha of **0.05**. This means that we reject the null hypothesis *statusquo* predicting vote outcome is 0.

In the Bayesian results, The HDI for both *age* and the *intercept* cross over zero. This supports failing to reject the null hypothesis. The HDI for *statusquo* does not cross over 0 (**2.948473** to **3.49959**). This supports the GLM model of rejecting the null.

---

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.

```
OddsHistogram <- function(mcmc_out, seq){
  logOdds <- as.matrix(mcmc_out[,3])
  odds <- apply(logOdds,1,exp)
  hist(odds, col="skyblue",
       main="Histogram of Statusquo Odds - Bayesian Analysis",
       xlab='statusquo odds')
  abline(v=quantile(odds,c(0.025)),col='black')
  abline(v=quantile(odds,c(0.975)),col='black')
}
OddsHistogram(dataBayes, 3)
```



Answer: I'll admit that I did not know exactly how to tackle this, so I plugged in the function from the homework help file. Looking at the function, I can see that it takes in 2 arguments: a Bayesian model and a sequence. From there, a matrix is constructed from the Bayesian model and the odds are calculated. Finally, a histogram plot is constructed and lines are placed on the chart to show the HDI range.