Matthew L. Pergolski
IST 772
Dr. Block
11/25/2021

Homework 9

Beginning Statement

"I produced the material below with no assistance [direct quote from IST 772 class syllabus]." Note: Homework questions from the book may have been copied/pasted into the document for both the student and viewer's convenience. Quotes may also have been used from the class textbook.

he homework for week 9 is exercises 1, 5, 6 and 7 on page 234.

Homework Question 1

Question:

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.

Answer/Student Response:

The following was generated and observed:

```r
#Data
?mtcars
mtcars
str(mtcars)
summary(mtcars)

#Model
mt.glm <- glm(vs ~ gear + hp, family = binomial(link = "logit"), data = mtcars)
mt.glm

plot(mt.glm)
summary(mt.glm)

cof <- coef(mt.glm)
cof
exp(cof)
```

```
>     summary(mt.glm)

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

From the model generated above, we do not see statistical significance in terms of the independent variable of 'gear' having a significant effect on the 'vs' variable (i.e., dependent variable). However, it appears as though the other independent variable, ‘hp,’ does have a p-value below the threshold at 0.0141, which is under the threshold of 0.05. Moreover, we can also indicate that the odds are 3.797461e-01:1 for 'gear' and 9.230734e-01 for 'hp.'

Homework Question 5

Question:
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.

Answer/Student Response:
The following was generated and observed:

```
>     PseudoR2(mt.glm)
       McFadden     Adj.McFadden        Cox.Snell       Nagelkerke McKelvey.Zavoina           Effron
      0.6349042        0.4525061        0.5811397        0.7789526        0.8972195        0.6445327
          Count        Adj.Count              AIC   Corrected.AIC
      0.8125000        0.5714286       22.0131402       22.8702830
```

According to the book, "the Nagelkerke comes out consistently larger than the others and Smith and McKenna (2012) suggest that it is the closest analog to the plain, old R-squared that is used in least-squares models." Although there is much debate in the community over which is best, the Nagelkerke value points to 0.7789526 which indicates significance. The fact that our data set is small one suggests that why this result may be more significant than previously thought.

Homework Question 6

Question:

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).

Answer/Student Response:

The following as generated and observed:

```
#data prep
data(Chile)

chile.df <- data.frame(Chile)
chile.N <- chile.df[chile.df$vote =='N',]
chile.Y <- chile.df[chile.df$vote =='Y',]
chile.YN <- rbind(chile.Y, chile.N)
chile.YN <- chile.YN[complete.cases(chile.YN),]
chile.YN$vote <- factor(chile.YN$vote, levels = c('N','Y'))

# conventional model
chile.lm <- glm(vote ~ age + statusquo, family = binomial(), data = chile.YN)
summary(chile.lm)

# bayesian model
chile.YN$vote <- as.numeric(chile.YN$vote) -1
chile.YN
str(chile.YN)
chile.bayesian <- MCMClogit(vote ~ age + statusquo, family = binomial(), data = chile.YN)
chile.bayesian

summary(chile.bayesian)
```

```
>     summary(chile.lm)

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

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

>
```

```
>     summary(chile.bayesian)

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

>
```

We see that the 'statusquo' variable is significant with a p-value of 2e-16;
Furthermore, we see that the HDI of the Bayesian perspective does not span across zero. Because of this observation, we can assume it's likely to also be significant. In summary, it appears as both the conventional and Bayesian tests agree.

Homework Question 7

Question:
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.

Answer/Student Response:
This question is optional and a ‘bonus’ question; therefore, it will not be graded for completeness or accuracy.