IST 777

Joe Hernandez

9/5/2019

Homework #9

1. Use "?mtcars" to display help about data set

```
?mtcars
Use logistic regression to predict vs using gear and hp

> # Basic Logistic Regression
> carOut <- glm(formula = vs ~ gear + hp, family = binomial(link="logit"), data = mtcars)
> summary(carOut)

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
```

With the median residual slightly negative at -0.00889 the distribution of residuals is positively skewed. The coefficients for the intercept is high at 13.43752 and the coefficients for gear and horsepower are -.096825 and -0.08005 respectively. The Null Hypothesis Significance Test, supported by the "Wald" z-test, suggests that horsepower is a better predictor coefficient for vs than gear.

5. From Exercise 1 results generate report and interpret pseudo-R squared value

```
> library("BaylorEdPsych")
> PseudoR2(carOut)
        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
```

The Nagelkerke pseudo-R-squared value is 0.7789526 and measures as the proportion of variance in a vehicle having vs based on the gear and horsepower. It would be worth considering adding variables for improvement.

6. Install and load car packages

```
> library("car")
```

Get access to Chile data set

```
> data("Chile")
```

Isolate cases with yes and no votes

```
ChileY <- Chile[Chile$vote == "Y", ]                         # Grab the Yes votes
ChileN <- Chile[Chile$vote == "N", ]                         # Grab the No votes
ChileYN <- rbind(ChileY, ChileN)                             # Make a new dataset
ChileYN <- ChileYN[complete.cases(ChileYN), ]                # Get rid of missing data
ChileYN$vote <- factor(ChileYN$vote, levels = c("N", "Y"))  # Simplify the factor
```

Replace income variable with statusquo as new predictor to model

```
> # General Linear Model
> chOut <- glm(formula = vote ~ age + statusquo, family = binomial(), data = ChileYN)
> summary(chOut)

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
```

The output shows the intercept is slightly different from 0 with a positive skewness representing a log-odds of "Yes" when age and statusquo are equal to 0. The age predictor would not be significant at this stage if we pursue a threshold of alpha = .05; however, the status quo is strongly significant where we reject the null hypothesis that the log-odds of status quo is 0 in the population but does not hold true for age in this model.

Install and load MCMCpack package

```
> library("MCMCpack")
> # Bayesian Analysis
> ChileYN$vote <- as.numeric(ChileYN$vote) - 1                # Adjust the outcome
variable
> bayesLogitOut <- MCMClogit(formula = vote ~ age + statusquo, data = ChileYN)
> summary(bayesLogitOut)
```

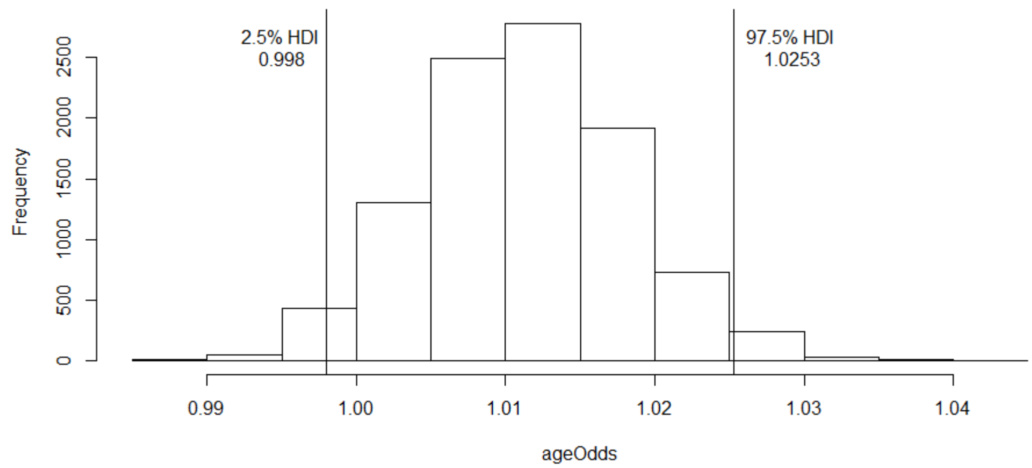
```
Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

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
```

This output focuses on describing the posterior distribution of parameters representing both the intercept and the coefficients on age and status quo. In contrast, the Highest Density Interval of age did not overlap with zero for status quo and did for age.

The AIC for this model is 740.52 compared to example from the book at 109.19. The exercise AIC will be weaker than the one provided in the chapter.

7. Develop function taking posterior distribution of coefficient from output to histogram of distributions of coefficient in terms of regular odds mark vertical lines on histogram to 95% HDI Convert log-odds to odds.

```
> Log2RegOdds <- function (x) # Get r from BayesFactor
+ {
+   ageLogOdds <- as.matrix(x[,"age"])                        # Creates matrix of
posterior distribution
output
+   ageOdds <- apply(ageLogOdds,1,exp)                        # Converts log-odds to
regular odds
+   hist(ageOdds, main=NULL)                                  # Creates histogram of
regular odds output
+   abline(v=quantile(ageOdds,c(0.025)),col="black")         # Draws line for 2.5% HDI
+   abline(v=quantile(ageOdds,c(0.975)),col="black")          # Draws line for 97.5% HDI
+   lowbd <- round(exp(quantile(x[,"age"],c(0.025))), digits = 4) # Gets actual value for
2.5% HDI
+   uppbd <- round(exp(quantile(x[,"age"],c(0.975))), digits = 4) # Gets actual value for
97.5% HDI
+   text((lowbd - (.003 * lowbd)), 2500, c("2.5% HDI \n \n", lowbd)) # Labels 2.5% HDI
+   text((uppbd + (.004 * lowbd)), 2500, c("97.5% HDI \n \n", uppbd)) # Labels 97.5% HDI
+ }
> Log2RegOdds(bayesLogitOut)
```



When converted to regular odds, the mean value of the posterior distribution for age was 1.011275 to 1, suggesting that for every additional year of age, an individual was about 1% more likely to vote to keep Pinochet.