Homework 9

Daniel Caley

8/28/2021

Contents

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

Question 1: Logistic Regression

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.

Interpretting the Results

- Horsepower with a Z value of -2.455 is significant with a P-value of 0.0141.
- Gear does not have a significant P-value
- The Chi Square test shows a difference of 26.4814 at a significant P-value of 0.00000026. Gear does not a significant P-Value
- The graph illustrate that for each unit change in the value of X, odds that Y=1 is the correct prediction increases by 3.78:1 for gears and 9.123:1 for hp.
- Said differently as hp increases by 1 unit, mpg deceases by 92 percent for every unit.

```
?mtcars

MyMtcars <- mtcars

LogitCars <- glm(vs ~ gear + hp, MyMtcars, family = binomial())
summary(LogitCars)
```

```
##
## Call:
## glm(formula = vs ~ gear + hp, family = binomial(), data = MyMtcars)
##
## 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
```

```r
anova(LogitCars, test = "Chisq")
```

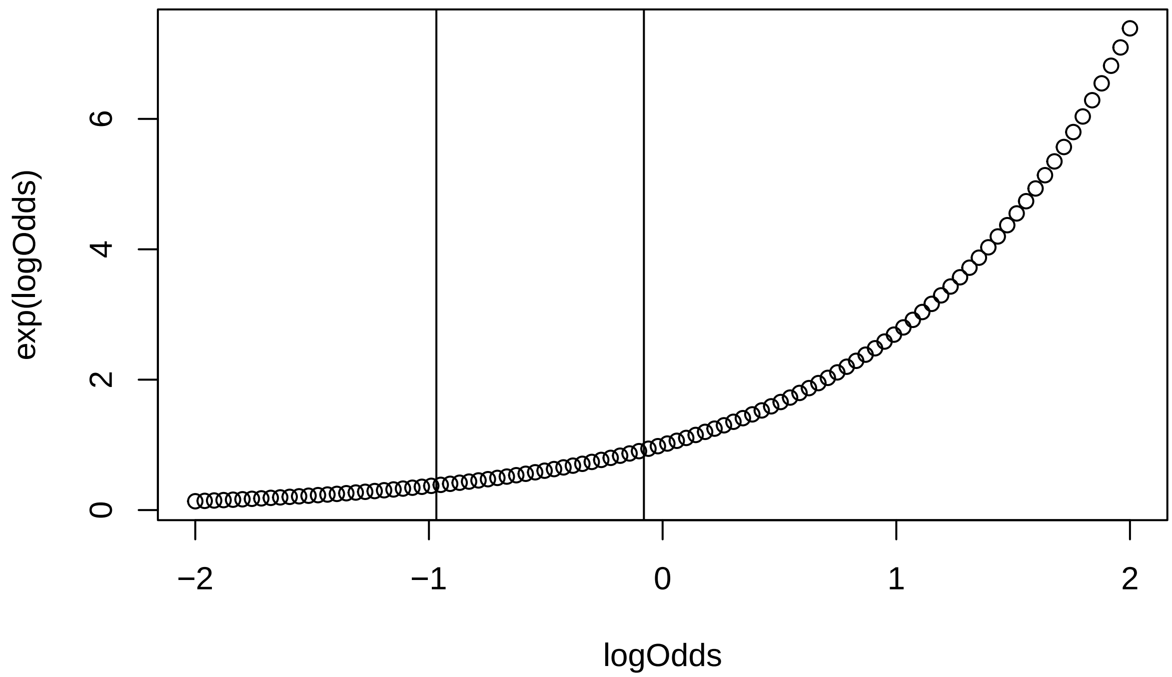
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: vs
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    31     43.860
## gear  1   1.3656        30     42.495    0.2426
## hp    1  26.4814        29     16.013 2.661e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
CarsCoef <-  exp(coef(LogitCars))
CarsCoef
```

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

```r
logOdds <- seq(from=-2, to = 2, length.out=100)

plot(logOdds, exp(logOdds))
abline(v = log(CarsCoef[2]))
abline(v = log(CarsCoef[3]))
```


exp(logOdds)
0
2
4
6
−2
−1
0
1
2
logOdds

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.

```
PseudoR2(LogitCars)
```

```
##         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
```

Interpreting the Results

The Suedo R-Sqaured shows a value of 0.7789. Said differently 78% of the predictor variables (HP and Gear) can explain the engine shape.

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).

```
MyChile <- Chile %>%
            filter(vote %in% c("Y","N")) %>%
            mutate(
                   vote =  factor(vote,levels=c("N","Y")),
                   vote = as.numeric(vote) - 1
                   )

MyChile <- MyChile[complete.cases(MyChile),] # Get rid of missing

LogitChile <- glm(vote ~ age + statusquo, MyChile,family = binomial())
summary(LogitChile)
```

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

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

```
## Analysis of Deviance Table
##
```

```
## Model: binomial, link: logit
##
## Response: vote
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       1702    2360.29
## age        1     34.2      1701    2326.09 4.964e-09 ***
## statusquo  1   1591.6      1700     734.52 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PseudoR2(LogitChile)
```

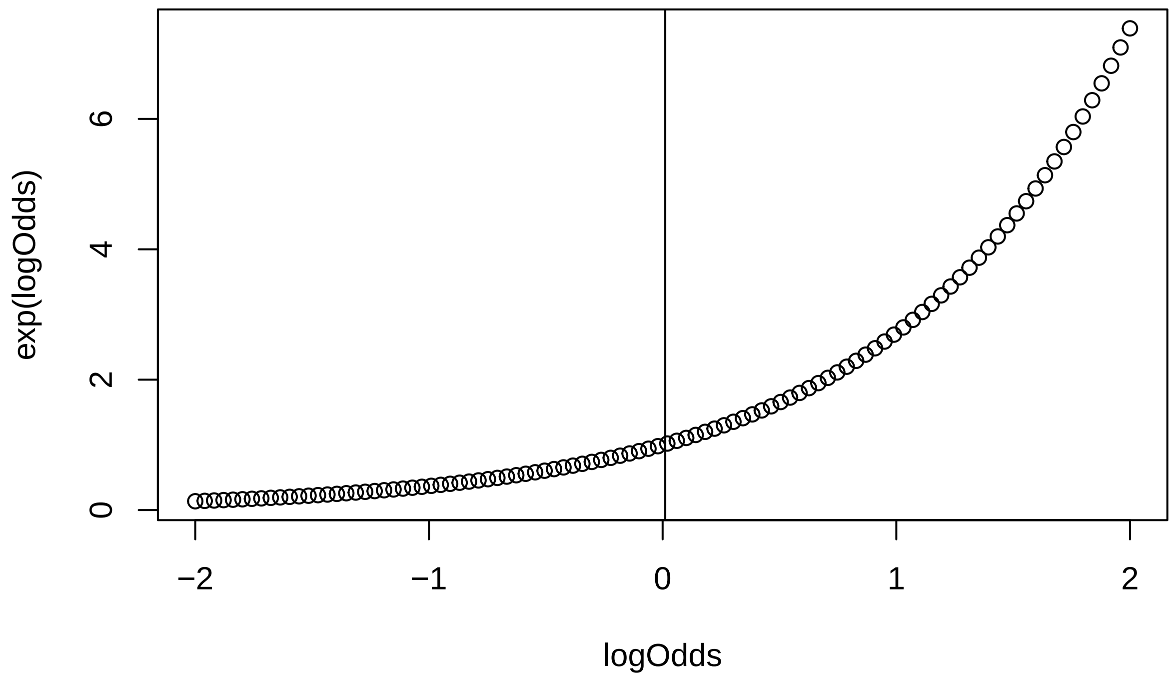
```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##        0.6888013        0.6854119        0.6150544        0.8201631
## McKelvey.Zavoina           Effron            Count        Adj.Count
##        0.7855565        0.7553412        0.9230769        0.8433014
##              AIC    Corrected.AIC
##      740.5206862      740.5348122
```

```
ChilCoef <-  exp(coef(LogitChile))
ChilCoef
```

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

```
logOdds <- seq(from=-2, to = 2, length.out=100)

plot(logOdds, exp(logOdds))
abline(v = log(ChilCoef[2]))
abline(v = log(ChilCoef[3]))
```




```
# Bayes Piece
set.seed(271) # Control randomization
bayesLogitOut <- MCMClogit(formula = vote ~ age + statusquo, data = MyChile)
summary(bayesLogitOut) # Summarize the results
```

```
##
## 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
```

```r
# Age
ageLogOdds <- as.matrix(bayesLogitOut[,"age"])
ageOdds <- apply(ageLogOdds,1,exp) # Transform with exp()
mean(ageOdds) # The point estimate for age in plain odds
```

```
## [1] 1.011319
```

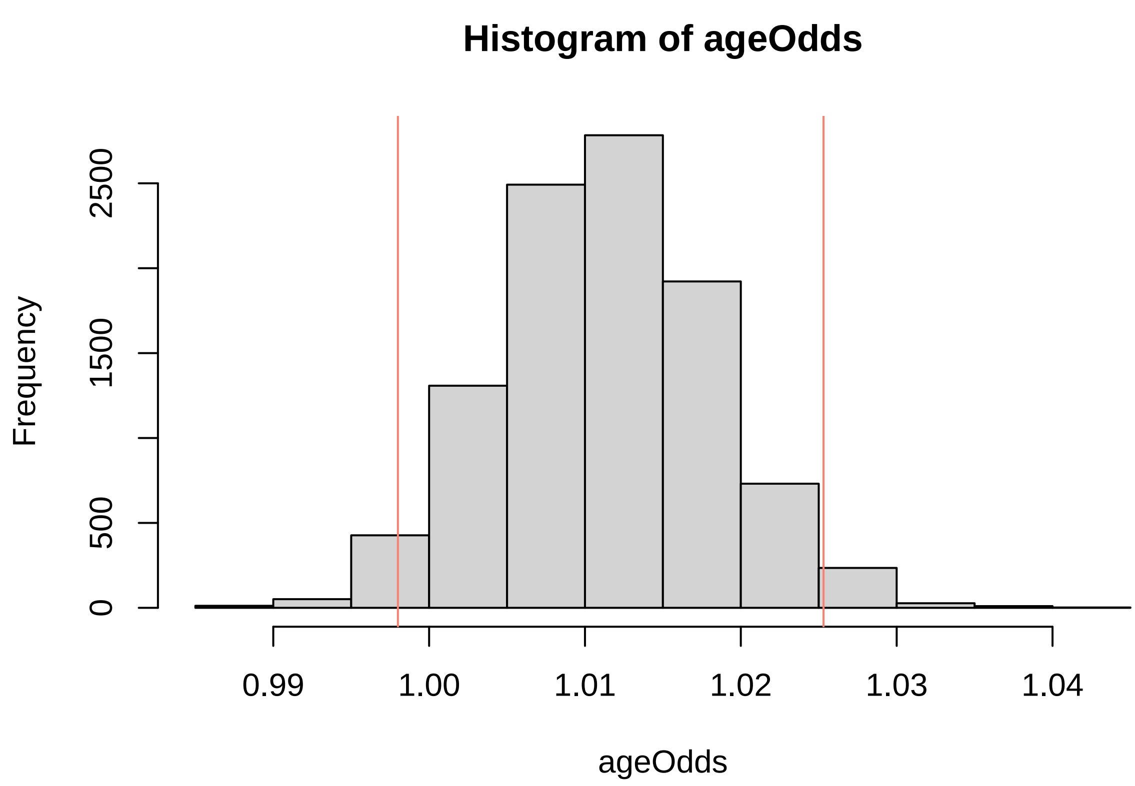
```r
quantile(ageOdds,c(0.025)) # Lower bound of HDI
```

```
##      2.5%
## 0.9979972
```

```r
quantile(ageOdds,c(0.975)) # Upper bound of HDI
```

```
##    97.5%
## 1.025307
```

```r
hist(ageOdds)
abline(v=quantile(ageOdds,c(0.025)), col='salmon')
abline(v=quantile(ageOdds,c(0.975)), col='salmon')
```



```r
actualVote <- MyChile$vote
predictedVote <- round(predict(LogitChile, type='response')) # round() splits probabilities at 0.5
ChiliConfus <-  table(predictedVote, actualVote)
ChiliConfus
```

```
##              actualVote
## predictedVote   0   1
##             0 810  74
##             1  57 762
```

```r
print("error rate:")
```

```
## [1] "error rate:"
```

```r
yvote <- ChiliConfus[2,1]
agevote <- ChiliConfus[1,2]

(yvote+agevote)/sum(ChiliConfus)
```

```
## [1] 0.07692308
```

Interpretting the results

- The Suedo R-squared is 68% indicating not a strong relationship between predicting if the Chilean plebiscite would be voted on.
- The p-values for age is 0.097 which is not under the 0.05 alpha which we would want.
- The p-value for status quo is under the 0.05 at 2-e16.

When interpretting the Bayes Theorom we notice the following.

We examined data from the 1988 Chilean plebiscite, to see if the age and statuesque of a voter could predict whether an individual would vote in favor of keeping Augusto Pinochet in office. We conducted a Bayesian logistic analysis, using age and statuesque to predict votes. The Highest Density Interval of age overlap with zero. When converted to regular odds, the mean value of the posterior distribution for age was 1.01 to 1, suggesting that for every additional year of age, an individual was about 1% more likely to vote to keep Pinochet. In addition with the p-value being so low for the Logit Regression model for age is not a significant determinant of voting behavior. Status Quo appears to be the better predictor. The confusion matrix showed that the overall error rate was 8% indicating that the logistic model for age was good at predicting votes. Though the HDI does overlap for 0 barely for age we fail to reject the null hypothesis.

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.

```
# Status Quo Funciton

PostDistroHist <- function(x,y){
  stutusQuoLogOdds <- as.matrix(x[,y])
  stutusQuoLogOdds <- apply(stutusQuoLogOdds,1,exp) # Transform with exp()
  mean(stutusQuoLogOdds) # The point estimate for age in plain odds
  quantile(stutusQuoLogOdds,c(0.025)) # Lower bound of HDI
  quantile(stutusQuoLogOdds,c(0.975)) # Upper bound of HDI



  hist(stutusQuoLogOdds)
  abline(v=quantile(stutusQuoLogOdds,c(0.025)), col='salmon')
  abline(v=quantile(stutusQuoLogOdds,c(0.975)), col='salmon')

}

PostDistroHist(bayesLogitOut, "statusquo")
```