Homework 9

Kwasi Mensah

2022-12-9

Author: Kwasi Mensah Homwork_Number: 9 Output: pdf_document Attribution statement:

I did the homework by myself, with help from the book and the professor and the following sources: https://www.r-tutor.com/elementary-statistics/multiple-linear-regression/estimated-multiple-regression-equation https://www.displayr.com/variance-inflation-factors-vifs/

#R Markdown #Run these three functions to get a clean test of homework code

```
dev.off() #Clear the graph window
```

```
## null device
##           1
```

```
cat('\014') #Clear the console
```

```
rm(list = ls()) #Clear user objects from the environment
```

#R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
##
##############

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.

```
?mtcars
```

```
## starting httpd help server ... done
```

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

##
## 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(MTout))

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

#     We immediately observe that the coefficient is significantly different
from zero, as this is supported by the "Wald" z-test (conceptually similar to
a t-test) and the associated p-value.The z-test value of 1.871 and the
associated p-value of 0.0613 of the intercept is not less than .05, thus we
will fail to reject the null hypothesis. The gear variable is not
statistically significant because it has a z-value of -0.858 and associated p
value of .3907 which is not less than our threshold of .05. thus we will fail
to reject the null hypothesis. The hp variable has a z value of -2.455 and
the associated p-value of 0.0141, this is less than our threshold of .05.
Because our result for this variable is statistically significant, we can
reject the null hypothesis and make the interpretation that as hp increases
the vs being straight decreases. For each unit change in X, odds of Y=1
increase by 9.230734e-01:1
```

5. As noted in the chapter, the BaylorEdPsych add-in package contains a procedure for generating pseudo-Rsquared 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.

```
install.packages("C:/Users/lmori/Downloads/BaylorEdPsych_0.5.tar.gz",repos =
NULL, type = "source")

## Installing package into 'C:/Users/lmori/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)

library(BaylorEdPsych)
PseudoR2(MTout)

##         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

#The Nagelkerke pseudo R-squared value is 0.7789526, this can be loosely
interpreted as the proportion of variance in the outcome variable, vs (v-
shaped, straight), accounted for by the predictor variables, hp and gear.
given that only the hp variable was statistically significant with a p-value
of 0.0141 being less than .05, the results suggest that hp variable has a
small role in accounting for the vs variable.
```

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

data(Chile)
ChileN=Chile[Chile$vote=='N',]
ChileY=Chile[Chile$vote=='Y',]
ChileYN=rbind(ChileN, ChileY)
ChileYN=ChileYN[complete.cases(ChileYN),]
ChileYN$vote=factor(ChileYN$vote, levels=c("N", "Y"))


CHglm <- glm(vote ~ age + statusquo, family=binomial(), data=ChileYN)
summary(CHglm)
```

```
##
## 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

library(MCMCpack)

## Warning: package 'MCMCpack' was built under R version 4.2.2

## 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)
## ##

ChileYN$vote=as.numeric(ChileYN$vote)-1
BayesChile <- MCMClogit(formula=vote~ age + statusquo, data=ChileYN)
summary(BayesChile)

##
## 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
```

```
#    In our analysis the intercept has a z-value of -0.716 and a p-value of
.4741 and in addition the age variable has a z-value of 1.659 and a p-value
of .0972. Both the intercept and age variable are over our threshold of .05
making the two not statistically significant, thus we will fail to reject the
null hypothesis the log odds of the two are 0. The statusquo variable has a
z-value of 22.0570 and a p-value of <2e-16, which is below our threshold of
.05, and we will reject the null hypothesis that the log-odds of statusquo
variable is 0.

#    The HDI for the intercept is -0.742761 to  0.34439 and the HDI for the
age variable is -0.002005 to 0.02499, both of these intervals passes 0 which
supports our regression model of the two not being statistically significant.
as a result we will fail to reject the null hypothesis.  The statusquo
variable has an HDI range of 2.914442 to 3.48698 , this range does not pass 0
meaning that it is statistically significant and we will reject the null
hypothesis. There is a 95% probability that the population log-odds for
statusquo variable falls within the given HDI range.

#     AIC is good for comparing nonnested models, we should always determine
the model with the lower AIC value as the better model. The AIC for our
analysis is 740.52 and the AIC for the analysis in the book is 2332, thus we
should choose the model with the statusquo variable. In addition the model
with the lower AIC achieves lower error reduction because it is less complex
compared to a model with a high AIC value.
```

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.

```
functionhw9<- function(BayesChile, seq){
  statusquologodds <- as.matrix(BayesChile[,'statusquo'])
```

```
  statusquoodds <- apply(statusquologodds,1,exp)
  hist(statusquoodds, col="orange")
  abline(v=quantile(statusquoodds,c(.025)),col='blue')
  abline(v=quantile(statusquoodds,c(.975)),col='blue')
}
functionhw9(BayesChile, 'statusquo')
```