COSC6323 - Exercise 9

Sachin Shubham

4/10/2021

Task 1:

Work with the project data. Build a regression model based on the article level data which is predicting the cross-disciplinary in CIP (XCIPp variable) Classification of Instructional Programs, using year of publication, log-transformed coauthors count, log-transformed Major MeSH count, regions count and total number of SAps (NSAp) as predictors. Get model summary, pseudo r-squared measures, odds ratio and comment about the results.

Solution:

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(modelr)
library(broom)

## Warning: package 'broom' was built under R version 4.0.4

##
## Attaching package: 'broom'

## The following object is masked from 'package:modelr':
##
##     bootstrap

library(ROCR)

## Warning: package 'ROCR' was built under R version 4.0.5

library(questionr)

## Warning: package 'questionr' was built under R version 4.0.5
```

```r
setwd("D:/Statistical Methods/Project")
df_article<-read.csv("ArticleLevel-RegData-
ALLSA_Xc_1_NData_655386_LONGXCIP2.csv")
df_article<-as_tibble(df_article)


model1 <- glm(XCIPp ~ Yp + log(Kp) + log(nMeSHMain) + NRegp + NSAp, data =
df_article, family=binomial(link='logit'))
summary(model1)
```

```
##
## Call:
## glm(formula = XCIPp ~ Yp + log(Kp) + log(nMeSHMain) + NRegp +
##     NSAp, family = binomial(link = "logit"), data = df_article)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2664  -0.4450  -0.3693  -0.2874   2.8338
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.157e+01  1.093e+00 -19.736  < 2e-16 ***
## Yp              7.752e-03  5.467e-04  14.178  < 2e-16 ***
## log(Kp)         6.198e-01  7.180e-03  86.322  < 2e-16 ***
## log(nMeSHMain) -3.984e-02  1.090e-02  -3.655 0.000257 ***
## NRegp           1.992e+00  7.411e-03 268.825  < 2e-16 ***
## NSAp            1.938e-01  4.416e-03  43.890  < 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: 542453  on 655385  degrees of freedom
## Residual deviance: 418616  on 655380  degrees of freedom
## AIC: 418628
##
## Number of Fisher Scoring iterations: 5
```

```r
pscl::pR2(model1)["McFadden"]
```

```
## fitting null model for pseudo-r2
```

```
##  McFadden
## 0.2282903
```

```r
output = odds.ratio(model1) # HEAVY COMPUTATIONAL!
```

```
## Waiting for profiling to be done...
```

```r
output = apply(output, 2, formatC, format="f", digits=4)
output
```

```
##                  OR       2.5 %    97.5 %   p       
## (Intercept)    "0.0000" "0.0000" "0.0000" "0.0000"
## Yp             "1.0078" "1.0067" "1.0089" "0.0000"
## log(Kp)        "1.8585" "1.8326" "1.8849" "0.0000"
## log(nMeSHMain) "0.9609" "0.9406" "0.9817" "0.0003"
## NRegp          "7.3321" "7.2265" "7.4395" "0.0000"
## NSAp           "1.2139" "1.2034" "1.2244" "0.0000"
```

Conclusion:

The summary of the model shows that:

1. All of the variables are highly significant in predicting cross-disciplinary in CIP with all p-values are quite less than 0.05. The number of co-authors and region the article was written appear to also have a close standard error.

2. As we know that values of pseudo r-squared [ρ2] between 0.2 to 0.4 represent an excellent fit.Our model pseudo r-squared [ρ2] is 0.2282903, which represent our model is excellent fit.

3. The odds ratio shows that from the data given, the region from where the article research took place was 7.33 times more successful in predicting whether the article had cross-disciplinary co-authors. The log number of co-authors was also successful in predicting whether the article had cross-disciplinary co-authors with a success-to-failure ratio of 1.85. All variables in the odds ratio were shown to be extremely significant.