# Exercise-9

Mohammad Imtiaz Nur

4/8/2021

## ID : 1878074

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(rcompanion)
library(questionr)
library(modelr)
library(broom)
```

```
##
## Attaching package: 'broom'

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

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
#### Filters
###### Filter Year [1970-2018]
###### Filter Kp >= 2 and Wp >= 2
df_article = df_article %>% filter(Yp >= 1970)
df_article = df_article %>% filter(Yp <= 2018)
df_article = df_article %>% filter(Kp >= 2)
df_article = df_article %>% filter(nMeSHMain >= 2)
df_article = df_article %>% filter(IRegionRefinedp > 0 & IRegionRefinedp < 7)

#### Convert Data types
df_article$eidsp = as.factor(df_article$eidsp)
df_article$Yp = as.integer(df_article$Yp)
df_article$Kp = as.integer(df_article$Kp)
df_article$XCIPp = as.factor(df_article$XCIPp)
df_article$NRegp = as.integer(df_article$NRegp)
df_article$NSAp = as.integer(df_article$NSAp)
df_article$NCIPp = as.integer(df_article$NCIPp)
df_article$nMeSHMain = as.integer(df_article$nMeSHMain)
df_article$IRegionRefinedp = as.factor(df_article$IRegionRefinedp)

## Model 1 - for X_CIP
options(scipen=2)
model1 <- glm(XCIPp ~ Yp + log(Kp) + log(nMeSHMain) + NRegp + NSAp,
              data = df_article, family=binomial(link='logit'))

# MeanZJp

# Here:
# XCIPp: binary indicator variable = 1 if any 2+ CIP are present, and 0 otherwise
# Yp: article's publication year
# Kp: article's coauthor count based upon author list in PubMed record
# NRegp: article's count variable indicating the total number of regions
# NSAp: article's count variable indicating the total number of SAps
```

## Model Summary

```
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
## -2.9327  -0.4484  -0.3764  -0.3047   2.7242
##
## Coefficients:
##                   Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)    -21.5304627   1.1289951 -19.070   < 2e-16 ***
```

```
## Yp               0.0077249   0.0005651  13.671   < 2e-16 ***
## log(Kp)          0.5919190   0.0076580  77.295   < 2e-16 ***
## log(nMeSHMain)  -0.0526264   0.0122992  -4.279 0.0000188 ***
## NRegp            2.0774934   0.0081746 254.141   < 2e-16 ***
## NSAp             0.1963843   0.0044915  43.723   < 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: 490224  on 602598  degrees of freedom
## Residual deviance: 395757  on 602593  degrees of freedom
## AIC: 395769
##
## Number of Fisher Scoring iterations: 5
```

From the above summary, all the p-values of the predictors are very much less than the significance value (0.05). So, all the variables are very much significant for the model for predicting the cross-disciplinary in CIP (XCIPp variable) Classification of Instructional Programs.

## Pseudo r-squared measures

```
nagelkerke(model1)
```

```
## $Models
##
## Model: "glm, XCIPp ~ Yp + log(Kp) + log(nMeSHMain) + NRegp + NSAp, binomial(link = \"logit\"), df_ar
## Null:  "glm, XCIPp ~ 1, binomial(link = \"logit\"), df_article"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                            0.192703
## Cox and Snell (ML)                  0.145097
## Nagelkerke (Cragg and Uhler)        0.260635
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff Chisq p.value
##       -5      -47234 94468       0
##
## $Number.of.observations
##
## Model: 602599
## Null:  602599
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitting with ML"
##
## $Warnings
## [1] "None"
```

Our model is compared against the null model. From the McFadden's pseudo r-squared values ranging from 0.2 to 0.4 is a good model fit. The pseudo r-squared values ranges from 0.19 to 0.26 for our model which means though it captures the variance of the data moderately but still not a strong fit.

## Odds ratio

```
output = odds.ratio(model1)
```

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

```
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" "1.0066" "1.0089" "0.0000"
## log(Kp)        "1.8075" "1.7805" "1.8348" "0.0000"
## log(nMeSHMain) "0.9487" "0.9261" "0.9719" "0.0000"
## NRegp          "7.9844" "7.8576" "8.1135" "0.0000"
## NSAp           "1.2170" "1.2063" "1.2278" "0.0000"
```

Odds ratio $> 1$ means greater likelihood of having the outcome while $< 1$ refers to lower likelihood of having the outcome.

From the output of the odds ratio, except Major Mesh all the predictors count have odds ratio values greater than 1 which refers to greater likelihood of having the outcome.

The odds ratio for Major Mesh count is less than 1 refers to the likelihood of predicting the outcome is lower by 6.68%.