H2

Carina\_Sun

2024-03-15

knitr::opts\_chunk$set(echo = TRUE)  
  
getwd()

## [1] "/Users/carina/Documents/Multivariant"

library(psych)  
library(stats)  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:psych':  
##   
## logit

library(tinytex)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ ggplot2::%+%() masks psych::%+%()  
## ✖ ggplot2::alpha() masks psych::alpha()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::recode() masks car::recode()  
## ✖ purrr::some() masks car::some()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ResourceSelection)

## ResourceSelection 0.3-6 2023-06-27

religion <- read.csv("RELIGION.csv")  
  
# 1: Check the data structure  
str(religion)

## 'data.frame': 626 obs. of 10 variables:  
## $ id : int 2 3 4 5 6 7 9 10 13 15 ...  
## $ sex : chr "female" "female" "female" "female" ...  
## $ age : int 30 32 51 18 37 18 90 70 57 72 ...  
## $ educ : chr "postgraduate work/degree" "postgraduate work/degree" "high school graduate (12)" "high school graduate (12)" ...  
## $ income : chr "$100,000-$149,999" "$40,000-$49,999" "$100,000-$149,999" "$15,000-$19,999" ...  
## $ relschol: chr "no" "no" "no" "no" ...  
## $ married : chr "married" "married" "married" "not married" ...  
## $ attend : int 6 5 2 6 6 5 5 5 6 5 ...  
## $ agesq : int 900 1024 2601 324 1369 324 8100 4900 3249 5184 ...  
## $ race : chr "white" "white" "white" "white" ...

# The table includes a mix of categorical variables (=sex, educ, married, race..) and numerical variables (age, attend, agesq). Some categorical variables are binomial - like race, school and relschol  
  
# 2: Re-run the data changing the all character strings to factors using the stringAsFactors = T argument when you read in the data.   
religion <- read.csv("RELIGION.csv", stringsAsFactors = TRUE)  
  
# 3. Describe the data   
describe(religion)

## vars n mean sd median trimmed mad min max range skew  
## id 1 626 503.49 289.64 496.5 503.62 381.77 2 1002 1000 -0.01  
## sex\* 2 626 1.42 0.49 1.0 1.39 0.00 1 2 1 0.34  
## age 3 623 47.08 17.40 44.0 46.09 19.27 18 99 81 0.48  
## educ\* 4 626 3.99 1.64 3.0 3.94 1.48 1 7 6 0.27  
## income\* 5 626 6.85 3.14 7.0 6.80 2.97 1 13 12 0.01  
## relschol\* 6 626 1.13 0.33 1.0 1.04 0.00 1 2 1 2.22  
## married\* 7 626 1.36 0.48 1.0 1.33 0.00 1 2 1 0.56  
## attend 8 626 4.54 1.13 5.0 4.66 1.48 1 6 5 -0.97  
## agesq 9 623 2518.50 1811.66 1936.0 2287.85 1535.97 324 9801 9477 1.05  
## race\* 10 626 1.84 0.37 2.0 1.92 0.00 1 2 1 -1.82  
## kurtosis se  
## id -1.21 11.58  
## sex\* -1.89 0.02  
## age -0.65 0.70  
## educ\* -1.43 0.07  
## income\* -0.43 0.13  
## relschol\* 2.95 0.01  
## married\* -1.69 0.02  
## attend 0.43 0.05  
## agesq 0.39 72.58  
## race\* 1.32 0.01

describeBy(religion, group = religion$relschol)

##   
## Descriptive statistics by group   
## group: no  
## vars n mean sd median trimmed mad min max range skew  
## id 1 546 503.06 292.82 500 503.64 386.96 2 1002 1000 -0.02  
## sex 2 546 1.41 0.49 1 1.39 0.00 1 2 1 0.36  
## age 3 543 48.49 17.99 46 47.77 20.76 18 99 81 0.32  
## educ 4 546 3.99 1.64 3 3.95 1.48 1 7 6 0.27  
## income 5 546 6.81 3.19 7 6.75 2.97 1 13 12 0.05  
## relschol 6 546 1.00 0.00 1 1.00 0.00 1 1 0 NaN  
## married 7 546 1.38 0.49 1 1.36 0.00 1 2 1 0.47  
## attend 8 546 4.50 1.15 5 4.62 1.48 1 6 5 -0.91  
## agesq 9 543 2674.14 1877.94 2116 2466.29 1802.84 324 9801 9477 0.88  
## race 10 546 1.86 0.35 2 1.95 0.00 1 2 1 -2.08  
## kurtosis se  
## id -1.23 12.53  
## sex -1.87 0.02  
## age -0.84 0.77  
## educ -1.43 0.07  
## income -0.48 0.14  
## relschol NaN 0.00  
## married -1.78 0.02  
## attend 0.26 0.05  
## agesq 0.00 80.59  
## race 2.33 0.01  
## ------------------------------------------------------------   
## group: yes  
## vars n mean sd median trimmed mad min max range skew  
## id 1 80 506.43 268.68 482.5 501.83 337.29 37 995 958 0.09  
## sex 2 80 1.45 0.50 1.0 1.44 0.00 1 2 1 0.20  
## age 3 80 37.49 7.58 38.0 37.36 5.93 19 60 41 0.26  
## educ 4 80 3.98 1.63 3.0 3.92 1.48 2 7 5 0.32  
## income 5 80 7.12 2.73 7.0 7.20 1.48 1 13 12 -0.29  
## relschol 6 80 2.00 0.00 2.0 2.00 0.00 2 2 0 NaN  
## married 7 80 1.23 0.42 1.0 1.16 0.00 1 2 1 1.29  
## attend 8 80 4.76 0.92 5.0 4.89 0.00 2 6 4 -1.47  
## agesq 9 80 1462.06 590.92 1444.0 1414.42 426.99 361 3600 3239 1.03  
## race 10 80 1.68 0.47 2.0 1.72 0.00 1 2 1 -0.73  
## kurtosis se  
## id -1.10 30.04  
## sex -1.99 0.06  
## age 0.61 0.85  
## educ -1.46 0.18  
## income -0.06 0.31  
## relschol NaN 0.00  
## married -0.33 0.05  
## attend 2.26 0.10  
## agesq 1.92 66.07  
## race -1.48 0.05

# 4. For the ordered categorical variables, income, we check the order of the levels. Check the order of income levels.   
levels(religion$income)

## [1] "" "$10,000-$14,999" "$100,000-$149,999"  
## [4] "$15,000-$19,999" "$150,000 or over" "$20,000-$29,999"   
## [7] "$30,000-$39,999" "$40,000-$49,999" "$50,000-$59,999"   
## [10] "$60,000-$69,999" "$70,000-$79,999" "$80,000-$99,999"   
## [13] "under $10,000"

#a. What is the problem with the income order and why is this a problem? - This is not in order. It is in order of the numbers and then alphabets rather than the numerical values. We need to reorder them.  
  
religion$income[religion$income==""] <- NA  
religion$income <-factor(religion$income,levels(religion$income)[c(13, 2,4,6:12,3,5)])  
# 4b.Check again, found in right order!  
levels(religion$income)

## [1] "under $10,000" "$10,000-$14,999" "$15,000-$19,999"   
## [4] "$20,000-$29,999" "$30,000-$39,999" "$40,000-$49,999"   
## [7] "$50,000-$59,999" "$60,000-$69,999" "$70,000-$79,999"   
## [10] "$80,000-$99,999" "$100,000-$149,999" "$150,000 or over"

# 5. Create a new variable called "income0" which treats income categories as a continuous scale.   
religion$income0 <- as.numeric(religion$income)  
# Yes, they are in the same order!  
  
# 6. The dichotomous outcome variable   
religion$relschol <- as.numeric(religion$relschol)-1  
religion$race <- as.numeric(religion$race)-1  
  
# 7.get rid of the NA’s   
religion <- religion %>% dplyr:: select(race, attend, income0,relschol) %>% na.omit()  
str(religion)

## 'data.frame': 590 obs. of 4 variables:  
## $ race : num 1 1 1 1 1 1 0 1 1 0 ...  
## $ attend : int 6 5 2 6 6 5 5 6 5 3 ...  
## $ income0 : num 11 6 11 3 6 3 1 4 2 1 ...  
## $ relschol: num 0 0 0 0 0 0 0 0 0 0 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:36] 7 38 48 51 56 59 62 74 75 83 ...  
## ..- attr(\*, "names")= chr [1:36] "7" "38" "48" "51" ...

# The number of observations decreased from 626 to 590 because we omitted the row with NA.  
  
# 8. Run a glm logit regression  
relschol\_logit <- glm(relschol ~ race + attend + income0, data = religion, family = binomial)  
  
#8a.  
anova(relschol\_logit, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: relschol  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 589 453.27   
## race 1 13.8791 588 439.39 0.0001949 \*\*\*  
## attend 1 5.4822 587 433.91 0.0192105 \*   
## income0 1 17.1132 586 416.79 3.522e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# This is good fit. The residual deviance of all factors are significantly lower than the null deviance, especially race and income.  
  
# 8b. Exponentiate  
exp(coef(relschol\_logit))

## (Intercept) race attend income0   
## 0.0277883 0.2754600 1.3932522 1.2222368

#For the race variable, an individual from the reference group (non-white) has odds of attending a religious school that are 72.4% less than the odds for white individuals  
#For every additional religious service attended, the odds of being admitted to a religious school increase by a factor of 1.393  
#For each one-unit increase in the income category, the odds of attending a religious school increase by a factor of 1.222  
  
#8c  
vif(relschol\_logit)

## race attend income0   
## 1.049275 1.020813 1.054785

# No evidence of multicollinearity since data is no greater 5  
#8d  
car::Anova(relschol\_logit, type=3)

## Analysis of Deviance Table (Type III tests)  
##   
## Response: relschol  
## LR Chisq Df Pr(>Chisq)   
## race 18.3399 1 1.848e-05 \*\*\*  
## attend 7.3148 1 0.006839 \*\*   
## income0 17.1132 1 3.522e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# All the independent variables are statistically significant to improve the model's fit.  
#8e  
hoslem.test(relschol\_logit$y, fitted(relschol\_logit))

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: relschol\_logit$y, fitted(relschol\_logit)  
## X-squared = 10.782, df = 8, p-value = 0.2143

# This is good fit, the p value is bigger than 0.05, so the model does fit the data well.  
  
# 9. Re-run the same model, taking income0 out of the model and assign it to the name relschol\_logit2  
relschol\_logit2 <- glm(relschol ~ race + attend, data = religion, family = binomial)  
#9a  
anova(relschol\_logit2, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: relschol  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 589 453.27   
## race 1 13.8791 588 439.39 0.0001949 \*\*\*  
## attend 1 5.4822 587 433.91 0.0192105 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# This is good fit. The residual deviance of both factors are significantly lower than the null deviance, especially race.  
# 9b. Exponentiate  
exp(coef(relschol\_logit2))

## (Intercept) race attend   
## 0.09414639 0.33150441 1.32367662

#The odds of attending a religious school for the white group decrease by a factor of about 0.332 compared to the non-white group   
#For each additional religious service attended, the odds of attending a religious school increase by a factor of 1.324  
#9c  
vif(relschol\_logit2)

## race attend   
## 1.004888 1.004888

# No evidence of multicollinearity since data is no greater 5  
#9d  
car::Anova(relschol\_logit2, type=3)

## Analysis of Deviance Table (Type III tests)  
##   
## Response: relschol  
## LR Chisq Df Pr(>Chisq)   
## race 14.5408 1 0.0001372 \*\*\*  
## attend 5.4822 1 0.0192105 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Both the independent variables are statistically significant to improve the model's fit.  
#9e  
hoslem.test(relschol\_logit2$y, fitted(relschol\_logit2))

## Warning in hoslem.test(relschol\_logit2$y, fitted(relschol\_logit2)): The data  
## did not allow for the requested number of bins.

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: relschol\_logit2$y, fitted(relschol\_logit2)  
## X-squared = 4.9859, df = 4, p-value = 0.2887

# This is good fit, the p value is over 0.05, so the model does fit the data well.  
  
# 10. Compare the models using   
AIC(relschol\_logit)

## [1] 424.793

AIC(relschol\_logit2)

## [1] 439.9062

anova(relschol\_logit, relschol\_logit2, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model 1: relschol ~ race + attend + income0  
## Model 2: relschol ~ race + attend  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 586 416.79   
## 2 587 433.91 -1 -17.113 3.522e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# relschol\_logit has lower residual deviance than relschol\_logit2 with p value lower than 0.05, indicating it is significantly a better fit.   
  
# 11. generate a new dataframe  
newdata <- data.frame(attend = 5, income0 = 4, race = 0)  
predict\_value<- predict(relschol\_logit, newdata, type = "response")  
predict\_value

## 1   
## 0.2456022

#students who attend religious services five times per month, have a family income of $20,000-$29,999, and are non-white have a 24.5% predicted probability of attending a religious school, according to the model