UCI Data Analysis

Julian Lucero

2024-05-13

###Upload Data from excel  
library(readxl)  
adult <- read\_excel("adult.xlsx")  
View(adult)

###Remove "?'s"  
New <- adult  
New[New == "?"] <- NA  
Data <- na.omit(New)  
print(dim(Data))

## [1] 45222 15

###Make subset with studied variables   
Age\_Data <- Data[, c("age", "race", "educational\_num", "gender", "hours\_per\_week", "income", "relationship")]  
print(Age\_Data)

## # A tibble: 45,222 × 7  
## age race educational\_num gender hours\_per\_week income relationship   
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 25 Black 7 Male 40 <=50K Own-child   
## 2 38 White 9 Male 50 <=50K Husband   
## 3 28 White 12 Male 40 >50K Husband   
## 4 44 Black 10 Male 40 >50K Husband   
## 5 34 White 6 Male 30 <=50K Not-in-family  
## 6 63 White 15 Male 32 >50K Husband   
## 7 24 White 10 Female 40 <=50K Unmarried   
## 8 55 White 4 Male 10 <=50K Husband   
## 9 65 White 9 Male 40 >50K Husband   
## 10 36 White 13 Male 40 <=50K Husband   
## # ℹ 45,212 more rows

#Mutate Data to age groups for Table 1  
library(dplyr)

##   
## Attaching package: 'dplyr'

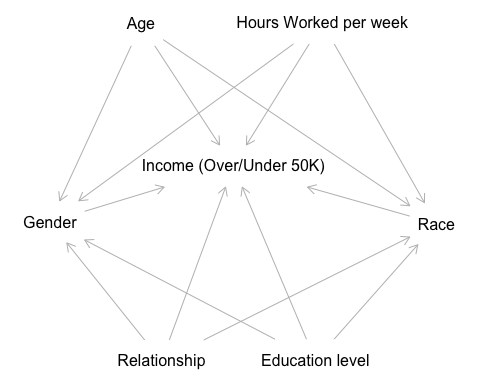
## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

Age\_Data <- Age\_Data %>%  
 mutate(age2 = case\_when(  
 age >= 17 & age <= 30 ~ 1,  
 age >= 31 & age <= 40 ~ 2,  
 age >= 41 & age <= 50 ~ 3,  
 age >= 51 & age <= 60 ~ 4,  
 age >= 61 & age <= 70 ~ 5,  
 age >= 71 ~ 6  
 ))  
  
head(Age\_Data)

## # A tibble: 6 × 8  
## age race educational\_num gender hours\_per\_week income relationship age2  
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr> <dbl>  
## 1 25 Black 7 Male 40 <=50K Own-child 1  
## 2 38 White 9 Male 50 <=50K Husband 2  
## 3 28 White 12 Male 40 >50K Husband 1  
## 4 44 Black 10 Male 40 >50K Husband 3  
## 5 34 White 6 Male 30 <=50K Not-in-family 2  
## 6 63 White 15 Male 32 >50K Husband 5

library(dagitty)  
  
dag <- dagitty('dag {  
"Education level" [pos="0.327,0.589"]  
"Hours Worked per week" [pos="0.387,-0.705"]  
"Income (Over/Under 50K)" [outcome,pos="-0.391,-0.156"]  
Age [pos="-1.199,-0.698"]  
Gender [exposure,pos="-1.995,0.058"]  
Race [exposure,pos="1.386,0.065"]  
Relationship [pos="-1.019,0.592"]  
"Education level" -> "Income (Over/Under 50K)"  
"Education level" -> Gender  
"Education level" -> Race  
"Hours Worked per week" -> "Income (Over/Under 50K)"  
"Hours Worked per week" -> Gender  
"Hours Worked per week" -> Race  
Age -> "Income (Over/Under 50K)"  
Age -> Gender  
Age -> Race  
Gender -> "Income (Over/Under 50K)"  
Race -> "Income (Over/Under 50K)"  
Relationship -> "Income (Over/Under 50K)"  
Relationship -> Gender  
Relationship -> Race  
}')  
  
plot(dag)



#Properly Label coded numbers for Table 1  
Age\_Data$age2 <- recode\_factor(Age\_Data$age2,  
 "1" = "17-30",  
 "2" = "31-40",  
 "3" = "41-50",  
 "4" = "51-60",  
 "5" = "61-70",  
 "6" = "71+")

### Table 1  
  
library(table1)

##   
## Attaching package: 'table1'

## The following objects are masked from 'package:base':  
##   
## units, units<-

label(Age\_Data$age2) <- "Age"  
label(Age\_Data$race) <- "Race"  
label(Age\_Data$educational\_num) <- "Level of Education"  
label(Age\_Data$gender) <- "Gender"  
label(Age\_Data$hours\_per\_week) <- "Hours per Week Worked"  
label(Age\_Data$income) <- "Income"  
label(Age\_Data$relationship) <- "Relationship"  
  
Epi\_Table <- table1::table1(  
 ~ gender + race + age2 + educational\_num + hours\_per\_week + hours\_per\_week + relationship |  
 income,  
 data = Age\_Data,  
 overall = c(left = "Total"),  
 topclass = "Rtable1-times",  
 caption = "Table1:Baseline Characteristics of Wealth, 1994"  
)  
Epi\_Table

## Get nicer `table1` .docx output by simply installing the `flextable` package

## Total <=50K >50K  
## 1 (N=45222) (N=34014) (N=11208)  
## 2 Gender   
## 3 Female 14695 (32.5%) 13026 (38.3%) 1669 (14.9%)  
## 4 Male 30527 (67.5%) 20988 (61.7%) 9539 (85.1%)  
## 5 Race   
## 6 Amer-Indian-Eskimo 435 (1.0%) 382 (1.1%) 53 (0.5%)  
## 7 Asian-Pac-Islander 1303 (2.9%) 934 (2.7%) 369 (3.3%)  
## 8 Black 4228 (9.3%) 3694 (10.9%) 534 (4.8%)  
## 9 Other 353 (0.8%) 308 (0.9%) 45 (0.4%)  
## 10 White 38903 (86.0%) 28696 (84.4%) 10207 (91.1%)  
## 11 Age   
## 12 17-30 14260 (31.5%) 13296 (39.1%) 964 (8.6%)  
## 13 31-40 12291 (27.2%) 8785 (25.8%) 3506 (31.3%)  
## 14 41-50 9990 (22.1%) 6131 (18.0%) 3859 (34.4%)  
## 15 51-60 5833 (12.9%) 3664 (10.8%) 2169 (19.4%)  
## 16 61-70 2212 (4.9%) 1628 (4.8%) 584 (5.2%)  
## 17 71+ 636 (1.4%) 510 (1.5%) 126 (1.1%)  
## 18 Level of Education   
## 19 Mean (SD) 10.1 (2.55) 9.63 (2.42) 11.6 (2.37)  
## 20 Median [Min, Max] 10.0 [1.00, 16.0] 9.00 [1.00, 16.0] 12.0 [1.00, 16.0]  
## 21 Hours per Week Worked   
## 22 Mean (SD) 40.9 (12.0) 39.4 (12.0) 45.7 (10.8)  
## 23 Median [Min, Max] 40.0 [1.00, 99.0] 40.0 [1.00, 99.0] 40.0 [1.00, 99.0]  
## 24 Relationship   
## 25 Husband 18666 (41.3%) 10159 (29.9%) 8507 (75.9%)  
## 26 Not-in-family 11702 (25.9%) 10474 (30.8%) 1228 (11.0%)  
## 27 Other-relative 1349 (3.0%) 1299 (3.8%) 50 (0.4%)  
## 28 Own-child 6626 (14.7%) 6521 (19.2%) 105 (0.9%)  
## 29 Unmarried 4788 (10.6%) 4486 (13.2%) 302 (2.7%)  
## 30 Wife 2091 (4.6%) 1075 (3.2%) 1016 (9.1%)

###Regression Model   
  
#Need to recode salary   
  
Age\_Data <- Age\_Data %>%  
 mutate(coded\_income = case\_when(  
 income == "<=50K" ~ 0,  
 income == ">50K" ~ 1  
 ))  
  
View (Age\_Data)

#Fit a model (unadjusted + adjusted) \*logistic regression  
Model1 <- glm(coded\_income ~ gender, data = Age\_Data)  
summary(Model1)

##   
## Call:  
## glm(formula = coded\_income ~ gender, data = Age\_Data)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.113576 0.003478 32.66 <2e-16 \*\*\*  
## genderMale 0.198901 0.004233 46.99 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.177747)  
##   
## Null deviance: 8430.2 on 45221 degrees of freedom  
## Residual deviance: 8037.7 on 45220 degrees of freedom  
## AIC: 50222  
##   
## Number of Fisher Scoring iterations: 2

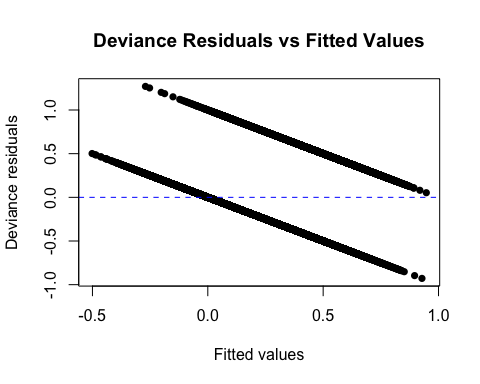
Model2 <- glm(coded\_income ~ race, data = Age\_Data)  
summary(Model2)

##   
## Call:  
## glm(formula = coded\_income ~ race, data = Age\_Data)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.121839 0.020599 5.915 3.34e-09 \*\*\*  
## raceAsian-Pac-Islander 0.161354 0.023790 6.782 1.20e-11 \*\*\*  
## raceBlack 0.004462 0.021632 0.206 0.837   
## raceOther 0.005640 0.030776 0.183 0.855   
## raceWhite 0.140531 0.020714 6.785 1.18e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1845732)  
##   
## Null deviance: 8430.2 on 45221 degrees of freedom  
## Residual deviance: 8345.8 on 45217 degrees of freedom  
## AIC: 51929  
##   
## Number of Fisher Scoring iterations: 2

model3 <-  
 glm(coded\_income ~ gender + age + race + educational\_num + hours\_per\_week + relationship,  
 data = Age\_Data)  
summary(model3)

##   
## Call:  
## glm(formula = coded\_income ~ gender + age + race + educational\_num +   
## hours\_per\_week + relationship, data = Age\_Data)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.4539202 0.0210319 -21.582 < 2e-16 \*\*\*  
## genderMale 0.0460702 0.0047833 9.631 < 2e-16 \*\*\*  
## age 0.0036744 0.0001441 25.501 < 2e-16 \*\*\*  
## raceAsian-Pac-Islander 0.0435335 0.0198936 2.188 0.028651 \*   
## raceBlack 0.0317947 0.0180584 1.761 0.078303 .   
## raceOther 0.0471798 0.0256912 1.836 0.066302 .   
## raceWhite 0.0572605 0.0173040 3.309 0.000937 \*\*\*  
## educational\_num 0.0485343 0.0006780 71.590 < 2e-16 \*\*\*  
## hours\_per\_week 0.0033192 0.0001508 22.007 < 2e-16 \*\*\*  
## relationshipNot-in-family -0.2995973 0.0048583 -61.667 < 2e-16 \*\*\*  
## relationshipOther-relative -0.2609424 0.0105677 -24.692 < 2e-16 \*\*\*  
## relationshipOwn-child -0.2751731 0.0063933 -43.041 < 2e-16 \*\*\*  
## relationshipUnmarried -0.2925494 0.0069569 -42.052 < 2e-16 \*\*\*  
## relationshipWife 0.1038042 0.0095717 10.845 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1284701)  
##   
## Null deviance: 8430.2 on 45221 degrees of freedom  
## Residual deviance: 5807.9 on 45208 degrees of freedom  
## AIC: 35552  
##   
## Number of Fisher Scoring iterations: 2

dev\_residuals <- residuals(model3, type = "deviance")  
  
# Plot deviance residuals against fitted values  
plot(fitted(model3), dev\_residuals,   
 xlab = "Fitted values", ylab = "Deviance residuals",  
 main = "Deviance Residuals vs Fitted Values",  
 pch = 16, col = "black")  
  
# Add a horizontal line at y = 0  
abline(h = 0, col = "blue", lty = 2)



library(ResourceSelection)

## ResourceSelection 0.3-6 2023-06-27

hoslem.test(model3$y, fitted(model3))

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: model3$y, fitted(model3)  
## X-squared = 108.97, df = 8, p-value < 2.2e-16

# Subset the data for white males aged 41-50  
WMale <- subset(Age\_Data, race == "White" & gender == "Male" & age2 == "41-50"& educational\_num == "12" & hours\_per\_week == "40")   
summary (WMale)

## age race educational\_num gender   
## Min. :41.00 Length:103 Min. :12 Length:103   
## 1st Qu.:43.00 Class :character 1st Qu.:12 Class :character   
## Median :45.00 Mode :character Median :12 Mode :character   
## Mean :44.78 Mean :12   
## 3rd Qu.:46.00 3rd Qu.:12   
## Max. :50.00 Max. :12   
## hours\_per\_week income relationship age2   
## Min. :40 Length:103 Length:103 17-30: 0   
## 1st Qu.:40 Class :character Class :character 31-40: 0   
## Median :40 Mode :character Mode :character 41-50:103   
## Mean :40 51-60: 0   
## 3rd Qu.:40 61-70: 0   
## Max. :40 71+ : 0   
## coded\_income   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.3981   
## 3rd Qu.:1.0000   
## Max. :1.0000

BFemale <- subset(Age\_Data, race == "Black" & gender == "Female" & age2 == "41-50" & educational\_num == "10" & hours\_per\_week == "40")  
summary (BFemale)

## age race educational\_num gender   
## Min. :41.00 Length:64 Min. :10 Length:64   
## 1st Qu.:42.00 Class :character 1st Qu.:10 Class :character   
## Median :44.00 Mode :character Median :10 Mode :character   
## Mean :44.77 Mean :10   
## 3rd Qu.:47.00 3rd Qu.:10   
## Max. :50.00 Max. :10   
## hours\_per\_week income relationship age2   
## Min. :40 Length:64 Length:64 17-30: 0   
## 1st Qu.:40 Class :character Class :character 31-40: 0   
## Median :40 Mode :character Mode :character 41-50:64   
## Mean :40 51-60: 0   
## 3rd Qu.:40 61-70: 0   
## Max. :40 71+ : 0   
## coded\_income   
## Min. :0.00000   
## 1st Qu.:0.00000   
## Median :0.00000   
## Mean :0.04688   
## 3rd Qu.:0.00000   
## Max. :1.00000

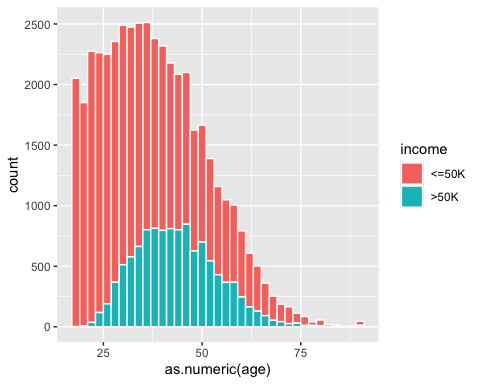
denom\_fit <- glm(coded\_income ~ gender + age + race + educational\_num + hours\_per\_week + relationship,  
 data = Age\_Data, family = binomial)  
  
num\_fit <- glm(coded\_income ~ 1, data = Age\_Data, family = binomial)  
  
pdenom <- predict(denom\_fit, type = "response")  
pnum <- predict(num\_fit, type = "response")  
  
Age\_Data$iptw <- ifelse(Age\_Data$coded\_income == 0, ((1 - pnum) / (1 - pdenom)), (pnum / pdenom))  
Age\_Data

## # A tibble: 45,222 × 10  
## age race educational\_num gender hours\_per\_week income relationship age2   
## <dbl> <chr> <dbl> <chr> <dbl> <chr> <chr> <fct>  
## 1 25 Black 7 Male 40 <=50K Own-child 17-30  
## 2 38 White 9 Male 50 <=50K Husband 31-40  
## 3 28 White 12 Male 40 >50K Husband 17-30  
## 4 44 Black 10 Male 40 >50K Husband 41-50  
## 5 34 White 6 Male 30 <=50K Not-in-family 31-40  
## 6 63 White 15 Male 32 >50K Husband 61-70  
## 7 24 White 10 Female 40 <=50K Unmarried 17-30  
## 8 55 White 4 Male 10 <=50K Husband 51-60  
## 9 65 White 9 Male 40 >50K Husband 61-70  
## 10 36 White 13 Male 40 <=50K Husband 31-40  
## # ℹ 45,212 more rows  
## # ℹ 2 more variables: coded\_income <dbl>, iptw <dbl>

### 

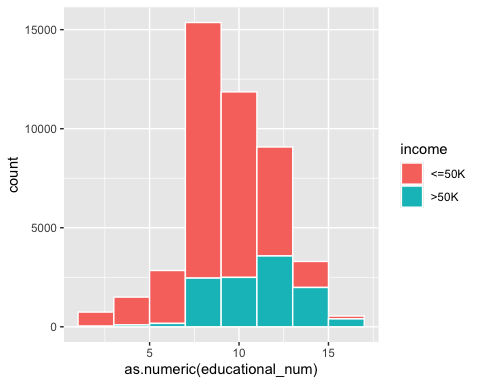
## Visualization

library(ggplot2)  
ggplot(Age\_Data) + aes(x=as.numeric(age), group=income, fill=income) +   
 geom\_histogram(binwidth=2, color='white')



IncomevRace <- data.frame(table(Age\_Data$income, Age\_Data$race, Age\_Data$gender, Age\_Data$educational\_num))  
names(IncomevRace) <- c('income', 'race', 'gender', 'education','count')  
IncomevRace

ggplot(Age\_Data) + aes(x=as.numeric(educational\_num), group=income, fill=income) +   
 geom\_histogram(binwidth=2, color='white')



# Recode "race" variable  
Age\_Data <- Age\_Data %>%  
 mutate(gender2 = case\_when(  
 gender == "Male" ~ 1,  
 gender == "Female" ~ 2  
 ))