Exam#2

Thakur Prasad Ghimire

11/18/2021

1. Hypothesis test Null hypothesis: people with various educational qualifications who are full or parttime have different fraction offered health insurance.
2. (20 points) Using a subset of the data (so you can do this question without R), the following coefficients are estimated in a logit model where employer offering health insurance depends on , , , and interaction of with the age terms.

|  |  |  |
| --- | --- | --- |
| Variable | Coeff | Std Error |
|  | -0.019 | 0.0018 |
|  | 0.00002 | 0.000002 |
|  | -0.470 | 0.1206 |
|  | 0.0082 | 0.0026 |
|  | -0.00001 | 0.000003 |
|  | 1.84 | 0.09 |

1. What is the predicted probability that a 30-year-old person (not a female) will have a job with benefits?

Log(p/((1-p) ))=-0.019*30+0.00002*(30)2-0.470*0+0.0082*30*1-0.00001*(30)2\*1+1.84=-1.993 p=0.0233

1. What is the predicted probability that a 30-year-old female person will have a job with benefits?

Log(p/((1-p) ))=-0.019\*30+0.00002\*(30)^2-0.470\*1+0.082\*1-0.00001\*1+1.84=2.002  
p=0.9772

1. Describe and explain the impact of the squared age terms in the model, for male and female.

Explore likelihood of people working at a job that offers health insurance.

#3

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.1.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(stats)  
library(dplyr)  
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(corrplot)

## corrplot 0.92 loaded

load("C:/Homework EcoB2000/Exam#2/NHIS2020\_data (2).RData")

Choose a subgroup of the sample to consider and provide summary statistics of that subgroup. #Choose the black race sub group

data\_NHIS\_subgrp = data\_NHIS[data\_NHIS$RACEA=="Black",]  
summary(data\_NHIS\_subgrp)

## REGION URBRRL AGE SEX   
## Northeast: 614 Large central metro:1574 Min. : 0.00 Male :1639   
## Midwest : 540 Large fringe metro : 855 1st Qu.: 28.00 Female :2275   
## South :2377 Med or small metro :1095 Median : 47.00 Refused : 1   
## West : 385 nonmetropolitan : 392 Mean : 48.07 dont know: 1   
## 3rd Qu.: 63.00   
## Max. :999.00   
##   
## SEXORIEN MARST FAMSIZE   
## straight :3032 never married:1201 Min. :1.000   
## NIU : 607 Married : 789 1st Qu.:1.000   
## NA : 111 Divorced : 614 Median :2.000   
## Lesbian or gay: 61 NIU : 607 Mean :2.368   
## dont know : 35 Widowed : 308 3rd Qu.:3.000   
## bisexual : 33 unknown : 180 Max. :6.000   
## (Other) : 37 (Other) : 217 NA's :7   
## RACEA HISPETH YRSINUS   
## Black :3916 Not Hispanic :3758 NIU :3474   
## white : 0 Mexican : 31 Less than 1 year in US: 4   
## Aleut Alaskan : 0 Other Hispanic: 125 1-5 years in US : 40   
## American Indian: 0 NA : 2 5-10 years in US : 51   
## Asian : 0 10-15 yr in US : 49   
## Other : 0 15 or more yr in US : 288   
## (Other) : 0 NA : 10   
## CITIZEN EDUC EMPSTAT   
## No not US citizen: 125 HS diploma :836 NIU : 607   
## yes US citizen :3660 NIU :607 Employed :1667   
## refused : 4 some college :602 not employed:1477   
## NA : 125 bachelors :554 dont know : 165   
## dont know : 2 assoc deg academic:325   
## masters :295   
## (Other) :697   
## HOURSWRK EMPHI   
## Min. : 0.00 NIU :2255   
## 1st Qu.: 0.00 no workplace did not offer health insurance: 398   
## Median : 0.00 yes workplace offer health insurance :1252   
## Mean :16.76 refused : 5   
## 3rd Qu.:40.00 NA : 3   
## Max. :95.00 dont know : 3   
## NA's :29   
## EMPFT FAMTOTINC HEALTH HEIGHT   
## NIU :2250 Min. : 0 excellent: 996 Min. : 0.00   
## parttime : 215 1st Qu.: 21000 very good:1013 1st Qu.:63.00   
## fulltime :1428 Median : 40000 good :1206 Median :66.00   
## refused : 5 Mean : 56360 fair : 565 Mean :60.69   
## NA : 17 3rd Qu.: 75000 poor : 133 3rd Qu.:69.00   
## dont know: 1 Max. :250000 refused : 3 Max. :76.00   
## dont know: 0 NA's :329   
## WEIGHT BMICALC HINOTCOVE   
## Min. : 0.0 Min. :12.8 has health insurance coverage:3546   
## 1st Qu.:142.0 1st Qu.:24.6 no health insurance coverage : 352   
## Median :175.0 Median :28.3 dont know : 18   
## Mean :165.9 Mean :29.0   
## 3rd Qu.:204.0 3rd Qu.:32.8   
## Max. :298.0 Max. :51.7   
## NA's :416 NA's :747   
## ARMEDFORCES   
## NIU : 607   
## No never active duty :2811   
## active only for training : 49   
## yes ever served in armed forces: 281   
## refused : 3   
## NA : 164   
## dont know : 1

Explain why this subgroup is interesting: This subgroup is mostly located in the south region, and most of the black people have high school diploma as their highest education level, and most of them have health insurance.

Form a hypothesis test about an interesting variable, explore whether your chosen subgroup differs from the rest of sample. Please provide both a p-value for the hypothesis test and a confidence interval. Write a short paragraph explaining the test (carefully noting what is the null hypothesis) and explaining the results of that test. The interesting variable is HOURSWRK. Null hypothesis: Hours of work is not correlated with someones age.

Now we will test this hypothesis on both the sample and the rest of the data to see if it holds.

With a linear probability model (OLS), can you find relevant information to predict an interesting outcome? How good is the model? Discuss. Include hypothesis tests about individual or joint coefficient values. Include some predicted values. Lets predict the hours worked by a black person in USA using age, sex, and citizenship. The null hypothesis is that these predictors have little or no effect on the number of hours worked by a black person in the US.

``{r}

logit\_model = glm(HOURSWRK ~ AGE + SEX + CITIZEN,data = data\_NHIS\_subgrp)  
summary(logit\_model)

##   
## Call:  
## glm(formula = HOURSWRK ~ AGE + SEX + CITIZEN, data = data\_NHIS\_subgrp)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -26.16 -16.59 -16.55 21.97 78.99   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.208e+01 1.966e+00 11.231 < 2e-16 \*\*\*  
## AGE -5.865e-04 5.823e-03 -0.101 0.9198   
## SEXFemale -1.468e+00 6.983e-01 -2.101 0.0357 \*   
## SEXRefused -1.806e+01 2.145e+01 -0.842 0.3999   
## SEXdont know -5.043e+00 2.145e+01 -0.235 0.8141   
## CITIZENyes US citizen -4.021e+00 1.951e+00 -2.061 0.0394 \*   
## CITIZENrefused 5.595e+00 1.253e+01 0.447 0.6551   
## CITIZENNA -2.118e+01 2.718e+00 -7.794 8.31e-15 \*\*\*  
## CITIZENdont know 3.915e+00 1.528e+01 0.256 0.7979   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 459.6088)  
##   
## Null deviance: 1823384 on 3886 degrees of freedom  
## Residual deviance: 1782363 on 3878 degrees of freedom  
## (29 observations deleted due to missingness)  
## AIC: 34871  
##   
## Number of Fisher Scoring iterations: 2

The null hypothesis as mentioned earlier that Hours of work is not affected by these variables (age, sex, citizen). The logit model performs the same way as the OLS model because the same variables like female sex, US citizen are the ones that are significant here too.

#Correlation test on the sub group

cor.test(data\_NHIS\_subgrp$HOURSWRK,data\_NHIS\_subgrp$AGE)

##   
## Pearson's product-moment correlation  
##   
## data: data\_NHIS\_subgrp$HOURSWRK and data\_NHIS\_subgrp$AGE  
## t = -0.83673, df = 3885, p-value = 0.4028  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04484286 0.01802333  
## sample estimates:  
## cor   
## -0.01342303

The test shows that there was no correlation between age and hours of work for black people.

#Rest of the data

rest\_df = data\_NHIS[data\_NHIS$RACEA!="Black",]

#Correlation test on the rest of the data

cor.test(rest\_df$HOURSWRK,rest\_df$HOURSWRK)

##   
## Pearson's product-moment correlation  
##   
## data: rest\_df$HOURSWRK and rest\_df$HOURSWRK  
## t = Inf, df = 33224, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 1 1  
## sample estimates:  
## cor   
## 1

The test on the rest of the data showed that there was correlation between hours of work and age. Therefore the sub group differed with the rest of the data here.

* With a linear probability model (OLS), can you find relevant information to predict an interesting outcome? How good is the model? Discuss. Include hypothesis tests about individual or joint coefficient values. Include some predicted values.

Lets predict the hours worked by a black person in USA using age, sex, and citizenship. The null hypothesis is that these predictors have little or no effect on the number of hours worked by a black person in the US.

LPM = lm(HOURSWRK ~ AGE + SEX + CITIZEN,data = data\_NHIS\_subgrp)  
summary(LPM)

##   
## Call:  
## lm(formula = HOURSWRK ~ AGE + SEX + CITIZEN, data = data\_NHIS\_subgrp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.16 -16.59 -16.55 21.97 78.99   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.208e+01 1.966e+00 11.231 < 2e-16 \*\*\*  
## AGE -5.865e-04 5.823e-03 -0.101 0.9198   
## SEXFemale -1.468e+00 6.983e-01 -2.101 0.0357 \*   
## SEXRefused -1.806e+01 2.145e+01 -0.842 0.3999   
## SEXdont know -5.043e+00 2.145e+01 -0.235 0.8141   
## CITIZENyes US citizen -4.021e+00 1.951e+00 -2.061 0.0394 \*   
## CITIZENrefused 5.595e+00 1.253e+01 0.447 0.6551   
## CITIZENNA -2.118e+01 2.718e+00 -7.794 8.31e-15 \*\*\*  
## CITIZENdont know 3.915e+00 1.528e+01 0.256 0.7979   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21.44 on 3878 degrees of freedom  
## (29 observations deleted due to missingness)  
## Multiple R-squared: 0.0225, Adjusted R-squared: 0.02048   
## F-statistic: 11.16 on 8 and 3878 DF, p-value: 1.024e-15

logit\_model = glm(HOURSWRK ~ AGE + SEX + CITIZEN,data = data\_NHIS\_subgrp)  
summary(logit\_model)

##   
## Call:  
## glm(formula = HOURSWRK ~ AGE + SEX + CITIZEN, data = data\_NHIS\_subgrp)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -26.16 -16.59 -16.55 21.97 78.99   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.208e+01 1.966e+00 11.231 < 2e-16 \*\*\*  
## AGE -5.865e-04 5.823e-03 -0.101 0.9198   
## SEXFemale -1.468e+00 6.983e-01 -2.101 0.0357 \*   
## SEXRefused -1.806e+01 2.145e+01 -0.842 0.3999   
## SEXdont know -5.043e+00 2.145e+01 -0.235 0.8141   
## CITIZENyes US citizen -4.021e+00 1.951e+00 -2.061 0.0394 \*   
## CITIZENrefused 5.595e+00 1.253e+01 0.447 0.6551   
## CITIZENNA -2.118e+01 2.718e+00 -7.794 8.31e-15 \*\*\*  
## CITIZENdont know 3.915e+00 1.528e+01 0.256 0.7979   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 459.6088)  
##   
## Null deviance: 1823384 on 3886 degrees of freedom  
## Residual deviance: 1782363 on 3878 degrees of freedom  
## (29 observations deleted due to missingness)  
## AIC: 34871  
##   
## Number of Fisher Scoring iterations: 2

The null hypothesis as mentioned earlier that Hours of work is not affected by these variables (age, sex, citizen). The logit model performs the same way as the OLS model because the same variables like female sex, US citizen are the ones that are significant here too.

Thakur Prasad ghimire City College of New york