PS6_12265092

12265092

23/02/2022

```r
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```r
#library(estimatr)
library(Rcpp)
library(readxl)
library(haven)
library(boot)
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.1.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(margins)
```

```
## Warning: package 'margins' was built under R version 4.1.2
```

```r
library(mfx)
```

```
## Warning: package 'mfx' was built under R version 4.1.2

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 4.1.2

## Loading required package: MASS

##
## Attaching package: 'MASS'

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

## Loading required package: betareg

## Warning: package 'betareg' was built under R version 4.1.2
```

```r
data <- read_dta("jtrain2.dta")
```

```r
summary(data)
```

```
##      train             age             educ          black
##  Min.   :0.0000   Min.   :17.00   Min.   : 3.0   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:20.00   1st Qu.: 9.0   1st Qu.:1.0000
##  Median :0.0000   Median :24.00   Median :10.0   Median :1.0000
##  Mean   :0.4157   Mean   :25.37   Mean   :10.2   Mean   :0.8337
##  3rd Qu.:1.0000   3rd Qu.:28.00   3rd Qu.:11.0   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :55.00   Max.   :16.0   Max.   :1.0000
##       hisp            married          nodegree        mosinex
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.000   Min.   : 5.00
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:14.00
##  Median :0.00000   Median :0.0000   Median :1.000   Median :21.00
##  Mean   :0.08764   Mean   :0.1685   Mean   :0.782   Mean   :18.12
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:23.00
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.000   Max.   :24.00
##       re74              re75             re78            unem74
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:0.0000
##  Median : 0.0000   Median : 0.000   Median : 3.702   Median :1.0000
##  Mean   : 2.1023   Mean   : 1.377   Mean   : 5.301   Mean   :0.7326
##  3rd Qu.: 0.8244   3rd Qu.: 1.221   3rd Qu.: 8.125   3rd Qu.:1.0000
##  Max.   :39.5707   Max.   :25.142   Max.   :60.308   Max.   :1.0000
##      unem75           unem78            lre74             lre75
##  Min.   :0.0000   Min.   :0.0000   Min.   :-0.8093   Min.   :-2.5991
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median :1.0000   Median :0.0000   Median : 0.0000   Median : 0.0000
```

```
##  Mean   :0.6494   Mean   :0.3079   Mean   : 0.4198   Mean   : 0.2771
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 0.0000   3rd Qu.: 0.1995
##  Max.   :1.0000   Max.   :1.0000   Max.   : 3.6781   Max.   : 3.2245
##      lre78            agesq          mostrn
##  Min.   :-3.107   Min.   : 289   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 400   1st Qu.: 0.000
##  Median : 1.309   Median : 576   Median : 0.000
##  Mean   : 1.136   Mean   : 694   Mean   : 7.688
##  3rd Qu.: 2.095   3rd Qu.: 784   3rd Qu.:15.000
##  Max.   : 4.099   Max.   :3025   Max.   :24.000
```

```r
#a

#summary(data)

summary(factor(data$train, labels = c(0,1)))
```

```
##   0   1
## 260 185
```

```r
#As we can see here, 185 men in the sample participated
#in the job training program

summary(data$mostrn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   7.688  15.000  24.000
```

```r
##As we can see here , the hugherst number of months a man actually participated
## in the program is 24.
```

```r
#b

data$train <- as.numeric(data$train)

reg_b = lm(formula = train ~ unem74 + unem75 + age + educ +
             black + hisp + married, data = data)


summary(reg_b)
```

```
##
## Call:
## lm(formula = train ~ unem74 + unem75 + age + educ + black + hisp +
##     married, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6024 -0.4196 -0.3437  0.5537  0.7669
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.338022   0.189445   1.784   0.0751 .
## unem74       0.020880   0.077294   0.270   0.7872
## unem75      -0.095571   0.071902  -1.329   0.1845
## age          0.003206   0.003403   0.942   0.3467
## educ         0.012013   0.013342   0.900   0.3684
## black       -0.081666   0.087732  -0.931   0.3524
## hisp        -0.200017   0.116971  -1.710   0.0880 .
## married      0.037289   0.064404   0.579   0.5629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4917 on 437 degrees of freedom
## Multiple R-squared:  0.02238,    Adjusted R-squared:  0.006722
## F-statistic: 1.429 on 7 and 437 DF,  p-value: 0.1915
```

```
#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)  0.338022   0.189445   1.784   0.0751 .
#unem74       0.020880   0.077294   0.270   0.7872
#unem75      -0.095571   0.071902  -1.329   0.1845
#age          0.003206   0.003403   0.942   0.3467
#educ         0.012013   0.013342   0.900   0.3684
#black       -0.081666   0.087732  -0.931   0.3524
#hisp        -0.200017   0.116971  -1.710   0.0880 .
#married      0.037289   0.064404   0.579   0.5629

#Residual standard error: 0.4917 on 437 degrees of freedom
#Multiple R-squared:  0.02238
#Adjusted R-squared:  0.006722
#F-statistic: 1.429 on 7 and 437 DF
#p-value: 0.1915

##H0 = All slope coefficients =0

#We observe a fstatistic of 1.429 on DF k =7 and n-k-1 = 437
#Critical value of f distribution at 5% level in these conditions is 2.01
#As the observed statistic value is less than the critical value, we cannot
#reject the null hypothesis. Thus using OLS we cannot say with evidence that the
#independent variables in the regression are significant in explaining the
#dependent variable train.
```

```
#c

probit_c <- glm(formula = train ~ unem74 +
                  unem75 + age + educ + black + hisp + married,
                    family = binomial(link = "probit"), data = data)

summary(probit_c)
```

```
##
## Call:
## glm(formula = train ~ unem74 + unem75 + age + educ + black +
##     hisp + married, family = binomial(link = "probit"), data = data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3620  -1.0421  -0.9159   1.2702   1.6962
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.424107   0.489506  -0.866   0.3863
## unem74       0.053026   0.198834   0.267   0.7897
## unem75      -0.247725   0.184806  -1.340   0.1801
## age          0.008344   0.008780   0.950   0.3419
## educ         0.031443   0.034657   0.907   0.3643
## black       -0.206930   0.224614  -0.921   0.3569
## hisp        -0.539777   0.307947  -1.753   0.0796 .
## married      0.096625   0.165503   0.584   0.5593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 604.20  on 444  degrees of freedom
## Residual deviance: 594.02  on 437  degrees of freedom
## AIC: 610.02
##
## Number of Fisher Scoring iterations: 4
```

```r
probit_c_1 <- glm(formula = train ~ 1,
                    family = binomial(link = "probit"), data = data)


summary(probit_c_1)
```

```
##
## Call:
## glm(formula = train ~ 1, family = binomial(link = "probit"),
##     data = data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.037  -1.037  -1.037   1.325   1.325
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.2128     0.0599  -3.553 0.000381 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 604.2  on 444  degrees of freedom
## Residual deviance: 604.2  on 444  degrees of freedom
## AIC: 606.2
##
## Number of Fisher Scoring iterations: 3
```

```r
lrtest(probit_c, probit_c_1)
```

```
## Likelihood ratio test
##
## Model 1: train ~ unem74 + unem75 + age + educ + black + hisp + married
## Model 2: train ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -297.01
## 2   1 -302.10 -7 10.182     0.1785
```

```r
#HO = Coefficients of all independent variables in regression = 0
# We observe a Chisquare statistic value of 10.18
# The critical value of chisquare distribution
#at 5% at the same condition is of value = 14.07
#We see that the observed statistic is less than the critical value
#Using above, using Probit , we cannot say that the independent variables are
#significant in explaining the dependent variable.


c_psuedoR2 <- 1 - (probit_c$deviance) / (probit_c$null.deviance)


#c_psuedoR2 <- 1 - logLik(probit_c)[1]/logLik(probit_c)[1]

c_psuedoR2
```

```
## [1] 0.01685271
```

```r
#d

#In b, we observed that the independent variables are not statistically
#significant in explaining the dependent variable
#Similarly, we observed in c.
#Hence, using results from b and c, participation in job training can be
#treated as exogenous for explaining the 1978 unemployment status.
```

```r
#e

reg_e <- lm(formula = unem78 ~ train, data = data)

summary(reg_e)
```

```
##
## Call:
## lm(formula = unem78 ~ train, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3538 -0.3538 -0.2432  0.6462  0.7568
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.35385    0.02849  12.419   <2e-16 ***
## train       -0.11060    0.04419  -2.503   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4594 on 443 degrees of freedom
## Multiple R-squared:  0.01394,    Adjusted R-squared:  0.01172
## F-statistic: 6.265 on 1 and 443 DF,  p-value: 0.01267
```

```
#Intercept = 0.35385
#Coefficient estimate of variable train = -0.1106

#Equation form
# unem78^ = 0.35385 -0.1106*train^


#Interpretaion of coefficient
#Participation in training has reduced the probability
#of being employment in 1978 by 11%

#F-statistic: 6.265 on 1 and 443 DF,  p-value: 0.01267
#H0 = Coefficient of train = 0
#At 5% significance, the Observed p-value is less than 0.05, so we can reject
#the null hypothesis
#We cannot reject the null hypothesis that the participation in training
#is significant in explaining the unemployment in 1978
```

```
#f

probit_f <- glm(formula = unem78 ~ train,
                family = binomial(link = "probit"), data = data)

summary(probit_f)
```

```
##
## Call:
## glm(formula = unem78 ~ train, family = binomial(link = "probit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9346  -0.9346  -0.7466   1.4414   1.6815
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.37496    0.07975  -4.702 2.58e-06 ***
## train       -0.32095    0.12848  -2.498   0.0125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 549.47  on 444  degrees of freedom
## Residual deviance: 543.17  on 443  degrees of freedom
## AIC: 547.17
##
## Number of Fisher Scoring iterations: 4
```

```r
#Observed p value = 0.0125
#At 5% level, observed p-value is less than 0.05
#Hence we can reject the null hypothesis that train variable is insignificant
#And we cannot reject the alternate hypothesis that participation in training
#is significant in explaining the unemployment in 1978



probit_f_1 <- glm(formula = unem78 ~ 1,
                family = binomial(link = "probit"), data = data)

summary(probit_f_1)
```

```
##
## Call:
## glm(formula = unem78 ~ 1, family = binomial(link = "probit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8579  -0.8579  -0.8579   1.5350   1.5350
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.50191    0.06221  -8.067 7.18e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 549.47  on 444  degrees of freedom
## Residual deviance: 549.47  on 444  degrees of freedom
## AIC: 551.47
##
## Number of Fisher Scoring iterations: 4
```

```r
lrtest(probit_f, probit_f_1)
```

```
## Likelihood ratio test
##
## Model 1: unem78 ~ train
## Model 2: unem78 ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   2 -271.58
## 2   1 -274.74 -1 6.3043    0.01204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#  #Df  LogLik Df  Chisq Pr(>Chisq)
#1   2 -271.58
#2   1 -274.74 -1 6.3043    0.01204 *


#H0 = Coefficient of train = 0
#Observed chi square statistic is 6.3043
#At df =1, 5% level, the critical value of chisquare distribution is 3.84
#As the observed statistic is less than the critical value, we can reject the
#null hypothesis that the variable train is insignificant
#And we cannot reject the alternate hypothesis that the variable tain is
#significant in explaining unem78 using the model

f_psuedoR2 <- 1 - (probit_f$deviance) / (probit_f$null.deviance)

f_psuedoR2
```

```
## [1] 0.01147336
```

```r
#It doesnt make sense to compare both the coefficients.
#Reason is: In linear model the interpretaion of coefficients is direct.
#In lm the coefficients are directly the marginal effect
#In probit model, marginal effect is not the coefficient
#To interpret the coefficient of probit model, we need to calculate
#the marginal effect of the regressor on the outcome while holding all other
#variables constant.
```

```r
#g

fitted_f <- predict(probit_f, type = "response")

summary(fitted_f)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2432  0.2432  0.3538  0.3079  0.3538  0.3538
```

```r
head(fitted_f)
```

```
##         1         2         3         4         5         6
## 0.2432432 0.2432432 0.2432432 0.2432432 0.2432432 0.2432432
```

```r
fitted_e <- predict(reg_e, type = "response")

summary(fitted_e)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2432  0.2432  0.3538  0.3079  0.3538  0.3538
```

```r
head(fitted_e)
```

```
##         1         2         3         4         5         6
## 0.2432432 0.2432432 0.2432432 0.2432432 0.2432432 0.2432432
```

```
#They are identical

#As we have only one binary variable train and only one outcome variable unem78
#which is also binary, we will get a perfect fit when we perform regression.
#This doesnt change with different regression functions. Thus they are identical.
#This simply means that the estimated probability of unem78 is actually the
#observed probability of the independent variable train.



#We prefer Probit or Logit models when evaluating binary independent variable.
#This is because using linear model to explain a binary variable will result
#in heteroskedasticity. Also the predicted values do not follow the
#boundary conditions of binary and usually go beyond 0 and beyond 1

#In Probit model, the resulted model is distributed normally. Thus the
#depended variable or the outcome predicted will be either 0 or 1. This
#satisfies the criteria of the dependant variable to be a binary. For this case,
#we can use the Probit model, but usually Logit has more benefits, below.


#Logit model assumption is that the model is logistically distributed.
#Thus the predicted outcome is again a 0 or 1, happens or doesnt happen. THis
#satisfies initial criteria of outcome variable to be a binary variable.
#While the marginal effects produced by both the Probit and Logit models
#are same, the coefficients differ by a factor of 1.6. The benefit that the
#Logit model has is that the transformations possible in the Logit model
#makes it easy to interpreting them.
```

```
#h


#variables from b
#unem74 + unem75 + age + educ + black + hisp + married

reg2_e <- lm(formula = unem78 ~ train + unem74 + unem75 + age +
               educ + black + hisp + married,
                family = binomial(link = "probit"), data = data)
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

```
summary(reg2_e)
```

```
##
## Call:
## lm(formula = unem78 ~ train + unem74 + unem75 + age + educ +
##     black + hisp + married, data = data, family = binomial(link = "probit"))
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.4106 -0.3546 -0.2428  0.5908  0.9709 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  1.632e-01  1.761e-01   0.927   0.3546  
## train       -1.117e-01  4.431e-02  -2.521   0.0121 *
## unem74       3.869e-02  7.160e-02   0.540   0.5892  
## unem75       1.596e-02  6.673e-02   0.239   0.8111  
## age          4.332e-05  3.155e-03   0.014   0.9891  
## educ         1.442e-04  1.237e-02   0.012   0.9907  
## black        1.888e-01  8.134e-02   2.322   0.0207 *
## hisp        -3.770e-02  1.087e-01  -0.347   0.7289  
## married     -2.544e-02  5.967e-02  -0.426   0.6701  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4554 on 436 degrees of freedom
## Multiple R-squared:  0.0462, Adjusted R-squared:  0.0287 
## F-statistic:  2.64 on 8 and 436 DF,  p-value: 0.007796
```

```
#Residual standard error: 0.4554 on 436 degrees of freedom
#Multiple R-squared:  0.0462,   Adjusted R-squared:  0.0287 
#F-statistic:  2.64 on 8 and 436 DF,  p-value: 0.007796

fitted2_e <- predict(reg2_e, type = "response")


summary(fitted2_e)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## -0.009249  0.243112  0.298227  0.307865  0.408703  0.410595
```

```
head(fitted2_e)
```

```
##          1          2          3          4          5          6 
## 0.27271826 0.07068344 0.29799657 0.29772238 0.29754956 0.29721730
```

```
probit_h <-  glm(formula = unem78 ~ train + unem74 + unem75 + age 
                  + educ + black + hisp + married, 
                      family = binomial(link = "probit"), data = data)

fitted_h <- predict(probit_h, type = "response")


summary(fitted_h)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.05507 0.23743 0.29585 0.30771 0.41491 0.43035
```

```r
head(fitted_h)
```

```
##          1          2          3          4          5          6
## 0.26857563 0.08942665 0.29254246 0.29249596 0.29584637 0.29263505
```

```r
#The fitted values are almost identical in the same varying by a small factor.

correlation_fitted <- cor(fitted2_e, fitted_h)

correlation_fitted
```

```
## [1] 0.9932445
```

```r
#Correlation between fitted values of linear model and the fitted values of
#the probit model is ~0.9932445
```

```r
#i

#summary(margins(probit2_f, variables = "train"))
#summary(effects(probit2_f, effect = "marginal",
 #    marg.type = aveacr, varlist = train) )
#summary(effects(probit2_f, effect = "discrete",
 #              marg.type = atmean, varlist = train)  )



probit_APE <- probitmfx(unem78 ~ train + unem74 + unem75+ age + educ + black
                        +hisp + married, data = data)

APE <-  probit_APE$mfxest[1]

APE
```

```
## [1] -0.1143574
```

```r
#Observed APE is -0.1143
#OLS estimate from part h = -0.117

#The estimate APE is almost same as the OLS estimate found in part h
#Interpretation
#Probit:Average partial effect on unemployment in 1978 is -0.1143 times of train
#Linear: Effect on unemployment in 1978 is -0.117 times of train
```