STATS 506 HW3

Minxuan Chen

2023-10-01

Table of contents

Github repo: https://github.com/PKUniiiiice/STATS_506

1 Problem 1 Vision

```

. do "K:\STATS_506\STATA\stata_hw3.Do"

. //load data
. cd "K:\STATS_506\STATA\"
K:\STATS_506\STATA

.
. //Part a ---------------------------------------------------------------------
> -----------------------
.
. //import first
. import sasxport5 VIX_D.XPT, clear

. //save as dta
. save "K:\STATS_506\STATA\VIX_D.dta"
file K:\STATS_506\STATA\VIX_D.dta saved

. //import second
. import sasxport5 DEMO_D.XPT, clear

. // Merge the second dataset using the SEQN variable
. merge 1:1 seqn using "K:\STATS_506\STATA\VIX_D.dta"

    Result                      Number of obs
    -----------------------------------------
    Not matched                         3,368
        from master                     3,368  (_merge==1)
        from using                          0  (_merge==2)

    Matched                             6,980  (_merge==3)
    -----------------------------------------


.
. // Keep only the matched records
. keep if _merge == 3
(3,368 observations deleted)

.
. //total sample size
. di _N
6980

.
. //End of Part a ----------------------------------------------------------------
> ------------------------------
.
. //Part b -----------------------------------------------------------------------
> ------------------------------
. //We use the variable 'VIX220 - Glasses/contact lenses worn for distance' (viq
> 220)
. //and  'RIDAGEYR - Age at Screening Adjudicated - Recode' (ridageyr)
.
. egen age_interval = cut(ridageyr), at(0(10)90) label

. table age_interval viq220, missing statistic(percent, across(viq220)) statisti
> c(frequency)

-----------------------------------------------------------------
              |      Glasses/contact lenses worn for distance
              |         1          2         9          .      Total
--------------+--------------------------------------------------
```

```
age_interval  |
  10-         |
    Percent   |     30.36     64.25                5.39     100.00
    Frequency |       670     1,418                 119      2,207
  20-         |
    Percent   |     29.97     61.80     0.20      8.03     100.00
    Frequency |       306       631        2        82      1,021
  30-         |
    Percent   |     32.89     58.80                8.31     100.00
    Frequency |       269       481                  68        818
  40-         |
    Percent   |     35.09     59.75                5.15     100.00
    Frequency |       286       487                  42        815
  50-         |
    Percent   |     53.09     43.42                3.49     100.00
    Frequency |       335       274                  22        631
  60-         |
    Percent   |     59.30     36.01                4.69     100.00
    Frequency |       392       238                  31        661
  70-         |
    Percent   |     63.75     31.56                4.69     100.00
    Frequency |       299       148                  22        469
  80-         |
    Percent   |     58.10     28.77               13.13     100.00
    Frequency |       208       103                  47        358
  Total       |
    Percent   |     39.61     54.15     0.03      6.20     100.00
    Frequency |     2,765     3,780        2       433      6,980
-----------------------------------------------------------------

. quietly collect layout (age_interval) (viq220#result[percent] viq220#result[fr
> equency])

. //If only want to see first column
. //https://grodri.github.io/stata/tables https://www.stata.com/manuals/tables.p
> df
.
. //percent
. collect layout (age_interval) (viq220[1]#result[percent] viq220[.m]#result[fre
> quency])

```

```
Collection: Table
      Rows: age_interval
   Columns: viq220[1]#result[percent] viq220[.m]#result[frequency]
   Table 1: 10 x 2

----------------------------------------------------------
             |   Glasses/contact lenses worn for distance
             |                    1                  Total
             |              Percent              Frequency
-------------+--------------------------------------------
age_interval |
  10-        |                30.36                  2,207
  20-        |                29.97                  1,021
  30-        |                32.89                    818
  40-        |                35.09                    815
  50-        |                53.09                    631
  60-        |                59.30                    661
  70-        |                63.75                    469
  80-        |                58.10                    358
  Total      |                39.61                  6,980
----------------------------------------------------------

.
. //End of Part b --------------------------------------------------------------
> ------------------------------
.
.
. //Part c ---------------------------------------------------------------------
> ------------------------------
. //For age, we use ridageyr
. //For race, we use ridreth1
. //For gender, we use riagendr
. //For Poverty Income ratio, we use indfmpir
. //We first check how many missing values are in these variables
. misstable summarize ridageyr ridreth1 riagendr indfmpir viq220
                                                               Obs<.
                                                +-------------------------------
               |                                | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values        Min         Max
  -------------+--------------------------------+-------------------------------
      indfmpir |       342               6,638  |    435          0           5
```

```
        viq220 |       433                 6,547  |      3          1           9
  -----------------------------------------------------------------------------

.
. //It seems that the proportion of missing value is not large, about 10%, so we
>  choose to directly delete them
. drop if missing(indfmpir) | missing(viq220)
(731 observations deleted)

. misstable summarize ridageyr ridreth1 riagendr indfmpir viq220
(variables nonmissing or string)

.
. ///We treat viq220==1 as "Yes, wear", and all other values as "No, don't wear"
> //recode viq220
. recode viq220 (1=1) (else=0), generate(viq220_bin)
(3,594 differences between viq220 and viq220_bin)

.
. //ref https://www.stata.com/manuals/rlogistic.pdf
. //Note that race and gender shoule be categorical variables and age and PIR ar
> e continuous variables
.
.
. // Fit the first logistic regression model (age only)
. logistic viq220_bin ridageyr

Logistic regression                                     Number of obs =  6,249
                                                        LR chi2(1)    = 403.63
                                                        Prob > chi2   = 0.0000
Log likelihood = -4058.8462                             Pseudo R2     = 0.0474

------------------------------------------------------------------------------
  viq220_bin | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ridageyr |   1.024531   .0012702    19.55   0.000     1.022044    1.027023
       _cons |   .2923673    .015974   -22.51   0.000     .2626769    .3254136
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

.
```

```
. // Store the results
. eststo model1

.
. // Fit the second logistic regression model (age, race, gender)
. logistic viq220_bin ridageyr i.ridreth1 i.riagendr

Logistic regression                                     Number of obs =  6,249
                                                        LR chi2(6)    = 584.06
                                                        Prob > chi2   = 0.0000
Log likelihood = -3968.6291                             Pseudo R2     = 0.0685

------------------------------------------------------------------------------
  viq220_bin | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ridageyr |     1.0226   .0013241    17.26   0.000     1.020009    1.025199
             |
    ridreth1 |
          2  |   1.169508   .1959093     0.93   0.350     .8421995    1.624021
          3  |   1.895064   .1363291     8.89   0.000     1.645846    2.182019
          4  |   1.293781   .1015763     3.28   0.001     1.109257    1.509002
          5  |   1.885095   .2612655     4.57   0.000     1.436681    2.473465
             |
  2.riagendr |   1.650228   .0891912     9.27   0.000     1.484357    1.834634
       _cons |   .1650721   .0132324   -22.47   0.000     .1410718    .1931555
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

.
. eststo model2

.
. // Fit the third logistic regression model (age, race, gender, Poverty Income
> ratio)
. logistic viq220_bin ridageyr i.ridreth1 i.riagendr indfmpir

Logistic regression                                     Number of obs =  6,249
                                                        LR chi2(7)    = 625.24
                                                        Prob > chi2   = 0.0000
Log likelihood = -3948.0387                             Pseudo R2     = 0.0734

```

```
------------------------------------------------------------------------------
  viq220_bin | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ridageyr |    1.02245    .001324    17.15   0.000     1.019858    1.025048
             |
    ridreth1 |
          2  |   1.124663   .1892328     0.70   0.485     .8087261    1.564023
          3  |   1.652417    .124123     6.69   0.000     1.426201    1.914514
          4  |    1.23222   .0975979     2.64   0.008      1.05504    1.439155
          5  |    1.70633   .2391229     3.81   0.000     1.296513    2.245688
             |
  2.riagendr |   1.673821   .0908852     9.49   0.000     1.504841    1.861777
    indfmpir |    1.12011   .0198248     6.41   0.000      1.08192    1.159647
       _cons |   .1330474   .0116811   -22.97   0.000     .1120144    .1580298
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

.
. eststo model3

.
. // Create a table to display results using esttab
. // https://repec.org/bocode/e/estout/hlp_esttab.html
. esttab model1 model2 model3, ///
>         con ///
>         not ///
>     stats(N r2_p aic) ///
>     eform ///
>     varwidth(15) ///
>     title("Logistic Regression Results") ///
>     label

Logistic Regression Results
---------------------------------------------------------------
                         (1)             (2)             (3)   
                RECODE of ~c    RECODE of ~c    RECODE of ~c   
---------------------------------------------------------------
RECODE of viq~c
Age at Screen~R        1.025***        1.023***        1.022***
Race/Ethnicit~1                            1               1   
Race/Ethnicit~2                        1.170           1.125   
Race/Ethnicit~3                        1.895***        1.652***
```

```
Race/Ethnicit~4                        1.294**         1.232**
Race/Ethnicit~5                        1.885***        1.706***
Gender=1                                   1               1
Gender=2                               1.650***        1.674***
Family PIR                                             1.120***
Constant               0.292***        0.165***        0.133***
---------------------------------------------------------------
N                         6249            6249            6249
r2_p                    0.0474          0.0685          0.0734
aic                     8121.7          7951.3          7912.1
---------------------------------------------------------------
Exponentiated coefficients
* p<0.05, ** p<0.01, *** p<0.001

.
. //End of Part c --------------------------------------------------------------
> ------------------------------
.
. //Part d ---------------------------------------------------------------------
> ------------------------------
.
. //Note that in the output table, the odds ratio of Gender=2 is significant, th
> erefore,
. //the odds of men and women being wears of glasses/contact lenses for distance
>  vision differs.
.
. //We use chi-square test (Pearson's Chi-Squared Test of Independence)
. tabulate riagendr viq220_bin, chi2

           |   RECODE of viq220
           |   (Glasses/contact
           |    lenses worn for
           |       distance)
    Gender |         0          1 |     Total
-----------+----------------------+----------
         1 |     1,919      1,134 |     3,053
         2 |     1,675      1,521 |     3,196
-----------+----------------------+----------
     Total |     3,594      2,655 |     6,249

          Pearson chi2(1) =  69.7397   Pr = 0.000
```

```
.
. //From the result, p-value is 0.000, therefore we conclude that gender and wea
> ring or not
. //are not independent, in other words, the proportion of wearers of glasses/co
> ntact lenses for distance vision differs between men and women
.
. //End of Part d --------------------------------------------------------------
> ------------------------------
.
.
.
.
.
.
end of do-file

.
```