Soc 73994, Homework #2: Basics of Logistic Regression

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised September 10, 2024

All answers should be submiutted through Canvas. Be sure your response includes your name, the date, and a clear title, e.g. Homework # 2. If there is a huge amount of output for any analyses you run yourself, you may want to be selective in what you copy and paste into your assignment (but make sure you include enough so it is clear what commands you executed, e.g. you might show all the commands but only parts of the output).

This assignment has two parts. First, you will interpret the results from a hypothetical logistic regression analysis (put another way, the data are fake). You will then conduct and interpret a similar analysis using the data set of your choice. If you don't understand what a command is doing, check the help file and/or the class handouts. Some of the output is deleted (or the command is run quietly) so feel free to run the command yourself if you want to see what was left out.

1. It is September 2032. After his stunning and decisive upset victory over Donald Trump Jr. in the Indiana primary, Republican Presidential Nominee Ted Cruz now faces the daunting challenge of taking on his immensely popular opponent: Democratic nominee Pete Buttigieg, the former mayor of South Bend, Indiana, who rose to the top of the ticket when Vice-President Tim Walz opted to become coach of the Green Bay Packers rather than pursue the Presidency.

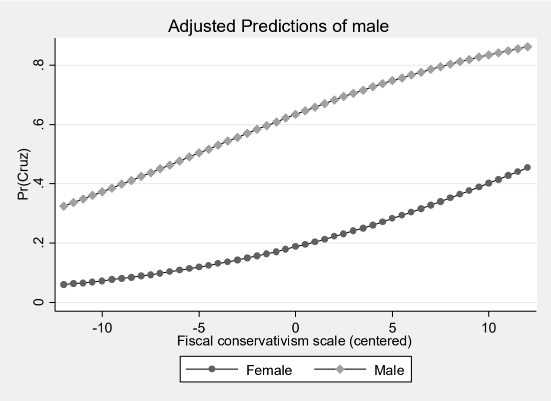
As Joe Biden's and Kamala Harris's Secretary of Transportation, Buttigieg has electrified rural voters with his successful efforts to build smart streets and roundabouts in every small town in America. Cruz, however, remains optimistic. He believes it is actually a very close race at the moment. Further, if he can identify which of his issues resonates most with the American people, he is confident he can win and provide the nation with the leadership it so desperately needs. His pollsters have therefore gathered the following information from over 4,000 likely voters:

Variable	*Description*
cruz	1 = supports Cruz, 0 = supports Buttigieg
male	1 = male, 0 = female (Cruz's people still don't appreciate that gender isn't simply binary)
socialcons	1 = respondent considers self conservative on social issues, 0 = not socially conservative
fiscalcons	Fiscal conservatism scale (continuous variable). The higher the score, the more fiscally conservative the respondent is. The scale has been centered to have a mean of zero. The scale potentially runs from -13 to +13, but the observed values are a little less extreme than that.
ses	Socio-Economic Status (ordinal scale). 1 = Lower Class, 2 = Middle Class, 3 = Upper Class. (Not used in the current analysis)

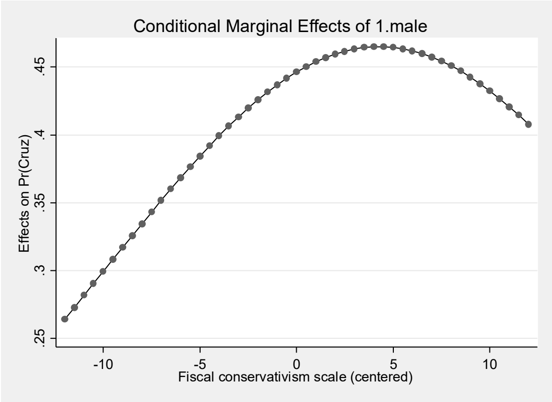
The study obtains the following results:

```
. version 13.1

. * HW 02 -- Ted Cruz problem
. use https://www3.nd.edu/~rwilliam/statafiles/cruz, clear

. * Descriptive analyses
. sum cruz male socialcons fiscalcons

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        cruz |      4,165    .3639856    .4812023          0          1
        male |      4,165    .5111645    .4999354          0          1
  socialcons |      4,165    .7097239    .4539442          0          1
  fiscalcons |      4,165   -1.71e-08    2.769073  -12.42706   11.16392

. * Estimate the lpm
. reg cruz i.male i.socialcons fiscalcons

      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(3, 4161)      =    317.63
       Model |  179.661361         3  59.8871204   Prob > F        =    0.0000
    Residual |  784.536478     4,161  .188545176   R-squared       =    0.1863
-------------+----------------------------------   Adj R-squared   =    0.1857
       Total |  964.197839     4,164  .231555677   Root MSE        =    .43422

--------------------------------------------------------------------------------------
                cruz |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------------+----------------------------------------------------------------
                male |
               Male  |   .4053313    .014197    28.55   0.000     .3774975    .4331651
                     |
          socialcons |
Social Conservative  |   .1692383    .015073    11.23   0.000     .1396871    .1987894
          fiscalcons |   .0197513   .0026037     7.59   0.000     .0146467     .024856
               _cons |   .0366822   .0144812     2.53   0.011     .0082914     .065073
--------------------------------------------------------------------------------------

. est store lpm

. * Estimate series of LRMs. Won't bother showing all the output though.
. logit cruz i.male, nolog

--- [Output deleted] ---

. est store lrm1

. logit cruz i.male i.socialcons, nolog

--- [Output deleted] ---

. est store lrm2
. logit cruz i.male i.socialcons fiscalcons, nolog

Logistic regression                             Number of obs     =      4,165
                                                LR chi2(3)        =     837.74
                                                Prob > chi2       =     0.0000
Log likelihood = -2312.0241                     Pseudo R2         =     0.1534

-------------------------------------------------------------------------------------
                cruz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+---------------------------------------------------------------
                male |
               Male  |   2.014584   .0816795    24.66   0.000     1.854495    2.174673
                     |
          socialcons |
Social Conservative  |   .9218104   .0846065    10.90   0.000     .7559847    1.087636
          fiscalcons |   .1068539   .0141381     7.56   0.000     .0791438    .1345641
               _cons |  -2.387975   .0931491   -25.64   0.000    -2.570544   -2.205406
-------------------------------------------------------------------------------------

. est store lrm3

. * Use lrtest commands to assess fit
. lrtest lrm1 lrm2, stats

Likelihood-ratio test                                 LR chi2(1)  =    159.97
(Assumption: lrm1 nested in lrm2)                     Prob > chi2 =    0.0000

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
        lrm1 |      4,165 -2730.894  -2421.107       2    4846.214   4858.883
        lrm2 |      4,165 -2730.894  -2341.121       3    4688.241   4707.245
-----------------------------------------------------------------------------
               Note: N=Obs used in calculating BIC; see [R] BIC note.

. lrtest lrm2 lrm3, stats

Likelihood-ratio test                                 LR chi2(1)  =     58.19
(Assumption: lrm2 nested in lrm3)                     Prob > chi2 =    0.0000

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
        lrm2 |      4,165 -2730.894  -2341.121       3    4688.241   4707.245
        lrm3 |      4,165 -2730.894  -2312.024       4    4632.048   4657.386
-----------------------------------------------------------------------------
               Note: N=Obs used in calculating BIC; see [R] BIC note.
```

```
. * Display all the models in a single table
. esttab lpm lrm1 lrm2 lrm3, nobase mtitles z scalar(chi2 df_m p) pr2 bic

----------------------------------------------------------------------------
                      (1)             (2)             (3)             (4)
                      lpm            lrm1            lrm2            lrm3
----------------------------------------------------------------------------
main
1.male              0.405***        1.700***        1.803***        2.015***
                  (28.55)         (23.46)         (24.11)         (24.66)

1.socialcons        0.169***                        1.015***        0.922***
                  (11.23)                         (12.18)         (10.90)

fiscalcons         0.0198***                                        0.107***
                   (7.59)                                          (7.56)

_cons              0.0367*         -1.531***       -2.334***       -2.388***
                   (2.53)        (-26.42)        (-25.44)        (-25.64)
----------------------------------------------------------------------------
N                    4165            4165            4165            4165
pseudo R-sq                         0.113           0.143           0.153
BIC                4900.1          4858.9          4707.2          4657.4
chi2                                619.6           779.5           837.7
df_m                    3               1               2               3
p               1.08e-185       9.25e-137       5.29e-170       2.82e-181
----------------------------------------------------------------------------
z statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001


. * Plot the differences between men and women who are social conservatives
. * at different values of fiscal conservatism
. quietly margins male, at(fiscalcons = (-12 (.5) 12) socialcons = 1)
. marginsplot, noci scheme(sj) name(graph1, replace)

  Variables that uniquely identify margins: fiscalcons male
```



```
. quietly margins , dydx(male)at(fiscalcons = (-12 (.5) 12) socialcons = 1)
. marginsplot, noci scheme(sj) name(graph2, replace)

  Variables that uniquely identify margins: fiscalcons
```



```
. * Classification table
. estat clas

Logistic model for cruz

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       902           529  |       1431
     -     |       614          2120  |       2734
-----------+--------------------------+-----------
   Total   |      1516          2649  |       4165

Classified + if predicted Pr(D) >= .5
True D defined as cruz != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   59.50%
Specificity                     Pr( -|~D)   80.03%
Positive predictive value       Pr( D| +)   63.03%
Negative predictive value       Pr(~D| -)   77.54%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   19.97%
False - rate for true D         Pr( -| D)   40.50%
False + rate for classified +   Pr(~D| +)   36.97%
False - rate for classified -   Pr( D| -)   22.46%
--------------------------------------------------
Correctly classified                        72.56%
--------------------------------------------------
```

Answer the following questions. Feel free to run additional analyses if you think it would help.

a. Cruz thinks he has almost as much support as Buttigieg. Do you agree?
b. Three nested logistic regression models were run. Cruz's staff thought model 3 was the best. Explain how both likelihood ratio chi-square statistics and BIC statistics support this decision.
c. According to the linear probability model the effect of male is .405. What does that number mean, i.e. how do we interpret it? Why is that number potentially problematic?
d. Various `margins` and `marginsplot` commands were run. What do they tell you about the effect of gender on Cruz's support? How does this differ from what the LPM said about the effect of gender?
e. What exactly does the classification table tell you? How many cases are classified correctly? How many cases would you classify correctly if you just predicted that nobody supported Cruz? Does the classification table do better? Indicate whether you think the

classification table is useful in this case. If you don't think it is useful, briefly explain why.

f. (Optional) Indicate anything else you think is worth noting. You can run additional analyses if you want. For example, you might look at the odds ratios or run diagnostic tests or do other things that were in the class handouts.

2. Now you will do similar analyses using a data set of your choice (preferably the data you want to use for your paper but you can pick anything, including the data you used in HW 1). From this data set you will need

- A binary dependent variable. If you are desperate, remember that continuous, ordinal, and count variables can be dichotomized if necessary, e.g. an ordinal variable might be recoded to 1 = agree, 0 = disagree.
- Three independent variables. At least one should be a binary variable (e.g. gender, if it is measured as binary in your data set) and at least 1 should be continuous (or if desperate, use an ordinal variable). (You can have more independent variables if you want but you will need to adapt the analysis a bit. I would suggest keeping it simple for now but you can do what you want.)
- Briefly explain what each of these variables are and how they are coded, even if you already did so in HW 1.

Now do the following:

a. Run descriptive statistics on your variables. Indicate how many cases are coded 1 on your dependent variable.

b. Estimate a linear probability model. Also estimate a series of nested logistic regressions. Summarize all the models in a single table as was done above – you don't have to use `esttab` but make sure the same or similar sorts of information gets presented. Use likelihood ratio chi square statistics and BIC tests to determine which LRM model is best. If the BIC and LR approaches prefer different models be sure to point that out. However, all of the subsequent analyses should be based on the model that has all the variables.

c. Using one of your binary independent variables, indicate what its LPM coefficient says about the variable's effect on your outcome variable.

d. Use `margins` and/or `marginsplot` commands to compute predicted probabilities that involve your binary independent variable. For example, you might see how the effect of gender varies across different values of your continuous variable. Note how this differs from what the LPM said the effect was.

e. Estimate the classification table for your model with all the variables. How many cases are classified correctly? How does this compare to what you would get if you just predicted that every case would be a 1 (or a 0)? Indicate whether you think the classification table is useful in this case. If you don't think it is useful, briefly explain why. If you feel ambitious, you might try out the Count R^2 and Adjusted Count R^2 measures that Long & Freese discuss.

f. (Optional) Indicate anything else you think is worth noting. You can run additional analyses if you want. For example, you might look at the odds ratios or run diagnostic tests or do other things that were in the class handouts.