Iterative Propensity Score Logistic Regression
Model Search Procedure (itpscore) Example File

The itpscore routine (Moore, Brand, and Shinkre 2021) implements the Imbens and Rubin (2015) algorithm that identifies a propensity score model by selecting covariates and their interactions that lead to the greatest gains in the logit log-likelihood function.

Users must define input variables specifying:

(1) treat: This input holds the name of a binary variable that is suitable for serving as an outcome in a logistic regression model. This is the treatment or causal variable for which the propensity score is calculated.

(2) cand: This input holds a list of three or more variable names that are candidate control variables in the logistic regression model. Variable names should be 15 characters or less in length.

Users can also define base (a set of baseline covariates), thr1, thr2, rand, keepint, and viewiter. These are all options that dictate how the iterative search process will be performed and what output will be shown.

The code and log file output below offers two examples of itpscore program runs. Each example utilizes Stata-provided data so that users can replicate results independently. Please see the following examples of program calls and output on subsequent pages:

Example 1: The first simple example only employs the required program inputs.

Example 2: The second example utilizes system settings “set seed” and “set maxiter.” The call also includes additional model required covariates (base), different user-defined convergence thresholds (thr1 and thr2), and 2 random covariates (rand).

Please reference the itpscore help file for further information on this package.

Example 1: Program call only utilizing required program inputs. Program call produces a propensity score model specification for the likelihood of labor union membership.

Example 1 Stata Code

```
clear
sysuse nlsw88

itpscore , treat(union) cand(hours tenure south age)
```

Example 1 Log File Output

```
-------------------------------------------------------------------------------------------------
Iterative Propensity Score Program Output
-------------------------------------------------------------------------------------------------
     ROUND 0: Initial Model: logit union
     ---> Linear Round 1
             Updated Best:  hours                          Log Likelihood = -1042.852420403815
             Updated Best:  tenure                         Log Likelihood = -1026.619754626618
             Updated Best:  south                          Log Likelihood = -1024.742855492293
     ---> Linear Round 2
```

```
            Updated Best:  hours                          Log Likelihood = -1019.44429320913
            Updated Best:  tenure                         Log Likelihood = -1006.310467188781
     ---> Linear Round 3
            Updated Best:  hours                          Log Likelihood = -1002.983999024951
     ---> Linear Round 4
            Updated Best:  age                            Log Likelihood = -1002.976049174116
     ---> Interaction Round 1
            Updated Best:  x_southtenure                  Log Likelihood = -1001.603310037596
            Updated Best:  x_southage                     Log Likelihood = -1001.373633361502
-------------------------------------------------------------------------------------------------
     Iterative Propensity Score Process Complete.
     Total Run Time  : 0 hours 0 minutes 1.191 seconds
     Program Executed: 18 Nov 2021 16:50:20
-------------------------------------------------------------------------------------------------
     Dependent Variable                    union
     Base Model
     Random Covariate(s)                   No
     Candidate Covariates                  hours tenure south age
     Total Models Estimated/Compared       30
     LL Improvement Threshold 1            2.71  (Applies to linear terms.)
     LL Improvement Threshold 2            3.84 (Applies to interaction terms.)

-------------------------------------------------------------------------------------------------
     Iterative Model Covariates
                                           south
                                           tenure
                                           hours
                                           age
                                           x_southage
-------------------------------------------------------------------------------------------------

Iteration 0:   log likelihood = -1042.3988
Iteration 1:   log likelihood = -1002.1085
Iteration 2:   log likelihood = -1001.3743
Iteration 3:   log likelihood = -1001.3736
Iteration 4:   log likelihood = -1001.3736

Logistic regression                                     Number of obs =  1,867
                                                        LR chi2(5)    =  82.05
                                                        Prob > chi2   = 0.0000
Log likelihood = -1001.3736                             Pseudo R2     = 0.0394

------------------------------------------------------------------------------
       union | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       south |   1.928822   1.499632     1.29   0.198    -1.010402    4.868046
      tenure |   .0474149   .0094258     5.03   0.000     .0289407    .0658892
       hours |   .0143931   .0058665     2.45   0.014      .002895    .0258912
         age |   .0260079   .0225085     1.16   0.248    -.0181079    .0701238
  x_southage |  -.0682668   .0382266    -1.79   0.074    -.1431897     .006656
       _cons |  -2.733793   .9167318    -2.98   0.003    -4.530554   -.9370314
------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------
     Description and Summary Statistics of Iterative Model Covariates

Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------
south           byte    %9.0g      southlbl   Lives in the south
tenure          float   %9.0g                 Job tenure (years)
hours           byte    %8.0g                 Usual hours worked
age             byte    %8.0g                 Age in current year
x_southage      float   %9.0g                 Auto Interaction: south*age

    Variable |        Obs        Mean    Std. dev.       Min        Max
```

```
-------------+--------------------------------------------------------
       south |      2,246    .4194123    .4935728          0          1
      tenure |      2,231     5.97785    5.510331          0   25.91667
       hours |      2,242    37.21811    10.50914          1         80
         age |      2,246    39.15316    3.060002         34         46
  x_southage |      2,246    16.43188    19.44246          0         45


     Iterative Propensity Score (itpscore) program output complete.
-------------------------------------------------------------------------------------------------
```

Example 2: Program call utilizing system settings "set seed" and "set maxiter," additional model required covariates (base), different user-defined convergence thresholds (thr1 and thr2), and two random covariates (rand). Program call produces a propensity score model specification for the likelihood of being married.

Example 2 Stata Code

```
clear
sysuse nlsw88
quietly tab race, gen(race_)

set maxiter 25
set seed 2021

itpscore , treat(married) cand(wage hours tenure south race_1 race_2 race_3 smsa) base(age
collgrad)  thr1(2.71) thr2(5.41) rand(2) keepint viewiter
```

Example 2 Log File Output

```
-------------------------------------------------------------------------------------------------
Iterative Propensity Score Program Output
-------------------------------------------------------------------------------------------------
     ROUND 0: Initial Model: logit married age collgrad
     ---> Linear Round 1
             Updated Best:  wage                             Log Likelihood = -1462.451029167081
             Updated Best:  hours                            Log Likelihood = -1436.618246190747
             Updated Best:  race_1                           Log Likelihood = -1416.950613666133
             Updated Best:  race_2                           Log Likelihood = -1414.322311628997
     ---> Linear Round 2
             Updated Best:  wage                             Log Likelihood = -1409.989022387958
             Updated Best:  hours                            Log Likelihood = -1388.616555638053
     ---> Linear Round 3
             Updated Best:  wage                             Log Likelihood = -1386.327700742496
             Updated Best:  tenure                           Log Likelihood = -1378.577980399158
     ---> Linear Round 4
             Updated Best:  wage                             Log Likelihood = -1376.287951394306
             Updated Best:  south                            Log Likelihood = -1373.452569695672
     ---> Linear Round 5
             Updated Best:  wage                             Log Likelihood = -1371.903856660413
     ---> Linear Round 6
             Updated Best:  race_1                           Log Likelihood = -1371.903233836569
             Updated Best:  smsa                             Log Likelihood = -1371.118854299245
     ---> Interaction Round 1
             Updated Best:  x_ageage                         Log Likelihood = -1371.074887981381
             Updated Best:  x_agecollgrad                    Log Likelihood = -1370.323855540372
             Updated Best:  x_collgradtenure                 Log Likelihood = -1369.998407753622
             Updated Best:  x_race_2hours                    Log Likelihood = -1365.96358340995
     ---> Interaction Round 2
             Updated Best:  x_ageage                         Log Likelihood = -1365.945855188618
             Updated Best:  x_agecollgrad                    Log Likelihood = -1365.300163185643
             Updated Best:  x_agewage                        Log Likelihood = -1365.29821439143
```

```
            Updated Best:  x_collgradtenure               Log Likelihood = -1365.030421334725
            Updated Best:  x_race_2wage                   Log Likelihood = -1364.229601296574
            Updated Best:  x_hourssmsa                    Log Likelihood = -1363.636772631846
            Updated Best:  x_tenurewage                   Log Likelihood = -1362.846930076877
     ---> Interaction Round 3
            Updated Best:  x_ageage                       Log Likelihood = -1362.825093706481
            Updated Best:  x_agecollgrad                  Log Likelihood = -1362.247158146627
            Updated Best:  x_collgrad_itpsrand2           Log Likelihood = -1361.873365405383
            Updated Best:  x_race_2wage                   Log Likelihood = -1361.190078866772
            Updated Best:  x_hourssmsa                    Log Likelihood = -1360.410140503021
-------------------------------------------------------------------------------------------------
     Iterative Propensity Score Process Complete.
     Total Run Time  : 0 hours 0 minutes 10.34 seconds
     Program Executed: 18 Nov 2021 16:53:24
-------------------------------------------------------------------------------------------------
     Dependent Variable                    married
     Base Model                            age collgrad
     Random Covariates                     Yes, 2
     Candidate Covariates
                                           wage
                                           hours
                                           tenure
                                           south
                                           race_1
                                           race_2
                                           race_3
                                           smsa

     Total Models Estimated/Compared       189
     LL Improvement Threshold 1            2.71  (Applies to linear terms.)
     LL Improvement Threshold 2            5.41 (Applies to interaction terms.)

-------------------------------------------------------------------------------------------------
     Iterative Model Covariates
                                           age
                                           collgrad
                                           race_2
                                           hours
                                           tenure
                                           south
                                           wage
                                           smsa
                                           x_race_2hours
                                           x_tenurewage
                                           x_hourssmsa
-------------------------------------------------------------------------------------------------

Iteration 0:   log likelihood = -1451.8393
Iteration 1:   log likelihood = -1361.8964
Iteration 2:   log likelihood = -1360.4142
Iteration 3:   log likelihood = -1360.4101
Iteration 4:   log likelihood = -1360.4101

Logistic regression                                     Number of obs =  2,227
                                                        LR chi2(11)   = 182.86
                                                        Prob > chi2   = 0.0000
Log likelihood = -1360.4101                             Pseudo R2     = 0.0630

------------------------------------------------------------------------------
     married | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -.0241541   .0152471    -1.58   0.113    -.0540378    .0057297
    collgrad |   .0879525   .1142376     0.77   0.441     -.135949     .311854
      race_2 |  -2.566076   .4815862    -5.33   0.000    -3.509968   -1.622184
       hours |  -.0227186    .009183    -2.47   0.013    -.0407169   -.0047203
      tenure |   .0388009   .0172295     2.25   0.024     .0050317    .0725701
```

```
        south |   .2659425   .1009575     2.63   0.008     .0680694    .4638156
         wage |    .007684   .0117399     0.65   0.513    -.0153257    .0306937
         smsa |   .7879545   .4173403     1.89   0.059    -.0300175    1.605926
x_race_2hours |   .0388366   .0122119     3.18   0.001     .0149018    .0627715
 x_tenurewage |  -.0041615   .0016834    -2.47   0.013    -.0074608   -.0008622
  x_hourssmsa |  -.0237082   .0106543    -2.23   0.026    -.0445903    -.002826
        _cons |   2.578275    .705677     3.65   0.000     1.195174    3.961377
------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------
     Description and Summary Statistics of Iterative Model Covariates

Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------------------------
------------------------------------------------------------------
age             byte    %8.0g                 Age in current year
collgrad        byte    %16.0g     gradlbl    College graduate
race_2          byte    %8.0g                 race==Black
hours           byte    %8.0g                 Usual hours worked
tenure          float   %9.0g                 Job tenure (years)
south           byte    %9.0g      southlbl   Lives in the south
wage            float   %9.0g                 Hourly wage
smsa            byte    %9.0g      smsalbl    Lives in SMSA
x_race_2hours   float   %9.0g                 Auto Interaction: race_2*hours
x_tenurewage    float   %9.0g                 Auto Interaction: tenure*wage
x_hourssmsa     float   %9.0g                 Auto Interaction: hours*smsa

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         age |      2,246    39.15316    3.060002         34         46
    collgrad |      2,246    .2368655    .4252538          0          1
      race_2 |      2,246    .2595726    .4384977          0          1
       hours |      2,242    37.21811    10.50914          1         80
      tenure |      2,231     5.97785    5.510331          0   25.91667
-------------+---------------------------------------------------------
       south |      2,246    .4194123    .4935728          0          1
        wage |      2,246    7.766949    5.755523   1.004952   40.74659
        smsa |      2,246    .7039181    .4566292          0          1
x_race_2ho~s |      2,242     9.87868    17.16993          0         70
x_tenurewage |      2,231    52.22341    66.48405          0   824.0607
-------------+---------------------------------------------------------
 x_hourssmsa |      2,242    26.36976    19.26008          0         80


     Iterative Propensity Score (itpscore) program output complete.
------------------------------------------------------------------------------------------------
```

References

Imbens, Guido W., and Donald B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

Moore, Ravaris L., Jennie E. Brand, Tanvi Shinkre. 2021. itpscore: Stata module to perform iterative propensity score estimation. Statistical Software Components [S######], Boston College Department of Economics, revised [date].