Be Wary of the Stepwise Regression Selection Procedures

Example: Suppose you are interested in associating socio-demographic factors with total medical expenditures for persons 19 to 94 years of age. You have data from the National Medical Expenditure Survey (1987). You are interested in looking at the following socio-demographic variables after adjusting for age and gender (known confounders):

Seat Belt Use as a surrogate for risk taking behavior:

1 – rare, 2 – some, 3 – always/almost always

Education Level:

1 – college grad, 2 – some college, 3 – hs grad, 4 – other

Marital Status:

1 – married, 2 – widowed, 3 – divorced, 4 – separated, 5 – never married

Our full model would be represented by the following output (assuming that its adequate to adjust for linear age):

```
. xi: regress TOTALEXP i.beltuse i.educate i.marital MALE LASTAGE
i.beltuse         _Ibeltuse_1-3       (naturally coded; _Ibeltuse_1 omitted)
i.educate         _Ieducate_1-4       (naturally coded; _Ieducate_1 omitted)
i.marital         _Imarital_1-5       (naturally coded; _Imarital_1 omitted)

      Source |       SS       df       MS              Number of obs =   22076
-------------+------------------------------           F( 11, 22064) =   78.96
       Model |  3.1044e+10    11  2.8222e+09           Prob > F      =  0.0000
    Residual |  7.8860e+11 22064  35741387.2           R-squared     =  0.0379
-------------+------------------------------           Adj R-squared =  0.0374
       Total |  8.1964e+11 22075  37129887.0           Root MSE      =  5978.4

------------------------------------------------------------------------------
    TOTALEXP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Ibeltuse_2 |  -254.0798    122.231    -2.08   0.038    -493.6612   -14.49831
 _Ibeltuse_3 |   16.58768   101.5703     0.16   0.870    -182.4973    215.6727
 _Ieducate_2 |   206.0112   140.3061     1.47   0.142     -68.9987    481.0211
 _Ieducate_3 |   139.6909   119.1392     1.17   0.241    -93.83043    373.2123
 _Ieducate_4 |  -9.276727    155.301    -0.06   0.952    -313.6778    295.1243
 _Imarital_2 |   346.6635   152.8893     2.27   0.023     46.98958    646.3374
 _Imarital_3 |   272.9113   147.7179     1.85   0.065    -16.62643    562.4489
 _Imarital_4 |    676.878   234.6771     2.88   0.004      216.894    1136.862
 _Imarital_5 |   372.4754   117.1905     3.18   0.001     142.7736    602.1772
        MALE |  -111.2689   83.61385    -1.33   0.183    -275.1581    52.62019
    LASTAGE |   63.67303    2.71129    23.48   0.000     58.35871    68.98735
       _cons |  -1212.247   189.8173    -6.39   0.000    -1584.303   -840.1919
------------------------------------------------------------------------------
```

Someone not too familiar with Stata may request the following stepwise regression procedure and obtain the result below:

```
. sw regress TOTALEXP _Ibeltuse_2 _Ibeltuse_3 _Ieducate_2 _Ieducate_3 _Ieducate_4 _Im
> arital_2 _Imarital_3 _Imarital_4 _Imarital_5 MALE LASTAGE, pr(0.20)
                    begin with full model
p = 0.9524 >= 0.2000  removing _Ieducate_4
p = 0.8605 >= 0.2000  removing _Ibeltuse_3

      Source |       SS       df       MS              Number of obs =   22076
-------------+------------------------------           F(  9, 22066) =   96.51
       Model |  3.1043e+10     9  3.4492e+09           Prob > F      =  0.0000
    Residual |  7.8860e+11 22066  35738203.5           R-squared     =  0.0379
-------------+------------------------------           Adj R-squared =  0.0375
       Total |  8.1964e+11 22075  37129887.0           Root MSE      =  5978.1
```

```
------------------------------------------------------------------------------
    TOTALEXP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Ibeltuse_2 |  -265.9251   99.83842    -2.66   0.008    -461.6155   -70.23463
        MALE |   -112.335   83.31994    -1.35   0.178    -275.6481    50.97801
 _Ieducate_2 |   210.5739   122.3683     1.72   0.085    -29.27661    450.4245
 _Ieducate_3 |   142.8822   94.09414     1.52   0.129    -41.54903    327.3134
     LASTAGE |   63.64699   2.649114    24.03   0.000     58.45453    68.83944
 _Imarital_2 |    344.719   152.1435     2.27   0.023     46.50685    642.9311
 _Imarital_3 |   271.7896   147.5858     1.84   0.066    -17.48914    561.0683
 _Imarital_4 |   674.7144   234.2095     2.88   0.004     215.6471    1133.782
 _Imarital_5 |   371.6564   116.9555     3.18   0.001     142.4153    600.8976
       _cons |   -1202.13   164.9624    -7.29   0.000    -1525.468   -878.7919
------------------------------------------------------------------------------
```

What does _Ibeltuse_2 mean now? What is the reference group for beltuse now?

What does _Ieducate_2 and _Ieducate_3 mean now? What is the reference group for educate now?

A more logical approach to would be to restrict testing to the factors (beltuse, educate, marital, MALE and LASTAGE). See this very different result below:

```
sw regress TOTALEXP (_Ibeltuse_2 _Ibeltuse_3) (_Ieducate_2 _Ieducate_3 _Ieducate_4)
>  (_Imarital_2 _Imarital_3 _Imarital_4 _Imarital_5) MALE LASTAGE, pr(0.20)
                  begin with full model
p = 0.3157 >= 0.2000  removing _Ieducate_2 _Ieducate_3 _Ieducate_4

      Source |       SS       df       MS              Number of obs =   22076
-------------+------------------------------           F(  8, 22067) =  108.13
       Model |  3.0918e+10     8  3.8647e+09           Prob > F      =  0.0000
    Residual |  7.8872e+11 22067  35742261.3           R-squared     =  0.0377
-------------+------------------------------           Adj R-squared =  0.0374
       Total |  8.1964e+11 22075  37129887.0           Root MSE      =  5978.5

------------------------------------------------------------------------------
    TOTALEXP |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Ibeltuse_2 |   -249.694    121.961    -2.05   0.041    -488.7462   -10.64174
 _Ibeltuse_3 |   14.15188   99.68587     0.14   0.887    -181.2396    209.5433
     LASTAGE |   62.79254    2.61119    24.05   0.000     57.67442    67.91065
 _Imarital_2 |   339.3327   152.2451     2.23   0.026     40.92151    637.7439
 _Imarital_3 |   276.5974   147.6817     1.87   0.061    -12.86931    566.0641
 _Imarital_4 |   680.2484   234.2978     2.90   0.004      221.008    1139.489
 _Imarital_5 |   373.5833   116.9656     3.19   0.001     144.3224    602.8441
        MALE |  -122.0456   83.36123    -1.46   0.143    -285.4395    41.34841
       _cons |   -1057.58   159.0009    -6.65   0.000    -1369.233   -745.9267
------------------------------------------------------------------------------
```