BOSTON COLLEGE
Department of Economics
EC 771: Econometrics
Spring 2011
Prof. Baum, Mr. Smith

Problem Set 3: Solutions

Point Distribution:
1), 2) : 5 points each
3), 4): 10 points each
5), 6): 15 points each

1) Denote the true value of β_1 and β_2 by β_{10} and β_{20}, respectively. Then.

$$c_i = \beta_{10} + \beta_{20} y_i^* + u_i^* = \beta_{10} + \beta_{20} y_i + (u_i^* - \beta_{20}\nu_i)$$

where we have used the equation $y_i = y_i^* + \nu_i$. Thus, if we run the regression of c_i on y_i and a constant, we have the error term in the regression, $u_i = u_i^* - \beta_{20}\nu_i$. Now, the covariance of y_i and u_i is calculated below:

$$\operatorname{cov}(y_i, u_i) = \mathrm{E}\left[(y_i^* + \nu_i)(u_i^* - \beta_{20}\nu_i)\right] - \mathrm{E}\left[(y_i^* + \nu_i)\right]\mathrm{E}\left[(u_i^* - \beta_{20}\nu_i)\right] \quad (1)$$
$$= -\beta_{20}\omega^2 \quad (2)$$

since $\operatorname{cov}(\nu_i, y_i^*) = 0$, $\operatorname{cov}(\nu_i, u_i^*) = 0$, and $\operatorname{cov}(u_i^*, y_i^*) = 0$, and since u_i^* and ν_i are mean-zero.

Since, $\beta_{20} > 0, \omega^2 > 0$, we have that $\operatorname{cov}(y_i, u_i) < 0$, which implies that the correlation is negative.

2)

```
. tsset date
        time variable:  date, 1967q1 to 1998q4
                delta:  1 quarter

. ivreg2 M Y L.M L2.M (R = L.R L2.R) if year>1967

Instrumental variables (2SLS) regression
----------------------------------------

                                                     Number of obs =       124
                                                     F(  4,    119) = 11585.54
                                                     Prob > F      =    0.0000
Total (centered) SS     =  13.09085971               Centered R2   =    0.9974
Total (uncentered) SS   =  15717.17658               Uncentered R2 =    1.0000
Residual SS             =   .033525694               Root MSE      =    .01644
```

```
------------------------------------------------------------------------------
           M |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           R |  -.0039326   .0008238    -4.77   0.000    -.0055473   -.0023179
           Y |   .0700983    .013829     5.07   0.000     .0429938    .0972027
           M |
         L1. |   1.396964    .078166    17.87   0.000     1.243762    1.550167
         L2. |  -.4528233   .0730188    -6.20   0.000    -.5959375    -.309709
       _cons |  -.2631494   .0821198    -3.20   0.001    -.4241013   -.1021975
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (identification/IV relevance test): 171.778
                                                   Chi-sq(2) P-val =    0.0000
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):          14.902
                                                   Chi-sq(1) P-val =    0.0001
------------------------------------------------------------------------------
Instrumented:         R
Included instruments: Y L.M L2.M
Excluded instruments: L.R L2.R
------------------------------------------------------------------------------

. ivendog

Tests of endogeneity of: R
H0: Regressor is exogenous
    Wu-Hausman F test:                    0.12547  F(1,118)    P-value = 0.72380
    Durbin-Wu-Hausman chi-sq test:        0.13171  Chi-sq(1)   P-value = 0.71666
```

The rejection of the null in the Sargan test indicates that the excluded instruments are not valid instruments. The Durbin-Wu-Hausman test of exogeneity of r_t i.e. of the appropriateness of OLS fails to reject this null hypothesis, suggesting that IV regression is not needed. Since one would expect that lagged values of a variable to be reasonable instruments, the results of these two tests suggests that the original model might have been misspecified, and that the lagged values of r_t that were used as instruments ought not to have been omitted from the regression model.

3)

```
. regress logQty X_2 X_3 logPrice

      Source |       SS       df       MS              Number of obs =     120
-------------+------------------------------           F(  3,   116) =  116.03
       Model |   9.1604655     3   3.0534885           Prob > F      =  0.0000
    Residual |  3.05268076   116  .026316213           R-squared     =  0.7500
-------------+------------------------------           Adj R-squared =  0.7436
       Total |  12.2131463   119  .102631481           Root MSE      =  .16222
```

```
------------------------------------------------------------------------------
      logQty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         X_2 |   .3830345   .0226551    16.91   0.000     .3381632    .4279058
         X_3 |   -.197489   .0343902    -5.74   0.000     -.265603   -.1293749
    logPrice |  -.3951405    .025181   -15.69   0.000    -.4450146   -.3452665
       _cons |     3.4532   .1234017    27.98   0.000     3.208788    3.697613
------------------------------------------------------------------------------

. ivreg2 logQty X_2 X_3 (logPrice = X_4 X_5)

IV (2SLS) estimation
--------------------

                                                      Number of obs =      120
                                                      F(  3,   116) =    97.07
                                                      Prob > F      =   0.0000
Total (centered) SS     =  12.21314626                Centered R2   =   0.6162
Total (uncentered) SS   =  3303.294588                Uncentered R2 =   0.9986
Residual SS             =  4.686889168                Root MSE      =    .1976

------------------------------------------------------------------------------
      logQty |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    logPrice |  -.5935736   .0389166   -15.25   0.000    -.6698487   -.5172985
         X_2 |    .491335   .0305378    16.09   0.000     .4314819     .551188
         X_3 |  -.2252342   .0420297    -5.36   0.000     -.307611   -.1428575
       _cons |   3.061598    .157588    19.43   0.000     2.752731    3.370465
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (underidentification test):         116.545
                                                   Chi-sq(2) P-val =    0.0000
------------------------------------------------------------------------------
Cragg-Donald F statistic (weak identification test):                    94.366
Stock-Yogo weak ID test critical values: 10% maximal IV size             19.93
                                         15% maximal IV size             11.59
                                         20% maximal IV size              8.75
                                         25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           0.975
                                                   Chi-sq(1) P-val =    0.3235
------------------------------------------------------------------------------
Instrumented:         logPrice
Included instruments: X_2 X_3
Excluded instruments: X_4 X_5
```

```
------------------------------------------------------------------------------

. ivendog

Tests of endogeneity of: logPrice
H0: Regressor is exogenous
    Wu-Hausman F test:                831.99234  F(1,115)    P-value = 0.00000
    Durbin-Wu-Hausman chi-sq test:    105.42755  Chi-sq(1)   P-value = 0.00000
```

The null of the Durbin-Wu-Hausman test of exogeneity of p_t is overwhelmingly rejected, which implies that the OLS estimation is not valid. This can be seen also by comparing the coefficients from the OLS regression with those from the IV regression. The differences between the two are much larger than the standard errors of the OLS coefficients, which should not be the case if the OLS estimates are consistent. The null of the Sargan test for validity of the instruments is not rejected at any reasonable level, which indicates that the instruments, and the IV regression, is valid.

```
. regress logPrice X_2 X_3 logQty

      Source |       SS       df       MS              Number of obs =     120
-------------+------------------------------           F(  3,   116) =  153.25
       Model |  52.6754551     3   17.558485           Prob > F      =  0.0000
    Residual |  13.2904765   116  .114573073           R-squared     =  0.7985
-------------+------------------------------           Adj R-squared =  0.7933
       Total |  65.9659316   119   .55433556           Root MSE      =  .33849

------------------------------------------------------------------------------
    logPrice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         X_2 |   .8337184   .0418204    19.94   0.000     .7508878    .9165491
         X_3 |  -.3845204   .0730633    -5.26   0.000    -.5292316   -.2398093
      logQty |  -1.720326   .1096305   -15.69   0.000    -1.937463   -1.503189
       _cons |   5.308666   .5204841    10.20   0.000     4.277782     6.33955
------------------------------------------------------------------------------

. ivreg2 logPrice X_2 X_3 (logQty = X_4 X_5)

IV (2SLS) estimation
--------------------

                                                      Number of obs =      120
                                                      F(  3,   116) =   145.70
                                                      Prob > F      =   0.0000
Total (centered) SS     =  65.96593159                Centered R2   =   0.7983
Total (uncentered) SS   =  294.8998539                Uncentered R2 =   0.9549
Residual SS             =  13.30781268                Root MSE      =     .333
```

```
------------------------------------------------------------------------------
    logPrice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logQty |  -1.677681   .1102461   -15.22   0.000    -1.893759   -1.461603
         X_2 |   .8265807   .0413213    20.00   0.000     .7455926    .9075689
         X_3 |  -.3784546   .0719554    -5.26   0.000    -.5194846   -.2374247
       _cons |    5.12815   .5211023     9.84   0.000     4.106808    6.149492
------------------------------------------------------------------------------
Anderson canon. corr. LR statistic (underidentification test):         378.000
                                                   Chi-sq(2) P-val =    0.0000
------------------------------------------------------------------------------
Cragg-Donald F statistic (weak identification test):                  1284.323
Stock-Yogo weak ID test critical values: 10% maximal IV size             19.93
                                         15% maximal IV size             11.59
                                         20% maximal IV size              8.75
                                         25% maximal IV size              7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           0.970
                                                   Chi-sq(1) P-val =    0.3246
------------------------------------------------------------------------------
Instrumented:         logQty
Included instruments: X_2 X_3
Excluded instruments: X_4 X_5
------------------------------------------------------------------------------

. ivendog

Tests of endogeneity of: logQty
H0: Regressor is exogenous
    Wu-Hausman F test:                    3.45111  F(1,115)    P-value = 0.06577
    Durbin-Wu-Hausman chi-sq test:        3.49624  Chi-sq(1)   P-value = 0.06151
```

The DWH test rejects the null hypothesis that the OLS regression is consistent at the 10% level, but not the 5% level. Therefore exogeneity of q_t is doubtful, but in contrast to the previous specification, the level of correlation between the suspect regressor and the error term is much lower. An examination of the difference in the parameter estimates between the OLS and the IV regressions supports the notion that the OLS estimates are only slightly biased. The Sargan test indicates that the instrument set used is valid, and so the IV regression is valid.

Rewriting the demand equation as an inverse demand equation, we obtain

$$p_t = -\frac{\beta_1}{\gamma} - \frac{\beta_2}{\gamma}x_{t2} - \frac{\beta_3}{\gamma}x_{t3} + \frac{1}{\gamma}q_t - \frac{1}{\gamma}u_t$$

Thus, estimating the inverse demand equation

$$p_t = \beta_1^* + \beta_2^* x_{t2} + \beta_3^* x_{t3} + \gamma^* q_t + v_t$$

we obtain the following relationships between the parameters of the two regression models:

$$\beta_i^* = -\frac{\beta_i}{\gamma}, \quad \gamma^* = \frac{1}{\gamma} \implies \beta_i = -\frac{\beta_i^*}{\gamma^*}$$

Thus, we have four estimates of γ:

$$\hat{\gamma}_{OLS} = -0.3951, \quad \hat{\gamma}_{2SLS} = -0.5936, \quad \frac{1}{\hat{\gamma}^*_{OLS}} = -\frac{1}{1.7203} = -0.5813, \quad \frac{1}{\hat{\gamma}^*_{2SLS}} = -\frac{1}{1.6777} = -0.5961$$

It is a little surprising that the estimate obtained from OLS estimation of the inverse demand function matches so closely the estimates obtained from the two IV(2SLS) regressions. This adds support to the earlier finding that the OLS estimation of the inverse demand equation is adequate. However, it is clear that OLS estimation of the demand equation yields biased parameter estimates.

4) a)

```
. regress lwage educ exper tenure married black south urban

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  7,   927) =   44.75
       Model |  41.8377677     7  5.97682396           Prob > F      =  0.0000
    Residual |  123.818527   927  .133569069           R-squared     =  0.2526
-------------+------------------------------           Adj R-squared =  0.2469
       Total |  165.656294   934  .177362199           Root MSE      =  .36547

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0654307   .0062504    10.47   0.000     .0531642    .0776973
       exper |    .014043   .0031852     4.41   0.000      .007792     .020294
      tenure |   .0117473    .002453     4.79   0.000     .0069333    .0165613
     married |   .1994171   .0390502     5.11   0.000     .1227801    .2760541
       black |  -.1883499   .0376666    -5.00   0.000    -.2622717   -.1144282
       south |  -.0909036   .0262485    -3.46   0.001     -.142417   -.0393903
       urban |   .1839121   .0269583     6.82   0.000     .1310056    .2368185
       _cons |   5.395497    .113225    47.65   0.000      5.17329    5.617704
------------------------------------------------------------------------------
```

Ceteris paribus, the approximate difference in the log wage of blacks and nonblacks is -0.18835, where blacks receive a lower wage. The difference is statistically significant, as the p-value of the t-test for significance is basically 0 i.e. the null of the coefficient on black being zero is rejected at 0.1% level. Going from log wages to wages, we obtain that ceteris paribus the ratio of wages of blacks to nonblacks is $e^{-0.18835} = 0.8283$ i.e. about blacks earn about 17% lower than non-blacks, all other things being equal.

b)

```
. gen blackXeduc = black * educ

. regress lwage educ blackXeduc exper tenure married black south urban

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  8,   926) =   39.32
       Model |  42.0055536     8  5.2506942            Prob > F      =  0.0000
    Residual |  123.650741   926  .133532117           R-squared     =  0.2536
-------------+------------------------------           Adj R-squared =  0.2471
       Total |  165.656294   934  .177362199           Root MSE      =  .36542

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0671153   .0064277    10.44   0.000     .0545008    .0797299
  blackXeduc |  -.0226237   .0201827    -1.12   0.263    -.0622327    .0169854
       exper |   .0138259   .0031906     4.33   0.000     .0075642    .0200876
      tenure |    .011787   .0024529     4.81   0.000     .0069732    .0166009
     married |   .1989077   .0390474     5.09   0.000     .1222761    .2755394
       black |   .0948094   .2553995     0.37   0.711    -.4064194    .5960383
       south |  -.0894495   .0262769    -3.40   0.001    -.1410187   -.0378803
       urban |   .1838523   .0269547     6.82   0.000      .130953    .2367516
       _cons |   5.374817   .1147027    46.86   0.000     5.149709    5.599924
------------------------------------------------------------------------------

. test black blackXeduc

 ( 1)  black = 0
 ( 2)  blackXeduc = 0

       F(  2,   926) =   13.13
            Prob > F =    0.0000
```

The returns to education for blacks is lower than that of whites, but the difference is not statistically significantly different from zero at any reasonable level. It is interesting to note that the negative effect of being black on the log wage is no longer negative or significant. Jointly testing to see if there is an effect of being black, we reject the hypothesis that the regression is stable over the category of race. Thus, even though neither black nor blackXeduc were significantly different from zero individually, they are jointly significantly different from zero.

c)

```
. gen MB = married * black

. gen mB = (1 - married) * black

. gen Mb = married * (1 - black)
```

```
. gen mb = (1 - married ) * (1 - black)

. regress lwage educ exper tenure MB Mb mB south urban

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  8,   926) =   39.17
       Model |  41.8849419     8  5.23561773           Prob > F      =  0.0000
    Residual |  123.771352   926  .133662368           R-squared     =  0.2528
-------------+------------------------------           Adj R-squared =  0.2464
       Total |  165.656294   934  .177362199           Root MSE      =   .3656

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0654751    .006253    10.47   0.000     .0532034    .0777469
       exper |   .0141462    .003191     4.43   0.000     .0078837    .0204087
      tenure |   .0116628   .0024579     4.74   0.000      .006839    .0164866
          MB |   .0094485   .0560131     0.17   0.866    -.1004788    .1193757
          Mb |   .1889147   .0428777     4.41   0.000     .1047659    .2730635
          mB |  -.2408201   .0960229    -2.51   0.012    -.4292678   -.0523724
       south |  -.0919894   .0263212    -3.49   0.000    -.1436455   -.0403333
       urban |   .1843501   .0269778     6.83   0.000     .1314053    .2372948
       _cons |   5.403793   .1141222    47.35   0.000     5.179825    5.627761
------------------------------------------------------------------------------

. lincom MB -Mb

 ( 1)  MB - Mb = 0

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.1794663   .0405386    -4.43   0.000    -.2590244   -.0999082
------------------------------------------------------------------------------
```

The log wage differential between married black and married nonblacks is -0.1795, and is significantly different from zero. In terms of wage levels, this difference translates to a 16.5% lower wage for married blacks than for married nonblacks ($e^{-0.1795} = 0.8357$).

5) a) One would expect that higher SAT scores would imply higher college GPA. Similarly, lower academic percentile (measured as the percentage who performed better than the student in question) would suggest better college performance as measured by college GPA. It is unclear how size (or the square of the size) of the high-school class would affect college GPA. One might suppose that being an athlete would detract from studying, which would lower college GPA. It is unclear how gender would affect GPA.

b)

```
. regress colgpa hsize hsizesq hsperc sat female athlete

      Source |       SS       df       MS              Number of obs =    4137
-------------+------------------------------           F(  6,  4130) =  284.59
       Model |  524.819305     6  87.4698842           Prob > F      =  0.0000
    Residual |  1269.37637  4130  .307355053           R-squared     =  0.2925
-------------+------------------------------           Adj R-squared =  0.2915
       Total |  1794.19567  4136  .433799728           Root MSE      =   .5544

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsize |  -.0568543   .0163513    -3.48   0.001    -.0889117   -.0247968
     hsizesq |   .0046754   .0022494     2.08   0.038     .0002654    .0090854
      hsperc |  -.0132126   .0005728   -23.07   0.000    -.0143355   -.0120896
         sat |   .0016464   .0000668    24.64   0.000     .0015154    .0017774
      female |   .1548814   .0180047     8.60   0.000     .1195826    .1901802
     athlete |   .1693064   .0423492     4.00   0.000     .0862791    .2523336
       _cons |   1.241365   .0794923    15.62   0.000     1.085517    1.397212
------------------------------------------------------------------------------
```

The estimated GPA differential between athletes and nonathletes is 0.169, and is statistically significant at the 0.1% level.

c)

```
. regress colgpa hsize hsizesq hsperc female athlete

      Source |       SS       df       MS              Number of obs =    4137
-------------+------------------------------           F(  5,  4131) =  191.92
       Model |  338.217123     5  67.6434246           Prob > F      =  0.0000
    Residual |  1455.97855  4131   .35245184           R-squared     =  0.1885
-------------+------------------------------           Adj R-squared =  0.1875
       Total |  1794.19567  4136  .433799728           Root MSE      =  .59368

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsize |  -.0534038   .0175092    -3.05   0.002    -.0877313   -.0190763
     hsizesq |   .0053228   .0024086     2.21   0.027     .0006007     .010045
      hsperc |  -.0171365   .0005892   -29.09   0.000    -.0182916   -.0159814
      female |   .0581231   .0188162     3.09   0.002     .0212333     .095013
     athlete |   .0054487   .0447871     0.12   0.903    -.0823582    .0932556
```

```
       _cons |   3.047698   .0329148    92.59   0.000     2.983167    3.112229
------------------------------------------------------------------------------
```

Yes, the effect is greatly lessened, though still positive. However, the estimate is not statistically significantly different from zero at any reasonable level.

Why would this happen? It suggests that athelte and sat are negatively correlated. What we have is omitted variable bias. Suppose the true specification of some process is

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Suppose instead we run the regression

$$y = \alpha_1 + \alpha_2 x_2 + \nu$$

Then, $\nu = \beta_3 x_3 + \epsilon$, and our estimate of α_2 will not be unbiased.

$$\begin{aligned} \hat{\alpha_2} &= \frac{\text{cov}\,(y, x_2)}{\text{var}\,(x_2)} = \frac{\text{cov}\,(\alpha_1 + \alpha_2 x_2 + \nu, x_2)}{\text{var}\,(x_2)} \\ &= \alpha_2 + \beta_3 \frac{\text{cov}\,(x_3, x_2)}{\text{var}\,(x_2)} = \alpha_2 + \beta_3 \rho_{x_2, x_3} \sqrt{\frac{\text{var}\,(x_3)}{\text{var}\,(x_2)}} \end{aligned}$$

So, what must be happening to have athlete become insignificantly different from 0 is that the negative correlation between the athlete and sat is biasing the estimated effect of begin an athlete in this variant of the model. The following unconditional correlation matrix and conditional correlation (via the regression) both indicate the negative correlation necessary for such a biased estimate result.

```
. corr athlete sat
(obs=4137)

             |  athlete      sat
-------------+------------------
     athlete |   1.0000
         sat |  -0.1851   1.0000

. regress sat athlete female white hsize hsizesq hsperc

      Source |       SS       df       MS              Number of obs =    4137
-------------+------------------------------           F(  6,  4130) =  153.28
       Model |  14637941.7     6  2439656.96           Prob > F      =  0.0000
    Residual |  65735904.6  4130  15916.6839           R-squared     =  0.1821
-------------+------------------------------           Adj R-squared =  0.1809
       Total |  80373846.3  4136  19432.7481           Root MSE      =  126.16

------------------------------------------------------------------------------
         sat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
     athlete |   -74.9073   9.679523    -7.74   0.000    -93.88437   -55.93022
      female |  -57.84925   3.999139   -14.47   0.000    -65.68971   -50.00878
       white |   106.3164   7.613348    13.96   0.000     91.39015    121.2427
       hsize |   1.211836     3.7214     0.33   0.745    -6.084113    8.507785
     hsizesq |   .3737667   .5118486     0.73   0.465    -.6297322    1.377266
      hsperc |  -2.495847   .1254639   -19.89   0.000    -2.741823    -2.24987
       _cons |   1002.015   9.763235   102.63   0.000     982.8739    1021.156
------------------------------------------------------------------------------
```

d)

```
. gen FA = female * athlete

. gen Fa = female * (1 - athlete)

. gen fA = (1 - female) * athlete

. gen fa = (1 - female) * (1 - athlete)

. regress colgpa hsize hsizesq hsperc sat FA Fa fA

      Source |       SS       df       MS              Number of obs =    4137
-------------+------------------------------           F(  7,  4129) =  243.88
       Model |  524.821272     7  74.9744674           Prob > F      =  0.0000
    Residual |   1269.3744  4129  .307429015           R-squared     =  0.2925
-------------+------------------------------           Adj R-squared =  0.2913
       Total |  1794.19567  4136  .433799728           Root MSE      =  .55446

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsize |  -.0568006   .0163671    -3.47   0.001    -.0888889   -.0247124
     hsizesq |   .0046699   .0022507     2.07   0.038     .0002573    .0090825
      hsperc |  -.0132114    .000573   -23.06   0.000    -.0143349    -.012088
         sat |   .0016462   .0000669    24.62   0.000     .0015151    .0017773
          FA |   .3297256   .0840593     3.92   0.000     .1649242    .4945271
          Fa |   .1546151   .0183122     8.44   0.000     .1187133    .1905168
          fA |   .1674185   .0484877     3.45   0.001     .0723564    .2624806
       _cons |   1.241575   .0795453    15.61   0.000     1.085623    1.397526
------------------------------------------------------------------------------

. test FA = Fa

 ( 1)  FA - Fa = 0
```

```
      F(  1,  4129) =    4.34
           Prob > F =    0.0372

. lincom FA - Fa

 ( 1)  FA - Fa = 0

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1751106   .0840258     2.08   0.037     .0103748    .3398464
------------------------------------------------------------------------------
```

The hypothesis that there is no difference between female athletes and female nonathletes is rejected. Female athletes have higher GPAs than female nonathletes.

e)

```
. regress colgpa hsize hsizesq hsperc sat femXsat female athlete

      Source |       SS       df       MS              Number of obs =    4137
-------------+------------------------------           F(  7,  4129) =  243.91
       Model |  524.867644     7  74.981092           Prob > F      =  0.0000
    Residual |  1269.32803  4129  .307417784           R-squared     =  0.2925
-------------+------------------------------           Adj R-squared =  0.2913
       Total |  1794.19567  4136  .433799728           Root MSE      =  .55445

------------------------------------------------------------------------------
      colgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsize |  -.0569121   .0163537    -3.48   0.001    -.0889741   -.0248501
     hsizesq |   .0046864   .0022498     2.08   0.037     .0002757    .0090972
      hsperc |   -.013225   .0005737   -23.05   0.000    -.0143497   -.0121003
         sat |   .0016255   .0000852    19.09   0.000     .0014585    .0017924
     femXsat |   .0000512   .0001291     0.40   0.692     -.000202    .0003044
      female |   .1023066   .1338023     0.76   0.445    -.1600179    .3646311
     athlete |   .1677568   .0425334     3.94   0.000     .0843684    .2511452
       _cons |   1.263743   .0974952    12.96   0.000       1.0726    1.454887
------------------------------------------------------------------------------
```

The effect of SAT scores do not differ by gender, since the coefficient on femXsat is statistically insignificantly different from 0.

6) a)

```
. regress nettfa e401k
```

```
      Source |       SS       df       MS              Number of obs =    9275
-------------+------------------------------           F(  1,  9273) =  196.22
       Model |  786249.663     1  786249.663           Prob > F      =  0.0000
    Residual |  37157139.8  9273  4007.02468           R-squared     =  0.0207
-------------+------------------------------           Adj R-squared =  0.0206
       Total |  37943389.5  9274   4091.3726           Root MSE      =  63.301

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e401k |   18.85832   1.346275    14.01   0.000     16.21933    21.49732
       _cons |   11.67677   .8430406    13.85   0.000     10.02423    13.32932
------------------------------------------------------------------------------
```

The average net total financial assets, which is measured in thousands of dollars, does differ by 401k eligibility, and the estimated difference is $18,858.

b)

```
. regress nettfa inc incsq age agesq male e401k

      Source |       SS       df       MS              Number of obs =    9275
-------------+------------------------------           F(  6,  9268) =  391.61
       Model |  7673992.51     6  1278998.75           Prob > F      =  0.0000
    Residual |    30269397  9268  3266.01176           R-squared     =  0.2022
-------------+------------------------------           Adj R-squared =  0.2017
       Total |  37943389.5  9274   4091.3726           Root MSE      =  57.149

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |  -.2702243   .0746105    -3.62   0.000    -.4164772   -.1239713
       incsq |    .010216   .0005871    17.40   0.000     .0090651    .0113669
         age |  -1.939771   .4834769    -4.01   0.000    -2.887492   -.9920497
       agesq |   .0345662   .0055482     6.23   0.000     .0236906    .0454418
        male |   3.369048   1.485813     2.27   0.023     .4565283    6.281569
       e401k |   9.713482   1.277127     7.61   0.000     7.210032    12.21693
       _cons |   21.19779   9.992211     2.12   0.034     1.610861    40.78472
------------------------------------------------------------------------------
```

Yes, both the quadratic terms included are statistically (and economically) significant. The estimated dollar effect of 401k eligibility is $9,713, and is statistically significant.

c)

```
. gen e401kXage41 = e401k * (age - 41)

. gen e401kXage41sq = e401k * (age - 41) * (age - 41)

. regress nettfa inc incsq age agesq e401kXage41 e401kXage41sq male e401k

      Source |       SS       df       MS              Number of obs =    9275
-------------+------------------------------           F(  8,  9266) =  297.95
       Model |  7763594.46     8  970449.308           Prob > F      =  0.0000
    Residual |    30179795  9266  3257.04673           R-squared     =  0.2046
-------------+------------------------------           Adj R-squared =  0.2039
       Total |  37943389.5  9274   4091.3726           Root MSE      =  57.071

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |  -.2705924   .0745119    -3.63   0.000    -.4166522   -.1245326
       incsq |   .0101878   .0005864    17.37   0.000     .0090383    .0113373
         age |  -2.287514   .5908919    -3.87   0.000    -3.445792   -1.129235
       agesq |   .0360854   .0067801     5.32   0.000     .0227948    .0493759
 e401kXage41 |   .6524833   .1313038     4.97   0.000      .395099    .9098676
e401kXage4~q |  -.0038891   .0116248    -0.33   0.738    -.0266762    .0188981
        male |   3.310739   1.483828     2.23   0.026     .4021098    6.219369
       e401k |   9.978824   1.718176     5.81   0.000     6.610821    13.34683
       _cons |   32.75766   12.21115     2.68   0.007     8.821123     56.6942
------------------------------------------------------------------------------
```

The linear interaction term between `e401k` and `age - 41` is significant, but not the quadratic interaction term. These interaction terms allow the effect of 401k eligibility to differ with age, centering around the age of 41. The difference in the effect of 401k eligibility between this interacted model with the previous model is not statistically significant, which is easy to see by noticing that the estimates have overlapping 95% confidence intervals.

d)

```
. replace fs1 = (fsize == 1)
(7258 real changes made)

. replace fs2 = (fsize == 2)
(7076 real changes made)

. replace fs3 = (fsize == 3)
(7446 real changes made)

. replace fs4 = (fsize == 4)
(7285 real changes made)
```

```
. replace fs5 = (fsize >= 5)
(424 real changes made)

. regress nettfa inc incsq age agesq male e401k fs2 fs3 fs4 fs5

      Source |       SS       df       MS              Number of obs =    9275
-------------+------------------------------           F( 10,  9264) =  237.03
       Model |   7730274.7    10   773027.47           Prob > F      =  0.0000
    Residual |  30213114.8  9264  3261.34659           R-squared     =  0.2037
-------------+------------------------------           Adj R-squared =  0.2029
       Total |  37943389.5  9274   4091.3726           Root MSE      =  57.108

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |  -.2412311   .0755141    -3.19   0.001    -.3892554   -.0932069
       incsq |   .0100501   .0005895    17.05   0.000     .0088946    .0112056
         age |  -1.504553   .4947967    -3.04   0.002    -2.474464   -.5346428
       agesq |    .029165   .0057032     5.11   0.000     .0179855    .0403444
        male |   1.323946   1.652795     0.80   0.423    -1.915896    4.563789
       e401k |   9.481517   1.278268     7.42   0.000      6.97583     11.9872
         fs2 |  -.3536808   1.924384    -0.18   0.854    -4.125898    3.418536
         fs3 |  -4.081595   2.013317    -2.03   0.043     -8.02814   -.1350509
         fs4 |  -5.696103   2.021384    -2.82   0.005     -9.65846   -1.733746
         fs5 |  -6.748335   2.235513    -3.02   0.003    -11.13043   -2.366237
       _cons |   15.74294     10.143     1.55   0.121     -4.13958    35.62545
------------------------------------------------------------------------------

. test fs2 = fs3 = fs4 = fs5 = 0

 ( 1)  fs2 - fs3 = 0
 ( 2)  fs2 - fs4 = 0
 ( 3)  fs2 - fs5 = 0
 ( 4)  fs2 = 0

       F(  4,  9264) =    4.31
            Prob > F =    0.0017
```

Yes, they are jointly significantly different from the zero vector. This implies that family size affects net total financial assets.

e)

```
. regress nettfa inc incsq age agesq male e401k _INfs_2 _INfs_3 _INfs_4 _INfs_5 _INfsXinc_2
> _INfsXinc_3 _INfsXinc_4 _INfsXinc_5 _IN2fsXincsq_2 _IN2fsXincsq_3 _IN2fsXincsq_4
```

```
> _IN2fsXincsq_5 _AfsXage_2 _AfsXage_3 _AfsXage_4 _AfsXage_5 _A2f sXagesq_2
> _A2fsXagesq_3 _A2fsXagesq_4 _A2fsXagesq_5 _MfsXmal_2_1 _MfsXmal_3_1 _MfsXmal_4_1
> _MfsXmal_5_1 _EfsXe40_2_1 _EfsXe40_3_1 _EfsXe40_4_1 _EfsXe40_5_1

      Source |       SS       df       MS              Number of obs =    9275
-------------+------------------------------           F( 34,  9240) =   72.18
       Model |  7962430.63    34  234189.136           Prob > F      =  0.0000
    Residual |  29980958.9  9240  3244.69252           R-squared     =  0.2099
-------------+------------------------------           Adj R-squared =  0.2069
       Total |  37943389.5  9274   4091.3726           Root MSE      =  56.962

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .7324251   .2465294     2.97   0.003      .249173    1.215677
       incsq |   .0004576   .0025611     0.18   0.858    -.0045627    .0054779
         age |  -1.593533   .9856013    -1.62   0.106    -3.525529    .3384627
       agesq |   .0289718   .0115668     2.50   0.012     .0062982    .0516454
        male |   2.468105   2.622154     0.94   0.347    -2.671897    7.608106
       e401k |   7.060432    2.76795     2.55   0.011      1.63464    12.48622
      _INfs_2 |  8.472165   27.89939     0.30   0.761     -46.2168    63.16113
      _INfs_3 |  7.819918    30.6623     0.26   0.799    -52.28495    67.92479
      _INfs_4 | -15.20342   32.39481    -0.47   0.639     -78.7044    48.29755
      _INfs_5 | -2.197631   39.65744    -0.06   0.956    -79.93497    75.53971
 _INfsXinc_2 |  -1.168416   .2829879    -4.13   0.000    -1.723134   -.6136971
 _INfsXinc_3 |  -.9147828   .2906207    -3.15   0.002    -1.484463   -.3451022
 _INfsXinc_4 |  -1.044873   .3030632    -3.45   0.001    -1.638944   -.4508028
 _INfsXinc_5 |  -1.380207    .338948    -4.07   0.000    -2.044619   -.7157936
_IN2fsXinc~2 |   .0118186   .0027563     4.29   0.000     .0064157    .0172214
_IN2fsXinc~3 |   .0081092   .0027966     2.90   0.004     .0026272    .0135912
_IN2fsXinc~4 |   .0098486   .0029118     3.38   0.001     .0041407    .0155564
_IN2fsXinc~5 |   .0131491   .0032414     4.06   0.000     .0067952    .0195029
  _AfsXage_2 |  -.0562306   1.349162    -0.04   0.967    -2.700886    2.588425
  _AfsXage_3 |   .3185739   1.506122     0.21   0.832    -2.633759    3.270906
  _AfsXage_4 |    1.54704   1.600657     0.97   0.334    -1.590602    4.684681
  _AfsXage_5 |   1.542528    1.92175     0.80   0.422    -2.224526    5.309583
_A2fsXages~2 |   .0062072   .0155267     0.40   0.689    -.0242286     .036643
_A2fsXages~3 |  -.0032207   .0175261    -0.18   0.854    -.0375758    .0311343
_A2fsXages~4 |  -.0190828   .0188308    -1.01   0.311    -.0559953    .0178298
_A2fsXages~5 |  -.0242004   .0223179    -1.08   0.278    -.0679484    .0195475
_MfsXmal_2_1 |    -3.4121   4.331588    -0.79   0.431    -11.90297    5.078768
_MfsXmal_3_1 |  -1.427265    5.08321    -0.28   0.779    -11.39148    8.536948
_MfsXmal_4_1 |  -.0454875   5.213619    -0.01   0.993    -10.26533    10.17436
_MfsXmal_5_1 |  -3.886421   6.284615    -0.62   0.536    -16.20565    8.432812
_EfsXe40_2_1 |   6.348744   3.796426     1.67   0.095     -1.09309    13.79058
_EfsXe40_3_1 |   .9601566   3.992241     0.24   0.810    -6.865516    8.785829
```

```
_EfsXe40_4_1 |   .8992764    3.88614     0.23   0.817    -6.718416    8.516969
_EfsXe40_5_1 |   4.071537   4.510328     0.90   0.367    -4.769702    12.91278
       _cons |   2.123567   19.91196     0.11   0.915    -36.90827    41.15541
------------------------------------------------------------------------------

. test _Ifs_2 _Ifs_3 _Ifs_4 _Ifs_5 _IfsXinc_2 _IfsXinc_3 _IfsXinc_4 _IfsXinc_5 _IfsXincsq_2 _IfsXincs
> _3 _IfsXage_4 _IfsXage_5 _IfsXagesq_2 _IfsXagesq_3 _IfsXagesq_4 _IfsXagesq_5 _IfsXmale_2 _IfsXmale_
> _3 _IfsXe401k_4 _IfsXe401k_5

 ( 1)  _Ifs_2 = 0
 ( 2)  _Ifs_3 = 0
 ( 3)  _Ifs_4 = 0
 ( 4)  _Ifs_5 = 0
 ( 5)  _IfsXinc_2 = 0
 ( 6)  _IfsXinc_3 = 0
 ( 7)  _IfsXinc_4 = 0
 ( 8)  _IfsXinc_5 = 0
 ( 9)  _IfsXincsq_2 = 0
 (10)  _IfsXincsq_3 = 0
 (11)  _IfsXincsq_4 = 0
 (12)  _IfsXincsq_5 = 0
 (13)  _IfsXage_2 = 0
 (14)  _IfsXage_3 = 0
 (15)  _IfsXage_4 = 0
 (16)  _IfsXage_5 = 0
 (17)  _IfsXagesq_2 = 0
 (18)  _IfsXagesq_3 = 0
 (19)  _IfsXagesq_4 = 0
 (20)  _IfsXagesq_5 = 0
 (21)  _IfsXmale_2 = 0
 (22)  _IfsXmale_3 = 0
 (23)  _IfsXmale_4 = 0
 (24)  _IfsXmale_5 = 0
 (25)  _IfsXe401k_2 = 0
 (26)  _IfsXe401k_3 = 0
 (27)  _IfsXe401k_4 = 0
 (28)  _IfsXe401k_5 = 0

       F( 28,  9240) =    3.17
            Prob > F =    0.0000
```

The null hypothesis of the Chow test, which is distributed as an F-statistic, is that the categories don't matter. However, the null is rejected at 5% (and even the 0.1%!) level, which implies that the regression is not stable over family size categories.