# ECON3150/4150 Spring 2015

## Seminar 5

Siv-Elisabeth Skjelbred

University of Oslo

Last updated: March 19, 2015

# Regular exercise 9.10

- A statistical analysis is said to have internal validity if the statistical inferences about casual effects are valid for the population being studied.
  - If the OLS estimator is unbiased and consistent
  - And standard errors are computed in a way that makes confidence intervals have the desired significance level.

# Threats to internal validity

- Omitted variable bias
- Misspecification of the functional form of the regression function
- Measurement error of the regressor
- Sample selection bias
- Simultaneous causality

# Sources of inconsistency of OLS standard errors

- Heteroskedasticity - solved by computing robust standard errors
- Correlation of the error term across observations

# External validity

- The analysis is said to have external validity if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.
- Threats to internal validity:
  - Differences in population
  - Differences in setting

# Population and setting

- Full time workers
- Age 30 to 64
- Random sample from the Current Population Survey
- Data from 2008

# Add E8.2a

```
1 . reg ed tuition dist bytest incomehi ownhome dadcoll ///
  > momcoll cue80 stwmfg black hispanic female, robust

Linear regression                               Number of obs =      3796
                                                F( 12,  3783) =    168.48
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2836
                                                Root MSE      =    1.5378
```

| ed | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| tuition | -.1910519 | .0985259 | -1.94 | 0.053 | -.3842209 | .0021171 |
| dist | -.0366613 | .0120749 | -3.04 | 0.002 | -.0603352 | -.0129874 |
| bytest | .0930377 | .003014 | 30.87 | 0.000 | .0871284 | .0989469 |
| incomehi | .3718305 | .0622177 | 5.98 | 0.000 | .2498471 | .4938138 |
| ownhome | .1385475 | .0649795 | 2.13 | 0.033 | .0111492 | .2659459 |
| dadcoll | .5709712 | .0763028 | 7.48 | 0.000 | .4213726 | .7205698 |
| momcoll | .3778102 | .0834999 | 4.52 | 0.000 | .214101 | .5415193 |
| cue80 | .0286753 | .0095229 | 3.01 | 0.003 | .0100049 | .0473458 |
| stwmfg80 | -.0425003 | .0199355 | -2.13 | 0.033 | -.0815857 | -.0034148 |
| black | .3506095 | .0674301 | 5.20 | 0.000 | .2184066 | .4828125 |
| hispanic | .3617649 | .0764184 | 4.73 | 0.000 | .2119397 | .5115902 |
| female | .1429742 | .0502718 | 2.84 | 0.004 | .0444118 | .2415366 |
| _cons | 8.920823 | .2434585 | 36.64 | 0.000 | 8.4435 | 9.398145 |

```
2 . predict yhat1
  (option xb assumed; fitted values)

3 . est sto reg1
```

If distance increase from 20 to 30 miles education is predicted to decrease by 0.037 years. If distance increase from 60 to 70 miles the same change is predicted.

# Add E8.2b

```
1 . gen lned =ln(ed)

2 . reg lned tuition dist bytest incomehi ownhome dadcoll ///
  > momcoll cue80 stwmfg black hispanic female, robust

Linear regression                               Number of obs =      3796
                                                F( 12,  3783) =    173.89
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2853
                                                Root MSE      =    .10918
```

| lned | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| tuition | -.0139382 | .0070081 | -1.99 | 0.047 | -.0276783 | -.0001982 |
| dist | -.0026072 | .0008651 | -3.01 | 0.003 | -.0043032 | -.0009111 |
| bytest | .0066561 | .0002133 | 31.21 | 0.000 | .0062379 | .0070742 |
| incomehi | .0265197 | .0044 | 6.03 | 0.000 | .0178931 | .0351463 |
| ownhome | .0098332 | .0046395 | 2.12 | 0.034 | .000737 | .0189295 |
| dadcoll | .0405374 | .0053518 | 7.57 | 0.000 | .0300446 | .0510302 |
| momcoll | .0266016 | .0058414 | 4.55 | 0.000 | .0151491 | .0380541 |
| cue80 | .0020357 | .0006768 | 3.01 | 0.003 | .0007088 | .0033626 |
| stwmfg80 | -.0028642 | .0014142 | -2.03 | 0.043 | -.0056368 | -.0000916 |
| black | .0261676 | .0048091 | 5.44 | 0.000 | .0167389 | .0355963 |
| hispanic | .0259986 | .0054098 | 4.81 | 0.000 | .0153922 | .0366049 |
| female | .0103059 | .0035664 | 2.89 | 0.004 | .0033137 | .0172981 |
| _cons | 2.265819 | .0172772 | 131.15 | 0.000 | 2.231946 | 2.299693 |

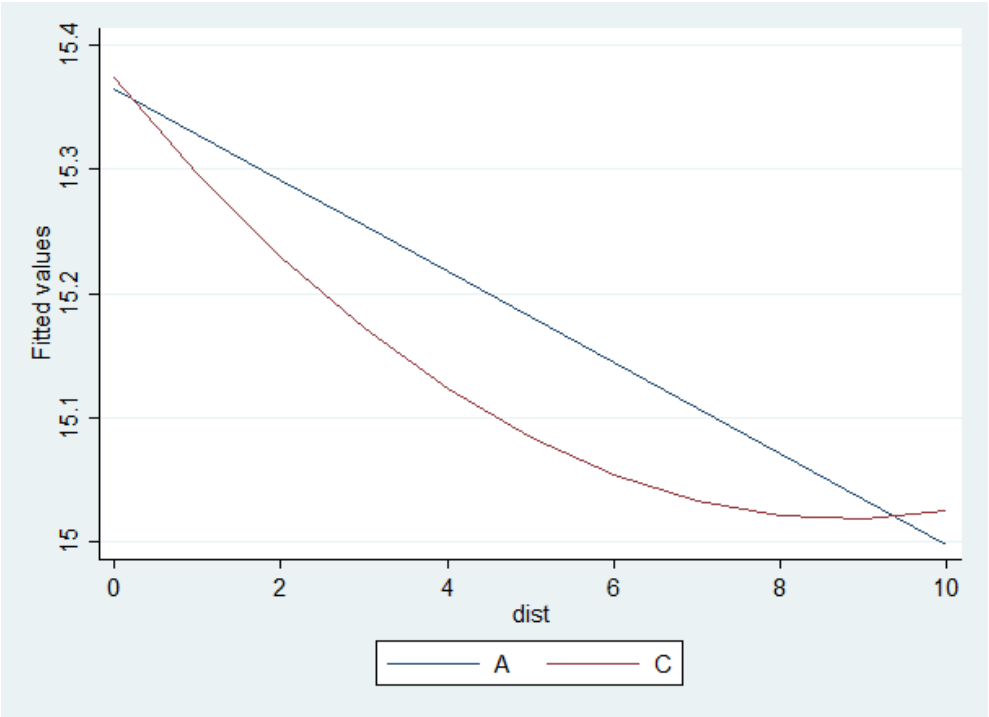
If distance increase by 10 miles the predicted increase in education is 0.26%.

# Add E8.2c and d

```
1 . reg ed tuition dist dist2 bytest incomehi ownhome dadcoll ///
  > momcoll cue80 stwmfg black hispanic female, robust

Linear regression                               Number of obs =      3796
                                                F( 13,  3782) =    155.93
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2844
                                                Root MSE      =    1.5372

------------------------------------------------------------------------------
             |               Robust
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tuition |   -.1928193   .0985524    -1.96   0.050    -.3860403    .0004016
        dist |   -.0811732   .0251112    -3.23   0.001    -.1304061   -.0319403
       dist2 |    .0046413   .0020542     2.26   0.024     .0006139    .0086687
     bytest |    .0926367   .0030243    30.63   0.000     .0867072    .0985661
   incomehi |    .3694975   .0623003     5.93   0.000     .2473521    .4916429
    ownhome |     .14327    .0648817     2.21   0.027     .0160636    .2704765
    dadcoll |    .5611581   .0765802     7.33   0.000     .4110157    .7113006
    momcoll |    .3777022   .0835025     4.52   0.000     .2139878    .5414166
      cue80 |    .0259537    .009587     2.71   0.007     .0071574    .0447499
    stwmfg80 |   -.0425539   .0199267    -2.14   0.033    -.081622   -.0034858
      black |    .3339309   .0683045     4.89   0.000     .2000136    .4678482
   hispanic |    .3333104   .0778789     4.28   0.000     .1806216    .4859991
     female |    .1433144   .0502511     2.85   0.004     .0447925    .2418363
      _cons |    9.012167   .2498793    36.07   0.000     8.522256    9.502078
------------------------------------------------------------------------------

2 .  matrix b=e(b)

3 .  display b[1,2]*3+b[1,3]*3^2-(b[1,2]*2+b[1,3]*2^2)
-.05796676

4 .  display b[1,2]*7+b[1,3]*7^2-(b[1,2]*6+b[1,3]*6^2)
-.0208364
```

Dist2 is statistically significant thus I prefer the regression in c.

# Add E8.2ei




- The quadratic regression in 3 is steeper for small values of dist than for larger values.
- The quadratic function is essentially flat when Dist=10 (100 miles).
- The only change in the regression function for a white male is that the intercept would shift.
- The functions would have the same slopes.

# Add E8.2e ii)

- The regression function becomes positively sloped for distance larger than 10.
- There are only 44 of the 3796 observations with distance larger than 10.
- This is approximately 1% of the sample.
- Thus this part of the regression function is very imprecisely estimated.

# Add E8.2f

```
2 . reg ed tuition dist dist2 bytest incomehi ownhome dadcoll ///
  > momcoll DadMomColl cue80 stwmfg black hispanic female, robust

Linear regression                               Number of obs =      3796
                                                F( 14,  3781) =    145.73
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2854
                                                Root MSE      =    1.5363
```

| ed | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| tuition | -.1939714 | .0985584 | -1.97 | 0.049 | -.3872042 | -.0007387 |
| dist | -.0810001 | .025094 | -3.23 | 0.001 | -.1301992 | -.0318011 |
| dist2 | .0046773 | .0020564 | 2.27 | 0.023 | .0006455 | .0087091 |
| bytest | .0925664 | .0030234 | 30.62 | 0.000 | .0866388 | .0984939 |
| incomehi | .3623156 | .0622537 | 5.82 | 0.000 | .2402615 | .4843697 |
| ownhome | .1412131 | .0649487 | 2.17 | 0.030 | .0138752 | .2685511 |
| dadcoll | .6538031 | .087084 | 7.51 | 0.000 | .483067 | .8245392 |
| momcoll | .5693549 | .1218052 | 4.67 | 0.000 | .3305445 | .8081652 |
| DadMomColl | -.3664802 | .1639813 | -2.23 | 0.025 | -.6879805 | -.0449799 |
| cue80 | .0257697 | .00959 | 2.69 | 0.007 | .0069677 | .0445716 |
| stwmfg80 | -.0415432 | .0199035 | -2.09 | 0.037 | -.0805658 | -.0025206 |
| black | .3305619 | .0683148 | 4.84 | 0.000 | .1966244 | .4644994 |
| hispanic | .3297465 | .0779131 | 4.23 | 0.000 | .1769907 | .4825024 |
| female | .1406184 | .0502133 | 2.80 | 0.005 | .0421707 | .2390661 |
| _cons | 9.00197 | .2500197 | 36.01 | 0.000 | 8.511783 | 9.492157 |

The estimated coefficient is -0.366 and measures the extra effect of education above and beyond the separate MomColl and DadColl effects, when both mother and father attended college.

# Add E8.2g

- Difference Jane and Mary is that Janes father attended college while Mary's did not.
- Difference between Alexis and Mary is that Alexis' mother attended college while Mary's did not.
- Difference Bonnie and Mary is that both of Bonnies parents attended college while neither of Mary's parents attended college.

# Add E8.2g

```
1 . di "Jane vs Mary"
Jane vs Mary

2 . lincom _b[dadcoll]

 ( 1)  dadcoll = 0
```

| ed | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| (1) | **.6538031** | **.0843096** | **7.75** | **0.000** | **.4885064** | **.8190998** |

```
3 . di "Alexis vs Mary"
Alexis vs Mary

4 . lincom _b[momcoll]

 ( 1)  momcoll = 0
```

| ed | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| (1) | **.5693549** | **.1172733** | **4.85** | **0.000** | **.3394298** | **.7992799** |

```
5 . di "Bonnie vs Mary"
Bonnie vs Mary

6 . lincom _b[dadcoll]+_b[momcoll]+_b[DadMomColl]

 ( 1)  dadcoll + momcoll + DadMomColl = 0
```

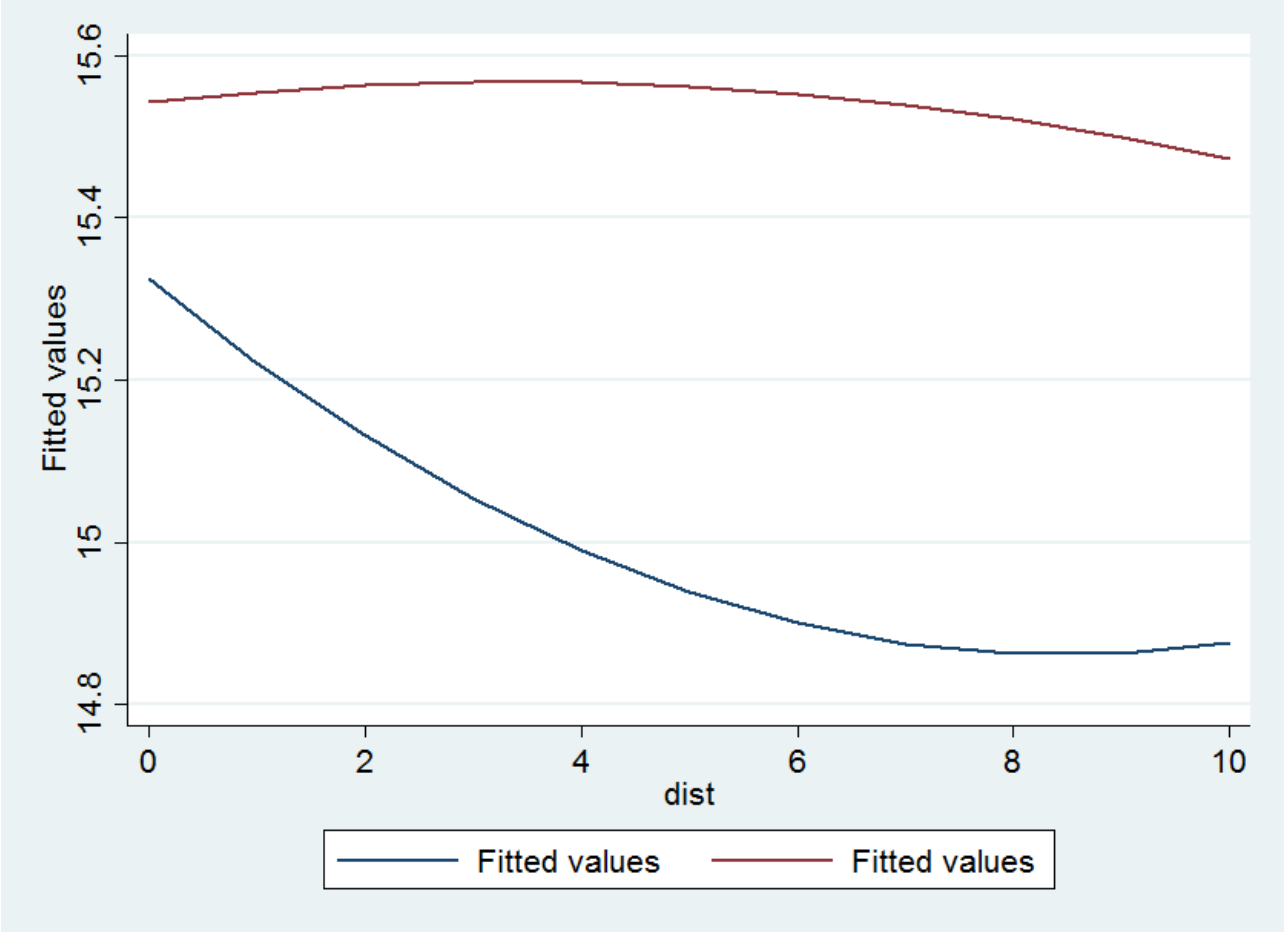
| ed | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| (1) | **.8566778** | **.0947159** | **9.04** | **0.000** | **.6709785** | **1.042377** |

# Add E8.2h

```
1 . *h
2 . gen incomedist = incomehi*dist

3 . gen incomedist2 = incomehi*dist2

4 . reg ed tuition dist dist2 female bytest income* ownhome dadcoll ///
  > momcoll cue80 stwmfg black hispanic DadMomColl, robust

Linear regression                               Number of obs =        3796
                                                F( 16,  3779) =      128.72
                                                Prob > F      =      0.0000
                                                R-squared     =      0.2863
                                                Root MSE      =      1.5357

------------------------------------------------------------------------------
             |               Robust
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tuition |  -.2099784   .0991537    -2.12   0.034    -.4043783   -.0155785
        dist |  -.1095309   .0281269    -3.89   0.000    -.1646763   -.0543855
       dist2 |   .0064744   .0022177     2.92   0.004     .0021264    .0108224
      female |    .141463   .0501943     2.82   0.005     .0430524    .2398736
      bytest |   .0927566   .0030201    30.71   0.000     .0868353    .0986778
    incomehi |   .2172968   .0897228     2.42   0.015      .041387    .3932065
  incomedist |   .1244186   .0620106     2.01   0.045     .0028412     .245996
 incomedist2 |   -.008659    .006246    -1.39   0.166    -.0209049     .003587
     ownhome |   .1437389   .0649888     2.21   0.027     .0163223    .2711554
     dadcoll |   .6627368   .0870109     7.62   0.000      .492144    .8333297
     momcoll |   .5674681   .1219911     4.65   0.000     .3282934    .8066428
       cue80 |   .0260482   .0095869     2.72   0.007     .0072522    .0448443
     stwmfg80|  -.0419249   .0198822    -2.11   0.035    -.0809058    -.002944
       black |    .333128   .0684285     4.87   0.000     .1989677    .4672883
    hispanic |   .3230637   .0777508     4.16   0.000     .1706261    .4755013
  DadMomColl |  -.3556964   .1642177    -2.17   0.030    -.6776602   -.0337326
       _cons |   9.042179   .2508048    36.05   0.000     8.550453    9.533905
------------------------------------------------------------------------------

5 . test incomedist=incomedist2=0

 ( 1)  incomedist - incomedist2 = 0
 ( 2)  incomedist = 0

       F(  2,  3779) =      2.34
            Prob > F =      0.0966
```

The coefficients are jointly significant at the 10% level, but not at the 5% level. (However, incomedist is individually significant so indication that there is a difference.)

# Add E8.2h




The above line is the one for high income parents. Thus there is an indication that there is no effect of distance on education for children of high income parents as the line is almost flat.

- The nonlinear effect of distance on years of education is statistically significant.
- The regression shows a slight effect for non-high income students, but essentially no effect for high income students.

# Add E9.2a

Internal validity:

- Omitted variable bias: Students from wealthier families might live closer to colleges and have higher average years of completed education. Incomehi and Ownhome attempts to control for this, but these are imperfect measures of wealth.
- Misspecification of the functional form: The previous exercise compared different functional forms to find the one that fit the data best.
- Errors in variables: Years of completed education may be imprecisely measured. (It is self-reported education as it is a survey)
- Sample selection: This is a random sample of high school seniors, so sample selection within this population is unlikely to be a problem. However, the results are not necessarily generalizable to the population of high school students as we have not included those who drop out before the senior year.

- Simultaneous causality: Parents who want to send their children to college may locate closer to a college. This is possible, but the effect is likely to be small.
- Inconsistency of standard errors: Heteroskedasticity-robust standard errors were used. The data represents a random sample so that correlation across the error terms is not a problem. Thus the standard errors should be consistent.

# ADD E9.2b

```
1 . u "http://wps.aw.com/wps/media/objects/3254/3332253/datasets2e/datasets/CollegeDistanceWest.dta", cl

2 . gen dist2 = dist^2

3 .
4 . gen incomedist = incomehi*dist

5 . gen incomedist2 = incomehi*dist2

6 . gen DadMomColl = dadcoll*momcoll

7 .
8 . reg ed tuition dist dist2 female bytest income* ownhome dadcoll ///
  > momcoll cue80 stwmfg black hispanic DadMomColl, robust

Linear regression                               Number of obs =         943
                                                   F( 16,   926) =    22.15
                                                   Prob > F      =   0.0000
                                                   R-squared     =   0.2312
                                                   Root MSE      =   1.4888
```

| ed | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| tuition | -.5226686 | .2426253 | -2.15 | 0.031 | -.9988279 | -.0465093 |
| dist | -.0916798 | .0448716 | -2.04 | 0.041 | -.1797416 | -.003618 |
| dist2 | .0040874 | .003112 | 1.31 | 0.189 | -.0020201 | .0101948 |
| female | .0505693 | .099252 | 0.51 | 0.611 | -.1442157 | .2453544 |
| bytest | .0732997 | .0064989 | 11.28 | 0.000 | .0605453 | .0860541 |
| incomehi | .4070321 | .168813 | 2.41 | 0.016 | .0757317 | .7383326 |
| incomedist | .0045501 | .0903815 | 0.05 | 0.960 | -.1728262 | .1819264 |
| incomedist2 | -.0000224 | .0056678 | -0.00 | 0.997 | -.0111457 | .0111008 |
| ownhome | .1992296 | .1266021 | 1.57 | 0.116 | -.0492307 | .4476898 |
| dadcoll | .4412696 | .1447355 | 3.05 | 0.002 | .1572219 | .7253173 |
| momcoll | .283049 | .2629621 | 1.08 | 0.282 | -.2330218 | .7991197 |
| cue80 | .0452626 | .0227165 | 1.99 | 0.047 | .0006809 | .0898444 |
| stwmfg80 | .0307996 | .044474 | 0.69 | 0.489 | -.056482 | .1180811 |
| black | .0671427 | .1816453 | 0.37 | 0.712 | -.2893415 | .423627 |
| hispanic | .1955382 | .1156337 | 1.69 | 0.091 | -.0313963 | .4224726 |
| DadMomColl | .1422522 | .3295428 | 0.43 | 0.666 | -.504485 | .7889895 |
| _cons | 9.227512 | .523652 | 17.62 | 0.000 | 8.19983 | 10.25519 |

# ADD E9.2b

```
1 . est sto reg6

2 . esttab reg5 reg6, se
```

| | (1) ed | (2) ed |
|---|---|---|
| tuition | -0.210* | -0.523* |
| | (0.0992) | (0.243) |
| dist | -0.110*** | -0.0917* |
| | (0.0281) | (0.0449) |
| dist2 | 0.00647** | 0.00409 |
| | (0.00222) | (0.00311) |
| female | 0.141** | 0.0506 |
| | (0.0502) | (0.0993) |
| bytest | 0.0928*** | 0.0733*** |
| | (0.00302) | (0.00650) |
| incomehi | 0.217* | 0.407* |
| | (0.0897) | (0.169) |
| incomedist | 0.124* | 0.00455 |
| | (0.0620) | (0.0904) |
| incomedist2 | -0.00866 | -0.0000224 |
| | (0.00625) | (0.00567) |
| ownhome | 0.144* | 0.199 |
| | (0.0650) | (0.127) |
| dadcoll | 0.663*** | 0.441** |
| | (0.0870) | (0.145) |
| momcoll | 0.567*** | 0.283 |
| | (0.122) | (0.263) |
| cue80 | 0.0260** | 0.0453* |
| | (0.00959) | (0.0227) |
| stwmfg80 | -0.0419* | 0.0308 |
| | (0.0199) | (0.0445) |
| black | 0.333*** | 0.0671 |
| | (0.0684) | (0.182) |
| hispanic | 0.323*** | 0.196 |
| | (0.0778) | (0.116) |
| DadMomColl | -0.356* | 0.142 |
| | (0.164) | (0.330) |
| _cons | 9.042*** | 9.228*** |
| | (0.251) | (0.524) |
| N | 3796 | 943 |

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- The sample from the West contains 943 observations compared to the 3796 observations in the Non-west sample. Thus the estimates in the sample from West will be more imprecise, i.e. larger standard errors.
- Because the samples are independent the standard errors for the estimated difference in the coefficients can be calculated as:

$$SE(\hat{\beta}_{NW} - \hat{\beta_W}) = \sqrt{SE(\hat{\beta}_{NW})^2 + SE(\hat{\beta_W})^2}$$

- Thus for Dist this is $\sqrt{0.028^2 + 0.045^2} = 0.053$
- The coefficients on Dist and Dist2 in the West are very similar to the values for the non-West.
- The interaction terms IncomeHi*dist and Incomehi*Dist2 looks different.
- In the Western sample the coefficients in the West for students with incomehi is very similar to the regression function for students with incomehi=0. (But this difference is not statistically significant)
- The only statistically significant coefficient across the two samples is the bytest score with a difference of (0.093-0.073)=0.20 and a standard error of $\sqrt{0.003^2 + 0.006^2} = 0.0067$