# The Multivariate Regression Model



## Example 1
## Determinants of College GPA

- Sample of 141 Freshman
- Collect data on College GPA (4.0 scale)
- Look at importance of ACT
- Consider the following model

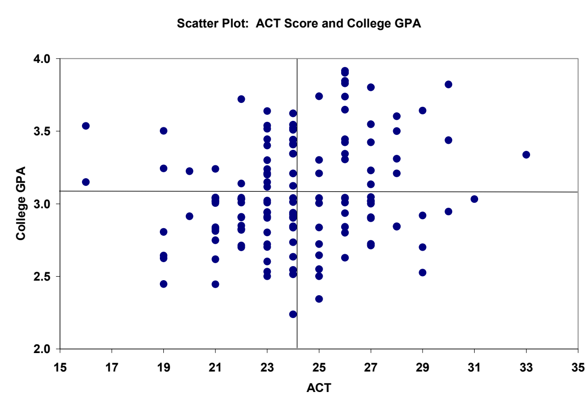
$$CGPA_i = \beta_0 + ACT_i\beta_1 + \varepsilon_i$$

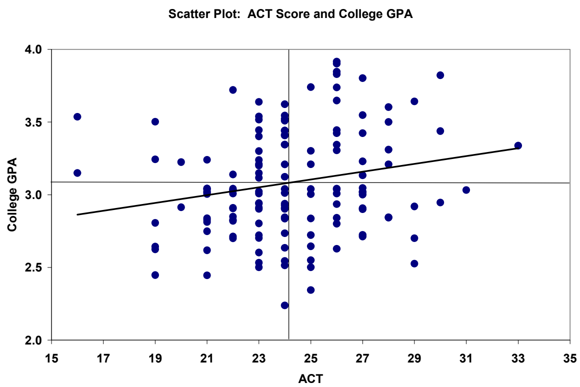


## ACT

- 4 tests
  - English/math/reading/science reasoning
- Composite scores from 1-36
- Average score in 2000 was 21
- Movement from 21 to 22 represents 7 percentage points in the distribution (56$^{th}$ to 63th percentile)








5

```
. *run regression with one variable
. reg college_gpa act

      Source |       SS       df       MS              Number of obs =     141
-------------+------------------------------           F(  1,   139) =    6.21
       Model |  .829558811     1  .829558811           Prob > F      =  0.0139
    Residual |  18.5765406   139  .133644177           R-squared     =  0.0427
-------------+------------------------------           Adj R-squared =  0.0359
       Total |  19.4060994   140  .138614996           Root MSE      =  .36557

------------------------------------------------------------------------------
 college_gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         act |    .027064   .0108628     2.49   0.014     .0055862    .0485417
       _cons |   2.402979   .2642027     9.10   0.000     1.880604    2.925355
------------------------------------------------------------------------------
```

Interpret the result:

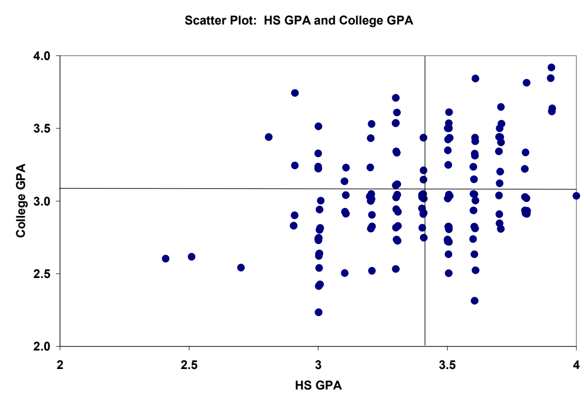
6

## Is this an accurate estimate of $\partial(\text{CGPA})/\partial(\text{ACT})$?

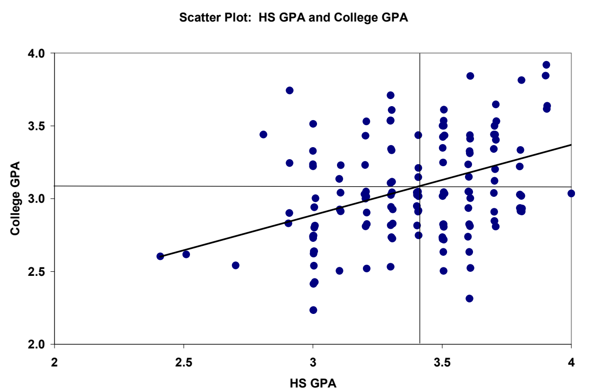
- ACT is but one measure of ability
- "Noisy" measure at best
- Are there other measures available?
- Consider another model (Think of this as the true model)

$$CGPA_i = \beta_0 + ACT_i\beta_1 + HSGPA_i\beta_2 + \varepsilon_i$$

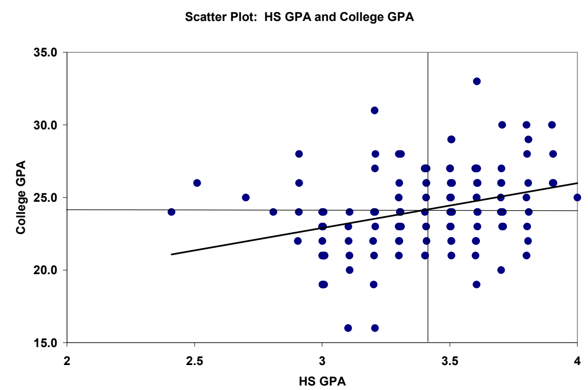
7




8




9




10

```
. * get correlations between key variables
. corr college_gpa act hs_gpa
(obs=141)

             | colleg~a      act   hs_gpa
-------------+---------------------------
 college_gpa |   1.0000
         act |   0.2068   1.0000
      hs_gpa |   0.4146   0.3458   1.0000
```

11

```
*run synthetic regression of hs_gpa on act
reg hs_gpa act

    Source |       SS       df       MS              Number of obs =     141
-----------+------------------------------           F(  1,   139) =   18.88
     Model |  1.71352621     1  1.71352621           Prob > F      =  0.0000
  Residual |   12.615835   139  .090761403           R-squared     =  0.1196
-----------+------------------------------           Adj R-squared =  0.1132
     Total |  14.3293612   140   .10235258           Root MSE      =  .30127

------------------------------------------------------------------------------
    hs_gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
       act |   .0388968    .008952     4.35   0.000     .0211971    .0565964
     _cons |   2.462537   .2177273    11.31   0.000     2.032051    2.893022
------------------------------------------------------------------------------
```

12

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \hat{\delta}_1$$

$$x_{2i} = \delta_0 + x_{1i}\delta_1 + \zeta_i$$

$$(7) \qquad \hat{\delta}_1 = \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1)}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$$



*we anticipate that* $\beta_2 > 0$

*and we have shown that* $\hat{\delta}_2 > 0$

$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \hat{\delta}_1$

*then*

$E[\tilde{\beta}_1] > \beta_1$

On average, the value we estimated in the "False" model will be greater than the one in the "true" model



```
. * run multivariate regression
. reg college_gpa act hs_gpa

      Source |       SS       df       MS              Number of obs =     141
-------------+------------------------------           F(  2,   138) =   14.78
       Model |  3.42365506     2  1.71182753           Prob > F      =  0.0000
    Residual |  15.9824444   138  .115814814           R-squared     =  0.1764
-------------+------------------------------           Adj R-squared =  0.1645
       Total |  19.4060994   140  .138614996           Root MSE      =  .34032

------------------------------------------------------------------------------
 college_gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         act |    .009426   .0107772     0.87   0.383    -.0118838    .0307358
      hs_gpa |   .4534559   .0958129     4.73   0.000     .2640047    .6429071
       _cons |   1.286328   .3408221     3.77   0.000      .612419    1.960237
------------------------------------------------------------------------------
```

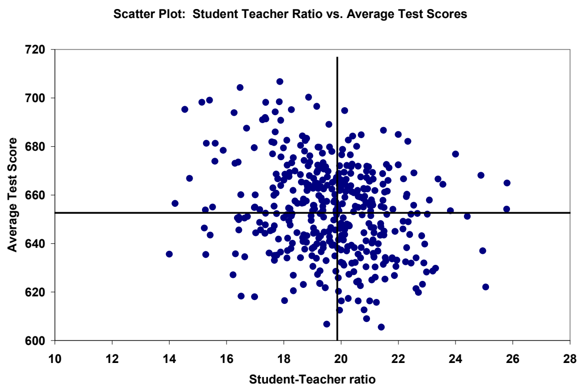
The coefficient on ACT in the "false" model was 0.039
The coefficient in the "True" model is 0.009—the coefficient falls by 77%



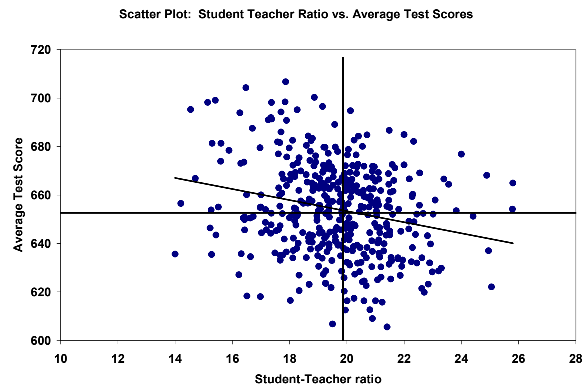
## Example 2: Class Size and Performance

- Data from 420 schools in CA
- Outcome is average on state test for reading and math in 6th grade
- Average scores around 650 for state
- Key covariate: student/teacher ratio

$$SCORE_i = \beta_0 + STR_i\beta_1 + \varepsilon_i$$




17




18

```
. * run regression with one variable
. reg average_score student_teacher

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  1,   418) =   22.58
       Model |  7794.11004     1  7794.11004           Prob > F      =  0.0000
    Residual |  144315.484   418  345.252353           R-squared     =  0.0512
-------------+------------------------------           Adj R-squared =  0.0490
       Total |  152109.594   419  363.030056           Root MSE      =  18.581

------------------------------------------------------------------------------
average_sc~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
student_te~r |  -2.279808   .4798256    -4.75   0.000    -3.22298   -1.336637
       _cons |    698.933   9.467491    73.82   0.000    680.3231    717.5428
------------------------------------------------------------------------------
```

# Omitted variables

- Class size is but one covariate we could add
- Consider others that might be correlated with X that are omitted from model
- Example: % ESL
  - These students tend to score lower on tests
  - If they are also more or less likely to be in more crowded schools, then results could be biased

20

$$SCORE_i = \beta_0 + STR_i\beta_1 + ESL_i\beta_2 + \varepsilon_i$$

Think of this as the "true" model

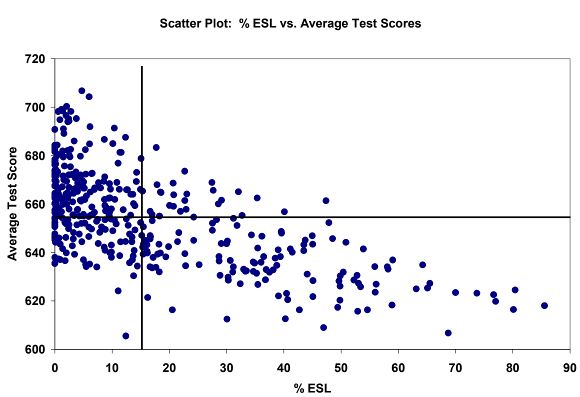
21

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2\hat{\delta}_1$$

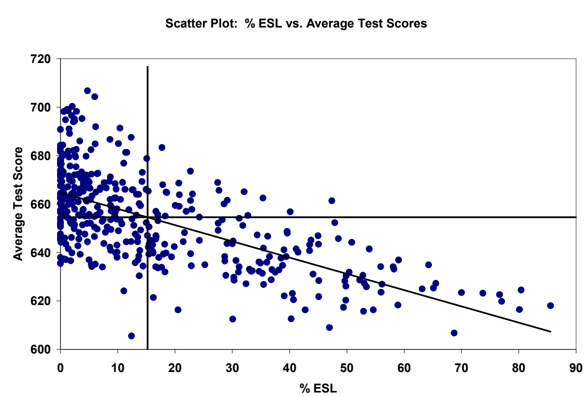
$$x_{2i} = \delta_0 + x_{1i}\delta_1 + \zeta_i$$

(7) $$\hat{\delta}_1 = \frac{\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)(x_{1i} - \overline{x}_1)}{\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)^2}$$

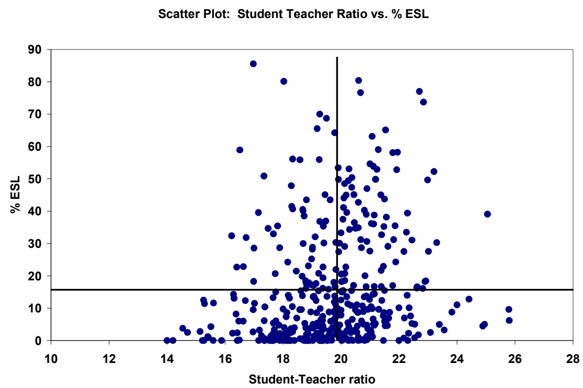
22



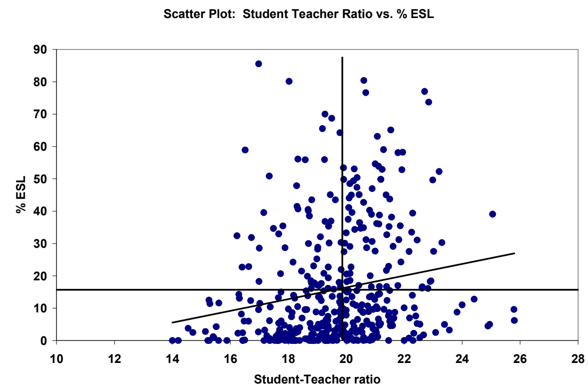

23




24




25




26

```
             | averag~e studen~r  esl_pct
-------------+---------------------------
average_sc~e |   1.0000
student_te~r |  -0.2264   1.0000
     esl_pct |  -0.6441   0.1876   1.0000
```

27

*we anticipate that* $\beta_2 < 0$

*and we have shown that* $\hat{\delta}_2 > 0$

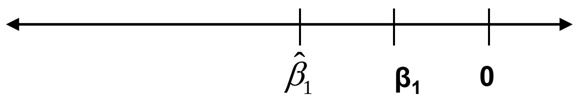
$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \hat{\delta}_1$$

*then*

$$E[\tilde{\beta}_1] < \beta_1$$

On average, the value we estimated in the “False” model will be smaller than the one in the “true” model

28




$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \hat{\delta}_1$$

$$(-) + (-)(+)$$

29

## Think of the prediction this way

- In the single variable model, the Student/teacher ratio is picking up two effects
  - Larger class sizes reduce performance
  - ESL students are more likely to be in more crowded schools, and they that tend to have lower scores
- Therefore, the model without ESL will estimate a too large of a negative number

30

```
. * run multivariate regression
. reg average_score student_teacher esl_pct

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  2,   417) =  155.01
       Model |  64864.3011     2  32432.1506           Prob > F      =  0.0000
    Residual |  87245.2925   417  209.221325           R-squared     =  0.4264
-------------+------------------------------           Adj R-squared =  0.4237
       Total |  152109.594   419  363.030056           Root MSE      =  14.464

------------------------------------------------------------------------------
average_sc~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
student_te~r |  -1.101296   .3802783    -2.90   0.004    -1.848797   -.3537945
     esl_pct |  -.6497768   .0393425   -16.52   0.000    -.7271112   -.5724423
       _cons |   686.0322   7.411312    92.57   0.000     671.4641    700.6004
------------------------------------------------------------------------------
```

**5 student increase in class size reduces test scores by 5(1.1) = 5.5 which is 5.5/654= 0.008 or .8% -- half the Estimate impact as before**

**A one percentage point increase in % ESL in school Will reduce average scores by .64 points**

31

```
. * demonstrate the partialing out
. * nature of mv regressions
.
. * run a regression of STR on ESL
. * output the residuals
. reg student_teacher esl

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  1,   418) =   15.25
       Model |  52.7997281     1  52.7997281           Prob > F      =  0.0001
    Residual |  1446.78109   418  3.46119878           R-squared     =  0.0352
-------------+------------------------------           Adj R-squared =  0.0329
       Total |  1499.58082   419  3.57895184           Root MSE      =  1.8604

------------------------------------------------------------------------------
student_te~r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     esl_pct |    .019413   .0049704     3.91   0.000     .0096429     .029183
       _cons |   19.33432   .1199307   161.21   0.000     19.09858    19.57006
------------------------------------------------------------------------------

.
. * output residuals
. predict res_str, residual
```

32

```
. * run a regression of test scores
. * on the student_teacher residuals
. reg average_score res_str

      Source |       SS       df       MS              Number of obs =     4
-------------+------------------------------           F(  1,   418) =    4.
       Model |  1754.73229     1  1754.73229           Prob > F      =  0.02
    Residual |  150354.861   418  359.700625           R-squared     =  0.01
-------------+------------------------------           Adj R-squared =  0.00
       Total |  152109.594   419  363.030056           Root MSE      =  18.9

------------------------------------------------------------------------------
average_sc~e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interva
-------------+----------------------------------------------------------------
     res_str |  -1.101296   .4986194    -2.21   0.028     -2.08141   -.12118
       _cons |   654.1565   .9254351   706.86   0.000     652.3375    655.97
------------------------------------------------------------------------------
```

Exact same number as before



# school_districts_2000.dta

- Data on spending/pupil and revenues/pupil for 10,279 school districts in 2000
- Schools are funded with local, state and federal dollars
- Local revenues are usually from the property tax
- State and federal dollars are usually transferred to districts based on need – poorer districts get more



# Variables

| Variable | Label |
|---|---|
| exp_pupil | Real expenditures per pupil |
| med_fam_inc | Real median family income |
| sf_rev_pupil | State/federal revenues per pupil |
| per_under_20 | Percent of the district population under 20 |
| schools | # of schools in the district |



# Summary Statistics

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| exp_pupil | 10,729 | 7718.393 | 2069.6 | 4098.78 | 40315.48 |
| med_fam_inc | 10,729 | 53101.24 | 19562.17 | 17453.32 | 215967.5 |
| sf_rev_pupil | 10,729 | 5139.467 | 2157.409 | 157.2581 | 22967.97 |
| schools | 10,729 | 7.773511 | 20.28197 | 1 | 1164 |
| per_under_20 | 10,729 | .2857616 | .0390008 | .0813769 | .5089928 |

```
      Source |       SS           df       MS      Number of obs   =    10,729
-------------+----------------------------------   F(1, 10727)     =   1580.02
       Model |  5.8993e+09         1  5.8993e+09   Prob > F        =    0.0000
    Residual |  4.0051e+10    10,727  3733693.91   R-squared       =    0.1284
-------------+----------------------------------   Adj R-squared   =    0.1283
       Total |  4.5951e+10    10,728  4283244.49   Root MSE        =    1932.3

------------------------------------------------------------------------------
   exp_pupil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 med_fam_inc |   .0379074   .0009537    39.75   0.000     .0360381    .0397768
       _cons |   5705.462   53.96718   105.72   0.000     5599.676    5811.248
------------------------------------------------------------------------------
```

Interpret the coefficient on med_fam_inc

37

$$exp\_pupil_i = \beta_0 + med\_fam\_inc_i \beta_1 + sf\_rev_i \beta_2 + \varepsilon_i$$

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$$

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \hat{\gamma}_1$$

what do we expect for $\beta_2$?

what do we expect for $\hat{\gamma}_1$?

38

```
. corr exp_pupil med_fam_inc sf_rev_pupil
(obs=10,729)

             | exp_pu~l med_fa~c sf_rev~l
-------------+---------------------------
   exp_pupil |   1.0000
 med_fam_inc |   0.3583   1.0000
sf_rev_pupil |   0.1988  -0.3871   1.0000
```

39

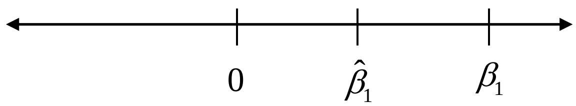
## Some text

- $\beta_2$ should be positive – districts will spend more if they receive more resources from state and federal sources
- What about $\hat{\gamma}_1$? State and Federal dollars are usually redistriobutionary. They tend top go to the districts with the highest need – so – we expect $\hat{\gamma}_1 < 0$

40

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \hat{\gamma}_1$$

$\beta_1 > 0$ $\quad \beta_2 > 0$ $\quad \hat{\gamma}_2 < 0$




$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \hat{\gamma}$$

$$(+) + (+)(-)$$

$$E[\tilde{\beta}_1] < \beta_1$$

41

```
      Source |       SS           df       MS      Number of obs   =    10,729
-------------+----------------------------------   F(2, 10726)     =   1907.22
       Model |  1.2054e+10         2  6.0272e+09   Prob > F        =    0.0000
    Residual |  3.3896e+10    10,726  3160196.73   R-squared       =    0.2623
-------------+----------------------------------   Adj R-squared   =    0.2622
       Total |  4.5951e+10    10,728  4283244.49   Root MSE        =    1777.7

------------------------------------------------------------------------------
   exp_pupil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 med_fam_inc |   .0541612   .0009515    56.92   0.000      .052296    .0560263
sf_rev_pupil |   .3807734   .0086279    44.13   0.000      .363861    .3976858
       _cons |   2885.396   80.92154    35.66   0.000     2726.775    3044.017
------------------------------------------------------------------------------
```

42

## Partialing our properties

```
 Source |       SS           df       MS      Number of obs   =    10,729
-------------+----------------------------------   F(4, 10724)     =   1228.13
       Model |  1.4436e+10         4  3.6091e+09   Prob > F        =    0.0000
    Residual |  3.1514e+10    10,724  2938671.34   R-squared       =    0.3142
-------------+----------------------------------   Adj R-squared   =    0.3139
       Total |  4.5951e+10    10,728  4283244.49   Root MSE        =    1714.3

------------------------------------------------------------------------------
   exp_pupil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 med_fam_inc |   .0541014   .0009176    58.96   0.000     .0523027       .0559
sf_rev_pupil |    .417348   .0084248    49.54   0.000     .4008338    .4338621
per_under_20 |  -12172.15   430.7222   -28.26   0.000    -13016.44   -11327.85
     schools |  -2.264741   .8165082    -2.77   0.006    -3.865248    -.664234
       _cons |   6196.533   140.2428    44.18   0.000     5921.631    6471.435
------------------------------------------------------------------------------
```

Remember the coef. on med_fam_inc which is 0.054

43

```
*regress med_fam_inc on other x's
reg med_fam_inc sf_rev_pupi per_under_20 schools

*output residuals
predict r_medfaminc, residuals

*Regress exp_pupil on residuals
reg exp_pupil r_medfaminc
```

44

```
  Source |       SS           df       MS      Number of obs   =    10,729
-------------+----------------------------------   F(1, 10727)     =   3066.57
       Model |  1.0216e+10         1  1.0216e+10   Prob > F        =    0.0000
    Residual |  3.5735e+10    10,727  3331310.01   R-squared       =    0.2223
-------------+----------------------------------   Adj R-squared   =    0.2222
       Total |  4.5951e+10    10,728  4283244.49   Root MSE        =    1825.2

------------------------------------------------------------------------------
   exp_pupil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 r_medfaminc |   .0541014    .000977    55.38   0.000     .0521864    .0560164
       _cons |   7718.393   17.62089   438.03   0.000     7683.853    7752.933
------------------------------------------------------------------------------

  Source |       SS           df       MS      Number of obs   =    10,729
-------------+----------------------------------   F(4, 10724)     =   1228.13
       Model |  1.4436e+10         4  3.6091e+09   Prob > F        =    0.0000
    Residual |  3.1514e+10    10,724  2938671.34   R-squared       =    0.3142
-------------+----------------------------------   Adj R-squared   =    0.3139
       Total |  4.5951e+10    10,728  4283244.49   Root MSE        =    1714.3

------------------------------------------------------------------------------
   exp_pupil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 med_fam_inc |   .0541014   .0009176    58.96   0.000     .0523027       .0559
sf_rev_pupil |    .417348   .0084248    49.54   0.000     .4008338    .4338621
per_under_20 |  -12172.15   430.7222   -28.26   0.000    -13016.44   -11327.85
     schools |  -2.264741   .8165082    -2.77   0.006    -3.865248    -.664234
       _cons |   6196.533   140.2428    44.18   0.000     5921.631    6471.435
------------------------------------------------------------------------------
```



# Interpretation

- The variation in $x_{1i}$ that is used to generate the estimate for $\beta_1$ is only that variation in $x_{1i}$ that is NOT predicted by the other variables in the system
- The less residual variation on $x_{1i}$ the more difficult it will be extract information about the impact of $x_1$ on y