Data Analysis and Statistical Software

Regression Models For Nonrandom Treatment Assignment, Selection Bias, and Unobserved Confounding Using Stata

Chuck Huber, PhD
StataCorp
chuber@stata.com

University of California, Los Angeles
February 22, 2018

Extended Regression Models (ERMs) in Stata 15

Chuck Huber, PhD
StataCorp
chuber@stata.com

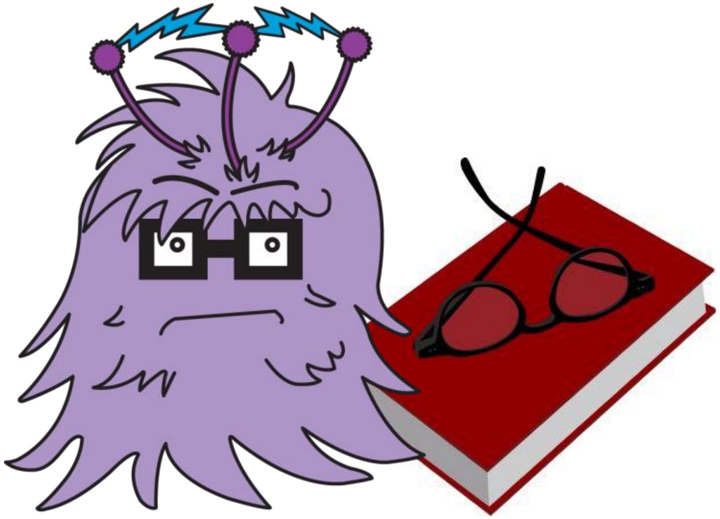
University of California, Los Angeles
February 22, 2018

Outline

- Description of the dataset
- Endogenous Covariates
- Nonrandom treatment assignment
- Missing not at random (MNAR) and selection bias
- Treatment effects

The Research Question

- Fictional State University (FSU) has developed a new study-skills program with the goal of improving the grade point averages of their students.



Stata Data Analysis and Statistical Software

The Data

```
. use gpa.dta, clear
(Simulated GPA Dataset for ERMs seminars)


. describe

Contains data from gpa.dta
  obs:         1,000                          Simulated GPA Dataset for ERMs seminars
 vars:             9                          22 Jan 2018 16:06
 size:        22,000                          (_dta has notes)
---------------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
---------------------------------------------------------------------------------------------
id              int     %9.0g                 Student Identification Number
gpa             float   %9.0g                 Final College Grade Point Average
hsgpa           float   %9.0g                 High School Grade Point Average
program         byte    %9.0g      YesNo      Student participated in the study skills program?
graduate        byte    %9.0g      YesNo      Did the student graduate college?
income          float   %9.0g                 Parent's Income (x $100,000)
hs_comp         float   %9.0g                 High School Competitiveness
roommate        byte    %9.0g      YesNo      Students's roommate is also a student?
scholarship     byte    %9.0g      YesNo      Student received scholarship funds?
---------------------------------------------------------------------------------------------
Sorted by: id
```

The Data

```
. summarize
```

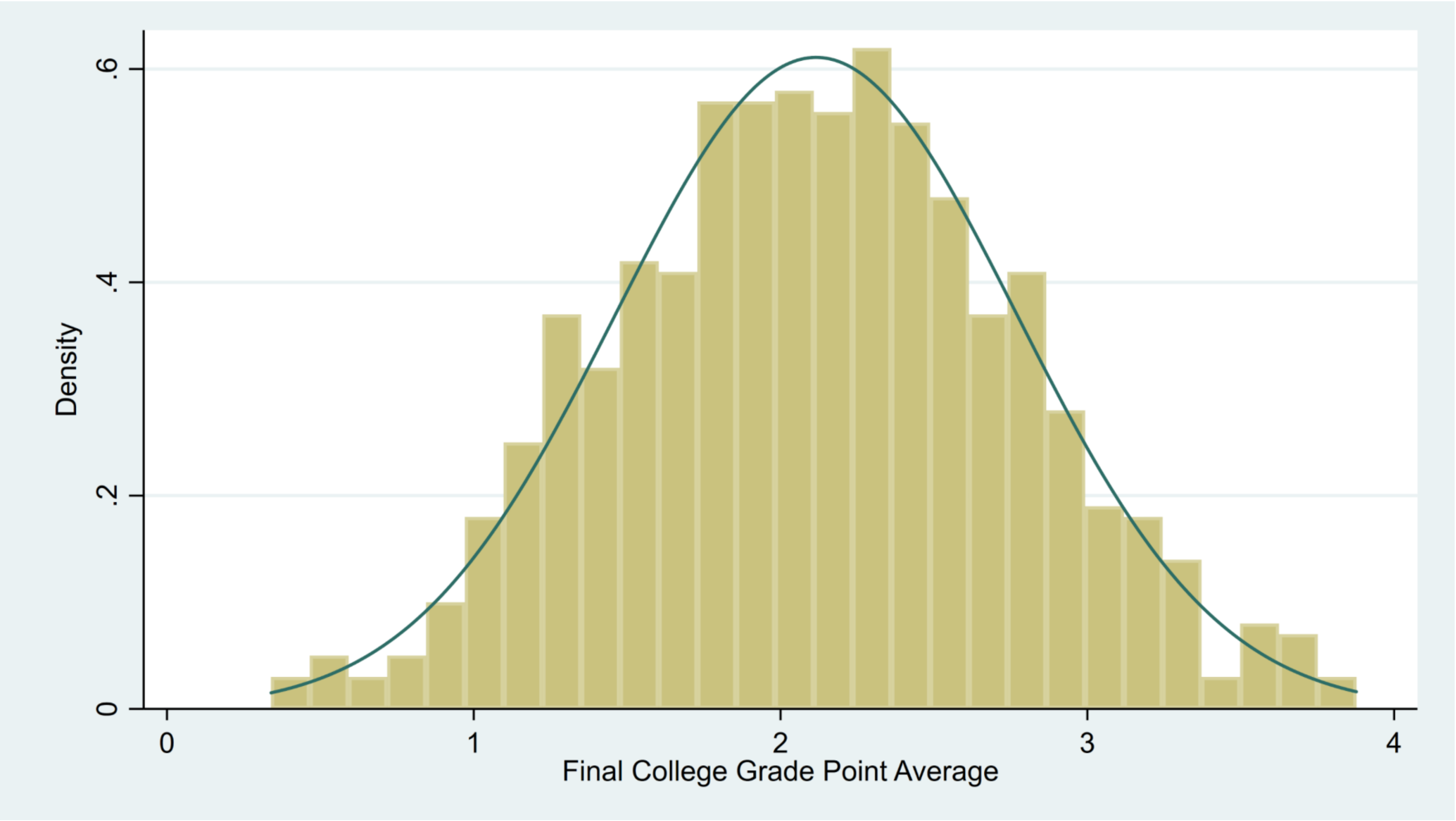
Variable	Obs	Mean	Std. Dev.	Min	Max
id	1,000	500.5	288.8194	1	1000
gpa	792	2.115962	.6529961	.3392706	3.876919
hsgpa	1,000	2.294384	.5714525	.6758502	3.786486
program	1,000	.3	.4584869	0	1
graduate	1,000	.792	.4060799	0	1
income	1,000	.5031867	.2848887	.0004344	.9969745
hs_comp	1,000	.4946027	.286164	.0001878	.9985294
roommate	1,000	.321	.4670944	0	1
scholarship	1,000	.32	.4667096	0	1

The Data

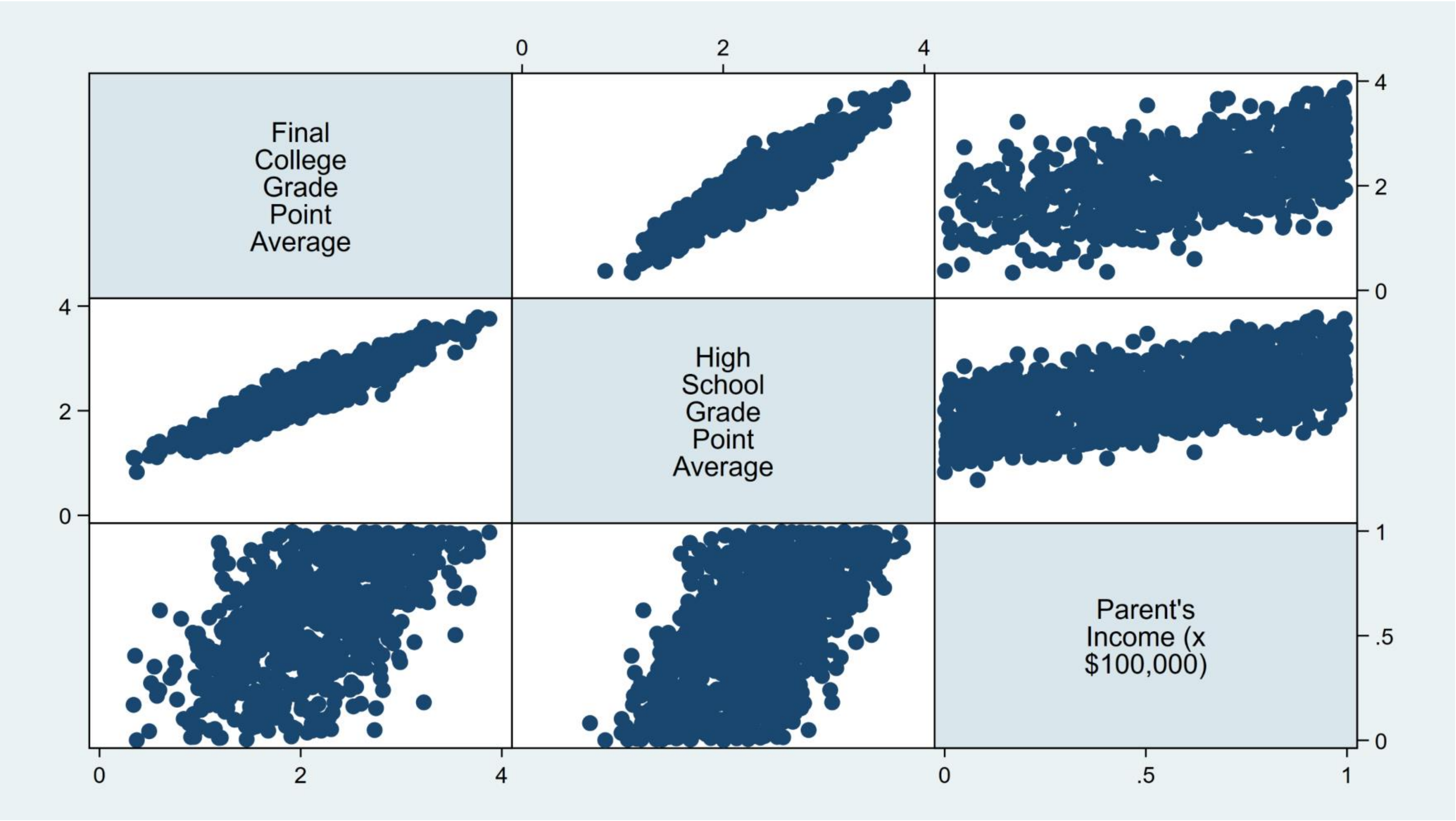
```
. tab graduate
```

Did the student graduate college?	Freq.	Percent	Cum.
No	208	20.80	20.80
Yes	792	79.20	100.00
Total	1,000	100.00	

The Data



The Data

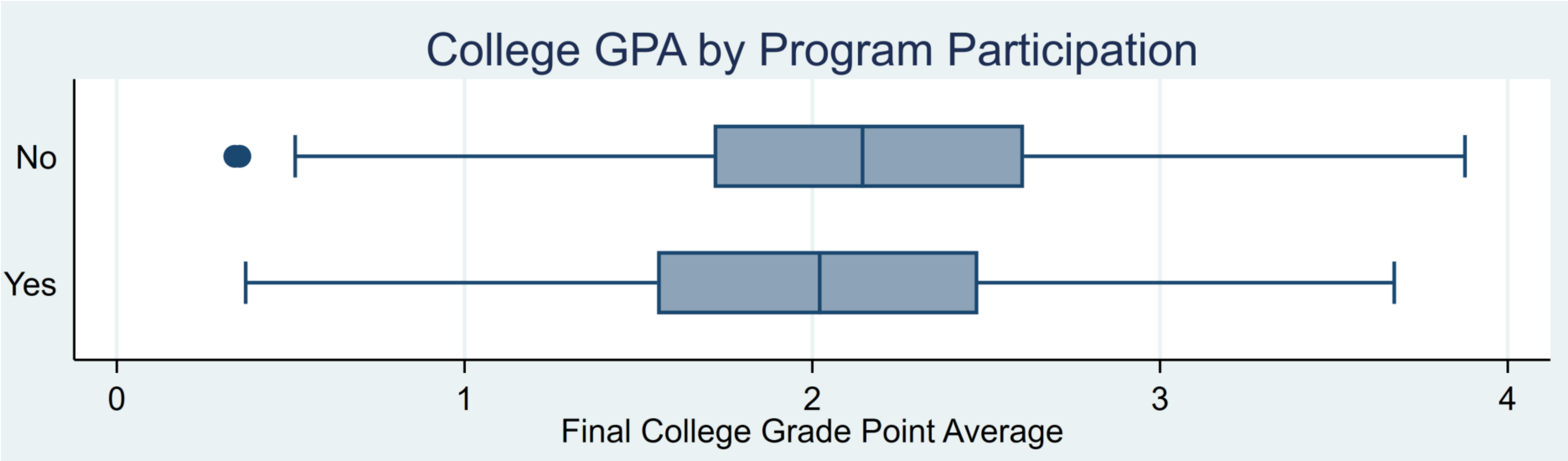

Final
College
Grade
Point
Average
High
School
Grade
Point
Average
Parent's
Income (x
$100,000)
0
2
4
.5
1

The Data

```
. tab program
```

Student participated in the study skills program?	Freq.	Percent	Cum.
No	700	70.00	70.00
Yes	300	30.00	100.00
Total	1,000	100.00	

The Data

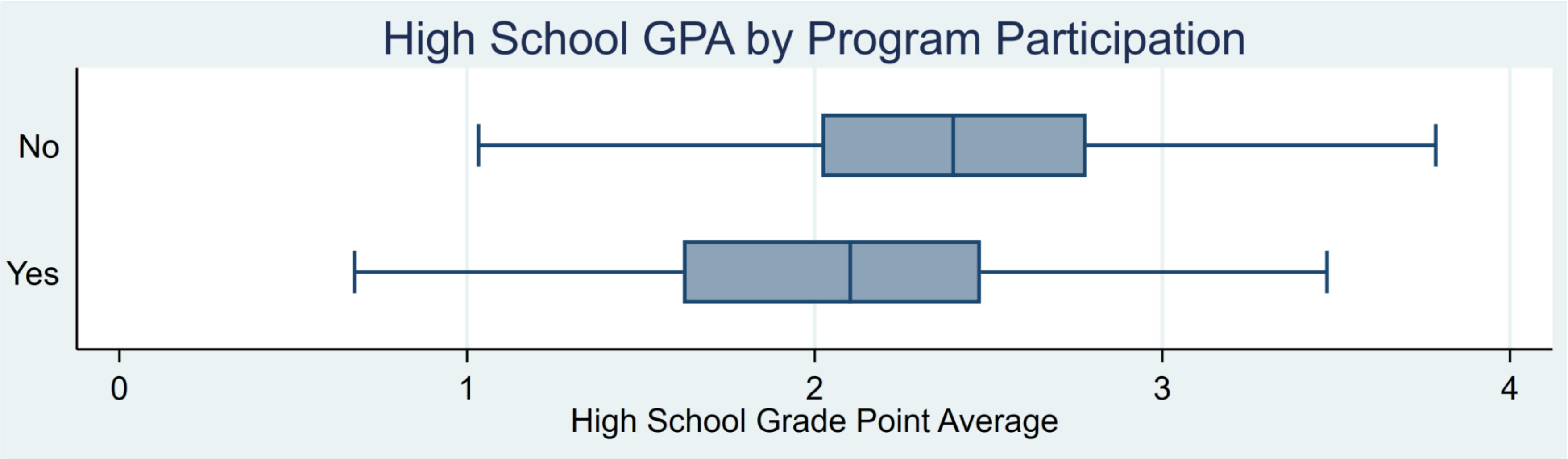

College GPA by Program Participation
No
Yes
0
1
2
3
4
Final College Grade Point Average

The Data

```
. regress gpa i.program

      Source |       SS           df       MS      Number of obs   =       792
-------------+----------------------------------   F(1, 790)       =      5.74
       Model |  2.43242384         1  2.43242384   Prob > F        =    0.0168
    Residual |  334.853048       790  .423864618   R-squared       =    0.0072
-------------+----------------------------------   Adj R-squared   =    0.0060
       Total |  337.285472       791  .426403884   Root MSE        =    .65105

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     program |
        Yes  |  -.1259343     .05257    -2.40   0.017    -.2291278   -.0227409
       _cons |   2.149036   .0269406    79.77   0.000     2.096152    2.201919
------------------------------------------------------------------------------

. estimates store univar
```

Students who participated in the program had **lower** GPAs?!?!?

The Data


High School GPA by Program Participation
No
Yes
0
1
2
3
4
High School Grade Point Average

The Data

```
. regress gpa i.program hsgpa

      Source |       SS           df       MS      Number of obs   =       792
-------------+----------------------------------   F(2, 789)       =   3350.31
       Model |  301.753841         2  150.876921   Prob > F        =    0.0000
    Residual |  35.5316304       789  .045033752   R-squared       =    0.8947
-------------+----------------------------------   Adj R-squared   =    0.8944
       Total |  337.285472       791  .426403884   Root MSE        =    .21221

------------------------------------------------------------------------------
         gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     program |
         Yes |   .2002776   .0175963    11.38   0.000     .1657364    .2348187
       hsgpa |   1.144457   .0140378    81.53   0.000     1.116901    1.172013
       _cons |  -.6744815    .035729   -18.88   0.000    -.7446166   -.6043464
------------------------------------------------------------------------------

. estimates store hsgpa
```

Students who participated in the program had higher GPAs when we account for high school GPA.

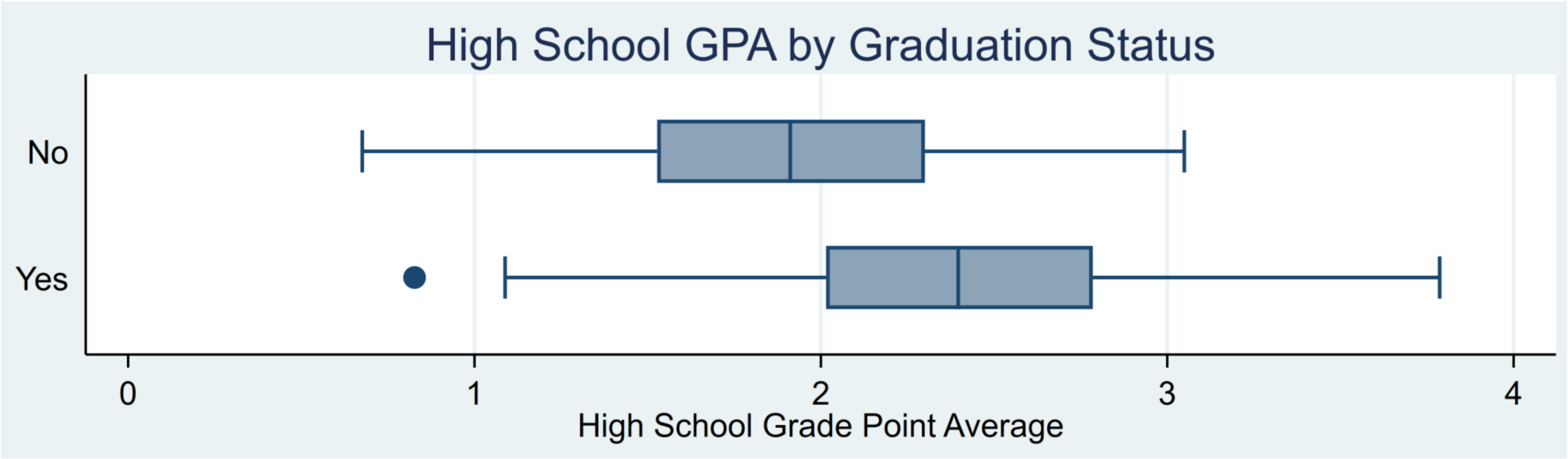
The Data

```
. tab program graduate, row
```

Key
frequency
row percentage

Student participated in the study skills program?	Did the student graduate college? No	Yes	Total
No	116 16.57	584 83.43	700 100.00
Yes	92 30.67	208 69.33	300 100.00
Total	208 20.80	792 79.20	1,000 100.00

The Data


High School GPA by Graduation Status
No
Yes
0
1
2
3
4
High School Grade Point Average

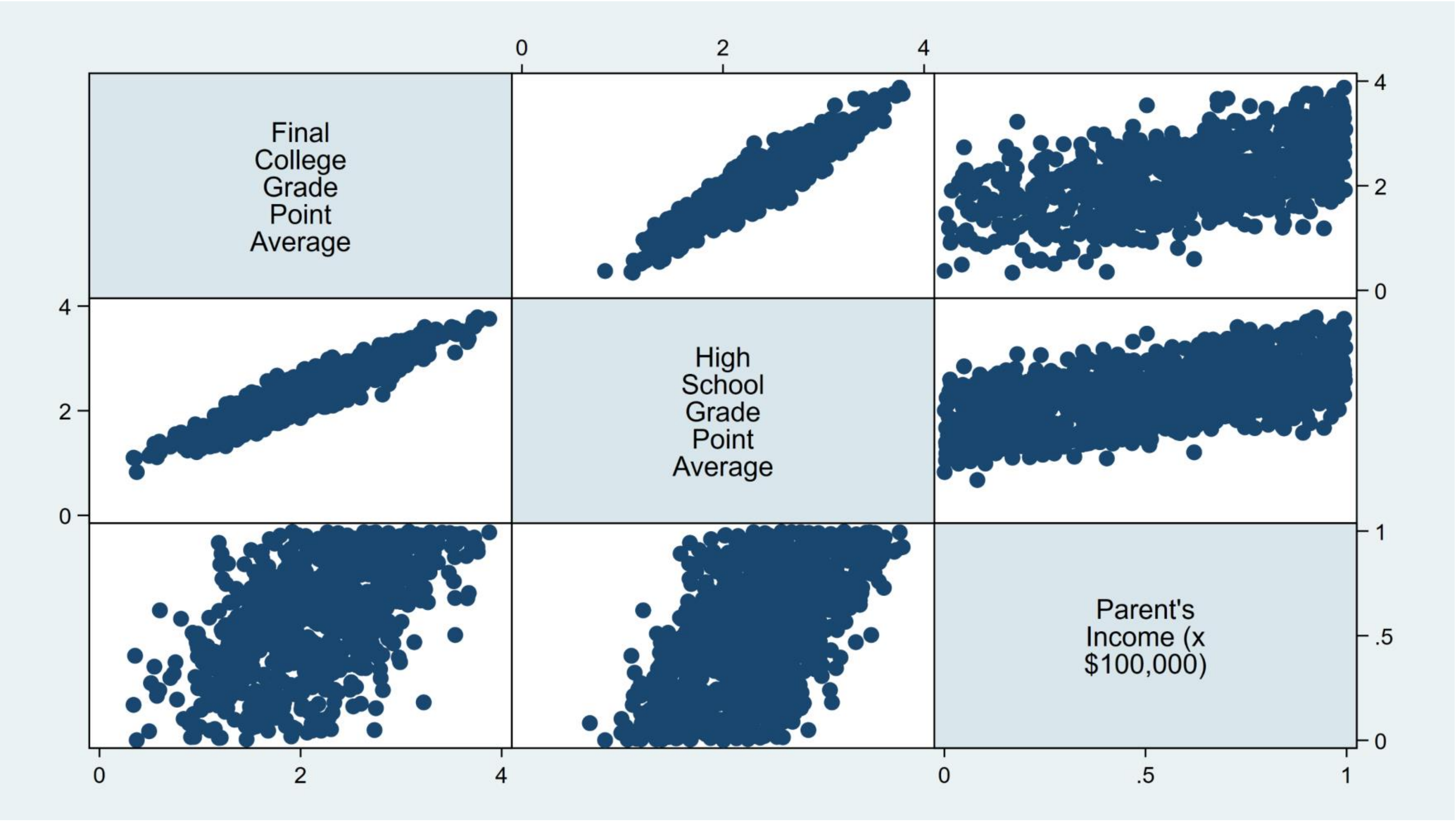
The Data

What was the effect of the study program on students GPAs?

Observational Data

- Observational data often have one or more of these issues:
  - Unobserved confounding and endogeneity.
  - Nonrandom treatment assignment (or exposure)
  - Data that are "missing not at random" (MNAR) which can lead to selection bias

The Data



Final College Grade Point Average
High School Grade Point Average
Parent's Income (x $100,000)
0
2
4
0
.5
1

Endogeneity and Endogenous Covariates

- The Problem
  - Unobserved Factors
  - Endogeneity
  - Omitted Variable Bias
  - Unobserved Confounding
- The Solution
  - Endogenous Covariates

Observed and Unobserved Factors

$$y = all\ factors\ that\ influence\ y$$

$$y = observed\ factors + unobserved\ factors$$

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

Endogeneity

"An explanatory variable in a multiple regression model that is correlated with the error term…" (Wooldridge*, pg 838).

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

$$\rho_{z\varepsilon} \neq 0$$

*Jeffrey M. Wooldridge (2009) Introductory Econometrics: A Modern Approach, 4th ed.

Omitted Variable Bias

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

$$\rho_{xz} \neq 0$$

$$y = \beta_0 + \beta_1 x + \varepsilon^*$$

$$\varepsilon^* = z + \varepsilon$$

$$y = \beta_0 + \beta_1 x + \varepsilon^*$$

$$\rho_{x\varepsilon^*} \neq 0$$

Confounding

"…X and Y are confounded when there is a third variable Z that influences both X and Y…" (Pearl*, pg 193).

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$$

*Judea Pearl (2009) Causality: Models, Reasoning, and Inference, 2nd ed.

Unobserved Confounding

$$y = \beta_0 + \beta_1 x + (z + \varepsilon)$$

Observed and Unobserved Factors

$$gpa = all\ factors\ that\ influence\ gpa$$

$$gpa = observed\ factors + unobserved\ factors$$

$$gpa = \left\{ \begin{matrix} \text{High school GPA} \\ \text{SAT Scores} \\ \text{Parents Income} \\ \text{Sex} \\ \text{etc…} \end{matrix} \right\} + \left\{ \begin{matrix} \text{Ability} \\ \text{Motivation} \\ \text{Sleep} \\ \text{Support} \\ \text{etc…} \end{matrix} \right\}$$

$$gpa = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

Unobserved Confounding

$$gpa = \beta_0 + \beta_1 income + \beta_2 hsgpa + \varepsilon_{total}$$

$$gpa = \beta_0 + \beta_1 income + \beta_2 hsgpa + (Ability + \varepsilon)$$

Endogenous Covariates

$$gpa = \beta_0 + \beta_1 income + \beta_2(hsgpa) + (Ability + \varepsilon_1)$$

$$hsgpa = \pi_0 + \pi_1 hs\_comp + (Ability + \varepsilon_2)^*$$

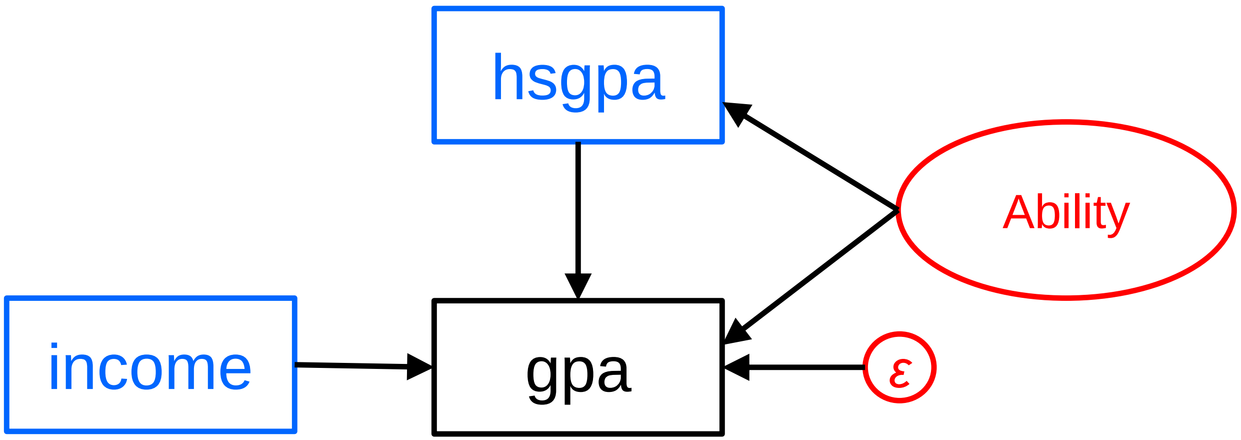
$$gpa = \beta_0 + \beta_1 income + \beta_2(\pi_0 + \pi_1 hs\_comp) + (Ability + \varepsilon_1)^*$$

$$where\ \rho_{\varepsilon_1^* \varepsilon_2^*} \neq 0$$

hsgpa = (factors NOT related to Ability) + (Ability + error)

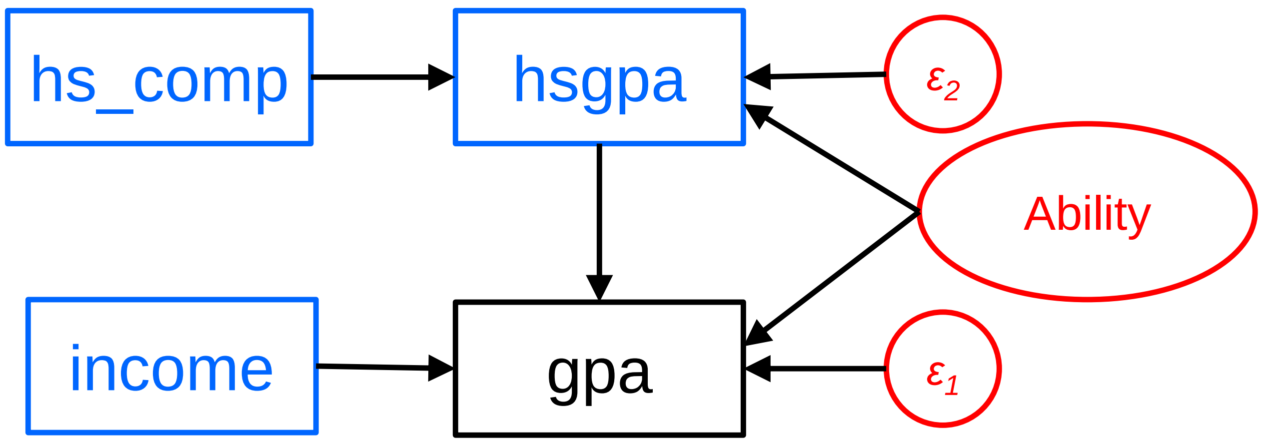
Endogenous Covariates

$$gpa = \beta_0 + \beta_1 income + \beta_2 hsgpa + (Ability + \varepsilon)$$



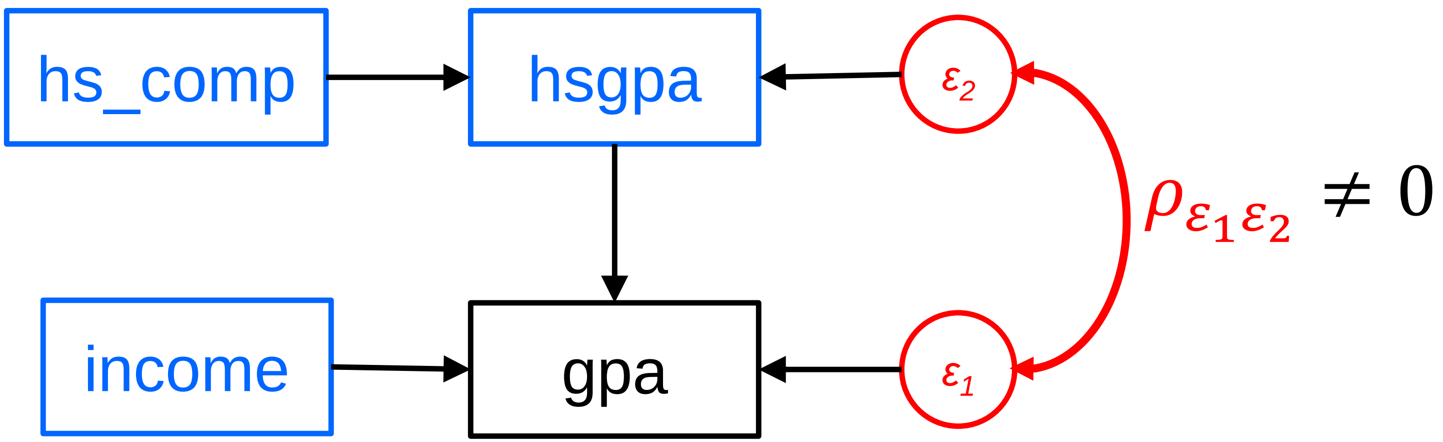
Endogenous Covariates

"Endogenous variables have arrows pointing to them and are variables within the system that are the effects of exogenous variables or causes of other endogenous variables within the system." (Mulaik, pg 120).

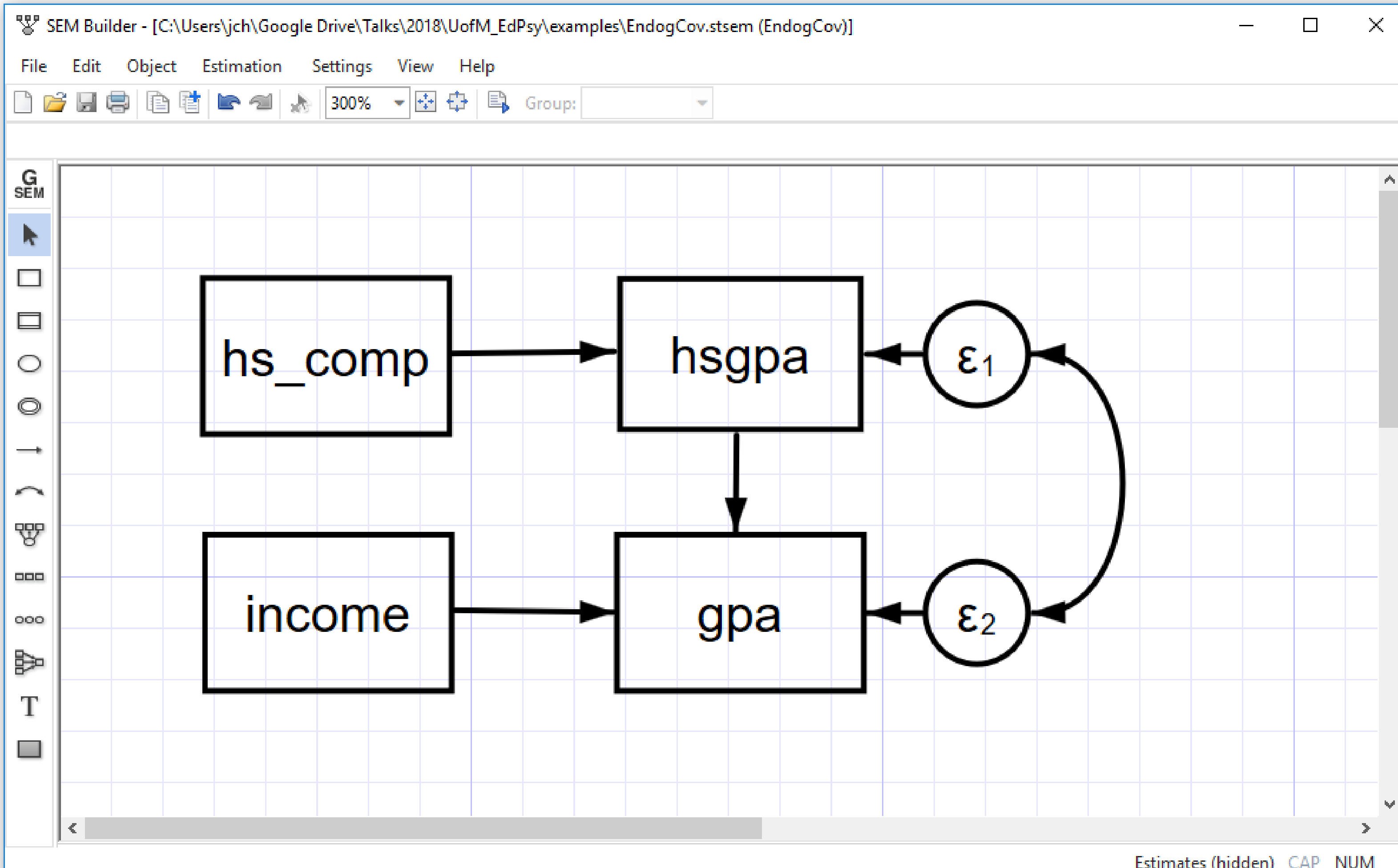


Stanley A. Mulaik (2009) Linear Causal Modeling With Structural Equations

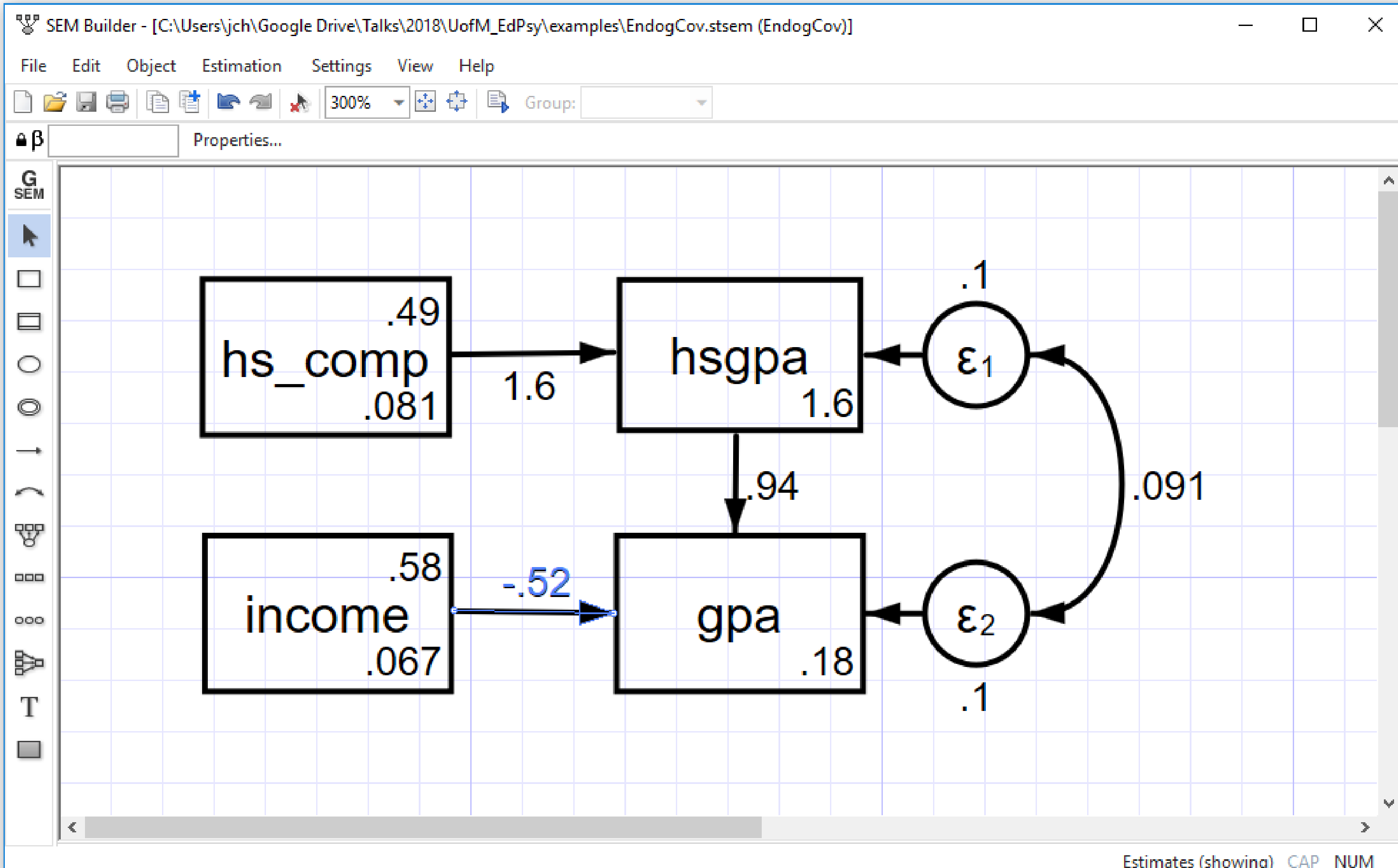
Endogenous Covariates



Endogenous Covariates



Endogenous Covariates



Endogenous Covariates

Primary model

```
eregress gpa income,                                        ///
            endogenous(hsgpa = hs_comp income)
```

Auxillary model

Endogenous Covariates

```
. eregress gpa income,                                    ///
>              endogenous(hsgpa = hs_comp income) nolog

Extended linear regression                      Number of obs     =        792
                                                Wald chi2(2)      =    3951.76
Log likelihood =  519.11827                     Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa						
income	.3702125	.0384969	9.62	0.000	.29476	.4456649
hsgpa	.9064316	.0193174	46.92	0.000	.8685702	.944293
_cons	-.2665693	.0397868	-6.70	0.000	-.3445499	-.1885887
hsgpa						
hs_comp	1.524814	.0244398	62.39	0.000	1.476913	1.572715
income	.9898417	.0268549	36.86	0.000	.9372069	1.042476
_cons	1.072201	.0203047	52.81	0.000	1.032404	1.111997
var(e.gpa)	.0554234	.0030877			.0496902	.061818
var(e.hsgpa)	.0381556	.0019174			.0345767	.0421049
corr(e.hsgpa,e.gpa)	.7503328	.0170348	44.05	0.000	.7149879	.7818521

```
. estimates store endog
```

Endogenous Covariates

var(e.gpa)	.0554234	.0030877			.0496902	.061818
var(e.hsgpa)	.0381556	.0019174			.0345767	.0421049
corr(e.hsgpa,e.gpa)	.7503328	.0170348	44.05	0.000	.7149879	.7818521

$$where\ \rho_{\varepsilon_1^* \varepsilon_2^*} \neq 0$$

Endogenous Covariates

```
. estimates table univar hsgpa endog, stats(N) equations(1) keep(#1:)  b(%9.4f)
```

Variable	univar	hsgpa	endog
program			
1	-0.1259	0.2003	
hsgpa		1.1445	0.9064
income			0.3702
_cons	2.1490	-0.6745	-0.2666
N	792	792	792

Outline

- ✓ Description of the dataset
- ✓ Endogenous Covariates
- Nonrandom treatment assignment
- Missing not at random (MNAR) and selection bias
- Treatment effects

Random Treatment Assignment

Study Program

No Study Program

Assign

Nonrandom Treatment Assignment

Study Program

No Study Program

Choice?

Nonrandom Treatment Assignment

A student's decision to enroll in the study program is based on observed and unobserved factors.

$$P(program = 1) = observed\ factors + unobserved\ factors$$

Unobserved Confounding

$$gpa = \beta_0 + \beta_1 income + \beta_2 program + \varepsilon_{total}$$

$$gpa = \beta_0 + \beta_1 income + \beta_2 program + (Ability + \varepsilon)$$

Endogenous Treatment

$$gpa = \beta_0 + \beta_1 income + \beta_2(program) + (Ability + \varepsilon_1)$$

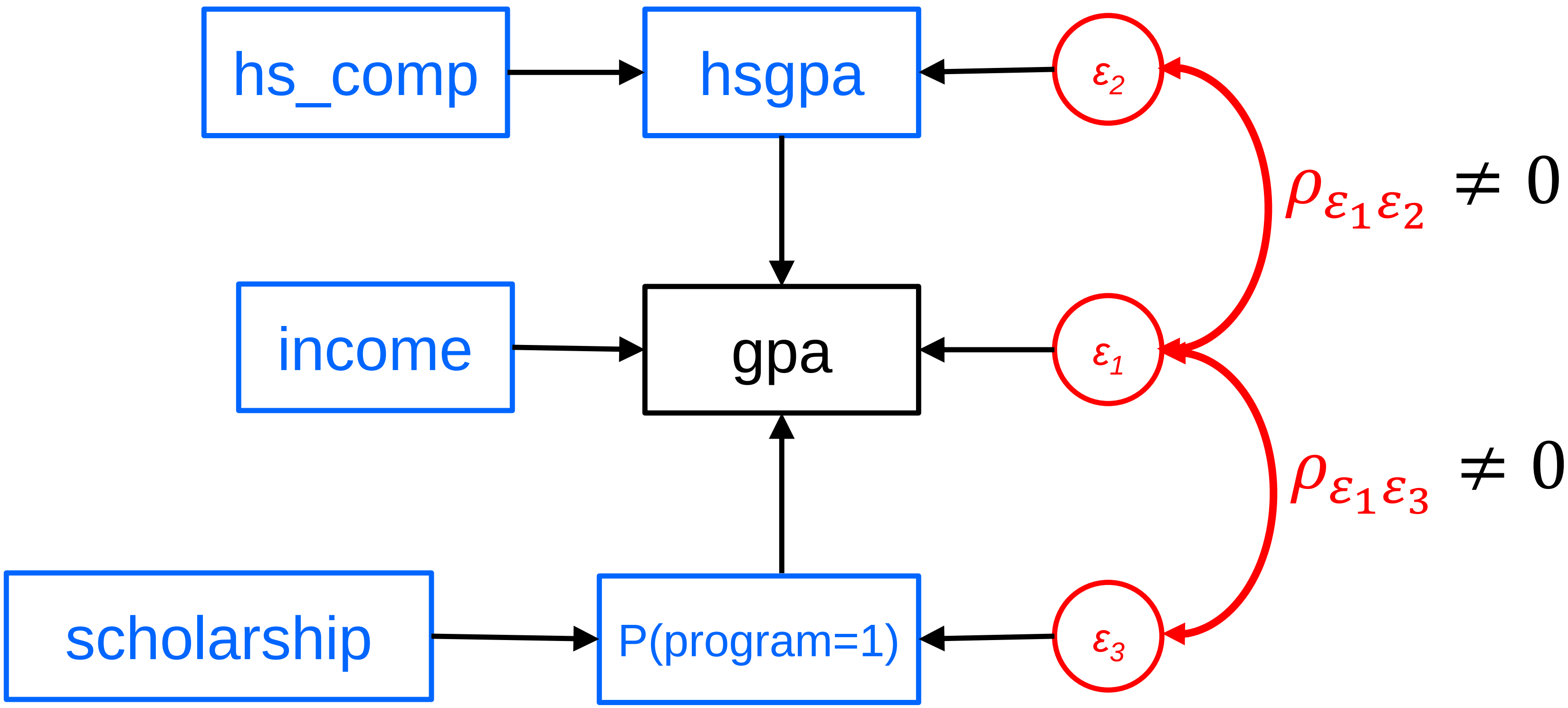
$$P(program = 1) = \pi_0 + \pi_1 scholarship + (Ability + \varepsilon_3)^*$$

$$gpa = \beta_0 + \beta_1 income + \beta_2(\pi_0 + \pi_1 scholarship) + (Ability + \varepsilon_1)^*$$

$$where\ \rho_{\varepsilon_1^* \varepsilon_3^*} \neq 0$$

P(program=1) = (factors NOT related to Ability) + (Ability + error)

Endogenous Treatment



Endogenous Treatment

Primary model

```
eregress gpa income,                                              ///
            endogenous(hsgpa = hs_comp income)                    ///
            entreat(program = income scholarship, nointeract)
```

Auxillary model

Endogenous Treatment

```
. eregress gpa income,                                            ///
>            endogenous(hsgpa = hs comp income)                   ///
>            entreat(program = income scholarship, nointeract)  nolog

Extended linear regression                   Number of obs     =        792
                                             Wald chi2(3)      =    6576.43
Log likelihood =   597.15048                 Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa						
income	.6876358	.0368212	18.67	0.000	.6154676	.7598041
hsgpa	.9021844	.0150525	59.94	0.000	.872682	.9316869
program						
Yes	.3040315	.019458	15.62	0.000	.2658944	.3421685
_cons	-.5198319	.0352035	-14.77	0.000	-.5888295	-.4508344
program						
income	-5.868551	.401813	-14.61	0.000	-6.65609	-5.081012
scholarship	1.814856	.1636187	11.09	0.000	1.49417	2.135543
_cons	1.503659	.1629857	9.23	0.000	1.184213	1.823106
hsgpa						
hs_comp	1.528458	.0236304	64.68	0.000	1.482144	1.574773
income	.989619	.0268526	36.85	0.000	.9369889	1.042249
_cons	1.070543	.0201056	53.25	0.000	1.031136	1.109949
var(e.gpa)	.0358984	.0020787			.0320469	.0402127
var(e.hsgpa)	.0381566	.0019175			.0345776	.0421062
corr(e.program,e.gpa)	.4511304	.0772058	5.84	0.000	.2877691	.5889813
corr(e.hsgpa,e.gpa)	.8093104	.0134908	59.99	0.000	.7811792	.8341618
corr(e.hsgpa,e.program)	.480631	.0565509	8.50	0.000	.3624217	.5836218

```
. estimates store entreat
```

Endogenous Treatment

var(e.gpa)	.0358984	.0020787			.0320469	.0402127
var(e.hsgpa)	.0381566	.0019175			.0345776	.0421062
corr(e.program,e.gpa)	.4511304	.0772058	5.84	0.000	.2877691	.5889813
corr(e.hsgpa,e.gpa)	.8093104	.0134908	59.99	0.000	.7811792	.8341618
corr(e.hsgpa,e.program)	.480631	.0565509	8.50	0.000	.3624217	.5836218

$$where\ \rho_{\varepsilon_1^* \varepsilon_3^*} \neq 0$$

Endogenous Treatment

```
. estimates table univar hsgpa endog entreat, stats(N) equations(1) keep(#1:) b(%9.4f)
```

Variable	univar	hsgpa	endog	entreat
program				
Yes	-0.1259	0.2003		0.3040
hsgpa		1.1445	0.9064	0.9022
income			0.3702	0.6876
_cons	2.1490	-0.6745	-0.2666	-0.5198
N	792	792	792	792

Outline

- ✓ Description of the dataset
- ✓ Endogenous Covariates
- ✓ Nonrandom treatment assignment
- Missing not at random (MNAR) and selection bias
- Treatment effects

No Missingness

Freshman

Graduate!

Missing Completely at Random (MCAR)

Freshman

Graduate!

Missing at Random (MAR)

Freshman

Graduate!

Missing Not at Random (MNAR)

Freshman

Graduate!

MNAR and Selection Bias

```
. tab graduate
```

Did the student graduate college?	Freq.	Percent	Cum.
No	208	20.80	20.80
Yes	792	79.20	100.00
Total	1,000	100.00	

Endogenous Selection

A student's decision to drop out of school is based on observed and unobserved factors.

$$P(graduate = 1) = {\color{blue} observed\ factors} + {\color{red} unobserved\ factors}$$

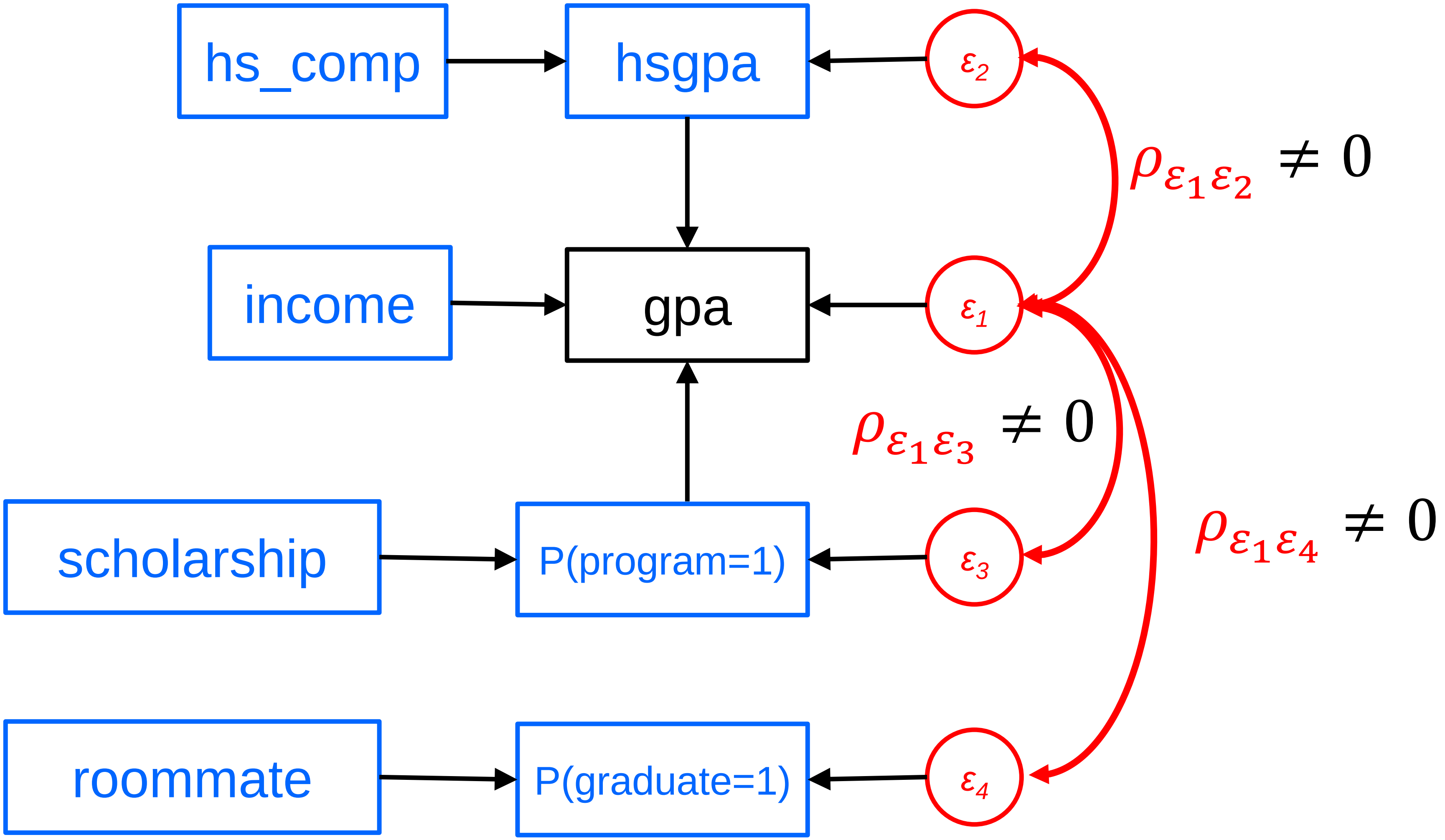
Endogenous Treatment

$$gpa = \begin{cases} \beta_0 + \beta_1 income + \beta_2(program) + (Ability + \varepsilon_1)* & \text{if graduate=1} \\ \text{missing} & \text{if graduate=0} \end{cases}$$

$$P(graduate = 1) = \pi_0 + \pi_1 roommate + (Ability + \varepsilon_4)^*$$

$$where\ \rho_{\varepsilon_1^* \varepsilon_4^*} \neq 0$$

Endogenous Covariates



Endogenous Treatment

Primary model

```
eregress gpa income,                                              ///
            endogenous(hsgpa = hs_comp income)                    ///
            entreat(program = income scholarship, nointeract)     ///
            select(graduate = income roommate)
```

Auxillary model

Endogenous Treatment

```
. eregress gpa income,                                              ///
>          endogenous(hsgpa = hs_comp income)                       ///
>          entreat(program = income scholarship, nointeract)        ///
>          select(graduate = income roommate)  nolog

Extended linear regression                      Number of obs     =      1,000
                                                      Selected    =        792
                                                      Nonselected =        208

                                                Wald chi2(3)      =    8866.38
Log likelihood =  323.23691                     Prob > chi2       =     0.0000

---------------------------------------------------------------------------------------
                         |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------+-------------------------------------------------------------
gpa                      |
                  income |   .8220509   .0333135    24.68   0.000     .7567576    .8873443
                   hsgpa |   .8935782   .0136619    65.41   0.000     .8668013    .9203551
                         |
                 program |
                     Yes |   .2976643   .0168041    17.71   0.000     .2647288    .3305998
                   _cons |  -.6071633    .029947   -20.27   0.000    -.6658583   -.5484682
-------------------------+-------------------------------------------------------------
graduate                 |
                  income |   4.010154   .2557017    15.68   0.000     3.508988     4.51132
                roommate |   1.412072   .1320596    10.69   0.000      1.15324    1.670904
                   _cons |  -1.053694   .1059937    -9.94   0.000    -1.261438   -.8459504
-------------------------+-------------------------------------------------------------
program                  |
                  income |  -4.889741   .2935974   -16.65   0.000    -5.465181     -4.3143
             scholarship |   1.791084   .1291875    13.86   0.000     1.537881    2.044287
                   _cons |   .8297874   .1047466     7.92   0.000     .6244878    1.035087
-------------------------+-------------------------------------------------------------
hsgpa                    |
                 hs_comp |   1.512085   .0202588    74.64   0.000     1.472378    1.551791
                  income |     1.0879   .0221946    49.02   0.000     1.044399      1.1314
                   _cons |   .9990863   .0161398    61.90   0.000     .9674529     1.03072
-------------------------+-------------------------------------------------------------
              var(e.gpa) |    .040487   .0023354                        .0361589    .0453332
            var(e.hsgpa) |   .0399236   .0017858                        .0365726    .0435817
-------------------------+-------------------------------------------------------------
     corr(e.graduate,e.gpa)|   .7609452   .0402982    18.88   0.000     .6700487    .8293596
      corr(e.program,e.gpa)|   .5402021   .0577087     9.36   0.000     .4175545    .6435181
        corr(e.hsgpa,e.gpa)|   .8221551   .0119073    69.05   0.000      .797394    .8441524
 corr(e.program,e.graduate)|     .85115   .0432119    19.70   0.000     .7411121    .9166561
   corr(e.hsgpa,e.graduate)|   .5633432   .0408602    13.79   0.000     .4780104    .6381415
    corr(e.hsgpa,e.program)|   .5265467   .0436265    12.07   0.000      .435811    .6066872
---------------------------------------------------------------------------------------

. estimates store endsel
```

Endogenous Treatment

var(e.gpa)	.040487	.0023354			.0361589	.0453332
var(e.hsgpa)	.0399236	.0017858			.0365726	.0435817
corr(e.graduate,e.gpa)	.7609452	.0402982	18.88	0.000	.6700487	.8293596
corr(e.program,e.gpa)	.5402021	.0577087	9.36	0.000	.4175545	.6435181
corr(e.hsgpa,e.gpa)	.8221551	.0119073	69.05	0.000	.797394	.8441524
corr(e.program,e.graduate)	.85115	.0432119	19.70	0.000	.7411121	.9166561
corr(e.hsgpa,e.graduate)	.5633432	.0408602	13.79	0.000	.4780104	.6381415
corr(e.hsgpa,e.program)	.5265467	.0436265	12.07	0.000	.435811	.6066872

Endogenous Covariates

```
. estimates table univar hsgpa endog entreat endsel, stats(N) equations(1) keep(#1:) b(%9.4f)
```

Variable	univar	hsgpa	endog	entreat	endsel
program					
Yes	-0.1259	0.2003		0.3040	0.2977
hsgpa		1.1445	0.9064	0.9022	0.8936
income			0.3702	0.6876	0.8221
_cons	2.1490	-0.6745	-0.2666	-0.5198	-0.6072
N	792	792	792	792	1000

<u>True Model (simulated)</u>

gpa = -0.6 + 0.3*treatment + 0.9*hsgpa + 0.8*income

ERM Postestimation

- **`estat teffects`**
- **`margins`**
- **`marginsplot`**
- **`predict`**

Treatment Effects

```
. estat teffects

Predictive margins                               Number of obs     =      1,000
Model VCE      : OIM
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
program (Yes vs No)	.2976643	.0168041	17.71	0.000	.2647288	.3305998

Note: Standard errors treat sample covariate values as fixed and not a draw from the population. If your interest is in population rather than sample effects, refit your model using **vce(robust)**.

Treatment Effects

```
. estat teffects, atet

Predictive margins                              Number of obs     =      1,000
                                                Subpop. no. obs   =        300
Model VCE    : OIM

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ATET         |
     program |
  (Yes vs No)|   .2976643    .0168041    17.71   0.000     .2647288    .3305998
------------------------------------------------------------------------------
```

ERM Postestimation

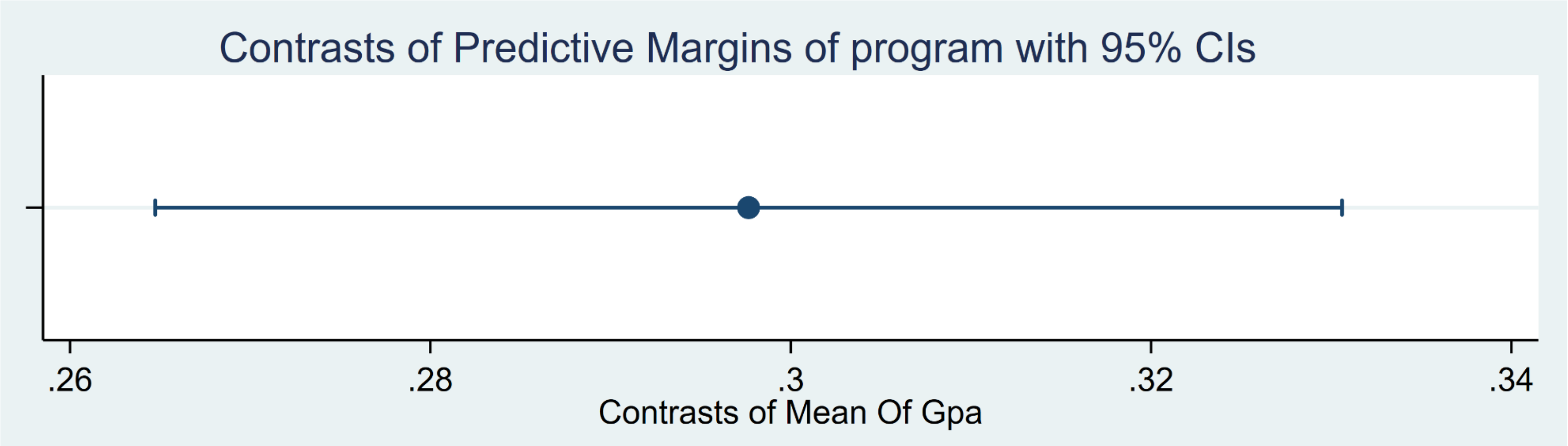
```
. generate programT = program

. margins r(0 1).program if program,        ///
>         predict(base(program=programT))   ///
>         contrast(effects nowald)

Contrasts of predictive margins
Model VCE    : OIM

Expression   : mean of gpa, predict(base(program=programT))
```

	Contrast	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
program						
(Yes vs No)	.2976643	.0168041	17.71	0.000	.2647288	.3305998

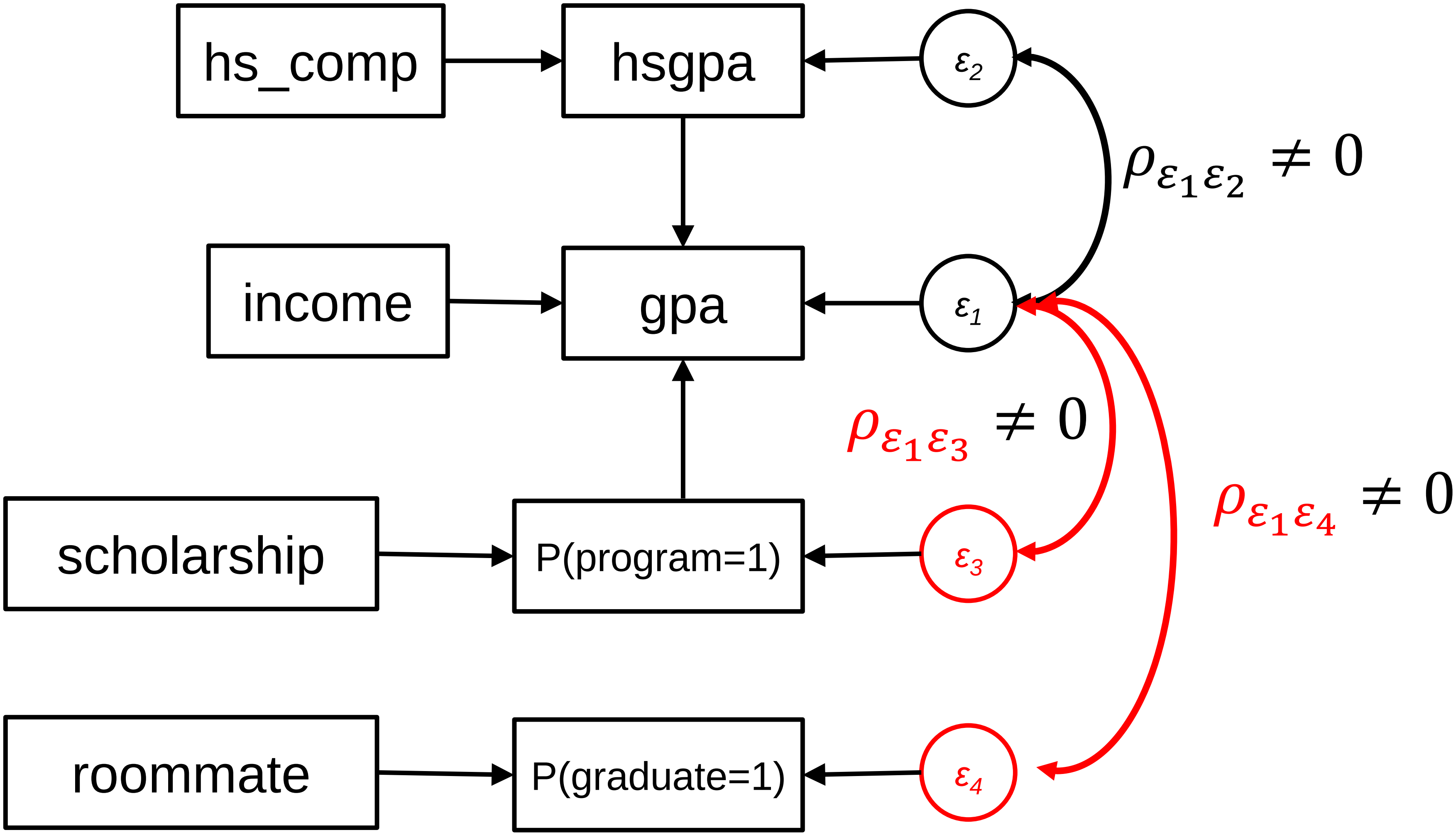
```
. marginsplot, horizontal aspectratio(0.2)
```

ERM Postestimation

```
. marginsplot, horizontal aspectratio(0.2)
```



Why don't we just use **gsem**?



More ERMs

- **`eregress`** – continuous outcomes
- **`eintreg`** – interval outcomes
- **`eprobit`** – binary outcomes
- **`eoprobit`** – ordinal outcomes

More ERMs

- ERMs can include:
  - polynomials of endogenous covariates
  - interactions of endogenous covariates
  - interactions of endogenous with exogenous covariates

Cautionary Note

- Nothing about ERMs magically extracts causal relationships.

- As with any regression analysis of observational data, the causal interpretation must be based on a reasonable underlying scientific rationale.

Thanks for coming!

Questions?

chuber@stata.com