# Extended regression models using Stata 15

Charles Lindsey

Senior Statistician and Software Developer
Stata

July 19, 2018

# Introduction

- Common problems in observational data

- endogenous sample selection
- endogenous covariates
- nonrandom treatment assignment

# Introduction

- Common problems in observational data

- endogenous sample selection
  - trials with informative droput
  - missing not at random (MNAR)
  - selection on unobservables
  - Heckman selection

# Introduction

- Common problems in observational data

- endogenous covariates
  - unobserved confounding variables
  - simultaneous causality, in linear models
  - any covariates correlated with the errors

# Introduction

- Common problems in observational data

- nonrandom treatment assignment
  - unobserved factors affecting outcome and treatment are related

# Introduction

- Common problems in observational data

- endogenous sample selection
- endogenous covariates
- nonrandom treatment assignment

# Introduction

- Common problems in observational data

  Solution: Extended Regression Model (ERM) commands

- endogenous sample selection

  `select()`
- endogenous covariates

  `endogenous()`
- nonrandom treatment assignment

  `entreat()`

# Introduction

- Some of you are shaking your heads up and down.
- You have encountered these complications often.
- Others may be less familiar with them.

# Introduction

- Some of you are shaking your heads up and down.
- You have encountered these complications often.
- Others may be less familiar with them.
- What if you wish to estimate the relationship between college GPA and high school GPA but have no measure of unobservable ability?

# Introduction

- Some of you are shaking your heads up and down.
- You have encountered these complications often.
- Others may be less familiar with them.
- What if you wish to estimate the relationship between college GPA and high school GPA but have no measure of unobservable ability?
- Ability affects both GPAs and those effects must be accounted for in order to estimate the relationship between the GPAs.

# Introduction

- Some of you are shaking your heads up and down.
- You have encountered these complications often.
- Others may be less familiar with them.
- What if you wish to estimate the relationship between college GPA and high school GPA but have no measure of unobservable ability?
- Ability affects both GPAs and those effects must be accounted for in order to estimate the relationship between the GPAs.
- ERMs can handle this problem if you also have a model for high school GPA.

# Introduction

- What if you further want to measure the relationship between the GPAs for everyone?

# Introduction

- What if you further want to measure the relationship between the GPAs for everyone?
- This includes those who do not even attend college.

# Introduction

- What if you further want to measure the relationship between the GPAs for everyone?
- This includes those who do not even attend college.
- ERMs can handle this problem if you also have a model for college attendance.

# Introduction

- What if you further want to measure the relationship between the GPAs for everyone?
- This includes those who do not even attend college.
- ERMs can handle this problem if you also have a model for college attendance.
- What if you want to see the effect of a voluntary program on college GPA?

# Introduction

- What if you further want to measure the relationship between the GPAs for everyone?
- This includes those who do not even attend college.
- ERMs can handle this problem if you also have a model for college attendance.
- What if you want to see the effect of a voluntary program on college GPA?
- Program participation is not randomly assigned.

# Introduction

- What if you further want to measure the relationship between the GPAs for everyone?
- This includes those who do not even attend college.
- ERMs can handle this problem if you also have a model for college attendance.
- What if you want to see the effect of a voluntary program on college GPA?
- Program participation is not randomly assigned.
- ERMs can handle this problem if you have a model for program assignment.

# Introduction

- Extended regression model (ERM) is a term that we developed to describe models that accomodate endogenous sample selection, nonrandom treatment assignment, and endogenous covariates.
- The term and the mascot monster are clearly made up, but the models themselves are not our invention.
- Stata has many commands for estimating models with these complications using maximum likelihood and other estimation methods.
- What makes ERMs different is that you can combine the complications in a single model.

# Introduction

- Extended regression model (ERM) is a term that we developed to describe models that accomodate endogenous sample selection, nonrandom treatment assignment, and endogenous covariates.
- The term and the mascot monster are clearly made up, but the models themselves are not our invention.
- Stata has many commands for estimating models with these complications using maximum likelihood and other estimation methods.
- What makes ERMs different is that you can combine the complications in a single model.
- You can have an endogenous covariate **and** endogenous sample selection.

# Introduction

- Extended regression model (ERM) is a term that we developed to describe models that accomodate endogenous sample selection, nonrandom treatment assignment, and endogenous covariates.
- The term and the mascot monster are clearly made up, but the models themselves are not our invention.
- Stata has many commands for estimating models with these complications using maximum likelihood and other estimation methods.
- What makes ERMs different is that you can combine the complications in a single model.
- You can have an endogenous covariate **and** endogenous sample selection.
- You can have an endogenous covariate **and** endogenous treatment assignment.

# Introduction

- Extended regression model (ERM) is a term that we developed to describe models that accomodate endogenous sample selection, nonrandom treatment assignment, and endogenous covariates.
- The term and the mascot monster are clearly made up, but the models themselves are not our invention.
- Stata has many commands for estimating models with these complications using maximum likelihood and other estimation methods.
- What makes ERMs different is that you can combine the complications in a single model.
- You can have an endogenous covariate **and** endogenous sample selection.
- You can have an endogenous covariate **and** endogenous treatment assignment.
- You might even have more than two complications.

# Introduction

- Nothing comes for free though.
- To handle any of these complications, ERMs require an additional model for the complication itself.
- The ERM commands estimate the parameters of these additional models and the model of the outcome using maximum-likelihood.

# ERM commands

- So ERM commands have options to deal with these common observational data issues.
- There are four ERM commands. All of which support these options.

# ERM commands

- So ERM commands have options to deal with these common observational data issues.
- There are four ERM commands. All of which support these options.
  - **eregress** for continuous outcomes

# ERM commands

- So ERM commands have options to deal with these common observational data issues.
- There are four ERM commands. All of which support these options.
  - **eregress** for continuous outcomes
  - **eintreg** for
    - interval-censored outcomes
    - right-censored outcomes
    - left-censored outcomes
    - tobit-type outcomes

# ERM commands

- So ERM commands have options to deal with these common observational data issues.
- There are four ERM commands. All of which support these options.
  - **eregress** for continuous outcomes
  - **eintreg** for
    - interval-censored outcomes
    - right-censored outcomes
    - left-censored outcomes
    - tobit-type outcomes
  - **eprobit** for binary outcomes

# ERM commands

- So ERM commands have options to deal with these common observational data issues.
- There are four ERM commands. All of which support these options.
  - **eregress** for continuous outcomes
  - **eintreg** for
    - interval-censored outcomes
    - right-censored outcomes
    - left-censored outcomes
    - tobit-type outcomes
  - **eprobit** for binary outcomes
  - **eoprobit** for ordinal outcomes

# ERM commands

- So ERM commands have options to deal with these common observational data issues.
- There are four ERM commands. All of which support these options.
  - **eregress** for continuous outcomes
  - **eintreg** for
    - interval-censored outcomes
    - right-censored outcomes
    - left-censored outcomes
    - tobit-type outcomes
  - **eprobit** for binary outcomes
  - **eoprobit** for ordinal outcomes
- Today we will explore how to use the ERM commands to make inference using data with these issues.

# Example

- Fictional State University is studying the relationship between high school grade point average (GPA) of admitted students and their final college GPA.
- Parental income is included as a covariate.

$$\text{gpa} = \beta_1 \text{hsgpa} + \beta_2 \text{income} + \beta_0 + \epsilon$$

- If we did not have not any complications, we could use linear regression through the `regress` command to estimate the parameters of this model.

# Example

- The syntax for `regress` is

```
. regress gpa income hsgpa
```

- For `eregress`, we have:

```
. eregress gpa income hsgpa

Iteration 0:   log likelihood = -1079.4282
Iteration 1:   log likelihood = -1079.4267
Iteration 2:   log likelihood = -1079.4267

Extended linear regression                      Number of obs     =      1,585
                                                Wald chi2(2)      =    1967.58
Log likelihood = -1079.4267                     Prob > chi2       =     0.0000
```

| gpa | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| income | .0227565 | .0043742 | 5.20 | 0.000 | .0141833 | .0313297 |
| hsgpa | 1.707055 | .0482858 | 35.35 | 0.000 | 1.612417 | 1.801694 |
| _cons | -2.270331 | .1346492 | -16.86 | 0.000 | -2.534238 | -2.006423 |
| var(e.gpa) | .2285902 | .00812 | | | .2132166 | .2450723 |

# Example

- We saw estimated coefficients and a variance estimate for the unobserved error $\epsilon$.
- Here `eregress` and `regress` will have the same coefficient estimates.
- However, the standard errors differ by a factor of $\sqrt{(N/(N-k))}$, where $N$ is the sample size and $k$ is the number of coefficients.
- We will not interpret the estimated coefficients in this model.
- The data suffers from some of those complications that we mentioned earlier.

# Endogenous sample selection

- Not all admitted students stayed in school.
- But even for those that dropped out, the administration wants to predict what their GPA would have been if they had remained in school.
- The unobserved factors that affect whether a student stays in school may be related to the unobserved factors that affect their GPA.
  - Family, social support system, etc.
- Using a standard linear regression here will provide inconsistent estimates.

# Endogenous sample selection

- In ERMs, we model this relationship by correlating the unobserved error of the outcome ($\epsilon$ here) with the unobserved error that affects selection into the sample.
- Whether the student has a roommate from the school is used as a selection covariate.

$$\text{inschool} = (\alpha_1 \text{income} + \alpha_2 \text{roommate} + \alpha_0 + \epsilon_{sel} > 0)$$

- When the correlation between $\epsilon$ and $\epsilon_{sel}$ is non-zero, we have **endogenous sample selection**.

# Example

- The existing `heckman` command could be used to estimate the parameters if endogenous sample selection was the only problem.

```
. heckman gpa income hsgpa, select(inschool=i.roommate income)
```

- For `eregress`, we have:

```
. eregress gpa income hsgpa, select(inschool=i.roommate income)
```

# Example

Extended linear regression — Number of obs = 2,000; Selected = 1,585; Nonselected = 415

Wald chi2(2) = 1602.57

Log likelihood = -1897.6514 — Prob > chi2 = 0.0000

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| gpa | | | | | | |
| income | .0341667 | .0066101 | 5.17 | 0.000 | .0212111 | .0471223 |
| hsgpa | 1.702159 | .0482049 | 35.31 | 0.000 | 1.607679 | 1.796639 |
| _cons | -2.379314 | .1433418 | -16.60 | 0.000 | -2.660259 | -2.098369 |
| inschool | | | | | | |
| 1.roommate | .7749166 | .0768935 | 10.08 | 0.000 | .6242081 | .9256251 |
| income | .2392745 | .0159158 | 15.03 | 0.000 | .2080801 | .2704689 |
| _cons | -.7127948 | .0912127 | -7.81 | 0.000 | -.8915684 | -.5340212 |
| var(e.gpa) | .2392988 | .0127984 | | | .2154843 | .2657452 |
| corr(e.ins~l, e.gpa) | .3886257 | .1592341 | 2.44 | 0.015 | .0425408 | .6514386 |

# Example

- So if you know how to use Stata's existing `heckman` command, you know how to use ERMs to model sample selection.
- In our online documentation, see [ERM] intro 7 for other examples comparing the ERM commands with existing Stata commands like `heckman`.
- The entire ERM manual is free on our website.
- Also see [ERM] intro 4 for an introduction to endogenous sample selection in the ERM framework.

# Example

- We have plenty of examples of endogenous sample selection in the documentation as well:
  - [ERM] example 1c Interval regression with endogenous covariate and sample selection
  - [ERM] example 4a Probit regression with endogenous sample selection
  - [ERM] example 4b Probit regression with endogenous treatment and sample selection
  - [ERM] example 6b Ordered probit regression with endogenous treatment and sample selection

# Example

- In the output we saw coefficient estimates for the outcome model and selection model.
- We also estimated the variance of the the unobserved outcome error $\epsilon$, and the correlation of this outcome error with the selection errors $\epsilon_{sel}$.
- We will wait to interpret the parameter estimates because our data also suffers from...

# Endogenous covariates

- The unobserved factors that affect high school GPA may also be related to the unobserved factors that affect college GPA.
  - Ability, family, social support system, etc.
- In this situation, standard linear regression is again faulty. `regress` will give us inconsistent estimates. So will `heckman`.
- In the extended linear regression model, we model this relationship by correlating the unobserved error that affects college GPA ($\epsilon$) with the unobserved error that affects high school GPA.

# Endogenous covariates

- We use high school competitiveness as a covariate for high school GPA.

$$\begin{aligned} \text{hsgpa} = \beta_{21}\text{income} &+ \beta_{22}(\text{hscomp=medium}) \\ &+ \beta_{23}(\text{hscomp=high}) + \beta_{20} + \epsilon_2 \end{aligned}$$

- When the correlation between $\epsilon$ and $\epsilon_2$ is non-zero, high school GPA is an **endogenous covariate**.

# Example

- The existing `ivregress` command could be used to estimate the parameters if an endogenous covariate was the only problem.

```
. ivregress liml gpa income (hsgpa=i.hscomp)
```

- For `eregress`, we have:

```
. eregress gpa income, endogenous(hsgpa=i.hscomp income)
```

# Example

Extended linear regression — Number of obs = 1,585
Wald chi2(2) = 630.97
Log likelihood = -1045.398 — Prob > chi2 = 0.0000

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| gpa | | | | | | |
| income | .0601803 | .0094922 | 6.34 | 0.000 | .0415759 | .0787847 |
| hsgpa | .8911469 | .1866711 | 4.77 | 0.000 | .5252784 | 1.257015 |
| _cons | -.0367553 | .5117771 | -0.07 | 0.943 | -1.03982 | .9663093 |
| hsgpa | | | | | | |
| hscomp | | | | | | |
| moderate | -.1433858 | .0134962 | -10.62 | 0.000 | -.1698379 | -.1169337 |
| high | -.2101839 | .0222694 | -9.44 | 0.000 | -.2538312 | -.1665367 |
| income | .0456505 | .0018832 | 24.24 | 0.000 | .0419595 | .0493414 |
| _cons | 2.849839 | .0161962 | 175.96 | 0.000 | 2.818095 | 2.881583 |
| var(e.gpa) | .2697688 | .0211392 | | | .2313615 | .3145519 |
| var(e.hsgpa) | .0569694 | .0020237 | | | .053138 | .0610772 |
| corr(e.hsgpa, e.gpa) | .4071113 | .0745743 | 5.46 | 0.000 | .2514341 | .542255 |

# Example

- So if you know how to use Stata's existing `ivregress` command, you know how to use ERMs to model endogenous covariates.
- In our online documentation, see [ERM] intro 3 for an introduction to endogenous covariates in the ERM framework.

# Example

- We have plenty of examples of endogenous covariates in the documentation as well:
  - [ERM] example 1a Linear regression with continuous endogenous covariate
  - [ERM] example 1b Interval regression with continuous endogenous covariate
  - [ERM] example 1c Interval regression with endogenous covariate and sample selection
  - [ERM] example 2a Linear regression with binary endogenous covariate
  - [ERM] example 3a Probit regression with continuous endogenous covariate
  - [ERM] example 3b Probit regression with endogenous covariate and treatment

# Example

- In the output, we saw coefficient estimates for the outcome model and endogenous covariate model.
- We also estimated the variance of the the unobserved outcome error $\epsilon$, the variance of the endogenous error $\epsilon_2$, and the correlation between them.
- We will not interpret the parameter estimates, because this model ignores the endogenous sample selection.

# Example

- Our data suffers from both endogenous sample selection and an endogenous covariate.
- We will use `eregress` to estimate the parameters of the model.
- The estimation output takes more than one page since we have two data complications.

# Header and main equation

```
. eregress gpa income,
           endogenous(hsgpa=i.hscomp income)
           select(inschool=i.roommate income)

Iteration 0:   log likelihood = -1820.8777
Iteration 1:   log likelihood = -1820.4304
Iteration 2:   log likelihood = -1820.4271
Iteration 3:   log likelihood = -1820.4271

Extended linear regression                      Number of obs     =      2,000
                                                      Selected    =      1,585
                                                      Nonselected =        415

                                                Wald chi2(2)      =     367.52
Log likelihood = -1820.4271                     Prob > chi2       =     0.0000
```

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| gpa | | | | | | |
| income | .0708905 | .0112158 | 6.32 | 0.000 | .0489079 | .0928731 |
| hsgpa | .8777339 | .185311 | 4.74 | 0.000 | .514531 | 1.240937 |
| _cons | -.1141296 | .5005744 | -0.23 | 0.820 | -1.095238 | .8669783 |

# Auxiliary equations and parameters

| | | | | | | |
|---|---|---|---|---|---|---|
| inschool | | | | | | |
| 1.roommate | .7628986 | .0697877 | 10.93 | 0.000 | .6261172 | .89968 |
| income | .2411492 | .0158024 | 15.26 | 0.000 | .2101771 | .2721213 |
| _cons | -.7124675 | .0873117 | -8.16 | 0.000 | -.8835953 | -.5413397 |
| hsgpa | | | | | | |
| hscomp | | | | | | |
| moderate | -.1390269 | .0116398 | -11.94 | 0.000 | -.1618404 | -.1162134 |
| high | -.2127761 | .0196419 | -10.83 | 0.000 | -.2512735 | -.1742787 |
| income | .0501507 | .0017217 | 29.13 | 0.000 | .0467762 | .0535252 |
| _cons | 2.793765 | .0136546 | 204.60 | 0.000 | 2.767002 | 2.820527 |
| var(e.gpa) | .2801667 | .0244111 | | | .2361842 | .3323397 |
| var(e.hsgpa) | .0581159 | .001838 | | | .0546228 | .0618324 |
| corr(e.ins~l, e.gpa) | .3466803 | .1429833 | 2.42 | 0.015 | .0431142 | .5916431 |
| corr(e.hsgpa, e.gpa) | .431405 | .0723976 | 5.96 | 0.000 | .2796273 | .5621463 |
| corr(e.hsgpa, e.inschool) | .3752079 | .0317998 | 11.80 | 0.000 | .3112529 | .4357796 |

# Correlations

| | | | | | | |
|---|---|---|---|---|---|---|
| corr(e.ins~l, e.gpa) | .3466803 | .1429833 | 2.42 | 0.015 | .0431142 | .5916431 |
| corr(e.hsgpa, e.gpa) | .431405 | .0723976 | 5.96 | 0.000 | .2796273 | .5621463 |
| corr(e.hsgpa, e.inschool) | .3752079 | .0317998 | 11.80 | 0.000 | .3112529 | .4357796 |

- These estimates tell about us about the relationship between the unobserved factors that affect college GPA, high school GPA, and whether the student stays in school.
- Clearly we have endogeneity, there is non-zero correlation between these unobserved factors.
- We can interpret the direction of relationship as well.
- For example, the unobserved factors that increase high school GPA tend to increase college GPA as well.

# Main equation

| | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| gpa | | | | | | |
| income | .0708905 | .0112158 | 6.32 | 0.000 | .0489079 | .0928731 |
| hsgpa | .8777339 | .185311 | 4.74 | 0.000 | .514531 | 1.240937 |
| _cons | -.1141296 | .5005744 | -0.23 | 0.820 | -1.095238 | .8669783 |

- In the extended linear regression model, we can directly interpret the model coefficients.
- For example, the difference in college GPA is about .88 points for students with a 1 point difference in high school GPA.

# Nonrandom treatment assignment

- Now we will extend this model even further to handle all three complications.
- The administration has implemented a new study skills training program.
- Students must elect to take part.
- So the assignment of the treatment (participation in the program) is not random.

# Potential outcomes

$$
\begin{aligned}
\text{gpa}_0 &= \beta_{01}\text{hsgpa} + \beta_{02}\text{income} + \beta_{00} + \epsilon_0 \\
\text{gpa}_1 &= \beta_{11}\text{hsgpa} + \beta_{12}\text{income} + \beta_{10} + \epsilon_1
\end{aligned}
$$

- This is a classic treatment effects framework.
- We observe $\text{gpa}_0$ for those who do not participate in the study program.
- We observe $\text{gpa}_1$ for those who do participate in the study program.

# Potential outcomes

$$
\begin{aligned}
\text{gpa}_0 &= \beta_{01}\text{hsgpa} + \beta_{02}\text{income} + \beta_{00} + \epsilon_0 \\
\text{gpa}_1 &= \beta_{11}\text{hsgpa} + \beta_{12}\text{income} + \beta_{10} + \epsilon_1
\end{aligned}
$$

- We wished that we observed $\text{gpa}_0$ for those who participated.
- However, we can use the model to predict the mean of $\text{gpa}_0$ for those who participated.
- Similarly, we can use the model to predict the mean of $\text{gpa}_1$ for those who did not participate.

# Potential outcomes

$$
\begin{aligned}
\text{gpa}_0 &= \beta_{01}\text{hsgpa} + \beta_{02}\text{income} + \beta_{00} + \epsilon_0 \\
\text{gpa}_1 &= \beta_{11}\text{hsgpa} + \beta_{12}\text{income} + \beta_{10} + \epsilon_1
\end{aligned}
$$

- We can estimate $E(\text{gpa}_1 - \text{gpa}_0)$ to determine the treatment effect of the program on college GPA.
- I am having to cover this concept pretty fast. There is much more information on the potential outcome framework in the Stata documentation on our website: [TE] teffects intro, [ERM] intro 5
- Remember that you will get a copy of these slides, and be able to access the links.

# Treatment assignment

- The unobserved factors that affect whether a student takes part in the study program may be related to the unobserved factors that affect their GPA.
- Ability, family, social support system, extracurricular activities.
- In ERMs, we again model this relationship by correlating the unobserved outcome errors ($\epsilon_0$ and $\epsilon_1$) with the unobserved error that affects treatment assignment.

# Treatment assignment

- Whether the student has a scholarship is used as a treatment covariate.

$$\text{program} = (\gamma_1 \text{income} + \gamma_2 \text{scholar} + \gamma_0 + \epsilon_{tr} > 0)$$

- When the correlation between $\epsilon_{tr}$ and $\epsilon_0$, $\epsilon_1$ is non-zero, we have **endogenous treatment assignment**.
- If the correlation is zero, we have **exogenous treatment assignment**.
- The ERM commands can handle both these cases.

# Command

```
eregress gpa income,
         entreat(program=scholar income)
         endogenous(hsgpa=i.hscomp income)
         select(inschool=i.roommate income)
         vce(robust)
```

# Header and main equation

```
Extended linear regression                      Number of obs     =      2,000
                                                      Selected    =      1,585
                                                      Nonselected =        415

                                                Wald chi2(6)      =   57650.13
Log pseudolikelihood =  -2396.361               Prob > chi2       =     0.0000
```

| | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| gpa | | | | | | |
| program# c.income | | | | | | |
| 0 | .0559082 | .0081052 | 6.90 | 0.000 | .0400223 | .0717942 |
| 1 | .0921056 | .0080322 | 11.47 | 0.000 | .0763629 | .1078483 |
| program# c.hsgpa | | | | | | |
| 0 | 1.142148 | .1282104 | 8.91 | 0.000 | .8908606 | 1.393436 |
| 1 | .9391335 | .131239 | 7.16 | 0.000 | .6819098 | 1.196357 |
| program | | | | | | |
| 0 | -1.051847 | .3449417 | -3.05 | 0.002 | -1.72792 | -.3757735 |
| 1 | -.0869778 | .3550886 | -0.24 | 0.806 | -.7829387 | .6089832 |

# Auxiliary equations

| | | | | | | |
|---|---|---|---|---|---|---|
| inschool | | | | | | |
| 1.roommate | .7493605 | .0691626 | 10.83 | 0.000 | .6138043 | .8849168 |
| income | .2412716 | .0151986 | 15.87 | 0.000 | .211483 | .2710603 |
| _cons | -.7051772 | .0864542 | -8.16 | 0.000 | -.8746244 | -.5357301 |
| program | | | | | | |
| scholar | 1.004336 | .0610865 | 16.44 | 0.000 | .8846087 | 1.124064 |
| income | -.0480899 | .0097213 | -4.95 | 0.000 | -.0671433 | -.0290364 |
| _cons | -.2931821 | .0631522 | -4.64 | 0.000 | -.416958 | -.1694061 |
| hsgpa | | | | | | |
| hscomp | | | | | | |
| moderate | -.1403685 | .0116822 | -12.02 | 0.000 | -.1632652 | -.1174718 |
| high | -.2112942 | .018883 | -11.19 | 0.000 | -.2483041 | -.1742842 |
| income | .0501522 | .0017847 | 28.10 | 0.000 | .0466543 | .0536502 |
| _cons | 2.794466 | .0135717 | 205.90 | 0.000 | 2.767866 | 2.821066 |

# Variance and correlation parameters

| | | | | | | |
|---|---|---|---|---|---|---|
| var(e.gpa) | .1369695 | .0125304 | | | .1144862 | .1638682 |
| var(e.hsgpa) | .0581203 | .0018605 | | | .0545859 | .0618837 |
| corr(e.ins~l, e.gpa) | .3495295 | .1134498 | 3.08 | 0.002 | .1111427 | .5498816 |
| corr(e.pro~m, e.gpa) | .3140963 | .0799182 | 3.93 | 0.000 | .1501581 | .4612241 |
| corr(e.hsgpa, e.gpa) | .4549455 | .0685265 | 6.64 | 0.000 | .3109127 | .5785514 |
| corr(e.pro~m, e.inschool) | .2068967 | .0448376 | 4.61 | 0.000 | .1175707 | .2929015 |
| corr(e.hsgpa, e.inschool) | .3763213 | .0318662 | 11.81 | 0.000 | .3122227 | .4370091 |
| corr(e.hsgpa, e.program) | .0989748 | .0283577 | 3.49 | 0.000 | .0431431 | .1541902 |

# Main equation

| | | | | | | |
|---|---|---|---|---|---|---|
| gpa | | | | | | |
| program# | | | | | | |
| c.income | | | | | | |
| 0 | .0559082 | .0081052 | 6.90 | 0.000 | .0400223 | .0717942 |
| 1 | .0921056 | .0080322 | 11.47 | 0.000 | .0763629 | .1078483 |
| program# | | | | | | |
| c.hsgpa | | | | | | |
| 0 | 1.142148 | .1282104 | 8.91 | 0.000 | .8908606 | 1.393436 |
| 1 | .9391335 | .131239 | 7.16 | 0.000 | .6819098 | 1.196357 |
| program | | | | | | |
| 0 | -1.051847 | .3449417 | -3.05 | 0.002 | -1.72792 | -.3757735 |
| 1 | -.0869778 | .3550886 | -0.24 | 0.806 | -.7829387 | .6089832 |

- We cannot directly interpret these coefficients.

# Average Treatment Effect (ATE)

- We can use `estat teffects` to estimate the ATE of the study program on college GPA

```
. estat teffects

Predictive margins                              Number of obs     =      2,000
```

| | Unconditional Margin | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ATE | | | | | | |
| program (1 vs 0) | .5620163 | .0478861 | 11.74 | 0.000 | .4681612 | .6558713 |

- The average college GPA is increased by .56 points if everyone participates in the study program instead of no one participating in the study program.
- The robust variance-covariance estimate allowed us to use unconditional standard errors.

# Average Treatment Effect (ATE)

- The standard error and confidence interval are for the population effect.

```
. estat teffects

Predictive margins                              Number of obs     =      2,000
```

| | Unconditional Margin | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ATE | | | | | | |
| program (1 vs 0) | .5620163 | .0478861 | 11.74 | 0.000 | .4681612 | .6558713 |

- When we estimate the ATE, we are using the observed values of the covariates. However, our sample is just one possible draw from the population.
- The population standard error and confidence interval account for this additional randomness when we are averaging over the observations in our sample to estimate the ATE.

# Average Treatment Effect on the Treated (ATET)

- We can also use `estat teffects` to estimate the ATET of the study program on college GPA

```
. estat teffects, atet

Predictive margins                              Number of obs     =      2,000
                                                Subpop. no. obs   =        856
```

| | Unconditional | | | | | |
|---|---|---|---|---|---|---|
| | Margin | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
| ATET | | | | | | |
| program (1 vs 0) | .5489433 | .0480846 | 11.42 | 0.000 | .4546992 | .6431874 |

- The average college GPA is .55 points higher for those who particpate in the program compared to what those students would have scored had they not participated.

# Example

- We have plenty of examples of nonrandom treatment assignment in the documentation:
  - [ERM] example 2b Linear regression with exogenous treatment
  - [ERM] example 2c Linear regression with endogenous treatment
  - [ERM] example 3b Probit regression with endogenous covariate and treatment
  - [ERM] example 4b Probit regression with endogenous treatment and sample selection
  - [ERM] example 5 Probit regression with endogenous ordinal treatment
  - [ERM] example 6a Ordered probit regression with endogenous treatment
  - [ERM] example 6b Ordered probit regression with endogenous treatment and sample selection

# Unobserved components

- We just estimated the parameters of a complex model.
- So far we have only very generally described how this works.
- We can gain some intuition about how ERMs work by using the unobserved component framework.

# Unobserved components

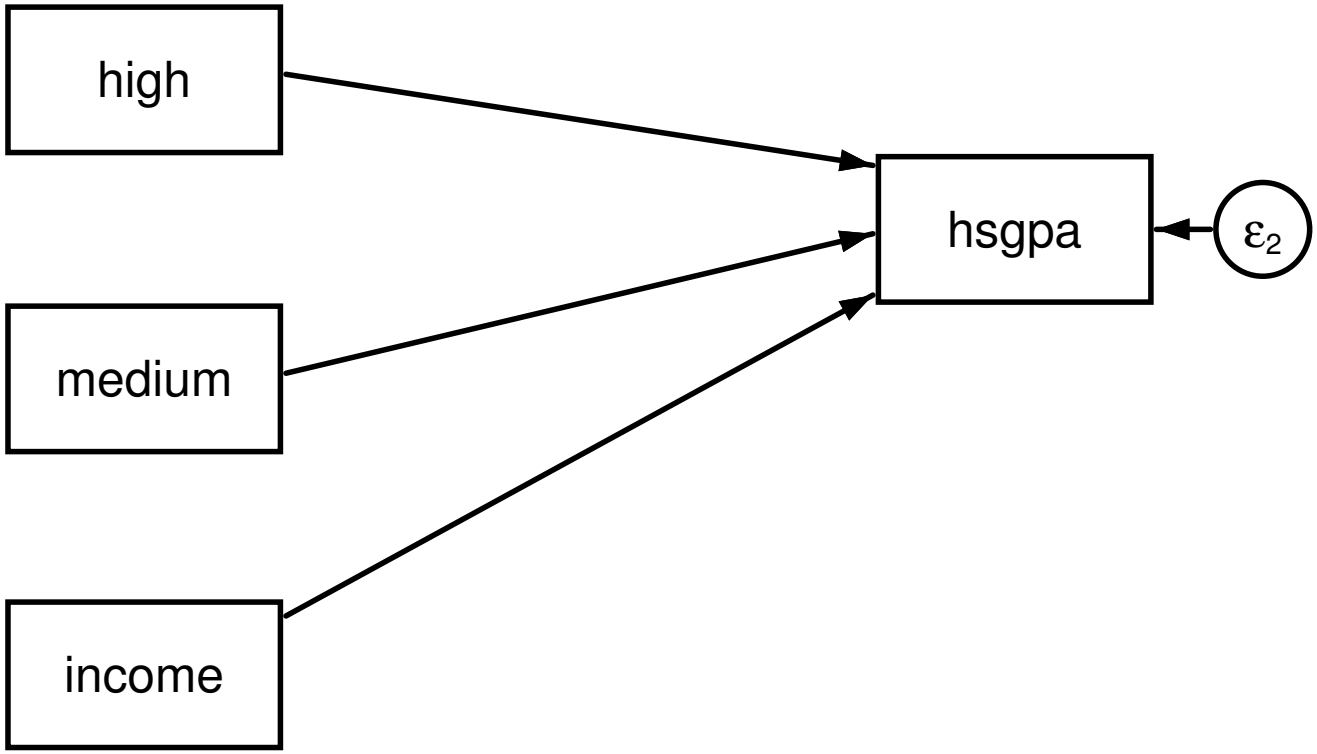
- Suppose an endogenous covariate was our only data issue.
- What if ability was the only unobserved factor that affected both college GPA and high school GPA?

# Unobserved components

- For high school GPA, we have

$$\begin{aligned} \text{hsgpa} = \beta_{21}\text{income} &+ \beta_{22}(\text{hscomp=medium}) \\ &+ \beta_{23}(\text{hscomp=high}) + \beta_{20} + \epsilon_2 \end{aligned}$$

# Unobserved components

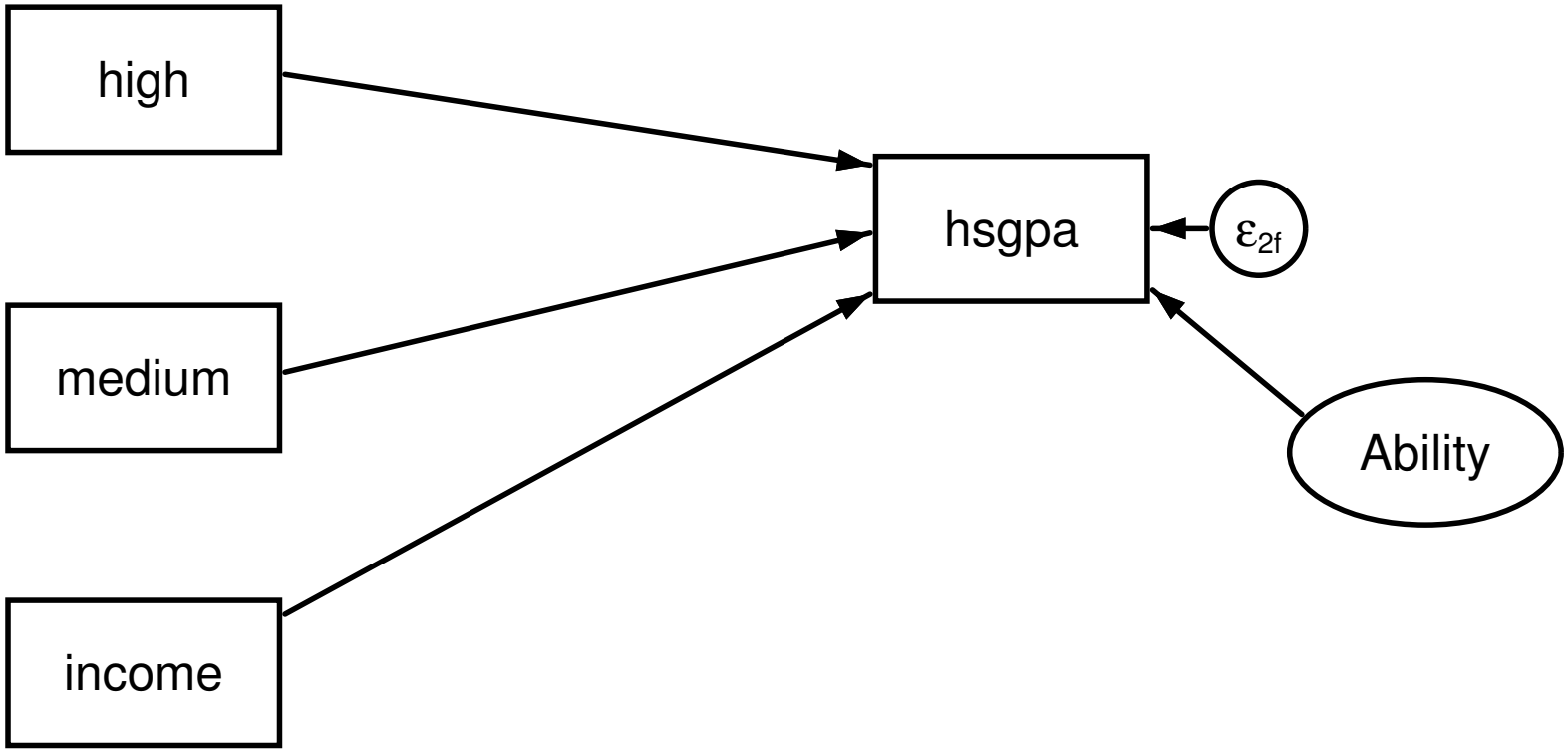


# Unobserved components

- We can decompose $\epsilon_2$ into ability and an independent error $\epsilon_{2f}$

$$\begin{aligned} \text{hsgpa} = \beta_{21}\text{income} &+ \beta_{22}(\text{hscomp=medium}) \\ &+ \beta_{23}(\text{hscomp=high}) + \beta_{20} + \text{ability} + \epsilon_{2f} \end{aligned}$$

# Unobserved components

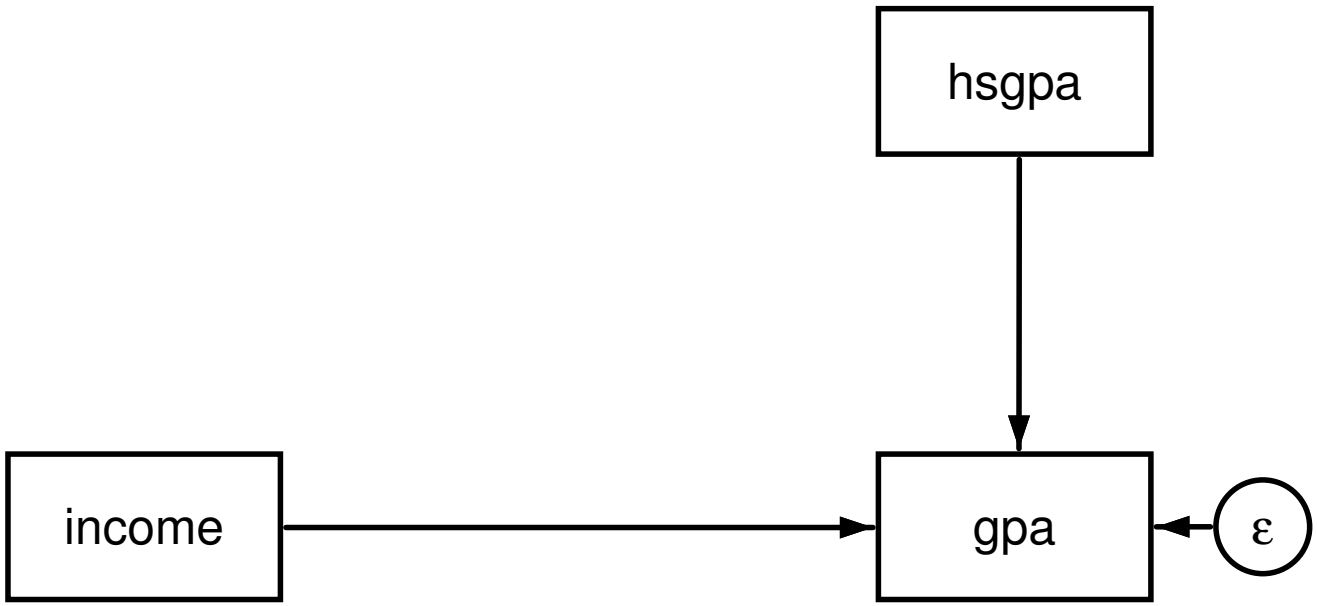


# Unobserved components

- For college GPA we have

$$\text{gpa} = \beta_1 \text{hsgpa} + \beta_2 \text{income} + \beta_0 + \epsilon$$

# Unobserved components

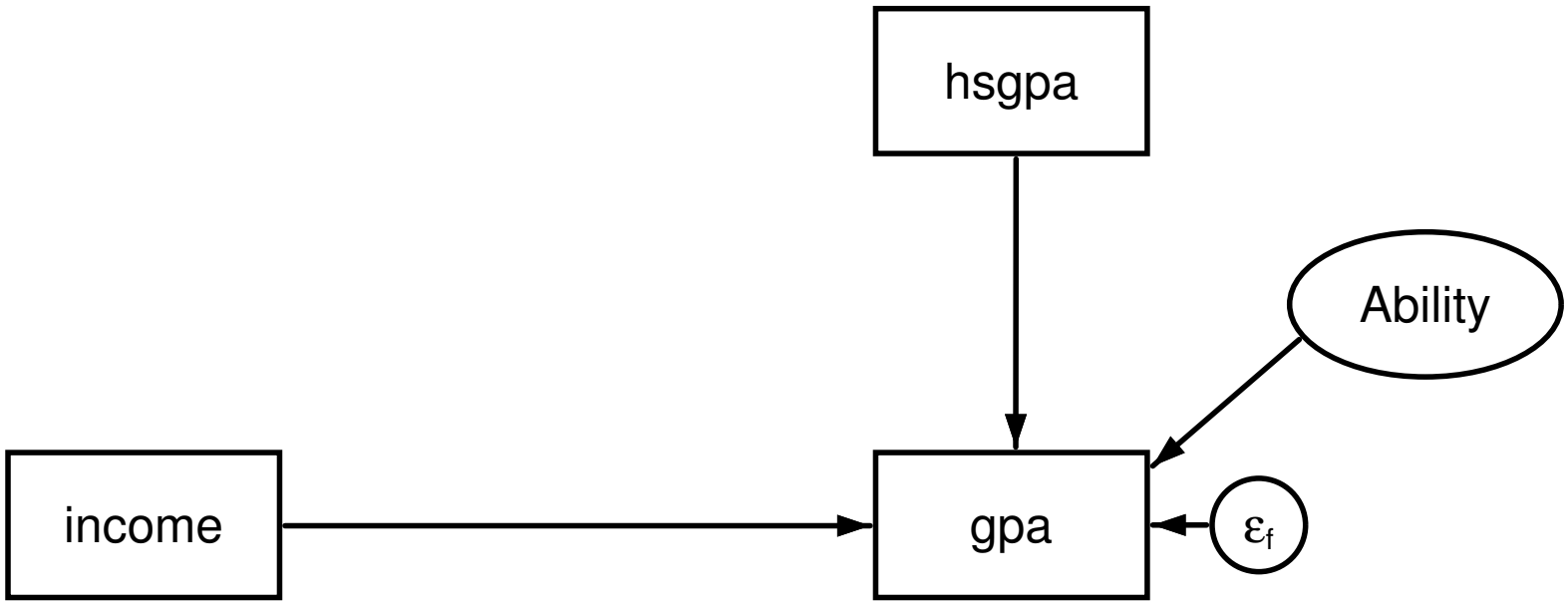


# Unobserved components

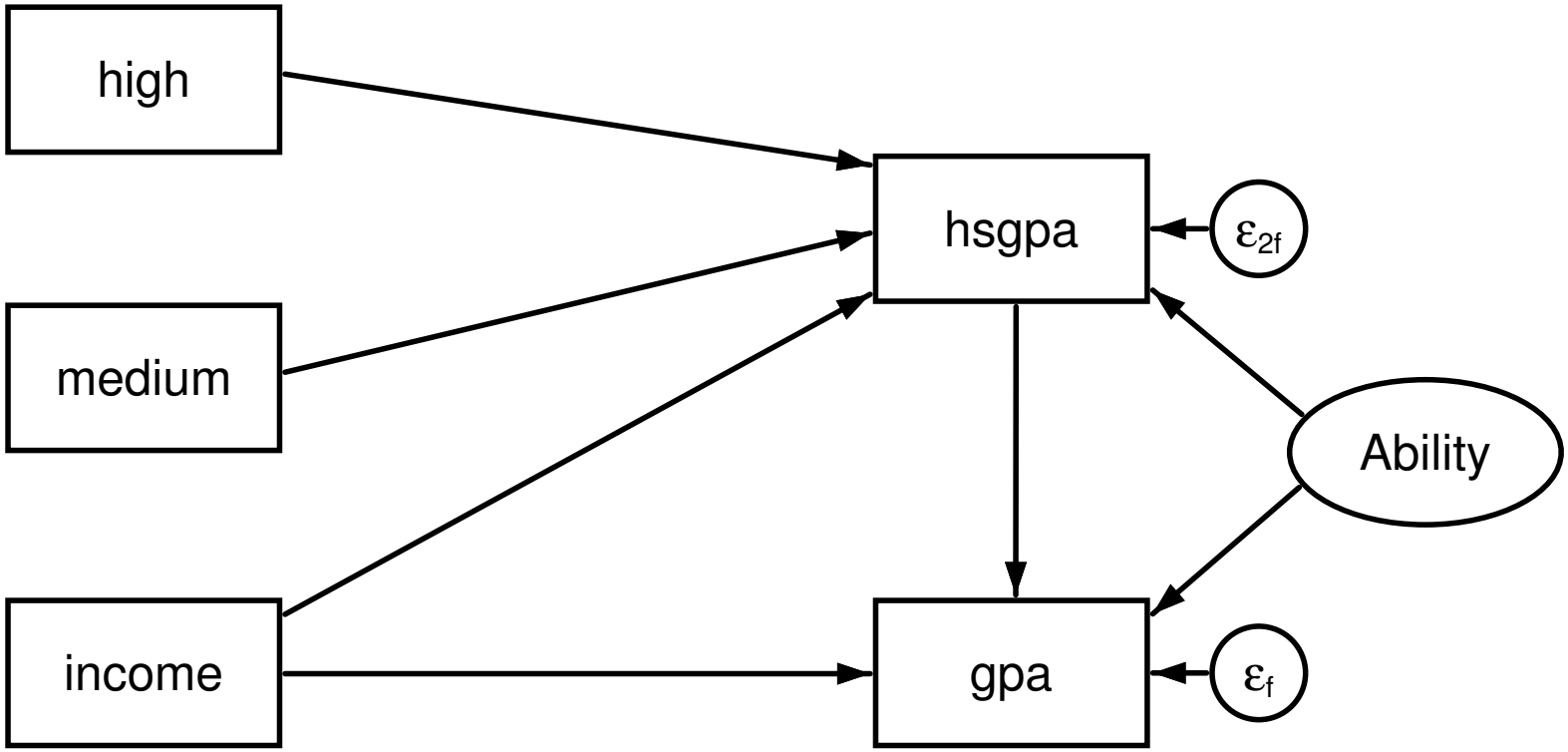
- We can decompose $\epsilon$ into ability and another independent error $\epsilon_f$

$$\text{gpa} = \beta_1 \text{hsgpa} + \beta_2 \text{income} + \beta_0 + \lambda \text{ability} + \epsilon_f$$

# Unobserved components



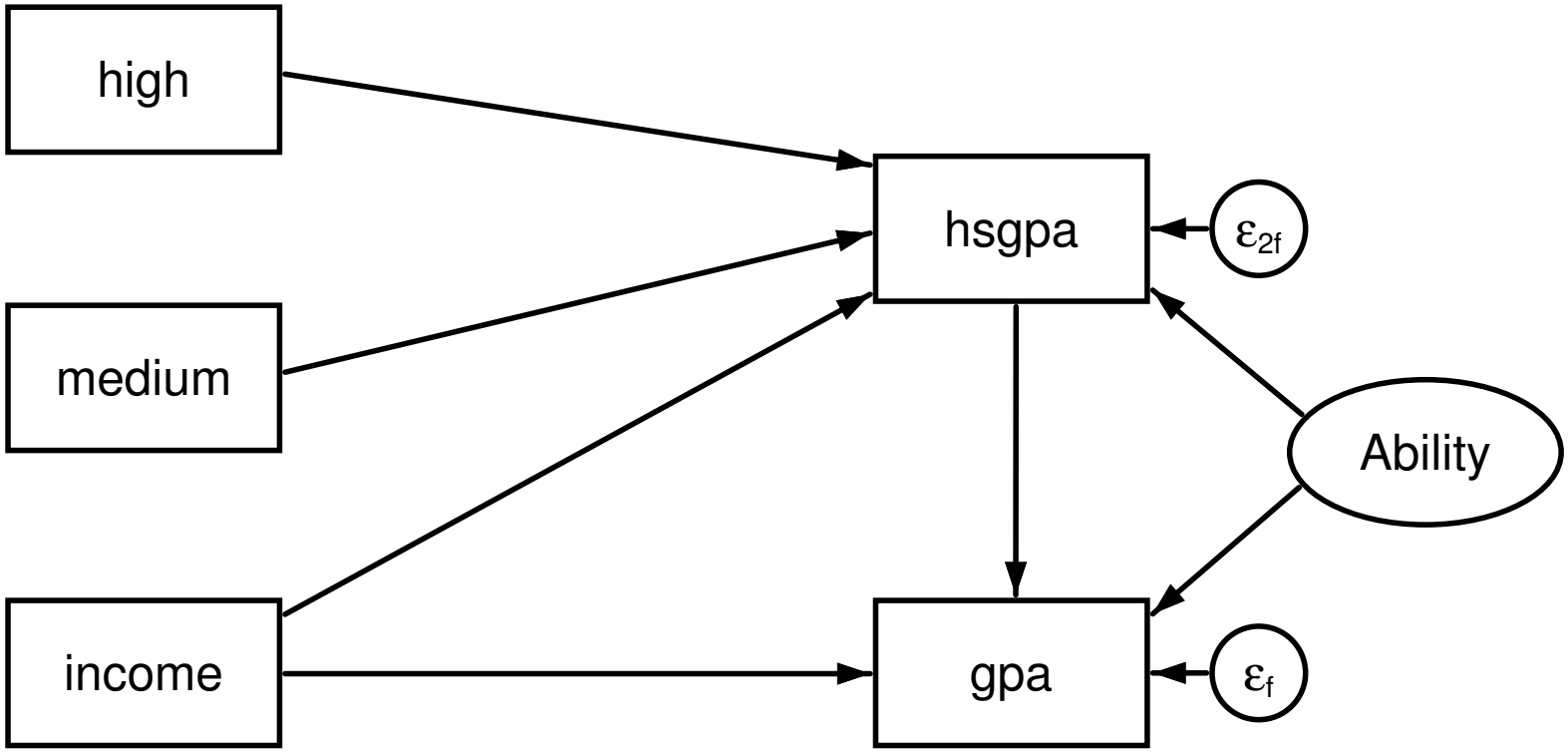
# Unobserved components



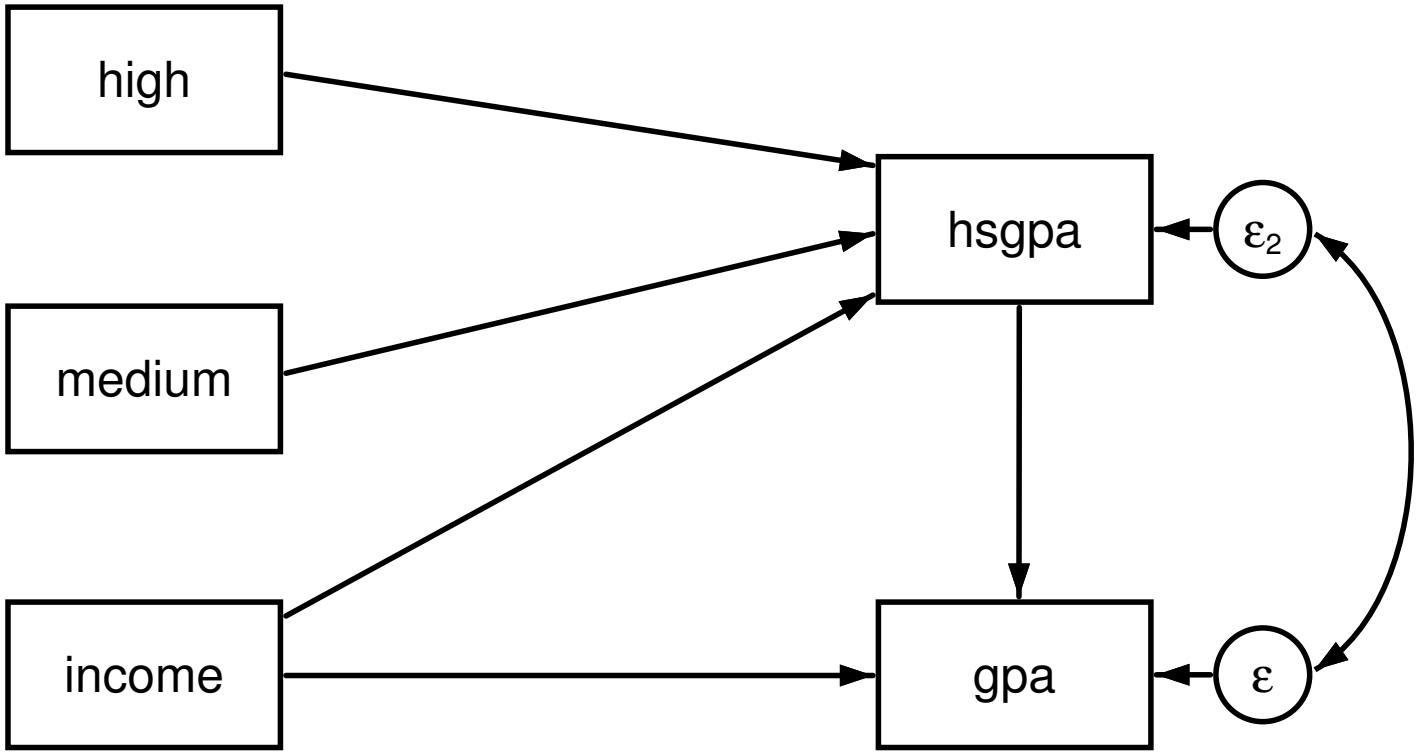
# Unobserved components

- We can do this with other unobserved factors as well.
- The factors would appear in each equation that they affect.
- This applies to the equations for endogenous selection and endogenous treatment as well.
- Our assumption that ability is the only unobserved component is not realistic, but it helps us to understand how the structure of the model is built.
- Intead of using unobserved components, we estimate correlations and variances that are summary parameters for all the unobserved components.
- The parameters are estimated using maximum likelihood.

# Unobserved components



# Unobserved components



# Summary

- I have shown you how to use **eregress** to estimate the parameters of models with endogenous sample selection, endogenous covariates, and nonrandom treatment assignment.
- We also learned about these observational data issues, and this knowledge can be applied to estimating other models.
- But there are many other things that ERM commands can do.
- Let me show you some more examples.

# Examples

- Now suppose that we did not measure the GPA of students with GPA's below 2.0.
- This is a standard tobit-type outcome.
- We have one dependent variable, that records the value 2.0 for anyone with a GPA of 2.0 or less.
- Can we perform this analysis?

# Examples

- Now suppose that we did not measure the GPA of students with GPA's below 2.0.
- This is a standard tobit-type outcome.
- We have one dependent variable, that records the value 2.0 for anyone with a GPA of 2.0 or less.
- Can we perform this analysis?
- Yes, we use `eintreg`.

# Examples

- First we transform our single censored GPA into two separate variables so that we can use interval regresssion.

```
generate gpal = gpa
replace gpal = . if gpa==2
```

# Examples

- Then we use `eintreg`

```
eintreg gpal gpa income,
        entreat(program=scholar income)
        endogenous(hsgpa=i.hscomp income)
        select(inschool=i.roommate income)
        vce(robust)
```

# Examples

- Suppose `graduate` is a binary indicator for whether the student graduated.
- Can we estimate the probability of graduation?

# Examples

- Suppose `graduate` is a binary indicator for whether the student graduated.
- Can we estimate the probability of graduation?
- Yes, we use `eprobit`.

```
eprobit graduate income,
        entreat(program=scholar income)
        endogenous(hsgpa=i.hscomp income)
        select(inschool=i.roommate income)
        vce(robust)
```

# Examples

- What if we wanted to estimate the probability of graduating with honors as well?
- Now suppose `graduate` has three values:
  - 0, did not graduate
  - 1, graduated without honors
  - 2, graduated with honors

# Examples

- What if we wanted to estimate the probability of graduating with honors as well?
- Now suppose `graduate` has three values:
  - 0, did not graduate
  - 1, graduated without honors
  - 2, graduated with honors
- We would use `eoprobit`.

```
eoprobit graduate income,
         entreat(program=scholar income)
         endogenous(hsgpa=i.hscomp income)
         select(inschool=i.roommate income)
         vce(robust)
```

# Examples

- Endogenous covariates can be binary as well as continuous.
- Suppose we wanted to model the effect of diet and exercise on the chance of having a heart attack.
- Diet and exercise are binary, and we suspect that they are endogenous.

# Examples

- Endogenous covariates can be binary as well as continuous.
- Suppose we wanted to model the effect of diet and exercise on the chance of having a heart attack.
- Diet and exercise are binary, and we suspect that they are endogenous.
- We would use `eprobit`.

```
eprobit attack i.exercise#i.diet#c.x,
        endogenous(exercise = x z1, probit)
        endogenous(diet = x z2, probit)
```

# Examples

- We just interacted two endogenous binary covariates.
- We can use interactions of continuous endogenous covariates as well.

# Examples

- We just interacted two endogenous binary covariates.
- We can use interactions of continuous endogenous covariates as well.
- For example,

```
eintreg yl yu x y2 c.y2#c.y2,
        endogenous(y2 = x z1)
```

# Examples

- We do not have to stop with quadratic terms either.

# Examples

- We do not have to stop with quadratic terms either.
- For example,

```
eoprobit y x c.y2#c.x c.y2#c.y2#c.y2 c.y2#c.y3 c.y3#i.b,
         endogenous(y2 = x z1)
         endogenous(y3 = x z2)
         endogenous(b = x z3, oprobit)
```

# Examples

- We do not have to stop with quadratic terms either.
- For example,

```
eoprobit y x c.y2#c.x c.y2#c.y2#c.y2 c.y2#c.y3 c.y3#i.b,
         endogenous(y2 = x z1)
         endogenous(y3 = x z2)
         endogenous(b = x z3, oprobit)
```

# Examples

- We do not have to stop with quadratic terms either.
- For example,

```
eoprobit y x c.y2#c.x c.y2#c.y2#c.y2 c.y2#c.y3 c.y3#i.b,
         endogenous(y2 = x z1)
         endogenous(y3 = x z2)
         endogenous(b = x z3, oprobit)
```

# Examples

- We do not have to stop with quadratic terms either.
- For example,

```
eoprobit y x c.y2#c.x c.y2#c.y2#c.y2 c.y2#c.y3 c.y3#i.b,
         endogenous(y2 = x z1)
         endogenous(y3 = x z2)
         endogenous(b = x z3, oprobit)
```

# Examples

- We do not have to stop with quadratic terms either.
- For example,

```
eoprobit y x c.y2#c.x c.y2#c.y2#c.y2 c.y2#c.y3 c.y3#i.b,
         endogenous(y2 = x z1)
         endogenous(y3 = x z2)
         endogenous(b = x z3, oprobit)
```

# Examples

- In our treatment effects example with the university, we assumed that the the variance of the potential outcome errors $\epsilon_0$ and $\epsilon_1$ was the same.
- We also assumed that the correlations between the potential outcome errors and the other equation errors were the same.
- Both these assumptions can be relaxed when we use the `povariance` and `pocorrelation` options in `entreat()`.

# Command

```
eregress gpa income,
         entreat(program=scholar income,
         povariance pocorrelation)
         endogenous(hsgpa=i.hscomp income)
         select(inschool=i.roommate income)
         vce(robust)
```

# Variance parameters

| | | | | |
|---|---|---|---|---|
| var(e.gpa) | | | | |
| program | | | | |
| 0 | .1262563 | .0127193 | .1036338 | .1538172 |
| 1 | .15904 | .0229821 | .1198129 | .2111101 |
| var(e.hsgpa) | .0581187 | .0018605 | .0545842 | .061882 |

# Correlation parameters

| | | | | | | |
|---|---|---|---|---|---|---|
| corr(e.ins~l, e.gpa) | | | | | | |
| program | | | | | | |
| 0 | .2243906 | .1860848 | 1.21 | 0.228 | -.1545344 | .5457665 |
| 1 | .4720304 | .097983 | 4.82 | 0.000 | .2595068 | .6409472 |
| corr(e.pro~m, e.gpa) | | | | | | |
| program | | | | | | |
| 0 | .3299157 | .1125316 | 2.93 | 0.003 | .0949503 | .530061 |
| 1 | .2922389 | .1053965 | 2.77 | 0.006 | .0750085 | .4829889 |
| corr(e.hsgpa, e.gpa) | | | | | | |
| program | | | | | | |
| 0 | .3318133 | .1040308 | 3.19 | 0.001 | .1152275 | .5182817 |
| 1 | .5876842 | .076013 | 7.73 | 0.000 | .4190482 | .7171271 |
| corr(e.pro~m, e.inschool) | .2072091 | .0447798 | 4.63 | 0.000 | .1179971 | .2931031 |
| corr(e.hsgpa, e.inschool) | .3766597 | .0318127 | 11.84 | 0.000 | .3126693 | .4372466 |
| corr(e.hsgpa, e.program) | .0993276 | .0282984 | 3.51 | 0.000 | .0436121 | .1544272 |

# To Summarize...

- Sample selection?

# To Summarize...

- Sample selection?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs
- Polynomial endogenous covariates?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs
- Polynomial endogenous covariates?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs
- Polynomial endogenous covariates?
  - Use ERMs
- Interactions with endogenous covariates?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs
- Polynomial endogenous covariates?
  - Use ERMs
- Interactions with endogenous covariates?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs
- Polynomial endogenous covariates?
  - Use ERMs
- Interactions with endogenous covariates?
  - Use ERMs
- Binary endogenous covariates?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs
- Polynomial endogenous covariates?
  - Use ERMs
- Interactions with endogenous covariates?
  - Use ERMs
- Binary endogenous covariates?
  - Use ERMs

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs
- Polynomial endogenous covariates?
  - Use ERMs
- Interactions with endogenous covariates?
  - Use ERMs
- Binary endogenous covariates?
  - Use ERMs
- Ordinal endogenous covariates?

# To Summarize...

- Sample selection?
  - Use ERMs
- Endogenous covariates?
  - Use ERMs
- Nonrandom treatment?
  - Use ERMs
- Continous, censored, binary, or ordinal outcomes?
  - Use ERMs
- Need fully conditional inferences?
  - Use ERMs
- Need ATEs or ATETs?
  - Use ERMs
- Polynomial endogenous covariates?
  - Use ERMs
- Interactions with endogenous covariates?
  - Use ERMs
- Binary endogenous covariates?
  - Use ERMs
- Ordinal endogenous covariates?
  - Use ERMs

# To Summarize...

- Any and all combinations of the above?

# To Summarize...

- Any and all combinations of the above?
  - Use ERMs

# Conclusion

- Now you have a taste of what the ERM commands can do.
- Our documentation has more examples and much more information: ERM manual

# Thank you!