Introduction to GSEM in Stata

Christopher F Baum

ECON 8823: Applied Econometrics

Boston College, Spring 2016

Generalized Structural Equation Modeling in Stata

We now present an introduction to Stata's `gsem` command, which extends the facilities of the `sem` command to implement a broader set of applications of structural equation modeling: thus, generalized structural equation modeling. As `gsem` has many capabilities, we can only discuss a limited subset of its features and give some illustrations of its use.

Generalized Linear Model

To understand Stata's extension of the SEM framework, we must introduce the concept of the Generalized Linear Model: something that has been a component of Stata for many years as the `glm` command.

The generalized linear model (GLM) framework of McCullaugh and Nelder (1989) is common in applied work in biostatistics, but has not been widely applied in econometrics. It offers many advantages, and should be more widely known.

GLM estimators are maximum likelihood estimators that are based on a density in the linear exponential family (LEF). These include the normal (Gaussian) and inverse Gaussian for continuous data, Poisson and negative binomial for count data, Bernoulli for binary data (including logit and probit) and Gamma for duration data.

GLM estimators are essentially generalizations of nonlinear least squares, and as such are optimal for a nonlinear regression model with homoskedastic additive errors. They are also appropriate for other types of data which exhibit intrinsic heteroskedasticity where there is a rationale for modeling the heteroskedasticity.

The GLM estimator $\hat{\theta}$ maximizes the log-likelihood

$$Q(\theta) = \sum_{i=1}^{N} \left[a\left(m(x_i, \beta)\right) + b(y_i) + c\left(m(x_i, \beta)\right) \right]$$

where $m(x, \beta) = E(y|x)$ is the conditional mean of y, $a(\cdot)$ and $c(\cdot)$ correspond to different members of the LEF, and $b(\cdot)$ is a normalizing constant.

For instance, for the Poisson, where the mean equals the variance, $a(\mu) = -\mu$ and $c(\mu) = \log(\mu)$. Given definitions of these two functions, the mean and variance are $E(y) = \mu = -a'(\mu)/c'(\mu)$ and $Var(y) = 1/c'(\mu)$. For the Poisson, $a'(\mu) = 1, c'(\mu) = 1/\mu$, so $E(y) = Var(y) = \mu$.

GLM estimators are consistent provided that the conditional mean function is correctly specified: that $E(y_i|x_i) = m(x_i, \beta)$. If the variance function is not correctly specified, a robust estimate of the VCE should be used.

To use the GLM estimator, you must specify two options: the `family()`, which defines the member of the LEF to be employed, and the `link()`, which is the inverse of the conditional mean function. The family option may be chosen as `gaussian, igaussian, binomial, poisson, nbinomial, gamma`.

The link function essentially expresses the transformation to be applied to the dependent variable. Each family has a canonical link, which is chosen if not specified: for instance, `family(gaussian)` has default `link(identity)`, so that a GLM with those two options would essentially be linear regression via maximum likelihood.

The `binomial` family has a default `link(logit)`, while the `poisson` and `nbinomial` families share `link(log)`. However, a number of other combinations of `family` and `link` are valid: for instance, `link(power` *n*) is valid for all distributional families.

The GLM and the GSEM

What, then, is Stata's Generalized Structural Equation Model, or `gsem`? Essentially, the combination of the `sem` modeling capabilities we have discussed thus far with the broader `glm` estimation framework, allowing us to build models that include latent variables as well as response variables that are not continuous measures.

`sem` fits standard linear SEMs, and `gsem` fits generalized SEMs.

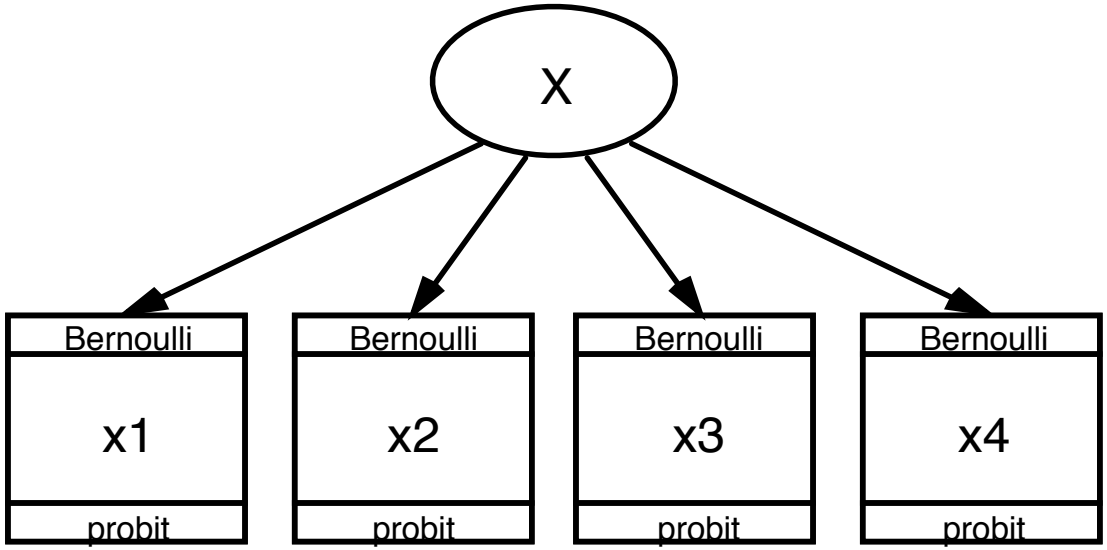
In `sem`, responses are continuous and models are linear regression.

In `gsem`, responses are continuous or binary, ordinal, count, or multinomial. Models are linear regression, gamma regression, logit, probit, ordinal logit, ordinal probit, Poisson, negative binomial, multinomial logit, and more.

`gsem` also has the ability to fit multilevel mixed SEMs. Multilevel mixed models refer to the simultaneous handling of group-level effects, which can be nested or crossed. Thus you can include unobserved and observed effects for subjects, subjects within group, group within subgroup, ... , or for subjects, group, subgroup, ... This extends Stata's `mixed` framework.

Models supported by GSEM

We now consider a number of models that are supported by the SEM methodology. The first is the *single-factor measurement model*, in which we consider several observed variables as influencing a single latent factor, as we considered earlier. The difference is that we now allow for a *generalized response*, rather than assuming that the response is continuous, driven by Gaussian errors. This can be graphically represented:


X
Bernoulli
x1
probit
Bernoulli
x2
probit
Bernoulli
x3
probit
Bernoulli
x4
probit

In this model, we have four observed factors, each of which is a binary (pass/fail) outcome. The latent factor, being related to only binary measurements, will have different properties than a model based on continuous measurements. Thus, the errors are presumed to follow a Bernoulli distribution, and the GLM link function is the probit. Notice that those specifications show up in the graphical diagram. We may implement this model using `gsem` as:

```
gsem (x1 x2 x3 x4 <-X), probit
```

If one or more of these measurements was continuous, we could use a different family and link for that part of the model. Say that measurement 4 was not only a pass/fail mark, but the score on a test. Then that equation would be fit with the `gsem` default of Gaussian errors and the Identity link.

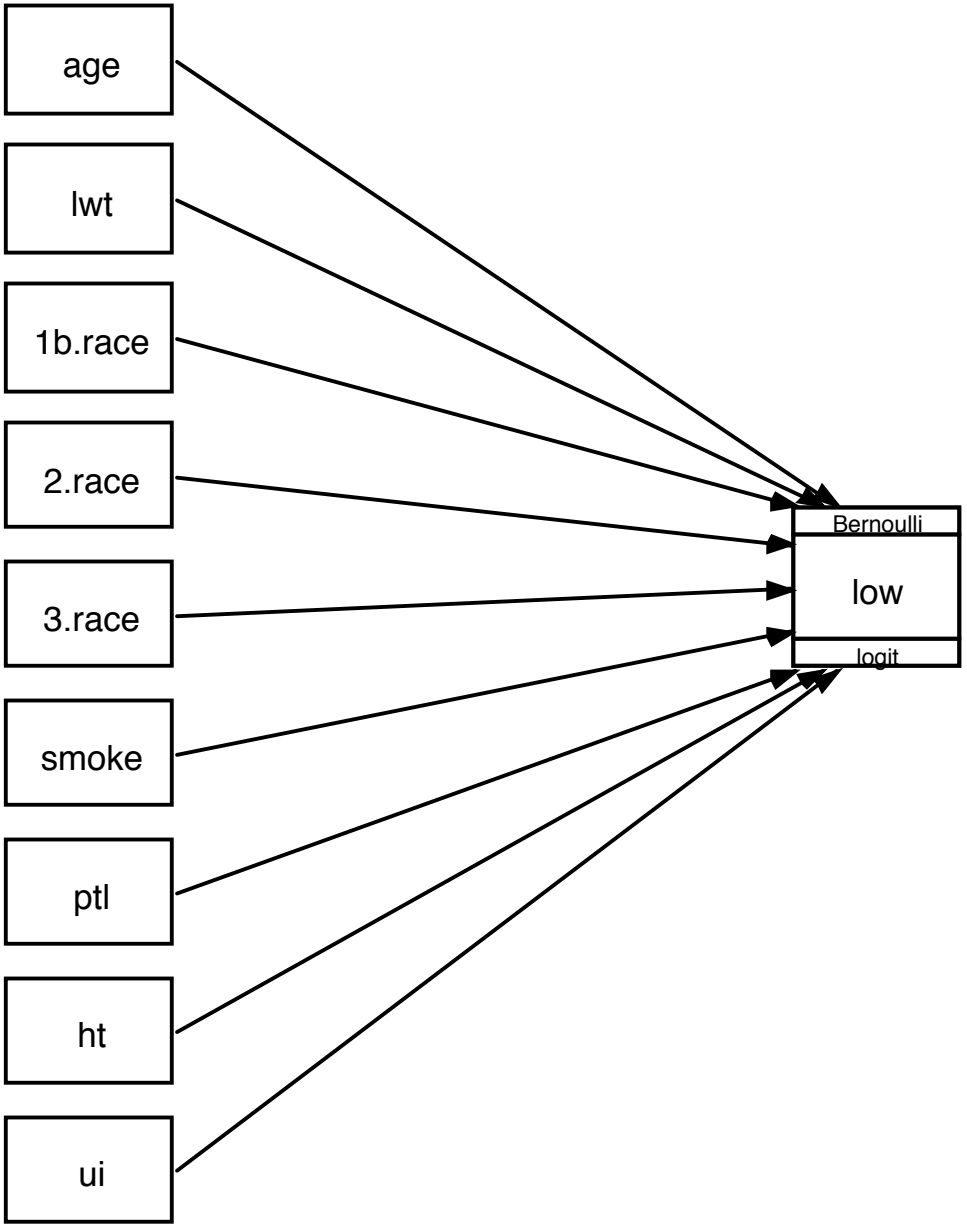
```
gsem (x1 x2 x3 <-X, probit) (s4<-X)
```

Logistic regression

We could use `gsem` to fit a standard logistic regression, which is equivalent to the logit model in the GLM framework. The model here considers the probability of low birth weight as related to a number of observed factors about the mother's medical condition, weight, race, and smoking status.
We may implement this model using `gsem` as:

```
gsem (low <- age lwt i.race smoke ptl ht ui), logit
```

where `i.race` is the standard factor variable notation, indicating that one race should be omitted and indicator variables created for each of the other race categories. Graphically:

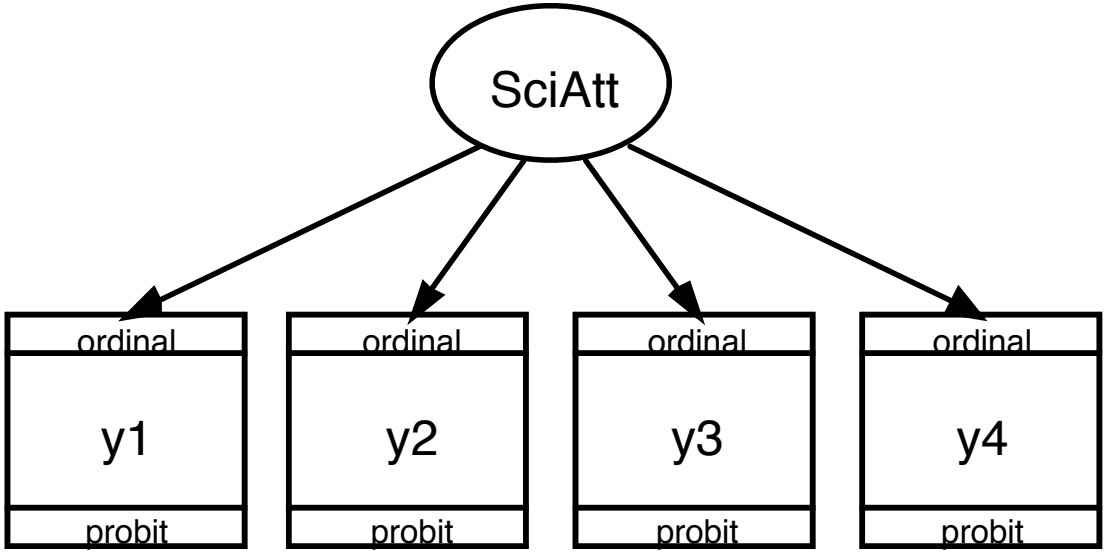

age
lwt
1b.race
2.race
3.race
smoke
ptl
ht
ui
Bernoulli
low
logit

Ordered probit and ordered logit

We can also use ordered probit or ordered logit models in the GSEM framework to deal with variables, such as responses on a Likert scale, where there is assumed to be an underlying factor, with ranges of that latent variable 'binned' into observed discrete categories.
We may implement this model for a latent factor, relating attitudes toward science in a pure measurement framework to four Likert sale variables, as:

```
gsem (y1 y2 y3 y4 <- SciAtt), oprobit
```

Ordered logit could also be used, yielding almost identical results, while making use of the logistic distribution rather than the Gaussian.

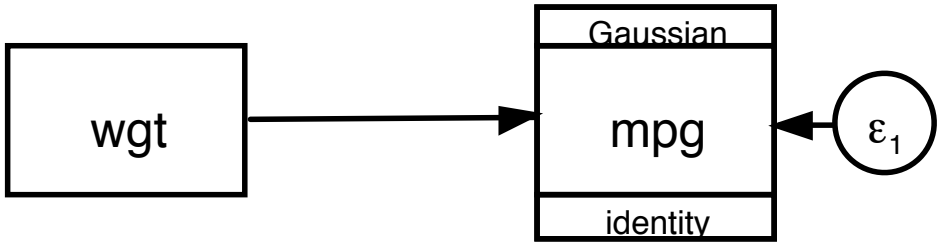

SciAtt
ordinal
y1
probit
ordinal
y2
probit
ordinal
y3
probit
ordinal
y4
probit

Tobit model

The Tobit regression model combines a binary outcome, which indicates censoring, and a continuous outcome for uncensored observations. Censoring may be from below, above or both. For instance, we may have a response to "how much did you spend on a new car last year?", where responses of 0 indicate non-purchase. This may be implemented as:

```
gsem mpg <- wgt, family(gaussian, lcensored(17))
```

where the `lcensored` option indicates that left-censoring at the value 17 is applied. Graphically:

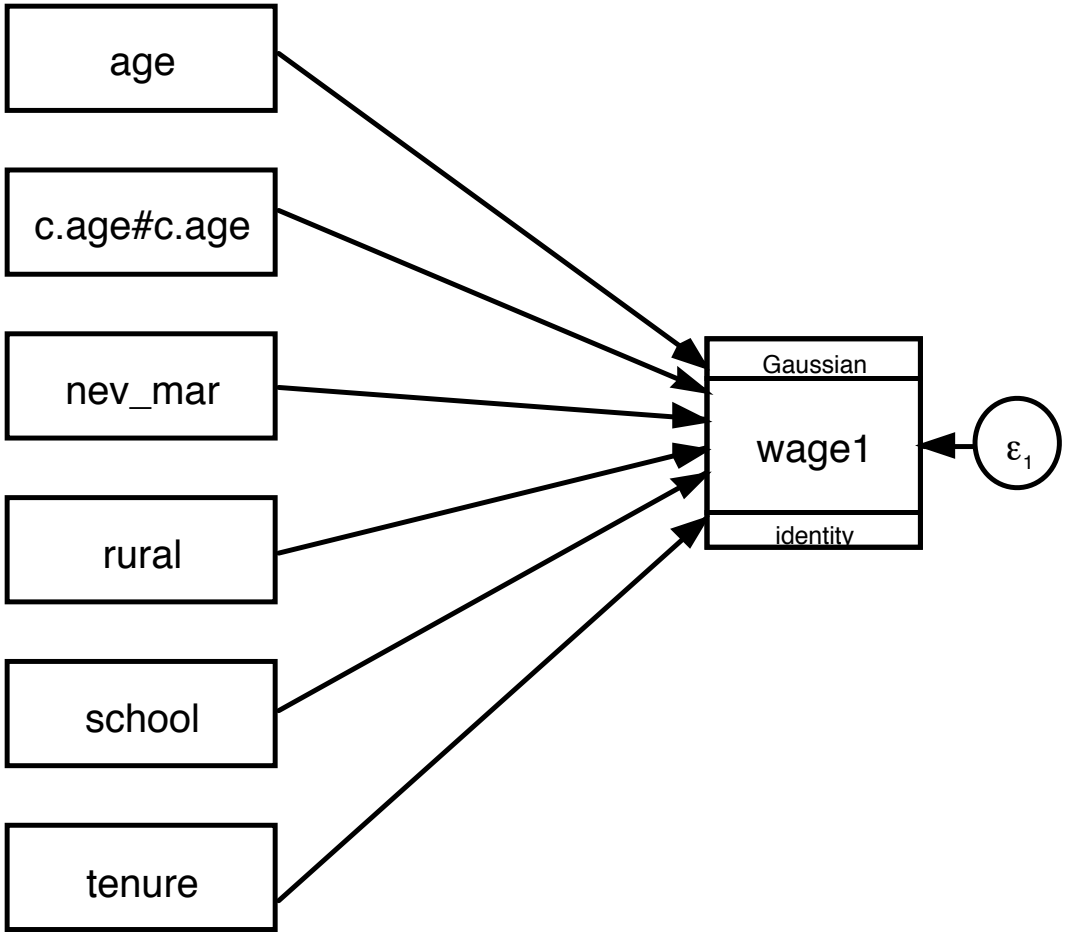

wgt
Gaussian
mpg
identity
ε_1

Interval regression

Interval regression (as implemented by Stata's `intreg`) fits a model where the response lies in an interval, as described by two dependent variables. For instance, from a survey we may have the information that a worker's wage is between $10.00 and $11.99, or between $12.00 and $13.99. Those values would appear as the lower and upper limits in the interval regression. The GSEM implementation of this model can be represented as:

```
gsem wage1 <- age c.age#c.age nev_mar rural school tenure, ///
family(gaussian, udepvar(wage2))
```

where `wage1` would be the lower-limit values, and the `udepvar` option specifies the variable containing the upper-limit values. Graphically:


age
c.age#c.age
nev_mar
rural
school
tenure
Gaussian
wage1
identity
ε1

Heckman selection model

The Heckman regression with selection, as implemented in `heckman`, can also be considered in the GSEM framework. This model deals with a continuous outcome that is observed only when another equation determines that the observation is selected, and the errors of the two equations are allowed to be correlated. Subjects often choose to participate in an event or medical trial or even the labor market, and thus the outcome of interest might be correlated with the decision to participate.

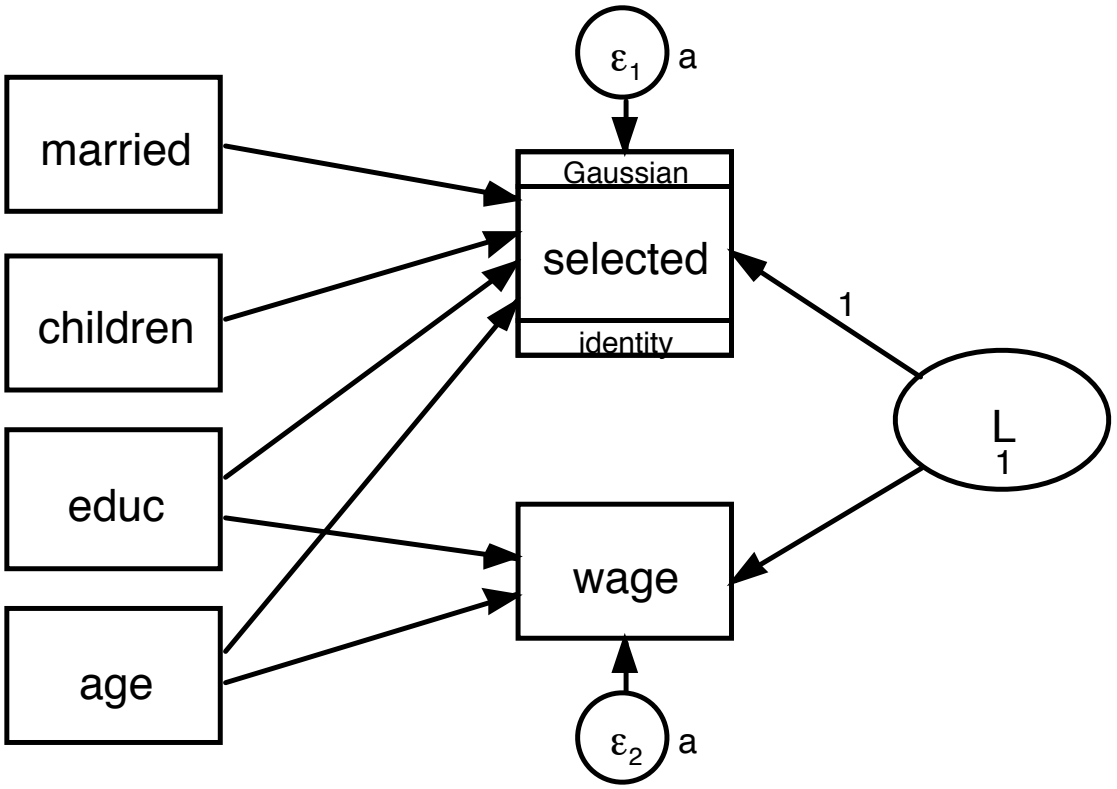
The Heckman selection model can be recast as a two-equation SEM—one linear regression (for the continuous outcome) and the other censored regression (for selection)—and with a latent variable L added to both equations. The latent variable is constrained to have variance 1 and to have coefficient 1 in the selection equation, leaving only the coefficient in the continuous-outcome equation to be estimated. For identification, the variance from the censored regression will be constrained to be equal to that of the linear regression.

This may be implemented as:

```
gsem (wage <- educ age L)
(selected <- married children educ age L@1,
family(gaussian, udepvar(notselected))),
var(L@1 e.wage@a e.selected@a)
```

where the variable `wage` is only observed when the `notselected` variable is 0. The `selected` and `notselected` variables are complements. The variables `married` and `children` are assumed to only affect the probability of labor force participation, while `educ` and `age` are presumed to affect both LFP and the level of the wage for working women.

Like Roodman's `cmp`, Stata considers missing values on an equation-by-equation basis, so the fact that `wage` is missing for non-working women is not a problem. Graphically:


ε1
a
married
Gaussian
selected
identity
children
1
L
1
educ
wage
age
ε2
a

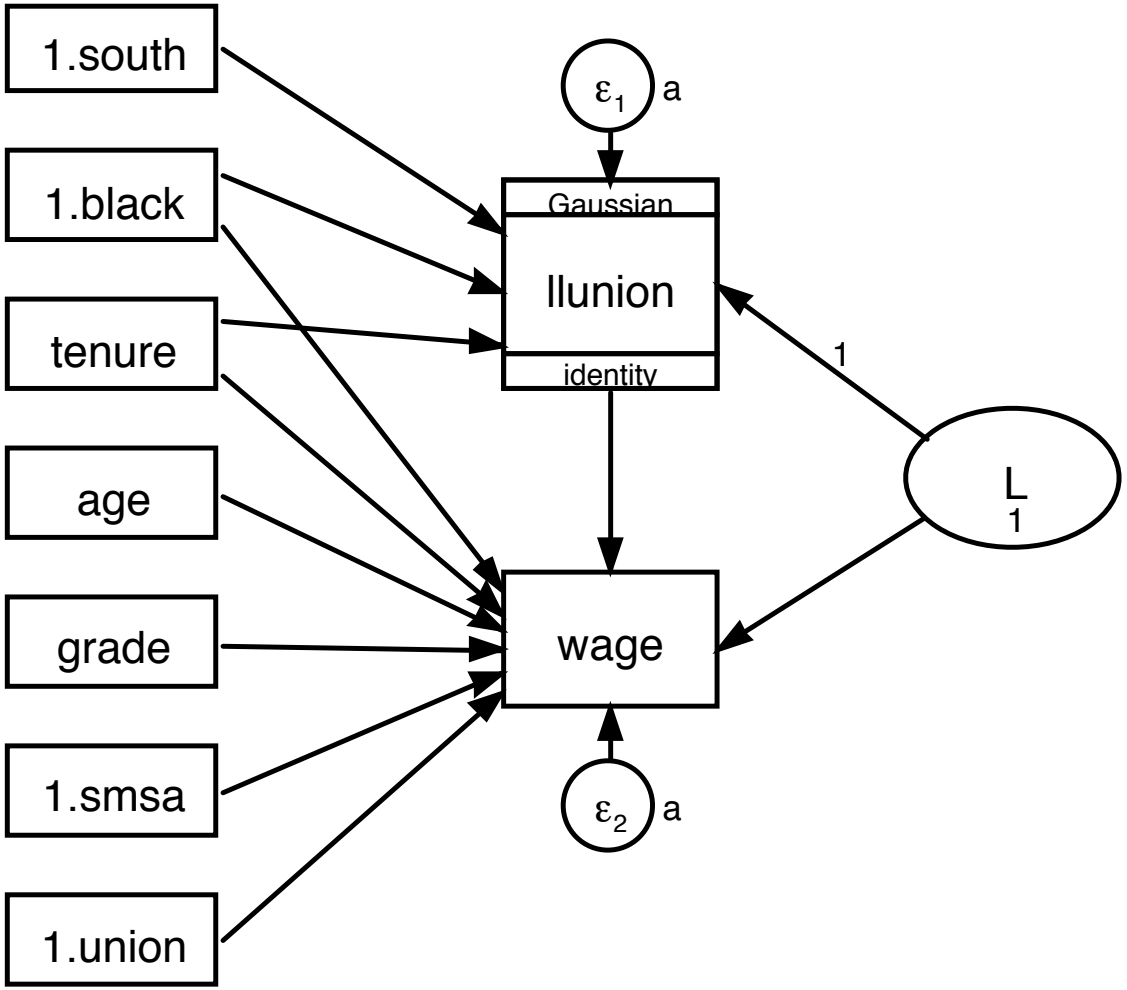
Endogenous treatment-effects model

The treatment-effects model attempts to measure the effect of a "treatment" on a continuous outcome. For instance, we might hypothesize that belonging to a labor union has an effect on wages, and we want to measure the effect. This differs from the Heckman selection model in that here we observe the outcome—the wage—for all observations.

The econometric problem is that those persons with certain characteristics, for instance, higher education, might be more or less likely to be 'treated'. Thus, we must take account of the non-experimental nature of the data at hand, rather than merely regressing `wage` on `union` with controls.

We may implement this, similar to the Heckman model, with a latent variable related to the probability of being treated. Variables `llunion` and `ulunion` are complements, reflecting the observed `union` indicator.

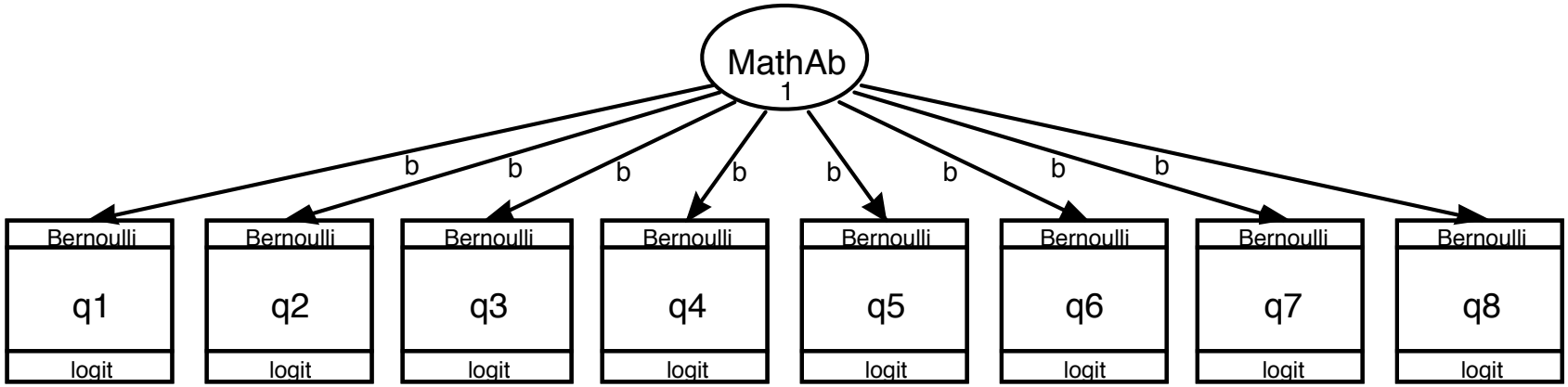
```
gsem (wage <- age grade i.smsa i.black tenure 1.union L)
(llunion <- i.black tenure i.south L@1,
family(gaussian, udepvar(ulunion))),
var(L@1 e.wage@a e.llunion@a)
```


1.south
1.black
tenure
age
grade
1.smsa
1.union
ε1
a
Gaussian
llunion
identity
1
L
1
wage
ε2
a

One-parameter IRT (Rasch) model

The GSEM framework may be used to implement Item Response Theory (IRT) models such as the Rasch model. In this example, we have eight binary measurements from a math test, and we want to generate a single latent factor, Math Ability, and evaluate how difficult each of the questions were. This can be done by constraining the effects of each observed variable on the latent factor to 1; the estimated intercepts then gauge difficulty. A logit link is used, as all of the observed variables follow a Bernoulli distribution.

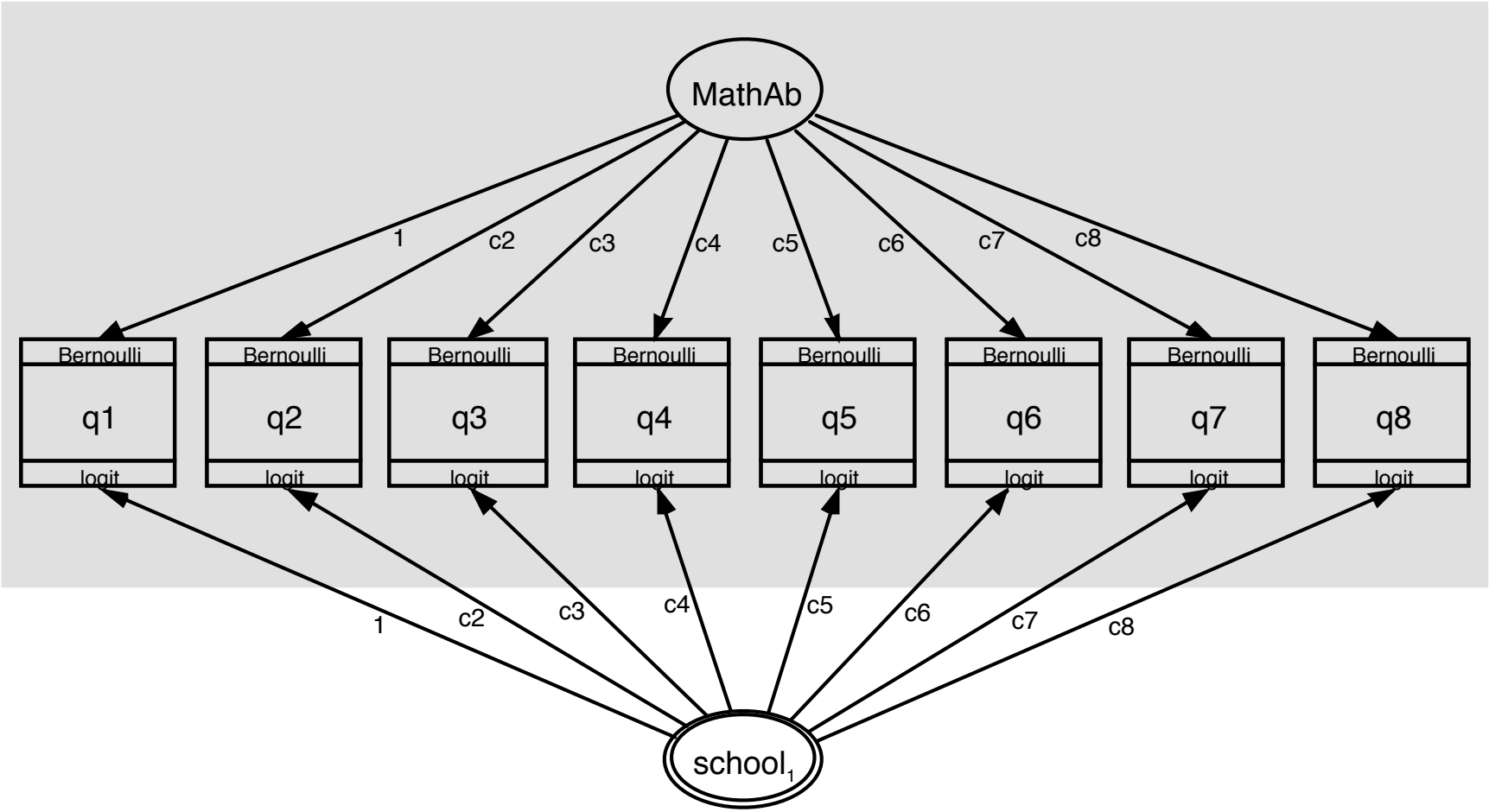
```
gsem (MathAb -> (q1-q8)@1), logit
```


MathAb
1
b
b
b
b
b
b
b
b
Bernoulli
q1
logit
Bernoulli
q2
logit
Bernoulli
q3
logit
Bernoulli
q4
logit
Bernoulli
q5
logit
Bernoulli
q6
logit
Bernoulli
q7
logit
Bernoulli
q8
logit

Two-level measurement model (multilevel, generalized response)

We consider the Math Ability problem, noting that students are nested within schools. We include a latent variable at the school level to account for possible school-by-school effects. This makes the estimation problem into a multilevel model. In the graphical representation, `school` shows up as a latent variable at the school level.

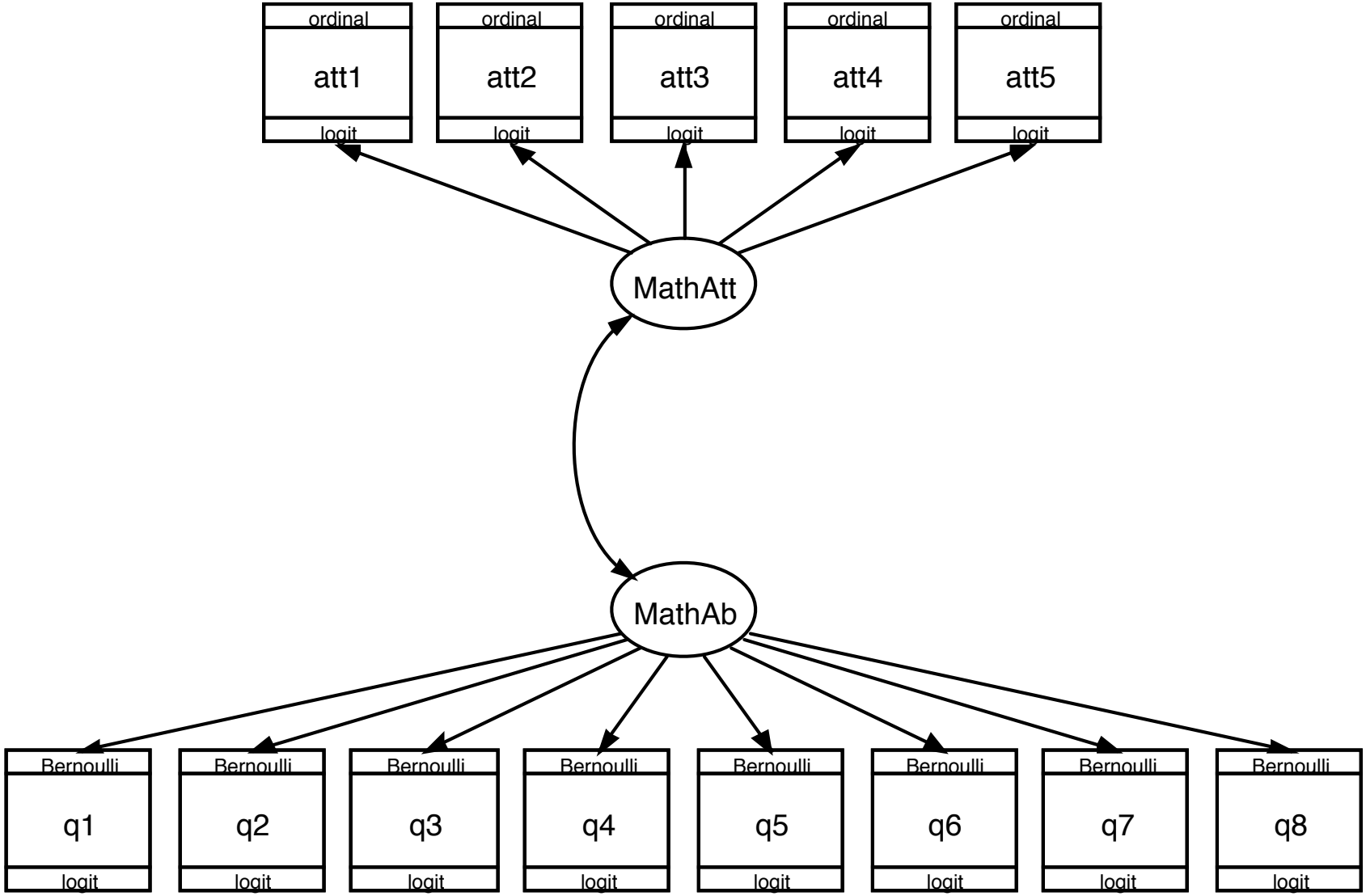
```
gsem (MathAb M1[school] -> q1-q8), logit
```


MathAb
1
c2
c3
c4
c5
c6
c7
c8
Bernoulli
q1
q2
q3
q4
q5
q6
q7
q8
logit
1
c2
c3
c4
c5
c6
c7
c8
school1

Two-factor measurement model (generalized response)

This two-factor measurement model contains two latent variables: one measuring Math Ability and a second measuring Math Attitude. While the observed variables related to Ability are pass/fail grades on an eight-question test, those underlying Math Attitude are Likert-scale variables, necessitating the use of an ordinal estimator. The two latent factors are assumed to be correlated. The code:

```
gsem (MathAb -> q1-q8, logit) ///
(MathAtt -> att1-att5, ologit)
```

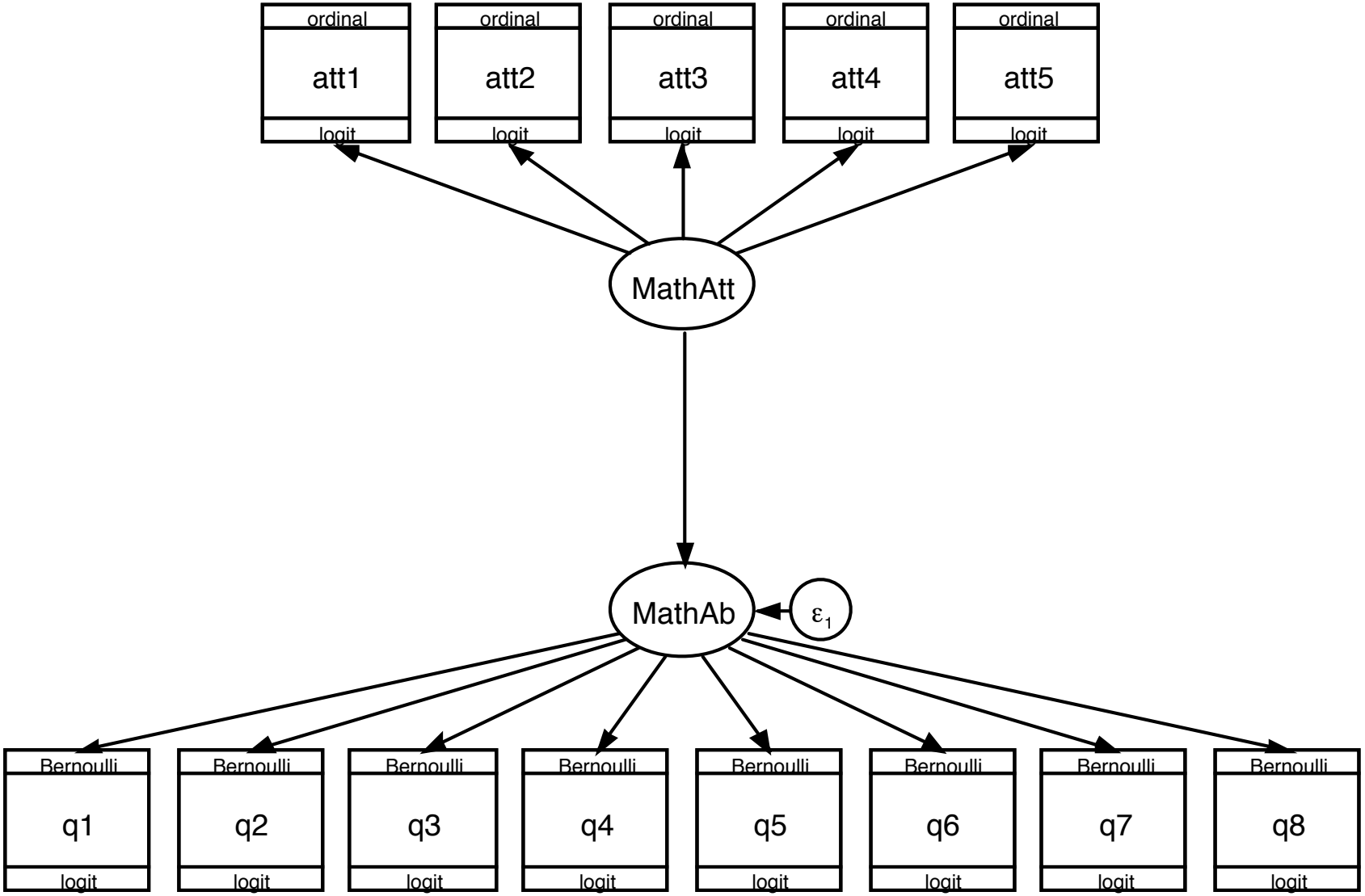

ordinal
att1
logit
ordinal
att2
logit
ordinal
att3
logit
ordinal
att4
logit
ordinal
att5
logit
MathAtt
MathAb
Bernoulli
q1
logit
Bernoulli
q2
logit
Bernoulli
q3
logit
Bernoulli
q4
logit
Bernoulli
q5
logit
Bernoulli
q6
logit
Bernoulli
q7
logit
Bernoulli
q8
logit

Full structural equation model (generalized response)

In the prior model, we allowed for (and estimated) a covariance between the two latent factors, Math Ability and Math Attitude. We now add a structural component to the model by assuming that there is a causal relationship between Math Attitude and Math Ability. This allows us to test a hypothesis regarding the way in which attitudes toward math may affect ability, as evidenced by the observed test scores.

```
gsem (MathAb -> q1-q8, logit) ///
(MathAtt -> att1-att5, ologit) ///
(MathAtt -> MathAb)
```

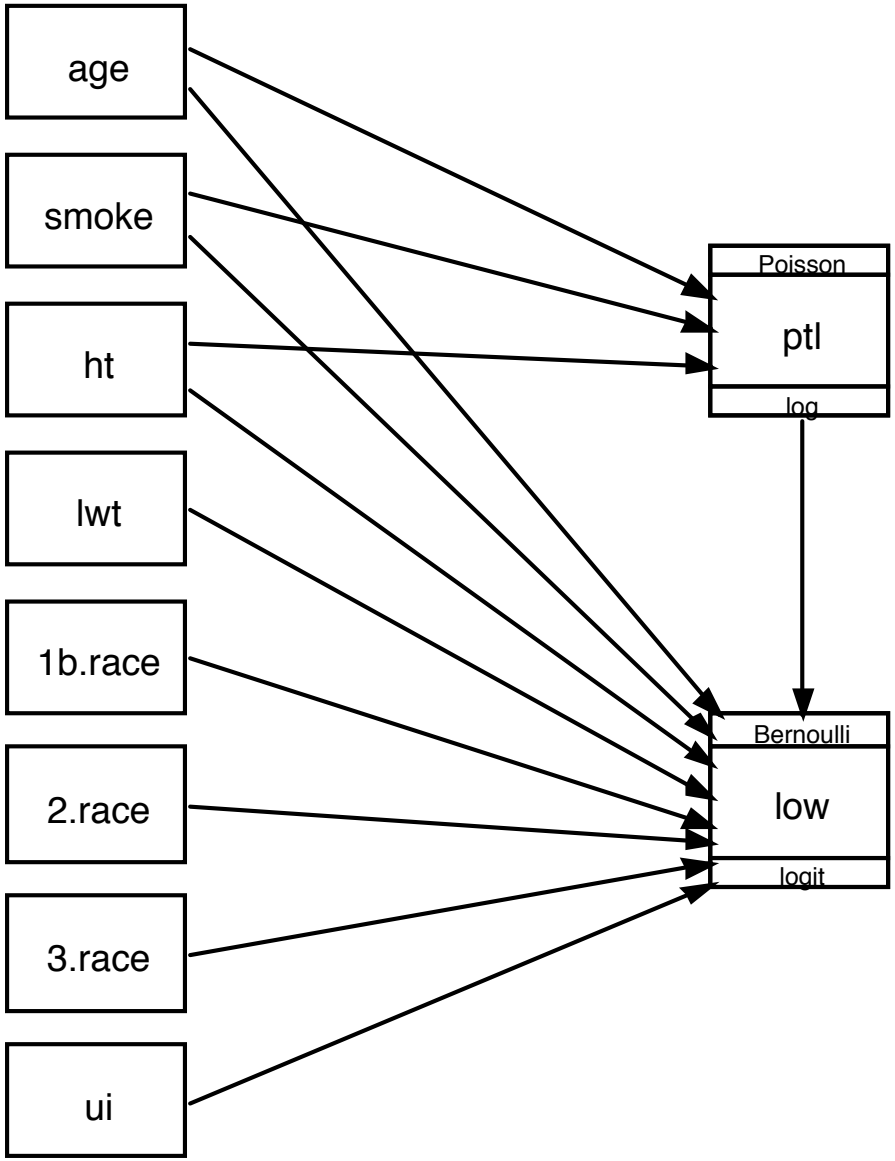
As MathAb is now endogenous, an error term appears in the specification.


ordinal
att1
logit
ordinal
att2
logit
ordinal
att3
logit
ordinal
att4
logit
ordinal
att5
logit
MathAtt
MathAb
ε_1
Bernoulli
q1
logit
Bernoulli
q2
logit
Bernoulli
q3
logit
Bernoulli
q4
logit
Bernoulli
q5
logit
Bernoulli
q6
logit
Bernoulli
q7
logit
Bernoulli
q8
logit

We now return to a framework similar to that implemented by Roodman's `cmp`. In this application, we combine a logit equation and a Poisson regression equation. The logit models the probability of low birthweight, while the Poisson counts the number of premature episodes of labor encountered by the mother. We posit a causal relationship between premature labor and low birth weight.

```
gsem (low <- ptl age smoke ht lwt i.race ui, logit) ///
(ptl <- age smoke ht, poisson)
```

Age, smoking status and an indicator of hypertension are assumed to affect both outcomes, where other controls only enter the logit equation.


age
smoke
ht
lwt
1b.race
2.race
3.race
ui
Poisson
ptl
log
Bernoulli
low
logit

Although we will not illustrate these models, the GSEM framework can also be used to implement:

- MIMIC model (generalized response)
- Multinomial logistic regression
- Random-intercept and random-slope models (multilevel)
- Crossed models (multilevel)

and a number of others. See Stata's [SEM] manual for details.