RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Università della Svizzera italiana

An introduction to

Latent Class Modelling

LCM with Stata: how does it work?

Lugano, 9th of September, 2019

Barbara Antonioli Mantegazzini

Outline of the presentation

RISIS

Premise: to run our LCM we use the **Stata 15** (gsem) package (updated version 15.1)

From theory to practice with **Stata**:

1. Premise: LCM a basic classification
2. Brief introduction to **Stata**.
3. LPA and LCRM with **gsem**
4. Run the model!
5. How to interpret results?

Università della Svizzera italiana

Outline of the presentation

RISIS

From theory to practice with **Stata**:

1. **Premise: LCM a basic classification**

Università della Svizzera italiana

1. Premise: LCM a basic classification

RISIS

LC Models are typically classified according to:

- Nature of the response/observed variables (discrete or continuous)
- Nature of the latent variables (discrete or continuous)

LPA/ LCA

- Inclusion or not of individual covariates

LCRM

What does could LCA tell us?

RISIS

LCA contains two parts

LCA fits the probabilities of which observations belongs to which class (probability class membership)

LCA describes the relationship between the classes and the observed variables

Università della Svizzera italiana

Advantages of Latent Class Analysis

RISIS

- A case can be classified into each class even if there are some missing data
- Parameters can be estimated even if there are missing data (all the available data will be used)
- Probabilistic assignement of cases into classes based on the higher Log Likelihood
- Statistical criteria to select the number of classes (ex: AIC criterion)

Two problems could arise:

RISIS

A. Too many latent classes, with an associated increased likelihood of a local maximum solution -> **local maximum solution**

B. Failure to account for local dependance among manifest variables -> **conditional dependence**

A. Local maximum solution

- Ideally the estimation algorithm will converge on the global maximum solution--the parameter values associated with the single largest log L.
- However no existing LCA algorithm can distinguish between a global maximum and a local maximum of log L. If a solution is reached that is locally optimal--such that a minor change in any parameter value decreases log L--the algorithm will terminate
- There is a potential risk.
  - How to avoid it?

Università della Svizzera italiana

How to avoid local maximum solutions

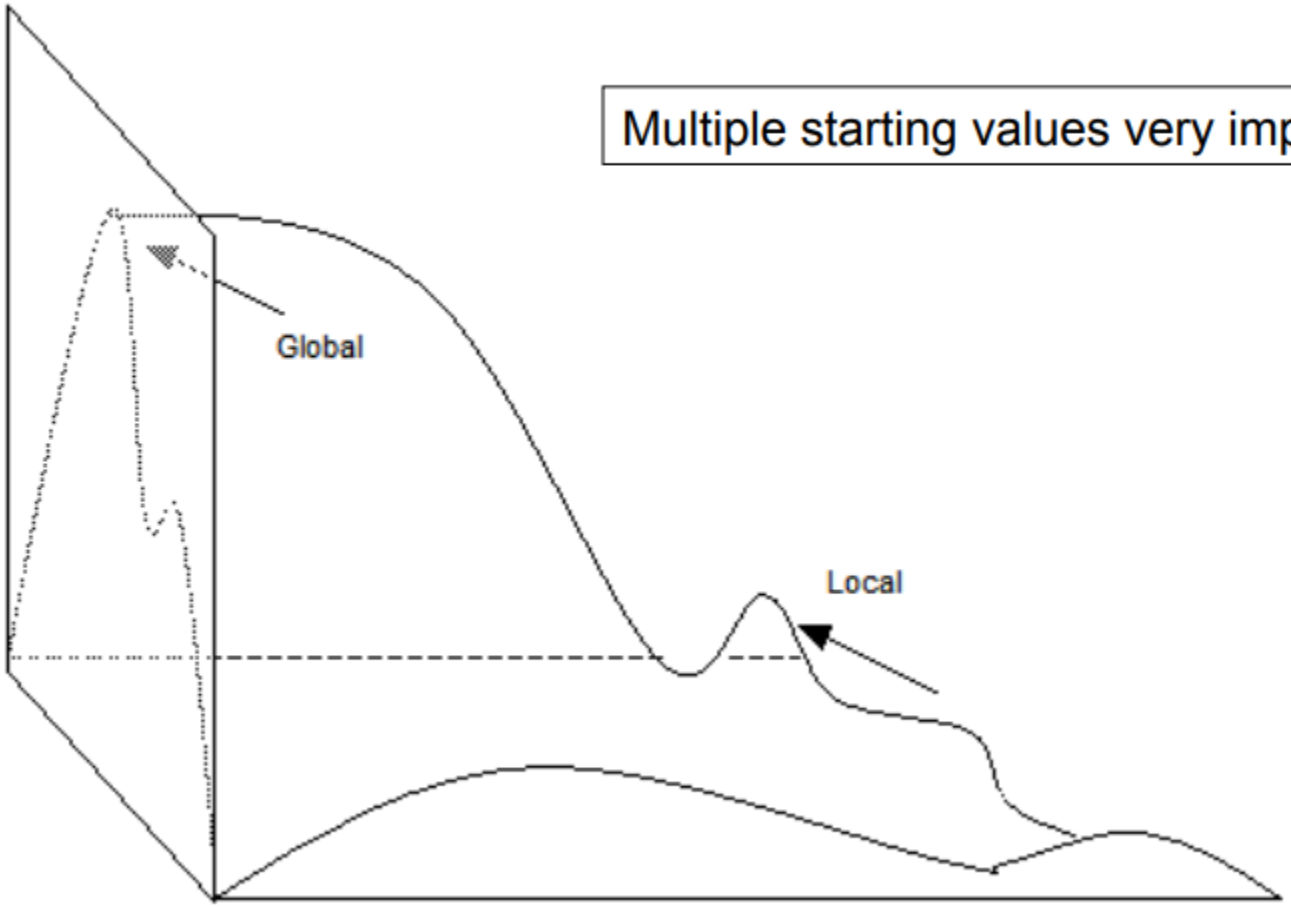
RISIS

- Keep number of latent classes as few as necessary
  - No larger models (AIC and BIC tests may help)
- Always test multiple start value
  - For any model being considered, run the program at least five different times using different random start values (**see Stata syntax command**)
  - If all five runs converge to the same solution, accept that as the global maximum. Otherwise, run the program another five or more times..
- When appropriate, use unidimensional latent class models
  - Also known as discrete latent trait models. It is meant that one can imagine the latent classes as corresponding to *gradations* of some underlying trait, such as disease severity (low, medium or high). For a given number of latent classes, unidimensional LCMs are less prone to local maximum solutions.

Università della Svizzera italiana

Global and local maxima

RISIS



Università della Svizzera italiana

To sum up:

RISIS

Model convergence and robustness:

- Sample with high number of observations and high degree of freedom > model design
- Highly discriminative variables (necessary to create classes/subgroups) > variable design/selection
- Local optima/dependence on starting conditions > test robustness with different starting conditions

Model usefulness:

- Classes interpretability

Università della Svizzera italiana

From LCA to Latent Class Regression Model

RISIS

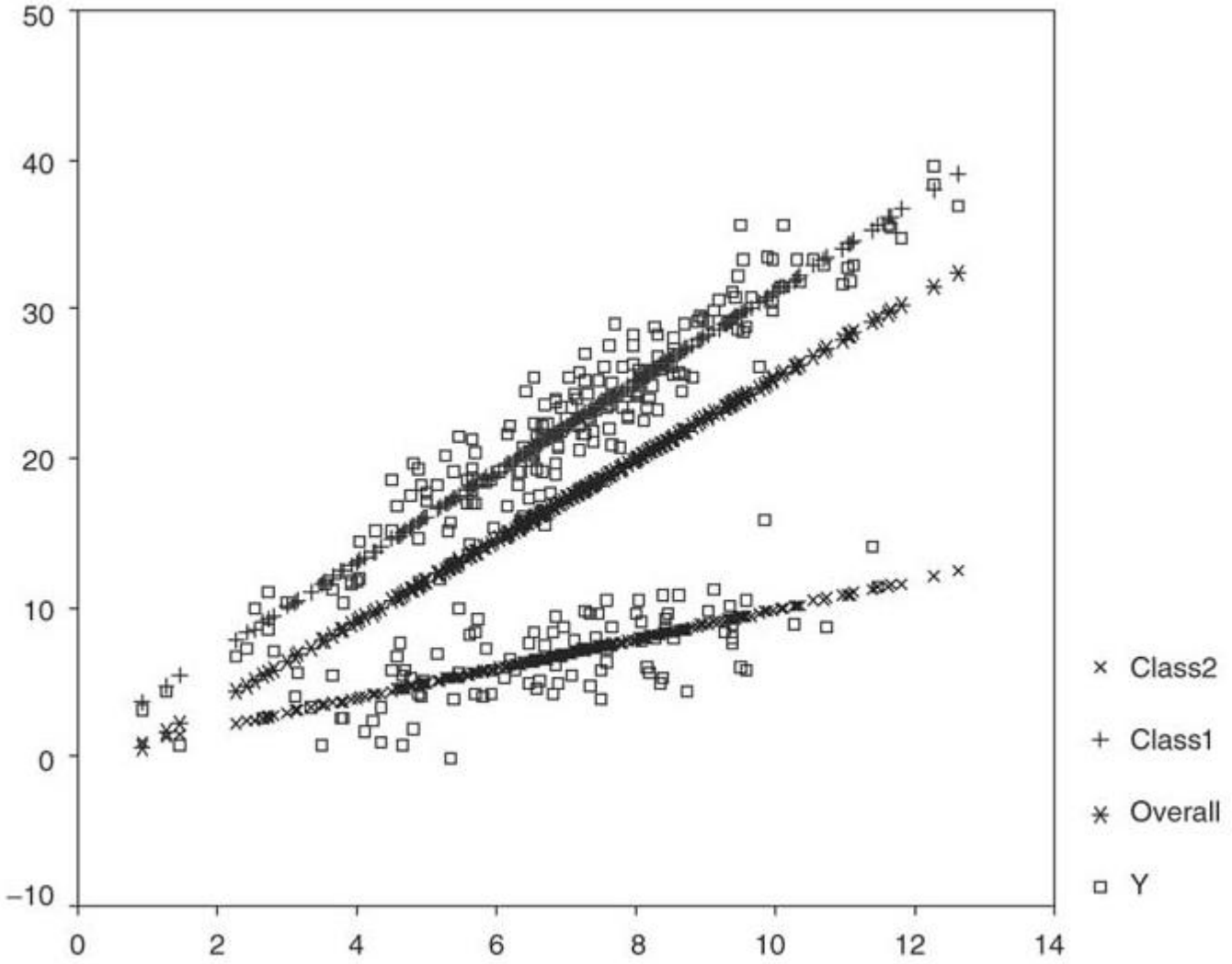
Basically: we add structural piece to LCA model where covariates predict class membership

- Regression -> used to predict a dependent variable as a function of predictor variables
- LC model -> includes a K-category latent variable X to cluster cases
- LC Regression Model -> Each category represents a homogeneous subpopulation (segment) having identical regression coefficients.
- The LCRM considers, jointly, the effect of covariates on the probability of belonging to a certain latent class.
- Thus, covariates can be added into the latent class model to predict the latent class membership probability.
- In the LCA model, it is assumed that every individual has the **same probabilities** of being in a latent class; however, in the LCR model it is assumed that latent class probabilities **differ** by individuals depending on their observed covariates.

Università della Svizzera italiana

Latent Class Regression Model

RISIS



Università della Svizzera italiana

Outline of the presentation

RISIS

From theory to practice with **Stata**:

1. Premise: LCM a basic classification
2. **Brief introduction to Stata.**

Università della Svizzera italiana

2. Brief introduction to Stata

RISIS

a) The Stata interface.

b) The menus and dialog boxes.

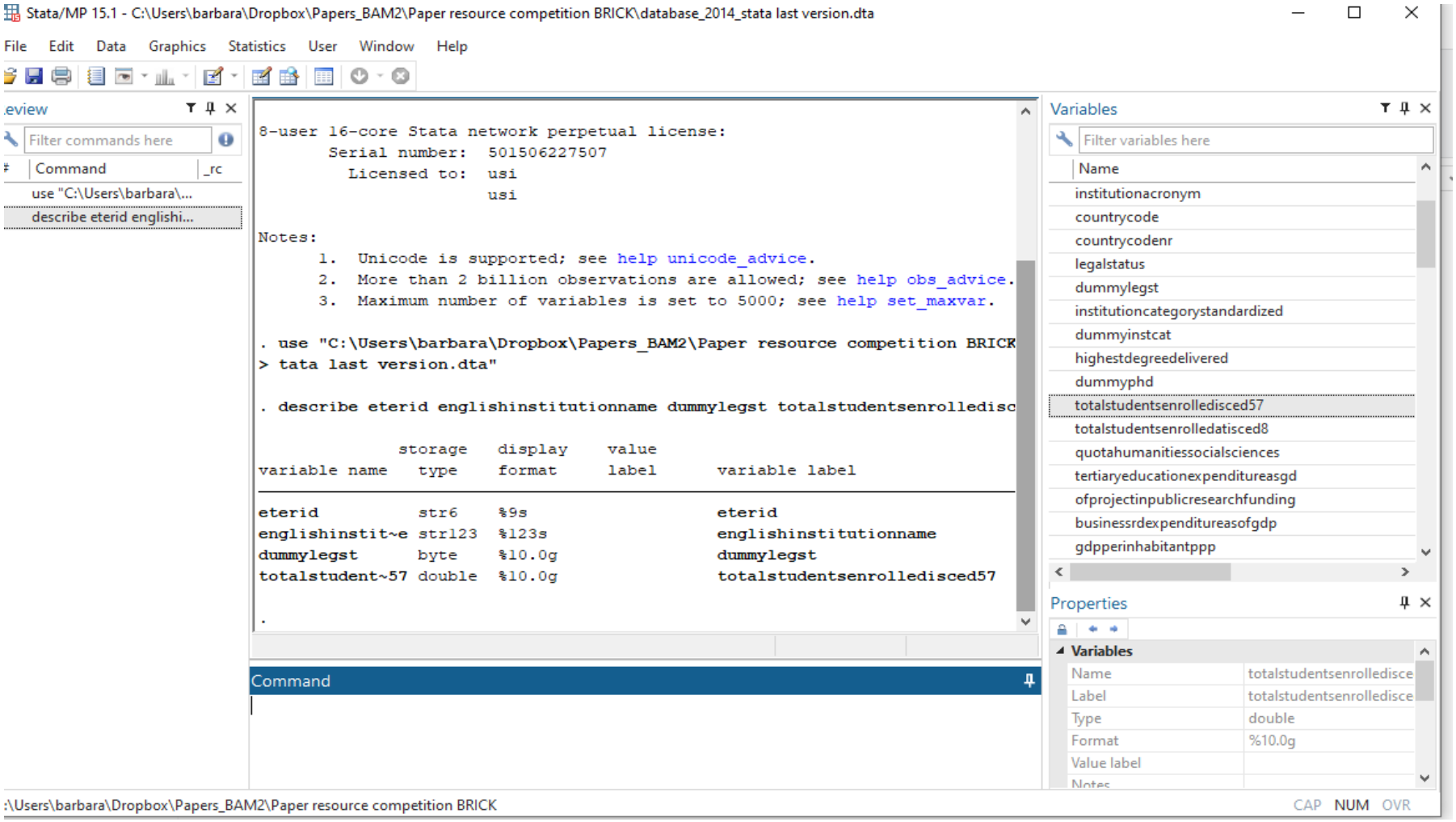
c) Stata command syntax.

d) The do-file editor.

Università della Svizzera italiana

a) The Stata interface

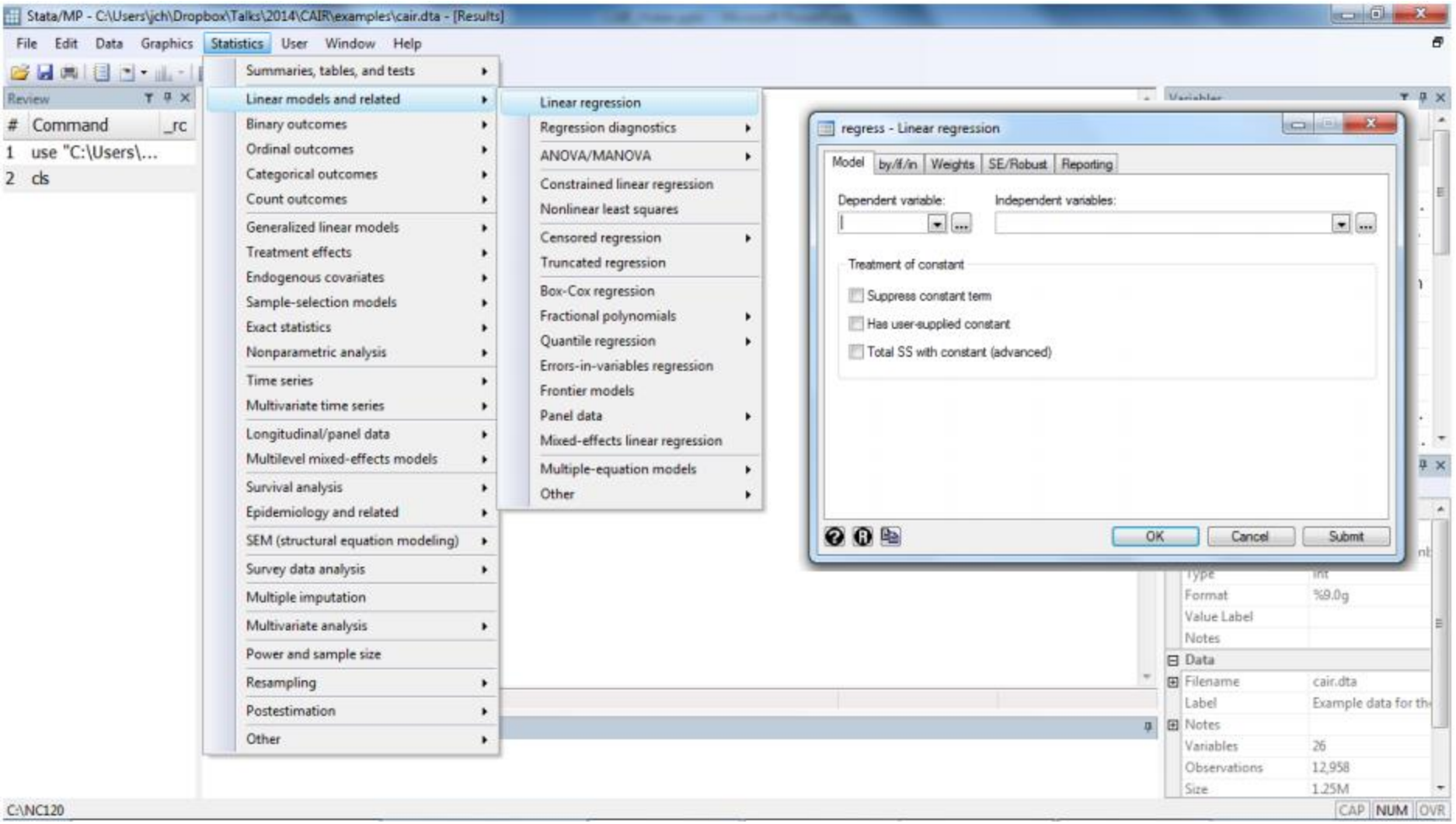
RISIS



Università della Svizzera italiana

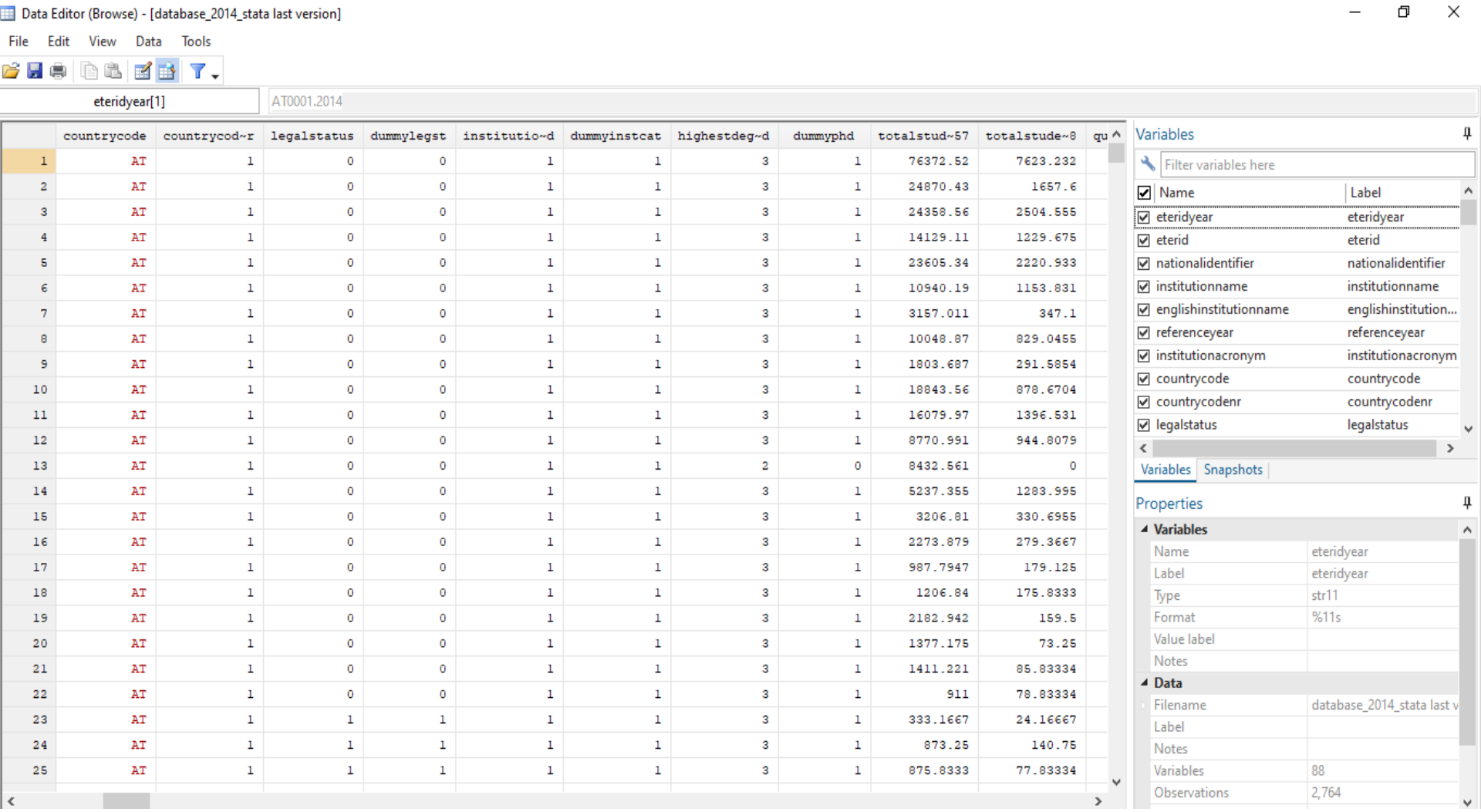
b) The Menus and Dialog Boxes

RISIS



c) The Data Editor

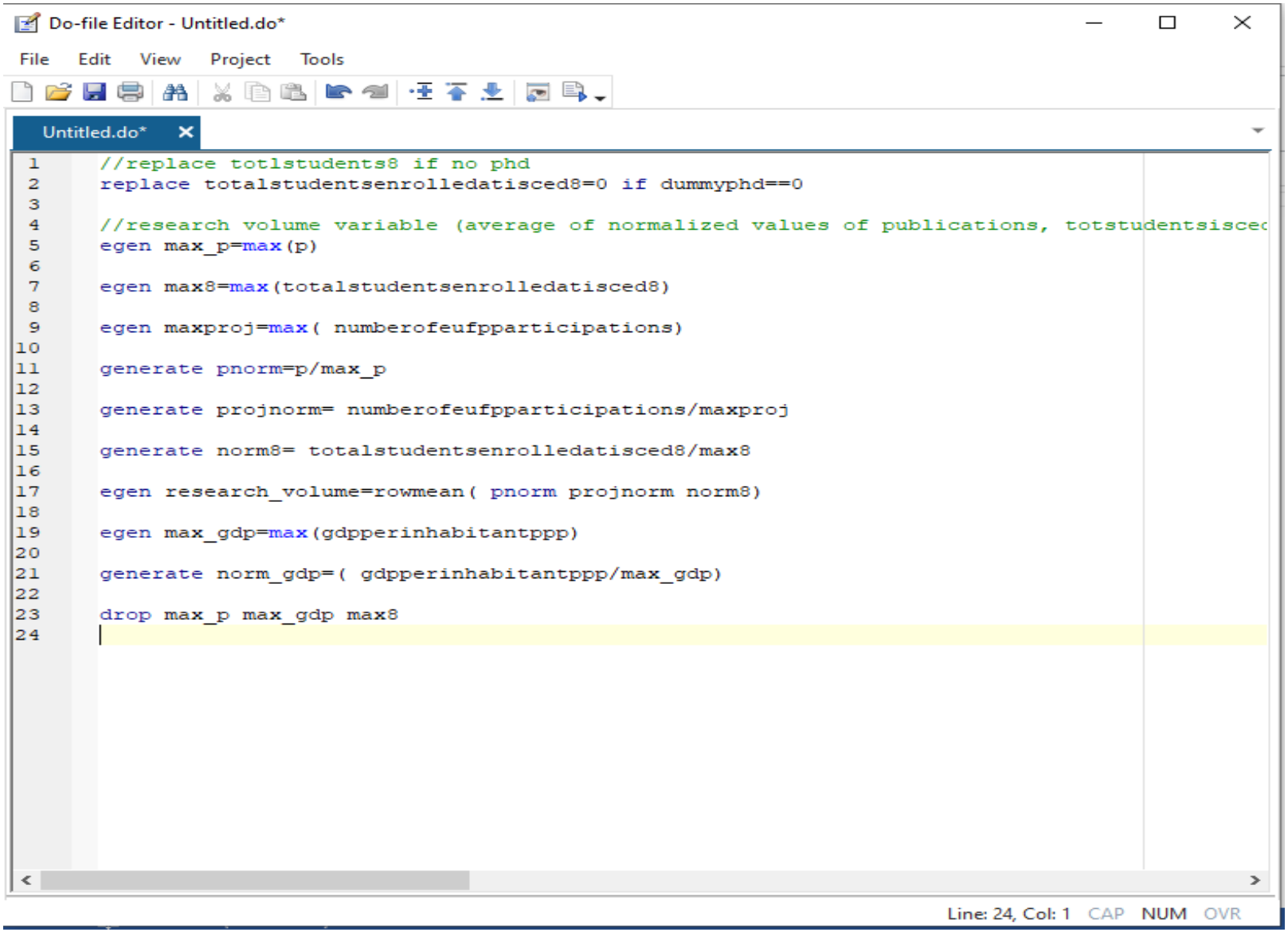
RISIS



	countrycode	countrycod~r	legalstatus	dummylegst	institutio~d	dummyinstcat	highestdeg~d	dummyphd	totalstud~57	totalstude~8	qu
1	AT	1	0	0	1	1	3	1	76372.52	7623.232	
2	AT	1	0	0	1	1	3	1	24870.43	1657.6	
3	AT	1	0	0	1	1	3	1	24358.56	2504.555	
4	AT	1	0	0	1	1	3	1	14129.11	1229.675	
5	AT	1	0	0	1	1	3	1	23605.34	2220.933	
6	AT	1	0	0	1	1	3	1	10940.19	1153.831	
7	AT	1	0	0	1	1	3	1	3157.011	347.1	
8	AT	1	0	0	1	1	3	1	10048.87	829.0455	
9	AT	1	0	0	1	1	3	1	1803.687	291.5854	
10	AT	1	0	0	1	1	3	1	18843.56	878.6704	
11	AT	1	0	0	1	1	3	1	16079.97	1396.531	
12	AT	1	0	0	1	1	3	1	8770.991	944.8079	
13	AT	1	0	0	1	1	2	0	8432.561	0	
14	AT	1	0	0	1	1	3	1	5237.355	1283.995	
15	AT	1	0	0	1	1	3	1	3206.81	330.6955	
16	AT	1	0	0	1	1	3	1	2273.879	279.3667	
17	AT	1	0	0	1	1	3	1	987.7947	179.125	
18	AT	1	0	0	1	1	3	1	1206.84	175.8333	
19	AT	1	0	0	1	1	3	1	2182.942	159.5	
20	AT	1	0	0	1	1	3	1	1377.175	73.25	
21	AT	1	0	0	1	1	3	1	1411.221	85.83334	
22	AT	1	0	0	1	1	3	1	911	78.83334	
23	AT	1	1	1	1	1	3	1	333.1667	24.16667	
24	AT	1	1	1	1	1	3	1	873.25	140.75	
25	AT	1	1	1	1	1	3	1	875.8333	77.83334	

d) The Do-File Editor

RISIS



Università della Svizzera italiana

Outline of the presentation

RISIS

From theory to practice with **Stata**:

1. Premise: LCM a basic classification
2. Brief introduction to Stata.
3. **LPA and LCRM with gsem**

Università della Svizzera italiana

3. LPA and LCRM with gsem

RISIS

a) Some general infos

b) Key concepts and assumptions

c) Differences in capabilities between sem and gsem

a) Some general infos

RISIS

Why using a **gsem** and, first of all, what is a **gsem**?

..Easy, a *Generalized* **sem**!

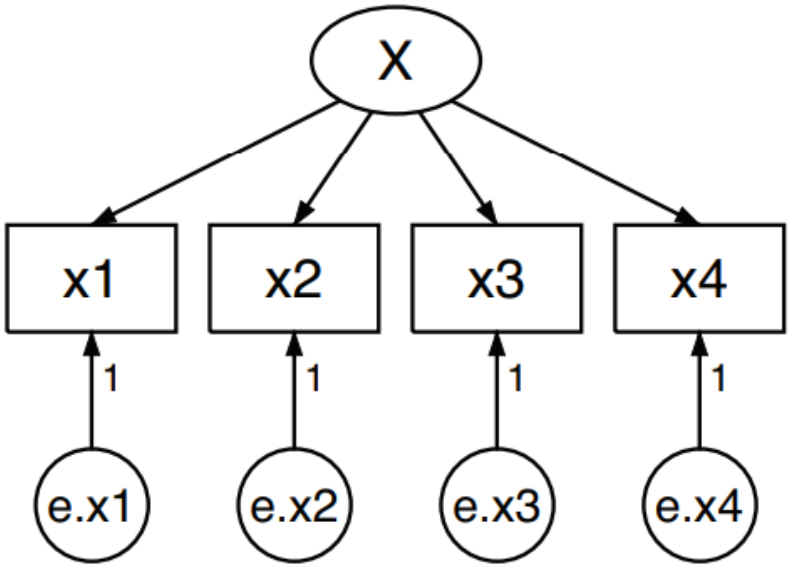
ok, so what is **sem**? And why we use it?

- **sem** = **s**tructural **e**quation **m**odelling = an extension of general linear modelling (GLM).
- **sem** was initially developed in genetics, econometrics and, later, sociology.
- **sem** encompasses a broad array of models from linear regression to measurement models to simultaneous equations.
- **sem** is not just an estimation method for particular model but a way of thinking, writing and estimating.

Università della Svizzera italiana

sem has its roots in path analysis (Wright, 1921).
Just few words to clear ideas..

RISIS



$$x_1 = \alpha_1 + \beta_1 X + e.x_1$$
$$x_2 = \alpha_2 + \beta_2 X + e.x_2$$
$$x_3 = \alpha_3 + \beta_3 X + e.x_3$$
$$x_4 = \alpha_4 + \beta_4 X + e.x_4$$

(Wright notation)

Boxes (x1 x4): observed data (continuous)

Circles (e.x1.. e.x4): unobserved, latent variables

Arrows or «paths: used to define causal relationship, with the variable at the tail of the arrow causing the variable at the point.

Università della Svizzera italiana

b) Key concepts and assumptions (1)

RISIS

- **sem** is a multivariate technique that allows us to estimate a system of equations. Variables in these equations may be measured with error. There may be variables in the model that cannot be measured directly (Latent).
- More precisely, according to (Hoyle, 1995), **sem** is a comprehensive statistical approach to testing hypotheses about relations among observed and latent variables. Multiple, related equations are solved simultaneously to determine parameters.
- What are the assumptions?
  - Large Sample Size
  - Multivariate Normality
  - Correct Model Specification

b) Key concepts and assumptions (2)

RISIS

- Large Sample Size
  - Necessary to obtain reliable parameter estimates.
  - A common rule of thumb is to have a sample size of more than 200 observations, although sometimes 100 is seen as adequate.
  - Several authors propose sample sizes relative to the number of parameters being estimated. Ratios of observations to free parameters from 5:1 up to 20:1 have been proposed.
- Multivariate normality
  - The likelihood that is maximized using ML is derived under the assumption that the observed variables follow a multivariate normal distribution.
- Correct Model Specification
  - No relevant variables are omitted from any equation in the model.
  - Omitted variable bias can arise in linear regression if an independent variable is omitted from the model and the omitted variable is correlated with other independent variables.
  - When fitting structural equation models with ML and all equations are fit jointly, errors can occur in equations other than the one with the omitted variable.

c) Differences in capabilities between sem and gsem (1)

RISIS

- **gsem** provides several abilities not provided by **sem**:
  - **gsem** allows for multilevel models
    - 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
  - **gsem** ML is able to use more observation in presence of missing values.
  - **gsem** fist SEMs containing generalized linear response variables
    - Generalized response variables means that the response variables can be specifications from the generalized linear model (GLM). These include probit, logistic regression, ordered probit and logistic regression, multinomial logistic regression, and more
  - **gsem** fits models with categorical latent variables

Università della Svizzera italiana

c) Differences in capabilities between sem and gsem (2)

- You may obtain different likelihood values when fitting the same model with sem and gsem
  - The likelihood for **sem** is derived including estimation of the means, variances and covariances of the observed exogenous variables.
  - The likelihood for the model fit by **gsem** is derived as conditional on the values of the observed exogenous variables.
  - Normality of observed exogenous variable is never assumed with **gsem**.

Outline of the presentation

RISIS

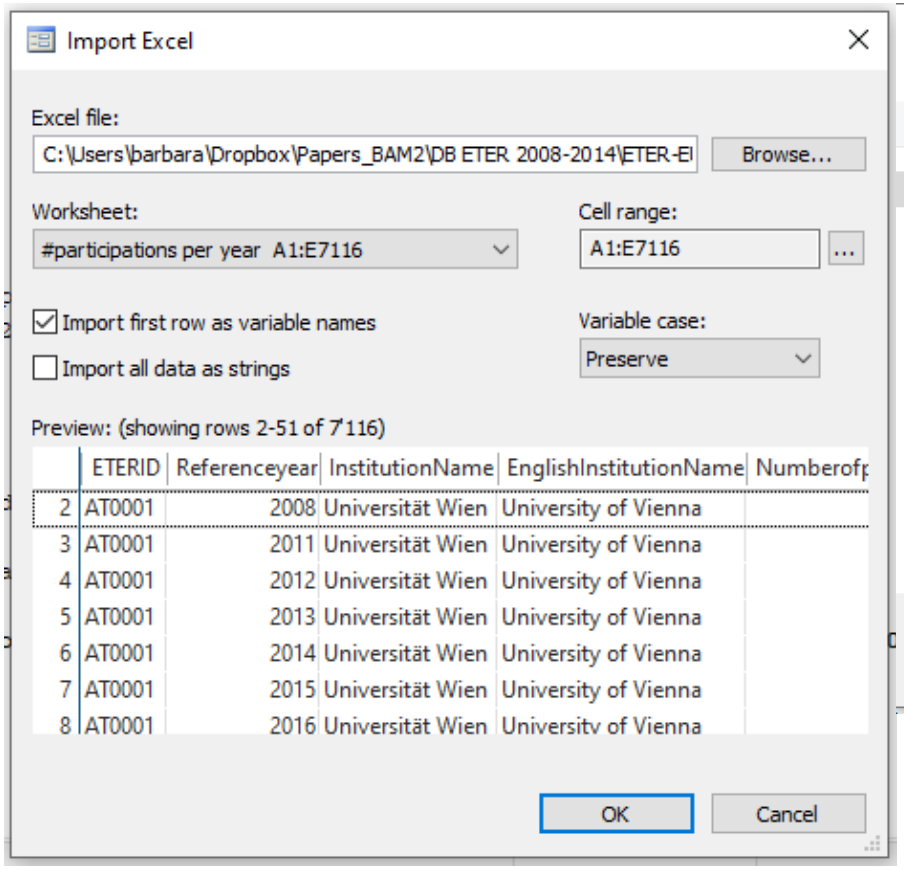
From theory to practice with **Stata**:

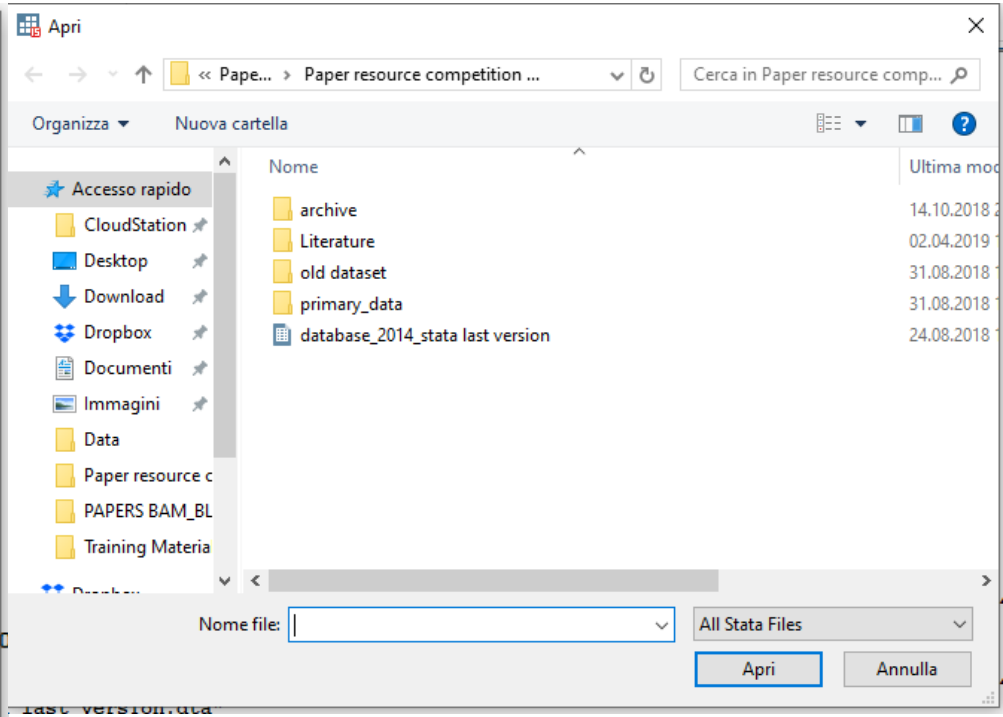
1. Premise: LCM a basic classification
2. Brief introduction to Stata.
3. LPA and LCRM with **gsem**
4. **Run the model!**

Università della Svizzera italiana

4. Run the model!

- Getting your data into Stata
  - Can import data using:
    - Import (.xls file, .csv file)
    - Open (.dta file)



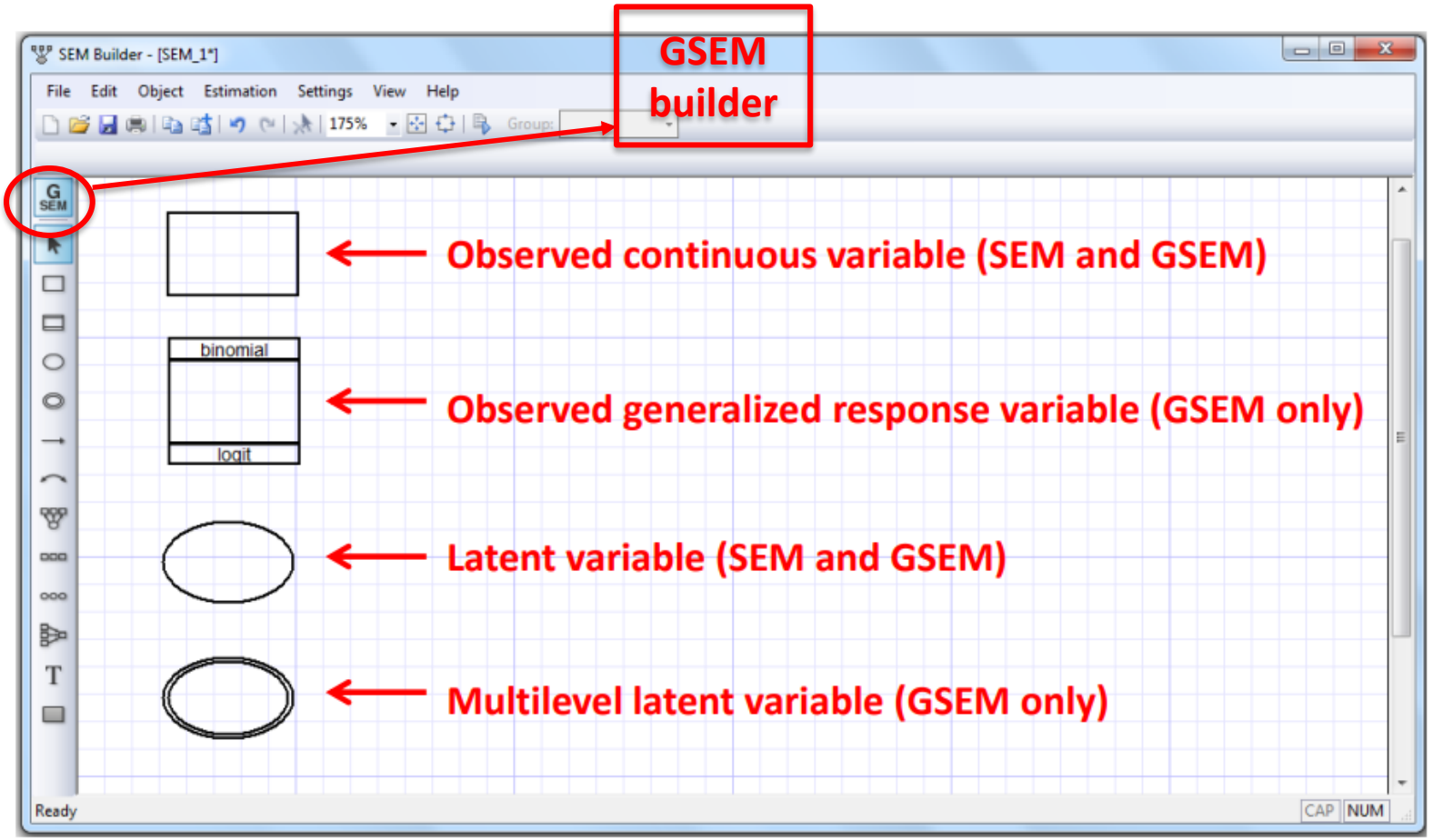


- **sem/gsem** builder and **gsem** syntax

RISIS

Drawing variables in Statas **sem/gsem** builder

In our examples we use **gsem** syntax; in any case, builders shapes are useful to sketch and understand the model to be fitted



Università della Svizzera italiana

But running what? LPA and LCRM

RISIS

- According to our premise, in our presentation and then in the group exercises we will run two different LCM:
  - **A Latent Profile Analysis** (or mixture modeling)
    - Continuous observed variables, discrete latent variables.
    - Every individual has the same probability of being in a latent class.
  - **A Latent Class Regression Model**
    - Considers, jointly, the effect of covariates on the probability of belonging to a certain latent class.
    - Covariates can be added into the latent class model to predict the latent class membership probability.
    - Latent class probabilities differ by individual depending on their observed covariates.

Università della Svizzera italiana

Example of classic LCA

(MacDonald K., StataCorp LLC, 2018)

RISIS

- Authors believe that there are different types of people who attend Stata conferences.

- They hypothesize that there are three groups. Their intuition tells us the groups might be characterized as:

1. **Stata promoters** those who love Stata, encourage others to use Stata, and provide resources for others

2. **Stata researchers** those who use Stata regularly for their own Research

3. **Stata novices** those who have used Stata for a short time and want to learn more

- They have a sample of individuals who have attended conferences around the world (576 people)

- They dont have a variable that records the whether each individual is a Stata promoter, researcher, or novice. Instead, **attendee classification can be considered a latent (unobserved) variable.**

Università della Svizzera italiana

Each conference attendee in the sample answered the following questions:

Questionnaire

1. Do you use Stata at least once per week? *(yes/no)*
2. Have you ever written and distributed a Stata command? *(yes/no)*
3. Have you used Stata for more than 5 years? *(yes/no)*
4. Have you presented at a previous Stata conference? *(yes/no)*
5. Do you teach a course using Stata? *(yes/no)*
6. Have you published a paper based on data analyzed using Stata? *(yes/no)*
7. Have you published an article in the Stata Journal? *(yes/no)*
8. Do you regularly participate in discussions on Statalist? *(yes/no)*
9. Do you live within 50 miles of the conference? *(yes/no)*

Some descriptive statistic

RISIS

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
weekly	576	.5208333	.5	0	1
command	576	.2986111	.4580467	0	1
years5	576	.4826389	.5001328	0	1
presenter	576	.3402778	.4742143	0	1
teacher	576	.4201389	.49401	0	1
published	576	.4930556	.5003863	0	1
sjauthor	576	.3142361	.4646144	0	1
statalist	576	.3628472	.4812392	0	1
location	576	.515625	.5001902	0	1

Università della Svizzera italiana

The syntax to fit the latent class model is

RISIS

(and so on) observed variables

STATA statistics

```
gsem (weekly command years5 presenter teacher published
sjauthor statlist location <- ), logit lclass(C 3)
```

The observed variables are all binary, so we use the logit option to model each one using a constant-only logistic regression.

The lclass(C 3) *option* specifies that we want to allow for differences in these logistic regression models across the levels of a categorical latent variable named C with three classes.

We will not look at the gsem output yet.
It is easier to interpret results using **estat lcprob** and **estat lcmean**.

estat lcprob: reports a table of the marginal predicted latent class probabilities

estat lcmean: reports a table of the marginal predicted means of the outcome within each latent class.

Based on this model, what are the expected proportions of the population in each group?

```
. estat lcprob
Latent class marginal probabilities             Number of obs     =        576
```

	Margin	Delta-method Std. Err.	[95% Conf.	Interval]
C				
1	.1057509	.0582876	.0341272	.2835627
2	.4187809	.0704887	.2900013	.5596688
3	.4754682	.0397848	.3987046	.5534088

- W[illegible] % is in class 3.
- But what do those classes represent?

We now have to try to find out the label of each class..
(importance of our hypothesis)

RISIS

- For individuals in Class 1, what is the probability of responding positively to each question?

```
. estat lcmean
Latent class marginal means                     Number of obs     =        576
```

	Margin	Delta-method Std. Err.	[95% Conf.	Interval]
1				
weekly	.5594732	.1144653	.338218	.759382
command	.703362	.1655266	.3336843	.9182112
years5	.9462668	.1009533	.2644505	.9988421
presenter	.5892076	.1128971	.3650511	.7815784
teacher	.596822	.0986313	.3986389	.7677449
published	.8785688	.0824458	.6140342	.9705049
sjauthor	.7467327	.1777284	.3185127	.9489785
statalist	.4410877	.1074878	.2513733	.6497189
location	.1202751	.0922665	.0241521	.4302775

What do these values tell us?

- The marginal probabilities of answering yes are high for all questions except the one about living nearby.
- This might be our hypothesized Stata Promoters group.

Università della Svizzera italiana

And what about individuals in Class 2?

RISIS

2				
weekly	.7953942	.0490352	.6829157	.8752613
command	.2682777	.0520701	.1789817	.3814271
years5	.7053751	.0461704	.6076852	.7872555
presenter	.5136087	.049906	.4165146	.6096865
teacher	.5796951	.0461948	.4874827	.6666613
published	.6302565	.0507412	.5266124	.7231388
sjauthor	.3026139	.051335	.2122123	.4114143
statalist	.5908731	.0555132	.479385	.6937391
location	.4509978	.0559189	.3454076	.5611936

- What do these values tell us?
- The marginal probabilities of using Stata weekly, having used Stata for more than five years, and publishing articles based on data analyzed in Stata are fairly large.
- These individuals are less likely to have written a Stata command or to have published in the Stata Journal.
- This class might be our hypothesized Stata Researchers.

Università della Svizzera italiana

And finally - what do we expect in Class 3?

RISIS

3				
weekly	.270413	.0382115	.2022746	.3513939
command	.2353055	.0288825	.1834426	.2965067
years5	.1833394	.0370618	.1214216	.2672279
presenter	.1322467	.0255786	.089635	.1908686
teacher	.2403093	.0312686	.1844201	.3067651
published	.2864695	.0349021	.2231754	.3594091
sjauthor	.2282789	.029189	.1761288	.290427
statalist	.1446059	.0295687	.0956889	.2126493
location	.6604777	.0334121	.592279	.7226114

- What do these values tell us?
- These individuals are likely to live close to the conference, but they have lower probabilities of answering yes to all other questions.
- This class might be our hypothesized Stata Novice group.

Università della Svizzera italiana

- Lets take a look at these predictions for some individuals in our sample

RISIS

```
. list in 1/2, abbrev(10)
```

1.	weekly	command	years5	presenter	teacher
	0	0	1	0	0
	published	sjauthor	statalist	location	sjeditor
	1	0	1	1	0
	cpost1	cpost2	cpost3	predclass	
	.0145142	.6011773	.3843085	2	

2.	weekly	command	years5	presenter	teacher
	1	1	1	1	1
	published	sjauthor	statalist	location	sjeditor
	1	1	1	0	1
	cpost1	cpost2	cpost3	predclass	
	.7521391	.2477402	.0001208	1	

- Attendee 1 has used STATA for more than 5 years, published a paper based on data analyzed using STATA, regularly participates in discussion on Statalist and live within 50 miles of the conference: -> cpost 0.6011773 - > about 60% probability of belonging to CLASS 2 (STATA researcher)
- Attendee 2 answered yes to all questions except the one on location: -> cpost 0,7521391 -> about 75% of probability of belonging to CLASS 1 (STATA promoter)
- And so on..

Università della Svizzera italiana

May I try with a different number of classes?
Is there a test that can help me in selecting the right C number?

RISIS

Of course!

We can compare the models fit using Akaikes information criterion (AIC) and Schwarzs Bayesian information criterion (BIC).

```
. estimates stats c2inv c3inv c4inv c5inv

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
c2inv	145	.	-1702.554	10	3425.108	3454.876
c3inv	145	.	-1653.238	14	3334.476	3376.15
c4inv	145	.	-1626.828	18	3289.656	3343.237
c5inv	145	.	-1578.207	22	3200.414	3265.902

```
Note: BIC uses N = number of observations. See [R] BIC note.
```

The model with five latent classes has the smallest values of both AIC and BIC and would be considered the best based on these information criteria.

Outline of the presentation

RISIS

Premise: to run our LCM we use the **Stata 15** (gsem) package (updated version 15.1)

From theory to practice with **Stata**:

1. Premise: LCM a basic classification
2. Brief introduction to **Stata**.
3. LPA and LCRM with **gsem**
4. Run the model!
5. **How to interpret results?**

- Now we are ready to have a look to the **gsem** output estimation results
- The command syntax for the **gsem** estimation was the following

gsem (weekly command years5 presenter teacher published sjauthor statalist location<-), logit lclass(C 3)

(multinomial logistic regression for latent categorical variable C and with 3 classes)

The first Class will be treated as the baseline.

```
Generalized structural equation model           Number of obs     =        576
Log likelihood = -3283.0567
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.C	(base outcome)					
2.C						
_cons	1.376261	.696632	1.98	0.048	.0108875	2.741635
3.C						
_cons	1.503213	.5577001	2.70	0.007	.4101412	2.596285

- Lets have a look also to results for each class. Tables report class-specific, constant-only logistic regression results for each of our observed variables.
- Class 1

Class : 1

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weekly						
_cons	.2390244	.464432	0.51	0.607	-.6712456	1.149294
command						
_cons	.8633593	.7933449	1.09	0.276	-.6915682	2.418287
years5						
_cons	2.868493	1.985474	1.44	0.149	-1.022964	6.75995
presenter						
_cons	.3606906	.4664361	0.77	0.439	-.5535073	1.274889
teacher						
_cons	.3922409	.4098956	0.96	0.339	-.4111397	1.195621
published						
_cons	1.978947	.7727922	2.56	0.010	.4643019	3.493592

- Class 2

Class : 2

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weekly						
_cons	1.357752	.3013059	4.51	0.000	.7672035	1.948301
command						
_cons	-1.003379	.2652515	-3.78	0.000	-1.523262	-.4834952
years5						
_cons	.8730265	.2221644	3.93	0.000	.4375923	1.308461
presenter						
_cons	.0544483	.1997721	0.27	0.785	-.3370978	.4459945
teacher						
_cons	.3215218	.1895961	1.70	0.090	-.0500796	.6931232
published						
_cons	.5333175	.2177424	2.45	0.014	.1065502	.9600848

Possible extensions

RISIS

- We can include continuous, binary, ordinal, categorical, count, fractional, and even survival-time observed variables.

- We can include **predictors** of the latent classes.

```
gsem (y1 y2 y3 y4 <- , logit) ///
(C <- x1), lclass(C 3)
```

Now x1 is included as a regressor in the multinomial logit model for C.

- We can allow **regression** models to vary across classes.

Università della Svizzera italiana

Time for another example?

RISIS

Classic LPA (from STATA manuals)

```
Contains data from https://www.stata-press.com/data/r16/gsem_lca2.dta
  obs:           145                          Latent profile analysis
 vars:             7                          18 Jan 2019 12:39
 size:         3,045                          (_dta has notes)
```

variable name	storage type	display format	value label	variable label
patient	int	%9.0g		Patient ID
relwgt	float	%9.0g		Relative weight
fglucose	int	%9.0g		Fasting plasma glucose
glucose	float	%9.0g		Glucose area (mg/10mL/hr)
insulin	float	%9.0g		Insulin area (mIU/10mL/hr)
sspg	float	%9.0g		Steady-state plasma glucose
cclass	byte	%17.0g	class	Clinical classification

- **One categorical latent variable** and **three observed continues variables** (glucose, insulin and sspg).
- The goal is to determine **categories of diabetes based on these three variables**.
- **Open stata and run with me the example!**

Università della Svizzera italiana

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Università della Svizzera italiana

.. Now let s have fun with our group exercise nr. 1: an example of LPA applied to the Higher Education sector!

«The heterogeneity of European Higher Education Institutions. A typological approach

(Lepori B., 2019)

But before.

Lunch in our wonderful mensa!