polychoric, by any other 'namelist'

Stas Kolenikov @StatStas

Abt SRBI @AbtSRBI

Stata Conference 2016

Motivation: methods

In many social, behavioral or health studies, there may be interest in summarizing multivariate ordinal data.

- Multivariate exploratory analysis:
  - Find structure in the data
  - Describe main features (e.g., principal components)
- Multivariate confirmatory analysis:
  - Regression-type models
  - Structural equation / latent variable models
- Data processing: construct a variable summarizing socio-economic status
  - No income or consumption variables available
  - Can only use HH assets

Motivation: data

Running example: Demographic and Health Surveys (DHS), Bangladesh 2014

Whether the household has:

- HV206 Electricity
- HV207 A radio
- HV208 A television
- HV209 A refrigerator

What the dwelling is made of:

- HV213 Main material of the floor (dirt, wood, cement, . . .)
- HV214 Main material of the walls (dirt, wood, tin, brick, . . .)
- HV215 Main material of the roof (straw, wood, tin, cement, . . .)

SES: solutions offered

- Historic procedure: break the categories into dummy variables, run PCA, score 1st component
- Polychoric procedure: maintain the ordinal nature, estimate polychoric correlation matrix (Olsson 1979), run PCA, score 1st component (Kolenikov & Angeles 2009)
- Utilize structural equation modeling treating SES as a latent variable (Bollen et al. 2007)

Goal of this talk

Compare and contrast the existing Stata tools, including the third party ones:

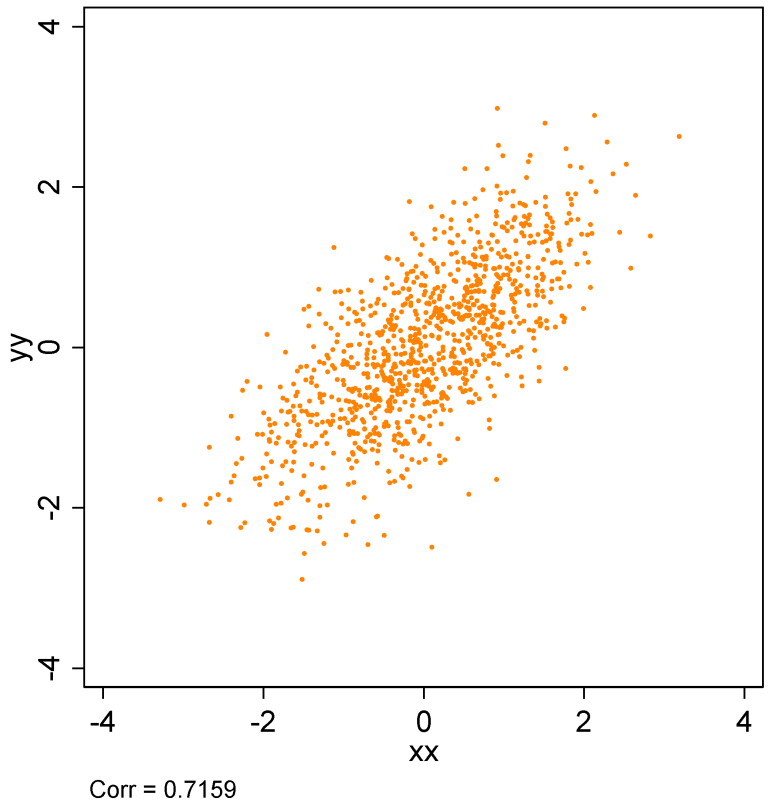
- `polychoric` (by yours truly)
- `cmp` (Roodman 2011)
- `gsem` (official Stata)

POLY...WHAT??

Polychoric correlation concept

Let us start with just two bivariate normal variables

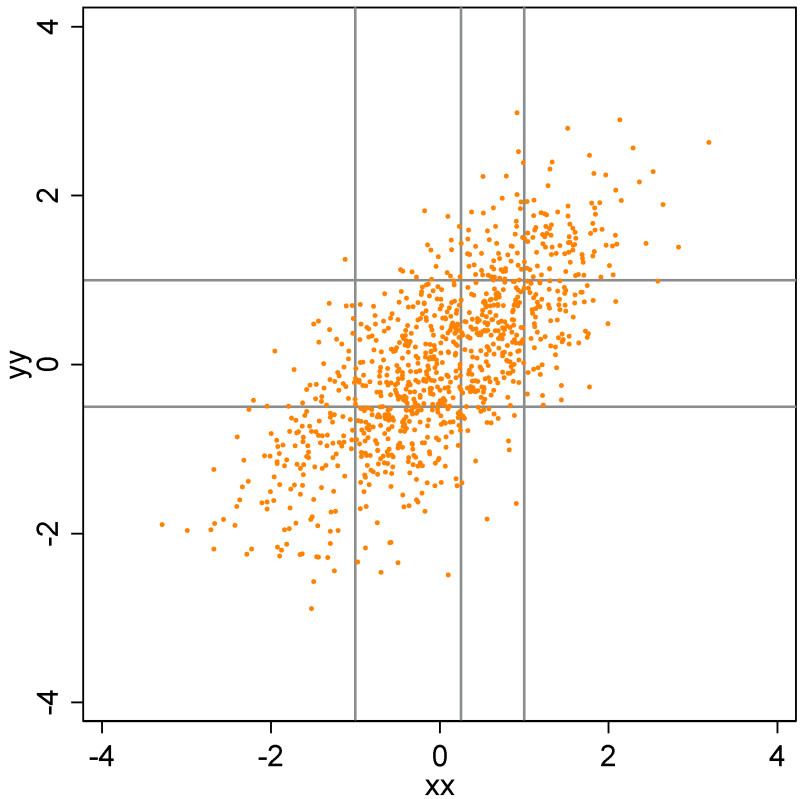
```
gen xx = rnormal()
gen yy = 1/sqrt(2)*xx + 1/sqrt(2)*rnormal()
```



Polychoric correlation concept

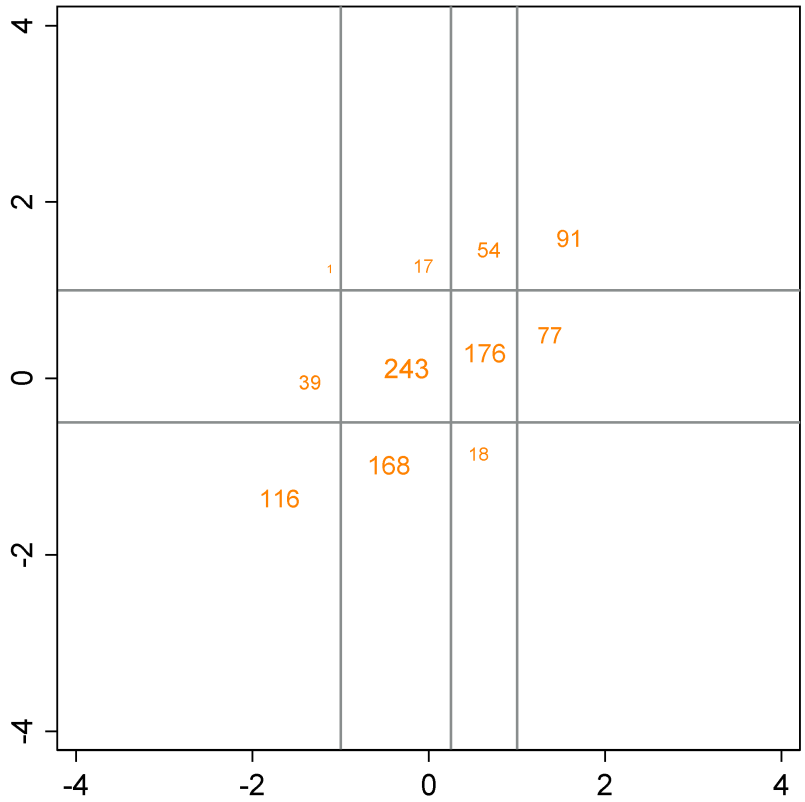
Now, let's bin both variables into a small number of ordinal categories

```
recode xx (-100/-1=1) (-1/0.25=2) (0.25/1=3) (1/100=4),
gen(x)
recode yy (-100/-0.5=1) (-0.5/1=2) (1/100=3), gen(y)
```



Polychoric correlation concept

Here's our contingency table on the original scale:



Polychoric correlation concept

Can we recover the original correlation from these ordinal variables now?

```
. tab y x
 RECODE of |                RECODE of xx
        yy |         1         2         3         4 |     Total
-----------+--------------------------------------------+----------
         1 |       116       168        18         0 |       302
         2 |        39       243       176        77 |       535
         3 |         1        17        54        91 |       163
-----------+--------------------------------------------+----------
     Total |       156       428       248       168 |     1,000
```

Polychoric correlation concept

Polychoric correlation:

1. Assume an underlying normal variate for each of the ordinal variables
2. Write up the likelihood for the cutoff and the correlation parameters
3. Estimate by maximum likelihood
4. (optional) Produce a likelihood ratio or a Pearson goodness of fit test for the table

Polychoric correlation concept

```
.  polychoric x y

Variables :  x y
Type :       polychoric
Rho          = .73385592
S.e.         = .01898606
Goodness of fit tests:
Pearson G2 = 10.842193, Prob( >chi2(5)) = .05460018
LR X2      = 6.8388022, Prob( >chi2(5)) = .23290749
```

`polychoric`: small print

The `polychoric` command is actually a partial/two-step information maximum likelihood estimator.

1. Estimate the thresholds from marginal distributions of each categorical variable only;
2. Estimate the correlation based on bivariate likelihood treating the thresholds as known.

Polychoric: a FIML implementation

Roodman (2011) `cmp`: every variable is a truncated/censored/categorized/missing normal

```
.  cmp setup
.  cmp (x=) (y=), ind($cmp_oprobit $cmp_oprobit)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/cut_1_1	-1.011033	.0475311	-21.27	0.000	-1.104192	-.9178735
/cut_1_2	.209776	.0397371	5.28	0.000	.1318928	.2876593
/cut_1_3	.9700072	.047006	20.64	0.000	.877877	1.062137
/cut_2_1	-.52586	.0415173	-12.67	0.000	-.6072325	-.4444876
/cut_2_2	.9859156	.0473324	20.83	0.000	.8931458	1.078685
rho_12	.7324529	.020709			.6892003	.770504

The tale of three correlations

```
bootstrap r(rho), reps(1000) :  corr yy xx
bootstrap r(rho), reps(1000) :  corr y x
```

Correlation	Estimate	Std. error
Pearson, original	0.7159	0.0146
Pearson, categorical	0.6222	0.0187
Polychoric, partial	0.7339	0.0190
Polychoric, FIML	0.7325	0.0207
Population	0.7071	

SOCIO-ECONOMIC STATUS

Principal component analysis

Given $\mathrm{Cov}[X] = \Sigma$, solve eigenproblem $\Sigma a = \lambda a$
Equivalent: find $a : \|a\| = 1$ s.t. $\lambda_1 \equiv \mathrm{Var}[a'X] \to \max$

- The method is useful as a quick multivariate exploratory summary of the data, or a data dimension reduction technique
- The first component is usually the measure of "size". In applications to socio-economic status, it is a measure of overall wealth.
- Subsequent components usually describe finer structure. In SES applications, these are often urban-rural distinction, sector of employment, etc.

SES as a latent variable

Bollen et al. (2007): socio-economic status is a latent variable, and it can be described in terms of:

- *internal validity*: the degree of measurement error in the ordinal measurements of household assets and dwelling quality
- *external validity*: if a substantive theory predicts a certain relation to behavioral/health outcomes, can test the strength of the relation
  - Fertility: more affluent women are expected to have lower fertility rates

SES as a latent variable

Pros:

- Deals properly with measurement error in SES measurement
- Simultaneous estimation $\Rightarrow$ correct standard errors

Cons:

- SES scores are specific to the model, and in particular to the dependent variable in the analysis

EMPIRICAL EXAMPLE

Back to Bangladesh data

Running example: Demographic and Health Surveys (DHS), Bangladesh 2014

Whether the household has:

- HV206 Electricity
- HV207 A radio
- HV208 A television
- HV209 A refrigerator

What the dwelling is made of:

- HV213 Main material of the floor (dirt, wood, cement, . . .)
- HV214 Main material of the walls (dirt, wood, tin, brick, . . .)
- HV215 Main material of the roof (straw, wood, tin, cement, . . .)

Polychoric analysis of Bangladesh data

```
view stataconf2016-kolenikov-02-bangla-dhs-polychor.smcl

view stataconf2016-kolenikov-03-bangla-dhs-cmp.smcl
```

Bangladesh DHS: women's data

- Age
- Education
- Religion
- Dates of births given

Dependent variable (per Bollen et al. (2007)): given birth in the past 3 years

Bangladesh DHS: women's data

```
. svy: logit _birth_in_3years i.v106 i.v013 _non_islam i.v101
```

_birth_in_3years	Coef.	Svy Std.Err.	t	P>|t|	[95% Conf.	Interval]
educ primary	.012397	.1214593	0.10	0.919	-.2261407	.2509348
secondary	.0401373	.1221026	0.33	0.742	-.1996639	.2799384
higher	.0468193	.1207769	0.39	0.698	-.1903782	.2840169
age 20-24	.0423951	.0814294	0.52	0.603	-.1175268	.2023169
25-29	-.4672969	.0705984	-6.62	0.000	-.6059474	-.3286465
30-34	-1.198171	.0809	-14.81	0.000	-1.357053	-1.039289
35-39	-2.181392	.1097832	-19.87	0.000	-2.396999	-1.965785
40-44	-3.637009	.1846494	-19.70	0.000	-3.999648	-3.274371
45-49	-4.523696	.3310319	-13.67	0.000	-5.17382	-3.873572
_non_islam	-.1690494	.0721039	-2.34	0.019	-.3106565	-.0274423
region dhaka	.0555269	.090483	0.61	0.540	-.1221755	.2332293
chittagong	.2545225	.0818472	3.11	0.002	.0937801	.415265
khulna	-.1834627	.0868905	-2.11	0.035	-.3541099	-.0128156
rajshahi	-.1270263	.0845166	-1.50	0.133	-.2930111	.0389585
rangpur	-.2078279	.0863269	-2.41	0.016	-.377368	-.0382877
sylhet	.5504491	.1489664	3.70	0.000	.2578891	.8430091

Bangladesh DHS: blunt regression

Approach 1:

1. (optional) recode the HH assets
2. Run PCA and get the principal component scores
3. Plug the principal component scores as regressors

Regression with PC scores

```
. svy: logit _birth_in_3years i.v106 i.v013 _non_islam i.v101 _pcw1 _pcw2 _pcw3
```

_birth_in_3years	Coef.	Svy Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ primary	.0782491	.0989555	0.79	0.429	-.1160927	.272591
secondary	.2608675	.0993257	2.63	0.009	.0657985	.4559365
higher	.4660717	.1182902	3.94	0.000	.2337579	.6983856
age 20-24	.0680425	.0714992	0.95	0.342	-.0723771	.2084621
25-29	-.4598912	.0736109	-6.25	0.000	-.604458	-.3153245
30-34	-1.186385	.0844865	-14.04	0.000	-1.352311	-1.020459
35-39	-2.047623	.1181851	-17.33	0.000	-2.279731	-1.815516
40-44	-3.486118	.1953435	-17.85	0.000	-3.869759	-3.102476
45-49	-4.518884	.3831279	-11.79	0.000	-5.271322	-3.766447
_non_islam	-.2132556	.0754696	-2.83	0.005	-.3614727	-.0650385
region (output omitted)						
_pcw1	-.07803	.0219485	-3.56	0.000	-.1211354	-.0349245
_pcw2	-.2834368	.0299997	-9.45	0.000	-.3423543	-.2245194
_pcw3	-.0499353	.0339053	-1.47	0.141	-.116523	.0166523

Latent variable analysis

```
sum _pcw1
svy, noisily:  gsem ///
    (SES -> _floor _water _toilet _wall _roof, ologit) ///
    (SES -> _ln_rooms_per_person) ///
    (SES -> _electricity _fridge _bank, logit) ///
    (_birth_in_3years <- ///
      i.v106 i.v013 _non_islam i.v101 SES, logit) ///
    , iter(20) difficult var(SES@‘=r(sd)^2’)
```

SEM: regression of interest

```
. svy, noisily: gsem ...
Survey: Generalized structural equation model
```

		Coef.	Svy Std. Err.	t	P>|t|	[95% Conf.	Interval]
_birth_in_3years <-							
educ	primary	.054513	.1228607	0.44	0.657	-.186777	.2958031
	secondary	.1475874	.1227421	1.20	0.230	-.0934697	.3886445
	higher	.2556996	.1294689	1.97	0.049	.0014314	.5099678
age	20-24	.057382	.0823479	0.70	0.486	-.1043436	.2191077
	25-29	-.4432318	.0721716	-6.14	0.000	-.584972	-.3014916
	30-34	-1.157909	.0829267	-13.96	0.000	-1.320771	-.9950465
	35-39	-2.128615	.1113795	-19.11	0.000	-2.347356	-1.909873
	40-44	-3.575322	.1864055	-19.18	0.000	-3.94141	-3.209234
	45-49	-4.45321	.3322289	-13.40	0.000	-5.105685	-3.800735
	_non_islam	-.2104908	.073272	-2.87	0.004	-.3543921	-.0665896
region	chittagong	.3034457	.0781044	3.89	0.000	.1500539	.4568374
	dhaka	.1344882	.0860672	1.56	0.119	-.0345419	.3035184
	khulna	-.1628592	.0844922	-1.93	0.054	-.3287961	.0030778
	rajshahi	-.1227685	.0829127	-1.48	0.139	-.2856035	.0400665
	rangpur	-.2408266	.0859807	-2.80	0.005	-.4096868	-.0719664
	sylhet	.5826718	.1460822	3.99	0.000	.2957762	.8695675
	SES	-.1020749	.0224082	-4.56	0.000	-.146083	-.0580668

SEM: measurement loadings

```
. svy, noisily: gsem ...
Survey: Generalized structural equation model
```

		Coef.	Svy Std. Err.	t	P>|t|	[95% Conf.	Interval]
_floor <-	SES	4.752314	.4200563	11.31	0.000	3.927352	5.577276
_water <-	SES	1.282591	.0990658	12.95	0.000	1.088032	1.47715
_toilet <-	SES	1.626373	.0709076	22.94	0.000	1.487115	1.76563
_wall <-	SES	1.387252	.065728	21.11	0.000	1.258167	1.516337
_roof <-	SES	1.366078	.0990119	13.80	0.000	1.171626	1.560531
_ln_rooms_per_person <-	SES	.0719725	.0082108	8.77	0.000	.0558469	.088098
_electricity <-	SES	1.398475	.0806419	17.34	0.000	1.2401	1.556851
_fridge <-	SES	1.60602	.0792796	20.26	0.000	1.45032	1.76172
_bank <-	SES	.8368984	.038785	21.58	0.000	.7607273	.9130695

WRAPPING UP

Conclusions?

- SES does reduce fertility rates
- It was an omitted variable in the initial analysis biasing the coefficients
- Coefficients in regression with PCA scores suffer from measurement error attenuation bias
- Analysis with `gsem` is about as fast as that with `polychoricpca` (in terms of computing time, not necessarily that of the person in front of the computer)
- However, analysis with several PCA scores produces a different story with urban/rural divide. . . which probably should have been modeled explicitly

Workflow

Robert Picard's `project`

Profiling:

- `polychoric` analysis:
  - 2 min 07 sec for full matrix
  - some analysis of misfitting variables
  - 62 sec final analysis and scoring
- `cmp` analysis: 1 hr 22 min
  - Utilized telescoping sample: start with a small sample, increase the sample size gradually to the full sample, pass previous parameter estimates along
- `gsem` analysis: 1 min 06 sec
  - . . . although Stas screwed up the model initially by choosing a poor scaling variable, so nothing converged

Repository

`http://bitbucket.org/stas_kolenikov/stataconf2016-polychoric`

Private; email me at skolenik@gmail.com with access requests

References

Bollen, K. A., Glanville, J. L. & Stecklov, G. (2007), 'Socio-economic status, permanent income, and fertility: A latent-variable approach', *Population Studies* **61**(1), 15–34.

Kolenikov, S. & Angeles, G. (2009), 'Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer?', *The Review of Income and Wealth* **55**(1), 128–165.

Olsson, U. (1979), 'Maximum likelihood estimation of the polychoric correlation', *Psychometrika* **44**, 443–460.

Roodman, D. (2011), 'Fitting fully observed recursive mixed-process models with cmp', *The Stata Journal* **11**(2), 159–206.