



# Pregled ostalih multivariatnih metod

## Multivariatna analiza

# Pregled ostalih multivariatnih metod

- Večrazmernostno lestvičenje
- Korespondenčna analiza
- Splošni linearni model
- Logistična regresija
- Posplošeni linearni model
- Analiza preživetja
- Večnivojski modeli

# Večrazmerno lestvičenje (VRL)

- VRL (Multidimensional scalling - MDS) poskuša v  $k$ -dimenzionalnem prostoru (pogosto 2) predstaviti enote tako, da se čim bolj ohranijo različnosti (podobnosti) med enotami.
- Uporablja se predvsem takrat, ko so že originalni podatki podobnosti (ali različnosti)
- Za razliko od metode glavnih komponent torej poskuša z  $k$  dimenzijami čim bolje pojasniti različnosti (podobnosti) med enotami in ne vrednosti enot pri posameznih spremenljivkah.

# Klasično metrično VRL

- Klasično metrično večrazmernostno lestvičenje minimizira tale „Stress“

$$STRESS_{klasično} = \frac{\sum_{i>j} (d_{ij} - \tilde{d}_{ij})^2}{\sum_{i>j} d_{ij}^2}$$

, kjer so  $d_{ij}$  originalne različnosti (podobnosti) med enotami,  $\tilde{d}_{ij}$  pa evklidska razdalja v ocenjenem  $k$ -dimenzionalnem prostoru

- Pri tej različici je mogoče rešitev dobiti tudi preko razcepa ustrezno popravljene matrike različnosti na lastne vektorje in lastne vrednosti.
- R: Funkcija `cmdscale {stats}`

- V primeru, da uporabimo evklidsko razdaljo, je prostor klasične metričnega VRL (razdalje med točkami) enak prostoru, ki ga vrne metoda glavnih komponent, če je število dimenzij pri MGK enako kot pri VRL.
- MGK moramo uporabiti na kovariančni matriki.
- Če MGK uporabimo na korelacijski matriki, moramo za podoben rezultat pri VRL evklidsko razdaljo izračunati na standardiziranih spremenljivkah.

# Splošno (ne-metrično) VRL

- Bolj splošno VRL (Non-metric MDS), ki se pogosteje uporablja, pa minimizira *STRESS1*:

$$STRESS1 = \frac{\sum_{i>j} (f(d_{ij}) - \tilde{d}_{ij})^2}{\sum_{i>j} \tilde{d}_{ij}^2}$$

Če je  $f(d_{ij}) = a + b \cdot d_{ij}$ , je to še vedno metrično VRL.

R: Funkcija **isoMDS** {**MASS**}

- Metoda „Sammon Mapping“ pa minimizira tole mero:

$$Sammonov\ STRESS = \frac{1}{\sum_{i>j} d_{ij}} \sum_{i>j} \frac{(d_{ij} - \tilde{d}_{ij})^2}{\sum_{i>j} d_{ij}}$$

R: Funkcija **sammon** {**MASS**}

- Uporabljata se tudi meri  $STRESS2$  in  $S - STRESS$ :

$$STRESS2 = \frac{\sum_{i>j} (f(d_{ij}) - \tilde{d}_{ij})^2}{\sum_{i>j} (f(\tilde{d}_{ij}) - \overline{f(\tilde{d})})^2}$$

$$S - STRESS(\text{ver 1}) = \sum_{i>j} \frac{(f(d_{ij})^2 - \tilde{d}_{ij}^2)^2}{f(d_{ij})^4}$$

Ali

$$S - STRESS(\text{ver 1}) = \sum_{i>j} w_{ij} (f(d_{ij})^2 - \tilde{d}_{ij}^2)^2$$

# VRL v SPSS-u

- V SPSS-u obstajata dva algoritma za VRL in sicer PROXSCAL in ALSCAL. Mi bomo uporabljali 1., saj drugi omeji število „objektov“ na 100.
- PROXSCAL v enostavni obliki minimizira nasledijo mero (STRESS1 z uteževanjem):

$$STRESS_{PROXSCAL} = \frac{\sum_{i < j} w_{ij} (f(d_{ij}) - \tilde{d}_{ij})^2}{\sum_{i < j} \tilde{d}_{ij}^2}$$

- Kjer je  $f$  monotona transformacija.



# Transformacije mer podobnosti

`mds/smacofSym {smacof}` in SPSS ponujajo več načinov obravnavanja podobnosti in sicer:

- Ratio → transforirane mere podobnosti so sorazmerne originalnim → **METRIČNO**
- Interval → transforirane mere podobnosti so sorazmerne originalnim + neka konstanta → **METRIČNO**
- Ordinal → transforirane mere podobnosti imajo enak vrsni red kot originalne → **NE-METRIČNO**  
**VRL**
- Spline → ne bomo obravnavali

# Lokalna optimizacija

- VRL je neke vrste lokalna optimizacija, zato imam pomemben vpliv na končni rezultat izbor začetnega stanja.
- Najbolj varna možnost je večkrat ponoviti postope z naključnimi konfiguracijami, je pa tudi najpočasnejša.

# Število dimenzij in kvaliteta rešitve

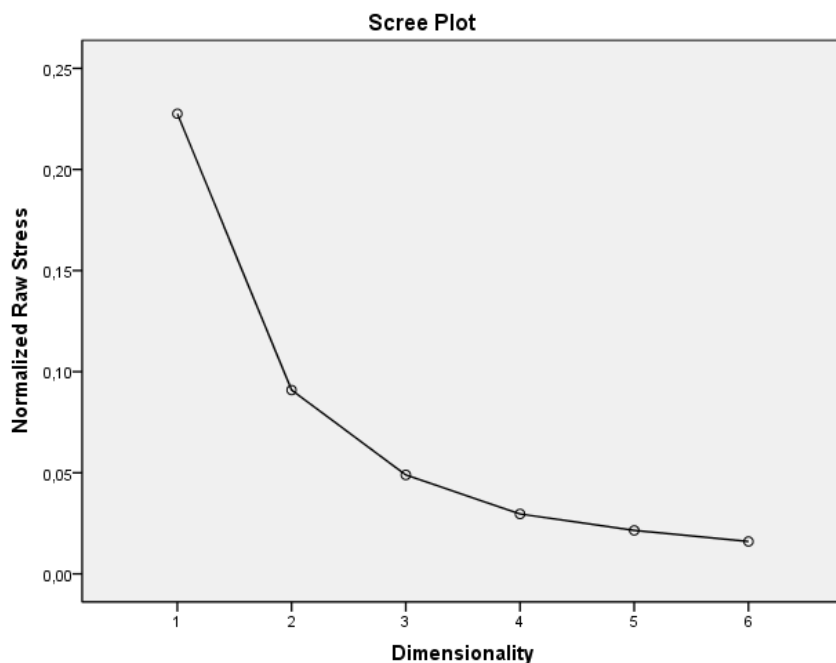
- Število dimenzij zaradi grafične predstavitve pogosto nastavimo na 2.
- Za določitev števila dimenzij lahko sicer uporabimo tudi „Scree plot“, kjer na x os nanašamo število dimenzij, na y pa stress.
- Interpretacija *STRESS1* (Kruskal 1964) :
  - 0,20 – slabo;
  - 0,10 – sprejemljivo;
  - 0,05 – dobro;
  - 0,025 – odlično;
  - 0 – idealno.

# Primer 1

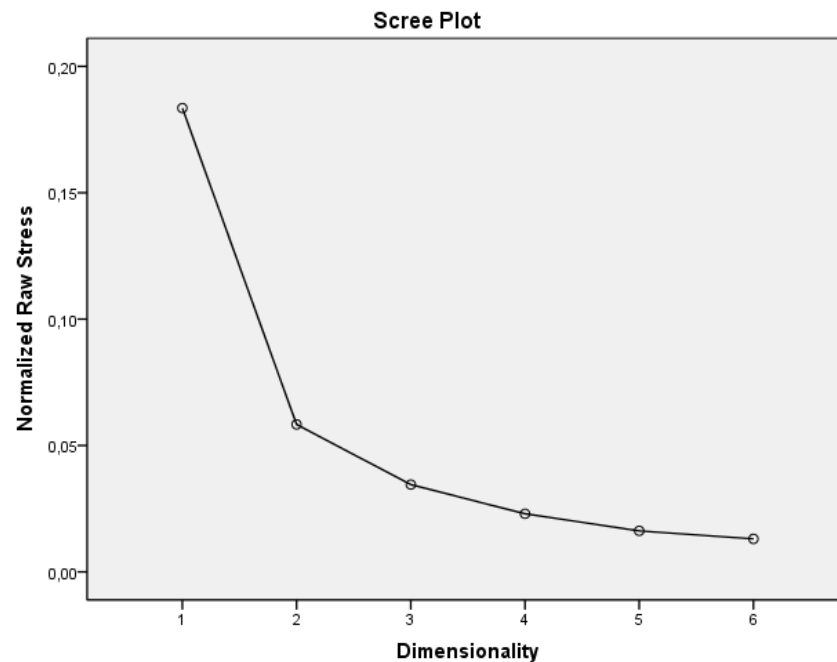
- Podatki so bili zbrani v okviru raziskave *Kakovost merjenja egocentričnih socialnih omrežij* (Ferligoj in drugi, 2000) leta 2000. Vzorec vsebuje 1033 prebivalcev Ljubljane. Analiza je bila narejena na 631 prebivalcih, ki so bili osebno intervjuvani.
- Enote bomo narisali v dvodimenzionalnem prostoru tako, da bomo čim bolje ohranili razlike med enotami pri spremenljivkah emocionalne stabilnosti in ekstrovertiranosti. Večrazmernostno lestvičenje bomo uporabili na evklidski razdalji, izračunani na standardiziranih spremenljivkah.

# Scree plot

## Ratio



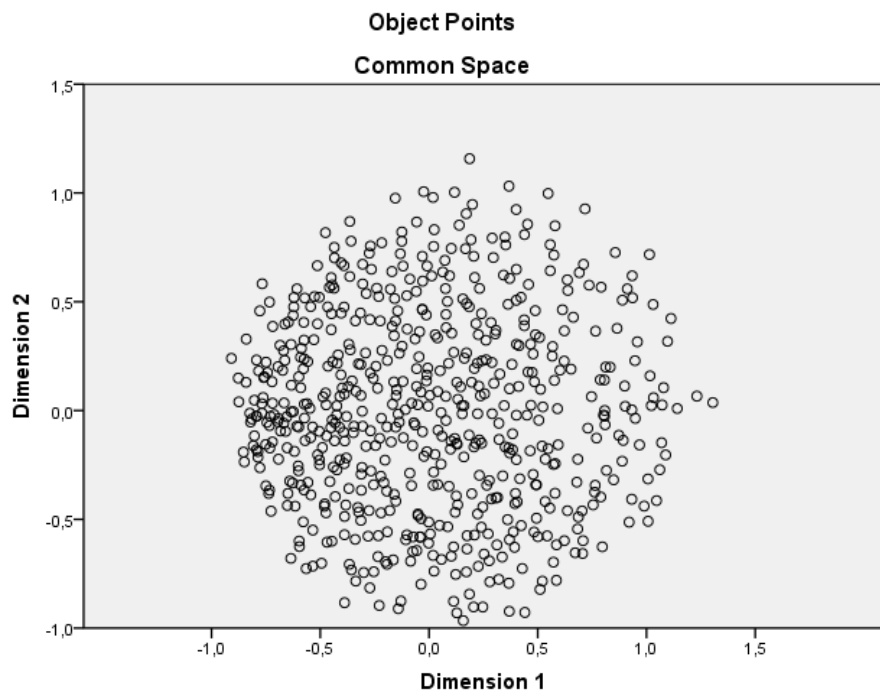
## Ordinal



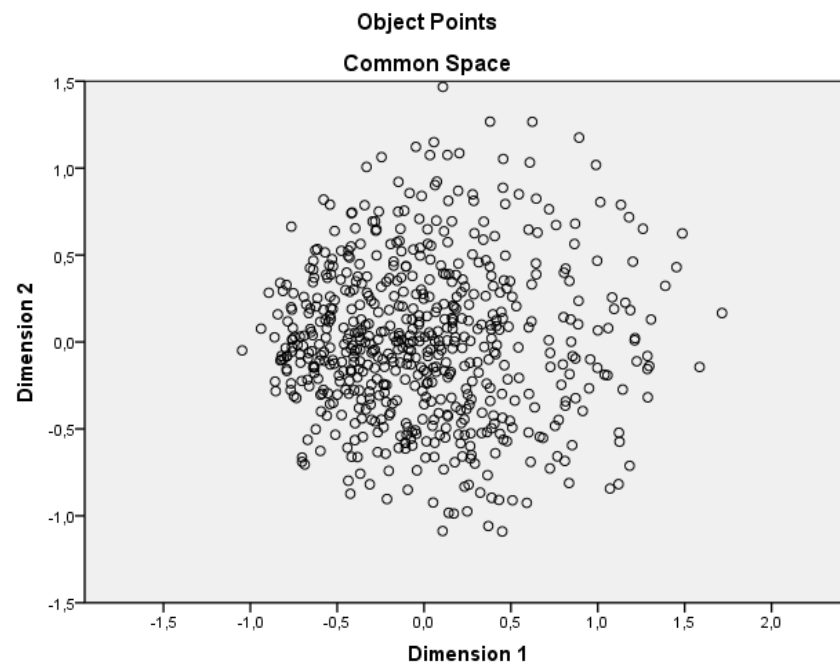
- Zaradi enostavne grafične predstavitve bomo uporabljali 2 dimenziji, kar pa se tudi nakazuje kot primerno na grafikonih

# Enote v prostoru

## Ratio



## Ordinal



# Stress

## Ratio

### Stress and Fit Measures

Dimensionality:2

Normalized Raw Stress	,09091
Stress-I	,30151 <sup>c</sup>
Stress-II	,69366 <sup>c</sup>
S-Stress	,17781 <sup>d</sup>
Dispersion Accounted For (D.A.F.)	,90909
Tucker's Coefficient of Congruence	,95346

PROXSCAL minimizes  
Normalized Raw Stress.

c. Optimal scaling factor = 1,100.  
d. Optimal scaling factor = ,910.

## Ordinal

### Stress and Fit Measures

Dimensionality:2

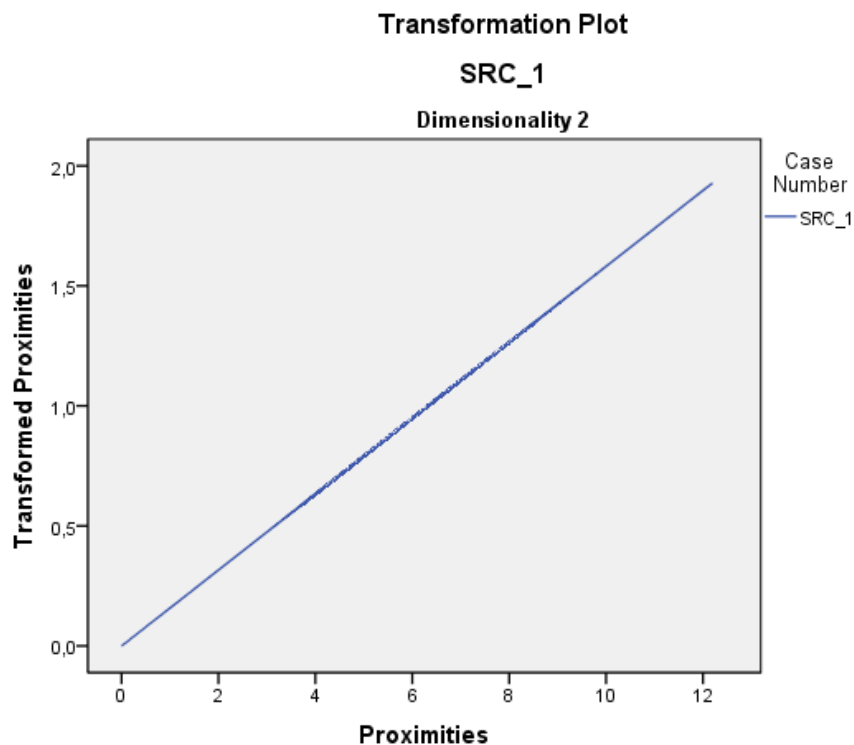
Normalized Raw Stress	,05834
Stress-I	,24153 <sup>c</sup>
Stress-II	,51670 <sup>c</sup>
S-Stress	,11196 <sup>d</sup>
Dispersion Accounted For (D.A.F.)	,94166
Tucker's Coefficient of Congruence	,97039

PROXSCAL minimizes  
Normalized Raw Stress.

c. Optimal scaling factor = 1,062.  
d. Optimal scaling factor = ,982.

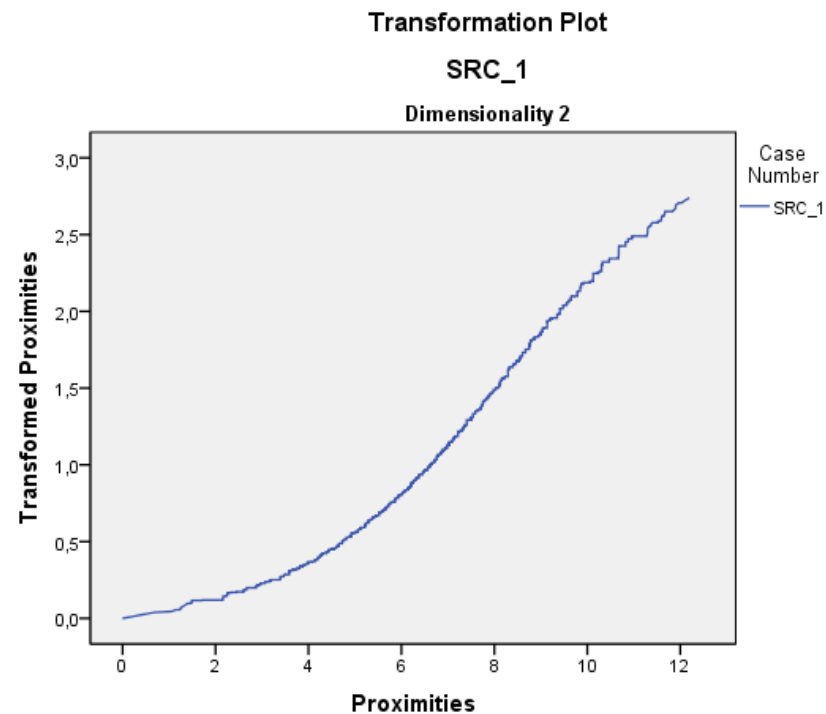
# Transformacija razdalj

## Ratio



Transformation: matrix conditional, ratio.

## Ordinal

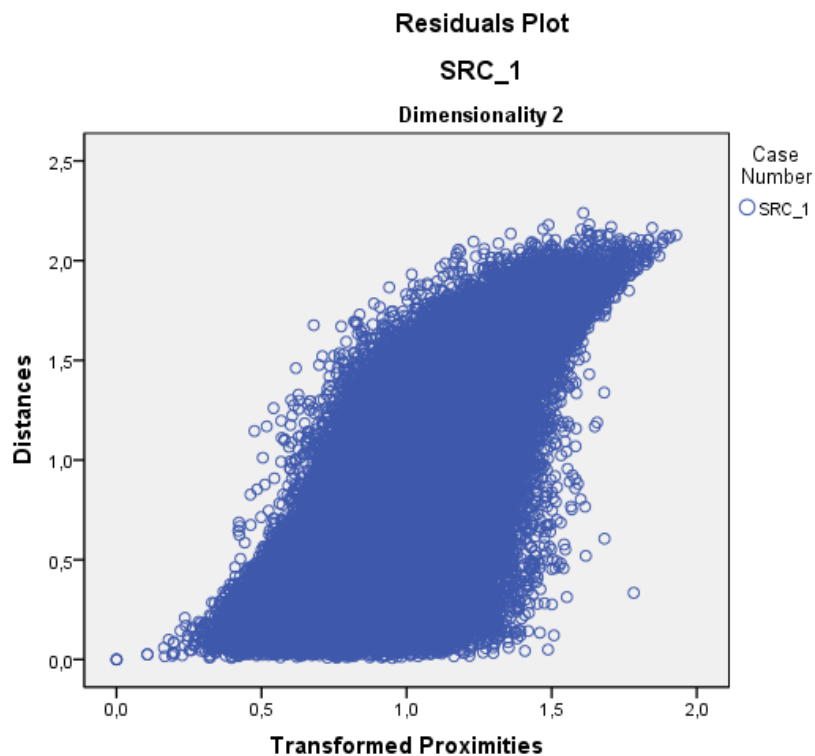


Transformation: matrix conditional, ordinal (ties kept tied).

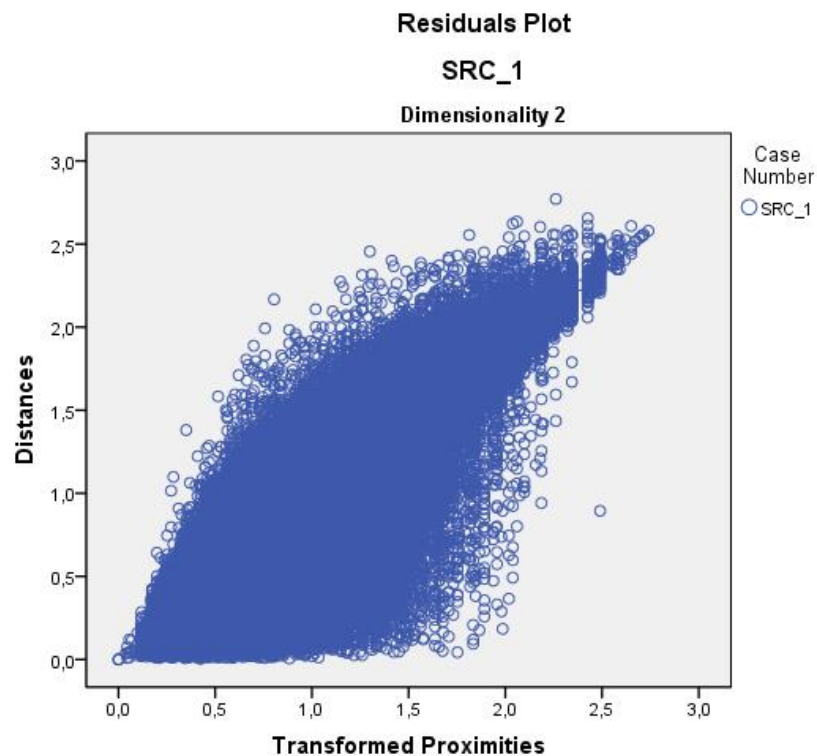


# Prileganje razdalj

Ratio

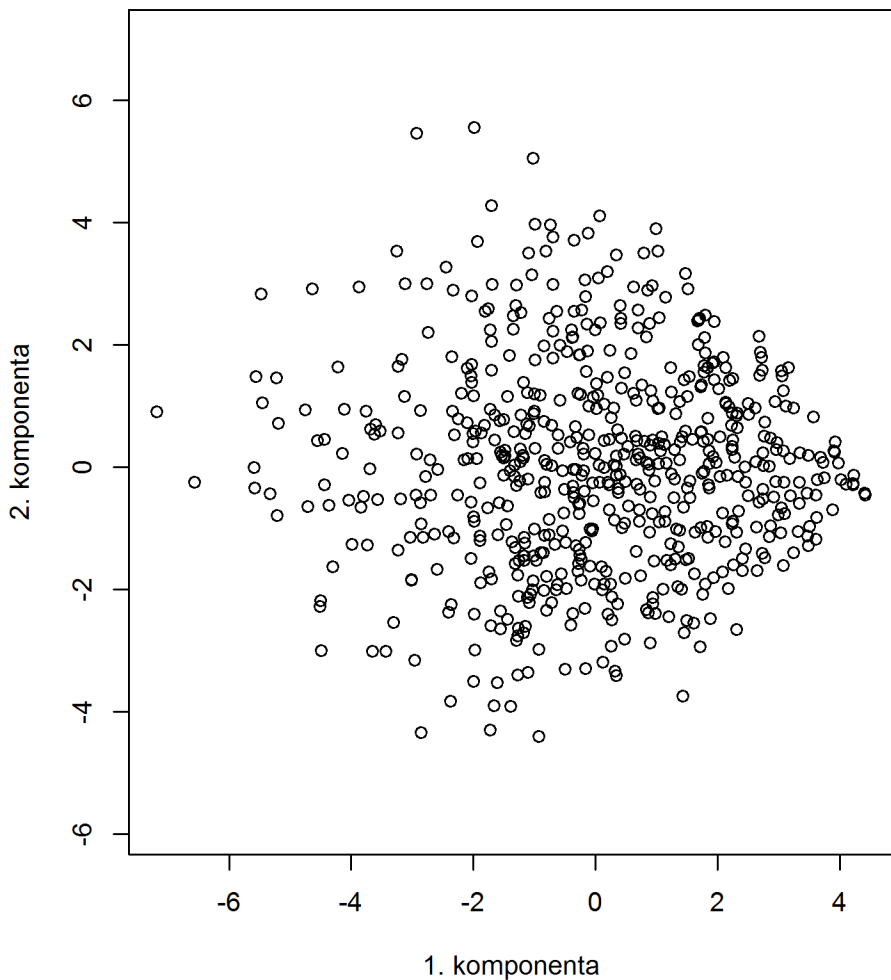


Ordinal

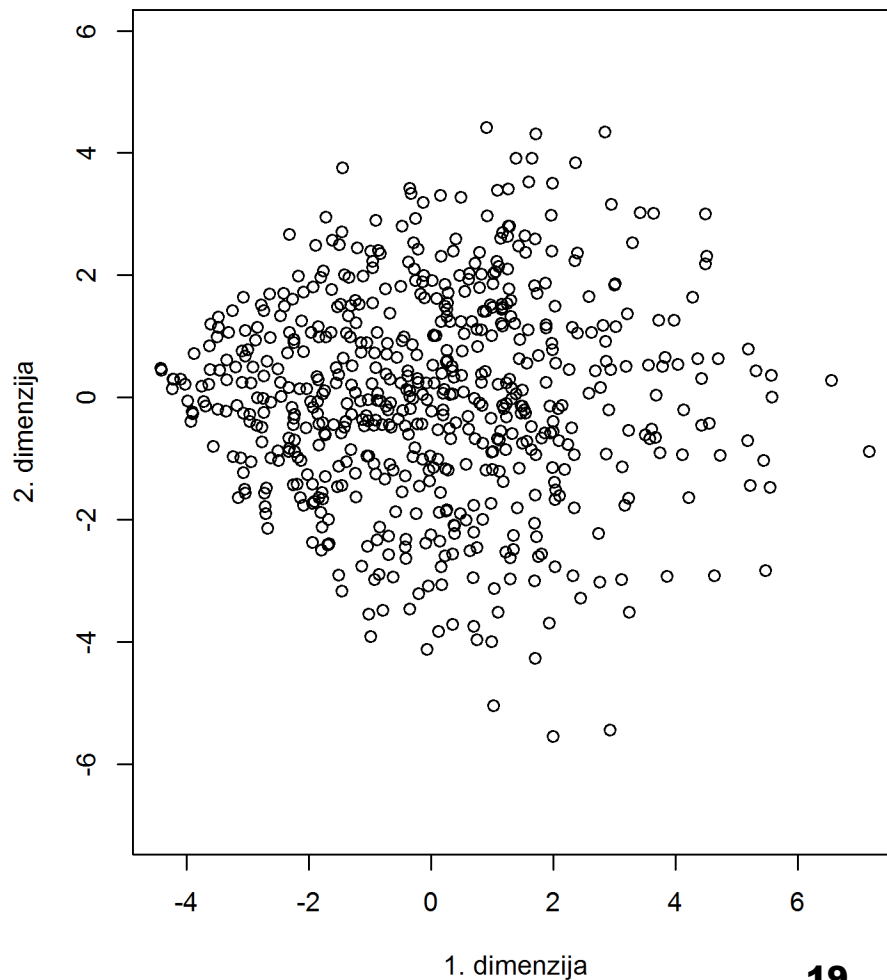


# Primerjava VRL in MGK

Metoda glavnih komponent



Klasično metrično večrazsežno lestvičenje



# Primer 2

- Uporabljamo podatke o podobnosti strank v slovenskem parlamentu leta 1994. Podatki so bili pridobljeni na spletni strani: <http://vlado.fmf.uni-lj.si/pub/networks/data/soc/samo/stranke94.htm>, več pa lahko najdete v: Samo Kropivnik and Andrej Mrvar: An Analysis of the Slovene Parliamentary Parties Network. Developments in Statistics and Methodology. (A. Ferligoj, A. Kramberger, editors) Metodološki zvezki 12, FDV, Ljubljana, 1996, p. 209-216.
- Enote bomo narisali v dvodimenzionalnem prostoru tako, da bomo čim bolje ohranili razlike in podobnosti med enotami.

# Opombe pred analizo

- Upoštevati moramo, da podatki predstavljajo podobnosti (negativne vrednosti podobnosti pa so seveda različnosti)
- Zaradi negativnih vrednosti moramo uporabiti vsaj „interval“ model. Tega smo tudi uporabili.

# Rezultati – Stress in koordiante

**Stress and Fit Measures**

Normalized Raw Stress	,00797
Stress-I	,08927 <sup>a</sup>
Stress-II	,18337 <sup>a</sup>
S-Stress	,02262 <sup>b</sup>
Dispersion Accounted For (D.A.F.)	,99203
Tucker's Coefficient of Congruence	,99601

PROXSCAL minimizes Normalized Raw Stress.

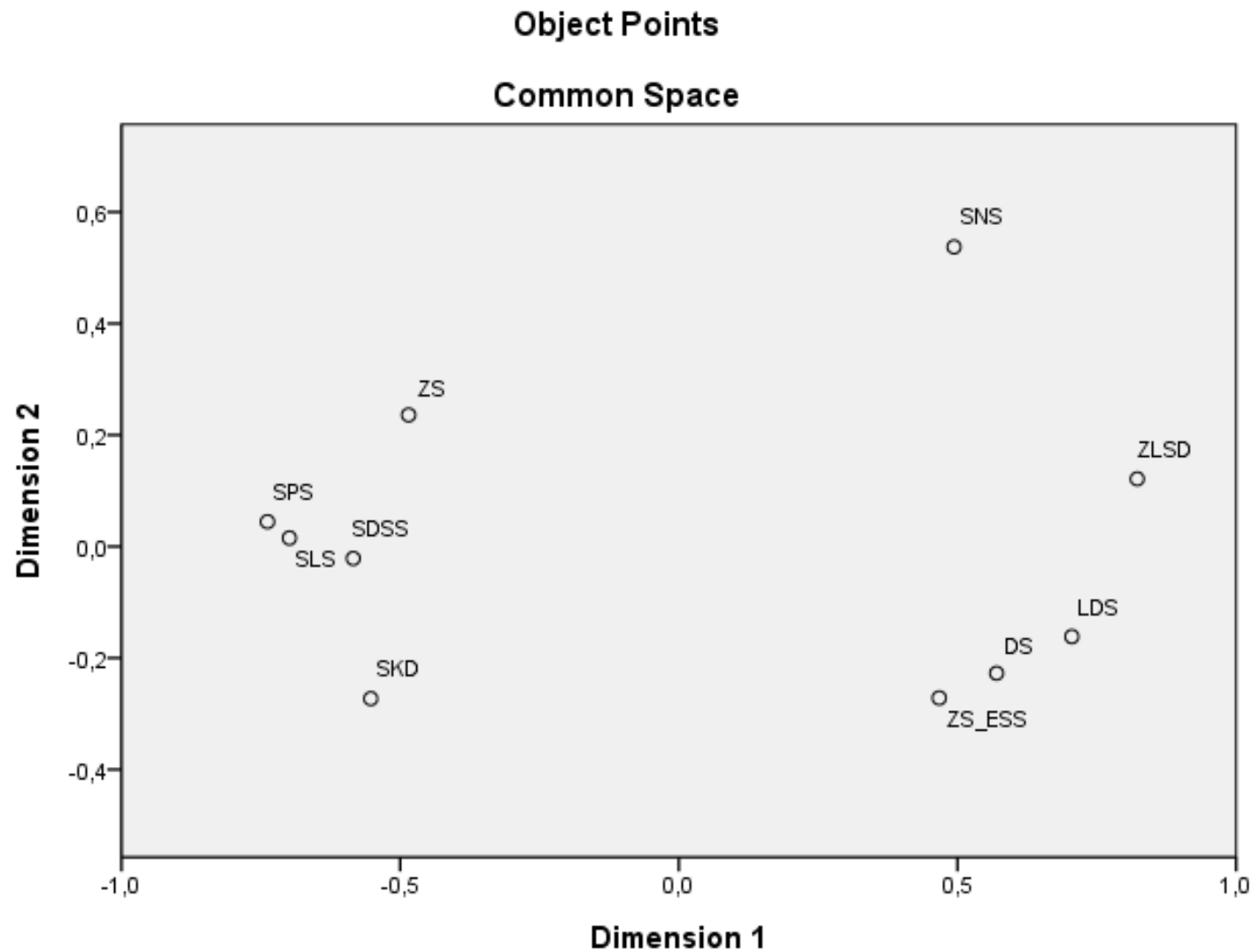
a. Optimal scaling factor = 1,008.

b. Optimal scaling factor = 1,001.

**Final Coordinates**

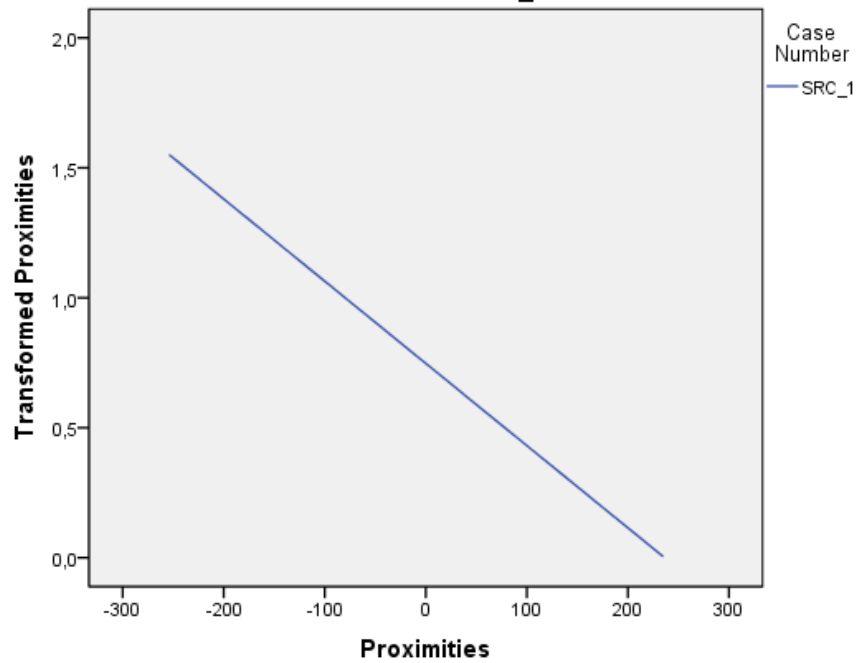
	Dimension	
	1	2
SKD SKD	-,553	-,273
ZLSD ZLSD	,823	,121
SDSS SDSS	-,584	-,021
LDS LDS	,705	-,162
ZS_ESS ZS-ESS	,467	-,272
ZS ZS	-,485	,236
DS DS	,570	-,227
SLS SLS	-,699	,015
SPS SPS	-,738	,045
SNS SNS	,494	,537

# Rezultati - slika



Transformation Plot

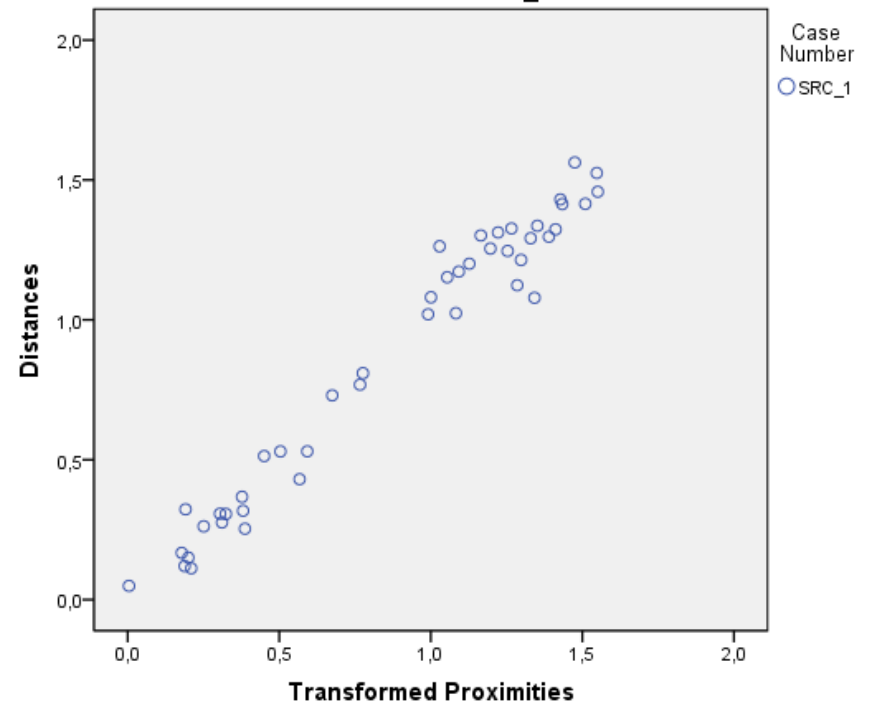
SRC\_1



Transformation: matrix conditional, interval.

Residuals Plot

SRC\_1



# Korespodnenčna analiza

- Cilj metode je prikazati povezanost med dvema nominalnima spremenljivkama oz. podobnosti med kategorijami teh dveh spremenljivk v  $k$ -dimenzionalnem „zveznem“ prostoru ( $k$  je skoraj vedno 2)
- Koordinate dobimo tako, da razčlenimo ustrezno „centrirano“ in normalizirano kontingenčno matriko z metodo razcepa s singularnimi vrednostmi (Singular Value Decomposition – SVD)



- Matrika, ki jo analiziramo z SVD je:

$$X_{ij} = \frac{n_{ij} / n - (n_{i.}/n)(n_{.j}/n)}{\sqrt{(n_{i.}/n)(n_{.j}/n)}} = \frac{n_{ij} - n r_i c_j}{n \sqrt{r_i c_j}}$$

,kjer  $r_i = n_{i.}/n$  in  $c_j = n_{.j}/n$

- Dobljene koordinate posameznih kategorij potem še delimo z koreni deleži posameznih kategorij.

# Primer

- Podatki so bili zbrani v okviru raziskave *Kakovost merjenja egocentričnih socialnih omrežij* (Ferligoj in drugi, 2000) leta 2000. Vzorec vsebuje 1033 prebivalcev Ljubljane. Analiza je bila narejena na 631 prebivalcih, ki so bili osebno intervjuvani.
- Zanima nas, kako so povezane posamezne kategorije izobrazbe in poklica.

# Kontingenčna tabela

Correspondence Table

POKLIC poklic	D9 izobrazba								
	nedokončana osnovna šola	osnovna šola	poklicna šola	štiriletna srednja šola	višja šola	visoka šola	magisterij	doktorat	Active Margin
manager,vodilni delavec, lastnik podjetja	0	0	0	1	0	5	1	0	7
srednji manager, vodstveni delavec	0	0	0	5	7	5	1	0	18
samostojni podjetnik, obrtnik	0	0	3	7	3	1	0	0	14
samozaposleni strokovnjak	0	0	1	4	2	11	0	1	19
zaposleni strokovnjak	0	0	4	23	18	61	4	6	116
uradnik	0	0	0	17	2	2	0	0	21
pisarniški delavec	0	1	2	25	8	3	0	0	39
delavec	1	12	31	23	2	1	0	0	70
Active Margin	1	13	41	105	42	89	6	7	304

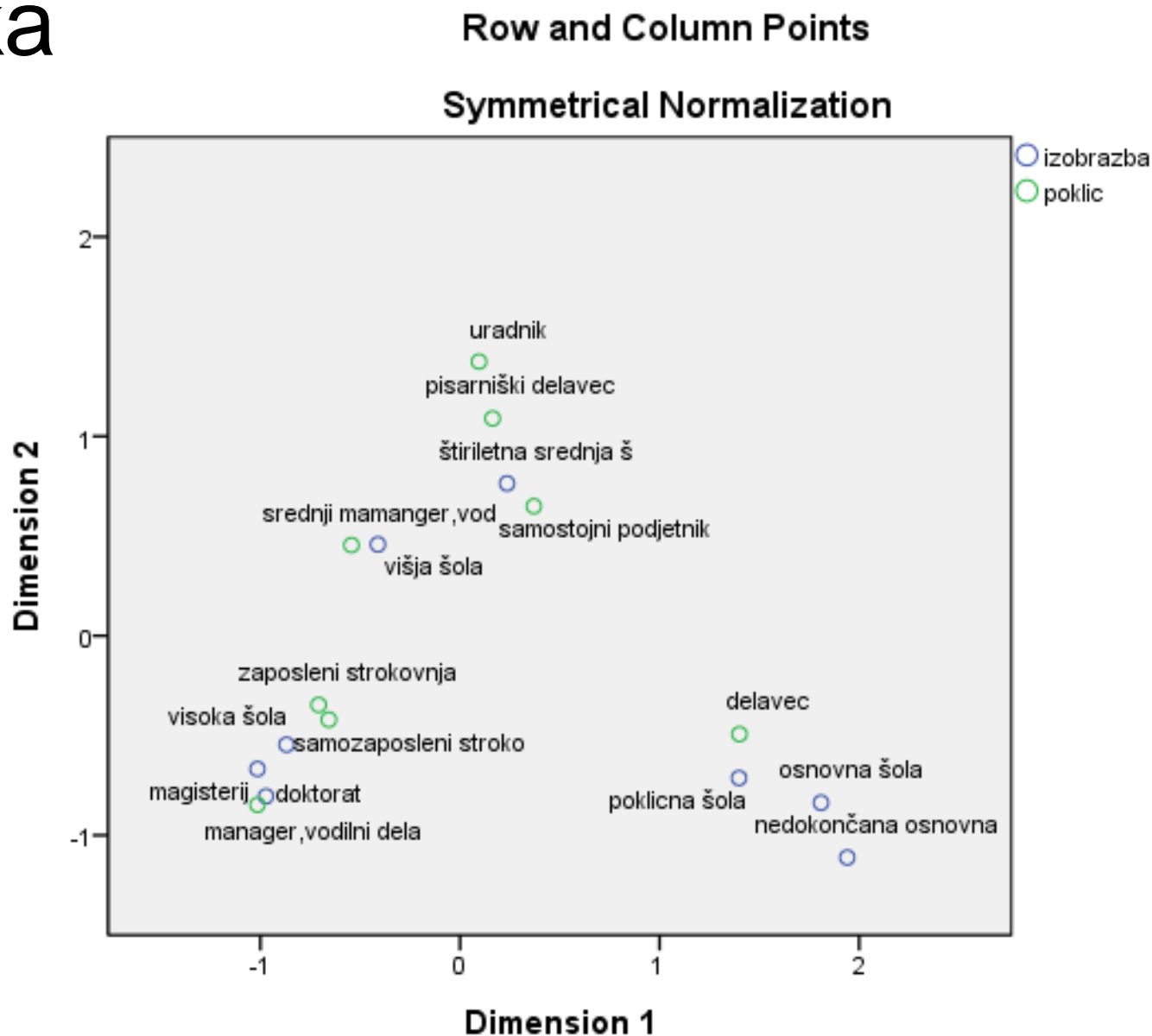
# Pojasnjevanje

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
								2
1	,721	,520			,663	,663	,033	,436
2	,444	,197			,251	,914	,047	
3	,199	,040			,050	,964		
4	,150	,022			,029	,993		
5	,068	,005			,006	,999		
6	,032	,001			,001	1,000		
7	,008	,000			,000	1,000		
Total		,785	238,680	,000 <sup>a</sup>	1,000	1,000		

a. 49 degrees of freedom

# Slika



# Multipla korespodnenčna analiza

- Cilj metode je prikazati povezanost med več nominalnimi spremenljivkama oz. podobnosti med kategorijami teh spremenljivk v  $k$ -dimenzionalnem „zveznem“ prostoru ( $k$  je skoraj vedno 2) ter v tem prostoru predstaviti tudi enote
- POZOR: V primeru samo dveh spremenljivk ne dobimo „navadne“ korepondenčne analize
- Koordinate dobimo tako, da izvedemo „navadno“ korespondečno analizo na matriki, kjer so spremenljivke predstavljene z „umetnimi“ (dummy) spremenljivkami (vse kategorije!)

# Primer

- Podatki so bili zbrani v okviru raziskave *Kakovost merjenja egocentričnih socialnih omrežij* (Ferligoj in drugi, 2000) leta 2000. Vzorec vsebuje 1033 prebivalcev Ljubljane. Analiza je bila narejena na 631 prebivalcih, ki so bili osebno intervjuvani.
- Zanima nas, kako so povezane posamezne kategorije izobrazbe, poklica, tipa hiše in zakonskega stanu.

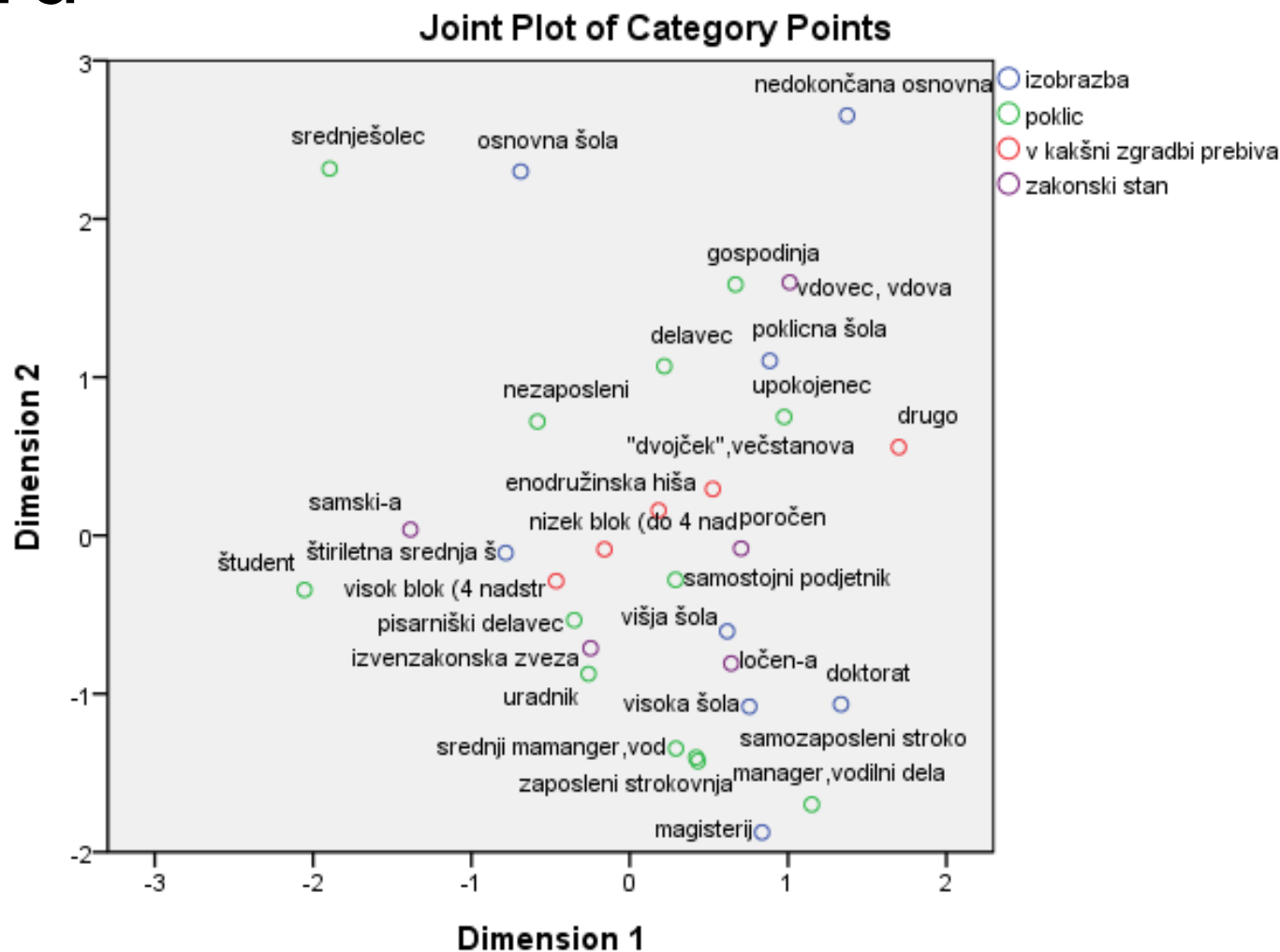
# Kvaliteta modela

**Model Summary**

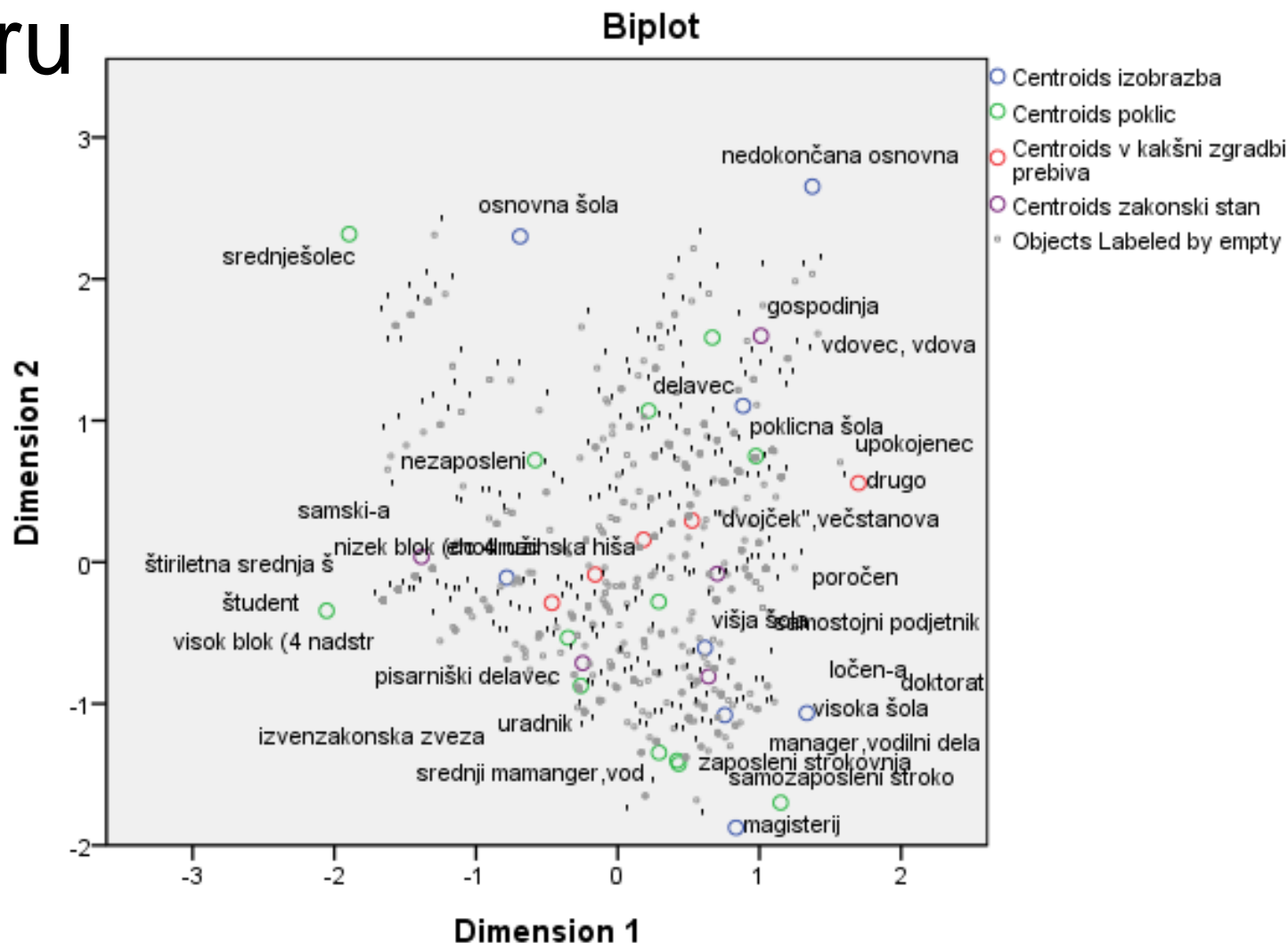
Dimension	Cronbach's Alpha	Variance Accounted For	
		Total (Eigenvalue)	Inertia
1	,626	2,004	,401
2	,515	1,700	,340
3	,457	1,577	,315
4	,407	1,482	,296
5	,299	1,315	,263
6	,245	1,244	,249
7	,212	1,204	,241
8	,197	1,187	,237
9	,169	1,157	,231
10	,111	1,097	,219
11	,087	1,075	,215
12	,074	1,063	,213
13	,022	1,018	,204
14	,001	1,001	,200
15	-,023	,982	,196
16	-,069	,949	,189



# Kategorije vseh spremenljivk v prostoru



# Enote in centoridi kategorij v prostoru



Symmetrical Normalization.

# Več o (multipli in navadni) korespodnečni analizi

- PREUČEVANJE POVEZANOSTI  
KATEGORIALNIH SPREMENLJIVK S  
KORESPONDENČNO ANALIZO

*prof.dr. Jože Rován*, Univerza v Ljubljani,  
Ekonomska fakulteta (12.12.2006) URL:


- <http://ibmi3.mf.uni-lj.si/ibmi/biostat-center/gradiva/RovanKorespAna1.pdf>
- <http://ibmi3.mf.uni-lj.si/ibmi/biostat-center/gradiva/RovanKorespAna2.pdf>

# Splošni linearni model

- Model, ki ga lahko zapišemo z enačbo

$$Y = XB + E$$

- Y predstavlja 1 ali več odvisnih spremenljivk
- X, predstavlja „design matrix“, nekakšno matriko učinkov. V enostavnem primeru je to kar matrika neodvisnih spremenljivk (vključno z umetnimi za nominalne spremenljivke)
- B je matrika parametrov, ki se jih ocenjuje
- E je matrika napaka (enakih dimenzij kot Y)

- 
- Posebni primeri splošnega linearnega modela so:
    - Linearna regresija (samo ena nedovisna spremenljivka)
    - Različice t-testa
    - ANOVA, ANCOVA, MANOVA, MANCOVA

# Binarna logistična regresija

- Podobno kot linearna regresija, kjer je odvisna spremenljivka binarna

$$\log \left( \frac{p(y)}{1 - p(y)} \right) = XB = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$$

- Uporablja se podobno kot diskriminanta analiza, le da je po navadi poudarek na pojasnjevanju (in ne napovedovanju).

# Binarna logistična regresija - primer

- Napovedovanje/pojasnjevanje udeležbe na volitvah
- Podatki: ESS 2004 za Slovenij
- Odvisna spremenljivka:
  - Ali ste volili na zadnjih volitvah (Da – 1, Ne – 0)
- Neodvisne spremenljivke:
  - starost
  - levoDesno (0 – Levo, 10 – desno)
  - Izobrazba (št. let šolanja)
  - zenska
  - Kraj (1 – Veliko mesto, 2 – Malo mesto, 3 – Podeželje)

# Binarna logistična regresija - primer

## ■ Ocenjevanje celotnega modela

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	103.187	6	.000
	Block	103.187	6	.000
	Model	103.187	6	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	989.109 <sup>a</sup>	.102	.150

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

- Domnevo, da so vsi koeficienti enaki 0, lahko zavrnamo pri zanemarljivi stopnji tveganja (enakovredno F-testu pri linearni regresiji)
- Tu ne moremo govoriti o pravem % pojasnjene variabilnosti, a tu sta dve podobni meri. Lahko bi rekli, da je za 10-15% boljši od modla brez neodvisnih spremenljivk.<sup>41</sup>



# Binarna logistična regresija - primer

## ■ Ocenjevanje celotnega modela

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.948	8	.542

- Precej vprašljiv test prileganja.  $H_0$ : Podatki se prilegajo modelu

Classification Table<sup>a</sup>

			Predicted		
			volil Ali ste volili na zadnjih voltivah		Percentage  Correct
			0 No	1 Yes	
Observed					
Step 1	volil Ali ste volili	0 No	37	208	15.1
	na zadnjih voltivah	1 Yes	27	691	96.2
Overall Percentage					75.6

a. The cut value is .500

- Klasifikacijska tabela (podobno kot pri diskriminatni analizi)
- Tabela je pristranska, saj je model ocenjen na istih podatkih

# Binarna logistična regresija - primer

## ■ Ocenjevanje regresijskih parametrov

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
starost	.045	.005	78.516	1	.000	1.046	1.036	1.057
levoDesno	.045	.036	1.587	1	.208	1.046	.975	1.123
izobrazba	.147	.027	29.954	1	.000	1.159	1.099	1.222
zenska	-.175	.157	1.235	1	.267	.840	.617	1.143
Kraj			1.879	2	.391			
Kraj (1)	.051	.200	.065	1	.799	1.052	.710	1.559
Kraj (2)	-.233	.197	1.391	1	.238	.792	.538	1.167
Constant	-2.711	.493	30.185	1	.000	.066		

a. Variable(s) entered on step 1: starost, levoDesno, izobrazba, zenska, Kraj.

# Binarna logistična regresija - primer

- Ocenjevanje regresijskih parametrov
- Pri 5% tveganju sta statistično značilna le vpliva starosti in izobrazbe.
- Primer interpretacije: Koeficient za starost (B) je 0,45,  $e^B = 1,046$ . Če se starost poveča za eno leto (in ostale spr. ostanejo nespremenjene), se „šanse“ za to, da je nekdo volil ( $p(\text{volil})/p(\text{ni volil})$ ) povečajo za 4,6 %.

# Logistična regresija za nominalne spremenljivke

- Podobna kot logistična regresija, le da skupaj ocenimo  $K - 1$  logističnih regresij, kjer je  $K$  število kategorij odvisne spremenljivke\*:

$$\log \left( \frac{p(y = k)}{p(y = K)} \right) = XB_k = \beta_0 + \beta_{k1}X_1 + \cdots + \beta_{km}X_m$$

- Verjetnosti ocenimo tako, da so (pri vseh vrednostih neodvisnih spremenljivk) vsote verjetnosti za vseh  $k$  kategorij enake 1.
- Lahko se ocenjuje preko posplošenega linearnega modela → glej nadaljevanje

\* Zgornja enačba predvideva, da smo kot referenčno kategorijo vzeli zadnjo kategorijo!

# Ordinalna logistična regresija

- Podobno kot binarna logistična regresija

$$\log \left( \frac{p(y \leq k)}{p(y > k)} \right) = \beta_{0k} - XB = \beta_{0k} - (\beta_1 X_1 + \dots + \beta_m X_m)$$

- $- \rightarrow$  ponekod se namesto  $-$  uporablja  $+$   $\rightarrow$  v tem primeru je interpretacija koeficientov ravno obratna.
- Primerjamo verjetnost, da ima sprem. manjšo ali enako vrednost z možnostjo, da ima večjo.
- Pri tem uporabimo omejitve, da so vsi regresijski koeficienti (torej razen konstante) enaki pri vseh regresijah/modelih ( $K - 1$  enačb/regresij).
- Lahko se ocenjuje preko posplošenega linearnega modela  $\rightarrow$  glej nadaljevanje

# Ordinalna logistična regresija - primer

- Pojasnjevanje druženja z prijatelji/sorodniki
- Podatki: ESS 2004 za Slovenijo
- Odvisna spremenljivka:
  - Kako pogosto se dobivate s prijatelji, sorodniki, sodelavci... (1 – nikoli, ..., 7 – vsak dan)
- Neodvisne spremenljivke:
  - starost
  - Velikost gospodinjstva (št. članov)
  - izobrazba (št. let šolanja)
  - spol (1 – moški, 2 – ženski)
  - Kraj (1 – Veliko mesto, 2 – Malo mesto, 3 – Podeželje)

# Ordinalna logistična regresija - primer

## ■ Celoten model

Model Fitting Information

	-2 Log Likelihood	Chi-Square	df	Sig.
Model				
Intercept Only	4781.470			
Final	4560.696	220.774	6	.000

Link function: Logit.

Pseudo R-Square

Cox and Snell	.148
Nagelkerke	.152
McFadden	.045

Link function: Logit.

- Vsi koeficienti niso enaki 0 ( $p < 0,0005$ )
- Model je za malce boljši od modela brez koeficientov (4,5% - 15,2 %)

# ■ Regresijski parametri

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[C2 = 1]	-6.439	.383	282.173	1	.000	-7.190	-5.687
	[C2 = 2]	-4.359	.331	173.887	1	.000	-5.007	-3.711
	[C2 = 3]	-3.421	.323	112.127	1	.000	-4.054	-2.788
	[C2 = 4]	-2.343	.317	54.714	1	.000	-2.963	-1.722
	[C2 = 5]	-1.502	.313	23.042	1	.000	-2.116	-.889
	[C2 = 6]	.068	.313	.047	1	.829	-.545	.680
Location	starost	-.041	.003	178.161	1	.000	-.047	-.035
	Izobrazba	-.029	.015	3.663	1	.056	-.059	.001
	velikostGosp	-.069	.037	3.529	1	.060	-.141	.003
	[spol=1]	.289	.097	8.948	1	.003	.100	.479
	[spol =2]	0 <sup>a</sup>	.	.	0	.	.	.
	[Kraj=1]	.343	.123	7.787	1	.005	.102	.585
	[Kraj=2]	.333	.124	7.204	1	.007	.090	.576
	[Kraj=3]	0 <sup>a</sup>	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.



# Ordinalna logistična regresija - primer

- Ocenjevanje regresijskih parametrov
- Pri 5% tveganju so statistično značilni vplivi starosti, spola in kraja bivanja.
- Primer interpretacije: Koeficient za starost (B) je -0,41,  $e^B = 0,960$ . Če se starost poveča za eno leto (in ostale spr. ostanejo nespremenjene), se obeti za biti nad določeno kategorijo v primerjavi z biti pod njo zmanjšajo za 4,0 %. Torej starejši se manj pogosto družijo s prijatelji, sorodniki, sodelavci.

# Posplošeni linearni model

- Posplošitev linearne regresije
- Predpostavlja se, da je porazdelitev  $Y$  iz družine eksponentnih porazdelitev
- Posplošitev je v točkah:

- „link function“, torej funkcija, ki povezuje „linearni komponento“ z pričakovano vrednostjo odvisne spremenljivke

$$E(Y) = \mu = g^{-1}(X\beta)$$

- Variabilnost napak je lahko odvisna od napovedne vrednosti za  $Y$  ( $\mu$ )


$$\text{Var}(Y) = V(\mu) = V(g^{-1}(X\beta))$$

# Posplošeni linearni model - logistična

- Logistična regresija je posebne primer posplošenega linearnega modela, kjer je:
  - „Link“ funkcija:  $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$  (pri binarnih podatkih  $(0,1)$  je  $\mu = p$ ).
  - Funkcija V (za izračuna variance):
$$Var(Y_i) = V(\mu_i) = \mu_i(1 - \mu_i)$$

# Analiza preživetja

- Analiza preživetja (tudi analiza zgodovine dogodkov) poskuša napovedati/pojasniti čas do enega (ali v razširjen obliki) več dogodkov.
- Zelo pogosto se uporablja pri medicini (kjer je dogodek smrt, bolezen, resni zapleti pri bolezni, ...) in v tehničnih vedah (dogodek je odpoved neke naprave, elementa, ...).
- Glavna značilnost podatkov je, da običajno za precejšen delež eno nimamo podatka, koliko časa je preteklo do dogodka, ampak le, da v določenem času ni prišlo do dogodka.

- 
- Razlog zato je, ker je enota „izstopila“ iz študije, ali pa se je študija končala, preden je nastopil dogodek pri vseh enotah. Temu rečemo, da so podatki desno „cenzurirani“.

- Primer 1: Recimo da proučujejo na nekem vzorcu, kako kajenje, pitje alkohola, telesna teža, ... vplivajo na nastanek rakavih obolenj. Študija recimo traja 10 let, v tem času pa seveda ne bodo vse osebe iz vzorca zbolele.
- Primer 2: Lahko tudi pri bolnikih z določeno boleznijo preučujemo, kako določeni vidiki zdravljenja, življenjskega sloga in lastnosti oseb vplivajo na preživetje oseb glede na to bolezen. Zopet vse osebe v času študije ne bodo umrle, nekatere pa bodo morda umrle zaradi drugih razlogov (prometna nesreča), kar se tudi upošteva kot cenzuriranje.
- Osebe pa lahko tudi izstopijo iz študije.

# Analiza preživetja - primer

- *Vir: Crowley J, Hu M. Covariance analysis of heart transplant data. J Amer Stat Assoc 1977; 72: 27-36.*
- Preživetje 65 bolnikov, ki so jim presadili srce:
  - Izid: smrt zaradi presaditve (reject=1, reject=0 pomeni krnitev)
  - Čas do dogodka: time
- Neodvisne spremenljivke:
  - starost (age)
  - Enak antigen dajalec in prejemnik (antigen)
  - Neujemanje dajalca in prejemnika (mismatch)

# Analiza preživetja - primer

## ■ Rezultati

Omnibus Tests of Model Coefficients<sup>a</sup>

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
175.735	19.852	3	.000	22.319	3	.000	22.319	3	.000

a. Beginning Block Number 1. Method = Enter

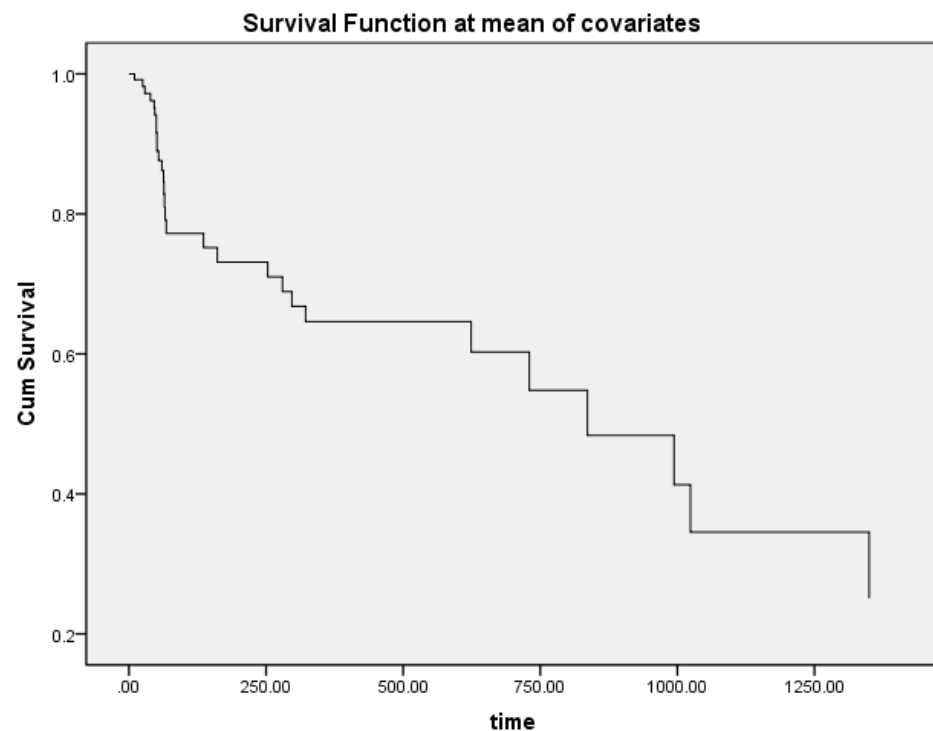
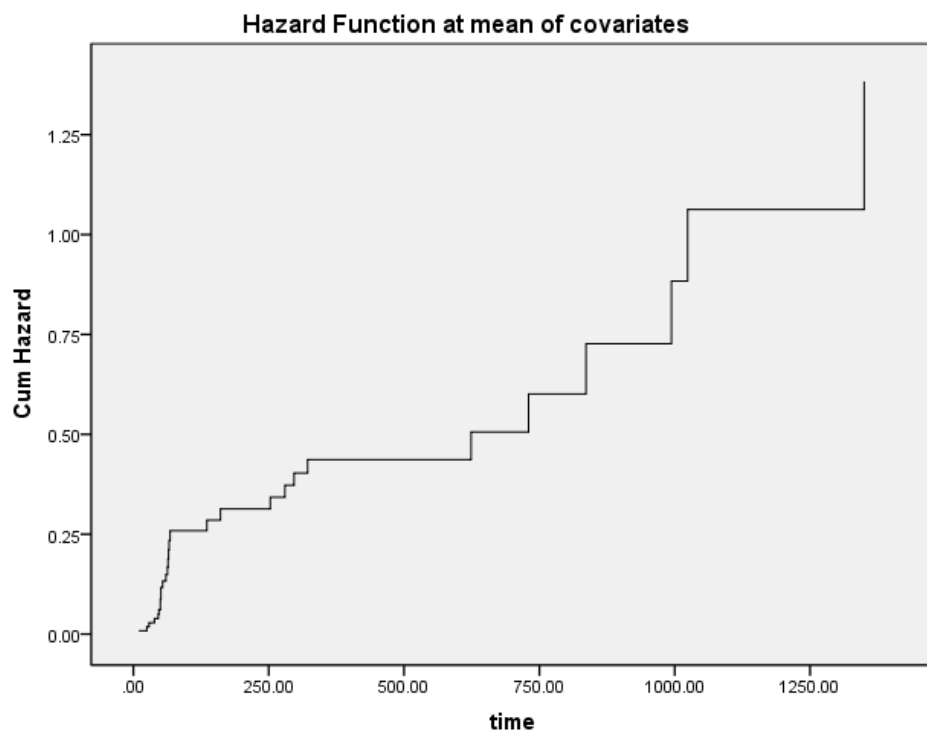
Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
age	.109	.033	10.738	1	.001	1.115	1.045	1.190
antigen	-.049	.472	.011	1	.918	.952	.378	2.400
mismatch	1.064	.395	7.268	1	.007	2.897	1.337	6.279



# Analiza preživetja - primer

## ■ Rezultati

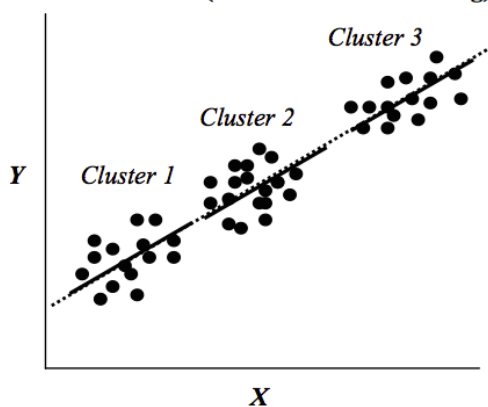


# Večnivojsko modeliranje

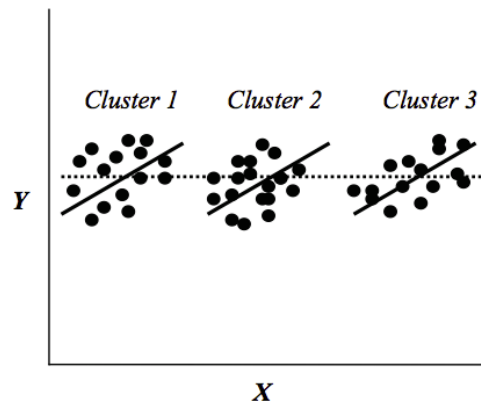
- Uporabljajo se tudi imena: Hierarhično modeliranje, mešani modeli, modeli z slučajnimi koeficienti, ...
- Gre za razširitev klasične linearne regresije, lahko pa tudi prej omenjenih kompleksnejših primerov (npr. posplošeni linearni model)
- Model upošteva, da imamo praktično v modelu enote, ki so iz različnih nivojev oz. da se „osnovne“ enote deli večjih enot in da so neodvisne spremenljivke merjene na različnih ravneh.

# Večnivojsko modeliranje

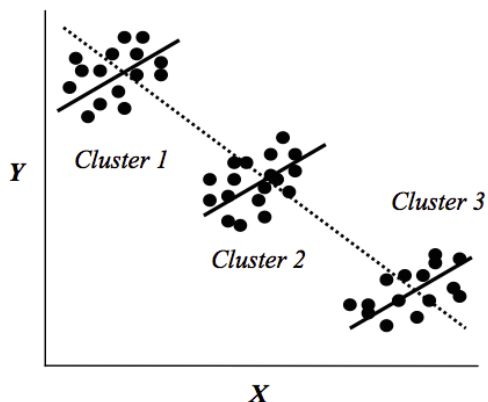
**A. Within-Cluster Effect = Between-Cluster Effect (No Cluster Confounding)**



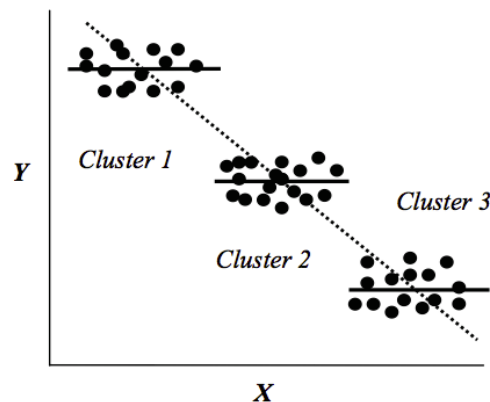
**B. Positive Within-Cluster Effect, Null Between-Cluster Effect**



**C. Positive Within-Cluster Effect, Negative Between-Cluster Effect**



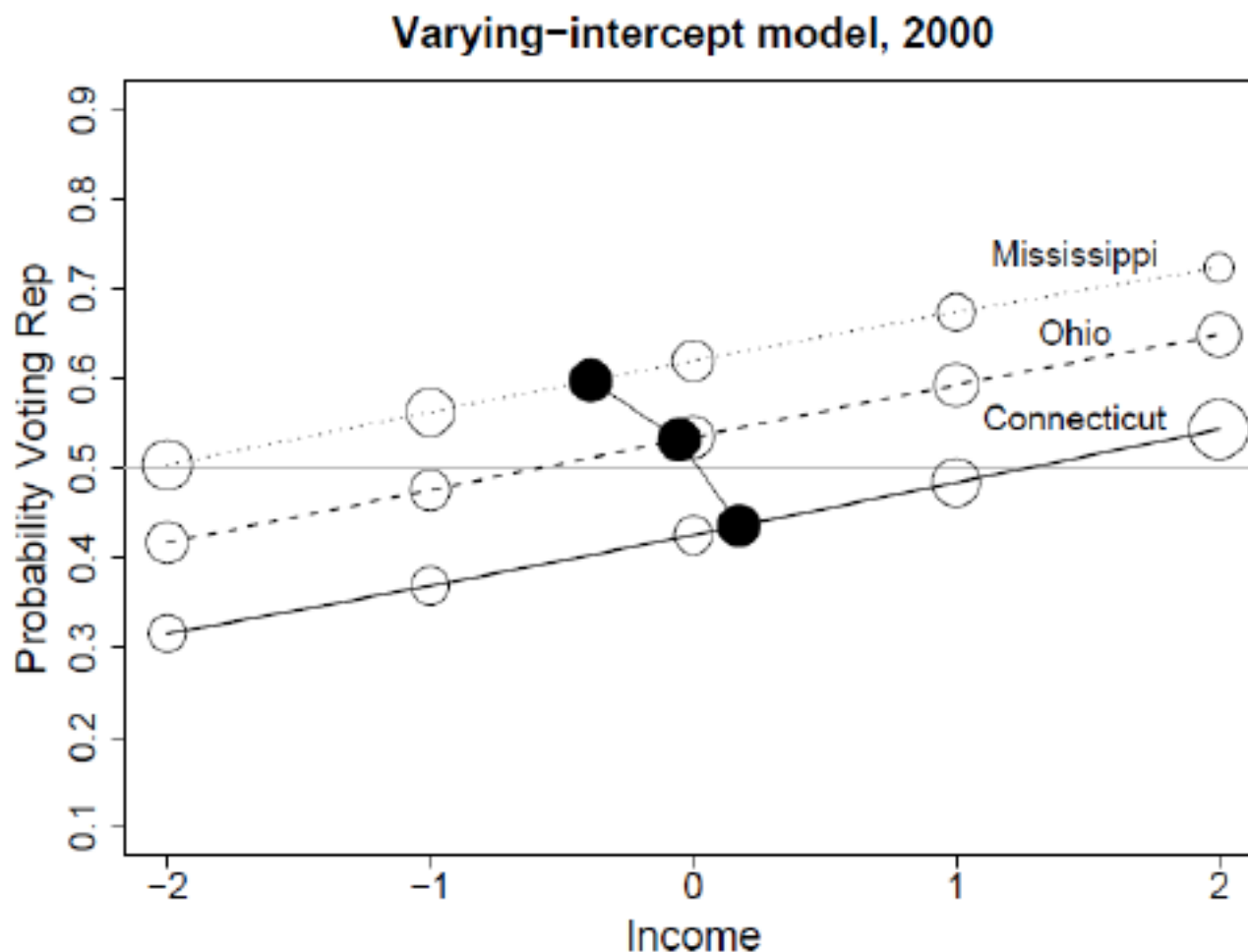
**D. Null Within-Cluster Effect, Negative Between-Cluster Effect**



Vir: Bartels, 2008.

# Večnivojsko modeliranje – primer 1

- Vpliv dohodka na volilne odločitve v ZDA.



Vir: Gelman, Andrew E., Boris Shor, Joseph Bafumi, and David K. Park. 2007. "Rich State, Poor State, Red State, Blue State: What's the Matter with Connecticut?" *Quarterly Journal of Political Science* 2 (4): 345–367.

# Večnivojsko modeliranje – primer 2

- Preučevali so vpliv učenčevega uspeha na matematičnem testu, ocene šole na testu (povprečje vseh učencev) in tipa šole na učenčevo samooceno znanja matematike.

Vir: Marsh, H.W., Trautwein, U., Lüdtke, O., Baumert, J., Köller, O., 2007. The Big-Fish-Little-Pond Effect: Persistent Negative Effects of Selective High Schools on Self-Concept After Graduation. *American Educational Research Journal* 44, 631–669. doi:10.3102/0002831207306728

*Table 1*  
**Study 1: Stability of the Big-Fish-Little-Pond Effect: Effects of School-Average Achievement,  
School Type, and Stability Over Time**

Variables	Model 1: School-Average Achievement			Model 2: School Type			Model 3: School-Average Achievement and School Type		
	Model 1A	Model 1B	Model 1C	Model 2A	Model 2B	Model 2C	Model 3A	Model 3B	Model 3C
	T1MSC <i>b</i> ( <i>SE</i> )	T2MSC <i>b</i> ( <i>SE</i> )	$\Delta$ T2MSC <i>b</i> ( <i>SE</i> )	T1MSC <i>b</i> ( <i>SE</i> )	T2MSC <i>b</i> ( <i>SE</i> )	$\Delta$ T2MSC <i>b</i> ( <i>SE</i> )	T1MSC <i>b</i> ( <i>SE</i> )	T2MSC <i>b</i> ( <i>SE</i> )	$\Delta$ T2MSC <i>b</i> ( <i>SE</i> )
Fixed effects									
Level 1: Individual student predictors									
Math test	.68 (.02)*	.63 (.02)*	.15 (.02)*	.65 (.01)*	.60 (.02)*	.14 (.02)*	.68 (.04)*	.64 (.02)*	.15 (.02)*
T1 math self-concept			.71 (.02)*			.71 (.02)*			.71 (.04)
Level 2: School-level predictors									
School-average math test	-.39 (.04)*	-.34 (.04)*	-.07 (.03)*				-.25 (.04)*	-.23 (.05)*	-.05 (.03)
School type				-.20 (.02)*	-.17 (.02)*	-.03 (.02)	-.12 (.02)*	-.10 (.02)*	-.01 (.02)
Residual variance components									
Level 2 school	.02 (.007)*	.01 (.006)	.00 (.003)	.01 (.006)*	.01 (.006)	.00 (.003)	.01 (.005)	.00 (.005)	.00 (.003)
Level 1 students	.63 (.019)*	.68 (.021)*	.36 (.013)*	.63 (.019)*	.68 (.021)*	.36 (.011)*	.63 (.019)*	.68 (.021)*	.36 (.011)*
Deviance summary	5,530.0	5,673.2	4,176.0	5,530.7	5,679.0	4,178.0	5,505.2	5,656.3	4,175.6

(continued)

*Note.* T1MSC = Time 1 math self-concept; T2MSC = Time 2 math self-concept ;  $\Delta$ T2MSC = Time 2 math self-concept controlling for T1MSC; school-average math test = school average of math achievement test scores; school type: 1 = selective Gymnasium, 0 = other. All parameter estimates are statistically significant when they differ from zero by more than 2 standard errors. All outcome and predictor variables were standardized ( $M = 0$ ,  $SD = 1$ ) at the individual student level so that parameter estimates are standardized in relation the mean and standard deviation of individual-level variables. Analyses are based on responses by 2,306 students who completed the math self-concept instrument at Time 2.

\*  $p < .01$ .