# *Faktor- és főkomponens analízis*

Informatikai Tudományok Doktori Iskola

## *Adatredukció*

**Olyan statisztikai módszerek tartoznak ide, melyek lehetővé teszik, hogy az adatmátrix méretét csökkentve kisebb költséggel értékelhessük ki a statisztikai sokaságot. A redukált adatmennyiségből levont statisztikai következtetések érvényesek maradnak az eredeti statisztikai sokaságra is. A csökkentés vonatkozhat az esetszám csökkentésére és a változók számának a csökkentésére egyaránt.**

- **Klaszteranalízis**
- **Ritkítás véletlenszám generálással**
- **Faktoranalízis**
- **Főkomponens-analízis**
- **Többdimenziós skálázás (MDS)**



## *A faktoranalízis problematikája*

**Nagyszámú, sztochasztikusan erősen összefüggő változónk van. A változók redundáns információt hordoznak. Ismeretlen, kisszámú faktorváltozót keresünk.**

- **Hogyan lehet a változók által közösen magyarázott információt korrelálatlan faktorokkal kifejezni?**
- **A faktorok milyen mértékben magyarázzák az eredeti változókat?**
- **Mely változók vannak ugyanazokkal a faktorokkal kifejezve?**
- **Hogyan lehet ezek alapján a változóinkat csoportosítani?**
- **Mi lehet az egyes faktorok jelentése?**

## *A faktoranalízis problematikája*

· **A változók számának csökkentése, de úgy, hogy ezáltal a megfigyelésekben rejlő információ ne csökkenjen lényegesen; lényegkiemelés.**

· **Nehezen megadható fogalmak (pl. gazdasági fejlettség) definiálása összetett mutatórendszerrel való jellemzés útján.**

· **Osztályozási (csoportosítási) feladatok: a csoportképző ismérvnek kijelölt változók nem függetlenek és nem azonos szórásúak, ezért nem lehet azonos súllyal venni figyelembe őket – a változókat kialakító közös faktorok alapján csoportosítunk.**



## *A módszerek jellemzői*

· **Számolásigényesek, számítógépes programcsomagok segítségével hajthatók végre.**

· **Többdimenziós normális eloszlású megfigyelések esetén optimumtulajdonságokkal rendelkeznek, de bármely más, véges szórású mintaeloszlás esetén is igaz, hogy természetesen definiált célfüggvényeket optimalizálnak.**

· **A klasszikus módszerek nem robusztusak, érzékenyek a kiugró és extrém értékekre, de léteznek nemparaméteres, robusztusabb változatok is, amelyek rangstatisztikákkal dolgoznak.**



## *A módszerek jellemzői*

**A módszert olyan esetekben lehet alkalmazni, amikor a sokaságot nagyszámú változóval jellemezzük, és feltételezhetően a változóink egymást átfedő (koherens) információt hordoznak. Az elemzés egyik célja éppen az, hogy a közös információt egymástól korrelálatlan faktorokkal jellemezzük.**
**A faktoranalízis módszere alapvetően abban különbözik a regresszió módszerétől, hogy a prediktor változók a vizsgálat megkezdődésekor nem ismertek, azok előállítása és értelmezése a feladat.**
**Csak akkor van esély jó faktorelemzésre, ha a vizsgálatba bevont változók között erős összefüggés van.**

## A VÁLTOZÓK KÖZÖTTI ÖSSZEFÜGGÉS EREJÉNEK MÉRÉSE

$$\mathrm{KMO} = \frac{\sum_{i=1}^{p}\sum_{\substack{j=1\\ i\neq j}}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{\substack{j=1\\ i\neq j}}^{p} \rho_{ij}^2 + \sum_{i=1}^{p}\sum_{\substack{j=1\\ i\neq j}}^{p} r_{ij}^2},$$

**Kaiser-Meyer-Olkin mérték**

$$\rho_{ij} = \frac{R_{ij}}{\sqrt{R_{ii}\cdot R_{jj}}}$$

**parciális korrelációs együttható**

$$r_{ij} = R(X_i, X_j)$$

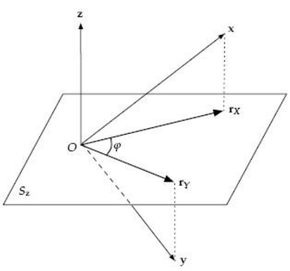
**korrelációs együttható**

| | |
|---|---|
| $0.9 \le \mathrm{KMO}$ | csodálatos (marvelous) |
| $0.8 \le \mathrm{KMO} < 0.9$ | dicséretes (meritorious) |
| $0.7 \le \mathrm{KMO} < 0.8$ | közepes middling) |
| $0.6 \le \mathrm{KMO} < 0.7$ | mérsékelt (mediocre) |
| $0.5 \le \mathrm{KMO} < 0.6$ | szánalmas (miserable) |
| $\mathrm{KMO} < 0.5$ | elfogadhatatlan (unacceptable) |



## Parciális korrelációs együttható

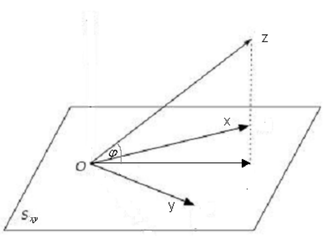
**Vesszük *x* és *y* *z*-re merőleges komponenseit ($r_x$-et és $r_y$-t), és ezek totális korrelációját vesszük. *x* és *y* parciális korrelációs együtthatója annak a lineáris kapcsolatnak az erősséget fejezi ki, ami nem magyarázható *z*-vel.**






## Többszörös korrelációs együttható

**A többszörös korrelációs együttható *z* korrelációs együtthatója az *x*,*y*-ra vett lineáris regressziójával. Ez a többszörös korrelációs együttható a maximális korreláció, amely a *z* változó és a többi változó tetszőleges lineáris kombinációja között előfordul.**



## A VÁLTOZÓK KÖZÖTTI ÖSSZEFÜGGÉS EREJÉNEK MÉRÉSE

$$\mathrm{MSA}_i = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} \rho_{ij}^2 + \sum_{i \neq j} r_{ij}^2} \qquad (i = 1,2,\ldots,p)$$

***measure of sampling adequacy***

**Az indulási *p* db változóból azokat érdemes elhagyni, amelyeknél az $MSA_i$ érték a legkisebb.**

**Elvégezhető még a Bartlett-féle gömb próba.
Itt az a nullhipotézis, hogy a vizsgált változók függetlenek egymástól. Akkor érdemes továbbmenni, ha ez a próba nem szignifikáns!**



## Bartlett-féle gömb próba

- **Azt a nullhipotézist teszteli, hogy a változóink korrelációs mátrixa egységmátrix-e. Ebben az esetben a változók páronként korrelálatlanok lennének, vagyis a változók nem hordoznának redundáns információt.**

- **A nullhipotézist akkor vetjük el, ha a próbastatisztika számított értéke nagy, azaz a próba szignifikancia-szintje nullához közeli érték.**

- **Amennyiben a próba szignifikáns, nincs értelme belefogni a faktorelemzésbe.**



## A k-FAKTOROS MODELL

**Adottak az** $X_1, X_2, \ldots, X_p$ **változók, a belőlük alkotott *p*-dimenziós vektor** $\underline{X}$ **.**

$$\underline{X} = \underline{\underline{A}} \cdot \underline{F} + \underline{U} + \underline{m}$$

$\underline{\underline{A}}$ ***pxk*-as átviteli mátrix**

$\underline{F}$ ***k*-dimenziós közös faktor-vektor**

$\underline{U}$ ***p*-dimenziós egyedi faktor-vektor**

$E\underline{X} = \underline{m}$ **várható érték vektor**

## A k-FAKTOROS MODELL FELTÉTELEI

$F_1, F_2, ..., F_k$ **páronként korrelálatlanok** $R(F_i, F_j) = 0$

$E(F_i) = 0, \quad D^2(F_i) = 1$

$U_1, U_2, ..., U_p$ **páronként korrelálatlanok** $R(U_i, U_j) = 0$

$E(U_i) = 0, \quad D^2(U_i) = \Psi_{ii}$

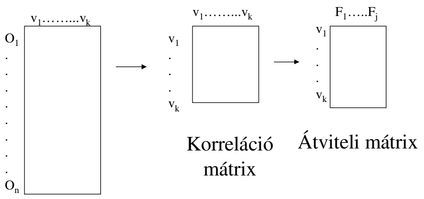
$F_1, F_2, ..., F_k$ **és** $U_1, U_2, ..., U_p$ **páronként korrelálatlanok:**

$$R(U_i, U_j) = 0$$



## Adat mátrix → átviteli mátrix

- A faktoranalízist teljesen behatárolja a változók korrelációs mátrixának felépítése.
- A faktoranalízis ténylegesen a korrelációs struktúrát tárja fel, írja le.




Adatmátrix



## Identifikáció

**Egy *k*-faktoros modell pontosan akkor oldható meg, ha**

$$\underline{\underline{\Sigma}} = \underline{\underline{A}} \cdot \underline{\underline{A}}^T + \underline{\underline{\Psi}}$$

$\underline{\underline{\Sigma}} = \mathrm{cov}(X_i, X_j)$ $\quad \underline{X}$ **kovarianciamátrixa**

$\underline{\underline{\Psi}}$ $\quad \underline{U}$

**Van *p(p+1)/2* egyenlet, és *p***

***(p+1)/2 > k+1*** **esetben az egy**

***(p+1)/2 < k+1*** **esetben az**

- Maximum likelyhoo
- Főkomponens-analízis
- A legkisebb négyzetek módszere

**:**

Aluldefiniált esetben különböző módon adhatunk meg kényszerfeltételeket, amelyek más-más eredményhez (átviteli mátrixhoz) vezethetnek! Ezek közül a legjobban magyarázható megoldást fogjuk választani.

## A k-FAKTOROS MODELL KOORDINÁNTÁNKÉNT

$$X_i = \sum_{j=1}^{k} a_{ij} F_j + U_i + m_i$$

**$\underline{F}$ koordinátái mindegyik $X_i$ előállításában szerepelnek**

**$\underline{U}$ koordinátái közül csak $U_i$ szerepel $X_i$ előállításában**

$$\underbrace{D^2 X_i}_{\text{Az } X_i \text{ varianciája}} = \underbrace{\sum_{j=1}^{k} a^2_{ij}}_{\text{A kummuláns}} + \underbrace{\Psi_{ii}}_{\text{Az egyedi variancia}}$$

$$\frac{\sum_{j=1}^{k} a^2_{ij}}{D^2 X_i}$$ **Ez az arány azt fejezi ki, hány %-ot magyaráznak a közös faktorok.**



## A k-FAKTOROS MODELL KOORDINÁNTÁNKÉNT

**Az átviteli mátrix együtthatóinak jelentése:**

$$\mathbf{cov}(X_i, F_j) = a_{ij}$$

$$R(X_i, F_j) = \frac{a_{ij}}{DX_i}$$

**A kommunalitás a változók varianciájának az a része, amit a közös faktorok magyaráznak:**

$$D^2 X_i = \sum_{j=1}^{k} a_{ij}^2 + D^2 U_i$$



## A FAKTOROK FORGATÁSA (ROTÁCIÓ)

**Az átviteli mátrixnak Lawley módszerével történő egyértelművé tétele a becslési eljárások matematikai elemzését segíti, de az az ára, hogy a kapott közös faktorok gyakran csak nehezen értelmezhetőek. Alkalmas elforgatással esetleg szemléletesebb jelentést adhatunk a faktoroknak. Ha például a faktorsúlyok között csak 0-hoz közeli vagy aránylag nagy értékek fordulnak elő, akkor a változók csoportosíthatók annak alapján hogy melyik faktor mely változókban játszik fontos szerepet, szerencsés esetben a változók halmaza akár diszjunkt osztályokra is bontható.**

## *A FAKTOROK FORGATÁSA (ROTÁCIÓ)*

$$\underline{\underline{\Sigma}} = \underline{\underline{A}} \cdot \underline{\underline{A}}^T + \underline{\underline{\Psi}} = \underline{\underline{A}}\,\underline{\underline{E}}\,\underline{\underline{A}}^T + \underline{\underline{\Psi}} = \underline{\underline{A}}\,\underline{\underline{G}}\,\underline{\underline{G}}^T \underline{\underline{A}}^T + \underline{\underline{\Psi}} = \underline{\underline{A}}^* \cdot \underline{\underline{A}}^{*T} + \underline{\underline{\Psi}}$$

$\underline{\underline{A}}^* = \underline{\underline{A}}\,\underline{\underline{G}}$ **Az új átviteli mátrix**

$\underline{F}^* = \underline{\underline{G}}\,\underline{F}$ **Az új faktorvektor**

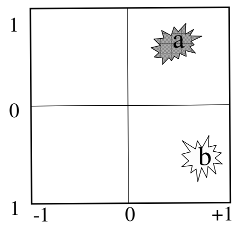
**$\underline{\underline{G}}$ ügyes megválasztásával a modell jobban magyarázható lesz!**

- **Varimax** – **azon változók száma kevés lesz, melyekhez sok faktor szerepel nagy súllyal**
- **Quartimax** – **a magyarázó faktorok számát minimalizálja**
- **Equamax** – **a két eljárás keverékét végzi**



## *A FAKTOROK FORGATÁSA (ROTÁCIÓ)*

**A rotáció szemléltetése egy egyszerű kétdimenziós példán:**




Az eredeti változók **a** és **b** csoportja a rotáció nélkül kapott mindkét faktoron jelentős faktorsúllyal rendelkezik.



## *A FAKTOROK FORGATÁSA (ROTÁCIÓ)*

**A rotáció szemléltetése egy egyszerű kétdimenziós példán:**


**Az eredeti változók a csoportja csak a rotációval kapott egyik, a b csoport pedig csak a másik faktoron rendelkezik jelentős faktorsúllyal.**

## Varimax-rotáció

**Célja, hogy minél több 0-hoz közeli faktorsúlyt állítson elő**

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left(b_{ij}-\overline{b}_j\right)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n} b_{ij}^2 - n\sum_{j=1}^{n}\overline{b}_j^2 \longrightarrow \max$$

**ahol** $b_{ij} = \dfrac{a_{ij}^2}{h_i^2}, \quad \overline{b}_j = \dfrac{1}{n}\sum_{i=1}^{n} b_{ij}$

$A = (a_{ij})$ **a keresett átviteli mátrix**

$h_i^2$ **az *i*-edik változó kommunalitása**



## Varimax-rotáció

**Mivel a forgatások az átviteli mátrix sorainak normáit nem változtatják meg, az egész eljárás során a kommunalitások változatlanok maradnak.**
**A fokozatos maximalizálás úgy történik, hogy minden lépésben csak egy-egy faktorpárt forgatunk el:**
**ha a $(j, m)$ párt választjuk $(1 \le j < m \le k)$, akkor csak az $a_{ij}$ és $a_{im}$ faktorsúlyok változnak:**

$$a'_{ij} = a_{ij}\cos\varphi - a_{im}\sin\varphi, \quad a'_{im} = a_{ij}\sin\varphi + a_{im}\cos\varphi$$

**ahol φ az elforgatás szöge; ez lépésenként csak egyváltozós szélsőérték-feladat megoldását jelenti. Minden ciklusban végighaladunk minden páron (összesen $k(k-1)/2$ pár van), és a ciklus végén ellenőrizzük a célfüggvény változását. Akkor állunk meg, ha már csak elhanyagolható mértékben változik.**



## Főkomponensanalízis

**A faktoranalízis speciális esete. Dimenziószám csökkentésre használható. Az eredetileg *p* változóval jellemzett statisztikai sokaságot $k \ll p$ változóval (főkomponensekkel) jellemezzük. A *k*-dimenziós statisztikai elemzések következtetései a *p*-dimenziós sokaságra is érvényesek lesznek. Ezzel jelentős költséget lehet megtakarítani. Lehetőség van a $p>3$ dimenziós sokaságot (ha $k<4$) pontfelhő grafikonon szemléltetni.**
**A főkomponensek terében a változók korrelálatlanok lesznek.**

**A főkomponens-transzformáció:**

$$\underline{F} = \underline{\underline{G}}^T(\underline{X} - \underline{m}) \quad \Rightarrow \quad \underline{X} = \underline{\underline{G}}\cdot\underline{F} + \underline{m}$$

$\underline{F} = (F_1, F_2, \ldots, F_p)$ **a főkomponens-vektor**

$\underline{\underline{G}} = (\underline{g}_1, \underline{g}_2, \ldots, \underline{g}_p)$ **a főirányok mátrixa**

## A FŐKOMPONENS-MODELL TULAJDONSÁGAI

• A főkomponensek korrelálatlanok: $R(F_i, F_j) = 0$

• A

$E(F_i$ … $0$

**A főkomponensek csökkenő jelentőségűek:**

**$F_1$ magyaráz a legtöbbet, $F_2$ a második legtöbbet,..., $F_p$ magyaráz a legkevesebbet $T$-ből.**

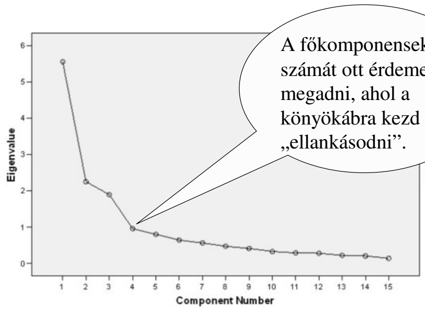
• A

$T = \sum_{i=1} \ldots \sum_{i=1} \ldots$

$\frac{\lambda_i}{T}$ **megmutatja, hány %-ot magyaráz $F_i$**



## *Könyökábra (scree- plot)*




A főkomponensek számát ott érdemes megadni, ahol a könyökábra kezd „ellankásodni”.



## A FŐKOMPONENS-MODELL TULAJDONSÁGAI

• **A főirányok jelentése:** $\underline{\underline{\Sigma}}\,\underline{g}_i = \lambda_i \underline{g}_i$

$\underline{g}_1$ **ebben az irányban a legnagyobb a variancia**

$\underline{g}_2$ **ebben az irányban a legnagyobb a variancia a $g_1$ irányra merőleges irányok között**

$\vdots$

• **Dimenziócsökkentés:**

**Ha $\underline{X}$ helyett az első k főfaktor-alkotta vektorral számolunk, az elvesztett információ csupán:** $1 - \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{T}$

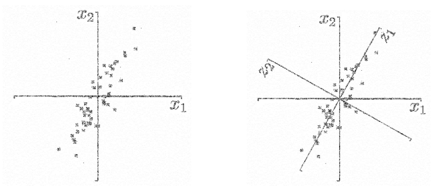
## Watanabe-tétele

**Belátható, hogyha *p* dimenziót lecsökkentünk *k*<*p* dimenzióra, akkor az összes lehetséges dimenziócsökkentési eljárással összevetve, a főkomponens analízissel végrehajtott dimenziócsökkentés minimalizálja az információ-veszteséget! Az eredeti változók totális varianciája és a *k* főfaktor totális varianciája van egymáshoz a legközelebb! Ezt az optimális arányt fejezi ki a kovariancia-mátrix sajátértékeiből számítható arány, amely jó esetben közel esik 1-hez:**

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$



## 2D FŐKOMPONENSANALÍZIS

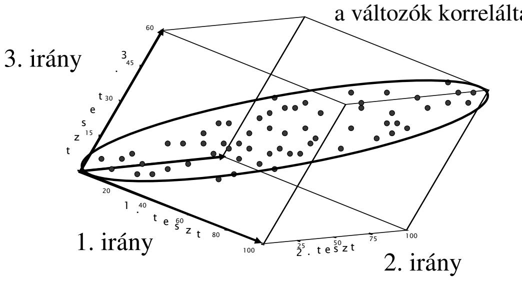



- Az első főkomponens irány ($z_1$) azon egyenes iránya az $X$ síkon, amely körül a legkisebb a pontok szóródása.
- A második főkomponens irány ($z_2$) erre merőleges.



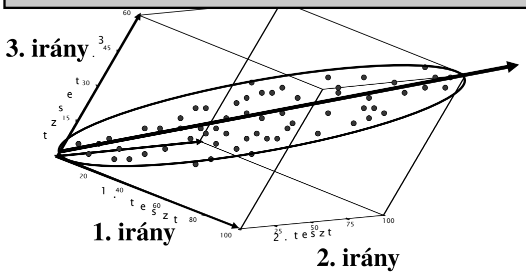
## 3D FŐKOMPONENSANALÍZIS

Tengelyek nem derékszögeket zárnak be: a változók korreláltak!



## *3D FŐKOMPONENSANALÍZIS*

**Ebben az irányban tudunk legjobban differenciálni a pontok között. A főkomponensek hosszát (fontosságát) az ún. sajátértékkel (*eigenvalue*) jellemezzük, ami az értelmezett variancia.**

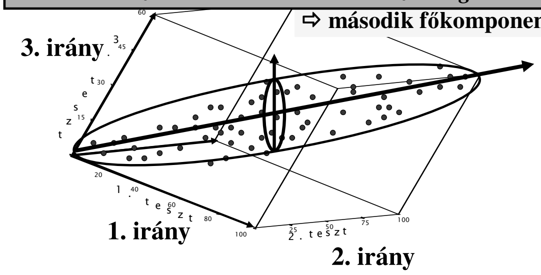





## *3D FŐKOMPONENSANALÍZIS*

Az eljárást folytatni lehetne a harmadik főkomponens megkeresésével, de ennek a konkrét esetben már nincs értelme, mivel ebben az irányban már jelentéktelen a szóródás ⇨ ***az adatok leírására 2 dimenzió elegendő!***

e
a
**⇨ második főkomponens**

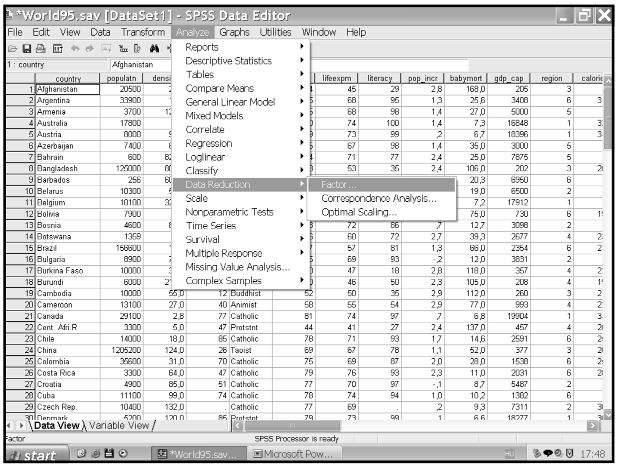


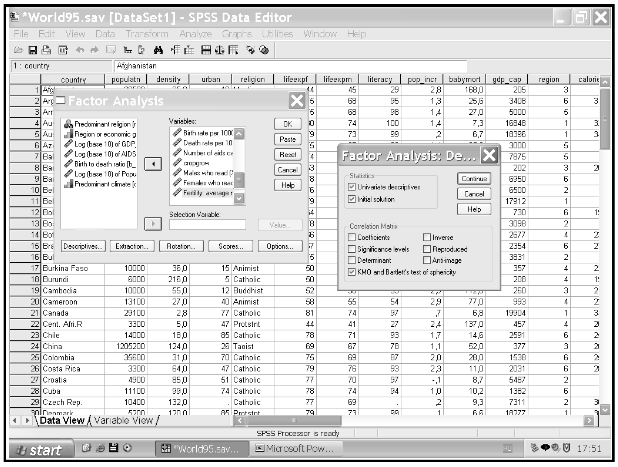


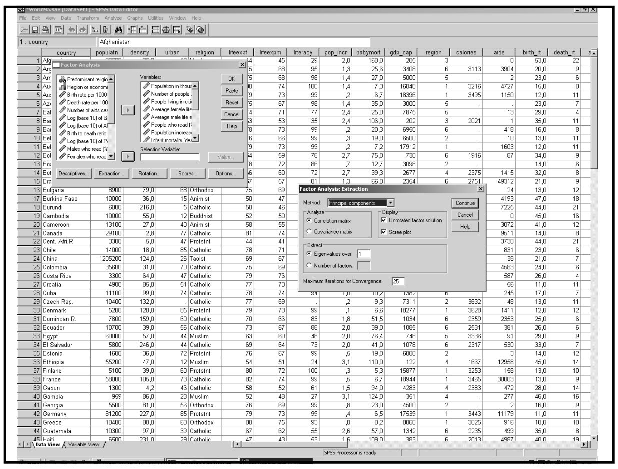

## *Példa a faktoranalízisre I.*

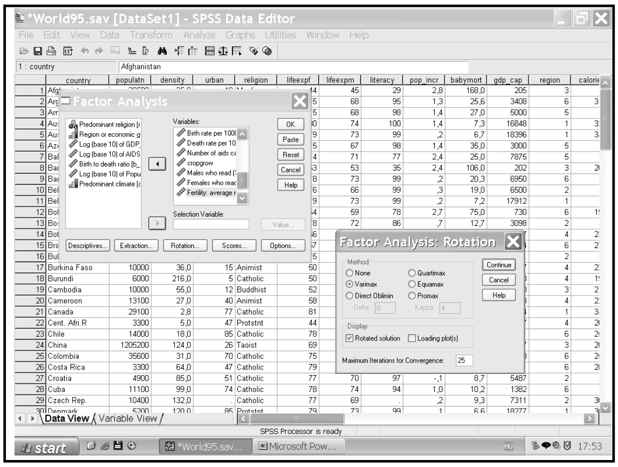
**Megvizsgáljuk, milyen kapcsolat van a *world 95* állomány változói között!**

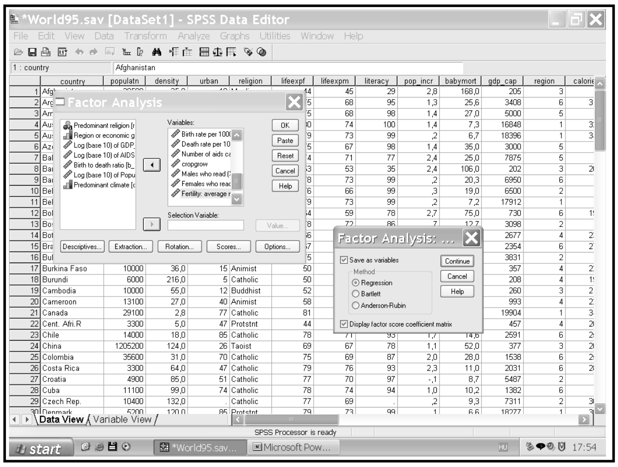
| | Name | Type | Width | Decimals | Label |
|---|---|---|---|---|---|
| 1 | country | String | 12 | 0 | |
| 2 | populatn | Numeric | 8 | 0 | Population in thousands |
| 3 | density | Numeric | 8 | 1 | Number of people / sq. kilometer |
| 4 | urban | Numeric | 5 | 0 | People living in cities (%) |
| 5 | religion | String | 8 | 0 | Predominant religion |
| 6 | lifeexpf | Numeric | 4 | 0 | Average female life expectancy |
| 7 | lifeexpm | Numeric | 5 | 0 | Average male life expectancy |
| 8 | literacy | Numeric | 4 | 0 | People who read (%) |
| 9 | pop_incr | Numeric | 5 | 1 | Population increase (% per year)) |
| 10 | babymort | Numeric | 6 | 1 | Infant mortality (deaths per 1000 live births) |
| 11 | gdp_cap | Numeric | 6 | 0 | Gross domestic product / capita |
| 12 | region | Numeric | 12 | 0 | Region or economic group |
| 13 | calories | Numeric | 6 | 0 | Daily calorie intake |
| 14 | aids | Numeric | 8 | 0 | Aids cases |
| 15 | birth_rt | Numeric | 5 | 1 | Birth rate per 1000 people |
| 16 | death_rt | Numeric | 6 | 0 | Death rate per 1000 people |
| 17 | aids_rt | Numeric | 8 | 2 | Number of aids cases / 100000 people |
| 18 | log_gdp | Numeric | 8 | 2 | Log (base 10) of GDP_CAP |
| 19 | lg_aidsr | Numeric | 8 | 2 | Log (base 10) of AIDS_RT |
| 20 | b_to_d | Numeric | 8 | 2 | Birth to death ratio |
| 21 | fertilty | Numeric | 6 | 1 | Fertility: average number of kids |
| 22 | log_pop | Numeric | 8 | 2 | Log (base 10) of Population |
| 23 | cropgrow | Numeric | 4 | 0 | |
| 24 | lit_male | Numeric | 4 | 0 | Males who read (%) |
| 25 | lit_fema | Numeric | 4 | 0 | Females who read (%) |
| 26 | climate | Numeric | 3 | 0 | Predominant climate |


*World95.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Factor...
Correspondence Analysis...
Optimal Scaling...
Data View
Variable View
SPSS Processor is ready


*World95.sav [DataSet1] - SPSS Data Editor
Factor Analysis
Variables:
Selection Variable:
OK
Paste
Reset
Cancel
Help
Descriptives...
Extraction...
Rotation...
Scores...
Options...
Factor Analysis: De...
Statistics
Univariate descriptives
Initial solution
Continue
Cancel
Help
Correlation Matrix
Coefficients
Inverse
Significance levels
Reproduced
Determinant
Anti-image
KMO and Bartlett's test of sphericity
Data View
Variable View
SPSS Processor is ready






# Factor Analysis

**Descriptive Statistics**

| | Mean | Std. Deviation[a] | Analysis N[a] | Missing N |
|---|---|---|---|---|
| Population in thousands | 47723,88 | 146726,364 | 109 | 0 |
| Number of people / sq. kilometer | 203,415 | 675,7052 | 109 | 0 |
| People living in cities (%) | 56,53 | 24,091 | 109 | 1 |
| Average female life expectancy | 70,16 | 10,572 | 109 | 0 |
| Average male life expectancy | 64,92 | 9,273 | 109 | 0 |
| People who read (%) | 78,34 | 22,670 | 109 | 2 |
| Population increase (% per year)) | 1,682 | 1,1976 | 109 | 0 |
| Gross domestic product / capita | 5859,98 | 6479,836 | 109 | 0 |
| Infant mortality (deaths per 1000 live births) | 42,313 | 38,0792 | 109 | 0 |
| Daily calorie intake | 2753,83 | 470,025 | 109 | 34 |
| Aids cases | 7914,26 | 40121,542 | 109 | 3 |
| Birth rate per 1000 people | 25,923 | 12,3609 | 109 | 0 |
| Death rate per 1000 people | 9,56 | 4,233 | 109 | 1 |
| Number of aids cases / 100000 people | 24,3794 | 48,76326 | 109 | 3 |
| cropgrow | 17,98 | 15,594 | 109 | 3 |
| Males who read (%) | 78,73 | 18,031 | 109 | 24 |
| Females who read (%) | 67,26 | 25,229 | 109 | 24 |
| Fertility: average number of kids | 3,563 | 1,8848 | 109 | 2 |

a. For each variable, missing values are replaced with the variable mean.

**KMO and Bartlett's Test[a]**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,840 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 1357,170 |
| | df | 153 |
| | Sig. | ,000 |

a. Based on correlation

A KMO statisztika „dicséretes" !

A Bartlett-féle függetlenségi teszt is sikertelen!



**Communalities**

| | Initial | Extraction |
|---|---|---|
| Population in thousands | 1,000 | ,450 |
| Number of people / sq. kilometer | 1,000 | ,975 |
| People living in cities (%) | 1,000 | ,751 |
| Average female life expectancy | 1,000 | ,935 |
| Average male life expectancy | 1,000 | ,880 |
| People who read (%) | 1,000 | ,851 |
| Population increase (% per year)) | 1,000 | ,775 |
| Infant mortality (deaths per 1000 live births) | 1,000 | ,916 |
| Gross domestic product / capita | 1,000 | ,705 |
| Daily calorie intake | 1,000 | ,653 |
| Aids cases | 1,000 | ,810 |
| Fertility: average number of kids | 1,000 | ,882 |
| cropgrow | 1,000 | ,711 |

Extraction Method: Principal Component Analysis.

A kummunalitás értékének százszorosa azt mutatja, hogy az egyes változók varianciáit a faktorok hány százalékban tudnak „megmagyarázni". A kis kummunalitású változók „lógnak ki" leginkább a faktortérből. Ha ezeket elhagyjuk, a maradék változókra jobb faktorelemzés adható.

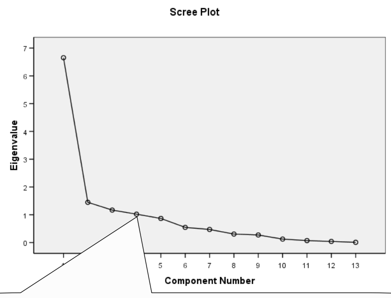


**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 6,654 | 51,186 | 51,186 | 6,654 | 51,186 | 51,186 | 6,515 | 50,115 | 50,115 |
| 2 | 1,448 | 11,138 | 62,324 | 1,448 | 11,138 | 62,324 | 1,511 | 11,625 | 61,740 |
| 3 | 1,169 | 8,991 | 71,316 | 1,169 | 8,991 | 71,316 | 1,184 | 9,109 | 70,849 |
| 4 | 1,022 | 7,860 | 79,176 | 1,022 | 7,860 | 79,176 | 1,083 | 8,327 | 79,176 |
| 5 | ,867 | 6,669 | 85,845 | | | | | | |
| 6 | ,546 | 4,196 | 90,041 | | | | | | |
| 7 | ,471 | 3,625 | 93,666 | | | | | | |
| 8 | ,306 | 2,357 | 96,023 | | | | | | |
| 9 | ,272 | 2,096 | 98,118 | | | | | | |
| 10 | ,125 | ,959 | 99,077 | | | | | | |
| 11 | ,071 | ,547 | 99,624 | | | | | | |
| 12 | ,040 | ,310 | 99,933 | | | | | | |
| 13 | ,009 | ,067 | 100,000 | | | | | | |

Extraction Method: Principal Component Analysis.

Négy faktorral majdnem 80%-os a magyarázhatóság, azaz 13 dimenziót 4-re lecsökkentve, „csak" az információ 20%-át veszítettük el!




Az egyes főkomponensek fontosságának csökkenését mutatja a „könyök-ábra”. Esetünkben az első négy főkomponenst tartottuk meg.



**Component Matrix**[a]

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Population in thousands | -,062 | ,563 | ,331 | ,140 |
| Number of people / sq. kilometer | ,185 | -,112 | -,179 | ,946 |
| People living in cities (%) | ,779 | -,371 | ,024 | ,079 |
| Average female life expectancy | ,956 | -,094 | -,061 | -,094 |
| Average male life expectancy | ,926 | -,125 | -,037 | -,065 |
| People who read (%) | ,898 | ,075 | -,089 | -,177 |
| Population increase (% per year)) | -,726 | -,469 | ,137 | -,096 |
| Infant mortality (deaths per 1000 live births) | -,950 | ,057 | ,070 | ,080 |
| Gross domestic product / capita | ,757 | -,117 | ,304 | ,164 |
| Daily calorie intake | ,765 | -,070 | ,239 | -,075 |
| Aids cases | ,097 | ,147 | ,878 | ,085 |
| Fertility: average number of kids | -,903 | -,237 | ,099 | -,015 |
| cropgrow | ,140 | ,795 | -,244 | ,001 |

Extraction Method: Principal Component Analysis.
a. 4 components extracted.

Ez a táblázat mutatja az átviteli mátrixot. Leolvasható, hogy az egyes változók előállításában a faktorok milyen súlyokkal vesznek részt.



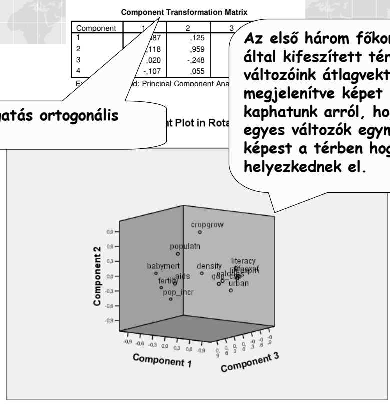
**Rotated Component Matrix**[a]

| | Component | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Population in thousands | -,136 | ,458 | ,470 | ,023 |
| Number of people / sq. kilometer | ,090 | ,012 | -,037 | ,982 |
| People living in cities (%) | ,804 | -,259 | -,031 | ,189 |
| Average female life expectancy | ,964 | ,039 | -,071 | ,019 |
| Average male life expectancy | ,935 | ,002 | -,051 | ,044 |
| People who read (%) | ,894 | ,197 | -,072 | -,081 |
| Population increase (% per year)) | -,648 | -,580 | -,014 | -,134 |
| Infant mortality (deaths per 1000 live births) | -,951 | -,077 | ,068 | -,029 |
| Gross domestic product / capita | ,749 | -,084 | ,310 | ,202 |
| Daily calorie intake | ,776 | -,035 | ,220 | -,027 |
| Aids cases | ,087 | -,060 | ,892 | -,051 |
| Fertility: average number of kids | -,860 | -,365 | ,013 | -,091 |
| cropgrow | ,039 | ,840 | -,043 | -,034 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Az elforgatás után a jobban értelmezhető modellt kapunk. Ez segít a faktorok értelmezésében, és a változók kapcsolat-rendszerének feltárásában egyaránt.

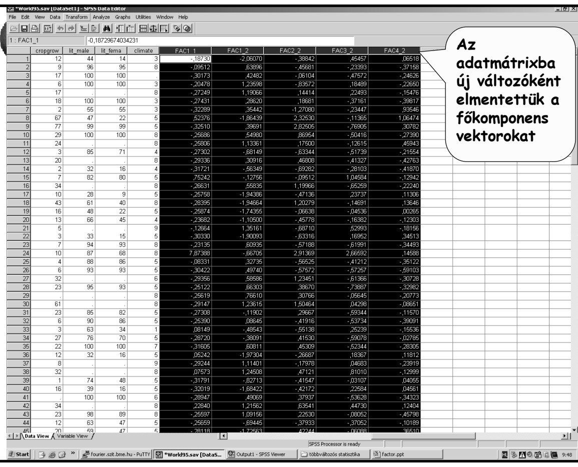
A varimax elforgatás után keletkezett átviteli-mátrix táblázata.



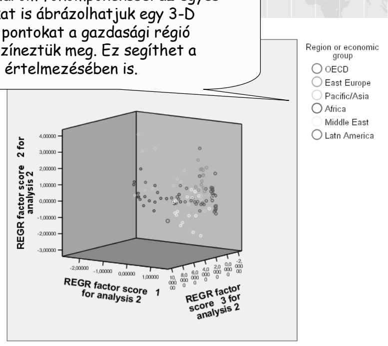

Az elforgatás ortogonális mátrixa

Az első három főkomponens által kifeszített térben a változóink átlagvektorait megjelenítve képet kaphatunk arról, hogy az egyes változók egymáshoz képest a térben hogyan helyezkednek el.





Az adatmátrixba új változóként elmentettük a főkomponens vektorokat



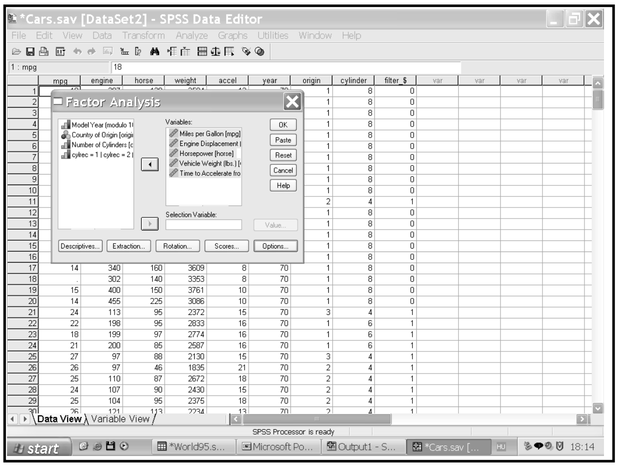

Az első három főkomponenssel az egyes országokat is ábrázolhatjuk egy 3-D ábrán. A pontokat a gazdasági régió szerint színeztük meg. Ez segíthet a faktorok értelmezésében is.

## Példa a faktoranalízisre II.

**Milyen kapcsolat van a gépkocsik jellemzői között?**

| | Name | Type | Widt | Decim | Label |
|---|---|---|---|---|---|
| 1 | mpg | Numeric | 4 | 0 | Miles per Gallon |
| 2 | engine | Numeric | 5 | 0 | Engine Displacement (cu. inches) |
| 3 | horse | Numeric | 5 | 0 | Horsepower |
| 4 | weight | Numeric | 4 | 0 | Vehicle Weight (lbs.) |
| 5 | accel | Numeric | 4 | 0 | Time to Accelerate from 0 to 60 mph (sec) |
| 6 | year | Numeric | 2 | 0 | Model Year (modulo 100) |
| 7 | origin | Numeric | 1 | 0 | Country of Origin |
| 8 | cylinder | Numeric | 1 | 0 | Number of Cylinders |
| 9 | filter_$ | Numeric | 1 | 0 | cylrec = 1 \| cylrec = 2 (FILTER) |






## Factor Analysis

**Warnings**

Only one component was extracted. Component plots cannot be produced.

**Descriptive S**

| | Mean |
|---|---|
| Miles per Gallon | 23 |
| Engine Displacement (cu. inches) | 194 |
| Horsepower | 104 |
| Vehicle Weight (lbs.) | 2969 |
| Time to Accelerate from 0 to 60 mph (sec) | 15,50 |

a. For each variable, missing values are replaced with the variable mean.

**Kaptunk egy figyelmeztetést, hogy csak egyetlen dimenziót tartottunk meg a beállításkor, tehát a 2-D és 3-D ábrák nem készülhetnek el.**

**KMO and Bartlett's Test**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,796 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 2215,912 |
| | df | 10 |
| | | ,000 |

**Csupán közepes a változók közötti összefüggés, viszont a Bartlett-féle függetlenségi próba sikertelen volt, azaz erős az összefüggés a változók között.**

| | | Extraction |
|---|---|---|
| | | ,743 |
| Engine Displacement (cu. inches) | 1,000 | ,913 |
| Horsepower | 1,000 | ,917 |
| Vehicle Weight (lbs.) | 1,000 | ,862 |
| Time to Accelerate from 0 to 60 mph (sec) | 1,000 | ,451 |

Extraction Method: Principal Component Analysis.

**A gyorsulás eléggé kilóg a faktortérből, csupán 45%-os a magyarázhatóság esetében.**



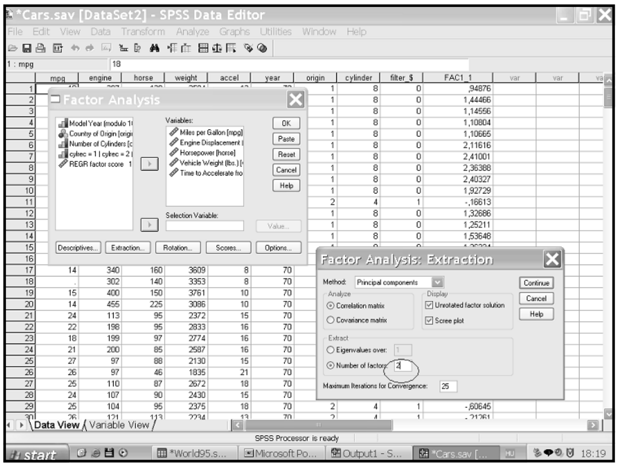
**Total Variance Explained**

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 3,885 | 77,701 | 77,701 | 3,885 | 77,701 | 77,701 |
| 2 | ,708 | 14,167 | 91,868 | | | |
| 3 | ,268 | 5,368 | 97,23 | | | |
| 4 | ,085 | 1,706 | 98,94 | | | |
| 5 | ,053 | 1,058 | 100,00 | | | |

Extraction Method: Principal Component Analysis.

**Az első főkomponens az információ több mint 77%-át magyarázza. Viszont az első két főfaktor már 91% feletti magyarázó erőt képvisel!**



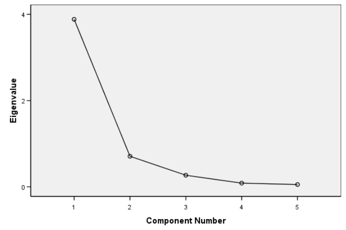


**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 3,885 | 77,701 | 77,701 | 3,885 | 77,701 | 77,701 | 3,136 | 62,723 | 62,723 |
| 2 | ,708 | 14,167 | 91,868 | ,708 | 14,167 | 91,868 | 1,457 | 29,145 | 91,868 |
| 3 | ,268 | 5,368 | 97,236 | | | | | | |
| 4 | ,085 | 1,706 | 98,942 | | | | | | |
| 5 | ,053 | 1,058 | 100,000 | | | | | | |

Extraction Method: Principal Component Analysis.

**Scree Plot**






**Component Matrix**[a]

| | Component 1 | Component 2 |
|---|---|---|
| Miles per Gallon | -,862 | -,257 |
| Engine Displacement (cu. inches) | ,955 | ,120 |
| Horsepower | ,958 | -,112 |
| Vehicle Weight (lbs.) | ,928 | ,283 |
| Time to Accelerate from 0 to 60 mph (sec) | -,672 | ,731 |

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

**Rotated Component Matrix**[a]

| | Component 1 | Component 2 |
|---|---|---|
| Miles per Gallon | -,878 | ,193 |
| Engine Displacement (cu. inches) | ,893 | -,359 |
| Horsepower | ,783 | -,563 |
| Vehicle Weight (lbs.) | ,949 | -,203 |
| Time to Accelerate from 0 to 60 mph (sec) | -,232 | ,966 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

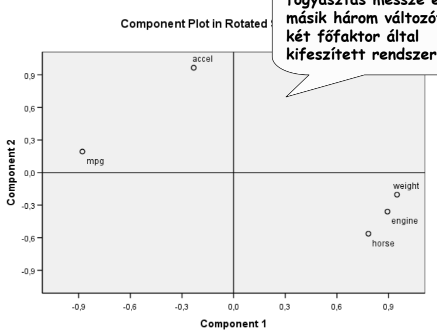
**Component Transformation Matrix**

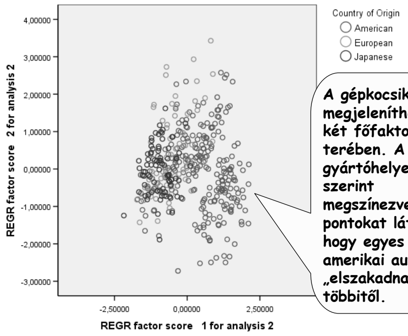
| Component | 1 | 2 |
|---|---|---|
| 1 | ,874 | -,486 |
| 2 | ,486 | ,874 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.



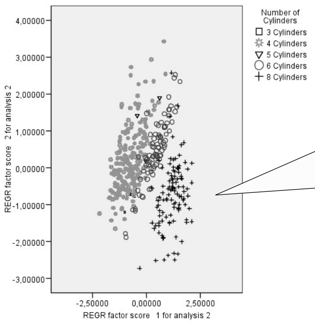
**A gyorsulás és a fogyasztás messze esik a másik három változótól a két főfaktor által kifeszített rendszerben!**






A gépkocsik megjeleníthetők a két főfaktor terében. A gyártóhelyek szerint megszínezve a pontokat látható, hogy egyes amerikai autók „elszakadnak” a többitől.






Ha a hengerek száma szerint is kiszínezzük az ábrát, láthatjuk, hogy az elkülönült csoport az amerikai 8 hengeres autóknak felel meg!



# *Példa a főkomponens analízisre*

**Elemezzük ki a banki ügyfelek 700 esetet tartalmazó állományát főkomponens analízissel! Csoportosítsuk az állomány változóit!**

**Variable Information**

| Variable | Label | Measurement Level |
|---|---|---|
| age | Age in years | Scale |
| ed | Level of education | Ordinal |
| employ | Years with current employer | Scale |
| address | Years at current address | Scale |
| income | Household income in thousands | Scale |
| debtinc | Debt to income ratio (x100) | Scale |
| creddebt | Credit card debt in thousands | Scale |
| othdebt | Other debt in thousands | Scale |
| default | Previously defaulted | Nominal |
| preddef1 | Predicted default, model 1 | Scale |
| preddef2 | Predicted default, model 2 | Scale |
| preddef3 | Predicted default, model 3 | Scale |

**age** (ügyfél életkora), **ed** (ügyfél iskolázottsága), **employ** (ügyfél hány éve van alkalmazásban jelenlegi munkaadójánál), **address** (ügyfél jelenlegi lakcíme), **income** (ügyfél háztartásának évi jövedelme ezer USD-ban), **debtinc** (ügyfél által felvett hitel aránya a jövedelméhez), **creddebt** (ügyfél hitelkártya tartozása ezer USD-ban), **othdebt** (ügyfél egyéb tartozása ezer USD-ban), **default** (ügyfél korábban megtagadta-e már a törlesztést).

## Példa a főkomponens analízisre

*bankloan.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

| | age | ed | employ | address | income | debtinc | creddebt | othdebt | default |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 41 | Some college | 17 | 12 | 176,00 | 9,30 | 11,36 | 5,01 | Yes |
| 2 | 27 | Did not complete high school | 10 | 6 | 31,00 | 17,30 | 1,36 | 4,00 | No |
| 3 | 40 | Did not complete high school | 15 | 14 | 55,00 | 5,50 | ,86 | 2,17 | No |
| 4 | 41 | Did not complete high school | 15 | 14 | 120,00 | 2,90 | 2,66 | ,82 | No |
| 5 | 24 | High school degree | 2 | 0 | 28,00 | 17,30 | 1,79 | 3,06 | Yes |
| 6 | 41 | High school degree | 5 | 5 | 25,00 | 10,20 | ,39 | 2,16 | No |
| 7 | 39 | Did not complete high school | 20 | 9 | 67,00 | 30,60 | 3,83 | 16,67 | No |
| 8 | 43 | Did not complete high school | 12 | 11 | 38,00 | 3,60 | ,13 | 1,24 | No |
| 9 | 24 | Did not complete high school | 3 | 4 | 19,00 | 24,40 | 1,36 | 3,28 | Yes |
| 10 | 36 | Did not complete high school | 0 | 13 | 25,00 | 19,70 | 2,78 | 2,15 | No |
| 11 | 27 | Did not complete high school | 0 | 1 | 16,00 | 1,70 | ,18 | ,09 | No |
| 12 | 25 | Did not complete high school | 4 | 0 | 23,00 | 5,20 | ,25 | ,94 | No |
| 13 | 52 | Did not complete high school | 24 | 14 | 64,00 | 10,00 | 3,93 | 2,47 | No |
| 14 | 37 | Did not complete high school | 6 | 9 | 29,00 | 16,30 | 1,72 | 3,01 | No |
| 15 | 48 | Did not complete high school | 22 | 15 | 100,00 | 9,10 | 3,70 | 5,40 | No |
| 16 | 36 | High school degree | 9 | 6 | 49,00 | 8,60 | ,82 | 3,40 | Yes |
| 17 | 36 | High school degree | 13 | 6 | 41,00 | 16,40 | 2,92 | 3,81 | Yes |
| 18 | 43 | Did not complete high school | 23 | 19 | 72,00 | 7,60 | 1,18 | 4,29 | No |
| 19 | 39 | Did not complete high school | 6 | 9 | 61,00 | 5,70 | ,56 | 2,91 | No |



## Példa a főkomponens analízisre

**Descriptive Statistics**

| | Mean | Std. Deviation | Analysis N |
|---|---|---|---|
| Age in years | 35,03 | 8,041 | 850 |
| Years with current employer | 8,57 | 6,778 | 850 |
| Years at current address | 8,37 | 6,895 | 850 |
| Household income in thousands | 46,6753 | 38,54305 | 850 |
| Debt to income ratio (x100) | 10,1716 | 6,71944 | 850 |
| Credit card debt in thousands | 1,5768 | 2,12584 | 850 |
| Other debt in thousands | 3,0788 | 3,39880 | 850 |

**KMO and Bartlett's Test**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | ,645 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 2988,784 |
| | df | 21 |
| | Sig. | ,000 |



## Példa a főkomponens analízisre

Az MSA$_i$ statisztikák

| | | Age in years | Years with current employer | [illegible] | [illegible] | [illegible] income ratio (x100) | Credit card debt in thousands | Other debt in thousands |
|---|---|---|---|---|---|---|---|---|
| Anti-image Covariance | Age in years | ,488 | -,159 | [illegible] | [illegible] | -,007 | ,012 | -,016 |
| | Years with current employer | -,159 | ,516 | [illegible] | [illegible] | ,048 | -,042 | -,033 |
| | Years at current address | -,287 | [illegible] | [illegible] | [illegible] | [illegible] | -,001 | ,008 |
| | Household income in thousands | -,035 | [illegible] | [illegible] | [illegible] | [illegible] | -,151 | -,161 |
| | Debt to income ratio (x100) | -,007 | [illegible] | [illegible] | [illegible] | [illegible] | -,183 | -,200 |
| | Credit card debt in thousands | ,012 | -,042 | [illegible] | [illegible] | [illegible] | ,406 | ,008 |
| | Other debt in thousands | -,016 | [illegible] | [illegible] | [illegible] | [illegible] | ,008 | ,279 |
| Anti-image Correlation | Age in years | ,743[a] | [illegible] | [illegible] | [illegible] | [illegible] | ,026 | -,043 |
| | Years with current employer | -,316 | ,857[a] | [illegible] | [illegible] | [illegible] | [illegible] | -,086 |
| | Years at current address | -,514 | -,003 | ,705[a] | [illegible] | ,015 | [illegible] | ,020 |
| | Household income in thousands | -,095 | -,245 | -,028 | ,586[a] | [illegible] | [illegible] | -,581 |
| | Debt to income ratio (x100) | -,016 | ,116 | ,015 | ,604 | ,359[a] | -,500 | -,659 |
| | Credit card debt in thousands | ,026 | -,093 | -,003 | -,452 | -,500 | ,727[a] | ,025 |
| | Other debt in thousands | -,043 | -,086 | ,020 | -,581 | -,659 | ,025 | ,646[a] |

a. Measures of Sampling Adequacy(MSA)

## Példa a főkomponens analízisre

**Communalities**

| | Initial | Extraction |
|---|---|---|
| Age in years | 1,000 | ,695 |
| Years with current employer | 1,000 | ,653 |
| Years at current address | 1,000 | ,519 |
| Household income in thousands | 1,000 | ,680 |
| Debt to income ratio (x100) | 1,000 | ,771 |
| Credit card debt in thousands | 1,000 | ,753 |
| Other debt in thousands | 1,000 | ,819 |

Extraction Method: Principal Component Analysis.



## Példa a főkomponens analízisre

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 3,269 | 46,700 | 46,700 | 3,269 | 46,700 | 46,700 | 2,631 | 37,591 | 37,591 |
| 2 | 1,622 | 23,174 | 69,874 | 1,622 | 23,174 | 69,874 | 2,260 | 32,282 | 69,874 |
| 3 | ,856 | 12,236 | 82,109 | | | | | | |
| 4 | ,442 | 6,315 | 88,424 | | | | | | |
| 5 | ,361 | 5,164 | 93,589 | | | | | | |
| 6 | ,325 | 4,643 | 98,232 | | | | | | |
| 7 | ,124 | 1,768 | 100,000 | | | | | | |

Extraction Method: Principal Component Analysis.



## Példa a főkomponens analízisre

**Component Matrix[a]**

| | Component | |
|---|---|---|
| | 1 | 2 |
| Age in years | ,699 | -,453 |
| Years with current employer | ,739 | -,328 |
| Years at current address | ,520 | -,498 |
| Household income in thousands | ,809 | -,161 |
| Debt to income ratio (x100) | ,342 | ,809 |
| Credit card debt in thousands | ,748 | ,440 |
| Other debt in thousands | ,796 | ,432 |

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

**Rotated Component Matrix[a]**

| | Component | |
|---|---|---|
| | 1 | 2 |
| Age in years | ,830 | ,080 |
| Years with current employer | ,782 | ,203 |
| Years at current address | ,718 | -,066 |
| Household income in thousands | ,733 | ,377 |
| Debt to income ratio (x100) | -,236 | ,846 |
| Credit card debt in thousands | ,311 | ,810 |
| Other debt in thousands | ,354 | ,833 |

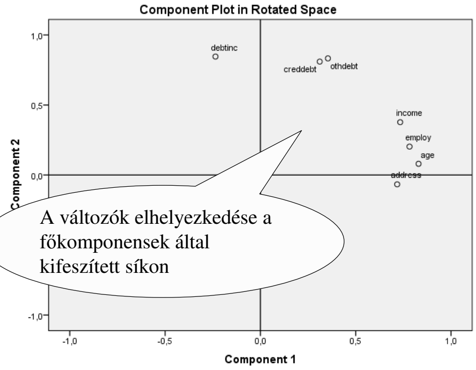
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

**Component Transformation Matrix**

| Component | 1 | 2 |
|---|---|---|
| 1 | ,783 | ,622 |
| 2 | -,622 | ,783 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

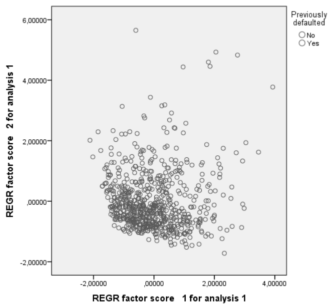
## *Példa a főkomponens analízisre*






## *Példa a főkomponens analízisre*

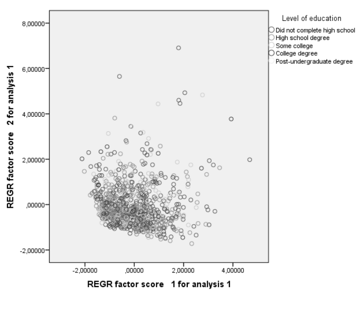
**A főkomponensek kétdimenziós terében ábrázoltuk az ügyfeleket a default (bedőlés) változó színeivel.**






## *Példa a főkomponens analízisre*

**A főkomponensek kétdimenziós terében ábrázoltuk az ügyfeleket a ed (isk. végzettség) változó színeivel.**



# *Példa a főkomponens analízisre*

**A főkomponensek kétdimenziós terében ábrázoltuk az ügyfeleket a korkategória színeivel.**
**1: harminc alatti**
**2: harminc és negyven közötti**
**3: negyven feletti**