Többváltozós statisztika jamovi-ban és R-ben

Abari Kálmán

2023. 03. 23.

Tartalomjegyzék

# Előszó

A statisztika alapfogalmai nagyon jól szemléltethetők az egyváltozós statisztikai eljárásokkal. Ezek az eljárások tipikusan egy (vagy két) változó vizsgálatával járulnak hozzá az empirikus vizsgálatok során felmerülő statisztikai jellegű kérdések megválaszolásához.

A kutatómunka során azonban szükség lehet egyszerre több változó bevonására az elemzésbe, ezeket az eljárásokat többváltozós statisztikai eljárásoknak nevezzük. Ilyen eljárás például:

* Lineáris regresszió ([1](#sec-linearis-regresszio))
* Főkomponens elemzés ([2](#sec-fokomponens-elemzes))
* Megbízhatóság elemzés ([3](#sec-megbizhatosag-elemzes))
* Feltáró faktorelemzés ([4](#sec-feltaro-faktorelemzes))
* Megerősítő faktorelemzés ([5](#sec-megerosito-faktorelemzes))
* Többszempontos varianciaelemzés ([6](#sec-tobbszempontos-varianciaelemzes))
* Klaszterelemzés ([7](#sec-klaszterelemzes))
* Diszkriminancia elemzés ([8](#sec-diszkriminancia-elemzes))
* Többváltozós varianciaelemzés ([9](#sec-tobbvaltozos-varianciaelemzes))
* Logisztikus regresszióelemzés ([10](#sec-logisztikus-regresszio))
* Többdimenziós skálázás ([11](#sec-tobbdimenzios-skalazas))

A jegyzet elkészítéséhez elsősorban a kurzus tankönyvét (Münnich és mtsai., 2006) használtuk fel, de támaszkodtunk egyéb forrásokra is (Csallner, 2015; Ketskeméty és Izsó, 2005; Malhotra és Simon, 2008; Moksony, 2006; Sajtos és Mitev, 2007; Székelyi és Barna, 2002; Takács, 2017; Varga, 2019).

# 1. Lineáris regresszió

A korrelációszámítás két változó szimmetrikus kapcsolatának erősségét és irányát vizsgálja csupán. Mivel az egyszerű lineáris regresszió két változó függvényszerű kapcsolatát vizsgálja, ez már nem szimmetrikus viszony, vagyis megkülönböztetjük a

* függő változót (célváltozót, -t), amely “elszenvedi” a független változó hatását, és a
* független változót (magyarázó változót, -et), amely befolyásolja a függő változót.

A többszörös lineáris regresszió annyiban tér el az egyszerű lineáris regressziótól, hogy a független változók száma egynél több. Itt is megkülönböztetjük a

* függő változót (célváltozót, -t), amelynek értékei a független változóktól függenek, és a
* független változókat (magyarázó változókat, -t) amelyek hatnak a függő változóra.

Két változó ( és ) között nem feltétlenül van szisztematikus kapcsolat, lehet a két változó független is egymástól. Ha van valamilyen szisztematikus kapcsolat és között, akkor az még számos formában megvalósulhat, ezek egyike a lineáris kapcsolat,

* amely olyan függvényszerű kapcsolat, amely megmondja, hogy milyen mértékű változás várható az változóban, ha adott mértéknyit változik.

## 1.1 Egyszerű lineáris regresszió

Az egyszerű lineáris regressziós modell: , amely egy egyenessel (regressziós egyenes) írja le a két változó függvényszerű kapcsolatát, ahol

* – tengelymeszet, a regressziós egyenes itt metszi az y tengelyt
* – meredekség, a regressziós egyenes és az x tengely szögének tangense
* – hibatag, amelyről feltételezzük, hogy normális eloszlású 0 várható értékkel.

A és populációbeli paramétereket a minta alapján becsüljük a legkisebb négyzetek módszere segítségével, így kapjuk a és becsléseket.

A regressziós egyenes birtokában tetszőleges értékhez tudunk értéket előre jelezni, vagyis jósolni bizonyos hibával: .

Például egy fiktív adatbázison vizsgálhatjuk a fizetés és a munkahellyel való elégedettség kapcsolatát (Münnich és mtsai., 2006).

d <- rio::import(file = "adat/lin\_reg\_fizetes\_elegedettseg\_02.xlsx")  
str(d)  
#> 'data.frame': 5 obs. of 2 variables:  
#> $ fizetes : num 44 66 89 155 130  
#> $ elegedettseg: num 30 45 60 100 85  
d  
#> fizetes elegedettseg  
#> 1 44 30  
#> 2 66 45  
#> 3 89 60  
#> 4 155 100  
#> 5 130 85

lm\_1 <- lm(elegedettseg ~ fizetes, data = d)  
summary(lm\_1)  
#>   
#> Call:  
#> lm(formula = elegedettseg ~ fizetes, data = d)  
#>   
#> Residuals:  
#> 1 2 3 4 5   
#> -0.8423 0.3420 0.8983 -0.5488 0.1508   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 3.210890 0.931791 3.446 0.0411 \*   
#> fizetes 0.627987 0.008873 70.774 6.22e-06 \*\*\*  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 0.8077 on 3 degrees of fre...  
#> Multiple R-squared: 0.9994, Adjusted R-squared: 0...  
#> F-statistic: 5009 on 1 and 3 DF, p-value: 6.216e-06

Jamovi-ban a Regression / Linear Regression menüpontot kell használnunk.

|  |
| --- |
| Fizetés és elégedettésg kapcsolata (N=5): együtthatók |

A fenti elemzés alapján például a konkrét formája:

becsült elégedettség = 3,211+ 0,628 \* fizetés

* A értelmezése: a zérus -hez tartozó érték.
* A értelmezése: az egy egységnyi növekedéséhez ilyen nagyságú változás tartozik.

Tudjuk, hogy az Pearson-féle korrelációs együttható, az és változók közötti kapcsolat erősségét és irányát mutatja meg. A és kapcsolatban áll:

* azonos az előjelük,
* az egy szórásnyi növekedéséhez tartozó változás megegyezik az szórásának szeresével (rövidebben, a populációbeli paraméterekkel megfogalmazva:

A determinációs együttható () a korrelációs együttható négyzete , amely szimmetrikus mutató, megmutatja, hogy varianciájának mekkora hányadát magyarázza varianciája, vagy fordítva, varianciájának mekkora hányadát magyarázza varianciája.

A fenti példában látható, hogy 99%-ban lehet a függő változó varianciáját magyarázni a független változóval (az arányt legtöbbször százalékos formában adjuk meg).

summary(lm\_1)$r.squared  
#> [1] 0.9994014

|  |
| --- |
| Fizetés és elégedettésg kapcsolata (N=5): determinációs együttható |

A és együtthatók értékét hipotézisvizsgálatokkal vizsgálhatjuk:

* : , Kérdés: origón átmenő a regresszió? ( megtartása esetén igen)
* : , Kérdés: függ -től? ( elfogadása esetén igen)

A példában látható, hogy nem origón átmenő a regresszió, és az elégedettség függ a fizetéstől.

summary(lm\_1)$coefficients  
#> Estimate Std. Error t value Pr(...  
#> (Intercept) 3.2108896 0.931790852 3.445934 4.10580...  
#> fizetes 0.6279867 0.008873152 70.773796 6.21643...

|  |
| --- |
| Fizetés és elégedettésg kapcsolata (N=5): hipotézisvizsgálat az együtthatókra |

## 1.2 Többszörös lineáris regresszió

A többszörös lineáris regressziós modell: .

Míg az egyszerű lineáris regresszió esetén a regressziós egyenes írta le a két változó kapcsolatát, a többszörös lineáris regresszió esetén a lineáris függvény egy dimenziós sík az dimenziós térben.

Az egyes együtthatók becslése itt is a legkisebb négyzetek elve alapján történik, így kapjuk a becsléseket.

A lineáris függvény birtokában tetszőleges értékekhez tudunk értéket előre jelezni, vagyis jósolni bizonyos hibával: .

d <- rio::import(file = "adat/lin\_reg\_fizetes\_eletkor\_elegedettseg\_01.xlsx")  
str(d)  
#> 'data.frame': 5 obs. of 3 variables:  
#> $ fizetes : num 44 66 89 155 130  
#> $ eletkor : num 25 65 21 35 40  
#> $ elegedettseg: num 37 36 61 92 76  
d  
#> fizetes eletkor elegedettseg  
#> 1 44 25 37  
#> 2 66 65 36  
#> 3 89 21 61  
#> 4 155 35 92  
#> 5 130 40 76

lm\_1 <- lm(elegedettseg ~ fizetes + eletkor, data = d)  
summary(lm\_1)  
#>   
#> Call:  
#> lm(formula = elegedettseg ~ fizetes + eletkor, data...  
#>   
#> Residuals:  
#> 1 2 3 4 5   
#> 0.28596 0.08556 -0.30015 0.71047 -0.78184   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 21.508055 1.292166 16.64 0.00359 \*\*   
#> fizetes 0.519198 0.008847 58.69 0.00029 \*\*\*  
#> eletkor -0.305549 0.023279 -13.12 0.00575 \*\*   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 0.8047 on 2 degrees of fre...  
#> Multiple R-squared: 0.9995, Adjusted R-squared: 0...  
#> F-statistic: 1841 on 2 and 2 DF, p-value: 0.000543

|  |
| --- |
| Elégedettség kapcsolata a fizetéssel és az élekorral (N=5): együtthatók |

A fenti példában a lineáris regresszió futtatása után azt mondhatjuk:

becsült elégedettség = 21,05 -0,306\*életkor + 0,519\*fizetés

Más szavakkal a fizetés tekintetében a magasabb fizetés nagyobb mértékű elégedettséggel jár, addig az életkor esetében az évek számának növekedése a munkahellyel való elégedetlenséget vonja maga után.

* A értelmezése: a csupa zérus -ekhez tartozó érték.
* A értelmezése: az hatása úgy, hogy a többi független változót is figyelembe vesszük.

A fenti többszörös lineáris regressziós együtthatók nem alkalmasak az egyes magyarázó változóktól való függés erősségének mérésére, ugyanis a nagyságuk függ a változó értékeinek nagyságától is. Ezért a standard lineáris regressziós együtthatókat használjuk, amelyek már mértékegység nélküli, egymással összehasonlítható arányszámok, így abszolút értékeiket összevetve megtudhatjuk, milyen relatív fontossággal bírnak az egyes független változók a függő változó magyarázásában.

lsr::standardCoefs(lm\_1)  
#> b beta  
#> fizetes 0.5191980 0.9677518  
#> eletkor -0.3055489 -0.2164358

|  |
| --- |
| Elégedettség kapcsolata a fizetéssel és az élekorral (N=5): standardizált együtthatók |

A fenti példában láthatjuk, hogy a fizetés erősebb kapcsolatban van az elégedettséggel, hiszen a standardizált együtthatójának értéke abszolút értékben nagyobb, mint az életkor standardizált együtthatójának abszolút értéke.

Többszörös lineáris regresszió esetén több hipotézisvizsgálat végezhető:

* minden együtthatót külön tesztelhetünk t-próbákkal szabadsági fokkal
  + , , Kérdés: függ -től? ( elfogadása esetén igen)
* a teljes modellt tesztelhetjük F-próbával szabadsági fokkal
  + , Kérdés: a modell bír valamilyen bejósló erővel? ( elfogadása esetén igen)

summary(lm\_1)$coefficients  
#> Estimate Std. Error t value Pr(...  
#> (Intercept) 21.5080549 1.29216584 16.64496 0.00358...  
#> fizetes 0.5191980 0.00884672 58.68819 0.00029...  
#> eletkor -0.3055489 0.02327903 -13.12550 0.00575...  
summary(lm\_1)$fstatistic  
#> value numdf dendf   
#> 1840.547 2.000 2.000  
pf(q = summary(lm\_1)$fstatistic[1], df1 = summary(lm\_1)$fstatistic[2],  
 df2 = summary(lm\_1)$fstatistic[3], lower.tail = F)  
#> value   
#> 0.0005430217

|  |
| --- |
| Elégedettség kapcsolata a fizetéssel és az élekorral (N=5): hipotézisvizsgálatok |

A lenti példában látható, hogy mindkét magyarázó változótól függ az elégedettség (életkor p-értéke: 0,006, a fizetés p-értéke p < 0,001), és a teljes modell bír magyarázó erővel (p-érték: p < 0,001).

A függő változó és a független változók közötti korreláció erősségének leírására több mennyiséget használhatunk

* többszörös korrelációs együttható: , amely a függő változó és a becsült értékek közötti korrelációs együttható értékével egyezik meg, azaz . Valójában a lineáris regresszió ennek a korrelációs együtthatónak az értékét maximalizálja, mikor az -t -ek speciális lineáris kombinációjaként előállítja.
* többszörös determinációs együttható: , amely a többszörös korrelációs együttható négyzete, és megmutatja, hogy a magyarázó változók a függő változó ingadozásának hányad részét magyarázzák.
* korrigált determinációs együttható: , amely kiküszöböli az azon tulajdonságát, hogy a magyarázó változók számának növekedésével, függetlenül azok hatásától, nő az értéke. Így alkalmas több modell esetén a magyarázó erők összehasonlítására, akkor is, ha azok eltérő számú független változót használnak.

summary(lm\_1)$r.squared  
#> [1] 0.999457  
summary(lm\_1)$adj.r.squared  
#> [1] 0.998914

|  |
| --- |
| Elégedettség kapcsolata a fizetéssel és az élekorral (N=5): magyarázó erő |

A fenti példában látható mindhárom fenti mutató. Az leolvasásával láthatjuk, hogy a két független változó, az életkor és a fizetés a függő változó 99%-át magyarázza.

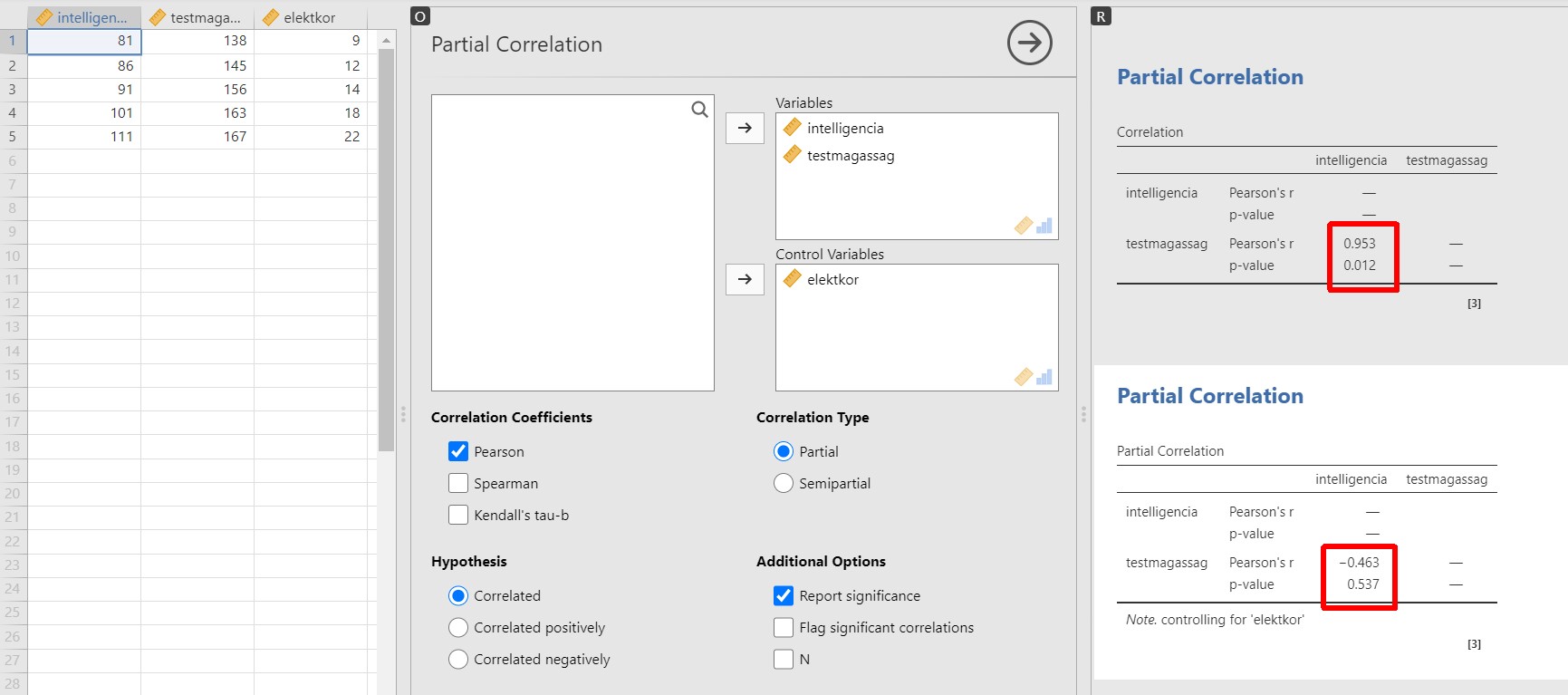
## 1.3 Parciális korrelációs együttható

Parciális korrelációs együttható: két változó () közötti korreláció mértéke, miután változók egy halmazának a két változó korrelációjára vonatkozó hatását többszörös lineáris regresszióval kiküszöböljük:

* , ahol és az és változó többszörös lineáris regresszióból származó becslése a magyarázó változók esetén.

d <- rio::import(file = "adat/lin\_reg\_intelligencia\_testmagassag\_eletkor\_01.xlsx")  
str(d)  
#> 'data.frame': 5 obs. of 3 variables:  
#> $ intelligencia: num 81 86 91 101 111  
#> $ testmagassag : num 138 145 156 163 167  
#> $ eletkor : num 9 12 14 18 22  
d  
#> intelligencia testmagassag eletkor  
#> 1 81 138 9  
#> 2 86 145 12  
#> 3 91 156 14  
#> 4 101 163 18  
#> 5 111 167 22

cor.test(d$intelligencia, d$testmagassag)  
#>   
#> Pearson's product-moment correlation  
#>   
#> data: d$intelligencia and d$testmagassag  
#> t = 5.4629, df = 3, p-value = 0.01205  
#> alternative hypothesis: true correlation is not equ...  
#> 95 percent confidence interval:  
#> 0.4463631 0.9970093  
#> sample estimates:  
#> cor   
#> 0.953235  
RcmdrMisc::partial.cor(d, tests = T)  
#>   
#> Partial correlations:  
#> intelligencia testmagassag eletkor  
#> intelligencia 0.00000 -0.46317 0.97875  
#> testmagassag -0.46317 0.00000 0.62856  
#> eletkor 0.97875 0.62856 0.00000  
#>   
#> Number of observations: 5   
#>   
#> Pairwise two-sided p-values:  
#> intelligencia testmagassag eletkor  
#> intelligencia 0.5368 0.0213   
#> testmagassag 0.5368 0.3714   
#> eletkor 0.0213 0.3714   
#>   
#> Adjusted p-values (Holm's method)  
#> intelligencia testmagassag eletkor  
#> intelligencia 0.7429 0.0638   
#> testmagassag 0.7429 0.7429   
#> eletkor 0.0638 0.7429

 A fenti példában látható, hogy míg szignifikáns erős pozitív kapcsolat van az intelligencia és a magasság között (korrelációs együttható: ), ez a kapcsolat eltűnik, ha figyelembe vesszük az életkor változót is (parciális korreláció: ). Vagyis sikerült az intelligencia és a testmagasság közötti kapcsolat erősségét megállapítani, miközben az életkor hatását erre a kapcsolatra kiküszöböltük.

A többszörös lineáris regressziós modell becsült paraméterei nagyban hasonlítanak a parciális korrelációs együtthatókra, mivel minden az és közötti kapcsolat erősségét írja le, miközben a többi magyarázó változó (, összesen db, nincs köztük) hatását kiküszöböljük a két változó korrelációjából.

A parciális korrelációs együtthatók és a többszörös lineáris regresszió együtthatói között annyira közvetlen a kapcsolat, hogy azonos p-érték tartozik hozzájuk, mint a lenti példában ez látható is lesz.

A lenti példában két modell szerepel, először az intelligencia és a testmagasság függvényszerű kapcsolatát vizsgáljuk és azt a meglepő dolgot tapasztaljuk, hogy minél magasabb valaki, annál intelligensebb , majd ha bevonjuk az életkor változót, akkor azt tapasztalhatjuk, hogy eltűnik az intelligencia és a testmagasság közötti kapcsolat .

lm\_1 <- lm(intelligencia ~ testmagassag, data = d)  
summary(lm\_1)  
#>   
#> Call:  
#> lm(formula = intelligencia ~ testmagassag, data = d)  
#>   
#> Residuals:  
#> 1 2 3 4 5   
#> 1.9228 0.3114 -5.0779 -1.6892 4.5328   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) -51.2613 26.6569 -1.923 0.1502   
#> testmagassag 0.9445 0.1729 5.463 0.0121 \*  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 4.202 on 3 degrees of freedom  
#> Multiple R-squared: 0.9087, Adjusted R-squared: 0...  
#> F-statistic: 29.84 on 1 and 3 DF, p-value: 0.01205  
lm\_2 <- lm(intelligencia ~ testmagassag + eletkor, data = d)  
summary(lm\_2)  
#>   
#> Call:  
#> lm(formula = intelligencia ~ testmagassag + eletkor...  
#>   
#> Residuals:  
#> 1 2 3 4 5   
#> 0.89121 -1.16330 -0.09995 0.21170 0.16034   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 73.1026 19.6042 3.729 0.0650 .  
#> testmagassag -0.1210 0.1637 -0.739 0.5368   
#> eletkor 2.6338 0.3902 6.750 0.0213 \*  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 1.055 on 2 degrees of freedom  
#> Multiple R-squared: 0.9962, Adjusted R-squared: 0...  
#> F-statistic: 259.3 on 2 and 2 DF, p-value: 0.003841

|  |
| --- |
| Az intelligencia és a testmagasság kapcsolata (N=5): két modell, életkor nélkül és életkorral |

## 1.4 A többszörös lineáris regresszió esetei

### 1.4.1 Egyetlen dichotóm magyarázó változó

A magyarázó változóink eddig kvantitatívak voltak, de kategorikus változók is lehetnek. Ha a kategorikus változónk csupán 2 értékű, akkor a becsült (, ) együtthatók értelmezése módosul. A tengelymetszet () a kategorikus változó referencia szintjén a függő változó átlagát tartalmazza, míg a () a kategorikus változó másik szintjén számolt átlag eltérését a -hoz képest.

d <- rio::import(file = "adat/lin\_reg\_magassag\_hajhossz\_nem\_01.xlsx")  
d$nem <- factor(d$nem, levels = c("nő", "férfi"))  
str(d)  
#> 'data.frame': 6 obs. of 3 variables:  
#> $ magassag: num 158 159 162 170 182 179  
#> $ hajhossz: num 28 25 20 1 1.5 3  
#> $ nem : Factor w/ 2 levels "nő","férfi": 1 1 1...  
d  
#> magassag hajhossz nem  
#> 1 158 28.0 nő  
#> 2 159 25.0 nő  
#> 3 162 20.0 nő  
#> 4 170 1.0 férfi  
#> 5 182 1.5 férfi  
#> 6 179 3.0 férfi

lm\_1 <- lm(magassag ~ nem, data = d)  
summary(lm\_1)  
#>   
#> Call:  
#> lm(formula = magassag ~ nem, data = d)  
#>   
#> Residuals:  
#> 1 2 3 4 5 6   
#> -1.6667 -0.6667 2.3333 -7.0000 5.0000 2.0000   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 159.667 2.687 59.413 4.81e-07 \*\*\*  
#> nemférfi 17.333 3.801 4.561 0.0103 \*   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 4.655 on 4 degrees of freedom  
#> Multiple R-squared: 0.8387, Adjusted R-squared: 0...  
#> F-statistic: 20.8 on 1 and 4 DF, p-value: 0.01033

|  |
| --- |
| A magasság és a nem kapcsolata |

A fenti példa a nem hatását vizsgálja testmagasságra. A p-érték alapján ez a hatás szignifikáns, tehát a függés fennáll, a paraméterek pedig a nők átlagáról és a férfiak és nők átlagának eltéréséről tájékoztatnak .

## 1.5 Modellválasztás

Előfordulhat, hogy egy jelenség vizsgálatakor több lineáris regressziós modellt is meg tudunk fogalmazni, nem csak egyetlen modell létezik. Ez a probléma leggyakrabban úgy jelenik meg, hogy rengeteg független változónk van, és nem tudjuk eldönteni, hogy elég egy kisebb modell, néhány változóval, vagy vegyük inkább a nagyobb modellt több változóval.

A megfelelő modell megtaláláshoz a modelleket összehasonlíthatjuk F-próba segítségével, szignifikáns eredmény esetén a két modell magyarázó ereje eltér egymástól. Ilyenkor a célunk a legszűkebb (legkevesebb magyarázó változót tartalmazó), de a legbővebbtől szignifikánsan nem különböző modell megtalálása.

A korrigált determinációs együttható is alkalmas mód a modellek összehasonlítására: az 1-hez legközelebbi értékkel bíró modell rendelkezik a legnagyobb magyarázó erővel. Léteznek már kritériumok is:

* AIC (Akaike-kritérium): minél kisebb az AIC értéke, annál nagyobb a modell magyarázó ereje.
* BIC (Bayes-kritérium): minél kisebb a BIC értéke, annál nagyobb a modell magyarázó ereje.
* RMSE (négyzetes középérték, Root Mean Square Error) az a mennyiség, amennyivel a vizsgált értékek eltérnek az előre megbecsült értékektől. Minél kisebb ez az érték, annál jobban becsül a modell.

d <- rio::import(file = "adat/lin\_reg\_fizetes\_eletkor\_elegedettseg\_01.xlsx")  
str(d)  
#> 'data.frame': 5 obs. of 3 variables:  
#> $ fizetes : num 44 66 89 155 130  
#> $ eletkor : num 25 65 21 35 40  
#> $ elegedettseg: num 37 36 61 92 76  
d  
#> fizetes eletkor elegedettseg  
#> 1 44 25 37  
#> 2 66 65 36  
#> 3 89 21 61  
#> 4 155 35 92  
#> 5 130 40 76

lm\_1 <- lm(elegedettseg ~ fizetes, data = d)  
summary(lm\_1)  
#>   
#> Call:  
#> lm(formula = elegedettseg ~ fizetes, data = d)  
#>   
#> Residuals:  
#> 1 2 3 4 5   
#> 4.249 -8.272 4.684 1.123 -1.785   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 9.71048 7.07562 1.372 0.26355   
#> fizetes 0.52365 0.06738 7.772 0.00443 \*\*  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 6.134 on 3 degrees of freedom  
#> Multiple R-squared: 0.9527, Adjusted R-squared: 0...  
#> F-statistic: 60.4 on 1 and 3 DF, p-value: 0.004432  
lm\_2 <- lm(elegedettseg ~ fizetes + eletkor, data = d)  
summary(lm\_2)  
#>   
#> Call:  
#> lm(formula = elegedettseg ~ fizetes + eletkor, data...  
#>   
#> Residuals:  
#> 1 2 3 4 5   
#> 0.28596 0.08556 -0.30015 0.71047 -0.78184   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 21.508055 1.292166 16.64 0.00359 \*\*   
#> fizetes 0.519198 0.008847 58.69 0.00029 \*\*\*  
#> eletkor -0.305549 0.023279 -13.12 0.00575 \*\*   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 0.8047 on 2 degrees of fre...  
#> Multiple R-squared: 0.9995, Adjusted R-squared: 0...  
#> F-statistic: 1841 on 2 and 2 DF, p-value: 0.000543  
anova(lm\_1, lm\_2)  
#> Analysis of Variance Table  
#>   
#> Model 1: elegedettseg ~ fizetes  
#> Model 2: elegedettseg ~ fizetes + eletkor  
#> Res.Df RSS Df Sum of Sq F Pr(>F)   
#> 1 3 112.864   
#> 2 2 1.295 1 111.57 172.28 0.005754 \*\*  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
performance::model\_performance(lm\_1)  
#> # Indices of model performance  
#>   
#> AIC | AICc | BIC | R2 | R2 (adj.) | RMS...  
#> ---------------------------------------------------...  
#> 35.773 | 59.773 | 34.601 | 0.953 | 0.937 | 4.75...  
performance::model\_performance(lm\_2)  
#> # Indices of model performance  
#>   
#> AIC | AICc | BIC | R2 | R2 (adj.) | RMSE ...  
#> ---------------------------------------------------...  
#> 15.436 | Inf | 13.873 | 0.999 | 0.999 | 0.509 ...

|  |
| --- |
| Elégedettség kapcsolata a fizetéssel és az élekorral (N=5): modellek összehasonlítása |

A fenti példán látható, hogy két modellt építettünk. Az 1. modell az elégedettséget a fizetés segítségével próbálja jósolni. A 2. modell az elégedettséget a fizetéssel és az életkorral. Láthatjuk a 2. modell szignifikánsan eltér magyarázó erőben a az 1. modelltől, valamint a modell “jóságát” leíró mutatók mindegyike kedvezőbb a 2. modell esetén: , , , .

## 1.6 Alkalmazási feltételek

A regressziós modellt ne használjuk, ha az alkalmazási feltételek valamelyike nem teljesül. Melyek ezek?

* Multikollinearitás: a független változók közötti korreláció. Multikollinearitás esetén elképzelhető, hogy a regressziós számítást el sem lehet végezni. Van mérőszám, amelynek a segítségével ki lehet szűrni az érintett változókat:
  + VIF (variancia infláció faktor): a nagy érték kollinearitást jelez az érintett változókat kihagyhatjuk, vagy származtatott adatokkal dolgozunk tovább (például főkomponenselemzéssel nyert adatokat)
* Autokorreláció: a hibatagok együttmozgása szignifikáns.

## 1.7 1. Példa

Befolyásolja-e a munkahellyel való elégedettséget a fizetés nagysága és az életkor? (Münnich és mtsai., 2006) [1.6.1 probléma]

d <- rio::import(file = "adat/lin\_reg\_elegedettseg.xlsx")  
str(d)  
#> 'data.frame': 30 obs. of 4 variables:  
#> $ fizetes : num 109 125 98 124 115 132 124 99...  
#> $ elegedettseg: num 69.2 90.8 71 90.1 77.8 ...  
#> $ kor : num 20 46.3 36.2 46 31 ...  
#> $ nem : chr "nő" "férfi" "nő" "férfi" ...  
psych::headTail(d)  
#> fizetes elegedettseg kor nem  
#> 1 109 69.15 20 nő  
#> 2 125 90.85 46.33 férfi  
#> 3 98 71.04 36.2 nő  
#> 4 124 90.12 45.95 férfi  
#> ... ... ... ... <NA>  
#> 27 129 96.32 53 férfi  
#> 28 135 97.86 49.4 férfi  
#> 29 145 100 46.9 férfi  
#> 30 120 81.08 32 férfi

lm\_1 <- lm(elegedettseg ~ fizetes + kor, data = d)  
summary(lm\_1)  
#>   
#> Call:  
#> lm(formula = elegedettseg ~ fizetes + kor, data = d)  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -10.248 -1.543 1.188 2.437 4.333   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 8.13712 3.26797 2.490 0.0192 \*   
#> fizetes 0.44404 0.02321 19.128 < 2e-16 \*\*\*  
#> kor 0.53361 0.07127 7.487 4.71e-08 \*\*\*  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 3.59 on 27 degrees of freedom  
#> Multiple R-squared: 0.953, Adjusted R-squared: 0....  
#> F-statistic: 273.9 on 2 and 27 DF, p-value: < 2.2e-16  
lsr::standardCoefs(lm\_1)  
#> b beta  
#> fizetes 0.4440418 0.8322319  
#> kor 0.5336111 0.3257287

|  |
| --- |
| Elégedettség kapcsolata a fizetéssel és lektkorra (N=30) |

A fenti outputból láthatjuk, hogy a fizetés és a kor változó is szignifikánsan befolyásolja az elégedettséget, hiszen a hozzájuk tartozó szignifikanciaszint . A teljes modell vonatkozó F-próba is szignifikáns. A fizetés változó együtthatója 0,44, a kor változó együtthatója pedig 0,53, ami arra utal, hogy pozitív kapcsolat van a változó között: minél magasabb a fizetés, és minél idősebbek az emberek, annál elégedettebbek a munkahelyükkel.

A pontos becslés a regressziós egyenlet alapján a következőképpen fest:

elégedettség = 8,14 + 0,44 \* fizetés + 0,53 \* kor

Mivel a többszörös regresszió esetében a független változók hatása csak a standardizált együtthatók mentén hasonlítható össze, így kiszámoltuk a standardizált együtthatókat is. Az adatok jól példázzák, hogy miért fontos a standardizált együtthatókat is vizsgálni, hiszen a nem standardizált együtthatók esetén a kor változó együtthatójának értéke a magasabb, míg a standardizált értékeknél fordítva. Vagyis, ha az egyes változók relatív fontosságának vizsgálatakor nem nézzük a dimenziómentes értékeket, akkor könnyen téves következtetésre juthatunk.

A négyzetes korrelációs együttható értéke 0,9, ami arra utal, hogy a független változók igen jól magyarázzák a függő változót.

## 1.8 2. Példa

Befolyásolja-e a kalandvágy a hivatásos katonai szolgálatnál eltöltött időt? (Münnich és mtsai., 2006) [1.6.2 probléma]

d <- rio::import(file = "adat/lin\_reg\_katonasag.xlsx")  
str(d)  
#> 'data.frame': 156 obs. of 4 variables:  
#> $ kaland : num 3 3 5 1 3 4 2 2 3 5 ...  
#> $ egyhangu: num 3 2 4 1 3 1 2 1 2 3 ...  
#> $ sport : num 1 1 2 1 2 2 2 1 2 2 ...  
#> $ evek : num 4 7 10 3 6 15 5 6 9 13 ...  
psych::headTail(d)  
#> kaland egyhangu sport evek  
#> 1 3 3 1 4  
#> 2 3 2 1 7  
#> 3 5 4 2 10  
#> 4 1 1 1 3  
#> ... ... ... ... ...  
#> 153 2 4 2 1  
#> 154 3 5 3 2  
#> 155 1 3 2 2  
#> 156 3 2 4 12

lm\_1 <- lm(evek ~ egyhangu + sport + kaland, data = d)  
summary(lm\_1)  
#>   
#> Call:  
#> lm(formula = evek ~ egyhangu + sport + kaland, data...  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -1.6611 -0.5925 -0.0798 0.2726 9.7833   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 0.63951 0.33263 1.923 0.0564 .   
#> egyhangu -2.28197 0.09160 -24.912 <2e-16 \*\*\*  
#> sport 1.52987 0.09447 16.194 <2e-16 \*\*\*  
#> kaland 3.17525 0.09884 32.125 <2e-16 \*\*\*  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 1.213 on 152 degrees of fr...  
#> Multiple R-squared: 0.9195, Adjusted R-squared: 0...  
#> F-statistic: 578.4 on 3 and 152 DF, p-value: < 2.2...  
lsr::standardCoefs(lm\_1)  
#> b beta  
#> egyhangu -2.281974 -0.5927751  
#> sport 1.529871 0.3821476  
#> kaland 3.175249 0.7676009

|  |
| --- |
| Befolyásolja-e a kalandvágy a hivatásos katonai szolgálatnál eltöltött időt? (N=156) |

A fenti output a többszörös lineáris regresszió eredményét mutatja:

* A lineáris regressziós modellt megtarthatjuk, hiszen az F-statisztika értékét tekintve a modell szignifikáns, a változók együtthatóinak az értéke nem nulla.
* A modell magyarázóértéke igen jó, hiszen a korrigált determinációs együttható értéke 0,92, vagyis a független változók a függő változó varianciájának kb. 92%-át magyarázzák.
* Minden egyes független változó hatással van a függő változóra, vagyis mind a kalandvágy, az extrém sportok szeretete és a nyugalom utáni vágy is befolyásolja azt, hogy mennyi időt tölt valaki a hivatásos katonai szolgálatban.
* Ellenben a vagyis a konstans értéke most nulla, hiszen a táblázatban szereplő érték nem szignifikáns.
* Maga a regressziós egyenlet a pontos együtthatók ismeretében a következőképpen alakul:

evek=3,18\*kaland+1,53\*sport-2,28\*egyhangu

* Vagyis minél jobban kedveli valaki a kalandos életet és az extrém sportokat, és minél jobban irtózik a szürke hétköznapoktól, annál több időt tölt a katonaság kötelékében.
* A standardizált változók alapján a kaland szeretetének a hatása a legerősebb (0,768), a második legerősebb hatás az egyhangúság kedvelése, ám hatásának iránya negatív (-0,593), leggyengébb hatása pedig az extrém sportok szeretetének van (0,382).

# 2. Főkomponens elemzés

A főkomponens elemzés a legegyszerűbb többváltozós statisztikai eljárások egyike. Olyan eljárás, amely az egy jelenségre vonatkozó méréseket úgy „összegzi”, hogy közben az óhatatlanul fellépő információ veszteséget a lehető legkisebb mértékűre csökkenti.

A módszer alapgondolata az, hogy vegyünk változót: , majd keressük meg ezek kombinációit, hogy ezáltal -vel jelölt indexeket kaphassunk, melyek egymással nem korrelálnak, továbbá , ahol a komponens varianciáját jelöli.

A korreláció hiánya a -kben hasznos tulajdonság, ugyanis azt jelenti, hogy az indexek az adatok különböző “dimenzióit” mérik.

A -t főkomponensnek nevezzük. Amikor főkomponens-analízist végzünk, mindig abban bízunk, hogy a legtöbb index varianciája elhanyagolhatóan kicsi. Ezáltal az adatok varianciája adekvátan leírható néhány olyan változóval, melyek varianciája nem elhanyagolható.

A főkomponens analízis működéséhez szükséges, hogy az eredeti változók korreláljanak egymással (akár pozitív, akár negatív irányban), ekkor elképzelhető az az eset, hogy 20-30 eredeti változót adekvátan reprezentálhat 2-3 főkomponens. Ha pedig ez teljesül, akkor a fontosabb főkomponensek (melyek varianciája elég nagy) lesznek csupán érdekesek, hiszen ezek fogják az adatok „dimenzióit” mérni. Természetesen nagyon fontos azt is tudnunk, hogy rengeteg eredeti változónk van, és legtöbbjük ugyanazt, vagy legalábbis hasonló dolgokat mér.

## 2.1 A főkomponens elemzés menete

Az eredeti változókból a lineáris kombinációk segítségével kapjuk meg a főkomponenseket, azzal feltétellel, hogy , és az egymás után létrejövő főkomponensek nem korrelálnak egymással.

Gyakran az változó standardizált értékeiből indulunk ki, hogy a változók arányosan fejtsék ki hatásukat a főkomponensekre. A jamovi is így végzi az elemzést. Ekkor a változok átlaga nulla, szórása és variancia pedig 1 lesz.

A részletek ismertetése nélkül a keresett együtthatók megtalálása egy sajátérték-sajátvektor keresési feladat az eredeti változók korrelációs mátrixában. A megtalált darab sajátérték $\_1\_2\_p>0 $sorrendjét feltételezve, az főkomponens varianciáját adja , és a megtalált darab sajátvektorból az egyes elemei lesznek a főkomponens együtthatói.

Fontos összefüggés, hogy a főkomponensek (-k) varianciájának az összege egyenlő az eredeti standardizált változók (-k) varianciájának összegével, azaz

## 2.2 A főkomponens elemzés alkalmazási feltételei

* A főkomponens elemzés során általában 4-5-ször (egyes szerzőknél 10-szer) nagyobb a mintaelemszám a vizsgált változók számánál.
* A faktoranalízis feltétele, hogy egymással korreláló változókból induljunk ki. A Bartlett-féle szferikus próba nullhipotézise, hogy a változók korrelálatlanok (vagyis a korrelációs mátrixnak a főátlón kívüli elemei csak véletlenül térnek el a nullától). A szignifikáns p-érték a kedvező a főkomponens elemzés számára. (Megjegyezzük, hogy a túlságosan magas egyirányú korrelációk sem jók, ugyanis ez azt okozhatja, hogy a főkomponens elemzésnek nem lesz megoldása, ugyanis minden változó egy faktorba kerül.)
* Az MSA (Measure of Sampling Adequecy) az egyes változók esetében mutatja meg, hogy mennyire van szoros kapcsolatban a többi változóval. Érdemes a 0,5 alatti MSA értékkel rendelkező változókat kizárni az elemzésből. Értéke 0 és 1 közötti lehet.
* A Kaiser-Meyer-Olkin- (KMO) kritérium az MSA értékek átlaga. Míg az MSA érték az egyes változókra vonatkozik, a KMO az összes változóra egyidejűleg. A KMO mutatószám jelentését a következőképpen ítélhetjük meg:
  + KMO ≥ 0,9 kiváló
  + KMO ≥ 0,8 nagyon jó
  + KMO ≥ 0,7 megfelelő
  + KMO ≥ 0,6 közepes
  + KMO ≥ 0,5 gyenge
  + KMO < 0,5 elfogadhatatlan.

## 2.3 A főkomponensek forgatása (rotáció)

A faktorkiválasztás (extrakció) során az elemzés elsődleges célja, hogy maximalizálja a főkomponensek varianciáját, amely eredményeként megkapjuk a rotálatlan faktorsúly-mátrixot. A faktorsúly az eredeti változó és az adott faktor közötti korrelációt mutatja, amelynek értéke a korrelációs együtthatókhoz hasonlóan -1 és 1 között változhat.

A faktorkiválasztás során azonban előfordulhat, hogy olyan változók fognak korrelálni egy adott faktorral, amelyeknek semmi közük egymáshoz, ezáltal lehetetlenné téve az értelmezést. Ezen a problémán segít a forgatás, vagy más néven rotáció. A faktor-rotáció azt jelenti, hogy a faktorok tengelyeit elforgatjuk úgy, hogy egyszerűbb és értelmezhetőbb faktormegoldáshoz vezessen.

A rotáció (forgatás) során nem változnak sem a kommunalitás, sem pedig az összes magyarázott variancia, csak a faktorok sajátértékei/magyarázott varianciái módosulnak.

A rotáláson belül két típust különböztetünk meg: a derékszögű (ortogonális) (Varimax, Equimax, Quartimax) és a hegyesszögű (nem ortogonális) (Direct Oblimin, Promax) forgatási módszereket.

A derékszögű esetében a tengelyek merőlegesen állnak egymásra, ezáltal a faktorok nem korrelálnak egymással, míg a hegyesszögű esetében ezek tetszőleges szöget zárnak be egymással, vagyis a faktorok korrelálni fognak egymással.

## 2.4 1. Példa

A példa Münnich és mtsai. (2006) 2.3.1. alfejezet alapján készült és a kapcsolódó jamovi állomány: fokomp\_elemzes\_tantargyak.omv.

**1. Határozzuk meg a korrelációs mátrixot (jamovi-ban: Regression / Correlation matrix)**

d <- rio::import(file = "adat/fokomp\_elemzes\_tantargyak.xlsx")  
str(d)  
#> 'data.frame': 9 obs. of 4 variables:  
#> $ matek : num 5 4 3 2 5 1 5 2 5  
#> $ fizika : num 5 5 3 3 4 2 4 3 5  
#> $ informatika: num 4 4 4 2 5 1 5 2 5  
#> $ kemia : num 5 5 3 3 5 1 5 3 5  
d  
#> matek fizika informatika kemia  
#> 1 5 5 4 5  
#> 2 4 5 4 5  
#> 3 3 3 4 3  
#> 4 2 3 2 3  
#> 5 5 4 5 5  
#> 6 1 2 1 1  
#> 7 5 4 5 5  
#> 8 2 3 2 3  
#> 9 5 5 5 5

cor(d)  
#> matek fizika informatika kemia  
#> matek 1.0000000 0.8712476 0.9492623 0.9499475  
#> fizika 0.8712476 1.0000000 0.7662496 0.9271176  
#> informatika 0.9492623 0.7662496 1.0000000 0.8867155  
#> kemia 0.9499475 0.9271176 0.8867155 1.0000000

|  |
| --- |
| Korrelációs mátrix meghatározása |

A korrelációs mátrix adatai arra utalnak, hogy szoros kapcsolat van a változók között. A korrelációs értékek nullánál nagyobbak, ami azonos irányú tendenciákra utal. E két mátrix is alátámasztja a feltételezésünket, hogy a változók szorosan együtt változnak.

**2. Ellenőrizzük le az adatok alkalmasságát (jamovi-ban: Factor / Principal Component Analysis)**

A változóink eleget tesznek a Bartlett-féle szferikus próbának, a korrelációs mátrix nem egységmátrix , az MSA értékek is nagyobban -nél és a KMO érték is megfelelő.

|  |
| --- |
| Alkalmazási feltételek ellenőrzése |

**3. Határozzuk meg a komponensek számát**

Elvileg annyi főkomponenst lehet kiszámolni, ahány változónk van, a célunk azonban a komponensek számának minimalizálása.

Több eljárás létezik a főkomponensek számának meghatározására:

* Horn-féle párhuzamos analízis (jamovi-ban: Based on parallel analysis): modern eljárás, amely szimuláció segítségével állapítja meg a főkomponensek számát (Horn, 1965).
* A priori meghatározás (jamovi-ban: Fixed number): korábbi ismerete alapján megadjuk a főkomponensek számát.
* Sajátértéken alapuló megoldás (jamovi-ban: Based on eigenvalue): tipikusan csak az 1-nél nagyobb sajátértékű faktorokat tartjuk bent a modellben. Az 1-nél kisebb varianciájú faktorok ugyanis nem jobbak mint az eredeti standardizált változók
* Sajátértékábrán (scree-plot, kőtörmelék ábra) alapuló meghatározás (jamovi-ban: Scree plot): a sajátérték ábra a sajátértékek ábrázolása a főkomponensek sorrendjében. Az ábra formája alapján lehet következtetni a főkomponensek számára: ahol a görbe meredekségében van egy határozott törés, meredekebb rész után laposabb jön. Ahol tehát a görbe laposodása elkezdődik, az a figyelembe vett főkomponensek megfelelő száma.
* Magyarázott varianciahányadon alapuló meghatározás (jamovi-ban: Component summary): ekkor az előállított főkomponensek számát úgy határozzuk meg, hogy a főkomponensek által magyarázott variancia kumulált százalékos értéke elérjen egy megfelelő szintet. A megfelelő szint (60%-95%-ig) a probléma jellegétől függ.

A Horn-féle párhuzamos elemzés 1 főkomponenst javasol.

|  |
| --- |
| Főkomponensek számának meghatározása |

**4. Válasszunk forgatást (jamovi-ban: Rotation)**

A jamovi alapértelmezés szerint a Varimax forgatást ajánlja, amely derékszögű koordinátatengelyeket eredményez és a legtöbb esetben ez a megfelelő választás. Lehetőségünk van ezen módosítani. Az összes lehetőség:

* None – rotálatlan elemzés
* Varimax
* Qartimax
* Promax
* Oblimin
* Simplimax

Mivel egyetlen főkomponensünk van, így nem változtatunk az alapértelmezett Varimax beállításon.

**5. A főkomponens elemzés eredménye**

*Komponens mátrix (jamoviban: Component loadings)*

A főkomponens elemzés eredménye a komponens mátrix (faktormátrix), amelynek soraiban az eredeti változók, oszlopaiban a kinyert főkomponensek vannak. A cellákban a komponens súlyok (faktorsúlyok) szerepelnek, amelyek a főkomponens és a változó közötti korrelációt jelentik. Ezek egyben a főkomponensek azon együtthatói, amelyekkel a standardizált változó a főkomponensekkel kifejezhető.

A magas abszolút értékű faktorsúly azt jelzi, hogy komponens és a változó szorosan összefügg.

A változókat tartalmazó sorok rendezhetők a faktorsúlyok csökkenő sorrendjében (jamovi-ban: Sort loading by size)

Az adott értéknél kisebb faktorsúlyok elrejthetők a táblázatban (jamoviban: Hide loadings below)

A Uniqueness oszlopban az egyes változók „egyediségét” is láthatjuk. Az egyediség a variancia azon aránya, amely „egyedi” a változóra nézve, és nem magyarázható a komponensekkel. Vegyük figyelembe, hogy minél nagyobb az „egyediség”, annál kisebb a változó relevanciája/hozzájárulása a modellben.

*A kezdő sajátértékek (jamovi-ban: Initial eigenvalues)*

A kezdő sajátértékek táblázat a sajátértékeket adja meg. A komponensek sajátértékei csökkenő nagyságúak, ahogy az 1. komponenstől a 4. komponensig haladunk. A komponens sajátértéke kifejezi a komponens által magyarázott teljes varianciát. A 4 komponens összvarianciája pontosan 4. A további két oszlopban ez alapján számoljuk a százalékos és a kummulált százalékos varianciát.

*A komponensek összegzése (jamovi-ban: Component summary)*

A komponensek összegzése táblázat tartalmazza a megtartott komponenseket, a magyarázott varianciát, illetve utóbbit százalékosan is kifejezve. Vegyük észre, hogy ez a sor teljesen megegyezik a kezdő sajátértékek táblázat első sorával. Az SS Loadings felirat magyarázata, hogy magyarázott variancia a komponenshez tartozó faktorsúlyok négyzetösszege (sum of square).

|  |
| --- |
| Komponensek összegzése |

**6. Főkomponens értékek kiszámítása**

A főkomponens elemzés célja az eredeti változók csökkentése. A főkomponens(ek) az eredeti változók lineáris kombinációjával kifejezhetők. Ez(ek) a főkomponens értékek (jamovi-ban: Component score) az adatbázisban is rögzíthetők, és további elemzések kiindulópontjai lehetnek.

|  |
| --- |
| Főkomponens értékek kiszámítása |

Sikerült tehát az érdemjegyeket egyetlen mérőszámmal kifejezni, a fenti főkomponens érték az, amely a lehető legjobban magában foglalja az egyes tantárgyakból szerzett jegyeket és ezáltal a reál tantárgyak iránti fogékonyság mérőszáma lehet. A legjobban a kilencedik személy teljesít a reál tárgyakból, legrosszabbul pedig a hatodik. Ezek az értékek standardizáltak, vagyis 0 átlagúak és 1 szórásúak.

|  |
| --- |
| Főkomponens értékek leíró statisztikája |

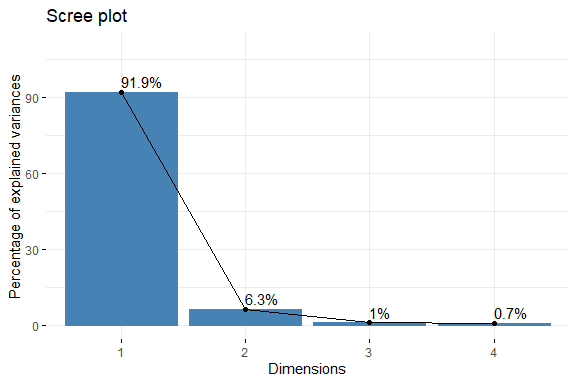
R-ben több lehetőségünk van a főlomponenselemzés elvégzésére.

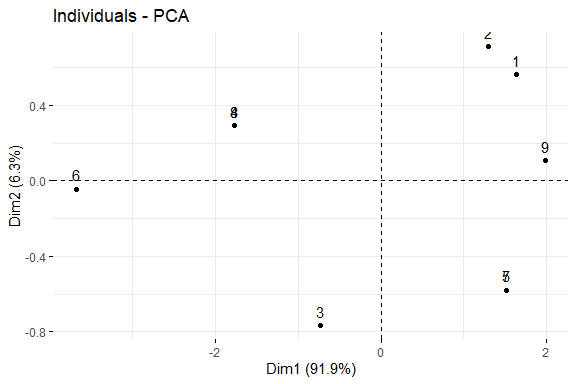
pca\_1 <- prcomp(d, scale. = TRUE)  
pca\_1  
#> Standard deviations (1, .., p=4):  
#> [1] 1.9177188 0.5017701 0.2045278 0.1695572  
#>   
#> Rotation (n x k) = (4 x 4):  
#> PC1 PC2 PC3  
#> matek -0.5129614 0.2231620 -0.02972158  
#> fizika -0.4843985 -0.7077528 0.50610146  
#> informatika -0.4900124 0.6474152 0.35260295  
#> kemia -0.5119731 -0.1736040 -0.78654250  
#> PC4  
#> matek -0.82836339  
#> fizika 0.09113404  
#> informatika 0.46520167  
#> kemia 0.29848968

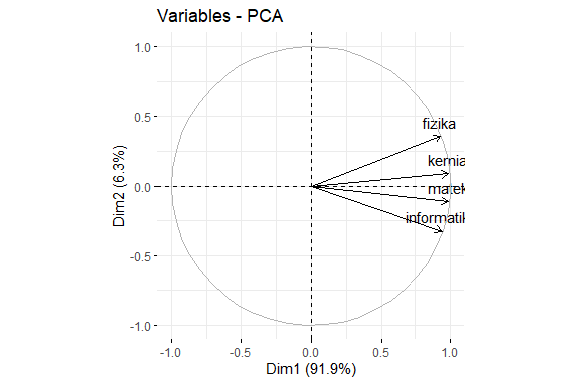
pca\_1 <- princomp(d, cor = TRUE)  
pca\_1  
#> Call:  
#> princomp(x = d, cor = TRUE)  
#>   
#> Standard deviations:  
#> Comp.1 Comp.2 Comp.3 Comp.4   
#> 1.9177188 0.5017701 0.2045278 0.1695572   
#>   
#> 4 variables and 9 observations.

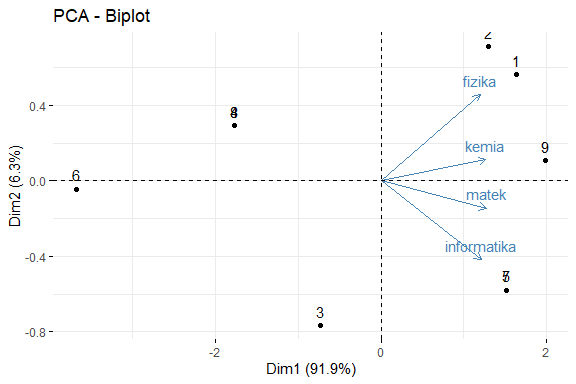
psych::pca(d, rotate = "varimax")  
#> Principal Components Analysis  
#> Call: principal(r = r, nfactors = nfactors, residua...  
#> rotate = rotate, n.obs = n.obs, covar = covar, ...  
#> missing = missing, impute = impute, oblique.sco...  
#> method = method, use = use, cor = cor, correct ...  
#> Standardized loadings (pattern matrix) based upon c...  
#> PC1 h2 u2 com  
#> matek 0.98 0.97 0.032 1  
#> fizika 0.93 0.86 0.137 1  
#> informatika 0.94 0.88 0.117 1  
#> kemia 0.98 0.96 0.036 1  
#>   
#> PC1  
#> SS loadings 3.68  
#> Proportion Var 0.92  
#>   
#> Mean item complexity = 1  
#> Test of the hypothesis that 1 component is sufficient.  
#>   
#> The root mean square of the residuals (RMSR) is 0.05   
#> with the empirical chi square 0.28 with prob < ...  
#>   
#> Fit based upon off diagonal values = 1

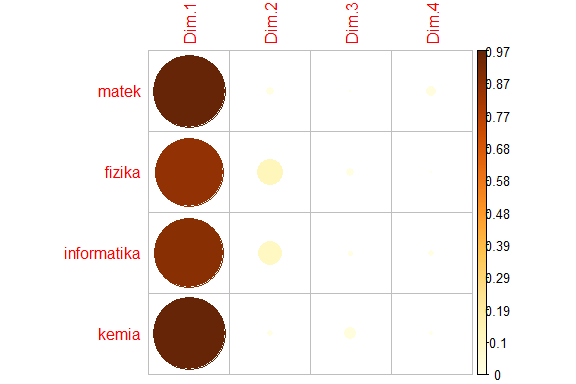
pca\_1 <- FactoMineR::PCA(d, graph = FALSE)  
pca\_1$eig  
#> eigenvalue percentage of variance  
#> comp 1 3.67764553 91.9411383  
#> comp 2 0.25177321 6.2943303  
#> comp 3 0.04183160 1.0457901  
#> comp 4 0.02874965 0.7187414  
#> cumulative percentage of variance  
#> comp 1 91.94114  
#> comp 2 98.23547  
#> comp 3 99.28126  
#> comp 4 100.00000  
pca\_1$var  
#> $coord  
#> Dim.1 Dim.2 Dim.3  
#> matek 0.9837158 -0.1119760 -0.006078888  
#> fizika 0.9289402 0.3551292 0.103511795  
#> informatika 0.9397061 -0.3248536 0.072117091  
#> kemia 0.9818205 0.0871093 -0.160869772  
#> Dim.4  
#> matek -0.14045500  
#> fizika 0.01545243  
#> informatika 0.07887831  
#> kemia 0.05061108  
#>   
#> $cor  
#> Dim.1 Dim.2 Dim.3  
#> matek 0.9837158 -0.1119760 -0.006078888  
#> fizika 0.9289402 0.3551292 0.103511795  
#> informatika 0.9397061 -0.3248536 0.072117091  
#> kemia 0.9818205 0.0871093 -0.160869772  
#> Dim.4  
#> matek -0.14045500  
#> fizika 0.01545243  
#> informatika 0.07887831  
#> kemia 0.05061108  
#>   
#> $cos2  
#> Dim.1 Dim.2 Dim.3  
#> matek 0.9676968 0.012538628 3.695288e-05  
#> fizika 0.8629298 0.126116721 1.071469e-02  
#> informatika 0.8830475 0.105529830 5.200875e-03  
#> kemia 0.9639714 0.007588031 2.587908e-02  
#> Dim.4  
#> matek 0.0197276074  
#> fizika 0.0002387777  
#> informatika 0.0062217871  
#> kemia 0.0025614817  
#>   
#> $contrib  
#> Dim.1 Dim.2 Dim.3 Dim.4  
#> matek 26.31294 4.980128 0.08833723 68.6185908  
#> fizika 23.46419 50.091398 25.61386842 0.8305413  
#> informatika 24.01122 41.914638 12.43288421 21.6412590  
#> kemia 26.21165 3.013836 61.86491013 8.9096089  
factoextra::fviz\_eig(pca\_1, addlabels = TRUE, ylim = c(0, 110))  
factoextra::get\_eigenvalue(pca\_1)  
#> eigenvalue variance.percent  
#> Dim.1 3.67764553 91.9411383  
#> Dim.2 0.25177321 6.2943303  
#> Dim.3 0.04183160 1.0457901  
#> Dim.4 0.02874965 0.7187414  
#> cumulative.variance.percent  
#> Dim.1 91.94114  
#> Dim.2 98.23547  
#> Dim.3 99.28126  
#> Dim.4 100.00000  
factoextra::get\_pca\_ind(pca\_1)  
#> Principal Component Analysis Results for individuals  
#> ===================================================  
#> Name Description   
#> 1 "$coord" "Coordinates for the individuals"   
#> 2 "$cos2" "Cos2 for the individuals"   
#> 3 "$contrib" "contributions of the individuals"  
factoextra::get\_pca\_var(pca\_1)  
#> Principal Component Analysis Results for variables  
#> ===================================================  
#> Name   
#> 1 "$coord"   
#> 2 "$cor"   
#> 3 "$cos2"   
#> 4 "$contrib"  
#> Description   
#> 1 "Coordinates for the variables"   
#> 2 "Correlations between variables and dimensions"  
#> 3 "Cos2 for the variables"   
#> 4 "contributions of the variables"  
factoextra::fviz\_pca\_ind(pca\_1)  
factoextra::fviz\_pca\_var(pca\_1)  
factoextra::fviz\_pca\_biplot(pca\_1)  
corrplot::corrplot(pca\_1$var$cos2, is.corr = FALSE)











1. Példa

Létezik a reál tárgyak iránti fogékonyság?

Korábban már foglalkoztunk azzal a felvetéssel, hogy néhány tantárgy eredményeit egyetlen mérőszámmal reprezentáljuk. Korábbi példánkban a matematika, fizika, informatika és kémia jegyek közötti összefüggéseket vizsgáltuk egy kisebb adatbázison, most egy sokkal nagyobb adatbázis segítségével mutatjuk be, hogyan végezhetünk főkomponens-analízist.

Összességében az adatok jól sűríthetők egyetlen mérőszámba, minimális információveszteséggel, ezt a mutatót pedig hívhatjuk a reál tárgyak iránti fogékonyság mutatójának.

1. Példa

Egy kérdőív szerkesztésének problémái

A kapott eredmények alapján az itemszelekciót ennél a lépésnél befejezhetjük. Az így kapott hat itemünk a statisztikai eredmények alapján egészen jól lefednek egy dimenziót, ezáltal használhatóak egy jelenség kérdőíves vizsgálatára.

# 3. Megbízhatóság elemzés

A pszichológiai tesztelés során használt mérőeszközök legfontosabb tulajdonsága a megbízhatóság (reliabilitás) és a validitás (Carver és Scheier, 2006; Nagy, 2006). A reliabilitás azt mutatja meg, hogy az eszköz mennyire mér megbízhatóan, pontosan, mennyire bízhatunk abban, hogy a mérés második és harmadik alkalommal is ugyanazt az eredményt adja, amit az első esetben. A validitás vagy érvényesség azt jelenti, hogy a mérőeszköz azt méri, amit mérni szeretnénk. A klasszikus ábrázolás szerint mérőeszközünk a megbízhatóság és az érvényesség alapján a lenti négy csoportok egyikébe is eshet:

|  |
| --- |
| Megbízhatóság és validitás esetei |

Egy mérőeszköz megbízhatóságát mindig úgy vizsgáljuk, hogy a mérés eredményét egy vagy több más eszköz eredményével hasonlítjuk össze. Az összehasonlítás mindig korrelációt típusú vizsgálatot jelent, és a magasabb korreláció egyben magasabb megbízhatóságot jelöl. Ennek megfelelően a megbízhatósági mutatók értéktartománya megegyezik a korrelációs együttható értéktartományával (feltehetően azonban 0 és 1 közötti lesz az értéke, negatív értéket ritkán kapunk és el is szeretnénk kerülni).

A megbízhatósággal kapcsolatban három aspektust érdemes vizsgálni:

* **belső konzisztencia** – az önjellemző skálák sok tételből állnak (melyik mindegyike külön mérőeszköznek tekinthető), ezek kitöltésével egyidejűleg végzünk egymással ekvivalens, párhuzamos méréseket.
* **időbeli stabiltás (teszt-reteszt reliabilitás)** – időben eltolva, ugyanazon mérést egy későbbi időpontban megismételve jutunk két mérési eredményhez, például, ha ugyanazt az önjellemző skálát mondjuk, egy nap eltéréssel felvesszük ugyanazon személyekkel.
* **értékelő megbízhatóság (inter-rater reliabilitás)** – amikor megfigyelő pontoz, akkor a megfigyelő személy a mérőeszköz, így a megfigyelői ítéleteket az értékelő megbízhatóságának meghatározásával ellenőrizzük.

## 3.1 Cronbach-alfa – belső konzisztencia mérése

Főkomponens elemzés segítségével könnyen tudunk több változót - viszonylag csekély veszteséggel - egyetlen változóba tömöríteni, ezért gyakran használják kérdőívek itemeinek szelekciójára, valamint megbízhatóság (reliabilitás) vizsgálatra. A klasszikus tesztelmélet keretein belül azonban a tesztek megbízhatóságának (reliabilitásának) több lehetséges mutatója is létezik.

Cronbach 1951-es munkájában publikálta azon nézetét, hogy a korábbi egyszerű tesztfelezéses eljárás helyett egy annál tökéletesebb mutatót kellene használni a tesztek megbízhatóságának indikátoraként. Ha az itemek száma alacsony vagy az itemek közötti átlagos korreláció alacsony, akkor csökkenni fog a Cronbach-féle alfa értéke is. Az is egyértelmű, hogy az itemek közötti alacsony korreláció arra enged következtetni, hogy a teszt itemjei nem egy és ugyanazon dolog vizsgálatára szolgálnak, a belőlük képzendő tesztérték nem alkalmas sem elméleti, sem pedig gyakorlati felhasználásra.

Az ómega (McDonald ) korrigálja a Cronbach-alfa torzítását, érdemes elvégezni az elemzést ezzel a mutatóval is (Kárász és mtsai., 2022; Malkewitz és mtsai., 2023).

## 3.2 Példa: Real tárgyak iránti fogékonyság

Egy fiktív adatbázis 9 tanuló iskolai jegyeit tartalamzza reál tantárgyakból (matematika, fizika, kémia, informatika) (megbizhatosag\_tantargyak.xlsx). Vizsgáljuk meg, ha a reál tantárgyak iránti fogékonyságot ezzel a 4 érdemjeggyel mérnénk, akkor ez megbízhatóság szempontjából alkalmas mérőeszköz lenne.

real <- rio::import(file = "adat/megbizhatosag\_tantargyak.xlsx")  
str(real)  
#> 'data.frame': 9 obs. of 4 variables:  
#> $ matek : num 5 4 3 2 5 1 5 2 5  
#> $ fizika : num 5 5 3 3 4 2 4 3 5  
#> $ informatika: num 4 4 4 2 5 1 5 2 5  
#> $ kemia : num 5 5 3 3 5 1 5 3 5

A Cronbach alfa meghatározását végezhetjük a {psych} csomag alpha() függvényével.

psych::alpha(real) # Cronbach-alfa  
#>   
#> Reliability analysis   
#> Call: psych::alpha(x = real)  
#>   
#> raw\_alpha std.alpha G6(smc) average\_r S/N ase mean  
#> 0.97 0.97 0.98 0.89 33 0.017 3.7  
#> sd median\_r  
#> 1.4 0.91  
#>   
#> 95% confidence boundaries   
#> lower alpha upper  
#> Feldt 0.91 0.97 0.99  
#> Duhachek 0.93 0.97 1.00  
#>   
#> Reliability if an item is dropped:  
#> raw\_alpha std.alpha G6(smc) average\_r S/N  
#> matek 0.94 0.95 0.95 0.86 18  
#> fizika 0.97 0.98 0.97 0.93 39  
#> informatika 0.96 0.97 0.97 0.92 33  
#> kemia 0.94 0.95 0.96 0.86 19  
#> alpha se var.r med.r  
#> matek 0.032 0.0070 0.89  
#> fizika 0.015 0.0013 0.95  
#> informatika 0.019 0.0016 0.93  
#> kemia 0.029 0.0084 0.87  
#>   
#> Item statistics   
#> n raw.r std.r r.cor r.drop mean sd  
#> matek 9 0.99 0.98 0.98 0.97 3.6 1.6  
#> fizika 9 0.92 0.93 0.91 0.88 3.8 1.1  
#> informatika 9 0.95 0.94 0.93 0.90 3.6 1.5  
#> kemia 9 0.98 0.98 0.98 0.96 3.9 1.5  
#>   
#> Non missing response frequency for each item  
#> 1 2 3 4 5 miss  
#> matek 0.11 0.22 0.11 0.11 0.44 0  
#> fizika 0.00 0.11 0.33 0.22 0.33 0  
#> informatika 0.11 0.22 0.00 0.33 0.33 0  
#> kemia 0.11 0.00 0.33 0.00 0.56 0

A McDonald értékét kiszámolhatjuk a {psych} csomag omega() függvényével.

psych::omega(real, plot = F) # McDonald-ómega  
#> Omega   
#> Call: omegah(m = m, nfactors = nfactors, fm = fm, k...  
#> digits = digits, title = title, sl = sl, labels...  
#> plot = plot, n.obs = n.obs, rotate = rotate, Ph...  
#> covar = covar)  
#> Alpha: 0.97   
#> G.6: 0.98   
#> Omega Hierarchical: 0.95   
#> Omega H asymptotic: 0.96   
#> Omega Total 0.99   
#>   
#> Schmid Leiman Factor loadings greater than 0.2   
#> g F1\* F2\* F3\* h2 u2 p2  
#> matek 0.97 0.28 0.99 0.01 0.94  
#> fizika 0.89 0.29 0.92 0.08 0.86  
#> informatika 0.91 0.28 0.95 0.05 0.87  
#> kemia 0.96 0.29 0.99 0.01 0.94  
#>   
#> With Sums of squares of:  
#> g F1\* F2\* F3\*   
#> 3.48 0.17 0.16 0.04   
#>   
#> general/max 21.07 max/min = 4.03  
#> mean percent general = 0.9 with sd = 0.04 and ...  
#> Explained Common Variance of the general factor = ...  
#>   
#> The degrees of freedom are -3 and the fit is 0   
#> The number of observations was 9 with Chi Square ...  
#> The root mean square of the residuals is 0   
#> The df corrected root mean square of the residuals ...  
#>   
#> Compare this with the adequacy of just a general fa...  
#> The degrees of freedom for just the general factor ...  
#> The number of observations was 9 with Chi Square ...  
#> The root mean square of the residuals is 0.05   
#> The df corrected root mean square of the residuals ...  
#>   
#> RMSEA index = 0.401 and the 10 % confidence inter...  
#> BIC = 0.75   
#>   
#> Measures of factor score adequacy   
#> g ...  
#> Correlation of scores with factors 0.98 ...  
#> Multiple R square of scores with factors 0.95 ...  
#> Minimum correlation of factor score estimates 0.90 ...  
#> F2\* ...  
#> Correlation of scores with factors 0.86 ...  
#> Multiple R square of scores with factors 0.74 ...  
#> Minimum correlation of factor score estimates 0.48 ...  
#>   
#> Total, General and Subset omega for each subset  
#> g ...  
#> Omega total for total scores and subscales 0.99 ...  
#> Omega general for total scores and subscales 0.95 ...  
#> Omega group for total scores and subscales 0.04 ...  
#> F2\* F3\*  
#> Omega total for total scores and subscales 0.99 NA  
#> Omega general for total scores and subscales 0.91 NA  
#> Omega group for total scores and subscales 0.08 NA

A fenti elemzéseket jamovi-ban a Factor / Reliability Analysis menüpont segítségével végezhetjük el.

|  |
| --- |
| Megbízhatóság elemzés jamovi-ban |

A fenti megbízhatósági elemzések azt mutatják, hogy a négy tantárgy alfa értéke 0,966, ami egy igen jó érték, hiszen közel van 1-hez (jamovi-ban: Scale Reliability Statistics). Az Item Reliability Statistics táblázat oszlopában szereplő értékek azt mutatják, mi történik, ha egy változót kiveszünk a modellből. Láthatjuk, hogy egyedül a fizika változó értéke növelné az alfát, de a növekedés mértéke elenyésző lenne, tehát nem éri meg eltávolítani a változót, hiszen minél több információnk van egy személyről, annál jobb.

# 4. Feltáró faktorelemzés

A feltáró faktorelemzést új faktorok létrehozására használjuk, a megerősítő faktorelemzést egy meglévő modell tesztelésére (lásd következő fejezet). Ebben a fejezetben faktorelemzés alatt a feltáró faktorelemzést értjük (Rózsa és mtsai., 2019; Watkins, 2018).

A lineáris regresszióelemzéstől eltérően a főkomponens- és a faktorelemzés nagy számú változó kölcsönös összefüggésén alapuló módszer, így nincsenek függő vagy független változóink. Elemzéskor a változók korrelációs mátrixából indulunk ki.

## 4.1 A főkomponens elemzés és a faktorelemzés összehasonlítása

* A főkomponens elemzés során az adatok teljes varianciáját vesszük figyelembe, míg faktorelemzés során a faktorokat csak a közös variancia alapján becsüljük. A két eljárás egyébként nagyon hasonló elvekre épül.
  + Főkomponens elemzés során a korrelációs mátrix átlójában lévő 1-esek összege adja teljes varianciát, ami teljes egészében bekerül a faktormodellbe. Ezért ez az eljárás akkor javasolt, ha a fő cél, hogy meghatározzuk azon főkomponensek (faktorok) legkisebb számát, amelyek a legtöbb varianciát magyarázzák. Ezek a faktorok később jól alkalmazhatók többváltozós elemzésekben. Összegezve: a főkomponens elemzésnél részinformációkat próbálunk összegezni a lehető legkisebb információveszteséggel (vagyis a variancia maximalizálásával).
  + Az ún. közös faktorelemzésnél a faktorokat csak a közös variancia alapján becsüljük, vagyis a kommunalitások kerülnek a korrelációs mátrix átlójába (ezek 1-nél kisebb számok). Összegezve a faktorelemzés általános célja egy látens, lineáris struktúra feltárása manifeszt változók segítségével.

## 4.2 Fogalmak

A főkomponens- és fakorelemzésben a következő fogalmak fordulnak elő (a komponens és a faktor szavak felcserélhetők, attól függően, hogy főkomponens- vagy fakorelemzésről van szó).

* **Kommuninalitás:** a variancia azon hányada, amelyen egy változó osztozik a többi elemzésbe vont változóval. Ez egyben a közös faktorok által magyarázott variancia aránya.
* **Sajátérték:** Az egyes faktorok által magyarázott teljes varianciát fejezi ki.
* **Faktorsúly:** A változók és a faktorok közötti közönséges korreláció.
* **Faktormátrix:** Valamennyi változónak az összes előállított faktorra vonatkozó faktorsúlyát tartalmazza.
* **Faktorértékek:** Az előállított faktoroknak az egyes megkérdezettekre vonatkozóan becsült értékei.
* **Sajátértékábra (scree-plot, kőtörmelék ábra):** A sajátértékek ábrázolása az előállított faktorok sorszámának függvényében.
* **Varianciahányad:** A teljes variancia egy adott faktornak tulajdonított része százalékban kifejezve.

## 4.3 A faktormodell

A főkomponens és faktorelemzés annyiban hasonlít a többszörös regressziószámításhoz, hogy minden változót kifejezhető a háttérben meghúzódó faktorok lineáris kombinációjaként. Minden egyes változó kifejezhető kisszámú közös faktor és egy egyedi faktor segítségével. Ezek a faktorok nem figyelhetők meg közvetlenül. Standardizált kiinduló változók esetén a faktormodell így írható fel:

* , ahol
  + az standardizált változó
  + az változó közös faktorra vonatkozó többszörös standardizált parciális regressziós együtthatója
  + a közös faktor
  + az változó egyedi faktorra vonatkozó többszörös standardizált parciális regressziós együtthatója
  + az változó egyedi faktora
  + a közös faktorok száma.

Az egyedi faktorok egymással és a közös faktorokkal is korrelálatlanok. A közös faktorok kifejezhetők a megfigyelt változók lineáris kombinációiként:

* , ahol
  + az faktor becslése
  + súly vagy a faktorérték együtthatója
  + a változók száma

A súlyokat vagy faktorérték együtthatókat úgy is meg lehet választani, hogy az első faktor magyarázza a teljes variancia legnagyobb részét, a második faktor a második legnagyobb részét és így tovább, valamint, hogy a faktorok korrelálatlanok legyenek egymással. Ez történik főkomponens elemzés esetén.

## 4.4 A faktorelmzés menete

**1. A probléma megfogalmazása**

A kutató arra volt kíváncsi, milyen előnyöket keresnek a fogyasztók a fogrémvásárlásnál. Egy 30 fős mintán a válaszadókat arra kérték, hogy jelezzék, mennyire értenek egyet a következő állításokkal (1 = egyáltalán nem ért egyet; 7 = teljes mértékben egyetért)

* v1: Fontos, hogy olyan fogkrémet vásároljak, amellyel megelőzhető a fogszuvasodás.
* v2: Az olyan fogkrémeket szeretem, amely fényessé teszi a fogaimat.
* v3: Egy fogkrémnek erősítenie kell a fogínyt.
* v4: Az olyan fogkrémeket szeretem, amely friss leheletet biztosít.
* v5: A fog romlásának megelőzése számomra nem fontos elvárás.
* v6: A legfontosabb szempont a fogkrém vásárlásánál a szép fog.

**2. A korrelációs mátrix előállítása**

A korrelációs mátrix előállítása során arra számítunk, hogy azok a változók, amelyek között magas a korreláció, ugyanazzal a faktorral fognak korrelálni.

|  |
| --- |
| Korrelációs mátrix |

A fogkrémvásárlás során keresett előnyök korrelációs mátrix tanulmányozásával látható:

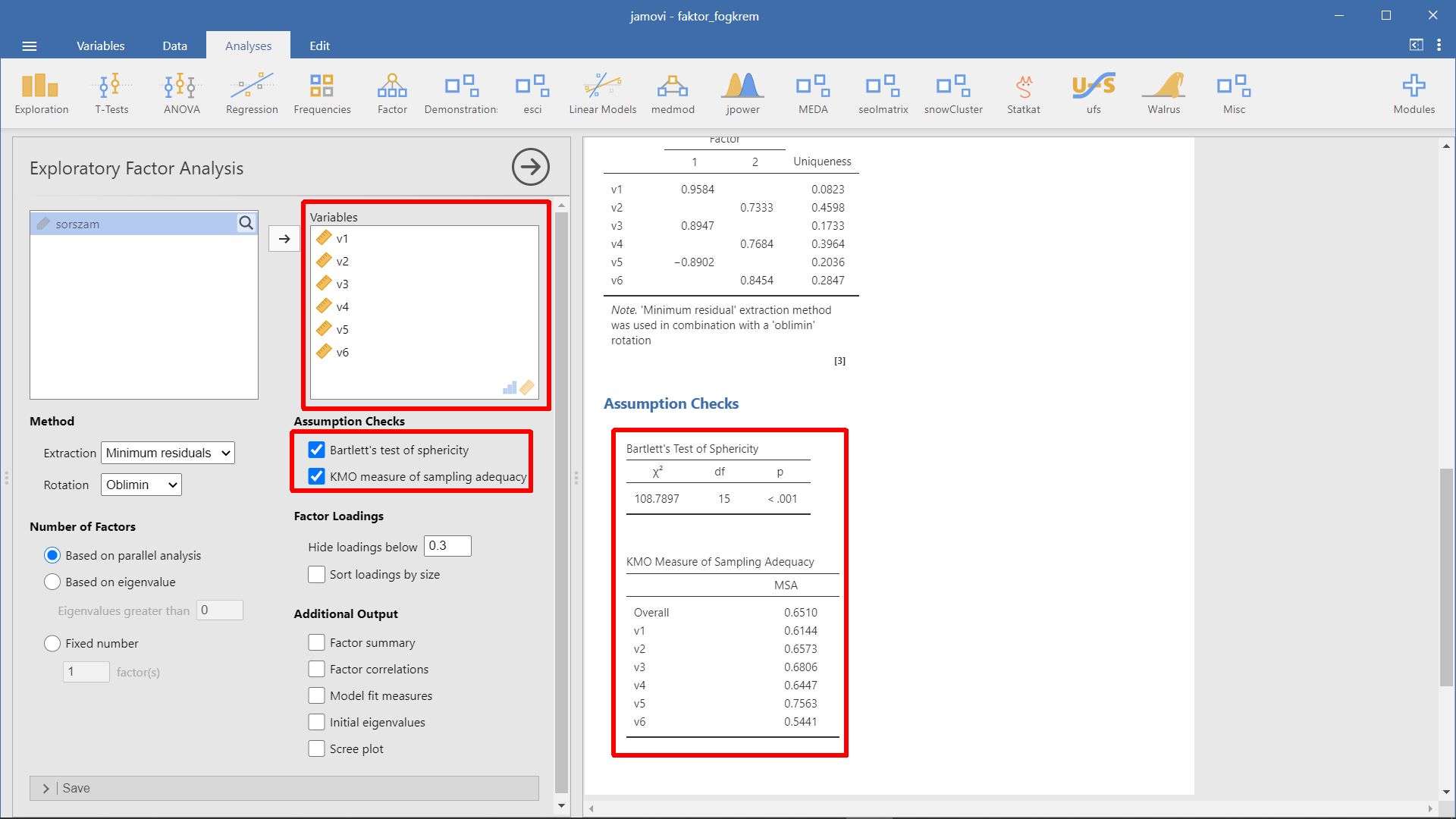
* viszonylag magas a korreláció a v1 (fogszuvasodás megelőzése), v3 (erős fogíny) és v5 (a fog romlásának megelőzése) között. Arra számítunk, hogy ezek a változók ugyanazokkal a faktorokkal fognak korrelálni.
* viszonylag magas a korreláció a v2 (fényes fogak), v4 (friss lehelet) és v6 (szép fogak) változók között, ezek is feltehetőleg ugyanazokkal a faktorokkal fognak korrelálni.

**3. Az alkalmazási feltételek ellenőrzése**

Ahhoz, hogy a faktorelemzés alkalmazható legyen, a változóknak korrelálniuk kell egymással. Erről meggyőződhetünk kétféle objektív módszerrel is:

* Bartlett-féle szferikus próba: nullhipotézise szerint a korrelációs mátrix egységmátrix (a változók korrelálatlanok), azaz az átlón kívül minden elem nulla. Amennyiben a nullhipotézis nem vethető el, a faktorelemzés alkalmazhatósága megkérdőjelezhető.
* Kaiser-Meyer-Olkin-féle megfelelőségi mutató: a megfigyelt korrelációs együtthatók nagyságát viszonyítja a parciális korrelációs együtthatók nagyságához. Az alacsony KMO-mutató azt jelzi, hogy a változópárok közötti korreláció nem magyarázható más változókkal, így a faktorelemzés nem megfelelő módszer. Általában 0,5 fölött érték kívánatos.

Jamovi-ban a Factor / Exploratory Factor Analysis menüpontban tudjuk a fenti vizsgálatokat elvégezni.

 A Bartlett-féle szferikus próba szerint a pupulációban a korrelációs mátrix nem egységmátrix (ez számunkra kedvező), valamint a KMO-érték 0,66, ameéy elég magas (>0,5), így megállapíthatjuk, hogy a faktorelemzés alkalmas módszer a korrelációs mátrix elemzésére.

**4. A faktorelemzés módszerének meghatározása**

A faktorelemzés egyes módszerei abban különböznek egymástól, hogy milyen módon határozzák meg a súlyokat vagy faktorérték együtthatókat.

A jamovi 3 módszert ismer a közös faktorok becslésére:

* Principal axis – főtengelyelemzés
* Minimum residuals
* Maximum likelihood

**5. A faktorok számának meghatározása**

A faktorelemzés akkor ér célt, ha a változók számánál kevesebb számú közös faktort hozunk létre. De mi legyen ez a szám. Több eljárás létezik. Ezeket részletesen a főkomponens elemzés során bemutattuk. Itt csak felsoroljuk őket:

* Horn-féle párhuzamos analízis (jamovi-ban: Based on parallel analysis)
* A priori meghatározás (jamovi-ban: Fixed number)
* Sajátértéken alapuló megoldás (jamovi-ban: Based on eigenvalue)
* Sajátértékábrán (scree-plot, kőtörmelék ábra) alapuló meghatározás (jamovi-ban: Scree plot)
* Magyarázott varianciahányadon alapuló meghatározás (jamovi-ban: Component summary)

**6. A faktorok forgatása**

A faktorelemzés fontos eredménye a faktormátrix (jamovi-ban: Factor Loadings).

|  |
| --- |
| A faktormátrix |

A faktormátrix tartalmazza azokat az együtthatókat, amelyekkel a standardizált változókat ki lehet fejezni a faktorokkal. Ezeket az együtthatókat faktorsúlyoknak nevezzük, a faktorok és a faktorsúlyok közötti korrelációt mutatják. A magas abszolút értékű együttható azt jelzi, hogy a faktor és a változó szorosan összefügg. A faktormátrix együttható alapján lehet a faktorokat értelmezni.

A kiinduló vagy rotálatlan faktormátrix jelzi ugyan az egyes változók és a faktorok kapcsolatát, de ritkán eredményez könnyen értelmezhető faktorokat. Ennek főképp az az oka, hogy a faktorok túl sok változóval korrelálnak. (A rotálatlan faktormátrix beállításához jamovi-ban a Rotation: None beállítást használjuk.)

A faktorok forgatásával a faktormátrix egyszerűbbé, könnyebben értelmezhetővé válik. A faktorok forgatásával szeretnénk elérni:

* minden faktor csak néhány vátozóra rendelkezzen szignifikánsan nem nulla súllyal
* minden változónak lehetőleg egy faktorral legyen nem nulla, azaz szignifikáns faktorsúlya.

A forgatás nem érinti a kommunalitásokat és a magyarázott varianciahányadot, azonban az egy faktor által magyarázott varianciahányad változik (és természetesen a faktorsúlyok is).

A forgatási módszereket érdemes jól megválasztani, mert más-más faktorok azonosításához vezetnek.

Az ortogonális (derékszögű) forgatási eljárások egymással nem korreláló faktorokat eredményeznek.

* Ezek közül az egyik legnépszerűbb a Varimax eljárás, amely minimalizálja a nagy faktorsúllyal rendelkező változók számát, így segíti a faktorok értelmezését. A magyarázott variancia egyenletesen próbálja elosztani a faktorok között.
* A Quartimax eljárás első faktorként egy általános faktort faktort hoz létre, amellyel szinte mindegyik változó magasan korrelál.

A ferdeszögű forgatási eljárások során a tengelyek hegyeszöget zárnak be egymással, és a kapott faktorok korrelálni fognak egymással. Ferdeszögű forgatást akkor kell használni, ha feltételezhető, hogy a sokaságban a faktorok erősen összefüggenek.

* A Promax eljárás gyorsan lefuttatható, amely főképp nagy adatbázisoknál jelent előnyt.
* A Simplimax a Promax egy módosított formája.
* Az Oblimin eljárás a tengelyek egymással bezárt szögét fokozatosan változtatja, ami egyben a faktorok korreláltságát is meghatározza.

**7. A faktorok értelmezése**

Az értelmezést megkönnyíti, ha meghatározzuk azokat a változókat, amelyeknek ugyanazon a faktorra nagy a súlyuk. A faktort a magas faktorsúlyú változók alapján lehet értelmezni.

Az 1. faktornak magasabb az együtthatói a v1 (fugszuvasodás megelőzése), v3 (erős fogíny) változókkal, negatív az együttható a v5 (a fog romlásának megeleőzése nem fontos) változó esetében. Ezt a faktort az “egészséggel kapcsolatos előnyöknek” nevezhetjük. A 2. faktor a v2 (fényes fogak), a v4 (friss lehelet), v6 (szép fogak) változókkal függ össze. Ezt a 2. faktort “társadalmi előnyök”-nek nevezhetjük.

Összegezve, a fogyasztók feltehetőleg két fő előnyt keresnek a fogkrémvásárlás során: egészséggel kapcsolatos és társadalmi előnyöket.

**8. A faktorértékek kiszámítása**

A faktorelemzésnek önmagában is van értelme, hiszen látens változók azonosításához vezet, azonban hasznos lehet a későbbi elemzések számára a faktorértékek kiszámítása minden egyes megkérdezettre. A faktor az eredeti változók lineáris kombinációja. A standardizált változó értékeinek és a megfelelő faktorérték-együtthatónak a szorzata adja a faktorétéket, amely jelen példában minden válaszadóra két faktorértéket jelent. A faktorérték csak főkomponens elemzés esetében lehet pontosan kiszámítani, egyébként csak közelítő értékeket kapunk.

|  |
| --- |
| Faktorértékek kiszámítása |

## 4.5 Illeszkedési mutatók

* **CFI** - összehasonlító illeszkedési mutató (Comparative Fit Index) - A CFI azt méri fel, hogy egy feltételezett hipotetikus modell milyen mértékben reprodukálja a valós adatokon nyugvó kovarianciamátrixot egy független modellhez képest.
* **TLI** - Tucker–Lewis-féle Illeszkedési mutató - A TLI a CFI-hez hasonló módon méri az illeszkedést, annyi különbséggel, hogy ez a mutató a modellben használt szabadságfokot is figyelembe veszi, így kiküszöböli a vizsgálati minta méretének befolyásoló szerepét

A CFI és TLI mutatók értéke 0 és 1 közötti tartományba eshet, ahol az 1-hez közeli érték jelzi a szoros illeszkedést. Kezdetben a mutatók elfogadhatósági kritériumának 0,90-et adtak meg, de az utóbbi időkben inkább a 0,95-ot tekintik az elfogadhatóság alsó határának.

* **RMSEA** - a becslési hiba négyzetes átlagának gyöke (Root-Mean-Square Error of Approximation) - A Steiger-féle RMSEA mutatót a modell populációs kovariancia mátrixhoz viszonyított illeszkedésének becsléséhez használjuk. Az RMSEA az elemszámtól függetlenül hasonlítja össze, hogy a valós és az optimális paraméterekkel rendelkező hipotetikus modell kovarianciamátrixa milyen mértékben illeszkedik. Az RMSEA a modell takarékosságának megbízható jelzője, a komplex modellek hibás specifikálásának hatékony mutatója. Az RMSEA értéke is 0 és 1 közé eshet, itt azonban a kisebb, 0-hoz közel eső érték jelzi a jobb illeszkedést. Browne és Cudeck (1993) szerint az RMSEA értékei 0,05-ig szoros illeszkedést jeleznek; 0,08-os értékig pedig megfelelő illeszkedést, elfogadható populáción belüli becslési hibákkal.

**Model Test**

Az adatok és a teoretikus modell egybeesésének vizsgálata. Az egyik leggyakrabban használt illeszkedési mutató a -próba mértéke, amelyet általában akkor tekinthetünk elfogadhatónak, ha a szabadságfokhoz viszonyított értéke alacsony (pl. kisebb, mint a szabadságfok kétszerese) és nem szignifikáns (p > 0,05). Ennek a mutatónak azonban több korlátja létezik. A legjellemzőbbek a többváltozós normalitás sérülésére és a mintanagyságra való érzékenység. Számos empirikus eredmény és szimulációs vizsgálat támasztja alá, hogy a normalitás sérülésekor vagy nagy elemszámú minta esetében a -próba kevésbé informatív, és a legtöbb esetben a modell elvetését jelzi. A mintanagyságból fakadó korlátot gyakran a -próba szabadságfokhoz mért arányával próbálják kompenzálni (/szabadságfok), amelynek ugyan nincs pontos kritériuma, de az ajánlások általában 2-től 5-ig terjednek, és a határérték alatti érték jelez megfelelő illeszkedést.

|  |
| --- |
| Illeszkedési mutatók |

## 4.6 1. Példa

Valóban szétválasztható a reál és a humán tárgyakhoz szükséges tudás?

Összefoglalva, az adatokra jól illeszkedik a kétfaktoros modell, vagyis azonosíthatjuk a humán és a reál tárgyakat az egyes tantárgyakból nyújtott eredmények alapján. Az egyes tárgyak faktorba történő besorolása összhangban van „hétköznapi”, előzetes tudásunkkal: a matek, informatika és a kémia sorolható a reál, míg az irodalom, nyelvtan és angol tárgyak a humán tárgyakhoz.

## 4.7 2. Példa

Milyen dimenziói vannak a kockázatvállalásnak és változik-e a korral a kockázatvállalás?

A példák is mutatják, hogy a kockázatvállalásokat csoportosíthatjuk a kockázatot jelentő tényezők alapján, ahol az egyik csoportban az emberek saját testi épségüket teszik kockára (mint az autóversenyzés esetében), de kockáztathatnak pénzt vagy valamilyen becsületbeli dolgot is (mint a kártyázás és a blöffölés esetében). Példánkban megnézzük, hogy a faktoranalízis alátámasztja-e feltevésünket, majd a faktoranalízis eredményeit felhasználva megnézzük, hogy a kockázatvállaló viselkedésre hatással van-e a kor.

Az adatbázisban szereplő adatokat úgy kaptuk, hogy a vizsgálati személyeknek különböző foglalkozású, illetve különböző tevékenységet végző embereket kellett megítélniük, hogy mennyire tartják őket szimpatikusnak egy 1-7 skálán, ahol a 7 jelentette azt, hogy nagyon szimpatikus. Ily módon megkaptuk a személyek kockázat iránti attitűdjét. A változók között olyan személyek szerepelnek, mint kártyajátékosok, autóversenyzők, üzletemberek (akik sok pénzt kockáztatnak), veszélyes sportot űző emberek, nagy pénzekben fogadó emberek, blöffölők és hivatásos katonák.

Ezen kívül két adat szerepel az adatbázisban: a nem és a kor.

Összefoglalva, sikerült a faktoranalízissel alátámasztanunk a kockázatvállalás két faktorát. Azt is megállapítottuk, hogy mindkét faktor függ a kortól: az anyagi/erkölcsi téren vállalt kockázat iránti attitűd az évek múlásával egyre kedvezőbbé válik, míg a testi épség terén vállalt kockázat iránti attitűd idővel egyre elutasítóbbá válik.

## 4.8 3. Példa

Egy kérdőív szerkesztésének problémái

Egy jó kérdőív kialakítása hosszas és nagyon alapos munkát igényel. Egyik lépése, hogy megnézzük, az itemek valóban egy dimenzióra illeszkednek-e. A nem odaillő itemeket pedig kivesszük a kérdőívből (itemszelekció). Ennek módszerei lehetnek a korábban már tárgyalt Cronbach-alfa, illetve ha főkomponens-analízist végzünk az itemekkel, akkor a Theta is. A következőkben egy főkomponens-analízisen és Cronbach-alfán alapuló itemszelekcióra nézünk példát.

# 5. Megerősítő faktorelemzés

Tehát az a kísérletünk, hogy az EFA segítségével azonosítsuk a mögöttes látens tényezőket a személyiségi elemkészletből gondosan kiválasztott kérdésekkel, meglehetősen sikeresnek tűnt. A személyiség hasznos mérőszámának kidolgozására irányuló törekvésünk következő lépése az eredeti EFA-ban azonosított látens tényezők ellenőrzése egy másik mintán. Azt akarjuk látni, hogy a tényezők helytállóak-e, meg tudjuk-e igazolni létezésüket különböző adatokkal. Ez egy szigorúbb ellenőrzés, mint látni fogjuk. És ezt Confirmatory Factor Analysisnek (CFA) hívják, mivel nem meglepő módon egy előre meghatározott látens faktorstruktúra megerősítésére törekszünk.

A CFA-ban ahelyett, hogy olyan elemzést végzünk, amelyben az adatok feltáró értelemben vett együttállását látjuk, ehelyett struktúrát írunk elő, mint a 15.13. ábrán, és megnézzük, hogy az adatok mennyire illeszkednek az előre megadott struktúránkhoz. Ebben az értelemben egy megerősítő elemzést végzünk, hogy megnézzük, mennyire erősítik meg a megfigyelt adatok egy előre meghatározott modellt.

A személyiségelemek egyszerű megerősítő faktoranalízise (CFA) ezért öt látens faktort határoz meg, amint azt a 15.13. ábra mutatja, mindegyiket öt megfigyelt változóval mérve. Mindegyik változó egy mögöttes látens tényező mértéke. Például az A1-et a mögöttes látens egyetértési tényező előrejelzi. És mivel A1 nem tökéletes mértéke az elfogadhatósági tényezőnek, van egy hibatag, , kapcsolódik hozzá. Más szavakkal, az A1 eltérését jelenti, amelyet az egyetértési tényező nem vesz figyelembe. Ezt néha mérési hibának nevezik.

# 6. Többszempontos varianciaelemzés

JÖN.

# 7. Klaszterelemzés

Az klaszterelemzés segítségével vizsgált probléma a következő: egy elemű adatbázisban minden egyes elemhez darab változó értékei kapcsolódnak; alakítsunk az elemekből csoportokat úgy, hogy a “hasonlóak” egy csoportba kerüljenek. Minden klaszter elemei viszonylag hasonlóak legyenek egymáshoz, de különbözzenek más klaszterek elemeitől.

Számos eljárás született a klaszteranalízis módszerén belül. Ebben a könyvben két eljárással foglalkozunk részletesen:

* a hierarchikus eljárásokkal és a
* K-középpontú klaszteranalízissel.

## 7.1 Hierarchikus eljárások

A hierarchikus eljárások az egyes személyek, objektumok, esetek közötti távolság meghatározásával kezdődnek. A csoportok, klaszterek kialakítása történhet összevonáson vagy felosztáson alapuló módszerekkel. Az összevonó módszerek abból indulnak ki, hogy minden egyes elem egy önálló csoportot alkot, majd fokozatosan vonják össze az egyelemes csoportokat egyetlen nagy csoportba. Ezzel szemben a lebontó módszerben az összes elem egyetlen csoportba tartozik, és ezt a csoportot osztjuk fel kettő, majd egyre több csoportra addig, amíg minden elem egy önálló csoportot nem alkot. Az összevonó módszernél kezdetben minden egyes elem külön klasztert alkot. A klaszterek a megfigyelési egységek egyre nagyobb klaszeterekbe csoportosításával alakulnak ki. A folyamat addig folytatódik, amíg egyetlen klaszter lesz az egész. A következőkben a legelterjedtebb klaszteralkotó módszereket soroljuk fel:

* Egyszerű lánc, avagy a legközelebbi szomszéd elve
* Teljes lánc, avagy a legtávolabbi szomszéd elve
* Átlagos távolság
* Variancia-módszerek
* Ward-féle eljárás
* Centroidmódszerek
* Szekvenciális küszöbérték módszer
* Párhuzamos küszöbérték módszer
* Optimális felosztás módszere

Jamovi-ban a snowCluster csomag segítségével végethetünk klaszterelemzéseket. A csomag telepítése után a snowCluster / Hiearchical Clustering vagy snowCluster / Clustering Dendogram menüpontokat használjuk.

A jamovi a következő módszereket ismeri:

* ward.D
* ward.D2
* single
* complete
* average

## 7.2 K-középpontú klaszterelemzés

JÖN.

# 8. Diszkriminancia elemzés

A diszkriminancia analízisben azt a problémát járjuk körül, hogyan lehet az emberek egyes csoportjait valamilyen vizsgált jellemzők alapján szétválasztani, az egyes csoportokat azonosítani, valamint a csoporttagságokat az előbb említett vizsgált jellemzők alapján előrejelezni.

Képzeljünk el, hogy szalagmunkásokkal végeznek alkalmasság-vizsgálatot. A szalagmunka általában sok figyelmet igényel, ugyanakkor meglehetősen monoton munka, éppen ezért jó figyelmi képességek és monotónia tűrés szükséges hozzá. A lenti szalagmunka adatmátrix 10 személy adatát tartalmazza.

szalagmunka <- rio::import(file = "adat/diszkriminancia\_alkalmassag.xlsx")  
szalagmunka$bevalt <- factor(szalagmunka$bevalt, labels = c("nem",  
 "igen"))  
szalagmunka  
#> bevalt figyelem monotonia\_tures  
#> 1 igen 1 2  
#> 2 igen 1 5  
#> 3 nem 2 1  
#> 4 igen 2 3  
#> 5 nem 3 2  
#> 6 igen 3 4  
#> 7 nem 4 3  
#> 8 nem 4 1  
#> 9 igen 4 6  
#> 10 nem 6 5

A fenti adatok egy részét a szalagmunkára való jelentkezéskor gyűjtötték:

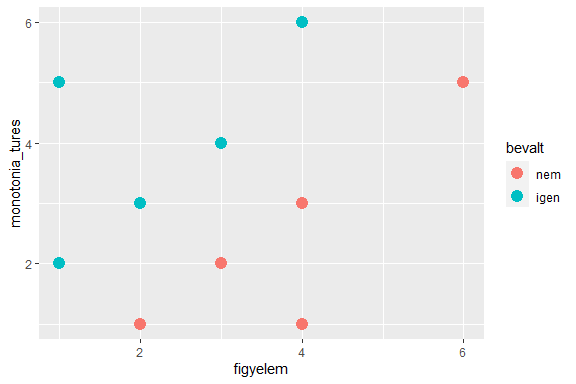
* figyelem: figyelmi képességükre és
* monotonia\_tures: monotónia-tűrésükre vonatkozó információk.

Mindkét változót 7 fokú skálán értékeltek (ahol a magasabb érték jobb képességeket jelent), valamint azt az információt is láthatjuk, hogy később beváltak-e vagy sem (bevalt).

Azt kellene megmutatnunk, hogy a figyelem és a monotónia-tűrés pontszámai alapján valóban lehet következtetést levonni a személy beválását illetően. Ha ezt sikerül egy objektív statisztikai módszerrel is igazolnunk, akkor az ezt követően szalagmunkára jelentkezőket figyelem és monotónia tűréssel vizsgálva tesztelhetjük, és egész jól ki lehet válogatni az alkalmasabb jelölteket.

Ha pontdiagramon ábrázoljuk az adatokat, és színezéssel jelöljük a beválást, akkor a két csoport szemmel láthatóan szétválik egymástól, ám sem a függőleges, sem a vízszintes tengely mentén nem lehet elkülöníteni a csoportokat.

library(ggplot2)  
ggplot(szalagmunka, aes(x = figyelem, y = monotonia\_tures, colour = bevalt)) +  
 geom\_point(size = 4)



**A diszkriminancia elemzés sajátossága, hogy a csoportokat a magyarázó változók együttes figyelembevételével tudja szétválasztani.** Ennek megfelelően ha önmagában tekintjük az egyik (például figyelem) vagy másik (monotonia\_tures) magyarázó változókat, akkor nem tudunk szignifikáns különbséget kimutatni a bevalt változó két csoportjában .

t.test(figyelem ~ bevalt, data = szalagmunka)$p.value  
#> [1] 0.1082333  
t.test(monotonia\_tures ~ bevalt, data = szalagmunka)$p.value  
#> [1] 0.1588974

Nézzük meg, hogy többváltozós variancia-analízissel (MANOVA) szét tudjuk-e választani a csoportokat, amikor a két magyarázó változót egyszerre vesszük figyelembe. Továbbra is arra keressük a választ, hogy beválás tekintetében valóban létezik-e a munkások két csoportja.

man\_1 <- manova(cbind(figyelem, monotonia\_tures) ~ bevalt, data = szalagmunka)  
summary(man\_1, test = "Wilks")  
#> Df Wilks approx F num Df den Df Pr(>F)   
#> bevalt 1 0.2708 9.4247 2 7 0.01033 \*  
#> Residuals 8   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A fenti output alapján megnyugodhatunk, a bevált és a nem bevált munkások csoportja valóban különbözik egymástól .

A diszkriminancia elemzéstől azonban ettől többet fogunk várni, például predikciót is végezhetünk a kapott modellben.

Ha diszkriminancia elemzést szeretnénk végrehajtani, akkor a {MASS} csomag lda() függvényét használhatjuk.

lda\_1 <- MASS::lda(bevalt ~ figyelem + monotonia\_tures, data = szalagmunka)  
lda\_1  
#> Call:  
#> lda(bevalt ~ figyelem + monotonia\_tures, data = sza...  
#>   
#> Prior probabilities of groups:  
#> nem igen   
#> 0.5 0.5   
#>   
#> Group means:  
#> figyelem monotonia\_tures  
#> nem 3.8 2.4  
#> igen 2.2 4.0  
#>   
#> Coefficients of linear discriminants:  
#> LD1  
#> figyelem -0.9981001  
#> monotonia\_tures 0.8365579

A fenti outputban látható, hogy akik nem váltak be, a monotónia-tűrés tesztben gyengébb teljesítményt nyújtottak, a figyelem tesztben pedig egy jobbat, míg akik beváltak, a monotónia-tűrés tesztben igen jó teljesítményt értek el, a figyelem tesztben pedig valamivel gyengébbet. A kanonikus diszkriminancia együtthatókat is láthatjuk, melyek alapján felírhatjuk a kanonikus diszkriminancia-függvényt a következő módon:

Z = -0,998 \* figyelem + 0,837 \* monotonia\_tures

Az eddig vizsgált 10 személyről tudjuk, hogy bevált-e vagy sem, vagyis ismertük a tényleges csoporttagságát. Ám a diszkriminancia elemzés fontos célja, hogy előre jelezzük a csoporttagságokat, vagyis a figyelem és monotónia-tűrés ismeretében megmondjuk egy személyről, hogy nagy valószínűséggel beválik-e vagy sem.

Tegyük fel, hogy az első személy a figyelem teszten 2, míg a monotónia-tűrés teszten 4 pontot kapott, míg a második személy pontszámai ebben a sorrendben 6 és 1.

newdata <- data.frame(figyelem = c(2, 6), monotonia\_tures = c(4,  
 1))  
newdata  
#> figyelem monotonia\_tures  
#> 1 2 4  
#> 2 6 1  
lda\_1\_pred <- predict(lda\_1, newdata = newdata)  
lda\_1\_pred  
#> $class  
#> [1] igen nem   
#> Levels: nem igen  
#>   
#> $posterior  
#> nem igen  
#> 1 0.00743262 9.925674e-01  
#> 2 0.99999931 6.861874e-07  
#>   
#> $x  
#> LD1  
#> 1 1.667346  
#> 2 -4.834728

A fenti output class és posterior része alapján láthatjuk, hogy az első személyt nagy valószínűséggel alkalmasnak, míg a másodikat alkalmatlannak ítélhetjük a szalagmunkára.

Utolsó lépésként összevetjük a tényleges és a becsült csoporttagságot, és megállapítjuk, az adatok mekkora részét tudjuk helyesen besorolni az alkotott modell alapján. Ezzel magát a modellt értékeljük.

lda\_2\_pred <- predict(lda\_1, method = "plug-in")  
tab\_1 <- table(lda\_2\_pred$class, szalagmunka$bevalt)  
tab\_1  
#>   
#> nem igen  
#> nem 5 0  
#> igen 0 5

Mivel az összes adat a főátlóban van, így megállapíthatjuk, hogy a modell alapján az összes adatot helyesen kategorizáltuk. A helyes besorolás arányát százalékosan is kiszámíthatjuk, ez az arány 100%.

100 \* sum(diag(tab\_1))/sum(tab\_1)  
#> [1] 100

## 8.1 Példa: Kikből lesznek a balesetezők?

Ebben a példában azt vizsgáljuk, mely tényezők járulnak hozzá a balesetekhez.

baleset <- rio::import(file = "adat/diszkriminancia\_baleset.xlsx")  
baleset$baleset <- factor(baleset$baleset, labels = c("nem volt baleste",  
 "volt baleste"))  
str(baleset)  
#> 'data.frame': 36 obs. of 5 variables:  
#> $ baleset : Factor w/ 2 levels "nem volt baleste...  
#> $ megosztott: num 7 6 5 6 7 3 6 7 5 6 ...  
#> $ pontossag : num 6 6 5 6 7 3 5 7 5 6 ...  
#> $ kockazat : num 2 3 1 2 4 7 2 1 3 2 ...  
#> $ eszleles : num 7 6 5 6 7 7 7 6 3 7 ...  
psych::headTail(baleset)  
#> baleset megosztott pontossag kockazat  
#> 1 nem volt baleste 7 6 2  
#> 2 nem volt baleste 6 6 3  
#> 3 nem volt baleste 5 5 1  
#> 4 nem volt baleste 6 6 2  
#> ... <NA> ... ... ...  
#> 33 volt baleste 3 3 5  
#> 34 volt baleste 2 2 7  
#> 35 volt baleste 3 3 4  
#> 36 volt baleste 4 4 6  
#> eszleles  
#> 1 7  
#> 2 6  
#> 3 5  
#> 4 6  
#> ... ...  
#> 33 4  
#> 34 1  
#> 35 4  
#> 36 4

Az adatbázisban a baleset változó azt rögzíti, hogy volt-e már balesete a személynek vagy sem. Ez lesz tehát a csoportosító változó. A többi változó, melyek segítségével próbáljuk a csoportok közötti különbséget jellemezni, olyan dolgot mérnek, mint a megosztott figyelem (megosztott változó), a figyelem pontossága (pontossag), kockázatvállalási hajlandóság (kockazat) és az észlelés gyorsasága (eszleles).

A diszkriminancia-analízisben az első lépés annak megállapítása, vajon valóban szét lehet-e választani a balesetezők és a nem balesetezők csoportját az adott változók alapján. Ehhez a Wilks-lambda tesztet használjuk a többváltozós variancia-analízis keretein belül.

man\_1 <- manova(cbind(megosztott, pontossag, kockazat, eszleles) ~  
 baleset, data = baleset)  
summary(man\_1, test = "Wilks")  
#> Df Wilks approx F num Df den Df Pr(>F)  
#> baleset 1 0.27605 20.324 4 31 2.645e-08  
#> Residuals 34   
#>   
#> baleset \*\*\*  
#> Residuals   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A fenti output tesztstatisztikájának szignifikanciaszintje azt mutatja, hogy a csoportok közötti különbségek szignifikánsak, vagyis valóban van különbség a balesetet szenvedett és a balesetmentes autóvezetők között.

Futtassuk le a diszkriminancia-analízis.

lda\_1 <- MASS::lda(baleset ~ megosztott + pontossag + kockazat +  
 eszleles, data = baleset)  
lda\_1  
#> Call:  
#> lda(baleset ~ megosztott + pontossag + kockazat + e...  
#>   
#> Prior probabilities of groups:  
#> nem volt baleste volt baleste   
#> 0.4722222 0.5277778   
#>   
#> Group means:  
#> megosztott pontossag kockazat eszl...  
#> nem volt baleste 5.941176 5.647059 2.588235 5.94...  
#> volt baleste 2.842105 2.684211 5.578947 3.10...  
#>   
#> Coefficients of linear discriminants:  
#> LD1  
#> megosztott -0.25764616  
#> pontossag -0.07708289  
#> kockazat 0.36270659  
#> eszleles -0.36702363

A fenti output alapján az előzetes valószínűsége annak, hogy valakinek még nem volt balesete 0,472, míg annak a valószínűsége, hogy már volt balesete a személynek 0,528. Ezután vizsgálhatjuk a csoportátlagokat. A balesetmentes vezetők esetében magasabb a megosztott figyelem, a figyelem pontosságának és az észlelés változójának az átlaga, míg a kockázatvállalásé alacsonyabb. Ugyanakkor a másik csoport esetében a kockázatvállalás változójának az átlaga magasabb, míg a másik három képesség változójának átlaga alacsonyabb. Vagyis a balesetmentes vezetők gyorsabban képesek észlelni és jobban meg tudják osztani a figyelmüket, figyelmük pontosabb. A balesetet szenvedett vezetők esetében ezek a képességek gyengébbek, míg jobban szeretnek kockázatot vállalni.

Végül a kanonikus diszkriminancia együtthatók segítségével felírhatjuk a kanonikus diszkriminancia-függvényt a következő módon:

Z = 0,3627 \* kockázat - 0,367 \* észlelés - 0,2567 \* megosztott-0,0771 \* pontosság

Utolsó lépésként pedig megnézhetjük, mennyire hatékony a diszkriminancia-analízis vagyis összevethetjük az eredeti csoporttagságokat a modell alapján alkotott besorolásokkal.

lda\_1\_pred <- predict(lda\_1, method = "plug-in")  
tab\_1 <- table(lda\_1\_pred$class, baleset$baleset)  
tab\_1  
#>   
#> nem volt baleste volt baleste  
#> nem volt baleste 16 2  
#> volt baleste 1 17  
100 \* sum(diag(tab\_1))/sum(tab\_1)  
#> [1] 91.66667

A fenti sorok elkészík a predikciót, majd egy táblázatban reprezentálják az eredeti és a becsült csoportba tartozásokat. A legtöbb adat a főátlóban helyezkedik el, ami igen magas helyes besorolási arányra utal. A helyes besorolások aránya 91,7%.

A példában a gépjárműbalesetek emberi okait vizsgáltuk. Az eredmények alapján a balesetmentes vezetők gyorsabban képesek észlelni és jobban meg tudják osztani a figyelmüket, figyelmük pontosabb is. Ellenben a balesetet szenvedett vezetők esetében ezek a képességek gyengébbek, míg jobban szeretnek kockázatot vállalni.

## 8.2 Példa: A szülés utáni depresszió vizsgálata

Ebben a példában a szülés utáni depresszió pszichés és szociális hátterét vizsgáljuk meg a diszkriminancia-analízis segítségével.

depresszio <- rio::import(file = "adat/diszkriminancia\_depresszio.xlsx")  
depresszio$ppdepresszio <- factor(depresszio$ppdepresszio, labels = c("nincs depresszió",  
 "van depresszió"))  
str(depresszio)  
#> 'data.frame': 20 obs. of 5 variables:  
#> $ ppdepresszio: Factor w/ 2 levels "nincs depressz...  
#> $ szeretet : num 7 6 2 6 7 4 6 7 6 7 ...  
#> $ tulvedes : num 4 2 8 3 9 5 3 5 3 5 ...  
#> $ kor : num 24 20 19 22 23 25 26 18 19 22...  
#> $ iskola : num 12 17 8 16 17 17 12 17 16 17 ...  
psych::headTail(depresszio)  
#> ppdepresszio szeretet tulvedes kor iskola  
#> 1 nincs depresszió 7 4 24 12  
#> 2 nincs depresszió 6 2 20 17  
#> 3 nincs depresszió 2 8 19 8  
#> 4 nincs depresszió 6 3 22 16  
#> ... <NA> ... ... ... ...  
#> 17 van depresszió 4 7 30 6  
#> 18 van depresszió 4 6 24 8  
#> 19 van depresszió 3 7 21 9  
#> 20 van depresszió 2 8 18 10

Az adatbázisban a ppdepresszio változó mutatja a depresszió jelenlétét, vagy hiányát. A magyarázó változók között a következő változók szerepelnek: a szeretet skála (szeretet változó), amely azt méri, hogy a személyek mennyire érzik, hogy a szüleik szeretik őket; tulvedes-sel jelölt túlvédés iránti tendencia azt mutatja, hogy mennyire hajlamosak arra a személyek, hogy túlságosan is burokban tartsák, túlvédjék gyerekeiket, illetve szeretteiket; ezeken kívül két szociológiai adat is a rendelkezésünkre áll, nevezetesen az életkor (kor változó) és az elvégzett iskolai osztályok száma (iskola).

A diszkriminancia-analízisben az első lépésében megvizsgáljuk, vajon valóban szét lehet-e választani a depressziósok és a nem depressziósok csoportját az adott változók alapján. Ehhez a Wilks-lambda tesztet használjuk a többváltozós variancia-analízis keretein belül.

man\_1 <- manova(cbind(szeretet, tulvedes, kor, iskola) ~ ppdepresszio,  
 data = depresszio)  
summary(man\_1, test = "Wilks")  
#> Df Wilks approx F num Df den Df P...  
#> ppdepresszio 1 0.29985 8.7561 4 15 0.00...  
#> Residuals 18 ...  
#>   
#> ppdepresszio \*\*\*  
#> Residuals   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A fenti output tesztstatisztikájának szignifikanciaszintje azt mutatja, hogy a csoportok közötti különbségek szignifikánsak, vagyis valóban van különbség a depressziós és a nem depressziós nők között az adott változókat vizsgálva.

Végezzük el a diszkriminancia elemzést!

library(MASS)  
lda\_1 <- lda(ppdepresszio ~ szeretet + tulvedes + kor + iskola,  
 data = depresszio)  
lda\_1  
#> Call:  
#> lda(ppdepresszio ~ szeretet + tulvedes + kor + isko...  
#>   
#> Prior probabilities of groups:  
#> nincs depresszió van depresszió   
#> 0.5 0.5   
#>   
#> Group means:  
#> szeretet tulvedes kor iskola  
#> nincs depresszió 5.8 4.7 21.8 14.9  
#> van depresszió 3.3 7.5 24.0 8.3  
#>   
#> Coefficients of linear discriminants:  
#> LD1  
#> szeretet -0.21900671  
#> tulvedes 0.18422053  
#> kor 0.03467147  
#> iskola -0.26661705

Az fenti output alapján az előzetes valószínűségek egyenlőek. A csoportátlagok közötti különbségek azt mutatják, hogy a nem depressziósok átlaga szeretet tekintetében magasabb (5,8), mint a depressziósoké (3,3), az iskolai végzettségük is magasabb (14,9), mint a depressziósoké (8,3). Ellenben a túlvédésnél a depressziósok értek el magasabb átlagot, ők az idősebbek is (24). Vagyis azok, akik postpartum depresszióban szenvednek, úgy érzik, a szüleik kevésbé szeretik őket, túlvédőbbek a gyerekeikkel szemben, idősebbek, és az iskolai végzettségük is alacsonyabb. A kanonikus diszkriminancia egyenlet alakja:

Z =0,1842 \* túlvédés + 0,0347 \* kor - 0,2666 \* iskola - 0,219 \* szeretet

Utolsó momentumként az analízis értékelésére van még szükség.

lda\_1\_pred <- predict(lda\_1, method = "plug-in")  
tab\_1 <- table(lda\_1\_pred$class, depresszio$ppdepresszio)  
tab\_1  
#>   
#> nincs depresszió van depresszió  
#> nincs depresszió 9 0  
#> van depresszió 1 10  
100 \* sum(diag(tab\_1))/sum(tab\_1)  
#> [1] 95

Láthatjuk, hogy a valódi és a modell alapján becsült csoporttagságok mátrixában a legtöbb adat a főátlóban helyezkedik el. Ez arra utal, hogy a becsült csoporttagságok nagyjából lefedik az eredetit, az arány 95%.

Vagyis azok, akik postpartum depresszióban szenvednek, úgy érzik, a szüleik kevésbé szeretik őket, túlvédőbbek a gyerekeikkel szemben, idősebbek, és az iskolai végzettségük is alacsonyabb.

## 8.3 Példa: Pszichoszomatikus megbetegedésekre hajlamosító tényezők

Ebben a példában a pszichoszomatikus megbetegedéseket vizsgáljuk a diszkriminancia-analízis segítségével.

pszichoszomatikus <- rio::import(file = "adat/diszkriminancia\_pszichoszomatika.xlsx")  
pszichoszomatikus$pszichoszomatika <- factor(pszichoszomatikus$pszichoszomatika,  
 labels = c("pszichoszomatikus megbetegedése van", "egészséges"))  
str(pszichoszomatikus)  
#> 'data.frame': 36 obs. of 4 variables:  
#> $ pszichoszomatika: Factor w/ 2 levels "pszichoszo...  
#> $ stressz : num 5 6 5 6 7 3 6 7 5 6 ...  
#> $ szorongas : num 6 6 5 6 7 3 3 7 5 6 ...  
#> $ coping : num 2 3 1 2 4 7 2 1 3 2 ...  
psych::headTail(pszichoszomatikus)  
#> pszichoszomatika stressz  
#> 1 pszichoszomatikus megbetegedése van 5  
#> 2 pszichoszomatikus megbetegedése van 6  
#> 3 pszichoszomatikus megbetegedése van 5  
#> 4 pszichoszomatikus megbetegedése van 6  
#> ... <NA> ...  
#> 33 egészséges 3  
#> 34 egészséges 2  
#> 35 egészséges 3  
#> 36 egészséges 4  
#> szorongas coping  
#> 1 6 2  
#> 2 6 3  
#> 3 5 1  
#> 4 6 2  
#> ... ... ...  
#> 33 3 5  
#> 34 2 7  
#> 35 3 4  
#> 36 4 6

Az adatbázisban most a pszichoszomatika változó jelzi, hogy valamilyen pszichoszomatikus megbetegedése van vagy nincs a személynek. A változók közt szerepel a személyt ért stressz mértéke (stressz), a szorongási szintje (szorongás) és a megküzdési stratégiáinak hatékonysága (coping).

A diszkriminancia-analízisben az első lépésében megvizsgáljuk, vajon valóban szét lehet-e választani a pszichoszomatikusok és a nem pszichoszomatikusok csoportját az adott változók alapján. Ehhez a Wilks-lambda tesztet használjuk a többváltozós variancia-analízis keretein belül.

man\_1 <- manova(cbind(stressz, szorongas, coping) ~ pszichoszomatika,  
 data = pszichoszomatikus)  
summary(man\_1, test = "Wilks")  
#> Df Wilks approx F num Df den Df  
#> pszichoszomatika 1 0.37974 17.423 3 32  
#> Residuals 34   
#> Pr(>F)   
#> pszichoszomatika 6.92e-07 \*\*\*  
#> Residuals   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A fenti output tesztstatisztikájának szignifikanciaszintje azt mutatja, hogy a csoportok közötti különbségek szignifikánsak, vagyis valóban van különbség a két csoport között az adott változókat vizsgálva.

Végezzük el a diszkriminancia elemzést.

library(MASS)  
lda\_1 <- lda(pszichoszomatika ~ stressz + coping + szorongas,  
 data = pszichoszomatikus)  
lda\_1  
#> Call:  
#> lda(pszichoszomatika ~ stressz + coping + szorongas...  
#>   
#> Prior probabilities of groups:  
#> pszichoszomatikus megbetegedése van   
#> 0.4722222   
#> egészséges   
#> 0.5277778   
#>   
#> Group means:  
#> stressz coping  
#> pszichoszomatikus megbetegedése van 5.764706 2.588235  
#> egészséges 2.842105 5.578947  
#> szorongas  
#> pszichoszomatikus megbetegedése van 5.529412  
#> egészséges 2.684211  
#>   
#> Coefficients of linear discriminants:  
#> LD1  
#> stressz -0.31309547  
#> coping 0.46637406  
#> szorongas -0.06258674

A fenti outputból láthatjuk, hogy a csoporttagságok előzetes valószínűsége a pszichoszomatikusok esetében kicsit kisebb (0,472). A két csoport összevetésénél azt láthatjuk, hogy a stressz és a szorongás változó átlaga a pszichoszomatikusok esetében, míg a coping változó átlaga az egészségesen esetében magasabb. Vagyis az egészséges személyeket kevesebb stressz éri, és azzal hatékonyabban is tudnak megküzdeni, mint a pszichoszomatikusok, illetve kevesebbet is szoronganak.

A kanonikus diszkriminancia egyenlet pedig a következő módon alakul:

Z = 0.4664 \* coping - 0,3131 \* stressz - 0,0626 \* szorongas

Utolsó lépésként az analízis értékelésére van még szükség.

lda\_1\_pred <- predict(lda\_1, method = "plug-in")  
tab\_1 <- table(lda\_1\_pred$class, pszichoszomatikus$pszichoszomatika)  
tab\_1  
#>   
#> pszichoszomat...  
#> pszichoszomatikus megbetegedése van ...  
#> egészséges ...  
#>   
#> egészséges  
#> pszichoszomatikus megbetegedése van 2  
#> egészséges 17  
100 \* sum(diag(tab\_1))/sum(tab\_1)  
#> [1] 91.66667

Az eredményen láthatjuk, hogy a valódi és a modell alapján becsült csoporttagságok mátrixában a legtöbb adat a főátlóban helyezkedik el. Ez arra utal, hogy a becsült csoporttagságok nagyjából lefedik az eredetit. Ez az arány 91,7%.

Ebben a példában a pszichoszomatikus megbetegedések lelki okait vizsgáltuk. Az diszkriminancia-analízis eredménye szerint az egészséges személyeket kevesebb stressz éri, és azzal hatékonyabban is tudnak megküzdeni, mint a pszichoszomatikusok, valamint kevesebbet is szoronganak.

## 8.4 Példa: Kik vásárolnak gyakran bio termékeket?

Utolsó példánk a marketingkutatás területére kalauzol minket. Azt próbáljuk megvizsgálni, hogy főként kik vásárolnak bio termékeket.

bio <- rio::import(file = "adat/diszkriminancia\_bio.xlsx")  
bio$vasarlas <- factor(bio$vasarlas, labels = c("soha nem vesz",  
 "időnként vesz", "gyakran vesz"))  
table(bio$vasarlas)  
#>   
#> soha nem vesz időnként vesz gyakran vesz   
#> 10 10 10  
str(bio)  
#> 'data.frame': 30 obs. of 5 variables:  
#> $ vasarlas: Factor w/ 3 levels "soha nem vesz",..:...  
#> $ ertek : num 2 4 2 3 1 1 2 3 4 6 ...  
#> $ attitud : num 2 4 2 3 1 3 5 3 1 2 ...  
#> $ fizetes : num 55 67 89 78 99 112 132 78 95 64 ...  
#> $ kor : num 32 56 59 48 44 39 37 40 44 43 ...  
psych::headTail(bio)  
#> vasarlas ertek attitud fizetes kor  
#> 1 soha nem vesz 2 2 55 32  
#> 2 soha nem vesz 4 4 67 56  
#> 3 soha nem vesz 2 2 89 59  
#> 4 soha nem vesz 3 3 78 48  
#> ... <NA> ... ... ... ...  
#> 27 gyakran vesz 6 6 62 19  
#> 28 gyakran vesz 9 9 69 27  
#> 29 gyakran vesz 9 9 78 28  
#> 30 gyakran vesz 8 8 73 30

Az adatbázisban a vasarlas változó mutatja a biotermékek vásárlásának gyakoriságát, amely három értéket vehet fel: a személy szinte soha nem vesz ilyen termékeket, időnként vesz, illetve gyakran vesz. A vásárlás gyakoriságát a következő változókkal próbáljuk előre jelezni: milyen értékeket tulajdonít ezeknek a termékeknek (ertek változó, minél nagyobb pontszámot kap a skálán, annál jobban értékeli a személy a bio termékeket); az attitud skála a termékek iránti attitűdöt méri, a magasabb értékek itt is kedvezőbb atttitűdöt jeleznek; ezen túl szerepel még a személy életkora (kor változó) és a fizetése is (fizetes).

A diszkriminancia-analízisben az első lépésében megvizsgáljuk, vajon valóban szét lehet-e választani a bio termékeket vásárlók három csoportját az adott változók alapján. Ehhez a Wilks-lambda tesztet használjuk a többváltozós variancia-analízis keretein belül.

man\_1 <- manova(cbind(ertek, attitud, fizetes, kor) ~ vasarlas,  
 data = bio)  
summary(man\_1, test = "Wilks")  
#> Df Wilks approx F num Df den Df Pr(>F)  
#> vasarlas 2 0.12109 11.242 8 48 8.599e-09  
#> Residuals 27   
#>   
#> vasarlas \*\*\*  
#> Residuals   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
summary.aov(man\_1)  
#> Response ertek :  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> vasarlas 2 146.07 73.033 27.656 2.915e-07 \*\*\*  
#> Residuals 27 71.30 2.641   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Response attitud :  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> vasarlas 2 211.67 105.833 57.495 1.853e-10 \*\*\*  
#> Residuals 27 49.70 1.841   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Response fizetes :  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> vasarlas 2 6427.5 3213.7 3.1857 0.05726 .  
#> Residuals 27 27237.5 1008.8   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Response kor :  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> vasarlas 2 1449.3 724.63 7.3001 0.002922 \*\*  
#> Residuals 27 2680.1 99.26   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A fenti elemzés tesztstatisztikájának szignifikanciaszintje azt mutatja, hogy a csoportok közötti különbségek szignifikánsak, vagyis valóban van különbség a három csoport között az adott változókat vizsgálva.

Végezzük el a diszkriminancia elemzést.

library(MASS)  
lda\_1 <- lda(vasarlas ~ ertek + attitud + fizetes + kor, data = bio)  
lda\_1  
#> Call:  
#> lda(vasarlas ~ ertek + attitud + fizetes + kor, dat...  
#>   
#> Prior probabilities of groups:  
#> soha nem vesz időnként vesz gyakran vesz   
#> 0.3333333 0.3333333 0.3333333   
#>   
#> Group means:  
#> ertek attitud fizetes kor  
#> soha nem vesz 2.8 2.6 86.9 44.2  
#> időnként vesz 5.3 5.6 106.5 34.9  
#> gyakran vesz 8.2 9.1 70.7 27.2  
#>   
#> Coefficients of linear discriminants:  
#> LD1 LD2  
#> ertek 0.278041839 -0.175361797  
#> attitud 0.578431017 0.066376341  
#> fizetes -0.003687657 -0.032311789  
#> kor -0.019282287 0.003972177  
#>   
#> Proportion of trace:  
#> LD1 LD2   
#> 0.9687 0.0313  
# klaR::greedy.wilks(vasarlas~ertek+attitud+fizetes+kor,data=bio,  
# niveau = 0.15)

A fenti output alapján az elemzés elején a három vásárlási gyakoriság valószínűsége egyenlő (0,333). Ha a csoportátlagokat vizsgáljuk akkor láthatjuk, hogy mind az értékek, mind az attitűd változójának tekintetében a soha sem vásárolók átlaga a legalacsonyabb (3 mindkét változó esetében), az időnként bio termékeket vásárlók csoport átlaga középen helyezkedik el mind a két változó esetében (5 és 6), és a gyakran vásárlók átlaga a legmagasabb (8 és 9). Életkor tekintetében egy kissé másképpen alakulnak a csoportok. A legidősebbek szinte sohasem vásárolnak bio termékeket, a legfiatalabbak pedig igen gyakran vásárolnak. Fizetés tekintetében nem figyelhető meg jól magyarázható összefüggés: a legalacsonyabb fizetésűek gyakran, míg a közepes fizetésűek szinte soha sem vásárolnak bio termékeket.

A két kanonikus diszkriminancia-egyenlet a következőképpen alakul:

Z1 = 0,278 \* ertek + 0,578 \* attitud - 0,019 \* kor - 0,004 \*fizetes  
Z2 = -0,175 \* ertek + 0,066 \* attitud + 0,004 \* kor - 0,032 \*fizetes

Utolsó lépésként az analízis értékelésére van még szükség.

lda\_1\_pred <- predict(lda\_1, method = "plug-in")  
tab\_1 <- table(lda\_1\_pred$class, bio$vasarlas)  
tab\_1  
#>   
#> soha nem vesz időnként vesz gyakran...  
#> soha nem vesz 9 1 ...  
#> időnként vesz 1 7 ...  
#> gyakran vesz 0 2 ...  
100 \* sum(diag(tab\_1))/sum(tab\_1)  
#> [1] 83.33333

A fenti eredményben láthatjuk, hogy a valódi és a modell alapján becsült csoporttagságok mátrixában a legtöbb adat a főátlóban helyezkedik el. Ez arra utal, hogy a becsült csoporttagságok nagyjából lefedik az eredetit. Ez az arány 83%.

Az utolsó probléma körében a bio termékek vásárlásának gyakoriságát vizsgáltuk. A kapott eredményeink alapján azok, akik gyakran vásárolnak ilyen termékeket, pozitívabbak értékelik és pozitívabb attitűdökkel rendelkeznek a bio termékek irányában, fiatalabbak, fizetésük viszont alacsonyabb.

## 8.5 Példa: Vezetési programok

Egy vállalat menedzsmentje szeretné megvizsgálni különböző vezetési programok hatását, ezért három különböző vezetési programot vezetett be három különböző stratégiai üzleti egységben (SÜE). Az első SÜE-ben bevezetett program az egyenlőséget és az individualizmust hangsúlyozta. A második SÜE-ben az egyenlőséget és a csoportmunkát helyzeték középpontba. A harmadik SÜE-ben a bevezetett program egy nagyon hierarchikus vezetési elvet alkalmazott. Később mindhárom SÜE dolgozóinak körében felmérést végeztek, és a kérdések között szerepelt a szervezettel való elkötelezettség mértéke (szelkot), a szervezettel való elégedettség nagysága (elegedett), illetve a rendszer egalitárius vagy tekintélyelvű (autokrata) jellege (rendszer).

vezetes <- rio::import(file = "adat/diszkriminancia\_vezetesi\_program.xlsx")  
vezetes$SUE <- factor(vezetes$SUE)  
str(vezetes)  
#> 'data.frame': 30 obs. of 4 variables:  
#> $ SUE : Factor w/ 3 levels "1","2","3": 1 1 1...  
#> $ szelkot : num 4 4 5 5 3 3 3 5 3 5 ...  
#> $ elegedett: num 1 3 4 1 2 2 4 2 1 1 ...  
#> $ rendszer : num 2 1 2 4 4 3 2 3 4 2 ...  
psych::headTail(vezetes)  
#> SUE szelkot elegedett rendszer  
#> 1 1 4 1 2  
#> 2 1 4 3 1  
#> 3 1 5 4 2  
#> 4 1 5 1 4  
#> ... <NA> ... ... ...  
#> 27 3 1 1 5  
#> 28 3 5 1 4  
#> 29 3 5 1 5  
#> 30 3 5 3 4

Végezzük el a többváltozós variancia elemzést.

man\_1 <- manova(cbind(szelkot, elegedett, rendszer) ~ SUE, data = vezetes)  
summary(man\_1, test = "Wilks")  
#> Df Wilks approx F num Df den Df Pr(>F...  
#> SUE 2 0.1894 10.815 6 50 1.102e-0...  
#> Residuals 27 ...  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
summary.aov(man\_1, test = "Wilks")  
#> Response szelkot :  
#> Df Sum Sq Mean Sq F value Pr(>F)  
#> SUE 2 2.867 1.4333 1.1057 0.3455  
#> Residuals 27 35.000 1.2963   
#>   
#> Response elegedett :  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> SUE 2 17.267 8.6333 8.4152 0.001444 \*\*  
#> Residuals 27 27.700 1.0259   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Response rendszer :  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> SUE 2 52.267 26.1333 48 1.287e-09 \*\*\*  
#> Residuals 27 14.700 0.5444   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
lm\_1 <- lm(szelkot ~ SUE, data = vezetes)  
car::Anova(lm\_1, test.statistic = c("Wilks"))  
#> Anova Table (Type II tests)  
#>   
#> Response: szelkot  
#> Sum Sq Df F value Pr(>F)  
#> SUE 2.867 2 1.1057 0.3455  
#> Residuals 35.000 27  
1 - summary(lm\_1)$r.squared # Wilks lambda  
#> [1] 0.9242958  
  
lm\_1 <- lm(elegedett ~ SUE, data = vezetes)  
car::Anova(lm\_1, test.statistic = c("Wilks"))  
#> Anova Table (Type II tests)  
#>   
#> Response: elegedett  
#> Sum Sq Df F value Pr(>F)   
#> SUE 17.267 2 8.4152 0.001444 \*\*  
#> Residuals 27.700 27   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
1 - summary(lm\_1)$r.squared # Wilks lambda  
#> [1] 0.6160119  
  
lm\_1 <- lm(rendszer ~ SUE, data = vezetes)  
car::Anova(lm\_1, test.statistic = c("Wilks"))  
#> Anova Table (Type II tests)  
#>   
#> Response: rendszer  
#> Sum Sq Df F value Pr(>F)   
#> SUE 52.267 2 48 1.287e-09 \*\*\*  
#> Residuals 14.700 27   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
1 - summary(lm\_1)$r.squared # Wilks lambda  
#> [1] 0.2195122

biotools::boxM(data = vezetes[c("szelkot", "elegedett", "rendszer")],  
 grouping = vezetes$SUE)  
#>   
#> Box's M-test for Homogeneity of Covariance  
#> Matrices  
#>   
#> data: vezetes[c("szelkot", "elegedett", "rendszer")]  
#> Chi-Sq (approx.) = 19.607, df = 12, p-value =  
#> 0.0749

library(MASS)  
lda\_1 <- lda(SUE ~ szelkot + elegedett + rendszer, data = vezetes)  
lda\_1  
#> Call:  
#> lda(SUE ~ szelkot + elegedett + rendszer, data = ve...  
#>   
#> Prior probabilities of groups:  
#> 1 2 3   
#> 0.3333333 0.3333333 0.3333333   
#>   
#> Group means:  
#> szelkot elegedett rendszer  
#> 1 4.0 2.1 2.7  
#> 2 4.7 3.4 1.5  
#> 3 4.1 1.6 4.7  
#>   
#> Coefficients of linear discriminants:  
#> LD1 LD2  
#> szelkot -0.07056752 0.4959142  
#> elegedett 0.10193972 0.8596451  
#> rendszer -1.32360357 0.5392379  
#>   
#> Proportion of trace:  
#> LD1 LD2   
#> 0.9616 0.0384  
# klaR::greedy.wilks(vasarlas~ertek+attitud+fizetes+kor,data=bio,  
# niveau = 0.15)

lda\_1\_pred <- predict(lda\_1, method = "plug-in")  
tab\_1 <- table(lda\_1\_pred$class, vezetes$SUE)  
tab\_1  
#>   
#> 1 2 3  
#> 1 4 1 0  
#> 2 3 9 0  
#> 3 3 0 10  
100 \* sum(diag(tab\_1))/sum(tab\_1)  
#> [1] 76.66667

## 8.6 Megjegyzések

Diszkriminancia analízis esetén az adatokat nem szükséges standardizálni, ennek oka, hogy az analízis eredményét nem befolyásolja jelentős mértékben az egyes változók mértékegysége.

A függő változónk tehát kategorikus, a függetlenek pedig numerikusak. Arra vagyunk kíváncsiak, hogy a függő változó által meghatározott csoportok mely független változókban különböznek egymástól, melyek különböztetik meg egy egymástól a függő változó kategóriáit.

Ha a kategorikus függő változónk csupán kétértékű, akkor kétváltozós diszkriminancia elemzésről beszélünk, több szint esetén többváltozós diszkiriminancia elemzésről.

## 8.7 Az alkalmazási feltételek

A fűggő változó kategorikus két vagy több szinttel. A független változók intervallum vagy arány skálájú változók, de használhatunk dichotóm változókat és a legalább 5 fokú likert skálán mért értékeket is. A függő változó csoportjaiban nagyjából azanosnak kell lennie a csoportnagyságnak, minden csoportnak legalább két adatsort tartalmaznia kell. A mintanagyságra is figyelnünk kell, a független változók számának kisebb kell lenni, mint a legkisebb csoport esetszáma, a teljes mintanagyság legalább 10-szer nagyobb a független változók számánál. A diszkriminancia elemzés feltételezi a független változók közötti lineáris kapcsolatot.

Az egyváltozós normalitás vizsgálatára a kiugró értékek vizsgálata javasolt, illetve megfelelő mérési skála (például nem dichotóm változó esetén) a Shapiro–Wilk-próbát is használhatjuk. A többváltozós normalitás vizsgálatához

A csoportok szétválasztásának egyik megközelítése a Mahalanobis-féle távolságot használja. Az eljárás lényege, hogy az csoportot tartalmazó minta átlagvektorával becsüljük a csoportok valódi átlagvektorát. Az egyes személyek csoportközéptől való átlagát számoljuk ki a Mahalanobis-féle távolsággal, és minden személyt abba a csoportba sorolunk be ez alapján, amelyhez közelebb esik. Ez lehet az a csoport, amelybe a személy valóban beletartozik, de lehet másik is. A helyes besorolások aránya világosan megmutatja, hogy mennyire jól lehet a csoportokat szétválasztani a használt változók alapján.

# remotes::install\_github('hyunsooseol/snowCluster')

library(snowCluster)  
vallalat <- rio::import(file = "adat/diszkriminancia\_vezetesi\_program.sav")  
snowCluster::disc(  
 data = vallalat,  
 dep = SUE,  
 covs = vars(szelköt, elégedett, rendszer),  
 gm = TRUE,  
 coef = TRUE,  
 prop = TRUE,  
 tes = TRUE,  
 plot = TRUE,  
 plot1 = TRUE,  
 plot2 = TRUE)  
  
  
str(vallalat)  
vallalat$SUE <- factor(vallalat$SUE)  
  
lda\_1 <- MASS::lda(SUE ~ szelköt + elégedett + rendszer, data = vallalat)  
lda\_1  
  
man\_1 <- stats::manova(cbind(szelköt, elégedett, rendszer)~SUE, data=vallalat)  
man\_1  
summary(man\_1, test="Wilks")  
summary.aov(man\_1)  
  
F <- 1.1057 # F próbastatisztika érték  
p <- 1 # függő változók száma  
n <- 30 # mintaelemszám  
k <- 3 # a független változó csoportjainak a száma  
   
Wilks\_1 <- 1 / (1 + (F \* p) / (n - k - 1 - p))  
Wilks\_1  
  
 1 - (F / (F + 27))  
  
ahol F az F-érték, df1 pedig az első szab  
anova(lm(szelköt~SUE, data=vallalat), test="Wilks")  
  
  
man\_1 <- manova(cbind(szelköt, elégedett, rendszer)~SUE, data=vallalat)  
summary(man\_1, test="Wilks")  
summary(man\_1)  
  
  
install.packages("klaR")  
gw\_1 <- klaR::greedy.wilks(SUE ~ szelköt + elégedett + rendszer, data = vallalat, output=T)  
unclass(gw\_1)  
plot(gw\_1)  
  
jmv::mancova(  
 data = vallalat,  
 deps = vars(szelköt, elégedett, rendszer),  
 factors = SUE,  
 multivar = "wilks",  
 boxM = TRUE,  
 shapiro = TRUE)  
  
??'Wilk'  
rrcov::Wilks.test(SUE ~ szelköt + elégedett + rendszer, data = vallalat)  
rrcov::Wilks.test(x = vallalat[2:4], grouping=vallalat$SUE)  
  
library(klaR)  
data(iris)  
library(MASS)  
iris.d <- iris[,1:4] # the data   
iris.c <- iris[,5] # the classes   
sc\_obj <- stepclass(iris.d, iris.c, "lda", start.vars = "Sepal.Width")  
sc\_obj  
plot(sc\_obj)  
  
## or using formulas:  
sc\_obj <- stepclass(Species ~ ., data = iris, method = "qda",   
 start.vars = "Sepal.Width", criterion = "AS") # same as above   
sc\_obj

data <- rio::import(file = "adat/diszkriminancia\_alkalmassag.xlsx")  
data <- rio::import(file = "adat/diszkriminancia\_baleset.xlsx")  
snowCluster::disc(data = data, dep = baleset, covs = vars(megosztott,  
 pontossag, kockazat, eszleles), gm = TRUE, coef = TRUE, prop = TRUE,  
 tra = TRUE, plot = TRUE, plot1 = TRUE, plot2 = TRUE)  
  
  
iris  
  
  
  
  
  
snowCluster::disc(data = iris, dep = Species, covs = vars(Sepal.Length,  
 Sepal.Width, Petal.Length, Petal.Width), gm = TRUE, coef = TRUE,  
 prop = TRUE, tra = TRUE, plot = TRUE, plot1 = TRUE, plot2 = TRUE)

# 9. Többváltozós varianciaelemzés

## 9.1 Elméleti háttér

A MANOVA a többváltozós varianciaelemzés angol megfelelőjéből képzett betűszó (Multivariate ANOVA vagy Multivariate Analysis of Variance). A szokásos ANOVA kiterjesztésének tekinthető, ahol nem egy, hanem kettő vagy több függő változóval dolgozhatunk, de a cél ugyanaz: a független változó több csoportja közötti különbségek elemzése.

Felmerülhet bennünk, hogy ha több függő változónk van, akkor mindegyikre végezzünk el külön egy-egy hagyományos ANOVA-t, azonban ez az elsőfajú hiba emelkedéséhez vezet. A MANOVA olyan megoldást kínál, amivel több függő változó kombinált információi alapján képes kimutatni a csoportkülönbségeket.

Mivel a MANOVA egynél több függő változót használ, a null- és az ellenhipotézisek kissé megváltoznak:

* : A csoportok várható érték vektorai minden csoportban azonosak.
* : A csoportok várható érték vektorainak legalább egyike eltér egy másiktól.

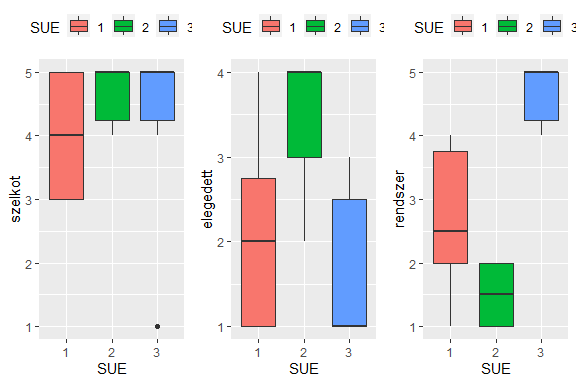
## 9.2 Példa: Vezetési programok

Egy vállalat menedzsmentje szeretné megvizsgálni különböző vezetési programok hatását, ezért három különböző vezetési programot vezetett be három különböző stratégiai üzleti egységben (SÜE). Az első SÜE-ben bevezetett program az egyenlőséget és az individualizmust hangsúlyozta. A második SÜE-ben az egyenlőséget és a csoportmunkát helyzeték középpontba. A harmadik SÜE-ben a bevezetett program egy nagyon hierarchikus vezetési elvet alkalmazott. Később mindhárom SÜE dolgozóinak körében felmérést végeztek, és a kérdések között szerepelt a szervezettel való elkötelezettség mértéke (szelkot), a szervezettel való elégedettség nagysága (elegedett), illetve a rendszer egalitárius vagy tekintélyelvű (autokrata) jellege (rendszer). Vizsgáljuk meg, hogy a SÜE három csoportja azonosnak tekinthető-e a vizsgált 3 kérdésre (szelkot, elegedett és rendszer) adott válaszok tekintetében.

vezetes <- rio::import(file = "adat/manova\_vezetesi\_program.xlsx")  
vezetes$SUE <- factor(vezetes$SUE)  
str(vezetes)  
#> 'data.frame': 30 obs. of 4 variables:  
#> $ SUE : Factor w/ 3 levels "1","2","3": 1 1 1...  
#> $ szelkot : num 4 4 5 5 3 3 3 5 3 5 ...  
#> $ elegedett: num 1 3 4 1 2 2 4 2 1 1 ...  
#> $ rendszer : num 2 1 2 4 4 3 2 3 4 2 ...  
psych::headTail(vezetes)  
#> SUE szelkot elegedett rendszer  
#> 1 1 4 1 2  
#> 2 1 4 3 1  
#> 3 1 5 4 2  
#> 4 1 5 1 4  
#> ... <NA> ... ... ...  
#> 27 3 1 1 5  
#> 28 3 5 1 4  
#> 29 3 5 1 5  
#> 30 3 5 3 4

Mivel a MANOVA a három stratégiai üzleti egységben (SÜE) a kérdőívek pontszámainak (vagyis a függő változók) átlagainak különbségire kérdez rá, így készítsünk dobozdiagramot mindhárom csoportban

library(ggplot2)  
p1 <- ggplot(vezetes, aes(x = SUE, y = szelkot, fill = SUE)) +  
 geom\_boxplot() + theme(legend.position = "top")  
p2 <- ggplot(vezetes, aes(x = SUE, y = elegedett, fill = SUE)) +  
 geom\_boxplot() + theme(legend.position = "top")  
p3 <- ggplot(vezetes, aes(x = SUE, y = rendszer, fill = SUE)) +  
 geom\_boxplot() + theme(legend.position = "top")  
gridExtra::grid.arrange(p1, p2, p3, nrow = 1)



Úgy tűnik, hogy a mindhárom csoport eléggé különbözik egymástól.

Végezzük el az egyszempontos többváltozós variancia elemzést. A manova() függvény a formula= argumentumában a kettő vagy több numerikus függő változót és legalább egy független változót vár. A függő változókat most a cbind() függvénnyel fűztük egymás mellé, a független változónk pedig a 3 szintű kategorikus SUE.

man\_1 <- manova(formula = cbind(szelkot, elegedett, rendszer) ~  
 SUE, data = vezetes)  
summary(man\_1)  
#> Df Pillai approx F num Df den Df Pr(>F)  
#> SUE 2 0.90947 7.2278 6 52 1.211e-05  
#> Residuals 27   
#>   
#> SUE \*\*\*  
#> Residuals   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Alapértelmezés szerint a MANOVA az R-ben a Pillai-féle tesztstatisztikáit használja. A p-érték gyakorlatilag nulla, ami azt jelenti, hogy nyugodtan elvethetjük a nullhipotézist: legalább egy csoportátlagvektor eltér a többitől.

Használhat más teszteket is, mint például a Wilk-lambda, a Roy-féle vagy a Hotelling-Lawley statisztikákat, de a Pillai-féle a legrobusztosabb.

summary(man\_1, test = "Wilks")  
#> Df Wilks approx F num Df den Df Pr(>F...  
#> SUE 2 0.1894 10.815 6 50 1.102e-0...  
#> Residuals 27 ...  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
summary(man\_1, test = "Hotelling-Lawley")  
#> Df Hotelling-Lawley approx F num Df den Df  
#> SUE 2 3.7577 15.031 6 48  
#> Residuals 27   
#> Pr(>F)   
#> SUE 1.375e-09 \*\*\*  
#> Residuals   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
summary(man\_1, test = "Roy")  
#> Df Roy approx F num Df den Df Pr(>F...  
#> SUE 2 3.6133 31.315 3 26 8.724e-0...  
#> Residuals 27 ...  
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A hatásnagyság kiszámítására MANOVA esetében a parciális Eta négyzet mutatót használhatjuk. Azt méri, hogy a független változó milyen hatással van a függő változókra. Ha az érték 0,14 vagy nagyobb, akkor azt mondhatjuk, hogy a hatás mérete nagy. Ez most 0,45, ami azt jelenti, hogy a hatás mérete nagy.

effectsize::eta\_squared(man\_1, partial = T)  
#> # Effect Size for ANOVA (Type I)  
#>   
#> Parameter | Eta2 (partial) | 95% CI  
#> -----------------------------------------  
#> SUE | 0.45 | [0.24, 1.00]  
#>   
#> - One-sided CIs: upper bound fixed at [1.00].  
effectsize::interpret\_eta\_squared(0.45, partial = T)  
#> [1] "large"  
#> (Rules: field2013)

Mivel a MANOVA szignifikáns lett, további kérdés, hogy melyik csoport átlagvektora különbözik a többitől? Post-hoc tesztet kell végeznünk, amely esetünkben a lineáris diszkriminancia elemzés és az egyváltozós ANOVA lesz.

## 9.3 Post-hoc teszt: LDA

A lineáris diszkriminancia elemzés (LDA) célja, hogy változók olyan lineáris kombinációját találja meg, amely a legjobban elválaszt két vagy több csoportot. Ezáltal képesek leszünk egy olyan pontdiagramot megjeleníteni, amely az X és Y tengely két lineáris diszkriminánst jeleníti meg, a pontokat pedig a független változónak (SUE) megfelelően fogjuk színezni.

A lineáris diszkriminancia elemzést R-ben a {MASS} csomag lda() függvényével végzünk.

library(MASS)  
lda\_1 <- lda(SUE ~ szelkot + elegedett + rendszer, data = vezetes)  
lda\_1  
#> Call:  
#> lda(SUE ~ szelkot + elegedett + rendszer, data = ve...  
#>   
#> Prior probabilities of groups:  
#> 1 2 3   
#> 0.3333333 0.3333333 0.3333333   
#>   
#> Group means:  
#> szelkot elegedett rendszer  
#> 1 4.0 2.1 2.7  
#> 2 4.7 3.4 1.5  
#> 3 4.1 1.6 4.7  
#>   
#> Coefficients of linear discriminants:  
#> LD1 LD2  
#> szelkot -0.07056752 0.4959142  
#> elegedett 0.10193972 0.8596451  
#> rendszer -1.32360357 0.5392379  
#>   
#> Proportion of trace:  
#> LD1 LD2   
#> 0.9616 0.0384  
# klaR::greedy.wilks(vasarlas~ertek+attitud+fizetes+kor,data=bio,  
# niveau = 0.15)

A fenti együtthatókból megtudhatjuk hogyan használják fel a függő változókat az LDA döntési szabályának kialakítására. Az LD1 a következőképpen számítható ki:

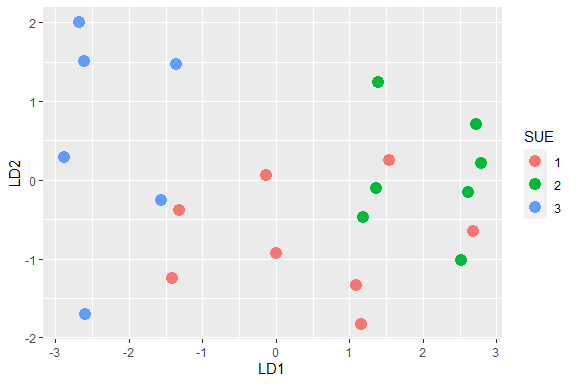
A vezetes adatmátrix numerikus változóira magunk is kiszámolhatjuk az LD1 és LDA2 értékét a predict() függvénnyel:

lda\_1\_pred <- predict(lda\_1, method = "plug-in")  
psych::headTail(lda\_1\_pred$x)  
#> LD1 LD2  
#> 1 1.16 -1.83  
#> 2 2.69 -0.65  
#> 3 1.39 1.25  
#> 4 -1.56 -0.25  
#> ... ... ...  
#> 27 -2.6 -1.7  
#> 28 -1.56 -0.25  
#> 29 -2.88 0.29  
#> 30 -1.35 1.47

A post-hoc teszt utolsó lépése a fenti a pontdiagram megjelenítése. Ideális esetben egy vagy több csoport kiemelkedik:

d <- data.frame(lda\_1\_pred$x, SUE = vezetes$SUE)  
psych::headTail(d)  
#> LD1 LD2 SUE  
#> 1 1.16 -1.83 1  
#> 2 2.69 -0.65 1  
#> 3 1.39 1.25 1  
#> 4 -1.56 -0.25 1  
#> ... ... ... <NA>  
#> 27 -2.6 -1.7 3  
#> 28 -1.56 -0.25 3  
#> 29 -2.88 0.29 3  
#> 30 -1.35 1.47 3

ggplot(d, aes(x = LD1, y = LD2, colour = SUE)) + geom\_point(size = 4)



A képen látható, hogy a harmadik SÜE csoport eltér mindkét másik csoporttól, míg a az első két csoport eltérése egymástól nem mondható markánsnak. Könnyen elképzelhető, hogy a SÜE harmadik csoportja volt a legnagyobb hatással a nullhipotézis elutasítására.

## 9.4 Post-hoc test: egyváltozós vizsgálatok

A statisztikailag szignifikáns egyszempontos MANOVA után az egyváltozós egyszempontos ANOVA-val is vizsgálódhatunk, amely minden függő változót külön-külön vizsgál. A cél az, hogy azonosítsuk azokat a konkrét függő változókat, amelyek hozzájárultak a jelentős globális hatáshoz. A klasszikus ANOVA mellett a Welch-féle változat és a Kruskal–Wallis-próba is használható, a feltételek egyre nagyobb csorbulása esetén. Most a nemparaméteres Kruskal–Wallis-próbát használjuk.

kruskal.test(szelkot ~ SUE, data = vezetes)$p.value  
#> [1] 0.2499069  
kruskal.test(elegedett ~ SUE, data = vezetes)$p.value  
#> [1] 0.003003241  
kruskal.test(rendszer ~ SUE, data = vezetes)$p.value  
#> [1] 1.690362e-05

Látjuk, hogy az elegedett és a rendszer függő változókban nem egyeznek a várható értékek a SÜE egyes csoportjaiban. Megjegyezzük, hogy mivel 3 függő változónk van, a Bonferroni-féle többszörös tesztelési korrekciót alkalmaznunk kell, vagyis a statisztikai szignifikancia szintet csökkenteni kell. Ez úgy történik, hogy a klasszikus alfa szintet (0,05) elosztjuk a tesztek (vagy függő változók, itt 3) számával. Ez p < 0,017-es szignifikancia elfogadási kritériumhoz vezet. A fenti próbák szignifikáns voltán ez most nem változtat.

A statisztikailag szignifikáns egyváltozós ANOVA-t (esetünkben Kruskal–Wallis-próbát) többszörös páronkénti összehasonlítás követi annak meghatározására, hogy mely csoportok különböznek egymástól. Most a Kruskal–Wallis-próba szokásos utóvizsgálatát a Dunn-próbát fogjuk használni.

library(DescTools)  
DunnTest(formula = szelkot ~ SUE, data = vezetes, method = "holm")  
#>   
#> Dunn's test of multiple comparisons using rank sum...  
#>   
#> mean.rank.diff pval   
#> 2-1 5.6 0.3179   
#> 3-1 4.0 0.4964   
#> 3-2 -1.6 0.6442   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '....  
DunnTest(formula = elegedett ~ SUE, data = vezetes, method = "holm")  
#>   
#> Dunn's test of multiple comparisons using rank sum...  
#>   
#> mean.rank.diff pval   
#> 2-1 8.75 0.0410 \*   
#> 3-1 -3.80 0.3143   
#> 3-2 -12.55 0.0027 \*\*   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '....  
DunnTest(formula = rendszer ~ SUE, data = vezetes, method = "holm")  
#>   
#> Dunn's test of multiple comparisons using rank sum...  
#>   
#> mean.rank.diff pval   
#> 2-1 -6.95 0.0694 .   
#> 3-1 10.85 0.0092 \*\*   
#> 3-2 17.80 9.9e-06 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '....

A fenti utóvizsgálatok világossá teszik, hogy a harmadik SÜE csoport a rendszer változó esetén mindkét másik csoporttól, az elegedett változó esetén pedig a második csoporttól szignifikánsan eltér. Legjelentősebb mértékben tehűt a harmadik csoport különül el a másik két csoporttól, tehát ez okozza a MANOVA nullhipotézisének elvetését.

## 9.5 Elemzés jamovi-ban

A fenti elemzés jamovi-ban is elvégezhető az ANOVA / MANCOVA menüpontok kiválasztásával.

|  |
| --- |
| MANOVA jamovi-ban |

## 9.6 Alkalmazási feltételek vizsgálata

A MANOVA statisztikai próbának számos szigorú alkalmazási feltétele van. Néhány az ANOVA-ból jön, például a megfigyelések függetlensége vagy a variancia homogenitása, azonban vannak újdonságok is.

* **Megfelelő mintanagyság.** Ökölszabály: a mintaelemszám mindegyik független változó csoportban nagyobb az függő változók számánál.

summarytools::freq(vezetes$SUE, cumul = FALSE)  
#> Frequencies   
#> vezetes$SUE   
#> Type: Factor   
#>   
#> Freq % Valid % Total  
#> ----------- ------ --------- ---------  
#> 1 10 33.33 33.33  
#> 2 10 33.33 33.33  
#> 3 10 33.33 33.33  
#> <NA> 0 0.00  
#> Total 30 100.00 100.00

Látható, hogy a függő változók számát (3) minden csoport elemszáma (10) meghaladja.

* **A megfigyelések függetlensége.** Minden személynek csak egy csoportba kell tartoznia. Az egyes csoportok megfigyelései között nincs kapcsolat. Az ismételt mérések nem megengedettek. A minta kiválasztásának teljesen véletlenszerűnek kell lennie.
* **Az egyváltozós vagy többváltozós kiugró értékek hiánya.**

Az egydimenziós kiugró értékek dobozdiagramokkal is ellenőrizhetők, ezt korábban elvégeztük, láttuk csak egyetlen részcsoportban van kiugró értékek (a szelkot változó változó esetén a SÜE harmadik csoportjában). Használhatjuk a kényelmes rstatix::identify\_outliers() függvényt is.

library(tidyverse)  
vezetes %>%  
 group\_by(SUE) %>%  
 rstatix::identify\_outliers(szelkot)  
#> # A tibble: 2 × 6  
#> SUE szelkot elegedett rendszer is.outlier is.ex...  
#> <fct> <dbl> <dbl> <dbl> <lgl> <lgl>...  
#> 1 3 1 1 5 TRUE TRUE ...  
#> 2 3 1 1 5 TRUE TRUE  
vezetes %>%  
 group\_by(SUE) %>%  
 rstatix::identify\_outliers(elegedett)  
#> [1] SUE szelkot elegedett rendszer   
#> [5] is.outlier is.extreme  
#> <0 rows> (or 0-length row.names)  
vezetes %>%  
 group\_by(SUE) %>%  
 rstatix::identify\_outliers(rendszer)  
#> [1] SUE szelkot elegedett rendszer   
#> [5] is.outlier is.extreme  
#> <0 rows> (or 0-length row.names)

A többváltozós kiugró értékek olyan adatpontok, amelyek szokatlan értékkombinációt tartalmaznak a kimeneti (vagy függő) változókon. A Mahalanobis távolságot általában a többváltozós kiugró értékek észlelésére használják. A távolság megmondja, milyen messze van egy megfigyelés a felhő középpontjától, figyelembe véve a felhő alakját (kovariancia) is. A rstatix::mahalanobis\_distance() függvény könnyen használható a Mahalanobis-távolság kiszámítására és a többváltozós kiugró értékek megjelölésére. A Mahalanobis-távolságot csoportonként kell kiszámítani:

vezetes %>%  
 group\_by(SUE) %>%  
 rstatix::mahalanobis\_distance() %>%  
 filter(is.outlier == TRUE) %>%  
 as.data.frame()  
#> [1] szelkot elegedett rendszer mahal.dist  
#> [5] is.outlier  
#> <0 rows> (or 0-length row.names)

Látható, hogy nincs többváltozós kiugró érték az adatbázisban.

* **Többváltozós normalitás.**

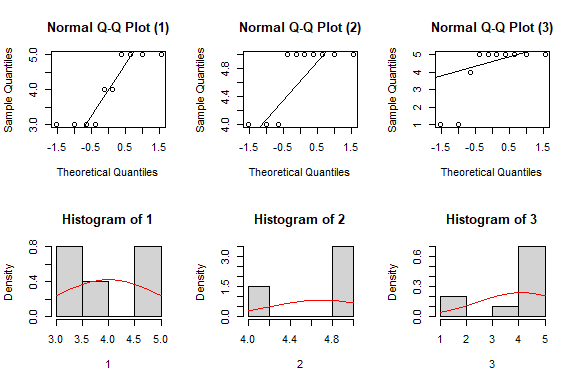
A többváltozós normalitás Shapiro-Wilk tesztjének végrehajtása:

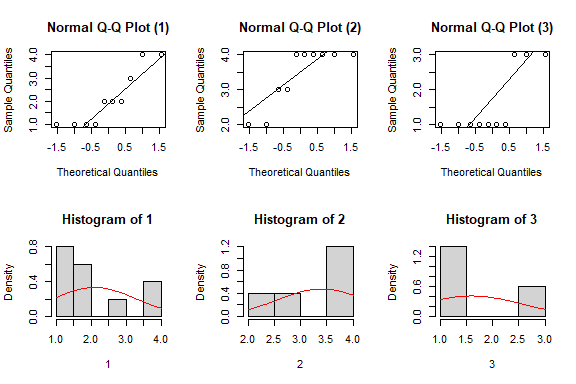
rstatix::mshapiro\_test(vezetes[c("szelkot", "elegedett", "rendszer")])  
#> # A tibble: 1 × 2  
#> statistic p.value  
#> <dbl> <dbl>  
#> 1 0.772 0.0000214

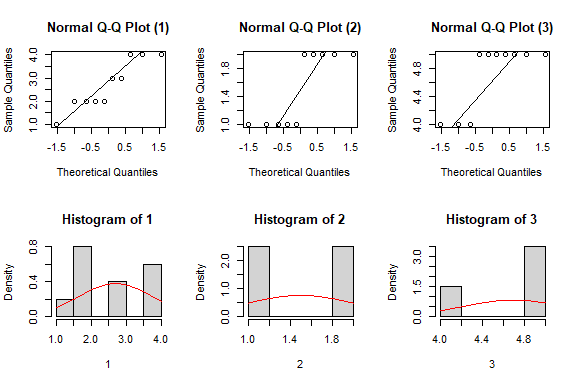
Látható, hogy ez az alkalmazási feltétel nem teljesül.

Az egyváltozós normalitásokat is érdemes lehet tesztelni:

# egyváltozós Shapiro–Wilk próba több csoportra  
library(onewaytests)  
nor.test(formula = szelkot ~ SUE, data = vezetes, method = "SW")  
#>   
#> Shapiro-Wilk Normality Test (alpha = 0.05)   
#> --------------------------------------------------   
#> data : szelkot and SUE   
#>   
#> Level Statistic p.value Normality  
#> 1 1 0.7685823 6.009970e-03 Reject  
#> 2 2 0.5941735 4.713464e-05 Reject  
#> 3 3 0.5876023 3.936679e-05 Reject  
#> --------------------------------------------------  
nor.test(formula = elegedett ~ SUE, data = vezetes, method = "SW")  
#>   
#> Shapiro-Wilk Normality Test (alpha = 0.05)   
#> --------------------------------------------------   
#> data : elegedett and SUE   
#>   
#> Level Statistic p.value Normality  
#> 1 1 0.8236140 2.802300e-02 Reject  
#> 2 2 0.7172415 1.425861e-03 Reject  
#> 3 3 0.5941735 4.713464e-05 Reject  
#> --------------------------------------------------  
nor.test(formula = rendszer ~ SUE, data = vezetes, method = "SW")  
#>   
#> Shapiro-Wilk Normality Test (alpha = 0.05)   
#> --------------------------------------------------   
#> data : rendszer and SUE   
#>   
#> Level Statistic p.value Normality  
#> 1 1 0.8737456 1.105101e-01 Not reject  
#> 2 2 0.6552710 2.539627e-04 Reject  
#> 3 3 0.5941735 4.713464e-05 Reject  
#> --------------------------------------------------







* **A multikollinearitás hiánya.** A függő (eredmény) változók nem korrelálhatnak túlságosan egymással. Egyetlen korreláció sem lehet r = 0,90 feletti.

Ideális esetben az eredményváltozók közötti korreláció mérsékelt, nem túl magas. A 0,9 feletti korreláció a multikollinearitást jelzi, ami a MANOVA esetében problematikus. Másrészt, ha a korreláció túl alacsony, fontolóra kell vennie külön egyszempontos ANOVA futtatását minden függő változóra.

Számítsuk ki a páronkénti Pearson-korrelációs együtthatókat a függő változók között.

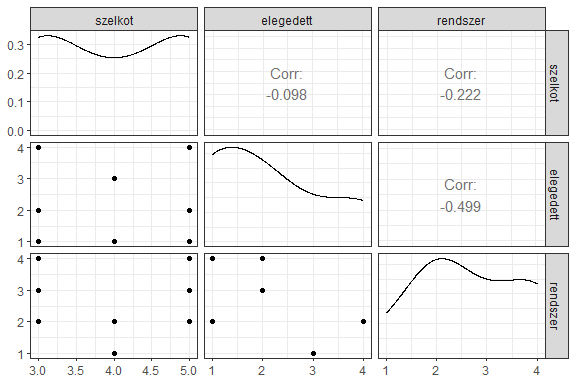
cor(vezetes[c("szelkot", "elegedett", "rendszer")])  
#> szelkot elegedett rendszer  
#> szelkot 1.0000000 0.1954881 -0.2528624  
#> elegedett 0.1954881 1.0000000 -0.6129079  
#> rendszer -0.2528624 -0.6129079 1.0000000

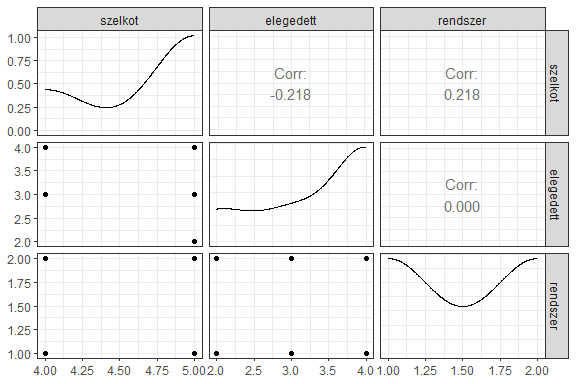
Látható, hogy a korrelációs együtthatók nem támogatják a multikollinearitás tényét.

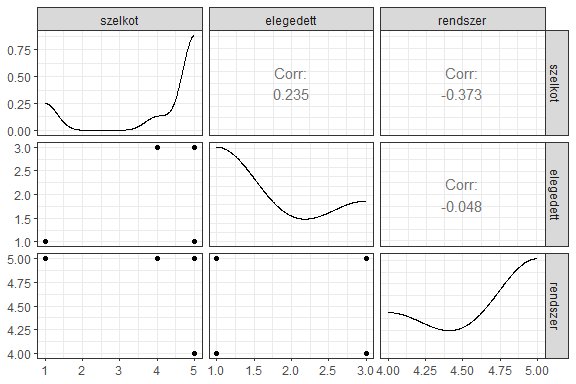
* **Linearitás az összes függő változó között minden csoportban.**

Mivel a függő változók közötti páronkénti kapcsolatnak lineárisnak kell lennie minden csoport esetében, ezért érdemes ezt a feltételt vizuálisan ellenőrizni. A {GGally} csomag ggpairs() függvényét használhatjuk.

library(GGally)  
res <- vezetes %>%  
 select(SUE, szelkot, elegedett, rendszer) %>%  
 group\_by(SUE) %>%  
 rstatix::doo(~ggpairs(.) + theme\_bw(), result = "plots")  
res$plots  
#> [[1]]  
#>   
#> [[2]]  
#>   
#> [[3]]







A fenti ábrák megkérdőjelezik a páronkénti lineáris kapcsolatok létezését.

* **A varianciák homogenitása.** A Levene-próba használható a csoportok közötti varianciák egyenlőségének tesztelésére. A Levene-próba nem szignifikáns értékei a varianciák homogenitását támogatják.

Az egyszempontos MANOVA mindegyik függő változó esetében azt feltételezi, hogy a csoportok között egyenlők a varianciák.

DescTools::LeveneTest(szelkot ~ SUE, data = vezetes)  
#> Levene's Test for Homogeneity of Variance (center =...  
#> Df F value Pr(>F)  
#> group 2 0.9755 0.3899  
#> 27  
DescTools::LeveneTest(elegedett ~ SUE, data = vezetes)  
#> Levene's Test for Homogeneity of Variance (center =...  
#> Df F value Pr(>F)  
#> group 2 0.4112 0.667  
#> 27  
DescTools::LeveneTest(rendszer ~ SUE, data = vezetes)  
#> Levene's Test for Homogeneity of Variance (center =...  
#> Df F value Pr(>F)   
#> group 2 5.6 0.009238 \*\*  
#> 27   
#> ---  
#> Signif. codes:   
#> 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Látható, hogy a szóráshomogenitás a rendszer változó kivételével teljesül.

* **Variancia-kovariancia mátrixok homogenitása.** A BoxM-próba használható a csoportok közötti kovariancia egyenlőségének ellenőrzésére. Ez egyenértékű a variancia többváltozós homogenitásával. Ez a teszt rendkívül érzékenynek tekinthető. Ezért ennek a tesztnek a szignifikanciáját alfa = 0,001 értéknél határozzuk meg. A {biotools} csomag megvalósított boxM() függvényét használhatjuk.

biotools::boxM(data = vezetes[c("szelkot", "elegedett", "rendszer")],  
 grouping = vezetes$SUE)  
#>   
#> Box's M-test for Homogeneity of Covariance  
#> Matrices  
#>   
#> data: vezetes[c("szelkot", "elegedett", "rendszer")]  
#> Chi-Sq (approx.) = 19.607, df = 12, p-value =  
#> 0.0749

A teszt statisztikailag nem szignifikáns (azaz p > 0,001), tehát az adatok nem sértették meg a variancia-kovariancia mátrixok homogenitásának feltételezését.

Kiegyensúlyozott a csoportelemszámok esetén nem probléma a variancia-kovariancia mátrixok homogenitásának megsértése miatt, de kiegyensúlyozatlan kialakításnál már problémás lehet.

# 10. Logisztikus regresszió

A logisztikus regresszió céljait tekintve megegyezik a diszkriminancia elemzéssel, de sokkal robusztusabb, azaz kevesebb alkalmazási feltétellel rendelkezik. Használható a logisztikus regresszió akkor is, ha a független változók között kategorikus változók is előfordulnak, illetve a normalitásra és homoszkedaszticitásra vonatkozó feltétel megsértésre sem érzékeny a módszer.

JÖN.

# 11. Többdimenziós skálázás

JÖN.

# Appendix A — Elérhető videók

* Alapozo jamovi és R videók
  + [Jamovi a gyakorlatban](https://www.youtube.com/watch?v=sZummF3Sd90&list=PLnmeQHnHYqv5_FoVOvX9tqE90EqoOW01o)
  + [R a gyakorlatban](https://www.youtube.com/watch?v=UBeiIKzE8VA&list=PLnmeQHnHYqv6ENGrdTXiE9YJrvHuxH2C9)
* Jamovi tutorial videók az összes tanult eljáráshoz
  + [datalab.cc](https://www.youtube.com/@datalabcc/videos)
  + [Alexander Swan](https://www.youtube.com/@AlexanderSwan/videos)
* Lineáris regresszió
  + [The linear regression model](https://www.youtube.com/watch?v=m88h75F3Rl8)
  + [Linear Regression, Clearly Explained](https://www.youtube.com/watch?v=nk2CQITm_eo)
  + [R-squared, Clearly Explained](https://www.youtube.com/watch?v=2AQKmw14mHM)
  + [Simple linear regression in Jamovi](https://www.youtube.com/watch?v=rUMWjRE8L1U)
* Főkomponens elemzés
  + [Principal Component Analysis (PCA)](https://www.youtube.com/watch?v=g-Hb26agBFg)
* Klaszterelemzés
  + [Flat and Hierarchical Clustering | The Dendrogram Explained](https://www.youtube.com/watch?v=ijUMKMC4f9I)
  + [K Means Clustering: Pros and Cons of K Means Clustering](https://www.youtube.com/watch?v=YIGtalP1mv0)
  + [Hierarchical Cluster analysis in Jamovi](https://www.youtube.com/watch?v=TnFfAQpq0o4)
  + [K Means Cluster analysis in Jamovi](https://www.youtube.com/watch?v=wW7TNlam77A)
* Többszempontos varianciaelemzés
  + [One Way ANOVA Post hoc test in Jamovi](https://www.youtube.com/watch?v=pECfK5IfB_o)
  + [One Way Repeated Measure ANOVA Repeated Measure ANOVA Within Subject ANOVA in Jamovi](https://www.youtube.com/watch?v=O6BiwVRj3BQ)
  + [Two Way ANOVA Post hoc test in Jamovi](https://www.youtube.com/watch?v=l6ClJFgxNXs)
  + [Three Way ANOVA Post hoc test in Jamovi](https://www.youtube.com/watch?v=WW8ALE6x3uQ)

# Appendix B — Adatbázisok

Jelen jegyzetben “könyv” alatt a következőt értjük:

* Münnich Ákos, Nagy Ágnes, Abari Kálmán. *Többváltozós statisztika pszichológus hallgatók számára.* Bölcsész Konzorcium, Debrecen, 2006. (http://psycho.unideb.hu/statisztika) ISBN 963 9704 04 0

## B.1 Megbízhatóság elemzés

* megbizhatosag\_tantargyak.xlsx - fiktív adatbázis 9 tanuló iskolai jegyeivel (Münnich és mtsai. (2006), 2.2. táblázat)
  + Az adatbázis szerkezete:
    - matek - matematika érdemjegy (numerikus: 1-5)
    - fizika - fizika érdemjegy (numerikus: 1-5)
    - informatika - informatika érdemjegy (numerikus: 1-5)
    - kemia - kémia érdemjegy (numerikus: 1-5)
  + Kapcsolódó állományok:
    - megbizhatosag\_tantargyak.omv - megbízhatóság elemzés jamovi-ban

## B.2 Többváltozós varianciaelemzés

* manova\_vezetesi\_program.xlsx - A szervezeti elkötelezettség, a szervezeti kultúra és az elégedettség eltér a vállalat 3 különböző vezetési irányelvét valló egységében? Az adatbázis Sajtos és Mitev (2007, o. 332) könyvéből származik.  
  Egy vállalat menedzsmentje szeretné megvizsgálni különböző vezetési programok hatását, ezért három különböző vezetési programot vezetett be három különböző stratégiai üzleti egységben (SÜE). Az első SÜE-ben bevezetett program az egyenlőséget és az individualizmust hangsúlyozta. A második SÜE-ben az egyenlőséget és a csoportmunkát helyzeték középpontba. A harmadik SÜE-ben a bevezetett program egy nagyon hierarchikus vezetési elvet alkalmazott. Később mindhárom SÜE dolgozóinak körében felmérést végeztek, és a kérdések között szerepelt a szervezettel való elkötelezettség mértéke (szelkot), a szervezettel való elégedettség nagysága (elegedett), illetve a rendszer egalitárius vagy tekintélyelvű (autokrata) jellege (rendszer).
  + Az adatbázis szerkezete:
    - SUE - a három különböző vezetési irányelvet követő stratégiai üzleti egység (nominális: 1-3)
    - szelkot - szervezet elkötelezettség mértéke (likert: 1-5)
    - elegedett - szervezettel való elégedettség (likert: 1-5)
    - rendszer - szervezet tekintélyelvű jellege (likert: 1-5)
  + Kapcsolódó állományok:
    - manova\_vezetesi\_program.omv - Többváltozós varianciaelemzés jamovi-ban

## B.3 Diszkriminancia elemzés

* diszkriminancia\_alkalmassag.xlsx - szalagmunkások adatai (Münnich és mtsai. (2006), 4.1. táblázat)
  + Az adatbázis szerkezete:
    - bevalt - a munkás beválásával kapcsolatos információ: bevált? (nominális: “igen”, “nem”)
    - figyelem - a munkás figyelmi képessége (likert: 1-7, a magasabb érték jobb képességeket jelent)
    - monotonia\_tures - a munkás monotónia tűrése (likert: 1-7, a magasabb érték jobb képességeket jelent)
* diszkriminancia\_baleset.xlsx - mely tényezők járulnak hozzá a balesetekhez (Münnich és mtsai. (2006), 4.11. R-forráskód)
  + Az adatbázis szerkezete:
    - baleset - volt már balesete a személynek vagy sem (nominális “nem volt balesete”, “volt baleste”)
    - megosztott - megosztott figyelem (intervallum/arány)
    - pontossag - a figyelem pontossága (intervallum/arány)
    - kockazat - kockázatvállalási hajlandóság (intervallum/arány)
    - eszleles - észlelés gyorsasága (intervallum/arány)
* diszkriminancia\_depresszio.xlsx - a postpartum depresszió pszichés és szociális háttere (Münnich és mtsai. (2006), 4.16. R-forráskód)
  + Az adatbázis szerkezete:
    - ppdepresszio - szülés utáni depresszió jelenléte (nominális: “nincs depresszió”, “van depresszió”)
    - szeretet - a személyek mennyire érzik, hogy a szüleik szeretik őket (intervallum/arány)
    - tulvedes - mennyire hajlamosak arra a személyek, hogy túlságosan is burokban tartsák, túlvédjék gyerekeiket (intervallum/arány)
    - kor - életkor (intervallum/arány)
    - iskola - az elvégzett iskolai osztályok száma (intervallum/arány)
* diszkriminancia\_pszichoszomatika.xlsx - a pszichoszomatikus megbetegedéseket vizsgálata (Münnich és mtsai. (2006), 4.21. R-forráskód)
  + Az adatbázis szerkezete:
    - pszichoszomatika - van valamilyen pszichoszomatikus megbetegedése a személynek? (nominális: “szichoszomatikus megbetegedése van”, ” egészséges”)
    - stressz - a személyt ért stressz mértéke (intervallum/arány)
    - szorongas - a szorongási szintje (intervallum/arány)
    - coping - a megküzdési stratégiáinak hatékonysága (intervallum/arány)
* diszkriminancia\_bio.xlsx - kik vásárolnak bio termékeket (Münnich és mtsai. (2006), 4.26. R-forráskód)
  + Az adatbázis szerkezete:
    - vasarlas - a biotermékek vásárlásának gyakorisága (ordinális: “soha nem vesz”, “időnként vesz”, “gyakran vesz”)
    - ertek - minél nagyobb pontszámot kap a skálán, annál jobban értékeli a személy a bio termékeket (intervallum/arány)
    - attitud - a magasabb értékek kedvezőbb atttitűdöt jelez a biotermékek iránt (intervallum/arány)
    - fizetes - a személy fizetésének nagysága (intervallum/arány)
    - kor - a személy életkora(intervallum/arány)
* diszkriminancia\_vezetesi\_program.xlsx - A szervezeti elkötelezettség, a szervezeti kultúra és az elégedettség alapján szétválasztható a vállalat 3 különböző vezetési irányelvét valló egysége? Az adatbázis Sajtos és Mitev (2007, o. 332) könyvéből származik.  
  Egy vállalat menedzsmentje szeretné megvizsgálni különböző vezetési programok hatását, ezért három különböző vezetési programot vezetett be három különböző stratégiai üzleti egységben (SÜE). Az első SÜE-ben bevezetett program az egyenlőséget és az individualizmust hangsúlyozta. A második SÜE-ben az egyenlőséget és a csoportmunkát helyzeték középpontba. A harmadik SÜE-ben a bevezetett program egy nagyon hierarchikus vezetési elvet alkalmazott. Később mindhárom SÜE dolgozóinak körében felmérést végeztek, és a kérdések között szerepelt a szervezettel való elkötelezettség mértéke (szelkot), a szervezettel való elégedettség nagysága (elegedett), illetve a rendszer egalitárius vagy tekintélyelvű (autokrata) jellege (rendszer).
  + Az adatbázis szerkezete:
    - SUE - a három különböző vezetési irányelvet követő stratégiai üzleti egység (nominális: 1-3)
    - szelkot - szervezet elkötelezettség mértéke (likert: 1-5)
    - elegedett - szervezettel való elégedettség (likert: 1-5)
    - rendszer - szervezet tekintélyelvű jellege (likert: 1-5)
  + Kapcsolódó állományok:
    - diszkriminancia\_vezetesi\_program.omv - diszkriminancia elemzés jamovi-ban

## B.4 Lineáris regresszió

* lin\_reg\_fizetes\_elegedettseg\_01.omv - konstans oszlopokkal nem tudunk számolni (Münnich és mtsai. (2006) 1.1/A táblázat)
* lin\_reg\_fizetes\_elegedettseg\_02.omv- az adatpontok szinte tökéletesen az egyenesre illeszkednek (Münnich és mtsai. (2006) 1.1/B táblázat)
* lin\_reg\_kapcsolatok\_01.omv - nem szisztematikus kapcsolat két változó között (Münnich és mtsai. (2006) 1.5. R-forráskód)
* lin\_reg\_kapcsolatok\_02.omv - szisztematikus (függvényszerű) kapcsolat két változó között (Münnich és mtsai. (2006) 1.6. R-forráskód)
* lin\_reg\_kapcsolatok\_03.omv - szisztematikus (függvényszerű) kapcsolat két változó között (Münnich és mtsai. (2006) 1.7. R-forráskód)
* lin\_reg\_kapcsolatok\_04.omv - szisztematikus (függvényszerű) kapcsolat két változó között (Münnich és mtsai. (2006) 1.8. R-forráskód)
* lin\_reg\_kapcsolatok\_05.omv - szisztematikus (függvényszerű) kapcsolat két változó között (Münnich és mtsai. (2006) 1.9. R-forráskód)
* lin\_reg\_elegedttseg.omv - a fizetés és a munkahellyel való elégedettség pontdiagramja, egyszerű lineáris regresszió (Münnich és mtsai. (2006) 1.10 R-forráskód)
* lin\_reg\_fizetes\_eletkor\_eledettseg\_01.omv - többszörös lineáris regresszió, 2 numerikus magyarázó változó (Münnich és mtsai. (2006) 1.2. táblázat)
* lin\_reg\_intelligencia\_testmagassag\_eletkor\_01.omv - többszörös lineáris regresszió, 2 numerikus magyarázó változó, parciális korreláció magyarázata
  + minél magasabb valaki, annál intelligensebb
  + ha bevonjuk az életkor változót, akkor eltűnik az intelligencia és a testmagasság közötti kapcsolat
* lin\_tizproba.omv - többszörös lineáris regresszió, a legjobb modell keresése, sok numerikus magyarázó változó
* lin\_college\_success\_02.omv - többszörös lineáris regresszió, sok numerikus magyarázó változó, GPA a függő változó, mi magyarázza az egyetemi teljesítményt
* lin\_reg\_elegedttseg\_02.omv - A férfiak vagy a nők elégedettebbek a munkahelyükkel? (Münnich és mtsai. (2006) 1.6.3 probléma), egyetlen kategorikus magyarázó változó 2 értékkel (nem: férfi, nő)
  + kapcsolat a kétmintás t-próbával
* lin\_reg\_magassag\_hajhossz\_nem\_01.omv - többszörös lineáris regresszió, parciális korreláció 1 numerikus és 1 kategorikus változóval (Münnich és mtsai. (2006) 1.2. táblázat)
  + a testmagasság és a hajhossz között kapcsolat van
  + ha a személyek nemét is figyelembe vesszük, egyáltalán nincs kapcsolat a testmagasság és a hajhosszúság között
* lin\_auction.omv - többszörös lineáris regresszió, Simpson paradoxon, párhuzamos regresszió, majd interakció bevonása.

## B.5 Főkomponens elemzés

* fokomp\_elemzes\_tantargyak.omv - 1 főkomponens létrehozása (Münnich és mtsai. (2006) 2.2. táblázat)
* fokomp\_real\_targyak.omv - példa kidolgozása, 1 főkomponens(Münnich és mtsai. (2006) 2.5.1 Probléma)
* fokomp\_kerdoivtervezet.omv - példa kidolgozása (Münnich és mtsai. (2006) 2.5.2 Probléma)
* fokomp\_munkahelyi\_tolarencia.omv - példa kidolgozása (Münnich és mtsai. (2006) 2.5.3 Probléma)
* fokomp\_munkahelyi\_elegedettseg.omv - példa kidolgozása (Münnich és mtsai. (2006) 2.5.4 Probléma)

## B.6 Faktorelemzés

* faktor\_szorongas.omv - példa (Münnich és mtsai. (2006) 3.1. R-forráskód)
* faktor\_real\_human\_targyak.omv - példa (Münnich és mtsai. (2006) 3.9. R-forráskód)
* faktor\_bigfive.omv - példa (Münnich és mtsai. (2006) 3.21. R-forráskód)
* faktor\_kockazat.omv - példa (Münnich és mtsai. (2006) 3.7.4 Probléma)
* faktor\_fogkrem.omv - példa (Malhotra és Simon (2008) 617. oldal)

## B.7 Feltáró faktorelemzés

* faktor\_fogkrem.xlsx - A kutató arra volt kíváncsi, milyen előnyöket keresnek a fogyasztók a fogrémvásárlásnál. Egy 30 fős mintán a válaszadókat arra kérték, hogy jelezzék, mennyire értenek egyet a következő állításokkal (1 = egyáltalán nem ért egyet; 7 = teljes mértékben egyetért)
  + Az adatbázis szerkezete:
    - sorszam: válaszadó sorszáma (id)
    - v1: Fontos, hogy olyan fogkrémet vásároljak, amellyel megelőzhető a fogszuvasodás. (likert: 1-7)
    - v2: Az olyan fogkrémeket szeretem, amely fényessé teszi a fogaimat. (likert: 1-7)
    - v3: Egy fogkrémnek erősítenie kell a fogínyt. (likert: 1-7)
    - v4: Az olyan fogkrémeket szeretem, amely friss leheletet biztosít. (likert: 1-7)
    - v5: A fog romlásának megelőzése számomra nem fontos elvárás. (likert: 1-7)
    - v6: A legfontosabb szempont a fogkrém vásárlásánál a szép fog. (likert: 1-7)
  + Kapcsolódó állományok:
    - efa\_fogkrem.omv - Feltáró faktorelemzés jamovi-ban

# Appendix C — Irodalomjegyzék

Carver, C. S. és Scheier, M. F. (2006). *Személyiségpszichológia*. Osiris Kiadó.

Csallner, A. E. (2015). *Bevezetés az SPSS statisztikai programcsomag használatába*. <http://www.jgypk.hu/tamop15e/tananyag_html/spss/index.html>

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. <https://doi.org/10.1007/BF02289447>

Kárász, J. T., Nagy, O. N., Széll, K. és Takács, S. (2022). Cronbach-alfa: vele vagy nélküle? *Magyar Pszichológiai Szemle*, *77*, 81–98. <https://doi.org/10.1556/0016.2022.00004>

Ketskeméty, L. és Izsó, L. (2005). *Bevezetés az SPSS programrendszerbe - Módszertani útmutató és feladatgyűjtemény statisztikai elemzésekhez*. ELTE Eötvös Kiadó.

Malhotra, N. K. és Simon, J. (2008). *Marketingkutatás*. Akadémiai Kiadó.

Malkewitz, C. P., Schwall, P., Meesters, C. és Hardt, J. (2023). Estimating reliability: A comparison of Cronbach’s α, McDonald’s ωt and the greatest lower bound. *Social Sciences & Humanities Open*, *7*, 100368. <https://doi.org/10.1016/j.ssaho.2022.100368>

Moksony, F. (2006). *Gondolatok és adatok. Társadalomtudományi elméletek empirikus ellenőrzése*. Aula Kiadó.

Münnich, Á., Nagy, Á. és Abari, K. (2006). *Többváltozós statisztika pszichológus hallgatók számára*. Bölcsész Konzorcium. <http://psycho.unideb.hu/statisztika>

Nagy, O. N. (2006). A pszichológiai tesztek reliabilitása. In S. Rózsa, O. N. Nagy, és A. Oláh (Szerk.), *A pszichológiai mérés alapjai. Elmélet, módszer és gyakorlati alkalmazás*. Bölcsész Konzorcium. <https://mek.oszk.hu/05500/05536/05536.pdf>

Rózsa, S., Hupuczi, E., Martin, L., Birkás, B., Hartung, I., Hargitai, R., Varga, J., Láng, A., Tiringer, I. és Kállai, J. (2019). A Tellegen Abszorpciós Skála részletes pszichometriai elemzése. *Mentálhigiéné és Pszichoszomatika*, *20*, 35–77. <https://doi.org/10.1556/0406.20.2019.003>

Sajtos, L. és Mitev, A. (2007). *SPSS kutatási és adatelemzési kézikönyv*. Alinea Kiadó.

Székelyi, M. és Barna, I. (2002). *Túlélőkészlet az SPSS-hez. Többváltozós elemzési technikáról társadalomkutatók számára*. Typotex Kiadó.

Takács, S. (2017). *Bevezetés a matematikai statisztikába 2. Többváltozós statisztikai módszerek*. Antarész Kiadó.

Varga, A. (2019). *Többváltozós statisztika dióhéjban: Változó-orientált módszerek*. Pólya Kiadó.

Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *44*, 219–246. <https://doi.org/10.1177/0095798418771807>