Eötvös Loránd Tudományegyetem
Természettudományi Kar

Szórás- és kovarianciaanalízis alkalmazása szociológiai adatokra

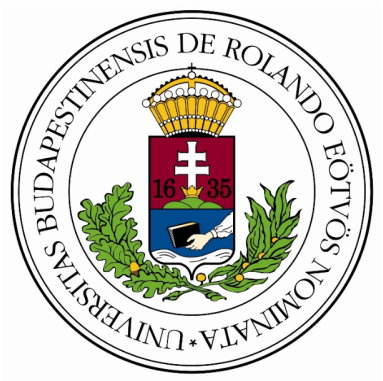
Szakdolgozat

Kelemen Kinga
Matematika BSc
Matematikai elemző szakirány

Témavezető:

Dr. Zempléni András

Valószínűségelméleti és Statisztika Tanszék



Budapest
2016

Tartalomjegyzék

1. Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőnek, Zempléni Andrásnak, hogy segítségével, hasznos tanácsaival és útmutatásával hozzájárult a szakdolgozatom elkészüléséhez. Külön köszönöm a konzultációkat, ahol mindig türelemmel fordult felém.
Ugyancsak köszönöm a TÁRKI-nak, hogy rendelkezésünkre bocsátotta a Háztartás Monitor vizsgálat adatait.
Hálás köszönettel tartozom a szüleimnek az egyetemi éveim alatt nyújtott kitartó támogatásukért.

2. Bevezetés

Szakdolgozatom témája a szórás- és kovarianciaanalízis alkalmazása szociológiai adatokra. A témaválasztás során elsődleges célom volt, hogy a matematika szerteágazó témakörei közül olyan témában mélyedjek el, amelyet más tudományok is alkalmaznak. A TÁRKI Háztartás Monitor adathalmazai lényegében a társadalom mindennapjairól szólnak. Ezeket az adathalmazokat egy- és többszempontos szórásanalízissel és kovarianciaanalízissel vizsgálom a szakdolgozatomban.
A két mintás t-próbák általánosításának tekinthető szórásanalízis, több, egyező szórású, normális eloszlású csoport átlagának összevetésére alkalmas statisztikai eljárás [1]. A szórásanalízisnek, mint statisztikai módszernek többféle elnevezése is van a szakirodalomban. Szóráselemzésként, varianciaanalízisként, varianciaelemzésként is nevezik, illetve több helyen csak ANOVA-ként hivatkoznak rá. Az ANOVA elnevezés az angol ANalysis Of VAriance kezdőbetűiből keletkezett rövidítés. Az első részben ismertetem a modell történeti hátterét, majd a következő fejezetben rendszerezem a változók különböző típusait, amely lényeges szempont az ANOVA alkalmazásánál. Ezután az alapfogalmak definiálása következik, majd rátérek az ANOVA feltételeire és azok ellenőrzésének módszereire. Ezt követően bemutatom az egy- és többszempontos szórásanalízis és kovarianciaanalízis matematikai, elméleti hátterét. Legvégül az ismertetett módszereket alkalmazom társadalomtudományi adatokra. Első lépésben azt nézem, hogy a vizsgált személyek neme hatással van-e a jövedelmükre. Ezt követően újabb változókat vonok be a vizsgálatba, az iskolai végzettséget és az életkort. Ezek külön-külön és együttes hatásait vizsgálom a függő változóra, a jövedelemre. Befejezésként a vizsgált időszak adatainak jövedelem és inflációs változását szemléltetem.

3. Az ANOVA modell történeti háttere

Először ismertetem az ANOVA módszer kialakulásának történetét, felhasználva főleg a [2], [4] és [5] forrásokat.
A varianciaanalízis a 20. században alakult ki, habár az előélete korábbi századokig nyúlik vissza. Ezalatt értendő a hipotézis vizsgálatok, a négyzetösszegek elkülönítése, egyéb kísérleti technikák és az additív modell(AM). Az első statisztikai hipotézisvizsgálat idejét nehéz pontosan meghatározni, de az időszámításunk előtti ősi Kroában történt népszámlálás alapján úgy gondolták a nemek születési aránya 50-50 %. Az 1700-as években több évtizednyi népszámlálási adat birtokában John Arbuthnot, angol matematikus rámutatott arra, hogy el kell vetni ezt a hipotézist, ez keltette fel Laplace érdeklődését is [3]. A legkisebb négyzetek elve fejlődése Gauss és Laplace nevéhez köthető, ennek segítségével fejlődött ki egy olyan módszer, ami a megfigyelések vizsgálatát segítette. Gyakorlati alkalmazásai megjelennek a geodéziában és az asztronómiában is. Így több tanulmány született a négyzetösszegekről. Laplace hamar rájött hogyan tudja megbecsülni a szórást a reziduális (inkább mint a totális) négyzetösszegekből. 1827-ben Laplace a legkisebb négyzetek módszerét használva feladatként azonosította az ANOVA problémát atmoszferikus árapály mérésekre vonatkozóan.
Az ANOVA alkotója egy brit statisztikus, Sir Ronald Aylmer Fisher, aki egy angliai mezőgazdasági kísérleti állomáson dolgozott. Fisher ismerte fel először, hogy a nullhipotézis, a H_0 úgy is vizsgálható több csoporton együtt végzett kísérletben, hogy egymástól függetlenül kiszámítjuk a minta varianciájának becslését kétféleképpen. Az egyik módszer, amikor a csoporton belüli szóródásból, a másik módszer, amikor a csoportok közötti szóródásból végzünk becslést. H_0 érvényessége esetén a két módszerrel számított becslés ugyanannak a mennyiségnek a becslése. Amennyiben a H_0-t elvetjük és az elsőfajú hiba valószínűsége kicsi, akkor a csoportok között nagy valószínűséggel van különbség.
A varianciaanalízis akkor vált széles körben ismertté, amikor megjelent 1925-ben Fisher könyve a Statistical Methods for Research Workers címmel. A varianciaanalízis kifejezést is ő alkotta meg. Az ANOVA használja a Fisher féle F eloszlást a statisztikai szignifikancia teszt részeként. Fisher híres írásai közé soroljuk a "On the mathematical foundations of theoretical statistics" cikkét, amely megjelent 1922-ben a Philosophical Transactions of the Royal Society tudományos folyóiratban, illetve az 1925-ben megjelent "Applications of Student's distribution" című írása is mérföldkőnek számít a módszer történetében.

4. Az ANOVA modell elméleti háttere

4.1. Változók típusai

A matematikában a változókat 4 féle csoportba oszthatjuk a mérési szintjüktől függően. Ebben a fejezetben főleg a [6], [7] és [8] forrásokat használtam fel.
Ez a 4 mérési szint:
Nominális mérési szint: a változót csoportokba osztjuk egy tulajdonság alapján, amelyek között nem tudunk felállítani sorrendiséget. Például nemi hovatartozás(férfi/nő), ebben az esetben a nem a változó. Nominális mérési szintű változónál átlag és medián számítást nem lehet vizsgálni, de móduszt lehet számolni.
Ordinális mérési szint: a változók kategorizálása mellett sorrendiséget tudunk felállítani a kategóriák között, de az ezek közötti különbséget nem tudjuk számszerűsíteni, azaz két értékpár távolságát nem tudjuk meghatározni. Például településtípus(tanya/falu/község/város/főváros). A nominális mérési szinthez képest itt már a medián számításról van értelme beszélni, de a számtani átlag itt sem értelmezhető.
Intervallumskála: a sorba rendezhetőség mellett itt már értelmezhető két értékpár távolsága, de ezek az értékek az arányosságot nem fejezik ki. A zérus megválasztása megegyezésen alapul, mint Celsius foknál a víz fagyáspontja. Pl. Celsius fok: a 40° nem kétszer melegebb a 20°-nál. Itt már van értelme átlagról beszélni.
Arányskála: az előbbi mérési szinthez képest itt már az arányosság is érvényes az értékek között és itt már a nullapont megválasztása nem megegyezésen alapul. Például jövedelem, súly, magasság stb.

	Nominális	Ordinális	Intervallumskála	Arányskála
Módusz számítás	igen	igen	igen	igen
Medián számítás	nem	igen	igen	igen
Átlag számítás	nem	nem	igen	igen
Osztás	nem	nem	nem	igen

1. táblázat.

A táblázatban szereplő "igenek" arra utalnak, hogy az adott mérési szinten végrehajthatóak-e az egyes statisztikai számítások, míg a "nemek" azt jelentik, hogy nem hajthatók végre.
Alacsony mérési szintűnek nevezzük a nominális és az ordinális mérési szintű változókat, illetve magas mérési szintűnek nevezzük az intervallumskála és az arányskála típusúakat.
Az ANOVA modellben a függő változókat szeretnénk megmagyarázni a független változók segítségével, azonban a társadalomtudományokban nem

olyan egyértelmű/egyszerű meghatározni a függő/független változókat, mint például egy fizikai jelenségnél. Előfordulhat, hogy két változó között csak "látszólagos" kapcsolat van és valójában egy harmadik változó bevonásával már teljesen más eredményt kapunk [8].
A változókat kategorizálhatjuk aszerint is, hogy diszkrét vagy folytonos változóról beszélünk.

4.1. Definíció. (Diszkrét valószínűségi változó) Értékkészlete legfeljebb megszámlálhatóan végtelen, azaz $\{x_1, \ldots, x_n, \ldots\}$ elemekből áll.

Például a családonkénti gyerekszám diszkrét valószínűségi változó [9].

4.2. Definíció. (Folytonos valószínűségi változó) Az X valószínűségi változó folytonos, ha az eloszlásfüggvénye folytonos függvény.

A jövedelem például egy folytonos valószínűségi változó [10].
A társadalomtudományokban jelentős szerepe van az elemzési egységnek, azaz hogy a vizsgálat középpontjában mi áll. Állhat az egyén, a család, kisebb közösség, település, régió, ország, kontinens stb. Az elemzési egységnek fontos szerepe van a társadalmi kutatásoknál, ugyanis egy adott közösségre vonatkozó jellemzőkből nem vonhatunk le következtetéseket az egyénre (ökológiai tévkövetkeztetés) [8].
A változók mérési szintjei a későbbiekben fontos szerepet játszanak, mivel a varianciaanalízis alkalmazásakor a magyarázó változók csak alacsony mérési szintűek lehetnek. Azonban magas mérési szintű változó diszkretizálás után már lehet faktor.

4.2. Alapfogalmak

Ez a fejezet főleg a [13] és a [15] forrásokon alapszik.
Faktor: a kutatásban vizsgált független változók pl. különböző iskolai végzettségűek.
Faktor szint: A faktor értékkészletének az eleme, amely beállítása mellett vizsgálhatjuk meg a függő változónkat pl. iskolai végzettség esetében az érettségivel rendelkezők.
Diszkretizálás: Folytonos változó esetében alkalmazható, amikor a folytonos tartományt intervallumokra bontjuk.
Homoszkedasztikusság, másnéven homogenitás: A csoportokon belül a függő változó szórása azonos, szignifikáns különbség nincs közöttük.
Bootstrap statisztikai eljárás: Újramintavételezési eljárás, becslések szórásának a vizsgálatára is alkalmazható.
Egyszempontos varianciaanalízis: Varianciaanalízis, ahol csak egy faktor van.
Többszempontos varianciaanalízis: Varianciaanalízis, ahol kettő vagy több faktor van.

Interakció: Többszempontos varianciaanalízis esetében az interakció azt jelenti, hogy a tényezők között van kölcsönhatás, tehát a szempontok hatása nem független.
ANCOVA, azaz a kovarianciaanalízis: olyan elemzéseket nevezzünk így, ahol még kovariánsokat(folytonos magyarázó változó) is bevonunk a vizsgálatba.
Egyszempontos szórásanalízis esetében a minta:
1.csoport $N(\mu_1, \sigma^2)X_{1,1}X_{2,1}\dots X_{n_1,1}$
2.csoport $N(\mu_2, \sigma^2)X_{1,2}X_{2,2}\dots X_{n_2,2}$
3.csoport $N(\mu_3, \sigma^2)X_{1,3}X_{2,3}\dots X_{n_3,3}$
...
k.csoport $N(\mu_k, \sigma^2)X_{1,k}X_{2,k}\dots X_{n_k,k}$
Ahol a csoportok normális eloszlásúak, a μ_i $(i = 1, \dots k)$ a csoportok várható értékét, a σ^2 pedig a szórásnégyzetet jelöli.

4.3. Az ANOVA feltételeinek ellenőrzése

Az osztályokba tartozó megfigyeléseket függetlennek, közös szórásúnak és normális eloszlásúnak feltételezzük, a várható érték az lehet különböző. Ezek a feltételek mind kellenek az F-próbák használatához. Ezeket a [15] és [16] források alapján közelítem meg.
A normalitás ellenőrzése történhet grafikusan vagy numerikusan:
Normalitás vizsgálatra többféle statisztikai teszt létezik. A numerikus tesztekhez soroljuk a Kolmogorov-Szmirnov tesztet, Cramér-von Mises-tesztet, Anderson-Darling-próbát és Shapiro-Wilk tesztet is. A leggyakrabban használt teszt eloszlásvizsgálatokra a Kolmogorov-Szmirnov teszt. A teszt előnye, hogy eloszlásfüggetlen, a hátránya viszont, hogy kicsi az ereje.
A normalitást szokás grafikusan is megjeleníteni, tesztelni. Érdemes hisztogramon ábrázolni az adott változót, így a hisztogram alakjáról lehet következtetéseket levonni. Az osztópontok sűrítésével a hisztogram nem lesz annyira durva, míg az osztópontok ritkításával nem lesz a hisztogramba olyan sok ugrás. Grafikus vizsgálatoknál elterjedt módszer a Q-Q (kvantilis-kvantilis) ábra készítése. A Q-Q ábra készítésénél először az alapadatok standardizálása történik, majd ezek alapján elkészül az empirikus eloszlásfüggvény. Végül az eloszlásfüggvény értékeit a normális eloszlásfüggvény szerint kell transzformálni. Amennyiben a vizsgált változó normális vagy közelítőleg normális (valós adatoknál jellemzőbb eset), akkor a pontok az origón átmenő 45 fokos egyenes körül szóródnak. Ez a fajta tesztelés nem annyira szigorú, mint a numerikus tesztek, mivel a döntésmeghozatal vizuálisan történik [22].
Szórás azonosság ellenőrzését Levene-teszttel lehet megvizsgálni. Az alkalmazás során a beépített függvényét fogom használni az R-ben. Amennyiben

a Levene-teszt szignifikáns, szimulációkkal fogom vizsgálni a szórásanalízis együtthatóit, illetve p-értékeit.
Az ANOVA hipotézisvizsgálatnál kulcsfontosságú statisztikák függetlenségét a Fisher-Cochran tétel biztosítja, a tétel kimondása és bizonyítása előtt azonban ismertettek pár definíciót és tételt a [19] alapján.

4.3. Definíció. (Kvadratikus alak) A kvadratikus alak egy homogén másodfokú polinom.

$$Q = Q(x_1, \ldots, x_n) = Q(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}x_i x_j, ahol \quad m_{ij} \in \mathbb{R}$$

A kvadratikus alak mátrixos felírása:

$$Q = \mathbf{x}^T M\mathbf{x}, ahol \quad M = (m_{ij}) \quad (i = 1, \ldots, n, j = 1, \ldots, n)$$

A későbbiekben M egy $n \times n$-es szimmetrikus mátrix. A kvadratikus alak rangja az M mátrix rangja.

4.4. Definíció. Legyenek X_1, X_2,..., X_n független, standard normális eloszlású valószínűségi változók. $X_j \sim N(\mu_j, 1)$, $j = 1, \ldots, n$. Ekkor az

$$Y_n = X_1^2 + X_2^2 + \ldots + X_n^2$$

valószínűségi változó n szabadságfokú χ^2-eloszlású.

4.5. Tétel. *Legyenek X_n és X_m független χ^2 eloszlású valószínűségi változók n, illetve m szabadsági fokkal. A két valószínűségi változó összege is χ^2 eloszlású, a szabadsági fokok pedig összeadódnak, vagyis $n + m$.*

Bizonyítás. Legyenek X_1, X_2,...,X_{n+m} független, standard normális eloszlásúak: $X_j \sim N(\mu_j, 1)$, ahol $j = 1, \ldots, n$

$$Y_n = X_1^2 + X_2^2 + \ldots + X_n^2 \quad Y_m = X_{n+1}^2 + X_{n+2}^2 + \ldots + X_{n+m}^2$$

$$Y_n + Y_m = X_1^2 + X_2^2 + \ldots + X_{n+m}^2$$

$n + m$ szabadsági fokú χ^2 eloszlású. □

4.6. Tétel. *Legyenek Q_j-k $(j = 1, \ldots, k)$ az x_i-k $(i = 1, \ldots, n)$ változók kvadratikus formái. Tegyük fel, hogy $rang(Q_j) = n_j$ és*

$$Q_1 + Q_2 + \ldots + Q_k = \sum_{i=1}^{n} x_i^2$$

Ha $n_1 + n_2 + \ldots + n_k = n$ akkor és csak akkor $\exists$ olyan M ortogonális mátrix, amelyre igaz az, hogy $\boldsymbol{b} = M\boldsymbol{x}$, ahol $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)^T$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ és

$$Q_1 = b_1^2 + \ldots + b_{n_1}^2, \quad Q_2 = b_{n_1+1}^2 + \ldots + b_{n_1+n_2}^2, \ldots$$

$$Q_k = b_{n_1+\ldots+n_{k-1}+1}^2 + \ldots + b_{n_1+\ldots+n_k}^2$$

tejesül.

4.7. Tétel. (Fisher-Cochran) *Legyen adva* $\boldsymbol{X} = (X_1, X_2, ..., X_n)^T$ *véletlen vektor, ahol* X_i $(i = 1, \ldots, n)$ *független, standard normális eloszlású valószínűségi változók, és definiáljuk a segítségükkel a* $Q = \boldsymbol{X}^T \boldsymbol{I}_n \boldsymbol{X} = \boldsymbol{X}^T \boldsymbol{X} = \sum_{i=1}^{n} X_i^2$ *és a* $Q_j = \boldsymbol{X}^T \boldsymbol{M}_j \boldsymbol{X} (j = 1, .., k)$ *kvadratikus alakokat, ahol* $\boldsymbol{M}_j$ *szimmetrikus* $n \times n$*-es mátrixok* $(j = 1, ..., k \leq n)$*. Tegyük fel, hogy érvényes*

$$Q = Q_1 + Q_2 + ... + Q_k$$

azonosság. Legyen Q_j *rangja: rang*$(M_j) = n_j$*. A* Q_j $(1 \leq j \leq k)$ *kifejezések független,* χ^2*- eloszlásúak* n_j $(1 \leq j \leq k)$ *szabadságfokkal, pontosan akkor, ha*

$$\sum_{j=1}^{k} n_j = n$$

teljesül.

Bizonyítás.
$\Rightarrow$
Legyenek a Q_j-k függetlenek és az eloszlásuk $\chi^2_{n_j}$, $(j = 1, \ldots, k)$. A 4.5 tételből tudjuk, hogy a $Q_1 + \ldots + Q_k$ eloszlása $\chi^2_{n_1+\ldots+n_k}$. Azonban azt is tudjuk, hogy $Q_1 + Q_2 + \ldots + Q_k = X_1^2 + X_2^2 + \ldots + X_n^2$, amelynek a 4.4. definíció szerint az eloszlása χ^2_n. Tehát $n_1 + n_2 + \ldots + n_k = n$.
$\Leftarrow$
Legyen $n_1 + n_2 + \ldots + n_k = n$. A 4.6 tétel alapján $\exists$ olyan M ortogonális mátrix, hogy az $\mathbf{Y} = M\mathbf{X}$-re, ahol $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$

$$Q_1 = Y_1^2 + Y_2^2 + \ldots + Y_{n_1}^2 \quad Q_2 = Y_{n_1+1}^2 + \ldots + Y_{n_1+n_2}^2, \ldots$$

$$Q_k = Y_{n_1+\ldots+n_{k-1}+1}^2 + \ldots + Y_{n_1+\ldots+n_k}^2.$$

Viszont $\mathbf{X} \sim N_n(0, I)$, ezért $\mathbf{Y} = M\mathbf{X} \sim N_n(M * 0, MM^T) = N_n(0, I)$, mivel M ortogonális mátrix.
Tehát az $\mathbf{Y}$ koordinátái $N(0, 1)$ eloszlásúak és függetlenek. A $Q_1, Q_2, \ldots, Q_k$ pedig n_j darab ilyenek négyzetösszege, vagyis $\chi^2_{n_j}$ eloszlású. A $Q_1, Q_2, \ldots, Q_k$ függetlenek, mivel különböző Q_j-k előállításában azonos n_j-k nem vesznek részt. $\square$

4.4. A modell felépítése

4.4.1. Egyszempontos szórásanalízis

Döntésmeghozatal előtt feltételezéseket fogalmazzunk meg és ezek igaz/hamis voltára vagyunk kíváncsiak. Ilyenkor segítenek a statisztikai hipotézisvizsgálatok, amelyek a minta alapján kiszámolhatóak egy megadott szignifikancia szint mellett. A szignifikanciaszintet (jelölése:α) a modell vizsgálatánál előre meg kell határozni. Általában $\alpha = 0,05$-nek szokás megválasztani.

A modell felépítése során a [11], [13], [12], [14], [15] és a [16] forrásokat használom.
Az ANOVA modellben azt vizsgáljuk meg, hogy egy faktornak, körülménynek van-e hatása a kimeneti változó várható értékére. A faktor különböző szintekre való beállítása után méréseket végzünk. Majd kimondjuk a nullhipotézist.
Egy(későbbiekben több) szempont alapján k csoportba osztjuk az adatokat. A csoportok létrehozatalánál fontos, hogy a faktor, tehát a szempont, amely alapján csoportokat csinálunk, az alacsony mérési szintű változó legyen. Folytonos változót csak diszkretizálás után tehetünk faktorrá. A csoportok mintaelemszáma nem feltétlenül egyezik meg, ezt jelölje n_i, ahol az i az i. csoportra utal, az teljes minta elemszáma pedig legyen $n = \sum_{i=1}^{k} n_i$. Az i. csoportban az $X_i \sim N(\mu_i, \sigma^2)$ valószínűségi változóra vett mintaelemeket

$$X_{ij} \sim N(\mu_i, \sigma^2) \qquad (j = 1, \ldots, n_i)$$

jelöli. A várható értékekre vezessünk be egy célszerű felbontást:

$$\mu_i = m + a_i$$

ahol az a_i az i. csoport hatása, az m pedig a várható értékek súlyozott átlaga, vagyis $m = \frac{1}{n} \sum_{i=1}^{k} n_i \mu_i$.
Az egyszempontos modell egyenlete:

$$X_{ij} = m + a_i + \varepsilon_{ij} \qquad (j = 1, \ldots, n_i, i = 1, \ldots, k)$$

ahol az ε_{ij} a véletlen hatást/hibát jelöli. A szóráselemzés egy lineáris modell, így

$$\mathbf{Y} = \mathbf{Ba} + \mathbf{1}m + \vec{\varepsilon}$$

ahol $\mathbf{Y} := (X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}, X_{k1}, \ldots, X_{kn_k})^T$, $\mathbf{a} := (a_1, \ldots, a_k)^T$ $\vec{\varepsilon} := (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \varepsilon_{21}, \ldots, \varepsilon_{2n_2}, \varepsilon_{k1}, \ldots, \varepsilon_{kn_k})^T$, $\mathbf{1} \in \mathbb{R}^n$ vektor és $\mathbf{B}$ pedig egy 0-1-esekből álló ún. struktúramátrix. Az egyszempontos varianciaanalízis esetében a mátrix oszlopainak a száma megegyezik a csoportok k számával. A sorok pedig az n_i-ket jelöli. A következő struktúramátrixban $k = 3$, $n_1 = 2$, $n_2 = 3$ és $n_3 = 4$.

$$\mathrm{B} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

A legkisebb négyzetek módszerét használva keressük a minimumát a

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-m-a_i)^2$$

kifejezésnek. Legyen a csoportátlag $\overline{X}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, ahol $i = 1,\ldots,k$, illetve legyen a teljes mintaátlag $\overline{X}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$. Tehát a korábban használt paraméterek becslései $\hat{m} = \overline{X}_{..}$ és $\hat{a}_i = \overline{X}_{i.} - \overline{X}..$, ahol $i = 1,\ldots,k$. Visszahelyettesítve a legkisebb négyzetes módszernél felírt egyenletbe:

$$SS_{csb} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\hat{m}-\hat{a}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i.})^2$$

A mintaelemek teljes mintaátlagtól vett eltéréseinek négyzetösszege (jelölése: SS) felbontható a csoporton belüli (jelölése: SS_{csb}) és a csoportok közötti (jelölése:SS_{csk}) részre:

$$SS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(X_{ij}-\overline{X}_{i.})+(\overline{X}_{i.}-\overline{X}_{..})]^2 =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i.})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i.})(\overline{X}_{i.}-\overline{X}_{..}) + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{X}_{i.}-\overline{X}_{..})^2 =$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i.})^2 + \sum_{i=1}^{k} n_i(\overline{X}_{i.}-\overline{X}_{..})^2 = SS_{csb} + SS_{csk}$$

H_0:(nullhipotézis)a várható értékek egyenlőek, azaz a faktornak nincs hatása $\mu_1 = \mu_2 = \ldots = \mu_k$
H_1:(ellenhipotézis)a várható értékek nem egyenlőek, van legalább 2 olyan várható érték, amely nem egyenlő $\exists\, i,j : \mu_i \neq \mu_j$

	Nullhipotézis igaz	Nullhipotézis hamis
Elfogadjuk a nullhipotézist	Helyes döntés	Másodfajú hiba
Elutasítjuk a nullhipotézist	Elsőfajú hiba	Helyes döntés

2. táblázat.

A szabadsági fokok:
A teljes szórásnégyzet szabadságfoka(jelölése: df): $n-1$
A csoporton belüli szórásnégyzet szabadságfoka: $n-k$
A csoportok közötti szórásnégyzet szabadságfoka: $k-1$

F-próbával ellenőrizzük a szórások egyezését.

$$F = \frac{SS_{csk}(n-k)}{SS_{csb}(k-1)}$$

$F > F_{n-k,k-1,\alpha}$-ra elutasítjuk a próbát. Amint azt a Fisher-Cochran tételnél láttuk, H_0 érvényessége esetén két független, χ^2 eloszlásból kapjuk a képletet, tehát valóban F eloszlású. A végeredményt táblázatban szokás összefoglalni. Egyszempontos ANOVA-tábla:

Forrás	SS	df	MS	F	p-érték
Hatás(csk)	SS_{csk}	$k-1$	$s^2_{csk} = \frac{SS_{csk}}{k-1}$	$\frac{s^2_{csk}}{s^2_{csb}}$	$P\left(F > \frac{s^2_{csk}}{s^2_{csb}}\right)$
Hiba(csb)	SS_{csb}	$n-k$	$s^2_{csb} = \frac{SS_{csb}}{n-k}$		

3. táblázat.

Az R^2 együttható:

$$R^2 = \frac{SS_{csk}}{SS}$$

Az R^2 együttható írja le, hogy mekkora a megmagyarázott szórásnégyzet részaránya.

4.4.2. Kétszempontos szórásanalízis interakcióval és anélkül

A kétszempontos varianciaanalízis vizsgálatánál két különböző szempont alapján vizsgálódunk. Az egyik szempont szerint legyen k, a másik szempont szerint pedig p lehetséges érték. Az egyszempontos esethez képest a lineáris modellben megjelenik egy újabb tag:

$$X_{ij} = m + a_i + b_j + \varepsilon_{ij} \qquad (i = 1, \ldots, k; j = 1, \ldots, p)$$

Az a_i-k jelölik az egyik, míg a b_j-k a másik szempont egymástól független hatását. A többi jelölést az egyszempontos esetben már definiáltam. A $\mathbf{B}$ struktúramátrix segítségével az előbbi lineáris modell:

$$\mathbf{Y} = \mathbf{B}\vec{ab} + \mathbf{1}m + \vec{\varepsilon}$$

ahol $\mathbf{Y} := (X_{11}, \ldots, X_{1p}, X_{21}, \ldots, X_{2p}, X_{k1}, \ldots, X_{kp})^T$, $\vec{ab} := (a_1, \ldots, a_k, b_1, \ldots, b_p)^T$ $\vec{\varepsilon} := (\varepsilon_{11}, \ldots, \varepsilon_{1p}, \varepsilon_{21}, \ldots, \varepsilon_{2p}, \varepsilon_{k1}, \ldots, \varepsilon_{kp})^T$, $\mathbf{1} \in \mathbb{R}^n$ vektor. A $\mathbf{B}$ struktúramátrix kétszempontos varianciaanalízis esetén kölcsönhatás nélkül:

$$\mathrm{B} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

A fenti struktúramátrix esetében $k = 3$ és $p = 3$. Az első 3 oszlop a k-t jelöli, az utolsó három oszlop pedig a p-t jelöli. Az egyszempontos esethez hasonlóan itt is a legkisebb négyzetek módszerével becsüljük a paramétereket:

$$\sum_{i=1}^{k}\sum_{j=1}^{p}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p}(X_{ij} - m - a_i - b_j)^2$$

A fenti kifejezésnek a minimumát szeretnénk meghatározni. Legyen az egyik szempont szerinti csoportátlag $\overline{X}_{i.} = \frac{1}{p}\sum_{j=1}^{p} X_{ij}$, ahol $i = 1, \ldots, k$, a másik szempont szerinti csoportátlag $\overline{X}_{.j} = \frac{1}{k}\sum_{i=1}^{k} X_{ij}$, ahol $j = 1, \ldots, p$, illetve legyen a teljes mintaátlag $\overline{X}_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} X_{ij}$. A paraméterek legkisebb négyzetes becslései:

$$\begin{aligned} \hat{m} &= \overline{X}_{..} \\ \hat{a_i} &= \overline{X}_{i.} - \overline{X}_{..}, \text{ahol} \quad \mathrm{i} = 1, \ldots, \mathrm{k} \\ \hat{b_j} &= \overline{X}_{.j} - \overline{X}_{..}, \text{ahol} \quad \mathrm{j} = 1, \ldots, \mathrm{p} \end{aligned}$$

Az előbb meghatározott paraméterek alapján a kifejezés minimuma:

$$SS_{csb} = \sum_{i=1}^{k}\sum_{j=1}^{p}(X_{ij} - \hat{m} - \hat{a}_i - \hat{b}_j)^2$$

Kétszempontos esetben is a mintaelemek teljes mintaátlagtól vett eltéréseinek a négyzetösszege felbontható csoportok közötti (SS_a, illetve SS_b a kétféle szempont szerinti csoportosításban) és a csoportokon belüli (SS_{csk}) reziduális részre.

$$SS = SS_a + SS_b + SS_{csb}$$

Ebben az esetben kétféle nullhipotézist fogalmazhatunk meg [18]:
$H_0^{(1)} : \mu_1^{(1)} = \mu_2^{(1)} = \ldots = \mu_k^{(1)} = 0$, vagyis az első szempont szerinti k csoport a függő változó átlagára nézve mind azonos, tehát az átlagok között nincs különbség.
A másik nullhipotézis a másik szempontra vonatkozik:

$H_0^{(2)} : \mu_1^{(2)} = \mu_2^{(2)} = \ldots = \mu_p^{(2)} = 0$, vagyis a második szempont szerinti p csoport a függő változó átlagára nézve mind azonos, az átlagok között nincs különbség.
Az eredményeket a kétszempontos ún. ANOVA-tábla foglalja magába:

Forrás	SS	df	MS	F	p-érték
a-hatás(csk)	SS_a	$k-1$	$s_a^2 = \frac{SS_a}{k-1}$	$\frac{s_a^2}{s_{csb}^2}$	$P\left(F > \frac{s_a^2}{s_{csb}^2}\right)$
b-hatás(csk)	SS_b	$p-1$	$s_b^2 = \frac{SS_b}{p-1}$	$\frac{s_b^2}{s_{csb}^2}$	$P\left(F > \frac{s_b^2}{s_{csb}^2}\right)$
Hiba(csb)	SS_{csb}	$(k-1)(p-1)$	$s_{csb}^2 = \frac{SS_{csb}}{(k-1)(p-1)}$		

4. táblázat.

A kétszempontos varianciaanalízis interakcióval esetben szintén kp csoport van, de itt a lineáris modellben egy újabb tag jelenik meg a $(ab)_{ij}$-k, amely az interakciókat jelöli.

$$X_{ijl} = m + a_i + b_j + (ab)_{ij} + \varepsilon_{ijl}$$

A $\mathbf{B}$ struktúramátrix segítségével az

$$\mathbf{Y} = \mathbf{B}\overrightarrow{ab(ab)} + \mathbf{1}m + \vec{\varepsilon}$$

lineáris modell alakját ölti. A $\mathbf{B}$ struktúramátrix alakja megváltozik az interakció nélküli esethez képest. Az alábbi mátrixnál $k = 2$ és $p = 3$, ez első két oszlop a k-t jelöli, a következő három oszlop a p-t és a többi kp darab oszlop az interakciót. Az alábbi példa mátrixban az ismétlések száma 2, ezért van minden sorból 2.

$$\mathrm{B} = \begin{pmatrix}
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

A legkisebb négyzetek módszerével történő paraméter becslés hasonló interakció nélküli esethez. A varianciafelbontás kiegészül egy újabb elemmel:

$$SS = SS_a + SS_b + SS_{ab} + SS_{csb}$$

A kétszempontos interakció nélküli varianciaanalízis esethez képest egy harmadik nullhipotézist is megfogalmazhatunk, amely azt állítja, hogy nincs interakció. Az ANOVA-táblában is vannak változások:

Forrás	SS	df	MS	F	p-érték
a-hatás(csk)	SS_a	$k-1$	$s_a^2 = \frac{SS_a}{k-1}$	$\frac{s_a^2}{s_{csb}^2}$	$P\left(F > \frac{s_a^2}{s_{csb}^2}\right)$
b-hatás(csk)	SS_b	$p-1$	$s_b^2 = \frac{SS_b}{p-1}$	$\frac{s_b^2}{s_{csb}^2}$	$P\left(F > \frac{s_b^2}{s_{csb}^2}\right)$
ab-interakció	SS_{ab}	$(k-1)(p-1)$	$s_{ab}^2 = \frac{SS_{ab}}{(k-1)(p-1)}$	$\frac{s_{ab}^2}{s_{csb}^2}$	$P\left(F > \frac{s_{ab}^2}{s_{csb}^2}\right)$
Hiba(csb)	SS_{csb}	$kp(n-1)$	$s_{csb}^2 = \frac{SS_{csb}}{kp(n-1)}$		

5. táblázat.

4.4.3. Többszempontos szórásanalízis

Három vagy több szempontra is működik az ANOVA. Háromszempontos modell esetében vizsgálni kell az összes kétszeres, illetve háromszoros interakciót. Három, illetve többtényezős kísérleteknél többféle módszer létezik: véletlen blokkelrendezés, kétszeresen osztott parcellás elrendezés (split-split-plot), osztott sávos elrendezés (split-strip plot) és latin négyzet elrendezés. A három szempont mindenféle kombinációját ismétlésen belül véletlenszerűen rendezzük el, amikor a véletlen blokkelrendezést használjuk. Ebben az esetben a matematikai modell egyenlete:

$$X_{ijkl} = m + a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{ij} + (abc)_{ijk} + \varepsilon_{ijkl}$$

A háromszempontos ANOVA-tábla:

Forrás	SS	df	MS	F	p-érték
a-hatás(csk)	SS_a	$k-1$	$s_a^2 = \frac{SS_a}{k-1}$	$\frac{s_a^2}{s_{csb}^2}$	$P\left(F > \frac{s_a^2}{s_{csb}^2}\right)$
b-hatás(csk)	SS_b	$p-1$	$s_b^2 = \frac{SS_b}{p-1}$	$\frac{s_b^2}{s_{csb}^2}$	$P\left(F > \frac{s_b^2}{s_{csb}^2}\right)$
c-hatás(csk)	SS_c	$c-1$	$s_c^2 = \frac{SS_c}{c-1}$	$\frac{s_c^2}{s_{csb}^2}$	$P\left(F > \frac{s_c^2}{s_{csb}^2}\right)$
ab-interakció	SS_{ab}	$(k-1)(p-1)$	$s_{ab}^2 = \frac{SS_{ab}}{(k-1)(p-1)}$	$\frac{s_{ab}^2}{s_{csb}^2}$	$P\left(F > \frac{s_{ab}^2}{s_{csb}^2}\right)$
ac-interakció	SS_{ac}	$(k-1)(c-1)$	$s_{ac}^2 = \frac{SS_{ac}}{(k-1)(c-1)}$	$\frac{s_{ac}^2}{s_{csb}^2}$	$P\left(F > \frac{s_{ac}^2}{s_{csb}^2}\right)$
bc-interakció	SS_{bc}	$(p-1)(c-1)$	$s_{bc}^2 = \frac{SS_{bc}}{(p-1)(c-1)}$	$\frac{s_{bc}^2}{s_{csb}^2}$	$P\left(F > \frac{s_{bc}^2}{s_{csb}^2}\right)$
abc-interakció	SS_{abc}	$(k-1)(p-1)(c-1)$	$s_{abc}^2 = \frac{SS_{abc}}{(k-1)(p-1)(c-1)}$	$\frac{s_{abc}^2}{s_{csb}^2}$	$P\left(F > \frac{s_{abc}^2}{s_{csb}^2}\right)$
Hiba(csb)	SS_{csb}	$kpc(n-1)$	$s_{csb}^2 = \frac{SS_{csb}}{kpc(n-1)}$		

6. táblázat.

4.4.4. Kovarianciaanalízis (ANCOVA)

A szóráselemzésnél egy magas mérési szintű változót vizsgálunk egy alacsony mérési szintű változó függvényében. A kovarianciaanalízis esete nagyon hasonló, itt azonban az alacsony mérési szintű változó mellett megjelenik egy folytonos (magas mérési szintű) változó is, ezt nevezzük kovariánsnak (jelölése: y). A legegyszerűbb esetben egy, bonyolultabb esetekben több kovariáns is bevonható a vizsgálatba. Az ANCOVA modellnek két alkalmazási feltétele van. Az egyik az, hogy a kovariáns lineáris kapcsolatban legyen a függő változóval. A másik szempont szerint a kovariáns értéke nem függhet az alkalmazott tényezőktől, szempontoktól. Ehhez a részhez a [17] forrást használtam.
A kétszempontos lineáris modell egy kovariáns bevonásával:

$$X_{ij\ell} = m + a_i + b_j + (ab)_{ij} + \beta y_{ij\ell} + \varepsilon_{ij\ell}$$

A fenti modellben $X_{ij\ell}$ a függő változó értéke, az m a fix hatású főátlag, az a_i és b_j az egyik, illetve a másik szempont szerinti hatás, a $(ab)_{ij}$ a két szempont kölcsönhatása, a β a függőváltozó és a kovariáns közötti lineáris regressziós együttható, y_{ijk} a kovariáns értékei és végül az $\varepsilon_{ij\ell}$ a hibát jelöli. A csoportok száma kp, mivel az egyik szempont k részre, a másik szempont pedig p részre osztja a mintát. Az n a minta elemszámát jelöli, n_j a csoportokban a megfigyelésszámot, így $\sum_{j=1}^{kp} n_j = n$. Az ANCOVA modell végrehajtásához több mindent ki kell számolni:
A mintákon belüli eltérés-négyzetösszegek y-ra, vagyis a kovariánsra nézve, a teljes mintára:

$$SS_{csb-y} = \sum_{j=1}^{kp} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2$$

A mintákon belüli eltérés-négyzetösszegek x-re, vagyis a függő változóra nézve, a teljes mintára:

$$SS_{csb-x} = \sum_{j=1}^{kp} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$$

A mintákon belüli eltérés-keresztszorzatok a összege a mintára:

$$SS_{csb-xy} = \sum_{j=1}^{kp} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)(x_{ij} - \overline{x}_j)$$

A teljes eltérés-négyzetösszeg y-ra:

$$SS_y = \sum_{j=1}^{kp} \sum_{i=1}^{n_j} (y_{ij} - \overline{y})^2$$

A teljes eltérés-négyzetösszeg x-re:

$$SS_x = \sum_{j=1}^{kp} \sum_{i=1}^{n_j} (x_{ij} - \overline{x})^2$$

A teljes keresztszorzat összeg:

$$SS_{xy} = \sum_{j=1}^{kp} \sum_{i=1}^{n_j} (y_{ij} - \overline{y})(x_{ij} - \overline{x})$$

Az SS_{csb-y_j} a mintákon belüli eltérés-négyzetösszegek az egyes mintákban, a SS_{csb-xy_j} a mintákon belüli eltérés-keresztszorzatok összege mintánként.
Az ANCOVA feltételeinek a vizsgálatához egy-egy F-próba szükséges.
Az első feltételnek a próbafüggvénye:

$$F = \frac{\left(\sum_{j=1}^{kp} \frac{(SS_{csb-xy_j})^2}{SS_{csb-y_j}} - \frac{(SS_{csb-xy})^2}{SS_{csb-y}} \right) / (kp-1)}{\left(SS_{csb-x} - \sum_{j=1}^{kp} \frac{(SS_{csb-xy_j})^2}{SS_{csb-y_j}} \right) / (n-2kp)}$$

Ha teljesül a feltétel, akkor a felírt változó $(kp-1, n-2kp)$ szabadságfokú F-eloszlású, tehát folytatható a vizsgálat.
A második feltétel próbafüggvénye:

$$F = \frac{(SS_{csb-xy})^2}{SS_{csb-x} SS_{csb-y} - SS_{csb-xy}} (n - kp - 1)$$

Ha a második feltétel is teljesül, akkor a felírt változó $(1, n - kp - 1)$ szabadságfokú F-eloszlású. A feltételek teljesülése után elvégezhetjük a kovarianciaanalízist. Az alacsony mérési szintű faktornak a hatását a függő változóra a következő próbafüggvény teszteli:

$$F = \frac{\left(SS_x - \frac{(SS_{xy})^2}{SS_y} - SS_{csb-x} + \frac{(SS_{xy})^2}{SS_y}\right)/(kp-1)}{(SS_{csb-x} - \frac{(SS_{cs-xy})^2}{SS_{csb-y}})/(n-kp-1)}$$

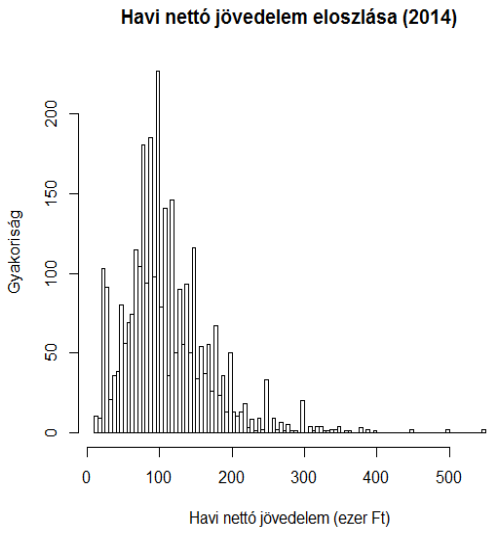
Ha a nullhipotézist elfogadjuk, akkor ez a változó egy $(kp - 1, n - kp - 1)$ szabadságfokú F-eloszlású.

5. Adatok és elemzés

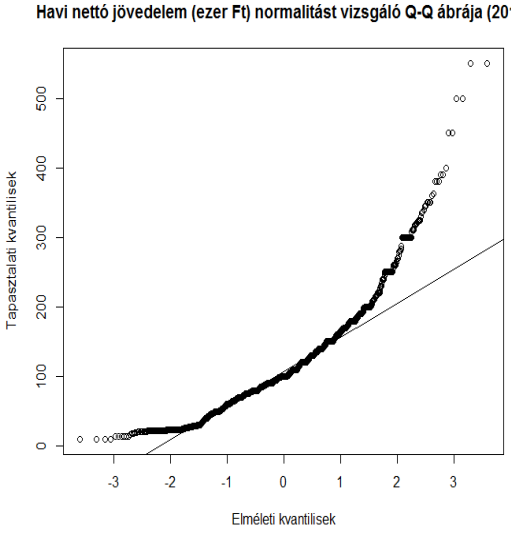
5.1. Az adatok ismertetése és előkészítése

A szakdolgozatomban a TÁRKI Háztartás Monitor felmérés adatait dolgozom fel. A Háztartás Monitor longitudinális keresztmetszeti háztartásvizsgálat. A háztartásvizsgálat sorozat 1998-ban kezdődött el, miután a Magyar Háztartás Panel (1992-1997) véget ért. Módszere: kérdőíves adatfelvétel. A személyes adatfelvétel során mintegy kétezer háztartásról és tagjairól gyűjtöttek adatokat úgy, hogy a háztartás minden 16 éven felüli tagját megkérdezték. Emellett a háztartás egészére jellemző adatokat is felvettek.
A vizsgálat központjában a munkaerőpiac és a jövedelmek témái állnak. A TÁRKI Háztartás Monitor alkalmas a teljes népességre vonatkozó következtetések levonására. Az eredmények értelmezését valamelyest nehezíti, hogy a leggazdagabbak és a legalacsonyabb jövedelműek válaszadási hajlandósága alacsony az ilyen típusú jövedelemvizsgálatok során [21].
A Háztartási Monitor 2001-es, 2003-as, 2005-ös, 2010-es, 2012-es és 2014-es adatait vizsgáltam meg szórás-és kovarianciaanalízis módszerével a szabad forráskódú R-program felhasználásával. A továbbiakban a 2014-es adatok eredményét mutatom be részletesebben. A 2014-es adathalmazban 4420 megfigyelés található, ezek közül azokat a rekordokat vonom be az elemzésbe, akik 18 éven felüliek, tehát a felnőtt lakosságot, illetve azokat, akik rendelkeznek jövedelemmel és a kérdőíves adatfelvétel során nyilatkoztak erről. Így a tényleges elemzést 3034 mintaelemre végeztem el.
Az ANOVA és az ANCOVA módszerekkel fogom megvizsgálni, hogy a jövedelemre hatással van-e a nem, az iskolai végzettség és az életkor. A Változók típusai alfejezetben említettem, hogy az ANOVA módszer alkalmazása során jelentős szerepet játszik, hogy a függő változó csak folytonos változó (magas mérési szintű) lehet, míg a faktorok csak alacsony mérési szintűek lehetnek, bár magas mérési szintű változó is lehet faktor diszkretizálás után. A kovarianciaanalízis estében a faktorok mellé még további

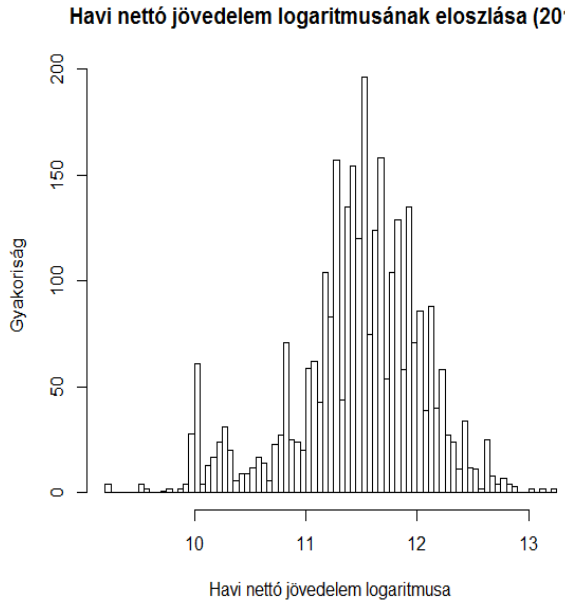
kovariánsok(folytonos változók) is bevonhatóak. Tehát a függő változóm a vizsgálat során végig a jövedelem, pontosabban a havi nettó jövedelem lesz, az adathalmazban ezt az attribútumot hgjobe0 kód jelöli. A faktorok kezdetben a nem(hgbnem0) lesz, majd az iskolai végzettség(hgiisk0) és végül az életkor(hgbszu0). Az iskolai végzettség egyes szintjeit blokkokba vontam össze. Az első szintbe tartozik a "Kevesebb, mint 8 általános", "8 általános" és a "Szakmunkásképző; szakképzés érettségi nélkül" megnevezésű iskolai végzettségek. A második szintbe tartozik a "Szakközépiskolai érettségi; szakképzést követő érettségi", "Gimnáziumi érettségi", "Érettségit követő, felsőfokra nem akkreditált szakképzés; technikum" és a "Akkreditált felsőfokú szakképzés; felsőfokú technikum". A harmadik szint pedig magába foglalja a "Főiskola", "Egyetem" és "Tudományos fokozat" elnevezésű iskolai végzettségeket. Látható, hogy az első szintbe kerültek az érettségivel nem rendelkező személyek, a második szintbe az érettségivel, illetve a legtöbb esetben szakképzéssel is rendelkezők, míg a harmadik szintbe a diplomások kerültek. Az életkor attribútum (2014-születési év) magas mérési szintű, ezért az ANOVA számításhoz diszkretizáltam ezt a változót. Négy részre osztottam, az első kategória a 18 és 35 év közöttiek, ahol az intervallum baloldali végpontja beletartozik a csoportba, de a jobboldali végpontja a következő csoporthoz fog tartozni. A második kategória a 35 és 50 év közöttiek, a harmadik szintbe az 50 és 65 év közöttiek tartoznak és végül a legutolsó szinten a 65 év felettiek vannak.
A szóráselemzés alkalmazhatóságának feltétele az, hogy a függő változó normális eloszlású legyen. A következő hisztogramon a havi nettó jövedelem eloszlása látható, amely egyáltalán nem hasonlít a normális eloszlásra, ezt mutatja a hisztogram után látható Q-Q ábra is. A hisztogramon megfigyelhető, hogy viszonylag sokan vannak, akik keveset keresnek(100 ezer forint alattiak). A jövedelem mediánja 100 ezer forint és ezután hosszan elnyúlik az eloszlás. A gyakorlatban többször tapasztalható, hogy a jövedelem lognormális eloszlású és valóban ez a hisztogram is hasonlít a lognormális eloszláshoz, ezért az adatokat transzformáltam, vagyis az e alapú logaritmusát vettem.



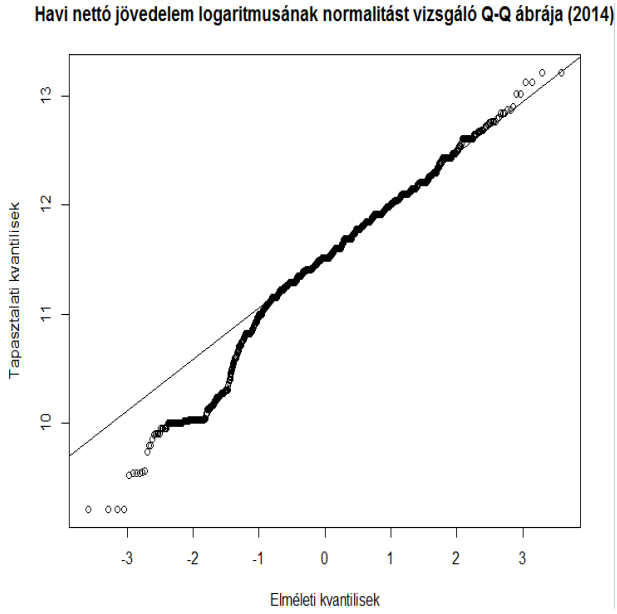
A következő Q-Q ábrán látható, hogy a normalitási feltevés nem igazolódik be, mivel a pontok jelentős mértékben eltérnek az ábrán látható egyenestől.



A Kolmogorov-Szmirnov teszt sem fogadja el a normalitást. A D értéke 0,1071, a D a tapasztalati és az elméleti eloszlásfüggvény abszolút eltérésének a maximuma. A p-érték kisebb, mint 2,2e-16.
Az alábbi hisztogram a havi nettó jövedelem logaritmusát ábrázolja, amely már jobban hasonlít a normális eloszlásra, mint az előbbi hisztogram, habár nem teljes mértékben követi, de ez talán nem is várható el ilyen nagyságú való életből vett mintánál, ahol torzítások és véletlen hibák is befolyásolják az adatok felvételét.

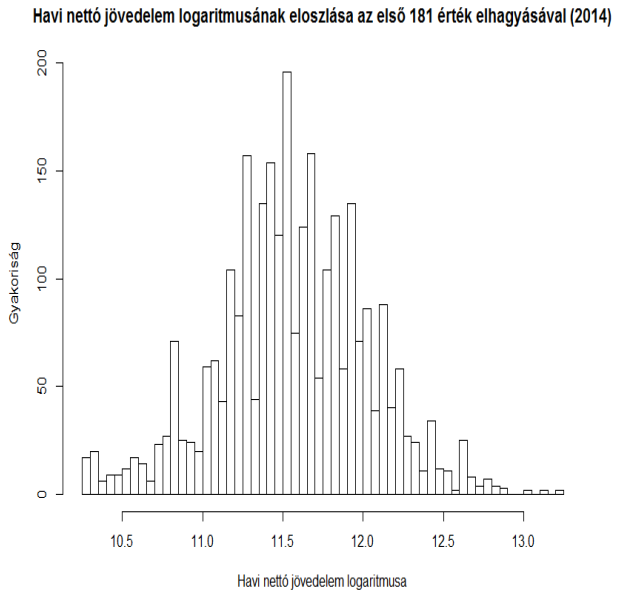


Az adatok transzformálása után a Q-Q ábra is megváltozott. Megfigyelhető, hogy a pontok jobban követik az egyenest, de az ábra alsó részénél eltér az egyenestől. Az ábra legalján van pár kiugró érték, majd egy kisebb csoport, akiknek nagyon kevés a jövedelmük. Ez alatt lehetnek a különböző segélyek, GYES, családi pótlék vagy nagyon alacsony összegű nyugdíj. Ez a kisebb csoport nem tekinthető véletlennek. Ezt a részt elhagyva meredeken közelítenek a pontok a kívánt egyeneshez.



Az adat transzformálás után a Kolmogorov-Szmirnov teszt értéke 0,085 és a p-értéke kisebb, mint 2,2e-16. Az adatok első 181 elemét elhagyva a következő hisztogramot láthatjuk. A 181 érték elhagyása után a Kolmogorov-

Szmirnov teszt D értéke 0,0343 és a p-értéke 0,002382. Tehát egyre jobban közelítjük a normális eloszlást.



5.2. Egyszempontos szórásanalízis alkalmazása

A 7. táblázatban az egyszempontos szóráselemzés eredményei láthatóak. A nullhipotézisünk az, hogy a havi nettó jövedelemre nincs hatással a 18 év feletti lakosok neme. Vagyis a férfiak és nők jövedelmének átlaga között nincs szignifikáns különbség. Az ANOVA-tábla F próbája szerint a nullhipotézist el kell vetni (szignifikancia 3,27e-15, vagyis a p<0,05). A férfiak havi nettó jövedelmének átlaga 122541, a nőké 103450. A p-érték szerint a férfiak és a nők jövedelmének átlaga közötti eltérést nehezen lehet véletlen ingadozással magyarázni. Az R^2 együttható 0,0202, csupán 2%-ot magyaráz meg személyek neme a jövedelmek szórásnégyzetéből.

Forrás	SS	df	MS	F	p-érték
Hatás(csk)	20,9	1	20,910	62,75	3,27e-15
Hiba(csb)	1010,4	3032	0,333		

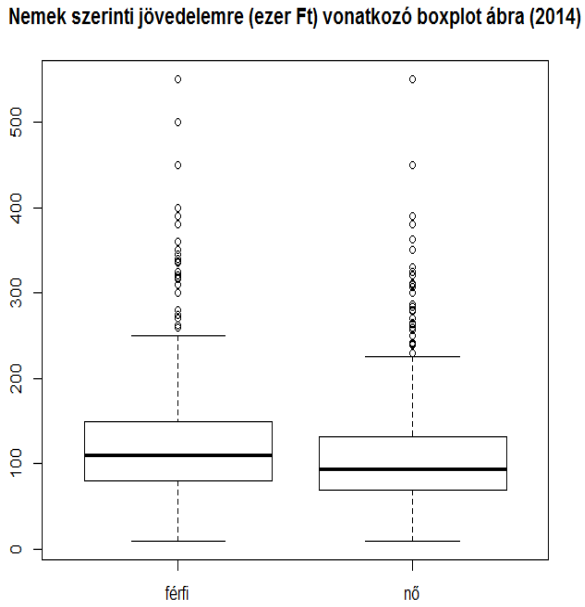
7. táblázat.

A 8. táblázatban az előbbi egyszempontos varianciaanalízis homoszkedasztikusság vizsgálatának eredménye látható, ahol a Levene-tesztet használtam. Látható, hogy a Levene-teszt statisztika értéke alacsony (0,014), a p-érték magas (0,906) a nullhipotézist megtarthatjuk, tehát a csoportokon belüli szórás megegyezik.

Forrás	SS	df	MS	F	p-érték
Hatás(csk)	0,0	1	0,00205	0,014	0,906
Hiba(csb)	444,1	3032	0,14648		

8. táblázat.

Az alábbi boxplot ábrán a férfiak és a nők havi nettó jövedelmeiről látható pár statisztikai mutató. Az egyes dobozok az alsó kvartilistől a felső kvartilisig tartanak. A dobozok középvonala a csoport mediánját jelöli. A férfiak havi nettó jövedelmének mediánja 110 ezer forint, a nők havi nettó jövedelmének mediánja 94 ezer forint. A vonalak a teljes terjedelmet felölelik, ha ez mindkét irányban nem nagyobb a kvartilisek közötti különbség 1,5-szeresénél [12]. Az ezen kívül eső pontokat (ún. outliereket) is megjeleníti az ábra. A boxplot ábra egy grafikus megjelenítést ad az adathalmaz jellegéről.



5.3. Kétszempontos szórásanalízis és szimulációs vizsgálatok

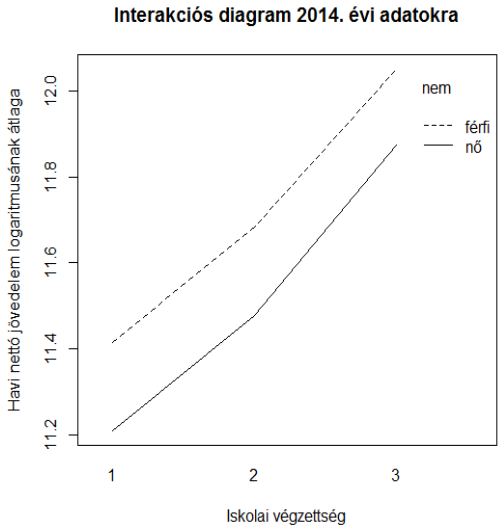
A 9. táblázatban egy kétszempontos szórásanalízis eredményei láthatóak. A továbbiakban a nemre vonatkozó hatást a-val jelölöm, míg az iskolai végzettségre vonatkozó hatást b-vel. Tehát az egyszempontos esethez képest itt az iskolai végzettséget is bevontam újabb faktorként. A táblázatban megjelenik még egy plusz sor, ami a nem és az iskolai végzettség közötti interakciót jellemzi. Mivel a nem p-értéke (kisebb, mint 2e-16) és az iskolai végzettség p-értéke (kisebb, mint 2e-16) is szignifikáns, mert mind-

kettő érték kisebb, mint 0,05, ezért el kell utasítani azt a nullhipotézist, miszerint a nem és az iskolai végzettség nincs hatással a jövedelemre. Az ANOVA-táblát megvizsgálva megállapítható, hogy szignifikáns interakciós hatás nem figyelhető meg a két faktor között (F=0,122, p-érték=0,885). Az egyes csoportok átlagai a következőképpen változnak. Az iskolai végzettség első szintjén a férfiak havi nettó jövedelmének átlaga 103376, szórása 51760, míg a nőknél az átlag 83040, a szórás pedig 38655. Az iskolai végzettség második szintjén a férfiak havi nettó jövedelmének átlaga 135278, szórása 65732 , míg a nők esetében az átlag 110807, a szórás 54541. Végül az iskolai végzettség harmadik szintjén a férfiak havi nettó jövedelmének átlaga 187918, szórása 81633, ezzel szemben a nőknél az átlag 157570, a szórás pedig 65534. A kétszempontos szórásanalízis esetében az R^2 együttható 0,16, vagyis 16 % a megmagyarázott szórásnégyzet-hányad.

Forrás	SS	df	MS	F	p-érték
a-hatás(csk)	29,8	1	29,80	105,005	<2e-16
b-hatás(csk)	142,0	2	71,00	250,153	<2e-16
ab-interakció	0,1	2	0,03	0,122	0,885
Hiba(csb)	859,4	3028	0,28		

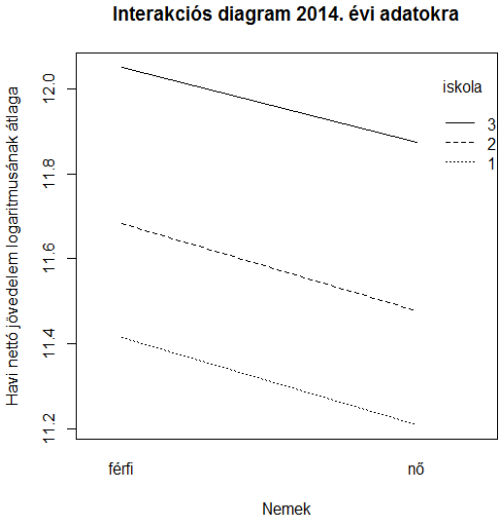
9. táblázat.

Az interakciót vagy annak hiányát grafikusan is lehet ábrázolni. A következő két ábra ezt mutatja. Az interakciós ábrának az x-tengelyén az egyik faktor szintjei láthatóak és az ábra a másik faktor viselkedését mutatja az első faktor függvényében.



Tehát a fenti interakciós ábrán a szaggatott vonal mutatja a férfiak jövedelem függését az iskolai végzettségtől, míg a folytonos vonal a nőkét írja

le. Mivel ez a két vonal nagyjából párhuzamos, ezért ebben az esetben nem beszélhetünk interakcióról. Az alábbi interakciós ábra ugyanazt írja le, de ebben az esetben az x-tengelyen a másik faktor szerepel. Ezek az egyenesek is nagyjából párhuzamosak.

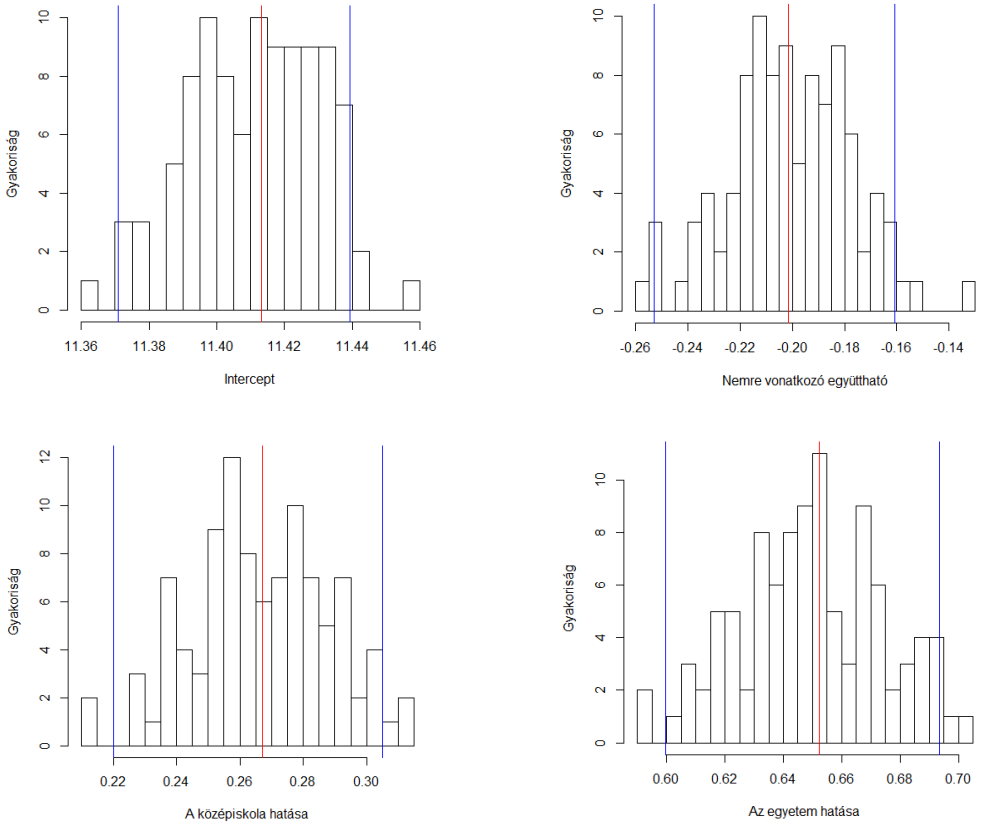


A 10. táblázatban a kétszempontos varianciaanalízis homogenitás vizsgálata látható, hasonlóan az egyszempontos esethez itt is a Levene-teszt alkalmazásával. A nullhipotézis az, hogy a csoportok szórásai egyenlőek, de mint azt a táblázat is mutatja az iskolai végzettség faktornak a p-értéke (0,00137) kisebb mint 0,05, a másik faktornál a p-érték megfelelő.

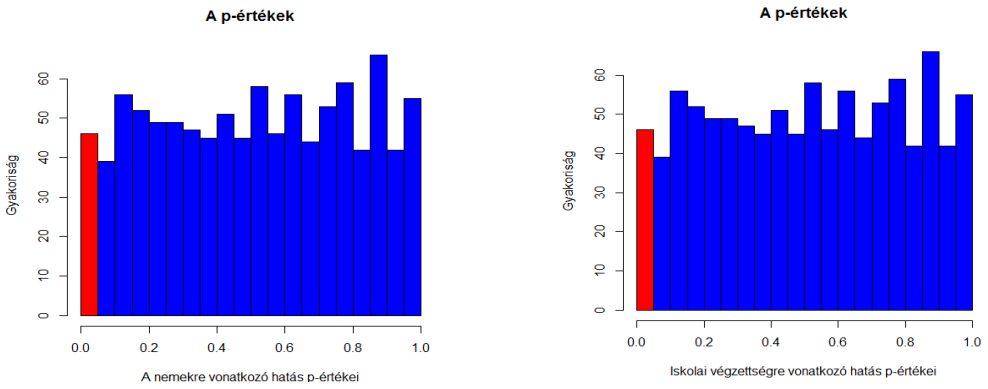
Forrás	SS	df	MS	F	p-érték
a-hatás(csk)	0,0	1	0,0059	0,045	0,83136
b-hatás(csk)	1,7	2	0,8556	6,611	0,00137
ab-interakció	0,2	2	0,1183	0,914	0,40103
Hiba(csb)	391,9	3028	0,1294		

10. táblázat.

Mivel a Levene-teszt szignifikáns, így szimulációval is vizsgálom a hatások szignifikanciáját. Az első négy hisztogramon a szimulált ANOVA együtthatói láthatóak. A hisztogramok bootstrap eljárással készültek. Minden egyes csoportból vettem mintát. Az első két hisztogramon a konstans és a nemre vonatkozó (nő) együtthatók eloszlása látható, a piros vonal az eredeti kétszempontos szórásanalízis együtthatójának az értéke, a kék vonalak a 2,5%-97,5%-os kvantiliseket jelölik. Végül az iskolára vonatkozó együtthatók eloszlása is látható (2-es és 3-as szinten lévő iskolai végzettségűek). Látszik az ábrákon a hatások szignifikanciája (semelyik esetben sem kerültek a 0 közelébe) és a konfidencia intervallum becslése is.

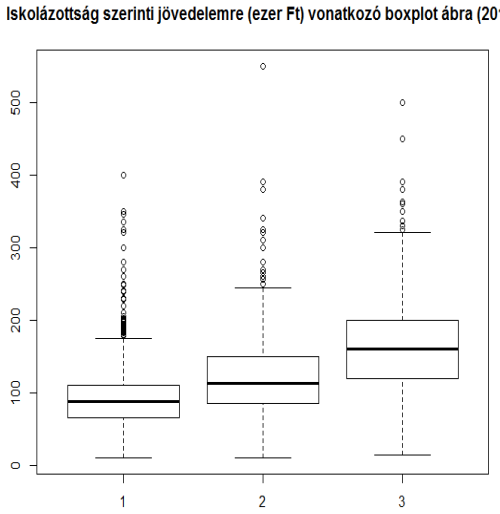


Végeztem egy másik szimulációt is, amikor a hatásokat 0-nak vettem. Itt arra voltam kíváncsi, hogy vajon a heteroszkedasztikusság nem befolyásolja-e túlságosan az alkalmazott teszt p-értékét. A nemre és az iskolai végzettségre vonatkozó szimulált p-értékek láthatóak. A pirossal jelölt rész a szignifikáns rész. Ennél a szimulációnál a korábban ismertetett lineáris modell egyenletét használtam, a véletlen hibát pedig normális eloszlásból generáltam úgy, hogy a szórások megegyezzenek a megfelelő csoport szórásával. Megállapítható, hogy az 1000 szimulációból közel 50 esetben találtunk szignifikánsnak tűnő hatást, és a hisztogram is mutatja, hogy a statisztika p-értékei közel vannak az egyenletes eloszláshoz. Tehát nem okozott számottevő eltérést a csoportonkénti szórások feltételezhető eltérése.



A hatás nélküli szimuláció p-értékei

Az alábbi boxplot ábra az iskolai végzettség szerint jövedelemre ad pár statisztikai mutatót. Az iskolai végzettség első szintje, vagyis az érettségivel nem rendelkezők mediánja 88 ezer forint.Ez a legalacsonyabb a többi szinthez képest. A második szint mediánja 112500 forint, a harmadik szint mediánja pedig 160 ezer forint. A mediánok mutatják, hogy a magasabb iskolai végzettséggel rendelkezők általában többet keresnek. Érdekessége az ábrának, hogy a felmérésben résztvevő személyek között a legmagasabb havi nettó jövedelem az 550 ezer forint volt és ennek a személynek az iskolai végzettsége második szinten van, vagyis az adathalmazban a legmagasabb havi nettó jövedelmű személynek nincs diplomája. Az ábrán az is megfigyelhető, hogy az érettségivel nem rendelkezők csoportjában sok a kiugró érték. Ennek egyik lehetséges oka lehet az is, hogy a szakmunkásképzőt végzett személyeknek jobbak a munkaerő-piaci kilátásai, mint a csak 8 általánost végzetteknek. A szakmunkásképzőt végzett embereket nem tettem külön csoportba, mivel kis számban jelentek meg az adathalmazban.



5.4. Háromszempontos szórásanalízis alkalmazása

A 11. táblázatban egy háromszempontos ANOVA-tábla látható, a korábban használt két faktorhoz bevontam egy harmadik faktort az életkort. Az életkornak a diszkretizált változatát használom ebben az elemzésben. Az életkort az ANOVA-táblában c-vel fogom jelöni, mint hatást. A háromszempontos szóráselemzésnél megjelennek a táblázatban az összes kétszeres interakcióra vonatkozó adatok, illetve a hármas interakciót is vizsgálja a módszer. A nullhipotézisünk azt mondja ki, hogy a havi nettó jövedelemre nincs hatással a nem, az iskolai végzettség és az életkor. A táblázat első három sora a három faktor külön-külön hatását nézi, mindegyiknek a p-értéke kisebb, mint 2e-16, amiből következik, hogy a nullhipotézist nem tudjuk elfogadni. Az interakciókból megállapítható, hogy az iskolai végzettség és nem hatása között nem tudunk interakcióról beszélni (p-érték=0,491) és ugyanez igaz a hármas esetben is (p-érték=0,618). Viszont az iskolai végzettség és az életkor hatása között szignifikáns az interakció (p-érték=2,38e-10) és a nem és az életkor hatása között is szignifikáns (p-érték=3,72e-05). Az R^2 együttható értéke 0,23, tehát 23 %-ot magyaráznak a faktorok a jövedelmek szórásnégyzetéből.

Forrás	SS	df	MS	F	p-érték
a-hatás(csk)	29,8	1	29,80	113,654	<2e-16
b-hatás(csk)	142,0	2	71,00	270,758	<2e-16
c-hatás(csk)	47,6	3	15,88	60,547	<2e-16
ab-interakció	0,4	2	0,19	0,712	0,491
ac-interakció	6,1	3	2,03	7,753	3,72e-05
bc-interakció	14,9	6	2,49	9,487	2,38e-10
abc-interakció	1,2	6	0,19	0,739	0,618
Hiba(csb)	789,3	3010	0,26		

11. táblázat.

5.5. Kovarianciaanalízis bemutatása a vizsgált adatokon

A 12. táblázatban annak a kovarianciaanalízisnek az eredménye látható, ahol egy faktor van a nem, és egy kovariáns az életkor. A kovarianciaanalízisnél az életkort folytonos változóként használom és nem a korábban említett diszkretizált változatát. Most az életkort k-val fogom jelölni a táblázatban, mint kovariánst. Ez a táblázatból kiderül, mivel az életkor szabadságfoka(df) 1, diszkretizált esetben 3 lenne. Megállapítható a táblázatból, hogy a nem p-értéke (2,42e-15) és a kovariáns p-értéke (4,66e-08)

is szignifikáns, vagyis elvetendő az a nullhipotézis miszerint a nem és az életkor nincs hatással a jövedelemre. Az R^2 együttható értéke 0,029, vagyis 2,9 %-ot magyaráz meg a faktor és a kovariáns a jövedelmek szórásnégyzetéből.

Forrás	SS	df	MS	F	p-érték
a-hatás(csk)	20,9	1	20,910	63,35	2,42e-15
k-hatás(csk)	9,9	1	9,904	30,01	4,66e-08
Hiba(csb)	1000,5	3031	0,330		

12. táblázat.

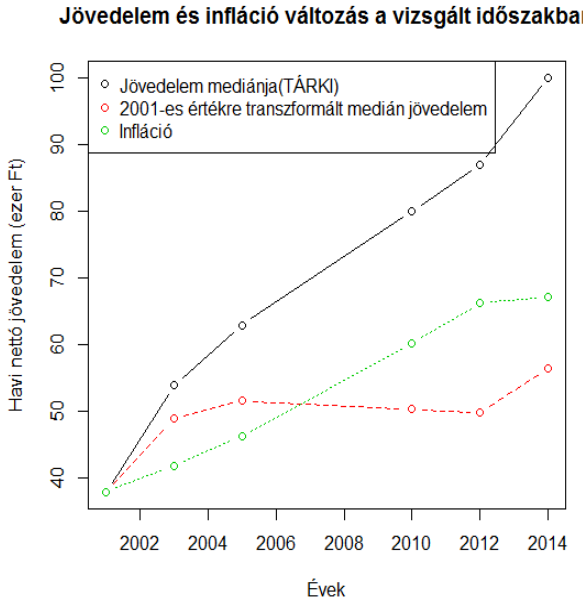
A 13. táblázatban szintén egy kovarianciaanalízis eredményei láthatóak. Itt két faktor van az iskolai végzettség és nem, illetve továbbra is egy kovariáns az életkor. A két faktor és a kovariáns esetében is a p-érték kisebb, mint 2e-16, vagyis mindegyik szignifikáns. A két faktor interakcióját vizsgáló p-érték (0,396) azt bizonyítja, hogy a két faktor között nem szignifikáns a kölcsönhatás. Az R^2 együttható értéke 0,198, vagyis 19,8 %-os a megmagyarázott szórásnégyzet-hányad. Látható, hogy a háromszempontos szórásanalízis esetében az R^2 együttható nagyobb volt (23%). Ez betudható annak is, hogy a kovarianciaanalízis esetében az életkor (vagyis a kovariáns) és a jövedelem kapcsolata nem lineáris. Kezdetben nő a jövedelem, majd egy idő után stagnál és nyugdíjas években lecsökken.

Forrás	SS	df	MS	F	p-érték
a-hatás(csk)	29,8	1	29,80	109,170	<2e-16
b-hatás(csk)	142,0	2	71,00	260,076	<2e-16
k-hatás	32,6	1	32,63	119,522	<2e-16
ab-interakció	0,5	2	0,25	0,927	0,396
Hiba(csb)	826,4	3027	0,27		

13. táblázat.

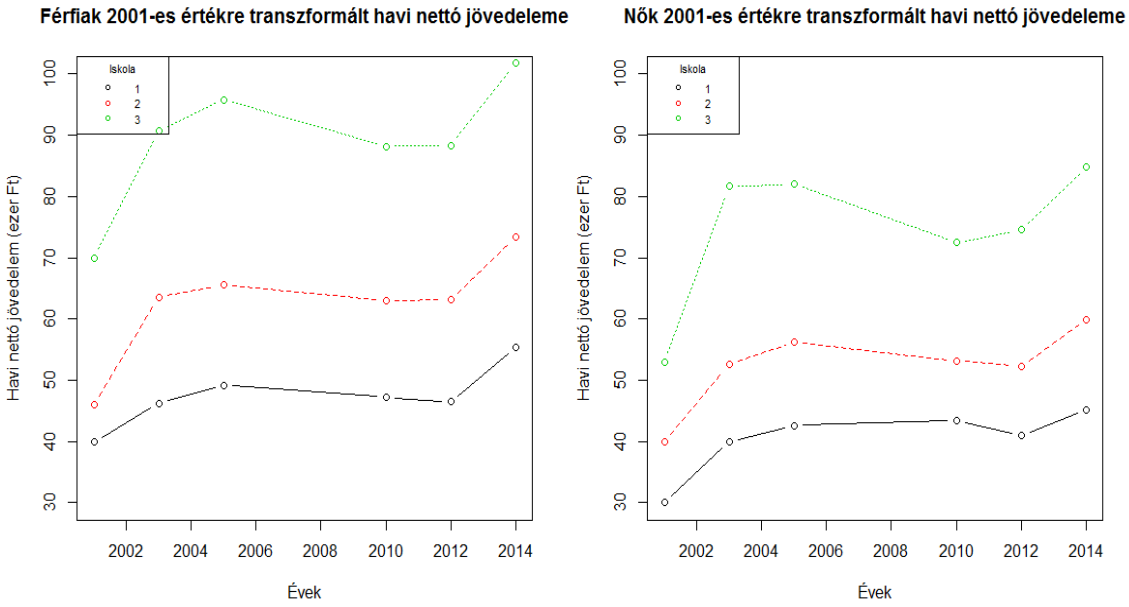
5.6. Eredmények összesítése

A következő grafikonon a jövedelem változása látható. A TÁRKI-s adathalmazokból kiszámolt havi nettó jövedelem mediánját az ábrán feketével jelöltem. Ezeket a jövedelem mediánokat visszatranszformáltam 2001-es értékekre, vagyis leosztottam az aktuális éves inflációk szorzatával (fogyasztóiárindex [22]), ezt piros színnel tüntettem fel az alábbi grafikonon. A 2001-es medián jövedelem aktuális értékét pedig zölddel jelöltem. Az ábrán látható, hogy 2005-től 2012-ig az inflációs görbe meredeken emelkedik, míg a 2001-es értékre visszatranszformált jövedelem elkezdett csökkenni, tehát ebben az időszakban a jövedelmeknek a vásárló értéke csökkent. 2012-től az inflációs görbe nagyon picit emelkedik csak, szinte stagnál és láthatóan a jövedelmek vásárló értéke elkezdett nőni.

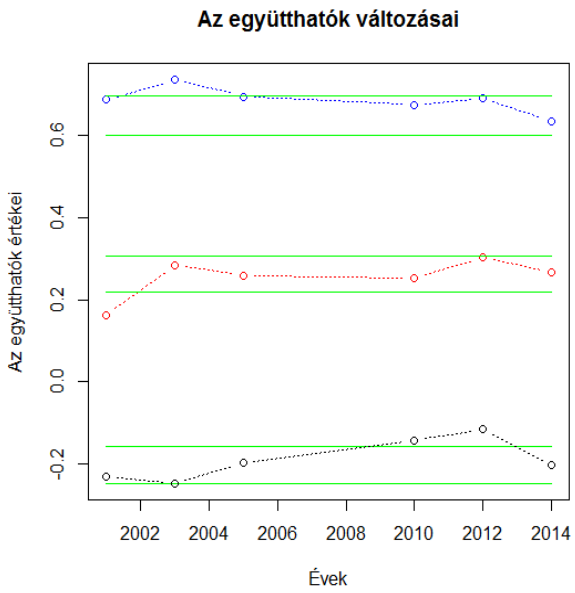


Az alábbi két ábrán a férfiak és a nők havi nettó jövedelmének változása látható, amelyeknél az inflációt figyelembe vettem. A zölddel jelölt a legmagasabb iskolai végzettség csoportba tartozókat ábrázolja, tehát a diplomásokat. A piros az érettségivel (esetleg még szakmával is) rendelkezők jövedelem változását mutatja. Végül a feketével jelöltem az érettségivel nem rendelkezők csoportjának jövedelem változásait. Az ábrán megfigyelhető elsősorban a csoportok jövedelmének 2001-es értékekre visszatranszformált mediánja alapján, hogy a magasabb iskolai végzettségűek általában többet keresnek, illetve hogy a férfiaknak is általában többet keresnek, mint a nők. Természetesen vannak kivételek, de ha a társadalomból vett minták mediánjait nézzük ezek az eredmények jönnek ki, összhangban az előző fejezetben bemutatott ANOVA elemzéssel, ahol szintén pozitív hatást jelentett az iskolai végzettség magasabb szintje. Ahogy az előző ábrán is, itt is látható, hogy 2005-től az egyes csoportok jövedelmének vásárló értéke

csökken, a legnagyobb mértékben a diplomásoké. Érdekes, hogy a nőknél az érettségivel nem rendelkező csoportról ez nem mondható el. Viszont 2012-től minden csoportnál megfigyelhető a jövedelmek vásárló értékének a növekedése.



A vizsgált időszakban a kétszempontos szórásanalízis együtthatóinak a változását a következő ábra mutatja. A konstans (intercept) nincs ábrázolva. A kékkel jelölt a diplomásokat ábrázolja, a pirossal jelölt az érettségivel rendelkezőket, a feketével jelölt pedig a nemre vonatkozó hatás (nők). A zöld egyenesek az előzőekben bemutatott 2014-re vonatkozó 95%-os bootstrap konfidenciaintervallumokat jelölik. Ezeket az előző fejezetben a szimulációknál határoztam meg. Az ábrán látható, hogy az évek múlásával nagyjából stabilak ezek az együtthatók.



A 14. táblázatban minden egyes csoportnak a becsült átlagfizetése látható, amely a kétszempontos szórásanalízis együtthatóinak a visszatranszformálásából keletkezett.

	2014
8 általánost végzett nők	73870 Ft
Érettségivel rendelkező nők	96517 Ft
Diplomás nők	143463 Ft
8 általánost végzett férfiak	90685 Ft
Érettségivel rendelkező férfiak	118554 Ft
Diplomás férfiak	171171 Ft

14. táblázat.

6. Összegzés

Szakdolgozatomban a szórás- és kovarianciaanalízis statisztikai eljárásokat mutattam be. Először a történeti, majd a matematikai, elméleti hátterét ismertettem. Az elméleti rész bemutatása után az R 3.1.2-es verziójával vizsgáltam a TÁRKI-tól kapott adatokat a fentebb ismertetett statisztikai módszerekkel. Látható volt, hogy a gyakorlatban több probléma is felmerült a módszerek alkalmazása során. A szórásanalízis feltételei közül a normális eloszlás és a csoportok közötti azonosság nem minden esetben teljesült. A normális eloszlás közelítése érdekében az adatokat transzformáltam, e alapú logaritmusát vettem. A homogenitás az egyszempontos szórásanalízis esetében teljesült, de a kétszempontos esetben nem. A homogenitás hiányát szimulációkkal vizsgáltam és elemeztem ezután. Megfigyelhető volt, hogy a havi nettó jövedelemre, a vizsgált változóra hatással van a nem, az iskolai végzettség és az életkor is. Pontosabban a kapott eredmények alapján a különbségeket nehezen lehetne csupán a véletlen ingadozással magyarázni. Legvégül a kapott hat év havi nettó jövedelem mediánjainak változását az infláció függvényében ábrázoltam a nem és az iskolai végzettség csoportokra bontása alapján. Az utolsó ábrán az együtthatók időbeni változását vizsgáltam meg. Látható volt, hogy a hatások nagyjából állandóak.

7. Irodalomjegyzék

Hivatkozások

[1] https://hu.wikipedia.org/wiki/Varianciaanal%C3%ADzis

[2] https://en.wikipedia.org/wiki/Analysis_of_variance#History

[3] Székely J. Gábor: *Paradoxonok a véletlenek matematikájában*, Typotex, Budapest, 2004, p134

[4] http://xenia.sote.hu/hu/biosci/docs/biometr/lecture/anova1.html

[5] http://clinfowiki.org/wiki/index.php/ANOVA#History

[6] Babbie, E.: A társadalomtudományi kutatás gyakorlata, Balassi Kiadó Budapest, 1995. p430-435

[7] http://psycho.unideb.hu/munkatarsak/balazs_katalin/matalapok/matalapok_ora2.pdf

[8] Németh Renáta, Simon Dávid: Társadalomstatisztika

http://www.tankonyvtar.hu/hu/tartalom/tamop425/0010_2A_21_Nemeth_Renata-Simon_David_Tarsadalomstatisztika_magyar_es_angol_nyelven/ch02s04.html

[9] http://www.cs.elte.hu/~vargal4/Elm_vsz1_14.pdf

[10] https://hu.wikipedia.org/wiki/Folytonos_val%C3%B3sz%C3%ADn%C5%B1s%C3%A9gi_v%C3%A1ltoz%C3%B3

[11] http://www.agr.unideb.hu/~baloghp/PhD%20anyagok/parameteres_elmelet.pdf

[12] Dr. Zempléni András, *Leíró és matematikai statisztika előadásjegyzet*

http://www.cs.elte.hu/~zempleni/

[13] Dr. Márkus László, *Idősorok és többdimenziós statisztika előadásjegyzet*

http://www.math.elte.hu/probability/markus/index.m.html

[14] http://www.tankonyvtar.hu/hu/tartalom/tamop425/0027_MSTE5/ch01s06.html

[15] Huzsvai László: *Variancia-analízisek az R-ben*, Seneca Books, Debrecen, 2013

[16] Bolla Marianna, Krámli András: *Statisztikai következtetések elmélete*, Typotex, Budapest, 2005,p15-61, p269-291

[17] `http://www.tankonyvtar.hu/hu/tartalom/tamop425/2011_0001_531_pedagogia/ch15s04.html`

[18] `http://www2.univet.hu/users/zslang/phd/ANOVA%20es%20elrendezesek.pdf`

[19] Fazekas István: Statisztika

`http://www.inf.unideb.hu/valseg/dolgozok/fazekasi/oktatas/statmobi.pdf`

[20] Hunyadi László: *Grafikus ábrázolás a statisztikában*, Statisztikai Szemle, 2002 január, p49

[21] `www.tarki.hu`

[22] `https://www.ksh.hu/docs/hun/xstadat/xstadat_eves/i_qsf001.html`