JANUS PANNONIUS
TUDOMÁNYEGYETEM
PÉCS

Herman – Pintér –
– Rappai – Rédey

STATISZTIKA II.

JANUS PANNONIUS TUDOMÁNYEGYETEM
KÖZGAZDASÁGITUDOMÁNYI KAR
Statisztika és Demográfia Tanszék

Herman – Pintér – Rappai – Rédey

STATISZTIKA II.

III. kiadás

Pécs, 1999

Írta:

Dr. Herman Sándor (8.1 - 8.5 alfejezetek)

Dr. Pintér József (7. fejezet)

Dr Rappai Gábor (6. fejezet)

Dr. Rédey Katalin (8.6 alfejezet)

Lektorálta:

Dr. Hunyadi László

Janus Pannonius Tudományegyetem, Közgazdaságtudományi Kar
Pécs 1999

Felelős kiadó: Dr. Farkas Ferenc dékán
Készült 1000 példányban

ISBN 963 641 356 8

Készült: Carbocomp Kft., Nyomda – Pécs
Felelős vezető: Pető Attila

6. SZTOCHASZTIKUS KAPCSOLATOK ELEMZÉSE

A gazdasági jelenségek beható elemzése gyakran igényli annak megállapítását, vajon két vagy több - az adott esemény szempontjából lényeges - statisztikai ismérv kapcsolatban áll-e egymással. A változók közötti kölcsönhatások típusait tanulmányaink során érintőlegesen már megismertük (lásd! TK. I. 1. fejezet). A jellemzők közötti kapcsolatokat számos szempont alapján csoportosíthatjuk; a tananyagunk szempontjából legfontosabb felosztás szerint az ismérvek lehetnek egymástól **függetlenek**; köztük a kapcsolat lehet **sztochasztikus**, illetve **determinisztikus**. A statisztika-tudomány a fentiek közül alapvetően a sztochasztikus kapcsolatok vizsgálatával foglalkozik, ám ez nem jelenti azt, hogy az ismérvek közötti függetlenség, vagy a determinisztikus kapcsolat - mint a sztochasztikus kapcsolat szélsőséges esete - nem képezheti vizsgálódásunk tárgyát[1].

A statisztikai ismérvek között tendenciaszerűen, valószínűségi jelleggel érvényesülő összefüggést sztochasztikus kapcsolatnak nevezzük. Ez a gyakorlatban annyit jelent, hogy egy egyednek az egyik ismérv egy adott ismérvváltozatához való tartozásából következtethetünk arra, hogy ez az egyed egy másik ismérv melyik változatához tartozik. Az esetek egy bizonyos százalékában azonban következtetésünk hibás lesz. Minél szorosabb az összefüggés a változók között, azaz minél közelebb áll a kapcsolat a függvényszerűhöz, annál kisebb a valószínűsége a tévedésnek. A fentiek alapján kézenfekvő, hogy a statisztika-tudomány megkísérli valamilyen eszközzel az ismérvek közötti kapcsolatok

[1]A változók közti függvényszerű (determinisztikus) kapcsolat elemzése elsősorban a matematika-tudomány tárgykörébe tartozik, így - ismerve e két módszertudomány szoros kapcsolatát - nem lepődhetünk meg azon, hogy számos a statisztikában használatos vizsgálati eszköz erősen matematikai indíttatású.

szorosságát (illetve egyáltalán a kapcsolat meglétét) számszerűsíteni, ezáltal a következtetés hibáját mérsékelni, de legalábbis meghatározni.

Könnyen belátható, hogy a jellemzők között meglévő összefüggések szorosságának számszerűsítése mellett, nagyon lényeges a köztük levő oksági viszonyok feltárása is. Két változó között alapvetően három fajta oksági kapcsolat képzelhető el:

- együttmozgó változók, ahol egyik változó sem oka a másiknak (ebben az esetben, ha módszereink az ismérvek között kapcsolatot mutatnak ki, úgy ez ún. látszat-kapcsolat, aminek idejekorán való feltárása számos későbbi, következtetésbeli hibától kímélhet meg);
- egyértelmű ok-okozati viszony;
- a változók kölcsönhatása esetén, köztük "visszacsatolási" (ismert angol szóval **feedback**) mechanizmus érvényesül .

A látszatkapcsolatok leggyakoribb oka valamilyen más, a vizsgált ismérveket egyaránt erősen befolyásoló, az elemzés során figyelembe nem vett változó megléte. A téves következtetések elkerülése érdekében ezért mindenképpen ajánlható, hogy a statisztikai módszerekkel meghatározott eredményeinket felhasználás előtt feltétlenül vessük alá közgazdasági kontrollnak.

Amennyiben az ismérvek között oksági összefüggés van, annak megállapítása, hogy melyik változó tekinthető tényező- (magyarázó) és melyik függő- (eredmény) változónak, általában szakmai ismeretek alapján elvégezhető. Könnyítheti helyzetünket, ha a változók időben megelőzik egymást, mivel ekkor az ok és az okozat elkülönítése nem jelent nehézséget. Az oksági kapcsolat fajtájának meghatározása azonban az esetek egy részében nem egyszerű dolog. Az összefüggések meglétét, illetve szorosságát vizsgáló, tisztán matematikai, valószínűségelméleti alapokon álló eszközök ebben nem mindig nyújtanak elegendő segítséget.

Előfordulhat, hogy előzetes információink alapján nem tudjuk eldönteni, melyik változó az ok, és melyik az okozat, mert a vizsgálatunkban szereplő ismérvek között mindkét irányú hatás érvényesülhet (gondoljunk például arra a közismert tényre, hogy az emberek matematikai és zenei érzéke között kapcsolat mutatható ki; az azonban nehezen dönthető el, hogy vajon a kitűnő matematikai érzék teszi-e könnyebbé a zene harmóniájának észrevételét, vagy a zenében megtalálható tökéletes arányok keltik fel a matematika iránti érdeklődést). Ilyenkor a legcélravezetőbb megoldás, ha a kapcsolat szorosságának számszerűsítését mindkét feltételezés alapján elvégezzük.

A sztochasztikus kapcsolatok csoportosítása leggyakrabban a bennük szereplő ismérvek típusa alapján történik. Ennek megfelelően beszélhetünk **asszociációs, vegyes** és **korrelációs** kapcsolatról. **Asszociációs kapcsolatnak a minőségi ismérvek; korrelációs kapcsolatnak a mennyiségi ismérvek közötti kapcsolatot nevezzük; vegyesnek pedig azon kapcsolatokat, melyben az ok szerepét minőségi, az okozat szerepét mennyiségi ismérvek töltik be.** Az elemzési eszközök kiválasztása szempontjából azonban az előbbi osztályozás rugalmasan kezelendő. Tudjuk, hogy míg a minőségi ismérvek általában nominális-, illetve ordinális skálán mérhetők, addig a mennyiségi ismérvek legtöbbször intervallum-, vagy arányskálán jeleníthetők meg. Előfordul, hogy egy adott ismérv többfajta mértéken is megjeleníthető. Például jellemezzük eredményesség szempontjából a gazdálkodó egységeket a vagyonarányos nyereségükkel. Ha ezt a tényleges, százalékos értékekkel tesszük akkor arányskálával dolgozunk; ha azt mondjuk, hogy az adott vállalat az 1.,2.,... stb. a vizsgált területen, akkor ordinális mértéket használunk, noha egyazon ismérv alapján végeztük az elemzést. Ilyen esetekben előfordulhat, hogy alapvetően minőségi ismérvek közötti kapcsolatra kidolgozott eszközöket alkalmazunk a mennyiségi ismérvek közötti összefüggés leírására[2].

---

[2] Nem szabad elfeledkeznünk azonban arról, hogy egy magasabb rendű skáláról alacsonyabb rendűre való áttérés szinte mindig információvesztéssel jár. Ezért a kapcsolatvizsgálatok

Az asszociációs kapcsolatot kombinációs-táblából kiindulva elemezzük, ezért praktikusan csak olyan ismérvek vizsgálatára nyílik lehetőségünk, amely változatainak száma korlátozott[3]. A vegyes kapcsolatok vizsgálata során a tényezőváltozóként szereplő minőségi ismérv alapján csoportosítjuk a sokaságot, ebből következik, hogy elvárjuk, hogy a tényezőváltozó vizsgálatba vont ismérvváltozatainak a száma ne legyen kezelhetetlenül nagy. A korrelációs kapcsolat elemzésekor általában két (vagy több) intervallum-, vagy arányskálán mért változó páronként (hármasonként stb.) összetartozó értékeiből indulunk ki, így itt elégséges megkövetelnünk a vizsgálatba vont adatpároktól, hogy számuk - számolástechnikai okok miatt - véges legyen. Láthatjuk, hogy az elemzés alapeszközei alkalmazkodnak az ismérvek tulajdonságaihoz, hiszen a kvalitatív ismérveknek általában kevesebb változata van, mint a kvantitatív változóknak.

Kézenfekvő csoportosítása a jellemzők közötti összefüggéseknek a bennük szereplő változók száma alapján történő osztályozás. Ennek megfelelően például a két ismérv közötti kapcsolatot kétváltozós, a három ismérv közötti összefüggést háromváltozós kapcsolatnak nevezzük (természetesen a sor folytatható). Tanulmányaink ezen szintjén csak olyan összefüggéseket vizsgálunk, melyben egy eredményváltozó és egy, vagy több magyarázó változó szerepel; azaz nem térünk ki az olyan problémákra, amelyekben a függő változók száma is meghaladja az egyet.

során is törekednünk kell arra, hogy mindig a lehető legtöbb információt tartalmazó mértékkel, és az ennek megfelelő módszerrel végezzük az elemzést.

[3] Könnyen belátható, hogy egy kombinációs tábla elemzésre alkalmatlan, ha az abban szereplő ismérvek oly sok változattal rendelkeznek, hogy a tábla rovatainak túlnyomó része 0, vagy 1 gyakoriságot tartalmaz.

6.1 A SZTOCHASZTIKUS KAPCSOLATOK VIZSGÁLATÁNAK STATISZTIKAI MÓDSZEREI

A statisztikai ismérvek közötti összefüggések csoportosítása után vizsgáljuk meg, milyen módszertani eszközök állnak rendelkezésünkre ezek elemzéséhez. Annak alapján, hogy vizsgálatunk adatbázisát milyen sokaság képezi, a sztochasztikus kapcsolatok analizálása során felhasznált statisztikai módszereket két csoportba oszthatjuk:

- alapsokaság elemzését kapcsolatszorossági mérőszámokkal végezzük (leíró statisztika);
- részsokaság (minta) elemzése során az ismérvek közötti kapcsolat szignifikáns voltát tesztelő hipotézisellenőrzési eljárásokat alkalmazunk (következtetéses statisztika).

Az első csoportba tartozó együtthatókkal a kapcsolat intenzitását kívánjuk számszerűsíteni, azaz valamilyen mértéket akarunk meghatározni, amelynek alapján eldönthető, hogy egy adott kapcsolat szorosabb-e, mint egy másik. A kapcsolatszorossági mérőszámokkal szemben négy olyan követelményt határozunk meg, melyek teljesülése egy széles körben, általánosan használandó mutatótól elvárható:

1) az együttható legyen **meghatározható**, azaz lehetőség szerint kevéssé legyen érzékeny a mutató számszerűsíthetősége az olyan extrém esetekre, melyben például az egyik ismérv szerinti megoszlás egy ismérvváltozat körül tömörül; vagy ha előfordulnak olyan ismérvváltozat-kombinációk, melyekhez tartozó gyakoriságok nullák;

2) az együttható **abszolút értéke legyen a [0;1] zárt intervallumban**, azaz teljesüljön az egyenlőtlenség:

$$0 \le |\Theta| \le 1 \qquad (6\text{-}1)$$

ahol Θ a kapcsolatszorossági együttható. Annak kikötése, hogy a mutató abszolút értéke szerepeljen a fenti intervallumban azért lényeges, mert bizonyos típusú kapcsolatokban (ahol az ismérvek ordinális-, intervallum-, vagy arányskálán mérhetők) értelmezzük a kapcsolat pozitív, illetve negatív irányát is. Az ilyen összefüggés szorosságát mérő mutatószámok a [-1;1] intervallumba kell, hogy essenek, ugyanis előjelük a kapcsolat irányáról tájékoztat. A mutató 0 értéke esetén a kapcsolat hiányáról, ±1 értéke esetén függvényszerű (determinisztikus) kapcsolatról beszélünk;

3) legyen a mutató **monoton**, vagyis szorosabb kapcsolat esetén a mérőszám abszolút értéke legyen magasabb;

4) teljesítse a mutató azon elvárásunkat, miszerint **akkor** és **csak akkor** veszi fel a szélsőértékeit, ha valóban a kapcsolat teljes hiányával, vagy függvényszerű kapcsolattal állunk szemben.

A fenti feltételek mellett az együtthatókkal szemben meg szokták határozni a **szimmetria** követelményét is, mely szerint ha egy adott módszerrel vizsgáljuk A és B változó kapcsolatát, akkor a mutató szerint A ugyanolyan szoros kapcsolatban álljon B-vel, mint B A-val. Ez látszólag logikus követelmény, ám ha a kapcsolat szorosságának vizsgálata során a változók között megkülönböztetjük a magyarázó-, illetve eredményváltozókat, akkor értelmezhető különböző szorosságú kapcsolat is. (Gondoljunk a változók közötti ok-okozati viszonyokról elmondottakra! Ha A változó oka B-nek, de B változó nem oka A-nak, akkor értelmezhető az az eredmény is, amikor A és B között szorosabb kapcsolatot találunk abban az esetben, ha modellünkben A magyarázza B-t, mint akkor, ha B magyarázza A-t.) Mivel a változók közötti kapcsolatok szorosságának vizsgálata régebbi eredetű, mint az oksági viszonyok elemzése, ezért a bemutatandó mutatók nagy része megfelel a szimmetria követelményének is.

A társadalmi-gazdasági elemzések során gyakran szembesülünk olyan problémával, melyben nem csupán kettő ismérv kapcsolatát kell elemeznünk. A kettőnél több változót tartalmazó összefüggések esetében a sztochasztikus kapcsolatokat számszerűsítő mutatószámoknak három típusát különböztethetjük meg, annak alapján, hogy milyen mértékben használják ki az összes vizsgálatba vont változóban rejlő információt.

Így beszélhetünk[4]:

- **totális** együtthatókról, amelyekkel csak két-két változó összefüggését elemezzük úgy, hogy teljesen figyelmen kívül hagyjuk a vizsgálatba vont további változó(k) hatását (Θ_{AB}, Θ_{AC} és Θ_{BC});
- **parciális** mutatókról, melyekkel két változó összefüggését úgy vizsgáljuk, hogy "kiszűrjük" a vizsgálatban szereplő összes többi változó hatását ($\Theta_{AB\,C}$, $\Theta_{AC\,B}$ és $\Theta_{BC\,A}$);
- **többszörös** együtthatóról, mellyel a vizsgálatban szereplő összes változónak a kapcsolatát számszerűsítjük (Θ_{ABC}).

Az előbbiekből kitűnik, hogy amennyiben vizsgálatunkban mindössze két változó szerepel, úgy ezek között mindig totális kapcsolatszorossági együtthatót számszerűsítünk. A másik lényeges összefüggés, hogy mindig a vizsgálatban szereplő ismérvek alapján határozható meg a mutatószám milyensége. Mivel a gazdasági

[4] A fenti csoportosítást tulajdonképpen a korrelációs kapcsolat elemzése során használjuk. Mindez analóg módon értelmezhető a másik két kapcsolattípusra is, ám a nominális skálán mért változók esetében az értelmezés kissé nehézkes. Zárójelben feltüntetjük a későbbiekben általánosan használt jelölését az együtthatónak. Példánkban három változót tételezünk fel, melyek közül A az eredményváltozó, B és C magyarázó változók.

jelenségek elemzése során egy adott változó alakulását befolyásoló összes tényező felsorolása, illetve vizsgálatba vonása szinte lehetetlen, ezért lényeges felsorolni, hogy pl. egy parciális együttható meghatározása során melyek azok a jellemzők, melyek hatását kiszűrtük. (Értelemszerűen törekednünk kell arra, hogy a vizsgálatunkba minél több lényeges ismérvet bevonjunk, ezáltal ugyanis elérhető, hogy a kimaradt változók hatása elenyésző legyen.)

A kapcsolatszorossági-együtthatók értelmezését mindig az adott vizsgálatnak, az elemzésben szereplő egyedeknek és az ismérvek körének ismeretében kell elvégeznünk. Általánosságban a mutatók értelmezése a következő séma alapján történik[5]:

- ha $\Theta = 0$, akkor a nincs kapcsolat a vizsgált ismérvek között;
- ha $0 < |\Theta| < 0{,}3$, akkor a kapcsolat gyenge ;
- ha $0{,}3 \leq |\Theta| \leq 0{,}7$, akkor közepes szorosságú a kapcsolat ;
- ha $0{,}7 < |\Theta| < 1$, akkor erős kapcsolatról beszélünk ;
- ha $|\Theta| = 1$, akkor a kapcsolat függvényszerű .

Az együtthatók abszolút értéke mellett értelmezhetjük a mutatók négyzetét is. A statisztikában bevett gyakorlat szerint a kapcsolatszorossági-mérőszámok négyzetét **determinációs együtthatónak** nevezzük, és magyarázó erőként interpretáljuk, vagyis Θ^2 megmutatja, hogy a tényezőváltozó(k) az egyedek eredményváltozó szerinti különbözőségének (szóródásának) hány százalékát magyarázza(k) meg.

Amennyiben nem áll módunkban a teljes alapsokaságot felmérni, úgy a sztochasztikus kapcsolatok elemzését minta alapján is elvégezhetjük. Amennyiben az

[5] A mérőszámok értéke erősen függ a vizsgált sokaság nagyságától és - esetenként - az ismérvváltozatok számától, így a kapcsolat intenzitásának értelmezése során erre is tekintettel kell lennünk.

előbb bemutatott leíró statisztikai eszközt (kapcsolatszorossági mérőszámot) egy részsokaságra vonatkoztatva határozzuk meg, mindössze arról kapunk információt milyen szoros a kapcsolat a mintában a vizsgálandó ismérvek között. Ahhoz, hogy az egész alapsokaságra vonatkozó megállapításokat tehessünk, szükséges valamilyen következtetéses statisztikai módszer használata. A kapcsolatok vizsgálatának (meglétének, illetve hiányának) második alapvető módszere a hipotézisellenőrzés. Ennek során tesztelni kívánjuk, vajon a vizsgálatba vont változók között a kapcsolat valóban szignifikáns-e, avagy az ismérvek közötti tendenciaszerűnek tűnő összefüggés a véletlen hatására alakult ki.

Az elemzésünk során alkalmazott hipotézisrendszer:

$$\begin{aligned} H_0 &: \Theta = 0 \\ H_1 &: |\Theta| > 0 \end{aligned} \qquad (6\text{-}2)$$

Láthatjuk, hogy a nullhipotézis elfogadása a sztochasztikus kapcsolat hiányát, elvetése a kapcsolat meglétét jelenti. Ennek értelmében adott szinten szignifikánsnak tekintjük a kapcsolatot, ha a próbafüggvény értéke a kritikus tartományba esik, vagyis meghaladja a táblabeli értéket.

A kapcsolatvizsgálati módszerekről általánosságban elmondottak után vizsgáljuk meg a különböző típusú összefüggések elemzésére szolgáló konkrét eljárásokat. A következőkben elsőként az asszociációs-, majd a vegyes-, végül a korrelációs kapcsolat szorosságát mérő együtthatókat, valamint az ezek szignifikáns voltát tesztelő hipotézisellenőrzési eljárásokat ismertetjük.

6.2 AZ ASSZOCIÁCIÓS KAPCSOLAT ELEMZÉSE

A kapcsolatok csoportosításakor említettük, hogy az asszociációs kapcsolat mérése kontingenciatábla alapján történik. A későbbiekben alkalmazandó jelölésrendszer egységesítése érdekében elevenítsük fel a kétdimenziós kontingenciatáblát.

A kétdimenziós kontingenciatábla sémája

6-1. tábla

"A" ismérv változatai	"B" ismérv változatai				
	B_1	B_2	...	B_t	$\sum$
A_1	f_{11}	f_{12}	...	f_{1t}	$f_{1.}$
A_2	f_{21}	f_{22}	...	f_{2t}	$f_{2.}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_s	f_{s1}	f_{s2}	...	f_{st}	$f_{s.}$
$\sum$	$f_{.1}$	$f_{.2}$	...	$f_{.t}$	n

Tudjuk, hogy a táblában f_{ab} gyakoriság megmutatja azon egyedek számát, akik az "A" változó szempontjából az a. (a=1,...,s), a "B" változó szempontjából a b. (b=1,...,t) ismérvváltozathoz tartoznak. A tábla elemeire érvényesek a következő összefüggések:

$$\sum_{b=1}^{t} f_{ab} = f_{a.} \qquad \sum_{a=1}^{s} f_{ab} = f_{.b}$$

$$\sum_{a=1}^{s}\sum_{b=1}^{t} f_{ab} = \sum_{a=1}^{s} f_{a.} = \sum_{b=1}^{t} f_{.b} = n$$

Az asszociációs kapcsolat vizsgálata elvégezhető a relatív gyakoriságok alapján is. Ebben az esetben a kombinációs tábla a következőképpen módosul:

Kétdimenziós kontingenciatábla a relatív gyakoriságok alapján

6-2. tábla

"A" ismérv változatai	"B" ismérv változatai				
	B_1	B_2	...	B_t	Σ
A_1	p_{11}	p_{12}	...	p_{1t}	$p_{1.}$
A_2	p_{21}	p_{22}	...	p_{2t}	$p_{2.}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_s	p_{s1}	p_{s2}	...	p_{st}	$p_{s.}$
Σ	$p_{.1}$	$p_{.2}$	...	$p_{.t}$	1

A fenti tábla elemeire a következő összefüggések érvényesek:

$$p_{ab} = \frac{f_{ab}}{n}$$

$$\sum_{b=1}^{t} p_{ab} = p_{a.} \qquad \sum_{a=1}^{s} p_{ab} = p_{.b}$$

$$\sum_{a=1}^{s}\sum_{b=1}^{t} p_{ab} = \sum_{a=1}^{s} p_{a.} = \sum_{b=1}^{t} p_{.b} = 1$$

A következőkben a fenti kétdimenziós kombinációs táblák alapján végezhető, két ismérv közötti asszociációs kapcsolat mérését és tesztelését mutatjuk be.

A minőségi (nominális, vagy ordinális skálán mérhető) ismérvek közötti kapcsolatok szorosságának mérésére számos mutatószámot dolgoztak ki a statisztika művelői. Tankönyvünkben ezek közül csak a legelterjedtebb, ún. **hagyományos asszociációs mérőszámokkal** foglalkozunk[6]. Az ismertetendő mérőszámokra jellemző, hogy megszerkesztésükkor a függetlenség követelményéből indulunk ki.

---

[6] A tananyagban nem szereplő asszociációs mérőszámok széles körét mutatja be Irodalom [37].

$$P(A \cap B) = P(A)P(B) \tag{6-3}$$

vagyis, amennyiben az A és a B esemény függetlenek, akkor együttes bekövetkezési valószínűségük megegyezik a egyedi bekövetkezési valószínűségeik szorzatával. Amennyiben az A és B ismérvek függetlenek egymástól, akkor annak valószínűsége, hogy egy egyed az a és a b ismérvváltozattal jellemezhető, kifejezhető a peremvalószínűségek szorzataként (lásd! TK. I. 3. fejezet). Tehát a **függetlenség esetére érvényes valószínűség**:

$$p^{*}_{ab} = p_a p_b \tag{6-4}$$

Természetesen nemcsak a relatív, hanem a **tényleges gyakoriságok függetlenség esetén várható értékét** is meghatározhatjuk:

$$f^{*}_{ab} = \frac{f_a f_b}{n} \tag{6-5}$$

Az első kísérletek az asszociációs kapcsolat szorosságának meghatározására abból az egyszerűsítésből indultak ki, hogy a vizsgálandó mindkét ismérv alternatív, azaz csak két ismérvváltozattal rendelkezik[7].

[7] A változók minden esetben alternatívvá transzformálhatók az általunk legfontosabbnak tartott ismérvváltozat kiválasztásával és a többi változat összevonásával; ám ez lényeges információk elvesztéséhez vezet. (Példa lehet erre a tantárgyak minősítése: az öt fokozatú skála (jeles, jó, közepes, elégséges, elégtelen - alternatívvá alakítható - megfelelt, nem felelt meg.)

Alternatív ismérveket tartalmazó kombinációs tábla sémája

6-3. tábla

A/B	B_1	B_2	Σ
A_1	f_{11}	f_{12}	$f_{1.}$
A_2	f_{21}	f_{22}	$f_{2.}$
Σ	$f_{.1}$	$f_{.2}$	n

A kétállapotú (bináris) változók közötti kapcsolat elemzése során használható első mutatószám kidolgozása **G.U.YULE** nevéhez fűződik. Mivel függetlenség esetén:

$$f_{11}=\frac{f_{1.}f_{.1}}{n} \quad ; \quad f_{12}=\frac{f_{1.}f_{.2}}{n} \quad ; \quad f_{21}=\frac{f_{2.}f_{.1}}{n} \quad ; \quad f_{22}=\frac{f_{2.}f_{.2}}{n}$$

ezért:

$$f_{11}f_{22}=\frac{f_{1.}f_{.1}f_{2.}f_{.2}}{n^2}=\frac{f_{1.}f_{.2}f_{2.}f_{.1}}{n^2}=f_{12}f_{21} \text{ vagyis:}$$

$$f_{11}f_{22}-f_{12}f_{21}=0$$

A fenti különbségről belátható, hogy értéke a $[-n^2/4;+n^2/4]$ intervallumban szóródik. Annak érdekében, hogy a mutató a kapcsolatszorossági mérőszámokkal szemben támasztott második követelménynek megfeleljen, transzformáljuk a különbséget, hogy abszolút értéke a $[0;1]$ intervallumba kerüljön:

$$Y=\frac{f_{11}f_{22}-f_{12}f_{21}}{f_{11}f_{22}+f_{12}f_{21}} \qquad (6\text{-}6)$$

A mutató értéke 0, ha az ismérvek függetlenek. Az együttható előjelét ordinális skálán mérhető változók esetén értelmezhetjük, a következő módon: pozitív előjel esetén az "A" változó szempontjából előbbre rangsorolt egyedek "vonzzák" a "B" változó szempontjából előbbre rangsorolt egyedeket; negatív előjel esetén "taszítják" ezeket[8]. Nézzünk erre egy példát!

Egy félév végi kollokvium előtt a tantárgy oktatója konzultációt tartott. A vizsgaköteles 100 hallgatóból 60 vett részt a konzultáción. A két nappal későbbi vizsgán 20 hallgató nem felelt meg, közülük 5 vett részt a konzultáción. Határozzuk meg, segítette-e a konzultáció a vizsgára való felkészülést, azaz milyen szoros az asszociációs kapcsolat a konzultáción való részvétel és a vizsga sikeressége között !

Két alternatív minőségi ismérvvel állunk szemben, ezért alkalmazhatjuk a Yule-mutatót. (Az ismérveket ordinális skálán mérhetőnek tekinthetjük, ha a konzultáción való részvételt nagyobb intenzitású felkészülésként, a vizsgán való megfelelést jobb eredményként értelmezzük.) Információink a következő táblába foglalhatók:

Konzultáción	Vizsgán megfelelt		Σ
	Nem	Igen	
nem vett részt	15	25	40
részt vett	5	55	60
Σ	20	80	100

Képezzük az Y-mutatót:

$$Y = \frac{15 \times 55 - 25 \times 5}{15 \times 55 + 25 \times 5} = \frac{825 - 125}{825 + 125} = \frac{700}{950} = 0{,}7368$$

[8] A nominális skálán mért változók közötti kapcsolat iránya nem értelmezhető, ilyenkor csak a mutató abszolút értéke lényeges a kapcsolat vizsgálata szempontjából.

Tehát a konzultáción való részvétel és a vizsgaeredmények között szoros kapcsolat mutatható ki, vagyis a konzultáció "segítette" a vizsgára való felkészülést.

A Yule-féle asszociációs együtthatóról megmutatható, hogy nem minden szempontból felel meg a kapcsolatszorossági-mérőszámokkal szemben támasztott követelményeknek. Láttuk, hogy csak alternatív ismérvek esetén alkalmazható, így olyankor, ha bármelyik ismérv kettőnél több ismérvváltozattal rendelkezik, az eredeti adatokból nem számítható. Szintén hiányossága a mérőszámnak, hogy értéke nem determinisztikus kapcsolat esetén is elérheti az 1-et. Könnyen belátható, hogy ha a táblában található egy 0 gyakoriság, akkor a mutató abszolút értéke 1 lesz, noha ebben az esetben nem feltétlenül határozható meg az egyik ismérv szerinti hovatartozás alapján a másik ismérv szerinti besorolás. (Determinisztikus kapcsolat esetén ugyanis vagy a tábla fő-, vagy mellékátlójának mindkét eleme nulla lenne.) A fenti problémák azt eredményezik, hogy a mutatót ritkán alkalmazzuk, általában csak előzetes tájékozódásra használjuk.

A Yule-együttható hiányosságai arra késztették a statisztika művelőit, hogy ennél általánosabban is használható, a követelményekhez jobban igazodó mutatót szerkesszenek. Amennyiben az ismérvek nem függetlenek egymástól, akkor a relatív gyakoriságok, illetve a tényleges gyakoriságok eltérnek az (6-4), illetve (6-5) képletekben meghatározottaktól. Ezen eltérések annál nagyobbak, minél távolabb van a kapcsolat a függetlenségtől. A fentiekből kiindulva megszerkeszthetjük a **négyzetes kontingencia** mutatóját, a relatív gyakoriságok alapján:

$$\chi^2 = n\sum_{a=1}^{s}\sum_{b=1}^{t}\frac{\left(p_{ab} - p_{a.}p_{.b}\right)^2}{p_{a.}p_{.b}} \qquad (6\text{-}7)$$

illetve a tényleges gyakoriságok alapján:

$$\chi^2 = \sum_{a=1}^{s}\sum_{b=1}^{t}\frac{\left(f_{ab} - \frac{f_{a.}f_{.b}}{n}\right)^2}{\frac{f_{a.}f_{.b}}{n}} \tag{6-8}$$

Az (6-7) és (6-8) képletek egyszerűsíthetők[9]:

$$\chi^2 = n\left(\sum_{a=1}^{s}\sum_{b=1}^{t}\frac{f_{ab}^2}{f_{a.}f_{.b}} - 1\right) = n\left(\sum_{a=1}^{s}\sum_{b=1}^{t}\frac{p_{ab}^2}{p_{a.}p_{.b}} - 1\right) \tag{6-9}$$

A mutató értéke az alábbi intervallumban található:

$$0 \leq \chi^2 \leq (s-1)n \qquad (s \leq t)$$

A mutató értékének felső határát a két változó ismérvváltozat számának minimuma befolyásolja. A kisebb ismérvváltozatszám természetesen tartozhat a "B" változóhoz is. A négyzetes kontingencia mellett gyakran használatos - a különböző megfigyelésszámot tartalmazó kapcsolatok összehasonlításának megkönnyítésére - az **átlagos négyzetes kontingencia** mutatója:

$$\Phi^2 = \frac{\chi^2}{n} \tag{6-10}$$

A mutató értéke a [0;(s-1)] intervallumban szóródik, vagyis maximális értéke csak abban az esetben nem nagyobb 1-nél, ha legalább az egyik ismérv alternatív. Ezen hiányosság kiküszöbölését számos kutató kísérelte meg. Az együtthatót előbb **K. PEARSON**, majd **A.A. CSUPROV** módosította, de ezek a változatok sem feleltek

[9] A bizonyítást lásd! pl. Irodalom [36].

meg minden követelménynek. A feltételeket minden szempontból kielégítő mérőszám megalkotása **H. CRAMER** nevéhez fűződik. A **Cramer-együttható**:

$$C = \sqrt{\frac{\Phi^2}{(s-1)}} = \sqrt{\frac{\chi^2}{n(s-1)}} \qquad (s \leq t) \qquad (6\text{-}11)$$

értéke 0, ha a két ismérv független; értéke 1, ha köztük a kapcsolat determinisztikus. Láthatjuk, hogy a gyökvonás következtében a mutató értéke mindig pozitív, így a kapcsolat irányának jelzésére még ordinális skálán mérhető változók esetén sem alkalmas. A Cramer-együttható használatát a következő példán mutatjuk be.

Egy terület népszámlálási adatai alapján a 14 éven felüli lakosság településtípus és iskolai végzettség szerinti megoszlásáról a következők ismeretesek (ezer fő):

Iskolai	Településtípus			
végzettség	Budapest	Vidéki város	Község	Összesen
8 általános	300	400	1.500	2.200
Szakmunkás	600	1.200	2.200	4.000
Érettségi	500	500	200	1.200
Felsőfokú végzettség	200	300	100	600
Összesen	1.600	2.400	4.000	8.000

Határozzuk meg milyen szoros a kapcsolat az iskolai végzettség és a különböző településtípusok között!

Az ismérvváltozatok száma 3 és 4, ezért a Cramer-féle mutató számszerűsítésével válaszolható meg a feladat. Határozzuk meg ehhez először a függetlenség esetén érvényes gyakoriságokat (f_{ab}•):

Iskolai végzettség	Településtípus			
	Budapest	Vidéki város	Község	Összesen
8 általános	440	660	1.100	2.200
Szakmunkás	800	1.200	2.000	4.000
Érettségi	240	360	600	1.200
Felsőfokú végzettség	120	180	300	600
Összesen	1.600	2.400	4.000	8.000

A tényleges-, illetve az ismérvek függetlenségének feltételezésével számított gyakoriságokból meghatározható a négyzetes kontingencia értéke, majd ebből a Cramer-mutató értéke:

$$\chi^2 = \frac{(300-440)^2}{440} + \frac{(400-660)^2}{660} + \ldots + \frac{(100-300)^2}{300} = 1.231{,}87$$

$$\chi^2 = 8.000\left(\frac{300^2}{1.600 \times 2.200} + \frac{400^2}{2.000 \times 2.200} + \ldots + \frac{100^2}{4.000 \times 600} - 1\right) = 1.231{,}87$$

$$\Phi^2 = \frac{1.231{,}87}{8.000} = 0{,}1540$$

$$C = \sqrt{\frac{0{,}1540}{(3-1)}} = 0{,}2775$$

Az asszociációs együttható alapján elmondható, hogy a település szerinti hovatartozás és az iskolai végzettség között gyenge sztochasztikus kapcsolat mutatható ki.

Az asszociációs kapcsolat vizsgálata során viszonylag ritkán nyílik lehetőségünk arra, hogy a teljes alapsokaság alapján végezzük el az elemzésünket. A mintából történő kapcsolatvizsgálat két módszerrel végezhető el:

- az asszociációs együttható becslésével;
- a kapcsolat szignifikáns voltára vonatkozó hipotézis tesztelésével.

Az asszociációs együtthatók mintából történő pont- és intervallumbecslésének bemutatása meghaladja a tananyag kereteit, ehhez ugyanis a hagyományos ismertetett - mérőszámok mellett további mutatók tárgyalása is szükséges lenne[10].

A kapcsolat szignifikáns voltára vonatkozó feltevésünket a **függetlenség-vizsgálat** módszerével válaszolhatjuk meg. A teszteljárás a négyzetes kontingencia mutatójára épül. A vizsgálat során alkalmazott hipotézis-rendszer:

[10] Lásd! Irodalom [18].

$$H_0: \chi^2 = 0$$
$$H_1: \chi^2 > 0 \qquad (6\text{-}12)$$

Tehát a nullhipotézis szerint az ismérvek függetlenek, míg az alternatív hipotézis elfogadása a sztochasztikus kapcsolat szignifikáns voltát jelenti.

Belátható, hogy a $\dfrac{\left(f_{ab} - \dfrac{f_{a.}f_{.b}}{n}\right)^2}{\dfrac{f_{a.}f_{.b}}{n}}$ valószínűségi változók standard normális eloszlást követő valószínűségi változók négyzetei, így az ezek összegeként képzett négyzetes kontingencia mutató χ^2-eloszlást követ. Az eloszlás szabadságfoka elvben (st-1), hiszen az összeg (st) tagból áll, melyek közül az utolsó a minta nagyságából eredően determinált. Mivel azonban az esetek túlnyomó részében a perem-valószínűségeket is becsülnünk kell, így a szabadságfok csökken a becsülendő paraméterek számával. Tudjuk, hogy a peremvalószínűségek összege sorok és oszlopok szerint is 1, így összesen csak (s-1) és (t-1) peremvalószínűséget kell a minta alapján meghatározni. A fentiekből következően, a nullhipotézis teljesülése esetén a négyzetes kontingencia χ^2-eloszlást követ, (st-1-(s-1+t-1)) = (s-1)(t-1) szabadságfokkal. Vagyis:

$$\chi^2_{emp} = \sum_{a=1}^{s}\sum_{b=1}^{t} \frac{\left(f_{ab} - \dfrac{f_{a.}f_{.b}}{n}\right)^2}{\dfrac{f_{a.}f_{.b}}{n}} \,\Big|\, H_0 \approx \chi^2_{(s-1)(t-1)} \qquad (6\text{-}13)$$

Ennek ismeretében a hipotézisellenőrzés könnyen elvégezhető:

1) Meghatározzuk az empirikus χ^2 értéket (χ^2_{emp}).

2) Kikeressük a választott szignifikancia-szinthez tartozó a kritikus értéket (χ^2_{elm}).

3) Amennyiben az általunk számított érték kisebb úgy a nullhipotézist, ellenkező esetben az alternatív hipotézist fogadjuk el.

A fentiek bemutatására tekintsük a következő példát!

Az orvostudomány régóta kutatja, melyek azok a tényezők, amelyek terhességi rendellenességeket idéznek elő. Feltételezhető, hogy a terhesség alatti dohányzás nehezíti a magzat kihordását, vagyis spontán vetélést vagy koraszülést okoz. Egy célzott vizsgálat során véletlenszerűen kiválasztott 1000, a gyermekét vállaló terhes adatait tartalmazza a következő tábla:

A terhesség	A terhes		Összesen
kimenetele	dohányzott	nem dohányzott	
Spontán vetélés	50	40	90
Koraszülés	70	30	100
Érett szülés	180	630	810
Összesen	300	700	1.000

Határozzuk meg a kapcsolat szorosságát mérő együtthatót, és vizsgáljuk meg, hogy 5 %-os szignifikancia-szinten függetlennek tekinthető-e a szülés kimenetele a terhesség alatti dohányzástól!

Elsőként a minta alapján határozzuk meg a négyzetes kontingencia becsült értékét:

$$\chi^2_{emp} = \frac{\left(50 - \frac{300 \times 90}{1.000}\right)^2}{\frac{300 \times 90}{1000}} + \ldots + \frac{\left(630 - \frac{700 \times 810}{1.000}\right)^2}{\frac{700 \times 810}{1000}} =$$

$$= 1.000\left(\frac{50^2}{300 \times 90} + \ldots + \frac{630^2}{700 \times 810} - 1\right) = 127{,}51$$

A (3-1) (2-1)=2 szabadságfokú χ^2-eloszlás különböző szignifikancia-szintekhez (α) tartozó eloszlásfüggvény-értékét tartalmazza a következő tábla (az adatok a χ^2-eloszlás táblájából származnak, lásd! TÁBLÁZATOK):

α	0,100	0,050	0,025	0,010
χ^2	4,605	5,991	7,378	9,210

A mintából számított érték 127,51, vagyis minden ésszerű szignifikancia-szinten meghaladja a táblabeli értéket, így a kapcsolat szignifikánsnak tekinthető.

6.3 A VEGYES KAPCSOLAT ELEMZÉSE

A gyakorlati elemzések során gyakran merül fel annak az igénye, hogy különböző típusú - kvalitatív és kvantitatív - ismérvek egymás közötti kapcsolatát elemezzük. Ez annyit jelent, hogy megkíséreljük a minőségi ismérv (ismérvek) alapján képzett csoportok mennyiségi ismérv szerinti különbözőségét számszerűsíteni. Az elemzés kiterjed a minőségi ismérv(ek) és a mennyiségi változó közötti kapcsolat szorosságának mérésére, valamint annak megállapítására, hogy a csoportosító minőségi ismérv szerint homogénnek tekinthető-e a sokaság. Tananyagunkban csak azokkal a vegyes kapcsolatokkal foglalkozunk, melyekben egy mennyiségi- és egy vagy több minőségi ismérv szerepel[11]. A vegyes kapcsolat jellemzése során alkalmazott eszközök - a szórásnégyzet összetevőkre bontása, illetve a varianciaanalízis - korábbi tanulmányainkból ismeretesek (lásd! TK. 1. 2. fejezet), így módszertanilag csak a több csoportosító ismérvet tartalmazó összefüggések elemzése lesz újdonság.

6.3.1 Az egy minőségi ismérvet tartalmazó vegyes kapcsolat elemzése

Az egy minőségi és egy mennyiségi ismérvet tartalmazó vegyes kapcsolat elemzését a következőképpen rendezett sokaságból végezzük:

A kétváltozós vegyes kapcsolat adatbázisa

6-4.tábla

Megfigyelés	A minőségi ismérv változatai			
	1.	2.	...	m.
1.	x_{11}	x_{21}	...	x_{m1}
2.	x_{12}	x_{22}		x_{m2}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n_i	x_{1n1}	x_{2n2}		x_{mnm}

A fenti tábla áttekintése során fontos megjegyezni, hogy az egyes csoportokhoz (a minőségi ismérv különböző változataihoz) eltérő számú gyakoriság is tartozhat, vagyis pl. n_1 nem feltétlenül egyezik meg n_m-mel. Mivel általában a mennyiségi ismérvek

[11] Az olyan vegyes kapcsolatok elemzése, melyekben több mennyiségi ismérv szerepel, a kvalitatív magyarázó változókat is tartalmazó regressziós modell segítségével végezhető el.

igen sok változattal rendelkeznek, a sokaság fenti módon való leírása nehézkes. Ezért gyakran az alapsokaságot minden egyedének felsorolása helyett a csoportok átlagával és szórásával jellemezzük. A következőkben bemutatandó eszközök mindegyike feltételezi az alapsokaságon és az egyes csoportokon belül a mennyiségi ismérv normális eloszlását.

A vegyes kapcsolat elemzéséhez írjuk fel a varianciaanalízis ismert modelljét (lásd! TK. I. 5. fejezet):

$$x_{ij} = \mu + \tau_j + \varepsilon_{ij} \qquad (6\text{-}14)$$

ahol a j-edik csoport i-edik eleme (x_{ij}), a teljes sokaságra vonatkozó várható érték (μ), a j-edik osztály csoporthatása (τ_j) és az ε_{ij} véletlen hatás összegeként adódik. Könnyen belátható, hogy minél szorosabb a kölcsönhatás a csoportosító minőségi ismérv és a mennyiségi ismérv között, annál jelentősebb a modellben az egyes csoportokhoz való tartozás hatása. Legyen a j-edik csoportra jellemző várható érték (μ_j):

$$\mu_j = \mu + \tau_j$$

Mivel:

$$\tau_j = \mu_j - \mu \quad \text{és} \quad \varepsilon_{ij} = x_{ij} - \mu_j$$

ezért az (6-14) képlet felírható:

$x_{ij} = \mu + (\mu_j - \mu) + (x_{ij} - \mu_j)$ Vagy némiképp átalakítva:

$$x_{ij} - \mu = (\mu_j - \mu) + (x_{ij} - \mu_j)$$

A fenti képlet a várható értékektől (átlagoktól) való eltérésekre ír fel összefüggést, így kézenfekvő a vegyes kapcsolat elemzését is ezen eltérések alapján elvégezni. Vegyük észre, hogy amennyiben az eltérések egész sokaságra vonatkozó átlagait kell meghatároznunk, akkor a feladat nem más, mint a sokasági teljes-, külső- és belső szórás kiszámítása. Minél szorosabb a vegyes kapcsolat, annál nagyobb - a csoportosítás hatását számszerűsítő - külső szórás részaránya a teljes szóráson belül. A vegyes kapcsolat szorosságát mérő mutatószám, a **szóráshányados** is erre az arányra épül:

$$H = \frac{\sigma_K}{\sigma} = \sqrt{\frac{\sigma_K^2}{\sigma^2}} = \sqrt{1 - \frac{\sigma_B^2}{\sigma^2}} \qquad (6\text{-}15)$$

A szóráshányadost a sztochasztikus kapcsolat szorosságát mérő mutatóként értelmezzük: amennyiben értéke 0, úgy a kapcsolat hiányáról, ha értéke 1, akkor függvényszerű kapcsolatról beszélünk. A mutató négyzetét (**szórásnégyzet-hányados**) magyarázó erőként értelmezzük, vagyis H^2 megmutatja, hogy a mennyiségi ismérv szóródásának hány százalékát magyarázza a csoportosító minőségi ismérv szerinti hovatartozás. (A fennmaradó ($100-H^2$) százalék a véletlen hatásának tudható be.)

A fentiek szemléltetésére tekintsük a következő példát!

Egy könyvkiadó vállalat 1990-es évre vonatkozó forgalmi adatai közül az alábbiak ismeretesek:

Jellemző	Az eladott könyvek típusa		
	Regény	Verses mű	Színmű
Kiadott művek száma	40	12	10
Átlagos példányszám	100.000	40.000	28.000
A példányszám szórása	20.000	15.000	10.000

Elemezzük a könyvek típusa és az eladási példányszám közötti sztochasztikus kapcsolatot!

Mivel a kapcsolatban a tényezőváltozó minőségi-, és a függő változó mennyiségi ismérv, így vegyes kapcsolatról van szó. A kapcsolat szorosságának vizsgálatát a szóráshányados mutatója alapján végezzük el. Ennek számítása:

$$\overline{x}_1 = 100.000 \quad \overline{x}_2 = 40.000 \quad \overline{x}_3 = 28.000$$

$$\overline{x} = \frac{40 \times 100.000 + 12 \times 40.000 + 10 \times 28.000}{40 + 12 + 10} = 76.775$$

$$\sigma_1 = 20.000 \quad \sigma_2 = 15.000 \quad \sigma_3 = 10.000$$

$$\sigma_B = \sqrt{\frac{40 \times 20.000^2 + 12 \times 15.000^2 + 10 \times 10.000^2}{40 + 12 + 10}} = 17.825$$

$$\sigma_K = \sqrt{\frac{40(100.000 - 76.775)^2 + \ldots + 10(28.000 - 76.775)^2}{40 + 12 + 10}} = 31.519$$

$$\sigma = \sqrt{17.825^2 + 31.519^2} = 36.210$$

$$H = \frac{31.519}{36.210} = \sqrt{1 - \frac{17.825^2}{36.210^2}} = 0,8704$$

$$H^2 = 0,7576$$

A fentiek alapján tehát megállapítható, hogy a kiadott művek típusa és a kiadási példányszám között szoros sztochasztikus kapcsolat található, a művek típusa a példányszám szerinti szóródás 75,76 %-át magyarázza meg.

A gazdasági jelenségek elemzése során gyakran nincs módunk a teljes alapsokaság felmérésére. Ilyenkor a vegyes kapcsolat vizsgálatát minta alapján kell elvégeznünk, ami az alapsokaságban meglevő kapcsolat szorosságának mérését nem teszi lehetővé, mindössze a kapcsolat szignifikáns voltának tesztelésére nyílik lehetőségünk. A probléma más módszerek esetén gyakran előkerül, például amikor egy rétegzett mintavétel során meg akarunk győződni arról, vajon megfelelő rétegképző ismérvet választottunk-e; vagy amikor azt vizsgáljuk, hogy egy adott sokaság bizonyos szempontból homogénnek tekinthető-e. A kérdés a gyakorlati életben is számos helyen felmerül, gondoljunk például arra, amikor el kívánjuk dönteni van-e különbség a különböző nemű dolgozók keresetében, vagy eltérőek-e a más-más iskolai végzettségűek fogyasztási szokásai.

A vegyes kapcsolat szignifikáns voltát a **varianciaanalízis**, már korábban megismert módszerével végezzük[12]. Hipotézisrendszerünk:

$$\begin{aligned} &H_0 : \forall\, \tau_j = 0 \ (\ \tau_j = 0 \text{ minden } j\text{ - re }) \\ &H_1 : \exists\, j,\ \tau_j \neq 0 \ (\text{ létezik } j,\text{ amelyre } \tau_j \neq 0\ ) \end{aligned} \qquad (6\text{-}16)$$

vagyis a nullhipotézis teljesülése esetén nincs szignifikáns kapcsolat a minőségi- és a mennyiségi ismérv között. Ismeretes, hogy a varianciaanalízis végrehajtásának előfeltétele, hogy a részsokaságok eloszlása normális, szórásuk pedig egyenlő legyen.

A feltételek teljesülése esetén, amennyiben a nullhipotézis igaz, az alábbi próbafüggvény F-eloszlást követ, (m-1) és (n-m) szabadságfokkal:

$$F = \frac{\dfrac{\sum_{j=1}^{m} n_j\left(\bar{x}_j - \bar{x}\right)^2}{m-1}}{\dfrac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2}{n-m}} \qquad (6\text{-}17)$$

Az F-próba elvégzéséhez szükséges értékeket foglaljuk össze az ún. varianciaanalízis-táblában:

---

[12] A módszert a kettőnél több sokasági várható érték egyezőségének tesztelése során vezettük be (lásd! TK. I.). Jelen problémára való alkalmazása - bár némiképp más szemléletből történik - semmiféle módsz... ni újdonságot nem tartalmaz.

Varianciaanalízis-tábla kétváltozós vegyes kapcsolat esetén

6-5.tábla

A szóródás oka	Szabadság-fok (szf)	Eltérésnégyzetösszeg (SS)	Átlagos eltérés-négyzetösszeg (MSS)
Minőségi ismérv	m-1	$S_K = \sum_{j=1}^{m} n_j(\bar{x}_j - \bar{x})^2$	$\frac{S_K}{m-1}$
Véletlen	n-m	$S_B = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$	$\frac{S_B}{n-m}$
Összesen	n-1	$S = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2$	$\frac{S}{n-1}$

Tekintsük a következő példát!

Egy a felsőoktatási intézmények hallgatóinak anyagi helyzetét célzó vizsgálat során 38 véletlenszerűen kiválasztott hallgatót kérdeztünk meg lakáshelyzetéről és havi átlagos kiadásáról. A válaszok alapján a havi kiadások a következőképpen csoportosithatók (Ft/hó):

Hallgató	Lakáshelyzet		
	Szülőknél	Kollégiumban	Albérletben
1.	1.600	1.500	2.000
2.	1.800	1.900	2.500
3.	1.800	2.000	3.500
4.	2.000	2.000	3.500
5.	2.000	2.200	4.000
6.	2.000	2.300	5.000
7.	2.000	2.500	6.000
8.	2.500	2.500	6.500
9.	2.900	2.500	
10.	3.800	2.800	
11.	4.000	3.000	
12.	5.700	3.000	
13.		3.000	
14.		3.500	
15.		4.000	
16.		4.200	
17.		4.200	
18.		4.500	
ÁTLAGOS KIADÁS	2.675	2.867	4.125

Szignifikánsnak tekinthető-e - 5 %-os szignifikancia-szinten - a kapcsolat a hallgatók lakáskörülményei és havi kiadásai között!

Alkalmazzuk a varianciaanalízis módszerét:

$$\bar{x} = \frac{12 \times 2.675 + 18 \times 2.867 + 8 \times 4.125}{12 + 18 + 8} = 3.071$$

$$F = \frac{\dfrac{12(2.675 - 3.071)^2 + \ldots + 8(4.125 - 3.071)^2}{3 - 1}}{\dfrac{(1.600 - 2.675)^2 + \ldots + (6.500 - 4.125)^2}{38 - 3}} = 3{,}38$$

A 2;35 szabadságfokú F-eloszlás kritikus értéke 5 %-os szignifikancia-szinten 3,27, azaz a nullhipotézist el kell vetnünk. A próba alapján - adott szinten - a hallgatók lakáskörülményei és kiadásai közötti összefüggést szignifikánsnak tekintjük.

Láthattuk, hogy az egy minőségi ismérvet tartalmazó vegyes kapcsolat tesztelése - bizonyos feltételek teljesülése esetén -visszavezethető a több sokasági várható érték egyezőségét ellenőrző varianciaanalízisre. Kézenfekvő, hogy a vegyes kapcsolat jelentős voltának tesztelését akkor is a sokasági várható értékek egyezőségének vizsgálatánál bemutatott módszerekkel végezzük, ha a feltételek nem teljesülnek. Így ha:

- a csoportokon belüli szórások nem egyeznek meg, akkor az ún. variancia-stabilizáló transzformációt[13];
- a csoportokon belüli eloszlás nem normális, akkor a **KRUSKAL-WALLIS** tesztet alkalmazzuk.

6.3.2 A több minőségi ismérvet tartalmazó vegyes kapcsolat elemzése

A több minőségi (csoportosító) ismérv és egy mennyiségi (eredmény) változó közti kapcsolat a következő tábla alapján vizsgálható:

[13] Lásd! Irodalom [45] 158-59 old.

A többváltozós vegyes kapcsolat adatbázisa

6-6.tábla

B_1			B_2			...	B_k		
A_1	...	A_m	A_1	...	A_m	...	A_1	...	A_m
x_{111}	...	x_{1m1}	x_{112}	...	x_{1m2}	...	x_{11k}	...	x_{1mk}
x_{211}		x_{2m1}	x_{212}		x_{2m2}		x_{21k}		x_{2mk}
.		.	.		.		.		.
.		.	.		.		.		.
.		.	.		.		.		.
x_{n11}		x_{nm1}	x_{n12}		x_{nm2}		x_{n1k}		x_{nmk}

A többváltozós vegyes kapcsolat hatékony elemzésének előfeltétele, hogy a minőségi ismérvek változatainak száma áttekinthetően kicsi legyen; valamint hogy az osztályozás során képződő csoportok elemszáma ne legyen nagyon csekély. (A gyakorlatban általában megköveteljük, hogy minden csoport legalább 5 megfigyelést tartalmazzon.) Látható, hogy csak olyan sokaságokkal foglalkozunk, melyekben az egyedi csoportok elemszáma azonos (n). A korlátozások természetesen feloldhatók, ám ennek bemutatása meghaladná tankönyvünk kereteit.

Két minőségi ismérv (magyarázó változó) esetén a (6-14) modell viszonylag egyszerűen kiterjeszthető:

$$x_{ijv} = \mu + \tau_j + \rho_v + (\tau\rho)_{jv} + \varepsilon_{ijv} \qquad (6\text{-}18)$$

ahol:

x_{ijv} - az "A" változó j-edik és a "B" változó v-edik ismérvváltozatához tartozó i-edik megfigyelés értéke;

μ - a sokasági várható érték;

τ_j - az "A" változó j-edik változatának csoporthatása;

ρ_v - a "B" változó v-edik változatának csoporthatása;

$(\tau\rho)_{jv}$ - a két minőségi ismérv "közös" hatása;

ε_{ijv} - a véletlen hatás .

Látható, hogy a minőségi ismérvek egyedi hatása mellett számszerűsítjük azt a hatást is, mely a két csoportosító ismérv együttes fellépésének tudható be. Ezt a hatást a továbbiakban "közös" hatásnak (interakciónak) nevezzük.

A modellre érvényesek az alábbi paramétermegszorítások:

$$\sum_{j=1}^{m}\tau_j = 0\ ;\ \sum_{v=1}^{k}\rho_v = 0$$

$$\sum_{j=1}^{m}(\tau\rho)_{jv} = 0\ (\ v=1,2,\ldots,k\ )\ ;\ \sum_{v=1}^{k}(\tau\rho)_{jv} = 0\ (\ j=1,2,\ldots,m\ )$$

Könnyen átlátható, hogy a vegyes kapcsolat intenzitása (illetve a modellben szereplő csoportosító változók magyarázó ereje) annál nagyobb, minél kisebb hányad tudható be x_{ijv} varianciájából a véletlennek (ε_{ijv}). Kézenfekvő, hogy alapsokaságból történő következtetés esetén a minőségi ismérvek és a mennyiségi jellemző közötti kapcsolat szorosságát ezen varianciákból képzett mutatóval mérjük:

$$H_{XAB} = \sqrt{1-\frac{\sigma_\varepsilon^2}{\sigma_x^2}}$$

A mutató kiszámításához vezessük be a következő jelöléseket:

$$\bar{x}_{jv} = \frac{1}{n}\sum_{i=1}^{n}x_{ijv}$$

$$\overline{x}_{.j.} = \frac{1}{kn}\sum_{v=1}^{k}\sum_{i=1}^{n} x_{ijv} \qquad \overline{x}_{..v} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} x_{ijv}$$

$$\overline{x}_{...} = \frac{1}{kmn}\sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n} x_{ijv}$$

Ekkor az (6-18) modell, némi átalakítás után felírható a következőképpen:

$$\varepsilon_{ijv} = x_{ijv} - \left[\overline{x}_{...} + \left(\overline{x}_{.j.} - \overline{x}_{...}\right) + \left(\overline{x}_{..v} - \overline{x}_{...}\right) + \left(\overline{x}_{.jv} - \overline{x}_{.j.} - \overline{x}_{..v} + \overline{x}_{...}\right)\right]$$

$$\varepsilon_{ijv} = x_{ijv} - \overline{x}_{.jv}$$

Az eltérésnégyzetösszegekre érvényes egyenlőség:

$$\sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \overline{x}_{...}\right)^2 = \sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\overline{x}_{.j.} - \overline{x}_{...}\right)^2 + \sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\overline{x}_{..v} - \overline{x}_{...}\right)^2 +$$
$$+\sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\overline{x}_{.jv} - \overline{x}_{.j.} - \overline{x}_{..v} + \overline{x}_{...}\right)^2 + \sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \overline{x}_{.jv}\right)^2$$

vagyis

$$\sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \overline{x}_{...}\right)^2 = kn\sum_{j=1}^{m}\left(\overline{x}_{.j.} - \overline{x}_{...}\right)^2 + mn\sum_{v=1}^{k}\left(\overline{x}_{..v} - \overline{x}_{...}\right)^2 +$$
$$+n\sum_{j=1}^{m}\sum_{v=1}^{k}\left(\overline{x}_{.jv} - \overline{x}_{.j.} - \overline{x}_{..v} + \overline{x}_{...}\right)^2 + \sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \overline{x}_{.jv}\right)^2$$

A fenti összefüggés alapján meghatározható a többszörös kapcsolatszorossági mutató, amely a két csoportosító ismérvet tartalmazó vegyes kapcsolat intenzitását számszerűsíti:

$$H_{XAB} = \sqrt{1 - \frac{\sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \bar{x}_{.jv}\right)^2}{\sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \bar{x}_{...}\right)^2}} \tag{6-19}$$

Felírhatjuk a parciális mutatók képletét is[14]:

$$H_{XA.B} = \sqrt{\frac{kn\sum_{j=1}^{m}\left(\bar{x}_{.j.} - \bar{x}_{...}\right)^2}{\sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \bar{x}_{...}\right)^2}}$$

$$H_{XB.A} = \sqrt{\frac{mn\sum_{v=1}^{k}\left(\bar{x}_{..v} - \bar{x}_{...}\right)^2}{\sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \bar{x}_{...}\right)^2}} \tag{6-20}$$

A mutatók négyzetei magyarázó erőként, százalékos formában értelmezhetők.

Részsokaság (minta) alapján végzett elemzés során megvizsgálhatjuk, vajon szignifikánsnak tekinthető-e a minőségi ismérvek és a mennyiségi ismérv közötti kapcsolat. A kapcsolat meglétét a **kétutas**, más néven **kétfaktorú varianciaanalízis** módszerével tesztelhetjük. Akárcsak az egyutas varianciaanalízisnél, a módszer

---

[14] Vegyük észre, hogy mivel a minőségi ismérvek általában nominális skálán mértek, így szórásuk nem értelmezhető. Ebből adódóan a parciális mutatók megegyeznek a kétváltozós totális mutatókkal.

alkalmazásának előfeltétele, hogy az egyes csoportokon belül a mennyiségi változó eloszlása normális legyen, valamint a csoportok ismeretlen szórása legyen egyenlő[15].

Az eltérésnégyzetösszegekre érvényes összefüggés alapján felírható a kétutas varianciaanalízis táblája, amelyben a teljes eltérésnégyzetösszeget kiváltó okai alapján bontottuk fel[16]:

Kétutas variancianalízis-tábla

6-7. tábla

Ok	szf	SS	MSS
A	m - 1	$S_A = kn\sum_{j=1}^{m}\left(\overline{x}_{.j.} - \overline{x}_{...}\right)^2$	$S_A/(m-1)$
B	k - 1	$S_B = mn\sum_{v=1}^{k}\left(\overline{x}_{..v} - \overline{x}_{...}\right)^2$	$S_B/(k-1)$
AB	(m-1)(k-1)	$S_{AB} = n\sum_{v=1}^{k}\sum_{j=1}^{m}\left(\overline{x}_{.jv} - \overline{x}_{.j.} - \overline{x}_{..v} + \overline{x}_{...}\right)^2$	$S_{AB}/((m-1)(k-1))$
E	mk(n-1)	$S_E = \sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ijv} - \overline{x}_{.jv}\right)^2$	$S_E/(mk(n-1))$
Σ	nmk - 1	$S = \sum_{v=1}^{k}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ijv} - \overline{x}_{...}\right)^2$	$S/(nmk-1)$

Belátható, hogy az átlagos eltérésnégyzetösszegek várható értéke a következő:

---

15 Amennyiben a feltételek nem teljesülnek, más eljárást kell alkalmazni (pl. nem normális eloszlású véletlen hatás esetén **FRIEDMANN**-próbát). Ezen eljárások leírása megtalálható pl. Irodalom [45]-ben.

16 A tábla fejrovatában alkalmazott rövidítések megegyeznek a 6-5. táblában alkalmazottakkal.

$$E(MSS) = \sigma^2$$

$$E(MSS_A) = \sigma^2 + kn\frac{\sum_{j=1}^{m}\tau_j^2}{m-1}$$

$$E(MSS_B) = \sigma^2 + mn\frac{\sum_{v=1}^{k}\rho_v^2}{k-1}$$ vagyis az átlagos eltérésnégyzetösszegek

$$E(MSS_{AB}) = \sigma^2 + n\frac{\sum_{v=1}^{k}\sum_{j=1}^{m}(\tau\rho)_{jv}^2}{(k-1)(m-1)}$$

$$E(MSS_E) = \sigma^2$$

torzitatlan becslései a sokasági varianciának, ha teljesülnek a következő nullhipotézisek:

$$H_{0_1}: \tau_1 = \ldots = \tau_m = 0 \qquad (6\text{-}21.1)$$

$$H_{0_2}: \rho_1 = \ldots = \rho_k = 0 \qquad (6\text{-}21.2)$$

$$H_{0_3}: (\tau\rho)_{11} = \ldots = (\tau\rho)_{mk} = 0 \qquad (6\text{-}21.3)$$

A (6-21.1), (6-21.2) és (6-21.3) nullhipotézisek tesztelése az 6-7. táblában szereplő mennyiségek alapján, a becsült szórásnégyzetek összehasonlítását célzó F-próbával végezhető.

A próbafüggvények és az F-eloszlás szabadságfokai rendre:

$$F = \frac{MSS_A}{MSS_E}; \qquad \nu_1 = m-1, \qquad \nu_2 = mk(n-1) \qquad (6\text{-}22.1)$$

$$F = \frac{MSS_B}{MSS_E} \qquad \nu_1 = k-1 \qquad \nu_2 = mk(n-1) \qquad (6\text{-}22.2)$$

$$F = \frac{MSS_{AB}}{MSS_E} \qquad \nu_1 = (m-1)(k- \quad \nu_2 = mk(n-1) \qquad (6\text{-}22.3)$$

A két csoportosító (magyarázó) változót tartalmazó vegyes kapcsolat elemzésére tekintsük az alábbi példát!

Sportorvosok tesztelni kívánják azt a feltevést, hogy az élsportolók testalkatában eltérés mutatható ki, az általuk űzött sportág függvényében. Ennek ellenőrzésére 18 férfi és 18 női sportoló testmagasságát mérték meg. Eredményeik (testmagasság cm-ben):

Sportoló	Férfi			Nő		
	Atléta	Tornász	Úszó	Atléta	Tornász	Úszó
1.	179	168	184	164	152	164
2.	180	172	186	166	153	164
3.	185	175	196	168	158	166
4.	185	175	202	170	158	166
5.	190	176	204	172	160	172
6.	195	178	204	174	162	175

Feltételezhető, hogy az egyes csoportokon belül a magasságok eloszlása normális, és csoportokon belüli szórása azonos. Elemezzük a sportolók neme (A) és sportága (B), valamint a magassága közötti kapcsolatot!

Határozzuk meg a kétutas varianciaanalízis modelljében szereplő mennyiségeket:

$$\bar{x}_{11} = \frac{1}{6}(179+...+195) = 185{,}66 \quad \bar{x}_{12} = \frac{1}{6}(164+...+174) = 169{,}00$$

$$\bar{x}_{21} = \frac{1}{6}(168+...+178) = 174{,}00 \quad \bar{x}_{22} = \frac{1}{6}(152+...+162) = 157{,}17$$

$$\bar{x}_{31} = \frac{1}{6}(184+...+204) = 196{,}00 \quad \bar{x}_{32} = \frac{1}{6}(164+...+175) = 167{,}83$$

$$\bar{x}_{1.} = \frac{1}{12}(179+...+174) = 177{,}33 \quad \text{(atléták átlagos magassága)}$$

$$\bar{x}_{2.} = \frac{1}{12}(168+...+162) = 165{,}58 \quad \text{(tornászok átlagos magassága)}$$

$$\bar{x}_{3.} = \frac{1}{12}(184+...+175) = 181{,}92 \quad \text{(úszók átlagos magassága)}$$

$$\bar{x}_{.1} = \frac{1}{18}(179+...+204) = 185{,}22 \quad \text{(férfi sportolók átlagos magassága)}$$

$$\bar{x}_{.2} = \frac{1}{18}(164+...+175) = 164{,}66 \quad \text{(női sportolók átlagos magassága)}$$

$$\bar{x}_{..} = \frac{1}{36}(179+...+175) = 174{,}94 \quad \text{(átlagmagasság)}$$

$$S_B = 12\left[(177{,}33 - 174{,}94)^2 + ... + (181{,}92 - 174{,}94)^2\right] = 1.703{,}4$$

$$S_A = 18\left[(185{,}22 - 174{,}94)^2 + ... + (164{,}66 - 174{,}94)^2\right] = 3.802{,}9$$

$$S_{AB} = 6\begin{bmatrix}(185{,}66 - 177{,}33 - 185{,}22 + 174{,}94)^2 + ... + \\ +(167{,}83 - 181{,}92 - 164{,}66 + 174{,}94)^2\end{bmatrix} = 260{,}7$$

$$S_E = (179 - 185{,}56)^2 + ... + (175 - 167{,}83)^2 = 904{,}9$$

$$S = (179 - 174{,}94)^2 + ... + (175 - 174{,}94)^2 = 6.671{,}9$$

A fentiek alapján meghatározhatók a kapcsolatszorossági mérőszámok:

$$H_{XAB} = \sqrt{1 - \frac{904,9}{6.671,9}} = 0,9297 \quad H^2_{XAB} = 0,8644$$

$$H_{XB.A} = \sqrt{\frac{1.703,4}{6.671,9}} = 0,5053 \quad H^2_{XB.A} = 0,2553$$

$$H_{XA.B} = \sqrt{\frac{3.802,9}{6.671,9}} = 0,7550 \quad H^2_{XA.B} = 0,5700$$

Elmondható tehát, hogy a minőségi ismérvek és a mennyiségi ismérv között a kapcsolat szoros; a két csoportosító ismérv az eredményváltozó szóródásának mintegy 86 %-át határozza meg. A sportágak és a magasság közötti sztochasztikus kapcsolat - eltekintve a nemek hatásától - közepes intenzitású. Az élsportolók nemek szerinti hovatartozása testmagasságuk szóródásának 57 %-át magyarázza meg, ha kiszűrjük a sportágak hatását.

A kapcsolatszorossági mutatók számszerűsítése után teszteljük az ismérvek közötti kapcsolatok szignifikáns voltát:

$$F_B = \frac{1.703,4/2}{904,9/30} = 28,23$$

$$F_A = \frac{3.802,9/1}{904,9/30} = 126,08$$

$$F_{AB} = \frac{260,7/2}{904,9/30} = 4,32$$

Mivel 5 %-os szignifikancia-szinten az F-eloszlás kritikus értéke ($F_{1;30} = 4,17$ $F_{2;30} = 3,32$) kisebb az általunk becsültnél így minden esetben el kell vetnünk a nullhipotézist. A nem, a sportág és a két minőségi ismérv "közös" hatása egyaránt szignifikáns kapcsolatban áll a sportolók testmagasságával.

A két magyarázó változót tartalmazó kapcsolat bemutatása után **végezzük el a modell általánosítását**! Könnyen átlátható, hogy három minőségi ismérv esetén három egyedi hatást, három kétszeres- és egy háromszoros interakciót tartalmaz a modell. Általánosítva egy **k** magyarázó változót (minőségi ismérvet) tartalmazó modellben

k egyedi hatás ;

$\binom{k}{2}$ kétszeres interakció ;

$\binom{k}{3}$ háromszoros interakció ; s.i.t.

$\binom{k}{k}$ vagyis egy k-szoros interakció szerepel.

A többváltozós varianciaanalízis lineáris modellje[17]

$$\mathbf{x}_{ij} = \mathbf{m} + (\mathbf{m}_j - \mathbf{m}) + (\mathbf{x}_{ij} - \mathbf{m}_j)$$

ahol:

$\mathbf{x}_{ij}$ a mennyiségi vektorváltozó i-edik esete a j-edik csoportban ($i=1,\ldots,n_j$; $j=1,\ldots,g$)

$\mathbf{m}$ a várható érték $(k\times 1)$-es vektora

$\mathbf{m}_j$ a j-edik csoport várható értékének $(k\times 1)$-es vektora

[17] Vegyük észre, hogy skaláris változók helyett a továbbiakban vektorváltozókkal dolgozunk!

Képezzük az egyes elemek eltérését a várható értéktől:

$$\left(\mathbf{x}_{ij} - \mathbf{m}\right) = (\mathbf{m}_j - \mathbf{m}) + (\mathbf{x}_{ij} - \mathbf{m}_j)$$

Az eltérések elemeinek keresztszorzatát képezve, majd ezeket összegezve jutunk a többváltozós varianciaanalízis alapegyenletéhez:

$$\sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(\mathbf{x}_{ij} - \mathbf{m}\right)\left(\mathbf{x}_{ij} - \mathbf{m}\right)^T = \sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(\mathbf{m}_j - \mathbf{m}\right)\left(\mathbf{m}_j - \mathbf{m}\right)^T + \sum_{j=1}^{s}\sum_{i=1}^{n_j}\left(\mathbf{x}_{ij} - \mathbf{m}_j\right)\left(\mathbf{x}_{ij} - \mathbf{m}_j\right)^T$$

A szokásos jelölésrendszernek megfelelően felírható a többváltozós varianciaanalízis alapegyenlete mátrix-jelöléssel:

$$\mathbf{T} = \mathbf{K} + \mathbf{B}$$

vagyis a teljes eltérésnégyzetösszeg és keresztszorzat mátrix felbontható két részre: külső eltérésnégyzetösszeg és keresztszorzat mátrixra, illetve belső eltérésnégyzetösszeg és keresztszorzat mátrixra. A mátrixok determinánsainak felhasználásával felírható a minőségi (csoportosító) ismérvek magyarázó ereje[18]:

$$H^2 = \frac{|\mathbf{K}|}{|\mathbf{T}|} \qquad (6\text{-}23)$$

A fenti többváltozós modell alapján tesztelhető az alábbi hipotézisrendszer:

$$H_0: \forall j \quad \mathbf{m}_j = \mathbf{m}$$
$$H_1: \exists j \quad \mathbf{m}_j \neq \mathbf{m}$$

---

[18] Komplementere $\Lambda = \frac{|\mathbf{B}|}{|\mathbf{T}|} = 1 - \frac{|\mathbf{K}|}{|\mathbf{T}|}$ az ún. Wilks-lambda.

A próbafüggvény:

$$F = \frac{1-\Lambda^{\frac{1}{s}}}{\Lambda^{\frac{1}{s}}} \cdot \frac{\nu_2}{\nu_1} \tag{6-24}$$

ahol

$$s = \sqrt{\frac{k^2(g-1)^2-4}{k^2+(g-1)^2-5}}$$

$$\nu_1 = k(g-1)$$

$$\nu_2 = s\left[(n-1)-\frac{k+(g-1)+1}{2}\right]-\frac{k(g-1)-2}{2}$$

a nullhipotézis teljesülése esetén ν_1, ν_2 szabadságfokú F-eloszlást követ.

6.4 A KORRELÁCIÓS KAPCSOLAT ELEMZÉSE

A mennyiségi ismérvek közötti összefüggést **korrelációs kapcsolatnak** nevezzük. Az ezen kapcsolat vizsgálata során alkalmazott eljárások lényegesen különböznek a korábban bemutatott asszociációs és vegyes kapcsolat elemzésére használt módszerektől. Ennek oka, hogy a mennyiségi ismérvek általában intervallum-, illetve arányskálán mérhetők, így esetükben nemcsak az ismérvváltozatok különbözőségét, hanem a különböző változóértékek távolságát, hányadosát is értelmezhetjük. Az ebben rejlő információk tökéletesebb kihasználása a korábbiaktól eltérő módszertani apparátus alkalmazását igényli.

A korrelációs kapcsolat alapvetően a következő két lényeges kérdésben tér el a korábban tárgyalt sztochasztikus kapcsolatoktól:

1) Mivel a kapcsolatban szereplő változók ismérvváltozatainak száma általában igen nagy, ezért az előforduló ismérvváltozat-kombinációk száma lényegesen meghaladja az asszociációs kapcsolatban keletkező kombinációk számát. Ennek következtében az egyes változatpárokhoz tartozó gyakoriságok kicsik (nagyságuk jellemző módon 1). Az előbbiek miatt a korrelációs kapcsolat vizsgálata nem kombinációs tábla, hanem a változópárok (-hármasok stb.) egyes megfigyelésekhez tartozó értékeinek felsorolása alapján történik.

A korrelációs kapcsolat adatbázisa

6-8.tábla

Megfigyelés	Változó				
	Y	X_1	X_2	...	X_k
1.	y_1	x_{11}	x_{21}	...	x_{k1}
2.	y_2	x_{12}	x_{22}	...	x_{k2}
.	.				.
.	.				.
.	.				.
n.	y_n	x_{1n}	x_{2n}	...	x_{kn}

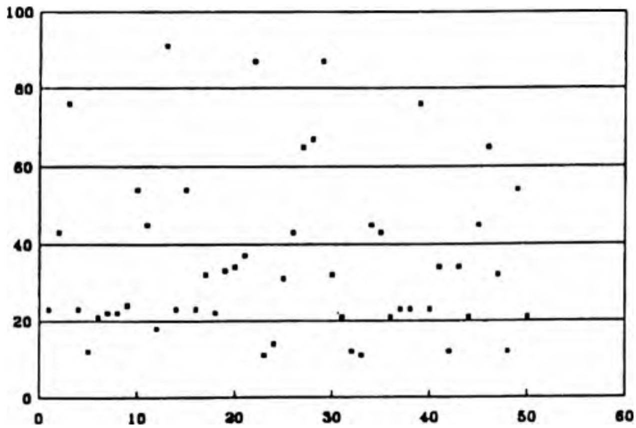
Az adatbázisról jegyezzük meg a következőket

- konvencionálisan a kapcsolat eredményváltozóját jelöljük y-nal, a magyarázó (tényező) változókat pedig x_j-vel (ahol: j=1,2...,k);
- az adatbázis minden változójához azonos (n) számú megfigyelés tartozik;
- a korrelációs kapcsolat megnevezése mindig az összes benne szereplő változó alapján történik, vagyis az y-t és x-t tartalmazó kapcsolat kétváltozós; az y-t, x_1-t és x_2-t tartalmazó háromváltozós, s.i.t. (praktikusan

ez annyit jelent, hogy a kapcsolat mindig eggyel több változós, mint ahány magyarázó változót tartalmaz).

2) A második jelentős eltérés a korábban bemutatott kapcsolatokhoz képest, hogy korreláció esetén a kapcsolat szorossága (intenzitása) és iránya mellett az összefüggés jellegét is értelmezzük. Az összefüggés jellege alatt azt a jellegzetes függvénytípust értjük, amellyel leírható, hogy a magyarázó ismérv(ek) változására hogyan reagál az eredményváltozó.

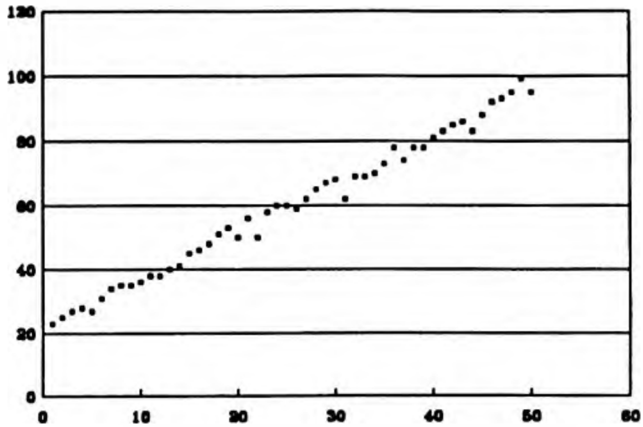
Ennek megvilágítását szolgálják a következő ábrák[19]:



6-1. ábra Függetlenség (kapcsolat hiánya)

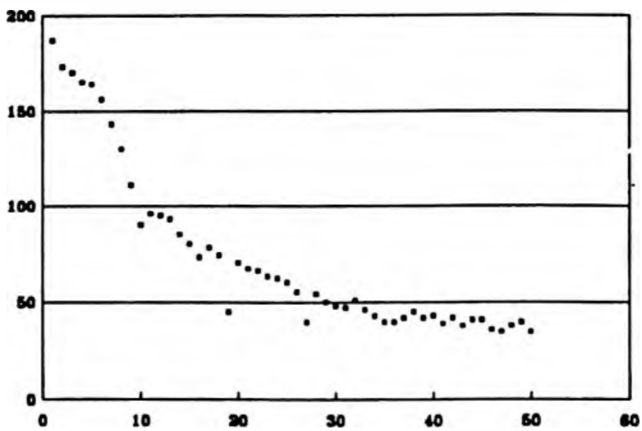
Az 6-1. ábra az ismérvek közötti kapcsolat **hiányát** mutatja. Az X változó különböző értékei semmiféle hatást nem gyakorolnak Y-ra, ennek szóródása csak a véletlennek tudható be.

[19] Ábráink kétváltozós kapcsolatot illusztrálnak, mivel síkban csak ez jeleníthető meg; a különböző kapcsolatjellegek többváltozós korreláció esetére is kiterjeszthetők.



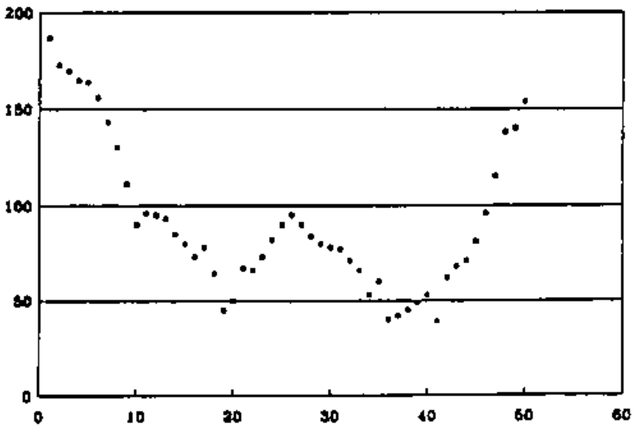
6-2. ábra Lineáris kapcsolat

Az 6-2. ábra **lineáris** kapcsolatot mutat. A kapcsolat jellemzője, hogy a magyarázó változó egységnyi változása az eredményváltozóban az esetek túlnyomó többségében azonos irányú, megközelítően konstans egységnyi változást idéz elő.



6-3. ábra Monoton kapcsolat

Az 6-3. ábra **monoton** összefüggést mutat X és Y között. Ennek lényege, hogy a kapcsolat iránya egyértelmű, de Y változásának mértéke X egy egységnyi elmozdulása esetén nem azonos. Vegyük észre, hogy a monoton kapcsolatok magukban foglalják a lineáris kapcsolatokat. Ennek következtében a monoton jellegű összefüggésekre kidolgozott eljárások elméletileg alkalmazhatók lineáris összefüggések esetén is, de ez fordítva nem érvényes.



6-4. ábra Határozatlan irányú kapcsolat

Az 6-4. ábra **határozatlan irányú** kapcsolatot mutat X és Y között. Látható, hogy nem függetlenségről van szó, hiszen az eredményváltozó alakulása x függvényében jól leírható (esetünkben polinom függvénnyel), de a kapcsolat iránya változó. Ebben az esetben tehát olyan mutatószámot kell alkalmaznunk, amely nem érzékeny a kapcsolat irányának változására.

A különböző jellegű kapcsolatok bemutatása után felmerülhet a kérdés, hogyan tudjuk a vizsgálat során eldönteni, hogy milyen típusú kapcsolattal állunk szemben. Erre alapvetően három megoldás kínálkozik:

- dönthetünk előzetes információ alapján (szakmai ismeret, korábbi vizsgálat eredménye stb.) ;
- ábrázolhatjuk a változókat és dönthetünk az ábra alapján;

- dönthetünk regresszióanalízis[20] alapján .

A továbbiakban azokat a módszertani eszközöket mutatjuk be, amelyekkel a különböző jellegű kapcsolatok vizsgálata elvégezhető. Az eljárások ismertetése után röviden kitérünk azokra a problémákra, melyeket a helytelen eszközválasztás idéz elő.

6.4.1 A kétváltozós korrelációs kapcsolat vizsgálata

A társadalmi-gazdasági jelenségek vizsgálata során gyakran merül fel annak igénye, hogy két mérhető (mennyiségi) ismérv közötti kapcsolat szorosságát vizsgáljuk, illetve megállapítsuk, van-e egyáltalán jelentős kapcsolat az adott ismérvek között. A tudomány számos területéről hozhatunk példákat ezen kérdésfeltevésre, ezek közül csupán néhány:

- az antropometria régi kérdése: van-e kapcsolat a testmagasság és a testsúly között;
- gyakran felmerülő kérdés, hogyan hat az iskolázottság (tanulással eltöltött idő; elvégzett osztályok száma) az intelligencia-hányadosra;
- a makroökonómia örök kérdése: milyen az összefüggés az inflációs- és a munkanélküliségi-ráta között;
- vállalatgazdaságtani elemzések sora kísérelte meg kideríteni, van-e összefüggés a kifizetett bérek és a nyereség között .

A fenti kérdések bármelyikének megválaszolása egy kétváltozós korrelációs kapcsolat elemzését igényli.

[20] A regresszióanalízissel a tankönyv következő fejezetében foglalkozunk.

A különböző jellegű kapcsolatok különböző mutatókkal írhatók le. Tananyagunkban a következő kapcsolatszorossági mérőszámokkal elemezzük a korrelációs kapcsolatot:

- lineáris kapcsolat esetén a **lineáris korrelációs együtthatóval**;
- monoton kapcsolat esetén a **Spearman-féle rangkorrelációs együtthatóval**;
- határozatlan irányú kapcsolat esetén a **korrelációs hányadossal**.

Lineáris korrelációs együttható

A lineáris korrelációs kapcsolat mérése a **kovariancia** mutatójára épül. Ismeretes[21], hogy a kovariancia elsőrendű vegyes, centrális momentum, melynek képlete:

$$C_{YX} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (6\text{-}25)$$

A kovariancia mutatója jól hasznosítható a mennyiségi ismérvek közötti kapcsolat irányának megállapítása során, hiszen a mutató értéke pozitív, ha az egyes változók saját átlaguktól való eltérésének iránya megegyezik; és negatív fordított esetben. Vagyis, ha minden i-re $x_i > \bar{x}$ és $y_i > \bar{y}$, vagy $x_i < \bar{x}$ és $y_i < \bar{y}$, akkor $(x_i - \bar{x})(y_i - \bar{y}) > 0$, azaz $C_{YX} > 0$. Illetve, ha minden i-re $x_i > \bar{x}$ és $y_i < \bar{y}$, vagy $x_i < \bar{x}$ és $y_i > \bar{y}$, akkor $(x_i - \bar{x})(y_i - \bar{y}) < 0$, azaz $C_{YX} < 0$.

[21] Lásd! TK. I.

A kovariancia mutatójáról belátható, hogy abszolút értéke nem lehet nagyobb a két változó szórásának szorzatánál, így ezzel normálva a mutatót származtatjuk a lineáris korrelációs együtthatót:

$$r_{YX} = \frac{C_{YX}}{\sigma_Y \sigma_X} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} \qquad (6\text{-}26)$$

Az így meghatározott korrelációs együttható abszolút értéke 0 és 1 között található, előjele a kapcsolat irányára utal, nagyobb abszolút értéke szorosabb kapcsolatot jelez.

A mutató - néhány kézenfekvő átalakítás után - más formában is felírható:

$$r_{YX} = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{yx}}{\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

A korrelációs együttható szimmetrikus, azaz ugyanolyan erősségű kapcsolatot mutat X magyarázó- és Y eredményváltozó között, mint fordítva. A mutató négyzetét - akárcsak korábban - magyarázó erőként értelmezzük, és **determinációs együtthatónak** nevezzük.

A mennyiségi ismérvek közötti kapcsolat vizsgálata során gyakran találkozunk azzal a problémával, hogy nem nyílik alkalmunk a teljes alapsokaság felmérésére, és így minta alapján kell elemzésünket elvégezni. A kapcsolat vizsgálata ekkor elvégezhető

- vagy a lineáris korrelációs együttható becsült értékének felhasználásával ;
- vagy a függetlenségre vonatkozó hipotézis tesztelésével .

A lineáris korrelációs együtthatóról belátható, hogy amennyiben mindkét változó eloszlása normális, úgy mintából számítva torzítatlan becslése az elméleti együtthatónak. A becslőfüggvény kisminta esetén **n-2 szabadságfokú t-eloszlást**, nagyminta esetén **normális eloszlást** követ. A becsült együttható standard hibája:

$$s_{\hat{r}_{YX}} = \sqrt{\frac{1-\hat{r}_{YX}^2}{n-2}} \tag{6-27}$$

A fentiek alapján, kismintából való következtetés esetén a mutató köré (1-α) megbízhatósági szinten konfidencia-intervallum szerkeszthető:

$$\Pr\left\{\hat{r}_{YX} + t_{\frac{\alpha}{2}}\sqrt{\frac{1-\hat{r}_{YX}^2}{n-2}} < r_{YX} < \hat{r}_{YX} + t_{1-\frac{\alpha}{2}}\sqrt{\frac{1-\hat{r}_{YX}^2}{n-2}}\right\} = 1-\alpha$$

Amennyiben a kapcsolat szorosságának konkrét számszerűsítése helyett pusztán csak a kapcsolat létéről kívánunk dönteni, adott szignifikancia-szinten ellenőrizhetjük, vajon van-e kapcsolat a két ismérv között. Hipotézisrendszerünk:

$$\begin{aligned} H_0 &: r_{YX} = 0 \\ H_1 &: r_{YX} \neq 0 \end{aligned} \tag{6-28}$$

A nullhipotézis értelmében a két ismérv független, ennek elvetése a korrelációs kapcsolat szignifikáns voltát igazolja[22]. A becsült korrelációs együtthatóra épülő próbafüggvényünk:

$$t = \frac{\hat{r}_{YX}\sqrt{n-2}}{\sqrt{1-\hat{r}_{YX}^2}} \qquad (6\text{-}29)$$

a nullhipotézis teljesülése esetén **(n-2) szabadságfokú t-eloszlást** követ.

A kétváltozós lineáris korrelációs kapcsolat elemzésének bemutatására tekintsük a következő példát:

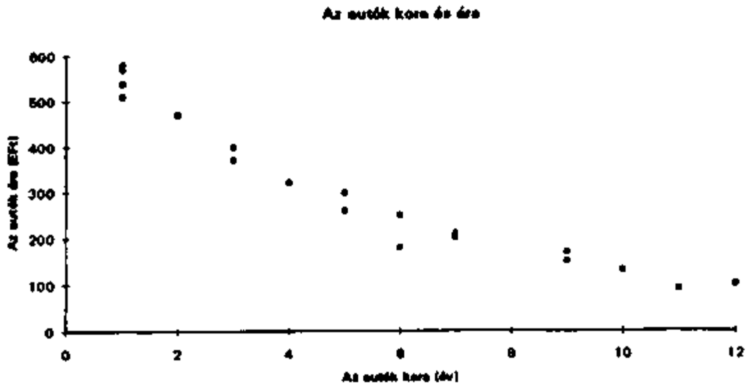
Egy használtautó-kereskedés kínálatából választottunk ki véletlenszerűen 20 azonos típusú személygépkocsit, hogy megvizsgáljuk, milyen szoros az összefüggés az autók életkora és ára között. Adataink:

Sorszám	Életkor (év)	Ár (EFt)	Sorszám	Életkor (év)	Ár (EFt)
1.	3	400	11.	12	100
2.	10	130	12.	7	200
3.	7	200	13.	6	250
4.	6	180	14.	5	260
5.	3	370	15.	7	210
6.	1	540	16.	4	320
7.	1	510	17.	1	570
8.	1	580	18.	9	170
9.	2	470	19.	11	90
10.	5	300	20.	9	150

Elemezzük a kapcsolatot a gépkocsik életkora és ára között!

[22] Vegyük észre, hogy mivel a mutató értéke pozitív és negatív egyaránt lehet, ezért alternatív hipotézisünkben az r≠0 hipotézis szerepel, vagyis kétoldalú próbát kell végeznünk!

A kapcsolat elemzésének kezdetén megállapíthatjuk, hogy esetünkben az autók kora az ok (magyarázó változó), és áruk az okozat (eredmény) szerepét betöltő változó. Ábrázoljuk a változókat, annak megállapítása érdekében, hogy milyen jellegű a kapcsolat!



6-5. ábra Összefüggés az autók kora és ára között

Ábránkból leolvasható, hogy az összefüggés jellege jól közelíthető lineárissal, ezért határozzuk meg a lineáris korrelációs együtthatót:

$$\bar{x} = 5{,}5 \quad s_x = 3{,}41 \quad \bar{y} = 300 \quad s_y = 157{,}51$$

$$r_{YX} = \frac{(3-5{,}5)(400-300)+\ldots+(9-5{,}5)(150-300)}{20 \times 157{,}51 \times 3{,}41} = -0{,}9533$$

A mutató alapján megállapítható, hogy a kapcsolat az autók kora és ára között szoros; az autók kora az ár szóródásának 90,88 %-át magyarázza meg. Az erős negatív irányú kapcsolat gyakorlatilag úgy értelmezhető, hogy a gépkocsik életkorának növekedése az vételáruk erőteljes csökkenését idézi elő. Mivel adatbázisunkat mintának kell tekinteni, így felmerülhet, hogy az általunk talált kapcsolat csak a véletlennek tudható be, ezért szükség van az ismérvek függetlenségére vonatkozó hipotézis tesztelésére is. A próbafüggvény:

$$t_{emp} = \frac{-0{,}9533\sqrt{20-2}}{\sqrt{1-(-0{,}9533)^2}} = -13{,}39$$

A 18 szabadságfokú t-eloszlás kritikus értékei, 5 %-os szignifikancia-szinten: ±2,10; vagyis a nullhipotézist el kell vetni: az ismérvek között jelentős korrelációs kapcsolat mutatható ki.

Belátható, hogy a lineáris korrelációs együttható érzékenyen reagál a kiugró értékekre, illetve a lineáristól eltérő függvényformára, ezért csak megfelelő körültekintéssel alkalmazható.

Rangkorrelációs együttható

Ha a mennyiségi ismérvek közötti kapcsolat nem egyenes vonalú, a lineáris korrelációs együttható a kapcsolat szorosságának torzított mérőszáma. Ha a két változó közötti kapcsolat monoton, akkor a kapcsolat szorosságának bizonyíthatóan robusztusabb mérőszáma a **Spearman-féle rangkorrelációs együtható**[23]. A mutató - a tényleges változóértékek helyett - a rangsorba rendezett változók rangszámaira alapoz (az intervallum-, vagy az arányskála helyett ordinálisat használ). Ennek következtében a mérőszám nem számol a tényleges értékekkel, vagyis nem használja ki az adatbázisban rejlő összes információt. Írjuk fel a rangszámokra érvényes lineáris korrelációs együtthatót:

[23] A robusztusságnak természetesen "ára" van, vagyis a mutató kevésbé hatékony (ARE=0.955). Ez azt jelenti, hogy lineáris kapcsolat esetén feltétlenül a lineáris korrelációs együtthatót kell alkalmazni.

$$\rho_{YX} = \frac{\sum_{i=1}^{n}\left[R(y_i) - \overline{R}(y_i)\right]\left[R(x_i) - \overline{R}(x_i)\right]}{\sqrt{\sum_{i=1}^{n}\left[R(x_i) - \overline{R}(x_i)\right]^2}\sqrt{\sum_{i=1}^{n}\left[R(y_i) - \overline{R}(y_i)\right]^2}}$$

ahol $R(y_i)$ és $R(x_i)$ a rangsorba rendezett változóértékek rangszámát jelöli. A mutató a következő egyszerűbb alakra hozható[24]:

$$\rho_{YX} = 1 - \frac{6\sum_{i=1}^{n}\left[R(y_i) - R(x_i)\right]^2}{n\left(n^2 - 1\right)} \qquad (6\text{-}30)$$

Mivel a mutató a rangszámokból képzett lineáris korrelációs együttható, így az összes r-re vonatkozó tulajdonság ρ-ra is érvényes. Ennek megfelelően a rangkorrelációs együttható abszolút értéke a [0;1] zárt intervallumban található. A mutató szimmetrikus; vagyis a két változó közötti kapcsolatot azonos intenzitásúnak méri, bármelyik legyen is az eredményváltozó. Szintén megegyezik a lineáris kapcsolatnál bemutatottal a kapcsolat szignifikáns voltának tesztelése. A rangkorrelációs együttható esetében is t-próbát alkalmazunk az alábbi hipotézisrendszer tesztelésére.

$$H_0 : \rho_{YX} = 0$$

$$H_1 : \rho_{YX} \neq 0$$

[24] Bizonyítás lásd! pl. Irodalom [45].

A próbafüggvény:

$$t = \frac{\rho_{YX}\sqrt{n-2}}{\sqrt{1-\rho_{YX}^2}} \qquad (6\text{-}31)$$

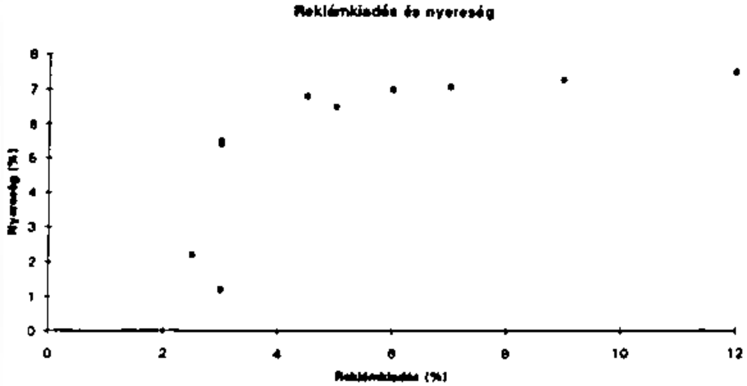
(n-2) szabadságfokú t-eloszlást követ.

10 kereskedelmi vállalat reklámkiadása (az összes kiadás százalékában) és vagyonarányos nyeresége közötti kapcsolatot vizsgáljuk. Adataink (%):

Vállalat	Reklámkiadás (x)	Nyereség (y)
1.	2,5	2,2
2.	3,0	5,4
3.	3,0	1,2
4.	3,0	5,5
5.	4,5	6,8
6.	5,0	6,5
7.	6,0	7,0
8.	7,0	7,1
9.	9,0	7,3
10.	12,0	7,5

Vizsgáljuk meg, milyen szoros a korrelációs kapcsolat a két változó között!

Ábrázoljuk a két változót:



6-6. ábra Összefüggés a reklámkiadás és a nyereség között

Az ábra alapján szembeötlő, hogy bár a magasabb reklámkiadáshoz az esetek többségében magasabb nyereség tartozik, semmiképpen sem tekinthető a kapcsolat lineárisnak. Számszerűsítsük a becsült rangkorrelációs együtthatót:

$$R(x_1)=1,\ R(x_2)=3,\ldots,R(x_{10})=10$$

$$R(y_1)=2,\ R(y_2)=3,\ldots,R(y_{10})=10$$

$$\rho_{YX}=1-\frac{6\left[(1-2)^2+(3-3)^2+\ldots+(10-10)^2\right]}{10\left(10^2-1\right)}=0{,}9515$$

A kapcsolat a reklámkiadások és a vagyonarányos nyereség között a mintában szoros pozitív irányú; a magyarázó változó a nyereség szóródásának 90,5 %-át határozza meg.

Elképzelhető az az eset is, amelyben a mennyiségi ismérv egy, vagy több változatához 1-nél nagyobb gyakoriság tartozik. Ilyenkor a rangsorba rendezés nem oldható meg egyértelműen, mivel több megfigyeléshez tartozna azonos rangszám. A bevett gyakorlat szerint ebben az esetben az ilyen megfigyelésekhez a rangszámok átlagát rendeljük.

A Spearman-féle rangkorrelációs mérőszámról meg kell említenünk, hogy alkalmazható minden olyan esetben, amikor a vizsgálatban szereplő ismérvek nehezen mérhetők (pl. képesség, tanulmányi eredmény stb.) A mutató tehát gyakran használatos eleve ordinális skálán mért ismérvek közötti kapcsolat vizsgálatában.

Korrelációs hányados

A korrelációs hányados alkalmazható olyan mennyiségi ismérvek közötti kapcsolat szorosságának mérésekor, amelyek esetében feltételezhető, hogy a kapcsolat iránya nem állandó. A mérőszám a vegyes kapcsolatnál bemutatott szóráshányados mutatójának mennyiségi ismérvek esetén használatos adaptációja. Lényege, hogy a magyarázó változó nem egyedi értékeivel szerepel a vizsgálatban, hanem csoportosító ismérvként. Ez annyit jelent, hogy ezt a mennyiségi ismérvet nominális skálán mért ismérvként kezelve csak az egyes csoportok eredményváltozó szerinti különbözőségét használjuk ki az elemzés során. Könnyen észrevehető, hogy a fenti eljárás jelentős információveszteséget okoz, mintegy ezzel "fizetünk" azért, hogy nem ismerjük a kapcsolatot leíró függvény típusát. A mutató számítása során használt adatbázis:

A korrelációs hányados adatbázisa

6-9. tábla

$x_{1min} - x_{1max}$	$x_{2min} - x_{2max}$	...	$x_{mmin} - x_{mmax}$
y_{11}	y_{21}		y_{m1}
y_{12}	y_{22}		y_{m2}
.	.		.
.	.	...	.
.	.		.
y_{1n_1}	y_{2n_2}		y_{mn_m}

A 6-9. táblában m csoportot képeztünk a magyarázó változó értékei alapján, és felsoroltuk az egyes csoportokban szereplő eredményváltozó értékeket. Vegyük észre, hogy az egyes csoportok gyakorisága nem feltétlenül egyezik meg!

A korrelációs hányados kiszámítása a továbbiakban teljesen megegyezik a szóráshányados mutatójának kiszámításával, képlete:

$$\eta_{Y|X} = \frac{\sigma_{K_Y}}{\sigma_Y} \qquad (6\text{-}32)$$

Könnyen belátható, hogy a mutató nem szimmetrikus, azaz

$$\eta_{Y|X} \neq \eta_{X|Y}$$

Ennek megfelelően a mutató esetében különösen kell ügyelnünk a magyarázó, illetve eredményváltozó idejekorán való elkülönítésére.

Láthattuk, hogy a korrelációs hányados meghatározása, szemlélete teljesen megegyezik a vegyes kapcsolatnál alkalmazott szóráshányados mutatójánál bemutatottal. Ebből következően nem meglepő, hogy a kapcsolat szignifikáns voltának tesztelését a vegyes kapcsolatnál ismertetett módon, varianciaanalízissel lehet elvégezni. A hipotézisrendszer:

$$H_0: \ \eta_{Y|X} = 0$$
$$H_1: \ \eta_{Y|X} > 0$$

A próbafüggvény:

$$F = \frac{\dfrac{S_{K(Y|X)}}{m-1}}{\dfrac{S_{B(Y|X)}}{n-m}} \qquad (6\text{-}33)$$

amely **(m-1) és (n-m) szabadságfokú F-eloszlást** követ.

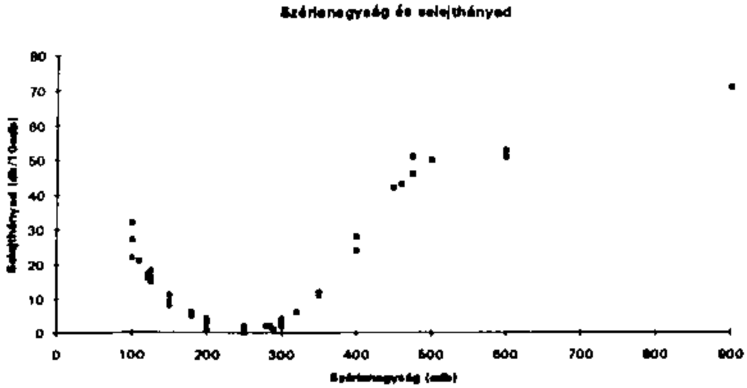
A mutató kiszámítását szemlélteti a következő példa:

Egy híradástechnikai termékeket gyártó mammutvállalat meg kívánja vizsgálni, milyen szoros a kapcsolat a gyártott termékek szérianagysága (ezer db; x) és selejthányada között (db/10000 termék; y). A vizsgálatba 50 terméket vontak be, a felmérés eredményeit a következő tábla tartalmazza:

i.	x	y	i.	x	y	i.	x	y	i.	x	y	i.	x	y
1.	100	27	11.	150	9	21.	200	3	31.	300	2	41.	400	24
2.	100	32	12.	150	11	22.	200	3	32.	300	3	42.	400	28
3.	100	22	13.	150	11	23.	200	1	33.	300	3	43.	450	42
4.	110	21	14.	150	8	24.	200	3	34.	300	4	44.	460	43
5.	120	16	15.	150	11	25.	250	0	35.	300	2	45.	475	51
6.	120	17	16.	180	5	26.	250	2	36.	300	2	46.	475	46
7.	125	15	17.	180	6	27.	250	1	37.	300	3	47.	500	50
8.	125	15	18.	200	4	28.	280	2	38.	320	6	48.	600	53
9.	125	18	19.	200	4	29.	285	2	39.	350	11	49.	600	51
10.	125	16	20.	200	3	30.	290	1	40.	350	12	50.	900	71

Vizsgáljuk meg, milyen szoros a korrelációs kapcsolat a szérianagyság és a selejtarány között!

Ábrázoljuk a selejthányadot a szérianagyság függvényében:



6-7. ábra Összefüggés a szérianagyság és a selejthányad között

Az ábrából kitűnik, hogy a két változó közötti kapcsolat iránya változó, ugyanis a selejtarány eleinte a szérianagyság növekedésével párhuzamosan csökken, később azonban nő. (A vállalati gazdaságtan ismeri ezt a problémát: lényege, hogy túl kis szériáknál a "betanulatlanság", túl nagy szériáknál a "munka monotonsága" növeli a selejthányadot.) A fentiek miatt a kapcsolat szorosságát a korrelációs hányados mutatójával kíséreljük meg leírni. Képezzünk a szérianagyság alapján csoportokat, és számítsuk ki a csoportokhoz tartozó fontosabb mutatókat:

Mutató	Szérianagyság (ezer db)			
	- 150	151 - 380	381 -	Σ
Gyakoriság	15	25	10	50
Átlag	16,60	3,52	45,90	15,92
Szórás	6,47	2,76	12,54	17,45

A fenti tábla alapján meghatározható a külső szórás:

$$\sigma_{K_{(Y|X)}} = \sqrt{\frac{15(16{,}60-15{,}92)^2 + 25(3{,}52-15{,}92)^2 + 10(45{,}90-15{,}92)^2}{50}} = 16{,}02$$

Képezzük a korrelációs hányados mutatóját:

$$\eta_{Y|X} = \frac{16{,}02}{17{,}45} = 0{,}918$$

A mutató alapján megállapítható, hogy a mintában a szérianagyság és a selejtarány között szoros korrelációs kapcsolat van; a szérianagyság a selejthányad szóródásának 84,3 %-át magyarázza meg. Mód nyílik - a varianciaanalízis módszerének alkalmazásával - annak tesztelésére is, vajon szignifikánsnak tekinthető-e a hipotézis, miszerint a szérianagyság hat a selejtarány alakulására. Eredményünk

$$F = \frac{\frac{S_K}{m-1}}{\frac{S_B}{n-m}} = 126{,}13$$

szerint a két változó közötti kapcsolat minden ésszerű szignifikancia-szinten jelentős ($\alpha < 0{,}0001$).

A korrelációs hányados kiszámításának sarkalatos pontja a magyarázó változó értékeinek helyes csoportosítása. Könnyen átlátható, hogy amennyiben nem homogén (vagyis az eredményváltozó szempontjából nem homogén) csoportokat képezünk, úgy egy esetleg szoros korrelációs kapcsolat esetén is viszonylag alacsony mutatóértéket kapunk. A helyes csoportosítás eléréséhez nehéz támpontot adni, de mindenképpen segít a változók megfigyelt értékeinek ábrázolása. Tulajdonképpen maga a mutatószám is egy mértéke az osztályozás helyességének (lásd! varianciaanalízisnél elmondottak!). A számítógépek elterjedésének korában lehetséges megoldás az összes ésszerűnek tűnő csoportosítás végrehajtása, és az ezek alapján számított mutatószámok közül a legnagyobb kiválasztása. (Ügyeljünk arra, hogy az egyes csoportokhoz tartozó megfigyelések száma ne legyen túl kicsi, valamint arra, hogy a csoportszám növelése a mutató értékének növekedését eredményezi!) A fenti nehézségek miatt a mutató használata csak akkor ajánlatos, ha meggyőződtünk arról, hogy a lineáris korrelációs-, illetve a rangkorrelációs együttható nem használható.

A megfelelő korrelációs mérőszám kiválasztása

Az előbbiekben bemutattunk három, kiszámítása mellett, szemléletében is más mutatószámot, melyekkel a kétváltozós korrelációs kapcsolat szorosságát számszerűsíthetjük. A következőkben röviden szólunk az egyes mutatók alkalmazásának előnyeiről és hátrányairól, valamint a helytelen mutatóválasztás következményeiről.

Az előzőekből következik, hogy közülük a korrelációs hányados alkalmazható a legszélesebb körben, valamint, hogy a lineáris korrelációs együttható felhasználhatósága a legkorlátozottabb[25]. Ez könnyen belátható, ha arra gondolunk, hogy a lineáris kapcsolat a monoton kapcsolatok egy speciális eseteként fogható fel; illetve, hogy a változó irányú kapcsolat általánosabb, mint a monoton (egyirányú) összefüggés. Mivel a kapcsolat jellegére vonatkozó ismeretünket előzetes információként is felfoghatjuk, így nem meglepő, hogy amennyiben a kapcsolat valóban lineáris, úgy a lineáris korrelációs együttható a leghatékonyabb mutatószám. Természetesen ugyanígy érvényes, hogy monoton kapcsolat esetén a rangkorrelációs együttható hatékonyabb mérőszám, mint a korrelációs hányados. Összefoglalásként elmondható, hogy célszerű mindig az adott jellegű kapcsolatnak megfelelő mérőszámot alkalmazni,

[25] A gyakorlatban ezzel ellentétes tendenciát tapasztalhatunk: mivel a kapcsolatok legnagyobb részénél - az egyszerűbb áttekinthetőség érdekében - élünk a linearitás feltételezésével, így a korrelációs kapcsolatok vizsgálatában leggyakoribb a lineáris korrelációs együttható alkalmazása.

de ha valami gátol ebben bennünket, akkor használható az általánosabb jellegű kapcsolatra kidolgozott koefficiens is.

Felmerülhet a kérdés: mi történik, ha helytelen specifikációt alkalmazunk, vagyis olyan mutatóval mérjük a kapcsolat szorosságát, amelyet más - kevésbé általános jellegű - kapcsolatra dolgoztak ki. Ennek szemléltetésére tekintsük a korábban bemutatott példákat. Kiszámítottuk mindhárom példa adatai alapján, mindhárom mutatószámot[26], eredményeink:

Mutatószám	Autók	Reklám	Selejthányad
r	-0,95327	0,68447	0,68025
ρ	-0,98071	0,95152	0,12378
η	0,91429	0,81467	0,91812

A tábla alapján egyértelműen leolvashatók a különböző mutatók alkalmazási sajátosságai. Lineáris kapcsolat esetén nem követünk el hibát egyik mutató használata esetén sem, mindhárom mutató szoros kapcsolatot jelez. Ügyelnünk kell azonban arra, hogy a korrelációs hányados értelemszerűen nem mutatja a kapcsolat irányát, hiszen ezt változó irányú kapcsolat esetére dolgozták ki. Monoton kapcsolat esetén (reklámkiadás és nyereség kapcsolata) a lineáris korrelációs együttható megtévesztő lehet (nem feltétlenül az!), általában "alábecsüli" a kapcsolat szorosságát. Az η-mutató itt használható, esetünkben a ρ-nál kisebb érték valószínűleg a nem tökéletesen megválasztott csoportosításnak tudható be. A változó irányú kapcsolat elemzésekor mind a lineáris korrelációs-, mind a rangkorrelációs együttható torz eredményre vezet, a kapcsolatot gyengébbnek tüntetik fel, és az általuk meghatározott kapcsolat irány is

[26] A korrelációs hányados esetében alkalmazott csoportosítás:
autó kora (év): - 3 ; 4 - 7 ; 8 - reklámköltség(%): - 3 ; 3.1 -

értelmezhetetlen. Az ilyen jellegű kapcsolatoknál fordul elő leggyakrabban, hogy a helytelenül megválasztott korrelációs mutatószám a kapcsolat hiányát mutatja, holott az ismérvek között szignifikáns összefüggést találhatunk.

6.4.2 A többváltozós korrelációs kapcsolat elemzése

A több magyarázó változót tartalmazó korrelációs kapcsolat elemzésének ismertetése során csak a lineáris kapcsolat vizsgálatának eszközeivel foglalkozunk. Tesszük ezt két okból:

- **egyrészt** a nemlineáris kapcsolatok elemzése vagy ennek alapján, vagy a korábban bemutatott módszerek segítségével elvégezhető (hiszen a rangkorrelációs együttható nem más, mint egy rangszámok alapján felírt lineáris korrelációs e-gyüttható, illetve több csoportosító ismérvet tartalmazó vegyes kapcsolat);
- **másrészt** többváltozós esetben, azért hogy az elemzés áttekinthető legyen, általában élünk a linearitás feltételezésével[27].

A többváltozós korrelációs kapcsolat vizsgálata a 6-8. táblában bemutatott sematikus adatbázis alapján történik, az elemzés a lineáris korrelációs együtthatón[28] alapul.

A jelenség vizsgálatának szükségessége akkor merül fel, amikor egy eredményváltozó alakulását több tényezőváltozó is befolyásolja, és meg akarjuk állapítani, hogy a függő változó szóródásának hány százaléka tudható be a vizsgálatba vont magyarázó

[27] A gyakorlatban szinte megvalósíthatatlan az olyan kapcsolatok elemzése, amelyekben a különböző változók között más és más jellegű kapcsolat van; például a X_1 és Y között lineáris, a X_2 és Y között monoton, X_1 és X_2 között változó irányú.

[28] A továbbiakban a statisztikai gyakorlatban bevett szóhasználatnak megfelelően lineáris korrelációs együttható helyett a korrelációs együttható elnevezést használjuk.

változóknak, és hány százaléka ettől független tényezőknek, vagyis - az elemzés szempontjából - a véletlennek. A gyakorlati közgazdaságtan feladata lehet például annak meghatározása, hogy milyen a kapcsolat a gazdálkodó egységek nyeresége, és az ezt befolyásoló tényezők (árbevétel, állóeszköz-állomány, dolgozói létszám stb.) között. Erre a kérdésre a többváltozós korrelációszámítás eszközeinek felhasználásával tudunk válaszolni.

Kettőnél több mennyiségi ismérv kapcsolatának elemzése az ún. **korrelációs mátrixon** alapul. A mátrix a totális, kétváltozós korrelációs együtthatókat tartalmazza:

$$\mathbf{R} = \begin{bmatrix} r_{yy} & r_{y1} & r_{y2} & . & . & . & r_{yk} \\ r_{1y} & r_{11} & r_{12} & . & . & . & r_{1k} \\ r_{2y} & r_{21} & r_{22} & . & . & . & r_{2k} \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ r_{ky} & r_{k1} & r_{k2} & . & . & . & r_{kk} \end{bmatrix}$$

Kihasználva, hogy a változók önmagukkal mért korrelációja értelemszerűen 1 ($r_{ii} = 1$), és a korrelációs együttható szimmetrikus, a mátrix felírható a következő egyszerűbb alakban is:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{y1} & r_{y2} & . & . & . & r_{ykp} \\ r_{y1} & 1 & r_{12} & . & . & . & r_{1k} \\ r_{y2} & r_{12} & 1 & . & . & . & r_{2k} \\ . & & & & & & \\ . & & & & & & \\ . & & & & & & \\ r_{yk} & r_{1k} & r_{2k} & . & . & . & 1 \end{bmatrix} \qquad (6\text{-}34)$$

A korrelációs mátrix segítségével kiszámíthatók a többváltozós kapcsolat esetében jelentős parciális-, illetve többszörös mutatószámok is. Legyen **Q** mátrix a korrelációs mátrix inverze, azaz:

$$\mathbf{R}^{-1} = \mathbf{Q} = \left[q_{ij}\right] \qquad (6\text{-}35)$$

Ekkor az Y és X_i közötti **parciális korrelációs** együttható felírható:

$$r_{yi.12...i-1,i+1...k} = \frac{-q_{yi}}{\sqrt{q_{yy}q_{ii}}} \qquad (6\text{-}36)$$

Az együttható két magyarázó változó esetén a totális korrelációs együtthatókból közvetlenül is kiszámítható (pl. $r_{y1.2}$):

$$r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{\left(1 - r_{y2}^2\right)\left(1 - r_{12}^2\right)}} \qquad (6\text{-}37)$$

A parciális korrelációs együtthatók négyzetét **parciális determinációs együtthatónak** nevezzük, és - akárcsak a totális mutató esetében - magyarázó erőként értelmezzük.

A parciális korrelációs együtthatók nullától való különbözősége, ugyanúgy mint a totális koefficiensek esetén, t-próbával tesztelhető, bár a próbafüggvény némiképp módosul:

$$t = \frac{r_{yi.12...i-1,i+1,...k}\sqrt{n-k-1}}{\sqrt{1-r^2_{yi.12...i-1,i+1,...k}}} \tag{6-38}$$

A nullhipotézis teljesülése esetén a próbafüggvény **(n-k-1) szabadságfokú t-eloszlást** követ.

Szintén a korrelációs mátrix inverzének felhasználásával számíthatjuk ki a **többszörös korrelációs együtthatót**, melyet R-rel jelölünk:

$$R = \sqrt{1 - \frac{1}{q_{yy}}} \tag{6-39}$$

Háromváltozós kapcsolat esetén a mutató kiszámításának egyszerűbb formája:

$$R = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2 \times r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}} \tag{6-40}$$

Vegyük észre, hogy a mutató számítása során alkalmazott gyökvonás miatt, R értéke mindig pozitív; így a **többváltozós kapcsolat iránya nem értelmezhető**[29]. A mutató négyzete a **többszörös determinációs együttható**, melynek főleg a regresszióanalízisben van kitüntetett szerepe.

[29] Ez a gyakorlatban érthető megszorítás, hiszen nem értelmezhető a kapcsolat iránya, ha van olyan magyarázó változó, amely pozitív- és olyan is, amely negatív kapcsolatban áll az eredményváltozóval.

A többszörös korrelációs együttható alapján tesztelhetjük azon nullhipotézisünket, miszerint a magyarázó változók együttesen nem gyakorolnak szignifikáns hatást az eredményváltozó alakulására.

Hipotézisrendszerünk:

$$H_0: \ R = 0$$
$$H_1: \ R > 0 \qquad (6\text{-}41)$$

A hipotézis tesztelése F-próbával történik:

$$F = \frac{\dfrac{R^2}{k}}{\dfrac{1-R^2}{n-k-1}} \qquad (6\text{-}42)$$

ahol a próbafüggvény a nullhipotézis teljesülése esetén k és **(n-k-1) szabadságfokú F-eloszlást** követ.

A többváltozós korrelációs kapcsolat elemzését szemlélteti a következő példa:

Folytatva az autós példát, a használt autók árát (y; EFt) befolyásoló tényezők vizsgálata során a gépkocsik életkora (x_1; év) mellett figyelembe vettük az eddig futott kilométert (x_2; Ekm) is. Adatbázisunk:

Gépkocsi	y	x_1	x_2
1.	400	3	55
2.	130	10	140
3.	200	7	130
4.	180	6	140
5.	370	3	70
6.	540	1	12
7.	510	1	25
8.	580	1	10
9.	470	2	24
10.	300	5	60
11.	100	12	200
12.	200	7	150
13.	250	6	100
14.	260	5	100
15.	210	7	120
16.	320	4	70
17.	570	1	14
18.	170	9	130
19.	90	11	250
20.	150	9	200
Átlag	300	5,5	100
Szórás	157,5	3,4	67,2

Elemezzük a kapcsolatot a gépkocsik ára és az ezt befolyásoló tényezők között!

Elsőként határozzuk meg a változók korrelációs mátrixát:

$$\mathbf{R} = \begin{bmatrix} 1{,}0000 & -0{,}9533 & -0{,}9308 \\ -0{,}9533 & 1{,}0000 & 0{,}9418 \\ -0{,}9308 & 0{,}9418 & 1{,}0000 \end{bmatrix}$$

A korrelációs mátrix vizsgálata azt mutatja, hogy mindkét magyarázó változó rendkívül szoros negatív kapcsolatban áll az eredményváltozóval, ám a köztük talált, szintén magas korrelációs együttható érték ezen eredmények óvatos kezelését sugallja.

Képezzük a mátrix inverzét:

$$\mathbf{R}^{-1} = \mathbf{Q} = \begin{bmatrix} 12,2499 & 8,3082 & 3,5775 \\ 8,3082 & 14,4854 & -5,9093 \\ 3,5775 & -5,9093 & 9,8954 \end{bmatrix}$$

Az inverzmátrix alapján képezhetők a változók közötti "tiszta" kapcsolatot jobban kifejező parciális mutatók:

$$r_{y1.2} = \frac{-8,3082}{\sqrt{12,2499 \times 14,4854}} = -0,6238$$

$$r_{y2.1} = \frac{-3,5775}{\sqrt{12,2499 \times 9,8954}} = -0,3249$$

Mivel háromváltozós kapcsolatról van szó, a parciális mutatók meghatározhatók közvetlenül a totális korrelációs együtthatók alapján is. Például:

$$r_{y2.1} = \frac{-0,9308 - \left(-0,9533 \times 0,9418\right)}{\sqrt{\left[1 - \left(-0,9533\right)^2\right]\left(1 - 0,9418^2\right)}} = -0,3249$$

A parciális korrelációs együtthatók megmutatják, hogy a totális korrelációs együtthatók alapján számszerűsített rendkívül szoros kapcsolatban - az autók kora és ára, illetve a futásteljesítménye és ára között - a magyarázó változók egymás közti szoros kapcsolatának is szerepe van. Amennyiben kiszűrjük a futásteljesítmény hatását, úgy a gépkocsik ára és kora között csupán közepes szorosságú, negatív irányú korrelációs kapcsolatot mutathatunk ki. Ennél is szembeötlőbben csökken a gépkocsik által eddig megtett útnak az árra gyakorolt magyarázó ereje, ha kiszűrjük az életkor hatását. Míg a totális mutatók alapján a futásteljesítmény magyarázó ereje közel 86,6 % volt, addig az életkortól "tisztított" magyarázó erő mindössze 10,6 %.

Mivel következtetéseinket minta alapján vonjuk le indokolt a fenti kapcsolatok szignifikáns voltának tesztelése:

$$t_{y1.2} = \frac{-0{,}6238\sqrt{20-2-1}}{\sqrt{1-(-0{,}6238)^2}} = -3{,}29$$

$$t_{y2.1} = \frac{-0{,}3249\sqrt{20-2-1}}{\sqrt{1-(-0{,}3249)^2}} = -1{,}42$$

A 17 szabadságfokú t-eloszlás 5 %-os szignifikancia-szinthez tartozó kritikus értékei ± 2,11. Ennek értelmében csak az autók életkora parciális hatásának esetében kell elvetnünk a nullhipotézist, vagyis a másik magyarázó változó hatását kiszűrve csak a gépkocsik életkora gyakorol szignifikáns hatást az autók árára.

Az egyedi hatások elemzése után vizsgáljuk meg, mekkora magyarázó erővel bír a mintában a két tényezőváltozó együttesen! A kérdésre a többszörös determinációs együttható meghatározásával adhatunk választ:

$$R^2 = 1 - \frac{1}{12{,}2499} = 0{,}9184$$

vagy a totális mutatókból közvetlenül:

$$R^2 = \frac{(-0{,}9533)^2 + (-0{,}9308)^2 - 2(-0{,}9533)(-0{,}9308)(0{,}9418)}{1-0{,}9418^2} = 0{,}9184$$

A determinációs együttható alapján megállapíthatjuk, hogy az autók kora és eddigi futásteljesítménye együttesen a vételár szóródásának 91,84 %-át magyarázzák meg. A magyarázó változók és az eredményváltozó közötti együttes hatás szoros (R = 0,9583).

Vizsgáljuk meg, hogy szignifikánsnak tekinthető-e a fenti többváltozós kapcsolat! A próbafüggvény:

$$F=\frac{\dfrac{0,9184}{2}}{\dfrac{1-0,9184}{20-2-1}}=95,67$$

A próbafüggvény empirikus értéke minden ésszerű szignifikancia-szinten meghaladja az elméleti értéket, ezért a nullhipotézist elvetjük, a kapcsolatot szignifikánsnak tartjuk.

A többváltozós kapcsolat elemzési eszközeinek ismertetése után röviden térjünk ki a különböző korrelációs együtthatók nagyságrendi viszonyaira is! Elmondható, hogy azt az extrém esetet kivéve, amikor a vizsgálatba vont további magyarázó változók nincsenek kapcsolatban az eredményváltozóval, a többszörös korrelációs (és determinációs) együttható mindig meghaladja a kapcsolatban szereplő totális mutatókat. A parciális és a totális korrelációs együttható között nem adhatunk meg egyértelmű nagyságrendi viszonyt, ezek lehetnek akár egyenlőek is, de közülük bármelyik meghaladhatja a másikat.

7. REGRESSZIÓANALÍZIS

Korábbi tanulmányainkban megismerkedtünk a mennyiségi ismérveknek a statisztikai elemzésben betöltött szerepével. Mivel a mennyiségi ismérvek általában intervallum,- illetve arányskálán mérhetők, a statisztika alkalmazott módszereinek köre jelentősen kibővül. Az előző fejezetben tanulmányoztuk a korrelációszámítás főbb módszereit, amelyek jelentős segítséget adtak a jelenségek, folyamatok közötti kapcsolatok megismeréséhez. Ebben a fejezetben a regresszióanalízis legfontosabb problémáit érintjük, szem előtt tartva az alkalmazás fontosságát.

Korábban megismerkedtünk a **korrelációszámítással**, amely - mint a mennyiségi ismérvek elemzésének matematikai-statisztikai módszere - a jelenségek, folyamatok szorosságának mérését, a kapcsolat létének felismerését és minősítését hivatott ellátni. A **regresszióanalízis** a kapcsolat létének felismerésén túl lehetőséget nyújt a korrelációs kapcsolat számszerű leírására, segítségével elemezhető a kapcsolat tartalma, iránya, az ismérvek egymásra gyakorolt hatásának iránya, mértéke, azaz az összefüggés.

A regresszióanalízis módszertana feltételezi azt, hogy a kapcsolatban rejlő törvényszerűségeket matematikai modellel írjuk le, ami az elemzési lehetőségek körét kibővíti.

A modell említése igényli a fogalom rövid magyarázatát. A bennünket körülvevő, rendkívül összetett világ megismerése során a különféle tudományágak gyakorta élnek a modellezés adta előnyökkel. A **modell** (amint a TK. I. 1. fejezetében megismertük) a valóság leegyszerűsített, kicsinyített mása, amely a vizsgálat szempontjából fontos hatásokat úgy tartalmazza, hogy kiemeli a lényeges, releváns jellemzőket és eltekint a kevésbé szignifikánsaktól, elhanyagolhatóktól. Minden modell - és ez érvényes a szimbolikus nyelven megfogalmazott modellekre is - különböző alkotó-elemekből épül fel. Az alkotó-elemek egyik csoportja a vizsgált rendszer legfontosabb belső (endogén) jegyeit hordozza, míg a másik csoport a rendszerre ható külső (exogén) szerepkört tölti be. A társadalmi-gazdasági folyamatok, valamint a pillanatnyi állapotot kifejező jelenségek vizsgálói, kutatói számára felbecsülhetetlen segítséget jelenthet a modell-alkotás. E modellezés nehézsége alapvetően abból adódik, hogy a társadalmi-gazdasági folyamatok állandóan változó körülmények között mennek végbe, ugyanúgy soha nem ismétlődnek meg. Mindez nagymértékben behatárolja a modellezés során felhasználható eszközök körét.

A társadalmi-gazdasági folyamatok, jelenségek vizsgálata során a **modell fogalmán a vizsgált valóságnak** - a vizsgálat szempontjából - **legfontosabb vonásait, összefüggéseit kifejező logikai, matematikai konstrukciót értünk.** Ennek értelmében a valóság szimbolikus, formalizált ábrázolását adhatjuk a modell segítségével, amely történhet egy egyenlet vagy akár egyenletek rendszere által.

Nagyon egyszerűen lehet formalizálni az eladott mennyiség (X) és az árbevétel (Y) összefüggését az egységár (p) figyelembevételével: Y=pX. Itt az egységár a konstans paraméter szerepét tölti be. Amennyiben n számú megfigyeléssel rendelkezünk, mind az X mind az Y változóra vonatkozóan - a feltételek változatlansága mellett - minden esetben pontosan meghatározható egy adott mennyiséghez tartozó árbevétel[1]. A fenti kapcsolat a két változó között determinisztikus (függvényszerű) összefüggést tételez fel. A függvény X változóját **független-, tényező- vagy magyarázóváltozónak**, az Y változót **függő-, eredményváltozónak** hívjuk. Amennyiben a fenti összefüggést kibővítjük egy véletlen taggal, egy a véletlen hatását reprezentáló valószínűségi változóval, **sztochasztikus függvényt** kapunk, amelyben már nem határozható meg egyértelműen az eredményváltozó. Az ily módon kibővített függvény egy ún.

---

[1] Természetesen itt ki kell zárni - a valóságban egyébként létező - árengedményeket, akciókat, a boltonkénti más-más árat.

statisztikai modell vázát adja. A statisztikai modellek nagyon lényeges csoportját alkotják a **regressziós modellek.**

Az általunk vizsgált regressziós modellekben egy eredményváltozó és egy vagy több tényezőváltozó van. A modell alapján megállapítható, hogy a tényezőváltozó(k) milyen módon és milyen törvényszerűség mellett fejti(k) ki hatását (hatásukat) az eredményváltozóra.

Ebben a fejezetben a két- és többváltozós regressziós modell-alkotás főbb kérdéseivel ismerkedünk meg; vizsgálódásaink tárgyát a gyakorlatban jól alkalmazható és még elméleti alapokkal is jól alátámasztható lineáris illetve lineárisra visszavezethető modellek képezik. Ezeknek a modelleknek közös sajátossága, hogy paramétereik viszonylag egyszerűen becsülhetők és az eredmények könnyen interpretálhatóak.

7.1 KÉTVÁLTOZÓS LINEÁRIS REGRESSZIÓ

A lineáris regresszió matematikai modellje képezi a regressziós modell építésének alapját. A korábban már említett elveknek megfelelően jelöljük X-szel a tényezőváltozót és Y-nal az eredményváltozót, a megfigyelt értékpárokat pedig jelölje x_i és y_i ($i=1,2,...n$). Minthogy sztochasztikus kapcsolatról van szó, a modell kialakítása során figyelembe kell venni a véletlen hatását, amit a modellben egy (ε-nal jelölt) véletlen változó, azaz valószínűségi változó képvisel. A nagyon leegyszerűsített modell az alábbi:

$$Y = f(X, \varepsilon) \tag{7-1}$$

A tényezőváltozó eredményváltozóra gyakorolt hatása egy konkrét matematikai függvény segítségével írható le, ennek a függvénynek a konkretizálása azonban a korábban tanult valószínűségelmélet ismeretek felidézését igényli.

7.1.1 Elméleti alapvetés

Tételezzük fel, hogy mindkét változónk (X,Y) valószínűségi változó. Ahhoz, hogy az Y változót jól jellemezhessük célszerű az X változó értékeit is figyelembe venni. A konkrét numerikus számításokhoz szükség van egy olyan formulára, függvényre, amelybe a független változó értékét behelyettesítve Y becslését kapjuk.

Abban az esetben, ha csak az Y változóra vonatkozóan rendelkeznénk információval, a várható érték kielégítő becslést eredményezne. Esetünkben azonban az X változóra vonatkozóan is rendelkezünk mért értékekkel, amelyek az Y változóra vonatkozóan fontos tartalommal bírnak. Ennek az információnak a felhasználásától nem célszerű eltekinteni, ezért hatékony megoldáshoz jutunk, ha Y becslését Y-nak X-re vonatkoztatott **feltételes várható értékével** közelítjük.

$$y = E(Y \mid X = x) \qquad (7\text{-}2)$$

A fenti függvényt **Y valószínűségi változó X-re vonatkoztatott regressziójának nevezzük.** Amennyiben ugyanis rögzítjük X értékét egy adott x_i szinten, az x_i-vel egyidejűleg fellépő Y értékek **feltételes eloszlást** alkotnak. Ezek alapján célszerűen járunk el, ha a feltételes valószínűség-eloszlások ismeretében kíséreljük meg a probléma megoldását.

Amennyiben X és Y diszkrét valószínűségi változók, amelyek x_1, x_2,...x_k valamint y_1,y_2,...y_m értékpárok minden kombinációiban előfordulhatnak, az Y várható értéke $X=x_i$ feltétel mellett a következőképpen írható fel:

$$E\left(Y \middle| x_i\right) = \sum_{j=1}^{m} y_j \Pr(Y = y_j | x_i) \qquad (7\text{-}3)$$

A folytonos együttes eloszlásból származó feltételes várható értéket az alábbi formula reprezentálja:

$$E(Y|x) = \int_{-\infty}^{+\infty} y\, g(y|x)dy \tag{7-4}$$

A $g(y|x)$ jelöli az Y valószínűségi változó X-re vonatkoztatott feltételes sűrűségfüggvényét. Ez az X és Y együttes sűrűségfüggvénye és az X változó sűrűségfüggvényének segítségével fejezhető ki:

$$g(y|x) = \frac{h(x,y)}{f(x)} \tag{7-5}$$

A gyakorlat szempontjából különös jelentőséggel bír a kétdimenziós normális eloszlás. A központi határeloszlás tételéből tudjuk azt, hogy nagy számú megfigyelés esetén, azonos eloszlások mellett várható a kétdimenziós normális eloszlás bekövetkezése. Ezek alapján sokszor számíthatunk arra, hogy a vizsgált gazdasági jelenségek és folyamatok is leírhatók a kétdimenziós normális eloszlással.

Az X és Y változók együttes kétdimenziós normális eloszlású sűrűségfüggvényének és az X változó sűrűségfüggvényének hányadosaként felírható feltételes sűrűségfüggvénynek, rögzített X értékek mellett, a feltételes várható értéke[2]:

$$E(Y|x) = E(Y) + \rho \frac{D(Y)}{D(X)}\left[x - E(X)\right] \tag{7-6}$$

A fenti képletből közvetlenül leolvasható, hogy normális együttes eloszlás esetén Y-nak X-re vonatkozó regressziója egy egyenes, amely keresztül megy a sík E(X), E(Y) koordinátájú pontjain és iránytangense:

[2] Bizonyítását lásd! Irodalom [11]

$$\rho\frac{D(Y)}{D(X)}.$$

A fenti összefüggésben D(X) és D(Y) a változók elméleti szórásai és ρ az elméleti korrelációs együttható. A fenti megállapítások közül talán a legfontosabb az, hogy a kétdimenziós normális eloszlás esetén **a regresszió lineáris függvény.** Mindez elméleti oldalról alátámasztja a lineáris regresszió széleskörű alkalmazását a gyakorlatban. Fontos továbbá, hogy függetlenség esetén az egyenes párhuzamos az X tengellyel, és mivel a szórások nem lehetnek nullák, a függetlenséget ρ=0 közvetíti.

A fentiekben összefoglalt elméleti összefüggésekből világosan kitűnik, hogy az elméleti regressziót, amelyet a feltételes várható érték fejez ki, a gyakorlatban csak közelíthetjük. A közelítés egyik egyszerű módja az empirikus regressziófüggvény. A gazdasági-társadalmi jelenségek, folyamatok vizsgálata során azonban általában nem elégedhetünk meg az empirikus regressziós függvényekkel, a közelítés másik módjához, egy konkrét függvényformula meghatározásához kell nyúlnunk. A tapasztalati adatokból meghatározható, de konkrét formulával kifejezett függvényt **analitikus regressziónak** hívjuk. Az analitikus függvény meghatározásához, és a lineáris forma használatának létjogosultságához adott elméleti támpontot az elméleti regresszió megismerése.

Könnyű annak a belátása, hogy az elméleti regresszió megfogalmazása során rögzített x_i értékekről és a hozzájuk tartozó Y feltételes várható értékekről beszéltünk. Ebből következik, hogy X nem szükségszerűen valószínűségi változó. Abban az esetben ha az n-dimenziós változók között egy vagy több olyan szerepel, amelyik nem valószínűségi változó, **regressziós modellről** van szó. A gyakorlatban a legtöbb módszert és eljárást a regressziós modellekre dolgozták ki.

A regressziós modellekben az X változó jelölhet rögzített (előre meghatározott ún. predeterminált) változót és időt[3] is. Amennyiben rögzített értékeket használunk a korreláció szorosságának mérése új megvilágítást kap. Például az idő szerepeltetése mellett nem mindig van értelme a korreláció szorossági mérőszámának. Más esetben használjuk a korrelációs mérőszámokat, és a segítségükkel levonható információk jelentősen gazdagítják a sztochasztikus modellekből nyerhető információkat.

A regresszióanalízis során általában a végtelen sokasághoz képest mindig mintával dolgozunk, erre épülnek modelljeink. Mintaként lehet felfogni a regresszióanalízishez felhasznált adatbázist, akár teljes körű, akár reprezentatív adatfelvételből származik, és a regressziós-, korrelációs modellekből levonható következtetéseket becslésekként kell értékelni. Mindezt könnyen beláthatjuk ha arra gondolunk, hogy a változók között valószínűségi változók szerepelnek.

Elképzelhető az is, hogy az X változó valószínűségi változó. Azt a modellt amelyben szereplő változók mindegyike (a tényezőváltozók és az eredményváltozó is) valószínűségi változó **korrelációs modellnek** hívjuk.

7.1.2 A regressziós függvény koncepciója

A regressziós függvény felépítését, a regresszióanalízis gondolatmenetét úgy érthetjük meg könnyen, ha egy nagyon egyszerű, szimplifikált példával próbáljuk jellemezni a valóság egy szeletét.

> Egy kereskedelmi vállalatnak 50 azonos profilú boltja van. A boltok alapterülete csupán 10 féle. Egy adott évben a boltok forgalmára vonatkozó adatok az alapterület függvényében az alábbiak:

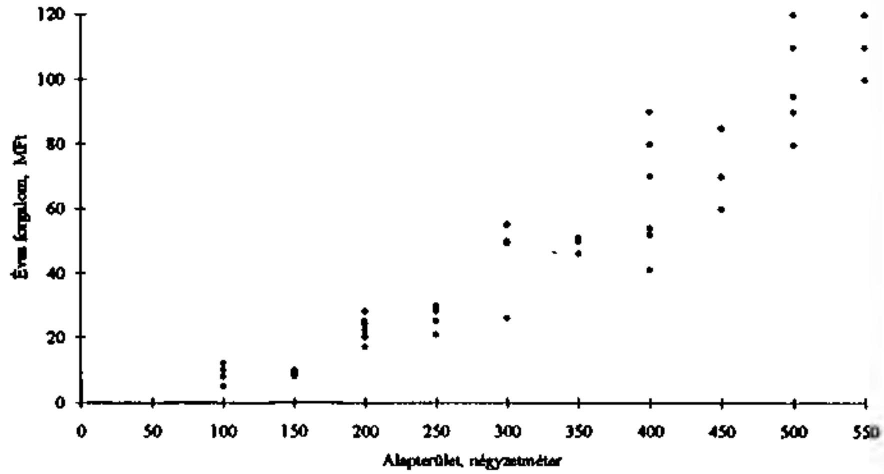
[3]Ebben az esetben a regressziófüggvényt analitikus trendnek nevezzük, amelyről bővebben a következő fejezetben beszélünk.

A kereskedelmi vállalat forgalmára vonatkozó adatok

7-1. tábla

Alap-terület, m^2	Forgalom, MFt						
100	12	10	8	5			
150	10	8	9	8	8		
200	24	28	17	20	25	22	
250	30	21	30	25	28	25	29
300	26	50	55	50	49		
350	50	46	50	51			
400	41	80	70	90	52	54	
450	60	70	70	85			
500	80	110	120	90	95		
550	100	100	120	110			

A fentiek például a következőképpen értelmezhetők: 100 m^2 alapterülető boltok száma négy, amelyek rendre 12, 10, 8, és 5 millió forint értékű forgalmat értek el az adott évben. A többi számérték hasonlóképp interpretálható. Az adatok a derékszögű koordináta rendszerben ponthalmaz segítségével ábrázolhatók, az 7-1. ábrának megfelelően. Könnyen belátható, hogy az ábrázolás eszközeként a pontdiagram azért felel meg alapvetően, mivel egy-egy X értékhez több Y érték tartozik.



7-1. ábra Alapterület és a boltok forgalma

Az ábrán világosan felismerhető az Y változó X változó rögzített értékei melletti feltételes valószínűség-eloszlása. Azt is megállapíthatjuk, hogy **tendenciájában** a nagyobb bolti alapterülethez várhatóan nagyobb éves forgalom járult. Célszerűnek tűnik, ha az Y valószínűségi változót fontos jellemzőjével a feltételes várható értékkel jellemezzük. Tételezzük fel, hogy nem járunk messze az igazságtól, ha arra gondolunk, hogy az X változó **lineárisan** fejti ki hatását az Y változóra. A feltételes várható érték, vagy másképpen regresszió az X magyarázó változó valamilyen, tetszőleges függvényeként is felírható, ami esetünkben egy egyenes:

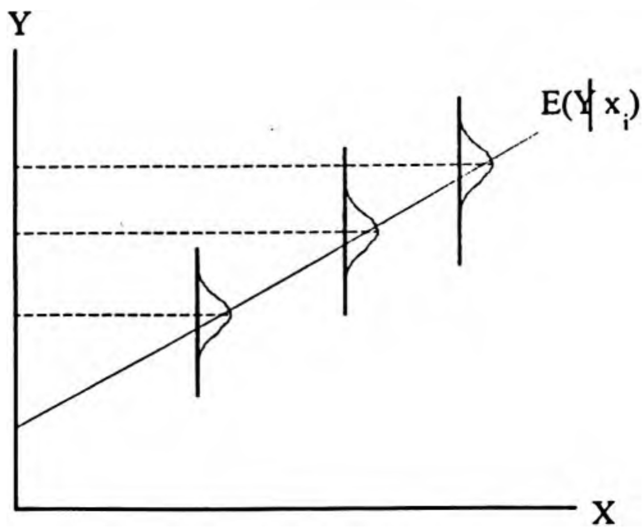
$$E(Y_i | x_i) = \beta_0 + \beta_1 x_i \qquad (7\text{-}7)$$

ahol β_0 és β_1 az általunk ismeretlen, de a valóságban létező lineáris kapcsolat paraméterei, a β_1 paramétert **regressziós együtthatónak** nevezünk. A (7-7) alatt jelzett függvényt **alap-regressziós függvénynek** hívjuk.

A regressziós egyenes körül azonban a pontok, a tényleges megfigyelés értékei szóródnak, egy adott valószínűség-eloszlásnak megfelelve helyezkednek el. Az egyes értékeket a regresszió és a **sztochasztikus hatást** kifejező véletlen változó a hibatag (vagy reziduális változó) ε, együttesen határozza meg. Y_i konkrét értékét rögzített x_i mellett az alábbi azonossággal jellemezhetjük:

$$Y_i = Y_i | x_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (7\text{-}8)$$

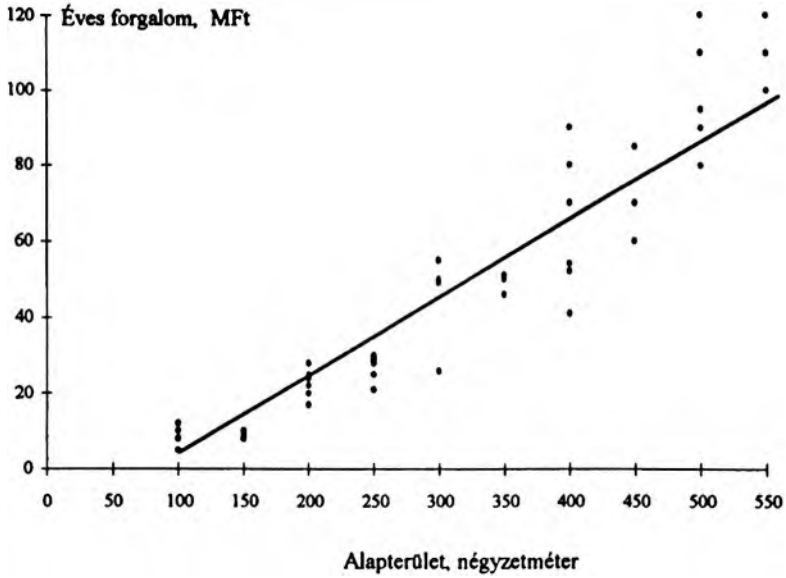
A fenti összefüggés az **alap-regressziós modell** felhasználásával egy konkrét megfigyelési érték megfogalmazását nyújtja. Mind az Y mind az ε értékek valószínűségi változók. Rögzített X mellett a konkrét megfigyelt értékek eloszlását az alábbi ábra szemlélteti.



7-2. ábra Regressziós modell

A fenti sztochasztikus kapcsolatban a véletlen hatást kifejező reziduális változóról feltettük, hogy eloszlása normális.

Tételezzük fel, hogy ismerjük a boltok alapterülete és forgalma közötti összefüggést leíró alap-regressziós függvényt.



7-3. ábra Alap-regressziós függvény

A forgalom lineárisan függ az alapterülettől és a véletlen változó hatásától. A 100 m^2-es boltok forgalma, ezek szerint a következőképpen modellezhető egy regressziós modell segítségével (ahol $x_1=x_2=...=x_4=100$):

$$y_1 = 12 = \beta_0 + \beta_1(100) + \varepsilon_1$$
$$y_2 = 10 = \beta_0 + \beta_1(100) + \varepsilon_2$$
$$y_3 = 8 = \beta_0 + \beta_1(100) + \varepsilon_3$$
$$y_4 = 5 = \beta_0 + \beta_1(100) + \varepsilon_4$$

A fenti modellből kitűnik a sztochasztikus hatást kifejező véletlen tényezőnek a szerepe az egyes konkrét értékek kialakulásában. Itt kell szólnunk arról, hogy a könnyebb megértés érdekében az ún. mérési hibáktól (adatfelvételi-, adat-feldolgozási hibáktól) fejtegetéseink során eltekintünk, azt külső adottságnak tekintjük. A véletlen hatást sok, különböző irányú és erejű, de egyenként nem jelentős tényező együttes hatásának tulajdonítjuk, amit a hibatag, más néven a reziduális változó közvetít. Fontos tudni azonban, hogy ez a véletlen hatás magában foglalja a figyelembe nem vett változók hatását is. Az Y eredményváltozó alakulásában szerepet játszanak ugyanis olyan tényezők, amelyektől eltekintünk a specifikáció[4] során, mivel pl. nem rendelkezünk róluk megbízható információkkal, illetve nem számszerűsíthetők. Ezért a leggondosabban végrehajtott specifikáció mellett is számítanunk kell arra, hogy változók maradnak ki a függvényből, ezért a hibatag vagy reziduális változó nem csak a tiszta véletlent tartalmazza. Kielégítő eredmény lehet, ha reziduális változó szisztematikus hatástól mentes, és így megközelítően véletlen változóként viselkedik.

A (7-8) alatt jelzett modellben szereplő változók várható értékeire is fennáll az azonosság

$$E(Y|x_i) = E[E(Y|x_i)] + E(\varepsilon|x_i) \qquad (7\text{-}9)$$

$$= E(Y|x_i) + E(\varepsilon|x_i)$$

Mindez természetesen azt is magában foglalja, hogy a véletlen változó feltételes várható értéke zéró:

$$E(\varepsilon|x_i) = 0 \qquad (7\text{-}10)$$

A regressziónak, mint az Y változó feltételes várható értékének számszerűsítése alapvető kérdés. Egyfajta "gyors" közelítést ad az ún. **empirikus regresszófüggvény**,

[4] Specifikáción a jelenséget leíró modellben szereplő függvények változóinak és a függvények konkrét formájának a meghatározását értjük. (A fogalmat később még részletesebben tárgyaljuk.)

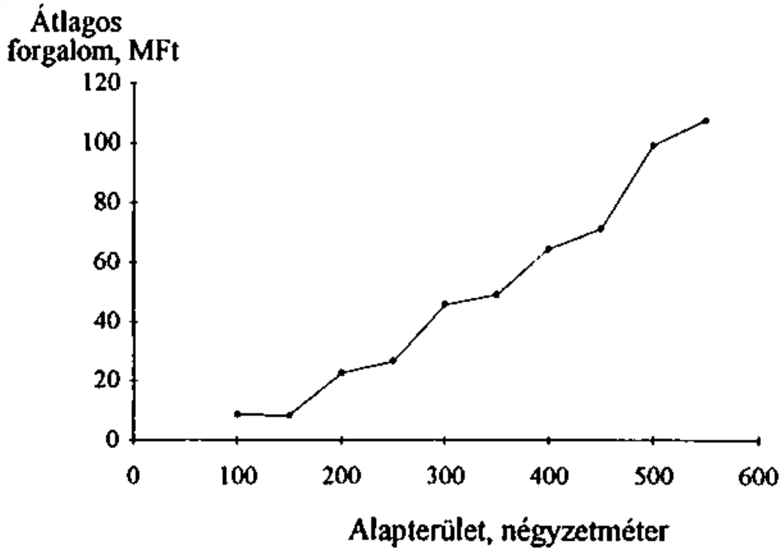
amely tulajdonképpen nem más, mint a feltételes várható értékek empirikus megfeleléseként a **részátlagok alkotta statisztikai sor.**

Empirikus regressziófüggvény

7-2. tábla

Alapterület, m^2	Átlagos forgalom, MFt
100	8,75
150	8,60
200	22,70
250	26.80
300	46,00
350	49,25
400	64,50
450	71,25
500	99,00
550	107,50

A 7-2. tábla adataiból kitűnik a két változó közötti pozitív irányú sztochasztikus kapcsolat, ezért az alapterület növekvő értékeihez egyre nagyobb átlagos forgalom tartozik. Az empirikus regressziófüggvényt ábrázolva, a vonaldiagram is jól mutatja a fent elmondottakat.



7-4. ábra Empirikus regresszió

Az empirikus regressziófüggvény a kapcsolat lényegét már tömörítve, sűrítve tükrözi. Segítségével információkat nyerhetünk a korrelációs kapcsolat természetéről, irányáról, azonban mivel részátlagok segítségével definiáltuk a függvényt, további elemzések, becslések céljaira nem tudjuk érdemben alkalmazni. Ezt a problémát hivatott feloldani az analitikus regresszió, amely tapasztalati adatokból meghatározható, konkrét formulával kifejezett függvényt használ. A függvény paramétereinek, az ún. regressziós együtthatóknak számszerűsítése, a regresszióanalízis segítségével, mért adatokból történik.

A valóságban létező alap-regressziós függvény paramétereit azonban közvetlenül nem tudjuk meghatározni, mivel a tapasztalati adatok teljes körével csak a legritkább esetben rendelkezünk. A gyakorlatban a rögzített X értékeknek megfelelő Y értékek valószínűségi változók, konkrét értékük a véletlen hatástól nagymértékben függ, amint ezt a fentiekben is érzékeltettük. Mintapéldánkban feltettük, hogy csupán 50 egység adatait vizsgáljuk. Az vizsgált jelenség (a boltok alapterülete és forgalma közötti összefüggés) azonban általános megállapításokat igényel. Ennek megfelelően a rendelkezésre álló adatbázis is csupán mintaként fogható fel, de gyakorlatban sűrűn előfordul, hogy csak kevesebb, - akárcsak példánkban 10 - adatpár képezi a vizsgálat adatbázisát. Általában feltesszük, hogy az Y értékei véletlen mintaként viselkednek. A valóság egy adott realizációjaként fogható fel a 7-1. tábla adataiból kiválasztható minta, aminek a konkrét értékeit a 7-3. táblában foglaltuk össze[5].

[5]Pédánk - didaktikai okokkal magyarázhatóan - csak kevés adatból áll. A valóság hitelesebb leírása, valamint néhány további statisztikai módszer alkalmazása nem nélkülözheti a nagyobb adatbázist, amint azt majd látni fogjuk.

Az alapterület és forgalom egy véletlen mintája

7-3. tábla

Alapterület, m^2	Forgalom, MFt
100	12
150	10
200	24
250	30
300	26
350	50
400	41
450	60
500	80
550	100

A forgalom adatai természetesen egy más mintában a fentiektől eltérő értékeket vesznek fel. A gyakorlatban azonban a legtöbb esetben csak egy minta áll az elemző rendelkezésére. Ezekből a konkrét adatokból kell levonni általános érvényű megállapításokat. Általában elmondhatjuk, hogy egy tényleges sokaságot végtelen sok egyedi adat alkot, és egy adott minta alapján számszerűsíthető sztochasztikus kapcsolat csak egyetlen a lehetséges összefüggések közül. Ezek alapján könnyen belátható, hogy bármely véges sokaság csak a minta szerepét tölti be. A konkrét adatok birtokában az alap-regresszió paramétereinek csupán **becslését** adhatjuk.

A fentiek alapján célszerűnek tűnik, ha bevezetjük a **minta-regressziós függvény** fogalmát, amely a valóságban létező regresszió mintabeli adatok alapján történő becslése. Az alap-regresszió és a minta-regresszió közötti különbözőséget a jelölés rendszerében is érzékeltetjük:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \hat{Y} + \varepsilon$$

$$y_i = b_0 + b_1 x_i + e_i$$

$$y_i = \hat{y}_i + e_i$$

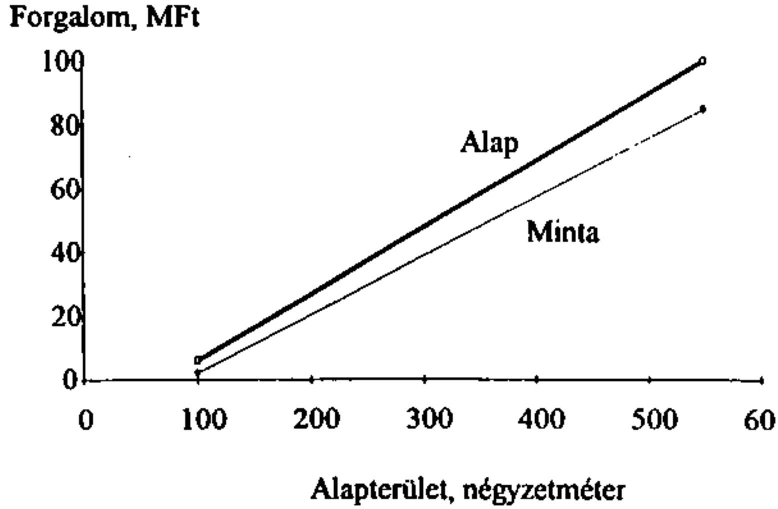
ahol:

$\hat{Y}$ becslése $\hat{y}$

β_0 becslése b_0

β_1 becslése b_1

Természetesen egy adott minta adataitól nem várhatjuk el, hogy az abból származó becslés pontosan lefedje az alap-regressziót. Amennyiben a 7-3. tábla adatait mintának tekintjük, a belőlük készített becslés kissé "lefelé becsüli" az alapsokaságból nyerhető regressziót, amint azt az alábbi ábra is szemlélteti.



7-5. ábra Az alap-regressziós függvény és a minta-regressziós függvény

A továbbiakban figyelmünket a mintából nyerhető becslésekre összpontosítjuk. Elsőként a lineáris regressziós függvény becslésével foglalkozunk. A linearitás feltétele természetesen a későbbiekben is fontos szerepet fog betölteni és ahol lehet a modelleket lineárisra vezetjük vissza.

7.1.3 A klasszikus lineáris regressziós modell

Az előzőekben láttuk, hogy a modell eredményváltozójának értéke a véletlen ingadozásától is függ. Tulajdonképpen a modellben szereplő mindkét valószínűségi változó[6] (Y és ε) olyan információkat hordoz, amelyek alapvető szerepet töltenek be a regressziós paraméterek becslése során. A regressziós modell specifikációja során feltevéseket teszünk a modell változóira, amelyekről csupán elvárásaink, közelítő információink lehetnek. Célszerűnek tűnik, ha számba vesszük ezeket. Itt szintén alá kell húzni azt, hogy a regressziós modellt első megközelítésben hipotézisnek tekintjük, amit - a statisztikai adatok birtokában - vagy elfogadunk, vagy elvetünk. A becslés módszerének kiválasztása nagymértékben függ a valószínűségi változókra tett feltevésektől, amelyeket a következő táblában foglaltunk össze.

A klasszikus lineáris regressziós modell feltételei

7-4. tábla

Feltétel	ε	Y
1.	$E(\varepsilon_i \mid x_i) = 0$	$E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$
2.	$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ $i \neq j$	$\mathrm{cov}(Y_i, Y_j) = 0$ $i \neq j$
3.	$\mathrm{var}(\varepsilon_i \mid x_i) = \sigma^2$	$\mathrm{var}(Y_i \mid x_i) = \sigma^2$

Amint az látható, a táblában a feltételeket mindkét valószínűségi változóra vonatkozóan megfogalmaztuk, és sok hasonlóságot állapíthatunk meg közöttük.

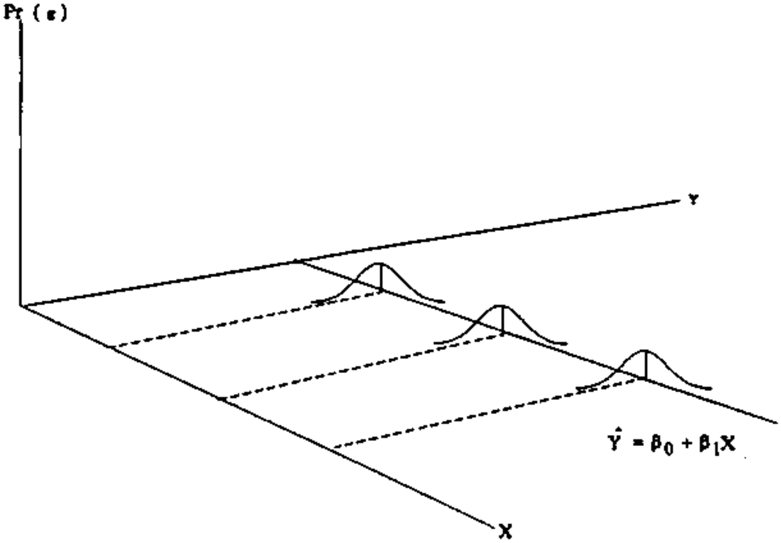
---

[6] Mivel tudjuk, hogy az X rögzített, fix változó.

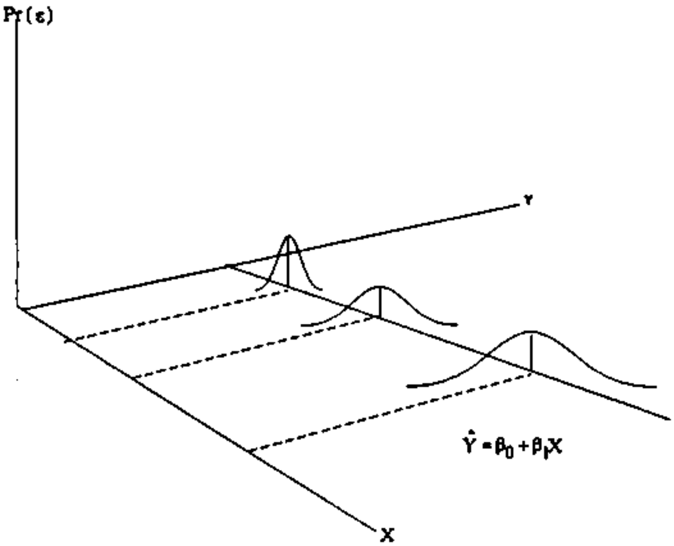
Az első feltétel szerint a **véletlen változó (ε) várható értéke nulla.** Amennyiben ez nem áll fenn, szisztematikus hatásra kell gyanakodni, ami specifikációs problémát takar. Ennek a feltételnek közvetlen következményeként az eredményváltozó (Y) feltételes várható értéke a lineáris kapcsolat (függvény) által kifejezhető konkrét x_i-ez rendelt Y_i értékkel egyenlő.

A második feltétel teljesülésének hiányát az **autokorreláció** jelenségeként szokták értelmezni. A klasszikus lineáris regressziós modell alapfeltevése szerint a **véletlen változó értékei páronként nem korrelálnak egymással.** Természetesen a szomszédos (vagy távolabbi) értékek függetlensége az Y eredményváltozóra is kiterjeszthető. Meg kell jegyezni, hogy az autokorreláció jelenlétére általában (de nem kizárólagosan) olyan esetekben kell számítani, amikor az ok-okozati viszonyokat feltáró kapcsolatrendszer modellezése során a változók megfigyelt értékei idősorokból származnak.

A harmadik feltétel fennállása esetén a modellt **homoszkedasztikusnak** nevezzük, míg ellenkező esetben a **heteroszkedaszticitás** jelenségével állunk szemben. A 7-1. táblában lévő szemléltető példában láttuk, hogy egy-egy x_i értékhez több Y_i tartozhat, tehát az értékek szóródnak. (Természetesen a korábban kifejtettek értelmében az Y_i szóródása a véletlen tényezőtől függ.) Amennyiben a **véletlen változók szórása állandó,** nem függ az x_i nagyságától, homoszkedasztikusnak tekintjük a modellt. Természetesen, ha a fentieket átgondoljuk, könnyű belátni, hogy mind az eredmény-, mind a reziduális változó esetén ugyanarról a szórásról van szó, amit feltételes szórásnak tekinthetünk. A jelenség megértését segíti az alábbi két ábra.



7-6./a ábra Homoszkedaszticitás



7-6./b ábra Heteroszkedaszticitás

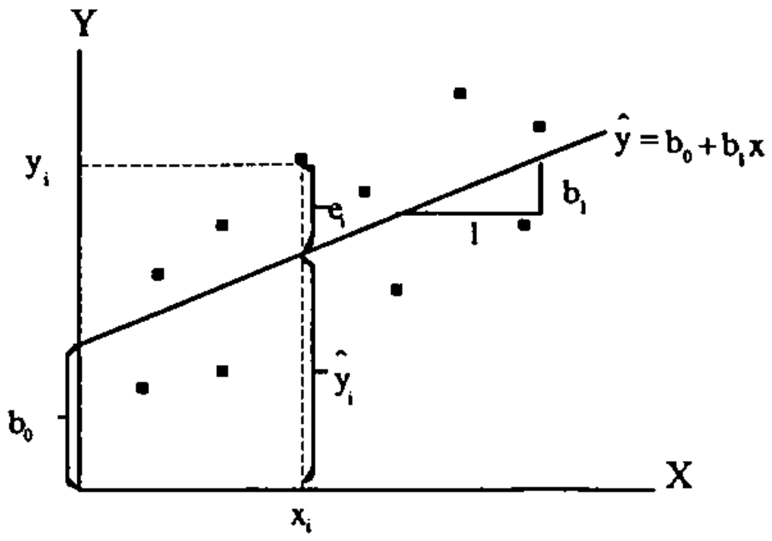
A fenti három feltétel együttes teljesülése esetén tekinthetjük a modellt **klasszikus lineáris regressziós modellnek,** és csak ekkor alkalmazhatjuk a paraméterek becslésére a legkisebb négyzetek módszerét.

A klasszikus legkisebb négyzetek módszere (KLNM)

A regressziós paraméterek becslése lineáris esetben nem más, mint az alábbi egyenes együtthatóinak meghatározása:

$$\hat{y}_i = b_0 + b_1 x_i \qquad (7\text{-}11)$$

A módszer a következő gondolatmenetre épül. Tudjuk, hogy az X és Y változóra vonatkozóan n számú megfigyeléssel n adatpárral rendelkezünk. Ezek az értékpárok a síkban n pontot határoznak meg. Önként adódik az a kérdés, hogy melyik az egyenes, amelyik ezekhez a pontokhoz legjobban illeszkedik? A probléma megválaszolását segíti az alábbi ábra.



7-7. ábra A klasszikus legkisebb négyzetek módszerének elve

Válasszunk ki egy tetszőleges x_i értéket és keressük meg a hozzá rendelhető eredményváltozó becsült és tényleges értékeit! Az y_i és $\hat{y}_i$ adatok általában eltérnek egymástól, amit e_i-vel jelölünk. Vegyük észre, hogy ez a különbség a hibatag, a reziduális változó empirikus megfelelője. A tapasztalati reziduumok előjele egyaránt lehet pozitív, illetve negatív, attól függően, hogy a tényleges értéket reprezentáló pont az egyenes felett vagy az egyenes alatt helyezkedik el. A feladat egy olyan egyenes meghatározása, amely a legjobban illeszkedik a ponthalmazhoz. Logikus, hogy az e_i különbségek abszolút értéke annál kisebb, minél közelebb van az egyenes a megfigyelt értékekhez. Vizsgálatunk szempontjából a pontoktól való eltérés iránya érdektelen, csupán a nagysága fontos. Kézenfekvő megoldásnak tűnik, ha az e_i négyzetekkel dolgozunk a továbbiakban, ezáltal a hibákat "súlyozva" vesszük figyelembe. Mivel a valamennyi megfigyelt értékhez legjobban illeszkedő egyenest keressük, az **eltérések négyzetösszegének minimuma** mellett nyerjük a legjobb megoldást.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \text{minimum} \qquad (7\text{-}12)$$

A fenti összefüggést alakítsuk át a (7-11) felhasználásával! Mivel x_i és y_i mintabeli értékek adottak, a négyzetösszeg a függvény megválasztásától függ, azaz a négyzetösszeg kizárólag b_0 és b_1 függvénye.

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \qquad (7\text{-}13)$$

A klasszikus legkisebb négyzetek módszerének logikája szerint a lineáris regressziófüggvényt azzal az egyenessel becsüljük, amelyre nézve a (7-13) négyzetösszeg minimális. Tulajdonképpen egy többváltozós szélsőérték számítási feladat megoldásával juthatunk a keresett függvényhez. A (7-13) függvénynek csak akkor lehet szélső értéke, ha ott mindkét parciális derivált nulla. Esetünkben a szélső érték minimum, mivel a második derivált pozitív. A parciális deriváltak b_0 és b_1 szerint:

$$\frac{\partial f}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) \qquad (7\text{-}14)$$

$$\frac{\partial f}{\partial b_1} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i$$

Ha a fenti parciális deriváltakat egyenlővé tesszük nullával és célszerűen átalakítjuk, a paraméterek kiszámítására az alábbi ún. **normálegyenleteket** nyerjük:

$$\sum_{i=1}^{n} y_i = b_0 n + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \qquad (7\text{-}15)$$

Ezek, és a belőlük származtatható formulák tulajdonképpen a **paraméterekre vonatkozó becslőfüggvények.**

Mivel az x_i és y_i, valamint a szükséges összegek, keresztszorzatok összegei és négyzetösszegek ismertek, a normálegyenletek kétismeretlenes első fokú egyenletrendszert alkotnak, ahol az ismeretlenek (b_0, b_1) a regressziós együtthatók pontbecslései. Könnyen adhatunk numerikus megoldást az egyenletrendszerrel is, de előállítható egy olyan célszerű transzformáció, amelyből közvetlenül nyerhetők a paraméterek becslőfüggvényei.

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum \left(x_i - \bar{x}\right)^2}$$
$$b_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \bar{y} - b_1 \bar{x} \qquad (7\text{-}16)$$

A b_1 regressziós együttható képletét szemlélve feltűnik a "rokonsága" a lineáris korrelációs együtthatóval. Megfelelő átalakítás után a két képlet:

$$r_{yx} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_y \sigma_x}$$
$$b_1 = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x^2} \qquad (7\text{-}17)$$

Ebből az összehasonlításból leolvasható a lineáris korrelációs együttható és a regressziós együttható közötti összefüggés:

$$b_1 = r_{yx} \frac{\sigma_y}{\sigma_x}, \text{ illetve } r_{yx} = b_1 \frac{\sigma_y}{\sigma_x} \qquad (7\text{-}18)$$

Az analitikus regressziófüggvény a (7-16) és a (7-18) összefüggés felhasználásával célszerűen átrendezhető az alábbiak szerint:

$$\begin{aligned}\hat{y} &= b_0 + b_1 x = \bar{y} - b_1 \bar{x} + b_1 x \\ &= \bar{y} + b_1 (x - \bar{x}) = \bar{y} + r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x})\end{aligned} \qquad (7\text{-}19)$$

Szembetűnő a hasonlóság a (7-6)-ban közölt összefüggéssel. Itt a valószínűségelméletben használt szimbólumok helyett a megfelelő statisztikai mutatószámok állnak. Általánosságban is megfogalmazhatjuk, hogy kétdimenziós normális eloszlás esetén a regresszió és a klasszikus legkisebb négyzetek módszerével számított analitikus regresszió egymással megegyezik.

Tudjuk, hogy a regressziószámítás kapcsán felhasznált adatbázist - az esetek döntő többségében - mintának tekintjük. A klasszikus legkisebb négyzetek módszere egy olyan eljárás, amely végül nagyon jó tulajdonságokkal rendelkező becslőfüggvényeket eredményez. Bebizonyítható ugyanis az, hogy a paraméterek becslése torzítatlan becslés, és ugyanakkor a regressziófüggvénnyel is torzítatlanul becsülhető a feltételes várható érték.

$$E(b_0) = \beta_0 \quad \text{és} \quad E(b_1) = \beta_1 \qquad (7\text{-}20)$$
$$E(\hat{y}_i) = E(Y|x_i)$$

Az ún. Gauss-Markov-tétel értelmében a b_0 és b_1 az alap-regresszió paramétereinek legjobb lineáris torzítatlan becslései, abban az értelemben, hogy szórásuk (standard hibájuk) a legkisebb.

Amennyiben a véletlen, reziduális változó normális eloszlást követ, az ún. **maximum likelihood becsléssel** is a klasszikus legkisebb négyzetek segítségével megismert becslőfüggvényeket kapjuk.

Természetesen bármilyen kedvező képet is festhetünk a klasszikus legkisebb négyzetek módszerével történő becslésről, soha nem szabad megfeledkezni arról, hogy az eredmény minden esetben csak becslés. A regressziós modellekkel nyerhető becslések pontosságának, megbízhatóságának vizsgálata ezért fokozott hangsúlyt kap. A pontosság vizsgálata előtt azonban néhány szót szólunk az eredmények szakmai értékeléséről.

A paraméterek értelmezése

A b_0 paraméter a geometriai értelmezésből adódóan az a hely, ahol az egyenes metszi az Y tengelyt, más megfogalmazásban az Y változónak az értéke, amelyet X=0 helyen vesz fel. Ebből következik, hogy értelmezése attól függ: a nulla beletartozik-e az X értékek halmazába, illetve logikailag az értelmezési tartomány részének tekinthető-e? Egy mezőgazdaságból vett példával világíthatjuk meg a fent elmondottakat. Valamely növény terméshozamára, mint függő-, vagy eredményváltozóra hat például a felhasznált műtrágya mennyisége. Ilyen esetben a b_0 paraméter arra ad választ, hogy mekkora termésátlag várható abban az esetben ha nem használnak fel műtrágyát. Ha azonban a termésátlag és a munka-időráfordítás kapcsolatát vizsgálnánk, a b_0 paraméternek nem tulajdoníthatunk tárgyi tartalmat, mivel a munkaidő-ráfordítás nulla értéke nem értelmezhető.

A b_1 regressziós paraméter geometriai értelemben az egyenes meredekségét meghatározó iránytangens, azaz $\frac{dy}{dx}$. A regressziós együttható segítségével megállapíthatjuk, hogy egy korrelációs kapcsolat esetén a tényezőváltozó egységnyi növekedése milyen átlagos változást idéz elő az eredményváltozóban. Vegyük észre ugyanis, hogy X értékének egy egységgel való növelése Y nagyságát várhatóan b_1 egységgel változtatja meg (növeli vagy csökkenti a kapcsolat irányától függően). A korreláció hiánya esetén b_1 értéke nulla, míg pozitív irányú kapcsolat esetén b_1 előjele pozitív, illetve negatív korreláció esetén b_1 negatív előjellel bír.

A közgazdasági elemzések során gyakorta találkozhatunk azzal az esettel, amikor a nemcsak a változók abszolút különbségeire, hanem relatív változásaira vagyunk kíváncsiak. A relatív (százalékos) változások sokszor egyértelműbbé, a vizsgálat szempontjából kezelhetőbbé, összehasonlíthatóbbá teszik az eredményeket. Mindez az analízisből és mikroökonómiából jól ismert elaszticitás fogalmához vezet. Az **elaszticitás (rugalmasság)** arra ad választ, hogy az egyik változó relatív változása a másik változó milyen mértékű relatív változását indukálja várhatóan.

Jelöljük X változó egy tetszőleges értékét x-szel, Y változó megfigyelt értékét y-nal! Egy adott helyről kiindulva x értékét Δx-szel növelve a függvénynövekmény nagysága Δy. Amennyiben a **relatív növekmények hányadosát** képezzük a rugalmassági együtthatóhoz jutunk:

$$\frac{\Delta y}{y} \div \frac{\Delta x}{x} \qquad (7\text{-}21)$$

Mivel a (7-21)-ben definiált hányados nem független Δx nagyságától, célszerűnek tűnik ha a rugalmasság mutatószámát határátmenet segítségével értelmezzük. Az így képzett mutató **Y-nak X-re vonatkozó elaszticitási együtthatója:**

$$El(y,x) = \lim_{\Delta x \to 0} \left(\frac{\Delta y}{y} \div \frac{\Delta x}{x} \right) \qquad (7\text{-}22)$$

Ha a regressziófüggvény differenciálható, az elaszticitási együttható kifejezhető a regressziós függvény segítségével:

$$El\left(\hat{y}, x\right) = \lim_{\Delta x \to 0} \left(\frac{\Delta y}{\Delta x} \div \frac{y}{x} \right) = \frac{dy}{dx} \div \frac{y}{x} = b_1 \frac{x}{y} \qquad (7\text{-}23)$$

tehát függ X nagyságától.

Egy függvény elaszticitását a változók logaritmikus transzformációjával is kifejezhetjük. Legyen v=log y és u=log x; a v-nek u szerinti differenciálhányadosa nem más, mint y-nak x-re vonatkozó elaszticitási együtthatója:

$$El(y,x)=\frac{dv}{du} \tag{7-24}$$

Az elaszticitási együttható számítható x átlagos értéke mellett, de bármely más konkrét x esetén is.

Tételezzük fel, hogy a korábban említett kereskedelmi vállalatnál a tíz különféle alapterülethez egy véletlen mintát rendeltek, amely tíz üzlet éves forgalmának adatait tartalmazza. (A 7-3. tábla tartalmazza ezeket az adatokat!) A rendelkezésre álló véletlen minta segítségével kívánjuk megbecsülni az alap-regresszió paramétereit, azaz az alapterület és az éves forgalom közötti sztochasztikus összefüggést leíró modell együtthatóit. A kiinduló adatokat és a részeredményeket az alábbi tábla tartalmazza.

Munkatábla

7-5. tábla

x_i	y_i	$x_i y_i$	x_i^2	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y-\bar{y})$
100	12	1.200	0.000	50.625	7.042,5
150	10	1.500	22.500	30.625	5.827,5
200	24	4.800	40.000	15.625	2.412,5
250	30	7.500	62.500	5.625	997,5
300	26	7.800	90.000	625	432,5
350	50	17.500	122.500	625	167,5
400	41	16.400	160.000	5.625	-172,5
450	60	27.000	202.500	15.625	2.087,5
500	80	40.000	250.000	30.625	6.422,5
550	100	55.000	302.500	50.625	12.757,5
3.250	433	178.700	1.262.500	206.250	37.975,0

A fenti tábla adatai lehetőséget nyújtanak a regressziós paraméterek meghatározására.

$$b_1 = \frac{10 \times 178.700 - 3.250 \times 433}{10 \times 1.262.500 - 3.250^2} = \frac{37.975}{206.250} = 0,1841 \text{ illetve}$$

$$b_0 = 43,3 - 0,1841 \times 325 = -16,54$$

Az paraméterek birtokában felírhatjuk a minta-regressziós függvényt:

$$\hat{y}_i = -16,54 + 0,1841 x_i$$

Az abszolút változás mérőszáma mellett természetesen a rugalmasság is értelmezhető:

$$El(\bar{y}, \bar{x}) = 0,1841 \frac{325}{43,3} = 1,38$$

Példánkban a b_0 paraméternek nem tulajdoníthatunk tárgyi értelmet, mivel az alapterület nem lehet nulla.

A b_1 regressziós együttható közgazdasági tartalommal bír. A függvény szerint az 1 m^2-rel nagyobb alapterület átlagosan 0,1841 MFt-os, azaz mintegy 184 EFt évi forgalmi többletet eredményez.

Természetesen az elaszticitásnak is tulajdoníthatunk közgazdasági tartalmat. Mind az x, mind az y értékek helyébe a mintabeli adatok átlagát írva, eredményül az átlagos elaszticitást számszerűsítettük. Átlagos alapterületű boltok esetén az alapterület 1 %-os növekedése (többlete), az éves forgalmat 1,38 %-kal növeli átlagosan. Az alapterület és a boltok forgalma között rugalmas kapcsolatot számszerűsítettünk, mivel az elaszticitási együttható 1 %-nál nagyobb, azaz az alapterület változására igen gyorsan és erőteljesen reagál a forgalom.

Minden megállapításunk kapcsán hallgatólagosan feltételeztük, hogy a körülmények, a külső tényezők változatlanok, illetve ezek megváltozásával a következtetések megfogalmazása során nem számoltunk.

A regressziós együtthatóból levonható információ minőségét nagymértékben behatárolja a változók közötti korrelációs kapcsolat szorossága. Itt is kiszámolhatjuk a korrelációs együtthatót, amelynek értéke r=0,945; tehát, szoros pozitív irányú a kapcsolat.

Hipotézisellenőrzés, intervallumbecslés

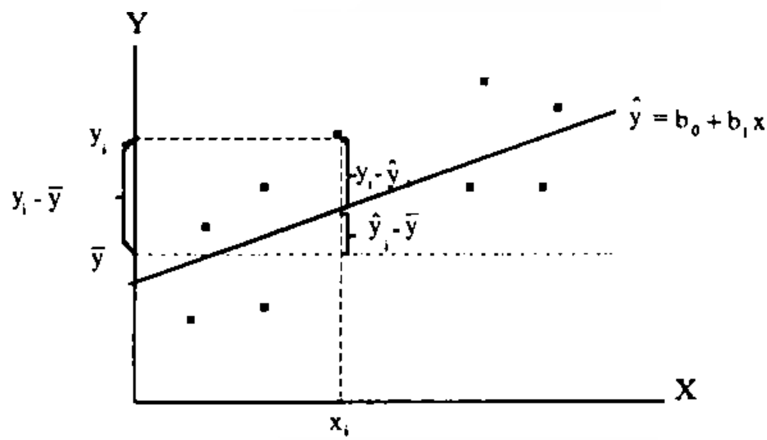
Tudjuk, hogy a megfigyelt Y értékek eltérését az $\hat{Y}$ regressziós becsléstől az ε_i véletlen, reziduális változó tapasztalati megfelelője a reziduum (e_i) reprezentálja. Ez a változó igen fontos szerepet tölt be a klasszikus legkisebb négyzetek módszerével becsült lineáris regressziós modell vizsgálatában. A megfigyelt, tapasztalati adatokra értelmezett regressziós azonosság:

$$y_i = \hat{y}_i + e_i \qquad (7\text{-}25)$$

Kiinduló feltételként elfogadhatjuk, hogy a reziuális változó értékeinek összege és így átlaga is nulla. (Csak a tiszta véletlen játszik szerepet.) Így nem tűnik meglepőnek, ha kijelentjük, hogy az $\hat{y}_i$ regressziós becslések összege (és átlaga) megegyezik a megfigyelt y_i értékek, az eredményváltozó összegével (illetve átlagával). Ezek előrebocsátása mellett vonjuk ki a (7-25) azonosság mindkét oldalából a közös átlagot!

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + e_i \qquad (7\text{-}26)$$

Az eltérés összetevőkre bontását az alábbi ábrával szemléltetjük.



7-8. ábra A függő változó átlagtól való eltérésének felbontása

Emeljük négyzetre a (7-26) kifejezés mindkét oldalát, és végezzük el az összeg-zéseket!

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2 + 2\sum (\hat{y}_i - \bar{y}) e_i \qquad (7\text{-}27)$$

Feltételezéseink szerint a reziduális változót kifejező reziduum nem korrelál az eredményváltozóval, valamint tudjuk, hogy a várható értéke nulla, a (7-27) kifejezés utolsó tagja értelemszerűen nullával lesz egyenlő. Ezek alapján a négyzetösszegek közötti összefüggés az alábbi:

SST SSR SSE

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum e_i^2 \qquad (7\text{-}28)$$

Ez az összefüggés egy kis átalakítással az alábbi formában is felírható (amire egy későbbi összefüggés megértése miatt lesz szükségünk):

$$\sum(y_i - \bar{y})^2 = b_i^2 \sum(x_i - \bar{x})^2 + \sum e_i^2 \qquad (7\text{-}29)$$

ugyanis $(\hat{y}_i - \bar{y}) = b_1(x_i - \bar{x})$.

Mivel itt és a (7-28)-ban eltérés-négyzetösszegekről van szó, használjuk az alábbi egyszerű jelölésmódot!

$$S_y = S_{\hat{y}} + S_e \qquad (7\text{-}30)$$

Az eltérés-négyzetösszegek ismerete lehetővé teszi egyrészt a korrelációs együttható és a lineáris regressziós fogalmak összekapcsolását, másrészt alapul szolgál a regresszióval kapcsolatos hipotézisellenőrzések végrehajtásához.

Amennyiben a (7-30) összefüggést n-nel végigosztjuk, a szórásnégyzet összetevőkre bontásának a regresszióanalízisben megfelelő eljárásához jutunk.

$$s_y^2 = s_{\hat{y}}^2 + s_e^2 \qquad (7\text{-}31)$$

A bal oldali tag a megfigyelt y értékek szórásnégyzete. A kifejezés jobb oldalán az első tag a becsült értékek szórásnégyzete, a regresszió által megmagyarázott rész, a **külső szórásnégyzet** szerepét tölti be; a második tag, a reziduális szórásnégyzet, képviseli az eredményváltozó ingadozásának az X változó által meg nem magyarázott részét, azaz a **belső szórásnégyzetet**. A korrelációs mérőszámok megismerése során már találkoztunk a fenti gondolatmenettel, amiből könnyen származtathatjuk azt a determinációs együtthatót, amely a lineáris korrelációs együttható négyzetének felel meg.

$$r^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2} \qquad (7\text{-}32)$$

A korábbi eszmefuttatások során láttuk, hogy a regresszióanalízis eljárásai becsléseknek tekinthetők, amelyek a rendelkezésre álló adatbázis (minta) birtokában közelítik az alap (elméleti) értékeket. A regressziós modellekkel nyerhető becslések pontosságának, ezáltal a modell megbízhatóságának a vizsgálata ezért különös hangsúllyal bír. A pontossági vizsgálatokhoz elméleti hátteret szolgáltat a KLNM-nek néhány tulajdonsága. Ezek szerint a KLNM-vel nyert becslések - amennyiben a normalitás feltételei fennállnak - torzítatlanok; minimális hibával rendelkező, azaz efficiens (abszolut hatásos) becslések; valamint konzisztens becslések. A regressziós paraméterek normális eloszlásúak (kis minta esetén azonban a Student-féle t-eloszlást kell használni). Az eredményváltozó (valamint ezáltal a reziduális változó) állandó szórással rendelkezik, a becsült[7] és tényleges variancia segítségével felírható $(n-2)s^2/\sigma^2$ hányados n-2 szabadságfokú χ^2 eloszlás. A regressziós együtthatók eloszlása független a varianciától.

A modellezés során lényeges kérdés az, hogy a magyarázó változó(k) valóban szignifikáns kapcsolatban legyenek az eredményváltozóval. A választ elsőként a **varianciaanalízis** regressziószámításban történő alkalmazásával közelítjük meg. A varianciaanalízis alkalmazását lehetővé teszi az eltérésnégyzet-összegek felbontásának

[7] A becsült varianciát s^2 jelöli.

(7-30)-ban közölt módja, az eltérésnégyzet-összegek alapján ugyanis elvégezhető a variancia (σ^2) különböző irányból történő becslése.

Varianciaanalízis a kétváltozós regressziós modellhez

7-6. tábla

Összetevő	Négyzetösszeg	Szabadságfok	Szórásnégyzet becslése
Regresszió	$S_{\hat{y}} = \sum(\hat{y}_i - \bar{y})^2$	1	$s_{\hat{y}} = S_{\hat{y}} / 1$
Maradék (reziduum)	$S_e = \sum(y_i - \hat{y}_i)^2$	n-2	$s^2 = S_e/(n-2)$
Teljes	$S_y = \sum(y_i - \bar{y})^2$	n-1	$s_y = S_y / n-1$

A szabadságfokokat, a (7-29) összefüggés alapján, az alábbi várható értékek determinálják:

$$E(S_y) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2 \qquad (7\text{-}33)$$
$$E(s^2) = \sigma^2$$

A hibatag, reziduális változó szórását nem ismerjük, azt a maradék négyzetösszegnek, a reziduumok négyzetösszegének (n-2)-vel való osztásával becsüljük torzítatlanul. A hibatag varianciájának ismerete kiemelkedően fontos szereppel bír mind a hipotézisellenőrzések, mind az intervallumbecslések során.

A varianciaanalízis táblájához szervesen kapcsolódnak a hipotézisek. Kétváltozós regresszió esetén a hipotézisek csupán a β_1 regressziós együtthatóra vonatkoznak.

$$H_0 : \beta_1 = 0 \qquad (7\text{-}34)$$
$$H_1 : \beta_1 \neq 0$$

A nullhipotézis a lineáris regresszió fennállásának tagadását jelenti és amennyiben igaz, úgy az eredményváltozó kizárólag a véletlen hatására szóródik; az alternatív hipotézis fennállása esetén a regressziós modellt elfogadhatónak ítéljük. A nullhipotézis ellenőrzésére az

$$F = \frac{S_{\hat{y}}}{S_e / (n-2)} \tag{7-35}$$

ahol: $v_1 = 1$

$v_2 = (n-2)$

próbafüggvény alkalmas.

A boltok forgalmát taglaló példánknál maradva ellenőrizzük a regressziós modellt a varianciaanalízis segítségével! A számításokat az alábbi tábla segítségével lehet követni.

Munkatábla

7-7. tábla

Összetevő	Négyzetösszeg	Szabadságfok	Variancia-becslés
Regresszió	6.992,0	1	6.992,0
Reziduum	836,1	8	104,5
Teljes	7.828,1	9	----

F=66,909

Látható, hogy 5 %-os szignifikancia szinten a nullhipotézist el kell vetnünk, tehát a regressziós modell a valóságban létezik.

A regressziós együtthatók szórásai a becslőfüggvények szórásainak, standard hibáinak felelnek meg. A standard hibák ismerete lehetővé teszi az egyes paraméterek külön-külön történő tesztelését. A regressziós paraméterek standard hibái tartalmazzák a reziduális szórás torzítatlan [(n-2)-vel számított] becslését, amit **korrigált reziduális szórásnak** nevezünk.

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

A β_0 paraméter **standard hibájának** becslése:

$$s_{b_0} = s\sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}} \qquad (7\text{-}36)$$

A β_1 paraméter **standard hibájának** becslése:

$$s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \qquad (7\text{-}37)$$

Ha a modellben szereplő véletlen változó (ε) illetve az eredményváltozó (Y) normális eloszlású, az alábbi statisztika

$$t = \frac{b_1}{s_{b_1}} \qquad (7\text{-}38)$$

n-2 szabadságfokú t-eloszlást követ. Ezzel a próbafüggvénnyel választ lehet adni az alábbi hipotézisekre:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Hasonlóan képezhetjük a β_0 paraméter ellenőrzésére szolgáló próbafüggvényt, a b_0 és standard hibájának hányadosaként.

A fenti t-próba a β_1 regressziós együttható parciális tesztelését szolgálja. A t- és F-próbafüggvény között felírható összefüggés[8]:

$$F = t^2 \qquad (7\text{-}39)$$

Amennyiben az F-próbafüggvényt alkalmazzuk, természetesen **1 és (n-2) szabadságfokú F-eloszlást** kell a hipotézisek ellenőrzésére felhasználni.

Ellenőrizzük számpéldánkban a regressziós paramétereket! A számításokhoz szükséges adatokat a 7-5. és 7-7. táblában találjuk meg.

$$s = \sqrt{\frac{836,1}{8}} = 10,22$$

$$s_{b_1} = \frac{10,22}{\sqrt{206.250}} = 0,0225$$

$$s_{b_0} = 10,22\sqrt{\frac{1.262.500}{10 \times 206.250}} = 7,99$$

A standard hibák módot adnak arra, hogy segítségükkel ellenőrizzük a regressziós paraméterekre vonatkoztatott hipotéziseket.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

$$t_{b_1} = \frac{0,1841}{0,0225} = 8,182$$

$$t_{0,05(8)} = 2,306$$

[8] Bővebben lásd! Irodalom [38]

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$
$$t_{b_0} = \frac{-16{,}54}{7{,}99} = -2{,}06$$
$$t_{0.05(8)} = 2{,}306$$

A fenti számításokból megállapítható, hogy a β_1 paraméter szignifikáns 5 %-os szignifikancia szinten. Ugyanez a β_0 paraméterre vonatkozóan ezen a szignifikancia szinten nem mondható el.

Az ökonometriai modellekben gyakorta szignifikánsnak tekintik a paramétereket és ezáltal az adott változót is, amennyiben a t-statisztika meghaladja az 1-et, ami azt is jelenti, hogy legalább a paraméterek előjele biztos.

Láttuk, hogy a regressziós együtthatók becslése pontbecslésnek minősíthető. A standard hiba birtokában lehetőség nyílik az intervallumbecslésre, azaz konfidencia intervallumot szerkeszthetünk az alap regressziós paraméterre. Egy tetszőleges β_j paraméterre vonatkozóan a következő valószínűségi megállapítást tehetjük:

$$\Pr\left(b_j - t_{\alpha/2}s_{b_j} < \beta_j < b_j + t_{\alpha/2}s_{b_j}\right) = 1 - \alpha \qquad (7\text{-}40)$$

A fenti összefüggésből közvetlenül adódik β_j regressziós paraméter konfidencia intervalluma 1-α valószínűségi szinten.

$$b_j \pm t_{\alpha/2}s_{b_j} \qquad (7\text{-}41)$$

Meg kell jegyeznünk, hogy amennyiben nagy mintánk van, a t-eloszlás helyett normális eloszlást használhatunk, a standard normális eloszlás megfelelő értékének behelyettesítése útján.

Számpéldánkban megszerkeszthetjük a regressziós együtthatók konfidencia intervallumait.

$$-16,54 \pm 2,306 \times 7,99$$
$$-34,96 - - - -1,88$$
$$0,1841 \pm 2,306 \times 0,0225$$
$$0,1321 - - - 0,2361$$

Látható, hogy a b_0 paraméter viszonylag magas standard hibája azt is eredményezi, hogy a konfidencia intervallum negatív értéktől pozitív értékig húzódik, igen széles sávban. A b_1 paraméterhez rendelt konfidencia intervallumból viszont leolvashatjuk azt, hogy a külső tényezők változatlanságának feltételezése mellett az üzletek területének 1 m^2-es növekedése várhatóan legalább 132 EFt, illetve legfeljebb 236 EFt-os évi átlagos forgalom növekedést indukál, 95 %-os megbízhatósági szinten.

A regressziós függvénybe egy konkrét x_0 értéket behelyettesítve a regresszió átlagos színvonalát becsülhetjük meg. A becslés a regressziófüggvény felhasználásával végezhető el, ami ugyancsak torzítatlan lesz:

$$\hat{y}_0 = b_0 + b_1 x_0$$

ahol: x_0 - a tényezőváltozó rögzített értéke

Ez az átlagos színvonal becslése a regresszióanalízisnek igen fontos felhasználási területe. (Egy adott alapterületű üzlethez megbecsülhetjük az átlagos forgalmat pl. ha x_0=350, a várható évi forgalom 47,9 MFt.) Általában a rögzített értékeket a regressziófüggvény értelmezési tartományából szoktuk választani, de nem meglepő, ha ettől eltérő értékre is kíváncsi az elemző. Ilyen esetben extrapolációról, prognózisról beszélhetünk, aminek az értelmezése fokozott figyelmet igényel.

A becsült átlagos színvonal standard hibájának képlete:

$$s_{\hat{y}} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \qquad (7\text{-}42)$$

A fenti képletből látható, hogy amennyiben a rögzített érték megegyezik az X változó megfigyelt értékeinek számtani átlagával a standard hiba minimális lesz.

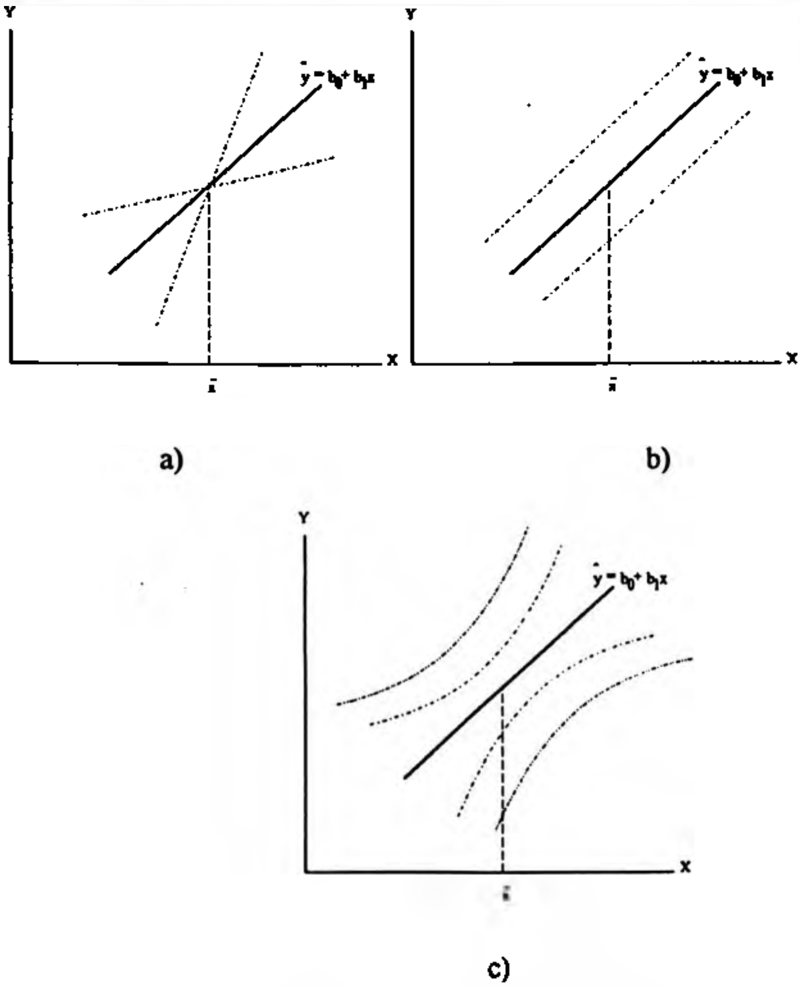
$$s_{\hat{y}} = \frac{s}{\sqrt{n}}$$

$$\text{mivel } x_0 = \bar{x}$$

Előfordulhat, hogy nem az átlagos szintet, hanem egy hiányzó egyedi adatot kívánunk becsülni. Ehhez is az $\hat{y}_0$ becslőfüggvényt (a regressziófüggvényt) használjuk fel, a két pontbecslés megegyezik, azonban a becslés standard hibája az előzőhöz képest nagyobb lesz.

$$s_{\hat{y}}' = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \qquad (7\text{-}43)$$

A standard hibák segítségével (adott valószínűség mellett) létrehozható konfidencia sávok sémáját, az eredményváltozóra vonatkozóan az alábbi ábrákon szemléltetjük:



7-9. ábra A konfidencia sávok sémái

Az a) ábra azt az esetet szemlélteti, amikor csak a b_1 paraméter rendelkezik hibával; a b) ábra a b_0 paraméter önálló hibasávját reprezentálja; míg a c) ábra a feltételes várható értékre (átlagokra) illetve az egyedi értékekre szerkesztett hibasávokat reprezentálja. Mindez jól szemlélteti azt, hogy milyen szerepet játszanak a különböző paraméterek becslési hibái Y értékének becslése során.

A példa adatait felhasználva számítsuk ki a 350 m^2-es üzlet várható éves forgalmát. A regresszió függvénybe behelyettesítve az átlagos érték becsült értéke: 47,9 MFt. Az átlagos érték standard hibája:

$$s_{\hat{y}} = 10{,}22\sqrt{\frac{1}{10} + \frac{(350 - 325)^2}{206.250}} = 3{,}28$$

A hibahatár 95 %-os valószínűségi szinten 7,56; ezek alapján a t-statisztika felhasználásával legalább 40,3 MFt, legfeljebb 55,4 MFt forgalom várható 350 m^2 -es üzlet esetén. Amennyiben az egyedi érték becslése lenne a cél a hibahatár értéke: 10,7 MFt, ami lényegesen szélesebb intervallumot eredményez az előbbinél.

Természetesen mind az átlagos színvonal, mind az egyedi értek becslése során kapott számértékek alkalmasak arra, hogy segítségükkel a paramétereknél megismert módon ellenőrizzük a modellre megfogalmazható hipotéziseinket.

A változók felcserélése

A gyakorlati alkalmazás során elsősorban azt kell megvizsgálni, hogy a változók kapcsolatában melyek azok a szignifikáns elemek, amelyek oknak tekinthetők és alapvető hatást gyakorolnak arra a változóra, amelyik az okozat szerepét tölti be. Előfordulhat az, hogy a hatások kétirányúak, például két változó esetén mindkét változó egyaránt betöltheti az ok és az okozat szerepét. A gyakorlati szempontok és érvek ilyen esetekben szükségessé tehetik a változók felcserélését. természetesen tágabban a felcserélhetőség igen fontos gyakorlati információkat eredményez mind a modellek szerkesztése, mind az elemző munka során is.

A változók felcserélésének gondolatmenetét a már megismert kétváltozós lineáris regressziós kapcsolatból kiindulva érthetjük meg jól. A jelenséget leíró regressziófüggvényt írjuk fel oly módon, hogy a regressziós paraméterek jelöljék az ok és okozat jegyeit is.

$$\hat{y}_i = b_{0(Y/X)} + b_{1(Y/X)}x_i \tag{7-44}$$

A változók szerepének felcserélése során a tényezőváltozó szerepét a korábbi eredményváltozó veszi fel és fordítva. Ily módon az új regressziófüggvény:

$$\hat{x}_i = b_{0(X/Y)} + b_{1(X/Y)}y_i \tag{7-45}$$

A klasszikus legkisebb négyzetek módszere alapján lehetőség nyílik a (7-44) paraméterbecslésének megfelelően a (7-45) egyenlet paramétereinek meghatározására az alábbi módon:

$$b_{1(X/Y)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(y_i - \bar{y})^2} \tag{7-46}$$

$$b_{0(X/Y)} = \bar{x} - b_{1(X/Y)}\bar{y}$$

> Példánkban a változók szerepének felcserélése nem ütközik nagy nehézségbe, mivel ugyan az üzletek alapterülete a forgalom meghatározó tényezője, azonban az is indokolt kérdés lehet, amely arra ad választ, hogy egy adott nagyságú forgalom, milyen területen zajlik le. A "fordított" módon számított regresszió:
>
> $$\hat{x}_i = 110{,}7 + 4{,}85y_i$$
>
> A b_1 paraméter (regressziós együttható) alapján megállapíthatjuk, hogy 1 MFt-nyi forgalomnövekedés átlagosan és várhatóan 4,85 m^2 alapterület-növekedést igényel.

A kétféle módon számított b_1 regressziós együttható között reciprok viszony áll fenn. Mindezt könnyen beláthatjuk, ha az eredeti gondolatmenetnek megfelelően (ahol az Y volt az eredményváltozó) módosítjuk a (7-45) összefüggést. Az átalakítást úgy oldhatjuk meg, ha a $b_{1(X/Y)}$ regressziós együtthatóval végigosztjuk az egyenletet és y-ra rendezzük azt.

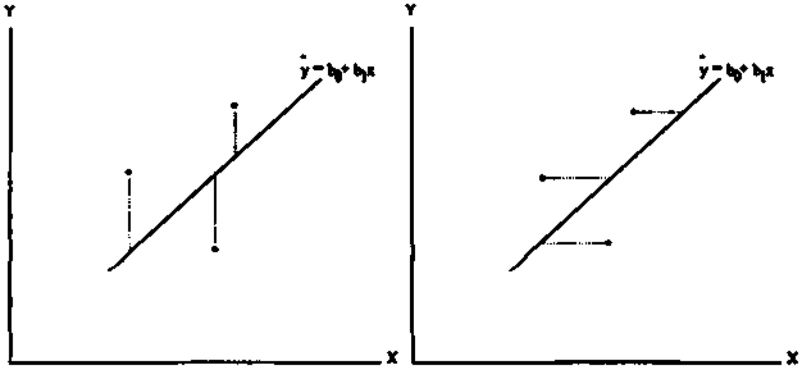
$$\hat{y}_i = -\frac{b_{0(X/Y)}}{b_{1(X/Y)}} + \frac{1}{b_{1(X/Y)}} x_i \qquad (7\text{-}47)$$

A kereskedelmi egységek forgalmával foglalkozó példánkban a fenti átalakítás után a becsült regressziófüggvény:

$$\hat{y}_i = 22{,}8 + 0{,}2062 x_i$$

Ebben a megközelítésben az 1 m^2-rel nagyobb bolti alapterület átlagosan mintegy 206 EFt-tal nagyobb forgalmat indukál. Az eredeti modellünkben ez az érték 184 EFt volt.

A példából látható, hogy a különféle megközelítési módon kiszámított paraméterek - az azonos dimenzióra vetítve - egymástól eltérnek. Az eltérésnek könnyen adhatjuk a magyarázatát, ha arra gondolunk, hogy az eredeti szemléletnek megfelelően a pontoknak az egyenestől mért távolságát a (7-45)-nek megfelelve vertikálisan, míg a (7-47) összefüggés logikáját követve horizontálisan mértük fel. Mindezt jól szemlélteti az alábbi ábra:



a) vertikális b) horizontális

7-10. ábra A pontok és az egyenes viszonya

Ha a kétféle módon számított regressziófüggvényt egy koordináta-rendszerben ábrázolnánk, egymást metsző egyeneseket kapnánk. Azonban, ha a kapcsolat függvényszerű, akkor joggal várhatnánk el, hogy a kétféle módon számított függvény egymástól ne különbözzön. Függetlenség esetén viszont joggal várható el, hogy a két egyenes egymásra merőleges legyen. Amennyiben a kapcsolat sztochasztikus - és ez a jellemző -, a két egyenes hegyesszöget zár be. Könnyen belátható, hogy a kétféle módon számított regressziós együtthatók és a determinációs együttható között az alábbi kapcsolat áll fenn:

$$r^2 = b_{1(Y/X)} b_{1(X/Y)} \qquad (7\text{-}48)$$

A kétféle megközelítést végiggondolva önkéntelenül adódik az az ötlet is, hogy a pontok távolságát az egyeneshez viszonyítva merőlegesen értelmezzük. Ebben az esetben azt az egyenest fogadjuk el a lehetséges egyenesek közül, amelyiknek a geometriai távolsága átlagosan a legkisebb. Az így számított függvényt **ortogonális regressziófüggvénynek** hívjuk. A ortogonális regressziófüggvény paramétereinek meghatározása a már megismert módszernél bonyolultabb algoritmusra épül, ami nem képezi tananyagunk tárgyát. Végezetül azt kell még elmondanunk, hogy az ortogonális regressziófüggvény b_1 együtthatójának értéke a vertikális illetve a horizontális regresszió paramétereinek értéke között foglal helyet. A fenti összefüggések természetének megemlítése nagymértékben ráirányítja a figyelmet az ok- és okozati kapcsolatok körültekintő meghatározására.

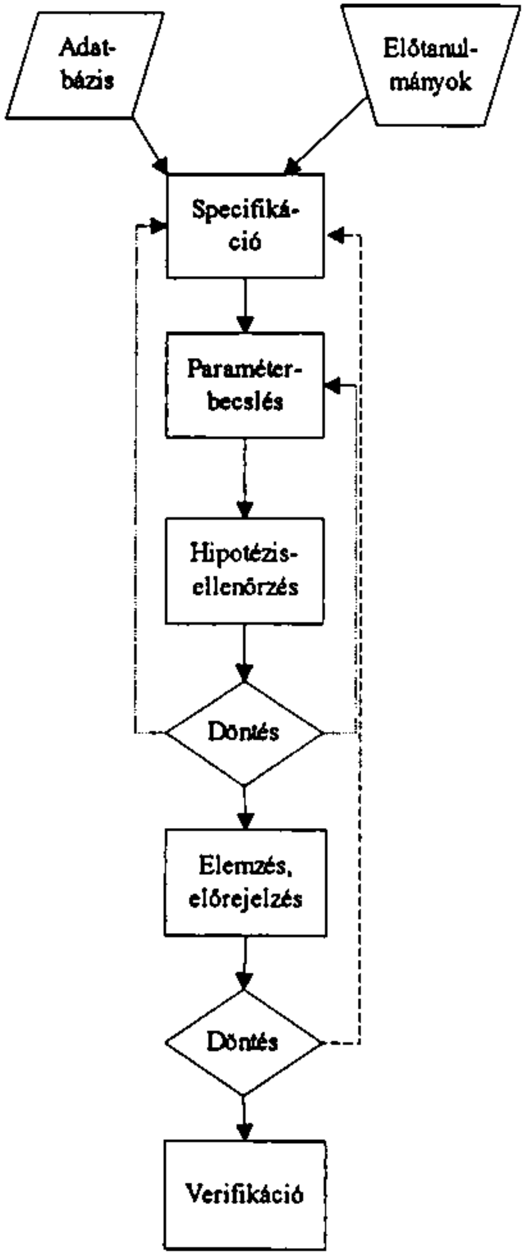
7.2 TÖBBVÁLTOZÓS LINEÁRIS REGRESSZIÓ

Az előző alfejezetben csak két mennyiségi ismérv kapcsolatát vizsgáltuk behatóan. A társadalmi-gazdasági jelenségek összefüggései általában sokkal bonyolultabbak annál, amit két jelenség, folyamat összefüggése jelezni képes. A korábban bemutatott számpéldát átgondolva megállapítható, hogy az üzletek forgalmára jelentősen hat az alapterület, azonban számos más tényező (felszereltség, reklám, vonzáskörzet, dolgozói létszám stb.) is szerepet játszik. Természetesen a modellezés során csak a lényeges befolyásoló tényezőket kívánjuk és tudjuk megragadni, azokat, amelyek valóban elsőrendű szerepet töltenek be e jelenségek, folyamatok alakulásában.

Mivel a jelenségek nem elszigetelten, környezetükből kiragadva játszódnak le, gyakorta számolnunk kell az összetettebb, többváltozós vizsgálatok szükségességével. A **többváltozós** vagy **többszörös regresszióanalízis** módot ad arra, hogy **egyszerre több tényezőváltozónak az eredményváltozóra gyakorolt hatását elemezzük.** A továbbiakban tehát olyan esetekkel foglalkozunk, amikor az egyik változónak az eredményváltozónak kitüntetett szerepe van, alakulását két vagy több tényezőváltozó segítségével magyarázzuk meg.

A többváltozós regresszióanalízis során igen nagy jelentőséget kell tulajdonítani a lényeges ismérvek kiválasztásának, a kapcsolat természetéről alkotott kép jó körülhatárolásának.

Ezen a helyen szeretnénk szólni a modellépítés általános menetéről, ami az alábbi lépéseket foglalja magában:



7-11. ábra A modellépítés folyamatábrája

A regressziós modell építésének első lépése a **specifikáció**, amin, - mint már ezt korábban jeleztük - a jelenséget leíró modellben szereplő változók, tényezőváltozók kiválasztását valamint a függvények konkrét formájának meghatározását értjük, de ugyanakkor specifikációnak nevezzük a véletlen változó eloszlására vonatkozó feltevést is. Ebben a munka-szakaszban a modell szerkesztője mindenekelőtt a priori információkra támaszkodik, **előtanulmányokat** végez, felhasználja azokat az elméleti, szakmai ismereteket, amelyek összefüggnek a vizsgálandó jelenséggel. A modell szerkesztőjének ebben a szakaszban meg kell ragadni a lényeges elemeket. Fontos szerepet játszik a specifikáció szakaszában az **adatbázis**, amelynek minősége, szerkezete nagymértékben befolyásolja a specifikáció érdemi munkáját. Mindig szem előtt kell tartani, hogy a specifikáció során tett megállapítások csupán hipotézisnek tekinthetők, amelyek további ellenőrzésre, felülvizsgálatra szorulnak.

A regressziós modell építésének fontos munkafázisa a **paraméterbecslés**, a regresszióföggvény paramétereinek meghatározása. Ez már tisztán statisztikai munkafázis, a paraméterek becslése szempontjából nem lényegtelen, hogy a regressziós kapcsolatot két vagy több változóval, egy egyenlettel vagy az egyenletek összefüggő rendszerével kívánjuk leírni. Gyakran felmerülő kérdés az, hogy a modell megfelel-e azoknak a különböző matematikai feltételeknek, amelyek a becslési módszereket (klasszikus legkisebb négyzetek módszere, vagy egyéb eljárások) determinálják. Végezetül le kell szögezni, hogy a legnagyobb körültekintéssel kiválasztott becslési eljárással is csak olyan eredményeket kapunk, amelyek hipotetikus jellegűek.

A modell statisztikai ellenőrzésének fázisában a modellező alapvetően a **hipotézisellenőrzés** módszereire támaszkodik. Ez a munkafázis visszahat mind a specifikációra, mind a paraméterbecslésre. Ebben a munkaszakaszban a modellező megállapítja, hogy egy adott megbízhatóság mellett mennyire fogadható el a modell. A próbákkal nyert információk alapján döntést hoz a modell esetleges megváltoztatásáról, vagy a becslési módszer módosításáról. Ezek a döntések természetesen visszahatnak a specifikációra és indokolt esetben az egész eljárás (specifikáció, becslés, hipotézisellenőrzés) megismétlését igényelhetik.

A regressziós modelleket alapvetően két elkülöníthető feladat megoldására használhatjuk:

a) jelenségek elemzésére, analízisére

b) előrejelzések (ex post, ex ante) **készítésére,**

azaz adott feltételek mellett a függő változó értékének, vagy a paramétereknek a becslésére. Az első esetben az ún. bázisvizsgálatok kapcsán a változók összefüggéseinek elemzése a cél, míg a második esetben - különösen a prognóziskészítéseknél - már az is fontos információ, hogy melyek azok a szignifikáns elemek, amelyek a jövőben is jelentős hatást gyakorolhatnak a folyamatra.

A **verifikálás** során a modellt szembesítjük valósággal. Mindez nem csupán egy technikai eljárás. A kétféle feladat jellegének megfelelve közgazdasági, szakmai, logikai kontroll mellett keressük ebben a szakaszban arra a választ, hogy helyesen választottuk-e ki a problémával kapcsolatos változókat, jól írtuk-e le a vizsgált jelenséget, folyamatot. Természetesen a verifikáció során sokszor kapunk olyan információkat amelyek visszahatnak a korábbi munkafázisokra, így a specifikációra is.

A továbbiakban a többváltozós modell legfontosabb jegyeivel ismerkedünk meg.

7.2.1 A többváltozós lineáris modell és a paraméterek becslése

A többváltozós lineáris regresszió a regresszióanalízisnek talán leggyakrabban használt módszere. Elterjedtségét alapvetően a jelenségek, folyamatok összetett jellegével magyarázhatjuk. Tulajdonképpen tiszta kétváltozós regressziót csak nagyon ritkán specifikálhatunk, az egyéb releváns változók hatásainak figyelembevételéről legtöbbször csak technikai okok miatt mondunk le. Általában élünk a kapcsolat linearitásának feltételezésével, amit alapvetően az alkalmazott módszertan kidolgozottsága támaszt alá. El kell mondanunk, hogy a közgazdasági elemző munkát nem érinti károsan ez az egyszerűsítés, a jelenségek, folyamatok természetének megismerése így sem szenved csorbát.

A többváltozós lineáris regresszió modellje az alábbi:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \qquad (7\text{-}49)$$

A fenti formula nem különbözik lényegesen a kétváltozós regressziótól, az ε változó a véletlen hatást reprezentálja, amelyről a továbbiakban is feltesszük, hogy várható értéke nulla és független a tényezőváltozók nagyságától; a k számú tényezőváltozót rögzített, nem véletlen változónak tekintjük, amelyek lineárisan kapcsolódnak az eredményváltozóhoz.

A többváltozós regresszióanalízis problémája **mátrixalgebra** segítségével áttekinthető és könnyen kezelhető formába önthető. A továbbiakban ezért, ahol célszerűnek tartjuk gyakran élünk majd ezzel a felírási móddal. Az eredményváltozó megfigyelt értékeit az y vektor tartalmazza:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

A tényezőváltozók megfigyelt értékeit egy n×m típusú[9] X mátrixban[10] foglalhatjuk össze, ahol a mátrix első oszlopa egy n elemű összegzővektor, a további oszlopvektorok pedig az X_1, X_2 stb. változók megfigyelt értékeiből épülnek fel. Itt szólunk arról is, hogy az összegzővektor segítségével a regressziófüggvény β_0 paraméterét becsüljük.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ . & & & & \\ . & & & & \\ . & & & & \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}$$

A regressziós paraméterek vektora az alábbi:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}$$

A fentiekhez hasonlóan írhatjuk fel a véletlen változót tartalmazó vektort is:

[9]Ahol: m=k+1

[10]Megjegyezzük, hogy a jelölés kissé eltér a mátrixalgebrában megszokott módtól, mivel itt (és az ökonometriában) az X érték első futóindexe a tényezőváltozó sorrendjére, míg a második a megfigyelés sorszámára utal.

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Az ily módon bevezetett jelölések lehetővé teszik azt, hogy a (7-49)-ben jelzett regressziós modellt egyszerűbb formában felírjuk:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (7\text{-}50)$$

Természetesen az alap-regressziós modellnek csupán becslését tudjuk előállítani a gyakorlatban. A minta-regressziós modell annyiban tér el az előbbitől, hogy a véletlen változót a "megfigyelhető" értékekkel valamint a paramétervektort a minta segítségével történő becsült paramétervektorral helyettesítjük.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \qquad (7\text{-}51)$$

ahol:

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ . \\ . \\ . \\ b_k \end{bmatrix} \text{ és } \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

A reziduális változó empirikus értékeit az alábbi egyenlet segítségével határozhatjuk meg:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

A tapasztalati reziduumok négyzetösszege a fenti kifejezésből előállítható az alábbi módon:

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T\mathbf{y} - 2\mathbf{b}\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b} \tag{7-52}$$

A paraméterbecslést a klasszikus legkisebb négyzetek módszerével úgy tudjuk megvalósítani, ha az $\mathbf{e}^T\mathbf{e}$ négyzetösszeg **minimumát** keressük.

$$\frac{\partial(\mathbf{e}^T\mathbf{e})}{\partial\mathbf{b}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b} \tag{7-53}$$

Amennyiben a (7-53) kifejezést egyenlővé tesszük nullával és elvégezzük a egyszerűsítéseket a **normálegyenletek** mátrixalgebrai felírásához jutunk:

$$\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})\mathbf{b} \tag{7-54}$$

A (7-54) kifejezés tulajdonképpen az alábbi egyenletrendszert tartalmazza:

$$\sum y = b_0 n + b_1 \sum x_1 + b_2 \sum x_2 + ..$$
$$\sum x_1 y = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + ...$$
$$\sum x_2 y = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 + ...$$
$$\vdots$$
$$\sum x_k y = ...$$

A megfelelő adatokat behelyettesítve, az egyenletrendszert megoldva meghatározhatjuk a paraméterek számszerű értékeit. Több változó esetén azonban áttekinthetőbb és a számítógépek számára "érthetőbb" a mátrix alakkal felírható megoldás. Tulajdonképpen a (7-54) egyenletből közvetlenül kapjuk a paramétervektort, ha az egyenlet mindkét oldalát $\mathbf{X}^T\mathbf{X}$ mátrix inverzével szorozzuk. Itt jegyezzük meg, hogy amennyiben a fenti inverz nem létezik, azaz a mátrix szinguláris, a paraméterbecslés sem végezhető el.

$$b = (X^T X)^{-1} X^T y \qquad (7\text{-}55)$$

A kétváltozós lineáris regresszió bemutatásakor megismert példánkat folytatva jól tudjuk szemléltetni a többváltozós lineáris regressziós függvény becslésének menetét.

Tételezzük fel, hogy a specifikáció során fontosságot tulajdonítunk - az alapterület változója mellett - a boltokban foglalkoztatottak létszámának is. A 10 üzlet alapterülete mellett ezért feltüntettük a létszámadatokat is. Mindkét változó, természetesen rögzített, nem véletlen változónak tekinthető. Az eredményváltozót azonban már, az alapfelevésünknek megfelelően, valószínűségi változónak tekinthetjük. A kibővített számpélda alapadatait tartalmazza az alábbi tábla:

Alapterület létszám és a forgalom

7-8. tábla

Alapterület, m^2 X_1	Létszám, fő X_2	Évi forgalom, MFt Y
100	8	12
150	7	10
200	10	24
250	8	30
300	12	26
350	10	50
400	15	41
450	22	60
500	25	80
550	32	100

A klasszikus legkisebb négyzetek becslési módszerét alkalmazva a többváltozós regressziófüggvény konkrét formája az alábbi:

$$\hat{y} = -13{,}72 + 0{,}0989x_1 + 1{,}67x_2$$

A többváltozós regressziós modell paraméterei igen fontos jelentéstartalommal bírnak, amelyek a gazdasági elemzésekben kiemelkedő szerepet töltenek be. A regressziós együtthatók értelmezéséhez az alábbi regressziós függvény szolgál kiindulásul.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Ha a fenti modellben pl. az x_1 értékét egy egységgel növeljük, miközben a többi változó értéke nem változik, az eredményváltozó értéke átlagosan b_1 értékével növekszik, ugyanis

$$\begin{aligned}\hat{y} &= b_0 + b_1(x_1 + 1) + b_2x_2 + \ldots + b_kx_k = \\ &= b_0 + b_1x_1 + b_1 + b_2x_2 + \ldots + b_kx_k\end{aligned}$$

Ez a megállapítás értelemszerűen a többi változóra is érvényes. Általánosságban is megfogalmazhatjuk, hogy a többváltozós lineáris regressziófüggvény együtthatói azt mutatják, hogy egy adott tényezőváltozó egységnyi növekménye átlagosan és várhatóan mekkora változást (növekedést vagy csökkenést) idéz elő az eredményváltozóban, miközben a modellben szereplő többi tényezőváltozó értéke nem változik. A többváltozós regressziós modellek együtthatóit - mivel egy-egy tényezőváltozónak a többitől elszigetelt változását mutatják - **parciális regressziós együtthatóknak** nevezhetjük.

A többváltozós lineáris regressziós modell b_0 együtthatójának értelmezése hasonló a kétváltozós esethez. A regressziós függvény ennek a paraméternek az értékét csak akkor veheti fel, ha a mindegyik tényezőváltozó értelmezési tartományában szerepel a nulla is, ellenkező esetben a paraméter közgazdasági értékelésétől eltekintünk.

Példánkban a b_0 paramétert nem értékeljük. A b_1 regressziós paraméter azt mutatja, hogy egy adott üzlet alapterületének egy egységnyi (1 m^2) növekedése mintegy 99 EFt-os évi átlagos forgalmi növekedést idéz elő, feltéve a másik tényező - a létszám - változatlan, azonos szintjét. A b_2 paraméter szerint egy üzlet létszámán egy fővel történő növelése átlagosan mintegy 1,67 MFt-tal növelheti üzlet forgalmát, feltéve természetesen az alapterület azonos mértékét.

Természetesen a többváltozós modell esetén is értelmezhetjük az elaszticitást, amit **parciális elaszticitás** mutatójának nevezünk. Ez a mutató arra ad választ, hogy egy adott tényezőváltozó egységnyi relatív növekedése várhatóan mekkora relatív változást idéz elő az eredményváltozóban, a többi tényezőváltozó állandó színvonala mellett. Elaszticitási mérőszámot a parciális differencálhányadosból származtathatunk:

$$El(y,x_j)=\frac{\partial y}{\partial x_j}\frac{x_j}{y} \qquad (7\text{-}56)$$

$$\text{ahol: } j=1,2\ldots k$$

Az y változó helyett, annak a többváltozós függvénnyel becsült értékeit alkalmazva, a j-edik tényezőváltozó parciális elaszticitási együtthatója:

$$El(\hat{y},x_j)=\frac{b_jx_j}{b_0+b_1x_1+b_2x_2+\ldots+b_kx_k} \qquad (7\text{-}57)$$

A parciális elaszticitás értéke függ a tényezőváltozók adott szintjétől. Gyakran használják az elaszticitási mérőszámot, ha valamennyi tényezőváltozó átlagos szintje mellett kívánják számszerűsíteni a relatív változást.

Számpéldánkban az alapterület változó (x_1) elaszticitásának átlagos szintje (a másik tényezőváltozó átlagos értéke mellett):

$$El(\hat{y},\bar{x}_1)=\frac{0,0989\times 325}{-13,72+0,0989\times 325+1,67\times 14,9}=1,29$$

Egy átlagos alapterülettel rendelkező bolt területének 1 %-os átlagos növekedése - a létszám változatlan, átlagos szintje mellett - 1,29 %-os forgalomnövekedést indukál várhatóan. A forgalom a terület növekedésére érzékenyen reagál.

7.2.2 Hipotézisellenőrzés, intervallumbecslés

A kétváltozós lineáris regressziós függvényhez hasonlóan végezhetjük el többváltozós esetben is, a modell egészének, illetve a modell paramétereinek külön-külön történő tesztelését, valamint az intervallumbecsléseket.

A modell egészének megbízhatóságát a varianciaanalízis segítségével ellenőrizhetjük. A már megismert algoritmusok szerint járunk el a többváltozós esetben is. A varianciaanalízis táblázata az alábbi:

Varianciaanalízis

7-9. tábla

Összetevő	Négyzet-összeg	Szabadság-fok	Szórásnégyzet becslése
Regresszió SSR	$bX^Ty - n\bar{y}$	k	$\frac{(bX^Ty - n\bar{y})}{k}$
Maradék SSE	$y^Ty - bX^Ty$	n-k-1	$\frac{(y^Ty - bX^Ty)}{n-k-1}$
Teljes SST	$y^Ty - n\bar{y}$	n-1	$\frac{(y^Ty - n\bar{y})}{n-1}$

Az F-próbafüggvény:

$$F = \frac{(bX^Ty - n\bar{y})}{k} \div \frac{(y^Ty - bX^Ty)}{n-k-1}$$

Mivel

$$R^2 = \frac{bX^Ty - n\bar{y}}{y^Ty - n\bar{y}} \qquad (7\text{-}58)$$

ezért a többszörös determinációs együttható segítségével is ellenőrizhetjük a modell egészét, az alábbi módon:

$$F = \frac{R^2 / k}{\left(1 - R^2\right) / \left(n - k - 1\right)} \qquad (7\text{-}59)$$

Természetesen - mint azt már korábban is láttuk - a (7-58) próbafüggvény az alábbi nullhipotézis ellenőrzésére szolgál:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \qquad (7\text{-}60)$$

Az ellenhipotézis elfogadása esetén azt állíthatjuk, hogy van legalább egy olyan tényezőváltozó, amely szignifikáns hatással rendelkezik.

Példánkban a számláló szabadságfoka 2; a nevező szabadságfoka 7, és ennek megfelelően a táblabeli érték $F_{0,05}$=4,74. Mivel a számított érték (F=51,4) jelentősen meghaladja a táblabeli értéket, a modell szignifikáns, valódi összefüggést fejez ki.

A regressziós együtthatók parciális tesztelése, valamint az intervallumbecslések a véletlen változó szórásának (σ) becslésére épülnek. Ezt a szórást a gyakorlati adatok alapján az alábbi reziduális szórással becsüljük:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{(n-k-1)} = \frac{e^T e}{(n-k-1)} \qquad (7\text{-}61)$$

A paraméterek standard hibáit az ún. regressziós együtthatók variancia-kovariancia mátrixából vezethetjük le, amelynek általános formáját az alábbiakban adjuk meg:

$$Var(b) = \begin{bmatrix} Var(b_1) & Cov(b_1 b_2) & \cdot & \cdot \\ Cov(b_2 b_1) & Var(b_2) & \cdot & \cdot \\ \cdot & & & \\ \cdot & & & \end{bmatrix} = \sigma^2 \left(X^T X \qquad (7\text{-}62)$$

A fenti összefüggésből látható, hogy az $X^T X$ mátrix inverze igen fontos szerepet tölt be, segítségével számíthatjuk ki az egyes paraméterek standard hibáit. A további összefüggések áttekinthetősége érdekében vezessük be az alábbi jelölést:

$$\left(X^T X\right)^{-1} = [z_{ij}]$$

ahol a z_{ij} a mátrix általános elemét jelöli.

Mindezek előrebocsátása mellett a j-edik paraméter standard hibáját a következő módon határozhatjuk meg:

$$s_{b_j} = s\sqrt{z_{jj}} \qquad (7\text{-}63)$$

Általában a paraméterek standard hibáit a programcsomagok szolgáltatják, így nem okoz nehézséget a regressziós együtthatók parciális tesztelése az alábbi t-statisztika segítségével:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \qquad (7\text{-}64)$$

$$t = \frac{b_j}{s_{b_j}}$$

Példánkat folytatva a klasszikus legkisebb négyzetek módszerével a többváltozós regressziófüggvényre az alábbi eredményeket kapjuk:

$$\begin{aligned} \hat{y} = -13{,}72 &+ 0{,}0989x_1 + 1{,}67x_2 \\ (6{,}7) \quad &(0{,}0431) \quad (0{,}763) \end{aligned}$$

A paraméterek alatt zárójelben a standard hibák szerepelnek[11]. A fenti számadatok ismeretében felírhatjuk a t-statisztikákat is, amelyek sorrendben: 2,05; 2,29; és 2,19.

Példánkban az n-k-1 szabadságfokú t-eloszlást figyelembe véve, 5 %-os szignifikancia szinten ($t_{0,05(7)} = 2,365$) a tényezőváltozók parciális hatását kifejező regressziós együtthatók nem szignifikánsak. A szignifikancia szintet 1 %-ra csökkentve a tényezőváltozók parciális hatása szignifikánsnak tekinthető.

A standard hibákat közvetlenül felhasználhatjuk - amint azt a kétváltozós esetben láttuk - a regressziós együtthatók konfidencia intervallumának előállítására, azaz intervallumbecslésére. Itt is hangsúlyozzuk, hogy a regressziós együtthatók torzítatlan becslését kapjuk a klasszikus legkisebb négyzetek módszerével.

A feltételes várható érték, illetve egy egyedi érték intervallumbecslését is elvégezhetjük a többváltozós lineáris regressziófüggvény segítségével. A számításokat itt is meg kell előznie a pontbecslésnek. A becsléshez a rögzített kiinduló értékeket az alábbi vektor tartalmazza:

[11] Itt hívjuk fel arra a figyelmet, hogy a számítógép által szolgáltatott output táblák általában a fenti információkat tartalmazzák. Természetesen az output táblák felépítése programcsomagonként változó, azonban főbb elemeikben megegyeznek.

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{10} \\ x_{20} \\ . \\ . \\ . \\ x_{k0} \end{bmatrix} \tag{7-65}$$

Az $\hat{Y}_0$ **feltételes várható értéknek** $(\hat{y})$ és az **egyedi** $\hat{Y}_0'$ **értéknek** $(\hat{y}')$ a **pontbecslése** $\mathbf{x}_0$ mellett (ahol $\mathbf{b}$ a becsült paramétervektor):

$$\hat{y} = \hat{y}' = \mathbf{x}_0^T \mathbf{b} \tag{7-66}$$

A korábban már megismert, (7-61) szerint becsült reziduális szórás, valamint a megfelelő vektorok és mátrixok segítségével meghatározhatók a becslés standard hibái.

A feltételes várható érték standard hibája:

$$s_{\hat{y}} = s\sqrt{\mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_0} \tag{7-67}$$

Az adott vektorhoz és így a kijelölt értékekhez tartozó **egyedi érték standard hibája**:

$$s_{\hat{y}}' = s\sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_0} \tag{7-68}$$

Látható, hogy az egyedi érték standard hibája a többváltozós esetben is meghaladja a feltételes várható érték hibáját.

Végezetül szólni kell arról, hogy a standard hibákat az t-eloszlás n-k-1 szabadságfokú értékével megszorozva a becslések hibahatárait és így végül a konfidencia intervallumokat állíthatjuk elő, adott megbízhatósági szint mellett.

7.3 NEMLINEÁRIS REGRESSZIÓ

A függvénytípus megfelelő megválasztása - mint láttuk - a specifikáció fontos eleme. Sikeres megvalósítása nagymértékben függ a kiválasztást megelőző közgazdasági-logikai elemzéstől.

A függvénytípust kezdetben csak hipotézisként kezeljük mindaddig, amíg a statisztikai-, és szakmai ellenőrzések választás helyességét nem igazolják.

Tanulmányunkban eddig előnyben részesítettük a lineáris függvényt, megállapításaink csak erre az esetre vonatkoznak. A gyakorlati elemző munka sok esetben elfogadja a lineáris modell feltételezését, és amikor nagyon indokolt, a bonyolultabb összefüggéseket egyszerű transzformációk segítségével lineárissá alakítják át.

A gazdasági-társadalmi folyamatok és jelenségek azonban gyakran igénylik a nemlineáris típusú függvények közvetlen használatát is. A regressziós modell-építés során a nemlineáris függvényeknek alapvetően két típusát különböztethetjük meg. Ezek a **változóiban nem lineáris** függvények és **paramétereiben nem lineáris** függvények. Az első esetben viszonylag könnyebb dolga van a felhasználónak, mivel megfelelő transzformáció után a függvény lineárissá alakítható, és alkalmazható a paraméterbecslésre a klasszikus legkisebb négyzetek módszere. Természetesen tudni kell azt, hogy az ilyen átalakítások a véletlen változóra vonatkozó pontosan definiált feltételezések esetén adnak valós eredményeket, azoknak a transzformáció után is értelmezhetőnek kell lenniük. A második esetben a paraméterek becslése a hagyományostól eltérő becslési módszerek megválasztását igényli. A véletlen változóra tett feltétel természetesen itt is alapvető fontossággal bír, azonban általában a priori feltehetjük, hogy a véletlen változó additív módon kapcsolódik a modellhez, eloszlása normális. Ilyen esetekben a transzformáció nem járható út és ekkor nyúlunk a bonyolultabb nemlineáris becslési eljárásokhoz, illetve a maximum likelihood módszerhez.

Az alkalmazható függvény típusát a gazdasági jelenségek természete gyakorta szigorúan behatárolja. A gazdasági elemzések során például a vállalatok, vállalkozások szintjén néha egy-egy előrebecslés során kifejezetten hamis információt szolgáltatna a lineáris modell. (Például egy optimális külső gazdasági feltételek mellett, dinamikusan fejlődő vállalkozás termelési inputjai és a termelés outputja között az összefüggés nem lineáris.) A számítógépek elterjedésével sokféle lehetőség nyílik a megfelelő függvénytípus megtalálására, de mindez csak nagyfokú szakmai kontroll mellett ad jó eredményt.

A nemlineáris regressziós modellek részletes tárgyalása meghaladja tananyagunk kereteit, ezért a továbbiakban csak azokkal a leggyakrabban alkalmazható függvénytípusokkal foglalkozunk, amelyek a klasszikus legkisebb négyzetek módszerével becsülhetők.

7.3.1 Hatványkitevős modell

A gyakorlatban sokszor használt nemlineáris modell az alábbi[12]:

$$Y = \beta_0 X^{\beta_1} e^{u} \qquad (7\text{-}69)$$

ahol: u -a reziduális, véletlen változó

[12]Meg kell jegyeznünk, hogy itt e = 2,718..., azaz az ún. természetes szám szerepel a képletben és ebben a fejezetben a véletlen változó jelölésére az u-t használjuk.

A modell sajátos vonása, hogy a β_1 az eredményváltozó X kitevőjében szerepel az egyes tagok kapcsolódása pedig szorzatszerű. Az ilyen modellt **hatványkitevős**, illetve **multiplikatív** modellnek nevezzük.

A β_1 regressziós együttható azt fejezi ki, hogy a tényezőváltozó egységnyi relatív változása mekkora relatív változást idéz elő az eredményváltozóban. Jól tudjuk, hogy az ilyen jellegű kérdésekre általános válasz adható az elaszticitási együtthatóval segítségével. A fenti modellben az elaszticitás állandó. Ennek megfelelve a modellt szokás **konstans elaszticitású modellnek** is nevezni.

A modell logaritmikus transzformációval lineáris alakra hozható:

$$\ln Y = \ln\beta_0 + \beta_1 \ln X + u \qquad (7\text{-}70)$$

A logaritmikus transzformációval lineáris alakra hozott modell nagy előnye, hogy a paraméterei már a klasszikus legkisebb négyzetek módszerével becsülhetők. Itt ugyanis az eredeti változók logaritmus értékei között lineáris összefüggés van. A logaritmizálás művelete nyomán a fenti modellt nevezik még az irodalomban **log-lineáris** vagy **dupla logaritmikus modellnek** is.

A logaritmikus transzformációból adódóan, ha a grafikus ábrázolás során mindkét tengelyen logaritmikus skálát alkalmazunk, a függvény képe egyenes vonal lesz. Mindez nagymértékben megkönnyíti a multiplikatív összefüggések felismerését.

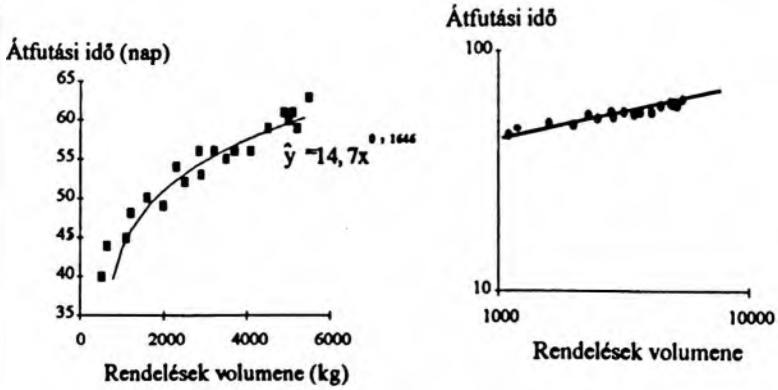
A hatványkitevős modellt egy egyszerű számpéldával illusztráljuk. Tételezzük fel, hogy egy textilipari vállalatnál a rendelésállományok elemzése során a rendelések volumenének és az átfutási időnek az összefüggéseit vizsgálták. 20 rendelés adatait az alábbi tábla tartalmazza:

A rendelések volumene és az átfutási idő

7-10. tábla

Sorszám	Rendelésenkénti súly (kg)	Rendelések átfutási ideje (nap)
1.	500	40
2.	650	44
3.	1.100	45
4.	1.200	48
5.	1.600	50
6.	2.000	49
7.	2.300	54
8.	2.500	52
9.	2.850	56
10.	2.900	53
11.	3.200	56
12.	3.500	55
13.	3.700	58
14.	4.100	56
15.	4.500	59
16.	4.900	61
17.	5.000	60
18.	5.100	61
19.	5.200	59
20.	5.500	63

Az adatokat ábrázolva (7-12 ábra) megállapíthatjuk, hogy létezik korrelációs kapcsolat a rendelési volumen és a termelés átfutási ideje között, a nagyobb volumenű rendelés hosszabb átfutási időt igényel, azonban ez a kapcsolat nem lineáris. A rendelés mennyiségének jelentősebb növekedése relatíve kisebb megmunkálási többletidőt igényel. A logaritmikus skálán ábrázolva az adatokat a pontok határozott egyenes körül helyezkednek el, ami alátámasztja a hatványkitevős, multiplikatív modell alkalmazását.



7-12. ábra A rendelések volumene és az átfutási idő természetes és logaritmikus skálán

A változókat logaritmizálva a klasszikus legkisebb négyzetek módszerével kiszámíthatók a regressziós paraméterek.

$$\ln \hat{y}_i = 2{,}6878 + 0{,}1646 \ln x_i$$

$$\hat{y}_i = 14{,}7 \times x_i^{0{,}1646}$$

Amennyiben X=1 beletartozik a tényezőváltozók értelmezési tartományába, ezen a helyen a b_0 együttható az adott értéket veszi fel. Általában, így példánkban sem értelmezzük szakmailag a b_0 együtthatót.

A b_1 regressziós együttható arra ad választ, hogy a tényezőváltozó egységnyi relatív növekedése várhatóan milyen relatív változást indukál az eredményváltozó értékében. Példánkban azt mondhatjuk, hogy - a 20 elemű minta alapján - a rendelés volumenének 1 %-os növekedése az átfutási idő mintegy 0,16 %-os emelkedését vonja maga után átlagosan.

A b_1 regressziós együttható a hatványkitevős regressziós modell esetén azonos az **elaszticitási együtthatóval**, ami állandó.

7.3.2 Féllogaritmikus modell

Előfordulhat a gyakorlatban olyan összefüggés, amelyben az egyik változó természetes értéke és a másik változó logaritmusa között írható fel a lineáris modell. Két ilyen esetet különböztethetünk meg:

$$\ln Y = \beta_0 + \beta_1 X + u \qquad (7\text{-}71)$$

illetve

$$Y = \beta_0 + \beta_1 (\ln X) + u$$

Az egyenes meredekségét kifejező regressziós paraméter az első esetben az Y változó **relatív változását** X változó **abszolút változásához** viszonyítja, míg a második esetben Y változó **abszolút változását méri** X változó **relatív változásához** képest.

Könnyen beláthatjuk, hogy az első eset az ún. **exponenciális modell** linearizált megfelelője, ami eredeti formájában:

$$Y = \beta_0 \beta_1^X e^u \qquad (7\text{-}72)$$

A fenti modell fő felhasználási területe a trendszámítás[13], mivel nem idősorokból származó ún. keresztmetszeti adatokból viszonylag ritkán számítunk exponenciális regressziót. Ugyanakkor el kell mondani, hogy a féllogaritmikus, vagy szemilogaritmikus regressziónak számos egyéb felhasználási lehetőségére nyílik mód különböző gazdasági számításokban.

[13]Lásd a 8. fejezetet.

7.3.3 Reciprok regressziós modell

A reciprok regressziós modellt (általában hiperbolát) akkor alkalmazzuk, ha a közgazdasági feltevések szerint az összefüggések természete reciprok jellegű kapcsolatra utal. Az ilyen típusú modell felhasználásának szinte iskolapéldája a termelési output mint X változó és az állandó költségeknek az összköltségen belüli aránya, mint Y változó közötti kapcsolatot számszerűsítő modell. A specifikáció során itt is alapvető fontossággal bír a szakmai ismeretek köre, de nem hanyagolható el a függvény típusának felismerése során a grafikus ábrázolás sem. A reciprok regresszió lineáris modellre visszavezethető. A gyakorlatban leggyakrabban használt két **hiperbola** függvény az alábbi:

$$Y = \frac{1}{\beta_0 + \beta_1 X} + u \qquad \text{illetve} \qquad Y = \beta_0 + \beta_1\left(\frac{1}{X}\right) + u \tag{7-73}$$

Az első esetben az $1/Y$ és X változó míg a második esetben az Y és az $1/X$ változók között áll fenn lineáris kapcsolat.

Külön szólni kell a második modellben a β_0 paraméter értelmezéséről. Ez az érték, amely felé a függvény aszimptotikusan közelít, ha X értéke minden határon túl nő.

$$\lim_{X \to \infty}\left(\beta_0 + \beta_1 \frac{1}{X}\right) = \beta_0$$

A b_0 paraméter tehát a regresszió minimális értékének becsléseként fogható fel. Nem ilyen kézenfekvő a b_1 paraméter értelmezése, ezért a gyakorlat sokszor eltekint magyarázatától. Megjegyezzük, hogy a hiperbolikus regresszió, nemlineáris jellegéből adódóan igen érzékeny a tényezőváltozó mértékegységére, annak megváltozása az eredményeket erőteljesen módosítja.

A reciprok regresszió szemléltetésére egy építőipari példát mutatunk be. Tételezzük fel, hogy valamely építőipari vállalatnál egy meghatározott időszakban a kivitelezett munkák számla szerinti értékét és a szállítási költségek önköltségen belüli arányát jellemzik az alábbi adatok:

A kivitelezett munkák értéke és a szállítási költség aránya

7-11. tábla

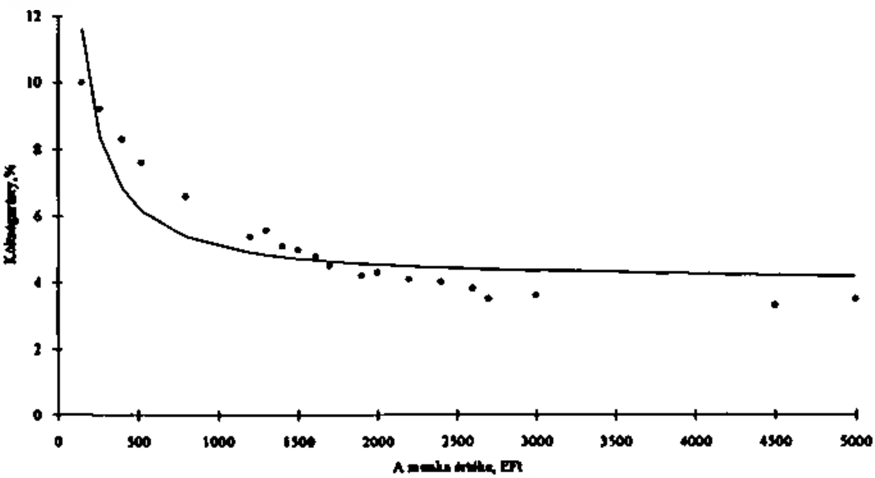
Sorszám	A munka értéke (EFt)	Szállítási költség aránya (%)
1.	150	10,0
2.	260	9,2
3.	400	8,3
4.	520	7,6
5.	800	6,6
6.	1.200	5,4
7.	1.300	5,6
8.	1.400	5,1
9.	1.500	5,0
10.	1.610	4,8
11.	1.700	4,5
12.	1.900	4,2
13.	2.000	4,3
14.	2.200	4,1
15.	2.400	4,0
16.	2.600	3,8
17.	2.700	3,5
18.	3.000	3,6
19.	4.500	3,3
20.	5.000	3,5

A fenti számadatokból a regressziós modell paramétereit a klasszikus legkisebb négyzetek módszerével nyerhetjük. Itt a (7-73)-ban másodikként jelzett modellt használtuk fel, amelynek számszerű értéke:

$$\hat{y}_i = 3{,}96 + 1.148{,}2\frac{1}{x_i}$$

Példánkban a regressziós függvény segítségével megállapíthatjuk, hogy az építési munkák értékének nagyarányú növekedése esetén az elérhető minimális szállítási költségarány mintegy 3,96 %.

A vizsgált jelenség természetét jól reprezentálja az alábbi ábra.



7-13. ábra Összefüggés az építési munka nagysága és a szállítási költség aránya között

Az ábrából láthatjuk, hogy a szállítási költségek aránya jelentősen különbözik pl. az 500-1.000 EFt illetve a 4.000-5.000 EFt közötti intervallumban. A fenti regressziós függvény segítségével határozzuk meg néhány tényezőváltozó értéke mellett az eredményváltozó értékét!

$$\text{ha } x = 500, \hat{y} = 6{,}26\ \%$$
$$\text{ha } x = 1.000, \hat{y} = 5{,}11\ \%$$
$$\text{ha } x = 4.000, \hat{y} = 4{,}25\ \%$$
$$\text{ha } x = 5.000, \hat{y} = 4{,}19\ \%$$

A fentiek alapján megállapíthatjuk, hogy ha a kivitelezett munkák értéke 500 EFt-ról 1.000 EFt-ra nő, a szállítási költségek aránya mintegy 1,15 %-kal csökken, míg 4.000 EFt-ról 5.000 EFt-ra történő növekedés esetén az arány csökkenése csupán 0,06 %-os.

7.3.4 Polinomiális regresszió

A nemlineáris regresszió gyakran használt típusa a polinomiális regresszió. Különösen a másodfokú polinom, a parabola, illetve a harmadfokú polinom közgazdasági alkalmazása gyakori. A parabola alkalmazását indokolja pl. a mezőgazdasági termelés során a felhasznált műtrágya mennyisége, amely egy adott szintig növeli a termésátlagot, majd ezen túl a több hatóanyag már alacsonyabb hozamot eredményez. A harmadfokú polinom felhasználása különösen a hagyományos összköltség-görbe illesztése során indokolt.

A polinomiális modell általános alakja[14]:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + u \qquad (7\text{-}74)$$

Az általános felírásból láthatjuk, hogy a lineáris függvény tulajdonképpen felfogható a polinomiális függvény speciális esetének. A többváltozós lineáris regressziós modell során megismert megoldási eljárások a polinomiális modell

[14]A polinomiális modell alkalmazásának fontos feltétele a polinom fokszámának helyes meghatározása, aminek sokféle módszere közismert. Meg kell azonban mondani, hogy mindegyik eljárás csupán közelíti a probléma megoldását.

esetében maradéktalanul alkalmazhatóak, ugyanis az X változó hatványai helyett új változókat vezethetünk be, ami tulajdonképpen a problémát lineáris esetre vezeti vissza:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots \qquad (7\text{-}75)$$

ahol

$$Z_1 = X;\ Z_2 = X^2 \text{ stb.}$$

7.3.5 A nemlineáris kapcsolat szorossága

A változók közötti korrelációs kapcsolatot kétféle módon, egyrészt a **transzformált változók** másrészt az **eredeti változók** alapján tudjuk mérni. A transzformált változók segítségével történő kapcsolatmérés csupán az eredeti változók között fennálló kapcsolat közelítéseként fogható fel. Az eredeti változók alapján történő kapcsolatmérésben hatékony segítséget ad az ún. **korrelációs index:**

$$I = \sqrt{1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}} \qquad (7\text{-}76)$$

A mutatóról könnyen belátható, hogy értéke 0 és 1 közé esik, valamint szembetűnő az a tulajdonsága, hogy lineáris kapcsolat esetén értéke megegyezik a lineáris korrelációs együtthatóval.

7.3.6 Többváltozós nemlineáris modell

A nemlineáris modellek között kiemelt fontossággal bír a Cobb-Douglas féle termelési függvény, ami már más közgazdasági diszciplínákból közismert. A függvény tulajdonképpen egy többváltozós hatványkitevős függvény, amelyben a véletlen változóra tett feltevés "eléggé speciális":

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} e^u \tag{7-77}$$

ahol:

Y - a termelés outputja,

X_1 - a munkaráfordítás,

X_2 - a tőkeráfordítás.

Amennyiben logaritmizáljuk a fenti függvényt, az könnyen kezelhetővé, becsülhetővé válik.

$$\ln Y = \ln\beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + u \tag{7-78}$$

Így már nincs akadálya annak, hogy a paramétereket a klasszikus legkisebb négyzetek módszerével becsüljük. Természetesen néha feltételeket, korlátokat is meg kell fogalmaznunk a kitevőkben lévő paraméterekre vonatkozóan. Ezek a feltételek a becslési módszerek további finomítását igénylik.

Amennyiben a véletlen változó nem a (7-77) modellben szereplő módon helyezkedik el, a nemlineáris legkisebb négyzetek módszerét , vagy a maximum likelihood becslést kell alkalmazni.

7.4 A MODELL FELTÉTELRENDSZERÉNEK VIZSGÁLATA

A regressziós modell specifikálása során fontos feltevésekkel élünk mind a reziduális változóra, mind a magyarázó változókra vonatkozóan. Az alábbiakban összefoglaljuk a lineáris regressziós modellek legfontosabb követelményeit. Ezek teljesülése esetén a klasszikus legkisebb négyzetek módszere segítségével közvetlenül jól alkalmazható becsléseket nyerhetünk, míg ellenkező esetben vagy más becslési módszert kell választani, vagy kellő kritikával kell fogadni a kapott eredményeket.

7.4.1 A regressziós modell feltételrendszere

A regressziós modell feltételei közül néhányat már a korábban megismertünk, míg mások csak később kerülnek vizsgálódásunk előterébe.

1. A reziduális változó várható értéke nulla.

$$E(\varepsilon_i) = 0 \qquad (7\text{-}79)$$

2. A véletlen tag, hibatényező varianciája állandó, a modell **homoszkedasztikus**.

$$Var(\varepsilon_i) = \sigma^2 \qquad (7\text{-}80)$$

3. A véletlen tag, hibatényező értékei páronként nem korrelálnak egymással, vagyis a modell **nem autokorrelált**, autokorrelálatlan.

$$Cov(\varepsilon_i \varepsilon_j) = 0 \quad \text{ahol: } i \neq j \qquad (7\text{-}81)$$

A második és **harmadik** feltételt **egyesíthetjük** az I **egységmátrix** segítségével:

$$E(\varepsilon\varepsilon^T) = \sigma^2 \mathbf{I} \qquad (7\text{-}82)$$

A fenti három - már korábban megismert - feltételt továbbiakkal egészítjük ki.

4. Ha az X mátrix értékeit, az $X_1, X_2 \ldots X_k$ magyarázó változókat **nem sztochasztikus**, azaz rögzített, fix értékeknek tekintjük, akkor ismételt mintavétel esetén, csupán a véletlen hatására jönnek létre az Y változó különböző értékei.

Az eddigiek során mindig feltételeztük, hogy a magyarázó változó nem sztochasztikus, a méréseket megismételve mindig ugyanezen értékeket nyerjük. A gazdasági, társadalmi modellek megfigyeléseit azonban nem kisérlet-sorozatból kapjuk, hanem azok tőlünk független folyamat eredményei. Mindez azt is jelenti, hogy az egyes magyarázó változók hatását az eredményváltozóra nem lehet külön-külön, tisztán megragadni, hiszen minden változó autonóm módon, egyszerre is változhat. Ezek fényében, kicsit tágabban értelmezve a magyarázó változók szerepét, azok sztochasztikus viselkedését is feltételezhetjük. Mindez értehetőbbé teszi a további feltételek interpretálását.

5. A magyarázó változók **függetlenek** a hibatényezőtől, a reziduális változótól.

$$E\left(X^T \varepsilon\right) = 0 \qquad (7\text{-}83)$$

6. Az $X_1, X_2 \ldots X_k$ magyarázó változók lineárisan függetlenek, nincs "tökéletes" vagy egzakt lineáris összefüggés a tényezőváltozók között. Nem létezik tehát olyan, nullától eltérő c_0, c_1, $c_2 \ldots c_k$ konstans paramétersor, amely teljesíti az alábbi feltételt:

$$c_0 + c_1 x_{1i} + c_2 x_{2i} + \ldots + c_k x_{ki} = 0 \qquad (7\text{-}84)$$
$$\text{ahol: } \forall\, x_{ji} \neq 0$$

Amennyiben a fenti összefüggés mégis fennáll, a modellt **multikollineárisnak** tekinthetjük.

7. A magyarázó változók számának kisebbnek kell lennie, mint a tapasztalati független megfigyelések száma (k<n); más megfogalmazásban az **X** mátrix rangja legyen kisebb a megfigyelések számánál, praktikusan egyezzen meg a változók számával.

$$\rho(X) = k \qquad (7\text{-}85)$$
$$\text{ahol: } k < n$$

A fent említett feltételek teljesülése esetén a regressziós modellt multikollinearitástól mentes **standard lineáris regressziós modellnek** tekinthetjük. Belátható, hogy ezek a néha igen "szigorú" feltételek, a modellépítés alapelvét képzik, amikre a munka során fokozottan figyelnünk kell. Mindez azt jelenti, hogy a modell konkretizálása során a feltételeket folyamatosan kell ellenőrizni, az eredményeket hipotetikus jellegűeknek kell tekinteni. Amennyiben a mintabeli adatokkal nem igazolhatók empirikusan az elméleti várakozások, a specifikáció módosítására, óvatos értelmezésre, illetve sajátos, új becslési módszer megválasztására van szükség. Az első öt feltétel teljesülése esetén, a klasszikus legkisebb négyzetek módszere segítségével kielégítő becsléseket nyerünk, mivel a Gauss-Markov tétel alapján bebizonyítható, hogy a klasszikus legkisebb négyzetek módszerének becslőfüggvénye valamennyi lineáris torzítatlan esetnél jobb, mivel a varianciája a legkisebb. Itt jegyezzük meg, hogy az első feltétel teljesülését tulajdonképpen már a becslési módszer helyes megválasztásával garantálhatjuk.

7.4.2 Multikollinearitás

A 6. és 7. feltételben megfogalmaztuk azt, hogy általában nem lehet lineáris összefüggés a tényezőváltozók között, ellenkező esetben multikollineárisnak kell tekinteni a modellt. Amennyiben a fenti két feltétel teljesül, egy adott tényezőváltozó értékét tetszőlegesen változtathatjuk, anélkül, hogy ez a változás a többi tényező hatását befolyásolná.

A tényezőváltozók közötti függetlenség matematikai következménye, hogy egy adott β_j paraméternek a klasszikus legkisebb négyzetek módszerével becsült b_j értéke független a modell többi tényezőváltozójától. Másként megfogalmazva, a modell tényezőváltozóinak, magyarázó változóinak körét bővíthetjük vagy szűkíthetjük, a már kiszámított regressziós együtthatók értékei minderre érzéketlenek.

A társadalmi-gazdasági jelenségek, folyamatok összefüggéseit kifejező modellekben a tényezőváltozók közötti kölcsönös függetlenség azonban nagyon ritkán nem fordul elő. A vizsgálatok során ugyanakkor természetes és jogos az a cél, hogy a komplex társadalmi jelenségeket és folyamatokat mind mélyrehatóbban ismerjük meg, ami azt is jelenti, hogy célszerű minél több hatótényezőnek, magyarázó változónak a figyelembevétele. Így sok esetben elkerülhetetlenné válik a változók közötti teljes függetlenség elvének a feladása.

A multikollinearitás két alapvető formában jelenhet meg:

1) a magyarázó változók között a **kapcsolat függvényszerű;**

2) a magyarázó változók között a **kapcsolat sztochasztikus**[15].

Az első esetben **teljes,** vagy **extrém** multikollinearitásról, azaz a tényezőváltozók között egzakt lineáris függőségről beszélhetünk. Eddigi tanulmányainkból világosan kitűnik, hogy a b regressziós együtthatókat a klasszikus legkisebb négyzetek módszerével csak akkor határozhatjuk meg egyértelműen, ha az $\mathbf{X}^T\mathbf{X}$ mátrix **nem szinguláris,** azaz egyértelműen invertálható. Ha a mátrix szinguláris, extrém multikollineartással kell számolnunk.

Az előbbinél gyakrabban számíthatunk a multikollinearitás "burkoltabb" formájával, amikor a magyarázóváltozók közötti kapcsolat **sztochasztikus.** Ebben az esetben a jelenség felismerése és lokalizálása is bonyolultabb feladat.

Előre kell bocsátanunk, hogy a **multikollinearitás nem specifikációs hiba,** hanem a felhasznált adatok, információk adottsága, a minta tulajdonsága. Ahhoz azonban, hogy kezelni tudjuk a multikollinearitást, meg kell ismerkednünk várható következményeivel, feltárásának, számszerűsítésének módozataival, és azokkal az

[15] Ebben az esetben természetesen sztochasztikusnak kell tekinteni a magyarázó változókat.

eljárásokkal, amiket az ilyen tulajdonsággal bíró minták adataiból történő becslések során alkalmazni lehet.

A multikollinearitás következményei, megítélése

A multikollinearitás sokszor nehezen észrevehető hatása során gyakorta kell számítani az alábbi következményekkel:

- A multikollinearitás növeli a regressziós együtthatók varianciáját, az egyes t-próba értékek nagyon alacsonyak lesznek.
- Az alacsony t-próba értékek miatt (mivel az egyes paraméterek nem szignifikánsak) gazdaságilag fontos magyarázó változókat indokolatlanul hagyunk ki a modellből.
- A paraméterek értékei mintánként jelentősen különbözhetnek. Érzékenyen reagálhatnak a paraméterbecslések eredményei a minta és a specifikáció változásaira.
- Általánosságban elmondhatjuk, hogy a multikollinearitás a modellel készült előrejelzéseket nem befolyásolja abban az esetben, ha a változók közötti, mintában megállapítható kapcsolat nem változik meg az előrejelzési periódusban. Ellenkező esetben számolni kell azzal, hogy az előrejelzés torzított lesz. A modell eredményeinek elemzésre történő felhasználhatóságát azonban megkérdőjelezi a multikollinearitás, ugyanis a fentiek értelmében nem tudjuk elkülöníteni az egyes magyarázó változók hatását.

A multikollinearitás feltárása

Az alábbiakban néhány egyszerűbb-bonyolultabb eljárást mutatunk be.

a) A **korrelációs mátrix** segítségével a multikollinearitás viszonylag egyszerűen felismerhető. A **tényezőváltozók közötti korrelációs mátrix** általános alakja az alábbi:

$$\mathbf{R}_k = \begin{bmatrix} 1 & r_{12} & . & . & . & r_{1k} \\ r_{21} & 1 & . & . & . & r_{2k} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ r_{k1} & r_{k2} & . & . & . & 1 \end{bmatrix} \qquad (7\text{-}86)$$

Az $\mathbf{R}_k$ mátrix (mérete: k×k) főátlón kívüli elemei a magyarázó változók páronkénti lineáris korrelációs együtthatói. A multikollinearitás meglétére számíthatunk nagy eséllyel, ha a két változó közötti korrelációt mérő együttható értéke abszolút értékben távol esik a nullától. Itt azonban két gondra kell felhívni a figyelmet: egyrészt ha két változópár között alacsony is a korreláció, több mint két változó között erőteljes multikollinearitás állhat fenn; másrészt a "károsnak" ítélt szint megállapítása meglehetősen önkényes. Az ún. **Klein-féle hüvelykujjszabály** szerint erőteljesnek kell tekinti a multikollinearitást, ha van olyan, a kétváltozós kapcsolatot mérő korrelációs együttható, amelynek értéke megközelíti a többszörös korrelációs együtthatót. Természetesen a korrelációs mátrix nem tájékoztat közvetlenül a magyarázó változók közötti többszörös kapcsolatról.

b) A multikollinearitás megléte vizsgálható az ún. **globális mérőszámokkal**.

Egyik ilyen közelítő jellegű mérőszám az M_1 **mutató**. Felépítésének gondolatmenete az alábbi. Ha a tényezőváltozók (magyarázó változók) lineárisan függetlenek, korrelálatlanok, a tényezőváltozók és az eredményváltozó közötti páronkénti, (totális) korrelációs együtthatók négyzeteinek összege egyenlő a többszörös determinációs együtthatóval:

$$\dot{R}^2 = \sum_{j=1}^{k} r_{yj}^{\,2} \tag{7-87}$$

A fenti módon előállítható $\dot{R}^2$, a többszörös determinációs együttható **feltételezett értékét** jelenti. Amennyiben a feltételezett és tényleges determinációs együttható értéke nem egyezik meg, a különbség a multikollinearitás létét jelzi:

$$M_1 = \dot{R}^2 - R^2 \tag{7-88}$$

Másik globális mérőszám az **M mutató**. A determinációs együtthatóra vonatkozó fontos összefüggés alapján - amelynek bizonyításától eltekintünk -, számszerűsíthetjük az ún. **parciális hatás** mértékét, amely azt fejezi ki, hogy mennyivel növekszik a többszörös determinációs együttható értéke, ha egy adott változót utolsóként kapcsolunk be a modellbe. A különféle lehetséges módon kiszámított parciális hatások összege, értelemszerűen megegyezik a determinációs együtthatóval, ha nincs multikollinearitás.

A j-edik változó parciális hatása (jele: f_j):

$$f_j = R^2 - R^2_{y_{12 \ldots j-1, j+1 \ldots k}}$$

ahol:

R^2 - a k számú változó alapján számított többszörös determinációs együttható,

$R^2_{y_{12 \ldots j-1, j+1 \ldots k}}$ - a j-edik változó hatását nem tartalmazó többszörös determinációs együttható

Ha az R^2-ből kivonjuk a parciális hatások összegét, a különbség az R^2 -nek azon része, amelyet a tényezőváltozók együttesen magyaráznak meg. A multikollinearitást tehát a tényezőváltozóknak ezzel a nem elkülöníthető hatásával mérjük.

$$M = R^2 - \sum_{j=1}^{k} f_j \qquad (7\text{-}89)$$

A többváltozós modell tárgyalása során megismert példánkban, ahol a boltok évi forgalmát az alapterület és a létszám függvényében magyaráztuk, vizsgáljuk meg a multikollinearitás meglétét az eddig megismert módszerek segítségével[16]! Az analízishez szükségünk van a korrelációs mátrixra:

A modell egészét jellemző többszörös determinációs együttható:

$$R^2 = 0{,}936$$

A tényezőváltozók közötti korrelációs együttható (értéke: 0,903) ugyan kisebb, mint a többszörös korrelációs együttható (értéke:0,967), azonban a

[16] Itt a példa kedvéért eltekintünk attól, hogy a magyarázó változók eredetileg nem valószínűségi változók, hanem rögzített értékek voltak.

totális korrelációs együttható nagysága már felveti a multikollinearitás fennállásának veszélyét.

Az M_1 mutató értéke:

$$\dot{R}^2 = 0{,}945^2 + 0{,}943^2 = 1{,}7822$$
$$M_1 = 1{,}7822 - 0{,}936 = 0{,}8462$$

Az M mutató alkalmazása feltételezi a parciális hatások számszerűsítését. Példánkban - mivel csupán két magyarázó változóval van dolgunk - az adott változó hatását nem tartalmazó többszörös determinációs együttható megegyezik az eredményváltozó és a másik változó közötti totális korrelációs együtthatóval.

Az X_2 változó, a létszám parciális hatása:

$$f_2 = 0{,}936 - 0{,}945^2 = 0{,}043$$

Az X_1 változó, az alapterület hatása:

$$f_1 = 0{,}936 - 0{,}943^2 = 0{,}047$$

A parciális hatások összege és az R^2 alapján az M mutatószám értéke:

$$M = 0{,}936 - (0{,}047 + 0{,}043) = 0{,}846$$

A fenti mutatószámok jelzik a modellben a multikollinearitás jelenlétét.

A globális mutatószámok értékei példánkban megegyeznek, mivel modellünkben csak két tényezőváltozó szerepel.

c) A multikollinearitás feltárása **statisztikai próbákkal**.

A multikollinearitás jelenlétére következtethetünk a tényezőváltozók korrelációs mátrixának **determinánsából** is. Igazolható ugyanis, hogy amennyiben a magyarázóváltozók lineárisan függetlenek egymástól a modellt **ortogonálisnak** tekinthetjük. Az ortogonális rendszert leíró mátrix determinánsa 1-gyel, teljes multikollinearitás esetén viszont 0-val egyenlő. Érvényes tehát az alábbi reláció:

$$0 \le |\mathbf{R}_k| \le 1$$

A determinánsértékek valószínűségi eloszlása, egy egyszerű transzformáció után megközelítően χ^2-eloszlású. Az így képzett próbafüggvény annak a H_0 hipotézisnek a tesztelésére szolgál, amely szerint a magyarázó változók lineárisan függetlenek. A próbafüggvény az alábbi:

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2k+5)\right]\log|\mathbf{R}_k| \qquad (7\text{-}90)$$

A fenti függvényben a tényezőváltozók korrelációs mátrixa determinánsának ($|\mathbf{R}_k|$) tizes alapú logaritmusával számolunk. A próbafüggvény szabadságfoka:

$$\frac{1}{2}k(k-1)$$

Adott szignifikancia szint mellett a χ^2 táblában megkeressük a megfelelő kritikus értéket. Amennyiben a (7-90) próbafüggvény alapján számított érték nagyobb mint, a táblából vett, számottevőnek tarthatjuk a modellben a multikollinearitást. Ellenkező esetben, H_0 hipotézis elfogadása esetén, azonban még nem vethetjük el teljesen a jelenség meglétének feltevését, mivel az összetettebb kapcsolat kimutatására nem kellően érzékeny a módszer. A korrelációs mátrix (a változók közötti olykor eltérő erejű és ellenkező előjelű kapcsolatok miatt) ugyanis nem mindig képes rámutatni arra, ha kettőnél több tényező között áll fenn a multikollinearitás.

Példánkat folytatva teszteljük a multikollinearitást.

A tényezőváltozók korrelációs mátrixa:

$$\mathbf{R}_2 = \begin{bmatrix} 1 & 0{,}903 \\ 0{,}903 & 1 \end{bmatrix}$$

A mátrix determinánsa segítségével a próbafüggvény:

$$\chi^2 = -\left[10 - 1 - \frac{1}{6}(2 \times 2 + 5)\right] \log 0{,}1846 = 5{,}503$$

ahol:

$$\chi_{0,05}{}^2 = 3{,}841$$

A próbafüggvény szabadságfoka: 1.

A fentiek alapján megállapíthatjuk, hogy 5 %-os szignifikancia szinten nagy eséllyel állíthatjuk azt a modellről, hogy zavaró multikollinearitást tartalmaz.

Természetesen a fenti módszereken kívül a multikollinearitás feltárására egyéb eljárásokat is használ a gyakorlat azonban ezek nem képezik tananyagunk tárgyát[17].

A multikollinearitás kiküszöbölése

Csupán vázlatosan említjük meg ezen a helyen a multikollinearitás kiszűrésének módszereit.

- A legegyszerűbb eljárás, ha elhagyjuk azt a változót, amely "okolható" a multikollinearitásért. Ez azonban jelentős információ-veszteséggel jár és a kihagyott változó szinte majdnem biztosan rontja a becslést. Amennyiben a

[17]A teljesség igénye nélkül csupán megemlítjük a sajátérték-sajátvektor vizsgálatokat, illetve a ún. sugárkévék módszerét.

modell specifikációja helyes, a multikollinearitás csak a megfelelő információk hiányának tudható be, tehát a minta tulajdonsága.

- Kézenfekvően adódik a minta nagyságának a növelése, amely - mint tudjuk - információ-többletet eredményez. Erre azonban sok esetben - különösen idősoros vizsgálatok esetén - nincs mód, illetve sokszor nem biztosítható az, hogy a minta a korábbitól valóban független információt szolgáltasson.

- A **főkomponensek módszere** a regresszió-analízis kibővítését adja azáltal, hogy az eredeti magyarázó változókat a belőlük képzett lineárisan független mesterséges változókkal helyettesíti. A módszer hatékonyan csökkenti a multikollinearitás hatását, azonban jelentős információveszteséggel jár.

- Az elmúlt évtizedekben az ún. **ridge-regresszió** alkalmazása terjedt el. A módszer tulajdonképpen egy torzító változónak a modellbe történő beépítésével próbálja megoldani problémát. Segítségével nyert eredmények meglepően hatékonynak bizonyultak a gyakorlatban. Az így felépített modellek torzított becslések eredményeként jönnek létre, és a paraméterek interpretálhatósága nagyon kérdéses lesz.

Végezetül elmondhatjuk, hogy a multikollinearitás problémájára megnyugtató megoldást nem tudunk javasolni, a jelenséggel együtt kell élni, de hatásának ismerete megvéd a félreérthető információk megfogalmazásától.

7.4.3 Autokorreláció

A gazdasági jelenségeket leíró regressziós modellek felépítése során gyakorta modellezünk olyan ok-okozati rendszert, amelyben a változók megfigyelt értékei idősorokból származnak. A dinamikus elemek bekapcsolása a modellbe azonban további speciális gondokat vet fel, nevezetesen az **autokorreláció**, és **reziduális autokorreláció** problémáit.

A klasszikus lineáris regressziós modellekkel szemben támasztott 3. számú feltétel szerint autokorrelálatlannak tekinthetjük a regressziós modellt, ha a különböző megfigyelésekhez tartozó reziduális változók korrelálatlanok. Mivel a feltételrendszerünknek megfelelve az X változók nem valószínűségi változók, hanem rögzített értékek, és ennek megfelelően egy ismételt mintavétel esetén a különböző Y_i értékek a véletlen tényező hatásaként jönnek létre, definiálhatjuk általános értelemben az autokorrelációt. Ezek szerint az Y_t (Y_i) idősor (adatsor)[18] elemei között fellépő sztochasztikus kapcsolatot **autokorrelációnak**, vagy más néven **szeriális korrelációnak** nevezzük. Az autokorreláció mértékét az **autokorrelációs együtthatóval** mérjük.

A **p-ed rendű** (p időegységgel **késleltetett**) elméleti autokorrelációs együtthatót az egymástól p időegységnyi távolságra álló elempárok korrelációs együtthatójaként definiálhatjuk:

[18]Az autokorrelációt nem szűkíthetjük le csak az idősori adatokra, a jelenség ugyanis ennél általánosabb. Mivel az elemek sorrendje itt nagy fontossággal bír, és ez az idősoroknál kézenfekvő, a továbbiakban az idősoroknál megismert t indexet használjuk az eredmény-, és tényezőváltozók jelölésére.

$$\rho_p = \frac{cov(Y_t, Y_{t+p})}{\sigma_{Y_t}\sigma_{Y_{t+p}}}$$

ahol:

$cov(Y_t, Y_{t+p})$ - az Y_t Y_{t+p} változók kovarianciája,

σ_{Y_t} $\sigma_{Y_{t+p}}$ - a megfelelő szórások

Ha p=1, elsőrendű autokorrelációs együtthatóról beszélünk.

Számos időbeli adatsor esetén könnyen, logikailag is igazolható, hogy az idősor egyes értékei függvényei a megelőző számértékeknek, és ebben a kapcsolatban ez a hatás nem helyettesíthető más magyarázó változóval. Például a GDP adott évben mért értékét determinálja az előző év (évek) értéke (értékei), az idősorban egyébként meglévő általános trendhatáson kívül. Ilyen esetben az idősor egymást követő értékei között autokorreláció áll fenn, és az idősor autoregresszív folyamatot alkot. Az **autoregresszív folyamat** egy sajátos regressziós modellnek fogható fel, ahol a véletlen változót egy tiszta véletlen hatás az ún. **fehér zaj** reprezentálja.

$$Y_t = \varphi_1 Y_{t-1} + \ldots + \varphi_p Y_{t-p} + v_t \qquad (7\text{-}91)$$

ahol:

$\varphi_1 \ldots \varphi_p$ - autoregressziós együtthatók

v_t - fehérzaj

Természetesen az autoregresszív folyamatot - mivel az X változó nem valószínűségi változó - a fenti jelölésekkel írhatjuk fel a véletlen változóra, a hibatagra vonatkozóan is:

$$\varepsilon_t = \varphi_1\varepsilon_{t-1} + \ldots + \varphi_p\varepsilon_{t-p} + v_t \qquad (7\text{-}92)$$

A gazdasági idősorokból felépített regressziós modellek esetén gyakran élünk az elsőrendű autoregresszív kapcsolat feltételezésével. Ilyen esetben az elsőrendű autoregresszív együttható (φ_1) megegyezik az elsőrendű autokorrelációs együtthatóval (ρ_1). Mindezek figyelembevételével a gyakorlatban sok esetben használt modell:

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + v_t \qquad (7\text{-}93)$$

Itt kell megjegyeznünk, hogy a magasabb rendű autoregresszív modellek esetén a paraméterek közötti összefüggés nem ilyen közvetlen. Az autoregresszivitás rendjének pontos meghatározása csak a folyamat belső összefüggésének további vizsgálata útján lehetséges, célszerű függvények alkalmazása mellett.

Az autokorreláció megítélése

Amennyiben egy regressziós modellben nem teljesül a modellekkel szemben megfogalmazható azon feltétel, amely szerint a hibatényező értékei páronként nem korrelálnak egymással, tudjuk, hogy a modell **autokorrelált**. Ez arra utal, hogy a vizsgált jelenség kapcsán valamilyen specifikációs hibát követtünk el. A specifikációs hiba típusának ismerete alapvető fontossággal bír a probléma megoldása során.

A legegyszerűbb feltevés szerint az autokorreláció csak abból ered, hogy a reziduális változók autokorreláltak, de a modell által megmagyarázott rész tökéletesen specifikált. Ebben az esetben, de csak ebben az esetben, a becslés módszerének módosításával a probléma könnyen megoldható.

Sok esetben azonban a valóság ennél bonyolultabb. Előfordulhat ugyanis, hogy az autokorreláció oka szorosan kapcsolódik a modell magyarázott részéhez. Különösen három ilyen eset[19] érdemel figyelmet: **kihagyott változók esete, függvényforma helytelen megválasztása,** és az **adattranszformációs hiba.** A különböző típusok felismerésében statisztikai próbák segítik a felhasználót, de alapvető minden esetben a vizsgálandó jelenség elméleti hátterének az ismerete is. Itt kell megemlíteni, hogy az adattranszformációs hiba az idősorokban pl. szezonális hatások szűrésénél, illetve differenciák alkalmazása kapcsán[20] lép fel.

Az autokorreláció jelensége nem szükségszerűen kötődik az idősorokból képzett modellekhez, - amint már utaltunk rá - keresztmetszeti vizsgálatok esetén is felléphet, azonban főleg a dinamikus specifikációk esetén kell számolnunk hatásával.

Általánosságban elmondhatjuk, hogy az autokorreláció jelenléte mellett készített paraméter-, és pontbecslések ugyan torzítatlanok maradnak, de nem lesznek hatásosak. Különösen óvatosan kell kezelni az autokorrelált modellt, ha segítségével előrebecsléseket kívánunk készíteni. Autokorrelált modellek esetében az együtthatók standard hibái torzítottak, így sem a standard hibákhoz kapcsolódó próbák, sem az előrebecslésekhez kapcsolódó konfidencia intervallumok nem használhatók fel.

Az autokorreláció kimutatása, valamint "kezelése" illetve hatásának figyelembevétele a becslés módszerének kiválasztása igen nagy körültekintést igényel. A fejezet további részében az autokorreláció egyik elterjedt tesztelési módszerével, és egy hatékony kezelési módjával ismerkedünk meg.

---

[19]Ezek az esetek a következő alfejezetben tárgyalandó heteroszkedaszticitás problémájánál is viszatérnek.
[20]Ezekkel a fogalmakkal a 8. fejezetben ismerkedünk meg.

Az autokorreláció tesztelése

Tananyagunkban csupán az elsőrendű autokorreláció kimutatására alkalmas statisztikai próbával ismerkedünk meg. Természetesen itt is fel kell hívni a figyelmet arra, hogy az autokorreláció kimutatására egyéb eljárások nagy számban találhatók az irodalomban.

Tételezzük fel, hogy érvényesül a lineáris megközelítési mód, és a t-edik időszak (időpont) eredményváltozója az alábbi modellel írható fel:

$$y_t = \mathbf{x}_t^T\beta + \varepsilon_t \qquad (7\text{-}94)$$

ahol:

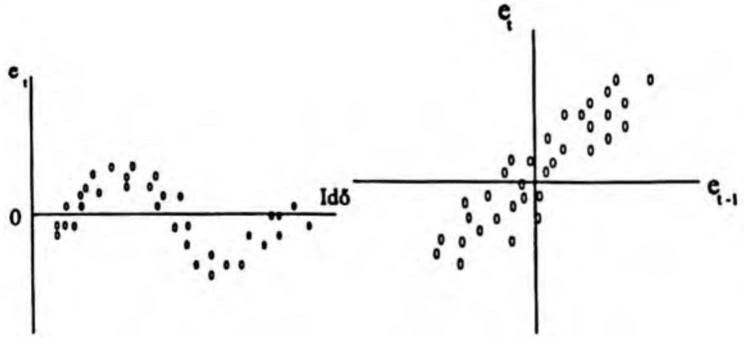
$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t \quad |\rho| < 1 \text{ és } E(v) = 0$$

A fenti autoregresszív modellben amennyiben a ρ autokorrelációs együtthatónak az értéke eltér nullától a regressziós modell autokorrelált, míg ellenkező esetben a regressziós modellben szereplő reziduális változó megegyezik a tiszta véletlen hatással (gyakorlatilag a ún. fehér zajnak felel meg), tehát a modell jól specifikált és autokorrelálatlan. Az autokorreláció tényének eldöntése tulajdonképpen az alábbi hipotézis tesztelésének felel meg:

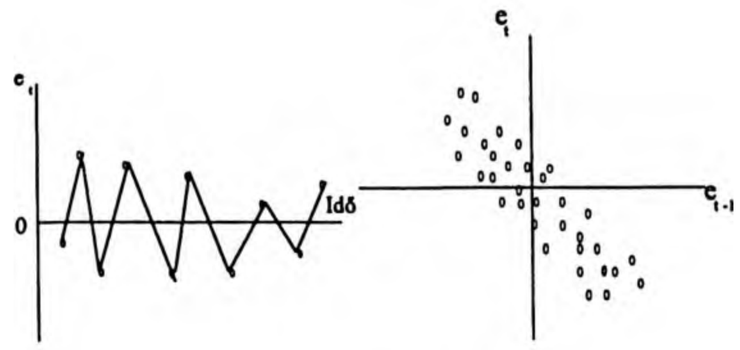
$$\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array} \qquad (7\text{-}95)$$

A nullhipotézis elfogadása azt jelenti, hogy a reziduális változó véletlen jellegű, a szomszédos értékek egymástól függetlenek.

A fenti kétoldalú próbával természetesen tesztelhető ρ pozitív és negatív értékének szignifikanciája is. Az első esetben H_1:ρ>0, míg a második esetben a H_1:ρ<0 a megfelelő alternatív hipotézis. Az autokorrelálatlanság tényét csak a kétoldalú próba alkalmazásával bizonyíthatjuk megnyugtató módon. A különféle előjelű autokerrelációt, vagy más megfogalmazásban az autokorrelációs folyamatot az alábbi ábrával próbáljuk meg szemléltetni:



a)



b)

7-14. ábra Pozitív (a) és negatív (b) autokorreláció

Az ábrából világosan kitűnik, hogy a pozitív autokorreláció esetén az egymást követő értékek "vonzzák" egymást, míg negatív autokorreláció esetén "taszítják" egymást. Ily módon megállapíthatjuk, hogy az előjelek váltása, azok számának alakulása végső soron jelzi a reziduális változók sorozatának véletlenszerűségét, vagyis az autokorreláltságot.

Az autokorreláció kimutatására leggyakrabban használt módszer az ún. **Durbin-Watson d-próba**. A próbafüggvény előállításához első lépésben ki kell számítani a tapasztalati reziduumokat (e_t), amelyeket a klasszikus legkisebb négyzetek módszerének alkalmazásával nyerünk. A próbafüggvény:

$$d=\frac{\sum_{t=2}^{n}\left(e_t-e_{t-1}\right)^2}{\sum_{t=1}^{n}e_t^{\,2}} \qquad (7\text{-}96)$$

A d-próbafüggvény értelmezéséhez hasznos segítséget ad a ρ autokorrelációs együttható és a d között felírható közelítő összefüggés. Mindenekelőtt írjuk fel a $\hat{\rho}$ mutatószámának becslőfüggvényét!

$$\hat{\rho}=\frac{\sum_{t=2}^{n}e_t e_{t-1}}{\sum_{t=2}^{n}e_t^{2}} \qquad (7\text{-}97)$$

Eléggé nagyszámú megfigyelés esetén a d próbafüggvény és a ρ autokorrelációs együttható között az alábbi közelítő összefüggés írható fel:

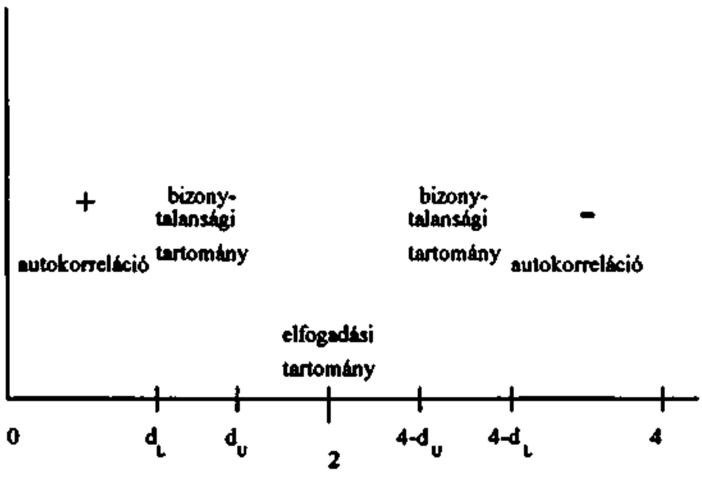
$$d\approx 2-2\hat{\rho}=2\left(1-\hat{\rho}\right) \qquad (7\text{-}98)$$

A fenti összefüggésből nyilvánvaló, hogy amennyiben a modell autokorrelálatlan (ρ=0), a kiszámított d=2. Pozitív autokorreláció erősödése esetén d értéke közelít nullához, míg erősödő negatív autokorreláció esetén a 4-hez tart.

A d-próba elvégzéséhez a mintából egy klasszikus legkisebb négyzetek módszerével történt becslés segítségével meghatározzuk d értékét, majd az egyenletben szereplő magyarázó változók és a megfigyelések számának, valamint a próba adott szignifikancia-szintjének megfelelően megkeressük a Durbin-Watson táblából a megfelelő értékeket. A táblázatban szereplő értékek közvetlenül a pozitív autokorreláció tesztelésére ($H_1: \rho > 0$) alkalmasak, amelyeket d_L (alsó) és d_U (felső) értékekkel valósíthatunk meg. A negatív autokorreláció ($H_1: \rho < 0$) tesztelése, illetve a kétoldalú hipotézisek ellenőrzése újabb értékek kiszámítását igényli, amelyek d eloszlásának szimmetrikus jellege miatt nem okoznak különös gondot:

$$\begin{aligned} d_L' &= 4 - d_U \\ d_U' &= 4 - d_L \end{aligned} \qquad (7\text{-}99)$$

A próba lehetséges kimeneteleit, a döntési sávokat jól szemlélteti az alábbi ábra:



7-15. ábra A Durbin-Watson d-próba döntési sávjai

A próba alkalmazásának viszonylagos hátránya az ún. bizonytalansági tartomány megléte, amely a gyakorlatban sok gondot okoz. Az irodalomban többféle módon igyekeztek a problémát feloldani, amelyek közül a legegyszerűbbnek tűnő megoldás az, amikor a bizonytalansági tartományt az elutasítási tartományhoz csatolják.

Az autokorreláció tesztelésének megértését az alábbi számpélda segítheti. 1970 és 1987 között vizsgáltuk a hazai összes fogyasztás és az adott évben megtermelt bruttó hazai termék (GDP) közötti sztochasztikus összefüggést, regressziós modell segítségével. A kiinduló adatokat az alábbi tábla szemlélteti:

GDP és fogyasztás

7-12. tábla

Év	GDP (MdFt)	Fogyasztás (MdFt)
1970	332,2	228,7
1971	362,5	250,3
1973	394,6	266,2
.	.	.
.	.	.
.	.	.
1985	1.012,3	753,9
1986	1.088,8	811,5
1987	1.226,4	904,8

Természetes kapcsolatnak tételezhetjük fel, ha a fogyasztást az adott év GDP értékével magyarázzuk. A klasszikus legkisebb négyzetek módszerével nyert regressziófüggvény az alábbi (zárójelben a standard hibák szerepelnek):

$$\hat{y} = -44,17 + 0,775x$$
$$(7,0) \quad (0,009)$$

A paraméterek a megszokott 5 %-os szignifikancia szinten valóságos összefüggést jeleznek. A d-statisztika meghatározásához szükséges adatok:

Munkatábla

7-13 tábla

Év	y	$\hat{y}$	e_t	e_t^2	$(e_t-e_{t-1})^2$
1970	228,7	213,4	15,2	233,88	—
1971	250,3	236,9	13,4	179,53	3,59
1972	266,2	261,8	4,4	19,44	80,81
1973	285,9	291,6	-5,7	32,98	103,07
1974	311,9	309,9	1,9	3,83	59,30
1975	335,3	329,2	6,1	37,56	17,39
1976	359,2	364,9	-5,7	32,68	140,31
1977	390,4	406,0	-15,6	243,73	97,92
1978	426,1	442,9	-16,8	285,53	1,65
1979	472,8	483,9	-11,1	124,05	33,17
1980	515,3	514,8	0,4	0,18	133,69
1981	556,8	566,9	-10,1	102,09	110,84
1982	599,3	608,0	-8,7	75,67	1,97
1983	642,0	637,5	4,5	19,88	173,13
1984	695,7	690,5	5,2	27,04	0,55
1985	753,9	740,7	13,2	173,06	63,28
1986	811,5	800,1	11,4	130,84	2,95
1987	904,8	906,7	-1,9	3,82	179,36
Σ	----	----	—	1725,80	1203,00

Behelyettesítve a d próbafüggvénybe a megfelelő számértékeket:

$$d = \frac{1.203}{1.725,8} = 0,697$$

A táblázatban n=18 mellett, egy magyarázóváltozó esetén 5 %-os szignifikancia-szinten a kritikus értékek:

$$d_L=1,16 \qquad \text{és} \qquad d_U=1,39$$

Mivel a táblabeli alsó értéknél a számított érték kisebb, a modellt autokorreláltnak tartjuk. A két számérték viszonyából azt is megállapíthatjuk, hogy modellünkben pozitív autokorreláció van. Az autokorreláció tényét támasztja alá a (7-97) alapján kiszámítható becsült autokorrelációs együttható is ($\hat{\rho} = 0,651$).

Becslés autokorrelált reziduum esetén

Amint a korábbiakban szóltunk róla szignifikáns autokorreláció esetén a klasszikus legkisebb négyzetek hagyományos módszerével nyerhető előrebecslések félreinformálhatnak. Amennyiben az autokorreláció eredete a modellben a nem megmagyarázott részben található, a modell módosítása helyett egy **iteratív paraméterbecslést** célszerű elvégezni. Ez a módszer feltételezi az autokorrelációs együttható előzetes a priori ismeretét. Az autokorrelációs együttható segítségével egy egyszerű transzformációt hajtunk végre, amelynek eredményeként a "káros" hatás mértéke csökken illetve az megszűnik. Az eljárást az alábbiakban foglalhatjuk össze:

$$y_t - \hat{\rho} y_{t-1} = b_0 (1 - \hat{\rho}) + b_1 (x_{1t} - \hat{\rho} x_{1t-1}) + \ldots + b_k (x_{kt} - \hat{\rho} x_{kt-1}) \qquad (7\text{-}100)$$

A fenti regressziós függvény becslése során előzetes feltevést teszünk az autokorrelációs együtthatóra vonatkozóan. A kiinduló érték meghatározása többféle módon történhet. Kézenfekvő megoldás, ha első lépésben az együtthatónak 0 kezdő értéket adunk (ekkor a paramétereket tulajdonképpen a klasszikus legkisebb négyzetek módszerével határozzuk meg!), és az így nyert egyenlet alapján nyerhető reziduumok képezik az autokorrelációs együttható kiszámításának alapját, amit a továbbiakban kiinduló értéknek kezelünk. Ezt az együtthatót egy következő iterációs lépésben behelyettesítjük az egyenletbe, ismét elvégezzük, most már a módosított becslést, aminek eredményeként új paramétereket számszerűsíthetünk. Az iteratív eljárást vagy addig folytatjuk, ameddig a regressziós együtthatók az egyik lépésről a másikra nem változnak lényegesen, vagy addig, amíg a modell már nem tartalmaz autokorrelált reziduumokat.[21] Itt igen jelentős szerep hárul, az eljárás során, a Durbin-Watson féle d-próbára, amelynek segítségével az autokorreláció lépésről-lépésre tesztelhető.

A fenti eljárást általánosított differenciák módszereként is ismerik, ami más megfogalmazásban - a klasszikus legkisebb négyzetek módszerénél általánosabb paraméterbecslési eljárásnak -, az ún. **az általánosított legkisebb négyzetek módszerének** felel meg, és a becsléshez szükséges $\hat{\rho}$ előállítsa történik az iteratív módszerrel[22].

További részletes ismertetés helyett csupán megjegyezzük, hogy ez a becslési módszer igen fontos szerepet tölt be az ökonometria módszerei között.

Az előző példánkban szignifikáns autokorrelációt állapítottunk meg. Ismerjük az elsőrendű autokorrelációs együttható becsült értékét, ($\hat{\rho} = 0{,}651$), amit ha behelyettesítünk a (7-100) összefüggésbe, és a paraméterbecslést ismételten elvégezzük az alábbi függvényt számszerűsítjük (zárójelben a standard hibákat közöljük):

$$\hat{y}_t = \underset{(6{,}8)}{24{,}75} + \underset{(0{,}03)}{1{,}27}x_t$$

Látható, hogy a paraméterek szignifikánsak.

Míg a b_1 regressziós együtthatót közvetlenül értékelhetjük, a b_0 paramétert át kell alakítani ahhoz, hogy az eredeti modellnek megfeleljen. Ugyanis

$$b_0(1-\hat{\rho}) = 24{,}75$$

átrendezve

$$\frac{24{,}75}{1-0{,}651} = 70{,}9$$

[21] Ezt az eljárást Cochrane-Ourcutt iterációs eljárás néven ismeri az irodalom.
[22] Lásd! Irodalom [31] 143. old.

Mindezek figyelembevételével a regressziófüggvény:

$$\hat{y}_t = 70,9 + 1,27x_t$$

A regressziós függvényt az autokorreláció szempontjából kedvezően ítélhetjük meg, ugyanis a módosított becslés után a d-statisztika értéke:

$$d = 1,72$$

$$d_U = 1,39 \text{ és } d_L' = 2,6$$

az elfogadási tartományba esik.

A példa számszerű értékei alapján látható, hogy amennyiben az autokorreláció hatását nem vettük volna figyelembe, a közvetlenül kapott regressziós paraméter a fentitől jelentősen eltérne. Természetesen a két modell között a vizsgálat céljának (elemzés vagy előrebecslés) ismeretében tudunk választani.

7.4.4 Heteroszkedaszticitás

A klasszikus lineáris regressziós modell 2. feltevése értelmében, amennyiben egy regressziós modell **hibatényező varianciája állandó, a modell homoszkedasztikus.** Ha ez a feltétel nem teljesül, **heteroszkedasztikus** reziduumokkal rendelkezik a modell, becslése különböző problémákat vet fel.

A regressziós becslésekkel foglalkozó szakkönyvek jelentős része megemlíti, hogy a reziduális változók szórásának állandósága a minta egészében nem teljesül a fogyasztás és jövedelem kapcsolatának vizsgálata során. Ugyanis a nagyobb jövedelemmel (X) rendelkező háztartások, egyének fogyasztása (Y) tágabb intervallumban szóródik, mint a kisebb jövedelemmel rendelkezőké. A fogyasztás varianciája ebben az esetben sztochasztikus kapcsolatban van a jövedelem nagyságával, így magától értetődik, hogy a fogyasztással összefüggő jövedelemrugalmasság szórása is a magasabb jövedelmi sávokban nagyobb. Ezt a jelenséget az irodalom heteroszkedaszticitásnak hívja. Megfigyelésére a modell reziduális változói adnak lehetőséget. Meg kell említenünk, hogy a heteroszkedaszticitás veszélye elsősorban a keresztmetszeti adatok felhasználása esetén nagy, de - mivel a reziduális változók tulajdonságáról van szó idősorokban is előfordulhat. (Sokszor a probléma gyökere rokon az autokorrelációval.)

A heteroszkedaszticitást előidéző okok sokfélék lehetnek. Illusztrációként megemlítünk néhány példát. A már ismertetett jövedelem-fogyasztás kapcsolata mellett számíthatunk a heteroszkedaszticitásra amikor valamilyen munkafolyamat gyakorlati idejének növekedése mellett csökken a megmunkálási idő szórása. Nagy eséllyel bekövetkezik a heteroszkedaszticitás, ha nagytömegű adatbázis esetén, az adatokat átlagukkal helyettesítjük (csoportosított adatok), ha helytelenül választjuk meg a függvényformát, kihagyunk releváns változókat, vagy ha különböző intervallumban eltérő a jelenség struktúrája. A felsoroltak egyben azt is jelentik, hogy a heteroszkedaszticitás gyakorta specifikációs probléma, így néha a modell-specifikáció revíziója is kielégítő eredményre vezethet.

A heteroszkedaszticitás jelensége esetén a regressziós együtthatók becslése ugyan torzítatlan, de a paraméterek varianciája már torzított, így a standard hibák használata megkérdőjelezhető, és a segítségükkel elvégzett próbák és becslések félreinformálhatnak.

A heteroszkedaszticitás feltárása

A tanulmányaink során már megismertük, hogy a standard lineáris regressziós modell változóira vonatkozó feltevések hipotetikus jellegűek, amelyek különféle eljárásokkal vagy megerősíthetők, vagy elvetendők. A reziduális változó varianciájára vonatkozó feltevés is hipotézis-rendszer formájában fogalmazható meg:

$$
\begin{aligned}
&H_0 : \sigma_1^2 = \ldots = \sigma_n^2 \\
&H_1 : \exists \sigma_i^2 \neq \sigma_j^2 \qquad (7\text{-}101) \\
&\text{ahol: } i \neq j
\end{aligned}
$$

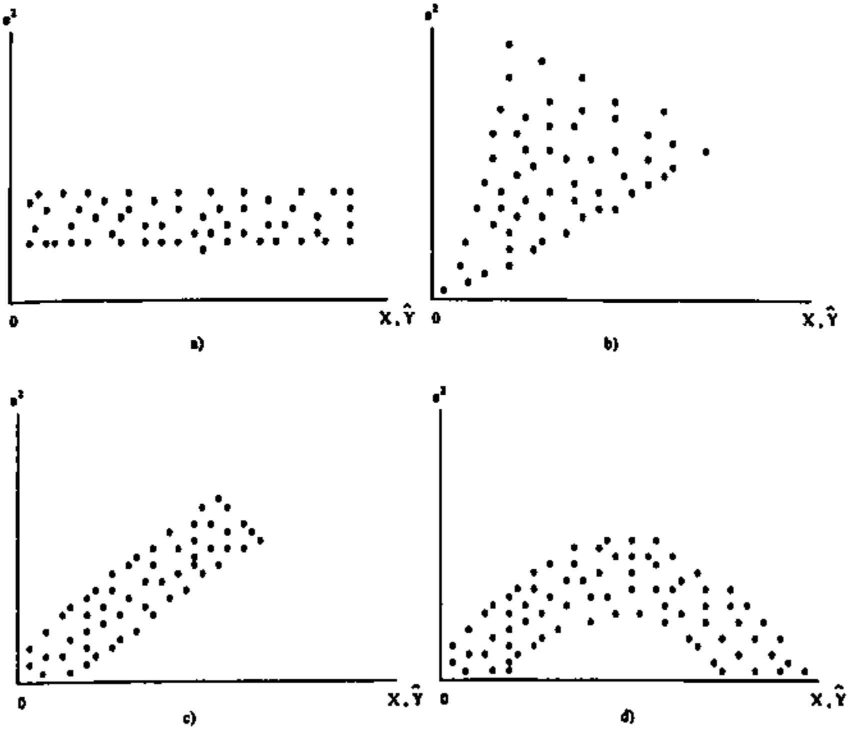
A nullhipotézis teljesülése egyben azt is jelenti, hogy a modell homoszkedasztikus, míg az alternatív hipotézis a heteroszkedaszticitás feltételezését szimbolizálja.

Az alábbiakban a heteroszkedaszticitás felismerésének néhány elterjedtebb módszerét mutatjuk be vázlatosan.

Grafikus módszer

A heteroszkedasztikus modellt legegyszerűbben a reziduális változók értékeinek grafikus ábrája segítségével ismerhetjük fel. Az eljárás feltételezi azt, hogy korábban, egy célszerűen megválasztott becslési módszerrel már meghatároztuk a modell paramétereit és segítségükkel előállítottuk a reziduumokat (e_i).

Az alábbi ábra néhány tipikus esetet mutat be.



7-16 ábra A becsült reziduumok négyzetének pontdiagramja

Az ábrán a reziduális változó négyzetét X, illetve $\hat{Y}$ változók függvényében ábrázoltuk. Itt kell megjegyeznünk, hogy amennyiben a modell csak egy magyarázó változót tartalmaz a kétféle ábrázolási mód megkülönböztetésének nincs értelme, míg többváltozós modell esetén az eredményváltozó függvényében ábrázolt reziduum a modell egészére ad információt, szemben a tényezőváltozók függvényében történő ábrázolással.

Az a) típusú pontdiagram esetén, nagy eséllyel feltételezhetjük, hogy a modell homoszkedasztikus. A b), c) és d) ábra a heteroszkedaszticitás tényére hívja fel a figyelmet, illetve esetleg valamilyen specifikációs hibára utal. Amennyiben c) típusú az ábra joggal vélhetjük azt, hogy lényeges változó maradt ki a modellből, míg a d) típus egy helytelenül megválasztott függvénytípus esetére vall (pl. lineáris függvény multiplikatív helyett). A c) esetben a reziduumok négyzete és a változó(k) között lineáris a kapcsolat, míg a d) esetben valamilyen négyzetes összefüggés áll fenn. Ezek a felismerések a jelenség kiküszöbölésének módját is behatárolják.

A fejezet további részében a heteroszkedaszticitás feltárására kidolgozott statisztikai próbák közül néhányat bemutatunk vázlatosan.

Goldfeld-Quandt-próba

Paraméteres próba segítségével jó hatékonysággal lehet diagnosztizálni a heteroszkedaszticitást. A Goldfeld-Quandt-próba kiinduló feltevése szerint a reziduális változó szórását nem ismerjük, de feltesszük, hogy az a magyarázó változók négyzetével arányosan változik. A próba menete az alábbi:

- A heteroszkedaszticitásban feltehetően meghatározó szerepet játszó X_j magyarázó változó növekvő sorrendbe rendezett értékei szerint írjuk fel a reziduumokat, és a többi változót.
- Kiválasztunk c számú középen elhelyezkedő értéket, amelyeknek megfelelő változók értékeit kihagyjuk a további számításokból. (Célszerű, ha pl. 30 megfigyelés esetén ez a szám nem haladja meg a 8-at!)
- A klasszikus legkisebb négyzetek módszere segítségével az első (n-c)/2 és az utolsó (n-c)/2 megfigyeléshez külön-külön regressziós függvényt illesztünk.
- Kiszámítjuk a két regressziófüggvény reziduális négyzetösszegét:

$$S_1^2 = \sum_{i=1}^{(n-c)/2} e_i^2$$

$$S_2^2 = \sum_{i=\frac{n+c}{2}+1}^{n} e_i^2$$

ahol i követi az X_j növekvő sorrendjét.

- A próbafüggvény:

$$F = \frac{S_2^2}{S_1^2} \qquad (7\text{-}102)$$

ahol mind a számláló mind a nevező szabadságfoka egyaránt: (n-c-2k)/2

- Amennyiben a számított F-érték nagyobb, mint egy α szignifikancia szinthez tartozó F_α érték a nullhipotézist elvetjük, a modellt heteroszkedasztikusnak tekintjük.

Glejser-próba

A Glejser-próba során a priori feltételezzük, hogy a reziduális változó szórásnégyzete, vagy szórása a tényezőváltozó(k) pontosan leírható függvénye.

Amennyiben feltételezzük, hogy a j-edik tényezőváltozó függvénye a **reziduális szórás**, azaz a reziduumok abszolut értéke X_j változó függvénye, akkor felírható egy pótlólagos regressziós egyenlet:

$$|e_i| = \alpha_0 + \alpha_1 x_{ji} + v_i \qquad (7\text{-}103)$$

ahol: x_{ji} a j-edik magyarázó változó i-edik értéke.

Az α_0 és α_1 regressziós együtthatók szignifikáns voltát a t-statisztika segítségével ellenőrizhetjük, felhasználva a paraméterek standard hibáit. Amennyiben egyik paraméter sem tér el szignifikánsan nullától a modell homoszkedasztikus; ha csak az α_1 tér el szignifikánsan nullától a Goldfeld-Quandt féle teória érvényesül (azaz a reziduális szórás a tényezőváltozó négyzetével arányos), ha mindkét paraméter szignifikánsan eltér nullától akkor a heteroszkedaszticitás egy kvadratikus függvénnyel jellemezhető (a reziduális szórás a tényezőváltozó egy összetett kvadratikus függvényével arányos).

Feltételezhetjük azt is, hogy a **reziduális variancia** a tényezőváltozó(k) függvényében változik. Ekkor az alábbi pótlólagos regressziós modell használható fel az elemzés céljára:

$$e_i^2 = \alpha_0 + \alpha_1 x_{ji} + v_i \qquad (7\text{-}104)$$

A levonható következtetések nagy vonásokban megegyeznek a fent leírt megállapításokkal.

Természetesen a fenti modellek kiterjeszthetők egyidőben k magyarázó változóra is. Ilyen esetben a modellt az F-próba segítségével teszteljük (szabadságfokok: k és n-k-1), és segítségével a heteroszkedaszticitás globális próbáját végezzük el.

Szroeter-féle próba

A heteroszkedaszticitás felismerésének a Szroeter-féle próba egy olyan eljárása, amely kötődik az autokorrelációnak a d-statisztika segítségével történő teszteléséhez.

A homoszkedaszticitás már megismert nullhipotézisével szemben az alábbi alternatív hipotézist fogalmazhatjuk meg:

$$H_1 : \sigma_1^2 \leq \ldots \leq \sigma_n^2 \qquad (7\text{-}105)$$

ahol legalább egy esetben teljesül az egyenlőtlenség.

Természetesen itt a varianciák növekvő sorozatát tételeztük fel (mint a Goldfeld-Quandt-próba esetén), de a hipotézis csökkenő sorozat esetére is felírható. Mindenesetre a további számítások céljából célszerű ha reziduumok növekvő sorrendjében írjuk fel a tényezőváltozókat.

A próbafüggvény az alábbi módon definiálható:

$$h = \frac{\sum_{i=1}^{n} h_i e_i^2}{\sum_{i=1}^{n} e_i^2} \qquad (7\text{-}106)$$

ahol a h_i súlyszámokat az alábbi képlet segítségével határozhatjuk meg:

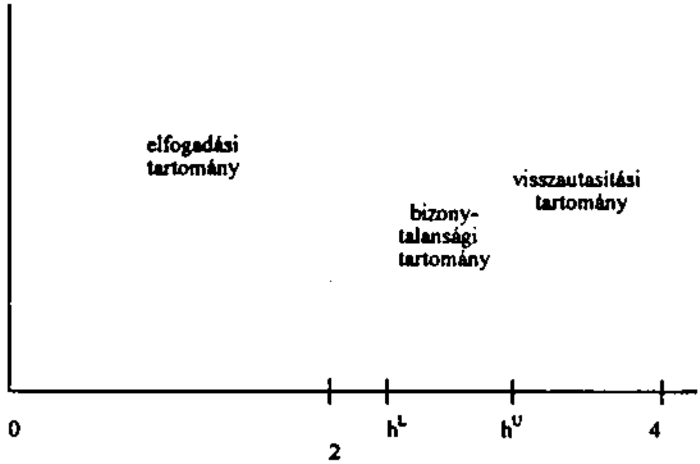
$$h_i = 2\left[1 - \cos\left(i \times \pi / n + 1\right)\right]$$

A próba kritikus értékeinek meghatározására többféle eljárás ismert, amelyek közül egyszerűsége miatt a Durbin-Watson táblázaton alapuló módszer érdemel említést. A két kritikus érték, amely egy bizonytalansági tartományt definiál az alábbi módon határozható meg:

$$h^L = 4 - d^U_{(n+1,k+1)}$$
$$h^U = 4 - d^L_{(n+1,k+1)}$$

ahol k a magyarázóváltozók száma.

Egyértelműen heteroszkedasztikusnak tekintjük a modellt, ha $h>h^U$, és homoszkedasztikusnak ha $h<h^L$. Amennyiben az empirikus adatok alapján kiszámított h érték a két kritikus érték közé esik, az alábbi ábrának megfelelve döntésünk bizonytalan.



7-17. ábra A Szroeter-próba döntési szabálya

A heteroszkedaszticitás felismerésének illusztrálására a korábban megismert kereskedelmi vállalati adatokat tartalmazó példánkat a megfigyelések számának növelésével kibővítettük. Erre alapvetően azért volt szükség, mivel a korábbi 10 elemszámú minta nagysága miatt nem alkalmas a heteroszkedaszticitás tesztelésére. A fentieknek megfelelve, a kereskedelmi vállalat 37 üzletére vonatkozóan vizsgáltuk meg az üzletek alapterülete és a forgalom közötti kapcsolatot az alábbi regressziós függvény segítségével:

$$\hat{y} = 9.604,9 + 171,6x$$
$$(3.883,3) \quad (25,56)$$

A modell és paraméterei szignifikáns összefüggésre utalnak ($t_{0,05(35)} = 2,03$), a korrelációs kapcsolat a közepesnél erősebb ($R^2 = 0,563$).

Az alábbiakban felsoroljuk a különféle próbák esetén számított, és a megfelelő táblabeli értékeket.

A Goldfeld-Quant-próba függvényének értéke: 5,57

ugyanakkor $F_{0,05} = 2,4$.

A Glejser-próba esetén a reziduumok abszolút értékei mellett illesztett függvény paraméterei alapján számított t-statisztikák:

$t_{\alpha_0} = -1,3$, $t_{\alpha_1} = 4,37$, ugyanakkor a $t_{0,05} = 2,03$

A Szroeter-féle próba értéke: $h = 3,26$ ugyanakkor a $h^L = 4 - 1,59 = 2,41$ és $h^U = 4 - 1,37 = 2,63$.

Megállapíthatjuk tehát, hogy mindegyik általunk használt próbával szignifikáns heteroszkedaszticitást diagnosztizáltunk.

Becslés heteroszkedasztikus modell esetén

A diagnosztizált heteroszkedaszticitás lehetőséget ad arra, hogy ismeretében "javítsuk" becslésünket. Kézenfekvő mód - ha erre lehetőség van - a modell újraspecifikálása (változók beválasztása illetve elhagyása, a függvényforma módosítása). Sokszor azonban a specifikáció módosítása nem vezet eredményre, ilyen esetben a becslési eljáráson kell változtatni. Különféle összetettebb paraméterbecslési eljárások, így például az **általánosított legkisebb négyzetek módszere**, továbbá az **instrumentális változók módszere** igen hatékony, jó becsléseket eredményez heteroszkedasztikus modellek esetén is. Mindezek tárgyalása azonban túlnő tananyagunk keretein.

8. AZ IDŐSOROK ELEMZÉSE

A számszerűsíthető társadalmi-gazdasági jelenségek statisztikai elemzése során nagy jelentősége van az időbeli összehasonlításnak, a különböző időpontokban, illetve időszakokban mért értékek összevetésének.

A 2.1 fejezetben a statisztikai sorok között már definiáltuk az idősort. Későbbi tanulmányaink során megismerkedhettünk a sztochasztikus kapcsolatokkal, így lehetőség nyílik arra, hogy az időben lejátszódó folyamatokat is mint sztochasztikus kapcsolatot közelítsük.

Az idősor felfogható Y valószínűségi változók olyan sorozatának, amelyet az időkoordináta T értékeihez rendelünk.

Az Y_1, Y_2, ...,Y_n változókat **elméleti idősornak** nevezzük, amelynek csak egyetlen realizációja ismert, a megvalósult idősori értékek sorozata. A regressziószámítással ellentétben nincs lehetőség elméletileg sem a minta növelésére sem új mintavételre.

A realizált idősori értékeket **tapasztalati idősornak** nevezzük és a továbbiakban y_t-vel (t=1,2,....,n) jelöljük[1].

Az idősorelemzésnek két fő megközelítési módja ismert.

A **determinisztikus idősorelemzés** abból a feltételezésből indul ki, hogy az idősort tartós hosszú távú tendencia (trend), tartósan ható szabályos jól modellezhető hullámmozgások (szezonalitás) határozzák meg és ezektől eseti-egyedi eltérítő hatást

[1]A t a latin tempus szóból származik.

eredményez a véletlen. A determinisztikus[2] jelzőt itt más értelmezésben használjuk, mint a sztochasztikus kapcsolatoknál megismert függetlenség, sztochasztikus függőség, determinisztikus meghatározottság hármas összefüggését. Tehát az idő múlása és az idősori realizált értékek között összefüggést vélelmezünk.

A **sztochasztikus idősorelemzés** abból a feltételezésből indul ki, hogy az aktuális idősori értékeket korábban realizálódott értékei és a véletlen hatás alakítja ki, a determinisztikus modellezés feltételezte hosszú távú tendencia befolyásoló szerepe ebben a megkülönböztetésben közvetlenül nem jelenik meg.

A továbbiakban részletesebben a determinisztikus modellezéssel foglalkozunk. Ennek két indoka van:

- történetileg ez alakult ki előbb, és felhasználhatóságát nem kérdőjelezi meg a sztochasztikus megközelítés kimunkálása és elterjedése,

- a determinisztikus elemzés és a regressziószámítás ismeretanyaga, kiegészítve a sztochasztikus elemzés alapjainak bemutatásával[3] - kellő alapismeretet nyújt a sztochasztikus modellezés szakirodalomból önképzéssel történő elsajátításához.

8.1 AZ IDŐSORELEMZÉS EGYSZERŰBB ESZKÖZEI

A **tapasztalati idősor**, mint az elemzés adatbázisa, már egyszerű rátekintéssel elárul valamit a vizsgált jelenség időbeli alakulásáról.

[2] A latin determinatio (be)határolás szóból ered.

[3] Röviden a 8.6. alfejezetben szólunk erről.

Az idősorok **grafikus megjelenítése** - koordinátarendszerben , vagy azon kívül - szemléletesebbé teszi a tendenciákat, így segítve az elemzést. Az egyszerűbb elemzési eszközök közé sorolhatók a **dinamikus viszonyszámok** (lásd! 2.2 fejezet).

A **bázis-**, és **láncviszonyszámok**, mint egyszerűen meghatározható mutatószámok, gyors, előzetes tájékoztatást adnak a vizsgált idősor alakulásáról.

Az egyszerűbb elemzési eszközök további csoportját alkotják az idősorelemzés területén alkalmazható **speciális átlagok**.

Az 1.5 fejezetben megismerkedtünk az álló- és mozgó sokaság fogalmával, illetve az ezekhez kötődő állapot és tartamidősorral.

Ha egy időszakra vonatkozó **tartamidősor** átlagos értékére vagyunk kíváncsiak, akkor fenntartások nélkül alkalmazható az **egyszerű számtani átlag**:

$$\bar{y}_t = \frac{\sum_{t=1}^{n} y_t}{n} \qquad (8\text{-}1)$$

Vizsgáljuk meg a magyarországi tehéntej-felvásárlás havi átlagos értékét 1991-ben![4].

[4]Forrás: Statisztikai Havi Közlemények 1992.

Tehéntej-felvásárlás (millió liter) adatai 1991-ben

8-1. tábla

Hónap	y_t
Január	157
Február	135
Március	151
Április	151
Május	169
Június	167
Július	164
Augusztus	156
Szeptember	146
Október	139
November	129
December	191

$$\bar{y}_t = \frac{1855}{12} = 154{,}583 \approx 155$$

Tehát a havi átlagos tehéntej-felvásárlás Magyarországon 1991-ben 155 millió liter volt.

Ha a hónapok eltérő hosszát a számítás során érvényesíteni szeretnénk, akkor - az idősor természetétől függően - a hónapok napjainak, vagy munkanapjainak számával súlyozott átlagszámítást végezhetnénk el. A súlyozás módosító ereje csekély, gyakorlati jelentősége nem túl nagy, ezért ritkán alkalmazzuk.

Abban az esetben, ha idősorunk időpontokhoz köthető állapotot fejez ki, vagyis **állapotidősort** elemzünk, az úgynevezett **kronologikus átlagot** célszerű használni:

$$\bar{y}_t^{\,k} = \frac{\frac{y_1}{2} + y_2 + \ldots + y_{n-1} + \frac{y_n}{2}}{n-1} \qquad (8\text{-}2)$$

Vegyük példának egy olyan idősort amelynek adatai az egyes negyedévek első napjára vonatkoznak. Könnyű belátni, hogy az I. negyedévhez rendelt adat valójában az elmúlt év november 15-től a tárgyév február 14-ig jellemzi leginkább az idősort, a II. negyedévi adat pedig február 15-től május 15-ig stb. Hogy egy adott évről teljes képet kapjunk a következő év I. negyedévének az adatait is figyelembe kell vennünk, hogy a tárgyév IV. negyedévének második felét is hatékonyan jellemezzük.

A fentiekből következik, hogy egy előre meghatározott időszak korrekt jellemzéséhez egy további, a vizsgált időszakon kívüli, megfigyelés szükséges, de az első és utolsó megfigyelés csak fél súllyal szerepel. Például egy negyedévekkel jellemzett teljes év megfigyeléséhez az következő év első negyedévének adata is szükséges.

Vizsgáljuk meg, hány főt foglalkoztattak hazánkban 1991-ben a kőolaj-feldolgozó iparban! (Az adatok a hónap első napjára vonatkoznak.)

A kőolajiparban foglalkoztatottak (ezer fő) száma 1991-ben

8-2. tábla

Hónap	y_t
Január	5,8
Február	5,8
Március	5,7
Április	5,7
Május	5,7
Június	5,7
Július	5,7
Augusztus	5,7
Szeptember	6,0
Október	6,0
November	6,0
December	5,9
Január	5,9

$$\bar{y}_t^{\,k} = \frac{\frac{5,8}{2} + 5,8 + \ldots + 5,9 + \frac{5,9}{2}}{12} = \frac{69,75}{12} = 5,8125 \approx 5,813$$

Megállapíthatjuk, hogy 1991-ben a kőolaj-feldolgozó iparban az átlagos foglalkoztatott létszám 5.813 fő. Ezen átlagformának is létezik az időszakok hosszával súlyozott változata, de igen ritkán használatos.

Az átlagszámítást felhasználhatjuk az időszakról időszakra, illetve időpontról időpontra történő változások tömör leírására is. Ezek átlagos értéke jellemzi a növekedés vagy csökkenés mértékét, az egész vizsgált időszakban. Ha feltételezhető, hogy a változások az időszakban abszolút nagyságukat tekintve állandóságot mutatnak, akkor az **átlagos abszolút változás** mutatóját célszerű számszerűsítenünk, amely az $y_t - y_{t-1}$ változások egyszerű számtani átlaga:

$$\bar{D} = \frac{\sum_{t=2}^{n} D_t}{n-1} = \frac{(y_2 - y_1) + (y_3 - y_2) + \ldots + (y_n - y_{n-1})}{n-1} = \frac{y_n - y_1}{n-1} \quad (8\text{-}3)$$

ahol $D_t = y_t - y_{t-1}$ a változások abszolút nagysága

Vizsgáljuk meg hazánk művelt mezőgazdasági területének átlagos abszolút változás mutatóját az 1966-1990-es idősoron![5]:

[5]Forrás: Statisztikai Évkönyv

A művelt mezőgazdasági terület (ezer hektár) alakulása 1966-1990.

8-3. tábla

Hó	y_t	D_t
1966	6.928	-
1967	6.913	-15
1968	6.903	-10
1969	6.888	-15
1970	6.875	-13
1971	6.855	-20
1972	6.846	-9
1973	6.835	-11
1974	6.772	-53
1975	6.770	-12
1976	6.757	-13
1977	6.730	-27
1978	6.698	-32
1979	6.651	-47
1980	6.627	-24
1981	6.601	-26
1982	6.582	-19
1983	6.570	-12
1984	6.554	-16
1985	6.540	-14
1986	6.524	-16
1987	6.511	-13
1988	6.497	-14
1989	6.484	-13
1990	6.473	-11

$$\overline{D} = \frac{6.473 - 6.928}{25 - 1} = \frac{-455}{24} = -18,9583 \approx -18,958$$

Elmondhatjuk tehát, hogy hazánkban a mezőgazdaságilag művelt terület 1966-1990 folyamán évente átlagosan 18.958 hektárral csökkent.

Gyakran előfordul, hogy az egymást követő megfigyelések hányadosai mutatnak viszonylagos állandóságot. Ekkor az **átlagos relatív változás** mutatóját célszerű számszerűsíteni, ami nem más, mint az $l_t = \frac{y_{t+1}}{y_t}$ láncviszonyszámok mértani átlaga:

$$\bar{l} = \sqrt[n-1]{\prod_{t=1}^{n-1} l_t} = \sqrt[n-1]{\frac{y_2}{y_1}\frac{y_3}{y_2}\dots\frac{y_n}{y_{n-1}}} = \sqrt[n-1]{\frac{y_n}{y_1}} \qquad (8\text{-}4)$$

Vizsgáljuk meg hazánkban az ismertté vált felnőtt bűnelkövetők 1974-1990-es idősorát!

Az ismertté vált felnőtt bűnözés (ezer fő) 1974-1990.

8-4. tábla

Év	y_t	l_t
1974	63	-
1975	65	1,032
1976	70	1,077
1977	72	1,029
1978	73	1,014
1979	67	0,918
1980	66	0,985
1981	70	1,061
1982	69	0,986
1983	75	1,087
1984	74	0,987
1985	76	1,027
1986	83	1,092
1987	83	1,000
1988	74	0,892
1989	80	1,081
1990	100	1,250

$$\bar{l} = \sqrt[16]{\frac{100}{63}} = \sqrt[16]{1,5873} = 1,0293$$

Az átlagos relatív változó mutatója alapján azt mondhatjuk, hogy a vizsgált időszakban a felnőtt bűnelkövetők száma évről-évre átlagosan 2,9 %-kal nőtt.

8.2 AZ IDŐSOROK ÖSSZETEVŐI

Az idősorok vizsgálatánál a következő hatásokat célszerű elkülöníteni:

Trendhatás vagy alapirányzat

A trendhatás olyan az idősort befolyásoló főbb hatások eredőjeként határozottan jelentkező tendencia, amely a vizsgált időszakban állandóan érvényesül, és stabilitást mutat. Ha az adott időintervallumon becsült tendenciát extrapolálással ki akarjuk terjeszteni a vizsgált intervallum határain kívülre, ezt csak azzal a feltételezéssel tehetjük, hogy ott is érvényesül ez a stabilitás.

Periodikus ingadozás

A szisztematikusan jelentkező hullámmozgással modellezhető hatásokat nevezzük így. Az **állandó periódushosszú** hullámzást mutató idősorok esetében a periódushossz szinte kizárólagosan egy év, vagy ennél rövidebb, amely eset ha fennáll, a jelenséget **szezonalitásnak** nevezzük. Az elnevezés hagyományosan a negyedéves és havi bontásos idősorokra vonatkozik, de használhatjuk a héten belüli napi bontású, és a napon belüli órabontású idősorokra is. Ezen hullámmozgások összetetten is jelentkezhetnek, pl. a külkereskedelmi forgalom, a közlekedés, az elektromos energiafogyasztás stb. idősorain.

A **változó periódushosszú** ingadozások köré sorolható a **konjunkturális ingadozás**, bár utóbbi részhalmaza előzőnek, szinonimaként is kezelhető, hiszen a változó periódushosszú ingadozások vizsgálati módszerei döntően a konjunktúrahatás területén születtek. E terület elemzési lehetőségeiről a prognosztika elnevezésű diszciplína ad tájékoztatást.

Véletlen ingadozás

A véletlen ingadozás alatt sok kisebb jelentőségű, egyedi, esetenként befolyásoló tényező együttes hatásának eredőjét értjük, amiről részletes elemzést tartalmaz jegyzetünk 7.4 alfejezete.

Strukturális törés

Strukturális töréseknek nevezzük az olyan egyszeri, jelentős tendenciaváltozásokat, melyek oly számottevően befolyásolják az adott időszak alakulását, hogy külön vizsgálatot igényelnek. Strukturális törés megléte esetén fontos cél a létrehozó ok, vagy okok feltárása, a hatás, vagy hatások tovagyűrűzésének, esetleges "elhalásának" elemzése. Ha a strukturális törések száma és jelentősége nagy, akkor a tárgyalt determinisztikus idősori megközelítési mód hatékonysága megkérdőjelezhető.

Ha a szezonális hullámmozgás kitérései, amplitúdói abszolút értelemben, vagy relatív (a trendhez viszonyítva) értelemben állandóságot mutatnak, akkor **állandó szezonalitásról**, ellenkező esetben **változó szezonalitásról** beszélünk.

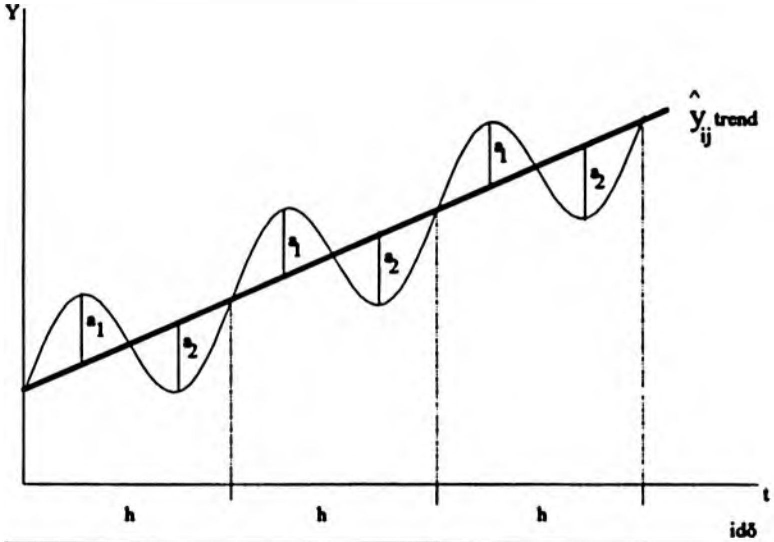
Az állandó szezonalitás modellezésének két alapformája ismeretes.

Az **additív modell**:

$$y_{ij} = \hat{y}_{ij} + s_j^* + v_{ij}^* \tag{8-5}$$

ahol: y_{ij} = a megfigyelt idősor értéke
$\hat{y}_{ij}$ = a trendérték
s_j^* = a szezonális eltérés
v_{ij}^* = a véletlen hatás
$i = 1,2,\ldots,n$ = a periódusok (pl.évek) száma
$j = 1,2,\ldots,m$ = a perióduson belüli időszakok, azaz a szezonok (pl.hónapok,negyedévek) száma

Az additív modellt az alábbi ábra szemlélteti:



8-1. ábra Az additív modell

ahol: h = a hullámhossz

a_1 = a periódusonkénti legnagyobb értéknél mért amplitúdó

a_2 = a periódusonkénti legkisebb értéknél mért amplitúdó

Ha az idősor értékei a 8-1. ábrán vázolt hullám-formával jól jellemezhetők, akkor **additív** modellről beszélhetünk. Mivel a szezonális eltérések az adott időpontokhoz tartozó amplitúdók átlagértékének megfeleltethetők, könnyű belátni, hogy $\sum_{j=1}^{m} s_j^* = 0$.

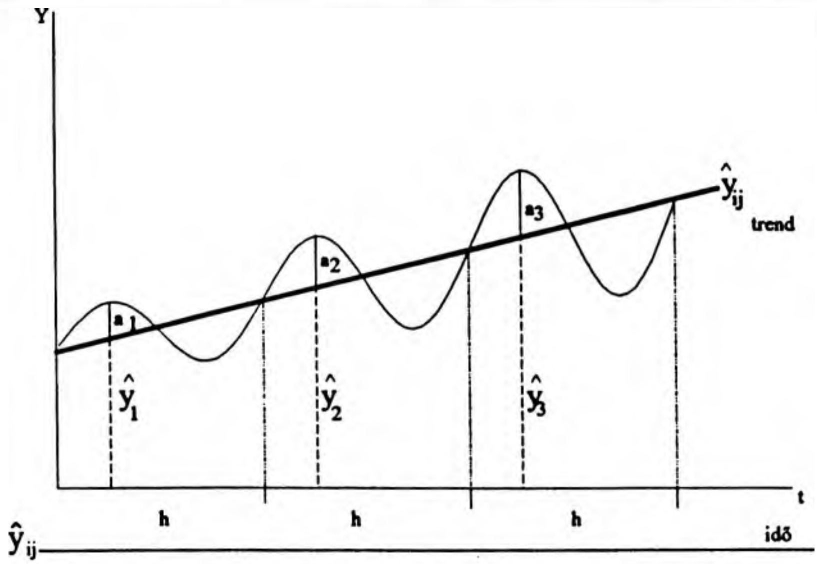
A multiplikatív modell:

$$y_{ij} = \hat{y}_{ij} \times s_j \times v_{ij} \qquad (8\text{-}6)$$

ahol: a már ismert jelölések mellett

s_j - a j-edik szezonhoz tartozó szezonális komponens, a szezonindex

A multiplikatív modellt az alábbi módon szemléltethetjük:



8-2 ábra A multiplikatív modell

Akkor beszélünk multiplikatív modellről, ha $\frac{a_1}{\hat{y}_1} = \frac{a_2}{\hat{y}_2} = \frac{a_3}{\hat{y}_3}$ igaz, és a megfelelő szezonoknál jelentkező hasonló hányadosok szintén állandóságot mutatnak.

Multiplikatív modell esetén a szezonindexre vonatkozó követelmény: $\frac{\sum_{j=1}^{m} s_j}{m} =$.

Összefoglalóan megállapítható, hogy a szezonalitás eltérítő hatása a megfelelő szezonoknál additív modellben **abszolút** állandóságot, multiplikatív modellben a trendhez mért **relatív** állandóságot mutat.

A két modell közötti választásnál segíthet:

- a vizsgált jelenséggel kapcsolatos szakmai ismeretanyag és
- a grafikus ábrázolás.

8.3 TRENDELEMZÉS

Az alaptendencia számszerűsítésének két alapvető módszere terjedt el.

1) Az ún. **mozgóátlagolásos eljárás** a számtani átlagszámítás speciális idősori alkalmazása.

2) Az **analitikus trendszámítás** analitikus függvénnyel írja le a vizsgált jelenség tartós tendenciáját, a függvény paramétereit a tapasztalati idősorból becsüli. Az analitikus trendszámítás gyakorta alkalmazza a regressziószámítás során már megismert klasszikus legkisebb négyzetek módszerét a paraméterek becslésére. A regressziószámítás során elkülönítettük az ún. alap-regresszió,- és a minta-regresszió függvény fogalmát, amelyeknek a paramétereit az előbbi sorrendben β_j, illetve b_j-vel jelöltük. Bár az idősorelemzésben is megkülönböztettük az elméleti,- és a tapasztalati idősor fogalmát, ezt a megkülönböztetést a jelölésben nem hangsúlyozzuk. Mindez magyarázható az X tényezőváltozó és a t időváltozó, mint speciális tényezőváltozó közötti eltérésekkel, amiket az alábbiakban foglaltunk össze:

X tényezőváltozó	**t időváltozó**
- a minta általában bővíthető;	- a vizsgált időtáv ritkán bővíthető;
- jellemző a több minta kiválasztásának a lehetősége;	- csak egy realizáció, egy minta képzelhető el;
- a mintába került konkrét x értékek elsődlegesen, mint egy valószínűségi változó értékei jelenek meg;	- az idő hatását természetes, egész számokkal jellemzett, rögzített sorrendű t változó fejezi ki;

A fentiek miatt a paraméterbecslést nem hibahatárral jellemezhető konfidencia intervallum becslésként, hanem pontbecslésként értelmezzük, annak ellenére, hogy a hibahatár képzésének technikai feltételei adottak. A becsléses jelleg tehát, csak a teoretikus, és nem a praktikus megközelítésből adódik.

8.3.1 Mozgóátlagolásos trendelemzés

A mozgóátlagolásos technika alkalmazásánál vesszük az idősor első k számú értékét, majd ezeket összegezzük. Az összeget k értékével osztjuk, és az így kapott számértéket az összeget szolgáltató időszak középső elemének megfeleltetjük. A következő lépés az előzőek olyan módon történő ismétlése, hogy az első összegzendő tagot elhagyjuk, és helyette az idősor következő megfigyelését vesszük, és újra elvégezzük az átlagolást. Ezt a műveletet addig végezzük, míg az utolsó idősori értéket is felhasználjuk.

Ha k páros, akkor az összeget szolgáltató időszak közepe két átlagolt érték közé esik. Ez esetben egy ismételt k=2-es mozgóátlagolást, más néven centrírozást végzünk, így a kétszeri eltolódás miatt az y_t és $\hat{y}_t$ értékek már egymásnak megfelelhetők. Az átlagolt értékek jelentik az idősornak az alapirányzat szerint várható értékeit, röviden az $\hat{y}_t$ trendértékeket.

A következő számpélda a mozgóátlagolásos eljárást mutatja be. Állítsuk elő k=4-es mozgóátlagolásos módszerrel egy nagyáruház 1984-1991-es összforgalmának trendértékeit!

Egy nagyáruház 1984-1991-es összforgalmának adatai (ezer Ft) negyedéves bontásban

8-5. tábla

Év	I.	II.	III.	IV.
	negyedév			
1984	96.323	110.081	119.029	156.040
1985	119.019	130.557	138.625	172.903
1986	136.559	139.963	160.600	210.812
1987	158.449	158.870	182.098	227.863
1988	176.904	190.627	188.501	235.614
1989	176.757	190.243	189.941	250.006
1990	202.205	191.694	198.541	261.029
1991	247.441	226.908	261.689	307.951

Részlet a mozgóátlagolás munkatáblájából

8-6.tábla

Sorszám	Megfigyelés	4 tagú átlag	Centrírozott érték
1.	96.323	-	-
2.	110.081	120.368	-
3.	119.029	128.542	124.455
4.	156.040	133.661	131.101
5.	129.019	.	.
6.	130.557	.	.
.	.	.	.
.	.	.	.
.	.	.	.

Első 4 tagú átlag = (96.323+110.081+119.029+156.040)/4 = 120.368

Első centrírozott érték = (120.368+128.542)/2 = 124.455

A műveletsor végeredményét mutatja a következő tábla:

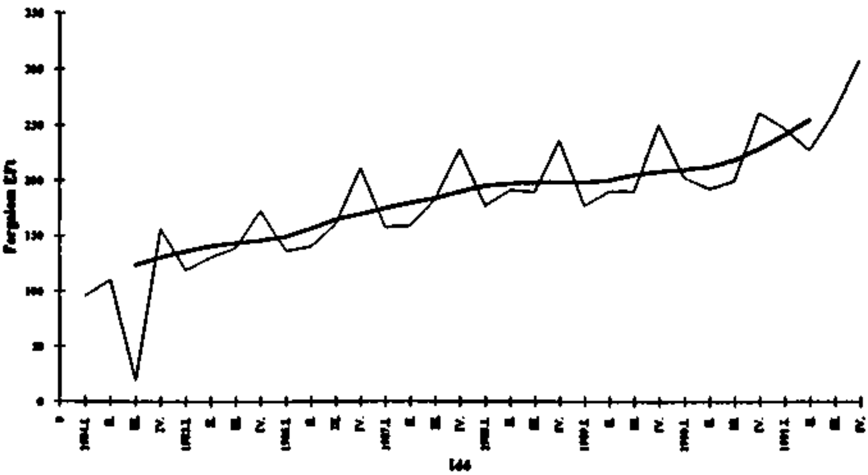
A mozgóátlagolásos trendértékek egy nagyáruház összforgalmi adataira

8-7.tábla

Év	I.	II.	III.	IV.
	negyedév			
1984	-	-	124.455	131.102
1985	136.111	140.668	143.723	145.847
1986	149.769	157.255	164.725	169.819
1987	174.870	179.689	184.127	190.403
1988	195.173	196.943	197.893	197.827
1989	197.959	199.938	204.918	208.280
1990	209.536	211.989	219.022	229.078
1991	241.373	255.132	-	-

Külön megfontolást kíván a mozgóátlagolás k tagszámának meghatározása. Minél nagyobb k értéke, annál biztosabban felszínre kerül a feltárni kívánt alaptendencia, annál eredményesebb a véletlen hatás kiszűrése. Korlátozó tényező viszont a számpéldában is észlelhető rövidülés.

A szezonalitást is tartalmazó idősoroknál a k megválasztásánál még egy szempontot figyelembe kell venni. Nevezetesen azt, hogy a k tagszám megegyezzen a szezonok számával, vagy az m-nek egész számú többszöröse kell, hogy legyen, a szezonhatás kiszűrése érdekében. Ellenkező esetben a gyenge, illetve az erős idények rendre nagyobb súlyt kapnak az összegképzésben, így egy zavaró hullámzást visznek be az $\hat{y}_t$ idősorába. Fentiek miatt a negyedéves idősorok lehetséges k tagszáma 4, 8, 12,..., a havi bontású idősorok esetén pedig 12, illetve 24,...



8-3. ábra Egy áruház forgalmi adatai 1984-1991 között és a mozgóátlagolású trend adatok

A bemutatott mozgóátlagolásos módszer súlyozatlannak tekinthető alapmódszer, hiszen az y_t idősori értékek az átlagolás során egyforma jelentőséggel szerepelnek. Megjegyzendő azonban, hogy a páros k-érték esetén alkalmazott centrírozás egyfajta súlyozásként is felfogható.

Ismeretesek súlyozott mozgóátlagolású módszerek is. A leginkább elterjedt módszer a **Spencer-féle, 15 és 21 tagú súlyozott mozgóátlagolás**. A 15 tagú mozgóátlagolás súlyai:

Tagszám[6]:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Súlyok:	-3	-6	-5	3	21	46	67	74	67	46	21	3	-5	-6	-3.

Látható, hogy a súlyrendszer a középső tagra szimmetrikus. Ugyanez a 21 pontos mozgóátlagolásnál is fennáll, ezért elegendő a súlyrendszert a középső tagig közölni:

[6]Tagszámon az egyes tagok sorszámát értjük.

Tagszám:	1	2	3	4	5	6	7	8	9	10	11	...
Súlyok:	-1	-3	-5	-5	-2	6	18	33	47	57	60	...

A Spencer-féle 15 pontos súlyozott mozgóátlagolás úgy is előállítható, hogy egymás után alkalmazunk két 4 tagú súlyozatlan, majd egy 5 tagú súlyozott mozgóátlagolást, ahol a súlyok rendre: -3, 3, 4, 3, -3. A Spencer-féle 21 pontos súlyozott mozgóátlagolás hasonló módon állítható elő. A bemutatott módszer alkalmazásra kerül a változó szezonalitás elemzésénél (lásd! 8.4.2 alfejezet).

8.3.2 Analitikus trendszámítás

Az analitikus trendszámítás esetében a vizsgált jelenség megfigyelt értékei y_t, és az idő hatását kifejező t, természetes egész számokból álló változó közötti kapcsolatot modellezzük. A lehetséges függvénykapcsolatok egy részével a 7. fejezetben is találkoztunk.

Az alkalmazható függvénytípusokból néhány:

1. Lineáris trendfüggvény

$$\hat{y}_t = b_0 + b_1 t \qquad (8\text{-}7)$$

2. Exponenciális trendfüggvény

$$\hat{y}_t = b_0 b_1^{\,t} \qquad (8\text{-}8)$$

3. Másodfokú polinom

$$\hat{y}_t = b_0 + b_1 t + b_2 t^2 \qquad (8\text{-}9)$$

4. Hatványkitevős függvény

$$\hat{y}_t = b_0 t^{b_1} \tag{8-10}$$

5. Hiperbolikus trendek

$$\hat{y}_t = b_0 + b_1 \frac{1}{t} \tag{8-11}$$

$$\hat{y}_t = b_0 + b_1 \frac{1}{t} + b_2 \frac{1}{t^2} \tag{8-12}$$

$$\hat{y}_t = \frac{1}{b_0 + b_1 t} \tag{8-13}$$

$$\hat{y}_t = \frac{1}{b_0 + b_1 t + b_2 t^2} \tag{8-14}$$

6. Törnquist függvény, telítődési görbe

$$\hat{y}_t = \frac{Kt}{t + b} \tag{8-15}$$

7. Logisztikus S formájú telítődési görbe

$$\hat{y}_t = \frac{K}{1 + e^{b_0 + b_1 t}} \tag{8-16}$$

A felsorolt függvénytípusok 1-től 6-ig logaritmikus transzformációkkal, illetve új változók bevezetésével a 7. fejezetben megismert módon linearizálhatók, és klasszikus legkisebb négyzetek módszerével becsülhetők. Nézzük a fenti függvénytípusok alkalmazásának néhány indokát!

ad 1. Akkor alkalmazzuk a lineáris formulát, ha feltételezhető, hogy egységnyi időváltozás hatására, a vizsgált folyamat változása, növekedése vagy csökkenése, az elemzett időtávon abszolút értelemben megközelítően állandó.

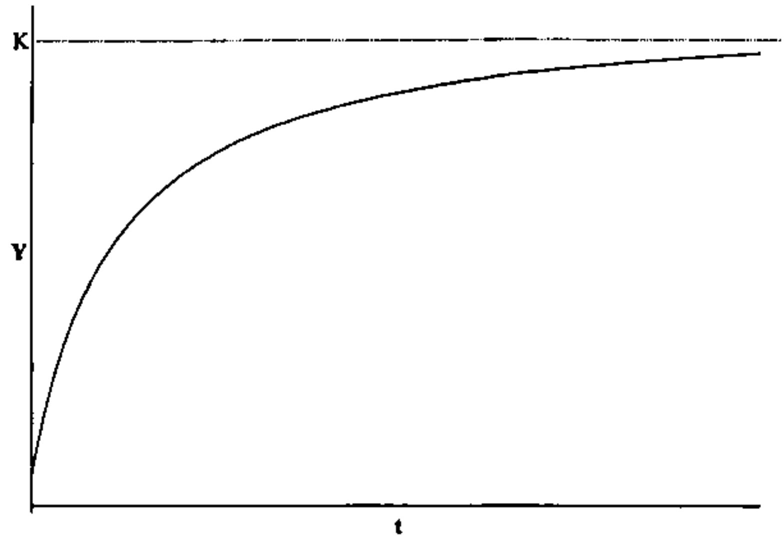
ad 2. Általában a közép- és hosszú távú gazdasági és társadalmi folyamatok jellemzésének alapmodellje. Akkor alkalmazzuk, ha feltételezhető, hogy egységnyi időváltozás hatására a folyamat változása relatíve állandó, azaz a vizsgált időszakban a megfigyelések az előző értékhez képest rendre megközelítően azonos százalékos növekedést vagy csökkenést mutatnak.

ad 3. Amennyiben az idősorban tendencia-váltás tapasztalható, vagyis növekedésből csökkenésbe (vagy fordítva) megy át, akár ismétlődően is, az idősor gyakran jól modellezhető p-ed fokú polinommal. Értelemszerű, hogy a fokszám növelése egyre jobb illesztést ad, de megállapítható, hogy p>3 fokszám alkalmazása már igen nehezen indokolható.

ad. 4. A regressziószámításban általánosan használt hatványkitevős formula alkalmazása esetén az időváltozó sajátos értelmezést nyerne. Ebben az esetben feladnánk a t természetes egész számokkal jellemzett időváltozó ekvidisztanciáját, vagyis az időt az ún. gyorsuló idő kategóriájával jellemeznénk. Mivel ez a közelítési mód - bár eredményei értelmezhetők - a statisztikai elemzési gyakorlatban nem terjedt el, így e formula csak kiegészítő vizsgálatra javasolható.

ad. 5. Gyakran előfordul, hogy az idősor aszimptotikusan közelít egy értéket. Ekkor trendfüggvényként valamelyik felsorolt hiperbolikus függvény alkalmazható. Az önköltséget, az árak alakulását jellemző folyamatok gyakran modellezhetők e módon.

ad.6. A formula kevéssé ismert[7], így közöljük ábráját:



8-4. ábra A Törnquist telítődési görbe

Azon folyamatok jellemezhetők így, ahol az idősor egy felső határértéket gyorsan induló, de lassúló mértékű növekedéssel közelít meg. Jó példa erre egy piacon bevezetett termék magasabb használati értékűvel történő kiváltása, pl. a mechanikus karórák kiváltása kvarcórával. A megszabott szolgáltatást helyettesíti egy jobb és alacsonyabb árú termék, ezért a felfutás a látványos kezdeti áresés hatására gyors.

[7]A Törnquist modellnek csak egy eleméről van itt szó.

A linearizálás menete:

$$\hat{y} = \frac{Kt}{t+b} \quad \Rightarrow \quad \frac{1}{\hat{y}} = \frac{1}{K} + \frac{b}{Kt}$$

ahol: K - a konstans telítődési szint
b- paraméter

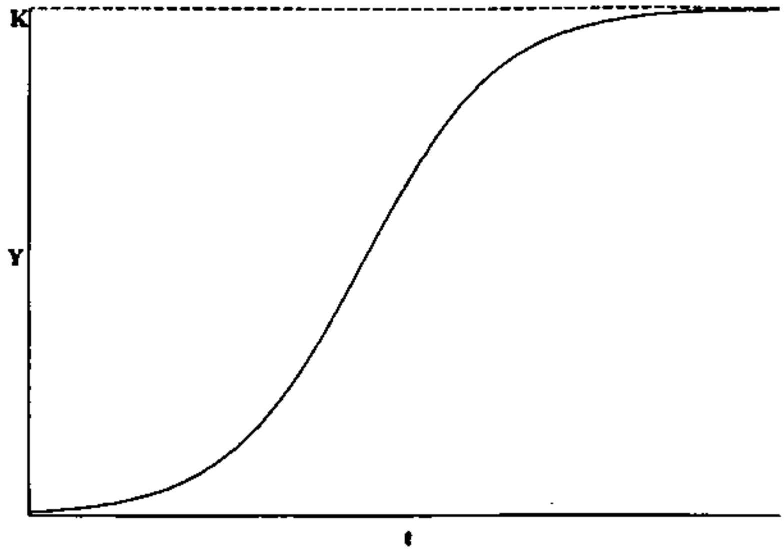
Legyen:

$$\frac{1}{K} = C, \quad \frac{b}{K} = B, \quad \frac{1}{\hat{y}} = \hat{Y}, \quad \frac{1}{t} = T.$$

vagyis a linearizált formula

$$\hat{Y} = C + BT$$

ad.7. Az S alakú telítődési görbékre tipikus példa a tartós fogyasztási cikkek termelése és forgalmazása. A lassú felfutás a termék szolgáltatásának megismerése, a kezdeti magas ár következménye. A középrétegek számára elérhető ár belépése, a fokozatos megismerés a kereslet felfutását eredményezi, majd bekövetkezik a telítődés, és a forgalom stabilizálódik az átlagos élettartam és a potenciális fogyasztói létszám függvényében. A szakirodalom számos formulát ismer leírására, becslése általában iterációs eljárásokkal történik.



8-5. ábra Az S alakú telítődési görbe

Az alkalmazandó trendtípus kiválasztásánál a következő módszereket, illetve szempontrendszert tartjuk szem előtt:

- Szakmai, logikai elemzés

  A vizsgált jelenség szakmai ismeretének birtokában dönthetünk a trendfüggvényként alkalmazásra kerülő függvénytípusról.

- Grafikus elemzés:

  Alátámaszthatja, vagy kiegészítheti a szakmai elemzést a grafikus ábrázolás, amely felszínre hozhatja az idősorban megmutatkozó változási tendenciát.

- Analitikus elemzés:

  A különböző szakmai és logikai elemzés alapján választható függvények illeszkedését az

$$s_e^2 = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n} = \frac{\sum_{t=1}^{n} e_t^2}{n} \tag{8-17}$$

reziduális szórásnégyzet mutatóval tudjuk mérni. Nyilvánvaló, hogy a mutató értéke a legjobban illeszkedő függvény esetén a legkisebb. A különböző fokszámú polinomiális illesztésnél, illetve más függvénytípusokkal történő összehasonlításnál, a szabadságfokkal korrigált reziduális szórást kell használnunk, a megfelelő összehasonlítás érdekében. A korrigált reziduális szórásnégyzet formulája:

$$s_e^2 = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n-p-1} = \frac{\sum_{t=1}^{n} e_t^2}{n-p-1} \tag{8-18}$$

ahol: p - a függvény paramétereinek a száma.

A továbbiakban részletesen csak a lineáris és az exponenciális trendszámítást ismertetjük.

A lineáris trend becslése

A cél az

$$\hat{y}_t = b_0 + b_1 t$$

függvény paramétereinek becslése, amelyet a 7. fejezetben megismert legkisebb négyzetek módszerével végezzük.

A minimalizálási feladat:

$$\sum_{t=1}^{n}\left(y_t-\hat{y}_t\right)^2 \Rightarrow \text{minimum} \tag{8-19}$$

amelynek megoldásával az ún. normálegyenleteket nyerjük:

$$\begin{aligned}\sum y_t &= b_0 n + b_1\sum t\\ \sum t y_t &= b_0\sum t + b_1\sum t^2\end{aligned} \tag{8-20}$$

A két egyenletből megkapjuk a b_0 és b_1 paramétereket. Ha a t változó alkalmas felvételével pl. t = 1, 2, 3, ... 9 helyett, -4, -3, -2, -1, 0, 1, 2, 3 4 értékekkel jellemezzük az idő múlását és így biztosítjuk a $\sum_{t=1}^{n} t = 0$ feltételt, akkor a számítások leegyszerűsödnek.

$$b_0=\frac{\sum_{t=1}^{n} y_t}{n} \tag{8-21}$$

$$b_1=\frac{\sum_{t=1}^{n} t y_t}{\sum_{t=1}^{n} t^2} \tag{8-22}$$

Látható, hogy b_0 ez esetben a megfigyelt idősori értékek számtani átlaga.

Megfelelő számítógépes háttér birtokában nincs gyakorlati jelentősége a fenti transzformációs megoldásnak.

A b_o paraméter a t = 0 időpontban modellezett alapirányzati idősor értéket mutatja. A b_1 paraméter az időpontról - időpontra, illetve időszakról-időszakra mért **átlagos abszolút változást** számszerűsíti.

A lineáris trend becslését a hazánkra vonatkozó művelt mezőgazdasági terület 1966-1990-es idősorán mutatjuk be.

A művelt mezőgazdasági terület (ezer hektár) 1966-1990.

8-8. tábla

Év	t	y_t	$\hat{y}_t$
1966	1	6.928	6.948
1967	2	6.928	6.927
1968	3	6.903	6.906
1969	4	6.888	6.885
1970	5	6.875	6.864
1971	6	6.855	6.843
1972	7	6.846	6.822
1973	8	6.835	6.801
1974	9	6.782	6.780
1975	10	6.770	6.759
1976	11	6.757	6.738
1977	12	6.730	6.717
1978	13	6.698	6.696
1979	14	6.651	6.674
1980	15	6.627	6.654
1981	16	6.601	6.633
1982	17	6.582	6.612
1983	18	6.570	6.590
1984	19	6.654	6.569
1985	20	6.540	6.548
1986	21	6.524	6.527
1987	22	6.511	6.506
1988	23	6.497	6.485
1989	24	6.497	6.485
1990	25	6.473	6.443

A normálegyenletek megoldásához szükséges számítási eredmények:

$$n = 25, \quad \sum t = 325, \quad \sum t^2 = 5.525$$

$$\sum y_t = 167.394 \quad \sum ty_t = 214.8754$$

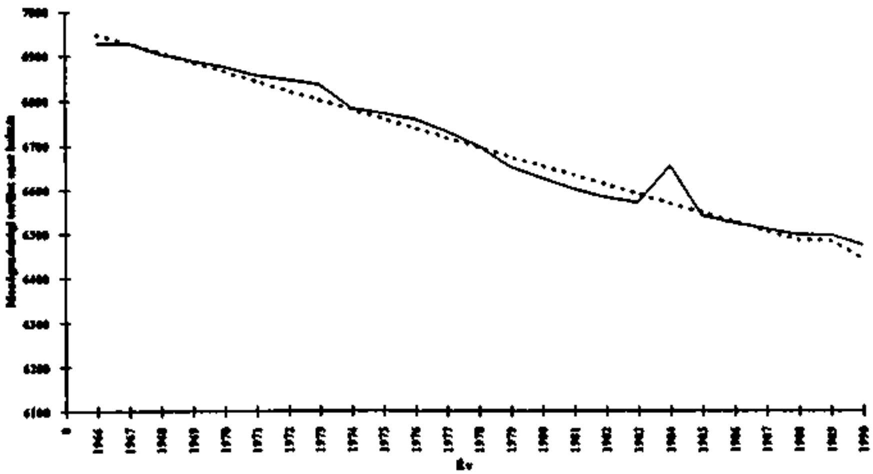
$$167.394 = b_0 25 + b_1 325$$
$$2.148.754 = b_0 325 + b_1 5.525$$

A becsült trendegyenlet:

$$\hat{y}_t = 6.969,44 - 21,05t$$

A b_0 paraméter szerint a fenti lineáris trendfüggvény a magyar mezőgazdasági művelt termőterületre 1965-ben 6.969 ezer hektárt becsül.

A b_1 paraméter értelmezése: A trendbecslés alapján a magyar mezőgazdasági művelt termőterület 1966-1990 között, évről-évre átlagosan 21.050 hektárral csökkent.



8-6. ábra A művelt mezőgazdasági terület 1966-1990. között és az illesztett lineáris trend

Az illeszkedés jóságának relatív mérésére szolgál a reziduális szórás mutatója. Az elemzéseknél felhasznált szoftverek általában szolgáltatják a regressziószámításnál megismert r lineáris korrelációs együttható értékét is, amely tulajdonképpen a korrelációs kapcsolatot méri (példánkban r = 0,9926) azonban jól tudjuk, hogy az idősor sajátossága miatt ez csak egy technikai jellegű mérőszám.

Az exponenciális trend becslése

Ebben az esetben célunk az

$$\hat{y}_t = b_0 b_1^t$$

függvény paraméterének becslése.

Az egyenletet logaritmizálva[8] lineáris alakra hozhatjuk:

$$\ln \hat{y}_t = \ln b_0 + t \ln b_1$$

legyen $\ln b_0 = B_0$ és $\ln b_1 = B_1$

így

$$\ln \hat{y}_t = B_0 + B_1 t \tag{8-23}$$

A formula már becsülhető a megismert legkisebb négyzetek módszerével. A normálegyenletek a következőképpen módosulnak:

$$\begin{aligned} \sum \ln y_t &= B_0 n + B_1 \sum t \\ \sum t \ln y_t &= B_0 \sum t + B_1 \sum t^2 \end{aligned} \tag{8-24}$$

A lineáris trend becslésénél bemutatott módon $\sum t = 0$ feltétel biztosítása esetén a becslő formulák a következők:

$$B_0 = \frac{\sum_{t=1}^{n} \ln y_t}{n} \quad \Rightarrow \quad e^{B_0} = b_0 \tag{8-25}$$

$$B_1 = \frac{\sum_{t=1}^{n} t \ln y_t}{\sum_{t=1}^{n} t^2} \quad \Rightarrow \quad e^{B_1} = b_1 \tag{8-26}$$

[8]Természetesen bármilyen alapú logaritmus használható.

A b_0 paraméter a $t = 0$ időpontban modellezett alapirányzati idősor értéket mutatja. A b_1 paraméter az időpontról időpontra mért **átlagos relatív változást** számszerűsíti.

Az exponenciális trend becslését az ismertté vált felnőtt bűnelkövetők magyarországi 1974-1990-es idősorán mutatjuk be:

Az ismertté vált felnőtt bűnelkövetők száma (ezer fő) 1974-1990.

8-9. tábla

Év	t	y_t	$\hat{y}_t$
1974	1	63	64
1975	2	65	65
1976	3	70	66
1977	4	72	67
1978	5	73	69
1979	6	67	70
1980	7	66	71
1981	8	70	72
1982	9	69	74
1983	10	75	75
1984	11	74	76
1985	12	76	78
1986	13	83	79
1987	14	89	81
1988	15	74	82
1989	16	80	84
1990	17	100	85

$$n = 17, \quad \sum t = 153, \quad \sum t^2 = 1.785,$$

$$\sum \ln y_t = 73{,}09, \quad \sum t \ln y_t = 665$$

$$73{,}09 = 17B_0 + 153B_1$$
$$665 = 153B_0 + 1.785B_1$$

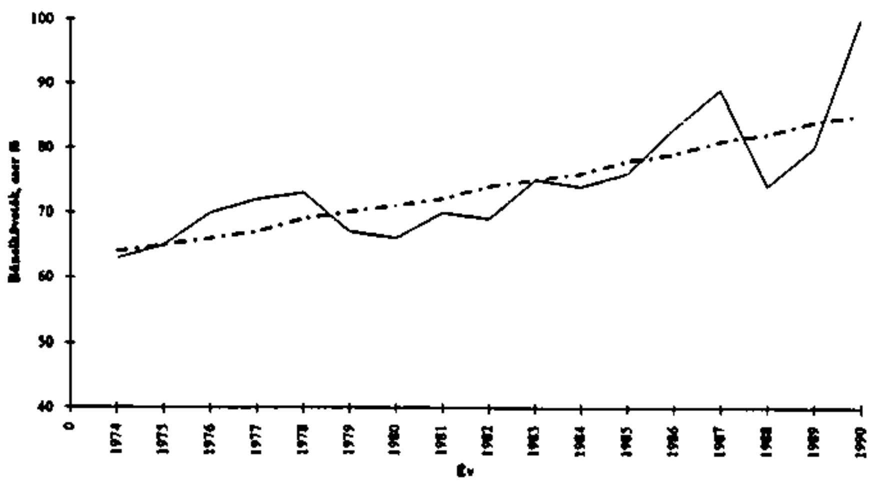
$$B_0 = 4{,}1366 \qquad b_0 = 62{,}59$$
$$B_1 = 0{,}0180968 \qquad b_1 = 1{,}018$$

Az exponenciális trendfüggvény:

$$\hat{y}_t = 62{,}59 \times 1{,}018^t$$

A b_o paraméter értelmezése: Az 1973-ra becsült alapirányzat érték, azaz az ismertté vált felnőtt bűnelkövetők becsült száma 62.590 fő.

A b_1 paraméter értelmezése: A trendbecslés alapján a felnőtt bűnelkövetők száma évről-évre átlagosan 1,018 szeresére, 101,8 %-ra nő, azaz átlagosan 11,8 %-kal növekszik.



8-7. ábra Az ismertté vált felnőtt bűnelkövetők (ezer fő) száma 1974-1990. között és az illesztett exponenciális trend

A 8-7. ábra alapján nem egyértelmű hogy az exponenciális illesztés hatékonyabb-e. Célszerű ilyen esetben a vizsgálatot kiegészíteni a lineáris illesztéssel, és az eredményeket összehasonlítva lehet állást foglalni abban, hogy melyik függvényt használjuk.

Az exponenciális és a lineáris illesztés összehasonlítása

8-10. tábla

Év	Exponenciális		Lineáris		
	e_t	e_t^2	$\hat{y}_t$	e_t	e_t^2
1974	-1	1	63	0	0
1975	0	0	64	1	1
1976	4	16	66	4	16
1977	5	25	67	5	25
1978	4	16	69	4	16
1979	-3	9	70	-3	9
1980	-5	25	72	-6	36
1981	-2	4	73	-3	9
1982	-5	25	75	-6	36
1983	0	0	76	-1	1
1984	-2	4	77	-3	9
1985	-2	4	79	-3	9
1986	4	16	80	3	9
1987	8	64	82	7	49
1988	-8	64	83	-9	81
1989	-4	16	85	-5	25
1990	15	225	86	14	196
$\sum$	----	514	----	----	527

A (8-17) mutatóval végzett összehasonlítás az alábbi eredményt adja:

$$s^2_{e(exp)} = \frac{514}{17} = 30,2$$

$$s^2_{e(lin)} = \frac{527}{17} = 31,0$$

$$s^2_{e(exp)} < s^2_{e(lin)}$$

A fenti eredmények az exponenciális illesztés realitását támasztják alá. Igaz, hogy a különbség csak igen kis mértékű, azonban a további szakmai megfontolások is az exponenciális függvény mellett szólnak, ugyanis az idősor növekedése az utolsó időszakban számottevően meghaladja az előzőket.

8.4 A SZEZONALITÁS ELEMZÉSE

A gazdasági-társadalmi folyamatok alakulásában gyakran tapasztalunk szabályos ingadozásokat. Havi és negyedéves bontású idősoroknál tapasztalható, hogy egyes negyedévekben, illetve hónapokban az alaptendenciához viszonyított szisztematikus visszaesés más negyedévekben, illetve hónapokban rendszeres kiugrás tapasztalható.

A szezonalitás létrejöttében meghatározó szerepe van bizonyos **természeti jelenségeknek**. Ilyen a föld forgása, meghatározott körforgása a Nap körül, amely a napi, havi, évszakonkénti szezonalitásnak fő okozója. Alapvetően éghajlati tényezőkön keresztül fejti ki hatását, elsősorban a hőmérséklet és a megvilágitottság megváltozása révén.

A fentiek összefüggésében alakultak ki olyan **társadalmi gazdasági** életben megfigyelt **jelenségek**, amelyek szintén szezonalitást okoznak. A társadalmi szokások, hagyományok, az ünnepek, a divat jelentős hullámzásokat hoz létre a kereskedelemben, a közlekedésben, a postaforgalomban, a hírközlésben, az áramfogyasztásban stb.

Nyilvánvaló, hogy a szezonalitás mind a munkaerő, mind a lekötött eszközök kapacitás-kihasználtságának vonatkozásában káros jelenség. Az okozott károk tompítására kétféle stratégia alkalmazása képzelhető el.

A csillapítás stratégiája

A szezonális hullámzás kiküszöbölésére, illetve csillapítására jelentős törekvések irányulnak. Sok helyen erőfeszítéseket tesznek az "év végi hajrá" kiküszöbölésére. A mezőgazdasági termelés az állattenyésztési ágazatban látványosan függetlenítette magát az évszakokhoz kötődő korábbi ciklikusságtól, pl. a hús és a tojástermelés területén. Az építőipar a paneltechnológiás ágazatban nagymértékben függetlenné vált az időjárási viszonyoktól. A kereskedelem és a vendéglátóipar is megpróbált tompítani a tevékenysége területén jelentkező ingadozáson (akciók holtidényben, árkedvezmény a gyenge szezonokban stb).

Az alkalmazkodás stratégiája

Sok esetben a szezonalitás megváltozhatatlan adottságként kezelhető. Ilyenkor az erőforrásoknak kell alkalmazkodni hozzá. Jó példát szolgáltat erre a vendéglátóipar. Összkomfortos szállodák építése helyett kifizetődőbb a csak főidényben, nálunk elsősorban nyáron üzemeltetett, fűtés nélküli szállodák, faházak fenntartása azokon a területeken, ahol az egész évi kihasználtságnak jelentős korlátai vannak.

8.4.1 Az állandó szezonalitás elemzése

Ha a szezonalitás abszolút, vagy relatív állandósága (lásd! additív, illetve multiplikatív modell) fennáll, akkor alkalmazhatjuk az állandó szezonalitás elemzésére kimunkált módszertani apparátust.

A szezontényező meghatározásának klasszikus módszere

A bemutatásra kerülő, a gyakorlatban leginkább elterjedt **szezonindex-számítási** módszer feltételezi az előzőekben ismertetett trendszámítási módszerek valamelyikével a trendértékek előzetes számszerűsítését.

A módszer lényege:

- a trendhatás leválasztása,
- a véletlen hatás kiszűrése.

A **multiplikatív modell** esetében

$$y_{ij} = \hat{y}_{ij} \times s_j \times v_{ij}$$

a **szezonindexek** számszerűsítése a következőképpen történik:

$$s_j = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{ij}}{\hat{y}_{ij}} \qquad (8\text{-}27)$$

A fenti szezonindexeket ún. **nyers szezonindexeknek** nevezzük amelyek tulajdonképpen az **alapirányzattól megtisztított** $\left(\frac{y_{ij}}{\hat{y}_{ij}}\right)$ **értékek** szezononkénti egyszerű számtani átlagai[9].

[9]Amennyiben a trendértékeket mozgóátlagolás segítségével határoztuk meg, vagy bármely más okból az idősor értékeinek sorozata hiányos, akkor előfordulhat, hogy nem rendelkezünk minden megfigyelésre vonatkozóan, becsült, illetve tényleges idősori értékkel. Ekkor értelemszerűen annyi taggal osztunk, ahány alapirányzattól tisztított érték a rendelkezésünkre áll, ez esetben az osztó tag kisebb n-nél.

Könnyen belátható, hogy s_j értékek átlagának nagysága együtthatós formában 1, százalékos formában 100 %. Az idősor rövidsége, és az alapirányzat számszerűsítésénél adódó kisebb pontatlanság miatt előfordulhat, hogy az átlag nem pontosan adja a szezonalitás definíciójának megfelelő elméleti értéket. Ekkor az s_j értékeket saját átlagukkal

$$\bar{s} = \frac{\sum_{j=1}^{m} s_j}{m} \qquad (8\text{-}28)$$

rendre elosztva, az elméleti elvárásnak megfelelő értékeket kapjuk. Amennyiben ez egy alkalmas kerekítéssel elérhető, a korrekciótól eltekinthetünk. Az elméleti elvárásnak megfelelő értékeket **tisztított szezonindexeknek** nevezzük.

Vizsgáljuk meg, hogyan érvényesül a szezonhatás a nagyáruház 1984-1991-es negyedéves bontású összforgalmi idősorában! Az alapadatok 8-5. táblában, a mozgóátlagolásos trendértékek a 8-7. táblában találhatók. A 8-11. táblában az alapirányzattól tisztított értékeket képezzük.

A szezonindex-számítás munkatáblája egy nagyáruház összforgalmi idősorában

8-11. tábla

Év	I.	II.	III.	IV.
	negyedév			
1984	-	-	0,95640	1,19022
1985	0,94789	0,92812	0,96452	1,18551
1986	0,91206	0,89003	0,97495	1,24138
1987	0,90639	0,88414	0,98898	1,19673
1988	0,90639	0,96793	0,95253	1,19101
1989	0,89289	0,95151	0,92691	1,20033
1990	0,96501	0,90426	0,90648	1,13947
1991	1,02513	0,88937	-	-
Átlag	0,93649	0,91648	0,95297	1,19209
s_j	0,93695	0,91692	0,95343	1,19267

Az átlag sorban a nyers szezonindexek találhatók, amelyeket korrigálva saját átlagukkal (0,99995) nyerjük az s_j-vel jelölt korrigált szezonindexeket.

Az átlagszám igen közel esik 1-hez, ami a korrekciót tulajdonképpen nem is indokolja, a művelet itt illusztrációs célokat szolgál, mivel egy kiskereskedelmi egységnél elegendő a százalékos pontosság is.

Az első negyedév kerekített 94 %-os szezonindexe úgy értelmezhető, hogy a szezonhatás 6 %-kal téríti el az idősor értékét a trendtől lefelé. A második és harmadik negyedévben az elmaradás 8, illetve 5 %-os. Mindezek a negyedik negyedévi mintegy 19 %-os többletforgalomban kompenzálódnak.

Az additív modell esetében

$$y_{ij} = \hat{y}_{ij} + s_j^* + v_{ij}^*$$

a szezonhatásnak **szezonális eltérés** az elnevezése. Számszerűsítése az

$$s_j^* = \frac{1}{n}\sum_{i=1}^{n} y_{ij} - \hat{y}_{ij} \qquad (8\text{-}29)$$

formulával történik.

Itt is ellenőriznünk kell, hogy a kapott nyers szezonális eltérésnek nevezett értékek megfelelnek-e a $\sum_{j=1}^{m} s_j^* = 0$ feltételnek. Ha ez nem teljesül, akkor képezzük az s_j^* értékek átlagát,

$$\bar{s}^* = \frac{1}{m}\sum_{j=1}^{m} s_j^* \qquad (8\text{-}30)$$

amit az s_j^* értékekből rendre levonva biztosítjuk a fenti feltétel teljesülését. Gyakran előfordul, hogy alkalmas kerekítéssel is elérhetjük célunkat.

Ha a szükséges korrekció jelentős mértékű, akkor modell-specifikációs problémákra is gyanakodhatunk. Előfordulhat, hogy csak a változó szezonalitást feltételező modell (lásd! 8.4.2 alfejezet) alkalmazása szolgáltat hatékony elemzési eredményeket.

> Additív modellt feltételezve az előző példa adatbázisán végezzük el a szezonális eltérések számítását is! Adatainkat a 8-5. tábla és a 8-7. tábla millió Ft-ra kerekített értékeiből nyertük, amelyek az alapirányzattól additív módon $(y_{ij} - \hat{y}_{ij})$ tisztított értékek.

A szezonális eltérések számításainak munkatáblája a vizsgált nagyáruházi példán (millió Ft)

8-12. tábla

Év	I.	II.	III.	IV.
	negyedév			
1984	-	-	-5	25
1985	-7	-10	-5	27
1986	-13	-17	-4	41
1987	-17	-21	-2	38
1988	-18	-6	-9	38
1989	-21	-10	-15	42
1990	-8	-20	-20	32
1991	6	-28	-	-
Átlag	-11,14	-16,00	-8,57	34,71
s_j^*	-10,89	-15,75	-8,32	34,96

A nyers szezonális eltérések (átlagok) átlaga -0,25 amit rendre levonva a nyers szezonális eltérésekből a korrigált szezonális eltéréseket kapjuk.

A szezonális eltérés például az I. negyedévben úgy értelmezhető, hogy mintegy 11 millió forinttal csökkenti, ugyanakkor a IV. negyedévben mintegy 35 millió forinttal növeli várhatóan a szezonhatás az áruház forgalmát a trendhez képest.

A szezontényező meghatározásának Pearson féle láncindex módszere

Az eddig bemutatott eljárások feltételezték a trend előzetes ismeretét. Azonban ennek hiányában is lehetőségünk van a Pearson féle láncindex módszer alkalmazására, amely a fentiektől eltérően, más módon számszerűsíti a szezonalitást.

A módszer menete a következő:

1. lépés: Az y_{ij} értékeket rendre osztjuk az előző megfigyelési értékkel, vagyis láncviszonyszámokat képezünk. Amennyiben lehetőség van rá, célszerű a vizsgált időszakot egy korábbi megfigyeléssel kiegészíteni az első láncviszonyszám képzése érdekében.

2. lépés: Az első lépésben képzett láncviszonyszámok periódusok (i) szerinti átlagolásával

$$\bar{l}_j = \frac{\sum_{i=1}^{n} l_{ij}}{n}$$

kiszűrjük a véletlen ingadozások hatását. Ezzel minden szezonhoz (j=1,...,m) egy átlagos láncviszonyszámot rendelünk.

3. lépés: A trend meghatározása és kiküszöbölése az átlagos bázisviszonyszámok képzésével történik. Ezt egy negyedéves idősornál az alábbiak szerint számítjuk:

$$b_I = \bar{l}_I$$
$$b_{II} = \bar{l}_I \times \bar{l}_{II}$$
$$b_{III} = \bar{l}_I \times \bar{l}_{II} \times \bar{l}_{III}$$
$$b_{IV} = \bar{l}_I \times \bar{l}_{II} \times \bar{l}_{III} \times \bar{l}_{IV}$$

Könnyen belátható, hogy az utolsó bázisviszonyszám, esetünkben a b_{IV} (általánosan a b_m) fejezi ki az idősorban megmutatkozó trendhatást. Ennek linearitását feltételezve, az egy idényre jutó trendhatás a következőképpen határozható meg:

$$r = \frac{b_{IV} - 100}{4}$$

Nyilvánvaló, hogy a szezonalitás számszerűsítése szükségessé teszi a trendhatás kiküszöbölését az alábbiak szerint:

$$s_I = b_I - r \qquad s_{III} = b_{III} - 3r$$
$$s_{II} = b_{II} - 2r \qquad s_{IV} = b_{IV} - 4r$$

4. lépés: Ebben a lépésben a már bemutatott szezonkorrekció következik. Ez a művelet formailag ugyan azonos az ott bemutatottal, tartalmában azonban jelentősen eltér. Míg az alapesetben sokszor jelentéktelen korrekció adódik, itt az utolsó idény bázisáról a trendbázisra történő áttérés miatt a korrekció általában lényegesen nagyobb mértékben hoz értékmódosulást.

Nézzük meg alappéldánkat, a vizsgált nagyáruház forgalmi idősora alapján a Pearson-féle láncindex módszer gyakorlati alkalmazását. Az alapadatok a 8-1. táblában találhatók.

A Pearson-féle láncindex módszer alkalmazásának munkatáblája (az adatok százalékban):

8-13. tábla

Év	I.	II.	III.	IV.
	negyedév			
1984	81	114	108	131
1985	83	101	106	125
1986	79	102	115	131
1987	75	100	115	125
1988	78	108	99	125
1989	75	108	100	137
1990	81	95	104	131
1991	95	92	115	118
Átlag	80,9	102,5	107,8	127,9
IV. n.év=100	80,9	82,9	89,4	114,3
Trend-korrekció	-3,58	-7,15	-10,73	-14,30
s_j	77,32	75,75	78,67	100,00
s'_j	93,20	91,30	94,90	120,60

Megjegyzés: Az adatbázist kiegészítettük az 1984. I. negyedévi láncviszonyszám képzése érdekében, az 1983. IV. negyedévi 118.917 ezer Ft-os értékkel.

A IV. negyedév = 100 %-os értékei a következőképpen adódnak:

$$82{,}9 = 0{,}809 \times 102{,}5$$

$$89{,}4 = 0{,}829 \times 107{,}8$$

$$114{,}3 = 0{,}894 \times 127{,}9$$

A trendkorrekció sora az $r = (114{,}3 - 100)/4 = 3{,}58$ értékből, illetve többszöröseiből adódik.

Az s_j értékek átlaga, azaz a korrekciós tényező: 82,94 %. Ennek alapján a korrigált szezonindexek (s_j') a tábla utolsó sorában találhatók (például az első negyedévben: $s_I = 77,32 / 0,8294 = 93,2\%$).

A Pearson-féle láncindex módszerrel kapott szezonindexek értelmezése az eltérő származtatási mód ellenére megegyezik a már tárgyalt klasszikus szezonindexekével, és mint látható, számértékeik is közel állnak egymáshoz.

A szezonalitás számszerűsítése lineáris regresszióval

A többváltozós lineáris regresszió segítségével is meghatározhatók a szezonalitás paraméterei. Ebben az esetben a szezonalitás paramétereinek meghatározása érdekében ún. **mesterséges változókat** kell bevezetnünk.

Tételezzük fel, hogy egy negyedéves idősorban lineáris trend érvényesül, a szezonális eltéréseket az I. negyedév bázisán fogjuk értelmezni, ezért a II, III, IV. negyedéveket az x_2, x_3, x_4 mesterséges változóval jellemezzük[10]. Ezek a változók 0 vagy 1 értéket vehetnek fel a például a következő módon:

[10]Az első negyedévet a teljes multikollinearitás miatt (lásd! 7. fejezet) kell elhagyni és ennek bázisán értelmezzük a többi negyedévet.

	x_2 (II. n.év)	x_3 (III. n.év)	x_4 (IV. n.év)
1984. I.	0	0	0
II.	1	0	0
III.	0	1	0
IV.	0	0	1
1985. I.	0	0	0
II.	1	0	0
III.	0	1	0
IV.	0	0	1
.	.	.	.
.	.	.	.
.	.	.	.

Ha a modell konstans tagot, a t időváltozót és három - a negyedéves szezonalitást jellemző - mesterséges változót tartalmaz az X tényezőváltozók mátrixa a következőképpen írható fel:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 0 & 1 & 0 \\ 1 & 4 & 0 & 0 & 1 \\ 1 & 5 & 0 & 0 & 0 \\ 1 & 6 & 1 & 0 & 0 \\ 1 & 7 & 0 & 1 & 0 \\ 1 & 8 & 0 & 0 & 1 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & n \times m & 0 & 0 & 1 \end{bmatrix}$$

Az y vektor tartalmazza az idősor megfigyelt értékeit. Az

$$\hat{y} = b_0 + b_1 t + b_2 x_2 + b_3 x_3 + b_4 x_4 \qquad (8\text{-}31)$$

egyenlet megoldását a

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

művelet adja. A szezonális eltérések ekkor az I. negyedév bázisán értelmezhetők. A b_1 paraméter a lineáris trend paramétere, az első szezonális eltérés pedig a következő módon képezhető[11]:

$$s_I = \frac{-b_2 - b_3 - b_4}{4}$$

A többi szezonális eltérés pedig

$$s_{II} = b_2 + s_I, \quad s_{III} = b_3 + s_I, \quad s_{IV} = b_4 + s_I,$$

módon állítható elő.

A bemutatott módszer a multiplikatív modell számszerűsítésére is alkalmas. Ekkor logaritmikus transzformációval linearizálhatjuk az alapmodellt[12].

> A módszer a nagyáruházra vonatkozó, már ismert számpéldákon a következő eredményeket adta:
>
> $\hat{y} = 114.850 + 5.220t - 3.430x_2 + 3.040x_3 + 47.600x_4$

[11]Bizonyítást lásd! Irodalom [32] II. 314-315 oldal.
[12]Ugyanott Irodalom [32] 278. oldal

a szezonális eltérések értékei pedig a következők:

$$s_I = -10.283$$

$$s_{II} = -15.234$$

$$s_{III} = -\ 8.758$$

$$s_{IV} = 37.317$$

Látható, hogy a fenti módon nyert szezonális eltérések igen közel állnak az additív modellel kapott eredményekhez.

8.4.2 A változó szezonalitás elemzése

Bár a szezonalitást létrehozó okok állandónak tűnnek, mégis gyakran előfordul, hogy megkérdőjelezhető a szezonalitás állandósága. Az okok állandósága mellett az arra reagáló emberi beavatkozás, illetve tevékenység (lásd! a csillapítás stratégiája) olyan hullámmozgásokat is eredményezhet, amikor sem a szezonalitás abszolút, sem relatív állandósága nem tételezhető fel. Mindenképpen célszerű, hogy legalább bevezető ismereteket szerezzünk a változó szezonalitás módszertanáról, hiszen e módszerekkel az állandó szezonalitás is - mint speciális eset - jellemezhető.

A változó szezonalitás modellezésére több módszer ismeretes. Most röviden bemutatjuk a **Census II. módszer** alkalmazásának alaplogikáját. Az idősori alapadatokon többször végzünk **súlyozott mozgóátlagolást** a tendencia számszerűsítésére. Több lépésben **kiküszöböljük** a mozgóátlagolásból származó **rövidülést. Az extrém értékek kiszűrése** után azokat a szomszédos adatok átlagával helyettesítjük. A fenti műveleteket több fázisban alkalmazva előállítjuk a **nyers**, majd ismételt korrekciók után a **végleges** változó szezonindexeket. Ezek jelölése, mivel minden periódus minden szezonjához külön-külön paramétert rendelünk a korábbi s_j helyett s_{ij}. A módszer számszerűsíti az alapirányzat $\hat{y}_{ij}$ értékét is. Egy évre lehet előrejelezni a módszerrel mind az alapirányzat, mind a változó szezonindexek várható alakulását.

A módszer havi bontásos idősorokra kimunkált változatát, és négyidényes változatát a szakirodalomban találhatja meg az olvasó.

Illusztratív példánk Dél-Dunántúl villamos energia fogyasztásának egy idősorát elemzi A vizsgált idősor a Dél-Dunántúli Áramszolgáltató Vállalat (DÉDÁSZ) adott havi csúcsterhelési adatait tartalmazza a magán és az ún. közületi fogyasztókra együttesen.

Villamos energia csúcsterhelés a DÉDÁSZ-nál megawattban

1972-1987 folyamán havi bontásban

8-14. tábla

Év	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII
	hónapok											
1972	192	185	185	181	182	176	185	190	186	194	195	205
1973	290	267	268	261	261	255	259	266	276	284	280	291
1974	288	284	295	282	278	281	285	296	307	312	321	324
1975	318	319	329	319	316	309	306	349	340	339	358	350
1976	357	355	351	344	335	332	348	353	369	375	386	390
1977	368	379	370	365	355	351	361	371	378	385	389	398
1978	385	389	384	382	374	376	393	405	404	426	432	435
1979	431	439	432	432	420	420	423	435	452	455	457	469
1980	464	469	459	454	455	447	462	468	472	499	520	510
1981	504	509	493	483	484	460	484	474	470	486	492	498
1982	494	494	486	486	446	440	459	462	466	504	410	532
1983	530	534	521	490	462	460	476	484	500	532	567	556
1984	578	581	580	564	521	493	505	509	531	554	557	568
1985	563	586	566	556	514	515	512	519	549	567	598	610
1986	604	611	601	575	521	523	532	545	551	602	619	620
1987	625	638	622	578	538	543	540	542	549	622	648	628

A minden egyes megfigyeléshez rendelt, több lépésben meghatározott változó szezonindexeket tartalmazza a következő tábla:

A változó szezonindexek értékei a DÉDÁSZ idősorán

8-15. tábla

Év	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII
	hónapok											
1972	1.117	1.07	1.046	0.994	0.984	0.992	0.926	0.936	0.968	1.002	0.999	1.033
1973	1.089	1.058	1.042	0.990	0.975	0.925	0.934	0.948	0.980	1.006	1.013	1.036
1974	1.057	1.041	1.034	0.986	0.965	0.929	0.946	0.964	0.994	1.010	1.030	1.030
1975	1.031	1.027	1.024	0.984	0.959	0.934	0.958	0.976	0 1.05	1.013	1.041	1.041
1976	1.020	1.022	1.012	0.983	0.955	0.937	0.966	0.983	1.007	1.019	1.044	1.045
1977	1.014	1.022	1.005	0.985	0.956	0.943	0.969	0.988	1.006	1.023	1.038	1.046
1978	1.014	1.024	1.000	0.985	0.957	0.947	0.970	0.989	1.001	1.026	1.037	1.044
1979	1.018	1.027	1.001	0.988	0.961	0.947	0.971	0.984	0.995	1.023	1.036	1.042
1980	1.025	1.032	1.005	0.991	0.960	0.944	0.971	0.976	0.982	1.022	1.041	1.046
1981	1.035	1.042	1.016	0.996	0.952	0.936	0.968	0.966	0.972	1.019	1.042	1.051
1982	1.047	1.052	1.029	1.004	0.994	0.924	0.957	0.956	0.965	1.020	1.044	1.054
1983	1.055	1.063	1.041	1.000	0.938	0.917	0.942	0.945	0.967	1.017	1.046	1.054
1984	1.058	1.069	1.040	1.013	0.935	0.913	0.929	0.938	0.969	1.010	1.048	1.057
1985	1.058	1.076	1.051	1.009	0.928	0.916	0.923	0.933	0.964	1.021	1.056	1.060
1986	1.060	1.079	1.053	1.005	0.921	0.917	0.921	0.932	0.953	1.030	1.064	1.060
1987	1.063	1.082	1.055	0.996	0.916	0.920	0.921	0.930	0.943	1.038	1.074	1057

A módszer egy összetett trendszámítással alapirányzat-becslést és előrejelzést is végez, valamint felismeri a változó szezonindexekben havonként a változás tendenciáját és egy évre előrebecsüli annak értékeit. Az előrebecsült trendértékek és változó szezonindexek segítségével hatékony előrejelzés is végezhető.

A DÉDÁSZ idősorának 1988-ra vonatkozó előrebecslése

8-16. tábla

Hó	Trend	Szezonindex	Előrebecslés
I.	596,9	1,064	635,3
II.	598,2	1,083	648,1
III.	599,4	1,056	633,5
IV.	600,7	0,992	596,4
V.	602,0	0,913	549,9
VI.	603,2	0,921	556,1
VII.	604,5	0,921	557,1
VIII.	605,8	0,929	563,2
IX.	607,0	0,937	569,3
X.	608,3	1,042	634,1
XI.	609,5	1,079	657,7
XII.	610,8	1,056	645,3

A változó szezonindexek alakulása jól jellemzi a szezonalitás módosító hatásának szisztematikus változásait. Ha például a januári szezonindexeket a vizsgált időszak során végig nyomon követjük, akkor a januári szezonalitásból következő többlet érték 1978-ig tartó csökkenését, majd egyenletes növekedését tapasztaljuk.

Az 1972. januári 1,117-es szezonindex úgy értelmezhető, hogy e hónapban a szezonhatás következtében a trendhez képest 11,7 %-os többlet-áramfogyasztás észlelhető. Ez az érték 1978-ra 1,4 %-ra csökkent, majd 1987-re 6,3 %-ra nőtt. A kapott eredmények elemezhetők abból a szempontból is, hogy a szezonalitás jelentősége évről-évre hogyan változott. Ha egy adott év 12 változó szezonindexéből szórásmutatót számolunk, úgy egy olyan átlagszámot kapunk, amellyel jól jellemezhetjük adott évben a szezonalitás erősödéséről, vagy tompulásáról. Ezen számértékeket tartalmazza a következő adatsor:

Év	Szórás
1972	0,056
1973	0,048
1974	0,040
1975	0,034
1976	0,033
1977	0,030
1978	0,029
1979	0,028
1980	0,032
1981	0,038
1982	0,045
1983	0,051
1984	0,056
1985	0,059
1986	0,062
1987	0,065

Látható, hogy 1979-ig a szezonalitás hullámzó mozgása tompult, majd újra szisztematikusan erősödni kezdett. Ez a jelenség a fűtésre használt elektromos energia felhasználásának változásaival függött össze. 1979-ig a fűtésre használt áramot más gazdaságosabb fűtési módokkal váltották ki. A hetvenes évek végén a nyolcvanas évek elején terjedtek el az éjszakai kedvezményes tarifával működtetett jó hatásfokú hőtárolós elektromos kályhák, és ez okozta a szezonalitás erősödését, hiszen ezek csak fűtési idényben használnak energiát.

A változó szezonalitás módszerének kiterjedt pontosság- és hatásosság vizsgálati eszköztára van[13].

[13]Lásd! Irodalom

8.5 IDŐSORI ELŐREJELZÉSEK

Az idősorelemzés alapvetően kettős, de egymással szoros összefüggésben lévő célt tűz ki maga elé. Az egyik a múlt feltárása, tömör jellemzése. E célt szolgálják a már megismert egyszerűbb eszközök, a trendszámítás és a szezonalitás vizsgálatának módszerei. A mozgó átlagolásos trend adatsora is jellemzi a vizsgált idősor múltbeli értékeit.

A múlt megismeréséből származó ismereteink közvetve segítik az idősori elemzések másik céljának elérését, a jövő megismerését. A múltbeli tendenciák előrevetítése alapul szolgálhat a vizsgált folyamatok jövőbeni értékeinek hatékony becsléséhez.

Ebben az alfejezetben összefoglaljuk ezirányú ismereteinket. A már megismert módszerek mellett bemutatunk egy kifejezetten az idősori előrejelzést szolgáló módszert is.

Exponenciális simítás

Az előrejelzés módszerei elkülöníthetők olyan módon, hogy a korábbi idősori megfigyelések egyforma jelentőséggel bírnak-e az előrebecsülni kívánt idősori értékek kialakulásában, illetve azzal a feltételezéssel élünk, hogy az előrejelzés időpontját, illetve időtartamát közvetlenül megelőző, már realizált idősori értékek nagyobb befolyásoló erővel rendelkeznek, és ez a befolyásoló erő időben visszafelé haladva fokozatosan csökken.

Könnyű belátni ezen utóbbi érvelés logikáját. A szakirodalomban széleskörű módszertani leírást találunk különböző típusú idősorok fenti szempontú elemzésére[14].

Jelen esetben csak a legegyszerűbb simító eljárást a kismértékű trendhatást tartalmazó időszak elemzésére alkalmas **exponenciális simítás** módszerét mutatjuk be.

Az előrebecslő formula a következő:

$$y_{n+1} = \alpha y_n + \alpha(1-\alpha)y_{n-1} + \alpha(1-\alpha)^2 y_{n-2} + \ldots + \alpha(1-\alpha)^{k-1} y_{n+1-k} \qquad (8\text{-}32$$

ahol: α = az ún. kiegyenlítési konstans, melyre igaz, hogy $0 \leq \alpha \leq$.

k = a súlyozott tagok tagszáma.

Az α kiválasztásánál - bár léteznek statisztikai módszerek az optimális értékének meghatározására - elsősorban elméleti megfontolásokból indulunk ki. 1-hez közeli α nagyobb súlyt ad az utolsó megfigyeléseknek, csökkentése pedig a korábbi megfigyelések befolyásoló erejét növeli, ami nem változtat azon a tényen, hogy a korábbi megfigyelésekhez tartozó súlyok minden esetben a mértani sornak megfelelően csökkennek.

A k megválasztása összefügg az α értékével is. Módszertani követelmény, hogy a súlyok összege egyhez közelítsen. A

$$\sum_{k=1}^{\infty} \alpha(1-\alpha)^{k-1} = 1$$

kifejezés tárja fel az α és k viszonyát.

---

[14]Igen elterjedt a gyakorlatban a Brown- és Winter féle exponenciális kiegyenlítési eljárás. (lásd! Irodalom).

A gyakorlatban $\alpha = 0,5$ körüli értékekkel és minimálisan k=5 értékkel célszerű számolnunk. Megjegyzendő, hogy az exponenciális simítás nemcsak a klasszikus értelemben vett **ex ante előrejelzésre** (a realizált idősoron kívül, jövőbeni idősori értékekre), hanem az **ex post előrejelzésre** is alkalmas, értelemszerű módosításokkal. Az ex post előrejelzés azt jelenti, hogy már realizálódott idősori értéket becsülünk korábbi megfigyelésekkel. Ilyen módon az exponenciális simítás a trendszámítási módszerekhez is csatolható.

Vizsgáljuk meg a 8.3.2 fejezet számpéldáján a művelt mezőgazdasági terület (ezer hektár) idősorán az exponenciális simítással történő előrebecslést.

Legyen: $\alpha = 0,5$ és $k = 7$

Munkatábla

8-17. tábla

Év	y_t	Súly
1984	6.554	0,007813
1985	6.540	0,015625
1986	6.524	0,031250
1987	6.511	0,062500
1988	6.497	0,125000
1989	6.484	0,250000
1990	6.473	0,500000

Exponenciális simítással végzett előrebecslés 1991-re:

$$\hat{y}_t = 0{,}007813 \times 6.554 + 0{,}015625 \times 6.540 + \ldots + 0{,}5 \times 6.473 =$$

$$= 6.433{,}829 \approx 6.434 \text{ ezer hektár}$$

Előrejelzések trendfüggvények alapján

Ebben az alfejezetben a már megismert trendfüggvények alapján történő előrejelzéseket foglaljuk össze. Az előrejelzések esetén mindenkor feltételezzük az elemzési időszakra megállapított alapirányzat változatlanságát.

Az átlagos abszolút változás mutatója ($\overline{D}$)alapján az előrebecslés a következő képletekkel végezhető el:

$$\hat{y}_{\overline{D}(n+1)} = y_1 + n \times \overline{D} = y_n + \overline{D} \qquad (8\text{-}33)$$

A 8.1 fejezetben bemutatott példában az átlagos abszolút változás mutatójára $\overline{D}$ = -18,958-as értéket kaptunk. Az 1991-re vonatkozó előrejelzés:

$$\hat{y}_{\overline{D}} = 6.928 + 25(-18{,}958) = 6.473 - 18{,}958 \approx 6.454 \text{ ezer hektá r}$$

A lineáris trendfüggvény alapján az előrebecslés az alábbi módon végezhető:

$$\hat{y}_{lin(n+1)} = b_0 + b_1(n+1) = \hat{y}_n + b_1 \qquad (8\text{-}34)$$

A 8.3.2. fejezetben szereplő példában a lineáris trendre vonatkozóan a következő egyenletet kaptuk:

$$\hat{y}_t = 6.969{,}44 - 21{,}05t$$

Az 1991-re vonatkozó előrejelzés :

$$\hat{y}_{lin} = 6.969{,}44 - 21{,}05 \times 26 = 6.443 - 21{,}05 \approx 6.422 \text{ ezer hektár}$$

Az előző három eredményt összevetve megállapíthatjuk, hogy jelentősen eltérnek egymástól. Egyértelmű hatékonysági rangsort - azon túl, hogy a trendbecslés hatékonyabb módszer a $\overline{D}$ mutatónál, hiszen minden egyes megfigyelést bevon a vizsgálatba - a szakirodalom nem állít fel, így csak az adott idősor alapos, szakmai ismeretekkel kiegészített elemzése adhat választ arra, mi a végső állásfoglalásunk az előrebecslést illetően. Célszerű a matematikai-statisztikai előrejelzéseket szakmai kontroll alá venni, és ha szükséges, akár felülbírálni. Inkább értékes kiinduló adatbázist, mintsem megfellebbezhetetlen végeredményt jelentenek ezek a becslések.

Folytassuk az előrejelzések tárgyalását az exponenciális trendbecslés és az átlagos relatív változás összehasonlító elemzésével. Mivel a módszer a relatív változás állandóságát vélelmezi, így értelemszerű összevethetőségük.

Az előrebecslés az **átlagos relatív változás mutatója** ($\bar{l}$) a következő képlettel végezhető el:

$$\hat{y}_{l(n+1)} = y_1 \times \bar{l}^n = y_n \times \bar{l} \qquad (8\text{-}35)$$

A 8.3.2. fejezetben vizsgált ismertté vált felnőtt bűnelkövetők adatsorán, 1991-re előrebecslésünk a következő:

$$\hat{y}_{\bar{l}} = 63 \times 1{,}0293^{17} = 100 \times 1{,}0293 = 102{,}93 \text{ ezer fő}$$

Az **exponenciális trendfüggvény** alapján az előrebecslés a következő képlettel végezhető el:

$$\hat{y}_{exp(n+1)} = b_0 b_1^{n+1} = \hat{y}_n b_1 \qquad (8\text{-}36)$$

A vizsgált idősoron a 8.3.2 fejezetben az exponenciális trendre vonatkozóan a következő egyenletet kaptuk:

$$\hat{y}_t = 62{,}59 - 1{,}018^t$$

az előrejelzett érték 1991-re

$\hat{y}_{exp} = 62{,}59 \times 1{,}018^{18} = 85 \times 1{,}018 \approx 86$ ezer fő

A kétféle előrejelzés nagymértékben eltér egymástól. Ha megfigyeljük az idősort, látható, hogy az utolsó év ugrásszerű növekedést hozott magával. Ha a jelenség ismeretében ezen erősödő növekedést tendenciózusnak véljük, akkor az $\bar{l}$-on alapuló előrejelzést, ellenkező esetben a trendbecsléses előrejelzést fogadjuk el attól függetlenül, hogy a trend-előrebecslés hatékonyabbnak tekinthető, hiszen az összes megfigyelést bevonja a vizsgálatba.

Előrejelzések szezonalitás esetén

Vizsgáljuk meg, hogyan alakul az előrejelzés a **szezonális hullámzást** mutató idősorok esetén. Természetesen az alapirányzat változatlansága mellett feltételezzük a szezonalitás változatlanságát is az előrejelzési időszakban.

Az előrejelzés **multiplikatív modell** esetén a j-edik szezonra:

$$\hat{y}_{lin} \times s_j = \hat{y}_{multipl} \qquad (8\text{-}37)$$

illetve **additív modell** esetén a j-edik szezonra

$$\hat{y}_{lin} + s_j^{*} = \hat{y}_{additív} \qquad (8\text{-}38)$$

A 8.3 és 8.4 fejezetben megismert nagyáruházi forgalom elemzését most egészítsük ki egy lineáris trendbecsléssel. A hatásos előrejelzés érdekében ugyanis előre kell becsülnünk a következő év negyedéves bontású forgalmi értékeit, és tananyagunkban nem tárgyalunk olyan eljárást, amelyik mozgóátlagolással előre tud becsülni.

A trendegyenlet becslésénél a következő eredményeket kaptuk (adatok ezúttal millió Ft-ban).

$$n = 32, \quad \sum t = 528, \quad \sum y_t = 5.927$$

$$\sum ty_t = 110.378, \quad \sum t^2 = 11.440, \quad \sum y_t^2 = 1.173.721$$

ezekből: $b_0 = 109,1149 \approx 109$ millió

$b_1 = 4,61235 \approx 4,61$ millió

tehát, $\hat{y}_t = 109 + 4,61 \times t$

Az 1992-es trend-előrejelzés tehát:

	t	y_{lin}
I.n.év	33	261
II.n.év	34	266
III.n.év	35	271
IV.n.év	36	275

A 8.4 fejezetben kapott szezonindexek rendre 94 %; 92 %; 95 %; és 119 %; a szezonális eltérések pedig -11, -16, -8, 35 millió forint.

Az előrejelzés például 1992. I. negyedévére a szezonindexszel

$$\hat{y}_{multipl} = 261 \times 0,94 = 245 \text{ millió Ft},$$

illetve a szezonális eltéréssel

$$\hat{y}_{additív} = 261 - 0,94 = 250 \text{ millió Ft}.$$

Az 1992. évre előrejelzett értékek (millió Ft):

Negyedév	$y_{multipl.}$	$y_{additív}$
I.	245	250
II.	245	250
III.	257	263
IV.	327	310

A kétfajta előrejelzés eltér egymástól. Megállapíthatjuk, hogy a szezonális hullámzást a multiplikatív modell erősebbre becsüli, mint az additív. Az $y_{multipl}$ értékből származott szórás 34 millió Ft, míg az $y_{additív}$ értékei 25 milliós szórást mutatnak. Ha visszagondolunk a multiplikatív és az additív modellt bemutató ábránkra, akkor könnyen magyarázni tudjuk a kapott eltérést.

Az előrejelzések tárgyalásakor csak felvillantani kívántuk a rendelkezésre álló módszertani apparátus néhány praktikusan alkalmazható eljárását, az érdeklődő a szakirodalomban talál további ismeretanyagot e témakörben.

8.6. SZTOCHASZTIKUS IDŐSORI MODELLEK

A sztochasztikus idősori modellezés leggyakrabban alkalmazott eljárásai amelyeket már sok számítógépes programcsomag tartalmaz - az **integrált autoregresszív és mozgóátlag (ARIMA) modellek** néven ismeretesek. E modellek elméleti tulajdonságainak, és a modellek illesztésével kapcsolatos eljárásoknak egységes rendszerbe foglalása **G.E.P. Box** és **G.M. Jenkins** nevéhez fűződik. Ezért szokás ezeket **Box-Jenkins módszernek** is nevezni.

Mivel a sztochasztikus idősori modellezés ismeretanyaga messze meghaladja e tananyag kereteit, ebben a fejezetben csak a sztochasztikus folyamatokkal kapcsolatos fontosabb fogalmakat, illetve az ARIMA modellezés folyamatának - számítógépes outputok alapján történő - bemutatását tudjuk megtenni.

Először az ARIMA modellek típusait, illetve a modellkészítés menetét röviden ismertetjük, majd a modellezés egyes fázisait, illetve az ezekhez kapcsolódó fogalmakat és eljárásokat tárgyaljuk.

8.6.1 Az ARIMA modellek típusai, a modellkészítés menete

Az **autoregresszív (AR) modell** az idősor t időszaki értékét (Y_t-t), saját korábbi t-1, t-2,.....t-p időszaki értékeinek lineáris kombinációjaként és egy véletlen ingadozást reprezentáló ε_t változóval fejezi ki:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t \qquad (8\text{-}39)$$

ahol: p az autoregresszivitás rendjét jelöli.

A mozgóátlag (MA) modell az idősor t időszaki értékét (Y_t-t) a t, t-1, t-2,.... t-q időszakokhoz tartozó véletlen változók lineáris kombinációjaként állítja elő

$$Y_t = \varepsilon_t + \Theta_1\varepsilon_{t-1} + \Theta_2\varepsilon_{t-2} + \ldots + \Theta_q\varepsilon_{t-q} \tag{8-40}$$

ahol: q a mozgóátlag-folyamat rendjét jelöli.

Az **autoregresszív és mozgóátlag (ARMA) modellek** a fenti AR és MA modellekből álló (p,q) rendű vegyes modellek

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \Theta_1\varepsilon_{t-1} + \ldots + \Theta_q\varepsilon_{t-q} \tag{8-41}$$

A fenti modellekben szereplő ε_t véletlen változókról feltételezzük, hogy várható értékük nulla, szórásnégyzetük konstans, és autokorrelálatlanok, azaz

$$E(\varepsilon_t) = 0 \quad \mathrm{var}(\varepsilon_t) = \sigma^2 \quad \rho_k = 0 \quad (k = 1,2,\ldots).$$

E feltételek teljesülése esetén **teljesen véletlen folyamatról**, ún. **fehér zaj** (white noise) **folyamatról** beszélünk.

Az autoregresszív és/vagy mozgóátlag (AR, MA, ARMA) modellek alapvetően a **stacionárius idősorok elemzésére** alkalmasak, amelyek egy konstans átlagszint körül ingadoznak, állandó szórással.

Mivel a gazdasági, társadalmi idősorok többsége jelentős fejlődést mutat, általában jellemző rájuk az emelkedő, vagy csökkenő tendencia ahhoz, hogy ezen idősorokhoz a fenti modellek illeszthetők legyenek, ezért valamilyen eljárással **stacionáriussá kell tenni** azokat. Ezen eljárások jelenthetnek pl. differencia képzést, vagy az Y_t idősor logaritmikus transzformációját.

A differencia képzéssel stacionáriussá tehető ún. **integrált** (I) idősorokra felirt ARMA modelleket **ARIMA modelleknek** nevezzük. E modellek dimenzióit az elnevezések sorrendjében az autoregresszivitás rendjével (p-vel), a differenciaképzés fokával (d-vel) és a mozgóátlag folyamat rendjével (q-val) adjuk meg, a következő módon: ARIMA(p,d,q).

A gyakorlati alkalmazások szerint az idősorok nagy része jól közelíthető olyan modellekkel, melyeknél az autoregresszivitás - és a mozgóátlag folyamat rendje (p és q), illetve a differenciaképzés foka (d) alacsony. Általában mindhárom dimenzió - a (p), a (d), és a (q) is 0, vagy 1, vagy 2 értéket vesz fel.

Olyan idősorok elemzésére, melyek szezonális ingadozást is tartalmaznak, az ún. **szezonális ARIMA modellek** alkalmasak, melyekkel pl. havi adatsorok (s=12), vagy negyedéves adatsorok (s = 4) elemezhetők.

Az eddig elmondottakból látható, hogy a sztochasztikus modellezés kiindulópontja annak megállapítása, hogy a vizsgálni kívánt idősorunk stacionárius-e, illetve, ha nem, akkor az, hogy alkalmas transzformációval stacionáriussá tehető-e. Ezzel eldöntöttük azt, hogy az adott idősorhoz illeszthető-e ARIMA modell, ha igen milyen (d) dimenzióval rendelkezik. A következő fontos kérdés annak megválaszolása, hogy **AR, MA, vagy ARMA modell** illesztésével próbálkozzunk-e, és milyen legyen az autoregresszivitás (p) és/vagy, a mozgóátlagolás (q) rendje. Erre a kérdésre a választ az **idősorok korrelációs stuktúrájának** feltárása adhat. Ez az idősorokban meglévő belső összefüggések vizsgálatát jelenti, amely az Y_t, Y_{t-1}, Y_{t-2},... változók sztochasztikus kapcsolatainak megállapításával végezhető el. E sztochasztikus kapcsolatok az **autokorrelációs együtthatókkal** mérhetők. A modellezés ezen fázisát, **modell azonosításnak (identifikációnak)** nevezi a szakirodalom.

Ezután a modellezés lépései alapvetően megfelelnek a már ismert lineáris regressziós modellezésnek. A választott **modell paraméterbecslése** után a **modell ellenőrzése** következik. Ezután döntünk arról, hogy felhasználható-e az illesztett modell **elemzésre, előrejelzésre**, vagy más modell választásával kell próbálkoznunk.

Mivel a sztochasztikus modellezés az idősorokban meglévő összefüggések feltárásán alapul, rövid idősorok nem alkalmasak a modellezésre. Legalább 30-50, illetve szezonális modellek esetén legalább 100 elemű idősorra van szükség.

8.6.2 A stacionárius idősor fogalma; a stacionaritás biztosítása

Először azt nézzük meg, hogy miért szükséges a stacionarítás feltétele a sztochasztikus modellezés során.

A sztochasztikus modellezés adatbázisát jelentő y_t (t= 1,2,....,n) tapasztalati idősor - mint ismeretes- az elméleti idősort alkotó Y_t (t = 1,2, ..., T) valószínűségi változóknak egyetlen realizációja. Így a tapasztalati idősor csak akkor tekinthető egy olyan n elemű mintának, melyből az elméleti idősort jelentő sztochasztikus folyamat jellemzői becsülhetők, ha az elméleti idősor jellemzői (várható értéke, szórásnégyzete, autokorrelációs együtthatói) **időben állandóak**, azaz függetlenek a t változótól. Az ilyen tulajdonságokkal rendelkező idősorokat **stacionáriusnak** nevezzük.

A stacionárius idősort alkotó változók várható értéke, szórásnégyzete és a k-ad rendű autokorrelációs együtthatója a következőképpen írható fel:

$$\begin{array}{lll} Y_1 & E(Y_1)=\mu & Var(Y_1)=\sigma^2 \\ Y_2 & E(Y_2)=\mu & Var(Y_2)=\sigma^2 \\ \vdots & \vdots & \vdots \\ Y_t & E(Y_t)=\mu & Var(Y_t)=\sigma^2 \end{array} \qquad (8\text{-}42)$$

$$\rho_k = \frac{Cov Y_t, Y_{t-k}}{\sigma^2}, \quad k=1,2,\ldots,K \qquad (8\text{-}43)$$

ahol:

$$\mathrm{Cov}Y_t, Y_{t-k} = E\left[(Y_t - \mu)(Y_{t-k} - \mu)\right]$$

$$\sigma^2 = \sigma_t \times \sigma_{t-k}, \text{ mivel } \sigma_t = \sigma_{t-k} = \sigma$$

A stacionárius idősorok tehát **nem tartalmaznak trendhatást**, az idősor értékei egy állandó átlagos szint körül ingadoznak, állandó szórással. Az állandó szórás azt jelenti, hogy az **ingadozások intenzitása időben nem változik** (nem növekszik vagy csökken). Mindezek általában a tapasztalati idősor ábrája alapján megállapíthatók.

Ezenkívül a stacionárius idősorokra jellemző az, hogy az idősorban lévő belső összefüggéseket reprezentáló ρ_k **autokorrelációs együtthatók időben állandóak**, nem függnek t-től, csak a változók egymás közötti távolságától, k-tól. A ρ_k autokorrelációs együtthatók k = 1, 2, ..., K függvényében az ún. **autokorrelációs függvényt** alkotják. Pl. k = 1 esetén, ρ_1 az elsőrendű autokorrelációs együttható az egymás után következő idősori értékek, k = 2 esetén ρ_2 a másodrendű autokorrelációs együttható, az egymástól két időegységnyi távolságra lévő idősori értékek korrelációját mutatják. Az autokorrelációs függvény értékei a k késleltetés különböző értékeihez rendelve, az alábbi módon foglalhatók össze:

k	1	2	3	...	K
ρ_k	ρ_1	ρ_2	ρ_3	...	ρ_K

A stacionárius idősorok gyakran tartalmaznak **rövid távú autokorrelációt**, amely pl. azt jelenti, hogy ρ_1 értéke viszonylag magas, ρ_2 értéke is szignifikánsan különbözik nullától, de k növekedésével a ρ_k értékek nullához tartanak.

A stacionárius idősorok jellemzői a tapasztalatai idősorból, a már ismert becslőfüggvények segítségével becsülhetők. Az **idősor μ várható értéke és σ^2 szórásnégyzete** a tapasztalati idősor átlaga és szórásnégyzete alapján becsülhető:

$$\bar{y} = \frac{\sum_{t=1}^{n} y_t}{n} \qquad (8\text{-}44)$$

$$s_y^{\,2} = \frac{\sum_{t=1}^{n} (y_t - \bar{y})^2}{n} \qquad (8\text{-}45)$$

Az $\bar{y}$ adja meg azt a konstans szintet, mely körül a tapasztalati adatok ingadoznak, s_y szórással.

A ρ_k autokorrelációs együtthatók becslése a tapasztalati idősorból a következőképpen történik (a lineáris korrelációs együttható - r - számításának megfelelően):

$$r_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{(n-k)s_y^{\,2}}, \quad k = 1,2,\ldots,K \qquad (8\text{-}46)$$

A késleltetés különböző értékeihez (k = 1,2,....,K) rendelt becsült autokorrelációs együtthatók autokorrelációs függvénye is felírható az alábbi módon:

k	1	2	3	...	K
r_k	r_1	r_2	r_3	...	r_K

A becsült első-, másod-, és harmadrendű autokorrelációs együttható számítási módja a következő:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \bar{y})(y_{t-1} - \bar{y})}{(n-1)s_y^2}$$

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y})(y_{t-2} - \bar{y})}{(n-2)s_y^2}$$

$$r_3 = \frac{\sum_{t=4}^{n}(y_t - \bar{y})(y_{t-3} - \bar{y})}{(n-3)s_y^2}$$

Látható, hogy r_1, az egymás után következő idősori értékek, r_2, az egymástól két, az r_3 pedig az egymástól három időegységnyi távolságra lévő idősori értékek korrelációs együtthatói. A magasabbrendű (pl. k = 4,5...) autokorrelációs együtthatók a fentieknek megfelelően számíthatók.

Annak megállapítása, hogy a vizsgált idősorunk stacionáriusnak tekinthető-e, a tapasztalati idősorunk grafikus ábrája alapján történhet. Ha azt tapasztaljuk, hogy az idősorunk egy konstans átlagszint körül ingadozik, megközelítőleg állandó szórással, akkor stacionárius idősorként kezeljük.

Korrelálatlan Y_t változókból álló stacionárius idősor esetén $\rho_k = 0$, bármilyen $k = 1, 2, ...,$ késleltetés esetén. Ilyenkor az r_1, r_2 ... becsült autokorrelációs együtthatók egyike sem különbözik szignifikánsan a nullától. Amennyiben az $\bar{y}$ konstans átlagszint sem különbözik szignifikánsan a 0-tól és állandó a szórás, az elméleti idősor - melyből tapasztalati idősorunk származik - véletlen azaz, **fehér zaj folyamatot** követ.

Mivel a stacionárius idősorok gyakran tartalmaznak rövid távú autokorrelációt, a tapasztalati idősorból becsült r_k autokorrelációs együtthatók k függvényében történő gyors csökkenése is jelzi a vizsgált idősor stacionaritását. Például az r_1 és r_2 még szignifikáns, az r_3, r_4, r_5,... értéke már nullának tekinthető.

Az r_k becsült autokorrelációs együtthatók szignifikancia-vizsgálata, az együtthatók standard hibája és konfidencia intervalluma alapján végezhető el, melynek ismertetésétől eltekintünk.

A trendhatást tartalmazó idősorok nagy részéből **differencia képzéssel** a trendhatás kiszűrhető. Amennyiben az alapirányzatra a polinomiális változás jellemző, az idősor elemeinek megfelelő fokú differenciái egy konstans szint körül ingadoznak. Pl. lineáris tendencia esetén ez az első differenciák esetén teljesül.

Egy idősor első differenciái (d = 1) az idősor szomszédos adatai közötti különbségeket jelentik:

$$\Delta Y_t = Y_t - Y_{t-1} \qquad (8\text{-}47)$$

A második differenciák (d = 2), az első differenciákból számíthatók:

$$\Delta^2 Y_t = \left(Y_t - Y_{t-1}\right) - \left(Y_{t-1} - Y_{t-2}\right) = Y_t - 2Y_{t-1} + Y_{t-2} \qquad (8\text{-}48)$$

és így tovább.

Amennyiben idősorunkból megfelelő fokú differenciaképzéssel a trendhatást sikerült kiszűrni, és az így létrejött differencia sor szórása állandó - tehát stacionáriusnak tekinthető idősort nyertünk -, akkor a továbbiakban ez a differencia sor képezi a modellezés adatbázisát. Ezután a differencia sorból becsüljük az autokorrelációs együtthatókat (r_k, ahol: k = 1, 2, ...)

Ha az idősor első differencia sorából számított autokorrelációs együtthatók egyike sem különbözik szignifikánsan a nullától, és a ΔY_t egy $\mu \neq 0$ várható érték körül ingadozik, állandó szórásnégyzettel, akkor a modell az alábbi formában írható fel:

$$\Delta Y = \mu + \varepsilon_t$$

Átalakítva:

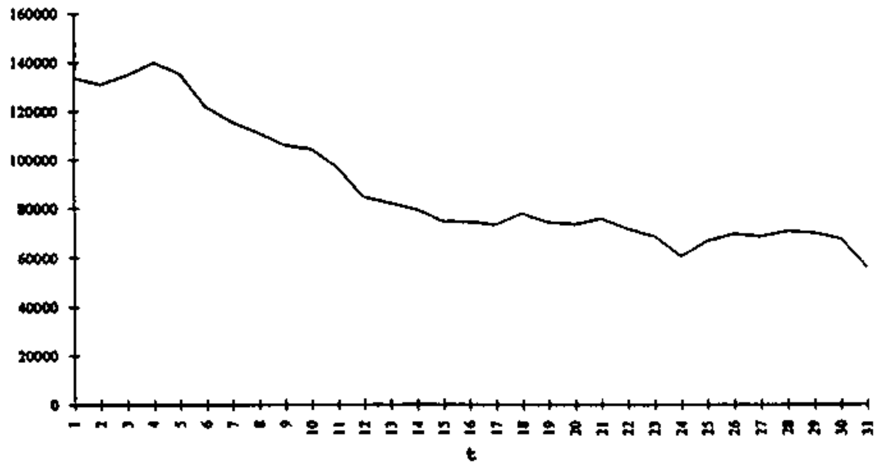
$$Y_t = Y_{t-1} + \mu + \varepsilon_t \qquad (8\text{-}49)$$

Ez az ún. **véletlen bolyongási folyamat** (random walk) egyik változata, amely a lineáris trend sztochasztikus megfelelője. A fenti modell szerint az idősor értéke egyik időszakról a másikra egy állandó,- és egy véletlen értékkel változik.

A szórás stabilizálására leggyakrabban a **logaritmikus transzformációt** használják[15]. Az idősor logaritmálása után, ha szükséges, megfelelő differencia képzéssel a konstans átlagszint elérhető.

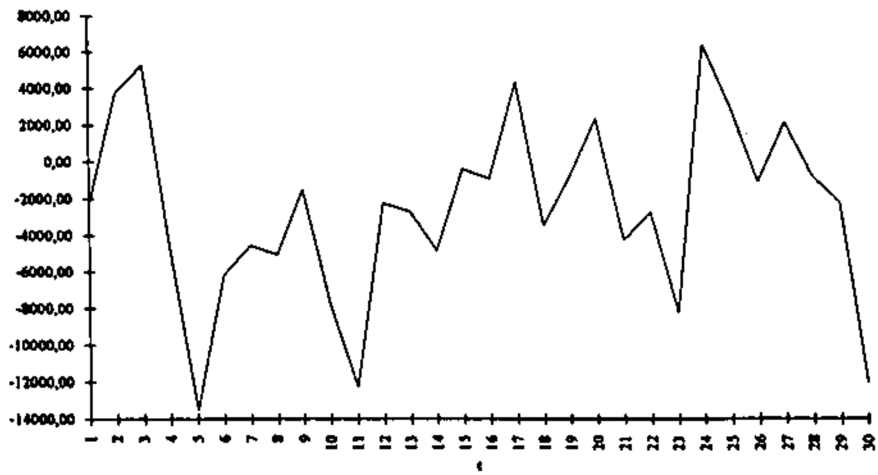
Az eddig elmondottak szemléltetésére a mozilátogatások számának alakulását vizsgáljuk 1957-1987-ig. (Forrás: Statisztikai Évkönyvek). A vizsgált idősor 8-8. ábrája alapján jelentős csökkenés tapasztalható, tehát nem stacionárius az idősor.

[15]Használatos még a Box-Cox transzformáció (lásd: Irodalom).



8-8. ábra A mozilátogatások számának alakulása (ezerben) 1957-1987-ig (t = 1,2,...,31)

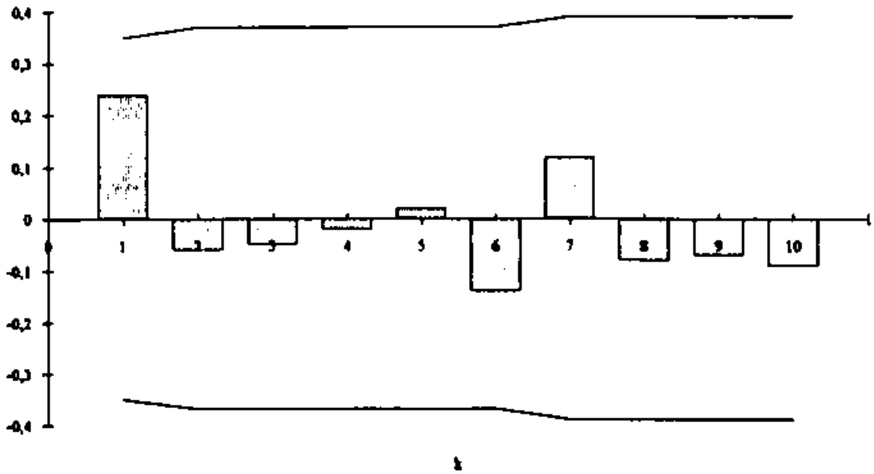
Első fokú differenciaképzéssel próbáltuk meg stacionáriussá transzformálni az idősort, melynek eredményét a 8-9. ábra tartalmazza.



8-9. ábra A mozilátogatások számának első differencia sora

A 8-9. ábra alapján megállapítható, hogy az első differenciák megközelítőleg egy átlagos szint körül ingadoznak, elég jelentős, de szabálytalan hullámzással. Elfogadva a szórás állandóságára tett feltételezést, az első differenciák idősorának átlagos értéke: -2.586; átlagbecslés standard hibája: 913; a t-próba értéke: -2,83. Ez azt jelenti, hogy a vizsgált idősorunk várható értéke 5 %-os szignifikancia szinten különbözik nullától. Az első differenciák idősora tehát $\overline{\Delta Y_t}$.= -2.586-os átlagos szint körül ingadozik, megközelítőleg konstans szórással, azaz stacionáriusnak tekinthető.

Az autokorreláció függvény értékeit ábrázolva, az idősor korrelogramját kapjuk. A 8-10. ábra, az első differencia sor korrelogramját mutatja, melyben az r_k autokorrelációs együtthatók konfidencia intervallumát is feltüntettük.



8-10. ábra A mozilátogatások számából képzett első differencia sor korrelogramja

A korrelogram az első tíz autokorrelációs együttható becsült értékét tartalmazza (r_k, k = 1, 2, ... 10). Az ábrán látható konfidencia sávon belül helyezkednek el az r_k értékek, így egyikük sem különbözik szignifikánsan a nullától, 5 %-os szignifikancia szinten. Ez azt jelenti, hogy a transzformált sor a $\overline{\Delta Y_t} = -2.586$-os érték körül véletlenszerűen ingadozik, azaz a modell

$$\Delta Y_t = Y_t - Y_{t-1} = -2.586 + \varepsilon_t$$

Az első differenciák, azaz az egyik időszakról a másik időszakra történő változások egy konstans értékkel és egy véletlen értékkel változnak.

Átrendezés után az alábbi egyenletet kapjuk:

$$Y_t = Y_{t-1} - 2.586 + \varepsilon_t$$

A konstans érték (-2.586 ezer, azaz kb. 2,6 millió) jelentése hasonló a lineáris trend b_1 paraméteréhez, az időegységenkénti változás átlagos nagyságát mutatja. Esetünkben a mozilátogatások száma évenként átlagosan mintegy 2,6 millióval csökkent 1957 és 1958 között. A fenti modell a lineáris trend sztochasztikus megfelelője.

8.6.3 A modell azonosítása

Miután meggyőződtünk arról, hogy idősorunk stacionárius, illetve valamilyen módon stacionáriussá alakítottuk, a modell azonosítása (identifikáció) következik. Ha a tapasztalati idősorunkat valamilyen transzformációval (logaritmálással és/vagy differencia képzéssel) tettük stacionerré, akkor a modellezés adatbázisát a továbbiakban ez a transzformált idősor képezi. Amennyiben differenciaképzés történt, az idősorunk integrált (I), amelynek ismert a d (differenciaképzés foka) dimenziója. A modell azonosítás azt jelenti, hogy megpróbáljuk megkeresni azt az MA(q), AR(p), vagy ARMA(p,q) modelltípust, **amelynek jellemzőire leginkább hasonlítanak a tapasztalati idősorunkból számított empirikus jellemzők.** Ehhez ismernünk kell a különböző modelltípusok elméleti jellegzetességeit, melyek részletes leírása a szakirodalomban megtalálható.

A legfontosabb információt a modell azonosításához az **autokorrelációs függvény** értékei (ρ_k, k = 1, 2, ..., K), illetve az ezek alapján meghatározható **parciális autokorrelációs függvény** értékei adnak. (A parciális autokorrelációk tárgyalásától eltekintünk; lényegét tekintve megegyeznek a többváltozós korrelációszámításban megismert parciális korrelációs együtthatókkal.)[16]

A mintából becsült autokorrelációs (és parciális autokorrelációs) együtthatók grafikus ábrája, a **korrelogram** alapján lehet a legkönnyebben az autokorrelációs együtthatók viselkedését - a késleltetés (k) függvényében - tanulmányozni. Ugyanis a r_k becsült autokorrelációs együtthatók konfidencia intervalluma alapján közvetlenül megállapíthatók a nullától szignifikánsan különböző r_k értékek. Ezek a konfidencia sávon kívül helyezkednek el.

A modell azonosítás bemutatására - a két legegyszerűbb modellel leírható folyamat - az MA(1) és az AR(1) folyamat autokorreláció függvényét nézzük meg.

Az MA(1) folyamat

$$Y_t = \varepsilon_t + \Theta_1 \varepsilon_{t-1} \qquad (8\text{-}50)$$

mindig stacionárius, autokorreláció függvényére pedig az jellemző, hogy csak az első (ρ_1) érték nem nulla.

Az MA(1) folyamat autokorrelációs együtthatói a Θ_1 paraméterből nyerhetők:

$$\begin{aligned} \rho_1 &= \frac{\Theta_1}{1+\Theta_1^2}, && \text{ha } k=1 \\ \rho_k &= 0, && \text{ha } k>1 \end{aligned} \qquad (8\text{-}51)$$

---

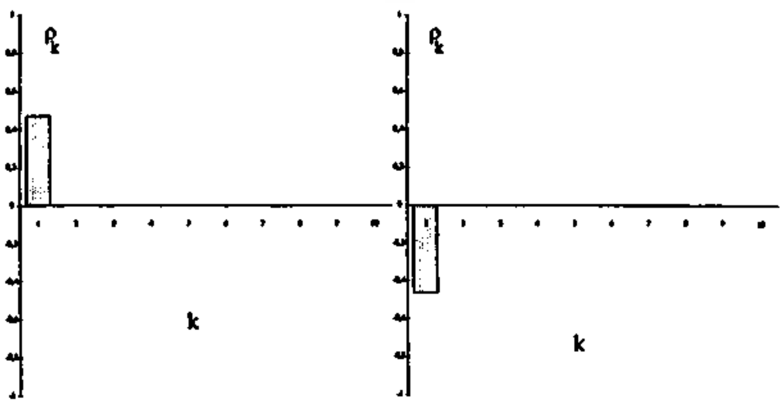
[16]A különböző modelltípusok autokorrelációs és parciális autokorrelációs függvényének sajátosságait foglalja össze a fejezet végén lévő Függelékben szereplő tábla.

Legyen például: a) $\Theta_1 = 0,7$, b) $\Theta_1 = -0,7$, akkor az elsőrendű autokorrelációs együttható értéke

$$\text{a.)}\ \rho_1 = \frac{0,7}{1+0,7^2} = 0,47$$

$$\text{b.)}\ \rho_1 = \frac{-0,7}{1+(-0,7)^2} = -0,47$$

A megfelelő korrelogrammokat a 8-11. ábra tartalmazza.



8-11.ábra Az MA(1) folyamat korrelogramja

a) $\Theta_1 = 0,7$ és b) $\Theta_1 = -0,7$ esetén

AZ AR(1) folyamat

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t \qquad (8\text{-}52)$$

akkor stacionárius, ha $|\phi_1| < 1$ fennáll.

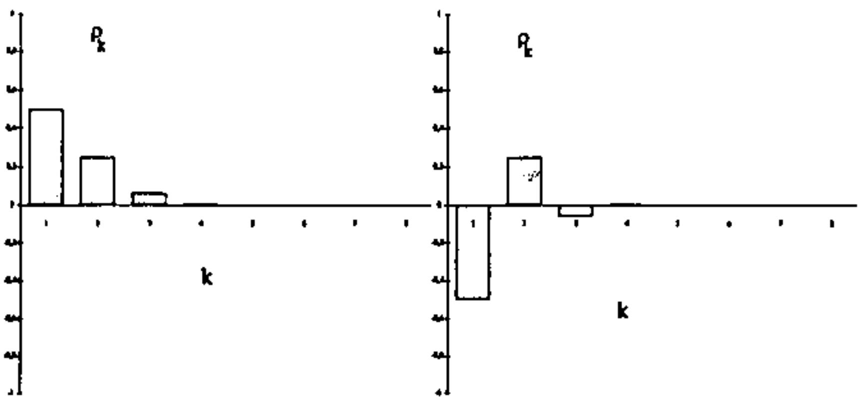
Autokorreláció függvénye exponenciálisan csökkenő, ha ϕ_1 pozitív érték ($\phi_1 > 0$), és csillapodó szinusz görbe szerint csökken, ha ϕ_1 negatív érték ($\phi_1 < 0$). Az AR(1) folyamat autokorrelációs együtthatói a ϕ_1 paraméterből származtathatók:

$$\rho_k = \phi_1^k \quad k = 1,2,\dots,K \tag{8-53}$$

Legyen például a) $\phi_1 = 0{,}5$ és b) $\phi_1 = -0{,}5$, akkor az autokorreláció függvény értékei:

	k	1	2	3	...
a)	ρ_k:	0,5	0,25	0,125	...
b)	ρ_k:	-0,5	0,25	-0,125	...

A megfelelő korrelogrammokat a 8-12. ábra tartalmazza.

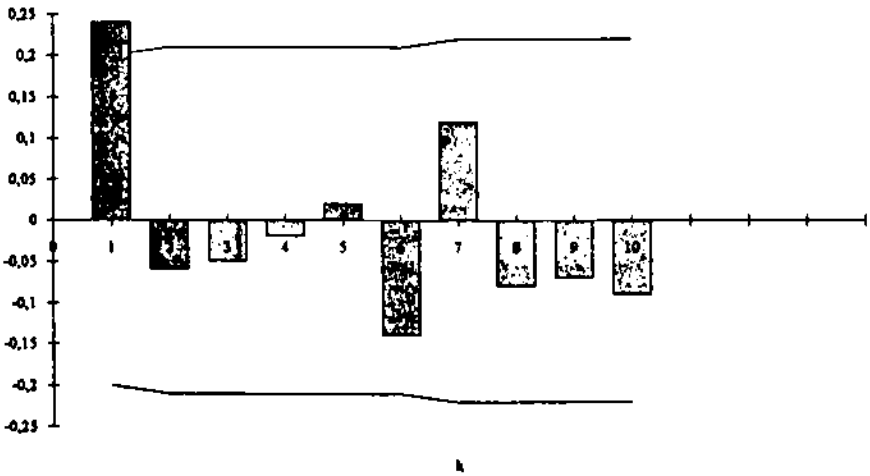


8-12. ábra Az AR(1) folyamat korrelogramja

a.) $\phi_1 = 0{,}5$ és b.) $\phi_1 = -0{,}5$ esetén

Ha a modell azonosítás során azt tapasztaljuk, hogy csak az első becsült autokorrelációs együttható (r_1) értéke szignifikáns, akkor feltételezhető, hogy elméleti idősorunk MA(1) folyamatot ír le. Ilyenkor MA(1) modell illesztésével célszerű próbálkozni. Ha azt tapasztaljuk, hogy az r_k értékek exponenciálisan, vagy csillapodó szinusz görbe szerint csökkennek, AR(1) modell illesztése tűnik megfelelőnek.

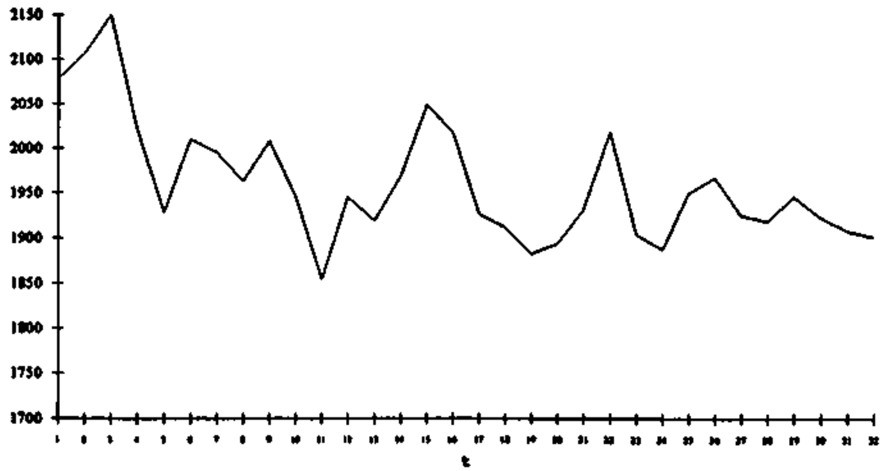
Nézzünk az elmondottakra példát! A mozilátogatások számának idősora alapján képzett első differenciák korrelogramjában (lásd! 8-10. ábra.) az első becsült autokorrelációs együttható (r_1) a legnagyobb értékű, így ez egy MA(1) típusú modellre "hasonlít" legjobban (bár látszik, hogy a szokásos 5 %-os szignifikancia szinten ez sem szignifikáns.) Az MA(1) típusú modell illusztrálása érdekében növeljük meg az 5 %-os szignifikancia szintet 10 %-ra! A konfidencia sáv így közelebb kerül az x tengelyhez és így az első becsült autokorrelációs együttható (r_1) szignifikánsan különbözik a nullától (8-13.ábra).



8-13. ábra A mozilátogatások transzformált idősorának korrelogramja módosított szignifikancia szinten

A 8-13. ábra alapján - mivel csak az r_1 szignifikáns, - MA(1) modell illesztésével célszerű próbálkoznunk.

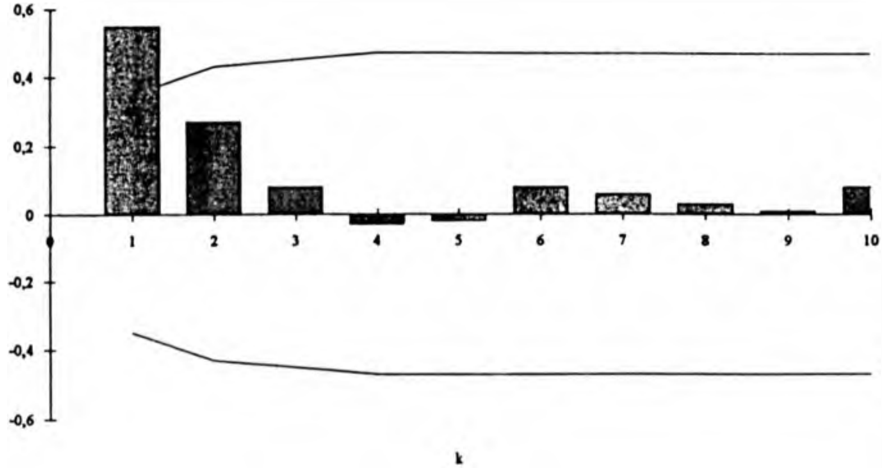
Másik példánkban a szarvasmarha állomány (ezer db) alakulását vizsgáljuk 1956 és 1987 között. A 8-14. ábra alapján megállapítható, hogy a trendhatást nem tartalmazó, stacionáriusnak tekinthető idősor, az $\bar{y}$=1.924 ezer db-nak megfelelő átlagos érték körül ingadozik.



8-14. ábra A szarvasmarha állomány (ezer darab) idősora 1956-1987.

(t= 1, 2, ..., 32)

Az idősor korrelogramja (8-15.ábra) az r_k értékek megközelítően exponenciális csökkenése alapján, AR(1) típusú folyamatra utal. Így az AR(1) típusú $Y_t = \phi_1 Y_{t-1} + \varepsilon_t$ modell illeszkedésével próbálkozhatunk.



8-15. ábra A szarvasmarha állomány idősorának korrelogramja

8.6.4 A modellek paraméterbecslése és ellenőrzése

Az illesztendő modellek ismeretlen ϕ és Θ paramétereinek becslésére a kidolgozott számítógépes programok többnyire a legkisebb négyzetek elvén alapuló iteratív eljárásokat alkalmazzák, egyéb speciális módszerek mellett (pl. backcasting módszer). Az iteratív eljárások a paraméterek kezdő értékéről indulva - amelyeket mi adhatunk meg, vagy amelyekről a program automatikusan indul - iterációként módosítja a becsléseket. Az iterációs lépéseket az algoritmus addig végzi, amíg paraméterbecslésünk különböző feltételeknek eleget nem tesz. (pl. $\sum e_t^2$ reziduális négyzetösszeg minimális csökkenése, a paraméter becslés minimális változása iterációnként.)

A mozilátogatások számából képzett első differenciák idősorát modellezzük, MA(1) típusú modellel.

Az MA(1) modell általános formája

$$Y_t = \varepsilon_t + \Theta_1 \varepsilon_{t-1}$$

melyet a $\Delta Y_t = Y_t - Y_{t-1}$ differenciákra felírva,

$$\Delta Y_t = Y_t - Y_{t-1} = \varepsilon_t + \Theta_1 \varepsilon_{t-1}$$

Átrendezés után, az

$$Y_t = Y_{t-1} + \varepsilon_t + \Theta_1 \varepsilon_{t-1}$$

modellt nyerjük, mely d = 1, q = 1 és p = 0 paraméterekkel rendelkező ARIMA (0,1,1) modell.

Elvégezve a paraméterbecslést, modellünk az alábbi:

$$Y_t = Y_{t-1} - 2.711 + \varepsilon_t + 0{,}3432\varepsilon_{t-1}$$

Eszerint a mozilátogatások száma egy adott időszakban a következőképpen alakul: a megelőző időszakban realizálódott értéknél (Y_{t-1}-nél) kevesebb az átlagos csökkenést jelző konstans értékével (kb. 2,7 millióval), ezt módosítja az ε_t véletlen hatás. Ezenkívül az előző időszaki véletlen hatás (ε_{t-1}) a jelenben is hat, $\hat{\Theta}_1$ = 0,3432 paraméterrel. (Természetesen a $\hat{\Theta}_1$ paraméter csak a megemelt szignifikancia szinten különbözik szignifikánsan a nullától.)

A modell ellenőrzésére a szokásos eljárások alkalmazhatók:

- a becsült paraméterek standard hiba számítása és szignifikancia vizsgálata (pl. t-próbával),
- az e_t tapasztalati reziduumok alapján az ε_t véletlen változók véletlen jellegének vizsgálata.

Mindezek mellett speciális tesztelési eljárások alkalmazására is sor kerül, amelyeket a számítógépes programok is tartalmaznak.

Amennyiben a választott és számszerűsített modellünk megfelel mindazon feltételeknek, melyekkel az illesztett modell "jóságát" ellenőrizhetjük, a modell felhasználható elemzésre és a tulajdonképpeni legfontosabb felhasználási területére, az **előrejelzések készítésére**. Ha modellünk nem felel meg a fenti feltételeknek (nem szignifikánsak a paraméterei, vagy az ε_t idősora nem véletlenszerűen alakul) a modellazonosítás fázisától újra indulva, más modelltípusok alkalmazásával próbálkozhatunk.

Becsült modelljeink az előrejelzett $\hat{Y}_t$ (t=n+1, n+2,...) értékek megállapításához egy-egy **rekurzív előrejelzési formulát** szolgáltatnak, amelyekhez különböző valószínűségi szinteken **konfidencia intervallumok** is rendelhetők. Az előrejelzések folyamatosan készíthetők úgy, hogy az új megfigyelt adatokkal - a modellszerkesztés és az előrejelzések ismételt elvégzése nélkül - az előrejelzések javíthatók, egy egyszerű korrekciós számítás alapján. Ezt az előrejelzések **korszerűsítésének** nevezik.

Szemléltetésül nézzük meg, például az AR(1) modell alapján milyen rekurzív előrejelzési formula adódik! Az y_t (t=1,2,....,n) idősorra felírt AR(1) modell általános alakja:

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t$$

Felírva ezt az idősor n-edik időszaki értékére:

$$Y_n = \phi_1 Y_{n-1} + \varepsilon_t$$

Előrejelzést készítve az idősor (n+1)-edik értékére,

$$\hat{Y}_{n+1} = \phi_1 Y_n \qquad (8\text{-}54)$$

az (n+2)-edik értékére

$$\hat{Y}_{n+2} = \phi_1 \hat{Y}_{n+1} = \phi_1 \left(\phi_1 \hat{Y}_n\right) = \phi_1{}^2 Y_n \qquad (8\text{-}55)$$

az (n+3)-dik értékére,

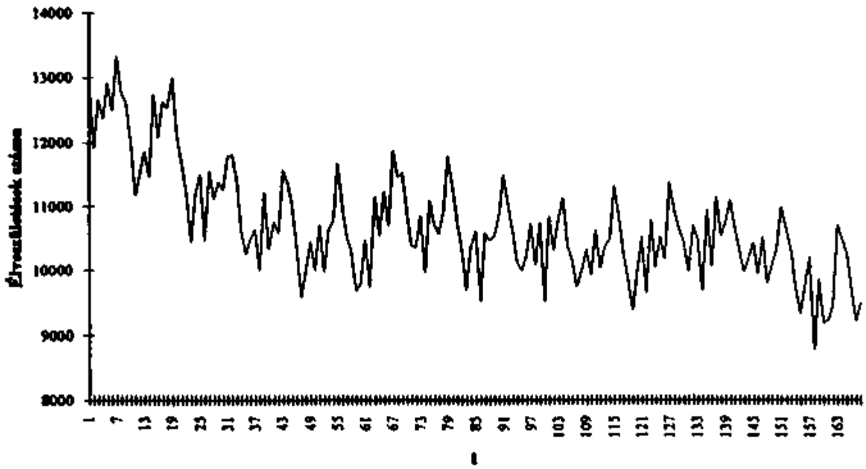
$$\hat{Y}_{n+3} = \phi_1 \hat{Y}_{n+2} = \phi_1^3 Y_n \qquad (8\text{-}56)$$

A számítógépes programok az illesztett modell alapján az előrejelzett értékeket is szolgáltatják, azok standard hibájával együtt.

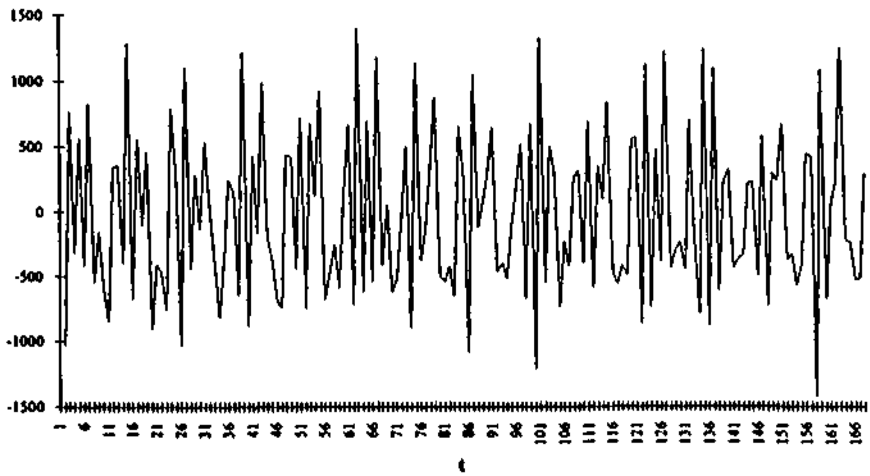
8.6.5 Szezonális ARIMA modellek

A szezonális ARIMA modellek - az idősorban érvényesülő kettős függőségi rendszer miatt - jóval bonyolultabbak. Lényegében két ARIMA modell összevonásából nyerjük az általánosan használt ún. **multiplikatív szezonális ARIMA modellt**, ARIMA(p,d,q)(P,D,Q)$_S$ dimenziókkal (ahol P,D,Q a modell szezonális részének dimenzióit, s pedig a szezonok számát jelöli.)

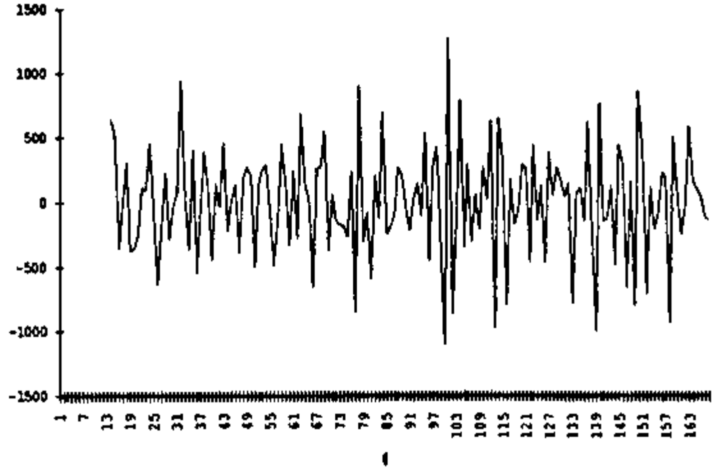
A szezonális ingadozást is tartalmazó idősorok modellezésére nézzük meg az élveszületések számának alakulását 1980. januártól 1993. decemberéig. (Forrás: Statisztikai Havi közlemények). A megfigyelt 102 tagú idősor grafikus ábráján (8-16. ábra) látható, hogy az idősor nem stacionárius, ezért első fokú differenciaképzést végeztünk (y_t - y_{t-1}) (8-17.ábra). Ez az idősor már egy konstans szint körül ingadozik, de a szórása nem állandó, a szezonalitás miatt. Ezt egy újabb differenciaképzéssel próbáljuk kiszűrni, azaz az (y_t - y_{t-12}) szezonális differenciákat is képezzük. A kétszeres (szezonális és nem szezonális) differenciaképzéssel nyert idősorunk (y_t-y_{t-12}) - (y_{t-1} -y_{t-13}) már stacionáriusnak tekinthető (8-18.ábra).



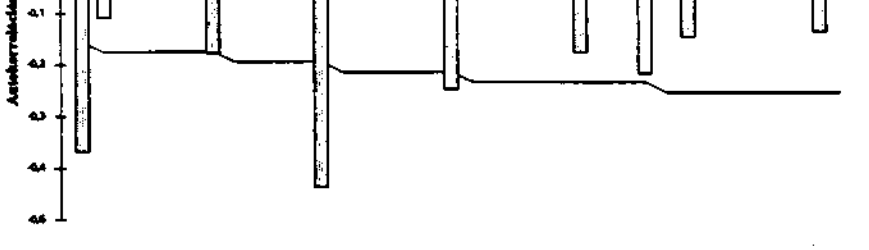
8-16. ábra Az élveszületések számának alakulása 1980-1993.



8-17. ábra Az élveszületések számának első differencia sora



8-18. ábra Az élveszületések számának kétszeres (szezonális és nem szezonális) differencia sora

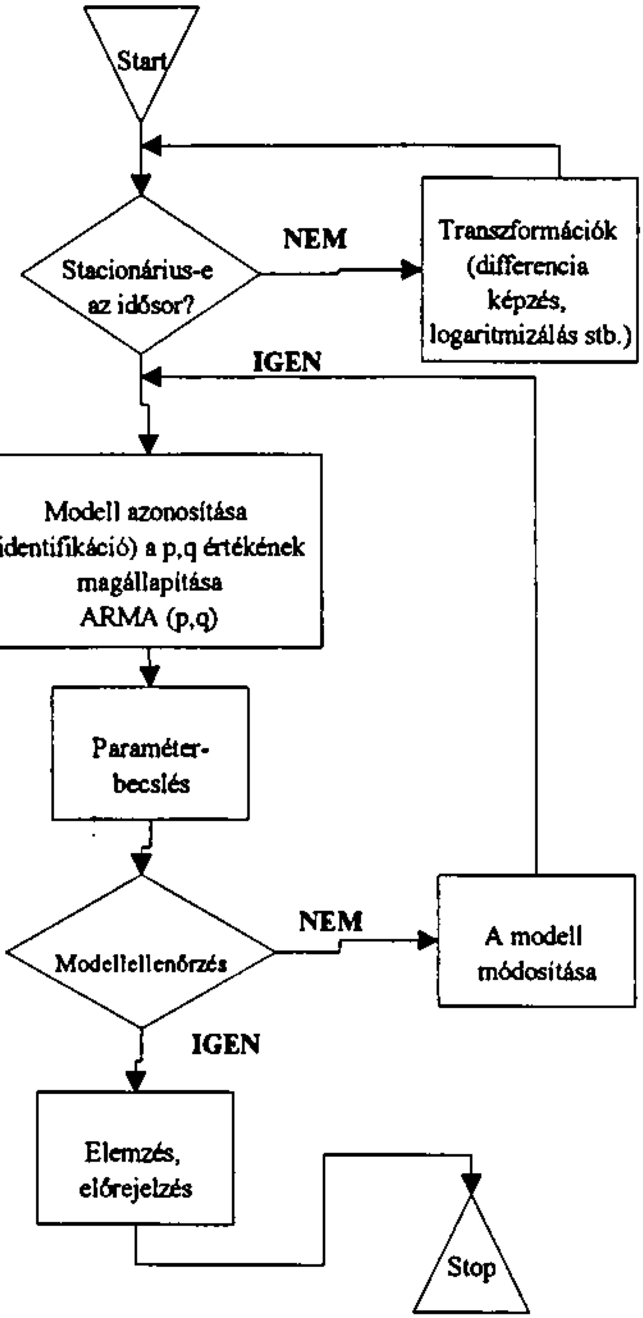


8-19. ábra Az élveszületések száma kétszeres differencia sorának korrelogramja

A modell azonosításhoz a stacionárius sor korrelogramját mutatja a (8-19. ábra) A k = 1 , k = 12 , k=18 és k=26 késleltetéssel számított r_k autokorrelációs együttható szignifikáns, mely "kétszeres" mozgóátlag folyamatra utal, azaz olyan MA modell illesztése látszik célravezetőnek, melyben a szezonalitás figyelembevételére is becsülünk mozgóátlag paramétert.

A modell felírásától továbbiakban eltekintünk, mivel célunk csak a sztochasztikus modellezés alapelveinek bemutatása volt.

Összefoglalásul közöljük az idősori modellkészítés folyamatábráját.



8-20. ábra Az idősori modellkészítés folyamatábrája

FÜGGELÉK

Az autokorrelációs és parciális autokorrelációs együtthatók értékeinek alakulása különböző típusú ARIMA modellekben

Modelltípus ARIMA (p,d,q)	Autokorrelációs együtthatók (ρ_k)	Parciális autokorrelációs együtthatók (ϕ_{kk})
(1, d, 0) AR(1)	Exponenciálisan csökken, ha $\rho_1 > 0$, csillapodó szinusz görbe szerint csökken, ha $\rho_1 < 0$	ϕ_{11} $k = 1$ 0 $k > 1$
(2, d, 0) AR (2)	Exponenciálisan és/vagy csillapodó szinusz görbe szerint csökken	ϕ_{11} $k = 1$ ϕ_{22} $k = 2$ 0 $k > 2$
(0, d, 1) MA(1), ha d = 0 vagy IMA (d, 1)	ρ_1 ha $k = 1$ 0 ha $k > 1$	Exponenciálisan, vagy csillapodó szinusz görbe szerint csökken
(0, d, 2) MA(2), ha d = 0 vagy IMA (d, 2)	ρ_1 ha $k = 1$ ρ_2 ha $k = 2$ 0 ha $k > 2$	Exponenciálisan és/vagy csillapodó szinusz görbe szerint csökken
(1, d, 1) ARMA (1, 1) ha d = 0, vagy ARIMA (1, d, 1)	Exponenciálisan, vagy csillapodó szinusz görbe szerint csökken a második értéktől kezdődően	Exponenciálisan, vagy csillapodó szinusz görbe szerint csökken a második értéktől kezdődően
(1, d, 2) ARMA (1, 2) ha d = 0 vagy ARIMA (1, d, 2)	$\rho_0 = 1$ és ρ_1 után exponenciálisan csökken	Exponenciálisan és/vagy csillapodó szinusz görbe szerint csökken
(2, d, 1) ARMA (2, 1) ha d = 0 vagy ARIMA (2, d, 1)	Exponenciálisan és/ vagy csillapodó szinusz görbe szerint csökken	$\phi_{11} = \rho_1$ és ϕ_{22} után exponenciálisan csökken
(2, d, 2) ARMA (2, 2) ha d = 0 vagy ARIMA (2, d, 2)	$\rho_0 = 1$ után exponenciálisan csökken	$\phi_{11} = \rho_1$ után exponenciálisan és/vagy csillapodó szinusz görbe szerint csökken

IRODALOM

1] ABRAHAM,B. - LEDOLTER,J.: *Statistical Methods for Forecasting,* John Wiley and Sons, Inc. New York, 1983. 445 old.

[2] BÁN ERVINNÉ - NÁDUDVARI ZOLTÁN: *A gazdálkodás statisztikai jellemzői az elektronikai termékeket gyártó vállalatoknál,* Statisztikai Szemle 1990. évi 2. sz. 145-155. old.

[3] BOOT,J.C.G. - COX,E.B.: *Statistical Analysis for Managerial Decision,,* Mc Graw-Hill Company, New York, 1970. 641 old.

[4] BOX,G.E.P. - COX,D.R.: *An analysis of transformations,* Journal Of the Royal Statistical Society, Series B, 26. 1964. évi 2. sz. 211-243. old.

[5] BOX,G.E.P. - JENKINS,G.M.: *Time series analysis. Forecasting and contol,* Holden Day, San Francisco, etc. 1970.

[6] BOX,G.E.P.-PIERCE,A.: *Distribution of residual autocorrelations in autoregressive moving average time series models,* Journal of the American Association, 65. 1509-1526. old.

[7] BÜNING,H. - TRENKLER,G.: *Nichtparametrische statistische Methoden,* Walter de Gruyter, Berlin, 1978. 435 old.

[8] CANAVOS,G.C.: *Applied Probability and Statistical Methods,* Little, Brown and Company, Boston Toronto 1984. 608 old.

[9] CHATTERJE,S.- PRICE,B.: *Regression Analysis by Example,* John Wiley, New York 1977. 228 old.

[10]DENKINGER GÉZA: *Valószínüségszámítás,* Tankönyvkiadó, Budapest, 1982. 278 old.

[11]ÉLTETŐ ÖDÖN - MESZÉNA GYÖRGY - ZIERMANN MARGIT: *Sztochasztikus módszerek és modellek,* Közgazdasági és Jogi Könyvkiadó Budapest, 1982. 420 old.

[12]ÉLTETŐ ÖDÖN: *Az ELAR minta és az 1984. évi mikrocenzus mintájának kiválasztási eljárása,* Statisztikai módszertani füzetek. KSH, Budapest, 1987. 59 old.

[13]ÉLTETŐ ÖDÖN: *Mintavétel véges alapsokaságból,* Bolyai János Matematikai Társulat, Budapest, 1970. 154 old.

[14]ENNS,P.G.: *Business Statistics,* Richard D. Irwin, Inc. Homewood, Illinois, 1985. 777 old.

[15]FRESCHL GYÖRGY: *Bevezetés az idősori módszerek gyakorlatába,* Statisztikai módszertani füzetek. KSH, Budapest, 1982.

[16]FÜSTÖS LÁSZLÓ - MESZÉNA GYÖRGY - SIMONNÉ MOSOLYGÓ NÓRA: *A sokváltozós adatelemzés statisztikai módszerei,* Akadémiai Kiadó, Budapest, 1986. 526 old.

[17] GIRAUD,R. - CHAIX,N.: *Econométrie*, Presses Universitaires de France, Paris, 1989. 304 old.
[18] GOODMAN,L.A. -KRUSKAL,W.H.: *Measures of associations for cross classifications*, Journal of the American Statistical Associations 1954. 12. 732-764. old.
[19] GUITTON,H.: *Statistique*, Dalloz, Paris, 1971. 351 old.
[20] GUJARATI, D.: *Basic Econometrics*, McGraw-Hill Book Company, New York, 1978. 462 old.
[21] HAJDU OTTÓ - HERMAN SÁNDOR - PINTÉR JÓZSEF -RÉDEY KATALIN: *Ökonometriai alapvetés*, Tankönyvkiadó, Budapest, 1987. 157 old.
[22] HEINCZINGER MÁRIA: *A multikollinearitás felismerése, mérése és kiszűrése*, Statisztikai Szemle 1983. évi 7.sz. 741-762. old.
[23] HOÓS JÁNOS DR.: *A magyar statisztika megújításának feladatai*, Statisztikai Szemle 1990. évi 6. sz. 437-453. old.
[24] HOÓZ ISTVÁN DR.: *Demográfia*, Tankönyvkiadó, Budapest,1987. 319 old.
[25] HORVATH,T.: *Basic Statistics for Behavioral Sciences*, Little, Brown and Company,Boston,1985. 456 old.
[26] HUANG,D.S.: *Regression and Econometric Methods*, John Wiley and Sons, Inc., New York, 1970. 273 old.
[27] HULYÁK KATALIN: *Idősorok sztochasztikus modelljei*, Ökonometriai füzetek. KSH, Budapest, 1977. 144 old.
[28] HUNYADI LÁSZLÓ - MUNDRUCZÓ GYÖRGY - VITA LÁSZLÓ: *Statisztika I.-II.* AULA, Budapest, 1991-1992. 430. és 402 old.
[29] KERÉKGYÁRTÓ GYÖRGYNÉ - MUNDRUCZÓ GYÖRGY: *Statisztikai módszerek a gazdasági elemzésben*, Közgazdasági és Jogi Könyvkiadó, Budapest 1987. 333 old.
[30] KORINEK LÁSZLÓ - PINTÉR JÓZSEF - SZŰCS ANDRÁSNÉ: *Statisztika I. rész*, Tankönyvkiadó, Budapest, 1976. 249 old.
[31] KŐRÖSI GÁBOR - MÁTYÁS LÁSZLÓ - SZÉKELY ISTVÁN: *Gyakorlati ökonometria*, Közgazdasági és Jogi Könyvkiadó, Budapest, 1990. 481 old
[32] KÖVES PÁL - PÁRNICZKI GÁBOR: *Általános statisztika I-II.*. Tankönyvkiadó,Budapest, 1981. 363 és 387 old.
[33] KÖVES PÁL: *Indexelmélet és közgazdasági valóság*, Akadémiai Kiadó, Budapest, 1981. 212 old.
[34] LJUNG,G.M. - BOX,G.E.P.: *On a measure of lach of fit in time series models*, Biometrika, 65. 297-303. old.
[35] MALINVAUD,E.: *Az ökonometria statisztikai módszerei*, Közgazdasági és Jogi Könyvkiadó, Budapest, 1974. 804 old.
[36] MESZÉNA GYÖRGY - ZIERMANN MARGIT: *Valószínűség-elmélet és matematikai statisztika*, Közgazdasági és Jogi Könyvkiadó, Budapest 1981. 554 old.

[37]MUNDRUCZÓ GYÖRGY: *A minőségi ismérvek közötti kapcsolatok vizsgálata I. II.*, Statisztikai Szemle 1982. évi 6. sz. 635-648. old. és 1982. évi 7. sz. 730-737. old.

[38]MUNDRUCZÓ GYÖRGY: *Alkalmazott regressziószámítás*, Akadémiai Kiadó, Budapest, 1981. 257 old.

[39]NEWBOLD,P.: *Statistics for Business and Economics*, Prentice-Hall, Inc. Englewood Cliffs New Jersey 1984. 895 old.

[40]NYITRAI FERENCNÉ - RÉDEY KATALIN: *Statisztika III.*, Tankönyvkiadó, Budapest, 1976. 195 old.

[41]PINTÉR JÓZSEF - RÉDEY KATALIN - SZŰCS ANDRÁSNÉ: *Statisztika II. rész*, Tankönyvkiadó, Budapest 1978. 258 old.

[42]PINTÉR JÓZSEF: *A heteroszkedaszticitás diagnosztizálása*, Statisztikai Szemle 1991. évi 1. sz. 16-36. old.

[43]PLANE,D.R. - OPPERMANN, E.B.: *Business and Economic Statistics*, Business Publications Inc. Plano(Texas), 1986. 711 old.

[44]RÉNYI ALFRÉD: *Valószínűségszámítás*, Tankönyvkiadó, Budapest 1954.

[45]SCHLITTGEN,R.: *Einführung in die Statistik (Analyse und Modellierung von Daten)*, Oldenbourg Verlag, München, 1991. 459 old.

[46]SZILÁGYI GYÖRGY DR.: *Fogyasztási kereslet vizsgálata nemzetközi összehasonlításban*, Statisztikai Szemle 1990. évi 2. sz. 123-132. old.

[47]TUSNÁDY - ZIERMANN: *Idősorok analízise*, Műszaki Könyvkiadó, Budapest, 1986.

[48]VAN MATRE,J.G. - GILBREATH,G.H.: *Statistics for Business and Economics*, Business Publications, Inc. Plano, Texas, 1983. 636 old.

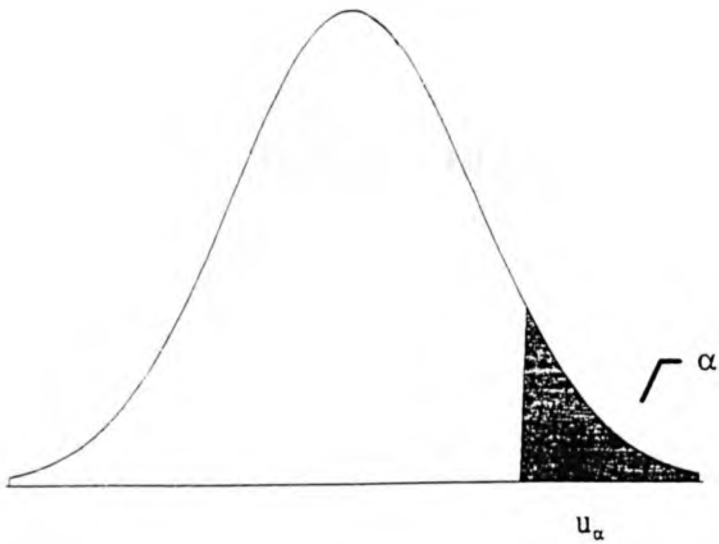
[49]VINCZE ISTVÁN: *Matematikai statisztika ipari alkalmazásokkal*, Műszaki Könyvkiadó, Budapest, 1975. 352 old.

[50]WONNACOTT,R.J. - WONNACOTT,T.H.: *Econometrics*, John Wiley and Sons, Inc., New York, 1979. 580 old.

[51]YULE,G.U. - KENDALL,M.G.: *Bevezetés a statisztika elméletébe*, Közgazdasági és Jogi Könyvkiadó, Budapest, 1964. 700 old.

TÁBLÁZATOK

Standard normális eloszlás kritikus értékei előre adott szignifikancia-szinten



Példa

Keressük a α=0,01 azaz 1 %-os szignifikancia-szinthez tartozó kritikus értéket egyoldalú ellenhipotézis esetén (H_1: $\mu>m_0$) :

$$u_{0,01} = 2,326$$

Keressük a α=0,01 azaz 1 %-os szignifikancia-szinthez tartozó kritikus értéket egyoldalú ellenhipotézis esetén (H_1: $\mu<m_0$) :

$$u_{0,99} = -u_{0,01} = -2,326$$

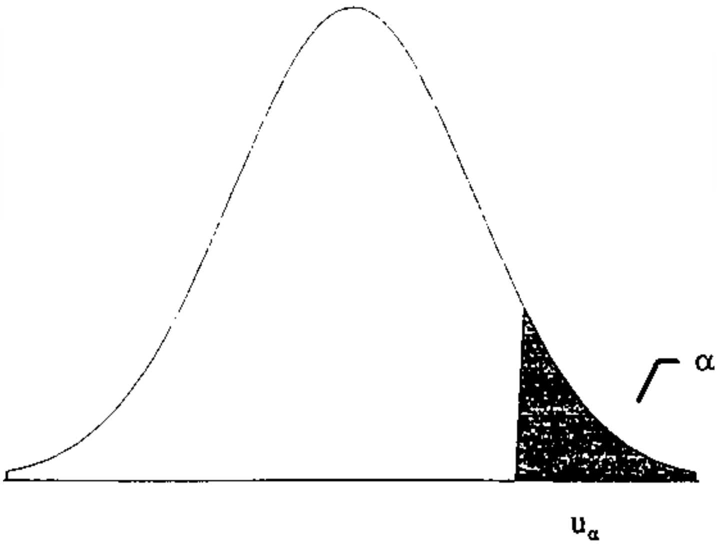
Keressük a α=0,01 azaz 1 %-os szignifikancia-szinthez tartozó kritikus értékeket kétoldalú ellenhipotézis esetén (H_1: $\mu\neq m_0$) :

$$u_{0,005} = 2,576$$
$$u_{0,995} = -u_{0,005} = -2,576$$

Standard normális eloszlás kritikus értékei előre adott szignifikancia-szinten

α	**0,1000**	**0,0500**	**0,0250**	**0,0225**	**0,0100**	**0,0050**
u	1,282	1,645	1,960	2,000	2,326	2,576

Standard normális eloszlásfüggvény szignifikancia-értékei



Példa

Keressük az u=1,54-hez tartozó szignifikancia-értéket:

$$\Pr\{1{,}54 \le \xi\} = 0{,}062$$

Keressük az u=-1,54-hez tartozó szignifikancia-értéket:

$$\Pr\{-1{,}54 \le \xi\} = 1 - \Pr\{1{,}54 \le \xi\} = 1 - 0{,}062 = 0{,}938$$

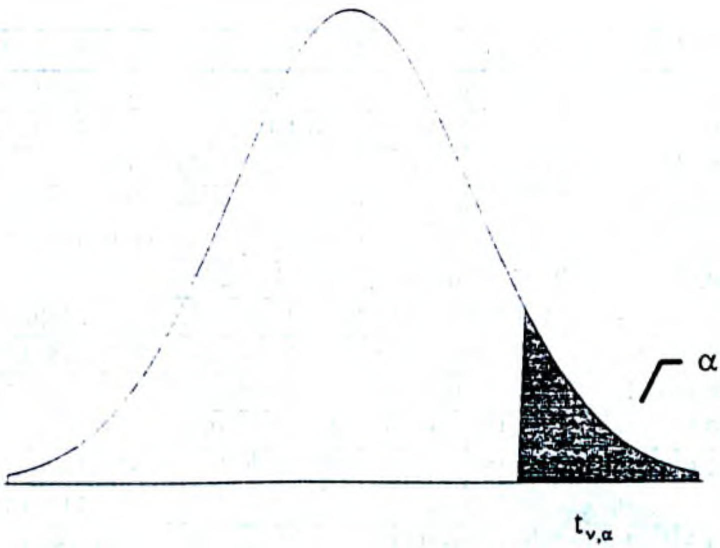
Keressük a (-1,8;1,2) intervallumba esés valószínűségét:

$$\Pr\{-1{,}8 \le \xi \le 1{,}2\} = \Pr\{-1{,}8 \le \xi\} - \Pr\{1{,}2 \le \xi\} = 1 - \Pr\{1{,}8 \le \xi\} - \Pr\{1{,}2 \le \xi\} =$$
$$= 1 - 0{,}036 - 0{,}115 = 0{,}849$$

Standard normális eloszlásfüggvény szignifikancia-értékei

u	0	1	2	3	4	5	6	7	8	9
0,0	0.500	0.496	0,492	0.488	0.484	0,480	0.476	0,472	0,468	0.464
0,1	0,460	0,456	0,452	0,448	0.444	0,440	0.436	0,433	0,429	0,425
0,2	0.421	0,417	0.413	0.409	0,405	0,401	0,397	0.394	0.390	0.386
0,3	0.382	0,378	0.374	0,371	0.367	0,363	0,359	0,356	0,352	0,348
0,4	0.345	0.341	0,337	0,334	0.330	0,326	0.323	0.319	0,316	0,312
0,5	0.309	0,305	0.302	0.298	0,295	0,291	0,288	0.284	0.281	0,278
0,6	0,274	0,271	0,268	0,264	0,261	0,258	0,255	0,251	0,248	0,245
0,7	0,242	0.239	0,236	0,233	0,230	0,227	0,224	0,221	0,218	0,215
0,8	0,212	0.209	0.206	0.203	0,200	0,198	0,195	0,192	0.189	0,187
0,9	0.184	0,181	0,179	0.176	0.174	0.171	0.169	0.166	0,164	0.161
1,0	0.159	0.156	0.154	0,152	0,149	0.147	0.145	0.142	0.140	0.138
1,1	0.136	0,133	0.131	0,129	0.127	0,125	0,123	0.121	0.119	0,117
1,2	0,115	0,113	0.111	0,109	0,107	0.106	0,104	0,102	0,100	0,099
1,3	0.097	0,095	0,093	0,092	0,090	0,089	0,087	0,085	0.084	0,082
1,4	0.081	0,079	0.078	0.076	0,075	0,074	0,072	0,071	0,069	0,068
1,5	0,067	0,066	0,064	0,063	0,062	0,061	0,059	0,058	0,057	0,056
1,6	0,055	0,054	0,053	0,052	0,051	0,049	0,048	0.047	0,046	0,046
1,7	0,045	0.044	0,043	0,042	0,041	0,040	0,039	0,038	0,038	0,037
1,8	0,036	0.035	0,034	0,034	0,033	0,032	0,031	0,031	0.030	0,029
1,9	0,029	0,028	0,027	0,027	0,026	0,026	0,025	0,024	0,024	0,023
2,0	0,023	0,022	0,022	0,021	0,021	0,020	0,020	0,019	0.019	0,018
2,1	0,018	0,017	0,017	0.017	0,016	0.016	0,015	0,015	0.015	0,014
2,2	0,014	0.014	0,013	0.013	0,013	0,012	0,012	0,012	0.011	0.011
2,3	0,011	0.010	0,010	0,010	0,010	0,009	0,009	0,009	0.009	0,008
2,4	0,008	0.008	0,008	0.008	0,007	0,007	0,007	0.007	0.007	0.006
2,5	0,006	0,006	0,006	0,006	0,006	0,005	0,005	0,005	0,005	0,005
2,6	0,005	0,005	0,004	0,004	0,004	0,004	0.004	0,004	0,004	0,004
2,7	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003	0,003
2,8	0,003	0,002	0,002	0,002	0,002	0,002	0,002	0,002	0.002	0,002
2,9	0,002	0.002	0,002	0,002	0,002	0,002	0.002	0,001	0.001	0,001
3,0	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001

A Student-féle t-eloszlás kritikus értékei előre adott szignifikancia-szinten



Példa

Keressük a ν=5 szabadságfokú t-eloszlás α=0,05 azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékét egyoldalú ellenhipotézis esetén (H_1: $\mu > m_0$) :

$$t_{5;0,05} = 2,015$$

Keressük a ν=5 szabadságfokú t-eloszlás α=0,05 azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékét egyoldalú ellenhipotézis esetén (H_1: $\mu < m_0$) :

$$t_{5;0,95} = -t_{5;0,05} = -2,015$$

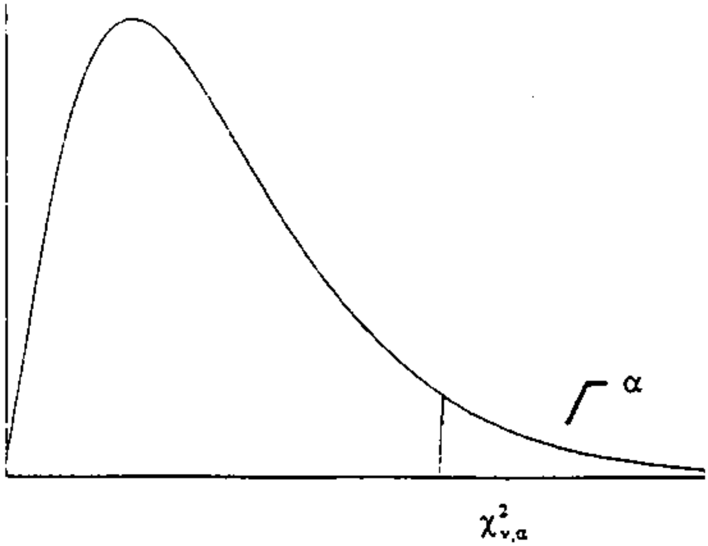
Keressük a ν=5 szabadságfokú t-eloszlás α=0,05 azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékeit kétoldalú ellenhipotézis esetén (H_1: $\mu \neq m_0$) :

$$t_{5;0,025} = 2,571$$
$$t_{5;0,975} = -t_{5;0,025} = -2,571$$

A Student-féle t-eloszlás kritikus értékei előre adott szignifikancia-szinten

Szabadság-fok	α				
	0,100	0,050	0,025	0,010	0,005
1	3,078	6,314	12,706	31,821	63,656
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
50	1,299	1,676	2,009	2,403	2,678
60	1,296	1,671	2,000	2,390	2,660
70	1,294	1,667	1,994	2,381	2,648
80	1,292	1,664	1,990	2,374	2,639
90	1,291	1,662	1,987	2,368	2,632
100	1,290	1,660	1,984	2,364	2,626
150	1,287	1,655	1,976	2,351	2,609
200	1,286	1,653	1,972	2,345	2,601

A χ^2-eloszlás kritikus értékei előre adott szignifikancia-szinten



Példa

Keressük a $v=3$ szabadságfokú χ^2-eloszlás $\alpha=0{,}05$ azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékét egyoldalú ellenhipotézis esetén (H_1: $\sigma > \sigma_0$) :

$$\chi^2_{3;0,05} = 7{,}815$$

Keressük a $v=3$ szabadságfokú χ^2-eloszlás $\alpha=0{,}05$ azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékét egyoldalú ellenhipotézis esetén (H_1: $\sigma < \sigma_0$) :

$$\chi^2_{3;0,95} = 0{,}352$$

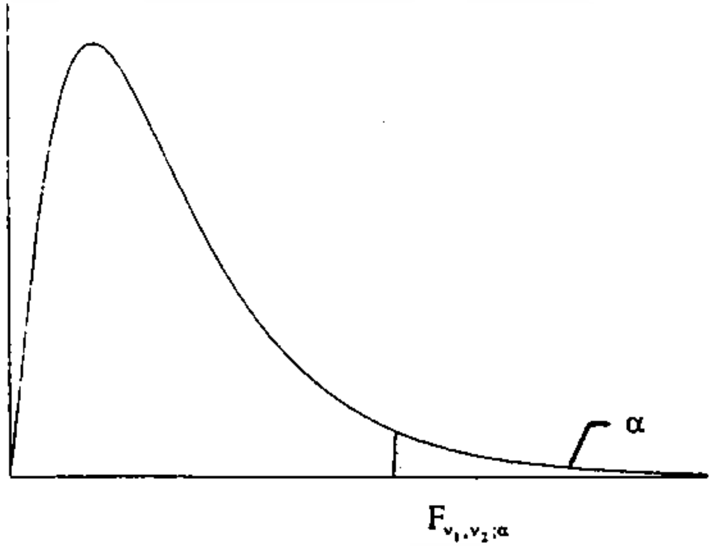
Keressük a $v=3$ szabadságfokú χ^2-eloszlás $\alpha=0{,}05$ azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékeit kétoldalú ellenhipotézis esetén (H_1: $\sigma \neq \sigma_0$) :

$$\chi^2_{3;0,025} = 9{,}348$$
$$\chi^2_{3;0,975} = 0{,}216$$

A χ^2-eloszlás kritikus értékei előre adott szignifikancia-szinten

Szabadság-fok	α							
	0,995	0,990	0,975	0,950	0,050	0,025	0,010	0,005
1	0,000	0,000	0,001	0,004	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	124,342	129,561	135,807	140,170
150	109,142	112,668	117,985	122,692	179,581	185,800	193,207	198,360
200	152,241	156,432	162,728	168,279	233,994	241,058	249,445	255,264

Az F-eloszlás kritikus értékei előre adott szignifikancia-szinten



Példa

Keressük a ν_1=20; ν_2=30 szabadságfokú F-eloszlás α=0,05 azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékét egyoldalú ellenhipotézis esetén ($H_1: \sigma_1^2 > \sigma_2^2$):

$$F_{20,30;0,05} = 1,93$$

Keressük a ν_1=20; ν_2=30 szabadságfokú F-eloszlás α=0,05 azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékét egyoldalú ellenhipotézis esetén ($H_1: \sigma_1^2 < \sigma_2^2$):

$$F_{20,30,0,95} = \frac{1}{F_{30,20;0,05}} = \frac{1}{2,04} = 0,49$$

Keressük a ν_1=20; ν_2=30 szabadságfokú F-eloszlás α=0,05 azaz 5 %-os szignifikancia-szinthez tartozó kritikus értékét kétoldalú ellenhipotézis esetén ($H_1: \sigma_1^2 \neq \sigma_2^2$):

$$F_{20,30;0,025} = 2,20$$

$$F_{20,30;0,975} = \frac{1}{F_{30,20;0,025}} = \frac{1}{2,35} = 0,426$$

Az F-eloszlás kritikus értékei 5 %-os szignifikancia-szinten

Nev.	Számláló szabadságfoka																			
szf.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	100
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	244.7	245.4	245.9	248.0	250.1	251.1	251.8	253.0
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.66	8.62	8.59	8.58	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.80	5.75	5.72	5.70	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.56	4.50	4.46	4.44	4.41
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.87	3.81	3.77	3.75	3.71
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.44	3.38	3.34	3.32	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.15	3.08	3.04	3.02	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.94	2.86	2.83	2.80	2.76
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.77	2.70	2.66	2.64	2.59
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.65	2.57	2.53	2.51	2.46
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.54	2.47	2.43	2.40	2.35
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.46	2.38	2.34	2.31	2.26
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.39	2.31	2.27	2.24	2.19
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.33	2.25	2.20	2.18	2.12
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.28	2.19	2.15	2.12	2.07
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.23	2.15	2.10	2.08	2.02
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.19	2.11	2.06	2.04	1.98
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.16	2.07	2.03	2.00	1.94
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.12	2.04	1.99	1.97	1.91
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.10	2.01	1.96	1.94	1.88
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.07	1.98	1.94	1.91	1.85
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20	2.18	2.15	2.13	2.05	1.96	1.91	1.88	1.82
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.03	1.94	1.89	1.86	1.80
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.01	1.92	1.87	1.84	1.78
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	1.99	1.90	1.85	1.82	1.76
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13	2.10	2.08	2.06	1.97	1.88	1.84	1.81	1.74
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	1.96	1.87	1.82	1.79	1.73
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.08	2.05	2.03	1.94	1.85	1.81	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.93	1.84	1.79	1.76	1.70
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.84	1.74	1.69	1.66	1.59
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.78	1.69	1.63	1.60	1.52
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.75	1.65	1.59	1.56	1.48
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.84	1.81	1.72	1.62	1.57	1.53	1.45
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88	1.84	1.82	1.79	1.70	1.60	1.54	1.51	1.43
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.90	1.86	1.83	1.80	1.78	1.69	1.59	1.53	1.49	1.41
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85	1.82	1.79	1.77	1.68	1.57	1.52	1.48	1.39
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.79	1.76	1.73	1.64	1.54	1.48	1.44	1.34
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80	1.77	1.74	1.72	1.62	1.52	1.46	1.41	1.32

Az F-eloszlás kritikus értékei 2,5 %-os szignifikancia-szinten

Nev. szf.	Számláló szabadságfoka 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	100
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,6	963,3	968,6	973,0	976,7	979,8	982,5	984,9	993,1	1001	1005	1008	1013
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,41	39,42	39,43	39,43	39,45	39,46	39,47	39,48	39,49
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,37	14,34	14,30	14,28	14,25	14,17	14,08	14,04	14,01	13,96
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,79	8,75	8,72	8,68	8,66	8,56	8,46	8,41	8,38	8,32
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,57	6,52	6,49	6,46	6,43	6,33	6,23	6,18	6,14	6,08
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,41	5,37	5,33	5,30	5,27	5,17	5,07	5,01	4,98	4,92
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,71	4,67	4,63	4,60	4,57	4,47	4,36	4,31	4,28	4,21
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,24	4,20	4,16	4,13	4,10	4,00	3,89	3,84	3,81	3,74
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,91	3,87	3,83	3,80	3,77	3,67	3,56	3,51	3,47	3,40
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,66	3,62	3,58	3,55	3,52	3,42	3,31	3,26	3,22	3,15
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,47	3,43	3,39	3,36	3,33	3,23	3,12	3,06	3,03	2,96
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,32	3,28	3,24	3,21	3,18	3,07	2,96	2,91	2,87	2,80
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,20	3,15	3,12	3,08	3,05	2,95	2,84	2,78	2,74	2,67
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,09	3,05	3,01	2,98	2,95	2,84	2,73	2,67	2,64	2,56
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	3,01	2,96	2,92	2,89	2,86	2,76	2,64	2,59	2,55	2,47
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,93	2,89	2,85	2,82	2,79	2,68	2,57	2,51	2,47	2,40
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,87	2,82	2,79	2,75	2,72	2,62	2,50	2,44	2,41	2,33
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,81	2,77	2,73	2,70	2,67	2,56	2,44	2,38	2,35	2,27
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,76	2,72	2,68	2,65	2,62	2,51	2,39	2,33	2,30	2,22
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,72	2,68	2,64	2,60	2,57	2,46	2,35	2,29	2,25	2,17
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,68	2,64	2,60	2,56	2,53	2,42	2,31	2,25	2,21	2,13
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,65	2,60	2,56	2,53	2,50	2,39	2,27	2,21	2,17	2,09
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,62	2,57	2,53	2,50	2,47	2,36	2,24	2,18	2,14	2,06
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,59	2,54	2,50	2,47	2,44	2,33	2,21	2,15	2,11	2,02
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,56	2,51	2,48	2,44	2,41	2,30	2,18	2,12	2,08	2,00
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,54	2,49	2,45	2,42	2,39	2,28	2,16	2,09	2,05	1,97
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,51	2,47	2,43	2,39	2,36	2,25	2,13	2,07	2,03	1,94
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,49	2,45	2,41	2,37	2,34	2,23	2,11	2,05	2,01	1,92
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,48	2,43	2,39	2,36	2,32	2,21	2,09	2,03	1,99	1,90
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,46	2,41	2,37	2,34	2,31	2,20	2,07	2,01	1,97	1,88
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,33	2,29	2,25	2,21	2,18	2,07	1,94	1,88	1,83	1,74
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32	2,26	2,22	2,18	2,14	2,11	1,99	1,87	1,80	1,75	1,66
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,22	2,17	2,13	2,09	2,06	1,94	1,82	1,74	1,70	1,60
70	5,25	3,89	3,31	2,97	2,75	2,59	2,47	2,38	2,30	2,24	2,18	2,14	2,10	2,06	2,03	1,91	1,78	1,71	1,66	1,56
80	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,35	2,28	2,21	2,16	2,11	2,07	2,03	2,00	1,88	1,75	1,68	1,63	1,53
90	5,20	3,84	3,26	2,93	2,71	2,55	2,43	2,34	2,26	2,19	2,14	2,09	2,05	2,02	1,98	1,86	1,73	1,66	1,61	1,50
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	2,12	2,08	2,04	2,00	1,97	1,85	1,71	1,64	1,59	1,48
150	5,13	3,78	3,20	2,87	2,65	2,49	2,37	2,28	2,20	2,13	2,08	2,03	1,99	1,95	1,92	1,80	1,67	1,59	1,54	1,42
200	5,10	3,76	3,18	2,85	2,63	2,47	2,35	2,26	2,18	2,11	2,06	2,01	1,97	1,93	1,90	1,78	1,64	1,56	1,51	1,39

Az F-eloszlás kritikus értékei 1 %-os szignifikancia-szinten

Nev.	Számláló szabadságfoka																		
szf.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	100
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6126	6143	6157	6209	6260	6286	6302	6334
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,41	99,42	99,42	99,43	99,43	99,45	99,47	99,48	99,48	99,49
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05	26,98	26,92	26,87	26,69	26,50	26,41	26,35	26,24
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,31	14,25	14,20	14,02	13,84	13,75	13,69	13,58
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,82	9,77	9,72	9,55	9,38	9,29	9,24	9,13
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,40	7,23	7,14	7,09	6,99
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,36	6,31	6,16	5,99	5,91	5,86	5,75
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,61	5,56	5,52	5,36	5,20	5,12	5,07	4,96
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,01	4,96	4,81	4,65	4,57	4,52	4,41
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,65	4,60	4,56	4,41	4,25	4,17	4,12	4,01
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,10	3,94	3,86	3,81	3,71
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,86	3,70	3,62	3,57	3,47
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,91	3,86	3,82	3,66	3,51	3,43	3,38	3,27
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,51	3,35	3,27	3,22	3,11
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,37	3,21	3,13	3,08	2,98
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,50	3,45	3,41	3,26	3,10	3,02	2,97	2,86
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,16	3,00	2,92	2,87	2,76
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,32	3,27	3,23	3,08	2,92	2,84	2,78	2,68
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,00	2,84	2,76	2,71	2,60
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,18	3,13	3,09	2,94	2,78	2,69	2,64	2,54
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,12	3,07	3,03	2,88	2,72	2,64	2,58	2,48
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	3,07	3,02	2,98	2,83	2,67	2,58	2,53	2,42
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	3,02	2,97	2,93	2,78	2,62	2,54	2,48	2,37
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,98	2,93	2,89	2,74	2,58	2,49	2,44	2,33
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,94	2,89	2,85	2,70	2,54	2,45	2,40	2,29
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,90	2,86	2,81	2,66	2,50	2,42	2,36	2,25
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99	2,93	2,87	2,82	2,78	2,63	2,47	2,38	2,33	2,22
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,96	2,90	2,84	2,79	2,75	2,60	2,44	2,35	2,30	2,19
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93	2,87	2,81	2,77	2,73	2,57	2,41	2,33	2,27	2,16
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,79	2,74	2,70	2,55	2,39	2,30	2,25	2,13
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,61	2,56	2,52	2,37	2,20	2,11	2,06	1,94
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,63	2,56	2,51	2,46	2,42	2,27	2,10	2,01	1,95	1,82
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,44	2,39	2,35	2,20	2,03	1,94	1,88	1,75
70	7,01	4,92	4,07	3,60	3,29	3,07	2,91	2,78	2,67	2,59	2,51	2,45	2,40	2,35	2,31	2,15	1,98	1,89	1,83	1,70
80	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,64	2,55	2,48	2,42	2,36	2,31	2,27	2,12	1,94	1,85	1,79	1,65
90	6,93	4,85	4,01	3,53	3,23	3,01	2,84	2,72	2,61	2,52	2,45	2,39	2,33	2,29	2,24	2,09	1,92	1,82	1,76	1,62
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,43	2,37	2,31	2,27	2,22	2,07	1,89	1,80	1,74	1,60
150	6,81	4,75	3,91	3,45	3,14	2,92	2,76	2,63	2,53	2,44	2,37	2,31	2,25	2,20	2,16	2,00	1,83	1,73	1,66	1,52
200	6,76	4,71	3,88	3,41	3,11	2,89	2,73	2,60	2,50	2,41	2,34	2,27	2,22	2,17	2,13	1,97	1,79	1,69	1,63	1,48

A Wilcoxon-próba kritikus értékei előre adott szignifikancia-szinten

Példa

Keressük 15 elemű minta alapján végzett hipotézis-ellenőrzés esetén, a α=0,01 azaz 1 %-os szignifikancia-szinthez tartozó kritikus értéket:

$$W_{15;0,01} = 20$$

A Wilcoxon-próba kritikus értékei előre adott szignifikancia-szinten

Minta nagyság	α				
	0,100	**0,050**	**0,025**	**0,010**	**0,005**
4	1	0	0	0	0
5	3	1	0	0	0
6	4	3	1	0	0
7	6	4	3	1	0
8	9	6	4	2	1
9	11	9	6	4	2
10	15	11	9	6	4
11	18	14	11	8	6
12	22	18	14	10	8
13	27	22	18	13	10
14	32	26	22	16	13
15	37	31	26	20	16
16	43	36	30	24	20
17	49	42	35	28	24
18	56	48	41	33	28
19	63	54	47	38	33
20	70	61	53	44	38

A Mann-Whitney-Wilcoxon-próba kritikus értékei előre adott szignifikancia-szinten

Példa

15, illetve 18 elemű minták alapján 5 %-os szignifikancia-szinten kívánjuk tesztelni a $H_0: \mu_1 - \mu_2 = 0$ nullhipotézist, egyoldalú alternatív hipotézis esetén ($H_0: \mu_1 - \mu_2 < 0$). A kritikus érték:

$$T_\alpha = T_{15,18;0,05} = 89$$

15, illetve 18 elemű minták alapján 5 %-os szignifikancia-szinten kívánjuk tesztelni a $H_0: \mu_1 - \mu_2 = 0$ nullhipotézist, egyoldalú alternatív hipotézis esetén ($H_0: \mu_1 - \mu_2 > 0$). A kritikus érték:

$$n_1 n_2 - T_\alpha = 15 \times 18 - T_{15,18;0,05} = 270 - 89 = 181$$

15, illetve 18 elemű minták alapján 5 %-os szignifikancia-szinten kívánjuk tesztelni a $H_0: \mu_1 - \mu_2 = 0$ nullhipotézist, kétoldalú alternatív hipotézis esetén ($H_0: \mu_1 - \mu_2 \neq 0$). A kritikus értékek:

$$T_{\alpha/2} = T_{15,18;0,025} = 81$$
$$n_1 n_2 - T_{\alpha/2} = 15 \times 18 - T_{15,18;0,025} = 270 - 81 = 189$$

Amennyiben valamelyik minta elemszáma nagyobb mint 20, akkor alkalmazható a következő közelítés:

$$T_\alpha = \frac{n_1 n_2}{2} + u_\alpha \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Példa

20, illetve 40 elemű minták esetén a kritikus érték 5 %-os szignifikancia-szinten:

$$T_{20,40;0,05} = \frac{n_1 n_2}{2} + u_{0,05} \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \frac{20 \times 40}{2} + 1{,}645 \sqrt{\frac{20 \times 40 (20 + 40 + 1)}{12}} = 504{,}9$$

A Mann-Whitney-Wilcoxon-próba kritikus értékei 5 %-os szignifikancia-szinten

n_1	n_2																		
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2	0	0	0	1	1	1	2	2	2	2	3	3	4	4	4	4	5	5	5
3	0	1	1	2	3	3	4	5	5	6	6	7	8	8	9	10	10	11	12
4	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19
5	1	2	3	5	6	7	9	10	12	13	14	16	17	19	20	21	23	24	26
6	1	3	4	6	8	9	11	13	15	17	18	20	22	24	26	27	29	31	33
7	1	3	5	7	9	12	14	16	18	20	22	25	27	29	31	34	36	38	40
8	2	4	6	9	11	14	16	19	21	24	27	29	32	34	37	40	42	45	48
9	2	5	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55
10	2	5	8	12	15	18	21	25	28	32	35	38	42	45	49	52	56	59	63
11	2	6	9	13	17	20	24	28	32	35	39	43	47	51	55	58	62	66	70
12	3	6	10	14	18	22	27	31	35	39	43	48	52	56	61	65	69	73	78
13	3	7	11	16	20	25	29	34	38	43	48	52	57	62	66	71	76	81	85
14	4	8	12	17	22	27	32	37	42	47	52	57	62	67	72	78	83	88	93
15	4	8	13	19	24	29	34	40	45	51	56	62	67	73	78	84	89	95	101
16	4	9	15	20	26	31	37	43	49	55	61	66	72	78	84	90	96	102	108
17	4	10	16	21	27	34	40	46	52	58	65	71	78	84	90	97	103	110	116
18	5	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103	110	117	124
19	5	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117	124	131
20	5	12	19	26	33	40	48	55	63	70	78	85	93	101	108	116	124	131	139

A Mann-Whitney-Wilcoxon-próba kritikus értékei 2,5 %-os szignifikancia-szinten

n_1	n_2																		
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2	0	0	0	0	0	0	1	1	1	1	2	2	2	2	2	3	3	3	3
3	0	0	0	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9
4	0	0	1	2	3	4	5	5	6	7	8	9	10	11	12	12	13	14	15
5	0	1	2	3	4	6	7	8	9	10	12	13	14	15	16	18	19	20	21
6	0	2	3	4	6	7	9	11	12	14	15	17	18	20	22	23	25	26	28
7	0	2	4	6	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35
8	1	3	5	7	9	11	14	16	18	20	23	25	27	30	32	35	37	39	42
9	1	3	5	8	11	13	16	18	21	24	27	29	32	35	38	40	43	46	49
10	1	4	6	9	12	15	18	21	24	27	30	34	37	40	43	46	49	53	56
11	1	4	7	10	14	17	20	24	27	31	34	38	41	45	48	52	56	59	63
12	2	5	8	12	15	19	23	27	30	34	38	42	46	50	54	58	62	66	70
13	2	5	9	13	17	21	25	29	34	38	42	46	51	55	60	64	68	73	77
14	2	6	10	14	18	23	27	32	37	41	46	51	56	60	65	70	75	79	84
15	2	6	11	15	20	25	30	35	40	45	50	55	60	65	71	76	81	86	91
16	2	7	12	16	22	27	32	38	43	48	54	60	65	71	76	82	87	93	99
17	3	7	12	18	23	29	35	40	46	52	58	64	70	76	82	88	94	100	106
18	3	8	13	19	25	31	37	43	49	56	62	68	75	81	87	94	100	107	113
19	3	8	14	20	26	33	39	46	53	59	66	73	79	86	93	100	107	114	120
20	3	9	15	21	28	35	42	49	56	63	70	77	84	91	99	106	113	120	128

A Mann-Whitney-Wilcoxon-próba kritikus értékei 1 %-os szignifikancia-szinten

n_1	n_2																		
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2
3	0	0	0	0	0	1	1	2	2	2	3	3	3	4	4	5	5	5	6
4	0	0	0	1	2	2	3	4	4	5	6	6	7	8	9	9	10	10	11
5	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
6	0	0	2	3	4	5	7	8	9	10	12	13	14	16	17	19	20	21	23
7	0	1	2	4	5	7	8	10	12	13	15	17	18	20	22	24	25	27	29
8	0	1	3	5	7	8	10	12	14	16	18	21	23	25	27	29	31	33	35
9	0	2	4	6	8	10	12	15	17	19	22	24	27	29	32	34	37	39	41
10	0	2	4	7	9	12	14	17	20	23	25	28	31	34	37	39	42	45	48
11	0	2	5	8	10	13	16	19	23	26	29	32	35	38	42	45	48	51	54
12	0	3	6	9	12	15	18	22	25	29	32	36	39	43	47	50	54	57	61
13	1	3	6	10	13	17	21	24	28	32	36	40	44	48	52	56	60	64	68
14	1	3	7	11	14	18	23	27	31	35	39	44	48	52	57	61	66	70	74
15	1	4	8	12	16	20	25	29	34	38	43	48	52	57	62	67	71	76	81
16	1	4	8	13	17	22	27	32	37	42	47	52	57	62	67	72	77	83	88
17	1	5	9	14	19	24	29	34	39	45	50	56	61	67	72	78	83	89	94
18	1	5	10	15	20	25	31	37	42	48	54	60	66	71	77	83	89	95	101
19	2	5	10	16	21	27	33	39	45	51	57	64	70	76	83	89	95	102	108
20	2	6	11	17	23	29	35	41	48	54	61	68	74	81	88	94	101	108	115

A Kolmogorov-Szmirnov-próba kritikus értékei előre adott szignifikancia-szinten

Példa

Keressük a 20 elemű mintához és 5 %-os szignifikancia-szinthez tartozó kritikus értéket:

$$D_{20;0,05} = 0{,}294$$

Keressük a 80 elemű mintához és 1 %-os szignifikancia-szinthez tartozó kritikus értéket:

$$D_{40;0,01} = \frac{1{,}63}{\sqrt{80}} = 0{,}182$$

A Kolmogorov-Szmirnov-próba kritikus értékei előre adott szignifikancia-szinten

n	α		
	0,100	**0,050**	**0,010**
1	0,950	0,975	0,995
2	0,776	0,842	0,929
3	0,642	0,708	0,828
4	0,564	0,624	0,733
5	0,510	0,565	0,669
6	0,470	0,521	0,618
7	0,438	0,486	0,577
8	0,411	0,457	0,543
9	0,388	0,432	0,514
10	0,368	0,410	0,490
11	0,352	0,391	0,468
12	0,338	0,375	0,450
13	0,325	0,361	0,433
14	0,314	0,349	0,418
15	0,304	0,338	0,404
16	0,295	0,328	0,392
17	0,286	0,318	0,381
18	0,278	0,309	0,371
19	0,272	0,301	0,363
20	0,264	0,294	0,356
25	0,240	0,270	0,320
30	0,220	0,240	0,290
35	0,210	0,230	0,270
n > 35	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

A Durbin-Watson-próba kritikus értékei 5 %-os szignifikancia-szinten

n	k=1		k=2		k=3		k=4		k=5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

A Durbin-Watson-próba kritikus értékei 1 %-os szignifikancia-szinten

	k=1		k=2		k=3		k=4		k=5	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

TARTALOMJEGYZÉK

A JANUS PANNONIUS TUDOMÁNYEGYETEM KÖZGAZDASÁGTUDOMÁNYI KARA ÁLTAL KIADOTT TANKÖNYVEK

Bélyácz Iván:	Vállalati tőkefinanszírozás	1991
Bélyácz Iván:	Amortizáció és pótlás	1991
Bélyácz Iván:	Tulajdonreform	1991
Vörös József:	Termelés management	1991
Bélyácz Iván:	Amortizáció elmélet	1992
Kaposi Zoltán:	Magyarország gazdaságtörténete	1992
Oroszi Sándor:	A makroökonómia alapvető elmélete	1992
Ulbert József:	A beruházások gazdaságtana	1992
Bélyácz Iván:	Privatizáció	1993
Papp László:	Könyvviteltan	1993
Balogh Sára:	Költséggazdálkodás-árak	1994
F. Elbert-Karoliny-Farkas-Poór:	Személyzeti emberi erőforrás management	1994
Barakonyi-Bencze-Ringhoffer:	Esettanulmány gyűjtemény Stratégiai tervezés és Stratégiai Menedzsment tantárgyakból	1994
Hajdu-Pintér-Rappai-Rédey:	Statisztika I.	1994
Hajdu-Pintér-Rappai-Rédey:	Stat. képl. és táblázatok gyűjteménye	1994
Rekettye Gábor:	Nemzetközi marketing	1994
Hanyecz Lajos:	Döntéshozatal, döntési modellek	1994
Komlósi Sándor:	Bevezetés egyensúlyi és optimalizáló modellek vizsgálatának matematikai módszereibe	1994
Ulbert József:	A vállalat értéke	1994
Mellár Tamás:	Stabilizáció, privatizáció, egyensúly	1994
Bélyácz Iván:	Tulajdon és gazdaság	1994
Herman-Pintér-Rappai-Rédey:	Statisztika II.	1995
Sipos Béla:	Vállalati prognosztika	1995
Orosdy Béla:	Koordináció, piac, marketing	1995
Hanyecz Lajos:	Tervezés és stratégia	1995
Varga József:	Mérték- és valószínűségi alapok	1995
László Gyula:	Emberi erőforrás gazdálkodás és munkaerőpiac	1995
Kaposi Zoltán:	Az európai gazdaság és társadalom fejlődése a 18-20. században	1995
Balogh S.-Bélyácz I.-László Gy.- -Marosi A.-Szerb L.-Ulbert J.:	Vállalati gazdaságtan	1995
Bélyácz Iván:	Tőkeberuházási és finanszírozási döntések	1995
Hoóz István:	Népesség és népesedés	1995
Danyi P.-Varró Z.:	Operációkutatás	1995
Berend Iván:	Új növekedési pálya felé	1995
Bélyácz Iván:	Érték- és tulajdon	1995
Török Ádám:	Piacműködés és iparvédelem	1995
Bessenyei István:	A gazdasági növekedés alapvető elmélete	1995
Fehér-Somogyváriné- Szabó-Papp:	Mérlegtan és mérlegelemzés	1996
Ulbert József:	Értékpapírértékelés	1996
Borgulya Istvánné:	Üzleti kommunikáció kulturák találkozásában	1996
Zeller Gyula:	Marketing stratégiák, marketingszervezés	1996
Szabó László:	Banküzemtan	1997
Bélyácz Iván:	Tőkefinanszírozási számítások	1997
Szira Tamás:	Gazdaság – társadalom – politika	1997
Sramó A. – Kruzslicz F.:	Rendszerfejlesztés CASE eszközökkel	1997
Danyi P. – Varró Z.:	Operációkutatás üzleti döntések megalapozásához	1997
Bugár Gyöngyi:	Portfolió elemzés	
Hoványi Gábor:	Menedzsment tanácsadás	
Mellár Tamás:	Alkalmazott makroökonómia	
Katits Etelka:	Vállalati válságkezelés pénzügyi módszerei	