Bevezetés az ökonometriába

Sándor Zsolt

T3 Kiadó, Sepsiszentgyörgy

T3 Kiadó, Sepsiszentgyörgy
ISBN: 978-973-1962-82-5

Bevezetés az ökonometriába

Sándor Zsolt

2019
Lektorálta: dr. Madaras Szilárd

Tartalomjegyzék

Bevezetés

Ezek a jegyzetek a madridi Carlos III Egyetemen és a Sapientia Erdélyi Magyar Tudományegyetemen oktatott alap- és mesteri szintű Ökonometria előadásokra készültek. Ebben a bevezető fejezetben vázoljuk, hogy miről lesz szó a jegyzetekben. Szó lesz röviden arról, hogy mi az Ökonometria tárgya és céljai, és hogy mi egy ökonometriai modell, ami a jegyzetek fő módszertani eleme. A fejezet végén felsoroljuk a különböző fejezetekben tárgyalt témákat.

Mi az Ökonometria?

Az Ökonometria tárgya a gazdasági adatok statisztikai elemzése. Három fő célját említjük:

- Bizonyos gazdasági változók közötti összefüggések meghatározása és számszerűsítése, például a kereslet, munkanélküliség és a GDP közötti összefüggés vagy keresletfüggvények, amelyek a piaci részesedést határozzák meg a termékek ára és a termékjellemzők függvényében.

- Fontos gazdasági változók értékeinek az előrejelzése, például GDP vagy foglalkoztatottság.

- Gazdasági elméletek vagy modellek tesztelése.

Az előbbi két cél a gazdasági gyakorlatot, az utóbbi a gazdaságtudományt szolgálja. Ezeket a célokat az ökonometria modellek szerkesztésével és a modellek adatok segítségével történő tanulmányozásával éri el.

Először röviden tárgyaljuk a modellekkel kapcsolatos fontosabb fogalmakat. A későbbi fejezetekben elsősorban a lineáris modellekkel fogunk foglalkozni, amelyekben az ismeretlen paraméterek lineárisan szerepelnek. Egy modell bizonyos változók közötti összefüggés. Modellezési szempontból a változók lehetnek:

- *bemenő* vagy *független*, amelyek olyan változók, melyek meghatározzák a modell többi változóját, vagy

- *kimenő* vagy *függő*, amelyek olyan változók, melyeket meghatároznak a bemenő változók.

Abból a szempontból, hogy az elemző rendelkezik-e adatokkal az illető változóról, a változók lehetnek:

- *észlelt*, vagy
- *nem észlelt* vagy *hibaváltozó* vagy *eltérésváltozó*.

A nem észlelt változók valószínűségi változókkal való helyettesítése vezetett az ökonometria és a statisztika valószínűség-elméleti megalapozásához. Ezért, az 1. fejezetet arra szenteljük majd, hogy áttekintsük a valószínűségi változókkal kapcsolatos fontosabb fogalmakat.

Példák modellekre

1. Autókeresleti modell. Tekintsük a következő autókereslet-függvényt, ahol egy adott autótípus iránti keresletet (k) az árán (a) kívül még az autó mérete (m) és az autó motorjának lóereje (ℓ) határozza meg:

$$k = \beta_1 + \beta_2 a + \beta_3 m + \beta_4 \ell + \varepsilon.$$

A modell változói: k, a, m, ℓ, ε míg a β_1,...,β_4 paraméterek.

- a, m, ℓ, ε független változók, k függő változó,
- a, m, ℓ, k észlelt változók, ε hibaváltozó, általában görög betűvel jelöljük.

Az autókereslet-függvényt meghatározzuk, ha becsüljük a paramétereket. A becsült modell segítségével előrejelezhetjük egy újonnan piacra dobott autótípus iránti keresletet. Egy ilyen modell becslését a 3. fejezetben (Többváltozós lineáris modell) tanulmányozzuk.

Felmerül a kérdés, hogy jó-e ez a modell, és hogyan lehetne javítani rajta? Tekintsük a következő példát.

2. Egy érdekesebb és bonyolultabb modell az, amelyben az i autótípus keresletfüggvénye

$$k_i = M \frac{e^{\beta_1+\beta_2 a_i+\beta_3 m_i+\beta_4 \ell_i+\varepsilon_i}}{1+\sum_{j=1}^{N} e^{\beta_1+\beta_2 a_j+\beta_3 m_j+\beta_4 \ell_j+\varepsilon_j}},$$

ahol M a vásárlók száma (egy országon vagy piacon belül) és N az autótípusok száma. Ebben az esetben

- a_i, m_i, ℓ_i, ε_i, M független, k_i függő változók,
- a_i, m_i, ℓ_i, k_i, M észlelt, ε_i hibaváltozó.

A keresletfüggvény képletében a tört

$$\frac{e^{\beta_1+\beta_2 a_i+\beta_3 m_i+\beta_4 \ell_i+\varepsilon_i}}{1+\sum_{j=1}^{N} e^{\beta_1+\beta_2 a_j+\beta_3 m_j+\beta_4 \ell_j+\varepsilon_j}}$$

az i autótípus piaci részesedése, és bizonyos feltevések mellett egyenlő annak a valószínűségével, hogy az N autótípus közül az i-t vásárolják meg. Ehhez hasonló modellekről az 5. fejezetben (Kvalitatív függővváltozójú modellek) lesz szó.

Az utóbbi modell jobb mint az 1. példa modellje, mert ebben a modellben egy autótípus keresletfüggvénye függ más autótípusok árától. Például, ha egy Opel-típus ára csökken, akkor várható, hogy a többi autótípus iránti kereslet csökken, és ez a modell teljesíti az említett tulajdonságot. Következésképpen, ezzel a modellel jobban előrejelezhető egy újonnan piacra dobott típus iránti kereslet.

3. Tőzsde-indexek. Számos gazdasági változó előrejelzése célszerűbb az illető változó múltbeli értékei alapján. Ilyenek például a tőzsde-indexek. Ez a fentiektől eltérő modellezést és statisztikai elemzést igényel. Ez a 6. fejezet (Idősorelemzés) témája.

Rövid áttekintés

A jegyzetek 1. fejezete, amint már említettük, a valószínűségi változókkal kapcsolatos fogalmakat tekinti át. Elsősorban folytonos valószínűségi változókról lesz szó, valamint ezek várható értékével és varianciájával kapcsolatos tulajdonságokról. Bevezetjük a statisztikai teszt fogalmát, amelyre példaként a populációs átlag tesztelését tárgyaljuk. Egy minta átlagával kapcsolatban még megemlítjük a matematikai statisztikai két alaptörvényét: a nagy számok törvényét és a központi határeloszlás tételt.

A 2. fejezet azt a lineáris modellt tárgyalja, amelyben egyetlen független változó szerepel. A paraméterek becslése céljából a legkisebb négyzetek módszere kerül bemutatásra, amellyel kapcsolatban bizonyos feltételek mellett tanulmányozzuk a becslőfüggvény statisztikai tulajdonságait valamint a paraméterek tesztelését. A 3. fejezet kibővíti a legkisebb négyzetek módszerét a többváltozós lineáris modellre. Itt szó van még a paraméterek egyidejű teszteléséről és előrejelzések tanulmányozásáról is.

A 4. fejezetben olyan eseteket tárgyalunk, amikor a többváltozós lineáris modellben nem teljesülnek a megadott feltételek. Három ilyen eset tanulmányozására kerül sor: heteroszkedaszticitás, autokorreláció és multikollinearitás. A legkisebb négyzetek módszerének a tulajdonságait és a helyes becslési és tesztelési módszereket tárgyaljuk.

Az 5. fejezetben néhány kvalitatív függővváltozójú modellt tárgyalunk: a logitot és a probitot. Itt elsősorban a lineáris modellhez viszonyítva tanulmányozzuk a modell paramétereinek az értelmezését és a modellek becslését.

A 6. fejezetben stacionárius idősormodelleket és azok elemzését tanulmányozzuk. Elsősorban az ARMA modellek változatait tanulmányozzuk és ezek azonosítását autokorrelációs függvények segítségével. Szó lesz még a modellszelekcióról és előrejelzésekről ARMA modellek alapján.

A jegyzetek tanulmányozásának javasolt sorrendje a következő. Az első négy fejezetet tanulmányozása egymás után javasolt, míg az 5. és a 6. fejezeteket az első négy után egymástól függetlenül lehet tanulmányozni. A fejezetek végén lévő gyakorlatok segítenek az ismeretek elmélyítésében.

1. fejezet

Valószínűségi változók

A valóságban és a gazdasági életben a legtöbb változó jövőbeli értéke ismeretlen. Például, egy diák államvizsga-jegye tanév közben, a sikeres államvizsgázók száma az egyetemen, egy diák jövedelme egy év múlva vagy a holnapi lej-euró árfolyam. Ezek az események ismeretlen kimenetűek. Ehhez hasonló példák elemzése céljából vezették be a valószínűségi változókat.

Ebben a fejezetben értelmezzük a valószínűségi változó fogalmát és ezen belül megemlítjük a folytonos valószínűségi változók legfontosabb tulajdonságait, értelmezzük a várható értéket és a varianciát, majd bemutatjuk a számunkra legfontosabb folytonos valószínűségi változókat. Az 1.2. alfejezet a statisztikai tesztek világába nyújt betekintést. Az 1.3. alfejezet kovarianciát és a korrelációs együtthatót mutatja be. Az 1.4. alfejezet az átlag statisztikai tulajdonságait tárgyalja nagy minták esetén. Az 1.5. alfejezet néhány számolásos gyakorlatot tartalmaz.

Egy valószínűségi változó (a továbbiakban rövidítve vv) egy olyan változó, amelynek ismerjük a lehetséges értékeit és az ezekhez rendelt valószínűségeket, de a megvalósulása előtt nem ismerjük a pontos értékét. Egy valószínűségi változó megvalósulása egy olyan esemény, amely során felveszi egy lehetséges értékét.

Példák

1. Dobókocka. A lehetséges értékek 1, 2, ..., 6 és ezek valószínűségei 1/6 mindegyik értékre.

2. A holnapi lej-euró árfolyam. Ennek a lehetséges értékei és a valószínűségeik megközelítőleg meghatározhatók a megvalósult értékek alapján.

A lehetséges értékek szerint a valószínűségi változók lehetnek diszkrétek és folytonosak. Egy folytonos valószínűségi változó egy vagy több intervallumból

minden értéket felvehet. Egy példa erre a holnapi lej-euró árfolyam. Egy diszkrét valószínűségi változó csak jól elhatárolt értékeket vehet fel, mint például a dobókocka.

A későbbi fejezetekben elsősorban folytonos valószínűségi változókkal fogunk dolgozni, ezért ezekre koncentrálunk a továbbiakban.

1.1. Folytonos valószínűségi változók

Egy X folytonos valószínűségi változót egyértelműen meg lehet határozni a hozzárendelt sűrűségfüggvény (f) által, ugyanis bármilyen a, b számokra az $(a < X < b)$ esemény valószínűsége megadható a sűrűségfüggvény segítségével mint:

$$P\left(a < X < b\right) = \int_a^b f(x)dx,$$

vagyis a valószínűség az f grafikonja és a vízszintes tengely közötti terület a és b között.

Egy fontos tualjdonsága az f sűrűségfüggvénynek, hogy nem vehet fel negatív értékeket, és az integrálja a valós számok halmazán 1:

$$f\left(x\right) \geq 0; \qquad \int_{-\infty}^{+\infty} f(x)dx = 1.$$

A sűrűségfüggvény grafikonja gyakorlatilag majdnem megegyezik egy hisztogrammal és a hisztogramhoz hasonlóan egy $[a, b]$ intervallumban előforduló lehetséges értékek relatív gyakoriságát fejezi ki.

Megjegyzés. $P\left(X = x\right) = 0$, és ezért $P\left(X < x\right) = P\left(X \leq x\right)$ bármely X folytonos valószínűségi változóra és x valós számra. Hasonlóan:

$$P\left[a \leq X \leq b\right] = P\left[a \leq X < b\right] = P\left[a < X \leq b\right] = P\left[a < X < b\right].$$

1.1.1. Várható érték és variancia

Legyen X folytonos valószínűségi változó. X várható értéke:

$$E\left[X\right] = \mu = \int_{-\infty}^{\infty} xf(x)dx.$$

A várható érték egy helyzetmutató, mely megadja, hogy az X valószínűségi változó milyen érték körül váltakozik véletlenszerűen, ugyanis megközelítőleg

az átlaggal egyenlő:

$$E[X] \approx \frac{1}{N}\sum_{i=1}^{N} X_i,$$

ahol $X_1, ..., X_N$ az X valószínűségi változó N megvalósult értéke. Ez a megközelítés annál pontosabb, minél nagyobb N.

Tulajdonság. Ha $g(X)$ az X vv egy függvénye, akkor

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

Például, ha $g(X) = X^2$, akkor $E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx$.

Az X valószínűségi változó varianciája (szórásnégyzete):

$$var(X) = \sigma^2 = E\left[(X-\mu)^2\right] = E\left[X^2\right] - \mu^2.$$

A fenti tulajdonság alapján

$$var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2.$$

A variancia egy szóródási mutató, amely méri, hogy az X valószínűségi változó lehetséges értékei mennyire szóródnak szét a várható érték körül.

Legyenek $X_1, X_2, ..., X_n$ megfigyelések (vagyis megvalósult értékek) egy bizonyos valószínűségi változóhoz (például a lej-euró napi árfolyam az elmúlt n munkanapon). Gyakorlati szempontból a várható érték az a szám amely körül csoportosulnak a megfigyelések. A várható érték egy becslőfüggvénye a megfigyelések átlaga:

$$\overline{X} \equiv \frac{X_1 + X_2 + ... + X_n}{n}.$$

A variancia egy pozitív szám, amely megmutatja, hogy a megfigyelések menynyire szóródnak szét a várható érték körül: minél nagyobb a variancia, annál nagyobb mértékű a szétszóródás. A variancia (egy) becslőfüggvénye:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

Az $\overline{X}$ és az s^2 kifejezések példák becslőfüggvényekre. Alább a 2. fejezetben majd tovább tisztázzuk a becslőfüggvény fogalmát.

1.1.2. Az egyenletes eloszlás

Az egyenletes eloszlást egy $[a, b]$ intervallumon értelmezzük, sűrűségfüggvénye

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{ha } a \leq x \leq b, \\ \\ 0 & \text{másképp.} \end{cases}$$

Itt megjegyezzük, hogy az „eloszlás“ fogalmat gyakran használják vvkra; tulajdonképpen felcserélhető a „valószínűségi változó“ fogalommal. Az eloszlás neve („egyenletes“) onnan származik, hogy egy ilyen eloszlású X vvra, annak a valószínűsége, hogy X az $[a, b]$ bizonyos részintervallumában van, csak az intervallum hosszától függ, és nem a helyétől, vagyis X egyenletesen oszlik el az $[a, b]$ intervallumban.

Várható értéke:

$$\mu = \int_{-\infty}^{\infty} x f\left(x\right) dx = \frac{a+b}{2},$$

varianciája:

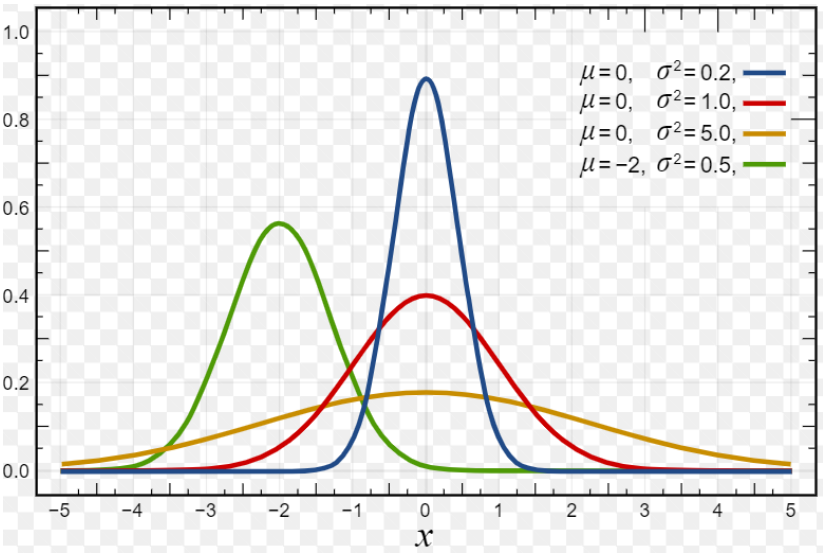
$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \frac{(b-a)^2}{12}.$$

1.1.3. A normális eloszlás

A **normális** eloszlást egyértelműen meg lehet határozni a μ várható értéke és a σ^2 varianciája segítségével. Egy ilyen eloszlású X vv jelölése $X \sim N\left(\mu, \sigma^2\right)$. Sűrűségfüggvénye:

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ ahol } -\infty < x < \infty,$$

melynek grafikonja:



A sűrűségfüggvényre fennállnak az összefüggések:

$$\int_{-\infty}^{\infty} f(x|\mu,\sigma)dx = 1,$$

$$E[X] = \int_{-\infty}^{\infty} x f(x|\mu,\sigma)dx = \mu,$$

$$var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x|\mu,\sigma)dx = \sigma^2.$$

A sűrűségfüggvény szimmetrikus a várható érték körül.

A normális eloszlás a fontosságát egyrészt abból nyeri, hogy egyértelműen meg lehet határozni a két legfontosabb statisztikai mutató (várható érték és variancia) segítségével, másrészt.a Központi határeloszlás-tételből, amit az 1.4. alfejezetben tárgyalunk.

Ha $\mu = 0$ és $\sigma^2 = 1$ a standard normális eloszlást kapjuk, melynek sűrűségfüggvénye

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \text{ ahol } -\infty < z < \infty.$$

A Z standard normális eloszlású vvra a $(Z \leq z)$ esemény valószínűsége

$$P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

Ennek a megközelítő értékeit táblázatba foglalták; pontosabb értékeket statisztikai programok segítségével lehet kapni.

Tulajdonság. Ha $X \sim N(\mu, \sigma^2)$ és a, b két szám, akkor $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

Egy nem standard normális eloszlású vvhoz rendelt valószínűségeket a következő tulajdonság segítségével számítjuk ki. Ha $X \sim N(\mu, \sigma^2)$, akkor a fenti tulajdonságból $a = 1/\sigma$ és $b = -\mu/\sigma$ értékekre

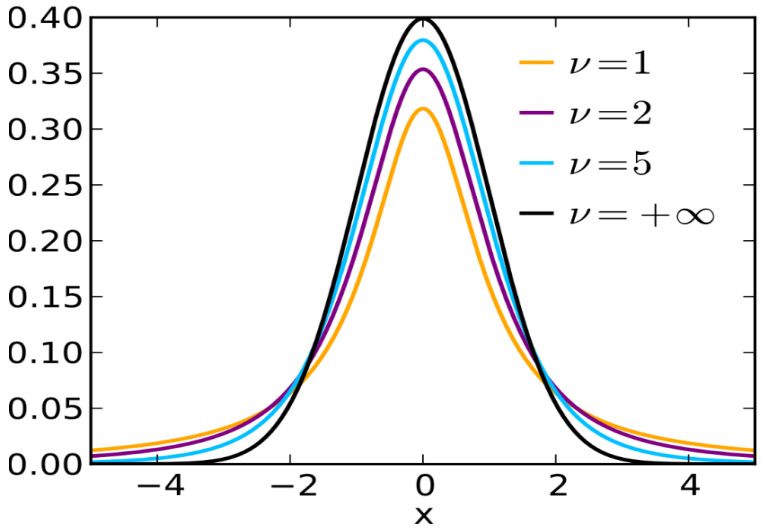
$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Ez alapján, mivel az $(X < x)$ és a $\left(Z < \frac{x-\mu}{\sigma}\right)$ események ugyanakkor fordulhatnak elő, a valószínűségük egyenlő

$$P(X < x) = P\left(\frac{X - \mu}{\sigma} < \frac{x - \mu}{\sigma}\right) = P\left(Z < \frac{x - \mu}{\sigma}\right).$$

1.1.4. A t eloszlás (Student t eloszlása)

A normális eloszláshoz hasonlóan, a várható értéke körül szimmetrikus eloszlás. A sűrűségfüggvényének a grafikonja:



A normális eloszlástól eltérően, a t eloszlás várható értéke mindig 0. Nevét a Student álnevet használó William Sealy Gosset után kapta, aki a dublini Guiness sörgyárban dolgozott, és aki egy 1908-as tanulmányban leírja,

hogy hogyan lehet ezt az eloszlást az árpa minőségének a meghatározására használni.

Ahogy a fenti grafikonon is látszik, egyetlen paraméter határozza meg, amelyet szabadsági foknak nevezünk (a grafikonon ν-vel jelölik). Az eloszlás jelölése: $t(\nu)$. Végtelen szabadsági fok esetén ($\nu = \infty$) a sűrűségfüggvénye megegyezik a standard normális eloszlás sűrűségfüggvényével (a grafikonon a fekete vonal). A t eloszlás bizonyos statisztikai tesztek miatt nagy fontosságú.

Példa: Megfigyelések átlagának a tesztelése

Tegyük fel, hogy az $X_1, X_2, ..., X_n$ megfigyelések normális eloszlásúak (ezt sejthetjük a hisztogram alapján), viszont nem tudjuk a várható értéküket és a varianciájukat. Azt szeretnénk tudni, hogy a megfigyelések 0 várható értekűek-e. Ezt a példát az alábbi alfejezetben folytatjuk. Ezt a példát az alábbi alfejezetben folytatjuk, ahol a tesztelés során hasznunkra lesz a t eloszlás is.

Megjegyzés. Azáltal, hogy bizonyos megfigyelésekről azt feltételezzük egyrészt, hogy valószínűségi változók, másrészt, hogy normális eloszlásúak, másfajta következtetéseket is le tudunk vonni, mint pusztán a leíró statisztika által.

1.2. Statisztikai tesztek

A vizsgált hipotézisek kiértékeléséhez a statisztikában egy úgynevezett statisztikai tesztet vagy próbát dolgoztak ki. A teszt logikai helyessége megkövetel néhány fogalmat és alapösszefüggést.

Egy statisztikai teszt legfontosabb összetevője a tesztstatisztika (vagy egyszerűen statisztika), ami a megfigyelésekből kiszámított olyan becslőfüggvény, amely alkalmas a probléma vizsgálatára. Ezt jelöljük T-vel; $T = T(X_1, X_2, ..., X_n)$.

Azt az állítást, amelyet vizsgálunk, nullhipotézisnek nevezzük, jele H_0. Úgy vizsgáljuk ezt az állítást, hogy megnézzük, mit kapnánk, ha érvényes lenne az állítás. Egy, a nullhipotézist kizáró állítást alternatív hipotézisnek nevezünk, jele H_1. A szignifikancia szint egy α-val jelölt kis szám, amelyet leginkább 0.05-nek vesznek (még gyakori értékek: $\alpha = 0.01$ vagy 0.1). Az elfogadási tartomány az a (c_1, c_2) intervallum, ahol c_1 és c_2 kritikus értékek, amelyre

$$P(c_1 < T < c_2) = 1 - \alpha.$$

Ezt az intervallumot a T statisztika eloszlása alapján határozzuk meg.

A statisztikai teszt egy eljárás, amelyben a fenti fogalmak mindegyikét felhasználjuk. A statisztikai teszt lépései:

1. megválasztjuk α-t,
2. meghatározzuk az elfogadási tartományt,
3. kiszámítjuk a T értékét a megfigyelésekből és H_0 felhasználásával,
4. a teszt eredménye: ha $T \in (c_1, c_2)$, nem utasítjuk el H_0-t; ha $T \notin (c_1, c_2)$, elutasítjuk H_0-t.

Azt nem mondjuk, hogy „elfogadjuk" H_0-t, ha $T \in (c_1, c_2)$, mert a teszt logikája alapján a H_0-t felhasználjuk a T kiszámításához, ezért a $T \in (c_1, c_2)$ esemény csak azt jelenti, hogy H_0 nem vezetett ellentmondáshoz.

A példa folytatása

Jelöljük a megfigyelések várható értékét μ-vel. Ezért a teszt jelöléseit használva, $H_0 : \mu = 0, \quad H_1 : \mu \neq 0$ (H_1 lehet $\mu > 0$ is, ebben az esetben az elfogadási tartományt másképp határozzuk meg). Legyen $\alpha = 0.05$. Az ehhez a nullhipotézishez kidolgozott statisztika:

$$T = \frac{\sqrt{n} \cdot \overline{X}}{s},$$

ahol $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$ a szórás becslőfüggvénye.

A T statisztika egy vv, minek eloszlása: $T \sim t\,(n-1)$ (vagyis, bebizonyították, hogy a T sűrűségfüggvénye megegyezik a $t\,(n-1)$ vv sűrűségfüggvényével). Mivel a T sűrűségfüggvénye szimmetrikus 0 körül, az elfogadási tartományt úgy határozzuk meg, hogy $c_1 = -c,\ c_2 = c$ és

$$P\,(-c < T < c) = 0.95.$$

Számszerű példák.

A. Most tegyük fel, hogy van egy 100 megfigyelésből álló mintánk, amire $\overline{X} = -0.088,\ s = 0.874$. A teszt lépései:

1. $\alpha = 0.05$
2. $c = 1.98$, tehát az elfogadási tartomány $(-1.98, 1.98)$
3. $T = \frac{\sqrt{n} \cdot \overline{X}}{s} = \frac{10 \cdot (-0.088)}{0.874} = -1.007$
4. Mivel $T = -1.007 \in (-1.98, 1.98)$, nem utasítjuk el H_0-t.

B. Most tegyük fel, hogy van egy másik 100 megfigyelésből álló mintánk, amire $\overline{X} = 0.774,\ s = 1.009$, és ugyanazt a nullhipotézis szeretnénk tesztelni. A teszt lépései:

1. $\alpha = 0.05$

2. $c = 1.98$, tehát az elfogadási tartomány $(-1.98, 1.98)$
3. $T = \frac{\sqrt{n}\cdot\overline{X}}{s} = \frac{10\cdot 0.774}{1.009} = 7.671$
4. Mivel $T = 7.671 \notin (-1.98, 1.98)$, elutasítjuk H_0-t.

1.3. Két vv közötti összefüggés

Az $X,\ Y$ vvk kovarianciája:

$$\begin{aligned} cov\,(X,Y) &= E\left[(X - E\left[X\right])\left(Y - E\left[Y\right]\right)\right] \\ &= E\left[XY\right] - E\left[X\right]E\left[Y\right] \end{aligned}$$

a két vv lineáris függőségét méri. Például, ha $cov\,(X,Y) > 0$, akkor X és Y együtt nő vagy csökken. Ha $cov\,(X,Y) < 0$, akkor X növekedése Y csökkenésével történik. Ha $cov\,(X,Y) \simeq 0$, akkor X és Y változása között kevés összefüggés van.

Az $X,\ Y$ korrelációs együtthatója:

$$corr\,(X,Y) = \frac{cov\,(X,Y)}{\sqrt{var\,(X)\,var\,(Y)}}.$$

A korrelációs együttható a kovarianciához hasonlóan méri a két vv változását, viszont az értéke le van szűkítve -1 és 1 közé. Ha $cov\,(X,Y) = 0$ vagy $corr\,(X,Y) = 0$ akkor azt mondjuk, hogy az $X,\ Y$ vvk nem korreláltak. Ha a korrelációs együttható közel van 1-hez vagy -1-hez, akkor erős lineáris összefüggés van a két vv között.

Alább felsorolunk néhány tulajdonságot, melyek a várható értékhez, varianciához és a kovarianciához kapcsolódnak. Ezek hasznosak lesznek a következő fejezetekben. Legyenek $X,\ Y,\ X_1, ..., X_n$ nem feltétlenül normális eloszlású vvk, α egy szám, ami nem vv. Itt megjegyezzük, hogy az olyan változókat, amelyek nem vvk determinisztikus változóknak is nevezik a szakirodalomban, tehát α determinisztikus. Ekkor fennállnak a következő tulajdonságok.

1. $E\left[\alpha\right] = \alpha,\ \ var\,(\alpha) = 0.$
2. $E\left[\alpha X\right] = \alpha E\left[X\right].$
3. $E\left[X + Y\right] = E\left[X\right] + E\left[Y\right]$ és $E\left[X_1 + \cdots + X_n\right] = E\left[X_1\right] + \cdots + E\left[X_n\right].$
4. $Var\left[\alpha X\right] = \alpha^2 Var\left[X\right].$
5. $Var\left[X \pm Y\right] = Var\left[X\right] + Var\left[Y\right] \pm 2Cov\left[X,Y\right].$
6. Ha X és Y nem korreláltak, akkor $E\left[XY\right] = E\left[X\right]E\left[Y\right]$ és $Var\left[X \pm Y\right] = Var\left[X\right] + Var\left[Y\right].$

7. $Var\left[X_1+\cdots+X_n\right]=Var\left[X_1\right]+\cdots+Var\left[X_n\right]+2\sum_{i<j}Cov\left[X_i,X_j\right].$
8. Ha $X_1,\cdots,X_n$ páronként nem korreláltak, akkor
$Var\left[X_1+\cdots+X_n\right]=Var\left[X_1\right]+\cdots+Var\left[X_n\right].$
9. Ha $X\sim N\left(\mu_1,\sigma_1^2\right)$, $Y\sim N\left(\mu_2,\sigma_2^2\right)$ és a,b két szám, akkor $aX+bY$ is normális eloszlású.
10. $Var\begin{pmatrix}X\\Y\end{pmatrix}=\begin{pmatrix}Var\left(X\right) & Cov\left(X,Y\right)\\ Cov\left(X,Y\right) & Var\left(Y\right)\end{pmatrix}.$
11. Két független vv korrelációja 0, fordítva csak normális eloszlású vv-kra igaz.

1.4. Az átlag tulajdonságai nagy minták esetén

A statisztika módszertani lényege, hogy megfigyelések alapján von le következtetéseket egy bizonyos populációról. Sok esetben nincs szükség és nagyon költséges pontosan meghatározni a populáció bizonyos jellemzőit.

Például, legyen a populáció a romániai munkavállalók fizetése, és ki szeretnénk számítani ennek az átlagát. A megfigyeléseink legyenek egy 100 fős minta, vagyis 100 véletlenszerűen kiválasztott munkavállaló fizetése. Azt sejtjük, hogy ha nagyobb mintát veszünk, akkor jobban megközelítjuk az átlagfizetést. A megfigyelések alapján az átlag és szórás segítségével tudunk olyan intervallumot szerkeszteni, amelyben nagy valószínűséggel benne van a populáció átlaga.

Azt, hogy az átlag használható a populáció átlagának meghatározására, a nagy számok törvénye támasztja alá. Azt, hogy a fent említett intervallum használható a populáció átlagának meghatározására, a központi határeloszlás-tétel támasztja alá.

Legyenek $X_1,\ldots,X_n$ megfigyelések, amelyekről azt feltételezzük, hogy független vvk μ várható értékkel és σ^2 varianciával. Az

$$\overline{X}=\frac{X_1+\cdots+X_n}{n}$$

átlagra fennállnak az alábbiak:

$$E\left(\overline{X}\right)=E\left(\sum_{i=1}^{n}\frac{1}{n}X_i\right)=\sum_{i=1}^{n}\frac{1}{n}E\left(X_i\right)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

$$var\left(\overline{X}\right)=\sum_{i=1}^{n}\frac{1}{n^2}var\left(X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}var\left(X_i\right)=\frac{\sigma^2}{n}.$$

Ebből következik, hogy a várható érték egyenlő a populációs átlaggal.

A nagy számok törvénye. Nagy minta esetén (nagy n értékre) az átlag jól megközelíti a várható értéket, vagyis $\overline{X} \simeq \mu$.

Ez abból is látszik, hogy $E\left(\overline{X}\right) = \mu$ és $var\left(\overline{X}\right) = \frac{\sigma^2}{n} \simeq 0$ nagy n értékre. Ezért elmondhatjuk, hogy nagy minta esetén az átlag $\left(\overline{X}\right)$ a populációs átlag (μ) körül váltakozik véletlenszerűen, viszont a váltakozás mértéke kicsi mivel az átlag varianciája kicsi.

Tudjuk, hogy ha $X_1, \ldots, X_n$ normális eloszlásúak, akkor az átlag is normális eloszlású, és ezért $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Ez akkor is így van megközelítőleg, ha $X_1, \ldots, X_n$ nem normális eloszlásúak és sok megfigyelésünk van.

Központi határeloszlás-tétel. Nagy minta esetén az átlag eloszlása megközelíti az $N\left(\mu, \frac{\sigma^2}{n}\right)$ eloszlást.

Ennek a határérték szempontjából helyes megfogalmazása a következő: $\dfrac{\sqrt{n} \cdot \left(\overline{X} - \mu\right)}{\sigma}$ statisztika eloszlása jól megközelíti az $N(0,1)$ eloszlást.

Hogyan tudunk olyan intervallumot szerkeszteni, amelyben nagy valószínűséggel benne van a populáció átlaga?

Jelölje μ a populáció átlagát. Ekkor fennáll, hogy

$$T = \frac{\sqrt{n} \cdot \left(\overline{X} - \mu\right)}{s} \sim t\left(n-1\right),$$

ahol $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$ a szórás becslőfüggvénye. Vagyis, a T statisztika eloszlása $t\left(n-1\right)$; ez úgy igazolható, hogy kiszámítjuk a T statisztika sűrűségfüggvényét, és abból láthatjuk, hogy megegyezik a $t\left(n-1\right)$ eloszlás sűrűségfüggvényével.

Meghatározzuk az elfogadási tartományt, vagyis a c kritikus értéket úgy, hogy

$$P\left(-c < T < c\right) = 0.95.$$

Fennebb láttuk, hogy ha $n = 100$ akkor $c = 1.98$.

Ekkor

$$P\left(-c < \frac{\sqrt{n} \cdot \left(\overline{X} - \mu\right)}{s} < c\right) = 0.95,$$

ami átírható a következő alakba

$$P\left(\overline{X} - c\frac{s}{\sqrt{n}} < \mu < \overline{X} + c\frac{s}{\sqrt{n}}\right) = 0.95.$$

Vagyis, a populáció átlaga 95% valószínűséggel a $\left(\overline{X} - c\frac{s}{\sqrt{n}}, \overline{X} + c\frac{s}{\sqrt{n}}\right)$ intervallumban van. Ezt a μ paraméter 95%-os konfidencia-intervallumának vagy intervallum-becslésének nevezzük.

Számszerű példák

A. $n = 100$, $\overline{X} = -0.088$, $s = 0.874$.

$\overline{X} - c\frac{s}{\sqrt{n}} = -0.088 - 1.98\frac{0.874}{10} = -0.261$,

$\overline{X} + c\frac{s}{\sqrt{n}} = -0.088 + 1.98\frac{0.874}{10} = 0.085$.

Tehát a populációs átlag 95% valószínűséggel a $(-0.261, 0.085)$ intervallumban van.

B. $n = 100$, $\overline{X} = 0.774$, $s = 1.009$.

$\overline{X} - c\frac{s}{\sqrt{n}} = 0.774 - 1.98\frac{1.009}{10} = 0.574$,

$\overline{X} + c\frac{s}{\sqrt{n}} = 0.774 + 1.98\frac{1.009}{10} = 0.974$.

Tehát a populációs átlag 95% valószínűséggel a $(0.574, 0.974)$ intervallumban van.

1.5. Gyakorlatok

Legyen $X, Y, Z \sim N(0, 1)$ független vvk. Határozzuk meg:

1. $E[2X + 1]$, $var(2X + 1)$, a $2X + 1$ eloszlását,
2. $E[2X + Y]$, $var(2X + Y)$, a $2X + Y$ eloszlását,
3. $E[X + Y + Z]$, $var(X + Y + Z)$,
4. $cov(X + Y, X + Z)$, $corr(X + Y, X + Z)$, az $(X + Y, X + Z)$ eloszlását.

2. fejezet

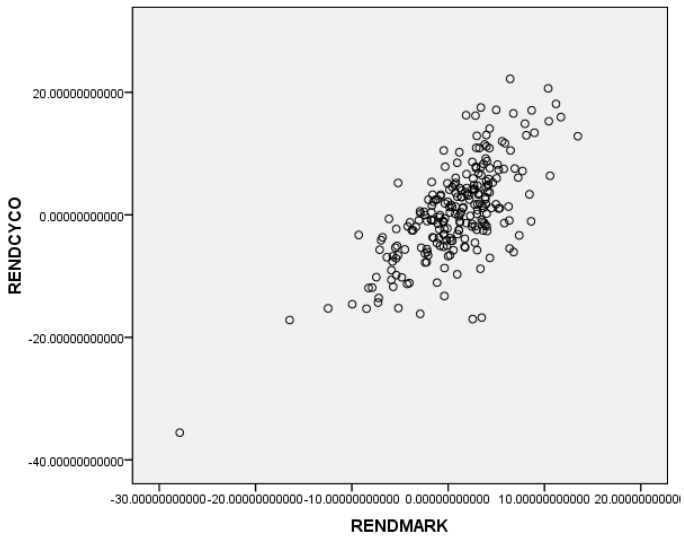
Az egyváltozós lineáris modell elemzése

Példaként tekintsük a tőkejavak árazási modelljét, amely összefüggést állít fel valamely értékpapír várható hozamának meghatározására. Ezen a modellen belül meg szeretnénk határozni az összes piacra számított nettó hozam (RENDMARK) hatását a ciklikus termékpiaci (tartós termékek, autók, ruhaneműk, sporttermékek piaca) nettó hozamra (RENDCYCO). (A nettó hozam az illető hozam és a banki kamat közötti különbség.)

Egy statisztikailag becsülhető modellt kell szerkesszünk, és ezt helyesen becsüljük. Ebben az esetben a függő változó a RENDCYCO és a független változó a RENDMARK. Az alábbi szóródási kép pontjai hozzávetőlegesen egy egyenes mentén helyezkednek el, ezért a két változó között megközelítőleg egy lineáris összefüggés áll fenn. Ezért a két változó közötti összefüggést a következő egyváltozós lineáris modellel határozzuk meg:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, ..., n,$$

amelyben
- y_i függő vagy kimenő vagy eredmény- vagy endogén változó,
- x_i független vagy bemenő vagy magyarázó vagy exogén vagy regresszor vagy kovariáns észlelt változók,
- ε_i nem észlelt vagy hibaváltozó vagy eltérésváltozó, (amely szintén bemenő változó),
- α és β paraméterek: α konstans együttható, β meredekségi együttható.



Mivel ε_i nem észlelt, ezért valószínűségi változóként modellezzük; erre részletesen kitérünk alább.

Az egyváltozós lineáris modell értelmezéséhez tegyük fel, hogy tudjuk a paraméterek értékeit. Például, legyen: $\alpha = 1,\ \beta = 1.5$, tehát

$$y_i = 1 + 1.5x_i + \varepsilon_i, \quad i = 1, ..., n.$$

Az α paraméter megadja a függőváltozó várható értékét, ha a független változó 0. Megadja a regressziós egyenes metszéspontját az Oy tengellyel.

A β paraméter méri a független változó hatását a függőváltozóra, vagyis β megadja a függőváltozó változását, ha a független változó 1-gyel nő, feltételezve, hogy a hibaváltozó ugyanannyi marad (ceteris paribus).

2.1. A közönséges legkisebb négyzetek módszere

A legkisebb négyzetek módszerének (rövidítve OLS az angol *ordinary least squares*-ből) a lineáris modellre való alkalmazását hívjuk „közönséges"-nek. Ez a jelző megkülönböztetőül szolgál a módszer más alkalmazásaival szemben, például a súlyozott legkisebb négyzetek módszerével szemben, amelyről a 4. fejezetben lesz szó.

Az OLS módszer meghatározza azt az $y = a + bx$ egyenest (vagyis az a és b számokat), amelyre a $y_i - a - bx_i$ különbségek négyzete minimális. Tehát,

az a és b számokat úgy határozzuk meg, hogy minimalizáljuk a következő kifejezést:

$$S(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 .$$

A megoldást a következő kijelentés adja meg.

Állítás. Az egyedüli értékek, amelyek minimizálják az $S(a,b)$ kifejezést, a következők:

$$\begin{aligned} a &= \overline{y} - b\overline{x}, \\ b &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \end{aligned} \tag{2.1}$$

ahol $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ és $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Itt megjegyezzük a számítások ellenőrzése céljából, hogy fennállnak a következő egyszerű összefüggések:

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \overline{x}) y_i &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}), \\ \sum_{i=1}^{n} (x_i - \overline{x}) x_i &= \sum_{i=1}^{n} (x_i - \overline{x})^2 . \end{aligned}$$

Tehát az OLS módszer pontos képleteket határoz meg az ismeretlen paraméterek kiszámítására, amelyeket az y_i és x_i, $i = 1, ..., n$, megfigyelések segítségével tudunk kiszámítani. Az a-t és b-t meghatározó képleteket becslőfüggvényeknek, míg ezek konkrét értékeit becsléseknek nevezzük. Az

$$\begin{aligned} an + b\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i, \end{aligned}$$

egyenletek a minimum kiszámításához szükséges első rendű derivált feltételekből származnak, és normális egyenleteknek nevezzük.

Megjegyzés. Az OLS becslőfüggvény kapcsolódik az x_i és y_i változók kovarianciájához, mivel igazolható, hogy

$$b = \frac{cov(x_i, y_i)}{var(x_i)}.$$

A szóródási képen látható adatokra a becsléseket a következő (Eviews ökonometriai programban készített) táblázat tartalmazza:

Dependent Variable: RENDCYCO
Method: Least Squares
Sample: 1980:01 1999:12
Included observations: 240

Variable	Coefficient
C	-0.447481
RENDMARK	1.171128

Ez alapján a becsült modell a következő egyenletként írható:

$$RENDCYCO = \quad -0.447 \quad + \quad 1.171 \cdot RENDMARK + e.$$

2.1.1. Az OLS becslőfüggvény tulajdonságai

Az a és b OLS becslőfüggvények az α és β paraméterek becslésére szolgálnak. A becslés bizonyos feltételek mellett jól működik, vagyis jól megközelíti az α és β valódi értékeit, de ez nem áll fenn minden esetben.

Fontos megjegyzések

1. Az α és β paraméterek számok, míg az a és b becslőfüggvények valószínűségi változók (alább meglátjuk miért). A becslések szintén számok, amelyek a becslőfüggvények megvalósult értékei.

2. A gyakorlatban az α és β paraméterek valódi értékeit csak akkor tudhatjuk meg, ha a populáció összes megfigyelésével rendelkezünk, de ez általában nem lehetséges. Amit megtudunk róluk azok a becsléseik (például az OLS becsléseik), amelyeket bizonyos adatokra ki tudunk számítani.

A becslések a paraméterek valódi értékeinek a megközelítései. Nagyon fontos tudni, hogy mennyire jók ezek a megközelítések. Az alábbiakban olyan feltételeket tanulmányozunk, amelyek mellett fogalmat alkothatunk az OLS becslések minőségéről.

Az egyváltozós lineáris modell feltételei

A független változókra vonatkozó feltételek:

A1. A független változó n megfigyelése $x_1, ..., x_n$ determinisztikus (nem valószínűségi változó) értékekből áll, amelyekre $\sum_{i=1}^{n} (x_i - \overline{x})^2 > 0$.

A hibaváltozókra vonatkozó feltételek:

A2. Az $\varepsilon_1, ..., \varepsilon_n$ hibaváltozók nulla várható értékű valószínűségi változók:

$$E[\varepsilon_i] = 0, \quad i = 1, ..., n.$$

A3. (Homoszkedaszticitás) Az $\varepsilon_1, ..., \varepsilon_n$ hibaváltozók varianciája egyenlő:

$$var(\varepsilon_i) = \sigma^2, \quad i = 1, ..., n.$$

A4. (Nem korrelált hibaváltozók) Bármely két különböző hibaváltozó $\varepsilon_i, \varepsilon_j$ nem korrelált:

$$cov(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j, \ i, j = 1, ..., n.$$

A modellre vonatkozó feltétel:

A5. (A valódi modell lineáris) Az $y_1, ..., y_n$ függő változókat az

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, ..., n$$

lineáris modell határozza meg.

A hibaváltozók eloszlására vonatkozó feltétel:

A6. (Normál eloszlás) Az $\varepsilon_1, ..., \varepsilon_n$ hibaváltozók normális eloszlásúak.

Megjegyzések

1. Az A5 feltétel, hogy a valódi modell lineáris, elméleti szempontból nagy fontosságú, ugyanis azt jelenti, hogy a függő változókat a paraméterek valódi értékei határozzák meg. Ez a feltétel nélkül lehetetlen felmérni a becslőfüggvények minőségét.

2. A későbbi fejezetekben tanulmányozunk olyan becslési módszereket, amelyeket arra az esetre lehet használni, amikor az A3. (homoszkedaszticitás) és A4. (nem korrelált hibák) feltételek nem teljesülnek. Idősorok esetén az utóbbi feltétel általában nem teljesül, tehát az OLS módszer nem lesz minden szempontból hasznos.

3. A feltételek alapján $y_i \sim N(\alpha + \beta x_i, \sigma^2)$, $i = 1, ..., n$, tehát a függővváltozó megfigyelései valószínűségi változók.

4. Az előző megjegyzés alapján világos, hogy a becslőfüggvények valószínűségi változók, ugyanis függnek az y_i megfigyelésektől.

A feltételek következményei

1. Az A1, A5 felételek mellett

$$b = \beta + \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \varepsilon_i$$
$$a = \alpha + \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\overline{x}\,(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \varepsilon_i.$$

Bizonyítás. (csak b-re) A1 és a b képlete alapján (A1 alapján oszthatunk a nevezővel)

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})\, y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

A5 alapján $y_i = \alpha + \beta x_i + \varepsilon_i$, tehát

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})\,(\alpha + \beta x_i + \varepsilon_i)}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$
$$= \beta \frac{\sum_{i=1}^{n} (x_i - \overline{x})\, x_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} + \frac{\sum_{i=1}^{n} (x_i - \overline{x})\, \varepsilon_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.$$

Mivel a β melletti kifejezés 1, megkapjuk az eredményt.

2. Az A1, A2, A5 feltételek mellett az a, b OLS becslőfüggvények torzítatlanok, vagyis,

$$E[a] = \alpha \quad \text{és} \quad E[b] = \beta.$$

(Egy $\widehat{\theta}$ becslőfüggvény torzítatlan becslése a θ valódi paraméternek ha a becslőfüggvény torzulása, $E\left[\widehat{\theta}\right] - \theta$, nulla.)

Bizonyítás. (csak b-re) Az 1. Következmény alapján

$$E[b] = \beta + E\left[\sum_{i=1}^{n} \frac{(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \varepsilon_i \right].$$

A1 alapján a szögletes zárójelben csak az ε_i-k valószínűségi változók, tehát

$$E[b] = \beta + \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} E[\varepsilon_i] \overset{A2}{=} \beta.$$

3. Az A1-A5 feltételek mellett az a, b OLS becslőfüggvények varianciája:

$$var\left(a\right) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} \right)$$
$$var\left(b\right) = \frac{\sigma^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}.$$

Ezek a varianciák a nullához tartanak, ha

$$\mu_x = \lim_{n\to\infty} \overline{x} < \infty \qquad \text{és}$$
$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 < \infty.$$

Bizonyítás. (csak b-re) A variancia tulajdonsága és az 1. Következmény alapján

$$var\left(b\right) = var\left(b - \beta\right) = var\left(\sum_{i=1}^{n} \frac{\left(x_i - \overline{x}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} \varepsilon_i \right).$$

A jelölés egyszerűsítése kedvéért, legyen

$$u_i = \frac{\left(x_i - \overline{x}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2};$$

ekkor

$$var\left(b\right) = var\left(\sum_{i=1}^{n} u_i \varepsilon_i \right).$$

Mivel u_i determinisztikus, A4 alapján

$$var\left(b\right) = \sum_{i=1}^{n} u_i^2 var\left(\varepsilon_i\right).$$

A4 alapján

$$var\left(b\right) = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{\left(\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2\right)^2} \sigma^2 = \frac{\sigma^2/n}{\frac{1}{n}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} \underset{n\to\infty}{\to} 0.$$

Azokat a becslőfüggvényeket, amelyek torzítatlanok és a varianciájuk a nullához tart, konszisztensnek nevezzük. A 3. következményben szereplő határérték-feltételeket megközelítőleg lehet ellenőrizni, ha kiszámítjuk a $\mu_x \approx \overline{x}$ és $\sigma_x^2 \approx \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n}$ kifejezéseket. Ha $\overline{x}$ nem túl nagy és $\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n}$ nem egy nullához közel levő szám, akkor a feltételek megközelítőleg teljesülnek.

A konszisztencia-tulajdonság következménye az, hogy ha az adataink elég nagy számú megfigyeléssel rendelkeznek, akkor a becslések jól megközelítik a paraméterek valódi értékeit. Ezért, helyes becslési módszeren azt értjük, hogy a módszer becslőfüggvénye konszisztens.

4. Gauss-Markov hatékonysági tétel
Az A1-A5 feltételek mellett az a, b OLS becslőfüggvények az α, β paraméterek legjobb lineáris torzítatlan becslőfüggvényei, vagyis bármely

$$\widehat{\alpha} = \sum_{i=1}^{n} q_i y_i \quad \text{és} \quad \widehat{\beta} = \sum_{i=1}^{n} r_i y_i,$$

alakú becslőfüggvények, ahol q_i és r_i, $i = 1, ..., n$ determinisztikusak, nagyobb varianciával rendelkeznek mint a, b, vagyis,

$$var\left(\widehat{\alpha}\right) \geq var\left(a\right) \quad \text{és} \quad var\left(\widehat{\beta}\right) \geq var\left(b\right).$$

5. Az A1-A6 feltételek mellett az a, b OLS becslőfüggvények eloszlása:

$$a \sim N\left(\alpha, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n \left(x_i - \overline{x}\right)^2}\right]\right)$$
$$b \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n \left(x_i - \overline{x}\right)^2}\right).$$

Ez az 1. Következményből és A6-ból jön, mert a és b lineárisan függ az ε_i-ktől. E tulajdonság segítségével lehet tesztelni a paramétereket, amint meg fogjuk látni később.

Megjegyzések

1. Az a, b OLS becslőfüggvények is lineárisak. Például,

$$b = \sum_{i=1}^{n} u_i y_i, \quad \text{ahol} \quad u_i = \frac{x_i - \overline{x}}{\sum_{i=1}^n \left(x_i - \overline{x}\right)^2}.$$

2. Az a, b OLS becslőfüggvények varianciáját (lásd 3. Következmény) meg tudjuk határozni, amennyiben becsüljük a hibaváltozók varianciáját, σ^2-t.

3. A következményekben a becslőfüggvények három tulajdonságát említettük: torzítatlanság, konszisztencia és hatékonyság. Ezek közül a konszisztenciát tartjuk a legfontosabbnak, mert ez elégséges ahhoz, hogy a becslések jól megközelítsék a paraméterek valódi értékét, ha az adataink elég nagy számú megfigyelést tartalmaznak.

2.2. A becsült modell elemzése

A becslések kiszámítása után tanulmányozni szeretnénk még bizonyos kérdéseket:

- A becslések pontosságát, amihez meg kell határozni a becslőfüggvények varianciáját. A 3. Következmény szerint, ehhez a σ^2-t kell becsülni.
- A becsült modell mennyire képes magyarázni a függőváltozó változását, vagyis, a becsült összefüggés mennyire szoros.
- Teszteljük a modell paramétereit.

Ezeket a kérdéseket csak úgy tudjuk tanulmányozni, ha elfogadjuk, hogy az A1-A6 feltételek teljesülnek. A későbbiekben (4. fejezet) tanulmányozzuk, hogy hogyan tudjuk ellenőrizni a feltételeket.

2.2.1. A hibaváltozók varianciájának becslése

A hibaváltozók σ^2 varianciájának egy becslőfüggvénye

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2,$$

ahol $e_i = y_i - a - bx_i, \quad i = 1, ..., n$, mennyiségeket reziduumoknak hívjuk. Az s mennyiséget a regresszió standard hibájának hívjuk. SPSS-ben „Std. Error of the Estimate" néven szerepel.

A reziduumokat a hibaváltozók becslésének tekintjük, mert

$$e_i = y_i - a - bx_i \approx y_i - \alpha - \beta x_i = \varepsilon_i$$

ha n elég nagy. A $\sigma^2 = var(\varepsilon_i) = E[\varepsilon_i^2] - E[\varepsilon_i]^2 = E[\varepsilon_i^2]$ tény magyarázattal szolgál az s^2 fenti definíciójára, ugyanis ez alapján a σ^2-t a hibaváltozók négyzetének átlagával tudjuk becsülni. Továbbá, a hibaváltozókat a reziduumokkal becsüljük, és az $\frac{1}{n-2}\sum_{i=1}^{n} e_i^2$ kifejezés nagy n esetén majdnem megegyezik az $\frac{1}{n}\sum_{i=1}^{n} e_i^2$ átlaggal. Itt fontos megjegyezni, hogy s^2 a σ^2 torzítatlan becslőfüggvénye, vagyis $E[s^2] = \sigma^2$; ez indokolja az s^2 képletét. A $\sum_{i=1}^{n} e_i^2$ összeget a reziduumok négyzetösszegének nevezzük.

2.2.2. A becsült összefüggés szorossága

Tegyük fel, hogy a megfigyeléseink (x_1, y_1), ... , (x_n, y_n), és ezekre egy lineáris modellt becsültünk. Meg szeretnénk állapítani, hogy mennyire szoros a becsült összefüggés.

A becsült összefüggést össze tudjuk hasonlítani egy másik becsült összefüggéssel. Ehhez becsüljük azt a modellt, amelynek csak egy konstans együtthatója van (nincs benne független változó):

$$y_i = \gamma + \omega_i, \quad i = 1, ..., n$$

ahol γ a konstans együttható és ω_i a hibaváltozó. A γ OLS becslőfüggvénye (alkalmazva a (2.1) képletet, amelyben az összes független változót 0-nak vesszük): $c = \overline{y}$.

Tehát a reziduumok négyzetösszege $\sum_{i=1}^{n} (y_i - \overline{y})^2$, ami nem kisebb mint $\sum_{i=1}^{n} e_i^2$, az eredeti modell reziduumainak négyzet-összegénél, mert a legkisebb négyzetek módszere azt jelenti, hogy

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

a lehető legkisebb, és ha a és b helyett bármilyen más számot helyettesítünk, az eredmény nagyobb lesz. Ugyanez történik, ha a helyett $c = \overline{y}$-t és b helyett 0-t helyettesítünk, amivel pontosan a $\sum_{i=1}^{n} (y_i - \overline{y})^2$ mennyiséget kapjuk.

Minél kisebb a reziduumok négyzetösszege, annál szorosabb a becsült összefüggés, mert annál közelebb helyezkednek el a megfigyelések a regressziós egyeneshez. Definiáljuk a két modell reziduum-négyzetösszegeinek a normalizált különbségét:

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}.$$

Ezt a számot determinancia-együtthatónak vagy egyszerűen R^2-nek hívjuk. Mivel a különbséget elosztottuk a nagyobbik reziduum-négyzetösszeggel, ezért $0 \leq R^2 \leq 1$.

A fenti gondolatmenet alapján minél nagyobb a számláló a nevezőhöz képest, annál szorosabb a becsült összefüggés. Tehát ha R^2 közel van 1-hez akkor levonhatjuk a következtetést, hogy szoros a becsült összefüggés. Ha R^2 a 0-hoz van közel, akkor meglehetősen tág a becsült összefüggés.

Megjegyzések

1. Be lehet bizonyítani, hogy

$$R^2 = [corr\,(x, y)]^2\,,$$

   vagyis, a determinancia-együttható megegyezik a függő- és a független változók közötti korreláció négyzetével.

2. Sok esetben R^2 azért van közel 0-hoz, mert az A5 (A valódi modell lineáris) feltétel nem teljesül, vagyis a változók közötti összefüggés nem lineáris. Ezért fontos becsléskor szóródási képet készíteni.

Példa: nettó hozam

A 2. megjegyzéssel kapcsolatban megvizsgálhatjuk, hogy mi történik, ha a nettó hozamok példában a független változót logaritmusban vesszük. Az első (Eviews-ban készült) táblázat bemutatja az eredeti modell OLS becslését és R^2-ét, vagyis ahol RENDCYCO a függővváltozó és RENDMARK a független változó, míg a második táblázat egy olyan OLS becslést mutat be ahol ugyancsak a RENDCYCO a függővváltozó de $\ln(\text{RENDMARK} + 30)$ a független változó. R^2 kisebb a második esetben, 0.41, az eredetihez képest, ami 0.5.

Dependent Variable: RENDCYCO	
Method: Least Squares	
Sample: 1980:01 1999:12	
Included observations: 240	
Variable	Coefficient
C	-0.447481
RENDMARK	1.171128
R-squared	0.503480

Dependent Variable: RENDCYCO	
Method: Least Squares	
Sample: 1980:01 1999:12	
Included observations: 240	
Variable	Coefficient
C	-74.66311
LNRENDMARK	22.04392
R-squared	0.414718

2.2.3. A paraméterek tesztelése

Vegyük megint a nettó hozamok példáját. Láttuk, hogy az összes piaci nettó hozam hatása a ciklikus termékek piaci nettó hozamára 1.17. Szeretnénk vizsgálni, hogy ez az érték szignifikánsan különbözik-e 1-től. Ezt a kérdést egy statisztikai teszttel (próbával) tudjuk megválaszolni. A teszt a becslőfüggvények eloszlását használja.

A nettó hozamok példájához vegyük a $H_0 : \beta = 1$ nullhipotézist és a $H_1 : \beta \neq 1$ vagy $H_1 : \beta < 1$ alternatív hipotéziseket. Aszerint, hogy az alternatív hipotézist hogyan definiáljuk, kétféle teszt létezik: kétoldalas és egyoldalas.

Egy kétoldalas tesztnél a H_1-ben a paraméter a tesztelt érték mindkét oldalán lehet. Például, $H_0 : \beta = 1$ és $H_1 : \beta \neq 1$. Ebben az esetben a H_1: $\beta < 1$ vagy $\beta > 1$.

Egy egyoldalas tesztnél a H_1-ben a paraméter a tesztelt érték egyik oldalán lehet. Például, $H_0 : \beta = 1$ és $H_1 : \beta < 1$ vagy $H_0 : \beta \geq 1$ és $H_1 : \beta < 1$.

Általában a teszteknél meg kell határozni a tesztstatisztikát, amit egy becslőfüggvényhez hasonlóan a megfigyelésekből számítunk ki, és amelynek az eloszlását meg tudjuk határozni a nullhipotézis alapján.

A β meredekségi együttható tesztelése

Példaként a β tesztelését tárgyaljuk; a konstans együttható (α) tesztelése hasonló. A többváltozós lineáris modell tárgyalásakor majd tanulmányozzuk több paraméter egyidejű tesztelését is. A β meredekségi együttható tesztelése az alábbi tényen alapul.

Állítás. Az A1-A6 feltételek mellett

$$t_b = \frac{b - \beta}{\sqrt{var\,(b)}} \sim t\,(n-2)\,,$$

vagyis, a t_b statisztika t eloszlású $n-2$ szabadsági fokkal.

Az $s_b = \sqrt{var\,(b)}$ értéket a b becslés standard hibájának nevezzük. Emlékezzünk vissza, hogy a t eloszlás sűrűségfüggvénye szimmetrikus 0 körül, és ahogy a szabadsági foka ∞-hez tart, megközelíti a standard normális eloszlást. A t_b kifejezésben az egyedüli ismeretlen a β, ugyanis b és $var\,(b)$ becsülhetőek a megfigyelések alapján. A nullhipotézis szerint viszont $H_0 : \beta = 1$, tehát felhasználva ezt kiszámítható t_b.

Kétoldalas teszthez egy $(-c_\alpha, c_\alpha)$ elfogadási tartományt határozzunk meg, amelyre teljesül a

$$P\,(-c_\alpha < t_b < c_\alpha) = 1 - \alpha$$

összefüggés, ahol α a szignifikancia szint (általában 0.05 vagy 0.01). A c_α kritikus értéket tábázatokból vagy statisztikai programokból (vagy internetről) kaphatjuk meg.

Azt mondjuk, hogy a nullhipotézist elutasítjuk α szignifikancia szinten ha $|t_b| > c_\alpha$; különben azt mondjuk, hogy a nullhipotézist nem utasítjuk el α szignifikancia szinten.

Egy egyoldalas teszthez, amelynek alternatív hipotézise $H_1 : \beta < 1$ (attól függetlenül, hogy $H_0 : \beta = 1$ vagy $H_0 : \beta \geq 1$), az elfogadási tartományt (c_α, ∞) formában határozzuk meg, amelyre teljesül:

$$P\left(t_b > c_\alpha\right) = 1 - \alpha.$$

A nullhipotézist elutasítjuk α szignifikancia szinten ha $t_b < c_\alpha$, és nem utasítjuk el, ha $t_b > c_\alpha$.

Szignifikancia-teszt

A b becslés szignifikancia-tesztje az a teszt, amire $H_0 : \beta = 0$ és $H_1 : \beta \neq 0$. Ez a teszt minden becslési folyamat velejárója, és arra ad választ, hogy a független változó hatása szignifikáns-e (vagyis elég erős-e) a függő változóra.

Ha elutasítjuk a nullhipotézist, akkor azt a következtetést vonhatjuk le, hogy a hatás szignifikáns, különben azt a következtetést vonjuk le, hogy a hatás nem szignifikáns. A t-statisztika értékeit a szignifikancia-tesztre megkaphatjuk az SPSS-ből.

p-érték

Egy tesztstatisztika p-értéke az a legkisebb szignifikancia szint, amire a nullhipotézist elutasítjuk. Ez tulajdonképpen a tesztstatisztika értékének „megfelelő" szignifikancia szint vagy valószínűség.

A p-értéket megkaphatjuk SPSS-ből „Sig." néven. A p-érték segítségével az elfogadási tartomány meghatározása nelkül tudunk dönteni a nullhipotézisről, tehát nincs szükség a kritikus értékek táblázatára.

Tegyük fel, hogy az SPSS megad egy bizonyos p p-értéket, és vegyünk egy α szignifikancia szintet. Ekkor:
- ha $\alpha < p$ nem utasítjuk el H_0-t (mivel α kisebb, mint p, ami a legkisebb szignifikancia szint, amelyre H_0-t elutasítjuk,
- ha $\alpha > p$ elutasítjuk H_0-t, mivel minden p-nél nagyobb szignifikancia szintre elutasítjuk.

Példa. A következő SPSS-táblázat a RENDCYCO és RENDMARK modelljének OLS becslés-eredményét, a t-statisztikákat és a p-értékeket mutatja:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.447	.363		-1.233	.219
	RENDMARK	1.171	.075	.710	15.535	.000

a. Dependent Variable: RENDCYCO

A becsült modell:

$$RENDCYCO = \underset{(0.363)}{-0.447} + \underset{(0.075)}{1.171} \cdot RENDMARK + e.$$

Legyen $\alpha = 0.05$ a szignifikancia-szint. Ekkor $0.05 < 0.219$, ami a konstansnak megfelelő p-érték, tehát a konstans együttható becslése nem szignifikáns. Mivel $0.05 > 0.000$, ami a meredekségi együtthatónak megfelelő p-érték, ezért a meredekségi együttható becslése szignifikáns.

Intervallum-becslés

Az együtthatóknak szerkesztett teszt-statisztikák alapján meg tudunk határozni egy olyan intervallumot, amely nagy valószínűséggel tartalmazza az együttható valódi értékét. Az α szignifikancia szint alapján meg tudjuk határozni a c_α kritikus értéket, amire

$$P\left(|t_b| < c_\alpha\right) = 1 - \alpha.$$

Ez alapján

$$P\left(b - c_\alpha s_b < \beta < b + c_\alpha s_b\right) = 1 - \alpha,$$

ami meghatároz egy intervallumot, amely $1-\alpha$ valószínűséggel tartalmazza a β-t. Ezt a $(b - c_\alpha s_b, b + c_\alpha s_b)$ intervallumot a β konfidencia intervallumának vagy intervallum-becslésének nevezzük.

2.3. Gyakorlatok

1. Legyen $x_i = y_{i-1}$, $y_0 = 1$ és

$$y_i = \alpha + \beta y_{i-1} + \varepsilon_i, \quad i = 1, ..., n,$$

ahol ε_i teljesíti az $E\left[\varepsilon_i\right] = 0$, $E\left[\varepsilon_i \varepsilon_j\right] = 0$ ha $i \neq j$, $E\left[\varepsilon_i^2\right] = \sigma^2$ feltételeket. Teljesül-e az A1 feltétel?

2. Egy egyváltozós lineáris modellben

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, ..., n,$$

ahol $\varepsilon_i = \omega_i - \frac{1}{2}\omega_{i-1}$ és az ω_i vv teljesíti az $E[\omega_i] = 0$, $E[\omega_i\omega_j] = 0$ ha $i \neq j$, $E[\omega_i^2] = \sigma^2$ feltételeket. Teljesül-e az A3 és az A4 feltétel?

3. Az y függő- és x független változókra a következő megfigyeléseink vannak $\begin{pmatrix} y: & 9 & 7 & 6 & 2 & 1 \\ x: & 2 & 4 & 3 & 6 & 10 \end{pmatrix}$.

   (a) Határozzuk meg a konstans és a meredekségi együttható OLS becslését. Mennyit változik y ha x 1-gyel nő?

   (b) Becsüljük az OLS becslések standard hibáit.

   (c) Számítsuk ki az R^2-t. Kommentáljuk a kapott eredményt.

   (d) Teszteljük a becslések szignifikanciáját.

   (e) Szerkesszünk 95%-os konfidencia intervallumokat a konstans és a meredekségi együtthatónak. Mennyire informatívak ezek az intervallum-becslések?

3. fejezet

A többváltozós lineáris modell elemzése

Több példában is láttuk, hogy a függővváltozót nem csak egy változó befolyásolja. Ez így volt az autókeresleti modellnél is. Ebben a modellben

$$k = \beta_1 + \beta_2 a + \beta_3 m + \beta_4 \ell + \varepsilon,$$

a keresletet (k) az áron (a) kívül még az autó mérete (m) és az autó motorjának lóereje (ℓ) határozza meg. Mivel ezek többváltozós modellek, a továbbiakban a többváltozós lineáris modell becslését tanulmányozzuk.

A többváltozós lineáris modellt a következő egyenletekkel fogjuk jelölni

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, ..., n.$$

A sok írás elkerülése végett gyakran vektor/mátrix jelölést fogunk használni:

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon},$$

vagyis,

$$y = X\beta + \varepsilon, \quad \text{ahol}$$

- y egy n elemű oszlopvektor ami tartalmazza a függővváltozó megfigyeléseit,
- X egy $n \times k$ mátrix ami tartalmazza a független változók megfigyeléseit és egy 1-esekből álló oszlopot a konstans együttható miatt,
- β egy k elemű oszlopvektor ami az ismeretlen paramétereket tartalmazza,
- ε egy n elemű oszlopvektor ami a hibaváltozókat tartalmazza.

3.1. A többváltozós lineáris modell paramétereinek értelmezése

A többváltozós lineáris modellt írhatjuk

$$y = \beta_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

formában is, ahol y és $x_2, ..., x_k$ változókra az y_i és $x_{2i}, ..., x_{ki}$, $i = 1, 2, ..., n$ megfigyeléseink vannak.

A β_j ($j \in \{2, 3, ..., k\}$) paramétert az x_j független változó parciális hatásának nevezzük mert

$$\frac{\partial y}{\partial x_j} = \beta_j.$$

A β_j paraméter azt fejezi ki, hogy mennyit változik a függővváltozó az x_j változó egységnyi növekedésére, ha a többi változó ugyanaz marad. Ez az angol szaknyelvben ceteris paribus-ként szerepel.

Ezt a lehetséges félreértések végett fontos tisztázni, ugyanis a gyakorlatban általában ritkán fordul elő, hogy ha változik egy megfigyelés akkor csak egyetlen független változó változik; ez azért van így, mert a független változók általában korrelálnak egymással.

Vegyük például az autókereslet-modellt:

$$k = \beta_1 + \beta_2 a + \beta_3 m + \beta_4 \ell + \varepsilon,$$

ahol a keresletet (k) az áron (a) kívül még az autó mérete (m) és a lóerő (ℓ) határozza meg. Ekkor β_2 az ár parciális hatása a keresletre, vagyis azt fejezi ki, hogy mennyit változik a kereslet, ha csak az ár változik. Ez a gyakorlatban is megfigyelhető árkedvezmények esetén. Viszont a lóerő parciális hatását a keresletre nehezebb megfigyelni a gyakorlatban, mert ha két különböző lóerejű autótípust veszünk, akkor valószínűleg az áraik is különböznek, tehát a kereslet változása nemcsak a parciális hatást tükrözi.

3.2. A modell becslése OLS-sel

Az egyváltozós lineáris modellhez hasonlóan, a többváltozós lineáris modellt is lehet becsülni a közönséges legkisebb négyzetek (OLS) módszerével. Ehhez minimalizáljuk a reziduumok négyzetösszegét:

$$S(b) = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_{2i} - ... - b_k x_{ki})^2 .$$

Mivel az $S(b)$ függvény mindegyik b_j-ben másodfokú, $j = 1, ..., k$, az $S(b)$ minimumát a parciális deriváltak 0-pontjainak meghatározásával kapjuk, vagyis ha megoldjuk $b_1, b_2, ..., b_k$ szerint a

$$\frac{\partial S(b)}{\partial b} = \begin{pmatrix} \frac{\partial S(b)}{\partial b_1} \\ \vdots \\ \frac{\partial S(b)}{\partial b_k} \end{pmatrix} = 0$$

egyenletrendszert. Az $S(b)$ függvény b_j szerinti parciális deriváltja

$$\frac{\partial S(b)}{\partial b_j} = \sum_{i=1}^{n} (-2) x_{ji} (y_i - b_1 - b_2 x_{2i} - ... - b_k x_{ki})$$

$$= -2 \begin{pmatrix} x_{j1} & \cdots & x_{jn} \end{pmatrix} \begin{bmatrix} y_1 - b_1 - b_2 x_{21} - ... - b_k x_{k1} \\ \vdots \\ y_n - b_1 - b_2 x_{2n} - ... - b_k x_{kn} \end{bmatrix}$$
$$= -2 \begin{pmatrix} x_{j1} & \cdots & x_{jn} \end{pmatrix} (y - Xb),$$

tehát

$$\frac{\partial S(b)}{\partial b} = \begin{pmatrix} \frac{\partial S(b)}{\partial b_1} \\ \vdots \\ \frac{\partial S(b)}{\partial b_k} \end{pmatrix} = -2 \begin{pmatrix} 1 & \cdots & 1 \\ x_{21} & \cdots & x_{2n} \\ \vdots & & \vdots \\ x_{k1} & & x_{kn} \end{pmatrix} (y - Xb)$$
$$= -2X'(y - Xb).$$

A b kiszámításához megoldjuk a

$$X'(y - Xb) = 0, \quad \text{vagyis} \quad X'y = X'Xb$$

egyenletet (tulajdonképpen egyenletrendszert). Tehát ebből megkapjuk a β paraméter-vektor OLS becslőfüggvényét:

$$b = (X'X)^{-1} X'y,$$

ha az $X'X$ mátrix invertálható.

Az $X'X$ mátrix

$$X'X = \begin{pmatrix} 1 & \cdots & 1 \\ x_{21} & \cdots & x_{2n} \\ \vdots & & \vdots \\ x_{k1} & & x_{kn} \end{pmatrix} \begin{pmatrix} 1 & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{2n} & \cdots & x_{kn} \end{pmatrix}$$
$$= \begin{pmatrix} n & \sum_{i=1}^n x_{2i} & \cdots & \sum_{i=1}^n x_{ki} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{2i}^2 & & \sum_{i=1}^n x_{2i}x_{ki} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_{ki} & \sum_{i=1}^n x_{2i}x_{ki} & \cdots & \sum_{i=1}^n x_{ki}^2 \end{pmatrix}.$$

Hasonlóan,

$$X'y = \begin{pmatrix} 1 & \cdots & 1 \\ x_{21} & \cdots & x_{2n} \\ \vdots & & \vdots \\ x_{k1} & & x_{kn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{2i}y_i \\ \vdots \\ \sum_{i=1}^n x_{ki}y_i \end{pmatrix}.$$

Az egyváltozós modellhez hasonlóan itt is bizonyos feltételek mellett vizsgáljuk az OLS becslés helyességét.

3.2.1. A többváltozós lineáris modell feltételei

A független változókra vonatkozó feltételek:

A1. A független változók megfigyeléseit tartalmazó X mátrix determinisztikus (nem vv) elemekből áll, amelyekre $n \geq k$ és $X'X$ invertálható.

A hibaváltozókra vonatkozó feltételek:

A2. Az $\varepsilon_1, ..., \varepsilon_n$ hibaváltozók nulla várható értékű valószínűségi változók.

A3. (Homoszkedaszticitás) Az $\varepsilon_1, ..., \varepsilon_n$ hibaváltozók varianciája egyenlő:

$$var\,(\varepsilon_i) = \sigma^2, \quad i = 1, ..., n.$$

A4. (Nem korrelált hibaváltozók) Bármely két különböző hibaváltozó $\varepsilon_i, \varepsilon_j$ nem korrelált:

$$cov\,(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j,\ i, j = 1, ..., n.$$

A modellre vonatkozó feltétel:

A5. (A valódi modell lineáris) Az $y = (y_1, ..., y_n)$ függővváltozókat az

$$y = X\beta + \varepsilon.$$

lineáris modell határozza meg.
A hibaváltozók eloszlására vonatkozó feltétel:
A6. (Normál eloszlás) Az $\varepsilon_1, ..., \varepsilon_n$ hibaváltozók normál eloszlásúak.

3.2.2. A feltételek következményei

1. A1, A5 feltételek mellett

$$b = \beta + (X'X)^{-1} X'\varepsilon.$$

Bizonyítás. A1 alapján a b becslőfüggvény kiszámítható:

$$b = (X'X)^{-1} X'y.$$

A5 alapján $y = X\beta + \varepsilon$, tehát

$$\begin{aligned} b &= (X'X)^{-1} X' (X\beta + \varepsilon) \\ &= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'\varepsilon \\ &= \beta + (X'X)^{-1} X'\varepsilon. \end{aligned}$$

2. A1, A2, A5 feltételek mellett b a β paraméter-vektor torzítatlan becslőfüggvénye, vagyis

$$E[b] = \beta.$$

Bizonyítás. Az 1. Következmény és A2 alapján,

$$E[b] = \beta + E\left[(X'X)^{-1} X'\varepsilon\right].$$

A1 alapján a szögletes zárójelben csak ε vv, tehát

$$E[b] = \beta + (X'X)^{-1} X'E[\varepsilon],$$

ami A2 alapján egyenlő β-val.

3. Az A1-A5 feltételek mellett a b OLS becslőfüggvény variancia mátrixa (vagy kovariancia mátrixa):

$$var(b) = \sigma^2 (X'X)^{-1}.$$

A variancia mátrix a 0 mátrixhoz tart, ha $\lim_{n\to\infty} \left(\frac{1}{n} X'X\right)$ invertálható.

Bizonyítás. A b vektor variancia mátrixa definíció szerint

$$var\left(b\right)=E\left[\left(b-E\left[b\right]\right)\left(b-E\left[b\right]\right)'\right].$$

A 2. Következmény alapján

$$var\left(b\right)=E\left[\left(b-\beta\right)\left(b-\beta\right)'\right].$$

Az 1. Következmény alapján

$$\begin{aligned}var\left(b\right)&=E\left[\left(\left(X'X\right)^{-1}X'\varepsilon\right)\left(\left(X'X\right)^{-1}X'\varepsilon\right)'\right]\\&=E\left[\left(X'X\right)^{-1}X'\varepsilon\varepsilon'X\left(X'X\right)^{-1}\right].\end{aligned}$$

A1 alapján a szögletes zárójelben csak ε vv, tehát

$$\begin{aligned}var\left(b\right)&=E\left[\left(X'X\right)^{-1}X'\varepsilon\varepsilon'X\left(X'X\right)^{-1}\right]\\&=\left(X'X\right)^{-1}X'E\left[\varepsilon\varepsilon'\right]X\left(X'X\right)^{-1}.\end{aligned}$$

A3, A4 alapján $E\left[\varepsilon\varepsilon'\right]=\sigma^2 I$ (lásd 1. Megjegyzés), tehát

$$\begin{aligned}var\left(b\right)&=\left(X'X\right)^{-1}X'\sigma^2 IX\left(X'X\right)^{-1}\\&=\sigma^2\left(X'X\right)^{-1}XX\left(X'X\right)^{-1}=\sigma^2\left(X'X\right)^{-1}.\end{aligned}$$

A variancia mátrix a 0 mátrixhoz tart, mivel

$$var\left(b\right)=\frac{\sigma^2}{n}\left(\frac{1}{n}X'X\right)^{-1}\underset{n\to\infty}{\to}\begin{pmatrix}0&\ldots&0\\&\ddots&\\0&\ldots&0\end{pmatrix}.$$

Megjegyzések

(a) A b OLS becslőfüggvény **helyes** mivel **konszisztens**, vagyis, torzítatlan (2. Következmény) és a varianciája a 0-hoz tart (3. Következmény).

(b) A b_j, ami a β paraméter-vektor β_j elemének a becslőfüggvénye, varianciáját a $var\left(b\right)$ variancia mátrix (j,j) elemeként kapjuk meg.

4. Gauss-Markov hatékonysági tétel

Az A1-A5 feltételek mellett a b OLS becslőfüggvény a β paraméter-vektor legjobb lineáris torzítatlan becslőfüggvénye abban az értelemben, hogy bármely

$$\widehat{\beta}=Ay,$$

alakú $\widehat{\beta}$ torzítatlan becslőfüggvény, ahol A egy $k \times n$ determinisztikus mátrix, varianciája nagyobb, vagyis,

$$var\left(\widehat{\beta}_j\right) \geq var\left(b_j\right)$$

mindegyik $j = 1, ..., k$-ra.

5. A1-A6 mellett a b OLS becslőfüggvény eloszlása:

$$b \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right).$$

Az 1. Következmény alapján b az ε_i hibaváltozók lineáris kombinációja, amikről A6 alapján tudjuk, hogy normál eloszlásúak. Ez a tulajdonság a teszt-statisztikák eloszlásának meghatározására szolgál.

A továbbiakban két számszerű példát tárgyalunk, amelyek közül az első egy 5 megfigyelésből álló példa, ezért minden idevágó mennyiséget könnyen ki lehet számítani, míg a második egy valószerű példa sok megfigyeléssel.

3.2.3. Első példa

Az y függő- és x_2, x_3 független változókra a következő megfigyeléseink vannak:

$y:$	2	3	5	8	12
$x_2:$	1	8	16	24	12
$x_3:$	2	4	3	6	10

Határozzuk meg a konstans és a két meredekségi együttható OLS becslését.
Megoldás. Legyen

$$y = \begin{pmatrix} 2 \\ 3 \\ 5 \\ 8 \\ 12 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 8 & 4 \\ 1 & 16 & 3 \\ 1 & 24 & 6 \\ 1 & 12 & 10 \end{pmatrix}.$$

Ekkor az OLS-becslőfüggvény $b = \left(X'X\right)^{-1} X'y$.

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 8 & 16 & 24 & 12 \\ 2 & 4 & 3 & 6 & 10 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 1 & 8 & 4 \\ 1 & 16 & 3 \\ 1 & 24 & 6 \\ 1 & 12 & 10 \end{pmatrix} = \begin{pmatrix} 5 & 61 & 25 \\ 61 & 1041 & 346 \\ 25 & 346 & 165 \end{pmatrix},$$

$$(X'X)^{-1} = \begin{pmatrix} 1.0215 & -0.0278 & -0.0965 \\ -0.0278 & 0.0039 & -0.0040 \\ -0.0965 & -0.0040 & 0.0291 \end{pmatrix},$$

$$X'y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 8 & 16 & 24 & 12 \\ 2 & 4 & 3 & 6 & 10 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 5 \\ 8 \\ 12 \end{pmatrix} = \begin{pmatrix} 30 \\ 442 \\ 199 \end{pmatrix}.$$

A becslések vektora

$$b = \begin{pmatrix} 1.0215 & -0.0278 & -0.0965 \\ -0.0278 & 0.0039 & -0.0040 \\ -0.0965 & -0.0040 & 0.0291 \end{pmatrix} \begin{pmatrix} 30 \\ 442 \\ 199 \end{pmatrix} = \begin{pmatrix} -0.846 \\ 0.094 \\ 1.125 \end{pmatrix},$$

ahol -0.846 a konstans és 0.094 és 1.125 a két meredekségi együttható OLS becslése.

3.2.4. Második példa

Ebben a példában azt vizsgáljuk, hogy a banki alkalmazottak tanulási éveinek a száma és a kezdőfizetésük hogyan befolyásolják a fizetésüket. Ehhez legyen
y a fizetések logaritmusa,
x_2 a tanulási évek száma és
x_3 a kezdőfizetések logaritmusa.

Az adatok a következőképpen vannak megadva:

$$\begin{pmatrix} \sum_{i=1}^n y_i^2 & \sum_{i=1}^n y_i & \sum_{i=1}^n y_i x_{2i} & \sum_{i=1}^n y_i x_{3i} \\ & n & \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{3i} \\ & & \sum_{i=1}^n x_{2i}^2 & \sum_{i=1}^n x_{2i} x_{3i} \\ & & & \sum_{i=1}^n x_{3i}^2 \end{pmatrix}$$

$$= \begin{pmatrix} 50917 & 4909 & 66609 & 47527 \\ & 474 & 6395 & 4583 \\ & & 90215 & 62165 \\ & & & 44376 \end{pmatrix}.$$

A β paraméter-vektor becslőfüggvénye $b = (X'X)^{-1} X'y$, ahol:

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{3i} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{2i}^2 & \sum_{i=1}^n x_{2i} x_{3i} \\ \sum_{i=1}^n x_{3i} & \sum_{i=1}^n x_{2i} x_{3i} & \sum_{i=1}^n x_{3i}^2 \end{pmatrix} = \begin{pmatrix} 474 & 6395 & 4583 \\ 6395 & 90215 & 62165 \\ 4583 & 62165 & 44376 \end{pmatrix},$$

$$X'y = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{2i}y_i \\ \sum_{i=1}^n x_{3i}y_i \end{pmatrix} = \begin{pmatrix} 4909 \\ 66609 \\ 47527 \end{pmatrix}.$$

Tehát

$$b = \begin{pmatrix} 474 & 6395 & 4583 \\ 6395 & 90215 & 62165 \\ 4583 & 62165 & 44376 \end{pmatrix}^{-1} \begin{pmatrix} 4909 \\ 66609 \\ 47527 \end{pmatrix} = \begin{pmatrix} 1.696 \\ 0.023 \\ 0.863 \end{pmatrix}.$$

Ez alapján a becsült modell: $y = 1.696 + 0.023x_2 + 0.863x_3 + e$.

Milyen hatással van a tanulási évek száma a banki fizetésekre? A tanulási évek száma (x_2) becsült együtthatója pozitív előjelű, ezért a tanulási évek számának a növekedése a banki fizetés növekedését eredményezi.

3.3. A becsült modell elemzése

Akárcsak az egyváltozós lineáris modellnél, a becslések kiszámítása után tanulmányozni szeretnénk még bizonyos kérdéseket, vagyis:

- a becslések pontosságát, amihez a σ^2-t kell becsülni,
- a becsült modell mennyire képes magyarázni a függővváltozó változását, vagyis, a becsült összefüggés mennyire szoros,
- teszteljük a modell paramétereit.

Ezen kérdések tanulmányozásához elfogadjuk, hogy az A1-A6 feltételek teljesülnek.

3.3.1. A hibaváltozók varianciájának becslése

Akárcsak az egyváltozós lineáris modellnél, σ^2 becslőfüggvénye a $\sum_{i=1}^n e_i^2$ reziduumok négyzetösszegén alapul, ahol

$$e_i = y_i - b_1 - b_2x_{2i} - ... - b_kx_{ki}, \quad i = 1, ..., n$$

a reziduumok. A σ^2 egy torzítatlan becslőfüggvénye a következő:

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-k}.$$

Az s mennyiséget, ami az s^2 négyzetgyöke, a regresszió standard hibájának nevezik. Az alábbi állítás szerint az A1-A5 feltételek elégségesek ahhoz, hogy s^2 a σ^2 torzítatlan becslőfüggvénye legyen.

Állítás. A1-A5 mellett $E\left[\sum_{i=1}^{n} e_i^2\right] = (n-k)\,\sigma^2$, vagyis s^2 a σ^2 torzítatlan becslőfüggvénye.

3.3.2. A becsült összefüggés szorossága

Akárcsak az egyváltozós lineáris modellnél, becsléskor szeretnénk tudni, hogy mennyire szoros a becsült összefüggés. Ehhez a becslést összehasonlítjuk a csak egy konstans együtthatót tartalmazó (nincs benne független változó) modell becslésével. Ez a modell:

$$y_i = \gamma + \omega_i, \quad i = 1, ..., n$$

ahol γ a konstans együttható és ω_i a hibaváltozó. A γ OLS becslőfüggvénye (alkalmazzuk a képletet): $c = \overline{y}$. A reziduumok négyzetösszege $\sum_{i=1}^{n} (y_i - \overline{y})^2$, ami minden esetben nagyobb (vagy egyenlő) mint $\sum_{i=1}^{n} e_i^2$, az eredeti modell reziduumainak négyzet-összegénél.

Az egyváltozós lineáris modellhez hasonlóan, definiáljuk a két modell reziduum-négyzetösszegeinek a normalizált különbségét, amit determinancia-együtthatónak vagy egyszerűen R^2-nek hívnak:

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \quad 0 \leq R^2 \leq 1.$$

Minél nagyobb a számláló a nevezőhöz képest, annál szorosabb a becsült összefüggés. Tehát ha R^2 közel van 1-hez akkor levonhatjuk a következtetést, hogy szoros a becsült összefüggés. Ha R^2 a 0-hoz van közel, akkor meglehetősen tág a becsült összefüggés.

A példák folytatása

Becsüljük az OLS becslések standard hibáit és számítsuk ki az R^2-t.

Megoldás. **(Első példa)** A b varianciájának becslése $\widehat{var\,(b)} = s^2\,(X'X)^{-1}$ ahol $s^2 = \dfrac{\sum e_i^2}{n-k}$.

$$e = y - Xb = \begin{pmatrix} 2 \\ 3 \\ 5 \\ 8 \\ 12 \end{pmatrix} - \begin{pmatrix} 1 & 1 & 2 \\ 1 & 8 & 4 \\ 1 & 16 & 3 \\ 1 & 24 & 6 \\ 1 & 12 & 10 \end{pmatrix} \begin{pmatrix} -0.846 \\ 0.094 \\ 1.125 \end{pmatrix} = \begin{pmatrix} 0.502 \\ -1.406 \\ 0.967 \\ -0.160 \\ 0.468 \end{pmatrix}$$

$\sum_{i=1}^{n} e_i^2 = 3.409$.

Tehát $s^2 = \frac{\sum e_i^2}{n-k} = \frac{1}{5-3}3.409 = 1.704$.

A b varianciájának becslése: $\widehat{var\,(b)} = 1.704 \cdot \begin{pmatrix} 1.0215 & -0.0278 & -0.0965 \\ -0.0278 & 0.0039 & -0.0040 \\ -0.0965 & -0.0040 & 0.0291 \end{pmatrix}$.

Az egyéni varianciák becslése

$$\widehat{var\,(b_1)} = 1.704 \cdot 1.0215 = 1.741$$
$$\widehat{var\,(b_2)} = 1.704 \cdot 0.0039 = 0.007$$
$$\widehat{var\,(b_3)} = 1.704 \cdot 0.0291 = 0.049$$

és a standard hibák

$$s_{b_1} = \sqrt{1.741} = 1.319$$
$$s_{b_2} = \sqrt{0.007} = 0.084$$
$$s_{b_3} = \sqrt{0.049} = 0.221.$$

Az R^2 determinancia-együttható:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{3.409}{66} = 0.948,$$

ahol $\sum_{i=1}^{n} (y_i - \overline{y})^2 = 66$.

(Második példa) Használjuk a következő képletet:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_{2i} y_i, \sum_{i=1}^{n} x_{3i} y_i \right) b.$$

Miért van ez így? Magyarázat:

$$\sum_{i=1}^{n} e_i^2 = (e_1, ..., e_n) \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = e'e = (y - Xb)' e$$
$$= y'e - b' \underbrace{X'e}_{=0} = y' (y - Xb) = y'y - y'Xb$$
$$= \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_{2i} y_i, \sum_{i=1}^{n} x_{3i} y_i \right) b.$$

Tehát $\sum_{i=1}^{n} e_i^2 = 50917 - \begin{pmatrix} 4909 & 66609 & 47527 \end{pmatrix} \begin{pmatrix} 1.696 \\ 0.023 \\ 0.863 \end{pmatrix} = 43.528;$

$s^2 = \dfrac{\sum e_i^2}{n-k} = \dfrac{43.528}{474-3} = 0.092.$

A b varianciájának becslése:

$\widehat{var\,(b)} = s^2 \left(X'X\right)^{-1} = 0.092 \cdot \begin{pmatrix} 2.0794 & 0.0167 & -0.2382 \\ 0.0167 & 0.0005 & -0.0024 \\ -0.2382 & -0.0024 & 0.0279 \end{pmatrix}$, ahol

$(X'X)^{-1} = \begin{pmatrix} 474 & 6395 & 4583 \\ 6395 & 90215 & 62165 \\ 4583 & 62165 & 44376 \end{pmatrix}^{-1} = \begin{pmatrix} 2.0794 & 0.0167 & -0.2382 \\ 0.0167 & 0.0005 & -0.0024 \\ -0.2382 & -0.0024 & 0.0279 \end{pmatrix}.$

Ez alapján a standard hibák

$$s_{b_1} = \sqrt{0.092 \cdot 2.0794} = 0.437$$
$$s_{b_2} = \sqrt{0.092 \cdot 0.0005} = 0.007$$
$$s_{b_3} = \sqrt{0.092 \cdot 0.0279} = 0.051.$$

Ekkor a becsült modellt így írhatjuk:

$$\begin{array}{cccccccc} y = & 1.696 & + & 0.023x_2 & + & 0.863x_3 & + & e. \\ & (0.437) & & (0.007) & & (0.051) & & \end{array}$$

Az $R^2 = 1 - \dfrac{\sum e_i^2}{\sum \left(y_i - \overline{y}\right)^2}$, ezért szükség van a $\sum \left(y_i - \overline{y}\right)^2$ összeg kiszámítására.

$\sum \left(y_i - \overline{y}\right)^2 = \sum y_i^2 - n\left(\overline{y}\right)^2 = \sum y_i^2 - n\left(\dfrac{\sum y_i}{n}\right)^2 = \sum y_i^2 - \dfrac{\left(\sum y_i\right)^2}{n} =$

$50917 - \dfrac{4909^2}{474} = 76.745.$

$R^2 = 1 - \dfrac{\sum e_i^2}{\sum \left(y_i - \overline{y}\right)^2} = 1 - \dfrac{43.528}{76.745} = 0.433.$ Ez azt jelenti, hogy közepesen szoros az összefüggés.

3.3.3. Kiigazított determinancia-együttható

Az R^2 elsődleges szerepe, hogy méri a becsült összefüggés szorosságát. Az R^2-t arra is használhatjuk, hogy két különböző modell közül kiválasszuk azt, amelyikben a független változók jobban megmagyarázzák a függővváltozó

változását, és ezért alkalmasabb előrejelzésekhez. Például, az R^2 segítségével eldönthetjük, hogy a független változók lineárisak vagy nem lineárisak legyenek, vagy hogy a modell tartalmazzon-e bizonyos változókat.

Amikor az R^2-t két modell összehasonlítására használjuk, figyelembe kell vegyük azt a tulajdonságát, hogy ha az eredeti és a leszűkített modelleket hasonlítjuk össze, akkor az előbbi R^2-e sosem kisebb. Vagyis, ha újabb független változókat adunk a modellhez, az R^2 növekszik. Ehhez legyenek

$$y = X_1\mathbf{b}_1 + X_2\mathbf{b}_2 + e \quad \text{és} \quad y = X_1\mathbf{b}_R + e_R$$

az eredeti és a leszűkített becsült modellek és R^2 és R_R^2 a megfelelő determinancia-együtthatók.

Állítás. $R^2 \geq R_R^2$.

Bizonyítás. A determinancia-együttható képlete alapján

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2} \quad \text{és} \quad R_R^2 = 1 - \frac{\sum_{i=1}^n e_{Ri}^2}{\sum_{i=1}^n (y_i - \overline{y})^2},$$

tehát azt kell bizonyítani, hogy $\sum_{i=1}^n e_i^2 \leq \sum_{i=1}^n e_{Ri}^2$. A közönséges legkisebb négyzetek módszere alapján a reziduumok négyzetösszege a lehető legkisebb, vagyis

$$\begin{aligned}\sum_{i=1}^n e_i^2 &= \min_{\mathbf{b}_1,\mathbf{b}_2} \sum_{i=1}^n (y_i - x_{1i}\mathbf{b}_1 - x_{2i}\mathbf{b}_2)^2 \\ &\leq \sum_{i=1}^n (y_i - x_{1i}\mathbf{b}_R - x_{2i}\cdot\mathbf{0})^2 = \sum_{i=1}^n e_{Ri}^2.\end{aligned}$$

Az R^2-nek ez a tulajdonsága miatt bevezették a kiigazított R^2-t, amit $\overline{R}^2$-tel fogunk jelölni:

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^n e_i^2/(n-k)}{\sum_{i=1}^n (y_i - \overline{y})^2/(n-1)} = 1 - \frac{n-1}{n-k}\left(1 - R^2\right).$$

A képlet szerint, ha a paraméterek száma $k = 1$ akkor $\overline{R}^2 = R^2$; ha k növekszik akkor $\frac{n-1}{n-k}$ is nő, ezért $\overline{R}^2$ nem változik egyertelműen.

Az $\overline{R}^2$-nek nincs meg a fenti állításban említett tulajdonsága, és ezért használható olyan modellek összehasonlítására, amelyeknek a függővváltozójuk

ugyanaz (például az eredeti és leszűkített modellek). A **második példa** adatai szerint

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}\left(1-R^2\right) = 1 - \frac{474-1}{474-3}\left(1-0.433\right) = 0.431.$$

3.3.4. A paraméterek tesztelése

A többváltozós lineáris modellnél a paraméterek egyéni tesztelésén kívül a paraméterek egyidejű tesztelését is tanulmányozzuk. Az változók egyéni szignifikanciáján kívül több változó egyidejű szignifikanciáját is tanulmányozzuk.

A paraméterek egyéni tesztelése

Akárcsak az egyváltozós modellnél, a paraméterek egyéni tesztelése a t eloszláson alapszik.

Legyen b a β paraméter-vektor OLS becslőfüggvénye és b_j ennek a j-edik komponense ($j \in \{1, ..., k\}$) , ami a β_j becslőfüggvénye. A teszt statisztikáját és ennek eloszlását a következő állítás adja meg.

Állítás. A1-A6 mellett,

$$t_{b_j} = \frac{b_j - \beta_j}{s_{b_j}} \sim t\left(n-k\right),$$

vagyis, a t_{b_j} statisztika t eloszlású $n-k$ szabadsági fokkal. ($s_{b_j} = \sqrt{\widehat{var}\left(b_j\right)}$ a b_j becslés standard hibája.)

A független változók egyéni szignifikanciájának tesztelésénél ugyanúgy járunk el, mint az egyváltozós lineáris modellnél.

Több paraméter egyidejű tesztelése

Tegyük fel, hogy a β paraméter-vektor két vektor-komponensre van osztva, vagyis $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$. Eszerint

$$y = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Vegyük a $H_0 : \beta_2 = 0$, $H_1 : \beta_2 \neq 0$ null- és alternatív hipotézisű tesztet. Ha β_2 egydimenziós, akkor a fent tárgyalt t-teszt szerint járhatunk el. Ha viszont β_2 egy h-dimenziós vektor ahol $h > 1$, a fenti t-teszt nem alkalmazható. Ez a teszt nagyon fontos a gyakorlatban, ugyanis ennek a segítségével tudjuk

kiválasztani azokat a független változókat, amelyek szignifikánsan befolyásolják a függőváltozót.

Bevezetünk néhány új fogalmat. A nullhipotézis az

$$y = X_1\beta_1 + \varepsilon$$

modellt határozza meg, amit leszűkített modellnek nevezünk. A leszűkített modellt tudjuk becsülni OLS-sel; a reziduumok legyenek

$$e_R = y - X_1\mathbf{b}_1^R,$$

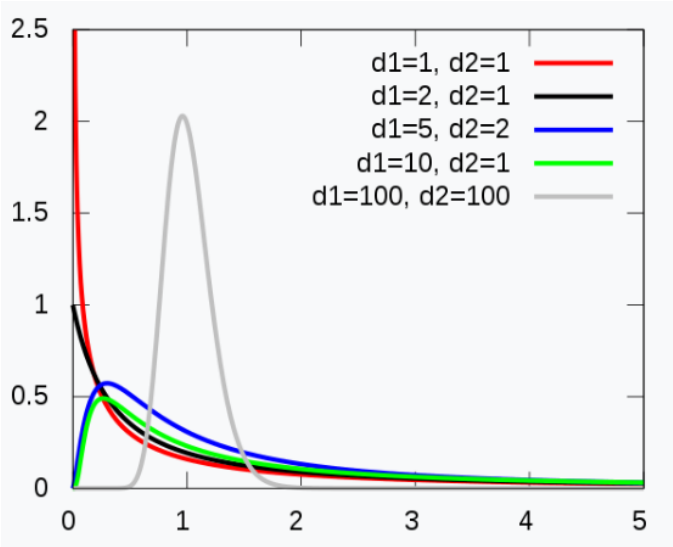
ahol $\mathbf{b}_1^R$ a β_1 paraméter-vektor OLS becslőfüggvénye. A fenti teszt statisztikáját a következő állítás adja meg.

Állítás. Ha $\beta_2 = 0$, akkor A1-A6 mellett

$$F = \frac{\frac{\sum_{i=1}^n e_{Ri}^2 - \sum_{i=1}^n e_i^2}{h}}{\frac{\sum_{i=1}^n e_i^2}{n-k}} \sim F\left(h, n-k\right),$$

vagyis, az F statisztika F eloszlású h és $n-k$ szabadsági fokokkal.

Az F eloszlás sűrűségfüggvénye:



F-teszt a $H_0 : \beta_2 = 0$ és $H_1 : \beta_2 \neq 0$ hipotézisekre

(1) Választunk egy α szignifikancia szintet.

(2) Az elfogadási tartomány a $(0, c_\alpha)$ intervallum, ahol c_α-t a

$$P\left(F < c_\alpha\right) = 1 - \alpha$$

összefüggésből határozzuk meg (c_α-t táblázatokból nézhetjük ki).

(3) Kiszámítjuk az F statisztika értékét a két modell OLS becsléséből.

(4) Ha $F < c_\alpha$, nem utasítjuk el H_0-t, különben elutasítjuk.

Másszóval, a nullhipotézist akkor nem utasítjuk el, ha az F statisztika értéke kellőképpen kicsi. Ezt úgy lehet magyarázni, hogy az F statisztika értéke akkor kicsi, ha $\sum_{i=1}^n e_{Ri}^2 - \sum_{i=1}^n e_i^2$ kicsi, vagyis ha a leszűkített és az eredeti modellek nagyjából megegyeznek. Ez pontosan akkor áll fenn, ha $\beta_2 \approx 0$.

3.3.5. A modell szignifikanciája

A statisztikai programcsomagok (mint az SPSS) általában megadják egy F statisztika és az ennek megfelelő p-érték értékét. Ez a statisztika annak a tesztnek az F statisztikája, amelyben a β_2 vektor a konstans kivételével az összes paramétert tartalmazza. Tehát a teszt az összes független változó szignifikanciájára vonatkozik, vagyis $H_0 : \beta_2 = ... = \beta_k = 0$.

Ebben az esetben a becsült leszűkített modell

$$y_i = \mathbf{b}_R + e_{Ri}, \quad i = 1, 2, ..., n$$

ahol $\mathbf{b}_R = \overline{y}$, tehát a reziduumok négyzetösszege

$$\sum_{i=1}^n e_{Ri}^2 = \sum_{i=1}^n \left(y_i - \overline{y}\right)^2.$$

Ekkor az F statisztika képlete ($h = k - 1$):

$$F = \frac{\frac{\sum_{i=1}^n \left(y_i - \overline{y}\right)^2 - \sum_{i=1}^n e_i^2}{k-1}}{\frac{\sum_{i=1}^n e_i^2}{n-k}} = \frac{n-k}{k-1} \cdot \frac{R^2}{1-R^2} \sim F\left(k-1, n-k\right),$$

ahol R^2 az eredeti modell determinancia-együtthatója. Következésképpen, a modell szignifikanciájának a teszteléséhez az F statisztikához elég az R^2-t kiszámítani.

A példák folytatása

Teszteljük a becslések egyéni szignifikanciáját és a modell szignifikanciáját.

Megoldás. (Első példa) A nullhipotézisek: $H_0 : \beta_1 = 0$ (konstans), $H_0 : \beta_2 = 0$ (első meredekségi együttható), $H_0 : \beta_3 = 0$ (második meredekségi együttható). A t-statisztikák:

$$t_{b_1} = \frac{-0.846}{1.319} = -0.641, t_{b_2} = \frac{0.094}{0.084} = 1.119, t_{b_3} = \frac{1.125}{0.221} = 5.090.$$

A t-statisztika szabadsági foka $5 - 3 = 2$. A kritikus érték 4.3 a táblázat alapján 0.05 szignifikanciára, ami a $(-4.3, 4.3)$ elfogadási tartományt eredményezi. Ezért csak b_3 különbözik 0-tól szignifikánsan, mert egyedül t_{b_3} nincs benne az elfogadási tartományban.

A modell szignifikanciájához a nullhipotézis $H_0 : \beta_2 = \beta_3 = 0$, az alternatív hipotézis $H_1 : \beta_2 \neq 0$ vagy $\beta_3 \neq 0$. A teszt statisztikája

$$F = \frac{n-k}{k-1} \cdot \frac{R^2}{1-R^2},$$

ami a nullhipotézis mellett $F(k-1, n-k)$ eloszlású, vagyis $F(3-1, 5-3) = F(2,2)$. A kritikus érték $c_\alpha = 19$, a statisztika

$$F = \frac{5-3}{3-1} \cdot \frac{0.948}{1-0.948} = 18.231,$$

ami kisebb a kritikus értéknél, tehát nem utasítjuk el a nullhipotézist, vagyis a modell nem szignifikáns.

(Második példa) A t-statisztikák:

$$t_{b_1} = \frac{1.696}{0.437} = 3.881, t_{b_2} = \frac{0.023}{0.007} = 3.285, t_{b_3} = \frac{0.863}{0.051} = 16.922.$$

A t-statisztika szabadsági foka $474-3 = 471$. A kritikus érték 1.97 a táblázat alapján 0.05 szignifikanciára, ami a $(-1.97, 1.97)$ elfogadási tartományt eredményezi. Ezért mindhárom becslés szignifikánsan különbözik 0-tól. Másképp megfogalmazva: $|t_{b_1}|, |t_{b_2}|, |t_{b_3}| > 1.97 \approx 2$, ezért mindhárom becslés szignifikánsan különbözik 0-tól.

A modell szignifikanciájához a nullhipotézis $H_0 : \beta_2 = \beta_3 = 0$, az alternatív hipotézis $H_1 : \beta_2 \neq 0$ vagy $\beta_3 \neq 0$. A teszt statisztikája

$$F = \frac{n-k}{k-1} \cdot \frac{R^2}{1-R^2},$$

ami a nullhipotézis mellett $F(k-1, n-k)$ eloszlású; $F(k-1, n-k) = F(2, 471)$. A kritikus érték $c_\alpha = 3$,

$$F = \frac{474-3}{3-1} \cdot \frac{0.433}{1-0.433} = 179.84 > 3$$

tehát elutasítjuk a nullhipotézist, vagyis a modell szignifikáns.

3.4. Előrejelzések

Az autókereslet-modell becslése után a becsült modellt felhasználhatjuk egy újonnan tervezett autótípus iránti kereslet előrejelzéséhez. Most azt tárgyaljuk, hogyan végezzünk előrejelzést egy becsült modell esetén. Tegyük fel, hogy a becsült modell $y = Xb + e$, és a függővváltozó $y_* = x_*\beta + \varepsilon_*$ értékét akarjuk előrejelezni, ha tudjuk a neki megfelelő k elemű x_* függetlenváltozó-vektort.

Az előrejelzés: $x_*b = b_1 + b_2x_{2*} + ... + b_kx_{k*}$, ami tulajdonképpen az $E[y_*] = x_*\beta$ becslése.

Akárcsak a paraméterek becslésekor, itt is szeretnénk meghatározni az előrejelzés pontosságát. Ehhez értelmezzük az előrejelzés hibáját:

$$f = y_* - x_*b.$$

Mivel ez a hiba nem számítható ki, ezért az f varianciáját fogjuk becsülni, és ezáltal meghatározzuk az előrejelzés pontosságát. Ehhez feltételezzük, hogy az y_*, x_* változók teljesítik az A1-A6 feltételeket. Az A5 alapján

$$y_* = x_*\beta + \varepsilon_*,$$

ahol az ε_* hibaváltozóról feltételezzük, hogy teljesíti az A2-A4, A6 feltételeket. Emellett, még tegyük fel, hogy $E[\varepsilon\varepsilon_*] = 0$, vagyis, a modell és az előrejelzés hibaváltozói nem korreláltak.

Az előrejelzés hibája így is felírható mint:

$$\begin{aligned} f &= x_*\beta + \varepsilon_* - x_*(X'X)^{-1}X'y \\ &= x_*\beta + \varepsilon_* - x_*(X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \varepsilon_* - x_*(X'X)^{-1}X'\varepsilon. \end{aligned}$$

Ezért $E[f] = 0$, vagyis az előrejelzés torzítatlan.

Az előrejelzés hibájának varianciája

$$\begin{aligned} var\left(f\right) &= E\left[f^2\right] - E^2\left[f\right] \\ &= E\left[\left(\varepsilon_* - x_*\left(X'X\right)^{-1}X'\varepsilon\right)^2\right] \\ &= \sigma^2\left(1 + x_*\left(X'X\right)^{-1}x_*'\right). \end{aligned}$$

A variancia becslését megkapjuk, ha σ^2-t helyettesítjük az $s^2 = \frac{e'e}{n-k}$ becsléssel.

Megjegyzés. Az A1-A5 feltételek mellett az x_*b előrejelzés a legjobb lineáris torzítatlan előrejelzése y_*-nak. Pontosabban, y_* bármely $\widehat{y}_* = Ly$ alakú torzítatlan előrejelzésére, ahol L egy nem vv elemű mátrix, fennáll a

$$var\left(f\right) \leq var\left(\widehat{f}\right)$$

egyenlőtlenség, ahol $\widehat{f} = y_* - \widehat{y}_*$ az $\widehat{y}_*$ előrejelzés hibája.

3.4.1. Intervallum-előrejelzések

Hasonlóan a paraméter-becslések konfidencia-intervallumához, tudunk olyan intervallumot szerkeszteni, amely nagy valószínűséggel tartalmazza az előrejelezendő valódi értéket. Ehhez szükség lesz az alábbi állításban megadott eloszlásra.

Állítás. A1-A6 mellett az előrejelzés hibája törve a becsült szórásával $t\left(n-k\right)$ eloszlású, vagyis

$$\frac{f}{s\sqrt{d}} \sim t\left(n-k\right),$$

ahol $d = 1 + x_*\left(X'X\right)^{-1}x_*'$.

Az $\frac{f}{s\sqrt{d}}$ statisztika alapján, amely t eloszlású, adott α szignifikancia szintre, meghatározzuk a c_α kritikus értéket. A

$$P\left(\left|\frac{f}{s\sqrt{d}}\right| < c_\alpha\right) = 1 - \alpha$$

valószínűség alapján $f = y_* - x_*b$ behelyettesítésével megkapjuk az y_* $(1-\alpha)$-**intervallum-előrejelzés**ét:

$$x_*b - c_\alpha s\sqrt{d} < y_* < x_*b + c_\alpha s\sqrt{d}.$$

Példa. A banki alkalmazottak fizetései logaritmusának az átlaga $\overline{y} = \frac{4909}{474} = 10.357$, a tanulási évei számának az átlaga $\overline{x}_2 = \frac{6395}{474} = 13.492$ és a kezdőfizetésük logaritmusának átlaga $\overline{x}_3 = \frac{4583}{474} = 9.669$. Vegyünk most egy banki alkalmazottat, akinek nem tudjuk a fizetését, viszont tudjuk a tanulási éveinek a számát: $x_{2*} = 17$ és a kezdőfizetése logaritmusát: $x_{3*} = 9.67$. Jelezzük előre a fizetésének a logaritmusát és határozzuk meg a 95%-os intervallum-előrejelzést.

Megoldás. Megszerkesztjük az $x_* = \begin{pmatrix} 1 & 17 & 9.67 \end{pmatrix}$ vektort.

Az előrejelzés

$$\begin{aligned} x_* b &= 1.696 + 0.023 \cdot 17 + 0.863 \cdot 9.67 \\ &= \begin{pmatrix} 1 & 17 & 9.67 \end{pmatrix} \begin{pmatrix} 1.696 \\ 0.023 \\ 0.863 \end{pmatrix} = 10.432. \end{aligned}$$

Tehát a fizetés logaritmusának az előrejelzése $x_* b = 10.432$. (Átszámítva: a fizetés előrejelzése $\exp(10.432) = 33928$, a fizetések átlaga $\exp(10.357) = 31477$; éves fizetésről van szó amerikai dollárban kifejezve.)

Kiszámítjuk a d kifejezést:

$$\begin{aligned} d &= 1 + x_* (X'X)^{-1} x_*' \\ &= 1 + \begin{pmatrix} 1 & 17 & 9.67 \end{pmatrix} \begin{pmatrix} 2.0794 & 0.0167 & -0.2382 \\ 0.0167 & 0.0005 & -0.0024 \\ -0.2382 & -0.0024 & 0.0279 \end{pmatrix} \begin{pmatrix} 1 \\ 17 \\ 9.67 \end{pmatrix} = 1.005. \end{aligned}$$

A kritikus érték $c_\alpha = 1.97$; $s = \sqrt{0.092} = 0.303$ (lásd fennebb), tehát a 0.95-intervallum-előrejelzés végpontjai:

$$\begin{aligned} x_* b - c_\alpha s \sqrt{d} &= 10.432 - 1.97 \cdot 0.303 \sqrt{1.005} = 9.834, \\ x_* b + c_\alpha s \sqrt{d} &= 10.432 + 1.97 \cdot 0.303 \sqrt{1.005} = 11.03. \end{aligned}$$

Az intervallum végpontjai átszámítva:

$$\begin{aligned} \exp\left(x_* b - c_\alpha s \sqrt{d}\right) &= \exp(9.834) = 18657, \\ \exp\left(x_* b + c_\alpha s \sqrt{d}\right) &= \exp(11.03) = 61698. \end{aligned}$$

A logaritmusban pontos intervallum-előrejelzés átszámítva már nem annyira pontos.

3.5. Rugalmasság

Az y függővváltozó rugalmassága az x_j független változóhoz viszonyítva az y százalékbeli változása az x_j 1%-nyi változása esetén (ceteris paribus). Példaként az árrugalmasságot, vagyis a kereslet árhoz viszonyított rugalmasságát említjük. Legyen a független változó új értéke x'_j, a függővváltozó új értéke y'. A rugalmasság képlete:

$$\frac{y'-y}{y} \Big/ \frac{x'_j - x_j}{x_j} = \frac{y'-y}{x'_j - x_j} \cdot \frac{x_j}{y} = \beta_j \frac{x_j}{y},$$

mivel $y = \beta_1 + \beta_2 x_2 + ... + \beta_j x_j + ... + \varepsilon_j$ és $y' = \beta_1 + \beta_2 x_2 + ... + \beta_j x'_j + ... + \varepsilon_j$, ezért $y' - y = \beta_j \left(x'_j - x_j\right)$.

A rugalmasságot, ha az y és x_j értékei nincsenek megadva, gyakran a megfigyelések átlaga alapján számítják ki; ebben az esetben: $b_j \dfrac{\overline{x}_j}{\overline{y}}$.

Példa. Határozzuk meg az y rugalmasságát az x_3-hoz viszonyítva.
Megoldás. A b_3 becslés 0.863, $\overline{y} = 10.357$, $\overline{x}_3 = 9.669$, ezért az y rugalmassága az x_3-hoz viszonyítva $0.863 \cdot \frac{9.669}{10.357} = 0.806$. Tehát átlagban az x_3 1%-nyi növekedésére az y szintén növekszik de csak 0.806%-kal.

Vegyük most az ún. log-lineáris modellt:

$$\ln y = \beta_1 + \beta_2 \ln x_2 + ... + \beta_k \ln x_k + \varepsilon.$$

A modell paramétereit OLS-szel becsülhetjük. Ebben az esetben az y függővváltozó rugalmassága az x_j független változóhoz viszonyítva

$$\begin{aligned} \frac{y'-y}{y} \Big/ \frac{x'_j - x_j}{x_j} &= \left(\frac{y'}{y} - 1\right) \Big/ \left(\frac{x'_j}{x_j} - 1\right) \approx \ln\left(\frac{y'}{y}\right) \Big/ \ln\left(\frac{x'_j}{x_j}\right) \\ &= \frac{\ln y' - \ln y}{\ln x'_j - \ln x_j} = \beta_j. \end{aligned}$$

Tehát az y függővváltozó rugalmassága az x_j független változóhoz viszonyítva konstans, egyenlő az $\ln x_j$ együtthatójával.

Példa. Határozzuk meg a banki fizetések rugalmasságát a kezdőfizetésekhez viszonyítva.
Megoldás. Mivel mindkét változó logaritmusban szerepel, ezért a rugalmasság $b_3 = 0.863$. Tehát, ha a kezdőfizetés 1%-kal nagyobb, akkor a fizetés 0.863%-kal lesz nagyobb.

ÖSSZEGZÉS

Ebben a fejezetben az

$$y = X\beta + \varepsilon$$

többváltozós lineáris modellt tanulmányoztuk. Láttuk, hogy az A1, A2, A5 feltételek mellett a β paraméter-vektor b OLS becslőfüggvénye torzítatlan,

$$E[b] = \beta,$$

és az A1-A6 feltételek mellett normál eloszlású,

$$b \sim N\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right).$$

A tény, hogy a becslőfüggvény normál eloszlású különböző hipotézisek tesztelését teszi lehetővé az egyéni paraméterekről és több paraméter egyidejű szignifikanciájáról, valamint konfidencia intervallumok és intervallum-előrejelzések szerkesztése is megvalósíthatóvá válik. Láttuk, hogy ezek a tesztek és intervallumok az A1-A6 feltételek mellett helyesek.

A következő fejezetben olyan gyakorlati szempontból fontos eseteket tanulmányozunk, amelyekben ezek a feltételek nem mind teljesülnek.

3.6. Gyakorlatok

Az angol font és a német márka közötti árfolyamra vegyük az 1975. január és 1983. április közötti hónaponkénti átlagokat, melyek logaritmusát jelöljük y-nal, valamint az angol és német fogyasztói árindexeket, amelyek logaritmusát jelöljük x_2-vel és x_3-mal. Ezekről az adatokról a következőket tudjuk:

$$\begin{pmatrix} \sum_{i=1}^n y_i^2 & \sum_{i=1}^n y_i & \sum_{i=1}^n y_i x_{2i} & \sum_{i=1}^n y_i x_{3i} \\ & n & \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{3i} \\ & & \sum_{i=1}^n x_{2i}^2 & \sum_{i=1}^n x_{2i} x_{3i} \\ & & & \sum_{i=1}^n x_{3i}^2 \end{pmatrix}$$

$$= \begin{pmatrix} 530.35 & 21.16 & 131.42 & 165.77 \\ & 100.00 & -5.87 & -13.61 \\ & & 80.57 & 31.55 \\ & & & 100.69 \end{pmatrix}.$$

1. Számítsuk ki az $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$ lineáris modell paramétereinek OLS becsléseit.
   Útmutatás.

$$\begin{pmatrix} 100.00 & -5.87 & -13.61 \\ -5.87 & 80.57 & 31.55 \\ -13.61 & 31.55 & 100.69 \end{pmatrix}^{-1} = \begin{pmatrix} 0.01019 & 0.00023 & 0.0013 \\ 0.00023 & 0.01415 & -0.0044 \\ 0.0013 & -0.0044 & 0.01148 \end{pmatrix}.$$

(Excelben is ki lehet számítani az inverzet.)

2. Számítsuk ki a reziduumok négyzetösszegét és a becslések standard hibáit.
   Útmutatás. A reziduumok négyzetösszege kiszámítható a

$$\sum_{i=1}^{n} e_i^2 = y'y - y'Xb = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \quad \sum_{i=1}^{n} y_i x_{2i} \quad \sum_{i=1}^{n} y_i x_{3i} \right) \cdot b$$

   képlettel.

3. Teszteljük a becslések szignifikanciáját egyénileg, használjuk a 2 kritikus értéket.

4. Teszteljük a modell szignifikanciáját, használjuk a 3.1 kritikus értéket.

5. Az 1983. májusi angol fogyaszói árindex logaritmusa -0.04 és a német fogyaszói árindex logaritmusa -0.07. Szerkesszünk 95%-os intervallum-előrejelzést az angol font és a német márka közötti árfolyam logaritmusára és ez alapján magára az árfolyamra. A kritikus értéket vegyük 1.98-nak; használjuk fel, hogy $d = 1 + x_* \left(X'X\right)^{-1} x_*' = 1.01$.

6. Határozzuk meg az angol font és a német márka közötti árfolyam rugalmasságát a német fogyasztói árindexhez viszonyítva.

4. fejezet

Általánosabb feltételek

A gyakorlatban az A1-A6 feltétel bármelyikével előfordulhat, hogy nem teljesül. Ebben a fejezetben részletesen tárgyaljuk azokat az eseteket, amikor az A3 (Homoszkedaszticitás) vagy az A4 (A hibaváltozók nem korreláltak) feltételek nem teljesülnek. Utána olyan esetekről lesz szó, amelyekben az A5 feltétel nem (mindig) teljesül. Ezek a gyakorlatban gyakran előfordulhatnak. Konkrétan, olyan eseteket tárgyalunk, amikor fontos változók maradnak ki a modellből, amikor olyan változók szerepelnek a modellben, amelyek nem befolyásolják a függőváltozót vagy amikor a paraméterek nem állandók a teljes mintára, amit strukturális törésnek neveznek. Végül a multikollinearitásról lesz szó, ami az a jelenség, amikor két vagy több független változó nagyon szorosan összefügg, ezért az A1 feltételből az, hogy $X'X$ invertálható megközelítőleg nem teljesül.

4.1. Heteroszkedaszticitás

A hibaváltozókat heteroszkedasztikusnak nevezik, ha az A3 feltétel nem teljesül, vagyis

$$var\left(\varepsilon_i\right) = E\left[\varepsilon_i^2\right] = \sigma_i^2,$$

ahol a σ_i^2, $i = 1, ..., n$ értékek nem mind egyenlők. A2 és A4 mellett ez az eset

$$var\left(\varepsilon\right) = E\left[\varepsilon\varepsilon'\right] = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix} = \Omega$$

formában foglalható össze.

Példa. A heteroszkedaszticitást nagyon gyakran az okozza, hogy a lineáris modellben különböző megfigyeléseknek különböző paraméterek felelnek meg, vagyis

$$y_i = \beta_{1i} + \beta_{2i}x_{2i} + \beta_{3i}x_{3i} + \varepsilon_i,$$

ahol ε_i teljesíti az A2-A4 és A6 feltételeket. A különböző paramétereknek lehet közös részük, amelyektől véletlenszerűen térnek el:

$$\beta_{1i} = \beta_1 + v_{1i},\ \beta_{2i} = \beta_2 + v_{2i},\ \beta_{3i} = \beta_3 + v_{3i},$$

ahol v_{i1}, v_{i2}, v_{i3} nulla várható értékű normál eloszlású egymással nem korrelált vvk. Ezért

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i + v_{1i} + v_{2i}x_{2i} + v_{3i}x_{3i}.$$

Ez egy olyan lineáris modell, amelynek hibaváltozója $\xi_i = \varepsilon_i + v_{1i} + v_{2i}x_{2i} + v_{3i}x_{3i}$, és ez a hibaváltozó teljesíti a

$$\begin{aligned} E[\xi_i] &= 0 \quad \text{és} \\ var(\xi_i) &= \sigma^2 + \omega_1^2 + \omega_2^2 x_{2i}^2 + \omega_3^2 x_{3i}^2 \end{aligned}$$

feltételeket, ahol $\omega_k^2 = var(v_{ki})$. Ha

$$\sigma_i^2 = \sigma^2 + \omega_1^2 + \omega_2^2 x_{2i}^2 + \omega_3^2 x_{3i}^2$$

a hibaváltozó varianciája, akkor ez a hibaváltozó heteroszkedasztikus.

4.1.1. Az OLS becslőfüggvény tulajdonságai

Az alábbi két állítás alapján az OLS becslőfüggvény bizonyos elég általános feltételek mellett helyes. Viszont a becslések pontosságának a kiszámítása az előző fejezet képlete alapján nem lesz helyes, ezért másik standard hiba képletre van szükség. Ezt a következő alfejezetben tárgyaljuk.

Állítás. Az A1, A2, A5 feltételek mellett a $b = (X'X)^{-1}X'y$ OLS becslőfüggvény torzítatlan, vagyis

$$E[b] = \beta,$$

és a varianciája

$$var(b) = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}. \tag{4.1}$$

Bizonyítás. A 3. Fejezetben láttuk, hogy ahhoz hogy az OLS becslőfüggvény torzítatlan legyen csak az A1, A2, A5 feltételekre van szükség, tehát a tulajdonság évényes az A3 (Homoszkedaszticitás) feltétel nélkül.
A variancia kiszámításához idézzük fel, hogy $b = \beta + (X'X)^{-1} X'\varepsilon$. Ezért

$$\begin{aligned} var(b) &= var\left(\beta + (X'X)^{-1} X'\varepsilon\right) = var\left((X'X)^{-1} X'\varepsilon\right) \\ &= E\left[(X'X)^{-1} X'\varepsilon\varepsilon' X (X'X)^{-1}\right] = (X'X)^{-1} X' E[\varepsilon\varepsilon'] X (X'X)^{-1} \\ &= (X'X)^{-1} (X'\Omega X) (X'X)^{-1}. \end{aligned}$$

Állítás. Heteroszkedaszticitás esetén, az A1, A2, A4, A5 feltételek mellett a $b = (X'X)^{-1} X'y$ OLS becslőfüggvény konszisztens, vagyis

$$var(b) \underset{n\to\infty}{\longrightarrow} 0,$$

ha a σ_i^2, $i = 1, ..., n$ varianciák és az $\frac{A'A}{n}$ mátrix elemei végesek maradnak, ha $n \to \infty$, ahol $A = (|x_{ij}|)_{\substack{i=1,...,n \\ j=1,...,n}}$.
Bizonyítás. A (4.1) képlet alapján

$$var(b) = \frac{1}{n}\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\Omega X}{n}\right)\left(\frac{X'X}{n}\right)^{-1}.$$

Mivel a

$$\frac{X'\Omega X}{n} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \sigma_i^2 & \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{2i} & \cdots & \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{ki} \\ \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{2i} & \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{2i}^2 & \cdots & \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{2i}x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{ki} & \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{2i}x_{ki} & \cdots & \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{ki}^2 \end{pmatrix},$$

mátrix elemei végesek maradnak, ha $n \to \infty$, ezért $var(b) \underset{n\to\infty}{\longrightarrow} 0$.

Ahhoz, hogy meggyőződjünk erről, vegyük például az $\frac{X'\Omega X}{n}$ mátrix $(2, k)$ elemét, ami $\frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{2i}x_{ki}$. Erre

$$\left|\frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_{2i}x_{ki}\right| \le \frac{1}{n}\sum_{i=1}^n \sigma_i^2 |x_{2i}x_{ki}| \le B\frac{1}{n}\sum_{i=1}^n |x_{2i}x_{ki}|,$$

ahol B egy olyan szám, amely mindegyik σ_i^2 varianciánál nagyobb. Vegyük észre, hogy $\frac{1}{n}\sum_{i=1}^n |x_{2i}x_{ki}| < \infty$ mivel az $\frac{A'A}{n}$ mátrix $(2, k)$ eleme.

4.1.2. A standard hibák kiszámítása

Az A1-A6 feltételek mellett a standard hibákat a becslőfüggvény varianciája alapján számítottuk ki. Tehát az A1, A2, A5 feltételek mellett a (4.1) képletet kell alkalmazzuk a standard hibák helyes kiszámításához. A statisztikai programcsomagok (például az SPSS) az OLS becsléshez az eredeti $var(b) = \sigma^2 (X'X)^{-1}$ képletet használják a standard hibák kiszámításához. Ha A3 nem teljesül, ez a képlet nem helyes!

A (4.1) képlet kiszámításánál felmerül az a probléma, hogy a gyakorlatban általában nem ismerjük az

$$\Omega = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$$

mátrix $\sigma_1^2, ..., \sigma_n^2$ elemeit. Lehet ezeket helyesen (konszisztensen) becsülni? Nem, mert n megfigyelésünk van és a β vektor elemeivel együtt $n+k$ ismeretlen paramétert kellene becsüljünk.

Viszont észrevesszük, hogy a $k \times k$ dimenziójú $X'\Omega X$ mátrixnak legfennebb $\frac{k(k+1)}{2}$ ismeretlen eleme van (mert szimmetrikus), ezért ha $n >> k$, akkor ezeket lehet helyesen becsülni. Ezt az ötletet alkalmazza a variancia White-féle becslőfüggvénye:

$$var(b) = (X'X)^{-1} \left(X'\widehat{\Omega}X \right) (X'X)^{-1},$$

ahol

$$\widehat{\Omega} = \begin{pmatrix} e_1^2 & & 0 \\ & \ddots & \\ 0 & & e_n^2 \end{pmatrix}.$$

Ezt az eljárást a következőképpen lehet megmagyarázni. Ugyanúgy, ahogy a korábban tárgyalt σ^2 becsléséhez a reziduumok négyzetét (e_i^2) használtuk, itt is felhasználjuk őket a σ_i^2 varianciák becsléséhez. Ebben az esetben e_i^2 nem konszisztens becslése σ_i^2-nek, és az $\widehat{\Omega}$ mátrix sem konszisztens becslése a Ω mátrixnak, viszont a $X'\widehat{\Omega}X = \sum_{i=1}^n e_i^2 x_i' x_i$ mátrix konszisztens becslőfüggvénye a $X'\Omega X = \sum_{i=1}^n \sigma_i^2 x_i' x_i$ mátrixnak.

Példa

Az y függő- és x_2, x_3 független változókra a következő megfigyeléseink vannak:

$$\begin{array}{lccccc} y: & 2 & 3 & 5 & 8 & 12 \\ x_2: & 1 & 8 & 16 & 24 & 12 \\ x_3: & 2 & 4 & 3 & 6 & 10 \end{array}$$

Számítsuk ki a $b = (X'X)^{-1} X'y$ OLS becslőfüggvény White-féle standard hibáit az

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

modellre és vessük össze a korábban kiszámított standard hibákkal:

$$\begin{aligned} s_{b_1} &= \sqrt{1.741} = 1.319 \\ s_{b_2} &= \sqrt{0.007} = 0.084 \\ s_{b_3} &= \sqrt{0.049} = 0.221. \end{aligned}$$

Megoldás. A b White-féle varianciája

$$\widetilde{var\,(b)} = (X'X)^{-1} X'\widehat{\Omega}X\,(X'X)^{-1}$$

ahol

$$\widehat{\Omega} = \begin{pmatrix} e_1^2 & & 0 \\ & \ddots & \\ 0 & & e_n^2 \end{pmatrix}.$$

Ezért $X'\widehat{\Omega}X = \begin{pmatrix} \sum_{i=1}^n e_i^2 & \sum_{i=1}^n e_i^2 x_{2i} & \sum_{i=1}^n e_i^2 x_{3i} \\ \sum_{i=1}^n e_i^2 x_{2i} & \sum_{i=1}^n e_i^2 x_{2i}^2 & \sum_{i=1}^n e_i^2 x_{2i} x_{3i} \\ \sum_{i=1}^n e_i^2 x_{3i} & \sum_{i=1}^n e_i^2 x_{2i} x_{3i} & \sum_{i=1}^n e_i^2 x_{3i}^2 \end{pmatrix}$, ahol

$$e = y - Xb = \begin{pmatrix} 2 \\ 3 \\ 5 \\ 8 \\ 12 \end{pmatrix} - \begin{pmatrix} 1 & 1 & 2 \\ 1 & 8 & 4 \\ 1 & 16 & 3 \\ 1 & 24 & 6 \\ 1 & 12 & 10 \end{pmatrix} \begin{pmatrix} -0.846 \\ 0.094 \\ 1.125 \end{pmatrix} = \begin{pmatrix} 0.502 \\ -1.406 \\ 0.967 \\ -0.160 \\ 0.468 \end{pmatrix},$$

$$(X'X)^{-1} = \begin{pmatrix} 1.0215 & -0.0278 & -0.0965 \\ -0.0278 & 0.0039 & -0.0040 \\ -0.0965 & -0.0040 & 0.0291 \end{pmatrix},$$

$\sum_{i=1}^n e_i^2 = 0.502^2 + (-1.406)^2 + 0.967^2 + (-0.16)^2 + 0.468^2 = 3.408,$

$\sum_{i=1}^n e_i^2 x_{2i} = 0.502^2 \cdot 1 + (-1.406)^2 \cdot 8 + 0.967^2 \cdot 16 + (-0.16)^2 \cdot 24 + 0.468^2 \cdot 12 =$

34.271,
$\sum_{i=1}^{n} e_i^2 x_{3i} = 0.502^2 \cdot 2 + (-1.406)^2 \cdot 4 + 0.967^2 \cdot 3 + (-0.16)^2 \cdot 6 + 0.468^2 \cdot 10 =$ 13.56,
$\sum_{i=1}^{n} e_i^2 x_{2i}^2 = 0.502^2 \cdot 1 + (-1.406)^2 \cdot 8^2 + 0.967^2 \cdot 16^2 + (-0.16)^2 \cdot 24^2 + 0.468^2 \cdot 12^2 =$ 412.44,
$\sum_{i=1}^{n} e_i^2 x_{2i} x_{3i} = 0.502^2 \cdot 1 \cdot 2 + (-1.406)^2 \cdot 8 \cdot 4 + 0.967^2 \cdot 16 \cdot 3 + (-0.16)^2 \cdot 24 \cdot 6 + 0.468^2 \cdot 12 \cdot 10 = 138.62$,
$\sum_{i=1}^{n} e_i^2 x_{3i}^2 = 0.502^2 \cdot 2^2 + (-1.406)^2 \cdot 4^2 + 0.967^2 \cdot 3^2 + (-0.16)^2 \cdot 6^2 + 0.468^2 \cdot 10^2 =$ 63.877.
Innen kapjuk, hogy

$$X'\widehat{\Omega}X = \begin{pmatrix} 3.408 & 34.271 & 13.56 \\ 34.271 & 412.44 & 138.62 \\ 13.56 & 138.62 & 63.877 \end{pmatrix}, \text{ és}$$

$$\widetilde{var\,(b)} = (X'X)^{-1} X'\widehat{\Omega}X \,(X'X)^{-1} = \begin{pmatrix} 1.0215 & -0.0278 & -0.0965 \\ -0.0278 & 0.0039 & -0.0040 \\ -0.0965 & -0.0040 & 0.0291 \end{pmatrix} \cdot$$

$$\cdot \begin{pmatrix} 3.408 & 34.271 & 13.56 \\ 34.271 & 412.44 & 138.62 \\ 13.56 & 138.62 & 63.877 \end{pmatrix} \begin{pmatrix} 1.0215 & -0.0278 & -0.0965 \\ -0.0278 & 0.0039 & -0.0040 \\ -0.0965 & -0.0040 & 0.0291 \end{pmatrix}$$

$$= \begin{pmatrix} 0.594 & -0.009 & -0.047 \\ -0.009 & 0.001 & -0.001 \\ -0.047 & -0.001 & 0.010 \end{pmatrix}. \text{ Tehát}$$

$s_{b_1} = \sqrt{0.594} = 0.771$,
$s_{b_2} = \sqrt{0.001} = 0.032$,
$s_{b_3} = \sqrt{0.01} = 0.1$.
A következő táblázat összegzi ezeket és az OLS-féle standard hibákat:

		OLS standard hibák	White standard hibák
b_1	-0.846	1.319	0.771
b_2	0.094	0.084	0.032
b_3	1.125	0.221	0.100

Példa. A nettó hozamok modelljére, amelyben a RENDMARK hatását vizsgáljuk a RENDCYCO-ra, SPSS-ben meghatároztuk az OLS becslést és a White standard hibákat. A kétféle standard hiba között van egy kis különbség.

SPSS-ben a White standard hibák kiszámításához a következő lépéseket követjük: Analyze–>Generalized Linear Models–>Response: Dependent variable: RENDCYCO–>

Predictors: Covariates: RENDMARK–>Model: Model: RENDMARK–>
Estimation: Covariance Matrix: Robust estimator.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.447	.363		-1.233	.219
	RENDMARK	1.171	.075	.710	15.535	.000

a. Dependent Variable: RENDCYCO

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-.447	.3421	-1.118	.223	1.711	1	.191
RENDMARK	1.171	.0679	1.038	1.304	297.264	1	.000
(Scale)	30.466[a]	2.7812	25.475	36.435			

Dependent Variable: RENDCYCO
Model: (Intercept), RENDMARK

a. Maximum likelihood estimate.

ÖSSZEGZÉS

- Heteroszkedaszticitás esetén a β paraméter-vektor OLS becslőfüggvénye helyes (konszisztens), de a standard hibák kiszámításához a variancia White-féle becslőfüggvényét kell használjuk. Ezeket White (-féle) standard hibáknak nevezik.

- A White standard hibákból kiszámított t-statisztikákat és az ehhez tartozó kritikus értékeket kell használjuk konfidencia intervallumok és intervallum-előrejelzések meghatározásához.

- Heteroszkedaszticitás esetén a β paraméter-vektor OLS becslőfüggvénye nem rendelkezik a legjobb lineáris torzítatlan becslőfüggvény tulajdonsággal.

- Heteroszkedaszticitás esetén az F-teszt a paraméterek egyidejű szignifikanciájára nem helyes az eredeti eljárás alapján.

4.1.3. Hogy ismerjük fel a heteroszkedaszticitást?

A heteroszkedaszticitás vizsgálata lényegében a hibaváltozók varianciáját kell vizsgálja. Mivel ez közvetett módon nem végezhető, ezért a reziduumok segítségével történik. Kezdjük a reziduumok vizsgálatával különböző ábrák

segítségével. Az ábrák segítséget nyújthatnak további pontosabb vizsgálatra statisztikai tesztek segítségével.

A heteroszkedaszticitás vizsgálata ábrák segítségével:
- A reziduumokról és a négyzetükről (e_i és e_i^2) vonalgrafikont készítünk. Ezek különösen idősorok esetén hasznosak, megmutatják a hibaváltozók varianciájának időbeni változását.
- A független változókról és a reziduumokról szóródási képet készítünk. Ezek megmutatják a hibaváltozók varianciájának változását a független változók függvényében.

A reziduumok használhatók a hibaváltozók helyett, mert ezek konszisztens becslései. Miért? Azért, mert tudjuk, hogy az OLS reziduumok megközelítőleg egyenlők a hibaváltozókkal:

$$e_i = y_i - x_i b \simeq y_i - x_i \beta = \varepsilon_i,$$

mert a 2. Következmény alapján A1, A2, A5 mellett az OLS becslőfüggvény konszisztens, ezért $b \simeq \beta$.

Ezért ezek az ábrák egy első benyomást nyújtanak arról, hogy a hibaváltozók heteroszkedasztikusak-e, és emellett arról is, hogy ha igen, akkor mi okozhatja ezt. Hogy pontosabb képet kaphassunk, formálisan kell teszteljük a hibaváltozók heteroszkedaszticitását. Ehhez egy modellt szerkesztünk a hibaváltozók varianciáira.

A hibaváltozók varianciáinak egy modellje

Láttuk, hogy nem lehet helyesen becsülni az összes σ_i^2 varianciát, $i = 1, ..., n$. Viszont gyakran ezek a varianciák csak néhány a modellhez kapcsolódó változótól függnek. Például, vegyük a már említett $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \xi_i$ modellt, ahol a hibaváltozó $\xi_i = \varepsilon_i + v_{1i} + v_{2i} x_{2i} + v_{3i} x_{3i}$, és erre a hibaváltozóra

$$\begin{aligned} E[\xi_i] &= 0 \quad \text{és} \\ var(\xi_i) &= \sigma^2 + \omega_1^2 + \omega_2^2 x_{2i}^2 + \omega_3^2 x_{3i}^2, \end{aligned}$$

vagyis a varianciák a független változók négyzetétől függnek.

Ilyen esetekben szerkeszthetünk egy lineáris modellt a varianciáknak:

$$\sigma_i^2 = z_i \gamma \equiv \gamma_1 + \gamma_2 z_{2i} + ... + \gamma_h z_{hi}$$

ahol $z_i = (1, z_{2i}, ..., z_{hi})$ olyan változókat tartalmaz, amelyek befolyásolják a varianciákat, és h egy n-hez képest kicsi szám. A fenti példában: $z_{2i} = x_{2i}^2$, $z_{3i} = x_{3i}^2$, $\gamma_1 = \sigma^2 + \omega_1^2$, $\gamma_2 = \omega_2^2$, $\gamma_3 = \omega_3^2$. Általában, a z_i változók

lehetnek az eredeti modell független változói, és ezek különböző hatványai vagy szorzata.

Az ábrák segítségével eldönthetjük, hogy milyen változókat vegyünk be a varianciák modelljébe. Ha a γ paraméter-vektort helyesen becsüljük, akkor a varianciák modelljét arra használhatjuk, hogy becsüljük a σ_i^2 varianciákat. Ezt egyrészt arra használhatjuk, hogy a változókat transzformáljuk úgy, hogy a hibaváltozók homoszkedasztikussá váljanak. Másrészt, a $\sigma_i^2 = z_i\gamma$ modellt felhasználhatjuk arra, hogy egy modell-szignifikancia F-teszt segítségével teszteljük a hibaváltozók heteroszkedaszticitását. Ezeknek a részleteit alább tárgyaljuk.

A varianciák modelljének becslése

A γ paraméter-vektort OLS-sel becsüljük az alábbi modell alapján:

$$e_i^2 = z_i\gamma + \eta_i,$$

vagyis, a σ_i^2 varianciát a neki megfelelő reziduum-négyzettel (e_i^2) helyettesítjük, és a modellhez egy hibaváltozót adunk. Ennek az indoklása hasonló mint az eddig tárgyalt két esetben (σ^2 becslése és a White standard hibák): az e_i^2 reziduum-négyzetet használhatjuk az ε_i^2 hibaváltozó-négyzet megközelítéseként mert

$$e_i^2 = (y_i - x_i b)^2 \simeq (y_i - x_i\beta)^2 = \varepsilon_i^2,$$

és mivel $\sigma_i^2 = var\,(\varepsilon_i^2)$, ezért e_i^2 hasznos információt tartalmaz az ismeretlen σ_i^2-ről.

Tehát, ha c-vel jelöljük a γ OLS becslőfüggvényét, akkor a σ_i^2 varianciákat az

$$s_i^2 = z_i c, \quad i = 1, ..., n$$

kifejezésekkel becsüljük.

Heteroszkedaszticitás tesztelése

Ha a varianciák modellje nem szignifikáns, akkor valószínű, hogy az eredeti modell homoszkedasztikus. Ez a megállapítás az alapja a heteroszkedaszticitás tesztelésének.

Formálisan, a nullhipotézis

$$H_0 : \sigma_1^2 = ... = \sigma_n^2$$

ekvivalens azzal, hogy $\gamma_2 = ... = \gamma_h = 0$.

Ezért, a homoszkedaszticitás tesztelése azt jelenti, hogy a fenti nullhipotézist teszteljük azzal az alternatív hipotézissel szemben, hogy legalább két σ_i^2 különbözik egymástól, ami ekvivalens azzal, hogy legalább egy $\gamma_j,\ j \in \{2, ..., h\}$ különbözik 0-tól. Ezt a tesztet Breusch-Pagan heteroszkedaszticitás-tesztnek nevezik.

A Breusch-Pagan teszt lépései:
(i) Becsüljük az eredeti $y = X\beta + \varepsilon$ modellt OLS-sel és kiszámítjuk a reziduumokat: $e = y - Xb$. Ábrák segítségével megvizsgáljuk a reziduumokat, és kiválasztjuk a $z_i = (1, z_{2i}, ..., z_{hi})$ változókat a varianciák $\sigma_i^2 = z_i\gamma$ modelljéhez.
(ii) Becsüljük OLS-sel a segédmodellt:

$$e_i^2 = z_i\gamma + \eta_i, \quad i = 1, ..., n.$$

(iii) A $\gamma_2 = ... = \gamma_h = 0$ nullhipotézist a modell-szignifikancia F-teszt segítségével teszteljük a segédmodellben. A tesztstatisztika

$$F = \frac{n-h}{h-1} \cdot \frac{R^2}{1-R^2} \approx F\left(h-1, n-h\right),$$

ahol R^2 a segédmodell R^2-e.

Ennek a tesztnek egy sajátos esete a White heteroszkedaszticitás-teszt. Ennél

$$z_i = \left(1, x_{2i}, ..., x_{ki}, x_{2i}^2, ..., x_{ki}^2\right),$$

ezért $h = 2k - 1$. Ez a White teszt szorzatok nélküli változata.
A White teszt egy másik változata tartalmazza még a független változók páronkénti szorzatait, vagyis $x_{ji}x_{li}$-t minden $j < l$-re. Ebben az esetben $h = 2k - 1 + \frac{k(k-1)}{2}$.

Példa: nettó hozamok

A tőkejavak árazási modellje összefüggést állít fel valamely értékpapír várható hozamának meghatározására. Ezen a modellen belül meg szeretnénk határozni az összes piacra számított nettó hozam (RENDMARK) hatását a ciklikus termékpiaci (tartós termékek, autók, ruhaneműk, sporttermékek piaca) nettó hozamra (RENDCYCO). A nettó hozam az illető hozam és a banki kamat közötti különbség.

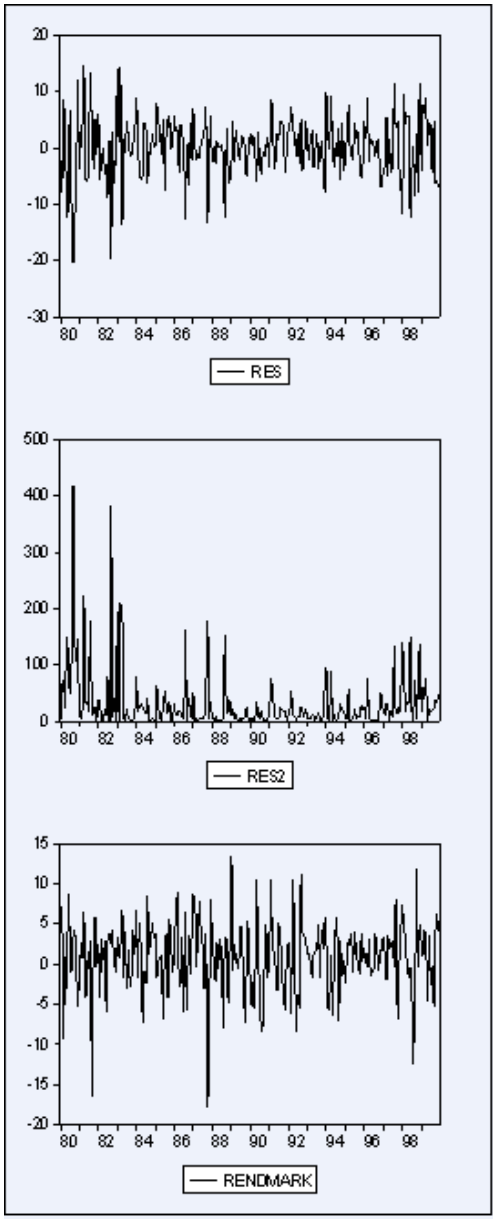
A függővváltozó a RENDCYCO, a független változó a RENDMARK, a megfigyelések 240 havi átlagot tartalmaznak 1980-1999 között. Az alábbi (Eviews-ban készült) táblázat mutatja a modell OLS-becslésének az eredményét.

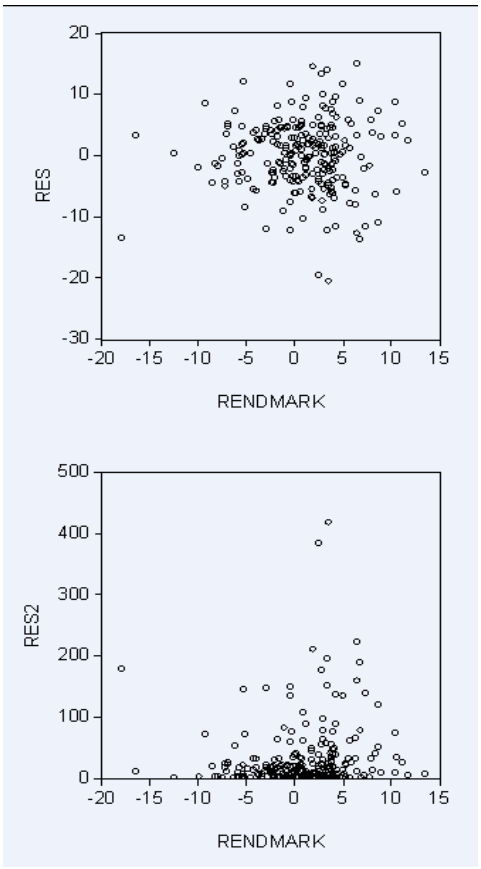
Dependent Variable: RENDCYCO				
Method: Least Squares				
Date: 11/21/06 Time: 11:11				
Sample: 1980:01 1999:12				
Included observations: 240				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
RENDMARK	1.210554	0.079929	15.14540	0.0000
C	-0.529811	0.368652	-1.437157	0.1520
R-squared	0.490782	Mean dependent var		0.499826
Adjusted R-squared	0.488642	S.D. dependent var		7.849594
S.E. of regression	5.613188	Akaike info criterion		6.296413
Sum squared resid	7498.875	Schwarz criterion		6.325418
Log likelihood	-753.5696	F-statistic		229.3832
Durbin-Watson stat	2.001640	Prob(F-statistic)		0.000000

A heteroszkedaszticitás vizsgálatához elkészítjük a vonalgrafikonokat és a szóródási képeket. A reziduumok és a négyzeteik vonalgrafikonja szerint a hibaváltozók varianciája nagyobb az 1980:11-1983:4 és 1997:08-1999:12 periódusokban. Ezért két dummy változót értelmezünk: $D_1 = 1$ az első periódusban és 0 más esetben; $D_2 = 1$ a második periódusban és 0 más esetben. (Egy dummy változó olyan típusú változó, amelynek mindegyik megfigyelése csak 0 vagy 1 lehet.) A RENDMARK vonalgrafioknját (baloldalt legalul) összehasonlíthatjuk a többi vonalgrafikonnal, hogy egyeztessük az esetleges nagyobb szórási periodusokat, de nehéz összefüggést találni.

A RENDMARK és a reziduumok (RES) közötti szóródási kép alapján nehéz bármilyen következtetést levonni a két változó közötti összefüggésről. A RENDMARK és a reziduumok négyzete közötti szóródási kép mutat némi összefüggést a két változó között (RES2 a reziduumok négyzete).

Összességében arra a következtetésre jutottunk, hogy az ábrák alapján nehéz egyértelmű következtetéseket levonni a hibaváltozók heteroszkedaszticitását illetően. Viszont a vonalgrafikonok alapján értelmezett dummy változókat felhasználhatjuk a varianciák modelljének a megszerkesztésére, amivel tudjuk pontosabban tesztelni a (Breusch-Pagan teszttel) a hibaváltozók heteroszkedaszticitását.





A következőkben teszteljük a heteroszkedaszticitást a Breusch-Pagan és a White tesztek segítségével. Ezt a segédmodellek OLS-becslésével valósítjuk meg.

A Breusch-Pagan teszt OLS-becslése (Eviewsban):

Dependent Variable: RES2
Method: Least Squares
Date: 11/21/06 Time: 11:20
Sample: 1980:01 1999:12
Included observations: 240

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D1	60.55578	8.572411	7.064032	0.0000
D2	28.25867	9.808829	2.880942	0.0043
RENDMARK	1.372169	0.699851	1.960659	0.0511
RENDMARK2	0.144705	0.085002	1.702368	0.0900
C	13.49269	4.161253	3.242459	0.0014

R-squared	0.205730	Mean dependent var	31.24531
Adjusted R-squared	0.192211	S.D. dependent var	54.25681
S.E. of regression	48.76446	Akaike info criterion	10.63249
Sum squared resid	558823.5	Schwarz criterion	10.70501
Log likelihood	-1270.899	F-statistic	15.21729
Durbin-Watson stat	2.051121	Prob(F-statistic)	0.000000

Az F-statisztika 15.22 és a p-értéke 0.000, ezért elutasítjuk a nullhipotézist, hogy a hibaváltozók homoszkedasztikusak.

A White teszt OLS-becslése (Eviewsban):

Dependent Variable: RES2
Method: Least Squares
Date: 11/21/06 Time: 12:03
Sample: 1980:01 1999:12
Included observations: 240

Variable	Coefficient	Std. Error	t-Statistic	Prob.
RENDMARK	1.535406	0.770801	1.991961	0.0475
RENDMARK2	0.172723	0.093518	1.846947	0.0660
C	26.26505	4.090727	6.420632	0.0000

R-squared	0.027067	Mean dependent var	31.24531
Adjusted R-squared	0.018857	S.D. dependent var	54.25681
S.E. of regression	53.74282	Akaike info criterion	10.81872
Sum squared resid	684524.8	Schwarz criterion	10.86223
Log likelihood	-1295.246	F-statistic	3.296727
Durbin-Watson stat	1.712122	Prob(F-statistic)	0.038708

Az F-statisztika 3.29 és a p-értéke 0.039, ezért $\alpha = 0.05$ szignifikancia szinten elutasítjuk a nullhipotézist, hogy a hibaváltozók homoszkedasztikusak.

4.1.4. Súlyozott legkisebb négyzetek módszere

Mivel tudjuk, hogy hogyan becsüljük, most tegyük fel, hogy ismerjük a hibaváltozók σ_i^2 varianciáit, és legalább kettő közülük nem egyenlő. Emlékezzünk

vissza, hogy ebben az esetben az OLS becslőfüggvény nem rendelkezik a legjobb lineáris torzítatlan becslőfüggvény tulajdonsággal, és az F-teszt nem helyes a becslések egyidejű szignifikanciájának tesztelésére, mivel az A3 feltétel nem teljesül.

Ennek a problémának a megoldására transzformáljuk a modell változóit úgy, hogy a hibaváltozók homoszkedasztikussá váljanak. Legyen az eredeti modell

$$y_i = x_i\beta + \varepsilon_i, \quad i = 1, ..., n,$$

ahol $E[\varepsilon_i^2] = \sigma_i^2$ ismertek. Transzformáljuk a változókat a következő módon. Legyen

$$y_i^* = \frac{y_i}{\sigma_i}, \quad x_i^* = \frac{x_i}{\sigma_i} \quad \varepsilon_i^* = \frac{\varepsilon_i}{\sigma_i}, \quad i = 1, ..., n.$$

Ha az eredeti modellben az A3-at kivéve a többi feltétel teljesül, akkor az

$$y_i^* = x_i^*\beta + \varepsilon_i^*, \quad i = 1, ..., n$$

modellben mind a 6 feltétel teljesül, mert

$$E[\varepsilon_i^{*2}] = \frac{E[\varepsilon_i^2]}{\sigma_i^2} = 1, \quad i = 1, ..., n,$$

vagyis a hibaváltozók varianciái egyenlőek. Ezért, ha OLS-sel becsüljük a transzformált modell β paraméter-vektorát, akkor a becslőfüggvény rendelkezik a legjobb lineáris torzítatlan becslőfüggvény tulajdonsággal, és az F-tesztek helyesek lesznek.

A transzformált modell OLS becslőfüggvénye:

$$\begin{aligned} b_{WLS} &= \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}y^* = \left(\sum_{i=1}^n x_i^{*\prime}x_i^*\right)^{-1} \sum_{i=1}^n x_i^{*\prime}y_i^* \\ &= \left(\sum_{i=1}^n \frac{1}{\sigma_i^2}x_i'x_i\right)^{-1} \sum_{i=1}^n \frac{1}{\sigma_i^2}x_i'y_i \\ &= \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}y, \end{aligned}$$

ahol

$$\Omega = E\left[\varepsilon\varepsilon'\right] = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}.$$

Mivel a b_{WLS} becslőfüggvényt a legkisebb négyzetek módszerével kapjuk, ahol a változókat $\frac{1}{\sigma_i}$-vel súlyozzuk, ezért ezt a módszert súlyozott legkisebb négyzetek módszerének (angol rövidítés szerint WLS-nek) nevezzük.
A becslőfüggvény varianciája:

$$var\left(b_{WLS}\right) = 1 \cdot \left(X^{*\prime}X^{*}\right)^{-1} = \left(X'\Omega^{-1}X\right)^{-1}.$$

A WLS becslőfüggvényt a következőképpen jellemezhetjük.

Állítás. Ha az $y = X\beta + \varepsilon$ modellben az A1, A2, A5 feltételek teljesülnek és

$$E\left[\varepsilon\varepsilon'\right] = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix},$$

akkor a β WLS becslőfüggvénye b_{WLS} torzítatlan, konszisztens és eloszlása $N\left(\beta, \left(X'\Omega^{-1}X\right)^{-1}\right)$.

Ez utóbbi állítás következtében az F-tesztek, amelyek statisztikáit a transzformált modell reziduumainak négyzetösszegéből vagy R^{*2} determinancia-együtthatójából számítjuk ki, helyesek.

A gyakorlatban Ω helyett az $\widehat{\Omega} = \begin{pmatrix} s_1^2 & & 0 \\ & \ddots & \\ 0 & & s_n^2 \end{pmatrix}$ mátrixot használjuk, ahol s_i^2-t, $i = 1, 2, ..., n$, a fent tárgyalt módon számítjuk ki. A kapott becslőfüggvényt megvalósítható WLS-nek (FWLS) nevezik:

$$b_{FWLS} = \left(X'\widehat{\Omega}^{-1}X\right)^{-1} X'\widehat{\Omega}^{-1}y.$$

b_{FWLS}-re szintén érvényes a fenti állítás.

4.2. Autokorreláció

Tekintsük a többváltozós lineáris modellt:

$$y_i = x_i\beta + \varepsilon_i, \quad i = 1, ..., n.$$

Azt a jelenséget amikor a hibaváltozók korreláltak, vagyis létezik legalább két $i \neq j$ megfigyelés amelyekre ε_i és ε_j kovarianciája nem nulla, vagyis,

$$E\left[\varepsilon_i\varepsilon_j\right] = \sigma_{ij} \neq 0,$$

autokorrelációnak nevezzük. Ebben az esetben az A4 feltétel nem teljesül, és a hibaváltozók variancia mátrixa

$$\Omega = E\left[\varepsilon\varepsilon'\right]$$

tartalmaz nemnulla elemeket a főátlón kívül.

4.2.1. Az OLS becslőfüggvény tulajdonságai

Tudjuk, hogy A1, A2, A5 mellett a b OLS becslőfüggvény torzítatlan és a variancia mátrixa

$$var\left(b\right) = \left(X'X\right)^{-1}\left(X'\Omega X\right)\left(X'X\right)^{-1}.$$

Ezen kívül, ugyanúgy mint heteroszkedaszticitás esetén, bizonyos elég általános feltételek mellett b konszisztens (vagyis $var\left(b\right) \underset{n\to\infty}{\longrightarrow} 0$).

Ugyanúgy, mint heteroszkedaszticitás esetén, A1, A2, A5, A6 mellett a b OLS becslőfüggvény normál eloszlású:

$$b \sim N\left(\beta, \left(X'X\right)^{-1}\left(X'\Omega X\right)\left(X'X\right)^{-1}\right).$$

A b varianciája miatt, ugyanúgy, mint heteroszkedaszticitás esetén, a becslések standard hibáit nem lehet a szokásos módon becsülni. A standard hibák becsléséhez előbb a $var\left(b\right)$ mátrixot kell becsülni, ami tartalmazza az ismeretlen Ω mátrixot.

Ugyanúgy, mint heteroszkedaszticitás esetén, az Ω mátrix elemeit nem lehet konszisztensen becsülni, viszont itt is lehetséges az $X'\Omega X$ mátrix konszisztens becslése. Ily módon, a $var\left(b\right)$ mátrix egy konszisztens becslése a Newey-West becslőfüggvény**:**

$$\widehat{var\left(b\right)} = \left(X'X\right)^{-1}\left(X'\widehat{\Omega}X\right)\left(X'X\right)^{-1},$$

ahol

$$\widehat{\Omega} = \begin{pmatrix} e_1^2 & w_1e_1e_2 & & w_{n-1}e_1e_n \\ w_1e_1e_2 & e_2^2 & & w_{n-2}e_2e_n \\ & & \ddots & \\ w_{n-1}e_1e_n & w_{n-2}e_2e_n & & e_n^2 \end{pmatrix}.$$

Akárcsak heteroszkedaszticitás esetén, ez a mátrix is a reziduumokat használja fel. Ebben az esetben a főátló kívüli elemek nem mind nullák és súlyozva vannak 1-nél kisebb súlyokkal:

$$w_i = \begin{cases} 1 - \frac{i}{B} & \text{ha } i < B \\ \\ 0 & \text{ha } i \geq B. \end{cases}$$

A B egy szám, ami $\approx n^{1/3}$, vagy nagyon nagy minta esetén $\approx n^{1/5}$. Ez a becslőfüggvény mind heteroszkedaszticitás, mind autokorreláció esetén konszisztens (ezért angolul a HAC rövidítés).

Megjegyzés. Ha ugyanúgy járnánk el mint heteroszkedaszticitás esetén, akkor $\widehat{\Omega}$

$$\begin{pmatrix} e_1^2 & e_1e_2 & & e_1e_n \\ e_1e_2 & e_2^2 & & e_2e_n \\ & & \ddots & \\ e_1e_n & e_2e_n & & e_n^2 \end{pmatrix}$$

lenne. Ez viszont nem jó, mert ekkor $\widehat{\Omega} = ee'$ és ezért $X'ee'X = 0$, vagyis $X'\widehat{\Omega}X = 0$. Ezért is szükséges $w_i < 1$ súlyokat használni. Megfigyelhető, hogy a súlyok csökkennek ahogy a megfigyelések indexei távolodnak egymástól.

Példa. Vegyünk hónaponkénti megfigyeléseket az 1950:01-1999:12 periódusra USA-beli 3 hónapos kincstári jegyek (US3MT) és AAA kötvények hozamaira, pontosabban, legyen DAAA az AAA kötvény és DUS3MT a három hónapos kincstári jegyek hozamának százalékbeli változása egyik hónapról a másikra. Az alább becsült modellben az utóbbi hatását vizsgáljuk az előbbire.OLS becslés és HAC standard hibák:

Dependent Variable: DAAA
Method: Least Squares
Sample: 1950:01 1999:12
Included observations: 600

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DUS3MT	0.274585	0.014641	18.75442	0.0000
C	0.006393	0.006982	0.915697	0.3602

R-squared	0.370346	Mean dependent var	0.008283
Adjusted R-squared	0.369293	S.D. dependent var	0.215322
S.E. of regression	0.171002	Akaike info criterion	-0.690952
Sum squared resid	17.48658	Schwarz criterion	-0.676296
Log likelihood	209.2857	F-statistic	351.7282
Durbin-Watson stat	1.446887	Prob(F-statistic)	0.000000

Dependent Variable: DAAA				
Method: Least Squares				
Sample: 1950:01 1999:12				
Included observations: 600				
Newey-West HAC Standard Errors & Covariance (lag truncation=5)				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
DUS3MT	0.274585	0.021187	12.95993	0.0000
C	0.006393	0.008309	0.769436	0.4419

ÖSSZEGZÉS

- Autokorreláció esetén a β paraméter-vektor OLS becslőfüggvénye helyes (konszisztens), de a standard hibák kiszámításához a variancia HAC (Newey-West) becslőfüggvényét kell használjuk. Ezeket HAC (Newey-West) standard_hibáknak nevezik.

- Ezen standard hibákból kiszámított t-statisztikákat és az ehhez tartozó kritikus értékeket kell használjuk konfidencia intervallumok és intervallum-előrejelzések meghatározásához.

- Autokorreláció esetén az OLS becslőfüggvény nem rendelkezik a legjobb lineáris torzítatlan becslőfüggvény tulajdonsággal. Ez viszont nem nagy probléma.

- Komolyabb probléma, hogy az F-tesztet (például egyidejű szignifikancia tesztelésére) nem tudjuk helyesen elvégezni az eredeti eljárás alapján. Emiatt tovább fogjuk tanulmányozni a lineáris modell becslését autokorreláció esetén.

Mi okozza az autokorrelációt?

Az autokorreláció, hasonlóan a heteroszkedaszticitáshoz, leggyakrabban azért fordul elő mert a modell nincs helyesen meghatározva, vagyis az A5 feltétel nem teljesül. Például, a modellből hiányoznak egyes független változók, amelyek leggyakrabban az eredeti modell változóinak késleltetett értékei.

Az y_i változó k-adik késleltetett értéke az y_{i-k} változó. Idősor-modelleknél a függővváltozó gyakran modellezhető a késleltetett értékei függvényében, és ezért, ha ezek a késleltetett értékek hiányoznak a modellből, a hibaváltozók korreláltak lesznek. Például, a fenti modellben elképzelhető, hogy az AAA kötvény hozamának változását nemcsak az USA-beli 3 hónapos kincstári

jegyek hozamának jelenlegi, hanem az egy hónappal korábbi változása is befolyásolja.

Példa. Tegyük fel, hogy az $y_i = x_i\beta + \varepsilon_i$ modellben a hibaváltozók korreláltak úgy, hogy $\varepsilon_i = \rho\varepsilon_{i-1} + \omega_i$, $i = 1, 2, ..., n$, ahol ω_i nem korreláltak, ugyanolyan eloszlásúak és függetlenek $\varepsilon_{i-1}, \varepsilon_{i-2}, ..., \varepsilon_0$-tól. Ekkor, ha az első késleltetett értékekből alkotott

$$y_{i-1} = x_{i-1}\beta + \varepsilon_{i-1}, \quad i = 2, ..., n$$

modellt beszorozzuk ρ-val és kivonjuk az eredetiből, a következőt kapjuk:

$$y_i = \rho y_{i-1} + x_i\beta - x_{i-1}\rho\beta + \omega_i, \quad i = 2, ..., n.$$

Ha a $\gamma = -\rho\beta$ jelölést használjuk, egy új lineáris modellt kapunk:

$$y_i = \rho y_{i-1} + x_i\beta + x_{i-1}\gamma + \omega_i, \quad i = 2, ..., n,$$

amelyben a hibaváltozók nem korreláltak (tehát A4 teljesül) és tartalmazza az eredeti modell függő- és független változójának a késleltetett értékeit. Annak ellenére, hogy az eredeti modellben (A1, A2, A5 mellett) helyesen tudjuk becsülni a β paraméter-vektort (vagyis, az x hatását az y-ra) jobb az utóbbi modellt becsülni, mert ebben a β OLS becslőfüggvénye rendelkezik a legjobb lineáris torzítatlan becslőfüggvény tulajdonsággal (amennyiben A3 is teljesül) és a standard hibák, F-tesztek és intervallum-előrejelzések helyesek lesznek. Fontos megjegyezni, hogy az új modellben A1 nem teljesül, mert az y_{i-1} független változó vv. Ennek ellenére az előbbi állítások megközelítőleg helyesek, ahol a megközelítés annál jobb, minél nagyobb a megfigyelések száma.

Megjegyzés. A fenti példában szereplő autokorreláción kívül van más típusú auto-korreláció is. Egy ilyen példa amikor a hibaváltozókat az

$$\varepsilon_i = \lambda\omega_i + \omega_{i-1}, \quad i = 1, 2, ..., n$$

összefüggés határozza meg, ahol ω_i, $i = 1, 2, ..., n$ nem korreláltak és normál eloszlásúak. Ebben az esetben $E[\varepsilon_i\varepsilon_{i-1}] = \lambda\sigma_\omega^2$ és $E[\varepsilon_i\varepsilon_{i-k}] = 0$ ha $k \geq 2$, ahol $\sigma_\omega^2 = var(\omega_i)$.

4.2.2. Hogy ismerjük fel az autokorrelációt?

Mint a heteroszkedaszticitás vizsgálatánál, az autokorreláció vizsgálata lényegében a hibaváltozók kovarianciáját kell vizsgálja. Mivel ez közvetett módon nem

végezhető, ezért a reziduumok segítségével történik. Ebben az esetben is ábrák segítségével kezdjük:
- a reziduumok vonalgrafikonján megnézzük, hogy az egymás után következő értékek összefüggnek-e,
- a reziduumok és a késleltetett értékeik szóródási képe megmutatja, hogy van-e összegüggés közöttük,
- kiszámítjuk az autokorrelációs együtthatókat (lásd alább).

Ezek az ábrák helyes képet mutatnak A1, A2, A5 mellett ha a megfigyelések száma elég nagy, mert ekkor a b OLS becslés konszisztens, és a reziduumok megközelítőleg egyenlőek a hibaváltozókkal:

$$e_i = y_i - x_i b \simeq y_i - x_i \beta = \varepsilon_i.$$

4.2.3. Autokorrelációs együtthatók

Az első rendű autokorrelációs együttható:

$$r_1 = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}.$$

Megközelítőleg egyenlő az e_i és az e_{i-1} korrelációs együtthatójával, ugyanis a reziduumok megfigyelései alapján felhasználva, hogy $var\,(e_i) \approx var\,(e_{i-1})$:

$$\begin{aligned} corr\,(e_i, e_{i-1}) &= \frac{cov\,(e_i, e_{i-1})}{\sqrt{var\,(e_i)\,var\,(e_{i-1})}} \\ &\approx \frac{\frac{1}{n-1}\sum_{i=2}^{n} e_i e_{i-1} - \left(\frac{1}{n-1}\sum_{i=2}^{n} e_i\right)\left(\frac{1}{n-1}\sum_{i=2}^{n} e_{i-1}\right)}{\frac{1}{n}\sum_{i=2}^{n} e_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} e_i\right)^2}. \end{aligned}$$

Tudjuk, hogy $\sum_{i=1}^{n} e_i = 0$, ezért $\frac{1}{n-1}\sum_{i=2}^{n} e_i \approx 0$; ezenkívül $\frac{n}{n-1} \approx 1$, tehát

$$corr\,(e_i, e_{i-1}) \approx \frac{n}{n-1}\frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=2}^{n} e_i^2} \approx \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=2}^{n} e_i^2}.$$

A k-ad rendű autokorrelációs együttható:

$$r_k = \frac{\sum_{i=k+1}^{n} e_i e_{i-k}}{\sum_{i=1}^{n} e_i^2}.$$

Megközelítőleg egyenlő az e_i és az e_{i-k} korrelációs együtthatójával. Az autokorrelációs együtthatók SPSS-ben kiszámíthatók. Ha a hibaváltozók nem korreláltak

(A4 teljesül), akkor az autokorrelációs együtthatók megközelítőleg normál eloszlásúak:

$$r_k \approx N\left(0, \frac{1}{n}\right).$$

Ezt lehet egy nem formális tesztnek tekinteni, amelyben a nullhipotézis az, hogy a hibaváltozók nem korreláltak: H_0-t nem utasítjuk el, ha $|r_k| < \frac{2}{\sqrt{n}}$. Ezt a nem formális tesztet arra használjuk, hogy kiválasszuk a legszignifikánsabb autokorrelációs együtthatókat.

A hibaváltozók autokorrelációs modellje

A felsorolt ábrák és autokorrelációs együtthatók egy első benyomást nyújtanak a hibaváltozók korrelációjáról és a legszignifikánsabb autokorrelációs együtthatókról. Ezek alapján:
- megszerkeszthetjük a hibaváltozók autokorrelációs modelljét,
- tesztelhetjük formálisan az autokorrelációt úgy, hogy teszteljük a reziduumok késleltetett értékeinek az egyidejű szignifikanciáját,
- a tesztek alapján kiigazíthatjuk a modellt késleltetett értékek figyelembe vételével, hogy nem korrelált hibaváltozójú modellt kapjunk.

Azt már láttuk, hogy nem lehet a hibaváltozók Ω variancia mátrixának (ami a kovarianciákat tartalmazza) elemeit becsülni. Ezért a fenti ábrák és autokorrelációs együtthatók alapján megszerkesztjük a hibaváltozók autokorrelációs modelljét. Ez az eljárás elvileg hasonló a heteroszkedaszticitásnál használt eljárásra.

Tegyük fel, hogy az ábrák és autokorrelációs együtthatók alapján úgy tűnik, hogy a hibaváltozók első h késleltetett értéke szignifikáns. Ezt formálisan tesztelhetjük a

$$e_i = x_i\beta + \gamma_1 e_{i-1} + ... + \gamma_h e_{i-h} + \eta_i$$

modell segítségével, ha az $e_{i-1}, ..., e_{i-h}$ egyidejű szignifikanciáját teszteljük, ahol e_i az eredeti modell reziduuma (tudjuk, hogy $e_i \approx \varepsilon_i$). Ezt a modellt a hibaváltozók autokorrelációs modelljének nevezik.

Az autokorreláció tesztelése

Ha a fenti modellben az $e_{i-1}, ..., e_{i-h}$ reziduumok késleltetett értékei egyidejűleg nem szignifikánsak, akkor az eredeti modell hibaváltozói valószínűleg nem korreláltak. Ez az alapja a Breusch-Godfrey autokorrelációs tesztnek. Tehát, a teszt nullhipotézise

$$H_0 : \gamma_1 = ... = \gamma_h = 0,$$

az alternatív hipotézis az, hogy legalább egyik γ_j különbözik 0-tól.

A Breusch-Godfrey teszt lépései:
(i) Becsüljük OLS-sel az $y_i = x_i\beta + \varepsilon_i$ eredeti modellt és kiszámítjuk a reziduumokat: $e_i = y_i - x_i b$, $i = 1, ..., n$. Megvizsgáljuk a reziduumokat a felsorolt ábrák és az autokorrelációs együtthatók segítségével, és kiválasztjuk a legszignifikánsabb késleltetett értékeket.
(ii) Becsüljük OLS-sel a segédmodellt:

$$e_i = x_i\beta + \gamma_1 e_{i-1} + ... + \gamma_h e_{i-h} + \eta_i, \quad i = h+1, ..., n.$$

(iii) Teszteljük a $\gamma_1 = ... = \gamma_h = 0$ nullhipotézist a segédmodellben egy F egyidejű szignifikancia teszttel. A teszt-statisztika

$$F = \frac{(SSR_R - SSR_U)/h}{SSR_U/(n-k-2h)} \approx F(h, n-k-2h),$$

ahol SSR_R és SSR_U a leszűkített és az eredeti modellek reziduumainak a négyzetösszege.

Durbin-Watson teszt
A statisztikája:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2},$$

amit Durbin-Watson statisztikának neveznek. A következő megfigyelésen alapul:

$$d = \frac{\sum_{i=2}^{n}\left(e_i^2 - 2e_{i-1}e_i + e_{i-1}^2\right)}{\sum_{i=1}^{n} e_i^2} = \frac{\sum_{i=2}^{n}\left(e_i^2 + e_{i-1}^2\right)}{\sum_{i=1}^{n} e_i^2} - 2\frac{\sum_{i=2}^{n} e_{i-1}e_i}{\sum_{i=1}^{n} e_i^2}$$
$$\simeq 2(1 - r_1),$$

ami annál pontosabb, minél nagyobb a megfigyelések száma. A $corr(\varepsilon_i, \varepsilon_{i-1}) = 0$ nullhipotézist teszteli. Mivel $-1 \le r_1 \le 1$, a Durbin-Watson statisztika 0 és 4 között van: $0 \le d \le 4$. A $d \approx 2$ érték $r_1 \approx 0$-nak felel meg, pontosabban, ha $1.6 < d < 2.4$, akkor az első rendű korreláció ≈ 0. Ahhoz, hogy az 1.6 és 2.4 kritikus értékek helyesek legyenek, a következő feltételek kell teljesüljenek:
- a hibaváltozók normál eloszlásúak,
- a független változók nem vvk.

Ezek miatt a Durbin-Watson teszt nem elég általános. Viszont az értékét könnyen megkaphatjuk SPSS-ben.

Box-Ljung teszt
Ez egy másik teszt, amit könnyen elvégezhetünk SPSS-ben. Nullhipotézise az, hogy az első h autokorrelációs együttható egyidejűleg nem szignifikáns. Statisztikáját Q-statisztikának nevezik:

$$Q = n \sum_{k=1}^{h} \frac{n+2}{n-k} r_k^2 \sim \chi^2 (h) .$$

A Durbin-Watson teszthez hasonlóan, nem helyes, ha az eredeti modell független változói vvk.

Tehát, ha a modell tartalmazza a függővváltozó késleltetett értékét, y_{i-1}-et, akkor a Durbin-Watson és a Box-Ljung tesztek nem helyesek. Tehát, a Breusch-Godfrey teszt általánosabb mint a másik kettő, de több számítást igényel, ugyanis a segédmodell OLS-becslését is el kell végezzük.

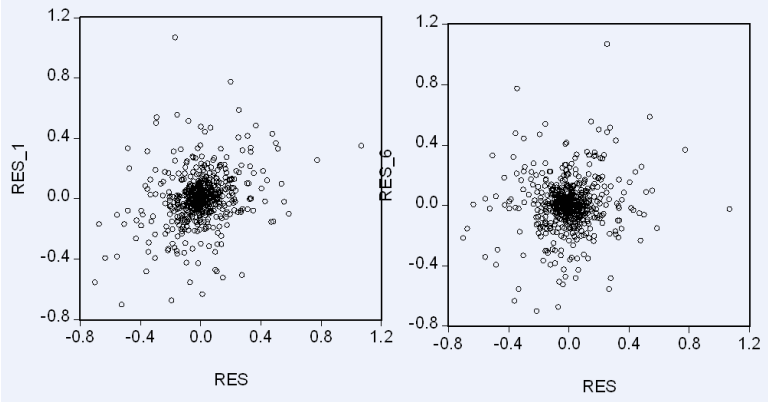
A modell kiigazítása
Ha a tesztek következtetése az, hogy a hibaváltozók korreláltak, akkor az eredeti modell változóinak késleltetett értékeit bevesszük a modellbe független változónak. Ezt az eljárást a modell kiigazításának nevezik.
A fenti példában a kiigazított modell:

$$y_i = \gamma_1 y_{i-1} + x_i \gamma_2 + x_{i-1} \gamma_3 + \omega_i, \quad i = 2, ..., n.$$

A kiigazított modellt becsüljük OLS-sel és teszteljük (a Breusch-Godfrey teszttel), hogy a hibaváltozói korreláltak-e. Ha igen, akkor tovább folytatjuk a modell kiigazítását addig, amíg olyan modellt kapunk, amelynek a hibaváltozói nem korreláltak.

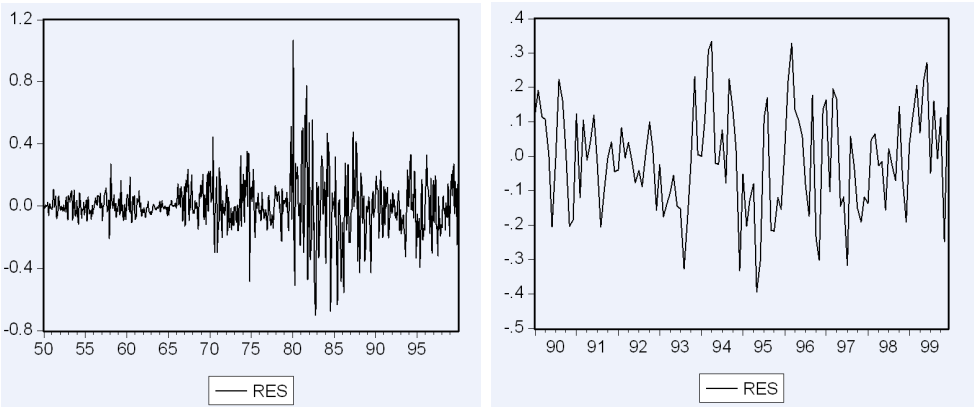
Példa: Az USA kincstári jegy és az AAA kötvény kapcsolata
Becsüljük a modellt OLS-sel (lásd az eredményt fennebb). A Durbin-Watson teszt ($D - W = 1.447$) szerint az első rendű autokorreláció nem nulla.

Ábrákat készítünk. A reziduumok és az első késleltetett értékeinek valamint a reziduumok és az hatodik késleltetett értékeinek a szóródási képe:



Az előbbi pontfelhője mintha enyhén enyúlt lenne, ami mutat egy gyenge pozitív lineáris összefüggést; az utóbbi alapján nem tudunk hasonló következtetést levonni.

A reziduumok vonalgrafikonjai közül a részletesebb (a jobboldali, ami csak az utolsó 10 év megfigyeléseit tartalmazza) mutat összefüggést az egymás utáni étékek között. (Vonalgrafikon alapján az autokorreláció jobban felismerhető, ha nem korrelált változók vonalgrafikonjához hasonlítjuk mint alább a 6. fejezetben.)



Az autokorrelációk közül az 1. és a 6. a legnagyobb (a kritikus érték $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{600}} = 0.082$), és csak ezek szignifikánsak.

Autocorrelations

Series:res

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	.276	.041	45.932	1	.000
2	-.076	.041	49.398	2	.000
3	.008	.041	49.441	3	.000
4	.034	.041	50.126	4	.000
5	.055	.041	51.939	5	.000
6	.101	.041	58.189	6	.000
7	.035	.041	58.934	7	.000
8	.049	.040	60.412	8	.000
9	.044	.040	61.610	9	.000
10	.008	.040	61.646	10	.000
11	.032	.040	62.289	11	.000
12	-.062	.040	64.624	12	.000

a. The underlying process assumed is independence (white noise).
b. Based on the asymptotic chi-square approximation.

Megszerkesztjük a hibaváltozók modelljét, amelybe a reziduumok ezen késleltetett értékeit vesszük be. A Breusch-Godfrey teszttel teszteljük az autokorrelációt:

$$F\left[\approx F\left(2,590\right)\right]=\frac{\left(17.487-15.956\right)/2}{15.956/\left(594-4\right)}=28.306.$$

Ez nagyobb mint a 3 kritikus érték, ezért elutasítjuk a nullhipotézist, tehát a hibaváltozók korreláltak. A teszthez szükséges reziduum-négyzetösszegeket alábbi OLS becslési eredményekből kaptuk SPSS-ben. A leszűkítetlen modell:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.530	3	.510	18.858	.000[a]
	Residual	15.956	590	.027		
	Total	17.486	593			

a. Predictors: (Constant), res__6, DUS3MT, res__1
b. Dependent Variable: res

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.000	.007		.038	.970
	DUS3MT	-.022	.014	-.063	-1.555	.120
	res__1	.285	.040	.285	7.063	.000
	res__6	.087	.040	.087	2.201	.028

a. Dependent Variable: res

A leszűkített modell:

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.000	1	.000	.000	1.000[a]
	Residual	17.487	598	.029		
	Total	17.487	599			

a. Predictors: (Constant), DUS3MT
b. Dependent Variable: res

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.725E-19	.007		.000	1.000
	DUS3MT	.000	.015	.000	.000	1.000

a. Dependent Variable: res

4.3. Fontos változók kihagyása

Ebben az alfejezetben azt vizsgáljuk, hogy mi történik, ha fontos változókat hagyunk ki a modellből. Tegyük fel, hogy a valódi modell

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon, \tag{4.2}$$

de mi az x_3 változót kihagyjuk, és az

$$y = \delta_1 + \delta_2 x_2 + \omega \tag{4.3}$$

modellt becsüljük (pl. elsősorban az x_2 hatására vagyunk kíváncsiak).

Ebben az esetben az A5 feltétel nem teljesül.

Kérdés: Milyen tulajdonságokkal rendelkezik a δ_2 OLS becslőfüggvénye, ha feltételezzük, hogy fennáll az

$$x_3 = \gamma_1 + \gamma_2 x_2 + u$$

összefüggés? Itt γ_1, γ_2 paraméterek, és u vv, tehát valamilyen lineáris összefüggés van x_2 és x_3 között.

A kérdésre választ kapunk, ha x_3-at behelyettesítjük a valódi modellbe:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 (\gamma_1 + \gamma_2 x_2 + u) + \varepsilon,$$

ahonnan az

$$y = \beta_1 + \beta_3 \gamma_1 + (\beta_2 + \beta_3 \gamma_2) x_2 + \varepsilon + \beta_3 u$$

összefüggést kapjuk. Ez egy lineáris modellként értelmezhető, amelynek független változója x_2, amelynek együtthatója $(\beta_2 + \beta_3\gamma_2)$.

Tehát, ha egy olyan modellben, amelyben mind az x_2 mind az x_3 befolyásolja a függőváltozót, csak az x_2-t vesszük figyelembe, akkor a β_2 paraméter helyett a $(\beta_2 + \beta_3\gamma_2)$-t becsüljük. Ezért a kapott OLS becslés nem a β_2-t hanem a $(\beta_2 + \beta_3\gamma_2)$-t közelíti meg. Ha $\beta_3\gamma_2 \neq 0$, akkor a becslés nem lesz helyes (konszisztens) és torzított lesz.

Fontos megjegyezni, hogy $\beta_3\gamma_2 \neq 0$ akkor fordul elő, ha $\gamma_2 \neq 0$, ami azt jelenti, hogy x_2 és x_3 között van valamilyen lineáris összefüggés és $\beta_3 \neq 0$, ami azt jelenti, hogy x_3 hatással van a függőváltozóra. Ha ezen feltételek közül valamelyik nem teljesül, akkor a (4.3) modell OLS becslése helyes lesz.

Példa. Ha fontos változót hagyunk ki a modellből, előfordulhat, hogy egy változó hatását a függőváltozóra helytelenül negatívnak becsüljük pozitív helyett. Ez, például, akkor fordulhat elő, ha $\beta_2, \beta_3 > 0$ és $\gamma_2 < 0$.

Például, vegyük az autótípusokra azt a modellt, amelyben az ár a függőváltozó. A megfigyelések 2014-re 571 autótípusra vonatkoznak Németországban. Ebben a példában az üzemanyagfogyasztás hatékonysága (km/l-ben; azt mutatja, 1 liter üzemanyaggal hány km-t tud megtenni) és az ár között negatív összefüggés van, de ha a méret is szerepel a modellben, akkor a km/l hatása az árra pozitívvá változik. Az alábbi táblázatokban Ar=ár; Lept=lóerő/tömeg; Mer=méret; Kmpl=km/l:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.732	.317		-2.311	.021
	Kmpl	-.305	.076	-.139	-3.999	.000
	Lept	1.142	.070	.570	16.340	.000

a. Dependent Variable: ar

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-6.600	.307		-21.480	.000
	Kmpl	.247	.055	.113	4.478	.000
	Lept	1.103	.047	.550	23.524	.000
	Mer	1.741	.066	.630	26.410	.000

a. Dependent Variable: ar

Az utóbbi táblázat alapján valóban $\beta_2, \beta_3 > 0$, míg a következő táblázat

azt mutatja, hogy ebben a példában $\gamma_2 < 0$:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.447	.089		38.678	.000
	Kmpl	-.327	.030	-.412	-10.796	.000

a. Dependent Variable: Mer

4.4. Irreleváns változók a modellben

Ez az alfejezet azt vizsgálja, hogy mi történik, ha olyan változók szerepelnek a modellben, amelyek nem befolyásolják a függővváltozót. Ebben az esetben a valódi modell

$$y = \beta_1 + \beta_2 x_2 + \varepsilon,$$

míg mi egy olyan modell-t becsülünk, amelyben az x_3 független változó is szerepel:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

Ebben az esetben az utóbbi modellben $\beta_3 = 0$, tehát az A5 feltétel teljesül. Ezért, ha a többi feltétel is teljesül, akkor az OLS becslőfüggvény torzítatlan, konszisztens és normál eloszlású (ezért a t és F-tesztek is helyesek lesznek).

Viszont nem rendelkezik a legjobb, lineáris torzítatlan tulajdonsággal, mert a becslőfüggvény varianciája (és a standard hibák) nagyobb mint a valódi modell becslése esetén. Tulajdonképpen minél jobban korrelál az x_3 változó a többi független változóval, annál nagyobb lesz a variancia (és a standard hibák).

Példa. Szélsőséges esetben tegyük fel, hogy a két változó korrelációja megközelítőleg 1. Ekkor az $(X'X)^{-1}$ mátrix főátlójának 2. és 3. eleme túl nagyok lesznek. Ezt a jelenséget hívják multikollinearitásnak; erről lesz szó a következő alfejezetben.

ÖSSZEGZÉS

- A fenti érvek alapján jobb több változót bevenni a modellbe független változónak mint kevesebbet, ugyanis ekkor nagyobb eséllyel elkerüljük, hogy az OLS becslőfüggvény nem konszisztens.

- OLS becslés után az egyéni és egyidejű szignifikancia tesztek segítségével el lehet dönteni, hogy egy vagy több változót kihagyjunk-e a modellből. A kiigazított R^2 is használható erre a célra.

- Viszont, ha a megfigyelések száma nem elég nagy, akkor nem ajánlatos túl sok független változót használni, mert a tesztek szabadsági foka $(n-k)$ nem lesz elég nagy, és ezért a tesztek pontatlanok lesznek, ami helytelen következtetéshez vezethet a szignifikanciákat illetően.

4.5. Multikollinearitás

Multikollinearitás esetén az A1-A6 feltételek teljesülnek, de az OLS becslések mégsem hasznosak olyan értelemben, hogy a nagy standard hibák miatt elutasítjuk a becslések szignifikanciáját. Ezt a jelenséget az okozza, hogy legalább két független változó nagyon korrelál egymással, és ezért az $X'X$ mátrix megközelítőleg egyenlő egy nem invertálható mátrixszal.

Példa. Vegyük az

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

modellt, ahol $x_3 = x_2 + a \cdot u$, ahol a fokozatosan közelít 0-hoz. Ekkor az $(X'X)^{-1}$ mátrix főátlójának 2. és 3. eleme egyre nő:

$$a = 0.5\text{-re} \qquad (X'X)^{-1} = \begin{pmatrix} 0.087 & 0.073 & -0.142 \\ 0.073 & 0.569 & -0.462 \\ -0.142 & -0462 & 0.475 \end{pmatrix}$$

$$a = 0.1\text{-re} \qquad (X'X)^{-1} = \begin{pmatrix} 0.070 & \cdot & \cdot \\ \cdot & 13.064 & \cdot \\ \cdot & \cdot & 12.828 \end{pmatrix}$$

$$a = 0.05\text{-re} \qquad (X'X)^{-1} = \begin{pmatrix} 0.068 & \cdot & \cdot \\ \cdot & 49.229 & \cdot \\ \cdot & \cdot & 49.083 \end{pmatrix}$$

$$a = 0.01\text{-re} \qquad (X'X)^{-1} = \begin{pmatrix} 0.076 & \cdot & \cdot \\ \cdot & 1364.455 & \cdot \\ \cdot & \cdot & 1362.399 \end{pmatrix}$$

Emlékezzünk vissza, hogy a standard hibák az $s^2 (X'X)^{-1}$ mátrix főátlóján lévő elemek négyzetgyöke. Ha túl nagyok az $(X'X)^{-1}$ mátrix főátlóján lévő elemek, akkor a standard hibák is túl nagyok lesznek.

Multikollinearitás más esetekben is előfordulhat. Az alábbi esetek mindegyikét multikollinearitásnak tekintjük:
- ha valamelyik független változó kevésbé váltakozik, ekkor ezen független változók vektora kollineáris az 1-esekből álló oszloppal,
- bármilyen két független változó, melyek nagyon korrelálnak egymással,
- több független változó lineárisan nagyon szorosan összefügg.
Mindegyik esetben az $(X'X)^{-1}$ mátrix viselkedése az előző oldalon szemléltetetthez hasonló.

Hogyan észleljük a multikollinearitást?

Ha a modell paramétereit OLS-sel becsüljük, és a standard hibák túl nagyok (még viszonylag nagy számú megfigyelés esetén is), akkor ezt okozhatja multikollinearitás. Egyszerű számítások útján megerősítést nyerhetünk a multikollinearitás feltételezésében. Például, kiszámítjuk a független változók szórását és páronkénti korrelációs együtthatójukat.

Azt az esetet, amikor a változók páronként nem korrelálnak hanem három változó között áll fenn multikollinearitás, másképp kell vizsgáljuk. Például, ha x_2 és x_3 két nem korrelált független változó, és $x_4 = x_2 + x_3$, akkor x_2 és x_4 valamint x_3 és x_4 nem biztos, hogy nagyon korrelál, de x_2, x_3 és x_4 kollineárisak mivel $x_4 = x_2 + x_3$.

A kérdés tehát az, hogy hogyan tudjuk felfeldezni, hogy fennáll-e szoros lineáris összefüggés a független változók között.

A megoldás az, hogy mindegyik független változóra becsüljük OLS-el azokat a modelleket, amelyben az illető független változót függőváltozónak vesszük, míg a többi a független változó, és kiszámítjuk az R^2-eket. Ha a legnagyobb R^2 közel van 1-hez, akkor az ennek megfelelő független változó okozza a multikollinearitást. Ha a legnagyobb R^2 nincs közel 1-hez, akkor nincs multikollinearitás.

Példa. Tegyük fel, hogy a lineáris modell

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon,$$

vagyis x_2, x_3, x_4, a független változók. Ekkor a következő három modell OLS

becslését határozzuk meg:

$$\begin{aligned} x_2 &= \alpha_1 + \alpha_2 x_3 + \alpha_3 x_4 + \omega, \quad R^2_{x_2} \\ x_3 &= \alpha_1 + \alpha_2 x_2 + \alpha_3 x_4 + \omega, \quad R^2_{x_3} \\ x_4 &= \alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \omega, \quad R^2_{x_4}, \end{aligned}$$

és kiválasztjuk a legnagyobb R^2-t.

Összefüggés a standard hibákkal

Legyen b_4 a β_4 OLS becslőfüggvénye a fenti modellben.Ekkor az A1-A5 feltételek mellett a b_4 variancáját az alábbi képlet adja meg:

$$var\left(b_4\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_{4i} - \overline{x}_4\right)^2\left(1 - R^2_{x_4}\right)}, \tag{4.4}$$

ahol $R^2_{x_4}$ annak a modellnek a determinancia együtthatója, amelyben x_4 a függőváltozó. Innen látszik, hogy ha $R^2_{x_4}$ közel van 1-hez, akkor $var\left(b_4\right)$ nagy, ezért a b_4 standard hibája is nagy.

Mi a teendő multikollinearitás esetén?

A következő lehetőségek közül választhatunk.

(i) Használjunk több megfigyelést. Ekkor $\sum_{i=1}^{n}\left(x_{4i} - \overline{x}_4\right)^2$ a (4.4) képleten nagyobb lesz, ezért $var\left(b_4\right)$ csökken (ha $R^2_{x_4}$ nem változik jelentősen). Ez az út gyakran nem járható. Ilyenkor a következő eljárás ajánlott.

(ii) Hagyjuk ki a modellből azt a független változót, amelyik okozza a multikollinearitást. Hogy megértsük, miért, tegyük fel, hogy az

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

modellben x_2, x_3, x_4, kollineárisak úgy, hogy

$$x_4 = \alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \omega,$$

ahol ω elhanyagolhatóan kicsi. Behelyettesítve ezt a modellbe,

$$\begin{aligned} y &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4\left(\alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \omega\right) + \varepsilon \\ &\simeq \left(\beta_1 + \beta_4\alpha_1\right) + \left(\beta_2 + \beta_4\alpha_2\right) x_2 + \left(\beta_3 + \beta_4\alpha_3\right) x_3 + \varepsilon, \end{aligned}$$

ahol elhagytuk $\beta_4\omega$-t. Tehát, ha x_4-et kihagyjuk a modellből, az x_2 és x_3 hatását y-ra helyesen tudjuk becsülni a $\gamma_1 = \left(\beta_1 + \beta_4\alpha_1\right)$, $\gamma_2 = \left(\beta_2 + \beta_4\alpha_2\right)$, $\gamma_3 = \left(\beta_3 + \beta_4\alpha_3\right)$ paraméterek becslésével.

4.6. Strukturális törés tesztelése

Előfordulhat, hogy olyan megfigyeléseink vannak a paraméterek becslésehez, amelyek hosszabb időszakra vonatkoznak. Ekkor felmerül a kérdés, hogy állandók maradnak-e a modell paraméterei.

Azt a jelenséget, amikor megváltoznak a modell paraméterei, strukturális törésnek hívják. Például, az autóárak modelljében, ha régebbi adatokat is használnánk, akkor elképzelhető, hogy változnak a paraméterek, ugyanis Németországban a gazdasági válság következményeként támogatást vezettek be 2010-ben új autók vásárlására. Ezek a támogatások 2014-ben már nem voltak érvényben. Ebben az esetben az A5 feltétel nem érvényes.

Legyen a modell az első és második periódusra

$$y_1 = X_1\beta_1 + \varepsilon_1, \quad y_2 = X_2\beta_2 + \varepsilon_2,$$

ahol y_1 és X_1 valamint y_2 és X_2 a függő és független változók az első és második periódusban. A célunk az, hogy teszteljük a $H_0 : \beta_1 = \beta_2$ nullhipotézist, ahol $H_1 : \beta_1 \neq \beta_2$.

Ha a két modellben teljesülnek az A1-A6 feltételek, akkor egy F-tesztet használhatunk az alábbi modellben

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

a $\beta_1 = \beta_2$ leszűkítésre. A tesztstatisztika

$$F = \frac{(S_0 - S_1 - S_2)/k}{(S_1 + S_2)/(n-2k)} \sim F(k, n-2k),$$

ahol
S_0 az eredeti (leszűkített) modell reziduumainak a négyzetösszege
S_1 az első modell reziduumainak a négyzetösszege
S_2 a második modell reziduumainak a négyzetösszege.

Ezt a tesztet Chow strukturális törési tesztjének hívják. A tesztstatisztikát az egyidejű szignifikancia tesztek alapján határozták meg.

ÖSSZEGZÉS

A 3. fejezetben bevezettünk 6 feltételt (A1-A6), amelyek mellett a többváltozós lineáris modell OLS becslése torzítatlan, konszisztens és normál eloszlású legjobb lineáris torzítatlan becslőfüggvény. Ebben a fejezetben olyan eseteket tanulmányoztunk, amelyekben egyes feltételek nem teljesülnek.

Heteroszkedaszticitás (A3 nem teljesül) vagy autokorreláció.(A4 nem teljesül) esetén a paraméter-vektor az OLS becslőfüggvénye helyes (konszisztens), de a standard hibák kiszámításához a variancia White (heteroszkedaszticitás) vagy HAC (Newey-West) (heteroszkedaszticitás és autokorreláció) becslőfüggvényét kell használjuk. Ezen standard hibákból kiszámított t-statisztikákat és az ehhez tartozó kritikus értékeket kell használjuk konfidencia intervallumok és intervallum-előrejelzések meghatározásához. Heteroszkedaszticitás vagy autokorreláció esetén az OLS becslőfüggvény nem rendelkezik a legjobb lineáris torzítatlan becslőfüggvény tulajdonsággal. Az F-teszteket nem tudjuk helyesen elvégezni az eredeti eljárás alapján. Ezért, hogy olyan modellt nyerjünk, amelyik teljesíti az A1-A6 feltételeket (esetleg A1 kivételével), transzformáljuk a modell változóit (heteroszkedaszticitás esetén) vagy a modellbe bevesszük az eredeti modell változóinak egyes késleltetett értékeit (auto-korreláció esetén).

Abban az esetben, ha fontos változókat hagyunk ki a modellből, az A5 feltétel nem teljesül. Ha a kihagyott változó korrelál valamelyik függőváltozóval, akkor az OLS becslőfüggvény torzított lesz. Ahhoz, hogy ezt elkerüljük, érdemes minél több függőváltozót bevenni a modellbe. Ezek közül egyéni és egyidejű szignifikancia-tesztekkel szabaduljunk meg a nem szignifikánsaktól. Vigyázzunk, hogy kerüljük el a multikollinearitást, ami a szignifikancia-tesztek eredményét torzíthatja.

4.7. Gyakorlatok

1. Vegyük azt a modellt, ami összefüggést állít az összes piaci nettó hozam (RENDMARK) és a nem ciklikus fogyasztói termékek piacának nettó hozama (RENDNCCO) között.

   (a) Becsüljük OLS-sel a

   $$RENDNCCO = \beta_1 + \beta_2 RENDMARK + \varepsilon$$

   modell paramétereit és számítsuk ki a standard hibákat.
   Használjuk fel, hogy
   $$\begin{pmatrix} \sum_{i=1}^n y_i^2 & \sum_{i=1}^n y_i & \sum_{i=1}^n y_i x_{2i} \\ & n & \sum_{i=1}^n x_{2i} \\ & & \sum_{i=1}^n x_{2i}^2 \end{pmatrix} = \begin{pmatrix} 6883.421 & 228.332 & 5220.464 \\ 228.332 & 240.000 & 194.132 \\ 5220.464 & 194.132 & 5562.901 \end{pmatrix},$$
   ahol $y = RENDNCCO$ és $x_2 = RENDMARK$.

   (b) Becsüljük a RENDNCCO rugalmasságát a RENDMARKhoz viszonyítva.

(c) Számítsuk ki a White standard hibákat, és teszteljük a becslések szignifikanciáját. Kritikus értéknek vegyünk $c_{0.05} = 2$-t. Használjuk fel, hogy
$X'\widehat{\Omega}X = \begin{pmatrix} 1975.184 & 4614.187 \\ 4614.187 & 58367.699 \end{pmatrix}$, ahol $\widehat{\Omega}$ az (1.) pontban kiszámított reziduumok négyzetének átlómátrixa.

(d) A White heteroszkedaszticitás-teszt statisztikája 3.9 és ennek p-értéke 0.022. Magyarázzuk el a White heteroszkedaszticitás-tesztet és teszteljük a hibaváltozók homoszkedaszticitását.

2. Az angol font és a német márka közötti árfolyamra vegyük az 1975:1 és 1983:4 közötti hónaponkénti átlagokat, melyek logaritmusát jelöljük y-nal, valamint az angol és német fogyasztói árindexeket, amelyek logaritmusát jelöljük x_2-vel és x_3-mal. Ezekről az adatokról a következőket tudjuk:

$$\begin{pmatrix} \sum_{i=1}^n y_i^2 & \sum_{i=1}^n y_i & \sum_{i=1}^n y_i x_{2i} & \sum_{i=1}^n y_i x_{3i} \\ & n & \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{3i} \\ & & \sum_{i=1}^n x_{2i}^2 & \sum_{i=1}^n x_{2i}x_{3i} \\ & & & \sum_{i=1}^n x_{3i}^2 \end{pmatrix}$$
$$= \begin{pmatrix} 530.35 & 21.16 & 131.42 & 165.77 \\ & 100.00 & -5.87 & -13.61 \\ & & 80.57 & 31.55 \\ & & & 100.69 \end{pmatrix}.$$

A 3. fejezet gyakorlatában becsültük OLS-sel az

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i, \quad i = 1, ..., 100$$

modell paramétereit, és egyebek mellett kiszámítottuk a reziduumok négyzetösszegét: $\sum_{i=1}^{100} e_i^2 = 147.13$.

(a) Vegyük annak a modellnek az OLS becslését, amelyben a fenti becsült modell reziduumai a függővváltozók, míg a független változók x_i és a reziduumok első két késleltetett értéke, vagyis $e_i = x_i\beta + \gamma_1 e_{i-1} + \gamma_2 e_{i-2} + \eta_i,\ i = 3, ..., n$. Felhasználva, hogy e modell OLS becslésénél a reziduumok négyzetösszege 101.82, míg e modell $H_0 : \beta = 0$ nullhipotézissel való leszűkítésénél a reziduumok négyzetösszege 147.13, teszteljük az eredeti modell hibaváltozóinak autokorrelációját. Magyarázzuk el a tesztet; használjuk a 3.1 kritikus értéket.

(b) Ha tudjuk, hogy $\sum_{i=2}^{n} e_i e_{i-1} = 50.02$ és $\sum_{i=3}^{n} e_i e_{i-2} = 36.78$, teszteljük az első két autokorrelációs együttható egyéni szignifikanciáját a nem formális teszt segítségével. Magyarázzuk el a tesztet.

(c) Felhasználva, hogy $\sum_{i=2}^{n} e_i e_{i-1} = 50.02$ és $\sum_{i=3}^{n} e_i e_{i-2} = 36.78$, teszteljük az első két autokorrelációs együttható egyidejű szignifikanciáját a Box-Ljung teszt segítségével. Magyarázzuk el a tesztet, használjuk az 5.9 kritikus értéket.

5. fejezet

Kvalitatív függővváltozójú modellek

Ebben a fejezetben azt vizsgáljuk, hogyan lehet bizonyos változók hatását mérni egy kvalitatív függővváltozóra. Az ökonometriai gyakorlatban gyakran szembesülünk olyan problémával, ahol erre van szükség. Például, választási vagy vásárlási helyzetek modellezésekor (egyetem, okostelefon, stb.), ahol a meglévő alternatívák száma véges; különböző egyének viselkedésének a modellezésekor (diákok elvégzik az egyetemet vagy nem, fogyasztók válaszolnak termékek hirdetésére vagy nem); vagy kategorizált változók modellezésekor (fizetés-intervallumok).

Néhány esetben (mint az utóbbi) azért van szükség kvalitatív modellezésre, mert nem elég részletesek a rendelkezésünkre álló adatok, de a legtöbb esetben a vizsgált jelenség kvalitatív jellegű, ezért nehéz lenne kvantitatív változójú megfigyelésekre szert tenni.

Mivel vvkkal tudunk csak dolgozni, ezért a kvalitatív változókat diszkrét vvkként használjuk. Elsősorban olyan modelleket tanulmányozunk amelyeknek a függővváltozói diszkrét vvk.

Egy diszkrét vv egymástól jól elkülönülő értékeket vesz fel (a folytonos vvtól eltérően, amely egy adott intervallum bármely elemét felveheti). Például, $Y = 1$ vagy $Y = 0$, ha fejet vagy írást dobunk egy pénzérmével. Ez egy Bernoulli vv, amelyre $p = P(Y = 1)$ az úgynevezett sikervalószínűség. Tudjuk, hogy

$$E[Y] = 1 \cdot p + 0 \cdot (1 - p) = p,$$

vagyis egy Bernoulli vv várható értéke a sikervalószínűséggel egyenlő; varianciája

$$var\left(Y\right)=E\left[Y^2\right]-E\left[Y\right]^2=1^2\cdot p+0^2\cdot\left(1-p\right)-p^2=p\left(1-p\right).$$

Egy másik diszkrét vv, jelöljük T-vel, a $0,1,...,J$ értékeket veszi fel, és a j valószínűsége $P\left(T=j\right)=p_j$, ahol $p_0+p_1+...+p_J=1$. A fenti Y bináris változó, mert két értéket vehet fel, míg a T multinomiális változó, mert több mint két értéket vehet fel.

A tanulmányozott modellekben bizonyos független változók hatását vizsgáljuk bináris vagy multinomiális függőváltozókra. Ezekben a modellekben a független változók a függőváltozó előfordulási valószínűségét határozzák meg. Ez alapján meg fogunk különböztetni logit és probit modelleket.

Tegyük fel, hogy a többváltozós lineáris modellhez hasonlóan, $k-1$ független változó hatását vizsgáljuk a függőváltozóra, amelyekre n megfigyelésünk van, és az első független változó mindegyik megfigyelésre 1. A független változók hatását mérő paraméter-vektor legyen β. Jelöljük x_i-vel a független változók i-edik megfigyelésének k-elemű sorvektorát; ekkor ennek a hatását $x_i\beta$-val mérjük.

5.1. Bináris függőváltozójú modellek

A bináris függőváltozójú modelleket különböző egyének viselkedésének a modellezésére használják (pl. a diákok elvégzik az egyetemet vagy nem, a fogyasztók válaszolnak termékek hirdetésére vagy nem). Ezért az i-edik egyénnek megfelelő függőváltozó y_i 0 vagy 1 lehet, és a független változók az egyének tulajdonságaira vonatkoznak. A modellezés célja felmérni a különböző tulajdonságok hatását a $P\left(y_i=1\right)$ valószínűségre.

A már említett logit és probit modelleken kívül itt a lineáris valószínűségű modellt is tanulmányozzuk.

5.1.1. Lineáris valószínűségű modell

Ritkán használják, alább meglátjuk, miért. Mi a lineáris modellel való összehasonlítás céljából tanulmányozzuk.Tulajdonképpen egy lineáris modell,

$$y_i=x_i\beta+\varepsilon_i=\beta_1+\beta_2x_{2i}+...+\beta_kx_{ki}+\varepsilon_i,$$

amelynek a függőváltozója csak a 0 és 1 értékeket veszi fel. A kiindulópont a

$$P\left(y_i=1\right)=E\left[y_i\right]=x_i\beta$$

lineáris valószínűségi összefüggés, amelyből kapjuk, hogy $E[\varepsilon_i] = 0$, vagyis a hibaváltozók nulla várható értékűek, mint a lineáris modell A2 feltétele szerint (lásd 3.2.1. alfejezet).

Ebben a modellben a paraméterek **értelmezése** a derivált alapján:

$$\frac{\partial P(y_i = 1)}{\partial x_{ji}} = \beta_j,$$

vagyis az x_{ji} egységnyi növekedésével az $y_i = 1$ valószínűsége β_j-vel nő.

A lineáris valószínűségű modell OLS becslőfüggvénye torzítatlan és konszisztens (helyes). Viszont

$$\varepsilon_i = 1 - x_i\beta, \text{ ha } y_i = 1, \text{ és}$$
$$\varepsilon_i = -x_i\beta, \text{ ha } y_i = 0,$$

ezért $var(\varepsilon_i) = E[\varepsilon_i^2] = (1 - x_i\beta)^2 x_i\beta + (-x_i\beta)^2 (1 - x_i\beta) = x_i\beta(1 - x_i\beta)$. Vagyis a hibaváltozók varianciái nem egyenlőek (heteroszkedaszticitás).

Emiatt az egyéni szignifikancia tesztek nem végezhetők el az OLS becslések alapján, és az OLS becslőfüggvény nem rendelkezik a legjobb lineáris torzítatlan becslőfüggvény tulajdonsággal. A 4. fejezetben tanulmányoztuk, hogy mi a teendő ebben az esetben.

Van még egy hiányossága ennek a modellnek, éspedig, hogy előfordulhat, hogy értelmetlen előrejelzést kapunk a független változók bizonyos értékeire. Jelöljük b-vel az OLS becslőfüggvényt; ekkor előfordulhat, hogy a független változó bizonyos x_* értékére $x_*b < 0$ vagy $x_*b > 1$ értékeket kapunk. Az előbbi azt jelenti, hogy $P(y_* = 1) < 0$, míg az utóbbi, hogy $P(y_* = 1) > 1$, amiknek nincs értelme.

Fontos megjegyezni, hogy a modell szignifikancia-teszt (melyre $H_0 : \beta_2 = ... = \beta_k$) elvégzéséhez alkalmas az OLS-nél használatos F-teszt mivel ez a teszt nullhipotézise alapján $var(\varepsilon_i) = \beta_1(1 - \beta_1)$, vagyis a hibaváltozók homoszkedasztikusak. Tehát a lineáris valószínűségű modell hasznos ebből a szempontból: könnyen tudjuk tesztelni (OLS-sel) a független változók egyidejű szignifikanciáját.

5.1.2. A logit modell

A bináris logit modellben a $P(y_i = 1)$ valószínűséget a

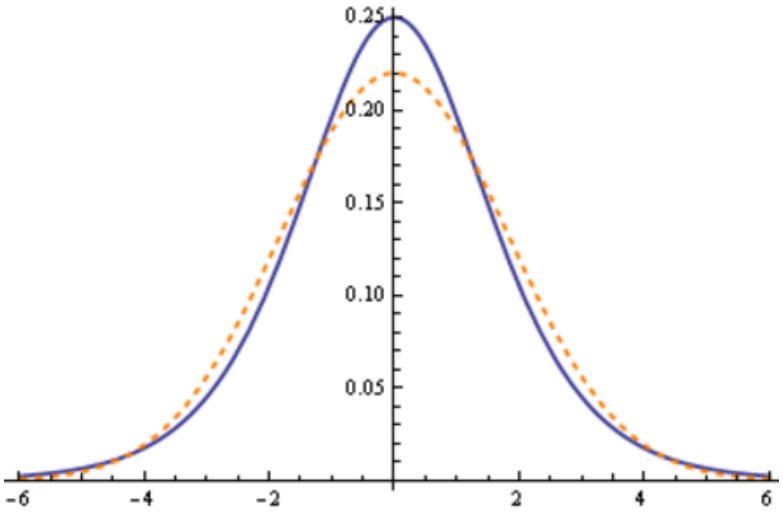
$$P(y_i = 1) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} = \frac{1}{1 + e^{-x_i\beta}}$$

kifejezéssel adjuk meg. A $\Lambda(z) = \dfrac{1}{1+e^{-z}}$ függvény szigorúan növekvő és $-\infty$-ben 0, $+\infty$-ben 1, ezért egy eloszlásfüggvény, tulajdonképpen logisztikus eloszlásfüggvénynek hívják. Megjegyezzük, hogy bármilyen β értékre és az x_i független változó bármilyen értékére $0 < \dfrac{e^{x_i\beta}}{1+e^{x_i\beta}} < 1$, ezért ennél a modellnél nem kaphatunk értelmetlen előrejelzést mint a lineáris valószínűségű modellnél.

Egy eloszlásfüggvény a sűrűségfüggvényhez hasonlóan egyértelműen meghatározza az eloszlást. Ha egy X vvnak F az eloszlásfüggvénye és f a sűrűségfüggvénye, akkor

$$P(X < x) = F(x) = \int_{-\infty}^{x} f(x)\,dx.$$

A $\lambda(z) = \Lambda'(z) = \dfrac{e^z}{(1+e^z)^2}$ az eloszlásfüggvény deriváltja, a logisztikus sűrűségfüggvény, szimmetrikus 0 körül. Ennek grafikonja:



Szaggatott vonal: standard normál; folytonos vonal: logisztikus.

Mivel a Λ függvény szigorúan növekvő, ezért
- $\beta_j > 0$ azt jelenti, hogy nagyobb x_{ji}-re a $P(y_i = 1)$ is nagyobb lesz,
- $\beta_j < 0$ azt jelenti, hogy nagyobb x_{ji}-re a $P(y_i = 1)$ kisebb lesz,
- $\beta_j = 0$ azt jelenti, hogy nagyobb x_{ji}-re a $P(y_i = 1)$ nem változik.

Ez látható az x_{ji} szerinti parciális deriváltból is:

$$\frac{\partial P\left(y_i=1\right)}{\partial x_{ji}}=\frac{e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2}\beta_j,$$

mert $\frac{e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2}$ mindig szigorúan pozitív. Tehát azt, hogy a $P\left(y_i=1\right)$ valószínűség nő vagy csökken az x_{ji} független változó változása nyomán, a β_j előjele határozza meg. A változás nagyságát viszont a β_j mellett az $\frac{e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2}$ kifejezés nagysága is befolyásolja, éspedig, ha $x_i\beta$ közel van 0-hoz, akkor az $\frac{e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2}$ kifejezés közel 0.25, másképp jóval kisebb.

A független változók hatásának mérése szempontjából egy fontos fogalom az esélyhányados:

$$\frac{P\left(y_i=1\right)}{P\left(y_i=0\right)}=e^{x_i\beta}.$$

A bináris logit modellben ennek a logaritmusa pontosan $x_i\beta$.

A bináris logit modell értelmezése

A modellnek két alkalmazási szempontból hasznos értelmezése ismert.

1. Egy lineáris modell függővváltozóinak csak az előjelét ismerjük.

A modellt értelmezhetjük egy

$$y_i^*=x_i\beta+\varepsilon_i$$

lineáris modellként, amelynek a függővváltozójáról nincsenek megfigyeléseink, viszont azt tudjuk, hogy a megfigyelések negatívak vagy pozitívak, vagyis vannak megfigyeléseink az y_i változókról úgy, hogy $y_i=1$, ha $y_i^*\geq 0$, és $y_i=0$, ha $y_i^*<0$.

Tegyük fel, hogy mindegyik ε_i logisztikus eloszlású. Ekkor

$$\begin{aligned}P\left(y_i=1\right)&=P\left(y_i^*\geq 0\right)=P\left(x_i\beta+\varepsilon_i\geq 0\right)\\&=P\left(\varepsilon_i\geq -x_i\beta\right)\overset{\lambda\ \text{szimmetrikus}}{=}P\left(\varepsilon_i\leq x_i\beta\right)=\Lambda\left(x_i\beta\right)=\frac{e^{x_i\beta}}{1+e^{x_i\beta}},\end{aligned}$$

ami pontosan a logit modell.

2. Hasznosságfüggvény segítségével.

A hasznosságfüggvénnyel való értelmezés közgazdaságtani szempontból fontos, mivel összhangban van azzal az elvvel, hogy a fogyasztók úgy döntenek, hogy maximalizáják a hasznosságukat, ami a mikroökonómia egyik alapelve.

Tegyük fel, hogy minden fogyasztó két lehetőség közül választhat: 1 és 0; jelöljük ezek hasznosságát az i fogyasztónak u_{i1}-gyel és u_{i0} -val. Tegyük fel, hogy

$$u_{i1} = x_i\beta + \varepsilon_{i1},$$
$$u_{i0} = \varepsilon_{i0}.$$

ahol ε_{i1} és ε_{i0} vvk eloszlásfüggvénye $e^{-e^{-z}}$ (ezt első típusú szelsőérték eloszlásnak hívják).

Ekkor a hasznosság-maximalizálás elve alapján az i fogyasztó akkor választja az 1-es terméket ($y_i = 1$), ha nagyobb hasznosságot nyújt, vagyis $u_{i1} \geq u_{i0}$; ennek valószínűsége:

$$P(y_i = 1) = P(u_{i1} \geq u_{i0}) = P(x_i\beta + \varepsilon_{i1} \geq \varepsilon_{i0}).$$

Felhasználva, hogy ε_{i1}, ε_{i0} eloszlásfüggvénye $e^{-e^{-z}}$, igazolható, hogy

$$P(x_i\beta + \varepsilon_{i1} \geq \varepsilon_{i0}) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}},$$

vagyis az 1-es termék választásának valószínűsége $\dfrac{e^{x_i\beta}}{1 + e^{x_i\beta}}$.

5.1.3. A probit modell

A logithoz hasonlóan adjuk meg a $P(y_i = 1)$ valószínűséget:

$$P(y_i = 1) = \Phi(x_i\beta),$$

tehát a különbség az, hogy a Λ logisztikus eloszlásfüggvény helyett a Φ standard normál eloszlásfüggvényt használjuk. A független változók hatását a $P(y_i = 1)$ valószínűségre hasonlóan mérjük mint a logitnál:

$$\frac{\partial P(y_i = 1)}{\partial x_{ji}} = \phi(x_i\beta)\beta_j,$$

ahol ϕ a standard normál sűrűségfüggvény.

Bináris függővváltozók esetén a logit modell használatosabb mint a probit, mert a logit valószínűség kifejezése egyszerűbb, és emiatt könnyebben értelmezhető (lásd például az esélyhányados logaritmusát, ami lineáris). A valódi modell

viszont állhat közelebb a probithoz mint a logithoz, ezért érdemes lehet a probitot is tanulmányozni egy becslés során.

A logit és a probit összehasonlítása során szem előtt kell tartani azt, hogy a logisztikus eloszlás szórása $\pi/\sqrt{3} \simeq 1.8$, míg a standard normál eloszlás szórása 1. Ezért a két modell becsléseinek összehasonlításakor a probit becsléseket 1.6−1.8-del szokták szorozni ($\phi(0)/\lambda(0) = 4/\sqrt{2\pi} \simeq 1.6$).

5.1.4. Bináris függővváltozójú modellek becslése

A becslés az úgynevezett maximum likelihood módszerrel történik, amely úgy határozza meg az ismeretlen β paraméter-vektort, hogy maximizálja a megfigyelt függővváltozók valószínűségét.

Legyenek a függővváltozó megfigyelései $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$. Tegyük fel, hogy $\bar{y}_i = 1$; tudjuk, hogy ennek a valószínűsége $P(y_i = 1) = p_i = F(x_i\beta)$, ahol F vagy Λ (logit) vagy Φ (probit). Ha $\bar{y}_i = 0$, tudjuk, hogy ennek a valószínűsége $P(y_i = 0) = 1 - p_i$.

Tehát annak a valószínűsége, hogy $y_i = \bar{y}_i$, ahol $\bar{y}_i$ a 0 vagy 1 értékeket veheti fel,

$$P(y_i = \bar{y}_i) = p_i^{\bar{y}_i}(1 - p_i)^{1-\bar{y}_i}.$$

Fontos szem előtt tartani, hogy y_i az i-edik függővváltozó, egy vv, míg $\bar{y}_i$ egy szám, az y_i lehetséges értéke.

Annak a valószínűsége, hogy a függővváltozó megfigyelései $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$ az úgynevezett likelihood függvény:

$$L(\bar{y}_1, ..., \bar{y}_n; \beta) = P(y_1 = \bar{y}_1, ..., y_n = \bar{y}_n).$$

Feltételezve, hogy a függővváltozók függetlenek egymástól, a likelihood függvény

$$\begin{aligned} L(\bar{y}_1, ..., \bar{y}_n; \beta) &= P(y_1 = \bar{y}_1) \cdot ... \cdot P(y_n = \bar{y}_n) \\ &= p_1^{\bar{y}_1}(1 - p_1)^{1-\bar{y}_1} \cdot ... \cdot p_n^{\bar{y}_n}(1 - p_n)^{1-\bar{y}_n}. \end{aligned}$$

Ennek a logaritmusa

$$\begin{aligned} \ln L(\bar{y}_1, ..., \bar{y}_n; \beta) &= \sum_{i=1}^{n} [\bar{y}_i \ln p_i + (1 - \bar{y}_i) \ln(1 - p_i)] \\ &= \sum_{i=1}^{n} [\bar{y}_i \ln F(x_i\beta) + (1 - \bar{y}_i) \ln(1 - F(x_i\beta))] \end{aligned}$$

a log-likelihood függvény. Ezzel egyszerűbb számolni, ezért ezt szokták maximalizálni a β függvényében.

A kapott becslőfüggvényt maximum likelihood (ML) becslőfüggvénynek nevezzük, és b-vel jelöljük. Fontos megjegyezni, hogy a log-likelihood függvény nem lineáris a paraméterek függvényében sem a logit sem a probit esetén. Ugyanez érvényes a likelihood függvényre is.

Az ML becslőfüggvény torzított (nem torzítatlan), vagyis $E[b] \neq \beta$. Ez a tulajdonság általában jellemző a nem lineáris modellek becslőfüggvényeire. Ha a független változók determinisztikusak, b konszisztens (helyes), a lehető legkisebb a varianciája (a legpontosabb), és megközelítőleg normál eloszlású, ahol a megközelítés annál pontosabb minél nagyobb a megfigyelések száma.

A logit esetén

$$b \approx N\left(\beta, var\left(b\right)\right),$$

ahol

$$var\left(b\right) = \left(\sum_{i=1}^{n} p_i\left(1-p_i\right) x_i x_i'\right)^{-1} = \left(\sum_{i=1}^{n} \frac{e^{x_i\beta}}{\left(1+e^{x_i\beta}\right)^2} x_i x_i'\right)^{-1}.$$

A fenti variancia kiszámításához a β-t b-vel helyesettesítjük:

$$var\left(b\right) \simeq \left(\sum_{i=1}^{n} \frac{e^{x_i b}}{\left(1+e^{x_i b}\right)^2} x_i x_i'\right)^{-1}.$$

Mivel $p_i\left(1-p_i\right) \leq \frac{1}{4}$ ezért $var\left(b\right) \geq 4\left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} = 4\left(X'X\right)^{-1}$.

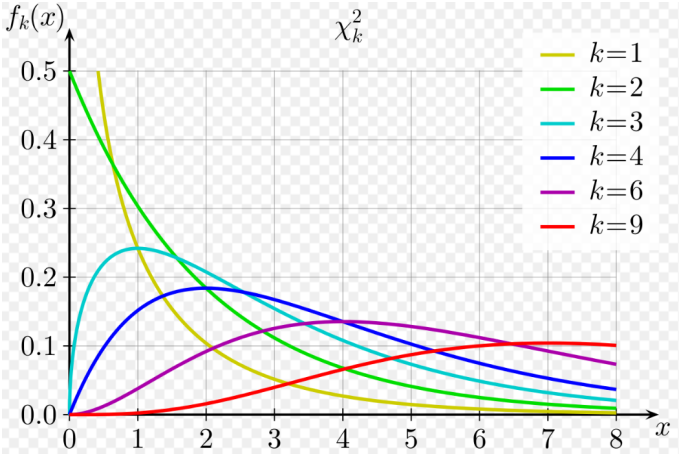
Emlékezzünk vissza, hogy a lineáris modell OLS becslőfüggvényének a varianciája $\sigma^2\left(X'X\right)^{-1}$, tehát a bináris logit ML becslőfüggvényének varianciája nagyobb az OLS becslőfüggvény varianciájánál, ha $\sigma^2 < 4$.

A becslések szignifikanciája

A becslések egyéni szignifikanciáját a lineáris modellhez hasonlóan végezzük. A j-edik paraméter szignifikanciájához $H_0 : \beta_j = 0$, $H_1 : \beta_j \neq 0$. A használt statisztika a $t_{b_j} = \frac{b_j}{s_{b_j}}$, ami a nullhipotézis mellett megközelítőleg standard normál eloszlású. Ezért használhatjuk a 2 kritikus értéket 0.05 szignifikancia szint esetén.

Az SPSS a Wald statisztikát és p-értékét adja meg (ezért a kritikus értékekre nincs szükség). A Wald statisztika $\left(\frac{b_j}{s_{b_j}}\right)^2 = t_{b_j}^2$ és eloszlása $\chi^2\left(1\right)$. A Wald teszt a t-teszttel megegyező következtetéshez vezet a nullhipotézist

illetően. A χ^2 egy pozitív értékeket felvevő eloszlás, amelynek egyetlen paramétere van, amit szabadsági foknak neveznek. Az alábbi ábra a χ^2 sűrűségfüggvényének a grafikonja különböző k szabadsági fokokra:



A becslések egyidejű szignifikanciáját a likelihood ratio (LR) teszttel végezhetjük. Legyen $H_0 : \beta_2 = 0, ..., \beta_j = 0$ és $H_1 : \beta_2 \neq 0$ vagy ... vagy $\beta_j \neq 0$. A LR teszt statisztika $2\left(\ln L_1 - \ln L_0\right) \approx \chi^2\left(j-1\right)$, ahol $\ln L_1$ és $\ln L_0$ az eredeti (leszűkítetlen) és a leszűkített modellek log-likelihoodok maximumai. (Adott becslésre az SPSS megadja a $-2\ln L$ értékét.) Az elfogadási tartomány $(0, c_{0.05})$, ahol $c_{0.05}$ a $\chi^2\left(j-1\right)$ által meghatározott kritikus érték (táblázatból nézhetjük ki). Ha $2\left(\ln L_1 - \ln L_0\right) > c_{0.05}$, elutasítjuk a nullhipotézist, másképp nem.

Ha H_0 az összes meredekségi együtthatót tartalmazza, akkor ezt a tesztet a modell szignifikancia tesztjének nevezzük. A LR teszt alkalmas egyéni szignifikancia tesztelésére is.

A becsült összefüggés szorossága

Felhasználva a maximum likelihood tulajdonságait, az OLS becsléshez hasonlóan szerkeszthetünk egy determinancia együtthatót. Legyen $\ln L_1$ a becsült modell log-likelihoodja, míg $\ln L_0$ annak a becsült modellnek a log-likelihoodja, amely nem tartalmaz egyetlen független változót sem. Ekkor $\ln L_0 \leq \ln L_1 < 0$, ezért $0 \leq 1 - \dfrac{\ln L_1}{\ln L_0} < 1$. Ez alapján értelmezhetjük a következő determinancia-együtthatót:

$$R^2_M = 1 - \frac{\ln L_1}{\ln L_0},$$

amely az OLS R^2-éhez hasonlóan mutatja a becsült modell szorosságát. Ezt McFadden-féle R^2-nek nevezik.

Egy másik determinancia-együttható, amely szorosabban kapcsolódik az LR tesztstatisztikához,

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_1}\right)^{2/n} = 1 - e^{-LR/n},$$

ahol $LR = 2\left(\ln L_1 - \ln L_0\right)$, és nagy n értékekre nem függ n-től. Ezt a Cox-Snell-féle R^2-nek nevezik (az SPSS megadja a statisztikák között).

A determinancia-együtthatókat nem használhatjuk arra, hogy eldöntsük, a logit vagy a probit modell megfelelőbb. Ez megvalósítható viszont az előrejelzési teljesítmény mérésével.

Az előrejelzési teljesítmény

A bináris függőváltozójú modelleknél az előrejelzési teljesítményt az osztályozási táblázattal mérhetjük. Ezt az SPSS is megadja a statisztikák között. A modell paramétereinek becslése alapján becsülhetjük a $p_i = P\left(y_i = 1\right)$ valószínűségeket a $\hat{p}_i = F\left(x_i b\right)$-vel mindegyik megfigyelésre. Ekkor az y_i-t $\hat{y}_i = 1$-nek jelezzük előre, ha $\hat{p}_i \geq 0.5$, és $\hat{y}_i = 0$-nak, ha $\hat{p}_i < 0.5$.

Ez alapján az előrejelzési teljesítményt a helyes előrejelzések arányával mérjük. Az a modell megfelelőbb, amelyiknek jobb az előrejelzési teljesítménye.

5.2. Multinomiális függőváltozójú modellek

A multinomiális függőváltozójú modelleket különböző egyének viselkedésének a modellezésére használják, amikor több mint két választás lehetséges (pl. több mint két munkahelytípus választásának modellezése, választás több mint két termék közül).

Jelöljük a választási lehetőségeket a $1, 2, ..., J$ számokkal. Ekkor az i-edik egyénnek megfelelő y_i függőváltozó $1, 2, ..., J$ lehet, és a független változók az egyének vagy a választási lehetőségek tulajdonságaira vonatkoznak. A modellezés célja felmérni a különböző tulajdonságok hatását a $P\left(y_i = j\right)$ valószínűségekre, ahol $j = 1, 2, ..., J$.

Részletesen a multinomiális logit modellt tárgyaljuk; ezen kívül még megemlítjük a feltételes logit és a multinomiális probit modelleket.

Multinomiális logit

Ebben a modellben a független változók kizárólag az egyének tulajdonságaira vonatkoznak. A bináris logithoz hasonlóan értelmezhető a hasznosságfüggvénnyel. Tegyük fel, hogy minden fogyasztó J lehetőség közül választhat, és az i fogyasztónak a j kategória hasznossága

$$u_{ij} = x_i\beta_j + \varepsilon_{ij}, \quad j = 1, ..., J,$$

ahol x_i független változók vektora (első eleme 1) és β_j paraméterek vektora; mindegyik ε_{ij} eloszlásfüggvénye $e^{-e^{-z}}$.

Ekkor a hasznosság-maximalizálás elve alapján

$$P\left(y_i = j\right) = P\left(u_{ij} \geq u_{ih},\ h = 1, ..., J\right).$$

Felhasználva, hogy ε_{ij} eloszlásfüggvénye $e^{-e^{-z}}$, igazolható, hogy

$$P\left(y_i = j\right) = \frac{e^{x_i\beta_j}}{\sum_{h=1}^{J} e^{x_i\beta_h}}.$$

Vegyük észre, hogy

$$P\left(y_i = 1\right) = \frac{e^{x_i\beta_1}}{\sum_{h=1}^{J} e^{x_i\beta_h}} = \frac{1}{1 + \sum_{h=2}^{J} e^{x_i(\beta_h - \beta_1)}},$$
$$P\left(y_i = j\right) = \frac{e^{x_i(\beta_j - \beta_1)}}{1 + \sum_{h=2}^{J} e^{x_i(\beta_h - \beta_1)}},$$

vagyis mind a J választási valószínűség kiszámítható csak a $\gamma_2 = \beta_2 - \beta_1, ..., \gamma_J = \beta_J - \beta_1$ paraméterekkel.

Ez azt jelenti, hogy a $\beta_1, \beta_2, ..., \beta_J$ ismeretlen paraméterek nem határozhatók meg egyértelműen, mert több értékükre kaphatjuk ugyanazokat a $\gamma_2, ..., \gamma_J$ paramétereket. Ez úgy is értelmezhető, hogy csak $J-1$ kategória paraméterei határozhatók meg az egyik (pl. $j = 1$ vagy $j = J$) kategóriához viszonyítva. Ezért, kiválasztunk egy referencia kategoriát (reference category az SPSS-ben), amelynek paramétereit 0-nak vesszük, és ehhez viszonyítjuk a többi paramétert.

Legyen például $j = 1$ a referencia kategória. Ekkor $\beta_1 = 0$, tehát

$$P\left(y_i = 1\right) = \frac{1}{1 + \sum_{h=2}^{J} e^{x_i\beta_h}},$$
$$P\left(y_i = j\right) = \frac{e^{x_i\beta_j}}{1 + \sum_{h=2}^{J} e^{x_i\beta_h}},$$

ahol $x_i\beta_j = \beta_{j1} + \beta_{j2}x_{i2} + ... + \beta_{jk}x_{ik}$.

A bináris logithoz hasonlóan itt is értelmezzük az esélyhányadost a j és 1 valamint a j és h kategóriákra:

$$\frac{P(y_i = j)}{P(y_i = 1)} = e^{x_i\beta_j} \quad \text{és} \quad \frac{P(y_i = j)}{P(y_i = h)} = e^{x_i(\beta_j - \beta_h)}, \quad \text{ha } j, h \neq 1.$$

Ezeknek a logaritmusai $x_i\beta_j$ és $x_i(\beta_j - \beta_h)$.

5.2.1. Feltételes logit

Tegyük fel, hogy minden fogyasztó J lehetőség közül választhat. Az i fogyasztónak a j kategória hasznossága

$$u_{ij} = x_{ij}\beta + \varepsilon_{ij}, \quad j = 1, ..., J,$$

vagyis a független változók vonatkozhatnak mind az egyénekre, mind a kategóriákra (ahol mindegyik ε_{ij} vv eloszlásfüggvénye $e^{-e^{-z}}$).

A hasznosság-maximizálás elv alapján a

$$P(y_i = j) = \frac{e^{x_{ij}\beta}}{\sum_{h=1}^{J} e^{x_{ih}\beta}}$$

valószínűségeket kapjuk. Ezt a modellt gyakran használják fogyasztói preferenciák és piaci kereslet modellezésére.

5.2.2. Multinomiális probit

A multinomiális logithoz hasonlóan értelmezhető. Az i fogyasztónak a j kategória hasznossága

$$u_{ij} = x_i\beta_j + \varepsilon_{ij}, \quad j = 1, ..., J,$$

ahol mindegyik ε_{ij} vv normál eloszlásúak. A $P(y_i = j)$ valószínűségek nem fejezhetők ki egyszerű képletekkel.

Viszont a multinomiális probit szofisztikáltabb választási folyamat modellezésére alkalmas azáltal, hogy a hibaváltozók korrelálhatnak egymással. Vagyis, tegyük fel, hogy

$$(\varepsilon_{i1}, ..., \varepsilon_{iJ}) \sim N(0, \Sigma),$$

ahol a

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & & \sigma_{1J} \\ \sigma_{12} & \sigma_2^2 & & \sigma_{2J} \\ & & \ddots & \\ \sigma_{1J} & \sigma_{2J} & & \sigma_J^2 \end{pmatrix}$$

variancia mátrix főátlón kívüli elemei lehetnek nemnullák. Ezáltal modellezhető bizonyos kategóriák hasznosságai közötti korreláció.

5.2.3. A multinomiális logit becslése

A becslés a maximum likelihood módszerrel történik akárcsak a bináris logit esetében, vagyis úgy határozzuk meg az ismeretlen $\beta_2, ..., \beta_J$ paraméter-vektorokat, hogy maximizáljuk a log-likelihood függvényt. Használjuk a $p_{ij} = P(y_i = j)$ jelölést és legyenek a függőváltozó megfigyelései $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$, amelyek mindegyike az $1, 2, ..., J$ értékeket veheti fel.

Legyen

$$\bar{y}_{ij} = \begin{cases} 1 & \text{ha } \bar{y}_i = j \\ 0 & \text{másképp.} \end{cases}$$

Ekkor a log-likelihood

$$\ln L(\bar{y}_1, ..., \bar{y}_n; \beta_2, ..., \beta_J) = \sum_{i=1}^{n} \sum_{j=1}^{J} \bar{y}_{ij} \ln p_{ij}.$$

A kapott maximum likelihood becslőfüggvényeket $b_2, ..., b_J$-vel jelöljük. Ezek a becslőfüggvények a bináris logithoz hasonló tulajdonságokkal rendelkeznek, ezért a szignifikancia-teszteket hasonlóan végezzük.

A becslések szignifikanciája

A becslések egyéni szignifikanciáját a lineáris modellhez hasonlóan végezzük. A j-edik paraméter szignifikanciájához $H_0 : \beta_{j\ell} = 0$, $H_1 : \beta_{j\ell} \neq 0$. A használt statisztika a $t_{b_{j\ell}} = \frac{b_{j\ell}}{s_{b_{j\ell}}}$, ami a nullhipotézisre megközelítőleg standard normál eloszlású. Ezért használhatjuk a 2 kritikus értéket 0.05 szignifikancia szint esetén.

Az SPSS a Wald statisztikát és p-értékét adja meg (ezért a kritikus értékekre nincs szükség). A Wald statisztika $\left(\frac{b_{j\ell}}{s_{b_{j\ell}}}\right)^2 = t_{b_{j\ell}}^2$ és eloszlása $\chi^2(1)$.

A Wald teszt a t-teszttel megegyező következtetéshez vezet a nullhipotézist illetően.

A becslések egyidejű szignifikanciáját a likelihood ratio (LR) teszttel végezhetjük. A nullhipotézis az, hogy az összes meredekségi együttható 0, míg az alternetív hipotézis az, hogy legalább egyikük nem 0. A LR teszt statisztika $2\left(\ln L_1 - \ln L_0\right) \approx \chi^2\left(m\right)$, ahol $\ln L_1$ és $\ln L_0$ az eredeti (leszűkítetlen) és a leszűkített modellek log-likelihoodainak maximumai, m a nullhipotézisben szereplő paraméterek száma. Az elfogadási tartomány $(0, c_{0.05})$, ahol $c_{0.05}$ a $\chi^2\left(m\right)$ által meghatározott kritikus érték (táblázatból nézhetjük ki). Ha $2\left(\ln L_1 - \ln L_0\right) > c_{0.05}$, elutasítjuk a nullhipotézist, másképp nem.

Ha H_0 az összes meredekségi együtthatót tartalmazza, akkor ezt a tesztet a modell szignifikancia tesztjének nevezzük. Az SPSS megadja a modell szignifikancia teszt statisztikáját és p-értékét.

Egy változó szignifikancia-tesztje egy bizonyos független változó egyidejű szignifikanciáját teszteli az összes választási valószínűségre. Nullhipotézise az, hogy az ezeknek megfelelő paraméterek egyidejűleg nullák (az alternatív hipotézis a szokásos). Ezt a tesztet is el lehet végezni a LR teszt segítségével a fent leírtak alapján. Az statisztikáját és a p-értékét megkaphatjuk SPSS-ben.

A becsült összefüggés szorossága

Mivel a bináris logithoz hasonlóan itt is kiszámítható $\ln L_0$, vagyis annak a becsült modellnek a log-likelihoodja, amely nem tartalmaz egyetlen független változót sem, könnyen kiszámítható a McFadden és a Snell-Cox féle determinancia-együttható.

Ha $\ln L_1$ a becsült modell log-likelihoodja, fennáll

$$0 \leq 1 - \frac{\ln L_1}{\ln L_0} < 1 \quad \text{és} \quad 0 \leq 1 - \left(\frac{L_0}{L_1}\right)^{2/n} < 1.$$

Ez alapján értelmezhetjük a McFadden R^2-t:

$$R^2_M = 1 - \frac{\ln L_1}{\ln L_0},$$

és a Cox-Snell R^2-t:

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_1}\right)^{2/n} = 1 - e^{-LR/n},$$

ahol $LR = 2\left(\ln L_1 - \ln L_0\right)$. Mindkettőt megkaphatjuk SPSS-ben.

Az előrejelzési teljesítmény

A multinomiális függőváltozójú modelleknél az előrejelzési teljesítményt az osztályozási táblázattal mérhetjük. Ezt az SPSS is megadja a statisztikák között. A modell paramétereinek becslése alapján becsülhetjük a $p_{ij} = P\left(y_i = j\right)$ valószínűségeket mindegyik megfigyelésre.

Ekkor az y_i-t $\hat{y}_i = j$-nek jelezzük előre, ha a becsült $p_{i1}, p_{i2}, ..., p_{iJ}$ valószínűségek közül p_{ij} a legnagyobb. Ez alapján az előrejelzési teljesítményt a helyes előrejelzések arányával mérjük (SPSS-ben ennek neve Classification Table). Az a modell megfelelőbb, amelyiknek jobb az előrejelzési teljesítménye.

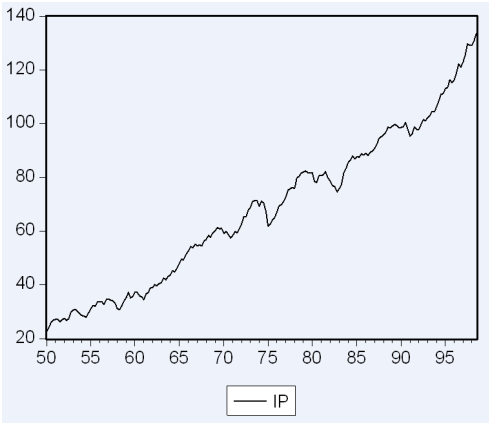
6. fejezet

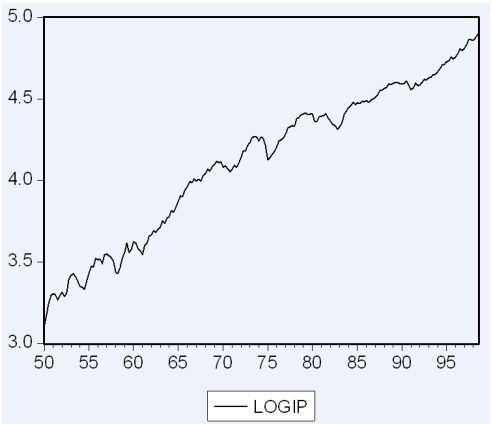
Idősorelemzés

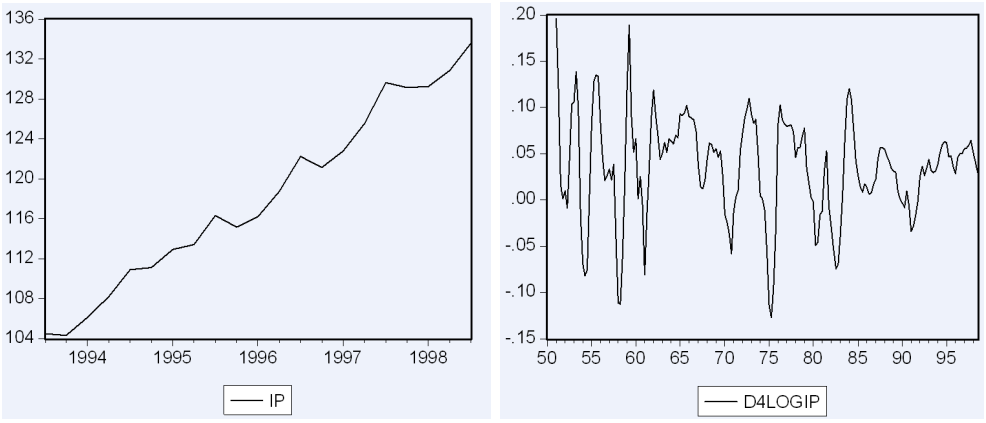
Egy idősor vagy folyamat bizonyos gyakoriságú megfigyelésekből áll. Ezért ebben a fejezetben az i helyett a t indexet fogjuk használni, tehát a megfigyeléseket általában y_t-vel fogjuk jelölni. Gyakoriságuk szerint a megfigyelések lehetnek naponkénti, hónaponkénti, évenkénti, (stb.) megfigyelések.

Bevezetésül, ábrák segítségével tanulmányozzuk a gazdasági idősorok fontosabb jellemvonásait. Két példát tanulmányozunk: ipari termelés és Dow-Jones index.

Ipari termelés. Az alábbi négy vonalgrafikon mutatja (1) a tulajdonképpeni megfigyeléseket (IP), amelyek negyedévenkénti megfigyelések az USA-beli ipari termelés indexére, (2) a logaritmusukat, (LOGIP), (3) az eredeti megfigyelések egy leszűkített (1993:03-1998:03) periódusban (IP) és a megfigyelések logaritmusának negyedik különbségét (D4LOGIP): $\Delta_4 \ln y_t = \ln y_t - \ln y_{t-4}$.



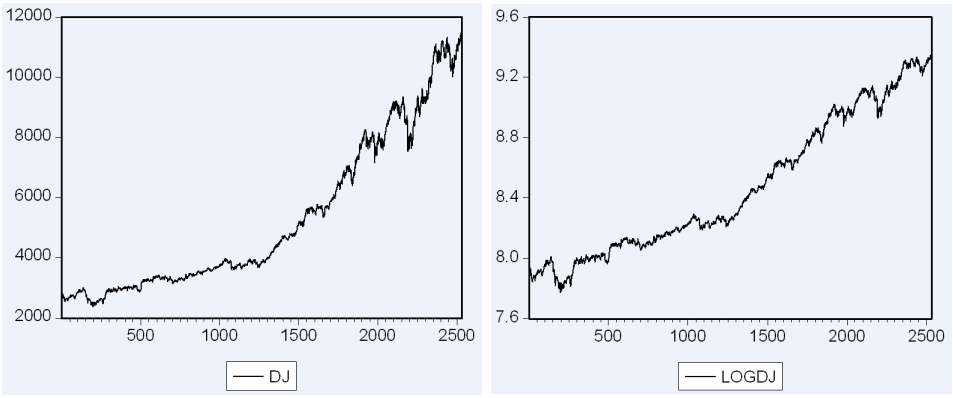


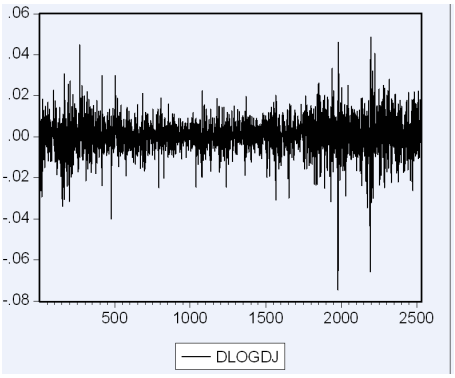


Dow-Jones index. A Dow-Jones index naponkénti zárási megfigyeléseit tanulmányozzuk az 1990. jan. 2.-1999. dec. 31 periódusban. A vonalgrafikonok (1) a tulajdonképpeni megfigyeléseket (DJ), (2) a logaritmusukat (LOGDJ) és (3) a logaritmusok első különbségét (DLOGDJ) mutatják. Megjegyezzük, hogy ez utóbbi megközelítőleg egyenlő a Dow-Jones index százalékbeli változásával, mert

$$\ln y_t - \ln y_{t-1} = \ln \frac{y_t}{y_{t-1}} = \ln \left(1 + \frac{y_t - y_{t-1}}{y_{t-1}} \right) \simeq \frac{y_t - y_{t-1}}{y_{t-1}}.$$

Itt a megközelítés annál pontosabb minél közelebb van $\frac{y_t - y_{t-1}}{y_{t-1}}$ a nullához.





Megjegyzések:
- mindkét idősor lényegében exponenciálisan változik, vagyis exponenciális trenddel rendelkezik,
- a logaritmusuk lényegében lineárisan változik: lineáris trenddel rendelkezik,
- mindkét különbség várható értéke konstans, $= 0$,
- az ipari termelés mutat némi szezonalitást (lásd: (3) az eredeti megfigyelések egy leszűkített (1993:03-1998:03) periódusban (IP)), de a logaritmusok negyedik különbsége már kevésbé,
- a DJ index varianciája változik.

Az idősorelemzés fő célja olyan modellek szerkesztése, amelyek segítségével minél pontosabb előrejelzéseket tudunk meghatározni egy bizonyos idősorra. Ebben a fejezetben stacionárius idősorok elemzését tanulmányozzuk, vagyis olyanokét, amelyeknek a főbb statisztikai tulajdonságaik nem változnak az idő során. A fő célunk olyan egyszerű modellek szerkesztése, amelyek tükrözik a vizsgált idősor korrelációs struktúráját. A korrelációs struktúra meghatározza a különböző megfigyelések közötti időbeni összefüggéseket, ami alapul szolgál az előrejelzésekhez.

Egyes idősor-modellek becslése elvégezhető az eddig tanult módszerekkel. A fennebb tanulmányozott két idősor nem stacionárius mert trenddel és/vagy szezonalitással rendelkezik. Az ilyenszerű idősorokat különböző transzformációkkal (pl. különbségekkel) átalakíthatjuk stacionáriussá.

Stacionárius idősorok

Egy y_t idősort **(gyengén)** stacionáriusnak nevezünk, ha a $\mu = E[y_t]$ várható értéke, $\gamma_0 = E\left[(y_t - \mu)^2\right]$ varianciája és k-adik $\gamma_k = E[(y_t - \mu)(y_{t-k} - \mu)]$, $k = 1, 2, ...,$ autokovarianciája nem változik az időindex függvényében. Emiatt egy stacionárius folyamat $\rho_k = \dfrac{\gamma_k}{\gamma_0}$ autokorrelációi sem változnak az idő

függvényében.

A stacionárius folyamatokra olyan modelleket tanulmányozunk, amelyek tükrözik a folyamat és a késleltetett értékei közötti összefüggéseket. Ezért a függővváltozó y_t és a független változók ennek különböző késleltetett értékei: y_{t-k}, $k \geq 1$. Ez más mint az eddig tanulmányozott modellek, mert azoknál a független változók általában a függővváltozótól eltérő változók.

A fő célunk olyan modellek szerkesztése, amelyekkel pontos előrejelzéseket végezhetünk, és a lehető legkevesebb ismeretlen paramétert tartalmazzák. A következő típusú modelleket tanulmányozzuk: autoregresszív, mozgóátlag és a kettő kombinációja.

6.1. Autoregresszív modellek

Egy p-rendű autoregresszív modellt, amelynek jelölése AR(p), a következőképpen értelmezünk:

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t, \quad t = p+1, ..., n. \qquad (*)$$

Az α, ϕ_1, ..., ϕ_p ismeretlen paraméterek és ε_t hibaváltozó, amit fehérzajnak feltételezünk, vagyis, egy olyan folyamat, amelynek várható értéke 0 (A2), varianciája konstans (A3) és 0 autokorrelációjú (A4):

$$\begin{aligned} E[\varepsilon_t] &= 0 \quad \text{minden } t\text{-re,} \\ var(\varepsilon_t) &= \sigma^2 \quad \text{minden } t\text{-re,} \\ cov(\varepsilon_s, \varepsilon_t) &= 0 \quad \text{minden } s \neq t\text{-re,} \end{aligned}$$

és ezeken kívül, ε_t kovarianciája az y_t minden késleltetett értékével 0, vagyis,

$$E[\varepsilon_t y_{t-k}] = 0 \quad \text{minden } k \geq 1.$$

Megfigyelhetjük, hogy egy AR(p) modell tulajdonképpen egy többváltozós lineáris modell, ahol a regresszorok (független változók) a függővváltozó késleltetett értékei.

Egy AR(p) modellel meghatározott y_t idősort AR(p) folyamatnak nevezzük. Egy ilyen AR(p) folyamat paraméterei alapján eldönthető, hogy a folyamat stacionárius-e.

Az AR folyamatok egyes statisztikai tulajdonságát könnyen levezethetjük. Ezek –elsősorban az autokorreláció– hasznosak lesznek amikor a folyamatok típusának azonosítását végezzük.

AR(p) folyamat várható értéke
Ha az AR(p) modellben mindkét oldalnak vesszük a várható értékét, akkor

$$\mu = \alpha + \phi_1\mu + ... + \phi_p\mu + 0, \quad \text{tehát} \quad \mu = E[y_t] = \frac{\alpha}{1 - \sum_{i=1}^{p}\phi_i}.$$

AR(1) folyamat varianciája és autokorrelációja
Egy AR(1) folyamat az

$$y_t = \alpha + \phi_1 y_{t-1} + \varepsilon_t, \quad t = 2, ..., n$$

alakban írható. Mivel $var(y_t) = var(y_{t-1}) = \gamma_0$ és $E[\varepsilon_t y_{t-1}] = 0$, ezért

$$\gamma_0 = \phi_1^2\gamma_0 + \sigma^2, \quad \text{tehát} \quad \gamma_0 = \frac{\sigma^2}{1 - \phi_1^2}.$$

($\phi_1^2 < 1$ mivel a folyamat stacionárius.)

Első rendű autokovariancia:

$$\begin{aligned}
\gamma_1 &= E[(y_t - \mu)(y_{t-1} - \mu)] = E[y_t y_{t-1}] - \mu^2 \\
&= E[(\alpha + \phi_1 y_{t-1} + \varepsilon_t) y_{t-1}] - \mu^2 \\
&= \alpha E[y_{t-1}] + \phi_1 E\left[y_{t-1}^2\right] + E[\varepsilon_t y_{t-1}] - \mu^2 \\
&= \alpha E[y_{t-1}] + \phi_1\left\{var(y_{t-1}) + \mu^2\right\} + 0 - \mu^2 \\
&= \alpha\mu + \phi_1\left(\gamma_0 + \mu^2\right) - \mu^2 \\
&= \phi_1\gamma_0 + \mu(\alpha + \phi_1\mu - \mu) \\
&= \phi_1\gamma_0 + \mu\left(\alpha + (\phi_1 - 1)\frac{\alpha}{1 - \phi_1}\right) \\
&= \phi_1\gamma_0 + \mu(\alpha - \alpha) = \phi_1\gamma_0.
\end{aligned}$$

Az első rendű autokorreláció

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\phi_1\gamma_0}{\gamma_0} = \phi_1.$$

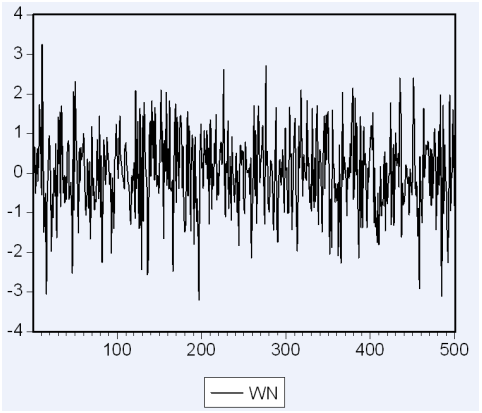
A k-adik autokorreláció

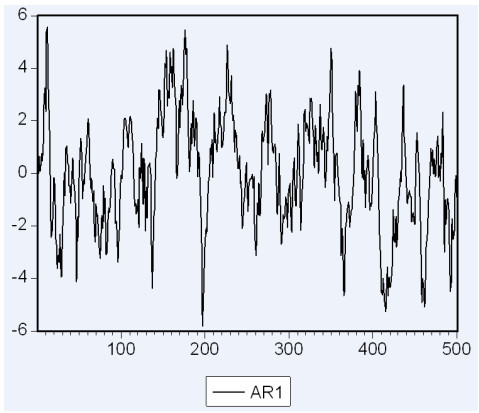
$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi_1^k.$$

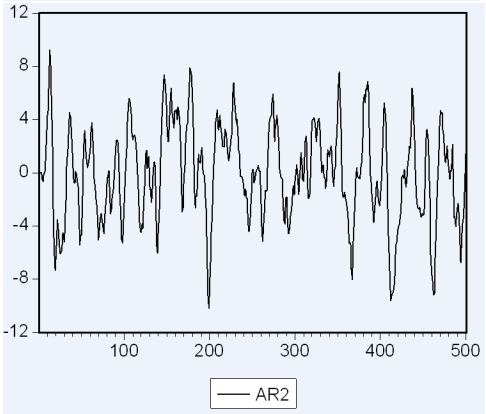
Mivel $|\phi_1| < 1$ (mert csak ekkor stacionárius), az autokorrelációk exponenciálisan csökkennek ha $k \to \infty$. Ez a tulajdonság hasznos lesz arra, hogy egy AR

folyamatot megkülönböztessünk más folyamatoktól.

Példák. A következő vonalgrafikonok egy (1 varianciájú) fehérzajt, az $y_t = 0.9y_{t-1} + \varepsilon_t$ AR(1) folyamatot és egy $y_t = 1.5y_{t-1} - 0.6y_{t-2} + \varepsilon_t$ modellel megadott AR(2) folyamatot ábrázolnak:







Az AR(2) váltakozása a legritkább, ami az egymás után következő megfigyelések közti összefüggés létezését tükrözi, az AR(1) váltakozása kicsit sűrűbb az AR(2) váltakozásánál.

6.2. Mozgóátlagok

Egy q-rendű mozgóátlag folyamatot (jelölése MA(q)) az

$$y_t = \alpha + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}, \quad t = q+1, ..., n,$$

modellel értelmezünk, ahol ε_t egy σ^2 varianciájú fehérzaj. Egy ilyen folyamat mindig stacionárius; várható értéke:

$$\mu = E[y_t] = \alpha,$$

varianciája:

$$\gamma_0 = \sigma^2 \left(1 + \sum_{i=1}^{q} \theta_i^2\right)$$

és autokovarianciája:

$$\gamma_k = \begin{cases} \sigma^2 \left(\theta_k + \sum_{i=k+1}^{q} \theta_i \theta_{i-k}\right) & \text{ha } k \leq q \\ 0 & \text{ha } k > q. \end{cases}$$

Következésképpen a k-adik autokorreláció 0 minden $k > q$-ra és általában nem 0 $k \leq q$-ra.

Példa. MA(1)-re $y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1}$,

$$\begin{aligned} \gamma_0 &= var(y_t) = var(\varepsilon_t) + \theta_1^2 var(\varepsilon_{t-1}) = \sigma^2\left(1 + \theta_1^2\right), \\ \gamma_1 &= E[(y_t - \mu)(y_{t-1} - \mu)] = E[(\varepsilon_t + \theta_1\varepsilon_{t-1})(\varepsilon_{t-1} + \theta_1\varepsilon_{t-2})] \\ &= E[\varepsilon_t\varepsilon_{t-1}] + \theta_1 E[\varepsilon_t\varepsilon_{t-2}] + \theta_1 E\left[\varepsilon_{t-1}^2\right] + \theta_1^2 E[\varepsilon_{t-1}\varepsilon_{t-2}] \\ &= 0 + 0 + \theta_1 E\left[\varepsilon_{t-1}^2\right] + 0 = \theta_1\sigma^2, \\ \gamma_2 &= 0. \end{aligned}$$

Mivel egy MA modellben az y_t folyamat az ε_t hibaváltozó jelen és késleltetett értékeinek a függvénye, ezért bizonyos értelemben egy MA és egy AR folyamatot egymás fordítottjának tekinthetjük, ugyanis az utóbbit értelmezhetjük úgy, hogy az ε_t az y_t hibaváltozó jelen és késleltetett értékeinek a függvénye:

$$\varepsilon_t = y_t - \alpha - \phi_1 y_{t-1} - ... - \phi_p y_{t-p}, \quad t = p+1, ..., n.$$

Ez a fajta értelmezés hasznos, ugyanis bizonyos esetekben azt is jelenti, hogy egy nagy rendű MA folyamat kifejezhető egy kis rendű AR folyamatként, vagyis kevesebb ismeretlen paraméterrel. A két folyamat kombinációja egy folyamatba egy olyan folyamatot eredményez, ami kevés paraméterrel komplexebb idősorokat képes leírni.

6.3. Autoregresszív mozgóátlag (ARMA)

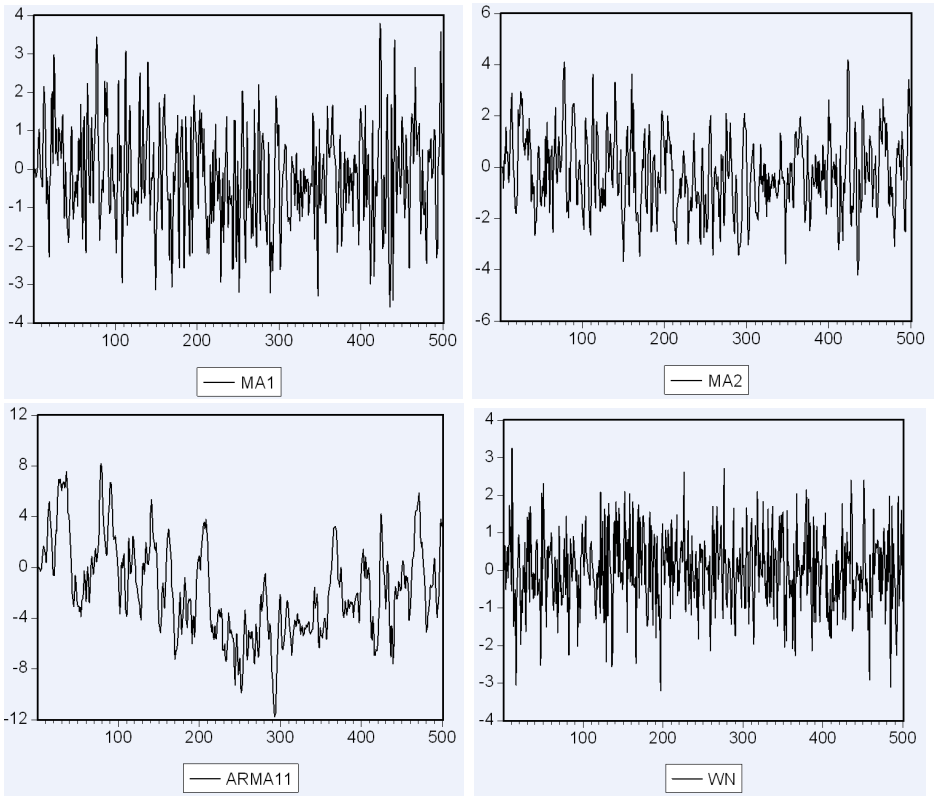
Az AR és MA modellek kombinációja. Egy y_t folyamatot (p, q)-rendű autoregresszív mozgóátlagnak (ARMA(p, q)) neveznek ha

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}.$$

Egy ARMA folyamat stacionárius, ha az AR alkotórésze stacionárius.

Az ARMA modellek hasznosak, ugyanis sok esetben kis rendű ARMA modellekkel pontosan megközelíthetünk jóval nagyobb rendű AR és MA modelleket. Más szóval, az ARMA modellek képesek kevés paraméter segítségével pontosan leírni bonyolult folyamatokat.

Példák. A következő vonalgrafikonok az $y_t = \varepsilon_t + 0.9\varepsilon_{t-1}$ MA(1), az $y_t = \varepsilon_t + 0.9\varepsilon_{t-1} + 0.8\varepsilon_{t-2}$ MA(2), az $y_t = 0.9y_{t-1} + \varepsilon_t + 0.8\varepsilon_{t-1}$ ARMA$(1, 1)$ folyamatot, és a fenti fehérzajt ábrázolják:



Az MA(1) váltakozása sűrű, majdnem eléri a fehérzaj váltakozásának sűrűségét; az MA(2) váltakozása ritkább, míg az ARMA$(1, 1)$-é egészen ritka.

A vonalgrafikonok alapján láthatjuk a folyamatok néhány jellegzetes vonását, de nehéz a folyamat típusát egyértelműen meghatározni. Például, az AR(2) és az ARMA(1, 1) eléggé hasonlítanak, de különböznek is abban, hogy az ARMA(1, 1) nem mindenhol 0 körül váltakozik. Ezért tovább tanulmányozzuk ezeket a folyamatokat az autokorrelációk és a parciális autokorrelációk segítségével.

6.4. Autokorreláció és parciális autokorreláció

Egy idősor egymás után következő értékeinek korrelációi hasznosak az idősor jövőbeni értékeinek az előrejelzésére. Most azt tanulmányozzuk, hogy az autokorreláció és a parciális autokorreláció hogy használható erre a célra. Amint már említettük, ezek arra használhatók, hogy megkülönböztessük a modellek típusát (AR, MA vagy ARMA).

Egy idősor autokorrelációs függvénye (ACF) az autokorrelációk sorozata:

$$\rho_k = corr\left(y_t, y_{t-k}\right) = \frac{\gamma_k}{\gamma_0}, \quad k = 1, 2, ...,$$

ahol $\gamma_k = E\left[\left(y_t - \mu\right)\left(y_{t-k} - \mu\right)\right]$ a k-ad rendű autokovariancia, $k = 0, 1, 2, ...$

MA folyamatok meghatározása az ACF segítségével:
Egy MA(q) folyamat

$$y_t = \alpha + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}, \quad t = q+1, ..., n,$$

alakban írható, ahol ε_t σ^2 varianciájú fehérzaj, stacionárius és az autokovarianciák $\gamma_k = 0$, $k > q$-ra, és $\gamma_q = \sigma^2\left(\theta_q + 0\right) = \sigma^2\theta_q \neq 0$. Tehát az MA($q$) ACF-e teljesíti azt, hogy $\rho_k = 0$, $k > q$ és $\rho_q \neq 0$.

Ennek a fordítottja is igaz: **Ha egy stacionárius folyamat ACF-e teljesíti azt, hogy** $\rho_k = 0$, $k > q$ **és** $\rho_q \neq 0$, **akkor a folyamat egy MA**(q) **folyamat.** Ez utóbbi tulajdonságot felhasználhatjuk arra, hogy egy MA folyamatot megkülönböztessünk más folyamatoktól (AR, ARMA), és meghatározzuk a folyamat rendjét.

AR folyamatok meghatározása a PACF segítségével
Láttuk, hogy egy AR(1) folyamat ρ_k autokorrelációi fokozatosan (exponenciálisan) csökkennek ha $k \to \infty$. Általában, bármilyen rendű AR folyamatnak megvan ez a tulajdonsága, viszont $p > 1$ rendű AR(p) folyamatoknál a csökkenés nem feltétlenül monoton. Ezzel a tulajdonsággal megkülönböztethetünk egy AR

folyamatot egy MA folyamattól, de nem biztos, hogy egy ARMA folyamattól is. Ezenkívül, ha tudjuk is, hogy a folyamat AR, ez a tulajdonság nem nyújt elég információt ahhoz, hogy a rendjét meghatározzuk.

Legyen az AR(p) stacionárius modell

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t, \quad t = p+1, ..., n.$$

Ebben az esetben, ha egy olyan modellt becsülünk, amelyben több késleltetett érték szerepel ($k > p$),

$$y_t = \alpha + \phi_{k1} y_{t-1} + ... + \phi_{kp} y_{t-p} + ... + \phi_{kk} y_{t-k} + \omega_t, \quad t = k+1, ..., n,$$

akkor $\phi_{kk} = 0$.

A ϕ_{kk} paramétert az y_t és y_{t-k} parciális autokorrelációs együtthatójának nevezik, és a ϕ_{kk} sorozatot, $k = 1, 2, ...$, az y_t parciális autokorrelációs függvényének (PACF). Tehát egy stacionárius AR(p) folyamat parciális autokorrelációira fennáll, hogy $\phi_{pp} \neq 0$ és $\phi_{kk} = 0$, $k > p$-re.

Ennek a fordítottja is igaz: **Ha egy stacionárius folyamat PACF-e teljesíti, hogy $\phi_{pp} \neq 0$ és $\phi_{kk} = 0$, $k > p$, akkor ez a folyamat egy AR(p) folyamat.** Ezt a tulajdonságot használhatjuk arra, hogy egy AR folyamatot megkülönböztessünk egy MA vagy egy ARMA folyamattól, és meghatározzuk a folyamat rendjét.

A gyakorlatban az ACF vagy a PACF nem ismert, ezért a megfigyelések alapján a minta-ACF (SACF) és a minta-PACF-fel (SPACF) becsüljük őket. A ρ_k autokorrelációk becslőfüggvénye (a k-ad rendű autokorrelációs együttható):

$$r_k = \frac{\sum_{t=k+1}^{n} (y_t - \overline{y})(y_{t-k} - \overline{y})}{\sum_{t=1}^{n} (y_t - \overline{y})^2}.$$

A ϕ_{kk} parciális autokorrelációkat OLS-sel becsüljük a következő modellben:

$$y_t = \alpha + \phi_{k1} y_{t-1} + ... + \phi_{kk} y_{t-k} + \omega_t, \quad t = k+1, ..., n.$$

Megjegyezzük, hogy különböző k-ra a modellek, amelyek alapján a ϕ_{kk}-t becsüljük különböznek, mert tartalmazzák y_{t-k}-t. Például, ϕ_{11}-t a

$$y_t = \alpha + \phi_{11} y_{t-1} + \omega_t, \quad t = 2, ..., n,$$

modellben, míg ϕ_{22}-t a

$$y_t = \alpha + \phi_{21} y_{t-1} + \phi_{22} y_{t-2} + \omega_t, \quad t = 3, ..., n,$$

modellben becsüljük, stb.

Példa. Az első 10 minta-autokorrelációt a hat bemutatott folyamatra a következő tábázatok tartalmazzák. Az első a fehérzajra, es a két AR folyamatra:

	WN		AR(1)		AR(2)	
Lag	SACF	SPACF	SACF	SPACF	SACF	SPACF
1	0.023	0.023	0.874	0.874	0.929	0.929
2	-0.007	-0.008	0.753	-0.050	0.777	-0.624
3	-0.041	-0.041	0.641	-0.026	0.594	0.011
4	-0.029	-0.028	0.547	0.006	0.412	0.007
5	-0.024	-0.024	0.466	0.000	0.252	-0.019
6	0.012	0.011	0.397	-0.002	0.122	-0.019
7	-0.020	-0.023	0.330	-0.036	0.021	-0.023
8	-0.038	-0.040	0.271	-0.008	-0.050	0.003
9	-0.025	-0.024	0.225	0.016	-0.095	-0.013
10	0.031	0.030	0.189	0.008	-0.123	-0.039

- A fehérzajra (WN) a SACF és a SPACF értékei általában kicsik (meg fogunk adni egy kritikus értéket, amely alapján eldönthető, hogy mennyi a "kicsi").
- Az AR(1) folyamatra az SACF értékek fokozatosan csökkennek, míg a SPACF kicsik a 2. késleltetett értéktől kezdve.
- Az AR(2) folyamatra az SACF értékek fokozatosan csökkennek, de abszolút értékben nem monotonul, míg a SPACF kicsik a 3. késleltetett értéktől kezdve.

A második táblázat a két mozgóátlag és az ARMA$(1,1)$ folyamat minta autokorrelációit és parciális autokorrelációit tartalmazza:

	MA(1)		MA(2)		ARMA(1,1)	
Lag	SACF	SPACF	SACF	SPACF	SACF	SPACF
1	0.465	0.465	0.622	0.622	0.941	0.941
2	-0.053	-0.342	0.298	-0.145	0.836	-0.434
3	-0.015	0.255	-0.046	-0.281	0.747	0.290
4	0.002	-0.204	-0.006	0.332	0.669	-0.185
5	-0.013	0.147	-0.029	-0.161	0.601	0.141
6	-0.019	-0.139	-0.028	-0.133	0.542	-0.080
7	-0.043	0.056	-0.033	0.225	0.492	0.080
8	-0.005	-0.010	0.003	-0.076	0.452	0.014
9	0.059	0.069	0.033	-0.055	0.418	-0.031
10	0.013	-0.086	-0.011	0.019	0.379	-0.038

- A MA(1) folyamatra a SACF kicsik a 2. késleltetett értéktől kezdve, míg a SPACF (nem monotonul) csökken.
- A MA(2) folyamatra a SACF kicsik a 3. késleltetett értéktől kezdve, míg a SPACF (nem monotonul) csökken.
- Az ARMA$(1,1)$ folyamatra a SACF és a SPACF is csökken, többnyire monotonul (abszolút értékben), de általában a monoton csökkenés nem jellemző az ARMA folyamatokra.

A SACF-t és a SPACF-t arra használjuk, hogy meghatározzuk egy stacionárius folyamat típusát (AR, MA vagy ARMA), és ezenkívül, ha a folyamat típusa AR vagy MA, akkor meghatározhatjuk ezek rendjét is. Egy ARMA folyamat rendjét nem tudjuk meghatározni az S(P)ACF segítségével.

6.4.1. Az autokorreláció és a parciális autokorreláció szignifikanciája

Ugyanúgy, mint a reziduumok autokorrelációjának tesztelésénél (4. fejezet), itt is fennáll, hogy, ha a nullhipotézis az, hogy a folyamat fehérzaj (tehát a vizsgált folyamat autokorrelációja 0), akkor

$$r_k, \widehat{\phi}_{kk} \approx N\left(0, \frac{1}{n}\right),$$

vagyis, a becsült autokorrelációk és parciális autokorrelációk megközelítőleg normál eloszlásúak 0 várható értékkel és $\frac{1}{n}$ varianciával. Ezért, a gyakorlatban elutasítjuk az autokorreláció és a parciális autokorreláció szignifikanciáját 5% szignifikancia szinten, ha $r_k, \left|\widehat{\phi}_{kk}\right| < \frac{2}{\sqrt{n}}$.

Tudjuk azt is tesztelni, hogy a modell AR vagy nem AR: ha a nullhipotézis az, hogy a folyamat AR(p), akkor

$$var\left(\widehat{\phi}_{kk}\right) \simeq \frac{1}{n}, \quad \text{ha } k > p,$$

ezért a gyakorlatban elutasítjuk a ϕ_{kk} ($k > p$) szignifikanciáját 5% szignifikancia szinten, ha $\left|\widehat{\phi}_{kk}\right| < \frac{2}{\sqrt{n}}$.

Tudjuk azt is tesztelni, hogy a modell MA vagy nem MA: ha a nullhipotézis az, hogy a folyamat MA(q), akkor

$$var\left(r_k\right) \simeq \frac{1 + 2\sum_{j=1}^{q} r_j^2}{n}, \quad \text{ha } k > q,$$

tehát a gyakorlatban elutasítjuk a ρ_k ($k > q$) szignifikanciáját 5% szignifikancia szinten, ha

$$|r_k| < \frac{2\sqrt{1 + 2\sum_{j=1}^{q} r_j^2}}{\sqrt{n}} \quad \text{minden } k > q\text{-ra.}$$

A hat bemutatott folyamatra az 5%-os kritikus érték fehérzaj nullhipotézisre és AR nullhipotézisre $\frac{2}{\sqrt{500}} = 0.089$. Az autokorrelációk és a parciális autokorrelációk

szignifikanciáját helyesen utasítjuk el vagy nem utasítjuk el a hat bemutatott folyamatra. A MA(1) nullhipotézisre a kritikus érték

$$\frac{2\sqrt{1+2r_1^2}}{\sqrt{n}} = \frac{2\sqrt{1+2\cdot 0.465^2}}{\sqrt{500}} = 0.107,$$

míg a MA(2) nullhipotézisre

$$\frac{2\sqrt{1+2\left(r_1^2+r_2^2\right)}}{\sqrt{n}} = \frac{2\sqrt{1+2\left(0.622^2+0.298^2\right)}}{\sqrt{500}} = 0.124.$$

Tehát az autokorrelációk szignifikanciáját helyesen utasítjuk el vagy nem utasítjuk el az MA(1) (és az MA(2)) folyamatra.

Az ARMA(1, 1) folyamatra több autokorreláció és parciális autokorreláció szignifikáns mindhárom típusú nullhipotézis mellett. Ebből azt a következtetést vonjuk le, hogy ezt a folyamatot nem tudjuk jól modellezni fehérzaj, AR vagy MA típusú folyamatok segítségével.

A gyakorlatban, ha nem AR vagy MA típusúnak találunk egy adott folyamatot, akkor azt a következtetést vonjuk le, hogy ARMA típusú. Viszont, mivel az S(P)ACF nem képes információt nyújtani az ARMA lehetséges p és q rendjéről, ezért különböző p és q értékekkel próbálkozunk (lásd alább az Idősor-modellek becslése alfejezetet).

6.5. Előrejelzések

ARMA típusú stacionárius idősorok előrejelzése a lineáris modell függőváltozójának előrejelzéséhez hasonló elveket követ, de mivel a hibaváltozók késleltetett értékei is szerepelhetnek a modellben, ezért az előrejelzések néhány új vonását is tárgyaljuk.

Az előrejelzéseket arra az esetre tárgyaljuk, amikor ismerjük a paraméterek és a megfigyelésekhez tartozó hibaváltozók értékét. Utána aztán tárgyaljuk azt is, hogy hogyan végzünk előrejelzést a gyakorlatban, amikor ezeket az értékeket nem ismerjük.

Tegyük fel, hogy egy

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}$$

ARMA(p, q) folyamatra az $y_1, y_2, ..., y_n$ megfigyeléseink vannak. Ennek a folyamatnak az egylépéses előrejelzése:

$$\widehat{y}_{n+1} = \alpha + \phi_1 y_n + ... + \phi_p y_{n+1-p} + 0 + \theta_1 \varepsilon_n + ... + \theta_q \varepsilon_{n+1-q},$$

míg az $n+1$-edik megfigyelés valódi értéke

$$y_{n+1} = \alpha + \phi_1 y_n + ... + \phi_p y_{n+1-p} + \varepsilon_{n+1} + \theta_1 \varepsilon_n + ... + \theta_q \varepsilon_{n+1-q}.$$

Az előrejelzésnél a fő elv az, hogy az ε_{n+1}, ε_{n+2},... jövőbeni hibaváltozók helyett 0-t teszünk. Az előrejelzés hibája

$$f_{n+1} = y_{n+1} - \widehat{y}_{n+1} = \varepsilon_{n+1},$$

aminek a varianciája $var\left(f_{n+1}\right) = \sigma^2$.

Ki tudunk számítani többlépéses előrejelzéseket is. Például, a fenti ARMA folyamat kétlépéses előrejelzése

$$\begin{aligned}\widehat{y}_{n+2} = \alpha + \phi_1 \widehat{y}_{n+1} + \phi_2 y_n + ... + \phi_p y_{n+2-p} \\ + 0 + \theta_1 \cdot 0 + \theta_2 \varepsilon_n + ... + \theta_q \varepsilon_{n+2-q},\end{aligned}$$

míg az $n+2$-edik megfigyelés valódi értéke

$$\begin{aligned}y_{n+2} = \alpha + \phi_1 y_{n+1} + \phi_2 y_n + ... + \phi_p y_{n+2-p} \\ + \varepsilon_{n+2} + \theta_1 \varepsilon_{n+1} + \theta_2 \varepsilon_n + ... + \theta_q \varepsilon_{n+2-q}.\end{aligned}$$

Ebben az esetben az előrejelzés hibája

$$\begin{aligned}f_{n+2} = y_{n+2} - \widehat{y}_{n+2} = \phi_1 \left(y_{n+1} - \widehat{y}_{n+1}\right) + \varepsilon_{n+2} + \theta_1 \varepsilon_{n+1} \\ = \phi_1 \varepsilon_{n+1} + \varepsilon_{n+2} + \theta_1 \varepsilon_{n+1} = \left(\phi_1 + \theta_1\right) \varepsilon_{n+1} + \varepsilon_{n+2}.\end{aligned}$$

Tehát $var\left(f_{n+2}\right) = \sigma^2 \left(\left(\phi_1 + \theta_1\right)^2 + 1\right)$.

Ha a hibaváltozók normál eloszlásúak és a σ^2 ismert, akkor

$$\Pr\left(\left|\frac{f_{n+2}}{\sqrt{var\left(f_{n+2}\right)}}\right| < 1.96\right) = 0.95,$$

és ez alapján a következő 95%-os intervallum-előrejelzést kapjuk:

$$\widehat{y}_{n+2} - 1.96\sqrt{var\left(f_{n+2}\right)} \leq y_{n+2} \leq \widehat{y}_{n+2} + 1.96\sqrt{var\left(f_{n+2}\right)}.$$

Kettőnél több-lépéses előrejelzéseket és intervallum-előrejelzéseket hasonlóan számítunk ki.

A gyakorlatban nem ismerjük a paramétereket és a hibaváltozókat, tehát a becsléseiket használjuk (az ARMA modell becslését nem tárgyaljuk). Tehát a gyakorlatban a fenti kétlépéses előrejelzés:

$$\widehat{y}_{n+2} = \widehat{\alpha} + \widehat{\phi}_1\widehat{y}_{n+1} + \widehat{\phi}_2 y_n + ... + \widehat{\phi}_p y_{n+2-p} + \widehat{\theta}_2\widehat{\varepsilon}_n + ... + \widehat{\theta}_q\widehat{\varepsilon}_{n+2-q},$$

ahol $\widehat{\alpha}, \widehat{\phi}_1, \widehat{\phi}_2, ..., \widehat{\phi}_p, \widehat{\theta}_2, ..., \widehat{\theta}_q$ a paraméterek becslései.

Megjegyzés. Mivel azt feltételezzük, hogy a paraméterek értékei ismertek, ezért amikor az előrejelzés hibáját számítjuk ki, nem vesszük figyelembe azt, hogy a paraméterek becslőfüggvényei vvk. Ezért az előrejelzés varianciája csak megközelítése a valódi varianciának. Ez a megközelítés annál jobb minél nagyobb a megfigyelések száma. Ugyanezt elmondhatjuk az intervallum-előrejelzésről is.

6.6. Idősor-modellek becslése

A gyakorlatban az idősor-modellezés és becslés egy több lépéses folyamat, amely a modell meghatározását, úgynevezett diagnosztikai vizsgálatot és a modell kiigazítását foglalja magába.

A lépések előzetes áttekintése:

1. lépés. Ábrák. Készítsünk vonalgrafikont a megfigyelésekről. Ennek a fő célja annak vizsgálata, hogy az idősor stacionárius-e. Ha szezonalitásra vagy valamilyen trendre utaló jelt vennénk észre, akkor készítsünk vonalgrafikont a negyedik különbségről vagy az első különbségről vagy a megfigyelések logaritmusáról (exponenciális trend esetén).

A következő lépésekben azt feltételezzük, hogy az idősor stacionárius.

2. lépés. Az ARMA modell meghatározása. A SACF és a SPACF meghatározásával eldönthetjük, hogy milyen rendű modell a legmegfelelőbb, amennyiben AR vagy MA modellt alkalmazhatunk. Ha egy ARMA(p, q) modell a legmegfelelőbb, akkor a SACF és a SPACF útmutatást nyújt a p és a q meghatározására.

3. lépés. A paraméterek becslése. A 2. lépésben meghatározott p és q ARMA rendekre becsüljük az ARMA(p, q) modell paramétereit.

4. lépés. Diagnosztikai vizsgálat. A becsült modellt megvizsgáljuk statisztikai tesztek segítségével. Ez abból áll, hogy teszteljük a modell hibaváltozóinak homoszkedaszticitását és autokorrelációját és a becslések szignifikanciáját.

Ha több modellünk van amelyek teljesítik a feltételeket, akkor összehasonlítjuk az előrejelzési teljesítményüket, és ez alapján választjuk ki a legjobb modellt, szem előtt tartva, hogy a modell minél kevesebb paramétert tartalmazzon.
5. lépés. A modell kiigazítása. Ha a 4. lépésben azt kapjuk, hogy a modell nem teljesíti a feltételeket, akkor kiigazítjuk a diagnosztikai vizsgálat eredménye alapján. Ha a hibaváltozók korreláltak, akkor növeljük az ARMA modell p és q rendjét. Ha a hibaváltozók nem korreláltak, és egyes becslések nem szignifikánsak, akkor csökkentjük a p-t vagy a q-t. Ezután a 3. lépéstől folytatjuk az eljárást.
6. lépés. A modell alkalmazása. Ha a 4. lépésben kiválasztottuk a legjobb modellt, ezt alkalmazhatjuk az idősor jövőbeni értékeinek az előrejelzésére.

Az első két lépést már tárgyaltuk. Most a 3. és 4. lépéseket tárgyaljuk részletesen.

6.6.1. Stacionárius ARMA modellek becslése (3. lépés)

ARMA modellek OLS becslőfüggvénye általában nem torzítatlan és nem konszisztens. Ezért az ARMA modellek becslését az OLS-nél bonyolultabb módszerrel végzik. A gyakorlatban ARMA modellek becslését SPSS-el végezhetjük.

Ez alól kivételek az AR modellek. Egy stacionárius AR(p) modell

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t, \quad t = p+1, ..., n$$

ε_t hibaváltozója fehérzaj, ezért

$$\begin{aligned} E[\varepsilon_t] &= 0 \quad \text{minden } t\text{-re,} \\ E\left[\varepsilon_t^2\right] &= \sigma^2 \quad \text{minden } t\text{-re,} \\ E[\varepsilon_s \varepsilon_t] &= 0 \quad s \neq t, \\ E[\varepsilon_t y_{t-k}] &= 0 \quad \text{minden } k \geq 1 \text{ és minden } t \geq k+1. \end{aligned}$$

Ezek a feltételek mellett, amint a 4. Fejezetben láttuk, az OLS becslőfüggvény torzítatlan, konszisztens és rendelkezik a legjobb lineáris torzítatlan becslőfüggvény tulajdonsággal.

6.6.2. Diagnosztikai vizsgálat (4. lépés)

A 3. lépésben becsüljük a modellt, és mivel az idősorelemzés lényege olyan modellek szerkesztése és becslése, amelyek segítségével előrejelzéseket végezhetünk,

a diagnosztikai vizsgálat során a becsült modell és a szignifikancia tesztek helyességét ellenőrizzük. Ezért megvizsgáljuk, hogy:
(i) a hibaváltozó fehérzaj-e,
(ii) a becslések szignifikánsak-e,
(iii) az (i) és (ii)-t teljesítő modellek közül melyiknek jobb az előrejelzési teljesítménye.

(i) Annak a vizsgálata, hogy a hibaváltozó fehérzaj-e hasonló a homoszkedaszticitás és az autokorreláció teszteléséhez a többváltozós lineáris modellnél. A reziduumok vonalgrafikonja nyújt némi információt a heteroszkedaszticitásról és az autokorrelációról. Ezenkívül használjuk a White vagy Breusch-Pagan heteroszkedaszticitás-tesztet.

A hibaváltozók autokorrelációját tesztelhetjük a reziduumok autokorrelációs együtthatóival (SPSS-ből):

$$r_k = \frac{\sum_{t=k+1}^{n} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t-k}}{\sum \widehat{\varepsilon}_t^2},$$

amelyek szignifikanciáját elutasítjuk 5% szignifikancia szinten ha $|r_k| < \frac{2}{\sqrt{n}}$. (Itt $\widehat{\varepsilon}_t$ jelöli a t megfigyelésnek megfelelő reziduumot.) A Box-Ljung teszt Q-statisztikája (SPSS-ből)

$$Q = n \sum_{k=1}^{h} \frac{n+2}{n-k} r_k^2 \approx \chi^2 (h - p - q)$$

használható az első h autokorreláció szignifikanciájának egyidejű tesztelésére ha $h > p + q$. Itt p, q azon ARMA modell rendjei, amelynek a hibaváltozóit vizsgáljuk. Vegyük észre, hogy ez a Q-statisztika eloszlása különbözik a lineáris modellek hibaváltozóinak autokorrelációjára alkalmazott teszt Q-statisztikájának eloszlásától, ami $\chi^2 (h)$.
Ha a hibaváltozók autokorrelációja szignifikáns, akkor kell növelni az ARMA modell p és q rendjét.

(ii) Ha olyan modellt kaptunk, amelyben a hibaváltozó fehérzaj, akkor erre a modellre teszteljük a becslések szignifikanciáját. Ezt egyéni (t) és egyidejű (F) szignifikancia-tesztek segítségével végezzük. Ha bizonyos becslések nem szignifikánsak akkor csökkenthetjük az ARMA modell p rendjét. Az ARMA modell q rendjének csökkentésére nem használhatjuk az F-tesztet.

(iii) Két modell közül kiválaszthatjuk a jobbikat az előrejelzési teljesítményük alapján. Egy modell előrejelzési teljesítményét felmérhetjük úgy, hogy csak

az első $n_0 < n$ megfigyelést $(y_1, ..., y_{n_0})$ használva becsüljük a modellt és a megmaradó $y_{n_0+1}, ..., y_n$ értékeket előrejelezzük: $\widehat{y}_{n_0+1}, ..., \widehat{y}_n$. Ezután kiszámítjuk az előrejelzések átlagos négyzetes eltérését (MSPE), ami megmutatja, hogy az előrejelzések mennyire közel vannak a valódi értékekhez:

$$MSPE = \frac{1}{m} \sum_{h=1}^{m} \left(y_{n_0+h} - \widehat{y}_{n_0+h} \right)^2 .$$

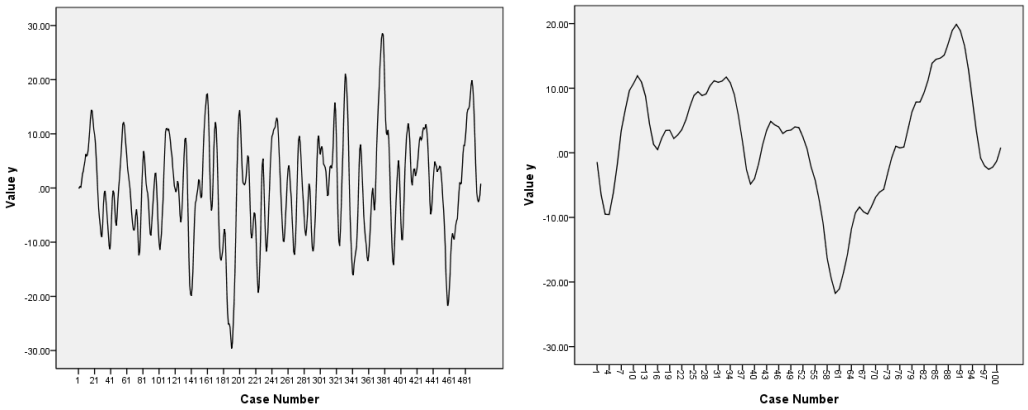
Annak a modellnek jobb az előrejelzési teljesítménye, amelyre kisebb az előrejelzések átlagos négyzetes eltérése.

6.6.3. Példa: Idősor-modellek becslési folyamata

Ebben a példában becsüljük egy y-nal jelzett $n = 500$ megfigyelésből álló idősor modelljét, követve a tárgyalt lépéseket.

1. lépés. Ábrák

Az alábbi két vonalgrafikon közül a baloldali a teljes idősort mutatja, míg a jobboldali az utolsó 100 megfigyelést. Az y_t idősor meglehetősen szabályosan váltakozik a 0 érték körül. Az utóbbin nem veszünk észre szezonalitást, úgyhogy nincs szükség a megfigyelések transzformációjára.



2. lépés. Az ARMA modell meghatározása

A kritikus érték fehérzaj és AR nullhipotézis esetén $\frac{2}{\sqrt{500}} = 0.089$. Az alábbi táblazatokban (SPSS-ből) mindkét autokorrelációs függvény csökken fokozatosan (a PACF csökkenése nem monoton). Másrészt mindkét autokorrelációnak több értéke is szignifikáns, ezért az idősor nem AR vagy MA, hanem ARMA. Próbálkozzunk az ARMA$(1, 1)$ modell becslésével.

Autocorrelations

Series:y

Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	.954	.045	457.505	1	.000
2	.830	.045	804.968	2	.000
3	.662	.044	1026.348	3	.000
4	.485	.044	1145.417	4	.000
5	.325	.044	1199.124	5	.000
6	.199	.044	1219.273	6	.000
7	.112	.044	1225.621	7	.000
8	.059	.044	1227.417	8	.000
9	.032	.044	1227.935	9	.000
10	.016	.044	1228.068	10	.000
11	.001	.044	1228.069	11	.000

a. The underlying process assumed is independence (white noise).
b. Based on the asymptotic chi-square approximation.

Partial Autocorrelations

Series:y

Lag	Partial Autocorrelation	Std. Error
1	.954	.045
2	-.876	.045
3	.571	.045
4	.025	.045
5	-.252	.045
6	.258	.045
7	-.051	.045
8	-.200	.045
9	.068	.045
10	-.033	.045
11	-.104	.045

.

3. lépés. Az ARMA$(1,1)$ paraméterek becslése

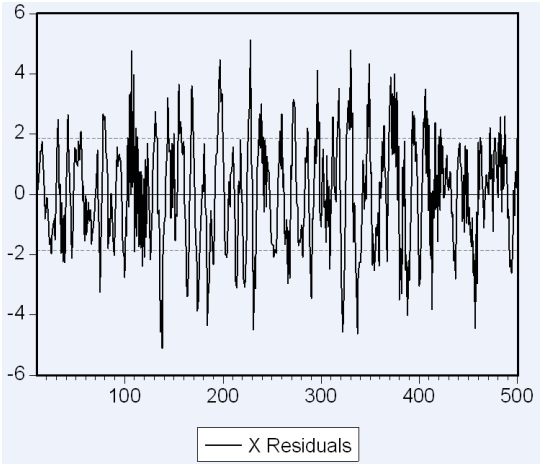
Az alábbi táblázat tartalmazza a becslési eredményt (SPSS-ből):

ARIMA Model Parameters

				Estimate	SE	t	Sig.
y-Model_1	y	No Transformation	Constant	.195	2.354	.083	.934
			AR Lag 1	.939	.015	61.561	.000
			MA Lag 1	-.781	.028	-27.418	.000

4. lépés. Diagnosztikai vizsgálat

A 3. lépésbeli becslés során kapott reziduumok vonalgrafikonja (Eviews-ból; SPSS-ben nem lehet) azt mutatja, hogy a reziduumok 0 körül váltakoznak, és a varianciájuk többnyire állandó:



A White-teszt (Eviews-ból; SPSS-ben nem lehet) nem utasítja el a homoszkedaszticitás nullhipotézist:

White Heteroskedasticity Test:				
F-statistic	0.059040	Probability		0.942676
Test Equation:				
Dependent Variable: RESID^2				
Method: Least Squares				
Sample: 11 500				
Included observations: 490				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.470092	0.243810	14.23279	0.0000
X(-1)	0.005290	0.020912	0.252982	0.8004
X(-1)^2	-0.000317	0.001468	-0.215818	0.8292
R-squared	0.000242	Mean dependent var		3.441042
Adjusted R-squared	-0.003863	S.D. dependent var		4.453430
S.E. of regression	4.462024	Akaike info criterion		5.835186
Sum squared resid	9696.005	Schwarz criterion		5.860866
Log likelihood	-1426.621	F-statistic		0.059040
Durbin-Watson stat	1.644567	Prob(F-statistic)		0.942676

A reziduumok autokorrelációi és parciális autokorrelációi, valamint a Box-Ljung teszt alapján a hibaváltozók korreláltak. Ez látható az alábbi táblázatból, ahol több (például az első, második, stb) autokorreláció is szignifikáns a fenti kritikus érték alapján:

Residual ACF

Model		1	2	3	4	5	6	7	8	9	10	11
y-Model_1	ACF	.539	.521	-.048	-.051	-.352	-.267	-.340	-.168	-.119	.026	.045
	SE	.045	.056	.065	.065	.065	.069	.071	.074	.075	.075	.075

Residual PACF

Model		1	2	3	4	5	6	7	8	9	10	11
y-Model_1	PACF	.539	.325	-.649	.094	.055	-.313	.097	.095	-.153	.023	.026
	SE	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045

A Box-Ljung teszt eredménye az alábbi táblázatban is mutatja az első 11 autokorreláció egyidejű szignifikanciáját:

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
y-Model_1	0	.963	481.181	16	.000	0

5. lépés. A modell kiigazítása

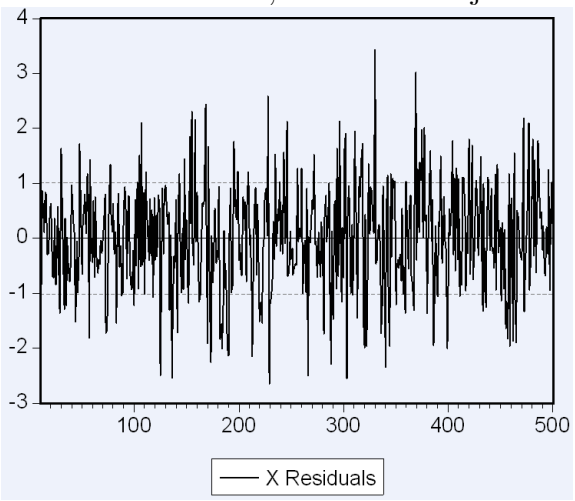
Mivel a hibaváltozók korreláltak (a reziduumok alapján), növeljük a modell AR és MA rendjét: legyen a modell ARMA$(3, 3)$. A 3. lépéstől folytatjuk.

3. lépés. Az ARMA$(3,3)$ paraméterek becslése
A becslés eredménye az alábbi táblázatban található:

ARIMA Model Parameters

					Estimate	SE	t	Sig.
y-Model_1	y	No Transformation	Constant		.194	1.207	.161	.872
			AR	Lag 1	2.247	.645	3.485	.001
				Lag 2	-1.751	.982	-1.784	.075
				Lag 3	.474	.406	1.167	.244
			MA	Lag 1	-.187	.656	-.285	.775
				Lag 2	-.186	.583	-.319	.750
				Lag 3	.534	.525	1.017	.310

4. lépés. Diagnosztikai vizsgálat
A reziduumok vonalgrafikonja (Eviews-ból) azt mutatja, hogy a reziduumok 0 körül váltakoznak, és a varianciájuk többnyire állandó.



A White-teszt nem elutasítja el a homoszkedaszticitás nullhipotézist:

White Heteroskedasticity Test:			
F-statistic	1.180778	Probability	0.305051

A reziduumok autokorrelációi és a Box-Ljung teszt alapján a hibaváltozók nem korreláltak. Az alábbi táblázat alapján az autokorrelációk mind kisebbek abszolút értékben mint a 0.089 kritikus érték:

Residual ACF

Model		1	2	3	4	5	6	7	8	9	10	11
y-Model_1	ACF	-.015	.019	.034	.000	-.010	-.027	-.034	-.003	.007	.067	.036
	SE	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045

Residual PACF

Model		1	2	3	4	5	6	7	8	9	10	11
y-Model_1	PACF	-.015	.019	.035	.001	-.012	-.028	-.035	-.002	.011	.070	.038
	SE	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045

A Ljung-Box teszt p-értéke 0.432, ami jóval meghaladja a 0.05 szignifikancia-szintet:

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
y-Model_1	0	.989	12.172	12	.432	0

5. lépés. A modell kiigazítása

Ez a modell teljesíti azokat a feltételeket, hogy a hibaváltozó fehérzaj legyen, de több változó nem szignifikáns az egyéni szignifikancia t-teszt alapján. Ezért a modellt úgy igazítjuk ki, hogy csökkentjük az AR és az MA rendeket 1-gyel. Tehát az új modell ARMA$(2, 2)$; folytatjuk a 3. lépéssel.

3. lépés. Az ARMA$(2, 2)$ paraméterek becslése

ARIMA Model Parameters

					Estimate	SE	t	Sig.
y-Model_1	y	No Transformation	Constant		.188	1.112	.169	.866
			AR	Lag 1	1.540	.037	42.090	.000
				Lag 2	-.648	.037	-17.725	.000
			MA	Lag 1	-.882	.030	-29.813	.000
				Lag 2	-.790	.029	-26.999	.000

4. lépés. Diagnosztikai vizsgálat

A reziduumok vonalgrafikonja (nincs feltüntetve) azt mutatja, hogy a reziduumok 0 körül váltakoznak, és a varianciájuk többnyire állandó. A White-teszt nem elutasítja el a homoszkedaszticitás nullhipotézist. A reziduumok autokorrelációi és a Box-Ljung teszt alapján a hibaváltozók nem korreláltak:

Residual ACF

Model		1	2	3	4	5	6	7	8	9	10	11
y-Model_1	ACF	-.001	.022	.033	-.016	-.015	-.025	-.036	.005	.021	.075	.047
	SE	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045

Residual PACF

Model		1	2	3	4	5	6	7	8	9	10	11
y-Model_1	PACF	-.001	.022	.033	-.016	-.017	-.025	-.034	.007	.023	.077	.044
	SE	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045	.045

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
y-Model_1	0	.989	13.518	14	.486	0

Ezenkívül, az ARMA$(2,2)$ modell összes változója szignifikáns. Az ARMA$(2,2)$ modellnek kevesebb paramétere van, ebből a szempontból kicsit jobbnak tekinthető mint az ARMA$(3,3)$ modell.

A két modell összehasonlítása előrejelzési teljesítményük alapján
Az ARMA$(3,3)$ és az ARMA$(2,2)$ modelleket becsüljük csak az első 495 megfigyelés alapján, és az utolsó 5 megfigyelésre többlépéses előrejelzést végzünk.
ARMA$(3,3)$:
A többlépéses előrejelzések képletei:

$$\begin{aligned}
\widehat{y}_{496} &= \widehat{\alpha} + \widehat{\phi}_1 y_{495} + \widehat{\phi}_2 y_{494} + \widehat{\phi}_3 y_{493} + \widehat{\theta}_1 \widehat{\varepsilon}_{495} + \widehat{\theta}_2 \widehat{\varepsilon}_{494} \\
&\quad + \widehat{\theta}_3 \widehat{\varepsilon}_{493} \\
\widehat{y}_{497} &= \widehat{\alpha} + \widehat{\phi}_1 \widehat{y}_{496} + \widehat{\phi}_2 y_{495} + \widehat{\phi}_3 y_{494} + 0 + \widehat{\theta}_2 \widehat{\varepsilon}_{495} + \widehat{\theta}_3 \widehat{\varepsilon}_{494} \\
\widehat{y}_{498} &= \widehat{\alpha} + \widehat{\phi}_1 \widehat{y}_{497} + \widehat{\phi}_2 \widehat{y}_{496} + \widehat{\phi}_3 y_{495} + 0 + 0 + \widehat{\theta}_3 \widehat{\varepsilon}_{495} \\
\widehat{y}_{499} &= \widehat{\alpha} + \widehat{\phi}_1 \widehat{y}_{498} + \widehat{\phi}_2 \widehat{y}_{497} + \widehat{\phi}_3 \widehat{y}_{496} + 0 + 0 + 0 \\
\widehat{y}_{500} &= \widehat{\alpha} + \widehat{\phi}_1 \widehat{y}_{499} + \widehat{\phi}_2 \widehat{y}_{498} + \widehat{\phi}_3 \widehat{y}_{497} + 0 + 0 + 0,
\end{aligned}$$

ahol $\widehat{\varepsilon}$ a reziduumok vektora és $\widehat{\gamma}$ a becslések vektora:

$$\widehat{\gamma} = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\phi}_1 \\ \widehat{\phi}_2 \\ \widehat{\phi}_3 \\ \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ \widehat{\theta}_3 \end{pmatrix} = \begin{pmatrix} -0.00338 \\ 2.35041 \\ -1.90505 \\ 0.53380 \\ 0.08528 \\ 0.09325 \\ -0.63081 \end{pmatrix}.$$

A megfigyelések és a becslések alapján (Eviews-ból)

n :	y :	$\widehat{\varepsilon}$:
493:	8.0057	-0.66926
494:	3.19463	0.22996
495:	-0.81385	-0.26080

A kapott előrejelzések és valódi értékek:

$$\begin{array}{l} \widehat{y}_{496} = -3.3074 \\ \widehat{y}_{497} = -4.6908 \\ \widehat{y}_{498} = -4.9978 \\ \widehat{y}_{499} = -4.5795 \\ \widehat{y}_{500} = -3.7500 \end{array} \qquad \begin{pmatrix} y_{496} \\ y_{497} \\ y_{498} \\ y_{499} \\ y_{500} \end{pmatrix} = \begin{pmatrix} -2.04982 \\ -2.55888 \\ -2.21506 \\ -1.27810 \\ 0.80268 \end{pmatrix}.$$

Az előrejelzés átlagos négyzetes eltérése:

$$MSPE = \frac{1}{5}\sum_{h=1}^{5}\left(y_{495+h} - \widehat{y}_{495+h}\right)^2 = 9.0992.$$

ARMA$(2,2)$:
A többlépéses előrejelzések:

$$\begin{aligned} \widehat{y}_{496} &= \widehat{\alpha} + \widehat{\phi}_1 y_{495} + \widehat{\phi}_2 y_{494} + \widehat{\theta}_1 \widehat{\varepsilon}_{495} + \widehat{\theta}_2 \widehat{\varepsilon}_{494} \\ \widehat{y}_{497} &= \widehat{\alpha} + \widehat{\phi}_1 \widehat{y}_{496} + \widehat{\phi}_2 y_{495} + 0 + \widehat{\theta}_2 \widehat{\varepsilon}_{495} \\ \widehat{y}_{498} &= \widehat{\alpha} + \widehat{\phi}_1 \widehat{y}_{497} + \widehat{\phi}_2 \widehat{y}_{496} + 0 + 0 \\ \widehat{y}_{499} &= \widehat{\alpha} + \widehat{\phi}_1 \widehat{y}_{498} + \widehat{\phi}_2 \widehat{y}_{497} + 0 + 0 \\ \widehat{y}_{500} &= \widehat{\alpha} + \widehat{\phi}_1 \widehat{y}_{499} + \widehat{\phi}_2 \widehat{y}_{498} + 0 + 0, \end{aligned}$$

ahol ebben az esetben

$$\widehat{\gamma} = \begin{pmatrix} \widehat{\alpha} \\ \widehat{\phi}_1 \\ \widehat{\phi}_2 \\ \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} = \begin{pmatrix} -0.00253 \\ 1.54832 \\ -0.65511 \\ 0.88059 \\ 0.78901 \end{pmatrix}.$$

A kapott előrejelzések:

$$\begin{array}{l} \widehat{y}_{496} = -3.3502 \\ \widehat{y}_{497} = -4.8351 \\ \widehat{y}_{498} = -5.2941 \\ \widehat{y}_{499} = -5.0320 \\ \widehat{y}_{500} = -4.3255. \end{array}$$

n :	y :	$\widehat{\varepsilon}$:
494:	3.19463	0.25929
495:	-0.81385	-0.22632

Az előrejelzés átlagos négyzetes eltérése:

$$MSPE = \frac{1}{5}\sum_{h=1}^{5}\left(y_{495+h} - \widehat{y}_{495+h}\right)^2 = 11.349 > 9.0992,$$

tehát az ARMA$(2,2)$ modell előrejelzési teljesítménye gyengébb mint az ARMA$(3,3)$ modellé.

Összefoglalás

A diagnosztikai vizsgálat alapján az ARMA$(2,2)$ és az ARMA$(3,3)$ modellek helyesek. A két modell közül az utóbbinak jobb az előrejelzési teljesítménye, de az előbbi kevesebb paramétert tartalmaz. Mivel az ARMA$(3,3)$ modell paramétereinek a száma nem túlságosan nagy, ezért ezt a modellt választjuk előrejelzések kiszámításához.

6.7. Gyakorlatok

Egy y_t stacionárius idősor 400 megfigyelésre meghatározott autokorrelációit (SACF) és parciális autokorrelációit (SPACF) a következő táblázat tartalmazza:

Lag	Autocorrelation (SACF)	Partial autocorrelation (SPACF)
1	.834	.834
2	.736	.134
3	.586	-.192
4	.493	.051
5	.405	.024
6	.336	-.023
7	.271	-.017
8	.202	-.057
9	.144	-.020
10	.081	-.047
11	.033	-.021
12	.007	.048
13	.011	.080
14	.001	-.045
15	.008	.016

1. Teszteljük az idősor autokorrelációját. Magyarázzuk el a tesztet.

2. Teszteljük azt, hogy az idősor AR(2).
3. Teszteljük azt, hogy az idősor MA(1), majd azt, hogy MA(2).
4. Az autokorrelációk és a parciális autokorrelációk alapján határozzuk meg az idősor modelljének típusát.

Irodalomjegyzék

[1] Damodar G. (2014), *Econometrics by Example*, Palgrave Macmillan.

[2] Duguleană L., Duguleană C. (1998), *Economie aplicată*, Editura Universităţii Transilvania, Braşov.

[3] Greene, W.H. (2017), *Econometric Analysis,* Pearson Education Limited, 8. kiadás.

[4] Heij C., de Boer P., Franses P.H., Kloek T., van Dijk H. (2004), *Econometric Methods with Applications in Business and Economics*, Oxford University Press.

[5] Kőrösi G., Mátyás L., Székely I. (1990), *Gyakorlati ökonometria*, KJK kiadó, Budapest.

[6] Ramanathan, R. (2003), *Bevezetés az ökonometriába*, Panem könyvkiadó, Budapest.