Ökonometria

Ferenci Tamás, tamas.ferenci@medstat.hu

2020. január 30.

Tartalom

Előszó

Az ökonometriai a társadalmi-gazdasági jelenségek számszerűsített, empirikus – azaz az tapasztalati, tényadatokon alapuló – vizsgálatának, modellezésének a tudománya. Szemben azzal, amit esetleg a név sugallhat, közel sem csak közgazdászoknak fontos: szociológusok, társadalomkutatók számára ugyanúgy alapvető az ökonometria ismerete. Sőt, maguk a módszerei még ennél is szélesebb körben, ahol empirikus adatok kezelésére szükség van, biostatisztikától a pszichometrián át az agrometriáig, felhasználhatóak. (Ahogy szokták is mondani: a „statisztika egy".)

Az ökonometriai a társadalmi-gazdasági jelenségek számszerűsített, empirikus – azaz tapasztalati, tényadatokon alapuló – vizsgálatának, modellezésének a tudománya

Ez a jegyzet ezeket a módszereket tárgyalja, az alapoktól kezdve. Nem célja mély matematikai részletek tárgyalása (noha a korszerű ökonometriára ez nagyon is jellemző), inkább a módszerek, alkalmazási területek, és eszközök sokféleségét kívánja bemutatni. Az elméleti szigorból azonban nem enged.

Tárgyalás matematikai részletek nélkül, inkább sok területet érintve (de elméletileg precízen)

Manapság ökonometria elképzelhetetlen számítógépes támogatás nélkül. Bár a jegyzetnek nem kifejezett célja ennek megtanítása, de igyekszik hozzá minden segítséget megadni: valamennyi eredmény előállítását is bemutatja a manapság egyre népszerűbb R statisztikai programcsomag alatt. (Az R általános statisztikai programcsomag, és bár klasszikus orientációja nem kimondottan ökonometriai, erre a célra is egyre jobban használható, kitűnő általános tulajdonságai és a megjelenő kiegészítő csomagok sokaságának köszönhetően.)

Számítógépes munka bemutatásához R statisztikai környezet alatti illusztrációk

A jegyzettel kapcsolatban minden visszajelzést, véleményt, kritikát a lehető legnagyobb örömmel veszek a tamas.ferenci@medstat.hu email-címen!

Minden visszajelzést örömmel veszek a tamas.ferenci@medstat.hu email-címen

A jegyzet weboldala (oktatási segédanyagokkal, technikai információkkal) a https://github.com/tamas-ferenci/FerenciTamas_Okonometria címen érhető el.

1. fejezet

Út az ökonometriához

Egy ilyen tudomány esetén az első feladat annak tisztázása, hogy egyáltalán mi az az ökonometria, mire szolgál, és mi szükség van rá a társadalmi-gazdasági jelenségek elemzése során. A kérdést két történetszálon fogjuk végigvezetni (persze mint minden valamirevaló kortás norvég regényben, a szálak végül össze fognak érni).

1.1. Történetünk első szála: néhány motiváló példa

Elsőként, ahelyett, hogy rögtön a definíciókra térnénk, talán érdemesebb pár példát átbeszélni, melyek már mutatni fogják az ökonometriai vizsgálatok fő problémáit.

1.1.1. Hogyan hat az osztálylétszám a tanulók teljesítményére?

Kalifornia, 1999: 420 iskolai körzet adatait gyűjtik be

A jobb tanár:diák arányú (kisebb létszámú) osztályokban jobb a teljesítmény

A közoktatásokkal kapcsolatos vizsgálatok egyik klasszikus kérdése, hogy az osztálylétszám hogyan hat a tanulók teljesítményére. Sokan amellett érvelnek, hogy a kisebb létszámú osztályokban több tanári figyelem jut egy diákra, így a tanulók teljesítménye jobb lesz. De vajon tényleg így van?

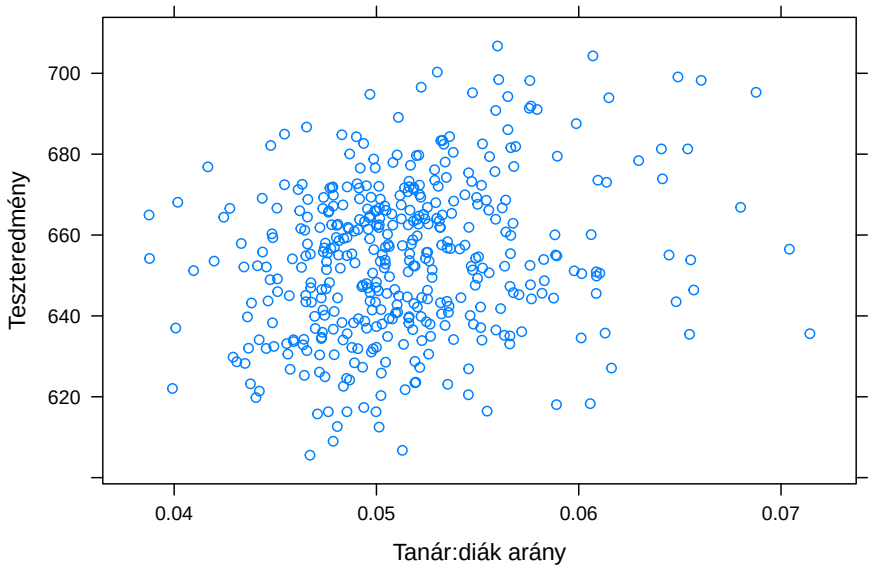
E kérdésre számos módon válaszolhatunk: felállíthatunk elméleti modelleket, papíron és ceruzával, készíthetünk interjúkat szakértőkkel, vizsgálhatunk analóg helyzeteket más területekről stb.

Mi azonban a továbbiakban egy módszerrel fogunk foglalkozni: ha empirikusan igyekszünk válaszolni a kérdésre. Empirikusan, annyi mint a tapasztalatok alapján, tehát való életbeli tényadatok begyűjtével. Elvégre az osztálylétszámokra csak van valamilyen adatgyűjtés, ha az országban futnak standardizált képességmérő felmérő-programok, akkor a tanulók teljesítményére is van adatunk – mi lenne, ha

1999-ben Kalifornia állam pontosan ezzel a kérdéssel szembesült. 420 iskolai körzetből gyűjtött adatokat, melyek – számos egyéb mellett – tartalmazták a tanulók és tanárok

létszámát, valamint az elért teszteredményeket[1]. Az `AER` csomag `CASchools` néven tartalmazza a tényleges adatokat. Lássuk is akkor az eredményt! Íme a tanár:diák arányok és a teszteredmények szóródási diagramok szemléltetve:

```
data("CASchools", package = "AER")
CASchools$tsratio <- with(CASchools, teachers/students)
CASchools$score <- with(CASchools, (math + read)/2)
lattice::xyplot( score ~ tsratio, data = CASchools,
                 xlab = "Tanár:diák arány", ylab = "Teszteredmény" )
```



Az eredmények első ránézésre megerősítik a sejtésünket: ha több tanár jut egy diákra (kisebbek az osztályok), akkor az jobb teszteredménnyel jár együtt. Nem túl erős az összefüggés, de azért egyértelműen létezik (vájtfülűek kedvéért: $r = 0.23$, $p < 0.001$) és némi munkával még az is kihozható, hogy ezen eredmények szerint ha egy századdal megnöveljük a tanár:diák arányt, akkor várhatóan 8.38 ponttal fog javulni a tesztpontszám.

Ezek az eredmények húsbavágóak. Ha lecsökkentjük az osztálylétszámokat, akkor több tanárt kell alkalmazni, több osztályt kell indítani, adott esetben több osztályteremre lesz szükség, de még az sem kizárt, hogy új iskolaépületre. Pontosan tudni kell tehát, hogy tényleg elérünk-e ezzel valamit. Sőt, valójában ennél többről van szó, azt is tudni kell, hogy *mennyit* érünk el ezzel, egész egyszerűen azért, hogy költség-haszon mérleget lehessen csinálni: a tanárok bérét meg az iskolafelújítások árát megmondják a kontrollerek, de mit rakunk a mérleg másik serpenyőjébe? Ehhez kell tudni a fenti

[1] Az adatok igazából nem osztály-szintűek, hanem körzetenkénti átlagok, de ez minket, a mostani kérdésünk szempontjából nem érint, így a továbbiakban az egyszerűség kedvéért osztályt fogok mondani.

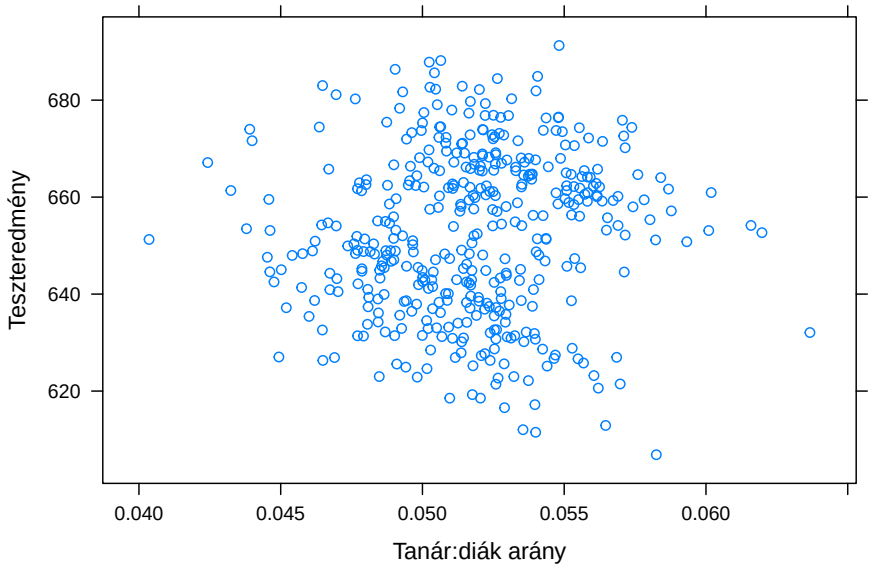
számot, hogy a kettőből együtt meg tudjuk mondani: hány millió dollárt kell költenünk 1 pont teljesítményjavításra. Hogy aztán ez megéri-e, az természetesen már nem statisztikai kérdés, függ a rendelkezésre álló büdzsétől, az egyéb feladatoktól, de még a kormány értékválasztásától is, ám a statisztikának kell ezt, mint inputadatot szolgáltatni a döntéshez.

De vajon biztos jól van így minden? Ha az ember elkezdi jobban nézni a problémát, esetleg „szociológiai" szemmel is igyekszik ránézni, akkor hamar szöget üthet a fejében valami. És nem is kell Kaliforniáig menni, magyar, vagy akár még konkrétabban budapesti viszonylatban is ugyanúgy érzékelhető a probléma: ha veszünk kis átlaglétszámú osztályokat (sztereotipikusan mondjuk 2. kerület) és nagy átlaglétszámú osztályokat (sztereotipikusan mondjuk 8. kerület), akkor hihető, hogy az előbbiek teljesítménye jobb, na de álljunk meg egy pillanatra! Csak és kizárólag az osztálylétszám nagyságában térnek el ezek az osztályok egymástól?! Dehogy! Akkor meg honnan tudjuk, hogy a tapasztalt különbség tényleg az osztálylétszámbeli eltérés miatt van…?

A rövid válasz – sajnos – az, hogy sehonnan! És itt van a bökkenő: igaz, hogy a 2. kerületi osztályok kisebbek mint a 8. kerületiek, de *egyúttal* másban is eltérnek, az oda járó gyerekek szocioökonómiai háttere tendenciájában jobb, tanulásra motiválóbb otthoni környezetből érkeztek, a szülők anyagilag is megengedhetik maguknak, hogy a gyermekeiket különórára járassák stb. E ponton viszont nagy baj van: innen kezdve fogalmunk sem lehet, hogy a tapasztalt különbség *tényleg* a kisebb osztálylétszám miatt van, vagy esetleg az osztálylétszámnak a világon semmi hatása nincs, csak egyszerűen a kisebb osztályokba jobb szocioökonómiai helyzetű diákok járnak és *ez* a valódi oka az ott tapasztalt jobb tesztpontszámnak?

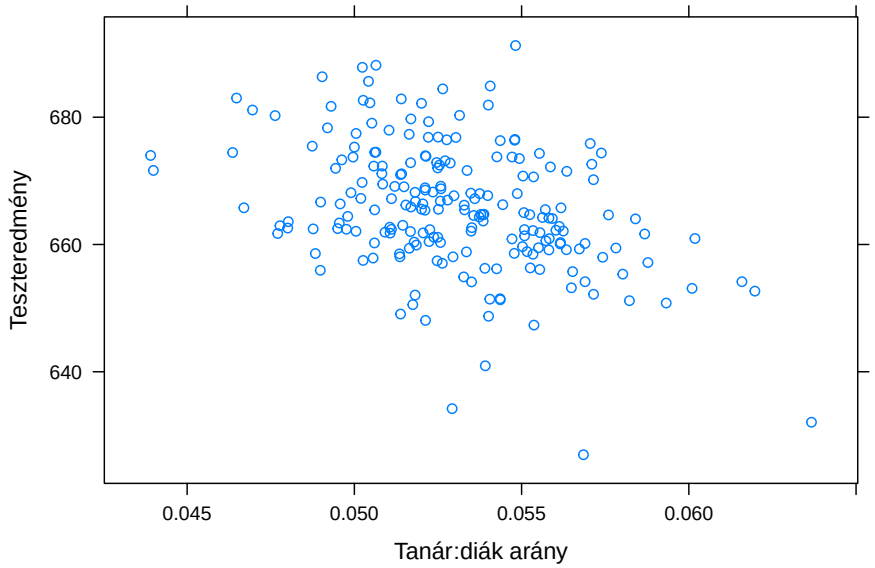
Sőt! Innentől kezdve még akár az is elképzelhető, hogy a kisebb osztálylétszám igazából kifejezetten *ront* a teljesítményen *önmagában*, csak épp a kisebb osztályokba annyival jobb szociális helyzetű diákok járnak, hogy az átfordítja a helyzetet.

Valaki nem hiszi el, hogy ez még is lehetséges? Nos, gyártsunk egy egyszerű szimulációt! Egyelőre nem árulom el, hogy hogyan készítettem, de íme a végeredménye:

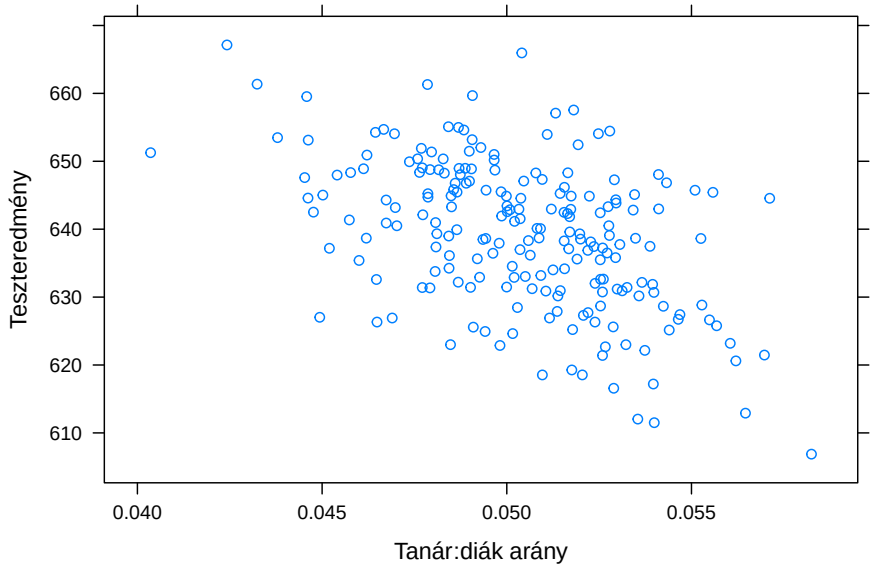


Első ránézésre nagyjából megfelel a korábbi képnek. De nézzük csak meg jobban mi történik itt! (Mivel ezt saját kezűleg generáltuk, így megtehetjük, hiszen tudjuk mi van a valóságban, az adatok hátterében). Az egyszerűség kedvéért mondjuk, hogy a szocioökonómiai státusz egy bináris változó, „jó” és „rossz” a két lehetséges értéke.

Nézzük a tanár:diák arány és a pontszám összefüggését a jó szocioökonómiai státuszú osztályok körében:

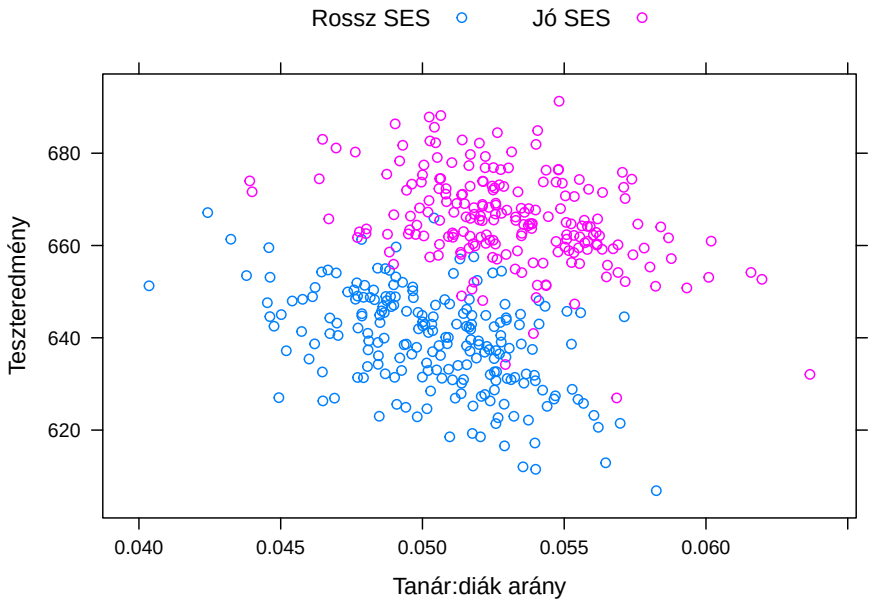


Érdekes! Itt fordított a kapcsolat. Na de mi a helyzet a rossz szocioökonómiai státuszú osztályokban? Íme:



Itt is negatív a kapcsolat!

Ez meg hogy a viharban lehet? – kérdezhetné valaki. A rossz szociális helyzetű csoporton belül is ront a kisebb osztálylétszám, a jó helyzetűeken belül is ront, de összességében meg javít?! Egyből világosabb a helyzet, ha egy ábrán ábrázoljuk a kettőt, csak eltérő színekkel:



Azonnal érthető, hogy mi történik, ha hozzávesszük a korábban mondottakat: a jobb tanár:diák arány igazából *ront* a helyzeten, de a jobb szociális helyzetű diákok által alkotott osztályok egyszerre kisebbek *és* – szociális helyzetük, nem az osztálylétszám miatt! – jobb teljesítményűek, és ez olyan erős effektus, hogy ha egyben vizsgáljuk az osztályokat, akkor a kisebb létszám rontó hatását átbillenti, hogy a kisebb osztályok a jobb szociális helyzetük miatt jobb teljesítményűek. Így összeségében azt fogjuk látni, ahonnan indultunk: hogy a kisebb létszámú osztályok jobb teljesítményűek.

A végeredmény tehát: a kisebb osztálylétszám rontja a teljesítményt *és* a kisebb létszámú osztályok jobb teljesítményűek. És ha valaki érti a fentieket, akkor azt is érti, hogy ebben a mondatban miért nincs semmi ellentmondás!

1.1.2. Csökkenti-e a korrupció mértékét a nők részvétele a politikában?

TODO

1.1.3. Csalnak-e az orosz választásokon?

TODO

1.1.4. További példák

Hosszasan sorolhatóak a további, hasonló példák a társadalmi-gazdasági elemzések világából. Kommentár nélkül még néhány kérdés, érdemes mindegyiket a fenti példákból leszűrődni kezdődő tanulságok szemüvegén keresztül végiggondolni:

- Hogyan hat a munkanélküliség a GDP-re?
- Hogyan hat az államadósság a növekedésre?
- Return on education: mekkora az oktatás haszna, tehát, ha egy évvel többet tölt valaki az iskolapadban, az mennyivel növeli a fizetését?
- Létezik-e „cigánybűnözés”?
- Az ökonometria-előadás haszna: ha többet tölt a hallgató az öko előadáson, jobb jegyet kap-e emiatt, és ha igen, mennyivel?
- Milyen tényezők hatnak arra, hogy egy országban hány terrortámadás történik?
- Hogyan hat a rendőri erők létszáma egy adott városban az ottani bűnözési rátákra?
- Cégeknek adott továbbképzési támogatás hogyan hat a termelékenységre?

(Igen, ezekre mind válaszolhatunk ökonometriai módszerekkel!)

1.2. A példák tanulságai: az empirikus adatok elemzésének legnagyobb problémája

A mintázat most már látszik. Valamilyen tényező hatására vagyunk kíváncsiak, az okozati hatására, szép szóval: a **kauzalitásra**, ez mindegyik példa lelke.

Valamilyen ok-okozati hatásra vagyunk kíváncsiak; a **kauzalitás** érdekel minket

Úgy döntünk, hogy a kérdést **empirikus** adatok elemzésével igyekszünk megoldani, azaz való életbeli tényadatokat gyűjtünk be. (A vizsgált tényezőről, a hatásról, esetleg egyéb fontos változókról.)

Számos vizsgálati módszer közül most az **empirikus** adatok elemzésével fogunk foglalkozni: tényadatokat gyűjtünk be, és ebből igyekszünk következtetni

Azért, hogy eldöntsük, hogy van-e okozati hatás (illetve, hogy lemérjük mekkora), csoportokat hasonlítunk össze, melyek eltérnek a vizsgált tényezőben. Csak épp közben a vizsgált tényezőbeli eltéréssel *automatikusan együtt járnak* egyéb tényezőbeli eltérések, és innentől kezdve bajban vagyunk, mert ha találunk is különbséget a csoportok között, nem fogjuk tudni, hogy az mi miatt van: a vizsgált tényezőbeli eltérés miatt, a vele együtt járó egyéb eltérések miatt, vagy esetleg ezek valamilyen keveréke miatt. Ezt a jelenséget, mely az empirikus adatok elemzésének legnagyobb problémája, hívják **confoundingnak**. (Angolul nagyon jó szó, amire nem sikerült hasonlóan találó magyar fordítást bevezetni. A confounding azt jelenti, hogy összemosódás, és csakugyan, az a probléma, hogy a vizsgált tényezőbeli eltérés összemosódik egyéb tényezőkbeli eltérésekkel.)

Vegyük észre, hogy a confounding fellépéséhez az kellett, hogy létezzen olyan tényező, amire két dolog *egyszerre* igaz: együttmozog a vizsgált tényezővel (összefügg vele) *és* önmagában – azaz a vizsgált tényező minden értéke mellett – hat az eredményváltozóra. Akkor van confounding, ha ez a kettő egyidejűleg fennáll. Ha bármelyik nincs jelen, akkor nincs probléma. Ha a szociális helyzet összefügg ugyan az osztálylétszámmal, de nem hat a teljesítményre, akkor nincs baj: igaz, hogy a kisebb osztályok jobb szociális helyzetűek, de ez nem befolyásolja a teljesítményt. Hasonlóképp, ha a szociális helyzet

befolyásolja ugyan a teljesítményt, de nem függ össze az osztálymérettel, akkor sincs gond: a kisebb osztályoknak nem tér el a szociális helyzete a nagyobbaktól. (Gondoljuk végig az összes többi példára is!) Azokat a változókat, amelyek ezt a két dolgot egyidejűleg tudják, tehát a vizsgált tényezővel együttmozognak *és* az eredményváltozóra is hatnak, ilyen módon okozzák a confoundingot, szokás zavaró változónak, vagy **confoundernek** nevezni.

A későbbiek szempontjából hasznos lesz ezt még másképp átfogalmazni. A probléma, hogy nem az érdekel minket, hogy egy osztály abban tér el, hogy kisebb a létszám, akkor ott jobb-e a teljesítmény, hanem az, hogy ha *csak* abban tér el, hogy kisebb a létszám, akkor ott jobb-e a teljesítmény. Ezt szokás **ceteris paribus elvnek** („minden mást változatlanul tartva") nevezni, ez a kulcs a kauzalitáshoz: az érdekel minket, hogy ha minden mást változatlanul tartva *csak* az osztálylétszám változik, akkor mi történik. A naiv elemzésben az osztálylétszám változásával együtt egyéb tényezők is változhatnak, így ebből nem tudunk a kauzalitásra következtetni. Figyeljünk a szóhasználatra: azt mondhatjuk, hogy a kisebb osztálylétszám jobb teljesítménnyel *jár együtt* (korreláció), de azt nem, hogy a kisebb osztálylétszám jobb teljesítményt *okoz* (kauzalitás).

Valaki esetleg azt mondhatja, hogy rendben, itt tényleg van valami módszertani gubanc, meg szép latin szavak[2] de igazából ez csak az ilyen módszertani kérdéseken szöszmötölő kutatóknak érdekes, a lényeg, hogy ha kisebb az osztálylétszám, akkor ott jobb a teljesítmény, ennyi a fontos, és pont. Nem! Ez az érvelés teljesen fals, az hogy mi hat mire, nem tudományos szőrszálhasogatás, hanem elsőrendű gyakorlati kérdés. Miért? A beavatkozás miatt! A *valódi* okozatiság felismerése ott válik kritikussá, ha beavatkozunk a rendszerbe, ha ugyanis rosszul állapítjuk meg az okozati kapcsolatok irányát, akkor ez teljesen félremehet. Például lecsökkentjük az osztálylétszámokat, adott esetben rengeteg pénzt elköltve, de ha a *valódi* oka a jobb teljesítménynek nem az osztálylétszám, hanem a jobb szociális helyzet, akkor ezzel semmit nem érünk el! Sőt, mint a későbbi példa mutatja, adott esetben még kimondottan árthatunk is!

Végezetül még egy megjegyzés. A confounding felismerése nem azt jelenti, hogy akkor igazából nincs hatás, végképp nem azt, hogy bizonyítottuk, hogy ellentétes irányú hatás van. Pusztán annyit jelent, hogy a confounding-gal terhelt adatok *nagyon gyenge* bizonyítékot jelentenek a hatás léte mellett. De ettől még lehet éppenséggel hatás! – csak az ilyen adatok nagyon kevéssé támasztják ezt alá.

1.3. A confounding megoldásai: kísérlet és megfigyelés

Most, hogy alaposan kivesézték a confounding problémáját, természetesen adódik a kérdés: na de mit tehetünk ez ellen? Azt könnyű lenne biztosítani, hogy a csoportok egy-két általunk megadott szempont szerint ne legyenek eltérőek, de azt, hogy *egyáltalán semmilyen* szempont szerint ne térjenek el (kivéve persze az vizsgált tényezőt), olyanok szerint sem, amikről eszünkbe sem jut, hogy lehet bennük eltérés, csak egy módon lehet: ez a **randomizálás**. Ahogy a szó is sugallja, a randomizálás lényege, hogy a megfigyelési egységeket *véletlenszerűen* sorsoljuk különböző csoportokba, majd

[2] Pedig még össze sem foglaltam a fentieket úgy, hogy a korreláció nem implikál kauzalitást!

ezeket a csoportokat tesszük ki a vizsgált tényezőnek. Például pénzfeldobással döntjük el az óvoda végén minden egyes gyermekről, hogy kis vagy nagy létszámú osztályba kerüljön. Ez azért jó, mert ilyen módon a két csoport között nem lesz szisztematikus különbség szocioökonómiai státuszban, de ami még fontosabb: *semmilyen* tényezőben nem lesz szisztematikus különbség, a kék szeműek vagy a balkezesek számában sem, hiszen a pénzfeldobás nyilván ezekre is érzéketlen. Ilyen módon a csoportok összehasonlíthatóak: ha találunk köztük különbséget a tanulmányi eredményben, az *tényleg* az osztálylétszámnak lesz betudható… hiszen másban nincs szisztematikus különbség.

A randomizálásnak egy baja van: akkor alkalmazható, ha a vizsgált tényezőt tudjuk irányítani. (Hiszen nekünk kell az egyik csoportba sorsolt gyerekeket kis, a másikat nagy létszámú osztályba helyezni.) Azokat a kutatásokat, ahol a kutatást végzők tudják irányítani a vizsgált tényezőt, **kísérletes (experimentális) kutatásnak** nevezzük. És itt értünk el a bökkenőhöz: a társadalmi-gazdasági jelenségek vizsgálata az a terület, ahol tipikusan *nem* lehet kísérletet végezni. Aligha lehet gyerekeket pénzfeldobással sorsolni osztályokba, vagy országokban pénzfeldobással meghatározni, hogy mennyi nő üljön a kormányban…

(Persze ez sincs kőbe vésve. Néha lehet kísérletet csinálni, ahogy a választási megfigyelők példája is mutatja. Másik oldalról, például az orvostudományban sokszor lehet kísérletet csinálni, ez a jellemző új gyógyszerek bevezetésénél, ahol a vizsgálat során véletlenszerűen kiválasztott alanyok kapnak gyógyszert, míg a többiek placebot, de ott is van olyan kérdés, ahol nem lehet kísérletet csinálni! Tipikusan ilyenek fordulnak elő az epidemiológiában: a vörös hús rákkeltő? Aligha lehet emberekkel pénzfeldobás alapján évtizedekig több vagy kevesebb vörös húst etetni… Innentől a probléma ott is ugyanaz: ha a vörös húst evők között több a rákos, az nagyon gyenge bizonyíték, mert a több vörös húst fogyasztó emberek milliónyi *egyéb* dologban is eltérnek a kevesebb vörös húst fogyasztó emberektől a vörös hús fogyasztás mértékén túl – és mi van, ha ezek közül valami növeli a rákkockázatot…?)

Azokat a vizsgálatokat, ahol a kutatást végzők nem tudják befolyásolni a vizsgált tényezőt, az alakul a maga rendje szerint, és a kutatók csak passzíve feljegyzik a történéseket külső szemlélőként, **megfigyeléses (obszervációs) vizsgálatnak** nevezzük. A társadalmi-gazdasági elemzések során tehát szinte mindig ilyenekkel lesz dolgunk. Márpedig ezeknél mindig fejünk felett fog lebegni a confounding problémája.

1.4. Történetünk második szála: az ökonometriai modellek és a regresszió

Folytassuk most valami – látszólag – teljesen más témával.

Minden fenti példában volt egy változó, mely az eredménye volt a vizsgálatunknak, a kimenet szerepét játszotta, tehát aminek az alakulását le kívántuk írni (tesztpontszám, korrupció mértéke, szavazati arány stb.). A továbbiakban ezt **eredményváltozónak** (vagy függő változónak, angolul response) fogjuk hívni, jele Y. Az első példánkban Y = Teszteredmény. Másrészről voltak változók, adott esetben nem is egy, amikkel

le akarjuk írni az eredményváltozó alakulását, amelyekről azt mondjuk, hogy hatnak, vagy hathatnak az eredményváltozóra; ezek neve **magyarázó változó** (vagy független változó, angolul predictor). Ezekből több is lehet (az első példában ilyen az osztálylétszám és a szocioökonómiai helyzet), jelöljük számukat k-val, és az egyes változókat X_i-vel ($i = 1, 2, \ldots, k$). Az első példában $k = 2$ és $X_1 =$ Tanár:diák arány, $X_2 =$ Szocioökonómiai státusz. Összefoglalva, az eredményváltozó a vizsgált kimenet, a magyarázó változók az azt – potenciálisan – befolyásoló tényezők (tehát a fontos, vizsgált változók és a – potenciális – confounderek egyaránt).

Az X-ek hatnak az Y-ra, vagy fordítva megfogalmazva, az Y függ az X-ektől – ragadjuk meg most ezt matematikailag. Szerencsére arra, hogy egy változó függ más változóktól, ismerünk egy jó matematikai objektumot, ez a függvény fogalma:

$$Y = f\left(X_1, X_2, \ldots, X_k\right)$$

A későbbiekben erre azt fogjuk mondani, hogy ez egy statisztikai modell. Ennek az általánosságával nehéz lenne vitatkozni, de egy baja mégis csak van.

A fő probléma, hogy a modell azt feltételezi, hogy az Y és az X-ek kapcsolata **determinisztikus**. Szinte teljesen mindegy is, hogy mi az Y és mik az X-ek, hogy mi a vizsgált probléma, a társadalmi-gazdasági jelenségek vizsgálata kapcsán lényegében általánosan kijelenthető, hogy ez irreális: bármilyen ügyesek vagyunk, soha az életben nem fogunk tudni determinisztikus modelleket alkotni társadalmi-gazdasági jelenségekre. (Aligha lehet olyan modellt alkotni, ami *pontosan*, hiba nélkül megmondja előre, hogy egy osztály milyen pontszámot fog elérni, vagy, hogy egy választáson pontosan hány szavazat érkezik egy pártra.) Ez legfeljebb középiskolás fizikában működik, a társadalmi-gazdasági jelenségekben szinte kizárt, hogy *függvényszerű* módon meghatározzák a magyarázó változók az eredményváltozót. Hiszen lesznek változók amiket nem ismerünk, rosszul mérünk, rosszul veszünk figyelembe, az, hogy egy gyerek hány pontot ír egy teszten, mindig függ a mi közelítési szintünk ténylegesen véletlen dolgoktól stb. A A valódi modell tehát **sztochasztikus** kell legyen:

$$Y = f\left(X_1, X_2, \ldots, X_k\right) + \varepsilon$$

Itt ε jelzi a fentiekből fakadó bizonytalanságot, a neve: **hibatag**.

Rövid jelölésként az X-eket gyakran egy vektorba vonjuk össze: $Y = f\left(\underline{X}\right) + \varepsilon$.

Az így kapott modellünk már egy teljes értékű statisztikai (ökonometriai) modell! Az ilyen f-et hívjuk (sokasági) **regressziófüggvénynek**.

Már most is fontos, hogy lássuk, hogy az f-nek van egy nagyon is földhözragadt értelmezése: ezt kell használni, ha szeretném „megtippelni” Y értékét X-ek ismeretében. Hogy ez miért lesz fontos, azt majd később fogjuk látni, de a feladat így is értelmes: ha ismerem egy osztály tanár:diák arányát és a szociális helyzetet, akkor ezek alapján mit mondhatok a teszteredményéről. Ha ezek tényleg hatnak rá, akkor valamit mondhatok, ezt fejezi ki az f-es rész, ε pedig azt, amit nem tudok ezek alapján megmondani, azaz ettől lesz ez csak tipp: mindenképp kell számolnom azzal, hogy a valóság ettől a fent említett okok miatt eltér.

Ez az egyenlet egy **sokasági modell**: azt írja le, hogy a valóság hogyan működik. Pontosan ugyanaz a helyzet, mint bármilyen következtető statisztikai kurzus alapjainál: van a sokaság, amit eloszlásokkal, valószínűségszámítási eszközökkel írunk le, de a tényleges vizsgálatokban mi sem ismerjük. (Tehát nem tudjuk, hogy ezek az eloszlások milyenek.) Ahhoz, hogy megismerjük veszünk egy mintát, ennek a kezeléséhez már statisztika kell, aminek a feladat épp az lesz, hogy következtessünk a sokaságra.

Most is hasonló a helyzet: mi sem tudhatjuk, hogy milyen *eloszlása* van a teszteredményeknek, csak van egy 420 elemű mintánk rá nézve; és hasonlóan a többi változóval. Most azonban egy pillanatig leszünk valszámos emberek: ne törődjünk azzal a problémával, hogy a sokaságot igazából nem ismerhetjük, játsszuk azt, hogy ismerjük (tudjuk mik ezek az eloszlások), és vizsgáljuk meg, hogy ebből mire jutunk! Ugyanúgy mint a következtető statisztikánál, ez nagyon hasznos lesz majd később, a számunkra igazán érdekes – statisztikai – feladat megoldásánál is.

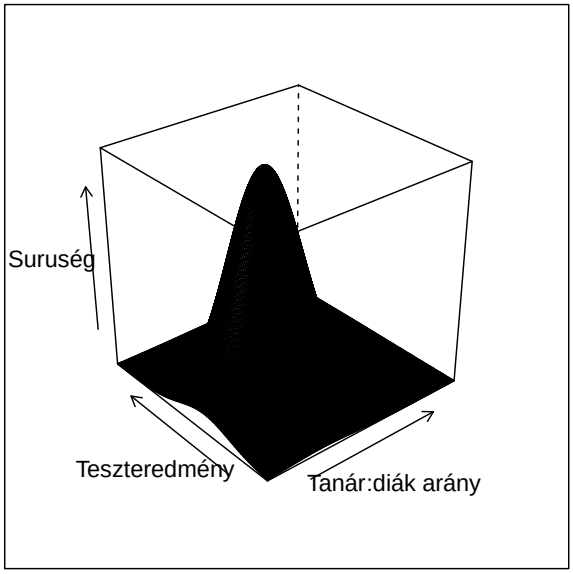
A nem-kísérleti jelleg miatt az az értelmes modell, ha mind az eredményváltozót, mind a magyarázó változókat – és így persze ε-t is – valószínűségi változónak vesszük. (Ezért használtam eddig is nagy betűket!) Bizonyos egyszerűsített tárgyalások úgy tekintik, mintha az X-ek nem valószínűségi változók lennének, hanem rögzített értékek. Ez a kísérletek világában rendben lehet, ahol mi be tudjuk *állítani* az X-ek értékét, de ökonometriában, a társadalmi-gazdasági elemzések világában még közelítő feltevésként is értelmetlen.

1.5. Regresszió a sokaságban

Elsőként tehát le kell írnunk a sokaságot: valszámos emberek leszünk, és úgy vesszük mintha ismernénk a sokaságot. Mit jelent ez, mit is ismerünk pontosan? Nem csak Y és $X_1, X_2, \ldots, X_k$ eloszlásait (külön-külön), hanem az együttes eloszlásukat is! Ekkor tudunk mindent ezekről (valószínűségszámítási értelemben).

Ezt úgy kell elképzelnünk, mint egy $k+1$ dimenziós teret: minden pont egy adott magyarázó- *és* eredményváltozó-kombináció. E fölött értelmezve van egy eloszlás, ami azt mutatja, hogy ha mintát veszünk ebből az eloszlásból, akkor milyen valószínűséggel esünk az adott pont kis környékére.

$k+1$ dimenziós terekben a legtöbb ember relatíve rosszul tájékozódik, úgyhogy ábrázoljunk egy olyan együttes eloszlást, amikor még átlátható a dolog!

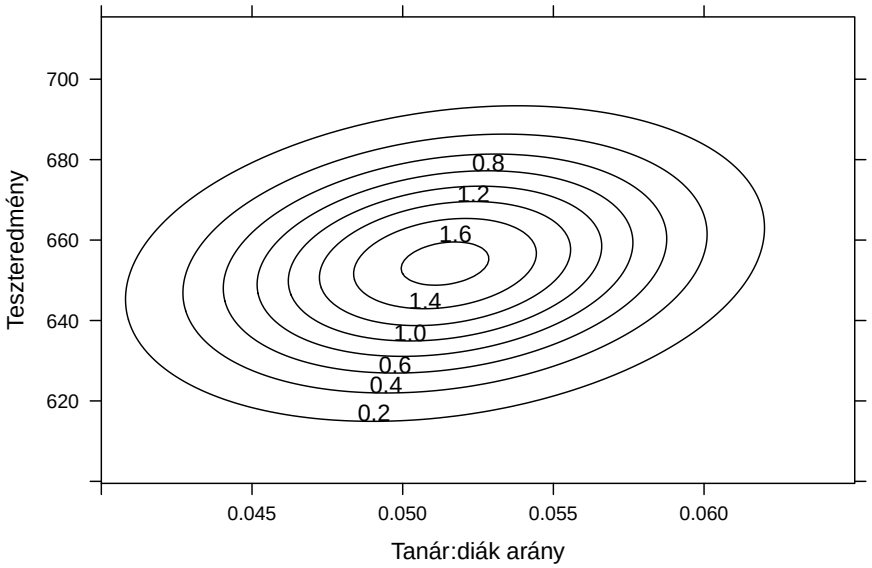


Ez egy kétváltozós eloszlás együttes sűrűségfüggvénye; itt az egyik változó játsza a magyarázó-, a másik az eredményváltozó szerepét. A mintavétel ebből az eloszlásból azt jelenti, hogy kiveszünk egy iskolát (tehát tanár:diák arányt *és* teszteredményt egyszerre!); ahol magasan fut a sűrűségfüggvény, arról a környékről gyakran veszünk ki, ahol alacsonyan, ott ritkábban. A hiba mibenléte is jól érthető erről az ábráról: ha kiválasztunk egy adott konkrét X_1-et, ahhoz csak egyetlen $f\left(X_1\right)$-et adhatunk, mégis Y minden értéket felvehet, tehát lehetetlen, hogy ne hibázzunk. (Csak egyetlen egy pont lesz a végtelen sok közül, ahol nem hibázunk.) Persze $f\left(X_1\right)$-et majd pont úgy lesz célszerű megválasztani, hogy oda rakjuk, ahol Y gyakran előfordul, hogy a gyakran előforduló esetekben hibázzunk picit, és csak a ritkábbakban nagyobbat – de erről majd kicsit később.

Elárulom, hogy ez az eloszlás többváltozós normális (később ennek majd jelentősége lesz), $\boldsymbol{\mu} = \begin{pmatrix} 654 \\ 0.0514 \end{pmatrix}$ várhatóérték-vektorral és $\mathbf{C} = \begin{pmatrix} 19{,}1^2 & 0{,}23 \cdot 19{,}1 \cdot 0{,}00515 \\ 0{,}23 \cdot 19{,}1 \cdot 0{,}00515 & 0{,}00515^2 \end{pmatrix}$ kovariancia-mátrixszal[3].

Sajnos ez az ábrázolás nehezen érzékelhető (pláne, ha nem interaktívan nézzük, és nincs módunk forgatni), jobban járunk, ha így rajzoljuk ki:

[3]Egyszerűen úgy választottam a paramétereket, hogy megfeleljen a kalifornaiai példának



Ez ugyanaz mint a fenti sűrűségfüggvény, de „szintvonalakkal” leírva (azaz különböző z magasságokban elmetszettük a sűrűségfüggvényt és a kapott metszeteket ábrázoltuk). Belátható, hogy többváltozós normális esetén ezek mindig ellipszisek[4]. Ezt az ábrázolást szokás „contour plot”-nak nevezni, előnye, hogy – a háromdimenziós érzékeltetéssel szemben – nem érzékeny a nézőpont megválasztására, részek nem takarnak ki másokat stb. (Ám cserében nyilván információ-vesztéssel jár, ami azzal arányos, hogy milyen sűrűn képezzük a metszeteket.)

Térjünk most vissza az alapkérdésünkre! Úgy vesszük, hogy ez az eloszlás adott, és le akarjuk írni mint $Y = f\left(\underline{X}\right) + \varepsilon$; de vajon mi f-re a legjobb választás? Persze egy ilyen kérdést hallva azonnal vissza kell kérdezni: mi a jóság mérőszáma? Hiszen csak ennek ismeretében mondható meg, hogy mi az optimális sokasági regressziófüggvény.

Mivel az ε hibát fejez ki, így azzal valószínűleg kevesen vitatkoznának, hogy az a legjobb f, amely mellett a hiba a legkisebb. Igen ám, de mi az, hogy a hiba a „legkisebb”? Ez nem olyan nyilvánvaló, ennek megértéséhez beszéljünk egy picit a hibáról. A helyzetet a fenti példán úgy kell elképzelnünk, hogy behúzzuk a $f\left(X_1\right)$ függvényt; ez olyan mintha rajzolnánk egy görbét az $X_1 - Y$ síkra. Ez után végigmegyünk a sík minden pontján, és megnézzük ott mekkora a hiba: mennyire van távol Y az $f\left(X_1\right)$-től; ez pedig akkora súllyal fog szerepet játszani a hiba eloszlásában, amilyen magasan fut az adott ponton a sűrűségfüggvény. Mindezek természetesen ugyanígy működnek az általános, $k+1$ dimenziós esetben is.

[4]Úgy, hogy az ellipszis középpontját a várhatóérték-vektor adja meg, a tengelyek a kovariancia-mátrix sajátvektorainak irányába mutatnak, féltengelyeik hossza pedig a kovariancia-mátrix megfelelő sajátértékeivel arányos.

A hibának tehát egy eloszlása van, így nem egyértelmű, hogy mikor a „legkisebb". Két dolgot kell tennünk, az egyik választás egyértelmű, de a másik már inkább döntés kérdése. Az első, hogy a hiba helyett annak $\mathbb{E}\varepsilon$ várható értékét tekintjük. Ez jó, mert így a valószínűségi változóból rögtön egy számot kapunk, amire pedig azonnal jobban értjük, hogy mit jelent az, hogy legyen a legkisebb. De igazán azért jó, mert ha összekombináljuk a várható érték fogalmát az előbbi bekezdés végén mondottakkal, akkor látjuk, hogy ez egy nagyon logikus dolgot mond: azt, hogy ott kevésbé számít a hibázás, ahová egyébként is ritkán esünk, és ott számít jobban a hibázás, ami gyakran előfordul!

Azonban még nem végeztük. Ha meggondoljuk, akkor rögtön látjuk, hogy $\mathbb{E}\varepsilon$ még nem lesz jó: a hiba lehet negatív is és pozitív is, de mi[5] nem mondhatjuk azt, hogy ha egyszer 10-zel fölé lőttünk, egyszer meg 10-zel alá, akkor tökéletesek voltunk. Magyarán: meg kell szabadulni az előjeltől. Itt már van választási lehetőségünk, hogy mit teszünk, most döntsünk úgy (és jelen jegyzet túlnyomó többségében ezt adottságnak fogjuk venni), hogy négyzetre emeléssel szabadulunk meg az előjeltől, hiszen a négyzetre emelés függvény tulajdonságai nagyon kellemesek.

Így tehát a megoldandó feladat:

$$\arg\min_{f} \mathbb{E}\left[Y - f\left(\underline{X}\right)\right]^2$$

Ez első ránézésre nagyon is ijesztően néz ki: optimalizációs feladat – az *összes létező* függvény terében?! Mert azt még érti az ember, hogy x felveszi az összes lehetséges valós számot, és mikor lesz $f(x)$ minimális, na de mi az, hogy valami felveszi az összes létező (k-változós) függvényt…? Hiszen semmi más megkötés nincs a világon, *akármilyen* k-változós függvény szóba jöhet, semmit nem mondtunk a függvényformáról, összeadhatjuk a változókat, összeszorozhatjuk, hatványozhatjuk, bármilyen műveletet végezhetünk, bármilyen konstanst belerakhatunk, és az összes ilyen közül mondjuk meg, hogy ez a kifejezés mikor lesz a legkisebb?!

Az érdekes az, hogy bármilyen abszurdan is néz ki, a dolognak van megoldása! Ráadásul a végeredmény nem is túl bonyolult: f legjobb megválasztása adott pontban Y feltételes várható értéke lesz az kérdéses pontban:

$$f_{\text{opt}}(\mathbf{x}) = \mathbb{E}\left(Y \mid \underline{X} = \mathbf{x}\right)$$

Bizonyítsuk is be ezt! Legyen f_{opt} a feltételes várható érték, f pedig egy tetszőleges k-változós függvényt. Alakítsuk át a kritériumfüggvényt:

$$\begin{aligned}\mathbb{E}\left[Y - f\left(\underline{X}\right)\right]^2 &= \mathbb{E}\left[Y - f_{\text{opt}}\left(\underline{X}\right) + f_{\text{opt}}\left(\underline{X}\right) - f\left(\underline{X}\right)\right]^2 = \\ &= \mathbb{E}\left[Y - f_{\text{opt}}\left(\underline{X}\right)\right]^2 + \mathbb{E}\left\{\left[Y - f_{\text{opt}}\left(\underline{X}\right)\right]\left[f_{\text{opt}}\left(\underline{X}\right) - f\left(\underline{X}\right)\right]\right\} + \\ &+ \mathbb{E}\left[f_{\text{opt}}\left(\underline{X}\right) - f\left(\underline{X}\right)\right]^2.\end{aligned}$$

[5] A statisztikusokról szóló viccekkel szemben.

A középső tag szerencsére nulla, ezt toronyszabállyal láthatjuk be:

$$\begin{aligned}
&\mathbb{E}\left\{\left[Y - f_{\text{opt}}\left(\underline{X}\right)\right]\left[f_{\text{opt}}\left(\underline{X}\right) - f\left(\underline{X}\right)\right]\right\} = \\
&= \mathbb{E}\left\{\mathbb{E}\left\{\left[Y - f_{\text{opt}}\left(\underline{X}\right)\right]\left[f_{\text{opt}}\left(\underline{X}\right) - f\left(\underline{X}\right)\right]\right\} \mid \underline{X}\right\} = \\
&= \mathbb{E}\left\{\left[f_{\text{opt}}\left(\underline{X}\right) - f_{\text{opt}}\left(\underline{X}\right)\right]\mathbb{E}\left[f_{\text{opt}}\left(\underline{X}\right) - f\left(\underline{X}\right)\right] \mid \underline{X}\right\} = 0,
\end{aligned}$$

így azt kaptuk, hogy

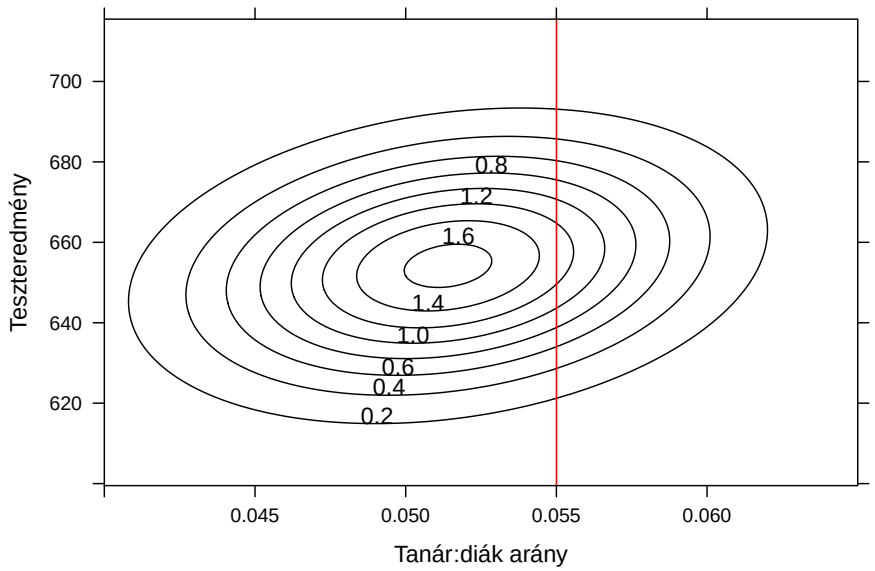
$$\mathbb{E}\left[Y - f\left(\underline{X}\right)\right]^2 = \mathbb{E}\left[Y - f_{\text{opt}}\left(\underline{X}\right)\right]^2 + \mathbb{E}\left[f_{\text{opt}}\left(\underline{X}\right) - f\left(\underline{X}\right)\right]^2,$$

amiből már csakugyan látható, hogy f_{opt} a legjobb választás, hiszen az első tagra nincsen ráhatásunk (mi ugye f-et állítjuk), a második tag pedig egy négyzet várható értéke, így 0-nál kisebb nem lehet, de az csakugyan elérhető, ha f-nek f_{opt}-ot választjuk.

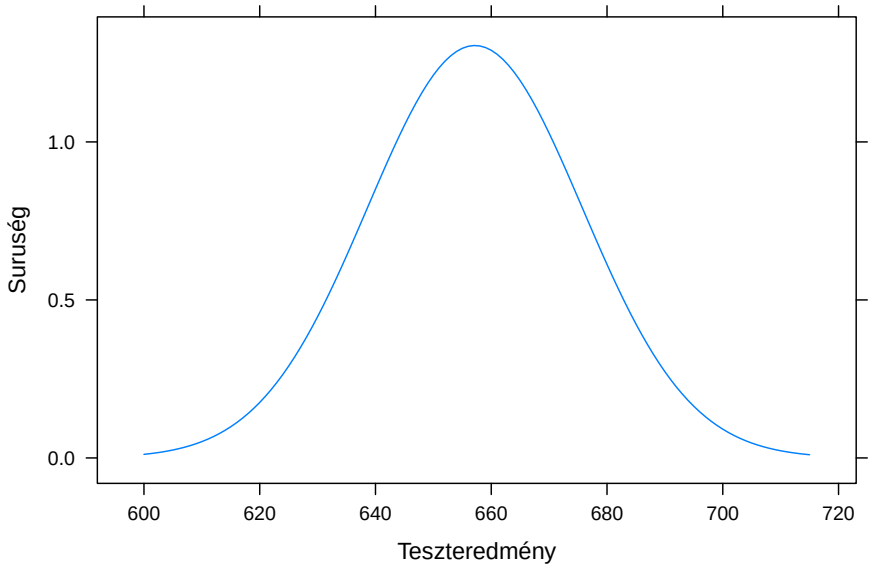
Látható tehát, hogy ez az eredmény *teljesen univerzális*, semmit nem tételeztünk fel f-ről!

Talán nem felesleges feleleveníteni ezen a pontok a feltételes várható érték fogalmát.

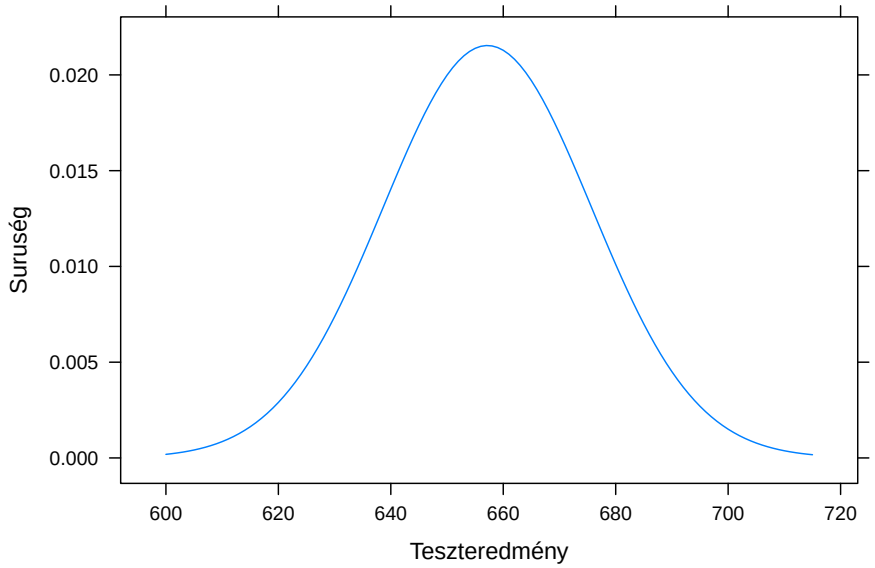
A kiindulópont a feltételes eloszlás, amit úgy kapunk, hogy fogjuk az együttes eloszlást, és egy adott ponton (ami a feltétel) átmetszük. Mondjuk legyen a feltétel az, hogy $X_1 = 0{,}055$:



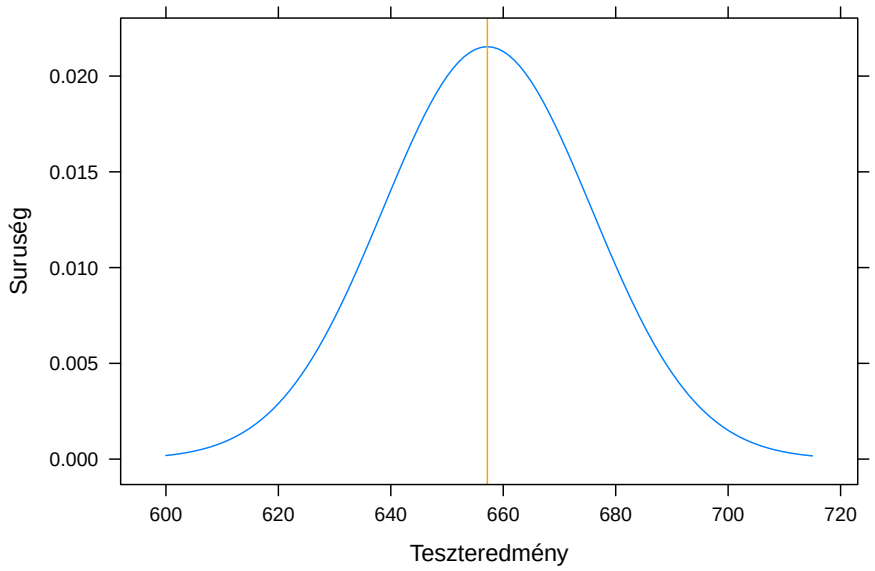
Az együttes sűrűségfüggvény, ne feledjük, egy hegy (aminek a szintvonalait mutatja az ábra), tehát arról van szó, hogy fogunk egy nagy kést, és a piros vonal mentén végigvágjuk a hegyet. Így ezt kapjuk:



Vigyázat, ez még nem sűrűségfüggvény, hiszen nem 1 a görbe alatti területe! De már majdnem megvagyunk, nincs más feladatunk, mint átnormálni (elosztani alkalmas konstanssal), hogy 1 legyen a görbe alatti terület, ez az alkalmas konstans persze a jelenlegi görbe alatti területe lesz, ami nem más, mint a vetületi eloszlás értéke a feltétel pontjában. Elvégezve ezt kapjuk a feltételes eloszlást:



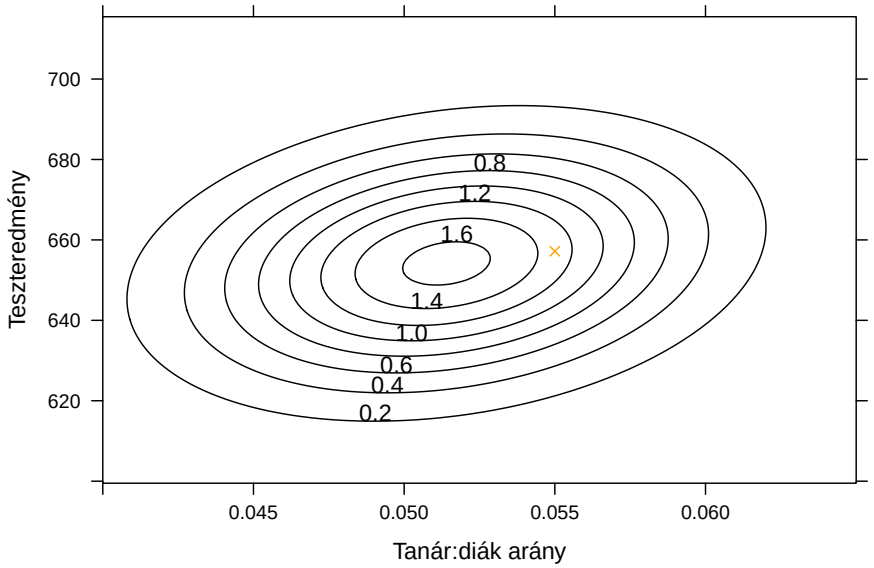
A feltételes várható érték nem más, mint a feltételes eloszlás várható érték – tehát ennek a fenti függvénynek a várható érték. Bejelölve rajta:



A feltételes várható érték tehát nagyjából 657, és ne feledjük, ez ahhoz a feltételhez tartozik, hogy a tanár:diák arány értéke 0,055. Ahogy az előbb megállapítottuk: ha

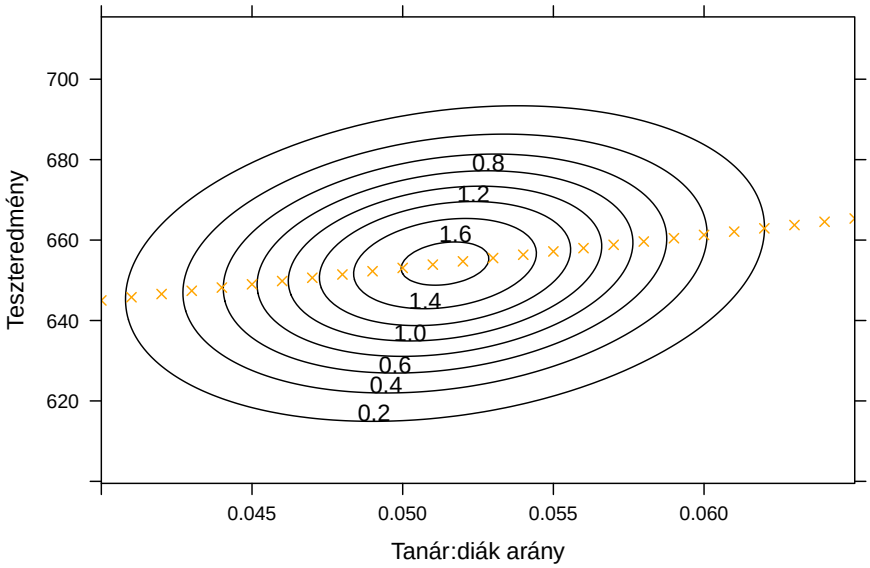
valaki azt kérdezi, hogy ekkora tanár:diák arány mellett mi a legjobb tippünk a teszteredményre, akkor válaszoljunk 657-et! Ezzel is hibázhatunk persze, de így is ekkor járunk a legjobban (elfogadva persze, hogy négyzetes hibázást minimalizálunk).

Jelöljük is be ezt az értéket az eredeti ábránkon:



Így ni: ha 0,055-ben kérdeznek meg minket, akkor ez a legjobb tippünk.

De az ember itt már vérszemet kap: vajon mi történik, ha kiszámoljuk az összes többi pontban is, hogy mi a legjobb tippünk, tehát a feltételes várhatóértéket?! Íme:



Nem lehet nem észrevenni: ezek mind egy egyenesre[6] illeszkednek! A dolog természetesen nem véletlen, és azért van így, mert az eloszlás többváltozós normális volt. Ez esetben az optimális sokasági regressziófüggvény csakugyan mindig lineáris, ez tételként kimondható[7]: ha Y és $\underline{X}$ együttes eloszlása normális[8], akkor

$$\mathbb{E}\left(Y \mid \underline{X}\right) = \mathbb{E}Y + \mathbf{C}_{Y\underline{X}}\mathbf{C}_{\underline{X}\underline{X}}^{-1}\left(\underline{X} - \mathbb{E}\underline{X}\right).$$

Egy pillanatra álljunk meg. Eddig feltételes eloszlást csak úgy írtunk, hogy a feltétel az egy konkrét érték (szám vagy vektor) volt: $\mathbb{E}\left(Y \mid \underline{X} = \mathbf{x}\right)$. De itt valami más szerepel! A magyarázathoz elevenítsünk fel egy valszám definíciót: a $\mathbb{E}\left(Y \mid \underline{X} = \mathbf{x}\right)$ egy h transzformációt definiál (hiszen adott $\mathbf{x}$-hez hozzárendel egy valós számot), és $\mathbb{E}\left(Y \mid \underline{X}\right)$ alatt $h\left(\underline{X}\right)$-et értjük. Tehát van értelme az $\mathbb{E}\left(Y \mid \underline{X}\right)$ objektumnak is, és ez egy valószínűségi változó lesz. Számunkra ebből annyi fontos, hogy ha $\mathbb{E}\left(Y \mid \underline{X}\right)$-t látunk, azt értsük úgy, mint valami, ami *minden* $\mathbf{x}$ esetén működik, bármikor beírható, hogy így $\mathbb{E}\left(Y \mid \underline{X} = \mathbf{x}\right)$ legyen belőle. Természetesen ez fontos, hogy ha egy egyenletben szerepel, akkor ezt az átírást mindenhol megtegyük, pl. írhatjuk, hogy $\mathbb{E}\left(Y \mid \underline{X} = \mathbf{x}\right) = \mathbb{E}Y + \mathbf{C}_{Y\underline{X}}\mathbf{C}_{\underline{X}\underline{X}}^{-1}\left(\underline{X} - \mathbf{x}\right)$ (hiszen $\mathbf{x}$ várható értéke saját maga). Ez a jelölés tehát egyfajta általánosítás.

[6] Érdemes megfigyelni (ez kétváltozós esetben jó szemmértékkel még érzékelhető vizuálisan is), hogy a regressziófüggvény *nem* az ellipszisek nagytengelye – tehát a korrelációs mátrix megfelelő sajátvektora – irányába mutat! Hanem az ellipszis „vízszintesen szélső" pontjain megy át.)

[7] A bizonyítást itt elhagyom, lásd például: Bolla-Krámli: Statisztikai következtetések elmélete. Typotex, 2005. 207-208.oldal.

[8] Jelölje $\mathbf{C}_{\underline{X}\underline{X}}$ az X-ek szokásos kovarianciamátrixát, $\mathbf{C}_{Y\underline{X}}$ pedig azt az oszlopvektort, amely sorban az összes X kovarianciáját tartalmazza Y-nal.

Egy tulajdonságát már ennyi alapján is rögtön láthatjuk a sokasági regressziófuggvénynek: hogy átmegy a várhatóértékek pontján. Hiszen ha a magyarázó változók értéke épp a várhatóértékük, akkor a második tag kiesik, és azt kapjuk, hogy a regressziófüggvény pont az eredményváltozó várható értékét veszi fel.

Visszatérve, ha bevezetjük a

$$\beta_0 = \mathbb{E}Y - \mathbf{C}_{Y\underline{X}}\mathbf{C}_{\underline{XX}}^{-1}\mathbb{E}\underline{X}$$

és a

$$\begin{pmatrix}\beta_1 & \beta_2 & \cdots & \beta_k\end{pmatrix}^T = \mathbf{C}_{Y\underline{X}}\mathbf{C}_{\underline{XX}}^{-1}\underline{X}$$

jelöléseket, akkor írhatjuk, hogy

$$\mathbb{E}\left(Y \mid \underline{X}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

Ebből talán még világosabban látszik a korábbi állítás: hogy többváltozós normális eloszlásnál speciálisan a regressziófüggvény lineáris lesz.

Érdemes megnézni, hogy a még áttekinthető kétváltozós (Y,$X_1 = X$) esetben ez az általános eredmény mire specializálódik: ekkor azt kapjuk, hogy $\mathbb{E}(Y \mid X) = \mathbb{E}Y + \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2X} \cdot \left(X_1 - \mathbb{E}X\right)$. Két dolgot vegyünk észre:

- Korreláció megjelenése:

$$\mathbb{E}(Y \mid X) = \mathbb{E}Y + \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2X}(X - \mathbb{E}X) = \mathbb{E}Y + \frac{\mathbb{D}Y}{\mathbb{D}X} \cdot \operatorname{corr}(X,Y) \cdot (X - \mathbb{E}X).$$

- A linearitás megjelenése itt:

$$\mathbb{E}(Y \mid X) = \mathbb{E}Y + \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2X}(X - \mathbb{E}X) = \left(\mathbb{E}Y - \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2X} \cdot \mathbb{E}X\right) + X \cdot \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2X},$$

azaz $\mathbb{E}(Y \mid X) = \beta_0 + \beta_1 X$, ha $\beta_0 = \left(\mathbb{E}Y - \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2X} \cdot \mathbb{E}X\right)$ és $\beta_1 = \frac{\operatorname{cov}(X,Y)}{\mathbb{D}^2X}$.

És most egy borzasztó fontos dolog következik. Rakjuk össze a puzzle darabjait: egyfelől tudjuk, hogy $Y = f\left(X_1, X_2, \ldots, X_k\right) + \varepsilon$, másrészt most már azt is megállapítottuk, hogy az itt szereplő f legjobb értéke $\mathbb{E}\left(Y \mid \underline{X}\right)$, és ez mindig[9] igaz. Tehát azt kaptuk, hogy:

$$Y = \mathbb{E}\left(Y \mid \underline{X}\right) + \varepsilon$$

Ez a dekompozíciót szokás a regresszió „hibaalakjának” (error form) nevezni. Lényegében arról van szó, hogy szétbontjuk az eredményváltozó alakulását egy „magyarázóváltozókkal elérhető legjobb becslés” (már láttuk: a feltételes várhatóérték) és egy „maradék hiba” részre (ami marad). A regresszióanalízis a *feltételes* eloszlásra koncentrál! Ezért elvileg olyasmit kéne írnunk, hogy „$(Y \mid X) = \mathbb{E}\left(Y \mid \underline{X}\right) + \varepsilon$”, de ezt nem tesszük (az

[9] Ha szigorúak akarunk lenni, akkor azért annyit hozzá kell tennünk, hogy *ha* létezik egyáltalán a feltételes várható érték. Vannak eloszlások, amiknek egyszerűen nem létezik várható értéke, úgyhogy ez elvileg nem mindegy, de mi most ilyen helyzetekkel nem fogunk foglalkozni.

$\left(Y \mid \underline{X}\right)$ objektumot nem szokás definiálni), ehelyett a bal oldalra simán Y-t írunk (de ne feledjük, hogy ez *feltételes*).

Nagyon fontos látni, hogy a regresszió *mindig* felírható így! És ráadásul ez *biztosan* optimális felírás. A mi választásunk az lesz, hogy majd $\mathbb{E}\left(Y \mid \underline{X}\right)$ helyébe mit írunk be, például *ha* tudjuk hogy minden változó együttes eloszlása többváltozós normális, akkor azt, hogy $\mathbb{E}\left(Y \mid \underline{X}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$, és így azt kapjuk, hogy

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon,$$

de vigyázat, ez már – szemben az előző formulával – *nem* univerzális, csak normalitás esetén érvényes.

Lehetne $\mathbb{E}\left(Y \mid \underline{X}\right)$ helyébe más is beírni, de mindaddig, amíg jellegre ilyen megoldást használunk, tehát megadjuk a függvény formáját, csak egy vagy több paraméter az, ami meghatározza a konkrét függvényt, szokás **paraméteres regresszióról** beszélni. Ez nem kötelező, lehetne az $\mathbb{E}\left(Y \mid \underline{X}\right)$ anélkül próbálni közelíteni, hogy bármilyen *konkrét* függvényforma mellett elköteleződnék, ekkor beszélünk nem-paraméteres regresszióról. Ilyenekkel most nem foglalkozunk, csak az érzékeltetés kedvéért egy lehetőség: TODO

Tegyünk még egy megállapítást, ami most nem tűnik nagyon izgalmasnak, de a későbbiekben rettentő fontos lesz. Annyit kell tudnunk hozzá, hogy a feltételes várható érték is lineáris, valamint, hogyha valaminek kétszer vesszük a várható értékét, az ugyanaz mintha egyszer vennénk, és ez nem csak a szokásos várható értékre, hanem a feltételes várható értékre is igaz:

$$\mathbb{E}\left(\varepsilon \mid \underline{X}\right) = \mathbb{E}\left(Y - \mathbb{E}\left(Y \mid \underline{X}\right) \mid \underline{X}\right) = \mathbb{E}\left(Y \mid \underline{X}\right) - \mathbb{E}\left[\mathbb{E}\left(Y \mid \underline{X}\right) \mid \underline{X}\right] = \mathbb{E}\left(Y \mid \underline{X}\right) - \mathbb{E}\left(Y \mid \underline{X}\right) = 0.$$

Azaz azt kapjuk, hogy $\mathbb{E}\left(\varepsilon \mid \underline{X}\right) = 0$. Nagyon fontos, hogy értsük, hogy most mit mondunk: *ha* (!) tényleg – valóságban helyes – $\mathbb{E}\left(Y \mid \underline{X}\right)$-t használjuk, *akkor* $\mathbb{E}\left(\varepsilon \mid \underline{X}\right) = 0$ *kell* legyen. Ez azért lesz izgalmas, mert $\mathbb{E}\left(Y \mid \underline{X}\right)$-t mi sem tudhatjuk biztosan, majd be kell valamit írnunk a helyébe (például azt, hogy $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$). Ez a megállapítás tehát azt mondja, hogy *ha* beletrafálunk a dologba, *akkor* $\mathbb{E}\left(\varepsilon \mid \underline{X}\right) = 0$ kell legyen. *De* ha nem (például ezt írjuk be, csak épp közben nem normális az eloszlás), akkor már ez egyáltalán nem biztos, hogy igaz lesz!

Összefoglalva, ott tartunk, hogy *ha* az eloszlás normális akkor $\mathbb{E}\left(Y \mid \underline{X}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ és így $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$. A bökkenő persze ott van, hogy azt mi magunk sem tudhatjuk, hogy milyen a változóink eloszlása. És itt jön egy fontos döntés. Munkánkat úgy fogjuk megkezdeni, hogy azt mondjuk *akármilyen* is az eloszlás, mi *mindenképp* az $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$ modellt használjuk! Ezt fogjuk **lineáris regressziónak** nevezni. Még egyszer: ez egy helyes döntés, ha normális a változóink eloszlása, de különben nem. Ha nem ismerjük a változóink eloszlását, akkor ez többé már nem egy matematikailag levezethető szükségszerűség, hanem egy *választás* a részünkről. De több ok szól e választás mellett:

- Többváltozós normalitásnál egzaktan ez a helyzet
- Más esetekben ugyan nem, de cserében nagyon kellemesek a tulajdonságai, különösen ami az interpretációt illeti

- Az is elmondható, hogy – a Taylor-sorfejtés logikáját követve – bármi más is a jó függvényforma, legalábbis *lokálisan* ez is jó közelítés kell legyen
- Bár első ránézésre ez vegytisztán lineáris, valójában majd látni fogjuk, hogy egy sor nemlineáris modell is *visszavezethető* erre a modellre
- És végezetül egy lényeges szempont: lesznek majd eszközeink arra, hogy észrevegyük, hogy rossz volt ez a választás, és megpróbáljuk kijavítani

De újfent nagyon fontos hangsúlyozni, hogy a valós munka során, ahol nem tudjuk mik az eloszlások, azt, hogy az $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$ fennáll, már nem kezelhetjük matematikai szükségszerűségnek, hanem mint feltételt fel kell tennünk!

1.6. A szálak összeérnek

Ezen a ponton összeérnek a szálak. Vegyük csak újra az előbbi alakot (és feltételezzük, hogy a szükséges feltevések teljesültek):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon.$$

Mi itt

2. fejezet

Regresszió a mintában: következtetés

Pár fogalmat talán érdemes feleleveníteni következtető statisztikából. Az alapprobléma: a halmaz amire a kérdésünk irányul, a **sokaság** sajnos azonban ennek minden elemét nem tudjuk megfigyelni (azaz lemérni), csak egy részét, a kisebb részhalmaz a **minta**. Ez előfordulhat akkor, ha a sokaság TODO

2.1. A hagyományos legkisebb négyzetek (OLS) elve

Ilyen becslési elv a hagyományos legkisebb négyzetek (ordinary least squares, OLS) elve. Mint általános becslési el, nem kell hozzá semmilyen regresszió, a legközönségesebb következtető statisztikai példán is elmondható. Példaként vegyük az egyik legelemibb kérdést: sokasági várható érték becslése normalitás esetén, tehát a sokaság eloszlása normális (az egyszerűség kedvéért legyen a szórás is ismert, tehát azt nem kell becsülnünk). Ami fontos: bár egy alap következtető statisztika kurzuson nem szokták mondani, de lényegében itt is az a helyzet, hogy egy *modellt* feltételezünk a sokaságra, jelesül: $Y \sim \mathcal{N}\left(\mu, \sigma_0^2\right)$, amit nem mellesleg úgy is írhatnánk, hogy $Y = \mu + \varepsilon$, ahol $\varepsilon \sim \mathcal{N}\left(0, \sigma_0^2\right)$. Most μ megbecslése céljából veszünk egy n elemű fae (független, azonos eloszlású) mintát a sokaságból; ekkor feltevésünk szerint $Y_i = \mu + \varepsilon_i$ lesz az i-edik mintaelem. (A feltevésünk igazából azt jelentette, hogy az ε_i változók függetlenek és azonos eloszlásúak). Figyeljünk oda a kis és nagybetűkre! A nagy betű valószínűségi változó, valami aminek eloszlása van, sokasági dolog. Kisbetű egy konkrét szám, nem valószínűségi, nincsen eloszlása, mintabeli dolog. Most valaki megkérdezhetné, hogy oké, azt értem, hogy Y miért nagy betű, de az Y_i miért az? Hiszen azt mondtuk, hogy az az egyik mintaelem…! Talán a legjobban úgy lehet ezt elképzelni, hogy a véletlen mintavétel az, hogy megkeverjük az urnát, hogy kihúzzunk belőle egy golyót. Megáll a keverés, nyúlunk bele az urnába, hogy húzzunk: ekkor számunkra az még egy véletlen dolog, hogy mi lesz az elsőként húzott elem, annak eloszlása van (fae mintavétel esetén

– tehát ha a golyókat mindig visszadobjuk, és az urnát mindig jól átkeverjük – ugyanaz, mint a sokaság, tehát mint az egész urna eloszlása). Ekkor ez még Y_1 számunkra. Ekkor kihúzzuk a golyót, és meglátjuk a konkrét értéket: ez lesz y_1. Kicsit matematikusabban szólva: kaptunk egy realizációt Y_1-ből, ez lesz az y_1.

A másik ami fontos: a modellből következik egy *becsült érték* minden mintabeli elemhez, jelen esetben, ha m egy feltételezett érték az ismeretlen sokasági várható értékre, akkor

$$\widehat{y}_i = m.$$

(Itt persze elvileg beszélni kellene arról, hogy még ha tudjuk is, hogy a sokasági várható érték m, miért pont az lesz a becslésünk is minden mintaelemre. Fogadjuk el intuitíve, egyébként olyan érvelést használhatnánk mint az előző fejezetben, úgy, hogy az egyetlen magyarázó változónk az $X_1 = 1$.)

Egy kis kitérő megjegyzés: ha jobban megnézzük a fentieket, akkor láthatjuk, hogy az OLS-elv alkalmazásához igazából nem is kell, hogy a sokasági eloszlást ismerjük, csak annyi a fontos, hogy legyen egy modellünk, és belőle tudjunk becsült értékeket származtatni a ténylegesen is ismert megfigyelésekhez.

És akkor jöhet az OLS-elv! Egy mondatban összefoglalva: az ismeretlen sokasági paraméterre az a becsült érték, amely mellett a tényleges mintabeli értékek, és az adott paraméter melletti, modellből származó becsült értékek közti eltérések négyzetének összege a legkisebb! A megoldandó – optimalizációs jellegű – feladat tehát matematikailag:

$$\widehat{\mu} = \arg\min_m \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 = \arg\min_m \sum_{i=1}^{n} \left(y_i - m\right)^2$$

És ennek megoldása természetesen $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$ ebben a példában.

Egyetlen kiegészítést kell tenni a fentiekhez. Megkaptunk ugyan a becslőt, csakhogy az $\overline{y}$ egyetlen konkrét szám. (Hát persze, mert egy konkrét mintához, a $\{y_1, y_2, \ldots, y_n\}$ mintához – kisbetűk! – tartozik.) Minket azonban alapvető fontossággal fog érdekelni a becslő **mintavételi eloszlása**, tehát, hogy ha újra meg újra mintát veszünk ugyanabból a sokaságból, és mindegyik mintából kiszámoljuk a becslőfüggvény értékét (jelen esetben a mintaátlagot), akkor annak mi lesz az eloszlása. A becslőfüggvényünk az igazából egy *transzformáció* a mintaelemekkel („add össze őket és oszd el a mintanagysággal"), de ha egyszer ez a transzformáció megvan, azt nyugodtan ráereszthetjük valószínűségi változókra is, nem csak számokra! Ami magyarán azt fogja jelenteni, hogy felírjuk ugyanazt – csak épp kisbetűk helyett nagybetűkkel. Jelen példában a becslőfüggvényünk: $\frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$, és íme, ennek már nagyon is eloszlása van, hiszen egy valószínűségi változó maga is – ez az eloszlás lesz a mintavételi eloszlás. Megvizsgálhatóak a tulajdonságai, megnézhetjük, hogy a várható értéke egyezik-e a sokasági paraméterrel (torzítatlanság), hogy mekkora a szórása (hatásosság), hogy hogyan viselkedik, ha n egyre nagyobb (konzisztencia) és így tovább.

2.2. Lineáris regresszió becslése OLS-elven

Most vegyük elő a lineáris regressziónkat! (Ahol ezt közszemérem-sértés veszélye nélkül megtehetjük.) Azt látjuk, hogy ott eddig a sokaságról beszéltünk, feltettünk egy modellt (*ugyanúgy mint az előbbi példában*), jó, lehet, hogy egy kicsit bonyolultabbat, de akkor is, ugyanúgy egy sokaságra vonatkozó modell, amiből, megint csak pontosan ugyanúgy mint az előbbi példában, tudunk egy becsült értéket előállítani minden mintaelemhez. Ez lehetővé teszi, hogy az ismeretlen paramétereket OLS-elven megbecsüljük!

Lássuk a részleteket. A változóink az $\left(Y, X_1, X_2, \dots, X_k\right)$, ezekre vegyünk egy n elemű mintát; az i-edik mintaelemet jelölje $\left(Y_i, X_{i1}, X_{i2}, \dots, X_{ik}\right)$. Természetesen a modellünk ezekre is igaz lesz, tehát írhatjuk, hogy

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i.$$

Ez minden i-re teljesül, tehát ha nagyon elszántak vagyunk, akkor n ilyen egyenletet írhatnánk fel:

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_k X_{1k} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \dots + \beta_k X_{2k} + \varepsilon_2 \\
&\dots \\
Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_k X_{nk} + \varepsilon_n
\end{aligned}$$

Az i-edik mintaelem realizációja az $\left(y_i, x_{i1}, x_{i2}, \dots, x_{ik}\right)$. (A minta egyelőre legyen fae – hogy ez mennyire jó feltevés, arról később még fogunk beszélni.)

Ha $b_0, b_1, \dots, b_k$-val jelöljük a feltételezett sokasági paramétereket, akkor a becslés

$$\widehat{y_i} = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik}$$

lesz az i-edik mintaelemre. (Itt szerencsére nincs mit gondolkozni, hiszen azt az előző fejezetben részletesen levezettük, hogy ez lesz a legjobb becslés adott $\mathbf{x}$ mellett.)

Most hogy megvannak a becsült értékek ($\widehat{y_i}$) és a tényleges értékek (y_i), betű szerint ugyanazt az optimalizációs feladatot kell felírnunk, mint az előbb, csak $\widehat{y_i}$ lesz kicsit hosszabb, ha kifejtjük:

$$\begin{aligned}
\left(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}, \dots, \widehat{\beta_k}\right) &= \underset{b_0, b_1, b_2, \dots, b_k}{\arg\min} \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2 = \\
&= \underset{b_0, b_1, b_2, \dots, b_k}{\arg\min} \sum_{i=1}^{n} \left[y_i - \left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik}\right)\right]^2
\end{aligned}$$

Annyi bonyolódottság van, hogy itt most *több* paramétert kell becsülni, de ez csak a kivitelezést nehezíti, elvileg teljesen ugyanaz a feladat.

Össze ne keverjük β_i-t, b_i-t és $\widehat{\beta_i}$-t! β_i a kérdéses sokasági paraméter valódi, tényleges értéke, egy adott, konkrét szám (csak mi nem tudjuk mennyi), b_i egy általunk feltételezett

érték rá, mi állítjuk be, választhatunk nagy számot is, kis számot is, tetszés szerint, a fenti optimalizációban végig fogunk vele futni az összes lehetséges értékén, $\widehat{\beta}_i$ pedig a megoldásként kapott *legjobb tippünk* β_i-re, de ettől még csak tipp, azaz eloszlása lesz, hiszen a mintától is függeni fog, mintáról mintára ingadozni fog (miközben a valódi érték ugyebár állandó – ez lesz a mintavételi hiba forrása).

Ezt az optimalizációs problémát kell tehát most megoldanunk. Ezt megtehetnénk a fenti formában is, de célszerűbb, ha már most áttérünk a vektoros/mátrixos jelölésekre. Ez eleinte kicsit kényelmetlennek tűnhet, de a magasabb absztrakciós szint később ki fog fizetődni: lehet, hogy most kicsit nehezebben indulunk, de a cserében a bonyolultabb problémák sem lesznek sokkal nehezebbek.

Fogjunk tehát össze mindent értelemszerű vektorokba és mátrixokba! A jelölésrendszer teljes bemutatása végett felírom a mintavétel előtti – valószínűségi változós – és a realizálódott értékes alakokat is[1]. Az eredményváltozók:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}$$

A magyarázó változókat nyilván mátrixba kell összefogni, de itt egy kis cselre lesz szükségünk: hozzáveszünk az elejéhez egy csupa 1 oszlopot. (Az így kapott mátrixot a regresszió **design mátrixának** szokás nevezni.) Íme:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \underline{\underline{X}} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}$$

Ez a csupa 1 oszlop azért lesz célszerű, mert ha a regressziós koefficienseket egy

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}$$

vektorba, a hibatagokbat pedig egy

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_1 \\ \dots \\ \varepsilon_k \end{pmatrix}$$

[1] A jelölésrendszer sajnos nem tökéletesen konzisztens, hiszen **X** nagybetű, és mégis kisbetűs dolgokat fog össze. Nem akartam szakítani a lineáris algebra hagyományával, hogy a mátrixot nagybetű jelöli, bár ez tényleg keveredik a valószínűségszámítás nagybetűjével. Abból azonban, hogy vastagítás vagy aláhúzás van, mindenképpen világos lesz, hogy valószínűségi változóról vagy realizálódott értékről van szó, még ha a kis és nagy betű nem is segít.

vektorba fogjuk össze, akkor a korábbi, n darab egyenletből álló, igencsak terjengős felírás helyett nemes egyszerűséggel ezt írhatjuk:

$$\underline{Y} = \underline{\underline{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

És ennyi, pontosan ugyanaz van leírva!

Látható tehát, hogy a csupa 1 oszlop azért kellett, hogy a vektorral való rászorzásnál az legyen a β_0 szorzója, így az egyenletben tényleg egyszerűen β_0 fog megjelenni.

Menjünk most vissza az OLS optimalizációs problémájára! Ezekkel a jelölésekkel a kezünkben ugyanis azt is sokkal egyszerűbben felírhatjuk:

$$\arg\min_{\mathbf{b}} \hat{\mathbf{y}}^T\hat{\mathbf{y}} = \arg\min_{\mathbf{b}} (\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b}),$$

hiszen számok négyzetösszegét megkapjuk, ha összefogjuk őket egy vektorba, és vesszük ezen vektor saját transzponáltjával vett szorzatát. ($\hat{\mathbf{y}}$ és $\mathbf{b}$ az értelemszerű vektorok, $\hat{y}_i$-ket és b_i-ket fogják össze.)

Az $(\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b})$ hibanégyzetösszeget ESS-sel (error sum of squares) is fogjuk jelölni[2].

És akkor essünk neki: oldjuk meg ezt az optimalizációt! Először alakítsuk át a célfüggvényt, bontsuk fel a zárójeleket:

$$\arg\min_{\mathbf{b}} (\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b}) = \arg\min_{\mathbf{b}} \left[\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right].$$

Itt egyszerű algebrai átalakításokat végzünk (és a definíciókat használjuk), hiszen a zárójeleket felbontani, műveleteket elvégezni, mátrixokkal/vektorokkal is hasonlóan kell mint valós számokkal. (A transzponálás tagonként elvégezhető, azaz $(\mathbf{a}-\mathbf{b})^T = \mathbf{a}^T - \mathbf{b}^T$.) Egyedül annyit kell észrevenni, hogy a $\mathbf{y}^T\mathbf{X}\mathbf{b}$ egy egyszerű valós szám, ezért megegyezik a saját transzponáltjával, $\mathbf{b}^T\mathbf{X}^T\mathbf{y}$-nal. Ezért írhattunk $-(\mathbf{X}\mathbf{b})^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{b}$ helyett egyszerűen – például – $-2\mathbf{b}^T\mathbf{X}^T\mathbf{y}$-t. (Itt mindenhol felhasználtuk, hogy a transzponálás megfordítja a szorzás sorrendjét: $(\mathbf{A}\mathbf{B})^T = \mathbf{A}^T\mathbf{B}^T$.)

Most jön a minimum megkeresése. Az ember rávágja, hogy deriválni kell, de itt ez picit zűrősebb, hiszen a függvényünk többváltozós (ráadásul az is határozatlan, hogy pontosan hányváltozós). Itt jelentkezik igazán a mátrixos jelölésrendszer előnye. A $\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{b} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}$ lényegében egy „másodfokú kifejezés" többváltozós értelemben (az ax^2+bx+c többváltozós megfelelője), és ami igazán szép: pont ahogy az ax^2+bx+c lederiválható a változója (x) szerint (eredmény $2ax+b$), ugyanúgy ez is lederiválható a változója (azaz $\mathbf{b}$) szerint… és az eredmény az egyváltozóssal teljesen analóg lesz, ahogy fent is látható! (Ez persze bizonyítást igényel! – lásd többváltozós analízisből.) Bár ezzel átléptünk egyváltozóról többváltozóra, a többváltozós analízisbeli eredmények

[2]Sajnos néhány irodalom az általunk használt ESS-re inkább az RSS-t (residual sum of squares) rövidítést használja, ami a jelölési zavarok legszerencsétlenebb típusa, ugyanis az RSS-t majd később mi is fogjuk használni, csak épp másra. Éppen ezért, ha ilyenekről olvasunk, mindig tisztázni kell, hogy a könyv vagy program írói mit értenek alatta.

biztosítanak róla, hogy formálisan ugyanúgy végezhető el a deriválás. (Ezt írja le röviden a „vektor szerinti deriválás" jelölése. Egy $\mathbf{b}$ vektor szerinti derivált alatt azt a vektort értjük, melyet úgy kapunk, hogy a deriválandó kifejezést lederiváljuk $\mathbf{b}$ egyes b_i komponensei szerint – ez ugye egyszerű skalár szerinti deriválás, ami már definiált! –, majd ez eredményeket összefogjuk egy vektorba. Látható tehát, hogy a vektor szerinti derivált egy ugyanolyan dimenziós vektor, mint ami szerint deriváltunk.) Ami igazán erőteljes ebben az eredményben, az nem is egyszerűen az, hogy „több" változónk van, hanem, hogy nem is kell tudnunk, hogy mennyi – mégis, általában is működik! Az eredmény tehát:

$$\frac{\partial}{\partial \mathbf{b}} \left[\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right] =$$
$$= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b} = 0 \Rightarrow \widehat{\boldsymbol{\beta}_{\mathrm{OLS}}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y},$$

ha $\mathbf{X}^T\mathbf{X}$ nem szinguláris.

Azt, hogy a megtalált stacionaritási pont tényleg minimumhely, úgy ellenőrizhetjük, hogy megvizsgáljuk a Hesse-mátrixot a pontban. A mátrixos jelölésrendszerben ennek az előállítása is egyszerű, még egyszer deriválni kell a függvényt a változó(vektor) szerint:

$$\frac{\partial^2}{\partial \mathbf{b}^2} \left[\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}\right] = \frac{\partial}{\partial \mathbf{b}} \left[-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b}\right] = 2\mathbf{X}^T\mathbf{X}.$$

Az ismert tétel szerint a függvénynek akkor van egy pontban ténylegesen is (lokális, de a konvexitás miatt egyben globális) minimuma, ha ott a Hesse-mátrix pozitív definit. Esetünkben ez minden pontban teljesül. A $\mathbf{X}^T\mathbf{X}$ ugyanis pozitív szemidefinit (ez egy skalárszorzat-mátrix, más néven Gram-mátrix, amelyek mindig pozitív szemidefinitek), a kérdés tehát csak a határozott definitség. Belátható azonban, hogy ennek feltétele, hogy $\mathbf{X}^T\mathbf{X}$ ne legyen szinguláris – azaz itt is ugyanahhoz a feltételhez értünk! Megjegyezzük, hogy ez pontosan akkor valósul meg, ha az $\mathbf{X}$ teljes oszloprangú. (Erre a kérdésre a modellfeltevések tárgyalásakor még visszatérünk.)

Végül egy számítástechnikai megjegyzés: az együtthatók számításánál a fenti formula direkt követése általában nem a legjobb út, különösen ha sok megfigyelési egység és/vagy változó van. Ekkor nagyméretű mátrixot kéne invertálni, amit numerikus okokból (kerekítési hibák, numerikus instabilitás stb.) általában nem szeretünk. Ehelyett, a különféle programok igyekeznek a direkt mátrixinverziót elkerülni, tipikusan az $\mathbf{X}$ valamilyen célszerű mátrix dekompozíciójával (QR-dekompozíció, Cholesky-dekompozíció). Extrém esetekben még az is elképzelhető, hogy az egzakt, zárt alakú megoldás előállítása helyett valamilyen iteratív optimalizálási algoritmus (gradiens módszer, Newton–Raphson-módszer) alkalmazása a gyakorlatban járható út, annak ellenére is, hogy elvileg van zárt alakban megoldása.

A kapott eredmény nem más, mintha $\mathbf{X}$ Moore-Penrose pszeudoinverzével szoroznánk $\mathbf{y}$-t.

TODO

Végezzük el a fenti műveleteket közvetlenül lekódolva R-ben a már látott kaliforniai iskolás példára, ha a pontszámot a tanár:diák arányt a pontszámmal és a jövedelemmel regresszáljuk:

```
y <- CASchools$score
X <- cbind( 1, CASchools$tsratio, CASchools$income )
solve( t(X)%*%X )%*%t(X)%*%y
```

```
##       [,1]
## [1,] 614.0
## [2,] 233.4
## [3,]   1.8
```

Egy mátrixot a `t` függvénnyel transzponálhatunk és a `solve` függvénnyel invertálhatunk, a `cbind` pedig vektorokat, mint oszlopvektorokat fűz egybe mátrixszá. (Valaki megkérdezheti, hogy akkor az `1` miért működik, hiszen az nem vektor: ez az R egyik jellemző – kétélű fegyverként viselkedő – tulajdonsága: megengedi a trehányságot, ugyanis érzékeli, hogy mi a helyzet, és automatikusan egymás alá rakja annyiszor, mint amilyen hosszúak a többi vektorok.)

Természetesen az R tartalmaz beépített parancsot regressziók becslésére:

```
lm( score ~ tsratio + income, data = CASchools)
```

```
##
## Call:
## lm(formula = score ~ tsratio + income, data = CASchools)
##
## Coefficients:
## (Intercept)      tsratio       income
##      613.98       233.41         1.84
```

Az `lm` a lineáris modell rövidítése. Első argumentumban a regressziós egyenletet kell megadnunk, mint egy R formula (tehát `~` felel meg az egyenlőségjelnek, bal oldalán az eredményváltozó, jobb oldalán a magyarázó változók felsorolása, `+` jellel elválasztva.) Az R konstans alapbeállításként rak a modellbe, azt kell külön kérnünk ha nem szeretnénk (egy `-1` hozzáfűzésével az utolsó magyarázó változó után). A `data` argumentum tartalma a szokásos: ha használjuk, akkor a formulában elég a változóneveket leírni, nem kell jelölni, hogy melyik adatkeretre vonatkoznak, mert az R úgy érti, hogy mind a `data` argumentumban megadottra értendő.