doi:10.20312/dim.2023.04

Többváltozós lineáris regresszió gyakorlati alkalmazásai STATISTICA13 programcsomag segítségével

Horváth-Szováti Erika
Soproni Egyetem, Informatikai és Matematikai Intézet
horvath-szovati.erika@uni-sopron.hu

ÖSSZEFOGLALÓ. Egyszerű mintapéldákkal szemléltetjük a többváltozós lineáris regresszió gyakorlati alkalmazását. A programcsomag által megadott eredmények értelmezése és egyszerű esetekben számításokkal történő igazolása hallgatóink számára hasznos tudást nyújthat.

ABSTRACT. Here are simple practical examples that illustrate the practical use of multivariate linear regression. Interpreting the results of the program package and verifying them with calculations in simple cases can provide useful knowledge for the students.

1. Bevezetés

A többváltozós statisztikában használt eljárások a XX. század első felében születtek. A klasszikus módszerek (pl. a regresszióanalízis, varianciaanalízis és diszkriminancia-analízis) elsősorban a lineáris algebra eredményeire épülnek, normális eloszlású valószínűségi változókat használnak. Napjainkban a munkaerőpiacon az adatelemzésé a jövő, emiatt nagyon fontos a statisztika nyelvét értő és azon kommunikálni képes szakemberek képzése. Azokban a felsőoktatási intézményekben, ahol jelenleg a matematika oktatása egyre hátrányosabb helyzetbe kerül, a többváltozós statisztika tárgyalása komoly nehézséget jelenthet. A hallgatók csekély matematikai előképzettsége miatt az eljárások elméleti háttérből csak a legszükségesebb részek, azok is nagyrészt leegyszerűsítve kerülnek szóba. A gyakorlati alkalmazásra fókuszálunk, a módszerek lényegét mintapéldákon keresztül mutatjuk be a STATISTICA13 programcsomag segítségével.

2. Regresszióanalízis

A regresszióanalízis alapgondolata az, hogy egyes változók között ok–okozati összefüggést feltételezünk, amelynek leírására egy függvénykapcsolatot keresünk. Fontos megjegyezni, hogy az a fő különbség a korrelációszámítás és a regresszióanalízis között, hogy korrelációszámítás során nincsen feltételezett ok-okozati kapcsolat, minden változót valószínűségi változónak tekintünk. A többváltozós regresszióanalízis során több ismérv eredményváltozóra gyakorolt hatását vizsgáljuk. A többváltozós regresszió az ismérvek száma szerint lehet három-, négy-, öt- stb. változós (hiszen már egyváltozós függvény illesztésekor is két ismérv szerepel), a függvény típusa szerint pedig lineáris és nemlineáris. A többváltozós regressziószámítás modelljében az okot a magyarázó változókkal, más néven független (independent) változókkal ($X_1, X_2, \dots, X_m$), az okozatot pedig az eredmény változóval, más néven függő (dependent) változóval (Y) jelöljük, ez utóbbit tekintjük valószínűségi változónak,

amit a továbbiakban a „kalap” jelölés is mutat. A cél annak az $\hat{Y} = f(X_1, X_2, \dots, X_m)$ függvénykapcsolatnak a megtalálása, amely „kellően jól” írja le a vizsgált ok-okozati összefüggést. Itt csak a többváltozós lineáris esettel foglalkozunk, tehát jelen esetben f lineáris függvény.

A többváltozós lineáris regresszió lépései a következők: 1. modellalkotás, 2. együtthatók meghatározása, 3. determinációs és korrelációs együtthatók kiszámítása, 4. megbízhatóság vizsgálata F-próbával, 5. együtthatók t-próbája, 6. validálás (esetleg további módszerek, mérőszámok, mutatók használata, amelyekkel meggyőződhetünk arról, hogy a modell megfelelő-e, majd jóváhagyás/elutasítás). Ebből a felsorolásból is látszik, hogy nem létezik egyetlen egy olyan mérőszám, ami megmutatná, hogy melyik függvény a „legjobb”. A végső modell kiválasztása a becslési pontosság és az egyszerűség kompromisszuma, nem elhanyagolva a szakmai (adott tudományterületre vonatkozó) szempontokat.

2.1. Korrelációs mátrix, többszörös determinációs együttható

A statisztika tanulmányainkból ismert, hogy két változó közötti lineáris korreláció mérőszámát r-rel jelöljük, $r \in [-1, 1]$. Ha $r = 1$ (vagy $r = -1$), akkor a mérési pontok 1 valószínűséggel egy növekvő (vagy csökkenő) egyenesen helyezkednek el, azaz lineáris kapcsolat van közöttük. Emiatt az r úgy is tekinthető, mint a linearitás mérőszáma. Ha két adatsor független egymástól, akkor $r = 0$, tehát a függetlenségből következik a korrelálatlanság. Ha $r = 0$, azaz korrelálatlanság áll fenn, ebből nem feltétlen következik a két változó függetlensége (csupán a lineáris kapcsolat hiánya). Kettőnél több változó esetén a korreláció szorosságáról háromféle értelemben beszélhetünk. Mérhető az eredményváltozó és az összes tényezőváltozó közötti kapcsolat szorossága, emellett vizsgálható az egyes változók közötti kapcsolat páronként (a páronkénti korrelációs együtthatók), továbbá páronként, de a többi változó hatásának kiszűrésével is (a parciális korrelációs együtthatók) is. Alább ezt a három lehetőséget is tárgyaljuk.

Többváltozós regresszió esetén jelöljük az i-edik és j-edik változó korrelációját r_{ij}-vel, és ezekből a páronkénti korrelációs együtthatókból összeállíthatjuk az ún. korrelációs mátrixot. A korrelációs mátrix minden elemére $r_{ij} \in [-1, 1]$, továbbá $r_{yy} = r_{11} = r_{22} = \dots = r_{mm} = 1$ (a regressziós modellben a független változók száma m), és a főátlójára szimmetrikus mátrix:

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{y1} & r_{y2} & \dots & r_{ym} \\ r_{1y} & 1 & r_{12} & \dots & r_{1m} \\ r_{2y} & r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{my} & r_{m1} & r_{m2} & \cdots & 1 \end{pmatrix}. \tag{1}$$

A korrelációs mátrixból megállapítható, hogy melyek azok a magyarázó változók, amelyek szorosabb kapcsolatban vannak a függő változóval, és melyek azok, amelyek kevésbé. A korrelációs mátrix inverze ($\boldsymbol{R}^{-1}$):

$$\boldsymbol{R}^{-1} = \begin{pmatrix} q_{yy} & q_{y1} & q_{y2} & \dots & q_{ym} \\ q_{1y} & q_{11} & q_{12} & \dots & q_{1m} \\ q_{2y} & q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{my} & q_{m1} & q_{m2} & \cdots & q_{mm} \end{pmatrix}. \tag{2}$$

A (2) mátrix első sora első elemének (q_{yy}) segítségével meghatározható az ún. többszörös korrelációs együttható:

$$R_{y.1,2,3,\dots m} = \sqrt{1 - \frac{1}{q_{yy}}}\ , \qquad R_{y.1,2,3,\dots m} \in [0,1]\ . \tag{3}$$

Ez valójában egy speciális kétváltozós korrelációs együttható (erre az indexében a felsorolást kettéválasztó pont karakter is utal), amely a mért Y-ok és az $X_1, X_2, \dots, X_m$ tényezőváltozók alapján becsült $\hat{Y}$ kapcsolatának szorosságát méri többváltozós lineáris modell esetén. Vannak olyan irodalmak, amelyekben a (3)-ban szereplő alsó indexre az $Y(X_1, X_2, \dots, X_m)$ jelölést használják, ez talán még szemléletesebben tükrözi a jelentését. A többszörös korrelációs együttható négyzetét többszörös determinációs együtthatónak nevezzük:

$$R^2_{y.1,2,3,\dots m} = 1 - \frac{1}{q_{yy}}\ , \qquad R^2_{y.1,2,3,\dots m} \in [0,1]\ . \tag{4}$$

Ennek használatával számszerűsíthető a regressziós függvénnyel történő „magyarázat jósága", az illeszkedés szorossága, az „előrejelzés hibájának nagysága". Ha $m = 1$, akkor a többszörös determinációs együttható a függő- és az egyetlen független változó közötti r^2 determinációs együtthatóval egyenlő, azaz $R^2_{y.1} = r^2$. Az r^2 egyben a linearitás mérőszáma is, értéke minél közelebb esik 1-hez, a megfigyelési értékek annál jobban tömörülnek egy egyenes mentén (annál inkább igazoltnak látszik a linearitási feltevés). Az $R^2_{y.1} = r^2$ egyenlőség bizonyítása alapszintű lineáris algebrai ismeretek segítségével is nagyon egyszerű, ezért ezt érdemes megtennünk. Célszerű már az elején bevezetni az $r_{y1} = r_{1y} = r$ jelölést. A nevezők miatt felmerülhet az $r \neq \pm 1$ kikötés szükségessége is, de az $r = \pm 1$ eset ellentmondana a többváltozós lineáris regresszió egyik alkalmazhatósági feltételének (multikollinearitás kizárása), így soha nem áll fenn a többváltozós lineáris korreláció alkalmazásakor.

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{y1} \\ r_{1y} & 1 \end{pmatrix} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \tag{5}$$

$$\boldsymbol{R}^{-1} = \begin{pmatrix} \frac{1}{1-r^2} & \frac{-r}{1-r^2} \\ \frac{-r}{1-r^2} & \frac{1}{1-r^2} \end{pmatrix}, \tag{6}$$

$$R^2_{y.1} = 1 - \frac{1}{\frac{1}{1-r^2}} = r^2. \tag{7}$$

Az egyváltozós függvény illesztésekor (tehát két változó esetén) használt determinációs együttható egyben azt is megmutatja, hogy a függő változó teljes szórásnégyzetének mekkora hányada magyarázható a független változóval, tehát a regressziós függvénnyel. Bizonyítható, hogy ugyanez a többszörös determinációs együtthatóra is igaz (ennek levezetésével itt nincs értelme foglalkoznunk), azaz felírható a többváltozós modellben alkalmazott eltérés négyzetösszegek hányadosaként is:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}\ . \tag{8}$$

A (7)-ben és azt megelőzően használt alsó indexek a gyakorlatban akár el is hagyhatók, (8)-ban már nem írtuk ki. A (8)-ban szereplő jelölések értelmezéséhez definiálnunk kell az alább olvasható négyzetösszegeket (SS =Sum of Squares), ahol n a mérések számát jelöli. Teljes

eltérés négyzetösszeg: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$, a regresszió által meghatározott eltérés négyzetösszeg: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, valamint a reziduális eltérés négyzetösszege (más néven maradékok négyzetösszege, ami a mérés „hibáiból" származtatható): $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. A függő változó átlagtól vett eltérés négyzetösszege felbontható a regressziós becslések átlagtól vett eltérés négyzetösszegének és a reziduális négyzetösszegnek az összegére: $SST = SSR + SSE$.

Az (1) korrelációs mátrixban szereplő páronkénti korrelációs együtthatókból is kiszámítható az R^2 többszörös determinációs együttható. Egy eredményváltozó és két magyarázó változó esetén a következő összefüggés használatos:

$$R^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}. \qquad (9)$$

A többszörös determinációs együttható azt adja meg, hogy a függő változó teljes szórásnégyzetéből mekkora a tényezőváltozókkal (tehát a regresszióval) magyarázható hányad. Másképp fogalmazva R^2 nagysága a regressziós függvény magyarázó erejét mutatja meg, $R^2 \in [0,1]$, értékét %-ban szokás megadni. Az $R^2 = 1$ eset akkor áll fenn, ha az $X_1, X_2, \dots, X_m$ változók determinisztikusan meghatározzák az Y -t. Ha $R^2 = 0$, akkor az Y értékeinek szóródását teljes egészében a véletlen határozza meg. Ha ugyanazon minta esetén a magyarázó változók számát növeljük a regressziós modellben, akkor R^2 értéke automatikusan nő, és túl pozitív képet mutat az illeszkedésről. Így tehát a magas R^2 önmagában még nem feltétlenül jelent „jó" regressziós függvényt, az optimális modell változóinak kiválasztását nem célszerű kizárólag az R^2 mutatók alapján végezni.

2.2. Parciális korrelációk, szabadságfokkal korrigált R^2

Az (1) korrelációs mátrixban szereplő páronkénti korrelációs együtthatókból meghatározhatók az ún. parciális korrelációk is. Számításuk során két meghatározott változó közötti korreláció mérése valósul meg úgy, hogy minden további változó konstansként szerepel. Két magyarázó változós lineáris regresszió esetén jelölésük: $r_{y1.2}$, $r_{y2.1}$, $r_{12.y}$. Jelentésük: $r_{y1.2}$ például az y és x_1 kapcsolatának szorosságát méri, miközben x_2 hatását kiszűrjük, a többi ehhez hasonlóan. Parciális korreláció alkalmazása főleg akkor célszerű, ha két adatsor között sejtjük a kapcsolatot, de nem tudjuk kimutatni, mert egy harmadik adatsor eltakarja az összefüggést. A kapott eredmény $r_{y1.2}$, $r_{y2.1}$, $r_{12.y} \in [-1,1]$, értelmezése ugyanúgy történik, mint ahogy két változó lineáris korrelációjánál szokásos. A parciális korreláció abszolút értéke a két kiválasztott változó lineáris kapcsolatának szorosságát, az előjel pedig az irányát mutatja meg (pozitív esetben az összefüggést jellemző egyenes emelkedő, negatív esetben csökkenő), miközben a többi változó hatását figyelmen kívül hagyjuk. Az alábbi képletekkel számíthatók:

$$r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)\cdot(1-r_{12}^2)}}, \quad r_{y2.1} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{(1-r_{y1}^2)\cdot(1-r_{12}^2)}}, \quad r_{12.y} = \frac{r_{12} - r_{y1}r_{y2}}{\sqrt{(1-r_{y1}^2)\cdot(1-r_{y2}^2)}}. \quad (10)\text{-}(12)$$

Mivel nem a tökéletes, hanem a minél jobb modellt keressük, gyakran a különböző modellváltozatok összehasonlítását is el szeretnénk végezni. A legegyszerűbb (legkevesebb magyarázó változót tartalmazó) modell felírása a cél, mert a magyarázó változók számának növekedésével a multikollinearitás veszélye is megnő. Egymással lineárisan összefüggő magyarázó változók nem használhatók, szelektálni kell őket. Ezt segítheti a Theil-féle, szabadságfokkal korrigált R^2 (adjusztált R^2) , amit minden jelentősebb regressziós programcsomag kiszámít:

$$\bar{R}^2 = 1 - \frac{n-1}{n-m-1}(1-R^2) \quad , \tag{13}$$

ahol n a mérések száma, m a magyarázó változók száma. Tulajdonságaiban hasonlít az eredeti R^2-hez, viszont „bünteti" azt, ha túl sok magyarázó változót vonunk be a modellbe. A javaslat szerint azt a modellt célszerű választani, amelyik esetén az $\bar{R}^2$ maximális. Annál jobb a modell, minél közelebb van egymáshoz a korrigált és a korrigálatlan R^2 érték. A korrigált R^2 használatának is vannak hátrányai. Ha az $\bar{R}^2$-et mintából számítjuk, akkor valószínűségi változó, eloszlása függ a modell többi változójától, ezért a különböző modellek $\bar{R}^2$mutatói nem hasonlíthatók közvetlenül össze. Belátható, hogy ha R^2 olyan kicsi, hogy $R^2 < \frac{m}{n-1}$, akkor a korrigált $\bar{R}^2$ negatív lesz, összehasonlításra ilyenkor nem alkalmas. Mindezek ellenére az $\bar{R}^2$ kritériumot széles körben használják a többváltozós lineáris regresszió során, és legtöbbször helyes modellt eredményez.

3. Mintapéldák

3.1. Mintapélda

Az adatsor 30 dolgozó tanulmányi idejét ($x_1, [év]$), a munkában töltött éveinek számát ($x_2, [év]$) és a nettó munkabérét ($y, [100\ ezer\ Ft]$) tartalmazza. Az adatokra STATISTICA13 programcsomaggal többváltozós lineáris regressziós függvényt illesztettünk, és az 1.a-d. táblázatokban lévő eredményeket kaptuk ($\alpha = 0{,}05$ szignifikancia szinten dolgoztunk, $v1\sim$ tanulmányi idő években, $v2\sim$ munkában töltött évek, $v3\sim$ fizetés).

Regression Summary for Dependent Variable: v3 (01_Többváltozós lin regr_példák_STATISTICA (C3:AF32))
R= ,96873287 R2= ,93844338 Adjusted R2= ,93388363
F(2,27)=205,81 p<,00000 Std.Error of estimate: ,33791

N=30	b*	Std.Err. of b*	b	Std.Err. of b	t(27)	p-value
Intercept			-2,19850	0,302236	-7,27410	0,000000
v1	0,775550	0,047874	0,34337	0,021196	16,19986	0,000000
v2	0,527007	0,047874	0,07714	0,007007	11,00824	0,000000

Analysis of Variance; DV: v3 (01_Többváltozós lin regr_példák_STATISTICA (C3:AF32))

Effect	Sums of Squares	df	Mean Squares	F	p-value
Regress.	47,00006	2	23,50003	205,8103	0,000000
Residual	3,08294	27	0,11418		
Total	50,08300				

Variables currently in the Equation; DV: v3 (01_Többváltozós lin regr_példák_STATISTICA (C3:AF32))

Variable	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(27)	p-value
v1	0,775550	0,952216	0,773511	0,994750	0,005250	16,19986	0,000000
v2	0,527007	0,904318	0,525622	0,994750	0,005250	11,00824	0,000000

Correlations (01_Többváltozós lin regr_példák_STATISTICA (C3:AF32))

Variable	v1	v2	v3
v1	1,000000	0,072457	0,813735
v2	0,072457	1,000000	0,583201
v3	0,813735	0,583201	1,000000

1. **a-d. táblázatok**

Feladatok, kérdések:

I. Írja fel a regressziós egyenletet és értelmezze a paramétereket! Használja az 1.a táblázat adatait!

II. Mekkora a becsült fizetése egy 11 év tanulmányi idővel rendelkező szakmunkásnak 15 munkában eltöltött év után?

III. Írja le a paraméterek F-próbával történő tesztelésének gondolatmenetét az 1.b táblázat adatainak felhasználásával (globális tesztelés), dolgozzon 5%-os szignifikancia szinten!

IV. Írja le a paraméterek t-próbával történő tesztelésének gondolatmenetét az 1.a (vagy 1.c) táblázat adatainak felhasználásával (parciális tesztelések), használjon 5%-os szignifikancia szintet! Számítással ellenőrizze a táblázatban szereplő próbastatisztikákat, és indokolja a programcsomag által használt szabadságfokot!

V. Határozza meg a többszörös korrelációs és determinációs együtthatót a (8), illetve a (9) összefüggés segítségével is, és értelmezze a kapott eredményt! A számításhoz szükséges adatok az 1.a-d táblázatokban megtalálhatók. (Ugyaninnen az eredmények is leolvashatók, ezt csak ellenőrzésre használja!)

VI. Számítsa ki a parciális korrelációkat a (10)-(12) képletekkel, és ellenőrizze az 1.a-d. táblázatok segítségével! Értelmezze a kapott eredményeket, és hasonlítsa össze őket a páronkénti (teljes) korrelációkkal, amelyek szintén az 1.a-d táblázatokban találhatók!

Megoldás:

I. A kétváltozós lineáris regressziós modell $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$. Az 1.a táblázat harmadik oszlopában lévő együtthatókat behelyettesítve (két tizedesjegyre kerekítve): $\hat{y} = -2{,}20 + 0{,}34x_1 + 0{,}08x_2$. A $b_1 = 0{,}34$ jelentése: ugyanannyi munkában töltött év esetén, ha a munkavállaló egy évvel több tanulmányi idővel rendelkezik, akkor kb. 34 ezer Ft-tal magasabb a fizetése. A $b_2 = 0{,}08$ értelmezése: azonos tanulmányi idő esetén egy évvel több munkában töltött idő kb. 8 ezer Ft-tal több fizetést eredményez.

II. A becsült fizetés $x_1 = 11,\ x_2 = 15$ esetén: $\hat{y} = -2{,}20 + 0{,}34 \cdot 11 + 0{,}08 \cdot 15 = 2{,}74$, azaz kb. 274000 Ft.

III. Az F-próba szokásos lépései (a statisztikai szakirodalomban megtalálható a lépések jelentése, itt nem részletezzük):
1. H_0: $\beta_1 = \beta_2 = 0$,
2. H_1: $\exists i: \beta_i \neq 0$, ahol $i = 1{,}2$.
3. Szignifikanciaszint: $\alpha = 0{,}05$.
4. A próbastatisztika: $F_0 = 205{,}8$, és a hozzá tartozó valószínűség („jobbról”): $p \approx 0$.
5. Döntés: $F_{krit} < F_0$, így elutasítjuk a H_0 hipotézist.
6. Következtetés: 5%-os szignifikancia szinten legalább az egyik magyarázó változó szignifikáns.

IV. A t-próba gondolatmenete (a statisztikai alapokat itt is ismertnek tételezzük fel).
1. H_0: β_1=0, ill. H_0: $\beta_2 = 0$,
2. H_1: $\beta_1 \neq 0$, ill. H_1: $\beta_2 \neq 0$.
3. Szignifikanciaszint: $\alpha = 0{,}05$.
4. Kétoldali t-próbák $\nu = n - m - 1 = 30 - 2 - 1 = 27$ szabadságfokkal (ahol n az adatok száma, m az ismeretlenek száma). A próbastatisztikák értékei a következő képletekkel számíthatók (ellenőrizni az 1.a táblázat $t(27)$ oszlopa segítségével lehet):

$$t_{0(b_1)} = \frac{b_1 - \beta_1}{\hat{S}_{b_1}} = \frac{0{,}34337 - 0}{0{,}02116} = 16{,}1998, \qquad (14)$$

$$t_{0(b_2)} = \frac{b_2 - \beta_2}{\hat{S}_{b_2}} = \frac{0{,}07714 - 0}{0{,}007007} = 11{,}0090. \quad (15)$$

Mindkét próbastatisztikához $p \approx 0$ valószínűség tartozik (ld. 1.a vagy 1.c táblázat).

5. Döntés: Mindkét esetben $t_{krit} < t_0$, így elutasítjuk mindkét H_0 hipotézist.
6. Következtetés: 5%-os szignifikancia szinten mindkét magyarázó változó szignifikáns.

V. A többszörös determinációs együttható

$$R^2 = \frac{SSR}{SST} = \frac{47{,}00006}{50{,}08300} \approx 0{,}93844, \quad (16)$$

ahol a (8) összefüggést és az 1.b táblázat 1. oszlopát használtuk. Ugyanez kiszámítható a (9) összefüggéssel az 1.d táblázatban megadott korrelációs mátrixban szereplő páronkénti korrelációs együtthatókból is. Mindkét esetben közelítőleg ugyanazt az eredményt kapjuk, jelentése: a felírt regressziófüggvényben a tanulmányi idő és a munkában töltött évek száma együttesen a fizetés szórásnégyzetének (ingadozásának) kb. 94%-át magyarázza. Ellenőrzésül szolgálhat az 1.a táblázat fejléce, ahol ez az adat megtalálható. Az előbbi eredmény négyzetgyöke a többszörös korrelációs együttható: $R_{y.1,2} \approx 0{,}96873$. Ez azt mutatja, hogy a felírt kétváltozós lineáris modell segítségével nagyon jól leírható a munkabér, a magyarázó változók és a fizetés között kifejezetten erős kapcsolat van. A programcsomag az 1.a táblázat fejlécében az R értéket is megadja.

VI. A munkabér (y) és a tanulmányi idő (x_1) közötti parciális korreláció, vagyis az y és x_1 közötti kapcsolat szorossága a munkában eltöltött évek száma (x_2) hatásának kiszűrésével a (10) képlettel számítható ki, az ehhez szükséges páronkénti korrelációk az 1.d táblázatban találhatók:

$$r_{y1.2} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1 - r_{y2}^2) \cdot (1 - r_{12}^2)}} \approx \frac{0{,}81374 - 0{,}58320 \cdot 0{,}07246}{\sqrt{(1 - 0{,}58320^2) \cdot (1 - 0{,}07246^2)}} \approx 0{,}952215. \quad (17)$$

Látható, hogy y és x_1 között rendkívül szoros pozitív parciális korreláció van, azaz ugyanannyi munkában töltött év esetén hosszabb tanulmányi időhöz szinte mindig magasabb fizetés tartozik. Ezt az eredményt az 1.c táblázat 2. oszlopában találjuk (azért használtunk a számolás során 5 tizedesjegyet, hogy jól látható legyen a két érték egyezése). Az y és x_1 közötti páronkénti (teljes) korrelációt és parciális korrelációt összehasonlítva azt tapasztaljuk, hogy a munkában töltött időn keresztül érvényesülő közvetett hatás kiszűrésével a munkabér és a tanulmányi idő közötti kapcsolat tovább erősödött $(r_{y1} < r_{y1.2})$. Az $r_{y2.1} \approx 0,9043$ parciális korreláció a (11) képlettel az előbbiekhez hasonlóan számítható, és az eredmény az 1.c táblázatból ellenőrizhető. Eszerint y és x_2 szoros pozitív parciális korrelációban állnak egymással, a tanulmányi idő hatásának kiszűrésével a munkabér és a munkában töltött idő hossza közötti kapcsolat még erősebb lett $(r_{y2} < r_{y2.1})$.

A (12) összefüggésbe történő behelyettesítéssel $r_{12.y} \approx -0{,}8517$. Ez a parciális korrelációs együttható azt mutatja, hogy ha a fizetés összegét rögzítjük, akkor a tanulmányi idő (x_1) és a munkában töltött évek száma (x_2) közötti lineáris korrelációs együttható előjele megfordul. Ez úgy értelmezhető, hogy az adatok kb. 85%-ára igaz, hogy ugyanannyi munkabér alacsonyabb iskolázottság esetén hosszabb munkában töltött idő mellett, illetve hosszabb tanulmányi idő esetén rövidebb munkaviszonnyal valósul meg.

3.2. Mintapélda

Egy mezőgazdasági területen megfigyelték a munkaráfordítást (x_1 [ezer óra/hektár]), az öntözővíz mennyiségét (x_2 [ezer m^3 /hektár]), valamint a termés mennyiségének (y [mázsa/hektár]) alakulását. A 22 db adatra STATISTICA13 programcsomaggal többváltozós lineáris regressziós függvényt illesztettünk, és a 2.a-d. táblázatokban lévő eredményeket kaptuk ($\alpha = 0{,}05$ szignifikancia szinten dolgoztunk).

N=22	Regression Summary for Dependent Variable: termés (mázsa/hektár) (többvált_lin_regr) R= ,96567839 R2= ,93253475 Adjusted R2= ,92543315 F(2,19)=131,31 p<,00000 Std.Error of estimate: ,17523					
	b*	Std.Err. of b*	b	Std.Err. of b	t(19)	p-value
Intercept			4,274380	0,375105	11,39515	0,000000
munka (ezer óra/hektár)	0,911254	0,064003	0,536311	0,037668	14,23772	0,000000
öntözővíz (ezer m3/hektár)	0,128686	0,064003	0,628222	0,312450	2,01063	0,058774

Effect	Analysis of Variance; DV: termés (mázsa/hektár) (többvált_lin_regr)				
	Sums of Squares	df	Mean Squares	F	p-value
Regress.	8,064306	2	4,032153	131,3132	0,000000
Residual	0,583421	19	0,030706		
Total	8,647727				

Variable	Variables currently in the Equation; DV: termés (mázsa/hektár) (többvált_lin_regr)						
	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(19)	p-value
munka (ezer óra/hektár)	0,911254	0,956192	0,848406	0,866819	0,133181	14,23772	0,000000
öntözővíz (ezer m3/hektár)	0,128686	0,418857	0,119811	0,866819	0,133181	2,01063	0,058774

Variable	Correlations (többvált_lin_regr)		
	munka (ezer óra/hektár)	öntözővíz (ezer m3/hektár)	termés (mázsa/hektár)
munka (ezer óra/hektár)	1,000000	0,364940	0,958217
öntözővíz (ezer m3/hektár)	0,364940	1,000000	0,461239
termés (mázsa/hektár)	0,958217	0,461239	1,000000

7. a-d. táblázatok

Feladatok, kérdések:

I. Írja fel a regressziós egyenletet és értelmezze a paramétereket!

II. Mekkora a modell alapján becsült termés egy hektáron 3000 munkaóra és 1500 m^3öntözővíz esetén ($x_1 = 3,\ x_2 = 1{,}5$)?

III. Tesztelje a paramétereket F-próbával 5%-os szignifikancia szinten (globális tesztelés), használja a programcsomag eredmény-táblázatát!

IV. Tesztelje a paramétereket t-próbával 5%-os szignifikancia szinten (parciális tesztelések)!

V. Számítsa ki a többszörös korrelációs és determinációs együtthatót, majd értelmezze a kapott eredményt!

VI. Számítsa ki a parciális korrelációkat, értelmezze a kapott eredményt, és hasonlítsa össze a teljes korrelációkkal!

VII. Hozzon döntést arról, hogy érdemes-e a felírt modellen változtatni, és ha igen, akkor az $\bar{R}^2$ (szabadságfokkal korrigált R^2) értékek kiszámítása alapján javasoljon másik modellt!

Megoldás:
Csak a IV., VI. és VII. kérdésre térünk ki részletesen, a többi esetén a 3.1. Mintapélda megoldása nyújthat segítséget.

IV. A t-próba menete is hasonló a 3.1. Mintapélda IV. pontjához, de az egyik próbastatisztikához tartozó valószínűség értéke, és emiatt a döntés és következtetés más. A 2.c táblázatból leolvasható, hogy a $t_{0(b_1)} \approx 14{,}24$ próbastatisztikához $p \approx 0$ valószínűség tartozik, emiatt a munkaóra (x_1) magyarázó változó 5%-os szignifikancia szinten szignifikáns (pirossal jelöli a programcsomag), de a $t_{0(b_2)} \approx 2{,}01$ próbastatisztikához tartozó valószínűség $p \approx 0{,}06$, tehát az öntözővíz mennyisége (x_2) magyarázó változó nem szignifikáns ugyanezen a szignifikancia szinten (fekete színnel jelöli az eredménytáblázat).

VII. A programcsomag által megadott korrelációs mátrixban lévő elemekkel számolunk (2.d táblázat). A termés (y) és a munkaórák (x_1) közötti parciális korreláció az öntözővíz (x_2) hatásának kiszűrésével $r_{y1.2} \approx 0{,}9562$, ami erős pozitív korrelációt mutat. Ha ezt az eredményt az *y* és x_1 közötti páronkénti (teljes) korrelációval összehasonlítjuk, akkor látható, hogy az öntözővíz mennyisége által érvényesülő közvetett hatást kiszűrve a termés és a munkaráfordítás közötti kapcsolat nem változott ($r_{y1} = 0{,}9582$; $r_{y1.2} \approx 0{,}9562$). Az $r_{y2.1} \approx 0,4189$, ami azt mutatja, hogy a munkaórák (x_1) hatását kiszűrve a terméshozam (y) és az öntözés (x_2) között az gyenge pozitív korrelációt tapasztalunk. Ezáltal a terméshozam és az öntözés közötti kapcsolat kis mértékben gyengült ($r_{y2} = 0{,}46123$; $r_{y1.2} \approx 0,4189$). Az x_1 és x_2 közötti parciális korreláció $r_{12,y} \approx -0{,}3035$, azaz, ha rögzítjük a termés mennyiségét, akkor x_1 és x_2 között csekély mértékű negatív korreláció van. Tehát az esetek kb. 30%-ára igaz az, hogy ugyanaz a terméshozam kevesebb öntözés és több munkaóra, vagy több öntözés és kevesebb munkaóra mellett tapasztalható.

VIII. Kiszámítható, hogy az egyes magyarázó változók, illetve az együttes hatásuk hogyan járul hozzá az R^2 értékhez. Az x_1 hozzájárulása R^2-hez (az R^2 értéke a 2.a táblázat fejlécében található):

$$R^2 - r_{y2}^2 = 0{,}9325 - (0{,}4612)^2 = 0{,}7198. \quad (18)$$

Az x_2 hozzájárulása R^2-hez:

$$R^2 - r_{y1}^2 = 0{,}9325 - (0{,}9582)^2 = 0{,}0144. \quad (19)$$

Az R^2-ből fennmaradó rész az x_1 és x_2 együttes hatása, ahogy ezt a 3. táblázat mutatja.

magyarázó változó	magyarázó változó hozzájárulása R^2 -hez
x_1	0,7198
x_2	0,0144
$x_1$és x_2 (együttes hatás)	0,1983
Összesen	$\mathbf{0,9325}$

8. táblázat

Látható, hogy ha az egyes változók hatásához hozzáadjuk az együttes hatást, visszakapjuk a páronkénti determinációs együtthatókat:

$$r_{y1}^2 = 0{,}7198 + 0{,}1983 = 0{,}9181, \qquad r_{y1} = 0{,}9582, \tag{20}$$

$$r_{y2}^2 = 0{,}0144 + 0{,}1983 = 0{,}2127, \qquad r_{y2} = 0{,}4612. \tag{21}$$

Arról, hogy érdemes-e esetleg elhagyni valamelyik változót, és ha igen, akkor melyiket, a korrigált R^2 értékek kiszámítása segítségével hozhatunk döntést. Két változóra a (13) képletbe történő behelyettesítéssel az alábbi eredményt kapjuk (adatok száma $n = 22$, változók száma $m = 2$), ezt a STATISTICA13 programcsomag is megadja a 2.a táblázat felső fejlécében „Adjusted R2" néven:

$$\bar{R}^2 = 1 - \frac{n-1}{n-m-1} \cdot (1 - R^2) = 1 - \frac{22-1}{22-2-1} \cdot (1 - 0{,}9325) = 0{,}9254 = \bar{R}_{12}^2. \tag{22}$$

A (22) korrigált determinációs együttható jelölésére itt most használjunk $\bar{R}_{12}^2$-t, hogy megkülönböztessük az alábbiakban kiszámolásra kerülő további korrigált determinációs együtthatóktól. Ha csak az x_1 (munkaóra) változót hagyjuk meg ($n = 22, m = 1$), akkor a korrigált determinációs együttható (jelölésére megkülönböztetésül bevezetjük az $\bar{R}_1^2$-t):

$$\bar{R}^2 = 1 - \frac{n-1}{n-m-1} \cdot \left(1 - r_{y1}^2\right) = 1 - \frac{22-1}{22-1-1} \cdot (1 - 0{,}9181) = 0{,}9140 = \bar{R}_1^2. \tag{23}$$

Hasonló gondolatmenettel, ha csak az x_2 (öntözővíz) változót hagyjuk meg ($n = 22, m = 1$), akkor az adjusztált R^2 (itt az $\bar{R}_2^2$ megkülönböztető jelölést vezetjük be):

$$\bar{R}^2 = 1 - \frac{n-1}{n-m-1} \cdot \left(1 - r_{y2}^2\right) = 1 - \frac{22-1}{22-1-1} \cdot (1 - 0{,}2127) = 0{,}1733 = \bar{R}_2^2. \tag{24}$$

A modellben használt változók számáról, illetve arról, hogy melyiket hagyjuk meg, úgy is dönthetünk, hogy a fentiek közül azt az esetet választjuk, amelyikre $\bar{R}^2$ értéke maximális. (Megj.: A képletekből látszik, hogy ha ugyanannyi a változók száma, akkor az az adjusztált R^2 lesz nagyobb, amelyikben nagyobb páronkénti determinációs együtthatóval számolunk, a kisebbhez tartozót emiatt felesleges kiszámolni.) A (22), (23) és (24) eredmények növekvő sorrendben:

$$\bar{R}_2^2 < \bar{R}_1^2 < \bar{R}_{12}^2\,, \tag{25}$$

tehát érdemes a kétváltozós modellt megtartani. Az viszont megfontolandó, hogy mivel az $\bar{R}_1^2$ és $\bar{R}_{12}^2$ között alig van különbség, és ha a kétváltozós függvény helyett az x_1 egy magyarázó változós regressziós modellt használjuk, akkor ez összhangban lesz a IV. kérdésre adott válasszal, amely szerint csak az x_1 magyarázó változó szignifikáns. Ilyen esetben a szakmai szempontok (konzultáció az adatokat szolgáltató szakemberekkel) is fontos lehet a végső modell kialakításában. Sokszor előfordul, hogy egy statisztikai értelemben kevésbé megbízható függvény jobban modellezi a vizsgált folyamatot, mint a másik, „jobb" statisztikai mutatószámokkal rendelkező. Egy ismételt mintavétel (másik adatsor) vizsgálta is célravezető lehet, de erre a gyakorlatban legtöbbször nincs lehetőség.

Összefoglalás

Az alkalmazott statisztika oktatása kiemelt fontosságúvá vált napjainkban. A kevés óraszám és a hallgatók megfelelő előképzettségének hiánya miatt nagyon fontos az egyszerű megközelítés és a szemléletes példák használata. A fenti két mintafeladat a többváltozós lineáris regressziószámítás néhány olyan apró részletére világít rá, amelyekkel konkrét problémák megoldása során találkozhatnak hallgatóink. A témakör tárgyalásakor természetesen még számtalan egyéb kérdés felmerülhet, amelyekre itt nem adtunk választ. Nem foglalkoztunk például a többváltozós lineáris regresszió alkalmazhatósági feltételeinek vizsgálatával, amelytől a gyakorlatban nem tekinthetünk el. Nagyon fontos ismételten hangsúlyozni, hogy gyakorlati/szakmai szempontból nem mindig azok a legjobb modellek, amelyeket a megfelelő statisztikai mutatók kiszámításával a legmegfelelőbbnek gondolunk. Előfordulhat, hogy a statisztikai értelemben gyengébbnek értékelt modell az adott szakterületen sokkal jobban leírja a vizsgált folyamatot, mint amelyiket a statisztikai mutatók alapján választanánk. A statisztikai modellalkotás éppen ettől szép és izgalmas feladat.

Irodalomjegyzék

[1] **Bolla M., Krámli A., Nagy-György J.:** Többváltozós statisztikai módszerek. Szegedi Tudományegyetem 2013. http://eta.bibl.u-szeged.hu/1327/1/tobbvaltozos_statisztikai_modszerek.pdf

[2] **Domán Cs.:** Többváltozós Korreláció- és regressziószámítás. Miskolci Egyetem Gazdaságtudományi Kar, Oktatási segédlet 2005. https://docplayer.hu/10101458-Tobbvaltozos-korrelacio-es.html

[3] **Kis-Tóth L.,** Lengyelné Molnár T., Tóthné Parázsó L.: Statisztikai programrendszerek. Eszterházy Károly Főiskola, Eger 2013. https://mek.oszk.hu/14100/14139/pdf/14139.pdf

[4] **Korpás A.:** Általános statisztika. Nemzeti tankönyvkiadó, 1996.

[5] **Mundruczó Gy.:** Alkalmazott regressziószámítás. Akadémiai Kiadó, Budapest 1981.