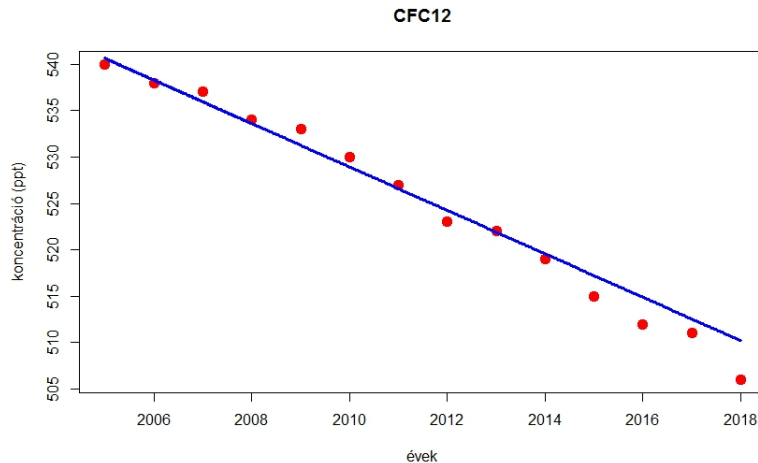


Matematikai statisztika előadás, 10. hét, április 22.
Lineáris modell egy- és többváltozós esetben

1. Lineáris regresszió

A lineáris regresszió során az a célunk, hogy egy $f(y) = x$ függvényt, melynek értékét néhány x_1, x_2, \dots, x_n pontban ismerjük, a „lehető legjobban” közelítsünk egy egyenessel (1. ábra). Ehhez szorosan fog kapcsolódni az együtthatók becslése az úgynevezett lineáris modellben.



1. ábra. A CFC-12 (freon) gáz koncentrációja az Antarktison és az adatokra illesztett egyenes (forrás: ESRL, USA)

Egyenes illesztése a **legkisebb négyzetek módszerével**. Adottak tehát $(x_1, y_1), \dots, (x_n, y_n)$ pontok. A leggyakrabban használt módszer esetén az illesztés hibája az egyenes által megadott $ax_i + b$ értékeknek és a mért y_i értékeknek a különbségének négyzetösszege. Az, hogy ez milyen a, b számokra a legkisebb, egy többváltozós szélsőérték-keresési feladat, melynek megoldását az alábbi állítás adja meg.

1.1. Állítás (Lineáris regresszió). *Legyenek $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ adott számpárok. Azokat az a és b együtthatókat keressük, melyre a*

$$h^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

mennyiség minimális. Ennek megoldása:

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{k=1}^n (x_k - \bar{x})^2}; \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

A példában: $\hat{a} = -2,63$; $\hat{b} = 5807,7$ (a b együttható neve: intercept)

1.1. Lineáris modell: példa R-ben

A reziduálisok az $y_i - (ax_i + b)$ hibák, ezekre vonatkozik egy összefoglaló statisztika.

```
> cfc12<-c(540, 538, 537, 534, 533, 530, 527, 523, 522, 519, 515, 512, 511, 506)
> ev<-c(seq(from=2005, to=2018, by=1))
> summary(lm(cfc12 ev))
Call:  lm(formula = cfc12 ev)
```

```

Residuals:      Min       1Q   Median       3Q      Max
               -1.8571  -0.8736  0.2088  0.8709  1.6483

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5807.73626   159.19290    36.48  1.15e-13 ***
ev           -2.62637     0.07914   -33.19  3.55e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:  1.194 on 12 degrees of freedom
Multiple R-squared:  0.9892, Adjusted R-squared:  0.9883
F-statistic:  1101 on 1 and 12 DF, p-value:  3.554e-13

```

2. Lineáris modell egyváltozós esetben

A lineáris modellben az az elképzelésünk, hogy az Y valószínűségi változó az X -ből úgy kapható, hogy vesszük X -nek egy lineáris függvényét ($aX+b$), majd egy normális eloszlású hibát hozzáadunk. Az (X, Y) pár eloszlásából vehetünk n elemű mintát, itt a párok egymástól már függetlenek. Ugyanakkor azért lesz egy statisztikai feladat, mert sem az a , sem a b együtthatókat nem ismerjük, sem pedig a hozzáadott hiba szórásának nagyságát. Az első feladat tehát ezen ismeretlen paraméterek becslése lesz a megfigyelt minta alapján.

2.1. Definíció (Lineáris modell). *Legyenek $X_1, X_2, \dots, X_n, Y_1, \dots, Y_n$ valószínűségi változók, és tegyük fel, hogy valamely a, b valós számokra*

$$Y_i = aX_i + b + \varepsilon_i,$$

ahol $\varepsilon_1, \dots, \varepsilon_n$ független $N(0, \sigma^2)$ normális eloszlású valószínűségi változók. Az így kapott (X_i, Y_i) párok együttes eloszlását lineáris modellnek nevezzük.

Az X_i valószínűségi változókat magyarázó változóknak, az ε_i valószínűségi változókat hibának szokták nevezni.

2.1. Becslések a lineáris modellben

A maximumlikelihood-módszer alkalmazható a lineáris modell együtthatóinak becslésére. A kapott eredmények megegyeznek a lineáris regresszióban a legkisebb négyzetek módszerével kapott egyenessel az együtthatók esetében. A szórás is ismeretlen paraméter, erre is adhatunk becslést.

A becslések szórása is kiszámítható: ez azt jelenti, hogy a kiszámított becslésnek (mely a minta függvénye, ezért véletlen), mint valószínűségi változónak mennyi a szórása. Ez természetesen a σ paramétertől függ. Ha a becslések szórásának nagyságrendjére vagyunk kíváncsiak, ebben a képletben σ helyére a $\hat{\sigma}$ maximumlikelihood-becslést írhatjuk.

2.1. Állítás. *A lineáris modellben az a, b együtthatók maximumlikelihood-becslése a következőképpen írható:*

$$\hat{a} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{k=1}^n (X_k - \bar{X})^2}; \quad \hat{b} = \bar{Y} - \hat{a}\bar{X}.$$

Továbbá, ezek a becslések torzítatlan becslései az a és b paramétereknek:

$$\mathbb{E}(\hat{a}) = a; \quad \mathbb{E}(\hat{b}) = b.$$

A hiba szórásának becslése:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2.$$

A becslések szórása:

$$D(\hat{a}) = \frac{\sigma}{\sqrt{\sum_{j=1}^n (X_j - \bar{X})^2}}; \quad D(\hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

A lineáris modell becslései érzékenyek a kiugró értékekre, így azokat a becslés előtt érdemes lehet eltávolítani.

2.2. Előrejelzés a lineáris modellben

2.2. Állítás. Legyen x^* adott szám. A lineáris modellből kapott előrejelzés az Y véletlen folyamat x^* pontban felvett értékére:

$$\hat{a}x^* + \hat{b}.$$

Az előrejelzés szórása:

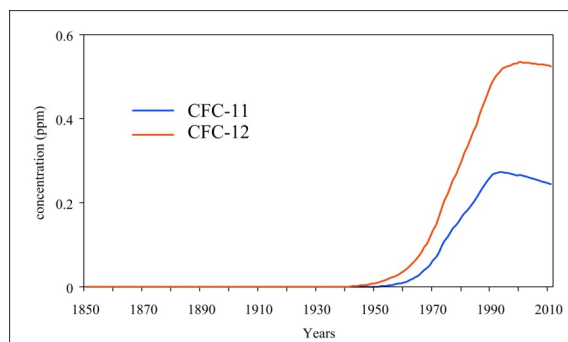
$$D(\hat{a}x^* + \hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

Az előrejelzés szórásának becslésekor a σ értéket gyakran $\hat{\sigma}$ -val helyettesítik, ez információt ad a becslés pontosságáról. Minél távolabbi pontra készítjük az előrejelzést, annál nagyobb lesz a szórás.

A példában: előrejelzés $x^* = 2019$ -re:

$$\hat{a} \cdot x^* + \hat{b} = -2,63 \cdot 2019 + 5807,7 = 497,7.$$

Ugyanakkor, ahogy a 2. ábra is mutatja, az, hogy egy folyamat egy rövidebb szakaszon jól közelíthető a lineáris modellel, nem jelenti, hogy nagyobb skálán is érvényes a közelítés (az emelkedés az ipari tevékenység következménye, a csökkenés az adott gázok gyártásának tiltásának következménye – bár a tiltás ellenére a gyártás nem állt le teljesen).



2. ábra. A CFC-11 és CFC-12 (freon) gáz koncentrációja (forrás: elte.promt.hu)

2.3. Reziduálisok és R^2

A különböző adatsorokra a lineáris modell természetesen különböző mértékben illeszkedik. Az illeszkedés pontosságát elsősorban a reziduálisokon keresztül, vagyis a közelítő egyenes és a megfigyelt érték különbségéből érthetjük meg.

Reziduálisok: $Y_i - \hat{a}X_i - \hat{b}$ (ezeknek a négyzetösszege minimális)

A teljes ingadozás (total sum of squares): $\sum_{j=1}^n (Y_j - \bar{Y})^2$.

Ezt összehasonlíthatjuk a reziduális négyzetösszeggel (residual sum of squares):

$$\sum_{j=1}^n (Y_j - \hat{a}X_j - \hat{b})^2.$$

A kettő hányadosát 1-ből levonva kapjuk az úgynevezett megmagyarázott ingadozás részarányát:

$$R^2 = 1 - \frac{\sum_{j=1}^n (Y_j - \hat{a}X_j - \hat{b})^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2}.$$

Ugyanis a reziduális négyzetösszeg és a teljes ingadozás hányadosa azt mutatja, hogy az Y tapasztalati szórásnégyzetéből milyen rész adódik az illesztett érték és a megfigyelt érték különbségéből. Az egyből levont érték tehát azt mutatja, hogy a modell jó illeszkedése esetén milyen szórásnégyzet adódna. A reziduális négyzetösszeget egy másik alakba írva látható, hogy az így kapott R^2 valójában a két minta tapasztalati korrelációs együtthatójának négyzete, ezt használjuk definíciónak.

2.2. Definíció. A megmagyarázott ingadozás részaránya (coefficient of determination):

$$R^2 = \frac{[\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})]^2}{[\sum_{k=1}^n (X_k - \bar{X})^2][\sum_{k=1}^n (Y_k - \bar{Y})^2]}.$$

Az R^2 értéke 0 és 1 közé esik.

Értelmezés: minél közelebb van 1-hez, annál inkább jó közelítést ad a lineáris modell. De ez nem minden szempontnak megfelelő mérőszám, és fordítva nem is feltétlenül igaz a következtetés. Például az R^2 érzékeny a kiugró értékekre, néhány kiugró esetén R^2 lecsökken. Vagyis az R^2 -ből nem tudjuk jól eldönteni, hogy a „tipikus” értékek sem illeszkednek jól, vagy esetleg néhány pont kivételével lényegében jó az illeszkedés.

A példában: $R^2 = 0,98$, vagyis jól illeszkedik a lineáris modell.

Az R kódban megadott adjusted R^2 : nem csak a reziduálisokat veszi figyelembe, hanem azt is, hogy hány paramétert használtunk (ennek többváltozós esetben van nagyobb jelentősége).

2.4. Konfidenciaintervallumok

Amikor az együtthatókat megbecsüljük, nem csak a becslést, hanem konfidenciaintervallumot is adhatunk. Ugyanis a becslés eloszlása t -eloszlás, és ez alapján a t -próba kritikus értékeinek segítségével adhatunk konfidenciaintervallumot.

$1 - \alpha$ megbízhatósági szintű konfidenciaintervallum a -ra:

$$\left(\hat{a} - t_{n-2,\alpha} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}, \hat{a} + t_{n-2,\alpha} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \right),$$

ahol $t_{n-2,\alpha}$ az $f = n - 2$ szabadsági fokú α szignifikanciaszintű kétoldali t -próba kritikus értéke.

Az x^* pontban az előrejelzett érték becslése $\hat{a} \cdot x^* + \hat{b}$.

$1 - \alpha$ megbízhatósági szintű konfidenciaintervallum $ax^* + b$ -re, azaz az x^* -ban felvett érték várható értékére:

$$\left(\hat{a}x^* + \hat{b} \pm t_{n-2,\alpha} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} \right).$$

Minél távolabbi pontban készítjük az előrejelzést, annál hosszabb lesz a konfidenciaintervallum.

2.5. Az egyenes meredekségére vonatkozó próbák

A lineáris modell fő egyenlete: $Y_i = aX_i + b + \varepsilon_i$. Állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan eltér 0-tól? A lineáris modellen belül ez a kérdés felel meg annak, hogy van-e egyáltalán összefüggés a két vizsgált mennyiség között.

$$H_0: a = 0 \quad H_1: a \neq 0$$

A nullhipotézis teljesülése esetén csak $Y_i = b + \varepsilon_i$, vagyis ez normális eloszlású, X_j eloszlásáról azonban nem volt feltételünk. Ezzel együtt, ha a nullhipotézis igaz, akkor az alábbi mennyiség t -eloszlású $f = n - 2$ szabadsági fokkal, ezért erre t -próbát végezhetünk.

Kétoldali t -próbát végezhetünk az alábbi próbastatisztikával és $f = n - 2$ szabadsági fokkal:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}.$$

Ha $|t| > t_{n-2,\alpha}$, azaz $p < \alpha$, akkor elutasítjuk H_0 -t, az egyenes meredeksége szignifikánsan eltér 0-tól (itt $t_{n-2,\alpha}$ az α szignifikanciaszintű $f = n - 2$ szabadsági fokú kétoldali t -próba kritikus értéke).

Ha $|t| \leq t_{n-2,\alpha}$, azaz $p \geq \alpha$, akkor elfogadjuk H_0 -t, az egyenes meredeksége nem tér el szignifikánsan 0-tól.

A korábbi példában (az 1. ábra):

$$t = -33,19; \quad \alpha = 0,05; \quad n = 14; \quad f = n - 2 = 12; \quad c_{\text{krit}} = 2,19.$$

Mivel $|t| = 33,19 > c_{\text{krit}} = 2,19$, elutasítjuk a nullhipotézist, az egyenes meredeksége **szignifikánsan eltér 0-tól**. A p -érték: $p = 3,6 \cdot 10^{-13} < 0,05 = \alpha$.

A t és a p is kiolvasható az R-kódból (1.1. példa).

Egy másik kérdés: állíthatjuk-e, hogy az egyenes meredeksége szignifikánsan nagyobb 0-nál? A modellen belül ez jelenti azt, hogy a vizsgált mennyiségek között pozitív irányú összefüggés van, minél nagyobb az X , annál nagyobb az Y értéke is (természetesen a fordított irányú összefüggés is hasonlóképpen tesztelhető lenne).

$$H_0: a \leq 0 \quad H_1: a > 0$$

Továbbra is használhatjuk, hogy $a = 0$ esetén az alábbi próbastatisztika t -eloszlású $f = n - 2$ szabadsági fokkal.

Egyoldali t -próbát végezhetünk az alábbi próbastatisztikával és $f = n - 2$ szabadsági fokkal:

$$t = \hat{a} \frac{\sqrt{(n-2) \sum_{i=1}^n (X_i - \bar{X})^2}}{\sqrt{\sum_{i=1}^n (Y_i - \hat{a}X_i - \hat{b})^2}}.$$

Ha $t > \bar{t}_{n-2,\alpha}$, azaz $p < \alpha$, akkor elutasítjuk H_0 -t, az egyenes meredeksége szignifikánsan több 0-nál (itt $\bar{t}_{n-2,\alpha}$ az α terjedelmű $f = n - 2$ szabadsági fokú egyoldali t -próba kritikus értéke α szignifikanciaszint mellett).

Ha $t \leq \bar{t}_{n-2,\alpha}$, azaz $p \geq \alpha$, akkor elfogadjuk H_0 -t, az egyenes meredeksége nem szignifikánsan pozitív.

3. Többváltozós lineáris regresszió (multiple linear regression)

Természetesen az is elképzelhető, hogy az Y mennyiség nem egy, hanem több változónak a lineáris függvénye, valamilyen hiba hozzáadásával.

Az Y változót fejezzük ki az X_1, \dots, X_p valószínűségi változók lineáris függvényeként, de az együttthatókat ismeretlennek tekintjük ($X_{i,p} \equiv 1$ lehet a konstans tag):

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i,$$

ahol ε_i független $N(0, \sigma^2)$ normális eloszlású valószínűségi változók.

Például: $X_{i,1}$ az év, $X_{i,2}$ a CFC-12 kibocsátás az i . mérésnél, és $X_{i,3} = b$ egy konstans tag (vagyis az $X_{i,1}$ évben Y a koncentráció, ami az időnek és a kibocsátásnak is a függvénye). Ekkor a lineáris modell:

$$\begin{aligned} Y_1 &= a_1 X_{1,1} + a_2 X_{1,2} + b + \varepsilon_1; \\ Y_2 &= a_1 X_{2,1} + a_2 X_{2,2} + b + \varepsilon_2; \\ &\dots \\ Y_n &= a_1 X_{n,1} + a_2 X_{n,2} + b + \varepsilon_n. \end{aligned}$$

Itt a_1 az, hogy milyen együtthatóval számít az év, a_2 az, hogy milyen együtthatóval számít a kibocsátás, b a konstans tag, ε pedig a véletlen hiba, amely évről évre független, azonos eloszlású.

Vektoros formában, visszatérve az általános esetre: $\underline{Y} = X\beta + \underline{\varepsilon}$, ahol X az $X_{i,j}$ megfigyelésekből készített mátrix, és $\beta = (a_1, a_2, \dots, a_p)^T$ az együtthatók oszlopvektora.

Ezután az a_1, \dots, a_p együtthatók becslése (torzítatlan, és ugyanaz a legkisebb négyzetek módszerével és maximumlikelihood-módszerrel):

$$\hat{\beta} = (X^T X)^{-1} X^T \underline{Y}.$$

Az egyváltozós esethez képest most a konstans tagot másképpen vettük figyelembe, ezt is egy valószínűségi változónak tekintettük, az X vektor része. Ennek következménye, hogy az együtthatók becslésében nem kell levonni az átlagot. A konstans tag nélkül (vagyis ha $b = 0$ lenne) ugyanazt kapnánk vissza, ha $p = 1$, hiszen ekkor $X^T X = \sum_{j=1}^n X_j^2$, és $X^T Y = \sum_{j=1}^n X_j Y_j$.

A megmagyarázott ingadozás részaránya:

$$R^2 = \frac{(X^T X)^{-1} X^T \underline{Y}}{\underline{Y}^T \underline{Y}}.$$

Ez a mennyiség azonban nem csak például a kiugró értékekre érzékeny, hanem nem veszi megfelelően figyelembe a p -től való függést, vagyis a becsült paraméterek számát. Ez azért okoz problémát, mert túl sok becsült paraméter esetén gyakran megfigyelhető a túltanulás (overfitting) jelensége, amikor a paraméterek becslései jobban függnek a megfigyelések véletlen hibából adódó komponensétől, mint a megfigyelt rendszer valódi szerkezetétől, és ezért valójában nem lesz jó a modellillesztés és az előrejelzés sem. Ezért az R^2 -nek az alábbi módosított (adjusted) változata is gyakran használt:

$$\tilde{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}.$$

Itt tehát $p = 0$ (ez nem is egy valódi modell) esetén az eredeti R^2 -et kapjuk vissza, és ha a megfigyelések száma nagy, p pedig kicsi, akkor szintén közel van a módosított érték az eredetihez. Ha azonban például $n = 100$ -as mintaelemszám mellett $p = 10$ -et használunk, akkor az R^2 -nek az 1-től való eltérése, amit figyelni szoktunk, nagyjából 10%-kal megnő, jelezve, hogy a mintaelemszámhoz képest túl sok lehet a paraméter.

3.1. Hipotézisvizsgálat a lineáris modellben

Ekkor is megfelelő próbat statisztikával t -próbbával tesztelhetők az $a_i = 0$ hipotézisek. Ennek jelentősége a következő. A modell eredeti felírásakor kiválasztottunk p mennyiséget, melyről feltételeztük, hogy ezek olyan értelemben meghatározzák Y viselkedését, hogy egy lineáris függvényükhöz már csak egy véletlen hiba adódik hozzá. Az a_i együttható mondja meg, hogy az i . mennyiség milyen súllyal szerepel, vagyis ha például

$$Y_j = 5X_{j,1} + 3X_{j,2} + 0,02X_{j,3} + \varepsilon_j,$$

és az $X_{j,i}$ valószínűségi változók várható értéke és szórása nagyjából megegyezik, akkor Y_j -re az első mennyiség hatása a legjelentősebb, a második is ugyanennyire fontos, ugyanakkor a harmadik mennyiség sokkal kisebb súllyal szerepel, felmerülhet, hogy ezt ne is vegyük figyelembe a modellezés során. Vagyis megtehetjük, hogy az elsőként felépített lineáris modellt leszűkítjük csak azokra a változókra, amiknél az együttható szignifikánsan különbözik 0-tól, vagyis aminek jelentős hatása van a megfigyelt Y mennyiségre, újra megbecsüljük a paramétereket, és csak ezzel a leszűkített modellel számolunk tovább, feltéve, hogy ott is még jó illeszkedés kapható. Ennek az az előnye, hogy elkerülhetjük az előző részben említett túltanulás jelenségét, amikor túl sok a becsült paraméter, és nem az illesztés nem tükrözi a megfigyelt rendszer valódi szerkezetét.

A fent megfogalmazott hipotézisnél általánosabb feladatot oldunk meg, így több együttható 0 volta is egyszerre tesztelhető például.

Többváltozós lineáris modell ($X_{i,p}$ lehet a konstans tag):

$$Y_i = a_1 X_{i,1} + a_2 X_{i,2} + \dots + a_p X_{i,p} + \varepsilon_i, \text{ azaz } \underline{Y} = X\beta + \underline{\varepsilon}.$$

Legyen H olyan $r \times p$ méretű mátrix, aminek a rangja r (itt $r < p$). Ekkor az alábbi hipotézisvizsgálati feladatot tekintjük:

$$H_0 : H\beta = 0 \qquad H_1 : H\beta \neq 0.$$

Ha például H egy sora a j . egységvektor, akkor βH egy eleme az a_j együttható, a nullhipotézis az $a_j = 0$ -t jelenti. Ha H -t különböző egységvektorból állítjuk össze, akkor tudjuk több együttható 0 voltát egyszerre tesztelni.

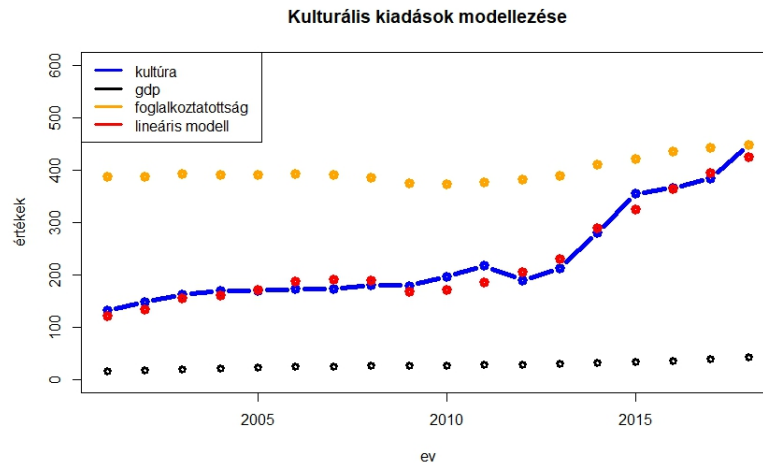
A valószínűséghányados próba (ami a Neyman–Pearson-lemmában szerepelt) próbastatisztikája:

$$F = \frac{(\underline{Y} - X\beta^*)^T(\underline{Y} - X\beta^*) - (\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})}{(\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})},$$

ahol β^* a β becslése a $H\beta = 0$ feltétel mellett a redukált lineáris modellben (a fenti példában ez annak felel meg, amikor bizonyos magyarázó változókat nem használhatunk).

Ha H_0 igaz, akkor $F \cdot (n - p)/r$ eloszlása F -eloszlás $(r, n - p)$ szabadsági fokkal. Ezért H_0 -t elutasítjuk, ha F értéke nagyobb ennek az F -próbának a kritikus értékénél, különben elfogadjuk H_0 -t.

3.2. Többváltozós lineáris regresszió: példa



3. ábra. A költségvetés kultúrára szánt kiadásai és lineáris modell a gdp, a foglalkoztatottság és az évszám figyelembevételével (az ábrán minden mennyiség valamilyen konstansszorosra látható, a valódi nagyságrendek eltérőek)

Az alábbi adatsorok (forrás: KSH) Magyarország kultúrára fordított költségvetési összegeit (milliárd forintban), gdp-jét (milliárd forintban), illetve a Magyarországon foglalkoztatottak számát (ezer fő) mutatják 2001-2018-ig. Olyan lineáris modellt építünk, ahol Y a kultúrára fordított éves kiadás, legyen X_1 az évszám, X_2 a gdp, X_3 a foglalkoztatottak száma, $X_4 \equiv 1$ a konstans tag:

$$Y = a_1X_1 + a_2X_2 + a_3X_3 + a_4 + \varepsilon,$$

ahol $\varepsilon \sim N(0, \sigma^2)$ normális eloszlású hiba. A magyarázó változók nagyságrendje eltérő, de a becslés szempontjából ez nem baj, ha valamelyik változót átskálázzuk, a becsült együttható is átskálázódik, de a például R^2 változatlan marad.

```
kultura<-c(132, 148, 163, 170, 170, 173, 173, 181, 179, 197, 217, 190, 213, 281, 355, 366, 384, 448)
```

```
ev<-2001:2018
```

```
gdp<-c(15399, 17434, 19134, 21078, 22549, 24316, 25701, 27217, 26458, 27269, 28371, 28848, 30290, 32694, 34785, 35896, 38835, 42662)
```

```
fogl<-c(3868, 3871, 3922, 3900, 3902, 3928, 3902, 3848, 3749, 3732, 3759, 3827, 3893, 4101,
4211, 4352, 4421, 4470)
summary(lm(kultura ev + gdp + fogl))
Call: lm(formula = kultura ev + gdp + fogl)
Residuals:
Min 1Q Median 3Q Max
-22.1858 -14.4101 0.9424 10.3284 27.5662
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.394e+03 9.580e+03 -0.876 0.396
ev 3.801e+00 4.788e+00 0.794 0.441
gdp 3.939e-03 3.896e-03 1.011 0.329
fogl 2.201e-01 3.351e-02 6.568 1.25e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 18.46 on 14 degrees of freedom
Multiple R-squared: 0.9683, Adjusted R-squared: 0.9615
F-statistic: 142.6 on 3 and 14 DF, p-value: 9.94e-11
A becslések alapján az illesztett modell (ez látható a 3. ábrán):
```

$$Y = 3,8X_1 + 0,0039X_2 + 0,22X_3 - 8394 + \varepsilon.$$

Az R^2 értéke 1-hez viszonylag közele, mondhatjuk, hogy jól illeszkedik a modell. A t -próba egyedül a foglalkoztatottak számánál mutat 0-tól való szignifikáns eltérést. Ha most csak ezt a változót tartjuk meg, és így illesztünk modellt:

```
> summary(lm(kultura fogl))
```

Ekkor az illesztett modell ez lenne: $Y = 0,37X_3 - 1261$, és $\tilde{R}^2 = 0,83$, ez tehát kevésbé jó illeszkedést jelent az előzőhöz képest.

Házi feladat április 29., 8:15-ig A KSH adatainak segítségével (<http://www.ksh.hu/stadat>) illesszünk lineáris modellt a kormányzati alrendszerek vagy a háztartások egészségügyi kiadásainak alakulására (2003-2017, éves adatok, 2.4.1. pont). A magyarázó változók száma legalább három legyen (az egyik lehet az évszám, a fenti példához hasonlóan), és ezeket is lehet például az általános gazdasági mutatók vagy más, az oldalon feldolgozott mérőszámok alapján választani. Keressünk úgy magyarázó változókat, hogy az \tilde{R}^2 értéke legalább 0,8 legyen. (A táblák a honlapról letölthetők excel-formátumban.)

Az adatokból az utolsó évet hagyjuk ki, így becsüljük meg a paramétereiket, majd a magyarázó változók utolsó évben mért adatainak segítségével készítsünk előrejelzést az utolsó év egészségügyi kiadásaira. Hasonlítsuk ezt össze a valódi adattal, számítsuk ki az abszolút és a relatív hibát is (a két érték különbségét, illetve ennek arányát a valódi értékhez képest).