

A normalitás tesztelése: Lilliefors-próba (11. előadás)

A normális eloszlás paramétereit először meg kell becsülni az adatok alapján.

H_0 : a minta normális eloszlásból származik (valamilyen m, σ paraméterekkel)

H_1 : a minta eloszlása nem normális eloszlás

Legyen \bar{X} a mintaátlag, s_n^* a korrigált tapasztalati szórás, \hat{G} pedig az m várható értékű és σ szórású normális eloszlás eloszlásfüggvénye: $\hat{G}(t) = \Phi((t - \bar{X})/s_n^*)$. Ekkor a próbastatisztika (ugyanaz, mint a Kolmogorov–Szmirnov-próbánál):

$$D_n = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}(t)|.$$

Ha $D_n > \bar{D}_{\text{krit}}$ (vagy $p < \alpha$), akkor elutasítjuk H_0 -t, a minta eloszlása szignifikánsan eltér a normális eloszlástól (itt \bar{D}_{krit} a megfelelő Lilliefors-próba kritikus értéke).

Ha $D_n < \bar{D}_{\text{krit}}$, (vagy $p > \alpha$) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés a normális eloszlástól.

A normalitás tesztelése: Lilliefors-próba

A korábbi ábrához tartozó, 96 elemű, testmagasságra vonatkozó példában:

```
require(nortest)
```

```
> lillie.test(testmagassag)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data: testmagassag
```

```
D = 0.0609, p-value = 0.5307
```

Mivel $0,068 = D < D_{\text{krit}} = 0,09$, illetve $p = 0,5307 > 0,05 = \alpha$, a szignifikanciaszintet $\alpha = 0,05$ -nek választva **elfogadható**, hogy a testmagasság normális eloszlású valamilyen paraméterekkel, nincs szignifikáns eltérés a normális eloszlástól.

A normalitás tesztelése: Lilliefors-próba

Ugyanakkor GDP volumenindexére vonatkozó példában

```
> lillie.test(gdp)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data:  gdp
```

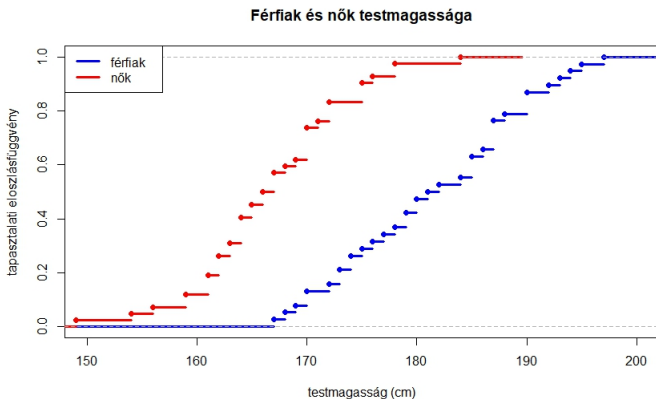
```
D = 0.2055, p-value = 0.01287
```

Itt $p < 0,05$, vagyis a nullhipotézist elutasítjuk, a volumenindex eloszlása szignifikánsan eltér a normális eloszlástól.

Megjegyzés: a rendezett minta kovarianciamátrixát használó Shapiro–Wilk-próbánál a testmagasság esetében $p = 0,36$, míg a gdp volumenindexe esetében $p = 0,0002$. Ilyenkor érdemes lehet részletesebben megnézni, hogy melyik próbánál mit kell feltelezni, milyen az adatsor (vannak-e például kiugró értékek).

Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

Állíthatjuk-e, hogy a férfiak és a nők testmagasságának **eloszlása** szignifikánsan eltérő? Ez a kérdés nem csak a várható értékre és a szórásra vonatkozik, hanem magára az eloszlásra.



A férfiak ($n = 38$ megfigyelés) és nők ($m = 42$ megfigyelés) testmagasságának tapasztalati eloszlásfüggvénye

Kolmogorov–Szmirnov-próba: homogenitásvizsgálat

H_0 : az X_1, \dots, X_n és Y_1, \dots, Y_m minták **ugyanabból az eloszlásból** származnak, azaz minden t valós számra teljesül, hogy $\mathbb{P}(X_j \leq t) = \mathbb{P}(Y_j \leq t)$.

H_1 : a minták **különböző eloszlásból** származnak, azaz van olyan t valós szám, amire $\mathbb{P}(X_j \leq t) \neq \mathbb{P}(Y_j \leq t)$.

A próbastatisztika, ami H_0 esetén Kolmogorov–Szmirnov-eloszlású:

$$D_{m,n} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|,$$

ahol \hat{F}_n az X , a \hat{G}_m pedig az Y minta tapasztalati eloszlásfüggvénye.

Ha $D_{m,n} > D_{\text{krit}}$ (vagy $p < \alpha$), akkor elutasítjuk H_0 -t, a minták eloszlása szignifikánsan különböző (itt D_{krit} a megfelelő Kolmogorov–Szmirnov-próba kritikus értéke). Ha $D < D_{\text{krit}}$, (vagy $p > \alpha$) akkor elfogadjuk a nullhipotézist, nincs szignifikáns eltérés a minták eloszlása között.

A kritikus értékek az alábbi összefüggés alapján közelíthetők:

$$\lim_{m,n \rightarrow \infty} \mathbb{P}\left(\sqrt{\frac{mn}{m+n}} D_{m,n} < y\right) = \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 y^2} \Rightarrow D_{\text{krit}} \approx \sqrt{\frac{m+n}{mn}} \sqrt{-\frac{1}{2} \log \alpha}.$$

Homogenitásvizsgálat: példa

H_0 : az X_1, \dots, X_n és Y_1, \dots, Y_m minták ugyanabból az eloszlásból származnak, vagyis a férfiak és nők testmagasságának eloszlása megegyezik.

H_1 : a minták különböző eloszlásból származnak, vagyis a férfiak és nők testmagasságának eloszlása eltérő.

A próbastatisztika a tapasztalati eloszlásfüggvények legnagyobb eltérése, az ábra alapján $t = 174$ környékén lehet:

$$D_{m,n} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|.$$

```
> ks.test(ferfi, no, alternative="two.sided")
```

```
Two-sample Kolmogorov-Smirnov test
```

```
data:  ferfi and no
```

```
D = 0.6754, p-value = 2.486e-08
```

```
alternative hypothesis:  two-sided
```

Homogenitásvizsgálat: példa

H_0 : az X_1, \dots, X_n és Y_1, \dots, Y_m minták ugyanabból az eloszlásból származnak, vagyis a férfiak és nők testmagasságának eloszlása megegyezik.

H_1 : a minták különböző eloszlásból származnak, vagyis a férfiak és nők testmagasságának eloszlása eltérő.

A próbastatisztika a tapasztalati eloszlásfüggvények legnagyobb eltérése, az ábra alapján $t = 174$ környékén lehet:

$$D_{m,n} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|.$$

```
> ks.test(ferfi, no, alternative="two.sided")
```

```
Two-sample Kolmogorov-Smirnov test
```

```
data: ferfi and no
```

```
D = 0.6754, p-value = 2.486e-08
```

```
alternative hypothesis: two-sided
```

A p -érték kisebb 0,05-nél, a nullhipotézist elutasítjuk, a férfiak és a nők **testmagasságának eloszlása szignifikánsan különböző**.

Egyoldali eset

H_0 : az X és Y valószínűségi változók ugyanabból az eloszlásból származnak

H_1 : minden t valós számra $F(t) = \mathbb{P}(X \leq t) \leq G(t) = \mathbb{P}(Y \leq t)$, ahol F az X , a G pedig az Y eloszlásfüggvénye. Azaz $X \geq Y$ sztochasztikusan.

$$D_{m,n}^- = \sup_{t \in \mathbb{R}} \hat{G}_n(t) - \hat{F}_m(t),$$

ahol \hat{F}_n az X , a \hat{G}_m pedig az Y minta tapasztalati eloszlásfüggvénye. A nullhipotézist elutasítjuk, ha D nagyobb az egyoldali Kolmogorov–Szmirnov-próba kritikus értékénél.

Egyoldali eset

H_0 : az X és Y valószínűségi változók ugyanabból az eloszlásból származnak

H_1 : minden t valós számra $F(t) = \mathbb{P}(X \leq t) \leq G(t) = \mathbb{P}(Y \leq t)$, ahol F az X , a G pedig az Y eloszlásfüggvénye. Azaz $X \geq Y$ sztochasztikusan.

$$D_{m,n}^- = \sup_{t \in \mathbb{R}} \hat{G}_n(t) - \hat{F}_m(t),$$

ahol \hat{F}_n az X , a \hat{G}_m pedig az Y minta tapasztalati eloszlásfüggvénye. A nullhipotézist elutasítjuk, ha D nagyobb az egyoldali Kolmogorov–Szmirnov-próba kritikus értékénél.

```
> ks.test(ferfi, no, alternative="less")
```

```
Two-sample Kolmogorov-Smirnov test
```

```
data: ferfi and no
```

```
 $D^- = 0.6754$ , p-value = 1.243e-08
```

```
alternative hypothesis: the CDF of x lies below that of y
```

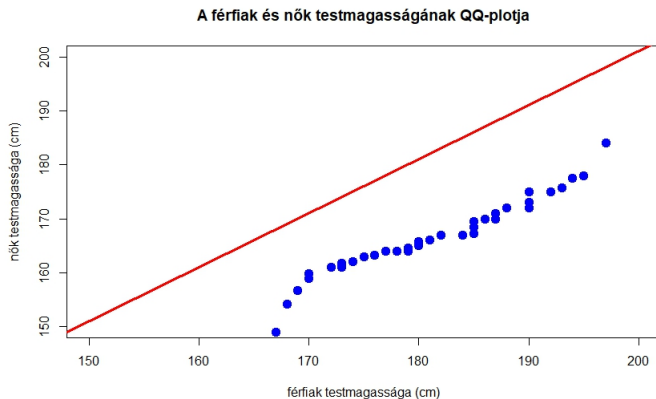
A férfiak testmagasságának eloszlása sztochasztikus értelemben szignifikánsan nagyobb, mint a nőké.

QQ-plot

Annak vizsgálatára, hogy két minta ugyanabból az eloszlásból származik-e (homogenitásvizsgálat) a leíró statisztikában a QQ-plot is gyakran használt ábrázolási mód.

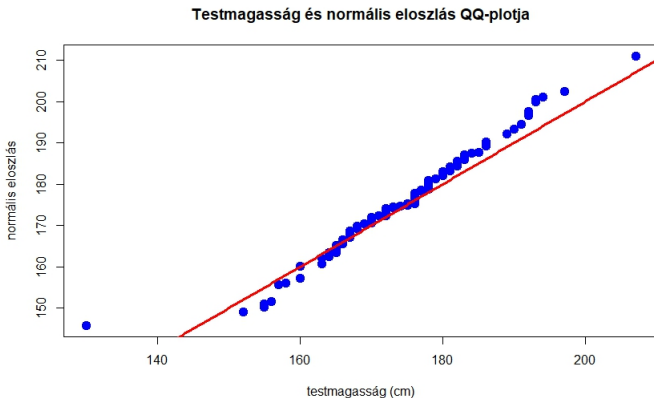
- a QQ-plot két minta eloszlásának az összehasonlítására szolgál, a kvantilisek összehasonlításával
- ha a tapasztalati z -kvantilis az első mintában q_1 , a másodikban q_2 , akkor a (q_1, q_2) pontba kerül egy pont
- minél inkább egyezik a két minta eloszlása, annál közelebb lesznek a tapasztalati eloszlásfüggvényik, ezért annál közelebb lesznek az ugyanahhoz a z -hez tartozó kvantiliseik egymáshoz, vagyis annál közelebb lesz a QQ-plot az $y = x$ egyeneshez.

QQ-plot



QQ-plot a férfiak és nők testmagasságának összehasonlítására, $n = 96$ elemű minta alapján

QQ-plot



A testmagasság adatok és egy szintén 96 elemű, $\bar{X} = 174,3$ várható értékű és $s_n^* = 11,5$ szórású normális eloszlású minta QQ-plotja

Előjelpróba

Legyen $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ olyan minta, melyben a felsorolt párok függetlenek egymástól (de a párok két eleme nem feltétlenül független), és folytonos eloszlásúak.

Kétoldali ellenhipotézis:

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

Legyen W az olyan párok száma, amikre $Y_i > X_i$. H_0 esetén ez binomiális eloszlású, n renddel és $p = 0,5$ paraméterrel, hiszen minden pár esetében egymástól függetlenül $0,5$ valószínűséggel teljesül az egyenlőtlenség. A binomiális eloszlást a centrális határeloszlástétel alapján normális eloszlással közelítve jutunk az alábbi eljáráshoz. Legyen

$$z = \frac{W - n/2}{\sqrt{n/4}}. \quad (1)$$

Elutasítjuk a nullhipotézist, ha $|z| > \Phi^{-1}(1 - \alpha/2)$, különben elfogadjuk. A p -érték, ugyanúgy, ahogy a z -próbánál, $2(1 - \Phi(|z|))$ lesz.

Az egyoldali esetben: $H_0 : \mathbb{P}(X > Y) \geq \mathbb{P}(X < Y)$. $H_1 : \mathbb{P}(X > Y) < \mathbb{P}(X < Y)$.

Elutasítjuk a nullhipotézist, ha $z > \Phi^{-1}(1 - \alpha)$, különben elfogadjuk. A p -érték ilyenkor $1 - \Phi(z)$.

Wilcoxon-próba

Az előző hipotézisvizsgálati feladatban egy másik eljárást is gyakran használnak.

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

- Hagyjuk el azokat a párokat, ahol $X_j = Y_j$. Marad k pár.
- A megmaradt k párt állítsuk az $|Y_j - X_j|$ szerint növekvő sorrendbe. Minél nagyobb az eltérés, annál nagyobb súllyal fog számítani.
- Minden párra számítsuk ki, hogy hányadik ebben a sorrendben, legyen ez R_j . Az 1 a legkisebb, k a legnagyobb különbség. Ha egyenlők vannak, mindegyik azonos sorszámot kapjon, a megfelelő sorszámok átlagát.
- Ezt az R_j rangot szorozzunk meg $Y_j - X_j$ előjével, majd ezeket adjuk össze:

$$W = \sum_{j=1}^k \text{sgn}(Y_j - X_j) \cdot R_j.$$

- A W -t a Wilcoxon-próba kritikus értékeihez hasonlíthatjuk.

Wilcoxon-próba

- Ha a mintaelemszám elég nagy, a

$$z = \frac{W}{\sqrt{\frac{k(k+1)(2k+1)}{6}}}$$

mennyiségre kétoldali z-próbát alkalmazhatunk, a kritikus érték ebben az esetben $1 - \Phi^{-1}(1 - \alpha/2)$, ahol α a szignifikanciaszint.

- Itt is lehet egyoldali ellenhipotézist is vizsgálni, akkor az egyoldali Wilcoxon-próba kritikus értékére van szükség, illetve a közelítő esetben az egyoldali z-próbának megfelelően járhatunk el.

Wilcoxon-próba, példa

Hat tóparti szálloda májusi és júniusi bevételét mutatja az alábbi táblázat (millió forintban). Az egyes szállodák bevételét egymástól függetlennek tekintjük, de a májusi és a júniusi érték egy szálloda esetében összefügghet. Nincsenek egyenlő értékek a párokon belül, így nem kell elhagyni mintaelemeket.

Kétoldali ellenhipotézist vizsgálunk. Legyen X a májusi, Y a júniusi bevétel:

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

szálloda	A	B	C	D	E	F
májusi bevétel (X_j)	20,3	19,3	16,5	22,4	23,8	18,5
júniusi bevétel (Y_j)	25,2	22,9	14,3	26,3	21,7	22,1
$ X_j - Y_j $	4,9	3,6	2,2	3,9	2,1	3,6
rang (R_j)	6	3,5	2	5	1	3,5
a különbség előjele ($\text{sgn}(Y_j - X_j)$)	+1	+1	-1	+1	-1	+1

Wilcoxon-próba, példa

Hat tóparti szálloda májusi és júniusi bevételét mutatja az alábbi táblázat (millió forintban). Az egyes szállodák bevételét egymástól függetlennek tekintjük, de a májusi és a júniusi érték egy szálloda esetében összefügghet. Nincsenek egyenlő értékek a párokon belül, így nem kell elhagyni mintaelemeket.

Kétoldali ellenhipotézist vizsgálunk. Legyen X a májusi, Y a júniusi bevétel:

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y).$$

$$H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y).$$

szálloda	A	B	C	D	E	F
májusi bevétel (X_j)	20,3	19,3	16,5	22,4	23,8	18,5
júniusi bevétel (Y_j)	25,2	22,9	14,3	26,3	21,7	22,1
$ X_j - Y_j $	4,9	3,6	2,2	3,9	2,1	3,6
rang (R_j)	6	3,5	2	5	1	3,5
a különbség előjele ($\text{sgn}(Y_j - X_j)$)	+1	+1	-1	+1	-1	+1

$$W = \sum_{j=1}^k \text{sgn}(Y_j - X_j) \cdot R_j = 6 + 3,5 - 2 + 5 - 1 + 3,5 = 15.$$

Wilcoxon-próba, példa

Bár most a mintaelemszám nem elég nagy, a példa kedvéért a közelítő összeget használva $k = 6$ -tal (hiszen hat pár van):

$$z = \frac{W}{\sqrt{\frac{k(k+1)(2k+1)}{6}}} = \frac{15}{\sqrt{6 \cdot 7 \cdot 136}} = 1,57.$$

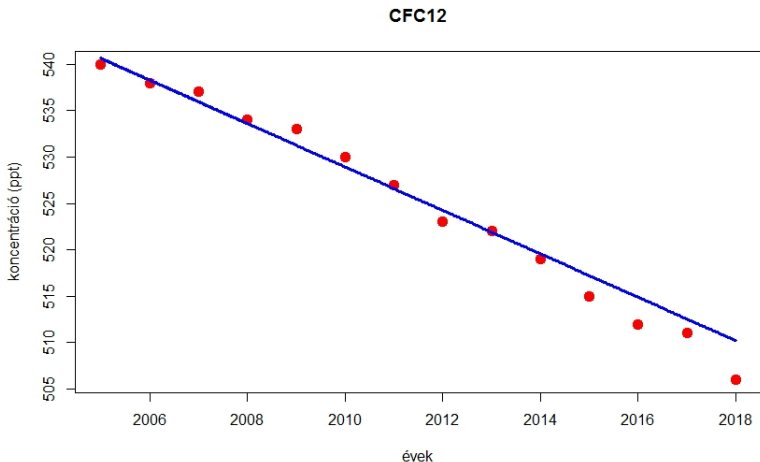
A kétoldali z-próba esetén az $\alpha = 0,05$ -höz tartozó kritikus érték: $1 - \Phi^{-1}(0,025) = 1,96$. Mivel tehát $|z|$ kisebb a kritikus értéknél, a nullhipotézist elfogadjuk, annak valószínűsége, hogy a májusi bevétel nagyobb a júniusinál, nem tér el szignifikánsan annak valószínűségétől, hogy a júniusi szignifikánsan nagyobb a májusinál.

Az előjelpróbával a (1) egyenlet alapján ($n = 6$ a párok száma):

$$z = \frac{W - n/2}{\sqrt{n/4}} = \frac{4 - 3}{\sqrt{6/4}} = 0,817$$

adódik, hiszen ott W az olyan párok száma, ahol $Y_j > X_j$. Erre is z-próbát végezhetünk, a kritikus érték most is $1,96$, ezzel az eljárással is elfogadjuk a nullhipotézist.

Lineáris regresszió



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

Lineáris regresszió

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)

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
```

Lineáris modell egyváltozós esetben

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 (amiknek tehát ugyanakkora a szórása). 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.

Becslések a lineáris modellben

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:

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

Előrejelzés a lineáris modellben

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:

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

ML-becslés a lineáris modellben

Legyenek $X_1 = x_1, \dots, X_n = x_n$ rögzítettek. Ebben az esetben a lineáris modellben az $Y_j = ax_j + b + \varepsilon_j$ valószínűségi változók függetlenek, normális eloszlásúak σ szórással, és Y_j várható értéke $ax_j + b$. Így a likelihood-függvény:

ML-becslés a lineáris modellben

Legyenek $X_1 = x_1, \dots, X_n = x_n$ rögzítettek. Ebben az esetben a lineáris modellben az $Y_j = ax_j + b + \varepsilon_j$ valószínűségi változók függetlenek, normális eloszlásúak σ szórással, és Y_j várható értéke $ax_j + b$. Így a likelihood-függvény:

$$L_{n,a,b,\sigma}(y_1, \dots, y_n) = \prod_{j=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} \cdot \exp \left(- \frac{(y_j - ax_j - b)^2}{2\sigma^2} \right) \right).$$

ML-becslés a lineáris modellben

Legyenek $X_1 = x_1, \dots, X_n = x_n$ rögzítettek. Ebben az esetben a lineáris modellben az $Y_j = ax_j + b + \varepsilon_j$ valószínűségi változók függetlenek, normális eloszlásúak σ szórással, és Y_j várható értéke $ax_j + b$. Így a likelihood-függvény:

$$L_{n,a,b,\sigma}(y_1, \dots, y_n) = \prod_{j=1}^n \left(\frac{1}{\sqrt{2\pi\sigma}} \cdot \exp \left(- \frac{(y_j - ax_j - b)^2}{2\sigma^2} \right) \right).$$

$$L_{n,a,b,\sigma}(y_1, \dots, y_n) = \frac{1}{(\sqrt{2\pi\sigma})^n} \exp \left(- \frac{\sum_{j=1}^n (y_j - ax_j - b)^2}{2\sigma^2} \right).$$

ML-becslés a lineáris modellben

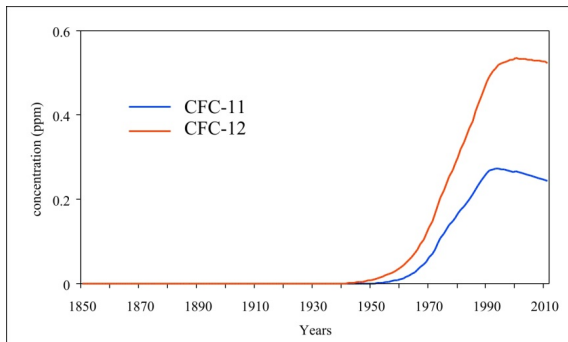
Legyenek $X_1 = x_1, \dots, X_n = x_n$ rögzítettek. Ebben az esetben a lineáris modellben az $Y_j = ax_j + b + \varepsilon_j$ valószínűségi változók függetlenek, normális eloszlásúak σ szórással, és Y_j várható értéke $ax_j + b$. Így a likelihood-függvény:

$$L_{n,a,b,\sigma}(y_1, \dots, y_n) = \prod_{j=1}^n \left(\frac{1}{\sqrt{2\pi\sigma}} \cdot \exp \left(- \frac{(y_j - ax_j - b)^2}{2\sigma^2} \right) \right).$$

$$L_{n,a,b,\sigma}(y_1, \dots, y_n) = \frac{1}{(\sqrt{2\pi\sigma})^n} \exp \left(- \frac{\sum_{j=1}^n (y_j - ax_j - b)^2}{2\sigma^2} \right).$$

Rögzített σ mellett ez akkor maximális, ha $\sum_{j=1}^n (y_j - ax_j - b)^2$ minimális. Ebből adódik az \hat{a} és \hat{b} maximumlikelihood-becslés.

Előrejelzés



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

A lineáris modell közelítése gyakran csak egy rövidebb időintervallumon jó. Minél távolabbra készítjük az előrejelzést, annál nagyobb lesz a bizonytalanság, amint azt a szórás is mutatja.

Reziduálisok és R^2

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:

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

Definíció

A megmagyarázott ingadozás részaránya (coefficient of determination):

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

Reziduálisok és R^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).

Konfidenciaintervallumok

$1 - \alpha$ 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.

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.

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

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.

Az egyenes meredekségére vonatkozó próbák

A korábbi példában (16. 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.

Az egyenes meredekségére vonatkozó próbák

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.

Többváltozós lineáris modell

A lineáris modellben több magyarázó változót is bevezethetünk.

Az Y változót fejezzük ki az X_1, \dots, X_p valószínűségi változók lineáris függvényeként, de az együtthatókat 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:

$$Y_1 = a_1 X_{1,1} + a_2 X_{1,2} + b + \varepsilon_1;$$

$$Y_2 = a_1 X_{2,1} + a_2 X_{2,2} + b + \varepsilon_2;$$

...

$$Y_n = a_1 X_{n,1} + a_2 X_{n,2} + b + \varepsilon_n.$$

Többváltozós lineáris modell

A lineáris modellben több magyarázó változót is bevezethetünk.

Az Y változót fejezzük ki az X_1, \dots, X_p valószínűségi változók lineáris függvényeként, de az együtthatókat 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:

$$Y_1 = a_1 X_{1,1} + a_2 X_{1,2} + b + \varepsilon_1;$$

$$Y_2 = a_1 X_{2,1} + a_2 X_{2,2} + b + \varepsilon_2;$$

...

$$Y_n = a_1 X_{n,1} + a_2 X_{n,2} + b + \varepsilon_n.$$

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

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

Az együtthatók becslése

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

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

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

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

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 megfelelő illeszkedés ellenőrzése

A megmagyarázott ingadozás részaránya:

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

- érzékeny a kiugró értékekre
- nem veszi figyelembe a becsült paraméterek számát
- elég sok paraméterrel megfigyelhető a **tútanulás (overfitting)** jelensége: valójában nem modellt illesztünk, hanem a véletlen hibákat külön-külön tanuljuk meg

A megfelelő illeszkedés ellenőrzése

A megmagyarázott ingadozás részaránya:

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

- érzékeny a kiugró értékekre
- nem veszi figyelembe a becsült paraméterek számát
- elég sok paraméterrel megfigyelhető a **túltanulás (overfitting)** jelensége: valójában nem modellt illesztünk, hanem a véletlen hibákat külön-külön tanuljuk meg

Ezért az R^2 -nek az alábbi módosított (adjusted) változata is gyakran használt:

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

Ha $p = 0$, visszakapjuk az eredetit (persze ez nem egy valódi modell).

Például: $n = 100$, $p = 10$ esetén jelzi, hogy a mintaelemszámhoz képest túl sok a paraméter.

Hipotézisvizsgálat a lineáris modellben, egyváltozós eset

Egyváltozós eset: $Y_i = aX_i + b + \varepsilon_i$

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

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.

Hipotézisvizsgálat a lineáris modellben

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

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

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

$$H_0 : H\beta = 0$$

$$H_1 : H\beta \neq 0.$$

Ha például $r = 3$, akkor a nullhipotézis három olyan típusú egyenletet jelent, hogy $5a_1 + 3a_2 - 2a_3 = 0$, vagyis az együtthatók valamely lineáris kombinációja 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 egységvektorokból állítjuk össze, akkor tudjuk több együttható 0 voltát egyszerre tesztelni.

Hipotézisvizsgálat a lineáris modellben

$$H_0 : H\beta = 0$$

$$H_1 : H\beta \neq 0.$$

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

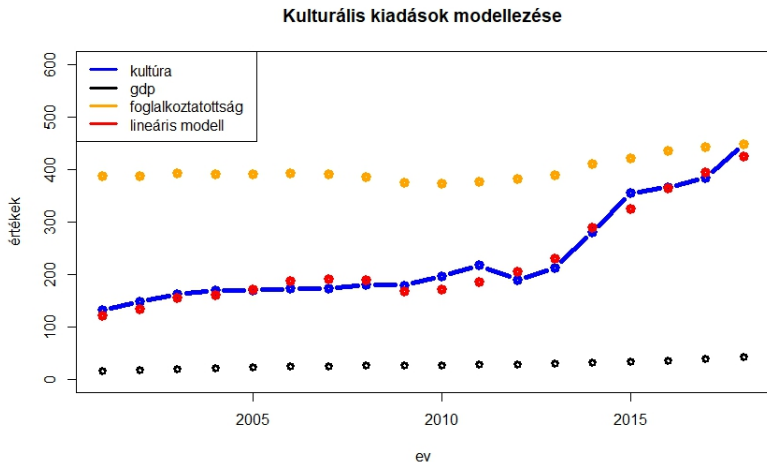
$$F = \frac{(\underline{Y} - X\beta^*)^T(\underline{Y} - X\beta^*) - (\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})}{(\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})},$$

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). Itt tehát a számláló első tagja a nullhipotézis esetén a maximumlikelihood-becslés, míg a nevezőben ugyanúgy maximumlikelihood-becslés szerepel, de a nullhipotézis feltétele nélkül, a teljes paramétertartományon.

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.

Ha $r = 1$ és $p = 2$, valamint a próbastatisztikából gyököt vonunk, akkor az egyváltozós eset próbastatisztikáját és egy t -eloszlás abszolút értékét kapjuk, így lesz ez a korábban látott módszer általánosítása.

Többváltozós lineáris modell: példa



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; forrás: KSH)

Többváltozós lineáris regresszió: példa

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.

```
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)
```

Többváltozós lineáris modell: példa

```
> 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 ***  
--
```

Többváltozós lineáris modell: példa

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:

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

Többváltozós lineáris modell: példa

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:

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

Az R^2 értéke 1-hez viszonylag közeli, 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:

Többváltozós lineáris modell: példa

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:

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

Az R^2 értéke 1-hez viszonylag közeli, 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 május 6., kedd, 10:15-ig

A félév elején gyűjtött adatok alapján $\alpha = 0,05$ szignifikanciaszinten elfogadhatjuk-e, hogy a sorozatnézéssel töltött idő normális eloszlású? (Az irreális adatokat távolítsuk el a próba elvégzése előtt, vagy váltsuk át megfelelő mértékegységre.)