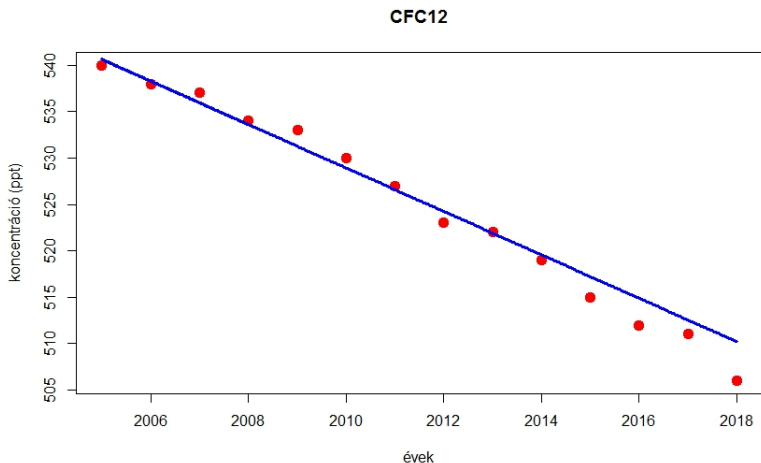


Lineáris regresszió (12. előadás)



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:

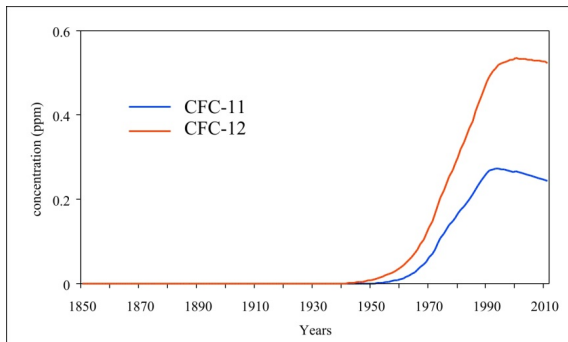
$$D(\hat{a}x^* + \hat{b}) = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}}.$$

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:

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

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.

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.)

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.)

```
> require("readxl")  
> adatok <- read_excel("sstadat-p.xlsx")  
> sorozat=adatok$sorozat  
> require("nortest")  
> lillie.test(sorozat)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: sorozat
```

```
D = 0.1823, p-value = 1.9e-10
```

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.)

```
> require("readxl")  
> adatok <- read_excel("sstadat-p.xlsx")  
> sorozat=adatok$sorozat  
> require("nortest")  
> lillie.test(sorozat)
```

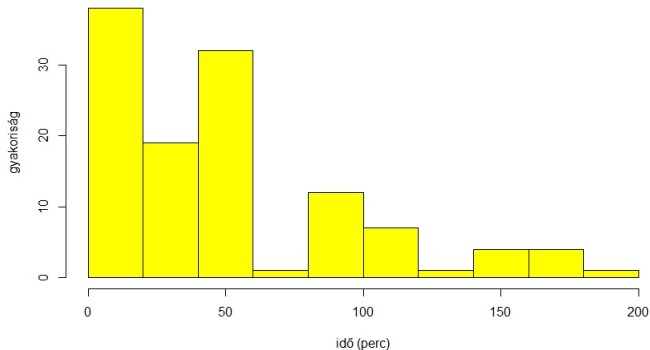
Lilliefors (Kolmogorov-Smirnov) normality test

data: sorozat

D = 0.1823, p-value = 1.9e-10

Szignifikáns eltérés van a normális eloszlástól.

Házi feladat május 6., kedd, 10:15-ig



Az egy hétköznapon sorozatnézéssel töltött idő hisztogramja $n = 120$ megfigyelésből

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 (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.

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 **tútanulás (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:

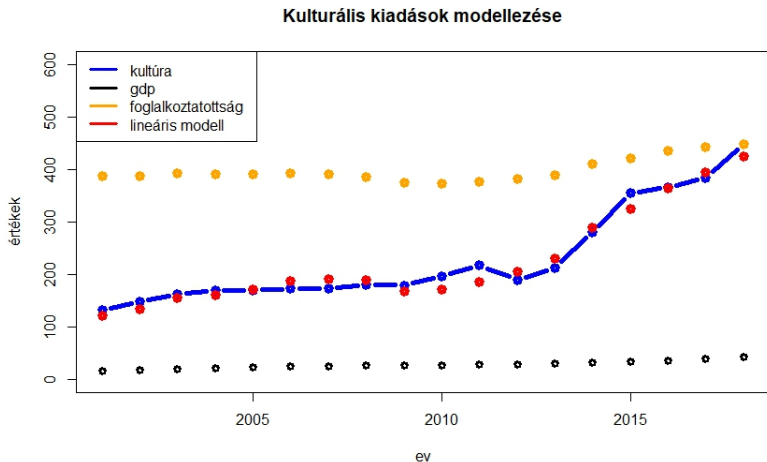
$$F = \frac{(\underline{Y} - X\beta^*)^T(\underline{Y} - X\beta^*) - (\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})}{(\underline{Y} - X\hat{\beta})^T(\underline{Y} - X\hat{\beta})},$$

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.

Szórásanalízis (analysis of variance, ANOVA)

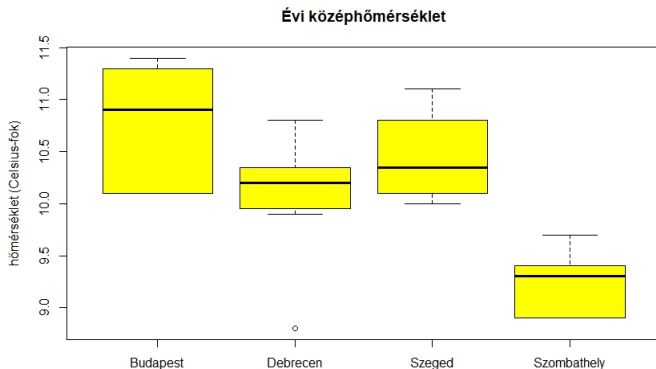
- hipotézisvizsgálati eljárás
- az egyedeket különböző csoportokba soroljuk
- egy **normális eloszlású** mennyiséget vizsgálunk: igaz-e, hogy ennek a várható értéke az egyes csoportokban azonos?
- **feltesszük, hogy a vizsgált mennyiség szórása az egyes csoportokban azonos**
- az egyes csoportokon belül és közöttük is függetlenek a megfigyelések
- két csoport esetén ez a kétmintás, Student-féle t -próba
- a többváltozós lineáris modell hipotézisvizsgálati feladatának speciális esete
- azt, hogy a megfigyelések különböző csoportokból származnak, úgy is szokták fogalmazni, hogy a mérés egy faktor különböző szintjein történik, és az a kérdés, hogy a faktornak van-e szignifikáns hatása a várható értékre

Szórásanalízis (analysis of variance, ANOVA)

Az alábbi táblázat néhány éves középhőmérséklet érték (forrás: Országos Meteorológiai Szolgálat), különböző évekből, különböző helyszínekről. A kérdés: elfogadható-e, hogy az egyes városokban az évi középhőmérséklet várható értéke megegyezik, vagy szignifikáns különbség mutatható ki? Ebben a példában a „faktor” a helyszín, és ennek négy „szintje” van.

	Budapest	Debrecen	Szeged	Szombathely
	10,8	8,8	11,1	8,9
	10,1	9,9	10,8	9,4
	11,4	10,0	10,1	8,9
	11,3	10,2	10,0	9,3
	11,0	10,4	10,4	9,7
	10,1	10,8	10,3	
		10,3		
átlag (\bar{X})	10,8	10,1	10,5	9,2
szórás (s_n^*)	0,57	0,63	0,42	0,34

Szórásanalízis (analysis of variance, ANOVA)



Boxplot ábra az egyes városok éves középhőmérséklet adataiból

Feltevések és kapcsolat a lineáris modellel

Legyenek X_{ij} független normális eloszlású valószínűségi változók, $i = 1, \dots, k$ és $j = 1, \dots, n_i$. Az X_{ij} valószínűségi változó várható értéke μ_i , szórása σ .

$$X_{ij} \sim N(\mu_i, \sigma) \quad (j = 1, 2, \dots, n_i).$$

Vagyis: k csoport van, és az i . csoportban μ_i a várható érték. Másképpen: egy faktor különböző szintjein történik mérés, az i . csoportban a faktor i . szintjének hatása μ_i .

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k.$$

$$H_1 : \mu_1 = \mu_2 = \dots = \mu_k \text{ nem teljesül.}$$

Másképpen:

H_0 : a faktornak nincs szignifikáns hatása

H_1 : a faktornak szignifikáns hatása van.

Kapcsolat a lineáris modellel

Ezt a feladatot a lineáris modell egy speciális esetének is tekinthetjük. A lineáris modell ez volt:

$$Y_j = a_1 X_{j,1} + a_2 X_{j,2} + \dots + a_k X_{j,k} + \varepsilon_j,$$

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

Most tegyük fel, hogy az $X_{j,i}$ valószínűségi változók értéke csak 0 vagy 1 lehet, sőt, hogy ezek közül mindig pontosan egy lesz 1, a többi 0 (a lineáris modellben a magyarázó változók függetlenségét nem kellett feltenni).

Ekkor ha $a_i = \mu_i$ (minden $i = 1, 2, \dots, k$ esetén), és az Y_j esetében, vagyis a j . mérésnél a k_j . valószínűségi változó 1, a többi 0, akkor $Y_j = \mu_{k_j} + \varepsilon_j$, azaz Y_j normális eloszlású μ_{k_j} várható értékkel és σ szórással.

Vagyis az Y_j -ket aszerint csoportosítva, hogy melyik $X_{j,k}$ értéke 1, éppen a p csoporthoz tartozó méréseket kapjuk vissza.

Kapcsolat a lineáris modellel

A többváltozós lineáris modellben $H\beta = 0$ alakú nullhipotéziseket tudunk tesztelni, ahol β az együtthatók vektora. Most tehát $\beta = (\mu_1, \dots, \mu_k)$, és lehet

$$H = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ & & \dots & & \\ 0 & \dots & 0 & 1 & -1 \end{pmatrix}.$$

Ekkor $H\beta = (\mu_1 - \mu_2, \mu_2 - \mu_3, \dots, \mu_{k-1} - \mu_k)^T$, így $H\beta = 0$ éppen azzal ekvivalens, hogy minden μ_j megegyezik, ami a szórásanalízis nullhipotézise volt.

A többváltozós lineáris modell esetében a megadott próbastatisztika F -eloszlású volt a nullhipotézis mellett és az F -próba kritikus értékeit használhattuk. Mivel tehát a szórásanalízis egy speciális eset, most is hasonlóképpen járhatunk el, a próbastatisztika pedig szintén megegyezik az ott látottal, bár most más alakban írjuk fel.

A szórásanalízis eljárása

X_{ij} valószínűségi változók, $i = 1, \dots, k$, $j = 1, \dots, n_i$. Vagyis k csoport van, és az i -ben n_i darab megfigyelés van.

Csoporton belüli átlagok: $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$.

Az összes megfigyelés száma: $n = n_1 + \dots + n_k$.

Teljes átlag: $\bar{\bar{X}} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}$.

Csoportokon belüli szóródás (hiba): $S_g = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$.

Csoportok közötti szóródás: $S_t = \sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2$.

Teljes szóródás: $S = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2 = S_t + S_g$.

A próbastatisztika:

$$F = \frac{S_t(n-k)}{S_g(k-1)}.$$

A szórásanalízis eljárása

Csoportokon belüli szóródás (hiba): $S_g = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2$.

Csoportok közötti szóródás: $S_t = \sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{\bar{X}})^2$.

A próbastatisztika:

$$F = \frac{S_t(n-k)}{S_g(k-1)}.$$

Legyen c_{krit} az $f_1 = k - 1$ és $f_2 = n - k$ szabadsági fokú F -próba kritikus értéke α terjedelem mellett.

Ha $F > c_{\text{krit}}$, akkor **elutasítjuk a nullhipotézist**, a várható értékek között szignifikáns eltérés van legalább egy pár esetében.

Ha $F < c_{\text{krit}}$, akkor **elfogadjuk a nullhipotézist**, a várható értékek között nincs szignifikáns eltérés.

Szórásanalízis: példa

A korábbi példára visszatérve feltételezzük, hogy a szórások az egyes városok esetében megegyeznek, és hogy a középhőmérséklet normális eloszlású, az egyes helyszínek esetében egymástól független (ez utóbbi nagyjából helyes is, mert az adatok mind különböző évekből származnak).

	Budapest	Debrecen	Szeged	Szombathely	összesen
	10,8	8,8	11,1	8,9	
	10,1	9,9	10,8	9,4	
	11,4	10,0	10,1	8,9	
	11,3	10,2	10,0	9,3	
	11,0	10,4	10,4	9,7	
	10,1	10,8	10,3		
		10,3			
átlag ($\bar{X}_{j.}$)	10,8	10,1	10,5	9,2	$\bar{\bar{X}} = 10,17$
hiba	1,62	2,36	0,89	0,47	$S_g = 5,34$

Szórásanalízis: példa

A csoportokon belüli szóródás kiszámítása:

$$\begin{aligned} S_g &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 = \\ &= ((10,8 - 10,8)^2 + (10,1 - 10,8)^2 + \dots + (10,1 - 10,8)^2) + \\ &\quad + ((8,8 - 10,1)^2 + (9,9 - 10,1)^2 + \dots + (10,3 - 10,1)^2) + \\ &\quad + ((11,1 - 10,5)^2 + (10,8 - 10,5)^2 + \dots + (10,3 - 10,5)^2) + \\ &\quad + ((8,9 - 9,2)^2 + (9,4 - 9,2)^2 + \dots + (9,7 - 9,2)^2) = 5,34. \end{aligned}$$

Itt az első sor Budapestnek (az $i = 1$ esetnek) felel meg, minden mérésnél a budapesti mérések átlagától vett különbség négyzetét számítjuk ki, és ezeket adjuk össze. A második sor, $i = 2$, Debrecen, ekkor az itteni átlagától vett eltérések négyzetét adjuk össze, majd hasonlóképpen az $i = 3$ és $i = 4$ esetekben is.

A csoportok közötti szóródás kiszámítása:

$$\begin{aligned} S_t &= \sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2 = 6 \cdot (10,8 - 10,17)^2 + 7 \cdot (10,1 - 10,17)^2 + \\ &\quad + 6 \cdot (10,5 - 10,17)^2 + 5 \cdot (9,2 - 10,17)^2 = 7,15. \end{aligned}$$

Szórásanalízis: példa

Teljes szóródás = csoportokon belüli + csoportok közötti:

$$S = S_g + S_t = 5,34 + 7,15 = 12,49.$$

Az előző példában: $n = 24$ a megfigyelések száma, $k = 4$ az osztályok száma.

A próbastatisztika:

$$F = \frac{S_t(n - k)}{S_g(k - 1)} = \frac{7,15 \cdot 20}{5,34 \cdot 3} = 8,77,$$

ahol n a megfigyelések száma, k a csoportok száma, és a csoportokon belüli szóródás (hiba): $S_g = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 = 5,43$, a csoportok közötti szóródás: $S_t = \sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X})^2 = 7,15$.

Az $f_1 = k - 1 = 3$ és $f_2 = n - k = 20$ szabadsági fokú F -próba kritikus értéke $\alpha = 0,05$ terjedelem mellett: $c_{\text{krit}} = 3,86$.

Mivel $F = 8,77 > c_{\text{krit}} = 3,86$, akkor **elutasítjuk a nullhipotézist**, a várható értékek között szignifikáns eltérés van.

Vagyis a helynek mint faktornak (tényezőnek) **szignifikáns hatása** van az évi középhőmérsékletre.

Házi feladat május 13., kedd, 10:15-ig

Tekintsük a félév elején gyűjtött adatokat, és illesztünk lineáris modellt úgy, hogy a sorozatnézéssel töltött időt írjuk fel az utazással töltött idő és a sportolások számának lineáris függvényeként (vagyis ez utóbbi kettő a magyarázó változók).

Mennyire jól illeszkedik a modell? A modell alapján mit gondolhatunk, mennyi időt tölt sorozatnézéssel valaki, aki naponta 60 percet utazik és hetente 4-szer sportol?