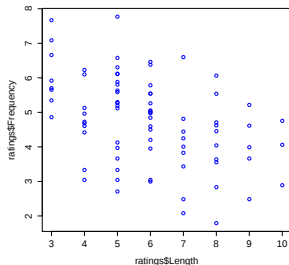


Regressziószámítás

Regressziószámítás, regressziós analízis

- ▶ Függvényszerű összefüggés keresése független (magyarázó, azaz általunk kontrollált) változó és egy vagy több függő (a kísérlet méréseit tartalmazó) változó között.
- ▶ Csak metrikus skálára alkalmazható, azaz az adatoknak ekvidisztánsnak kell lennie.
- ▶ Van egyváltozós és többváltozós regresszió.
- ▶ A regresszió lehet lineáris (elsőfokú) vagy magasabb rendű.



Regressziós egyenes

Hogyan találhatjuk meg azt az egyenest, amivel a legjobban kifejezhető az összefüggés?

Kovariancia: ha két mérőszám (x, y) függ egymástól, akkor ha x_i eltér x átlagától pozitív vagy negatív irányba, akkor y_i is el fog térni y átlagától ugyanabba az irányba.

Kovariancia kiszámítása: a két változó (x és y) értékeire kiszámítjuk az átlagtól való eltérések szorzatát, azaz

$$\sigma_{x,y} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}) * (y_k - \bar{y})$$

Regresszió: nem a változók **korrelációját**, hanem x független (azaz kontrollált) változó y függő (azaz kísérletben mért) változóra gyakorolt hatását akarjuk megállapítani, függvényyszerű kapcsolattal kifejezve. Vagyis nem az összefüggés erőssége érdekel, mint a korrelációnál, hanem hogy a független változó hogyan hat a függő változó értékére.

Példa: ratings dataframe. Elérhető a languageR csomag részeként. 81 angol növény- és állat nevét tartalmazza, valamint hozzájuk kapcsolódó becsült és mért értékeket.

Csomag telepítése: `install.packages("languageR")`
A felkínált tükrök (mirror) bármelyikét lehet választani.

Word: a növény vagy állat neve.

Class: növény vagy állat kategória.

Length: a szó hossza betűkben.

Frequency: a szó átlagos gyakorisága a vizsgált szövegekben (logaritmikus érték).

meanFamiliarity: az adott növény vagy állat saját becslésű ismertsége, a megkérdezettek között átlagolva.

Közönséges legkisebb négyzetek

A szóhossz (hány betű) adott, ez az x , független változó. A szógyakoriság a függő változó, mert függ a vizsgálatba véletlenszerűen bevont szövegektől függ. Ez az y -tengelyen ábrázolt változó.

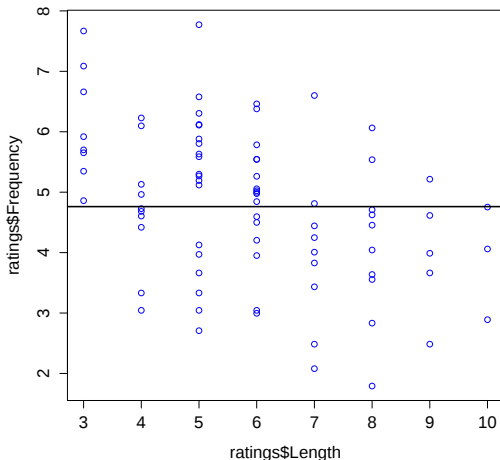
Cél: az ismert, független változó (szóhossz) alapján jósolni az ismeretlen, függő változó (gyakoriság) értékét.

Eljárás: meghatározzuk az y átlagát leképező egyenest, aminek az értéke független x -től: mindegy, hány betűből áll a szó, a vizsgált növény/állat gyakorisága azonos.

Regressziós egyenes: az átlagok egyenesét úgy döntjük meg, hogy a lehető legjobban leképezze x hatását y értékére.

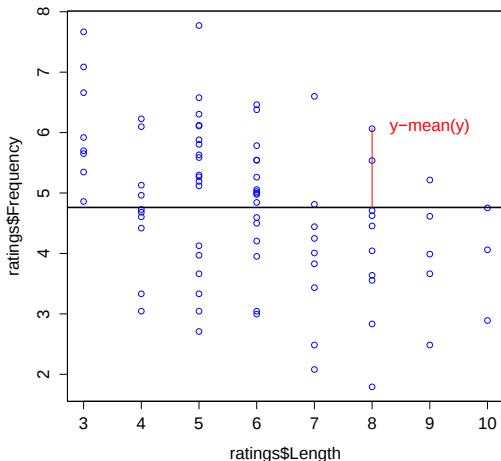
Közönséges legkisebb négyzetek

Lépések: (1) kiszámítjuk az y értékek átlagát, (2) minden egyes y érték átlagtól való eltérését (reziduum), és ezek négyzetének összegét.

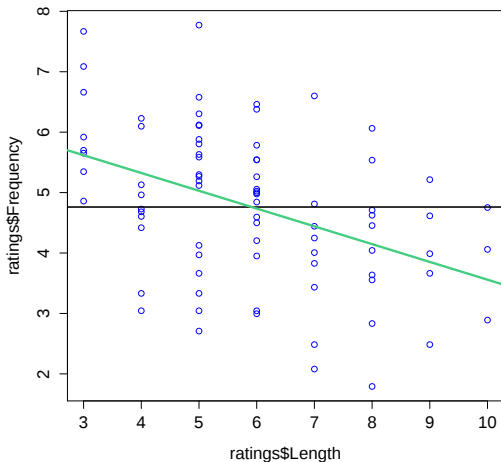


Közösleges legkisebb négyzetek

Lépések: (1) kiszámítjuk az y értékek átlagát, (2) minden egyes y érték átlagtól való eltérését (reziduum). A reziduumokat négyzetre emeljük (mindegy, hogy pozitív vagy negatív az átlagtól való eltérés), és összegezzük a négyzeteket.



Keressük azt az egyenest, amelytől az y értékek függőleges négyzetes eltérése (reziduuma, maradéka) a **legkisebb**. Angolul *ordinary least square*, *OLS* az eljárás neve.



Egyenes képlete

$$y = a + bx$$

Regressziós együtthatók:

a : egyenes metszéspontja az y -tengelyen,

b : egyenes meredeksége.

Keresett érték:

$$OLS = \sum_{k=1}^n (y_i - (a + bx_i))^2 = \min$$

ahol $OLS = \textit{Ordinary Least Square}$.

Regressziószámítás az R-ben

```
lm(függőváltozó~függetlenváltozó)
```

kimenet: a és b regressziós együtthatók

Érdemes az eredményt eltárolni egy változóban, mert így hozzáférünk a számított értékekhez:

```
ratings.lm = lm(ratings$Frequency~ratings$Length)
```

`coef(ratings.lm)` vagy `ratings.lm$coefficients`: vektor a két együtthatóval.

`fitted(ratings.lm)`: az egyeneshez igazított (hipotetikus) y értékek.

`resid(ratings.lm)`: reziduumok, a hipotetikus y értékektől való eltérések.

Egyéb elérhető adatok listázása:

```
str(ratings.lm)
```

Regressziós egyenes ábrázolása

R-függvény:

```
abline(intercept,slope)
```

1. argumentum: y -tengely metszéspontja, 2. argumentum: meredekség.

```
plot(ratings$Length,ratings$Frequency,cex.axis=1.3,cex.lab=1.3,col=4)  
abline(coef(ratings.lm))
```

hiszen a `coef(ratings.lm)` paranccsal épp a két szükséges együttthatót kapjuk meg (a metszéspont és b meredekség).

Mivel az `abline()` függvény mindig egy már meglévő grafikonokba rajzol egyenest, itt nem kell a `par(new=T)` függvényt megadni.

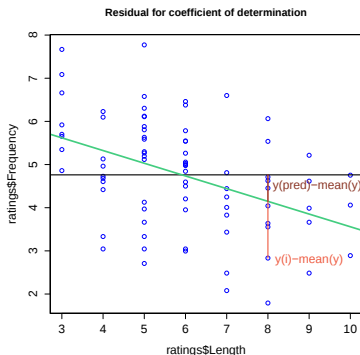
Színek gyors definiálásához lehet 1 és 9 közötti szám kódokat használni: 1 fekete, 2 piros, 3 zöld, 4 kék stb.

Determinációs együttható

A regressziós egyenes a legjobb **közelítés** az összefüggés leírására.

Hogyan jellemezhetjük, hogy **mennyire** jó a közelítés (*goodness of fit*)?

A reziduumok négyzetes összegével, de most nem az **átlagtól** való eltérés alapján, hanem egyrészt az **átlagtól**, másrészt a **regressziós egyenestől** való távolság alapján.



Determinációs együttható

Jele: R^2 . Kiszámítása:

$$D = R^2 = \frac{SS_R}{SS_T}$$

SS: *sum of squares* (négyzetes összeg). R: regressziós egyenestől való eltérés, T: y átlagtól való teljes eltérés.

Vagyis: a reziduumok (jósolt érték és megfigyelt érték közötti különbség) négyzeteinek összege és a regressziós egyenes NÉLKÜL, pusztán y átlagától számolt eltérések négyzetének aránya.

Értelmezése: az y teljes variabilitásából az x -től való függés az értékek hányad részét (hány százalékát) magyarázza meg. Értéke 0 és 1 között van.

Lineáris adatok esetén az érték megegyezik a Pearson-féle korrelációs együttható négyzetével, azaz r^2 -tel.

Ez az ún. **hatásnagyság** (*effect size*) számításának egyik lehetséges

Determinációs együttható számítása az R-ben

```
ratings.lm = lm(ratings$Frequency~ratings$Length)
```

R^2 értéke:

```
summary(ratings.lm)
```

Multiple R-squared: 0.1833

Megfelel ez a Pearson-féle korrelációs együttható, r négyzetének?

Pearson-féle $r = -0.4281$

$r^2 = 0.1833$

Vagyis a determinációs együtthatót jól megközelíti a korrelációs együttható négyzete, azaz $R^2 = r^2$.

Hasznos függvények az ábrázoláshoz

Mindkettő már létrehozott grafikonhoz ad hozzá további információt. Grafikon koordinátái „ismertek”, és felhasználhatók az elhelyezésben.

`text(x,y,"my text")`: szöveg elhelyezése a grafikonban megadott pozícióban, pl.:

`text(9,6,"y(i)-mean(y)")`

Alapbeállítás: szöveg **középpontja** esik a megadott koordinátákra.

`legend()`: jelmagyarázat

Számos opció, kötelező argumentumok: pozíció

("center", "topleft", "bottom" stb.), magyarázatok

(`legend=c("növény", "állat")`), szín vagy satírozás

(`col=c("red", "blue")`), ha `lwd` (vonaltvastagság) definiálva van, akkor vonal kerül elé, és az színes, stb.

Gyakorlás

ratings adatmátrix: hogyan függ össze a szógyakoriság (Frequency) és a növények és állatok ismertsége (meanFamiliarity)?
Végezzük el a következő számításokat külön a növényekre és az állatokra:

1. korrelációs együtthatók kiszámítása,
2. lineáris regresszió együtthatóinak kiszámítása,
3. eredmények ábrázolása egyazon ábrán,
4. determinációs együtthatók kiszámítása,
5. determinációs együttható megadása az ábrán,
6. jelmagyarázat készítése.

Hasznos függvények: `cor()`, `lm()`, `par(new=T)`, `text()`,
`abline()`, `plot(..., xlim=, ylim=, main=, xlab=, ylab=, col=)`