Regressziószámítás

Regressziószámítás, regressziós analízis

- Függvényszerű összefüggés keresése független (általunk kontrollált) változó és egy vagy több függő változó között.
- Csak metrikus skálára alkalmazható: tehát az adatoknak ekvidisztánsaknak 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 a korreláció?

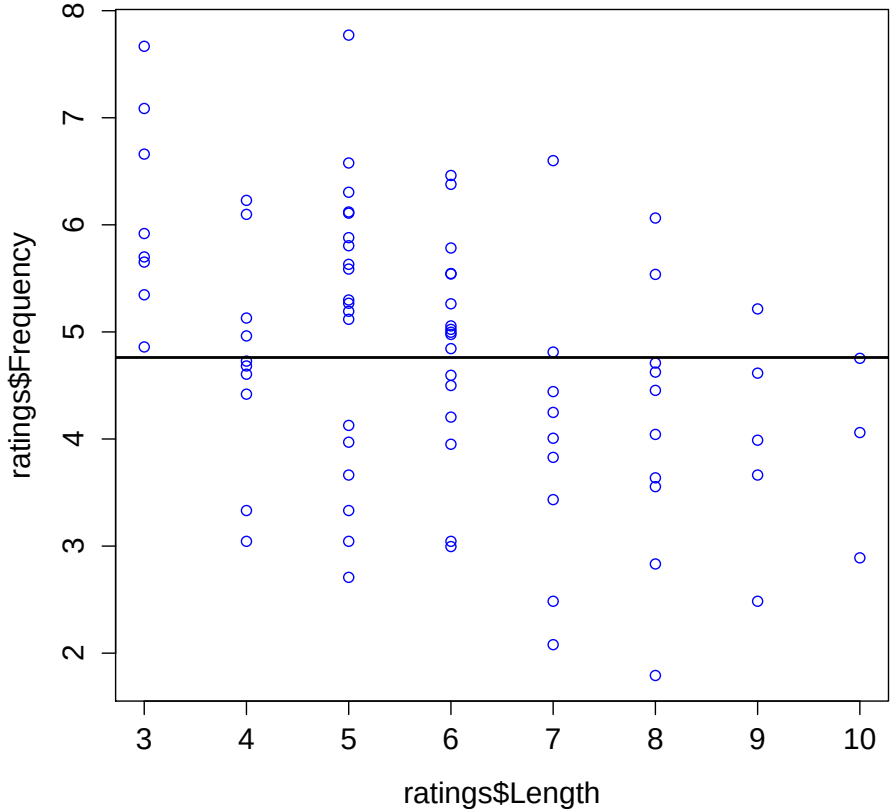
Kovariancia: ha két mérőszám (x, y) függ egymástól, akkor ha x eltér x átlagától, akkor y is el fog térni y átlagától pozitív vagy negatív irányba.
Számítása: a két változó (x és y) értékeire kiszámítjuk az átlagtól való eltérések szorzatát, azaz

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

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

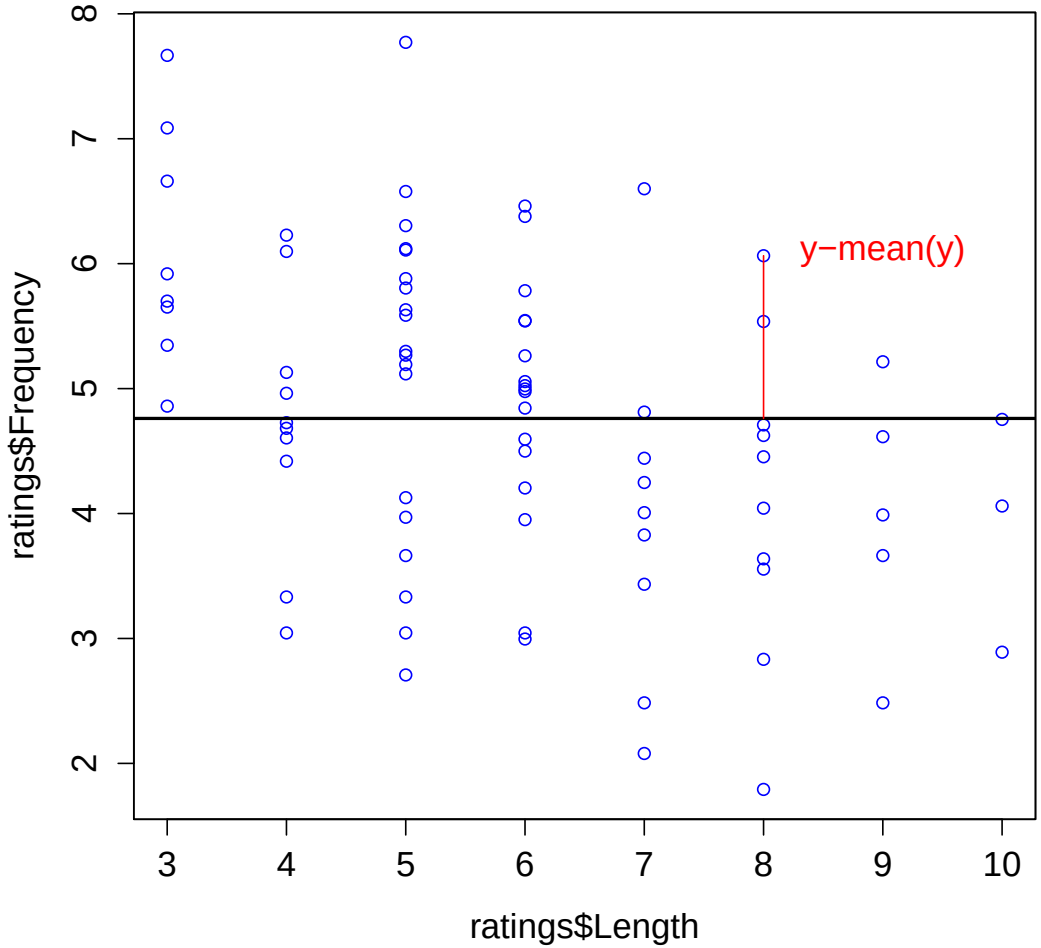
Közönséges legkisebb négyzetek

A szóhossz (hány betű) adott, ez az x, független változó. A szógyakoriság a vizsgálatba véletlenszerűen bevont szövegektől függ, vagyis az y-tengelyen ábrázolt függő változó.
Kiindulás: (1) kiszámítjuk az y értékek átlagát (fekete egyenes), (2) minden egyes y érték átlagtól való eltérését (reziduum), és ezek négyzetének összegét.

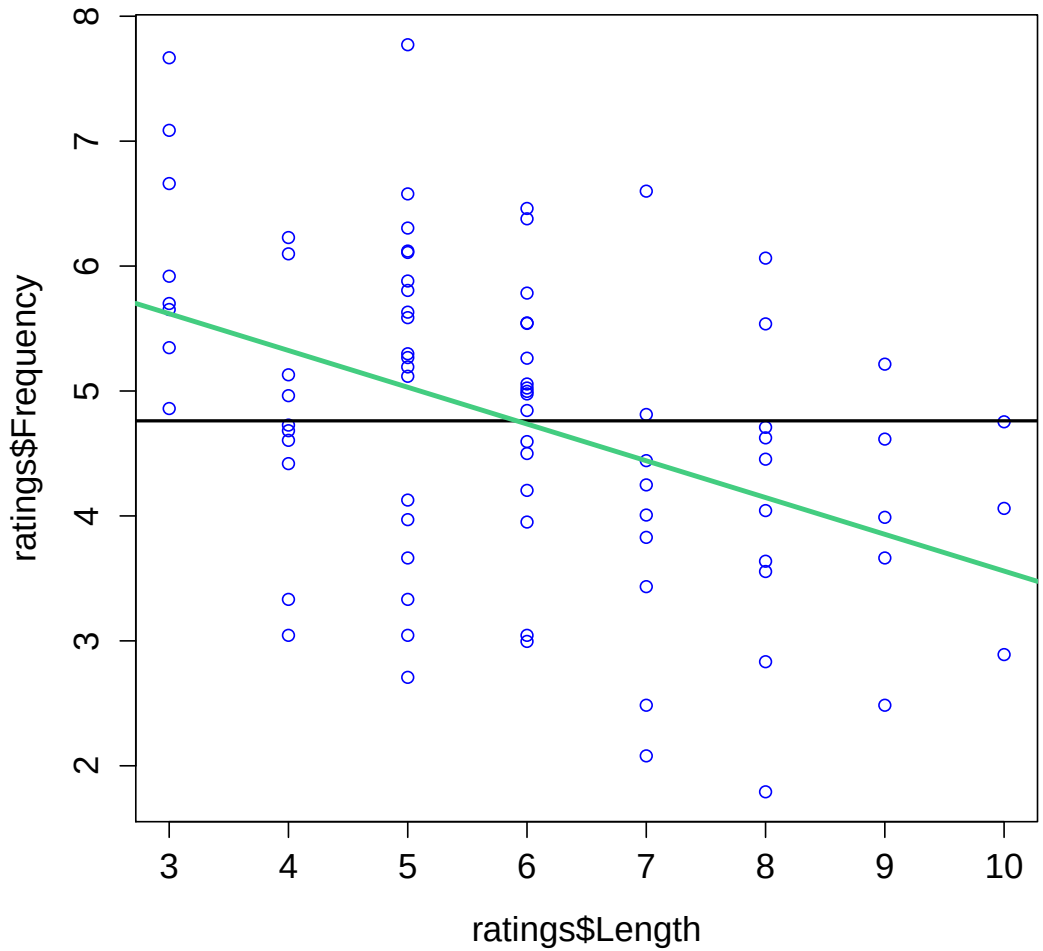


Közönséges legkisebb négyzetek

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



Keressük azt az egyenest, amelytől az y értékek függőleges négyzetes eltérése (reziduuma, maradéka) a legkisebb.



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 = *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)` parancccsal épp a két szükséges együtthatót kapjuk meg (a metszéspont és b meredekség).

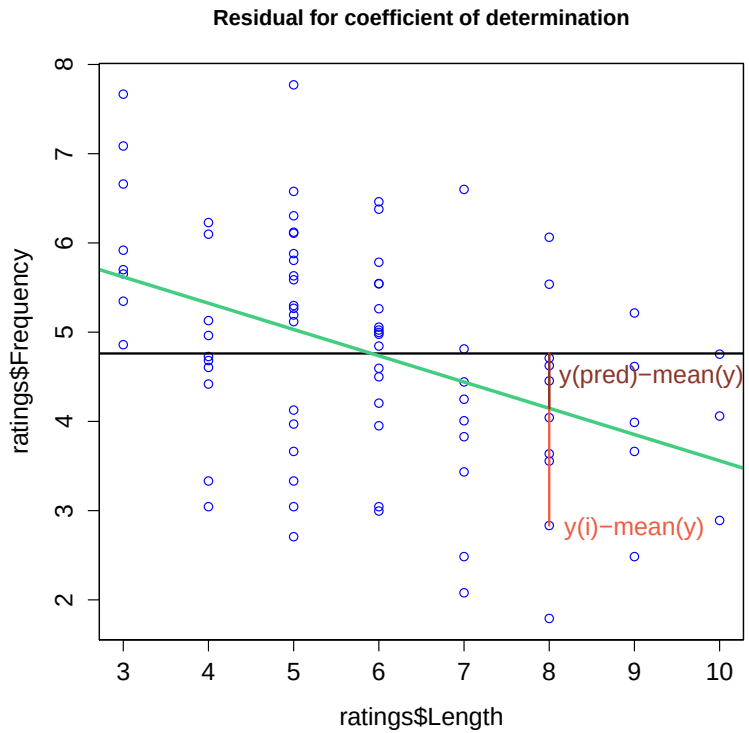
Mivel az `abline()` függvény mindig egy már meglévő grafikonokba rajzol egyenest, itt nem kell a `par(new=T)` függvényt megadni.

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 az átlagtól plusz a jobb modellként szolgáló regressziós egyenestől való távolság alapján.



Determinációs együttható

Kiszámítása:

$$R^2 = \frac{SS_R}{SS_H}$$

ahol SS: *sum of squares*, R: regresszió, H: hiba.

Ha az adatok lineárisak, az érték megegyezik a korrelációs együttható négyzetével, azaz r^2-tel.

Értelmezése: az y teljes variabilitásából az x-től való függés az értékek hányad részét (pl. hány százalékát) magyarázza meg.

Ez az ún. hatásnagyság (*effect size*) számításának egyik lehetséges módszere. Azt mutatja, hogy a pontbecslés (a regressziós egyenes mentén jósolt értékek) mennyire pontosak.

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 igaz, hogy 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 (vonalvastagság) 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.