Regressziószámítás

Házi feladatok

1. Hasonlítsuk össze a bikák és üszők születési súlyának varianciáit a megfelelő tesztekkel. Milyen különbségeket látunk?

Házi feladatok

1. Hasonlítsuk össze a bikák és üszők születési súlyának varianciáit a megfelelő tesztekkel. Milyen különbségeket látunk?

Az adatokra csak a `var.test()` (F-próba) végezhető el, mert a Levene-próba és a Bartlett-próba csak azonos elemszámra alkalmazható. Az F-próba szerint a varianciák azonosak.

2. Igaz-e, hogy a ratings mátrixban szereplő állatnevek gyakorisága alacsonyabb mértékű, mint a növényeké? És az ismertségük? Hogyan változik az eredmény, ha kétoldali próbát alkalmazunk?

2. Igaz-e, hogy a ratings mátrixban szereplő állatnevek gyakorisága alacsonyabb mértékű, mint a növényeké? És az ismertségük? Hogyan változik az eredmény, ha kétoldali próbát alkalmazunk?

```
z = ratings$Class == "plant"
t.test(ratings$Frequency[z],ratings$Frequency[!z],
alternative="less")
```

```
t.test(ratings$meanFamiliarity[z],
ratings$meanFamiliarity[!z],alternative="less")
```

Mindkét különbség szignifikáns, de fordított irányban: a növénynevek gyakoribbak, az állatnevek pedig ismertebbek.

Regressziószámítás, regressziós analízis

- Függvényszerű kapcsolat keresése egy vagy több folytonos magyarázó vagy független változó és egy függő változó között.
- Csak metrikus skálára alkalmazható.
- 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?

Regressziós egyenes

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

Korrelációs együttható számítása: átlagtól való eltérések négyzeteinek összege, azaz

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

Regressziós egyenes

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

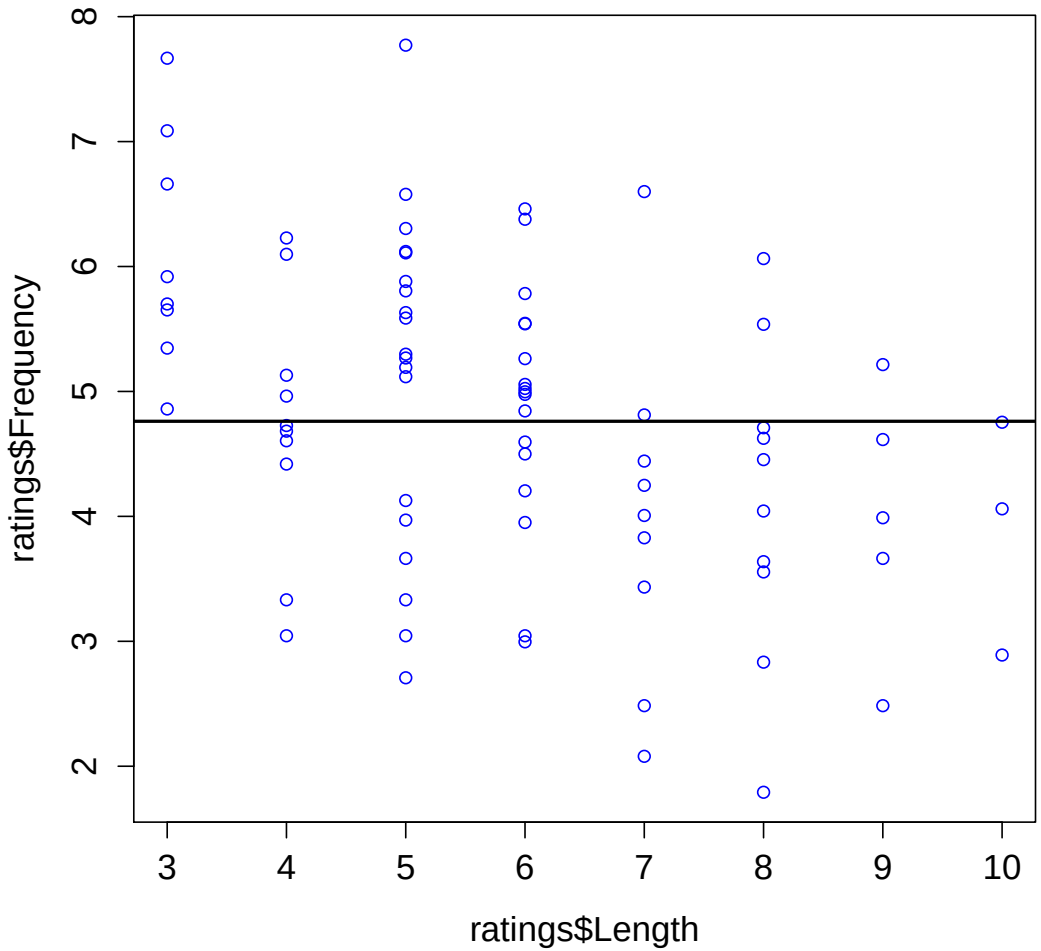
Korrelációs együttható számítása: átlagtól való eltérések négyzeteinek összege, azaz

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

Itt: nem a változók összefüggését, hanem x független változó y függő változóra gyakorolt hatását akarjuk megállapítani, függvényszerű kapcsolattal kifejezve.

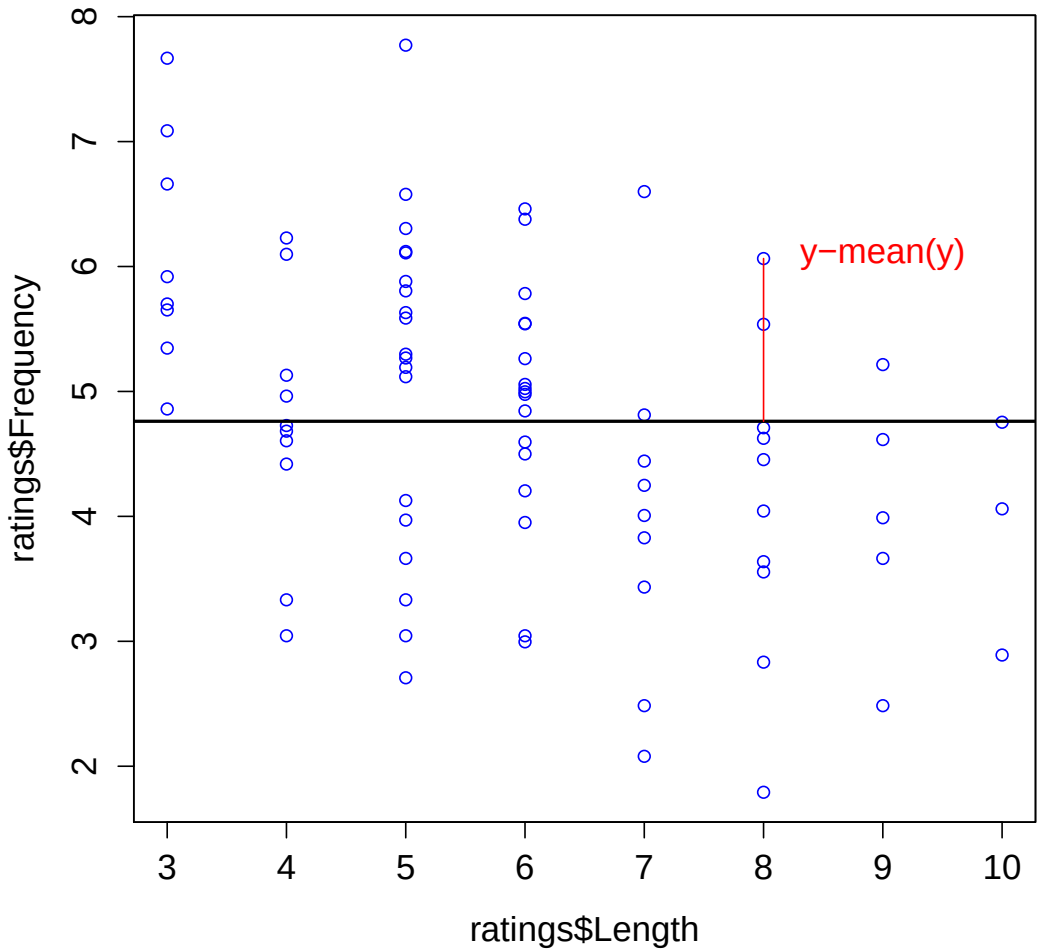
Közönséges legkisebb négyzetek

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

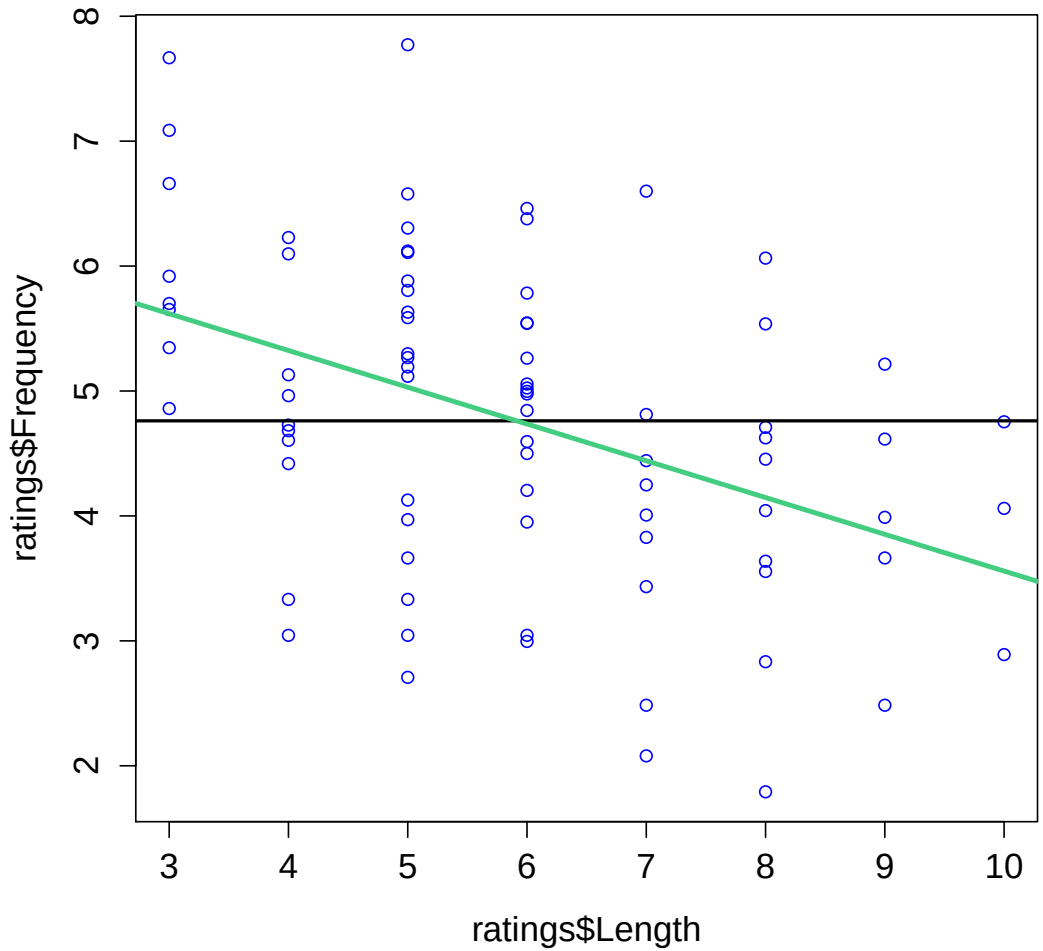


Közönséges legkisebb négyzetek

Kiindulás: (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

$$y = a + bx$$

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(függővváltozó~függetlenváltozó)`
kimenet: a és b regressziós együtthatók

Regressziószámítás az R-ben

`lm(függővváltozó~függetlenváltozó)`
kimenet: *a* és *b* regressziós együtthatók

Érdemes az eredményt eltárolni egy változóban, mert így hozzáférünk a számított értékekhez:
lmcoef = `lm(ratings$Frequency~ratings$Length)`

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

`fitted(lmcoef)`: az egyeneshez igazított (hipotetikus) *y* értékek.

`resid(lmcoef)`: reziduumok, a hipotetikus *y* értékektől való eltérések.

Egyéb elérhető adatok listázása:
`str(lmcoef)`

Regressziós egyenes ábrázolása

R-függvény:
`abline(intercept,slope)`

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

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(lmcoef))`

hiszen a `coef(lmcoef)` parancccsal épp a két szükséges együtthatót kapjuk meg.

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.

Determinációs együttható

A regressziós egyenes a legjobb közelítés az összefüggés leírására.

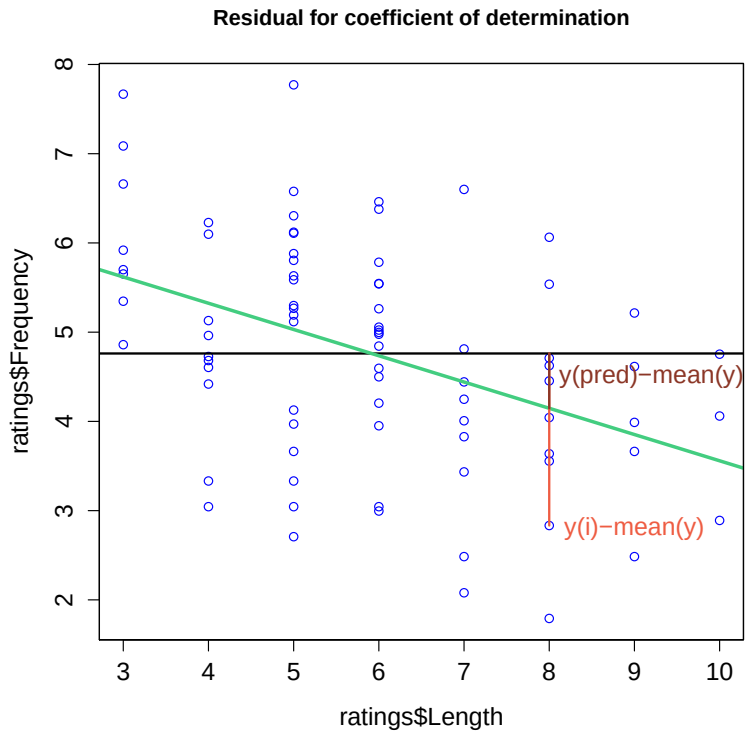
Hogyan jellemezhetjük, hogy mennyire jó a közelítés (*goodness of fit*)?

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 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 normális eloszlásúak, 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 magyarázza meg.

Determinációs együttható számítása az R-ben

```
lm(y~x)
summary(lmcoef)
```

```
lmcoef = lm(ratings$Frequency~ratings$Length)
summary(lmcoef)
```

```
Multiple R-squared:  0.1833
```

Pearson-féle $r = -0.4281$
Itt igaz az, hogy $R^2 = r^2$.

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

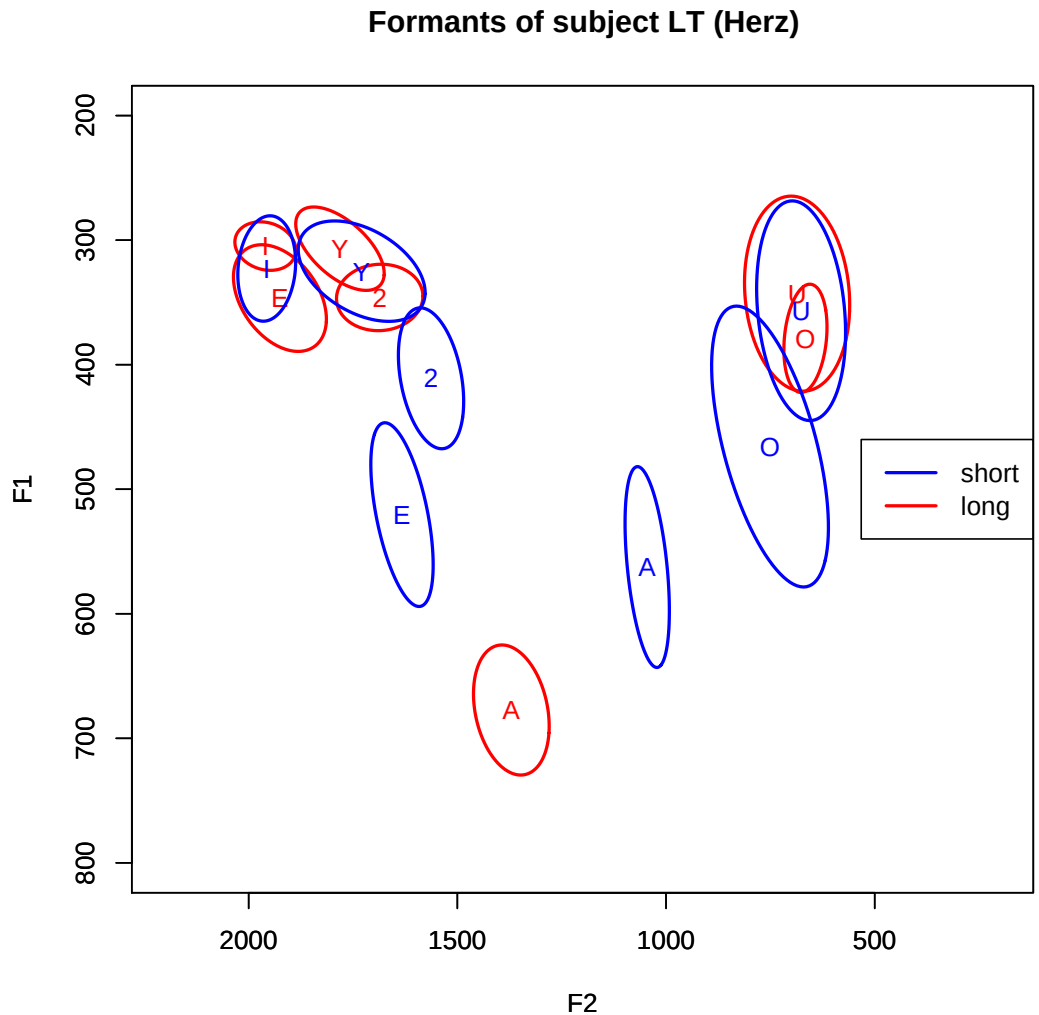
`ratings` adatmátrixban külön a növényekre és az állatokra:

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

További feladat

Háttér:

A hátsó magánhangzók 2. formánsa periférikus képzés esetén alacsonyabb, centralizáció esetén magasabb.



További feladat

Igaz-e az *á*, *ó* és *ú* magánhangzóra, hogy minél hosszabbak, annál periférikusabb az ejtésük, azaz annál alacsonyabb az F2 értékük?

longvow.RData adatmátrix letöltése a clara.nytud.hu/∼mady oldalról, a 7. óra diái mellől.
A mátrix hosszú magyar *á*, *ó* és *ú* magánhangzók tartamát és 2. formánsát tartalmazza 14 férfi ejtésében, különböző pozíciókban.

Feladat: a 26. dián megadott együtthatók kiszámítása, és a három magánhangzóra külön vagy közös ábra készítése.