

Töltsétek le a munkakönyvtárba a kuiper.RData file-t, majd nyissátok meg és futtassátok le a negyedikhazihoz2016kirak.R scriptet. Többek között a summary() függvény kerül meghívásra, ami tartalmazza az adatfile-on lefuttatott lineáris regresszió R² statisztikáját. Azt kérem, hogy a következő feladatokat megoldó parancsokkal egészítsétek ki a scriptet:

- Kérjétek korrelációt a Price és a Priceb oszlopok között a cor() függvény segítségével. A cor() függvény outputjából nyerjétek ki a korrelációt majd emeljétek négyzetre. Írjátok bele a scriptbe megjegyzésként, hogy mit tapasztaltatok!
- Ugyan a summary() kiszámolta, de hasznos ha a mi magunk írunk for ciklussal egy a Price és a Priceb változókból R² statisztikát számoló függvényt. Arra kérlek titeket, hogy ezt tegyétek meg. Egész pontosan a letöltött scriptben egészítsétek ki az üres függvényt úgy, hogy az alatta levő parancs működjön. Emlékeztetőül ideraktam az R² statisztika képletét. Kalap jelöli a becsült értékeket, míg a felülvonás átlagot jelöl. Addig próbálkozzatok amíg meg nem kapjátok a summary()-nél található értéket.

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$
- Ezután az órán látott módszertől kicsit eltérően bontsátok fel modell építő és tesztelő részre. Adjátok hozzá egy (0,1)-en egyenletes eloszlásból generált oszlopot majd a rank() függvénnyel adjátok hozzá a szimulált egyenletes változó rangszámaiból álló változót (bemenete és a kimenete is egy vektor, minden bemenetbeli elem helyére azt írja, hogy hányadik legnagyobb). Végül az építő rész álljon azokból a megfigyelésekből ahol az új rang változó 1 és 500 között van. Ennek a módszernek az az előnye az órai módszerrel szemben, hogy így pontosan tudjuk szabályozni azt, hogy hány esetet rakunk az építő és a tesztelő részbe.
- Ezután a fenti regressziót csak az építő részen futtassátok le, majd a kapott modellt a tesztelő részen használjátok előrejelzésre (legyen Priceb a becsléseket tartalmazó változó neve)
- Végül számoljátok ki a Price és a Priceb változók korrelációjának a négyzetét teszt adathalmazban. Ezután a kézzel R²-t számoló függvénnyel is számoljátok ki az R² statisztikát. Mit vesztek észre?

Megoldásképpen egy a megoldás minden lépését tartalmazó, kommentekkel ellátott R scriptet kérek.

Ha bármi kérdésetek van, keressetek nyugodtan! Jó munkát kívánok!