Gr. 232

Grigore Clara-Alexandra

Pandelea Adina-Maria

DOCUMENTATIE PROIECT

‘Swiss data set’

Setul de date “swiss” contine date pentru standardizarea masurarii fertilitatii si determinarea indicatorilor socio-economici pentru 47 de provincii vorbitoare de limba franceza din Elvetia.

Data set-ul consta in 47 observatii peste 6 variabile:

1. Fertility – rata fertilitatii
2. Agriculture (%) – cat la suta din barbati sunt implicati in agricultura
3. Examination(%)- cat la suta din cei recrutati pentru armata primesc punctaj maxim la examinarea militara
4. Education(%)-cat la suta din cei recrutati au studii peste scoala primara
5. Catholic(%)- cat la suta sunt catolici(opusul la protestanti)
6. Infant mortality- copii care traiesc < 1 an
7. Analizam media, dispersia, boxplot-ul si quartilele pentru fiecare camp, pentru intreg setul de date și de asemenea pentru fiecare variabilă am evidențiat quartilele , mediana și mustățile cu funcția summary.

Interpretare boxplot:

Cu cât cutia este mai mare, cutie care reprezintă propriu-zis doar 50% din subiecți, cu atât variabilitatea variabilei este mai mare, cum ar fi in cazul “catholic”. Opus acestui aspect avem “infant mortality”.

In cazul medianei, care reprezintă tentința central, observăm la “catholic” și “education” că aceasta se află mai aproape de marginea de jos, lucru care înseamnă că avem puține cazuri cu valori mari, dar extreme(predomină valorile mici).

Capetele ‘mustăților’ sunt valorile minime și maxime, exceptând valorile aberante(outliers).

Valorile aberante nu se înadrează în comportarea uzuală a unei variabile.

1. Pentru regresia simplă alegem ca și variabilă răspuns “fertility” și ca variabilă predictor “education”, considerând educația un factor important în vederea aproximării ratei fertilitatii.

Am calculat însă și cu variabilă predictor “agriculture“.

Am observat, comparând AIC-urile și revizuind și imaginele că într-adevăr folosind “education” putem estima mai bine rata.

Pentru regresia multipla, păstrăm “education” si adaugăm “catholic”.

Am ales aceasta variabila deorece ne axăm pe acești factori sociali care au mare efect asupra fertilității unei populații. O populatie cu un numar mai mare de catolici decat de protestanti are o rata mai mica, deoarece religia catolica este mai putin permisiva in legatura casatoriilor, de aici rezultând un număr mai mic de familii=>rata mai mică. În cazul protestanților, biserica lor ajută cetățenii în dezvoltare socială.

AIC-ul pe această regresie este și mai mic decât în cazurile regresiei simple, deci conduce la creșterea preciziei predicției variabilei “fertility”.

Regresia optima este bineînțeles cea multiplă.

1. **Repartiție Fisher**

Pentru verificarea ipotezei egalităţii dispersiilor de sondaj, obţinute în două eşantioane independente, este repartiţia Fisher.

