**2. Multipla regresija**

Primer 1: napovedovanje delovne uspešnosti nižjih vodij.

DU napovedujemo na osnovi dosežkov na testu nizov (TN; spl. inteligentnost) ter dosežkov na vprašalnikih prijaznosti (PRI), natančnosti (NAT) in vztrajnosti (VZT).

Nastavimo delovno mapo in odpremo podatke:

X <- read.table("mr1.txt", header=T) #ker imamo cela števila, ni treba dodati dec = ","

Osnovna regresijska analiza:

S funkcijo lm (»linear model«) izdelamo »regresijski objekt«:

lmr <- lm(DU ~ TN + PRI + NAT + VZT, data=X)

# splošna oblike formule: o.s. ~ n.s.1 + ns.2 itd.; konstanta je samodejno vključena!

# argument »data = X« potrebujemo, ker so spremenljivke del podatkovnega okvirja

Povzemimo ključne rezultate:

s <- summary(lmr)

Elemente povzetka lahko »izluščimo« kot nove objekte, npr.:

Delež pojasnjene variance:

R2 <- s$r.squared

Multipla korelacija:

R <- sqrt(R2)

Delež p.v. z Wherryjevim popravkom:

R2c <- s$adj.r.squared

Standardna napaka napovedi:

snn <- s$sigma

Vektor z F statistiko in prostostnimi stopnjami:

s$fstatistic

Koeficienti (s stand. napakami, t in p vrednostmi):

s$coefficients

Za obsežnejši povzetek modela uporabimo funkcijo regtabela (v datoteki z dodatnimi funkcijami; zahteva paket car):

rt <- regtabela(lmr)

Objekt rt vsebuje naslednje podatke:

* b: regresijski koeficienti;
* SE: njihove standardne napake;
* t: t vrednosti (Waldov test stat. značilnosti regresijskih koeficientov);
* p: p vrednost za dvostranski test stat. značilnosti regr. koeficientov;
* IZ: spodnja in zgornja meja 95% intervala zaupanja za regr. koeficiente
* beta: standardizirani regresijski koeficienti;
* kspr: kvadrirana semiparcialna korelacija;
* VIF: faktor povečanja kvadrata stand. napake (*variance inflation factor*);
* Tol.: toleranca.

Vse inferenčne statistike v funkciji regtabela predpostavljajo normalno porazdelitev ostankov!

Ocenjevanje parametrov z zankanjem

Intervalno oceno (po percentilni metodi) dobimo s funkcijo regzank(lmr)

Rezultati v tabeli:

* regresijski koeficienti;
* parametrične standardne napake;
* pristranskost bootstrap porazdelitve (b-b')/SE;
* standardne napake, dobljene z zankanjem;
* interval zaupanja, dobljen z zankanjem.

Primerjava modelov

Iz modela izločimo napovednik z neznačilnim nagibom:

lmr2 <- lm(DU ~ TN + PRI + VZT, data=X)

Ali je zmanjšanje pojasnjene variance statistično značilno?

anova(lmr2,lmr)

Postopna gradnja modela

Samo v izrazito eksploratornih situacijah! (Nimamo hipotez o relativni pomembnosti napovednikov.)

Funkcija step(lmr) poišče model, ki optimizira Akaikejev informacijski kriterij (AIC) – optimalno ravnotežje med parsimoničnostjo in pojasnjevalno močjo.

Napovedovanje

predict(regresijski objekt, podatki) # podatki morajo biti podatkovni okvir!

Dodajmo osebo s povprečnimi vrednostmi na napovednikih:

Xn<-rbind(c(100,50,50,50,NA),X)

predict(lmr,Xn)

Napovedane vrednosti + napovedni interval za vrednost o.s.:

predict(lmr,Xn,interval="prediction")[1,]

Napovedane vrednosti + interval zaupanja za napovedano vrednost:

predict(lmr,Xn,interval="confidence")[1,]