

Opgave 1: Graft vs. Host disease

Denne opgave er baseret på opgave 12.3 fra DG Altman, p. 361. Data omhandler knoglemarvstransplantation af 37 leukæmipatienter, og outcome er forekomst af graft versus host disease, GvHD. Formålet med opgaven er at forsøge at prediktere, hvem der får GvHD.

Filen `gvhd.txt` indeholder følgende variable:

pnr Patientnummer

rcpage Patientens alder (recipienten)

donage Donors alder

type Leukæmitype:

1:AML (akut myeloid leukæmi)

2:ALL (akut lymfatisk leukæmi)

3:CML (kronisk myeloid leukæmi)

preg Indikator for hvorvidt donor har været gravid (1:ja, 0:nej)

index Index for epidermal celle-lymfocyt reaktion (kvantitativ måling)

gvhd Forekomst af GvHD (1:ja, 0:nej)

1. Lav først en tabel, der viser hyppigheden af GvHD for de tre sygdomstyper. Er der evidens for en forskel på disse tre hyppigheder?
2. Lav nu en logistisk regression med `gvhd` som outcome og med de øvrige variable som forklarende variable, idet I dog først logaritmetransformerer `index`.
3. Fjern nu de forklarende variable successivt fra modellen, idet den mindst signifikante fjernes først (backwards elimination). Overvej gerne, om dette er en fornuftig fremgangsmåde.
 - (a) Undervejs i denne proces skal I være opmærksomme på, om estimer og P-værdier for nogle kovariater ændrer sig voldsomt ved fjernelse af andre. Se specielt på forskellen blandt leukæmigrupperne.
 - (b) I stedet for at se på leukæmitype som en kategorisk variabel med 3 niveauer kunne man have set på de to forskellige aspekter: 'kronisk versus akut' og 'lymfatisk versus myeloid'. Definer to nye 0/1-variable, der kan bruges til at vurdere hvert af de to aspekter.

I R kan man f.eks. benytte variablen

```
cml = (type==3)
```

for sammenligningen 'kronisk versus akut', idet der er den konvention, at værdien af et logisk udsagn er 1, hvis det er sandt, og 0, hvis det er falsk, så variablen ovenfor bliver 1 for CML-patienter og 0 for de øvrige.

Tilsvarende kan man benytte

```
all = (type==2)
```

for sammenligningen 'lymfatisk versus myeloid'.

Erstat den kategoriske variabel `type` med de to nye variable, og foretag en ny backwards elimination.

- (c) Overvej forskellen på resultaterne fra de to analyser i 3a og 3b.
- Find odds-ratioerne associeret med de binære forklarende variable i slutmodellen fra 3b, med tilhørende 95% konfidensintervaller.
 - Giv en verbal fortolkning af koefficienten til det logaritmerede index, gerne en, du kunne anvende, hvis du talte til en kongres.
Hvis du på baggrund af dette hellere ville have anvendt en anden logaritmefunktion, så skift den ud nu.
 - Stod der mon noget i protokollen om at undersøge nogle interessante interaktioner? Prøv f.eks. interaktionen mellem det logaritmerede index og `preg`, og giv en beskrivelse af resultatet.

Opgave 2: Mordsager

Over en årrække har man indsamlet materiale om mordsager i USA. Vi skal her se på en del af dette materiale, nemlig de sager, hvor den tiltalte blev dømt for mord. Der var i alt 191 sorte og 483 hvide, der blev dømt for mord, og af disse fik 15 sorte og 53 hvide en dødsdom.

- Opstil en 2×2 -tabel til sammenligning af sandsynlighederne for en dødsdom for hvide hhv. sorte, og kvantificer forskellen, med konfidensinterval. Ser konklusionen fornuftig ud?

Hvis vi opdeler data efter offerets race, får vi nedenstående to 2×2 -tabeller:

Sort offer:

Dømte	Dødsstraf?		Total
	nej	ja	
Sort	139	4	143
Hvid	16	0	16

Hvidt offer:

Dømte	Dødsstraf?		Total
	nej	ja	
Sort	37	11	48
Hvid	414	53	467

- Kan offerets race tænkes at være en confounder for problemstillingen?
- Lav en logistisk regressionsmodel for sandsynligheden for en dødsdom, med såvel offerets race som gerningsmandens race som forklarende variable. Sammenlign til resultatet i spørgsmål 1.
- Er der grund til at tro, at der kan være interaktion mellem offerets race og gerningsmandens race? Og er der evidens for en sådan i disse data?