ST2304 Statistisk modellering for biologer og bioteknologer

Oppgave 1

a) Normalfordelingsantagelsen kan diskuteres, ikke åpenbart at vi har en en-toppet, symmetrisk fordeling. Heller ikke at vi har uavhengige observasjoner fra identisk fordeling. Utvalget er nok i minste laget til at vi kan være sikre på at det er en god antagelse.

`pnorm(q=17.5,mean=16.14,sd=0.63,lower.tail=FALSE)` gir en sannsynlighet på 0,015.

b) Median: f.eks. `qnorm(p=0.50,mean=16.14,sd=0.63)` gir naturligvis 16.14
Øvre 95% kvantil: `qnorm(p=0.95,mean=16.14,sd=0.63,lower.tail=TRUE)` gir 17.18

Simulering (flere måter å gjøre dette på):

```
x<-rnorm(n=10000, mean=16.14,sd=0.63)
length(x[x>17.5])/ length(x[x>17])
```

c) Hvis vi fortsatt antar at hemoglobin er normalfordelt kan vi modellere variasjonen i hemoglobinnivå vha. en multippel lineær regresjonsmodell (lm). Kan så sjekke om antagelsen om lineære sammenhenger kan forsvares ved å plotte hemoglobin mot de forskjellige forklaringsvariablene. Hvis vi kaller de forskjellige aktuelle forklaringsvariablene X1, X2, X3 osv., kan kommandoen i R f.eks. se slik ut: `hem.mod<-lm(Hem~X1+X2+…+Xn)`. Kan også vurdere å ta med interaksjonsledd i modellen.

Oppgave 2

a) Vi har her en telling, så variabelen kan bare ha ikke-negative heltallsverdier, som kan stemme med en diskret Poisson-fordeling. Kravene til en Poisson-prosess er uavhengige forekomster, at forventede antall forekomster pr år pr 1000 personer er konstant innen kategori og at vi ikke kan ha to forekomster pr person pr. år. Alle disse forutsetningene skal diskuteres.

Siden antallet personer i hver alderskategori varierer, trenger vi å ta hensyn til dette ved å introdusere log(Pers) som offset variabel.

b) Generelt så trenger ikke andelene observert i hver kategori være korrekte i henhold til andelene i populasjonen siden vi modellerer raten i Poissonfordelingen for hver kategori. Men når røykere er overrepresenterte, og røyking ikke er med i modellen men kan antas å øke raten, kan raten overestimeres for hver alderskategori. Hvis det i tillegg er en avhengighet mellom røyking og alder (f.eks. flere eldre som røyker) vil det kunne påvirke den estimerte effekten av alder.

Dødsrate 40-44 år: exp(-3.39572) =0.03352, dvs. forventer 0.03352 dødsfall fra lungekreft pr 1000 personer i aldersgruppa 40-44 år.

Dødsrate 70-74 år: exp(-3.39572+2.20577)=0.30424

c) Overdispersjon – mer variasjon i datasettet enn en glm modell med (her) Poissonfordelt respons skulle tilsi.

Tegn til overdispersjon i vår modell? Ja, residual deviance = 191.72 er mye større enn forventet fra modellen (df=27) noe som kan tyde på overdispersjon. Kan testes ved `pchisq(191.72,df=27,lower.tail=FALSE)`

Hva kan ha forårsaket overdispersjonen:

- Ikke uavhengige observasjoner? Sannsynligvis liten effekt
- Viktige forklaringsvariable ikke med i modellen? Ja, sannsynlig.
- Den funksjonelle sammenhengen mellom alder og død som følge av lungekreft en annen enn den modellerte? Bør sjekkes.

Hva kan gjøres for å forbedre modellen? Ta med røyking som forklaringsvariabel, siden vi regner med at denne har mye å si. Andre potensielle forklaringsvariable? Kan korrigere for overdispersjon ved å bruke `family=quasipoisson`, hvis vi ikke har tilgang til flere aktuelle forklaringsvariable.

d) Matematisk notasjon: $\log(\lambda) = \beta_0 + \alpha_{A,i} + \alpha_{R,j} + \log(\text{Pers})$, hvor λ er forventet antall dødsfall som følge av lungekreft per 1000 personer per år, β_0 er skjæringspunkt, $\alpha_{A,i}$ er parameter for aldersgruppe *i*, $\alpha_{R,j}$ er parameter for røykekategori *j*, og log(Pers) er offset-variabelen.

Dummy variable. Variable som er 1 hvis en observasjon er i denne kategorien, og 0 ellers. Disse er nødvendige for å representere effekten av en faktorvariabel i en lineær modell

e) Modellen med røyking bedre enn modellen med kun alder? Ja – AIC avtar kraftig, effekten av å røyke sigaretter er svært signifikant og overdispersjonen er ikke lenger noe vi trenger å ta hensyn til.
R-uttrykk for om røyking skal være med i modellen? Kan f.eks. bruke drop1(Alder.Røyk.mod,test= «Chisq»). Flere alternative måter å gjøre dette på.

f) Overdispersjon? Ingen tegn til overdispersjon lenger, residual deviance er mindre enn antall frihetsgrader, noe som indikerer at p-verdien er større enn 0.5
Raten for dødsfall fra lungekreft for en person som røyker sigaretter er exp(0.41696)=1.52 ganger større enn for en person som ikke røyker.

g) Residualanalyse: Plottet f.eks. residualer mot «fitted values» - ser vi noe mønster eller trender? Sjekke om residualene har rett fordeling. Outliers, dvs. svært avvikende observasjoner?
Forslag til videre forbedringer:

- Kan evt. slå sammen kategorier (Sigar og pipe med ikke-røyk)? Som kan henge sammen med hvor mye en person røyker (de som røyker sigaretter røyker mer?) -> ha med en forklaringsvariabel på hvor mye de røyker?
- Kan vi heller ha alder som kontinuerlig variabel i modellen? Trenger vi da en transformasjon av alder (ref. figur 2)?
- Flere forklaringsvariable som kan være aktuelle, f.eks. visse yrkesgrupper som er mer utsatt pga. arbeidsmiljø eller om det er noen i familien som har hatt lungekreft (genetisk disponert)?
- Undersøke for interaksjonseffekt alder og røyking?