CMM, Predictive Modeling og Machine Learning Øvelsesopgaver, uge5

Peter Dalgaard

1. marts 2023

Opgave 1

Vi ser på datasættet SAhearts, som kan downloades fra https://web.stanford.edu/~hastie/ElemStatLearn eller i Canvas

- 1. Indlæs datasættet til en dataframe med fx read.csv. Lav et summary og check at det ser OK ud. (Kik evt i scriptet til forelæsningen i uge 4, hvis det driller.)
- 2. Fit en logistisk regression med naturlige kubiske splines for de prædiktorer der benyttes i bogens Figure 5.4 (undtagen famhist).
- 3. Tegn de fittede kurver svarende til hver prædiktor for (passende valgte) fastholdte værdier at de øvrige prædiktorer.
- 4. Funktionen predict() giver mulighed for at man kan få beregnet standard error for de fittede værdier via argumented se.fit=TRUE. Benyt dette til at tilføje approksimative 95% punktvise konfidensintervaller eller standard error bands, ligesom i Figure 5.4. (NB: Returværdien fra predict() bliver lidt anderledes end når se.fit=FALSE, så koden skal justeres tilsvarende.).
- 5. Udfør tests for om prædiktorerne kan droppes fra modeller og se om I kan reproducere Table 5.1. Forsøg evt. også test for linearitet, som i scriptet fra forelæsningen

Opgave 2

Vi ser på datasættet phoneme, som ligeledes kan downloades fra https://web.stanford.edu/~hastie/ElemStatLearn eller Canvas.

1. Indlæs datasættet til en dataframe og lav lidt præprocessering så vi får samlet de mange frekvenser i en matrix og fjernet andre lyde end "aa" og "ao". Følgende kode skulle kunne gøre det¹

- 2. Den sidste linje i ovenstående kode fitter en model med de 256 søjler i X som prædiktorer. Plot koefficienterne i fit256 (pånær den første).
- 3. I stedet for de 256 individuelle koefficienter vil vi gerne bruge en naturlig spline med 12 knudepunkter over frekvensværdierne 1,...,256. Generer en matrix ns12 af basisvektorer til dette (ns() funktionen).
- 4. For at fitte modellen med spline-parametriseringen, skal man blot danne matrixproduktet af X og ns12 og benytte dette i stedet for X i glm() kaldet fra før. Kald den fittede model fit12.
- 5. Tegn en kurve der svarer til den fittede spline-kurve af koefficienter.
- 6. Beregn AIC for de to modeller og diskuter resultatet. Prøv også at fitte modeller med flere end 12 basisfunktioner.
- 7. I det ovenstående brugte vi hele datasættet. I bogen bruges et træningsdatasæt, som er en sample på 1000 tilfældige rækker. Resultater herfra kan sammenlignes med et testdatasæt som er de øvrige rækker. Afprøv dette i praksis. Her er lidt kode til at komme i gang

```
N <-nrow(ph2)
train <- sample(1:N, 1000)
test <- setdiff(1:N, train)
fit12tr <- glm(g~X %*% ns12, family = binomial, data=ph2, subset=train)
predtst <- predict(fit12tr, newdata=ph2[test,]) >.5
table(predtst, ph2$g[test]
```

¹Det er hensigtsmæssigt eksplicit at sætte levels for g, fordi dansk "locale" kan finde på at sortere aa sammen med å, sidst i alfabetet.

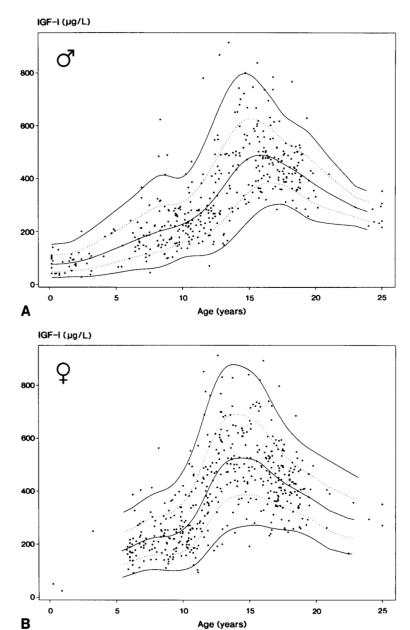
Opgave 3

Vi ser på juul datasættet i ISwR pakken (lige som til forelæsningen i uge 4)

- 1. Fit en smoothing spline for pigernes niveau af sqrt(igf1), i den gruppe er det nok bedst at begrænse sig til alder mellem 6 og 20 år (meget få er yngre end det).
- 2. Tegn data og den fittede spline kurve
- 3. Et problem i disse data er at stredningen omkring kurven ikke er konstant. Specielt er der større spredning i puberteten fordi væksten ikke sætter ind samtidig for alle børn. I Juul et al. (1994)² konstrueredes et normalområde efter følgende opskrift:
 - Fit en smoothing spline til sqrt(igf1).
 - Find residualerne i forhold til splinen (resid()).
 - Fit en spline til de kvadrerede residualer.
 - Den estimerede lokale standardafvigelse SD(x) fås ved at tage kvadratroden af den nye spline
 - Et approksimativt 95% normalområde fås ved at lægge $\pm 2SD(x)$ til den oprindeligt estimerede spline

Prøv at gentage denne konstruktion.

²Juul A, Bang P, Hertel NT, Main K, Dalgaard P, Jørgensen K, Müller J, Hall K, Skakkebaek NE. Serum insulin-like growth factor-I in 1030 healthy children, adolescents, and adults: relation to age, sex, stage of puberty, testicular size, and body mass index. J Clin Endocrinol Metab. 1994 Mar;78(3):744-52. doi: 10.1210/jcem.78.3.8126152. PMID: 8126152.



Age (years)

FIG. 2. A, The age distribution of serum IGF-I levels (micrograms per L) in healthy boys. The solid lines represent the mean as well as -2 and 2 SD, corresponding to a 95% prediction interval. The dotted lines represent -1 and 1 SD. B, The age distribution of serum IGF-I levels (micrograms per L) in healthy girls. girls.