Prøveeksamen STK2100 - vår 2017

Geir Storvik

Vår 2017

Oppgave 1
Anta en lineær regresjonsmodell

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$$

Vi kan skrive denne modellen på vektor/matrise-form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

La $\hat{\boldsymbol{\beta}}$ være minste kvadraters estimatet for $\boldsymbol{\beta}$. Vi vil i denne oppgaven se på størrelsen $\text{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ der $\widehat{Y} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$.

(a) Vis at

$$\boldsymbol{E} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = [\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}$$

der $\boldsymbol{I}_n$ er diagonalmatrisen av størrelse $n \times n$.

Hva blir forventningsvektoren og kovariansmatrisen til $\boldsymbol{E}$?

(b) Vis at

$$E[\text{RSS}] = \sigma^2(n - p - 1)$$

$\text{RSS} = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}})$.

Hint: Vis at $\text{RSS} = \text{trase}(\boldsymbol{E}\boldsymbol{E}^T)$ der trasen til en matrise er summen av diagonalleddene. Du kan videre bruke at $\text{trase}(\boldsymbol{A}\boldsymbol{B}) = \text{trase}(\boldsymbol{B}\boldsymbol{A})$ for $\boldsymbol{A}$ og $\boldsymbol{B}$ matriser av matchende størrelse.

(c) Vis at $\text{Cov}(\hat{y}_i, E_j) = 0$ for alle i, j. Diskuter dette resultatet.

Hint: Det kan her være lettere å jobbe direkte med vektorene $\hat{\boldsymbol{y}}$ og $\boldsymbol{E}$. Kryss-kovariansmatrisen mellom to stokastiske vektorer $\boldsymbol{U}$ og $\boldsymbol{V}$ kan skrives $\text{Cov}(\boldsymbol{U}, \boldsymbol{V})$ og vi har det generelle resultatet $\text{Cov}(\boldsymbol{A}\boldsymbol{U}, \boldsymbol{B}\boldsymbol{V}) = \boldsymbol{A}\text{Cov}(\boldsymbol{U}, \boldsymbol{V})\boldsymbol{B}^T$.

Oppgave 2
Vi vil her se på en regresjonssetting der vi har noen forklaringsvariable $\boldsymbol{x}$ og antar

$$Y = f(\boldsymbol{x}) + \varepsilon$$

Vi ønsker å predikere Y med en tapsfunksjon $L(y, \hat{y}) = (y - \hat{y})^2$. Vi har som vanlig data $\{(y_i, \boldsymbol{x}_i), i = 1, ..., n\}$.

(*a*) Vis at den optimale prediktor i denne situasjonen er $\widehat{Y} = f(\boldsymbol{x})$. Spesifiser spesielt hva som menes med optimalt her og hvilke antagelser dette resultatet bygger på.

(*b*) Anta nå $\widehat{Y}(\boldsymbol{x}_0) = \hat{f}(\boldsymbol{x}_0)$ for et nytt punkt $\boldsymbol{x}_0$. Vis at forventet tap kan skrives som

$$E[(Y - \widehat{Y}(\boldsymbol{x}_0))^2 | \boldsymbol{x}_0] = (f(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0)])^2 + E[(\widehat{f}(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0)|\boldsymbol{x}_0])^2 | \boldsymbol{x}_0] + \mathrm{Var}(\varepsilon)$$

Gi en fortolkning av de ulike ledd på venstre side

(*c*) La nå $\hat{f}_1(\boldsymbol{x})$ være en prediktor basert på en ganske restriktiv metode/modell mens $\hat{f}_2(\boldsymbol{x})$ er basert på en mer fleksibel tilnærming. Diskuter de ulike leddene i likningen ovenfor i denne settingen.

(*d*) Diskuter ulike metoder for å estimere forventet tap. Ta spesielt opp styrker og svakheter med ulike metoder.

Oppgave 3
I denne oppgaven skal vi se på et dataset `frogs` der interessevariabelen er tilstedeværelse av en spesifikk type frosker på ulike lokasjoner (0/1 variabel der 1 svarer til tilstedeværelse). Det er 9 ulike forklaringsvariable, alle numeriske.

En logisisk regresjonstilpasning ga følgende resultater:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−1.635e+02	2.153e+02	−0.759	0.44764
northing	1.041e−02	1.654e−02	0.630	0.52901
easting	−2.158e−02	1.268e−02	−1.702	0.08872
altitude	7.091e−02	7.705e−02	0.920	0.35745
distance	−4.835e−04	2.060e−04	−2.347	0.01893
NoOfPools	2.968e−02	9.444e−03	3.143	0.00167
NoOfSites	4.294e−02	1.095e−01	0.392	0.69482
avrain	−4.058e−05	1.300e−01	0.000	0.99975
meanmin	1.564e+01	6.479e+00	2.415	0.01574
meanmax	1.708e+00	6.809e+00	0.251	0.80198

Log-likelihood verdier for denne tilpasningen ble -97.83. Hvis y_i er responsvariabelen for observasjon i og $\hat{y}_i$ er tilhørende prediksjon, ga det følgende tabell (eller *forvirringsmatrise*):

$y \backslash \hat{y}$	0	1
0	113	20
1	24	55

(*a*) Basert på resultatene fra tilpasningen med den logistiske regresjonen, vil dette være en modell du er fornøyd med? Begrunn svaret.

En alternativ modell også basert på logistisk regresjon ga følgende resultater:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.916e+01  1.611e+01  -4.293 1.76e-05
easting     -9.236e-03  4.479e-03  -2.062  0.03921
altitude     3.217e-02  8.049e-03   3.997 6.41e-05
distance    -5.099e-04  1.837e-04  -2.776  0.00550
NoOfPools    2.969e-02  9.091e-03   3.266  0.00109
meanmin      8.916e+00  2.030e+00   4.391 1.13e-05
```

med log-likelihood verdi lik -98.71. Tilhørende forvirringsmatrise ble i dette tilfellet

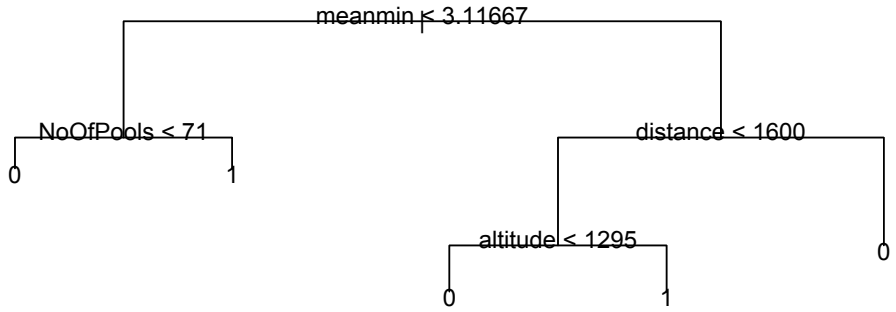
$y\backslash\hat{y}$	0	1
0	112	21
1	24	55

(*b*) Forklar hvorfor log-likelihood verdier for den nye modellen er lavere enn for den første modellen.

Bruk disse log-likelihood verdiene til å gjøre et valg mellom de to typer modeller. Spesifiser hvilket kriterie du bruker for dette valget.

(*c*) Forklar hva P-verdiene gitt i siste kolonne i de to tabeller betyr. Diskuter de faktiske verdier som kommer ut av de to modellene. Gi også en mulig forklaring på at for de forklaringsvariable som er med i begge tabeller så er P-verdiene ganske forskjellige.

Vi vil nå se på klassifikasjonstrær. Plottet nedenfor viser et estimert tre med 5 noder:



$y\backslash\hat{y}$	0	1
0	117	16
1	24	55

(*d*) Forklar hvorfor en likelihood funksjon for et klassifikasjonstre kan skrives på formen

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

der $p_i = c_m$ for $\boldsymbol{x}_i \in R_m$.

(*e*) For det spesifikke klassifikasjonstre blir log-likelihood verdien lik -90.21. Hva blir antall parametre i dette tilfelle?

Bruk dette til å beregne AIC verdien for klassifikasjonstreet og vurder denne modellen i forhold til tidligere modeller.

La oss nå i stedet se på bruk av kryss-validering. Tabellene nedenfor gir forvirringsmatriser for logistisk regresjon med 5 forklaringsvariable samt for klassifikasjonstre med 5 endenoder

	Logistisk regresjon		Klassifikasjonstre	
$y \backslash \hat{y}$	0	1	0	1
0	109	24	100	33
1	24	55	24	55

(*f*) Kommenter forskjeller mellom disse forvirringsmatrisene og de vi fikk tidligere.

Basert på disse nye forvirringsmatrisene, hvilken metode vil du foretrekke?

Tilslutt vil vi se på bagging. Forvirringsmatrisene nedenfor er basert på *out-of-bag* estimering.

	Logistisk regresjon		Klassifikasjonstre	
$y \backslash \hat{y}$	0	1	0	1
0	109	24	115	18
1	24	55	32	47

(*g*) Forklar hvordan bagging kan benyttes i forbindelse med både logistisk regresjon og klassifikasjonstrær.

Forklar hvordan out-of-bag ideen kan brukes for å klassifisere hver observasjon og dermed gi forvirringsmatriser.

Diskuter resultatene, både i forhold til kryss-valideringen tidligere og logistisk regresjon kontra klassifikasjon.

Oppgave 4

Anta $Y = f(x) + \varepsilon$ der $f(x)$ er et stykkevis kvadratisk polynom:

$$f(x) = \begin{cases} \beta_{0,1} + \beta_{1,1}x + \beta_{2,1}x^2 & \text{for } x < c \\ \beta_{0,2} + \beta_{1,2}x + \beta_{2,2}x^2 & \text{for } x \geq c \end{cases}$$

(*a*) Anta vi ønsker å legge inn begrensninger på $f(x)$ ved at funksjonen er både kontinuerlig og at den har kontinuerlig derivert. Hvilke begrensninger legger det på $\beta_{j,m}$-ene?

Hvor mange *effektive* (eller frie) parametre ender vi da opp med?

(*b*) La nå

$$g(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 (x-c)_+^2$$

der $(x-c)_+^2 = (x-c)^2$ hvis $x > c$ og 0 ellers.

Vis at $g(x)$ er kontinuerlig, har kontinuerlig derivert og er kvadratisk innenfor hver av intervallene $(-\infty, c)$ og $[c, \infty)$.

(*c*) Vis at vi kan oppnå $f(x) = g(x)$ for passende valg av $\theta_j, j = 0, ..., M+1$.

Anta nå $Y = f(x) + \varepsilon$ der $f(x)$ er et stykkevis kvadratisk polynom delt opp i flere intervaller:

$$f(x) = \beta_{0,m} + \beta_{1,m} x + \beta_{2,m} x^2 \quad \text{for } c_{m-1} \leq x < c_m$$

for $m = 1, ..., M$, $c_0 = -\infty < c_1 < \cdots < c_{M-1} < c_M = \infty$

(*d*) Anta igjen vi ønsker å legge inn begrensninger på $f(x)$ ved at funksjonen er både kontinuerlig og at den har kontinuerlig derivert. Hvilke begrensninger legger det på $\beta_{j,m}$-ene?

Hvor mange *effektive* (eller frie) parametre ender vi da opp med?

(*e*) Hvordan kan estimering av parametrene utføres?

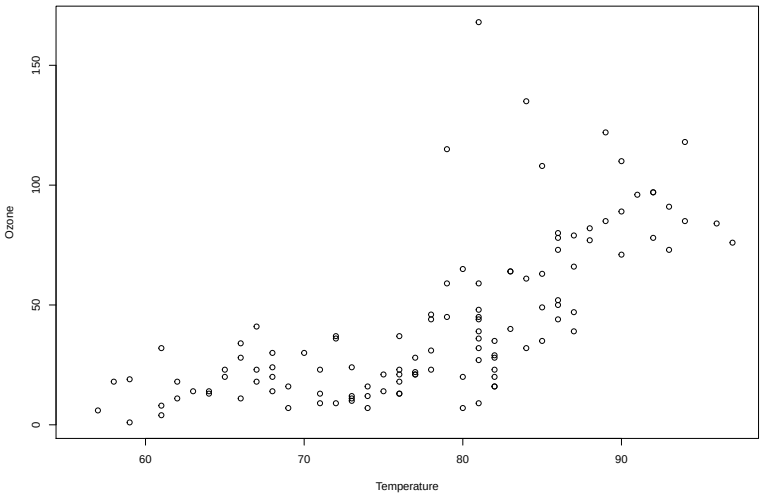
Du behøver ikke å utføre selve beregningene, bare beskrive hva slags metode som kan brukes.

Oppgave 5

Vi skal i denne oppgaven se på et datasett om luftkvalitet. Datasettet er fra mai til september 1973 og måler ozon nivå (skala er ppb=parts per billion) i New York sammen med flere andre forklaringsvariable. Vi vil i første omgang konsentrere oss om temperatur (Fahrenheit). Figuren nedenfor viser plott av ozon mot temperatur. Vi vil anta en modell

$$Y = f(x) + \varepsilon$$

der x er temperatur og Y er ozon nivå.



Vi vil i første omgang konsentrere oss om lokal regresjon i en dimensjon. La $K(x_i, x_0)$ være en vektfunksjon som angir hvor mye vekt vi skal gi til observasjon i når vi ønsker å predikere x_0. Matematisk kan dette beskrives ved at vi minimerer

$$\sum_{i=1}^{n} K(x_i, x_0)(y_i - \beta_0(x_0) - \sum_{j=1}^{d} \beta_j(x_0) x_i^j)^2$$

mhp $\beta_0(x_0), ..., \beta_d(x_0)$. Vi får da en prediksjon i punktet x_0 gitt ved

$$\hat{f}(x_0) = \hat{\beta}_0(x_0) + \sum_{j=1}^{d} \hat{\beta}_j(x_0) x_i^j$$

(a) For $d = 1$, utled de optimale estimater for $\beta_0(x_0), \beta_1(x_0)$ (det er nok å sette opp et likningssystem som estimatene må tilfredsstille).

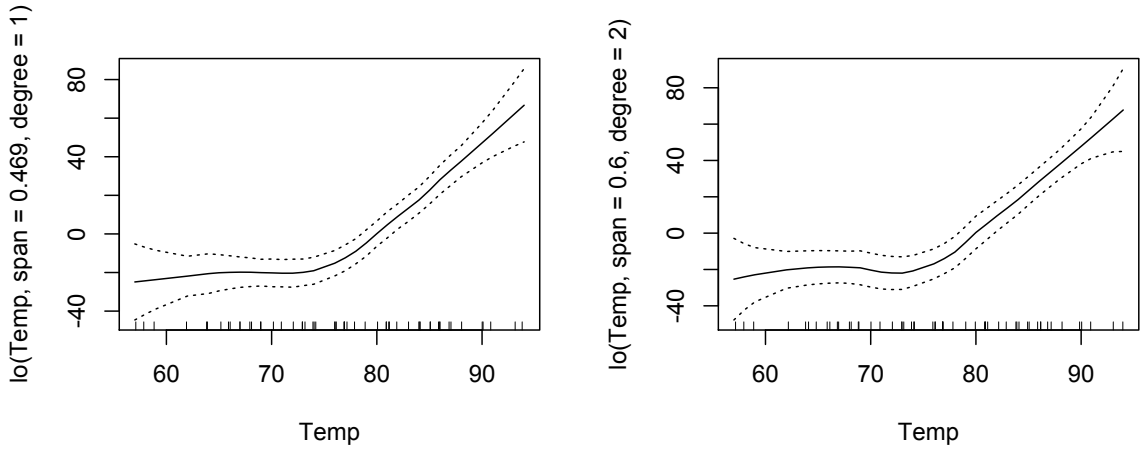
Blir $\hat{f}(x_0)$ en forventningsrett prediktor? Begrunn svaret.

Vis at $\hat{y}_i = \hat{f}(x_i) = \sum_{j=1}^{n} S_{ij} y_j$ for alle i. Argumenter for at dette også gjelder for $d = 2$.

(b) Nedenfor er plott av estimert sammenheng mellom temperatur og ozon basert på lokal regresjon med $d = 1$ (venstre) og $d = 2$ (høyre).

For disse to regresjonene er de tilhørende kjernefunksjoner $K(x_i, x_0)$ valgt slik at $\sum_{i=1}^{n} S_{ii}$ er omtrent lik for $d = 1$ og $d = 2$. Hvorfor er dette rimelig å gjøre?

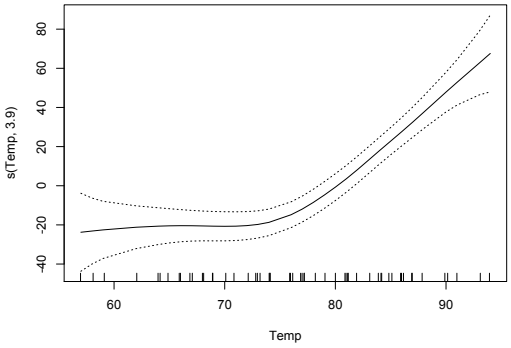
Estimert feilrate (basert på et uavhengig test sett) var 688.96 for $d = 1$ og 698.71 for $d = 2$. Basert på plottet nedenfor, argumenter hvorfor dette er rimelig i dette tilfellet.



(c) Et alternativ til lokal regresjon er splines. Nedenfor er et estimat av f ved bruk av glattingsspline (smoothing spline). Også her er $\hat{y}_i = \hat{f}(x_i) = \sum_{j=1}^{n} S_{ij} y_j$ (dette behøver du ikke å vise) og $\sum_{i=1}^{n} S_{ii}$ er omtrent lik det vi fikk for lokal regresjon.

Den estimerte funksjonen ser her ut til å være noe glattere enn hva tilfellet var for lokal regresjon. Argumenter for hvorfor dette kan være rimelig.

Feilraten på testsettet ble i dette tilfellet 694.66. Basert på disse resultatene, hvilken metode vil du foretrekke?



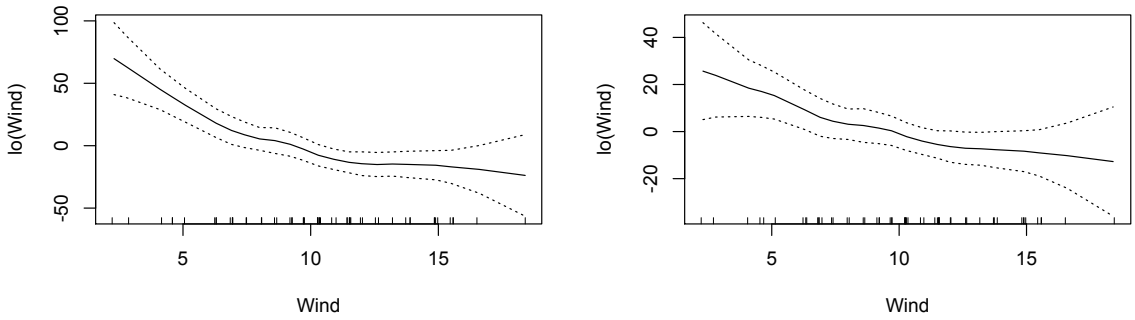
(d) Vi skal nå utvide modellen til også å inkludere vind. Vi vil nå anta en modell

$$Y = f_1(x_1) + f_2(x_2) + \varepsilon$$

der x_1 svarer til temperatur og x_2 svarer til vind. $f_1(\cdot)$ og $f_2(\cdot)$ er glatte funksjoner.

Hvilken klasse av metoder faller denne modellen innenfor?

Nedenfor viser tilpasning av $f_2(x_2)$ der $f_1(x_1) = 0$ (venstre) og der $f_1(x_1)$ er tilpasset simultant. Diskuter likeheter/forskjeller mellom de to estimatene av $f_2(x_2)$.



(*e*) Anta du har en god metode for å tilpasse en modell $Y = f(x) + \varepsilon$ der $f(\cdot)$ er en glatt funksjon. Forklar hvordan du kan bruke denne metoden for å tilpasse en modell med *to* glatte funksjoner (som i punkt (*d*)).

Oppgave 6

(*a*) Nevn minst 3 metoder relatert til regresjon eller klassifikasjon som enkelt parallelliseres?

(*b*) Vil noen av disse metodene også være minne-besparende?

Oppgave 7

Spam filtre er ofte basert på statistiske klassifikasjonsmetoder for å skille mellom reelle og spam mail. Slike metoder baserer seg bl.a. på frekvensen av ulike ord i mail. Hvis vi lar W være et ord som vi tenker opptrer oftere i spam enn i reell mail, er en mulig prosedyre å basere seg på

$$\Pr(V|S) = q > p = \Pr(V|R)$$

der V er begivenheten at W er et ord i mailen, S at en mail er spam mens R er at en mail er reell.

Anta videre at r er andelen av mail som er reelle.

(*a*) Utled en klassifikasjonsregel der du klassifiserer til spam hvis sannsynligheten for at det er en spam-mail er større enn 0.5.

(*b*) Anta nå at vi vil se mer alvorlig på reelle mail som klassifiseres som spam enn på spam mail som klassifiseres som reelle mail. Forklar hvordan dette kan formuleres matematisk og utled en optimal klassifikasjonsregel i dette tilfellet.

Anta nå at du har et sett av ord $W_1, ..., W_M$ som ofte forekommer. La V_m være begivenheten at ordet W_m forekommer i en mail. $\boldsymbol{V} = (V_1, ..., V_M)$ vil da være en vektor av binære variable (der 1 svarer til at ordet forekommer i mailen). La videre $\Pr(V_m|\mathrm{S}) = q_m$ og $\Pr(V_m|\mathrm{R}) = p_m$.

(*c*) Anta nå at $\Pr(\boldsymbol{V}|S) = \prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m}$ og $\Pr(\boldsymbol{V}|R) = \prod_{m=1}^{M} p_m^{V_m}(1-p_m)^{1-V_m}$. Hva slags antagelse bygger disse sannsynlighetene på?

Utled en klassifikasjonsregel også i dette tilfellet.

(*d*) Ofte ser en ikke bare på enkeltord men også par av ord. La $V_{m,m'}$ angi begivenheten at både ordene W_m og $W_{m'}$.

Diskuter fordeler og ulemper med en slik tilnærming.