Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

TMA4240 Statistikk
Høst 2017

Anbefalt øving 12

Denne øvingen består av oppgaver om enkel lineær regresjon. De handler blant annet om å estimere modellparametre, tolke tilpassede modeller, og predikere fremtidige observasjoner.

Oppgave 1

Figuren viser vinnertidene på 800 m løping for menn i alle Olympiske Leker (OL), altså de 28 offisielle sommerlekene mellom 1896 og 2016, samt ekstralekene i 1906.

Totalt er det $n = 29$ vinnertider. Vi lar Y_i være vinnertiden i OL nummer i, og x_i årstallet for OL nummer i. Vi antar følgende regresjonsmodell for vinnertidene:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2).$$

I tillegg antas at støyleddene $\epsilon_1, \ldots, \epsilon_n$ er uavhengige.

a) Gi en kort forklaring av minste kvadraters metode (også kalt minste kvadratsums metode eller method of least squares) for linjetilpasning.

Vis at denne metoden gir estimatorer:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}, \quad \text{og} \quad \hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

der gjennomsnittene er $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ og $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

En alternativ skrivemåte for stigningstall-estimatoren er $\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Det oppgis at $\bar{Y} = 109.0114$, $\bar{x} = 1956.6$, $\sum_{i=1}^n (x_i - \bar{x})Y_i = -6364.6$ og $\sum_{i=1}^n (x_i - \bar{x})^2 = 40\,169$. Et estimat for variansen til støyleddene er $s^2 = 3.32^2$.

b) Det kan vises at

$$T = \frac{(\hat{\beta} - \beta)}{s/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}$$

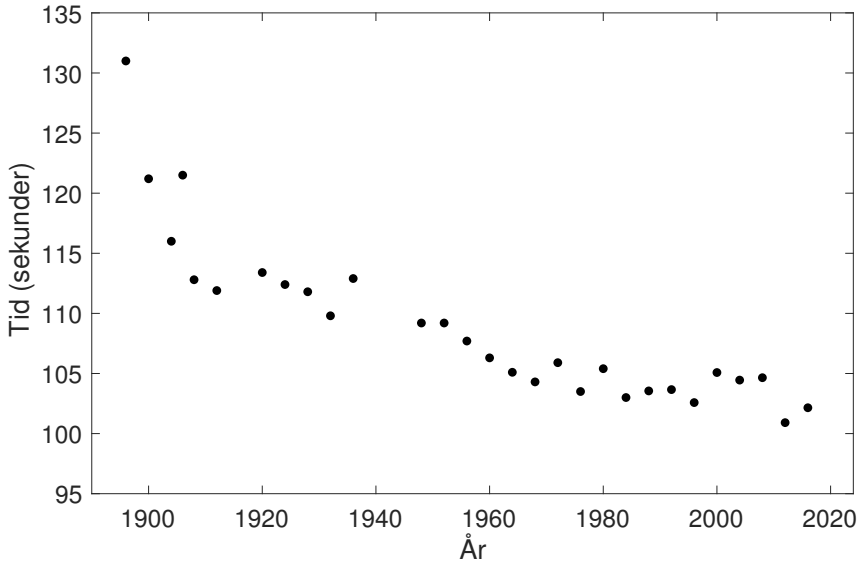
Bruk dette resultatet til å utlede et 95% konfidensintervall for β.

Regn ut konfidensintervallet ved bruk av tall oppgitt over.

Vi vil predikere vinnertiden i neste OL, altså Tokyo 2020.

c) Regn ut predikert vinnertid i 2020.

Finn et 95% prediksjonsintervall for vinnertiden i 2020.



d) Bruk modellen til å anslå året da 90-sekundersgrensen brytes, altså det første året da vinnertiden er under 90 sekunder.

Vurder modellantakelsene som gjøres. Hvilke metoder kan brukes for å undersøke antakelsene?

Oppgave 2

Kari har nylig kjøpt seg en ny bil. Nå ønsker hun å undersøke bilens bensinforbruk ved landeveiskjøring. La x være lengden av en tur (i mil) og Y tilhørende bensinforbruk (i liter). Kari forutsetter at hun selv velger lengden på turene og betrakter derfor ikke x som en stokastisk variabel, mens hun antar at Y er en normalfordelt stokastisk variabel med

$$\mathrm{E}(Y) = \beta x \quad \text{og} \quad \mathrm{Var}(Y) = x\sigma^2.$$

Dessuten antar hun at bensinforbruk på forskjellig turer er uavhengige stokastiske variable. Du skal i hele oppgaven forutsette det kjent at $\sigma^2 = 0.1^2$.

a) Hvilken tolkning har parameteren β i modellen?

Som et alternativ til antagelsen om $\mathrm{E}(Y)$ gitt over, kunne en satt $\mathrm{E}(Y) = \alpha + \beta x$. Hvorfor er det, slik Kari har gjort, mest rimelig å velge $\alpha = 0$.

Hvorfor er det rimelig å anta at variansen til Y er proporsjonal med x?

b) Anta i dette punktet at $\beta = 0.75$.

Hva er sannsynligheten for at Kari på en 5 mil lang tur vil bruke mer enn 4 liter bensin?

Betrakt to kjøreturer på henholdsvis $x_1 = 5$ og $x_2 = 10$ mil. Hva er sannsynligheten for at totalt bensinforbruk på de to turene er mindre enn 12 liter?

Betrakt igjen to kjøreturer på henholdsvis $x_1 = 5$ og $x_2 = 10$ mil og la Y_1 og Y_2 være tilhørende bensinforbruk på de to turene. Hva er sannsynligheten for at bensinforbruket

på turen på 10 mil er mer enn dobbelt så stor som bensinforbruket på turen på 5 mil? (dvs. finn $\mathrm{P}(Y_2 - 2Y_1 > 0)$)

For å undersøke bilens bensinforbruk, kjører Kari $n = 6$ turer av forskjellig lengde og måler bensinforbruket for hver tur. Målingene hennes gir følgende resultat

Lengde (mil)	5	10	20	50	100	150
Bensinforbruk (liter)	2.73	5.97	11.64	30.20	59.16	85.92

For å estimere β, betrakter Kari to estimatorer,

$$\widehat{\beta} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i} \qquad \text{og} \qquad \widetilde{\beta} = \frac{1}{n}\sum_{i=1}^n \frac{Y_i}{x_i},$$

der x_i og Y_i er henholdsvis lengde av og bensinforbruk på tur nr i.

c) Vis at

$$\mathrm{E}(\widehat{\beta}) = \beta \qquad \text{og} \qquad \mathrm{Var}(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n x_i}$$

Finn også forventningsverdi og varians for estimatoren $\widetilde{\beta}$.

Hvilken av de to estimatorene vil du foretrekke? (Begrunn svaret)

Selgeren som solgte bilen til Kari opplyste at bilens bensinforbruk ved landeveiskjøring var 0.56 liter/mil. Kari ønsker å benytte sine observasjoner til å sjekke om det er grunnlag for å påstå at bilens bensinforbruk er høyere enn hva selgeren opplyste.

d) Formuler dette som et hypotesetestingsproblem. Ta utgangspunkt i estimatoren $\widehat{\beta}$ og lag en test med signifikansnivå 5%. Hva blir konklusjonen på testen med observasjonene gitt over?

e) Ta utgangspunkt i estimatoren $\widehat{\beta}$ og utled et 95%-konfidensintervall for β. Hva blir intervallet med observasjoner som gitt over?

Oppgave 3

Når prøver fra ett og samme sted i en sølvåre analyseres med hensyn på sølvinnholdet, fåes analyseresultater som vi skal anta er uavhengige og normalfordelte med forventning μ (g/tonn) og varians σ^2.

a) Anta i dette punktet at $\mu = 500$ og at $\sigma = 80$.

La Y betegne en slik måling. Hva blir sannsynligheten for at Y skal overskride 550?

La Y_1 og Y_2 være 2 slike målinger. Hvor stor er sannsynligheten for at disse skal avvike fra hverandre med minst 80 g/tonn.

Den nevnte sølvåren er 40 meter lang og rettlinjet og går fra vest mot øst. Det er av interesse å anslå hvor mye sølv som finnes i sølvåren. Erfaringer fra andre sølvårer av tilsvarende type tilsier at sølvinnholdet i store trekk endrer seg lineært fra den ene enden av sølvåren til den andre.

La Y_j betegne målt sølvinnhold i en prøve som er tatt x_j meter fra den vestlige enden, $j = 1, 2, \ldots, n$. Vi skal anta at $Y_1, \ldots, Y_n$ er uavhengige og normalfordelte med samme ukjente varians σ^2 og forventningsverdi

$$E(Y_j) = \alpha + \beta x_j$$

der α og β er ukjente konstanter. Minste kvadratsums-estimatorene for α og β er da gitt ved, henholdsvis (du skal ikke vise dette):

$$\begin{aligned} B &= \frac{\sum_{j=1}^n (x_j - \bar{x}) Y_j}{\sum_{j=1}^n (x_j - \bar{x})^2} \\ A &= \bar{Y} - B\bar{x} \end{aligned}$$

der $\bar{x} = \frac{1}{n}\sum_{j=1}^n x_j$ og $\bar{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$.

Av praktiske årsaker bestemmer en seg for å ta 5 prøver av sølvinnholdet i hver ende av sølv åren. La $Y_1, \ldots, Y_5$ betegne målt sølvinnhold i den vestre enden ($x_i = 0$) og $Y_6, \ldots, Y_{10}$ betegne målt sølvinnhold i den østre enden ($x_i = 40$).

b) Vis at da er

$$B = \frac{\sum_{j=6}^{10} Y_j - \sum_{j=1}^{5} Y_j}{200}$$

Hva blir det tilsvarende uttrykket for A?

Finn variansen til B uttrykt ved σ^2.

Resultatet av de 10 målingene i g/tonn er gitt nedenfor:

x_i	0	0	0	0	0	40	40	40	40	40
y_i	248	176	52	98	76	682	854	870	838	806

La $\bar{y}_V$ være gjennomsnittet av målingene i den vestre enden og $\bar{y}_E$ være gjennomsnittet av målingene i den østre enden. Til hjelp for videre utregninger får du opplyst at

$$\bar{y}_V = 130,\ \bar{y}_E = 810,\ \sum_{j=1}^{5}(y_j - \bar{y}_V)^2 = 26064,\ \sum_{j=6}^{10}(y_j - \bar{y}_E)^2 = 22720$$

c) Finn et estimat for σ^2 basert på disse dataene.

Fra erfaring med andre sølvårer med relativt lite sølv i den ene enden, blir det fra økonomisk hold uttalt at β nok må være større enn 12 for at sølvåren skal være lønnsom. Gir dataene grunnlag for å påstå at $\beta > 12$? Formuler spørsmålsstillingen som en hypotesetest og utfør testingen. Hva blir konklusjonen når signifikansnivået settes til 5%?

d) En av personene i ledelsen for firmaet som eier sølvåren, hevder at det hadde blitt et sikrere estimat for β om de 10 prøvene hadde blitt tatt med noenlunde jevne mellomrom langs sølvåren. Anta at dette hadde blitt gjort, og at en fremdeles hadde $\bar{x} = 20$. Ville variansen til estimatoren for β i den gitte modellen da blitt mindre? Begrunn svaret.

Personen insisterte på at det måtte tas en ekstra prøve midt i sølvåren, dvs. for $x = 20$. Resultatet av denne ble 600 g/tonn. Vurder om denne verdien er rimelig ut i fra modellen

ved å se om den er inneholdt i et 95% prediksjonsintervall for en slik prøve. Du får opplyst at

$$Var(Y_0 - \hat{Y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)$$

der Y_0 er en ny observasjon for sølvinnholdet i et punkt x_0, $\hat{Y}_0 = A + Bx_0$, og n er antall observasjonspunkter som estimeringen er basert på.

Oppgave 4 Betrakt følgende lineære modell

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad i = 1, \ldots, 50$$

der ϵ_i er er uavhengige og normalfordelte stokastiske variabler med forventning 0 og varians σ^2. Verdier for x_i og tilhørende y_i er lagret i filen `anb12.txt` som kan lastes ned fra hjemmesiden til kurset. Du kan laste inn dataene på følgende måte:

```
A = load('anb12.txt');
x = A(:, 1);
y = A(:, 2);
```

a) Benytt Matlab til å plotte x mot y, og avgjør om korrelasjonen er positiv, negativ eller omtrent null.

Tilpass den foreslåtte lineære modellen ved å benytte Matlab. Hva er verdien til estimatene $\hat{\alpha}$ og $\hat{\beta}$? En ønsker å teste hypotesen $\mathrm{H}_0 : \alpha = 0$ mot den alternative hypotesen $\mathrm{H}_1 : \alpha \neq 0$. Benytt resultatet fra Matlab til å finne p-verdien for denne testen. Vil du forkaste H_0 ved 5% signifikansnivå?

Benytt for eksempel et normalsannsynlighetsplott og et residualplott til å diskutere om antakelsene for lineær regresjon er oppfylt.

Fasit

1. **b**) $[-0.1925, -0.1244]$ **c**) 98.88, $[91.62, 106.14]$ **d**) 2076 eller 2080

2. **b**) $0.131, 0.974, 0.5$ **c**) $\mathrm{E}(\widetilde{\beta}) = \beta, \mathrm{Var}(\widetilde{\beta}) = (\sigma^2/n)\sum_{i=1}^{n}(1/x_i)$, foretrekker $\widehat{\beta}$ **d**) $H_0 : \beta = 0.56$ mot $H_1 : \beta > 0.56$, Forkast H_0 **e**) $[0.573, 0.595]$

3. **a**) 0.266, 0.48 **b**) $\sigma^2/4000$ **c**) $s^2 = 6098$ **d**) $[281.1, 658.9]$

4. **a**) positiv, 0.043855