



Norges teknisk-naturvitenskapelige universitet  
Institutt for matematiske fag

## TMA4245 Statistikk Vår 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:

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

I tillegg antas at støyledene  $\epsilon_1, \dots, \epsilon_n$  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 støyledene er  $s^2 = 3.32^2$ .

- b) Det kan vises at

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

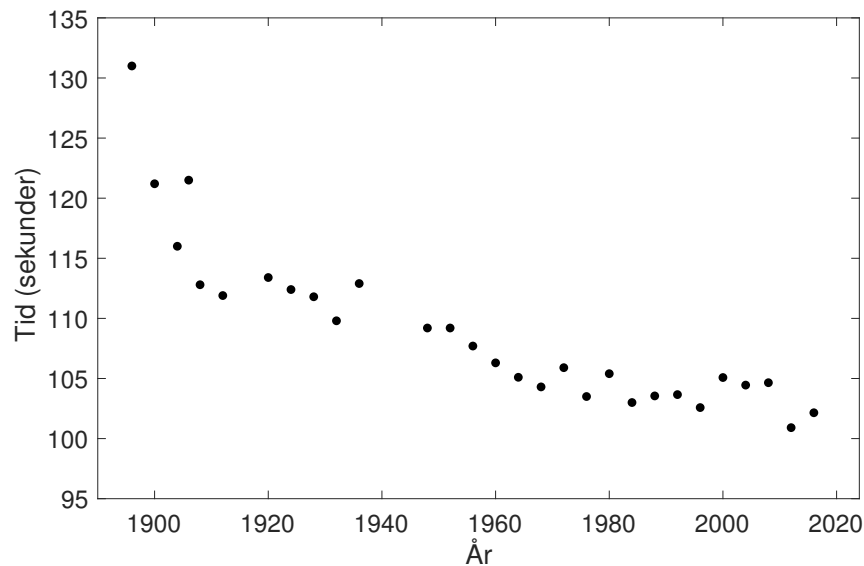
Bruk dette resultatet til å utlede et 95% konfidensintervall for  $\beta$ .

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

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  $\mu$  (g/tonn) og varians  $\sigma^2$ .

- 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,  $j = 1, 2, \dots, n$ . Vi skal anta at  $Y_1, \dots, Y_n$  er uavhengige og normalfordelte med samme ukjente varians  $\sigma^2$  og forventningsverdi

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

der  $\alpha$  og  $\beta$  er ukjente konstanter. Minste kvadratsums-estimatorene for  $\alpha$  og  $\beta$  er da gitt ved,

henholdsvis (du skal ikke vise dette):

$$B = \frac{\sum_{j=1}^n (x_j - \bar{x}) Y_j}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

$$A = \bar{Y} - B\bar{x}$$

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  $Y_1, \dots, Y_5$  betegne målt sølvinnhold i den vestre enden ( $x_i = 0$ ) og  $Y_6, \dots, Y_{10}$  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  $\sigma^2$ .

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  $\sigma^2$  basert på disse dataene.

Fra erfaring med andre sølvårer med relativt lite sølv i den ene enden, blir det fra økonomisk hold uttalt at  $\beta$  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  $\beta$  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  $\beta$  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 3

En viktig vitenskapelig oppdagelse fant sted i 1929 da Edwin Hubble oppdaget at universet er ekspanderende. Hubble's tallmateriale bestod blant annet av;  $x_i$  = avstanden til galakse  $i$  (målt i millioner lysår), og  $y_i$  = hastigheten til galakse  $i$  (målt i 1000 km/s). Verdiene Hubble benyttet i en av sine analyser er som følger:

Navn	Avstand, $x_i$	Hastighet, $y_i$
Virgo	22	1.2
Pegasus	68	3.8
Perseus	108	5.1
Coma Berenices	137	7.5
Ursa Major 1	255	14.9
Leo	315	19.2
Corona Borealis	390	21.4
Gemini	405	23.0
Bootes	685	39.2
Ursa Major 2	700	41.6
Hydra	1100	60.8

Hubble foreslo en modell for hastighet som funksjon av avstand på formen  $y = \beta x$ , der  $\beta$  senere har blitt kalt Hubble's konstant. En statistisk versjon av ligningen kan gis ved:

$$Y_i = \beta x_i + \varepsilon_i, \quad i = 1, \dots, 11, \quad (3.1)$$

der  $\varepsilon_i$ ,  $i = 1, \dots, 11$ , er uavhengige og normalfordelte stokastiske variabler med forventning 0 og varians  $\sigma^2$ .

- a) Bruk minste kvadraters metode (method of least squares) til å estimere  $\beta$  med utgangspunkt i ligning (3.1), og vis at estimatoren for  $\beta$  da blir gitt ved  $\hat{\beta} = \frac{\sum_{i=1}^{11} x_i Y_i}{\sum_{i=1}^{11} x_i^2}$ .

Finn uttrykk for forventning og varians til  $\hat{\beta}$ .

- b) Benytt Matlab til å tilpasse den lineære modellen foreslått i (3.1). Hva er estimatet for  $\hat{\beta}$ ?

**Hint:** funksjonen `fitlm(x, y, 'Intercept', false)` tilpasser den foreslåtte lineære modellen.

En ønsker å teste hypotesen  $H_0 : \beta = 0$  mot  $H_1 : \beta \neq 0$ . Hva er p-verdien til testen? Vil du forkaste  $H_0$  ved et 1% signifikansnivå?

- c) Residualene til den tilpassede modellen er gitt som  $y_i - \hat{y}_i$  for  $i = 1, \dots, 11$ . Lag et normalsannsynlighetsplott og et residualplott og diskuter om det er rimelig å anta at residualene er normalfordelte med konstant varians.

- d) Anta at en annen galakse befinner seg en avstand  $x_0 = 900$  millioner lysår borte.

Benytt Matlab til å finne predikert hastighet,  $\hat{y}_0$ , til denne galaksen.

Finn også et 95% prediksjonsintervall for en måling av hastigheten til denne galaksen.

**Hint:** funksjonen `predict(modell, 900, 'Prediction', 'observation')` kan være nyttig.

**Oppgave 4** Betrakt følgende lineære modell

$$Y_i = \alpha + \beta x_i + \epsilon_i \quad i = 1, \dots, 50$$

der  $\epsilon_i$  er uavhengige og normalfordelte stokastiske variabler med forventning 0 og varians  $\sigma^2$ . 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  $H_0 : \alpha = 0$  mot den alternative hypotesen  $H_1 : \alpha \neq 0$ . 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. **a)** 0.266, 0.48 **b)**  $\sigma^2/4000$  **c)**  $s^2 = 6098$  **d)** [281.1, 658.9]
3. **b)** 0.056691,  $4.7825 \times 10^{-16}$  **d)** 51.0220, (48.4969 ,53.5471)
4. **a)** positiv, 0.043855