TMA4245 Statistikk
Vår 2016

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Anbefalte oppgaver 12, blokk II

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).

Totalt er det $n = 28$ 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:

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

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 estimatoren av stigningstallet 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 $\bar{Y} = 109.26$, $\bar{x} = 1954.5$, $\sum_{i=1}^n (x_i - \bar{x})Y_i = -5942$ og $\sum_{i=1}^n (x_i - \bar{x})^2 = 36517$. Et estimat på variansen til støyleddene er $s^2 = 3.40^2$.

b) Det kan vises at

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

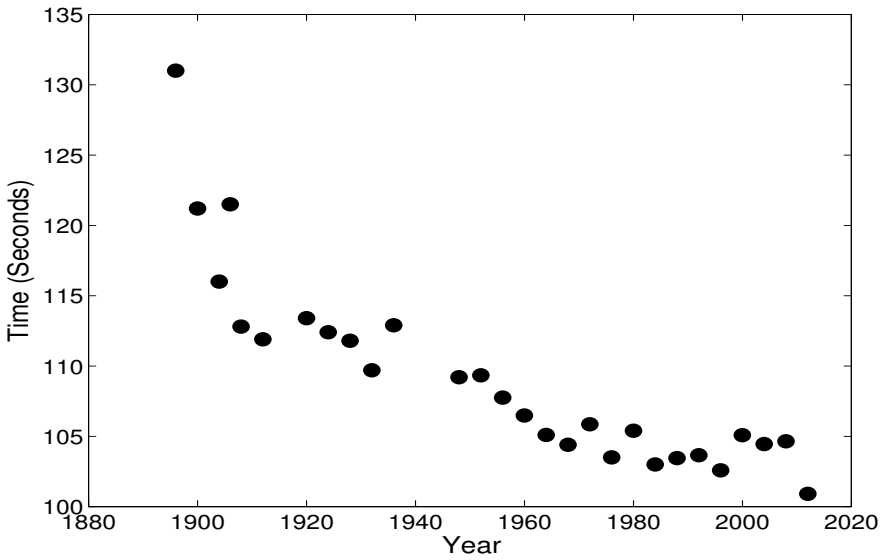
Bruk dette resultatet til å utlede et 95 prosent konfidensintervall for β.

Regn ut konfidensintervallet ved bruk av tall oppgitt over.

Vi vil predikere vinnertiden i neste OL: 2016 i Brasil.

c) Regn ut predikert vinnertid i 2016.

Finn et 95 prosent prediksjonsintervall for vinnertiden i 2016.



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 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 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

Det oppgis her at $\sum_{i=1}^{11} x_i = 4185$, $\sum_{i=1}^{11} y_i = 237.7$, $\sum_{i=1}^{11} x_i^2 = 2685141$ og $\sum_{i=1}^{11} x_i y_i = 152224$.

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

$$Y_i = \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, 11, \tag{3.1}$$

der ε_i, $i = 1, \ldots, 11$, er uavhengige og normalfordelte stokastiske variabler med forventning 0 og varians σ^2.

a) Vi vil i første omgang finne en estimator for β.

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

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

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

Finn predikert hastighet, $\hat{y}_0$, til denne galaksen.

Utled et 95% prediksjonsintervall for en måling av hastigheten til denne galaksen. Det oppgis at $\sum_{i=1}^{11}(y_i - \hat{y}_i)^2 = 9.87$, der $\hat{y}_i = \hat{\beta} x_i$.

Fasit

1. **b**) $[-0.199, -0.126]$ **c**) 99.3,$[91.8, 106.7]$ **d**) 2073

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

3. **a**) 0.0567 **b**) 51.03, (48.5,53.5)