TMA4240 Statistikk
Høst 2015

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Øving nummer 12, blokk II

I denne siste øvingen fokuserer vi på lineær regresjon, der vi har kjente kovariater x, og stokastiske responsvariabler Y, typisk beskrevet av modellen $Y_i = \alpha + \beta x_i + \varepsilon$. I denne øvingen ser vi på estimering av parameterne α og β, samt hypotesetester og konfidensintervall for disse. Vi regner også på prediksjon av en fremtidig måling y. Parameterestimering, konfidensintervall, prediksjonsintervall og hypotesetester har vi sett i tidligere øvinger, og i denne øvingen bruker vi disse konseptene i lineær regresjon.

Oppgave 1

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 2

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:

$$Y_i = \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, 11, \tag{2.1}$$

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 (2.1), 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 $\hat{y}_i = \hat{\beta} x_i$.

Oppgave 3

I medisin er det nyttig å studere vekten til nyfødte som funksjon av deres terminalder (*gestational age* eller tid siden unnfangelse). Data er her terminalder x_i (uker) og vekt y_i (gram) for $i = 1, \ldots, n$ babyer, og $n = 24$.

For dette datasettet har vi $\sum_{i=1}^{n} x_i y_i = 2\ 752\ 667$, $\sum_{i=1}^{n} x_i^2 = 35\ 727$, $\sum_{i=1}^{n} x_i = 925$ og $\sum_{i=1}^{n} y_i = 71\ 194$.

Anta en lineær regresjonsmodell: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, der ϵ_1, ..., ϵ_n antas uavhengige og normalfordelte med forventning 0 og varians σ^2.

a) Bruk oppsummeringen av tallmateriale gitt over til å regne ut estimatene for skjæringspunkt og stigningstallet for regresjonsmodellen: $\hat{\beta}_0$ og $\hat{\beta}_1$.

Vi regner ut et estimat for σ^2 ved $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 194^2$.

Regn ut et 95 prosent konfidensintervall for stigningstallet.

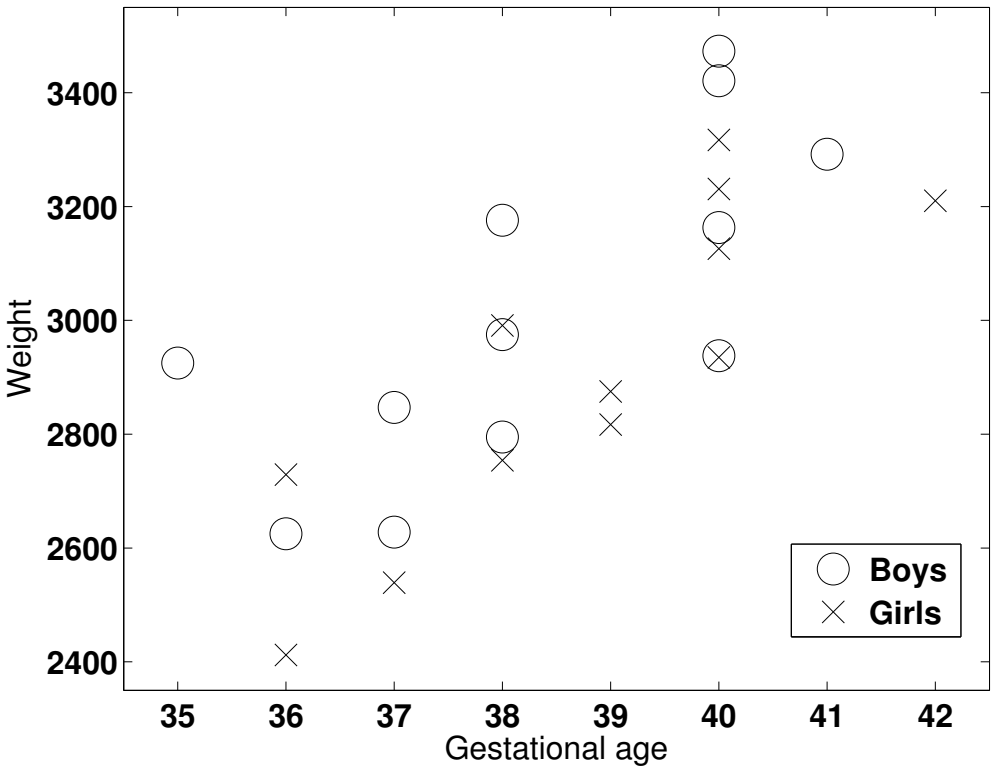
b) Bruk data til å finne et 90 prosent prediksjonsintervall for vekten til en nyfødt i terminuke 40.

Hvor bredt er 90 prosent prediksjonsintervallet for terminuke 42 sammenlignet med det vi fant for uke 40?

Figur 1 viser et kryssplott av terminuke og vekt. I dette plottet er data delt inn i to grupper: gutter og jenter. Det er $n_b = 12$ gutter (nummerert 1 til n_b) og 12 jenter (nummerert $i = n_b+1$ til n).

Vi foreslår følgende modell for data

$$Y_i = \beta_b + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n_b.$$



Figur 1: Kryssplott av terminuke og fødselsvekt for 12 gutter og 12 jenter.

$$Y_i = \beta_g + \beta_1 x_i + \epsilon_i, \quad i = n_b + 1, \ldots, n.$$

der vi fortsatt antar at ϵ_1, ..., ϵ_n er uavhengige og normalfordelte med forventning 0 og varians σ^2.

c) Bruk plottet til å forklare hvorfor denne modellen kan være hensiktsmessig. Forklar videre hvilke elementer av modellen som kan være uønsket.

Regn ut minste kvadratsums estimater (eller maximum likelihood estimater) for parametrene i modellen, her benevnt ved $\hat{\beta}_b$, $\hat{\beta}_g$ og $\hat{\beta}_1$. I tillegg til summene gitt tidligere i oppgaven har vi at $\sum_{i=1}^{n_b} y_i = 36\ 258$, $\sum_{i=1}^{n_b} x_i = 460$, $\sum_{i=n_b+1}^{n} y_i = 34\ 936$ og $\sum_{i=n_b+1}^{n} x_i = 465$.

Fasit

1. **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**) $[0.573, 0.595]$

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

3. **a**) -1465, 115, [69,161] **b**) [2789,3481], 690, 732 **c**) -1587, -1747, 120.22