Stat100, vår 2018

**Oppgaver til Uke 18**

**Oppgave 1 (du kan regne alt for hand, men det går utrolig mye raskere i R-commander.)**

Anta at du har følgene data.

x 1 2 3

Y 1 3 2

Anta den lineære regresjonsmodellen

Yi =  + xi + i, der i = 1,2 3, og i ~N(0, ). Alle feilledd () er uavhengige.

**Gjort uke 17:**

Finn korrelasjonskoeffisienten mellom x og Y.

Tilpass en rett linje med hjelp av minste kvadraters metode.

Beregn tilpassede Y-verdiene.

Beregn residualene.

Beregn SSReg, SSE og SST.

Estimer .

Finn R2.

**Uke 18:**

Finn standardfeilen til .

Lag et 95 % Konfidensintervall for .

Test om det er sammenheng mellom x og Y.

Test om det er positiv sammenheng mellom x og Y.

Finn testens p-verdi(R-commander)

Estimer forventet Y når x = 1,5

Lag et 95 % konfidensintervall for E(Y|x=1,5)

Prediker Y når x = 1,5.

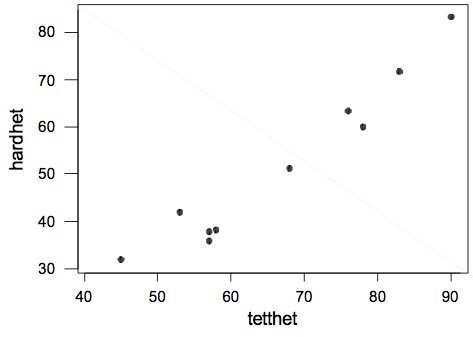
Finn 90 % prediksjonsintervall for Y når x = 1,5.

#### Oppgave 2:

Hardhet i tre er en særdeles viktig egenskap knyttet til kvalitet av for eksempel parkett. Hardhet bestemmes av hvilken kraft en må bruke for å presse en kule med diameter 5.64 millimeter helt ned i treverket. Problemet er at hardhet er vanskelig å måle. Derimot er tetthet (100\*vekt/volum, det vil si vann har tetthet 100) enkelt å måle. En håper derfor at sammenhengen mellom hardhet og tetthet er så god at en kan bruke tetthet til å predikere hardhet.  Følgende data gjelder for forskjellige treslag.

http://canvastest.bibsys.no/assessment_questions/2372/files/2820/download?verifier=0CjsuaP2jWs7DAZ2GgG88STxid1jmtOQ56bnhlah&wrap=1

 Sammenhengen ble undersøkt grafisk ved hjelp av spredningsdiagrammet (scatter plot) under.

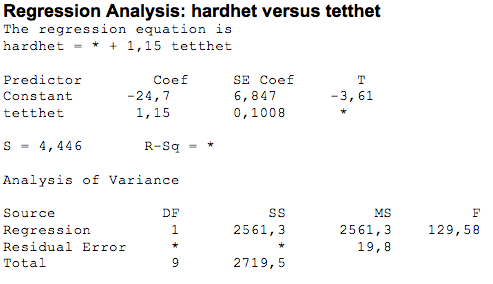


1. Sett opp den lineære regresjonsmodellen.

Tolk alle parameterene i modellen.

Synes du modellen passer til data?

Analysen ble utført i Minitab (en statistikkpakke som minner om R-commander), og dette ga følgende resultat, men der enkelte tall er erstattet med en \*.



**b)**  Fyll ut de manglende verdiene og oppgi alle parameterestimatene.

Er det grunn til å tro at det er en sikker sammenheng mellom tetthet og hardhet? (Formuler dette som et hypotesetestingsproblem og utfør testen.)

#### Finn et 95 % prediksjonsintervall for hardhet for et tre med tetthet lik gjennomsnittlig tetthet for det de trær som allerede er med i modellen.

**Oppgave 3.**

Gjør e, f, og h oppgave 3 uke 17.

**Oppgave 4.**

Gjør b og c oppgave 4 uke 17.

**Oppgave 5**

Gjør b oppgave 5 uke 17

**Oppgave 6.**

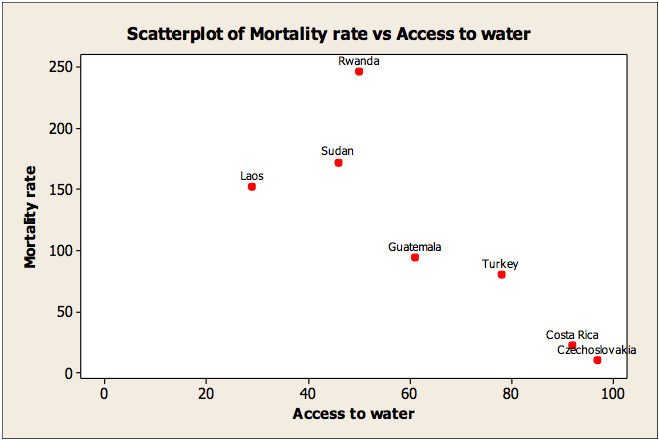
Gjør b oppgave 6 uke 17

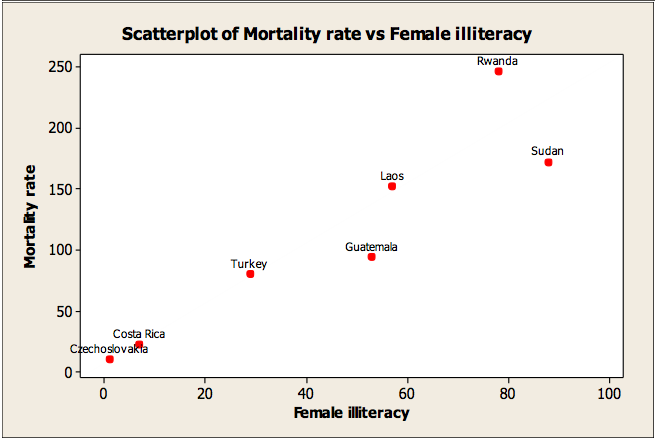
**Oppgave 7**

Gjør c og d oppgave 7 uke 17

**Oppgave 8:**

Årlig publiserer FN en rapport som kalles «Tilstand for verdens barn». Fra 2007-utgaven har vi plottet informasjon fra sju svært ulike land:





*Mortality rate* er årlige antall på døde barn under 5 år per 10 000 levende fødte. Senere i oppgava kalt barnedødelighet.

*Access to water* er prosentandel av befolkningen som har tilgang til rent drikkevann.

*Female illiteracy*er prosentandel av kvinner over 15 år som *ikke*kan lese.

Vi vil ha *Mortality rate* som responsvariabel, men prøver med to ulike forklaringsvariable.

Modell 1: La*access to water* være forklaringsvariabel

Modell 2: La*Female illiteracy*være forklaringsvariabel.

På slutten av denne oppgava finner du to Minitabutskrifter.

**a)**

For modell 1: Sett opp en lineær regresjonsmodell og tolk alle parametrene. Hvilket fortegn synes du det er naturlig at regresjonskoeffisienten () har?

For modell 2: Sett opp en lineær regresjonsmodell.

Hvordan vil du estimere standardavviket i barnedødelighet for alle land med samme prosentandel kvinnelige analfabeter?

Hvordan ville du estimert forventet barnedødelighet dersom alle kvinner kunne lese?

**b)**

Test påstanden om at det er sammenheng mellom tilgang på rent vann og barnedødelighet.

Test påstanden om at det er sammenheng mellom kvinnelig analfabetisme og barnedødelighet.

**c)**

Finn determinasjonskoeffisienten (R2) for Modell 2.

Hva betyr det at R2 er større i Modell 2 enn i Modell 1?

d)

I Vietnam har 42 % av befolkningen tilgang på rent vann mens 84 % av kvinnene kan

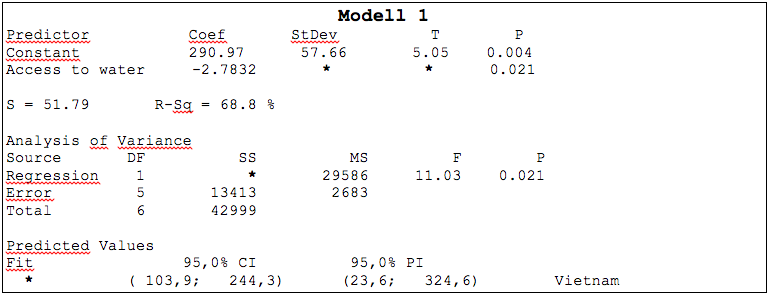
lese. I Norge har 99 % tilgang til rent vann.

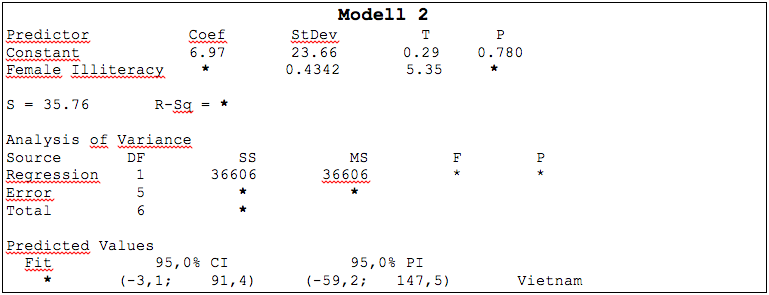
For begge modeller anslå barnedødelighet for Vietnam.

Hvorfor tror du modellene gir så forskjellige anslag?

Hva er den viktigste årsaken til at prediksjonsintervallet blir smalest for Modell 2?

Hvorfor vil et prediksjonsintervall for Modell 1 bli bredere for Norge enn for Vietnam?





**Oppgave 9:**

Denne oppgaven dreier seg om et datasett over ville bjørner, som er blitt bedøvet og deretter veid og målt på ulike måter. Et formål med studien er å kunne si noe om bjørners vekt basert på data om kroppsstørrelse og kjønn. Dette kan være nyttig ved framtidige observasjoner av bjørner som kan måles, men hvor en vekt ikke er tilgjengelig. I denne oppgaven skal vi benytte bjørnenes brystomkrets og deres nakkeomkrets som mulige forklaringsvariable (begge målt i cm). Vekt (målt i kg) er responsen.

Vi bruker modellene:

Modell 1: Yi = xi + i, der alle i er uavhengige, og der i ~ N(,). (Brystomkrets)

Modell 2: Yi = xi + i, der alle i er uavhengige, og der i ~ N(,). (Nakkeomkrets)

Yi er vekt på bjørn i, og xi er henholdsvis brystomkrets og nakkeomkrets på bjørn i.

i = 1, 2, . . . . . , 24.

**a)**  For begge modeller:

Gi en tolkning av alle parametre (3 i hver modell).

Estimer begge regresjonsparametre ( og )

Vis at for Modell 1 er = 21,12 og for Modell 2 er= 28,01.

**b**) Velg den modellen du synes passer best.

Lag et 95 % konfidensintervall for  i denne modellen.

Forklar hvordan du ved hjelp av dette konfidensintervallet kan teste hypotesen om at det er en sammenheng mellom vekt og forklaringsvariabel.

**c**) Vi skyter en ny bjørn og måler brystomkrets til 90 cm og nakkeomkrets til 54.3 cm.

Bruk begge modeller til å anslå vekt på denne bjørnen.

Bruk begge modeller til å lage et 99 % prediksjonsintervall for vekt på denne bjørnen.

Hvorfor blir prediksjonsintervallet bredere ved bruk av Modell 2 framfor Modell 1?

**d**) I Figuren neste side har vi plottet residualer mot tilpassede y-verdier for Modell 1. På plottet er det også markert forskjellig kjønn.

Forklar hvordan vi kan bruke et residualplot til modellevaluering.

Kan du ved hjelp av Figur 1 se noen svakheter ved modellen?

Brystomkrets 90.0

Nakkeomkrets 54.3

*Tabell 1:* Gjennomsnittsverdier for brystomkrets og nakkeomkrets.

Coefficients:

Estimate Std. Error t value

(Intercept) -133.52 16.68 -8.005

Brystomkrets 2.73 0.18 15.276

Df Sum Sq

Brystomkrets 1 104124

Residuals 22 9816

*Tabell 2:* Resultater for **Modell 1**.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Coefficients:

Estimate Std. Error t value

(Intercept) -118.23 21.57 -5.482

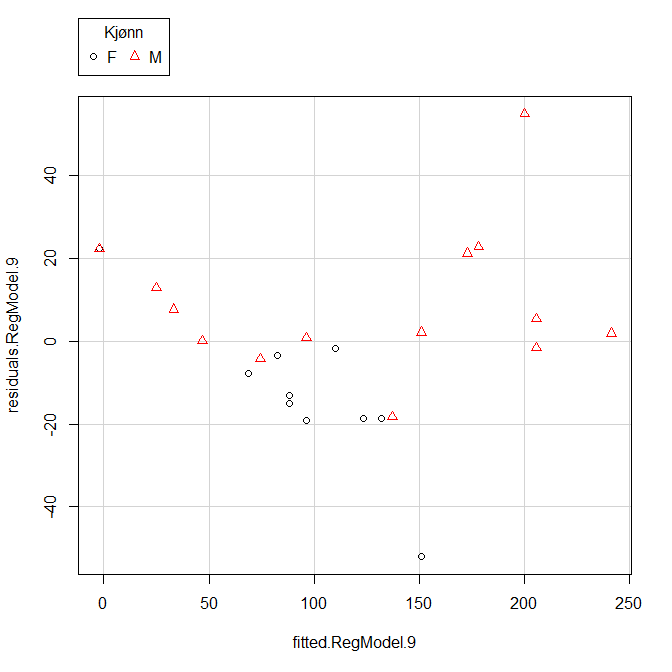
Nakkeomkrets 4.25 0.38 11.100

Df Sum Sq

Nakkeomkrets 1 96678

Residuals 22

*Tabell 3:* Resultater for **Modell 2.**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**Blant mange andre så tar disse eksamensoppgavene opp prediksjon og modellkritikk.**

Eksamen høst 2011 nr. 2 særlig 2d og 2e,

Eksamen høst 2010 nr. 3

Eksamen vår 2010 nr. 2 særlig 2c og 2d.

Eksamen vår 2007 nr. 2

Eksamen vår 2015 nr. 1.

Eksamen høst 2015 nr. 2

Eksamen vår 2106 nr. 1

Eksamen høst 2016 tekstoppgave 2

