#### Stat100, vår 2018

#### Oppgaver til uke 17

**Teorioppgave kun for matematisk orienterte studenter.**

Anta modell Yi =  + xi + i, da vil i = Yi – ( + xi)

Minste kvadraters metode går ut på å finne de verdier av  og  som minimerer kvadratsummen av feileddene.

Minimer: Q = 

Partiell-deriver Q med hensyn på  og , sett begge partiell-deriverte til null og løs ligningssystemet.

#### Oppgave 1. Alle skal prøve på denne. Bare lommeregner er tillatt.

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.

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.

**Oppgave 2** Anta den lineære regresjonsmodellen

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

Gi en tolkning av **alle** parametere (hvis de kan tolkes) og av i i de følgende situasjonene:

* Vekt i kg er respons, høyde i cm er forklaringsvariabel (gjelder voksne menn).
* Høyde i cm er respons, alder i år er forklaringsvariabel (gjelder for barn opp til 14 år)
* Bensinforbruk liter/mil (for bil) er respons, motorstørrelse i hestekrefter er forklaringsvariabel
* Eksamenspoeng i stat100 er respons, ukentlig timer nedlagt i stat100 forberedelser gjennom semesteret er forklaringsvariabel.

**Oppgave 3 (R-commander, data finner du på fronter bil.Rdata, oppgaven er nesten identisk med kollokvieoppgave 19. april)**

En dame vil selge bilen sin på internett. Hun eier en Toyota Yaris 2001 modell som har kjørt 85000 km. På internett finner hun 10 biler som er like hennes bil, bortsett fra årsmodell og km. stand. Under ser du data for de bilene hun har funnet. Prisen er gitt i antall tusen kr, mens tusenkm er i antall tusen km bilen har kjørt. Aar viser hva slags års modell det er, gitt som antall år etter år 2000

bilnr pris aar tusenkm

1 89 4 49

2 179 6 53

3 145 5 50

4 115 4 49

5 95 0 100

6 120 3 65

7 75 0 123

8 60 0 164

9 110 2 68

10 125 4 37

Lag to spredningsplot (scatterplot) der respons er pris og forklaringsvariabel enten er alder eller kjørte km.

1. Tilpass to modeller, der den ene forklarer pris med alder, den andre forklarer pris med kjørte kilometer. Hvilken av de to modellene ser ut til å ha best tilpasningsevne til data?
2. Estimer alle parametere i begge modeller, og gi estimatene en tolkning som en potensiell bilkjøper kan forstå.
3. Anta at hun en dag kjører 100 kilometer i bilen sin. Hvor mye ville du estimere prisfallet til å bli som en følge av denne turen?
4. Hun vurdere å selge bilen sin, men vil vente et halvt år, hvor mye vil du estimere prisfallet til å bli i forhold til å selge den med en gang?
5. **Vent til uke 18**

Beregn et 98% konfidensintervall for stigningstallet i den modellen du mener passer best av de 2 tilpassede modellene. Hvordan vil du forklare dette intervallet? (

1. **Vent til uke 18**

Bruk begge de tilpassede modellene til å anslå hvilken pris hun bør sette på bilen sin.

Gir modellene svært forskjellig svar?

Lag et 99 % prediksjonsintervall for pris for hennes bil ved bruk av begge modellene.

**g)** Bruk begge de tilpassede modellene til å anslå hva hun må betale i mellomlegg hvis hun bytter bort bilen sin med en bil som er maken til hennes bortsett fra at det er en 2005 modell som har kjørt 37000 km.

**h)**  **Vent til uke 18**

En selger påstår at denne biltypen i snitt faller med mer enn 10 000 kroner hvert år. Anser du dette som bevist (Test på 5 % nivå, dette får du ikke direkte ut av datamaskin).

1. *For spesielt interesserte (stat200 pensum)*. Dersom du vil inkludere både år og kjørte kmi modellen (som x-variable), prøv å gjøre dette i R-commander. Prøv å sette opp modellen. Kan du tolke paramet**e**re? Hvorfor tror du kjørte km ikke lenger har særlig betydning?

**Oppgave 4**

**(R-commander. Data finner du på fronter Slakt.Rdata)**

I et slakteri har nyslaktede dyr typisk en høy pH-verdi. Vi ønsker å utføre en spesiell type kjøttbehandling på slaktet, men denne behandlingen er avhengig av at pH verdien er sunket til rundt 6,0. Siden det er veldig ressurskrevende å overvåke pH-verdien til hvert enkelt slakt, ønsker vi å kunne predikere pH basert på hvor lang tid det har gått fra slakting. For å finne sammenheng mellom pH og tid fra slakting tok en 10 forskjellige okseslakt, lot disse henge et visst antall timer (Tid) og målte pH verdier. Dette ga:

Tid pH­­­­­­­­­­­­­­­­­­­­­­­­­

1 7.02

1 6.93

2 6.42

2 6.51

4 6.07

4 5.99

6 5.59

6 5.80

8 5.51

8 5.36

**a)**      Sett opp den lineære regresjonsmodellen der pH er respons, og tid er forklaringsvariabel.

Gi en tolkning av alle parametere i modellen.

Hvordan vil du estimere forventet pH i nyslakta dyr?

**b)**   **Vent med b og c til uke 18**

Lag et 95 % konfidensintervall for regresjonskoeffisienten.

Gi en forklaring på intervallet til slakteriarbeiderne.

Bruk intervallet til å teste om pH endrer seg med tiden etter slakting.

Hvilket nivå tester du på?

**c)**   Hva er et residual? Svar konkret for denne undersøkelsen?

Finn residualene for de to slaktene som har hengt i en time.

Bruk residualene til å sjekke, linearitet, konstant varians og normalfordeling.

**d**) Hvor lenge bør et slakt henge i følge denne modellen hvis det skal få optimal Ph?

En prøvde seg deretter med logaritmen til tid etter slakting (med naturlig grunntall) som forklaringsvariabel (kall dette Modell 2). Bruk: Data – Manage variables in active dataset – compute new variables. Merk at for logaritme ned naturlig grunntall skriv: log(), der du klikker tid inn i parentesen.

**e)** Vis at R2 ved Modell 2 er 0,98.

Gi to begrunnelser for å fortrekke Modell 2.

1. (**Vanskelig oppgave, kun for spesielt interesserte)**

Anta Modell 2.

Hvorfor kan du ikke estimere pH for nyslakta dyr?

Estimer pH reduksjon pr. time.

Hvor lenge bør i følge denne modellen et slakt henge for å oppnå en pH på 6?

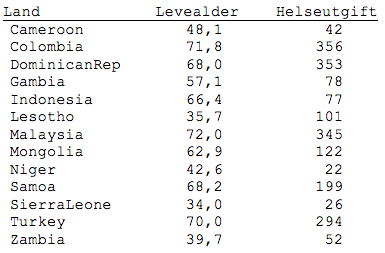
For alle delspørsmål: Sammenlign med tilsvarende resultat dersom du bruker tid som forklaringsvariabel.

#### Oppgaver med ferdig utskrift (eller deler av utskrift).

#### De fleste av disse er tidligere eksamensoppgaver, men som er utvidet med enkelte eksamensrelevante spørsmål.

#### Oppgave 5:

Det ble foretatt en verdensomspennende undersøkelse om helse og levealder. I utgangspunktet var alle land med, men vi har plukket ut 13 land.

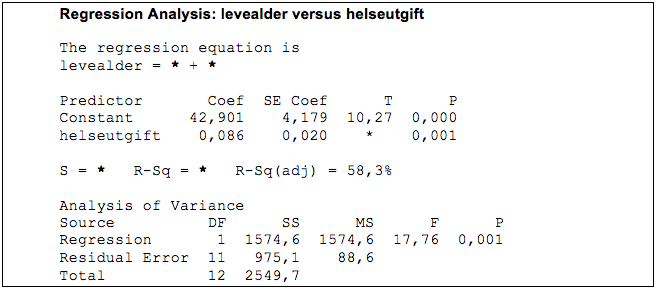


La responsvariabel (Y) være levealder og la forklaringsvariabel (x) være helseutgift, målt i innsats pr person (i 2001).

Vi antar følgende modell: Yi = xi  + \epsilon_i,

der \epsilon_i -ene er uavhengige og normalfordelt med forventning 0 og standardavvik .

Vi kjørte modellen i Minitab (gammelt dataprogram brukt tidligere) og fikk følgende resultat:



**a)** Gi en tolkning av alle parametere i modellen og finn de estimerte verdiene.

En del av verdiene i utskriften er erstattet med en **\***. Fyll ut de riktige verdiene.

**b)** (**Vent med b og c til uke 18.)**

Lag et 99 prosent konfidensintervall for , og forklar hva dette intervallet sier.

**c)** På Figur 1 ser du et scatterplott med inntegnet estimert regresjonslinje.

Ser det ut til at modellen passer til data?

Hvordan tror du et residualplot vil se ut? (Bare skisser det).

*En påstand er slik: En økning av helse budsjettet på en dollar i et gir langt større gevinst i et fattig land enn i et rikt land.*

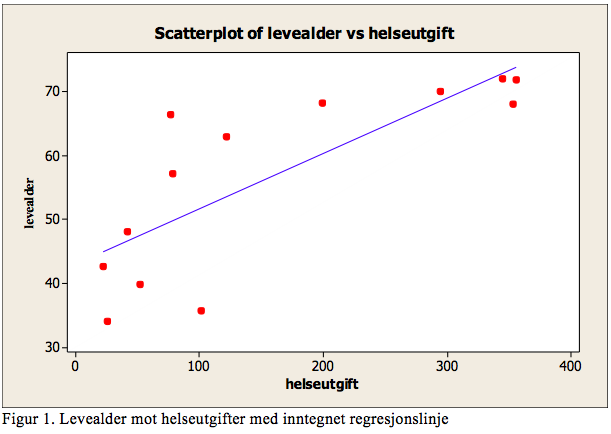
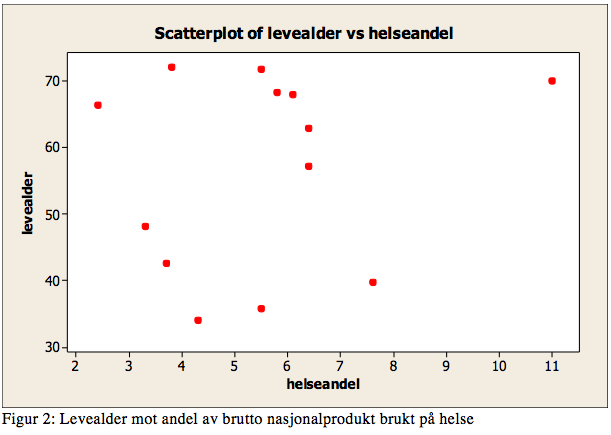
Er du enig i påstanden?

Dersom påstanden er riktig, hvorfor kan ikke en lineære modell brukes?

**d)** Vi prøvde også å bruke ”Andel av brutto nasjonalprodukt brukt på helse” som forklaringsvariabel i stedet for ”Helseutgift pr person”. Et plott av dette ser du i Figur 2.

Tror du dette er en brukbar forklaringsvariabel?

Hva tror du om størrelsen på R2 her?



**Oppgave 6:**

Vi ønsker å undersøke prisen på sykkeldekk til vinterbruk for terrengsykler. For 5 forskjellige typer dekk for terrengsykler har vi skaffet informasjon om pris i butikk og tallet på pigger. Resultatet ble:

Tallet på pigger:          296      160      72        106      0

Pris i kroner:               498      299      249      299      249

**a)**  Skriv opp en passende modell med tilhørende forutsetninger.

Tolk parametrene i modellen.

Finn deretter estimatene til parametrene.

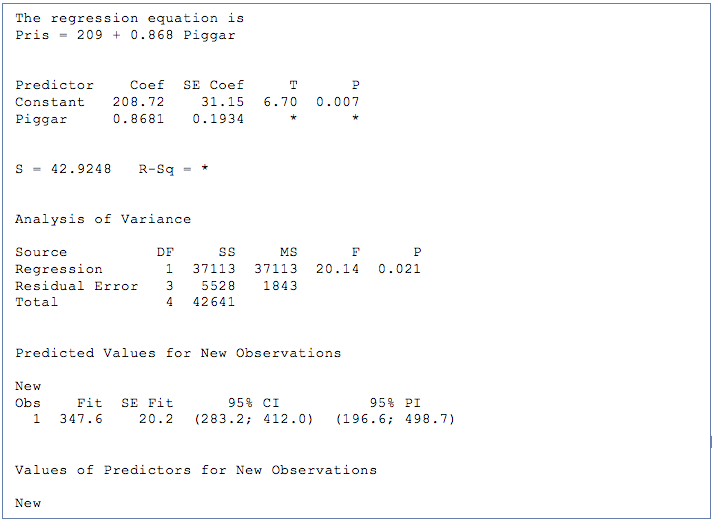
**b)**  **Vent med denne til uke 18**

Test på 5 % nivå om det er en positiv sammenheng mellom tallet på pigger og pris. Presiser nullhypotesen og alternativet.

Konkludèr slik at butikkeieren forstår det.

**c)** Hvor stor del av variasjonen i pris kan forklares av modellen?

**d)** Dersom pris var 348 kroner, hvor mange pigger tror du det var det i dette dekket?



**Oppgave 7(Bare a og b uke 17, resten uke 18):**

Det er viktig å så korn så tidleg som mogleg om våren, då kan ein auke avlinga om hausten. I eit forsøk registrerte vi såtid (målt i dagar frå 1. april) for ein spesiell byggsort og avling (målt i kg pr. mål) av denne om hausten.

http://canvastest.bibsys.no/assessment_questions/2372/files/2864/download?verifier=1Ae8nt0qBSsuVd378lzWJVdQuxHOSI6lo12j8BPI&wrap=1

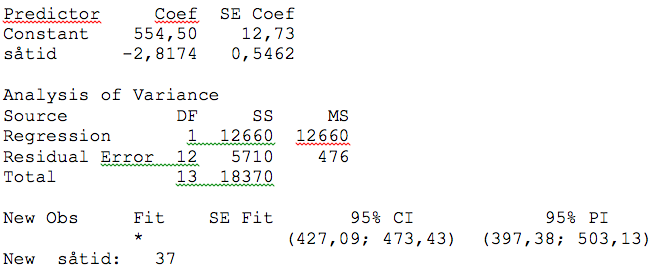
I slutten av denne oppgåva finn du eit spreiingsplot (scatterplot) og delar av ei datautskrift. Desse kan du bruke til å svare på fleir av dei fylgjande spørsmåla.

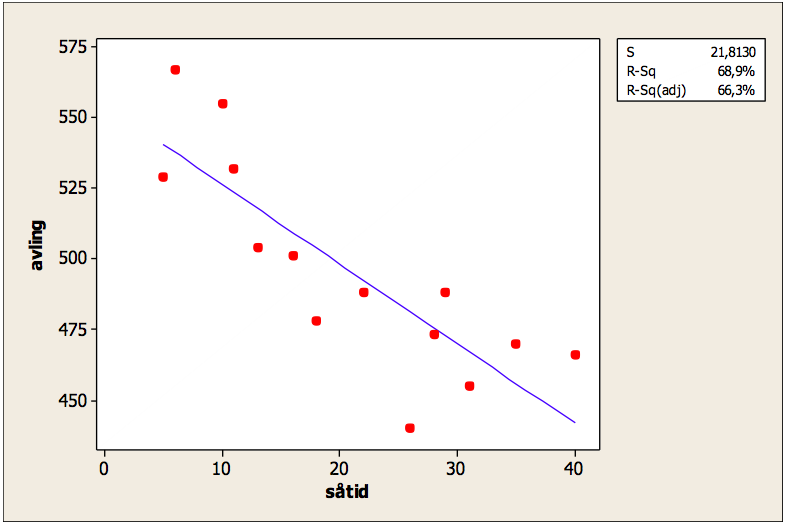
**a)** Sett opp ein passande statistisk modell for dette forsøket. Estimer alle parametrane i modellen.  Gi ei tolking av estimata.

**b)** Passar modellen til data? Kor mykje av variasjon i avling kan forklarast ved variasjon i såtid?

**c)**Sett opp ein statistisk hypotesetest for å avgjere om tidlig såtid verkar positivt på avling. Utfør testen og konkluder.

**d)** Eit år sådde en 6. mai. Kva vil du anslå avlinga til å bli dette året. Gje og eit 95 % prediksjonsintervall for dette anslaget. Gi en forklaring på kva intervallet betyr.





**Eksamensoppgaver i variansanalyse**

Et lite utplukk av eksamensoppgaver i variansanalyse. (Merk at samtlige eksamenssett har oppgaver i variansanalyse).

Høst 2015: Oppgave 1+ Flervalg 14-16

Vår 2016: F10-F17

Vår 2015: F6-F15

Vår 2012: Oppgave 1

Høst 2011: Oppgave 3