Stat100, vår 2018

**Løsning: Oppgaver til Uke 18**

**Oppgave 1 se også løsning uke 17**

**Residuals:**

**1 2 3**

**-0.5 1.0 -0.5**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) **1.000** 1.871 0.535 0.687

x **0.500** **0.866** 0.577 **0.667**

**s: 1.225** on 1 degrees of freedom (dette er sigmahatt)

Multiple R-squared**: 0.25**, (korrelasjon er kvadratrota er 0,5)

** = 1 + 0,5x**

x Y **Yhatt**  residual

1 1 **1.5**  -0.5

2 3 **2.0** 1.0

3 2 **2.5** -0.5

Response: y

Df Sum Sq

x 1 **0.5**

Residuals 1 **1.5**

**pre**dict\_CI\_PI(RegModel.10, data.frame('x'=1.5), level=0.95)

fit lwr.CI upr.CI lwr.PI upr.PI

1.75 -8.78 12.28 -17.04 20.54

#### Oppgave 2

**a)**Yi = \beta_0 + \beta_1xi + \epsilon_i, der \epsilon_i er normalfordelt (0, \sigma).

Der Yi er hardhet i tre nummer  i, xi er tetthet i tre nr *i, i* = 1, 2, . .  ., 10

\beta_0:  ingen fysisk tolkning

\beta_1: Forventet økning i hardhet dersom tetthet øker med 1 enhet

**\sigma:**Spredning (som standardavvik) hardhet for alle trær med samme tetthet.

Modellen passer godt til data.

**b)**

The regression equation is hardhet = -24,7 + 1,15 tetthet

Predictor        Coef     SE Coef       T

Constant      -24,7       6,847      -3,61

tetthet        1,15       0,1008      11,4

S = 4,446       R-Sq = 0.942

Analysis of Variance

Source                 DF          SS          MS         F

Regression           1      2561,3      2561,3    129,58

Residual Error     8        158.2        19,8

Total                9      2719,5

H0:  \beta_1= 0   H1:**\beta_1** ≠ 0

H0 Forkastes på 1 % nivå, fordi T = 11.4.  Dette er mye større enn t,0,005 med 8 frihetsgrader (3.355).

#### c. Bruk formel for 95 % PI som dersom ny verdi er identisk med gjennomsnittet av de gamle er gitt ved

#### ,der og er snittet av tetthetene i tabellen.

#### Oppgavene 3 -7: Se løsning uke 17

**Oppgave 8**

**a)** **Modell 1:**

La Yi være barnedødelighet i land i og xi være andelen med tilgang til reint vann i land i.

Vi antar Yi = \alpha + xi+ \epsilon_i, der \epsilon_i ~ N(0, \sigma) i = 1,2, . . .,7. Alle Yi-er er uavhengige.

\alpha er forventet barnedødelighet dersom ingen har tilgang til reint vann.

\beta er forventet økning i barnedødelighet dersom 1 % får økt tilgang til reint vann (vi får for all del håpe at denne er negativ)

\sigma er spredning (standardavvik) i barnedødelighet for alle land med samme tilgang til reint vann.

Modell 2: Som Modell 1, men xi er andelen kvinnelige analfabeter.

Estimert standardavvik i barnedødelighet for alle med samme andel analfabeter er 35,76. Hvis alle kan lese er analfabetisme 0 dermed er estimert forventet dødelighet 7.

**b)** For begge modeller betyr dette å teste:

H0:   = 0       ingen sammenheng

H1:  ≠ 0.       Sammenheng

**Modell 1**: H0 forkastes, p-verdien er 0,021. Vi påstår sammenheng mellom tilgang til reint vann og barnedødelighet.

**Modell 2:** H0 forkastes. T = 5,35. som er større enn alle mulige tabellverdier for t fordeling med 5 frihetsgrader. Vi påstår sammenheng mellom analfabetisme og barnedødelighet.

**c)** Modell 2: Det er samme SStot som i Modell 1, fordi SSTot ser kun på responsen.

R2 = SSreg/SStot 36606/42999  = 0,85

R2 blir størst når punktene ligger tett rundt linja. Da er s minst.

Da vil variasjon i x forklare mer av variasjon i Y.

**d)** Modell 1: Predikert verdi Vietnam: 291 – 2,78\*42 = 174

Modell2: Predikert verdi Vietnam: 7 + 2,32\*16 = 44

Merk at i Modell 2 må en finne estimert regresjonskoeffisient ved  = S.E.()\*T

Vietnam er å betrakte som et U-land ved vanntilgang, men som et I-land ved lesekyndighet.

Smalest for Modell 2 fordi s er minst.

Bredere for Norge fordi 99 % er langt fra gjennomsnittverdiene av vanntilgang som modellen bygger på.

**Oppgave 9:**

**a)**  For begge modeller har \alpha ingen fysisk tolkning da vi ikke kan tenke oss bjørner uten brystomkrets eller nakkeomkrets.

For Modell 1 tolker vi \beta som forventet (gjennomsnittlig for alle bjørner i populasjonen) økning i vekt dersom brystomkrets øker med 1 cm, og \sigma som standardavvik i vekt for alle bjørner med samme brystomkrets.

For Modell 2 tolker vi \beta som forventet (gjennomsnittlig for alle bjørner i populasjonen) økning i vekt dersom nakkeomkrets øker med 1 cm, og \sigma som standardavvik i vekt for alle bjørner med samme nakkeomkrets.

Modell 1: \hat{\alpha}= -133,52, \hat{\beta}= 2,73,

finner du ved å ta kvadratroten av (SSresidual/22)

Modell 2: \hat{\alpha}= -118,23, \hat{\beta}= 4,25.

Siden det er samme respons er SSTotal den samme for begge modeller: SSTotal = 113940, dermed blir SSresidual : 113940 – 96678 = 17262.

* finner du ved å ta kvadratroten av (SSresidual/22).

**b**)

Modell 1: SSTotal = 113940. R2 = SSREG/SSTotal = 0,91.

Modell 2: Samme SSTotal som i Modell 1. R2 = SSREG/SSTotal = 0,85

R2 måler andel variasjon i vekt som er forklart av henholdsvis variasjon i brystomkrets og variasjon i nakkeomkrets.

Et 95 % KI for b er gitt ved: \hat{\beta} ± t0,025,22S.E.(\hat{\beta} ). Dette gir:

Modell 1: 2,73 ± 2,074\*0,18 eller (2,36,   3,10)

Alternativt: Modell 2: 4,25 ± 2,074\*0,38 eller (3,46,   5,04)

Siden ingen av intervallene dekker 0, kan vi forkaste hypotesen H0: \beta = 0, og tro på alternativet H1: \beta ≠ 0 på 5 % signifikansnivå.

Med andre ord vi har gjennom KI påvist sammenheng mellom vekt og brystomkrets og alternativt mellom vekt og nakkeomkrets.

**c)**

Merk at for begge modeller har vi bjørner med gjennomsnittlig x verdi.

Intervallet blir bredere fordi s (****) er større.

Alt annet som bestemmer bredden er likt. s finner du ved kvadratrota av

(SSResidual /(n-2)).

SSresidual for nakke finner du ved SSResidaul = SSTotal – SSRegression.

SSTotal finner du fra modellen for brystomfang. (SSTotal = SSReggression + SSNakke)

Husk at SSTotal er uavhengig av modellvalg.

**d)**

Siden feilleddene er antatt å kun være støy uten struktur, vil vi heller ikke vente å finne struktur i residualene.

Vi kan avsløre ikke-linearitet, ikke-konstant varians hvis vi plotter residualer mot tilpassede verdier (eller mot x).

Et histogram over residualer kan si oss noe om hvorvidt normafordelingsantagelsen er rett.

Svakheter ved modellen:

Ser klar tendens til en bue, dermed har vi problemer med ikke-lineære data.

I tillegg er det slik at hannbjørner får positive residualer, mens hunnbjørner får negative residualer.

Vi predikere hanner for lette og hunner for tunge. Dette indikerer at vi burde bruke en modell for hvert kjønn.