Norges teknisk–naturvitenskapelige universitet
Institutt for matematiske fag

Faglig kontakt under eksamen:
Mette Langaas (988 47 649)

BOKMÅL

EKSAMEN I TMA4255 ANVENDT STATISTIKK

Fredag 7. juni 2013
Tid: 9:00–13:00

Antall studiepoeng: 7.5.

Tillatte hjelpemidler: Alle trykte og håndskrevne hjelpemidler. Spesiell kalkulator.

Sensurfrist: 28. juni 2013.

Eksamensresultatene annonseres fra `http://studweb.ntnu.no/`.

Merk deg følgende:

- I utskrift fra MINITAB er komma brukt som desimalseparator.
- Signifikansnivå 5% skal brukes hvis ikke annet er spesifisert.
- Alle svar må begrunnes.

Oppgave 1 Kroppsfettprosent og fotbehandling

Kroppsfettprosenten til en person kan måles ved å sende et svakt elektrisk signal gjennom kroppen for å måle kroppens impedans. Signalet ledes gjennom vannet i kroppen. Muskelmasse inneholder mer vann enn fettvev, og dette kan brukes til å beregne et mål på kroppsfettprosent.

En masteroppgave ved Obesitasklinikken ved St. Olavs Hospital hadde som mål å undersøke om pedikyrbehandling ville påvirke målingen av kroppsfettprosent (som beskrevet over).

Totalt deltok 40 personer i studien, alle med kroppsmasseindeks (vekt delt på kvadratet av høyde) over 30 kg/m^2. Alle deltakerne ble målt med impedansteknologien to ganger, før og etter en pedikyrbehandling. For person i, la X_{1i} angi kroppsfettprosenten målt før pedikyrbehandlingen og X_{2i} angi kroppsfettprosenten målt etter pedikyrbehandlingen, $i = 1, \ldots, 40$.

Spredningsplott og normalplott av dataene er presentert i figur 1.

I figur 2 finner du utskrifter fra følgende fire statistiske analyser:

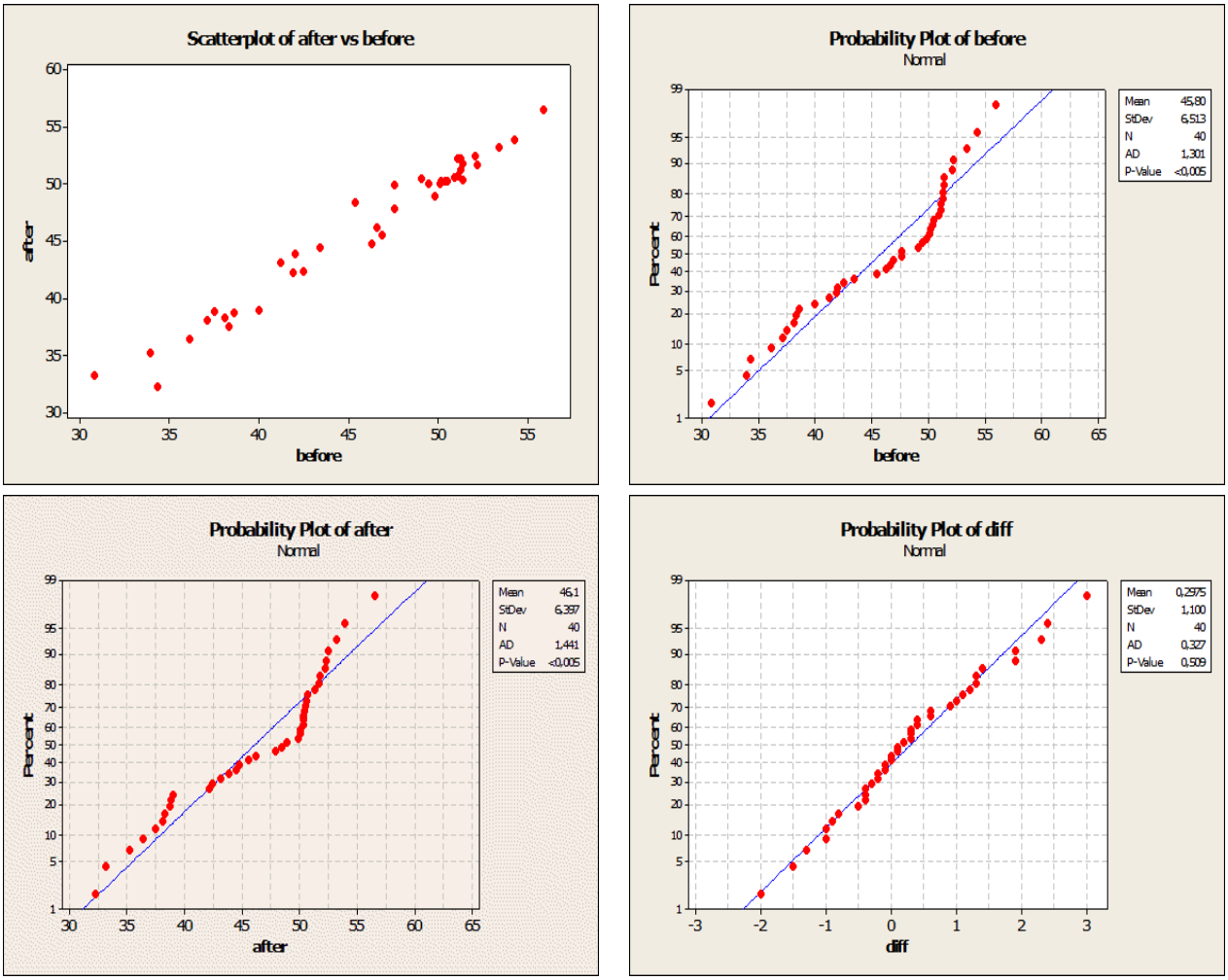
A en ett-utvalgs t-test basert på differansene $X_{2i} - X_{1i}$,

B en to-utvalgs t-test (uten å anta lik varians) basert på målingene før og etter pedikyren,

C en Wilcoxon signed-rank test basert på differansene $X_{2i} - X_{1i}$, og

D en Wilcoxon rank-sum test (Mann–Whitney) basert på målingene før og etter pedikyren.

a) Er de to utvalgene (før og etter pedikyr) uavhengige?
Vil du ut fra normalplottene i figur 1, si at målingene før pedikyren, etter pedikyren eller differansen $(X_{2i} - X_{1i})$ kan sees på som utvalg fra en normafordelt populasjon?
Basert på svarene dine over, hvilken av de fire analysene A–D synes du passer best til forskningsspørsmålet og dataene?
For den valgte analysen, skriv ned null og alternativ hypotese, list opp antagelsene du må gjøre for å gjennomføre analysen og rapporter resultatet fra hypotesetesten.



Figur 1: Spredningsplott og normalplott av kroppsfettprosent-dataene. Målingene kalt `diff` er $X_{2i} - X_{1i}$.

```
A: One-Sample T: diff
--------------------
Test of mu = 0 vs not = 0

Variable   N   Mean  StDev  SE Mean       95% CI          T      P
diff         40  0,298  1,100    0,174  (-0,054; 0,649)  1,71  0,095

B: Two-sample T for after vs before
---------------------------------
             N   Mean  StDev  SE Mean
after     40  46,10   6,40      1,0
before  40  45,80   6,51      1,0

Difference = mu (after) - mu (before)
Estimate for difference:  0,30
95% CI for difference:  (-2,58; 3,17)
T-Test of difference = 0 (vs not =):
T-Value = 0,21  P-Value = 0,837  DF = 77

C: Wilcoxon Signed Rank Test: diff
-----------------------------
Test of median = 0,00 versus median not = 0,00

        N     N for Wilcoxon   Test Statistic      P         Median
diff  40     38                      470,0          0,151     0,2500

D: Mann-Whitney Test and CI: after; before
-----------------------------------------
            N    Median
after     40   48,650
before  40   47,600

Point estimate for ETA1-ETA2 is 0,200
95,1 Percent CI for ETA1-ETA2 is (-1,998;2,699)
W = 1644,0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0,8211
The test is significant at 0,8211 (adjusted for ties)
```

Figur 2: Utskrift fra statistiske analyser av kroppsfettprosentdataene.

Oppgave 2 Prosesskontroll med resistorer

Ved en produksjonsenhet blir det tatt et utvalg på tre resistorer hver time, og resistansen, i ohm, blir målt. La X_{ij} være målingen av resistans for resistor j, og utvalg i, hvor $j = 1, 2, 3$ og $i = 1, 2, \ldots, k$. Videre, $\bar{X}_i = \frac{1}{3}\sum_{j=1}^{3} X_{ij}$, $S_i = \sqrt{\frac{1}{2}\sum_{j=1}^{3}(X_{ij} - \bar{X}_i)^2}$, $\bar{\bar{X}} = \frac{1}{k}\sum_{i=1}^{k} \bar{X}_i$, og $\bar{S} = \frac{1}{k}\sum_{i=1}^{k} S_i$.

Basert på $k = 30$ utvalg, som vi antar er i kontroll, finner vi $\bar{\bar{x}} = 5.095$ og $\bar{s} = 0.058$.

a) Konstruer et S-chart og et $\bar{X} - S$-chart (med 3σ grenser).
For et nytt utvalg finner vi at $\bar{x} = 5.15$ og $s = 0.10$. Er prosessen i kontroll for dette nye utvalget?

Hva er fordelen med å bruke et cusum-chart istedenfor et Shewhart-chart (som konstruert her)?

Oppgave 3 Treff fra flyvende bomber

Vi skal studere et klassisk datasett over treff fra flygende bomber i den sørlige delen av London under andre verdenskrig.

Byen ble delt inn i små områder, hvert område på en kvart kvadratkilometer. La oss tenke oss at vi velger ut ett slikt område tilfeldig, og la X være antallet bomber som treffer dette området. Basert på statistisk teori kan det være naturlig å undersøke om X er en possion-fordelt tilfeldig variabel. I vår situasjon betyr det at sannsynligheten for at det valgte området blir truffet nøyaktig k ganger, $P(X = k)$, er gitt ved

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, \ldots$$

for en gitt intensitet $\lambda > 0$.

Vi vil se på observasjoner av antallet bombetreff for hver av $n = 576$ små områder i London, og vi antar at intensiteten, λ, er den samme for alle områdene. Tabell 1 viser antallet områder der det er observert 0, 1, 2, 3, 4 eller flere treff. Totalt traff 537 bomber, slik at gjennomsnittlig antall bombetreff per lite område var $\hat{\lambda} = 537/576 = 0.93$.

Treff (k)	0	1	2	3	4 eller flere	Totalt
Observert antall små områder med k treff	229	211	93	35	8	576

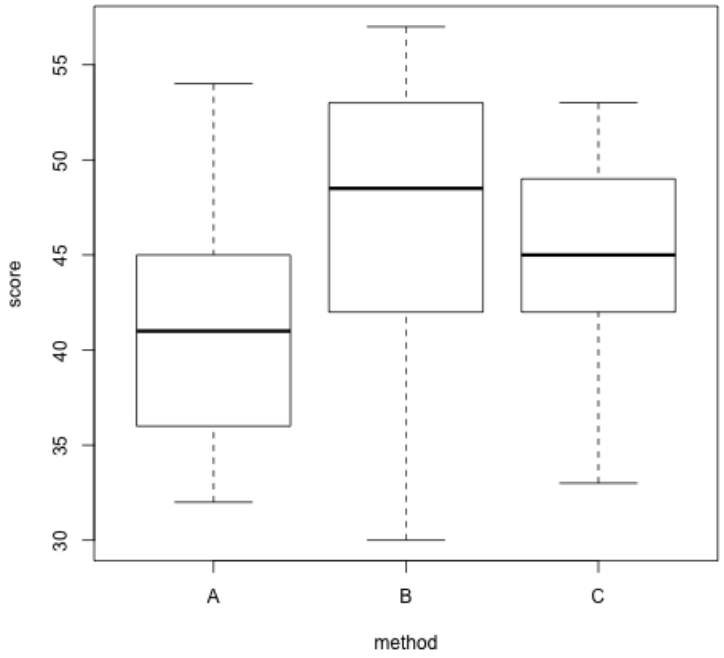
Tabell 1: Observert antall små områder med k bombetreff.

a) Utfør en hypotesetest for å svare på følgende spørsmål: Kan observasjonene i tabell 1 sees som et tilfeldig utvalg fra poissonfordelingen? Du kan bruke estimatet $\hat{\lambda} = 0.93$ i beregningene dine.
Hva er din konklusjon fra hypotesetesten?

Oppgave 4 Å lære å lese

I en randomisert studie var målet å sammenligne tre undervisingsmetoder for å lære å lese. En av metodene var den som ble brukt i undervisningen nå (A), og så var det to nye metoder (B og C). Totalt var 66 elever med i studien, og de ble tilfeldig tilordnet til en av de tre metodene, 22 elever for hver metode.

Vi vil se på data der lesescore er målt. Lesescore er en numerisk verdi, der en høy lesescore er foretrukket. Boksplott av data er presentert i figure 3, og deskriptive mål er gitt i tabell 2.



Figur 3: Boksplott for data for lesescore

Metode	Utvalgsstørrelse	Gjennomsnitt	Standardavvik
A	22	41.05	5.636
B	22	46.73	7.388
C	22	44.27	5.767
Total	66	44.02	

Tabell 2: Deskriptive mål for lesescorer.

a) Vi vil undersøke om forventet lesescore varierer mellom undervisningsmetodene. Skriv ned null- og alternativ hypotese og utfør én enkelt hypotesetest basert på de deskriptive målene gitt i tabell 2.
Hvilke antagelser må du gjøre for å bruke denne testen?
Hva blir konklusjonen fra testen?

Vi skal nå sammenligne de to nye undervisningsmetodene, metode B og C.

b) La μ_B og μ_C være forventet lesescore for læringsmetode B og C. Vi vil se på forholdet, γ, mellom de to forventede lesescorene,

$$\gamma = \frac{\mu_B}{\mu_C}.$$

Foreslå en estimator, $\hat{\gamma}$, for γ.

Bruk Taylor-metoder til å approksimere forventingsverdien og standardavviket til denne estimatoren, det vil si $\mathrm{E}(\hat{\gamma})$ og $\mathrm{SD}(\hat{\gamma}) = \sqrt{\mathrm{Var}(\hat{\gamma})}$.
Bruk relevante data fra tabell 2 til å beregne $\hat{\gamma}$ og de estimerte approksimerte verdiene for $\mathrm{E}(\hat{\gamma})$ og $\mathrm{SD}(\hat{\gamma})$ numerisk.

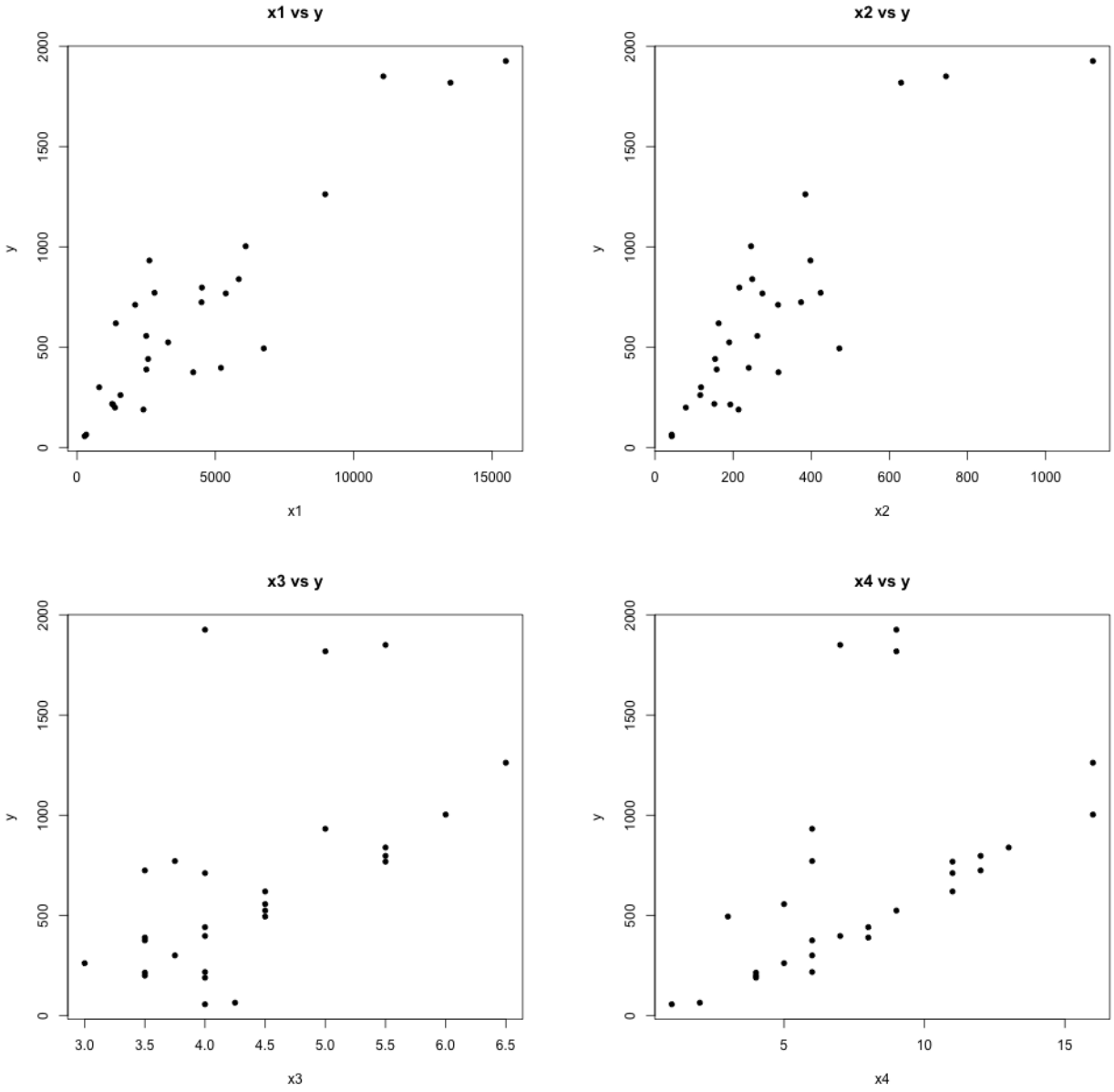
Oppgave 5 Betong

I en artikkel i *Journal of Computing in Civil Engineering* var målet å lage en prediksjonsmodell for mengden betong, y, som skulle brukes i konstruksjon av et silokompleks. Prediksjonsmodellen skulle brukes i designfasen av konstruksjonsjobben. Totalt fantes det 23 mulige forklaringsvariabler, og vi vil se på fire av disse. Følgende beskrivelse er gitt.

- y, `concrete`. Mengen av betong målt i m^3.
- x_1, `volume`. Volumet av silokomplekset.
- x_2, `perimeter`. Omkretsen av silokomplekset.

- x_3, `waste`. Prosentandel betong som ble kastet.
- x_4, `steel`. Antallet arbeidslag som jobbet med stålarmering.

Vi har data tilgjengelig fra 28 konstruksjonsjobber. Spredningsplott finnes i figur 4.



Figur 4: Spredningsplott for betongdataene.

```
Predictor      Coef  SE Coef      T      P
Constant    -574,2    190,1  -3,02  0,006
x1         0,02670  0,02142   1,25  0,225
x2          1,3612   0,3241   4,20  0,0003
x3          124,08    49,39   2,51  ?
x4           23,22    10,28   2,26  0,034

S = 158,231   R-Sq = ? %   R-Sq(adj) =  90,56 %

Analysis of Variance

Source          DF       SS       MS      F      P
Regression       4  6586535  1646634  65,77  0,000
Residual Error  23   575852    25037
Total           27  7162388
```

Figur 5: Utskrift fra statistiske analyser for modell A for betongdataene.

En multippel lineær regresjonsmodell ble tilpasset til dataene med y som respons og x_1, x_2, x_3 og x_4 som forklaringsvariabler. La $(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i})$ angi en observasjon fra jobb i, hvor $i = 1, \ldots, 28$. Definer den fulle modellen (modell A):

$$\text{Modell A: } y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$$

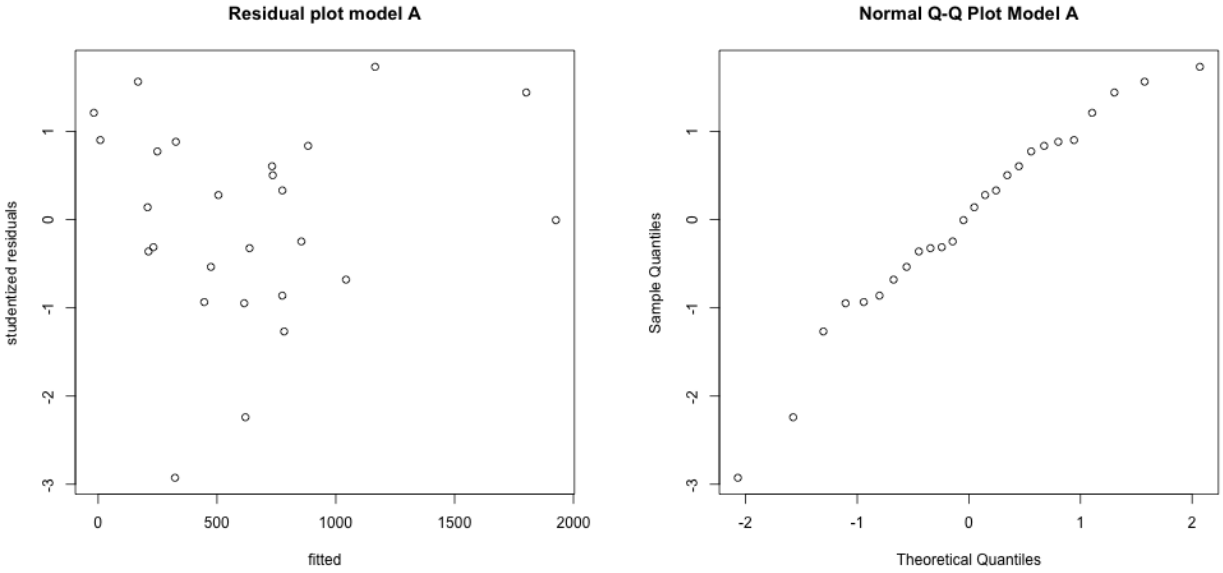
hvor ε_i er u.i.f. $N(0, \sigma^2)$ for $i = 1, \ldots, 28$. Utskrift fra statistisk analyse finnes i figur 5 og plott av studentiserte residualer finnes i figur 6. To av de numeriske verdiene i utskriften er erstattet med spørsmålstegn.

a) Skriv ned den estimerte regresjonsligningen.

Nå skal vi se på den estimerte regresjonskoeffisienten for x_3, `waste`, i denne modellen. Hvordan vil du forklare denne verdien til «mannen i gata» (som ikke kjenner til lineær regresjon)?
Er effekten av x_3, `waste`, signifikant i denne modellen?

Beregn R^2 og forklar hvordan du kan tolke denne verdien.

Tror du at modell A er en god modell for betongdataene? For å svare må du kommentere utskriften fra den statistiske analysen og residualplottene i figur 6.



Figur 6: Residualplott (studentiserte residualer mot tilpassede verdier til venstre, normalplott av studentiserte residualer til høyre) for modell A for betongdataene.

Vi vil nå sammenligne den fulle regresjonsmodellen (modell A) med en redusert model (kalt modell B), som bare inneholder x_2 (`perimeter`) og x_3 (`waste`).

$$\text{Modell B: } y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Resultater fra å tilpasse modell B til betongdataene finnes i figur 7.

b) Kommenter hva som er de viktigste forskjellene mellom modell A og modell B.

Modell A og modell B kan sammenlignes ved å undersøke følgende hypoteser.

$$H_0\colon \beta_1 = \beta_4 = 0 \text{ vs. } H_1\colon \beta_1 \text{ og } \beta_4 \text{ er ikke begge lik null}$$

Utfør hypotesetesten og konkluder.

Vil du foretrekke å bruke modell A eller modell B?

```
Predictor         Coef  SE Coef         T            P
Constant    -815,6       174,1    -4,68       8,45e-05
x2            1,7575     0,1521    11,55      1,62e-11
x3            219,65      40,09      5,48      1,09e-05

S = 176,717   R-Sq = 89,1%   R-Sq(adj) = 88,2%

Analysis of Variance
Source           DF       SS       MS       F      P
Regression        2  6381667  3190833  102,18  0,000
Residual Error   25   780721    31229
Total            27  7162388
```

Figur 7: Utskrift fra statistisk analyse for modell B for betongdataene.

Vi skal nå bruke modell B.

c) Anta at en ny konstruksjonsjobb er under planlegging og at vi forventer at $x_2 = 300$ og $x_3 = 4.4$ for denne nye jobben. Hva er den beste prediksjonen for mengden betong som skal brukes?

La X angi designmatrisen til modell B, og la $\boldsymbol{x}_0$ være vektoren med forklaringsvariabler som vi ønsker å bruke i prediksjonen, $\boldsymbol{x}_0^T = [1\ 300\ 4.4]$. Det oppgis at da er $\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0 = 0.0357$. Bruk denne informasjonen til å konstruere et 95% prediksjonsintervall for mengden betong som som vil bli brukt når $x_2 = 300$ og $x_3 = 4.4$.
Hva er fortolkningen av dette intervallet?