Mat-2.104 Tilastollisen analyysin perusteet

Mellin/Tentti 8.5.2005

Kirjoita *selvästi jokaiseen koepaperiin* alla mainitussa järjestyksessä: - **Mat-2.104 Tap/Tentti 8.5.2004** - **opiskelijanumero + kirjain, TEKSTATEN sukunimi, kaikki etunimet** - **koulutusohjelma, vuosikurssi** - **mahdolliset entiset nimet ja koulutusohjelmat** - **nimikirjoitus**

OHJEITA

(i) **Tehtäviä on 5 kpl.**

(ii) **Tehtävien tulostukset on tuotettu STATISTIX-ohjelmalla.**

(iii) **Yhden tehtävistä *saa korvata* harjoitustyöllä.**

Korvattava tehtävä on ilmaistava vastauspaperissa selvästi kokonaislukuna.

(iv) **Vastaa lyhyesti ja ytimekkäästi, mutta esitä *perustelut.***

1. Ihmiskehon rasvaprosentin määräämiseen käytetään yleisesti kahta menetelmää, A ja B. Menetelmä A perustuu kehon ominaispainon mittaamiseen ja on tarkka, mutta sitä on työläs soveltaa. Menetelmä B perustuu kehon sähköisen ominaisvastuksen mittaamiseen ja se ei ole yhtä tarkka kuin menetelmä A, mutta sen soveltaminen on helppoa.

Menetelmien vertaamiseksi rasvaprosentit mitattiin molemmilla menetelmillä 10 koehenkilöltä. Tavoitteena oli selvittää antavatko menetelmät keskimäärin samat tulokset. paired,

Tutkijat X ja Y sovelsivat kerättyyn aineistoon erilaisia testausmenetelmiä. Tulokset tutkijoiden X ja Y tekemistä testeistä on annettu seuraavalla sivulla.

Tehtävät:

(a) Mitä testejä tutkijat X ja Y sovelsivat?
Kuvaa testejä ja niiden käyttöä lyhyesti.

(b) Toinen tutkijoista X ja Y sovelsi testiä, jota *ei saa* soveltaa tehtävän testausasetelmassa. Kumpi? Perustele valintasi.

(c) Tee johtopäätökset siitä testistä, jonka valitsit (b)-kohdassa.

Tutkija X:

```
STATISTIX FOR WINDOWS                                    RASVAPROS

PAIRED T TEST FOR A - B

NULL HYPOTHESIS: DIFFERENCE =  0
ALTERNATIVE HYP: DIFFERENCE <> 0

MEAN         -1.5000
STD ERROR     0.5217
LO 95% CI    -2.6803
UP 95% CI    -0.3197
T              -2.87
DF                 9
P             0.0183

CASES INCLUDED 10    MISSING CASES 0
```

Tutkija Y:

```
STATISTIX FOR WINDOWS                                    RASVAPROS

TWO-SAMPLE T TESTS FOR A VS B

                          SAMPLE
VARIABLE        MEAN      SIZE        S.D.         S.E.
---------    ---------    ------    ---------    ---------
A              17.400       10        6.9634       2.2020
B              18.900       10        6.2619       1.9802
DIFFERENCE    -1.5000

NULL HYPOTHESIS: DIFFERENCE =  0
ALTERNATIVE HYP: DIFFERENCE <> 0

ASSUMPTION              T        DF         P      95% CI FOR
DIFFERENCE
-----------------    ------    ------    ------    --------------------
EQUAL VARIANCES       -0.51      18      0.6186      (-7.7217, 4.7217)
UNEQUAL VARIANCES     -0.51      17.8    0.6187      (-7.7267, 4.7267)

                         F       NUM DF    DEN DF       P
TESTS FOR EQUALITY    -------    ------    ------    ------
      OF VARIANCES      1.24        9         9      0.3784

CASES INCLUDED 20    MISSING CASES 0
```

2. Erästä vakavaa tautia vastaan on kehitetty rokote. Rokotuksen tehon selvittämiseksi järjestettiin rokotuskoe. Kokeen kohteiksi valitut henkilöt jaettiin satunnaisesti kahteen ryhmään:

Ryhmä 1 (CASE = 1): Rokotetut

Ryhmä 2 (CASE = 2): Ei-rokotetut

Kokeessa rekisteröitiin rokotusta seuranneen vuoden aikana sairastuneiden ja ei-sairastuneiden lukumäärät.

Kokeen tulokset on annettu alla olevassa 2×2-frekvenssitaulukossa.

	VARIABLE	
CASE	SAIRASTUI	TERVE
1	8	42
2	16	34

Kokeen tekijät halusivat tutkia tilastollisesti ovatko rokotus ja sairastuminen riippumattomia tekijöitä. Tulokset tehdystä tilastollisesta analyysista on annettu seuraavalla sivulla.

Huomautus:

Painovirhepaholainen halusi estää vastaamisen ja korvasi osan tulostuksen luvuista kysymysmerkeillä.

Paholainen ei kuitenkaan tiennyt, että puuttuvat luvut voidaan laskea jäljelle jääneistä luvuista.

Puuttuvat luvut ovat *havaintojen kokonaislukumäärä*, solun (CASE = 1, SAIRASTUI) *odotettu frekvenssi*, *solun* (CASE = 2, SAIRASTUI) χ^2-*arvo*, frekvenssitaulua vastaava χ^2-*testisuureen arvo* ja *vapausasteiden lukumäärä*.

Tehtävät:

(a) Mitä testiä sovellettiin?

Kuvaa testiä ja sen käyttöä lyhyesti.

(b) Laske puuttuvat luvut.

(c) Tee johtopäätökset tilastollisen analyysin tuloksista.

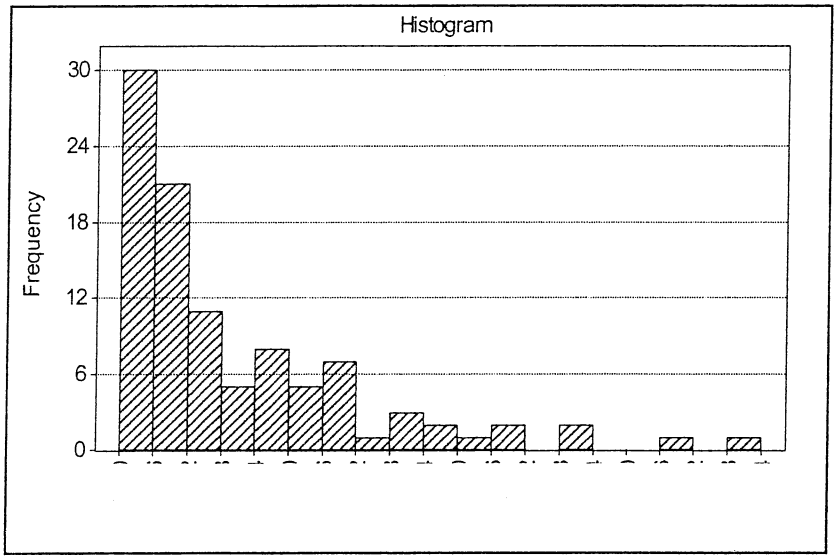
Suosittelisitko rokotusta analyysituloksen perusteella?

Pohdi asiaa siinä valossa, että ko. tauti on vakava.

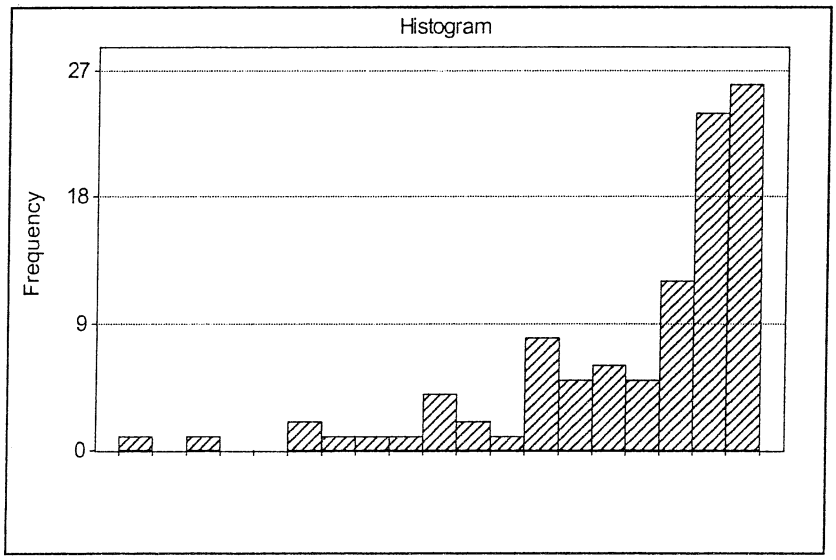
```
STATISTIX FOR WINDOWS

CHI-SQUARE TEST FOR HETEROGENEITY OR INDEPENDENCE

                          VARIABLE
CASE                 SAIRASTUI      TERVE
                    +-----------+-----------+
1       OBSERVED    |      8    |     42    |      50
        EXPECTED    |    ?????  |   38.00   |
     CELL CHI-SQ    |     1.33  |    0.42   |
                    +-----------+-----------+
2       OBSERVED    |     16    |     34    |      50
        EXPECTED    |   12.00   |   38.00   |
     CELL CHI-SQ    |    ????   |    0.42   |
                    +-----------+-----------+
                         24          76          ???

OVERALL CHI-SQUARE      ????
P-VALUE                0.0610
DEGREES OF FREEDOM          ?

CASES INCLUDED 4     MISSING CASES 0
```

3.1. Alla on esitetty histogrammit kahdesta aineistosta:

Kuvio 1



Kuvio 2



(a) Kumpi kuvioista on vino *vasemmalle*, kumpi *oikealle*?

(b) Aineistoista laskettiin myös niiden aritmeettiset keskiarvot, mediaanit ja vinoudet. Tulokset on annettu alla olevassa taulukossa.

STATISTIX FOR WINDOWS

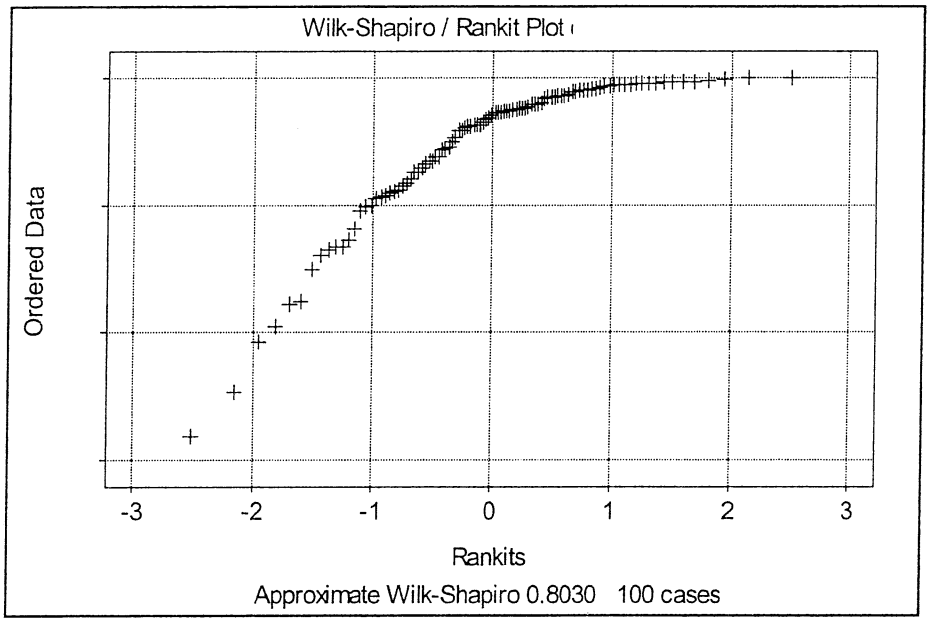
DESCRIPTIVE STATISTICS

VARIABLE	MEAN	MEDIAN	SKEW
A	2.0878	1.1429	1.7359
B	7.9122	8.8571	-1.7359

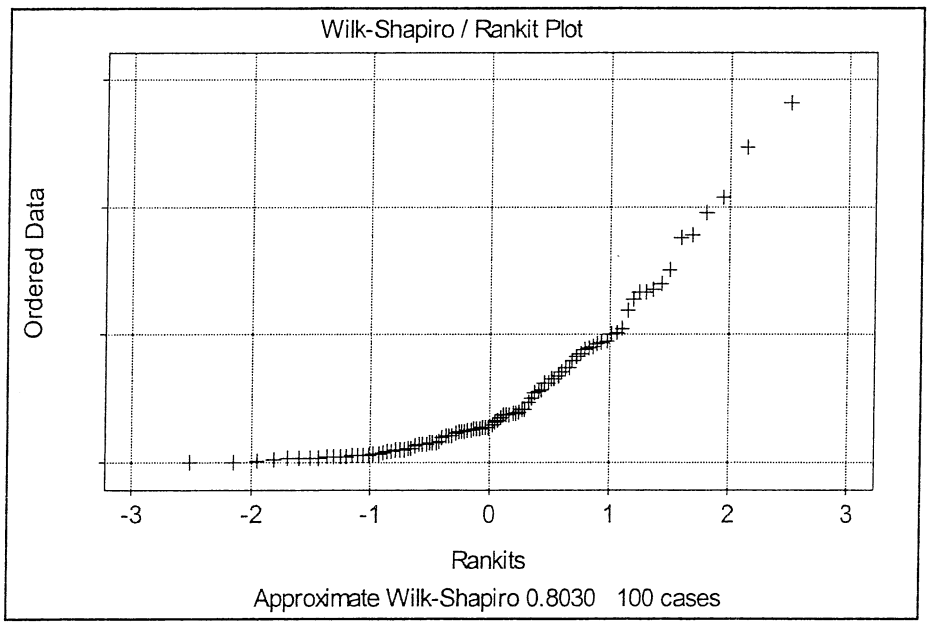
Päättele tulostuksista kumpi muuttujista A ja B liittyy Kuvion 1 histogrammiin ja kumpi Kuvion 2 histogrammiin.

(c) Aineistoista muodostettiin myös ns. Rankit Plot –kuviot:

Kuvio 3



Kuvio 4



Kumpi kuvioista liittyy Kuvion 1 histogrammiin ja kumpi Kuvion 2 histogrammiin?

3.2. Seuraavalla sivulla on esitetty kolme kahden muuttujan pistediagrammia, joihin liittyvät otoskorrelaatiokertoimet ovat *umpimähkäisessä järjestyksessä*

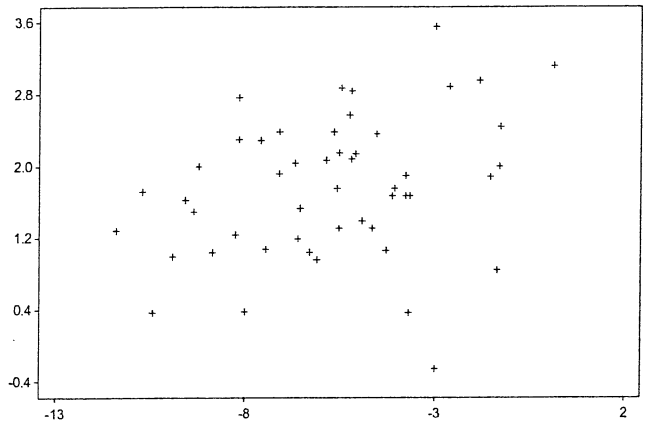
–0.79, 0.95, 0.29

Oletetaan, että kutakin kuviota vastaavasta aineistosta estimoidaan tavallinen yhden selittäjän lineaarinen regressiomalli, jossa pystyakselin muuttujaa selitetään vaaka-akselin muuttujalla ja vakioselittäjällä.

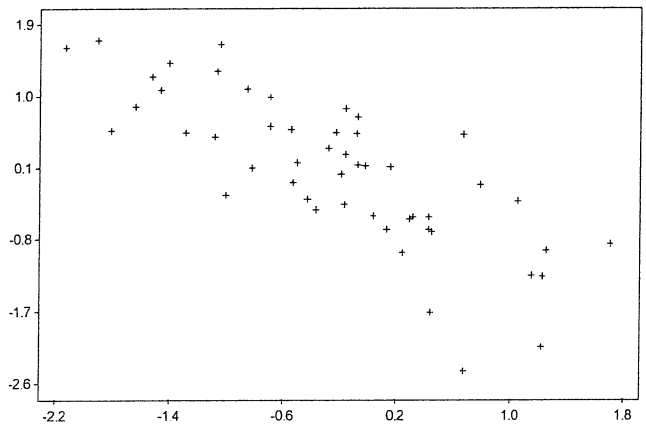
Tehtävä:

Määrää kuvioiden 5 – 7 aineistoihin liittyvien regressiomallien selitysasteet.

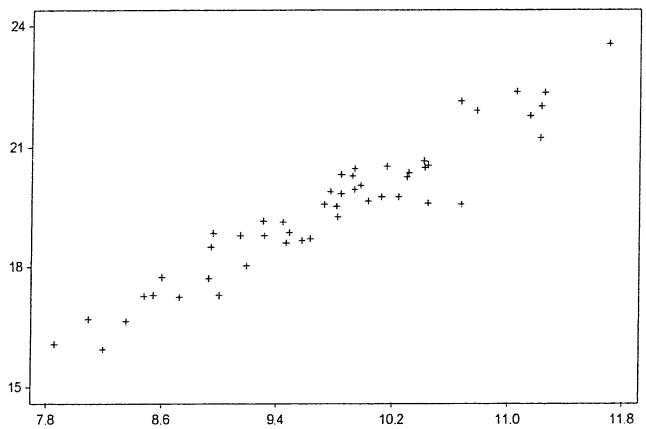
Kuvio 5



Kuvio 6



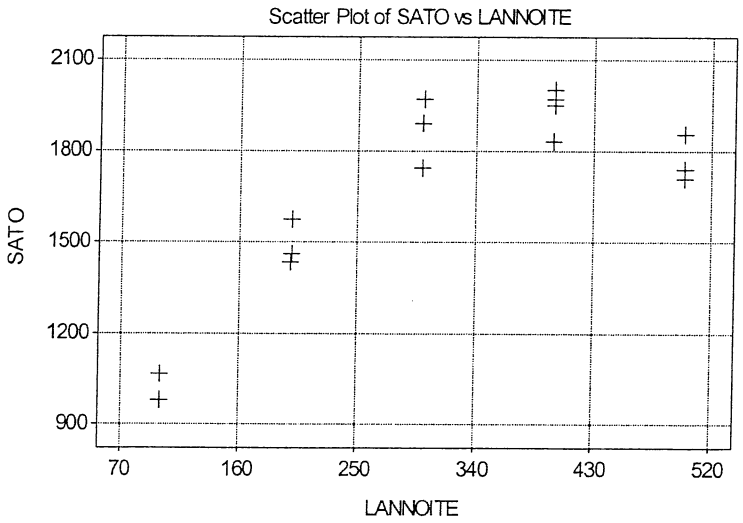
Kuvio 7



4. Eräässä rukiinviljelykokeessa tutkittiin käytetyn lannoiteaineen määrän (LANNOITE; kg/ha) vaikutusta vehnän satoon (SATO; kg/ha). Kokeessa oli mukana 15 samanlaista peltolohkoa, joille käytetyt lannoiteaineen määrät arvottiin. Lohkoihin kohdistettiin kaikissa muissa suhteissa samanlaiset käsittelyt.

Koetulokset ja sadon riippuvuutta lannoiteaineen määrästä kuvaava pistediagrammi on annettu alla.

CASE	LANNOITE	SATO
1	100	980.64580
2	100	1061.6804
3	200	1573.1315
4	200	1462.7448
5	200	1435.8953
6	300	1740.4902
7	300	1969.8052
8	300	1885.9458
9	400	1947.5419
10	400	1829.8867
11	400	1971.5341
12	400	2002.9903
13	500	1851.4186
14	500	1740.7020
15	500	1706.3952



Sadon riippuvuutta käytetyn lannoiteaineen määrästä tutkittiin kahdella lineaarisella regressiomallilla.

Mallissa 1 muuttujan SATO selittäjänä käytettiin (vakioselittäjän lisäksi) lannoiteaineen määrää (LANNOITE).

Mallissa 2 muuttujan SATO selittäjinä käytettiin (vakioselittäjän lisäksi) lannoiteaineen määrää (LANNOITE) ja lannoiteaineen määrän neliötä (LANNOITE2).

Seuraavilla sivuilla on annettu estimointitulokset molemmista regressiomalleista sekä estimoituja malleja vastaavat residuaalidiagrammit.

Huomautus:

Painovirhepaholainen halusi estää vastaamisen ja korvasi osan malliin 2 liittyvän tulostuksen luvuista kysymysmerkeillä.

Paholainen ei kuitenkaan tiennyt, että puuttuvat luvut voidaan laskea jäljelle jääneistä luvuista.

Puuttuvat luvut ovat estimoidun mallin muuttujaa LANNOITE vastaava *t-testisuureen arvo*, *selitysaste*, *mallineliösumma* ja sitä *vastaava keski-neliövirhe* sekä *F-testisuureen arvo*.

Tehtävät:

(a) Selitä mallien 1 ja 2 roolit aineiston analysoinnissa.

(b) Laske puuttuvat luvut mallin 2 tulostuksessa.

(c) Poimi tulostuksista varianssianalyysihajotelmat ja niille tulkinnat sekä vertaile mallin 1 ja 2 hajotelmia. Selitä mistä johtuvat mahdolliset erot ja yhtäläisyydet.

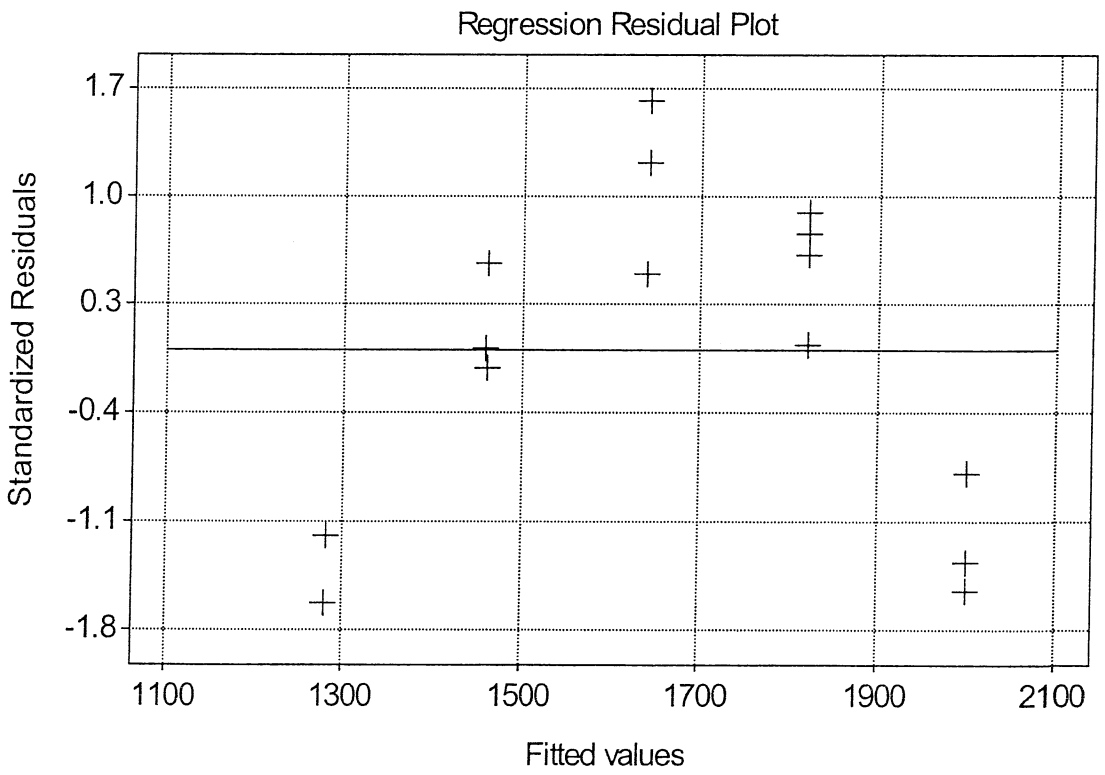
(d) Mitä tarkoittavat tulostuksessa 2 esiintyvät VIF-luvut?
Mitä johtopäätöksiä niistä voidaan tehdä?

(e) Tee johtopäätökset tilastollisen analyysin tuloksista.

(f) Kumpi malleista 1 ja 2 on parempi? Miksi?

Malli 1

```
STATISTIX FOR WINDOWS

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF SATO

PREDICTOR
VARIABLES     COEFFICIENT     STD ERROR      STUDENT'S T        P        VIF
---------     -----------     ---------      -----------     ------     -----
CONSTANT          223.807       114.408             1.96     0.0741
LANNOITE          8.84699       0.82298            ?????     0.0000      27.2
LANNOITE2        -0.01148       0.00132            -8.72     0.0000      27.2

R-SQUARED            ??????     RESID. MEAN SQUARE (MSE)     6568.26
ADJUSTED R-SQUARED   0.9360     STANDARD DEVIATION           81.0448

SOURCE          DF        SS           MS          F         P
----------     ---    ----------   ----------   -----     ------
REGRESSION       2      ???????      ??????    ??????    0.0000
RESIDUAL        12      78819.1     6568.26
TOTAL           14      1436706

CASES INCLUDED 15    MISSING CASES 0
```



Malli 2

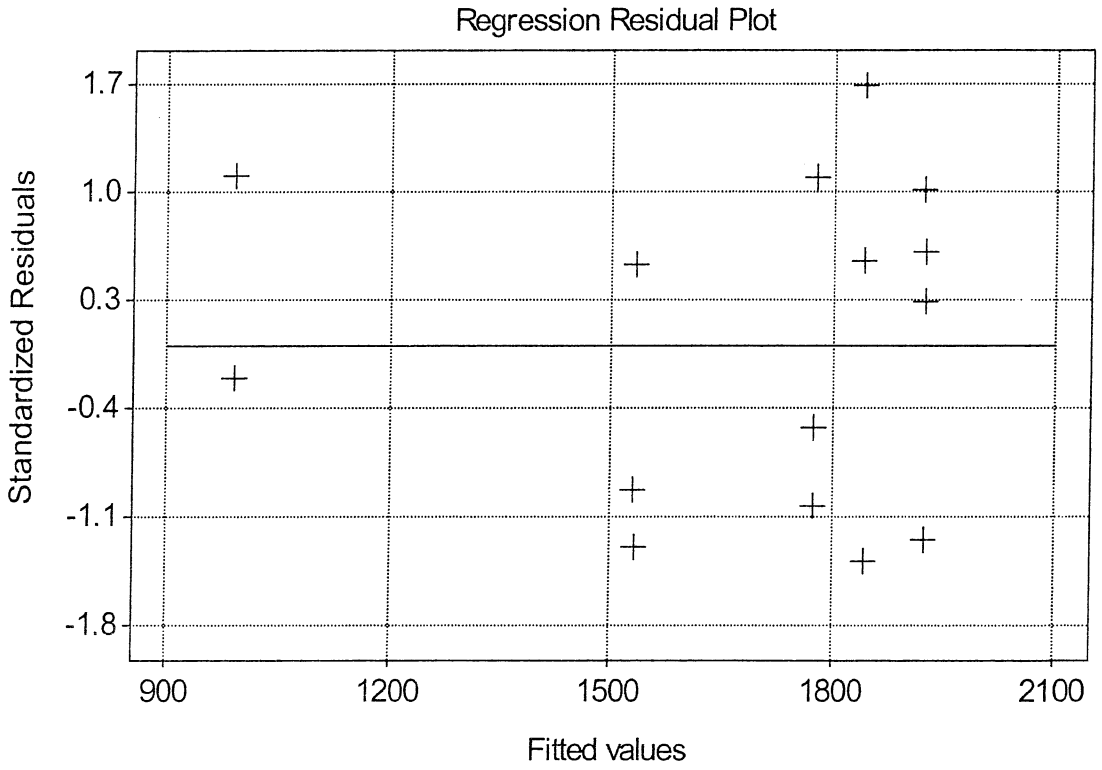
UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF SATO

PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P	VIF
CONSTANT	223.807	114.408	1.96	0.0741	
LANNOITE	8.84699	0.82298	?????	0.0000	27.2
LANNOITE2	-0.01148	0.00132	-8.72	0.0000	27.2

R-SQUARED	??????	RESID. MEAN SQUARE (MSE)	6568.26
ADJUSTED R-SQUARED	0.9360	STANDARD DEVIATION	81.0448

SOURCE	DF	SS	MS	F	P
REGRESSION	2	1357887	678943	?????	??????
RESIDUAL	12	???????	???????		
TOTAL	14	1436706			

CASES INCLUDED 15 MISSING CASES 0



5. Tehtaalla on kolme sähkölamppuja valmistavaa konetta, kone A, kone B ja kone C. Koneiden valmistamien lamppujen paloajat vaihtelevat satunnaisesti jonkin verran noudattaen normaalijakaumaa.

Tehtaan laadunvalvontaosasto halusi tutkia koneiden toimintaa ja poimi koneen A, koneen B ja koneen C valmistamien lamppujen joukosta toisistaan riippumattomat yksinkertaiset satunnaisotokset, joiden koko oli 15. Jokaisesta lampusta mitattiin sen paloaika (yksikkö = 1 h).

Otosten perusteella haluttiin selvittää palavatko eri koneiden tekemät lamput keskimäärin yhtä kauan.

Tulostukset tilastollisesta analyyseista on annettu seuraavalla sivulla.

Huomautus:

Painovirhepaholainen halusi estää vastaamisen ja korvasi osan tulostuksen luvuista kysymysmerkeillä.

Paholainen ei kuitenkaan tiennyt, että puuttuvat luvut voidaan laskea jäljelle jääneistä luvuista.

Puuttuvat luvut ovat *ryhmien välistä vaihtelua kuvaava neliösumma* ja sitä *vastaava keskineliövirhe*, *ryhmien sisäiseen vaihteluun liittyvä vapausasteiden lukumäärä* sekä menetelmässä käytetyn *F-testisuureen arvo*.

Tehtävät:

(a) Mitä tilastollista menetelmää on käytetty?
Kuvaa käytettyä menetelmää lyhyesti.

(b) Mikä on menetelmällä testattu nollahypoteesi?
Mikä on vaihtoehtoinen hypoteesi?

(c) Mikä on Bartlettin testin rooli menetelmän soveltamisessa.

(d) Mitä johtopäätöksiä voidaan tehdä tulostuksesta 2?

(e) Laske puuttuvat luvut.

(f) Poimi tulostuksesta ns. varianssianalyysihajotelma ja esitä sille tulkinta.

(g) Tee johtopäätökset analyysin tuloksista.

$\frac{n-k}{k-1} \cdot \frac{SSM}{SSE}$

Tulostus 1

```
STATISTIX FOR WINDOWS

ONE-WAY AOV FOR: PUTKIA PUTKIB PUTKIC

SOURCE     DF        SS          MS          F        P
-------   ----   ---------   ---------   ------   ------
BETWEEN     2      ???????      ??????     ????   0.0136
WITHIN     42      7077204      168505
TOTAL      44      8685820

                      CHI-SQ     DF        P
BARTLETT'S TEST OF    ------   ------   ------
   EQUAL VARIANCES      0.71      2     0.7007

COCHRAN'S Q                      0.4146
LARGEST VAR / SMALLEST VAR       1.5785

COMPONENT OF VARIANCE FOR BETWEEN GROUPS     42386.9
EFFECTIVE CELL SIZE                             15.0

                          SAMPLE      GROUP
VARIABLE        MEAN       SIZE     STD DEV
---------   ----------   ------   ----------
PUTKIA           10103       15       457.81
PUTKIB           10009       15       403.91
PUTKIC          9663.5       15       364.39
TOTAL           9925.3       45       410.49

CASES INCLUDED 45    MISSING CASES 0
```

Tulostus 2

```
BONFERRONI COMPARISON OF MEANS

                          HOMOGENEOUS
VARIABLE        MEAN      GROUPS
---------   ----------   -----------
LAMPPUA          10103    I
LAMPPUB          10009    I I
LAMPPUC         9663.5    .. I

THERE ARE 2 GROUPS IN WHICH THE MEANS ARE
NOT SIGNIFICANTLY DIFFERENT FROM ONE ANOTHER.

CRITICAL T VALUE                    2.494   REJECTION LEVEL   0.050
CRITICAL VALUE FOR COMPARISON      373.78
STANDARD ERROR FOR COMPARISON      149.89
```