TILM 3558 Harjoitustyö, Tehtävä 2
Lasse Rintakumpu, 63555

Numeeristen vastemuuttujien mallinnus: a. Varianssianalyysitehtävä

Tarkastellaan havaintoaineiston (`elinolo.sav`) perusteella kuntamuodon sekä sukupuolen yhteyttä asunnon pinta-alaan. Ollaan kiinnostuttu kysymyksistä:

1. Onko kuntamuotojen välillä eroa asunnon pinta-alassa?
2. Onko sukupuolten välillä eroa asunnon pinta-alassa?
3. Onko sukupuolten välinen ero erilaista kuntamuodoittain?

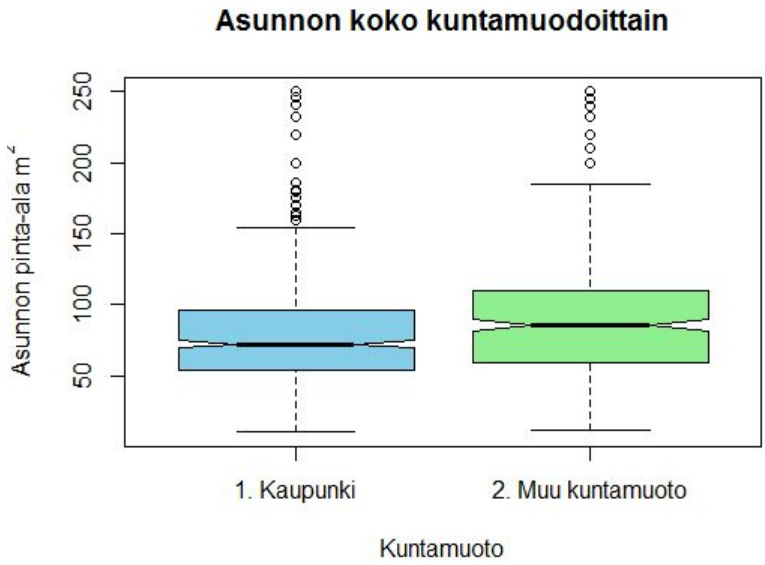
Havaintoaineisto sisältää $n_{kaikki} = 2199$ tilastoyksikköä, joista tähän tarkasteluun on valittu satunnaisesti n = 1000 tilastoyksikköä. [1] Havaintoaineisto sisältää 11 muuttujaa, joista kiinnostuksen kohteena ovat asunnon pinta-ala (`pala`, numeerinen diskreetti, pinta-ala neliömetreinä), kuntamuoto (`kumu`, kategorinen dikotominen, 0 = kaupunki, 1 = muu kuntamuoto) sekä sukupuoli (`supu`, kategorinen dikotominen, 1 = mies, 2 = nainen).

Kuvaillaan ensin havaintoaineistoa graafisesti, jonka jälkeen testataan havaintoainestossa mahdollisesti esiintyvien erojen tilastollista merkitsevyyttä.

1a. Kuntamuodon vaikutus asunnon pinta-alaan

Kuvaillaan havaintoaineistoa Tukeyn laatikko-janakuviolla.

[1] Tarkasteltavat tilastoyksiköt on poimittu SPSS-ohjelmalla. Itse analyysissa on käytetty R-ohjelmistoa. Poimintaan käytetty SPSS-syntaksi sekä analyysiin käytetty R-koodi tehtävän liitteenä.



Laatikko-janakuvion perusteella muualla kuin kaupungissa asuvien asuntojen pinta-ala on keskimääräisesti ja tilastollisesti merkitsevästi suurempi kuin kaupungissa asuvien. Kuvion perusteella molemmat kuntaluokat sisältävät poikkeavia havaintoja, eivätkä jakaumat vaikuta normaaleilta. Testataan jakaumien normaalisuutta Shapiro-Wilkin normaalijakaumatestillä. Tehdään testi tasolla 0.05. Valitaan hypoteesipariksi

H_0: Pinta-ala noudattaa normaalijakaumaa.
H_v: Pinta-ala ei noudata normaalijakaumaa.

Shapiro-Wilkin testillä saadaan W-testisuureen arvoiksi $W_{kaupunki}$= 0.9254 sekä W_{muu} = 0.9379 ja näitä vastaaviksi p-arvoiksi $p_{kaupunki}$ < 2.2e-16 ja p_{muu} = 5.117e-11. Hylätään nollahypoteesi kummankin luokan kohdalla tasolla 0.05. Pinta-alat eivät kummassakaan tapauksessa noudata normaalijakaumaa.

Käytetään luokkien välisen eron testaamiseen Kruskal-Wallis-testiä. [2] Tehdään testi tasolla 0.05. Valitaan hypoteesipariksi

H_0: Luokkien välillä ei ole eroa sijainnissa
H_v: Luokkien välillä on ero sijainnissa.

[2] Luokkien koot $n_{kaupunki}$ = 645 ja n_{muu} = 355 sekä samansuuntaiset vinoudet mahdollistaisivat myös t-testin käytön.

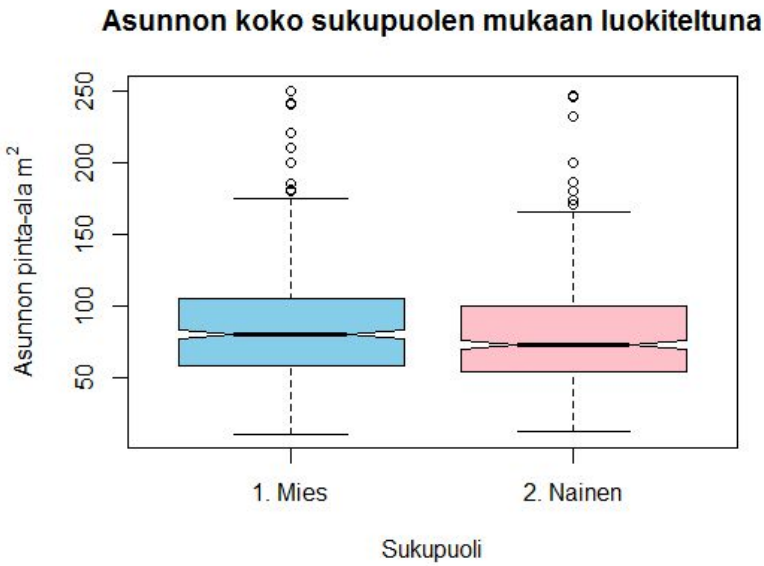
Saadaan tulokseksi

```
data:  pala and kumu
Kruskal-Wallis chi-squared = 28.4112, df = 1, p-value = 9.81e-08.
```

Testin perusteella voidaan nollahypoteesi hylätä riskitasolla 0.05. Aineiston perusteella asuntojen pinta-aloissa on tilastollisesti merkitsevä ero kaupungissa ja muun muotoisissa kunnissa asuvien välillä niin, että kaupungissa asuvien asunnot ovat keskimäärin pienempiä.

2a. Sukupuolen yhteys asunnon pinta-alaan

Tarkastellaan seuraavaksi sukupuolen yhteyttä asunnon kokoon. Kuvaillaan havaintoaineistoa Tukeyn laatikko-janakuviolla.



Kuvion perusteella sukupuolten välillä vaikuttaa olevan tilastollisesti merkitsevää eroa asunnon pinta-aloissa. Siirrytään testaamaan luokkien välillä havaittujen erojen merkitsevyyttä.

Silmämääräisesti sukupuoliluokat eivät vaikuta normaalisti jakautuneilta. Tarkastellaan normaalisuutta Shapiro-Wilk-testillä. Tehdään testi tasolla 0.05. Valitaan hypoteesipariksi

H_0: Pinta-ala noudattaa normaalijakaumaa.
H_v: Pinta-ala ei noudata normaalijakaumaa.

Shapiro-Wilkin testillä saadaan W-testisuureen arvoiksi W_{mies}= 0.9315 sekä W_{nainen} = 0.9239 ja näitä vastaaviksi p-arvoiksi p_{mies} = 5.64e-14 sekä p_{nainen} = 1.294e-15. Hylätään nollahypoteesi molempien sukupuoliluokkien kohdalla tasolla 0.05. Pinta-alat eivät kummassakaan luokassa noudata normaalijakaumaa.

Käytetään luokkien välisen eron testaamiseen Kruskal-Wallis-testiä. [3] Tehdään testi tasolla 0.05. Valitaan hypoteesipariksi

H_0: Sukupuolten välillä ei ole eroa sijainnissa
H_v: Sukupuolten välillä on ero sijainnissa.

Saadaan tulokseksi

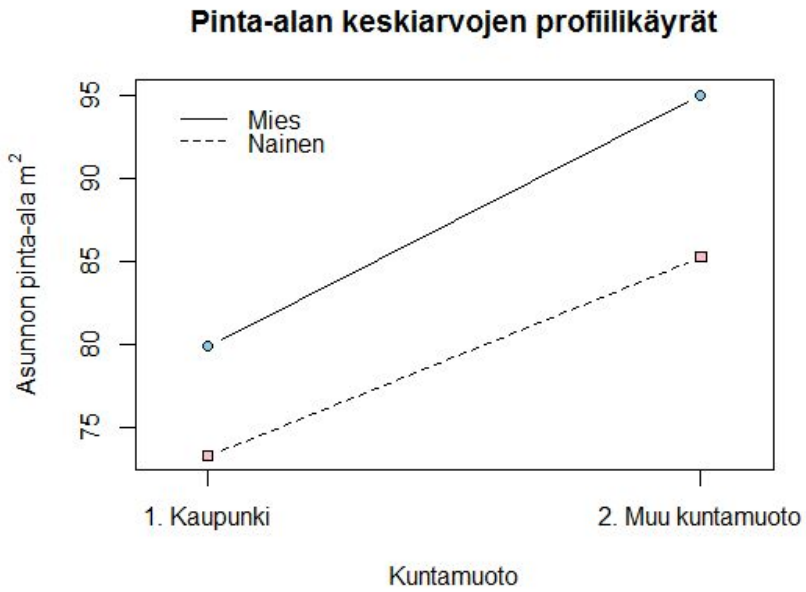
```
data:  pala and supu
Kruskal-Wallis chi-squared = 12.8887, df = 1, p-value = 0.0003306
```

Testin perusteella voidaan nollahypoteesi hylätä riskitasolla 0.05. Aineiston perusteella asuntojen pinta-aloissa on tilastollisesti merkitsevä ero miesten ja naisten välillä, miesten hyväksi. Miehet asuvat siis aineiston perusteella keskimääräisesti isommissa asunnoissa kuin naiset.

3a. Kuntamuodon ja sukupuolen yhdysvaikutus

Tarkastellaan vielä onko sukupuolen ja asunnon pinta-alan välinen yhteys erilaista kaupungissa ja muun muotoisissa kunnissa asuvien välillä. Piirretään pinta-alan keskiarvojen profiilikäyrät.

[3] Luokkien koot n_{mies} = 477 ja n_{nainen} = 523 sekä samansuuntaiset vinoudet mahdollistaisivat myös t-testin käytön.



Kuvaajan perusteella sukupuolien väliset erot ovat samansuuntaisia sekä kaupungissa että muun muotoisissa kunnissa asuvien kesken. Testataan kaksisuuntaisen varianssianalyysin F-testillä sukupuolen ja kuntamuodon yhdysvaikutuksen merkitsevyyttä mallissa. Koska varianssianalyysin oletukset eivät ole voimassa, sovelletaan ensin malliin *aligned rank transform* -muunnosta. [4] Tehdään testi tasolla 0.05. Valitaan hypoteesipariksi:

H_0: Sukupuolen ja kuntamuodon yhdysvaikutus ei ole tilastollisesti merkitsevä
H_v: Sukupuolen ja kuntamuodon yhdysvaikutus on tilastollisesti merkitsevä.

Saadaan tulokseksi:

```
Aligned Rank Transform Anova Table (Type III tests)
Response: art(pala)
Sum             Sq  Df   Df.res F value  Pr(>F)
supu        842829  1    996     10.2084  0.001442
kumu       2312695  1    996    28.4415  1.195e-07
supu:kumu     4186  1    996    0.0501   0.822976.
```

Mallin merkitsevyystestauksessa yhdysvaikutuksen supu*kumu F-testisuureen arvoksi saadaan 0.0501 ja sitä vastaavaksi p-arvoksi 0.823. Testin perusteella sukupuolen ja asunnon yhdysvaikutus ei ole tilastollisesti merkitsevä. Sukupuolen ja asunnon pinta-alan välistä yhteyttä voidaan siis pitää tilastollisesti merkitsevästi samanlaisena kuntamuodosta riippumatta.

[4] Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J.: *The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures*. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11). New York: ACM Press, ss. 143-146. 2011.

4a. Yhteenveto

Aineiston perusteella sekä kuntamuodolla että sukupuolella on tilastollisesti merkitsevä yhteys asunnon pinta-alaan. Muualla kuin kaupungissa asuvien asunnon pinta-ala on keskimääräisesti isompi kuin kaupungissa asuvien. Miesten asunnon pinta-ala on keskimääräisesti isompi kuin naisten. Aineiston perusteella sukupuolen ja kuntamuodon välillä ei ole tilastollisesti merkitsevää yhdysvaikutusta. Miesten ja naisten väliset erot asumispinta-alassa ovat samanlaisia kuntamuodosta riippumatta.

Numeeristen vastemuuttujien mallinnus: b. Regressioanalyysitehtävä

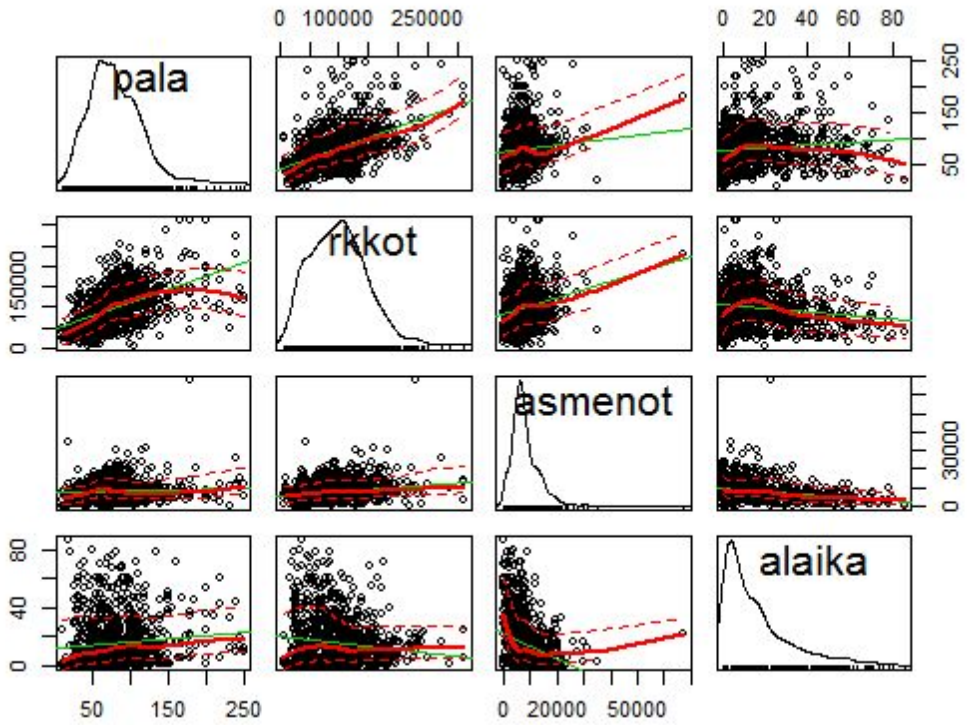
Tarkastellaan havaintoaineiston (`elinolo.sav`) perusteella onko ruokakunnan käytettävissä olevilla tuloilla, asumismenoilla yhteensä ja alueella asumisajalla yhteyttä asunnon pinta-alaan. Havaintoaineisto sisältää $n_{kaikki} = 2199$ tilastoyksikköä. Käytetään tarkastelussa a-kohdassa satunnaisesti valittua $n = 1000$ tilastoyksikköä.

Havaintoaineisto sisältää 11 muuttujaa, joista kiinnostuksen kohteena ovat asunnon pinta-ala (`pala`, numeerinen diskreetti, pinta-ala neliömetreinä), ruokakunnan käytettävissä olevat tulot (`rkkot`, numeerinen diskreetti, euroa vuodessa), asumismenot yhteensä (`asmenot`, numeerinen diskreetti, euroa vuodessa) sekä alueella asumisaika (`alaika`, numeerinen diskreetti, vuotta?).

Pyritään selittämään pinta-alaa lineaarisella regressiomallilla. Lähdetään liikkeelle valitsemalla malliin soveltuvat selittävät muuttujat. Mallin valinnan jälkeen edetään valitun mallin sopivuuden tarkasteluun selitysasteen sekä jäännöstarkastelun pohjalta.

1b. Mallin oletusten ja muuttujien tarkastelu

Aloitetaan analyysi tarkastelemalla muuttujien välisiä riippuvuuksia graafisesti sirontakuvion avulla:



Sirontakuvion perusteella ainoastaan pinta-alan ja ruokakunnan tulojen välinen riippuvuus vaikuttaa lineaariselta. Lasketaan muuttujien väliset korrelaatiokertoimet (Pearsonin korrelaatiokertoimet diagonaalin yläpuolella, Spearmanin korrelaatiokertoimet diagonaalin alapuolella):

	pala	rkkot	asmenot	alaika
pala	**1**	**0.52**	**0.08**	**0.11**
rkkot	0.56	**1**	**0.21**	**-0.13**
asmenot	0.06	0.21	**1**	**-0.27**
alaika	0.19	-0.03	-0.28	**1**

Korrelaatiokertoimista havaitaan, että ainoastaan ruokakunnan tulot korreloivat voimakkaan lineaarisesti ja positiivisesti pinta-alan kanssa, vaikkakin Spearmanin korrelaatiokerroin on hieman Pearsonia isompi. Pinta-alan ja asumismenojen sekä alueella asumisajan välinen lineaarinen korrelaatio on positiivista ja heikkoa.

Sirontakuvion perusteella minkään muuttujan jakauma ei vaikuta normaalilta. Muuttujien pala ja rkkot kaikki arvot ovat positiivisia kokonaislukuja, joten niiden normaalisuutta voidaan parantaa potenssimuunnoksella. Normaalisuuden kannalta optimeiksi potenssimuunnoksiksi löydetään pala^0.3128892 ja rkkot^0.5585493. Normaalisuuden lisäksi muunnosten jälkeen myös Pearsonin korrelaatiokerroin paranee hieman, 0.52:stä 0.56:een . Sovelletaan dataan muunnokset ja jatketaan analyysia muunnetulla datalla.

2b. Mallin valinta

Valitaan lineaarisen regression malli käyttäen taaksepäin askeltavaa mallinvalintaa (R:n MASS-paketista löytyvällä stepAIC-funktiolla). Aloitetaan täydellisestä mallista pala~rkkot+asmenot+alaika ja päädytään malliin:

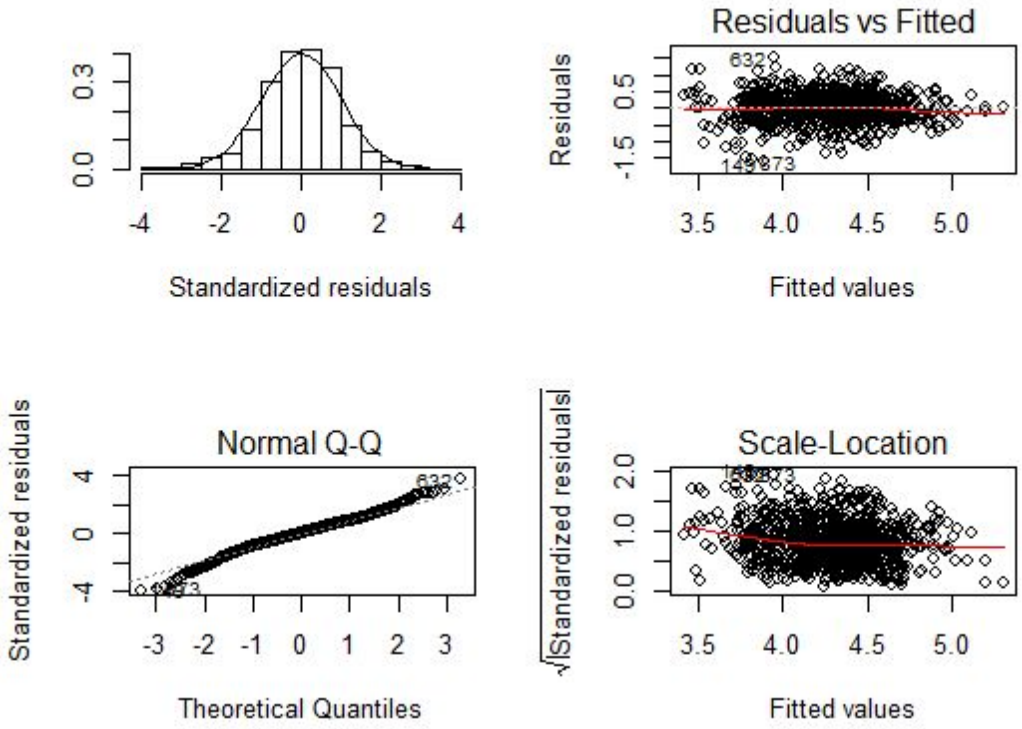
```
Estimate    Std.       Error       t value  Pr(>|t|)
(Intercept) 3.1764747  0.0497292   63.88    < 2e-16
rkkot       0.0017090  0.0000750   22.79    < 2e-16
alaika      0.0053713  0.0007841    6.85    1.29e-11

Residual standard error: 0.4016 on 997 degrees of freedom
Multiple R-squared:  0.3488,  Adjusted R-squared:  0.3475
F-statistic:   267 on 2 and 997 DF,  p-value: < 2.2e-16.
```

Jossa pinta-alan merkitseviksi selittäjiksi jäävät ruokakunnan tulot sekä alueella asumisaika. Mallin p-arvoksi saadaan < 2.2e-16, selitysasteeksi 0.3488 ja vertailukelpoiseksi selitysasteeksi 0.3475. Eteenpäin mallista pala~rkkot askeltaen päädytään samaan malliin. Vertailun vuoksi mallin pala~rkkot vertailukelpoiseksi selitysasteeksi saadaan 0.3175 ja p-arvoksi < 2.2e-16. Valitaan siis jatkotarkasteluun tilastollisesti merkitsevä (p< 2.2e-16) malli pala~rkkot+alaika, jonka regressioyhtälö on pala=rkkot*0.0017+alaika*0.0054+3.1765.

3b. Mallin jatkotarkastelu ja yhteenveto

Tarkastellaan mallin sopivuutta graafisesti. Piirretään mallin standardoitujen jäännösten histogrammi, normaalikvantiilikuvaaja sekä jäännösten jakaumia kuvaavat sirontakuviot (ennustetut arvot*jäännökset sekä ennustetut arvot*standardoitujen jäännösten neliöjuuri)):



Standardoitujen jäännösten jakauma on symmetrinen nollan molemmin puolin ja noudattaa muutamia poikkeavia havaintoja lukuunottamatta kohtalaisesti normaalijakaumaa. Jäännösten hajonta myös vaikuttaa vakiolta sovitteiden arvoalueella. Graafisen jäännöstarkastelun perusteella malli pala~rkkot+alaika on sopiva pinta-alan selittämiseen ja sen avulla voidaan pinta-alan vaihtelusta selittää noin 35%. Mallin sopivuutta voitaisiin todennäköisesti parantaa poistamalla datasta joitakin vahvasti poikkeavia havaintoja.

Liitteet

Liitteenä tilastoyksikköjen valinnassa käytetty SPSS-syntaksi sekä analyysissa käytetty R-koodi.

Liite 1. SPSS-syntaksi

```
SET SEED=63555.
DATASET ACTIVATE DataSet1.
USE ALL.
do if $casenum=1.
compute #s_$_1=1000.
compute #s_$_2=2199.
end if.
```

```
do if #s_$_2 > 0.
compute filter_$=uniform(1)* #s_$_2 < #s_$_1.
compute #s_$_1=#s_$_1 - filter_$.
compute #s_$_2=#s_$_2 - 1.
else.
compute filter_$=0.
end if.
VARIABLE LABELS filter_$ '1000 from the first 2199 cases (SAMPLE)'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

Liite 2. R-koodi

```
########################################
# TILM3558 Harjoitustyö, Osa 2, R-koodi #
# Lasse Rintakumpu, 63555               #
# 25.8.2015                             #
########################################

# Asetetaan työhakemisto
wd <- "D:/Dropbox/Edu/Statistics/TILM 3558 Harjoitustyö" #Hipatlaptop
setwd(wd)

# Funktio kirjastojen asentamiselle / lataamiselle
lataa_kirjasto <- function(kirjasto) {
  if(kirjasto %in% rownames(installed.packages()) == FALSE)
  {install.packages(kirjasto)}
  library(kirjasto, character.only = TRUE)
}

# Ladataan/asennetaan käytetyt kirjastot
lapply(c("moments", "car", "ARTool", "corrplot", "MASS"), lataa_kirjasto)

# Ladataan havaintoaineisto
elinolo <-
read.csv("https://raw.githubusercontent.com/rintakumpu/tilm3558/master/elinol
o_filtered.csv", sep=",", dec=".", header=TRUE, row.names=NULL, fileEncoding
= "UTF-8-BOM");

# Tallennetaan käsiteltävä data omiin muuttujiinsa,
# valitaan samalla tiedot vain niiltä 1000 riviltä, jotka on satunnaisesti
# valittu tarkasteluun koko datasta

pala <- elinolo$pala[elinolo$filter==1] # Asunnon pinta-ala, neliömetreinä
kumu <- as.factor(elinolo$kumu[elinolo$filter==1]) # Luokitteleva muuttuja
```

```
kuntamuoto. 0: Kaupunki, 1: Muu kuntamuoto
supu  <- as.factor(elinolo$supu[elinolo$filter==1]) # Luokitteleva sukupuoli,
1: Mies, 2:Nainen

###############################################
# 1a. Kuntamuodon yhteys asunnon pinta-alaan #
###############################################

# Tarkastellaan asunnon pinta-alaa kuntomuodoittain

length(pala[kumu==0]) # 645
length(pala[kumu==1]) # 355
boxplot(pala[kumu==0], pala[kumu==1],
        notch = T, xlab = "Kuntamuoto", ylab = expression(paste("Asunnon
pinta-ala m  " ^{2})), xlab="ahtas", main="Asunnon koko kuntamuodoittain",
        col=rep(c("skyblue","lightgreen")), xaxt="n")
axis(side=1, labels=c("1. Kaupunki", "2. Muu kuntamuoto"), at=c(1,2))
pdf('boxplot_pala_kumu.pdf')
dev.off()

# Kuvan perusteella asunnon pinta-aloissa vaikuttaa olevan tilastollisesti
# merkitsevä (95%-CI) ero kuntamuotojen välillä

# Lasketaan kuntamuotojen piste-estimaatit

mean(pala[kumu==0], na.rm=T) # Kaupunki: 76.30698
mean(pala[kumu==1], na.rm=T) # Muu kuntamuoto: 90.27887

# Jakaumien vinous

skewness(pala[kumu==0]) # 1.298357
skewness(pala[kumu==1]) # 1.045394

# Testataan jakaumien normaalisuutta

shapiro.test(pala[kumu==0])
#data:  pala[kumu == 0]
#W = 0.9254, p-value < 2.2e-16

shapiro.test(pala[kumu==1])
#data:  pala[kumu == 1]
#W = 0.9379, p-value = 5.117e-11

# Jakaumia ei voida pitää normaaleina, siirrytään epäparametriseen
testaukseen

# Käytetään testaamiseen Kruskal-Wallis-testiä
# h0: Luokkien välillä ei ole eroa sijainnissa.
# hv: Luokkien välillä on ero sijainnissa.
```

```
kruskal.test(pala, kumu)

# Kruskal-Wallis rank sum test
# data:  pala and kumu
# Kruskal-Wallis chi-squared = 28.4112, df = 1, p-value = 9.81e-08

##############################################
# 2b. Sukupuolen yhteys asunnon pinta-alaan #
##############################################

# Tarkastellaan asuntojen pinta-alaa ryhmittelevänä tekijänä sukupuoli
par(mar=c(5,5,4,2), xpd=T, xaxt="s")
boxplot(pala[supu==1], pala[supu==2],
        notch = T, xlab = "Sukupuoli", ylab = expression(paste("Asunnon
pinta-ala m  " ^{2})), xlab="ahtas", main="Asunnon koko sukupuolen mukaan
luokiteltuna",
        xaxt="n", col=rep(c("skyblue","pink")),
        at=c(1,2))
axis(side=1, labels=c("1. Mies", "2. Nainen"), at=c(1,2))
pdf('boxplot_supu_pala.pdf')
dev.off()

# Kuvan perusteella asunnon pinta-aloissa vaikuttaa olevan tilastollisesti
# merkitsevä (95%-CI) ero sukupuolien välillä.

length(pala[supu==1]) # 477
length(pala[supu==2]) # 523

# Siirrytään tilastolliseen testaukseen. Testataan ensin jakaumien
# normaalisuutta.

shapiro.test(pala[supu==1])

# Shapiro-Wilk normality test
# data:  pala[supu == 1]
# W = 0.9315, p-value = 5.64e-14

shapiro.test(pala[supu==2])

# data:  pala[supu == 2]
# W = 0.9239, p-value = 1.294e-15

# Jakaumia ei voida pitää normaaleina, siirrytään epäparametriseen
testaukseen

# Käytetään testaamiseen Kruskal-Wallis-testiä
# h0: Luokkien välillä ei ole eroa sijainnissa.
# hv: Luokkien välillä on ero sijainnissa.
```

```
kruskal.test(pala, supu)

# data:  pala and supu
# Kruskal-Wallis chi-squared = 12.8887, df = 1, p-value = 0.0003306

##################################################
# 3c. Asumisahtauden ja kuntamuodon yhdysvaikutus #
##################################################

interaction.plot(kumu, supu, pala, legend = FALSE, xlab = "Kuntamuoto", lty =
1:2,
                 ylab = expression(paste("Asunnon pinta-ala m  " ^{2})),
                 xtick=TRUE, main = "Pinta-alan keskiarvojen profiilikäyrät",
                 xaxt="n", type="b", pch=c(21,22), bg=c("skyblue","pink"))

axis(side=1, labels=c("1. Kaupunki", "2. Muu kuntamuoto"), at=c(1,2))

legend(x=0.9,y=95,legend=c("Mies", "Nainen"), inset=1,
       bg = par("bg"), bty = "n", lty = c(1,2))

pdf('interaction_plot_supu_kumu.pdf')
dev.off()

# Käytetään aligned rank transform -muunnosta dataan, jotta
# voidaan testata yhteisvaikutuksen merkitsevyyttä.

anova(art(pala~supu+kumu+supu:kumu))

# Aligned Rank Transform Anova Table (Type III tests)
# Response: art(pala)
# Sum Sq Df Df.res F value    Pr(>F)
# supu       842829  1    996 10.2084  0.001442 **
# kumu      2312695  1    996 28.4415 1.195e-07 ***
# supu:kumu    4186  1    996  0.0501  0.822976

##################################################
# 1b. Mallin oletusten ja muuttujien tarkastelu #
##################################################

# Ladataan data omiin muuttujiinsa
elinolo <-
read.csv("https://raw.githubusercontent.com/rintakumpu/tilm3558/master/elinol
o_filtered_b.csv", sep=",", dec=".", header=TRUE, row.names=NULL,
fileEncoding = "UTF-8-BOM");
rkkot <- elinolo$rkkot[elinolo$filter==1] # Kotitalouden tulot, euroa per
vuosi?
asmenot <- elinolo$asmenot[elinolo$filter==1] # Asumismenot, euroa per vuosi?
alaika <- elinolo$alaika[elinolo$filter==1] # Alueella asumisaika, kuukautta?
```

```
# Ladataan data matriisiin
elinolo_matrix <- matrix(c(pala,rkkot,asmenot,alaika), ncol=4, byrow=FALSE,
                         dimnames=list(c(1:1000),c("pala", "rkkot",
"asmenot", "alaika" )))


# Piirretään sirontakuvio
scatterplotMatrix(~pala+rkkot+asmenot+alaika, data=elinolo_matrix)
pdf ('scatterplot_pala_rkkot_asmenot_alaika.pdf')
dev.off()

# Piirretään korrelaatiomatrisi
corrplot(cor(elinolo_matrix, method="pearson"), method="number", col="black",
type="upper", tl.pos="tl", diag=TRUE, cl.pos="n", tl.col="black")
corrplot(cor(elinolo_matrix, method="spearman"), col="#777777", add=TRUE,
method="number", type="lower", tl.pos="n", cl.pos="n", tl.col="black",
diag=FALSE)

pdf('corrplot_pala_rkkot_asmenot_alaika.pdf')
dev.off()

# Riippuvuus vain palan ja rkkotin välillä lineaarista, Rs > Rp.
# Jakaumat eivät normaaleja, sovelletaan muunnoksia dataan

pala_pt <- powerTransform(pala) # 0.3128892
rkkot_pt <- powerTransform(rkkot) # 0.5585493

elinolo_matrix_transform <- matrix(c(pala^0.3128892,rkkot^0.5585493,
asmenot,alaika), ncol=4, byrow=FALSE,
                         dimnames=list(c(1:1000),c("pala", "rkkot",
"asmenot", "alaika" ))

cor(elinolo_matrix_transform[,"pala"],elinolo_matrix_transform[,"rkkot"],
method="pearson")
# 0.5640465
cor(elinolo_matrix_transform[,"pala"],elinolo_matrix_transform[,"rkkot"],
method="spearman")
# 0.5567287

######################
# 2b. Mallin valinta #
######################

malli1 <- lm(pala~rkkot+asmenot+alaika,
data=as.data.frame(elinolo_matrix_transform))
summary(stepAIC(malli1, direction="backward"))
malli2 <- lm(pala~rkkot+alaika, data=as.data.frame(elinolo_matrix_transform))
summary(stepAIC(malli2, direction="backward"))
```

```
# Coefficients:
# Estimate Std. Error t value Pr(>|t|)
# (Intercept) 3.1764747  0.0497292   63.88  < 2e-16 ***
#  rkkot       0.0017090  0.0000750   22.79  < 2e-16 ***
#  alaika      0.0053713  0.0007841    6.85 1.29e-11 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 0.4016 on 997 degrees of freedom
# Multiple R-squared:  0.3488,  Adjusted R-squared:  0.3475
# F-statistic:   267 on 2 and 997 DF,  p-value: < 2.2e-16

# Eteenpäin askeltaen
malli3 <- lm(pala~rkkot, data=as.data.frame(elinolo_matrix_transform))
malli4 <- lm(pala~rkkot+asmenot,
data=as.data.frame(elinolo_matrix_transform))
summary(stepAIC(malli3, direction="forward"))
summary(stepAIC(malli4, direction="forward"))
summary(stepAIC(malli2, direction="forward")) # Päädytään samaan malliin

summary(malli3)

# Residual standard error: 0.4107 on 998 degrees of freedom
# Multiple R-squared:  0.3181,  Adjusted R-squared:  0.3175
# F-statistic: 465.7 on 1 and 998 DF,  p-value: < 2.2e-16

# Jatketaan mallilla 2
coefficients(malli2)
malli <- malli2
# 3.176474690 0.001708954 0.005371325
# Regressioyhtälö:
# pala=rkkot*0.0017+alaika*0.0054+3.1765.

##############################################
# 3b. Mallin jatkotarkastelu ja yhteenveto #
##############################################

# Tarkastellaan valitun mallin (malli3) jäännöksiä
par(mfrow=c(2,2))

# Jäännösten histogrammi
hist(rstandard(malli), freq=FALSE, main="", ylab="", xlab="Standardized
residuals")
xfit<-seq(min(rstandard(malli)),max(rstandard(malli)),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
plot(malli)
```