Konsta Nyman
523834

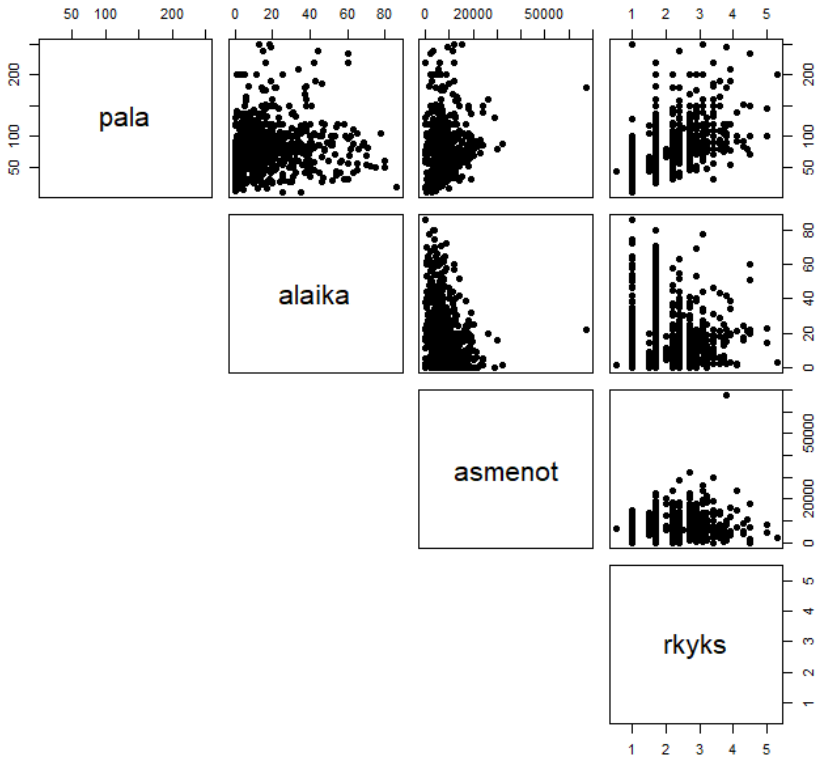
Tilastollisten mallien peruskurssi – harjoitustyö

Alustetaan 800:n kokoinen otos elinolo-tilastosta:

```
library(foreign)
ht1.dat<-read.spss("elinolo2020.sav", to.data.frame=TRUE)
attach(ht1.dat)

set.seed(523834)
# 800 kokoinen otos
oma.otos1<-ht1.dat[sample(nrow(ht1.dat), 800), ]
attach(oma.otos1)
```

1. Regressiomalli

Luodaan sirontakuvio:



Tarkasteltaessa asunnon pinta-alan suhdetta muihin, vaikuttaa siltä, että ainoastaan kuluttajayksiköiden lukumäärällä on yhteys siihen.

Konsta Nyman
523834

Korrelaatiot:

```
> cor.test(rkyks, pala, method="pearson")

        Pearson's product-moment correlation

data:  rkyks and pala
t = 17.616, df = 798, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4773269 0.5772761
sample estimates:
      cor
0.5291344

> cor.test(asmenot, pala, method="pearson")

        Pearson's product-moment correlation

data:  asmenot and pala
t = 2.2387, df = 798, p-value = 0.02545
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.009739594 0.147506855
sample estimates:
      cor
0.07900042

> cor.test(alaika, pala, method="pearson")

        Pearson's product-moment correlation

data:  alaika and pala
t = 3.7046, df = 798, p-value = 0.0002263
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06126555 0.19756113
sample estimates:
      cor
0.1300275
```

```
> cor.test(rkyks, pala, method="spearman")

        Spearman's rank correlation rho

data:  rkyks and pala
S = 35575139, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5831032

Warning message:
In cor.test.default(rkyks, pala, method = "spearman") :
  Cannot compute exact p-value with ties
> cor.test(asmenot, pala, method="spearman")

        Spearman's rank correlation rho

data:  asmenot and pala
S = 82549384, p-value = 0.3568
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.03262289

Warning message:
In cor.test.default(asmenot, pala, method = "spearman") :
  Cannot compute exact p-value with ties
> cor.test(alaika, pala, method="spearman")

        Spearman's rank correlation rho

data:  alaika and pala
S = 65008585, p-value = 8.814e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2381795

Warning message:
In cor.test.default(alaika, pala, method = "spearman") :
  Cannot compute exact p-value with ties
```

Pearsonin korrelaatiokertoimien mukaan kaikki selittäjät ovat tilastollisesti merkitseviä, mutta ainoastaan kuluttajayksiköiden lukumäärän korrelaatio asunnon pinta-alaan on edes jokseenkin suuri.

Myös Spearmanin korrelaatioista ainoastaan kuluttajayksiköiden lukumäärän voidaan sanoa korreloivan asunnon pinta-alan kanssa, vaikka myös alueella asumisaika on tilastollisesti merkitsevä selittäjä. Voidaan sanoa, että alueella asumisajan ja asunnon pinta-alan välillä on heikkoa ei-lineaarista riippuvuutta.

Suoraviivauksien tarkastelu:

Tilastollisesti merkitseville selittäjille Spearmanin korrelaatiot ovat hieman suurempia, joten ei ole syytä ajatella yhteyksien olevan suoraviivaisia.

Konsta Nyman
523834

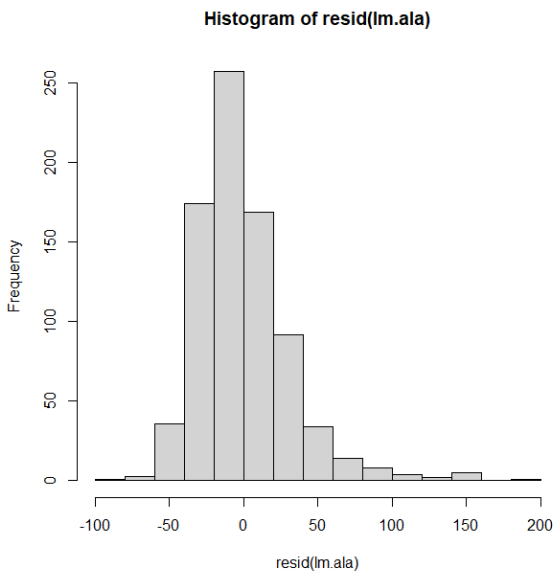
Yksinkertainen regressiomalli:

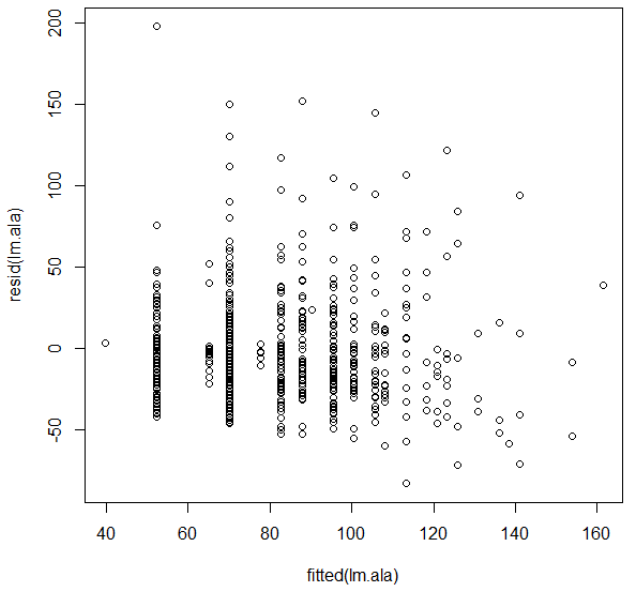
```
> # Yksinkertainen regressiomalli
> lm.ala <- lm(pala~rkyks)
> summary(lm.ala)

Call:
lm(formula = pala ~ rkyks)

Residuals:
    Min      1Q  Median      3Q     Max
-83.209 -20.779  -5.403  14.900 197.651

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   26.991      3.334   8.096 2.12e-15 ***
rkyks         25.358      1.440  17.616  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.36 on 798 degrees of freedom
Multiple R-squared:   0.28,     Adjusted R-squared:  0.2791
F-statistic: 310.3 on 1 and 798 DF,  p-value: < 2.2e-16
```

Mallin selitysaste: 0.28, eli asunnon pinta-ala seuraa kuluttajayksiköiden lukumäärää 28 %:n tarkkuudella

p<0.001, joten malli on merkitsevä

Jäännöstarkastelu:





Jäännösten hajonta vaikuttaa hieman laskevan. Jakauma on myös positiivisesti vino, eli huomataan että yhteys ei ole suoraviivainen.

Konsta Nyman
523834

2. Toistomittausmalli

Normaalijakaumatestit kaikille:

```
> shapiro.test(Functional_M1)

        Shapiro-Wilk normality test

data:  Functional_M1
W = 0.65794, p-value < 2.2e-16

> shapiro.test(Functional_M2)

        Shapiro-Wilk normality test

data:  Functional_M2
W = 0.86477, p-value < 2.2e-16
```

Kaikille p-arvo<0.001, joten kaikki poikkeaa normaalijakaumasta merkittävästi.

Dataa on kuitenkin huomattavan paljon, joten voidaan suorittaa toistettujen mittausten varianssianalyysi.

Normaalijakaumatestit naisille erikseen:

```
> shapiro.test(Functional_M1)

        Shapiro-Wilk normality test

data:  Functional_M1
W = 0.62249, p-value < 2.2e-16

> shapiro.test(Functional_M2)

        Shapiro-Wilk normality test

data:  Functional_M2
W = 0.85073, p-value < 2.2e-16
```

Normaalijakaumatestit miehille erikseen:

```
> shapiro.test(Functional_M1)

        Shapiro-Wilk normality test

data:  Functional_M1
W = 0.70835, p-value < 2.2e-16

> shapiro.test(Functional_M2)

        Shapiro-Wilk normality test

data:  Functional_M2
W = 0.89278, p-value = 6.369e-10
```

Konsta Nyman
523834

Toistettujen mittausten varianssianalyysi:

Kaikki:

```
> fit1 = aov(Functional_M1 ~ Functional_M2, data=oma.oto
s2)
> summary(fit1)
              Df Sum Sq Mean Sq F value   Pr(>F)
Functional_M2   1   9.92   9.923   29.95 7.08e-08 ***
Residuals     492 163.02   0.331
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
6 observations deleted due to missingness
```

F-testin p-arvo<0.001, eli on tilastollisesti merkitseviä eroja mielipiteiden keskiarvot eroavat toisistaan.

Naiset:

```
> fit1 = aov(Functional_M1 ~ Functional_M2, data=oma.oto
s2)
> summary(fit1)
              Df Sum Sq Mean Sq F value   Pr(>F)
Functional_M2   1   9.92   9.923   29.95 7.08e-08 ***
Residuals     492 163.02   0.331
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
6 observations deleted due to missingness
```

F-testin p-arvo<0.001, eli on tilastollisesti merkitseviä eroja mielipiteiden keskiarvot eroavat toisistaan.

Miehet:

```
> fit1 = aov(Functional_M1 ~ Functional_M2, data=oma.oto
s2)
> summary(fit1)
              Df Sum Sq Mean Sq F value   Pr(>F)
Functional_M2   1   9.92   9.923   29.95 7.08e-08 ***
Residuals     492 163.02   0.331
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
6 observations deleted due to missingness
```

F-testin p-arvo<0.001, eli on tilastollisesti merkitseviä eroja mielipiteiden keskiarvot eroavat toisistaan.

Konsta Nyman
523834

3. Kategoristen vastemuuttujien mallitus

```
> # logistinen binäärinen regressio
> logr_tyotilan <- glm(d32 ~ d2+ika, data=oma.otos3, family=binomial)
> logr_tyotilan

Call:  glm(formula = d32 ~ d2 + ika, family = binomial, data = oma.otos3)

Coefficients:
(Intercept)     d2Nainen          ika
    0.04360     -0.02946     -0.04028

Degrees of Freedom: 782 Total (i.e. Null);  780 Residual
  (17 observations deleted due to missingness)
Null Deviance:      628.3
Residual Deviance: 583.5        AIC: 589.5
```

```
> summary(logr_tyotilan)

Call:
glm(formula = d32 ~ d2 + ika, family = binomial, data = oma.otos3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9048  -0.6042  -0.4140  -0.3231   2.2970

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.043596   0.307851   0.142    0.887
d2Nainen    -0.029465   0.213615  -0.138    0.890
ika         -0.040283   0.006389  -6.305 2.89e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 628.26  on 782  degrees of freedom
Residual deviance: 583.50  on 780  degrees of freedom
  (17 observations deleted due to missingness)
AIC: 589.5

Number of Fisher Scoring iterations: 5
```

sukupuoli p=0.890, ei ole merkitsevä selittäjä

ikä p<0.001, on merkitsevä selittäjä

Konsta Nyman
523834

Merkitsevien selittäjien (ainoastaan ikä) malli:

```
                   OR     2.5 %    97.5 %
(Intercept) 1.0445600 0.5693737 1.9072751
d2Nainen    0.9709652 0.6382134 1.4770582
ika         0.9605178 0.9482632 0.9723609
```

selittäjä	**p**	**OR**	**95 % luottamusvälin alaraja OR:lle**	**95 % luottamusvälin yläraja OR:lle**
ikä	<0.001	0.960	0.948	0.972

Mallin selitysaste (Nagelkerke): 0.101

```
$R2
[1] 0.1007193
```

Konsta Nyman
523834

4. Monimuuttujamenetelmät

Pääkomponenttianalyysi:

```
> summary(pca)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     3.4368 2.02391 1.87203 1.57263 1.34644 1.31805 1.22340 1.15048 1.10279 1.05957
Proportion of Variance 0.2881 0.09991 0.08548 0.06032 0.04422 0.04237 0.03651 0.03228 0.02966 0.02738
Cumulative Proportion  0.2881 0.38799 0.47347 0.53379 0.57801 0.62038 0.65688 0.68917 0.71883 0.74621
                         PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18   PC19    PC20
Standard deviation     1.00819 0.94187 0.92815 0.85943 0.84235 0.80617 0.75938 0.74451 0.7301 0.70810
Proportion of Variance 0.02479 0.02164 0.02101 0.01801 0.01731 0.01585 0.01406 0.01352 0.0130 0.01223
Cumulative Proportion  0.77100 0.79264 0.81365 0.83166 0.84897 0.86482 0.87889 0.89241 0.9054 0.91764
                         PC21    PC22    PC23    PC24    PC25    PC26    PC27    PC28    PC29    PC30
Standard deviation     0.68918 0.65229 0.62300 0.58455 0.56470 0.5358 0.52260 0.47338 0.40550 0.33992
Proportion of Variance 0.01158 0.01038 0.00947 0.00833 0.00778 0.0070 0.00666 0.00547 0.00401 0.00282
Cumulative Proportion  0.92922 0.93960 0.94907 0.95740 0.96518 0.9722 0.97884 0.98431 0.98832 0.99114
                         PC31    PC32    PC33    PC34    PC35    PC36    PC37    PC38    PC39    PC40
Standard deviation     0.31029 0.26654 0.23848 0.20956 0.20128 0.15525 0.12091 0.09486 0.06215 0.05642
Proportion of Variance 0.00235 0.00173 0.00139 0.00107 0.00099 0.00059 0.00036 0.00022 0.00009 0.00008
Cumulative Proportion  0.99348 0.99522 0.99661 0.99768 0.99866 0.99925 0.99961 0.99983 0.99992 1.00000
                          PC41
Standard deviation     1.525e-15
Proportion of Variance 0.000e+00
Cumulative Proportion  1.000e+00
```

Valitaan kolme pääkomponenttia, joista lataukset:

```
> pca.chosen <- pca$rotation[,1:3]
> pca.promax <- promax(pca.chosen)
> pca.promax
$loadings

Loadings:
                                PC1    PC2    PC3
autom_lainan_perinta_luok      -0.204
lainojen_lukumaara_luok        -0.241
asuntoluotot1_luok             -0.223
automaattinostoja_luok         -0.131  0.235
vakuutus_a_luok
asuntolaina_a_kpl_luok         -0.185
asuntolaina_b_kpl_luok                -0.150
vakuutus_b_luok                 0.128
vakuutus_c_luok                 0.126
korkeakork_kpl_luok             0.221  0.125
rahasto_a1_luok                       -0.112
pankkikorttilkm_luok                   0.279
luottokortteja_yhteensa_luok
maaraaikaistileja_luok          0.285  0.104
maksuautomaattitapahtumia_luok -0.150  0.118
kayttotili_tal_luok             0.299  0.181
kayttotili_vel_luok            -0.112
asuntolaina_c_kpl_luok         -0.155
osakkeet_euroa_1_luok                        -0.353
eri_osakesarjoja_luok                        -0.420
rahasto_b1_luok                              -0.381
ottoja_luok                    -0.113  0.278
pkorttimaksuja_luok                    0.314
panoja_luok                            0.362
asuntolaina_d_kpl_luok         -0.125
palveluja_kpl_luok             -0.139  0.228
rahastolajeja_luok                           -0.422
lainarastit_luok               -0.191 -0.161  0.168
saastotililla_luok              0.320  0.141
asuntolaina_e_kpl_luok
suoraveloituksia_luok           0.245  0.392
netissa_maksut_luok            -0.177  0.157
maksupalvelussa_maksut_luok     0.293  0.297  0.142
tiskilla_maksut_luok            0.248  0.276  0.169
tilinylityspaivat_luok         -0.193
toimeksianto_a_kpl_luok
toimeksianto_b_kpl_luok
kv_maksukortit_luok                    0.162 -0.187
rahasto_c1_luok                              -0.393
korttiluotot1_luok             -0.128
kulutusluotot1_luok            -0.181         0.156
```