Uni

Johannes Piipponen

9 10 2019

Tämä dokumentti on tehty kokonaan R-ohjelmalla

Datana toimii itse keräämäni päivittäinen unidata (havaintoja 587 päivältä).
Sisällys 1) Datan kuvaus ja muokkaaminen 2) Frekvenssit, korrelaatiot, riippumattomuustestit, keskiarvot 3) OLS regressio (selittävät muuttujat dikotomisia) sekä OLS oletusten tarkastelu 4) OLS regressio 0 1 2 datalla 5) ANOVA ja Tukey multiple pairwais comparison 6) Logistinen regressio 7) Pari kuviota

Ladataan pari pakettia joilla päästään alkuun ja unidata excelistä

Filtteröidään tyhjät rivit pois, määritetään että aika-sarake on varmasti aikaformaatissa, katsotaan hieman miltä data näyttää.

```
library(readxl)     #luetaan excel
library(tidyverse) #piiputus, kuvat, kaikki
library(lubridate) #aggrekoidaan vuodet
library(pander)    #siistit taulukot

options(digits = 3)

df=read_excel("loki.xlsx", col_names =TRUE,na="") %>%
  filter(complete.cases(.))  %>%
  mutate(aika=as.Date(aika))

df %>% head() %>% pander()
```

Table 1: Table continues below

aika	vknpv	unituntia	unettomuus	myohaan	urheilu	aivotyo
2017-12-01	pe	7.5	0	0	0	1
2017-12-02	la	6	1	0	1	1
2017-12-03	su	7.5	0	0	1	0
2017-12-04	ma	7	0	0	1	0
2017-12-05	ti	7.5	0	0	0	0
2017-12-06	ke	6.3	1	0	1	0

tukevaruoka	alko	kahvi	ressi	kipea	hyvin	dota	sauna	suihku	kvalo
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0
0	1	0	0	0	1	0	0	0	0
0	0	0	0	0	1	0	0	0	0
1	0	0	0	0	1	0	0	0	0
1	1	0	0	0	1	0	0	0	0

Kuvataan data

muuttuja	selite
aika	pvm jolloin tapahtumat mitataan
vknpv	viikonpaiva
unituntia	unen maara tunteina
unettomuus	ei unettomuutta=0, selvasti pitka ja turhauttava aika ennen kuin nukahtaa =1, stressiheraaminen aamulla=2
myohaan	ennen klo23 saa arvon 0, 23-00 arvon 1 ja 00–>arvon 2
urheilu	ei urheilua=0, urheilu paivan aikana=1, kova urheilu illalla klo 18 jalkeen=2
aivotyo	ei aivotyota illalla klo 19 jalkeen=0, vaativaa ajattelua klo 19 jalkeen=1
tukevaruoka	ei tuhtia illallista=0, raskas ruoka klo 19 jalkeen
alko	ei alkoholia=0, alkoholia enemman kuin siemaus=1
kahvi	ei kahvia=0, kahvi aamulla=1, kahvi klo12 jÃ¤lkeen 2
ressi	ei stressia=0, selvÃ¤ stressintuntu seuraavasta pÃ¤ivasta=1
kipea	terve kuin pukki=0, kipea tai selva puolikunto=1
hyvin	riidoissa vaimon kanssa=0, menee hyvin=1
dota	ei dotaa=0,dota 19-21 =1, dota 21 –> =2
sauna	ei saunaa=0, sauna klo 19-21=1, sauna klo 21–>=2
suihku	suihku illalla ennen nukkumaanmenoa
kvalo	kirkasvalo tunteina aamulla ennen klo 09:30

Muutetaan kaikki data binaariseksi

Aineistossa siirryttiin jossain välissä 0 1 datasta 0 1 2 dataan. Muutetaan tässä vaiheessa kaikki data 0 1 muotoon siten, että 2 –> 1

Frekvenssejä

Kahvia juodaan harvoin, mutta juonti lisää selvästi unettomuutta. Kun kahvia juodaan, unettomuuden yleisyys kasvaa kolmenkertaiseksi.

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |          Column Percent |
## |-------------------------|
##
## Total Observations in Table:  587
```

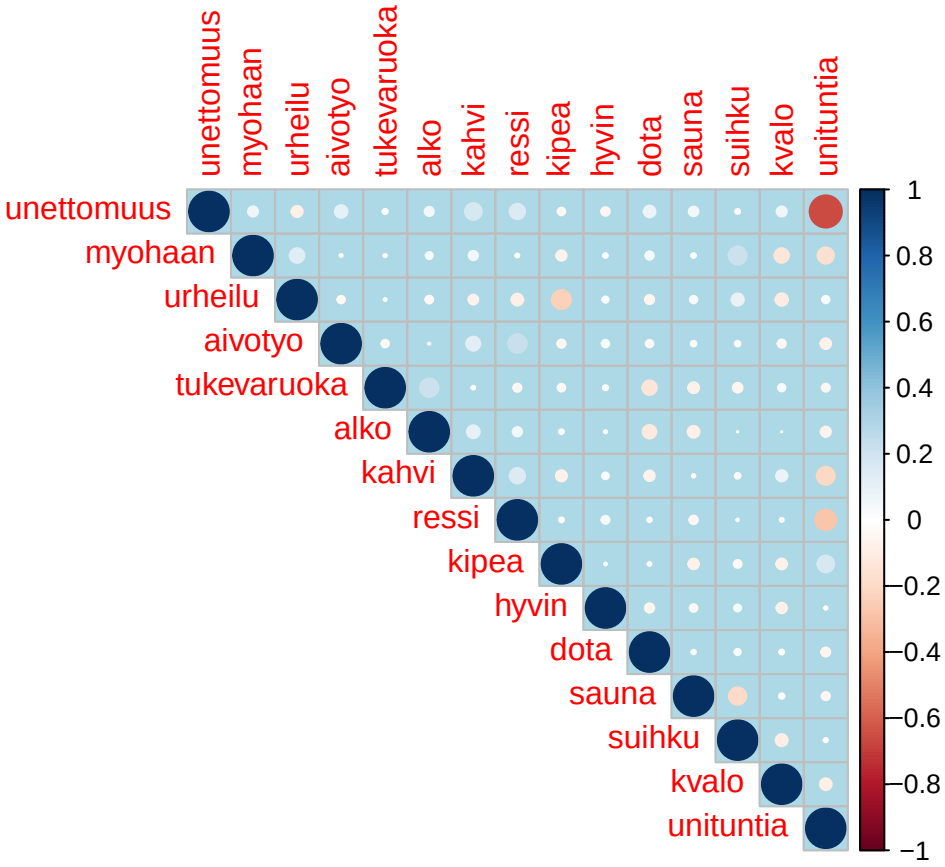
```
##
##                  | df01$kahvi
## df01$unettomuus |        0  |        1  | Row Total |
## ----------------|-----------|-----------|-----------|
##               0 |      463  |       44  |      507  |
##                 |     91.3% |      8.7% |     86.4% |
##                 |     88.5% |     68.8% |           |
## ----------------|-----------|-----------|-----------|
##               1 |       60  |       20  |       80  |
##                 |     75.0% |     25.0% |     13.6% |
##                 |     11.5% |     31.2% |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |      523  |       64  |      587  |
##                 |     89.1% |     10.9% |           |
## ----------------|-----------|-----------|-----------|
##
##
```

Tutkitaan korrelaatioita

Vahvaa negatiivista korrelaatiota muuttujien unettomuus ja unituntia kanssa. Tämä on toki ilmiselvää, koska herätys on samaan aikaan riippumatta siitä saako illalla nukahdettua vai ei. Toisin sanoen, korrelaatio on niin korkea, että voidaan puhua multikollineaarisuudesta. Negatiivista korrelaatiota myös muuttujien kahvi ja unituntia sekä ressi ja unituntia välillä. Positiivista korrelaatiota muuttujien aivotyö ja ressi, unettomuus ja kahvi sekä kipeä ja unituntia välillä.

Nähdään selvästi, että mikäli jokin muuttuja korreloi negatiivisesti unituntia-muuttujan kanssa, se korreloi positiivisesti muuttujan unettomuus kanssa.

```
## corrplot 0.84 loaded
```



Test of independence

Tutkitaan onko unituntia-keskiarvo eri luokissa kahvi=0 ja kahvi=1. Huomataan selvä (ja merkitsevä) ero!

Table 4: Welch Two Sample t-test: `df01$unituntia` by `df01$kahvi` (continued below)

Test statistic	df	P value	Alternative hypothesis
3.86	70.6	0.0002486 * * *	two.sided

mean in group 0	mean in group 1
7.117	6.331

Keskiarvoja

Muuttujana unen määrä tunteina. Havainnot suurimmaksi osaksi välillä 6.6 - 7.8 h. Lauantaina nukkuu yleensä eniten

0%	25%	50%	75%	100%
1.2	6.6	7.2	7.8	10

ke	la	ma	pe	su	ti	to
6.887	7.584	6.763	7.295	6.989	6.692	7.024

Otetaan uusi 012 data talteen (alkaa 2019-01-20)

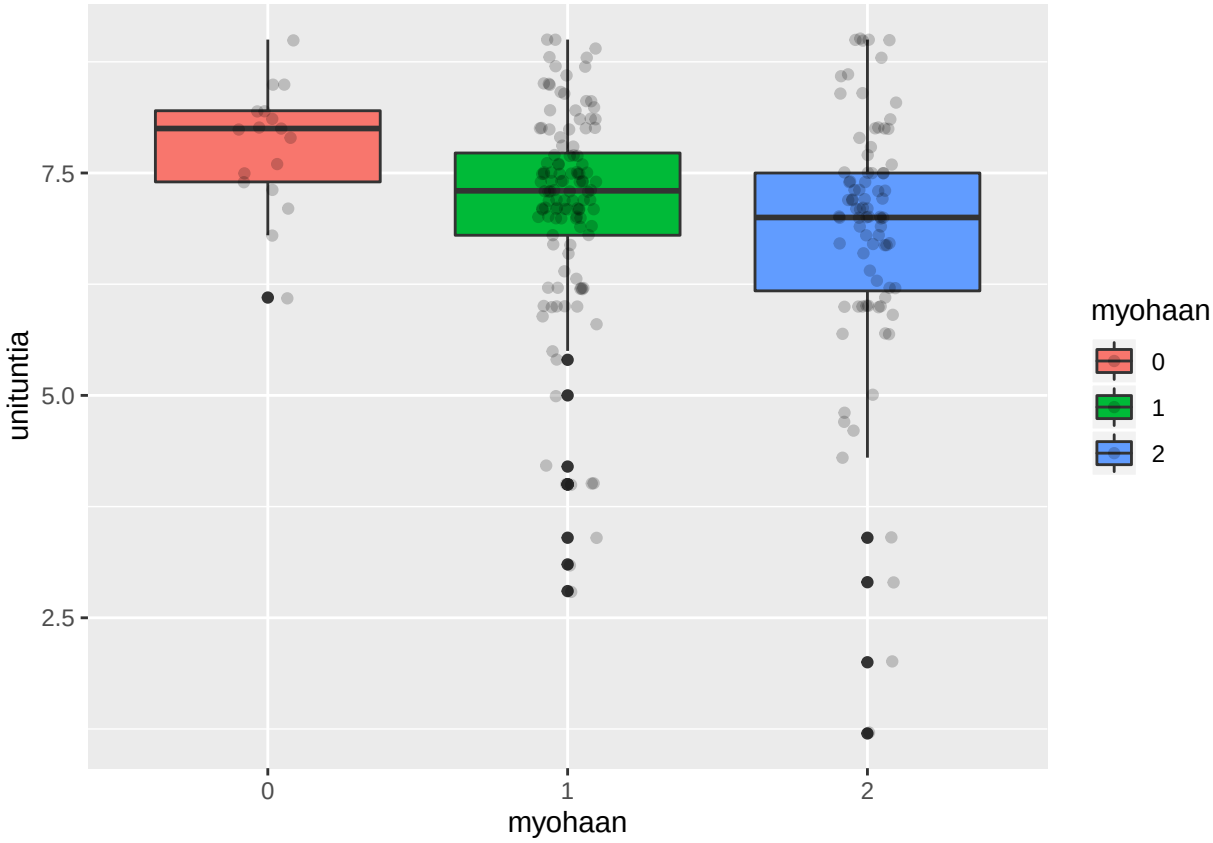
Valitaan data eteenpäin ajasta 2019-01-20 Sitten 012 data factoreiksi.

Anova

Käytetään nyt 0 1 2 dataa. Valitaan muuttujat myöhään (ennen klo23 saa arvon 0, 23-00 arvon 1 ja 00–>arvon 2) ja unituntia. Piirretään ensin boxplot-kuviot. Havaintoja eniten ryhmässä myöhään=1. Unimäärä näyttää vähenevän kun nukkumaanmenoaika myöhästyy.

Tehdään varsinainen anova-taulukko Huomataan, että muuttujien välillä on merkittävä riippuvuus, mutta muuta ei vielä tiedetäkään.

```
ggplot(df012, aes(x=myohaan, y=unituntia, fill=myohaan))+
  geom_boxplot() + geom_jitter(width=0.1,alpha=0.2)
```




```
res.aov<-aov(unituntia ~ myohaan, data=df012)
res.aov %>% summary() %>% pander()
```

Table 8: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
myohaan	2	13.51	6.756	4.131	0.01734
Residuals	218	356.5	1.635	NA	NA

Tukey multiple pairwais comparison

Ero ainoastaan ryhmien myohaan=0 ja myohaan on 2 valillä!

```
#http://www.sthda.com/english/wiki/one-way-anova-test-in-r
TukeyHSD(res.aov) %>% pander()
```

```
## Warning in pander.default(.): No pander.method for "TukeyHSD", reverting to
## default.No pander.method for "multicomp", reverting to default.
```

- **myohaan**:

	diff	lwr	upr	p adj
1-0	-0.6846	-1.467	0.09752	0.09948
2-0	-0.9539	-1.756	-0.1512	0.01512
2-1	-0.2693	-0.6986	0.16	0.3024

```
# plot(res.aov, 2) #outlierit riveillä 151 13 181
```

OLS regressio 01 datalla

Myöhään meno vähentää unen määrää. Samoin ressi, kahvi ja valitettavasti dotakin (tietokonepeli).. Kipeänä nukkuu sitä vastoin hyvin!

Mukautettu R^2 ~0.15, eli ei kovin kehuttava. Kokeillaan stepwise proseduuria alempana.

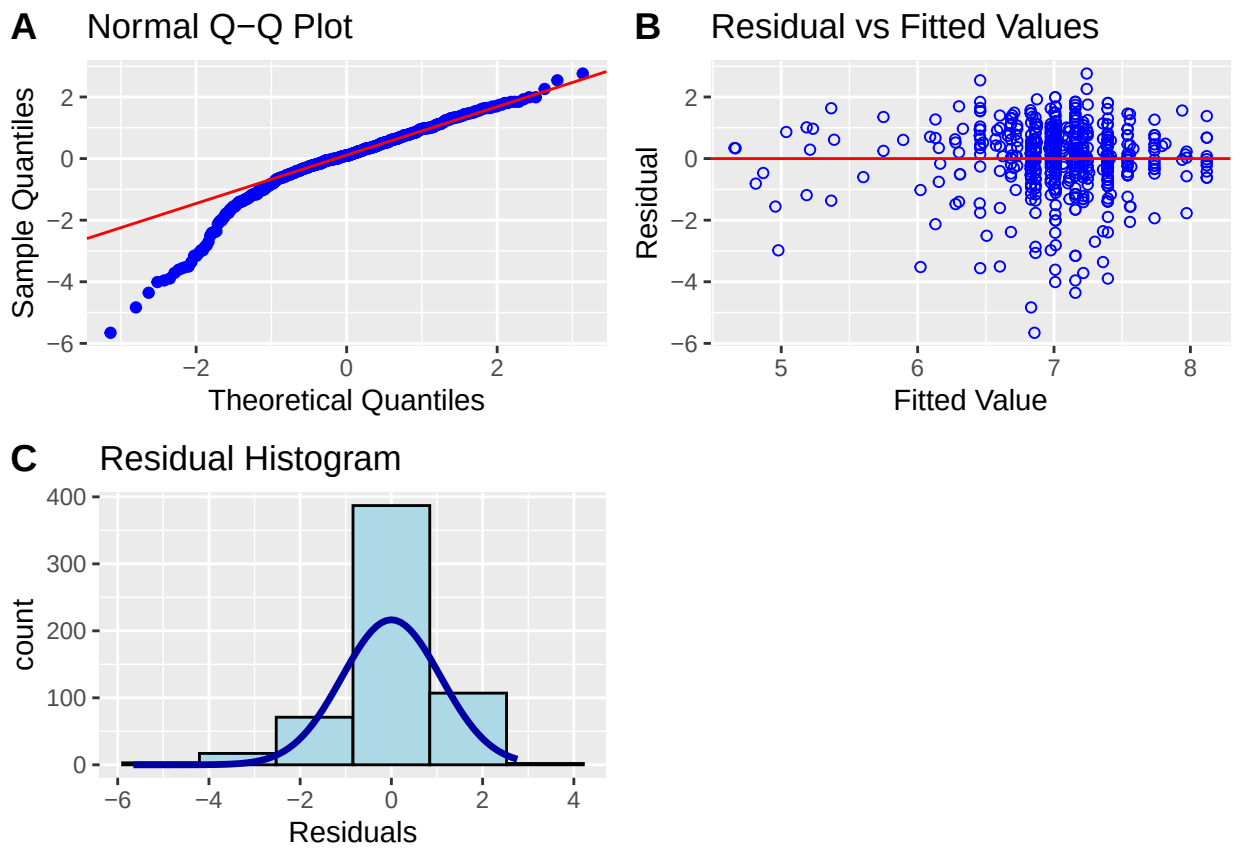
```
##
## Call:
## lm(formula = unituntia ~ myohaan + urheilu + aivotyo + tukevaruoka +
##     alko + kahvi + ressi + kipea + dota + sauna, data = df01)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.657 -0.416  0.105  0.643  2.760
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.3951     0.0992   74.52  < 2e-16 ***
## myohaan      -0.3830     0.0964   -3.97  8.0e-05 ***
## urheilu       0.1445     0.0964    1.50  0.13450
## aivotyo       0.0212     0.3048    0.07  0.94458
## tukevaruoka  -0.2820     0.2089   -1.35  0.17758
## alko         -0.1554     0.1761   -0.88  0.37797
```

```
## kahvi        -0.5535     0.1489   -3.72  0.00022 ***
## ressi        -1.6456     0.2649   -6.21  1.0e-09 ***
## kipea         0.7249     0.1777    4.08  5.1e-05 ***
## dota         -0.1803     0.1073   -1.68  0.09358 .
## sauna        -0.1478     0.1024   -1.44  0.14934
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.09 on 576 degrees of freedom
## Multiple R-squared:  0.165,  Adjusted R-squared:  0.15
## F-statistic: 11.4 on 10 and 576 DF,  p-value: <2e-16
```

Tutkitaan edellä tehdyn mallin oletuksia

Residuaalit eivät näytä normaalijakautuneilta (kuvat A-C). Shapiro-Wilk ja muut normaalisuustestit hylkäävät nollahypoteesin. Toisin sanoen, malli ei näytä täyttävän normaalisuusvaatimuksia. Huom! Ainoastaan error termien pitää olla norm jakautuneita, ei muuttujan unituntia!

```
ggarrange(a,b,c,labels = c("A", "B","C" ),  ncol = 2, nrow = 2)
```




```
ols_test_normality(model)  #Test for detecting violation of normality assumption
```

```
## Warning in ks.test(y, "pnorm", mean(y), sd(y)): ties should not be present
## for the Kolmogorov-Smirnov test
```

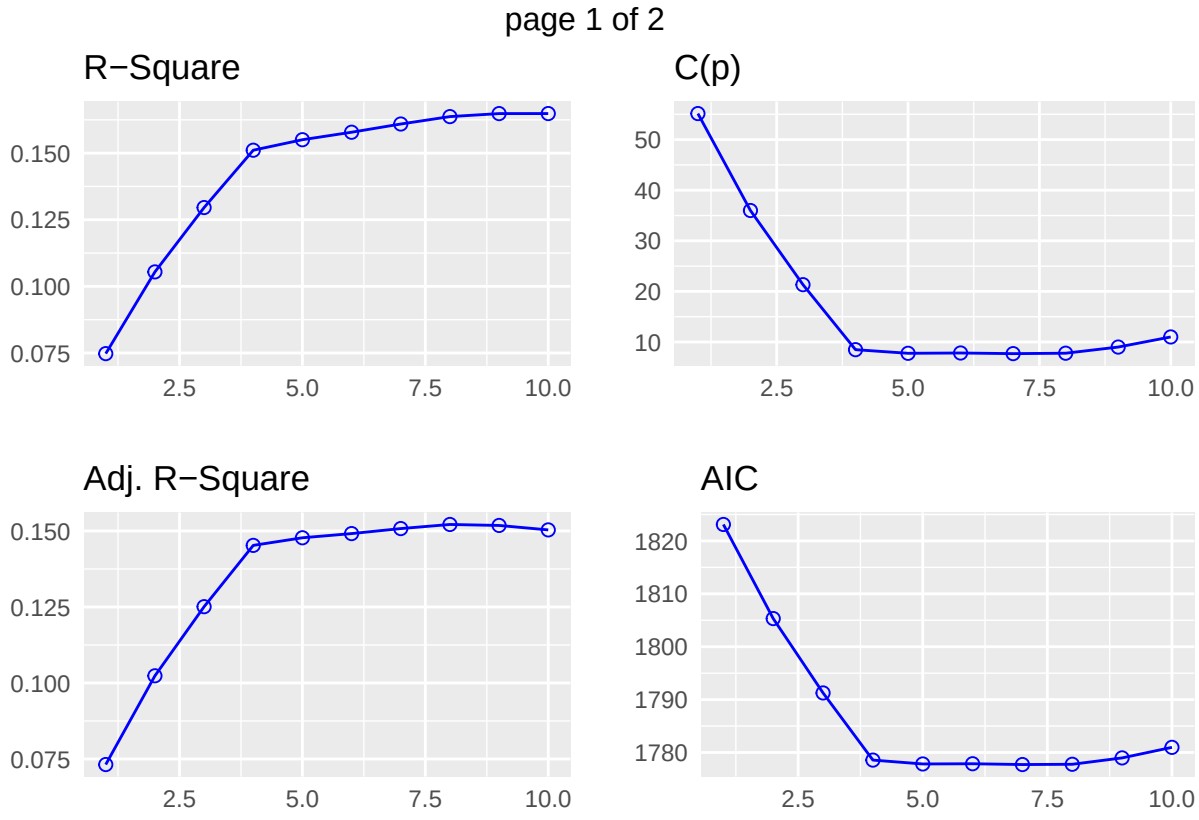
```
## ------------------------------------------------
##        Test             Statistic       pvalue
## ------------------------------------------------
## Shapiro-Wilk              0.9195         0.0000
## Kolmogorov-Smirnov        0.1012         0.0000
## Cramer-von Mises         35.0178         0.0000
## Anderson-Darling          10.052         0.0000
## ------------------------------------------------
```

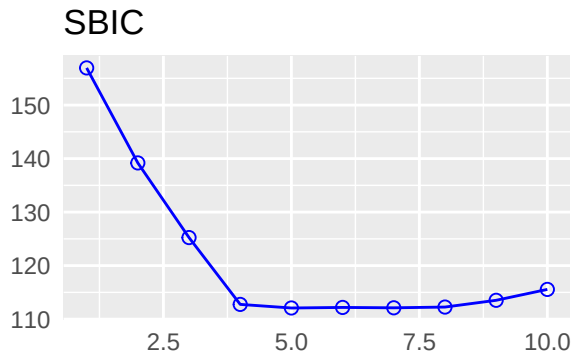
```
#ols_test_breusch_pagan(model)
#samanlaiset kuvat saisi plot(model) komennolla..
#ols_coll_diag(model) #interpretation?
```

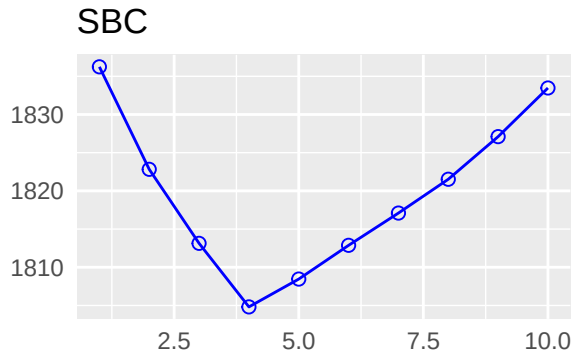
Stepwise regression

Malli ei näytä paranevan juuri lainkaan viidennen selittävän muuttujan jälkeen. Toisin sanoen ainoastaan muuttujat myohaan, urheilu, kahvi, ressi, kipea kannattaa ottaa mukaan.

```
ols_step_best_subset(model) %>% plot()
```







OLS regressio 012 datalla

Loppuviikosta to-la nukkuu parhaiten. Mitä myöhempään menee nukkumaan, sitä vähemmän unta saa. Huom! Kahvia juotu vain todella huonosti nukuttujen öiden jälkeen –> seuraavana yönä nukkuu vaikka aamulla puoli kuppia joisikin.

```
lm(unituntia ~vknpv+myohaan+urheilu+aivotyo+tukevaruoka+alko+kahvi+ressi+kipea+dota+sauna, data=df012) %
```

```
##
## Call:
## lm(formula = unituntia ~ vknpv + myohaan + urheilu + aivotyo +
##     tukevaruoka + alko + kahvi + ressi + kipea + dota + sauna,
##     data = df012)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.904 -0.393  0.210  0.653  2.100
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6964     0.3670   20.97  < 2e-16 ***
## vknpvla       1.1800     0.3357    3.52  0.00054 ***
## vknpvma       0.0668     0.3041    0.22  0.82646
## vknpvpe       0.7885     0.3017    2.61  0.00964 **
```

```
## vknpvsu       0.2568     0.3073    0.84  0.40425
## vknpvti       0.0333     0.3043    0.11  0.91299
## vknpvto       0.6047     0.3070    1.97  0.05021 .
## myohaan1     -0.8290     0.3099   -2.68  0.00809 **
## myohaan2     -1.3900     0.3296   -4.22  3.8e-05 ***
## urheilu1      0.3309     0.1969    1.68  0.09439 .
## urheilu2      0.0370     0.2502    0.15  0.88247
## aivotyo       0.5285     0.6285    0.84  0.40136
## tukevaruoka  -0.3771     0.4700   -0.80  0.42330
## alko         -0.6383     0.3591   -1.78  0.07700 .
## kahvi        -0.3292     0.2612   -1.26  0.20903
## ressi        -1.8070     0.5105   -3.54  0.00050 ***
## kipea         0.6446     0.3863    1.67  0.09673 .
## dota1         0.1343     0.3239    0.41  0.67876
## dota2        -0.1911     0.2285   -0.84  0.40402
## sauna1       -0.1221     0.3307   -0.37  0.71232
## sauna2       -0.3458     0.2245   -1.54  0.12500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.16 on 200 degrees of freedom
## Multiple R-squared:  0.276,  Adjusted R-squared:  0.204
## F-statistic: 3.81 on 20 and 200 DF,  p-value: 4.79e-07
```

Logistinen regressio 01 datalla

Koetetaan selvittää mitkä tekijät lisäävät iltaunettomuuden todennäköisyyttä. Huomataan kahvin, ressin, dotan ja saunan lisäävän iltaunettomuutta. Saunan kohdalla tosin vajaa 9 prosentin riski että ollaan väärässä.. Muut selittävät muuttujat ovat merkityksettömiä.
Ressi ja kahvi kolminkertaistavat unettomuuden todennäköisyyden.

```
library(ISLR) #kannattais kokeilla pakettia  blorr
malli_glm01 <- glm(unettomuus~myohaan+urheilu+aivotyo+tukevaruoka+alko+kahvi+ressi+kipea+dota+sauna, fa
malli_glm01 %>% summary()
```

```
##
## Call:
## glm(formula = unettomuus ~ myohaan + urheilu + aivotyo + tukevaruoka +
##     alko + kahvi + ressi + kipea + dota + sauna, family = binomial(link = "logit"),
##     data = df01)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.412  -0.545  -0.459  -0.372   2.325
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.600      0.306   -8.49   <2e-16 ***
## myohaan        0.401      0.284    1.41   0.1585
## urheilu       -0.434      0.277   -1.57   0.1171
## aivotyo        0.800      0.653    1.23   0.2205
## tukevaruoka    0.498      0.550    0.90   0.3657
```

```
## alko           0.366      0.447    0.82   0.4131
## kahvi          1.064      0.330    3.22   0.0013 **
## ressi          1.204      0.542    2.22   0.0263 *
## kipea         -0.337      0.559   -0.60   0.5465
## dota           0.708      0.280    2.53   0.0114 *
## sauna          0.469      0.273    1.72   0.0857 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 467.44  on 586  degrees of freedom
## Residual deviance: 429.50  on 576  degrees of freedom
## AIC: 451.5
##
## Number of Fisher Scoring iterations: 5
```

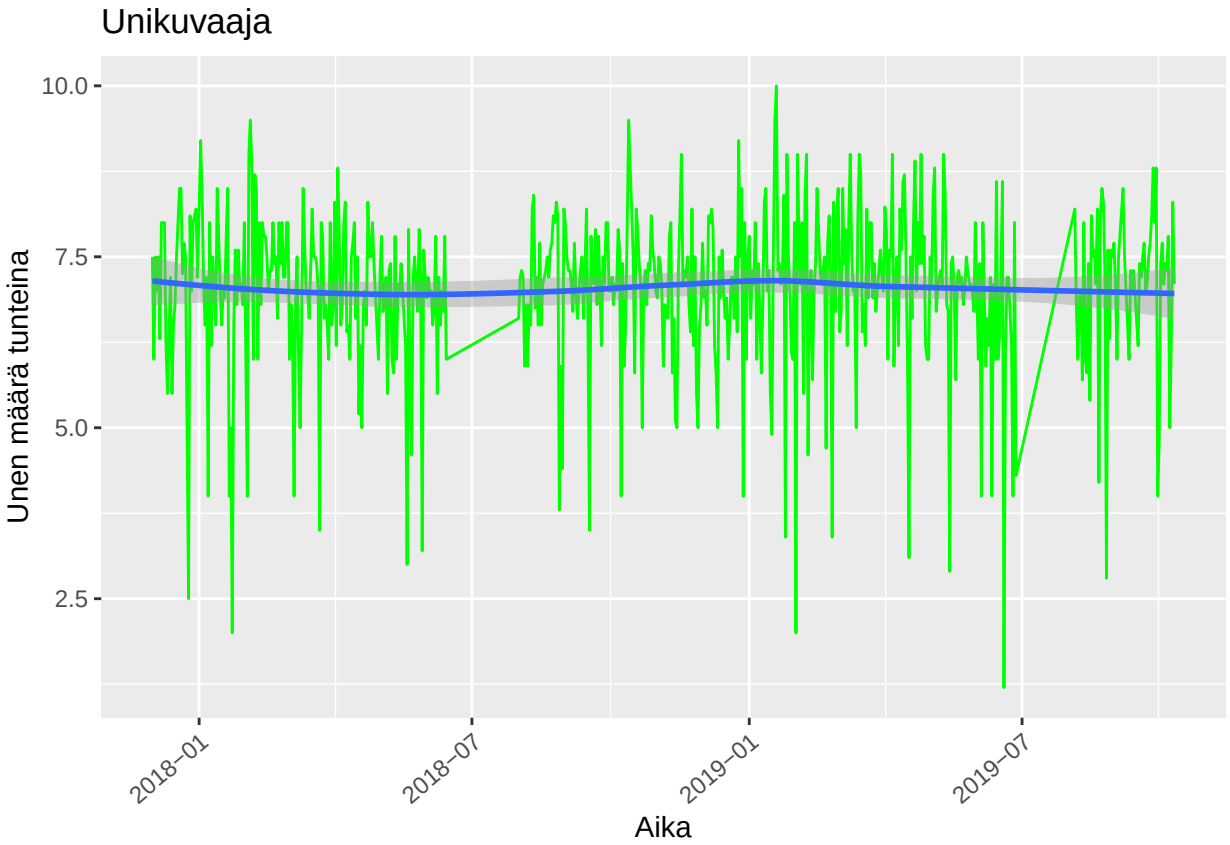
```
exp(malli_glm01$coefficients)
```

```
## (Intercept)     myohaan     urheilu     aivotyo tukevaruoka        alko
##      0.0743      1.4933      0.6480      2.2258      1.6447      1.4413
##       kahvi       ressi       kipea        dota       sauna
##      2.8967      3.3325      0.7137      2.0304      1.5979
```

Piirretään kuvioita

Unen määrä tunteina

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


Unikuvaaja
10.0
7.5
5.0
2.5
Unen määrä tunteina
2018–01
2018–07
2019–01
2019–07
Aika