Ühefaktoriline dispersioonanalüüs
(*One-way ANOVA*)

Vaatame, mida teha, kui soovime prognoosimisel kasutada sellise tunnuse abi, mis polegi pidev.

Mudel 1:

Saak = 3205,5*I(sort==“A“) + 3572,9*I(sort==“B“) + prognoosiviga

Mudel 1 prognoos:
sort A saagikusele: 3205,5
sort B saagikusele: 3572,9

Mudel 2 (samaväärne):

Saak = 3205,5 + 367,4*I(sort==“B“) + prognoosiviga

Mudel 2 prognoos:
sort A saagikusele: 3205,5
sort B saagikusele: 3205,5+367,4 = 3572,9

Faktortunnused

Vahel soovime prognoosida tunnuse *Y* väärtuseid, aga tunnus, mille abil me prognoosime, *X*, pole pidev. Näiteks soovime prognoosida põllult saadavat saaki, teades, kumba sorti – A-d või B-d sellel põllul kasvatati.

Mida saame teha?

```
> by(saak, sort, mean)
INDICES: A
[1] 3205.455
------------------------------------------------------------
INDICES: B
[1] 3572.889
```

Kuidas leida 2. mudeli parameetreid?

Teeme esmalt abitunnuse

$$I_{sortB} = \begin{cases} 0, \text{ kui sort} \neq \text{B} \\ 1, \text{ kui sort} = \text{B} \end{cases}.$$

Kasutame nüüd saadud abitunnust regressioonimudelis argumenttunnusena.

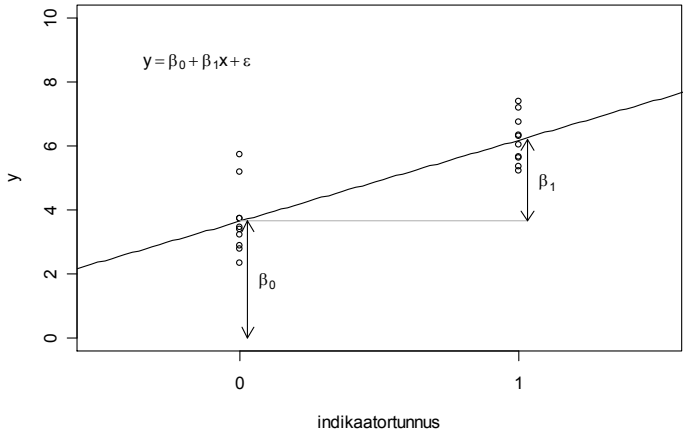
```
> I_sortB=1*(sort=="B")
> lm(saak~I_sortB)

Coefficients:
(Intercept)      I_sortB
     3205.5        367.4
```

Võime lasta ka statistikaprogrammil endal indikaatortunnuse teha:

```
> lm(saak~factor(sort))

Coefficients:
  (Intercept)  factor(sort)B
       3205.5          367.4
```



Millist rühma valida võrdluse aluseks?

Paneme tähele, et vaatlustulemuste kirjeldamise seisukohast on järgmised kolm mudelit täpselt sama head:

Mudel 1:
Saak = 3205*I(sort==“A“) + 3573*I(sort==“B“) + prognoosiviga

Mudel 2 (võrdluse aluseks sort A):
Saak = 3205 + 367*I(sort==“B“) + prognoosiviga

Mudel 3 (võrdluse aluseks sort B):
Saak = 3573 - 367*I(sort==“A“) + prognoosiviga

Mudel 4 (võrdluse aluseks valimi keskmine):
Saak = 3371 - 166*I(sort==“A“) + 202*I(sort==“B“) + prognoosiviga

Mudel 5 (võrdluse aluseks sortide keskmiste saakide keskmine):
Saak = 3389 - 184*I(sort==“A“) + 184*I(sort==“B“) + prognoosiviga

Tähelepanu!

Erinevad programmid valivad võrdluse aluse erinevalt! Kui analüüsime sama andmestikku erinevate statistikaprogrammide abil, võime saada vastuseks vägagi erinevaid numbreid. Tulemuste interpretatsioon jääb aga alati samaks.

SAS

```
proc glm data=saak;
  class sort;
  model saak=sort /solution; run;
```

```
Parameter          Estimate             Error    t Value    Pr > |t|
 Intercept       3572.888889 B     35.72819340     100.00      <.0001
 sort      A     -367.434343 B     48.17588615      -7.63      <.0001
 sort      B        0.000000 B       .                .         .
```

R

```
summary(lm(saak~factor(sort)))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3205.45      32.32  99.187  < 2e-16 ***
factor(sort)B   367.43      48.18   7.627 4.81e-07 ***
```

Testimine – faktortunnusel 2 taset

Kui faktortunnusel on kõigest 2 taset, jõuame samade tulemusteni nii t-testi, ANOVA kui ka regressioonanalüüsi abil:

```
> t.test(SAAK~factor(SORT), var.equal=TRUE)                    Studenti t-test
        Two Sample t-test
data:  SAAK by factor(SORT)
t = -7.6269, df = 18, p-value = 4.805e-07
mean in group A mean in group B
       3205.455        3572.889
> summary(lm(SAAK~factor(SORT)))                ANOVA ehk dispersioonanalüüs
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3205.45      32.32  99.187  < 2e-16 ***
factor(SORT)B   367.43      48.18   7.627 4.81e-07 ***

Residual standard error: 107.2 on 18 degrees of freedom
Multiple R-Squared: 0.7637,     Adjusted R-squared: 0.7506
F-statistic: 58.17 on 1 and 18 DF,  p-value: 4.805e-07
> summary(lm(saak~I_sortB))                                  regressioonanalüüs
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3205.45      32.32  99.187  < 2e-16 ***
I_sortB         367.43      48.18   7.627 4.81e-07 ***
```

Rohkem kui kaks faktortunnuse taset

Kui faktortunnuses on rohkem kui kaks taset (lisaks sortidele A ja B on vaadeldud ka sorte C, D ja E), siis tuleb teha ka rohkem indikaatortunnuseid.

Andmestiku näide (võrreldakse sordiga A):

Saak	sort	I_{sortB}	I_{sortC}	I_{sortD}	I_{sortE}
3450	A	0	0	0	0
3567	A	0	0	0	0
3256	B	1	0	0	0
3345	B	1	0	0	0
3890	C	0	1	0	0
3925	C	0	1	0	0
3300	D	0	0	1	0
3123	D	0	0	1	0
3800	E	0	0	0	1
3850	E	0	0	0	1

Üks vähegi viisakas statistikaprogramm teeb indikaatortunnused muidugi ise ka valmis.

Näide

```
> summary(lm(saak~factor(sort)))

Residuals:
    Min      1Q  Median      3Q     Max
-309.00 -108.12   13.52   90.05  290.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3205.45      43.90  73.025  < 2e-16 ***
factor(sort)B   367.43      65.44   5.615 1.16e-06 ***
factor(sort)C   170.05      67.65   2.514  0.01559 *
factor(sort)D  -216.45      62.08  -3.487  0.00110 **
factor(sort)E    33.00      62.08   0.532  0.59762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 145.6 on 45 degrees of freedom
Multiple R-Squared: 0.6578,     Adjusted R-squared: 0.6274
F-statistic: 21.63 on 4 and 45 DF,  p-value: 5.249e-10
```

Kordajad on tõestatavalt nullist erinevad, sortide B, C ja D keskmine saagikus erineb sordi A keskmisest saagikusest

Sordi E saagikuse keskväärtus võib olla sama, mis sordi A keskmine saagikus

Testimine – kuid me soovisime testida midagi muud?

Esimene küsimus, millele otsime vastust: kas üldse eksisteerib erinevust sortide vahel?

Kui sort A saagikus on $Saak_1$ js sort B saagikust tähistame $Saak_2$ jne, siis meid huvitava hüpoteesi võiks sõnastada järgmiselt:

H_0: E $Saak_1$ = E $Saak_2$ = E $Saak_3$ = … = E $Saak_k$
H_1: H_0 ei kehti (eksisteerivad vähemalt kaks sorti, mille keskmised saagikused pole võrdsed).

Alljärgnevalt vaatame erinevaid võimalusi kontrollida hüpoteese keskväärtuste võrdsuse kohta R-is.

Kolm sarnast testi

```
> summary(lm(SAAK~factor(SORT)))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3205.45      43.21  74.189  < 2e-16 ***
factor(SORT)B    367.43      64.41   5.705 8.55e-07 ***
factor(SORT)C    116.82      61.10   1.912   0.0623 .
factor(SORT)D   -268.73      61.10  -4.398 6.62e-05 ***
factor(SORT)E     36.42      66.59   0.547   0.5871
Residual standard error: 143.3 on 45 degrees of freedom
Multiple R-Squared: 0.694,      Adjusted R-squared: 0.6668
F-statistic: 25.52 on 4 and 45 DF,  p-value: 4.456e-11

> drop1(lm(SAAK~factor(SORT)), test="F")
             Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                     924079     501
factor(SORT)  4   2095865 3019944     552  25.516 4.456e-11 ***

> summary(aov(lm(SAAK~factor(SORT))))
             Df  Sum Sq Mean Sq F value     Pr(>F)
factor(Variety)  4 2095865  523966  25.516 4.456e-11 ***
Residuals       45  924079   20535
```

Kas mudel on hea?

Kas me vajame tunnust „SORT“?

„Traditsiooniline“ ANOVA tabel, võib osutuda eksitavaks mittetasakaaluliste andmestike korral (aga ühefaktorilise dispersioonanalüüsi korral on OK.)

Testimine – üks tähelepanek

```
> m1=lm(Saak~factor(Sort)); summary(m1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     4370.6      152.8  28.607   <2e-16 ***
factor(Sort)B   -202.3      161.0  -1.256    0.212
factor(Sort)C    162.0      162.4   0.997    0.321
factor(Sort)D    103.8      162.9   0.637    0.526
factor(Sort)E   -162.5      160.8  -1.011    0.315

Residual standard error: 264.6 on 98 degrees of freedom
Multiple R-Squared: 0.2693,     Adjusted R-squared: 0.2394
F-statistic: 9.028 on 4 and 98 DF,  p-value: 2.993e-06


> drop1(m1, test="F")
Single term deletions

Model:
Saak ~ factor(Sort)
             Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                    6862583    1154
factor(Sort)  4   2528847 9391430    1178  9.0282 2.993e-06 ***
```