Mitme tunnuse koosanalüüs

Allan Sims

31.01.2025

Mitme tunnuse koosanalüüs

Lineaarne regressioon

Lineaarne regressioon on üks viis, kuidas uurida, kuidas kaks või enam asja on omavahel seotud. Selle abil saame aru saada, kuidas ühe tunnuse muutumine mõjutab teist. Näiteks, kui me teame, kuidas ühe tunnuse väärtus muutub, saame ennustada, mis võib juhtuda teise tunnusega.

Lineaarne regressioon lähtub eeldusest, et kahe asja vaheline suhe on sirgejooneline. See tähendab, et kui üks tunnus muutub, siis teine asi muutub kindlal viisil, mida saab joonistada sirgjoonena. Selle meetodi eesmärk on leida kõige sobivam sirgjoon, mis näitab, kuidas need kaks tunnust on omavahel seotud. Lineaarse regressioonimudeli võib esitada võrrandina:

$$Y = a + bX + \epsilon$$

Kus:

- Y on funktsioontunnus
- X on argumenttunnus
- a on lõikepunkt, mis näitab, kui suur on Y väärtus, kui X väärtus on 0
- b on regresseerimiskordaja, mis näitab, kui palju muutub Y väärtus, kui X väärtus suureneb ühe ühiku võrra
- ϵ on veaväärtus, mis tähistab juhuslikke kõrvalekaldeid mudelist

Lineaarse regressiooni puhul kasutatakse mitmesuguseid meetodeid, et hinnata, kui hästi meie sirgjoon mudelina toimib. Üks selline meetod on vähim ruutude meetod. See tähendab, et me otsime sirgjoont, mis käitub nii, et meie ennustuste ja tegelikkuses mõõdetud väärtuste erinevused oleksid võimalikult väikesed. Lihtsamalt öeldes, me püüame leida sirgjoont, mis sobib kõige paremini meie andmetega, minimeerides vigu ennustustes.

Lineaarse regressiooni usaldusväärsed tulemused sõltuvad teatud eeldustest. Need eeldused aitavad tagada, et meie analüüs peegeldaks tegelikkust võimalikult täpselt. Siin on need eeldused lihtsustatult:

- **Lineaarsus.** Eeldatakse, et kahe tunnuse vaheline suhe on sirgjooneline. See tähendab, et kui üks tunnus muutub, siis teise tunnus vastus on alati sellega proportsioonis, sõltumata olukorrast.
- **Normaaljaotus.** Hälve ehk ennustuste ja tegelike tulemuste erinevused järgivad kindlat mustrit, mis vastab normaaljaotusele.
- **Homoskedastilisus.** Sõltuva tunnuse väärtuste hajuvus on ühtlane kogu sõltumatu tunnuse ulatuses, mis tähendab, et erinevused ei suurene ega vähene sõltumatu tunnuse väärtuse kasvades.
- **Autokorrelatsiooni puudumine.** Vaatluste vead on omavahel sõltumatud, st et ühe vaatluse viga ei mõjuta teise vaatluse viga.
- **Multikollineaarsuse puudumine.** Analüüsitavad tunnused on sõltumatud, st ühe tunnuse väärtus ei sõltu teise tunnuse väärtusest ega ole nendega liiga tihedalt seotud.

Need eeldused aitavad tagada, et lineaarne regressioonimudel oleks usaldusväärne ja et selle abil tehtud järeldused peegeldaksid tegelikke seoseid.

Lineaarne regressioon R keskkonnas

```
# Näide R keskkonnas
# Laadime peatükis kasutatavad paketid
library(dplyr)
library(ggplot2)
library(readxl)
library(car)
library(ggpmisc)
```

Loeme andmed failist naited.xlsx töölehelt "prt_andmed" ja salvestame selle tabelisse:

```
# Näide R keskkonnas
andmed <- read_excel("data/naited.xlsx","prt_andmed") %>% filter(h > 0)
```

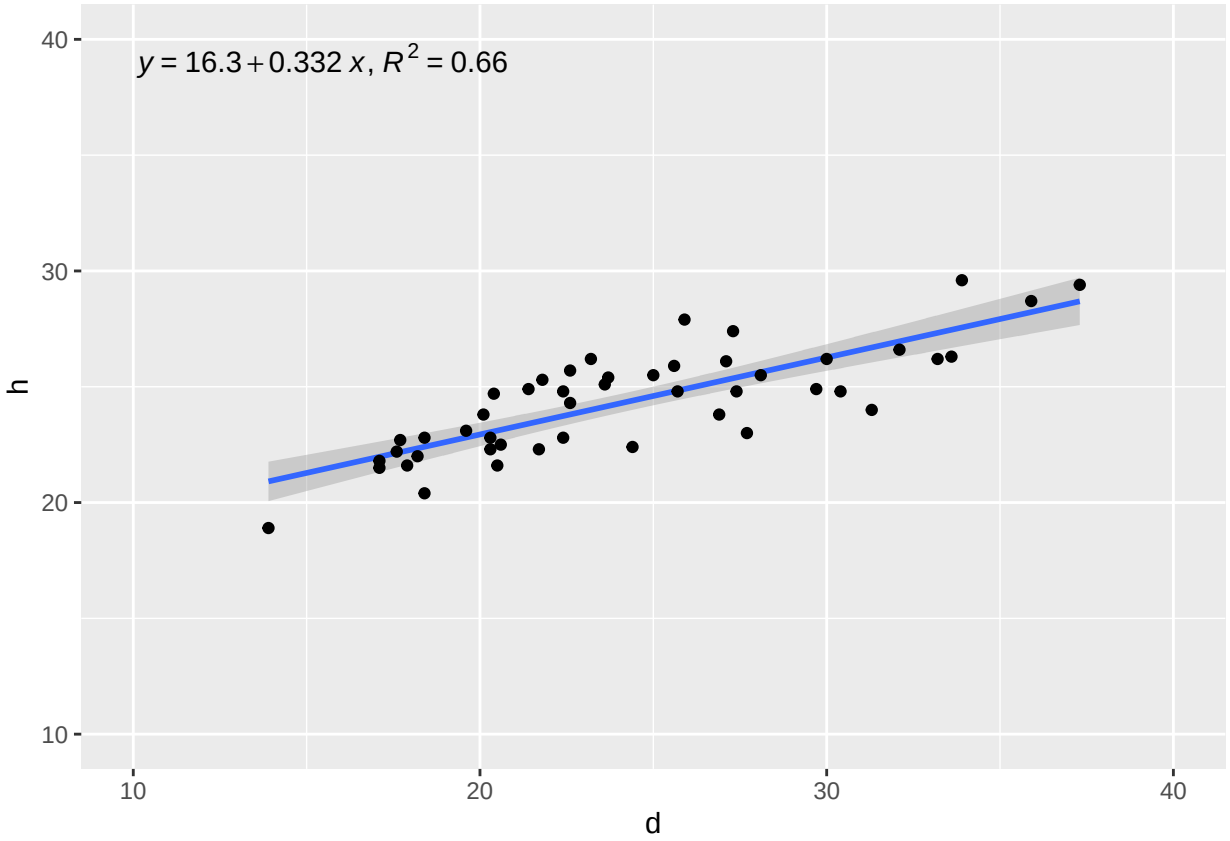
Teeme lineaarse regressiooni tunnuste D ja H vahel:

```
# Näide R keskkonnas
print(lm_model <- lm(h ~ d, data = andmed))
```

```
##
## Call:
## lm(formula = h ~ d, data = andmed)
##
## Coefficients:
## (Intercept)            d
##     16.2950       0.3323
```

Järgnevalt visualiseerime nii andmeid kui ka mudelit funktsiooni `ggplot()` abil.

```
# Näide R keskkonnas
ggplot(andmed, aes(x = d, y = h)) +
  stat_poly_line() +
  stat_poly_eq(use_label(c("eq", "R2"))) +
  geom_point() +
  xlim(c(10, 40)) + ylim(c(10, 40))
```



Järgnevalt käsu selgitus:

- `ggplot(andmed, aes(x = d, y = h))`: See algatab ggplot-objekti, määrates andmestiku (data) ja esteetilised vastavused. See tähendab, et andmestiku veerg `d` vastendatakse x-teljega ja veerg `h` y-teljega.
- `stat_poly_line()`: See lisab graafikule polünoomilise regressioonijoone. Vaikimisi kasutab see lineaarset mudelit (astmega 1), kuid seda saab muuta, kasutades argumenti `formula`. Näiteks `stat_poly_line(formula = y ~ x^2)` lisaks graafikule ruutvõrrandi.
- `stat_poly_eq(use_label(c("eq", "R2")))`: See lisab graafikule polünoomilise regressioonivõrrandi ja R-ruudu väärtuse. `use_label()` funktsiooniga saab määrata, milliseid tunnuseid kuvatakse. Antud juhul kuvatakse võrrand (`eq`) ja R-ruut (`R2`).
- `geom_point()`: See lisab graafikule punktid, mis tähistavad andmepunkte.
- `xlim(c(10, 40))`: See seab x-telje piirid vahemikku 10 kuni 40.
- `ylim(c(10, 40))`: See seab y-telje piirid vahemikku 10 kuni 40.

```r
# Näide R keskkonnas
ggplot(andmed, aes(x = d, y = h)) +
  stat_poly_line(formula = y ~ log(x)) +
  stat_poly_eq(formula = y ~ log(x), use_label(c("eq", "R2"))) +
  geom_point() +
  xlim(c(10, 40)) + ylim(c(10, 40))
```

R-is on võimalik lihtalt lisada mudelisse täiendavad tunnuseid, mis arvutakse otse mudeli arvutamise käigus ning seetõttu ei ole vajalik eraldi eelnevalt täiendavat tunnust välja arvutada. Funktsiooniga `I()` saab luua uusi tunnuseid olemasolevate tunnuste põhjal. Uus tunnus arvutatakse iga andmerea jaoks mudeli arvutamise käigus.

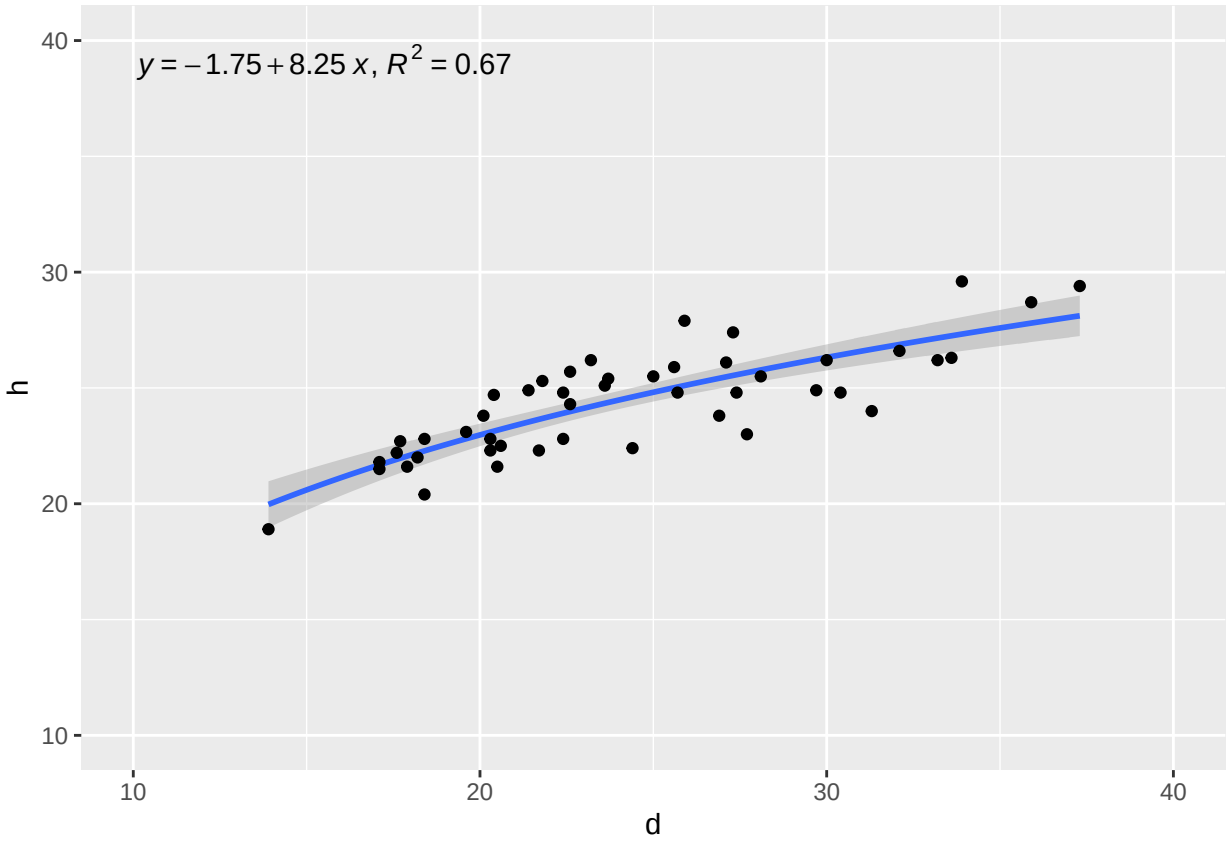


Figure 1: Lineaarne regressioon argumenttunnuse logistilise teisendusega

Järgnevas näites on muudetud mudelit: `lm(h ~ log(d)+I(pl == "MA"), data = andmed)`

- **`log(d)`**: Funktsioon `log()` arvutab `d` naturaallogaritmi, mis aitab lineariseerida seost diameetri ja kõrguse vahel.
- **`I(pl == "MA")`**: See on teine sõltumatu tunnus, mis luuakse jooksvalt funktsiooni `I()` abil. See on loogiline tunnus, mis on `TRUE`, kui tunnus `pl` väärtus on "MA" , ja `FALSE` vastasel juhul. Sisuliselt lisab see mudelisse teavet selle kohta, kas puu on mänd. Arvutustes kasutatakse antud tunnust vastavalt 1 või 0 väärtusena.

```
# Näide R keskkonnas
# täiendatud argumenttunnustega regressioon
summary(lm_model_pl <- lm(h ~ log(d)+I(pl == "MA"), data = andmed))
```

```
##
## Call:
## lm(formula = h ~ log(d) + I(pl == "MA"), data = andmed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1686 -0.8569  0.0484  0.6647  3.4213
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -3.5829     2.7247  -1.315   0.1953
## log(d)              8.9235     0.8718  10.236 3.24e-13 ***
## I(pl == "MA")TRUE  -0.9776     0.4300  -2.274   0.0279 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.268 on 44 degrees of freedom
## Multiple R-squared:  0.7074, Adjusted R-squared:  0.6941
## F-statistic: 53.19 on 2 and 44 DF,  p-value: 1.81e-12
```

Antud juhul on näha, et mänd on teistest statistiliselt oluliselt erinev.

Lineaarne regressioon MS Exceli keskkonnas

MS Exceli keskkonnas saab lineaarse regressiooni parameetreid arvutada mitmel viisil.

- Kui argumenttunnuseid on vaid üks, siis saab kasutada funktsioone `intercept()` ja `slope()`.
- Kui argumenttunnuseid on mitu (või ka ainult üks), siis saab kasutada funktsiooni `linest()`, mis annab tulemuseks juba ka täpsemad mudeli parameetrite hinnangud.
- Ühe ja mitme argumenttunnuse puhul saab kasutada *Analysis Toolpak* protseduuri *Regression*.

Avame faili naited.xlsx töölehel "linreg", kus on samad andmed nagu "prt_andmed", aga sellel on juba välja filtreeritud ainult mudelpuude andmed. MS Exceli keskkonnas on üldiselt vajalik filtreeritud andmed viia eraldi tabelisse, et regressiooni käigus muid andmeid ei kasutataks.

Analysis Toolpak abil regressioonanalüüsi tegemiseks tuleb menüüs "Andmed" valida "Data Analysis", mis järel avaneb analüüsivahendite aken.

Kordame samasugust analüüsi ka MS Exceli keskkonnas nagu R-is, arvutame diameetrist naturaallogaritmi funktsiooniga `ln()` ning kasutame `OnMänd` tunnust.

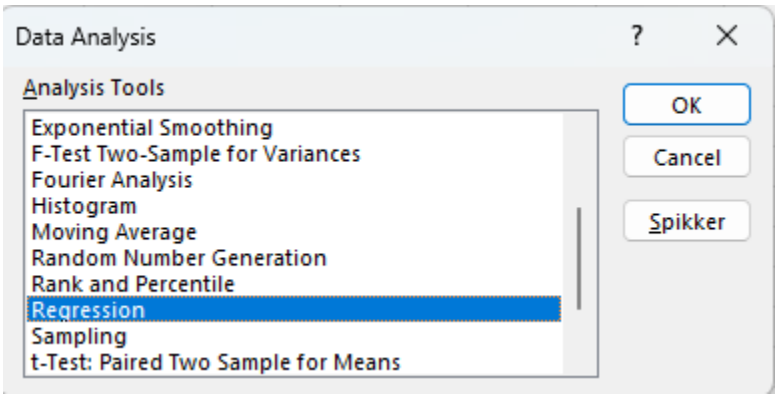
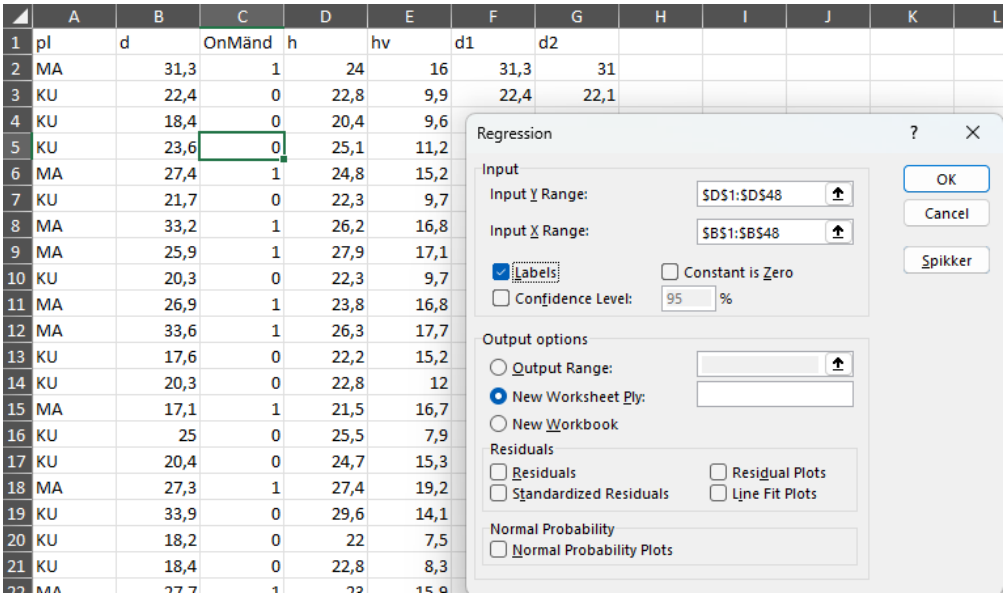
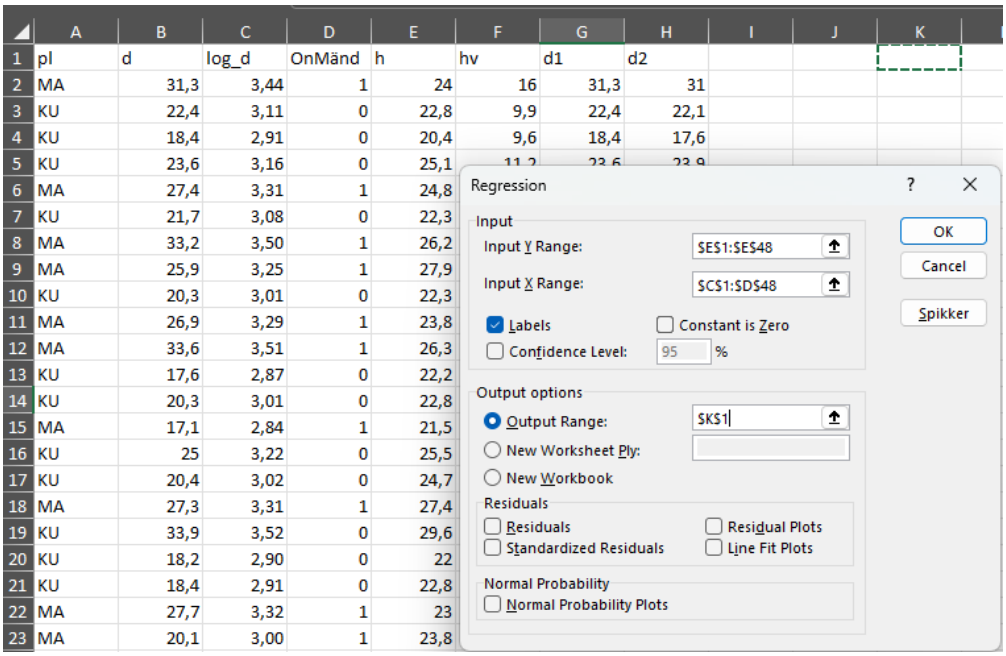


Figure 2: Data Analysis aken MS Excelis.



	A	B	C	D	E	F	G
1	pl	d	OnMänd	h	hv	d1	d2
2	MA	31,3	1	24	16	31,3	31
3	KU	22,4	0	22,8	9,9	22,4	22,1
4	KU	18,4	0	20,4	9,6		
5	KU	23,6	0	25,1	11,2		
6	MA	27,4	1	24,8	15,2		
7	KU	21,7	0	22,3	9,7		
8	MA	33,2	1	26,2	16,8		
9	MA	25,9	1	27,9	17,1		
10	KU	20,3	0	22,3	9,7		
11	MA	26,9	1	23,8	16,8		
12	MA	33,6	1	26,3	17,7		
13	KU	17,6	0	22,2	15,2		
14	KU	20,3	0	22,8	12		
15	MA	17,1	1	21,5	16,7		
16	KU	25	0	25,5	7,9		
17	KU	20,4	0	24,7	15,3		
18	MA	27,3	1	27,4	19,2		
19	KU	33,9	0	29,6	14,1		
20	KU	18,2	0	22	7,5		
21	KU	18,4	0	22,8	8,3		

Figure 3: Regressioonanalüüsi aken MS Excelis.



	A	B	C	D	E	F	G	H
1	pl	d	log_d	OnMänd	h	hv	d1	d2
2	MA	31,3	3,44	1	24	16	31,3	31
3	KU	22,4	3,11	0	22,8	9,9	22,4	22,1
4	KU	18,4	2,91	0	20,4	9,6	18,4	17,6
5	KU	23,6	3,16	0	25,1	11,2	23,6	23,9
6	MA	27,4	3,31	1	24,8			
7	KU	21,7	3,08	0	22,3			
8	MA	33,2	3,50	1	26,2			
9	MA	25,9	3,25	1	27,9			
10	KU	20,3	3,01	0	22,3			
11	MA	26,9	3,29	1	23,8			
12	MA	33,6	3,51	1	26,3			
13	KU	17,6	2,87	0	22,2			
14	KU	20,3	3,01	0	22,8			
15	MA	17,1	2,84	1	21,5			
16	KU	25	3,22	0	25,5			
17	KU	20,4	3,02	0	24,7			
18	MA	27,3	3,31	1	27,4			
19	KU	33,9	3,52	0	29,6			
20	KU	18,2	2,90	0	22			
21	KU	18,4	2,91	0	22,8			
22	MA	27,7	3,32	1	23			
23	MA	20,1	3,00	1	23,8			

Figure 4: Regressioonanalüüs kahe tunnuse alusel.

Table: Lineaarse regressiooni tulemus MS Excelis.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,841
R Square	0,707
Adjusted R Square	0,694
Standard Error	1,268
Observations	47

ANOVA

	df	**SS**	**MS**	**F**	**Significance F**
Regression	2	171,139	85,570	53,191	0,000
Residual	44	70,783	1,609		
Total	46	241,922			

	Coefficients	**Standard Error**	**t Stat**	**P-value**
Intercept	-3,5829	2,7247	-1,3150	0,1953
log_d	8,9235	0,8718	10,2358	0,0000
OnMänd	-0,9776	0,4300	-2,2737	0,0279

Regressioonanalüüsi tulemuste tõlgendamine

Lineaarse regressioonimudeli kokkuvõte sisaldab mitmeid olulisi statistilisi näitajaid, mis aitavad hinnata mudeli sobivust andmetele.

Determinatsioonikordaja, tuntud ka kui R^2 (R-ruut), on statistiline mõõdik, mida kasutatakse lineaarse regressiooni kontekstis, et hinnata, kui suures ulatuses argumenttunnuse variatsioon on seletatav sõltumatute tunnuste abil mudelis. R^2 väärtus varieerub vahemikus 0 kuni 1, kus 0 tähendab, et mudel ei seleta argumenttunnuse variatsiooni üldse, samas kui 1 tähendab, et mudel seletab argumenttunnuse variatsiooni täielikult.

R^2 arvutamiseks kasutatakse järgmist valemit:

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

kus:

- $\mathrm{SS}_{\mathrm{res}}$ on jääkide ruutude summa (residual sum of squares), mis mõõdab mudeli poolt seletamata variatsiooni.
- $\mathrm{SS}_{\mathrm{tot}}$ on kogu variatsiooni ruutude summa (total sum of squares), mis mõõdab argumenttunnuse variatsiooni võrreldes tema keskmisega.

Kõrge R^2 väärtus (lähedane 1-le) näitab, et mudeli poolt seletatav variatsioon on suur, mis tähendab, et mudel sobib andmetele hästi. See viitab sellele, et suur osa argumenttunnuse variatsioonist on seletatav mudeli sõltumatute tunnustega.

Madal R^2 väärtus (lähedane 0-le) näitab, et mudel seletab argumenttunnuse variatsioonist väikese osa, mis tähendab, et mudeli sobivus andmetele on kehv. See võib viidata sellele, et mudelis kasutatavad sõltumatud tunnused ei ole sõltuva tunnuse variatsiooni seletamisel tõhusad.

Oluline on mõista, et kuigi R^2 on kasulik mõõdik mudeli sobivuse hindamisel, ei tähenda kõrge R^2 automaatselt, et mudel on õige või, et argumenttunnused põhjustavad funktsioontunnuse käitumist. Samuti ei pruugi madal R^2 alati tähendada mudeli ebatõhusust, eriti keerukamate või mittelineaarsete suhete puhul. Lisaks tuleks mudeli hindamisel arvesse võtta ka muid tegureid, nagu mudeli tõlgendatavus, argumenttunnuste olulisus ja korrigeeritud R^2.

Lineaarse regressioonimudeli **nihutamata hinnang**, tuntud ka kui korrigeeritud R^2 (inglise keeles "adjusted R^2"), on statistiline mõõdik, mida kasutatakse mudeli seletusvõime hindamiseks, võttes arvesse mudeli sõltumatute tunnuste arvu. Erinevalt lihtsast R^2-st, mis mõõdab, kui suur osa argumenttunnuse variatsioonist on seletatav mudeli abil, arvestab korrigeeritud R^2 mudeli keerukust, arvestab mudeli liigsete sõltumatute tunnuste lisamise eest. See on oluline, kuna mudelisse liiga palju tunnusid lisades võib R^2 väärtus eksitavalt tõusta, viidates paremale mudeli sobivusele, kuigi tegelikult võivad lisatud tunnused olla statistiliselt mitteolulised.

Korrigeeritud R^2 arvutatakse järgmise valemi järgi:

$$R^2_{\text{adj}} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

kus:

- R^2 on mudeli lihtne determinatsioonikoefitsient,
- n on vaatluste arv,
- k on sõltumatute tunnuste arv mudelis.

Valemist nähtub, et R^2_{adj} väheneb, kui lisatakse tunnusid, mis ei paranda mudeli seletusvõimet proportsionaalselt nende arvuga. Seega, erinevalt lihtsast R^2-st, pakub korrigeeritud R^2 ausamat hinnangut mudeli sobivusele, eriti kui võrrelda mudeleid, mis sisaldavad erinevat arvu sõltumatuid tunnusid. Korrigeeritud R^2 on eriti kasulik mudelite võrdlemisel ja mudeli valimisel, aidates vältida ülemudelitamist ja eelistades mudelit, mis annab parima tasakaalu seletusvõime ja keerukuse vahel.

- P-väärtused näitavad, kas argumenttunnus on statistiliselt oluline mudelis. Kui P-väärtus on väiksem kui 0,05, siis võib järeldada, et argumenttunnus mõjutab oluliselt sõltuvat tunnust.
- Regressioonikordaja väärtused annavad teavet sõltumatute tunnuste mõju kohta sõltuvale tunnusle.

Kokkuvõttes aitab väljund mõista, kuidas mudel sobitub andmetega ja millist infot see pakub funktsioon- ja argumenttunnuste seoste kohta.

Homoskedastilisuse testimine

Statistilise regressioonimudeli kontekstis viitab homoskedastiivsus olukorrale, kus jääkliikmete (ehk mudeli veatermini) dispersioon on konstantne sõltumata seletavate muutujate väärtustest. See tähendab, et mudeli ennustusvead on ühtlase hajuvusega kogu vaadeldava andmevahemiku ulatuses, ega ole süstemaatiliselt seotud ennustatud väärtustega. Homoskedastiivsus on kriitiline eeldus mitmetes statistilistes meetodites, eriti vähimruutude meetodil baseeruva lineaarse regressiooni puhul. Selle eelduse rikkumine, tuntud kui heteroskedastiivsus, võib moonutada mudeli parameetrite hinnanguid, viies ebatäpsete usaldusvahemike ja potentsiaalselt vigaste järeldusteni hüpoteeside testimisel.

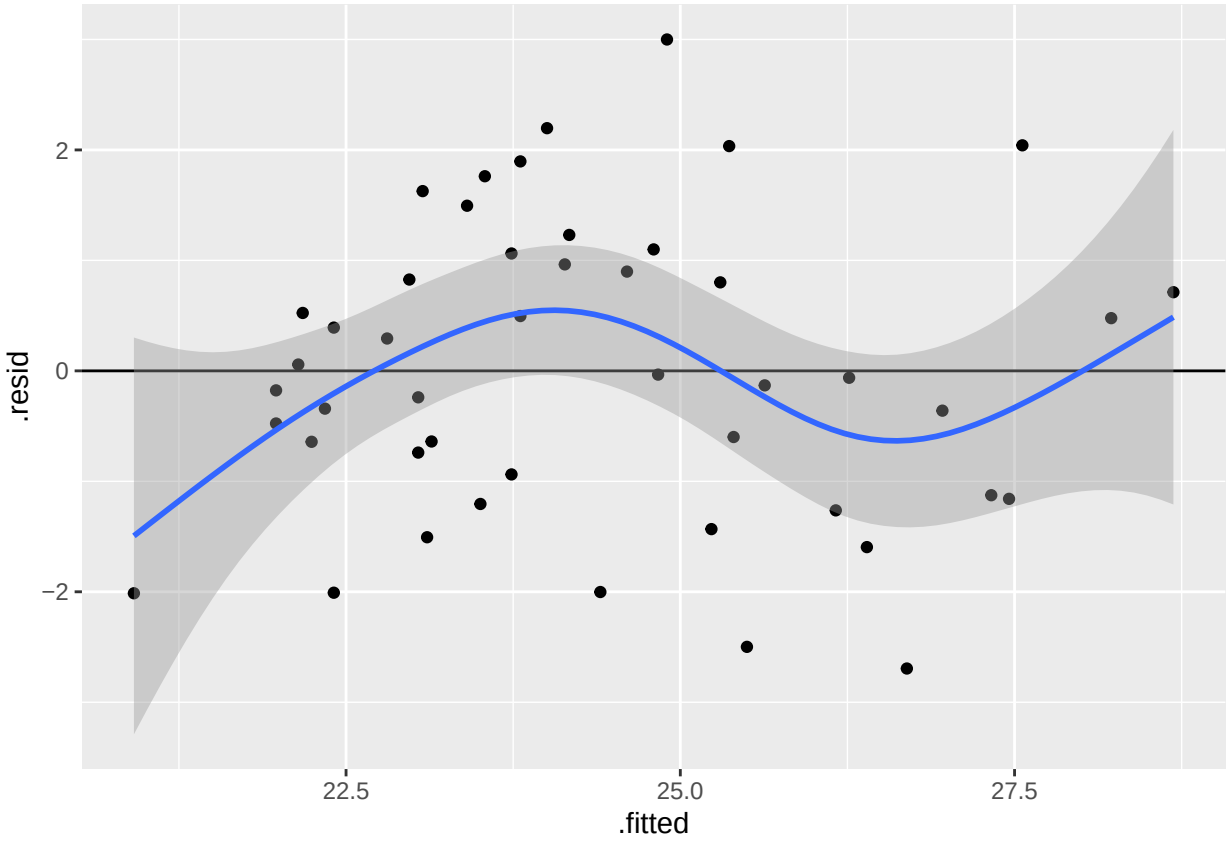
On mitmeid viise, kuidas testida, kas andmetes esineb homoskedastiivsust:

1. **Visuaalne analüüs:** Lihtsaim viis on vaadata jääkliikmete graafikut seletavate muutujate või ennustatud väärtuste suhtes. Kui jääkliikmed on ühtlaselt hajunud nulljoone ümber, on homoskedastiivsus tõenäoline. Kui tekib mingi muster (näiteks koonusekujuline hajumine), on tõenäoline heteroskedastiivsus.
2. **Formaliseeritud testid:** On olemas mitmed statistilised testid, näiteks Breusch-Pagani test ja White'i test, mis testivad heteroskedastiivsust formaalselt. Need testid kontrollivad, kas jääkliikmete dispersioon on seotud seletavate muutujatega.

3. **Robustsed meetodid:** Kui andmetes esineb heteroskedastiivsust, võib kasutada robustseid regressioonimeetodeid, mis on vähem tundlikud heteroskedastiivsuse suhtes.

```r
# Näide R keskkonnas

ggplot(lm_model, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)+
  geom_smooth(method = "gam", formula = y ~ s(x))  # Lisab sujuva joone GAM mudeli abil
```



Koodi selgitus:

`ggplot(lm_model, aes(x = .fitted, y = .resid))`

- See käsk loob aluse graafikule ggplot2 paketiga.
- `lm_model` määrab andmeallika, milleks on lineaarne mudel.
- `aes(x = .fitted, y = .resid)` määrab x-teljele mudeli sobitatud väärtused (.fitted) ja y-teljele jääkväärtused (.resid).

`geom_point()`

- See lisab graafikule punktid, mis esindavad iga andmepunkti jääki ja sobitatud väärtust.

`geom_hline(yintercept = 0)`

- See lisab horisontaalse joone y-telje väärtusel 0. See joon aitab visuaalselt hinnata jääkide jaotust - ideaalses mudelis peaksid punktid jaotuma ühtlaselt joone ümber.

`geom_smooth(method = "gam", formula = y ~ s(x))`

- See lisab graafikule sujuva joone, mis on arvutatud GAM mudeli (Generalized Additive Model) abil.

- `method = "gam"` määrab joone arvutamise meetodiks GAM.
- `formula = y ~ s(x)` määrab mudeli valemi, kus y (jääkväärtused) on seotud x-i (sobitatud väärtused) sujuva funktsiooniga s(x).

Lineaarse mudeli homoskedastilisuse testimiseks saab kasutada Breush-Pagan või White'i testi. Need testid hindavad, kas jääkide varieeruvus on seotud sõltumatute tunnustega ning kas see varieeruvus on ühtlane.

```
# Näide R keskkonnas
lmtest::bptest(lm_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_model
## BP = 0.87129, df = 1, p-value = 0.3506
```

Paketi `lmtest` funktsioon `bptest()` testib regressioonimudeli jääkide heteroskedastsuse olemasolu. See tähendab, et jääkide varieeruvus ei ole konstantne üle kõigi sõltumatute tunnuste väärtuste ning võib tekitada probleeme statistiliste hüpoteeside testimisel.

Testi tulemusi võib lugeda järgmiselt:

- Nullhüpotees (H0): jääkide varieeruvus on konstantne (st heteroskedastsust ei esine).
- Alternatiivhüpotees (H1): jääkide varieeruvus ei ole konstantne (st heteroskedastsus esineb).

Testist saadav p-väärtus aitab hinnata, kas on piisavalt tõendeid, et lükata ümber nullhüpotees heteroskedastsuse puudumise kohta. Väike p-väärtus (<0,05) viitab sellele, et saame nullhüpoteesi ümber lükata ja järeldada, et heteroskedastsust esineb.

Jääkide normaalsuse test

Üks oluline eeldus, mida lineaarses regressioonis kasutatakse, on see, et mudeli jäägid vastaksid normaaljaotusele. Normaaljaotus tähendab, et andmete jaotus on sümmeetriline ning keskmise jaotusega. Kui jäägid on normaaljaotusele lähedased, siis on regressioonimudeli ennustused usaldusväärsemad ja täpsemad. Paljud statistilised testid eeldavad, et andmed vastavad normaaljaotusele. Kui jäägid ei vasta normaaljaotusele, võivad testide tulemused olla ebatäpsed või valed. Normaaljaotuse eeldustele mitte vastamine võib viidata sellele, et mudel ei sobi antud andmetele hästi ning on vajalik leida alternatiivseid mudeleid.

Normaliseeritud jääkide graafik on üks viis, kuidas hinnata, kas jäägid vastavad normaaljaotusele. See graafik aitab visualiseerida, kas jäägid on juhuslikult hajutatud ümber nulli ja kas nende jaotus on ligilähedane normaaljaotusele.

Samuti saab vaadata jääkide jagunemist tihedusfunktsiooni abil koostatud graafikul.

```
# Näide R keskkonnas

ggplot(lm_model, aes(x = .resid)) +
  geom_density() +
  geom_vline(
    xintercept = mean(lm_model$residuals),
    color = "blue",
    lwd = 1
  )
```

Sellise protseduuri abil saad hinnata, kui hästi sobib loodud lineaarne regressioonimudel andmetele ning kas jääkide jaotus vastab normaalsuse eeldusele.

Järgnevalt vaatame mudeli ANOVA funktsiooniga `anova()` ja parameetrite usalduspiire funktsiooniga `confint()`.

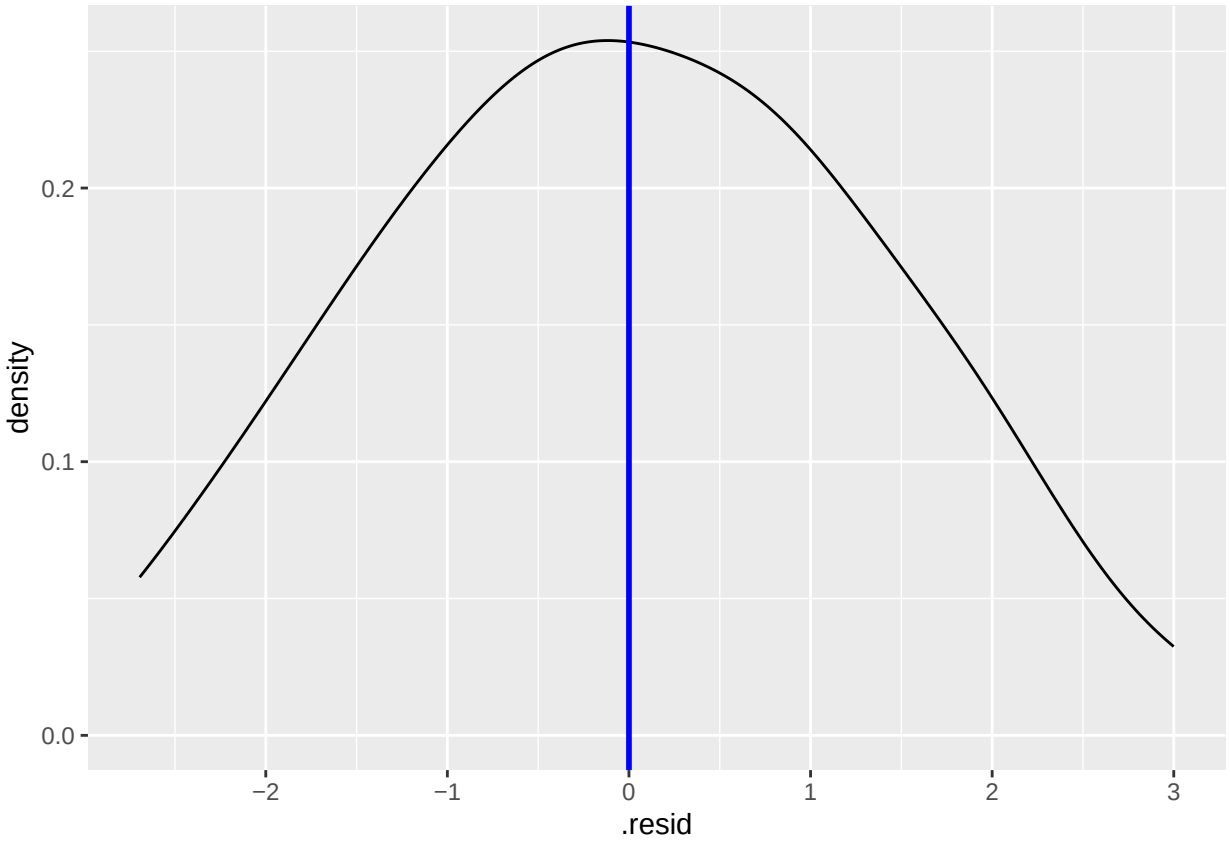


Figure 5: Jääkide tihedusfunktsioon

```r
# Näide R keskkonnas
# ANOVA
anova(lm_model)
```

```
## Analysis of Variance Table
##
## Response: h
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## d          1 158.487 158.487  85.479 5.723e-12 ***
## Residuals 45  83.435   1.854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
confint(lm_model)
```

```
##                   2.5 %     97.5 %
## (Intercept) 14.4914076 18.0986396
## d            0.2598733  0.4046344
```

Multikollineaarsus

Lineaarses regressioonis tähistab multikollineaarsus olukorda, kus kaks või enam sõltumatut tunnust (selgitajat) on omavahel tugevalt seotud või korreleerunud. See tähendab, et ühe selgitaja väärtusi saab suure täpsusega ennustada teiste selgitajate väärtuste põhjal. Multikollineaarsus võib olla probleemiks, sest:

Mõjutab koefitsientide hinnangute täpsust: multikollineaarsus suurendab koefitsientide hinnangute varieeruvust, mis muudab need ebastabiilseks. Väiksed muudatused andmestikus võivad tuua kaasa suuri muutusi koefitsientide hinnangutes, muutes mudeli interpretatsiooni keeruliseks.

Raskendab tunnuste tähtsuse hindamist: kõrge multikollineaarsuse korral võib olla raske kindlaks teha, milline argumenttunnus mõjutab sõltuvat tunnust, sest statistilised testid ei pruugi olla usaldusväärsed. See võib viia oluliste tunnuste ebaõiglase kõrvalejätmiseni või ebaoluliste tunnuste mudelisse lisamiseni.

Vähendab mudeli interpretatsiooni selgust: kui argumenttunnused on omavahel tugevalt seotud, muutub mudeli interpretatsioon keerulisemaks, kuna keeruline on eristada iga tunnus unikaalset mõju argumenttunnusele .

Multikollineaarsuse tuvastamiseks lineaarses regressioonis kasutatakse tavaliselt järgmisi meetodeid:

- **Korrelatsioonimatriks**: Vaadeldakse sõltumatute tunnuste vahelisi korrelatsioonikoefitsiente. Kõrge korrelatsioon näitab võimalikku multikollineaarsust.
- **Variance Inflation Factor (VIF)**: VIF mõõdab, kui palju tunnus variatiivsus suureneb multikollineaarsuse tõttu. Tavaliselt peetakse VIF väärtust üle 5 või 10 viitavaks tugevale multikollineaarsusele.
- **Tolerants**: Tolerants on VIF-i pöördväärtus ja näitab, kui suur osa tunnus variatsioonist ei ole seletatav teiste mudeli tunnustega. Madal tolerants viitab multikollineaarsusele.

Multikollineaarsuse lahendamiseks võib kasutada mitmeid lähenemisviise, näiteks tunnuste eemaldamist mudelist, tunnuste kombineerimist või andmete transformeerimist. Samuti võib kasulikuks osutuda regulaarimismeetodite, nagu ridge regressioon või lasso regressioon, kasutamine, mis aitavad vähendada koefitsientide varieeruvust, karistades suuri koefitsiente.

Vaatame järgmist näidet, kus puude kõrgust ennustame mõlema diameetri alusel. Kui varasemalt oli näha, et üks diameeter on oluline, siis antud mudelid ei ole enam kumbki oluline, sest nende tunnuse vahel on korrelatsioon.

```r
# Näide R keskkonnas
summary(lm.col <- lm(h ~ d1 + d2, data = andmed))
```

```
##
```

```
## Call:
## lm(formula = h ~ d1 + d2, data = andmed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7091 -0.8934 -0.1999  0.7867  3.1426
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.13901    0.89024  18.129   <2e-16 ***
## d1          -0.02309    0.24311  -0.095    0.925
## d2           0.36428    0.24656   1.477    0.147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.344 on 44 degrees of freedom
## Multiple R-squared:  0.6714, Adjusted R-squared:  0.6565
## F-statistic: 44.95 on 2 and 44 DF,  p-value: 2.324e-11
```

Tehes täiendavalt ANOVA, on näha, et diameeter, mis lineaarses regressioonimudelis ei olnud oluline, on dispersioonanalüüsi järgi peamine varieeruvuse kirjeldaja.

```
# Näide R keskkonnas
anova(lm.col)
```

```
## Analysis of Variance Table
##
## Response: h
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## d1         1 158.487 158.487 87.7259 4.837e-12 ***
## d2         1   3.943   3.943  2.1828    0.1467
## Residuals 44  79.491   1.807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Arvutame VIF väärtuse multikollineaarsuse tuvastamiseks
vif(lm.col)
```

```
##       d1       d2
## 46.96857 46.96857
```

Kui VIF väärtus on üle 5, siis on tunnuste vahel tugev korrelatsioon, mida saame kontrollida ka korrelatsioonimaatriksiga.

```
# Näide R keskkonnas
cor(andmed[,c("d1","d2")])
```

```
##           d1        d2
## d1 1.0000000 0.9892973
## d2 0.9892973 1.0000000
```

Sellest on näha, et need on tugevas korrelatsioonis (0,986). See mõjutab ka lineaarse regressiooni parameetrite arvutamist.

Dispersioonanalüüs

Dispersioonanalüüs (ANOVA) on statistiline meetod, mida kasutatakse rohkem kui kahe rühma keskväärtuste võrdlemiseks. See analüüs võimaldab hinnata, kas faktortunnuse erinevatel tasemetel (ehk rühmadel) on

oluline mõju uuritavale tunnusväärtusele. Dispersioonanalüüs jagab koguhajuvuse osadeks, mis on seletatavad erinevate faktoritega, võimaldades nii tuvastada, millised faktorid mõjutavad oluliselt uuritavat tunnust.

Dispersioonanalüüsi põhimõtted ja eeldused

Dispersioonanalüüs põhineb koguhajuvuse jagamisel kaheks peamiseks komponendiks: * **Seletatud hajuvus (rühmadevaheline hajuvus):** See on hajuvus, mis on tingitud erinevustest rühmade keskväärtuste vahel. Suur rühmadevaheline hajuvus viitab sellele, et faktortunnusel on oluline mõju uuritavale tunnusele. * **Seletamata hajuvus (rühmasisene hajuvus):** See on hajuvus, mis esineb iga rühma sees, tulenedes juhuslikest põhjustest. Väike rühmasisene hajuvus viitab sellele, et rühmade keskväärtused on omavahel sarnased.

Dispersioonanalüüsi rakendamiseks on vaja täita mõningaid eeldusi: 1. **Juhuslikud vead on üksteisest sõltumatud:** Iga vaatluse viga ei tohiks mõjutada teiste vaatluste vigu. 2. **Juhuslike vigade keskväärtus on null ja dispersioon on faktori kõikidel tasemetel ühesugune (homoskedastiivsus):** Vigade hajuvus (dispersioon) peaks olema konstantne kõigi rühmade korral. See on oluline eeldus, mille rikkumine võib viia valede järeldusteni. Meie varasemas vestluses rääkisime pikemalt homoskedastiivsusest ja selle olulisusest. Kui homoskedastiivsuse eeldus ei kehti, tuleb kasutada alternatiivseid meetodeid või transformeerida andmeid. 3. **Kõigi faktorite juhuslikud vead on normaaljaotusega:** Veaterminid peavad olema ligikaudu normaalselt jaotunud igas rühmas. See eeldus on eriti oluline väikeste valimite korral. Suuremate valimite korral ei ole see eeldus nii kriitiline tänu tsentraalse piirteoreemi mõjule. 4. **Kõik faktorite mõjud on aditiivsed:** Faktortunnuste mõjud liituvad lineaarselt, mis tähendab, et ühe faktori mõju ei sõltu teise faktori tasemest. Kui see eeldus ei kehti, võib kaaluda mitte-aditiivseid mudeleid.

Kui mõni nendest eeldustest ei ole täidetud, võib kasutada mitteparameetrilisi meetodeid, nagu Kruskal-Wallise test, mis ei eelda normaaljaotust ja homoskedastiivsust.

Dispersioonanalüüsi tüübid

Dispersioonanalüüsi saab liigitada mitmel viisil:

1. **Ühefaktoriline dispersioonanalüüs (One-way ANOVA):** Kasutatakse, kui uuritakse ühe faktortunnuse mõju funktsioontunnusele. Näiteks võrdlemaks müügitulemusi erinevates piirkondades, kus piirkond on faktortunnus.
2. **Kahefaktoriline dispersioonanalüüs (Two-way ANOVA):** Kasutatakse, kui uuritakse kahe faktortunnuse samaaegset mõju funktsioontunnusele. Näiteks analüüsimaks nii reklaamikampaania tüübi kui ka piirkonna mõju müügitulemustele. See võimaldab uurida ka faktortunnuste koosmõju.
3. **Mitme faktoriline dispersioonanalüüs (Multifactor ANOVA):** Kasutatakse, kui uuritakse rohkem kui kahe faktortunnuse samaaegset mõju funktsioontunnusele.
4. **Fikseeritud ja juhuslikud faktorid:** Faktortunnus võib olla fikseeritud (mõõtmised on tehtud kõigil võimalikel tasemetel) või juhuslik (mõõdetud tasemed on juhuslik valim kõigi võimalike tasemete hulgast).

Dispersioonanalüüsi teostamine

Dispersioonanalüüsi teostamiseks arvutatakse erinevad ruutude summad ja keskmised ruudud:

- **Kogu ruutude summa (SST):** See mõõdab kogu hajuvust andmetes, mis on arvutatud vaatluste ja üldkeskmise hälvete ruutude summana.
- **Rühmadevaheline ruutude summa (SSR või MST):** See mõõdab hajuvust, mis on tingitud erinevustest rühmade keskväärtuste vahel. See on seletatud hajuvus.
- **Rühmasisene ruutude summa (SSE või MSE):** See mõõdab hajuvust iga rühma sees, mis on tingitud juhuslikest põhjustest. See on seletamata hajuvus.

Keskmised ruudud (MS) saadakse, jagades ruutude summad vastavate vabadusastmete arvuga: * **Rühmadevaheline keskmine ruut (MSR või MST):** MSR = SSR / (k - 1), kus k on rühmade arv. * **Rühmasisene keskmine ruut (MSE):** MSE = SSE / (n - k), kus n on vaatluste koguarv.

Teststatistik F arvutatakse järgmiselt: * **F = MSR / MSE**

F-statistik võrreldakse F-jaotuse kriitilise väärtusega, et otsustada, kas nullhüpotees (rühma keskväärtused on võrdsed) lükatakse tagasi. Kui F-statistik on suurem kui kriitiline väärtus, siis on piisavalt tõendeid nullhüpoteesi tagasilükkamiseks ja sisuka hüpoteesi toetamiseks (et rühmade keskväärtused on erinevad).

Dispersioonanalüüsi tulemuste interpretatsioon

Dispersioonanalüüsi tulemuste põhjal saab teha järgmisi järeldusi:

1. **F-statistik ja p-väärtus:** F-statistik ja sellega seotud p-väärtus näitavad, kas rühmade keskväärtuste erinevused on statistiliselt olulised. Väike p-väärtus (tavaliselt alla 0,05) viitab sellele, et nullhüpotees lükatakse tagasi.
2. **Rühma keskmised:** Kui nullhüpotees lükatakse tagasi, siis tuleb uurida, milliste rühmade keskväärtused erinevad. Selleks võib kasutada post-hoc teste (nt Tukey HSD, Bonferroni).
3. **Mõju suurus:** Lisaks statistilisele olulisusele on oluline hinnata ka mõju suurust, mis näitab, kui palju faktortunnus mõjutab uuritavat tunnust. Mõju suuruse mõõdikud on näiteks eta ruut (η^2) või osaline eta ruut.

MS Exceli keskkonnas saab ühefaktorilise dispersioonanalüüsi tegemiseks kasutada protseduuri Anova (Analysis of Variance). Selleks tuleb andmed eelnevalt salvestada rühmade kaupa tulpadesse või veergudesse. Ühefaktorilise dispersioon¬analüüsi protseduuri saab kasutada nii tasakaalustatud kui ka tasakaalustamata mudeli jaoks. Protseduur käivitatakse menüüst: Andmed, Data Analysis, Anova: Single Factor.

Avanevas sisestusaknas tuleb määrata: Input Range – algandmete plokk, Grouped by – määratakse, kas rühmad on orienteeritud veerge pidi (Columns) või ridu pidi (Rows), Labels in First Row – märgitakse faktori tasemete väärtuste olemasolu korral andme-ploki esimeses reas (või esimeses veerus, kui andmed on orienteeritud ridu pidi), Output options – määratakse, kas vastus salvestatakse samale töölehele (Output Range), uuele töölehele (New Worksheet Ply) või uude faili (New Workbook). Protseduuri tulemusena väljastatakse funktsioontunnuse aritmeetilised keskmised ja dispersioonid rühmade kaupa ning dispersioonanalüüsi tabel.

Erinevate väetiste mõju selgitamiseks männipoogendite kasvule korraldati Mõisaküla seemlas katse, kus ühe ja sama klooni poogendeid väetati lämmastikväetisega (N), fosforväetisega (P) ja täisväetisega (NPK). Kontrollala ei väetatud. Kolmandal aastal peale väetamist mõõdeti iga katseala poogenditelt 10 juhuslikult valitud (ladvakasvu) okka pikkused millimeetrites.

|N|60 | 63 | 64 | 65 | 58 | 60 | 64 | 61 | 65 | 58 P | 51 | 49 | 52 | 48 | 54 | 52 | 52 | 51 | 50 | 50 NPK | 56 | 56 | 55 | 56 | 56 | 53 | 58 | 59 | 58 | 53 Kontroll | 61 | 54 | 56 | 58 | 55 | 63 | 59 | 53 | 54 | 56

Kas saab tõestada, et erinevate väetamise variantide puhul on poogendite keskmised okka pikkused erinevad? Olulisuse nivooks valime 0,05. Funktsioontunnuseks on okka pikkus. Faktoriks on väetamise variant, millel on 4 taset: N, P, NPK ja kontroll. Nullhüpoteesiks on väide, et kõigi nelja väetamise variandi korral on poogendite okka pikkuste keskväärtused ühesugused. Sisukaks hüpoteesiks on väide, et vähemalt ühe väetamise variandi korral on okka pikkuse kesk¬väärtus teistest erinev. MS Exceli dispersioonanalüüsi protseduuri tulemusena saame järgmised tabelid.

Anova: Single Factor | | | | | | SUMMARY | | | | | | Groups | Count | Sum | Average | Variance | | N | 10 | 618 | 61,8 | 7,51 | | P | 10 | 509 | 50,9 | 2,99 | | NPK | 10 | 560 | 56,0 | 4,00 | | Kontroll | 10 | 569 | 56,9 | 10,77 | | ANOVA | | | | | | Source of Variation | SS | df | MS | F | P-value | F crit Between Groups | 598,2 | 3 | 199,40 | 31,57 | 3,5E-10 | 2,87 Within Groups | 227,4 | 36 | 6,32 | | | | Total | 825,6 | 39 | | | | Dispersioonanlüüsi tabelist on näha, et arvutatud F-statistiku väärtus on 31,57. Kuna see on suurem kui F-jaotuse 0,05-täiendkvantiil (F0,05, 3, 36 = 2,87), võime lugeda tõestatuks, et erinevate väetamisvariantide korral on vähemalt ühe variandi poogendite keskmine okka pikkus teistest erinev. Lisaks ühefaktorilisele dispersioonanalüüsile on MS Exceli keskkonnas kaks kahe-faktorilise dispersioonanalüüsi protseduuri: kordsusteta ja kordsustega andmestike jaoks. Mõlemal juhul peab olema tegemist tasakaalustatud mudeliga.

Näide 5.7. Olgu funktsioontunnuseks teatud metsatüübist juhuslikult valitud kümne puistu hektaritagavara ja kõrguse suhte aritmeetiline keskmine Eesti riigimetsa 50 aastastes puistutes. Antud suurus on teatud mõttes metsatüübi puistute tiheduse näitaja, mis iseloomustab puidutagavara kõrgusühiku kohta. Alljärgnevas tabelis on see suurus arvutatud kolme peamise puuliigi (männikute, kuusikute ja kaasikute) ja nelja kasvu¬koha¬tüübi (kastikuloo, jänesekapsa, sinilille ja kõdusoo) jaoks.

KKT | MA | KU | KS KL | 8,44 | 7,70 | 6,25 JK | 11,66 | 12,04 | 10,09 SL | 10,17 | 10,55 | 9,16 KS | 11,40 | 10,95 | 8,65

Meid huvitab, kas tagavara ja kõrguse suhe on antud puuliikidel erinev ning kas see suhe on antud kasvukohatüüpides erinev. Käivitame MS Excelis kahefaktorilise dispersioonanalüüsi protseduuri kordsusteta andmestiku jaoks, valides menüüst: Tools, Data Analysis, Anova: Two-Factor Without Replication. Avanevas sisestusaknas tuleb määrata: Input Range – algandmete plokk risttabeli kujul, Labels – märgitakse faktorite tasemete väärtuste olemasolu korral andmeploki esimeses reas ja esimeses veerus, Output options – määratakse, kas vastus salvestatakse samale töölehele (Output Range), uuele töölehele (New Worksheet Ply) või uude faili (New Workbook). Protseduuri tulemusena väljastatakse funktsioontunnuse aritmeetilised keskmised ja dispersioonid mõlema faktori kõigi tasemete jaoks ning dispersioonanalüüsi tabel.

Table 4: Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
KL	3	22,39	7,46	1,24
JK	3	33,79	11,26	1,07
SL	3	29,88	9,96	0,52
KS	3	31,01	10,34	2,17
MA	4	41,67	10,42	2,16
KU	4	41,24	10,31	3,42
KS	4	34,15	8,54	2,68

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	23,70	3	7,90	43,70	0,0002	4,76
Columns	8,92	2	4,46	24,67	0,0013	5,14
Error	1,08	6	0,18			
Total	33,70	11				

Antud ülesandes on ridades (Rows) faktori “kasvukohatüüp” tasemed ja veergudes (Columns) faktori “puuliik” tasemed. Kahefaktorilise dispersioonanalüüsi tabelis kontrollitakse mõlema faktori mõju funktsioontunnusele. Et mõlemal faktori korral on F statistiku väärtus (F) suurem kui F-jaotuse 0,05-täiendkvantiil (F crit), oleme tõestanud kaks väidet (olulisuse nivooga 0,05): 1) puistute tagavara ja kõrguse suhe ei ole männikutes, kuusikutes ja kaasikutes ühesugune; 2) puistute tagavara ja kõrguse suhe ei ole antud kasvu¬koha¬tüüpides ühesugune.

Näide 5.8. Pauska seemlas korraldati katse, kus ühe ja sama klooni kuusepoogendeid väetati lämmastikväetisega (N), fosforväetisega (P) ja täisväetisega (NPK). Kontrollala ei väetatud. Kahel järjestikusel käbiaastal: 1989/90 ja 1992/93 korjati iga väetis-variandiga alal seitsmelt poogendilt kõik käbid ja leiti iga puu keskmine käbi kaal grammides.

Table 6: Väetise andmed

N	P	NPK	Kontroll
1989/90	28,2	28,0	27,0
28,6	26,9	28,3	28,2
17,7	26,5	29,2	29,8
28,1	24,2	23,2	30,6
29,0	21,6	26,4	20,0
30,5	23,9	23,4	20,2

N	P	NPK	Kontroll
25,1	26,2	22,5	21,4
1992/93	20,7	23,1	25,5
22,8	24,3	22,2	23,6
20,6	23,0	23,3	21,7
23,3	24,9	22,0	24,6
24,5	26,0	23,8	24,8
23,6	25,7	24,0	23,3
23,7	22,2	23,4	21,9

Meid huvitab, kas poogendi keskmine käbi kaal on erinevate väetamisvariantide korral erinev. Samuti soovime teada saada, kas keskmine käbi kaal on vaadeldavatel käbiaastatel erinev.

Käivitame MS Excelis kahefaktorilise dispersioonanalüüsi protseduuri kordsustega andmestiku jaoks, valides menüüst: Tools, Data Analysis, Anova: Two-Factor With Replications. Avanevas sisestusaknas tuleb määrata: Input Range – algandmete plokk ülal esitatud tabeli kujul, Rows per sample – kordsus (mõõtmiste arv) igal tasemel (antud näites 7), Alpha – olulisuse nivoo (vaikimisi = 0,05), Output options – määratakse, kas vastus salvestatakse samale töölehele (Output Range), uuele töölehele (New Worksheet Ply) või uude faili (New Workbook). Protseduuri tulemusena väljastatakse funktsioontunnuse aritmeetilised keskmised ja dispersioonid mõlema faktori kõigi tasemete jaoks ning dispersioonanalüüsi tabel.

Anova: Two-Factor With Replication

SUMMARY	N	P	NPK	Kontroll	Total
1989/90					
Count	7	7	7	7	28
Sum	187,2	177,3	180,0	176,3	720,8
Average	26,74	25,33	25,71	25,19	25,74
Variance	18,52	4,83	7,16	21,10	11,85
—					
1992/93					
Count	7	7	7	7	28
Sum	159,2	169,2	164,2	164,3	656,9
Average	22,74	24,17	23,46	23,47	23,46
Variance	2,30	2,11	1,39	1,59	1,91
Total					
Count	14	14	14	14	
Sum	346,4	346,5	344,2	340,6	
Average	24,74	24,75	24,59	24,33	
Variance	13,92	3,56	5,32	11,27	

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Sample	72,91	1	72,91	9,89	0,003	4,04
Columns	1,63	3	0,54	0,07	0,974	2,80
Interaction	15,89	3	5,30	0,72	0,546	2,80
Within	353,95	48	7,37			
Total	444,39	55				

Antud ülesandes on ridades (Sample) faktori "käbiaasta" tasemed ja veergudes (Columns) faktori "väetamisvariant" tasemed. Antud dispersioonanalüüsi tabelis kontrollitakse lisaks mõlema faktori mõjule ka nende faktorite koosmõju olemasolu. Antud näiteülesande dispersioonanalüüsi tabelist saab (olulisuse nivool $\alpha = 0{,}05$) välja lugeda järgmist: 1) vaadeldavatel käbiaastatel (1989/90 ja 1992/93) on kuusepoogendite keskmine käbi kaal erinev (käbiaasta faktori real (Sample) on F-statistiku väärtus (9,89) suurem kui F jaotuse 0,05-täiendkvantiil (4,04)); 2) väetamisvariandi mõju kuusepoogendi keskmisele käbi kaalule ei saa antud andmes-tiku põhjal tõestada (väetamisvariandi faktori real (Columns) on F-statistiku väärtus (0,07) väiksem kui F jaotuse 0,05-täiendkvantiil (2,80)); 3) ka väetamisvariandi ja käbiaasta koosmõju kuusepoogendi keskmisele käbi kaalule ei saa antud andmestiku põhjal tõestada (real Interaction on F-statistiku väärtus (0,72) väiksem kui F jaotuse 0,05-täiendkvantiil (2,80)).