Matemaatiline statistika ja modelleerimine

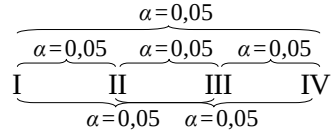
Dispersioonanalüüs

EMÜ doktorikool
DK.0007

Tanel Kaart

Mitmene võrdlus

Võrdleme näiteks 4 gruppi, lubades iga üksikvõrdluse puhul eksimist 5% tõenäosusega.



Tõenäosus, et üksikvõrdlusel viga ei tehta, on $1-\alpha=0{,}95$.

Tõenäosus, et kuuel üksikvõrdlusel kokku ei eksita, on $(1-\alpha)^6=0{,}95^6\approx 0{,}735$.

Mistõttu tõenäosus teha üks (või mitu) vale otsus(t) 4 grupi paarikaupa võrdlemisel on $1-0{,}735=0{,}265$ (eksimise tõenäosus on üle 25%!).

Otsite näiteks põhjust, mis võiks soodustada taimedel haiguse tekkimist. Viite läbi uuringu ja fikseerite haigetel ja tervetel taimedel 100 potentsiaalselt haigestumist mõjutava tunnuse väärtused (a'la mulla toitaineterikkus, eelmisel kuul sadanud vihma kogus, taimi kasvata taluniku pikkus jne).

Iga potentsiaalse haigusega seotud tunnuse osas võrdlete haigeid ja terveid taimi kasutades olulisuse nivood 0,05.

Kui nüüd eeldada, et tegelikult ei mõjuta ükski valitud 100-st tunnusest haigestumist, siis sellest hoolimata võiksite antud uuringu puhul lugeda tõestatuks umbes 5 haiguse tekkimist soodustavat tegurit.

Mitmene võrdlus

Bonferroni meetod: piiramaks k üksikvõrdluse puhul ühe või enama vea tegemise tõenäosust olulisuse nivooga α, tuleb kõigil üksikvõrdlustel võtta olulisuse nivooks α/k.

Näiteks 4 grupi võrdlemisel, garanteerimaks kuue võrdluse peale kokku eksimist mitte üle 5%-lise tõenäosusega, tuleb üksikvõrdlustel võtta olulisuse nivooks $\alpha^* = \alpha/k = 0{,}05/6 \approx 0{,}0083$.

Bonferroni-Holmi meetod: teostatakse kõik testid ja järjestatakse saadud olulisuse tõenäosused, $p_1 \le p_2 \le \ldots \le p_k$; otsused nullhüpoteesi kasuks või kahjuks tehakse kasutades olulisuse nivoosid α/k, $\alpha/(k-1)$, $\alpha/(k-2)$, …, $\alpha/2$, α.

Kui teostatavate testide arv kasvab, väheneb kasutatav olulisuse nivoo kiiresti ja alternatiivse hüpoteesi tõestamine osutub sageli äärmiselt raskeks (nõuab tohutu hulga vaatluste olemasolu).

Seetõttu ei ole statistilised meetodid mitte eriti sobivad katse/eksituse-meetodil teaduse tegemiseks (proovime, kas midagi õnnestub)!

False Discovery Rate (“valeavastuste määr”)

Idee: piirata (kontrollida) ekslikult vastuvõetud alternatiivsete hüpoteeside osakaalu kõigi vastuvõetud alternatiivsete hüpoteeside seas (piiriks näiteks 5%).

Üks lihtne moodus antud lähenemist ise kasutada on järgmine:

1. Järjesta testide poolt raporteeritavad olulisuse tõenäosused kasvavalt, $p_1 \le p_2 \le \ldots \le p_k$.
2. Kirjuta välja kriitilised suurused α/k, $2*\alpha/k$, $3*\alpha/k$, …, α, kus α näitab maksimaalset nn *False Discovery Rate*’i.
3. Võrdle iga olulisuse tõenäosust temale vastava (sama jrk numbriga) kriitilise väärtusega, kuni jõuad paarini, kus olulisuse tõenäosus on suurem kriitilisest väärtusest. Loe kõigi eelnenud hüpoteeside jaoks alternatiivne hüpotees tõestatuks ja temale järgnevate hüpoteeside puhul jää nullhüpoteesi juurde.

Keskmiste mitmene võrdlus

On k gruppi, mille keskmist taset tahame võrrelda.

Sellisel juhul on sageli otstarbekas $k(k–1)/2$ paariviisilist võrdlust (t-testi) asendada üheainsa hüpoteeside paari kontrollimisega.

Viimase võib sõnastada kujul:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
H_1: leiduvad sellised grupid i,j, et $\mu_i \neq \mu_j$

Eeldustel, et

¤ vaatlused on sõltumatud,

¤ uuritav (sõltuv) tunnus on normaaljaotusega ja

¤ uuritava tunnuse varieeruvus võrreldavais gruppides on ühesugune,

on taolise hüpoteeside paari korral rakendatavaks analüüsimeetodiks **dispersioonanalüüs**.

Dispersioonanalüüs

Dispersioonanalüüsil jagatakse tunnused vastavalt nende rollile kaheks:

¤ tunnus, mille keskmisi võrrelda soovitakse, on **uuritav tunnus** e **funktsioontunnus** (lehma piimatoodang, forelli kasvukiirus, talle mass, sea pekipaksus, jne);

¤ (diskreetne või mittearvuline) tunnus, mille väärtuste alusel võrreldavad grupid moodustatakse, on **faktortunnus** (tõug, lüpsiseade, laudatüüp jne).

Dispersioonanalüüsi tulemuste tõlgendamisel räägitaksegi enamasti faktor-tunnuse **mõjust** uuritavale tunnusele.

Näiteks, tõu või lüpsiseadme või laudatüübi vm mõju piimatoodangule, kasvanduse mõju forellide kasvukiirusele, omaniku mõju talle massile, genotüübi (teatud geenikombinatsioonide) mõju sigade pekipaksusele jne.

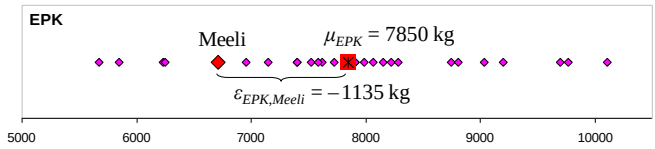
Dispersioonanalüüsi mudel

¤ Iga rühma i (kus $i=1,\ldots,k$) iseloomustab keskmine uuritava tunnuse väärtus μ_i, mistõttu mõõtmistulemused saab esitada mudeliga

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

kus y_{ij} on uuritava tunnuse väärtus i. rühma kuuluval j. objektil ja ε_{ij} on juhuslik mõju (objekti omapära).

Näiteks EPK-tõugu lehm Meeli 1. laktatsiooni piimatoodang 6715 kg on väljendatav kui uuritud EPK-tõugu lehmade 1. laktatsiooni keskmise toodangu $\mu_{EPK}=7850$ kg ja Meeli tõusisese erinevuse $\varepsilon_{EPK,Meeli}=-1135$ kg summa.



Dispersioonanalüüsi mudel

¤ Faktortunnuse mõju uurimiseks esitatakse mudel kujul

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

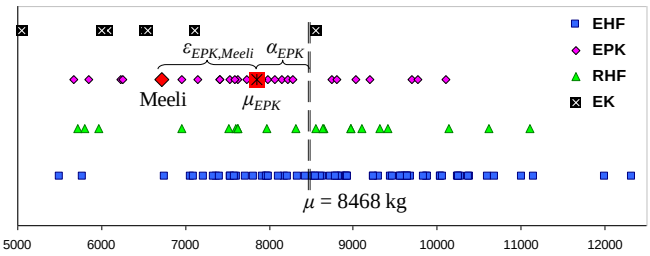
kus μ tähistab üldkeskmist ja α_i on faktori i. taseme poolt põhjustatud kõrvalekalle üldkeskmisest (i. taseme mõju), $\mu_i = \mu + \alpha_i$.

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
H_1: leiduvad grupid i,j, et $\mu_i \neq \mu_j$

$\Leftrightarrow$

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$
H_1: leidub grupp i, et $\alpha_i \neq 0$

Näiteks EPK-tõugu lehm Meeli 1. laktatsiooni piimatoodang 6715 kg on väljendatav kui kõigi uuritud lehmade keskmise 1. laktatsiooni piimatoodangu $\mu=8468$ kg, EPK-tõu mõju (EPK-tõugu lehmade 1. laktatsiooni keskmise toodangu erinevus üld-keskmisest) $\alpha_{EPK}=-618$ kg ja Meeli tõusisese erinevuse $\varepsilon_{EPK,Meeli}=-1135$ kg summa.



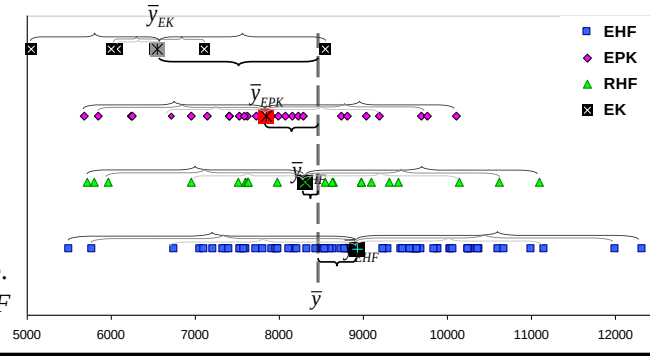
Dispersioonanalüüsi tööpõhimõte

Dispersioonanalüüsi tööpõhimõte seisneb uuritava tunnuse rühmadesisese (nn juhusliku) varieeruvuse $SSA = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ ja rühmadevahelise (faktori mõjust tingitud) varieeruvuse $SSE = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$ võrdlemises – kui rühmadevaheline erinevus on suurem kui rühmadesisene erinevus, on tegu ilmse tõendiga faktortunnuse mõju olemasolu kohta.

Siit ka analüüsi nimetus – dispersioonanalüüs [*analysis of variance*, ANOVA].

Näide.

$i = EK, EPK, RHF, EHF$



Dispersioonanalüüsi tabel

Dispersioonanalüüsiga seotud arvutused koondatakse tavaliselt alljärgnevasse nn. dispersioonanalüüsi tabelisse.

Varieeruvuse allikas	Hälvete ruutude summa	Vabadusastmeid	Keskruut	*F*-suhe	Olulisustõenäosus
Faktor	$SSA = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$	$k-1$	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$	p
Viga	$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$n-k$	$MSE = \frac{SSE}{n-k}$		
Kokku	$SS = SSA + SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$	$n-1$			

Juhul, kui faktortunnuse mõjule vastav keskmine **gruppide vaheline varieeruvus** *MSA* on suurem, kui uuritavate objektide omapärale vastav keskmine **gruppide sisene varieeruvus** *MSE*, on *F*-statistiku väärtus ühest suurem.

Piisavalt suure *F*-suhte väärtuse korral võib lugeda tõestatuks sisuka hüpoteesi – leiduvad vähemalt 2 teineteisest selgelt eristuvat gruppi.

Dispersioonanalüüs

Näide. Uuritakse ühes katsefarmis peetava 121 lehma 1. laktatsiooni piimatoodangu sõltuvust tõust (EHF, RHF, EPK, EK). Kontrollitav hüpoteeside paar on kujul:

$H_0: \mu_{EHF} = \mu_{RHF} = \mu_{EPK} = \mu_{EK}$ ⇔ H_0: tõul ei ole mõju
H_1 : leiduvad tõugrupid i,j, et $\mu_i \neq \mu_j$ H_1 : tõul on mõju

Dispersioonanalüüsi mudel on kujul $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,
kus μ on kõigi farmi lehmade keskmine 1. laktatsiooni piimatoodang, α_i on i. tõu keskmine erinevus sellest (i. tõu mõju, $i = EHF, RHF, EPK, EK$) ning y_{ij} ja ε_{ij} on vastavalt i. tõugu j. lehma mõõdetud piimatoodang ja selle erinevus tõu keskmisest (lehma "omapära", $j = 1, \ldots, n_i$, n_i on lehmade arv i. tõus).

Tõug	n_i	$\bar{y}_i = \hat{\mu}_i$	$s_i^2 = \hat{\sigma}_i^2$
EHF	68	8929,9	1776652
RHF	20	8311,5	2265768
EPK	27	7850,1	1335053
EK	6	6549,3	1419571
Kokku	121	8468,1	2093541

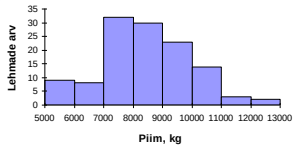
Tõugude mõjud on kõrvaloleva keskmiste toodangute tabeli alusel leitavad kujul

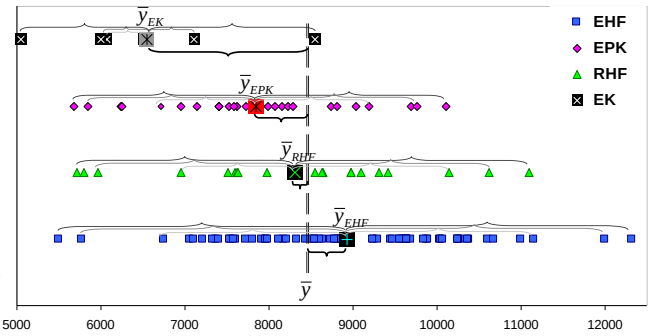
$\alpha_{EHF} = 460{,}8$ kg; $\alpha_{RHF} = -156{,}6$ kg;
$\alpha_{EPK} = -618{,}0$ kg ja $\alpha_{EK} = -1918{,}8$ kg.

Nende mõjude erinevuse kontrollimiseks tuleb läbi viia dispersioonanalüüs [viimase eeldused dispersioonide võrdsuse ja normaaljaotuse (vt ka järgmine lk) osas on enamvähem täidetud].

Dispersioonanalüüs

Näide. Uuritakse ühes katsefarmis peetava 121 lehma 1. laktatsiooni piimatoodangu sõltuvust tõust.





Varieeruvuse allikas	Hälvete ruutude summa	Vabadusastmete arv	Keskruut	F-suhe	p
Tõug	47330434	3	15776811	9,053	**0,000019**
Viga	203894493	117	1742688		
Kokku	251224927	120			

< 0,05 => H_1: tõul on mõju

Dispersioonanalüüs

Juhul, kui võrreldavaid gruppe on vaid kaks, on dispersioonanalüüsi tulemused identsed võrdsete dispersioonide eeldusel läbi viidud *t*-testiga.

Näide. Võrreldakse kahest erinevast tõust sigade ööpäevast juurdekasvu.

Andmed:

	Ööpäevane juurdekasv (g)			
Tõug 1	520	550	560	530
Tõug 2	630	690	700	680

MS Exceli protseduuride *t-Test: ...Equal Variances* ja *Anova: Single Factor* väljatrükid.

t-Test: Two-Sample Assuming Equal Variances

	Tõug 1	*Tõug 2*
Mean	540	675
Variance	333,33	966,67
Observations	4	4
Pooled Variance	650	
Hypothesized Mean Difference		
df	6	
t Stat	-7,4885	
P(T<=t) one-tail	0,00015	
t Critical one-tail	1,9432	
P(T<=t) two-tail	0,00029	
t Critical two-tail	2,4469	

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Tõug 1	4	2160	540	333,33
Tõug 2	4	2700	675	966,67

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	36450	1	36450	56,0769	0,00029	5,9874
Within Groups	3900	6	650			
Total	40350	7				

$p < 0{,}05$
=>
erinevat tõugu sead kasvavad erineva kiirusega

Keskväärtuste võrdlemine

Keskväärtuste võrdlemine

1 grupi keskmise võrdlus konstandiga	**2 grupi keskmiste võrdlus**	**3 või enama (*k*) grupi keskmiste võrdlus**
$H_0: \mu = c$ $H_1: \mu \neq c$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ H_1: leiduvad i,j: $\mu_i \neq \mu_j$
Usalduspiirid; normaaljaotuse eeldusel *t*-test; suurte valimite (n>60) või teadaoleva dispersiooni σ^2 korral *z*-test	*t*-test (kolm erinevat!), eeldused: uuritav tunnus normaaljaotusega (või suur valim); teadaolevate dispersioonide σ_1^2 ja σ_2^2 korral *z*-test	Dispersioonanalüüs, eeldused: ¤ uuritav tunnus normaaljaotusega, ¤ uuritava tunnuse varieeruvus võrreldavais gruppides ühesugune

Mitmefaktoriline dispersioonanalüüs

Kui vaatlusobjekte saab rühmitada mitme tunnuse (faktortunnuse) järgi, võib osutuda mõttekaks analüüsidagi korraga mitme faktortunnuse mõju (näiteks igal lehmal võib olla fikseeritud tema tõug ja farm, igal kalal tema sugu ja püügikoht).

Dispersioonanalüüsi mudel, mis hõlmab kahe faktortunnuse mõjusid, on kujul:

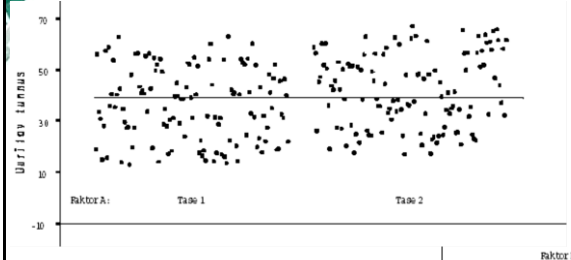
$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

kus μ tähistab üldkeskmist,
α_i-d ja β_j-d märgivad uuritava tunnuse keskmise muutust vastavalt esimese ja teise faktori väärtuste muutumisele (α_i on esimese faktori i. taseme mõju ja β_j on teise faktori j. taseme mõju),
y_{ijk} ning ε_{ijk} on vastavalt esimese faktori i. tasemel ja teise faktori j. tasemel sooritatud k. mõõtmise väärtus ning selle erinevus sama väärtuste kombinatsiooni keskmisest (vaatluse omapära, mudeli viga).

Mitmefaktoriline dispersioonanalüüs

Miks seda vaja on?

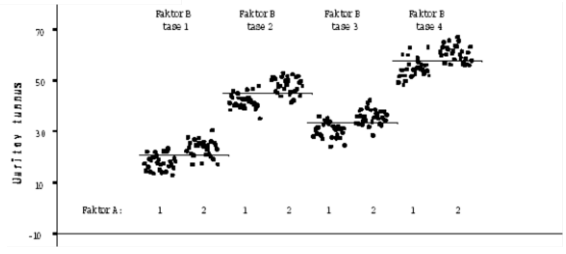
1) Hinnangute ja otsustuste täpsus võib paraneda.



Illustratsiooniks kaks samu andmeid illustreerivat hajuvusdiagrammi.

(joonised M. Mõlsi konspektist)

Osutub, et vaadeldes uuritava tunnuse väärtusi homogeensete gruppide kaupa (faktori B järgi), võib huvipakkuva faktori (A) mõju selgemalt esile tõusta.



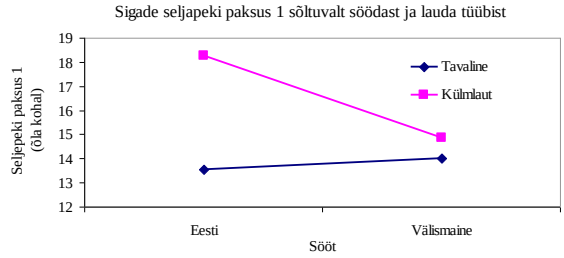
Mitmefaktoriline dispersioonanalüüs

Miks seda vaja on?

2) Võimalik selgemalt väljendada uuritava tunnuse ja faktorite vahelisi seoseid.

3) Interaktsioonid e koosmõjud – uuritava tunnuse väärtused muutuvad ühe faktori tasemete vahel erinevalt, sõltuvalt teise faktori väärtustest.

4) Ilma ei pruugi mudel olla korrektne (jääkliige ei puugi olla normaal-jaotusega).



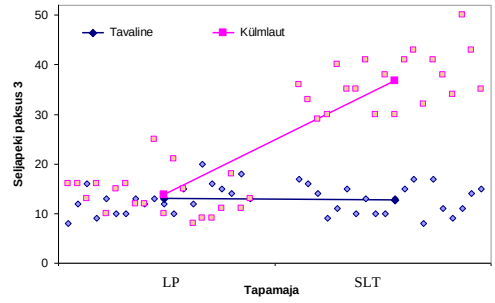
Kahefaktoriline faktoritevahelist interaktsiooni arvestav mudel esitatakse kujul

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

kus γ_{ij} märgib esimese faktori i. taseme ja teise faktori j. taseme koosmõju.

Mitmefaktoriline dispersioonanalüüs

Näide. 80-st seast 40 peeti tavalistes ja 40 välitingimustes. Mõlemast grupist pooled tapeti kohalikus tapamajas (LP) ja pooltel eelnes "paremale jahimaadele siirdamisele" stressirohke üle 200 kilomeetrine transport auto ja praami abil (SLT).



The SAS System

The GLM Procedure

Dependent Variable: BackFat3 BackFat3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8295.450000	2765.150000	166.77	<.0001
Error	76	1260.100000	16.580263		
Corrected Total	79	9555.550000			

R-Square	Coeff Var	Root MSE	BackFat3 Mean
0.868129	21.34672	4.071887	19.07500

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Kasvukoht	1	3050.450000	3050.450000	183.98	<.0001
Tapamaja	1	2553.800000	2553.800000	154.03	<.0001
Kasvukoht*Tapamaja	1	2691.200000	2691.200000	162.31	<.0001

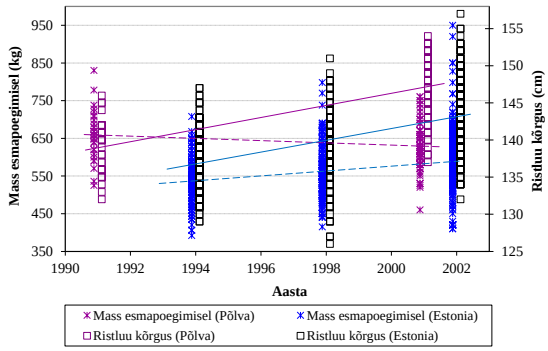
H_0 : mudel ei ole parem võrreldes konstantse mudeliga
H_1 : mudel on parem võrreldes konstantse mudeliga

Hüpoteeside kontroll mudeli iga faktori (ja nende kombinatsiooni) kohta

$H_0: a_1 = \alpha_2 = \ldots = 0$,
H_1: leidub i, et $a_i \neq 0$;
$H_0: \beta_1 = \beta_2 = \ldots = 0$,
H_1: leidub j, et $\beta_j \neq 0$;
$H_0: \gamma_{11} = \gamma_{12} = \ldots = 0$,
H_1: leiduvad i,j, et $\gamma_{ij} \neq 0$.

Kovariatsioonanalüüs

Näide. Sooviti uurida, kas lehmade suurus on aastate jooksul muutunud.
Kasutada olid kahe aasta andmed ühest ja kolme aasta andmed teisest majandist.
Uuritavad tunnused: mass esmapoegimisel, ristluu kõrgus.
Faktorid: farm (diskreetne, 2 taset), aasta (pidev).
Mudel: $y_{ij} = \mu + \alpha_i + b * x_{ij} + \varepsilon_{ij}$



Faktor	Sõltuv tunnus	
	Esmapg. mass	Ristluu kõrgus
Aasta	0,013	<0,001
Farm	<0,001	0,84
Aasta*Farm	<0,001	0,81