Tartu Ülikool

Matemaatika-informaatikateaduskond

Matemaatilise statistika instituut

Mari Liiva

Binaarsete tunnuste polügeense determineerituse hindamine

Magistritöö

Juhendajad: Tanel Kaart

Mare Vähi

Tartu 2012

Sisukord

Sissejuhatus

Põllumajandusloomade aretusteoorias eeldatakse enamasti, et selektsiooni aluseks olevad tunnused jaotuvad normaaljaotuse seaduspärade kohaselt. Juhul, kui mõõdetavat tunnust mõjutavad paljud geenid ja iga üksiku geeni efekt eraldi võetuna on tühine, ongi see vastavalt tsentraalsele piirteoreemile nii. Nii on näiteks lehmade piimatoodangu ja lammaste kehamassi jaotuseks normaaljaotus, samuti on normaaljaotusega inimeste pikkus ja vererõhk. Taoliste, nn polügeensete tunnuste statistiliseks modelleerimiseks on aastakümneid kasutatud üldiseid lineaarseid mudeleid ja geneetilise determineerituse uurimiseks üldiseid lineaarseid segamudeleid (Kaart, 2006;Mrode, 2005).

Kaheväärtuseliste, nn binaarsete tunnuste (näiteks haige/terve, tiinestub/ei tiinestu) kujunemise osas on loomulik eeldada vaid ühe geeni rolli – üks geenivariant põhjustab haigestumist, teine geenivariant aga mitte. Siis aga peaks binaarsete tunnuste pärandumine järgima Mendeli seaduseid (Viikmaa, 1998). Paljude haiguste puhul see nii ka on (näiteks Huntingtoni tõbi inimestel või BLAD - leukotsüütide kleepumise puudulikkuse sündroom - holsteini tõugu veistel). Siiski on enamus binaarseid tunnuseid oma olemuselt polügeensed, nende avaldumine sõltub paljude geenide ja ka keskkonna koosmõjust, silmaga nähtav või registreeritav väärtus on aga binaarne. Et saada üle binaarsete tunnuste ja nende polügeense determineerituse vastuolust, on geneetikud võtnud kasutusele nn alustunnuse kontseptsiooni (Falconer ja Mackay, 1996).

Binaarsete tunnuste polügeense determineerituse statistilisel modelleerimisel on kaks varianti – kas jätta tähelepanuta tunnuse tegelik normaaljaotusest erinev jaotus ja rakendada tavalisi üldiseid lineaarseid segamudeleid või võtta kasutusele tunnuse tegeliku jaotusega arvestavad üldistatud lineaarsed segamudelid. Et geneetiliste parameetrite hindamine üldistatud lineaarsete segamudelite abil suurtes populatsioonides on komplitseeritud ning parameetrite olemus võrreldes tavalisest segamudelist hinnatutega võib olla teistsugune, rakendatakse populatsioonipõhisel põllumajandusloomade geneetilisel hindamisel ka binaarsete tunnuste puhul sageli normaaljaotust eeldavaid üldiseid lineaarseid segamudeleid. Näiteks on sellisel viisil hinnatud piimaveiste surnultsünni aretusväärtuseid nii Eestis kui ka Soomes (Uba, 2010; Niskanen ja Juga, 1998).

Käesoleva magistritöö eesmärgiks on esitada binaarsete tunnuste polügeense determineerituse kontseptsioon nö geneetilise mudeli kontekstis, anda metoodiline ülevaade geneetiliste

parameetrite hindamisel kasutatavatest üldistest ja üldistatud lineaarsetest segamudelitest ning üldistatud lineaarsete segamudelite rakendamise võimalustest SAS-s, viia läbi modelleerimiseksperimendid selgitamaks, kas ja kuivõrd on erinevate mudelitega leitud binaarsete tunnuste geneetilised parameetrid erinevad, ja rakendada erinevaid mudeleid lehmade tiinestuvuse geneetilise determineerituse uurimisel.

Töö esimeses peatükis antakse ülevaade nii pidevate kui ka binaarsete tunnuste geneetilise determineerituse olemusest ja nende modelleerimisest nö geneetilise mudeli kontekstis. Töö teises peatükis tuuakse ära üldise lineaarse mudeli, üldise lineaarse segamudeli, üldistatud lineaarse mudeli ja üldistatud lineaarse segamudeli esitused ja kovariatsioonistruktuurid ning selgitatakse mudelite vahelisi erinevusi eeldustes ja parameetrite tähendustes. Samuti tuuakse ära lävendimudeli statistiline esitus logit- ja probit-mudeli kaudu. Teise peatüki viimases alapunktis tutvustatakse binaarsete tunnuste modelleerimist üldistatud lineaarse mudeliga SAS-i protseduuri GLIMMIX abi. Töö kolmandas peatükis viiakse läbi rida modelleerimiseksperimente selgitamaks juhuslike geneetiliste faktorite mõjude ja vastavate dispersioonikomponentide hinnangute täpsust ja sõltuvust rakendatavast mudelist ning binaarse tunnuse väärtuste kujunemisest. Magistritöö neljandas peatükis rakendatakse nii üldist lineaarset segamudelit kui ka logit- ja probit-seosefunktsiooniga üldistatud lineaarseid segamudeleid mullikate ning 1. ja 2. laktatsiooni lehmade tiinestuvuse uurimiseks. Hinnatakse juhuslike faktorite “Isa“ ja “Seemendaja“ mõju, vastavaid dispersioonikomponente ja isa kui geneetilise faktori mõju alusel mullikate tiinestuvuse päritavust. Töö lisas on ära toodud modelleerimiseksperimendi läbiviimisel kasutatud SAS-programm.

1. Geneetiline mudel

Suurem osa morfomeerilisi (arvulisi, mõõdetavaid) tunnuseid, mis iseloomustavad fenotüüpi, sõltuvad samaaegselt paljudest geenidest, st on polügeensed. Et iga üksiku geeni mõju on enamasti tühine, käsitletakse polügeenseid tunnuseid mõjutavaid geene ühtse tervikkomplektina. Polügeensete tunnuste väärtus on juhuslik, sõltudes ühelt poolt geenide konkreetsest komplektist genotüübis, nende omavahelisest paigutusest ja koosmõjudest, teisalt aga ka keskkonnatingimustest. Järgnev geneetiliste mudelite kirjeldus baseerub materjalidel Kaart ja Möls (2011) ning Falconer ja Mackay (1996).

1.1. Pidevate tunnuste geneetiline mudel

Klassikaline geneetiline mudel eeldab, et iga indiviidi fenotüübiväärtus P on kujunenud genotüübiväärtuse G ja keskkonnamõju E koostoime tulemusel:

$$P = G + E. \tag{1}$$

Iga isendi geneetilist potentsiaali iseloomustab tema genotüübiväärtuse erinevus populatsiooni keskmisest:

$$g = G - \bar{G}.$$

Kuna on loomulik eeldada, et kõigi populatsiooni kuuluvate indiviidide summaarne ja seega ka keskmine keskkonnamõju on null, siis võrdub populatsiooni keskmine fenotüübiväärtus seda moodustavate isendite keskmise genotüübiväärtusega, $\bar{P} = \bar{G}$, ja genotüübiefekt g avaldub kujul

$$g = G - \bar{P}.$$

Teisalt on iga indiviidi genotüübiefekt lahtikirjutatav summana

$$g = A + D + I,$$

kus A on indiviidi aditiivne geneetiline efekt (geenide summaarne efekt ehk **aretusväärtus**), D on alleelide lookustesisesest interaktsioonist ehk dominantsusest tingitud efekt ja I on geenide lookustevahelisest koostoimest tulenev efekt ehk epistaasi efekt.

Et vanematelt päranduvad järglastele vaid üksikalleelid, mitte aga nende koosmõjud, on järglaste järgi enamasti hinnatavad üksnes geenide summaarsed mõjud. Seda arvestades omandab mudel (1) kuju

$$P = \overline{P} + A + E\,, \tag{2}$$

kus E sisaldab juba kõiki mitteaditiivgeneetilisi mõjusid. Et dominantsi ja epistaasi efekt on enamasti tühine, mõistetakse suuruse E all siiski eelkõige keskkonnaefekti.

Nii keskkonnamõjud E kui ka geenide summaarsed mõjud A kujutavad enesest suure hulga otseselt mittemõõdetavate väikeste efektide summat – võivad ju mingi tunnuse täpse väärtuse kujunemist samaaegselt mõjutada tuhanded geenid ja hoomamatu hulk mitteaditiivgeneetilisi tegureid. Sestap on loomulik käsitleda nii aditiivgeneetilist efekti A, keskkonnaefekti E kui ka fenotüübiväärtust P juhuslike suurustena, mille jaotuseks on vastavalt tsentraalsele piirteoreemile normaaljaotus. Et nii aditiivgeneetiline efekt A kui ka keskkonnamõjud E näitavad erinevust populatsiooni keskmisest, on nende keskmine väärtus võrdne nulliga.

Loomakasvatuses hinnatakse loomade aditiivgeneetilised efektid ehk aretusväärtused enamasti järglaste järgi. Et ühel isasloomal võib tänu kunstlikule seemendusele olla sadu või tuhandeid järglasi enam kui emasloomal, on eriti oluline isasloomadelt järglastele kanduva geneetilise efekti hindamine. Et iga vanem pärandab järglasele pooled oma geenidest ja seeläbi ka poole oma aretusväärtusest, siis võib mudeli (2) ümber kirjutada ka kujul

$$P = \overline{P} + \tfrac{1}{2} A_s + E = \overline{P} + S + E\,, \tag{3}$$

kus $½A_s = S$ on pool isa aretusväärtusest ehk isamõju, mis kujutab enesest isalt järglasele pärandunud geenide summaarset efekti, kõik ülejäänud geneetilised mõjud, sh emalt pärandunud geenide mõju, loetakse kuuluvaks juhuslike mitteaditiivgeneetiliste mõjude E hulka.

Eeldades keskkonna ja genotüübi sõltumatust, on populatsiooni kuuluvate indiviidide fenotüübiväärtuste dispersioon mudelist (1) avaldatav kujul

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2$$

ja mudelist (2) kujul

$$\sigma_P^2 = \sigma_A^2 + \sigma_E^2. \tag{4}$$

Üks olulisemaid populatsioonigeneetikas ja loomade aretuses kasutatavaid geneetilisi parameetreid on **päritavuskoefitsient**, mis näitab, kui suur osa indiviidide vahelisest fenotüübilisest erinevusest on tingitud nende geneetilisest erinevusest ehk on päritav. Päritavuskoefitsient avaldub aditiivgeneetilise dispersiooni σ_A^2 ja fenotüübidispersiooni σ_P^2 suhtena:

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}.$$

Valemist (3) järeldub, et isaefektide dispersioon σ_s^2 võrdub ¼-ga isade aretusväärtuste dispersioonist ehk aditiivgeneetilisest dispersioonist. Eeldades, et geneetiline varieeruvus isade hulgas on sama suur kui kogu populatsiooni aditiivgeneetiline varieeruvus, võib kirjutada, et $\frac{1}{4}\sigma_A^2 = \sigma_s^2$. Viimasest võrdusest tuleneb omakorda, et $\sigma_A^2 = 4\sigma_s^2$ ja päritavuskoefitsient avaldub isaefektide dispersiooni σ_s^2 kaudu kujul

$$h^2 = \frac{4 \cdot \sigma_S^2}{\sigma_P^2}. \qquad (5)$$

1.2. Diskreetsete tunnuste geneetiline mudel

Vastavalt eelmises punktis esitatud klassikalisele geneetilisele mudelile ei saa polügeenne tunnus omada diskreetseid väärtuseid. Diskreetsete tunnuste puhul on loomulik eeldada, et nende väärtused on täielikult määratud üksikute, mõnikord ka ainult ühe, geeni poolt. Näiteks üks geenivariant põhjustab haigestumist, teine geenivariant aga mitte. Sellisel juhul aga peaks diskreetsete tunnuste pärandumine järgima Mendeli seaduseid. Mõnikord see nii ka on. Siiski on enamus diskreetseid tunnuseid oma olemuselt polügeensed, nende avaldumine sõltub paljude geenide ja keskkonna koosmõjust, kuigi silmaga nähtav või registreeritav väärtus on sellest hoolimata diskreetne. Et saada üle diskreetsete tunnuste ja nende polügeense determineerituse vastuolust, on kasutusele võetud nö **alustunnuse** (*liability*) kontseptsioon. Selle kontseptsiooni kohaselt eeldatakse, et diskreetse tunnuse taga on tegelikult pidev otseselt mitte mõõdetav alustunnus. Alustunnus on see, mis on tegelikult mõjutatud nii geenidest kui ka keskkonnast ja mille väärtuste kujunemine on modelleeritav eelmises punktis toodud mudelite kohaselt. Reaalse diskreetse tunnuse väärtused kujunevad alustunnuse baasil nö **läviväärtuste** (*threshold*) läbi – kui alustunnuse väärtus ületab lävendi, omandab diskreetne fenotüüp järgmise väärtuse. Binaarse fenotüübi tarvis on kirjeldatud kontseptsioon esitatud joonisel 1 ja selle võib lühidalt võtta kokku mudeliga kujul

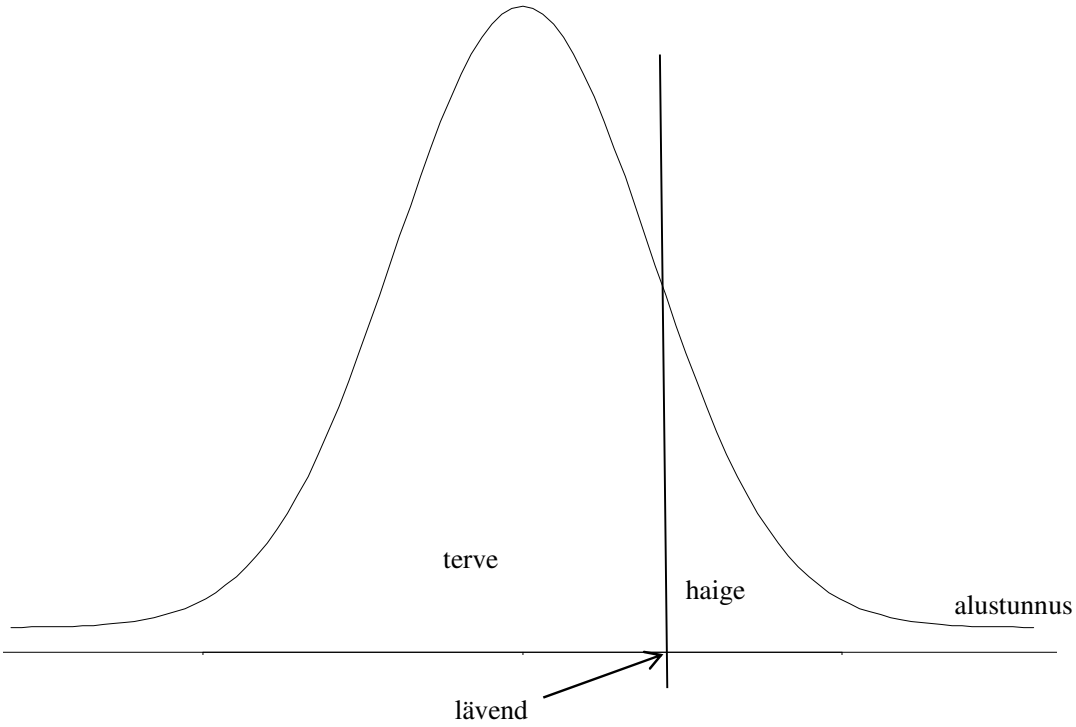
$$P = \begin{cases} 1, & L > \text{lävend}, \\ 0, & L \le \text{lävend}. \end{cases}, \tag{6}$$

Mudelis (6) märgib L alustunnuse väärtust ja P binaarset fenotüüpi. Nii loomade aretusväärtused kui ka päritavuskoefitsiendi väärtused eeldatakse tegelikult eksisteerivat alustunnuse skaalal, st, et alustunnuse väärtused eeldatakse esituvat klassikalise geneetilise mudeli kaudu kujul

$$L = \overline{L} + A_L + E_L,$$

kus A_L ja E_L on vastavalt aditiivgeneetiline efekt ja keskkonnaefekt alustunnuse skaalal. Muidugi eeldatakse siis ka alustunnuse jaotumist vastavalt normaaljaotuse seaduspäradele, millest tulenevalt kujutavad lävendiväärtused enesest normaaljaotuse kvantiile ja mis statistiliste mudelite kontekstis viib loomulikul teel välja probit-mudeliteni.

Diskreetsete tunnuste väärtuseid kirjeldatud kontseptsioonist lähtuvalt modelleerivaid mudeleid nimetatakse **lävendi mudeliteks** (*threshold model*).



Joonis 1. Alustunnuse ja binaarse fenotüübi omavaheline seos.

2. Statistiline mudel

Populatsiooni geneetiliste parameetrite hindamisel kasutatakse suurt hulka erinevaid statistilisi mudeleid, millest enamus on liigitatavad üldiste või üldistatud lineaarsete mudelite alla. Iga üldise või üldistatud lineaarse mudeli püstitamine eeldab

a) tõenäosusjaotuslike eelduste tegemist uuritava tunnuse ja mudeli liikmete kohta,

b) mudeli esitust faktorite mõjude lineaarkombinatsioonina ning

c) uuritava tunnuse ning faktorite keskväärtuste ja dispersioonistruktuuri fikseerimist vastavalt andmestiku ülesehitusele, faktorite olemusele ja tehtud tõenäosusjaotuslikele eeldustele.

Järgnev ülevaade on koostatud materjalide McCulloch ja Searle (2001), Engel (1997), SAS Institute Inc. (2006) ja Kaart (2012) põhjal.

2.1. Üldine lineaarne mudel (LM)

Eeldame, et uuritava tunnuse väärtused mõõdetuna erinevatel objektidel on sõltumatud ja jaotuvad normaaljaotuse järgi:

$$y_i \sim N(\mu, \sigma_y^2)$$

ehk

$$\boldsymbol{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$$

kus $\boldsymbol{y}$ on mõõtmistulemuste $n\times1$ vektor: $\boldsymbol{y} = (y_1\ y_2\ \dots\ y_n)^{\mathrm{T}}$, y_i on mõõtmistulemus objektil i, $\boldsymbol{\mu}$ on mõõtmistulemuste keskväärtuste vektor: $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu}$, ja $\mathbf{V}$ on mõõtmistulemuste $n\times n$ dispersioonimaatriks: $\mathrm{var}(\boldsymbol{y}) = \mathbf{V}$.

Uuritavat tunnust potentsiaalselt mõjutavate faktorite kohta eeldame, et nende kõik tasemed on (või saavad põhimõtteliselt olla) andmetes esindatud, pakuvad iseseisvat huvi ja on valitud mittejuhuslikult. Sellisel juhul käsitletakse faktorite erinevate tasemete mõjusid tundmatute konstantidena ja faktoreid endid nimetatakse **fikseeritud faktoriteks**.

Näiteks uurides lehmade piimatoodangut kolmes erinevat tüüpi farmis, on loomulik käsitleda farmi mõju fikseerituna, kuna uuringu eesmärgiks ongi võrrelda just andmetes esindatud kolme farmitüüpi.

Üldine lineaarne mudel (inglise keeles *general linear model*), modelleerimaks mõõtmistulemusi fikseeritud faktorite lineaarkombinatsioonina, on esitatav kujul

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{7}$$

kus

$\boldsymbol{\beta}$ on fikseeritud faktorite mõjude $p\times1$ vektor ja p on fikseeritud faktorite tasemete arv,

$\mathbf{X}$ on igal objektil teostatud mõõtmistulemust temale vastavate faktorite tasemetega siduv plaani- ehk disainimaatriks, mis sisaldab üht rida iga uuritava objekti ja üht veergu iga fikseeritud efekti kohta (seega on tegu $n\times p$ maatriksiga),

$\boldsymbol{\varepsilon}$ on mudeli juhuslike vigade $n\times1$ vektor.

Üldise lineaarse mudeli alternatiivne esitus uuritava tunnuse keskväärtuse $\mathrm{E}(\boldsymbol{y})$ tarvis on kujul

$$\mathrm{E}(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta} \tag{8}$$

ning objekti i tarvis kujul

$$\mathrm{E}(y_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta},$$

kus y_i on uuritava tunnuse väärtus objektil i ja $\boldsymbol{x}_i^{\mathrm{T}}$ on fikseeritud faktoritele vastava plaanimaatriksi i. rida.

Viimased esitused tulenevad loomulikust eeldusest, et mudeli jääkide keskväärtus $\mathrm{E}(\varepsilon_i) = 0$.

Kuna erinevatel objektidel teostatud mõõtmised on eelduse kohaselt sõltumatud ning fikseeritud faktorite mõjude näol on tegu konstantidega, mille dispersioon on 0, siis

$$\mathrm{var}(y_i) = \sigma_y^2 = \sigma_\varepsilon^2 \text{ ja } \mathrm{cov}(y_i, y_j) = 0$$

ning mõõtmistulemuste dispersioonimaatriks $\mathbf{V}$ on diagonaalmaatriks:

$$\mathrm{var}(\boldsymbol{y}) = \mathbf{V} = \sigma_y^2\mathbf{I}.$$

2.2. Üldine lineaarne segamudel (LMM)

Juhul, kui faktoril on potentsiaalselt väga palju (lõpmatu hulk) tasemeid ja andmetes on neist tasemetest esindatud juhuslik valim, nimetatakse faktorit **juhuslikuks faktoriks**. Andmetes esindatud juhusliku faktori tasemete mõjusid käsitletakse kui mingi teoreetilise jaotusega juhuslike suuruste realiseerunud väärtuseid, kusjuures selle teoreetilise jaotuse all mõistetakse üldjuhul normaaljaotust.

Näiteks uurides kolme erineva pulli järglaste piimatoodangut ja mõistes pulli mõju all temalt järglastele pärandunud geenikomplekti mõju, on loomulik käsitleda pulli juhusliku faktorina – iga pull saab oma järglastele pärandada suure hulga erinevaid geenikombinatsioone, andmetes esindatud järglastele pärandunud geenikombinatsioonide mõjud kujutavad enesest juhuslikku valikut (juhuslikku realisatsiooni) kõikvõimalike geenikombinatsioonide mõjudest.

Kui mudel sisaldab nii fikseeritud kui ka juhuslikke faktoreid, nimetatakse mudelit üldiseks lineaarseks segamudeliks (*general linear mixed model*) ja esitatakse kujul

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{9}$$

kus

$\boldsymbol{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$ on mõõtmistulemuste $n\times1$ vektor: $\boldsymbol{y} = (y_1\ y_2\ \dots\ y_n)^{\mathrm{T}}$, y_i on mõõtmistulemus objektil i, $\boldsymbol{\mu}$ on mõõtmistulemuste keskväärtuste vektor: $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu}$, ja $\mathbf{V}$ on mõõtmistulemuste $n\times n$ dispersioonimaatriks: $\mathrm{var}(\boldsymbol{y}) = \mathbf{V}$;

$\boldsymbol{\beta}$ on fikseeritud faktorite mõjude $p\times1$ vektor (p on fikseeritud faktorite tasemete arv), $\mathbf{X}$ on igal objektil teostatud mõõtmistulemust temale vastavate fikseeritud faktorite tasemetega siduv $n\times p$ plaani- ehk disainimaatriks;

$\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{G})$ on juhuslike faktorite mõjude $r\times1$ vektor (r on juhuslike faktorite tasemete arv) ja $\mathbf{Z}$ on igal objektil teostatud mõõtmistulemust temale vastavate juhuslike faktorite tasemetega siduv $n\times r$ plaani- ehk disainimaatriks,

$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$ on mudeli juhuslike vigade $n\times1$ vektor.

Kuna eelduste kohaselt $\mathrm{E}(\boldsymbol{u}) = \mathbf{0}$ ja $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, siis

$$\mathrm{E}(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta}.$$

Juhuslike efektide ja mudeli juhuslike vigade sõltumatuse eeldusest tuleneb ka, et

$$\mathrm{var}\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{pmatrix} = \begin{pmatrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} \end{pmatrix}$$

ning

$$\mathrm{var}(\boldsymbol{y}) = \mathbf{ZGZ}^{\mathrm{T}} + \mathbf{R}.$$

Seega sõltub fikseeritud efektidest vaid uuritava tunnuse keskväärtus ja juhuslikest efektidest vaid uuritava tunnuse dispersioon:

$$\boldsymbol{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZGZ}^{\mathrm{T}} + \mathbf{R}).$$

Kui nii erinevatel objektidel kui ka erinevatel juhuslike faktorite tasemetel sooritatud mõõtmised on sõltumatud, on dispersioonimaatriksid **R** ja **G** diagonaalmaatriksid kujul

$$\mathbf{R} = \mathbf{I}_n \sigma_\varepsilon^2$$

Ja

$$\mathbf{G} = \begin{pmatrix} \mathbf{I}_{n1}\sigma_1^2 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{I}_{rs}\sigma_s^2 \end{pmatrix}$$

kus r_i on juhusliku faktori u_i tasemete arv, $r = \sum_{i=1}^{s} r_i$

$$\mathrm{var}(\mathbf{y}) = \mathbf{ZGZ}^{\mathrm{T}} + \mathbf{R} = \sum_{i=1}^{s} \mathbf{Z}_i \mathbf{Z}_i^{\mathrm{T}} \sigma_i^2 + \mathbf{I}_n \sigma_\varepsilon^2 .$$

Tähistades $\mathbf{V}_0 = \mathbf{I}_{\mathrm{n}}$, $\sigma_0^2 = \sigma_\varepsilon^2$ ja $\mathbf{V}_i = \mathbf{Z}_i \mathbf{Z}_i^{\mathrm{T}}$ saab uuritava tunnuse dispersiooni kirjutada välja kujul

$$\mathbf{V} = \mathrm{var}(\boldsymbol{y}) = \sum_{i=0}^{s} \mathbf{V}_i \sigma_i^2 ,$$

millest

$$\mathrm{var}(y_i) = \sigma_y^2 = \sigma_u^2 + \sigma_\varepsilon^2.$$

Sarnaselt üldisele lineaarsele mudelile esitatakse ka üldine lineaarne segamudel sageli hoopis uuritava tunnuse keskväärtuse kaudu. Seejuures peab arvestama, et kuigi juhuslike efektide

keskväärtuse nulliga võrdumise tõttu sõltub uuritava tunnuse keskväärtus vaid fikseeritud efektidest, on mudeli seisukohalt oluline võtta arvesse ka juhuslike efektide realiseerunud väärtuseid. St, et üldine lineaarne segamudel on defineeritav ka kui uuritava tunnuse $\boldsymbol{y}$ tinglik keskväärtus juhusliku suuruse $\mathbf{U}$ realiseerunud väärtuste $\boldsymbol{u}$ korral:

$$\mathrm{E}(\boldsymbol{y}|\boldsymbol{u}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u}. \qquad (10)$$

Esitus $\mathrm{E}(\boldsymbol{y}|\boldsymbol{u})$ on lihtsalt lühem vorm esitusest $\mathrm{E}(\boldsymbol{y}|\mathbf{U}=\boldsymbol{u})$.

Sarnaselt vaatluste vektori $\boldsymbol{y}$ tinglikule keskväärtusele on esitatav ka tinglik dispersioon:

$$\mathrm{var}(\boldsymbol{y}|\boldsymbol{u}) = \mathrm{var}(\boldsymbol{\varepsilon}) = \mathbf{R}.$$

Objekti i tarvis on üldine lineaarne segamudel esitatav kujul

$$\mathrm{E}(y_i|\boldsymbol{u}) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u},$$

kus y_i on uuritava tunnuse väärtus objektil i, $\boldsymbol{x}_i^{\mathrm{T}}$ on fikseeritud faktoritele vastava plaanimaatriksi i. rida ja $\boldsymbol{z}_i^{\mathrm{T}}$ on juhuslikele faktoritele vastava plaanimaatriksi i. rida.

Kaks peamist lineaarsete mudelite omadust, mis kehtivad sõltumata juhuslike faktorite olemasolust, on järgmised (Engel, 1997):

1) vaatluste y_i keskväärtused ja tinglikud keskväärtused esituvad faktorite mõjude lineaarkombinatsioonina:

$$\mathrm{E}(y_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$$

ja

$$\mathrm{E}(y_i|\boldsymbol{u}) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u};$$

2) mudeli parameetrite dispersioonid ja kovariatsioonid on keskväärtustest sõltumatud, st et näiteks ühe juhusliku faktoriga mudeli korral

$$\mathrm{var}(y_i) = \sigma_u^2 + \sigma_\varepsilon^2$$

ja $\mathrm{cov}(y_i, y_j) = \sigma_u^2$, kui i. ja j. mõõtmine on sooritatud juhusliku faktori samal tasemel, ning $\mathrm{cov}(y_i, y_j) = 0$ muudel juhtudel.

Näide.

Võtame vaatluse alla järgmise andmetabeli, mis sisaldab kahe pulli kokku kaheksal järglasel teostatud teatud verenäitaja kontsentratsiooni mõõtmise tulemusi. Seejuures on teada ka, et kaheksast järglasest 4 olid pärit ühest ja 4 teisest farmist.

Loom	Isa	Farm	Tunnus
1	1	A	38,8
2	1	B	24,9
3	1	B	6,3
4	1	B	56,8
5	2	B	76,4
6	2	A	22,1
7	2	A	39,1
8	2	A	49,7

Mudel, hindamaks korraga nii fikseeritud faktori 'Farm' erinevate tasemete kui ka juhusliku faktori 'Isa' erinevate tasemete mõju, on kujul

$$y_{ijk} = \mu + H_i + s_j + \varepsilon_{ijk},$$

kus y_{ijk} on farmist i pärit j-nda isa k-nda järglane, H_i on i-nda farmi mõju, s_j j-nda isa mõju ja ε_{ijk} farmist i pärit j-nda isa k-nda järglase juhuslik viga.

Sama mudel on maatrikskujul järgmine:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{s} + \boldsymbol{\varepsilon},$$

ehk, kirjutades maatriksite tähistuste asemele nende sisu:

$$\begin{pmatrix} 38.8 \\ 24.9 \\ 6.3 \\ 56.8 \\ 76.4 \\ 22.1 \\ 39.1 \\ 49.7 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} \mu \\ H_1 \\ H_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,1,1} \\ \varepsilon_{2,1,2} \\ \varepsilon_{2,1,3} \\ \varepsilon_{2,1,4} \\ \varepsilon_{2,2,5} \\ \varepsilon_{1,2,6} \\ \varepsilon_{1,2,7} \\ \varepsilon_{1,2,8} \end{pmatrix}.$$

Mudeli kovariatsioonistruktuur on kirjeldatav maatriksitega

$$\mathbf{G} = \begin{pmatrix} \sigma_S^2 & 0 \\ 0 & \sigma_S^2 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 \end{pmatrix}$$

ja

$$\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} \sigma_S^2 & 0 \\ 0 & \sigma_S^2 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} + \sigma_\varepsilon^2 \cdot \mathbf{I}_8 =$$

$$= \begin{pmatrix} \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & 0 & 0 & 0 & 0 \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & 0 & 0 & 0 & 0 \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & 0 & 0 & 0 & 0 \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 \\ 0 & 0 & 0 & 0 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 \\ 0 & 0 & 0 & 0 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 \\ 0 & 0 & 0 & 0 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 \end{pmatrix} + \begin{pmatrix} \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_\varepsilon^2 \end{pmatrix} =$$

$$= \begin{pmatrix} \sigma_S^2 + \sigma_\varepsilon^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & 0 & 0 & 0 & 0 \\ \sigma_S^2 & \sigma_S^2 + \sigma_\varepsilon^2 & \sigma_S^2 & \sigma_S^2 & 0 & 0 & 0 & 0 \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 + \sigma_\varepsilon^2 & \sigma_S^2 & 0 & 0 & 0 & 0 \\ \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 + \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_S^2 + \sigma_\varepsilon^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 \\ 0 & 0 & 0 & 0 & \sigma_S^2 & \sigma_S^2 + \sigma_\varepsilon^2 & \sigma_S^2 & \sigma_S^2 \\ 0 & 0 & 0 & 0 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 + \sigma_\varepsilon^2 & \sigma_S^2 \\ 0 & 0 & 0 & 0 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 & \sigma_S^2 + \sigma_\varepsilon^2 \end{pmatrix}.$$

2.3. Üldistatud lineaarne mudel (GLM)

Eeldame, et uuritava tunnuse väärtused mõõdetuna erinevatel objektidel on sõltumatud ja nende jaotus kuulub eksponentsiaalsete jaotuste perre.

Kui uuritava tunnuse väärtused y_i ei jaotu normaaljaotuse kohaselt, ei ole uuritava tunnuse keskväärtus

$$\mu_i = \mathrm{E}(y_i)$$

avaldatav faktorite mõjude lineaarkombinatsioonina. Küll aga on faktorite mõjude lineaarkombinatsioonina esitatav mingi monotoonne funktsioon uuritava tunnuse keskväärtusest $\mathrm{g}(\mu_i)$. Taolist funktsiooni $g(.)$ nimetatakse **seosefunktsiooniks** ja üldistatud lineaarne mudel (*generalized linear model*) esitatakse kujul:

$$\mathrm{g}(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \tag{11}$$

kus $\boldsymbol{\beta}$ on fikseeritud faktorite mõjude $p\times 1$ vektor, p on fikseeritud faktorite tasemete arv, ja $\boldsymbol{x}_i^{\mathrm{T}}$ on fikseeritud faktoritele vastava plaanimaatriksi i. rida.

Kuna seosefunktsioon $g(.)$ on mingi mittelineaarne funktsioon (va identsusseose $\eta = \mu$ korral), on üldistatud lineaarsete mudelite näol tegu mittelineaarse seosega uuritava tunnuse ja argumenttunnuste vahel.

Tavaliselt tähistatakse seosefunktsiooni väärtust kohal μ_i tähega η_i:

$$\eta_i = g(\mu_i)$$

ehk maatrikskujul

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}).$$

Seega kujutab üldistatud lineaarne mudel $\boldsymbol{\mu}$ tarvis enesest tavalist lineaarset mudelit $\boldsymbol{\eta}$ tarvis.

Juhul, kui uuritav tunnus on binaarne (0/1-tunnus), st Bernoulli jaotusega: $y_i \sim \mathrm{Be}(p_i)$, kus $p_i \equiv \mu_i = \mathrm{P}(y_i{=}1)$ on meid huvitava sündmuse toimumise tõenäosus i. objektil, on levinuim seose-funktsioon **logit-funktsioon** kujul:

$$\eta_i = \mathrm{logit}(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right),$$

millest

$$\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{-\eta_i}}. \qquad (12)$$

Alternatiivne binaarsete tunnuste modelleerimisel kasutatav seosefunktsioon on **probit-funktsioon**, so standardse normaaljaotuse jaotusfunktsioon Φ:

$$\eta_i = \Phi(\mu_i),$$

millest

$$\mu_i = \Phi^{-1}(\eta_i). \qquad (13)$$

Nii logit- kui ka probit-funktsioon garanteerivad prognoositavate tõenäosuste $\mu_i = P(y_i=1)$ jäämise vahemikku (0, 1).

Üldistatud lineaarsete mudelite puhul ei kehti ka lineaarsete mudelite teine, dispersioonide kohta käiv omadus, sest enamuste eksponentsiaalsete jaotuste perre kuuluvate jaotuste puhul ei ole dispersioon ja keskväärtus sõltumatud. Sestap esitataksegi üldistatud lineaarsete mudelite korral vaatluste dispersioon $\text{var}(y_i)$ mingi keskväärtusest sõltuva funktsioonina, nn **dispersioonifunktsioonina**:

$$\text{var}(y_i) = V(\mu_i).$$

Dispersioonifunktsiooni kuju sõltub üksnes uuritava tunnuse jaotusest. Näiteks binaarse, Bernoulli jaotusega tunnuse korral

$$V(\mu_i) = \mu_i(1 - \mu_i).$$

Kuna mitte alati ei ole võimalik esitada vaatluste dispersiooni jaotusele vastava dispersioonifunktsioonina, lisatakse dispersioonifunktsioonile nn **skaalaparameeter** φ ning dispersioonifunktsioon saab kuju

$$\text{var}(y_i) = \varphi V(\mu_i).$$

Lisatud parameeter võimaldab esitada vaatluste dispersiooni keskväärtusest sõltuva funktsioonina ka võimaliku üle- või aladispersiooni (*over-* ja *under-dispersion*) puhul. Seega avaldub binaarse tunnuse dispersioon kujul

$$\mathrm{var}(y_i) = \varphi \times \mu_i(1 - \mu_i), \tag{14}$$

kusjuures enamasti eeldatakse, et $\varphi = 1$.

Oluline on märkida ka, et kuna $\mu_i = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ väärtused on iga faktorite tasemete kombinatsiooni puhul erinevad, on erinevalt üldisest lineaarsest mudelist iga faktorite tasemete kombinatsiooni puhul erinev ka uuritava tunnuse dispersioon.

2.4. Üldistatud lineaarne segamudel (GLMM)

Nii nagu üldine lineaarne segamudel saadakse tavalisest üldisest lineaarsest mudelist juhuslikele faktoritele vastavate efektide lisamise teel, saadakse ka üldistatud lineaarne segamudel (*generalized linear mixed model*) tavalisest üldistatud lineaarsest mudelist juhuslikele faktoritele vastavate efektide lisamise teel.

Tähistades uuritava tunnuse y tingliku keskväärtuse

$$\mathrm{E}(y_i|\boldsymbol{u}) = \mu_i,$$

on üldistatud lineaarne segamudel esitamaks uuritavat tunnust fikseeritud ja juhuslike faktorite kaudu kujul

$$\eta_i = \mathrm{g}(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}, \tag{15}$$

kus $\boldsymbol{x}_i^{\mathrm{T}}$ on fikseeritud faktoritele vastava plaanimaatriksi i. rida, $\boldsymbol{\beta}$ on fikseeritud faktorite mõjude $p\times 1$ vektor (p on fikseeritud faktorite tasemete arv), $\boldsymbol{z}_i^{\mathrm{T}}$ on juhuslikele faktoritele vastava plaanimaatriksi i. rida ja $\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{G})$ on juhuslike faktorite mõjude $r\times 1$ vektor (r on juhuslike faktorite tasemete arv).

Binaarse tunnuse puhul kasutatava logit-seosefunktsiooni korral on vastav üldistatud lineaarne segamudel kujul:

$$\mathrm{logit}(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_i^{\mathrm{T}}\mathbf{u}\,,$$

millest

$$\mu_i = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\mathbf{x_i^T\beta}-\mathbf{z_i^T u}}}; \tag{16}$$

ja probit-seosefunktsiooni korral kujul:

$$\eta_i = \Phi(\mu_i),$$

millest

$$\mu_i = \Phi^{-1}(\eta_i) = \Phi^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{u}), \tag{17}$$

kus $\Phi(.)$ on standardse normaaljaotuse jaotusfunktsioon.

Analoogselt üldistatud lineaarsele mudelile avaldub ka üldistatud lineaarse mudeli puhul uuritava tunnuse dispersioon enamasti mingi funktsioonina keskväärtusest, ainult tegu on juhuslike faktorite realiseerunud väärtuste suhtes tingliku dispersiooniga:

$$\mathrm{var}(y_i|\boldsymbol{u}) = \varphi \mathrm{V}(\mu_i),$$

siis φ tähistab jällegi skaalaparameetrit.

Binaarse tunnuse puhul

$$\mathrm{var}(y_i|\boldsymbol{u}) = \varphi \times \mu_i(1-\mu_i), \tag{18}$$

kusjuures vaikimisi $\varphi = 1$.

2.5. Lävendimudel

Binaarse lävendimudeli kohaselt omandab uuritav tunnus y väärtuse 1, kui alustunnuse w väärtus on suurem mingist piirist λ, ja väärtuse 0 vastupidisel juhul. Üldsust kitsendamata võib eeldada, et $\lambda = 0$. Olgu η alustunnuse w tinglik keskväärtus juhusliku faktori realiseerunud väärtuste $\mathbf{u}$ suhtes: $\eta = \mathrm{E}(w|\mathbf{u})$. Ning olgu mudeli jääkide jaotus alustunnuse skaalal tähistatud F-ga: $w - \eta = \varepsilon \sim F$. Siis on uuritava tunnuse tinglik keskväärtus μ esitatav kujul

$$\mu = \mathrm{P}(y=1|\mathbf{u}) = \mathrm{P}(w>0|\mathbf{u}) = 1 - \mathrm{P}(w \le 0|\mathbf{u}) = 1 - \mathrm{P}(w-\eta \le -\eta|\mathbf{u}) = 1 - F(-\eta),$$

millest järeldub omakorda, et

$$\eta = -F^{-1}(1-\mu).$$

Kuna üldistatud lineaarse segamudeli korral $\eta = g(\mu)$, siis kujutab lävendimudel enesest üldistatud lineaarset segamudelit seosefunktsiooniga $g(\mu) = -F^{-1}(1-\mu)$. Kui ε-i jaotus on sümmeetriline, siis viimane valem lihtsustub, saades kuju $g(\mu) = F^{-1}(\mu)$.

Tõenäosused μ ei muutu, kui korrutada alustunnuse w väärtuseid mingi positiivse konstandiga. Sestap võib mudeli jääkide dispersiooni σ_ε^2 fikseerida mistahes sobiva konstantse väärtusena.

Kui jääkide jaotus on standardne normaaljaotus, siis $g(\mu) = \Phi^{-1}(\mu)$ – seosefunktsiooniks on probit-funktsioon ja jääkvarieeruvus $\sigma_\varepsilon^2 = 1$.

Kui jääkide jaotus on standardne logistiline jaotus, st $F(\varepsilon) = 1/(1+e^{-\varepsilon})$, siis on seose-funktsiooniks on logit-funktsioon: $g(\mu) = \text{logit}(\mu)$, ja jääkvarieeruvus $\sigma_\varepsilon^2 = \pi^2/3$.

Seega – hinnates binaarse tunnuse päritavust mudelist, kus geneetilise varieeruvuse allikas on isa, avaldub päritavuskoefitsiendi (5) hinnang probit-seosefunktsiooniga üldistatud lineaarse segamudeli puhul valemist

$$h^2 = \frac{4\sigma_s^2}{\sigma_s^2 + 1} \tag{19}$$

ja logit-seosefunktsiooniga üldistatud lineaarse segamudeli puhul valemist

$$h^2 = \frac{4\sigma_s^2}{\sigma_s^2 + \pi^2/3}, \tag{20}$$

σ_s^2 on juhuslike isamõjude dispersioon.

2.6. Üldistatud lineaarne segamudel paketis SAS

Võtame vaatluse alla peatüki 2.2 lõpus esitatud näiteandmestiku ja oletame täiendavalt, et verenäitaja y väärtuste alusel on loom klassifitseeritav haigeks, kui $y > 40$. Sellest lähtuvalt saame lisada andmestikku binaarse tunnuse π, mille väärtuseks on 1 haige looma ja 0 terve looma puhul.

Täiendatud andmestik on esitatud järgmises tabelis.

Loom	Isa	Farm	Tunnus	Binaarne tunnus
1	1	A	38,8	0
2	1	B	24,9	0
3	1	B	6,3	0
4	1	B	56,8	1
5	2	B	76,4	1
6	2	A	22,1	0
7	2	A	39,1	0
8	2	A	49,7	1

Eesmärgiks on hinnata farmi ja isa mõju haigestumisele, kusjuures isa mõju tuleks käsitleda juhuslikuna.

Kuna uuritava tunnuse näol on tegu binaarse tunnusega, on sobivateks seosefunktsioonideks logit- ja probit-seosefunktsioonid. Seega on mudel esitatav kujul

$$\text{logit}(\pi_{ijk}) = \mu + H_i + s_j + \varepsilon_{ijk}$$

või

$$\Phi^{-1}(\pi_{ijk}) = \mu + H_i + s_j + \varepsilon_{ijk},$$

kus π_{ijk} on farmist i pärit j. isa järglasena sündinud k. looma haigestumus, H_i on i-nda farmi fikseeritud mõju, s_j on isa j juhuslik efekt, $s_j \sim N(0, \sigma_s^2)$, ε_{ijk} on farmist i pärit j. isa järglasena sündinud k. looma omapära, $i = 1, 2$, $j = 1, 2$, $k = 1, \ldots, 8$.

Fikseeritud efektide μ, H_1 ja H_2, juhuslike efektide s_1 ja s_2 ning isamõjude dispersiooni σ_s^2 hindamiseks võib SAS-s kasutada protseduuri GLIMMIX või protseduuri NLMIXED. Neist esimene on loodud spetsiaalselt üldistatud lineaarsete segamudelite hindamiseks, teine võimaldab hinnata mittelineaarsete segamudelite parameetreid, mistap on see protseduur kohandatav ka üldistatud lineaarsete segamudelite tarvis.

Esmalt tuleb analüüsitav andmestik lugeda SAS-i sisse (joonis 2).

```
data naide;
input loom isa farm$ y biny;
datalines;
1 1 A 38.8 0
2 1 B 24.9 0
3 1 B 6.3 0
4 1 B 56.8 1
5 2 B 76.4 1
6 2 A 22.1 0
7 2 A 39.1 0
8 2 A 49.7 1
;
```

Joonis 2. Näiteandmestiku sisselugemine SAS-i.

Protseduuri GLIMMIX süntaks on sarnane teiste SAS-s leiduvate üldiste ja üldistatud lineaarsete mudelite analüüsimiseks mõeldud protseduuridega. Lauses CLASS tuleb ära nimetada kõik klassifitseerivad faktorid, juhuslikud faktorid määratakse lauses RANDOM. Lauses MODEL tuleb defineerida hinnatav mudel, seejuures tuleb ära tuua vaid fikseeritud faktorid, uuritava tunnuse järel saab käsuga REF määrata, millise uuritava tunnuse taseme suhtes mudel hinnatakse. Lause MODEL lisavalikuna peab ette andma ka uuritava tunnuse jaotuse ja vastava seosefunktsiooni. Lausete MODEL ja RANDOM lisavalikud SOLUTION tellivad väljundisse ka fikseeritud ja juhuslike faktorite tasemete mõjude hinnangud. Näiteandmetele rakendatud protseduurid on esitatud joonistel 3 ja 4, esimene neist kasutab probit-seosefunktsiooni ja teine logit-seosefunktsiooni.

```
proc glimmix data=naide;
class isa farm;
model biny (ref=first)= farm / dist=binary link=probit solution;
random isa / solution;
run;
```

Joonis 3. GLIMMIX protseduur kasutades probit-seosefunktsiooni.

```
proc glimmix data=naide;
class isa farm;
model biny (ref=first) = farm / dist=binary link=logit solution;
random isa / solution;
run;
```

Joonis 4. GLIMMIX protseduur kasutades logit-seosefunktsiooni.

Analüüside tulemusena väljastatakse esmalt mudeli üldinformatsioon. Joonistelt 5 ja 6 on näha, et andmestiku nimi, mida analüüsil kasutatakse, on *naide* ja et uuritavaks tunnuseks on *biny*. Mudeli üldinformatsioonis tuuakse ära, millisest jaotusest eeldatakse uuritav tunnus olevat (*Response Distribution*) ja millist seosefunktsiooni (*Link Function*) ning dispersioonifunktsiooni (*Variance Function*) kasutatakse. Nii probit- kui ka logit-seosefunktsiooniga mudeli korral eeldatakse binaarset uuritavat tunnust.

The GLIMMIX Procedure

Model Information

Data Set	WORK.NAIDE
Response Variable	biny
Response Distribution	Binary
Link Function	Probit
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Residual PL
Degrees of Freedom Method	Containment

Joonis 5. Probit-seosefunktsiooniga mudeli üldinformatsioon

The GLIMMIX Procedure

Model Information

Data Set	WORK.NAIDE
Response Variable	biny
Response Distribution	Binary
Link Function	Logit
Variance Function	Default
Variance Matrix	Not blocked
Estimation Technique	Residual PL
Degrees of Freedom Method	Containment

Joonis 6. Logit-seosefunktsiooniga mudeli üldinformatsioon

Järgmisena kuvatakse väljundis nii probit- kui ka logit-seosefunktsiooniga mudeli korral seletavate faktortunnuste kohta käiv info, mitu taset ja millised need on (joonis 7). Tuuakse ära ka vaatluste arv kokku ja uuritava tunnuse erinevatel tasemetel – antud näite korral on uuritava tunnuse väärtusteks ’1’ ja ’2’, ning ’1’ esineb andmestikus viis korda ja ’2’ kolm korda, kokku on kaheksa vaatlust. Märgitakse ära ka uuritava tunnuse see tase, mille suhtes mudeli parameetrid hinnatakse (praegu ’1’). Lisaks on kirjas maatrikskujul oleva mudeli

liikmete dimensioonid. *G-side* kovariatsioon tähistab juhuslike faktorite kovariatsiooni, juhusliku vea kovariatsiooni nimetatakse *R-side* kovariatsiooniks.

Class Level Information

Class	Levels	Values
isa	2	1 2
farm	2	A B

Number of Observations Read	8
Number of Observations Used	8

Response Profile

Ordered Value	biny	Total Frequency
1	0	5
2	1	3

The GLIMMIX procedure is modeling the probability that biny='1'.

Dimensions

G-side Cov. Parameters	1
Columns in X	3
Columns in Z	2
Subjects (Blocks in V)	1
Max Obs per Subject	8

Joonis 7. Faktortunnused ja nende tasemed, samasugused nii probit- kui ka logit-seose-funktsiooniga mudelite korral.

Edasi tuleb info optimiseerimise kohta (joonis 8), algul üldinfo, millist optimiseerimis protsessi kasutatakse ja mitu parameetrit on vaja optimiseerida. Edasi tuleb iteratsiooni sammude info ja lõpuks info, kas protsess koondus või mitte. Mõlema mudeli korral kasutatakse Newton Raphsoni meetodit, see on vaikeväärtuseks, kui uuritav tunnus on binaarne. Ise saab optimiseerimismeetodit valida NLOPTIONS käsuga MODEL lause järel. Praegusel juhul on vaja optimiseerida üle ühe parameetri. Mõlemad optimiseerimis protsessid koondusid, probit-seosefunktsiooniga mudel peale 14 sammu ja logit -seosefunktsiooniga peale 8 sammu, kuna väljatrükid on küllaltki sarnased, siis on siinkohal väljatoodud vaid probit-seosefunktsiooniga mudeli väljatrükk (joonis 8).

Optimization Information

Optimization Technique	Newton-Raphson with Ridging
Parameters in Optimization	1
Lower Boundaries	1
Upper Boundaries	0
Fixed Effects	Profiled
Starting From	Data

Iteration History

Iteration	Restarts	Subiterations	Objective Function	Change	Max Gradient
0	0	4	24.981833247	0.34504192	5.115E-6
1	0	4	25.028828717	0.12914433	3.239E-7
2	0	3	25.108523356	0.05580130	7.883E-7
3	0	2	25.143200064	0.02355244	8.519E-6
4	0	2	25.157961971	0.00980828	6.024E-7
5	0	2	25.164129821	0.00407528	4.279E-8
6	0	1	25.166692836	0.00169446	1.748E-6
7	0	1	25.16775779	0.00069685	2.960E-7
8	0	1	25.168196935	0.00028808	5.060E-8
9	0	1	25.16837829	0.00011916	8.664E-9
10	0	1	25.168453275	0.00004930	1.488E-9
11	0	0	25.168484294	0.00000183	6.781E-6
12	0	0	25.16848784	0.00000016	7.370E-6
13	0	0	25.16848812	0.00000001	7.421E-6
14	0	0	25.168488144	0.00000000	7.425E-6

Convergence criterion (PCONV=1.11022E-8) satisfied.

Joonis 8. Iteratsiooni protsessi info probit-seosefunktsiooniga mudeli korral.

Peale optimiseerimisprotsessi koondumist väljastatakse mudeli sobivuse statistikud (joonised 9 ja 10). Kõigepealt viimase iteratsioonisammu järgne kahekordne negatiivne logaritmitud tõepärafunktsiooni väärtus, mis on probit-seosefunktsiooniga mudeli korral 25,17 ja logit-seosefunktsiooniga mudeli korral 31.30. Edasi tulevad üldistatud χ^2-statistiku väärtused. Probit-seosefunktsiooniga mudeli korral tuleb üldistatud χ^2-statistiku väärtuseks 6,92 ja logit-seosefunktsiooniga mudeli korral 6,91, mis on põhimõtteliselt sama. Üldistatud χ^2-statistik jagatis vabadusastmete arvuga on mõlema mudeli korral võrdne 1,15-ga. See suhe näitab allesjäänud varieeruvust andmestikus, mis peaks olema eelduse kohaselt 1

Fit Statistics

-2 Res Log Pseudo-Likelihood	25.17
Generalized Chi-Square	6.92
Gener. Chi-Square / DF	1.15

Joonis 9. Mudeli sobivuse statistikud probit-seosefunktsiooniga mudeli korral.

Fit Statistics	
-2 Res Log Pseudo-Likelihood	31.30
Generalized Chi-Square	6.91
Gener. Chi-Square / DF	1.15

Joonis 10. Mudeli sobivuse statistikud logit-seosefunktsiooniga mudeli korral.

Järgnevalt tuuakse ära juhuslike faktorite dispersiooni hinnangud koos standardvigadega (joonised 11 ja 12). Eelduse kohaselt oli vaja hinnata vaid üks dispersiooni maatriksi element σ_S^2, mille hinnanguks saame probit mudeli korral 0,797 ja logit mudeli korral 2.219. Need hinnangud on erinevad, kuna mudelite parameetrid on hinnatud erinevatel skaaladel. Seega esituvad probit-seosefunktsiooniga mudeli juhuslike efektide kovariatsioonimaatriksid kujul $\mathbf{G} = 0{,}797 \cdot \mathbf{I}_2$, $\mathbf{R} = \mathbf{I}_8$ ($\mathbf{I}_2$ ja $\mathbf{I}_8$ on vastava dimensiooniga ühikmaatriksid). Logit-seosefunktsiooniga mudeli korral on vastavad maatriksid aga kujul $\mathbf{G} = 2.219 \cdot \mathbf{I}_2$, $\mathbf{R} = \pi^2/3 \cdot \mathbf{I}_8$.

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
isa	0.7965	2.1889

Joonis 11. Juhusliku faktori dispersiooni hinnang probit-seosefunktsiooniga mudeli korral.

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
isa	2.2193	6.3486

Joonis 12. Juhusliku faktori dispersiooni hinnang logi-seosefunktsiooniga t mudeli korral.

Edasi tulevad fikseeritud parameetrite hinnangud ja nende olulisuse tõenäosused (joonised 13 ja 14). Mõlema seosefunktsiooni korral saame, et farmi mõju ei ole statistiliselt oluline, aga see on nii väikse andmestiku korral oodatav tulemus. Parameetrite hinnangud on mudelite puhul jällegi erinevad, sest tulemused on erinevatel skaaladel.

Solutions for Fixed Effects

Effect	farm	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		0.2478	0.9345	1	0.27	0.8350
farm	A	−1.2198	1.1045	5	−1.10	0.3197
farm	B	0	.	.	.	.

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
farm	1	5	1.22	0.3197

Joonis 13. Fikseeritud faktorite mõju hinnangud probit-seosefunktsiooniga mudeli korral.

Solutions for Fixed Effects

Effect	farm	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		0.3944	1.5511	1	0.25	0.8415
farm	A	−1.9662	1.8824	5	−1.04	0.3441
farm	B	0	.	.	.	.

Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
farm	1	5	1.09	0.3441

Joonis 14. Fikseeritud faktorite mõju hinnangud logit-seosefunktsiooniga mudeli korral.

Ning lõpuks väljastatakse juhuslike faktorite mõjude hinnangud koos olulisustõenäosustega (joonised 15 ja16). Sarnaselt fikseeritud efektide hinnangutega on ka juhuslike efektide hinnangud erinevad tingituna mudelite erinevatest seosefunktsioonidest. Juhusliku faktori "Isa 1" hinnang on probit-seosefunktsiooni korral -0,453 ja logit-seosefunktsiooni korral -0,740, seega on esimese isa korral tõenäosus haigestuda väiksem, kui teise isa korral, ja see seos kehtib mõlema mudeli korral.

Solution for Random Effects

Effect	isa	Estimate	Std Err Pred	DF	t Value	Pr > |t|
isa	1	-0.4527	0.7691	5	-0.59	0.5817
isa	2	0.4527	0.7691	5	0.59	0.5817

Joonis 15. Juhuslike faktorite mõjude hinnangud probit-seosefunktsiooniga mudeli korral.

Solution for Random Effects

Effect	isa	Estimate	Std Err Pred	DF	t Value	Pr > |t|
isa	1	-0.7401	1.2929	5	-0.57	0.5918
isa	2	0.7401	1.2929	5	0.57	0.5918

Joonis 16. Juhuslike faktorite mõjude hinnangud logit-seosefunktsiooniga mudeli korral.

Probit-seosefunktsiooniga mudelist hinnatud haigestumistõenäosused erinevates farmides on vastavalt

$$\hat{\pi}_A = \Phi(0,248 - 1,220) = 0,165 \text{ ja } \hat{\pi}_B = \Phi(0,248) = 0,598$$

ning logit-seosefunktsiooniga mudelist hinnatud haigestumistõenäosused erinevates farmides on vastavalt

$$\hat{\pi}_A = \frac{1}{1+\exp\{-0,394+1,966\}} = 0,172 \text{ ja } \hat{\pi}_B = \frac{1}{1+\exp\{-0,394\}} = 0,597 .$$

Tulemused on üsna sarnased.

Haigestumistõenäosused erinevates farmides saab lasta välja arvutada ka protseduuril GLIMMIX lausega

LSMEANS farm / CL ILINK;

Üldiselt väljastab lause LSMEANS tõenäosuse hinnangud logit või probit-skaalal, mõõdetud tunnuse skaalal, antud juhul siis tõenäosuse skaalal hinnanguid saab tellida lisavalikuga /ILINK, lause CL lisab tõenäosuste hinnangutele ka usalduspiirid. Joonisel 17 on esitatud vastava analüüsi tulemus probit-mudeli korral – nagu näha, on tulemused identsed eelnevalt käsitsi arvutatutega.

farm Least Squares Means

farm	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
A	-0.9720	0.9948	5	-0.98	0.3734	0.05	-3.5292	1.5852
B	0.2478	0.9345	5	0.27	0.8014	0.05	-2.1543	2.6500

farm Least Squares Means

farm	Mean	Standard Error Mean	Lower Mean	Upper Mean
A	0.1655	0.2475	0.000208	0.9435
B	0.5979	0.3615	0.01561	0.9960

Joonis 17. Keskmised haigestumise väärtused probit-skaalal (veerus *Estimate*) ja tõenäosuse skaalal (veerus *Mean*).

Vastavalt valemile (18) on arvutatavad uuritava tunnuse tinglikud dispersioonid tõenäosuse skaalal. Probit-seosefunktsiooniga mudelist saame, et farmis A ja 1. isa järglaste seas on uuritava tunnuse tinglik dispersioon juhuslike isamõjude vektori $\boldsymbol{s}$ suhtes

$$\mathrm{var}(y_{A,1}|\boldsymbol{s}) = \varphi \times \pi_{A,1}(1 - \pi_{A,1}) = 1 \times 0{,}077 \cdot (1 - 0{,}077) = 0{,}071;$$

farmis A ja 2. isa järglaste seas on uuritava tunnuse tinglik dispersioon

$$\mathrm{var}(y_{A,2}|\boldsymbol{s}) = \varphi \times \pi_{A,2}(1 - \pi_{A,2}) = 1 \times 0{,}302 \cdot (1 - 0{,}302) = 0{,}211;$$

farmis B ja 1. isa järglaste seas on uuritava tunnuse tinglik dispersioon

$$\mathrm{var}(y_{B,1}|\boldsymbol{s}) = \varphi \times \pi_{B,1}(1 - \pi_{B,1}) = 1 \times 0{,}419 \cdot (1 - 0{,}419) = 0{,}243$$

ja farmis B ja 2. isa järglaste seas on uuritava tunnuse tinglik dispersioon

$$\mathrm{var}(y_{B,2}|\boldsymbol{s}) = \varphi \times \pi_{B,2}(1 - \pi_{B,2}) = 1 \times 0{,}758 \cdot (1 - 0{,}758) = 0{,}183.$$

Logit-seosefunktsiooniga mudelist saame, et farmis A ja 1. isa järglaste seas on uuritava tunnuse tinglik dispersioon

$$\mathrm{var}(y_{A,1}|\boldsymbol{s}) = \varphi \times \pi_{A,1}(1 - \pi_{A,1}) = 1 \times 0{,}090 \cdot (1 - 0{,}090) = 0{,}082;$$

farmis A ja 2. isa järglaste seas on uuritava tunnuse tinglik dispersioon

$$\mathrm{var}(y_{A,2}|\boldsymbol{s}) = \varphi \times \pi_{A,2}(1 - \pi_{A,2}) = 1 \times 0{,}303 \cdot (1 - 0{,}303) = 0{,}211;$$

farmis B ja 1. isa järglaste seas on uuritava tunnuse tinglik dispersioon

$$\mathrm{var}(y_{B,1}|\boldsymbol{s}) = \varphi \times \pi_{B,1}(1 - \pi_{B,1}) = 1 \times 0{,}414 \cdot (1 - 0{,}414) = 0{,}243;$$

ja farmis B ja 2. isa järglaste seas on uuritava tunnuse tinglik dispersioon

$$\mathrm{var}(y_{B,2}|\boldsymbol{s}) = \varphi \times \pi_{B,2}(1 - \pi_{B,2}) = 1 \times 0{,}757 \cdot (1 - 0{,}757) = 0{,}184.$$

Juhuslike isamõjude varieeruvuse alusel saab nii probit kui ka logit-seosefunktsiooniga mudeleist hinnata päritavuskoefitsent väärtust. Võttes aluseks valemid (19) ja (20), tulevad päritavuskoefitsiendi hinnangud vastavalt probit- ja logit-mudelist

$$h^2_{\text{probit}} = \frac{4 \cdot 0{,}797}{0{,}797 + 1} = 1{,}773$$

ja

$$h^2_{\text{logit}} = \frac{4 \cdot 2{,}219}{2{,}219 + \pi^2/3} = 1{,}611\,.$$

Mõlemad hinnangud on võrdselt ebaõiged, sest päritavuskoefitsiendi väärtus ei saa kunagi olla ühest suurem. Antud juhul on ebaõigete hinnangute põhjuseks liiga väike andmestik.

3. Modelleerimiseksperiment

Käesoleva modelleerimiseksperimendi ülesandeks oli uurida binaarse tunnuse puhul juhuslike isamõjude, vastavate dispersioonikomponentide ja päritavuskoefitsiendi hinnangute täpsust ja sõltuvust hindamismudelist ning binaarse tunnuse väärtuste kujunemisest. Andmete genereerimisel võeti aluseks peatükis 1.2 käsitletud lävendimudel. Esmalt genereeriti alustunnuse väärtused lähtudes loengukonspektis Kaart ja Möls (2011) esitatud simulatsiooniprogrammist. Eeldati, et alustunnuse väärtused on määratud 100 geeni poolt ning see, kumb alleelidest isalt ja emalt järglasele pärandub, on juhuslik. Kokku genereeriti 20 isa 1000 järglase standardse normaaljaotuse järgi jaotuva alustunnuse väärtused, seejuures eeldati igal isal olevat võrdselt 50 järglast. Täiendavalt paigutati kõik järglased juhuslikult ühte kahest erineva diskreetse mõjuga farmist. Viimase etapina moodustati alustunnuse väärtuste alusel binaarse tunnuse väärtused, seejuures kasutati kolme erinevat lävendiväärtust, mis vastasid 10%-lisele, 50%-lisele ja 90%-lisele juhtude arvule binaarse tunnuse osas. Samuti muudeti üksikute alleelide mõjude varieeruvust nii, et alustunnuse päritavus oleks kas 0,2, 0,5 või 0,8. Igal lävendi ja päritavuskoefitsiendi väärtuste kombinatsiooni puhul tehti 1000 modelleerimist. Andmete genereerimisel kasutatud SAS-i programm on toodud lisas 1.

Genereeritud andmetele rakendati nelja erinevat mudelit: üldist lineaarset segamudelit, kus uuritavaks tunnuseks oli normaaljaotuse järgi jaotuv alustunnus, ning binaarse uuritava tunnusega üldist lineaarset segamudelit, logit-seosefunktsiooniga üldistatud lineaarset segamudelit ja probit-seosefunktsiooniga üldistatud lineaarset segamudelit. Kõigi mudelite puhul oli fikseeritud faktoriks „Farm“ ja juhuslikuks faktoriks „Isa“. Lisas 1 toodud programm sisaldab ka mudelite rakendamiseks ning tulemuste kokkukoondamiseks ja analüüsimiseks kasutatud käske.

Modelleerimiseksperimendi tulemused päritavuskoefitsiendi väärtuse 0,5 korral on toodud tabelis 1 ja joonistel 18, 19 ja 20.

Tabelist 1 on näha, et kõigi lävendite korral alahindab binaarsele tunnusele rakendatud üldine lineaarne segamudel päritavuskoefitsiendi väärtust, täpseima hinnangu annab binaarsele tunnusele rakendatud üldine lineaarne segamudel lävendi 0,5 korral. Kõikide lävendite korral annab täpsema hinnangu loomulikult pidevale alustunnusele rakendatud üldine lineaarne segamudel, aga praktikas pole võimalik binaarse tunnuse taga olevat pidevat alustunnust

mõõta, mistap on see mudel vaid teoreetiline. Üldistatud lineaarse segamudeli, kus on logit-seosefunktsioon, korral on päritavuskoefitsiendi väärtused äärmuslike lävendiväärtuste puhul pisut ülehinnatud, samas kui lävendi 0,5 korral on päritavuskoefitsiendi väärtused selgelt alahinnatud. Seega sobib logit-mudel päritavuse hindamiseks kõige halvemini olukorras, kus lävend on 0,5 ehk binaarse tunnuse väärtused on võrdtõenäolised. Probit-seosefunktsiooniga üldistatud lineaarne segamudel on stabiilselt hea kõigi lävendite korral.

Tabel 1. Keskmised (standardhälve) dispersiooniparameetrite ja päritavuskoefitsiendi h^2 väärtused erinevate mudelite ja lävendi väärtuste korral. Analüüsid baseeruvad 1000-l modelleerimisel tegeliku pideval skaalal (alustunnuse skaalal) päritavuse 0,5 korral.

	LMM	LMMb	GLMM_l	GLMM_p
		Lävend 0,1		
σ_s^2	50,037 (18.527)	0,004 (0,003)	0,497 (0,243)	0,136 (0,068)
σ_e^2	349.221 (15.657)	0,086 (0,013)	$\pi^2/3$	1
h^2	0,494 (0,159)	0,187 (0,103)	0,511 (0,215)	0,467 (0,203)
		Lävend 0,5		
σ_s^2	49.443 (18.443)	0,019 (0,007)	0,362 (0,150)	0,139 (0,056)
σ_e^2	350,366 (16.095)	0,229 (0,007)	$\pi^2/3$	1
h^2	0,489 (0,161)	0,313 (0,114)	0,391 (0,145)	0,481 (0,170)
		Lävend 0,9		
σ_s^2	50,152 (18.822)	0,004 (0,003)	0,490 (0,247)	0,134 (0,068)
σ_e^2	350,855 (16.035)	0,087 (0,012)	$\pi^2/3$	1
h^2	0,493 (0,161)	0,187 (0,102)	0,505 (0,220)	0,461 (0,205)

LMM – üldine lineaarne segamudel normaaljaotuse järgi jaotuva uuritava tunnusega,

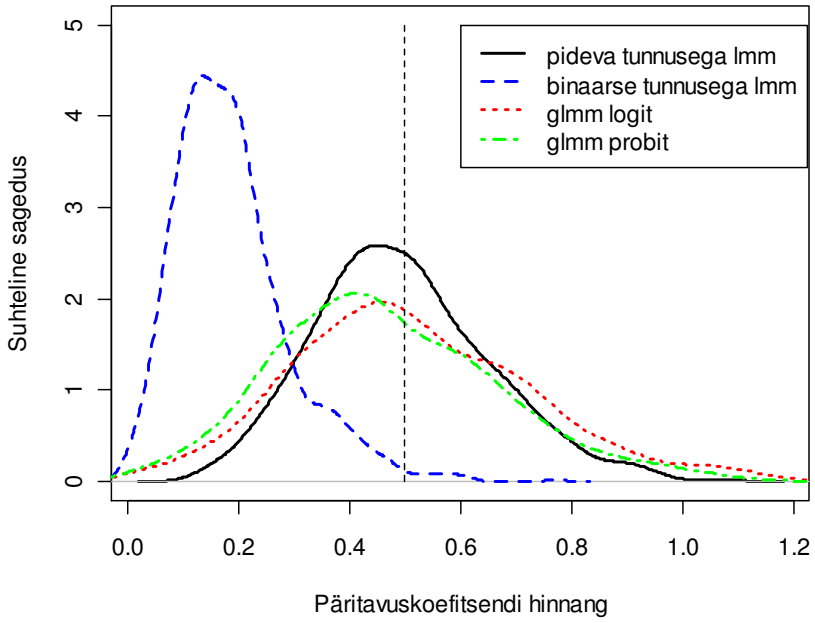
LMMb – üldine lineaarne segamudel binaarse uuritava tunnusega,

GLMM_l – üldistatud lineaarne segamudel logit-seosefunktsiooniga,

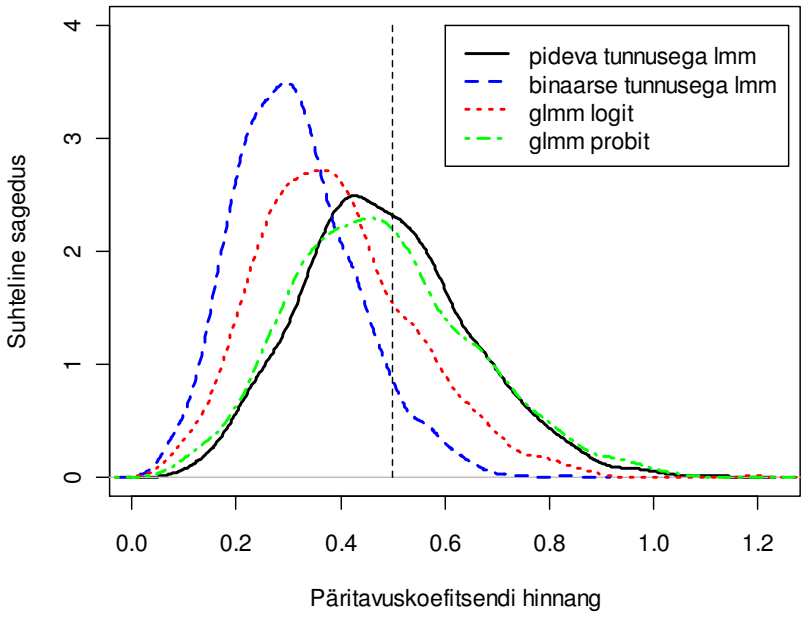
GLMM_p – üldistatud lineaarne segamudel probit-seosefunktsiooniga.

Joonistelt 18, 19 ja 20 on näha päritavuskoefitsiendi hinnangute paiknemise ebasümmeetrilisus, empiirilistel jaotustel on raske parempoolne saba. Joonistelt on näha, et binaarsele tunnusele rakendatud üldise lineaarse segamudeli korral saadud päritavus-koefitsiendi hinnangud on teiste meetoditega saadud hinnangutest väga erinevad ja väga

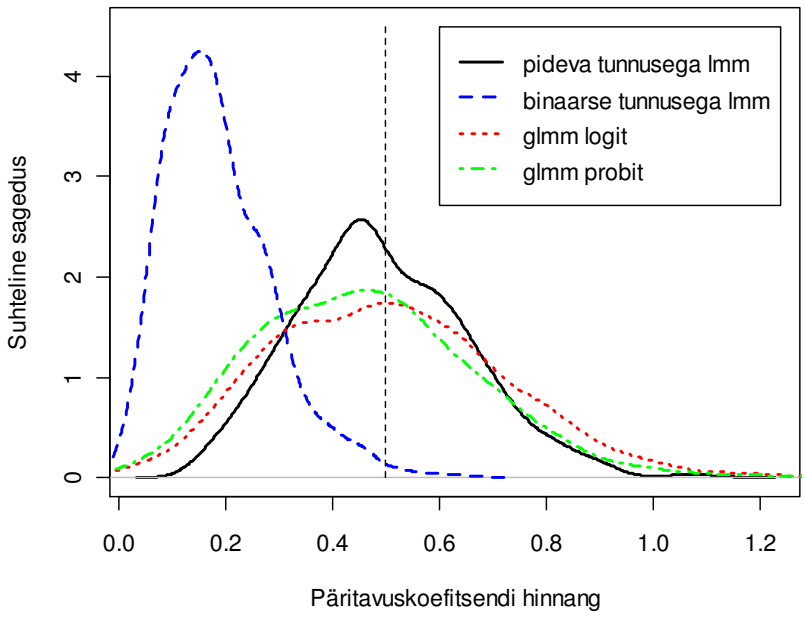
tugevalt alahinnatud. Pisut sagedamini esineb tegelikust väärtusest väiksemaid päritavuskoefitsiendi hinnanguid ka teiste mudelite puhul, aga nagu ilmnes ka tabelist 1, ei ole need erinevused kuigi suured. Ainsana eristuvad selgelt logit-seosefunktsiooniga mudelist leitud päritavuskoefitsiendi hinnangud lävendi 0,5 korral.



Joonis 18. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,5 ja lävendi 0,1 korral.



Joonis 19. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,5 ja lävendi 0,5 korral.



Joonis 20. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,5 ja lävendi 0,9 korral.

Tabelites 2 ja 3 ning joonistel 21-26 on näha, kuidas töötavad erinevad mudelid kui tegelik päritavus on 0,2 ja 0,8. Kõige täpsemaid hinnanguid annab ikka üldine lineaarne segamudel rakendatuna pidevale alustunnusele, aga kuna reaalsetes uuringutes pole võimalik alustunnust mõõta, siis ei ole võimalik reaalselt seda mudelit kasutada, küll on see hea võrdlusmaterjal hindamaks teiste mudelite headust. Üldine lineaarne segamudel rakendatuna binaarsele uuritavale tunnusele annab kõigi lävendite ja kõigi tegelike päritavuskoefitsiendi väärtuste korral alahinnangu, parima võimaliku lähendi saab selle mudeliga kui lävend on 0,5 aga ka siis on tegemist tugeva alahinnanguga päritavusele. Logit-seosefunktsiooniga üldistatud lineaarne segamudel annab äärmuslikemate lävendiväärtuste korral väga täpseid hinnanguid, aga lävendi 0,5 korral annab see tugeva alahinnangu, arvatavasti on põhjus selles, et logit-seosefunktsiooniga mudeli puhul eeldatakse lävendimudeli kontseptisooni kohaselt mudeli jääkide jaotumist standardse logistilise jaotuse seaduspärade kohaselt, mis aga lävendi 0,5 korral pole väga hästi täidetud. Probit-seosefunktsiooniga mudelist saadud hinnangute keskmised natuke alahindavad päritavuskoefitsienti, aga üldiselt on need kõige stabiilsemad üle erinevate lävendite ja päritavuste.

Tabel 2. Keskmised (standardhälve) dispersiooniparameetrite ja päritavuskoefitsiendi h^2 väärtused erinevate mudelite ja lävendi väärtuste korral. Analüüsid baseeruvad 1000-l modelleerimisel tegeliku pideval skaalal (alustunnuse skaalal) päritavuse 0,2 korral.

	LMM	LMMb	GLMM_l	GLMM_p
		Lävend 0,1		
σ_s^2	12,637 (5,689)	0,002 (0,001)	0,180 (0,129)	0,049 (0,036)
σ_ε^2	237,441 (10,692)	0,089 (0,010)	$\pi^2/3$	1
h^2	0,201 (0,087)	0,068 (0,051)	0,202 (0,137)	0,183 (0,126)
		Lävend 0,5		
σ_s^2	12,670 (5,739)	0,008 (0,004)	0,134 (0,070)	0,052 (0,027)
σ_ε^2	237,859 (10,631)	0,240 (0,004)	$\pi^2/3$	1
h^2	0,201 (0,087)	0,127 (0,06)	0,155 (0,078)	0,197 (0,098)
		Lävend 0,9		
σ_s^2	12,477 (5,685)	0,002 (0,001)	0,189 (0,128)	0,052 (0,035)
σ_ε^2	237,180 (10,611)	0,090 (0,010)	$\pi^2/3$	1
h^2	0,198 (0,087)	0,073 (0,052)	0,213 (0,136)	0,193 (0,125)

LMM – üldine lineaarne segamudel normaaljaotuse järgi jaotuva uuritava tunnusega,

LMMb – üldine lineaarne segamudel binaarse uuritava tunnusega,

GLMM_l – üldistatud lineaarne segamudel logit-seosefunktsiooniga,

GLMM_p – üldistatud lineaarne segamudel probit-seosefunktsiooniga.

Tabel 3. Keskmised (standardhälve) dispersiooniparameetrite ja päritavuskoefitsiendi h^2 väärtused erinevate mudelite ja lävendi väärtuste korral. Analüüsid baseeruvad 1000-l modelleerimisel tegeliku pideval skaalal (alustunnuse skaalal) päritavuse 0,8 korral.

	LMM	LMMb	GLMM_l	GLMM_p
		Lävend 0,1		
σ_s^2	202,935 (69,220)	0,007 (0,004)	0,850 (0,367)	0,230 (0,100)
σ_ε^2	801,964 (35,850)	0,083 (0,014)	$\pi^2/3$	1
h^2	0,794 (0,218)	0,308 (0,152)	0,797 (0,272)	0,727 (0,256)
		Lävend 0,5		
σ_s^2	200,010 (70,071)	0,032 (0,010)	0,651 (0,255)	0,246 (0,092)
σ_ε^2	801,538 (37,074)	0,218 (0,010)	$\pi^2/3$	1
h^2	0,785 (0,220)	0,508 (0,158)	0,647 (0,209)	0,772 (0,231)
		Lävend 0,9		
σ_s^2	202,500 (72,821)	0,007 (0,004)	0,849 (0,373)	0,230 (0,101)
σ_ε^2	797,956 (35,805)	0,083 (0,015)	$\pi^2/3$	1
h^2	0,795 (0,228)	0,312 (0,154)	0,796 (0,275)	0,726 (0,258)

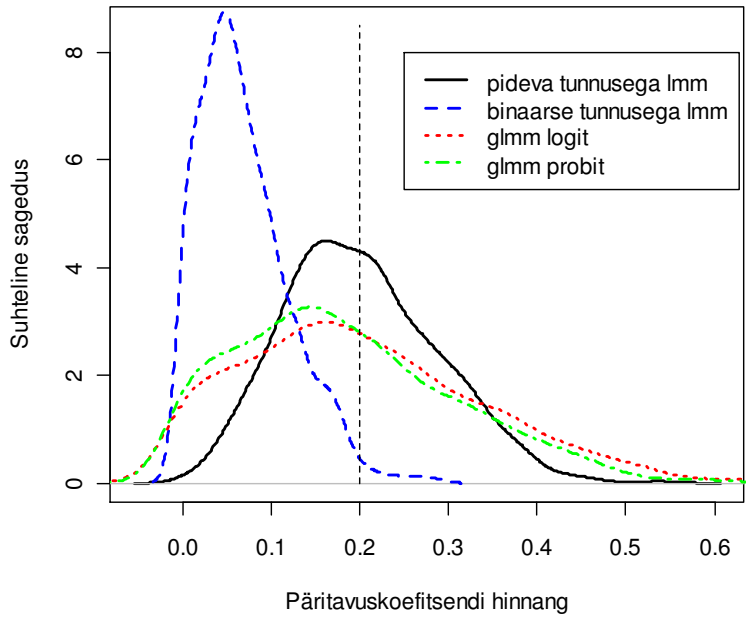
LMM – üldine lineaarne segamudel normaaljaotuse järgi jaotuva uuritava tunnusega,

LMMb – üldine lineaarne segamudel binaarse uuritava tunnusega,

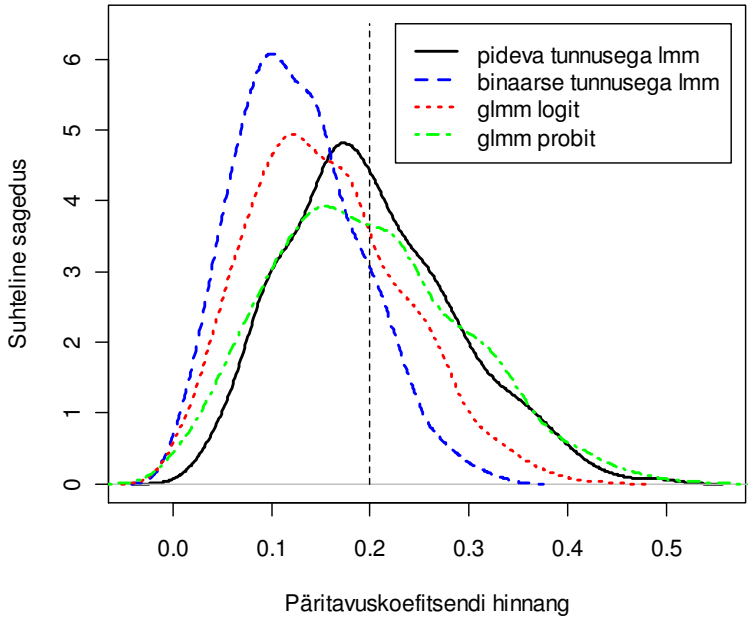
GLMM_l – üldistatud lineaarne segamudel logit-seosefunktsiooniga,

GLMM_p – üldistatud lineaarne segamudel probit-seosefunktsiooniga.

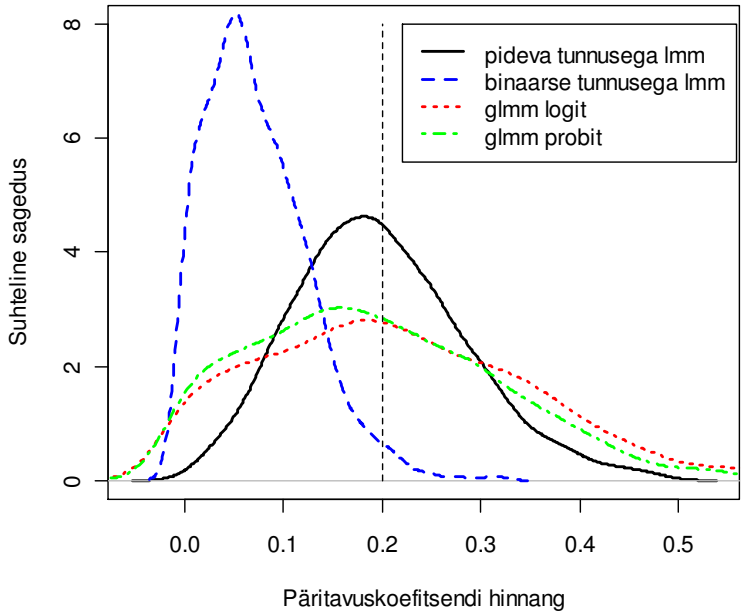
Joonistel 21-23 on ära toodud päritavuskoefitsiendi empiirilised jaotused lävendite 0,1, 0,5 ja 0,9 ning tegeliku päritavuse 0,2 korral. Nagu tegeliku päritavuse 0,5 korral, annab ka siin binaarsele tunnusele rakendatud üldine lineaarne segamudel kõikide lävendite korral alahinnangu päritavusele. Logit- ja probit-seosefunktsioonidega üldistatud lineaarsete segamudelite hinnangute empiirilistes jaotustes on näha suuremat hajuvust, kui tegeliku päritavuse 0,5 juures, kuid jaotused on endiselt parempoolse sabaga.



Joonis 21. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,2 ja lävendi 0,1 korral.



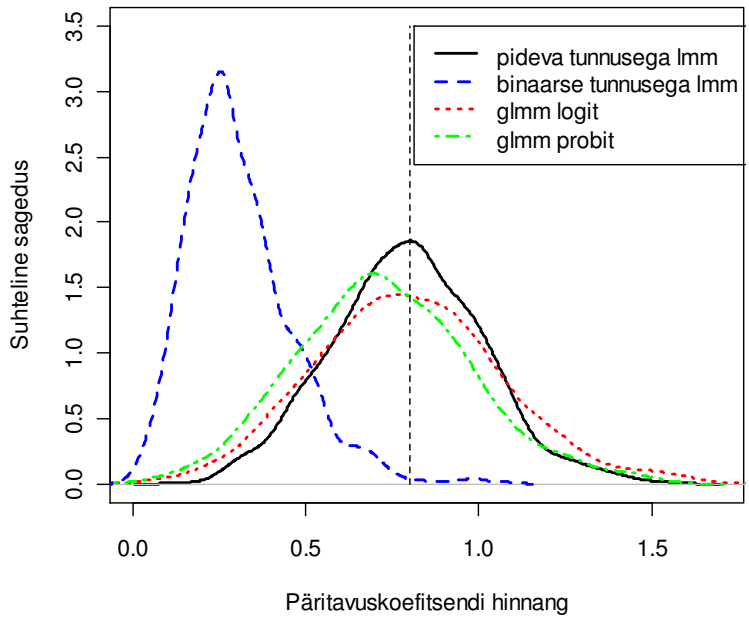
Joonis 22. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,2 ja lävendi 0,5 korral.



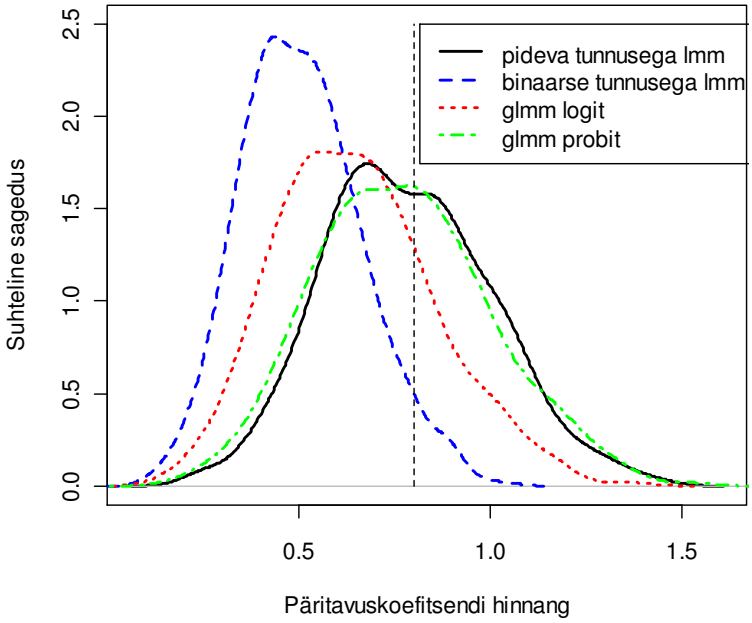
Joonis 23. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,2 ja lävendi 0,9 korral.

Joonistel 24-26 on näha päritavuskoefitsiendi empiirilised jaotused lävendite 0,1, 0,5 ja 0,9 ning tegeliku päritavuse 0,8 korral. Nagu kõigi teiste tegelike päritavuste korral annab ka tegeliku päritavuse 0,8 korral binaarsele tunnusele rakendatud üldine lineaarne segamudel väga tugevaid alahinnanguid – saadud hinnangud on keskmiselt poole väiksemad päritavuskoefitsiendi tegelikust väärtusest, eriti suured alahinnangud tekivad siis, kui binaarse uuritava tunnuse väärtuste osakaalud on väga palju erinevad 0,5-st.

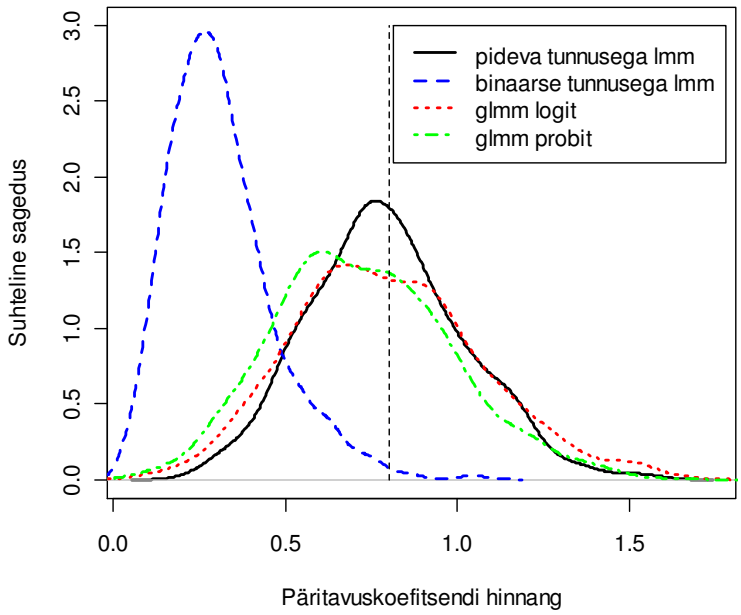
Võrreldes probit-mudeliga saadud hinnanguid erinevate lävendiväärtuste ja päritavus-koefitsienti väärtuste korral, ilmneb ka, et erinevalt logit-mudelist, mille korral olid päritavus-koefitsiendi hinnangud täpsemad äärmuslike päritavuse väärtuste puhul, annab probit-mudel pisut täpsemaid hinnanguid keskmise lävendiväärtuse korral.



Joonis 24. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,8 ja lävendi 0,1 korral.



Joonis 25. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,8 ja lävendi 0,5 korral.



Joonis 26. Päritavuskoefitsiendi hinnangu empiiriline jaotus tegeliku päritavuskoefitsiendi väärtuse 0,8 ja lävendi 0,9 korral.

Lisaks dispersioonikomponentide ja päritavuskoefitsiendi hindamisele leiti kõigist mudelitest ka juhuslike isaefektide hinnangud. Erinevate mudelite võrdlemiseks arvutati igal modelleerimisel välja erinevatest mudelitest hinnatud isaefektide vahelised Pearsoni lineaarsed korrelatsioonikordajad ja Spearmani astakkorrelatsioonikordajad. Kõigi antud parameetrite kombinatsioonil teostatud modelleerimiste lõpuks leiti keskmised korrelatsioonikordajad, mis tegeliku päritavuse 0,5 korral läbi viidud modelleerimiste kohta on toodud tabelis 4. Nagu näha, on Pearsoni ja Spearmani korrelatsioonikordajate keskmised väärtused enam-vähem samad ja üldiselt on kõigi mudelite vahelised korrelatsioonikordajad väga kõrged. Modelleerimisel oli näha, et kui kõik mudelid koondusid, siis olid korrelatsioonikordajad peaaegu ühed, seega enamus isaefektide vahelistest erinevustest tulid sellest, et üldistatud lineaarsed segamudelid just äärmuslike lävendiväärtuste korral mõni kord ei koondunud. Teiste tegelike päritavuskoefitsiendi väärtuste korral olid mudelite vahelised korrelatsioonid väga sarnased ja neid ei ole eraldi ära toodud. Kokkuvõttes, hoolimata päritavuskoefitsientide hinnangute suurest erinevust olid erinevatest mudelitest leitud juhuslike isamõjude hinnangud ja isade paremusjärjestus peaaegu täpses lineaarses seoses. Seega ei tohiks põllumajandusloomade geneetilise potentsiaali hindamisel ja paremate

loomade selekteerimisel binaarsele tunnusele rakendatud tavalise üldise lineaarse segamudeli tulemuste alusel vigu tekkida – valedel eeldustel baseeruva mudeli kasutamine ei too kaasa valesid aretusotsuseid ja Jõudluskontrolli Keskuse poolt hinnatavad poegimis- ja sigimistunnuste aretusväärtused võib lugeda usaldusväärseteks. Teine asi on muidugi aretusprogrammide ja selektsioonistrateegia paika panekuks vajalike geneetiliste parameetrite nagu aditiivgeneetiline dispersioon ja päritavuskoefitsient väärtustega – nende valedel eeldustel baseeruvatest mudelitest hindamine võib viia selektsiooniefekti alahindamisele. Analoogsele tõdemusele jõudsid ka Meijering ja Gianola oma 1985. aastal publitseeritud uuringus. Samas Mrode (2005) leidis oma raamatus, et juhul, kui järglaste arv erinevatel isadel on väga erinev ja kui hindamismudel sisaldab mitmeid mittetasakaalulisi fikseeritud faktoreid, ei pruugi tavalisest üldisest lineaarsest mudelist hinnatud isade aretusväärtused vastata tegelikkusele ja nende alusel loomade selekteerimine võib pärssida aretusedu.

Tabel 4. Erinevatest mudelitest hinnatud isaefektide vahelised keskmised (standardhälve) Pearsoni (allpool peadiagonaali) ja Spearmani (ülalpool peadiagonaali) korrelatsiooni-kordajad erinevate lävendi väärtuste korral. Analüüsid baseeruvad 1000-l modelleerimisel tegeliku pideval skaalal (alustunnuse skaalal) päritavuse 0,5 korral.

	LMM	LMMb	GLMM_l	GLMM_p
		Lävend 0,1		
LMM	1	0,847 (0,077)	0,844 (0,093)	0,844 (0,094)
LMMb	0,859 (0,058)	1	0,997 (0,059)	0,992 (0,063)
GLMM_l	0,869 (0,079)	0,984 (0,054)	1	0,989 (0,085)
GLMM_p	0,868 (0,082)	0,982 (0,063)	0,993 (0,084)	1
		Lävend 0,5		
LMM	1	0,941 (0,032)	0,941 (0,319)	0,939 (0,050)
LMMb	0,959 (0,019)	1	1,000 (0,001)	0,997 (0,043)
GLMM_l	0,959 (0,019)	0,999 (0,001)	1	0,997 (0,042)
GLMM_p	0,957 (0,045)	0,997 (0,043)	0,998 (0,043)	1
		Lävend 0,9		
LMM	1	0,841 (0,078)	0,833 (0,113)	0,837 (0,099)
LMMb	0,858 (0,058)	1	0,990 (0,093)	0,991 (0,068)
GLMM_l	0,861 (0,104)	0,978 (0,090)	1	0,982 (0,114)
GLMM_p	0,866 (0,086)	0,982 (0,065)	0,987 (0,111)	1

LMM – üldine lineaarne segamudel normaaljaotuse järgi jaotuva uuritava tunnusega,

LMMb – üldine lineaarne segamudel binaarse uuritava tunnusega,

GLMM_l – üldistatud lineaarne segamudel logit-seosefunktsiooniga,

GLMM_p – üldistatud lineaarne segamudel probit-seosefunktsiooniga.

4. Lehmade tiinestumise analüüs

4.1. Andmestiku kirjeldus

Tegelikel andmetel baseeruva uuringu aluseks on ühes Eesti suurfarmis aastatel 2003 kuni 2010 toimunud seemenduste andmed ja uuritavaks tunnuseks on mullikate ning esimese ja teise laktatsiooni lehmade tiinestuvus esimesest seemendusest. Mullikate ning esimese ja teise laktatsiooni lehmade tiinestuvust uuriti eraldi, mistap moodustati algsest andmestikust kolme erineva vanusegrupi tarvis eraldi alamandmestikud. Potentsiaalsete tiinestumist mõjutavate faktoritena on teada loomade seemendamisaasta, esmasseemendusvanus (mullikatel) või taastumisperioodi pikkus (so ajavahemik poegimisest esimese seemendamiseni, lehmadel), seemendustehniku kood ja geneetilise faktorina loomade isa. Usaldusväärsemate hinnangute saamise ja hindamisalgoritmide koondumise huvides on andmestikku kaasatud vaid 12 enamkasutatud pulli järglased, samuti on erinevate vanusegruppide tarvis koostatud alamandmestikest jäetud kõrvale kõik seemendused, mis on teostatud seemendustehnikute poolt, kes antud vanusegrupis teostasid viis või vähem seemendust.

Mullikate andmestikus olid kokku andmed 1080, esimese laktatsiooni lehmade andmestikus 888 ja teise laktatsiooni lehmade andmestikus 441 esimese seemenduse kohta. Tabelis 5 on toodud tiinestuvus vanusegrupiti - esimese seemenduse järel jäi tiineks 47% mullikatest, 1. ja 2. laktatsiooni lehmadel on see näitaja väiksem, vastavalt 29% ja 25%.

Tabelist 6 on näha, et erinevate isade tütred jäävad tõesti tiineks erineva edukusega, näiteks isa 65022 järglastest tiinestus mullikana esmaseemenduse järgselt tervelt 63%, samas isal 62998 on sama näitaja vaid 28%. Isade järglaste tiinestuvus varieerub ka vanusegrupiti – näiteks selle sama isa 65022, kelle tütarde tiinestuvus oli kõige kõrgem mullikate seas, 1. laktatsioonile jõudnud tütardest ei tiinestunud esimese seemenduse järel ükski.

Tabel 5. Tiinestuvus vanusegrupiti (osakaal vanusegrupiti).

	mullikad	1. laktatsiooni lehmad	2. laktatsiooni lehmad
Ei tiinestunud	569 (0,527)	629 (0,708)	330 (0,748)
Tiinestus	511 (0,473)	259 (0,292)	111 (0,252)

Tabel 6. Tiinestuvus isade kaupa ja vanusegrupiti (osakaal isade kaupa ja vanusegrupiti).

	Mullikad		1. laktatsiooni lehmad		2. laktatsiooni lehmad	
Isa	Ei tiinestunud	Tiinestus	Ei tiinestunud	Tiinestus	Ei tiinestunud	Tiinestus
56172	190 (0,576)	140 (0,424)	189 (0,690)	85 (0,310)	136 (0,727)	51 (0,273)
56274	26 (0,684)	12 (0,316)	28 (0,966)	1 (0,034)	0	0
62294	19 (0,543)	16 (0,457)	22 (0,629)	13 (0,371)	8 (1.000)	0 (0,000)
62313	36 (0,522)	33 (0,478)	46 (0,742)	16 (0,258)	8 (1.000)	0 (0,000)
62708	14 (0,636)	8 (0,364)	15 (0,750)	5 (0,250)	6 (1.000)	0 (0,000)
62998	18 (0,720)	7 (0,280)	19 (1.000)	0 (0,000)	0	0
65022	15 (0,366)	26 (0,634)	9 (1.000)	0 (0,000)	0	0
65210	120 (0,498)	121 (0,502)	148 (0,688)	67 (0,312)	94 (0,764)	29 (0,236)
65303	43 (0,537)	37 (0,463)	40 (0,533)	35 (0,467)	36 (0,632)	21 (0,368)
65575	32 (0,410)	46 (0,590)	55 (0,846)	10 (0,154)	9 (1.000)	0 (0,000)
65642	26 (0,464)	30 (0,536)	26 (0,510)	25 (0,490)	31 (0,756)	10 (0,244)
65699	30 (0,462)	35 (0,538)	32 (0,941)	2 (0,059)	2 (1.000)	0 (0,000)

4.2. Rakendatud mudelid

Hindamaks isa kui geneetilise faktori mõju, võttes samaaegselt arvesse ka võimalike mittegeneetiliste faktorite mõjud, konstrueeriti mullikate ning esimese ja teise laktatsiooni lehmade tarvis eraldi mudelid. Kõik mudelid sisaldasid juhuslikke faktoreid „Isa“ ja „Seemendustehnik“. Kuna erinevatel aastatel oli kasutatud erinevaid seemendustehnikuid, hõlmas seemendustehniku mõju ka aasta mõju, mistap aastat kui fikseeritud faktorit mudelisse lisada ei olnud vaja. Mullikate tiinestuvuse analüüsil ei osutunud statistiliselt oluliseks looma vanus seemendamisel (esmaseemendusiga), mistap jäeti see argument mudelist välja. Küll sisaldas esimese laktatsiooni lehmade analüüsil rakendatud mudel pideva argumendina lehmade taastumisperioodi pikkust, mille mõju oli statistiliselt oluline, teise laktatsiooni lehmade analüüsil ei tulnud taastumisperioodi pikkus statistiliselt oluline ja seega jäeti see argument mudelist välja. Kõigi kolme vanusegrupi puhul modelleeriti tiinestuvust üldise lineaarse segamudeliga (LMM), logit-seosefunktsiooniga üldistatud lineaarse sega-mudeliga (logit GLMM) ja probit-seosefunktsiooniga üldistatud lineaarse segamudeliga (probit GLMM).

Kokkuvõttes rakendati mullikate ja teise laktatsiooni lehmade tiinestuvuse uurimisel mudelit kujul

$$g(y_{ijk}) = \mu + S_i + T_k + \varepsilon_{ijk}$$

ja esimese laktatsiooni lehmade tiinestuvuse uurimisel mudelit kujul

$$g(y_{ijk}) = \mu + S_i + T_k + b^*tp_{ijk} + \varepsilon_{ijk},$$

kus y_{ijk} on *i*. isa järglasena sündinud *j*. seemendustehniku poolt seemendatud *k*. looma tiinestuvus, $S_i \sim N(0, \sigma_s^2)$ on isa *i* juhuslik mõju, $T_j \sim N(0, \sigma_T^2)$ on seemendustehniku *j* juhuslik mõju, tp_{ijk} on lehma taastumisperioodi pikkus ja *b* taasutmisperioodi tiinestumisega seostav regressioonikordaja. Funktsioon $g(.)$ on samasusteisendus üldise lineaarse segamudeli korral, logit-teisendus logit-seosefunktsiooniga üldistatud lineaarse segamudeli korral ja probit-teisendus probit-seosefunktsiooniga üldistatud lineaarse segamudeli korral.

Isa ja seemendustehniku mõjule vastavate dispersioonikomponentide σ_S^2 ja σ_T^2 ja jääkvarieeruvuse σ_ε^2 alusel arvutati kõigi mudelite puhul päritavuskoefitsiendi hinnagud valemist $h^2 = 4\sigma_S^2 / (\sigma_S^2 + \sigma_T^2 + \sigma_\varepsilon^2)$ ja seemendustehniku mõju osakaalud valemist $c^2 = \sigma_T^2 / (\sigma_S^2 + \sigma_T^2 + \sigma_\varepsilon^2)$. Vastavalt lävendimudeli statistilisele esitusele võeti jääkvarieeruvus σ_ε^2 logit-seosefunktsiooniga üldistatud lineaarse segamudeli korral võrdseks suurusega $\pi^2/3$ ja probit-seosefunktsiooniga üldistatud lineaarse segamudeli korral võrdseks ühega (vt ka valemid 19 ja 20).

Analüüsid üldise lineaarse segamudeliga teostati SAS-i protseduuriga MIXED ning analüüsid üldistatud lineaarse segamudeliga SAS-i protseduuriga GLIMMIX.

4.3. Tulemused

Ainsa fikseeritud faktorina 1. laktatsiooni lehmade tiinestumise modelleerimisel arvesse võetud taastumisperioodi mõju hinnangutest selgus, et mida pikem on taastumisperiood seda väiksem on tõenäosus tiinestuda esimesel seemendusel. Probit-seosefunktsiooniga mudeli korral tuli regressioonikordaja hinnanguks -0,006 (p=0,005), logit-seosefunktsiooniga mudeli korral -0,010 (p=0,006) ja üldise lineaarse segamudeli korral -0,002 (p=0,005). Logit-seosefunktsiooniga mudeli tulemuste alusel saab seega konstanteerida, et taastumisperioodi

pikenemisel ühe päeva võrra muutub šanss tiinestuda $e^{-0,01}$ = 0,990 korda väiksemaks, mis on muidugi imeväike muutus.

Päritavuskoefitsiendi hinnang mullikate jaoks on probit-seosefunktsiooniga mudeli korral 0,08, logit-seosefunktsiooniga mudeli korral 0,06 ja lineaarse segamudeli korral 0,05 (tabel 7). Modelleerimiseksperimendile tuginedes on viimase väärtuse näol arvatavasti tegemist alahinnanguga. Kuna keskmine tiinestuvus mullikate puhul on 47%, siis võib modelleerimis-eksperimendi alusel arvata, et ka logit-seosefunktsiooniga mudelist saadud hinnang on alahinnang ning et tegelik päritavus on 8% ringis. Esimese laktatsiooni lehmade analüüsil tulid päritavuskoefitsiendi hinnangud vastavalt 0,19, 0,19 ja 0,07, kuna logit- ja probit-seosefunktsiooniga mudelitest saadud päritavuskoefitsiendi hinnangud on võrdsed, siis võib suhteliselt veendunult väita, et umbes 20% 1. laktatsiooni lehmade esimese seemenduse järgsest tiinestuvusest on päritav. Teise laktatsiooni lehmade analüüsil tuli kõigi mudelite korral isaefektide dispersioon null, mis näitab nagu ei mängiks geneetika 2. laktatsiooni lehmade tiinestumise juures enam mingit rolli. Tegelikkuses see ilmselt siiski nii ei ole, pigem on 2. laktatsiooni lehmade andmestik isa geneetilise mõju hindamiseks liiga väike.

Seemendustehniku mõju osakaal kasvab lehma vanuse suurenedes (tabel 7). Mullikate korral on seemendustehniku mõju osakaal probit-seosefunktsiooniga mudeli korral 3.8% ja logit-seosefunktsiooniga mudeli korral 3%, samas 1. laktatsiooni lehmade puhul on vastavad osakaalud juba 19,8% ja 16,2% ning 2. laktatsiooni lehmade korral on osakaalu hinnangud veelgi kõrgemad vastavalt 30,6% ja 28,3%. Binaarsele tunnusele rakendatud tavalise lineaarse segamudeli korral hinnatakse seemendustehniku mõju osakaaluks mullikate korral 2,5%, 1. laktatsiooni lehmade korral 11,9% ja 2. laktatsiooni lehmade korral 11,6% – sarnaselt päritavusele on needki osakaalud alahinnatud.

Tabel 7. Isa ja seemendustehniku mõjule vastavate dispersioonikomponentide σ_s^2 ja σ_T^2, jääkvarieeruvuse σ_ε^2, päritavuskoefitsiendi $h^2 = 4\sigma_s^2 / (\sigma_s^2 + \sigma_T^2 + \sigma_s^2)$ ja seemendustehniku mõju osakaalu $c^2 = \sigma_T^2 / (\sigma_s^2 + \sigma_T^2 + \sigma_s^2)$ hinnangud erinevas vanuses loomade ja erinevate statistiliste mudelite korral.

Hinnatav parameeter	Mudel		
	LMM	Logit GLMM	Probit GLMM
Mullikad			
σ_s^2	0,003	0,054	0,021
σ_T^2	0,006	0,103	0,041
σ_ε^2	0,243	$\pi^2/3$	1
h^2	0,054	0,062	0,080
c^2	0,025	0,030	0,038
I laktatsiooni lehmad			
σ_s^2	0,004	0,200	0,063
σ_T^2	0,025	0,674	0,263
σ_ε^2	0,178	$\pi^2/3$	1
h^2	0,068	0,192	0,190
c^2	0,119	0,162	0,198
II laktatsiooni lehmad			
σ_s^2	0	0	0
σ_T^2	0,022	1.298	0,440
σ_ε^2	0,168	$\pi^2/3$	1
h^2	0	0	0
c^2	0,116	0,283	0,306

LMM – üldine lineaarne segamudel,

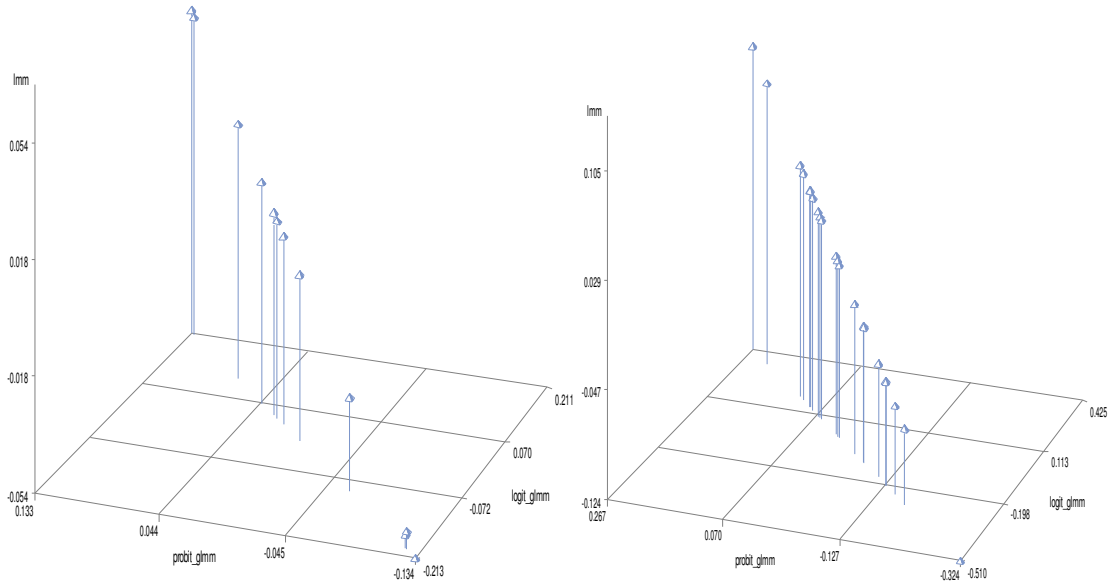
Logit GLMM – üldistatud lineaarne segamudel logit-seosefunktsiooniga,

Probit GLMM – üldistatud lineaarne segamudel probit-seosefunktsiooniga.

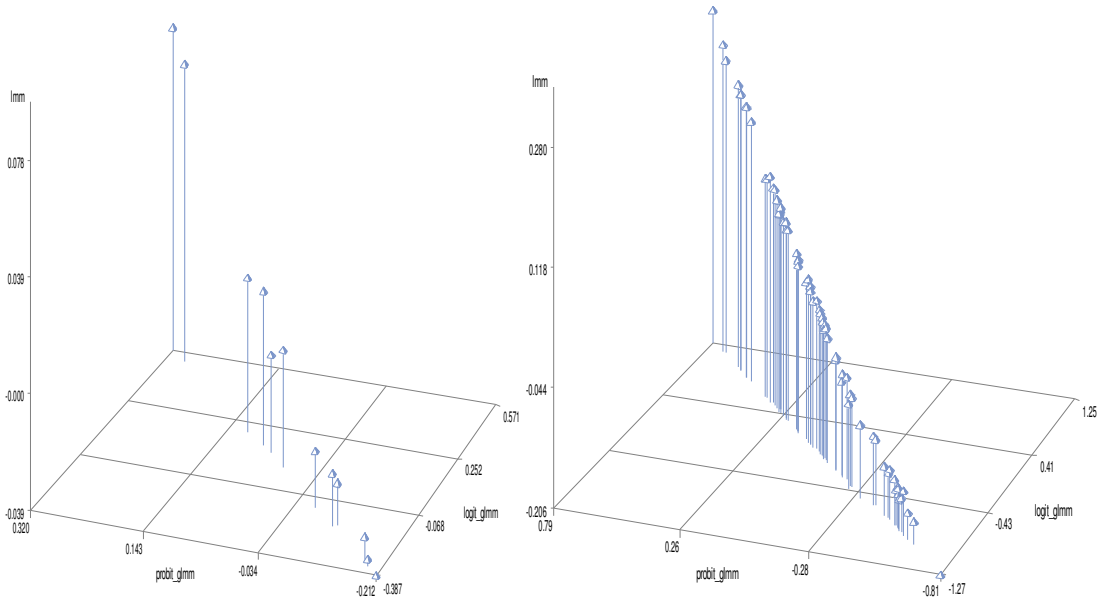
Jõudluskontrolli Keskus on Eesti piimaveiste populatsiooni baasil hinnanud üldisest lineaarsest segamudelist päritavuse kordusseemenduse puudumisele 56 päeva jooksul peale esmaseemendust (sisuliselt mõõdab see tunnus tiinestuvust esimesest seemendusest) ja saanud tulemuseks 0,012 (Veterinaar- ja Toiduamet, 2011). Antud väiksema mastaapsusega

uuringus saadi mullikate ja 1.laktatsiooni lehmade korral märksa kõrgemad päritavuskoefitsiendi hinnangud, samas 2. laktatsiooni lehmade puhul hinnati tiinestuvuse päritavuseks 0. Arvestades, et Jõudluskontrolli Keskuse uuringus on päritavus hinnatud kõigi vanusegrupi loomade tarvis ühiselt, siis vastav päritavuskoefitsiendi hinnangu kaalutud keskmine antud uuringus tuleb lineaarse segamudeli korral 0,036 ning logit- ja probit-seosefunktsiooniga mudelite korral vastavalt 0,084 ja 0,087. Kuna Jõudluskontrolli Keskus kasutas päritavuse hindamiseks uuritava tunnuse normaalsust eeldavat üldist lineaarset segamudelit, on nende poolt saadud väärtus ilmselt alahinnatud ja tegelik tiinestuvuse päritavus Eesti piimaveiste populatsioonis on kõrgem kui 1,2%. See teadmine loob paremad eeldused piimalehmade tiinestuvuse parandamiseks selektsiooni teel.

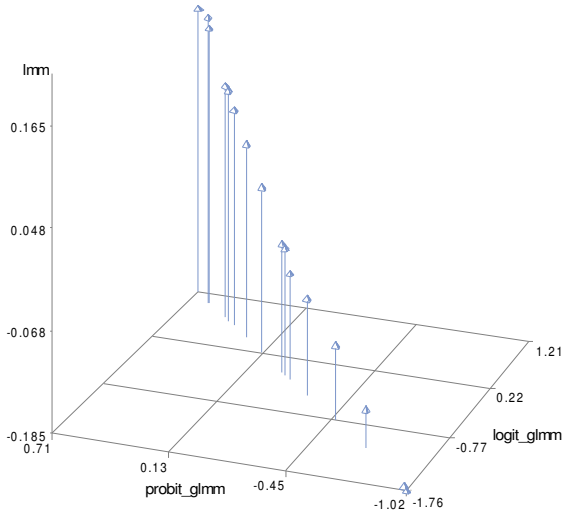
Isade ja seemendustehnikute mõjud hinnatuna eri mudelitega erinevate vanusegruppide korral on esitatud joonistel 27-30. Nagu nendelt joonistelt on näha, on kõigi vanusegruppide puhul nii erinevate mudelitega hinnatud isamõjud kui seemendustehnikute mõjud peaaegu täpses lineaarses vastavuses, enamus lineaarseid korrelatsioonikordajaid olid võrdsed ühega, nõrgim seos Pearsoni korrelatsioonikordajaga 0,976 ilmnes logit-seosefunktsiooniga mudelist ja lineaarse segamudelist hinnatud isamõjude vahel 1. 1aktatsiooni lehmadel.



Joonis 27. Isade ja seemendustehnikute mõjud eri mudelitega mullikate analüüsil (vastavalt vasak- ja parempoolne joonis).



Joonis 28. Isade ja seemendustehnikute mõjud eri mudelitega 1. laktatsiooni lehmade analüüsil (vastavalt vasak- ja parempoolne joonis).



Joonis 29. Seemendustehnikute mõjud eri mudelitega 2. laktatsiooni lehmade analüüsil.

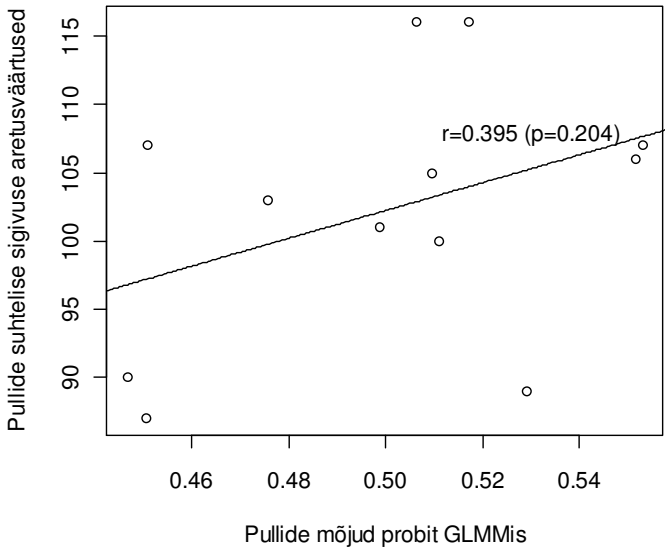
4.4. Seos isa mõju ja sigivuse suhtelise aretusväärtuse vahel

Alates 2008. aasta augustist hindab ja avaldab Jõudluskontrolli Keskus eesti punast ja eesti holsteini tõugu piimaveistele ka sigivuse suhtelist aretusväärtust SGAV (Jõudluskontrolli Keskus, 2008). SGAV-i arvutamiseks hinnatakse esmalt mitmemõõtmelisest mudelist aretusväärtused kordusseemenduse puudumisele 56 päeva jooksul peale esmaseemendust, taastumisperioodi pikkusele ja seemendusperioodi pikkusele. Taastumis- ja seemendus-perioodi pikkuste aretusväärtuste summana avaldatakse uuslüpsiperioodi aretusväärtus, mis omakorda teisendatakse punktiskaalale keskmisega 100 ja standardhälbega 12 punkti ja mis kujutabki enesest sigivuse suhtelist aretusväärtust (Veterinaar- ja Toiduamet, 2011). Sigivuse suhtelise aretusväärtuse seos tiinestuvusega esimesest seemendusest on loomulik, sest mida suurem hulk mingi pulli järglastest esimese seemenduse järgselt tiinestub, seda lühem on nende seemendusperiood ja seeläbi ka uuslüpsiperiood.

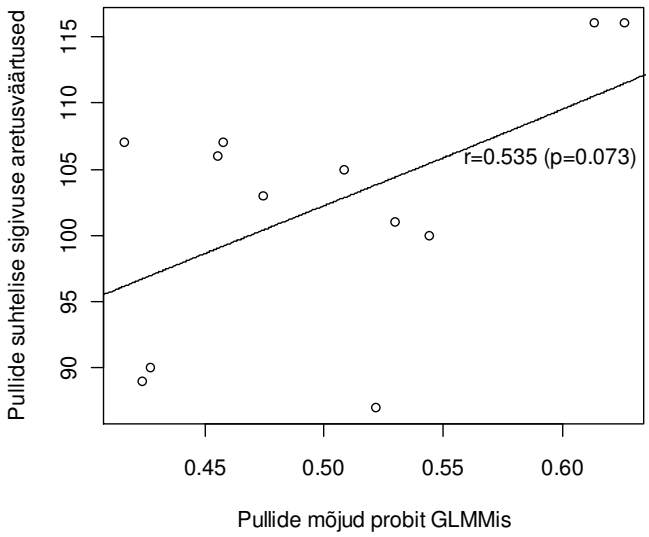
Uurimaks, kuivõrd kajastub pullide kogu Eesti piimaveiste populatsiooni baasil hinnatud geneetiline võimekus konkreetsetes tingimustes ja otseselt hindamise aluseks mitte olnud tunnuse korral, kõrvutati pullide suhtelisi sigivuse aretusväärtuseid nende tütarde logit- ja probit-seosefunktsiooniga üldistatud lineaarseist segamudeleist hinnatud tiinestuvustega antud farmis (tabel 8). Tütarde tiinestuvuse hinnangud tõenäosuse skaalal arvutati mudelite parameetrite hinnangute alusel valemitest (16) ja (17). Lisaks arvutati pullide SGAV-de ja probit-mudelist hinnatud mõjude vahelised Pearsoni korrelatsioonikordajad – vastavaid seoseid illustreerivad joonised 30 ja 31. Nii mullikate kui ka 1. laktatsiooni lehmade korral on seos positiivne – mida kõrgem on pulli suhteline sigivuse aretusväärtus, seda paremini tema tütred antud farmis tiinestuvad. Seos on pisut nõrgem mullikate ja tugevam 1. laktatsiooni lehmade korral (korrelatsioonikordajad vastavalt 0,395 ja 0,535). See erinevus on vastavuses antud andmete alusel hinnatud päritavuskoefitsientidega – mullikate puhul on päritavus-koefitsiendi väärtus ja seega ka isalt pärandunud geenide roll tütarde tiinestuvuses madalam ning väljendub see ka nõrgemas seoses isa geneetilise potentsiaali (SGAV) ja tütarde tiinestu-vuse vahel. Esimese laktatsiooni lehmade puhul aga peegeldab kõrgem päritavuskoefitsiendi väärtus tugevamat seost isa SGAV-i ja tütarde tiinestuvuse vahel. See, et ei mullikate ega ka 1. laktatsiooni lehmade puhul ei ole isade SGAV-i ja tütarde tiinestuvuse vaheline seos tugev, näitab ühelt poolt, et suhteline sigivuse aretusväärtus peegeldab vaid osaliselt tiinestuvust, ja teiselt poolt, et kõigi pullide SGAV ei pruugi nende tütardel avalduda kõigis farmides ühte moodi.

Tabel 8. Pullide suhtelised sigivuse aretusväärtused SGAV ning logit- ja probit-seosefunktsiooniga üldistatud lineaarseist segamudeleist hinnatud tütarde tiinestuvused antud farmis.

Pull	SGAV	Probit GLMM		Logit GLMM	
		mullikad	1. laktatsiooni lehmad	mullikad	2. laktatsiooni lehmad
56172	87	0,450	0,522	0,451	0,528
56274	90	0,447	0,427	0,447	0,417
62294	101	0,499	0,529	0,499	0,535
62313	105	0,510	0,509	0,509	0,511
62708	103	0,476	0,474	0,476	0,470
62998	107	0,451	0,416	0,451	0,404
65022	107	0,553	0,457	0,553	0,450
65210	100	0,511	0,455	0,511	0,549
65303	116	0,506	0,613	0,506	0,626
65575	106	0,552	0,456	0,551	0,454
65642	116	0,517	0,626	0,517	0,639
65699	89	0,529	0,424	0,529	0,412



Joonis 30. Mullikate tiinestuvuse analüüsis probit-mudelist hinnatud isamõjude ja isade SGAV-ide vaheline seos; joonisele on kantud ka regressioonisirge ja selle juurde Pearsoni korrelatsioonikordaja ja vastava olulisuse tõenäosuse väärtus.



Joonis 31. Esimese laktatsiooni lehmade tiinestuvuse analüüsis probit-mudelist hinnatud isamõjude ja isade SGAV-ide vaheline seos; joonisele on kantud ka regressioonisirge ja selle juurde Pearsoni korrelatsioonikordaja ja vastava olulisuse tõenäosuse väärtus.

Kokkuvõte

Käesolev magistritöö tegeleb teoreetilises mõttes paradoksaalse aga reaalses elus tavalise olukorra, kus polügeense tunnuse väärtused avalduvad binaarsel skaalal, modelleerimisega. Kui geneetika seisukohast on binaarse tunnuse polügeenne determineeritus seletatav alustunnuse (*liability*) kontseptsiooni kaudu, siis olukorra statistilisel modelleerimisel on kasutatavad mitmed erinevad mudelid.

Modelleerimiseksperimendist selgus, et põllumajandusloomade populatsioonipõhistes uuringutes laialt kasutatav uuritava tunnuse normaaljaotust eeldav üldine lineaarne segamudel alahindab tugevalt binaarse tunnuse päritavust. Spetsiaalselt binaarsete tunnuste modelleerimiseks mõeldud logit- ja probit-seosefunktsiooniga üldistatud lineaarsed segamudelid annavad seevastu küllaltki täpsed hinnangud päritavuskoefitsiendile, ja seda sõltumata päritavuskoefitsiendi tegelikust väärtusest ja binaarse tunnuse lävendiväärtusest. Vaid logit-seosefunktsiooniga mudel alahindab binaarse tunnuse päritavust võrdtõenäoliste väärtuste korral (lävend on siis 0,5). Aga ka need hinnangud on märksa täpsemad, kui binaarsele tunnusele rakendatud üldisest lineaarsest segamudelist saadud. Isade geneetiliste mõjude hinnangud ja isade paremusjärjestus on kõigi mudelite puhul peaaegu 100%-lises vastavuses. Seega binaarsele tunnusele rakendatud üldine lineaarne segamudel ei sobi geneetilise determineerituse määra hindamiseks, küll aga saab sellega hinnata aretusväärtuseid.

Neljandas peatükis rakendatakse tiinestumise modelleerimiseks üldist lineaarset segamudelit ning logit- ja probit-seosefunktsiooniga üldistatud lineaarseid segamudeleid. Analoogselt modelleerimis-eksperimendile ilmnes, et isade geneetilist mõju hindasid kõik mudelid võrdselt hästi, samas kui päritavuskoefitsiendi ja seemendustehniku mõju osakaalu hinnangud tulid uuritava tunnuse normaaljaotust eeldava üldise lineaarse segamudeli korral selgelt väiksemad, kui teiste mudelite korral. Isa geneetiline mõju ilmnes selgemalt 1. laktatsiooni lehmade tiinestumises. Täiendavalt uuritud seos pullide suhtelise sigivuse aretusväärtuse ja tütarde tiinestuvuse vahel oli keskmise tugevusega ja sarnaselt päritavuskoefitsiendiga oli ka siin seos tugevam 1. laktatsiooni lehmade puhul ja nõrgem mullikate puhul.

Käesoleva töö tulemused annavad põllumajandusloomade aretajatele aluse usaldada ka üldistest lineaarsetest segamudelitest hinnatud binaarsete tunnuste aretusväärtuseid, samas ilmneb töö tulemustest selgelt, et valedel matemaatilistel eeldustel baseeruvad mudelid alahindavad binaarsete tunnuste geneetilise determineerituse määra. Et just viimasel baseeruvad selektsioonistrateegia-alased otsused ja aretusprogrammide sisu, on kindlasti vajalik jätkata uuringuid ka teiste normaaljaotusest erinevate tunnuste, nagu näiteks järglaste arv või eluiga, geneetilise determineerituse hindamise osas. Veel enam, aretajaid ei huvita mitte ainult üksikute tunnuste geneetiline determineeritus, vaid ka nende tunnuste geneetiline seotus. Viimase hindamine eeldab aga erineva jaotusega tunnuste ühte

mitmemõõtmelisse mudelisse inkorporeerimist, mis muudab statistilise modelleerimise veel komplitseeritumaks.

Genetic evaluation of polygenetic binary traits

Master Thesis

Mari Liiva

Summary

In this master thesis the general linear mixed models and generalized linear mixed models are introduced and their uses in estimating the most essential genetic parameters as variance components and heritability coefficients are studied. As usually in animal breeding, the variance component caused by sire is analyzed.

In the first chapter genetic models are introduced for normally distributed data and for binary data, in the second chapter relating statistical models are introduced. For convenience all four models are presented: general linear model, general linear mixed model, generalized linear model and generalized linear mixed model.

The third chapter is dedicated to simulating experiment to compare four different models: general linear mixed model applied to normally distributed liability of binary data, general linear mixed model applied to binary data, generalized linear mixed model with logit link and generalized linear mixed model with probit link applied to binary data. The program generates 1000 offspring from 20 sires; every one of them is paired with 50 dams. Data were generated and models were estimated for heritability coefficient values 0.2, 0.5 and 0.8 and for binary data thresholds 0.1, 0.05 and 0.9. General linear mixed model applied to binary data underestimated heritability coefficient strongly on every occasion. The general linear mixed model applied to normally distributed data gives always most accurate estimates but in real life it's useless as we do not know the underling normally distributed liability data and only the binary data is measurable. Logit and probit link models are quite accurate, except logit link models applied to binary data with threshold close to 0.5.

In the fourth chapter real data is analyzed. The dataset contains information about the first insemination of heifers and first and second parity cows in one Estonian dairy farm. The additive genetic effect of sire and random effect of insemination technician on non-return rate during 90 days after the insemination was estimated. The data was divided into three subgroups according animals' age and those subgroups were analyzed separately. For all subgroups three models were estimated and the heritability coefficient and the effects of sire

and insemination technician were estimated. The real data analysis confirm the previous simulation experiment, the general linear mixed model obviously underestimates the heritability while the generalized linear mixed models with logit and probit link operate well. The heritability of non-return rate for heifers was estimated as 0.06-0.08 and for first parity cows as 0.19, which are much higher values than corresponding heritability reported by Estonian Animal Recording Centre. As the Estonian Animal Recording Centre uses the general linear mixed model instead more suitable generalized linear mixed model, the official value is obviously underestimated. There were no differences in effects of sire and insemination technician got with different models.

Kasutatud kirjandus

Engel, B. 1997. Extending generalized linear models with random effects and components of dispersion. Landbouwuniversiteit Wageningen, 191 lk.

Falconer, D. S., Mackay, T. F. C. 1996. Introduction to Quantitative Genetics. Fourth Edition. Longman, Harlow, UK, 465 lk.

Jõudluskontrolli Keskus. 2008. Täiendav info aretuspullide valikuks. Jõudluskontrolli Keskuse Infoleht, Nr 19.

Kaart, T. 2006. The Reliability of Linear Mixed Models in Genetic Studies. PhD Dissertation. Tartu, Tartu University Press, 124 lk.

Kaart, T. 2012. Üldine lineaarne mudel. Loengukonspekt. http://www.eau.ee/~ktanel/VL_0192/pt6_2012.pdf (1.05.2012).

Kaart, T., Möls, T. 2011. Populatsioonigeneetika fenotüüpide tasemel. Loengukonspekt. http://www.eau.ee/~ktanel/MTMS_02_007/loeng_02_2010web.pdf (1.05.2012).

McCulloch, E. C. ja Searle, R. S. 2001. Generalized, linear, and mixed models. John Wiley & Sons, 321 lk.

Meijering, A., Gianola, D. 1985. Observations on sire evaluation with categorical data using heteroscedastic mixed linear models. Journal of Dairy Science, 68, 1226-1232.

Mrode, R. A. 2005. Linear models for the prediction of animal breeding values. 2nd Edition. CAB International, Wallingford, Oxon OX10 8DE, UK.

Niskanen, S., Juga, J. 1998. Calving difficulties and calf mortality in Finnish dairy cattle population. Interbull Bulletin, 18, 17-20.

SAS Institute Inc. 2006. The GLIMMIX Procedure, June 2006. http://support.sas.com/rnd/app/papers/glimmix.pdf (1.05.2012).

Uba, M. 2010, Poegimistunnuste aretusväärtus karja tulemuslikumaks taastootmiseks. JKK Sõnumid, 24, 2-3.

Veterinaar- ja Toiduamet. 2011. Piimaveiste põlvnemis- ja jõudlusandmete kogumise, nende õigsuse kontrollimise, töötlemise ja säilitamise kord.

http://www.vet.agri.ee/static/files/597.Piimaveistep6lvnemisjaj6udlusandmetekogumisenende 6igusekontrollimiset88tlemisejas2ilitamisekord.pdf (1.05.2012).

Viikmaa, M. (1998). Klassikalise geneetika leksikon. [http://biomedicum.ut.ee/~martv/genolex.html] (01.05.2012)

Lisad

Lisa 1

Modelleerimisprogrammi kood, kus päritavus on 0,5 ja lävend on 0,5.

```
%MACRO MSc_mod(kordusi);
%LOCAL u;
%DO u=1 %TO &kordusi;

data mod_data (keep = y biny sire farm);

array msire[1:20, 1:2, 1:100]; /* 20 isa 2 kromosoomi 100-s lookuses
paiknevate alleelide mõjude massiiv */
array mdam[1:20, 1:50, 1:2, 1:100]; /* 20x50 ema 2 kromosoomi 100 lookuse
mõjude massiiv */
array moffspring[1:20, 1:50, 1:2, 1:100]; /* 20x50 järglase 2 kromosoomi
100 lookuse mõjude massiiv */
array mfarm[1:2](2 -2); /* 2 farmi mõjud */

va=1; my=0; ve=200; /* va - üksikute alleeliväärtuste dispersioon; my -
üldkeskmine; ve - jääkdispersioon; kui näiteks va=0.5 ja ve=200, siis
var(y)=200+2*100*0.5=300, var(isa)=25 ja h2=4*var(isa)/var(y)=0,333 */

* Vanemate genotüüpide genereerimine;
do sire=1 to 20; do chrom=1 to 2; do loc=1 to 100;
msire[sire, chrom, loc] = sqrt(va)*normal(0);
end; end; end;

do sire=1 to 20; do dam=1 to 50; do chrom=1 to 2; do loc=1 to 100;
mdam[sire, dam, chrom, loc] = sqrt(va)*normal(0);
end; end; end; end;

* Järglaste genotüüpide moodustamine (1000 järglase kokku 20000 geeni) -
mõlema vanema iga lookuse korral valitakse juhuslikult, kumb alleelidest
järglasele pärandub;
do sire=1 to 20; do dam=1 to 50; do loc=1 to 100;
moffspring[sire, dam, 1, loc] = msire[sire, floor(2*ranuni(0))+1, loc];
moffspring[sire, dam, 2, loc] = mdam[sire, dam, floor(2*ranuni(0))+1, loc];
end; end; end;

* Järglaste fenotüüpide arvutamine (1000 fenotüüpi);
do sire=1 to 20; do dam=1 to 50;
farm = mfarm[floor(2*ranuni(0))+1];
y = my + farm + sqrt(ve)*normal(0);

do loc = 1 to 100;
y = y + moffspring[sire, dam, 1, loc] + moffspring[sire, dam, 2, loc];
end; /* Kokkuvõttes E(y)=0 ja var(y)=ve+2*100*va */

if y<sqrt(ve+2*100*va)*probit(0.5) then biny=0; else biny=1; /* Binaarse
tunnuse biny tekitamine,funktsiooni probit() argument näitab, millisest
kvantiilist väiksemate y-i väärtuste korral võetakse biny=0 */
output;
end; end;
run;

* Analüüsid;

ods output SolutionR=solR_MixedY CovParms=var_MixedY;
```

```
proc mixed data=mod_data;
class sire farm;
model y = farm;
random sire / s;
run;

ods output SolutionR=solR_Mixed CovParms=var_Mixed;
proc mixed data=mod_data;
class sire farm;
model biny = farm;
random sire / s;
run;

ods output SolutionR=solR_GlimmixL CovParms=var_GlimmixL;
proc glimmix data=mod_data pconv=0.00001;
class sire farm;
model biny = farm / dist=binomial link=logit solution;
random sire / s;
run; quit;

ods output SolutionR=solR_GlimmixP CovParms=var_GlimmixP;
proc glimmix data=mod_data pconv=0.00001;
class sire farm;
model biny = farm / dist=binomial link=probit solution;
random sire / s;
run; quit;


* Analüüside tulemuste failidesse koondamine;

data var_lmmY; set var_MixedY; k=&u;
%if &u=1 %then %do; data var_lmmY_out1; set var_lmmY; %end;
%else %do; data var_lmmY_out1; set var_lmmY_out1 var_lmmY; %end;

data var_lmm; set var_Mixed; k=&u;
%if &u=1 %then %do; data var_lmm_out1; set var_lmm; %end;
%else %do; data var_lmm_out1; set var_lmm_out1 var_lmm; %end;

data var_glmmL; set var_GlimmixL; k=&u;
%if &u=1 %then %do; data var_glmmL_out1; set var_glmmL; %end;
%else %do; data var_glmmL_out1; set var_glmmL_out1 var_glmmL; %end;

data var_glmmP; set var_GlimmixP; k=&u;
%if &u=1 %then %do; data var_glmmP_out1; set var_glmmP; %end;
%else %do; data var_glmmP_out1; set var_glmmP_out1 var_glmmP; %end;


data solR_lmmY; set solR_MixedY; sol_s_lmmY=estimate; keep sire sol_s_lmmY;
run;
data solR_lmm; set solR_Mixed; sol_s_lmm=estimate; keep sire sol_s_lmm;
run;
data solR_glmmL; set solR_GlimmixL; sol_s_glmmL=estimate; keep sire
sol_s_glmmL; run;
data solR_glmmP; set solR_GlimmixP; sol_s_glmmP=estimate; keep sire
sol_s_glmmP; run;

proc sort data=solR_lmmY; by sire; run;
proc sort data=solR_lmm; by sire; run;
proc sort data=solR_glmmL; by sire; run;
proc sort data=solR_glmmP; by sire; run;
```

```
data solR_out1; merge solR_lmmY solR_lmm solR_glmmL solR_glmmP; by sire;
k=&u; run;
%if &u=1 %then %do; data solR_out2; set solR_out1; %end;
%else %do; data solR_out2; set solR_out2 solR_out1; %end;


DM 'OUTPUT;CLEAR'; /* tühjendab väljundiakna */

%END;
%MEND;


%MSc_mod(1000); /* simulatsioonide arv */


/* Erinevate mudelitega hinnatud dispersiooniparametrite alusel
päritavuskoefitsientide arvutamine ja tulemuste faili kirjutamine */
/* k on modelleerimiskord */

proc sort data=var_lmmY_out1; by k CovParm; run;
data var_lmmY_out2; set var_lmmY_out1;
by k;
select (CovParm);
when ('sire') do; sig_s_lmmY=estimate; retain sig_s_lmmY; end;
when ('Residual') do; sig_e_lmmY=estimate; retain sig_e_lmmY; end;
end;
keep k sig_s_lmmY sig_e_lmmY h2_lmmY;
if last.k then do;
h2_lmmY=4*sig_s_lmmY/(sig_s_lmmY+sig_e_lmmY);
output; end;
run;

proc sort data=var_lmm_out1; by k CovParm; run;
data var_lmm_out2; set var_lmm_out1;
by k;
select (CovParm);
when ('sire') do; sig_s_lmm=estimate; retain sig_s_lmm; end;
when ('Residual') do; sig_e_lmm=estimate; retain sig_e_lmm; end;
end;
keep k sig_s_lmm sig_e_lmm h2_lmm;
if last.k then do;
h2_lmm=4*sig_s_lmm/(sig_s_lmm+sig_e_lmm);
output; end;
run;

proc sort data=var_glmmL_out1; by k CovParm; run;
data var_glmmL_out2; set var_glmmL_out1;
by k;
select (CovParm);
when ('sire') do; sig_s_glmmL=estimate; retain sig_s_glmmL; end;
end;
keep k sig_s_glmmL h2_glmmL;
if last.k then do;
h2_glmmL=4*sig_s_glmmL/(sig_s_glmmL+constant('pi')**2/3);
output; end;
run;

proc sort data=var_glmmP_out1; by k CovParm; run;
data var_glmmP_out2; set var_glmmP_out1;
by k;
select (CovParm);
```

```
when ('sire') do; sig_s_glmmP=estimate; retain sig_s_glmmP; end;
end;
keep k sig_s_glmmP h2_glmmP;
if last.k then do;
h2_glmmP=4*sig_s_glmmP/(sig_s_glmmP+1);
output; end;
run;


/* Erinevate mudelitega hinnatud dispersioonikomponentide ja päritavuste
ühte tabelisse koondamine */
proc sort data=var_lmmY_out2; by k; run;
proc sort data=var_lmm_out2; by k; run;
proc sort data=var_glmmL_out2; by k; run;
proc sort data=var_glmmP_out2; by k; run;

data var_out; merge var_lmmY_out2 var_lmm_out2 var_glmmL_out2
var_glmmP_out2; by k; run;

/* Erinevate mudelitega dispersioonikomponentide ja päritavuskoefitsientide
hinnangute statistika */
proc means data=var_out;
var sig_s_lmmY sig_e_lmmY h2_lmmY sig_s_lmm sig_e_lmm h2_lmm sig_s_glmmL
h2_glmmL sig_s_glmmP h2_glmmP;
run;


/* Erinevate mudelitega hinnatud isaefektide vahelised korrelatsioonid */

ods exclude SimpleStats VarInformation PearsonCorr SpearmanCorr; /*
analüüsi tulemusi output-aknasse ei trükita */
ods output PearsonCorr=p_out_05_05 SpearmanCorr=s_out_05_05;
/* vastavalt pearsoni ja spearmani korrelatsioonikordajate väljundfail
lävendi 0,5 ja päritavuse 0,5 korral */
proc corr data=solR_out2 pearson spearman;
var sol_s_lmmY sol_s_lmm sol_s_glmmL sol_s_glmmP;
by k;
run;

/* Korrelatsioonikordajate ühte faili ühendamine - esmalt Pearsoni, siis
Spearmani korrelatsioonikordajad, ja siis failid kokku */
data p_out_05_05_summary; set p_out_05_05;
by k;
if Variable='sol_s_lmmY' then do;
pcor_lmmY_lmm=sol_s_lmm; pcor_lmmY_glmmL=sol_s_glmmL;
pcor_lmmY_glmmP=sol_s_glmmP; retain pcor_lmmY_lmm pcor_lmmY_glmmL
pcor_lmmY_glmmP;
end;
if Variable='sol_s_lmm' then do;
pcor_lmm_glmmL=sol_s_glmmL; pcor_lmm_glmmP=sol_s_glmmP; retain
pcor_lmm_glmmL pcor_lmm_glmmP;
end;
if Variable='sol_s_glmmL' then do;
pcor_glmmL_glmmP=sol_s_glmmP; retain pcor_glmmL_glmmP;
end;
keep k pcor_lmmY_lmm pcor_lmmY_glmmL pcor_lmmY_glmmP pcor_lmm_glmmL
pcor_lmm_glmmP pcor_glmmL_glmmP;
if last.k then output;
run;

data s_out_05_05_summary; set s_out_05_05;
```

```
by k;
if Variable='sol_s_lmmY' then do;
scor_lmmY_lmm=sol_s_lmm; scor_lmmY_glmmL=sol_s_glmmL;
scor_lmmY_glmmP=sol_s_glmmP; retain scor_lmmY_lmm scor_lmmY_glmmL
scor_lmmY_glmmP;
end;
if Variable='sol_s_lmm' then do;
scor_lmm_glmmL=sol_s_glmmL; scor_lmm_glmmP=sol_s_glmmP; retain
scor_lmm_glmmL scor_lmm_glmmP;
end;
if Variable='sol_s_glmmL' then do;
scor_glmmL_glmmP=sol_s_glmmP; retain scor_glmmL_glmmP;
end;
keep k scor_lmmY_lmm scor_lmmY_glmmL scor_lmmY_glmmP scor_lmm_glmmL
scor_lmm_glmmP scor_glmmL_glmmP;
if last.k then output;
run;
data cor_out;
merge p_out_05_05_summary s_out_05_05_summary;
by k;
run;

proc means data=cor_out;
var pcor_lmmY_lmm pcor_lmmY_glmmL pcor_lmmY_glmmP pcor_lmm_glmmL
pcor_lmm_glmmP pcor_glmmL_glmmP scor_lmmY_lmm scor_lmmY_glmmL
scor_lmmY_glmmP scor_lmm_glmmL scor_lmm_glmmP scor_glmmL_glmmP;
run;
```