TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Carmen Taimre

# Laenuvõtja maksejõuetuse modelleerimine

Bakalaureusetöö (9 EAP)

Juhendaja: Märt Möls, PhD

TARTU
2015

# Laenuvõtja maksejõuetuse modelleerimine

Käesoleva bakalaureusetöö eesmärk on selgitada elulemusanalüüsi olemust ning rakendada seda laenude andmestikul. Esimeses peatükis selgitatakse elulemusanalüüsi mõistet. Edasi kirjeldatakse, kuidas kasutada Kaplan-Meieri meetodit elulemusfunktsioonile hinnangu leidmiseks ning *log-rank* testi elulemuskõverate erinevuse tuvastamiseks. Seejärel antakse ülevaade Coxi võrdeliste riskide mudelist. Teises peatükis rakendatakse eelnimetatud meetodeid Bondora laenude andmestikul, analüüsimaks klientide maksejõuetuks muutumist. Lisaks viiakse läbi näide elulemusfunktsiooni rakendamisest laenude tootluste arvutamisel.

**Märksõnad:** *elukestusanalüüs, elulemus, statistilised meetodid, mudelid, R (programmeerimiskeel)*

# Modelling borrower's insolvency

The aim of this thesis is to explain the nature of survival analysis and to apply that on a loan dataset. The first chapter explains the concept of survival analysis, describes how to use Kaplan-Meier method to estimate a survival function and provides an overview of the log-rank test which is used to compare survival curves. Finally, the chapter introduces the Cox proportional hazards model. In the second chapter the previously mentioned methods are applied on Bondora's loan dataset to analyse the insolvency of a borrower. In addition, an example of using survival function for calculating internal rates of return on loans is carried out.

**Keywords:** *survival analysis, survival rate, statistical methods, models, R (programming language)*

# Sisukord

# Sissejuhatus

Otselaenamisettevõtted, mis pakuvad alternatiivi pangalaenule, võimaldavad läbi vastavate portaalide nii laene taotleda kui ka rahastada. Investoritel on eraisikulaenudesse investeerimine kasulik, sest tehinguid on lihtne sooritada ning võimalik on teenida suuremat tootlust kui näiteks pangas hoiustades. Samas kaasnevad sellega ka teatud riskid, millest olulisim on kliendi maksejõuetus. Kuna laenud ei ole ühegi hüvitamisskeemiga kindlustatud, riskib investor kogu investeeritud rahaga.

Käesolevas bakalaureusetöös uuritakse Bondora (isePankur AS) klientide maksejõuetuks muutumist, kasutades elulemusanalüüsi. Esimeses peatükis selgitatakse elulemusanalüüsi mõistet ning seejärel kirjeldatakse, kuidas Kaplan-Meieri meetodiga elulemusfunktsioonile hinnangut leida. Edasi selgitatakse, kuidas kasutada *log-rank* testi elulemuskõverate võrdlemiseks, ning viimaks antakse ülevaade Coxi võrdeliste riskide mudelist.

Teises peatükis rakendatakse eelnimetatud meetodeid Bondora laenude andmestikul, keskendudes põhiliselt 24 kuu pikkustele laenudele. Esmalt vaadatakse, milline näeb välja klientide maksejõulisust kirjeldav elulemuskõver ning seejärel uuritakse, kuidas muutuvad elulemuskõverad sõltuvalt klientide vanusest, riigist ja haridustasemest. Lisaks kasutatakse *log-rank* testi, et kõverate vahel erinevust tuvastada. Järgmisena luuakse Coxi võrdeliste riskide mudel, mis mitmete tunnuste abil, nt kliendi vanus ja haridustase, prognoosib tõenäosust, et klient on makseperioodi lõpus endiselt maksevõimeline. Viimaks tuuakse näide elulemusfunktsiooni rakendamisest laenude tootluste arvutamisel.

Töö on kirjutatud tekstitöötlusprogrammiga Microsoft Office Word 2013 ning analüüs on läbi viidud rakendustarkvaras R (versioon 3.1.2).

Autor tänab juhendajat Märt Mölsi rohkete nõuannete ja asjalike soovituste eest.

# 1 Elulemusanalüüs

## 1.1 Elulemusanalüüsi mõiste

Elulemusanalüüs (*survival analysis*) on kogum statistilistest protseduuridest, kus huvipakkuv tunnus on aeg sündmuse esinemiseni (Kleinbaum & Mitchel, 2005, lk 4). Olenevalt valdkonnast, kus analüüsi rakendatakse, mõistetakse sündmuse all näiteks surma, haigusjuhtumeid, masina rikkiminemisi, abielulahutusi, laenude pankrotistumisi jne. Elulemusanalüüs on levinud meditsiinis, bioloogias, mehaanikas, majanduses, sotsioloogias, demograafias ja mujal.

Aega kindla ajaintervalli algusest kuni sündmuse esinemiseni nimetatakse elulemusajaks (*survival time*) ning see on juhuslik suurus (Kleinbaum & Mitchel, 2005, lk 8). Kliinilistes uuringutes võidakse jälgida näiteks patsientide elulemusaegu, eesmärgiga tuvastada mõne uue ravimi mõju või teada saada, kaua mingi konkreetse haigusega inimesed elavad.

Elulemusanalüüsi puhul tuleb arvestada, et uuringud on ajaliselt piiratud ning jälgida saab vaid neid sündmusi, mis leiavad aset uuringu jooksul. Ometi võib objektil sündmus esineda ka pärast uuringu lõppu. Seda fakti võetakse arvesse tsenseerimisega (*censoring*). Tsenseerimine esineb siis, kui ei teata objekti täpset elulemusaega ning selle kohta on olemas vaid osaline informatsioon (Kleinbaum & Mitchel, 2005, lk 5). Näiteks teame, et patsient elas kauem kui 2 aastat, aga me ei tea täpselt kui kaua, sest uuring lõppes enne tema surma. See patsient on tsenseeritud ning tema jälgimisaeg on 2 aastat, s.o aeg, mille vältel ta uuringus viibis.

Enamasti on kasutusel parem-tsenseerimine, mis tähendab, et objekti täpne elulemusaeg n-ö lõigatakse ära paremalt poolt, kui uuring lõpeb või objekt uuringus osalemast loobub. Seega on tsenseeritud objekti jälgimisaeg lühem kui reaalne elulemusaeg. (Kleinbaum & Mitchel, 2005, lk 7) Kui aga tsenseerimist ei kasutataks ning sellised vaatlused üldse uuringust välja jäetaks, oleks tulemus süngem, kui see tegelikult on, sest analüüsimiseks jääksid vaid need objektid, kelle elulemusaeg on teada ehk kellel sündmus esineb varem.

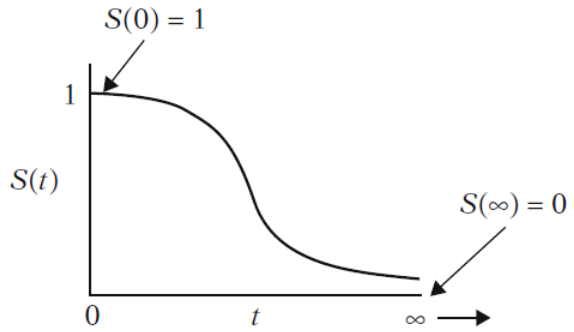
Juhuslikku suurust objekti elulemusaja jaoks tähistatakse $T$-ga, kus $T \geq 0$ ning elulemusaja $T$ realiseerinud väärtust $t$-ga. Sündmuse staatuse jaoks on kasutusel indikaatortunnus $\delta \in (0, 1)$, kus $\delta = 1$ tähendab, et objektil esines sündmus uuringu jooksul, ja $\delta = 0$ tähendab, et

objekt on tsenseeritud ehk objektil ei esinenud uuringus viibitud aja jooksul sündmust. (Kleinbaum & Mitchel, 2005, lk 8)

Elulemusfunktsiooni tähistatakse $S(t)$ ning see näitab üleelamistõenäosust ajahetkel $t$ ehk tõenäosust, et objekti elulemusaeg on suurem kui $t$. Elulemusfunktsiooni puhul kehtivad üldiselt järgmised väited. (Kleinbaum & Mitchel, 2005, lk 8-9)

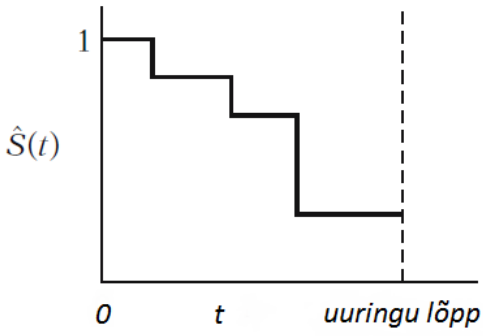
1. $S(t)$ on mittekasvav funktsioon.
2. Ajahetkel $t = 0$, $S(t) = 1$ ehk uuringu alguses, kui ühelgi objektil sündmust pole esinenud, on tõenäosus, et objekti elulemusaeg on suurem kui 0, üks.
3. Ajahetkel $t = \infty$, $S(t) = 0$ ehk kui uuringu pikkus kasvaks ajaliselt piiramatult, ei oleks lõpuks ühtegi objekti elus ning elulemusfunktsioon saaks väärtuse 0.

Paneme tähele, et 3. punkt ei pruugi kehtida näiteks laenude kontekstis, sest isegi kui uuringu pikkus kasvaks piiramatult, on võimalik, et kõik laenud ei pankrotistu.




**Joonis 1.1 Elulemusfunktsioon (Kleinbaum & Mitchel, 2005, lk 9)**

Elulemusfunktsiooni hinnang omandab lõpliku valimi korral enamasti treppfunktsiooni kuju. Kuna uuringu pikkus on ajaliselt piiratud, on võimalik, et igal objektil sündmust ei esine ning funktsioon ei lähe nulli. Näide elulemusfunktsiooni hinnangust on joonisel 1.2.




**Joonis 1.2. Näide elulemusfunktsiooni hinnangust (Kleinbaum & Mitchel, 2005, lk 9)**

Peale elulemusfunktsiooni, mis keskendub sündmuse mitteesinemisele, on kasutusel ka riskifunktsioon (*hazard function*), mis põhineb just sündmuse esinemisel. Riskifunktsiooni tähistatakse *h*(*t*) ning see näitab tõenäosust, et objektil esineb sündmus lõpmatult väikeses ajavahemikus $[t, t + \Delta t]$ tingimusel, et objekti elulemusaeg on vähemalt *t*. Riskifunktsiooni valem on kujul

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t},$$

kus $\Delta t$ tähistab lühikest ajavahemikku. (Kleinbaum & Mitchel, 2005, lk 9-10) Mida suurem on riskitihedus (*hazard rate*) ehk riskifunktsiooni väärtus ajahetkede $t_1$ ja $t_2$ vahel, seda suurem on tõenäosus, et objektil esineb selles ajaintervallis sündmus (Ritesh & Mukhopadhyay, 2011).

Kuna riskifunktsioon võib omada väärtusi $[0, \infty)$, siis on see pigem määr kui tõenäosus. Riskifunktsiooni väärtus sõltub sellest, kas aeg on mõõdetud päevades, nädalates, kuudes vm. Olgu näiteks $P = P(t \leq T < t + \Delta t | T \geq t) = \frac{1}{3}$. Kui $\Delta t = \frac{1}{2}$ päeva, siis on riskifunktisooni väärtus $\frac{P}{\Delta t} = \frac{2}{3} = 0.67$ päeva kohta. Kui $\Delta t = \frac{1}{14}$ nädalat, on aga riskifunktsiooni väärtus $\frac{P}{\Delta t} = \frac{14}{3} = 4.67$ nädala kohta. (Kleinbaum & Mitchel, 2005, lk 11)

## 1.2 Kaplan-Meieri meetod

Kaplan-Meieri meetodit kasutatakse elulemusfunktsioonile hinnangu leidmiseks. Olgu antud objekti jälgimisaja pikkuse $t_{(j)}$ järgi järjestatud andmestik, kus ajahetkel $t_{(j)}$ esinenud

sündmuste arv on $m_j$, ajavahemikus $[t_{(j)}, t_{(j+1)})$ tsenseeritud objektide arv on $q_j$ ja riskigrupi suurus $n_j$, mis sisaldab objekte, mille jälgimisaeg on vähemalt $t_{(j)}$. (Kleinbaum & Mitchel, 2005, lk 50)

**Tabel 1.1. Näide andmestikust, mille põhjal leitakse Kaplan-Meieri hinnang**

| järjestatud jälgimisajad, $t_{(j)}$ | sündmuste esinemiste arv, $m_j$ | tsenseeritud objektide arv vahemikus $[t_{(j)}, t_{(j+1)})$, $q_j$ | riskigrupi suurus, $n_j$ |
|---|---|---|---|
| $t_{(0)} = 0$ | $m_0 = 0$ | $q_0$ | $n_0$ |
| $t_{(1)}$ | $m_1$ | $q_1$ | $n_1$ |
| $t_{(2)}$ | $m_2$ | $q_2$ | $n_2$ |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| $t_{(k)}$ | $m_k$ | $q_k$ | $n_k$ |

Kaplan-Meieri hinnang elulemusfunktsioonile saadakse järgmise arvutusvalemiga (Tableman & Kim, 2005, lk 28):

$$\hat{S}(t_{(j)}) = \prod_{t_{(i)} \le t_{(j)}} \frac{n_i - m_i}{n_i}, \tag{1.1}$$

kus riskigrupp $n_j = n_{j-1} - m_{j-1} - q_{j-1}$ ning kuna ajahetkel $t$ = 0 kuuluvad kõik valimi objektid riskgruppi, siis on $n_0$ võrdne valimimahuga.

**Näide 1.1.** Kaplan-Meieri hinnang elulemusfunktsioonile

Vaatluse all on 12 laenuvõtjat ning uuritakse nende elulemusaegu – aega laenu saamisest pankrotistumiseni (päevades). Uuringu alguses kuuluvad kõik riskigruppi, st igaühel neist võib huvipakkuv sündmus esineda. Kliendid, kelle elulemusaja kohta on vaid osaline informatsioon, sest nad on laenu võtnud hiljuti ning uuring lõppes enne, kui saadi teada, kas neil sündmus esines, on tsenseeritud.

Klientide jälgimisajad järjestatakse kasvavalt ning märgitakse, mitu sündmust või tsenseerimist vastaval ajahetkel toimus. Nagu tabelist 1.2 on näha, on kõige lühem jälgimisaeg 27 päeva ning vastav klient on tsenseeritud. Elulemusfunktsiooni hinnangu väärtus on ikka 1, sest ükski klient pole antud hetkeks pankrotistunud.

Pärast seda kuulub riskigruppi 11 inimest, sest tsenseeritud kliendi kohta enam teavet pole. Kahe järgneva kliendi jälgimisaeg on 49 päeva: üks neist pankrotistus ja teine on tsenseeritud. Elulemusfunktsiooni hinnangu väärtus on nüüd 0.9091, mis on saadud eelneva ajahetke elulemusfunktsiooni väärtuse $\hat{S}(27) = 1$ korrutamisel $\frac{11-1}{11}$-ga, sest 11-st riskigruppi kuuluvast kliendist 1 pankrotistus.

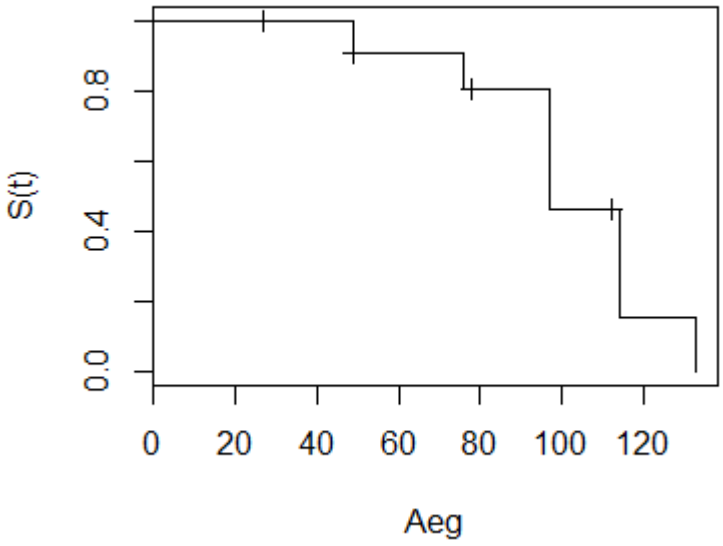
Vaatleme järgmisena jälgimisaega pikkusega 97 päeva. Näeme, et riskigruppi kuulub sel hetkel 7 inimest, kellest 3 pankrotistus. Elulemusfunktsiooni hinnangu väärtus saadakse valemi (1.1) abil: $\frac{12-0}{12} \cdot \frac{11-1}{11} \cdot \frac{9-1}{9} \cdot \frac{8-0}{8} \cdot \frac{7-3}{7} = 0.4618$ või lühemalt, eelneva ajahetke elulemusfunktsiooni väärtuse $\hat{S}(78) = 0.8081$ korrutamisel $\frac{7-3}{7}$-ga.

Viimasel kliendil kulus pankrotistumiseni 133 päeva. Kuna ta on ainuke, kes riskigruppi kuulub, sest ülejäänud kliendid on selleks hetkeks pankrotistunud või oleme nende kohta informatsiooni kaotanud, saab elulemusfunktsiooni hinnang tema pankrotistumisega väärtuse 0, $P(T > 133) = 0$.

**Tabel 1.2. Selgitav tabel elulemusfunktsiooni hinnangute leidmisest**

| $t_{(j)}$ | $n_j$ | $m_j$ | $q_j$ | $\hat{S}(t_{(j)})$ |
|---|---|---|---|---|
| 0 | 12 | 0 | 0 | 1 |
| 27 | 12 | 0 | 1 | $1 \cdot \frac{12-0}{12} = 1$ |
| 49 | 11 | 1 | 1 | $1 \cdot \frac{11-1}{11} = 0.9091$ |
| 76 | 9 | 1 | 0 | $0.9091 \cdot \frac{9-1}{9} = 0.8081$ |
| 78 | 8 | 0 | 1 | $0.8081 \cdot \frac{8-0}{8} = 0.8081$ |
| 97 | 7 | 3 | 0 | $0.8081 \cdot \frac{7-3}{7} = 0.4618$ |
| 112 | 4 | 0 | 1 | $0.4618 \cdot \frac{4-0}{4} = 0.4618$ |
| 114 | 3 | 2 | 0 | $0.4618 \cdot \frac{3-2}{3} = 0.1539$ |
| 133 | 1 | 1 | 0 | $0.1539 \cdot \frac{1-1}{1} = 0$ |

Andmetele vastav elulemusfunktsioon on joonisel 1.3.




**Joonis 1.3. Elulemuskõver 12 laenuvõtja andmete põhjal**

## 1.3 *Log-rank* test elulemuskõverate võrdlemiseks

Üks võimalus välja selgitamaks, kas kaks või enam elulemuskõverat on statistiliselt oluliselt erinevad, on *log-rank* test. *Log-rank* test on hii-ruut testi vorm, mille statistik kasutab sündmuste esinemiste tegelike ja oodatud arvude vahet igal erineval järjestatud jälgimisajal, mis analüüsitavasse andmestikku kuulub. (Kleinbaum & Mitchel, 2005, lk 58)

### 1.3.1 Kahe elulemuskõvera võrdlemine

Järgnev alapeatükk põhineb Kleinbaumi ja Mitcheli raamatul „Survival Analysis: A Self-Learning Text“ (2005, lk 58-61).

Vaatleme esmalt juhtu, kus võrreldakse kahte elulemuskõverat. Kui elulemus gruppides ei erine, siis peaks mõlemas grupis igal ajahetkel aset leidnud sündmuste arv olema proportsionaalne vastava riskigrupi suurusega sel ajahetkel. Oodatud sündmuste esinemiste arvud leitakse mõlemas grupis igal jälgimisajal järgmiste valemitega:

$$e_{1j} = \left(\frac{n_{1j}}{n_{1j} + n_{2j}}\right) \cdot \left(m_{1j} + m_{2j}\right)$$

$$e_{2j} = \left(\frac{n_{2j}}{n_{1j} + n_{2j}}\right) \cdot \left(m_{1j} + m_{2j}\right),$$

kus *j* tähistab jälgimisaja järjekorranumbrit, $\frac{n_{1j}}{n_{1j}+n_{2j}}$ ja $\frac{n_{2j}}{n_{1j}+n_{2j}}$ näitavad kahe riskigrupi proportsioone igal jälgimisajal ning $m_{1j} + m_{2j}$ sündmuste esinemiste arvu kahes grupis kokku igal jälgimisajal.

Seejärel leitakse mõlemas grupis sündmuste esinemiste tegelike ja oodatud arvude vahede summa üle kõigi jälgimisaegade valemiga $O_i - E_i = \sum_{j=1}^{k}\left(m_{ij} - e_{ij}\right)$, kus $i$ = 1, 2 tähistab võrreldavaid gruppe ja *k* seda, mitu erinevat jälgimisaega andmestikus on.

*Log-rank* statistiku arvutamisel võib kasutada emba-kumba gruppi, sest statistiku väärtus on mõlema grupi puhul sama, ning valem on kujul

$$L = \frac{(O_i - E_i)^2}{Var(O_i - E_i)},$$

kus $i$ = 1, 2.

Dispersioon summeeritud sündmuste tegelike ja oodatud arvude vahele leitakse valemiga

$$Var(O_i - E_i) = \sum_j \frac{n_{1j} n_{2j} (m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{\left(n_{1j} + n_{2j}\right)^2 (n_{1j} + n_{2j} - 1)},$$

kus $i$ = 1, 2 tähistab võrreldavaid gruppe, $j$ jälgimisaja järjekorranumbrit, $n_{ij}$ riskigrupi suurust ja $m_{ij}$ sündmuste esinemiste arvu *i*-ndas grupis *j*-ndal jälgimisajal.

Kontrollitav hüpoteesipaar on järgmine:

$H_0$: elulemuskõverad ei ole statistiliselt oluliselt erinevad

$H_1$: elulemuskõverad on statistiliselt oluliselt erinevad

*Log-rank* statistik on $H_0$ kehtides ligikaudu hii-ruut jaotusega, vabadusastmete arvuga 1.

### 1.3.2 Rohkem kui kahe elulemuskõvera võrdlemine

Järgnev alapeatükk põhineb Kleinbaumi ja Mitcheli raamatul „Survival Analysis: A Self-Learning Text“ (2005, lk 61-62, 82).

Rohkem kui kahe elulemuskõvera võrdlemisel on *log-rank* statistik keerulisem, sisaldades nii $O_i - E_i$ dispersioone kui kovariatsioone iga grupi jaoks, ning selle valem esitatakse enamasti maatrikskujul. *Log-rank* statistik on hii-ruut jaotusega, vabadustastmete arvuga $G - 1$, kus $G$ tähistab võrreldavate kõverate (gruppide) arvu.

Olgu $i = 1, 2, ..., G$ võrreldavate kõverate arv ja $j = 1, 2, ..., k$ erinevate jälgimisaegade arv. Tähistagu $n_{ij}$ riski all olevate objektide arvu *i*-ndas grupis *j*-ndal järjestatud jälgimisajal, $m_{ij}$ sündmuste esinemiste arvu *i*-ndas grupis *j*-ndal järjestatud jälgimisajal ja $e_{ij}$ oodatud sündmuste arvu *i*-ndas grupis *j*-ndal järjestatud jälgimisajal. Ajahetke *j* korral on kogu riskigrupi suurus $n_j = \sum_{i=1}^{G} n_{ij}$ ja kogu sündmuste esinemiste arv $m_j = \sum_{i=1}^{G} m_{ij}$. Sündmuste esinemiste tegelike ja oodatud arvude vahe igas grupis avaldub endiselt $O_i - E_i = \sum_{j=1}^{k}(m_{ij} - e_{ij})$. Selle dispersioon ja kovariatsioon on aga kujul

$$Var(O_i - E_i) = \sum_j \frac{n_{ij}(n_j - n_{ij})m_j(n_j - m_j)}{{n_j}^2(n_j - 1)},$$

$$Cov(O_i - E_i,\ O_l - E_l) = \sum_j \frac{-n_{ij}n_{lj}m_j(n_j - m_j)}{{n_j}^2(n_j - 1)}.$$

Komponentide vektor $(O_1 - E_1, O_2 - E_2, \dots, O_G - E_G)$ on lineaarselt sõltuv, sest $\sum_i(O_i - E_i) = 0$. Teststatistiku konstrueerimisel valitakse neist elementidest $G - 1$ tükki ja saadakse vektor $\boldsymbol{d} \coloneqq (O_1 - E_1, O_2 - E_2, \dots, O_{G-1} - E_{G-1})$.

Olgu $\boldsymbol{V} = (v_{il})$: $(G - 1) \times (G - 1)$ kovariatsioonmaatriks, kus $v_{ii} = Var(O_i - E_i)$, $v_{il} = Cov(O_i - E_i,\ O_l - E_l)$ ning $i, l = 1, 2, ..., G - 1$.

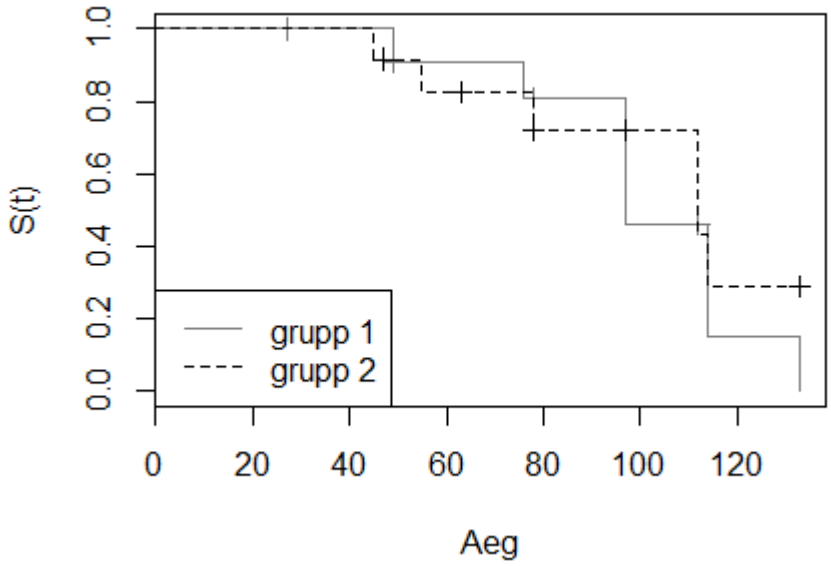
*Log-rank* statistik avaldub siis kujul $L = \boldsymbol{d}\boldsymbol{V}^{-1}\boldsymbol{d}^T$, mis on $H_0$ kehtides ligikaudu hii-ruut jaotusega, vabadusastmete arvuga $G - 1$.

**Näide 1.2.** *Log-rank* testi kasutamine elulemuskõverate erinevuse tuvastamiseks

Olgu vaatluse all kaks gruppi laenuvõtjaid, kummaski 12 klienti. Huvipakkuv sündmus on laenu pankrotistumine. Gruppidele vastavad elulemuskõverad on kujutatud joonisel 1.4. Eesmärk on välja selgitada, kas elulemuskõverad on statistiliselt oluliselt erinevad, kasutades selleks *log-rank* testi. Vastav hüpoteesipaar on järgmine:

$H_0$: elulemuskõverad ei ole statistiliselt oluliselt erinevad

$H_1$: elulemuskõverad on statistiliselt oluliselt erinevad




**Joonis 1.4. Kahte gruppi kuuluvate laenuvõtjate elulemuskõverad**

*Log-rank* statistiku valem on kujul $L = \frac{(O_i - E_i)^2}{Var(O_i - E_i)}$ , kus $i$ = 1, 2 tähistab võrreldavaid gruppe. Kuna statistiku väärtus on mõlema grupi puhul sama, võime selle leidmiseks kasutada näiteks esimest gruppi ehk $L = \frac{(O_1 - E_1)^2}{Var(O_1 - E_1)}$ .

Lugeja $(O_1 - E_1)^2$ leidmiseks on vaja summeerida realiseerunud sündmuste arv ja oodatud sündmuste arv esimeses grupis. Oodatud sündmuste arvu leidmiseks kasutame eespool tutvustatud valemit $e_{1j} = \left(\frac{n_{1j}}{n_{1j}+n_{2j}}\right) \cdot \left(m_{1j} + m_{2j}\right)$.

Tabelist 1.3 on näha, et esimese grupi lühim jälgimisaeg on 27 päeva ning vastaval ajahetkel sündmust ei toimunud, järelikult on klient tsenseeritud. Riski alla kuuluvad selle ajahetke

alguses kõik kliendid, mõlemas grupis 12 klienti. Oodatud sündmuste arv esimeses grupis ajahetkel 27 on seega $e_{11} = \frac{12}{(12+12)} \cdot (0 + 0) = 0$.

Järgmisel jälgimisajal, mille pikkus on 45 päeva, ei toimunud esimeses grupis ühtegi sündmust ega tsenseerimist, küll aga pankrotistus teises grupis üks klient. Riskigrupi suurus on selle ajahetke alguses esimese grupi jaoks 11, sest üks klient on tsenseeritud, ning teise grupi jaoks 12, sest ühegi kliendiga pole selleks hetkeks midagi juhtunud. Oodatav sündmuste arv teisel ajahetkel esimese grupi jaoks on $e_{12} = \frac{11}{(11+12)} \cdot (0 + 1) = 0.4783$.

Esimeses grupis toimus esimene sündmus 49. päeval, teises grupis samal ajal sündmusi ei toimunud. Riskigrupi suurused on selleks ajahetkeks vastavalt 11 ja 10 klienti. Oodatud sündmuste arv esimeses grupis on $e_{14} = \frac{11}{(11+10)} \cdot (1 + 0) = 0.5231$.

Vaatleme järgmisena 11. jälgimisaega, mille pikkus on 114 päeva. Sel hetkel kuulub riski alla nii esimeses kui teises grupis 3 klienti. Esimeses grupis toimub 2 pankrotistumist, teises grupis 1 pankrotistumine. Oodatud sündmuste arv esimeses grupis on $e_{1,11} = \frac{3}{(3+3)} \cdot (2 + 1) = 1.5$.

Nagu tabelist 1.3 näha, toimus esimeses grupis kokku 8 sündmust ehk $O_1 = \sum_{j=1}^{12} m_{1j} = 8$. Oodatud sündmuste arv oli $E_1 = \sum_{j=1}^{12} e_{1j} = 6.8428$. Lugeja väärtus on seega $(O_1 - E_1)^2 = (8 - 6.8428)^2 = 1.1572^2 = 1.3391$.

Nimetaja leidmiseks kasutame eespool kirjeldatud valemit $Var(O_i - E_i) = \sum_j \frac{n_{1j} n_{2j} (m_{1j}+m_{2j})(n_{1j}+n_{2j}-m_{1j}-m_{2j})}{(n_{1j}+n_{2j})^2 (n_{1j}+n_{2j}-1)}$ , mille väärtus on mõlema grupi puhul sama. Leiame ka selle esimese grupi jaoks ehk $Var(O_1 - E_1) = \frac{11 \cdot 12(0+1)(11+12-0-1)}{(11+12)^2(11+12-1)} + \frac{11 \cdot 10(1+0)(11+10-1-0)}{(11+10)^2(11+10-1)} + \cdots + \frac{1 \cdot 2(1+0)(1+2-1-0)}{(1+2)^2(1+2-1)} = 2.9730$.

*Log-rank* statistiku väärtus on seega $L = \frac{1.3391}{2.9730} = 0.4504$, mis peaks $H_0$ kehtides olema realisatsioon vabadusastmete arvuga 1 hii-ruut jaotusest. Vastav *p*-väärtus on 0.5021. Kuna $p > 0.05$, tuleb harikult kasutatava olulisusnivoo korral jääda $H_0$ juurde ehk elulemuskõverate vahel ei saa tõestada statistiliselt olulist erinevust.

**Tabel 1.3. Selgitav tabel *log-rank* statistiku leidmisest**

| | | sündmuste esinemiste arv | | riskigrupi suurus | | oodatud sündmuste arv | | tegelik - oodatud | |
|---|---|---|---|---|---|---|---|---|---|
| $j$ | $t_j$ | $m_{1j}$ | $m_{2j}$ | $n_{1j}$ | $n_{2j}$ | $e_{1j}$ | $e_{2j}$ | $m_{1j} - e_{1j}$ | $m_{2j} - e_{2j}$ |
| 1 | 27 | 0 | 0 | 12 | 12 | (12/24)·0 | (12/24)·0 | 0 | 0 |
| 2 | 45 | 0 | 1 | 11 | 12 | (11/23)·1 | (12/23)·1 | -0.4783 | 0.4783 |
| 3 | 47 | 0 | 0 | 11 | 11 | (11/22)·0 | (11/22)·0 | 0 | 0 |
| 4 | 49 | 1 | 0 | 11 | 10 | (11/21)·1 | (10/21)·1 | 0.4762 | -0.4762 |
| 5 | 55 | 0 | 1 | 9 | 10 | (9/19)·1 | (10/19)·1 | -0.4737 | 0.4737 |
| 6 | 63 | 0 | 0 | 9 | 9 | (9/18)·0 | (9/18)·0 | 0 | 0 |
| 7 | 76 | 1 | 0 | 9 | 8 | (9/17)·1 | (8/17)·1 | 0.4706 | -0.4706 |
| 8 | 78 | 0 | 1 | 8 | 8 | (8/16)·1 | (8/16)·1 | -0.5 | 0.5 |
| 9 | 97 | 3 | 0 | 7 | 6 | (7/13)·3 | (6/13)·3 | 1.3846 | -1.3846 |
| 10 | 112 | 0 | 2 | 4 | 5 | (4/9)·2 | (5/9)·2 | -0.8889 | 0.8889 |
| 11 | 114 | 2 | 1 | 3 | 3 | (3/6)·3 | (3/6)·3 | 0.5 | -0.5 |
| 12 | 133 | 1 | 0 | 1 | 2 | (1/3)·1 | (2/3)·1 | 0.6667 | -0.6667 |
| Kokku | | 8 | 6 | | | 6.8428 | 7.1572 | 1.1572 | -1.1572 |

**Näide 1.3.** Eelnev näide R-is läbiviiduna

R-i funktsioon `survdiff` sooritab vaikimisi *log-rank* testi. Argumendiks vajab ta `Surv` objekti. Loeme näitele vastava andmestiku R-i ja rakendame funktsiooni `survdiff`. Veendume, et saame samasuguse tulemuse nagu käsitsi arvutades.

```
> aeg <- c(27,49,49,76,78,97,97,97,112,114,114,133, #grupp1
          45,47,55,63,78,78,97,112,112,114,133,133)  #grupp2
> pankrot <- c(0,1,0,1,0,1,1,1,0,1,1,1,
               1,0,1,0,1,0,0,1,1,1,0,0) #1-sündmus,
                                        #0-tsenseerimine
> grupp <- c(rep(1,12),rep(2,12))
> survtest <- survdiff(Surv(aeg,pankrot)~grupp)
```

```
            N Observed Expected (O-E)^2/E (O-E)^2/V
grupp=1    12        8     6.84     0.196      0.45
grupp=2    12        6     7.16     0.187      0.45
Chisq= 0.5 on 1 degrees of freedom, p= 0.502
```

Näeme, et *log-rank* statistiku väärtus on ka R-i tulemustes 0.45 ning *p*-väärtus 0.502.

## 1.4 Coxi võrdeliste riskide mudel

Coxi võrdeliste riskide mudel (*Cox proportional hazards model*) on laialt kasutatav mudel elulemusanalüüsis. Selle eesmärk on leida elulemust prognoosivaid tunnuseid ning nende mõju riskifunktsioonile (Walters, 2009). Coxi mudel esitataksegi riskifunktsiooni kaudu, mis on kujul:

$$h(t|\boldsymbol{X}) = h_0(t)e^{\sum_{i=1}^{p}\beta_i X_i},$$

kus $t$ on ajahetk, mille jaoks objekti riskitihedust arvutatakse, $h_0(t)$ on baasriskifunktsioon, $\mathbf{X} = (X_1, X_2, \dots, X_p)$ on kirjeldavate tunnuste vektor ja $\beta_i$, $i$ = 1, 2, …, $p$ on regressioonikordajad (Kleinbaum & Mitchel, 2005, lk 94). Paneme tähele, et baasriskifunktsioon sõltub ajast $t$ ja mitte kirjeldavatest tunnustest $X_i$, $i$ = 1, 2, …, $p$. Seevastu eksponentosa sõltub kirjeldavatest tunnustest, mitte aga ajast $t$ ehk kirjeldavad tunnused on ajast sõltumatud.

Coxi mudeli oluline eeldus on riskitiheduste võrdelisus, mis tähendab, et riskitiheduste suhe on ajas muutumatu. Selle selgitamiseks vaatleme hinnangut riskitiheduste suhtele Coxi mudeli kehtides ja veendume, et see on konstantne.

$$\widehat{HR} = \frac{\hat{h}(t, \boldsymbol{X}^*)}{\hat{h}(t, \boldsymbol{X})},$$

kus $\boldsymbol{X}^* = (X_1^*, X_2^*, \dots, X_p^*)$ ja $\mathbf{X} = (X_1, X_2, \dots, X_p)$ on kahe objekti kirjeldavate tunnuste vektorid. Eelnevalt kirja pandud Coxi riskitiheduse valemi põhjal saamegi, et

$$\widehat{HR} = \frac{\hat{h}(t, \boldsymbol{X}^*)}{\hat{h}(t, \boldsymbol{X})} = \frac{\widehat{h_0(t)}\exp(\sum \widehat{\beta}_i {X_i}^*)}{\widehat{h_0(t)}\exp(\sum \widehat{\beta}_i X_i)} = \exp\left[\sum_{i=1}^{p} \widehat{\beta}_i (X_i^* - X_i)\right] = const$$

ning ei sõltu ajahetkest *t*. (Kleinbaum & Mitchel, 2005, lk 107)

Coxi võrdeliste riskide mudel on poolparameetriline mudel, sest baasriskifunktsioon on määratlemata. Hoolimata sellest annab Coxi mudel küllaltki häid hinnanguid regressioonikordajatele ja riskitiheduste suhetele ning need on ligilähedased tulemustele, mis saadaks õiget parameetrilist mudelit kasutades. (Kleinbaum & Mitchel, 2005, lk 95-96)

### 1.4.1 Parameetrite hindamine

Regressioonikordajad $\beta_i$, $i = 1, 2, \ldots, p$ hinnatakse osalise tõepära meetodil, mida käsitletakse samamoodi nagu suurima tõepära meetodit (Klein & Moeschberger, 2003, lk 253). Osaliseks nimetatakse seda sellepärast, et Coxi mudeli puhul ei ole täpsustatud uuritava tunnuse jaotus ning tõepära põhineb järjestatud sündmuste esinemiste aegadel, mitte aga nende jaotusel (Kleinbaum & Mitchel, 2005, lk 111).

Osalise tõepära funktsioon on kujul

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{k} \frac{\exp[\sum_{i=1}^{p} \beta_i X_{(j)i}]}{\sum_{h \epsilon R(t_{(j)})} \exp[\sum_{i=1}^{p} \beta_i X_{hi}]},$$

kus $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ on hinnatavate parameetrite vektor, $j = 1, 2, \ldots, k$ on sündmuste esinemisaegade järjekorranumbrid, $i = 1, 2, \ldots, p$ on mudeli kirjeldavate tunnuste arv, $X_{(j)i}$ on objekti, kelle sündmus esineb ajahetkel $t_j$, *i*-nda tunnuse väärtus, $h \epsilon R(t_{(j)})$ tähistab riskigruppi kuuluvaid objekte ajahetkel $t_j$ ja $X_{hi}$ on *h*-nda objekti *i*-nda tunnuse väärtus. Paneme tähele, et osalise tõepära funktsioon ei sisalda baasfunktsiooni $h_0(t)$, seega ei ole seda vaja parameetrite hindamiseks teada. (Klein & Moeschberger, 2003, lk 253)

Osalise tõepära funktsioonist leitakse logaritmiline tõepärafunktsioon $l(\boldsymbol{\beta}) = \ln[L(\boldsymbol{\beta})]$, millest võetakse osatuletised $\beta_b$, $b = 1, 2, \ldots, p$ järgi. Osalise tõepära hinnangud parameetritele leitakse, kui lahendatakse võrrand, kus osatuletised on võrdsustatud nulliga iga $b = 1, 2, \ldots, p$ jaoks. Seda tehakse mõnda iteratiivset meetodit, nt Newtoni meetodit kasutades. (Klein & Moeschberger, 2003, lk 253-254)

Regressioonikordajad, mille hinnatud väärtused on positiivsed, suurendavad riskitihedust. See tähendab, et vastavad tunnused suurendavad objekti sündmuse esinemise tõenäosust lõpmatult väikeses ajavahemikus $[t, t + \Delta t]$, tingimusel, et objekti elulemusaeg on vähemalt *t*. Seevastu negatiivsed regressioonikordajad vähendavad riskitihedust ning mida suuremad on vastavate tunnuste väärtused, seda väiksem on tõenäosus, et objektil esineb lõpmatult väikeses ajavahemikus sündmus.

### 1.4.2 Parameetrite olulisus

Kõige levinumad statistikud Coxi mudeli parameetrite olulisuse leidmiseks on Waldi statistik ja tõepärasuhte statistik. Waldi statistik on kujul $z = \frac{\widehat{\beta}}{\widehat{SE}(\widehat{\beta})}$ ning see on $H_0$ kehtides standardse normaaljaotusega. Ka rakendustarkvara R väljastab Coxi mudeli puhul automaatselt Waldi statistiku ja sellele vastava *p*-väärtuse. Tõepärasuhte statistik on kujul $LR = 2\ln\left(\frac{L_0}{L_1}\right) = 2\ln(L_0) - 2\ln(L_1)$, kus $L_0$ on esialgse mudeli tõepärafunktsiooni väärtus ning $L_1$ lihtsama mudeli, mis on saadud esialgsest mudelist mõne parameetri fikseerimisel, tõepärafunktsiooni väärtus. (Kleinbaum & Mitchel, 2005, lk 89)

Tõepärasuhte statistik on $H_0$ kehtides hii-ruut jaotusega, vabadusastmega *k*, kus *k* tähistab parameetrite arvu, mis on eemaldatud esialgsest mudelist, saamaks lihtsam mudel (Kleinbaum & Mitchel, 2005, lk 90). Kui vastav *p*-väärtus on väiksem kui olulisuse nivoo $\alpha = 0.05$, siis kummutatakse nullhüpotees ja öeldakse, et esialgne mudel on oluliselt parem kui lihtsam mudel ehk parameetrite eemaldamine ei olnud õigustatud. Seega kui esialgsest mudelist eemaldatakse lihtsama mudeli saamiseks ainult üks parameeter, saamegi *p*-väärtuse põhjal teada eemaldatud parameetri olulisuse.

Waldi ja tõepärasuhte statistikud ei pruugi alati samu vastuseid anda. Küll aga on teada, et tõepärasuhte statistik on paremate statistiliste omadustega, seega tasub kahtluse korral eelistada just seda. (Kleinbaum & Mitchel, 2005, lk 90)

### 1.4.3 Riskitiheduste suhe

Lisaks parameetrite hinnangutele ja nende olulisusele ollakse huvitatud ka riskitiheduste suhte hinnangust. See näitab, mitu korda erinevad kahe võrreldava grupi riskitihedused ehk mitu korda erineb ühte gruppi kuuluva objekti sündmuse esinemise tõenäosus lõpmatult väikeses ajavahemikus $[t, t + \Delta t]$ võrreldes teise grupi objektiga, tingimusel, et objekt on elanud ajahetkeni *t*.

Näiteks mõne uue ravimi testimisel võidakse soovida hinnata platseebogrupi ja ravigrupi objektide riskitiheduste suhet. Kui riskitiheduste suhte hinnang on näiteks $\widehat{HR} = \frac{platseebogrupp}{ravigrupp} = 1.5$, siis see tähendab, et kui platseebogrupi objekt on elanud mingi kindla ajahetkeni, on tal ravigrupi objektiga võrreldes 1.5 korda suurem tõenäosus, et järgmises lõpmatult väikeses ajavahemikus ta sureb. (Duerden, 2009, lk 6)

Üldiselt on riskitiheduste suhte hinnang leitav eespool kirjeldatud valemiga $\widehat{HR} = \exp\left[\sum_{i=1}^{p} \widehat{\beta}_i (X_i^* - X_i)\right]$, kus $\boldsymbol{X}^* = (X_1^*, X_2^*, \dots, X_p^*)$ ja $\mathbf{X} = (X_1, X_2, \dots, X_p)$ on kahe objekti kirjeldavate tunnuste vektorid. Kui aga huvipakkuvaks tunnuseks on vaid (0, 1) tunnus, nt grupp, kuhu objekt kuulub, siis valem lihtsustub ja jääb kujule $\widehat{HR} = \exp\left[\widehat{\beta_1}(1 - 0)\right] = e^{\widehat{\beta_1}}$. (Kleinbaum & Mitchel, 2005, lk 100-101) See tähendab, et kui kõik ülejäänud tunnused on fikseeritud, on gruppi 1 kuuluvatel objektidel lõpmatult väikeses ajavahemikus $e^{\widehat{\beta_1}}$ korda suurem tõenäosus sündmuse esinemiseks.

**Näide 1.4.** Coxi võrdeliste riskide mudeli kasutamisest R-is

Olgu vaatluse all 40 laenuvõtjat, kes on jagatud vanuse järgi kahte gruppi: grupis 0 on need kliendid, kes on laenu võtnud 35-aastaselt või varem, ning grupis 1 on kliendid, kes on laenu võtnud hiljem kui 35-aastaselt. Iga kliendi kohta on märgitud tema jälgimisaeg ja info pankrotistumise kohta (0 – tsenseeritud, 1 – pankrotistus). Samuti on teada laenusumma, mis igale kliendile väljastati. Näide andmetest on lisas 1.

Loome esiteks mudeli, kus on tunnused vanusgrupp, väljastatud laenusumma ja vanusgrupi ning väljastatud laenusumma koosmõju.

```
> cox1 <- coxph(Surv(jalgimisaeg,pankrot) ~
     factor(vanus) + laenusumma + factor(vanus):laenusumma)
> cox1
                              coef exp(coef) se(coef)    z     p
factor(vanus)1           -18.8585 6.45e-09 10.94805 -1.72 0.085
laenusumma                 0.0131 1.01e+00  0.00519  2.53 0.011
factor(vanus)1:laenusumma 0.0192 1.02e+00  0.01329  1.44 0.150
```

Väljatrükist näeme, et vanusgrupi ja laenusumma koosmõju ei ole oluline, sest $p = 0.150$. Nagu eelnevalt öeldud, väljastab R automaatselt Waldi statistiku. Vaatame, millise tulemuse saame tõepärasuhte statistikuga. Selleks loome lihtsama mudeli, kus on vaid tunnused vanusgrupp ja laenusumma. Valemi $LR = 2\ln(L_0) - 2\ln(L_1)$ põhjal, kus $L_0$ on esialgse mudeli `cox1` tõepärafunktsiooni väärtus ning $L_1$ lihtsama mudeli `cox2` tõepärafunktsiooni väärtus, leiame tõepärasuhte statistiku väärtuse. See peaks $H_0$ kehtides olema realisatsioon vabadusastmega 1 hii-ruut jaotusest, sest esialgsest mudelist eemaldati lihtsama mudeli saamiseks üks parameeter. Seejärel leiame vastava *p*-väärtuse.

```
> cox2 <- coxph(Surv(jalgimisaeg,pankrot) ~
                    factor(vanus) + laenusumma)
> 2*cox1$loglik[2] - 2*cox2$loglik[2] # vastus 2.5701
> 1 - pchisq(2.5701,1) # p = 0.1089008
```

Tõepärasuhte statistikule vastav *p*-väärtus on ligikaudu 0.109 ja see ei ole võrdne Waldi statistikule vastava *p*-väärtusega, mis on 0.150. Mõlema statistiku puhul võetakse aga vastu sama otsus: vanusgrupi ja laenusumma koosmõju ei ole mudelis statistiliselt oluline ning eelistada tasub lihtsamat mudelit. Lihtsamale mudelile vastavad parameetrite hinnangud ja *p*-väärtused on järgnevad:

```
> cox2
                          coef exp(coef) se(coef)    z       p
factor(vanus)1         -3.3271   0.0359  0.76163 -4.37 1.3e-05
laenusumma              0.0174   1.0176  0.00468  3.73 1.9e-04
```

Selles mudelis on mõlemad parameetrid olulised: *p*-väärtused on väiksemad kui olulisusnivoo $\alpha = 0.05$. Näeme, et vanusgrupile vastava parameetri hinnang on -3.3271. Leiame riskitiheduste suhte hinnangu kahe vanusgrupi jaoks. Eelnevast teame, et kui huvi pakub vaid (0, 1) tunnus, siis on riskitiheduse suhte valem kujul $\widehat{HR} = e^{\widehat{\beta_1}}$. Seega meil $\widehat{HR} = e^{-3.3271} = 0.0359$, mis tähendab, et sellise andmestiku põhjal on vanusgruppi 0 kuuluval isikul lõpmatult väikeses ajavahemikus $\frac{1}{0.0359} = 27.86$ korda suurem tõenäosus pankrotistumiseks.

Tunnusele *laenusumma* vastava parameetri hinnang on 0.0174. Kuna see on positiivne arv, võib öelda, et mida suurem on kliendile väljastatud laen, seda suurem on riskitihedusfunktsiooni väärtus ehk seda suurem on kliendi pankrotistumise tõenäosus lõpmatult väikeses ajavahemikus. Vaatleme kahe samas vanusgrupis oleva kliendi riskitiheduste suhte hinnangut. Olgu ühele kliendile väljastatud 800 euro suurune laen ja teisele 700 euro suurune laen. Siis $\widehat{HR} = \exp\left[\widehat{\beta_1}(X_1^* - X_1)\right] = \exp[0.0174(800 - 700)] = 5.70$ ehk sellise andmestiku põhjal on 800 euro laenajal 5.7 korda suurem tõenäosus lõpmatult väikeses ajavahemikus pankrotistuda kui 700 euro laenajal.

# 2 Bondora (isePankur AS) laenude analüüs

## 2.1 Ülevaade andmestikust

Andmestik pärineb Bondora (isePankur AS) veebileheküljelt avalikust andmebaasist (Bondora, 2014). Bondora on ettevõte, mis pakub teenuseid nii väikelaenu laenajatele kui investoritele. Tegemist on mugavama alternatiiviga pangalaenule, mis võimaldab kiiresti taotleda tagatiseta väikelaenu. Bondora ise laene ei rahasta, seda teevad teised kasutajad.

Andmestikus on Bondora laenude toorandmed seisuga 01.12.2014, sisaldades 162 tunnust iga 22 447 laenuvõtja kohta. Huvipakkuvaid tunnuseid on 19:

- laenutaotluse rahastatus, kus 0 – ei rahastatud, 1 – rahastati (*WasFunded*)
- laenu pikkus (*LoanDuration*)
- laenu väljastamise kuupäev (*LoanDate*)
- laenu täieliku tagastamise kuupäev (*MaturityDate_Original*)
- laenu pankrotistumise kuupäev (*Default_StartDate*)
- maksimaalne intressimäär, mida laenutaotlus lubas (*Interest*)
- laenuvõtja sugu (*Gender*)
- laenuvõtja elukohariik (*Country*)
- laenuvõtja vanus (*Age*)
- laenuvõtja haridustase, kus 1 – algharidus, 2 – põhiharidus, 3 – kutseharidus, 4 – keskharidus, 5 – kõrgharidus (*education_id*)
- laenuvõtja töösuhe (*employment_status_id*)
- laenuvõtja on laenu pikendanud, kus 0 – ei ole pikendanud, 1 – on pikendanud (*CurrentLoanHasBeenExtended*)
- laenuvõtja Bondora krediidiajalugu, kus 0 – kliendil oli vähemalt 3 kuud varasemat krediidiajalugu, 1 – kliendil ei olnud varasemat krediidiajalugu (*NewCreditCustomer*)
- laenuvõtja tööstaaž (*work_experience*)
- laenuvõtja kogusissetulek (*income_total*)
- laenuvõtja vaba raha pärast igakuiste kohustuste täitmist (*FreeCash*)
- laenu otstarve (*UseOfLoan*)
- laenuvõtja saadud summa (*FundedAmount*)
- laenu kuumakse (*NewLoanMonthlyPayment*)

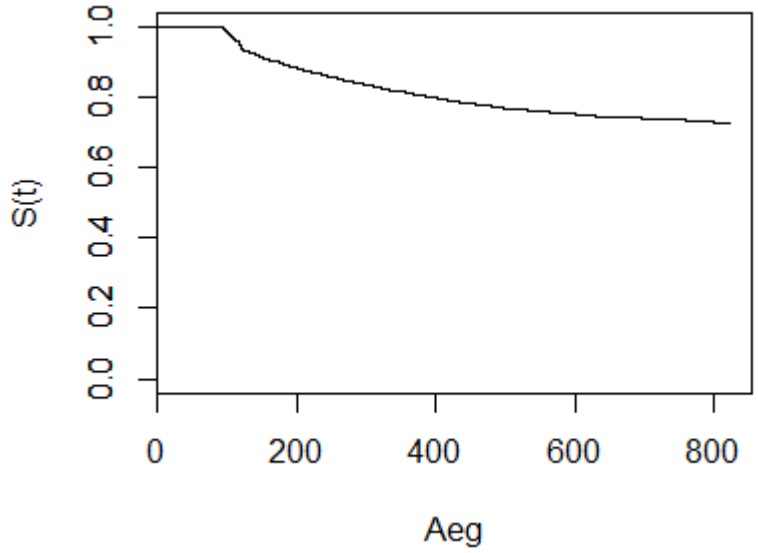
Analüüsiks moodustame sellest alamandmestiku, kus on vaid need kliendid, kelle laenutaotlus rahastati ning kellel pole puuduvaid väärtusi kuupäevadel, mil laen väljastati ja mil laen peaks täielikult tasutud saama. Alamandmestikus on 12 826 klienti.

Olgu välja toodud, et järgneva analüüsi käigus ei käsitleta eraldi laenude pikendajaid ning ei arvutata nende jaoks uusi jälgimisaegu. Samuti ei võeta arvesse pankrotistunud laenude taastumisi ehk summasid, mis nõutakse sisse pärast kliendi maksejõuetuks muutumist. Laenu ennetähtaegselt tagastajate korral loetakse teadaolevaks, et laen ei ole pankrotistunud enne kogu tagasimaksmise kuupäeva.

## 2.2 Elulemuskõverad

Esiteks vaatame, millise kujuga on kõikidele andmestikus olevatele laenudele vastav elulemuskõver, mis iseloomustab mittepankrotistunud laenude osakaalu. Selle jaoks märgime iga kliendi jaoks, kas ta pankrotistus või mitte, ning seejärel leiame iga kliendi jälgimisaja. Seejuures tuleb meeles pidada, et andmestik on seisuga 01.12.2014 ning kliendid, kes ei ole selleks ajaks pankrotistunud ja kelle laenu täieliku tagastamise kuupäev on pärast seda, on tsenseeritud. Lisaks on tsenseeritud tagasimaksjad. Kuna teame, et nendega pärast laenu tagastamist enam sündmust juhtuda ei saa, kuid soovime neid ikkagi uuringusse kaasata, märgime nende jälgimisajaks vaatlusperioodi pikkuse.

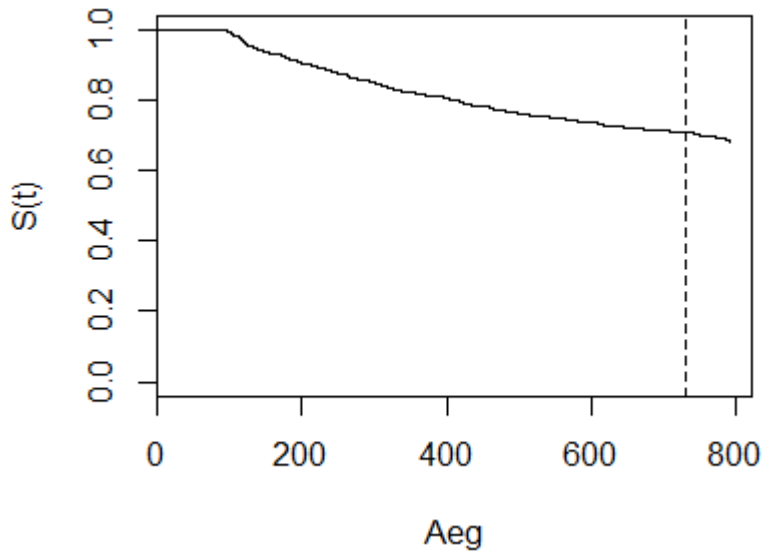
Seejärel visualiseerime andmetele vastava elulemuskõvera esimesed 2 aastat ja 3 kuud. Selline periood on valitud põhjusel, et varasemalt anti maksimaalselt 2 aasta pikkuseid laene ning esimesi pikemaid laene hakati väljastama 2012. aasta oktoobris. Seega on ka pikemad laenud saanud kesta vaid veidi üle 2 aasta. Üksikud hilisem punktid on erindid ning neid joonisele ei märgi.




**Joonis 2.1. Kaplan-Meieri hinnang kõikide laenude elulemuskõverale**

Vaatlusperioodi lõpuks on pankrotistunud hinnanguliselt 27% klientidest. Pärast 2 aastat ja 3 kuud on alles veel 32 klienti (0.25% esialgsetest klientidest), kellega pole sündmust toimunud. Neid võib käsitleda erinditena. Üle poolte neist on sellised, kelle laenu pikkuseks on määratud 24 kuud või vähem, kuid lepingu kohaselt võimaldatakse tagasimaksmiseks oluliselt pikemat perioodi, ulatudes 4 aastani. Samuti on nende hulgas kliente, kes pankrotistuvad, kuid mitte veidi pärast kaheaastast tagasimaksmise perioodi vaid oluliselt hiljem. Seega on alust arvata, et need kliendid on laenu pikendanud või on oma käitumiselt muudmoodi erilised, nt jätavad aeg-ajalt makseid tegemata, kuid mitte piisavalt, et neid kohe pankrotistujateks lugeda, ning seetõttu nende jälgimisaeg pikeneb.

Edaspidi vaatleme lähemalt neid kliente, kelle laenu pikkuseks on määratud 24 kuud. Seda põhjusel, et kliendid, kes on pikemaid laene võtnud, ei ole pidanud neid veel täielikult tagasi maksma. Jälgime klientide käitumist 26 kuu jooksul, sest lepingu sõlmimisel ei määrata tagasimaksmise perioodiks täpselt 730 päeva, periood võib olla ka veidi pikem ning pankrotistumine võib toimuda ka 2 kuud pärast laenu tähtaega. Vastav elulemuskõver on joonisel 2.2. Punktiirjoonega on märgitud 24 kuu piir.

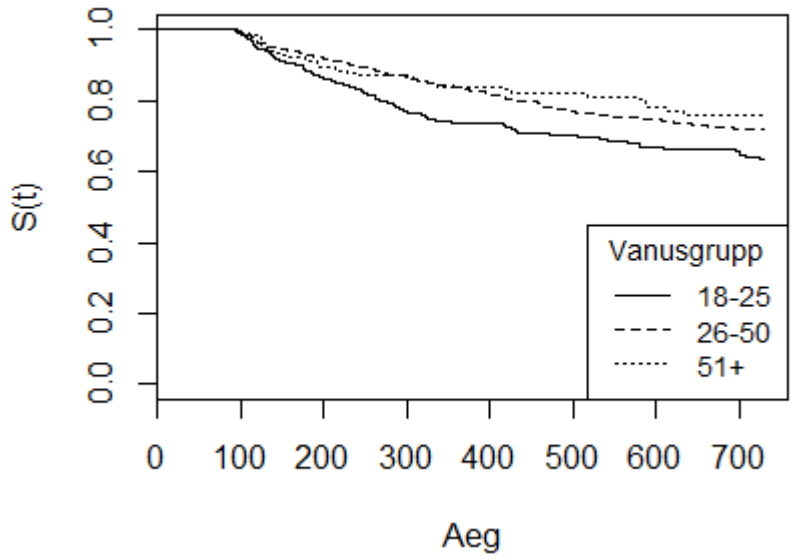



**Joonis 2.2. Kaplan-Meieri hinnang 24-kuuliste laenude elulemuskõverale**

Kahe aasta lõpuks pankrotistub hinnanguliselt 29% esialgsetest klientidest ning vaatlusperioodi lõpuks 32% esialgsetest klientidest. Pärast 791 päeva möödumist on alles veel 29 klienti (1.2% esialgsetest klientides), kes pole selleks ajaks laenu tagasi maksnud ega ka pankrotistunud. Neist 52% on laenu pikendanud ning 28% ei ole pikendanud ja pankrotistuvad hiljem. Seega võib järeldada, et enamiku klientidega toimub sündmus 26 kuu jooksul ning vähe on neid, kes laenu sellest kaugemale pikendavad või pärast 26 kuud pankrotistuvad.

## 2.3 *Log-rank* test elulemuskõverate erinevuse tuvastamiseks

Vaatleme joonistelt, kuidas erinevad elulemuskõverad 24 kuu pikkuste laenude hulgas sõltuvalt klientide vanusest, riigist ja haridustasemest. Seejärel kontrollime *log-rank* testiga, kas erinevus tõepoolest eksisteerib. Vaatlusperioodiks olgu nüüd täpselt 24 kuud.

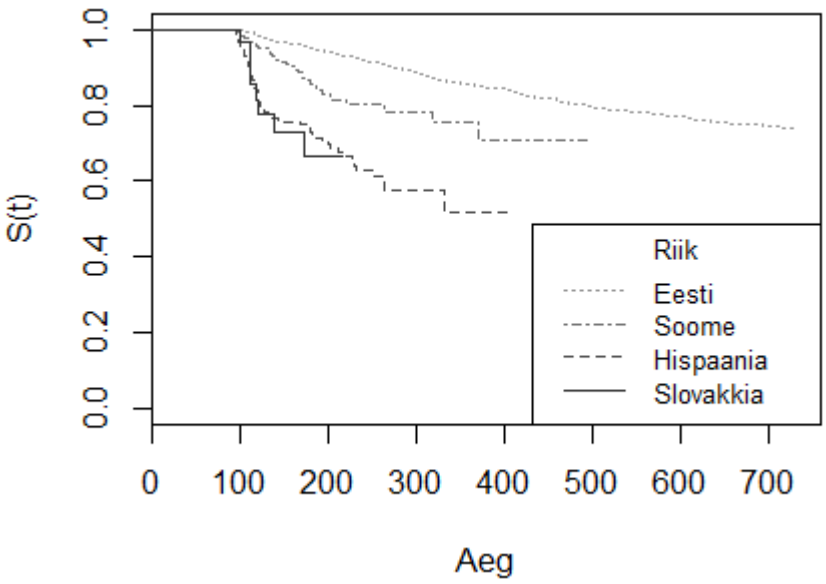



**Joonis 2.3. Erinevatest vanusgruppidest klientidele antud 24 kuu pikkuste laenude elulemuskõverad**

Jooniselt 2.3 näeme, et elulemusfunktsiooni hinnangu väärtused on kõige väiksemad 18–25-aastaste klientide puhul ning vaatlusperioodi lõpuks on maksejõulisi kliente hinnanguliselt 63% kõigist sellesse vanusgruppi kuuluvatest klientidest, kellele väljastati 24 kuu pikkune laen. 26–50-aastaste ja üle 50-aastaste klientide elulemuskõverad näivad sarnasemad olevat ning perioodi lõpuks on 26–50-aastaste klientide hulgas maksejõulisi 72% ja üle 50-aastaste klientide hulgas 76% esialgsetest klientidest.

Kõverate erinevust saame R-is testida käsuga `survdiff`, mis kasutab *log-rank* statistikut (vt koodi lisast 2). Statistikule vastav *p*-väärtus on ligikaudu 0.001 ning see kinnitab, et erinevatest vanusgruppidest klientidele antud laenude elulemuskõverad on statistiliselt oluliselt erinevad.

Kontrollime, kas tõestatav erinevus leidub ka 26–50-aastastele ja üle 50-aastastele klientidele antud laenude elulemuskõverate vahel. Kuna *log-rank* statistikule vastav *p*-väärtus on ligikaudu 0.62 (vt lisa 2), siis võib öelda, et nende kõverate vahel statistiliselt olulist erinevust ei ole.

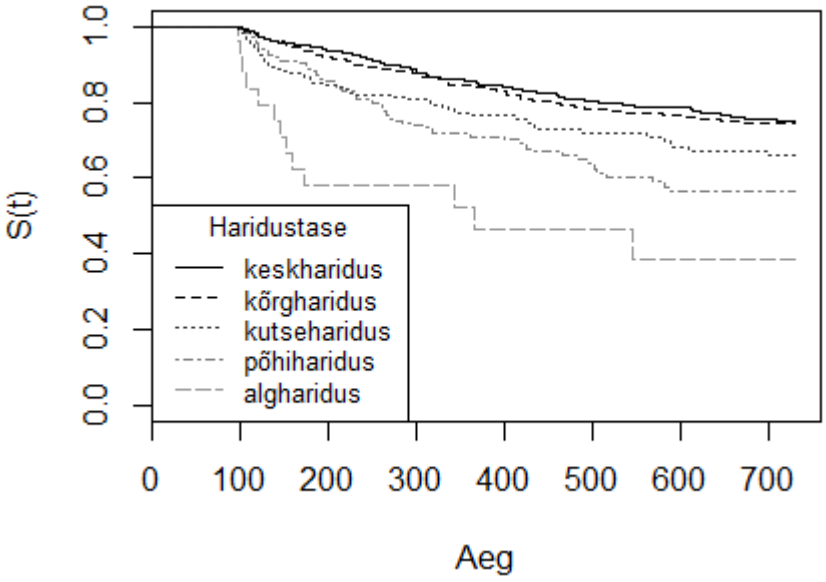



**Joonis 2.4. Erinevatest riikidest pärit klientidele antud 24 kuu pikkuste laenude elulemuskõverad**

Joonisel 2.4 on erinevate elukohariikidega klientidele väljastatud laenude elulemuskõverad. Näeme, et kaheaastase vaatlusperioodi lõpuni on väldanud vaid Eestist pärit klientide laenud. Põhjus on selles, et klientidele, kelle elukohariik on Soome ja Hispaania, anti esimesed laenud 2013. aasta juulis ja oktoobris ning Slovakkiast pärit klientidele 2014. aasta aprillis.

Sellest hoolimata on graafikult näha, et elulemuskõverad on esimese 215 päeva jooksul erinevad. Seda kinnitab ka *log-rank* statistik, millele vastav *p*-väärtus on 0 (vt lisa 2). Eestist pärit klientide hulgas on 215 päeva möödudes maksejõulisi kliente hinnanguliselt 93%, Soomest pärit klientide hulgas 81%, Hispaaniast pärit klientide hulgas 68% ja Slovakkiast pärit klientide hulgas 67% esialgsetest klientidest.

Kui võrdleme elulemuskõveraid kahe riigi kaupa kõikidest riikidest ja kasutame Bonferroni mitmese võrdlemise meetodit (Napierala, 2012), selgub, et olulisusnivool $\alpha = \frac{0.05}{6} = 0.0083$ ei erine omavahel Soomest ja Slovakkiast ning Hispaaniast ja Slovakkiast pärit klientidele antud laenude elulemuskõverad. Vastavad *p*-väärtused on ligikaudu 0.01 ja 0.76 (vt lisa 2).




**Joonis 2.5. Erineva haridustasemega klientidele antud 24 kuu pikkuste laenude elulemuskõverad**

Joonisel 2.5 on kujutatud erineva haridustasemega klientidele väljastatud laenude elulemuskõverad. Näeme, et kõige väiksem on mittepankrotistunud laenude osakaal algharidusega klientide hulgas ning suurim keskharidusega klientide hulgas. Vaatlusperioodi lõpuks on algharidusega klientide hulgas maksejõulisi kliente 39% kõigist sellesse gruppi kuuluvatest klientidest, kellele väljastati 24 kuu pikkune laen. Põhiharidusega klientide hulgas on selleks ajaks maksejõulisi kliente 57%, kutseharidusega klientide hulgas 66%, kõrgharidusega klientide hulgas 74% ja keskharidusega klientide hulgas 75% esialgsetest klientidest.

*Log-rank* testi põhjal võib öelda, et kõverad on statistiliselt oluliselt erinevad: vastav *p*-väärtus on ligikaudu 0 (vt lisa 2).

## 2.4 Coxi võrdeliste riskide mudel

Järgmisena loome 24 kuu pikkuste laenude jaoks Coxi võrdeliste riskide mudeli, mille abil on võimalik prognoosida erinevate kliendigruppide maksujõulisuse tõenäosust erinevatel ajahetkedel. Mudeli põhjal saab otsustada, millised kliendid on usaldusväärsemad ehk kes on vaatlusperioodi lõpus suurema tõenäosusega maksejõulised.

Esiteks lisame mudelisse kõik tunnused, mis võivad mõjutada laenu pankrotistumist. Need on laenuvõtja vanus, sugu, elukohariik, haridustase, töösuhe, tööstaaž, kogusissetulek, vaba raha pärast igakuiste kohustuste täitmist, laenuvõtja Bondora krediidiajalugu ning maksimaalne intressimäär, mida laenutaotlus lubas. Seejärel hakkame ükshaaval mudelist ebaolulisi tunnuseid eemaldama, jättes igal sammul välja kõige suurema *p*-väärtusega tunnus (vt lisa 3).

Lõplikus mudelis on tunnused laenuvõtja vanus, elukohariik, haridustase, kogusissetulek, vaba raha pärast igakuiste kohustuste täitmist ning maksimaalne intressimäär, mida laenutaotlus lubas. Mudel ja tunnused on statistiliselt olulised, vastavad *p*-väärtused on näha lisas 3. Samuti on täidetud võrdeliste riskitiheduste eeldus, mida saab R-is kontrollida käsuga `cox.zph` (vt lisa 3). Lõplik mudel on järgmine:

| | | kordaja | exp (kordaja) | st.viga (kordaja) | z-stat | p-väärtus |
|---|---|---|---|---|---|---|
| Vanus | | -0.020421 | 0.98 | 0.005768 | -3.540 | 4.0e-04 |
| Elukohariik | Hispaania | 2.166613 | 8.73 | 0.173360 | 12.498 | 0.0e+00 |
| Elukohariik | Soome | 1.264015 | 3.54 | 0.264029 | 4.787 | 1.7e-06 |
| Elukohariik | Slovakkia | 2.267203 | 9.65 | 0.376388 | 6.024 | 1.7e-09 |
| Intressimäär | | 0.084457 | 1.09 | 0.012030 | 7.020 | 2.2e-12 |
| Haridus | Alg | -0.371191 | 0.69 | 1.012650 | -0.367 | 7.1e-01 |
| Haridus | Põhi | 0.636176 | 1.89 | 0.167564 | 3.797 | 1.5e-04 |
| Haridus | Kutse | 0.257636 | 1.29 | 0.184449 | 1.397 | 1.6e-01 |
| Haridus | Kõrg | 0.111702 | 1.12 | 0.158631 | 0.704 | 4.8e-01 |
| Kogusissetulek | | -0.000487 | 1.00 | 0.000208 | -2.347 | 1.9e-02 |
| Vaba_raha | | 0.000671 | 1.00 | 0.000274 | 2.450 | 1.4e-02 |

Näeme, et vanusele ja kogusissetulekule vastavate parameetrite hinnangud on negatiivsed. See tähendab, et mida suurem on kliendi vanus ja kogusissetulek, seda väiksem on riskitihedus, kui ülejäänud tunnuste väärtused on samad. Seevastu intressimäärale ja vabale rahale vastavate parameetrite hinnangud on positiivsed, seega mida suuremad väärtused on vastavatel tunnustel, seda suurem on kliendi riskitihedus, arvestades, et muude tunnuste väärtused on samad.

Tunnuse *elukohariik* baastase on Eesti ning ülejäänud tasemed on sellega võrreldes statistiliselt olulised: vastavad *p*-väärtused on väiksemad kui olulisusnivoo $\alpha = 0.05$.

Võrreldes Eestiga on Soomest pärit klientidel $e^{1.26} = 3.53$ korda suurem riskitihedus, Hispaaniast pärit klientidel $e^{2.17} = 8.76$ korda suurem riskitihedus ning Slovakkiast pärit klientidel $e^{2.27} = 9.68$ korda suurem riskitihedus, kui ülejäänud tunnuste väärtused on samad.

Tunnuse *haridus* baastasemeks on valitud keskharidus, sest selles grupis on kõige rohkem kliente. Baastasemega võrreldes on pankrotistumisrisk oluliselt erinev vaid põhiharidusega klientidel, vastav *p*-väärtus on 0.00015. Võrreldes põhi- ja keskharidusega klientide riskitihedusi, saab öelda, et põhiharidusega klientide riskitihedus on $e^{0.64} = 1.90$ korda suurem kui keskharidusega klientidel, kui ülejäänud tunnuste väärtused on samad.

Edasi rakendame mudelit kolme erineva kliendi peal ning kasutame R-i funktsiooni `predictSurvProb`, millega leiame iga kliendi jaoks tõenäosuse, et ta on 24. kuu lõpus maksejõuline. Tõenäosustele anname juurde ka usaldusvahemikud. Vastav R-i kood on lisas 4. Olgu vaatluse all:

1. 35-aastane Eestist pärit keskharidusega klient, kelle laenutaotlus lubab maksimaalselt 30%-st intressimäära. Sissetulek on kliendil 1300 eurot ning vaba raha 200 eurot.
2. 50-aastane Eestist pärit põhiharidusega klient, kelle laenutaotlus lubab maksimaalselt 32%-st intressimäära. Sissetulek on kliendil 1200 eurot ning vaba raha 400 eurot.
3. 40-aastane Hispaaniast pärit kõrgharidusega klient, kelle laenutaotlus lubab maksimaalselt 35%-st intressimäära. Sissetulek on kliendil 1500 euro ning vaba raha 500 eurot.

Esimene klient jääb kogu laenuperioodi jooksul maksujõuliseks tõenäosusega 0.83 (95% usaldusintervall tõenäosusele on 0.78…0.89), teine klient tõenäosusega 0.70 (95% usaldusintervall tõenäosusele on 0.60…0.81) ja kolmas klient tõenäosusega 0.07 (95% usaldusintervall tõenäosusele on 0.02…0.22). Tulemuste põhjal saab öelda, et esimesele ja teisele kliendile on kindlam laenu anda kui kolmandale, sest nende tõenäosused kogu laenuperioodi vältel maksujõuliseks jääda on märksa kõrgemad kui kolmandal kliendil.

## 2.5 Näide elulemusfunktsiooni rakendamisest

Laenudele vastavat elulemusfunktsiooni kasutatakse näiteks ka siis, kui soovitakse leida laenude tootlusi. Tootluse arvutamisel on vaja teada, palju kliente igas kuus tagasimakseid

sooritas, ning seda infot saab elulemusfunktsioonilt. Enne laenude tootluste leidmist tutvume aga mõistega sisemine rentaablus, mille valemit läheb edaspidi tarvis.

### 2.5.1 Sisemine rentaablus

Investeeringu sisemine rentaablus (*internal rate of return*) ehk sisemine tulumäär näitab, kui suurt tulu saab investeeringuobjekti paigutatud rahalt. Laenude kui investeeringute kontekstis tähendab see intressimäära, mille korral on laenusumma võrdne tagasimaksete nüüdisväärtuste summaga. Kui laen kogu ulatuses tagasi makstakse, on sisemine rentaablus intressimäär, millega laen väljastati (Broverman, 2010, lk 126).

Sisemise rentaabluse leidmiseks võrdsustatakse järgnev valem nulliga ning seejärel avaldatakse *r*.

$$NPV = \sum_{n=0}^{N} \frac{CF_n}{(1+r)^n},$$

kus $NPV$ on rahavoogude nüüdispuhasväärtus, $n = 0, 1, \ldots, N$ on perioodi pikkus, mil makseid tehakse, $CF_n$ on rahavoo suurus perioodil $n$ ning $r$ on sisemine rentaablus. (Schmidt, 2015) Võrrandi lahendamiseks $r$-i suhtes kasutatakse numbrilisi meetodeid.

### 2.5.2 Laenude tootluste arvutamine

Esmalt arvutame 24 kuu pikkuste laenude teoreetilise intressimäära, mis realiseeruks siis, kui pankrotistujaid poleks. Selleks leiame kõigepealt väljastatud laenusummade kogusuuruse $CF_0$, mis on väljaminev rahavoog, ning seejärel igakuiste tagasimaksete suurused $CF_n$, $n = 1, 2, \ldots, 24$, mis on sissetulevad rahavood. Asendame leitud suurused $NPV$ valemisse ning võrdsustame selle nulliga. Avaldame võrrandist $r$-i, kasutades selleks R-i funktsiooni `uniroot`. Tulemus on ligikaudu 0.0247 (vt lisa 5).

Kuna makseid sooritatakse kuiselt, on võrrandi lahend 0.0247 kuine intressimäär. Aastane intressimäär ja $m$ korda aastas arvutatav intressimäär on omavahel seotud valemiga

$i = (1 + \frac{i^{(m)}}{m})^m - 1$. Seega meil $i = (1 + 0.0247)^{12} - 1 = 0.3402$, mis tähendabki, et 24 kuu pikkuste laenude teoreetiline intressimäär on ligikaudu 34%.

Leiame nüüd 24 kuu pikkuste laenude tegeliku intressimäära. Väljaminev rahavoog $CF_0$ on sama mis enne. Selleks, et teada saada *n*-nda realiseerunud tagasimakse suurust, peame teadma, palju kliente vastavat makset sooritas. Kuna pankrotistujateks loetakse kliente, kelle kaks järjestikust makset on võlas, tuleb iga teoreetiline makse $CF_n$ läbi korrutada elulemusfunktsiooni väärtusega, mis realiseerus 2 kuud pärast konkreetset kuumakset. Seega, et leida realiseerunud tagasimakse suurust ajahetkel *m*, peame teadma, palju kliente oli alles ajahetkel *m* + 61. Arvutuste tegemiseks kasutame Kaplan-Meieri hinnangut 24-kuuliste laenude elulemuskõverale, mis on joonisel 2.2.

Intressimäära arvutamise protsess on sarnane eelnevalt läbitehtuga, erineb vaid see, et tagasimaksete suurused ei ole konstantsed, vaid sõltuvad elulemusfunktsiooni väärtustest. Tulemuseks saame, et tegelik kuine intressimäär on 0.00654, mis võrdub aastase intressimääraga 0.0814. See tähendab, et 24 kuu pikkuste laenude tegelik tootlus on ligikaudu 8%. Arvutuste tegemiseks kasutatud R-i kood on lisas 5.

Järgmisena vaatame, millised oleksid 24 kuu pikkuste laenude tootlused erinevate intressimäärade korral, mis laenudele määratakse. Selle jaoks fikseerime *NPV* valemis *r*-i ehk soovitud intressimäära ning võrdsustame valemi nulliga, kusjuures väljastatud laenusummade kogusuurus $CF_0$ on sama mis enne. Avaldame võrrandist teoreetilised tagasimaksed $CF_n$, misjärel leiame elulemusfunktsiooni väärtuseid kasutades realiseerunud tagasimaksed (vt lisa 5). Pannes need nüüd *NPV* valemisse ja avaldades seejärel *r*-i, saame teada tootlused. Tulemused viie erineva intressimäära korral on toodud tabelis 2.1.

**Tabel 2.1. Realiseerunud tootlused erinevate kokkulepitud intressimäärade korral**

| intressimäär (%) | tegelik tootlus (%) |
|---|---|
| 20 | -3.050 |
| 25 | 0.947 |
| 30 | 4.937 |
| 35 | 8.944 |
| 40 | 12.932 |

Lisaks arvutame tootlused kahe erineva kliendi puhul:

1. 24-aastane Eestist pärit kutseharidusega klient, kelle sissetulek on 900 eurot ning vaba raha pärast igakuiste kohustuste täitmist 200 eurot;
2. 46-aastane Soomest pärit kõrgharidusega klient, kelle sissetulek on 2200 eurot ning vaba raha pärast igakuiste kohustuste täitmist 600 eurot.

Elulemuskõverad leiame eelnevalt loodud Coxi mudeli abil, kasutades intressimäärasid 20, 25, 30, 35 ja 40 protsenti. Realiseerunud tootlused on tabelis 2.2.

**Tabel 2.2. Realiseerunud tootlused kahe erinevate kliendi puhul**

| | tegelik tootlus (%) | |
|---|---|---|
| intressimäär (%) | klient 1 | klient 2 |
| 20 | 9.713 | 6.297 |
| 25 | 8.832 | 3.663 |
| 30 | 4.935 | -2.69 |
| 35 | -3.027 | -13.703 |
| 40 | -16.022 | -29.894 |

Näeme, et tegelikud tootlused on iga intressimäära korral suuremad Eestist pärit kliendi puhul (klient 1). Põhjus on selles, et temale antud laenu elulemusfunktsiooni hinnangud ehk maksejõulisuse tõenäosused on erinevatel ajahetkedel suuremad kui Soomest pärit kliendile antud laenu puhul. Lisas 6 on klientidele antud 20-, 30- ja 40%-se intressimääraga laenude elulemuskõverad.

Peale selle selgub tabelist 2.2, et intressimäära kasvades tegelik tootlus väheneb. Teame eelnevalt loodud Coxi mudeli põhjal, et mida suurem on intressimäär, seda suurem on riskitihedus ehk pankrotistumise tõenäosus lõpmatult väikeses ajavahemikus $[t, t + \Delta t]$, tingimusel, et klient on ajahetkel *t* maksejõuline. Lisas 6 olevatelt joonistelt on samuti näha, et mida suurem on intressimäär, seda väiksemad on elulemusfunktsiooni hinnangud erinevatel ajahetkedel.

# Kokkuvõte

Käesolevas bakalaureusetöös analüüsiti Bondora laenuvõtjate maksejõuetust. Selleks kasutati elulemusanalüüsi valdkonda kuuluvaid meetodeid, nagu Kaplan-Meieri meetod, *log-rank* test ja Coxi võrdeliste riskide mudel.

Analüüsi käigus saadi teada, et 24 kuu pikkuste laenude puhul pankrotistub 26. kuu lõpuks hinnanguliselt 32% esialgsetest klientidest. Ülejäänud klientidest on enamik selleks ajaks laenu tagasi maksnud ning üksikud pikendavad laenu või pankrotistuvad hiljem.

Lisaks selgus, et 24 kuu pikkuste laenude puhul on erinevatest vanusgruppidest klientidele väljastatud laenude elulemuskõverad statistiliselt oluliselt erinevad: 18–25-aastaste klientide hulgas on pankrotistunud laenude osakaal suurem kui 26–50-aastaste ja üle 50 aastaste klientide hulgas. Laenude elulemuskõverad on statistiliselt oluliselt erinevad ka erineva elukohariigi ja haridustasemega klientide puhul. Nimelt pankrotistuvad kõige vähem eestlased ning kesk- ja kõrgharidusega kliendid.

Analüüsi käigus loodi Coxi võrdeliste riskide mudel, kus on tunnused laenuvõtja vanus, elukohariik, haridustase, kogusissetulek, vaba raha pärast igakuiste kohustuste täitmist ning maksimaalne intressimäär, mida laenutaotlus lubas. Nende abil prognoositi kolme erineva kliendi maksejõulisuse tõenäosusi makseperioodi lõpus.

Analüüsi viimases osas arvutati laenude tootlusi, kasutades sisemise rentaabluse valemit. Selgus, et 24 kuu pikkuste laenude teoreetiline intressimäär on 34%, tegelik tootlus aga 8%, arvestamata seejuures pankrotistunud laenudelt sissenõutavaid summasid. Edasi vaadeldi, milliseid tootluseid on võimalik erinevate kokkulepitud intressimäärade korral saavutada ning viimaks arvutati tootlusi kahe erineva kliendi puhul, kasutades eelnevalt loodud Coxi mudelit.

# Kasutatud kirjandus


Bondora. (2014). *Andmete eksportimine*. Allikas: Bndora by isePankur: https://www.bondora.ee/et/et/datasets. Vaadatud 28.03.2015

Broverman, S. A. (2010). *Mathematics of Investment and Credit, 5th Edition.*

Duerden, M. (2009). *What are hazard ratios?* Allikas: Hayward Medical Communications: http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/what_are_haz_ratios.pdf. Vaadatud 28.03.2015

Klein, J. P., & Moeschberger, M. L. (2003). *Sysrvival Analysis: Techniques for Censored and Truncated Data, Second Edition.*

Kleinbaum, D. G., & Mitchel, K. (2005). *Survival Analysis: A Self-Learning Text, Second Edition.*

Napierala, M. A. (2012). *What Is the Bonferroni Correction?* Allikas: The American Academy of Orthopaedic Surgeons: http://www.aaos.org/news/aaosnow/apr12/research7.asp. Vaadatud 28.03.2015

Ritesh, S., & Mukhopadhyay, K. (2011). *Survival analysis in clinical trials: Basics and must know areas*. Allikas: PubMed Central: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227332/. Vaadatud 15.04.2015

Schmidt, M. (2015). *Internal Rate of Return IRR and Modified IRR Explained: Definition, Meaning, and Example Calculations*. Allikas: Solution Matrix Limited: https://www.business-case-analysis.com/internal-rate-of-return.html. Vaadatud 29.03.2015

Tableman, M., & Kim, J. S. (2005). *Survival Analysis using S: Analysis of Time-to-Event Data.*

Walters, S. J. (2009). *What is a Cox model?* Allikas: Hayward Medical Communications: http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/cox_model.pdf. Vaadatud 28.03.2015

# Lisad

## Lisa 1. Andmed Coxi võrdeliste riskide mudeli näite jaoks

| grupp 0 | | | grupp 1 | | |
|---|---|---|---|---|---|
| jälgimisaeg | pankrot | laenusumma | jälgimisaeg | pankrot | laenusumma |
| 65 | 1 | 785 | 84 | 1 | 800 |
| 67 | 1 | 790 | 89 | 1 | 840 |
| 68 | 1 | 770 | 95 | 1 | 815 |
| 70 | 1 | 865 | 96 | 1 | 825 |
| 72 | 1 | 900 | 97 | 1 | 860 |
| 73 | 1 | 880 | 104 | 1 | 810 |
| 74 | 1 | 880 | 89 | 0 | 665 |
| 75 | 1 | 825 | 94 | 0 | 780 |
| 79 | 1 | 890 | 98 | 0 | 760 |
| 80 | 1 | 895 | 99 | 0 | 785 |
| 81 | 1 | 855 | 102 | 0 | 630 |
| 82 | 1 | 745 | 103 | 0 | 680 |
| 84 | 1 | 830 | 106 | 0 | 650 |
| 86 | 1 | 840 | 109 | 0 | 640 |
| 86 | 1 | 740 | 110 | 0 | 640 |
| 87 | 1 | 810 | 111 | 0 | 740 |
| 88 | 1 | 840 | 114 | 0 | 765 |
| 92 | 1 | 760 | 115 | 0 | 650 |
| 93 | 1 | 740 | 119 | 0 | 790 |
| 95 | 1 | 730 | 120 | 0 | 740 |

## Lisa 2. Kõverate erinevuse tuvastamine R-is

```
>#kõverate erinevus erinevate vanusgruppide puhul
> survtest <- survdiff(Surv(aeg,pankrot)~vanusgrupp)
> 1 - pchisq(survtest$chisq,2) #df=2
[1] 0.001089563 #p-väärtus

>#kõverate erinevus vanusgruppide 26-50 ja 51+ puhul
> survtest2 <- survdiff(Surv(aeg,pankrot)~vanusgrupp12)
> 1 - pchisq(survtest2$chisq,1)
[1] 0.616958 #p-väärtus

>#kõverate erinevus kõikide riikide puhul
> survtest3 <- survdiff(Surv(aeg,pankrot)~riik)
> 1 - pchisq(survtest3$chisq,3)
[1] 0 #p-väärtus

>#kõverate erinevus Slovakkia ja Soome puhul
> survtest4 <- survdiff(Surv(aeg,pankrot)~riik_SK_FI)
> 1 - pchisq(survtest4$chisq,1)
[1] 0.01421117 #p-väärtus

>#kõverate erinevus Slovakkia ja Hispaania puhul
> survtest5 <- survdiff(Surv(aeg,pankrot)~riik_SK_ES)
> 1 - pchisq(survtest5$chisq,1)
[1] 0.7632719 #p-väärtus

>#kõverate erinevus erinevate haridustasemete puhul
> survtest6 <- survdiff(Surv(aeg,pankrot)~haridustase)
> 1 - pchisq(survtest6$chisq,4)
[1] 1.280975e-12 #p-väärtus
```

## Lisa 3. Coxi võrdeliste riskide mudeli loomine

```
>#esialgne mudel
> cox1 <- coxph(Surv(time2,kas_pankrotistus) ~
          Age+factor(Country)+Interest+factor(education_id)+
          factor(employment_status_id)+factor(Gender)+
          factor(NewCreditCustomer)+work_experience+
          laenud5$income_total+FreeCash+
          factor(UseOfLoan),data=laenud5)
> drop1(cox1,test="Chisq")
                            Df    AIC     LRT  Pr(>Chi)
<none>                         3827.6
Age                          1 3833.1   7.518  0.006108 **
factor(Country)              3 3939.0 117.348 < 2.2e-16 ***
Interest                     1 3870.9  45.233  1.75e-11 ***
factor(education_id)         4 3830.1  10.479  0.033090 *
factor(employment_status_id) 4 3821.9   2.260  0.688039
factor(Gender)               2 3824.1   0.501  0.778593
factor(NewCreditCustomer)    1 3825.6   0.004  0.951660
work_experience              5 3827.0   9.337  0.096366 .
laenud5$income_total         1 3831.9   6.283  0.012193 *
FreeCash                     1 3832.6   7.020  0.008062 **
factor(UseOfLoan)            8 3818.7   7.099  0.526025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>#factor(NewCreditCustomer) eemaldamine
                            Df    AIC     LRT  Pr(>Chi)
<none>                         3825.6
Age                          1 3831.1   7.515  0.006120 **
factor(laenud5$Country)      3 3942.2 122.573 < 2.2e-16 ***
Interest                     1 3870.1  46.457 9.365e-12 ***
factor(education_id)         4 3828.1  10.494  0.032881 *
factor(employment_status_id) 4 3819.9   2.277  0.684950
factor(Gender)               2 3822.1   0.502  0.778005
work_experience              5 3825.2   9.601  0.087353 .
income_total                 1 3829.9   6.292  0.012127 *
FreeCash                     1 3830.7   7.023  0.008048 **
factor(UseOfLoan)            8 3816.8   7.214  0.513724
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #factor(Gender) eemaldamine
```

| | Df | AIC | LRT | Pr(>Chi) | |
|---|---|---|---|---|---|
| <none> | | 3822.1 | | | |
| Age | 1 | 3828.5 | 8.351 | 0.003856 | ** |
| factor(Country) | 3 | 3958.9 | 142.746 | < 2.2e-16 | *** |
| Interest | 1 | 3866.3 | 46.135 | 1.104e-11 | *** |
| factor(education_id) | 4 | 3825.1 | 10.925 | 0.027418 | * |
| factor(employment_status_id) | 4 | 3816.6 | 2.467 | 0.650602 | |
| work_experience | 5 | 3821.7 | 9.617 | 0.086849 | . |
| income_total | 1 | 3826.0 | 5.906 | 0.015092 | * |
| FreeCash | 1 | 3827.2 | 7.028 | 0.008026 | ** |
| factor(UseOfLoan) | 8 | 3813.5 | 7.351 | 0.499271 | |

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #factor(employment_status_id) eemaldamine
```

| | Df | AIC | LRT | Pr(>Chi) | |
|---|---|---|---|---|---|
| <none> | | 3816.6 | | | |
| Age | 1 | 3823.3 | 8.736 | 0.00312 | ** |
| factor(Country) | 3 | 3962.7 | 152.141 | < 2.2e-16 | *** |
| Interest | 1 | 3862.8 | 48.188 | 3.872e-12 | *** |
| factor(education_id) | 4 | 3820.1 | 11.500 | 0.02149 | * |
| work_experience | 5 | 3816.9 | 10.262 | 0.06815 | . |
| income_total | 1 | 3820.2 | 5.590 | 0.01807 | * |
| FreeCash | 1 | 3821.0 | 6.412 | 0.01133 | * |
| factor(UseOfLoan) | 8 | 3808.1 | 7.475 | 0.48640 | |

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #factor(UseOfLoan) eemaldamine
```

| | Df | AIC | LRT | Pr(>Chi) | |
|---|---|---|---|---|---|
| <none> | | 3808.1 | | | |
| Age | 1 | 3814.7 | 8.614 | 0.003336 | ** |
| factor(Country) | 3 | 3953.1 | 151.027 | < 2.2e-16 | *** |
| Interest | 1 | 3855.4 | 49.353 | 2.138e-12 | *** |
| factor(education_id) | 4 | 3812.0 | 11.943 | 0.017778 | * |
| work_experience | 5 | 3807.5 | 9.437 | 0.092868 | . |
| income_total | 1 | 3811.2 | 5.158 | 0.023140 | * |
| FreeCash | 1 | 3811.9 | 5.828 | 0.015774 | * |

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #work_experience eemaldamine
```

| | Df | AIC | LRT | Pr(>Chi) | |
|---|---|---|---|---|---|
| <none> | | 3807.5 | | | |
| Age | 1 | 3818.9 | 13.359 | 0.0002572 | *** |
| factor(Country) | 3 | 3947.1 | 145.605 | < 2.2e-16 | *** |
| Interest | 1 | 3857.0 | 51.463 | 7.297e-13 | *** |
| factor(education_id) | 4 | 3813.5 | 14.005 | 0.0072795 | ** |
| income_total | 1 | 3811.5 | 5.997 | 0.0143339 | * |
| FreeCash | 1 | 3811.6 | 6.063 | 0.0138064 | * |

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> #Mudelis on nüüd ainult olulised tunnused.

> summary(cox2) #mudel on oluline

Concordance= 0.755  (se = 0.018 )
Rsquare= 0.107   (max possible= 0.856 )
Likelihood ratio test= 233.5  on 11 df,   p=0
Wald test            = 245.9  on 11 df,   p=0
Score (logrank) test = 300.4  on 11 df,   p=0

> cox.zph(cox2) #riskitiheduste võrdelisuse eeldus täidetud
```

| | rho | chisq | p |
|---|---|---|---|
| Age | -0.01262 | 0.052923 | 0.818 |
| factor(Country)ES | -0.09234 | 2.263737 | 0.132 |
| factor(Country)FI | -0.02268 | 0.157203 | 0.692 |
| factor(Country)SK | -0.03132 | 0.275556 | 0.600 |
| Interest | 0.00719 | 0.013547 | 0.907 |
| factor(education_id)1 | -0.04560 | 0.598388 | 0.439 |
| factor(education_id)2 | -0.03117 | 0.281618 | 0.596 |
| factor(education_id)3 | -0.03380 | 0.335688 | 0.562 |
| factor(education_id)5 | 0.03392 | 0.322330 | 0.570 |
| income_total | 0.00161 | 0.000713 | 0.979 |
| FreeCash | -0.00351 | 0.003448 | 0.953 |
| GLOBAL | NA | 4.756140 | 0.942 |

## Lisa 4. Üleelamistõenäosuste ja usaldusvahemike leidmine

```
> predictSurvProb(cox2, newdata = data.frame(Age=35,
          Country="EE", Interest=30, education_id=4,
          income_total=1300, FreeCash=200),times=730)
[1] 0.8334533
```

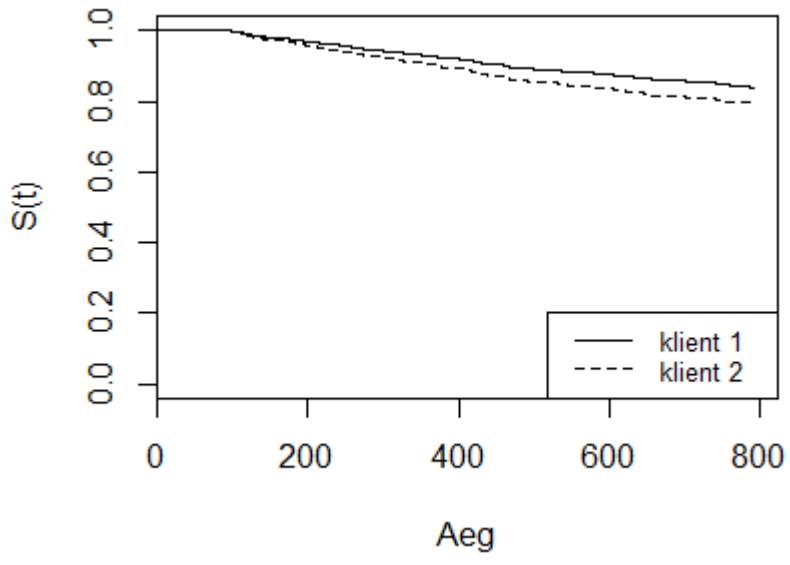
```
> predictSurvProb(cox2, newdata = data.frame(Age=50,
          Country="EE", Interest=32, education_id=2,
          income_total=1200, FreeCash=400),times=730)
[1] 0.697555

> predictSurvProb(cox2, newdata = data.frame(Age=40,
          Country="ES", Interest=35, education_id=5,
          income_total=1500, FreeCash=500),times=730)
[1] 0.06607362

> #usaldusvahemikud tõenäosustele
> klient1 <- survfit(cox2, newdata = data.frame(Age=35,
          Country="EE", Interest=30, education_id=4,
          income_total=1300, FreeCash=200))
> tail(klient1$lower,1) #0.78
> tail(klient1$upper,1) #0.89

> klient2 <- survfit(cox2, newdata = data.frame(Age=50,
          Country="EE", Interest=32, education_id=2,
          income_total=1200, FreeCash=400))
> tail(klient2$lower,1) #0.60
> tail(klient2$upper,1) #0.81

> klient3 <- survfit(cox2, newdata = data.frame(Age=40,
          Country="ES", Interest=35, education_id=5,
          income_total=1500, FreeCash=500))
> tail(klient3$lower,1) #0.02
> tail(klient3$upper,1) #0.22
```

## Lisa 5. Laenude tootluste arvutamine

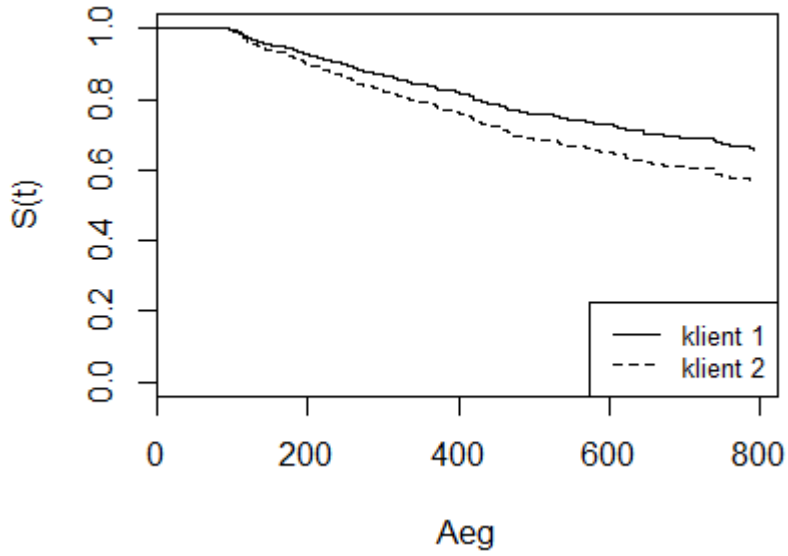
```
>#24 kuu pikkuste laenude teoreetiline intressimäär
> CF0 = -4326176 #kogu laenusumma suurus
> CFn = 241027.4 #igakuised tagasimaksed
> npv <- function(r){
  CF0 + sum(CFn/(1+r)**(1:24)) }
> uniroot(npv,c(0,1)) #c(0,1) on lõik, kust r-i otsitakse
$root
[1] 0.02469495 #r

>#24 kuu pikkuste laenude tegelik intressimäär
>#kõigepealt leitakse realiseerunud tagasimaksete suurused
> tegelikud_kuumaksed <- function(andmestik,kuumaksed){
  tegelik = vector() #esialgu tühi tagasimaksete vektor
  for(i in 1:24){
    tn <- andmestik$tn[andmestik$aeg==floor((i+2)*30.41667)]
    #elulemusfn-i väärtus tagasimakse ajast 2 kuud hiljem
    tegelikud_kuumaksed = tn*kuumaksed
    tegelik <- c(tegelik,tegelikud_kuumaksed) }
  return(tegelik) }

>#NPV funktsioon, mis võrdsustatakse nulliga
> npv <- function(kuumakse,r){
  -4326176 + sum(kuumakse[1:24]/(1+r)**(1:24)) }

>#NPV = 0, leiab r-i
> uniroot(npv,c(0,1),kuumakse=tegelikud_kuumaksed(kmlaenud2,
  241027.4)) #npv-sse pannakse nüüd realiseerunud maksed
$root
[1] 0.006535111 #tegelik kuine intressimäär
```

```
>#r antakse ette, uniroot arvutab kuumakse, mille korral fun=0
> kuumakse <- function(maksed,r){
  sum(maksed/(1+r)**(1:24))-4326176 }

>#funktsioon tootluse arvutamiseks
>#aastane intressimäär ja elulemusf-ni andmestik antakse ette
>#s.o andmestik, kus on pankrotistumise tn-sed igal päeval
> tootlus <- function(andmestik,pr){
  r=(1+0.01*pr)**(1/12)-1 #kuine intressimäär
  teor_kuumakse = uniroot(kuumakse,c(100000,300000),extendInt
  = "yes",r)$root #kuumakse(maksed,r)=0-st leiab kuumakse
  #NPV=0-st leiab r-i
  tegelik_r = uniroot(npv,c(0,1),extendInt = "yes",kuumakse =
  tegelikud_kuumaksed(andmestik,teor_kuumakse))$root
  tegelik_pr=((1+tegelik_r)**12-1)*100
  return(tegelik_pr) }
```

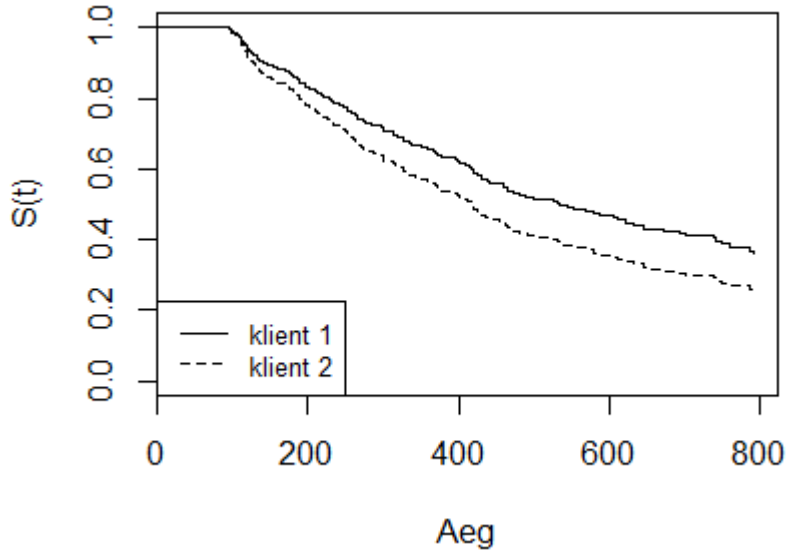
## Lisa 6. Kahe kliendi elulemuskõverad erinevate intressimäärade korral




**Joonis 6.1 Elulemuskõverad 20%-se intressimäära korral**




**Joonis 6.2 Elulemuskõverad 30%-se intressimäära korral**




**Joonis 6.3 Elulemuskõverad 40%-se intressimäära korral**

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina, Carmen Taimre,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Laenuvõtja maksejõuetuse modelleerimine“, mille juhendaja on Märt Möls,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace'is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2 üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **27.04.2015**