Elinaikapuihin perustuvat menetelmät FINRISKI1997-tutkimuksen analysoinnissa

SSL syysseminaari 29.10.2013

Juha Hyssälä

Lähtökohta

- Lääketieteellisessä tutkimuksessa on perinteisesti käytetty elinaika-analyysissä Coxin suhteellisen vaaran mallia ja/tai tämän johdannaisia.
- Kyseinen malli kuitenkin olettaa elinajan ja muuttujan välille suhteellisen yhteyden.
- Tutkittaessa erityisesti erilaisten biomerkkiaineiden yhteyttä elinaikaan, ei suhteellisuusoletus välttämättä toteudu. Näiden jakaumat saattavat olla vinoja ja mahdollinen epälineaarinen yhteys ilmetä vasta jonkin kynnysarvon ylittymisen jälkeen.

Elinaikapuihin perustuvat menetelmät

- Perusideana jakaa aineistoa samanlaisen selviytymistaipumuksen omaaviin osiin.
- Käytännössä tämä tapahtuu määrittämällä jokaiselle muuttujalle optimaalisin jakokohta, jonka avulla mallin tuoma informaatio aineistosta paranee eniten.
- Optimaalisen jakokohdan löytämisen jälkeen vertaillaan jokaista muuttujaa keskenään, ja suurimman lisäinformaation tuova muuttuja päätetään ottaa jakomuuttujaksi.

Elinaikapuihin perustuvat menetelmät

- Havainnot jaetaan tämän jakokriteerin mukaan kahteen tytärsolmuun, jonka jälkeen jakoa jatketaan edelleen jokaisessa uudessa tytärsolmussa aina ennalta määriteltyjen lopetuskriteerien täyttymiseen asti.

Puumallien käyttö

- Puumallit ovat varteenotettava analyysimenetelmä varsinkin suoritettaessa aineiston eksploratiivista analyysiä ja tutkittaessa muuttujien yhteyden luonnetta päätetapahtumaan
- Puumallien avulla saadaan helposti selville aineiston rakenne sekä tärkeimmät elinaikaan vaikuttavat muuttujat

Puumallien käytön edut

- Tarvitaan vähemmän oletuksia kuin Coxin mallissa. Muuttujien vaikutusten ei tarvitse olettaa olevan suhteellisia eikä lineaarisia, jolloin mallinnus on joustavaa.
- Pystytään käsittelemään aineiston rakennetta paremmin, jolloin muutamien poikkeavien havaintojen arvot eivät vaikuta sekoittavasti analyyseihin.
- Tulokset ovat helposti tulkittavissa ja ne pystytään näkemään suoraan erilaisista kuvista.
- Epälineaariset yhteydet tulevat esille helpommin.

Puumallien käytön haitat

- Hypoteesin testausta ei pystytä suorittamaan kuten Coxin mallissa, jolloin ei voida puhua muuttujien tilastollisesta merkitsevyydestä.
- Puumallit ovat aineiston pilkkomisen ja ryhmittelyn vuoksi luonteeltaan epävakaita, jolloin lineaaristen yhteyksien tapauksessa Coxin mallin avulla saadaan luotettavampia tuloksia.

Random Survival Forests

- Puumallien epästabiiliusongelmaan on kuitenkin löydetty ratkaisu koneoppimiseen perustuvista menetelmistä (ensemble-menetelmistä), jotka parantavat tuntuvasti mallien ennustetarkkuutta.
- Kyseisten menetelmien avulla tuodaan satunnaistaminen mukaan yksittäisen puun rakentamiseen ja toistot lopullisen analyysin suorittamiseen, jolloin mallin ennustekyky paranee ja ennustevirhe pienenee.

Algoritmin toimintaperiaate 1

- Vedetään alkuperäisestä aineistosta *B* kappaletta bootstrap-otoksia. Otos muodostetaan valitsemalla kaikkien havaintojen joukosta *N* kappaletta havaintoja käyttäen yksinkertaista satunnaisotantaa takaisinpanolla. Kukin otos on siis alkuperäisen aineiston (*N*) kokoinen otos, joka sisältää keskimäärin 63% alkuperäisistä havainnoista, joista osa sisältyy otokseen useammin kuin kerran. Ne havainnot, jotka eivät sisälly otokseen, muodostavat out-of-bag -datan (OOB-data). Havainnot, jotka puolestaan kuuluvat otokseen, muodostava in-bag -datan.

Algoritmin toimintaperiaate 2

- Kasvatetaan elinaikapuu kullekin *B*–kappaleelle bootstrap-otoksia. Kasvatettaessa puuta, käytetään jokaisessa jakokohdassa ainoastaan p kappaletta satunnaisesti valittuja muuttujia. Muuttujista valitaan parhaan jakokriteerin arvon tuottava muuttuja ja suoritetaan jako tämän mukaan.
- Kasvatetaan puu täyteen kokoonsa siten, että kukin päätesolmu täyttää ehdon, jonka mukaan missään päätesolmussa ei saa olla vähempää kuin d_0 kuolemaa.

Algoritmin toimintaperiaate 3

- Lasketaan kumulatiivisen vaarafunktion arvo kullekin *B* kappaleelle puuta. Koko analyysin kumulatiivisen vaarafunktion arvo saadaan keskiarvoistamalla kunkin puun kumulatiivisen vaarafunktion arvo yli kaikkien puiden.
- Lasketaan kumulatiiviselle vaarafunktiolle ennustevirhe käyttäen OOB-dataa, jolloin ennustevirheen estimaatin arvo on luotettavampi kuin in-bag -datasta laskettaessa.

Jakokriteerin määrittäminen 1

- Jakokritereitä on olemassa lukuisia erilaisia, joista parhaiten sopivaksi graduaineistoon osoittautui log-rank -jakokriteeri. Se määritellään muotoon:

$$L(x,c) = \frac{\sum_{i=1}^{N} \left(d_{i,1} - Y_{i,1} \frac{d_i}{Y_i} \right)}{\sqrt{\sum_{i=1}^{N} \frac{Yi,1}{Y_i} \left(1 - \frac{Yi,1}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}}$$

jossa $d_{i,j}$ on ajanhetkenä $t_{(i)}$ kuolleiden henkilöiden lukumäärä, $Y_{i,j}$ on ajanhetkenä $t_{(i)}$ vaarassa olevien henkilöiden lukumäärä solmussa j=1,2.

Jakokriteerin määrittäminen 2

- Tällöin suureen $|L(x,c)|$ arvo on mittari solmun jaolle. Mitä suuremman arvon $|L(x,c)|$ saa, sitä suurempi selviytymisero kahden tytärsolmun välillä on. Puolestaan mitä suurempi selviytymisero on, sitä parempi jako on kyseessä. Käytännössä paras jako solmussa h voidaan määritellä muuttujan x^* sekä jakoarvon c^* avulla siten, että $|L(x^*,c^*)| \geq |L(x,c)|$ kaikille x:n ja c:n arvoille.

Tulosten tulkinta 1

- Kun jakoa ei enää voida jatkaa missään solmussa, on puuhun muodostunut *h* kappaletta päätesolmuja.
- Kussakin päätesolmussa estimaattina kumulatiiviselle vaarafunktiolle käytetään *Nelson–Aalen* –estimaattoria, joka on muotoa

$$\tilde{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}}.$$

jossa $d_{l,h}$ on kuolemien lukumäärä ja $Y_{l,h}$ vaarassa olevien henkilöiden lukumäärä solmussa *h*

Tulosten tulkinta 2

- Koska kukin henkilö on tasan yhdessä päätesolmussa, voidaan kullekin henkilölle määrittää kumulatiivisen vaarafunktion arvoksi Nelson–Aalen –estimaatin antama arvo
- Päämääränä ei kuitenkaan ole laskea yksittäisen henkilön tai puun kumulatiivisen vaarafunktion arvoa, vaan lasketaan funktion arvo keskiarvoistamalla se yli kaikkien B elinaikapuuta sisältävän analysointikierroksen.

Tulosten tulkinta 3

- OOB-datasta lasketun kumulatiivisen vaarafunktion arvo yli *B*:n puun on tällöin

$$H_e(t|\mathbf{x}_i) = \frac{\sum_{b=1}^{B} I_{i,b} H_b^*(t|\mathbf{x}_i)}{\sum_{b=1}^{B} I_{i,b}}$$

jossa $I_{i,b}=1$ jos *i* on OOB-tapaus tietylle puulle *b* (henkilöä ei ole käytetty puun *b* rakentamiseen) ja $H_b^*(t|x_i)$on edellisen kaavan perusteella määritelty kumulatiivisen vaarafunktion arvo *b*:nnelle kasvatetulle puulle

Tulosten tulkinta 4

- Yksittäisten puiden ja henkilöiden kumulatiiviset vaarafunktiot eivät kuitenkaan ole kovin mielekkäitä tulosmuuttujan mittareita. Tämän vuoksi RSF-malleissa lasketaan muuttujille *mortaliteetteja*. Mortaliteettien laskeminen perustuu conservation-of-events -periaatteeseen, jossa yleisesti voimassa olevien oletusten mukaan kumulatiivisten vaarafunktioiden arvo summattuna yli ajan on yhtä suuri kuin kuolemien lukumäärä kussakin päätesolmussa.

Tulosten tulkinta 5

- Conservation-of-events periaatteen perusteella voidaan määritellä mortaliteetti, joka sanallisesti voidaan kuvailla odotetuksi kumulatiivisen vaarafunktion arvoksi yli ajan , kun ehdollistutaan ainoastaan tietyille kovariaattien arvoille. OOB-dataa käyttäen laskettu ensemble mortaliteetti henkilölle *i* määritellään muotoon

$$M_i = \mathbb{E}_i \left(\sum_{j=1}^{n} H(t_j | \mathbf{x}_i) \right)$$

Mallien vertailu

- RSF-menetelmän tuottamia tuloksia vertaillaan ennustevirheen avulla, jonka laskemiseen käytetään Harrelin C-indeksiä
- Se estimoi todennäköisyyttä, että satunnaisesti valitusta havaintoparista ensiksi kuolevan ennustettu lopputulema on ollut huonompi. C-indeksi on hyvä mittari vertailtaessa tuloksia, sillä se ei riipu yksittäisestä kiinnitetystä ajanhetkestä. Se ottaa huomioon myös havaintojen sensuroinnin, jonka lisäksi se voidaan tulkita virheluokittelutodennäköisyydeksi, joka on luonteva mittari erilaisten mallien vertailussa.

Mallien vertailu

- Mitä pienempi on ennustevirheen arvo, sitä parempia ovat mallin antamat ennusteet. Ennustevirheen arvo 0 tarkoittaisi täydellistä ennustekykyä ja yli 0.5:n olevat arvot sitä, että malli toimisi systemaattisesti huonommin kuin satunnainen arvaaminen.

Muuttujien valinta

- RSF-algoritmin tulosten tulkinta tapahtuu laskemalla muuttujille VIMP-arvoja (Variable Importance), joka käytännössä kuvaa kunkin muuttujan tärkeyttä RSF-mallissa.
- VIMP-arvo muodostetaan pudottamalla jokainen kunkin puun OOB-henkilö puussa alas aina päätesolmuun asti. Jako tehdään aina satunnaisesti siinä kohdassa, jossa jako olisi tehty VIMP-arvon laskemisen kohteena olleen muuttujan mukaan.
- Tämän jälkeen lasketaan koko RSF:n ennustevirhe ja verrataan tätä alkuperäiseen ennustevirheeseen

Bundling-menetelmä

- Kaikessa yksinkertaisuudessaan RSF-menetelmän sekä Coxin mallin yhdistelmä
- Lasketaan ensiksi Coxin mallin avulla lineaaripredиktorin arvo kullekin henkilölle, jonka jälkeen lisätään tämä muuttujaksi RSF-analyysiin
- Suoritetaan aineiston analysointi RSF-mallinnuksen tapaan

FINRISKI1997

- Gradussa käytetty aineisto on peräisin FINRISKI1997-tutkimuksesta, joka on laaja väestötutkimus kroonisten, ei-tarttuvien tautien riskitekijöistä. FINRISKI-tutkimus toteutettiin ensimmäisen kerran vuonna 1972, minkä jälkeen tutkimus on toteutettu aina viiden vuoden välein.

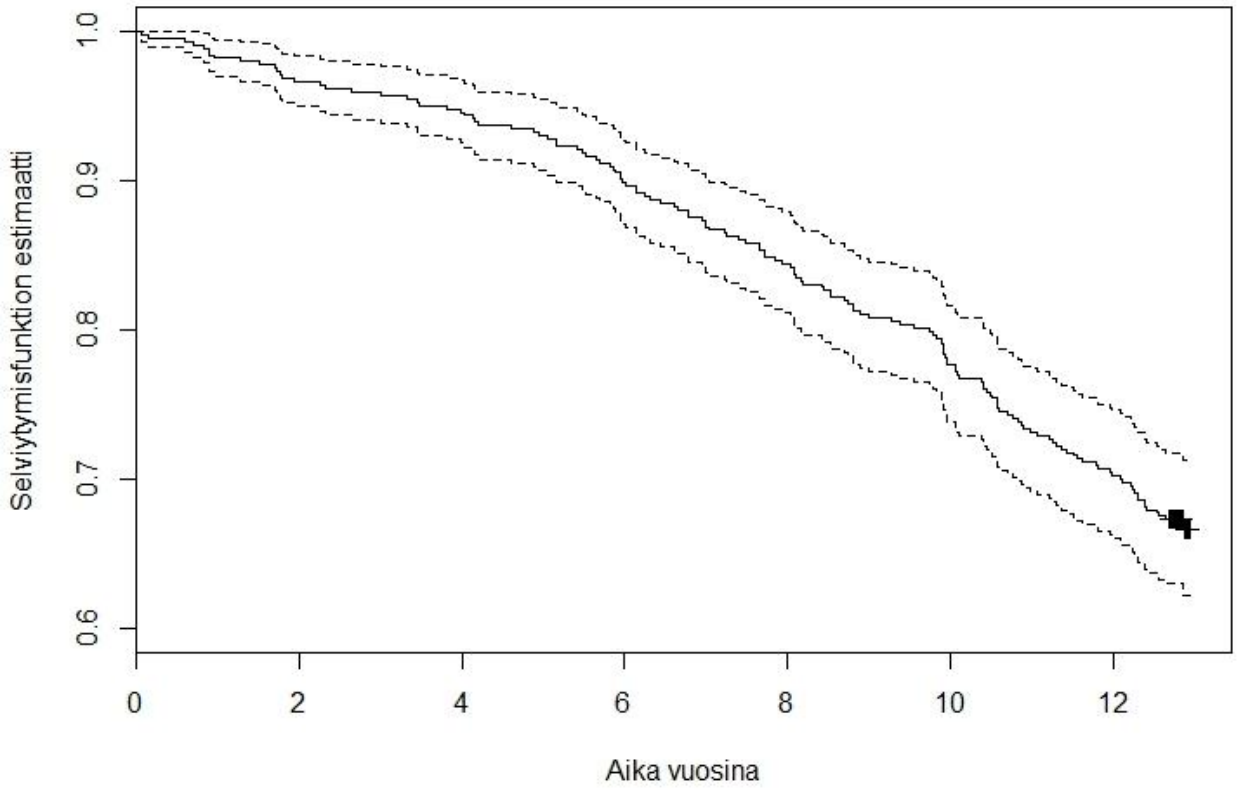
FINRISKI1997

- Yleisesti nimitystä biomerkkiaine (biomarker) käytetään kaikista ihmisen elintoimintoihin vaikuttavista aineista, joiden pitoisuus elimistössä voidaan määritellä.
- Näitä biomerkkiaineita on olemassa useita satoja, mutta nyt keskityttiin tutkimaan ainoastaan erilaisten sytokiinien, erityisemmin kasvutekijöiden (growth factors), verestä mitattavien pitoisuuksien vaikutuksia henkilöiden selviytymiseen.

FINRISKI1997

- Painotuttiin eniten kuolemia sisältävään osa-aineistoon, eli yli 65-vuotiaisiin miehiin
- Mukana analyysissä olivat myös henkilön ikä, painoindeksi, systolinen verenpaine, kokonaiskolesterolipitoisuus, HDL-kolesterolipitoisuus, tupakointistatus sekä alkoholinkulutus.
- Alkoholinkulutuksesta ja verenpaineesta luotiin mielenkiinnon vuoksi kategoriset muuttujat.

Kaplan–Meier –menetelmän sovittaminen aineistoon

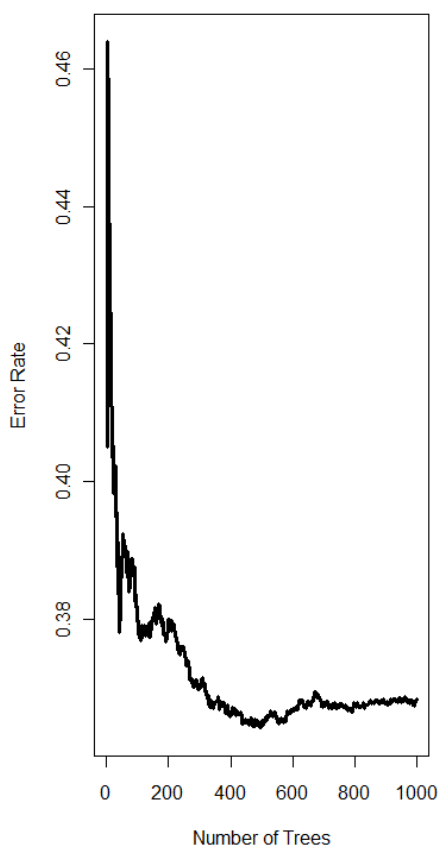


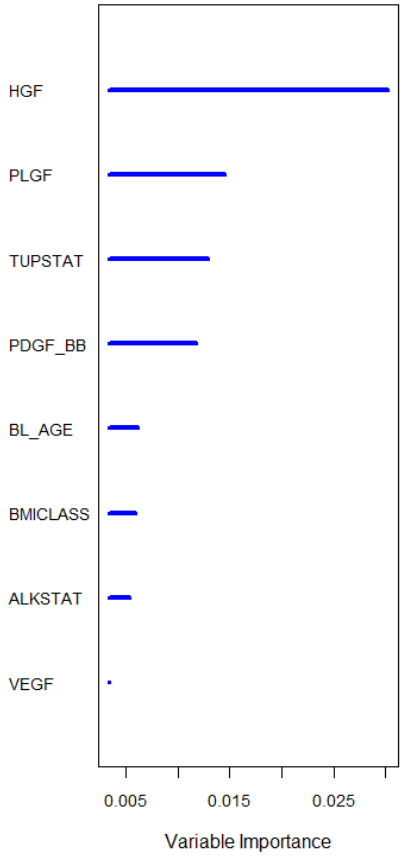
Coxin mallin tuloksia

Taulukko 3: Parhaan Coxin mallin sovittamisen tulokset

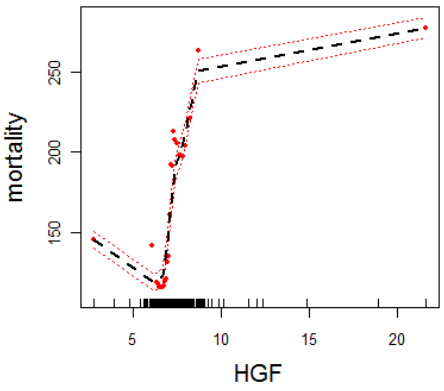
Muuttuja	β	$\exp(\beta)$	$se(\beta)$	Wald	p
Ikä	0.0708	1.073	0.0326	2.17	0.03
Tupakointistatus 1	0.4165	1.517	0.2048	2.03	0.04
Tupakointistatus 2	1.0127	2.753	0.2253	4.49	<0.001
HGF	0.2678	1.307	0.0754	3.55	<0.001
PLGF	0.9631	2.620	0.4487	2.15	0.03
GM-CSF	-0.4873	0.614	0.1786	-2.73	0.006

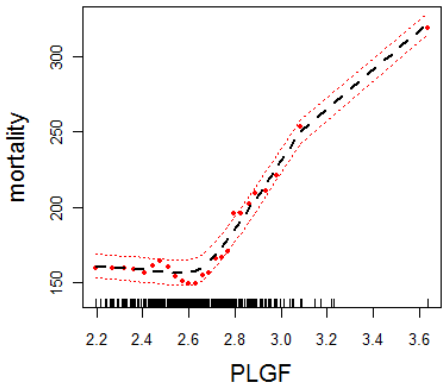
RSF-mallin tuloksia

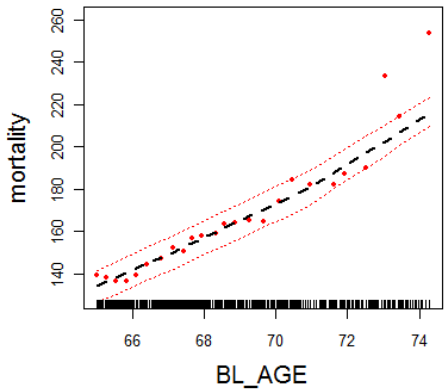


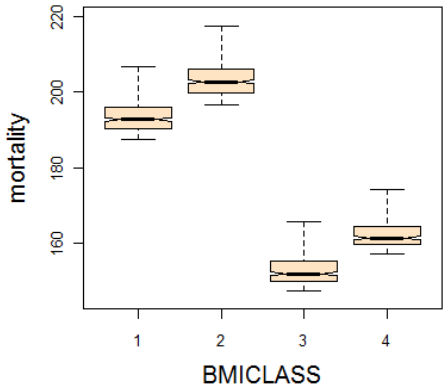


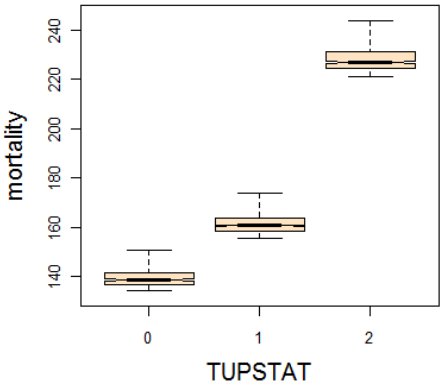
RSF-mallin mortaliteetteja

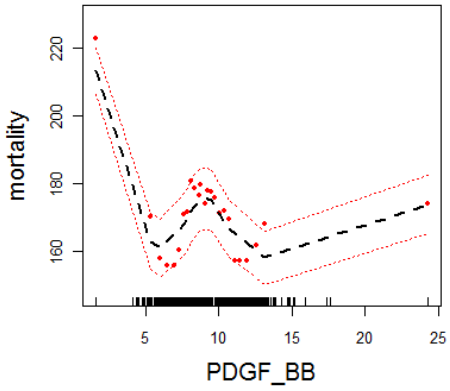


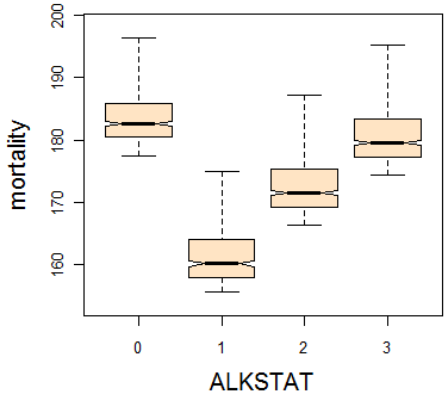


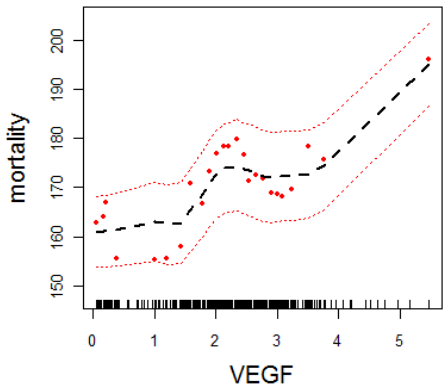












Bundling-mallin tuloksia

Taulukko 8: Lopullisen Bundling-mallin muuttujien VIMP-arvot

Muuttuja	Absoluuttinen VIMP	Suhteellinen VIMP
Lineaariprediktori	0.0388	1.0000
HGF	0.0138	0.3569
PDGF-BB	0.0132	0.3410
BMI-luokka	0.0092	0.2366
VEGF	0.0084	0.2173
Alkoholistatus	0.0077	0.1989
HDL-kolesteroli	0.0070	0.1804
PLGF	0.0055	0.1431

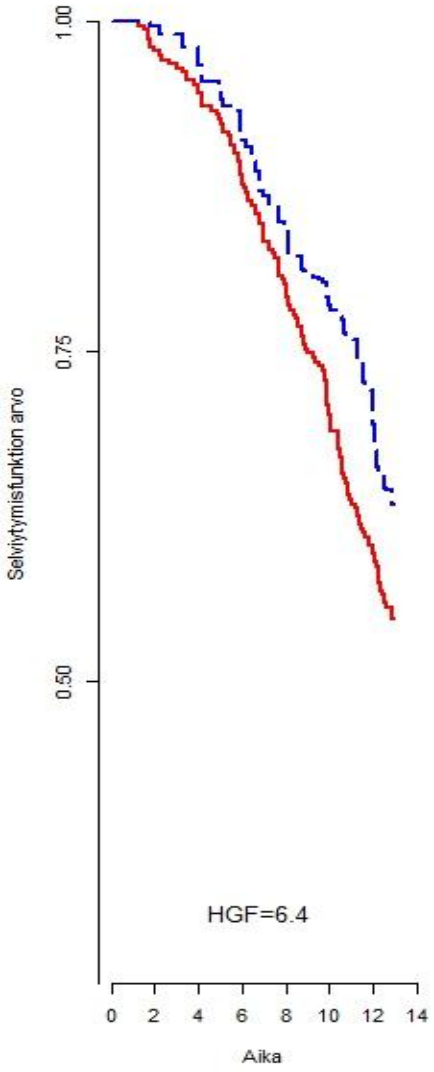
Taulukko 11: RSF-analyysiin ohjautuvien mallien ennustevirheen keskimääräiset arvot laskettuna yli 100 analysointikierroksen

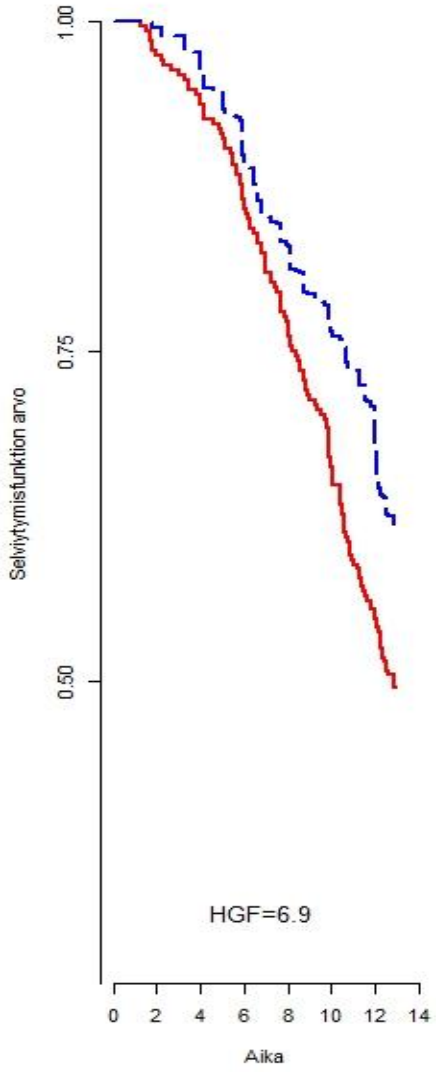
Menetelmä	Ennustevirheen suuruus
Täysi RSF	38.7%
Paras RSF	36.6%
Bundling	35.1%

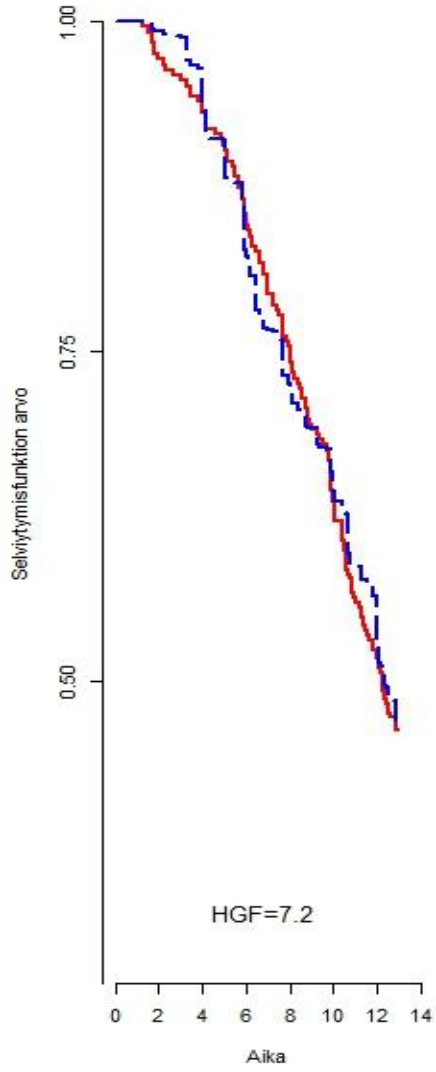
Mallien vertailu

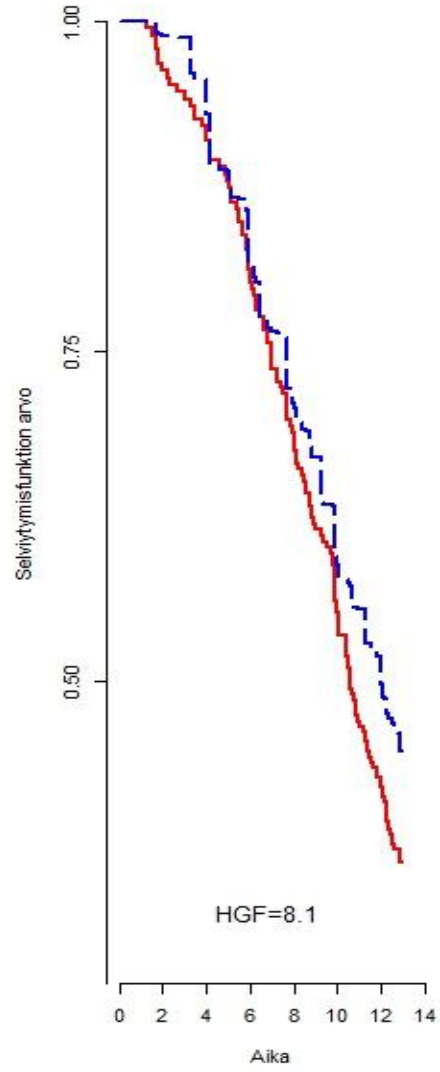
- Kasvutekijällä HGF näytti olevan vahva yhteys kuolemaan, havainnollistetaan analyysimenetelmien eroja HGF:n selviytymisfunktion kuvaajan avulla.
- Lasketaan neljälle kuvitteelliselle henkilölle selviytymisfunktion arvot eri HGF-arvoilla muiden muuttujien arvojen pysyessä vakioina.

Mallien vertailu









Mallien ennustekyvyn vertailu

- Koska eri menetelmiä ei pystytä vertailemaan keskenään suoraviivaisesti, otetaan menetelmien hyvyyden vertailua varten käyttöön integroitu Brier-pistemäärä (IBP).
- Tämän avulla pystytään mittaamaan henkilökohtaisen ennustetun selviytymisfunktion arvon epätarkkuutta verrattaessa tätä havaittuun päätepisteeseen.
- Perustuu quadratic loss -tappiofunktion käyttöön

Mallien ennustekyvyn vertailu

- Määritelmäksi saadaan:

$$\mathrm{BP}(t) = \frac{1}{N}\sum_{i=1}^{N}\Big((\hat{S}(t|\boldsymbol{X}_i))^2 I(t_i \le t, \delta_i = 1)\hat{G}^{-1}(t_i) + (1-\hat{S}(t|\boldsymbol{X}_i))^2 I(t_i > t)\hat{G}^{-1}(t)\Big),$$

- Josta integroiduksi Brier-pistemääräksi saadaan

$$\mathrm{IBP} = \frac{1}{\max(t_i)}\int_0^{\max(t_i)} \mathrm{BP}(t)dt$$

jossa $\hat{S}(t|\mathbf{X}_i)$ on mistä tahansa menetelmästä saatu selviytymisfunktion arvo, $G^{\wedge -1}$ on sensurointijakauman arvo ja δ_i on sensurointi-indeksi

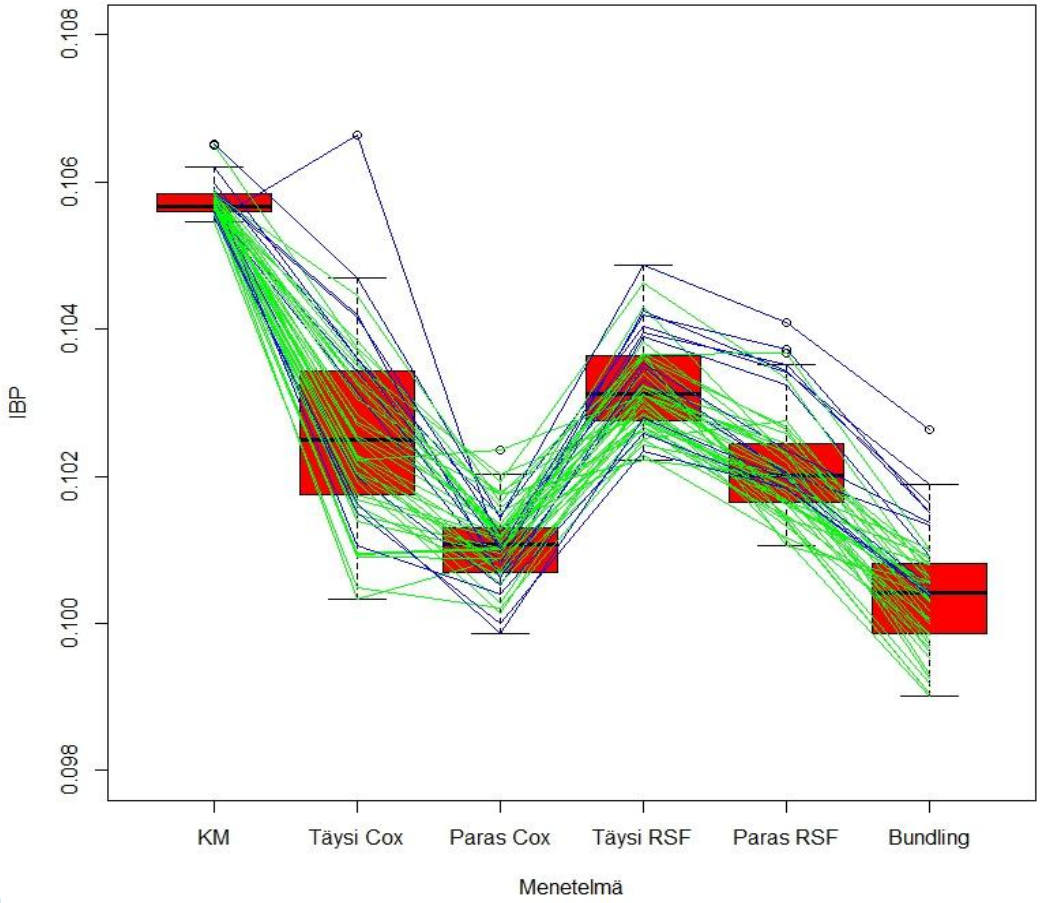
Mallien ennustekyvyn vertailu

- Mitä pienemmän arvon IBP saa, sitä parempi on mallin ennustekyky
- Brier-pistemäärän arvo 0.33 kertoisi mallin olevan yhtä hyvä, kuin jos jokaiselle henkilölle annettaisiin selviytymistodennäköisyys tasaisesta jakaumasta parametreilla 0 ja 1.
- Brier-pistemäärän arvo 0.25 kertoisi mallin olevan yhtä hyvä mallin kanssa, jossa kaikille henkilöille annettaisiin selviytymistodennäköisyys 0.5.

Mallien vertailu

- Suoritettiin 10-kertaisella ristiinvalidoinnilla kullekin menetelmälle
- Suoritettiin 50 kertaa
- Jokaisella kierroksella muodostettiin yksittäiset Coxin mallit kullekin ristiinvalidointiotokselle sekä 1000 puuta sisältävät RSF- ja Bundling-mallit.

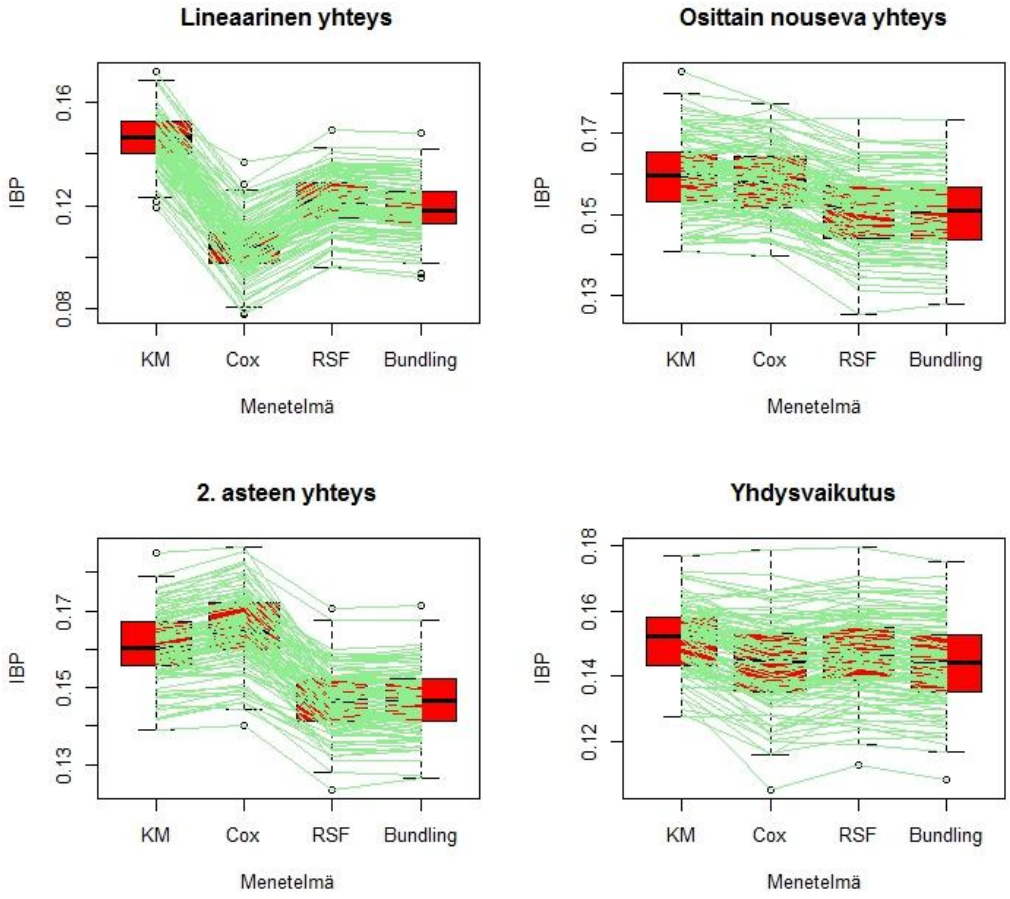
Mallien vertailu, tulokset



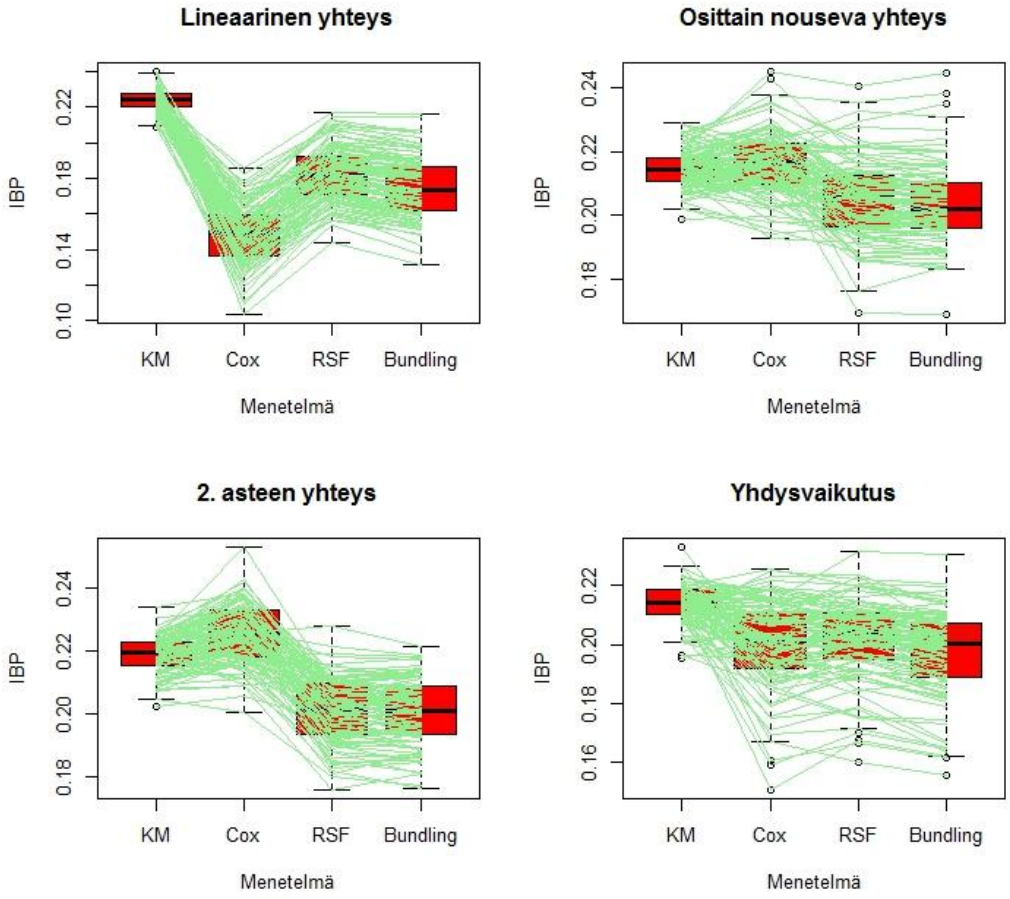
Menetelmien vertailu simuloiduissa aineistoissa

- Muodostettiin aineisto, johon luotiin neljä erilaista yhteyttä muuttujien ja elinajan välille.
  - Lineaarinen yhteys
  - Osittain lineaarinen yhteys
  - Toisen asteen yhteys
  - Kahden muuttujan yhdysvaikutus
  - Kaikki yhdessä
- Lisäksi malleihin sisällytettiin kolme kohinamuuttujaa normaalijakaumasta.
- Simuloinnit suoritettiin myös kahdella eri sensurointi-osuudella (10% ja 60%) luotettavien lopputulosten aikaansaamiseksi

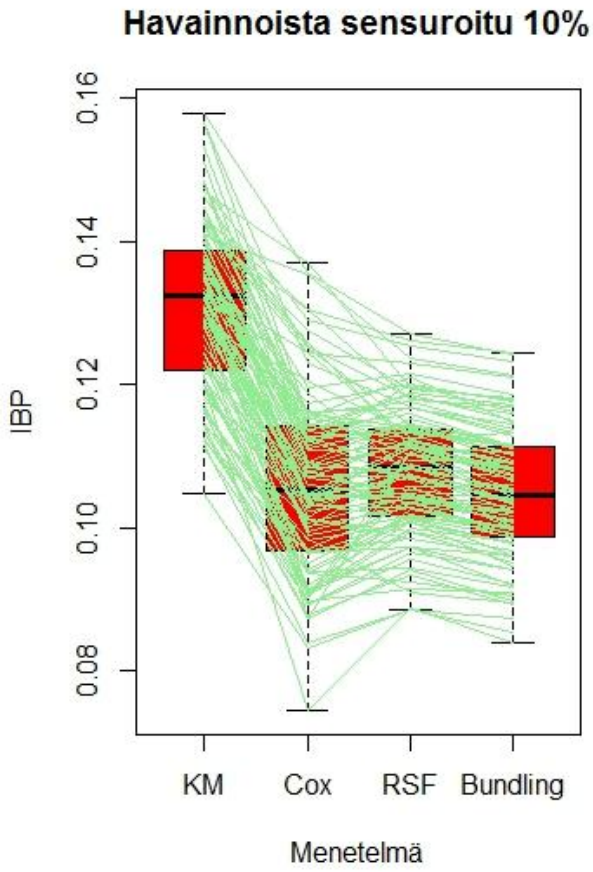
Simuloinnin tuloksia, kun 10% havainnoista on sensuroitu

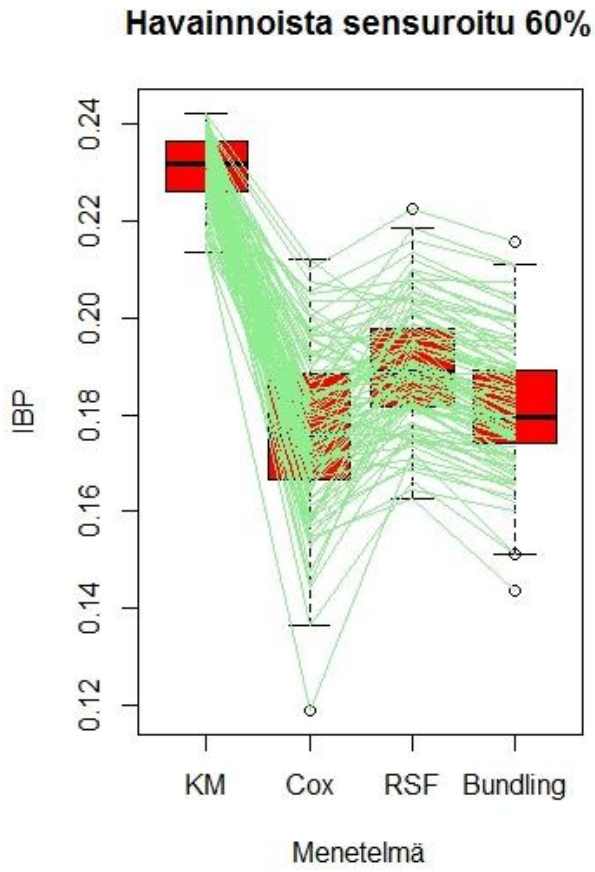


Simuloinnin tuloksia, kun 60% havainnoista on sensuroitu

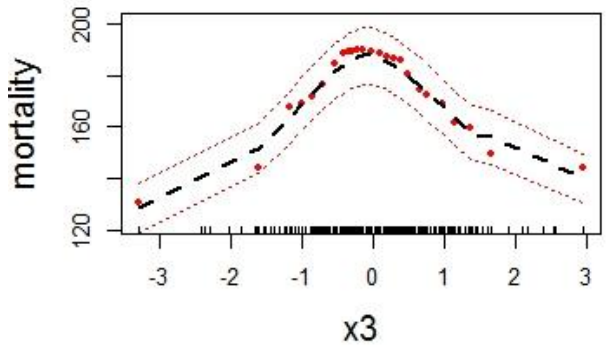


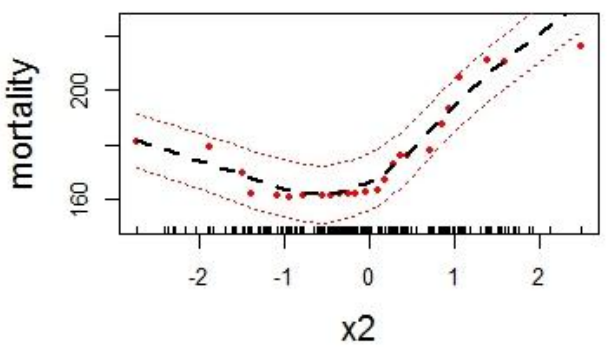
Simuloinnin tuloksia, kun mukana kaikki yhteydet samalla kertaa

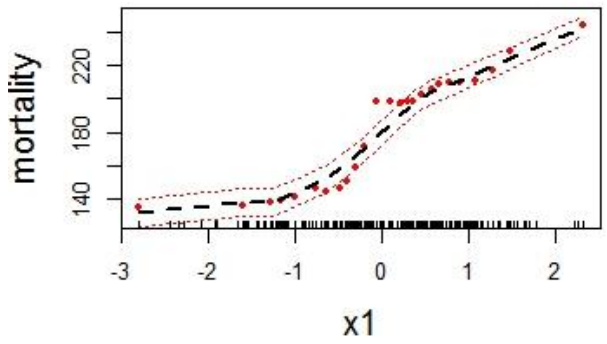


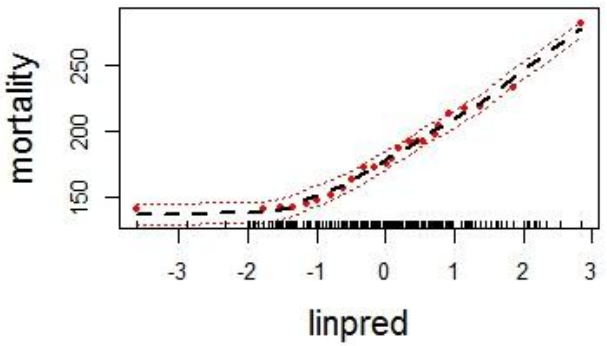


Bundling-mallin mortaliteetteja









Loppupäätelmät

- Bundling-malli toimi aineistossa parhaiten
- Epälineaarisuudet tulivat selvästi esille RSF-pohjaisissa menetelmissä
- Coxin malli toimi selvästi parhaiten lineaaristen yhteyksien ollessa kyseessä
- RSF-pohjaiset menetelmät ovat loistava työkalu tutkimaan epälineaarisia yhteyksiä
- Jos aikoo opetella käyttämään kunnolla R:ää, kannattaa varata paljon aikaa ja hyvät hermot.