UiA Fakultet for teknologi og realfag

MA-223-G 24V Statistikk - Prosjekt

HSCT-behandling av pediatriske pasienter: regresjonsanalyse av variabler

Philip Haugen

Kandidatnummer: 720

Inspera gruppenummer: 2308

Vår 2024

Innhold

1 Introduksjon

Dette prosjektet tar sikte på å utforske autolog hematopoietisk stamcelletransplantasjon (HSCT) gjennom anvendelse av statistiske metoder. HSCT er en livreddende prosedyre som brukes i behandlingen av ulike hematologiske lidelser, når mer skånsomme behandlinger har begrenset effekt eller gir store bivirkninger[1]. Datasettet som brukes i dette prosjektet inneholder informasjon om pediatriske pasienter som gjennomgikk HSCT, inkludert alder, vekt, blodtype til giver og mottaker, sykdomstyper, behandlingsrelaterte faktorer og post-transplantasjonsresultater. Ved å analysere dette datasettet, er målet å finne prediktorer for HSCT-variabler, og variabelene som relaterer til utfallet ved bruk av lineær og logistisk regresjon, samt trebasert regresjon.

Det er viktig å understreke at hensikten med dette prosjektet er åpenbart ikke medisinsk forskning, og ikke gir medisinske råd eller anbefalinger, men snarere en statistisk analyse av åpen, tilgjengelig data drevet av en personlig interesse for medisinsk statistikk.

2 Data, materialer og metoder

2.1 Data

Dataen er hentet fra UCI Machine Learning Repository[2], som har fått den donert fra skaperne Marek Sikora og Lukasz Wróbel fra henholdsvis Silesian University of Technology og Institute of Innovative Technologies EMAG. I følge studien av Sikora, Wróbel, & Gudyś[3] så beskriver dataen 187 pediatriske pasienter (75 kvinnelige og 112 mannlige) med flere hematologiske sykdommer: 155 maligne sykdommer (bl.a. 67 pasienter med akutt lymfoblastisk leukemi, 33 pasienter med leukemi, 33 med akutt myelogen leukemi, 25 med kronisk myelogen leukemi, 18 med myelodysplasi kronisk myelogen leukemi, 18 med myelodysplastisk syndrom) og 32 ikke-maligne tilfeller (blant annet 13 pasienter med alvorlig aplastisk anemi, 5 med Fanconi-anemi, 4 med X-bundet adrenoleukodystrofi, 4 med adrenoleukodystrofi). Alle pasientene ble utsatt for hematopoietisk stamcelletransplantasjon fra en ubeslektet donor.

Det er usikkert om hele mengden av data ble samlet inn ved hjelp av bekvemmelighetsutvalg, som innebærer det at utvalget kanskje ikke er representativt for den hele populasjonen[4, s.34]. Dette kan introdusere utvalgsbias og begrense generaliserbarheten av eventuelle konklusjoner trukket fra analysen. Videre, hvis det er slikt at de forklarende variablene ikke ble randomisert før registrering av responsvariablene, blir det utfordrende å etablere årsakssammenhenger mellom variablene. Dette skyldes at det kan være forstyrrende variabler eller ubemerkede faktorer som påvirker både de forklarende variablene og responsvariablene. Uten randomisering er det vanskelig å utelukke alternative forklaringer på eventuelle observerte sammenhenger.

2.2 Materialer og metoder

I dette prosjektet ble programvaren R benyttet som et verktøy for analyse og utforskning av data. R er et åpen-kildekode programvaresystem spesielt utviklet for statistisk analyse. Her ble dataene først importert inn i R, deretter ble de rensket for nullverdier og utforsket for å forstå egenskapene. Deretter ble nevnte statistiske metoder anvendt for å undersøke sammenhenger mellom variabler og identifisere mønstre. R ble brukt for brorparten av alle utregninger ved bruk av innebygde regnefunksjoner. Flere pakker ble benyttet for å utføre disse ulike analysene og utregningene. Alle disse finner man i kodesnuttene under vedlegget.

To store primærkilder utenom pensumboken som ble brukt for informasjon om de forskjellige statistiske metodene, samt det å finne, formulere, og utlede de nødvendige matematiske ligningene var først boken "Medical Statistics at a Glance, 2nd Edition" av Aviva Petrie og Caroline Sabin. Den andre boken var "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" av Trevor Hastie, Robert Tibshirani, og Jerome Friedman. Disse blir ofte henvist til i dette prosjektet da de var rett og slett øyeåpnere; svært nyttige for få en mye bredere forståelse av bayesianske og frekventistiske teknikker, og statistikken som blir jobbet med.

2.3 Variabler

Disse tabellene beskriver alle variablene i dataen. Ikke alle kommer til å bli brukt i analysen, men det gir en klinisk oversikt.

VARIABEL	BESKRIVELSE	TYPE
Recipientgender	Pasientens biologiske kjønn, 1 for mannlig og 0 for kvinnelig	Nominal
Stemcellsource	Kilde til hematopoietiske stamceller (Perifert blod - 1, Beinmarg - 0)	Nominal
Donorage	Alder av giveren på tidspunktet for aferese av hematopoietiske stamceller	Kontinuerlig
Donorage35	Alderen til giveren < 35 - 0, > = 35 - 1	Nominal
IIIV	Utvikling av akutt graft versus host sykdom stadium II eller III eller IV (Ja - 1, Nei - 0)	Nominal
Gendermatch	Kompatibilitet mellom donor og mottaker i henhold til deres kjønn (Kvinne til Mann - 1, Annet - 0)	Nominal
DonorABO	ABO-blodgruppe av giveren av hematopoietiske stamceller (0 - 0, 1, A, B = -1, AB = 2)	Ordinal
RecipientABO	ABO-blodgruppe av mottakeren av hematopoietiske stamceller (0 - 0, 1, A, B = -1, AB = 2)	Ordinal
RecipientRh	Tilstedeværelse av Rh-faktoren på mottakerens røde blodceller ('+' - 1, '-' - 0)	Nominal
ABOmatch	Kompatibilitet mellom donor og mottaker av hematopoietiske stamceller i henhold til ABO-blodgruppe (tilpasset - 1, ikke tilpasset - 0)	Nominal
CMVstatus	Serologisk kompatibilitet mellom giver og mottaker av hematopoietiske stamceller i henhold til cytomegalovirus	Nominal
DonorCMV	Tilstedeværelse av cytomegalovirusinfeksjon hos giver av hematopoietiske stamceller før transplantasjon (tilstedeværelse - 1, fraværende - 0)	Nominal
RecipientCMV	Tilstedeværelse av cytomegalovirusinfeksjon hos mottaker av hematopoietiske stamceller før transplantasjon (tilstedeværelse - 1, fraværende - 0)	Nominal
Disease	Type sykdom (ALL, AML, kronisk, ikke-malign, lymfom)	Nominal
Riskgroup	Risikogruppe, høy risiko - 1, lav risiko - 0	Nominal
Txpostrelapse	Den andre benmargstransplantasjonen etter tilbakefall (Nei - 0; Ja - 1)	Nominal
Diseasegroup	Type sykdom (malign - 1, ikke-malign - 0)	Nominal
HLAmatch	Kompatibilitet av antigener av hovedhistokompatibilitetskomplekset til giveren og mottakeren av hematopoietiske stamceller i henhold til ALL internasjonale BFM SCT 2008-kriterier (10/10 - 0, 9/10 - 1, 8/10 - 2, 7/10 - 3 (allel/antigener)	Ordinal
HLAmismatch	HLA tilpasset - 0, HL uoverensstemmende - 1	Nominal
Antigen	I hvor mange antigener er det forskjell mellom giveren og mottakeren (-1 - ingen forskjeller, 0 - én forskjell, 1 (2) - to (tre) forskjeller)	Ordinal
Allele	Hvor mange alleler er det forskjell mellom giveren og mottakeren (-1 ingen forskjeller, 0 - én forskjell, 1 (2) (3) - to, (tre, fire) forskjeller)	Ordinal
HLAgrI	Forskjellstypen mellom giveren og mottakeren (HLA-matched - 0, forskjellen er bare i ett antigen - 1, forskjellen er bare i ett allel - 2, forskjellen er bare i DRB1-celle - 3, to forskjeller (to alleler eller to antigener) - 4, to forskjeller (to alleler eller to antigener) - 5)	Nominal
Recipientage	Alder av mottakeren av hematopoietiske stamceller på tidspunktet for transplantasjonen	Kontinuerlig
Recipientage10	Mottaker alder < 10 - 0, Mottaker alder >= 10 - 1	Nominal
Recipientageint	Mottaker alder i (0,5] - 0, (5, 10] - 1, (10, 20] - 2	Ordinal

VARIABEL	BESKRIVELSE	TYPE
Relapse	Gjentakelse av sykdommen (Nei - 0, Ja - 1)	Nominal
aGvHDIIIIV	Utvikling av akutt graft versus host sykdom stadium III eller IV (Ja - 0, Nei - 1)	Nominal
extcGvHD	Utvikling av omfattende kronisk graft-versus-host sykdom (Ja - 0, Nei - 1)	Nominal
CD34kgx10d6	CD34+ celle dose per kg av mottakerens kroppsvekt, 10^6/kg	Diskret
CD3dCD34	CD3+ celle til CD34+ celle-forholdet.	Kontinuerlig
CD3dkgx10d8	CD3+ celle dose per kg av mottakerens kroppsvekt, 10^8/kg	Diskret
Rbodymass	Kroppsmasse til mottakeren av hematopoietiske stamceller på tidspunktet for transplantasjonen	Diskret
ANCrecovery	Tid til gjenoppretting av nøytrofiler definert som nøytrofiltelling >0.5 x 10^9/L	Diskret
PLTrecovery	Tid til platelettoppbygging definert som platelettelling >50000/mm3	Diskret
time_to_aGvHD_III_IV	Tid til utvikling av akutt graft-versus-host-sykdom stadium III eller IV	Diskret
survival_time	Tidspunkt for observasjon (hvis levende) eller tid til hendelse (hvis død) i dager	Kontinuerlig
survival_status	Overlevelsesstatus (0 - levende, 1 - død)	Nominal

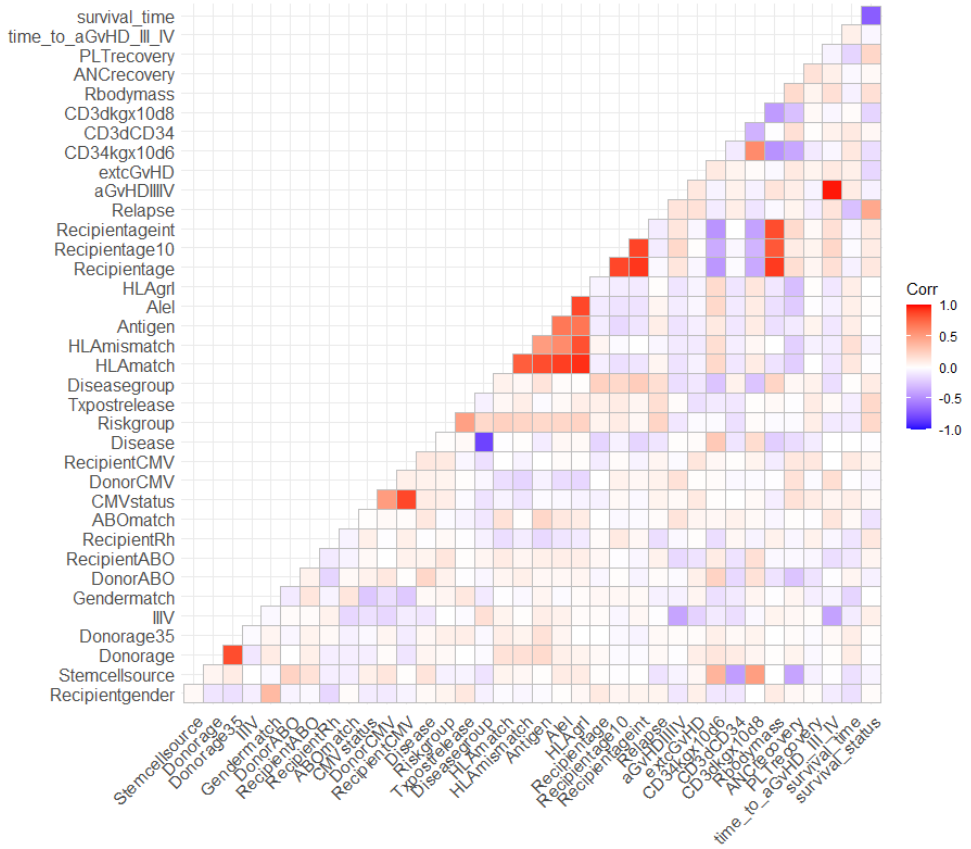
3 Analyse av data

3.1 Trening og testing

Målet er å lage en modell som er i stand til å forutsi en responsvariabel, gitt et sett med prediktorvariabler. For å teste generaliseringsevnen til modellen, holdes en del av den tilgjengelige dataen ut som en testmengde. Denne mengden er nyttig for å evaluere prediksjonsevnen til en modell på usett data. For å lage en testmengde, velges det 20% tilfeldig data fra datasettet. De resterende 80% regnes som treningsmengden, dvs. dataene som brukes til å bygge og validere modellene.

3.2 Utforskende dataanalyse

Her blir det gjort en enkel utforskende dataanalyse av noen variabler i datasettet. Først så lages det en korrelasjonsmatrise som representer et "varmekart", basert på Pearsons korrelasjonskoeffisient[4, s.67]. Korrelasjonskoeffisienten r varierer mellom -1 og 1. Den måler styrken av den lineære sammenhengen mellom X og Y. Hvis X og Y er lineært positivt korrelerte, er $r_{XY} > 0$, slik at verdier av X større enn gjennomsnittet er assosiert med Y-verdier også større enn gjennomsnittet. Omvendt, hvis X og Y er lineært negativt korrelerte, er $r_{XY} < 0$. En korrelasjonskoeffisient nær 0 antyder at det ikke er noen lineær sammenheng mellom X og Y.



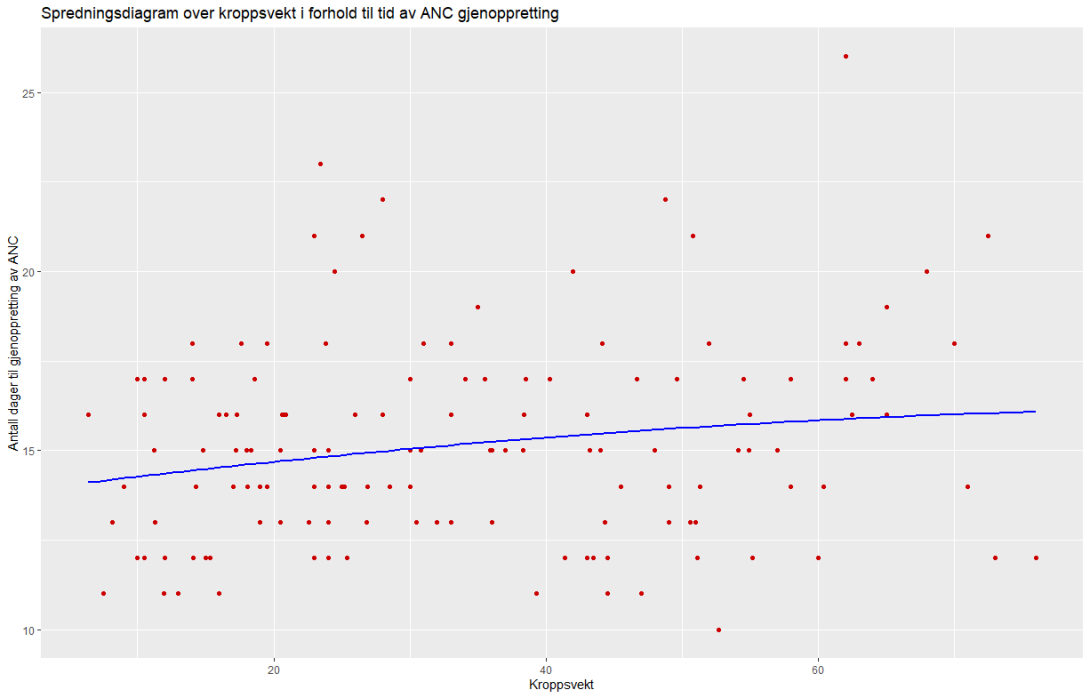
Figur 1: Korrelasjonsmatrise for hele datasettet

Denne matrisen tar i bruk alle variabler, men her skal det fokuseres på variabelen ANCrecovery. Ut fra figuren kan man se at det som korrelerer er for eksempel alderen til mottakeren, kroppsvekten, samt tilstedeværelse av cytomegalovirusinfeksjon hos giveren. For å illustrere dette nærmere:

ANCrecovery	Rbodymass	Recipientageint	Recipientage	CD3dCD34	PLTrecovery	DonorCMV
1.00000000	0.19010278	0.18990683	0.17152403	0.15813923	0.15071566	0.14919204
CMVstatus	RecipientCMV	extcGvHD	Recipientage10	aGvHDIIIIV	time_to_aGvHD_III_IV	Donorage
0.14708748	0.12029755	0.10990636	0.09850270	0.09234387	0.07939697	0.06412109

Figur 2: Variabler som korrelerer med ANCrecovery i synkende rekkefølge

Hvis man for eksempel lager en graf som sammenligner ANCrecovery med en korrelert variabel, her blir det kroppsvekt (Rbodymass), så vil man se at linjen følger en gradvis tilpasning da variablene følger hverandre tett:



Figur 3: Spredningsdiagram med kroppsvekt i forhold til tid for gjenoppretting av ANC

Ut fra denne enkle analysen så korrelerer visse variabler med ANCrecovery, men er det kausalitet? Dette blir tatt opp videre i prosjektet.

4 Lineær regresjon

Det første forsøket på å forutsi variabelen for ANCrecovery basert på de tilgjengelige variablene er ved å lage en modell for enkel lineær regresjon, og detter bruke multivariat regresjon, som en del av den generelle lineære modellen i underseksjon 3 "Generell lineær modell".

4.1 Teoretisk ramme

4.1.1 Enkel lineær regresjon

For å utforske forholdet mellom to diskrete variabler, x og y, måles verdiene for disse variablene for hver av de n individene i utvalget. Den enkle lineære regresjonslinjen estimeres ved hjelp av ligningen[4, s.70]:

$$Y = a + bx$$

Hvor i følge kilden:

- x er den uavhengige variabelen, kjent som prediktor eller forklaringsvariabel.

- For hver verdi av x, representerer Y den tilhørende verdien av den avhengige variabelen (responsvariabel). Denne verdien ligger på den estimerte regresjonslinjen.
- a er skjæringspunktet til regresjonslinjen med y-aksen. Denne verdien indikerer hvor regresjonslinjen krysser y-aksen når x er null.
- b er stigningen i regresjonslinjen. Denne verdien viser endringen i gjennomsnittlig Y for hver økning i x.

4.1.2 Generell lineær modell

Først undersøkes det en matrise X, der x_{ij} er den j-te prediktoren for den i-te prøven[5, s.10-12]. Matrisen har dimensjonene $n \times p$, der n er antall prøver og p er antall prediktorer. Målet er å forklare den sannsynlige responsen til utfallene y, som observasjoner av de tilfeldige variablene Y.

Forholdet mellom Y og X har denne formelen[5, (3.23), s.52]:

$$Y = X\beta + \epsilon$$

Hvor β er vektoren med ukjente parametere; parametere som er målet for statistisk inferens, og ϵ er en vektor med uavhengige og identisk fordelt tilfeldige variabler som representerer feilene (for eksempel naturlig variabilitet eller målefeil) addert med $X\beta$. Selv om det ikke er mulig å observere feilene, kan det gjøres noen antakelser[4, s.70]:

- Feilene er sentrert rundt 0.
- Feilene har samme varians (homoskedastisitet).
- Feilene er ukorrelerte.
- Feilene er normalfordelte.

Feilene er fordelt som en multivariat normalfordeling $\epsilon \sim \mathcal{N}(0, \sigma^2)$[5, s.47]. Dette antyder at målevektoren også er fordelt som en multivariat normalfordeling, der hver tilfeldig variabel Y_i har samme varians.

4.1.3 Sannsynlighetsmaksimering

Det er ikke mulig å nøyaktig vite de sanne verdiene av parameterne β, men ved å bruke et datasett trukket fra utvalget så kan disse estimeres. En vanlig tilnærming er å bruke sannsynlighetsmaksimeringsestimatoren for β_i ved å derivere rimelighetsfunksjonen $\mathcal{L}$ og bruke disse estimatene til å beregne $\hat{\beta}_i$.

Målet er å maksimere rimelighetsfunksjonen $\mathcal{L}$ ved å holde X konstant for å finne parameterne som gir maksimale verdier til observasjonene $y = (y_1, \ldots, y_n)'$. Rimelighetsfunksjonen er utledet og har følgende formel[5, (8.20), s.267]:

$$\mathcal{L}(\beta \mid y) \propto e^{-\frac{1}{2\sigma^2}\|y - X\beta\|^2}$$

Denne formelen representerer sannsynlighetsfordelingen til parametervektoren β gitt observasjonene $y_1, \ldots, y_n$. $\propto$ vil si at formelen er proporsjonal, og det til høyre er kjernen i sannsynlighetsfordelingen, og det er en eksponentialfunksjon som er avhengig av den kvadrerte euklidske avstanden mellom observasjonene y og forventede verdier $X\beta$. Her er y en vektor av observasjoner, og som tidligere nevnt er X matrisen som inneholder prediktorvariablene, og β er vektoren av parametere som skal estimeres. Dette er for å maksimere $\mathcal{L}$, som tilsvarer å minimere $\|y - X\beta\|^2$, som er et minste kvadraters metode (OLS) minimeringsproblem[4, s.70]. Her representerer $\|\cdot\|$ et euklidsk rom[5, (4.48), s.132]:

$$\min_{\beta} \|y - X\beta\|^2$$

Hvis $X'X$ er invertibel, er sannsynlighetsmaksimeringsestimatoren for β gitt ved[5, s.12]:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Samplingsfordelingen av $\hat{\beta}$ er multivariat normal, siden den er en lineær transformasjon av den avhengige variabelen y. Dette kommer fra sentralgrenseteoremet, som sier at summen av uavhengige og identisk fordelte tilfeldige variabler vil tilnærme seg en normalfordeling når utvalgsstørrelsen øker[4, s.26].

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(X'X)^{-1}\right)$$

Denne formelen er en multivariat normalfordeling med gjennomsnittlig vektor og varians–kovariansmatrise[5, (3.10), s.47]. Den er også er en objektiv estimator, noe som betyr at den i gjennomsnitt verken overestimerer eller underestimerer parameterne. Estimeringen skal brukes til å finne parameterne $\hat{\beta} = (X'X)^{-1}X'Y$ basert på en vektor y av observasjoner av Y oppnådd ved å bruke matrisen X. Dette er for å lage estimater for et nytt sett med prediktorverdier.

4.2 Lineær regresjon

Som en første "naiv" tilnærming til problemet, forsøkes det å forutsi hvor mange dager det tar å gjenopprette nøytrofiler ved å bruke kroppsvekten til mottakeren som eneste prediktor. Dette er fordi kroppsvekten kan påvirke ulike fysiologiske prosesser, inkludert metabolisme, immunfunksjon og hematopoese. Forskjeller i kroppsvekt kan påvirke produksjonen av nøytrofiler fra benmargen, og potensielt påvirke tiden det tar før nøytrofiltellingen når en bestemt terskel. En studie av Aoyama et al.[6] undersøker innvirkningen av kroppsmasseindeks (BMI) på 5-årsoverlevelsesrater hos pasienter som gjennomgikk HSCT. Studien fant at BMI kan være assosiert med forskjeller i langsiktige overlevelsesresultater etter HSCT. Med dette som grunnlag skal det bli sett på om det finnes en korrelasjon mellom vekt og antallet nøytrofiler, og regresjonsanalyse på disse to variabelene hjelper med å få kjennskap til fenomenet som studeres.

Denne enkle modellen forsøker å tilpasse en linje $\hat{\beta}_0 + \hat{\beta}_1 x$ til observasjonene. Med andre ord antar den at det er omtrent en lineær sammenheng mellom den avhengige variabelen Y (tid til gjenoppretting av nøytrofiler) og forklaringsvariabelen X (kroppsvekt).

For enkel lineær regresjon gir OLS følgende estimatorer[5, (3.24), s.52]:

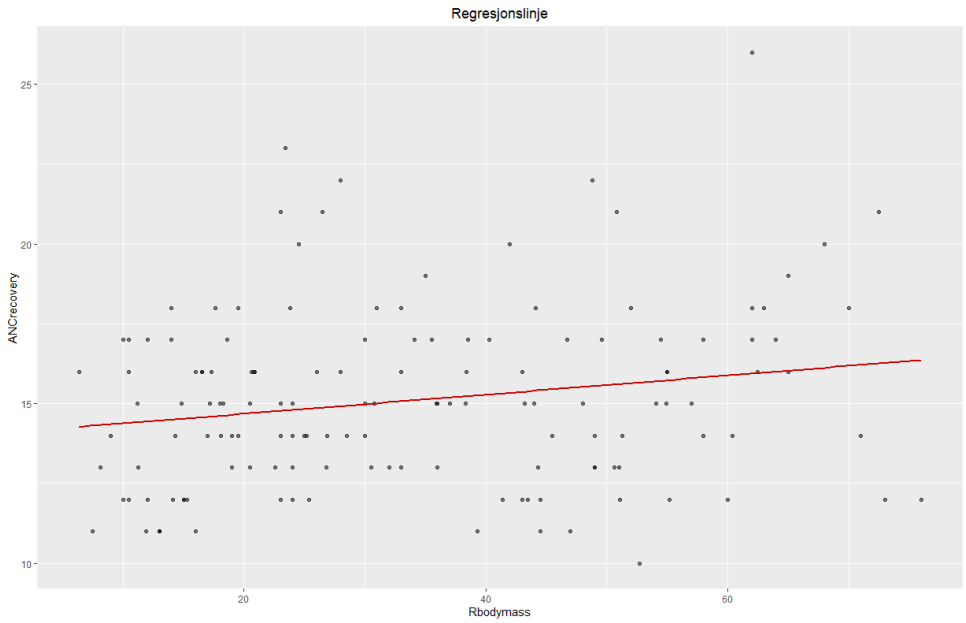
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$\hat{\beta}_0$ representerer skjæringspunktet, mens $\hat{\beta}_1$ tilsvarer stigningen på den tilpassede linjen. Oppsummeringen av den tilpassede enkle lineære modellen viser at estimatene er:

$$\hat{\beta}_0 = 14.08040$$
$$\hat{\beta}_1 = 0.03001$$

Dette betyr i gjennomsnitt at estimeringen er for hver kilo økning i kroppsvekt til mottakeren, forventes tiden for gjenoppretting av nøytrofiler å øke med omtrent 0.03001 dager, når alle andre variabler holdes konstante.

Denne følgende figuren viser den tilpassede regresjonslinjen. Intuitivt er det mulig å si at kroppsvekt som eneste variabel ikke er tilstrekkelig for å forutsi gjenopprettingstiden, da dataen består av mange kliniske variabler, og menneskekroppen er et komplekst system. Oppsummeringen viser også en p-verdi på 0.02345.

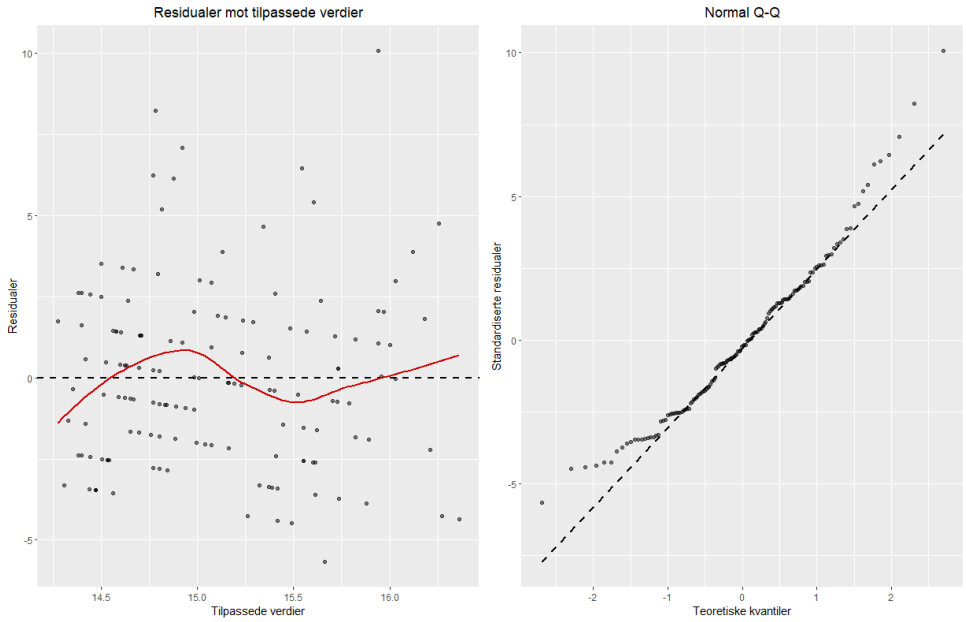


Figur 4: Lineær regresjonslinje for den første modellen

4.2.1 Analyse av residualer

En residual (eller feilleddet/restleddet) e_i er forskjellen mellom den observerte verdien y_i og den estimerte verdien $\hat{y}_i = x_i\hat{\beta}$[4, s.70]. Residualer er forskjellige fra feil, som er definert som forskjellen mellom den observerte verdien y_i og den sanne (ikke-observerbare) verdien, men det er mulig å betrakte residualer som en empirisk tilnærming til feil.

Den lineære modellen gir kvartiler av fordelingen av residualer. For at den lineære modellen skal anses som pålitelig, bør residualene se ut som om de er trukket fra uavhengige og identiske fordelte feil $\epsilon \sim \mathcal{N}(0, \sigma^2)$. For å vurdere passformen til modellen nærmere, plottes residualene:



Figur 5: Analyse av residualer

Plottet til venstre i Figur 2 viser residualer mot de tilpassede verdiene. Dette plottet bør vise en sky av punkter uten et spesifikt mønster, ellers kan antagelsene om uavhengighet eller samme varians av feilene være feil. Legg merke til den røde linjen, som gir en visuell indikasjon på om det er noen mønstre i residualene som ikke blir fanget opp av modellen. Her har den først en stigende trend (underestimasjon), så en synkende trend (overestimasjon), og deretter en stigende trend igjen. Dette kan tyde på at modellen har problemer med å tilpasse seg dataene på en god måte over hele spekteret av tilpasse verdier.

Plottet til høyre i Figur 2 viser et Q-Q plott av standardiserte residualer. Dette plottet er nyttig for å sjekke om fordelingen av residualer har samme form som en normalfordeling, der punktene vil ligge på den teoretiske linjen. I dette tilfelle, hvor datapunktene starter på nesten -5 på y-aksen og sprer seg oppover i en ganske rett fasong, indikerer det at de standardiserte residualene ikke følger en normalfordeling. Istedenfor følger de en skjev fordeling eller har tydelig heteroskedastisitet (variasjonen i residualene endrer seg med den uavhengige variabelen).

4.2.2 Inferens på koeffisienter

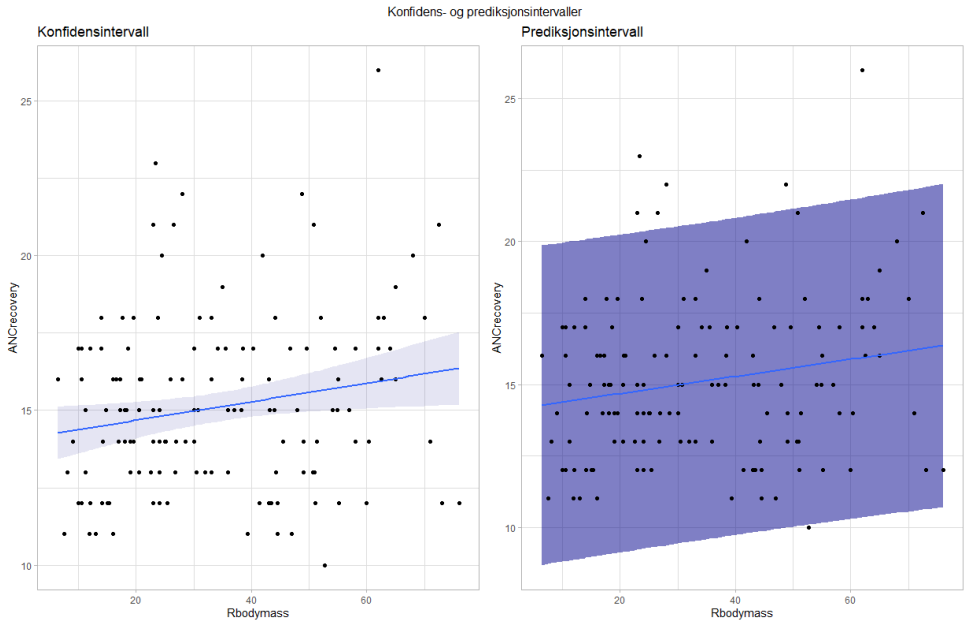
Dersom det antas at $\hat{\beta}$ har en multivariat normalfordeling er det mulig å beregne konfidens- og prediksjonsintervaller, og utføre hypotesetesting på koeffisientene. For å utlede en estimator for standardfeilen, er det først nødvendig å finne en estimator for feilvariansen σ^2. En objektiv estimator er middelkvadratet av residualer (MSR), som er lik kvadratsummen til residualene (RSS) delt på antall frihetsgrader.

$$\text{MSR} = \frac{\text{RSS}}{n - p}$$

Etter å ha introdusert en god estimator for σ^2, er det nå mulig å beregne konfidensintervaller for parameterne. De 95% konfidensintervallene for $\hat{\beta}_0$ og $\hat{\beta}_1$, ifølge dataene, er følgende:

	2.5%	97.5%
(Intercept)	13.086726038	15.07406995
Rbodymass	0.004113736	0.05590805

Plotter 95% konfidens- og prediksjonsintervaller:



Figur 6: Konfidens- og prediksjonsintervaller

I tillegg til dette er det mulig å lage disse hypotesetestene:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

Nullhypotesen tilsvarer å si at den tilknyttede i-te prediktoren ikke påvirker gjennomsnittlig utfall. Med andre ord ønskes det å forkaste nullhypotesen om at den i-te prediktoren ikke er nyttig og bør fjernes fra modellen. Den alternative hypotesen tilsvarer det motsatte, at den tilknytte i-te prediktoren påvirker utfallet, og at denne prediktoren er nyttig for å forklare variasjonen.

Disse hypotesetestene bruker t-statistikken. For hver koeffisient i modellen viser R-utdataen verdien av t-statistikken og dens p-verdi. I dette tilfellet er p-verdien for koeffisienten til kroppsvekt regnet som 0.02345, noe som antyder at denne prediktorvariabelen har en stor effekt på det gjennomsnittlige utfallet. Med denne p-verdien sammenlignet med et vanlig signifikansnivå på 0.05, finnes det et argument for at det er bevis nok til å forkaste nullhypotesen om at kroppsvektskoeffisienten er null. Dette antyder at denne koeffisienten har en stor sannsynlig påvirkning, men samtidig er det viktig å være klar over at korrelasjon ikke nødvendigvis antyder kausalitet. Det kan være andre påvirkende variabler som ikke har blitt tatt hensyn til i denne relativt enkle modellen, og dermed kan den observerte sammenhengen være indirekte eller direkte bestemt av disse variablene.

4.2.3 Passform

Resultatene av den enkle lineære regresjonen viser følgende målinger. Disse hjelper med å forstå ytelsen til modellen.

- Residual standard error (RSE), en objektiv estimator for standardavviket til feilene, som er kvadratroten av MSR.
$$\text{RSE} = \sqrt{\text{MSR}}$$
- Determinasjonskoeffisienten R^2, definert som forholdet mellom variasjonen i y forklart av lineær regresjon på prediktorene. Med andre ord er det forholdet mellom mengden variasjon i utfallet forklart av regresjonen og den totale mengden variasjon i utfallet, dvs. residual sum av kvadrater som tilsvarer nullmodellen.
- Resultater av en global test, nemlig verdien av F-statistikken og dens p-verdi. Nullhypotesen for den globale testen er $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$, dvs. alle koeffisientene bortsett fra skjæringspunktet er lik null. Hvis p-verdien er liten nok, kan nullhypotesen forkastes da regresjonen ikke er nyttig.

Modellen gir som sagt en p-verdi på 0.02345 for den globale testen, i tilegg en R^2-verdi på 0.03614. Dette indikerer at det finnes en korrelasjon mellom variablene, men modellen forklarer kun en liten del av variansen i responsvariabelen. Igjen, dette kan skyldes at modellen ikke fanger opp alle relevante variabler som påvirker responsvariabelen, eller at det er stor tilfeldig variasjon i dataene som ikke kan forklares av modellen.

4.2.4 Måleparametere

I tillegg brukes det følgende måleparametere for å evaluere den generelle prediktive ytelsen til modellene:

- Gjennomsnittlig absolutt feil (MAE):
$$\text{MAE} = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$
En mindre MAE-verdi resulterer i en bedre prediktiv modell. Denne måleparameteren er ganske robust mot ekstremverdier, så derfor kan en modell med relativt liten MAE fortsatt regne ut svært høy feil på ekstremverdier.
- Kvadratisk gjennomsnittlig feil (RMSE):
$$\text{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$
Denne måleparameteren er mer følsom for ekstremverdier enn MAE. En mindre RMSE tilsvarer en bedre modell.

Merk at RMSE og RSE bare skiller seg fra hverandre i nevneren. RMSE er en subjektiv estimator for standardavviket til feilene, i motsetning til RSE som er objektiv.

Denne tabellen viser verdiene for MAE og RMSE på treningsmengden for den enkle modellen. Disse verdiene gir en måling av nøyaktigheten til modellen og skal senere brukes til å evaluere og sammenligne ytelsen til en annen modell.

MAE	RMSE
2.189202	2.777606

4.3 Generell lineær modell

Selv om den enkle modellen for lineær regresjon kunne gi en predikasjon innenfor en akseptabel ramme, er målet å utvikle en mer kvantitativ modell. Den nye modellen tar hensyn til alle passende prediktorer (korrelasjon ≥ 0.15 fra korrelasjonsmatrisen i analysedelen) og kombinerer variansanalyse og regresjonsanalyse. Målet med denne modellen er å gi en grundigere beskrivelse av den kvantitative sammenhengen mellom responsvariabelen Y og de andre variablene som påvirker den.

```
Call:
lm(formula = ANCrecovery ~ Rbodymass + Recipientageint + CD3dCD34 +
    PLTrecovery, data = stemcelldata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4542 -1.9965 -0.2175  1.4807 10.4699

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.388e+01  5.166e-01  26.867   <2e-16 ***
Rbodymass       1.434e-02  2.436e-02   0.589   0.5570
Recipientageint 3.720e-01  5.250e-01   0.709   0.4798
CD3dCD34        4.602e-02  2.384e-02   1.930   0.0556 .
PLTrecovery     2.770e-06  1.614e-06   1.716   0.0885 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.756 on 137 degrees of freedom
Multiple R-squared:  0.08418,   Adjusted R-squared:  0.05744
F-statistic: 3.148 on 4 and 137 DF,  p-value: 0.01638
```

Figur 7: Generell lineær modell med variabler med ≥ 0.15 lineær korrelasjon

Den globale testen har en p-verdi på 0.01638. Selv om dette er en lav p-verdi, og nullhypotesen kan forkastes, så er den likevel ikke veldig nær null. Dette betyr at denne modellen kanskje fortsatt ikke er i stand til å forutsi responsen nøyaktig ved hjelp av de tilgjengelige prediktorene.

For å teste om den fullstendige modellen er bedre enn den forrige enkle lineære modellen, kan ANOVA[4, s.55] brukes. Denne metoden er nyttig for å sammenligne en stor modell med en mindre modell, forutsatt at de er "nestet".

```
Analysis of Variance Table

Model 1: ANCrecovery ~ Rbodymass
Model 2: ANCrecovery ~ Rbodymass + Recipientageint + CD3dCD34 + PLTrecovery
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    140 1095.5
2    137 1040.9  3    54.609 2.3957 0.07093 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figur 8: ANOVA med begge modellene

Resultatene fra ANOVA tyder på at nullhypotesen kan forkastes om at den enkle lineære modellen er tilstrekkelig sammenlignet med den fullstendige modellen. F-statistikkens p-verdi er ganske lav, men den er ikke så lav som ønsket. Dette antyder at den fullstendige modellen muligens gir en bedre forklaring på variasjonen i responsvariabelen enn den enkle lineære modellen, men det er fortsatt rom for forbedringer i modellen, egentlig begge modellene.

5 Logistisk regresjon

Det er fullt mulig å tilnærme seg problemet med å forutsi variablene i dataen på en annen måte, gitt de forskjellige datatypene i datasettet. En betydelig del av datasettet består av binære variabler, og derfor kan logistisk regresjon være en mer passende løsning, da den er egnet for slike variabler.

5.1 Teoretisk ramme

Her består utfallsvektoren av binære observasjoner som er trukket fra en Bernoulli-fordeling, der for eksempel Y_i representerer om pasient i døde ($Y_i = 1$) eller overlevde ($Y_i = 0$). Dette kan uttrykkes som:

$$Y_i \sim \text{Bernoulli}(p_i)$$

der $p_i \in [0, 1]$ er sannsynligheten for at pasient i overlevde eller ikke.

En lineær regresjonsmodell, uttrykt som $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, forsøker å forutsi Y_i som en lineær funksjon av en prediktor X_i, men den lineære modellen er ikke passende for binære utfall. Siden Y_i er binær, vil forventet verdi av Y_i være p_i. Samtidig er en lineær regresjonsmodell ikke begrenset til å gi verdier mellom 0 og 1 for p_i. Dette forårsaker ugyldige prediksjoner når p_i overstiger 1 eller faller under 0, og derfor er det nødvendig med logistisk regresjon, som sikrer at prediksjonene faller innenfor det gyldige området for sannsynligheter, dvs. $p_i \in [0, 1]$[4, s.79].

5.1.1 Generalisert lineær modell (GLM)

Den generelle lineære modellen antar at utfallsvariabelen Y er kvantitativ og feilene er normalfordelte. Den generaliserte lineære modellen har derimot ikke disse antagelsene. Derfor kan denne brukes når responsvariabelen har en Bernoulli-fordeling[4, s.86], altså at den er binær.

Aspektene i den generaliserte lineære modellen er[5]:

- en responsvektor $Y = (Y_1, \ldots, Y_n)'$ med gjennomsnitt $\mu = (\mu_1, \ldots, \mu_n)'$
- en lineær prediktor $\eta = X\beta$ uttrykt som en lineær kombinasjon av ukjente parametere
- en lenkefunksjon $g(\cdot)$ slik at $g(\mu_i) = \eta_i$

Merk at når $g(\mu_i) = \mu_i$, dvs. lenkefunksjonen er identitetsfunksjonen, samsvarer dette rammeverket med lineær regresjon[5, s.296]. I dette tilfellet er $\eta = \mu = X\beta$.

Av grunnene som ble forklart i forrige avsnitt, kan det ikke brukes en identitetsfunksjon som en lenke. Et mer passende valg er da logit-lenkefunksjonen[4, p.79]

$$\text{logit}(p) = \log \frac{p}{1-p}$$

Logiten kalles også logodds-funksjonen, siden den er logaritmen til oddsen $\frac{p}{1-p}$ der p er en sannsynlighet.

5.2 Logistisk regresjonsmodell

Den første logistiske regresjonsmodellen som lages er en enkel modell som forsøker å forutsi utfallet av behandlingen med tanke på overlevelse, ved å gjøre en tredimensjonal analyse med prediktorer for tid for gjenoppretting av nøytrofiler i blodet, samt om pasienten er i risikogruppen.

	2.5%	97.5%
(Intercept)	ANCrecovery	Riskgroup
-0.77250495	0.01910159	0.63196179

Disse koeffisientene gir innsikt i sammenhengen mellom variablene og overlevelsesstatus. Intercept-koeffisienten representerer den forventede logaritmen til oddsen for overlevelse når alle andre prediktorer er null. Koeffisienten for ANCrecovery indikerer hvordan endringer i risikogruppen påvirker logaritmen til oddsen for overlevelse, mens koeffisienten for Riskgroup sammenligner logaritmen til oddsen for overlevelse mellom de to variablene definert av denne variabelen. Disse koeffisientene gir altså et grunnlag for å forstå hvordan de ulike variablene påvirker sannsynligheten for overlevelse.

Den første estimerte koeffisienten er Riskgroup. Formelen for estimatet av logiten for den i-te observasjonen:

$$\hat{\text{logit}}(p_i) = x_i\hat{\beta} = c + 0.632 \cdot \text{Riskgroup}$$

Her avhenger konstanten c av verdien til de andre prediktorene for den i-te prøven, som antas forblir konstante. Denne tilpassede logistiske regresjonsmodellen anslår at $\hat{\text{logit}}(p_i)$ er lik c+0.632 hvis risikogruppen er 1, dvs. høy risiko, og til c ellers.

Dette antyder at parameterestimatet 0.632 er forskjellen mellom logiten når overlevelsesstatusen er lik 1, som her er $\hat{\text{logit}}(p_i)_1$, og logiten når prediktoren er lik 0, som her er $\hat{\text{logit}}(p_i)_0$. Denne forskjellen er lik odds ratio[4, s.79].

$$0.632 = \hat{\text{logit}}(p_i)_1 - \hat{\text{logit}}(p_i)_0 = \log \frac{\left(\frac{\hat{p_i}}{1-\hat{p_i}}\right)_1}{\left(\frac{\hat{p_i}}{1-\hat{p_i}}\right)_0}$$

Jo høyere odds ratio er, desto større er forskjellen i sannsynligheten for at responsen er lik 1 når risikogruppen er 1 sammenlignet med når den er 0.

For å kvantifisere den differensielle effekten av denne binære prediktoren på en bestemt prøve i, kan det beregnes de estimerte sannsynlighetene. Hvis det antas at den i-te observasjonen ble tatt av en pasient som hadde tid for gjenoppretting av nøytrofiler på 15 dager som tilhørte risikogruppen men deretter overlevde, og det ønskes å estimere effektene av overlevelsesstatusen på responsen, blir det følgende:

$$\hat{\text{logit}}(p_i) = x_i\hat{\beta} = -0.773 - 0.019 \cdot \text{ANCrecovery} + 0.632 \cdot \text{Riskgroup}$$

$$\hat{\text{logit}}(p_i) = x_i\hat{\beta} = -0.773 - 0.019 \cdot 15 + 0.632 \cdot \text{Riskgroup}$$
$$= -1.058 + 0.632 \cdot \text{Riskgroup}$$

Ved å ta inversen av logitten, beregnes det en sannsynlighet som avhenger av verdien til overlevelsesstatus:

$$\hat{p_i} = \begin{cases} \frac{1}{1+e^{-(-1.058+0.632)}} = 0.395 & \text{hvis Riskgroup} = 1 \\ \frac{1}{1+e^{-1.058}} = 0.742 & \text{ellers} \end{cases}$$

Da sier modellen at den estimerte sannsynligheten for overlevelse for en pasient med en spesifikk tid for gjenoppretting av nøytrofiler, men som tilhører risikogruppen, er omtrent 0.395. På den andre siden, hvis pasienten ikke tilhører risikogruppen, så er den estimerte sannsynligheten for overlevelse omtrent 0.742. Dette er en markant forskjell, og noe som muligens var forventet.

Hvis man ser på koeffisienten for nøytrofiler, betyr en økning på 1 dag en estimert økning på omtrent 0.019 i logoddsene for overlevelse. Dette antyder at pasienter med høyere verdier har en litt høyere sannsynlighet for overlevelse. Samlet sett viser modellen at både ANCrecovery og Riskgroup er viktige prediktorer for overlevelse, og kan brukes til å estimere sannsynligheten for overlevelse.

6 Trebasert regresjon

I denne seksjon skal det gjøres trebasert regresjon med Decision Tree og Random Forest, for å se om det gir bedre resultater enn vanlige regresjonsmetoder. Decision Tree er en maskinlæringsalgoritme som brukes både for klassifiserings- og regresjonsoppgaver. Den fungerer ved at dataene deles inn i mindre undergrupper basert på verdiene til attributtene. Dette skjer gjennom en serie av beslutninger basert på prediktorvariablene. Målet med algoritmen er å skape et tre som gir nøyaktige prediksjoner ved å segmentere dataene i så homogene grupper som mulig[7].

Random Forest er en maskinlæringsalgoritme som er bygget på Decision Trees. Det er en ensemble-metode som kombinerer flere trær for å forbedre nøyaktigheten og stabiliteten til modellen. Random Forest lager et stort antall trær, hver trent på et tilfeldig utvalg av data og variabler. Den endelige prediksjonen er et gjennomsnitt av prediksjonene fra alle trærne i skogen[8].

6.1 Teoretisk ramme

For å lage et regresjonstre så først deles prediktorene inn i J forskjellige, ikke-overlappende delområder $R_1, \ldots, R_J$ ved hjelp av beslutningsregler. Deretter gjøres den samme prediksjonen $\hat{y}_{R_j}$ for hver observasjon som faller innenfor den j-te delområde. Prediksjonen for den j-te rommet er gjennomsnittet av utfallene til "prøvene" inne i den.

Delområdene velges slik at de minimerer RSS. I prinsippet er jo lavere RSS, jo bedre er splitten, siden prediksjonene er nærmere det faktiske utfallet. Formelen for RSS er[5, s.12]:

$$\text{RSS} = \sum_{j=1}^{J} \sum_{i} (y_i - \hat{y}_{R_j})^2$$

Siden det er ikke mulig å finne alle partisjoner av funksjonsrommet for å finne den som minimerer RSS, brukes det rekursiv binærsplitting, en grådig tilnærming som starter på toppen av treet og deler opp prediktorområdet i rekkefølge[5, s.305]. Ved hvert trinn velger algoritmen den beste prediktoren og skjæringspunktet for å dele området, slik at splitten gir minst mulig RSS. Den er grådig fordi den ikke tar hensyn til mulige fremtidige splitter som kan gi en enda mindre RSS i fremtiden. Hvert trinn deler en eksisterende node inn i to noder, og denne prosessen gjentas til et stoppkriterium er nådd (f.eks. maksimal dyp eller maksimalt antall noder).

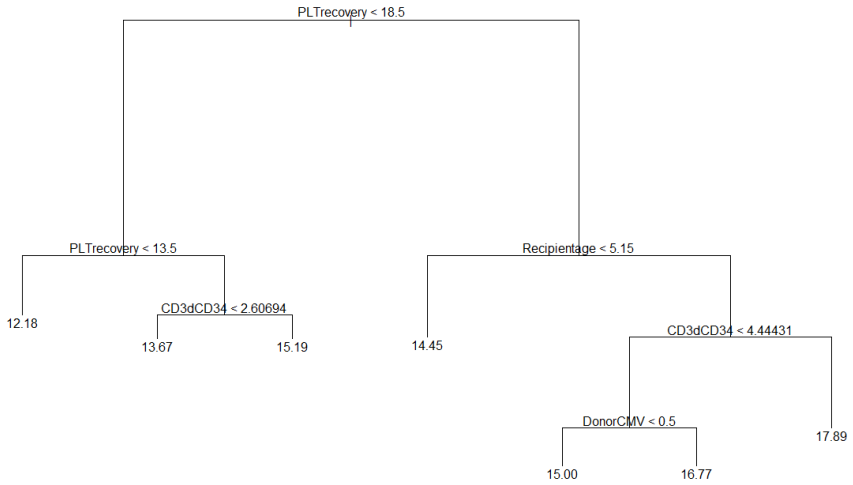
6.2 Decision Tree

Den første trebaserte algoritmen som brukes er Decision Tree (eller beslutningstre). Her blir det minimale antallet observasjoner som skal inkluderes i hver barne-node satt til 15. Dette begrenser antall bladnoder for å redusere overtilpasning og gjøre treet enklere å tolke.

```
Regression tree:
tree(formula = ANCrecovery ~ Rbodymass + Recipientage + CD3dCD34 +
    PLTrecovery + DonorCMV + RecipientCMV + extcGvHD, data = stemcells,
    mincut = 15)
Variables actually used in tree construction:
[1] "PLTrecovery"  "CD3dCD34"     "Recipientage" "DonorCMV"
Number of terminal nodes:  7
Residual mean deviance:  5.65 = 791.1 / 140
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.0000 -1.6670 -0.1765 -0.0246  1.2800  9.2270
```

Figur 9: Oppsummering av modellen for Decision Tree

Modellsummeringen sier at bare et delsett av alle tilgjengelige prediktorer ble valgt av algoritmen for å konstruere treet; dette er fordi algoritmen mener at disse variablene er gode nok for å gi prediksjoner om målvariabelen ANCrecovery. Utdataen gir også tilleggsinformasjon om residualer, nemlig fordelingen av residualer i kvartiler og residualmiddelavviket, som er lik MSR som tidligere definert. En av de viktigste fordelene med å bruke et enkelt regresjonstre er at som nevnt den er lett å tolke. Her plottes treet for å se hvordan det tar beslutningene sine, basert på verdien av prediktorene:



Figur 10: Grafisk fremstilling av modellen for Decision Tree

Her tilsvarer den en venstre grenen utfallene der indeksverdien er mindre enn grenseverdien, mens høyre gren tilsvarer utfallene der den er lik eller høyere enn grenseverdien. Rotnoden deler prediktorområdet basert på verdien av PLTrecovery, som er tid til oppbygging av blodplater. I følge dette treet så jo høyere tid for oppbygging av blodplater, så blir det gjennomsnittlig høyere tid på gjenoppretting av nøytrofiler i blodet (vist ved tallene nederst i treet, som representerer antall dager i gjennomsnitt). Den andre splitten gjøres på alderen til mottakeren, Recipientage, og setter en grenseverdi på 5.15 for å ytterligere dele området for prediktorer der PLTrecovery er større enn 18.15. Videre splitter den på variabler den mener er relevant som CD3dCD34 (forholdet mellom CD3+ celle og CD34+ celle), og DonorCMV (tilstedeværelse av virusinfeksjon hos giver) for å predikere gjennomsnittlig verdi av ANCrecovery.

For å evaluere ytelsen til dette beslutningstreet, beregnes det MAE og RMSE på treningsmengden:

MAE	RMSE
1.800135	2.359448

Dette beslutningstreet gjør det rimelig greit på testmengden, med enn lavere MAE og RMSE enn for lineær regresjon. Dette kan skyldes at beslutningstreet har funnet noen ikke-lineære forholder mellom variablene, og har ingen antakelser om residualene som lineær regresjon har.

6.3 Random Forest

En av de større ulempene ved enkle beslutningstrær er høy varians. Dette indikerer at ved å bruke et annet utvalg fra den samme populasjonen, kan det resulterende treet variere betydelig fra det som ble laget tidligere. For å løse denne utfordringen, benyttes det en teknikk kjent som heter aggregert bootstrapping[5, s.282], som er metode som brukes spesielt innenfor maskinlæring, spesielt for ensemble-metoder som Random Forest. Denne bygger på den statistiske teknikken bootstrapping[4, s.28].

I følge delen om Random Forest i The Elements of Statistical Learning[5, s. 587–593] så tar Random Forest tar ideen om bootstrapping et skritt videre ved å bruke en prosedyre som fjerner korrelasjon fra trærne, som går ut på at det bare vurderes et begrenset antall prediktorer ved hvert skille. For regresjon er dette antallet vanligvis satt til $m = \frac{p}{3}$[5, s.592]. Dermed velges det ved hvert skille av hvert tre tilfeldig bare et delsett av prediktorene, og vurderer bare disse for å evaluere mulige splitt. Siden trærne med høy varians blir satt sammen, og trærne blir tvunget til å bruke forskjellige prediktorer, er prediksjonene fra Random Forest mindre korrelerte, noe som fører til mindre varians.

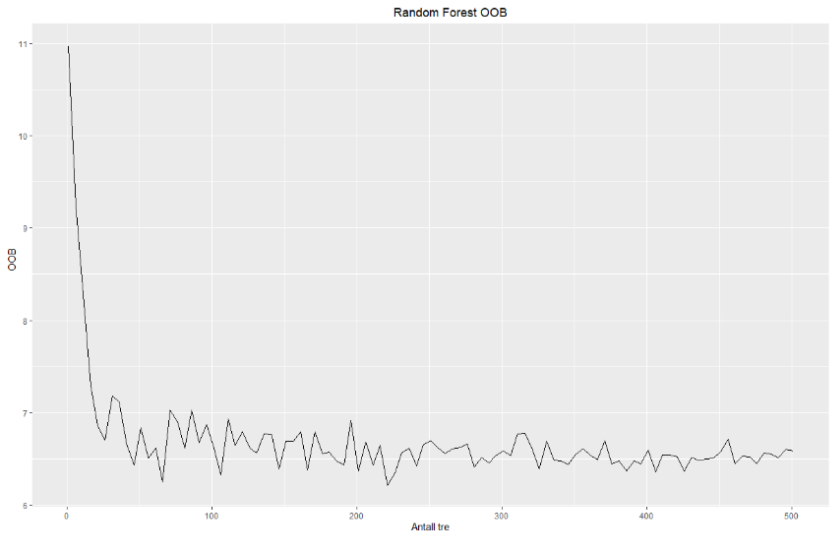
Denne utdataen tilpasser en tilfeldig skog med 500 trær og 4 prediktorer som velges tilfeldig for splitting ved hver node:

```
Type:                             Regression
Number of trees:                  500
Sample size:                      142
Number of independent variables:  7
Mtry:                             4
Target node size:                 5
Variable importance mode:         none
Splitrule:                        variance
OOB prediction error (MSE):       6.546919
R squared (OOB):                  0.1878413
```

Figur 11: Random Forest med 500 trær og 4 prediktorer

Utdataen viser to viktige måleparametere, nemlig MSR og R^2. Merk at begge disse måleparameterne ikke direkte beregnes på treningsmengden, men heller på out-of-bag (OOB) observasjoner.

Som nevnt bruker hvert tre omtrent to tredjedeler av observasjonene fra den opprinnelige treningsmengden. Naturligvis betyr dette at et gitt tre i gjennomsnitt ikke ser en tredjedel av observasjonene, og dette er OOB-observasjonene[5, s.593]. Derfor er det mulig å få et estimat av feilen på usette data forutsi responsen til den i-te observasjonen fra treningsmengden ved å gjennomsnittet prediksjonene fra trærne der den i-te observasjonen var out-of-bag. Disse prediksjonene kan brukes til å beregne OOB-måleparameterne; "OOB prediction error", eller gjennomsnittlig feil, ser man i utdraget med en verdi på omtrent 6.547. Denne følgende figuren viser denne verdien som en funksjon av antall tilpassede trær:

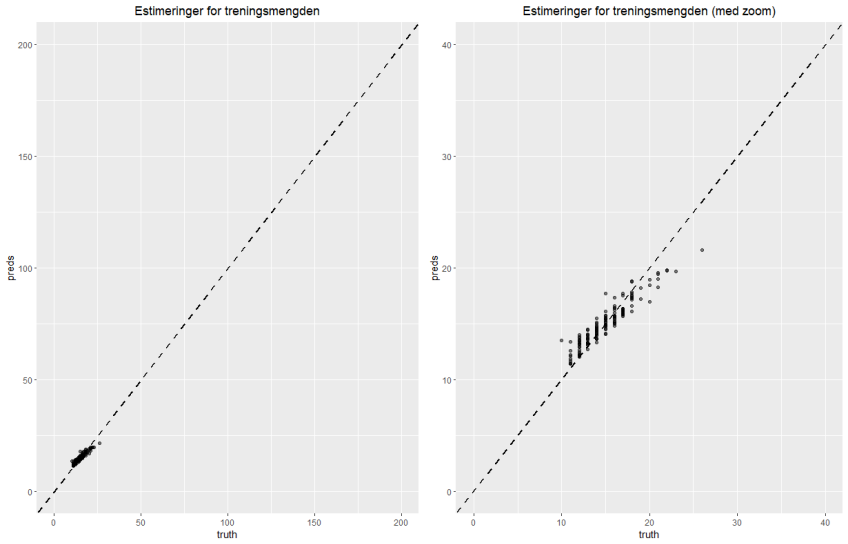


Figur 12: Graf for OOB prediction error

Grafen viser at feilerroren når sitt minimum rundt 200 trær, og forblir stabilt deretter. Med det finnes det nå en prediktiv modell, som kan brukes til å observere treningsfeilen. Derfor beregnes det MAE og RMSE på treningsmengden:

MAE	RMSE
0.885462	1.147544

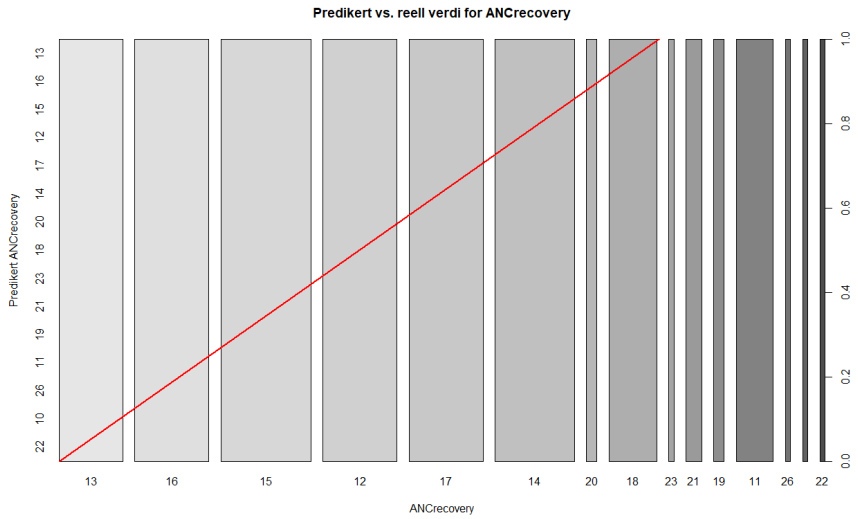
Denne tilfeldige skogen gir en treningsfeil lavere enn den "beste" lineære modellen. For å visualisere størrelsen på feilene på treningsmengden, plottes prediksjonene fra modellen mot den faktiske verdien.



Figur 13: Estimeringer for feilene på treningsmengden

Modellen ser ut til å passe bra, da punktene ligger tett rundt den 1:1-linjen (linjen der de predikerte verdiene er nøyaktig like de faktiske verdiene), og det antyder at modellen gjør nokså nøyaktige prediksjoner.

For å teste dette i praksis, lages det en enkel modell for å se om den klarer å predikere verdier for ANCrecovery:



Figur 14: Random Forest predikasjoner

Her viser den at den har en nøyaktighet på 100%, da alle datapunktene ligger på den røde linjen. Selv om dette er et utmerket resultat, så kan det være forårsaket av overtilpasning. Overtilpasning skjer når en modell lærer treningsmengden for godt, og fanger opp støy og tilfeldige svingninger i dataene i stedet for de underliggende mønstrene[9]. Som et resultat kan en overtilpasset modell prestere utmerket på akkurat denne treningsmengden, men mislykkes i å generalisere til usette data. Overtilpasning oppstår vanligvis når en modell er for "kompleks" i forhold til mengden treningsdata som er tilgjengelig, som kan være problemet her da det er ikke mye data i treningsmengden.

7 Konklusjon

I dette prosjektet ble det forsøkt å forutsi forskjellige variabler i forbindelse med HSCT-behandling i pediatriske pasienter. De statistiske metodene som ble brukt var lineær regresjon, logistisk regresjon, samt maskinlæringsalgoritmene Decision Trees og Random Forest. Bruken av disse ulike metodene ga muligheten til å se hvilke variabler som påvirker utfallet, og gi en bedre forståelse av sammenhengen.

Det var Random Forest-metoden som klart ga best resultat og de laveste måleparameterne, selv med overtilpasning, noe som indikerer dens overlegenhet i forhold til de andre modellene i prosjektet. Dette indikerer at den metoden bedre klarte å fange opp de underliggende mønstrene og sammenhengene i dataen, og dermed gi mer nøyaktige forutsigelser for responsvariabelen Y. Dette gir mening da denne metoden er spesielt effektiv når det er store sammenhenger mellom variabler, og erfaringsmessig har tidligere vist seg å være god i prediktive analyser.

Disse resultatene støtter bruken av Random Forest som en effektiv modell for prediktive analyser, spesielt når man håndterer store datasett med flere variabler. Hvis jeg skulle ha gjort prosjektet større og mer utfyllende, så ville jeg ha optimalisert den på en annen måte, og eventuelt sett på konfigurasjoner av den, som for eksempel endret metoden av splitting for optimalisering, eller tilpasset en hybridmodell som kombinerer Random Forest med andre maskinlæringsalgoritmer for å dra nytte av styrkene de har og forbedre ytelsen ytterligere. Dette kan inkludere bruk av K-Nearest Neighbors (KNN) innenfor Random Forest-ensemblet eller bruk av et nevralt nettverk for å behandle utdataene fra Random Forest, og dermed lage et simuleringsprosjekt.

Videre, et forbedringspotensiale er bruken av variabler. For den generelle lineære modellen kunne det vært mulig å laget dummyvariabler, og her kunne det blitt brukt flere, og analysen kunne vært mer utfyllende. Hvis jeg skulle ha sett på andre variabler så ville jeg ha sett på de kategoriske og eventuelt brukt en Chi-square test for å se på uavhengigheten mellom to kategoriske variabler, og for å evaluere hvor godt de observerte dataene passer til en forventet fordeling, for eksempel en normalfordeling[4, s.61].

Når det gjelder hypotesetestingen så var valget av signifikansnivået et usikkert valg. Valg av signifikansnivå i hypotesetesting er avgjørende, da det balanserer risikoen for Type I- og Type II-feil (falsk positiv/falsk negativ)[4, s.44]. Vanligvis er signifikansnivået $\alpha = 0.05$, men denne verdien kan være upassende for dette prosjektet. For å løse dette, kunne sammenligning med lignende studier, samt en generell grundigere litteraturgjennomgang, ha løst dette.

Uansett så er jeg fornøyd med prosjektet, og gjennom prosjektet fikk jeg en praktisk forståelse av statistiske metoder, da det å arbeide med virkelige datasett og konkrete problemstillinger var opplysende. Jeg forsøkte å gå utenfor pensum ved anvending av maskinlæringsalgoritmer vi lærte i tidligere semestere for å få andre perspektiver på problemløsning innenfor statistikk. I henhold til læringsutbyttene for MA-223, så lærte jeg det å kunne sette seg inn i og forstå statistiske rapporter, anvende sannsynlighetsregning og statistisk inferens på data og beregning av usikkerhet, risiko og pålitelighet. Alt i alt var prosjektarbeidet en lærerik opplevelse som forberedte meg bedre til å møte utfordringer innen statistikk og dataanalyse i fremtiden.

8 Vedlegg

8.1 Kode

Biblioteker, innstillinger for grafer og laste inn data

```
library(tidyverse)
library(ggplot2)
library(ggcorrplot)
library(cowplot)
library(tree)
library(ranger)
library(caret)
library(gridExtra)
library(reshape2)

point_alpha <- 0.5
line_color <- "red3"

stemcelldata <- read.csv("stemcell.csv")
```

Konvertere variabler til numeriske verdier og fjerne nullverdier

```
stemcelldata$Disease <- as.numeric(as.factor(stemcelldata$Disease))
stemcelldata[] <- lapply(stemcelldata, as.numeric)
stemcelldata <- na.omit(stemcelldata)
```

Splitte i testing- og treningsmengder

```
#dataen splittes i henhold til vanlig metode innenfor maskinl ring
trening_nrow <- floor(0.8 * nrow(stemcelldata))
set.seed(42)
trening_idx <- sample(seq_len(nrow(stemcelldata)), size=trening_nrow)
stemcells <- stemcelldata[trening_idx, ]
```

Konvertere dataen til en dataframe

```
if (!is.data.frame(stemcelldata)) {
  stemcelldata <- as.data.frame(stemcelldata)
}
```

Analyse

```
#korrelasjonsmatrise fra innebygd funksjon
str(stemcelldata)
cm <- cor(stemcelldata)
ggcorrplot(cm, type="lower", lab=FALSE)
```

```
cor_with_ANC <- cm[, "ANCrecovery"]
sorted_cor_anc <- sort(cor_with_ANC, decreasing = TRUE)
print(sorted_cor_anc)
```

```
ggplot(stemcelldata, aes(x = Rbodymass, y = ANCrecovery)) +
    geom_point(color = "red3") +
    geom_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE, color = "
    blue") +
    labs(title = "Spredningsdiagram over kroppsvekt i forhold til tid av ANC
    gjenoppretting",
         x = "Kroppsvekt", y = "Antall dager til gjenoppretting av ANC")
```

Lineær regresjon

```r
lm <- lm(ANCrecovery ~ Rbodymass, data=stemcelldata)
summary(lm)

ggplot(data = stemcelldata, mapping = aes(x = Rbodymass, y = ANCrecovery)) +
  geom_point(alpha = point_alpha) +
  geom_smooth(method = "lm", color = line_color, se = FALSE) +
  ggtitle("Regresjonslinje med ANCrecovery som avhengig variabel") +
  theme(plot.title = element_text(hjust = 0.5))
```

Analyse av residualer

```r
res_fit_df <- data.frame(
  "residuals" = lm$residuals,
  "fitted" = lm$fitted.values
)

res_fit <- ggplot(data=res_fit_df, mapping=aes(x=fitted, y=residuals)) +
  geom_abline(slope=0, intercept=0, color="black", linetype=2, size=1) +
  geom_point(alpha=point_alpha) +
  geom_smooth(color=line_color, se=FALSE) +
  ggtitle("Residualer mot tilpassede verdier") +
  ylab("Residualer") +
  xlab("Tilpassede verdier") +
  theme(plot.title = element_text(hjust = 0.5))

qq <- ggplot(data=res_fit_df, mapping=aes(sample=residuals)) +
  geom_qq(alpha=point_alpha) +
  stat_qq_line(color="black", linetype=2, size=1) +
  ggtitle("Normal Q-Q") +
  ylab("Standardiserte residualer") +
  xlab("Teoretiske kvantiler") +
  theme(plot.title = element_text(hjust = 0.5))

plot_grid(nrow=1, ncol=2, res_fit, qq)
```

Plotte konfidens- og prediksjonsintervaller

```r
conf.set = predict(lm, stemcelldata, interval="confidence")
conf.set = data.frame(conf.set)
stemcelldata = stemcelldata %>%
  mutate(lower.ci = conf.set$lwr, upper.ci = conf.set$upr)

plot.confidence = stemcelldata %>%
  ggplot(aes(x=Rbodymass,y=ANCrecovery, ymin=lower.ci, lwr, ymax=upper.ci)) +
  geom_ribbon(fill="darkblue",alpha = 0.1) +
  geom_point() +
  geom_smooth(method="lm",se = FALSE) +
  labs(x = "Rbodymass", y="ANCrecovery", title = "Konfidensintervall") +
  theme_light()

pred.set = predict(lm, stemcelldata,interval="predict")
pred.set = data.frame(pred.set)

stemcelldata = stemcelldata = stemcelldata %>%
  mutate(lower.ci = pred.set$lwr, upper.ci = pred.set$upr )

plot.prediction = stemcelldata %>%
  ggplot(aes(x=Rbodymass,y=ANCrecovery, ymin=lower.ci, lwr, ymax=upper.ci)) +
  geom_ribbon(fill="darkblue",alpha = 0.5) +
  geom_point() +
  geom_smooth(method="lm",se = FALSE) +
  labs(x = "Rbodymass", y="ANCrecovery", title = "Prediksjonsintervall") +
  theme_light()
```

```
27
28 gridExtra::grid.arrange(plot.confidence, plot.prediction, nrow = 1, ncol = 2,
       top = grid::textGrob("Konfidens- og prediksjonsintervaller"))
```

Inferens på koeffisienter, og måleparametere

```
1 confint(lm)
2
3 mae <- function(truth, preds) {
4   mae <- mean(abs(truth - preds))
5   return(mae)
6 }
7
8 rmse <- function(truth, preds) {
9   rmse <- sqrt(mean((truth - preds)^2))
10  return(rmse)
11 }
12
13 preds <- (predict(lm, stemcelldata, type="response")
14
15 print(data.frame("MAE"=mae(stemcelldata$ANCrecovery, preds), "RMSE"=rmse(
       stemcelldata$ANCrecovery, preds), row.names=c("lm")))
```

Komplett modell og ANOVA

```
1 komplett.lm <- lm(ANCrecovery ~ Rbodymass + Recipientage + CD3dCD34 +
       PLTrecovery + DonorCMV + RecipientCMV + extcGvHD, data=stemcelldata)
2 summary(komplett.lm)
3
4 print(anova(lm, komplett.lm))
```

Modell for logistisk regresjon

```
1 tw.glm <- glm(survival_status ~ ANCrecovery + Riskgroup, data=stemcelldata,
       family=binomial(link="logit"))
2 coefficients(tw.glm)
```

Decision Tree

```
1 set.seed(23)
2 stemcells <- subset(stemcells, ANCrecovery <= 30)
3 complete.tree <- tree(ANCrecovery ~ Rbodymass + Recipientage + CD3dCD34 +
       PLTrecovery, data=stemcells, mincut=15)
4 summary(complete.tree)
```

```
1 plot(complete.tree)
2 text(complete.tree, pretty=0)
```

```
1 complete.tree.preds <- predict(complete.tree, stemcells)
2
3 print(data.frame(
4   "MAE"=mae(stemcells$ANCrecovery, complete.tree.preds),
5   "RMSE"=rmse(stemcells$ANCrecovery, complete.tree.preds),
6   row.names=c("complete.tree")
7 ))
```

Random Forest

```
1 set.seed(23)
2
3 default.rf <- ranger(
4   ANCrecovery ~ Rbodymass + Recipientage + CD3dCD34 + PLTrecovery,
```

```
  data=stemcells,
  num.trees=500,
  mtry=4,
  splitrule="variance",
  max.depth=0
)

default.rf
```

```
ntrees <- seq(1, 501, 5)
oobpe <- vector("numeric", length(ntrees))

for (i in 1:length(ntrees)) {
    current.rf <- ranger(
        ANCrecovery ~ Rbodymass + Recipientage + CD3dCD34 + PLTrecovery,
        data=stemcells,
        num.trees=ntrees[i],
        mtry=12/3,
        splitrule="variance",
        max.depth=0
    )

    oobpe[i] <- current.rf$prediction.error
}

ggplot(data=data.frame(ntrees=ntrees, oobpe=oobpe)) +
    geom_line(mapping=aes(x=ntrees, y=oobpe)) +
    ylab("OOB prediction error") +
    xlab("Antall tre") +
    ggtitle("Random Forest OOB prediction error") +
    theme(plot.title = element_text(hjust = 0.5))
```

```
default.rf.preds <- predict(default.rf, stemcells)$predictions
print(data.frame(
  "MAE"=mae(stemcells$ANCrecovery, default.rf.preds),
  "RMSE"=rmse(stemcells$ANCrecovery, default.rf.preds),
  row.names=c("default.rf")
))
```

```
rfdtrp <- ggplot(data=data.frame(truth=stemcells$ANCrecovery, preds=default.rf.
    preds)) +
  geom_point(mapping=aes(x=truth, y=preds), alpha=point_alpha) +
  geom_abline(mapping=aes(intercept=0, slope=1), color="black", linetype=2, size
    =1) +
  coord_cartesian(xlim=c(0, 200), ylim=c(0, 200)) +
  ggtitle("Estimeringer for treningsmengden") +
  theme(plot.title = element_text(hjust = 0.5))

rfdtrp_zoom <- ggplot(data=data.frame(truth=stemcells$ANCrecovery, preds=default
    .rf.preds)) +
  geom_point(mapping=aes(x=truth, y=preds), alpha=point_alpha) +
  geom_abline(mapping=aes(intercept=0, slope=1), color="black", linetype=2, size
    =1) +
  coord_cartesian(xlim=c(0, 40), ylim=c(0, 40)) +
  ggtitle("Estimeringer for treningsmengden (zoomet inn)") +
  theme(plot.title = element_text(hjust = 0.5))

plot_grid(rfdtrp, rfdtrp_zoom, nrow=1, ncol=2)
```

```
rf_model <- randomForest(formula = ANCrecovery ~ Rbodymass + Recipientage +
    CD3dCD34 + PLTrecovery, data = stemcells)
predictions <- predict(rf_model, stemcells)
accuracy <- mean(predictions == stemcells$ANCrecovery)
print(accuracy)
```

Referanser

[1] L. Giske, V. Lauvrak, A. Stoinska-Schneider mfl., «Autolog hematopoietisk stamcelletransplantasjon ved multippel sklerose,» *Journal Name*, årg. 23, 2015, ISSN: 1890-1298.

[2] *Bone Marrow Transplant Children*, UCI Machine Learning Repository, Retrieved from `https://archive.ics.uci.edu/dataset/565`.

[3] M. Sikora, Ł. Wróbel og A. Gudyś, «GuideR: a guided separate-and-conquer rule learning in classification, regression, and survival settings,» *arXiv preprint arXiv:1806.01579*, 2018, Available at `https://arxiv.org/abs/1806.01579`. arXiv: `1806.01579 [cs.LG]`.

[4] A. Petrie og C. Sabin, *Medical Statistics at a Glance*, English, 2nd. Wiley-Blackwell, sep. 2005, s. 160, ISBN: 9781405127806.

[5] T. Hastie, R. Tibshirani og J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, English, 2. utg. Springer, feb. 2009, ISBN: 0387848576.

[6] T. Aoyama, A. Notsu, K. Ichimaru mfl., «Impact of Body Mass Index on 5-Year Survival Rates in Patients Undergoing Allogeneic Hematopoietic Stem Cell Transplantation,» *Nutr Metab Insights*, årg. 15, s. 11 786 388 221 128 362, 2022. DOI: `10.1177/11786388221128362`.

[7] Y.-Y. Song og Y. Lu, «Decision tree methods: applications for classification and prediction,» *Shanghai Archives of Psychiatry*, årg. 27, nr. 2, s. 130–135, apr. 2015. DOI: `10.11919/j.issn.1002-0829.215044`.

[8] A. Cutler, D. R. Cutler og J. R. Stevens, «Random Forests,» i *Ensemble Machine Learning: Methods and Applications*, Springer, 2011, kap. 5, s. 157–176. DOI: `10.1007/978-1-4419-9326-7_5`.

[9] X. Ying, «An Overview of Overfitting and its Solutions,» *Journal of Physics Conference Series*, årg. 1168, nr. 2, s. 022 022, 2019, ISSN: 1742-6596. DOI: `10.1088/1742-6596/1168/2/022022`.