Norges miljø- og
biovitenskapelige
universitet

Masteroppgave 2020 30 stp
Fakultet for realfag og teknologi

Statistisk analyse og optimering av kyllingproduksjon

(Statistical Analysis and Optimization of Poultry Production)

Morten Kvamsdal
Industriell økonomi og teknologiledelse, energi fysikk

Forord

Jeg vil gjerne takke de ansatte ved Nortura som har gjort tilgjengelig alt datamaterialet og assistert meg gjennom oppgaven med nødvendig dokumentasjon. Spesielt vil jeg takke Liv Marit Biltvedt for å være tilgjengelig for spørsmål gjennom hele prosessen.

Det har vært en ære og hatt med Kristian Hove Liland som bi-veileder, og Trygve Almøy som rådgiver. Kristian sin kunnskap om statistisk programmering har tidvis vært skremmende, men mest av alt veldig lærerikt. Trygve med sin utømmelige kunnskap og ærlige tilbakemeldinger har presset meg til å legge ned ekstra arbeid. Foruten om meg selv er Hilde Vilje, min veileder, den som har vært mest dedikert til denne oppgaven. Hilde har vist et engasjement langt utover hva man kan forvente av en veileder. Hilde har med sin kunnskap, smittsomme humør og omtanke gjort arbeidet denne våren til en lærerik og lystbetont prosess. Tusen takk!

Oslo, mai 2020

Morten Kvamsdal

Sammendrag

Slaktekyllingproduksjon er en svært marginal bransje hvor små forskjeller kan gi store utslag på produsentenes inntjening. I denne oppgaven er det undersøkt om noen produsenter jevnt og gjentakende har bedre dekningsbidrag per kylling enn andre produsenter, og eventuelt hvilke faktorer som er bidragsytende til dette resultatet. Norturas database, Fjørfekontrollen (01.01.18 – 01.12.19) er benyttet som datagrunnlag for analysene. Det relevante datamaterialet er konvertert til passende format og avviksverdier, støy og feilføringer er fjernet fra videre analyse. Grunnet store mengder manglende data er store deler av dataene enten fjernet eller imputert. Datasettets akkumuleringsnivå er lagt til innsett. Det vil si hvert innsett, eller produksjonsenhet har én observasjon per variabel. Preprosesseringen av dataene reduserte datamengden med over 90%.

For å kunne svare på oppgavens problemstilling er det benyttet ulike statistiske metoder for tilpasning av lineære modeller. Modellene ble trent opp til å predikere variasjonen i dekningsbidrag, før modellene ble evaluert og testet mot nye data. Modellen som presterte best inkluderer 68 variabler og evnet å forklare omkring 60% av variansen i dekningsbidraget gitt nye data. ANOVA og Tukey-testing bekreftet oppgavens hypotese om at noen produsenter har bedre dekningsbidrag enn andre produsenter. Selv etter å ha inkludert effekt fra faktorer som produksjonstype, produksjonsdato, ulike slakteri og rugeri (blokkfaktorer) viser resultatene at kyllingprodusent tydelig har effekt på dekningsbidraget. Modellkoeffisientenes estimater og forklaringsevne har igjen dannet grunnlaget for evalueringen av hvilke av faktorene som har effekt på dekningsbidraget. De mest lønnsomme produksjonstypene er McDonalds-kylling, Landkylling, Kyllinggården og Liveche. Økt dyrevelferd ser også ut til å ha positiv effekt på dekningsbidraget. Disse variablene sammen med en rekke andre faktorer ser tilsynelatende ut til å påvirke kyllingproduksjonens lønnsomhet.

Abstract

Poulty production is a marginal industry where small differences potential could have massive impact on a producers net profit. The object of this master thesis is to investigate if some producers repeatedly have a significantly higher/lower profit than others, and what factors that influences this result. The data material used for analysis is gathered form the database of Nortura, Fjørfekontrollen (01.01.18 - 01.12.19). Before starting the analysis all relevant data is converted to a preferable format before removing outliers and other erroneous registrations. Every batch of chickens has one row of observations, which mean that every production unit or batch has one observation for each variable. Due to the substantial amount of missing data a considerable amount is either removed or imputed. Preprocessing resulted in delation of over 90% of the data.

To answer the aim of this thesis there are used multiple different statistical methods for fitting different linear models. The models are trained with the aim of predicting producers expected contribution margin, before evaluated and tested agains new data. The best model included 68 variables and explained a little over 60% of the variation in the contribution margin given new data. ANOVA and Tukey testing confirmed the hypothesis of this thesis stating that some producers are in fact repeatedly preforming better/worse than other with regards to contribution margin. The result was still evident even after including the effects of production type, date of production, slaughterhouse, and hatchery (block factors) in the model. Based on the estimated model coefficients and their explainability the factors which is most influential are determined and evaluated. The production types McDonalds-kylling, Landkylling, Kyllinggården and Liveche seem to be the most profitable. Increased animal welfare also seem to have a positiv effect on the contribution margin. These variables together with a group of others is seemingly effecting the producers net profit.

Innhold

Kapittel 1

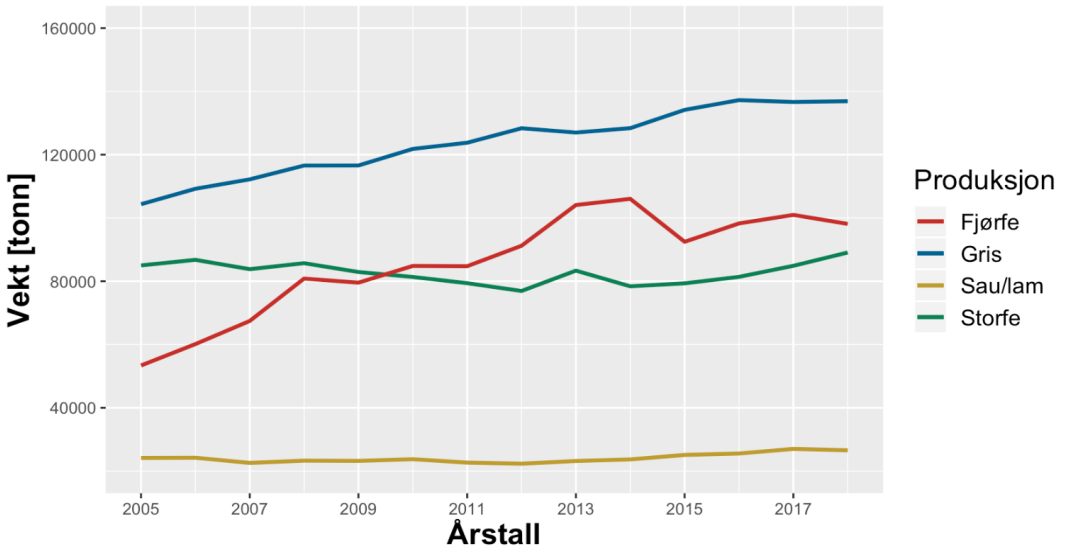
Innledning

1.1 Bakgrunn

Siden begynnelsen av 60 -tallet har verdens produksjon av fjørfekjøtt hatt en kraftig økning. I kjølvannet av finanskrisen i 2008 var økningen størst med over 4 millioner tonn i året frem til 2010 (Nortura, 2019b). Etter 2012 har utviklingen stagnert, og i 2018 var produksjonen 121,6 mill. tonn, noe som utgjorde 36,3 % av verdens totale kjøttproduksjon (Rye, Jenssen & Wenstøp, 2019). Fjørfe er en billig proteinkilde, og sammenlignet med andre kjøttsorter er det lett å tilpasse det tradisjonelle kjøkkenet i ulike land. Kina er i dag den største fjørfeprodusenten i verden. Ifølge amerikanske Food and Agriculture Organization (FAO) sine prognoser vil produksjon av fjørfekjøtt fortsette å øke i årene som kommer og være det kjøttslaget som er i raskest vekst (Nortura, 2019b).

Også i Norge har fjørfeproduksjonen opplevd sterk vekst, fra 38 000 tonn i 2002 til 90 000 tonn i 2018 (Almaas, Bjørkhaug & Almås, 2018; Nortura, 2019b). Fra å være en tilleggsproduksjon har slaktekyllingproduksjon gått over til å bli en hovedproduksjon for mange bønder i Norge. Kyllingkjøtt dominerer den norske fjørfeproduksjonen. Produksjonen er blitt modernisert, effektivisert, og industrialisert, der både vekt per kylling og antall kyllinger per produksjonsenhet har økt (Nortura, 2019b).

Til tross for jevn økning i produksjon har etterspørselen variert, spesielt i periodene 2008-2009 og 2014-2015 registrerte man svekket etterspørsel (Rye et al., 2019). I perioden 2014-2015 skyldtes den svekkede etterspørselen økt medieoppmerksomhet rundt bruken av narasin i fôret. Dette førte til overproduksjon og prisnedgang de påfølgende årene. I dag har etterspørselen tatt seg opp, og er igjen opp mot de samme nivåene som før 2015 (se figur 1.1).



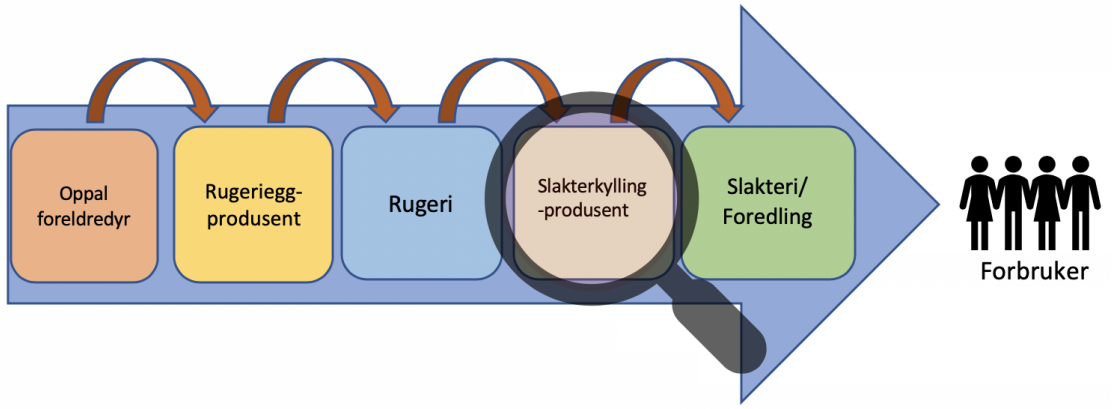
Figur 1.1: Produksjonsutvikling i Norge for ulike kjøttyper i perioden 2005-2018. Produksjonsmengden er oppgitt i antall tonn. Fjørfe inkluderer kyllinger, and, hane, gås og høns. Hvor kyllingproduksjon står for over 90%. Data er hentet fra digitaliseringsdirektoratet (Digitaliseringsdirektoratet, 2020)

Kosthold, mat og måltider har betydning for trivsel og helse, og er en viktig del av hverdagen. De siste tiårene har norske forbrukere fått et mer bevisst forhold til næringsinnhold og kvalitet, og ønsker matvarer som er raske og enkle å tilberede. Som følge av dette har industrien økt satsingen på produktutvikling. I dag kan man finne flere ulike kyllingvarianter i butikkhyllene som alle er tilpasset forbrukernes nye forbruksvaner. Hvor man tidligere solgte hele kyllinger, har man i dag kyllingbryst, -vinger, -lår, og -kjøttdeig, noe som har ført til økt inntjening for slaktekyllingprodusentene. Industrialiseringen, økt automasjon, integrert logistikk, bedre utnyttelse av energi og areal, samt god fôreffektivitet har gjort kyllingprodukter til en rimelig matvare. Kombinasjonen med kvalitet, produkttilpasning og lave priser har ført til at kylling er en ettertraktet matvare i det norske hjem. Klimaavtrykket knyttet til kyllingproduksjon er også mye lavere enn for eksempel storfe (se tabell 1.1). Det kan tenkes at den økte bevisstheten knyttet til klimautfordringene kan føre til enda større etterspørsel i årene som kommer.

Tabell 1.1: Utslippsintensiteter per kg protein i ulike produksjonstyper på verdensbasis (Alvseike et al., 2015).

Dyreslag	Produksjonstype	kg CO_2-ekv/kg protein
Fjørfe	Kjøtt og egg	40.8
Sau	Kjøtt	190.8
Storfe	Kjøtt	342.6
Svin	Kjøtt	51.8

Slaktekyllingprodusenten

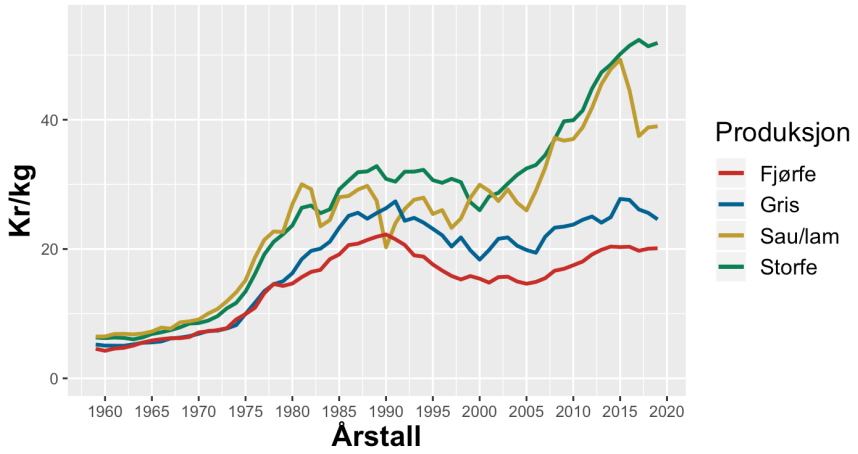


Figur 1.2: De ulike leddene i verdikjeden til kyllingproduksjon, fra oppal av foreldredyr (fra spesialavlede hybrider) til butikkhyllene. Det er oppfôringen av slaktekyllingene som er hovedfokuset i denne oppgaven.

Verdikjeden i kyllingproduksjon består av produksjon og import av rugeegg til foreldredyr, rugerier for foreldredyr, oppal av foreldredyr, rugeggprodusenter for slaktekyllinger, rugerier som produserer daggamle kyllinger, framfôring av slaktekylling, slakterier som slakter kyllingen og videreforedler kyllingslaktet til produkter som omsettes i dagligvarehandelen. Figur 1.2 gir en oversikt over den norske delen av verdikjeden.

Det er oppfôringen av slaktekyllinger som er hovedfokus i denne oppgaven og derfor er det videre lagt mest fokus på dette leddet i verdikjeden.

I Norge er det hovedsakelig fem kyllingraser til konvensjonell produksjon, nisje- og spesialproduksjon (Rye et al., 2019). Fra figur 1.3 kan man se at prisutviklingen de siste 20 årene har vært relativt flat sammenlignet med storfe, sau og lam. Realprisveksten for kylling er negativ i samme periode. For å finne pristoppen må man helt tilbake til 1990, da man fikk 22.25 kroner per kilo kylling (Budsjettnemda, 2019). Etter 1990 sank prisen gradvis, før den igjen steg opp mot 1990-nivå i 2014. En dobling av konsesjonsgrensen til 280 000 dyr i 2014, førte igjen til overproduksjon og prisfall i markedet (se figur 1.3) dette gjorde at flere produsenter gikk konkurs (Bondelag, 2019).



Figur 1.3: Prisutvikling, kr per kilo, for de vanligste kjøttsortene i Norge fra 1960 til 2019. Utviklingen for, spesielt gris og fjørfe, har vært marginal de siste 30 årene, men det har vært en noe mer naturlig utvikling for storfe og sau. Data er hentet fra Norsk institutt for bioøkonomi (Budsjettnemda, 2019).

En krevende markedssituasjon med overproduksjon og sterk konkurranse har de senere årene gjort kyllingproduksjon til en svært marginal bransje.

Ifølge Statistisk Sentralbyrå var det i 2017 kun 13,5 % av alle bønder, fra fjørfe- til storfeprodusenter, som kun næret seg på inntektene fra gårdsbruket (SSB, 2019).

1.2 Problemstilling

Oppgavens hypotese er at noen slaktekyllingprodusenter stabilt har bedre inntjening per kylling, enn andre. Stemmer dette, og i så fall hvilke faktorer er avgjørende for at noen produsenter *alltid* lykkes, mens andre *aldri* gjør det?

1.3 Formål

Hovedformålet med denne oppgaven er å analysere hvilke faktorer som påvirker lønnsomheten til slaktekyllingprodusentene knyttet til Nortura. Lønnsomheten er vurdert etter dekningsbidrag, det vil si salgsinntekter per kylling fratrukket de variable kostnadene.

Oppgavens delmål er å:

1. Vise at det er gjentagende, signifikant forskjell i dekningsbidrag mellom ulike kyllingprodusenter.
2. Imputere manglende data.
3. Belyse relevant teori.

4. Finne, estimere, og analysere mulige sammenhenger mellom de ulike faktorene og kyllingproduksjonens lønnsomhet.

1.4 Begrensninger

Hovedfokus i denne oppgaven er lagt på analysedelen. Selv om det er benyttet betydelig tid på å rydde i datasettet, er det i delkapittel 3.3 derfor kun valgt å inkludere noen metoder for håndtering av manglende data. Det finnes mange flere metoder enn hva som er diskutert og belyst her, men det tilhører en annen oppgave. Dette kapittelet er kun ment som et overblikk over de nevnte metodene. Kvaliteten på de opprinnelige datasettene har ført til økt usikkerhet knyttet til analysens konklusjoner.

Det er kun valgt å tilpasse og analysere lineære modeller. Dette er gjort i et forsøk på å begrense oppgavens omfang.

1.5 Oppgavens struktur

Innholdsmessig er oppgaven bygget opp for å, i tur og orden, besvare de ulike delmålene som tilslutt danner grunnlaget for å trekke en konklusjon rundt oppgavens hovedmål.

Innledningsvis vil det i kapittel 2 bli gitt en introduksjon til fjørfeproduksjon, hvor det er fokusert spesielt på norsk kyllingproduksjon. Det er sett på selve produksjonen og de økonomiske aspektene knyttet til denne. I kapittel 3 skraper man i overflaten og ser på utfordringer knyttet til usikkerheten rundt datainnsamling og manglende data, og strategier for å håndtere disse. Kapittel 4 vil først ta for seg oppbygningen av en generell lineær modell. Deretter vil det bli gått igjennom aspekter, kriterier, og metoder for å evaluere og optimalisere modellene. Før det i kapittel 4.3 -4.9 er presentert ulike statistiske metoder for å tilpasse og estimere modellenes parametere. Tilslutt er det sett på metoder for å vurdere variansen i datasettet. I kapittel 5 vil det bli gitt en kort presentasjon av datamaterialet som er brukt i denne oppgaven. Grunnet mengden finnes det flere detaljer om de ulike datasettene i vedlegg C. Siden datamaterialet eies av Nortura og inneholder konfidensiell informasjon, vil det ikke være mulig for leseren å oppdrive de faktiske dataene uten Norturas samtykke. Kapittel 6 beskriver først et overblikk over oppgavens metodikk, før det i detalj er gått igjennom fremgangsmåtene og metodene som er benyttet i oppgaven. Resultatene fra analysen er presentert i kapittel 7. I kapittel 8 diskuteres resultatene i lys av oppgavens problemstilling før det trekkes konklusjoner som tar sikte på å svare på problemstillingen.

Kapittel 2

Fjørfeproduksjon

Slaktekyllingproduksjon er vertikalt organisert, det vil si før de daggamle kyllingene leveres til slaktekyllingprodusenten, ruges de og klekkes hos en egen rugeriprodusent. De ulike leddene i verdikjeden har dermed et kjøper/selger forhold. Det tar omlag tre uker før eggene klekkes. De nyfødte kyllingene sprayes med vaksiner mot de vanligste sykdommene før de kjøres ut til kyllingprodusentene.

Den moderne slaktekyllingen er fostret innendørs i store haller hvor inneklimaet er nøye justert og monitorert. Der går kyllingene fritt rundt på strø, trespon, og har tilgang til mat og vann. Belysningen tilpasses dyrenes behov, slik at de får passe mengder med lys og mørke (Nortura, 2019a). Her er det store forskjeller mellom de ulike produsentene, men de fleste har et antall lystimer på mellom 16 og 18 timer. Et kull, også kalt innsett, med kyllinger varierer fra produsent til produsent, men ligger normalt opp mot 20 000 kyllinger. Ifølge norske reguleringer kan det maksimalt være 36 kilo kylling per kvadratmeter. Grensene i de fleste andre land i EU er 39 eller 42 kilo (Nortura, 2019a). Noen produksjonstyper har også strengere plassrestriksjoner for å øke dyrevelferden til dyrene. Det finnes i alt 10 ulike produksjonstyper (se vedlegg B). De ulike typene skiller seg blant annet ved hvilket fôr kyllingene spiser, ønsket størrelse, økologisk produksjon, og som nevnt, plassrestriksjoner.

I perioden 1948- 1951 ble det i USA avholdt konkurranser for å fostre opp den største og beste kyllingen i løpet av en 12-ukers periode (Shrader, 1952). I starten førte dette til at man beholdt de beste avlsdyrene og dannet slektstre fra disse. Etterhvert begynte man å krysse ulike raser, som danner grunnlaget for slaktekyllingen slik vi kjenner den i dag. De fleste av Norturas kyllinger stammer i dag fra et fåtall avlsdyr fra Skottland - hybrider spesialoppfostret for å vokse raskt og få fyldige og kjøttrike bryst og lår. Kyllingene når i dag slaktevekt etter 5 - 9 uker, alt etter

hvilken produksjonstype man har (se tabell 2.1). Vanligst er en slaktealder på mellom 29 - 34 dager - da veier kyllingen normalt 1,0 – 1,6 kilo.

Tabell 2.1: Oversikt over de fem vanligste hybridrasene som brukes i kommersiell norsk kyllingproduksjon. Ulike hybrider benyttes til ulike produksjonstyper. Produksjonstid og slaktevekt varierer også mellom de ulike produksjonstypene (Animalia, 2019)

Hybrid	Produksjonstype	Slaktealder	Slaktevekt
Ross 308	Ordinær produksjon	29 - 35 dager	ca. 1.0 - 1.5 kg
Cobb	Ordinær produksjon	30 - 33 dager	ca. 1.0 - 1.3 kg
Sasso	Spesial-, Økologisk produksjon	60- 70 dager	ca. 1.7 - 2.5 kg
Rowan	Spesial-, Økologisk produksjon	40 - 77 dager	ca. 1.8 - 3.0 kg
Hubbard	Spesial-, Økologisk produksjon	50 - 70 dager	ca. 1.6 - 2.1 kg

Ved siden av spesialavlede hybrider, er fôrkvaliteten også tilpasset for å optimalisere vekstraten hos kyllingen. I løpet av oppfôringsperioden får kyllingene normalt tre ulike typer fôr - startfôr, vekstfôr og sluttfôr. Startfôret har høyt innhold av energi og protein, samt at fôrstrukturen er tilpasset dyrets evne til fordøyelse. Vekstfôret inneholder mer energi og mindre protein og er designet for å øke veksten ytterligere. Tidligere fikk produsentene betalt for antall kilo kylling, noe som førte til produksjon av svært store kyllinger. I dag har man en målvekt og man får betalt etter hvor godt man treffer denne (se vedlegg fil:190114avregningsprisliste slaktekylling). En produsents evne til å treffe i området rundt målvekten kan derfor være essensielt for lønnsomheten.

Før kyllingene sendes til slakteri må de samles, eller plukkes, inn i transportbiler fra oppfôringslokalene. Dette gjøres manuelt og koster normalt produsentene en fast rate per kylling (se vedlegg B.3). Noen produsenter har særavtaler, mens andre igjen velger å gjøre dette selv, og har derfor ingen kostnad knyttet til plukking.

2.1 Kvalitetskontroll

Selv om kyllingene i Norge ikke får antibiotika gjøres det et grundig arbeid for at kyllingene skal være friske og av høy kvalitet. Denne prosessen er svært viktig da man håndterer rått kjøtt som potensielt kan være smittebærende. For å hindre sykdom blir alle kyllingene vaksinert hver syvende dag under oppfôringsperioden. Mange produsenter har også egne smittesluser for å hindre å dra med seg potensielle smittekilder når de skal inn og ut av produksjonslokalet. I tillegg har hver produsent en veterinærer som besøker produksjonslokalene og ser til syke dyr. Nortura følger et «alt inn, alt ut»-prinsipp som innebærer at hvert produksjonslokale kun har et innsett om gangen. På denne måten opprettholdes smittebarrieren og minsker risikoen

for smitte fra omgivelsene. Etter at dyra er levert til slakt, blir husdyrrommet og tilførselsrom rengjort og desinfisert. I etterkant har huset også en «karantenetid» før neste besetning setts inn (Nortura, 2019a). På denne måten unngår man eventuell smitte fra et innsett til et annet. I tillegg er det strenge krav til temperatur, fyringskapasitet, luftfuktighet, CO_2-innhold og ventilasjon.

Ved slakt blir kyllingene sjekket for en rekke faktorer som misvekst, farge, wooden breast, Aceties (tarmsykdom), og en rekke andre feil før de eventuelt blir godkjent for salg. Dersom kyllingen ikke skulle passere kontrollen blir den kassert, og kyllingprodusenten blir trukket en sum fra sluttoppgjøret.

2.2 Dyrevelferd

Ved siden av reguleringene knyttet til plassrestriksjoner, innførte Nortura i 2018 tiltak for at kyllingen skal få mer variasjon og økt velferd. Trivselstiltak som flisballer, torvstrø og små aktivitetshus de kan klatre på bidrar til mer lek, aktivitet, hvile og naturlig adferd (Animalia, 2017). Selv om reguleringer tillater en maksimal kyllingtetthet på 36 kg/m^2 må likevel flere krav til dyrevelferd opprettholdes. Tråputepoeng er en skala som er med på å avgjør hvor høy kyllingtetthet produsenten til enhver tid er tillatt ha i huset. Tråputene under kyllingenes føtter er en dyrevelferdsindikator – et mål på hvor godt miljø produsenten har lykkes å skape i kyllinghuset. Høy fuktighet og/eller ammoniakk i strøbedet, kan gi skader under kyllingens føtter (Animalia, 2017). Tråputeprogrammet gjennomføres på alle innsett og innebærer at minst 100 føtter bedømmes. Vurderingen klassifiseres i 3 ulike klasser:

- **Klasse 0: Uten anmerkninger**
- **Klasse 1: Lett skade**
- **Klasse 2: Grov skade**

Tråputepoengene beregnes ved at antall Klasse 0 -føtter multipliseres med 0, Klasse 1 -føtter med 1 og Klasse 2 -føtter med 2 for så å summeres sammen. Dermed får hvert innsett en score i intervallet 0-200. Skalaen skaleres dersom flere enn 100 føtter undersøkes. Poengsummene deles inn i 3 nivåer:

Tabell 2.2: Tråputepoengskalaen. Tråputepoengsystemet er i tråd med EUs rådsdirektiv 2007/43/EF, også kalt "slaktekyllingdirektivet", et tiltak for å sikre kyllingenes dyrevelferd. (Animalia, 2017)

Nivå	Bedømming	Poengsum
A	Tilfredsstillende	0-80
B	Ikke tilfredsstillende	81-120
C	Uakseptabelt	121-200

Ved dårlige resultater, må kyllingprodusenten redusere kyllingtettheten ved neste innsett. Produsenter som har fått restriksjoner på grunn av dårlig tråputescore, må gjentatte ganger levere gode resultater for å få lov å øke tettheten igjen.

2.3 Økonomi

2.3.1 Produksjonsinntekter

Produksjonsinntektene fra kyllingproduksjonen kommer hovedsakelig i form av salgsinntekter. I tillegg kommer distriktstilskudd for kyllingkjøtt for deler av landet, fra Hordaland og nordover (Holien, 2009). Videre kan man søke om avløsertilskudd ved ferie og fritid. Tilskuddet beregnes på bakgrunn av det dyretallet og areal hver enkelt produsent disponerer (Lovdata, 2015). Tilskuddet kan dog ikke overstige produsentens faktiske utgifter knyttet til avløsning. Tabell 2.3 viser en oversikt over gjennomsnittlig salgsinntekt per kylling for de vanligste produksjonstypene hos Nortura. Man kan tydelig se store forskjeller i salgsinntekt per kylling for de ulike produksjonene. Økologisk kylling har et snitt på rundt 90 kroner, mens de resterende produksjonstypene varierer fra rundt 22 - 42 kroner per kylling.

Tabell 2.3: Gjennomsnittlig salgsinntekt per kylling for ulike produksjonstyper og antall innsett av den aktuelle produksjonstypen i perioden 01.01.18 – 01.12.19. Merk at utgiftene ikke er tatt med i beregningene. Tallene tar utgangspunkt i data fra Norturas database, fjørfekontrollen.

Produksjonstype	Antall Innsett	Kr/kylling
Foredlingskylling	1986	25.9
Grillkylling	910	22,7
Kyllinggården	553	27.0
McDonalds-kylling	348	27.1
Liveche	81	41.6
Hubbard	57	36.7
Økologisk kylling	11	89.6

2.3.2 Variable kostnader

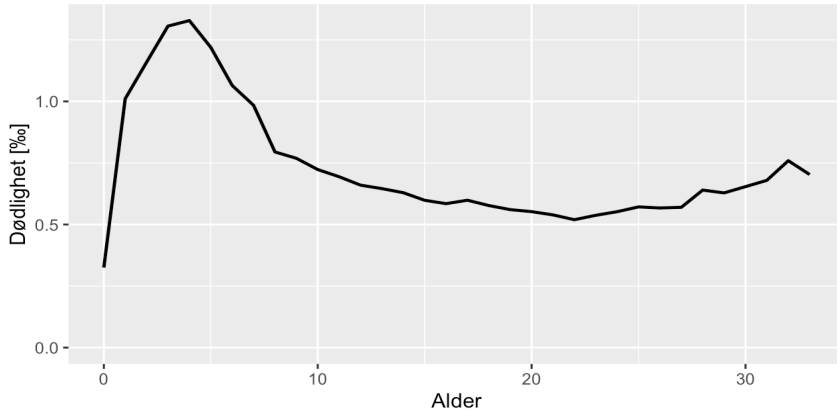
De største variable kostnadene knyttet til kyllingproduksjonen er kostnader til innkjøp av daggamle kyllinger og kraftfôr. Utgifter til kraftfôr utgjør omlag 50% av salgsinntektene uavhengig av produksjonstype. Ulike fôringsrutiner og svinn kan, i tillegg til pris, påvirke og skape store forskjeller mellom produsentene. En daggammel kylling koster omlag 5.4 kroner. I tillegg kommer kostnader til oppvarming, strø, forsikring, plukking, desinfeksjon, med mer. Til sammen regner Nortura med at andre variable kostnader utgjør ca. 3.2 kroner per kylling, ekskludert kostnader til fôr og innkjøp av daggamle kyllinger. En fullstendig oversikt over kostnadene som inngår i denne beregningen finnes i vedlegg B.2.

2.3.3 Dødelighet

De første levedagene, og da spesielt under transport, er kyllingene ekstra utsatt for bakterier og sykdommer som bidrar til økt dødelighet den første uken av innsettet. Det kan også være en utfordring å få kyllingene til å spise og drikke i starten, noe som gjør at noen kyllinger sulter/tørster ihjel. I et forsøk på å stimulere kyllingene til å spise legges normalt fôret på papir frem til de venner seg til å spise fra det ordinære fôringssystemet. Mange produsenter forsøker også å redusere fôrkostnadene ved å tidlig ta livet av de kyllingene som man ser ikke kommer til å overleve, grunnet misdannelser eller lignende. Dødeligheten i produksjonsperioden er i gjennomsnitt på 2.89% (Kjos et al., 2019).

Det å sende uskikkede kyllinger til slakt medfører som nevnt ekstra kostnader for produsenten. Ekstra fokus og årvåkenhet inn mot slaktedato bidrar derfor til at dødeligheten øker noe mot slutten av innsettet (se figur 2.1). Kyllinger som likevel sendes til slakt, men som må kasseres koster produsenten 4 kr per stykk. Det kan derfor være avgjørende for resultatet hvor dyktig produsenten er til å fjerne uegnede kyllinger, tidlig. Kassasjonsprosenten hos slakteri var i 2018, 2.60 % (Kjos et al., 2019).

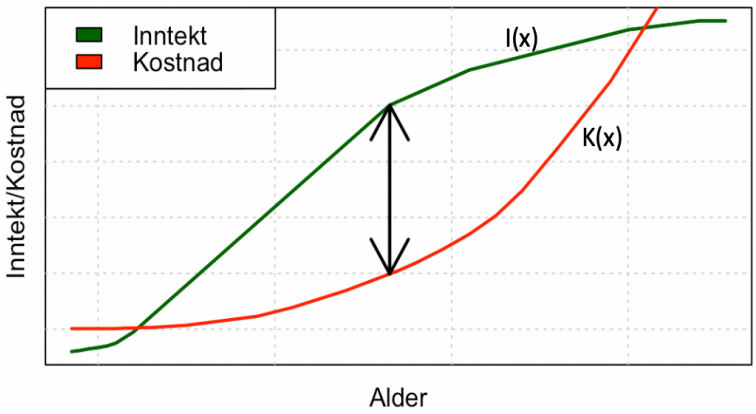
Dødelighet knyttet til transport fra kyllingprodusent til slakteri har blitt halvert fra 0.16% i 2010 til 0.08% i 2018 (Kjos et al., 2019) og utgjør en svært liten del av frafallsbilde.



Figur 2.1: Antall døde kyllinger som promille av totalt antall kyllinger i innsett vs. alder. Illustrerer hvordan gjennomsnittlig dødelighet blant kyllingene varierer per dag gjennom innsettet. Data er fra Norturas kyllingprodusenter i perioden 01.01.18 - 01.12.19.

2.3.4 Bruttofortjeneste

Slaktekyllinger slaktes i ung alder, normalt etter noen å tretti dager. Dette fordi de raskt oppnår ønsket størrelse og kjøttfylde sammenlignet med andre dyreslag. De fleste kyllinghybridene er avlet frem for å utnytte fôrressursene mest mulig effektivt, slik at de gir mest mulig kjøtt for minst mulig fôr. Etterhvert som kyllingene blir større, øker også fôrinntaket. Siden fôrforbruket utgjør omlag 50% av salgsinntekten ønsker man å finne den optimale alderen/størrelsen på dyrene før fôrkostnadene spiser en større del av inntektene. Gitt at I(x) og K(x) er henholdsvis inntjening og akkumulert fôrkostnad som funksjon av alder, x, ønsker man å oppfôre kyllingene til den alderen som maksimerer avstanden mellom kost- og inntektskurven for å maksimere profitten, P (se figur 2.2).



Figur 2.2: Illustrerer hvordan forventet inntekt og akkumulert fôrkostnad utvikler seg som funksjon av alder/produksjonslengde, x. For å maksimere profitten ønsker man å slakte kyllingene når avstanden mellom de to kurvene er størst, her markert ved to piler, og er normalt når kyllingene er 30-34 dager gamle.

Dekningsbidrag er produksjonsinntekter fratrukket de variable kostnadene. Nortura forholder seg til to ulike dekningsbidrag (se tabell 2.4) i deres lønnsomhetsberegninger. Dekningsbidrag 2 gir et mer helhetlig bilde over lønnsomheten til kyllingproduksjonen, mens dekningsbidrag 1 gjør det lettere å sammenligne produsenter. Årsaken til at dekningsbidrag 1 er lettere å sammenligne er at man ikke inkluderer plukkekostnader, hvor enkelte har spesialavtaler. De andre variable kostnadene kan også være vanskelig å holde styr på, og vil variere utover standardsatsen satt av Nortura.

Som man kan se av eksempelet tabell 2.5 er kyllingproduksjon et marginalt virke, hvor små utslag kan være avgjørende for lønnsomheten. Verdiene benyttet i eksempelet er gjennomsnittsverdier fra data i perioden 01.01.2018 - 01.12.2019.

Tabell 2.4: Oppsett over hvordan Nortura beregner de to dekningsbidragene. DB1 inkluderer kun fôrkostnader og anskaffelseskostnaden av daggamle kyllinger, mens DB2 også inkluderer kostnaden knyttet til plukking av kyllinger fra produksjonslokalet til transportbil og andre variable kostnader. Oversikt over andre variable kostnader finnes i vedlegg B.2

Dekningsbidrag 1 (DB1)	Dekningsbidrag 2 (DB2)
Salgsinntekter	Salgsinntekter
– Fôrkostnader	– Fôrkostnader
– Kostnad daggamle	– Kostnad daggamle
	– Plukkekostnader
	– Andre variable kostnader
Sum	**Sum**

Tabell 2.5: Eksempel: Beregning av dekningsbidrag 2 for foredlingskyllingproduksjon. Verdiene er gjennomsnittsverdier fra perioden 01.01.2018 - 01.12.2019.

Dekningsbidrag 2 (DB2)	Kroner/kylling [kr]
Salgsinntekt	25.9
– Fôrkostnader	– 12.95
– Kostnad daggamle	– 5.35
– Plukkekostnader	– 0.67*
– Andre variable kostnader	– 3.19*
Sum	**3.74**

* Verdier hentet fra vedlegg /Users/mortenkvamsdal/Documents/Master/190114.avregningsprisliste slaktekylling.xlsx og /Users/mortenkvamsdal/Documents/Master/Oversikt variabler kostnader per m2.xlsx.

Kapittel 3

Stukturering og analyse av ustrukturert data

Strukturering og vasking av data er en viktig prosess for å klargjøre ustrukturert datamateriell for analyse. Gjennom å fjerne - feil, ufullstendige, uinteressante, eller dobbeltføringer av data vil man bedre kunne modellere de virkelige sammenhengene mellom forklaringsvariabler og respons. Bedre datagrunnlag vil føre til bedre prediksjonsmodeller. I følge New York times bruker en dataanalytiker mellom 50 og 80 prosent av tiden på vasking av data (Lohr, 2014). Spesielt ved registreringsdata, hvor et populasjonsutvalg selv registrerer dataene, kan dette være en utfordrende prosess. Årsaken til at registreringsdata kan være ekstra utfordrende skyldes eksempelvis inkonsistent samling og føring av data.

3.1 Vasking av data

Det finnes flere ulike måter å håndtere ustrukturert data på, likevel er det noen retningslinjer som går igjen:

1. **Uønskede verdier.** Det første steget er å fjerne uønskede verdier. Uønskede verdier kan være dobbeltføringer, umulige verdier, eller irrelevante observasjoner som åpenbart ikke er relevant for sammenhengen man ønsker å modellere.

2. **Avviksverdier.** En avviksverdi er en observasjon som er mye lavere eller høyere enn forventet. I søken etter det ekstraordinære skal man være forsiktig med å ekskludere slike ekstremverdier, kanskje er det akkurat disse verdiene som forklarer det man ønsker å modellere. Hvis alle modellantagelsene (se kapittel 4.1) holder, vil omlag 68% av observasjonenes standardiserte residualer

ligge mellom ±1, 95% mellom ±2, og 99.7% mellom ±3 (Dean, Voss & Draguljic, 2017). Det er med andre ord svært usannsynlig å finne observasjoner med standardisert residual over/under ±3. Likevel bør man generelt ha god grunn for å kunne definere disse verdiene som avviksverdier, for eksempel hvis man mistenker at observasjonen er umulig.

3. **Skrivefeil.** Ofte, spesielt i registreringsdata kan det forekomme skrivefeil. Skrivefeil kan være vanskelige å oppdage, men bør undersøkes og forsøkes å rettes opp i.

I tillegg til de overnevnte punktene kan store mengder manglende data by på utfordringer som må håndteres før man kan starte på selve analysen.

3.2 Manglende data

Manglende data kan defineres som verdier som ikke er lagret eller registrert for en variabel som er interessant for analyse. Problemet med manglende data er svært vanlig i nesten all forskning og kan ha betydelig effekt på resultatet. Siden de fleste statistiske modeller ikke aksepterer manglende data, må disse rettes opp i før man kan starte analysen. Det er flere utfordringer knyttet til manglende data. For det første kan manglende data føre til skjevhet i estimatene av modellens parametere. For det andre, reduseres sannsynligheten for å finne støtte til hypotesene man ønsker å teste, selv om de i realiteten stemmer. I tillegg kan testutvalget reduseres, ved at man må slette data grunnet ufullstendig rader/kolonner.

Det kan være flere grunner til at man har manglende verdier. Normalt deles manglende verdier inn i ulike kategorier på bakgrunn av årsaken til deres eksistens, eller mangelen på sådan. Dette har betydning for modellens validitet og skjevhet, og hvordan man bør behandle dataene videre.

3.2.1 Missing completely at random

Dersom sannsynligheten for manglende verdier ikke er relatert til verken responsvariabelen, eller de andre prediksjonsvariablene, defineres det som Missing completely at random (MCAR) (Graham, 2012). MCAR er en ideell antagelse da den ikke fører til skjevhet i analysen, dessverre er det ofte urealistisk.

Eksempel: Man ønsker å kartlegge hvilke faktorer som avgjør hvor mye arbeidstakere har i lønn. Faktorene kan være bransje, stillingstype, alder, kjønn osv. Dersom yngre mennesker oftere valgte å avstå fra å rapportere sin lønn sammenlignet med de eldre ville MCAR antagelsen være brutt, og modellen ville vært forventningsskjev.

Eldre mennesker, med mer erfaring, tjener normalt mer enn nyutdannede. Om man delte datasettet inn i ulike aldersgrupper og man likevel ikke kan finne en signifikant forskjell i den gjennomsnittlige inntekten til de forskjellige gruppene kan man konkludere med at MCAR antagelsen likevel holder.

3.2.2 Missing at random

Missing at random (MAR) er ofte en mer realistisk antagelse. MAR er sannsynligheten for at en verdi mangler, ikke kan relateres til responsvariabelen etter at man har inkludert de andre prediksjonsvariablene (Graham, 2012).

Eksempel: MAR antagelsen vil holde selv om sannsynligheten for at en yngre arbeidstaker ikke har rapportert sin lønn er høyere, men innen arbeidstakerens aldersgruppe er det ingen sammenheng mellom innrapportering av lønn, og lønn. For at antagelsen skal holde må da alder inkluderes som en parameter i modellen.

3.2.3 Not missing at random

Ved Not missing at random (NMAR) er manglende data avhengig av responsvariabelen (Graham, 2012). Med andre ord er det et mønster knytte til de manglende dataene.

Eksempel: Arbeidstakere med høy lønn velger å ikke rapportere sin lønn.

3.2.4 Imputering eller sletting av data

Det finnes hovedsakelig to hovedprinsipp for håndtering av manglende data:

1. **Slette** manglende data. Dette er en sub-optimal løsning da man ender opp med å slette informasjon som kan vise seg å være signifikant.
2. **Imputering** vil si å erstatte den manglende verdien med estimerte verdier basert på de andre registrerte verdiene, eller annen bakgrunnsinformasjon om problemet. Dette vil også være sub-optimalt da man ikke tilfører datasettet noen virkelige verdier, men forsterker mønstre som allerede ligger i datasettet.

Begge fremgangsmåtene vil uansett ikke være optimale, da begge fører til skjevhet i datasettet og forringer analyseresultatet sammenlignet med et fullstendig datasett.

En alternativ løsning, da det også er informasjon i manglende data, er å lage en ny klasse for manglende verdier. Dette er mest aktuelt for kategoriske variabler.

Dersom enten MCAR eller MAR antagelsen holder kan man forsvare å fjerne de manglende observasjonene, men dersom man har NMAR kan fjerning av de

manglende verdiene påføre modellen ytterligere skjevhet. Manglende verdier bør ved NMAR inkluderes som en del av modellen (Lee & Jr., 2011). Nytteverdien ved imputering er størst ved MCAR og MAR, og andeler manglende data opp mot 30 % vil, avhengig av datasett og imputeringsmetode, bety liten økning i prediksjonsfeilen til modellen (Kokla, Virtanen, Kolehmainen, Paananen & Hanhineva, 2019). Prediksjonsfeil defineres normalt som summen av den kvadrerte differansen mellom modellens estimerte verdi, og virkelig verdi (se kapittel 4 for mer om prediksjonsfeil). Ved NMAR vil man ved imputering derimot forvente signifikant økning i prediksjonsfeil sammenlignet med MCAR og MAR, da det vil påføre datasettet økt skjevhet. Hvor stor andel manglende data man bør tillate er altså avhengig av årsaken til de manglende datene. Det er viktig å bemerke at imputering ikke nødvendigvis sikrer et bedre analyseresultat.

3.3 Metoder for å håndtere manglende data

Det finnes mange metoder og fremgangsmåter for å håndtere manglende data. I dette delkapittelet gis det et overblikk over noen av de mer konvensjonelle metodene til noen utvalgte få som er litt mer komplekse.

3.3.1 Konvensjonell metoder

Sletting av rader. Dersom det mangler verdier i en eller flere av variablene til en observasjon, slettes hele observasjonen. Ulempen er at man kan ende opp med å ekskludere store deler at datamaterialet.

Sletting av kolonner. Dersom noen forklaringsvariabler har store mangler, eller ikke er ført på en konsistent måte, kan variabelen slettes. Ulempen kan være at man risikerer å fjerne en viktig variabel. Fordelen vil være at man beholder en større del av observasjonene.

Imputering ved bruk av gjennomsnitt. Man imputerer manglende verdier ved hjelp av gjennomsnittsverdier fra de allerede eksisterende verdiene. Ulempen er at det fører til forskyvning i estimatene av varians og kovariansen.

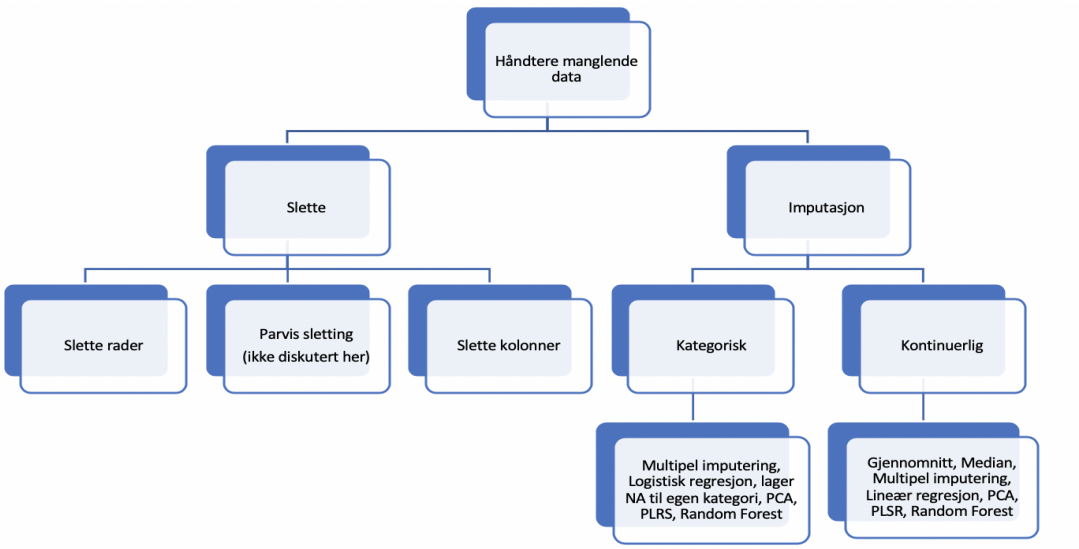
Imputering ved hjelp av lineær regresjon. Man benytter de eksisterende dataene til å lage en modell. Videre benytte man modellen, med de estimerte parameterne, til å predikere den manglende verdien. Dersom datasettet har mange manglende verdier bør man ikke bruke denne metoden (Kang, 2013).

3.3.2 Multippel imputering

I 1987 utviklet D.B.Rubin en metode for å håndtere ekstra støy i dataene som følge av imputering (Rubin, 2004). Metoden baser seg på å ta gjennomsnittsverdiene fra flere imputerte datasett for så å danne ett datasett. Metoden følger tre steg (Allison, 2000):

1. **Imputering** - På samme måte som ved konvensjonell imputering, imputeres de manglende verdiene. Forskjellen er at de imputerte dataene trekkes tilfeldig fra verdienes sannsynlighetsfordeling. Dette fører til at man får ulike resultat hver gang. Dette gjøres m-antall ganger, som vil si at man får m ulike imputerte datasett. Man kan også benytte bootstraping (se delkapittel 4.4.1) for å danne ulike imputerte datasett.
2. **Analyse** - Hvert av de imputerte datasettene analyseres slik at man har m ulike analyser.
3. **Resultat** - resultatene av alle analysene kombineres til et datasett.

Siden det vil være variasjon i de imputerte dataene vil det også være variasjon i parameternes estimater, noe som fører til realistiske estimater av varians og p -verdier (Allison, 2000). Til multippel imputering kan mange ulike statistiske metoder benyttes. Mer om noen av metodene i Kapittel 4. En oversikt over ulike strategier for å håndtere manglende data er oppsummert i Figur 3.1.



Figur 3.1: Oversikt over strategier for å håndtere manglende data. Noen statistiske metoder, som PCA, PLSR og Random Forest kan håndtere blandede datatyper (både kategoriske og kontinuerlige variabler).

Kapittel 4

Statistiske metoder

I vitenskapelige studier har man sjeldent data fra hele populasjonen man ønsker å analysere, derfor tar man ofte utgangspunkt i et *utvalg* av en større populasjon. Man ønsker da å bruke informasjonen fra utvalget til å trekke konklusjoner om hele populasjonen. Siden man kun analyserer et mindre utvalg og ikke hele populasjonen er konklusjonene beheftet med usikkerhet. Statistiske metoder tar sikte på å beskrive disse usikkerhetene ved å legge til grunn en statistisk modell for populasjonen. En modell kan inneholde én eller flere parametere som estimeres ved hjelp av dataene samlet i utvalget. Gjennom hypotesetesting kan modellparameterne, med deres tilhørende usikkerhet, benyttes til å trekke konklusjoner for hele populasjonen.

Før man går videre inn på de ulike statistiske metodene som er benyttet for å tilpasse modellene, vil det først introduseres en generell modell, samt noen statistiske begreper og prinsipper som er relevante for forståelsen og evalueringen av modellene.

4.1 Modellbygging

En statistisk modell er en forenklet representasjon av en observert del av virkeligheten. Modellen består av tilfeldige variabler og parametre:

Tilfeldig variabler

- Y: Responsvektoren som inneholder den observerte responsen til de n observasjonene. Y er en avhengig variabel, og er variabelen som påvirkes av de uavhengige variablene. Y kalles ofte for respons eller utfallsvariabel.
- X: En $n \times p$ matrise av forklaringsvariabler, hvor n er antall observasjoner og p er antall variabler. X består av uavhengige variabler, og er de man er interessert i virkningen av.

- ϵ: Et mål på avstanden mellom de observerte verdiene og forventningsverdien til modellen.

Parametere

- β_0: Er forventet respons dersom alle forklaringsvariablene er null.
- β_i: Regresjonskoeffisient, er effekten variabel i har på responsen. Det vil si gjennomsnittlig endring i responsvariabel når forklaringsvariabel i øker med 1 i verdi.
- σ^2: Variansen til ϵ.

Responsvariabelen Y kan observeres, mens parameterne estimeres på bakgrunn av de observerte dataene. Denne oppgaven vil stort sett konsentrere seg om lineære multivariable-univariat-modeller, det vil si multiple forklaringsvariabler som er lineære i sine parametere, og med én responsvariabel. Den generelle modellen vil da være på formen:

$$\begin{aligned} Y &= \beta_0 + f(X) + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \\ &= \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon_j \end{aligned} \tag{4.1}$$

hvor j representerer den j-te variabelen.

Ved lineær regresjon gjør man følgende antakelser:

1. ϵ er normalfordelt med forventningsverdi 0 og konstant varians σ^2 for alle verdier av X ($\epsilon \sim N(0, \sigma^2)$).
2. Differansen mellom modellenes predikerte verdier og de observerte verdiene, $y_i - \hat{y}_i = e_i$, også kalt residualene, er tilfeldig og uavhengige. Det vil si at de er tilnærmet normalfordelt rundt sin forventing, og ikke korrelert.
3. Lineær sammenheng mellom respons og forklaringsvariabler.

I de fleste praktiske tilfeller vil σ^2 være ukjent, og må estimeres ved hjelp av dataene fra utvalget. Det beste forventningsrette estimatet av σ^2 er s^2 og beregnes ved å dele Sum og squares estimate of error (SSE),

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.2}$$

på antall frihetsgrader. Antall frihetsgrader i dette tilfelle vil være antall observasjoner minus antall estimerte parametere. s^2 kalles ofte for Mean squared error (MSE) og s for Root mean squared error (RMSE).

Dersom man tar utgangspunkt i at antagelsene for lineær regresjon holder, vil prediksjonsverdien til responsen være gitt ved,

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{p} X_i \hat{\beta}_i, \tag{4.3}$$

hvor $\hat{\beta}_0$ og $\hat{\beta}_i$ vil være de estimerte verdiene for henholdsvis skjæringspunkt og regresjonskoeffisientene. Mer om hvordan estimere β-verdiene i delkapittel 4.3.

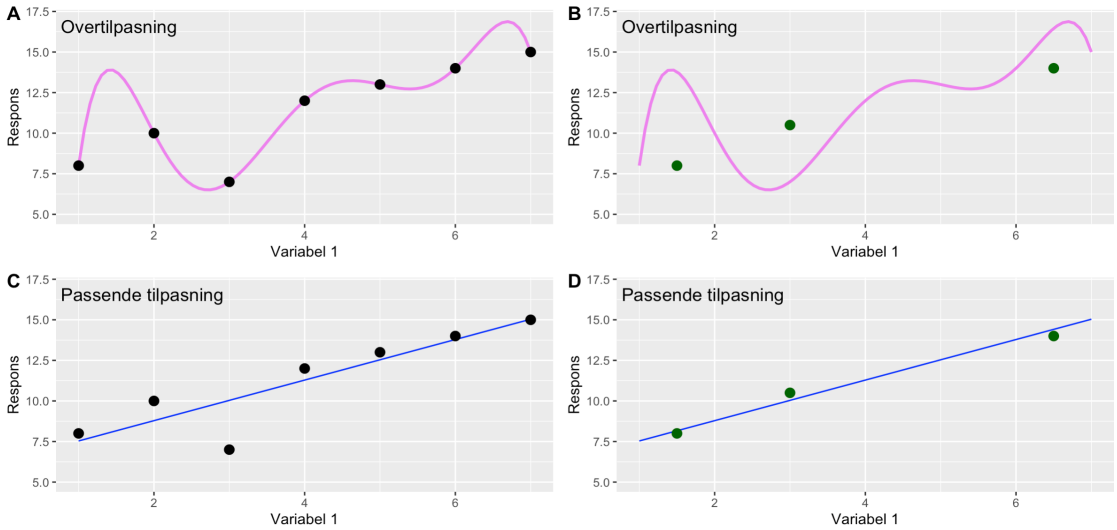
I tilfeller hvor sammenhengen mellom $\hat{Y}$ og en vilkårlig forklaringsvariabel er avhengig av verdien til de andre forklaringsvariablene, vil man inkludere interaksjonsledd $(X_i X_j)$ til modellen. I andre tilfeller hvor sammenhengen er krummet kan man inkludere ledd med høyere ordens polynomer. Ligning 4.4 er et eksempel på en modell hvor man både har interaksjon og krumning.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_1 X_2 + \hat{\beta}_4 X_1^2 + \hat{\beta}_5 X_2^2 \tag{4.4}$$

Ved å inkludere flere variabler eller kombinasjoner av disse vil modellen bedre kunne passe de observerte dataene. En mulig fallgruve er likevel at man modellerer tilfeldig støy og ikke reelle sammenhenger mellom forklaringsvariabel og responsvariabel. Man kan altså risikere å inkludere for mange regresjonskoeffisienter og lage modellen *for* kompleks.

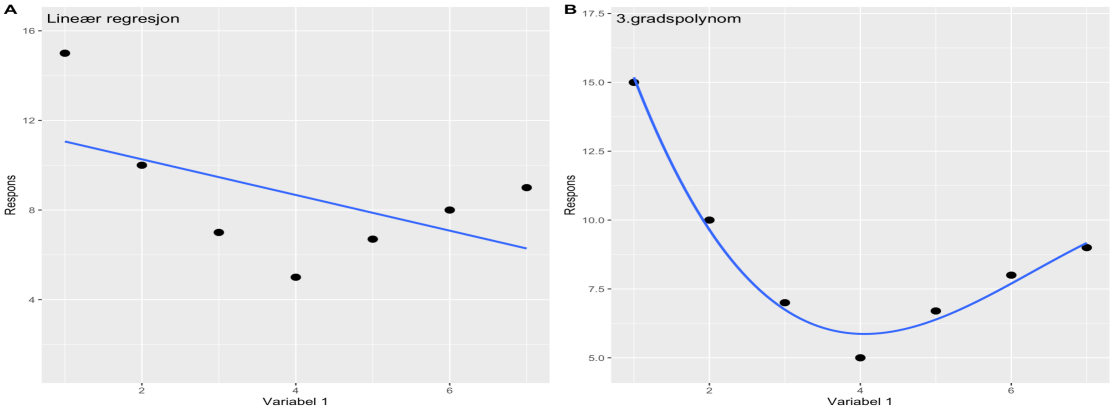
4.1.1 Over- og undertilpasning

Tilfeller hvor modellen er for kompleks og lærer for mye av detaljene i et gitt datasettet, kalles overtilpasning. Modellen vil da prestere veldig godt på datasettet, men dårligere ved introduksjon av nye data. Ved overtilpasning modellerer man noe av den tilfeldige variasjonen som ligger naturlig i utvalget, og modellen bør ikke generaliseres til å modellere hele populasjonen. Ikke-lineære modeller har mer fleksibilitet og er mer utsatt for overtilpasning. Et eksempel på dette kan sees i Figur 4.1.



Figur 4.1: Plott A og C illustrerer to ulike modeller som er tilpasset samme treningsdata. Modell A ser tilsynelatende ut til å fungere svært godt, men i plott B ser man at det ikke er tilfelle når nye observasjoner (grønne punkt) også er inkludert. Modellen er overtilpasset.

I motsatt fall har man det man kaller undertilpasning. Undertilpasning vil si at modellen verken kan forklare relasjonene i de observerte dataene eller ved nye data. Figur 4.2 A illustrerer et eksempel på undertilpasning.



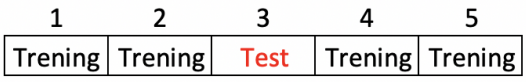
Figur 4.2: Plottet til venstre viser hvordan en modell ikke tar nok hensyn til treningsdataene og i liten grad klarer å forklare sammenhengen mellom forklaringsvariabel og respons (undertilpasning). Plottet til høyre viser en modell som mye bedre forklarer sammenhengen mellom variabel og respons.

4.1.2 Validering

Både over- og undertilpasning kan føre til at modellen ikke evner å forklare den virkelige sammenhengen i populasjonen. Undertilpasning er lett å oppdage, da modellen presterer dårlig, også på de observerte dataene. Overtilpasning derimot, kan prestere svært godt på de observerte dataene og derfor være vanskelig å oppdage (se figur 4.1 A). For å sikre at den tilpassede modellen kan generaliseres til hele populasjonen kan man benytte kryssvalidering, eller dele datasettet i trenings- og testsett, eller begge deler.

Oppdeling i trenings- og testsett
For å teste modellens prestasjon mot nye data setter man til side en del av det observerte datasettet. Dette gjøres for å sikre at modellen har lært, og kan modellere, den virkelige sammenhengen mellom forklaringsvariablene og responsen, ikke bare sammenhengen mellom dataene i datasettet. For å bruke en analogi; en elev skal lære multiplikasjon. Læreren bruker den lille gangetabellen for å undervise eleven, før eleven i etterkant skal testes i multiplikasjon. Læreren tenker å teste eleven i 7×6 og 4×8 og fjerner derfor disse eksemplene fra undervisningen. Når eleven da skal testes kan han kun svaret dersom han virkelig har lært seg prinsippet ved multiplikasjon. Dersom læreren hadde inkludert hele den lille gangetabellen i undervisningen, kunne eleven potensielt ha pugget denne, uten å kunne forklare resultatene. Dersom man har tilstrekkelig med observasjoner, deles normalt datasettet i to - ett treningssett hvor modellen tilpasses, og ett testsett hvor man tester den tilpassede modellen. Treningssettet er normalt større enn testsettet. I tilfeller hvor man ønsker å teste ulike modeller opp mot hverandre, kan datasettet også deles i tre deler. Man benytter da én del til å trene modellen, én del til å sammenligne prestasjonen til de ulike modellene, og én del til å teste den valgte/beste modellen.

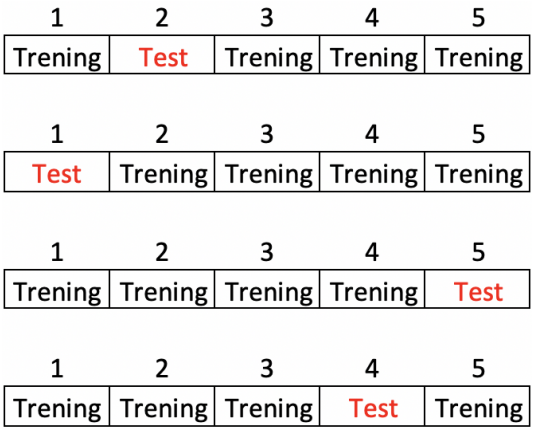
Kryssvalidering
Ofte har man ikke tilstrekkelig data til å sette av en egen del til å teste modellen, da kan man i stedet bruke kryssvalidering. Ved kryssvalidering deler man også datasettet i trenings- og testsett. Dette gjøres K-antall ganger med ulikt trenings- og testsett for hver gang. For eksempel, for K = 5, kan inndelingen se slik ut:



I dette eksempelet benytter man delsett 1, 2, 4, og 5 til å estimere parameterne i modellen, og tester modellen mot delsett 3 for å estimere modellens prediksjonsfeil. Som et mål på prediksjonsfeilen kan man benytte Mean squared error of prediction(MSEP),

$$MSEP = \frac{1}{n_k} \sum_{i \in F_k} (y_i - \hat{f}^{-k}(x_i))^2, \tag{4.5}$$

hvor n_k er antall observasjoner i testsett k, F_k er testsett k, $\hat{f}^{-k}(x)$ er modellen som er tilpasset treningssett k, x_i er forklaringsvariablenes i-te verdier fra testsettet, og y_i er de observerte responsverdiene i testsettet. Dette er en iterativ prosess hvor man tilpasser K nye modeller og tester mot K testsett:



Ved å trene og teste modellen K ganger vil man, selv med færre antall observasjoner, bedre kunne estimere modellens virkelige prediksjonsfeil sammenlignet med hvis man kun tilpasset én modell. Via kryssvalidering er estimatet av prediksjonsfeilen gitt ved (Friedman, Hastie & Tibshirani, 2001),

$$CV(\hat{f}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in F_k} (y_i - \hat{f}^{-k}(x_i))^2. \tag{4.6}$$

Hvor K er antall iterasjoner. Det er vanlig å velge $K = 5$ eller $K = 10$. Kryssvalidering brukes ofte også til å bestemme kompleksitetsparametere til modeller for å finne den optimale modellkompleksiteten som minimerer prediksjonsfeilen. Mer om kompleksitetsparametere i delkapittel 4.3.

4.1.3 Kvalitetsmål

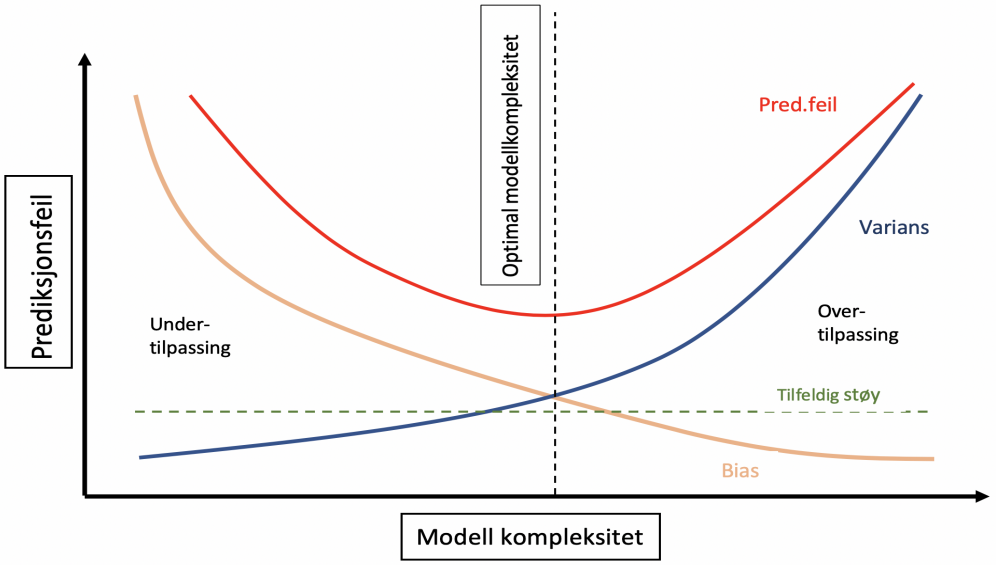
Hvis vi antar den sanne modellen $Y = f(X) + \epsilon$, hvor $E(\epsilon) = 0$ og $Var(\epsilon) = \sigma_\epsilon^2$ kan en dele prediksjonsfeilen i tre deler,

$$MSEP = Bias^2 \ + \ Varians \ + \ tilfeldig \ sty \tag{4.7}$$

Tilfeldig støy, σ^2, er et uttrykk for måleusikkerheten til responsen, og kan ikke reduseres. Bias, eller forventningsskjevhet, er forventet forskjell mellom populasjonens virkelige parametere, β, og modellens estimerte parametere, $\hat{\beta}$. Det vil si et mål på den estimerte modellens systematiske feil.

$$Bias(\hat{\beta}) = E(\hat{\beta}) - \beta \tag{4.8}$$

Variansen derimot er et spredningsmål, eller usikkerheten, til de estimerte parameterne til modellen. Bias og varians er avhengig av modellens kompleksitet. Man ønsker å finne den optimale modellkompleksiteten for å balansere bias og varians slik at prediksjonsfeilen minimeres (se figur 4.3). For å finne den optimale modellkompleksiteten kan man for flere statistiske metoder (PCR, PLS, Lasso, Random Forest, mm.) benytte kryssvalidering til å bestemme den kompleksitetsparameteren som minimerer prediksjonsfeilen. Man kryssvaliderer da innad i treningssettet.



Figur 4.3: Illustrasjon av hvordan prediksjonsfeil (rød), varians (blå) og forventningskjevheten (beige) varierer som funksjon av modellkompleksitet. Den tilfeldige støyen (grønn) vil være konstant uavhenging av modellkompleksitet. En modell kan bli for kompleks og forsøke å modellere støy noe som kun fører til økt prediksjonsfeil, eller for simpel og ikke inkludere relevante parametere. Man ønsker derfor å finne den optimale kompleksiteten som minimerer prediksjonsfeilen.

4.2 Modellkriterier

Det finnes ulike metoder og kriterier for å evaluere modellen, identifisere signifikante forklaringsvariabler, og finne den optimale modellkompleksiteten. Med signifikante forklaringsvariabler menes det variabler hvor man med høy sikkerhet kan påvise sammenheng mellom forklaringsvariabel og respons. De ulike metodene skiller seg ved terskelen for å inkludere en gitt forklaringsvariabel i modellen. Ved siden av MSEP som er beskrevet tidligere, vil det i dette delkapittelet presenteres noen av de vanligste modellevalueringskriteriene.

4.2.1 R^2 og $R^2_{justert}$

R^2 er et statistisk mål på hvor godt modellen passer datapunktene. R^2 kan defineres som andelen av variansen til responsen Y som kan forklares av den lineære modellen.

$$\begin{aligned} R^2 &= 1 - \frac{Var(modell)}{Var(total)} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}{\sum_{i=1}^{n}(y_i - \bar{Y})^2/n}, \end{aligned} \tag{4.9}$$

Hvor $\bar{Y}$ er gjennomsnittsverdien til responsvektoren. Legg merke til at R^2 alltid ligger mellom 0 og 1, hvor $R^2 = 0$ vil si at modellen ikke forklarer noe av variansen til Y, og $R^2 = 1$ vil si at modellen forklarer all variansen til Y. En utfordring med å benytte R^2 til å vurdere de ulike modellene, er at uansett hvilke forklaringsvariabler med tilhørende regresjonskoeffisient man tilføyer i modellen, så vil R^2 -verdien øke. Selv i tilfeller hvor det ikke finnes noen reel sammenheng mellom forklaringsvariabel og respons, vil man kunne estimere en regresjonskoeffisient som modellerer noe av den tilfeldige støyen i dataene. For å ta høyde for dette kan man heller benytte $R^2_{justert}$ for å evaluere modellen.

$R^2_{justert}$ reduseres dersom man inkluderer variabler som ikke forklarer tilstrekkelig av variansen til å kompensere for økningen i antall parametere.

$$R^2_{justert} = 1 - (1 - R^2)\frac{n-1}{n-m-1} \tag{4.10}$$

Hvor n er antall observasjoner, og m er antall parametere som er inkludert i modellen. På denne måten vil $R^2_{justert}$ reduseres dersom de ekstra modellparameterne ikke forklarer tilstrekkelig av variansen i Y. Derfor er $R^2_{justert}$ et bedre mål for å finne den optimale modellkompleksiteten.

Selv i tilfeller hvor man har funnet den optimale modellkompleksiteten, er det ikke dermed sagt at modellen evner å modellere virkeligheten på en god måte. Sum of squares error (SSE) kan fortsatt være høy relativt størrelsene i datasettet. Generelt kan man si at en modell som passer dataene godt, og har lav varians og lav bias, er en god modell. Hva som defineres som lavt kan være forskjellig fra datasett til datasett og kan være vanskelig å evaluere. I så måte kan R^2 og $R^2_{justert}$ være enklere å tolke enn SSE for å evaluere hvor godt modellen klarer å representere virkeligheten.

Det er viktig å presisere at selv i tilfeller hvor R^2 og $R^2_{justert}$ er lav, kan man trekke nyttig informasjon fra modellen dersom man har funnet signifikante sammenhenger mellom forklaringsvariabler og respons. Det er heller ikke slik at høye verdier for R^2 og $R^2_{justert}$ nødvendigvis utelukkende er bra. Eksempelvis kan begge R-verdiene være høye, mens modellen systematisk under-, eller overestimerer. For å se etter slike sammenhenger kan man plotte residualer mot tilpassede verdier, for å sjekke om variansen er tilfeldig fordelt. Dersom man oppdager mønstre i residualplottet bryter modellen også med modellantagelsene og man bør forsøke å transformere datasettet slik at man oppnår en tilfeldig spredning.

4.2.2 PRESS og R^2_{pred}

En annen metode som kan brukes til å evaluere modellene, er Predicted error sum of squares (PRESS). PRESS beregnes på følgende måte,

$$\begin{aligned} PRESS &= \sum_{i \in F_k} (y_i - \hat{f}^{-k}(x_i))^2 \\ &= n_k MSEP \end{aligned} \tag{4.11}$$

Dette er en iterativ prosess som gjøres K-antall ganger - én for hvert test- og treningssett. Man ønsker en modell med lav PRESS-verdi. Siden modeller som er overtilpasset tenderer til å gi små SSE-verdier for observasjoner som er inkludert i modelltilpasningen, men høye verdier for observasjoner som er ekskludert, kan PRESS være en god måte å sjekke om modellen er overtilpasset.

R^2_{pred} er et annet evalueringskriterium, gitt ved,

$$R^2_{pred} = 1 - \frac{PRESS}{\sum_{i=1}^{n} (y_i - \bar{Y})^2} \tag{4.12}$$

R^2_{pred} er generelt mer intuitiv en PRESS og kan sammenlignes med R^2 for å sjekke

modellen for overtilpasning. R^2 og R^2_{pred} har ganske lik form, men kan ha svært ulike verdier. Høy R^2-verdi relativt til R^2_{pred} indikerer at modellen er overtilpasset.

4.2.3 BIC

Bayesian informasjonskriterium (BIC)(Schwarz, 1978) er en annen metode for å test hvor godt modellen er tilpasset datasettet uten å overtilpasse modellen. Man ønsker en så lav BIC-verdi som mulig. I motsetning til $R^2_{justert}$ øker BIC dersom man inkluderer variabler som ikke forklarer tilstrekkelig av variansen til å kompensere for økningen i antall parametere. BIC vil på den måten indikere hva som er modellens optimale kompleksitet. BIC kan beregnes på denne måten,

$$BIC = -2\ ln(SSE) + m\ ln(n) \tag{4.13}$$

hvor m er antall parametere i modellen.

BIC benyttes til å vurdere den relative forskjellen mellom modellene, men sier lite om modellens evne til å representere virkeligheten.

4.3 Metoder for regresjon

Frem til nå har det vært antatt at modellparameterne har vært estimert og tilpasset datasettet. Dette delkapittelet vil konsentrere seg mer om ulike metoder for å estimere og selektere modellens signifikante modellparametere.

4.3.1 Minste kvadraters metode

Minste kvadraters metode (Ordinary least square - OLS) er den enkleste og mest kjente regresjonsmetoden. OLS tar sikte på å finne modellen, $f(X)$, ved å estimere de sanne populasjonsparameterne, β, ved å minimerer SSE,

$$\begin{aligned} SSE(\hat{\beta}) &= \sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j)^2 \\ &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}). \end{aligned} \tag{4.14}$$

Ved å derivere uttrykket 4.14 med hensyn på $\hat{\beta}$ og sette dette lik 0,

$$\frac{\partial SSE}{\partial \hat{\beta}} = -2X^T(Y - X\hat{\beta}) \tag{4.15}$$

$$X^T(Y - X\hat{\beta}) = 0 \tag{4.16}$$

kan man ved å omrokerer på ligning 4.16 finne uttrykket for $\hat{\beta}$ som minimerer SSE,

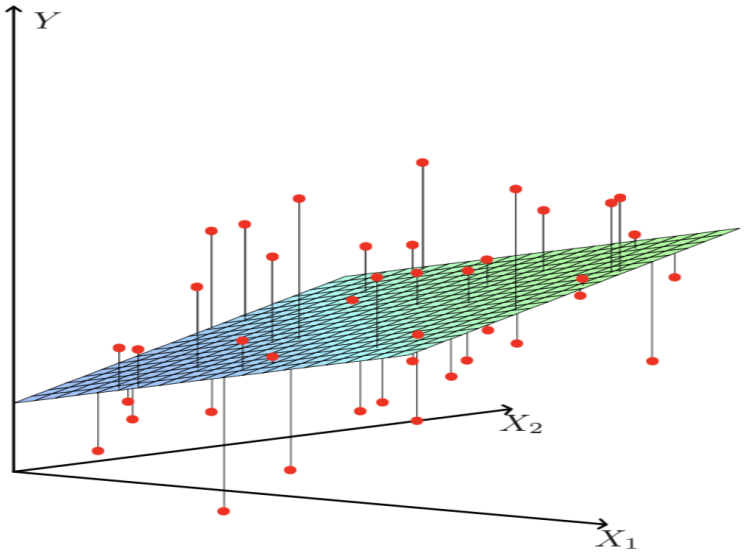
$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{4.17}$$

Den forventede responsvektoren, $\hat{Y}$, gitt matrisen, X, av forklaringsvariabler er da gitt ved

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y. \tag{4.18}$$

Legg merke til at dette er den samme modellen som er presentert i delkapittel 4.1, men her er $\hat{\beta}_0$ inkludert i koeffisientvektoren, $\hat{\beta}$. Samtidig er X-matrisen utvidet til en $n \times (p+1)$, hvor den siste kolonnen består av enere. Figur 4.4 illustrerer et eksempel hvor OLS er benyttet til å tilpasse en modell til et tilfeldig datasett med to variabler.

I tilfeller hvor $p > n$ vil kovariansmatrisen $X^T X$ være singulær og ikke-invertibel som igjen fører til at man ikke kan finne en unik løsning og dermed ikke klarer å estimere parameterne. For å kunne benytte Minste kvadraters metode må derfor $n > p$ for å finne en entydig løsning. Dersom noen av forklaringsvariablene er eksakte lineærkombinasjoner av andre vil dette også medføre at kovariansmatrisen er singulær og ikke-invertibel.



Figur 4.4: Minste kvadraters metode benyttet for å lage et todimensjonalt plan som minimerer SSE for de observerte verdiene. Figuren er hentet fra (Friedman et al., 2001)

.

4.3.2 Variabelseleksjon

Som forklart tidligere kan den fullstendige modellen inneholde én eller flere parametere som har lav forklaringsverdi. Foruten faren for å overtilpasse modellen vil det i mange sammenhenger også være upraktisk å overvåke og registrere store mengder data. Som en løsning finnes det flere metoder og fremgangsmåter som reduserer antall modellparametere, her under presenteres tre av metodene.

Best-Subset seleksjon
Ved Best-subset regresjon tilpasses modeller med alle mulige variabelkombinasjoner fra null-modellen, som kun inneholder den tilfeldige variabelen, $\bar{Y}$, til fullstendig modell hvor alle de estimerte parameterne er inkludert. For å avgjøre hvilken modell som er den beste, kan man eksempelvis bruke et av kriteriene beskrevet i delkapittel 4.2.

Best-Subset seleksjon blir upraktisk dersom man har mange forklaringsvariabler, da det etterhvert blir svært mange modellkombinasjoner å sammenligne. En alternativ fremgangsmåte er derfor å benytte enten forlengs seleksjon, eller baklengs eliminasjon.

Forlengs seleksjon
Ved forlengs seleksjon starter man med null-modellen og legger til den parameteren

som forbedrer modellen mest, inntil det ikke er hensiktsmessig å legge til flere variabler. Det vil si at man legger til den variabelen som forklarer mest av gjenværende SSE. Variabelen som skal inkluderes kan bestemmes ved hjelp kriteriene beskrevet i delkapittel 4.2, eller eksempelvis ved F-testing,

$$F_{inn} = \frac{SSE_p - SSE_{p+1}}{SSE_{p+1}/(n-p-2)}. \tag{4.19}$$

Hvor SSE_p og SSE_{p+1} er sum of squares error for henholdsvis modeller med p og $p+1$ antall variabler. F_{inn} testes så mot $F_{\alpha,1/n-p-2}$, hvor α er testens signifikansnivå.

Ulempen med denne metoden sammenlignet med best-subset er at modellene vil være nøstet inne i hverandre. Det er en grådig algoritme som alltid leter etter lokalt beste løsning gitt hva som allerede er inkludert i modellen. Fordelen er derimot at man kan benytte forlengs eliminasjon selv om man har mange variabler, p. Selv når $p > n$.

Baklengs eliminasjon
Ved baklengs eliminasjon starter man med den fullstendige modellen, det vil si alle modellparameterne, og eliminerer den parameteren som har minst påvirkning på modellen. Det vil si den variabelen som forklarer minst av variasjonen til responsvariabelen. Dette gjentas inntil de gjenværende parameterne anses å være signifikante. Tilsvarende som for forlengs seleksjon kan man benytte kriteriene fra delkapittel 4.2, eller eksempelvis F-testing for å avgjøre om en parameter skal ekskluderes fra modellen,

$$F_{ut} = \frac{SSE_{p-1} - SSE_p}{SSE_p/(n-p-1)}. \tag{4.20}$$

Hvor SSE_{p-1} og SSE_p er sum of squares for henholdsvis modeller med $p-1$ og p antall forklaringsvariabler. F_{ut} testes mot $F_{\alpha,1/n-p-2}$.

Dersom $n < p$ vil kovariansmatrisen $X^T X$ være singulær og ikke-invertibel. Baklengs eliminasjon kan derfor kun benyttes hvis $n > p$.

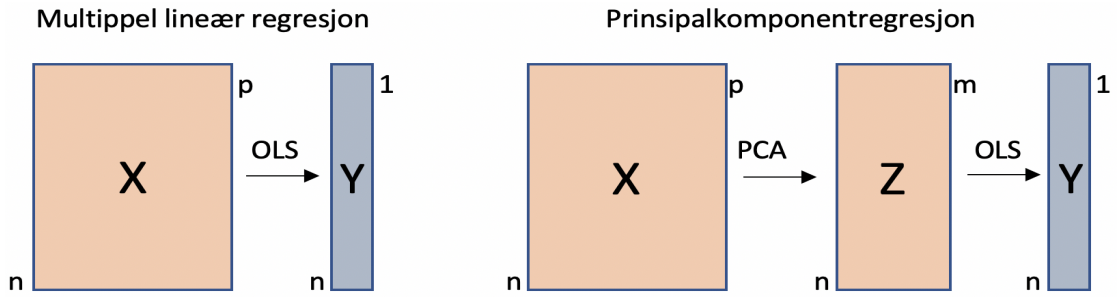
4.3.3 Dimensjonsreduserende metoder

Dersom to eller flere av forklaringsvariablene er korrelerte med hverandre, multikollinearitet, kan dette føre til høy varians i modellparameternes estimater. Små endringer i data kan derfor føre til svært ulike modelltilpasninger. Siden de korrelerte variablene innehar noe av den samme informasjonen er det ubetydelig for modellen hvilken av de estimerte parameterne som forteller informasjonen, gjerne

forteller de litt hver. Ideelt sett ønsker man at alle variablene og modellparameterne skal komme med unik informasjon.

Prinsipalkomponentregresjon
Prinsipalkomponentanalyse (PCA)(Hotelling, 1933) tar sikte på å erstatte p, mer eller mindre, korrelerte forklaringsvariabler, med $m < p$ ukorrelerte lineære kombinasjoner av de originale variablene. Man transformerer datasettet slik at hver av de nye variablene inneholder unik informasjon.

Prinsipalkomponentregresjon (PCR) benytter OLS på det transformerte datasettet fra PCA for å estimere modellparameterne. Figur 4.5 illustrerer hvordan PCR skiller seg fra vanlig OLS.



Figur 4.5: Illustrerer forskjellen på vanlig multippel lineær regresjon og Prinsipalkomponentregresjon (PCR). Ved vanlig multippel lineær regresjon tilpasses modellens parametere til det opprinnelige datasettet, mens for PCR transformeres datasettet slik at hver av forklaringsvariablene inneholder unik informasjon før man benytter multippel lineær regresjon på dette datasettet. Siden de ulike prinsipalkomponentene forklarer mindre og mindre av variansen i x-datene vil det potensielt være unødvendig å inkludere alle.

For å transformere X-matrisen med forklaringsvariabler dekomponerer man kovariansmatrisen, $X^T X$ slik at man finner egenvektorene, V_j, og egenverdiene, λ_j. Egenvektorene gir retningen til en vektor i et rom, mens egenverdien angir lengden, det vil si variansen, til vektoren. Alle egenvektorene står ortogonalt på hverandre slik at hver av de utspenner et unikt underrom. Sorterer man egenvektorene etter størrelsen på de tilhørende egenverdiene vil den retningen som forklarer mest av variansen være først, den som forklarer nest mest nummer 2, og så videre. Prinsipalscoringene til X blir da,

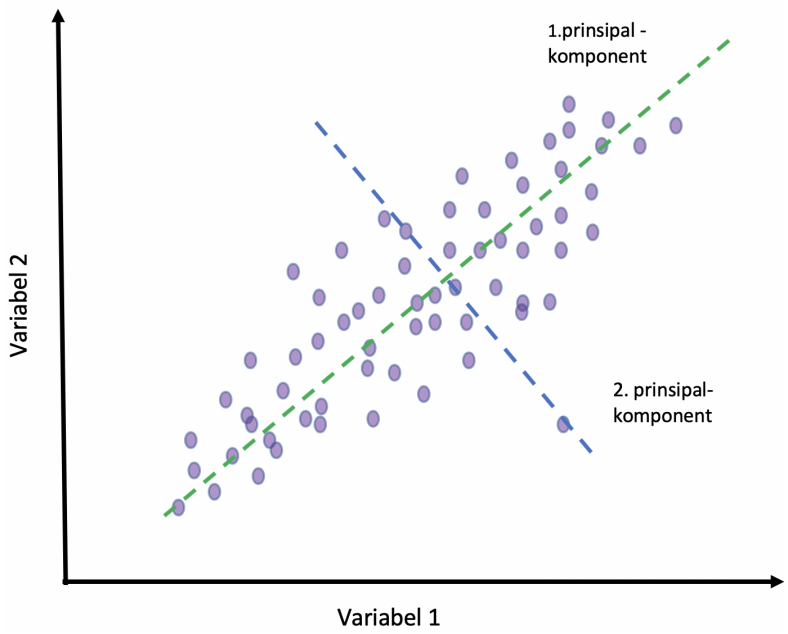
$$Z = XV \tag{4.21}$$

hvor,

$$Var(Z_i) = \lambda_i \text{ og}$$

$$Cov(Z_i Z_j) = 0 \text{ for } i \neq j$$

For at egenvektorene, også kalt prinsipalkomponentene, skal rangeres rettferdig bør variablene standardiseres. PCA benytter seg kun av forklaringsvariablene, X, og ikke av responsen, Y. Figur 4.6 illustrerer hvordan prinsipalkomponentene ville sett ut for et tilfeldig valgt datasett med to variabler.



Figur 4.6: Prinsipalkomponentene til et tilfeldig datasett. Første prinsipalkomponent er i den retningen som forklarer det meste av variansen i dataene, mens andre prinsipalkomponent er i retningen som forklarer den resterende variansen. Prinsipalkomponent 1 har høyest egenverdi.

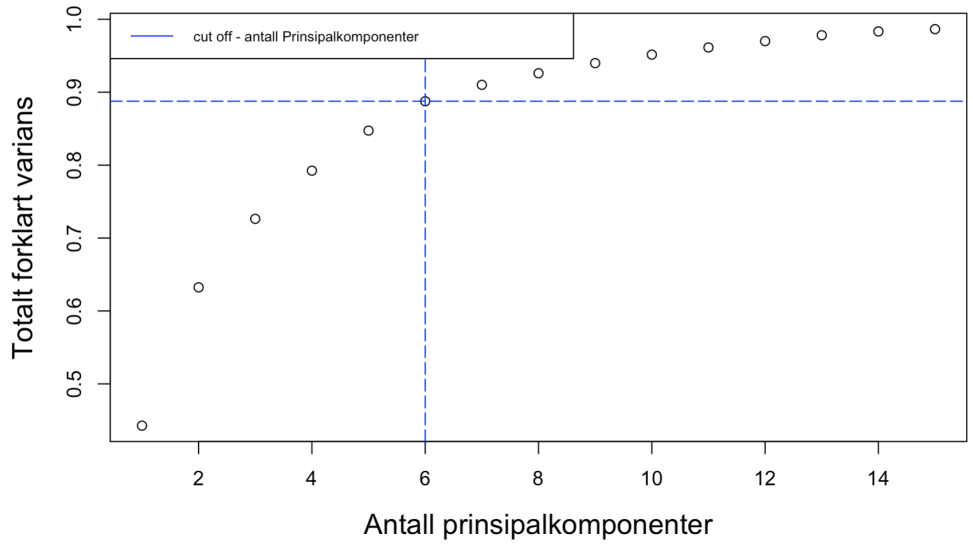
Ved å dele egenverdien til en spesifikk prinsipalkomponent på summen av alle egenverdiene finner man hvor mye av variansen i dataene de ulike prinsipalkomponentene forklarer. I figur 4.7 kan man se hvor mye av variansen i X-dataene som forklares etterhvert som man øker antallet prinsipalkomponenter for et tilfeldig datasett.

Siden de ulike prinsipalkomponentene forklarer mindre og mindre av variansen i X-datene vil det potensielt være unødvendig å inkludere alle. Man benytter derfor en kompleksitetsparameter m, for å avgjøre hvor mange komponenter som skal inkluderes.

PCR benytter deretter OLS på det transformerte datasettet for å estimere parameterne,

$$\hat{\beta}_{pcr} = V_m(V_m^T X^T X V_m)^{-1} V_m^T X^T Y \tag{4.22}$$

hvor V_m er en matrise av de m første prinsipalkomponentene, og Y er responsen.



Figur 4.7: Kumulativ forklart varians som funksjon av antall prinsipalkomponenter for et tilfeldig datasett. Man ser at de første prinsipalkomponentene forklarer mye av den variansen man finner i dataene. Forklaringsverdien ved å legge til flere komponenter avtar etterhvert som egenverdiene synker. I dette eksempelet er forklaringsverdien til den syvende prinsipal komponenten kun 2%, man kan da vurdere om det er hensiktsmessig å inkludere denne.

Siden man kun benytter m komponenter vil estimatet være forventningsskjevt (bias). m kan eksempelvis bestemmes ved å kryssvalidere mean squared error of prediction (MSEP) (se delkapittel 4.1.2). I tilfeller hvor $m = p$, vil $\hat{\beta}_{pcr} = \hat{\beta}_{OLS}$.

Partial least square regresjon

En mulig fallgruve med PCR er at de valgte prinsipalkomponentene ikke har noen relasjon til responsen. Et alternativ til PCR er derfor Partial least square regresjon (PLSR) (H. Wold, 1982; S. Wold, Sjöström & Eriksson, 2001), som tar sikte på å danne ortogonale lineærkombinasjoner, Z_m, av de originale observasjonene X_j som best forklarer responsen Y. Det transformerte datasettet består derfor av transformerte forklaringsvariabler, Z_m, som er rangert etter hvilke som best forklarer responsen. Z_m bestemmes ut i fra variansen i X, men også korrelasjonen mellom X og Y. Siden PLSR også tar hensyn til responsvektoren, Y, trengs det ofte færre PLSR -komponenter for å forklare responsen sammenlignet med PCR.

Til tross for fordelene med PLSR hvor man unngår multikollinearitet og kan håndtere mange forklaringsvariabler, kan tolkningen av PLSR-modellen være vanskelig da den ikke foretar variabelseleksjon. Variable Importance in Projection (VIP) (S. Wold, Johansson & Cocchi, 1993) er et estimat av viktigheten til hver av forklaringsvariablene i komponentene brukt i PLSR-modellen. En variabel med VIP score betydelige lavere en 1 er mindre viktige og man kan vurdere å ekskludere denne fra modellen. På denne måten kan man gjennomføre variabelseleksjon ved en

kombinasjon av PLSR og VIP score. I følge G.Chong og C.H.Jun bør en variabel inkluderes dersom den har VIP score over 0.83 - 1.21 (Chong & Jun, 2005).

4.3.4 Krympingsmetoder

Ved å benytte forlengs variabelseleksjon eller baklengs eliminasjon (se delkapittel 4.3.2) kan man ikke si noe om variablene som er fjernet sin påvirkning på responsen. Man kan si at man setter disse regresjonskoeffisientene til null. Som et alternativ kan man i stede krympe disse regresjonskoeffisientene.

Ridge regresjon
Ved Ridge regresjon (Hoerl & Kennard, 1970) krymper man regresjonskoeffisientene mot 0. På denne måten øker modellens bias, mens variansen reduseres betydelig sammenlignet med OLS. Modellenes regresjonskoeffisienter beregnes ved hjelp av ligning 4.23,

$$\hat{\beta}_{ridge} = argmin_{\beta}\bigg(\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\bigg), \tag{4.23}$$

på matriseform blir uttrykket,

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y. \tag{4.24}$$

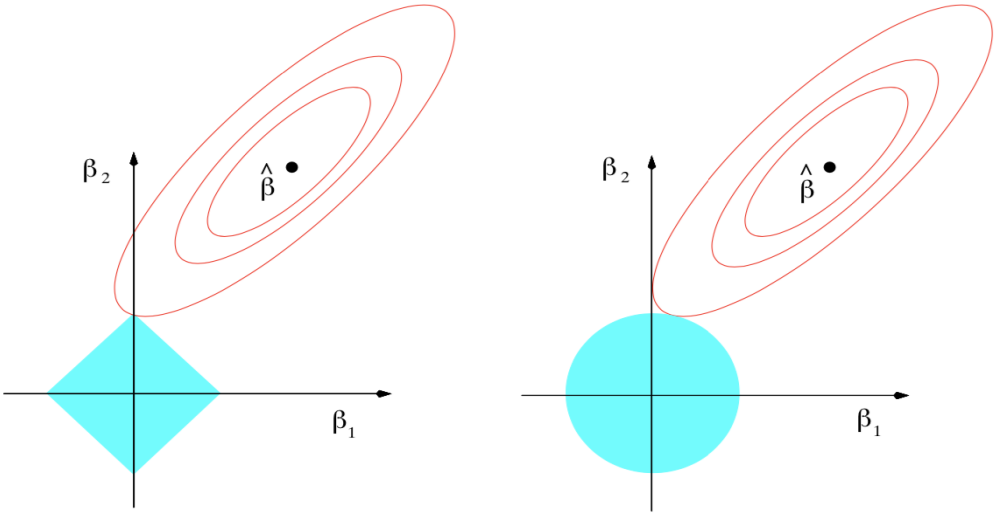
Som man kan se legges et ekstra ledd til ligning 4.17 som ble benyttet til å estimere regresjonskoeffisientene ved OLS. I er en $p \times p$ identitetsmatrise, og λ er kompleksitetsparameteren som bestemmer hvor mye regresjonskoeffisientene skal reduseres. På samme måte som for PCR må datasettet standardiseres for at kompleksitetsparameteren skal ha lik innvirkning på alle variablene. For å komme frem til den kompleksitetsparameteren som gir den optimale avveiningen mellom bias og varians benyttes kryssvalidering. Ved $\lambda = 0$ vil man oppnå samme modell som ved OLS, mens ved $\lambda \to \infty$ vil regresjonskoeffisientene gå mot null og man står igjen med kun konstantleddet, $\hat{\beta}_0 = \bar{Y}$. Ridge regresjon kan ikke benyttes til variabelseleksjon, men presterer spesielt godt dersom noen av modellens sanne parametere er små, eller til og med null. Den presterer ikke like godt dersom alle de sanne parameterne er av moderat størrelse (Tibshirani, 2013a).

Lasso

Selv om man ved ridge regresjon ikke reduserer noen av koeffisientene helt til null, skader tilsynelatende ikke dette modellens prediksjonsevne (Tibshirani, 2013b). Modellen kan derimot være vanskelig å tolke, spesielt dersom det er mange variabler. Lasso (Tibshirani, 1996) er konseptuelt svært likt ridge regresjon, men kan redusere regresjonskoeffisientene som har lav påvirkning på responsen til null, og dermed benyttes til variabelseleksjon. Ved Lasso legges et ekstra ledd til ligning 4.17, men istedenfor å kvadrere som ved ridge, bruker Lasso absoluttverdien til koeffisientene (se ligning 4.25). Dette gjør at ved høye λ-verdier vil flere parametere reduseres til null (se figur 4.8). Det vil aldri skje ved ridge regresjon.

$$\hat{\beta}^{lasso} = argmin_{\beta} \left(\frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=i}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{4.25}$$

I motsatt fall, dersom λ er liten vil parameterne være de samme som ved vanlig OLS ($\hat{\beta}_j = \hat{\beta}_j^{ols}$). Om, og eventuelt hvor mye, modellen reduseres er altså avhengig av kompleksitetsparameteren, λ. For å finne den optimale kompleksitetsparameteren kan man, som tidligere, benytte kryssvalidering. Figur 4.8 illustrerer forskjellen mellom lasso og ridge regresjon.



Figur 4.8: Illustrerer forskjellen mellom ridge regresjon og Lasso. $\hat{\beta}$ er summen av de to estimerte regresjonskoeffisientene, β_1 og β_2 fra minste kvadraters metode. Ellipsene rundt $\hat{\beta}$ illustrerer sammensetninger av ulike verdier for β_1 og β_2 som resulterer i samme SSE. Ved ridge regresjon (til høyre) tvinger man summen at disse koeffisientene til å ligge innenfor begrensningsområdet (det turkise området). Som man kan se av figuren til høyre reduseres både β_1 og β_2 betraktelig, men ingen av koeffisientene reduseres til null. Det er samme prinsipp for Lasso (til venstre), men som man kan se er begrensningsområdet endret. I dette eksempelet er β_1 redusert til null ved lasso. Figuren er hentet fra (Friedman et al., 2001)

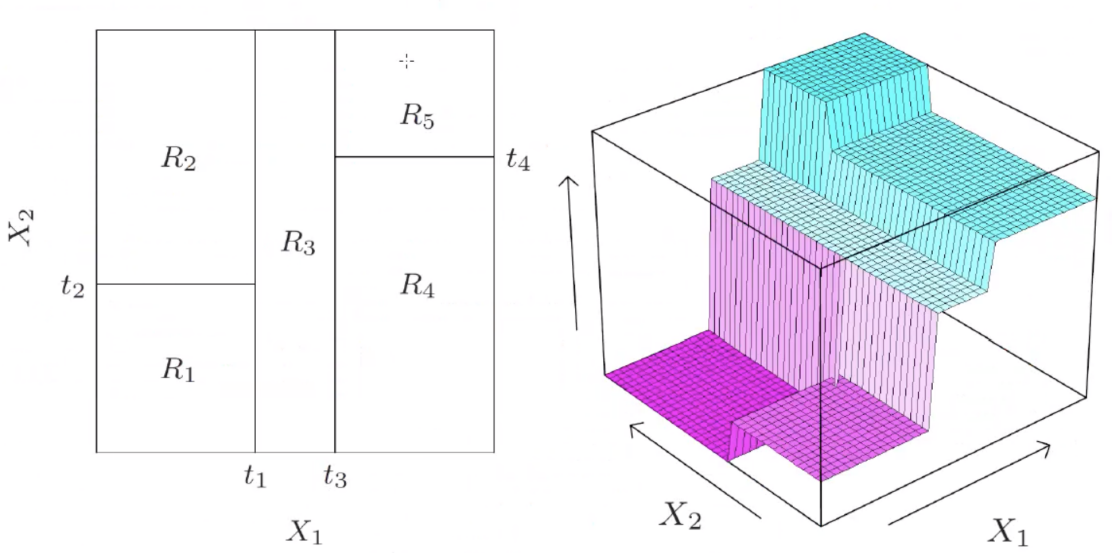
4.4 Beslutningstrær

Et beslutningstre er en repeterende binær oppdeling av utfallsrommet. Det vil si ved hver node deles en region i to eller flere regioner. Dette gjentas helt til man når et stoppkriterium. Til slutt står man igjen med et inputrom som er delt i M ulike regioner, R_m. Hver av regionen har hver sin tilhørende estimerte verdi, $\hat{c}_m$, for eksempel gjennomsnittsverdien i denne regionen. Den endelige prediksjonsmodellen blir da (Friedman et al., 2001),

$$\hat{f}(X) = \sum_{m=1}^{M} \hat{c}_m I(X \epsilon R_m) \tag{4.26}$$

hvor $\hat{f}(X)$ er den tilpassede modellen og X er element av det m-dimensjonale rommet.

Figur 4.9 er et eksempel på et beslutningstre med to variabler, hvor de ulike nivåene i grafen illustrerer verdien til denne regionen, altså c_m.



Figur 4.9: To ulike metoder for å illustrere prinsippet med beslutningstre. Hver av regionene R_m har en egen verdi, c_m. Verdien til c_m for de ulike regionene er illustrert i 3D-plottet til høyre. Denne fremstillingsmåten blir problematisk dersom treet har flere enn to variabler. Illustrasjonen er hentet fra (Friedman et al., 2001).

Beslutningstreet velger å dele de ulike variablene slik at man minimerer tapsfunksjonen. Man ønsker altså å hente ut mest mulig informasjon ved å splitte variablene i det punktet som minimerer SSE. I prinsippet kunne man splittet variablene ved hver observasjon og tilpasset et fullstendig tre, men dette ville ført til kraftig

overtilpasning av modellen. I stedet starter man med hele datasettet, og deler hver variable, X_j i to, slik at man får to nye halve hyperplan (regioner).

$$\begin{aligned} R_1(j,s) &= \{X|X_j < s\} \\ R_2(j,s) &= \{X|X_j > s\} \end{aligned} \tag{4.27}$$

De to nye regionene er splittet ved variabel j og ved punkt s. I figur 4.9 er variabel X_1 delt ved t_1 og t_3, mens X_2 er delt ved t_2 og t_4. Beslutningstre er lette å tolke og man trenger ikke ha en bakgrunn fra statistikk for å forstå prinsippet. For å avgjøre hvilken variabel og ved hvilket punkt man skal splitte finnes det ulike kriterier, her er det valgt å benytte 4.28,

$$\min_{j,s}[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2] \tag{4.28}$$

hvor c_1 og c_2 er gjennomsnittsverdien for de to regionene.

Videre er utfordringen hvor stort tre, eller modell, man skal tilpasse. Det er tidligere illustrert viktigheten av å finne en balanse mellom kompleksitet og anvendelighet (se figur 4.3). Det finnes to ulike fremgangsmåter for å avgjøre størrelsen på treet. Den ene er å stoppe modellbyggingen når man ikke reduserer SSE tilstrekkelig ved å inkludere den spesifikke noden. Problemet med denne fremgangsmåten er at man kan risikere å miste signifikant informasjon lengre nede i hierarkiet. En bedre metode er derfor å starte med et stort tre, enten med et prespesifisert antall noder, eller et fullstendig tre, for så å fjerne de nodene som ikke tilfører modellen noen forklaringsverdi. For å avgjøre når man skal stoppe bruker man en funksjon, kostkompleksitetskriteriet (cost-complexity criterion, (Friedman et al., 2001))

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T| \tag{4.29}$$

hvor,

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2. \tag{4.30}$$

N_m er antall observasjoner, mens $|T|$ er antall noder treet har. Siden c_m blir mer og mer lik den observerte verdien, reduseres ligning 4.30 når antall noder øker. Derimot øker $\alpha|T|$, som gjør at kostkompleksitetskriteriet finner en balanse mellom kompleksitet og anvendelighet for modellen. Den optimale kompleksitetsparameteren, α, finnes ved kryssvalidering. Større α, gir mindre tre, mens $\alpha = 0$ gir et fullstendig tre.

Beslutningstrær kan også benyttes ved klassifikasjon, forskjellen ligger i evalueringskriteriet. Beslutningstrær er lette å tolke, håndterer avviksverdier og irrelevant data godt (siden disse får sin egen avgrening). Ulempene er, som man kan se av figur 4.9, at overgangen mellom de ulike regionene er brå. beslutningstrær er generelt svært ustabile, har høy varians og egner seg dårlig for prediksjon(Tibshirani, 2013c).

4.4.1 Random Forest

For å håndtere ulempene med *ett* beslutningstre, tilpasser man flere trær og aggregerer resultatet. En av metodene som benytter denne fremgangsmåten kalles Random Forest og ble introdusert av Breiman (Breiman, 2001). Dette gjøres ved å tilfeldig trekke observasjoner fra datasettet, med tilbakelegging, for å danne nye datasett (bootstrap sample). Ulempen med bootstrap sampling er at man ikke har nye observasjoner, men danner nye datasett fra det samme datasettet. Datasettene er derfor korrelerte og vil forsterke relasjonen som ligger i det opprinnelige datasettet. For å ta høyde for dette benytter Random Forest seg kun av et utvalg av variablene ved hver node. På denne måten lager man ulike beslutningstre fra hvert av de nye datasettene. Når beslutningstrærne skal aggregeres i resultatet finnes det ulike framgangsmåter. For regresjon benytter man normalt et gjennomsnitt over prediksjonene fra trærne, mens for klassifikasjon benyttes ofte konsensus, eller sannsynlighet. Eksempelet i tabell 4.1 forklarer de to fremgangsmåtene for klassifikasjon på en god måte:

Tabell 4.1: Eksempel: Man har to mulig utfall, A og B. For å predikere hvilket utfall som er det reelle er det valgt å lage 3 beslutningstrær (se tabell). To av trærne stemmer på utfall B og ett på utfall A, dersom man benytter konsensus, ville utfallet blitt B - to stemmer mot én. Derimot hvis man legger sammen snittverdien av sannsynligheten ville utfallet blitt A ($\sum_{i=1}^{3} P_i(X = A) = 0.6$).

Tre nr.	P(X = A)	P(X = B)
1	0.9	0.1
2	0.45	0.55
3	0.45	0.55

Hvilke fremgangsmåter man bør bruke for klassifikasjon kommer an på datasettet.

I eksempelet fra tabell 4.1 kan man se at tre nr. 1 er svært sikker på at utfallet er A, mens de andre trærne er omlag 50/50. I dette tilfellet ville det derfor vært fornuftig å aggregere sannsynligheten for å avgjøre utfallet.

De observasjonene som ikke benyttes til å danne et beslutningstre kalles for Out of bag (OOB) observasjoner. Disse observasjonene kan benyttes for å estimere modellens prediksjonsfeil. Out of bag score kan sammenlignes med kryssvalidering (se delkapittel 4.1.2) og er en metode for å validere Random Forest-modellen. Scoren estimeres ved at beslutningstrærne som er tilpasset bootstrap-datasettene estimerer og imputerer OOB-observasjonene og sammenligner de imputerte verdiene med de virkelige verdiene i OOB-datasettet. For kategoriske variabler estimeres OOB-scoren som andelen feilklassifiserte variabler. For å estimere OOB-scoren ved kontinuerlige variabler kan man benytte normalized root mean squared error (NRMSE)(Stekhoven & Stekhoven, 2012):

$$NRMSE = \sqrt{\frac{1}{n}\frac{(X_i - f(X_i))^2}{Var(X_i)}}, \tag{4.31}$$

hvor X_i er de virkelige verdiene, $f(X_i)$ er de imputerte verdiene, og n er antall OBB observasjoner. $Var(X_i)$ er variansen til dataene som er i OOB-datasettet (Stekhoven & Stekhoven, 2012). Dersom Random Forest benyttes til imputering av manglende data kan OOB-scoren benyttes for å vurdere kvaliteten på de imputerte dataene.

4.5 Variansanalyse

4.5.1 Variansanalyse

Variansanalyse, populært kalt ANOVA (fra engelsk "Analysis of variance") er en statistisk metode som sammenligner forskjeller i ulike gruppers forventningsverdier. Dette gjøres ved å se på variasjonen i dataene, og hvor denne variasjonen finnes. Spesifikt sammenligner ANOVA variasjonen mellom grupper og variasjonen innad i gruppene. Tenk at man ønsker å undersøke om en medisin reduserer kroppstemperaturen til pasienter. Figur 4.10 illustrerer de fiktive resultatene for henholdsvis med (blå/grønn), og uten (rød), medisin. I graf A er differansen i forventningsverdi mellom gruppene tydelig og det kan se ut til at de ulike gruppene/medisinene har effekt. I graf B er differansen mellom gruppenes forventningsverdi lav og det er vanskeligere å se en tydelig effekt av medisin. Variasjonen innad i de ulike gruppene er lav (standardavvik = 0.05), men ved høyere variasjon innad i gruppene vil sannsynlighetsfordelingenes utfallsrom øke og potensielt overlappe. Graf C er

et eksempel på sistnevnte. Til tross fra at forventningsverdien i graf C er den samme som i A er det vanskelig, ut ifra grafene, å se om medisinene har effekt på temperaturen. Dette eksempelet illustrerer viktigheten av å ta hensyn til både variasjonen mellom grupper, men også innad i hver gruppe.

For å teste om grupper har effekt på responsen benyttes en F-test. Man tester da om det ikke er effekt av gruppe (null -hypotesen, H_0) mot alternativet, at det er effekt av gruppe (alternativ hypotesen, H_1). Matematisk kan dette illustreres på denne måten:

$$H_0 : \ \mu_1 = \mu_2 = ... = \mu_v$$

$$H_1 : \ \mu_j \neq \mu_m$$

hvor μ_j og μ_m er populasjonsgjennomsnittet til to vilkårlige grupper av alle gruppene som testes. Man forkaster null-hypotesen, og påstår at det er signifikant forskjell mellom gruppene dersom,

$$\begin{aligned} F_{\alpha,v-1,n-v} &< \frac{Variansen\ mellom\ grupper}{Variansen\ innad\ i\ gruppe} \\ &< \frac{MST}{MSE} \end{aligned} \tag{4.32}$$

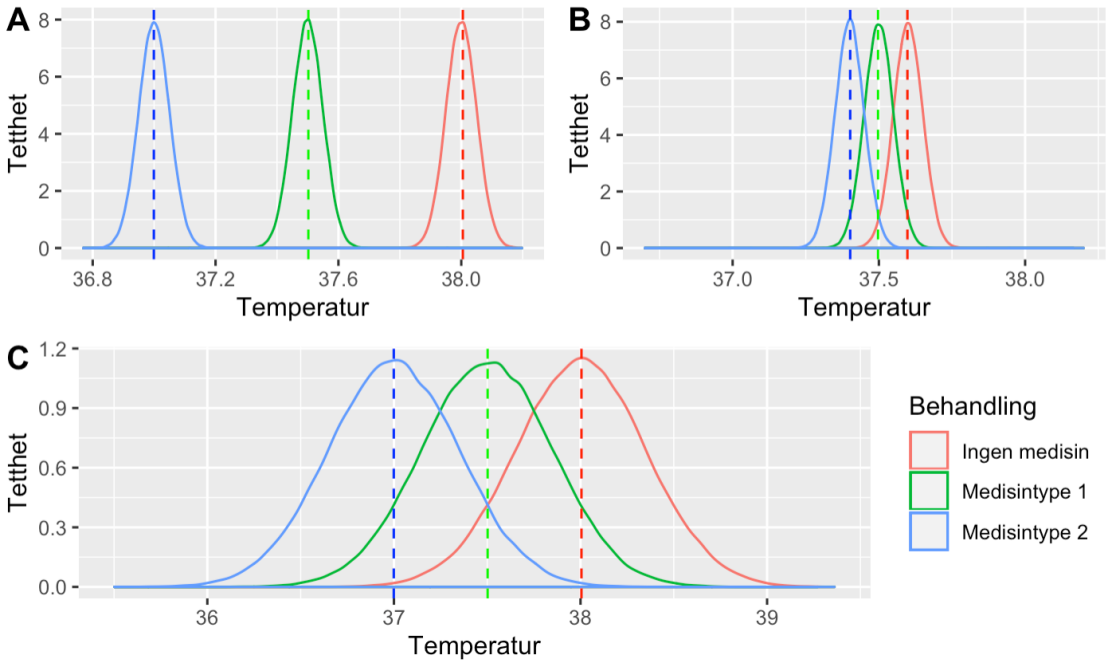
hvor,

$$\begin{aligned} MST &= \frac{SST}{v-1} \\ &= \frac{\sum_{j=1}^{v} n_j(\bar{Y_j} - \bar{Y_0})^2}{v-1} \end{aligned} \tag{4.33}$$

og,

$$MSE = \frac{\sum_{j=1}^{v}\sum_{i=1}^{n}(Y_{ji} - \bar{Y_j})^2}{n-v}. \tag{4.34}$$

Hvor $\bar{Y_j}$ og n_j er henholdsvis gjennomsnitt og antall observasjoner i gruppe j, Y_{ji} er den i-te observasjonen i gruppe j, mens $\bar{Y_0}$ enten er forventningsverdiene til en gitt referansegruppe eller gjennomsnittsverdien for alle gruppene.



Figur 4.10: Simulerte observasjoner av pasienters kroppstemperatur ved bruk av medisintype 1, medisintype 2, eller ingen medisin. For grafene A og B er variansen innad i hver gruppe er satt til 0.05 grader Celsius, mens variansen mellom gruppene er endret fra graf A og graf B. I graf C er variansen innad i gruppene satt til 0.35 grader celsius, mens variansen mellom gruppene er den samme som i graf A.

4.5.2 Tukey

Resultatet fra en ANOVA indikerer kun om det er signifikant gruppeeffekt, men ikke hvor denne signifikansen ligger. I etterkant av en ANOVA kan man benytte en Tukey-test (Tukey, 1977) for å finne hvilke grupper som er signifikant forskjellig. Dette forutsetter selvsagt en kategorisk variabel og at ANOVA indikerer signifikant forskjell mellom grupper. En Tukey-test parvis sammenligner alle de ulike gruppenes gjennomsnitt for å finne hvilke som er signifikant ulike. Siden man har $\binom{v}{2}$ ulike kombinasjoner å sammenligne ville en vanlig t-test økt sannsynligheten for minst én type I feil, det vil si konkluderer med at null-hypotesen er feil, selv om den egentlig er rett.

$$P(minst\ en\ Type\ I\ feil) = 1 - (1 - \alpha)^v. \tag{4.35}$$

Ved Tukey holdes det totale sannsynligheten for å gjøre minst én type I feil til α. Ligning 4.36 viser hvor stor differansen mellom to gruppers gjennomsnitt minst må være for å kunne påstå signifikant forskjell,

$$\left|\bar{Y}_i - \bar{Y}_s\right| > \frac{q_{\alpha(v,n-v)}}{\sqrt{2}}\sqrt{\frac{MSE}{r_i} + \frac{MSE}{r_s}}. \tag{4.36}$$

hvor r_i og r_s er antall observasjoner i henholdsvis gruppe i og s, og $q_{\alpha(v,n-v)}$ er øvre kritiske verdi i en Studentisert utvalgsfordeling.

4.5.3 Blokkfaktor

Variabler som påvirker sammenhengen mellom forklaringsvariabler og respons, uten å selv være av interesse, bør introduseres som blokkfaktor. For eksempel kan romtemperatur, sykehus, måleutstyr eller når på dagen medisinen ble gitt, alle påvirke kroppstemperaturen til pasienten, uten at det er disse effektene man ønsker å analysere. Ved å introdusere en blokkfaktor danner man homogene grupper (blokker) innen hver blokkfaktor, slik at denne faktoren holdes konstant for hver gruppe, mens den mulig interessante forklaringsvariabelen får variere fritt. Slik kan man analysere den mulige effekten av den forklaringsvariabelen som er interessant, uten å tenke på variasjoner i blokkfaktoren. Modellen med blokkfaktor vil da være,

$$Y = \beta_0 + B(X_j) + f(X_{(-j)}) + \epsilon, \quad hvor\ \epsilon \sim N(0, \sigma^2) \tag{4.37}$$

hvor $X_{(-j)}$ er de forklaringsvariablene man ønsker å analysere effekten av (alle utenom blokkfaktor j). $B(X_j)$ er effekten av å være i blokk j, mens $f(X_{(-j)})$ er effekten av de interessante forklaringsvariablene.

4.5.4 VIF

Multikollinearitet, høy korrelasjon mellom forklaringsvariablene, reduserer treffsikkerheten til estimatene av modellkoeffisientene som igjen øker standardavviket til $\hat{\beta}_j$. For å undersøke for multikollinearitet mellom forklaringsvariablene kan man benytte Variance inflation factor (VIF) (Daniel & Wood, 1980). VIF estimerer hvor mye variansen til den estimerte regresjonskoeffisienten øker som følge av multikollinearitet. Den minste mulige VIF-verdien er 1, som betyr fravær av multikollinearitet. Som en tommelfingerregel indikerer VIF-verdier over 5-10, problemer med multikollinearitet, og en av variablene bør vurderes slettet (James, Witten, Hastie & Tibshirani, 2013).

Kapittel 5

Datamateriell

Alt datamaterialet som er analysert kommer fra Norturas database og tilhører Nortura. Dataene er todelt; én fil, *slakterkylling.r* inneholder data fra 2013 til februar 2018, mens de resterende datasettene er fra Norturas nye registreringssystem, Fjørfekjøttkontrollen, som ble tatt i bruk fra og med februar 2018. Dataene før og etter februar 2018 er ikke ført etter samme mal, og inneholder få av de samme registreringsvariablene.

Alle datasettene inneholder en eller flere "nøkler" som kan koble flere datasett sammen. De som starter med "PK" er primærnøkler, altså en unik nøkkel for det aktuelle datasettet som kan kobles til andre sekundærnøkler "FK" i andre datasett. Eksempelvis har datasettet *Produksjonstype.r* en unik nøkkel "PK_produksjonstype" som kan kobles til "FK_produksjonstype" i de andre datasettene.

Mest informasjon finner man i datasettene, *Innsett.r*, *Produksjon.r*, og *Husinfo.r*;

Innsett.r inneholder informasjon om hvert innsett, det vil si hvert produksjonskull med kyllinger. I dette datasettet finnes informasjon om slaktealder, slaktevekt, dato for innsett og slakting, antall kyllinger, dødelighet, fôrforbruk, med mer. Datasettet har en unik ID for hvert innsett "PK_Innsett" som kan kobles mot andre datasett, hvor den kobles mot "FK_Innsett". Hvert innsett har én rad med observasjoner. Det vil si et akkumuleringsnivå på ca. 30 dager (fra innsettsdato til slaktedato). Det er i alt registrert 4.672 innsett og 87 forskjellige variabler. Alle produsentene er lagt inn med en fiktiv ID for å opprettholde deres anonymitet.

Produksjon.r inneholder de daglige produksjonsdataene, som daglige vektmålinger av kyllingene, CO_2 - nivå i produksjonslokalet, luftfuktighet, og temperatur. Det er også tall på vannforbruk, fôrforbruk og dødelighet per dag. Akkumuleringsnivået er én rad per dag. Man har i alt da 157.547 antall registrerte dager og 68 variabler.

Datasettet har tidvis store mengder manglende data. Det er uvisst om det skyldes manglende måleutstyr, eller manglende manuell registrering.

Husinfo.r inneholder de tekniske spesifikasjonene ved kyllinghuset, som areal, varme-, lufting- og fôringssystem, med mer, samt byggeår og eventuelt renovasjonsår. De tekniske spesifikasjonene er stort sett konstant for samme produsent med unntak av hvis produksjonslokalene har vært renovert. I alt er det registrert husdata for 382 kyllingprodusenter, hvor noen har flere registreringer dersom de har renovert produksjonslokalene i løpet av de siste 30 årene.

I tabell 5.1 finner man en oversikt over de øvrige datasettene. Fullstendig oversikt over de ulike datasettene og deres variabler finnes i tillegg C.

Tabell 5.1: Oversikt over de viktigste datasettene, og deres innhold, som er behandlet i denne oppgaven. Det er også tatt med datasettenes dimensjoner for å gi et innblikk i datamengden.

Liste over datasett		
Datasett	**Innhold**	**Observasjoner og Variabler** $(n \times p)$
Aktivitet.r	Inneholder informasjon om aktiviteter produsenten utfører i kyllinghuset, slik som vasking og diverse trivselstiltak.	8419×14
Aktivitetstype.r	Kobles til "Aktivitet" for å gi mer informasjon om de ulike aktivitetene.	50×9
Avregning.r	Avregningsresultat, samt tråputepoeng, for hvert innsett.	4176×33
DaggamleKyllinger.r	Informasjon om de daggamle kyllingene som leveres til produsenten på dag én av innsettsperioden. Inneholder klokkeslett, produksjonstype, kyllinghybrid, anskaffelseskostnad og hvor mange kyllinger som døde ved ankomst.	6463×15
Fôrbeholdning.r	Informasjon om hvilke fôrtyper som de ulike innsettene har fått. Hvert innsett har flere rader, da de har mottatt ulike typer fôr gjennom oppfôringsperioden.	19390×15
Forblanding.r	Kobles til "Fôrbeholdning" for å gi mer informasjon om de ulike fôrtypene, blant annet fôrprodusent og fôrtype (start-,vekst-, eller sluttfôr).	135×6
Foreldredyr.r	Informasjon om alder til foreldrene til de daggamle kyllingene, samt gjennomsnittsvekt på de daggamle.	8707×9
Kassasjon.r	Data over kassasjonsårsaker ved slakt. Det vil si dyr som av ulike årsaker ikke er blitt akseptert for videre foredling og derfor kassert.	3862×33
LeveringTilSlakt.r	Informasjon om slaktetidspunkt med mer.	6669×21
Rugeri.r	Hvilket rugeri kyllingene kommer fra.	5×4
Vektgruppe.r	Antall kyllinger i hver vektgruppe, Viser fordelingen av flokken på ulike slaktevekter.	82343×13

Ved siden av de nevnte datasettene er det også en avregningsliste for de største produksjonstypene. Avregningslisten avgjør hvor mye produsentene får betalt per kilo etter hvor godt hver enkelt kylling treffer målvekten. Avregningsprisene endres én til flere ganger per år, og fôrprisene hver måned. Forutsetningene for dekningsbidraget endres dermed flere ganger i løpet av et år.

En utfordring som er gjennomgående for flere av datasettene er de store mengdene med manglende data. Manglene data skyldes manglende måleutstyr hos produsenten, eller at produsenten manuelt ikke har ført inn sine registrerte data i Noraturas registreringssystem, Fjørfekjøttkontrollen. I vedlegg C finnes det en fullstendig oversikt over andelen manglende data for de ulike variablene. Det har også vært en utfordring hvor de samme variablene, men fra ulike datasett, inneholde ulike registrerte verdier. På toppen av dette er det indikasjoner på at noen av variablene er ført på svært inkonsistent måte.

Vasket datasett for analyse

Datasettet (heretter kalt *Ferdig_ data*) som er brukt til imputering og videre analyse i denne oppgaven er en sammensetning av alle de nevnte datasettene, hvor det er valgt å beholde 39 av forklaringsvariablene. 19 av disse er kategoriske variabler med ulikt antall nivåer, mens de resterende 20 er kontinuerlige variabler. For de kategoriske variablene er det i noen tilfeller få observasjoner per nivå. Det er derfor valgt å sette en generell grense på maks 8 nivå per variabel, hvor de resterende nivåene er plassert i "annen"-gruppe/nivå. Dette er gjort for å redusere sjansen for type 1-feil, altså at man påstår effekt av forklaringsvariabel på responsvariabel i tilfeller hvor det i realiteten ikke er noen. Eksempelvis kan man tenke seg at det kun er én produsent som har en spesifikk type oppvarming. Dersom denne produsenten presterer godt/dårlig uten at noen av de andre forklaringsvariablene forklarer denne effekten, vil oppvarmingstype forsøke å forklare effekten. Man kan da risikere å påstå effekt av forklaringsvariabel på respons, hvor det i realiteten kan det være et utall faktorer, som ikke er registrert, som er den reelle årsaken til effekten. Ulempen ved å sette en øvre grense for antall nivå er at man mister informasjon - det kan være nettopp valg av oppvarmingstype som gjør at denne produsenten presterer godt/dårlig. 8 nivå er det høyeste antall nivå hvor en ikke opplever singularitet ved estimering av modellparameterne knyttet til datasettet.

Akkumuleringsnivå er lagt til innsett, og observasjoner som oppfattes som feil, eller mistenkelige, er fjernet. Dimensjonene til *Ferdig_ data* er 3564×39, men inneholder betydelige mengder manglende data. Etter imputering og vasking av data består datasettet (heretter kalt *fullstendig_ data*) av 3514 observasjoner og 39 variabler. En oversikt over forklaringsvariablene som er inkludert i dette datasettet finnes i tabell 5.2. Mer om bakgrunnen for utvelgelsen av både variabler og observasjoner, og vurderingene knyttet til denne prosessen finnes i delkapittel 6.2.

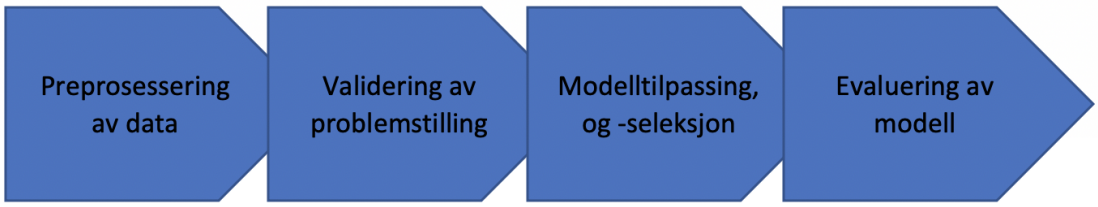
Tabell 5.2: Oversikt over de forklaringsvariablene som er inkludert i det endelige datasettet. Angir hvilken informasjon variabelen inneholder, samt om den er kontinuerlig eller kategorisk. Dersom variabelen er kontinuerlig er minimums- og maksimumsverdiene oppgitt, mens det for kategoriske variabler er oppgitt opprinnelig antall nivå.

Liste over forklaringsvariabler			
Variabel	**Innhold**	**Kategorisk/ kontinuerlig**	**Antall nivå / Min- og maksverdi**
Alder	Slaktealder	Kontinuerlig	28 - 63
AndelDøde	Andel døde ila. innsettet	Kontinuerlig	0.00 - 0.19
Antall	Antall kyllinger	Kontinuerlig	315 - 47700
AntallVinduer	Vinduer i produksjonslokalet	Kontinuerlig	0 - 14
AntMørkeperioder	Antall perioder uten lys	Kontinuerlig	1 - 6
Ant.Drikkenippler	Drikkenippler per kylling	Kontinuerlig	0.005 - 0.240
DrikkesystemMerke	Drikkesystemprodusent	Kategorisk	32
Fôrfirma	Fôrleverandør	Kategorisk	10
FôrsystemMerke	Fôringssystemprodusent	Kategorisk	21
Fylke	Lokasjon produsent	Kategorisk	11
LuftInn	Luftsirkulasjonssystem (inn)	Kategorisk	8
LuftUt	Luftsirkulasjonssystem (ut)	Kategorisk	4
Lyskilde	Lyskilde	Kategorisk	4
Lystimer	Antall timer med lys/døgn	Kontinuerlig	8 - 23
MaksLuftfuktighet	Høyeste reg. luftfuktighet	Kontinuerlig	30 - 100
MaksTemp	Maks temp. i prod.lokalet	Kontinuerlig	21 - 69
MinLuftfuktighet	Laveste reg. luftfuktighet	Kontinuerlig	3 - 75
MinTemp	Laveste reg. temperatur	Kontinuerlig	5 - 35
Oppvarmingskilde	Oppvarming i prod.lokalet	Kategorisk	23
Produksjonstype	Type produksjon	Kategorisk	7
Rugeri	Daggamle-produsent	Kategorisk	3
Slakteri	Hvor kyllingene er slaktet	Kategorisk	4
Slakteperiode	Tidsperiode for slakt	Kontinuerlig	3882 - 4536
Spillkopp	Spillkopp under vanningsnippel for å hindre vann på gulvet	Kategorisk	2
Startfôr	Startfôrforbruk per kylling	Kontinuerlig	0.038 - 4.248
Takkledning	Type takkledning	Kategorisk	15
Tråputepoeng	Velferdscore	Kontinuerlig	0 - 200
VannPerKylling	Vannkonsumet per kylling	Kontinuerlig	0.720 - 11.120
Varmekapasitet	Oppvarmingspotensialet	Kontinuerlig	15 - 485
Veggkledening	Type veggkledning	Kategorisk	14
Vekstfôr	Vekstfôrforbruk per kylling	Kontinuerlig	0.162 - 6.420
Vektvariasjon	Vektvariasjonen i slaktevekt	Kontinuerlig	12 - 47453
Vent.kapasitet	Vent.kapasitet per kylling	Kontinuerlig	2.719 - 24.773
Vent.systemMerke	Ventilasjonssystemprodusent	Kategorisk	14

Kapittel 6

Metode

Metodikken i denne oppgaven er delt opp i fire faser: preprosessering av datamateriell, validering av problemstilling, modelltilpasning og -seleksjon, og evaluering av modell. Preprosessering av datamaterialet innebærer opprydding og vasking av datasettet slik at det er klart for analyse. Videre testes det for signifikante forskjeller mellom produsentenes dekningsbidrag for å validere oppgavenes problemstilling. Datasettet deles så inn i trenings- og testsett, før ulike statistiske modeller tilpasses treningssettet. Treningsprosessen involverer også modellsammenligning, optimalisering, og valg av kompleksitetsparametere. Til slutt vil modellene ved hjelp av testsettet bli analysert og evaluert.



Figur 6.1: Oversikt over metodikkens fire faser.

6.1 Programvare

All programmering og plotting er gjort i RStudio versjon 1.2.1335 som baserer seg på programmet R versjon 3.6.1. R er et gratis, nedlastbart, programmeringsspråk for statistisk analyse og design. Det finnes også utallige tilleggspakker med forskjellige funksjoner man kan laste ned å benytte seg av. I denne oppgaven er det tatt i bruk pakkene *MissForest* (Stekhoven & Stekhoven, 2012) for imputering av data, og pakkene *leaps* (Lumley & Miller, 2017), *pls* (Mevik, Wehrens & Liland, 2019), *car*

(Fox & Weisberg, 2019), *mixlm* (Liland, 2018), *glmnet* (Jerome Friedman, 2010) for modelltilpasning og variabelseleksjon. For visualisering av resultater er *ggplot2* (Wickham, 2016) benyttet. Oppgaveteksten er skrevet i LaTeX.

6.2 Preprosessering av datamateriell

Da kyllingprodusentene kun *må* registrere fire variabler (*Selvdøde*, *Vannforbruk*, *Vekt* og *Fôrforbruk*) er mange av kolonnene i de ulike datasettene ufullstendige. Siden registreringen gjøres manuelt finnes det flere åpenbare feilføringer som er enkle å oppdage, mens det i andre tilfeller er det vanskelig å avgjøre om verdiene er feilføringer eller avvikende verdier. Det er også tilfeller av inkonsistent bruk av størrelsesorden som gjør det vanskelig å bedømme noen av variablene. Som nevnt i kapittel 5 er akkumuleringsnivået lagt til innsett, slik at hvert innsett har sin egen rad i datasettet som skal analyseres. De datasettene som har hatt andre akkumuleringsnivå er transformert slik at de alle har samme antall observasjoner, n. I tilfeller hvor man eksempelvis har temperaturmålinger for hver dag, er maksimums- og minimumsverdiene for innsettet beholdt. En oversikt over alle vurderingene som er gjort knyttet til de ulike variablene i datasettene finnes i vedlegg C. Vurderingene baserer seg på fremgangsmåten beskrevet i delkapittel 3.1.

Uønskede verdier
I de datasettene som inneholder dobbeltføringer er det fjernet de verdiene som er naturlige å fjerne, enten fordi de er nøyaktig kopier, eller fordi det er forsøkt å rette opp en tidligere dobbeltføring ved å legge til en rad/observasjon med negative verdier. I tilfeller hvor det åpenbart er umulige verdier, er disse fjernet. Dersom det har vært tvil rundt noen av verdiene, er det valgt å beholde disse. Forklaringsvariabler hvor man umulig kan tenke seg noen forklaringsverdi knyttet til dekningsbidrag 1 (DB1) er også fjernet, da disse kan komme i skade for å forklare noe av den tilfeldige støyen i observasjonene (se delkapittel 4.1.3). Forklaringsvariabler hvor spredningen i registrerte verdier er urealistisk stor, er det i samråd med Nortura konkludert med at det ikke har vært konsistent bruk av dataenes størrelsesorden. Disse forklaringsvariablene er også fjernet fra videre analyse.

Avviksverdier
Hva som er blitt definert som avviksverdier har variert mellom de ulike variablene. I de fleste tilfellene er det kun fjernet verdier dersom det har vært klar mistanke om umulige observasjoner. I de resterende tilfellene hvor man kan tenke seg en tydelig sammenheng mellom to eller flere variabler, er verdiene vurdert som avviksverdier dersom deres standardiserte residualer har en absoluttverdi på over 3. Det er benyttet tradisjonell OLS (se delkapittel 4.3.1) for å lage modellene for å estimere de

standardiserte residualene. Det er også, i samråd med Nortura, fjernet observasjoner hvor dekningsbidraget (respons) er under 2 kroner per kylling. Bakgrunnen for at disse observasjonene er fjernet skyldes uregelmessigheter, det være seg feilføringer, produksjonsfeil, ekstrem smitte/dødelighet, eller lignende, i de aktuelle innsettene.

Skrivefeil
Skrivefeil kan være vanskelig å oppdage, og eventuelt verifisere. Terskelen for å fjerne observasjoner som følge av mistanke om skrivefeil, er derfor svært høy og det er fjernet få verdier som konsekvens av dette. I flere tilfeller ser man at produsentene selv har forsøkt å rette opp feilene.

6.2.1 Manglende data

I prinsippet vil store mengder manglende data føre til økt sjanse for prediksjonsfeil, noe som igjen vil gjøre det vanskeligere å oppdage eventuelle reelle sammenhenger mellom forklaringsvariabler og respons (se delkapittel 3.2). Siden maksimalt anbefalt andel manglende data er avhengig av årsaken til deres eksistens, er det sett etter sammenheng mellom manglende data og responsvariabelen, DB1. Det er laget en matrise hvor de manglende dataene er markert og responsvariabelen er sortert i stigende rekkefølge for å se etter mønstre i de manglende dataene. Variabler som bryter Missing at random (MAR) antagelsen og/eller har høy andel manglende data er fjernet fra videre analyse.

6.2.2 Imputering av data

Etter å ha vasket dataene og oppnådd ønsket struktur er de ulike datasettene satt sammen til ett felles datasett, *Ferdig_ data*, før *MissForest* ble tatt i bruk. *MissForest* benytter prinsippene ved Random Forest for å imputere verdier for manglende data. Datasettene ble samlet før imputering for å bedre informasjonsgrunnlaget og derav imputeringskvaliteten til modellen. Imputeringsmetoder som baserer seg på PCA og Multippel korrespondanseanalyse ble også vurdert, men grunnet tidligere studier ((Kokla et al., 2019), (Starkweather, 214), (Stekhoven & Bühlmann, 2011)) er det valgt å gå videre med *MissForest*. Fordelen med *MissForest* er at den fanger opp komplekse interaksjoner mellom ikke-lineære variabler av ulik skala og type - både kategoriske og kontinuerlige (Stekhoven & Bühlmann, 2011).

For å bestemme de optimale kompleksitetsparameterne, antall trær i hver skog, og variabler ved hver node, er det gjennomført over 40 imputeringsrunder. Dette har resultert i over 40 imputerte datasett. OOB-scoren er benyttet for å evaluere og sammenligne kvaliteten til de imputerte dataene i de ulike datasettene. De kompleksitetsparameterne som gir lavest OOB-score danner grunnlaget for valg

av kompleksitetsparameterne som er benyttet i imputeringsmodellen for å oppnå datasettet *Fullstendig_ data*. For å finne den optimale skogstørrelsen, er det sammenlignet skogstørrelser på mellom 1-100 trær i hver skog, hvor antall variabler holdes konstant. Selv om det ikke vil skade å tilpasse flere trær i hver skog vil det kun føre til små forbedringer i modellens prestasjon (Probst & Boulesteix, 2017). Det er av praktiske årsaker derfor valgt å sette en øvre begrensning ved 100 trær. Modellen stopper å lete etter bedre imputeringsmodeller enten etter 10 iterasjoner, eller ved at de imputerte verdiene gir høyere OOB-score enn ved forrige iterasjon (Stekhoven & Stekhoven, 2012). Det er også undersøkt hvordan antallet variabler ved hver node påvirker OOB-scoren, da ved konstant skogstørrelse. Antall variabler som er inkludert ved hver node er generell og gjelder alle nodene i en gitt modell. Det er sammenlignet modeller med mellom 1-50 variabler ved hver node. Det er valgt å ikke sette en øvre begrensning for antall noder i hvert tre, slik at modellen selv kan optimalisere denne parameteren. Som konsekvens kan noen trær ha blitt overtilpasset (Friedman et al., 2001).

Etter imputering, med valgte kompleksitetsparametere, er datasettet *Fullstendig_ data* vasket etter beste evne for tilsynelatende uriktige verdier som følge av imputeringen. Disse observasjonene er da fjernet i sin helhet fra datasettet. Deretter er de variable kostnadene standardisert for å oppnå et mest mulig objektivt resultat for alle kyllingprodusentene. Gjennomsnittsprisen per kilo fôr er delt inn i start-, vekst-, og sluttfôr og videre gruppert etter fôrprodusent. Det vil si alle produsentene som benyttet samme fôrleverandør er antatt å betale samme pris for de respektive fôrtypene. Dette er gjort for å unngå sesongvariasjoner, samt skalafordeler - eksempelvis kan man tenke seg at storskalaproduksjon har gunstigere innkjøpspriser enn småskalaproduksjon. Det samme er gjort for anskaffelseskostnaden av daggamle kyllinger hvor gjennomsnittsprisen per kylling er delt inn etter hybridtype. For plukkekostnad er Norturas egne satser for de ulike produksjonstypene tatt i bruk. Det er både fordeler og ulemper ved å standardisere kostnadssatsene. Fordelen er at man eliminerer eventuelle konkurransefortrinn mellom produsentene, ulempen er at man potensielt mister informasjon. Siden man har konsesjonsgrense (se delkapittel 1.1) i Norge, og derav ikke fleksibilitet til å øke produksjonsstørrelsen om man skulle ønske det, er det vurdert som hensiktsmessig å standardisere kostnadssatsene.

6.3 Validering av problemstilling

Norturas påstand er at de samme kyllingprodusentene stabilt presterer bedre/dårligere enn andre produsenter, gitt samme tidsperiode og daggamle kyllinger. Dekningsbidrag 1 (DB1) er benyttet som mål på produsentens suksess. Dekningsbidraget er satt til kroner per innsatte kylling ved innsettstart for å kunne

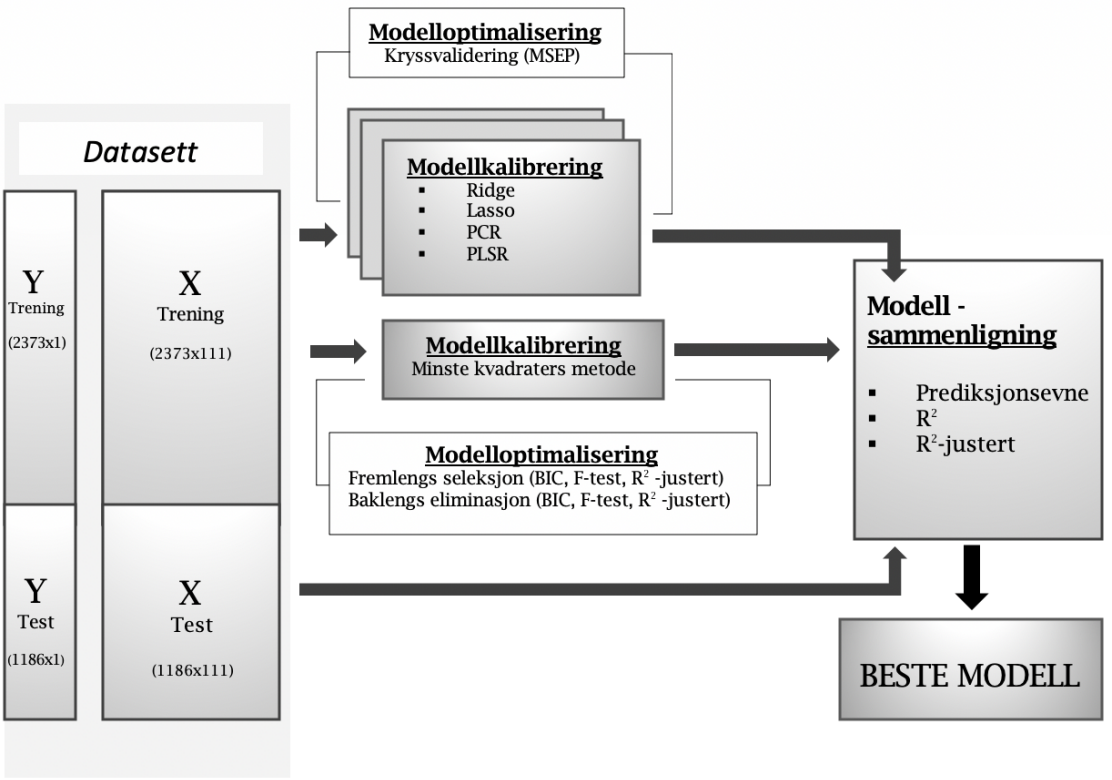
sammenligne produksjoner av ulik skala. For å verifisere påstanden til Nortura er det gjennomført en ANOVA for å se om det finnes belegg for å kunne påstå effekt av produsent på DB1. Deretter er ulike blokkfaktorer introdusert for å verifisere at produsent har effekt på DB1, selv ved relativt like forutsetninger. Nortura mener enkelte produksjonstyper er mer lønnsomme enn andre, men ønsker å beholde produksjonsmangfoldet som finnes i dag, derfor er produksjonstype, sammen med slakteri, rugeri og produksjonsdato introdusert som blokkfaktorer for å eliminere eventuelle effekter disse variablene har på dekningsbidraget.

Etter ANOVA-analysen ble det utført en Tukey-test for å se mellom hvilke produsenter det er signifikant forskjell i dekningsbidrag, og om det finnes grunnlag for Norturas påstand. Testens signifikansnivå er satt til 0.05.

6.4 Modelltilpasning og evaluering av modeller

Dataene er delt inn i treningssett (2/3 av observasjonene) og testsett (1/3 av observasjonene). Observasjonene for henholdsvis trenings- og testsett er trukket tilfeldig fra det vaskede imputerte datasettet, *Fullstendig_ data*. De kategoriske variablene er gjort om til dummyvariabler, det vil si at hver kategori/nivå har sin egen kolonne bestående av verdiene 1 og 0, alt ettersom kategorien er tilstedeværende eller ikke. Ved å gjøre de kategoriske variablene om til dummyvariabler er datasettet utvidet fra 39 variabler til 111 variabler. Videre er modellparameterne estimert ved bruk av treningssettet og tilpasningsmetodene beskrevet i kapittel 4.3. Ved bruk av Ridge, Lasso, PCR eller PLSR, er MSEP og/eller R^2_{pred} ved kryssvalidering benyttet som evalueringskriterium for å finne den optimale kompleksitetsparameteren for de ulike modellene. Det er også tilpasset en PLSR modell hvor det er benyttet VIP score med nedre terskel på 0.83 for variabelseleksjon (se kapittel 4.3.3). Ved OLS er F-test, BIC og $R^2_{justert}$ benyttet som utvelgelseskriterier ved forlengs variabelseleksjon og baklengs eliminasjon. Den fullstendige OLS-modellen er også beholdt for videre analyse.

Etter tilpasning og optimalisering av de ulike modellene er deres prediksjonsevne testet på testsettet for å avgjøre hvilken modell, med tilhørende modellparametere, som predikerer DB1 best. RMSE og $R^2_{justert}$ er benyttet som evalueringskriterier for å bestemme den beste modellen. Figur 6.2 viser en oversikt over prosessen.



Figur 6.2: Oversikt over modelltilpasning- og evalueringsprosessen for de ulike statistiske metodene. Den modellen som etter evalueringskriteriene er best, beholdes for videre analyse.

Kapittel 7

Resultat

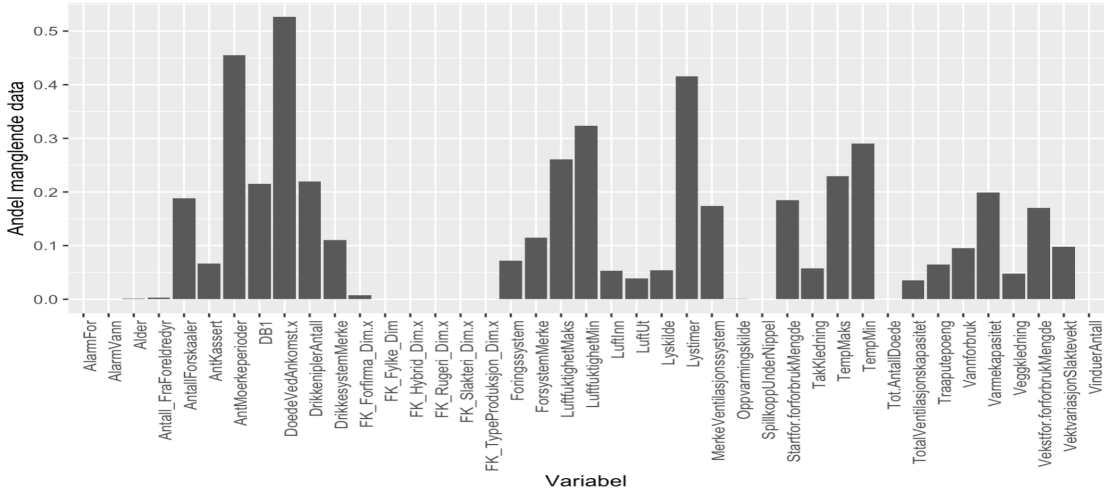
Dette kapittelet presentere resultatene fra de ulike fremgangsmåtene og analysene presentert i kapittel 6. Resultatene vil bidra til å kunne trekke kvalifiserte konklusjoner rundt oppgavens problemstilling.

7.1 Missing at random

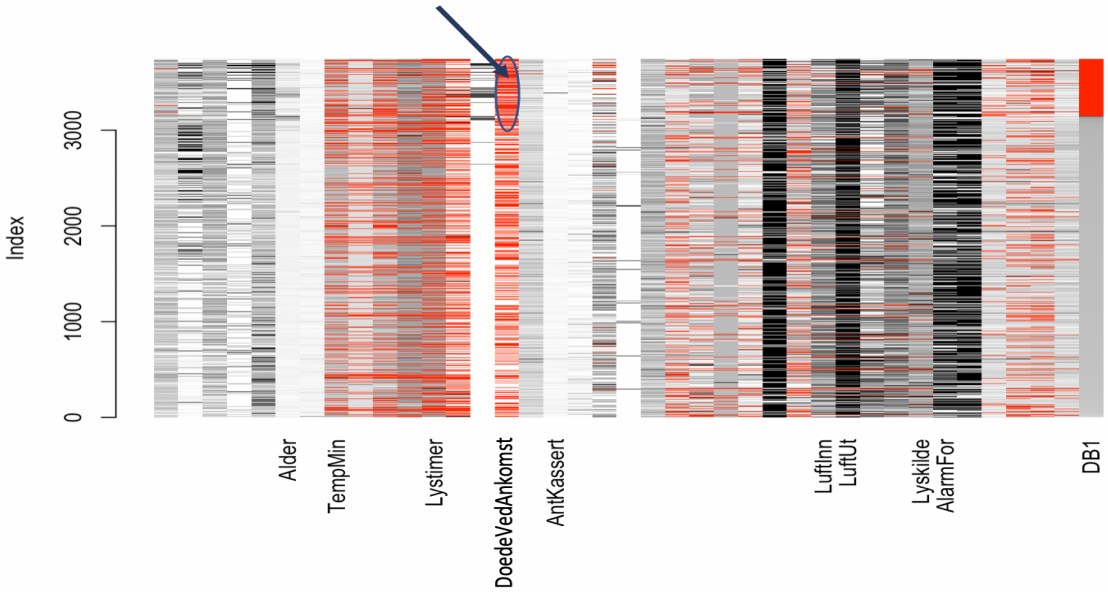
En av hovedutfordringene i denne oppgaven er de store mengdene manglende data. Figur 7.1 viser en oversikt over andelen manglende data for de ulike variablene i datasettet *Ferdig_ data*. Som tidligere nevnt er det viktig å undersøke årsaken til de manglende dataene da dette påvirker hvordan en bør behandle de videre (se delkapittel 3.1).

Matrisen i figur 7.2 illustrerer alle cellene i datasettet *Ferdig_ data*. Observerte data er vist som en kontinuerlig skala fra grå til svart (mørkere farge tilsvarer høyere verdi), mens manglende data er røde. Dataene er sortert etter responsvariabelen, DB1, for enklere å avdekke eventuelle mønster mellom manglende data og DB1. På bakgrunn av resultatene (figur 7.2) kan det se ut til at forklaringsvariabelen *antall døde ved ankomst* (markert med pil), har høyere frekvens manglende data i tilfeller hvor også DB1 mangler, og dermed bryter Missing at random (MAR) antagelsen. Siden *antall døde ved ankomst* også mangler over 50% av observasjonene (se figur 7.1) er variabelen fjernet fra videre analyse. Ut ifra matrisen ser de resterende manglende dataene ut til å være tilfeldig spredt, uten sammenheng mellom forklaringsvariabel og respons. Det er derfor antatt at MAR-antagelsen holder for de resterende variablene. Det vil si at de manglende verdiene ikke kan relateres til responsvariabelen etter at man har inkludert de andre prediksjonsvariablene. Det valgt å tillate 30% manglende data for de resterende forklaringsvariablene, med unntak av variablene, *lystimer* og *mørketimer*, hvor det er tillatt opp mot 50 % manglende data. Årsaken til at disse variablene vurderes annerledes er deres felles

relasjon som gjør at man kan beregne de manglende verdiene dersom én av de nevnte variablene er kjent. Variabelen *AntMørkeperioder* er også fjernet fra videre analyse med en andel manglende data på rundt 45 %.



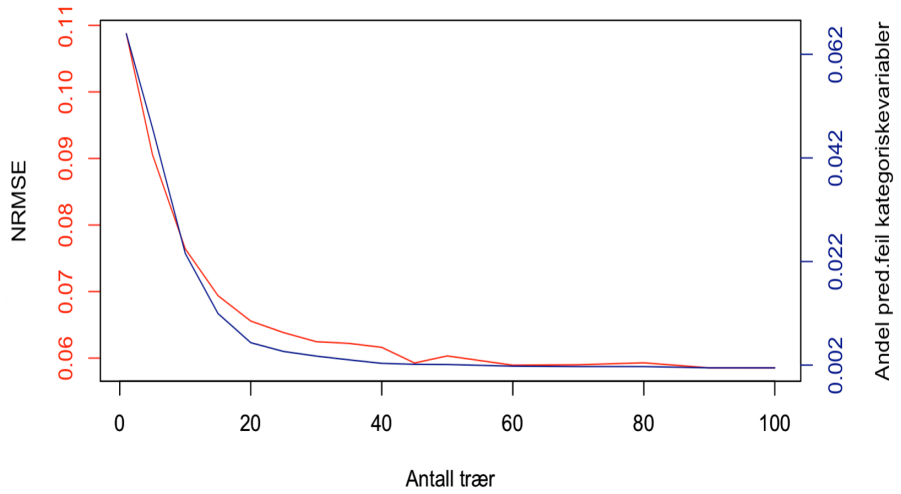
Figur 7.1: Oversikt over andelen manglende data for de ulike variablene i datasettet *Ferdig_ data*. Variablene *Lystimer*, *Antall døde ved ankomst*, og *AntMørkeperioder*, mangler alle over 40% av observasjonene.



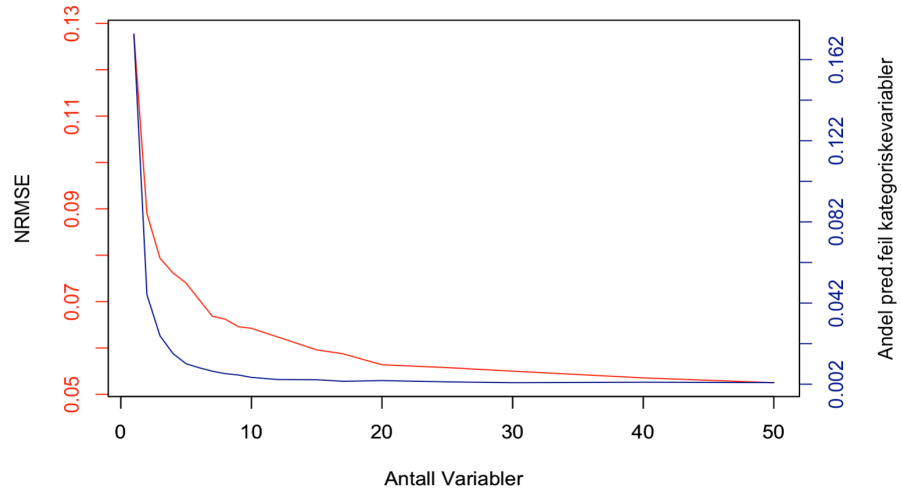
Figur 7.2: Oversikt over observerte data og manglende data (rød). De observerte dataene er skalert fra lav til høy (lys til mørk). Observasjonene for de ulike forklaringsvariablene er sortert etter størrelsen på responsvariabelen, DB1. Man har lavest DB1 nederst til høyre, før det gradvis stiger oppover. Siden det relativt sett er marginale forskjeller i DB1, sett i forhold til noen av forklaringsvariablene, ser denne skalaen relativt lik ut (grå) for hele utfallsrommet. Øverst finnes de radene hvor DB1 mangler (rød). Det kan se ut som variabelen "antall døde ved ankomst"har tettere frekvens (blå pil) manglende data i tilfeller hvor det også mangler DB1.

7.2 Imputering av data

Siden de manglende dataene tilsynelatende oppfyller MAR-antagelsen, samt ikke overstiger 30 %, anses det som forsvarlig å imputere de manglende dataene. Figur 7.3 og 7.4 viser henholdsvis hvordan ulike skogstørrelser, og ulikt antall variabler ved hver node, påvirker OOB-scoren. I figur 7.4 ser man at prediksjonsfeilen reduseres svært lite ved å inkludere flere enn 40 variabler ved hver node. Man kan også se at modellen raskere reduserer imputeringsfeilen for kategoriske variabler, enn for de kontinuerlige. Det er likevel valgt å sette en grense ved 20 variabler ved hver node i imputeringsmodellen som er benyttet for å imputere data til datasettet *Fullstendig_datasett*. Dette er gjort i et for forsøk på å nøytralisere eventuell skjevhet i dataene som konsekvens av overtilpasning av imputeringsmodellen. Ved å inkludere flere variabler ville man redusert OOB-scoren ytterligere (se figur 7.4), men økt risikoen for at de imputerte dataene forsterket de observerte dataenes iboende sammenhenger, selv i tilfeller hvor det i realiteten ikke er noen. Antall trær i hver skog er satt til 60 da det tilsynelatende ikke forbedrer imputeringsmodellen nevneverdig å tilpasse flere trær (se figur 7.3). Bakgrunnen for begrensningene som er gjort er en avveining mellom tidsbruk, modellens forbedringspotensial, og risikoen for overtilpasning.



Figur 7.3: Viser hvordan imputeringsfeilen for kategoriske (blå linje og høyre y-akse) og kontinuerlige (rød linje og venstre y-akse) varierer som følge av antall tilpassede trær i hver skog.



Figur 7.4: Viser hvordan imputeringsfeilen for kategoriske (blå linje og høyre y-akse) og kontinuerlige (rød linje og venstre y-akse) varierer som følge av antall variabler ved hver node i hvert tre. Det er kun tilpasset 20 trær i hver skog.

7.3 Validering av problemstilling

For å validere problemstillingen ble en ANOVA gjennomført for å sammenligne effekten kyllingprodusent har på dekningsbidrag 1 (DB1). Den tilpassede modellen kan da skrives som,

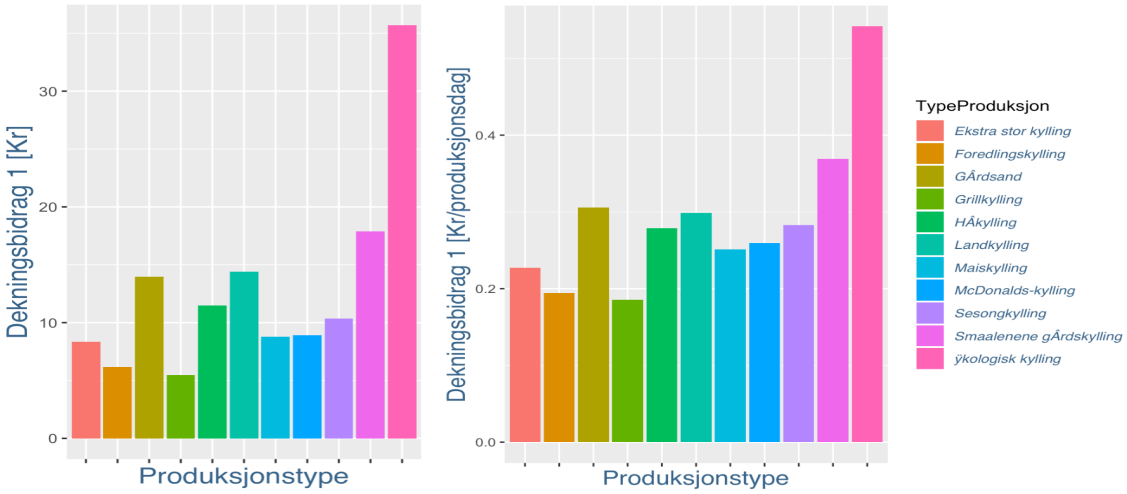
$$\begin{aligned} \hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_{324} X_{324} \\ &= \hat{\beta}_0 + \hat{f}(X) \end{aligned}$$

hvor $\hat{\beta}_0 = \bar{Y}$, og $\hat{\beta}_j$ er den estimerte effekten produsent j har på dekningsbidraget, Y. Det er totalt 324 ulike produsenter. Analysen av variansen viste at effekten produsent, uten blokkfaktorer, har på DB1 er signifikant med P-verdi $< 1.0e^{-10}$.

Av figur 7.5a og 7.5b er det tydelige forskjeller i gjennomsnittlig dekningsbidrag for ulike produksjonstyper.

For å hensyn ta effekten produksjonstype har på DB1 er denne variabelen introdusert som blokkfaktor, sammen med produksjonsdato, slakteri og rugeri. Den nye modellen blir da seende slik ut,

$$\hat{Y} = \hat{\beta}_0 + \hat{B}_1(X_j) + \hat{B}_2(X_k) + \hat{B}_3(X_l) + \hat{B}_4(X_m) + \hat{f}(X_n)$$



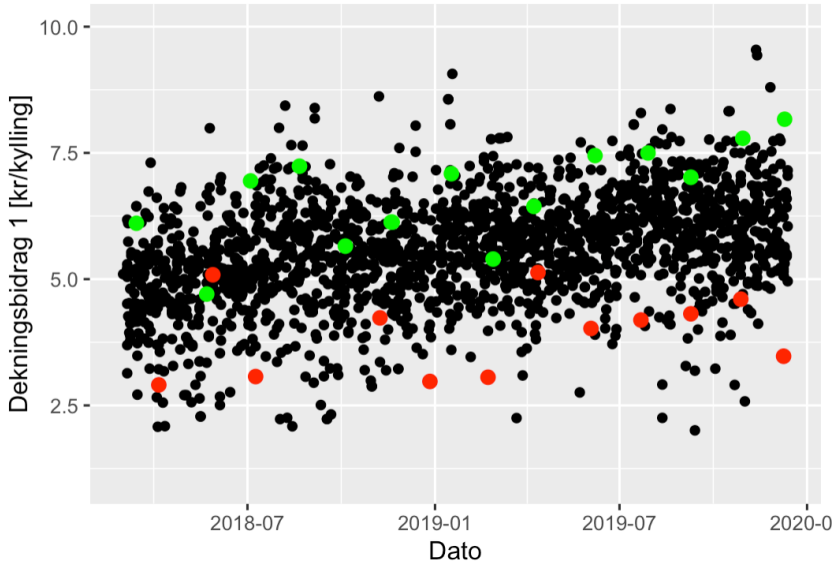
(a) Dekningsbidrag per kylling for ulike produksjonstyper.

(b) Dekningsbidrag per kylling delt på produksjonslengde (dager).

Figur 7.5: Det gjennomsnittlige dekningsbidraget for ulike produksjonstyper. Siden enkelte produksjonstyper har betydelig lengre produksjonstid enn andre, er det i figur (b) tatt hensyn til dette ved å dele på antall produksjonsdager, altså innsettet varighet.

hvor $\hat{B}_1(X_j)$, $\hat{B}_2(X_k)$, $\hat{B}_3(X_l)$, $\hat{B}_4(X_m)$ er den estimerte effekten til henholdsvis produksjonstype, produksjonsdato, slakteri og rugeri, på DB1. $\hat{f}(X_n)$ er estimert effekt produsent har på DB1 etter å ha inkludert blokkfaktorene i modellen. Analysen av variansen viste at produsent fortsatt har signifikant påvirkning på DB1 med P-verdi $< 1.0^{-10}$.

En Tukey-test ble gjennomført for å undersøke mellom hvilke produsenter man kan se signifikant forskjell i DB1. Figur 7.6 illustrerer DB1 for alle Norturas produsenter som driver med produksjonstypen "foredlingskylling" i perioden 01.01.2018 - 01.12.2019. Siden det vil være upraktisk å vise resultatet av sammenligningen mellom de over 300 ulike produsentene er kun ett av de mange resultatene hvor det er signifikant forskjell mellom produsenters DB1 presentert her (se figur 7.6). De uthevede observasjonene tilhører henholdsvis produsent nr.495 (grønn) og produsent nr 391 (rød). Tukey-testen viste signifikant forskjell mellom de to produsentenes gjennomsnittlige dekningsbidrag, med en forskjell på 3.31 kroner per kylling, resulterende i en q-verdi $= 1.45e^{-6}$ (se ligning 4.36).



Figur 7.6: Viser DB1 (kroner/kylling) for foredlingskyllingprodusenter i perioden 01.01.2018 - 01.12.2019. De uthevede observasjonene er ett av resultatene fra Tukey-testen, hvor det er påvist signifikant forskjell mellom to produsenter. Tukey-testen ble gjennomført etter at ANOVA, med blokkfaktorer, indikerte signifikant effekt av leverandør på DB1. Man kan tydelig se at produsent nr. 495 (grønn) generelt, og stabilt, har betydelig høyere dekningsbidrag enn produsent nr. 391 (rød).

7.4 Modelltilpasning og -seleksjon

Den tilpassede lineære modellen for å beskrive produsentenes dekningsbidrag (DB1) kan skrives som,

$$DB1 = \hat{\beta}_0 + \hat{f}(X) \tag{7.1}$$

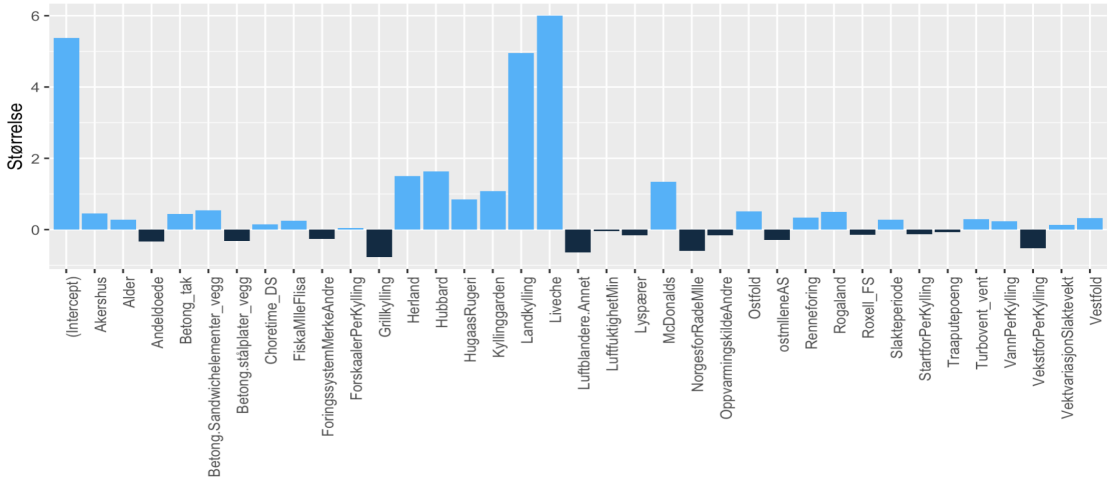
hvor f er en lineærfunksjon av regresjonskoeffisientene β til alle forklaringsvariablene med deres tilhørende nivåer (se tabell 5.2 for oversikt over de ulike variablene). Siden det i alt er 111 variabler, dummyvariabler inkludert, er det i et forsøk på å redusere antall forklaringsvariabler blant annet benyttet forlengs seleksjon og baklengs eliminasjon for modelltilpasning.

7.4.1 Variabelseleksjon

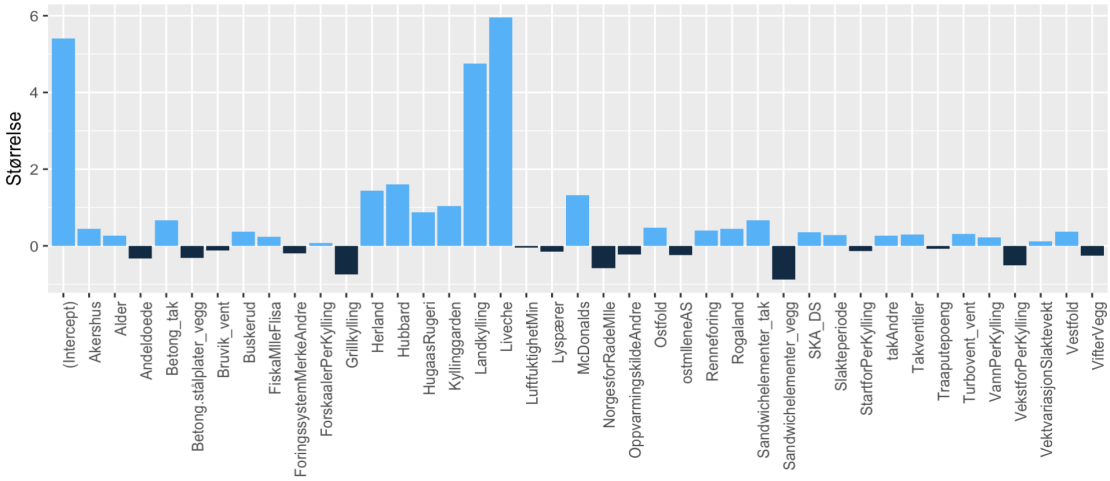
Det er benyttet modelltilpasning ved forlengs seleksjon og baklengs eliminasjon basert på modellevalueringskriteriene F-testing, $R^2_{justert}$, samt BIC (delkapittel 4.2 og 4.3.2).

F-test

Ved bruk av F-test med signifikansnivå, $\alpha = 0.05$, ble det inkludert 36 og 40 variabler for henholdsvis forlengs seleksjon og baklengs eliminasjon. Figur 7.7 og 7.8 viser en oversikt over modellkoeffisientene som er inkludert i de to modellene. Disse er inkludert her for å gi leseren en følelse av de estimerte modellkoeffisientenes størrelsesorden. Videre vises det til vedlegg A for alle de ulike modellenes koeffisienter.



Figur 7.7: Plottet viser de estimerte modellkoeffisientene etter forlengs variabelseleksjon. De 36 variablene som nå er inkludert i modellen er lagt langs x-aksen. F-test med signifikansnivå 0.05 er benyttet som seleksjonsmetode.

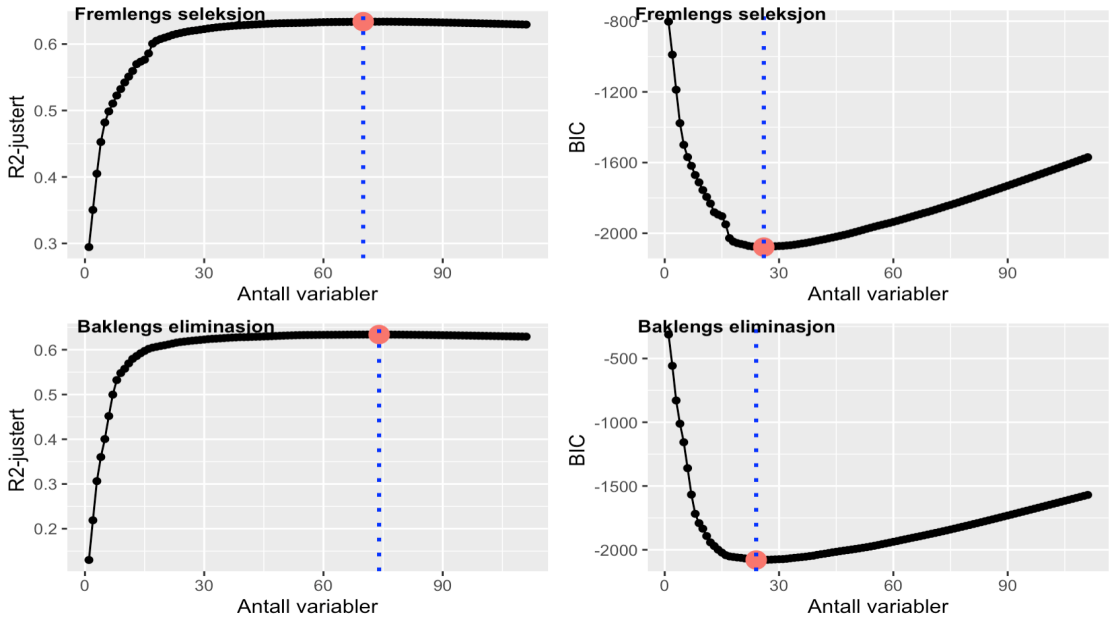


Figur 7.8: Plottet viser de estimerte modellkoeffisientene etter baklengs variabelseleksjon. De 40 variablene som nå er inkludert i modellen er lagt langs x-aksen. F-test med signifikansnivå 0.05 er benyttet som seleksjonsmetode.

BIC og $R^2_{justert}$

Ved å benytte $R^2_{justert}$ som evalueringskriterium inneholder den optimale modellen 71 og 75 forklaringsvariabler for henholdsvis forlengs seleksjon og baklengs eliminasjon, men som man kan se av figur 7.9 er det svært små endringer i $R^2_{justert}$ etter rundt 30 variabler. En fullstendig oversikt over hvilke variabler som er inkludert i de ulike modellene finnes i vedlegg A.

Bayesian informasjonskriterium (BIC) er et strengere evalueringskriterium (se ligning 4.13) som normalt vil redusere antall variabler som inkluderes i modellen ytterligere. Figur 7.9 viser hvordan BIC utvikler seg som funksjon av antall variabler som er inkludert. Ved forlengs seleksjon inkluderes 27 variabler, mens det for baklengs eliminasjon kun inkluderes 25. De blå stiplede linjene markerer de modellene som minimerer BIC-scoren. Som man kan se av grafene øker BIC-scoren betydelig ved å inkludere flere variabler sammenlignet med den optimale modellen.

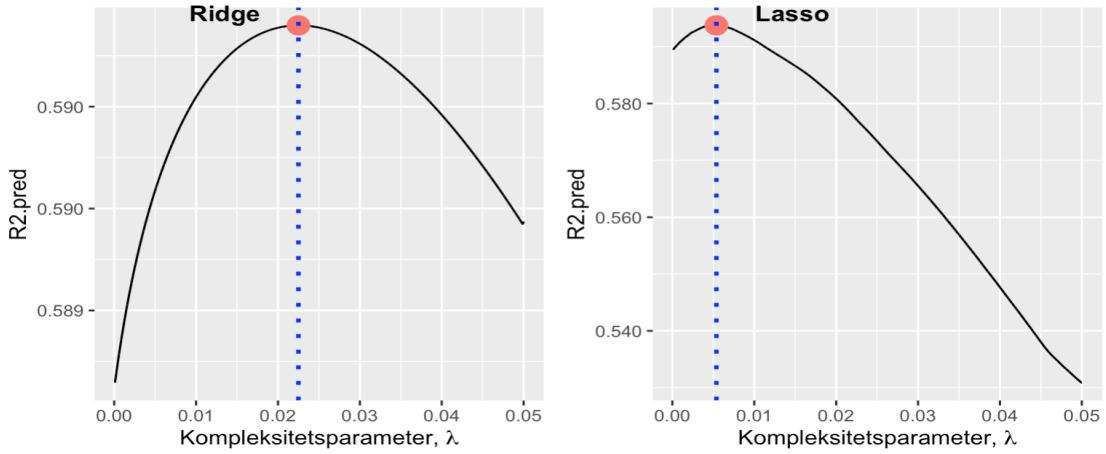


Figur 7.9: Illustrerer hvordan $R^2_{justert}$ og BIC varierer ved å øke/redusere antall forklaringsvariabler ved forlengs seleksjon og baklengs eliminasjon. Den blå stiplede linjen markerer antall forklaringsvariabler hvor $R^2_{justert}$ er på sitt høyeste / BIC-scoren er på sitt laveste. Ved bruk av $R^2_{justert}$ som evalueringskriterium ble 71 og 75 variabler inkludert i modellene, mens det ved bruk av BIC ble inkludert 27 og 25 variabler, for henholdsvis forlengs seleksjon og baklengs eliminasjon.

7.4.2 Krympingsmetoder

Som tidligere nevnt utfører ikke ridge variabelseleksjon, likevel fungerer metoden godt dersom noen av de virkelige parameterne er null (se delkapittel 4.3.4). For å finne den optimale kompleksitetsparameteren λ, er det benyttet kryssvalidering med 10 oppdelinger (K = 10) for å estimere R^2_{pred} for λ mellom 0 og 100. Ligning 4.12 er benyttet for å begrene R^2_{pred}. Figur 7.10 viser hvordan R^2_{pred} utvikler seg som funksjon av λ. Figuren viser kun λ-verdier fra 0-0.05, da grafen ikke vil ha vendepunkt utenfor dette området. Optimal λ for ridge regresjon er 0.022 hvor R^2_{pred} er benyttet som evalueringskriterium.

I motsetning til ridge, kan lasso krympe modellparametere til null og kan derfor brukes til variabelseleksjon (se delkapittel 4.3.4). Tilsvarende som for ridge er det benyttet kryssvalidering med 10 oppdelinger for å estimere R^2_{pred} for å finne den optimale kompleksitetsparameteren. Optimal λ for lasso er 0.005 ved R^2_{pred} som evalueringskriterium. Som konsekvens krympes 29 koeffisienter til null. Oversikt over de estimerte koeffisientene for ridge og de gjenværende koeffisientene for lasso finnes i vedlegg A. I figur 7.10 kan man se hvordan λ påvirker modellens R^2_{pred}.



Figur 7.10: Illustrerer hvordan R^2_{pred} utvikler seg for ulike modelltilpasninger hvor man varierer kompleksitetsparameteren λ for tilpasningsmetodene ridge og lasso. Merke at y-aksene til de to grafene er svært ulike.

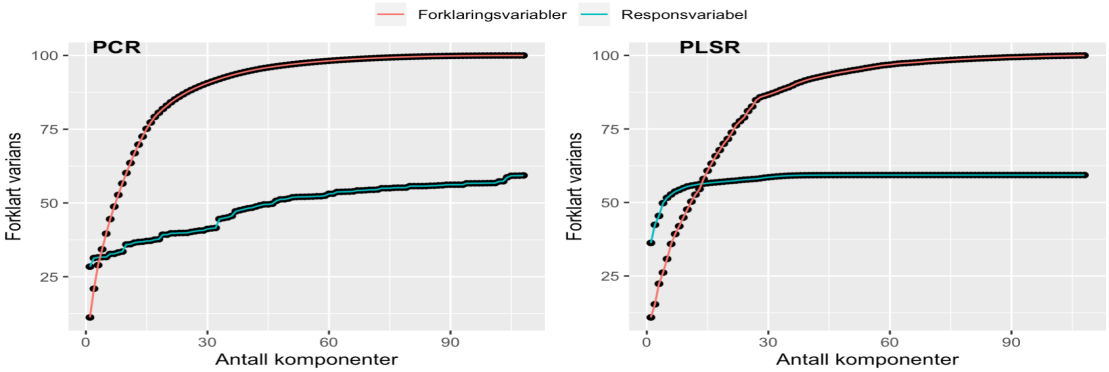
7.4.3 Dimensjonsreduserende metoder

Ved å benytte Prinsipalkomponentregresjon (PCR) danner man et nytt sett med ortogonale, ukorrelerte forklaringsvariabler, også kalt prinsipalkomponenter, før man benytter OLS på et utvalg av komponentene (se delkapittel 4.3.3). Prediksjonsmodellen basert på prinsipalkomponentene i stede for de originale variablene, reduserer ikke bare modellens kompleksitet, men også eventuelle problemer med multikollinearitet. Figur 7.11 illustrerer hvor mye av variansen i forklaringsvariablene og i DB1 (responsvariabelen) som forklares som funksjon av antall prinsipalkomponenter. Variansen i forklaringsvariablene forklares raskere enn for responsvariabelen, dette er ikke unaturlig da prinsipalkomponentene ved PCR tar sikte på å forklare mest mulig av variansen til nettopp forklaringsvariablene. Av figur 7.11 kan man derimot se at det trengs mange komponenter for å nå dataenes forklaringspotensial knyttet til DB1.

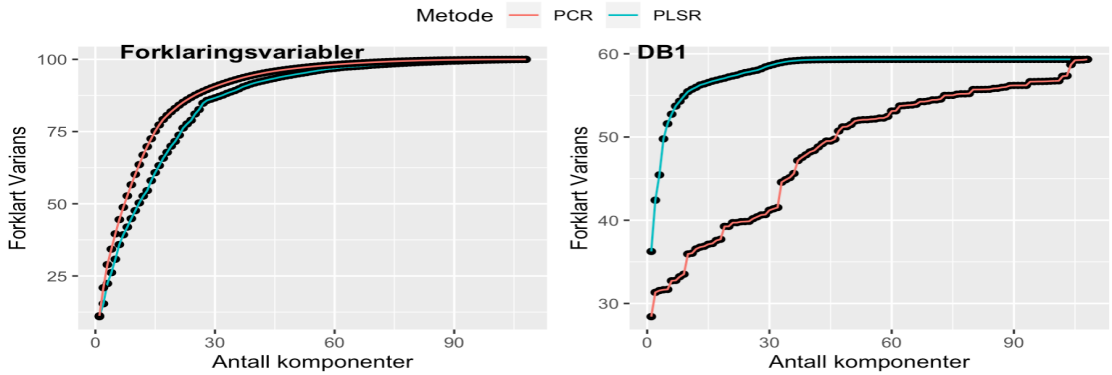
Ved Partial least square regresjon (PLSR) dannes komponentene ikke bare ved å ta hensyn til variansen i forklaringsvariablene, men også ved å ta hensyn til variansen til responsvariabelen. Som man kan se av figur 7.11 er det derfor ikke uventet at man trenger færre komponenter for å nå dataenes forklaringspotensial knyttet til DB1. Figur 7.12 sammenligner forklaringsevnen til komponentene knyttet til forklaringsvariabler og respons for de to metodene. Av figuren ser man at PCR trenger færre komponenter for å forklare variansen til forklaringsvariablene, mens PLSR trenger færre komponenter for å forklare variansen til responsen. Ved å inkludere nok komponenter vil begge modellene forklare like mye av variansen både for forklaringsvariabler og respons.

For å finne det antallet komponenter som optimaliserer modellens prediksjonsevne er det benyttet kryssvalidering med MSEP som evalueringskriterium. For PLSR oppnår man lavest MSEP ved 57 komponenter, mens ved PCR inkluderes hele 110 komponenter for å danne den optimale modellen. Likevel, som man kan se av MSEP-estimatenes usikkerhet i figur 7.13, kan man muligens klare seg med enda færre komponenter og oppnå samme resultat, spesielt ved PLSR. Modellkoeffisienter hvor det er benyttet henholdsvis 57 og 110 komponenter for å estimere koeffisientenes størrelse finnes i vedlegg A.

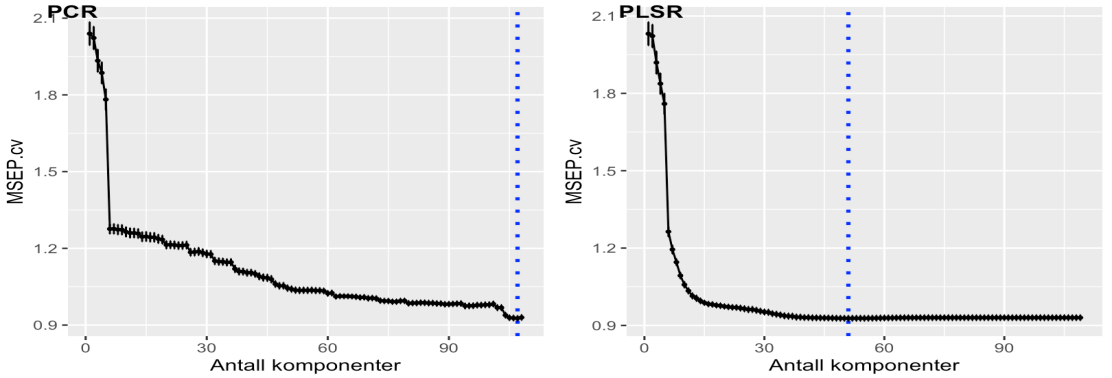
I tillegg til de nevnte modellene er det også tilpasset en modell hvor det er benyttet VIP score for å foreta variabelseleksjon. Terskelen for om en forklaringsvariabel skal inkluderes er satt til 0.83. I alt ble 24 forklaringsvariabler inkludert i denne modellen.



Figur 7.11: Viser hvor mye av variansen som er forklart gitt antall komponenter som er inkludert i modelltilpasningen for PCR (venstre graf) og PLSR (høyre graf). Rød linje representerer forklaringsvariablene, men grønn linje representerer responsen (DB1).



Figur 7.12: Sammenligner PLSR og PCR og deres forklaringsevne knyttet til både forklaringsvariablene (venstre graf) og responsvariabelen (høyre graf).

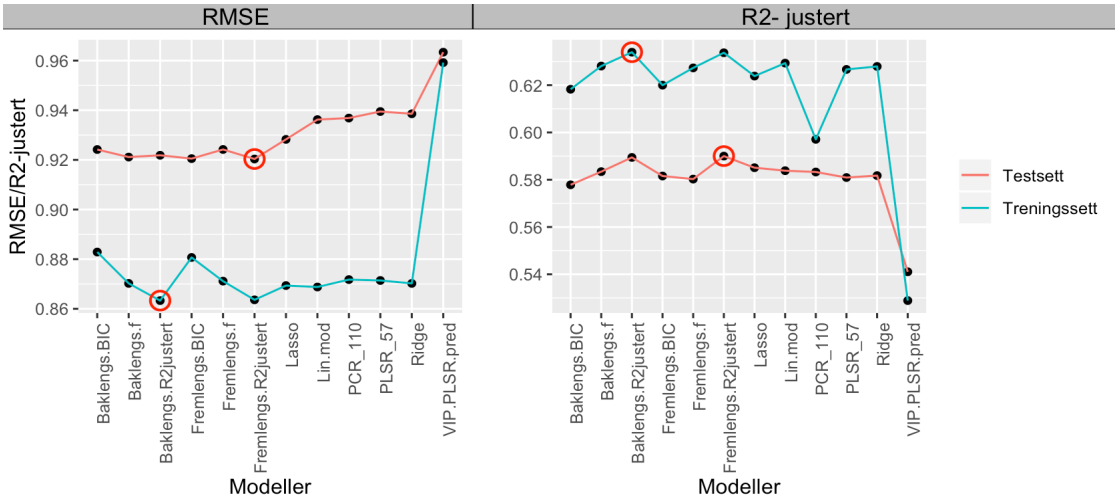


Figur 7.13: Sammenligner estimert MSEP ved kryssvalidering som funksjon av antall komponenter inkludert i modellen for PLSR (høyre) og PCR (venstre). Antall komponenter som fører til lavest MSEP-estimat er, markert med blå vertikal stiplet linje, 110 og 57 for henholdsvis PCR og PLSR. Standardavviket til MSEP-estimatet er også markert i hvert punkt. Som man kan se er det kun små endringer som skal til for at den optimale modellen som minimerer MSEP har enten færre eller flere komponenter, spesielt ved PLSR.

7.5 Modellevaluering

For prediksjon er tolv modeller tatt med videre for å teste deres evne til å predikere DB1 gitt nye data (testsettet). Modellene for hver av tilpasningsmetodene, er valgt på bakgrunn av resultatene i delkapittel 7.4. Tabell 7.1 viser en oversikt over de ulike modellene og deres tilhørende R^2, $R^2_{justert}$ og RMSE verdier knyttet til trenings- og testsett.

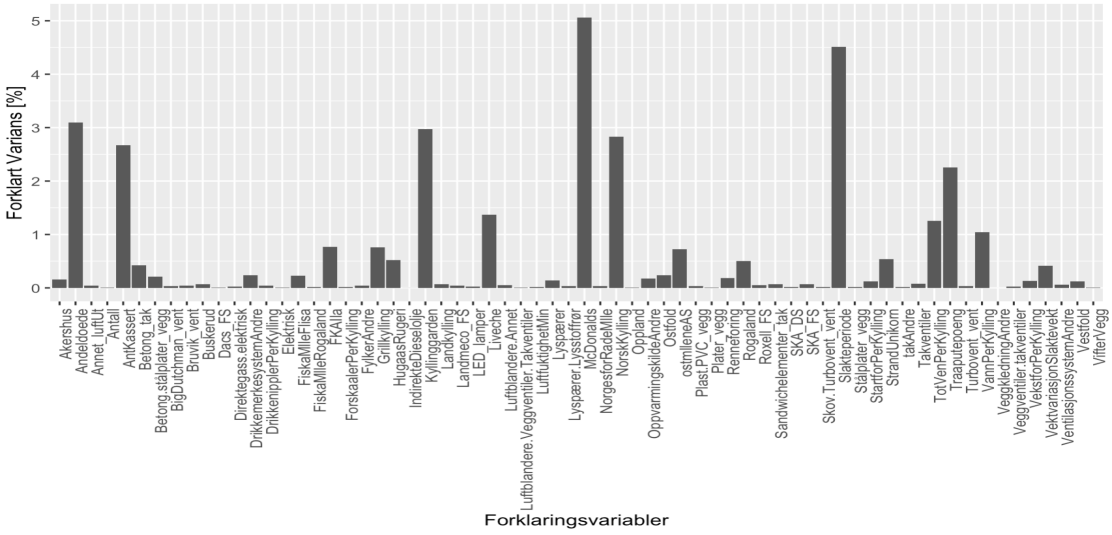
For å evaluere de tilpassede modellenes prediksjonsevne er de testet mot testsettet. De aktuelle modellenes predikerte DB1-verdier sammenlignes da mot testsettets observerte DB1-verdier. Root mean sqeared error (RMSE), og $R^2_{justert}$ er benyttet for å sammenligne modellene og som mål på modellenes prediksjonsevne. Som forventet, og som man kan se av figur 7.14, vil modellene ha høyere RMSE og lavere $R^2_{justert}$ verdier når de evalueres mot nye data sammenlignet med treningsdataene. Siden man er opptatt av modellenes prediksjonsevne, basert på resultatene (se figur 7.14), er modellen hvor det er benyttet fremlengs variabelseleksjon med $R^2_{justert}$ som evalueringskriterium ansett for å være den beste modellen. Modellen gir $RMSE = 0.920$ og $R^2_{justert} = 0.590$ og inkluderer 71 variabler. Variablene *Elverum*, *Dacs_ Drikkesystem* og *Sandwichelementer_ vegg* er fjernet fra modellen grunnet høy VIF-factor (>10) uten at det gir utslag på modellens forklaringsverdi. Den endelige modellen inneholder da 68 variabler. Med en R^2-verdi på 0.603 vil det si at modellen forklarer litt over halvparten av den totale variasjonen i DB1 (se ligning 4.9). Ved siden av kyllingenes alder, som forklarer 29 % av variansen, viser figur 7.15 en oversikt over forklaringsevnen til de resterende koeffisientene. Tabell 7.2 viser en oversikt over de 15 koeffisientene, med tilhørende estimater, som forklarer mest av variansen i DB1.



Figur 7.14: Sammenligning av modeller basert på deres RMSE og $R^2_{justert}$ verdier for henholdsvis trenings- og testsett.

Modell	Ant.koef	Treningssett			Testsett		
		R^2	$R^2_{justert}$	RMSE	R^2	$R^2_{justert}$	RMSE
Lineær full modell	111	0.647	0.629	0.869	0.604	0.584	0.936
Fremlengs f-test	37	0.633	0.627	0.871	0.591	0.583	0.924
Baklengs f-test	41	0.634	0.628	0.870	0.565	0.557	0.921
Fremlengs $R^2_{justert}$	71	0.645	0.634	0.864	0.602	0.590	0.921
Baklengs $R^2_{justert}$	75	0.646	0.634	0.863	0.603	0.589	0.922
Fremlengs BIC	27	0.624	0.618	0.883	0.586	0.582	0.921
Baklengs BIC	25	0.622	0.618	0.883	0.582	0.578	0.924
Ridge	111	0.646	0.628	0.870	0.602	0.582	0.939
Lasso	82	0.642	0.624	0.869	0.600	0.581	0.928
PLSR 57 komponenter	111	0.644	0.627	0.871	0.600	0.589	0.939
PCR 110 komponenter	111	0.616	0.597	0.871	0.603	0.583	0.937
PLSR VIP	24	0.551	0.529	0.959	0.546	0.541	0.963

Tabell 7.1: Oversikt over hvilke modeller som er tilpasset treningsdataene, samt antall koeffisienter, R^2, $R^2_{justert}$ og RMSE verdier for modellene med hensyn på treningssettet.



Figur 7.15: Viser forklaringsevnen til hver av variablene ved fremlengs seleksjon hvor $R^2_{justert}$ er benyttet som evalueringskriterium. Modellen inneholder i alt 68 variabler. Forklaringsevnen er oppgitt som prosent av total varians. Variabelen kyllingenes *Alder* forklarer 29% av variansen, men er ikke inkludert her for å lettere kunne lese grafen.

Tabell 7.2: Oversikt over de 15 variablene i den valgte modellen (Fremlengs $R^2_{justert}$) som forklarer mest av variasjonen i DB1. Estimat er regresjonskoeffisientenes estimerte verdier.

Koeffisient	Var. forklart [%]	Estimat
Alder	29.30	0.357
McDonalds	5.06	1.287
Slakteperiode	4.51	0.251
Andel Døde	3.09	-0.293
Kyllinggården	2.97	1.054
Norsk Kylling	2.83	2.753
Ant. Kassert	2.67	-0.224
Tråputepoeng	2.25	-0.049
Liveche	1.37	5.462
Vann Per Kylling	1.25	0.223
FKA Ila	0.77	0.109
Grillkylling	0.759	-0.730
Østmøllene AS	0.724	-0.196
Strand Unikorn	0.534	0.218
Hugaas Rugeri	0.524	0.818

Kapittel 8

Diskusjon og konklusjon

Dette kapittelet diskuterer resultatene fra kapittel 7 før det kulminerer i en konklusjon som forsøker å svare på oppgavens problemstilling:

Hypotesen er at noen slaktekyllingprodusenter har stabilt bedre inntjening per kylling, enn andre. Stemmer dette, og i så fall hvilke faktorer er avgjørende for at noen produsenter alltid *lykkes, mens andre* aldri *gjør det?*

8.1 Diskusjon

Preprosessering av data
I denne oppgaven er det valgt å legge akkumuleringsnivå til innsett. Dette har ført til at mye av de daglige observasjonene er gått tapt. Det kunne også vært interessant å sett på hvordan forklaringsvariablenes variasjon gjennom innsettet påvirket DB1. Dette ville muligens tegnet et mer nyansert bilde av de ulike variablene som påvirker DB1. Utfordringen ved å legge akkumuleringsnivå til hver dag er de store mengdene manglende data. Dersom man skulle imputert de manglende dataene for hver dag ville det i flere tilfeller vært svært få observasjoner å basere imputeringsverdiene på, som igjen kunne økt sjansen for uriktige slutninger. Som et alternativ kunne man valgt å slette de ufullstendige observasjonene. Dette ville ført til betydelig redusert datamateriale og gjentak av produsent, som igjen ville ført til et forventningsskjevt resultat.

Det er totalt over 300 registrerte forklaringsvariabler om man legger sammen alle de ulike datasettene. Det datasettet som er benyttet for modelltilpasning inneholder kun 39 forklaringsvariabler. Selv om store deler av de 300 opprinnelige variablene tydelig ikke har noen forklaringsverdi knyttet til DB1, kan man risikere at noen

forklaringsvariabler er fjernet fra videre analyse selv om de faktisk har effekt på DB1.

For å unngå gjentak av produsent er det satt en grense på maks 8 nivåer for hver av de kategoriske variablene, hvor de resterende nivåene er samlet i en gruppe. Et annet, og muligens bedre alternativ, ville vært og gitt en begrensning om at hvert nivå minst må inneholde observasjoner fra et gitt antall produsenter. Datasettets kompleksitet, bestående av mange kontinuerlige og kategoriske variabler, samt store mengder manglende data har gjort denne prosessen til en vanskelig avveining hvor man på den ene siden risikerer å miste viktig informasjon, mens man på den andre siden har høy risiko for betydelig gjentak innad i gruppe og multikollinearitet mellom forklaringsvariablene.

Det er valgt å imputere de manglende dataene, da et alternativ ved å fjerne de manglende dataene ville ført til at man enten kun hadde fire forklaringsvariabler eller en betydelig reduksjon i antall observasjoner. Dette ville muligens gitt høyere troverdighet knyttet til slutningene rundt disse variablene, men det ville redusert modellens totale forklaringsevne. Kostnaden i form av tapt informasjon ville vært enorm. Andelen manglende data hvor det ikke vil være lønnsomt å imputere dataene kan være avhengig av datasettet og valg av imputeringsmetode. Det er satt en grense på 30%, men dersom man hadde undersøkt dataenes underliggende strukturer nærmere kan det tenkes at det hadde påvirket valg av grense og imputeringsmetode. Ved imputering av manglende data risikerer man å underestimere tilpassede modellers prediksjonsfeil da man ikke har nye observasjoner, men imputerte verdier som belager seg på eksisterende observasjoner. Man risikerer dermed å forsterke de relasjonene som finnes i de observerte dataene, noe som igjen kan føre til at man påstår effekt av variabel selv om det i realiteten ikke finnes. For å optimalisere imputeringsmodellen ble kompleksitetsparameterne, antall trær i hver skog og antall forklaringsvariabler ved hver node vurdert hver for seg, i stedet for alle mulige kombinasjoner av de to. Muligens kunne man ved å vurdere alle mulige kombinasjoner tilpasset en imputeringsmodell som ville gitt lavere OOB-verdi. For å unngå overtilpasning ble uansett ikke den imputeringsmodellen som gav lavest OOB-verdi valgt for imputering av det endelige datasettet. Dette i et forsøkt på å ivareta både tilfeldigheten og nøyaktigheten ved de imputerte verdiene, uten at det finnes noen garantier for at valgene som er gjort oppnår dette.

Modelltilpasning

Det er kun tilpasset lineære modeller i denne oppgaven. Selv om eksempelvis *Alder* tilsynelatende har en positiv lineær effekt på DB1 er det lite trolig at denne utviklingen vil holde seg slik langt ut over det analyserte utfallsrommet. Resultatet gir altså ikke grunnlag for å øke produksjonslengdene til de ulike produksjonstypene.

For de tilpassede modellene er det ikke tatt hensyn til multikollinearitet før man valgte den beste modellen. Hadde man tatt hensyn til multikollinearitet for alle modellene kunne dette ført til et annet valg av beste modell.

Det er ikke evaluert alle mulige variabelkombinasjoner fra null-modellen til den fullstendige modellen. Det er mulig andre variabelkombinasjoner ville gitt bedre resultat enn de valgte modellene ved fremlengs seleksjon og baklengs eliminasjon.

Ved tilpasning av PLSR VIP-modellen ble det tatt utgangspunkt i PLSR-modellen med 57 komponenter og sett på variablenes VIP-verdier for variabelutvelgelse. Som et alternativ kunne man benyttet kryssvalidering med ulikt antall komponenter (1 til 111), sett på tilhørende VIP-verdier, og deretter vurdert de tilpassede modellenes prediksjonsevne. På denne måten ville man optimalisert modellen i forhold til VIP-modellen, ikke PLSR-modellen.

Modellantagelser
Det er sjekket modellantagelsene for alle modellene. I tillegg A.1 kan man se resultatene fra evalueringen av modellantagelsene fra en av modellene, fremlengs variabelseleksjon med F-test ($\alpha = 0.05$) som utvelgelses metode. Det er små forskjeller mellom de ulike modellene, så resultatene her er representative for alle modellene. Vanligvis starter man med tilpasningen av vanlig lineære modeller før man eventuelt gjøre disse mer kompleks, eller vurdere andre ikke-lineære tilpasningsmetoder. Av figur A.1 kan det tilsynelatende se ut til at residualene $(y_i - \hat{y}_i)$ er tilfeldig spredt rundt null for hele utfallsrommet, som tyder på at en lineær modell er passende for datasettet.

Som man kan se av figur A.2 finnes det noen ekstremverdier som modellen ikke klarer å forklare. Det kan tenkes at disse verdiene skyldes ekstraordinære produksjonsproblemer, eller at produsentene har prestert usedvanlig godt/dårlig. En annen mulig forklaring kan være feilføringer. Som et alternativ for å oppnå normalitet kunne man fjernet ekstremverdiene, men dette ville tilskrevet modellen kunstig høy forklaringsevne. I følge Gelman og Hill er normalitet i residualene den minst viktige antagelsen, spesielt dersom man har mange observasjoner (Gelman & Hill, 2006).

Fra figur A.3 kan det se ut til at antagelsen om konstant varians holder for DB1-verdier i området $2-8$ kroner. For DB1-verdier over dette kan det tilsynelatende se ut til at variansen øker og modellantagelsen vil være brutt. Selv om det kan være mange årsaker til varierende varians, endres variansen ofte proporsjonalt med en faktor (Frost, 2019). I dette tilfellet kan det se ut til produksjonstypene med lavt forventet DB1 (foredlingskylling og grillkylling) har konstant varians, mens produksjonstyper hvor man forventer høyere DB1 har større varians. Ved å bryte antagelsen om konstant varians reduseres estimeringspresisjonen av modellkoeffisientene. Lavere

presisjon øker sjansen for at koeffisientens estimat er lengre unna koeffisientens virkelige verdi.

Siden hver produsent har mellom 5-14 registrerte innsett, vil ikke alle observasjonene være fullstendig uavhengige da man vil oppleve gjentak av produsent. Dette kan føre til kunstig lave p-verdier og at man påstår effekt av forklaringsvariabel selv om det i realiteten ikke er noen.

Modellevaluering
Fra resultatene i delkapittel 7.5 ble modellen hvor det ble benyttet fremlengs seleksjon med $R^2_{justert}$ som evalueringskriterium ansett for å være den beste modellen. Modellen gir en R^2-verdi = 0.602, som vil si at de 68 modellkoeffisientene kun forklarer litt over halvparten av variasjonen i dataene. Dette tyder på at det finnes forklaringsvariabler som ikke er registrert som har stor påvirkning på DB1. I Norturas tidligere undersøkelser konkluderes det med at spesielt røktet, altså hvordan produsentene drifter produksjonen, har stor effekt på DB1. I så fall er det ikke unaturlig at man kun har en R^2-verdi på rundt 0.6. Av tabell 7.2 kan man se at flere av produksjonstypene som, *McDonalds, Kyllinggården, Liveche* og *Grillkylling* har høy forklaringsverdi, mens eksempelvis *Økologisk kylling* ikke engang er inkludert i modellen selv om denne produksjonstypen tydelig har høyere DB1 enn andre (se figur 7.5a). Årsaken til at *Økologisk kylling* ikke er inkludert i modellen kan skyldes variabelens effekt på DB1 tett kan relateres til effekten *Alder* har på DB1. Ellers er det gledelig å se at forventet DB1 øker med dyrevelferd (lavere *tråputepoeng*), men siden det ikke er tatt hensyn til eventuelle økte kostnader som følge av dyrevelferdstiltak, er det vanskelig å si om økt dyrevelferd faktisk øker lønnsomheten. Det er interessant å se at slakteriet *Norsk Kylling* og rugeriet *Hugaas Rugeri* har så høye verdier. Datasettet har kun 26 og 73 observasjoner fra henholdsvis Norsk Kylling slakteri og Hugaas rugeri, noe som kan ha ført til betydelig gjentak av produsent og produksjonstype. Deler av effekten kan også skyldes at kasseringskontrollørene ved Norsk Kylling ikke er like strenge som ved andre slakterier, og at kvaliteten på de daggamle kyllingene fra Hugaas er høyere. Foruten om *Vann per kylling*, som muligens kan relateres til drikkesystem, er ingen andre koeffisienter som beskriver produksjonslokalet blant de 15 koeffisientene som forklarer mest av variasjonen i DB1. Uansett ser *Vann per kylling* ut til å ha positiv effekt på dekningsbidraget. Det er naturligvis viktig at kyllingene tar til seg tilstrekkelig med vann, likevel kan det tenkes at effekten av *Vann per kylling* ikke er lineær utenfor det observerte utfallsrommet. Det er med andre ord lite trolig at DB1 vil fortsette å øke lineært dersom man får kyllingene til å drikke ekstreme mengder vann. Det kan også tenkes at denne positive effekten av *Vann per kylling* kan trekkes tilbake til effekten *Alder* og/eller *Produksjonstype*, da eldre kyllinger naturlig vil ha et høyere totalt vannforbruk. LED-lamper ser ut til å ha positiv

effekt på DB1 sammenlignet med lyspærer og lysstoffrør. Dessverre finnes kun 4 observasjoner, alle fra samme produsent, med LED-lamper. Denne variabelen bør derfor undersøkes nærmere før man drar noen endelige konklusjoner. I samtaler med Nortura ble det tidlig i prosessen skissert at *Lystimer* hadde effekt på DB1, likevel er det ingen av modellene hvor det er fortatt variabelseleksjon som har inkludert denne variabelen. Det kan dog tenkes at kyllingene trenger ulik mengde lys gjennom innsettet og at et akkumuleringsnivå per dag ville gitt et annet resultat.

Av figur 7.15 kan man se at flertallet av variablene som er inkludert i modellen forklarer under 0.5 % av variansen i DB1. Dersom det ikke er hensiktsmessig å forholde seg til 68 variabler, kan man vurdere å velge en modell med færre variabler, som for eksempel Baklengs eliminasjon med BIC som evalueringskriterium. Denne modellen inneholder 25 variabler, uten at det reduserer forklaringsevnen i særlig grad ($R^2 = 0.582$). På den andre siden er kyllingproduksjon en svært marginal bransje hvor små forskjeller kan utgjøre forskjellen mellom det å lykkes eller ikke.

8.2 Konklusjon

Det er forskjell mellom kyllingprodusenter. Noen produsenter oppnår jevnt og stabilt bedre dekningsbidrag enn andre produsenter. Store deler av denne forskjellen ligger i kyllingenes alder og produksjonstype. Disse variablene er også sterkt korrelert da flere produksjonstyper har ulike produksjonslengde, men produksjonstypene McDonalds, Landkylling, Kyllinggården og Liveche virker til å være ekstra lønnsomme selv etter å ha tatt hensyn til alder. På den andre siden virker grillkylling og foredlingskylling å være de produksjonstypene som fører til lavest forventet DB1. Kyllingprodusentene som benytter fôrleverandørene FKA Ila, Strand Unikorn, Fiskå Mølle Rogaland, og Fiskå Mølle Flisa har bedre DB1 sammenlignet med produsentene som benytter Østmøllene AS, Norges råde mølle og FKA Kambo. Lysstoffrør som lyskilde ser ut til å ha positiv effekt på DB1 sammenlignet med lyspærer. Effekten av LED-lamper ser ut til å være best, men bør undersøkes nærmere da det finnes for få observasjoner til å kunne trekke noen endelig konklusjon. Fôringssystemene fra Roxell og Dacs gir, med bakgrunn i det tilgjengelige datamaterialet, dårligere forventet DB1 sammenlignet med Big Dutchman. Som oppvarmingskilde burde man benytte direkte gass kombinert med elektrisk oppvarming da dette vil øke forventet DB1 sammenlignet med andre typer oppvarming. Det er dog ikke tatt hensyn til variasjon i kostnader knyttet til energiforbruket for de ulike teknologiene. DB1 for kyllingprodusenter med ventilasjonsanlegg fra Skov eller Turbovent er høyere enn de som benyttet Bruvik eller Big Dutchman.

Som man kan se av størrelsen på de estimerte modellkoeffisientene er kyllingproduksjon en svært marginal bransje. Siden innsettsperioden er kort finnes det små

rom for å rette opp eventuelle feil/avvik. Røktet anses derfor som å være den viktigste faktoren for et godt resultat. Konkret bør man gjøre nødvendige tiltak for å redusere fare for smitte, ta ut syke/svake dyr så tidlig som mulig, fokusere på strøbedet og dyrevelferden til dyrene, samtidig som man sørger for at kyllingene har optimale forutsetninger for å ta til seg fôr og ikke minst vann. Spesielt den første uken, før kyllingene begynner å finne seg til rette, bør det legges ned ekstra fokus og innsats.

8.3 Videre arbeid

For videre arbeid er det valgt å dele mellom (1) anbefalinger med tanke på datainnsamling og øvrige (2) anbefalinger.

For (1) anbefales det å:

(a) Gjennomgå registreringsrutiner for data i fjørfekontrollen med kyllingprodusentene for å standardisere hvordan dataene blir ført.

(b) Gjennomføre forsøk under kontrollerte forhold hvor man vurderer virkningen av de ulike variablene ved å endre en eller flere variabler om gangen.

(c) Gi insentiver til produsentene for å registrere data. Spesielt da de faktorene som man med erfaring tenker kan ha positiv/negativ effekt på resultatet.

(d) Forklare ovenfor kyllingprodusentene nytteverdien av gode data.

For (2) anbefales det å:

(a) Undersøke muligheter for å øke standardiseringen av røktet slik at man utjevner eventuelle forskjeller denne variabelen medfører.

(b) Utvide modellen og vurdere ikke-lineære sammenhenger og interaksjon mellom variabler.

(c) Undersøke hvordan daglige variasjoner eventuelt påvirker DB1.

Bibliografi

Allison, P.D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods & Research*, *28*(3), 301-309. Hentet fra `https://doi.org/10.1177/0049124100028003003` doi: 10.1177/0049124100028003003

Almaas, H.E., Bjørkhaug, H. & Almås, R. (2018). Norsk kyllingproduksjon 2013-2017. *Agrispace*(1/18), 1-4.

Alvseike, O.A., Kjos, A.K., Nafstad, O., Reksnes, H.O., Ruud, T.A. & Saltnes, T. (2015, November). Status i norsk kjøtt- og eggproduksjon. *Kjøttets tilstand*, 11-12.

Animalia. (2017, Januar). Dyrevelferdsprogram slaktekylling [programvarehåndbok]. Hentet 12.03.20 fra `https://www.animalia.no/no/Dyr/fjorfe/slaktekylling---helse-og-velferd/helse-og-velferd-hos-slaktekylling/`

Animalia. (2019, juni). *Slaktekylling – informasjon om hybrider.* Hentet 16.03.20 fra `https://www.animalia.no/no/Dyr/fjorfe/slaktekylling---helse-og-velferd/slaktekylling--informasjon-om-hybrider/`

Bondelag, N. (2019, April). *Konsesjon og kvote.* Hentet 12.03.20 fra `https://www.bondelaget.no/konsesjonogkvote/`

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Bruce, P. & Bruce, A. (2017). *Practical statistics for data scientists: 50 essential concepts*. O'Reilly Media, Inc.

Budsjettnemda, f.j. (2019). *Totalkalkylen - statistikk.* (Tilgjengelig fra: https://www.nibio.no/tjenester/totalkalkylen-statistikk?locationfilter=true)

Chong, I.-G. & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, *78*(1-2), 103–112.

Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Daniel, C. & Wood, F.S. (1980). *Fitting equations to data: computer analysis of multifactor data*. John Wiley & Sons, Inc.

Dean, A., Voss, D. & Draguljic. (2017). *Design and analysis of experiments* (2. utg.; R.DeVeaux, S. Fienberg & I.Olkin, red.). Springer International Publishing.

Digitaliseringsdirektoratet. (2020, 02). *Leveranse til slakteri (2005 til 2018) [.csv fil].* (Tilgjengelig fra https://data.norge.no/los/slakteri)

Fox, J. & Weisberg, S. (2019). *An R companion to applied regression* (Third utg.). Thousand Oaks CA: Sage. Hentet fra `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`

Friedman, J., Hastie, T. & Tibshirani, R. (2001). *The elements of statistical learning* (vol. 1) (nr. 10). Springer series in statistics New York.

Frost, J. (2019). Regression analysis. an intuitive guide for using and interpreting linear models. *ebook*.

Gelman, A. & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Graham, J.W. (2012). Missing data theory. I *Missing data: Analysis and design* (s. 3-46). New York, NY: Springer New York.

Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Holien, S.O. (2009). Økonomi og arbeidsforbruk i produksjon av slaktekylling. *Notat 2009-18*, 1-13.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning* (vol. 112). Springer.

Jerome Friedman, R.T., Trevor Hastie. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1-22. Hentet fra `http://www.jstatsoft.org/v33/i01/`

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiolgy*, *64*(5), 402-406. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/)

Kjos, A.K., Nafstad, O., Reksnes, H.O., Ruud, T.A., Saltnes, T. & Ytterdahl, M. (2019). Status i norsk kjøtt- og eggproduksjon. *Kjøttets tilstand*, 60-90.

Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J. & Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing lc-ms metabolomics data: a comparative study. *BMC Bioinformatics*, *20*(1), 492. Hentet fra `https://doi.org/10.1186/s12859-019-3110-0` doi: 10.1186/s12859-019-3110-0

Lee, J.H. & Jr., J.H. (2011, september). *Multiple imputation with large proportions of missing data: How much is too much?* (United Kingdom Stata Users' Group Meetings 2011 nr. 23). Stata Users Group. Hentet fra `https://ideas.repec.org/p/boc/usug11/23.html`

Liland, K.H. (2018). mixlm: Mixed model anova and statistics for education [programvarehåndbok]. Hentet fra https://CRAN.R-project.org/package=mixlm

Lohr, S. (2014, August). *For big -data scientists, janitor work is key hurdle to insights.* (https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html)

Lovdata. (2015, Juni). *Forskrift om produksjonstilskudd mv. i jordbruket.* https://lovdata.no/dokument/SF/forskrift/2014-12-19-1817.

Lumley, T. & Miller, A. (2017). leaps: Regression subset selection [programvarehåndbok]. Hentet fra https://CRAN.R-project.org/package=leaps

Mevik, B.-H., Wehrens, R. & Liland, K.H. (2019). *pls: Partial least squares and principal component regression.* Hentet fra https://CRAN.R-project.org/package=pls

Nortura. (2019a). *Fjørfehold i norge.* Hentet 16.03.20 fra http://www.nortura.no/naturlig-kvalitet-fra-norske-bonder/kyllinghold/

Nortura. (2019b). Grunnlagsdokument storfe, sau/lam og egg - 2.halvår 2019. *Totalmarked*, *2*, 28-29.

Probst, P. & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research*, *18*(1), 6673–6690.

Rubin, D.B. (2004). *Multiple imputation for nonresponse in surveys* (vol. 81). John Wiley & Sons.

Rye, S.K.P., Jenssen, E. & Wenstøp, Y.Q. (2019, Oktober). Økonomien i produksjon av slaktekylling. *Norsk institutt for bioøkonomi (NIBIO)*, *5*(88), 1-30.

Schwarz, G. (1978, 03). Estimating the dimension of a model. *Ann. Statist.*, *6*(2), 461–464. Hentet fra https://doi.org/10.1214/aos/1176344136 doi: 10.1214/aos/1176344136

Shrader, H. (1952). The chicken-of-tomorrow program; its influence on "meat-type" poultry production. *Poultry Science*, *31*(1), 3 - 10. Hentet fra http://www.sciencedirect.com/science/article/pii/S0032579119513013 doi: https://doi.org/10.3382/ps.0310003

SSB. (2019). *Fakta om jordbruk.* Hentet 18.03.20 fra https://www.ssb.no/jord-skog-jakt-og-fiskeri/faktaside/jordbruk

Starkweather, J. (214, July). A new recommended way of dealing with multiple missing values: Using missforest for all your imputation needs. *Benchmarks RSS Matters*, 1-9.

Stekhoven, D.J. & Bühlmann, P. (2011, 10). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118. Hentet fra https://doi.org/10.1093/bioinformatics/btr597 doi: 10.1093/bioinformatics/btr597

Stekhoven, D.J. & Stekhoven, M.D.J. (2012). Package 'missforest'.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R. (2013a, Mars). *Modern regression 1: Ridge regression.* (Powerpoint page 18-20)

Tibshirani, R. (2013b, Mars). *Modern regression 2: The lasso.* (Powerpoint page 2)

Tibshirani, R. (2013c, April). *Tree-based methods for classification and regression.* (Powerpoint)

Tukey, J.W. (1977). *Exploratory data analysis* (vol. 2). Reading, Mass.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (nr. 978-3-319-24277-4). Springer-Verlag New York. Hentet fra `https://ggplot2.tidyverse.org`

Wold, H. (1982). Soft modeling: the basic design and some extensions. *Systems under indirect observation*, *2*, 343.

Wold, S., Johansson, E. & Cocchi, M. (1993). 3d qsar in drug design: theory, methods and applications. *ESCOM, Leiden, Holland*, 523–550.

Wold, S., Sjöström, M. & Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, *58*(2), 109–130.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway

Tillegg

Tillegg A

Tilpassede modeller

En oversikt over alle de tilpassede modellene finnes ved å følge linken:

https://www.dropbox.com/sh/7u97fgvywlbu1s5/AABdmVYgYDnA1l1JFoWm-H4xa?dl=0.

Tabell A.1-A.2 viser oversikten over de estimerte koeffisientene til den beste modellen, fremlengs variabelseleksjon med $R^2_{justert}$ som evalueringskritere. Tabell A.3 viser en oversikt over referansenivåene i modellene for de ulike kategoriske variablene. Tabell A.4 viser hvilke variabler som "alltid" ble inkludert i modellene (>9/12).

Koeffisient	Estimat	Std.error	P-verdi
Intercept	5.473	0.084	$< 1e^{-10}$
Alder	0.357	0.068	$1.87e^{-7}$
Vann per kylling	0.223	0.0267	$< 1e^{-10}$
Luftfuktighet min	-0.036	0.019	0.069
Antall	0.121	0.029	$3.87e^{-5}$
Startfôr Per kylling	-0.106	0.025	$2.77e^{-5}$
Vekstfôr per kylling	-0.535	0.039	$< 1e^{-10}$
Tråputepoeng	-0.049	0.021	0.019
Forskåler per kylling	0.100	0.024	$3.72e^{-5}$
Drikkenippler per kylling	-0.044	0.027	0.105
Tot. ventilasjon per kylling	0.043	0.025	0.083
Slakteperiode	0.251	0.019	$< 1e^{-10}$

Tabell A.1: Modellkoeffisientene til modellen Fremlengs variabelseleksjon ved $R^2_{justert}$ som evalueringskriterium. Denne modellen er er valgt som den beste modellen på bakgrunn av dens prekdiksjonsevne sammenlignet med de andre modellene. Prediksjonsevenen er esimert mot et ukjent testsett. Tabellen fortsetter på neste side.

Koeffisient	Estimat	Std.error	P-verdi
Antall kassert	-0.225	0.023	$< 1e^{-10}$
Andel døde	-0.293	0.020	$< 1e^{-10}$
Østmøllene AS	-0.196	0.078	0.012
Strand Unikorn	0.214	0.081	0.008
Fiskå mølle Flisa	0.281	0.087	0.001
FKA Ila	0.109	0.108	0.314
Norges fôr råde mølle	-0.485	0.092	$1.48e^{-7}$
Fiskå mølle Rogaland	0.111	0.118	0.348
Grillkylling	-0.730	0.101	$6.26e^{-13}$
Kyllinggården	1.055	0.067	$< 1e^{-10}$
McDonalds	1.297	0.074	$< 1e^{-10}$
Landkylling	4.355	0.469	$< 1e^{-10}$
Liveche	5.462	0.440	$< 1e^{-10}$
Norsk Kylling	2.753	0.335	$< 1e^{-10}$
Hugås Rugeri	0.821	0.169	$1.26e^{-6}$
Elektrisk oppvarming	-0.169	0.102	0.098
Indirekte Dieselolje	-0.245	0.128	0.066
Direkte gass / elektrisk	0.146	0.130	0.264
Andre oppvarmingskilder	-0.262	0.070	0.001
Landmeco Fôrsystem	-0.080	0.061	0.188
Roxell fôrsystem	-0.238	0.076	0.002
Dacs Fôrsystem	-0.258	0.107	0.016
Rennefôring	0.549	0.153	0.001
SKA drikkesyst.	0.430	0.327	0.189
Drikkesyst. andre	-0.126	0.068	0.064
Bruvik vent.syst.	-0.131	0.058	0.025
Big Dutchman vent.syst.	-0.151	0.070	0.031
Turbovent vent.syst	0.195	0.126	0.122
Skov turbovent vent.syst	0.251	0.140	0.074
Vent.syst. andre	-0.210	0.122	0.086
Takventiler	0.339	0.141	0.016
Tak- + veggventiler	0.199	0.184	0.279
Vifter vegg	-0.207	0.092	0.025
Annet luft ut	-0.718	0.369	0.062
Veggplater	-0.096	0.064	0.138
Stålplater vegg	-0.107	0.065	0.097
Plast/PVC vegg	-0.128	0.087	0.139
Betong/stålplater vegg	-0.442	0.151	0.004
Veggkledning andre	0.185	0.134	0.168
Sandwichelementer tak	-0.187	0.109	0.085
Betongtak	0.759	0.137	$3.52e^{-8}$
Lyspærer	-0.152	0.056	0.007
Lyspærer/lysstoffrør	0.134	0.092	0.145
LED lamper	0.541	0.442	0.221

Tabell A.2: Fortsettelse av modellkoeffisienten til den valgte modellen.

Kategori	Referanse nivå
Produksjonstype	Foreldingskylling
Fôrprodusent	FKA Kambo
Merke fôringssytem	Big dutchman
Rugeri	Nortura samvirkekylling
Oppvarmingskilde	Direkte gass
Fôrsystem	Skålfôring
Drikkesystem Merke	Big Dutchman
Vent.syst. merke	Skov
Vent.syst ut	Vifter over tak og vegg
Vent.syst. inn	Veggventiler
Veggkledning	Betong
Takkledning	Stålplater
Lyskilde	Lysrør

Tabell A.3: Oversikt over referansenivåene til de kategoriske variablene som inngår i modellen.

Variabel	Estimat (min-maks)
Alder	0.105-0.383
Vann per kylling	0.223-0.242
Antall	0.108-0.129
Startfôr per kylling	-0.157-(-0.094)
Tråputepoeng	-0.066-(-0.046)
Forskåler per kylling	0.076-0.096
Slakteperiode	0.250-0.257
Antall kassert	-0.269-(-0.235)
Andel døde	-0.302-(-0.279)
Kyllinggården	1.017-1.107
McDonalds kylling	1.274-1.336
Landkylling	4.07-5.077
Liveche	5.343-6.044
Norsk kylling	2.451-3.114
Haugås rugeri	0.791-0.919
Bruvik ventilasjon	-0.180-(-0.107)
Takventiler	0.283-0.411
Betongtak	0.392-0.791
Lyspærer	-0.173-(-0.153)
Betong/stålplater vegg	-0.456 -(-0.315)

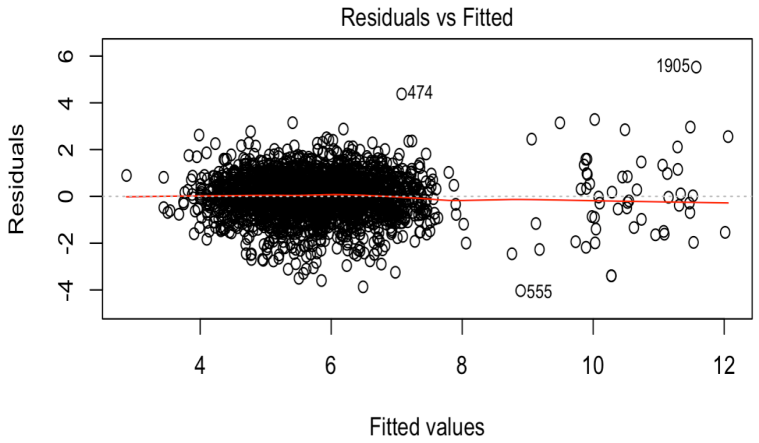
Tabell A.4: Oversikt over variablene som "alltid"(>9/12) blir inkludert i modellene. Det er også inkludert det høyeste og laveste estimatet av den tilhørende modellparameteren.

A.1 Modellantagelser

Det er sjekket modellantagelsene for alle modellene. Her er kun resultatene fra fremlengs variabelseleksjon med f-testing inkludert. Det er små forskjeller mellom de ulike modellene, så resultatene her utgjør et representativt utvalg.

Linearitet

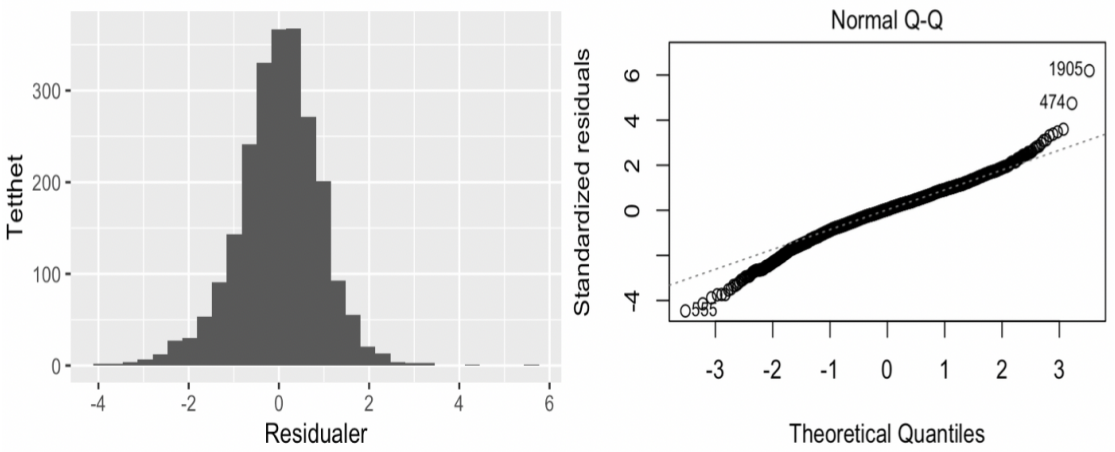
Man ønsker ikke å se et mønster om man plotter modellens tilpassede verdier (Fitted values) mot residualer (Residuals). Det vil si, den røde linjen bør være omtrent horisontal rundt null. Dersom man kan se et mønster kan det tyde på ikke-lineære sammenhenger i dataene. Et alternativ kan da være å transformere forklaringsvariablene (se delkapittel 4.1). Figur A.1 viser residualplottet for fremlengs variabelseleksjon hvor f-test er benyttet som evalueringskriterium. I følge residualplottet kan det se ut som det er lineær sammenheng mellom forklaringsvariabler og responsvariabel.



Figur A.1: Residualplot. De modelltilpassede verdiene (Fitted values) er plottet mot residualene for å se etter-reventuelle mønster. Dersom man oppdager mønster i residualene kan det type på ikke-lineære sammenhenger mellom forklaringsvariabler og respons.

Normalitet

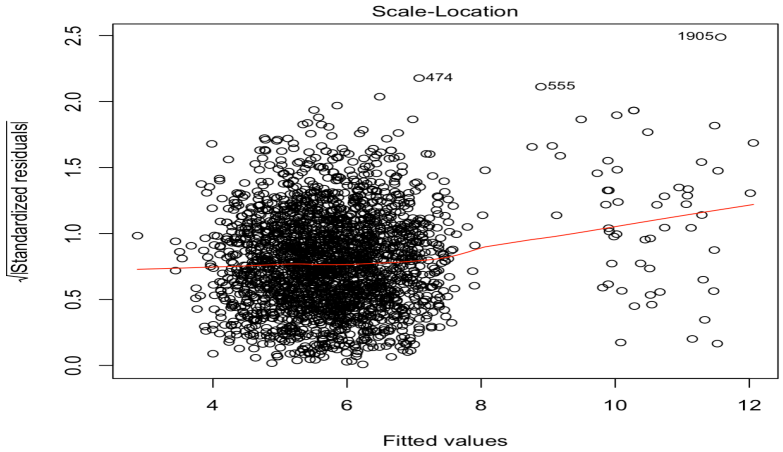
Av histogrammet i figur A.2 kan det se ut til at residualene er tilnærmet normalfordelt rundt 0, men man kan se antydninger til ekstremverdier som fører til "tunge haler". Dette blir enda tydeligere av Normal Q-Q plottet (se figur A.2). Cooks distanse er et mål på observasjoners innflytelse på modellen (Cook, 1977). En tommelfingerregel er om Cooks distanse overstiger $4/(n-p-1)$ så vil observasjonen ha høy innflytelse på modelltilpasningen (Bruce & Bruce, 2017). Dersom man vurderer Cooks distanse til de aktuelle observasjonene ser man at observasjonene også har høy innflytelse på modelltilpasningen.



Figur A.2: Histogram og Normal Q-Q plot benyttes for å sjekke om residualene er normalfordelt med forventningsverdi 0. Ideelt ligger alle observasjonene på den stiplede linjen i Normal Q-Q plottet.

Konstant varians

Man kan sjekke om residualene har konstant varians ved å undersøke et scale-location plot. Variansen er konstant om man har en horisontal linje med likt spredde punkter. Av figur A.3 ser man at variansen øker ved høyere tilpassede verdier. Man har altså ikke konstant varians.



Figur A.3: Scale-location plot benyttes for å sjekke om variansen er konstant over hele utfallsrommet til de tilpassede verdiene.

Tillegg B

Kyllingproduksjon

B.1 Produksjonstyper

Oversikt over produksjonstypene som er under Nortura.

Produksjonstype
Foredlingskylling
Grillkylling
Landkylling
Maiskylling
Økologisk kylling
Livêche
Forsøk SPR
McDonald‘s kylling
Hubbard
Kyllinggården

Tabell B.1: Oversikt over de 10 ulike produksjonstypene man har i Nortura

B.2 Variable kostnadssatser

Oversikt over de standardiserte variable kostnadene knyttet til kyllingproduksjon hos Nortura.

Satser hentet fra regnskapsfører	Kroner per år	Kroner per kylling
Oppvarming	217257	1.30
Strøm, andel fjørfehus	51639	0.31
Satser beregnet fra investering på teknisk utstyr		
Rep. vedlikehold tekn utstyr	37694	0.23
Satser hentet fra forsikringsselskap		
Forsikring dyr	38112	0.23
Satser ut i fra felles standard		
Traktor, andel fjørfehus	52323	0.31
Desinfiseringsmidler/vaskemidler	22747	0.14
Skadedyravtale	8094	0.05
Regnskap, andel fjørfe	9506	0.06
Vann	22245	0.13
Flis/strø	25566	0.15
Div. rekvisita/småverktøy	6331	0.04
Bil, andel fjørfe	7953	0.05
Abonnement telefon-utringninger	2283	0.01
Telefon	3425	0.02
Veterinæravtale	7434	0.04
Veterinær	1595	0.01
Arbeidsklær	2968	0.02
Brannvarsling	6910	0.04
Kontigenter	3442	0.02
Septik	5083	0.03
Sum variable kostnader	532605	3.19

Tabell B.2: Variable kostnader til bruk i prisberegninger 2.halvår 2018 og 1.halvår 2019. Reperasjon og vedlikehold av bygningsmasse fjørfehus ligge ikke i de variable kostnadene.

B.3 Avregningsliste

Norturas avregningsliste for ulike produksjonstyper finnes ved å føle linken:

https://www.dropbox.com/sh/7u97fgvywlbu1s5/AABdmVYgYDnA1l1JFoWm-H4xa?dl=0.

Tillegg C

Datasett

Oversikt over variabler, andel manglende data, og vurderinger som er gjort knyttet til inkluderingen av variabler i analysen finnes ved å følge linken:

https://www.dropbox.com/sh/7u97fgvywlbu1s5/AABdmVYgYDnA1l1JFoWm-H4xa?dl=0.

Tillegg D

R-kode

R-koden som er benyttet til analysedelen av oppgaven finnes ved å følge linken:
https://www.dropbox.com/sh/7u97fgvywlbu1s5/AABdmVYgYDnA1l1JFoWm-H4xa?dl=0.