Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Notat

Estimering av mangel på arbeidskraft: Modell og brukermanual

Notatnr SAMBA/48/16

Forfatter Clara-Cecilie Günther
Anders Løland

Dato 13. mars 2017

Norsk Regnesentral

Norsk Regnesentral (NR) er en privat, uavhengig stiftelse som utfører oppdragsforskning for bedrifter og det offentlige i det norske og internasjonale markedet. NR ble etablert i 1952 og har kontorer i Kristen Nygaards hus ved Universitetet i Oslo. NR er et av Europas største miljøer innen anvendt statistisk-matematisk modellering og har et senter for forskningsdrevet innovasjon, Big Insight, med finansiering fra Norges forskningsråd, bedrifter og offentlige partnere. Innen statistikk jobbes det med et bredt spekter av problemstillinger, for eksempel finansiell risiko, jordobservasjon, estimering av fiskebestander, helse og beskrivelse av geologien i petroleumsreservoarer. NR er ledende i Norge innen utvalgte deler av informasjons- og kommunikasjonsteknologi. Innen IKT-området har NR innsatsområdene e-inkludering, informasjonssikkerhet og smarte informasjonssystemer.

NRs visjon er forskningsresultater som brukes og synes.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Tittel	**Estimering av mangel på arbeidskraft: Modell og brukermanual**
Forfatter	**Clara-Cecilie Günther** <Clara-Cecilie.Gunther@nr.no> **Anders Løland** <Anders.Loland@nr.no>
Dato	13. mars 2017
Publikasjonsnummer	SAMBA/48/16

Sammendrag

NAV utfører jevnlig en bedriftsundersøkelse som skal gi innsikt i etterspørselssiden på arbeidsmarkedet. NR har implementert en modell i R som estimerer mangelen for alle bedriftene i landet innen ulike yrker, næringer, fylker og yrker. Rapporten beskriver modellen og gir veiledning i bruk av versjon 2.2 av programmet (Günther og Løland, 2017).

Forsidebilde: ©www.photos.com 2012

Emneord	Poissonregresjon, generalisert lineær modell, R, prognoser
Målgruppe	NAV
Tilgjengelighet	Konfidensiell
Prosjekt	Bedriftsundersøkelsen: Utvikling av ny modell og programvare
Prosjektnummer	220534
Satsningsområde	Teknologi, industri og forvaltning
Antall sider	32

Innhold

1 Innledning

NAV gjennomfører to ganger årlig en bedriftsundersøkelse der et utvalg bedrifter angir om de mangler arbeidskraft, og i tilfelle innen hvilke yrker. Hver bedrift kan angi mangel i inntil ti yrker. Basert på tallene fra bedriftene i utvalget, estimeres mangelen på arbeidskraft innen forskjellige næringer, yrker og fylker for alle bedriftene i landet.

NAV har tidligere benyttet to poissonmodeller, en for fylke og næring og en for fylke og yrke, som beskrevet av Schweder (2002). Fylke-næring-modellen ble estimert for seks ulike næringsgrupper mens fylke-yrke-modellen ble estimert for hvert enkelt yrke. Totalt sett ga ikke disse to modellene konsistente estimater. Etter en evaluering gjennomført av NR høsten 2011 (Günther og Løland, 2011), ble det besluttet å utvikle en ny modell og implementere denne i R. NR har implementert en poissonmodell som er felles for yrke og næring, og som estimeres en gang for hver yrkesgruppe, der flere yrker inngår. I kapittel 2 beskrives den statistiske modellen, kapittel 3 beskriver R-programmet og hvordan det skal brukes, mens kapittel 4 viser eksempler på resultater fra programmet for en gitt bedriftsundersøkelse.

2 Modell

Basert på svarene fra bedriftene i utvalget ønsker man å predikere mangel på arbeidskraft for alle bedriftene i populasjonen per næring, fylke og yrke. For hver bedrift har vi informasjon om antall ansatte i bedriften, geografisk beliggenhet (fylke) og næringsvis tilhørighet. Denne informasjonen utgjør forklaringsvariablene. Næringen en bedrift tilhører kan beskrives ved de såkalte nace-kodene, der nace1 er en grovere inndeling av nace2. Vi vil her kun benytte nace2-inndelingen.

Antall ansatte er en viktig forklaringsvariabel siden en liten bedrift sjelden vil mangle like mange innenfor et yrke som en stor bedrift, og den estimerte mangelen må justeres i forhold til dette. For å unngå at bedrifter med et stort antall ansatte skal få for stor påvirkningskraft, bruker vi logaritmen til antall ansatte som forklaringsvariabel i modellen. For å kunne estimere mangelen i hvert enkelt yrke uten å lage en separat modell for hvert av yrkene, må yrke inngå som en forklaringsvariabel i modellen. Av samme årsak ønsker vi å inkludere fylke og næring i modellen. Dette medfører at den estimerte mangelen for en bedrift i et gitt yrke avhenger av hvilken næring bedriften tilhører og hvilket fylke bedriften ligger i, og man unngår at helt ulike bedrifter med samme antall ansatte får estimert den samme mangelen i det gitte yrket.

I bedriftsundersøkelsen er antall bedrifter i utvalget og antall yrker med minst én registrert mangel så høyt, spesielt i vårundersøkelsen, at et fullstendig datasett med en observasjon per yrke per bedrift blir for krevende å estimere i R. Vi velger derfor å dele opp datasettet i yrkesgrupper. Yrkesgruppene dannes ved å samle yrker med lignende yrkeskoder, se kapittel 3.2.3.

Vi vil estimere mangelen i hver yrkesgruppe som funksjon av forklaringsvariablene antall ansatte i bedriften, næring, fylke og yrke. Responsvariabelen er derfor mangelen i et gitt yrke for en bedrift i utvalget. Siden mangelen er gitt som antall personer, er verdiene til responsvariabelen et heltall større eller lik null. Vi antar at mangelen i et yrke for en bedrift er poissonfordelt.

La $Y_{i,j,k,l,g}$ være observert mangel i yrke i for bedrift j i fylke k og næring l. Vi antar videre at yrke i tilhører yrkesgruppe g. La $\mu_{i,j,k,l,g}$ være forventet mangel i yrke i i yrkesgruppe g for bedrift j i fylke k og næring l. Denne forventningen vil modelleres ved hjelp av forklaringsvariablene `log(ant.ans)`, `yrke`, `fylke` og `nace2`. Vi antar en lineær sammenheng mellom logaritmen til forventningsverdien, $\log(\mu_{i,j,k,l,g})$, og forklaringsvariablene, og modellen kan da skrives som

$$\log \mu_{i,j,k,l,g} = \beta_{1,g} \log(\texttt{ant.ans}_j) + \beta_i \cdot \texttt{yrke}_i + \beta_{k,g} \cdot \texttt{fylke}_{k,g} + \beta_{l,g} \cdot \texttt{nace2}_{l,g}. \quad (1)$$

Dette er en såkalt glm (generalisert lineær modell, (McCullagh og Nelder, 1989)). $\beta_{1,g}$ beskriver effekten av antall ansatte i bedriften. Denne parameteren vil ha en verdi for hver yrkesgruppe g. β_i beskriver yrkeseffekten, og denne vil totalt sett ha like mange nivåer som antall yrker med registrert mangel. $\texttt{fylke}_k$ og $\texttt{nace2}_l$ er indikatorvariable som er 1 for det fylket og den næringen bedriften tilhører og 0 ellers. $\beta_{k,g}$ har et nivå per

fylke for hver yrkesgruppe, mens $\beta_{l,g}$ vil ha 23 nivåer innen hver yrkesgruppe g, et for hver næring. β-ene er parametrene i modellen som må estimeres. Dette gjøres ved hjelp av *glm*-funksjonen i R.

I poissonfordelingen er variansen lik forventningsverdien, og dermed konstant. I mange tilfeller vil man observere at variansen ikke er konstant, men i stedet øker med størrelsen på observasjonene (mangelen). Dette kalles overdispersjon, og gjør at usikkerheten i den estimerte modellen underestimeres. Man kan velge å inkludere en overdispersjonsparameter i modellen, dersom poissonantagelsen ikke holder. Dersom man fortsatt antar poissonfordeling vil overdispersjonen ikke påvirke estimatene av β-ene, men bare estimatene av usikkerheten i disse. Siden vi benytter bootstrapping til å beregne usikkerheten i den estimerte mangelen og ellers ikke tar hensyn til om variablene er signifikante eller ikke, finner vi det ikke nødvendig å inkludere overdispersjon i modellen. Vi har forsøkt å tilpasse en modell der man antar negativ binomisk fordeling, og dermed har en ekstra parameter slik at variansen kan estimeres uavhengig av forventningsverdien, men det er vanskelig å oppnå konvergens i tilpasningsalgoritmen med denne fordelingen for våre data, og vi velger derfor å beholde poissonfordelingen.

2.1 Prediksjon

Modellen predikerer mangelen i hvert yrke for bedriftene i populasjonen som ikke er med i utvalget. For bedrift p i populasjonen, predikerer man forventet mangel i yrke i i yrkesgruppe g gitt at bedriften ligger i fylke k og tilhører næring l. Logaritmen til denne forventingen er gitt som

$$\log \hat{\mu}_{i,p,k,l,g} = \hat{\beta}_{1,g} \log(\texttt{ant.ans}_p) + \hat{\beta}_i \cdot \texttt{yrke}_i + \hat{\beta}_{k,g} \cdot \texttt{fylke}_{k,g} + \hat{\beta}_{l,g} \cdot \texttt{nace2}_{l,g}, \quad (2)$$

der $\hat{\beta}$-ene er de estimerte parametrene fra modellen (1). Den predikerte mangelen i yrke i for bedrift p er dermed gitt som $\hat{Y}_{i,p,k,l,g} = e^{\hat{\mu}_{i,p,k,l,g}}$.

Man predikerer en mangel i hvert yrke for hver bedrift, og finner mangelen innen hver næring eller hvert fylke ved å summere over alle yrker for alle bedrifter innen hver næring eller hvert fylke. Når man summerer over alle yrker summerer man samtidig over alle yrkesgrupper. Den predikerte mangelen innen hvert yrke finnes ved å summere den predikerte mangelen for yrket over alle bedrifter. Den totale estimerte mangelen innen hver næring, hvert fylke eller yrke er gitt som summen av den observerte og den predikerte mangelen.

Som et eksempel vil den totale estimerte mangelen i fylke 1 og næring 3 være gitt som

$$\sum_{j=1}^{n_b} \sum_{i=1}^{n_i} Y_{i,j,1,3} + \sum_{p=1}^{n_p} \sum_{i=1}^{n_i} \hat{Y}_{i,p,1,3}$$

der n_b er antall bedrifter i utvalget, n_p er antall bedrifter i resten av populasjonen og n_i er antall yrker det er registrert mangel for.

2.2 Usikkerhet

Den predikerte mangelen avhenger av parameterestimatene, som igjen avhenger av den observerte mangelen i utvalget. Dersom vi hadde hatt et annet utvalg av bedrifter ville

den predikerte mangelen være annerledes. Vi ønsker derfor å variere utvalget som brukes når man estimerer modellen, og deretter beregne usikkerheten i den estimerte mangelen. Dersom vi hadde hatt et stort utvalg eller dersom de fleste bedriftene i utvalget hadde registrert mangel i flere yrker kunne vi tilpasset modellen til forskjellige delmengder av utvalget og vurdert usikkerheten ut fra prediksjonene man da fikk. Siden utvalget vårt er såpass lite og har relativt få observerte mangler, vil vi i stedet benytte en metode som kalles bootstrapping (Efron og Tibshirani, 1993). Da trekker vi såkalte bootstraputvalg, det vil si et sett av bedrifter fra utvalget, der antallet er lik antall bedrifter i det opprinnelige utvalget, men hvor trekningen skjer med tilbakelegging, slik at samme bedrift kan trekkes ut flere ganger. Deretter estimerer vi modellen for hver yrkesgruppe for bedriftene i bootstraputvalget, predikerer mangelen i hvert yrke for bedriftene som ikke er med i det opprinnelige utvalget og summerer den observerte og predikerte mangelen innen hver næring, fylke og yrke. Dette gjøres et gitt antall ganger, og man får da et sett med observasjoner for estimert mangel innen hvert yrke, fylke og næring som man kan beregne standardavviket til. Jo flere ganger man gjentar bootstrappingen, jo sikrere vil resultatet bli. I andre situasjoner vil man ofte trekke 1 000 eller 10 000 bootstraputvalg. Dette er imidlertid tidkrevende for vår modell, og av praktiske hensyn kan man måtte nøye seg med færre.

3 Programmet

For å bruke R-programmet må man ha innfiler i riktig format. Disse spesifiseres i kapittel 3.1, sammen med en beskrivelse av R-filene som inngår i programmet og filene som skrives ut fra programmet. Bruken av programmet beskrives i kapittel 3.2.

3.1 Filer

3.1.1 Innfiler

For å kjøre R-programmet trenger man fire innfiler. Disse er

- Fil med resultater fra bedriftsundersøkelsen, kalt besvartfil.
- Fil med data for populasjonen som angir hvilke bedrifter som er med i utvalget, kalt koblingsfil.
- Fil med kolonnenavn og kolonnebredder[1] i besvartfilen, kalt NavnViddeBesvart-fil.
- Fil med kolonnenavn og kolonnebredder i koblingsfilen, kalt NavnViddeKobling-fil.

De to førstnevnte filene er .dat-filer og lages av NAV fra deres SPSS-filer. De to sistnevnte filene er .txt-filer og skal være faste, det vil si at de skal ikke endres og de samme filene kan brukes til alle bedriftsundersøkelser. Innholdet i `NavnViddeBesvart.txt` skal se slik ut:

```
row.names    navn    vidde
1    bedrnr    10
2    ant.ans     8
3    nace1     2
4    nace2     2
5    fylke     2
6    probyrk1     5
7    prb.ant1     5
8    probyrk2     5
9    prb.ant2     5
10    probyrk3     5
11    prb.ant3     5
12    probyrk4     5
13    prb.ant4     5
14    probyrk5     5
15    prb.ant5     5
16    probyrk6     5
17    prb.ant6     5
18    probyrk7    5
19    prb.ant7    5
20    probyrk8    5
21    prb.ant8    5
22    prbobyrk9    5
23    prb.ant9    5
24    probyrk10    5
25    prob.ant10   5
```

Innholdet i `NavnViddeKobling.txt` skal se slik ut:

1. Fra og med versjon 2.2 av programmet benyttes ikke kolonnebreddene ved innlesning av filer, siden variablene er skilt fra hverandre med en eller flere blanke (Günther og Løland, 2017).

```
row.names    navn    vidde
1    bedrnr    10
2    ant.ans     8
3    nace1     2
4    nace2     2
5    fylke     2
6    benytt     1
```

Kolonne	Navn	Antall tegn	Beskrivelse
1	bedrnr	10	Organisasjonsnummer
2	ant.ans	8	Antall ansatte
3	nace1	2	Næringsgruppe, nace1
4	nace2	2	Næringsgruppe, nace2
5	fylke	2	Fylke
6	probyrk1	5	Yrkeskode for første yrke med mangel
7	prb.ant1	5	Mangel i antall personer for første yrke med mangel
8	probyrk2	5	Yrkeskode for andre yrke med mangel
9	prb.ant2	5	Mangel i antall personer for andre yrke med mangel
10	probyrk3	5	Yrkeskode for tredje yrke med mangel
11	prb.ant3	5	Mangel i antall personer for tredje yrke med mangel
12	probyrk4	5	Yrkeskode for fjerde yrke med mangel
13	prb.ant4	5	Mangel i antall personer for fjerde yrke med mangel
14	probyrk5	5	Yrkeskode for femte yrke med mangel
15	prb.ant5	5	Mangel i antall personer for femte yrke med mangel
16	probyrk6	5	Yrkeskode for sjette yrke med mangel
17	prb.ant6	5	Mangel i antall personer for sjette yrke med mangel
18	probyrk7	5	Yrkeskode for sjuende yrke med mangel
19	prb.ant7	5	Mangel i antall personer for sjuende yrke med mangel
20	probyrk8	5	Yrkeskode for åttende yrke med mangel
21	prb.ant8	5	Mangel i antall personer for åttende yrke med mangel
22	probyrk9	5	Yrkeskode for niende yrke med mangel
23	prb.ant9	5	Mangel i antall personer for niende yrke med mangel
24	probyrk10	5	Yrkeskode for tiende yrke med mangel
25	prb.ant10	5	Mangel i antall personer for tiende yrke med mangel

Tabell 1. Kolonner i besvartfilen. Hver bedrift kan oppgi mangel i opptil ti yrker. Verdiene i kolonne 7–25 er 0 dersom bedriften ikke har oppgitt mangel i noen yrker.

Besvartfilen inneholder 25 kolonner, som vist i tabell 1. Det er viktig at kolonnerekkefølgen stemmer overrens med det som står i NavnViddeBesvart-filen. De seks første kolonnene inneholder generell informasjon om bedriftene i utvalget, av samme type som er tilgjengelig for alle bedrifter i populasjonen. Kolonne 6–25 inneholder bedriftenes svar på spørsmålene i spørreskjemaet som angår mangel på arbeidskraft. Hver bedrift kan oppgi mangel i inntil ti yrker, og for hvert felt i spørreskjemaet er det derfor en variabel `probyrk` som angir i hvilket yrke bedriften mangler arbeidskraft og en tilhørende variabel `prb.ant` som angir antall personer bedriften mangler i det oppgitte yrket. Dersom en

Kolonne	Navn	Antall tegn	Innhold
1	bedrnr	10	Organisasjonsnummer
2	ant.ans	8	Antall ansatte
3	nace1	2	Næringsgruppe, nace1
4	nace2	2	Næringsgruppe, nace2
5	fylke	2	Fylke
6	benytt	1	Utvalgsindikator

Tabell 2. Kolonner i koblingsfilen. Utvalgsindikatoren er 1 for bedriftene i utvalget og tom ellers.

bedrift ikke mangler arbeidskraft innen noen yrker eller i færre enn ti yrker, vil alle eller opptil ni av disse settene av `probyrk` og `prb.ant` være lik null. Slik kan for eksempel de første radene se ut:

```
810094532     24 6 17  9   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
810392312     46 2 12  3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
810441682      4 7 19  8   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
810547472     40 3  9  6   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
810574232      8 7 19 19   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811167452    175 8 23 12   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811280852     15 3  5 11   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811563382      4 6 17 16   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811626562      2 3  7  2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811660612      8 4 13  4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811662682      4 8 20  7   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811673862      8 5 14  7   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811722332     13 5 14 14   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811742732      1 1  1  3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
811833592      2 6 17 10   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

Dersom spørreskjemaet endres og bedriften skal oppgi mangel i flere eller færre enn ti yrker kan fortsatt programmet brukes uten at man må gjøre endringer i koden. Man må da enten fjerne eller legge til rader på slutten av `NavnViddeBesvart.txt`-filen tilsvarende det antallet færre/flere kolonner besvartfilen har. De første seks kolonnene må imidlertid være de samme som før og innholdet i disse må ikke endres, ellers vil ikke programmet fungere.

Koblingsfilen inneholder seks kolonner med variable for alle bedriftene i populasjonen. Variablene er beskrevet i tabell 2. Variabelen `benytt` angir hvilke bedrifter som er med i utvalget, slik at disse kan tas ut når man skal predikere mangelen for resten av bedriftene i utvalget. Verdiene i denne kolonnen er tomme for bedriftene som ikke er med i utvalget og lik 1 for bedriftene i utvalget. Kolonnerekkefølgen i koblingsfilen må stemme overrens med det som står i NavnViddeKobling-filen. De første radene i filen kan for eksempel se slik ut:

```
810094532     24 6 17  9 1
810098252      8 7 19 18
810130822      3 5 14  9
810182482      7 5 14  9
810363142      3 7 19  6
810387882     53 7 19  5
810392312     46 2 12  3 1
810438452     29 5 14 10
810441682      4 7 19  8 1
810490772     19 3  3  8
810547472     40 3  9  6 1
810574232      8 7 19 19 1
```

```
810723262       3 7 19  3
810724412      13 5 14  6
810859482      99 3  8 15
810876832       4 3 10  1
811025542       3 5 14  3
811060402       5 3  6  2
811141402       5 5 14  8
811167452     175 8 23 12 1
```

3.1.2 R-filer

Programmet består av følgende R-filer:

- `brukerdata.R`
- `kjorprogram.R`
- `funksjoner.R`

I `brukerdata.R` spesifiseres innfilene, hvor resultatfilene skal skrives ut, arbeidsområdet for R og innparametre til programmet. `kjorprogram.R` inneholder de nødvendige kommandoene for å kjøre programmet, og filen `funksjoner.R` inneholder funksjonene programmet benytter. Disse er dokumentert i selve filen. Den eneste filen brukeren trenger å endre på er `brukerdata.R`, men denne må til gjengjeld endres hver gang programmet skal kjøres for et nytt datasett.

3.1.3 Utfiler

Underveis i programmet skrives det ut flere filer. De fleste skrives til et midlertidig dataområde brukeren har definert, og kan slettes etter at hele programmet er kjørt. De endelige resultatfilene skrives ut til et resultatområde definert av brukeren. Disse er markert med * i listen nedenfor. Listen viser filene i den rekkefølgen de skrives ut:

- `InputMangelPerYrke.csv`
- `beddat.txt`
- `ObsMangelPerFylkeNace2.txt`
- `ObsMangelPerFylkeYrke.txt`
- `bedrnr.txt`
- `popdat.txt`
- `bed.dat.gruppeX.txt`
- `gruppeinndeling.txt`
- `predsperfylkeyrkeX.txt`
- `predsperfylkenaringX.txt`
- `koeffX.txt`
- `EstimertMangelPerFylkeNaring.csv`*

- EstimertMangelPerFylkeYrke.csv*
- bootstrapestimater.txt
- antallbootstrapgjennomfort.txt
- EstimertMangelPerFylkeNaringStdavvik.csv*
- EstimertMangelPerFylkeYrkeStdavvik.csv*
- EstimertMangelPerNaringStdavvik.csv*
- EstimertMangelPerFylkeStdavvik.csv*
- EstimertMangelPerYrkeStdavvik.csv*

InputMangelPerYrke.csv er en tabell med en kolonne per yrke det er registert mangel for, og en rad per bedrift i utvalget. Denne filen skrives primært ut for at brukeren skal kunne kontrollere at oppgitt mangel i hvert yrke har blitt registrert korrekt av programmet.

beddat.txt inneholder objektet som brukes for å tilpasse modellen i R, det er en såkalt data.frame som inneholder kolonnene bedrnr, respons, ant.ans, fylke, nace1, nace2, yrke og yrkegr, der respons angir observert mangel for hver bedrift i yrket angitt i yrke. Det er en rad per yrke med observert mangel per bedrift, og filen kan derfor bli ganske stor, spesielt for vårundersøkelsene.

ObsMangelPerFylkeNace2.txt inneholder en tabell med observert mangel summert over alle fylker (rader) og næringer (kolonner). Tallene i tabellen legges senere sammen med den predikerte mangelen innen hvert fylke og næring for bedriftene som ikke er med i utvalget.

ObsMangelPerFylkeYrke.txt inneholder en tabell med observert mangel summert over alle fylker (rader) og yrker (kolonner). Tallene i tabellen legges senere sammen med den predikerte mangelen innen hvert fylke og yrke for bedriftene som ikke er med i utvalget.

bedrnr.txt inneholder en liste med bedriftsnumrene til alle bedriftene i utvalget. Denne benyttes når usikkerheten i estimert mangel skal beregnes.

popdat.txt inneholder et objekt med alle bedriftene det skal predikeres mangel for, en såkalt data.frame med kolonnene bedrnr, ant.ans, fylke, nace1 og nace2. I motsetning til i beddat.txt inneholder dette objektet kun en rad per bedrift.

bed.dat.gruppeX.txt inneholder deler av den opprinnelige filen beddat.txt, det vil si de observasjonene som hører til yrkesgruppe X. Det genereres en slik fil for hver yrkesgruppe det tilpasses en modell for, altså minimum ni grupper dersom den groveste inndelingen kan brukes, og et større antall dersom en finere inndeling må benyttes.

gruppeinndeling.txt inneholder en liste over hvilke yrkesgrupper modellen tilpasses for.

`koeffX.txt` inneholder en liste over koeffisientene i den tilpassede modellen for yrkesgruppe X. Disse brukes som startverdier i modelltilpasningen når usikkerheten beregnes.

`predsperfylkeyrkeX.txt` inneholder en tabell med predikert mangel for bedriftene i populasjonen som ikke var med i utvalget, summert over fylker (rader) og yrker (kolonner). Det skrives ut en fil for hver yrkesgruppe det tilpasses en modell for, og yrkene i hver tabell tilsvarer yrkene som tilhører den gitte yrkesgruppen.

`predsperfylkenaringX.txt` inneholder en tabell med predikert mangel for bedriftene i populasjonen som ikke var med i utvalget, summert over fylker (rader) og næringer (kolonner). Det skrives ut en fil for hver yrkesgruppe modellen tilpasses for, og tallene i hver tabell tilsvarer summen over yrkene som tilhører den gitte yrkesgruppen.

`EstimertMangelPerFylkeNaring.csv` inneholder en tabell med total estimert mangel, det vil si observert pluss predikert mangel, for alle bedriftene i utvalget og for alle yrker, summert over alle fylker (rader) og næringer (kolonner).

`EstimertMangelPerFylkeYrker.csv` inneholder en tabell med total estimert mangel, det vil si observert pluss predikert mangel, for alle bedriftene i utvalget, summert over alle fylker (rader) og alle yrker (kolonner) det var registrert mangel for.

`bootstrapestimaterkjoring.txt` inneholder en tabell med estimert mangel i alle fylker, næringer og yrker som en vektor (rader) for hvert bootstraputvalg (kolonner) i usikkerhetsestimeringen.

`antallbootstrapgjennomfort.txt` inneholder ett tall som angir hvor mange bootstraputvalg programmet har trukket.

`EstimertMangelPerFylkeNaringStdavvik.csv` inneholder en tabell med estimert standardavvik for den totale estimerte mangelen i hvert fylke (rader) og hver næring (kolonner).

`EstimertMangelPerFylkeYrkeStdavvik.csv` inneholder en tabell med estimert standardavvik for den totale estimerte mangelen i hvert fylke (rader) og i hvert yrke (kolonner) det er registrert mangel for.

`EstimertMangelPerNaringStdavvik.csv` inneholder estimert standardavvik for den totale estimerte mangelen i hver næring, summert over alle fylker. I tillegg er også det estimerte standardavviket for den totale estimerte mangelen for industrinæringene 3–11 slått sammen, summert over alle fylker, og det estimerte standardavviket for den totale estimerte mangelen i hele landet, summert over alle fylker og næringer.

`EstimertMangelPerFylkeStdavvik.csv` inneholder estimert standardavvik for den totale estimerte mangelen i hvert fylke, summert over alle næringer.

`EstimertMangelPerYrkeStdavvik.csv` inneholder estimert standardavvik for den totale estimerte mangelen i hvert yrke, summert over alle fylker.

3.2 Bruk av programmet

Programmet er opprinnelig skrevet for R-versjon 2.14.1, og det anbefales å installere den nyeste versjonen. Programmet kjøres ved hjelp av ferdig definerte funksjoner som ikke skal endres av brukeren, og brukeren kan derfor ikke påvirke hva som gjøres annet enn hvilke deler av programmet som kjøres.

For å bruke programmet må man spesifisere hvor innfilene ligger og hvor resultatene skal skrives ut. For å gjøre dette åpner man `brukerdata.R` og endrer verdiene slik at de passer til analysen som skal utføres. Et eksempel på hvordan den kan se ut er gitt her:

```
#Arbeidsområdet til R hvor R-filene ligger
setwd("M:/NAV/Rkode")

#Område hvor de endelige resultatene skal skrives ut
resultatomraade<-"M:/NAV/Rkode/Resultater/H10/"

#Område der datafilene ligger
dataomraade<-"M:/NAV/Data/H10/"

#Område der midlertidige filer fra programmet skrives ut
tmpomraade <- "M:/NAV/Rkode/Tmp/H10/"

memory.limit(4000)

#navn på fil med svar fra bedriftsundersøkelsen
besvartfil<- "bedfil.dat"

#navn på koblingsfil
koblingsfil<-"kobling.dat"

#Maksimal tid i minutter programmet kan bruke paa en runde med usikkerhetsestimering
MaksKjoretid <- 240

#Antall bootstraputvalg i usikkerhetsestimeringen
AntallBootstrapUtvalg <- 100

#Angi om det er første runde av usikkerhetsberegningen (TRUE) eller ikke (FALSE)
forste<-TRUE

# Antall fylker, 19 før Nord- og Sør-Trøndelag ble slått sammen
AntallFylker <- 19

source("funksjoner.R")
source("kjorprogram.R")
```

Først skriver man inn stien til katalogen der kodefilene til programmet ligger. Dette er stedet R skal kjøres fra, og kommandoen *setwd()* setter dette området. Videre spesifiserer man resultatområdet, en katalog der tabellene med totalt estimert mangel og estimert

usikkerhet skal skrives ut, dataområdet, katalogen hvor innfilene til programmet ligger, og området de midlertidige filene skal skrives ut til. De tre områdene kan være like, og de kan også være det samme området som man kjører R fra, dette er valgfritt. Spesielt dersom man ønsker å slette alle de midlertidige filene fra programmet kan det være greit at det midlertidige området er et annet område enn resultatområdet, men ellers kan man tilpasse dette slik det passer best. Det er likevel to ting man bør passe på:

> For å unngå at programmet leser inn feil filer er det viktig at man før en analyse av en ny bedriftsundersøkelse lager nye og tomme filområder for de midlertidige filene og resultatfilene.

> Mens programmet kjører må ingen av inn- eller utfilene til programmet være åpne.

Man kan eventuelt kopiere filene til et annet område og åpne de derfra for å se på resultatene underveis.

Kommandoen *memory.limit(4000)* benyttes for at R skal kunne utnytte så mye av det tilgjengelige minnet på maskinen som mulig.

Navnet på besvartfilen med resultatene fra bedriftsundersøkelsen og navnet på koblingsfilen med alle bedriftene i populasjonen må spesifiseres. Siden disse antas å ligge på dataområdet man har spesifisert, skal man bare angi selve filnavnet her, og ikke hele stien.

MaksKjoretid angir hvor lang tid man har til rådighet for usikkerhetsestimeringen. Denne angis i minutter. Usikkerhetsestimeringen kan tidsmessig deles opp i så mange runder man ønsker og tiden kan variere fra runde til runde. Man kan for eksempel første gang sette tiden til 120 minutter, neste gang til 300 minutter og så videre.

AntallBootstrapUtvalg er en parameter som angir hvor mange bootstraputvalg man ønsker å basere usikkerhetsestimeringen på. Dersom programmet ikke rekker å gå gjennom alle bootstraputvalgene innenfor kjøretiden som er angitt, vil det neste gang man starter usikkerhetsberegningen fortsette der det slapp inntil man når det totale antallet. Parameteren skal være den samme i alle rundene av usikkerhetestimeringen, vi anbefaler at man setter den til minst 100 og så høyt som mulig innenfor tiden man har til rådighet.

forste er en indikator som angir om det er første gang programmet beregner usikkerheten i nåværende analyse eller ikke. Denne settes til enten 'TRUE' eller 'FALSE'. Hvis den settes til 'TRUE' vil programmet lage en ny fil med bootstrapresultater som eventuelt overskriver filen som måtte ligge der fra før. Hvis den settes til 'FALSE' vil programmet lese inn filen med bootstrapresultatene fra de første bootstraputvalgene og legge til resultatene fra repetisjonene i denne runden i samme fil.

AntallFylker angir antall fylker og skal være lik 19 før Nord- og Sør-Trøndelag ble slått sammen. Det er ikke nødvendig å benytte de originale fylkesnumrene (Østfold=1, Akers-

hus=2 og så videre), så lenge fylkesnumrene er unike.

Kommandoen *source("funksjoner.R")* leser filen med funksjonene programmet trenger, mens *source("kjorprogram.R")* kjører selve programmet.

Etter å ha spesifisert navnene og parametrene i `brukerdata.R`, satt arbeidsområdet og lagret endringene skriver man *source("brukerdata.R")*, og selve programmet vil da kjøres i sju trinn:

1. Innlesing av filer og generering av filer for modelltilpasning.
2. Innlesing av filer og generering av fil for prediksjon.
3. Inndeling av datasett i yrkesgrupper.
4. Modellestimering og prediksjon.
5. Oppsummering av resultater.
6. Beregning av usikkerhet.
7. Oppsummering av resultater for usikkerhet.

Hvert av stegene tilsvarer et funksjonskall som angitt i `kjorprogram.R` og kan kjøres separat gitt at stegene før er utført i forkant:

```
lesInnOgForberedTilModell(besvartfil,dataomraade,tmpomraade)
lesInnOgForberedTilPrediksjon(koblingsfil,dataomraade,tmpomraade,AntallFylker)
delInnDatasettIGrupper(tmpomraade)
estimerModell(tmpomraade,AntallFylker)
oppsummerResultater(tmpomraade,resultatomraade,AntallFylker)
beregnUsikkerhet(tmpomraade,Makskjoretid,AntallBootstrapUtvalg,forste,AntallFylker)
oppsummerResultaterUsikkerhet(tmpomraade,resultatomraade,AntallFylker)
```

Dette gjør det mulig å kjøre ulike deler av programmet til ulike tidspunkt, og dersom programmet skulle henge og man må avbryte, trenger man ikke å starte helt fra begynnelsen, men fra funksjonen man kjørte sist. Dersom man ønsker å kun kjøre deler av programmet kan man sette inn et #-tegn i `kjorprogram.R`-filen foran funksjonen(e) man ønsker å utelukke, lagre filen, og deretter kjøre programmet som før. Her er et eksempel der man har valgt å utelate estimeringen av usikkerhet:

```
lesInnOgForberedTilModell(besvartfil,dataomraade,tmpomraade)
lesInnOgForberedTilPrediksjon(koblingsfil,dataomraade,tmpomraade,AntallFylker)
delInnDatasettIGrupper(tmpomraade)
estimerModell(tmpomraade,AntallFylker)
oppsummerResultater(tmpomraade,resultatomraade,AntallFylker)
#beregnUsikkerhet(tmpomraade,Makskjoretid,AntallBootstrapUtvalg,forste,AntallFylker)
#oppsummerResultaterUsikkerhet(tmpomraade,resultatomraade,AntallFylker)
```

Her er et eksempel på hvordan man kan kjøre bare usikkerhetsestimeringen (gitt at man på et tidligere tidspunkt har tilpasset modellen):

```
#lesInnOgForberedTilModell(besvartfil,dataomraade,tmpomraade)
#lesInnOgForberedTilPrediksjon(koblingsfil,dataomraade,tmpomraade,AntallFylker)
#delInnDatasettIGrupper(tmpomraade)
```

```
#estimerModell(tmpomraade,AntallFylker)
#oppsummerResultater(tmpomraade,resultatomraade,AntallFylker)
beregnUsikkerhet(tmpomraade,Makskjoretid,AntallBootstrapUtvalg,forste,AntallFylker)
#oppsummerResultaterUsikkerhet(tmpomraade,resultatomraade,AntallFylker)
```

3.2.1 Innlesing av fil og generering av filer for modelltilpasning

Funksjonen *lesInnOgForberedTilModell* leser inn besvartfilen fra dataområdet, og finner hvilke yrker det er registrert en mangel i. For hvert yrke med registrert mangel, finner man mangelen i dette yrket for hver bedrift i utvalget. For bedrifter som ikke har registrert mangel i det gitte yrket, vil mangelen være 0. Mangelen oppsummeres i en tabell som skrives ut i filen `InputMangelPerYrke.csv` til det midlertidige filområdet.

Deretter lages R-objektet som skal brukes til gruppeinndelingen i steg tre. Det er en såkalt *data.frame* som inneholder kolonnene `bedrnr`, `respons`, `ant.ans`, `fylke`, `nace1`, `nace2`, `yrke` og `yrkegr`. For hvert bedrift (angitt ved `bedrnr`) er det like mange rader som antall yrker det er registrert mangel for totalt. Responsen, gitt ved mangelen i hvert yrke for den aktuelle bedriften, er antall personer bedriften mangler i hvert av yrkene med registrert mangel. Kolonnen `yrke` angir hvilket yrke det dreier seg om. `yrkegr` angir hvilken hovedyrkesgruppe yrket tilhører. De resterende kolonnene, `ant.ans`, `fylke`, `nace1` og `nace2`, vil ha de samme verdiene for alle rader innen en bedrift. Dette objektet skrives til det midlertidige området som en tekstfil, med navn `beddat.txt`. Dette filnavnet må ikke endres før modelltilpasningen, ellers vil ikke programmet finne igjen objektet.

I tillegg skrives de to filene `ObsMangelPerFylkeNace2.txt` og `ObsMangelPerFylkeYrke.txt` med observert mangel summert over fylke, næring og yrke ut til det midlertidige filområdet. Disse leses senere inn igjen av funksjonen *oppsummerResultater* der de legges sammen med predikert mangel slik at man får ut den totale estimerte mangelen.

3.2.2 Innlesing av fil og generering av fil til prediksjon

Funksjonen *lesInnOgForberedTilPrediksjon* leser inn koblingsfilen fra dataområdet og tar ut bedriftene det ikke skal predikeres mangel for. Disse er bedrifter med ukjent antall eller færre enn tre ansatte, bedrifter med manglende eller ugyldig nace2-verdi, bedrifter med manglende eller ugyldig fylkeverdi og bedriftene som er med i utvalget. For de resterende bedriftene lages et objekt med kolonnene `bedrnr`, `ant.ans`, `nace1`, `nace2` og `fylke` som skrives til filen `popdat.txt` på det midlertidige området. Denne filen leses senere inn igjen av funksjonen *estimerModell*, der den brukes til å predikere mangelen for bedriftene som ikke er med i utvalget for alle yrkesmodellene.

3.2.3 Inndeling av datasett i yrkesgrupper

Siden datasettet er for stort til at en enkelt modell kan tilpasses for alle yrker simultant, må modellen estimeres for mindre yrkesgrupper, der antall yrker i hver av modellene vil være antall yrker som tilhører hver av yrkesgruppene. Gruppene genereres utfra en øvre og nedre grense. Den øvre grensen er satt så stor som mulig, men slik at modellen minnemessig kan estimeres for den resulterende yrkesgruppen. Etter vår erfaring er grensen litt under 300 000, og vi har derfor satt denne til 250 000. Størrelsen på et datasett

vil være gitt av antall yrker og antall bedrifter i utvalget. Det fulle datasettet vil ha like mange rader som

$$\text{antall yrker det er registrert mangel i} \times \text{antall bedrifter i utvalget.}$$

For vårundersøkelsen 2011 blir det 3 420 568 rader. I hver yrkesgruppe bestemmes således størrelsen på datasettet av antall yrker som tilhører yrkesgruppen, som ganges med antall bedrifter i utvalget. Det er dette tallet som sjekkes mot den maksimale gruppestørrelsen.

I utgangspunktet benyttes de ni hovedyrkesgruppene 1–9, basert på første siffer i yrkeskoden. Yrker med tresifret yrkeskode blir med i gruppe 1. For hver av gruppene sjekkes det om størrelsen på datasettet er mindre enn antall yrker ganger antall bedrifter i utvalget. Dersom denne betingelsen er oppfylt, lages det et datasett for hver yrkesgruppe, disse skrives til dataområdet med navn `bed.dat.gruppeX.txt`, der $X = 1, \ldots, 9$. Dersom betingelsen ikke er oppfylt, splittes yrkesgruppen opp i undergrupper, basert på andre siffer i yrkeskoden for yrkene i yrkesgruppen. For hver av undergruppene sjekkes det om antall yrker i undergrupper ganger antall bedrifter i utvalget er mindre enn grenseverdien. Dersom betingelsen er oppfylt lages det et datasett for undergruppen som skrives til dataområdet med navn `bed.dat.gruppeX.txt`, og X er nå for eksempel er 7U1, for undergruppe 1 av yrkesgruppe 7. Dersom betingelsen ikke er oppfylt deles undergruppen inn i enda mindre grupper, denne gang etter tredje siffer i yrkeskoden. Deretter lages et datasett for undergruppen av undergruppen som skrives til dataområdet med navn `bed.dat.gruppeX` der `X` nå for eksempel er 7U1U1, for undergruppe 1 av undergruppe 1 av yrkesgruppe 7.

I tillegg til at datasettene må være små nok til at analysen kan kjøres, bør de også være så store som mulig for å unngå problemer med tilpasningen av modellen. Slike problemer kan oppstå for eksempel dersom det kun er ett yrke i yrkesgruppen og bare en bedrift har mangel i dette yrket. Etter hvert steg i oppdelingsrutinen sjekkes det derfor om de genererte yrkegruppene er større enn en minimumsgrense. Dersom minst én gruppe er mindre enn grensen, settes den minste gruppen sammen med den minste av nabogruppene og dette gjentas inntil alle grupper er store nok. Vi velger å sette sammen nabogrupper fordi vi antar at disse ligner mer på hverandre enn grupper som er langt fra hverandre. Som minimumsgrense har vi valgt 50 000 dersom det er minst 10 000 bedrifter i utvalget (typisk for en vårundersøkelse), og at minst fem yrker skal være representert i hver yrkesgruppe dersom det er færre enn 10 000 bedrifter i utvalget.

Disse gruppeinndelingene medfører typisk at man for høstundersøkelsen kun bruker hovedyrkesgruppene, men at noen av dem settes sammen dersom de har få yrker med mangel. For vårundersøkelsen er problemet som regel at gruppene er for store, og man ender opp med mange yrkesgrupper som delvis består av hovedyrkesgrupper og delvis av sammenslåtte undergrupper. Datasettene kan få navn av typen
`bed.dat.gruppe3U4U2.3.4.5.6.txt`, og i dette eksemplet betyr det at undergruppe 2, 3, 4, 5 og 6 av undergruppe 4 av hovedyrkesgruppe 3 er slått sammen. U-en i navnet refererer til undergruppe, mens punktum binder sammen de gruppene som er slått sammen.

Det første tallet angir hovedyrkesgruppe (første siffer i yrkeskode), tallet etter første U angir underyrkesgruppe (andre siffer i yrkeskode), og tallene etter andre U angir undergruppene av underyrkesgruppen. Yrkene i denne gruppen er dermed yrker som begynner på 342, 343, 344, 345 og 346, for eksempel yrkene 3421, 3422, 3423, 3431, 3433, 3441, 3450 og 3460.

3.2.4 Modelltilpasning og prediksjon

I denne delen av programmet tilpasses en modell til hvert av datasettene generert i forrige steg. Det sjekkes først om det er mer enn ett yrke i datasettet. Dersom dette er tilfelle tilpasses en poissonmodell med `log(ant.ans)`, `fylke`, `nace2` og `yrke` som forklaringsvariable, som vist i (1) i kapittel 2. Dersom det kun er ett yrke i datasettet utelates `yrke` som forklaringsvariabel. For hver modell som tilpasses skrives det ut en fil, `koeffX.txt`, til det midlertidige filområdet med koeffisientene i den tilpassede modellen. Disse benyttes som startverdier i modelltilpasningen når usikkerheten skal beregnes.

Deretter bruker man den tilpassede modellen til å predikere mangelen for bedriftene som ikke er med i utvalget. Den predikerte mangelen summeres over næring, fylke og yrke, og det skrives ut to tabeller til det midlertidige filområdet, `predsperfylkeyrkeX.txt` og `predsperfylkenaring.txt`, der *X* angir den aktuelle yrkesgruppen. Det vil være ett sett av slike filer for hver av filene `bed.dat.gruppeX.txt`. Dette steget utføres i funksjonen *estimerModell*.

3.2.5 Oppsummering av resultater

Funksjonen *oppsummerResultater* leser inn de midlertidige resultatfilene med predikert mangel for hver yrkesgruppe og summerer den over næring, fylke og yrke, legger til den observerte mangelen og skriver ut to filer til resultatområdet, `EstimertMangelPerFylkeNaring.csv` og `EstimertMangelPerFylkeYrke.csv`. Disse filene inneholder tabeller med den totale estimerte mangelen (observert pluss predikert) innen hver næring, fylke og yrke.

3.2.6 Beregning av usikkerhet

For å beregne usikkerheten i den estimerte mangelen benyttes såkalt bootstrapping. Dette gjøres i funksjonen *beregnUsikkerhet*. Her gir man inn verdien `AntallBootstrapUtvalg`, det vil si det antall bootstraputvalg usikkerheten skal baseres på. Jo høyere tall, jo mer pålitelig vil usikkerhetsestimatet være. Generelt anbefaler vi å bruke minst 100 repetisjoner, men i tilfeller der estimeringen tar lang tid, som for vårundersøkelsene, kan man eventuelt bruke 50 repetisjoner hvis man ikke har tid til flere. I tillegg til antall repetisjoner angir man hvor lenge man vil la denne funksjonen kjøre. Hvis man for eksempel angir at funksjonen kan kjøre i 240 minutter (4 timer) vil programmet gå gjennom så mange bootstraputvalg det rekker innenfor denne tiden. Før et nytt utvalg trekkes beregnes det hvor mye tid som er igjen og om det er tid nok til å estimere modellen for det nye utvalget. Når tiden er brukt opp, skrives en tabell til filen `bootstrapestimater.txt`. Denne tabellen inneholder like mange kolonner som det antallet bootstraputvalg funksjonen rakk gjennom, og en rad per celle i fylke-næring og fylke-yrke-tabellene. Tabellen

vil oppdateres for hver gang man kjører denne funksjonen, ved at det legges til kolonner for hvert bootstraputvalg. Man kan dermed velge å dele opp kjøringene i flere runder slik det passer best tidsmessig. Det er viktig å huske på å sette parameteren *forste* til FALSE for alle påfølgende runder etter den første, ellers vil programmet overskrive filen og kun bootstrapestimatene fra den siste runde vil bli lagret.

I hver runde tilpasses modellene for alle yrkesgruppene for hvert utvalg av bedrifter. For å redusere estimeringstiden benyttes koeffisientene fra modelltilpasningen av det opprinnelige utvalget som startverdier i estimeringsrutinen. Deretter predikeres mangelen for hver bedrift i resten av populasjonen og det summeres over næring, fylke og yrke. Den predikerte mangelen legges til den observerte slik at man får en total estimert mangel per næring, fylke og yrke for hvert bootstraputvalg.

3.2.7 Oppsummering av resultater for usikkerhet

Det siste steget i programmet beregner standardavvikestimatene for den estimerte mangelen basert på resultatene fra bootstrapestimeringen. Funksjonen *oppsummerResultater-Usikkerhet* leser inn filen med bootstrapestimater, `bootstrapestimater.txt` som ble generert i steg seks. Standardavviket beregnes for den estimerte mangelen for hver kombinasjon av fylke og næring og fylke og yrke. I tillegg beregnes standardavviket for den totale estimerte mangelen innen hver næring, hvert fylke og hvert yrke, for næring 3–11 slått sammen og for den totale estimerte mangelen i hele landet. Det skrives deretter ut fem filer til resultatområdet, `EstimertMangelPerFylkeNaringStdavvik.csv`, `EstimertMangelPerFylkeYrkeStdavvik.csv`, `EstimertMangelPerNaringStdavvik.csv`, `EstimertMangelPerFylkeStdavvik.csv` og `EstimertMangelPerYrkeStdavvik.csv` som inneholder tabeller med den estimerte usikkerheten i form av standardavvik for næringer, fylker og yrker. I noen tilfeller vil modellen på grunn av numeriske problemer predikere en svært høy mangel for enkelte bootstraputvalg som gjør at standardavviket blir mye større enn i normale tilfeller. Derson det estimerte standardavviket er mer enn 10 ganger så stort som det robuste medianabsoluttavviket til medianen (MAD) (Venables og Ripley, 2002), vil derfor MAD-verdien skrives ut i filen i stedet for standardavviket for de aktuelle estimatene.

4 Eksempel på bruk av program med resultater

I dette kapittelet viser vi resultatet fra høstundersøkelsen 2010 analysert med den nye modellen. Filen med svarene fra undersøkelsen har vi kalt `bedfil.dat` og koblingsfilen heter `kobling.dat`. Vi ønsket å bruke 100 bootstraputvalg til å beregne usikkerheten og vi ville la beregningen bruke opptil 12 timer, det vil si 720 minutter. Verdiene i `bruker-data.R` ble derfor spesifisert som følger:

```
#Arbeidsområdet til R hvor R-filene ligger
setwd("M:/NAV/Rkode")
#Område hvor de endelige resultatene skal skrives ut
resultatomraade<-"M:/NAV/Rkode/Resultater/H10/"
#Område der datafilene ligger
dataomraade<-"M:/NAV/Data/H10/"
#Område der midlertidige filer fra programmet skrives ut
tmpomraade <- "M:/NAV/Rkode/Tmp/H10/"
memory.limit(4000)
#navn på fil med svar fra bedriftsundersøkelsen
besvartfil<- "bedfil.dat"
#navn på koblingsfil
koblingsfil<-"kobling.dat"
#Maksimal tid i minutter programmet kan bruke paa en runde med usikkerhetsestimering
MaksKjoretid <- 720
#Antall bootstraputvalg i usikkerhetsestimeringen
AntallBootstrapUtvalg <- 100
#Angi om det er første runde av usikkerhetsberegningen (TRUE) eller ikke (FALSE)
forste<-TRUE
source("funksjoner.R")
source("kjorprogram.R")
```

Underveis i programmet skrives det ut informasjon i R-vinduet om hva som foregår i hvert enkelt steg, i tillegg vil eventuelle feilmeldinger skrives ut her. For å vise hva som blir skrevet ut går vi gjennom hvert av stegene i programmet. Første steg er å lese inn besvartfilen og gjøre klar filer til modellestimering:

```
************************************************************
Les inn og forbered til modell.
Leser inn filen M:/NAV/Data/H10/bedfil.dat
OBS! Bedrift  973000462 har registrert mangel i ukjent yrke. Mangelen settes til 0.
OBS! Bedrift  973000462 har registrert mangel i ukjent yrke. Mangelen settes til 0.
Beregner mangel i alle yrker for hver bedrift.
Datasettet skrives til filen M:/NAV//Rkode/Tmp/H10/beddat.txt
************************************************************
```

Vi får to advarsler om at det er registert mangel i ukjent yrke. Det skjer fordi de to bedriftene, 973000462 og 974740621, har hver sin observasjon der `prb.ant3` er større enn 0 (henholdsvis 20 og 4), mens den tilhørende `probyrk3` er 0. I neste steg leser programmet inn koblingsfilen og lager prediksjonsdatasettet. Da skriver programmet ut:

```
************************************************************
Les inn og forbered til prediksjon.
Leser inn filen M:/NAV/Data/H10/kobling.dat

 OBS! Bedrift  871850992 har nace2=88 og tas ut av datasettet

 OBS! Bedrift  871881332 har nace2=88 og tas ut av datasettet

 OBS! Bedrift  871905002 har nace2=88 og tas ut av datasettet

 OBS! Bedrift  871942102 har nace2=88 og tas ut av datasettet

 OBS! Bedrift  871977992 har nace2=88 og tas ut av datasettet

 OBS! Bedrift  872064052 har nace2=88 og tas ut av datasettet

 OBS! Bedrift  872070532 har nace2=88 og tas ut av datasettet

 OBS! Bedrift  872079092 har nace2=88 og tas ut av datasettet

 OBS! Bedrift  872094652 har nace2=88 og tas ut av datasettet

...

 OBS! Bedrift  995764229 har nace2=88 og tas ut av datasettet
Skriver datasettet til filen M:/NAV/Rkode/Tmp/H10/pop.dat.txt.
************************************************************
```

Her får vi advarsler om at en del bedrifter har `nace2=88` og at disse derfor tas ut av datasettet siden denne verdien betyr at næringen er ukjent. Etter å ha lest inn besvartfilen og koblingsfilen, skal programmet dele opp datasettet i besvartfilen i yrkesgrupper. Utskriften fra programmet er:

```
************************************************************
Del inn datasett i grupper.
Yrkesgruppene som skal brukes i modellen er:
1 2 3 4 5.6 7 8.9
************************************************************
```

I dette tilfellet er ingen av hovedyrkesgruppene for store slik at man måtte ha brukt andre eller tredje siffer i yrkeskoden til å definere gruppene, men to av gruppene er for små til å være egne grupper. Yrkesgruppe 6 inneholder ett yrke med mangel, mens yrkesgruppe 9 inneholder tre yrker med mangel. Siden en gruppe må bestå av minst fem yrker må gruppe 6 og 9 slås sammen med andre grupper. Gruppe 6 kan slås sammen med enten gruppe 5 eller 7, og slås sammen med den minste av disse, gruppe 5. Siden gruppe 9 kun har én nabogruppe, gruppe 8, er det gruppe 8 og 9 som slås sammen. Da sitter man igjen med sju yrkesgrupper, 1, 2, 3, 4, 5.6, 7 og 8.9.

Etter å ha definert de ulike yrkesgruppene skal det tilpasses en modell til hver av dem. Programmet gjør dette og skriver samtidig ut hvilke yrker som er i de ulike gruppene:

```
************************************************************
Estimer modell.
Tilpasser modell for yrkesgruppe 1, for yrkene
1210 1222 1223 1228 1231 1232 1312
Tilpasser modell for yrkesgruppe 2, for yrkene
2114 2130 2141 2142 2145 2147 2149 2221 2230 2310 2340 2359 2412 2413 2419 2511
2512 2521 2541 2545 2553
Tilpasser modell for yrkesgruppe 3, for yrkene
3111 3113 3114 3115 3116 3119 3120 3141 3142 3231 3310 3341 3349 3411 3412 3415
3418 3432 3460 3471 3493
Tilpasser modell for yrkesgruppe 4, for yrkene
4113 4114 4121 4131 4141 4222
Tilpasser modell for yrkesgruppe 5.6, for yrkene
5122 5123 5131 5132 5141 5164 5169 5221 5223 5224 6310
Tilpasser modell for yrkesgruppe 7, for yrkene
7121 7122 7124 7125 7127 7128 7129 7131 7132 7134 7141 7142 7212 7213 7214 7216
7221 7231 7233 7234 7237 7241 7242 7244 7311 7313 7322 7341 7350 7412 7421 7432 7436
Tilpasser modell for yrkesgruppe 8.9, for yrkene
8113 8212 8213 8251 8263 8321 8322 8323 8331 8341 9132 9152 9320
************************************************************
```

Neste steg er å oppsummere resultatene ved at observert og predikert mangel summeres per næring, yrke og fylke:

```
************************************************************
Oppsummer resultater.
Leser inn midlertidige resultater.
Endelige resultater for fylke og næring ligger i filen
M:/NAV/Rkode/Resultater/H10/EstimertMangelPerFylkeNaring.csv
Endelige resultater for fylke og yrke ligger i filen
M:/NAV/Rkode/Resultater/H10/EstimertMangelPerFylkeYrke.csv
************************************************************
```

Når man nå har estimert mangelen, kan man beregne usikkerheten. Programmet skriver ut hvilket bootstraputvalg som trekkes og hvilken yrkesgruppemodell som tilpasses. I dette tilfellet rakk programmet å gå gjennom alle bootstraputvalgene i løpet av kjøretiden som ble spesifisert, og det er derfor tilstrekkelig med én kjøring av denne delen av programmet:

```
************************************************************
Beregn usikkerhet.
Programmet kan kjøre i opptil 720 minutter i denne runden.
Dette er første runde.
Bootstraputvalg nr. 1
Tilpasser modell for yrkesgruppe 1
Tilpasser modell for yrkesgruppe 2
Tilpasser modell for yrkesgruppe 3
Tilpasser modell for yrkesgruppe 4
```

```
Tilpasser modell for yrkesgruppe 5.6
Tilpasser modell for yrkesgruppe 7
Tilpasser modell for yrkesgruppe 8.9
Bootstraputvalg nr. 2
Tilpasser modell for yrkesgruppe 1
Tilpasser modell for yrkesgruppe 2
Tilpasser modell for yrkesgruppe 3
Tilpasser modell for yrkesgruppe 4
Tilpasser modell for yrkesgruppe 5.6
Tilpasser modell for yrkesgruppe 7
Tilpasser modell for yrkesgruppe 8.9
...
Alle 100 repetisjoner er gjennomført.
Resultatet skrives til filen M:/NAV/Rkode/Tmp/H10/bootstrapestimater.txt.
************************************************************
```

Helt til slutt oppsummeres estimatene for usikkerhet, og tabeller med standardavvik for estimert mangel per næring, fylke og yrke skrives ut.

```
************************************************************
Resultatene for usikkerhetsberegningen oppsummeres.
Standardavvik for estimert mangel per fylke og næring ligger i filen
M:/NAV/Rkode/Resultater/H10/EstimertMangelPerFylkeNaringStdavvik.csv.
Standardavvik for estimert mangel per fylke og yrke ligger i filen
M:/NAV/Rkode/Resultater/H10/EstimertMangelPerFylkeYrkeStdavvik.csv.
Standardavvik for estimert mangel totalt per næring ligger i filen
M:/NAV/Rkode/Resultater/H10/EstimertMangelPerNaringStdavvik.csv.
Standardavvik for estimert mangel totalt per fylke ligger i filen
M:/NAV/Rkode/Resultater/H10/EstimertMangelPerFylkeStdavvik.csv.
Standardavvik for estimert mangel totalt per yrke ligger i filen
M:/NAV//Rkode/Resultater/H10/EstimertMangelPerYrkeStdavvik.csv.
************************************************************
```

4.1 Sammenligning av gammel og ny modell

Den estimerte mangelen for fylke og næring basert på høstundersøkelsen 2010 er gitt i tabell 3, mens tabell 4 viser den estimerte mangelen per fylke i noen utvalgte yrker. Total estimert mangel summert over alle næringer er 45 586, noe som er lavere enn den estimerte mangelen med begge de gamle modellene, se Günther og Løland (2011). Også for de andre datasettene vi har hatt til rådighet estimerer den nye modellen en lavere mangel enn den gamle, men i samme størrelsesorden. Siden mangelen per næring og fylke i den nye modellen er avhengig av mangelen i hvert enkelt yrke, vil dette gi lavere estimert mangel i næringer og fylker der mangelen i yrkene i modellen er lav. I den gamle fylke-næringmodellen inngikk ikke yrkevariabelen direkte, og dette gjorde sannsynligvis at den estimerte mangelen innen enkelte næringer og fylker ble større (og for stor).

Den estimerte mangelen i de ulike yrkene med den nye modellen avviker tildels mye fra den estimerte mangelen med den gamle (korrigerte) modellen. Dette er naturlig siden den gamle fylke-yrkemodellen er ganske ulik den nye modellen. I den gamle fylke-

yrkemodellen var det en modell for hvert yrke og forklaringsvariablene var `log(ant.ans)` og `region` der `region` kunne være enten fylke, landsdel eller hele landet. For å studere forskjellene nærmere kan vi se på yrke 7121 og 7127. Med den nye modellen er den estimerte mangelen i disse yrkene henholdsvis 1773 og 148, mens den var 638 og 889 med den gamle modellen. Tabell 5 viser hvilke bedrifter som hadde mangel i disse to yrkene. Den totale observerte mangelen i disse to yrkene var 24 og 2, noe som indikerer at yrke 7121 bør få en høyere estimert mangel enn yrke 7127, spesielt siden det kun var én bedrift som har registrert mangel i yrke 7127 mot 4 i yrke 7122. På grunn av kodefeil i det gamle programmet fikk yrke 7122 riktignok en observert mangel på 14 i stedet for 24, men bedriftene med registrert mangel var de samme. Siden ingen av disse yrkene har mangel i alle fylker eller regioner, ble fylke-yrke-modellen med landsinndeling tilpasset til begge, og det var kun antall ansatte som påvirket den estimerte mangelen. `log(ant.ans)` hadde en positiv effekt for yrke 7122, mens den var negativ for yrke 7127, men siden konstantleddet i fylke-yrkemodellen var mer negativt for yrke 7122 enn for yrke 7127, fikk yrke 7127 en høyere predikert mangel i bedrifter med få ansatte. Medianen av antall ansatte i populasjonen er 4, og derfor får man totalt sett høyere estimert mangel i yrke 7127 enn i yrke 7122. I den nye modellen har vi tilstrekkelig med observasjoner til å estimere en fylkeseffekt, i tillegg har vi med næring i modellen. For denne yrkesgruppen er fylkeseffektene små, men flere av næringsgruppene har en positiv effekt, for `nace2=13` er den estimerte effekten 1,73. Blant yrkeseffektene har yrke 7122 en stor positiv effekt på 2,08, mens effekten for yrke 7127 er –0,41. Blant bedriftene i populasjonen i fylke 8 er det mange som tilhører næring 13, og derfor får vi en høy estimert mangel i yrke 7122 i dette fylket. Dette fylket får også en ganske høy estimert mangel i yrke 7127 i forhold til de andre fylkene, men siden yrkeseffekten for yrke 7127 er mindre enn for yrke 7122 blir den estimerte mangelen lavere.

Tabellen viser at det ikke for noen av yrkene i yrkesgruppe 7 er estimert mangel i fylkene 19 og 20. Dette skyldes at det ikke er observert noen mangel i disse fylkene for denne yrkesgruppen, og fylkeseffektene derfor blir små. Effekten av de andre variablene `log(ant.ans)` og `nace2` veier ikke opp for de to fylkeseffektene for disse yrkene. Den nye modellen vil derfor ikke predikere noen mangel i de to fylkene. Med den gamle modellen vil man derimot få estimert en mangel i alle fylker for de yrkene der landsinndelingen benyttes, der det kun er antall ansatte i bedriftene som avgjør hvor stor den estimerte mangelen blir.

Tabell 6 og 7 viser de estimerte standardavvikene til den estimerte mangelen i tabell 3 og 4. Begge tabellene viser at usikkerheten i den estimerte mangelen er stor for mange kombinasjoner av næring, fylke og yrke. Dette er det vanskelig å unngå i høstundersøkelsen ettersom det er relativt få bedrifter med observert mangel, og den observerte mangelen varierer fra 1 til 30, med medianen lik 2. Enkeltbedrifter med høy observert mangel kan derfor få stor påvirkningskraft i den estimerte modellen, og dersom noen av bedriftene med observert mangel utelates, kan den resulterende estimerte mangelen bli noe helt annet. Dette reflekteres i beregningen av usikkerheten, der noen bedrifter utelates i utvalget modellen baserer seg på hver gang, og variasjonen i den estimerte mangelen blir

	Næring																							
Fylke	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Sum
1	23	2	4	1	2	14	1	17	26	55	1	1	458	416	62	77	35	4	338	12	13	76	119	1757
2	9	7	1	1	1	2	0	10	8	62	2	1	418	37	20	0	104	8	682	19	32	101	2	1527
3	4	3	3	3	1	38	1	15	9	43	6	1	698	627	100	231	258	39	3299	47	29	89	330	5874
4	13	1	0	0	7	1	0	6	6	11	0	1	198	14	1	0	20	1	116	8	9	49	1	463
5	0	2	0	1	2	1	0	5	1	6	0	0	52	0	64	0	6	1	56	3	3	10	3	216
6	10	3	1	1	2	4	4	18	4	321	3	1	136	216	70	44	56	18	926	38	48	170	79	2173
7	15	4	4	2	3	4	2	14	5	139	6	1	82	418	72	78	80	18	1062	43	59	183	140	2434
8	32	6	3	2	9	12	2	34	42	127	1	5	1407	310	106	47	24	5	419	35	21	73	81	2803
9	3	1	0	2	1	1	0	6	3	43	0	0	95	6	9	0	3	4	101	10	11	15	0	314
10	22	1	1	1	5	5	0	18	22	55	0	1	673	45	46	0	8	0	115	10	5	15	2	1050
11	103	101	9	5	8	20	0	50	90	470	6	5	1782	357	234	54	108	14	1319	42	53	145	105	5080
12	51	28	3	6	9	23	0	41	47	311	8	3	1564	169	499	16	50	15	898	64	65	120	46	4036
14	116	23	8	16	16	8	3	74	58	638	1	5	1466	183	325	21	36	47	1315	156	136	233	46	4930
15	117	14	19	6	12	17	2	109	89	509	0	7	2555	861	18	128	12	7	548	57	33	59	234	5413
16	32	15	5	2	12	18	1	48	18	55	3	3	827	324	479	89	40	15	698	21	28	57	137	2927
17	31	1	1	0	3	1	0	8	15	23	0	1	421	30	2	0	0	0	44	10	3	0	0	594
18	30	5	3	1	4	3	0	19	17	73	0	4	724	48	71	0	66	3	363	40	32	155	7	1668
19	9	1	0	0	0	4	0	2	0	20	2	1	63	0	2	18	55	2	658	10	22	128	2	999
20	10	10	6	0	1	4	0	6	1	19	0	0	84	369	165	102	17	4	279	20	16	47	168	1328
Sum	630	228	71	50	98	180	16	500	461	2980	39	41	13703	4430	2345	905	978	205	13236	645	618	1725	1502	45586

Tabell 3. Estimert mangel per fylke og næring, høstundersøkelsen 2010.

	Yrke										
Fylke	6310	7121	7122	7124	7125	7127	7128	7129	7131	7132	7134
1	23	8	66	5	13	5	5	26	13	26	62
2	0	8	63	5	13	5	5	26	13	26	63
3	44	12	100	8	21	8	8	42	21	42	100
4	0	4	29	2	6	2	2	12	6	12	29
5	0	0	0	0	0	0	0	0	0	0	0
6	13	1	8	1	2	1	1	3	2	3	8
7	26	0	0	0	0	0	0	0	0	0	0
8	14	25	200	17	42	17	17	83	42	93	200
9	0	2	12	1	3	1	1	5	3	5	12
10	0	13	103	9	22	11	9	53	22	43	103
11	15	40	314	25	62	25	27	123	62	123	296
12	5	25	202	17	44	17	17	84	47	84	204
14	6	24	195	16	40	16	16	80	40	80	193
15	42	30	238	20	50	20	20	99	50	99	258
16	22	9	75	6	17	6	6	31	16	31	75
17	0	8	67	6	14	6	6	28	14	28	67
18	0	13	101	10	23	8	8	42	21	42	103
19	0	0	0	0	0	0	0	0	0	0	0
20	26	0	0	0	0	0	0	0	0	0	0
Sum	236	222	1773	148	372	148	148	737	372	737	1773

Tabell 4. Estimert mangel per fylke og yrke for utvalgte yrker, høstundersøkelsen 2010.

Bedrift	Mangel	Yrke	Antall ansatte	Fylke	Nace2
971837292	2	7122	30	14	8
974103680	15	7122	470	11	13
974861623	4	7122	41	1	8
975132161	3	7122	55	11	8
978452418	2	7127	3	10	1

Tabell 5. Observert mangel i yrkene 7122 og 7127, høstundersøkelsen 2010.

stor. For vårundersøkelsen der vi har flere bedrifter med observert mangel, er usikkerheten mindre. Generelt er usikkerheten størst for næringer, fylker og yrker med høyest estimert mangel, slik vi forventer. I en del tilfeller er standardavviket 0. Dette skyldes enten at det ikke er estimert noen mangel for denne næringen, fylket eller yrket for noen av bootstraputvalgene, eller at den estimerte mangelen er nesten lik for alle bootstraputvalgene slik at standardavviket blir rundet av til 0.

	Næring																						
Fylke	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	22	2	9	1	1	8	2	10	13	31	1	1	257	480	66	139	39	5	229	9	12	56	142
2	11	9	1	1	2	2	1	9	9	37	1	1	364	88	13	0	89	11	472	15	26	87	5
3	4	4	4	1	1	23	2	15	6	42	10	1	406	638	67	138	192	29	2449	25	22	67	148
4	19	1	0	0	4	1	0	6	5	8	0	1	240	0	0	0	23	0	95	8	11	37	0
5	0	3	0	1	1	1	0	6	2	7	0	0	77	0	83	0	7	2	61	5	6	14	5
6	10	5	2	2	2	3	2	32	6	108	5	2	101	246	40	47	91	21	683	32	34	173	66
7	17	4	4	1	3	3	2	17	7	92	8	2	74	518	57	58	84	25	547	50	39	125	109
8	39	4	4	3	10	10	1	40	41	113	3	8	1322	233	109	52	32	4	322	40	35	52	57
9	4	2	1	0	1	0	0	4	3	62	0	0	77	14	11	0	5	7	136	12	13	25	0
10	21	1	2	1	5	4	0	14	17	45	0	3	707	94	47	0	33	0	146	14	11	45	7
11	83	35	7	2	5	12	1	24	43	139	8	4	992	310	103	66	126	15	571	30	38	94	107
12	38	16	3	8	4	13	1	32	30	108	21	4	691	194	337	24	71	20	463	35	45	91	40
14	89	20	7	19	18	7	6	45	31	469	1	7	879	191	205	41	47	44	1023	115	86	197	39
15	104	6	18	5	10	11	2	83	67	335	0	13	2262	648	17	124	23	8	347	49	29	57	218
16	29	9	5	1	6	12	1	27	13	37	25	3	502	317	300	71	90	13	1326	21	22	63	109
17	48	0	0	0	3	1	0	9	10	18	0	0	488	0	0	0	0	0	0	0	0	0	0
18	27	4	4	1	3	2	0	9	11	34	1	4	381	70	63	0	61	4	218	31	20	84	7
19	0	1	0	0	0	4	0	1	0	10	1	1	70	0	0	0	54	2	347	9	17	73	3
20	12	15	9	2	1	4	0	5	1	19	0	0	116	282	159	110	28	4	200	18	18	50	119

Tabell 6. Estimert standardavvik for estimert mangel per fylke og næring, høstundersøkelsen 2010.

	Yrke										
Fylke	6310	7121	7122	7124	7125	7127	7128	7129	7131	7132	7134
1	0	7	49	5	9	7	56	37	13	34	37
2	0	9	61	7	13	7	56	33	1	45	39
3	57	12	63	8	16	8	82	47	22	39	57
4	0	0	42	0	7	0	40	0	0	0	28
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	8	0	1	0	0	0	0	0	5
7	33	0	0	0	0	0	0	0	0	0	0
8	13	0	206	0	49	0	212	117	0	193	147
9	0	3	14	2	2	2	19	9	0	7	14
10	0	15	171	14	24	14	110	92	3	69	77
11	0	34	244	25	40	24	194	121	60	121	129
12	0	21	134	16	30	17	130	86	77	82	86
14	0	21	139	21	34	19	189	92	48	119	129
15	52	50	309	46	73	35	723	178	73	196	492
16	37	10	71	9	17	8	99	36	26	42	69
17	0	0	96	0	16	0	10	0	0	0	70
18	0	11	72	12	19	7	91	47	21	52	64
19	0	0	0	0	0	0	0	0	0	0	0
20	29	0	0	0	0	0	0	0	0	0	0

Tabell 7. Estimert standardavvik for estimert mangel per fylke og yrke for utvalgte yrker, høstundersøkelsen 2010.

Referanser

Efron, B. og Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York London: Chapmann & Hall.

Günther, C.-C. og Løland, A. (2011). Evaluering av NAVs metodegrunnlag og programvare. NR Notat SAMBA/41/2011, Norsk Regnesentral.

Günther, C.-C. og Løland, A. (2017). NAVs bedriftsundersøkelse: Programendringer. NR Notat, Norsk Regnesentral.

McCullagh, P. og Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall CRC, 2 edition.

Schweder, T. (2002). Prosjekt: Analyse av Bedriftsundersøkelsen 2002.

Venables, W. N. og Ripley, B. D. (2002). *Modern Applied Statistics with S*, chapter 5. Springer, 4 edition.