BOKMÅL

NHH

HJEMMEEKSAMEN MET4

Vår 2021

Dato: 12. mai 2021

Tidsrom: 09:00 - 13:00

Antall timer: 4

BESVARELSEN SKAL LEVERES I WISEFLOW

På våre nettsider finner du informasjon om hvordan du leverer din besvarelse: https://www.nhh.no/for-studenter/eksamen/innlevering-individuelt-og-i-gruppe/

Kandidatnummer blir oppgitt på StudentWeb i god tid før innlevering. Kandidatnummer skal være påført på alle sider øverst i høyre hjørne (ikke navn eller studentnummer). Ved gruppeinnlevering skal alle gruppemedlemmers kandidatnummer påføres.

Samarbeid mellom individer eller grupper om utarbeidelse er ikke tillatt, og utveksling av egenprodusert materiale til andre individer eller grupper skal ikke forekomme. En besvarelse skal bestå av individets, eller gruppens egne vurderinger og analyse. All kommunikasjon under hjemmeeksamen er å anse som fusk. Alle innleverte oppgaver blir behandlet i Urkund, NHHs datasystem for tekst- og plagiatkontroll

UTFYLLENDE BESTEMMELSER OM EKSAMEN

https://www.nhh.no/globalassets/for-studenter/forskrifter/utfyllende-bestemmelser-til-forskrift-om-fulltidsstudiene-ved-nhh.pdf

Antall sider, inkludert forside og vedlegg: 9

Antall vedlegg: 3 (Alle vedlegg følger etter oppgavene)

Introduksjon

Norge eksporterte laks og ørret for 74 milliarder kroner i 2020 (`www.seafood.no`), og oppdrettsfisk utgjør dermed en av våre viktigste eksportvarer. Oppdrettsnæringen har likevel en rekke utfordringer, blant annet knyttet til lakselus, som er en naturlig forekommende parasitt på laksefisk. I følge Fiskehelserapporten for 2020[1] representerer lus et av de mest alvorlige problemene i fiskeoppdrett i Norge i dag.

Hver uke skal norske oppdrettsanlegg rapportere totalt antall lus på 20 tilfeldig valgte fisk. Vi har hentet slike uketellinger fra 757 norske anlegg i tidsperioden 2011-2020. Datasettet er levert av BarentsWatch (`https://www.barentswatch.no/fiskehelse`).

Oppgave 1

I tillegg til totalt antall lus på 20 tilfeldig valgte fisk, må oppdretterne rapportere om de har gjort tiltak mot lus den uken. For enkelhets skyld skal vi anta at lusetellingen i en gitt uke skjer *etter* en eventuell behandling den uken.

Vi skal først se på effekten av to ulike behandlinger:

- Mekanisk rens av lus, der laksen spyles eller børstes ren for lus.
- Medisinsk bad, der det tilsettes et legemiddel i vannet der laksen oppholder seg.

I tabellen under finner du deskriptiv statistikk for *økningen* i antall lus per 20 fisk etter behandlingen ble gjennomført. Negative verdier betyr at lusemengden har gått ned etter behandling.

Behandling	Gj. snitt	St. avvik	Min	25%	Median	75%	Max	N
Mekanisk rens	-3.26	10.90	-225.0	-5.6	-2	0.4	131.0	8983
Medisinsk bad	-2.89	19.11	-197.6	-6.6	-1	1.4	379.6	6604

(a) Gjennomfør en test av hypotesen om at mekanisk rens ikke har en effekt mot lus. Vurder om forutsetningene for testen er oppfylt i dette tilfellet.

(b) Gjennomfør en test av hypotesen om at de to behandlingsformene har lik effekt. Vurder om forutsetningene for testen er oppfylt i dette tilfellet.

Vi ønsker å undersøke om behandlingsregimet er ulikt i forskjellige landsdeler. Vi deler oppdrettsanleggene opp i tre kategorier etter følgende regel: Sør-Norge består av alle anlegg sør for breddegrad 62.5 (ca. ved Ålesund), Midt-Norge består av anleggene mellom breddegrad 62.5 og 67.3 (ca. ved Bodø), mens Nord-Norge defineres som anleggene som ligger nord for breddegrad 67.3.

Antall ukentlige behandlinger (inkludert ingen behandling) i hele perioden fordeler seg på landsdel på følgende måte:

[1] Utarbeidet av Veterinærinstituttet:
`https://www.vetinst.no/rapporter-og-publikasjoner/rapporter/2021/fiskehelserapporten-2020`

```
         behandling
landsdel Ingen behandling Medisinsk bad Mekanisk rens
    Midt           102107          1775          3278
    Nord            99921          1275           794
    Sør            138840          2889          4911
```

I utskriften under har vi gjennomført en statistisk test på denne tabellen:

```
    Pearson's Chi-squared test

data:  kontigenstabell
X-squared = 2006.8, df = 4, p-value < 2.2e-16
```

(c) Sett opp null- og alternativhypotese for testen over, skisser utregningen av testobservatoren, konkluder, og tolk resultatet.

Oppgave 2

Vi skal nå forklare variasjon i lusenivå ved hjelp av en regresjonsmodell som bruker variablene som er beskrevet i tabellen under. Vi har én observasjon per oppdrettsanlegg per uke, og indeksen i brukes til å telle over alle kombinasjonenene av oppdrettsanlegg og uke.

Variabel	Forklaring
`loglice`	Lusenivået representert på log-skala: $\log(Y_i + 1)$, der Y_i er antall lus per 20 fisk observert i kombinasjon i av oppdrettsanlegg og uke.
`loglice_lag1`	Lusenivået på logskala i dette oppdrettsanlegget *uken før*.
`action_mechanical`	Dummyvariabel som tar verdien 1 dersom det ble utført en mekanisk avlusing i dette oppdrettsanlegget denne uken, og 0 ellers.
`action_mechanical_lag1`	Dummyvariabel som tar verdien 1 dersom det ble utført en mekanisk avlusing i dette oppdrettsanlegget *uken før*, og 0 ellers.
`action_mechanical_lag2`	Dummyvariabel som tar verdien 1 dersom det ble utført en mekanisk avlusing i dette oppdrettsanlegget *for to uker siden*, og 0 ellers.
`action_medical`	Dummyvariabel som tar verdien 1 dersom det ble utført en medisinsk avlusing i dette oppdrettsanlegget denne uken, og 0 ellers.
`action_medical_lag1`	Dummyvariabel som tar verdien 1 dersom det ble utført en medisinsk avlusing i dette oppdrettsanlegget *uken før*, og 0 ellers.
`action_medical_lag2`	Dummyvariabel som tar verdien 1 dersom det ble utført en medisinsk avlusing i dette oppdrettsanlegget *for to uker siden*, og 0 ellers.
`temperatures`	Temperaturen i sjøen ved lusetellingen.

Responsvariabelen i regresjonsmodellen i Vedlegg 1 er lusenivået `loglice`.

NB: Vi legger merke til at lusenivået er representert på en noe uvanlig log-skala $\log(Y_i + 1)$, der Y_i er antall lus per 20 fisk ved tellingen. Grunnen til at vi legger til 1 inne i logaritmen er at situasjonen med null lus ($Y_i = 0$) da blir oversatt til $\log(Y_i + 1) = \log(1) = 0$. Legg også merke til følgende punkter:

- Symbolet "log" betyr den naturlige logaritmen: $\log(x) = \ln(x)$.

- Når vi bruker denne log-skalaen så holder *ikke* den vanlige prosent-tolkningen ved log-transformasjoner. Du trenger derfor ikke knytte en bestemt måleenhet til regresjonskoeffisientene i denne oppgaven, men heller fokusere på fortegn og innbyrdes størrelsesforskjeller.
- Vi kan enkelt regne frem og tilbake mellom de to skalaene, for eksempel:
  - 5 lus per 20 fisk svarer til $\log(5+1) = 1.79$ på log-skalaen.
  - En log-skalaverdi på 1.5 svarer til $e^{1.5} - 1 = 3.48$ lus per 20 fisk.
- Du kan måtte regne frem og tilbake mellom disse skalaene i oppgavene som følger. Vær oppmerksom på hvilken skala du bruker i utregningene!

(a) Gi en praktisk fortolkning av det estimerte konstantleddet i regresjonsmodellen i Vedlegg 1.

(b) Gi en *kortfattet* og praktisk fortolkning av den resterende informasjonen i regresjonsutskriften i Vedlegg 1 (Max 200 ord).

(c) Bruk residualplottene i Vedlegg 2 til å vurdere om forutsetningene for vanlig lineær regresjon er oppfylt.

Du eier et oppdrettsanlegg der du gjennomførte en mekanisk avlusing i forrige uke. Du har ikke gjort en medisinsk avlusing i løpet av de to siste ukene. Temperaturen i sjøen er 10 grader, og i forrige ukes telling talte du 1.5 lus per 20 fisk.

(d) Prediker antall lus per 20 fisk ved tellingen denne uken ved hjelp av regresjonsmodellen i Vedlegg 1.

Det generelle kravet i lovverket er at antall lus ikke skal overstige 4 per 20 fisk[2] (dette kravet kan variere med geografisk beliggenhet og årstid). Dersom man overstiger grensen utløser det en plikt til å enten foreta en lusebehandling, eller å slakte ned bestanden og la anlegget ligge brakk en periode.

(e) Bruk resultatet fra den forrige oppgaven og estimer sannsynligheten for at lusetellingen denne uken overstiger terskelverdien på 4 lus per 20 fisk. Bruk et eller flere av diagnoseplottene i Vedlegg 2 til å vurdere kvaliteten på dette estimatet. Begrunn svaret.

Hint: Dersom du ikke fikk til oppgave **(d)** kan du i denne oppgaven anta at predikert lusenivå per 20 fisk i oppgave **(d)** er 3, uten å få poengtrekk her.

Lusetelling foretas på et lite utvalg av fiskene i oppdrettsanlegget, og er i så måte bare et *estimat* av det sanne lusenivået. Når variablene i en regresjonsmodell observeres med usikkerhet kalles det *målefeil*. Regresjonsmodellen i Vedlegg 1 har målefeil både i responsvariabelen `loglice` og i en av forklaringsvariablene, `loglice_lag1`. Det er mulig at disse to målefeilene også avhenger av hverandre, noe som kompliserer den statistiske inferensen av de estimerte regresjonskoeffisientene i Vedlegg 1.

La oss ta for oss et enklere problem, der vi analyserer en enkel lineær regresjonmodell med en forklaringsvariabel X^* og en responsvariabel Y,

$$Y = \beta_0 + \beta_1 X^* + \epsilon,$$

[2] Forskrift om bekjempelse av lakselus i akvakulturanlegg

men der vi i stedet for forklaringsvariabelen X^* observerer variabelen X, som er definert som X^* pluss tilfeldig målefeil:

$$X = X^* + e,$$

der e er en støyvariabel som er normalfordelt med forventningsverdi 0 og varians σ_e^2. Anta at X^*, e og ϵ er ukorrelerte. Det går da an å bevise at minste kvadraters estimat av koeffisienten β_1 har en forventningsverdi som er gitt ved

$$\mathrm{E}(\hat{\beta}_1) = \frac{\mathrm{Cov(X,Y)}}{\mathrm{Var}(X)} = \frac{\beta_1}{1 + \sigma_e^2/\sigma_{x^*}^2}, \tag{1}$$

der $\sigma_{x^*}^2$ er variansen til X^*.

(f) Vis at det siste likhetstegnet i ligning (1) holder. Bruk dette resultatet til å tolke konsekvensen av målefeil.

Hint: Merk at denne deloppgaven har to deler. Du kan svare på den andre delen uten å ha gjort den første delen.

En analytiker i Mattilsynet undersøker effektene ulike lusebehandlinger har på lusenivået. Han fatter interesse for regresjonsmodellen i Vedlegg 1 som vi har analysert i denne oppgaven, og mener at denne modellen bør legges til grunn i en videre regulering av oppdrettsnæringen.

(g) Gi en kortfattet kommentar til analytikeren. Hva kan vi bruke modellen til? Hva kan vi *ikke* bruke modellen til? Begrunn svaret ditt, og bruk gjerne resultater fra tidligere deloppgaver.

Oppgave 3

I Vedlegg 3 har vi plottet autokorrelasjonsfunksjonen til et stort antall observasjoner simulert fra fire forskjellige tidsrekkemodeller.

(a) Klassifiser de fire modellene som enten AR, MA eller hvit støy. Begrunn svarene dine.

Vi tar for oss de ukentlige tellingene av lus i et enkelt oppdrettsanlegg, og bruker en statistisk programvarepakke til å søke etter den beste ARIMA-modellen for denne tidsrekken. På samme måte som i oppgave 2 så modellerer vi lusetellingen på log-skalaen $\log(Y + 1)$, der Y er antall lus per 20 fisk. Vi får ut følgende resultat:

```
Series: ts_df$loglice
ARIMA(4,1,0)

Coefficients:
          ar1      ar2      ar3      ar4
      -0.6629  -0.5187  -0.2756  -0.1919
s.e.   0.0472   0.0547   0.0551   0.0475

sigma^2 estimated as 0.5103:  log likelihood=-480.08
AIC=970.16   AICc=970.3   BIC=990.78
```

(b) Hvilken modell er dette? Skriv den opp.

Vi har nettopp gjennomført lusetellingen i dette anlegget for uke t. De fem siste lusetellingene (oppgitt i antall lus per 20 fisk og på log-skala) for dette anlegget er:

Uke:	$t-4$	$t-3$	$t-2$	$t-1$	t
Telling (Y_t):	6.4	6	2	3.6	3.6
$\log(Y_t+1)$:	1.9	1.8	0.69	1.3	1.3

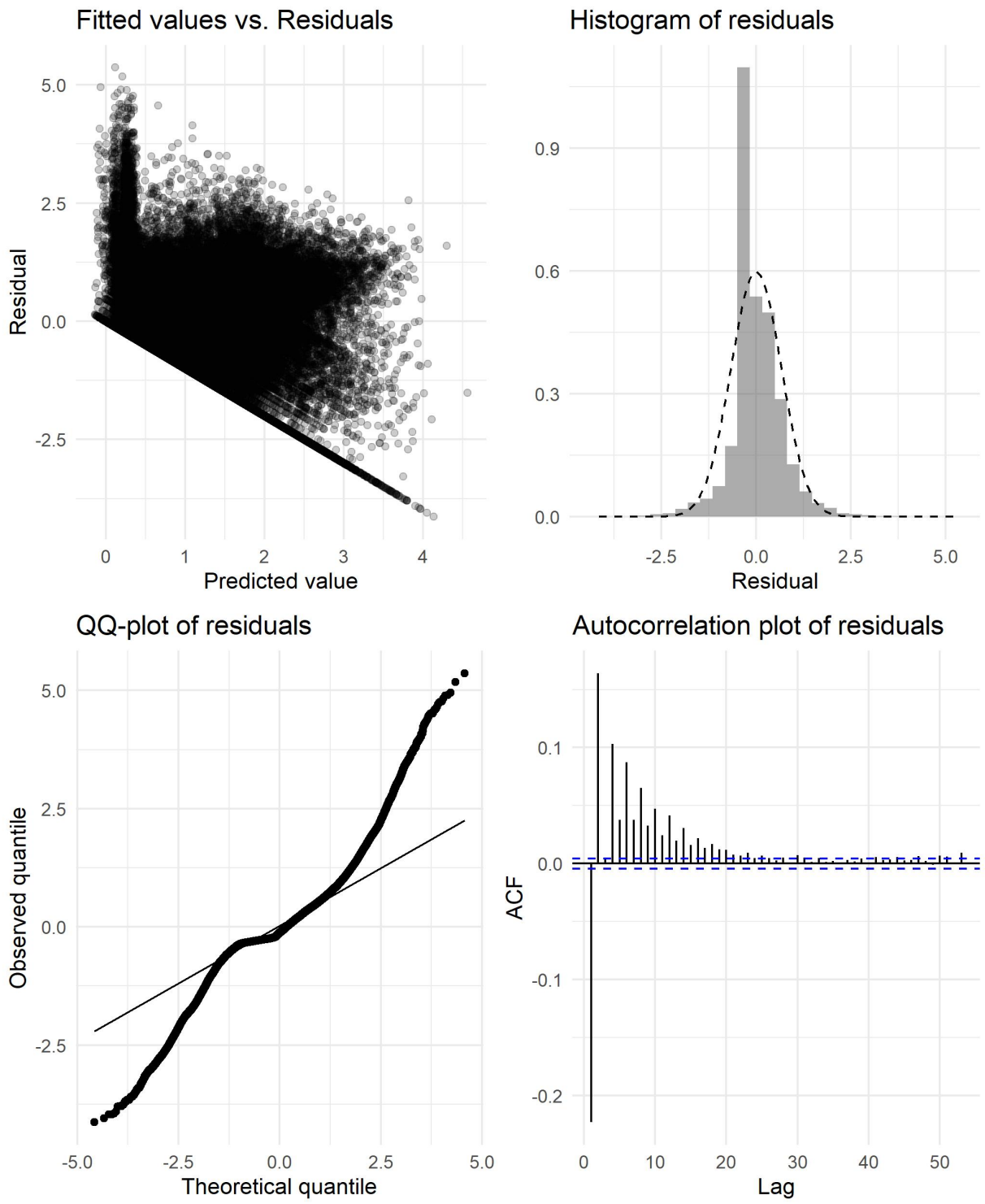
(c) Bruk modellen fra oppgave (b) til å predikere antall lus per 20 fisk i uke $t+1$. Husk å rapportere svaret på riktig skala.

Vedlegg 1: Regresjonsutskrift

```
Call:
lm(formula = loglice ~ ., data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1356 -0.3070 -0.1235  0.3498  5.3670

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.1917106  0.0041522   46.17   <2e-16 ***
loglice_lag1            0.6914138  0.0017037  405.84   <2e-16 ***
action_medical         -0.1506485  0.0087312  -17.25   <2e-16 ***
action_mechanical      -0.1942505  0.0076719  -25.32   <2e-16 ***
action_medical_lag1    -0.1137535  0.0087094  -13.06   <2e-16 ***
action_mechanical_lag1  0.1057053  0.0076200   13.87   <2e-16 ***
action_medical_lag2     0.0889638  0.0086168   10.32   <2e-16 ***
action_mechanical_lag2  0.2202553  0.0075746   29.08   <2e-16 ***
temperatures            0.0100295  0.0004047   24.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6678 on 208258 degrees of freedom
  (147523 observations deleted due to missingness)
Multiple R-squared:  0.4731,    Adjusted R-squared:  0.4731
F-statistic: 2.338e+04 on 8 and 208258 DF,  p-value: < 2.2e-16
```

Kommentar: Utskrift fra en regresjonsmodell med lusenivå på logskala i kombinasjonen av anlegg og uke som responsvariabel. Manglende observasjoner av lusetellinger skyldes at det aktuelle oppdrettsanlegget ligger brakk den aktuelle uken, og dermed ikke har fisk i sjøen.

Vedlegg 2: Residualplott



Kommentar: Diagnoseplott for residualene fra regresjonsmodellen i Vedlegg 1. Den stiplede linjen i histogrammet er tetthetsfunksjonen til en normalfordeling med samme forventningsverdi og varians som residualene.

Vedlegg 3: Fire autokorrelasjonsplott