LØSNINGSFORSLAG MET4 V21

Til sensor: Hver deloppgave teller likt og gir maksimalt 10 poeng. Se kommentarer under hver oppgave for eksempler på poengtrekk.

Oppgave 1

(a) La μ være forventet økning i antall lus etter mekaninsk rens. Vi gjør en ett-utvalgs t-test av nullhypotesen om ingen effekt mot den ensidige alternativhypotesen om at forventet økning i lusemengde er negativ:

$$H_0 : \mu = 0 \qquad \text{mot} \qquad H_A : \mu < 0.$$

Testobservator er gitt ved

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{-3.26 - 0}{10.90/\sqrt{8983}} = -28.35,$$

der $\overline{X}$ og s er henholdsvis gjennomsnittet og standardavviket til observasjonene. Med så mange observasjoner så forkaster vi den ensidige t-testen dersom testobservatoren er mindre enn -1.645. Vi har dermed en klar forkastning av nullhypotesen, og konkluderer med at forventet forandring i lusetellingen etter mekanisk rens av fisk er negativ.

Vi ser fra den deskriptive statistikken at fordelingen til observasjonene er skjev (gjennomsnitt og median er forskjellige) og har tunge haler (min/max er svært liten/stor), så observasjonene er ikke normalfordelte. Siden utvalgsstørrelsen er stor kan vi ut fra sentralgrenseteoremet likevel regne med at gjennomsnittet er nær normalfordelt. Vi må også regne med at det er avhengighet både i tid og mellom anlegg som ligger nær hverandre, men igjen gjør det store antallet observasjoner at vi får en soleklar forkastning, så vi kan med rimelighet si at forskjellen er statistisk signifikant.

Til sensor: Utførelse av test og diskusjon om antakelser teller likt, og gir 5p hver. For første del trekkes 2p for tosidig test, 2p for regnefeil men ellers korrekt oppsett, og 4p for riktig regning men gal konklusjon.

(b) La μ_1 og μ_2 være forventet forskjell i lusetelling etter henholdsvis medisinsk bad og mekanisk rens. Vi skal teste

$$H_0 : \mu_1 = \mu_2 \qquad \text{mot} \qquad H_A : \mu_1 \neq \mu_2.$$

Den ene variansen er omtrent fire ganger så stor som den andre ($20^2/10^2 = 4$). Med så mange observasjoner vil enhver forskjell være statistisk signifikant, så vi antar ulik varians her (det er selvsagt ok å sette opp denne testen formelt).

Testobservatoren er

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{-2.89 - (-3.26)}{\sqrt{\frac{19.11^2}{6604} + \frac{10.90^2}{8983}}} = 1.41.$$

Eksakt antall frihetsgrader er gitt ved Welch's formel, men i dette tilfellet er antall observasjoner så stort at vi uansett får kritisk verdi på 1.96. Siden $|T| < 1.96$ forkaster vi *ikke* nullhypotesen om ingen forskjell. Det er samme konklusjon om antakelsene som i (a).

Til sensor: Utførelse av test teller 8p og diskusjon om antakelser teller 2p. For første del trekkes 3p for ensidig test, 3p for regnefeil men ellers korrekt oppsett, og 6p for riktig regning men gal konklusjon.

(c) Vi har gjennomført en χ^2-test for uavhengighet:

$$H_0 : \text{Behandlingsmetoder og landsdel er uavhengige kjennetegn}$$
$$H_A : \text{Behandlingsmetoder og landsdel er ikke uavhengige kjennetegn}$$

I oppgaven har vi fått oppgitt de observerte frekvensene f_{ij} for alle kombinasjoner av landsdel og behandlingstype. Vi trenger også de gruppevise summene, og setter opp følgende tabell:

	Ingen behandling	Medisinsk bad	Mekanisk behandling	Sum
Midt	102 107	1 775	3 278	107 160
Nord	99 921	1 275	794	101 990
Sør	138 840	2 889	4 911	146 640
Sum	340 868	5 939	8 983	355 790

For å regne ut testobservatoren trenger vi også de *forventede* frekvensene e_{ij} under nullhypotesen om uavhengighet, som vi regner ut som $e_{ij} = f_{i\bullet} f_{\bullet j}/n$. For Midt-Norge og ingen behandling blir det $e_{11} = 340868 \cdot 107160/355790 = 102666$. Testobservatoren er gitt ved

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \frac{(102107 - 102665)^2}{102665} + \cdots = 2006.8,$$

der vi henter verdien for testobservatoren fra utskriften i oppgaven. Under nullhypotesen er testobservatoren χ^2-fordelt med $(3-1)(3-1) = 4$ frhihetsgrader, så kritisk verdi for en test på 95% signifikansnivå er 9.49, så vi har en klar forkastning av nullhypotesen om uavhengighet mellom landsdel og hyppighet av de ulike behandlingene. Dette kan vi også se direkte fra p-verdien i utskriften, som er nesten null.

Det ser dermed ut som at det er en sammenheng mellom landsdel og hvor often man bruker de ulike behandlingsmetodene.

Til sensor: De fire delspørsmålene i denne oppgaven teller likt.

Oppgave 2

(a) Konstantleddet i denne regresjonsmodellen er et estimat av forventningsverdien til $\log(Y+1)$, der Y er antall lus på 20 fisk, ved et anlegg som ikke har utført noen lusebehandlinger denne uken eller de to foregående ukene, der det var null lus ved tellingen i forrige uke ($\log(0+1) = 0$), og der temperaturen i sjøen var null grader.

Vi ser av utskriften at konstantleddet er positivt og statistisk signifikant forskjellig fra null, og hvis vi oversetter det fra log-skalaen får vi $e^{0.1917106} - 1 = 0.21$, og en rimelig fortolkning av denne størrelsen er kanskje at den representerer et slags underliggende "lusepress"; altså at ved et "rent" og ubehandlet anlegg kan vi likevel forvente at vi neste uke vil telle 0.21 lus på 20 fisk.

Til sensor: Det viktigste er at man klarer å få frem at konstantleddet er forventet $\log(Y+1)$ der alle forklaringsvariablene er lik null – forklart på en måte som viser forståelse for situasjonen vi har for oss. Fortolkningen er "praktisk" dersom det poengteres at prediksjonen er på log-nivå, eller oversettes til antall-lus-skalaen. En ren generisk opplisting av variabelnavn og lite praktisk forståelse gir maksimalt 5 poeng. Det siste poenget om lusepress er ikke nødvendig for full pott.

(b) Alle koeffisientene i regresjonsmodellen er statistisk signifikant forskjellig fra null. Det er en klar positiv statistisk sammenheng i lusenivået fra en uke til den neste (en AR-effekt om du vil), og de to behandlingene har en statistisk negativ sammenheng med lusetellingen som gjennomføres samme uke (det var det vi så i oppgave 1). Den mekaniske behandlingen henger sammen med færre lus også

uken etter, men har en klar *positiv* "effekt" etter to uker (ikke nødvendigvis kausal). Den medisinske behandlingen har positiv koeffisient både etter en og to uker, noe som *kan* bety at effekten er mer kortvarig. Det er uansett mye jobb igjen før vi kan snakke om kausale effekter her, og det er rimelig å tenke seg at høyt lusepress fører til behandling og ikke motsatt. Det er også en positiv statistisk sammenheng mellom antall lus og temperatur i sjøen.

Forklaringskraften er forholdsvis stor på nesten 50%. Den er stort sett drevet av `loglice_lag1`, men det kan vi ikke se rett fra utskriften.

Til sensor: Legg vekt på en forståelse for hva som er den praktiske betydningen av disse koeffisientestimatene. Man kan egentlig komme veldig langt ved å bare se på fortegn. Trekk etter skjønn for generiske fortellinger som bare er av typen "en enhets økning i ditt og datt fører til ...". Vær spesielt på vakt mot veldig kausale fortolkninger, som i dette tilfellet faktisk kan bære veldig galt av sted.

(c) Residualplottene viser at antakelsene om feilleddet ikke ser ut til å være oppfylt. Spredningsplottet viser et klart mønster som ikke fanges opp i modellen. Det ser ut til å være en "rett linje" av punkter som ganske sikkert representerer observasjoner der man ikke har sett noen lus. I tillegg er det en stor gruppe observasjoner til venstre i plottet som ser ut til å ha en helt annen fordeling enn resten. Det er neppe konstant varians.

Histogrammet ser forsåvidt symmetrisk og klokkeformet ut, men med untak av en stor opphopning av residualer rundt 0 (Dette skyldes nok det store antallet uker med 0 lus). QQ-plottet avslører at fordelingen til residualene ser ut til å ha tykkere haler enn normalfordelingen, så det er nok en del ekstreme observasjoner modellen ikke klarer å predikere.

Til slutt ser vi at det er autokorrelasjon i residualserien. Det følger nok av at uketellinger ligger etter hverandre i datasettet, og at det er en mer kompleks avhengighet enn det vi klarer å fange opp med `loglice_lag1`-leddet vårt. Det er ganske sikkert også avhengighet mellom oppdrettsanlegg som ligger nærme hverandre geografisk, men det er ikke helt klart om det er mulig å se i autokorrelasjonsplottet vårt. I alle tilfeller: antakelsen om uavhengige feilledd er helt klart brutt.

Alt i alt passer denne lineære regresjonsmodellen dårlig til datasettet. Vi trenger et mer avansert maskineri for å kunne modellere lakselus på en skikkelig måte.

Til sensor: Vi ser etter en konklusjon om at det er mange ulike mønstre og effekter som ikke er fanget opp av regresjonsmodellen. Residualene er åpenbart ikke uavhengige trekninger fra en normalfordeling med konstant varians.

(d) Det er tre forklaringsvariabler som ikke er null; `loglice_lag1` $= \log(1.5+1) = 0.9163$, `temperatures` $= 10$, og `action_mechanical_lag1` $= 1$. Prediksjonen for $\log(Y_i+1)$ blir

$$\widehat{\log(Y_i+1)} = 0.1917 + 0.6914 \cdot 0.9163 + 0.1057 \cdot 1 + 0.01 \cdot 10 = 1.03093.$$

Det svarer til et predikert (om enn ikke forventningsrett) lusenivå på $e^{1.03093} - 1 \approx 1.8$ lus per 20 laks.

Til sensor: 7 poeng for korrekt prediksjon på log-nivå (4 poeng dersom man har glemt å logtransformere lusenivået fra forrige uke), full pott dersom det er oversatt korrekt til lus per 20 fisk. Pass på å ikke trekke dersom svaret er noe annerledes på grunn av avrundinger.

(e) I følge den lineære regresjonsmodellen vår vil neste ukes måling være en trekning (på log-skala) fra normalfordelingen med forventningsverdi 1.031 (se forrige oppgave) og med standardavvik 0.6678 (se regresjosutskriften i vedlegg 1). Terskelverdien på 4 lus per 20 fisk på log-skalaen blir: $\log(4+1) = 1.61$. Vi bruker notasjonen $W = \log(Y+1)$ og oversetter til en standard normalfordelt variabel for å bruke tabell i læreboken:

$$\begin{aligned} P(W > 1.61) &= P\left(\frac{W - 1.031}{0.6678} > \frac{1.61 - 1.031}{0.6678}\right) \\ &= P(Z > 0.87) = 1 - P(Z < 0.87) = 1 - 0.8078 = 19.22\% \approx 20\%. \end{aligned}$$

Vi har fra vår analyse av residualplottene (histogrammet med normafordelingen og QQ-plottet) at halene til resiualfordelingen er *tykkere* enn i normalfordelingen, samtidig som det er en opphopning av residualer rundt 0. Normalantagelsen som ligger til grunn for utregningen over er derfor brutt og vi bør derfor ikke fatte viktige beslutninger basert på dette estimatet. Merk at det ikke går an å argumentere med sentralgrenseteoremet, da utregningen hviler på normalantagelsen for feilleddet.

Dersom vi bruker den oppgitte verdien på lusepredikskjon får vi i stedet en prediksjon på logskala på $\log(3+1) = 1.386$, og når vi setter det inn i utregningen får vi

$$\begin{aligned} P(W > 1.61) &= P\left(\frac{W - 1.386}{0.6678} > \frac{1.61 - 1.386}{0.6678}\right) \\ &= P(Z > 0.36) = 1 - P(Z < 0.36) = 1 - 0.6406 = 35.94\% \approx 36\%. \end{aligned}$$

Til sensor: 7p for riktig utregnet sannsynlighet, 3p for god diskusjon om påliteligheten til estimatet.

(f) Generelle regler for kovarians som brukes i denne oppgaven:

- Kovariansen mellom en konstant og en variabel er alltid 0.
- $\text{cov}(aX, Y) = a\text{cov}(X, Y)$
- $\text{cov}(X, X) = \text{var}(X)$
- $\text{cov}(\text{X, Y + Z}) = \text{cov}(\text{X, Y}) + \text{cov}(\text{X, Z})$
- $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$, dersom X og Y er ukorrelerte.

Vi starter med å regne ut variansen til $X = X^* + e$. Siden X^* og e er ukorrelerte har vi at

$$\text{var}(X) = \text{var}(X^* + e) = \text{var}(X^*) + \text{var}(e) = \sigma^2_{X^*} + \sigma^2_e$$

Tilsvarende, siden X^*, e og ϵ er ukorrelerte, har vi at kovariansen mellom X og Y er

$$\begin{aligned} \text{cov}(X, Y) &= \text{cov}(X^* + e, \beta_0 + \beta_1 X^* + \epsilon) \\ &= \text{cov}(X^*, \beta_0) + \text{cov}(X^*, \beta_1 X^*) + \text{cov}(X^*, \epsilon) \\ &+ \text{cov}(e, \beta_0) + \text{cov}(e, \beta_1 X^*) + \text{cov}(e, \epsilon) \\ &= \beta_1 \text{cov}(X^*, X^*) = \beta_1 \sigma^2_{X^*} \end{aligned} \tag{1}$$

Da følger det at

$$\text{E}(\hat{\beta}_1) = \frac{\text{cov}(X, Y)}{\text{var}(X)} = \frac{\beta_1 \sigma^2_{X^*}}{\sigma^2_{X^*} + \sigma^2_e} = \frac{\beta_1}{1 + \sigma^2_e / \sigma^2_{X^*}}$$

Tolkning: Vi husker at en estimator er forventningsrett dersom forventningen til estimatoren er lik populasjonsverdien vi prøver å estimere (f.eks $\text{E}(\bar{X}) = \mu$). Her ser vi at ved målefeil så vil ikke minstekvadraters metode gi et forventningsrett estimat av β_1 siden

$$\text{E}(\hat{\beta}_1) = \frac{\beta_1}{1 + \sigma^2_e / \sigma^2_{X^*}} < \beta_1$$

ettersom varianser alltid er positive. Utrykket forteller oss at vi systematisk underestimerer den sanne β_1 ved målefeil. Vi ser også at jo mindre målefeilen er (jo mindre σ^2_e er) jo mindre blir skjevheten i estimatet.

Til sensor: Utregning av utrykk gir 7 poeng, mens korrekt tolkning gir ytterligere 3 poeng og gis selv om ikke utregningen er gjort riktig.

(g) Det fremstår som en svært dårlig idè å bruke denne regresjonsmodellen til å fatte beslutninger. Fra oppgave **(c)** har vi at datasettet ikke tilfredsstiller noen av antakelsene som vi har for lineær regresjon (uavhengige, homoskedastiske og normalfordelte feilledd) som gjør at den statistiske inferensen blir upresis. Videre har vi fra oppgave **(f)** at målefeilen etter alt å dømme fører til at de estimerte regresjonskoeffisientene ikke en gang er forventningsrette.

Som en kausal analyse er en slik observasjonsstudie verdiløs. Det er ikke klart for oss hva som forårsaker hva: er det lusenivået som styrer behandlingsregimet, eller er det behandlingene som styrer lusenivået? Sannsynligvis er det en kombinasjon av begge disse effektene (og kanskje andre sammenhenger), så vi kan virkelig ikke si noe om den kausale *effekten* av de ulike behandlingene ut fra denne modellen.

Det eneste bruksområdet som fremstår som noenlunde fornuftig er å behandle modellen som en ren predisksjonsmodell, men selv der må vi være forskiktig med å regne ut prediksjonsintervaller og sannsynliheter siden feilleddene er heteroskedastiske og ikke normalfordelte (som vi så i oppgave **(e)**).

Til sensor: Gi full pott for korrekt forståelse, ellers trekk etter skjønn.

Oppgave 3

(a) Autokorrelasjonen til dataene fra Modell 1 avtar sakte når antall lag øker, noe som kjennetegner en AR modell. Autokorrelasjonen til dataene fra Modell 2 er lav for alle lag og med få untak ikke signifikant forskjellig fra 0, noe som kjennetegner en hvit støy modell. Autokorrelasjonen til dataene fra Modell 3 varierer i størrelse og er bare signifikant forskjellig fra 0 de fire første lag. Dette tyder på en MA modell, nærmere bestemt en MA(4) modell. Absoluttverdien til autokorrelasjonen til dataene fra Modell 4 avtar sakte når antall lag øker, noe som kjennetegner en AR-prosess. Det alternerende fortegnet kan bety at dette er en AR(1) prosess med negativ koeffisient.

Til sensor: Fire deler, lik vekt.

(b) Her er de $\log(Y_t + 1)$ transformerte dataene tilpasset en ARIMA(4,1,0) modell som kan skrives:

$$\begin{aligned}\Delta\log(Y_t+1) = &-0.6629\Delta\log(Y_{t-1}+1) - 0.5187\Delta\log(Y_{t-2}+1) \\ &- 0.2756\Delta\log(Y_{t-3}+1) - 0.1919\Delta\log(Y_{t-4}+1) + u_t\end{aligned}$$

der $\Delta\log(Y_t + 1) = \log(Y_t + 1) - \log(Y_{t-1} + 1)$ og hvor u_t er hvit støy med forventning 0 og varians 0.5103.

Til sensor: 2 poeng for navn på modell og 8 for korrekt modellformulering. Dersom en ARIMA(4,1,0) modell er skrevet opp for en generisk tidsrekke (på Y_t skala) gir dette 3 poengs trekk, med mindre det går klart frem at denne representer den transformerte tidsrekken. Vær liberal når det gjelder notasjon for differensiering, riktig lag etc.

(c) ARMA(4,1,0) modellen brukes først til å predikere førstedifferansen ved tidspunkt $t + 1$:

$$\begin{aligned}\widehat{\Delta\log}(Y_{t+1}+1) &= -0.6629\Delta\log(Y_t+1) - 0.5187\Delta\log(Y_{t-1}+1) - 0.2756\Delta\log(Y_{t-2}+1) - 0.1919\Delta\log(Y_{t-3}+1) \\ &= -0.6629(1.3-1.3) - 0.5187(1.3-0.69) - 0.2756(0.69-1.8) - 0.1919(1.8-1.9) \\ &= 0.0087\end{aligned}$$

Vi kan så få en prediksjon av antall lus per 20 fisk ved å løse ligningen

$$\widehat{\Delta\log}(Y_{t+1}+1) = \log(\hat{Y}_{t+1}+1) - \log(Y_t+1) = 0.0087$$

m.h.p. $\hat{Y}_{t+1}$:

$$\log(\hat{Y}_{t+1}+1) - \log(Y_t+1) = \log(\hat{Y}_{t+1}+1) - 1.3 = 0.0087$$
$$\Downarrow$$
$$\hat{Y}_{t+1} = \exp(1.3 + 0.0087) - 1 = 2.7014$$

Vi predikerer altså 2.7 lus per 20 fisk ved tidspunkt $t+1$.

Til sensor: Gi delvis poeng for riktige utregninger underveis.