LØSNINGSFORSLAG MET4 H20

Oppgave 1

(a) Dette er en test for sammenligning av to andeler. La p_1 og p_2 være de sanne andelene som vil gjespe i løpet av intervjuet i de to gruppene, og la $\widehat{p}_1 = 0.25$ og $\widehat{p}_2 = 0.294$ være de to observerte andelene. Den estimerte andelen som gjesper under nullhypotesen dersom det er ingen forskjell mellom gruppene er $\widehat{p} = (4+10)/(16+34) = 0.28$. Vi ønsker å teste følgende ensidige hypotese:

$$\mathrm{H}_0 : p_1 = p_2 \qquad \text{mot} \qquad \mathrm{H}_1 : p_1 < p_2$$

Testobservatoren er gitt ved

$$Z = \frac{\widehat{p}_2 - \widehat{p}_1}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\widehat{p}(1-\widehat{p})}} = \frac{0.294 - 0.25}{\sqrt{\left(\frac{1}{34} + \frac{1}{16}\right)0.28(1-0.28)}} = 0.323.$$

Testobservatoren er tilnæmet normalfordelt under nullhypotesen, så kritisk verdi for en ensidig test på 5% signifikansnivå er 1.645. Observert verdi av testobservatoren er 0.323, så vi forkaster ikke nullhypotesen om ingen forskjell mellom gruppene.

Programlederen tar feil når han påstår at forskjellen mellom de to gruppene er statistisk signifikant. Vi kan ikke slå fast at gjesping er smittsomt etter dette eksperimentet.

(b) Besvarelsen begynner bra. Formuleringene vitner om at studenten har forstått både denne spesifikke oppgaven, men også mer generelt hva det vil si å gjennomføre en hypotesetest. Det er positivt at vedkommende ser ut til å forstå eksplisitt forskjellen mellom populasjonsverdiene p_1, p_2 og deres estimater $\widehat{p}_1, \widehat{p}_2$. Studenten har valgt å utføre en tosidig test selv om oppgaven indikerer at en ensidig test er mer passende. Så lenge dette er gjort riktig er dette ikke så alvorlig. Så begår studenten desverre en grov regenfeil ved å glemme en "$(1/n_1 + 1/n_2)$" i nevneren på testobservatoren. Det gjør at tallverdien blir alt for liten. Akkurat i dette tilfellet blir konklusjonen likevel korrekt (se forrige oppgave), men det er bare flaks. Denne feilen er større enn en uskyldig tastefeil på kalkulatoren, så det vil neppe kunne få mer enn 4–5 av 10 poeng ved sensur.

(c) Denne besvarelsen er riktigere rent regneteknisk ved at vi får ut en korrekt verdi av testobservatoren, men den har flere mangler:

- Den er tynn. Studenten skriver ikke opp null- og alternativhypotesene, og viser liten forståelse for hva det *betyr* å gjøre en hypotesetest i denn konkrete situasjonen. Besvarelsen bærer derimot preg av å være resultat av ren formelpugging.
- Studenten er veldig uforsiktig med fortegn/absoluttverdi og det er ikke klart om dette er en ensidig eller tosidig test bortsett fra at at vedkommende bruker kritisk verdi 1.96. Vi har en negativ verdi av testobservatoren (fordi $\widehat{p}_1 - \widehat{p}_2$ står i telleren i stedet for $\widehat{p}_2 - \widehat{p}_1$) noe som selvsagt er helt greit siden Z- og t-fordelingene er symmetriske. Hvis vi antar at vedkommende utfører en to-sidig test må vi være nøye med absoluttverdi, og heller skrive "Siden $|-0.32| < 1.96$ ... ". Hva hadde skjedd om vi fikk ut en verdi på -3? Det skal jo føre til forkastning av nullhypotesen siden den ligger utenfor $[-1.96, 1.96]$ (eller $[-1.645, 1.645]$ for ensidig test), men denne studenten ville kanskje ikke fått det med seg, og skrevet at "Siden $-3 < 1.96$ forkaster vi ikke nullhypotesen...".
- Vi får enda et hint om manglende forståelse i siste setning, der det påstås at "nullhypotesen om ingen forskjell er bevist". Dette er helt feil. Hensikten med hypotesetesting er *ikke* å *bevise* noe som helst, men heller å sjekke om avviket mellom nullhypotesen og det vi faktisk observerer er så stort at vi må *forkaste* nullhypotesen. Hvis ikke det er tilfelle har vi ikke bevist at nullhypotesen stemmer, vi har bare ikke nok bevis til å slå fast at den ikke stemmer.

Besvarelsen får nok litt poeng for riktig formel og riktig regning, men trekkes for manglende forståelse. Neppe mer enn 4 av 10 poeng.

Alt i alt er begge besvarelsene svake, men på hvert sitt vis. Den første ødelegger et ellers godt svar med en grov forglemmelse i en formel, og den andre klarer ikke formidle mer enn helt overfladisk forståelse, selv om regningen i seg selv er riktig.

Oppgave 2

Merk: I denne oppgaven har det bevisst vært variasjoner i hvilken verdi koeffisientestimatet $\widehat{\beta}_{10}$ til `poltot` har i Vedlegg 1 (-0.0329, -0.0299, -0.0179, -0.0419, -0.0539). Det sanne tallet for dette datasettet er -0.0329, og løsningsforslaget er gitt for dette tallet, men med løsninger for de andre tallene i parantes i oppgave a), c) og d)

(a) I alle variasjonene av $\widehat{\beta}_{10}$ er denne effekten signifikant negativ (ved 5% nivå).

Siden `poltot` er en indeks som øker med hvor streng asylpolitikk ankomstlandet fører, så betyr en negativ effekt at vi forventer at en slik økning vil føre til færre asylsøkere.

For å være helt presis i denne tolkningen, må vi huske at responsvariabelen er logaritmen til andel asylsøkere, mens `poltot` ikke er på log-skalaen. Vi husker da tolkningen av en såkalt "log-lin-transformasjon" at 1 enhets økning i `poltot` ca. svarer til en -3.29% (-2.99%, -1.79%, -4.19%, -5.39%) avtagning i andel asylsøkere.

(b) Residualplottet avslører en litt spesiell konveks form noe som kan tyde på at sammenhengen mellom respons- og forklaringsvariablene ikke er helt lineær. Videre virker det å være antydninger til heteroskedastisitet siden det er noe større spredning for store prediksjoner av responsvariabelen. Histogrammet og QQ-plottet viser derimot ingen dramatiske avvik fra normalitet, og det er uansett så mange observasjoner at et evt. avvik ville hatt mindre å si. Autokorrelasjonsplottet indikerer at observasjonene ikke er uavhengige. Dette kommer av at både ankomstland og avreiseland går igjen i landparene. Dersom det f.eks var mange asylsøkere fra Syria i 2012 totalt sett, vil andel asylsøkere fra alle landpar hvor Syria er avreiseland ligge systematisk høyt og derfor være avhengige. Vi skal derfor være litt forsiktige med denne modellen både når det gjelder tolkning og inferens.

(c) Merk først at at med det oppgitte utrykket så er prediksjonen av `lnapps` gitt ved

$$\begin{aligned}\widehat{\texttt{lnapps}} = \widehat{\beta}_0 + \widehat{\beta}_1\texttt{bdbest} + \widehat{\beta}_2\texttt{fhcl} + \widehat{\beta}_3\texttt{fhpr} + \widehat{\beta}_4\texttt{lndist} + \widehat{\beta}_5\texttt{lngdpdest} + \widehat{\beta}_6\texttt{lngdpsource} \\ + \widehat{\beta}_7\texttt{lnsttot} + \widehat{\beta}_8\texttt{pt} + \widehat{\beta}_9\texttt{unp} + \widehat{\beta}_{10}\texttt{poltot} \\ = -10.3 + \widehat{\beta}_{10}\texttt{poltot}\end{aligned}$$

Vi får oppgitt at `poltot` $= 3.5$ og fra Vedlegg 1 finner vi at effekten av `poltot` er $\widehat{\beta}_{10} = -0.0329$. Da er prediksjonen av `lnapps` gitt ved

$$\begin{aligned}\widehat{\texttt{lnapps}} = -10.3 + (-0.0329)(3.5) = &-10.4151 \\ &(-10.4046, -10.3626, -10.4466, -10.4886)\end{aligned}$$

(d) Fra oppgave e) har vi allerede funnet en prediksjon på

$$\widehat{\texttt{lnapps}} = \widehat{\log(Y)} = -10.4151$$

Siden Y er andel søkere fra Syria og vi vet det er 17700000 innbyggere i landet så kan vi skrive $\log(Y) = \log(U/17700000)$ hvor U være antall søkere fra Syria. Vi kan da få et estimat på antall

asylsøkere $\widehat{U}$ ved å løse ligningen

$$\widehat{\texttt{lnapps}} = \log(\widehat{U}/17700000) = -10.4151$$

m.h.p $\widehat{U}$. Da får vi at

$$\widehat{U} = 17700000\exp(-10.4151) \approx 531$$

Vi predikerer altså 531 *(*536*,* 559*,* 514*,* 493*) asylsøkere fra Syria i 2020.*

Vi bemerker at dette i utgangspunktet ikke er et forventningsrett estimat av antall asylsøkere ($E(\hat{U}) = E\left(17700000\exp(\widehat{\texttt{lnapps}})\right) \neq 17700000\exp(\texttt{E}(\widehat{\texttt{lnapps}}))$) selv om $\widehat{\texttt{lnapps}}$ er forventningsrett. Det kan allikevel vises at denne estimatoren er tilnærmet forventningsrett når antall observasjoner øker. Det samme gjelder i oppgave f).

(e) I denne konkrete modellen representerer v_t-leddene globale variasjoner i antall søknader over tid. En kan f.eks tenke seg at det for et gitt år med global uro er spesielt mange søknader for alle land-par og da vil v_t for dette året være stor. De faste effektene α_i representerer det individuelle søknadsnivået for land-par i og skal ta hånd om at det vil være systematiske forskjeller mellom hvor mange søkere det er for de forskjellige land-parene. Vi kan f.eks se for oss at ankomstlandet for et gitt land-par har høy status/godt rykte i avreiselandet av andre grunner enn de kontrollvariablene vi har tatt med og at det derfor er en større andel søkere her i alle årene.

(f) Merk først at med det oppgitte utrykket for $\widehat{\eta}_{i,t}$ så er en kortere måte å formulere en prediksjon av $\log(Y_{i,t})$ gitt ved

$$\widehat{\log(Y_{i,t})} = \widehat{\eta}_{i,t} + \widehat{\beta}_{10}\texttt{poltot}_{i,t}.$$

Siden $Y_{i,t}$ er andelen av befolkningen i avreiselandet som søker asyl i ankomstlandet ved tid t så kan vi skrive $\log(Y_{i,t}) = \log(U_{i,t}/A_{i,t})$, hvor $U_{i,t}$ er antall asylsøkere mellom landpar i ved tid t. Vi kan derfor få et estimat på hvor mange asylsøkere det er mellom landpar i ved tid t ved å løse følgende ligning m.h.p. $\widehat{U}_{i,t}$:

$$\log(\widehat{U}_{i,t}/A_{i,t}) = \widehat{\eta}_{i,t} + \widehat{\beta}_{10}\texttt{poltot}_{i,t}$$

Da får vi at:

$$\widehat{U}_{i,t} = A_{i,t}\exp(\widehat{\eta}_{i,t} + \widehat{\beta}_{10}\texttt{poltot}_{i,t}) = A_{i,t}\exp(\widehat{\eta}_{i,t})\exp(\widehat{\beta}_{10}\texttt{poltot}_{i,t})$$

For å få et estimat på det totale antall asylsøkere for Norge i 2012 må vi derfor summere $\widehat{U}_{i,t}$ over alle landpar hvor Norge er ankomstland for $t = 2012$:

$$\text{Totalt antall} = \sum_{i=1}^{m}\widehat{U}_{i,t} = \sum_{i=1}^{m}A_{i,2012}\exp(\widehat{\eta}_{i,2012})\exp(\widehat{\beta}_{10}\texttt{poltot}_{i,2012})$$

Siden Norge er ankomstland for alle landpar i hvor $i = 1, \ldots, m$), vil ikke $\texttt{poltot}_{i,2012}$ variere med i. Denne svarer nemlig til Norges asylpolitikk i 2012 for hvert av disse landparene og vi kaller den derfor bare $\texttt{poltot}_{2012}$. Med andre ord kan vi skrive:

$$\text{Totalt antall} = \exp(\widehat{\beta}_{10}\texttt{poltot}_{2012})\sum_{i=1}^{m}A_{i,2012}\exp(\widehat{\eta}_{i,2012})$$

Vi kan derfor først regne ut en prediksjon av antall søkere når $\texttt{poltot}_{2012} = 3.5$. Fra Vedlegg 3 finner vi at $\beta_{10} \approx -0.0464$, og så får vi oppgitt at $\sum_{i=1}^{m} A_{i,2012} \exp(\widehat{\eta}_{i,2012}) = 13456.83$. Dermed får vi

$$\text{Totalt antall} = \exp(-0.0464 \times 3.5) \times 13456.83 = 11441$$

Dersom forslaget ble vedtatt og $\texttt{poltot}_{2012}$ blir endret til 5 får vi

$$\text{Totalt antall} = \exp(-0.0464 \times 5) \times 13456.83 \approx 10672$$

Vi estimerer altså at det ville blitt $11441 - 10672 \approx 769$ **færre asylsøkere i 2012 dersom forslaget hadde blitt vedtatt.**

Som en tilleggsopplysning her er den største `poltot` verdien observert for et OSED-land 11 og resultatet over viser at det skal være ganske så drakoniske tilstander i asylpolitikken for å endre bunken med søknader. Her kan det jo tenkes at noen av kontrollvariablene faktisk også kan betraktes som noe en kan påvirke, og kanskje dette vil være mer effektivt.

(g) En ting er å sette opp statistiske modeller og jobbe med datasett. Vi må også huske på at det bak tallene er ekste mennesker som påvirkes av politiske beslutninger. Alle betraktninger i denne retningen vil gi uttelling ved sensur.

Oppgave 3

Merk: I denne oppgaven er det bevisst to variasjoner i verdiene til y'ene i tabellen og disse gir motsatt konklusjon i klassifiseringen i a). b) vil ha samme konklusjon

(a) Vi begynner med å regne ut den euklidske avstanden mellom $(3, 3)$ og alle punktene (x_1, x_2) i datasettet vårt. F.eks er avstanden mellom $(3, 3)$ og $(3, 4)$

$$d((3, 3), (3, 4)) = \sqrt{(3-3)^2 + (3-4)^2} = 1$$

Vi kan så legge disse avstandene inn i en egen kolonne i tabellen:

Table 1: Utregnede avstander

y	x1	x2	avstand
0	3	4	1.000
0	4	5	2.236
1	5	3	2.000
0	3	6	3.000
1	4	3	1.000
1	6	2	3.162

Vi ser da at observasjon 1,3 og 5 med avstander på h.h.v. 1, 2 og 1 er de tre nærmeste naboene, og blant dem er det 2 mot 1 i flertall for å klassifisere y som en 1'er. Altså er $\hat{y} = 1$.

I den andre varianten av tabellen er det også observasjonene 1,3 og 5 som er de nærmeste naboene bare at det er 2 mot 1 i flertall for å klassifisere y som en 0'er.

Dersom oppgaven er løst visuelt, vel dette også gi uttelling.

(b) Siden vi bare har seks observasjoner vil alle verdier av $k \geq 6$ fullstendig ignorere informasjonen som ligger i forklaringsvariablene. Klassifiseringen vil da bare være basert på om det totalt sett er mest 1'ere eller 0'ere. I dette spesifikke datasettet har vi totalt tre 1'ere og tre 0'ere, så enhver majortetsavstemning

med $k \geq 6$ blir uavgjort. Altså vil det her ikke være mulig å oppnå flertall for verken 0'er eller 1'er for store verdier av k.