LØSNINGSFORSLAG MET4 H22

Til sensor: Hver deloppgave teller likt og gir maksimalt 10 poeng.

Oppgave 1

(a) Ut fra kategoriene "uncertainty, insecurity", "fear, shock" og "helplessness, perplexity" ser vi at gruppe 1 er mer opptatt av emosjonelle konsekvenser enn gruppe 2. Ut fra kategoriene "evacuation", debris, mud, water, dust" og "effort for cleaning up" er gruppe 1 også mer fokusert på praktiske problemer knyttet til flom. Gruppe 2 er derimot er svært opptatt av ødeleggelser/fare ved flom, noe vi ser ut fra kategoriene "casualties, deaths", "destruction (house, landscape)" og "material loss" sammenlignet med Gruppe 1.

(b) Vi skal for begge spørsmålene teste nullhypotesen om lik varians:

$$\sigma_{af}^2 = \sigma_{naf}^2 \quad \text{mot} \quad \sigma_{af}^2 \neq \sigma_{naf}^2$$

der σ_{af}^2 og σ_{naf}^2 er populasjonsvariansen i besvarelsene for henholdsvis gruppe 1 (affected) og gruppe 2 (not affected). Her er det lurte å sette den største estimerte variansen i telleren av testobservatoren som for begge kategoriene er gruppe 2 (not affected), slik at en en kun trenger å sammenligne testobservatoren mot kritisk verdi i høyre hale av F-fordeling (0.975 percentilen).

For spørsmålet 3 om "Financial loss" blir verdien av testobservatoren

$$F = \frac{S_{naf}^2}{S_{af}^2} = 1.38^2/1.34^2 = 1.06$$

Kritisk verdi er gitt ved 0.975 percentilen i en F-fordeling med $93-1$ og $103-1$ frihetsgrader. Her kan vi bruke en tabell eller R:

```
qf(1 - 0.05/2, df1 = 93 - 1, df2 = 103 - 1)
```

```
[1] 1.490043
```

Siden $F = 1.06 < k_{0.975}^{92,102} = 1.49$, kan vi ikke forkaste nullhypotesen om lik varians.

For spørsmålet 4 om "Time and effort" blir verdien av testobservatoren

$$F = \frac{S_{naf}^2}{S_{af}^2} = 1.39^2/1.18^2 = 1.39$$

Siden vi har like mange observasjoner som over er kritisk verdi ennå lik 1.49.

Siden $F = 1.38 < k_{0.975}^{92,102} = 1.49$ kan vi heller ikke for dette spørsmålet forkaste nullhypotesen om lik varians.

(c) Det fremgår ikke av oppgaven at vi ønsker å teste i en spesifikk retning, så det er ikke noen spesiell grunn til å gjøre ensidige tester. For begge spørsmålene utfører vi derfor følgende tosidige hypotesetest:

$$\mu_{af} = \mu_{naf} \quad \text{mot} \quad \mu_{af} \neq \mu_{naf}$$

der μ_{af} og μ_{naf} er forventningenverdien til besvarelsene i henholdsvis gruppe 1 (affected) og gruppe 2 (not affected). Fra oppgave b) vet vi at det er rimelig å anta lik varians i begge spørsmålene når vi skal utføre en to-utvalgs t-test.

For spørsmålet 3 om "Financial loss" får vi følgende testobservator:

$$T = \frac{\overline{X}_{naf} - \overline{X}_{af}}{\sqrt{S_P^2\left(\frac{1}{n_{naf}} + \frac{1}{n_{af}}\right)}} = \frac{3.95 - 3.51}{\sqrt{1.8471\left(\frac{1}{93} + \frac{1}{103}\right)}} = 2.26,$$

der vi har regnet ut en felles varians ved hjelp av følgende formel:

$$S_P^2 = \frac{(n_{naf} - 1)S_{naf}^2 + (n_{af} - 1)S_{af}^2}{n_{naf} + n_{af} - 2} = \frac{(93 - 1)1.38^2 + (103 - 1)1.34^2}{93 + 103 - 2} = 1.8471$$

Under nullhypotesen er testobsevatoren t-fordelt med $93 + 103 - 2 = 194$ frihetsgrader så kritisk verdi for en tosidig test er 1.97. Vi har så mange observasjoner at testobservatoren er tilnærmet standardnormalfordelt så det skal ikke trekkes poeng om en bruker 1.96 som kritisk verdi. Testobservatoren er lik 2.26, så **vi forkaster nullhypotesen om lik besvarelse av spørsmål 3 i de to gruppene.** Ut fra den målte effekten $3.95 - 3.51 = 0.44$, tyder dette på at flompåvirkede personer bekymrer seg mindre over finansielle tap enn gruppen som er ikke er påvirket.

For spørsmålet 4 om "Time and effort for cleaning" får vi følgende testobservator:

$$T = \frac{\overline{X}_{naf} - \overline{X}_{af}}{\sqrt{S_P^2\left(\frac{1}{n_{naf}} + \frac{1}{n_{af}}\right)}} = \frac{4.24 - 5.10}{\sqrt{1.6483\left(\frac{1}{93} + \frac{1}{103}\right)}} = -4.68,$$

der vi har regnet ut en felles varians ved hjelp av følgende formel:

$$S_P^2 = \frac{(n_{naf} - 1)S_{naf}^2 + (n_{af} - 1)S_{af}^2}{n_{naf} + n_{af} - 2} = \frac{(93 - 1)1.39^2 + (103 - 1)1.18^2}{93 + 103 - 2} = 1.6483$$

Vi bruker her samme kritisk verdi som over. Absoluttverdien av testobservatoren er lik 4.68, så **vi forkaster nullhypotesen om lik besvarelse av spørsmål 4 i de to gruppene.** Ut fra den målte effekten $4.24 - 5.10 = -0.86$ tyder dette på at flompåvirkede personer bekymrer seg mer over bruk av tid ved flom enn gruppen som er ikke er påvirket.

Oppgave 2

(a) β_1 representerer effekten av å ha publikum til stede. Hvis vi tror på modellen så svarer verdien av β_1 til forventet økning i målforskjell til fordel for hjemmelaget dersom kampen ikke er stengt for publikum. β_2 representerer en tilleggseffekt fra antallet tilskuere. Dersom $\beta_2 > 0$, for eksempel, så betyr det at vi forventer at forventet hjemmefordel øker med antall tilskuere. Konstantleddet β_0 svarer til en hjemmefordel som ikke har noe med tilstedeværelsen av publikum å gjøre (for eksempel at hjemmelaget spiller i kjente omgivelser, slipper å reise, etc).

Hvis det ikke finnes noen hjemmefordel så er $\beta_0 = \beta_1 = \beta_2 = 0$.

(b) Vi ser at koeffisienten for dummyvariabelen `opendoors` faktisk er svakt negativt, men at den ikke er statistisk signifikant forskjellig fra null. Vi kan ikke slå fast at koronanenstengningen hadde en egen effekt på hjemmefordelen. Koeffisienten til tilskuerantallet er derimot positiv og statistisk signifikant forskjellig fra null; en økning på 10 000 tilskuere ser ut til å henge sammen med at forventet målforskjell øker med omtrent 0.1 mål i favør hjemmelaget.

Videre fanger konstantleddet opp en generell hjemmefordel på 0.2 mål som ser ut til å være til stede uansett om det er tilskuere eller ikke.

Forklaringsgraden er likevel svært lav, bare 0.007, så mesteparten av variasjonen i målforskjellen blir ikke fanget opp av en slik lineær modell.

(c) Feilleddet ser ut til å være heteroskedastisk. Variansen ser ut til å være større for små predikerte målforskjeller enn for store predikerte målforskjeller.

(Dette kan ha følgende forklaring, som ikke kreves for full pott: Stor forventet målforskjell betyr at det er mange tilskuere på kampen, så da er det gjerne snakk om et "stort" lag som spiller på hjemmebane, og "store" lag er typisk også "gode" lag, som vi kan forvente vinner på hjemmebane med større sikkerhet, dvs med lavere varians, enn "små" lag.)

Ellers ser histogrammet ut til å ha omtrent den samme formen som normalfordelingen. QQ-plottet viser likevel at residualene har litt tyngre haler enn det vi hadde forventet om de var normalfordelte.

Datasettet er et slags panel, der vi har observert mange lag ved flere tidspunkt. Et autokorrelasjonsplott ville ikke vært lett å tolke (og vi har det heller ikke oppgitt), men det er å forvente at observasjonene ikke er uavhengige fra hverandre, siden lagene kan ha bølger med god eller dårlig form.

Alt i alt er ikke resultatene fra modellen veldig overbevisene slik den er satt opp nå. Den har lav forklaringskraft, og forutsetningene for OLS er heller ikke oppfylt.

(d) I de to modellene er utsagnet "det er ingen sammenheng mellom hjemmefordel og tilstedeværelse av publikum" ekvivalent med "$\beta_1 = \beta_2 = 0$". Det er nettopp denne hypotesen som F-testen tar seg av. Vi ser i utskriften at hypotesen blir forkastet i modell (1), men ikke i modell (2).

Vi var inne på en mulig forklaring i forrige spørsmål (men vi er nøye med å legge til at vi ikke krever noen spesiell forståelse av veldig fotballspesifikke fenomener for å få full pott), nemlig at "antall tilskuere" ikke bare måler "tilskuereffekten", men også hvor "stort" (og dermed til en viss grad hvor "godt") et hjemmelag er, og da til slutt hvilket resultat vi kan forvente å få.

Når vi i stedet forklarer målforskjellen med hvor full stadion er, så forsvinner denne effekten; det er nå ikke noen statistisk signifikant sammenheng mellom tilskuervariablene og målforskjellen.

(e) Modellene i Vedlegg 3 har svært lav forklaringskraft. Vi kan forvente at hvilke lag som faktisk spiller vil ha mye å si for hvilket resultat vi kan forvente å få. Ved å kontrollere for variasjonen som forklares av hvem som spiller tar vi høyde for at enkelte lag er bedre enn andre, og vi kan ha et håp om at eventuelle tilskuereffekter kommer klarere frem.

Det er heller ikke utenkelig at ulike dommere i større eller mindre grad systematisk favoriserer hjemmelaget, og at vi kan få en bedre modell ved å inkludere også det som faste effekter.

(f) Når vi tar med faste effekter for hjemmelag, bortelag og dommer forsvinner hjemmefordelen helt fra modell (1) i Vedlegg (5). En formell sjekk for dette ville vært å teste om $\beta_0 = \beta_1 = \beta_2 = 0$, men vi har ikke oppgitt informasjonen som trengs for å gjøre det.

I modell (2) ser vi at koeffisienten til dummyvariabelen `yellowdiff` er signifikant negativ. Det kan tyde på at dommerne faktisk lar seg påvirke av hjemmepublikumet til å være strengere med bortelaget enn med hjemmelaget. Etter koronanedstengningen ser det ut til at forventet forskjell i antall gule kort gitt til hjemme- og bortelaget gikk ned med 0.23, selv etter at vi har kontrollert for hjemmelag, bortelag og dommer. Forklaringsgraden til de to modellene i Vedlegg 5 er betraktelig større enn modellene i Vedlegg 3 som ikke hadde faste effekter.

Oppgave 3

(a) Vi skriver for enkelhets skyld $p = P(\text{lykkes} \mid \text{kapital} = x_1, \text{markedsbehov} = x_2, \text{konkurrenter} = x_3)$. Den estimerte modellen er da:

$$p = \frac{\exp(z)}{1+\exp(z)}, \quad z = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 = -3.521 + 0.015x_1 + 0.217x_2 - 0.248x_3.$$

For konstantleddet vil $e^{-3.521} = 0.029$ være oddsen for at startupen lykkes dersom de andre forklaringsvariablene er 0. Sannsynligvis finnes det ingen slike startups ettersom vi da må ha en startup

som har et produkt med 0 markedsbehov. Videre ser vi at en 1000 NOK økning i startkapital er assosiert med at oddsen for at en startupen lykkes øker med en multiplikativ faktor $e^{0.015} = 1.015$ (1.5%). En enhets økning i markedsbehov er assosiert med at den samme oddsen øker med en multiplikativ faktor $e^{0.217} = 1.24$ (24%). Til slutt ser vi at en ekstra konkurrent er assosiert med at oddsen avtar med en multiplikativ faktor $e^{-0.248} = 0.78$ (-22%). For disse tolkningene antar vi at de andre forklaringsvariablene er uendret.

(b) Vi regner først ut den lineære komponenten:

$$z = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 = -3.521 + 0.015 \cdot 40 + 0.217 \cdot 6 - 0.248 \cdot 1 = -1.867.$$

Da er sannsynligheten for å lykkes gitt ved

$$p = \exp(-1.867)/(1 + \exp(-1.867)) = 0.1339$$

(c) Kapitalen (x_3) er ukjent, men vi vet fra a) at for denne startupen er markedsbehovet $x_2 = 6$ og at den har en konkurrent $x_3 = 1$. Det lineære leddet kan da skrives som:

$$z = -3.521 + 0.015x_1 + 0.217 \cdot 6 - 0.248 \cdot 1 = -2.467 + 0.015x_1.$$

Videre har vi fått oppgitt at sannsynligheten for å lykkes skal være $p = 1/2$. Da skal vi altså løse følgende ligning m.h.p. kapital (z er en funksjon av x_1 som vi ser over):

$$1/2 = \frac{\exp(z)}{1 + \exp(z)} = \frac{1 + \exp(z) - 1}{1 + \exp(z)} = 1 - \frac{1}{1 + \exp(z)},$$

Flytter litt rundt:

$$1 - 1/2 = 1/2 = \frac{1}{1 + \exp(z)}$$

Siden det står 1 i telleren på begge sider må nevnerne være like:

$$2 = 1 + \exp(z) \rightarrow \exp(z) = 1 \rightarrow z = \log(1) \rightarrow z = 0,$$

Hvis vi nå plugger inn vårt utrykk for z får vi at

$$-2.467 + 0.015x_1 = 0 \rightarrow x_1 = 2.467/0.015 = 164.467$$

Med andre ord bør startupen ha $164.467 \times 1000 = 164467$ NOK i oppstartskapital for å ha 50% sannsynlighet for å lykkes (i følge vår modell vel og merke).

(d) I modell (2) ser vi at koeffisienten for markedsbehov endrer seg markant sammenlignet med modell (1) når vi utelater konkurrenter som forklaringsvariabel. Dette er et tegn på at markedsbehov ikke er en eksogen (Se siste oppgave i seminar 3) forklaringsvariabel i modell (2) . Det er rimelig å anta at antall konkurrenter er positivt assosiert med hvor stort markedsbehovet er. Koeffisenten for markedsbehovet i modell (2) fanger da opp hvordan både markedsbehovet og antall konkurrenter samlet påvirker sannsynligheten for å lykkes. Fra modell (1) ser vi at disse to effektene har motsatt påvirkning på denne sannsynligheten så i modell (2) hvor begge effektene fanges opp av markedsbehovet kanselleres disse og vi får en koeffisient som er nær 0 og ikke signifikant.