LØSNINGSFORSLAG MET4 V22

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

Oppgave 1

(a) La σ_1^2 og σ_2^2 være poengvariansen til elever i henholdsvis behandlingsgruppen og kontrollgruppen. For både USA og Kina utfører vi en F-test for nullhypotesen

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{mot} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

Siden det ikke er oppgitt noe signifikansnivå, velger vi å utføre testen med et 5% signifikansnivå.

For Kina blir testobservatoren

$$F = \frac{S_1^2}{S_2^2} = \frac{3.52^2}{2.95^2} = 1.42$$

Siden $F > k_{0.975}^{F_{322,332}} = 1.24$ forkaster vi H_0. Det er ikke grunn til å tro at variansen er lik i kontrollgruppen og behandlingsgruppen i Kina.

For USA blir testobservatoren

$$F = \frac{S_1^2}{S_2^2} = \frac{6^2}{5.64^2} = 1.13$$

Siden $F < k_{0.975}^{F_{219,225}} = 1.30$ kan vi ikke forkaste H_0, så den observerte forskjellen i standardavvik er ikke stor nok til å forkaste nullhypotesen om at de sanne standardavvikene er like.

(b) La μ_1 og μ_2 være forventet poengsum til elever i henholdsvis behandlingsgruppen og kontrollgruppen. For både USA og Kina utfører vi en to-utvalgs T-test for den tosidige hypotesen

$$H_0 : \mu_1 = \mu_2 \quad \text{mot} \quad H_1 : \mu_1 \neq \mu_2$$

Fra oppgave a) kan vi ikke anta lik varians i Kina. Testobservatoren for Kina blir derfor:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} = \frac{20.23 - 20.50}{\sqrt{3.52^2/323 + 2.95^2/333}} = -1.06$$

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 $|T| < 1.96$ kan vi ikke forkaste H_0 for Kina. Det ser ikke ut til at insentivet påvirker poengsummen kinesiske studenter får på prøven.

Fra oppgave a) kan vi anta lik varians for USA. Felles estimat for varians blir derfor

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(220 - 1)6^2 + (227 - 1)5.64^2}{220 + 227 - 2} = 33.87$$

og testobservatoren er gitt ved

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} = \frac{12.15 - 10.22}{\sqrt{33.87(1/220 + 1/227)}} = 3.51$$

Igjen er antall observasjoner så stort at vi uansett får kritisk verdi på 1.96. Siden $|T| > 1.96$ kan vi forkaste H_0 for USA. Det ser ut til at insentivet påvirker poengsummen amerikanske studenter får på prøven.

Merk at vi her har et kontrollert eksperiment, så vi kan faktisk hevde at insentiver har en positiv effekt på poengresultatet fra USA.

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

H_0 : Evnenivå og skoletilhørighet er uavhengige kjennetegn
H_A : Evnenivå og skoletilhørighet er ikke uavhengige kjennetegn

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

	[5,10]	(10,15]	(15,20]	(20,25]	Sum
skole 1	0	7	18	5	30
skole 2	0	6	49	50	105
skole 3	1	2	31	99	133
skole 4	1	2	24	38	65
Sum	2	17	122	192	333

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 skole 1 og intervallet (5, 10] blir det $e_{11} = 2 \cdot 30/333 = 0.18$. Testobservatoren er gitt ved

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \frac{(0 - 0.18)^2}{0.18} + \cdots = 56.2$$

der vi henter verdien for testobservatoren fra utskriften i oppgaven. Under nullhypotesen er testobservatoren χ^2-fordelt med $(4-1)(4-1) = 9$ frihetsgrader, så kritisk verdi for en test på 95% signifikansnivå er 16.91, så vi har en klar forkastning av nullhypotesen om uavhengighet mellom skoletilhørighet og evnenivå. 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 evnenivå og skoletilhørighet.

(d) Fra histogrammet er de to mest iøynefallende forskjellene mellom Kina og USA sentrum og spredningen: Mens USA har resultater mer eller mindre jevnt fordelt og kanskje noe mer i den nedre halvdelen av poengskalaen, har Kina de fleste resultatene konsentrert i den øvre halvdelen av poengskalaen. Dette er også reflektert av den deskriptive statistikken hvor både gjennomsnitt og median er nesten dobbelt så stort i Kina sammenlignet med USA, samt at standardavvikene for USA (5.65 og 6.00) er betydelig større enn standardavvikene i Kina (2.95 og 3.52). Vi legger også merke til at minimumsverdiene for Kina (9 og 7) er mye større enn for USA (1 og 0). Med andre ord virker det som at Kina gjør det stabilt bra på prøven, mens USA gjør det dårligere og med mye større variasjon.

Det er mye vanskeligere å se effekten av insentiver ut fra histogrammene og den deskriptive statistikken. Innad i hvert land er gruppene tilsynelatende veldig like. For USA kan vi ut fra histogrammet muligens se en ørliten forskyvning til høyre av fordelingen til behandlingsgruppen sammenlignet med kontrollgruppen.

Avvik fra normalitet er nok mest kritisk for resultatene fra Kina hvor histogrammet viser en høyreskjev fordeling. Dette er støttet av at gjennomsnitt/median og maksimumsverdier er svært nære hverandre. Vi skal generelt være skeptisk til normalitet når variablene er begrenset til et intervall (i dette tilfellet 0-25 poeng) og det er en opphopning ved en eller begge grensene. Likevel er resultatene i oppgave b) gyldige på grunn av størrelsen på utvalget og sentralgrenseteoremet.

Oppgave 2

(a) I begge modellene er koeffisienten for alder ikke statistisk signifikant forskjellig fra null, mens koeffisienten for kjønn (female) er statistisk signifikant på 5% nivå i Kina (men F-testen forkaster ikke nullhypotesen om $\beta_1 = \beta_2 = \beta_3 = 0$ på 5% nivå). Insentiver (treatment) har en klar positiv effekt i USA, mens den tilsvarende koeffisienten for Kina ikke er statistisk signifikant forskjellig fra null. Dette er et designet eksperiment med en behandling- og kontrollgruppe. Vi kan derfor hevde at insentiver har en kausal positiv effekt for poengresultatet i USA, men ikke i Kina.

Forklaringskraften til modellen er svært liten (1.1 % og 3.3%) så variasjonen i poengresultat kan ikke forklares kun av alder, kjønn og insentiv/ikke insentiv.

(b) Spredningsplottet viser at residualene er forholdsvis symmetrisk fordelt, men med en liten avtagende trend for høye prediksjonsverdier.

Histogrammet viser tydelige avvik fra normalitet og viser at residualene er venstreskjevt fordelt. I forhold til normalfordelingen er venstre hale av fordelingen for lang, mens høyre hale er for kort. Dette ser vi også igjen i QQ-plottet hvor de observerte kvantilene i venstre hale er mer negative enn det de skulle vært under normalfordelingen, mens de observerte kvantilene i høyre hale er for små forhold til det de skulle vært.

Samlet sett er det spesielt normalantagelsen til feilleddet som ikke stemmer. Det betyr at vi ikke kan stole på prediksjonsintervall som er laget ut fra modellen.

(c) Her skal vi bruke modellen for USA og setter inn forklaringsvariablene som blir oppgitt for å få ut prediksjonen:

$$\hat{Y} = 12.421 + 1.897 \times 1 - 0.107 \times 16 - 0.948 \times 0 = 12.61$$

Prediksjonsintervallet representerer usikkerheten til prediksjonen av poengresultatet for nettopp denne eleven. Hvis et stort antall elever med disse verdiene av forklaringsvariablene (16 år, gutt, behandlingsgruppe) tar prøven, forventer vi ut fra modellen at om lag 95% vil få poengsum i et 95% prediksjonsintervall.

(d) Formel for konfidensintervall for en regresjonskoeffisient β er $\widehat{\beta} \pm t_{\alpha/2,n-k-1} S(\widehat{\beta})$, der $S(\widehat{\beta})$ er det estimerte standardavviket til koeffisientestimatet. I dette tilfellet har vi såpass mange observasjoner at vi kan sette t-kvantilen til 1.96 for et 95% konfidensitervall, og får

$$[1.897 \pm 1.96 \cdot 0.552] = \mathbf{[0.82, 2.98]}.$$

(e) La $\hat{\beta}_1$ og $\hat{\beta}_1^*$ være koeffisienten for `treatment` i henholdsvis USA og Kina. Selv om vi har etablert at $\hat{\beta}_1$ er statistisk forskjellig fra 0, mens $\hat{\beta}_1^*$ ikke er statistisk forskjellig fra null, så kan vi ikke slutte at de to koeffisientene er forskjellige fra *hverandre*. Differansen $\hat{\beta}_1 - \hat{\beta}_1^*$, som da måtte vært statistisk signifikant forskjellig fra 0, har en annen (t-)fordeling og kan både ha større eller mindre varians enn hvert av leddene hver for seg p.g.a. mulig kovarians.

Den enkleste måten å undersøke dette på ville vært å sette opp en regresjonsmodell med USA (eller Kina) som en dummy variabel og et interaksjonsledd mellom USA (eller Kina) og `treatment`:

$$\texttt{score} = \beta_0 + \beta_1 \cdot \texttt{treatment} + \beta_2 \cdot \texttt{female} + \beta_3 \cdot \texttt{age} + \beta_4 \cdot \texttt{USA} + \beta_5 \cdot \texttt{treatment} \cdot \texttt{USA} + \epsilon,$$

Her representer β_5 nettopp den nevnte differansen, og først hvis β_5 i denne modellen er estimert til å være statistisk signifikant forskjellig fra null kunne vi hevdet at effekten av insentiv er forskjellig i de to landene. Leddet $\beta_5 \cdot \texttt{USA}$ tar vi med for å ta hensyn til at nivået i de to landene er så forskjellig.

Oppgave 3

(a) Studenten begynner bra og legger korrekt merke til den økende trenden og at dette er en indikasjon på at tidsrekken er ikke-stasjonæritet. Det er derimot fullt mulig å analysere ikke-stasjonære tidsrekker. To metoder som vi har lært om i dette kurset som egner seg for ikke-stasjonære tidsrekker av denne typen er dekomponering (trend, sesong og residual) og bruk av ARIMA-modeller.

(b) Vi ser at det er en statistisk signifikant positiv autokorrelasjon ved lag en, to og tre i denne tidsrekken.

(c) Utskriften viser en ARMA(1,3) modell med et konstantledd:

$$y_t = 0.7454 + 0.5885 \cdot y_{t-1} - 0.3528 \cdot u_{t-1} + 0.0846 \cdot u_{t-2} + 0.1739 \cdot u_{t-3} + u_t$$

Oppgave 4

(a) Histogrammet viser at signifikante resultater hvor $|Z| > 1.96$ i mye større grad blir publisert enn ikke-signifikante resultater hvor $|Z| < 1.96$.

En slik publikasjonsbias kalles gjerne 'skrivebordsskuff-effekten' der navnet henviser til at forskere bare legger ikke-signifikante resultater i skrivebordsskuffen i den tro at de er uinteressante eller at vitenskapelige tidsskrifter ikke vil akseptere et slikt resultat.