LØSNINGSFORSLAG MET4 H24

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

Oppgave 1

(a) Vi skal her teste nullhypotesen om lik varians:

$$H_0 : \sigma^2_{sykehus} = \sigma^2_{ikke} \quad \text{mot} \quad H_1 : \sigma^2_{sykehus} \neq \sigma^2_{ikke}$$

der $\sigma^2_{sykehus}$ og σ^2_{ikke} er populasjonsvariansen til personer som var innlagt på sykehus og ikke innlagt på sykehus.

Først må vi regne ut hva de empiriske standard avvikene er for de to populasjonene.

$$\text{Standard avvik} = \text{Standard error} \times \sqrt{N}$$
$$S_{sykehus} = 0.014 \times \sqrt{7774} = 1.23$$
$$S_{ikke} = 0.003 \times \sqrt{90049} = 0.90$$

Her er det lurte å sette den største estimerte variansen i telleren av testobservatoren slik at en kun trenger å sammenligne testobservatoren mot kritisk verdi i høyre hale av F-fordelingen (0.975 percentilen).

Testobservatoren tar verdien

$$F = \frac{S^2_{sykehus}}{S^2_{ikke}} = 1.23^2/0.90^2 = 1.86$$

Kritisk verdi er gitt ved 0.975 percentilen i en F-fordeling med $7774 - 1$ og $90049 - 1$ frihetsgrader. Her kan vi bruke følgende R-utskrift i Vedlegg 1:

```
qf(0.975, df1 = 7773, df2 = 90048)
```

```
## [1] 1.033076
```

Siden $F = 1.86 > k^{7773,90048}_{0.975} = 1.033076$**, kan vi forkaste nullhypotesen om lik varians.**

Til sensor: 2p for å sette opp riktig hypotese, 2p for riktig testobservator, 6p for riktig sammenligning med kritisk verdi og konklusjon. -2p for feil kritisk verdi. -3p dersom omregning til SD ikke er gjort og SE er brukt i stedet.

(b) Her er det mulig å argumentere for både en ensidig (med ulikhet begge veier) og en tosidig test. Vi undersøker her om de som har vært innlagt på sykehus har i snitt dårligere helse (lavere helseutfall). Vi gjennomfører derfor en ensidig test hvor vi tester om helseutfallene til dem som ikke ble innlagt på sykehus er høyere enn dem som ble lagt inn på sykehus.

$$H_0 : \mu_{sykehus} = \mu_{ikke} \quad \text{mot} \quad H_1 : \mu_{sykehus} < \mu_{ikke}$$

der $\mu_{sykehus}$ er forventet helseutfall for dem som ble innlagt på sykehus og μ_{ikke} er forventet helseutfall for dem som ikke ble innlagt på sykehus. Fra oppgave a) vet vi at det er rimelig å anta ulik varians. Vi bruker derfor følgende testobservator:

$$T = \frac{\bar{X}_{sykehus} - \bar{X}_{ikke}}{\sqrt{S^2_{sykehus}/n_{sykehus} + S^2_{ikke}/n_{ikke}}} = \frac{3.21 - 3.94}{\sqrt{1.23^2/7774 + 0.90^2/90043}} = -51.16$$

Eksakt antall frihetsgrader er gitt ved Welch's formel, men i dette tilfellet er antall observasjoner så stort at vi uansett får kritisk verdi svært nær -1.64. Siden $T < -1.64$ kan vi forkaste H_0. Det er forskjell i helseutfallene for de to gruppene, og det ser ut til at de som ikke ble innlagt på sykehus har bedre helseutfall med $3.94 - 3.21 = 0.73$ i snitt.

Til sensor: Det finnes en viss tolkningsfrihet i oppgaveteksten når det gjelder om testen skal være ensidig eller tosidig. Så lenge studenten begrunner valget sitt, er dette helt akseptabelt og bør gi full uttelling. 2p for et skikkelig argument for oppsett av hypotese. 2p for riktig testobservator, 6p for riktig sammenligning med kritisk verdi og riktig konklusjon. -2p for feil kritisk verdi. Kommentar om målt effekt kreves ikke. OBS: Ikke følgefeil for bruk av SE istedet for SD.

(c) Testen vi gjennomførte i oppgave **1b** viste oss at de som ble innlagt på sykehus hadde **dårligere** helseutfall enn dem som ikke ble innlagt på sykehus, som tilsynelatende kan lede til konklusjonen om at sykehus ikke gjør pasientene friskere. Problemet med denne analysen er at det ikke er tilfeldig hvem som blir innlagt på sykehus, hvor de som blir innlagt trolig er sykere enn de som ikke blir innlagt. Dette kalles en seleksjonsbias, og gjør at sammenligningen vi gjennomfører ikke kan svare på spørsmålet om sykehus gjør **disse** pasientene friskere. Dette betyr at gruppene vi sammenligner ikke ville hatt lik utvikling i helseutfall ved likt helsetilbud, og de er dermed ikke sammenlignbare.

Til sensor: Teksten over er veiledene.

Oppgave 2

(a) Ettersom dette forholdet mellom donasjoner og lønn er på log-log skala kan vi tolke at en 1% økning i lønn henger sammen med en β_1% økning i konsum. I dette tilfellet vil en ett prosent økning i husholdningsinntekt henge sammen med nesten en 0.8% økning i gaver til veldedige organisasjoner, så vi ser det er en positiv sammenheng mellom inntekt og donasjoner til veldedige organisasjoner som er signifikant på et 1% signifikans nivå.

Dette indikerer at de med mer lønn gir mer til veldedige organisasjoner, noe som kanskje ikke er så overaskende.

Til sensor: 3 Poeng dersom en kun poengterer en positiv sammenheng mellom lønn og donasjoner.

(b) Ettersom det er log-lin skala mellom lønn og dummien for det å være selvstendig næringsdrivende kan vi tolke at det å være selvstendig næringsdrivende henger sammen med $\beta_2 \times 100\% = 0.427 \times 100\% = 42.7\%$, **mer** i forbruk selv når en justerer for inntekt og andre kontrollvariabler. Denne effekten er signfikant på et 1% signifikans nivå.

Ettersom denne beregningen er en tilnærming, kan den eksakte økningen i forbruk beregnes ved å bruke den eksakte formelen:

$$(e^{\beta} - 1) \times 100\% = (e^{0.427} - 1) \times 100\% = 53\%$$

Vi ser at når koffisienten er såpass stor er approksimeringen mindre presis.

Dette indikerer at selvstendig næringsdrivende i det minste bruker mer av sin oppgitte inntekt på veldedige donasjoner og kan være et tegn på den faktiske inntekten er større enn hva som er oppgitt.

Hvis vi legger merke til at regresjonslikningen inkluderer variabelen $SExMale$ vil dette indikerer om husholdningen er selvstendig næringsdrivende og hvor den selvstendige er mann. Derfor vil koffisienten til SE faktisk reflektere forbruket hos kvinnlig selvstendig næringsdrivende.

Til sensor: Her godtar vi begge metodene for utregning. Det er ikke lagt vekt på approksimeringen i kurset.

(c) Formel for konfidensintervall for en regresjonskoeffisient β er $\widehat{\beta} \pm t_{1-\alpha/2} \cdot S(\widehat{\beta})$, der $S(\widehat{\beta})$ er det estimerte standardavviket til koeffisientestimatet. I dette tilfellet har vi såpass mange observasjoner at vi kan sette t-kvantilen $t_{1-0.05/2}$ til 1.96 for et 95% konfidensitervall, og får at konfidensintervallet for β_1 er

$$[0.777 \pm 1.96 \cdot 0.004] = [\mathbf{0.76916, 0.78484}].$$

Til sensor: Ok om ikke 4-desimalen er riktig eller om de har avrundet litt feil. Ellers en rimelig "rett-eller-gal" oppgave.

(d) Dette er en regresjonsmodell for panel data (de samme husholdningene målt på flere tidspunkt) med husholdningsfaste effekter indikert ved variabelen η_i.

Hvis det er slik at selvstendig næringdrivende har ulike preferanser for veldedighet enn lønnstakere som ikke endrer seg over tid (stabile preferanser) vil vi nå justere for disse forskjellene. Dermed har vi nå kontrollert for forskjeller i stabile preferanser for veldedighet mellom de to ulike gruppene, og dermed vil ikke forskjeller i disse preferansene lengre være en utelatt variabel. Vi er dermed et steg nærmere for å måle "overkonsum" hos selvstendig næringsdrivende som kan skyldes underrapportert inntekt og skatteunndragelse.

Til sensor: Besvarelsen over er veiledende.

(e) Modellen følger av å trekke fra tidsgjennomsnittet av modell (1) hver side i modell (1) (Se video 04-11 Estimering av faste effekter):

$$\begin{aligned}\widetilde{\log(Y_{it})} = \log(Y_{it}) - \overline{\log(Y_i)} &= \beta_0 - \beta_0 + \beta_1 \left(\log(\texttt{Wage}_{it}) - \overline{\log(\texttt{Wage}_i)}\right) + \beta_2 \left(\texttt{SE}_{it} - \overline{\texttt{SE}}_i\right) \\ &\quad + \beta_3 \left(\texttt{SE}_{it} \times \texttt{Male}_{it} - \overline{(\texttt{SE} \times \texttt{Male})_i}\right) + \eta_i - \eta_i + \epsilon_{it} - \bar{\epsilon}_i \\ &= \beta_1 \widetilde{\log(\texttt{Wage}_{it})} + \beta_2 \widetilde{\texttt{SE}_{it}} + \beta_3 \texttt{SE}_{it} \widetilde{\times \texttt{Male}_{it}} + \beta \widetilde{\texttt{X}_{it}} + \widetilde{\epsilon_{it}}\end{aligned}$$

Her vil tidsgjennomsnittet av η_i bare være η_i siden dette leddet ikke varierer med tiden. Tilsvarende er tidsgjennomsnittet av konstanten β_0 bare β_0. Derfor kanselerer disse to størrelsene seg i modellen.

Dersom alle husholdningene enten er lønnstakere ($SE_{it} = 0$) eller selvstendig næringsdrivende ($SE_{it} = 1$) gjennom alle 5 årene, så har vi heller ikke noe variasjon i SE_{it} over tid og $(\texttt{SE}_{it} - \overline{\texttt{SE}}_i) = 0$ i ligningen over. Det er da ikke mulig å estimere β_2 (og forsåvidt også β_3 siden kjønn stort sett er tidsinvariant) på denne måten. Derfor er vi avhengig av å ha husholdninger som bytter status mellom selvstendig næringsdrivende og lønnstakere i løpet av perioden i datasettet for å kunne estimere β_2. Disse husholdningene gir informasjon om hvordan endringer i arbeidsstatus påvirker konsumet, uavhengig av de faste egenskapene til husholdningen.

Til sensor: 5 poeng for å vise modell. 5 poeng for å poengtere hvilke husholdninger som må være i datasettet.

Oppgave 3

(a) La $p = P(\text{unemployed\_2007} \mid \texttt{race}, \texttt{age}, \texttt{age2}, \texttt{earnwke}, \texttt{married}, \texttt{female}, \texttt{educ})$. Den estimerte modellen er da:

$$p = \frac{\exp(z)}{1+\exp(z)}, \quad z = \widehat{\beta}_0 + \widehat{\beta}_1 \texttt{race2} + \widehat{\beta}_2 \texttt{race3} + \widehat{\beta}_3 \texttt{age} + \widehat{\beta}_4 \texttt{age}^2 + \widehat{\beta}_5 \texttt{earnwke} + \widehat{\beta}_6 \texttt{married}$$
$$+\widehat{\beta}_7 \texttt{female} + \widehat{\beta}_8 \texttt{educ\_hs} + \widehat{\beta}_9 \texttt{educ\_somecol} + \widehat{\beta}_{10} \texttt{educ\_aa} + \widehat{\beta}_{11} \texttt{educ\_bac} + \widehat{\beta}_{12} \texttt{educ\_adv}$$

$$= -1.522 + 0.346\texttt{race2} - 0.332\texttt{race3} - 0.037\texttt{age} + 0.0004\texttt{age}^2 - 0.001\texttt{earnwke} - 1.377\texttt{married}$$
$$-0.642\texttt{female} - 0.324\texttt{educ\_hs} - 0.652\texttt{educ\_somecol} - 0.176\texttt{educ\_aa} - 1.072\texttt{educ\_bac} - 0.700\texttt{educ\_adv}$$

Tolkning av `race2`: Dersom de andre forklaringsvariablene er uendret vil det at personen er svart være assosiert med at oddsen for arbeidsledighet øker med en multiplikativ faktor $e^{0.346} = 1.413$ (en økning på ca. 41%) sammenlignet med en som er hvit (referanse kategorien). Men denne sammenhengen er ikke statistisk signifikant forskjellig fra 0 og vi konkluderer med at hudfarge ikke påvirker sannsynligheten for å bli arbeidsledig.

Til sensor: Formulering med logit-funksjonen er også helt greit. Typisk feil her er at kandidaten skriver opp en log-lineær regresjonsmodell og da gir vi 0p. Noen vil også svare at "en enhets økning i race2. . . .": - 3p for det.

(b) Vi regner først ut den lineære komponenten:

$$z = \widehat{\beta}_0 + \widehat{\beta}_1 \texttt{race2} + \widehat{\beta}_2 \texttt{race3} + \widehat{\beta}_3 \texttt{age} + \widehat{\beta}_4 \texttt{age}^2 + \widehat{\beta}_5 \texttt{earnwke} + \widehat{\beta}_6 \texttt{married}$$
$$+\widehat{\beta}_7 \texttt{female} + \widehat{\beta}_8 \texttt{educ\_hs} + \widehat{\beta}_9 \texttt{educ\_somecol} + \widehat{\beta}_{10} \texttt{educ\_aa} + \widehat{\beta}_{11} \texttt{educ\_bac} + \widehat{\beta}_{12} \texttt{educ\_adv}$$

$$z = -1.522 + 0.346 \cdot 1 - 0.037 \cdot 27 + 0.0004 \cdot 27^2 - 0.001 \cdot 400 - 1.377 \cdot 1$$
$$-0.642 \cdot 1 - 1.072 \cdot 1 = -5.3744$$

Da angir modellen følgende sannsynligheten for å at personen beskrevet blir arbeidsledig:

$$p = \exp(-5.3744)/(1+\exp(-5.3744)) = 0.0046$$

Altså er det veldig lav sannsynelighet for at hun blir arbeidsledig, ifølge modellen under en halv prosent sannsynelighet.

Til sensor: Dukker fort opp noen numeriske regnefeil her. Minus 3p for det dersom alt annet er riktig.

(c) Den største forskjellen mellom normale tider og finanskrisen ser ut til å være blant arbeidstakere med og uten utdanning. Under finanskrisen hadde alle utdanningsnivåer en beskyttende effekt mot fremtidig arbeidsledighet sammenlignet med de som ikke hadde fullført videregående skole. Videre ser vi at alder har en statistisk signifikant og ikke-lineær sammenheng med arbeidsledighet under finanskrisen, der sammenhengen er konveks (fordi age < 0 og age$^2 > 0$). Det å være gift og kvinne gir fortsatt en viss beskyttelse mot arbeidsledighet, men denne sammenhengen er noe svakere enn i normale perioder. Dette kan tyde på at finanskrisen reduserte beskyttelsen som faktorer som sivilstatus tidligere ga, sannsynligvis på grunn av omfattende jobbkutt som rammet bredere deler av befolkningen. Likevel var utdanning en viktigere beskyttelsesfaktor under krisetider.

Til sensor: Besvarelsen over er veiledende.