LØSNINGSFORSLAG MET4 H21

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

Oppgave 1

(a) La μ_1 og μ_2 være forventet antall barn i de to gruppene henholdsvis. Vi skal gjøre to t-tester for

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

en for tidspunktet rett før kvinnene i den ene gruppen mister jobben, og en for tidspunktet 11 år etter den første gruppen mister jobben. Det fremgår av oppgaven at vi er generelt er interessert i sammenhengen mellom fertilitet og arbeidsledighet, så det er ikke noen spesiell grunn til å gjøre ensidige tester. Videre ser vi at de empiriske standardavvikene til de to gruppene er forholdsvis like mellom de to gruppene ved begge tidspunkt, så vi antar lik varians i begge testene. Ved det første tidspunktet får vi følgende testobservator:

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{1.124 - 1.104}{\sqrt{2.328\left(\frac{1}{7011} + \frac{1}{249894}\right)}} = 1.08,$$

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

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(7011 - 1) \cdot 1.483^2 + (249894 - 1) \cdot 1.527^2}{7011 + 249894 - 2} = 2.328.$$

Med et så stort antall observasjoner, er kritisk verdi for en tosidig t-test 1.96. Testobservatoren er lik 1.08, så **vi forkaster ikke nullhypotesen om forskjelling forventet antall barn i de to gruppene ved det første tidspunktet.**

Vi gjør den samme operasjonen for det andre tidspunktet, og får:

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(7011 - 1) \cdot 0.850^2 + (249894 - 1) \cdot 0.791^2}{7011 + 249894 - 2} = 0.628,$$

og

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{1.580 - 1.611}{\sqrt{0.628\left(\frac{1}{7011} + \frac{1}{249894}\right)}} = -3.23,$$

som gir en testobservator som er klart utenfor forkastningsgrensen på -1.96. **Vi forkaster nullhypotesen om at de to gruppene har samme forventet antall barn 11 år etter at den ene gruppen har mistet jobben**.

De to gruppene har ikke signifikant forskjellig fertilitet mens alle var i jobb, men 11 år etter at den ene gruppen mistet jobben har de fått signifikant færre barn enn gruppen som ikke mistet jobben.

(b) Vi kan godt tenke oss at det finnes personlige egenskaper som påvirker både evnen til å holde på en jobb, og antall barn man får. Derfor burde vi strengt tatt gjennomført et kontrollert eksperiment der vi gav en tilfeldig valgt gruppe sparken, mens den andre gruppen fikk beholde jobben. Da kan vi med stor sikkerhet tilskrive en eventuell forskjell i antall barn den til kausale effekten av å miste jobben. Det er selvsagt umulig å gjøre i praksis, men i denne studien har man fått til det kanskje nest beste: et naturlig eksperiment.

I og med at gruppen som mistet jobben havnet der først og fremst fordi *hele arbeidsplassen* ble lagt ned, så kan vi være rimelig sikker på at det ikke er personlige egenskaper ved den enkelte som gjør at man er i den ene eller andre gruppen. Vi kan i alle fall til en viss grad tolke forskjellen i fertilitet som vi fant i oppgave (a) som et kausalt resultat av at den ene gruppen mistet jobben.

Dersom vi aksepterer dette ser vi at hvert individ som mistet jobben har i gjennomsnitt fått $1.611 - 1.58 = 0.031$ færre barn som følge av nedleggelsene, **totalt $0.031 \cdot 7011 \approx 217$ barn**.

Det er likevel ikke perfekt. Vi kan tenke oss at det har vært for eksempel systematiske regionale forskjeller i hvilke fabrikker som er lagt ned, som også kan henge sammen med fertilitetstall.

(c) Vi ser fra figuren at det først og fremst er blant kvinner at arbeidsløshet fører til færre barn. Hos menn er det en svak nedgang i forventet antall barn i gruppen som mistet jobben, men denne effekten er ikke statistisk signifikant som vi ser ved at alle konfidensinterallene inneholder null.

Oppgave 2

(a) Vi kan sjekke to ting ut fra tabellen i oppgavesettet; symmetri og ekstremverdier. Variablene `fixed` og `interestBurden` har begge en viss forskjell mellom gjennomsnitt og median, og deres gjennomsnitt ligger begge mye nærmere 75-prosentilen enn 25-prosentilen, som tyder på at de har assymmetriske fordelinger, noe som ikke er forenelig med normalitet. Disse to variablene har også svært høye maksimumsverdier, noe som også gjelder variabelen `profitability`, som også tyder på at de ikke er normalfordelte. For normalfordelte data er det svært skjelden å se observasjoner som er mer enn 4-5 standardavvik fra forventningsverdien.

(b) Alle variablene er forhold mellom to pengeverdier og dermed enhetsløse. Vi ser fra regresjonsutskriften at alle variablene bortsett fra `industrial` har signifikant negative regresjonskoeffisienter, som tyder på at en økning i disse variablene henger sammen med lavere lønnsomhet for selskapene i utvalget vårt. Den siste variabelen er ikke statistisk signifikant forskjellig fra null. I denne modellen er $R^2 = 0.511$, som betyr at omtrent halvparten av variasjonen i lønnsomhet er forklart ved hjelp av den lineære modellen og de fem forklaringsvariablene.

(c) I residualplottet er det et tydelig tegn på heteroskedastisitet. Histogrammet til residualene ser også ut til å være noe forskjellig fra normalfordelingen, ved at det har en spissere topp og muligens for mye data i halene. Det ser vi tydeligere i QQ-plottet. I plottet for Cooks avstand så ser vi at det er to relativt innflytelsesrike observasjoner i datasettet, som kan burde dobbeltsjekke og muligens fjernet. Vi har ikke informasjon om autokorrelasjon, men det er sannsynligvis ikke relevant uansett siden observasjonene er gjort samtidig.

Den lineære modellen gir ikke en spesielt god tilpasning til dette datasettet.

(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

$$[4.838 \pm 1.96 \cdot 4.864] = \mathbf{[-4.70, 14.37]}.$$

Dersom den lineære modellen er sann, og dersom vi hadde hatt muligheten til å trekke nye datasett fra den samme populasjonen, ville disse konfidensintervallene inneholde den sanne verdien β i 95% av tilfellene. Dette konfidensintervallet inneholder 0, som betyr at estimatet ikke er statistisk signifikant forskjellig fra null.

(e) Vi finner gjennomsnittsverdien for de ulike variablene i tabellen med deskriptiv statistikk og setter inn i den estimerte regresjonsmodellen:

$\widehat{Y} = 13.412 - 13.93 \cdot 0.74 - 0.535 \cdot 1.08 - 5.906 \cdot 0.03 + 4.838 \cdot 0.01 - 0.035 \cdot 8.43 = \mathbf{2.10}$.

Et konfidensintervall rundt denne prediksjonen representerer usikkerheten for den forventede (populasjons gjennomsnittlige) lønnsomheten til en gjennomsnittlig bedrift (siden vi putter inn gjennomsnittlige

forklaringsvariabler), og dette intervallet vil krympe mot denne forventningsverdien når utvalget øker. Et prediksjonsintervall representerer usikkerheten for en prediksjon av lønnsomheten til en ny bedrift som har gjennomsnittlige forklaringsvariabler. Prediksjonsintervallet inneholder informasjon om usikkerheten fra estimering av modellen (altså det som konfidensintervallet sier noe om), *og* usikkerhet om hvilken verdi feilleddet vil ta for denne bedriften, som er en usikkerhet som alltid vil være der uansett hvor mye data vi har.

Prediksjonsintervallet er dermed alltid større enn konfidensintervallet, og det vil ikke krympe mot et punkt selv med mye data.

(f) Den lineære modellen ser ut til å passe litt bedre til det transformerte datasettet. For det første ser vi at justert R^2 er gått noe opp fra modell (1) til modell (2). For det andre ser residualplottene noe bedre ut; det er mindre heteroskedastisitet, og residualene er nærmere normalfordelingen, spesielt i høyre hale. Det ser fortsatt ut til å være en særlig innflytelsesrik observasjon.

(g) Den lineære regresjonsmodellen passer klart bedre når vi tar ut de mest ekstreme observasjonene. Justert R^2 er enda større, og residualplottene ser klart bedre ut. Det tyder på at det er et relativt lavt antall observasjoner som gjør at det er mønstre i residualene til modell (1) og til en viss grad modell (2).

Selve koeffisientestimatene forandrer seg lite på tvers av modellene, men med ett viktig unntak: den estimerte koeffisienten til `industrial` er klart positiv, og signifikant forskjellig fra null i den siste modellen. Det tyder på at dette faktisk *er* en viktig forklaringsvariabel for lønnsomhet, men at denne effekten ikke er synlig når vi tilpasser en lineær modell til hele datasettet.

Det ser ut til at vi har laget en modell som passer godt til "normale" selskaper. Et naturlig neste steg kan være å lage en bedre modell som gjelder for hele datasettet. Kanskje vi kan finne en forklaring på hvorfor enkelte bedrifter er så forskjellige fra resten av utvalget? Det kan for eksempel være en regional forklaring, eller det kan være snakk om bedrifter i en bestemt industriklasse. Dette kan i så fall kodes som en dummyvariabel og inkluderes i modellen.

Oppgave 3

(a) Vi skriver for enkelhets skyld $p = P(\text{Defekt} \mid \text{Vektindeks} = x_1, \text{Varmekapasitet} = x_2)$. Den estimerte modellen er da:

$$p = \frac{\exp(z)}{1+\exp(z)}, \quad z = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 4.142 - 1.1295 x_1 - 3.9108 x_2.$$

Evt.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = 4.142 - 1.1295 x_1 - 3.9108 x_2.$$

(b) For en generell logistisk regresjonsmodell med responsvariabel Y og to forklaringsvariabler, klassifiserer vi $Y = 1$ dersom $\widehat{p} > \delta$, og $Y = 0$ dersom $\widehat{p} < \delta$, for en eller annen verdi δ. Grensen mellom disse to hendelsene har vi ved $\widehat{p} = \delta$, og i følge den logistiske regresjonsmodellen er $\widehat{p} = \delta$ for alle x_1 og x_2 som er slik at

$$\log\left(\frac{\delta}{1-\delta}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$$

som definerer en rett linje i x_1-x_2-koordinatsystemet siden venstre side av ligningen over bare er en konstant.

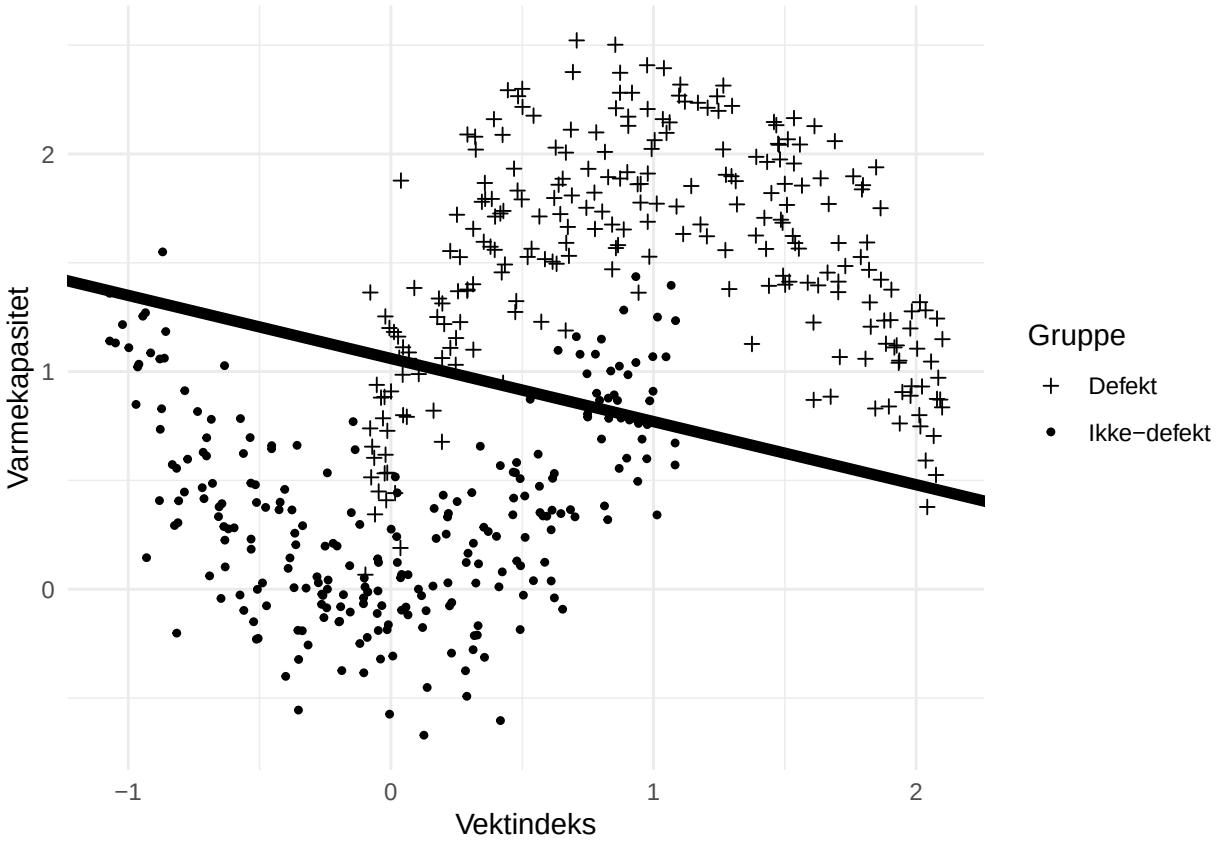
I dette tilfellet er $\delta = 0.5$ slik at vi klassifiserer en skrue som defekt dersom $\widehat{p} > 0.5$, og som ikke-defekt dersom $\widehat{p} < 0.5$, slik at linja da er gitt ved

$$\begin{aligned}
\log\left(\frac{0.5}{1-0.5}\right) &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 \\
\log(1) &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 \\
0 &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2,
\end{aligned}$$

Hvis vi vil tegne linjen er det nyttig å løse med hensyn på x_2:

$$x_2 = -\frac{\widehat{\beta}_0}{\widehat{\beta}_2} - \frac{\widehat{\beta}_1}{\widehat{\beta}_2} x_1 = -\frac{4.142}{-3.9108} - \frac{-1.1295}{-3.9108} x_1 = 1.06 - 0.29 x_1,$$

som gir oss skjæringspunktet med x_2-aksen og stigningstallet. Linjen ser slik ut i det opprinnelige koordinatsystemet (her er det nok å skissere selve linjen i koordinatsystemet uten å nødvendigvis tegne inn observasjonene):



(c) Vi ser fra spredningsplottet av observasjonene i oppgaveteksten at det ikke er mulig å separere de to gruppene særlig godt ved hjelp av en rett linje. Uansett hvordan vi legger linjen så vil det være grupper av skruer som blir systematisk feilklassifisert (og det skal være mulig fra forrige oppgave å se at det gjelder de tyngste ikke-defekte, og de letteste defekte skruene i dette tilfellet).

Det kan da være nyttig å bruke en klassifiseringsregel som ikke lager en grense med en bestemt form, for eksempel kNN. Med et godt valg av k vil vi her kunne få en S-formet klassifiseringsgrense mellom de to gruppene slik at de blir separert med en større presisjon enn det vi får til med logistisk regresjon.