BOKMÅL

NHH

HJEMMEEKSAMEN MET4

Høst 2021

Dato: 26. november 2021

Tidsrom: 09:00 - 12:00

Antall timer: 3

Foreleser/emneansvarlig kan kontaktes av eksamensvakt på telefon: *99385583*

TILLATTE HJELPEMIDLER:
Alle trykte/egenskrevne hjelpemidler, kalkulator.

Ordbok: én tospråklig ordbok tillatt.

Antall sider, inkludert forside og vedlegg: 9

Oppgave 1

Nyttårsaften 2018 holdt statsminister Erna Solberg en tale der hun blant annet ba nordmenn lage flere barn:

> *For å opprettholde folketallet må hver kvinne føde litt over to barn i gjennomsnitt. I dag er det tallet på 1.6. Da blir det relativt sett færre unge som skal bære en stadig tyngre velferdsstat på sine skuldre.*

Det er godt kjent at familieplanlegging henger sammen med økonomiske og samfunnsmessige forhold, men dersom samfunnet skal legge til rette for at det skal fødes flere barn må vi også undersøke relevante årsakssammenhenger. Det er ikke så lett, fordi vi for eksempel kan tenke oss at arbeidsledighet påvirker beslutningen om å få barn, samtidig som antall barn kan påvirke arbeidsdeltakelsen.

En finsk studie fra 2016[1] undersøker effekten av arbeidsledighet på fertilitet gjennom å følge et stort antall kvinner som i 1991 var ansatt på en arbeidsplass i privat sektor med mellom 5 og 1000 ansatte. Kvinnene ble så delt inn i to grupper:

- Gruppe 1: De som i perioden 1991-1993 mistet jobben som følge av at *hele arbeidsplassen ble lagt ned* (det vil si at alle ansatte mistet jobben samtidig).
- Gruppe 2: De som ikke mistet jobben i denne perioden.

Kvinner som sluttet i jobben av andre grunner er ikke med i denne studien. På denne måten kan vi argumentere for at kvinnene i gruppe 1 ikke mistet jobben på grunn av sine egne personlige egenskaper.

Før kvinnene i gruppe 1 mistet jobben kunne vi observere følgende deskriptive statistikk over antall barn i de to gruppene.

	Gjennomsnittlig antall barn	Standardavvik	Antall observasjoner
Gruppe 1	1.124	1.483	7011
Gruppe 2	1.104	1.527	249894

Tabell 1: Fertilitetstall for et utvalg finske kvinner **før** gruppe 1 mister jobben som følge av at arbeidsplassen der de jobber legges ned.

I 2004, dvs. ca 11 år etter kvinnene i gruppe 1 mistet jobben, observerer vi følgende deskriptive statistikk over antall barn i de to gruppene:

	Gjennomsnittlig antall barn	Standardavvik	Antall observasjoner
Gruppe 1	1.580	0.850	7011
Gruppe 2	1.611	0.791	249894

Tabell 2: Fertilitetstall for det samme utvalget finske kvinner **11 år etter** gruppe 1 mistet jobben som følge av at arbeidsplassen der de jobber ble lagt ned.

[1] "The Effect of Job Displacement on Couples' Fertility Decisions". Kristiina Huttunen og Jenni Kellokumpu, *Journal of Labour Economics* (2016). Standardavvikene som brukes i denne oppgaven er ikke oppgitt i studien, men er satt til verdier som gir samme konklusjoner.

(a) Test om forventet antall barn er likt i de to gruppene, både før gruppe 1 mister jobben, og 11 år etter gruppe 1 mister jobben. Begrunn eventuelle valg du gjør underveis (du trenger ikke utføre test for lik varians). Kommenter resultatene.

Professor i økonomi ved Universitetet i Oslo, Andreas Moxnes, kommenterer denne studien i *Morgenbladet Nr.11 (2021)*. Han skriver at "*... økonomi er avgjørende for det som skjer på soverommet*".

(b) Kan vi bruke analysen over til å argumentere for at det er en kausal sammenheng fra arbeidsledighet til fertilitet? I så fall, estimer hvor mange færre barn som ble født i Finland som følge av at arbeidsplassene til kvinnene i gruppe 1 ble lagt ned.

I den aktuelle studien ble to tilsvarende grupper med menn fulgt gjennom den samme perioden. Differansen i antall barn mellom de to gruppene ble estimert hvert år i 11 år etter at gruppe 1 mistet jobben, samt i tre år *før* gruppe 1 mistet jobben (vist som negative tall på x-aksen i figuren under). I figuren under er disse differansene vist som funksjon av tid både for kvinner og for menn. Gråfargen er 95% konfidensintervaller for estimatene.

(c) Diskuter kort hva vi lærer av denne figuren.

Oppgave 2

Du er ansatt i et stort investeringsfond, og blir bedt om å undersøke om det er bestemte egenskaper ved selskaper som forklarer variasjon i lønnsomheten. I en innledende fase har du hentet inn data fra 1405 italienske selskaper fra et enkelt regnskapsår[2]. For hvert selskap har du data på følgende variabler:

Tabell 3: Variabelbeskrivelser for italienske regnskapsdata.

`profitability`	Lønnsomhet, målt som avkastning delt på omsetning
`labourShare`	Lønnskostnad delt på profitt
`fixed`	Verdien av fysiske driftsmidler delt på profitt
`intangible`	Verdien av immatrielle eiendeler som andel av verdien av alle eiendeler
`industrial`	Verdien av industrielt utstyr som andel av verdien av alle eiendeler
`interestBurden`	Renteutgifter som andel av netto kapitalbeholdning

I tabellen under har vi grunnleggende deskriptiv statistikk for variablene i datasettet:

Tabell 4: Deskriptiv statistikk for variablene i regnskapsdatasettet

	Gj.snitt	Std.avvik	Min	25% Kv.	Median	75% Kv.	Max
`profitability`	2.10	5.30	-39.19	-0.32	2.23	4.48	27.25
`labourShare`	0.74	0.26	-0.89	0.58	0.74	0.89	1.95
`fixed`	1.08	1.57	-2.74	0.19	0.52	1.30	18.15
`intangible`	0.03	0.06	0.00	0.00	0.01	0.03	0.76
`industrial`	0.01	0.02	0.00	0.00	0.00	0.01	0.19
`interestBurden`	8.43	9.72	1.18	2.94	5.11	9.59	84.84

(a) Vurder kort i hvilken grad variablene i datasettet er normalfordelt.

Vi kjører en lineær regresjon med `profitability` som responsvariabel, og får resultatet som er gitt i kolonne **(1)** i vedlegg 1.

(b) Gi en kort fortolkning av regresjonsutskriften i kolonne (1) i vedlegg 1.

Du finner diagnoseplott for denne estimerte regresjonsmodellen i figurkolonne **(1)** i Vedlegg 2.

(c) Vurder kort om den estimerte regresjonsmodellen (1) oppfyller forutsetningene for lineær regresjon.

(d) Lag et 95% konfidensintervall for koeffisienten til variabelen `industrial` i regresjonsmodell (1) i vedlegg 1. Hva er fortolkningen av dette konfidensintervallet?

(e) Bruk modell (1) i vedlegg 1 til å predikere lønnsomheten til en bedrift med gjennomsnittlige verdier av de fem forklarlingsvariablene. Forklar så forskjellen på et *konfidensintervall* og et *prediksjonsintervall* for dette punktestimatet (du skal ikke regne ut intervallene).

[2]Datasettet er presentert og analysert i "The analysis of transformations for profit-and-loss data" av Atkinson, Riani, og Corbellin,*Journal of the Royal Statistical Society: Series C* (2020).

Vi ønsker å lage en bedre modell for sammenhengen mellom lønnsomheten til et selskap og de fem forklaringsvariablene som er listet opp i Tabell 3. Et klassisk triks for å få til en bedre tilpasning med den lineære regresjonsmodellen er å transformere responsvariabelen med den såkalte Box-Cox-transformasjonen[3]. Uheldigvis kan denne transformasjonen bare brukes på positive tall, noe som ikke vil fungere for oss siden flere selskaper har negativ lønnsomhet (se Tabell 4). Dette problemet kan løses ved å heller bruke Yeo-Johnson-transformasjonen[4], som også kan transformere negative tall.

I kolonne **(2)** i vedlegg 1 har vi kjørt den samme regresjonen på nytt, men der responsvariabelen `profitability` er Yeo-Johnson-transformert (vi skriver ikke opp selve transformasjonen her). I kolonne **(2)** i vedlegg 2 finner du tilhørende residualplott.

(f) Vurder kort om den lineære regresjonsmodellen passer bedre til det transformerte datasettet.

Du tar utgangspunkt i det transformerte datasettet og bruker en automatisk algoritme til å identifisere uteliggere. Algoritmen plukker opp 143 ekstreme observasjoner som du fjerner fra datasettet før du tilpasser regresjonsmodellen på nytt. Den estimerte modellen er presentert i kolonne **(3)** i vedlegg 1, med tilhørende residualplott i kolonne **(3)** i vedlegg 2.

(g) Hvilken effekt har det at vi fjerner uteliggerene fra datasettet? Hva forteller dette resultatet oss om sammenhengen mellom forklaringsvariablene og responsvariabelen i denne konteksten?

Oppgave 3

En fabrikk som produserer små skruer for bruk i elektroniske komponenter har den siste tiden fått tilbakemeldinger fra kundene sine om at et stort antall skruer blir ødelagt når de monteres. Man har etter grundige undersøkelser funnet frem til det sannsynlige problemet. For å herde metallegeringen, må skruene varmes opp til en kritisk temperatur for så å kjøles ned med en helt bestemt hastighet. Ovnen som gjør denne jobben ser ut til å ha problemer med å holde korrekt temperatur, slik at herdingen av og til ikke blir fullført.

Uheldigvis så er det ikke mulig å se på skruene om de er defekte eller ikke, og den problematiske ovnen har ikke en loggefunksjon som kan avsløre om herdeprosessen ikke ble fullført. Fabrikken trenger da en måte å skille mellom ikke-defekte og defekte skruer mens de venter på reparasjon av den problematiske ovnen.

En ingeniør har tatt for seg et utvalg skruer med kjent status, halvparten defekte og halvparten ikke-defekte, og oppdaget følgende sammenhenger:

- Defekte skruer har en tendens til å være noe tyngre enn ikke-defekte skruer.
- Man kan måle varmekapasiteten til skruene ved å se hvor fort de når romtemperatur etter herdingen. Det er en kjent sammenheng at varmekapasiteten er lavest ved gjennomsnittsvekt, og at den går noe opp for lettere og tyngre skruer. På grunn av den ufullstendige herdingen er denne sammenhengen motsatt for defekte skruer; varmekapasiteten er *størst* ved normalvekt, og så *minker* den for lettere og tyngre skruer.

[3] "An analysis of transformations", Box og Cox, *Journal of the Royal Statistical Association, Series B* (1964).
[4] "A New Family of Power Transformations to Improve Normality or Symmetry", Yeo og Johnson, *Biometrika* (2000).

Det er fort gjort å måle disse to variablene, så ingeniøren foreslår å lage en statistisk modell som kan brukes til å klassifisere skruene som enten defekt eller ikke-defekt før de pakkes og sendes ut.

I vedlegg 3 finner du et spredningsplott for skruene i utvalget til ingeniøren, med en standardisert vektindeks på x-aksen og standardisert varmekapasitet på y-aksen. Ikke-defekte skruer er tegnet inn som prikker ($\bullet$), og defekte skruer er tegnet inn som pluss-tegn ($+$)[5]. I det samme vedlegget finner du utskriften for en logistisk regresjonsmodell for sammenhengen mellom sannsynligheten for at en skrue er ikke-defekt, og de to forklaringsvariablene.

(a) Skriv opp den logistiske regresjonsmodellen som er estimert i vedlegg 3.

Ingeniøren bruker den estimerte logistiske regresjonsmodellen i vedlegg 3 til å klassifisere nye skruer. Terskelverdien settes til 0.5, som betyr at vi velger å klassifissere en skrue som defekt dersom den estimerte sannsynligheten for at den er defekt er større enn 0.5. Det betyr at den estimerte modellen definerer to områder i koordinatsystemet vist i vedlegg 3; Et område der nye skruer blir predikert til å være defekte, og et område der nye skruer blir definert til å være ikke-defekte.

(b) Vis at grensen mellom disse to områdene er en rett linje når vi bruker logistisk regresjon. Regn ut formelen for denne linjen ved bruk at den estimerte logistiske regresjonsmodellen som er vist i vedlegg 3, og tegn den inn i et koordinatsystem.

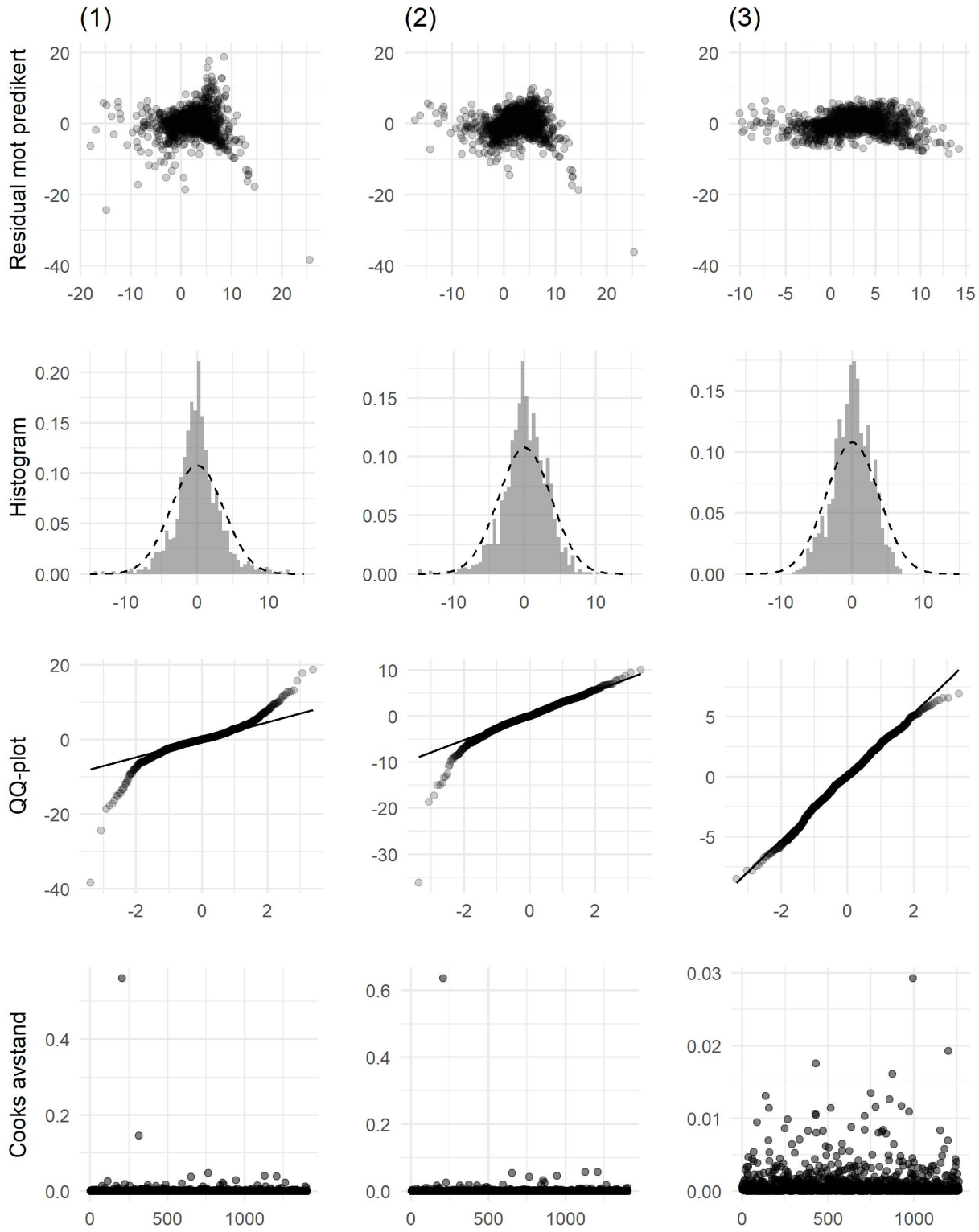
(c) Forklar hvorfor kNN (*k nearest neighbours*) sannsynligvis er en bedre klassifiseringsmodell i dette tilfellet.

[5]Datasettet er simulert, og opptrer i en annen sammenheng i "Pairwise local Fisher and naive Bayes: Improving two standard discriminants". Otneim, Jullum og Tjøstheim. *Journal of Econometrics* (2020)

Vedlegg 1: Estimerte regresjonsmodeller

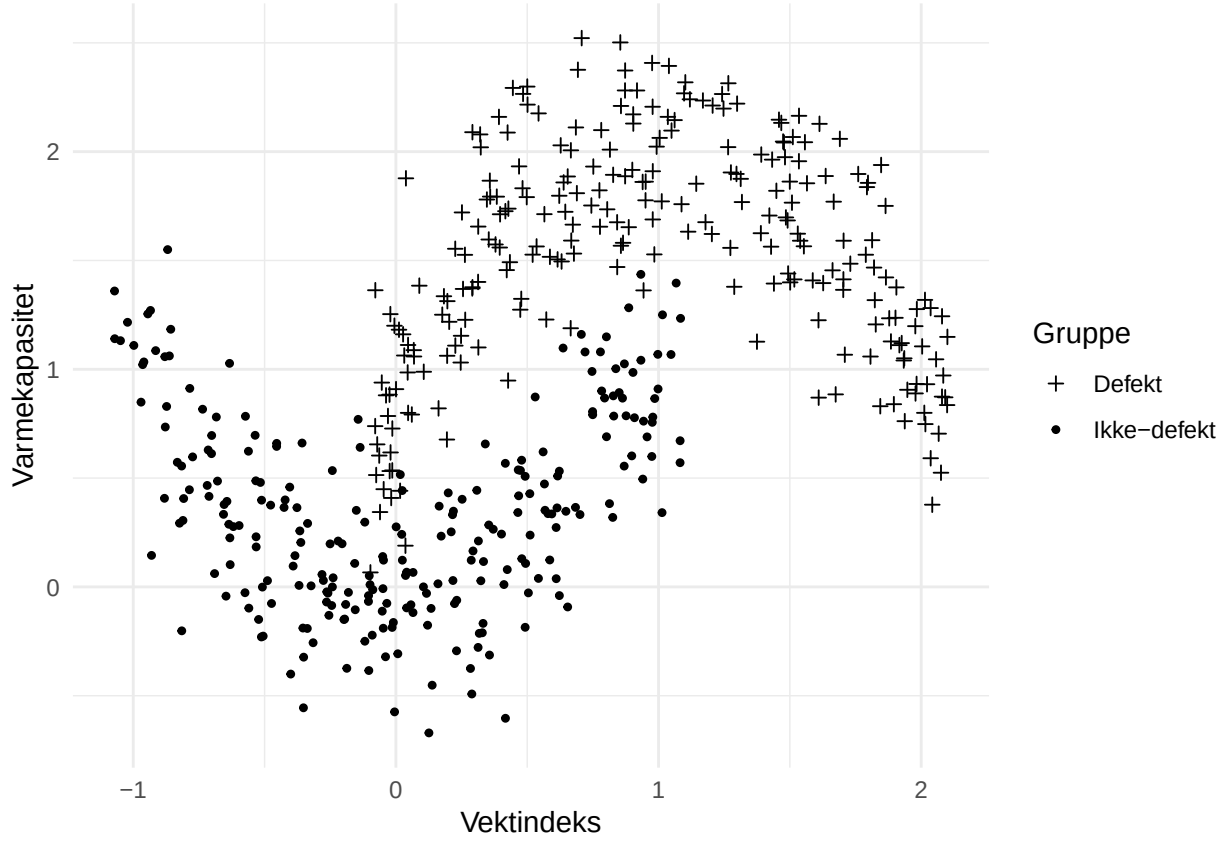
	Dependent variable:		
	profitability		
	(1)	(2)	(3)
labourShare	−13.930***	−13.636***	−15.406***
	(0.389)	(0.345)	(0.332)
fixed	−0.535***	−0.505***	−0.673***
	(0.064)	(0.057)	(0.069)
intangible	−5.906***	−6.305***	−5.155***
	(1.658)	(1.468)	(1.984)
industrial	4.838	2.783	15.242***
	(4.864)	(4.308)	(5.893)
interestBurden	−0.035***	−0.023**	−0.027**
	(0.010)	(0.009)	(0.011)
Constant	13.412***	13.330***	14.866***
	(0.319)	(0.283)	(0.277)
Observations	1,405	1,405	1,262
R^2	0.511	0.559	0.646
Adjusted R^2	0.509	0.557	0.644
Residual Std. Error	3.709 (df = 1399)	3.285 (df = 1399)	2.652 (df = 1256)
F Statistic	292.612*** (df = 5; 1399)	354.153*** (df = 5; 1399)	457.766*** (df = 5; 1256)
Note:			*p<0.1; **p<0.05; ***p<0.01

Vedlegg 2: Residualplott



Kommentar: Residualplott for de tre estimerte regresjonsmodellene i Vedlegg 1. Hver kolonne med plott svarer til tilsvarende kolonne i regresjonstabellen. Øverst er residualene plottet mot predikert verdi. I raden under har vi plottet histogram for de observerte residualene sammen med normalfordelingen med samme forventningsverdi og varians som residualene. I rad tre finner du QQ-plot for de tre regresjonsmodellene, og den nederste raden av plott viser Cooks distanse for alle observasjonene i de tre estimerte regresjonsmodellene.

Vedlegg 3: Spredningsplott og logistisk regresjonsmodell




```
Call:
glm(formula = as.factor(population) ~ vektindeks + varmekapasitet,
    family = binomial(link = "logit"), data = x2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.83185  -0.28624  -0.00133   0.29579   2.28255

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      4.1420     0.4070  10.176  < 2e-16 ***
vektindeks      -1.1295     0.2002  -5.643 1.67e-08 ***
varmekapasitet  -3.9108     0.4007  -9.760  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 693.15  on 499  degrees of freedom
Residual deviance: 254.82  on 497  degrees of freedom
AIC: 260.82

Number of Fisher Scoring iterations: 6
```